Handout 5
Bivariate Regression *

Instructor: Vira Semenova Note author: Danylo Tavrov, Vira Semenova

1 Linear Regression Basics

1.1 Motivation

As econometricians, we are interested in recovering (causal) relationships between variables: if one variable is changed by this many units, how will another variable change?

Example. Suppose we are interested in the following question: if an arbitrarily chosen person is given an extra year of education, by how much would their hourly wage increase? This is a *ceteris paribus* question: we are interested in the effect of an additional year of schooling, *keeping all other (relevant) factors fixed*. In other words, we want to look at the impact of education assuming that other individual characteristics such as family background, "innate ability," etc. remain unchanged.

The simplest model to describe the parameters that shows the magnitude of this impact is the *bivariate linear model*.

1.2 Linear Function

A linear function

$$f(x) = \beta_0 + \beta_1 \cdot x$$

is determined by two parameters, an *intercept* β_0 and a *slope* β_1. The *intercept* β_0 shows the function value at $x = 0$:

$$\beta_0 = \beta_0 + \beta_1 \cdot 0 = f(0) \, .$$

The *slope* β_1 shows the change in the outcome due to the change in x from x_0 to x_1[1]

$$\beta_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0} \, , \quad x_1 \neq x_0 \, .$$

1.3 Linear Regression Model

The expression

$$Y = \beta_0 + \beta_1 X + u \, , \tag{1.1}$$

is called a *linear regression model*, where:

- Y is the *outcome*, or *dependent variable*;
- X is the *covariate*, *regressor*, or *input variable*;
- u is the *residual*, *error*, or *disturbance*.

*We thank Prof. Anna Mikusheva and numerous former ECON 140 and 141 GSIs for their course notes.

[1] Alternatively, β_1 can be treated as *derivative* of $f(x)$ w.r.t. its argument x.

Equation (1.1) is also called a *data generating process* (DGP) [2], as it shows how the data, in particular the outcome variable, are generated.

The error term u rationalizes why the relationship between X and Y is not perfect, or why the same values of X result in different values of Y. The error term is interpreted differently in hard and social sciences:

- in hard sciences, u stands for the measurement error of a device in a laboratory setting;
- in social sciences, u stands for all unexplained factors affecting Y other than X.

We always tacitly assume that $\mathbb{E}[u_i] = 0$. Indeed, suppose that $\mathbb{E}[u_i] = c \neq 0$ with some constant c. This model is *observationally equivalent* to a model with intercept $\beta_0 + c$ and $\mathbb{E}[u_i] = 0$, in the sense that it is impossible to distinguish between the two models using the data. Thus, we always assume that $c = 0$.

Most economic relationships are not exact, and hence the actual values of Y do not lie on a straight line. The difference between the true linear relationship and the observed relationship between X and Y is illustrated in Fig. 1.1.

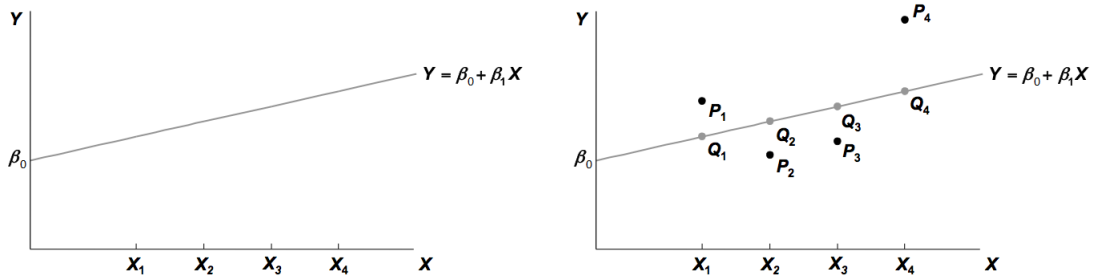


Figure 1.1: Linear relationship between Y and X: (a) original DGP; (b) exact (Q's) and not exact (P's) relationship

Each value of Y thus has a *non-random* component[3], $\beta_0 + \beta_1 X$, and a *random* component, u. In Fig. 1.2, the first observation has been decomposed into these two components.

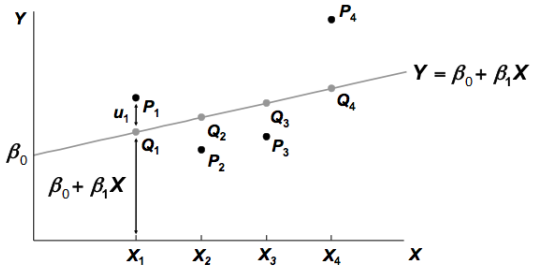


Figure 1.2: Decomposition of an observation P_1 into non-random, $\beta_0 + \beta_1 X_1$, and random, u_1, components

1.4 Regression Coefficients and Population Moments

Suppose that the conditional expectation of the random variable Y is *assumed to be* a linear function of X:

$$\mathbb{E}[Y \mid X = x] = \beta_0 + \beta_1 x.$$

[2] Strictly speaking, to complete the description of the DGP, we also need information about the (joint) probability distribution of all the variables involved. But in practice we are not interested in those, and our main interest is the slope coefficient in a linear model.

[3] In the sense that if X is held fixed, Y can be determined using its value and values of β_0 and β_1.

One immediate consequence is that in (1.1), the error term u obeys the *conditional mean restriction*:

$$\mathbb{E}[u \mid X] = 0 \,. \tag{1.2}$$

Taking expectation of Y, we get:

$$\beta_0 = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] - \mathbb{E}[u] = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[u \mid X]] = \mathbb{E}[Y] - \beta_1 \mathbb{E}[X] \;, \tag{1.3}$$

where we have used the law of iterated expectations.
Taking covariance of X and Y, we get:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \beta_0 + \beta_1 X + u) = 0 + \beta_1 \mathrm{Cov}(X, X) + \mathrm{Cov}(X, u) = \beta_1 \mathrm{Var}(X) \;,$$

where

$$\mathrm{Cov}(X, u) = \mathbb{E}[Xu] - \mathbb{E}[X]\,\mathbb{E}[u] = \mathbb{E}[\mathbb{E}[Xu \mid X]] - 0 = \mathbb{E}[X \cdot \mathbb{E}[u \mid X]] = 0 \,,$$

where we have used the law of iterated expectations. Therefore, the slope parameter is

$$\beta_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} \equiv \frac{\sigma_{XY}}{\sigma_X^2} \,. \tag{1.4}$$

Multiplying and dividing this expression by the standard deviation σ_Y gives

$$\beta_1 = \frac{\sigma_{XY}}{\sigma_X^2} \cdot \frac{\sigma_Y}{\sigma_Y} = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \;,$$

where ρ_{XY} is the correlation between X and Y.

1.5 Interpretation of Regression Equation

Continuous covariate. To interpret coefficients β_0 and β_1 in (1.1), consider the following example.

Example. Assume that average students' test scores in a school district and student-to-teacher ratio (STR, roughly class size) in that district are related in the following linear way:

$$\text{test score} = \beta_0 + \beta_1 \cdot \text{STR} + u \,.$$

In words, average test scores are assumed to depend linearly on the STR and other factors that are gathered in the error term u. The scalars β_0 and β_1 are the parameters of the model. The first parameter shows the level of test scores when there are no students in a district. Ceteris paribus, i.e. when the change in u is zero (equivalently, when $\mathbb{E}[u \mid \text{STR}] = 0$), the second parameter measures the strength and direction of the relationship between (expected, average over all districts) test scores and STR:

$$\frac{\partial \mathbb{E}[\text{test score} \mid \text{STR}]}{\partial \text{STR}} = \beta_1 \,.$$

The parameter β_1 can thus be interpreted as the *causal effect* of an additional student in a class on the (expected, average) hourly wage rate.

Similar questions arise in many fields in economics. Knowing the returns to education (effect of additional year of education on one's earnings) is an important question in labor economics. In gender and racial studies, figuring out the extent of gender and racial discrimination is key. In macroeconomics, we want to figure out the causal impact of macroeconomic policies (monetary, fiscal, etc.).

The above discussion assumes that we can perform a ceteris paribus analysis, in particular, that we can fix u when assessing the effect of changing X on Y, or, equivalently, that $\mathbb{E}[u \mid X] = 0$. If $\mathbb{E}[u \mid X] \neq 0$, then $\mathbb{E}[Y \mid X]$ is not a linear of function of X, and thus change in X will not be described only by β_1.

Binary covariate. Suppose STR is a binary variable (*dummy* variable) indicating whether the STR is high: STR $= 1$ if the classes are large, and 0 if they are small. In this case,

$$\begin{aligned}\mathbb{E}\left[\text{test score} \mid \text{STR} = 0\right] &= \beta_0 \ , \\ \mathbb{E}\left[\text{test score} \mid \text{STR} = 1\right] &= \beta_0 + \beta_1 \ .\end{aligned}$$

Therefore, β_0 is the (expected) test score for small STR, and β_1 is the (expected) difference between test scores in large and small classes.

In the case of a dummy variable, assumption (1.2) no longer acts as a linearity restriction, because for each pair $(\mathbb{E}\left[\text{test score} \mid \text{STR} = 0\right], \mathbb{E}\left[\text{test score} \mid \text{STR} = 1\right])$, it is always possible to find two numbers β_0 and β_1 such that the relationship above holds.

