Review of Bivariate Regression

A.Colin Cameron
Department of Economics
University of California - Davis
accameron@ucdavis.edu

October 27, 2006

Abstract

This provides a review of material covered in an undergraduate class on OLS regression with a single regressor. It presents introductory material that is assumed known in my Economics 240A course on multivariate regression using matrix algebra.

Contents

1 Introduction

Bivariate data analysis considers the **relationship between two variables**, such as education and income or house price and house size, rather than analyzing just one variable in isolation.

In principle the two variables should be treated equally. In practice one variable is often viewed as being caused by another variable. The standard notation used follows the notation of mathematics, where y is a function of x. Thus the variable y is explained by the variable x. [It is important

Sale Price	Square feet
375000	3300
340000	2400
310000	2300
279900	2000
278500	2600
273000	1900
272000	1800
270000	2000
270000	1800
258500	1600
255000	1500
253000	2100
249000	1900
245000	1400
244000	2000
241000	1600
239500	1600
238000	1900
236500	1600
235000	1600
235000	1700
233000	1700
230000	2100
229000	1700
224500	2100
220000	1600
213000	1800
212000	1600
204000	1400

Figure 1: House Sale Price in dollars and House Size in square feet for 29 houses in central Davis.

to note, however, that without additional information the roles of the two variables may in fact be reversed, so that it is x that is being explained by y. Correlation need not imply causation.]

This chapter introduces bivariate regression, reviewing an undergraduate course. Some of the results are just stated, with proof left for the multiple regression chapter.

2 Example: House Price and Size

Figure 1 presents data on the price (in dollars) and size (in square feet) of 29 houses sold in central Davis in 1999. The data are ordered by decreasing price, making interpretation easier.

It does appear that higher priced houses are larger. For example, the

Figure 2: House Sale Price and House Size: Two-way Scatter Plot and Regression Line for 29 house sales in central Davis in 1999.

five most expensive houses are all 2,000 square feet or more, while the four cheapest houses are all less than 1,600 square feet in size.

Figure 2 provides a scatterplot of these data. Each point represents a combination of sale price and size of house. For example, the upper right point is for a house that sold for $375,000 and was 3,400 square feet in size. The scatterplot also suggests that larger houses sell for more.

Figure 2 also includes the line that best fits these data, based on the least squares regression method explained below. The estimated regression line is

$$y = 115017 + 73.77x,$$

where y is house sale price and x is house size in square feet.

A more complete analysis of this data using the Stata command regress yields the output

```
. regress salepric sqfeet

      Source |       SS       df       MS              Number of obs =      29
```

```
-------------+------------------------------           F(  1,    27) =   43.58
       Model |  2.4171e+10     1  2.4171e+10           Prob > F      =  0.0000
    Residual |  1.4975e+10    27   554633395           R-squared     =  0.6175
-------------+------------------------------           Adj R-squared =  0.6033
       Total |  3.9146e+10    28  1.3981e+09           Root MSE      =   23551

------------------------------------------------------------------------------
    salepric |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      sqfeet |   73.77104   11.17491     6.60   0.000     50.84202    96.70006
       _cons |   115017.3   21489.36     5.35   0.000     70924.76    159109.8
------------------------------------------------------------------------------
```

The bottom results on the slope coefficient are of most interest. A one square foot increase in house size is associated with a $73.77 increase in price. This estimate is reasonably precise, with a standard error of $11.17 and a 95% confidence interval of ($50.84, $96.70). A test of the hypothesis that house size is not associated with house price (i.e. the slope coefficient is zero) is resoundingly rejected as the p-value (for a two-sided test) is 0.000.

The top results include a measure of goodness of fit of the regression, with R^2 of 0.6175.

The remainder of this review answers questions such as (1) How is the estimated line obtained? (2) How do we interpret the estimates? (3) How do allow for a different sample of house sales leading to different estimates?

3 Ordinary Least Squares Regression

Regression is the data analysis tool most used by economists.

3.1 Regression Line

The **regression line** from regression of y on x is denoted

$$\widehat{y} = b_1 + b_2 x, \tag{1}$$

where

- y is called the **dependent** variable
- $\widehat{y}$ is the **predicted** (or **fitted**) dependent variable

- x is the **independent** variable or **explanatory** variable or **regressor** variable or **covariate**.
- b_1 is the estimated intercept (on the y-axis)
- b_2 is the estimated slope coefficient

Later on for multiple regression we will denote the estimates as $\widehat{\beta}_1$ and $\widehat{\beta}_2$, rather than b_1 and b_2.

3.2 Interpretation

Interest lies especially in the slope coefficient. Since

$$\frac{d\widehat{y}}{dx} = b_2, \tag{2}$$

the slope coefficient b_2 is easily interpreted as the increase in the predicted value of y when x increases by one unit.

For example, for the regression of house price on size, $y = 115017 + 73.77x$, so house price is predicted to increase by 73.77 units when x increases by one unit. The units of measurement for this example are dollars for price and square feet for size. So equivalently a one square foot increase in house size is associated with a \$73.77 increase in price.

3.3 Least Squares Method

The regression line is obtained by choosing that line closest to all of the data points, in the following sense.

Define the residual e to be the difference between the actual value of y and the predicted value $\widehat{y}$. Thus the **residual**

$$e = y - \widehat{y}.$$

This is illustrated in Figure 3.

For the first observation, with subscript 1, the residual is $e_1 = y_1 - \widehat{y}_1$, for the second observation the residual is $e_2 = y_2 - \widehat{y}_2$, and so on. The **least squares method** chooses values of the intercept b_1 and slope b_2 of the line to **make as small as possible the sum of the squared residuals**, i.e. minimize $e_1^2 + e_2^2 + \cdots + e_n^2$. For a representative observation, say the i^{th} observation, the residual is given by

$$\begin{aligned} e_i &= y_i - \widehat{y}_i \\ &= y_i - b_1 - b_2 x_i. \end{aligned} \tag{3}$$

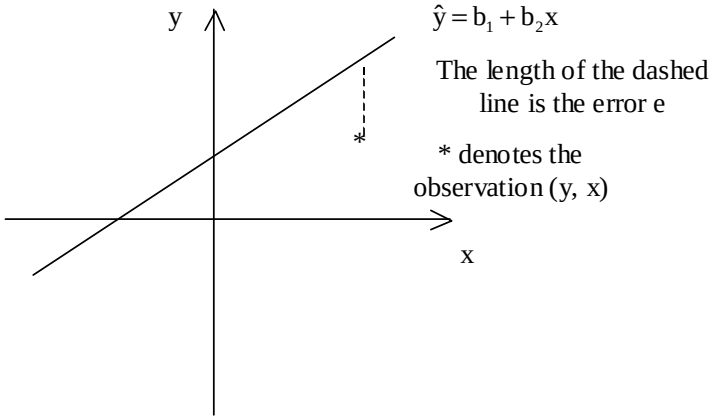

Figure 3: Least squares residual. The graph gives one data point, denoted *, and the associated residual e which is the length of the vertical dashed line between * and the regression line. Here the residual is negative since $y - \widehat{y}$ is negative. The regression line is the line that makes the sum of squared residuals over all data points as small as possible.

