An Introduction to Econometrics

Lecture notes

Jaap H. Abbring*

Department of Economics

The University of Chicago

First complete draft (v1.04)
March 8, 2001

Preface

These are my lecture notes for the Winter 2001 undergraduate econometrics course at the University of Chicago (Econ 210).

Some technical details are delegated to end notes for interested students. These are *not* required reading, and can be skipped without problems.

Comments and suggestions are most welcome. These notes are freshly written, in a fairly short amount of time, so I am particularly interested in any errors you may detect.

[*] I have benefitted from Jeffrey Campbell's econometrics lecture notes.

Contents

1 Introduction

Statistics studies analytical methods for uncovering regular relationships from experiments contaminated by "chance".

Example 1. We may conjecture that a particular coin is fair, in the sense that it ends with heads up with probability 1/2 if tossed. To study whether a coin is fair, we may toss it a number of times, say 100 times, and count the number of heads. Suppose we find heads 52 out of 100 times. If we have no a priori information on whether the coin is fair, it is intuitively clear that a good *estimate* of the probability that the coin ends with heads up is 52/100=0.52. Does this imply that the coin is not fair? Not necessarily: this depends on the precision of our estimate. As our tossing experiment is contaminated by chance, we could find a different number of heads each time we repeat the experiment and toss the coin 100 times. We will occasionally find less than 50 heads and occasionally more than 50 heads. Mathematical statistics provides a rigorous theory that allows us to determine the precision of our estimate of the probability of the outcome "head", and to test for the fairness of the coin.

Example 2. A related example is the prediction of the outcome of the presidential election by polls. If we would know the *population* of all votes cast in the election, we would know the outcome of the election (if we could count the votes without error). If we want to predict the outcome before all votes are counted, we can ask a *random sample* of voters exiting the polling stations whether they have voted for Gore or Bush. The experiment here is sampling a given number of votes from the population of all votes. Again, statistics provides methods to estimate the outcome and assessing the possible error in the estimated outcome based on the sample of votes. The results of such analyses are broadcast by news channels, as we have witnessed recently in the US.

Econometrics applies statistical methods to the analysis of economic phenomena. The existence of econometrics as a separate discipline is justified by the fact that straightforward application of statistical methods usually does not answer interesting economic questions. Unlike in some of the physical sciences, economic problems can rarely be studied in a fully controlled, experimental environment. In contrast, economists usually have to infer economic regularities from real world data. Economic theory can be used to provide the additional structure needed to analyze such data. Also, economic theory is usually applied to structure economic research questions and allow for a useful economic interpretation of results. We clarify this with some examples.

Example 3. A lot of research focuses on the disadvantaged position of African-Americans, in terms of wages, employment and education, in the US economy. One can conjecture that this is the result of discrimination against blacks. An economic model can give some substance to this conjecture. Human capital theory predicts that workers that are similar with respect to characteristics like ability, education and experience should be paid the same wages. So, economics seems to tell us that we should compare wages of blacks and whites that are similar with respect to these characteristics to see whether there is discrimination.

However, economics suggests there is more to this story. If discrimination would imply that blacks, as opposed to whites, do not receive the full return to schooling and work experience, they will invest less in schooling and work experience. This suggests that we should not just be interested in wage differentials between blacks and whites that are similar with respect to human capital variables, but also in the indirect earnings effects of the induced differences in human capital accumulation. In other words, the *unconditional*, or overall, wage differential between blacks and whites may be a better measure of the overall effect of discrimination than the *conditional* (on human capital characteristics) differential if the only source of human capital differences is discrimination.

This is a good example of how economic theory structures an empirical analysis of an economic problem. A similar story can be told for male-female wage differentials.

Example 4. A related issue is the comparison of wages between groups over time or space. The black-white (median) wage differential has narrowed from around 50% of white wages in the 1940s to around 30% in the 1980s (this is an example taken from professor Heckman's 2000 Nobel lecture). This seems to indicate that there has been an improvement in the economic position of African-Americans.

However, blacks have dropped out of the labor force, and therefore these wage statistics, at a much higher rate than whites over this period. It is intuitively clear that we somehow want to include these drop-outs in the comparison of the two groups if it is to say anything about the relative economic development of the groups. Statistics is agnostic about a correction for selection into employment. Economics suggests that individuals at the lower end of the wage distribution drop out of the labor force. This provides a statistical assumption that can be used to correct the results for selective drop-out. The correction wipes out the improvement in the relative position of African-Americans.

A related story can be told for comparison of wages across space, notably between the US and Europe.

Example 5. Another important research topic is the return to schooling, the effect of schooling on employment and wages. It is clearly important to know the return to schooling if you have to decide on investment in schooling yourself. Estimates of the return to schooling are also relevant to many public policy decisions.

Economic theory can guide us to a useful definition of the return. It could be a wage increase per extra year of schooling, or an earnings increase per year of schooling, etcetera. Having established which is the useful measure of the return to schooling, we face the problem of measuring it. Ideally, we may want to investigate this in a controlled experiment, in which we can randomly allocate different schooling levels to different individuals, and directly infer the effect on their earnings, etcetera. Obviously, we cannot do this, and we have to use real world data on actual schooling levels and earnings.

Now suppose that agents are heterogeneous with respect to ability. Assume that high ability individuals have relatively high returns to schooling, but also earn more at any given level of schooling than low returns individuals. Under some conditions, economic theory predicts that high ability individuals choose high schooling levels, and low ability individuals choose low schooling levels. If we compare the earnings or wages of low schooling and high schooling individuals, we do not just capture the return to schooling, but also the inherent differences in ability between these groups. The central problem here is again that we cannot control the “explanatory” variable of interest, schooling, as in a physics experiment. Instead, it is the outcome of choice. Economic theory can be used to further structure this problem.

Example 6. At various stages in recent history different countries have considered legalizing so called hard drugs (pre-WW-II *Opiumregie* in the Dutch East Indies, heroin and other drugs in present-day Europe). A major concern is that legalization would increase the use of the legalized substances, which is considered to be bad. Economics provides a framework for analyzing and evaluating this.

To keep things (ridiculously) simple, one could envision a simple static market for drugs. Prohibition decreases supply at given prices, and drives up prices and down demand in equilibrium. There are costs of implementing prohibition, *e.g.* the costs of enforcement and crime. Legalization reduces these costs, but also increases supply at given prices. This leads to a reduction in prices and an increase in demand, and therefore quantities, with the size of these effects depending on the elasticity of demand. It is not obvious that this is bad from an economic efficiency point of view. After all, the use of drugs and the subsequent addiction (in a dynamic model) could be outcomes of rational choice, weighting all pros and cons of consuming drugs (this is an issue studied by professor Becker and co-authors). However, there may be political reasons to be uncomfortable with an increase in drug use anyhow. So, the elasticity of the demand for drugs is a crucial parameter if we are concerned about the effects of legalization on substance use.

A statistical problem is that we can typically not directly experiment with demand under different prices. Instead, we only observe market outcomes, jointly determined by a demand and a supply relation between quantities and prices. This is an example of the problem of simultaneous equations. Obviously, as most of these markets are illegal, not much data and empirical analyses are available. We may however study data on the *Opiumregie* in the Dutch East Indies (roughly present-day Indonesia) in one of the problem sets.

Example 7. If stock markets are efficient, stock prices should reflect all available, relevant information. Economic theory suggest models of stock prices to test for efficiency. If we detect inefficiencies, we can exploit these and become rich (arbitrage). This is another example on how econometrics combines economic theory and statistics to formulate and analyze an interesting economic question. The kind of data used, *time series* of stock prices, is typical (although not unique) to economics (GDP, inflation, aggregate unemployment, money supply, interest rates). Econometricians have developed many techniques to deal with time series.

2 Quick review of probability and statistics

We start the course with a quick review of statistics. As mathematical statistics makes extensive use of *probability theory*, this includes a review of probability. The material in this review can be found in many introductionary probability and statistics texts. An easy-to-read introduction to probability theory is Ross (1998). A basic introduction to mathematical statistics is Wonnacott and Wonnacott (1990).

2.1 Probability spaces

In order to develop statistical methods, we have to be able to model random ("chance") experiments. The basic concept from probability theory that we need is that of a probability space.

Definition 1. A *probability space* consists of

(i). a *sample space* Ω of all distinct, possible outcomes of an experiment (*sample points*);

(ii). a *collection of events*[1] $\mathcal{F}$, where *events* are subsets of Ω; and

(iii). a *probability measure* $P : \mathcal{F} \to [0, 1]$ giving the "probability" $P(E)$ of each event E in $\mathcal{F}$.

Example 8. As an example, consider the experiment of tossing a fair coin. In this case, the sample points are that heads (H) and tails (T) prevail, and the sample space is $\Omega = \{H, T\}$. Possible events are that neither H nor T occurs ($\emptyset$), either H or T occurs ($\{H, T\}$), H occurs ($\{H\}$), and that T occurs ($\{T\}$), so we can take $\mathcal{F} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$. As the coin is fair, $P(\{H\}) = P(\{T\}) = 1/2$. Also, intuitively $P(\emptyset) = 0$ and $P(\{H, T\}) = 1$.

In this example, the specification of the probability measure P corresponds intuitively to the notion of a fair coin. More in general, P should satisfy certain properties for it to correspond to our intuitive notion of probability. In particular, we demand that the so called "axioms of probability" hold.

Definition 2. The *axioms of probability* are

A1. For all $E \in \mathcal{F}$: $P(E) \geq 0$;

A2. $P(\Omega) = 1$;

A3. For all sequences $E_1, E_2, \ldots$ of disjoint events in $\mathcal{F}$, $P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

Recall that two sets A and B are *disjoint* if their intersection is empty: $A \cap B = \emptyset$. Also, $\bigcup_{i=1}^{\infty} E_i = E_1 \cup E_2 \cup \ldots$ is the union of all sets in the sequence $E_1, E_2, \ldots$, or the event that an outcome in any of the sets $E_1, E_2, \ldots$ occurs.[2] $\sum_{i=1}^{\infty} P(E_i) = P(E_1) + P(E_2) + \ldots$ is the sum over the sequence of probabilities $P(E_1), P(E_2), \ldots$.

It is easily checked that the probability measure in Example 8 satisfies Axioms A1–A3 (check!). More in general, the axioms A1–A3 have intuitive appeal. Probabilities should be nonnegative (A1), and the probability that any outcome in the set of all possible outcomes Ω occurs is 1 (A2). Also, the probability that any of a collection of disjoint events occurs is the sum of the probabilities of each of these events (A3).

One may wonder whether these three axioms are also sufficient to ensure that some other desirable properties of probabilities hold. For example, probabilities should not be larger that 1 (for all $E \in \mathcal{F}$: $P(E) \leq 1$), and the probability that the chance experiment has no outcome at all should be 0 ($P(\emptyset) = 0$). It is easily checked that A1–A3 indeed imply these properties. The proof of this result is left as an exercise.

2.2 Conditional probability and independence

We frequently have to determine the probability that an event A occurs given that another event B does. This is called the "conditional probability of A given B".

Definition 3. If $P(B) > 0$, the *conditional probability of A given B* is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

We should check whether this definition corresponds to our intuitive notion of a conditional probability. It is easily checked that, for given B and as a function of A, $P(A|B)$ is indeed a probability measure, *i.e.* satisfies Axioms A1–A3. This is left as an exercise. The definition of $P(A|B)$ also has intuitive appeal, as the following example illustrates.

Example 9. Suppose we throw a fair die. We can take $\Omega = \{1, 2, 3, 4, 5, 6\}$, and $\mathcal{F}$ the collection of all subsets of Ω (including $\emptyset$ and Ω). As the die is fair, we take P such that $P(\{1\}) = \cdots = P(\{6\}) = 1/6$, $P(\{1, 2\}) = P(\{1, 3\}) = \cdots = 1/3$, etcetera. Now consider the event $B = \{1,2,3\}$. Then, $P(\{1\}|B) = P(\{2\}|B) = P(\{3\}|B) = 1/3$: the probability that a 1 (or a 2, or a 3) is thrown conditional on either one of $\{1,2,3\}$ being thrown is 1/3. Also, $P(\{4\}|B) = P(\{5\}|B) = P(\{6\}|B) = 0$: the probability that a 4 (or a 5, or a 6) is thrown conditional on either one of $\{1,2,3\}$ being thrown is 0. Obviously, we can also take events A consisting of more than one sample point. For example, $P(\{1,2,4,5,6\}|B) = P(\{1,2\})/P(\{1,2,3\}) = 2/3$: the probability that 3 is *not* thrown conditional on either one of $\{1,2,3\}$ being thrown is 2/3.

Definition 4. Two events $A, B \in \mathcal{F}$ are said to be (stochastically) *independent* if $P(A \cap B) = P(A)P(B)$, and *dependent* otherwise.

If A and B are independent then $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Intuitively, knowledge of B does not help to predict the occurrence of A, and vice versa, if A and B are independent.

Example 10. In Example 9, let $A = \{1,2,3\}$ and $B = \{4,5,6\}$. Obviously, $0 = P(A \cap B) \neq P(A)P(B) = 1/4$, and A and B are dependent. This makes sense, as A and B are disjoint events. So, given that a number in A is thrown, a number in B is thrown with zero probability, and vice versa. In conditional probability notation, we have that $P(B|A) = 0 \neq P(B)$ and $P(A|B) = 0 \neq P(A)$.

Example 11. Suppose we toss a fair coin twice. The sample space is $\Omega = \{H,T\} \times \{H,T\} = \{(H,H),(H,T),(T,H),(T,T)\}$, in obvious notation. Again, each subset of Ω is an event. As the coin is fair, the associated probabilities are $P(\{(H,H)\}) = \cdots = P(\{(T,T)\}) = 1/4$, $P(\{(H,H),(H,T)\}) = P(\{(H,H),(T,H)\}) = \cdots = 1/2$, etcetera. The events "the first toss is heads", $\{(H,H),(H,T)\}$, and "the second toss is heads", $\{(H,H),(T,H)\}$, are independent. This is easily checked, as $P(\{(H,H),(H,T)\} \cap \{(H,H),(T,H)\}) = P(\{(H,H)\}) = 1/4 = P(\{(H,H),(H,T)\})P(\{(H,H),(T,H)\})$. In this coin tossing experiment, the result from the first toss does not help in predicting the outcome of the second toss. Obviously, we have implicitly *assumed* independence in constructing our probability measure, in particular by choosing $P(\{(H,H)\}) = \cdots = P(\{(T,T)\}) = 1/4$, etcetera.

2.3 Random variables

2.3.1 Random variables and cumulative distribution functions

Usually, we are not so much interested in the outcome (*i.e.*, a sample point) of an experiment itself, but only in a function of that outcome. Such a function $X : \Omega \to \mathbb{R}$ from the sample space to the real numbers is called a *random variable.*[3] We usually use capitals like X and Y to denote random variables, and small letters like x and y to denote a possible value, or *realization*, of X and Y.

Example 12. Suppose we play some dice game for which it is only relevant whether the result of the throw is larger than 4, or smaller than 3. If the die is fair, we can still use the probability space from Example 9, with sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$, to model this game. We can define a random variable $X : \Omega \to \mathbb{R}$ such that $X(\omega) = 1$ if $\omega \in \{4, 5, 6\}$ and $X(\omega) = 0$ if $\omega \in \{1, 2, 3\}$. X is an indicator function that equals 0 if 3 or less is thrown, and 1 if 4 or more is thrown. Using our probability model, we can make probability statements about X. For example, the outcome $X = 1$ corresponds to the event $\{\omega : X(\omega) = 1\} = \{4, 5, 6\}$ in the underlying probability space. So, we can talk about the "probability that $X = 1$", and we will sometimes simply write $P(X = 1)$ instead of $P(\{\omega : X(\omega) = 1\})$. Note that $P(X = 1) = P(\{4, 5, 6\}) = 1/2$.

Example 13. Consider a game in which 2 dice are thrown, and only the sum of the dice matters. We can denote the sample space by $\Omega = \{1, 2, \ldots, 6\} \times \{1, 2, \ldots, 6\} = \{(1, 1), (1, 2), \ldots, (1, 6), (2, 1), \ldots, (6, 1), \ldots, (6, 6)\}$. The sum of both dice thrown is given by $X(\omega) = \omega_1 + \omega_2$, where $\omega = (\omega_1, \omega_2) \in \Omega$. This is a random variable, and its distribution is easily constructed under the assumption that the dice are fair. For example, $P(X = 2) = P(\{(1, 1)\} = 1/36 = P(\{6, 6\}) = P(X = 12)$. Also, $P(X = 3) = P(\{(1, 2), (2, 1)\}) = 2/36$, etcetera.

It is clear from the examples that we can make probability statements about random variables without bothering too much about the underlying chance experiment and probability space. Once we have attached probabilities to various (sets of) realizations of X, this is all we have to know in order to work with X. Therefore, in most practical work, and definitely in this course, we will directly work with random variables and their distributions.

Definition 5. The *cumulative distribution function* (c.d.f.) $F_X : \mathbb{R} \to [0,1]$ of a random variable X is defined as

$$F_X(x) = P(X \leq x) = P(\{\omega : X(\omega) \leq x\}).$$

Note that F_X is non-decreasing, and right-continuous: $\lim_{u \downarrow x} F(u) = F(x)$. Also, $F_X(-\infty) = \lim_{x \to -\infty} F_X(x) = P(\emptyset) = 0$ and $F_X(\infty) = \lim_{x \to \infty} F_X(x) = P(\Omega) = 1$. The c.d.f. F_X fully characterizes the stochastic properties of the random variable X. So, instead of specifying a probability space and a random variable X, and deriving the implied distribution of X, we could directly specify the c.d.f. F_X of X. This is what I meant before by "directly working with random variables", and this is what we will usually do in practice.

Depending on whether the random variable X is discrete or continuous, we can alternatively characterize its stochastic properties by a probability mass function or a probability density function.[4]

2.3.2 Discrete distributions and probability mass functions

Definition 6. A *discrete random variable* is a random variable that only assumes values in a countable subset of $\mathbb{R}$.

A set is *countable* if its elements can be enumerated one-by-one, say as $x_1, x_2, \ldots$. A set that is not countable is called *uncountable*. A special case of a countable set is a set with only a finite number of elements, say $x_1, x_2, \ldots, x_n$, for some $n \in \mathbb{N}$. Here, $\mathbb{N} = \{1, 2, \ldots\}$. In the sequel, we denote the values a discrete random variable X assumes by $x_1, x_2, \ldots$, irrespective of whether X assumes a finite number of values or not.

Definition 7. The *probability mass function* (p.m.f.) p_X of X is

$$p_X(x) = P(X = x) = P(\{\omega : X(\omega) = x\}),$$

p_X simply gives the probabilities of all realizations $x \in \mathbb{R}$ of X. For a discrete random variable, we have that $0 < p_X(x) \leq 1$ if $x \in \{x_1, x_2, \ldots\}$, and $p_X(x) = 0$ otherwise. The p.m.f. is related to the c.d.f. by

$$F_X(x) = \sum_{i \in \mathbb{N} : x_i \leq x} p_X(x_i)$$

Note that $F_X(\infty) = 1$ requires that X assumes values in $\{x_1, x_2, \dots\}$ with probability 1, or $\sum_{i=1}^{\infty} p_X(x_i) = 1$. p_X fully characterizes the stochastic properties of X if X is discrete, just like the corresponding F_X. This c.d.f. of a discrete random variable is a step function, with steps of size $p_X(x_i)$ at x_i, $i = 1, 2, \dots$.

Example 14. Let X be a discrete random variable that takes only one value, say 0, with probability 1. Then, $p_X(0) = P(X = 0) = 1$, and $p_X(x) = 0$ if $x \neq 0$. We say that the distribution of X is *degenerate*, and X is not really random anymore. The corresponding c.d.f. is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \text{ and} \\ 1 & \text{if } x \geq 0. \end{cases}$$

Note that F_X is right-continuous, and has a single jump of size 1 at 0. Both p_X and F_X fully characterize the stochastic properties of X.

Example 15. In the dice game of Example 12, in which X indicates whether 4 or more is thrown, $p_X(0) = p_X(1) = 1/2$. X has a discrete distribution, and assumes only a finite number (2) of values. The corresponding c.d.f. is

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1/2 & \text{if } 0 \leq x < 1, \text{ and} \\ 1 & \text{if } x \geq 1. \end{cases}$$

Note that F_X is again right-continuous, and has jumps of size 1/2 at 0 and 1. A random variable of this kind is called a Bernouilli random variable.

Example 16. X is a Poisson random variable with parameter $\lambda > 0$ if it assume values in the countable set $\{0, 1, 2, \dots\}$, and $p_X(x) = P(X = x) = \exp(-\lambda)\lambda^x/x!$ if $x \in \{0, 1, 2, \dots\}$, and $p_X(x) = 0$ otherwise. It is easily checked that $\sum_{x=0}^{\infty} p_X(x) = 1$, so that p_X is a p.m.f. and X is a discrete random variable that can assume infinitely many values. Now, F_X jumps at each element of $\{0, 1, 2, \dots\}$, and is constant in between.

2.3.3 Continuous distributions and probability density functions

A continuous random variable X has $p_X(x) = 0$ for all $x \in \mathbb{R}$. It assumes uncountably many values. In particular, it can possibly assume any value in $\mathbb{R}$, which is an uncountable set. Clearly, the distribution of a continuous variable cannot be represented by its p.m.f., as $p_X(x) = 0$ for all $x \in \mathbb{R}$. Instead, we need the concept of a probability density function.

Definition 8. An *(absolutely) continuous random variable* is a random variable X such that[5]

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du, \tag{1}$$

for all x, for some integrable function $f_X : \mathbb{R} \to [0, \infty)$.

Definition 9. The function f_X is called the *probability density function* (p.d.f.) of the continuous random variable X.

So, instead of specifying a p.m.f. for a continous random variable X, which is useless as we have seen earlier, we specify the probability $P(X \leq x)$ of $X \leq x$ as the integral in equation (1). The probability of, for example, $x' < X \leq x$, for some $x' < x$, can then be computed as

$$P(x' < X \leq x) = F_X(x) - F_X(x') = \int_{x'}^{x} f_X(u)du,$$

which corresponds to the surface under the graph of f_X between x' and x.

A continuous random variable X is fully characterized by its p.d.f. f_X. Note that $F_X(\infty) = 1$ requires that $\int_{-\infty}^{\infty} f_X(x)dx = 1$. Also, note that equation (1) indeed implies that $p_X(x) = F_X(x) - F_X(x-) = 0$ for all x. Here, $F_X(x-) = \lim_{u \uparrow x} F_X(u) = P(X < x)$.

Example 17. X has a *uniform* distribution on $(0, 1)$ if $f_X(x) = 1$ for $x \in (0, 1)$ and $f_X(x) = 0$ otherwise. Then,

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ x & \text{if } 0 \leq x < 1, \text{ and} \\ 1 & \text{if } x \geq 1. \end{cases}$$

Example 18. X has a *normal* distribution with parameters μ and $\sigma > 0$ if

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),$$

for $-\infty < x < \infty$. For $\mu = 0$ and $\sigma = 1$, we get the *standard normal probability density function*, which is frequently denoted by $\phi(x)$. The corresponding c.d.f. is denoted by $\Phi(x) = \int_{-\infty}^{x} \phi(u)du$. The normal p.d.f. is related to $\phi(x)$ by $f_X(x) = \sigma^{-1}\phi\left((x-\mu)/\sigma\right)$, and the normal c.d.f. F_X to Φ through $F_X(x) = \Phi\left((x-\mu)/\sigma\right)$.

2.3.4 Joint, marginal and conditional distributions

Frequently, we are interested in the joint behavior of two (or more) random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$. For example, in econometrics X may be schooling and Y may be earnings.

Example 19. Recall Example 11, in which a fair coin is flipped twice, with sample space is $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$. We can define two random variables by $X(x) = 1$ if $x \in \{(H,H),(H,T)\}$ and $X(x) = 0$ otherwise, and $Y(y) = 1$ if $y \in \{(H,H),(T,H)\}$ and $Y(y) = 0$. X and Y indicate, respectively, whether the first toss is heads and whether the second toss is heads.

If we just specify the distributions F_X and F_Y of X and Y separately (their marginal distributions; see below), we cannot say much about their joint behavior. We need to specify their joint distribution. The joint distribution of X and Y can be characterized by their joint cumulative distribution function.

Definition 10. The *joint cumulative distribution function* $F_{X,Y} : \mathbb{R}^2 \to [0,1]$ of a pair of random variables (X, Y) is defined as

$$F_{X,Y}(x,y) = P(X \leq x \text{ and } Y \leq y) = P(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq y\}).$$

The c.d.f. $F_{X,Y}(x,y)$ is non-decreasing in x and y. Also, $F_{X,Y}(-\infty,-\infty) = P(X \leq -\infty \text{ and } Y \leq -\infty) = P(\emptyset) = 0$ and $F_{X,Y}(\infty,\infty) = P(X \leq \infty \text{ and } Y \leq \infty) = P(\Omega) = 1$. Here, we denote $F_{X,Y}(-\infty,-\infty) = \lim_{x\to-\infty}\lim_{y\to-\infty} F_{X,Y}(x,y)$ and $F_{X,Y}(\infty,\infty) = \lim_{x\to\infty}\lim_{y\to\infty} F_{X,Y}(x,y)$.

If X and Y are discrete, *i.e.* assume (at most) countably many values $x_1, x_2, \ldots$ and $y_1, y_2, \ldots$, respectively, we can alternative characterize their joint distribution by the joint probability mass function.

Definition 11. The *joint probability mass function* $p_{X,Y}$ of two discrete random variables X and Y is

$$p_{X,Y}(x,y) = P(X = x \text{ and } Y = y) = P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\}).$$

Example 20. For the random variables in Example 19, we have that $p_{X,Y}(1,1) = P(\{H,H\}) = 1/4$, $p_{X,Y}(1,0) = P(\{H,T\}) = 1/4$, $p_{X,Y}(0,1) = P(\{T,H\}) = 1/4$, and $p_{X,Y}(0,0) = P(\{T,T\}) = 1/4$.

The joint p.m.f. is related to the joint c.d.f. by

$$F_{X,Y}(x,y) = \sum_{i\in\mathbb{N}:x_i\leq x}\ \sum_{j\in\mathbb{N}:y_j\leq y} p_{X,Y}(x_i,y_j).$$

So, we compute the joint c.d.f. from the joint p.m.f. by simply summing all probability masses on points (x_i,y_j) such that $x_i \leq x$ and $y_j \leq y$. Again, the sum of the probability masses on all points (x_i,y_j) should be 1 for $p_{X,Y}$ to be a p.m.f.. So, $P(X\leq\infty$ and $Y\leq\infty) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} p_{X,Y}(x_i,y_j) = 1$.

It should perhaps be noted that $p_{X,Y}(x_i,y_j)$ may be equal to 0 for some $i,j\in\mathbb{N}$, even if we pick $x_1,x_2,\ldots$ and $y_1,y_2,\ldots$ such that $p_X(x_i)>0$ and $p_Y(y_j)>0$ for all $i,j\in\mathbb{N}$. Even if X and Y assume all values x_i and y_j with positive probability, some particular combinations (x_i,y_j) may have zero probability.

As in the univariate case, we say that X and Y are jointly (absolutely) continuous if we can characterize their joint distribution as an integral over a joint probability density function.

Definition 12. The *joint probability density function* of two jointly (absolutely) continuous random variables X and Y is an integrable function $f_{X,Y}:\mathbb{R}^2\to[0,\infty)$ such that

$$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)dvdu.$$

For both discrete and continuous joint distribution functions we can define marginal cumulative distribution functions.

Definition 13. The *marginal cumulative distribution function* of X is given by $F_X(x) = P(X\leq x) = P(X\leq x$ and $Y\leq\infty) = F_{X,Y}(x,\infty)$. The marginal c.d.f. of Y is $F_Y(x) = P(Y\leq y) = P(X\leq\infty$ and $Y\leq y) = F_{X,Y}(\infty,y)$.

To these marginal c.d.f.'s correspond marginal p.m.f.'s in the discrete case and a marginal p.d.f.'s in the continuous case.

Definition 14. For discrete X and Y, the *marginal probability mass function* of X is given by $p_X(x) = P(X=x) = \sum_{i=1}^{\infty} P(X=x$ and $Y=y_i) = \sum_{i=1}^{\infty} p_{X,Y}(x,y_i)$. The marginal p.m.f. of Y is given by $p_Y(y) = P(Y=y) = \sum_{i=1}^{\infty} P(X=x_i$ and $Y=y) = \sum_{i=1}^{\infty} p_{X,Y}(x_i,y)$.

Example 21. Continuing Example 20, we have that $p_X(0) = p_X(1) = 1/2$ and $p_Y(0) = p_Y(1) = 1/2$.

Definition 15. For jointly continuous X and Y, the *marginal probability density function* of X is given by $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$. The marginal p.d.f. of Y is $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$.

Note that marginal probability mass and density functions are just univariate probability mass and density functions. So, they are related to marginal c.d.f.'s just like univariate probability mass and density functions are related to univariate c.d.f.'s:

$$F_X(x) = \begin{cases} \sum_{i:x_i \leq x} p_X(x_i) = \sum_{i \in \mathbb{N}:x_i \leq x} \sum_{j=1}^{\infty} p_{X,Y}(x_i, y_j) & \text{in the discrete case, and} \\ \int_{-\infty}^{x} f_X(u)du = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(u,y)dydu & \text{in the continuous case.} \end{cases}$$

The following definition of independent random variables closely follows our earlier Definition 4 of independent events in Subsection 2.2.[6]

Definition 16. Two random variables X and Y are *independent* if $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$. In the discrete case we can equivalently require that $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for all $x, y \in \mathbb{R}$, and in the continuous case that $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all $x, y \in \mathbb{R}$.

Note that X and Y are always independent if Y is degenerate. Suppose that $P(Y = c) = 1$ for some real constant c. Then $F_{X,Y}(x,y) = 0$ and $F_Y(y) = 0$ if $y < c$ and $F_{X,Y}(x,y) = F_X(x)$ and $F_Y(y) = 1$ if $y \geq c$. So, $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for all $x, y \in \mathbb{R}$.

Example 22. Recall again Example 19, in which a fair coin is flipped twice, and random variables X and Y indicate whether heads was thrown in the first and the second toss, respectively. It is easily checked that X and Y are independent. Indeed, we have already checked before that the events $\{\omega : X(\omega) = 1\}$ and $\{\omega : Y(\omega) = 1\}$ are independent, which is equivalent to saying that $p_{X,Y}(1,1) = p_X(1)p_Y(1)$. Similarly, $p_{X,Y}(x,y) = p_X(x)p_Y(y)$ for other values of (x,y).

Using our earlier definition of conditional probabilities in Subsection 2.2, we can also derive conditional distributions. We frequently want to talk about the conditional distribution of, for example, X for a single given value of Y, say y. If X and Y are discrete, this is straightforward. In this case, we can directly apply Definition 3 to $P(X = x|Y = y)$, for a value y such that $P(Y = y) > 0$. This gives $P(X = x|Y = y) = P(X = x \text{ and } Y = y)/P(Y = y)$. We call this a conditional probability mass function.

Definition 17. For discrete X and Y, the *conditional probability mass function* of X given $Y = y$ is given by $p_{X|Y}(x|y) = P(X = x|Y = y) = p_{X,Y}(x, y)/p_Y(y)$, for y such that $p_Y(y) > 0$. The conditional p.m.f. of Y given $X = x$ is given by $p_{Y|X}(y|x) = P(Y = y|X = x) = p_{X,Y}(x, y)/p_X(x)$, for x such that $p_X(x) > 0$.

If X and Y are continuous, we face the problem that $p_Y(y) = 0$ even if Y can assume the value y and we may want to condition on it. We have not defined conditional probabilities for conditioning events that have probability 0, *i.e.* we cannot directly apply Definition 3. Instead of formally discussing how to derive an appropriate conditional distribution in this case, we appeal to intuition, and give the following definition of a conditional probability density function.[7]

Definition 18. For jointly continuous X and Y, the *conditional probability density function* of X given $Y = y$ is given by $f_{X|Y}(x|y) = f_{X,Y}(x, y)/f_Y(y)$, for y such that $f_Y(y) > 0$. The conditional p.d.f. of Y given $X = x$ is given by $f_{Y|X}(y|x) = f_{X,Y}(x, y)/f_X(x)$, for x such that $f_X(x) > 0$.

Conditional probability mass and density functions are related to conditional c.d.f.'s as we expect. For example, the conditional distribution of X given $Y = y$ is given by

$$F_{X|Y}(x|y) = \begin{cases} \sum_{i\in\mathbb{N}:x_i\leq x} p_{X|Y}(x_i|y) & \text{in the discrete case, and} \\ \int_{-\infty}^{x} f_{X|Y}(u|y)du & \text{in the continuous case.} \end{cases}$$

Obviously, if X and Y are independent, then $p_{X|Y}(x|y) = p_X(x)$ in the discrete case and $f_{X|Y}(x|y) = f_X(x)$ in the continuous case.

Example 23. It is easy to check this for our coin flipping example.

2.3.5 Expectation and moments

Random variables can be (partially) characterized in terms of their moments. We start by more generally defining the expectation of a function of a random variable. Consider a function $g : \mathbb{R} \to \mathbb{R}$. Under some technical conditions, $g(X)$ is a random variable.[8] After all, if X assumes different values depending on the outcome of some underlying chance experiment, so does $g(X)$. So, we can define the expected value of $g(X)$. This is the average value of $g(X)$ in the population described by our probability model.