Categorical covariate. Suppose that we are interested in establishing a relationship between test scores and the state the school district is located in. In this case, our model would look like

$$\text{test score} = \beta_0 + \beta_1 \cdot \text{state} + u \ ,$$

where state $= 1$ if someone lives in Alabama, state $= 2$ if someone lives in Alaska, and so on. In this case, β_1 would be interpreted as the (average) difference in test scores between any two states encoded consecutively:

$$\begin{aligned}\beta_1 &= \mathbb{E}\left[\text{test score} \mid \text{state} = AK\right] - \mathbb{E}\left[\text{test score} \mid \text{state} = AL\right] \\ \beta_1 &= \mathbb{E}\left[\text{test score} \mid \text{state} = AZ\right] - \mathbb{E}\left[\text{test score} \mid \text{state} = AK\right] \\ &\vdots\end{aligned}$$

This makes no sense on two grounds:

- it is implausible that the difference in test scores between the states is the same regardless of what states are being compared;
- more importantly, it makes no sense to order the states alphabetically and then manipulate them as if they were points on a numerical scale.

This issue arises because state is a *categorical variable*, i.e. it assumes a finite number of values that have *no order*. A more correct approach to analyzing relationships between categorical variables and the output variable is to include separate binary variables for each category as follows:

$$\text{test score} = \beta_1 \text{Alabama} + \beta_2 \text{Alaska} + \ldots + u \ .$$

However, this is already not a bivariate regression, and we will return to discussing such models later on in this course.

2 Estimation of Regression Coefficients

2.1 Sample Analogues of Population Parameters

In practice, we typically observe not the whole population, but a sample $(X_i, Y_i)_{i=1}^n$ of n pairs of X_i and Y_i, which are assumed to be *independent and identically distributed* (i.i.d.), meaning that:

- (X_i, Y_i) is independent of (X_j, Y_j), $i \neq j$;
- each (X_i, Y_i) obeys the same distribution described by a joint CDF $F_{X,Y}$, although we don't know what it is (and don't care).

Using the sample at hand, we can *estimate* the population parameters in the linear model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \,, \quad i = 1, \ldots, n \,,$$

as the *sample analogues* of the corresponding population expressions (1.3) and (1.4):

$$\begin{aligned} \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \,, \\ \hat{\beta}_1 &= \frac{\widehat{\mathrm{Cov}}\,(X,Y)}{\widehat{\mathrm{Var}}\,(X)} \,, \end{aligned} \tag{2.1}$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean of X_i (likewise for $\overline{Y}$), $\widehat{\mathrm{Var}}\,(X) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is the sample variance of X_i[4], and $\widehat{\mathrm{Cov}}\,(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$.

Note that while the estimands β_0 and β_1 are fixed values, the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are *random variables* because their values depend on the sample we have (in other words, the estimators are functions of data).

2.2 The Ordinary Least Squares

We can arrive at the same result as in (2.1) from a different perspective. Suppose we have a sample (the P points in Fig. 2.1), and seek to draw a straight line that would be the "best" approximation to the true line in our DGP:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \,. \tag{2.2}$$

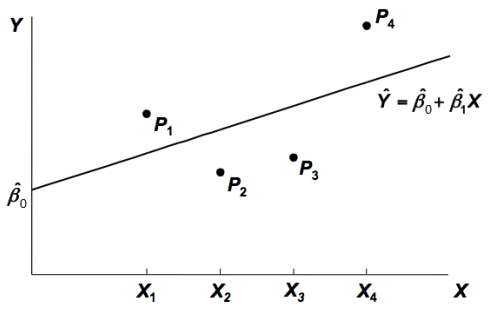


Figure 2.1: Line fitted through the observations generated by the DGP

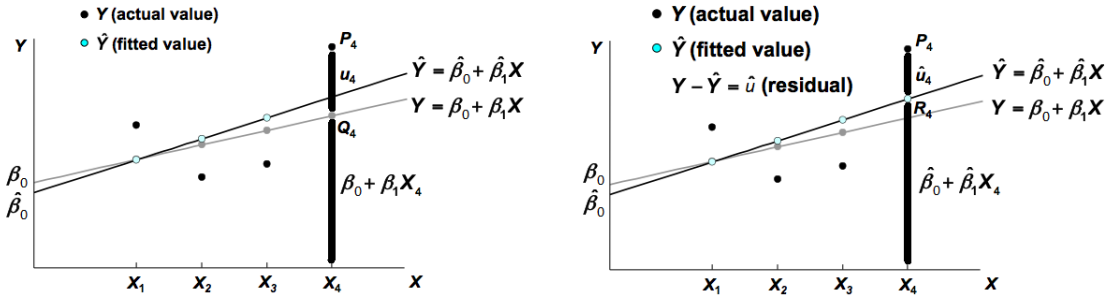
Equation (2.2) is called the *fitted model*, and the values of $\hat{Y}$ predicted using (2.2) are called the *fitted values* of Y. The discrepancies between the actual and fitted values of Y are known as *estimated residuals*, $\hat{u} \equiv Y - \hat{Y}$.

If the fit of the model is a good one, the residuals $\hat{u}$ and the values of the disturbance term u will be sufficiently similar. However, these two concepts are different. To illustrate this, consider two conceptually different ways to decompose values of Y:

- Y can be decomposed into its nonstochastic component $\beta_0 + \beta_1 X$ and its random component u (Fig. 2.2(a)). This is a *theoretical* ("ex ante") decomposition, because we do not know neither the values of β_0 or β_1, nor the values of the disturbance term u;
- Y can be decomposed with reference to the fitted line, using the fact that the actual value of $Y = \hat{Y} + \hat{u}$ (Fig. 2.2(b)). This is an *operational* ("ex post") decomposition.

Thus, our task is to find $\hat{\beta}_0$ and $\hat{\beta}_1$ in (2.2), such that the fitted line is the line that is going to *minimize* the overall *distance* between the actual and the predicted values. One of the ways to interpret this is to minimize the sum of

[4] Note that for the OLS estimator to be well-defined, the sample variance cannot be zero, in other words, X_i cannot be constant, there has to be variation in X_i.



(a) (b)

Figure 2.2: Decomposition of Y: (a) ex ante; (b) ex post

squared vertical distances (*sum of squared residuals*, SSR) between data points and the equation line—hence the name *ordinary least squares* (OLS)[5].

Formally, we have the following minimization problem:

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{b_0,b_1} \text{SSR}(b_0, b_1) = \arg\min_{b_0,b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 = \arg\min_{b_0,b_1} \sum_{i=1}^{n} \hat{u}_i^2 \ .$$

The first order conditions (FOC) for this problem are as follows:

- for b_0:

$$\frac{\partial \text{SSR}}{\partial b_0} = 2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) \cdot (-1) \ .$$

  Setting this expression equal to 0, we get:

$$\begin{aligned} -2 \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) &= 0 \\ \overline{Y} - b_0 - b_1 \overline{X} &= 0 \\ b_0 &= \overline{Y} - b_1 \overline{X} \ ; \end{aligned}$$

- for b_1:

$$\frac{\partial \text{SSR}}{\partial b_1} = 2 \sum_{i=1}^{n} ((Y_i - b_0 - b_1 X_i) \cdot (-X_i)) \ .$$

  Setting this expression equal to 0, we get:

$$\begin{aligned} -2 \sum_{i=1}^{n} ((Y_i - b_0 - b_1 X_i) \cdot X_i) &= 0 \\ \overline{XY} - b_0 \overline{X} - b_1 \overline{X^2} &= 0 \\ \overline{XY} - (\overline{Y} - b_1 \overline{X}) \overline{X} - b_1 \overline{X^2} &= 0 \\ b_1 &= \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} \ . \end{aligned}$$

[5] There are other criteria that can be used to draw the fitted line, e.g. find such $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of *absolute deviations*, $\sum_{i=1}^{n} \left| Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right|$, which leads to the *least absolute deviations* (LAD) estimator. It is more robust to outliers in the data but its analysis is more complicated.

Thus, we obtain[6]:

$$\begin{aligned} \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} \,, \\ \hat{\beta}_1 &= \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} \,. \end{aligned} \tag{2.3}$$

The first order conditions given above allow us to recognize the following facts:

- recalling the definition of residuals, it is evident that

$$\frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \right) = \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i = 0 \,, \tag{2.4}$$

  meaning that the *sample average* of residuals must be 0. In other words, the average deviation from true values Y_i to predicted values $\hat{Y}_i$ is 0;

- it is also evident that

$$\frac{1}{n} \sum_{i=1}^{n} \left((Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) \cdot X_i \right) = \frac{1}{n} \sum_{i=1}^{n} X_i \hat{u}_i = 0 \,, \tag{2.5}$$

  meaning that the *sample covariance* between the residuals and X_i must be 0. In other words, X_i cannot be used to predict the residuals;

- $\hat{\beta}_1$ is the *sample covariance* of X_i and Y_i divided by the *sample variance* of X_i. First, note that

$$\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) = \frac{1}{n} \sum_{i=1}^{n} X_i - \overline{X} = 0 \,,$$

  and therefore

$$\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right) = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) Y_i - \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) \cdot \overline{Y} = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) Y_i \,.$$

  Finally,

$$\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right) = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) Y_i = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i - \frac{1}{n} \sum_{i=1}^{n} X_i \cdot \overline{Y} = \overline{XY} - \overline{X} \cdot \overline{Y} \,.$$

  Also, note that

$$\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{2}{n} \sum_{i=1}^{n} X_i \overline{X} + \left(\overline{X} \right)^2 = \overline{X^2} - (\overline{X})^2 \,.$$

  Thus, we have the slope estimate

$$\begin{aligned} \hat{\beta}_1 &= \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - \left(\overline{X} \right)^2} \\ &= \frac{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right)}{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2} \\ &= \frac{\widehat{\mathrm{Cov}} \left(X_i, Y_i \right)}{\widehat{\mathrm{Var}} \left(X_i \right)} \,, \end{aligned}$$

  which is exactly the same as (2.1).