Given a sample of size n with data $(y_1, x_1), ..., (y_n, x_n)$, the **ordinary least squares (OLS)** method chooses b_1 and b_2 to minimize the sum of squares of the residuals. Thus b_1 and b_2 minimize **the sum of the squared residuals**

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2. \tag{4}$$

This is a calculus problem. Differentiating with respect to b_1 and b_2 yields two equations in the two unknowns b_1 and b_2

$$-2 \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i) = 0 \tag{5}$$

$$-2 \sum_{i=1}^{n} x_i (y_i - b_1 - b_2 x_i) = 0. \tag{6}$$

These are called the **least squares normal equations**.

Some algebra yields the **least squares intercept** as

$$b_1 = \bar{y} - b_2 \bar{x}, \tag{7}$$

and the **least squares slope coefficient** as

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{8}$$

We now obtain these results. First, manipulating (5) yields

$$\begin{aligned} & \textstyle\sum_{i=1}^{n}(y_i - b_1 - b_2 x_i) = 0 \\ \Rightarrow \quad & \textstyle\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} b_1 - b_2 \sum_{i=1}^{n} x_i = 0 \\ \Rightarrow \quad & n\bar{y} - nb_1 - b_2 n\bar{x} = 0 \\ \Rightarrow \quad & \bar{y} - b_1 - b_2\bar{x} = 0, \end{aligned}$$

so that $b_1 = \bar{y} - b_2\bar{x}$, as stated in (7). Second, plugging (7) into (6) yields

$$\begin{aligned} & \textstyle\sum_{i=1}^{n} x_i(y_i - [\bar{y} - b_2\bar{x}] - b_2 x_i) = 0 \\ \Rightarrow \quad & \textstyle\sum_{i=1}^{n} x_i(y_i - \bar{y}) = b_2 \sum_{i=1}^{n} x_i(x_i - \bar{x}) \\ \Rightarrow \quad & \textstyle\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = b_2 \sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}), \end{aligned}$$

and solving for b_2 yields (8). Note that the last line follows since in general

$$\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z}) = \sum_{i=1}^{n} x_i(z_i - \bar{z}) - \bar{x}\sum_{i=1}^{n}(z_i - \bar{z}) = \sum_{i=1}^{n} x_i(z_i - \bar{z}) \text{ as } \sum_{i=1}^{n}(z_i - \bar{z}) = 0.$$

The second-order conditions to ensure a minimum rather than a maximum will be verified in matrix case.

3.4 Prediction

The regression line can be used to predict values of y for given values of x. For $x = x^*$ the **prediction** is

$$\widehat{y} = b_1 + b_2 x^*. \tag{9}$$

For example, for the house price example we predict that a house of size 2000 square feet will sell for $\$263,000$, since $\widehat{y} \simeq 115000 + 74 \times 2000 = 263000$.

Such predictions are more reliable when forecasts are made for x values not far outside the range of the x values in the data. And the better the fit of the model, that is the higher the R-squared, the better will be the forecast. Prediction can be **in-sample**, in which case $\widehat{y}_i = b_1 + b_2 x_i$ is a prediction of y_i, $i = 1, ..., n$. If prediction is instead **out-of-sample** it becomes increasingly unreliable the further the prediction point x^* is from the sample range of the x values used in the regression to estimate b_1 and b_2.

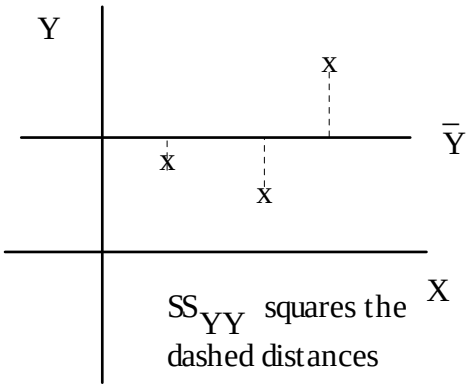

Figure 4: Total Sum of Squares: Variability around the sample mean.

3.5 R-Squared for Goodness of Fit

The regression line does not fit perfectly. The standard measure of closeness of the data points to the fitted regression line is **R-squared**, also called the **coefficient of determination**. This is a number between 0 and 1 that indicates the proportion of the variability in y, about its sample mean $\bar{y}$, explained by regression on x. If $R^2 = 1$ then all the variability is explained and the fit is perfect. If $R^2 = 0$ there is no explanation at all.

We begin by defining measures of variability for the data y before and after regression. For variability in y before regression we use the **total sum of squares**,

$$\text{TSS} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

This is the sum of squared deviations of the data points around the sample mean $\bar{y}$, as displayed in Figure 4.

As measure of the variability in y after regression we use the **error sum of squares**

$$\text{ESS} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2,$$

This is the sum of squared deviations of the data points around the value $\widehat{y}_i$ predicted by the regression line, as displayed in Figure 5. The error sum of squares is also called the **residual sum of squares**.

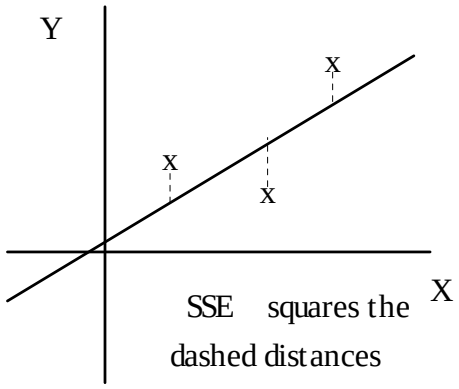

Figure 5: Error Sum of Squares: Variability around the regression line

The **R-squared** is defined as

$$R^2 = 1 - \frac{\text{Error SS}}{\text{Total SS}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{10}$$

Clearly $R^2 = 1$ if the model fit is perfect as then Error SS $= 0$.

It can be shown that R^2 can be equivalently defined in terms of deviations of the fitted values from the sample mean, the **regression sum of squares**, compared to deviations of the data from the sample mean (provided the model includes an intercept). Thus

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}} = \frac{\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \tag{11}$$

The regression sum of squares is also called the **explained sum of squares**. This leads to the interpretation of R^2 giving the proportion of the variability in y explained by regression on x. Equivalently, since s_y^2 =TSS$/(n-1)$, the R^2 gives the fraction of the variance explained by regression on x, and $100 \times R^2$ gives the percentage of the variance of y explained by variation in x.