Definition 19. The *expected value* $\mathbb{E}[g(X)]$ of a function $g : \mathbb{R} \to \mathbb{R}$ of a random variable X is

$$\mathbb{E}[g(X)] = \begin{cases} \sum_{i=1}^{\infty} g(x_i) p_X(x_i) & \text{if } X \text{ is discrete, and} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{if } X \text{ is continuous.} \end{cases}$$

This general definition is useful, as we can pick $g(X) = X^k$, which gives the moments of a random variable.

Definition 20. The *k-th moment* of a random variable X is $\mathbb{E}(X^k)$.

The first moment of X is very important and is sometimes called the mean of X. The mean is a measure of the *center* of a random variable.

Definition 21. The *expected value* or *mean* of a random variable X is $\mathbb{E}(X)$.

Another choice of g is the squared deviation of X from its mean, $g(X) = [X - \mathbb{E}(X)]^2$, which gives the variance of X.

Definition 22. The *variance* of a random variable X is the *centralized* second moment of X: $\text{var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$.

Definition 23. The *standard deviation* σ of a random variable X is $\sigma = \sqrt{\text{var}(X)}$.

Note that $[X - \mathbb{E}(X)]^2 = 0$ if $X = \mathbb{E}(X)$ and $[X - \mathbb{E}(X)]^2 > 0$ if $X < \mathbb{E}(X)$ or $X > \mathbb{E}(X)$. So, $\text{var}(X) > 0$, unless X is degenerate at $\mathbb{E}(X)$, *i.e.* $P(X = \mathbb{E}(X)) = 1$. The variance is a measure of the *spread* or dispersion around the center $\mathbb{E}(X)$.

Similarly, an interpretation can be given to higher (centralized) moments of X. For example, the third centralized moment is related to the *skewness* (lack of symmetry) of a distribution. The fourth moment is related to the *kurtosis* (peakedness or flatness) of a distribution. We will need these later on, so see your text book (Gujarati, 1995, Appendix A) for details.

Example 24. Suppose that X is uniform on $(0, 1)$. Then $\mathbb{E}(X) = \int_0^1 x dx = 1/2$ and $\mathbb{E}(X^2) = \int_0^1 x^2 dx = 1/3$, so that $\text{var}(X) = 1/3 - (1/2)^2 = 1/12$.

Example 25. Recall the notation introduced for normal distributions in Example 18. The mean and the variance of a normal random variable are μ and σ^2, respectively. If X is a normal random variable, then $(X-\mu)/\sigma$ is called a *standard normal random variable.* A standardized random variable has expectation 0 and variance 1, and is sometimes denoted by Z. The standard normal distribution Φ is the c.d.f. of a standard normal random variable Z, and ϕ is its p.d.f..

It is important to understand that moments do not necessarily exist.

Example 26. The distribution of income X within countries is frequently modeled as a Pareto distribution with parameters $A > 0$ and $\gamma > 0$, with c.d.f.

$$F_X(x) = \begin{cases} 0 & \text{if } x \leq A, \text{ and} \\ 1 - \left(\frac{x}{A}\right)^{-\gamma}, & \end{cases}$$

and p.d.f. $f_X(x) = 0$ if $x \leq A$ and $f_X(x) = \gamma x^{-\gamma-1}A^\gamma$ if $x > A$. Now, note that

$$\int_{-\infty}^{z} x f_X(x)dx = \int_{A}^{z} x f_X(x)dx = \frac{\gamma}{\gamma-1}\left(A - A^\gamma z^{-\gamma+1}\right)$$

converges if $\gamma > 1$ and diverges if $\gamma < 1$, as $z \to \infty$. So, if $\gamma > 1$ then $\mathbb{E}(X) = \gamma A/(\gamma-1)$, but if $\gamma < 1$ then the expectation does not exist (in this case, is “infinite”).

Definition 19 can be straightforwardly extended to the case in which we have a function $g : \mathbb{R}^2 \to \mathbb{R}$ of two random variables X and Y.

Definition 24. The *expected value* $\mathbb{E}[g(X,Y)]$ of a function $g : \mathbb{R}^2 \to \mathbb{R}$ of two random variables X and Y is

$$\mathbb{E}[g(X,Y)] = \begin{cases} \sum_{i=1}^\infty \sum_{j=1}^\infty g(x_i, y_i) p_{X,Y}(x_i, y_i) & \text{if } X \text{ and } Y \text{ are discrete, and} \\ \int_{-\infty}^\infty \int_{-\infty}^\infty g(x,y) f_{X,Y}(x,y)dxdy & \text{if } X \text{ and } Y \text{ are continuous.} \end{cases}$$

Similarly, we can extend the definition further to expected values of functions of more than two random variables.

Various specific choices of the function g lead to useful results. It is easy to check that $\mathbb{E}(a+bX+cY) = a + b\mathbb{E}(X) + c\mathbb{E}(Y)$ and that $\operatorname{var}(a+bX) = b^2 \operatorname{var}(X)$ if a, b, c are real constants. Also, if X and Y are independent and $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$, then $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\mathbb{E}[h(Y)]$ and $\operatorname{var}(X+Y) = \operatorname{var}(X-Y) = \operatorname{var}(X) + \operatorname{var}(Y)$. As we have seen before, the variance of a degenerate random variable, or a real constant, is 0. These results are easily generalized to sequences of random variables $X_1, X_2, \dots, X_n$ and real constants $c_1, c_2, \dots, c_n$. Particularly useful is that

$$\mathbb{E}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_i c_i \mathbb{E}\left(X_i\right).$$

Also, if the X_i are independent, *i.e.* if $P(X_1 \leq x_1, X_2 \leq x_2 \dots, X_n \leq x_n) = P(X_1 \leq x_1)P(X_2 \leq x_2) \cdots P(X_n \leq x_n)$ for all $x_1, x_2, \dots, x_n$, then

$$\operatorname{var}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_i c_i^2 \operatorname{var}\left(X_i\right).$$

We will see a generalization to dependent X_i later.

There are various ways to characterize joint distributions in terms of moments. If we let $g(X,Y) = [X - \mathbb{E}(X)][Y - \mathbb{E}(Y)]$, and take expectations, we get the covariance of X and Y.

Definition 25. The *covariance* $\operatorname{cov}(X,Y)$ of two random variables X and Y is

$$\operatorname{cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Note that $\operatorname{var}(X) = \operatorname{cov}(X,X)$. The covariance is a measure of linear dependence between two random variables. If X and Y are independent, then $\operatorname{cov}(X,Y) = \mathbb{E}[X - \mathbb{E}(X)]\mathbb{E}[Y - \mathbb{E}(Y)] = 0$.

The covariance depends on the scale of the random variables X and Y. If a, b, c and d are real constants, then $\operatorname{cov}(a + bX, c + dY) = bd\operatorname{cov}(X,Y)$. A normalized measure of linear dependency is the correlation coefficient.

Definition 26. The *correlation coefficient* $\rho(X,Y)$ of two random variables is given by

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

It is easy to check that $-1 \leq \rho(X,Y) \leq 1$, and that indeed $\rho(a + bX, c + dY) = \rho(X,Y)$. We have that $\rho(X,Y) = 0$ if X and Y are (linearly) independent. Otherwise, we say that X and Y are *correlated.*

For general random variables X and Y, we have that

$$\begin{aligned}\operatorname{var}(X+Y) &= \operatorname{var}(X)+\operatorname{var}(Y)+2\operatorname{cov}(X,Y), \text{ and}\\ \operatorname{var}(X-Y) &= \operatorname{var}(X)+\operatorname{var}(Y)-2\operatorname{cov}(X,Y).\end{aligned}$$

Note that this reduces to the earlier equations without the covariance term if X and Y are (linearly) independent, and $\operatorname{cov}(X,Y)=0$. For a sequence $X_1, X_2, \ldots, X_n$ of, possibly dependent, random variables we have

$$\operatorname{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \operatorname{var}(X_i) + 2\sum_{i=1}^n \sum_{j:j>i} \operatorname{cov}(X_i, X_j).$$

2.3.6 Conditional expectation and regression

Suppose we know the realization x of some random variable X, and would like to give some prediction of another random variable $g(Y)$. For example, Y could be earnings, $g(Y)$ log earnings, and X years of schooling. We would be interested in predicting log earnings $g(Y)$ for a given level of schooling x. In particular, we could focus on the conditional expectation of $g(Y)$ given that $X=x$. The easiest way to introduce such conditional expectations is as expectations with respect to a conditional distribution.

Definition 27. The *conditional expectation* $\mathbb{E}[g(Y)|X=x]$ of a function $g:\mathbb{R}\to\mathbb{R}$ of a random variable Y conditional on $X=x$ is

$$\mathbb{E}[g(Y)|X=x] = \begin{cases} \sum_{i=1}^\infty g(y_i) p_{Y|X}(y_i|x) & \text{if } X \text{ and } Y \text{ are discrete, and}\\ \int_{-\infty}^\infty g(y) f_{Y|X}(y|x) dy & \text{if } X \text{ and } Y \text{ are continuous.}\end{cases}$$

Note that $\mathbb{E}[g(Y)|X=x]$ is only well-defined if $p_{Y|X}(y|x)$ is well-defined in the discrete case, which requires $p_X(x)>0$, and if $f_{Y|X}(y|x)$ is well-defined in the continuous case, which demands that $f_X(x)>0$. The conditional expectation $\mathbb{E}[g(X)|Y=y]$ of a random variable $g(X)$ conditional on $Y=y$ can be defined analogously.

We can define conditional means, conditional higher moments and conditional variances as before, by choosing $g(Y)=Y$, $g(Y)=Y^k$ and $g(Y)=(Y-\mathbb{E}(Y))^2$, respectively.

Note that $\mathbb{E}[g(Y)|X=x]$ is a real-valued function of x. If we evaluate this function at the random variable X, we get the *conditional expectation* of $g(Y)$ conditional on X, which we simply denote by $\mathbb{E}[g(Y)|X]$. Note that $\mathbb{E}[g(Y)|X]$ is a random variable, as it is a function of the random variable X, and assumes different values depending on the outcome of the underlying chance experiment. As $\mathbb{E}[g(Y)|X]$ is a random variable, we can take its expectation. A very useful result is the *law of the iterated expectations*, which states that

$$\mathbb{E}\left[\mathbb{E}[g(Y)|X]\right] = \mathbb{E}[g(Y)].$$

Checking this result is left as an exercise.[9] The law of the iterated expectations is very useful in practice, as it allows us to compute expectations by first computing conditional expectations, and then taking expectations of these conditional expectations. We will see that this can simplify things a lot.

We started this subsection by saying that we are often interested in predicting some random variable Y given that we know the value of some other random variable X. This is the domain of regression theory, and conditional expectations play a central role in this theory. The conditional expectation $\mathbb{E}[Y|X]$ is sometimes called the *regression of Y on X*. It is the function of X that minimizes the expected quadratic "prediction error"

$$\mathbb{E}\left[(Y - h(X))^2\right] \tag{2}$$

among all possible functions $h(X)$ of X that may be used as "predictors" of Y. In other words, the choice $h(X) = \mathbb{E}[Y|X]$ is the best choice if you want to minimize the criterion in equation (2). A simple proof, which exploits the law of the iterated expectations, can be found in Ross (1998).[10] We will return to this if we discuss regression models later on.

Finally, note that conditional expectation $\mathbb{E}[Y|X]$, or $\mathbb{E}[Y|X = x]$ for that matter, is another way of summarizing the stochastic relationship between X and Y. We have earlier discussed the covariance between X and Y and the correlation coefficient of X and Y.

2.3.7 The normal and related distributions and the central limit theorem

In Examples 15, 16, 17, 18 and 26, we have seen some special distributions: an example of the Bernouilli distribution, the Poisson distribution, the uniform distribution, the normal distribution and the Pareto distribution. Of course, there are many other special distributions, but we will not discuss all these distributions here. If you ever need to know what a particular distribution looks like, you can usually find discussions of that distribution in a probability or statistics text book, like Ross (1998).

The normal distribution, however, is so important in this course that we take a closer look at it here. In Examples 18 and 25, we have seen that the standard normal p.d.f. is given by

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right),$$

for $-\infty < x < \infty$. This is the p.d.f. of a standard normal random variable, *i.e.* a normal random variable with expectation 0 and variance 1. We have denoted the corresponding standard normal c.d.f. by $\Phi(x) = \int_0^x \phi(u)du$.

If X is a standard normal random variable and μ and $\sigma > 0$ are real constants, then $Y = \mu + \sigma X$ is normally distributed with expectation μ and variance σ^2:

$$F_Y(y) = P(\mu + \sigma X \leq y) = P\left(X \leq \frac{y-\mu}{\sigma}\right) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

The corresponding (normal) p.d.f. of Y is

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \sigma^{-1}\phi\left(\frac{y-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right),$$

which is indeed the normal p.d.f. as we have introduced it in Example 18.

An important characteristic of the normal p.d.f. is that it is *symmetric* around the mean μ. We have that $\phi(x) = \phi(-x)$ for all $x \in \mathbb{R}$, so that $f_X(\mu + x) = f_X(\mu - x)$ for all $x \in \mathbb{R}$ if X is normal with expectation μ. In turn this implies that $P(X > \mu + x) = P(X \leq \mu - x)$ for all $x \in \mathbb{R}$ if X is normal with expectation μ.

One of the main reasons that the normal distribution is so important in statistics is that we frequently encounter sums of random variables $\sum_{i=1}^n X_i$, and that the normal distribution is very convenient if we work with such sums.

Example 27. Recall the coin tossing experiment in Example 1 of the introduction. Suppose we toss a coin n times. Define a sequence of random variables $X_1, X_2, \ldots, X_n$ so that $X_i = 1$ if the i-th toss is heads and $X_i = 0$ if the i-th toss is tails. Let $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$, for some $0 \leq p \leq 1$ and all i. Also, make the natural assumption that the outcomes of the tosses, and therefore $X_1, X_2, \ldots, X_n$, are independent. This fully characterizes the distribution of $(X_1, X_2, \ldots, X_n)$. Note that we could have formally defined the random variables as functions on some underlying probability space, but, as I said before, we are happy to work directly with the distribution of random variables. Example 1 suggested "estimating" p as the fraction of heads. This involves a sum of random variables, as the fraction of heads is Y_n/n, with $Y_n = \sum_{i=1}^n X_i$ the number of heads in our n flips of the coin. We postpone a discussion of such issues as "estimation" of an unknown parameter to Subsection 2.4. We only discuss it here to stress that such sums naturally arise in statistics.

One important result is that sums of independent and normally distributed random variables are again normally distributed. From Subsection 2.3.5, we already know that $\mathbb{E}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mu_i$ if $X_1, X_2, \ldots$ is a sequence of random variables such that $\mathbb{E}(X_i) = \mu_i$, $i = 1, \ldots, n$. Furthermore, if the X_i are independent, and if $\text{var}(X_i) = \sigma_i^2$, $i = 1, \ldots, n$, then $\text{var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \sigma_i^2$. If the X_i are not only independent, but also normal, then it is also true that $\sum_{i=1}^n X_i$ is normal, with expectation $\sum_{i=1}^n \mu_i$ and variance $\sum_{i=1}^n \sigma_i^2$.

The normal distribution also appears naturally if the X_i are not normally distributed. The result that links more general sums of random variables to the normal distribution is the *central limit theorem.* We give a simple version of this theorem (see, for example, Ross, 1998).

Proposition 1. *Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) random variables, each having mean μ and variance $\sigma^2 < \infty$, with $\sigma > 0$. Let $Y_n = \sum_{i=1}^n X_i$. Then, the distribution of $(Y_n - n\mu)/(\sqrt{n}\sigma)$ tends to the standard normal distribution as $n \to \infty$. More precisely,*

$$\lim_{n\to\infty} P\left(\frac{Y_n - n\mu}{\sqrt{n}\sigma} \leq y\right) = \Phi(y).$$

Using the results from Subsection 2.3.5, it is easy to see that $n\mu$ is the expected value of Y_n. Also, because the X_i are independent, $n\sigma^2$ is the variance of Y_n, so that $\sqrt{n}\sigma$ is the standard deviation of Y_n. Thus, $(Y_n - n\mu)/(\sqrt{n}\sigma)$ is the standardized version of Y_n, and has mean 0 and variance 1. If the X_i are normally distributed, Proposition 1 is trivial. After all, we have just seen that in this case $(Y_n - n\mu)/(\sqrt{n}\sigma)$ is a standard normal random variable for all n. Proposition 1, however, does not require normality of the X_i. It tells us that, in general, the distribution of the standardized Y_n looks more and more like a standard normal distribution if n increases. This is result is frequently used in statistics to approximate distributions of random variables in cases in which it is hard to derive the exact distributions.

Example 28. Let Y_n again be the number of heads in an experiment involving n (independent) flips of a coin, as in Example 27. We have that $\mathbb{E}(X_i) = p$ and $\text{var}(X_i) = p(1-p)$, $i = 1, \dots, n$. As the X_i are nondegenerate and i.i.d. and have (finite) means and variances, Proposition 1 tells us that the distribution of the standardized number of heads, $(Y_n - np)/(\sqrt{np(1-p)})$, converges to a standard normal distribution. In statistical practice, a result like this is used to approximate the distribution of, in this case, Y_n in large experiments, *i.e.* for large n. For example, suppose that $n = 400$ and $p = 1/2$. Then, $\mathbb{E}(Y_n) = 200$ and $\text{var}(Y_n) = 100$. So,

$$P(Y_n < 190) = P\left(\frac{Y_n - 200}{10} < -1\right) \approx \Phi(-1) \approx 0.16.$$

The first approximation is based on the central limit theorem. By invoking the central limit theorem, we avoid deriving the exact distribution of Y_n. 0.16 is the approximate value of $\Phi(-1)$ as it can be found in a statistical table of the normal distribution (see Gujarati, 1995, Appendix D, and the discussion below).

It should be clear by now that the normal distribution plays a central role in statistics. We finish this subsection by mentioning some important related distributions. We introduce these distributions by giving their relation to the normal distribution:

(i). If $X_1, X_2, \dots, X_n$ are i.i.d. standard normal random variables, then $\sum_{i=1}^n X_i^2$ has a so called *chi-square (χ^2) distribution with n degrees of freedom.* A random variable with this distribution is often denoted by χ_n^2.

(ii). If X_1 is standard normal, X_2 is χ^2 with n degrees of freedom, and X_1 and X_2 are independent, then the *(Student) t-ratio* $X_1/\sqrt{X_2/n}$ has a *(Student) t-distribution with n degrees of freedom.* A random variable with this distribution is often denoted by T_n (or t_n).

(iii). If X_1 and X_2 are χ^2 distributed with n_1 and n_2 degrees of freedom, respectively, and X_1 and X_2 are independent, then the *(Snedecor) F-ratio* $(X_1/n_1)/(X_2/n_2)$ has a *(Snedecor) F-distribution with degrees of freedom parameters n_1 and n_2.* A random variable with this distribution is often denoted by F_{n_1,n_2}.

The "degrees of freedom" n, n_1 and n_2 end up being parameters of the various distributions that are introduced, just like μ and σ are parameters of the normal distribution.

We have not explicitly given the c.d.f.'s or p.d.f.'s of the χ^2, t and F distributions. Instead, we have focused on the relation of these distributions to the normal distribution. If we discuss statistical and econometric applications later, we will frequently deal with normal random variables, and we will often encounter sums of squared i.i.d. normal random variables (χ^2), t-ratios and F-ratios. Instead of explicitly using the corresponding c.d.f.'s or p.d.f.'s to compute probabilities that these random variables take certain values, we will usually search for these probabilities in tables. Appendix D of Gujarati (1995) provides tables for the normal, χ^2, t and F distributions. So, often we do not need to know what exactly the χ^2, t and F distributions are. If necessary, however, they can be found in many probability and statistics text books.

2.4 Classical statistics

In the introduction, we noted that statistics studies analytical methods for uncovering regular relationships from experiments contaminated by "chance". The probability theory we have discussed so far allows us to formally model such chance experiments. In the remainder of this section, we will discuss the cornerstones of classical statistics: sampling from a population, estimation, and hypothesis testing. We will not discuss an alternative approach to statistics, Bayesian statistics.

2.4.1 Sampling from a population

Example 29. In Example 2, we discussed how exit polls are used to predict the presidential election outcome. For expositional convenience, suppose that Bush and Gore are the only two contenders. Also, suppose that we are interested in the popular vote, *i.e.* the shares of Bush and Gore votes in the population of votes.

We can model the *population* of votes as a Bernouilli random variable, *i.e.* a random variable X such $P(X = 1) = p$ and $P(X = 0) = 1 - p$, for some $0 \leq p \leq 1$. Here, $X = 1$ corresponds to a vote for Bush, and $X = 0$ to a Gore vote. Note that we could think of X as being defined on an underlying sample space $\Omega = \{\text{Bush}, \text{Gore}\}$, with $P(\text{Bush}) = p = 1 - P(\text{Gore})$. p is simply the share of Bush votes in the population, and (the distribution F_X of) X fully describes this population. In classical statistics, we want to learn about the population distribution F_X of votes. In this case, we would actually like to know $p = \mathbb{E}(X)$, which is a numerical property of the population distribution. This is called a *parameter*.

To learn about the parameter p, we randomly sample n votes from the population of votes. This means that we ask n randomly selected voters whom they have voted for. Denoting the i-th vote sampled by X_i, we can model the resulting *random sample* of n votes as a vector of n independent random variables $(X_1, \dots, X_n)$, each distributed as the population of votes X. After all, if sampling is truly random, each vote is an independent draw from the distribution F_X of votes.

Of course, if we ask voter i what he or she has voted for, this voter will actually tell us his or her realized vote, which is either Bush ($x_i = 1$) or Gore ($x_i = 0$). So, if we take a single random sample of n votes from the population of votes, we end up with a vector of realizations $(x_1, \dots, x_n)$ of $(X_1, \dots, X_n)$. As in the coin tossing Example 1, the share of realized Bush votes in the sample of votes seems a good estimate of p. We can formally denote this estimate by $n^{-1}\sum_{i=1}^{n} x_i$.

To judge whether this is a good estimate, we use the concept of *repeated sampling*. If we would take another sample of n votes, we would end up with another sequence of realized votes $x'_1, \dots, x'_n$, and another estimate of p, $n^{-1}\sum_{i=1}^{n} x'_i$. It is matter of chance that we end up with the first estimate of p, and not the second, if we only take the first sample. We can actually think of sampling n votes from the population of votes many times, which would give us an array of estimates of p. The properties of this array of estimates are the properties of the random variable $n^{-1}\sum_{i=1}^{n} X_i$. After all, the x_i are realizations of the X_i, so that the estimates $n^{-1}\sum_{i=1}^{n} x_i$ are realizations of the random variable $n^{-1}\sum_{i=1}^{n} X_i$. $n^{-1}\sum_{i=1}^{n} X_i$ is called an *estimator* of p, which is typically denoted by $\hat{p}$. It is also called a *statistic*, and it is a real function of the sample $(X_1, \dots, X_n)$.

Note that $\mathbb{E}(\hat{p}) = \mathbb{E}\left(n^{-1}\sum_{i=1}^{n} X_i\right) = \mathbb{E}(X) = p$. In expectation, our estimator $\hat{p}$ equals the "true" parameter p. So, if we repeatedly sample (many times) from the population of votes and compute an estimate of p, on average our estimates will be on target, *i.e.* equal to p. We say that our estimator is *unbiased*.

This is of course a desirable property, but we may be worried that our estimates are imprecise, in the sense that they vary a lot between different realized samples. We can actually evaluate the variability of our estimates between repeated samples by computing the variance of our estimator:

$$\operatorname{var}(\hat{p}) = \operatorname{var}\left(n^{-1}\sum_{i=1}^{n} X_i\right) = \frac{p(1-p)}{n}.$$

Clearly, if the number of votes n we sample is sufficiently large, the variance of our estimator will be sufficiently small to be confident in our estimate from a single realized sample.

Example 30. Suppose we are interested in the distribution of income over individuals in the US. This example is very much related to the previous example and we only discuss it briefly. We can define a random variable X which describes the distribution of income over the US population. We could, for instance, assume a Pareto distribution for the population distribution of income F_X in the US, with parameters A and γ (see Example 26).

We are interested in learning more about this distribution, or actually about the parameters A and γ. The Census Bureau randomly samples n individuals and asks them to report their income. We assume, for now, that they truthfully report their actual income. If the Census Bureau draws a truly random sample from the population, the sample is a vector of independent random variables $(X_1, \dots, X_n)$ that all have distribution F_X. If the Census Bureau provides us with the results from a single interviewing session, it will provide us with n realized income levels $(x_1, \dots, x_n)$, which is a realization of $(X_1, \dots, X_n)$. Perhaps, we can construct statistics $\hat{A}(X_1, \dots, X_n)$ and $\hat{\gamma}(X_1, \dots, X_n)$ that are good estimators of A and γ. If the Census bureau provides us with this single array of realized income levels $(x_1, \dots, x_n)$, our estimates of A and γ will then be $\hat{A}(x_1, \dots, x_n)$ and $\hat{\gamma}(x_1, \dots, x_n)$.

We will now discuss these ideas more formally.

Definition 28. The *population* is a random variable X, with c.d.f. F_X.

In general, X can be a vector of random variables. For expositional convenience, we restrict attention to a univariate random variable here.

Definition 29. A *random sample* from the population with c.d.f. F_X is a vector of independent random variables $(X_1, \dots, X_n)$ such that each X_i has is distributed with c.d.f. F_X.

A *parameter* is a real constant that describes some characteristic of the population distribution F_X. Examples are μ and σ in the case that F_X is a normal distribution. More in general, moments $\mathbb{E}(X^k)$ are parameters (if they exist). Unless parameters are known, *i.e.* specified with the model, we have to learn about the parameters from sample statistics.

Definition 30. If $g : \mathbb{R}^n \to \mathbb{R}$, and $(X_1, \dots, X_n)$ is a sample, then $g(X_1, \dots, X_n)$ is called a *sample statistic.*

Note that a statistic is a random variable. Important examples of statistics are sample moments.

Definition 31. If $(X_1, \dots, X_n)$ is a random sample, then the *k-th sample moment* is given by $n^{-1} \sum_{i=1}^n X_i^k$. In particular, the *sample mean* $\bar{X}_n$ is given by

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}.$$

2.4.2 Estimation

Suppose we want to estimate the value of an unknown parameter θ, using a sample $(X_1, \dots, X_n)$. To this end, we choose a particular sample statistic, which is a function of the sample $(X_1, \dots, X_n)$, and estimate θ to equal this sample statistic. A statistic that we use in this manner is called a *(point) estimator* of θ, and is typically denoted by $\hat{\theta}$ (*i.e.*, we use the same symbol as for the parameter itself, but add a hat). As an estimator is a sample statistic, it is a random variable. It assumes different values for different actually realized samples or data sets $(x_1, \dots, x_n)$. A realization of an estimator $\hat{\theta}$ for a particular data set is called an *estimate* of θ. If there is no risk of confusion, we will sometimes denote an estimate of θ by $\hat{\theta}$ as well. An estimate is *not* a random variable, but a particular real number that you report as your actual guess of the value of the parameter θ.

Definition 32. An estimator $\hat{\theta}$ of a parameter θ is *unbiased* if $\mathbb{E}(\hat{\theta}) = \theta$.

Example 31. Suppose we want to estimate the parameter $\mu = \mathbb{E}[X]$, the mean of the population X, from a random sample $(X_1, \dots, X_n)$. In analogy of Examples 29 (and 1), in which we focused on $p = \mathbb{E}(X)$, it seems reasonable to estimate μ by the *sample mean* $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$. This estimator satisfies $\mathbb{E}[\bar{X}_n] = \mu$ and $\text{var}[\bar{X}_n] = n^{-1}\sigma^2$, where we assume that the population variance $\sigma^2 = \text{var}(X) < \infty$. So, "on average", our estimator equals the population parameter that we want to estimate. Furthermore, the variance of our estimator decreases as our sample size increases.

Example 32. In the previous example, suppose we want to estimate the σ^2. In analogy to the previous example, a good estimator seems to be $n^{-1}\sum_{i=1}^n (X_i - \mu)^2$. Indeed, $\mathbb{E}\left[n^{-1}\sum_{i=1}^n (X_i - \mu)^2\right] = \sigma^2$. However, as we typically do not know μ, this estimator cannot be computed, *i.e.* it is not *feasible.* It seems reasonable to replace μ by $\bar{X}_n$, and try $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$. This estimator is feasible, as it is a known function of the sample, *i.e.* does not depend on unknown parameters. We have that

$$\begin{aligned}\mathbb{E}(\hat{\sigma}^2) &= n^{-1}\mathbb{E}\left[\sum_{i=1}^n \left(X_i - \mu - (\bar{X}_n - \mu)\right)^2\right]\\ &= n^{-1}\mathbb{E}\left[\sum_{i=1}^n \left((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_n - \mu) + (\bar{X}_n - \mu)^2\right)\right]\\ &= \mathbb{E}\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{n}\right] - 2\mathbb{E}\left[\frac{(\bar{X}_n - \mu)\sum_{i=1}^n (X_i - \mu)}{n}\right] + \mathbb{E}\left[(\bar{X}_n - \mu)^2\right]\\ &= \sigma^2 - \text{var}(\bar{X}_n) = \frac{n-1}{n}\sigma^2.\end{aligned}$$

So, $\hat{\sigma}^2$ is not an unbiased estimator of the population variance. Of course, an unbiased estimator is easily constructed by multiplying the $\hat{\sigma}^2$ by $n/(n-1)$:

$$\mathbb{E}\left[\frac{n}{n-1}\hat{\sigma}^2\right] = \frac{n}{n-1}\mathbb{E}(\hat{\sigma}^2) = \sigma^2.$$

We will denote the second, unbiased estimator of the variance by S_n^2, and call it the sample variance.

Definition 33. The *sample variance* S_n^2 is defined by

$$S_n^2 = \frac{\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2}{n-1}.$$

The square root S_n of the sample variance is called the *sample standard deviation.*

Note that if n is large, $n/(n-1)$ is close to 1 and both estimators in Example 32 are much alike. Actually, it is easy to show that the bias in $\hat{\sigma}^2$ disappears as $n \to \infty$:

$$\lim_{n\to\infty} \mathbb{E}(\hat{\sigma}^2) = \lim_{n\to\infty} \frac{n-1}{n}\mathbb{E}(S_n) = \sigma^2.$$

An estimator with this property is called *asymptotically unbiased.*

Unbiasedness is a desirable property of an estimator $\hat{\theta}$, as it guarantees that it equals the population parameter θ "on average". However, even if our estimator is unbiased, and equals the θ "on average", it may still be imprecise in the sense that it is often very different from θ in particular realized samples. So, we would also like to know what the spread of the estimator around the population parameter is. In other words, if we would repeatedly draw a random sample, how variable would the estimates be?

Two measures of the dispersion of an estimator are its variance and its mean squared error. The variance of $\hat{\theta}$ is simply $\text{var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right]$.

Definition 34. The *mean squared error* $\text{MSE}(\hat{\theta})$ of an estimator $\hat{\theta}$ is the expected squared "prediction" error:

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right].$$

The mean squared error can be decomposed as

$$\text{MSE}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}) + \mathbb{E}(\hat{\theta}) - \theta)^2\right] = \text{var}(\hat{\theta}) + \left[\mathbb{E}(\hat{\theta}) - \theta\right]^2.$$

The second term is the square of the *bias* $\mathbb{E}(\hat{\theta}) - \theta$ of $\hat{\theta}$. If $\hat{\theta}$ is unbiased, $\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta})$. Otherwise, $\text{MSE}(\hat{\theta}) > \text{var}(\hat{\theta})$.

If we provide an estimate of a parameter, we typically like to add a measure of the precision of that estimate. If the estimator is unbiased, a natural choice is the variance of the estimator. The variance of an estimator usually depends on unknown population parameters, and has to be estimated.

Example 33. Consider the problem of estimating a sample mean of Example 31. The variance of the unbiased estimator $\bar{X}_n$ was shown to be $\text{var}[\bar{X}_n] = n^{-1}\sigma^2$. So, this variance depends on the population variance σ^2, which is typically unknown. In Example 32 we have developed an unbiased estimator S_n^2 of σ^2. So, an unbiased estimator of $\text{var}[\bar{X}_n]$ is $n^{-1}S_n^2$. We typically do not just report the estimate of the parameter itself, but also the estimate of the variance of the estimator used. In this way, we can judge how much value to attach to our parameter estimate.

Suppose we focus on unbiased estimators. If we can choose between two unbiased estimators, we would like to choose the most "precise" of the two estimators. As the mean squared error and the variance of the estimator are the same in this case, we could simply choose the estimator that has the lowest variance. This estimator is sometimes called the more *efficient* of the two estimators.

Definition 35. Let $\theta \in \mathbb{R}$ be a parameter, and $\hat{\theta}$ and $\hat{\theta}'$ be two unbiased estimators of θ. Then, $\hat{\theta}$ is called *efficient* relative to $\hat{\theta}'$ if $\text{var}(\hat{\theta}) \leq \text{var}(\hat{\theta}')$.