[6] To make sure that these values are indeed minimizers, we need to check the *second-order* condition: $\left(\begin{array}{cc} \frac{\partial \mathrm{SSR}}{\partial b_0^2} & \frac{\partial \mathrm{SSR}}{\partial b_0 \partial b_1} \\ \frac{\partial \mathrm{SSR}}{\partial b_0 \partial b_1} & \frac{\partial \mathrm{SSR}}{\partial b_1^2} \end{array} \right) \Big|_{(b_0 = \hat{\beta}_1, b_1 = \hat{\beta}_1)}$ is positive definite. This is left as an exercise for the reader.

2.3 Difference Between Linear Regression Model and Projection

In Sect. 1.5, we showed that causal interpretation of the regression model coefficients hinges on assumption (1.2). Using this assumption, we showed that β_0 and β_1 have interpretation in terms of population moments given by (1.3)–(1.4). It is important to make distinction between regression model and *projection*.

When we want to *project* Y_i on a set of variables (a constant and X_i in our case), we seek to find such γ_0 and γ_1 that the average distance between Y_i and $\gamma_0 + \gamma_1 X_i$ is minimal[7]:

$$(\gamma_0, \gamma_1) = \arg\min_{c_0, c_1} \mathbb{E}\left[(Y_i - c_0 - c_1 X_i)^2\right] .$$

Reasoning similar to the one given above leads to the following solution of this problem:

$$\begin{aligned} \gamma_0 &= \mathbb{E}[Y_i] - \gamma_1 \mathbb{E}[X_i] , \\ \gamma_1 &= \frac{\operatorname{Cov}(X_i, Y_i)}{\operatorname{Var}(X_i)} . \end{aligned} \tag{2.6}$$

Then, we can always write

$$Y_i = \gamma_0 + \gamma_1 X_i + e_i ,$$

where, for the reasons analogous to (2.4) and (2.5), $\mathbb{E}[e_i] = 0$ and $\mathbb{E}[X_i e_i] = 0$. In this case, e_i is called the *projection error*.

Note the conceptual difference between (1.3)–(1.4) and (2.6). It is not true in general that $\mathbb{E}[e_i \mid X_i] = 0$, therefore coefficients in (2.6) have *no causal interpretation*, these are simply coefficients that correspond to the linear projection of Y_i on a constant and X_i. Crucially, even if in (1.1), it is not true that $\mathbb{E}[u \mid X] = 0$, then it is still possible to compute (2.6), but in that case, in general, $\beta_0 \neq \gamma_0$ and $\beta_1 \neq \gamma_1$. The equality will hold if assumption (1.2) is imposed, i.e. if the true model is indeed linear and thus coincides with the projection.

3 Properties of the OLS Estimator

3.1 Assumptions

Note that it is always possible to compute OLS coefficients for (1.1) (provided X_i is not constant) from a purely algebraic perspective—it is possible to draw a straight line through any dataset. To add content to the model, we need to introduce certain assumptions. The most important assumption was already introduced in (1.2), and we repeat it here for convenience.

Assumption 1. The conditional distribution of errors given regressors equals zero:

$$\mathbb{E}[u_i \mid X_i] = 0 .$$

Under this assumption, we assume that $\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i$, i.e. the conditional expectation of Y_i given X_i is a linear function of X_i. This assumption allows us to interpret the OLS coefficients in a way described in Sect. 1.5.

Note that the main idea of this assumption is to impose linear relationship between the conditional expectation and the regressor. If, for example, $\mathbb{E}[u_i \mid X_i] = c$, where $c \neq 0$ is some constant, then we would have $\mathbb{E}[Y_i \mid X_i] = \beta_0 + \beta_1 X_i + c$, i.e. the linear relationship will be preserved and coefficient β_1 will have the same causal interpretation. But in this case, the models with $\mathbb{E}[u_i \mid X_i] = c$ and $\mathbb{E}[u_i \mid X_i] = 0$ will be *observationally equivalent* because a priori it is impossible to tell whether the true model has intercept β_0 or $\beta_0 + c$. Therefore, we always use Assumption 1 in the form stated above and never consider the possibility $c \neq 0$.

Other assumptions are necessary to enable statistical inference (test appropriate statistical hypotheses).

Assumption 2. (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed.

Assumption 3. Large outliers are unlikely:

$$0 < \mathbb{E}\left[X_i^4\right] < \infty , \quad 0 < \mathbb{E}\left[Y_i^4\right] < \infty .$$

These assumptions allow us to apply large sample tools such as the (weak) law of large numbers (LLN) and the central limit theorem (CLT).

[7] Basically we are trying to minimize the population equivalent of SSR.

Homoskedasticity and heteroskedasticity. Another assumption that is sometimes imposed on the model is the assumption of homoskedasticity.

Assumption 4. The variance of the errors does not depend on the data:

$$\text{Var}\left(u_i \mid X_i\right) = \mathbb{E}\left[u_i^2 \mid X_i\right] = \sigma_u^2 \ .$$

In this case, errors are called *homoskedastic*. This assumption is useful to derive certain theoretical properties but it is of limited practical value and is taught out of historic stickiness, as in practice, virtually every possible model under analysis will have *heteroskedastic* errors (the ones that depend on the data). The following two examples provide some intuition.

Example 1. Consider a model for studying effect of education on wages, where only city average values for years of education, $\overline{\text{educ}}_c$, and income, $\overline{\text{wage}}_c$, are observed:

$$\overline{\text{wage}}_c = \beta_0 + \beta_1 \overline{\text{educ}}_c + u_c \ ,$$

where c is the index for cities. We know that

$$\overline{\text{educ}}_c = \frac{1}{M_c} \sum_i \text{educ}_{c,i}$$

and

$$\overline{\text{wage}}_c = \frac{1}{M_c} \sum_i \text{wage}_{c,i} \ ,$$

where i is the index for individuals and M_c is the population of city c. Suppose the error terms associated with each individual are homoskedastic, i.e. $\text{Var}\left(u_{c,i} \mid \text{educ}_{c,i}\right) = \sigma^2$. Then, the error terms of our (city-level) observations become heteroskedastic, because

$$\text{Var}\left(u_c \mid \text{educ}_c\right) = \text{Var}\left(\frac{1}{M_c} \sum_i u_{c,i} \mid \text{educ}_c\right) = \text{Var}\left(\frac{1}{M_c} \sum_i u_{c,i} \mid \text{educ}_{c,i}\right) = \frac{\sigma^2}{M_c} \ .$$

In simple words, variance of each (city-level) observation depends on the population of the associated city, M_c. Observations with larger population have smaller variance. This is a typical example of heteroskedasticity caused by *grouped data*.

Example 2. Consider a model where food_i is expenditure on food and inc_i is household income:

$$\text{food}_i = \beta_0 + \beta_1 \text{inc}_i + u_i \ .$$

Since food is a normal good, we would expect a positive relationship between income and food expenditure, so that, on average, higher income corresponds to higher food expenditure. In addition, we would expect the variation in food expenditure among high-income households to be higher than among low-income households: while low-income households spend the majority of their disposable income on food, high-income households can decide to allocate their income to other goods. This is illustrated in Fig. 3.1. Note that the variance of the dependent variable, food expenditure, increases with income.

We can find many similar examples in other fields: families with higher income have larger variance in vacation expenditures they make, sales of larger firms might be more volatile than sales of smaller firms, etc. In these cases, variance of the error term increases as the regressor becomes larger. This is a very common type of heteroskedasticity, but there exist other cases, e.g. when the variance of the error term decreases with the regressor, or the variance is larger for more extreme values of the regressor.

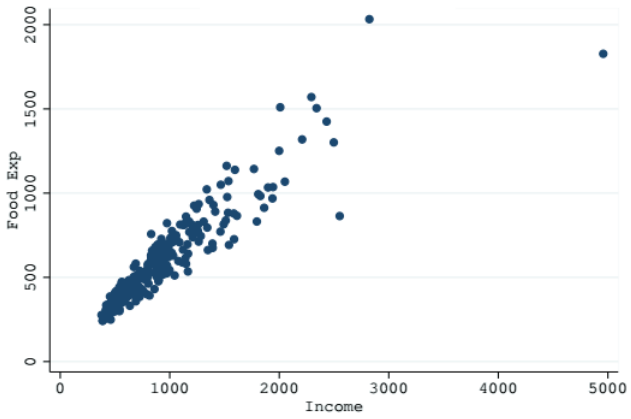


Figure 3.1: Illustrative data on food expenditure

3.2 Finite-Sample Properties

Linearity. The OLS estimator is linear in Y_i, i.e. it can be expressed as $\hat{\beta}_1 = \sum_{i=1}^n w_i(X_1, \ldots, X_n) Y_i$. Indeed, from (2.1), we have:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\widehat{\text{Cov}}\left(X_i, Y_i\right)}{\widehat{\text{Var}}\left(X_i\right)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&\equiv \sum_{i=1}^n w_i(X_1, \ldots, X_n) \,,
\end{aligned}$$

with

$$w_i(X_1, \ldots, X_n) = \frac{X_i - \overline{X}}{\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \,. \tag{3.1}$$