Output from the Stata command regress given earlier includes Regression SS (called Model SS), error SS (called Residual SS) and total SS. It yields

$$R^2 = 1 - \frac{1.4975 \times 10^{10}}{3.9146 \times 10^{10}} = 0.6175.$$

Thus 61.75 percent of the variation in house price is associated with variation in house size. This is viewed as a good fit, though still with room for improvement as can be seen from Figure 2.

A variation on R^2 is $\bar{R}^2$, the **adjusted** R^2, which is defined later.

3.6 Correlation

A standard way to measure association between x and y, one that predates regression, is the correlation coefficient.

The **sample covariance** between x and y is defined to be

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

This statistic is positive if x and y tend to move together in the same direction, and negative if x and y tend to move together in the opposite direction.

To see this, note that $s_{xy} > 0$ if the cross-product $(x_i - \bar{x})(y_i - \bar{y})$ is mostly positive. This happens if most observations have both $(x_i - \bar{x}) > 0$ and $(y_i - \bar{y}) > 0$ or both $(x_i - \bar{x}) < 0$ and $(y_i - \bar{y}) < 0$. This is positive association since above-average values of x tend to be associated with above-average values of y, and below-average values of x tend to be associated with below-average values of y.

The situation is illustrated in Figure 6 for the house price and house size data. The vertical line is $\bar{x} = 1883$, and the horizontal line is $\bar{y} = 253,910$. The top-right quadrant, denoted (+), has positive value of $(x_i - \bar{x})(y_i - \bar{y})$ since in this quadrant $(x_i - \bar{x}) > 0$ and $(y_i - \bar{y}) > 0$. Similar considerations lead to the signs in the other three quadrants. The covariance is positive, as most of the observations lie in the two positive quadrants. In fact, for these data $s_{xy} = 11298109.39 > 0$.

The sample covariance between x and itself equals the sample variance of x, since $s_{xx} = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = s_x^2$. Similarly $s_{yy} = s_x^2$.

A weakness of the sample covariance is that its magnitude its not easily interpreted. For the house price and house size data $s_{xy} = 11,298,109.39$. This is a large number, but it does not necessarily imply that the association between x and y is large since s_{xy} is not scale-free.

The **sample correlation coefficient** is a transformation of the sample covariance that is a standardized or unitless measure of association between x and y. It is defined by

$$\begin{aligned} r_{xy} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \\ &= \frac{s_{xy}}{\sqrt{s_{xx} \times s_{yy}}}, \end{aligned}$$

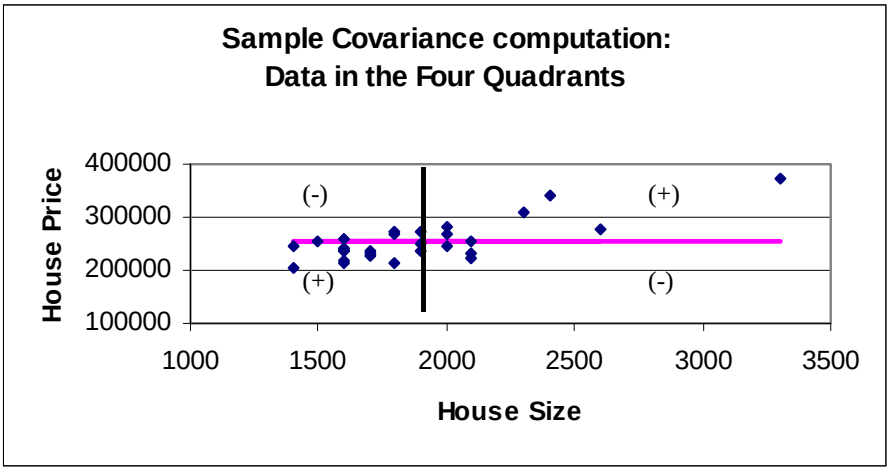

Figure 6: Sample Covariance Computation.

where the second second equality follows by division of both numerator and denominator by $(n-1)$.

The advantage of the correlation coefficient is that it can be shown to bounded between -1 and 1. The correlation coefficient equals one if x and y move exactly together in the same direction, and it equals minus one if x and y move exactly together in the opposite direction. In summary

$r_{xy} = 1$	perfect positive correlation
$0 < r_{xy} < 1$	positive correlation
$r_{xy} = 0$	no correlation
$-1 < r_{xy} < 0$	negative correlation
$r_{xy} = -1$	negative correlation

The correlation coefficient treats x and y symmetrically. It is clear from the definition that $r_{xy} = r_{yx}$. And while the correlation coefficient detects association, it is neutral on whether it is x that is causing y or y that is causing x.

For the house data $r_{xy} = 0.7857$, so there is high positive association between house sale price and house size.

3.7 Correlation versus Regression

The slope coefficient can be re-expressed in terms of the sample correlation coefficient as

$$b_2 = r_{xy} \times \sqrt{\frac{s_{yy}}{s_{xx}}}.$$

Thus if there is positive correlation, i.e. $r_{xy} > 0$, then the slope coefficient b_2 is positive, and similarly $r_{xy} < 0$ implies a negative slope coefficient. So the sample correlation coefficient and the slope coefficient always lead to the same conclusion regarding whether the association between the two variables is positive or negative. In economic data analysis regression is the most commonly-used method. In some other disciplines correlation analysis is more commonly used. The two methods lead to exactly the same conclusions regarding association between x and y.

It can be shown that

$$R^2 = r_{xy}^2,$$

i.e. R^2 equals the square of the sample correlation coefficient between y and x. Thus one by-product of regression analysis is to obtain the squared sample correlation coefficient. The definition of R^2 in terms of sums of squares, rather than in terms of the correlation coefficient, has the advantage of an easy physical interpretation. R^2 can also be easily extended to regression with additional regressors.

4 Moving from Sample to Population

Regression curve fitting is relatively easy. Now we move on to the more difficult topic of extrapolation from the sample to the population.

Recall that in univariate statistics the sample mean $\bar{x}$ was used to make inference about the population mean μ. Similarly the sample fitted regression line $b_1 + b_2 x$ can be used to make inference about the true population line, which is defined below and is denoted $\beta_1 + \beta_2 x$.

Different samples will lead to different fitted regression lines, due to different random departures in the data from the line. If the slope is greater than zero in our single sample, suggesting that y increases as x increases, does this necessarily hold in the population? Or is it just this particular sample that has this relationship?

For example, interest may lie in making statements about the relationship between earnings and education for all 30-year old males in the United States, given a sample of two hundred observations. If we observe a positive

association in our sample, how confident can we be that such a relationship really exists in the population, and is not just an artifact of the particular sample drawn?