Example 34. A somewhat trivial example can be constructed from Example 31. If $(X_1, \ldots, X_n)$ is a random sample, then $(X_1, \ldots, X_m)$, with $1 \leq m < n$, is a random sample as well. So instead of estimating μ by $\bar{X}_n$, we could discard the last $n - m$ observations and estimate μ by $\bar{X}_m$. Both estimators are unbiased. However,

$$\text{var}(\bar{X}_n) = \frac{\sigma^2}{n} < \frac{\sigma^2}{m} = \text{var}(\bar{X}_m),$$

so $\bar{X}_n$ is more efficient than $\bar{X}_m$. This makes sense, as we have simply thrown away information in constructing the alternative estimator $\bar{X}_m$.

In Example 31 we have seen that the sample mean is an unbiased estimator of the population mean, and, if $\sigma^2 < \infty$, that the variance of the sample mean decreases with the sample size n, and actually converges to 0 as $n \rightarrow \infty$. We may wonder whether, in some sense, the sample mean "converges" to μ and all uncertainty disappears if the sample size grows large. Formally, the concept we need is consistency.

For the sake of the definition, we leave Example 31 aside for a while and return to a general parameter $\theta \in \mathbb{R}$. Denote the estimator of θ in a sample of size n by $\hat{\theta}_n$.

Definition 36. An estimator $\hat{\theta}_n$ is *(weakly) consistent* if $\hat{\theta}_n$ *converges in probability* to θ as $n \rightarrow \infty$:

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| > \epsilon) = 0$$

for all $\epsilon > 0$, and all possible θ.

Consistency implies that $\hat{\theta}_n$ is very unlikely to be far away from θ in large samples.

Now return to Example 31. A useful and directly applicable result is the *(weak) law of large numbers.*

Proposition 2. *If $X_1, \dots, X_n$ is a sequence of i.i.d. random variables such that $\mathbb{E}[|X_i|] < \infty$ and $\mathbb{E}[X_i] = \mu$, $i = 1, \dots, n$, then $\bar{X}_n$ converges in probability to μ, or*

$$\lim_{n \to \infty} P(|X_n - \mu| > \epsilon) = 0$$

for all $\epsilon > 0$

The law of large numbers immediately implies that the sample mean is a consistent estimator of the population mean if $\mathbb{E}[|X_i|] < \infty$. The assumption that $\sigma^2 < \infty$ implies that $\mathbb{E}[|X_i|] < \infty$.

So far, we have focused on estimating a parameter by a single number. For example, we estimate a population mean by a sample mean. This is a called a *point estimator.* Alternatively, we could provide an interval of possible values of the parameter in which the parameter lies with some prescribed probability. In the case of estimating a mean, we could for example provide some interval such that we can say that the mean lies in that interval with probability 0.95. This is called an *interval estimator*, or confidence interval.

Definition 37. A *confidence interval* for a parameter θ is an interval $[\hat{\theta} - d, \hat{\theta} + d]$ such that

$$P(\hat{\theta} - d \leq \theta \leq \hat{\theta} + d) = 1 - \alpha,$$

for some $0 < \alpha < 1$, and sample statistics $\hat{\theta}$ and d, with $d \geq 0$. $1 - \alpha$ is called the *confidence level.*

It is important to stress that $\hat{\theta}$ and d, and therefore the end points of the confidence interval, are sample statistics. So, they are functions of the sample that we use to estimate θ, and they are random variables. We have defined the confidence intervals to be symmetric around $\hat{\theta}$. We can think of $\hat{\theta}$ to be some point estimator of θ. This is why I use the suggestive notation $\hat{\theta}$: we have indeed earlier used this symbol for a point estimator of θ.

A confidence interval provides an idea of the value of a population parameter, just like a point estimator does. It is also indicative of the uncertainty we are facing in estimating the parameter. If the confidence interval is very wide, we are very uncertain about the parameter. If it is small, we can say with some confidence that our parameter has any of a small number of values.

It is useful to compare this explicitly with a typical point estimation strategy. Recall that when using a point estimator, we give our best shot at estimating the parameter by providing a single (point) estimate, but usually also provide a measure of the precision of this estimate by providing an estimate of the variance of the estimator used.

Example 35. Consider a normal population X with mean μ and variance σ^2. Let $(X_1, \dots, X_n)$ be random sample from this population. Consider again the problem of estimating μ. Assume first that we know σ^2. We know from Example 31 that $\bar{X}_n$ is an unbiased estimator of μ, and therefore is a natural candidate for the center of our confidence interval. Because of the normality assumption, X_n is normally distributed with mean μ and variance σ^2/n. Thus, $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ has a known distribution, the standard normal distribution. Furthermore, it only involves μ, a statistic, and known numbers n and σ. So, we should be able to use $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$ to construct a confidence interval for μ.

To this end, Let $n_{1-\alpha/2}$ denote the $(1-\alpha/2)$-*quantile* of the standard normal distribution, *i.e.* the real number $n_{1-\alpha/2}$ such that $\Phi(n_{1-\alpha/2}) = 1 - \alpha/2$. As the standard normal distribution is symmetric around 0, $-n_{1-\alpha/2}$ is the $\alpha/2$-quantile of the standard normal distribution. Thus, as

$$P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq x\right) = \Phi(x),$$

we have that

$$P\left(-n_{1-\alpha/2} \leq \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq n_{1-\alpha/2}\right) = \Phi(n_{1-\alpha/2}) - \Phi(-n_{1-\alpha/2}) = 1 - \alpha.$$

Rearranging terms within the argument on the left hand side gives

$$P\left(\bar{X}_n - \frac{\sigma n_{1-\alpha/2}}{\sqrt{n}} \leq \mu \leq \bar{X}_n + \frac{\sigma n_{1-\alpha/2}}{\sqrt{n}}\right) = 1 - \alpha,$$

which shows that $[\bar{X}_n - \sigma n_{1-\alpha/2}/\sqrt{n}, \bar{X}_n + \sigma n_{1-\alpha/2}/\sqrt{n}]$ is a confidence interval for μ with confidence level $1 - \alpha$. In terms of the notation of Definition 37, $\hat{\theta} = \bar{X}_n$ and $d = \sigma n_{1-\alpha/2}/\sqrt{n}$. Note that in this case (σ known) d is not a random variable.

The confidence interval shrinks if the sample size n increases. This indicates that we can make more precise statements about μ in larger samples. This is closely related to the result that the variance of $\bar{X}_n$, an unbiased point estimator of μ, decreases with n.

This confidence interval depends on σ. If we abandon the assumption that we know σ, it is not feasible to base the construction of a confidence interval on $(\bar{X}_n - \mu)/(\sigma/\sqrt{n})$. A natural solution seems to be to replace σ by the sample standard deviation S_n, and base our confidence interval on $(\bar{X}_n - \mu)/(S_n/\sqrt{n})$. It can be shown that $(\bar{X}_n - \mu)/(S_n/\sqrt{n})$ has a t-distribution with $n-1$ degrees of freedom.[11] Furthermore, it only involves μ, statistics, and a known number n. So, just as in the known-variance case, we should be able to use $(\bar{X}_n - \mu)/(S_n/\sqrt{n})$ to construct a confidence interval for μ. Indeed, it is easy to see that we will find the same expression for the confidence interval, with σ replaced by S_n and the quantile $n_{1-\alpha/2}$ replaced by the $(1-\alpha/2)$-quantile of a t_{n-1}-distribution. This uses that the t-distribution is symmetric around 0, just like the standard normal distribution, so that again the $\alpha/2$-quantile is minus the $(1-\alpha/2)$-quantile.

We will frequently encounter the t-distribution if we test hypotheses. Quantiles for the t and other distributions can be found in tables in text books (for example, Gujarati, 1995, Appendix D). It is useful to know that the t_n-distribution converges to the normal distribution as $n \to \infty$. As a rule of thumb, we can safely use standard normal quantiles if $n = 120$ or higher.

Example 36. Results for the election poll in Example 29 can be reported as a confidence interval. For example, the estimate of the share of Bush votes could be $49\% +/- 3\%$ with a confidence level of 95%. This means that $P(46\% \leq p \leq 52\%) = 0.95$.

2.4.3 Hypothesis testing

One of the main goals of statistics is to confront *hypotheses* to data. In terms of a statistical model, an hypothesis is a conjecture about a parameter.

Example 37. In Examples 1 and 27, we were interested to establish the fairness of a coin. One hypothesis is that the coin is fair, or $p = 1/2$. Another hypothesis is that the coin is not fair, or $p \neq 1/2$. The hypothesis $p = 1/2$ is an example of a *simple hypothesis.* It is called simple because it completely specifies the population (and sample) distribution, by conjecturing a single value for the parameter p that fully characterizes this distribution. $p \neq 1/2$ is called a *composite* hypothesis, as it only states that p takes any of a range values.

In Example 29 we want to know whether Bush or Gore wins the popular vote. One hypothesis is that Bush wins, or $p > 1/2$. Alternatively, we can hypothesize that Gore wins, or $p < 1/2$. Both hypothesis are composite hypotheses.

Hypothesis testing is concerned with choosing between two competing hypotheses:

(i). a *null hypothesis*, denoted by H_0, and

(ii). an *alternative hypothesis*, denoted by H_1.

The two hypotheses are not treated symmetrically. The null hypothesis is favored in the sense that it is only rejected if there is strong evidence against it. The null hypothesis typically summarizes our prior belief about the true value of a parameter. The alternative hypothesis corresponds to our a priori idea about how the null hypothesis could be wrong.

Example 38. We typically belief that a coin is fair, unless proven otherwise. So, in the coin tossing example, our null hypothesis would be $H_0 : p = 1/2$, and our alternative hypothesis could be $H_1 : p \neq 1/2$. This gives a *two-sided test*, as H_1 is a two-sided alternative hypothesis. Alternatively, we may believe that the coin is fair, but also suspect that it is biased towards heads if biased at all. We would still pick $H_0 : p = 1/2$, but take $H_1 : p > 1/2$ as our alternative hypothesis (recall that p was the probability of heads). This is an example of a *one-sided test*. In the election poll example, we may a priori believe that Bush wins, and pick $H_0 : p > 1/2$. We would then maintain this hypothesis, unless there is strong evidence in favor of $H_1 : p \leq 1/2$.

More in general, suppose that we are concerned with a parameter $\theta \in \mathbb{R}$. In this review, we consider hypotheses like $H_0 : \theta = \theta_0$ and $H_1 : \theta \in \mathcal{T}$, where $\theta_0 \in \mathbb{R}$ is some hypothesized true value of θ and $\mathcal{T} \subset \mathbb{R}$ is a set of alternative values of θ that does not include θ_0 ($\theta_0 \notin \mathcal{T}$). A test procedure for such a set of hypotheses H_0 and H_1 proceeds as follows.

(i). First, we need a *test statistic* T that is somehow informative on θ, *i.e.* to some extent discriminates between H_0 and H_1. We typically take T to be a sample statistic with a known distribution under H_0.

(ii). Then, we choose a *significance level* α, with $0 < \alpha < 1$. The significance level is the probability of a *type-I error*: rejection of H_0 if it is true. It is also called the *size* of the test. Typical values of α are 0.01, 0.05 and 0.10.

(iii). Next, we construct a *critical (or rejection) region* Γ_α. This is a set of possible values of the test statistic that contains the test statistic with probability α under H_0, *i.e.* $P(T \in \Gamma_\alpha) = \alpha$ under H_0.

(iv). Finally, if the test statistic assumes a value in the critical region, $T \in \Gamma_\alpha$, this is considered to be strong evidence against H_0, and we *reject* H_0 in favor of H_1. In this case, we say that the result of our test is *(statistically) significant*. Otherwise, we conclude that we *fail to reject*, or just *not reject*, H_0.

The distribution of T typically depends on the parameter θ. If not, T would not be a good statistic to test hypotheses about θ. This implies that we only know the distribution of T if we pick a particular value of the unknown θ. The term "under H_0" refers to such a choice. It means that we use distribution functions evaluated at the parameter value hypothesized by H_0, *i.e.* $\theta = \theta_0$. Note that in (*ii*) and (*iii*) above, we can compute the probability of a type-I error (the size of the test) for a given test statistic and any critical region if we know the distribution of the statistic under H_0.

Typically, there are many critical regions that are consistent with a given size of the test. Some critical regions are better than others. This is where the alternative hypothesis H_1 comes in. Given a particular size of the test, we would like a critical region that leads to relatively many rejections of H_0 if H_1 is true. In other words, for a given probability of a type-I error, the size of the test, we would like to minimize the probability of a *type-II error*, failure to reject H_0 if H_1 is true. We say that we want to maximize the *power* of the test. We will return to this later. In the examples, we will see that it is usually intuitively clear which critical region to choose.

The fact the we are primarily concerned with limiting the type-I error reflects the conservative attitude towards rejecting the null hypothesis alluded to above. Again, we do not want to reject the null hypothesis, unless there is strong evidence against it.

The terminology "reject H_0" and "fail to reject H_0" used for conclusions drawn from tests reflects the asymmetric treatment of H_0 and H_1. We never say "accept H_1" (or "fail to reject H_1") instead of "reject H_0". A statistical test is centered around the null hypothesis H_0, and is not designed to judge whether H_1 can be accepted.

Also, we preferably do not say "accept H_0" instead of "fail to reject H_0". Tests typically have a considerable probability of a type-II error. "Accepting H_0" seems to suggest we are not willing to reconsider our test result if more data come in later.

Example 39. Consider again a normal population X with mean μ and variance σ^2. Let $(X_1, \dots, X_n)$ be random sample from this population. Consider the one-sided test $H_0 : \mu = 0$ against $H_1 : \mu > 0$. An appropriate test statistic seems to be the sample mean $\bar{X}_n$, or actually

$$Z_0 = \frac{\bar{X}_n}{\sigma/\sqrt{n}},$$

which is a standard normal random variable under H_0. The alternative hypothesis H_1 is more likely to be true if Z_0 is large, so it seems appropriate to construct a critical region of the form $\Gamma_\alpha = (c_\alpha, \infty)$. In this case, the number c_α is called a *critical point.* Given a significance level α, c_α should be such that $Z_0 \in (c_\alpha, \infty)$, *i.e.* H_0 is rejected, with probability α. So, we choose c_α such that (under H_0) $P(Z_0 > c_\alpha) = 1 - \Phi(z) = \alpha$. So, c_α should be the $(1-\alpha)$-quantile $n_{1-\alpha}$ of the standard normal distribution. For example, if $\alpha = 0.05$, we can find in a table of the standard normal distribution that $c_{0.05} = n_{0.95} \approx 1.645$ (Gujarati, 1995, last line of Table D.2). If we find that $Z_0 > 1.645$, we reject $H_0 : \mu = 0$. A test like this, involving a standard normal test statistic, is sometimes called a *Z-test.*

Usually, we do not know σ and a Z-test is not *feasible.* As in the construction of a confidence interval in Example 35, we can substitute the sample standard deviation S_n for σ, which gives the *t-statistic* (see Example 35)

$$T_{n-1} = \frac{\bar{X}_n}{S_n/\sqrt{n}}.$$

We can construct a critical region (c_α, ∞) as for the Z-test. The only difference is that we now pick c_α to be the $(1-\alpha)$-quantile of the t_{n-1}-distribution. A test involving a t-statistic is usually called a *t-test.*

In this example, we report a test result by saying whether H_0 is rejected or not at a given significance level α. In the case of the Z-test, we say the H_0 is rejected if $Z_0 \in (n_{1-\alpha}, \infty)$. Now note that that the probability of rejecting H_0 increases with α. The higher the level of significance α, the less conservative we are with respect to rejecting H_0. So, for a given realization z_0 of Z_0, we could compute the lowest significance level consistent with rejecting H_0. This is the idea behind the so called p-value (probability value) or exact significance of a test.

Definition 38. Suppose we have constructed a critical region Γ_α for a test statistic T. For a given realization t of T, the *p-value* is the lowest significance level α such that H_0 is rejected, *i.e.* the lowest significance level α such that $t \in \Gamma_\alpha$.[12]

So, instead of rejecting H_0 or not at a given significance level, the p-value gives the lowest significance level, corresponding to the most conservative attitude towards rejection of H_0, that would still lead to rejection of H_0. If the given significance level α is higher than the p-value, then we would reject H_0 at a significance level of α. Otherwise, we wouldn't. An example can clarify this.

Example 40. Consider again the Z-test for the case that σ is known in Example 39. The critical region for this test is $\Gamma_\alpha = (n_{1-\alpha}, \infty)$. Suppose we have a realization z_0 of our test statistic Z_0. Then, we reject H_0 as long as α is such that $z_0 > n_{1-\alpha}$. The value of α for which we switch from rejecting to not rejecting H_0 is the number p such that $z_0 = n_{1-p}$. We would reject H_0 for all significance levels $\alpha \in (p, \infty)$ (note that this does not include the boundary case $\alpha = p$; this technical detail is discussed in note 12).

A numerical example is useful. Suppose we have computed that $z_0 = 1.960$. According to Example 39, if we choose a significance level of 0.05, we should reject H_0 if $Z_0 > 1.645 \approx n_{0.95}$. So, with this particular realization, we would reject H_0, as $z_0 = 1.960 > 1.645$. The realized p-value of this test is the value p such that $z_0 = 1.960 = n_{1-p}$. From the last line of Table D.2 in Gujarati (1995), we know that this gives $p = 0.025$ (1.960 is the 0.975-quantile $n_{0.975}$ of the normal distribution). So, we would still have rejected H_0 for values of α below 0.05, but above 0.025.

Instead, suppose that $z_0 = 1.282$. Now, $z_0 = 1.282 < 1.645$, and we would not reject H_0 at a 0.05 significance level. The p-value corresponding to this realization z_0 is the number p such that $z_0 = 1.282 = n_{1-p}$. Again from the last line of Table D.2 in Gujarati (1995), we know that this gives $p = 0.10$ (1.282 is the 0.90-quantile $n_{0.90}$ of the normal distribution). So, we would have rejected H_0 if we would have been slightly less conservative, and had set the significance level α to some level $\alpha > p = 0.10$.

So far, we have been rather informal about the choice of the critical region Γ_α. It seemed appropriate to pick a region (c_α, ∞) for our one-sided test in Example 39, as the sample mean typically takes relatively large values under H_1. We can formalize this by considering the *power* of a test, the probability of rejecting H_0 when H_1 is true. Clearly, we want the power of a test to be as large as possible, for a given significance level. Note that this is equivalent to saying that we want the probability of a type-II error to be as small as possible for a given probability of a type-I error. One difficulty in assessing the power of a test is that we may not know the distribution of the test-statistic under H_1. In particular, if the test concerns a parameter θ, H_1 may specify a range of values for this parameter. Then, as the distribution of the test statistic typically depends on θ, it is not clear what this distribution is under H_1. To deal with this, we use the power function.

Definition 39. Suppose we have two hypotheses H_0 and H_1 concerning a parameter θ. The *power function* $\pi(\theta)$ of this test is the probability that H_0 is rejected as a function of the parameter θ.

Example 41. Consider again the one-sided test $H_0 : \mu = 0$ against $H_1 : \mu > 0$ from Example 39. Suppose that σ^2 is known, and that we use the Z-statistic Z_0 with critical region $(n_{1-\alpha}, \infty)$. We will derive the corresponding power function, say π_r. For a given value of μ,

$$Z_\mu = Z_0 - \frac{\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

is standard normal. Now, as the probability of $Z_0 > n_{1-\alpha}$ (rejecting H_0) equals the probability of $Z_\mu > n_{1-\alpha} - \sqrt{n}\mu/\sigma$, it follows that

$$\pi_r(\mu) = 1 - \Phi\left(n_{1-\alpha} - \frac{\sqrt{n}\mu}{\sigma}\right),$$

First, note that $\pi_r(0) = 1 - \Phi(n_{1-\alpha})$ is the probability of rejecting H_0 under $H_0 : \mu = 0$ (type-I error). This is simply the significance level α of the test.

Next, note that $\pi_r(\mu)$ is smaller at all values of μ if the significance level α is smaller, and the critical point $n_{1-\alpha}$ is higher. This highlights the trade-off between the type-I and type-II errors of the test: a higher size (higher probability of type-I error) corresponds to a higher power (lower probability of type-II error) of the test.

To further judge the power of the test, we evaluate $\pi_r(\mu)$ at values of μ consistent with H_1, *i.e.* $\mu > 0$. $\pi_r(\mu)$ is increasing in μ. For μ just above 0, $\pi_r(\mu)$ is only slightly higher than α. As $\mu \to \infty$, the power converges to 1: the Z-test is very likely to reject H_0 if μ is large. Finally, note that the power of the test increases with the sample size n. If n is very large, $\pi_r(\mu)$ is close to 1, even if μ is fairly small.

So far, we have restricted attention to a critical region of the form $(n_{1-\alpha}, \infty)$. We finish this example by contrasting this choice to the alternative critical region $(-\infty, -n_{1-\alpha})$. Note that $P(Z_0 < -n_{1-\alpha}) = P(Z_0 < n_\alpha) = \alpha$ under H_0, so that this alternative corresponds to the same significance level α. We can derive the corresponding power function π_l as before. The probability of $Z_0 < -n_{1-\alpha}$ (rejecting H_0) equals the probability of $Z_\mu < -n_{1-\alpha} - \sqrt{n}\mu/\sigma$, so that

$$\pi_l(\mu) = \Phi\left(-n_{1-\alpha} - \frac{\sqrt{n}\mu}{\sigma}\right)$$

for a given parameter value μ. Again $\pi_l(0) = \alpha$, as it should be. However, $\pi_l(\mu)$ is decreasing in both μ and n. So, for all values of μ consistent with $H_1 : \mu > 0$, the power is smaller than α, and for very large μ the power is near 0. Also, for large n, the power is close to 0 for most $\mu > 0$. Clearly, this alternative critical region is much worse than our original choice: at the same significance level, the power is much lower.

Note that $\pi_l(\mu) = \pi_r(-\mu)$. As we should expect (because of symmetry of the normal distribution), $(-\infty, -n_{1-\alpha})$ is as good a critical region for the one-sided test $H_0 : \mu = 0$ versus $H_1 : \mu < 0$ as $(n_{1-\alpha}, \infty)$ is for the test we have considered here, $H_0 : \mu = 0$ versus $H_1 : \mu > 0$.

We end this review with an example of a two-sided test.

Example 42. Now suppose we want to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. We can again use the Z-statistic Z_0. However, our intuition and the power discussion suggest that we have to adjust the shape of the critical region to reflect the fact that we want our test to have power for both $\mu < 0$ and $\mu > 0$. It seems reasonable to reject H_0 for both very low and very high values of Z_0, so that the critical region is $(-\infty, -c'_\alpha) \cup (c_\alpha, \infty)$, for some $c'_\alpha < c_\alpha$. A common approach to choosing c'_α and c_α is to divide the size α of the test evenly between both tails, and construct a *symmetric* test. So, we make sure that $P(Z_0 \in (-\infty, -c'_\alpha) \cup (c_\alpha, \infty)) = \alpha$ by picking c'_α and c_α such that (under H_0)

$$P(Z_0 < -c'_\alpha) = \frac{\alpha}{2} = P(Z_0 > c_\alpha).$$

This gives $c_\alpha = -c'_\alpha = n_{1-\alpha/2}$. We reject H_0 if $Z_0 > n_{1-\alpha/2}$ or $Z_0 < -n_{1-\alpha/2}$. We can again evaluate the power of the test as before, but leave that for an end note.[13]

There is a close connection between two-sided tests and confidence intervals. For example, suppose you want to test $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$, where μ is the mean of a random variable X of which the variance σ^2 is known, and μ_0 is some hypothesized value of μ. As σ is known, we can use the Z-statistic $Z_{\mu_0} = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ as our test-statistic. Note that Z_{μ_0} is indeed standard normal under H_0, *i.e.* if $\mu = \mu_0$. With a level of significance α, we would reject H_0 if Z_{μ_0} falls in the critical region $(-\infty, -n_{1-\alpha/2}) \cup (n_{1-\alpha/2}, \infty)$. It is easily checked that this indeed occurs with probability α under H_0. We can alternatively say that we would *not* reject H_0 if

$$-n_{1-\alpha/2} \leq Z_{\mu_0} \leq n_{1-\alpha/2}.$$

Substituting $Z_{\mu_0} = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ and rearranging, we find that this is equivalent to

$$\bar{X} - \frac{\sigma n_{1-\alpha/2}}{\sqrt{n}} \leq \mu_0 \leq \bar{X} + \frac{\sigma n_{1-\alpha/2}}{\sqrt{n}},$$

or μ_0 falling inside the confidence interval for μ at a confidence level $(1-\alpha)$ (see Example 35).

So, there are two ways to perform a two-sided test like this. We can either check whether the test statistic falls in the critical region of the test, or whether the hypothesized value of the parameter under H_0 falls inside the confidence interval for the parameter. In this example, we reject H_0 either if Z_{μ_0} falls in the critical region at a significance level α, or if μ_0 falls outside the $(1-\alpha)$-confidence interval for μ.

3 The classical simple linear regression model

Warning: I use upper case for random variables, and lower case for their realizations. I do not use lower case for variables in deviations from their means.

3.1 Introduction

In Subsection 2.4 we have focused on the statistical analysis of a single variable. The techniques developed there have some interesting econometric applications, for example the analysis of the income distribution (see Example 30). More often, however, we are interested in relating various random variables. A very simple example is found in problem set 2. There, we compare mean earnings between males and females, so we are jointly analyzing earnings and sex. In this and the next sections, we develop more advanced techniques of relating two random variables.

Example 5 discusses research into the returns to schooling. The returns to schooling can, for example, be defined as the gain in earnings in response to an additional year of schooling. We noted that schooling and earnings can be related for other reasons than the direct effect of schooling on earnings, *i.e.* the returns to schooling. For this reason, measuring the returns to schooling is a difficult problem. As a first step, we will discuss how we can characterize the relation between schooling and earnings in data, without interpreting this relation in terms of the returns to schooling.

Formally, denoting schooling by X and log earnings by Y, we can model schooling and earnings in the population by the joint distribution of (X, Y). To investigate how schooling and earnings are related in the US working age civilian population, we can use the 1995 CPS abstract of problem set 2. This is a realization $((x_1, y_1), \ldots, (x_n, y_n))$ of a random sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ of schooling levels and log earnings for n individuals from this population. Here, x_i is the actual years of schooling of individual i, and y_i are individual i's actual log earnings (note that we had to generate log earnings from the wage variable ourselves). So, $(x_1, y_1), \ldots, (x_n, y_n)$ are the actual numbers stored in the STATA data set, with x_1 the years of schooling for the first observation, y_1 the corresponding log wage, etcetera. Our goal is to somehow characterize the relation between log earnings and schooling in this sample. In particular, we would like to know whether earnings are higher for individuals with higher levels of education.

As a start, we can tell STATA to produce a scatter-plot of the sample with log earnings on the vertical axis and schooling on the horizontal axis. If we do so, we find a mildly positive relation, but with a lot of variation in earnings for any given level of schooling. The plot is not very convincing, and we would like to have more formal measures of the relation between earnings and schooling.

In Subsection 2.3.5 we have seen that we can summarize the relation between two random variables X and Y by the correlation coefficient $\rho(X,Y)$. So, it seems reasonable to characterize the linear relation between log earnings and schooling in our sample by the sample equivalent of this correlation coefficient,

$$\hat{\rho}_{X,Y} = \frac{s_{XY}}{s_X s_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

where $\bar{x} = n^{-1}\sum_{i=1}^n x_i$ and $\bar{y} = n^{-1}\sum_{i=1}^n y_i$ are the sample means of schooling and log earnings, respectively. $s_X^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})$ and $s_Y^2 = (n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y})^2$ are the sample variances of X and Y, s_X and s_Y the corresponding sample standard deviations, and $s_{XY} = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{x})$ the sample covariance of X and Y. Like the population correlation and covariance, the sample correlation is symmetric: $\hat{\rho}_{X,Y} = \hat{\rho}_{Y,X}$.

Note that this approach, estimating $\rho(X,Y)$ by replacing population moments in the definition of $\rho(X,Y)$ by sample moments, is similar to our approach to estimating means and variances in Subsection 2.4.2. Also, note that $\hat{\rho}_{X,Y}$ is an *estimate* of the population correlation here, as it is computed from an actual data set (realized sample). The sample correlation $\hat{\rho}_{X,Y}$ between schooling and log earnings is the number that is reported by STATA if you compute correlations (as in problem set 2). If $\hat{\rho}_{X,Y} > 0$, then high levels of schooling and high earnings go, in some sense, hand in hand in our sample. This is actually what we find in the 1995 CPS data abstract. If we would have found that $\hat{\rho}_{X,Y} < 0$, many years of schooling and low earnings would have coincided.

The sample correlation found, $\hat{\rho}_{X,Y} > 0$, confirms that there is some positive linear dependence between both variables. As an alternative to computing the sample correlation, we could draw a straight line $y = a + bx$ through the data, and "predict" each y_i by $a + bx_i$. As y is log earnings, the parameter b of this line can be interpreted as the (average) percentage change in earnings corresponding to one year of schooling (see problem set 2).

It is immediately clear from the scatter plot that it is impossible to find a straight line that cuts through all data points. So, at the very best, we can choose an intercept a and slope b such that our line $a + bx$ is as close as possible to the data. In particular, we could try to somehow minimize the average distance between the actual log earnings levels y_i and the corresponding points $a + bx_i$ on the straight line (the prediction errors or *residuals*). Exactly which line this produces depends on the measure of "closeness" we choose. Obviously, there is some arbitrariness involved in this choice. However, one particular criterion, the *sum of squared residuals*

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2 \tag{3}$$

will later be shown to be particularly natural in the context of a regression model.

So, suppose we choose a and b to minimize (3), and denote the corresponding values of a and b by $\hat{\alpha}$ and $\hat{\beta}$. From calculus, we know that we can find the minimum of (3) by taking derivatives with respect to a and b, and equating these derivatives to 0. So, $\hat{\alpha}$ and $\hat{\beta}$ should satisfy the first order conditions (also known as the *normal equations*)

$$\sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right) = 0. \tag{4}$$

From this, we can derive that

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}.$$

Substituting in the second equation gives

$$\sum_{i=1}^{n} x_i \left[y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})\right] = \sum_{i=1}^{n} (x_i - \bar{x}) \left[y_i - \bar{y} - \hat{\beta}(x_i - \bar{x})\right] = 0.$$

Rearranging gives

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{XY}}{s_X^2},$$

provided that $s_X > 0$, *i.e.* that there is some variation in the schooling levels in the sample. So, the "best" intercept according to our sum of squared residuals criterion is simply the intercept that ensures that the average residual is 0 ($\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$). The slope $\hat{\beta}$ is closely related to the sample correlation coefficient. This should come as no surprise, as we have simply characterized the linear relation between both variables in an alternative way. The only difference is that it is not the covariance standardized by the product of the sample standard deviations of both variables, but is instead divided by the sample variance of schooling.

3.2 The simple linear regression model

As will be become clear soon, fitting a straight line to the data by minimizing the sum of squared residuals ("least squares") is closely related to the regression concept introduced in Subsection 2.3.6. Regressions (conditional expectations) can be used to characterize the relation between random variables just like correlation coefficients. The main advantage of regression is that it can easily be extended to more than 2 variables. For example, earnings may not only depend on schooling but also on work experience, and we may want to analyze the dependence of earnings on schooling and experience simultaneously. An extension of the two-variable regression considered so far, *multiple regression*, can handle this problem, and will be discussed in the Section 4. In this section, however, we first focus on two-variable, or *simple* regression.

So, suppose we want to characterize the relation between Y and X in some population characterized by the joint distribution of (X, Y).

Definition 40. *The population regression of Y on X is given by $\mathbb{E}(Y|X)$. The disturbance or error term of the regression is defined by $U = Y - \mathbb{E}(Y|X)$. X is called the regressor or explanatory variable. Y is called the regressand or explained variable.*

Definition 40 implies that we can write the population regression as

$$Y = \mathbb{E}(Y|X) + U.$$

If we interpret $\mathbb{E}(Y|X)$ as a prediction of Y for given X, then U is the prediction error. The following results are easy to derive.

(i). $\mathbb{E}(U|X) = 0$. This result uses that $\mathbb{E}[\mathbb{E}(Y|X)|X] = \mathbb{E}(Y|X)$, so that

$$\mathbb{E}(U|X) = \mathbb{E}[Y - \mathbb{E}(Y|X)|X] = \mathbb{E}(Y|X) - \mathbb{E}[\mathbb{E}(Y|X)|X] = 0. \tag{5}$$

The law of the iterated expectations immediately implies (ii). $\mathbb{E}(U) = 0$.

Equation (5) also implies that

$$\mathbb{E}(XU) = \mathbb{E}[\mathbb{E}(XU|X)] = \mathbb{E}[X\mathbb{E}(U|X)] = 0, \tag{6}$$

where we again exploit the law of the iterated expectations. So, we have

(iii). $\operatorname{cov}(X, U) = \mathbb{E}(XU) = 0$.

In the sequel, we restrict attention to linear regressions. In the *simple linear regression model*, we assume that

$$\mathbb{E}(Y|X) = \alpha + \beta X, \tag{7}$$

for some *intercept* parameter α and *slope* parameter β. Note that α and β are parameters as they give some numerical properties, moments as we will see soon, of the population distribution of (X, Y). If U again denotes the error term, we can alternatively write

$$Y = \alpha + \beta X + U \quad \text{and} \quad \mathbb{E}(U|X) = 0.$$

Note that (7) is both linear in the parameters α and β, and linear in the regressor X. More generally, we will allow (7) to be nonlinear in X. For example, the methods that will be developed in this course can handle $\mathbb{E}(Y|X) = \alpha + \beta X^2$ as well. However, we will restrict attention to regression models that are linear in the parameters.