Unbiasedness. First, we show that the OLS estimator is *conditionally* unbiased, given X_i. Continue from the above expression and expand:

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right) Y_i}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)\left(\beta_0 + \beta_1 X_i + u_i\right)}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \beta_0 \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} + \beta_1 \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right) X_i}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} + \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right) u_i}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \\
&= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right) u_i}{\frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2} \,,
\end{aligned} \tag{3.2}$$

because $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)=0$ and $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)X_i=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(X_i-\overline{X}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$.
Then,

$$\begin{aligned}\mathbb{E}\left[\hat{\beta}_1 \mid X_1,\ldots,X_n\right] &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\mathbb{E}\left[Y_i \mid X_1,\ldots,X_n\right]}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\\ &= \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\mathbb{E}\left[Y_i \mid X_i\right]}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \quad \text{(by Assumption 2)}\\ &= \beta_1+\frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\mathbb{E}\left[u_i \mid X_i\right]}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}=\beta_1\ ,\end{aligned}$$

because $\mathbb{E}\left[u_i \mid X_i\right]=0$ by Assumption 1.
Then, by the law of iterated expectations,

$$\mathbb{E}\left[\hat{\beta}_1\right]=\mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_1 \mid X_1,\ldots,X_n\right]\right]=\beta_1\ ,$$

which shows that the OLS estimator is also *unconditionally* unbiased.
Using this result, we can show unbiasedness of $\hat{\beta}_0$:

$$\mathbb{E}\left[\hat{\beta}_0 \mid X_1,\ldots,X_n\right]=\mathbb{E}\left[\overline{Y}\right]-\mathbb{E}\left[\hat{\beta}_1\overline{X} \mid X_1,\ldots X_n\right]=\mathbb{E}\left[Y_i\right]-\mathbb{E}\left[\overline{X}\mathbb{E}\left[\hat{\beta}_1 \mid X_1,\ldots X_n\right]\right]=\mathbb{E}\left[Y_i\right]-\beta_1\mathbb{E}\left[X_i\right]=\beta_0\ ,$$

where we have used the fact the sample mean is an unbiased estimator of the population mean.
As a result,

$$\mathbb{E}\left[\hat{\beta}_0\right]=\mathbb{E}\left[\mathbb{E}\left[\hat{\beta}_0 \mid X_1,\ldots,X_n\right]\right]=\beta_0\ .$$

Conditional variance. Consider the conditional variance of the OLS estimator of the slope, $\hat{\beta}_1$:

$$\begin{aligned}\mathrm{Var}\left(\hat{\beta}_1 \mid X_1,\ldots,X_n\right) &= \mathrm{Var}\left(\frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)Y_i}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \mid X_1,\ldots,X_n\right)\\ &= \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)^2}\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)Y_i \mid X_1,\ldots,X_n\right)\\ &= \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\mathrm{Var}\left(Y_i \mid X_i\right) \quad \text{(by Assumption 2)}\\ &= \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)^2}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\mathrm{Var}\left(\beta_0+\beta_1X_i+u_i \mid X_i\right)\\ &= \frac{\frac{1}{n^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\mathrm{Var}\left(u_i \mid X_i\right)}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)^2}\\ &= \frac{\frac{1}{n^2}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\mathbb{E}\left[u_i^2 \mid X_i\right]}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)^2}\ ,\end{aligned} \tag{3.3}$$

because, in particular, $\mathrm{Var}\left(\beta_1X_i \mid X_i\right)=0$, as, conditional on X_i, β_1X_i is a constant.
Under Assumption 4, this expression can be further simplified to read:

$$\mathrm{Var}\left(\hat{\beta}_1 \mid X_1,\ldots,X_n\right)=\frac{\sigma_u^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\ . \tag{3.4}$$

Conditional variance of $\hat{\beta}_0$ is not pedagogically instructive and is omitted.

The Gauss-Markov theorem. Under assumptions stated above, including homoskedasticity, we can show that the OLS estimator is the *best linear (conditionally) unbiased estimator* (BLUE), in the sense that it has the smallest (conditional) variance among all linear and (conditionally) unbiased estimators of the regression coefficients. The theorem that claims that the OLS estimator is BLUE is called the *Gauss-Markov theorem* and requires the following conditions:

1. $\mathbb{E}[u_i \mid X_1, \dots X_n] = 0$. This is true because of Assumptions 1 and 2.
2. $\text{Var}(u_i \mid X_1, \dots X_n) = \sigma_u^2$, $0 < \sigma_u^2 < \infty$. The first part is true because of Assumptions 2 and 4, and the finite variance is implied by a stronger Assumption 3.
3. $\mathbb{E}[u_i u_j \mid X_1, \dots X_n] = 0$, $i \neq j$. This is true because, by Assumption 2, $\mathbb{E}[u_i u_j \mid X_1, \dots X_n] = \mathbb{E}[u_i u_j \mid X_i, X_j] = \mathbb{E}[u_i \mid X_i]\,\mathbb{E}[u_j \mid X_j]$. This product is equal to zero because of Assumption 1.

We will prove the theorem only for $\hat{\beta}_1$. The proof for $\hat{\beta}_0$ is not very instructive.

Proof. Consider another linear OLS estimator of β_1:

$$\tilde{\beta}_1 = \sum_{i=1}^{n} \tilde{w}_i Y_i = \beta_0 \sum_{i=1}^{n} \tilde{w}_i + \beta_1 \sum_{i=1}^{n} \tilde{w}_i X_i + \sum_{i=1}^{n} \tilde{w}_i u_i \ .$$

Taking expectations and invoking the first Gauss-Markov condition,

$$\mathbb{E}\left[\tilde{\beta}_1 \mid X_1, \dots, X_n\right] = \beta_0 \sum_{i=1}^{n} \tilde{w}_i + \beta_1 \sum_{i=1}^{n} \tilde{w}_i X_i \ .$$

Since $\tilde{\beta}_1$ has to be (conditionally) unbiased,

$$\beta_1 = \mathbb{E}\left[\tilde{\beta}_1 \mid X_1, \dots, X_n\right] = \beta_0 \sum_{i=1}^{n} \tilde{w}_i + \beta_1 \sum_{i=1}^{n} \tilde{w}_i X_i \ ,$$

from which we see that it must be the case that

$$\sum_{i=1}^{n} \tilde{w}_i = 0 \ , \tag{3.5}$$

$$\sum_{i=1}^{n} \tilde{w}_i X_i = 1 \ . \tag{3.6}$$

This means that

$$\tilde{\beta}_1 = \beta_1 + \sum_{i=1}^{n} \tilde{w}_i u_i \ ,$$

and thus

$$\text{Var}\left(\tilde{\beta}_1 \mid X_1, \dots, X_n\right) = \text{Var}\left(\sum_{i=1}^{n} \tilde{w}_i u_i \mid X_1, \dots, X_n\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{w}_i \tilde{w}_j \text{Cov}(u_i, u_j \mid X_1, \dots, X_n) \ .$$

From the second and the third Gauss-Markov conditions, we see that all the cross-terms in the double summation vanish:

$$\text{Var}\left(\tilde{\beta}_1 \mid X_1, \dots, X_n\right) = \sigma_u^2 \sum_{i=1}^{n} \tilde{w}_i^2 \ .$$

We need to show that this variance is at least as big as (3.4). Recall (3.1) and notice that (3.4) can be rewritten as

$$\text{Var}\left(\hat{\beta}_1 \mid X_1, \dots, X_n\right) = \sigma_u^2 \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}{\left(\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2\right)^2} \equiv \sigma_u^2 \sum_{i=1}^{n} w_i^2 \ .$$

We can always represent

$$\tilde{w}_i = w_i + d_i \ .$$

Note that since (3.6) must hold for both sets of weights,

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \tilde{w}_i - \sum_{i=1}^{n} w_i = 0 \ ,$$
$$\sum_{i=1}^{n} d_i X_i = \sum_{i=1}^{n} \tilde{w}_i X_i - \sum_{i=1}^{n} w_i X_i = 1 - 1 = 0 \ ,$$

and thus

$$\sum_{i=1}^{n} w_i d_i = \sum_{i=1}^{n} \frac{X_i - \overline{X}}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} d_i = \frac{\sum_{i=1}^{n} d_i X_i - \overline{X} \sum_{i=1}^{n} d_i}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = 0 \ .$$

Consider

$$\begin{aligned}
\text{Var}\left(\tilde{\beta}_1 \mid X_1, \ldots, X_n\right) - \text{Var}\left(\hat{\beta}_1 \mid X_1, \ldots, X_n\right) &= \sigma_u^2 \sum_{i=1}^{n} \left(\tilde{w}_i^2 - w_i^2\right) \\
&= \sigma_u^2 \sum_{i=1}^{n} \left((w_i + d_i)^2 - w_i^2\right) \\
&= \sigma_u^2 \sum_{i=1}^{n} \left(2 w_i d_i + d_i^2\right) \\
&= \sigma_u^2 \sum_{i=1}^{n} d_i^2 \geq 0 \ .
\end{aligned}$$

It is clear that the two variances will be the same if $d_i = 0$, $i = 1, \ldots, n$, i.e. if the weights coincide. Otherwise, the inequality will be strict, which shows that the (conditional) variance of $\hat{\beta}_1$ is the smallest in the class of linear (conditionally) unbiased estimators. ■

3.3 Asymptotic (Large-Sample) Properties

To be able to conduct inference for the OLS estimator, we need to know its (sampling) distribution. Since in practice, most of the datasets are of considerable size, the modern approach is not to study the small-sample distribution of the OLS estimator, but to consider its *asymptotic* distribution, i.e. the distribution when the sample size n is large. In particular, it is important to show that the OLS estimator is *consistent*, i.e. that when the sample size is sufficiently large, the distribution of the estimator will be tightly centered around the true population parameter[8].