4.1 Population and Sample

Statistical inference is based on sampling theory that is in turn based on assumptions about the population, or equivalently about the process generating the data. Without such assumptions statistical inference is not possible. And with different assumptions the statistical inference may need to be adapted.

4.2 Population Assumptions

We suppose that, given x, the data on y are generated by the model

$$y = \beta_1 + \beta_2 x + u. \tag{12}$$

There are two components. First a **population line** with formula $\beta_1 + \beta_2 x$. The parameters β_1 and β_2, where β is the Greek letter beta, denote unknown population values of the intercept and slope. Second, **randomness** is introduced through an **error term** or disturbance that is denoted u (some authors instead use ε, the Greek letter epsilon).

The OLS coefficients b_1 and b_2 (later denoted $\widehat{\beta}_1$ and $\widehat{\beta}_2$) are sample estimates of the population parameters β_1 and β_2. This is directly analogous to the univariate case where $\bar{x}$ denotes the sample estimate of the population mean μ.

For simplicity we assume that **the regressors x are nonstochastic** (i.e. fixed not random). This assumption is relaxed for multiple regression. The following **population assumptions** are made:

1. The population model is the linear model, so $y_i = \beta_1 + \beta_2 x_i + u_i$.
2. There is variation in the x variables, so $\sum_i (x_i - \bar{x})^2 \neq 0$.
3. The error has mean zero: $\mathrm{E}[u_i] = 0$.
4. The errors for different observations have constant variance $\mathrm{V}[u_i] = \sigma_u^2$.
5. The errors for different observations are independent $\mathrm{Cov}[u_i, u_j] = \sigma_u^2$.
6. The errors are normally distributed.

More succinctly, assumptions 3-6 together imply that the errors are independently and identically normally distributed with mean 0 and variance σ_u^2, or

$$u_i \sim \mathcal{N}[0, \sigma_u^2]. \tag{13}$$

Adding in assumption 1 we have that the y are independently and identically distributed with

$$y_i \sim \mathcal{N}[\beta_1 + \beta_2 x_i, \sigma_u^2]. \tag{14}$$

The setup is therefore very similar to that for univariate statistical inference. The one change is that the population mean of y is no longer a constant μ. Instead it varies with the value of x. Formally, the **population regression line** or the **conditional mean of y given x** is

$$\mathrm{E}[y_i|x_i] = \beta_1 + \beta_2 x_i. \tag{15}$$

$\mathrm{E}[y|x]$ varies with x, whereas assumption 4 implies that the conditional variance of y given x is a constant σ_u^2 which does not vary with x. In the special case that $\beta_2 = 0$ the population model simplifies to that used for univariate statistics, with $\mu = \beta_1$.

4.3 Example of Population versus Sample

As an example, let the data generating process be $y = 1 + 2x + \varepsilon$ where $\varepsilon \sim \mathcal{N}[0, 4]$. Suppose $x_1 = 1$ and that the first draw from the $\mathcal{N}[0, 4]$ distribution is $\varepsilon_1 = -4.55875$. Then $y_1 = 1 + 2 \times 1 - 4.55875 = -1.55875$. Other observations are similarly obtained. If $x_2 = 2$ and $\varepsilon_2 = 1.00969$ then $y_2 = 6.00969$, and so on. Suppose that this data generating process yields the following sample of five observations

Observation	x	$\varepsilon \sim \mathcal{N}[0, 4]$	$y = 1 + 2x + \varepsilon$
$i = 1$	$x_1 = 1$	$\varepsilon_1 = -4.55875$	$y_1 = -1.55875$
$i = 2$	$x_2 = 2$	$\varepsilon_2 = 1.00969$	$y_2 = 6.00969$
$i = 3$	$x_3 = 3$	$\varepsilon_3 = -1.31399$	$y_3 = 5.68601$
$i = 4$	$x_4 = 4$	$\varepsilon_4 = 4.04800$	$y_4 = 13.04800$
$i = 5$	$x_5 = 5$	$\varepsilon_5 = -0.62484$	$y_5 = 10.37516$

Figure 7 presents the true regression line $\mathrm{E}[y|x] = 1 + 2x$ and the five generated observations for $y = 1 + 2x + \varepsilon$. Note that the sample points deviate from the true regression line, due to the error term ε.

For the five data points given in Figure 7 the formulae for the least squares slope and intercept estimates lead to $b_2 = 3.0906$ and $b_1 = -2.5598$.

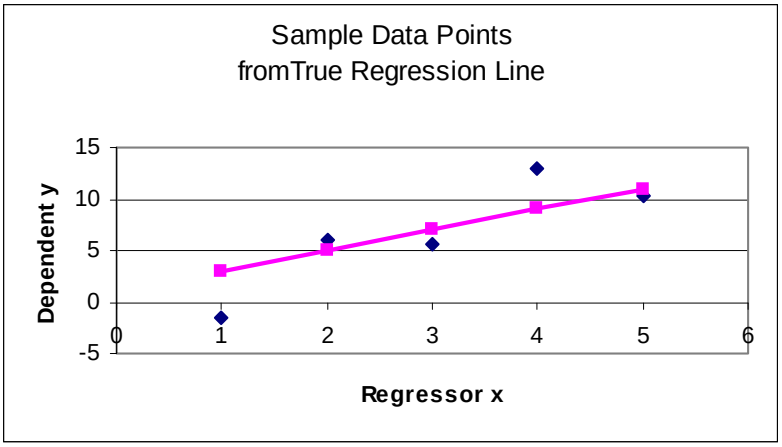

Figure 7: Population regression line $\beta_0 + \beta_1 x$ and five sample data points generated from this line using $y = \beta_0 + \beta_1 x + \varepsilon$ for $x = 1$, 2, 3, 4 and 5 and corresponding $\varepsilon = -4.55875$, 1.00969, -1.31399, 4.04800 and -0.62484.

So the fitted regression line $b_1 + b_2 x = -2.5598 + 3.0906x$ differs from the true regression line $\beta_1 + \beta_2 x = 1 + 2x$. This is shown in Figure 8 where the solid line is the true regression line and the dashed line is the fitted regression line.

The fitted regression line clearly differs from the true regression line, due to sampling variability. This chapter is concerned with inference on the true regression line, controlling for this sampling variability.

5 Finite Sample Properties of the OLS Estimator

We focus on the slope coefficient.

5.1 A Key Result

We consider the OLS slope coefficient

$$b_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{16}$$

as an estimator of population slope coefficient β_2.