We can derive some further properties of the regression model under the linear regression assumption in (7). First, note that $\operatorname{var}(Y) = \beta^2 \operatorname{var}(X) + \operatorname{var}(U)$, because X and U are uncorrelated. We can simply decompose the variance of Y in the variance "explained" by X and the variance of the error term. This result will be useful later.

Next, note that

$$\mathbb{E}(U) = \mathbb{E}(Y - \alpha - \beta X) = 0 \quad \text{and} \quad \mathbb{E}(XU) = \mathbb{E}[X(Y - \alpha - \beta X)] = 0$$

are the population counterparts to the normal equations (4). The first equation, $\mathbb{E}(U) = 0$, implies that

$$\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X).$$

Also, again using $\mathbb{E}(U)=0$, it is easy to see that the second normal equation implies that

$$0=\mathbb{E}(XU)=\mathbb{E}\left[(X-\mathbb{E}(X))U\right]=\mathbb{E}[(X-\mathbb{E}(X))(Y-\alpha-\beta X)].$$

Substituting α gives

$$\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y)-\beta(X-\mathbb{E}(X))]=0,$$

which implies that

$$\beta=\frac{\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))]}{\mathbb{E}[(X-\mathbb{E}(X))(X-\mathbb{E}(X))]}=\frac{\text{cov}(X,Y)}{\text{var}(X)}.$$

α and β are the population counterparts to the intercept $\hat{\alpha}$ and slope $\hat{\beta}$ of our "best" straight line through the data. This should come as no surprise, as α and β satisfy the population counterparts to the normal equations (4) that define $\hat{\alpha}$ and $\hat{\beta}$.

An other perspective at this is offered by recognizing that $\hat{\alpha}$ and $\hat{\beta}$ were chosen using a *least squares* criterion. From Subsection 2.3.6, we know that conditional expectations are "best" predictors according to a similar, population least squares criterion. In particular, this implies that the α and β equal the a and b that minimize

$$\text{var}(U)=\mathbb{E}\left[(Y-a-bX)^2\right].$$

After all, in the terminology of Subsection 2.3.6, $\mathbb{E}(Y|X)=\alpha+\beta X$ is the predictor $h(X)$ that minimizes

$$\mathbb{E}\left[(Y-h(X))^2\right].$$

over all appropriate, including linear, functions h.

So, $\hat{\alpha}$ and $\hat{\beta}$ are natural estimates of the unknown parameters α and β from our 1995 CPS sample. Because they follow from minimizing a squared error criterion, they are called *(ordinary) least squares* (OLS) estimates. Before we can develop some theory of least squares estimation, we need some additional assumptions.

3.3 The classical assumptions

Suppose we have a data set $((x_1,y_1),\dots,(x_n,y_n))$, which is a realization of a sample $((X_1,Y_1),\dots,(X_n,Y_n))$ from some population distribution $F_{X,Y}$. In the previous section, we have introduced the linear regression model $Y = \alpha + \beta X + U$. We have shown that the regression model implies that $\mathbb{E}(U|X) = 0$, and therefore $\mathbb{E}(U) = 0$. We can apply this regression model directly to the sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ by assuming that each pair (X_i, Y_i) satisfies the model. To this end, stack all regressors in the random sample into a vector $\mathbf{X} = (X_1, \ldots, X_n)$. The linear regression model then gives

Assumption 1. (linear regression) $\mathbb{E}[Y_i|\mathbf{X}] = \alpha + \beta X_i$.

We already know that this implies that $\mathbb{E}(U_i|\mathbf{X}) = 0$, $\mathbb{E}(U_i) = 0$ and $\text{cov}(U_i, X_i) = \mathbb{E}(U_i X_i) = 0$. We also make some assumptions on the second moments of the errors.

Assumption 2. (spherical errors) The errors are *homoskedastic*: $\text{var}(U_i|\mathbf{X}) = \sigma^2$, for some $\sigma > 0$, for all $i = 1, \ldots, n$. Furthermore, they are *uncorrelated*: $\text{cov}(U_i, U_j|\mathbf{X}) = 0$ for all $i, j = 1, \ldots, n$ such that $i \neq j$.

Our actual data set $((x_1, y_1), \ldots, (x_n, y_n))$ should be sufficiently large, and we need sufficient variation in the regressor.

Assumption 3. (sufficient variation) $n > 2$ and $s_X^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2 > 0$.

For most of the course, we simplify the analysis considerably by assuming that the regressors are non-random.

Assumption 4. (deterministic regressors) $X_1, \ldots, X_n$ are deterministic, *i.e.* fixed to the values $x_1, \ldots, x_n$ in repeated sampling.

At a later stage, we will make the additional assumption that the errors are normally distributed, but for now we restrict ourselves to Assumptions 1–4. Before we proceed, the assumptions deserve some additional discussion.

First, whenever we condition on regressors, we condition on *all* regressors $\mathbf{X}$, and never on a single regressor X_i. This makes an, admittedly subtle, difference. For example, Assumption 1 implies that the mean of Y_i only depends on X_i and not on the regressors corresponding to observations other than i. Similar implications hold for the variance, etcetera. This all follows naturally from random sampling, but is slightly weaker.[14]

If the homoskedasticity assumption in Assumption 2 is violated, we say that there is *heteroskedasticity*, in which case $\text{var}(U_i|\mathbf{X})$ would be different for different observations i.

Assumption 3 ensures that we have sufficiently many observations. First of all, we need at least two points to fit a straight a line, which has two parameters. We will also see later that we need at least one more observation to estimate the variance of the error term. The assumption also requires that there is some variation in the regressor. Without variation in the regressor, it would be a constant, so that β is superfluous in the sense that it can only change the level $\alpha + \beta x$, just like α.

Finally, note that with deterministic regressors, repeated samples of n observations all have the same vector of regressors $(x_1, \dots, x_n)$, but different realizations of $(Y_1, \dots, Y_n)$ corresponding to different realizations of the errors $(U_1, \dots, U_n)$. The non-random regressors assumption is usually defended as being valid if the regressor values are chosen by a scientist in an experimental setting.

As $\mathbf{X}$ is non-random in this case, all random variables, in particular the errors, are independent of $\mathbf{X}$. So, in this case the conditioning in Assumptions 1 and 2 has no bite and can be omitted without changing the assumptions. In econometrics we typically deal with observational, or non-experimental, data, and we may be worried that this is not a very appropriate assumption. With random regressors we would draw new realizations of $(X_1, \dots, X_n)$ for each new sample, and $\mathbf{X}$ would be truly random. The notation above already suggests how we can deal with random regressors in the linear regression framework. All assumptions are taken to be *conditional* on the regressors, and so are all the results derived from that. If necessary, the law of the iterated expectations can be applied to translate conditional results into unconditional results. We save a discussion of the more general case of random regressors for later, and suppress the conditioning on $\mathbf{X}$ in the coming subsections. Instead, we will make Assumption 4 explicit by using lower case symbols $x_1, \dots, x_n$ for the regressors throughout.

3.4 Least squares estimation: the Gauss-Markov theorem

In Subsection 3.2 we proposed estimating α and β by ordinary least squares (OLS). The corresponding first order conditions, or normal equations, are given by (4). For the random sample discussed in the previous subsection, the normal equations are

$$\sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta} x_i \right) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i \left(Y_i - \hat{\alpha} - \hat{\beta} x_i \right) = 0, \tag{8}$$

which leads to the OLS estimators

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \quad \text{and} \quad \hat{\beta} = \frac{S_{XY}}{s_X^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}. \tag{9}$$

Here I use lower case x_i to stress that the regressors are assumed to be nonrandom. Note that $\hat{\alpha}$ is simply the sample mean of the regressand if $\bar{x} = 0$, *i.e.* if the regressor is taken in deviation from its mean. In this case, the sample average of $\hat{\beta}x_i$ is independent of $\hat{\beta}$.

We will now derive some properties of these estimators. First note that the estimators are *linear*, in the sense that they are linear functions of the random variables $Y_1, \ldots, Y_n$ for given values of the regressors $x_1, \ldots, x_n$. This is very convenient, as it allows us to apply the various results for linear combinations of random variables we have seen earlier.

The main result is the *Gauss-Markov theorem*.

Proposition 3. *Under the classical Assumptions 1–4, the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are the* best linear unbiased estimators *(BLUE).*

Here, "best" means "most efficient", or "minimum variance". The Gauss-Markov theorem states that the OLS estimators are unbiased, and the most precise among all possible linear, unbiased estimators. We will discuss the unbiasedness, efficiency (in the class of linear estimators), and some other properties of the OLS estimators in some more detail next.

3.4.1 Unbiasedness

If Assumptions 1, 3 and 4 hold, $\hat{\alpha}$ and $\hat{\beta}$ are unbiased estimators of α and β. *Note that we do not need spherical errors (or normality) for unbiasedness.*

Denoting $\bar{U} = n^{-1}\sum_{i=1}^n U_i$, we have that

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(\alpha + \beta x_i + U_i - \alpha - \beta\bar{x} - \bar{U})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})(U_i - \bar{U})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta + \frac{\left[\sum_{i=1}^n (x_i - \bar{x})U_i\right] - \left[\bar{U}\sum_{i=1}^n (x_i - \bar{x})\right]}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta + \frac{\sum_{i=1}^n (x_i - \bar{x})U_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.
\end{aligned} \tag{10}$$

Note that we need Assumption 3 for the estimator to be well defined, *i.e.* the denominator to be positive. Assumption 1 is used in the second equality. Taking expectations gives

$$\begin{aligned}\mathbb{E}(\hat\beta) &= \mathbb{E}\left[\beta + \frac{\sum_{i=1}^n (x_i-\bar x)U_i}{\sum_{i=1}^n (x_i-\bar x)^2}\right]\\ &= \beta + \mathbb{E}\left[\frac{\sum_{i=1}^n (x_i-\bar x)U_i}{\sum_{i=1}^n (x_i-\bar x)^2}\right]\\ &= \beta + \frac{\sum_{i=1}^n (x_i-\bar x)\mathbb{E}(U_i)}{\sum_{i=1}^n (x_i-\bar x)^2} = \beta.\end{aligned}$$

The third equality follows from the non-randomness of the regressor (Assumption 4), and the last equality from Assumption 1.

Similarly, we have for $\hat\alpha$ that

$$\hat\alpha = \bar Y - \hat\beta \bar x = \alpha - \bar x(\hat\beta - \beta) + \bar U, \tag{11}$$

so that

$$\mathbb{E}(\hat\alpha) = \mathbb{E}\left[\alpha - \bar x(\hat\beta-\beta) + \bar U\right] = \alpha.$$

The unbiasedness of $\hat\beta$ is used in the last equality.

3.4.2 Efficiency

The Gauss-Markov theorem states that the OLS estimators are BLUE. Note again that the Gauss-Markov theorem only states that the OLS estimators are (the most) efficient in the class of linear unbiased estimators. It does not exclude the possibility that there are non-linear or biased estimators that have lower variance than the OLS estimators. I will give an example at the end of this subsection. Gujarati (1995), Section 3A.6, provides some discussion. We will not prove the efficiency part of the Gauss-Markov theorem, but only give an example that provides some intuition.

Example 43. Consider again the simpler example of estimating the mean μ of a random variable X with finite variance σ^2. Suppose we have an i.i.d. sample $(X_1, \ldots, X_n)$ from the distribution F_X of X. In Example 31, we proposed estimating μ by the sample mean $\bar X_n$. We also showed that $\text{var}(\bar X_n) = \sigma^2/n$ in this case.

We wonder whether there exists an estimator $\hat\mu$ that is more efficient than $\bar X_n$. This estimator should be unbiased, $\mathbb{E}(\hat\mu) = \mu$, and have lower variance than $\bar X_n$, $\text{var}(\hat\mu) <$

$\text{var}(\bar{X}_n)$. In general, this is a difficult question to answer. However, if we restrict attention to linear estimators $\hat{\mu}$, so that $\hat{\mu} = \sum_{i=1}^n w_i X_i$ for some weights $w_i \in \mathbb{R}$, the problem becomes quite manageable. Note that $\bar{X}_n$ is a special case in which $w_i = n^{-1}$.

The simpler question now is whether there exists weights $w_1, \dots, w_n$ such that

$$\mathbb{E}\left(\sum_{i=1}^n w_i X_i\right) = \mu \quad \text{and} \quad \text{var}\left(\sum_{i=1}^n w_i X_i\right) < \text{var}\left(\bar{X}_n\right).$$

The first requirement, unbiasedness, demands that $\sum_{i=1}^n w_i = 1$ (actually, unbiasedness also ensures that we cannot add an additional "free" constant to our estimator). The variance of $\hat{\mu}$ is given by $\text{var}(\sum_{i=1}^n w_i X_i) = \sigma^2 \sum_{i=1}^n w_i^2$. Now, note that

$$\begin{aligned}
\text{var}\left(\sum_{i=1}^n w_i X_i\right) &= \sigma^2 \sum_{i=1}^n w_i^2 \\
&= \sigma^2 \sum_{i=1}^n \left(w_i - \frac{1}{n} + \frac{1}{n}\right)^2 \\
&= \sigma^2 \sum_{i=1}^n \frac{1}{n^2} + \sigma^2 \sum_{i=1}^n \left(w_i - \frac{1}{n}\right)^2 + \sigma^2 \sum_{i=1}^n \left(w_i - \frac{1}{n}\right) \frac{1}{n} \\
&= \text{var}\left(\bar{X}_n\right) + \sigma^2 \sum_{i=1}^n \left(w_i - \frac{1}{n}\right)^2 \geq \text{var}\left(\bar{X}_n\right).
\end{aligned}$$

So, the variance of each linear unbiased estimator of μ is at least as large as the variance of $\bar{X}_n$: $\bar{X}_n$ is BLUE.

Note that it is crucial to restrict attention to unbiased estimators. A counterexample is an estimator that is always 0. This is a (trivial) linear estimator. It has zero variance, and therefore lower variance than the OLS estimator. However, it is not unbiased.

3.4.3 Standard errors and covariance

From equation (10), we have that

$$\begin{aligned}
\operatorname{var}(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^n \frac{(x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2} U_i\right)^2\right] \\
&= \frac{\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n (x_i-\bar{x})(x_j-\bar{x}) U_i U_j\right]}{\left[\sum_{i=1}^n (x_i-\bar{x}_n)^2\right]^2} \\
&= \frac{\sum_{i=1}^n (x_i-\bar{x})^2 \mathbb{E}(U_i^2)}{\left[\sum_{i=1}^n (x_i-\bar{x}_n)^2\right]^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x}_n)^2} = \frac{\sigma^2}{(n-1)s_X^2}
\end{aligned}$$

Here, we are again exploiting the non-randomness of the regressor. The fourth equality follows from the uncorrelatedness of the errors, and the fifth equality from the homoskedasticity in Assumption 2.

The *standard error* of $\hat{\beta}$ is just its standard deviation, the square root of $\operatorname{var}(\hat{\beta})$. Note that the precision of $\hat{\beta}$, as measured by its variance, decreases with the variance of the error σ^2, and increases with the sample variance of the regressor s_X^2 and the sample size. This makes sense. If the variation in the regressor is large relative to the error variance, it is easier to learn about the (linear) relation with the dependent variable. Also, if we have a larger sample, we have more information about this relation. This is not unlike the inverse dependence of the variance of the sample mean on the sample size (see Example 31).

Using (11), we find for $\hat{\alpha}$ that

$$\begin{aligned}
\operatorname{var}(\hat{\alpha}) &= \mathbb{E}\left[(\hat{\alpha}-\alpha)^2\right] \\
&= \mathbb{E}[(-\bar{x}(\hat{\beta}-\beta)+\bar{U})^2] \\
&= \operatorname{var}(\bar{U}) + \bar{x}^2 \operatorname{var}(\hat{\beta}) - 2\bar{x}\mathbb{E}\left[\bar{U}(\hat{\beta}-\beta)\right] \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x}_n)^2},
\end{aligned}$$

because

$$\begin{aligned}\operatorname{cov}(\bar{U},\hat{\beta}) = \mathbb{E}\left[\bar{U}(\hat{\beta}-\beta)\right] &= \mathbb{E}\left[\bar{U}\frac{\sum_{i=1}^n (x_i-\bar{x})U_i}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]\\ &= \frac{\sum_{i=1}^n (x_i-\bar{x})\mathbb{E}[U_i\bar{U}]}{\sum_{i=1}^n (x_i-\bar{x})^2}\\ &= \frac{\sigma^2}{n}\frac{\sum_{i=1}^n (x_i-\bar{x})}{\sum_{i=1}^n (x_i-\bar{x})^2} = 0.\end{aligned}$$

(Check where the various assumptions are invoked in these derivations!) Again, the precision of $\hat{\alpha}$ depends inversely on the error variance, and increases with the sample size. We have seen that if $\bar{x}=0$, $\hat{\alpha}$ is simply a sample mean, so that its variance is the variance of the sample mean, σ^2/n. If $\bar{x}\neq 0$, the $\operatorname{var}(\hat{\alpha})$ also depends on $\operatorname{var}(\hat{\beta})$, and therefore on s_X^2. We will give some intuition for that next.

First, note that the covariance of $\hat{\alpha}$ and $\hat{\beta}$ is given by

$$\begin{aligned}\operatorname{cov}(\hat{\alpha},\hat{\beta}) &= \mathbb{E}\left[(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\right]\\ &= \mathbb{E}\left[(\bar{U}-\bar{x}(\hat{\beta}-\beta))(\hat{\beta}-\beta)\right]\\ &= \mathbb{E}\left[\bar{U}(\hat{\beta}-\beta)\right]-\bar{x}\operatorname{var}(\hat{\beta})\\ &= -\bar{x}\operatorname{var}(\hat{\beta}) = -\bar{x}\frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x}_n)^2}\end{aligned}$$

If $\bar{x}>0$ ($\bar{x}<0$), then $\hat{\alpha}$ and $\hat{\beta}$ are negatively (positively) related. This makes sense. Unless $\bar{x}=0$, the average value of $\hat{\beta}x_i$ in the sample depends on $\hat{\beta}$. So, if $\hat{\beta}$ changes, this has to be compensated by a change in $\hat{\alpha}$. If $\bar{x}=0$, for example because the x_i are taken to be in deviation from their sample mean, then $\hat{\alpha}$ and $\hat{\beta}$ are not correlated.

The variances of the estimators depend on the unknown parameter σ^2. As in case of estimating a simple mean (Example 33), we can estimate the variance of the estimators by substituting an unbiased estimator for σ^2. This is discussed in Subsection 3.6.

3.4.4 Asymptotic properties: consistency and asymptotic normality

Proving asymptotic properties is well beyond the scope of this course. It is however useful to be aware of some nice properties of OLS estimators in large samples. Under some additional conditions, notably on the way the vector of regressors $(x_1,\dots,x_n)$ grows if we increase the sample size, OLS estimators are consistent. The proof of this result exploits a law of large numbers like Proposition 2.

In the next subsection, we show that the OLS estimators are normally distributed if the errors are normal. However, even if the errors are not assumed to be normal, it can be proven that the estimators are "asymptotically normal", again under some auxiliary conditions. This means that their distribution can be approximated by a normal distribution in sufficiently large samples, even if we do not assume that the errors are normally distributed. The proof of this result exploits a central limit theorem like Proposition 1. See Example 28 for a simple example.

3.4.5 Additional results for normal models

In the classical linear regression model, it is often assumed that the error terms are (jointly) normally distributed. As uncorrelated (jointly) normal random variables are independent, Assumption 2, with Assumption 1, then implies

Assumption 5. (normality) Conditional on $\mathbf{X}$, the errors $U_1, \ldots, U_n$ are i.i.d. and normally distributed with mean 0 and variance σ^2.

Normality is sometimes defended by referring to a central limit theorem like Proposition 1. If the error term is the sum of many omitted variables or other small errors, it will, under some conditions, be approximately normal.

Under normality, two additional results can be derived. First, the OLS estimators are (jointly) normally distributed. If the errors U_i are normal, the regressands Y_i are normal as well. As $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of the Y_i, for given regressors $x_1, \ldots, x_n$, they are normal too, with the means and variances derived before. We can actually determine the joint distribution of $(\hat{\alpha}, \hat{\beta})$ in this case: $(\hat{\alpha}, \hat{\beta})$ is bivariate normal. The bivariate normal distribution is fully characterized by the means and variances of the two estimators, and their covariance (see Subsection 3.4.3 and Gujarati, 1995, Exercise 4.1).

Second, the OLS estimators are *best unbiased estimators*, and not just best *linear* unbiased estimators. So, in the normal classical model, there are no other unbiased estimators, either linear or non-linear, that are more efficient, *i.e.* have lower variance.

3.5 Residual analysis and the coefficient of determination

So far, we have shown that the OLS estimators are unbiased and relatively efficient, and we have derived their standard errors and covariance. We have also seen that the estimators are (jointly) normal if the errors are assumed to be normal. Now that we have learned quite a bit about estimation of the simple linear regression model, we are ready to discuss some further analysis of the regression model.

Once we have estimated the model, we can use the model to predict the regressand for given values of the regressor. We denote the predicted, or fitted, value of Y_i for a given value of x_i by

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} x_i.$$

The corresponding *OLS (or fitted) residual* is then given by

$$\hat{U}_i = Y_i - \hat{Y}_i.$$

For a given data set $((x_1, y_1), \dots, (x_n, y_n))$, the actual predictions and residuals are denoted by $\hat{y}_i$ and $\hat{u}_i$, respectively. From the normal equations (8), it immediately follows that

$$\sum_{i=1}^{n} \hat{U}_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i \hat{U}_i = 0. \tag{12}$$

Note that these *are* the normal equations defining the OLS estimators. (Obviously, the normal equations also hold for the OLS estimates and realizations $\hat{u}_i$ of the residuals.)

For the linear population regression of Subsection 3.2, we found that $\text{var}(Y) = \beta^2 \,\text{var}(X) + \text{var}(U)$. Crucial for the derivation of this simple variance decomposition was that $\mathbb{E}(XU) = 0$ if U is a regression error. As the sample equivalent of this condition, $\sum_{i=1}^{n} x_i \hat{U}_i = 0$, holds as well, it seems that we can find an equally simple decomposition of the sample variance. Fortunately, this is indeed the case.

The result is usually stated in terms of a decomposition of the *total sum of squares* (*i.e.*, without dividing by $n-1$),

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2,$$

into the *explained (or fitted) sum of squares*,

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2,$$

and the *residual sum of squares*,

$$RSS = \sum_{i=1}^{n} (\hat{U}_i - \bar{\hat{U}})^2 = \sum_{i=1}^{n} \hat{U}_i^2.$$

$\bar{\hat{Y}}$ and $\bar{\hat{U}}$ are just the sample means of $\hat{Y}_i$ and $\hat{U}_i$. Note that by the first normal equation in (12), $\bar{\hat{U}} = 0$, so that $\bar{\hat{Y}} = \bar{Y} - \bar{\hat{U}} = \bar{Y}$. We will show that, in analogy to the result for the population variance,

$$TSS = ESS + RSS. \tag{13}$$

First, note that

$$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y} + \hat{U}_i)^2 = ESS + RSS + \sum_{i=1}^{n} \hat{U}_i(\hat{Y}_i - \bar{Y}).$$

This reduces to (13), as

$$\sum_{i=1}^{n} \hat{U}_i(\hat{Y}_i - \bar{Y}) = \hat{\beta} \sum_{i=1}^{n} x_i \hat{U}_i + (\hat{\alpha} - \bar{Y}) \sum_{i=1}^{n} \hat{U}_i = 0,$$

because of the normal equations (12).

The share of the total variance that is explained by the regressor is called the *coefficient of determination*, and denoted by

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}$$

Because of (13), and as $ESS \geq 0$ and $RSS \geq 0$,

$$0 \leq R^2 \leq 1.$$

The coefficient of determination is a measure of the "fit" of the regression, and is usually reported along with the parameter estimates. If R^2 is close to 1, almost all variation in the regressand can be explained (linearly) by the regressor. We can say that the fit is very good. If R^2 is close to 0, the regressor hardly predicts the regressand, and just predicting the regressand by a constant would not have been much worse. The fit is bad in this case.

It is interesting to note that the coefficient of determination can be reinterpreted as a squared sample correlation coefficient. Note that

$$\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}) - \sum_{i=1}^{n} \hat{U}_i(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n} (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y}),$$

so that

$$\begin{aligned}R^2 &= \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2} \\ &= \frac{\left[\sum_{i=1}^n (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})\right]^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2 \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2},\end{aligned}$$

which is the squared sample correlation between the actual and the fitted values of Y_i. We will denote this, nonnegative, sample correlation coefficient (and *not* minus this coefficient) by R. In the simple linear regression model we are studying here, it is easy to show that $R = |\hat{\rho}_{X,Y}|$, the absolute value of the sample correlation between the regressor and the regressand.[15] As the sample correlation coefficient is symmetric, this implies that R, and therefore the coefficient of determination R^2, are the same for the regression of Y on X and the reverse regression of X on Y.

In the multiple regression of Section 4, in which we regress Y on more than one regressor, this equivalence and symmetry break down. After all, we can still define R^2 as the fraction of the variance of the regressand explained by the regressors and R as the correlation between the predicted and the actual values of the regressand. However, the sample correlation coefficient itself does not extend to more than 2 of the variables involved in the regression. So, even though R is just the absolute value of the sample correlation between X and Y in the simple regression model, it will serve as a natural extension of the correlation coefficient in the multiple regression model. For this reason, R is called the *multiple correlation coefficient*. We will return to this in Section 4.

3.6 Estimating the variance of the error term

In Subsection 3.4, we have derived the variances of the estimators. These can be used to compute and report standard errors with our estimates. One problem is that the variances depend on the unknown variance σ^2 of the error term. As in Example 33, we can estimate the variance of the estimators by substituting an unbiased estimator for σ^2. In this subsection, we derive such an estimator.

As σ^2 is the variance of U_i, we could naively propose to estimate σ^2 by

$$n^{-1} \sum_{i=1}^n U_i^2 = n^{-1} \sum_{i=1}^n (Y_i - \alpha - \beta x_i)^2.$$

As in Example 32, this estimator is unbiased but not feasible, as we do not know α and β. If we substitute the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$, we get the RSS, $\sum_{i=1}^n \hat{U}_i^2$, divided by n, which is a known function of the sample. Using (11), we find that

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^n \hat{U}_i^2\right] &= \mathbb{E}\left[\sum_{i=1}^n(-(\hat{\alpha}-\alpha)-x_i(\hat{\beta}-\beta)+U_i)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n(-(x_i-\bar{x})(\hat{\beta}-\beta)+U_i-\bar{U})^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^n((x_i-\bar{x})^2(\hat{\beta}-\beta)^2+(U_i-\bar{U})^2-2(x_i-\bar{x})(\hat{\beta}-\beta)(U_i-\bar{U}))\right] \\
&= \left[\sum_{i=1}^n(x_i-\bar{x})^2\right]\text{var}(\hat{\beta})+(n-1)\sigma^2-2\mathbb{E}\left[\sum_{i=1}^n(x_i-\bar{x})(\hat{\beta}-\beta)(U_i-\bar{U})\right] \\
&= \sigma^2+(n-1)\sigma^2-2\sigma^2 \\
&= (n-2)\sigma^2.
\end{aligned}$$

Here, I have used that $\mathbb{E}[(U_i-\bar{U})^2]=(n-1)\sigma^2$ (see Example 32), and that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{i=1}^n(x_i-\bar{x})(\hat{\beta}-\beta)(U_i-\bar{U})\right] \\
&\quad= \mathbb{E}\left[\sum_{i=1}^n(x_i-\bar{x})\frac{\sum_{j=1}^n(x_j-\bar{x})U_j}{\sum_{j=1}^n(x_j-\bar{x})^2}\left(U_i-\bar{U}\right)\right] \\
&\quad= \left[\frac{\sum_{i=1}^n(x_i-\bar{x})^2\mathbb{E}\left({U_i}^2\right)}{\sum_{i=1}^n(x_i-\bar{x})^2}-\sum_{i=1}^n(x_i-\bar{x})\frac{\sum_{j=1}^n(x_j-\bar{x})\mathbb{E}\left(U_j\bar{U}\right)}{\sum_{j=1}^n(x_j-\bar{x})^2}\right] \\
&\quad= \sigma^2.
\end{aligned}$$

So, an unbiased estimator of σ^2 is

$$\hat{\sigma}^2=\frac{\sum_{i=1}^n \hat{U}_i^2}{n-2}.$$

Note that in Example 32, we only substituted the sample mean in the sum of squares, and we divided by $n-1$. We now need to estimate two parameters to compute the residuals, and divide by $n-2$. We will see similar results in the multiple regression model later on.

3.7 Some practical specification issues

This is related to Gujarati (1995), Chapter 6.

3.7.1 Regression through the origin

Some economic models suggest linear regressions without an intercept term α, so that

$$Y = \beta X + U \quad \text{and} \quad \mathbb{E}(U|X) = 0. \tag{14}$$

We could estimate (14) with OLS by choosing $\hat{\beta}$ as to maximize a least squares criterion. The first order condition is

$$\sum_{i=1}^{n} x_i \left(Y_i - \hat{\beta} x_i \right) = 0.$$

This is equivalent to the second normal equation in (8) without an intercept term. The corresponding least squares estimator of β is

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2},$$

which only deviates from the slope estimator in (9) in that x_i and Y_i are not in deviation from their sample means.

As we do not have the first normal equation in (8), which corresponds to the omitted intercept parameter α, the residuals $\hat{U}_i = Y_i - \hat{\beta} x_i$ do not necessarily add to zero. For the same reason, the analysis of the coefficient of determination in Subsection 3.5 is not valid anymore.

We could spend some energy on deriving a coefficient of determination for the model without an intercept. However, in most cases, a better approach is to estimate a model with an intercept, even if economic theory predicts a model without an intercept. After all, the standard regression model, with an intercept, contains the model without an intercept as a special case, $\alpha = 0$. So, if the theory is correct, we should expect the estimate of the intercept, $\hat{\alpha}$, to be close to zero. If it is far away from zero, we may suspect our theory is not right. The testing procedures discussed in Subsection 3.8 can be used to formalize this idea.

Example 44. In the capital asset pricing model (CAPM), the expected return on a security in excess of the risk-free rate of return is proportional to the expected excess return on an appropriately chosen market portfolio. Denoting the returns on a particular security i (or portfolio) over period t by $R_{i,t}$, and the returns on the market portfolio over the same period by $R_{M,t}$, we can write

$$R_{i,t} - r_{f,t} = \beta_i (R_{M,t} - r_{f,t}) + U_{i,t}. \tag{15}$$

Here, $r_{f,t}$ is the risk-free rate of return over the period t. Each security i has its own *beta-coefficient* β_i, which is a measure of the *systematic* or *non-diversifiable risk* of a security.

Obviously, if the CAPM is true, and if we are somehow involved in trading securities, we would be very interested in knowing the beta-coefficients of the securities we are trading. In principle, we could use equation (15) to estimate the beta-coefficient β_i of each security i. Suppose we have data on the returns $R_{i,t}$ on a particular common stock i, for example IBM, at various times t. We could use a stock index to approximately measure the returns $R_{M,t}$ on the market portfolio, and the returns on some U.S. Treasury bill to measure $r_{f,t}$. If we observe IBM, market, and risk-free returns for sufficiently many periods, we could estimate β_i in (15) by OLS without an intercept.

However, instead of estimating an equation like (15), we may actually prefer to estimate a standard model with an intercept, like

$$R_{i,t} - r_{f,t} = \alpha_i + \beta_i (R_{M,t} - r_{f,t}) + U_{i,t}.$$

After all, it includes (15) as a special case, $\alpha_i = 0$. So, by estimating this model, we can actually check whether the CAPM is correct by checking whether $\hat{\alpha}_i$ is sufficiently close to 0. This is in the domain of hypothesis testing, which we will study in Subsection 3.8. If we find that $\hat{\alpha}_i$ is far away from 0, the CAPM specification (15) is likely to be wrong. In this case, we should be happy that we didn't restrict attention to that specification, even though it is suggested by economic theory.

Warning: this example is different from the first CAPM example in Gujarati (1995), Section 6.1. Here, we focus on estimating β_i. Gujarati assumes that we already know β_i from some other source, and includes it as a regressor. See instead Gujarati's second example and the exercises to which he refers.

3.7.2 Scaling

It is intuitively clear that the regression results depend on the scaling of the variables. For example, we expect different estimates in the earnings and schooling example if we measure earnings in dollar cents instead of dollars, or if we measure schooling in months instead of years. To investigate this, consider the linear regression of Y on X,

$$Y = \alpha + \beta X + U \quad \text{and} \quad E(U|X) = 0. \tag{16}$$

Suppose we rescale Y and X by multiplying Y by some number w_Y and X by some number w_X. Denote the rescaled variables by

$$Y^* = w_Y Y \quad \text{and} \quad X^* = w_X X.$$

Then, multiplying both sides of (16) by $w_Y w_X$, we have that

$$w_X Y^* = w_Y w_X \alpha + \beta w_Y X^* + w_Y w_X U.$$

so that we have the rescaled model

$$Y^* = \alpha^* + \beta^* X^* + U^* \quad \text{and} \quad E(U^*|X^*) = 0, \tag{17}$$

with $\alpha^* = w_Y \alpha$, $\beta^* = w_Y \beta / w_X$, and $U^* = w_Y U$.