Consistency. Consider expression (3.2):

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right) u_i}{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} \equiv \beta_1 + \frac{N_{OLS}}{D_{OLS}} \ .$$

Let us consider the denominator:

$$\begin{aligned}
D_{OLS} &= \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \frac{2}{n} \sum_{i=1}^{n} X_i \overline{X} + \frac{1}{n} \sum_{i=1}^{n} \left(\overline{X}\right)^2 \\
&\xrightarrow{p} \mathbb{E}\left[X_i^2\right] - 2\left(\mathbb{E}\left[X_i\right]\right)^2 + \left(\mathbb{E}\left[X_i\right]\right)^2 \\
&= \mathbb{E}\left[X_i^2\right] - \left(\mathbb{E}\left[X_i\right]\right)^2 \\
&= \text{Var}\left(X_i\right) \ .
\end{aligned}$$

[8] We have shown that the OLS estimator is unbiased even in small samples, but later on in the course we will deal with the estimators that are *not* unbiased but that are still consistent. Consistency thus is a more important property than unbiasedness, especially if we seek an estimator with low mean squared error, i.e. we are willing to trade unbiasedness for lower variance of the estimator.

Here, we used LLN to show that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}\left[X_i^2\right] ,$$
$$\overline{X} \xrightarrow{p} \mathbb{E}\left[X_i\right] ,$$

and we used the continuous mapping theorem (CMT) to show that

$$\left(\overline{X}\right)^2 \xrightarrow{p} \left(\mathbb{E}\left[X_i\right]\right)^2$$

and that the sum of the respective addends adds up to the sum of their probability limits.
By similar logic, we can show that the numerator converges to zero[9]:

$$N_{OLS} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)u_i = \frac{1}{n}\sum_{i=1}^{n} X_i u_i - \overline{X}\cdot\frac{1}{n}\sum_{i=1}^{n} u_i \xrightarrow{p} \mathbb{E}\left[X_i u_i\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[u_i\right] = 0$$

by Assumptions 1–2 and the law of iterated expectations. Therefore, by another application of CMT[10], we have:

$$\frac{N_{OLS}}{D_{OLS}} \xrightarrow{p} \frac{0}{\mathrm{Var}\left(X_i\right)} = 0 ,$$

and

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 ,$$

showing that $\hat{\beta}_1$ is consistent.
Consistency of $\hat{\beta}_0$ follows from the following observation:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X} \xrightarrow{p} \mathbb{E}\left[Y_i\right] - \beta_1\mathbb{E}\left[X_i\right] = \beta_0 + \beta_1\mathbb{E}\left[X_i\right] - \beta_1\mathbb{E}\left[X_i\right] = \beta_0 ,$$

where we have used LLN, CMT, and Assumption 1.

Asymptotic normality. To establish the asymptotic distribution of the OLS estimator, let us go back to N_{OLS} and D_{OLS}. Consider

$$\begin{aligned}
\sqrt{n}N_{OLS} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)u_i \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right] + \mathbb{E}\left[X_i\right] - \overline{X}\right)u_i \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right)u_i + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\mathbb{E}\left[X_i\right] - \overline{X}\right)u_i \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right)u_i + \sqrt{n}\overline{u}\left(\mathbb{E}\left[X_i\right] - \overline{X}\right) .
\end{aligned}$$

By the central limit theorem[11],

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(X_i - \mathbb{E}\left[X_i\right]\right)u_i \xrightarrow{d} N\left(0, \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)^2 u_i^2\right]\right) ,$$

[9] Note that it would be wrong to simply claim that $\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)u_i \xrightarrow{p} \mathbb{E}\left[\left(X_i - \overline{X}\right)u_i\right]$ because the variables in the sequence must be i.i.d. for LLN to be applied, and $\left(X_i - \overline{X}\right)u_i$ is not independent of $\left(X_j - \overline{X}\right)u_j$ because of the common term $\overline{X}$.
[10] This is allowed here because $\mathrm{Var}\left(X_i\right) \neq 0$, otherwise the OLS estimator would not exist.
[11] Note that $\left(X_i - \mathbb{E}\left[X_i\right]\right)u_i$, $i = 1, \ldots, n$, are i.i.d. random variables, which validates application of the theorem.

because $\mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right) u_i\right] = 0$ by LIE and Assumption 1. The second term asymptotically goes to 0[12]. Therefore, by the Slutsky theorem,

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) = \sqrt{n}\frac{N_{OLS}}{D_{OLS}} \xrightarrow{d} N\left(0, \frac{\mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)^2 u_i^2\right]}{\left(\mathrm{Var}\left(X_i\right)\right)^2}\right) \equiv N\left(0, \mathrm{aVar}\left(\hat{\beta}_1\right)\right) . \tag{3.7}$$

In other words, slightly informally, we can claim that

$$\hat{\beta}_1 \approx N\left(\beta_1, \frac{\mathrm{aVar}\left(\hat{\beta}_1\right)}{n}\right) .$$

Similar to (3.4), we can show that under Assumption 4, (3.7) simplifies to

$$\sqrt{n}\left(\hat{\beta}_1 - \beta_1\right) \xrightarrow{d} N\left(0, \frac{\sigma_u^2}{\mathrm{Var}\left(X_i\right)}\right) . \tag{3.8}$$

Estimator of the intercept, $\hat{\beta}_0$, also is asymptotically normal, but derivation of this fact is not pedagogically instructive.

Relationship to projection coefficients. Note that to prove consistency and asymptotic normality (although not unbiasedness), we only used the fact that $\mathrm{Cov}\left(X_i, u_i\right) = 0$, more precisely, that $\mathbb{E}\left[X_i u_i\right] = 0$ and that $\mathbb{E}\left[u_i\right] = 0$. Recall that these are the properties of projection coefficients (2.6) introduced in Sect. 2.3. In other words, we have thus shown that the OLS estimator is also a consistent estimator of the projection coefficients.

Therefore, in practice, it is always possible to estimate OLS coefficients, which will be consistent for (2.6). It is only Assumption 1, $\mathbb{E}\left[X_i \mid u_i\right] = 0$, that adds content to the model and helps distinguish between "merely" projections and causal relationships.

4 Goodness of Fit

To measure how good the fit of the linear model we estimate is, we make use of the following quantities:

- *sum of squared residuals* (SSR) already introduced above, which is a measure of how much of the data cannot be explained by our estimated linear regression[13]:

$$\mathrm{SSR} = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 ;$$

- *explained sum of squares* (ESS) is a measure of how much of the data can be explained by the regression:

$$\mathrm{ESS} = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 . \tag{4.1}$$

  In other words, ESS shows the variation due to X_i, because $\hat{Y}_i$ is a function of X_i only;

- *total sum of squares* (TSS) is the total amount of information in the data, or a measure of the total spread of Y:

$$\mathrm{TSS} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 . \tag{4.2}$$

[12] To see this, perhaps slightly informally, consider that $\overline{u} \xrightarrow{p} \mathbb{E}\left[u_i\right] = 0$ by LLN, $\sqrt{n}\left(\mathbb{E}\left[X_i\right] - \overline{X}\right) \xrightarrow{d} -N(0, \mathrm{Var}\left(X_i\right))$ by CLT, and thus, by the Slutsky theorem, $\sqrt{n}\overline{u}\left(\mathbb{E}\left[X_i\right] - \overline{X}\right)$ goes to zero.

[13] This is also sometimes referred to as *residual sum of squares*, RSS.

We can rewrite (4.1) as follows:

$$\text{ESS} = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{\beta}_0 + \hat{\beta}_1 X_i - \hat{\beta}_0 - \hat{\beta}_1 \overline{X}\right)^2 = \left(\hat{\beta}_1\right)^2 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 . \tag{4.3}$$

It can be shown that

$$\text{TSS} = \text{ESS} + \text{SSR} . \tag{4.4}$$

Indeed, consider:

$$\begin{aligned}
\text{TSS} &= \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y}\right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right) + \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n} \hat{u}_i \left(\hat{Y}_i - \overline{Y}\right) + \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 \\
&= \text{ESS} + \text{SSR} + 2\sum_{i=1}^{n} \hat{u}_i \left(\hat{Y}_i - \overline{Y}\right) .
\end{aligned}$$

Observe that

$$\sum_{i=1}^{n} \hat{u}_i \overline{Y} = \overline{Y} \sum_{i=1}^{n} \hat{u}_i = 0 ,$$

using (2.4).

Also observe that

$$\sum_{i=1}^{n} \hat{u}_i \hat{Y}_i = \sum_{i=1}^{n} \hat{u}_i \left(\hat{\beta}_0 + \hat{\beta}_1 X_i\right) = \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i X_i = 0 ,$$

using (2.4) and (2.5). Therefore, the last addend above is zero, completing the proof of (4.4).

In particular, (4.4) means that if TSS = ESS, then the fitted line perfectly explains the data, i.e. all data points lie precisely on the regression line.