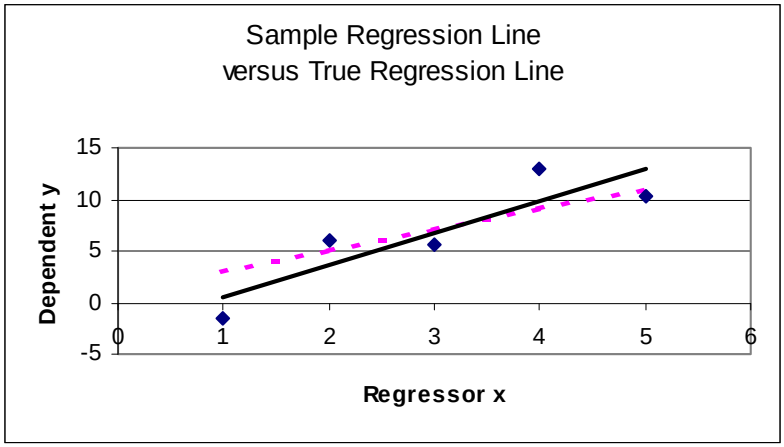

Figure 8: Population regression line $\beta_0 + \beta_1 x = 1 + 2x$ is the solid line. The fitted regression line for five sample data points is $b_1 + b_2 x = -2.5598 + 3.0906x$ and is given by the dashed line.

The key result is that given assumptions 1-2, the OLS slope coefficient b_2 can be expressed as β_2 plus a weighted sum of the errors u_i:

$$b_2 = \beta_2 + \sum_{i=1}^{n} w_i u_i, \tag{17}$$

where the weights

$$w_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{18}$$

For proof see the appendix. It follows that if we condition on the $x_i's$, so that the weights w_i are constants, then the properties of b_2 follow directly from the assumptions on u_i.

5.2 Unbiasedness of OLS Slope Estimate

Assumption 3, that $\mathrm{E}[u_i] = 0$, implies that $\mathrm{E}[w_i u_i] = 0$ so

$$\mathrm{E}[b_2] = \beta_2, \tag{19}$$

see the appendix.

Thus b_2 has the attractive property that it is **unbiased** for β_2. This means that if we had many samples of size n on y and x, and computed b_2

for each sample, then the average of the estimates of b_2 equals β_2. More simply, on average b_2 equals β_2. The inference problem is that for any particular sample b_2 will differ from β_2.

5.3 Variance of OLS Slope Estimate

Assumptions 4 and 5 then yield that

$$\mathrm{V}[b_2] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \tag{20}$$

see the appendix.

The OLS estimator has some desirable properties. It will be shown in the multiple regression section that this variance is the smallest among all linear unbiased estimators (of the form $\sum a_i y_i$) under assumptions 1-5. Under assumptions 1-6 this variance is the smallest among all unbiased estimators so OLS is then the best estimator.

5.4 Standard Error of OLS Slope Estimate

The formula for $\mathrm{V}[b_2]$ depends on the unknown parameter σ_u^2. An unbiased estimate for σ_u^2 is

$$s_u^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \widehat{y}_i)^2, \tag{21}$$

a result that will be proven for multiple regression. The square root, s_u, is called the **standard error of the regression** or the **root MSE**.

Replacing σ_u^2 by s_u^2 in (20) and taking the square root yields the **standard error of** b_2

$$\mathrm{se}[b_2] = \sqrt{\frac{s_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}, \tag{22}$$

where the term standard error means estimated standard deviation.

The standard error $\mathrm{se}[b_2]$ measures the precision of b_2 as an estimate of β_2, so precision is better:

1. the closer the data are to the true regression line (then s_u^2 is small);
2. the larger is the sample size n (then $\sum_i (x_i - \bar{x})^2$ is larger);
3. the more widely scattered are the regressors x (then $\sum_i (x_i - \bar{x})^2$ is larger).

The second property leads to an inverse square root rule for precision. If the sample size n quadruples, the sum $\sum_i (x_i - \bar{x})^2$ quadruples, its reciprocal is one-fourth as large and taking the square root $\text{se}[b_2]$ is halved. More generally, if the sample is m times larger then the standard error of b_2 is $1/\sqrt{m}$ times as large.

6 Finite Sample Inference

We consider linear regression in finite samples with normal errors. As for univariate statistics, hypothesis tests and confidence intervals are based on a t-statistic.

6.1 The t-statistic

So far we have shown that under assumptions 1-5 b_2 has mean β_2 and variance $\text{V}[b_2]$ given in (20). If additionally the errors u_i are normally distributed then b_2 is normal, since $b_2 = \beta_2 + \sum_{i=1}^n w_i u_i$ is then a linear combination of normals. Thus under assumptions 1-6

$$b_2 \sim \mathcal{N}\left[\beta_2, \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]. \tag{23}$$

This implies that

$$z = \frac{b_2 - \beta_2}{\sqrt{\sigma_u^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} \sim \mathcal{N}[0,1]. \tag{24}$$

This statistic depends on the unknown σ_u^2. Replacing σ_u^2 by s_u^2 defined in (20) leads to a slight change in the distribution, with

$$t = \frac{b_2 - \beta_2}{\sqrt{s_u^2 / \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{b_2 - \beta_2}{\text{se}[b_2]} \sim t(n-2). \tag{25}$$

Proof is deferred to multiple regression.

Here $t(n-2)$ denotes Students t-distribution with $(n-2)$ degrees of freedom. The $t(n-2)$ density is similar to a standard normal density. It is bell-shaped with mean 0 but has fatter tails with mean $n/(n-2) > 1$. In Stata the command ttail(df, t) gives $\Pr[T > t | T \sim t(df)]$. The command invttail$(df, a)$ gives the reverse. e.g. ttail$(27, 2.0) = 0.028$ and invttail$(27, 0.05) = 1.703$.

6.2 Confidence Intervals

Different samples yield regression lines with different slopes. We consider confidence intervals for the population slope coefficient β_2, based on the sample estimate b_2. A $100(1-\alpha)$ **percent confidence interval for** β_2 is

$$b_2 \pm t_{\alpha/2;n-2} \times \text{se}[b_2], \tag{26}$$

where $\text{se}[b_2]$ is defined in (22) and $t_{\alpha/2;n-2}$ varies with the level of confidence and denotes the value t^* such that the $\Pr[T_{n-1} > t^*] = \alpha/2$ where T_{n-2} is t distributed with $(n-2)$ degrees of freedom. In Stata $t_{\alpha/2;n-2}$ can be calculated using the command invttail$(n-2, \alpha/2)$.