Now, suppose we have a sample $((x_1, Y_1), \dots, (x_n, Y_n))$ corresponding to the original model in (16) and a rescaled sample $((x_1^*, Y_1^*), \dots, (x_n^*, Y_n^*)) = ((w_X x_1, w_Y Y_1), \dots, (w_X x_n, w_Y Y_n))$ corresponding to the rescaled model in (17). Denote the OLS estimators for both models by $\hat\alpha$, $\hat\beta$, $\hat\sigma^2$ and $\hat\alpha^*$, $\hat\beta^*$, $\hat{\sigma^*}^2$, respectively. It is easy to derive that

$$\begin{aligned}
\hat\beta^* &= \frac{S_{X^*Y^*}}{s_{X^*}^2} = \frac{\sum_{i=1}^n (x_i^* - \bar x^*)(Y_i^* - \bar Y^*)}{\sum_{i=1}^n (x_i^* - \bar x^*)^2} = \frac{\sum_{i=1}^n w_X w_Y (x_i - \bar x)(Y_i - \bar Y)}{\sum_{i=1}^n w_X^2 (x_i - \bar x)^2} \\
&= \frac{w_Y}{w_X} \frac{S_{XY}}{s_X^2} = \frac{w_Y}{w_X} \hat\beta, \\
\hat\alpha^* &= \bar Y^* - \hat\beta^* \bar x^* = w_Y \bar Y^* - \frac{w_Y}{w_X} \hat\beta w_X \bar x = w_Y \hat\alpha, \text{ and} \\
\hat{\sigma^*}^2 &= \frac{\sum_{i=1}^n \hat{U^*}_i^2}{n-2} = \frac{\sum_{i=1}^n \left(Y_i^* - \hat\alpha^* - \hat\beta^* x_i^*\right)^2}{n-2} = w_Y^2 \frac{\sum_{i=1}^n \left(Y_i - \hat\alpha - \hat\beta x_i\right)^2}{n-2} = w_Y^2 \hat\sigma^2,
\end{aligned}$$

so that

$$\text{var}(\hat\alpha^*) = w_Y^2 \,\text{var}(\hat\alpha) \quad \text{and} \quad \text{var}(\hat\beta^*) = \left(\frac{w_Y}{w_X}\right)^2 \text{var}(\hat\beta).$$

Also, the R^2 corresponding to both regressions is the same. After all, using that $\hat{Y}_i^* = \hat{\alpha}^* + \hat{\beta}^* x_i^* = w_Y(\hat{\alpha} + \hat{\beta} x_i) = w_Y \hat{Y}_i$, we have that

$$R^{*2} = \frac{\sum_{i=1}^n (\hat{Y}_i^* - \bar{Y}^*)^2}{\sum_{i=1}^n (Y_i^* - \bar{Y}^*)^2} = \frac{\sum_{i=1}^n (w_Y \hat{Y}_i - w_Y \bar{Y})^2}{\sum_{i=1}^n (w_Y Y_i - w_Y \bar{Y})^2} = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = R^2.$$

These results are useful if we want to translate estimation results for one model to results for a rescaled version of that model, without estimating the rescaled model. For example, if we have estimated a regression of earnings in dollars on years of schooling, we can translate the estimation results directly into results for a regression of earnings in dollar cents on months of schooling.

However, the main purpose of this exposition is to show that appropriate scaling of the variables under study may be important. For example, if you would specify schooling in seconds and hourly earnings in billions of dollars, β will be measured in billions of dollars per second of schooling, and is very small. It may be so small that it causes numerical problems. So, even though you should get exactly the same regression results up to scale, your computer may have problems dealing with the very small and large numbers it encounters. So, always make sure to appropriately scale variables in a regression, and never forget to mention the units of measurement with the regression results.

3.7.3 Specifying the regressor in deviation from its mean

We have already seen some advantages of specifying the regressor in deviation from its sample mean. If the regressor is taken in deviation from its mean, the intercept estimator $\hat{\alpha}$ equals the sample average of the regressand, and is uncorrelated with the slope estimator $\hat{\beta}$. The slope estimator $\hat{\beta}$ itself is unaffected.

In general, if we do not take the regressor in deviation from its mean, the intercept estimate takes fairly arbitrary values depending on the scale of the regressor and the slope estimate. In such cases, we may prefer the simple interpretation of the intercept after taking deviations from the mean.

There are cases, however, where it is easier to interpret the intercept without taking deviations from the mean. In problem set 3, you have estimated a regression of log earnings on a sex “dummy”, a regressor that is 1 for females and 0 for males. In this case, we have seen that the expected log earnings are α for males and $\alpha + \beta$ for females.

3.7.4 Transforming the regressand and the regressor

We only require the linear regression model to be linear in the parameters. Indeed, we have already seen examples of linear models that were non-linear in the variables. For example, in the regression of earnings on schooling, we specified earnings in logs. This allowed for an attractive interpretation of the schooling coefficient as the percentage increase in earnings per additional year of schooling. In this subsection, we briefly review some variable transformations that can be useful in econometric research.

The most common transformation is the logarithmic transformation, which can be applied to positive variables, like wages or prices. As we have seen earlier, changes in logs of variables correspond to *relative* changes of the variables. For small changes Δx of some variable x,

$$\ln(x + \Delta) - \ln(x) \approx \frac{\Delta x}{x}$$

gives the corresponding percentage change (divided by 100) of x. More formally, if we let Δ go to 0, we get the derivative $d\ln(x)/dx = 1/x$, and the exact relation

$$d\ln(x) = \frac{dx}{x}.$$

In problem set 3, we have estimated a regression of earnings on schooling. In a first regression, we specified both earnings and schooling linearly, which gives the regression model

$$W = \alpha + \beta S + U,$$

where W is average hourly earnings (in US dollars) and S is years of schooling. In this model, β simply gives the additional earnings in dollars per extra year of schooling in the population. As an alternative, we specified earnings in logs, which gives

$$\ln(W) = \alpha + \beta S + U,$$

In this model, 100β is the (expected) percentage increase in earnings corresponding to one extra year of schooling in the population. We may have a preference of one model over the other because of these kinds of differences in interpretation. In particular, one model may be consistent with an economic theory we have, and the other may not.

There is also a statistical reason why we may prefer one model over the other. Sometimes, we want to assume that the error term U has a normal distribution. A normal random variable assumes all values between $-\infty$ and ∞. So, if we assume that U has a normal distribution, then, for any given S, the left hand side variable assumes all values between $-\infty$ and ∞. In the first, linear, model, this implies that earnings are sometimes negative. Therefore, if earnings are always positive, we may prefer to use log earnings as the regressand. After all, log earnings can assume values between $-\infty$ and ∞, which is consistent with the normality assumption.

If we specify both the regressand and the regressor in logs, the slope parameter gives the (expected) percentage (relative) change in the regressor corresponding to a percentage change in the regressand. For, example, if Q is the quantity demanded of some good at a price P, we can specify a demand equation

$$\ln(Q) = \alpha + \beta \ln(P) + U.$$

In this *log-linear* model, β is the price elasticity of demand, the (expected) percentage change in demand Q corresponding to a percentage change in price P. If we forget the error term for a second, and write $\ln(q) = \alpha + \beta \ln(p)$ for some given quantity q and price p, we have that

$$\beta = \frac{d\ln(q)}{d\ln(p)} = \frac{dq/q}{dp/p}.$$

Typically, we would expect that $\beta < 0$. Again, the slope parameter β has a nice economic interpretation after taking appropriate transformations of the variables.

It should be noted that the log-linear demand model implicitly assumes that the price elasticity does not vary with price. We could alternatively specify

$$Q = \alpha + \beta \ln(P) + U.$$

If we again forget the error term and write $q = \alpha + \beta \ln(p)$, we have that $\beta = dq/(dp/p)$. So, β/q is the price elasticity of demand, which now decreases (in absolute value) with demand, or, if $\beta < 0$, increases (in absolute value) with price.

Gujarati (1995), Chapter 6, discusses some more variable transformations that are useful in econometrics. You are invited to read this chapter, but we will not discuss this any further in class.

We end this subsection with an important warning. The coefficient of determination R^2 *cannot* be used to compare the fit of models with different regressands. So, if we want to know whether a regression of log earnings on schooling fits the data better than a regression of earnings on schooling, we cannot simply compare the coefficients of determination of both regressions. The coefficient of determination can, however, be used to compare the fit of models with the same regressand, but different (transformations of) regressors.

3.8 Interval estimation and hypothesis testing

Warning: in the context of testing in linear regression models, I use α for both the intercept parameter and the test size.

So far, we have concentrated on point estimation of the parameters in the linear regression model. Typically, we are not just interested in estimating the parameters, but also in testing hypotheses we have formulated about the parameters in the model.

Example 45. The CAPM in Example 44 predicts that (expected) excess returns on a particular security and (expected) excess market returns are proportional. This implies that the intercept $\alpha = 0$ in a linear regression of excess returns of a security on excess market returns. We suggested to estimate a model that includes an intercept α anyhow, and to subsequently test whether α_0. In this case, it is natural to maintain the hypothesis that the CAPM is right until we find strong evidence against it. So, our null hypothesis is $H_0 : \alpha = 0$. If we have no a priori idea whether the security will be underpriced or overpriced relative to the CAPM if H_0 is rejected, we could pick $H_1 : \alpha \neq 0$ as our alternative hypothesis. This gives a two-sided test. Alternatively, if we suspect that a security is overpriced, and we would like our test to have much power against that alternative, we could pick $H_1 : \alpha > 0$.

This is an example of a test involving hypotheses regarding the intercept α. Often, we also want to test hypotheses about the slope β. For example, if we regress earnings on schooling, we may want to test whether earnings and schooling are related in the data, which gives $H_0 : \beta = 0$.

Subsection 2.4.3 provided a first introduction to hypothesis testing. We have seen a couple of tests involving hypotheses about population means. We also discussed the close connection between two-sided tests and interval estimation, which was introduced at the end of Subsection 2.4.2. In all cases, we assumed normality of the data, leading to convenient normal test statistics (Z-tests), or test statistics with related distributions like the t-distribution (t-tests).

Fortunately, hypothesis testing in the normal linear regression model is very similar. After all, the OLS estimators $\hat{\alpha}$ and $\hat{\beta}$ are also (jointly) normally distributed if, in addition to the classical assumptions, we assume that the regression errors U_i are (jointly) normally distributed. So, the Z-statistics

$$Z_\alpha = \frac{\hat{\alpha}-\alpha}{\sqrt{\text{var}(\hat{\alpha})}} = \frac{\hat{\alpha}-\alpha}{\sqrt{\sigma^2(1/n+\bar{x}^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2)}}$$

and

$$Z_\beta^* = \frac{\hat{\beta}-\beta}{\sqrt{\text{var}(\hat{\beta})}} = \frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2}},$$

have standard normal distributions. If we would know σ, we could again construct confidence intervals based on Z_α and Z_β^*. Also, a test of, for example, $H_0 : \alpha = 0$ against $H_1 : \alpha \neq 0$ could again be based on the test statistic Z_0.

Typically, we do not know σ^2, but we can substitute the unbiased OLS estimator $\hat{\sigma}^2$, which gives the t-statistics

$$T_\alpha = \frac{\hat{\alpha}-\alpha}{\sqrt{\hat{\sigma}^2(1/n+\bar{x}^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2)}} \quad \text{and} \quad T_\beta^* = \frac{\hat{\beta}-\beta}{\sqrt{\hat{\sigma}^2/\sum_{i=1}^n(x_i-\bar{x}_n)^2}}. \tag{18}$$

The only substantial difference with the examples discussed in class is that these statistics have t-distributions with $n-2$ degrees of freedom instead of $n-1$.[16] Note that, not coincidentally, we also divided the sum of squared residuals by $n-2$ to get the unbiased estimator $\hat{\sigma}^2$ of σ^2.

Statistical packages like STATA typically not only report estimates of the OLS parameters, but also estimates of the corresponding standard errors. As the denominators of the t-statistics in (18) are simply these standard errors, confidence intervals and test statistics are readily computed. For example, suppose that we run the regression suggested by the CAPM in Example 44 for a particular security. A test of $H_0 : \alpha = 0$ against, say, $H_1 : \alpha \neq 0$, as in Example 45, can be based on t_0. This t-statistic can simply be computed as $\hat{\alpha}$ divided by the corresponding standard error, both of which can be read directly from the regression output. Note that standard regression output typically reports this particular t_0, and also t_0^*, and 95%-confidence intervals based on t_α and t_β^*. However, if we want to perform other t-tests, or compute other confidence intervals, we have to make some of these simple computations ourselves.

Gujarati (1995), Chapter 5, also discusses tests of hypotheses involving the error variance σ^2. A test statistic can be based on the unbiased estimator $\hat{\sigma}^2$ of σ^2. We can use that $(n-2)\hat{\sigma}^2/\sigma^2$ has a χ^2-distribution with $n-2$ degrees of freedom (see also note 16).

In addition, Gujarati discusses so called *specification tests*. For example, you can test whether the normality assumption is valid. We postpone discussion of such tests to the multiple regression model.

4 The classical multiple linear regression model

4.1 Introduction

So far, we have restricted attention to relations between two random variables. Although this is often useful in practice, we are frequently interested in relating more than two variables.

Example 46. Consider the relation between earnings and schooling. So far, we have investigated this relation in isolation from any other variables that may affect earnings. One such variable is work experience. Work experience can be expected to raise earnings just like schooling. Ideally, if we want to assess the effect of schooling on earnings, we would like to compare earnings levels of individuals with different educational levels, but the same work experience. This is what is called a *ceteris paribus* ("other things being equal") effect of schooling on earnings in economics. In a typical data set, however, work experience and schooling are inversely related: you can accumulate more work experience if you leave school early. So, if we would compare earnings of individuals with high and low levels of schooling, we would be comparing individuals with not just different levels of schooling, but also different levels of work experience. In this particular example, individuals with more education would have higher earnings because they have more education, but (on average) lower earnings because they have less work experience. So, our simple comparison does not provide us with the *ceteris paribus* effect of schooling we are interested in. Instead, we underestimate that effect. So, we need tools to analyze the relation between schooling and earnings holding work experience constant, *i.e.* to disentangle the effects of schooling and work experience on earnings.

As mentioned before, the sample correlation coefficient only deals with relationships between two variables. Fortunately, regression analysis can easily be generalized to relationships between more than two random variables. Here, we will first develop the intuition for the simplest case of two regressors, *i.e.* the three-variable linear regression model. In Subsection 4.8 we show that all results extend to the general (k-variable) multiple linear regression model.

Suppose we want to relate three random variables X_1, X_2 and Y. For example, Y could be earnings, X_1 years of schooling, and X_2 years of work experience. Then, the relation between earnings, schooling and work experience in the population can be modeled by the joint distribution of (X_1, X_2, Y). The 1995 CPS abstract of problem set 2 provides us with data to analyze this relation. The data set is a realization $((x_{11}, x_{21}, y_1), \dots, (x_{1n}, x_{2n}, y_n))$ of a random sample $((X_{11}, X_{21}, Y_1), \dots, (X_{1n}, X_{2n}, Y_n))$ of schooling, work experience and earnings levels for n individuals from the population. Note that X_{1i} (x_{1i}) now denotes observation i of the variable X_1, and X_{2i} (x_{2i}) observation i of X_2.

To summarize the relation between the three variables, we could again approximate the relation in the data by a linear function. This is a bit more involved than in Subsection 3.1, though, as the plot of the data is not a simple cloud of points in $\mathbb{R}^2$. Instead, it now is a cloud of points in $\mathbb{R}^3$. So, instead of a straight line, we have to find a two-dimensional plane that best fits the cloud of data points. As in the previous section, we could predict each y_i by $a + b_1 x_{1i} + b_2 x_{2i}$ for some a, b_1 and b_2. We can now visualize this by plotting the data points in a graph with 3 axes, for X_1, X_2 and Y respectively, and then drawing the two-dimensional plane $a + b_1 x_1 + b_2 x_2$. The predicted values of y_i all lie on this plane, which cuts through the cloud of data points.

Obviously, we again would like to choose a, b_1 and b_2 that minimize, in some way, the distance between our linear model, *i.e.* the two-dimensional plane, and the data. As in the simple regression model, we will use a least squares criterion, and have a, b_1 and b_2 equal OLS estimators $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. This gives "best" predictions $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ of y_i. We return to this later.

Summarizing the cloud of data points this way is useful for various reasons. Most importantly, it allows us to distinguish *partial* effects of regressors on the regressand. Note that the slopes $\hat{\beta}_1$ and $\hat{\beta}_2$ in $\hat{y} = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$ are the partial derivatives of $\hat{y}$ with respect to x_1 and x_2 respectively. So, they have the same interpretation as partial derivatives. For example, $\hat{\beta}_1 = \partial \hat{y} / \partial x_1$ is the change in $\hat{y}$ for a unit change in x_1, *holding x_2 constant.*

Therefore, in the earnings example, the estimate $\hat{\beta}_1$ has a nice interpretation as the effect of schooling on earnings, holding work experience constant. Similarly, the estimate $\hat{\beta}_2$ can be interpreted as the effect of experience on earnings, holding schooling constant. These partial effects are, of course, the *ceteris paribus* effects we were after.

Before we develop the three-variable regression model more formally, we first present a few more examples.

Example 47. Suppose you are a car producer and have to decide upon the design of a new model. Obviously, you would like to know what the demand for the new model will be, *i.e.* what price you could charge and how many cars you would sell at that price. As you are not selling the car yet and it is a new model, you cannot use historical sales data for the particular model. However, you could see the new car as a new bundle of existing car characteristics that are traded in other "bundles" (car models) in the automobile market (unless you are a real innovator and not just a smart assembler). Examples of such characteristics are size, fuel efficiency, engine power (maximum speed), etcetera. If you could fully characterize the models you are considering by listing all its properties, you could use historical market data on existing models to find out how consumers value these characteristics. Perhaps, this will allow you to predict how they would value any of your potential new models, which are, after all, nothing more than new bundles of existing characteristics. Obviously, you will typically not be able to fully characterize a car by a single characteristic, so you will need multiple regression to deal with this problem. Using multiple regression, you would be able to determine how the valuation of a particular model changes if you would change its maximum speed without changing its fuel efficiency or size.

Econometric models like this, in which goods are seen as bundles of characteristics, and demand and prices are determined in terms of these characteristics are called *hedonic* models. These models can be applied in many fields.

Example 48. House prices may depend on the quality of the local public school system, air quality, the proximity to airports and rail tracks and the corresponding noise levels, local crime rates, etcetera. A simple comparison of prices of houses at the end of an airport runway, and houses in a nice and quiet environment wouldn't necessarily be very informative on the value attached to silence. After all, if it is so horrible to live at the end of an airport runway, maybe different types of houses are built there, for example smaller ones. Then, you would actually mix up the effect of noise and size of the house. By estimating multiple regression models for house prices, we may be able to properly distinguish all the effects on house prices and determine how much individuals value clean air, public school quality, noise reduction, and protection against crime. This information can, for example, be used to optimally regulate air pollution and airport noise levels.

Example 49. Hedonic models can also be used to correct price indices for a given class of products for changes in quality. For example, casual inspection of computer store ads suggest that PC prices remain fairly constant over time. However, the performance of PCs, in various dimensions, increases constantly and rapidly. So, in some sense computing power seems to get cheaper and cheaper. One way around this paradox is to view a PC as a bundle of computer characteristics like processor speed, memory size, hard disk speed and size, bus speed, etcetera. From 1999 data, you could estimate a multiple linear regression of PC prices on a, hopefully, exhaustive list of indicators of PC characteristics. Then, instead of viewing PCs as a homogeneous commodity, and comparing price tags of PCs in 2000 and 1999, you could compute what a PC with a 2000 bundle of characteristics would have cost in 1999 by evaluating the year 1999 price equation at the year 2000 characteristics.

4.2 The three-variable linear regression model

The three-variable classical linear regression model is a straightforward extension of the two-variable model. Suppose we have a data set $((x_{11}, x_{21}, y_1), \ldots, (x_{1n}, x_{2n}, y_n))$, which is a realization of a sample $((X_{11}, X_{21}, Y_1), \ldots, (X_{1n}, X_{2n}, Y_n))$ from some population distribution $F_{X_1,X_2,Y}$.

As we have seen in Subsection 2.3.6, a regression is simply a conditional expectation. There is no reason why we couldn't condition on two variables instead of one. So, we can straightforwardly extend Definition 40 to

Definition 41. The *population regression* of Y on X_1 and X_2 is given by $\mathbb{E}(Y|X_1, X_2)$. The *disturbance* or *error term* of the regression is defined by $U = Y - \mathbb{E}(Y|X_1, X_2)$.

The model can again be rewritten as

$$Y = \mathbb{E}(Y|X_1, X_2) + U,$$

in which $\mathbb{E}(Y|X_1, X_2)$ can be interpreted as a "best" prediction of Y given (X_1, X_2), and U as the corresponding prediction error.

As the results from Subsection 3.2 only exploit the properties of conditional expectations, these results directly apply to the three-variable regression. In particular, $\mathbb{E}(U|X_1, X_2) = 0$. By the law of the iterated expectations, this implies that $\mathbb{E}(U|X_1) =$

$\mathbb{E}[\mathbb{E}(U|X_1, X_2)|X_1] = 0$ and $\mathbb{E}(U|X_2) = \mathbb{E}[\mathbb{E}(U|X_1, X_2)|X_2] = 0$, and therefore that $\mathbb{E}(U) = 0$, $\text{cov}(X_1, U) = \mathbb{E}(X_1 U) = 0$ and $\text{cov}(X_2, U) = \mathbb{E}(X_2 U) = 0$.

Again, we will restrict attention to linear regressions

$$\mathbb{E}(Y|X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2,$$

or, equivalently,

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + U \quad \text{and} \quad E(U|X_1, X_2) = 0.$$

In this linear model, we have that $\text{var}(Y) = \text{var}(\beta_1 X_1 + \beta_2 X_2) + \text{var}(U)$ (why can I drop α?). So, the variance of the regressand can again be decomposed in a predicted part related to the regressors and an unpredicted part related to the errors. The covariance term is 0, as the error term and the regressors are uncorrelated. This suggests that we can extend our analysis of the coefficient of determination to the three-variable case. We will indeed do so in Subsection 4.7.

We can also again derive population normal equations directly from the regression assumption. This assumption implies that $\mathbb{E}(U) = 0$, $\mathbb{E}(UX_1) = 0$ and $\mathbb{E}(UX_2) = 0$, which boils down to

$$\begin{aligned} &\mathbb{E}(U) = \mathbb{E}(Y - \alpha - \beta_1 X_1 - \beta_2 X_2) = 0, \\ &\mathbb{E}(X_1 U) = \mathbb{E}[X_1(Y - \alpha - \beta_1 X_1 - \beta_2 X_2)] = 0, \text{ and} \\ &\mathbb{E}(X_2 U) = \mathbb{E}[X_2(Y - \alpha - \beta_1 X_1 - \beta_2 X_2)] = 0. \end{aligned} \tag{19}$$

The first normal equation can be rewritten as

$$\alpha = \mathbb{E}(Y) - \beta_1 \mathbb{E}(X_1) - \beta_2 \mathbb{E}(X_2). \tag{20}$$

Substituting this in the second and third equations, we find that

$$\begin{aligned} \mathbb{E}(X_1 U) &= \mathbb{E}[X_1(Y - \mathbb{E}(Y) - \beta_1(X_1 - \mathbb{E}(X_1)) - \beta_2(X_2 - \mathbb{E}(X_2))] \\ &= \mathbb{E}[(X_1 - \mathbb{E}(X_1))(Y - \mathbb{E}(Y) - \beta_1(X_1 - \mathbb{E}(X_1)) - \beta_2(X_2 - \mathbb{E}(X_2))] \\ &= \text{cov}(X_1, Y) - \beta_1 \,\text{var}(X_1) - \beta_2 \,\text{cov}(X_1, X_2) = 0, \\ &\text{and} \\ \mathbb{E}(X_2 U) &= \mathbb{E}[X_2(Y - \mathbb{E}(Y) - \beta_1(X_1 - \mathbb{E}(X_1)) - \beta_2(X_2 - \mathbb{E}(X_2))] \\ &= \mathbb{E}[(X_2 - \mathbb{E}(X_2))(Y - \mathbb{E}(Y) - \beta_1(X_1 - \mathbb{E}(X_1)) - \beta_2(X_2 - \mathbb{E}(X_2))] \\ &= \text{cov}(X_2, Y) - \beta_1 \,\text{cov}(X_1, X_2) - \beta_2 \,\text{var}(X_2) = 0. \end{aligned}$$

This can be reorganized into

$$\begin{aligned}
\beta_1 &= \frac{\operatorname{var}(X_2)\operatorname{cov}(X_1,Y) - \operatorname{cov}(X_1,X_2)\operatorname{cov}(X_2,Y)}{\operatorname{var}(X_1)\operatorname{var}(X_2) - \operatorname{cov}(X_1,X_2)^2} \\
&= \frac{\operatorname{var}(X_2)\operatorname{cov}(X_1,Y) - \operatorname{cov}(X_1,X_2)\operatorname{cov}(X_2,Y)}{\operatorname{var}(X_1)\operatorname{var}(X_2)\left[1-\rho(X_1,X_2)^2\right]} \\
&\text{and} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (21) \\
\beta_2 &= \frac{\operatorname{var}(X_1)\operatorname{cov}(X_2,Y) - \operatorname{cov}(X_1,X_2)\operatorname{cov}(X_1,Y)}{\operatorname{var}(X_1)\operatorname{var}(X_2) - \operatorname{cov}(X_1,X_2)^2} \\
&= \frac{\operatorname{var}(X_1)\operatorname{cov}(X_2,Y) - \operatorname{cov}(X_1,X_2)\operatorname{cov}(X_1,Y)}{\operatorname{var}(X_1)\operatorname{var}(X_2)\left[1-\rho(X_1,X_2)^2\right]},
\end{aligned}$$

provided that X_1 and X_2 are not degenerate and not perfectly correlated, so that $\operatorname{var}(X_1) > 0$, $\operatorname{var}(X_2) > 0$ and $|\rho(X_1,X_2)| < 1$, and the denominators are strictly positive.

This looks all quite prohibitive compared to the simple linear regression model. It may be of some comfort to know that it is much more convenient to use matrix algebra to develop the multiple regression model. In Subsection 4.8, we will see, that, in matrix notation, the results for the k-variable linear regression model closely resemble those for the simple model. Of course, on the downside, it requires some knowledge of matrix algebra to actually see that.

Even though equations (20) and (21) are more difficult to read than their simple regression counterparts, it is not so hard to develop some intuition. First, note that β_1 and β_2 are the partial derivatives of $\mathbb{E}(Y|X_1,X_2)$ with respect to X_1 and X_2, respectively. So, for example, β_1 can again be interpreted as the effect on $\mathbb{E}(Y|X_1,X_2)$ of a unit change in X_1, holding X_2 constant (*ceteris paribus*).

Now, suppose that X_1 and X_2 are uncorrelated, so that $\operatorname{cov}(X_1,X_2) = 0$. Then, (21) reduces to

$$\beta_1 = \frac{\operatorname{cov}(X_1,Y)}{\operatorname{var}(X_1)} \quad \text{and} \quad \beta_2 = \frac{\operatorname{cov}(X_2,Y)}{\operatorname{var}(X_2)},$$

the regression parameters for simple regressions of Y on respectively X_1 and X_2. This makes sense. If X_1 and X_2 are uncorrelated, there is no risk of confusing a linear relation between Y and one of the regressors with a relation with the other regressor.

In general, the partial effects are not the same as the overall effects. In other words, the parameters in equations (20) and (21) are generally not the parameters of simple regressions of Y on X_1 and X_2. We will provide more intuition in Subsection 4.5.

We end this subsection by pointing out that the normal equations have more in common than the discussion of the simple linear regression model may have suggested. First note that the second and third normal equation in (19) have the same form, only with X_1 and X_2 interchanged. Now, we could see the constant as just another regressor, say X_0, that is always one. With this notation, the linear regression equation can be rewritten as $Y_i = \alpha X_0 + \beta_1 X_1 + \beta_2 X_2$, and α would be a "slope" parameter for the constant X_0. The first normal equation above is just $\mathbb{E}(X_0 U) = \mathbb{E}(1U) = 0$, and is not fundamentally different from the second and third equations. We will exploit this thought further if we present the k-variable model in matrix notation in Subsection 4.8.

4.3 The classical assumptions revisited: multicollinearity

We can again apply the population regression model directly to the sample $((X_{11}, X_{21}, Y_1), \ldots, (X_{1n}, X_{2n}, Y_n))$ by assuming that each triple (X_{1i}, X_{2i}, Y_i) satisfies the model. To this end, collect all regressors in the random sample into a $(n \times 2)$-matrix $\mathbf{X}$ with i-th row equal to (X_{1i}, X_{2i}). Assumption 1 can be extended straightforwardly to

Assumption 1*. (linear regression) $\mathbb{E}[Y_i|\mathbf{X}] = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i}$.

Assumption 2 needs no change, but we repeat it for completeness.

Assumption 2*. (spherical errors) The errors are *homoskedastic*: $\text{var}(U_i|\mathbf{X}) = \sigma^2$, for some $\sigma > 0$, for all $i = 1, \ldots, n$. Furthermore, they are *uncorrelated*: $\text{cov}(U_i, U_j|\mathbf{X}) = 0$ for all $i, j = 1, \ldots, n$ such that $i \neq j$.

We have to extend Assumption 3 to

Assumption 3*. (sufficient variation) $n > 3$. Furthermore, the constant and the regressors x_{1i} and x_{2i} are *not perfectly multicollinear.*

We discuss Assumption 3* below. Finally, Assumption 4 can be extended directly into

Assumption 4*. (deterministic regressors) $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ are deterministic, *i.e.* fixed to the values $(x_{11}, x_{21}), \ldots, (x_{1n}, x_{2n})$ in repeated sampling.

Only Assumption 3* has substantially new content due to the introduction of a second regressor. First, it now requires that $n > 3$. In the simple regression model we only needed $n > 2$, but we now need an extra observation as we have to estimate one extra parameter. Again, it is intuitively clear that we need at least 3 data points to pin down a two-dimensional plane in $\mathbb{R}^3$. As will be shown later, we need one more data point to be able to estimate the variance of the error term.

More importantly, it now uses a new word, "multicollinearity", to describe the required variation in the regressors.

Definition 42. The constant and the regressors x_{1i} and x_{2i} are said to be *perfectly multicollinear* if there exist real numbers c_0, c_1 and c_2, with at least one of these numbers nonzero, such that

$$c_0 + c_1 x_{1i} + c_2 x_{2i} = 0 \text{ for all } i. \tag{22}$$

Assumption 3* excludes such perfect multicollinearity. This is best understood by going through some examples.

First, suppose that the sample variance of X_1 is zero, *i.e.* $s_{X_1} = 0$. Recall that this excluded by the similar Assumption 3 for the simple regression model. If $s_{X_1} = 0$, then x_{1i} is constant and equal to its sample mean. So, $x_{1i} = \bar{x}_1$ for all i. In this case, equation (22) is satisfied for $c_0 = -\bar{x}_1$, $c_1 = 1$ and $c_2 = 0$, and we have perfect multicollinearity. So, Assumption 3* excludes that $s_{X_1} = 0$ (or $s_{X_2} = 0$) just like Assumption 3. Another perspective on this is that the second part of Assumption 3 for the simple regression model can be rephrased as excluding perfect multicollinearity of the constant and the regressor.

Second, suppose that $x_{1i} = x_{2i}$ for all i. Then, equation (22) is satisfied for $c_0 = 0$, $c_1 = 1$ and $c_2 = -1$. More in general, we have perfect multicollinearity of the constant and the regressors x_{1i} and x_{2i} (*i.e.*, equation (22) holds for some c_0, c_1 and c_2 not all zero) if one regressor is a linear function of the other regressor, possibly including a constant term.

If we have perfect multicollinearity, we cannot distinguish the partial effects of the regressors. To see this, suppose that $x_{1i} = a + bx_{2i}$ for some real numbers a and b. Then, we have perfect multicollinearity, with $c_0 = -a$, $c_1 = 1$ and $c_2 = -b$ (check!). Also, we can rewrite the regression model as

$$\begin{aligned} Y_i &= \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + U_i \\ &= \alpha + \beta_1 (a + b x_{2i}) + \beta_2 x_{2i} + U_i \\ &= \alpha + \beta_1 a + (\beta_1 b + \beta_2) x_{2i} + U_i \\ &= \alpha^* + \beta_1^* x_{1i} + \beta_2^* x_{2i} + U_i, \end{aligned}$$

with $\alpha^* = \alpha + \beta_1 a$, $\beta_1^* = 0$ and $\beta_2^* = \beta_1 b + \beta_2$. This gives two equivalent characterizations of the same linear regression. Similarly, due to the multicollinearity of the regressors, we can rewrite the linear regression equation in many other ways, reallocating the slope on one regressor to the other. So, there is no way to discern the separate, partial relations between Y and respectively X_1 and X_2. This makes sense, as there is no independent variation in the regressors in the sample. For this reason we have to exclude perfect multicollinearity of the regressors.