Using the concepts just introduced, we define the *coefficient of determination*

$$R^2 = \frac{\text{ESS}}{\text{TSS}} = 1 - \frac{\text{SSR}}{\text{TSS}} . \tag{4.5}$$

R^2 measures the proportion of variation in Y_i that can be explained by the variation in X. It measures how accurately the predicted values $\hat{Y}_i$ fit the data. It is thus a measure of the *predictive* strength of the regression. Because of (4.4), it ranges from 0 to 1 by construction, with 1 indicating the perfect match.

By definition, R^2 is higher for regressions with:

- lower variance in residuals (smaller spread around the regression line);
- higher variance of Y_i (greater spread of Y_i values).

From Fig. 4.1, it is clear that the smaller the proportion of SSR over TSS, the better the fit is. Therefore, R^2 is often used as a measure of "goodness of fit." Note that, nevertheless, maximizing R^2 is *never* an objective of causal regression design.

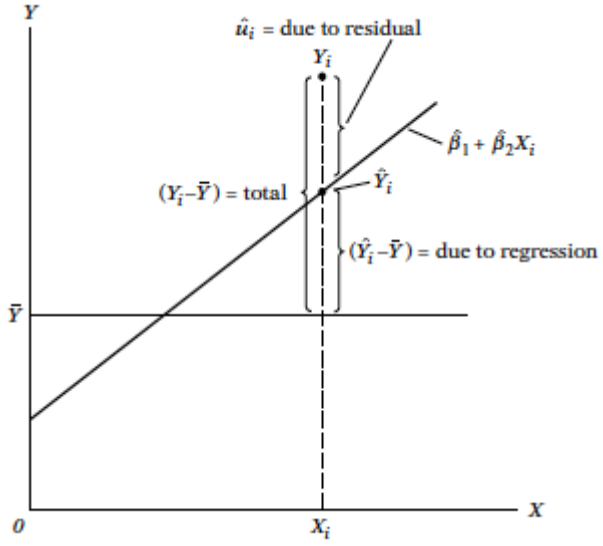


Figure 4.1: Graphical illustration of ESS, TSS, and SSR

Standard error of the regression. The *standard error of the regression* is an estimator of the standard deviation of the regression errors u_i:

$$\text{SER} = \hat{\sigma}_u \,, \quad \hat{\sigma}_u^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{\text{SSR}}{n-2} \,, \tag{4.6}$$

where division by $n-2$ is due to the degrees of freedom correction (two parameters, intercept and slope, has been estimated to fit the regression line). Estimator $\hat{\sigma}_u^2$ is a (conditionally) unbiased (under homoskedasticity assumption) and consistent estimator of $\text{Var}\left(u_i \mid \mathbf{x}_i\right)$.

SER measures the "typical" spread of the data around the regression line.

5 Hypothesis Testing of Regression Estimates

Since estimators of coefficients in a regression model are random variables (with the randomness coming from a random nature of a given sample), we can observe only their realizations (corresponding to this sample). Therefore, it is important not only to produce an estimate (function of the data) of a regression coefficient (a population parameter) but also to make *inference* about the population parameter, in particular, test statistical hypotheses about the true value of this regression coefficient.

As was discussed previously, to perform hypothesis testing, we need to know the (asymptotic) distribution of the coefficient we are making inference about. However, computing (3.7) is infeasible because we never observe the true errors. We start our discussion with the problem of estimating the asymptotic variance of the OLS coefficients.

5.1 Asymptotic Variance Estimation

A natural estimator of (3.7) is obtained by replacing unobserved true errors u_i with the estimated residuals $\hat{u}_i$ and all the population moments with their sample analogues:

$$\widehat{\text{aVar}}\left(\hat{\beta}_1\right)^{HC0} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right)^2} \,. \tag{5.1}$$

This is a consistent estimator for $\text{aVar}\left(\hat{\beta}_1\right)$[14] and is called the *HC0 estimator* (from "heteroskedasticity-consistent"). Recall from (4.6) that a degrees of freedom correction is often applied in practice to reduce bias in the estimator of the variance of the true errors. This leads to the following *HC1 estimator* of the asymptotic variance of the OLS estimator:

$$\widehat{\text{aVar}}\left(\hat{\beta}_1\right)^{HC1} = \frac{n}{n-2}\widehat{\text{aVar}}\left(\hat{\beta}_1\right)^{HC0} . \tag{5.2}$$

Although scaling by $n/(n-2)$ is a little bit informal, HC1 is recommended in practice over HC0. Sometimes such scaling is called *small-sample correction*.

Default estimators in software packages. Please note that such widely used software packages as Stata and R by default estimate (3.8), i.e. the standard errors they produce assume homoskedasticity. To produce heteroskedasticity-robust standard errors from the HC0 or HC1 estimators as given above, the researcher needs to specify this explicitly. In particular, in Stata, adding "`, robust`" option asks the system to compute the HC1 standard errors.

5.2 Hypothesis Testing Steps

Hypothesis testing for the regression coefficients involves making decisions about null hypotheses, which can be broken down into several steps.

1. Propose a null hypothesis $H_0 : \beta_j = \beta_j^0$. Select an alternative hypothesis:
   - a two-sided alternative of the form $H_1 : \beta_j \neq \beta_j^0$;
   - a one-sided alternative of the form $H_1 : \beta_j < \beta_j^0$;
   - a one-sided alternative of the form $H_1 : \beta_j > \beta_j^0$.

   Often (though not always), we take $\beta_j^0 = 0$.
2. Construct a test statistic (t-statistic):
$$t = \frac{\hat{\beta}_j - \beta_j^0}{\text{se}\left(\hat{\beta}_j\right)} \sim t_{n-2} , \tag{5.3}$$
   where $\text{se}\left(\hat{\beta}_j\right) = \sqrt{\widehat{\text{aVar}}\left(\hat{\beta}_j\right)/n}$ is the *standard error* of $\hat{\beta}_j$, and t is known to follow, under the null, the t-distribution with $n-2$ *degrees of freedom*, where n is the number of observations and 2 is the number of parameters we estimate in the regression model. In large samples, as was shown above, we approximate the distribution of (5.3) with the standard normal distribution (especially considering that the t-distribution is very close to the standard normal distribution in large samples).
3. Find t_{critical}, given a significance level α, for example, 1%, 5%, or 10%, such that (using symmetry of standard normal):
   - for two-sided hypotheses,
$$\alpha = \mathbb{P}\left(t < t_{\text{critical}} \mid \beta_j^0\right) + \mathbb{P}\left(t > t_{\text{critical}} \mid \beta_j^0\right) = 2\mathbb{P}\left(t < t_{\text{critical}} \mid \beta_j^0\right) = 2\Phi\left(t_{\text{critical}}\right) ,$$
     i.e. $t_{\text{critical}} = \Phi^{-1}\left(\frac{\alpha}{2}\right) \equiv q_{\frac{\alpha}{2}}$, the $\alpha/2$-quantile of the standard normal distribution;
   - for one-sided hypotheses of the form $H_1 : \beta_j < \beta_j^0$,
$$\alpha = \mathbb{P}\left(t < t_{\text{critical}} \mid \beta_j^0\right) = \Phi\left(t_{\text{critical}}\right) ,$$
     i.e. $t_{\text{critical}} = \Phi^{-1}\left(\alpha\right) \equiv q_{\alpha}$, the α-quantile of the standard normal distribution;

[14] We will not prove it here because it is tedious.

- for one-sided hypotheses of the form $H_1 : \beta_j > \beta_j^0$,

$$\alpha = \mathbb{P}\left(t > t_{\text{critical}} \mid \beta_j^0\right) = 1 - \Phi\left(t_{\text{critical}}\right) = \Phi\left(-t_{\text{critical}}\right) \ ,$$

  i.e. $t_{\text{critical}} = -\Phi^{-1}(\alpha) \equiv q_{1-\alpha}$, a $(1-\alpha)$-quantile of the standard normal distribution.

In particular, in large samples, for the 5% significance level and a two-sided alternative, $t_{\text{critical}} = \pm q_{0.025} = \pm 1.96$[15], and for one-sided alternatives, $t_{\text{critical}} = \pm q_{0.05} = \pm 1.64$.

For 1% level, $t_{\text{critical}} = \pm 2.58$ for two-sided alternatives and ± 2.33 for one-sided alternatives.

4. Compare the t-statistic (5.3) to t_{critical} and reject H_0:
   - for two-sided hypotheses, when $|t| > |t_{\text{critical}}|$;
   - for one-sided hypotheses of the form $H_1 : \beta_j < \beta_j^0$, when $t < t_{\text{critical}}$;
   - for one-sided hypotheses of the form $H_1 : \beta_j > \beta_j^0$, when $t > t_{\text{critical}}$.

Equivalently, after Step 2, one can calculate the corresponding p-value:

- for two-sided alternative:
$$p = \mathbb{P}\left(|t| > |\hat{t}| \mid \beta_j^0\right) = 2\Phi(-|\hat{t}|) \ ;$$
- for one-sided alternative of the form $H_1 : \beta_j < \beta_j^0$:
$$p = \mathbb{P}\left(t < \hat{t} \mid \beta_j^0\right) = \Phi(\hat{t}) \ ;$$
- for one-sided alternative of the form $H_1 : \beta_j > \beta_j^0$:
$$p = \mathbb{P}\left(t > \hat{t} \mid \beta_j^0\right) = \Phi(-\hat{t}) \ ,$$

where t is (5.3), and $\hat{t}$ is its realization calculated using the sample at hand assuming that H_0 is true. Then one can use this p-value to decide whether to reject H_0.