To obtain this result, note that

$$\begin{array}{lll}
& \Pr\left[-t_{\alpha/2} < T < t_{\alpha/2}\right] = 1-\alpha & \text{for general } T \\
\Rightarrow & \Pr\left[-t_{\alpha/2} < \frac{b_2-\beta_2}{\text{se}[b_2]} < t_{\alpha/2}\right] = 1-\alpha & \text{given (25)} \\
\Rightarrow & \Pr\left[-t_{\alpha/2} \times \text{se}[b_2] < b_2 - \beta_2 < t_{\alpha/2} \times \text{se}[b_2]\right] = 1-\alpha & \text{multiply by se}[b_2] \\
\Rightarrow & \Pr\left[-b_2 - t_{\alpha/2} \times \text{se}[b_2] < -\beta_2 < -b_2 + t_{\alpha/2} \times \text{se}[b_2]\right] = 1-\alpha & \text{subtract } b_2 \\
\Rightarrow & \Pr\left[b_2 + t_{\alpha/2} \times \text{se}[b_2] > \beta_2 > b_2 - t_{\alpha/2} \times \text{se}[b_2]\right] = 1-\alpha & \text{multiply by } -1 \\
\Rightarrow & \Pr\left[b_2 - t_{\alpha/2} \times \text{se}[b_2] < \beta_2 < b_2 + t_{\alpha/2} \times \text{se}[b_2]\right] = 1-\alpha & \text{rearrange}
\end{array}$$

A common example is a 95 percent confidence interval, in which case $(1-\alpha) = 0.95$, so $\alpha = 0.05$ and the critical value is $t_{.025;n-2}$. Furthermore $t_{.025;n-2} \simeq 2$ for most values of n, so a **95 percent confidence interval** is approximately

$$b_2 \pm 2 \times s_{b_2},$$

or the slope coefficient plus or minus two standard errors.

As an example, consider the house price and size data for Central Davis. The confidence interval can be computed from first principles as follows. The output gives $b_2 = 73.77$ and $\text{se}[b_2] = 11.17$, while $t_{\alpha/2;n-2} = t_{.025;27} = 2.052$. So the 95 percent confidence interval is

$$b_2 \pm t_{.025;27} \times s_{b_2} = 73.77 \pm 22.93 = (50.84,\ 96.70).$$

As expected, this is the interval given by Stata in Section 3.

6.3 Hypothesis Tests

Estimation yields a sample estimate of the population slope coefficient β_2 or a confidence interval of values for β_2. Hypothesis tests ask whether a specified value of β_2 is plausible, given the sample estimate b_2.

As an example, consider the claim that the population mean sale price of a house increases by \$50 per additional square foot in size, i.e. that $\beta_2 = 50$. This would be of interest if enlarging an existing house costs \$50 per square foot. The regression slope coefficient $b_2 = 73.77$ for a sample of 29 houses. Is this far enough away from 50 to reject the hypothesis that $\beta_2 = 50$? Or could this difference from 50 be merely an artifact of sampling variability?

For hypothesis tests on the slope coefficient the test statistic used is the t **statistic**

$$t_2 = \frac{b_2 - \beta_2^*}{\text{se}[b_2]}. \tag{27}$$

where β_2^* is the hypothesized value of β_2 under H_0. Large values of t support rejection of H_0, since they arise if b_2 differs greatly from the hypothesized value β_2^*. If indeed $\beta_2 = \beta_2^*$ then from (27) the statistic t_2 is $t(n-2)$ distributed, making it possible to obtain p values and critical values for the t test. We distinguish between two-sided and one-sided tests.

6.4 Two-Sided Tests

Let β_2^* be a hypothesized value for β_2, such as $\beta_2^* = 0$ or $\beta_2^* = 50$. A **two-sided test** or **two-tailed test** on the slope coefficient is a test of

$$\begin{aligned} & H_0 : \beta_2 = \beta_2^* \\ \text{against} \quad & H_a : \beta_2 \neq \beta_2^*, \end{aligned} \tag{28}$$

where H_0 denotes the null hypothesis and H_a denotes the alternative hypothesis. The t **statistic** t_2 defined in (27) is used. This statistic is distributed as t_{n-2} under the null hypothesis that $\beta_2 = \beta_2^*$ and assumptions 1-6 hold.

The statistic $t_2 \neq 0$ if the sample slope estimate $b_2 \neq \beta_2^*$, the hypothesized value of β_2. There are two reasons this may occur. It may be that $\beta_2 \neq \beta_2^*$, or it may be that $\beta_2 = \beta_2^*$ but due to sampling variability $b_2 \neq \beta_2^*$. Hypothesis testing proceeds as follows:

- Assume that $\beta_2 = \beta_2^*$, i.e. the null hypothesis is true.
- Obtain the probability (or significance level) of observing a t statistic equal to or greater than the sample value, where this probability is calculated under the assumption that $\beta_2 = \beta_2^*$.
- Reject the null hypothesis only if this probability is quite small.

Given a computed t statistic there are two ways to implement this significance level approach to testing. We present in order the p value approach and the critical value approach. They produce exactly the same conclusion.

6.4.1 Rejection using p values

As an example, suppose $t = 1.5$ for a sample of size 29. Since the t statistic is t distributed with $n-2$ degrees of freedom, the probability of observing a value of 1.5 or larger in absolute value is $\Pr[|T_{27}| > 1.5] = 0.145$, using $\text{ttail}(27, 1.5) = 0.0726$. Thus while $t = 1.5 \neq 0$ suggests that $\beta_2 \neq \beta_2^*$ there is a nontrivial probability that this departure from 0 is by chance, since even if $b_2 = \beta_2^*$ we expect the t statistic to exceed 1.5 with probability 0.145.

More generally $H_0 : \beta_2 = \beta_2^*$ is rejected if the t statistic is large in absolute value. The probability of just rejecting H_0 given a calculated value t is called the p **value**, with

$$p = \Pr[|T_{n-2}| \geq |t|\ ], \tag{29}$$

where $T_{n-2} \sim t_{n-2}$ and t is the sample value of the t statistic. The quantity $\Pr[|T| \geq |t|]$ equals $\Pr[T \geq t]$ for $t > 0$ and $\Pr[T \leq t]$ for $t < 0$. The decision rule is then

$$\text{Reject } H_0 \text{ if } p < \alpha,$$

where α is the significance level of the test. We do not reject the null hypothesis at level α if $p \geq \alpha$. In Stata $p = 2{*}\text{ttail}(n-2, |t|)$.

The standard testing approach is to take the conservative stance of rejecting H_0 only if the p value is low. The most common choice of significance level is $\alpha = .05$.

6.4.2 Rejection using critical values

The p value approach requires access to a computer, in order to precisely compute the p value. An alternative approach requires only tables of the t distribution for selected values of α, and was the method used before the advent of ready access to computers.

This alternative approach defines a **critical region**, which is the range of values of t that would lead to rejection of H_0 at the specified significance level α. H_0 is rejected if the computed value of t falls in this range.

For a two-sided test of $H_0 : \beta_2 = \beta_2^*$ and for specified α, the **critical value** is c such that $\Pr[|T_{n-2}| \geq c] = \alpha$, or equivalently

$$c = t_{\alpha/2;n-2}, \tag{30}$$

where α is the pre-chosen significance level, often $\alpha = 0.05$. The decision rule is then

$$\text{Reject } H_0 \text{ if } |t| > c$$

and do not reject otherwise. In Stata $c = \text{invttail}(n-2, \alpha/2)$.