The perfect multicollinearity problem relates to variation in the sample. Note however that we have seen a related problem in the population model in the previous section. The population slope parameters in equation (21) are not determined if the denominators in equation (21) are 0. This happens if either $\text{var}(X_1) = 0$, $\text{var}(X_2) = 0$, or $|\rho(X_1, X_2)| = 1$. In a way, these are the population equivalents to the examples above: a constant x_{1i}, a constant x_{2i}, and perfectly linearly related x_{1i} and x_{2i}.

Finally, note again that α can just be seen as another "slope" coefficient on a very specific regressor that equals 1 for all observations. In other words, we could slightly rewrite the regression as

$$Y_i = \alpha x_{0i} + \beta_1 x_{1i} + \beta_{2i} x_{2i} + U_i,$$

where $x_{0i} = 1$ for all i. In this notation, equation (22) can be written as $c_0 x_{0i} + c_1 x_{1i} + c_2 x_{2i} = 0$ for all i. Instead of speaking of "perfect multicollinearity of the constant and the regressors", we could then simply say "perfect multicollinearity of the regressors". Again, this insight will prove useful if we discuss the k-variable model in matrix notation later.

4.4 Least squares estimation

4.4.1 The OLS estimators

The OLS estimators $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ equal the a, b_1 and b_2, respectively, that minimize the sum of squared residuals

$$\sum_{i=1}^{n} \left(Y_i - a - b_1 x_{1i} - b_2 x_{2i}\right)^2 . \tag{23}$$

The first order conditions for this minimization problem are again found by setting the derivatives of (23) with respect to a, b_1 and b_2 to 0, and evaluating at $a = \hat{\alpha}$, $b_1 = \hat{\beta}_1$ and $b_2 = \hat{\beta}_2$. As we are now minimizing with respect to three variables, this gives three normal equations (note that we can cancel the "-2" without changing the equations),

$$\begin{aligned}
\sum_{i=1}^{n} \hat{U}_i &= \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right) = 0,\\
\sum_{i=1}^{n} x_{1i}\hat{U}_i &= \sum_{i=1}^{n} x_{1i}\left(Y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right) = 0, \text{ and}\\
\sum_{i=1}^{n} x_{2i}\hat{U}_i &= \sum_{i=1}^{n} x_{2i}\left(Y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right) = 0.
\end{aligned} \tag{24}$$

Here, $\hat{U}_i = Y_i - \hat{Y}_i$ is the OLS residual, where $\hat{Y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i}$ is again the predicted value of Y_i.

The normal equations (24) imply that

$$\hat{\alpha} = \bar{Y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2,$$

which is a straightforward extension of the corresponding equation for the simple regression model. Note that again $\hat{\alpha} = \bar{Y}$ if we take both regressors in deviation from their sample means ($\bar{x}_1 = 0$ and $\bar{x}_2 = 0$).

Substituting $\hat{\alpha}$ into the remaining two normal equations, we get

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_{1i} \left(Y_i - \bar{Y} - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)\right) \\ &= \sum_{i=1}^{n} (x_{1i} - \bar{x}_1) \left(Y_i - \bar{Y} - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)\right) \end{aligned}$$

and

$$\begin{aligned} 0 &= \sum_{i=1}^{n} x_{2i} \left(Y_i - \bar{Y} - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)\right) \\ &= \sum_{i=1}^{n} (x_{2i} - \bar{x}_2) \left(Y_i - \bar{Y} - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2)\right). \end{aligned}$$

Using the notation for sample (co-)variances introduced earlier, this is more concisely written as

$$S_{X_1Y} - \hat{\beta}_1 s^2_{X_1} - \hat{\beta}_2 s_{X_1X_2} = 0 \;\text{ and }\; S_{X_2Y} - \hat{\beta}_1 s_{X_1X_2} - \hat{\beta}_2 s^2_{X_2} = 0.$$

It takes a few steps to rewrite this into explicit expressions for $\hat{\beta}_1$ and $\hat{\beta}_2$,

$$\begin{aligned} \hat{\beta}_1 &= \frac{s^2_{X_2} S_{X_1Y} - s_{X_1X_2} S_{X_2Y}}{s^2_{X_1} s^2_{X_2} - s^2_{X_1X_2}} = \frac{s^2_{X_2} S_{X_1Y} - s_{X_1X_2} S_{X_2Y}}{s^2_{X_1} s^2_{X_2} \left[1 - \hat{\rho}^2_{X_1X_2}\right]} \;\text{ and} \\ \hat{\beta}_2 &= \frac{s^2_{X_1} S_{X_2Y} - s_{X_1X_2} S_{X_1Y}}{s^2_{X_1} s^2_{X_2} - s^2_{X_1X_2}} = \frac{s^2_{X_1} S_{X_2Y} - s_{X_1X_2} S_{X_1Y}}{s^2_{X_1} s^2_{X_2} \left[1 - \hat{\rho}^2_{X_1X_2}\right]}. \end{aligned} \tag{25}$$

These look again a bit messy. However, note that $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are indeed again the sample equivalents to α, β_1 and β_2 in (20) and (21). Perfect multicollinearity would render the denominators in (25) 0, and leave $\hat{\beta}_1$ and $\hat{\beta}_2$, and in general also $\hat{\alpha}$, undetermined. Note that this is a straightforward extension of the simple regression case, in which the estimators are undetermined if $s^2_X = 0$ and there is no variation in the regressor.

4.4.2 Properties of the OLS estimators

The properties derived for OLS estimators in the simple linear regression model of Section 3 also hold for the three-variable model. We will not prove these properties here. They are more conventiently derived in the more general k-variable model of Subsection 4.8 using matrix notation. So, we only list the properties again.

First note that the OLS estimators are again linear functions of the random variables $Y_1, \ldots, Y_n$. After all, $\hat{\alpha}$ is a linear function of $\bar{Y}$, and therefore $Y_1, \ldots, Y_n$, $\hat{\beta}_1$ and $\hat{\beta}_2$. In turn, the denominators in (25) do not depend on the Y_i. In the enumerator, the Y_i enter linearly through the covariance terms, and so $\hat{\beta}_1$ and $\hat{\beta}_2$ are linear in $Y_1, \dots, Y_n$ as well.

The main result is again the *Gauss-Markov theorem.*

Proposition 4. *Under the classical Assumptions 1*–4*, the OLS estimators $\hat{\alpha}$, $\hat{\beta}_1$ and $\hat{\beta}_2$ are the* best linear unbiased estimators *(BLUE).*

It is possible to derive the variances and covariances of the estimators, but this is rather tedious. With the use of matrix algebra it is much easier, so we postpone a full derivation to the k-variable model. Here, we just give the variances and covariances for the special case in which the regressors are taken in deviation from their sample means, so that $\bar{x}_1 = 0$ and $\bar{x}_2 = 0$. In this case, again $\hat{\alpha} = \bar{Y}$. So,

$$\operatorname{var}(\hat{\alpha}) = \frac{\sigma^2}{n}, \quad \operatorname{cov}(\hat{\alpha}, \hat{\beta}_1) = 0 \quad \text{and} \quad \operatorname{cov}(\hat{\alpha}, \hat{\beta}_2) = 0.$$

Furthermore, it can (and will) be shown that

$$\begin{aligned}
\operatorname{var}(\hat{\beta}_1) &= \frac{\sigma^2}{(n-1)s^2_{X_1}\left[1 - \hat{\rho}^2_{X_1X_2}\right]}, \\
\operatorname{var}(\hat{\beta}_2) &= \frac{\sigma^2}{(n-1)s^2_{X_2}\left[1 - \hat{\rho}^2_{X_1X_2}\right]}, \\
&\text{and} \\
\operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) &= \frac{-\sigma^2 s_{X_1X_2}}{(n-1)s^2_{X_1}s^2_{X_2}\left[1 - \hat{\rho}^2_{X_1X_2}\right]}.
\end{aligned}$$

Note that the latter set of slope variances and covariances would have been the same if we had not taken the regressors in deviation from their sample means. After all, this does not affect the slope estimators, only the intercept estimator.

The variances of $\hat{\beta}_1$ and $\hat{\beta}_2$ are larger if the error variance is larger relative to the regressor variances. In a sense, this corresponds to more "noise" relative to "useful" regressor variation in the data. Also, the variances are again smaller if the sample size increases. The only difference with the simple regression model is that the variances now depend (inversely) on the squared sample correlation of the regressors. The intuition for this result is that it is harder to unravel the partial effects of X_1 and X_2 if there is less independent variation in X_1 and X_2. If the regressors are uncorrelated, $s_{X_1X_2} = 0$ and $\hat{\beta}_1$ and $\hat{\beta}_2$ are uncorrelated. In this special case, $\hat{\rho}^2_{X_1X_2} = 0$ and $\operatorname{var}(\hat{\beta}_1)$ and $\operatorname{var}(\hat{\beta}_2)$ reduce to their simple regression counterparts.

The variances and covariances depend on the variance of the error σ^2. An unbiased estimator of σ^2 now is

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{U}_i^2}{n-3}.$$

Now, we have to divide by $n-3$ as we have 3 parameters in the regression equation.

Also, under some regularity conditions, the OLS estimators are consistent and asymptotically normal. Finally, if we assume normality of the error term from the outset, they are (jointly) normal anyhow. Also, the OLS estimators are the best unbiased estimators, and not just BLUE, under the normality assumption.

4.5 Omitted variable bias

In the introduction to this section, we argued that omitting variables could lead to a bias. We now have a closer look at that.

Suppose that $\mathbb{E}(Y|X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2$ with $\beta_2 \neq 0$, so that X_2 is a relevant variable. Now, suppose we omit X_2, for example because we cannot measure it or because we do not know it should be included, and we estimate β_1 by $\tilde{\beta}_1 = S_{X_1 Y}/s^2_{X_1}$. This is the simple OLS estimator for a regression of Y on X_1 only. This is generally not an unbiased estimator of β_1. After all,

$$\begin{aligned}
\tilde{\beta}_1 &= \frac{S_{X_1 Y}}{s^2_{X_1}} \\
&= \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(Y_i - \bar{Y})}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)\left[\beta_1 (x_{1i} - \bar{x}_1) + \beta_2 (x_{2i} - \bar{x}_2) + U_i - \bar{U}\right])}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} + \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) U_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} - \frac{\bar{U} \sum_{i=1}^n (x_{1i} - \bar{x}_1)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \\
&= \beta_1 + \beta_2 \frac{s_{X_1 X_2}}{s^2_{X_1}} + \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1) U_i}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2},
\end{aligned}$$

so that

$$\mathbb{E}(\tilde{\beta}_1) = \beta_1 + \beta_2 \frac{s_{X_1 X_2}}{s^2_{X_1}} + \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)\mathbb{E}(U_i)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} = \beta_1 + \beta_2 \frac{s_{X_1 X_2}}{s^2_{X_1}}. \tag{26}$$

The second, bias, term only disappears if X_1 and X_2 are not correlated in the sample. The intuition of equation (26) is clear. The bias term picks up any effect of the omitted regressor that can be captured by the included regressor X_1. The included regressor X_1 can only compensate for some of the omitted regressor X_2 if it is correlated with X_2.

For example, suppose that X_1 and X_2 are positively correlated in the sample. Also, let the partial effect of X_2 on Y be positive, *i.e.* $\beta_2 > 0$. Then, omitting X_2 from the regression attributes some of the positive relation of X_2 and Y to X_1, which leads to an upward bias in $\tilde{\beta}_1$. $\tilde{\beta}_1$ now not only captures the "true" partial regression effect β_1 of X_1, but also part of the effect of X_2.

Example 50. Suppose we want to know how PC price depends on computer speed, as measured by the number of calculations per second. Suppose we want to compare a wide range of speeds, and use data for 1986 and 2000 PCs. Just pooling the 1986 and 2000 data and running a simple regression of PC price on computer speed may be deceptive. In either 1986 and 2000, faster PCs are more expensive. However, speed was much lower in 1986, but in general PC prices were much higher at any given computer speed. So, from a pooled regression we may well find that price and speed are negatively related, but, at least intuitively, this is wrong. The problem is that we are not controlling for other differences between 1986 and 2000 PCs. We could run the same regression, but including a (dummy) variable that is 1 if it is a 2000 PC, and 0 if it is a 1986 PC. This dummy regressor will pick up the price difference between 1986 and 2000 PCs due to other reasons than computer speed differences. In a sense, we allow for regression lines with different intercepts for 1986 and 2000 (draw graph). The slopes are however the same, and correspond to the (partial) marginal effect of speed on price, which is assumed to be the same in 1986 and 2000. Estimating this three-variable regression model would give you a positive estimate of the effect of speed on price.

In terms of the analysis in this section, the omitted regressor, the 2000 dummy, is negatively related to the regressand, price, and positively related to the remaining regressor, computer speed. This leads to a downward bias in the estimate of the processor speed coefficient.

4.6 Estimation with irrelevant variables

If omitting variables may lead to biases, we may be tempted to always include as many variables as possible. However, there is a downside to this strategy. Including irrelevant variables generally leads to an efficiency loss.

To see this, suppose that the parameter $\beta_2 = 0$, so that X_2 is an irrelevant variable in the regression. We could now estimate β_1 from a simple regression of Y on X_1, which gives the OLS estimator

$$\tilde{\beta}_1 = \frac{S_{X_1Y}}{s^2_{X_1}},$$

If X_2 is irrelevant, this is an unbiased estimator of β_1 with variance

$$\text{var}(\tilde{\beta}_1) = \frac{\sigma^2}{(n-1)s^2_{X_1}}.$$

Of course, the OLS estimator $\hat{\beta}_1$ of β_1 for the three-variable regression is also unbiased, but it has variance

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{(n-1)s^2_{X_1}\left[1-\hat{\rho}^2_{X_1X_2}\right]} \geq \frac{\sigma^2}{(n-1)s^2_{X_1}} = \text{var}(\tilde{\beta}_1).$$

Only if $\hat{\rho}_{X_1X_2} = 0$, the two variances are the same. Otherwise, the variance of the simple regression estimator $\tilde{\beta}_1$ is strictly smaller than that of $\hat{\beta}_1$.

4.7 The coefficient of determination

In Subsection 4.2, we have already seen that we can decompose the population variance of the regressand in predicted and residual components, or $\text{var}(Y) = \text{var}(\beta_1X_1 + \beta_2X_2) + \text{var}(U)$, as in the simple regression model. This follows directly from the population normal equations (19). In particular, the fact that X_1 and X_2 on the one hand and U on the other hand are uncorrelated ensures that the covariance term is 0.

The normal equations (24) are the sample counterpart of (19). They imply that the sample counterpart of the variance decomposition,

$$TSS = ESS + RSS,$$

holds, where again

$$\begin{aligned} TSS &= \sum_{i=1}^n (Y_i - \bar{Y})^2, \\ ESS &= \sum_{i=1}^n (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2, \text{ and} \\ RSS &= \sum_{i=1}^n (\hat{U}_i - \bar{\hat{U}})^2 = \sum_{i=1}^n \hat{U}_i^2. \end{aligned}$$

We will derive this result for the general k-variable case in Subsection 4.8.4. For now, just note that the derivation is analogous to that for the simple regression model in Subsection 3.5.

The definition of the coefficient of determination needs no change. It is $R^2 = ESS/TSS$ and satisfies $0 \leq R^2 \leq 1$. Also, the multiple correlation coefficient $|R|$ is still the squared sample correlation between Y_i and $\hat{Y}_i$, which is now a true generalization of the (simple) sample correlation coefficient.

We end this subsection by relating the R^2 of a three-variable regression to the R^2 of a simple regression on only one of the regressors. To be concrete, suppose we regress Y on X_1 only. The OLS estimators of this simple regression, say $\tilde{\alpha}$ and $\tilde{\beta}_1$, are the values of a and b_1 that minimize the corresponding sum of squared residuals,

$$\sum_{i=1}^n (Y_i - a - b_1 x_{1i})^2,$$

which gives a (minimal) residual sum of squares

$$\widetilde{RSS} = \sum_{i=1}^n \left(Y_i - \tilde{\alpha} - \tilde{\beta}_1 x_{1i}\right)^2.$$

The OLS estimators $\hat{\alpha}$, $\hat{\beta}_1$, and $\hat{\beta}_2$ of a regression of Y on both X_1 and X_2 on the other hand equal the values c, d_1 and d_2 that minimize

$$\sum_{i=1}^n (Y_i - c - d_1 x_{1i} - d_2 x_{2i})^2, \tag{27}$$

which gives a (minimal) residual sum of squares

$$RSS = \sum_{i=1}^n \left(Y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i}\right)^2,$$

Now, note that equation (27) would be $\sum_{i=1}^{n}(Y_i - c - d_1 x_{1i} - d_2 x_{2i})^2 = \widetilde{RSS}$ if we set $c = \tilde{\alpha}$, $d_1 = \tilde{\beta}_1$ and $d_2 = 0$. So, one possible choice of c, d_1 and d_2 in (27) gives the same sum of squared residuals as the minimum sum of squared residuals $\widetilde{RSS}$ in the simple regression model. This implies that

$$RSS \leq \widetilde{RSS}.$$

As the total sum of squares TSS is the same in both regressions, as they have the same regressand Y, this implies that

$$R^2 = 1 - \frac{RSS}{TSS} \geq 1 - \frac{\widetilde{RSS}}{TSS} = \tilde{R}^2.$$

4.8 The k-variable multiple linear regression model

4.8.1 The population regression

The multiple linear regression model is most easily presented in matrix notation. Using matrix notation, we can allow for an arbitrary number of regressors. Obviuously, this includes the simple and the three-variable regression models discussed so far.

If all is well, you have refreshed your matrix algebra in the last TA session. Also, you should have read Gujarati (1995), Appendix B, by now.

So, suppose that we have $k-1$ regressors $X_1, \ldots, X_{k-1}$. As discussed before, we can treat the constant as a k-th regressor "X_0" that is always 1. So, stack the constant and the $k-1$ regressors in a $(1 \times k)$-vector

$$X = (1 \; X_1 \; X_2 \cdots X_{k-1}).$$

Be aware that we now use X to denote the constant and all $k-1$ regressors, whereas we used it to denote the single regressor in the simple regression model.

The linear regression of Y on $X_1, \ldots, X_{k-1}$ is

$$\mathbb{E}(Y|X) = \alpha + \beta_1 X_1 + \cdots + \beta_{k-1} X_{k-1} = (1 \; X_1 \cdots X_{k-1}) \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_{k-1} \end{pmatrix} = X\beta,$$

where $\beta = (\alpha\ \beta_1 \cdots \beta_{k-1})'$ is a $(k \times 1)$-vector stacking all the regression parameters. We can again write

$$Y = X\beta + U, \text{ with } \mathbb{E}(U|X) = 0.$$

$\mathbb{E}(U|X) = 0$ implies that $\mathbb{E}(X'U) = 0$, which gives the vector normal equation

$$\mathbb{E}[X'(Y - X\beta)] = \mathbb{E}\begin{bmatrix} 1(Y - X\beta) \\ X_1(Y - X\beta) \\ \vdots \\ X_{k-1}(Y - X\beta) \end{bmatrix} = \begin{pmatrix} \mathbb{E}[1(Y - X\beta)] \\ \mathbb{E}[X_1(Y - X\beta)] \\ \vdots \\ \mathbb{E}[X_{k-1}(Y - X\beta)] \end{pmatrix} = 0. \qquad (28)$$

Note that $X'U$, and therefore $\mathbb{E}(X'U)$, is a $(k \times 1)$ vector. So here, "0" is a $(k \times 1)$-vector of zeros. We use "0" interchangeably to denote the real number 0 and a real vector of zeros.

We can rewrite the normal equation (28) as $\mathbb{E}(X'Y) - \mathbb{E}(X'X\beta) = 0$, or $\mathbb{E}(X'X)\beta = \mathbb{E}(X'Y)$. Provided that $\mathbb{E}(X'X)$ is invertible, we can premultiply this equation by $\mathbb{E}(X'X)^{-1}$ to get

$$\beta = \mathbb{E}(X'X)^{-1}\mathbb{E}(X'Y).$$

Here, we have used that $\mathbb{E}(X'X)^{-1}\mathbb{E}(X'X)\beta = I_k\beta = \beta$, with I_k a $(k \times k)$-matrix with ones on the diagonal and zeros elsewhere (an identity matrix).

4.8.2 The classical assumptions

Suppose we have a sample of n observations of (X, Y). Stack the regressors, including constants, in a $(n \times k)$-matrix $\mathbf{X}$, and stack the regressand in a $(n \times 1)$-vector $\mathbf{Y}$. So, we have

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{(k-1)1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1n} & \cdots & X_{(k-1)n} \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}.$$

Here, X_{ij} is observation j of regressor i. Each row of $\mathbf{X}$ and $\mathbf{Y}$ corresponds to an observation in the sample, and each column of $\mathbf{X}$ to a regressor variable. The first column of $\mathbf{X}$ is reserved for the constant. We also write $\mathbf{x}$ for the corresponding realization of $\mathbf{X}$ (*i.e.*, the matrix of the regressors in an actual data set on your PC).

We can now give the classical assumptions in matrix notation.

Assumption 1†. (linear regression) $\mathbb{E}[\mathbf{Y}|\mathbf{X}] = \mathbf{X}\beta$.

We can again write

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{U} \quad \text{and} \quad \mathbb{E}(U|\mathbf{X}) = 0,$$

with $\mathbf{U} = (U_1 \cdots U_n)'$ the $(k \times 1)$-vector of regression errors. Alternatively, if we denote the i-th row of $\mathbf{X}$ by $\mathbf{X}_i$, we can say that $\mathbb{E}[Y_i|\mathbf{X}] = \mathbf{X}_i\beta$, or that $Y_i = \mathbf{X}_i\beta + U_i$ and $\mathbb{E}(U_i|\mathbf{X}) = 0$, for all i.

Assumption 2†. (spherical errors) The errors are spherical, *i.e.* $\mathbb{E}(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \sigma^2 I_n$ for some $\sigma > 0$.

This just says, in very compact matrix notation, that the errors should be homoskedastic and uncorrelated. This can be seen by expanding the matrix notation a bit, which gives

$$\begin{aligned}
\mathbb{E}[\mathbf{U}\mathbf{U}'|\mathbf{X}] &= \mathbb{E}\left[\left.\begin{pmatrix} U_1U_1 & \cdots & U_1U_n \\ \vdots & & \vdots \\ U_nU_1 & \cdots & U_nU_n \end{pmatrix}\right| \mathbf{X}\right] \\
&= \begin{bmatrix} \operatorname{cov}(U_1,U_1|\mathbf{X}) & \cdots & \operatorname{cov}(U_1,U_n|\mathbf{X}) \\ \vdots & & \vdots \\ \operatorname{cov}(U_n,U_1|\mathbf{X}) & \cdots & \operatorname{cov}(U_n,U_n|\mathbf{X}) \end{bmatrix} \\
&= \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma^2 & 0 \\ 0 & \cdots & 0 & 0 & \sigma^2 \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix} = \sigma^2 I_n.
\end{aligned}$$

Here we use that $\mathbb{E}(U|\mathbf{X}) = 0$, which ensures that $\mathbb{E}[UU'|\mathbf{X}]$ is a $(n \times n)$-matrix of (conditional) covariances between U_i and U_j. Note that this gives variances on the diagonal, as $\operatorname{cov}(U_i, U_i|\mathbf{X}) = \operatorname{var}(U_i|\mathbf{X})$. For this reason, this matrix is also called the *variance-covariance matrix* of $\mathbf{U}$ (conditional on $\mathbf{X}$).

Assumption 3†. (sufficient variation) $n > k$ and no perfect multicollinearity of the regressors.

The intuition for $n > k$ is the same as before. We have now fully replaced the conditions for sufficient variation of the regressors by a general condition excluding perfect multicollinearity. This condition, which in matrix algebra notation is written as $\text{rank}(\mathbf{x}) = k$, is satisfied if there does *not* exist a nonzero $(k \times 1)$-vector c such that $\mathbf{x}c = 0$. Sometimes, this is called *linear independence* of the k columns of $\mathbf{x}$.

The final assumption boils down to

Assumption 4†. (deterministic regressors) $\mathbf{X}$ is deterministic, *i.e.* fixed to the value $\mathbf{x}$ in repeated sampling.

Because of Assumption 4†, conditioning on $\mathbf{X}$ is again irrelevant. Therefore, we will not condition on the regressors in the following analysis of the OLS estimator. The reason to explicitly condition on $\mathbf{X}$ in Assumptions 1† and 2†, even though we make Assumption 4†, is that we will relax Assumption 4† in Subsection 5.1.

4.8.3 Least squares estimation

The OLS estimator $\hat{\beta}$ of β is the vector b that minimizes the sum of squared residuals

$$(\mathbf{Y} - \mathbf{x}b)' (\mathbf{Y} - \mathbf{x}b) = \sum_{i=1}^{n} (Y_i - \mathbf{x}_i b)^2 .$$

Here, $\mathbf{x}_i$ is the i-th row of $\mathbf{x}$.

Finding the minimum now requires taking derivatives with respect to the vector b, and equating these derivatives to 0. Doing so, we find that $\hat{\beta}$ should satisfy the normal equations (first order conditions)

$$\mathbf{x}'\hat{\mathbf{U}} = \mathbf{x}' \left(\mathbf{Y} - \mathbf{x}\hat{\beta} \right) = 0,$$

where $\hat{\mathbf{U}} = \mathbf{Y} - \mathbf{x}\hat{\beta}$ is the $(n \times 1)$-vector of OLS residuals. Solving for $\hat{\beta}$ gives

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{Y} = (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}' (\mathbf{X}\beta + \mathbf{U}) = \beta + (\mathbf{x}'\mathbf{x})^{-1} \mathbf{x}'\mathbf{U}.$$

The inverse $(\mathbf{x}'\mathbf{x})^{-1}$ exists because there is no perfect multicollinearity.

It is now easy to derive some properties. The most important of these is that the Gauss-Markov theorem again holds: the OLS estimator $\hat{\beta}$ is BLUE. We repeat this important result in

Proposition 5. *Under the classical Assumptions $1^\dagger$–$4^\dagger$, the OLS estimator $\hat\beta$ is the* best linear unbiased estimator *(BLUE).*

First note that $\hat\beta$ is again linear, *i.e.* it is a linear function of the random variables Y_1, … ,Y_n. Also, using that $(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$ is non-random and that $\mathbb{E}(\mathbf{U}) = 0$, we have that

$$\mathbb{E}(\hat\beta) = \mathbb{E}\left[\beta + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U}\right] = \beta + \mathbb{E}\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U}\right] = \beta + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbb{E}(\mathbf{U}) = \beta,$$

so that $\hat\beta$ is unbiased.

The Gauss-Markov theorem further tells us that $\hat\beta$ is efficient relative to all linear unbiased estimators. We should shortly discuss the meaning of the word "efficient" here. Definition 35 only defines efficiency for the case in which we are estimating a single parameter, but we are now estimating a vector of k parameters. Of course, we could have equally well have wondered about this when discussing the Gauss-Markov theorem for the simple and the three-variable regression models. However, efficiency of an estimator of a parameter vector is more easily defined in matrix notation. In general, we have

Definition 43. Let $\theta \in \mathbb{R}^k$ be a parameter, and $\hat\theta$ and $\tilde\theta$ two unbiased estimators of θ. Then, $\hat\theta$ is called *efficient* relative to $\tilde\theta$ if $\text{var}(c'\hat\theta) \leq \text{var}(c'\tilde\theta)$ for all $c \in \mathbb{R}^k$.

So, the Gauss-Markov theorem states that $\text{var}(c'\tilde\beta) \geq \text{var}(c'\hat\beta)$ for all $c \in \mathbb{R}^k$ if $\tilde\beta$ is an other linear unbiased estimator of β. So any linear combination of the elements of any other linear unbiased estimator $\tilde\beta$ has a variance at least as high as the same linear combination of the elements of the OLS estimator $\hat\beta$. In particular, we can choose c to be any of the unit vectors $(1\ 0\ 0 \cdots 0)'$, $(0\ 1\ 0 \cdots 0)'$, … , $(0 \cdots 0\ 0\ 1)'$. So, the standard errors of each the elements of $\tilde\beta$ are at least as large as the standard errors of the corresponding OLS estimators.

We will not prove efficiency of OLS estimators in this course. The proof is along the lines of the simple Example 43. If you are hungry for a proof, you can check any more advanced econometrics text book.

The standard errors of $\hat\beta$, or actually the variance-covariance matrix of $\hat\beta$ (after all, it is a random vector), can easily be derived. Denote this variance-covariance matrix by

$V(\hat{\beta})$. So, the (i,j)-th entry of $V(\hat{\beta})$ is $V_{ij}(\hat{\beta}) = \text{cov}(\hat{\beta}_i, \hat{\beta}_j)$. We have that

$$\begin{aligned} V(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] \\ &= \mathbb{E}\left[((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U})((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U})'\right] \\ &= \mathbb{E}\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U}\mathbf{U}'\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1}\right] \\ &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbb{E}[\mathbf{U}\mathbf{U}']\,\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1} \\ &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\sigma^2 I_n\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1} \\ &= \sigma^2\,(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1} \\ &= \sigma^2\,(\mathbf{x}'\mathbf{x})^{-1}. \end{aligned}$$

An unbiased estimator of the variance of the error term is again the sum of squared residuals divided by the appropriate degrees of freedom, which is now $n-k$. So, with

$$\hat{\sigma}^2 = \frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}}{n-k}$$

we have that $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$. An unbiased estimator of the variance-covariance matrix $V(\hat{\beta})$ is therefore given by $\hat{V}(\hat{\beta}) = \hat{\sigma}^2\,(\mathbf{x}'\mathbf{x})^{-1}$. The estimator of the standard error of the i-th element of $\hat{\beta}$ is simply the square root of the i-th diagonal element of $\hat{V}(\hat{\beta})$, $\sqrt{\hat{V}_{ii}(\hat{\beta})}$.

The other properties found for the simple OLS estimator can be extended as well. In particular, under some additional regularity conditions, $\hat{\beta}$ is consistent and asymptotically normal.

Finally, if the errors are assumed to be jointly normally distributed, then the OLS estimator $\hat{\beta}$ is not only BLUE, but even the best unbiased estimator (*i.e.*, it is efficient relative to *all* unbiased estimators, and not just relative to *linear* unbiased estimators). Also, $\hat{\beta}$ has a multivariate normal distribution distribution with mean β and variance-covariance matrix $V(\hat{\beta})$. In terms of the univariate normal distribution with which we are more familiar, this implies that each element of $\hat{\beta}$ is normally distributed with the corresponding parameter value as mean and the corresponding diagonal entry of $V(\hat{\beta})$ as variance. Also, the covariance between any two elements of $\hat{\beta}$ can be found at the relevant entry of $V(\hat{\beta})$.

4.8.4 Residual analysis and the coefficient of determination

Denote the vector of OLS predictions of $\mathbf{Y}$ by $\hat{\mathbf{Y}}$. So, $\hat{\mathbf{Y}} = \mathbf{x}\hat{\beta}$, and $\mathbf{Y} = \hat{\mathbf{Y}} + \hat{\mathbf{U}}$. Let $\bar{\mathbf{Y}}$ be the $(n \times 1)$-vector of sample averages of $\mathbf{Y}$ and therefore $\hat{\mathbf{Y}}$ (why?). So, if ι_n is a $(n \times 1)$-vector of ones and $\bar{Y}$ is the sample mean of Y, then $\bar{\mathbf{Y}} = \iota_n \bar{Y}$.

By the normal equations $\mathbf{x}'\hat{\mathbf{U}} = 0$, we again have that

$$\hat{\mathbf{U}}'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) = \hat{\mathbf{U}}'(\mathbf{x}\hat{\beta}) - \hat{\mathbf{U}}'(\iota_n \bar{Y}) = (\mathbf{x}'\hat{\mathbf{U}})'\hat{\beta} - (\iota_n'\hat{\mathbf{U}})\bar{Y} = 0\hat{\beta} + 0\bar{Y} = 0,$$

because ι_n is the first column of $\mathbf{x}$. Using this result, we find that

$$(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}) = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}} + \hat{\mathbf{U}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}} + \hat{\mathbf{U}}) = (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) + \hat{\mathbf{U}}'\hat{\mathbf{U}}.$$

Disguised as matrix algebra, you may not immediately recognize this, but this is nothing more or less than the decomposition of the sample variance of Y, or rather the total sum of squares, of Subsection 3.5:

$$TSS = ESS + RSS.$$

To see this, check that

$$\begin{aligned} TSS &= (\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}), \\ ESS &= (\hat{\mathbf{Y}} - \bar{\mathbf{Y}})'(\hat{\mathbf{Y}} - \bar{\mathbf{Y}}) \text{ and} \\ RSS &= \hat{\mathbf{U}}'\hat{\mathbf{U}} \end{aligned}$$

are consistent with the corresponding formulas for the simple regression model in Subsection 3.5. The coefficient of determination is again defined as $R^2 = ESS/TSS = 1 - RSS/TSS$, and the multiple correlation coefficient as $|R|$ (the positive square root of R^2).

In Subsection 4.7 we have seen that the R^2 never decreases if we add an additional regressor to a simple linear regression. For this reason, sometimes a coefficient of determination that is corrected for the "degrees of freedom" is reported. This *adjusted* R^2 is defined as

$$R_a^2 = 1 - \frac{RSS/(n-k)}{TSS/(n-1)} = 1 - \frac{\hat{\mathbf{U}}'\hat{\mathbf{U}}/(n-k)}{(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})/(n-1)} = 1 - \frac{\hat{\sigma}^2}{S_Y^2}.$$

It is important to note that $R_a^2 \leq 1$ like R^2, but that it is possible that $R_a^2 < 0$. The idea behind the adjustment is to create a statistic that can be used to decide upon the inclusion of additional regressors. However, this comes at a cost. Unlike R^2, R_a^2 has no nice interpretation as the fraction of the sample variance of the regressand explained by the regressor.