A confidence interval for β_j^0 with $1-\alpha$ degree of confidence is derived from the fact that

$$\mathbb{P}\left(\frac{\hat{\beta}_j - \beta_j^0}{\text{se}\left(\hat{\beta}_j\right)} \in \left[-q_{1-\frac{\alpha}{2}}; q_{1-\frac{\alpha}{2}}\right]\right) = 1 - \alpha \ ,$$

and, using the symmetry of the standard normal distribution, is given by

$$\text{CI}_{1-\alpha}\left(\beta_j^0\right) = \left[\hat{\beta}_j - q_{1-\frac{\alpha}{2}} \cdot \text{se}\left(\hat{\beta}_j\right); \hat{\beta}_j + q_{1-\frac{\alpha}{2}} \cdot \text{se}\left(\hat{\beta}_j\right)\right] \ . \tag{5.4}$$

For instance, for a 95% confidence interval, we have:

$$\text{CI}_{0.95}\left(\beta_j^0\right) = \left[\hat{\beta}_j - 1.96 \cdot \text{se}\left(\hat{\beta}_j\right); \hat{\beta}_j + 1.96 \cdot \text{se}\left(\hat{\beta}_j\right)\right] \ .$$

[15] For that reason, the rule of thumb is sometimes to use $t_{\text{critical}} = \pm 2$ for the 5% significance level.

5.3 Testing for Difference in Means

Consider a sample $(X_i, Y_i)_{i=1}^n$, with X_i binary. Suppose the problem is to test whether the population means for two different groups are the same: $\mathbb{E}[Y_i \mid X_i = 0] = \mathbb{E}[Y_i \mid X_i = 1]$.

One approach would be to consider a test statistic $\overline{Y}_1 - \overline{Y}_0$, where $\overline{Y}_j$ is the sample average taken over $X_i = j$. We know from previous lectures that the sampling distribution of the sample average is normal:

$$\overline{Y}_j \sim N\left(\mathbb{E}[Y_i \mid X_i = j], \frac{\text{Var}(Y_i \mid X_i = j)}{n_j}\right) ,$$

where n_j is the number of observations in the sample with $X_i = j$.

The difference in means is therefore also a normal variable:

$$\overline{Y}_1 - \overline{Y}_0 \sim N\left(\mathbb{E}[Y_i \mid X_i = 1] - \mathbb{E}[Y_i \mid X_i = 0], \frac{\text{Var}(Y_i \mid X_i = 1)}{n_1} + \frac{\text{Var}(Y_i \mid X_i = 0)}{n_0}\right) .$$

This allows us to apply the standard hypothesis testing procedure to a new test statistic,

$$t = \frac{\overline{Y}_1 - \overline{Y}_0}{\sqrt{\frac{\text{Var}(Y_i|X_i=1)}{n_1} + \frac{\text{Var}(Y_i|X_i=0)}{n_0}}} .$$

An alternative way to test for difference in means would be to consider the regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i ,$$

and observe that

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \arg\min_{b_0, b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2 = \arg\min_{b_0, b_1} \sum_{i:X_i=0} (Y_i - b_0)^2 + \sum_{i:X_i=1} (Y_i - b_0 - b_1)^2 ,$$

which we can reparameterize with $\gamma = \beta_0 + \beta_1$ as follows:

$$\left(\hat{\beta}_0, \hat{\gamma}\right) = \left(\arg\min_{b_0} \sum_{i:X_i=0} (Y_i - b_0)^2 , \arg\min_{c} \sum_{i:X_i=1} (Y_i - c)^2\right) .$$

The first order condition for the first problem gives

$$-2 \sum_{i:X_i=0} (Y_i - b_0) = 0 ,$$

which gives $\hat{\beta}_0 = \overline{Y}_0$. Likewise, $\hat{\gamma} = \overline{Y}_1$. Recalling the definition of γ, we get that $\hat{\beta}_1 = \hat{\gamma} - \hat{\beta}_0 = \overline{Y}_1 - \overline{Y}_0$. In other words, to test the difference in the two means, we can estimate the above regression model and test the hypothesis $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

6 Monte Carlo Simulations

To illustrate statistical properties of the OLS estimator, we will perform a Monte Carlo exercise. Suppose that we have two variables, STR and average test scores in a school district, related as follows:

$$\text{test score}_i = 660 - 2 \cdot \text{STR}_i + u_i .$$

This is our DGP, with $\beta_0 = 660$ and $\beta_1 = -2$. We will also assume that $\text{STR}_i \sim N(20, 4)$ and, to model heteroskedasticity, that $u_i \mid \text{STR}_i \sim N\left(0, 0.25 \cdot (\text{STR}_i - 15)^2\right)$. Observe that $\mathbb{E}[u_i \mid \text{STR}_i] = 0$.

One typical sample, of size $n = 1000$, generated according to this DGP is given in Fig. 6.1.

We will conduct three simulations:

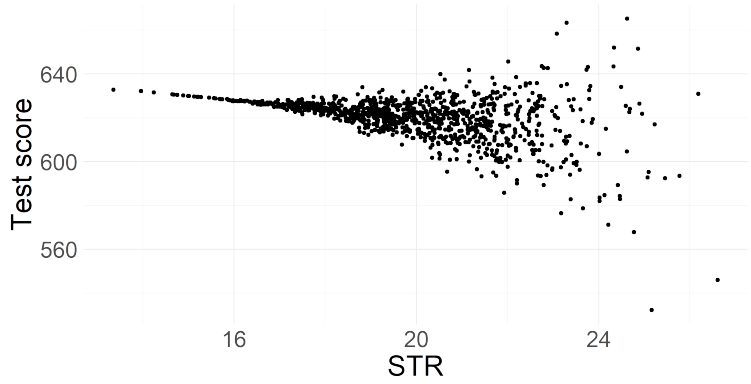


Figure 6.1: A typical sample of size $n = 1000$ generated according to the DGP for the Monte Carlo simulations

- to illustrate that $\left(\hat{\beta}_0, \hat{\beta}_1\right)^\top$ is conditionally unbiased, we will draw a (small) sample of STR_i, $i = 1, \ldots, 50$, and simulate $T = 10{,}000$ different samples $(\text{STR}_i, \text{test score}_i)_{i=1}^{n=50}$ by drawing u_i from its distribution. The resulting histograms should be centered around the true values;
- to illustrate that $\left(\hat{\beta}_0, \hat{\beta}_1\right)^\top$ is *un*conditionally unbiased, we will perform the same exercise but will generate a new sample of STR_i each time instead of keeping it fixed for all T simulation runs;
- to illustrate consistency of $\left(\hat{\beta}_0, \hat{\beta}_1\right)^\top$, we will increase the sample size n to 1000 and repeat the second exercise.

Histograms for $\hat{\beta}_0$ and $\hat{\beta}_1$ are given in Fig. 6.2. Several comments are in order:

- the OLS estimators for both coefficients are both conditionally and unconditionally unbiased, as the corresponding histograms are centered around the true population values;
- the OLS estimators for both coefficients are consistent, which is evident from their histograms for the larger sample. The distributions are still centered around the true population values, and they get tighter as the sample size increases.

To illustrate importance of using heteroskedasticity-robust standard errors (obtained from (5.2)) instead of the homoskedastic ones (obtained from an estimator of (3.4)), for each simulation out of $T = 10{,}000$, we compute both standard errors and test the following hypothesis: $H_0 : \beta_1 = -2$ vs. $H_1 : \beta_1 \neq -2$ at significance level $\alpha = 0.05$. Under the null, we expect that our test will reject H_0 around 5% of the time. For $n = 1000$ with random STR_i, we obtain the following result:

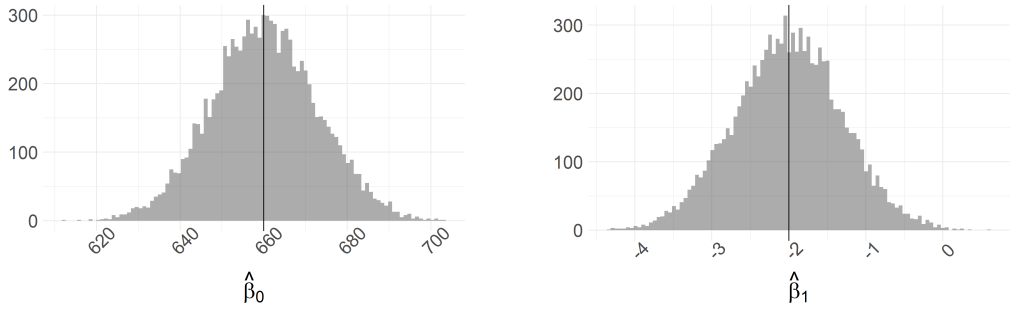
- using homoskedastic standard errors, H_0 is rejected 17.36% of the time;
- using heteroskedasticity-robust standard errors, H_0 is rejected 5.21% of the time.

As we can see, homoskedastic standard errors reject the null much too often, meaning that the probability of making a Type I error is higher that we expect it to be by picking up a critical value corresponding to $\alpha = 0.05$. Heteroskedasticity-robust standard errors, on the contrary, are more conservative and thus we reject less.