6.4.3 Example of Two-sided Test

Consider a test of whether or not house prices increase by \$50 per square foot, i.e. whether $\beta_2 = 50$ in a regression model of house price and house size. The Excel output cannot be immediately used, unlike the simpler test of $\beta_2 = 0$. Instead the appropriate t statistic must be calculated. This is

$$t = (b_2 - \beta_2^*)/s_{b_2} = (73.77 - 50)/11.17 = 2.13.$$

The p value is $p = \Pr[|T| >= 2.13] = 0.042$, using $2\times\text{ttail}(27, 2.13) = 0.042$. H_0 is rejected at significance level 0.05, as $0.042 < 0.05$.

Alternatively, and equivalently, the critical value $c = t_{.025;27} = 2.052$, using $\text{invttail}(27, 0.025) = 2.052$. Again H_0 is just rejected, since $t = 2.13 \geq 2.052$.

The conclusion, at significance level $\alpha = .05$, is that the effect of an extra square foot is to increase the sale price by an amount other than \$50.

6.4.4 Relationship to Confidence Interval

Two-sided tests can be implemented using confidence intervals. If the null hypothesis value β_2^* falls inside the $100(1-\alpha)$ percent confidence interval then do not reject H_0 at significance level α. Otherwise reject H_0 at significance level α.

For the house price data, the 95 percent confidence interval for β_2 is $(50.84, 96.70)$. Since this interval does not include 50 we reject $H_0 : \beta_2 = 50$ at significance level 0.05.

6.5 One-Sided Tests

Again let β_2^* be a specified value for β_2, such as $\beta_2^* = 0$ or $\beta_2^* = 50$. One-sided or one-tailed tests can be either an **upper one-tailed alternative test**

$$\begin{aligned} & H_0 : \beta_2 \leq \beta_2^* \\ \text{against} \quad & H_a : \beta_2 > \beta_2^*, \end{aligned}$$

or a **lower one-tailed alternative test**

$$\begin{aligned} & H_0 : \beta_2 \geq \beta_2^* \\ \text{against} \quad & H_a : \beta_2 < \beta_2^*. \end{aligned}$$

For one-sided tests it is not always clear which hypothesis should be set up as the null and which as the alternative. The convention for one-sided

tests is to **specify the alternative hypothesis to contain the statement being tested.** For example, if it is asserted that house prices increase by at least \$50 for each extra square foot of house, one tests $H_0 : \beta_2 \leq 50$ against $H_a : \beta_2 > 50$. This counterintuitive approach of determining H_0 and H_a was justified in detail in the univariate case and is not repeated here. Essentially setting up the problem this way makes it more difficult to confirm the claim.

Inference is based on the same calculated t statistic $t = (b_2 - \beta_2^*)/s_{b_2}$ as used for two-sided tests.

6.5.1 Upper one-tailed test

For an upper one-tailed test, large positive values of t are grounds for rejection of H_0, as then b_2 is much larger than β_2^*, suggesting that $\beta_2 > \beta_2^*$. Thus

$$\begin{aligned} p &= \Pr[T_{n-2} >= t] \\ c &= t_{\alpha;n-2}. \end{aligned}$$

H_0 is rejected at significance level α if $p < \alpha$ or if $t > c$.

In Stata, the p value can be computed using $p = \text{ttail}(n-2, t)$. The critical value is $c = \text{invttail}(n-2, \alpha)$.

6.5.2 Lower one-tailed test

For a lower one-tailed test, large negative values of t lead to rejection of H_0, as then b_2 is much less than β_2^*, suggesting that $\beta_2 < \beta_2^*$. Thus

$$\begin{aligned} p &= \Pr[T_{n-2} <= t] \\ c &= -t_{\alpha;n-2}. \end{aligned}$$

H_0 is rejected at significance level α if $p < \alpha$ or if $t < c$.

For lower one-tailed tests t is usually negative. In Stata we use $p = 1 - \text{ttail}(n-2, t)$. The critical value is $c = -\text{invttail}(n-2, \alpha)$.

6.5.3 Example of One-Sided Test

Suppose the claim is made that house prices increase by at least \$50 per additional square foot. Then the claim is made the alternative hypothesis and the appropriate test is an upper one-sided test of $H_0 : \beta_2 \leq 50$ against $H_a : \beta_2 > 50$.

The t statistic is the same as in the two-sided case, so $t = 2.13$. Now $p = \Pr[T >= 2.13] = 0.021$, using Stata function ttail(27,2.13). So H_0 is strongly rejected at significance level 0.05. Equivalently, the critical value $c = t_{.05;27} = 1.703$, using invttail$(27, 0.05)$. H_0 is again rejected since $t = 2.13 \geq 1.703$.

We therefore reject H_0 and conclude that, at significance level $\alpha = .05$, the effect of an extra square foot is to increase the sale price by at least \$50.

Note that compared to the similar two-sided test at level $\alpha = 0.05$ the rejection is stronger in the one-sided test. This is because it is easier to determine that β_2 lies in the narrower alternative hypothesis region $\beta_2 > 50$ than in the broader region $\beta_2 \neq 50$.

6.6 Tests of Statistical Significance

The most common hypothesis test is a test of whether or not a slope parameter equals zero.

This is called a **test of the statistical significance of a regressor**. It answers the question of whether or not x has an effect on y in the population. If it has no population effect, then clearly $\beta_2 = 0$ and the population regression model $y = \beta_1 + \beta_2 x + \varepsilon$ reduces to $y = \beta_1 + \varepsilon$, so that y bounces around a mean value of β_1.

Formally the test is a two-sided test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$.Regression packages automatically print out statistics to enable tests of this hypothesis. In particular, many packages give the p value for this test. H_0 is rejected at level α, and statistical significance is confirmed, if $p < \alpha$. If instead $p > \alpha$ then H_0 is not rejected and it is concluded that there is no statistically significant relationship or equivalently, that the regressor is statistically insignificant. The most common choice of α is 0.05, followed by 0.01 and 0.10. e.g. for the house price example in section 2, $p = 0.000 < 0.05$ so house size is a highly statistically significant regressor at significance level 0.05.

In some cases there may be a prior belief that the slope coefficient is positive or that it is negative. Then a one-sided test is used, which usually requires halving the printed p value. Unfortunately there can be ambiguity in the statement that "the regressor is statistically significant at significance level 0.05" as it will not always be clear whether a one-sided or two-sided test was performed. Clearer is to say that a regressor is positively statistically significant if the prior belief that β_2 is positive is supported by an appropriate hypothesis test, and negatively statistically significant if the prior belief that β_2 is negative is supported.