4.9 Some specification issues

4.9.1 Dummy regressors

In the problem sets, we have sometimes used so called *dummy variables*, *i.e.* variables that take only values 0 and 1. For example, we have used a dummy variable for sex in a wage equation, which allowed us to estimate and test the difference in wages between males and females. In this subsection, we discuss this in somewhat more detail.

Consider first the simplest case, in which we want to contrast two groups, say 1986 and 2000 PCs (see Example 50). We can construct a dummy variable D that equals 1 for PCs sold in 2000 and 0 for PCs sold in 1986. We could specify a regression of log prices (Y) on computer speed (X_1) and the year dummy (D) by

$$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 d_i + U_i.$$

This is just the three-variable regression model with $x_{2i} = d_i$. As discussed in Example 50, the dummy allows for a year-specific intercept (log price at 0 computer speed). The regression model assumes the same effect of computer speed on log price in both years. In the example, we discussed how including such a dummy helps to avoid omitted variable bias.

It is important to note that we only introduce one dummy variable for two years, just like we only introduced one dummy for two sexes in the problem set. We could of course specify another dummy variable, D^*, which equals 1 for 1986 computers and 0 for 2000 PCs. However, in a model with a constant, one of D and D^* is superfluous. After all, by construction we always have that $D + D^* = 1$, as each computer is either a 1986 or a 2000 model. So, in the sample $d_i + d_i^* = 1$ for all i, and a model with both dummies would suffer from perfect multicollinearity.

In general, suppose we want to use a set of $k-1$ dummies $D_1, \ldots, D_{k-1}$ as regressors.

We typically combine the dummies with other (continuous) regressors, but, for now, we concentrate on the dummies only. So, we regress Y on $X = (1 \;\; D_1 \cdots D_{k-1})$.

The dummies could correspond to a single categorical variable with k categories. For example, if we have wage data, we may actually observe the occupation of individuals. In a wage regression, we may want to condition on occupation. As occupation is a categorical variable, we could include dummies for each possible occupation. If we also include a constant, we would omit a dummy for one occupation to avoid multicollinearity. For example, if we distinguish 20 occupations, we would include dummies for 19 of these.

The dummies could also correspond to more than one categorical variable. For example, if we want to condition on both sex and occupation in a wage regression, we would include sex and occupation dummies. If we include a constant, we only want to include one dummy for the 2 sexes and 19 dummies for the 20 occupations. If we would include a dummy for each sex, or dummies for each of the 20 occupations, respectively the sex and the occupation dummies would again add to 1, and would be perfectly multicollinear with a constant.

In general, we should make sure to avoid perfect multicollinearity. We have perfect multicollinearity if $c_0 + c_1 d_{1i} + \cdots + c_{k-1} d_{(k-1)i} = 0$ for all i, for some $c_0, \ldots, c_1$ not all 0. As the examples above suggest, perfect multicollinearity typically arises if we include too many dummies. If we include a constant in the regression, we should never include a "full" set of dummies for a categorical variable. After all, as the categorical variable takes one and only one value for each observation, exactly one dummy in the full set of dummies is one. So, the dummies in a full set of dummies always add to 1, and a constant with a full set of dummies are perfectly multicollinear.

Example 51. A common application of dummies is correction for seasonality. If we are analyzing ice cream sales using quarterly price and sales data, we may worry that sales depend not only on price but also on the prevailing season. So, we may want to include a dummy that is 1 for observations in the July–September quarter, and 0 otherwise, because this quarter is usually hot with a lot of demand for ice cream at any price. We may also want to include dummies for the relatively cold January–March quarter because we expect sales to be low at a given price in the Winter. Now that we are at it, we could actually decide to include dummies for each quarter. If the first observation is for some January–March quarter and the last for some October-December quarter, this would give the following matrix of regressors:

$$\mathbf{x} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & p_1 \\ 1 & 0 & 1 & 0 & 0 & p_2 \\ 1 & 0 & 0 & 1 & 0 & p_3 \\ 1 & 0 & 0 & 0 & 1 & p_4 \\ 1 & 1 & 0 & 0 & 0 & p_5 \\ \vdots & & & & \vdots & \\ 1 & 1 & 0 & 0 & 0 & p_{n-3} \\ 1 & 0 & 1 & 0 & 0 & p_{n-2} \\ 1 & 0 & 0 & 1 & 0 & p_{n-1} \\ 1 & 0 & 0 & 0 & 1 & p_n \end{pmatrix}.$$

Each row corresponds to an observation for a particular quarter. The first column contains the constant, the next four column the quarter dummies, and the last column prices.

This is not a very smart choice, as the 4 quarter dummies always add to one. More formally, we have perfect multicollinearity because $\mathbf{x}c = 0$ for $c = (-1\ 1\ 1\ 1\ 1\ 0)'$. In yet other words, the rank of $\mathbf{x}$ is only 5 (if there is sufficient price variation; otherwise it would even be 4). So, we should exclude one quarter dummy (remove one of the dummy columns) if we include a constant in the regression.

We end this subsection with a short discussion of the interpretation of a dummy in a regression in logs. As an example, suppose we run a log wage regression including a sex dummy that is one for females. The parameter on this dummy is the mean difference between female and male mean log wages, holding all other regressors constant (*ceteris paribus*). For expositional convenience, suppose that the sex dummy is the only regressor, (apart from the constant). Then, we can further omit the *ceteris paribus* qualification, and the parameter on the dummy is simply the difference in mean log wages.

In problem set 2, we have seen that

$$\ln(w_m) - \ln(w_f) = \ln\left(\frac{w_m}{w_f}\right) = \ln\left(1 + \frac{w_m - w_f}{w_f}\right) \approx \frac{w_m - w_f}{w_f}, \tag{29}$$

for small values of $(w_m - w_f)/w_f$. So, if w_m is a male wage and w_f is a female wage, $\ln(w_m) - \ln(w_f)$ is approximately the difference between these two wages as a fraction of the female wage. So, the parameter on the dummy is approximately the mean percentage difference between male and female wages (divided by 100).

The advantage of using this approximation is that it allows for a direct interpretation of the parameter estimates. However, with dummy variables, we run into a problem that we did not have in Subsection 3.7.4 when discussing the case of continuous regressors. If we regress, for example, log wages on years of schooling we can think of the corresponding parameter as the relative change in wages in response to a very small change in schooling, for small and large values of the schooling parameter alike. In contrast, we cannot think of very small changes in sex. Usually, we consider a person to be either a man or a woman, and we can only change the sex dummy from 0 to 1, and *vice versa*. So, in the case of a sex dummy, or any other dummy variable, the approximation is only valid if the wage difference between the sexes, which is directly related to the corresponding parameter value, is small.

If we are worried about the error resulting from the approximation in (29) because the coefficient on the dummy is large, we could use that

$$\exp\left[\ln(w_m) - \ln(w_f)\right] - 1 = \exp\left[\ln\left(\frac{w_m}{w_f}\right)\right] - 1 = \frac{w_m}{w_f} - 1 = \frac{w_m - w_f}{w_f}. \tag{30}$$

This suggests an alternative approximation of the relative difference between female and male wages that is sometimes used, $\exp(\beta) - 1$. Here, β is the coefficient on the sex dummy in a log wage regression. The problem is that (30) only tells us that the

$$\text{expected value of}\quad \exp(\text{difference in log wages}) - 1$$

gives the expected relative wage difference between females and males. In contrast,

$$\exp(\beta) - 1 = \exp(\text{expected difference in log wages}) - 1,$$

which is not the same thing. We could however see $\exp(\beta) - 1$ as the relative wage difference between females and males conditional on U (*i.e.*, holding U constant; without taking expectations over the error term).

4.9.2 Higher order regressor terms

The linear regression model only has to be linear in the parameters, not in the variables. In problem set 5 we have seen that we can include both experience and experience squared as regressors if we want the partial effect of experience to vary with the level of experience.

For example, we may suspect that the marginal effect of experience on earnings is smaller at higher levels of experience.

We can also allow for interaction effects of regressors. For example, in problem set 5 we have included a regressor that equals years of schooling times experience. This allows the return to schooling to vary with experience. Of course, in that case the return to experience also varies with years of schooling. Intuitively, the level of schooling may be irrelevant once you have accumulated a lot of work experience, but may be very important for new labor market entrants.

To be more specific, consider the regression

$$\mathbb{E}(Y|X) = \alpha + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_1 X_2.$$

The partial effect of changing X_1 is $\partial \mathbb{E}(Y|X)/\partial X_1 = \beta_1 + 2\beta_2 X_1 + \beta_4 X_2$ and the partial effect of changing X_2 is $\partial \mathbb{E}(Y|X)/\partial X_2 = \beta_3 + \beta_4 X_1$. So, the partial effect of each regressor depends on the value of the other regressor because of the interaction term. Also, the effect of X_1 varies with X_1 because of the squared term. We can also allow for higher order polynomials and interaction terms. Note that we should include X_1 and X_2 itself to control for the overall level of the partial effects.

By interacting continuous regressors with dummy regressors, we can allow for different slopes for different groups or different periods in time (see problem set 6). For example, if we include years of schooling multiplied with a sex dummy in a log wage regression, we allow for different returns to schooling for males and females. Note that we also have to include a sex dummy and years of schooling itself. So, we would have

$$\mathbb{E}(Y|X, D) = \alpha + \beta_1 X + \beta_2 D + \beta_3 X D,$$

where Y the the log wage, X is years of schooling, and D is a dummy that is 1 for females and 0 for males. Then, β_1 is the coefficient on schooling for males, and $\beta_1 + \beta_3$ is the coefficient on schooling for females. Note that if we would have omitted X itself, we would have implicitly assumed that the coefficient on schooling for males is 0.

4.10 Hypothesis testing

As in Subsection 3.8, we assume, in addition to the classical assumptions, that the errors are normally and independently distributed. This ensures that the regression parameters are jointly normally distributed. As we know the distribution of the parameters, we can construct critical regions and confidence intervals.

4.10.1 Tests involving a single parameter or linear combination of parameters: t-tests

If we consider a test involving only a single parameter, for example the intercept α, or one of the slope parameters, then we can directly apply the results of Subsection 3.8. If we know the variance of the error term σ^2, we can construct appropriate Z-tests. If we do not know σ^2, we can substitute $\hat{\sigma}^2$ and construct a t-test. With k variables, the t-tests would now have t-distributions with $n-k$ degrees of freedom.[17] With two-sided hypotheses, we can alternatively work with confidence intervals.

As an example, suppose we test $H_0 : \beta_1 = \beta_{10}$ against $H_1 : \beta_1 \neq \beta_{10}$, where β_1 is the first slope parameter in a k-variable regression and β_{10} is some real number, for example 0. Then, we can base a test on

$$T_{\beta_{10}} = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\widehat{\text{var}(\hat{\beta}_1)}}},$$

where $\widehat{\text{var}(\hat{\beta}_1)}$ is the estimator of $\text{var}(\hat{\beta}_1)$. This is the second diagonal element of $\hat{V}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}$. As $\hat{\sigma}^2 = \hat{\mathbf{U}}'\hat{\mathbf{U}}/(n-k)$, $T_{\beta_{10}}$ has a t-distribution with $n-k$ degrees of freedom under H_0. If we choose a significance level, we can construct a two-sided critical region based on the appropriate quantiles of the t-distribution with $n-k$ degrees of freedom. Alternatively, we can construct a confidence interval for β_1 from T_{β_1}.

So, tests involving only a single parameter are not substantially different from similar tests in the simple regression model. It is also straightforward to test hypothesis involving a linear combination of regressor parameters. This is because any linear combination of OLS estimators is normal in the normal model.

First, suppose we want to test $H_0 : r_1\beta_1 + r_2\beta_2 = r_0$ against $H_1 : r_1\beta_1 + r_2\beta_2 \neq r_0$, for some real numbers r_1, r_2 and r_0. Here, β_1 and β_2 are the first two slope parameters in a k-variable regression model. It is intuitively clear that a test statistic can be based on the corresponding OLS estimators, or, more precisely, on $r_1\hat{\beta}_1 + r_2\hat{\beta}_2$. This is a linear combination of jointly normal random variables. As an extension of the result for sums of independent normal random variables in Subsection 2.3.7, it can be shown that any such linear combination of jointly normal random variables is again normal. In this case, $r_1\hat{\beta}_1 + r_2\hat{\beta}_2$ is normal with mean $r_1\beta_1 + r_2\beta_2$ and variance (see Subsection 4.4.2)

$$\begin{aligned}
&r_1^2 \operatorname{var}(\hat{\beta}_1) + r_2^2 \operatorname{var}(\hat{\beta}_2) + 2r_1r_2 \operatorname{cov}(\hat{\beta}_1, \hat{\beta}_2) \\
&= \frac{r_1^2\sigma^2}{(n-1)s^2_{X_1}\left[1-\hat{\rho}^2_{X_1X_2}\right]} + \frac{r_2^2\sigma^2}{(n-1)s^2_{X_2}\left[1-\hat{\rho}^2_{X_1X_2}\right]} - \frac{2r_1r_2\sigma^2 s_{X_1X_2}}{(n-1)s^2_{X_1}s^2_{X_2}\left[1-\hat{\rho}^2_{X_1X_2}\right]} \\
&= \frac{\sigma^2}{(n-1)}\left[\frac{r_1^2}{s^2_{X_1}\left[1-\hat{\rho}^2_{X_1X_2}\right]} + \frac{r_2^2}{s^2_{X_2}\left[1-\hat{\rho}^2_{X_1X_2}\right]} - \frac{2r_1r_2 s_{X_1X_2}}{s^2_{X_1}s^2_{X_2}\left[1-\hat{\rho}^2_{X_1X_2}\right]}\right].
\end{aligned}$$

So, we can construct a Z-statistic

$$Z_{r_0} = \frac{r_1\beta_1 + r_2\beta_2 - r_0}{\sqrt{\sigma^2\left[r_1^2\frac{1}{(n-1)s^2_{X_1}\left[1-\hat{\rho}^2_{X_1X_2}\right]} + r_2^2\frac{1}{(n-1)s^2_{X_2}\left[1-\hat{\rho}^2_{X_1X_2}\right]} - 2r_1r_2\frac{s_{X_1X_2}}{(n-1)s^2_{X_1}s^2_{X_2}\left[1-\hat{\rho}^2_{X_1X_2}\right]}\right]}}.$$

If we do not σ^2, we can construct the corresponding t-statistic T_{r_0} by substituting $\hat{\sigma}^2$ for σ^2. Alternatively, we can construct confidence intervals for $r_1\beta_1 + r_2\beta_2$ from $Z_{r_1\beta_1+r_2\beta_2}$ or $T_{r_1\beta_1+r_2\beta_2}$.

We can write this in matrix notation by introducing the $(1\times k)$-vector $R = (0 \;\; r_1 \;\; r_2 \;\; 0\cdots 0)$. Recalling that $\beta = (\alpha \quad \beta_1\cdots\beta_{k-1})'$, we can write the hypotheses as $H_0 : R\beta = r_0$ and $H_1 : R\beta \neq r_0$. Also, it is easy to check that the variance of $R\hat{\beta}$ is

$$\operatorname{var}(R\hat{\beta}) = R\operatorname{var}(\hat{\beta})R' = R(\sigma^2(\mathbf{x}'\mathbf{x})^{-1})R' = \sigma^2 R(\mathbf{x}'\mathbf{x})^{-1}R'.$$

If we know σ^2, we can construct the Z-statistic

$$Z_{r_0} = \frac{R\hat{\beta} - r_0}{\sqrt{\sigma^2 R(\mathbf{x}'\mathbf{x})^{-1}R'}},$$

which is standard normal under H_0. If we do not know σ^2, we can construct the t-statistic

$$T_{r_0} = \frac{R\hat{\beta} - r_0}{\sqrt{\hat{\sigma}^2 R(\mathbf{x}'\mathbf{x})^{-1}R'}},$$

which has a t-distribution with $n-k$ degrees of freedom under H_0. Alternatively, we can construct confidence intervals based on $Z_{R\beta}$ or $T_{R\beta}$.

Of course, the latter derivation in matrix notation is not specific to the particular R of our example. In general, we can use a t-statistic for any test involving a $(1\times k)$-vector R with corresponding null hypothesis $H_0 : R\beta = r_0$.

Example 52. The Cobb-Douglas production function is given by

$$Y = F(K, L; U) = \exp(\alpha) K^{\beta_1} L^{\beta_2} \exp(U), \tag{31}$$

where Y is output, K is capital input, L is labor input, and U is a productivity shock such that $\mathbb{E}(U|K, L) = 0$. Suppose we change the inputs from (K, L) to $(\lambda K, \lambda L)$, for some $\lambda > 0$. Then, new output is related to old output by

$$F(\lambda K, \lambda L; U) = \exp(\alpha)(\lambda K)^{\beta_1}(\lambda L)^{\beta_2} \exp(U) = \lambda^{\beta_1+\beta_2} F(K, L).$$

So, $F(K, L; U)$ has *constant returns to scale* if $\beta_1 + \beta_2 = 1$. In other words, if $\beta_1 + \beta_2 = 1$, then an $x\%$ increase in both capital and labor inputs leads to an $x\%$ increase in output.

The returns to scale in production are important in economics and management. If there are constant returns to scale, it doesn't matter whether production takes place in a couple of big plants, or many small plants. However, with increasing returns to scale, you would prefer to concentrate production in a single unit.

So, suppose that we have data on capital and labor inputs and output, and that we want to test for constant returns to scale. We can specify the null hypothesis of constant returns to scale as $H_0 : \beta_1 + \beta_2 = 1$. Taking logs in (31) gives the linear population regression

$$\ln(Y) = \alpha + \beta_1 \ln(K) + \beta_2 \ln(L) + U \quad \text{and} \quad \mathbb{E}(U|\ln(K), \ln(L)) = 0.$$

So, H_0 involves a linear combination of slope parameters in a linear regression model, and fits the setup of this section.

4.10.2 Joint hypotheses: F-tests

In this section, we first consider *joint null hypotheses* of the form

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{k-1} = 0$$

in a k-variable regression model. We test these against the alternative

$$H_1 : \beta_i \neq 0 \text{ for at least one } i \ (1 \leq i \leq k-1).$$

This is a test on the *joint significance* of the slope parameters. H_0 is a joint hypothesis, as it requires that jointly $\beta_1 = 0$,...,$\beta_{k-2} = 0$ and $\beta_{k-1} = 0$. In the previous subsection, we have already seen how to test each of these hypotheses separately. However, this is generally not the same as testing the joint hypothesis H_0. This is because the estimators $\hat\beta_1$, $\hat\beta_2$,...,$\hat\beta_{k-1}$, and therefore the corresponding test statistics for the separate tests, are typically dependent.

A joint test can be based on the slope estimators $\hat\beta_1$, $\hat\beta_2$,...,$\hat\beta_{k-1}$. Under H_0, all these estimators have expected value 0. So, it is natural to base a test on the difference between the elements of the vector $(\hat\beta_1 \quad \hat\beta_2 \cdots \hat\beta_{k-1})$ and 0. This is less straightforward than the tests in the previous subsection, which only involved the distance between one scalar estimator and an hypothesized scalar value of the corresponding parameter. Now, we need an appropriate measure of the "distance" of a vector to 0.

To develop some intuition for this problem, consider the three-variable case $k = 3$. In this case, we are testing

$$H_0 : \beta_1 = \beta_2 = 0 \quad \text{against} \quad H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

We base the test statistic on $\hat\beta_1$ and $\hat\beta_2$. The variance-covariance matrix of $(\hat\beta_1 \quad \hat\beta_2)'$ is

$$\begin{aligned} V\begin{pmatrix} \hat\beta_1 \\ \hat\beta_2 \end{pmatrix} &= \begin{pmatrix} \mathrm{var}(\hat\beta_1) & \mathrm{cov}(\hat\beta_1,\hat\beta_2) \\ \mathrm{cov}(\hat\beta_1,\hat\beta_2) & \mathrm{var}(\hat\beta_2) \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} \frac{1}{(n-1)s^2_{X_1}\left[1-\hat\rho^2_{X_1X_2}\right]} & \frac{-s_{X_1X_2}}{(n-1)s^2_{X_1}s^2_{X_2}\left[1-\hat\rho^2_{X_1X_2}\right]} \\ \frac{-s_{X_1X_2}}{(n-1)s^2_{X_1}s^2_{X_2}\left[1-\hat\rho^2_{X_1X_2}\right]} & \frac{1}{(n-1)s^2_{X_2}\left[1-\hat\rho^2_{X_1X_2}\right]} \end{pmatrix}, \end{aligned}$$

First, suppose that the regressors are uncorrelated, so that $s_{X_1X_2} = 0$. Then, $\hat\beta_1$ and $\hat\beta_2$ are uncorrelated, and even independent because uncorrelated jointly normal random variables are independent. A measure of the distance between $\hat\beta_1$ and 0 is $(\hat\beta_1 - 0)^2 = \hat\beta_1^2$, which is 0 if $\hat\beta_1 = 0$ and positive if $\hat\beta_1 < 0$ or $\hat\beta_1 > 0$. Similarly, a measure of the distance between $\hat\beta_2$ and 0 is $\hat\beta_2^2$. Under H_0, both $\hat\beta_1$ and $\hat\beta_2$ should be small. We can combine both distance measures in a single statistic,

$$\chi^2 = \left[\frac{\hat\beta_1}{\sqrt{\mathrm{var}(\hat\beta_1)}}\right]^2 + \left[\frac{\hat\beta_2}{\sqrt{\mathrm{var}(\hat\beta_2)}}\right]^2.$$

Under H_0, $\hat\beta_1/\sqrt{\mathrm{var}(\hat\beta_1)}$ and $\hat\beta_2/\sqrt{\mathrm{var}(\hat\beta_2)}$ are independent Z-statistics, *i.e.* standard normal random variables. So, χ^2 is the sum of two independent standard normal random variables squared. We have seen in Subsection 2.3.7 that such a statistic has a χ^2-distribution with 2 degrees of freedom.

Obviously, χ^2 is always nonnegative. If H_0 is true, we expect $\hat{\beta}_1$ and $\hat{\beta}_2$ to be close to 0, so that χ^2 is small. If H_0 is false, we expect χ^2 to be large. So, if we know σ^2, and therefore $\text{var}(\hat{\beta}_1)$ and $\text{var}(\hat{\beta}_2)$, we can base a test on χ^2. This would be a one-sided test: we would reject H_0 if χ^2 exceeds some critical value. Appropriate critical values can be found in statistical tables of the χ^2-distribution.

Unfortunately, we usually do not know σ^2, so that we cannot compute χ^2-statistics. Like Z-statistics, χ^2-statistics are typically not feasible. In the case of Z-statistics, this was easily solved by substituting an estimator $\hat{\sigma}^2$ for σ^2, giving t-statistics. This suggests substituting an estimator of σ^2 in our χ^2-statistic, which gives

$$F = \frac{1}{2}\left[\frac{\hat{\beta}_1^2}{\widehat{\text{var}(\hat{\beta}_1)}} + \frac{\hat{\beta}_2^2}{\widehat{\text{var}(\hat{\beta}_2)}}\right], \tag{32}$$

which can be shown to be an F-statistic with 2 and $n-3$ degrees of freedom under H_0 (see Subsection 2.3.7 for a discussion of the F-distribution).[18] Note that I did not only substitute $\hat{\sigma}^2$ for σ^2, but that I also divided by 2, the number of restrictions tested, in order to get an F-statistic. Like the χ^2-statistic, the F-statistic is nonnegative. It can be expected to be small if H_0 is true, and large if H_1 is true. So, we should again reject H_0 if F exceeds a critical value. An appropriate critical value can be found in statistical tables of the F-distribution.

So far, we have focused on uncorrelated estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. However, the test statistics extend directly to the general case in which $s_{X_1X_2}$ may be nonzero. In particular

$$\begin{aligned} F &= \frac{1}{2}\begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_2 \end{pmatrix}\left[\hat{V}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}\right]^{-1}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \\ &= \frac{1}{2}\begin{pmatrix} \hat{\beta}_1 & \hat{\beta}_2 \end{pmatrix}\begin{pmatrix} \widehat{\text{var}(\hat{\beta}_1)} & \widehat{\text{cov}(\hat{\beta}_1,\hat{\beta}_2)} \\ \widehat{\text{cov}(\hat{\beta}_1,\hat{\beta}_2)} & \widehat{\text{var}(\hat{\beta}_2)} \end{pmatrix}^{-1}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \end{aligned}$$

is still an F-statistic with 2 and $n-3$ degrees of freedom under H_0. It is easy to check that it reduces to the simpler F-statistic in equation (32) if $\widehat{\text{cov}(\hat{\beta}_1,\hat{\beta}_2)} = 0$.

In the general k-variable case, let R be a $(k \times k)$-matrix that is zero except for the second until the last diagonal entry, which are 1, so that $R\hat{\beta} = (0 \quad \hat{\beta}_1 \cdots \cdots \hat{\beta}_{k-1})$ is the vector of slope estimators. Then, $\hat{V}(R\hat{\beta}) = \hat{\sigma}^2 R(\mathbf{x}'\mathbf{x})^{-1}R'$, which is the estimated variance-covariance matrix $\hat{V}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}$ of $\hat{\beta}$ with the first column and row replaced by zeros. Then, the hypotheses can be written as $H_0 : R\beta = 0$ and $H_1 : R\beta \neq 0$. The corresponding F-statistic

$$F = \frac{(R\hat{\beta})' \left[\hat{V}(R\hat{\beta})\right]^{-1} R\hat{\beta}}{k-1} = \frac{(R\hat{\beta})' \left[R(\mathbf{x}'\mathbf{x})^{-1}R'\right]^{-1} R\hat{\beta} \Big/ (k-1)}{\hat{\sigma}^2},$$

can be shown to have an F-distribution with $k-1$ and $n-k$ degrees of freedom under H_0.

F-statistics can also be constructed for other tests involving more than one linear restriction on the regression parameters. We can specify a null hypothesis that a subset of the slope parameters equals 0 (joint significance of a subset of the parameters). In problem set 6, for example, we introduced a set of regional dummies in an opium consumption regression. The STATA regression procedure `areg` does not report the estimates of the corresponding parameters, but only an F-statistic for their joint significance. In this case, with 22 regions and 21 dummies, this gives an F-statistic with 21 and $n-k$ degrees of freedom, where k is the total number of regression parameters (including those for the dummies). If this F-statistic is sufficiently large, we reject the null hypothesis that all 21 region dummies are 0.

More generally, we can combine various linear restrictions on the parameters in joint null hypotheses. The notation above suggests a straightforward extension. If we write the restrictions in matrix notation as $H_0 : R\beta = r_0$, for some $((k-l) \times k)$-matrix R, then

$$F = \frac{(R\hat{\beta} - r_0)' \left[\hat{V}(R\hat{\beta})\right]^{-1} (R\hat{\beta} - r_0)}{k-l} = \frac{(R\hat{\beta} - r_0)' \left[R(\mathbf{x}'\mathbf{x})^{-1}R'\right]^{-1} (R\hat{\beta} - r_0) \Big/ (k-l)}{\hat{\sigma}^2}$$

has an F-distribution with $k-l$ and $n-k$ degrees of freedom under H_0. Here, $k-l$ is the rank of R, the number of restrictions in H_0, with $0 \leq l < k$. R could for example be a matrix with zeros, except for a subset of $k-l$ of the diagonal entries, for a significance test of a subset of $k-l$ parameters. Also, R could combine $k-l$ different linear restrictions of the type discussed in the previous subsection. For now, it is sufficient to understand that a wide variety of joint tests can be written in this form, and that we can generally construct F-statistics for such tests in the classical normal model.

Gujarati (1995) shows that these F-tests can also be written as the (relative) difference of the RSS of an unrestricted model and the RSS of a model on which the restrictions $R\beta = r_0$ are imposed. It is good to take note of this, but we will not discuss it in this course.

5 Extensions of the classical framework

5.1 Stochastic regressors

So far, we have assumed that the regressors are deterministic, *i.e.* fixed in repeated sampling (Assumptions 4, 4^* and $4^\dagger$). We formulated the classical assumptions in Subsections 3.3, 4.3 and 4.8.2 conditional on the regressors, but subsequently dropped the conditioning in the analyses. After all, there is no purpose in conditioning on deterministic variables. Instead, we have just taken the regressors as given non-random numbers throughout.

The reason for conditioning on the regressors in the first place is that it allows us to relax Assumption $4^\dagger$ without changing the other assumptions. To see this, reconsider the general multiple regression model of Subsection 4.8, and suppose we drop Assumption $4^\dagger$ (obviously, all results specialize to the simple and three-variable models). So, we allow the regressor matrix $\mathbf{X}$ to be random. We can maintain the other classical assumptions as they are, as we formulated each of them conditional on $\mathbf{X}$. Obviously, if we would like to add the normality assumption, we would assume normality conditional on $\mathbf{X}$, as in Assumption 5. Normality is however not relevant to the argument below.

The analysis in Subsection 4.8.3 takes the regressor matrix $\mathbf{x}$ as a given, non-random matrix. Alternatively, without Assumption $4^\dagger$, we can view the entire analysis as being *conditional on* $\mathbf{X}=\mathbf{x}$. After all, even if $\mathbf{X}$ is stochastic, we can take it as given and equal to $\mathbf{x}$ if we condition on $\mathbf{X}=\mathbf{x}$. So, the analysis is the same. By evaluating the conditional results at the random $\mathbf{X}$ and taking expectations, we get unconditional results by the law of the iterated expectations.

Let me illustrate this by replicating some of the analysis for the case of stochastic regressors. The OLS estimator of β is

$$\hat{\beta}=\beta+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{U}.$$

Here, I use capital $\mathbf{X}$ to make explicit that we now allow for stochastic regressors. We now have that

$$\mathbb{E}(\hat{\beta}|\mathbf{X})=\beta+\mathbb{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{U}|\mathbf{X}\right]=\beta+\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbb{E}(\mathbf{U}|\mathbf{X})=\beta,$$

so that $\mathbb{E}(\hat{\beta})=\mathbb{E}[\mathbb{E}(\hat{\beta}|\mathbf{X})]=0$ by the law of the iterated expectations.

Also, recall that variances and covariances are just expectations of appropriate functions of random variables. So, we can talk about conditional variances and covariances, and derive

$$V(\hat{\beta}|\mathbf{X}) = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

The only difference with the case of deterministic regressors is that $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$ is a random variable, unlike $\sigma^2 \left(\mathbf{x}'\mathbf{x}\right)^{-1}$. By the law of the iterated expectations, we have that

$$V(\hat{\beta}) = \mathbb{E}[V(\hat{\beta}|\mathbf{X})] = \sigma^2 \mathbb{E}[\left(\mathbf{X}'\mathbf{X}\right)^{-1}].$$

Now, we do not know $\mathbb{E}[\left(\mathbf{X}'\mathbf{X}\right)^{-1}]$, and we also need some additional assumptions to ensure it exists. However, for the purpose of this course, it suffices to understand that it is easy to extend the analysis to stochastic regressors by subsequently exploiting conditioning and the law of the iterated expectations.

5.2 Non-spherical errors and generalized least squares

In this subsection we investigate the consequences of non-spherical errors, *i.e.* of violation of Assumption $2^{\dagger}$ (or Assumption 2 or 2^{*}). In the first two subsections, we consider two special, but common, cases, heteroskedasticity (without autocorrelation) and first-order autoregressive errors (with homoskedasticity).

As we will see, the cases have much in common. We can effectively deal with non-spherical errors by transforming the model into a model with spherical errors, and estimating this model by OLS. The corresponding estimator is called the *generalized least squares* (GLS) estimator. The GLS estimator minimizes a weighted sum of squared residuals, just like the OLS estimator minimizes an unweighted sum of squared residuals. We end with some general considerations along these lines in Subsection 5.2.3.

5.2.1 Heteroskedasticity

First, suppose we replace Assumption $2^{\dagger}$ by

Assumption $2^{\ddagger}$. (heteroskedasticity) The errors are heteroskedastic, but uncorrelated, *i.e.* $\text{var}(U_i) = \sigma_i^2$ for some $\sigma_i > 0$, $i = 1, \ldots, n$, and $\text{cov}(U_i, U_j) = 0$ for all

$i \neq j$, or, in matrix notation,

$$\mathbb{E}[\mathbf{U}\mathbf{U}'|\mathbf{X}] = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_{n-1}^2 & 0 \\ 0 & \cdots & 0 & 0 & \sigma_n^2 \end{pmatrix} = \Sigma.$$

In this section, we use the notation $\Sigma = V(\mathbf{U})$ for the variance-covariance matrix of the error vector $\mathbf{U}$.

The question now is whether we should still use the OLS estimator to estimate β. Fortunately, even with heteroskedasticity, the OLS estimator $\hat{\beta}$ is unbiased. After all, $\hat{\beta}$ is linear in the errors U_i, so that the expectation of $\mathbb{E}(\hat{\beta})$ only involves first moments of U_i. Assumption $2^\dagger$ is concerned with the second moments, *i.e.* with variances and covariances, and is irrelevant to the derivation of the expectation of $\hat{\beta}$.