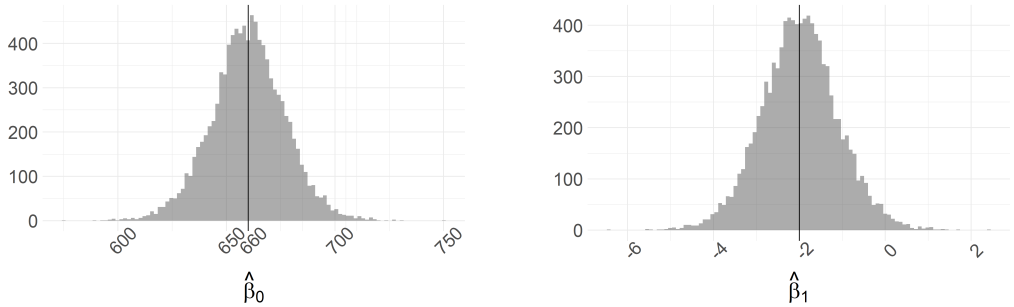
The R code for this exercise is given in Listing 1.

Listing 1: R code used to perform Monte Carlo simulations for bivariate regression

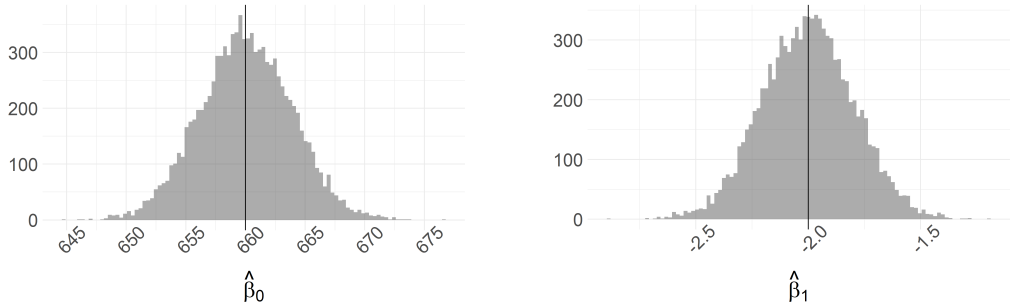
```
library(ggplot2)
library(gridExtra)
library(latex2exp)
```



(a) $n = 50$, STR fixed



(b) $n = 50$, STR random



(c) $n = 1000$, STR random

Figure 6.2: Histograms of the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ for the Monte Carlo simulations with different sample sizes (true parameter is indicated by a vertical bar)

```
5
# Function to generate a sample of X's
generate.X <- function(n){
  X <- rnorm(n, 20, sd = 2)
}
10

# Function to generate a dataset of size n
generate.data <- function(n, X = NULL, beta.0, beta.1){
  # generate X, if it wasn't passed inside
15  if (is.null(X)){
    X <- generate.X(n)
  }

  # errors:
20  u <- rnorm(n, 0, sd = 0.25*(X - 15)^2)

  # simulate Y using u and given X
  Y <- beta.0 + beta.1*X + u

25  # save data in a big dataframe
  data <- data.frame(Y, X, u)

  return(data)
}
30

# Function to compute OLS coefficients
compute.OLS <- function(X, Y){
  hat.beta.1 <- (mean(X*Y) - mean(X)*mean(Y)) / (mean(X^2) - mean(X)^2)
35  hat.beta.0 <- mean(Y) - hat.beta.1*mean(X)

  return(list("hat.beta.1" = hat.beta.1, "hat.beta.0" = hat.beta.0))
}

40

# Function to compute the (estimator of the) asymptotic variance of beta.1
avar.beta.1 <- function(data, hat.beta.0, hat.beta.1){
  n <- nrow(data)

45  # generate predicted residuals
  u.hat <- data$Y - hat.beta.0 - data$X*hat.beta.1

  # compute estimate of Var(u)
  s2 <- sum(u.hat^2) / (n - 2)
50
  # compute the estimator
  # homoskedastic
  aVar0 <- s2 / mean((data$X - mean(data$X))^2)

55  # heteroskedastic
  aVar <- n/(n - 2) *
    mean((data$X - mean(data$X))^2 * u.hat^2) / (mean((data$X - mean(data$X))^2))^2

  return(list("homo" = aVar0, "hetero" = aVar))
60 }


# Helper function to plot one histogram
plot.histogram <- function(data, true.value, n, xlab){
65  plot <- ggplot(data, aes_string(x = names(data)[1])) +
    geom_histogram(bins = 100, alpha = 0.5, position = "identity") +
    geom_vline(xintercept = true.value) +
    labs(x = TeX(xlab), y = "") +
    scale_x_continuous(breaks = c(pretty(data[, 1]), true.value)) +
70  theme_minimal() +
```

```
    theme(
      text = element_text(size = 25),
      plot.margin = margin(0.5, 0.5, 0.5, 2, "cm"),
      axis.text.x = element_text(angle = 45)
    )

  return(plot)
}


# Helper function to plot one pair of histograms
plot.histogram.both.coefficients <- function(
    data, true.values, n, file.suffix
){
  plot.beta.0 <- plot.histogram(
    data[names(data)[1]], true.value = true.values[1],
    n, xlab = "$\\hat{\\beta}_{0}$"
  )

  plot.beta.1 <- plot.histogram(
    data[names(data)[2]], true.value = true.values[2],
    n, xlab = "$\\hat{\\beta}_{1}$"
  )

  plot.both <- grid.arrange(plot.beta.0, plot.beta.1, ncol = 2)

  ggsave(
    paste0("histcoef_", file.suffix, "_", n, ".png"),
    plot.both, width = 5000, height = 1500, units = "px"
  )
}


# Function performing a Monte Carlo simulation
# for a data sample of size n with T repetitions
monte.carlo <- function(n, T, true.beta.0, true.beta.1, alpha){
  print(paste0("Sample size: ", n))

  # allocate memory
  hat.beta.0.cond <- rep(0, T)
  hat.beta.1.cond <- rep(0, T)

  hat.beta.0.uncond <- rep(0, T)
  hat.beta.1.uncond <- rep(0, T)

  reject.H0.homo <- rep(0, T)
  reject.H0.hetero <- rep(0, T)

  # get critical value
  t.critical <- qnorm(alpha/2)

  # generate X
  X <- generate.X(n)
  for (i in 1:T){
    # generate a dataset with given X
    data.cond <- generate.data(
      n = n, beta.0 = true.beta.0, beta.1 = true.beta.1, X = X
    )

    # generate a dataset with random X
    data.uncond <- generate.data(
      n = n, beta.0 = true.beta.0, beta.1 = true.beta.1
    )

    # for each sample, we compute regression coefficients
    coefs.cond <- compute.OLS(data.cond$X, data.cond$Y)
    hat.beta.0.cond[i] <- coefs.cond$hat.beta.0
```

```
    hat.beta.1.cond[i] <- coefs.cond$hat.beta.1

    coefs.uncond <- compute.OLS(data.uncond$X, data.uncond$Y)
    hat.beta.0.uncond[i] <- coefs.uncond$hat.beta.0
    hat.beta.1.uncond[i] <- coefs.uncond$hat.beta.1

    # compute estimator of the variance of beta.1
    aVar <- avar.beta.1(data.uncond, hat.beta.0.uncond[i], hat.beta.1.uncond[i])

    # compute test statistics for beta.1
    t.homo <- (hat.beta.1.uncond[i] - true.beta.1) / sqrt(aVar$homo / n)
    t.hetero <- (hat.beta.1.uncond[i] - true.beta.1) / sqrt(aVar$hetero / n)

    # perform hypothesis test
    reject.H0.homo[i] <- 1*(abs(t.homo) > abs(t.critical))
    reject.H0.hetero[i] <- 1*(abs(t.hetero) > abs(t.critical))
  }

  # report percentage of (incorrectly) rejected null hypotheses
  print("Percent of (incorrectly) rejected null hypotheses:")
  print(
    sprintf("     homoskedastic standard errors: %0.4f", mean(reject.H0.homo))
  )
  print(
    sprintf("     heteroskedastic standard errors: %0.4f", mean(reject.H0.hetero))
  )

  # organize the data in dataframes
  df.cond <- data.frame(beta.0 = hat.beta.0.cond, beta.1 = hat.beta.1.cond)
  df.uncond <- data.frame(beta.0 = hat.beta.0.uncond, beta.1 = hat.beta.1.uncond)

  # plot the histograms
  plot.histogram.both.coefficients(
    df.cond, true.values = c(true.beta.0, true.beta.1),
    n, file.suffix = "cond"
  )
  plot.histogram.both.coefficients(
    df.uncond, true.values = c(true.beta.0, true.beta.1),
    n, file.suffix = "uncond"
  )
}


# set random number generator seed for reproducibility
set.seed(100)

# set working directory to the current file
setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

# set constants
max.T <- 10000
alpha <- 0.05
# DGP:
# test.score = 660 - 2*STR + u
true.beta.0 <- 660
true.beta.1 <- -2

# generate one dataset and visualize it
data <- generate.data(n = 1000, beta.0 = true.beta.0, beta.1 = true.beta.1)

ggplot(data, aes(x = X, y = Y)) +
  geom_point() +
  theme_minimal() +
  labs(x = "STR", y = "Test score") +
  theme(
    text = element_text(size = 25),
    plot.margin = margin(0.5, 0.5, 0.5, 2, "cm")
```

```
  )
ggsave(
  paste0("sample_", 1000, ".png"),
  width = 3000, height = 1500, units = "px"
)

# run Monte Carlo simulations with different sample sizes
monte.carlo(
  n = 50, T = max.T,
  true.beta.0 = true.beta.0, true.beta.1 = true.beta.1,
  alpha = alpha
)
monte.carlo(
  n = 1000, T = max.T,
  true.beta.0 = true.beta.0, true.beta.1 = true.beta.1,
  alpha = alpha
)
```