However, the expression for the variance-covariance matrix of the OLS estimator derived earlier is no longer valid. To see this, recall that the OLS estimator is

$$\hat{\beta} = \beta + (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U},$$

so that

$$\begin{aligned} V(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'\right] \\ &= \mathbb{E}\left[((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U})((\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U})'\right] \\ &= \mathbb{E}\left[(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{U}\mathbf{U}'\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1}\right] \\ &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbb{E}[\mathbf{U}\mathbf{U}']\,\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1} \\ &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\Sigma\mathbf{x}\,(\mathbf{x}'\mathbf{x})^{-1}. \end{aligned} \tag{33}$$

So, it would not be appropriate to base the standard errors on the expression of the variance-variance matrix that we derived earlier for the homoskedastic case.

All in all, this seems to suggest that we can still use the OLS estimator $\hat{\beta}$, as long as we base our estimates of the standard errors on the proper variance-covariance matrix in (33). OLS procedures in statistical packages typically allow you to compute such *heteroskedasticity-corrected* standard errors instead of the usual standard errors (see for example the discussion of the White (1980) covariance-matrix-estimator in Gujarati, 1995).

So, if you are estimating a model by OLS and you suspect that the errors are heteroskedastic, it is easy to provide a correct estimate of the standard errors.

However, the Gauss-Markov theorem uses all classical assumptions, including homoskedasticity. We have no guarantee that the OLS estimator is BLUE in the case of heteroskedasticity. Indeed, it can be shown that $\hat{\beta}$ is, in general, not efficient anymore (in the class of linear unbiased estimators). To see this, note that we can rewrite the linear regression model

$$Y_i = \mathbf{x}_i\beta + U_i \quad \text{and} \quad \mathbb{E}(U_i|\mathbf{X}) = 0$$

into

$$\frac{Y_i}{\sigma_i} = \frac{\mathbf{x}_i}{\sigma_i}\beta + \frac{U_i}{\sigma_i} = \frac{\mathbf{x}_i}{\sigma_i}\beta + U_i^* \quad \text{and} \quad \mathbb{E}(U_i^*|\mathbf{X}) = 0, \tag{34}$$

where $U_i^* = U_i/\sigma_i$. Note that $\text{var}(U_i^*) = \text{var}(U_i)/\sigma_i^2 = 1$, so that $\mathbb{E}(\mathbf{U}^*\mathbf{U}^{*\prime}|\mathbf{X}) = I_n$.

So, the errors in the transformed regression equation are spherical. If we know $\sigma_1, \dots, \sigma_n$, the Gauss-Markov theorem implies that we can efficiently estimate β by OLS on the transformed regression. Let

$$Q = \begin{pmatrix} \frac{1}{\sigma_1} & 0 & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \frac{1}{\sigma_{n-1}} & 0 \\ 0 & \cdots & 0 & 0 & \frac{1}{\sigma_n} \end{pmatrix}.$$

Note that $Q'Q = \Sigma^{-1}$. In this notation, we can write the transformed regression concisely as

$$Q\mathbf{Y} = Q\mathbf{x}\beta + \mathbf{U}^* \quad \text{and} \quad \mathbb{E}(\mathbf{U}^*|\mathbf{X}) = 0, \tag{35}$$

where $\mathbf{U}^* = Q\mathbf{U}$. This is just repeating (34) in matrix notation. The OLS estimator of β in the transformed regression (35) is

$$\begin{aligned}\tilde{\beta} &= \left[(Q\mathbf{x})'(Q\mathbf{x})\right]^{-1}(Q\mathbf{x})'Q\mathbf{Y} \\ &= \left(\mathbf{x}'Q'Q\mathbf{x}\right)^{-1}\mathbf{x}'Q'Q\mathbf{Y} \\ &= \left(\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{Y}.\end{aligned}$$

This estimator $\tilde{\beta}$ is a weighted version of the OLS estimator of the original model, weighting each observation i by the inverse standard deviation σ_i^{-1}. It is a special case of a *generalized least squares* (GLS) estimator. It is BLUE, as it is the OLS estimator of a (transformed) model that satisfies the classical Assumptions $1^{\dagger}$–$4^{\dagger}$. In turn, this implies that the OLS estimator of the original model is not BLUE, unless $\Sigma = \sigma^2 I_n$ for some some $\sigma > 0$ and $\tilde{\beta} = \hat{\beta}$ (check!).

The variance-covariance matrix of the GLS estimator follows directly by using the formula for the variance-covariance matrix of the OLS estimator of the transformed model (35),

$$V(\tilde{\beta}) = \mathbb{E}[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] = (\mathbf{x}'Q'Q\mathbf{x})^{-1} = \left(\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-1}.$$

So far, we have assumed that we know Σ. In practice, this is usually not the case, and GLS is not feasible. Our earlier experience with similar problems suggests that we first estimate Σ, say by $\hat{\Sigma}$, and then estimate β by

$$\hat{\tilde{\beta}} = \left(\mathbf{x}'\hat{\Sigma}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}'\hat{\Sigma}^{-1}\mathbf{Y}.$$

This two-step procedure is called *feasible GLS*. The first step typically entails an OLS regression, which delivers an unbiased estimate $\hat{\beta}$ of β. The residuals $\hat{U}$ of this regression can then be used to estimate Σ. This requires some additional assumptions on the nature of the heteroskedasticity. Typically, some relation between σ_i and the regressors is assumed that only depends on a few unknown parameters. The estimation of Σ in the first step then boils down to estimating these parameters from the OLS residuals.

We will not discuss the details of this procedure. For now, it is sufficient to understand the general idea, so that you can understand what a statistical package like STATA does when you let it compute (feasible) GLS estimates.

Given that feasible GLS is substantially more burdensome than OLS, we may first want to test for the presence of heteroskedasticity in the data. Like the feasible GLS estimator, heteroskedasticity tests often assume a simple relation between the variances and the regressors. The tests then reduce to simple tests on the significance of these relations. An example is the Goldfeld-Quandt test discussed in Gujarati (1995). Please read the relevant sections of the book to get some feel for the way these tests work.

5.2.2 Autocorrelation

Next, suppose we have homoskedastic but correlated errors. This typically occurs in time-series data, in which the sample consists of consecutive observations of a random variable over time. In the case of time-series data, it is common to index the sample by t ("time") instead of i. Using that notation, we replace Assumption $2^{\dagger}$ by

Assumption $2^{\diamondsuit}$. (first-order autoregressive errors) The errors are homoskedastic and first-order autoregressive ($AR(1)$):

$$U_t = \rho U_{t-1} + V_t, \quad -1 < \rho < 1, \tag{36}$$

where $\mathbb{E}(\mathbf{V}\mathbf{V}'|\mathbf{X}) = \sigma_v^2 I_n$ for some $\sigma_v > 0$.

So, $\mathbf{V} = (V_1 \cdots V_n)'$ is assumed to be spherical, but the errors U_i themselves are correlated if $\rho \neq 0$.

We will first derive the variance-covariance-matrix of $\mathbf{U}$. Note that the variance σ^2 of U_t does not depend on t.[19] Furthermore, σ^2 is not the same as the variance σ_v^2 of V_t that figures in Assumption $2^{\diamondsuit}$. Using equation (36), we find that

$$U_t = \sum_{i=0}^{\infty} \rho^i V_{t-i}. \tag{37}$$

Together with the assumption that V_t is not autocorrelated, this implies that $\mathbb{E}(U_{t-k}V_t) = 0$ for all $k \geq 1$. As a consequence,

$$\sigma^2 = \mathrm{var}(U_t) = \mathrm{var}(\rho U_{t-1} + V_t) = \rho^2 \,\mathrm{var}(U_{t-1}) + \mathrm{var}(V_t) = \rho^2\sigma^2 + \sigma_v^2,$$

so that

$$\sigma^2 = \frac{\sigma_v^2}{1-\rho^2}.$$

For $k \geq 1$, we can again use (37) and derive that

$$\begin{aligned}
\mathrm{cov}(U_t, U_{t-k}) &= \mathrm{cov}(\rho^k U_{t-k} + \sum_{i=0}^{k-1} \rho^i V_{t-i}, U_{t-k}) \\
&= \mathrm{cov}(\rho^k U_{t-k}, U_{t-k}) + \mathrm{cov}(\sum_{i=0}^{k-1} \rho^i V_{t-i}, U_{t-k}) \\
&= \rho^k \,\mathrm{var}(U_{t-k}) + \sum_{i=0}^{k-1} \rho^i \,\mathrm{cov}(V_{t-i}, U_{t-k}) \\
&= \rho^k \sigma^2.
\end{aligned}$$

So, the correlation between U_t and U_{t-k} is ρ^k. Taken together, this implies that the variance-covariance matrix of $\mathbf{U}$ is given by $\Sigma = V(\mathbf{U}) = \sigma_v^2 \Omega$, where

$$\Omega = \frac{1}{1-\rho^2} \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \rho^{n-2} & \cdots & \rho & 1 & \rho \\ \rho^{n-1} & \cdots & \rho^2 & \rho & 1 \end{pmatrix}.$$

Again, the OLS estimator of β in this model is linear and unbiased, but generally not BLUE. As in the previous subsection, we can transform the model so that the transformed errors are spherical. In particular, note that

$$\begin{aligned} Y_t - \rho Y_{t-1} &= \mathbf{x}_t\beta + U_t - \rho(\mathbf{x}_{t-1}\beta + U_{t-1}) \\ &= (\mathbf{x}_t - \rho\mathbf{x}_{t-1})\beta + U_t - \rho U_{t-1} \qquad (38) \\ &= (\mathbf{x}_t - \rho\mathbf{x}_{t-1})\beta + V_t. \end{aligned}$$

If we know ρ, this suggests that we estimate β by the OLS estimator of the transformed model (38), as in the previous subsection. There is one problem though: we cannot use data for $t = 1$ as we do not observe Y_0 and $\mathbf{x}_0$. So, we use only data for $t = 2, \dots, n$, which gives

$$\tilde{\beta}^* = \left[\sum_{t=2}^{n} (\mathbf{x}_t - \rho\mathbf{x}_{t-1})'(\mathbf{x}_t - \rho\mathbf{x}_{t-1})\right]^{-1} \sum_{t=2}^{n} (\mathbf{x}_t - \rho\mathbf{x}_{t-1})'(Y_t - \rho Y_{t-1}).$$

I am using the slightly different symbol $\tilde{\beta}^*$ for the estimator here, as $\tilde{\beta}^*$ is not really the GLS estimator. After all, we are not using data on the first observation. So, even though we are now "correcting" for autocorrelation, $\tilde{\beta}^*$ cannot be efficient (is not BLUE), as it does not use all available information. It captures the main idea of the GLS estimator though, and we will argue below that it is pretty much the same if the sample size is large enough.

The "true" GLS estimator is most easily derived in matrix notation. The inverse of the matrix Ω defined above can be shown to be

$$\Omega^{-1} = \begin{pmatrix} 1 & -\rho & 0 & \cdots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & 0 \\ 0 & -\rho & 1+\rho^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -\rho & 0 \\ 0 & \cdots & 0 & -\rho & 1+\rho^2 & -\rho \\ 0 & 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}.$$

You can verify this by checking that $\Omega^{-1}\Omega = I_n$. Furthermore, it is easy to check that $\Omega^{-1} = Q'Q$, with

$$Q = \begin{pmatrix} \sqrt{1-\rho^2} & 0 & 0 & \cdots & 0 \\ -\rho & 1 & 0 & \cdots & 0 \\ 0 & -\rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & -\rho & 1 \end{pmatrix}.$$

If we transform the regression model into

$$Q\mathbf{Y} = Q\mathbf{x}\beta + \mathbf{U}^* \quad \text{and} \quad \mathbb{E}(\mathbf{U}^*|\mathbf{X}) = 0,$$

as in (35), then it can be shown that $V(\mathbf{U}^*|\mathbf{X}) = \sigma_v^2 I_n$.[20] So, the transformed model again satisfies the classical assumptions, and the OLS estimator

$$\begin{aligned}\tilde{\beta} &= \left[(Q\mathbf{x})'(Q\mathbf{x})\right]^{-1}(Q\mathbf{x})'Q\mathbf{Y} \\ &= \left(\mathbf{x}'Q'Q\mathbf{x}\right)^{-1}\mathbf{x}'Q'Q\mathbf{Y} \\ &= \left(\mathbf{x}'\Omega^{-1}\mathbf{x}\right)^{-1}\mathbf{x}'\Omega^{-1}\mathbf{Y}.\end{aligned}$$

of the transformed model is again BLUE by the Gauss-Markov theorem.

We could as well have transformed the model by multiplying by $\sigma_v^{-1}Q$. Note that $(\sigma_v^{-1}Q)'(\sigma_v^{-1}Q) = \sigma_v^{-2}\Omega^{-1} = \Sigma^{-1}$. So, in line with the previous subsection, this would have given an transformed error term with variance-covariance matrix I_n, and a GLS estimator $\left(\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{Y}$. However, this estimator is equivalent to $\tilde{\beta}$ above, as σ_v^2 cancels. As we have to estimate the unknown parameters in the feasible GLS procedure below, we better get rid of as many of these as we can, and use the expression for $\tilde{\beta}$ above.

The variance-covariance matrix of the GLS estimator again follows by using the formula for the variance-covariance matrix of the OLS estimator of the transformed model. It has exactly the same form as the variance-covariance matrix of the previous subsection,

$$V(\tilde{\beta}) = \left(\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-1}.$$

Note that Σ^{-1} and not Ω^{-1} enters this expression (check!).

$\tilde{\beta}$, as opposed to $\tilde{\beta}^*$, is the GLS estimator of β, but the two estimators are very closely related. If we inspect the matrix Q that we use to transform our model to derive $\tilde{\beta}$, it is clear that the second up to the last row simply deliver the $n-1$ "differenced" equations in (38). The first row only involves the first observation, and was omitted from (38). It is intuitively clear that in practice, if n is large, the simpler estimator $\tilde{\beta}^*$ will not perform much worse than the GLS estimator. Theoretically though, $\tilde{\beta}$ is superior.

As in the previous subsection, GLS estimation is typically not feasible. In this case, the "weighting" matrix Ω^{-1} depends on the autocorrelation parameter ρ. Again, we could, for example, estimate ρ from OLS residuals in a first stage, and then use the estimated value of ρ in a second GLS stage.

There are also various tests on autocorrelation that are based on OLS residuals. The most well-known example is the Durbin-Watson test. See Gujarati (1995) for details.

5.2.3 Generalized least squares

It is clear from the last two subsections that the approaches to estimation with heteroskedastic and AR(1) errors are quite similar. Those of you who like a general summary should read the following. I will not bother you about this section on the exam.

In general, note that GLS can be used if we replace Assumption 2[†] by

Assumption 2[§]. (general error structure) $\mathbb{E}(\mathbf{U}\mathbf{U}'|\mathbf{X}) = \Sigma = \sigma^2\Omega$, for some $\sigma > 0$,

where Ω is an appropriate $(n \times n)$-matrix.[21] This includes the cases of homoskedastic, heteroskedastic, and $AR(1)$ errors that we have studied so far. Note that Ω is not a correlation matrix in the case of heteroskedastic errors.

The OLS estimator for this model is still linear and unbiased, provided we maintain the other classical assumptions. Its variance-covariance matrix has changed, however, along the lines discussed earlier. Also, the Gauss-Markov theorem fails in this general case, and the OLS estimator is not (relatively) efficient.

The GLS estimator $\tilde{\beta}$ is the vector b that minimizes the weighted sum of squared residuals

$$(\mathbf{Y}-\mathbf{x}b)'\,\Omega^{-1}\,(\mathbf{Y}-\mathbf{x}b)\,.$$

The first order conditions or normal equations for this problem are given by

$$\mathbf{x}'\Omega^{-1}\tilde{\mathbf{U}}=\mathbf{x}'\Omega^{-1}\left(\mathbf{Y}-\mathbf{x}\tilde{\beta}\right)=0,$$

where $\tilde{\mathbf{U}}=\mathbf{Y}-\mathbf{x}\tilde{\beta}$ is the $(n\times 1)$-vector of GLS residuals. Solving for $\tilde{\beta}$ gives

$$\tilde{\beta}=\left(\mathbf{x}'\Omega^{-1}\mathbf{x}\right)^{-1}\mathbf{x}'\Omega^{-1}\mathbf{Y}.$$

As we have seen in the special cases of heteroskedastic and $AR(1)$ errors, we can alternatively view this estimator as an OLS estimator of an appropriately transformed model. Note (again) that we could have replaced Ω^{-1} by Σ^{-1} throughout, without changing $\tilde{\beta}$ and any of the results that follow.

The main result is an extension of the Gauss-Markov theorem,

Proposition 6. *Under Assumptions $1^\dagger$, $2^\S$, $3^\dagger$ and $4^\dagger$, the GLS estimator $\tilde{\beta}$ is the* best linear unbiased estimator *(BLUE).*

This follows directly from the Gauss-Markov theorem, and the fact that the GLS estimator is the OLS estimator of an appropriately transformed model.

The variance-covariance matrix of the GLS estimator is

$$V(\tilde{\beta})=\left(\mathbf{x}'\Sigma^{-1}\mathbf{x}\right)^{-1}.$$

Typically, we only know Ω up to some unknown parameters. For example, in the heteroskedasticity case, these are the parameters that link the variances to the regressors. In the case of $AR(1)$ errors, this is the autocorrelation parameter ρ. The (two-step) feasible GLS estimator first estimates these parameters of Σ from OLS residuals, and then evaluates the GLS estimator at the estimated value of Ω. So, feasible GLS consists of the following steps.

(i). Estimate the model by OLS, and compute the OLS residuals;

(ii). Estimate the unknown parameters of Ω using the OLS residuals, and construct an estimator $\hat{\Omega}$ of Ω;

(iii). Estimate β by $\hat{\tilde{\beta}} = \left(\mathbf{x}'\hat{\Omega}^{-1}\mathbf{x}\right)^{-1}\mathbf{x}'\hat{\Omega}^{-1}\mathbf{Y}$.

(iv). Estimate σ^2 by $\hat{\tilde{\sigma}}^2$ from the GLS residuals and estimate $V(\hat{\tilde{\beta}})$ by $\hat{V}(\hat{\tilde{\beta}}) = \left(\mathbf{x}'(\hat{\tilde{\sigma}}^2\hat{\Omega})^{-1}\mathbf{x}\right)^{-1}$.

The statistical properties of feasible GLS estimators are typically studied using so called asymptotic theory, *i.e.* in terms of approximations that hold if the sample size is sufficiently large. This is well beyond the scope of this course.

Notes

[1] In the simple example below, in which the sample space consists of a finite number (2) of sample points, we simply take the set of events $\mathcal{F}$ to be all (4) subsets of Ω. In general, there is some freedom in the choice of the set of events in the model. Probability theory does however require that $\mathcal{F}$ satisfies certain properties. Formally, it is required that $\mathcal{F}$ is a so called *σ-algebra*. This, for example, requires that $E^c \in \mathcal{F}$ if $E \in \mathcal{F}$, where E^c (or Ω/E) is the set of all point in Ω not in F. This is natural, as it requires that to each event "the outcome of the experiment is a point in E" corresponds the complementary event "the outcome of the experiment is not a point in E". Note that the simple example above satisfies this requirement. We will not further discuss these requirements, and simply assume that they are satisfied by $\mathcal{F}$.

[2] That $\bigcup_{i=1}^{\infty} E_i$ is an event in $\mathcal{F}$ if $E_1, E_2, \ldots \in \mathcal{F}$ is guaranteed by the requirement that $\mathcal{F}$ is a σ-algebra. See note 1.

[3] There is an additional requirement on a random variable that we will ignore here. In Example 12, each "event" defined in terms of X, *i.e.* $\{\omega : X(\omega) = 1\}$, $\{\omega : X(\omega) = 0\}$, $\{\omega : X(\omega) \in \{0, 1\}\}$ and $\{\omega : X(\omega) \in \emptyset\}$, is an event in $\mathcal{F}$: $\{4, 5, 6\}$, $\{1, 2, 3\}$, Ω and $\emptyset$, respectively. So, to each statement in terms of X we can assign a probability using our original probability model. In general, not each function $X : \Omega \to \mathbb{R}$ has the property that $\{\omega : X(\omega) \in B\} \subset \mathcal{F}$ if $B \subset \mathbb{R}$ (or, actually, if B is some set in an appropriate class of subsets of $\mathbb{R}$, the so called *Borel-σ-algebra*). So, we need an additional requirement on random variables, called *measurability*. This is way beyond the scope of this review, and we will not worry about this problem in the sequel.

[4] Note that a random variable could be neither discrete nor continuous.

[5] If f_X is continuous, we also have that $f_X(x) = dF_X(x)/dx$ for all x. In general case of absolutely continuous F_X, we can only say that $f_X(x) = dF_X(x)/dx$ *almost everywhere*.

[6] In the definition in terms of the c.d.f., each pair (x, y) corresponds to events $\{\omega : X(\omega) \leq x\}$ and $\{\omega : Y(\omega) \leq y\}$. According to Definition 4, $\{\omega : X(\omega) \leq x\}$ and $\{\omega : Y(\omega) \leq y\}$ are independent if $F_{X,Y}(x, y) = P(\{\omega : X(\omega) \leq x\} \cap \{\omega : Y(\omega) \leq y\}) = P(\{\omega : X(\omega) \leq x\})P(\{\omega : Y(\omega) \leq y\}) = F_X(x)F_Y(y)$. So, independence of the random

variables X and Y simply requires that all pairs of events $\{\omega : X(\omega) \leq x\}$ for $x \in \mathbb{R}$ and $\{\omega : Y(\omega) \leq y\}$ for $y \in \mathbb{R}$ are independent.

[7] Loosely, one argument is the following. For both the discrete and continuous cases, the conditional probability of $\{X \leq x\}$ given $\{y' < Y \leq y\}$ is

$$\begin{aligned} P(X \leq x|y' < Y \leq y) &= P(\{\omega : X(\omega) \leq x\}|\{\omega : y' < Y(\omega) \leq y\}) \\ &= \frac{F_{X,Y}(x,y) - F_{X,Y}(x,y')}{F_Y(y) - F_Y(y')}, \end{aligned} \tag{39}$$

provided that $P(y' < Y \leq y) > 0$ (this requires at the very least that $y' < y$).

If X and Y are discrete and y is such that $p_Y(y) > 0$, we can let $y' \uparrow y$ in equation (39), giving

$$\lim_{y' \uparrow y} P(X \leq x|y' < Y \leq y) = P(X \leq x|Y = y),$$

which fits Definition 3 of conditional probability as $p_Y(y) > 0$. $P(X \leq x|Y = y)$ is the conditional c.d.f. corresponding to the conditional p.m.f. defined in the text.

If X and Y are continuous, a limit argument just like the argument for the discrete case can be used to derive a conditional distribution for the continuous case that makes sense.

[8] Recall that $X : \Omega \to \mathbb{R}$, so that $g \circ X : \Omega \to \mathbb{R}$ is a function from the sample space to the real numbers as well. Just like the function X, this new function $g \circ X$ has to satisfy measurability conditions to ensure that probability statements in terms of $g \circ X$ can be restated in terms of events in the underlying probability space (see note 3). Measurability of $g \circ X$ requires that g is a measurable function as well. In the sequel, it is silently understood that any functions of random variables we discuss should be measurable.

[9] Note that conditional probabilities are conditional expectations of *indicator functions*. For example, $P(Y \leq y|X = x) = \mathbb{E}[I_{(-\infty,y]}(Y)|X = x]$, with $I_{(-\infty,y]}(u) = 1$ if $u \leq y$ and $I_{(-\infty,y]}(u) = 0$ if $u > y$. Again, if we evaluate $P(Y \leq y|X = x)$ at X, we get a random variable $P(Y \leq y|X)$. The law of the iterated expectations can be applied, and gives $\mathbb{E}[P(Y \leq y|X)] = P(Y \leq y)$. It is easy to check that this is consistent with our definitions of conditional distributions and expectations in Subsections 2.3.4 and 2.3.5.

[10] For those that are curious: the proof proceeds as follows. It is convenient to first condition on X and then apply the law of the iterated expectations. So, note that

$$\begin{aligned}\mathbb{E}\left[(Y-h(X))^2|X\right] &= \mathbb{E}\left[(Y-\mathbb{E}(Y|X)+\mathbb{E}(Y|X)-h(X))^2|X\right]\\ &= \mathbb{E}\left[(Y-\mathbb{E}(Y|X))^2|X\right]+\mathbb{E}\left[(\mathbb{E}(Y|X)-h(X))^2|X\right]\\ &\quad +2\mathbb{E}\left[(Y-\mathbb{E}(Y|X))(\mathbb{E}(Y|X)-h(X))|X\right].\end{aligned}\tag{40}$$

Now, for given X, $\mathbb{E}(Y|X)-h(X)$ is a given number, so that $\mathbb{E}(Y|X)-h(X)$ and $Y-\mathbb{E}(Y|X)$ are independent. So, the expectation in third term in the right-hand side of equation (40) reduces to

$$\begin{aligned}\mathbb{E}\left[(Y-\mathbb{E}(Y|X))(\mathbb{E}(Y|X)-h(X))|X\right] &= \mathbb{E}[Y-\mathbb{E}(Y|X)|X][\mathbb{E}(Y|X)-h(X)]\\ &= [\mathbb{E}(Y|X)-\mathbb{E}(Y|X)][\mathbb{E}(Y|X)-h(X)]=0.\end{aligned}$$

Therefore, equation (40) reduces to

$$\begin{aligned}\mathbb{E}\left[(Y-h(X))^2|X\right] &= \mathbb{E}\left[(Y-\mathbb{E}(Y|X))^2|X\right]+\mathbb{E}\left[(\mathbb{E}(Y|X)-h(X))^2|X\right]\\ &\geq \mathbb{E}\left[(Y-\mathbb{E}(Y|X))^2|X\right],\end{aligned}$$

for all measurable functions h. Applying the law of the iterated expectations shows that

$$\begin{aligned}\mathbb{E}\left[(Y-h(X))^2\right] &= \mathbb{E}\left[\mathbb{E}\left[(Y-h(X))^2|X\right]\right]\\ &\geq \mathbb{E}\left[\mathbb{E}\left[(Y-\mathbb{E}(Y|X))^2|X\right]\right]=\mathbb{E}\left[(Y-\mathbb{E}(Y|X))^2\right]\end{aligned}$$

for all measurable functions h.

[11] The proof of this result is somewhat involved for a review at this level, but we can give some intuition based on the discussion in Subsection 2.3.7. Note that we can write

$$\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}=\frac{\dfrac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S_n^2/\sigma^2}{n-1}}}.\tag{41}$$

We already know that the enumerator of this ratio is a standard normal random variable, say Z. Furthermore, note that $(n-1)S_n^2/\sigma^2$ is a quadratic expression in standard normal random variables. It can be shown that it is distributed as a χ^2_{n-1} random variable.

Finally, it can be shown that these Z and χ^2_{n-1} are independent. From Subsection 2.3.7, we know that this implies that the ratio in equation (41) is a t-ratio $Z/\sqrt{\chi^2_{n-1}/(n-1)}$ with $n-1$ degrees of freedom.

[12] Formally, it is not correct to use the term "lowest" here. To be precise, the p-value is $\inf\{\alpha : t \in \Gamma_\alpha\}$. See also the example. If we use $t = T$, the p-value is again a random variable.

[13] The probability of $Z_0 \in (-\infty, -n_{1-\alpha/2}) \cup (n_{1-\alpha/2}, \infty)$ (rejecting H_0) equals the probability of $Z_\mu \in (-\infty, -n_{1-\alpha/2} - \sqrt{n}\mu/\sigma) \cup (n_{1-\alpha/2} - \sqrt{n}\mu/\sigma, \infty)$. So, the power function $\pi_2(\mu)$ corresponding to this two sided test is

$$\pi_2(\mu) = \Phi\left(-n_{1-\alpha/2} - \sqrt{n}\mu/\sigma\right) + 1 - \Phi\left(n_{1-\alpha/2} - \sqrt{n}\mu/\sigma\right).$$

We should now evaluate the power function for both $\mu < 0$ and $\mu > 0$ to assess the power. $\pi_2(\mu)$ is decreasing on $(-\infty, 0)$ and increasing on $(0, \infty)$. As again $\pi_2(0) = \alpha$, this implies that $\pi_2(\mu) > \alpha$ for all $\mu \neq 0$. Furthermore, $\pi_2(\mu) \to 1$ as either $\mu \to \infty$ or $\mu \to -\infty$. Finally, if the sample size n grows large, $\pi_2(\mu)$ is close to 1 for most values of μ.

[14] With random sampling we would actually have that the (Y_i, X_i) are independent between observations. In this case, conditioning on all regressors $\mathbf{X}$ instead of only the relevant regressor X_i would be pointless. However, Assumptions 1 and 2 only require that the mean and variance of U_i for given X_i do not depend on X_j, for $j \neq i$. Assumption 2 also requires (conditionally) uncorrelated errors. Various properties of least squares estimators can be derived without stronger independence assumptions.

[15] Using that $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$ and $\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{x}$, R can be rewritten as

$$\begin{aligned} R &= \frac{\sum_{i=1}^n (Y_i - \bar{Y})(\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}\bar{x})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^n (\hat{\alpha} + \hat{\beta}x_i - \hat{\alpha} - \hat{\beta}\bar{x})^2}} \\ &= \frac{\hat{\beta}}{\sqrt{\hat{\beta}^2}} \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{\hat{\beta}}{\sqrt{\hat{\beta}^2}} \hat{\rho}_{X,Y} = |\hat{\rho}_{X,Y}|, \end{aligned}$$

where $|\hat{\rho}_{X,Y}|$ is the absolute value of the sample correlation between X and Y. The last equation uses that $\sqrt{\hat\beta^2} = |\hat\beta|$, so that

$$\frac{\hat\beta}{\sqrt{\hat\beta^2}} = \frac{\hat\beta}{|\hat\beta|} = \frac{\hat\beta\dfrac{s_X}{S_Y}}{|\hat\beta|\dfrac{s_X}{S_Y}} = \frac{\dfrac{S_{X,Y}}{s_X^2}\dfrac{s_X}{S_Y}}{\dfrac{|S_{X,Y}|}{s_X^2}\dfrac{s_X}{S_Y}} = \frac{\dfrac{S_{X,Y}}{s_X S_Y}}{\left|\dfrac{S_{X,Y}}{s_X S_Y}\right|} = \frac{\hat\rho_{X,Y}}{|\hat\rho_{X,Y}|}$$

and $R = \hat\rho^2_{X,Y}/|\hat\rho_{X,Y}| = |\hat\rho_{X,Y}|$.

[16] This is closely related to the discussion in note 11. For example, note that

$$T^*_\beta = \frac{\hat\beta - \beta}{\sqrt{\hat\sigma^2/\sum_{i=1}^n (x_i - \bar x_n)^2}} = \frac{\dfrac{\hat\beta-\beta}{\sqrt{\mathrm{var}(\hat\beta)}}}{\dfrac{\sqrt{(n-2)\hat\sigma^2/\sigma^2}}{n-2}} = \frac{Z^*}{\sqrt{\dfrac{(n-2)\hat\sigma^2/\sigma^2}{n-2}}}.$$

We already know that the Z-statistic in the enumerator is a standard normal random variable. Furthermore, $(n-2)\hat\sigma^2/\sigma^2$ is a quadratic expression in standard normal random variables, and can be shown to be χ^2_{n-2}-distributed. Finally, it can be shown that the enumerator and denominator are independent. From Subsection 2.3.7, we know that this implies that the ratio in equation (41) is a t-ratio with $n-2$ degrees of freedom.

[17]The intuition for this is similar to the intuition for the simple regression case provided earlier in note 16.

[18]Some intuition for this can be derived from an analysis similar to that for the t-statistic in notes 11 and 16. Note that

$$F = \frac{\left[\dfrac{\hat\beta_1^2}{\mathrm{var}(\hat\beta_1)} + \dfrac{\hat\beta_2^2}{\mathrm{var}(\hat\beta_2)}\right]\Big/2}{\dfrac{(n-3)\hat\sigma^2}{\sigma^2}\Big/(n-3)}.$$

The enumerator is the χ^2-statistic derived before, divided by its degrees of freedom, 2. Also, $(n-3)\hat\sigma^2/\sigma^2$ has a χ^2-distribution with $n-3$ degrees of freedom, and can be shown to be independent from the χ^2-statistic in the enumerator. So, under H_0, F is simply the ratio of two independent χ^2-statistics divided by their degrees of freedom. In Subsection 2.3.7, we have seen that such a statistic has an F-distribution.

[19] This follows from the stationarity of the $AR(1)$-proces.

[20] This uses that $V(\mathbf{U}^*|\mathbf{X}) = V(Q\mathbf{U}) = Q\Sigma Q' = \sigma_v^2 Q(Q'Q)^{-1}Q' = \sigma_v^2 QQ^{-1}Q'^{-1}Q' = \sigma_v^2 I_n$.

[21] In particular, Ω should be symmetric and positive definite.

References

Gujarati, D.N. (1995), *Basic Econometrics*, third edition, McGraw-Hill, New York.

Ross, S. (1998), *A First Course in Probability*, fifth edition, Prentice Hall, Upper Saddle River, N.J.

Wonnacott, T.H. and R.J. Wonnacott (1990), *Introductory Statistics for Business and Economics*, fourth edition, Wiley, New York.