6.7 One-sided versus two-sided tests

One-sided tests and two-sided tests use the same computed value t of the t statistic, but different p values or critical values.

For a one-sided test we use $p = \Pr[T >= t]$ or $p = \Pr[T <= t]$, which is one-half of the two-sided p value of $p = \Pr[|T| >= t]$.

In particular, for one-sided tests of statistical significance one halves the p value reported for a two-sided test. In addition one needs to verify that $b_2 > 0$ for tests against $H_a : \beta_2 > 0$ and that $b_2 < 0$ for tests against $H_a : \beta_2 < 0$.

6.8 Presentation of Regression Results

Published articles can differ in the method of presentation of regression results. They always report the intercept and slope coefficients b_1 and b_2. They usually report model fit using R^2. But there can be great variation in which combinations of the standard error, t statistic (for test that the population coefficient equals zero), and p value are reported. Given knowledge of one of these three, and knowledge of the slope coefficient, it is always possible to compute the other two. For example, given b_2 and s_{b_2}, we can compute $t = b_2/s_{b_2}$ and $p = \Pr[|T_{n-2}| \leq t]$. Similarly, given b_2 and t we can compute $s_{b_2} = b_2 \times t$.

It is easiest if all four of b, s_b, t and p are reported. But for space reasons, especially if there are several different models estimated or if the models have additional regressors, it is quite common for published studies to report only b and one of s_b, t and p.

Thus for the house price regression we might report the coefficients and standard errors

$$\underset{}{HPRICE} = \underset{(21489)}{115017} + \underset{(11.17)}{73.77} \times HSIZE \qquad R^2 = 0.79.$$

Or we may report the coefficients and t statistics for population coefficient equal to zero

$$\underset{}{HPRICE} = \underset{(5.35)}{115017} + \underset{(6.60)}{73.77} \times HSIZE \qquad R^2 = 0.79.$$

Or just the coefficients and p values may be reported

$$\underset{}{HPRICE} = \underset{(0.000)}{115017} + \underset{(0.000)}{73.77} \times HSIZE \qquad R^2 = 0.79.$$

Using any of these three alternatives we can verify that the slope coefficient is statistically significant at level 0.05. And while we have focused

on the slope coefficient it is clear from this output that the intercept is also statistically significant at level 0.05.

The p value approach is the simplest testing approach, though some care is needed for one-sided tests. Many studies instead report t statistics. These can be interpreted as follows. Testing is most often done at significance level $\alpha = 0.05$, and $t_{.025} \geq 1.960$ for all degrees of freedom. So a rough guide is that the null hypothesis is rejected at significance level 0.05 for a two-sided test if the t statistic exceeds approximately 2.0 in absolute value. Similarly, for one-sided tests at 5 percent a rough guide is to reject for an upper alternative test if $t > 1.645$, and to reject for an upper alternative test if $t < 1.645$, since $t_{.05} \geq 1.645$.

7 Large Sample Inference

As sample size $n \to \infty$ the least squares estimator b_2 collapses on β_2, since its variance goes to zero. Formally this property is called convergence in probability and the OLS estimator is said to be consistent.

Also it can be shown that as $n \to \infty$ the "t **statistic**"

$$t_2 = \frac{b_2 - \beta_2^*}{\text{se}[b_2]}, \tag{31}$$

is actually standard normal distributed (rather than t-distributed). This requires only assumptions 1-5, i.e. the regression model errors need not be normally distributed. Then inference is based on the $\mathcal{N}[0,1]$ distribution rather than the $t(n-2)$ distribution.

So wherever we use $t_{\alpha;n-2}$ or $t_{\alpha/2;n-2}$ in the preceding we can use z_α or $z_{\alpha/2}$, where z denotes a standard normal variable. This will be detailed for multiple regression.

8 Multivariate Regression

Multivariate regression is a conceptually straightforward. The **regression line** from regression of y on an intercept and x_2, x_3, ... , x_k is denoted

$$\widehat{y} = b_1 + b_2 x_2 + \cdots + b_k x_k. \tag{32}$$

The least squares estimator now minimizes wrt b_1, ..., b_k the sum of squared residuals

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i - \cdots - b_k x_k)^2. \tag{33}$$

The main jump is that it is not possible to express the resulting estimates using summation notation. Instead we use matrix algebra.

9 Summary

Some key equations / results for bivariate regression are

OLS method	b_1, b_2 minimize $\sum_{i=1}^n (y_i - (b_1 + b_2 x_i)^2$
OLS slope estimate	$b_2 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \ / \ \sum_{i=1}^n (x_i - \bar{x})^2$
Population model	$y = \beta_1 + \beta_2 x_2 + u$
Population assumptions	$1 - 6$
Unbiasedness of b_2	$\mathrm{E}[b_2] = 0$
Standard error of b_2	$\mathrm{se}[b_2] = s_u / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$
Standard error of regression	$s_u^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$
Confidence interval for β_2	$\beta_2 \in b_2 \pm t_{n-2;\alpha/2} \times \mathrm{se}[b_2]$
t statistic for $H_0 : \beta_2 = \beta_2^*$	$t = \frac{b_2 - \beta_2}{\mathrm{se}[b_2]} \sim t_{n-2}$.

10 Appendix: Mean and Variance for OLS Slope Coefficient

First show $b_2 = \beta_2 + \sum_{i=1}^n w_i u_i$, where $w_i = (x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$.

b_2	$= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$	
	$= \sum_{i=1}^n (x_i - \bar{x}) y_i / \sum_{i=1}^n (x_i - \bar{x})^2$	as $\sum_{i=1}^n (x_i - \bar{x})\bar{y} = 0$
	$= \sum_{i=1}^n [(x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2] y_i$	rearranging
	$= \sum_{i=1}^n w_i y_i$	for w_i defined above
	$= \sum_{i=1}^n w_i \{\beta_1 + \beta_2 x_i + u_i\}$	by assumption 1
	$= \beta_1 \sum_{i=1}^n w_i + \beta_2 \sum_{i=1}^n w_i x_i + \sum_{i=1}^n w_i u_i$	
	$= \beta_2 + \sum_{i=1}^n w_i u_i$	using $\sum_{i=1}^n w_i = 0$ and $\sum_{i=1}^n w_i x_i = 1$.

The last line uses

$$\sum_{i=1}^n w_i = \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0 \text{ as } \sum_{i=1}^n (x_i - \bar{x}) = 0,$$

and

$$\sum_{i=1}^n w_i x_i = \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} x_i = \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{\sum_{i=1}^n (x_i - \bar{x}) x_i} = 1.$$

Next obtain the mean and variance

$$\begin{aligned}
\mathrm{E}[b_2] &= \mathrm{E}[\beta_2 + \textstyle\sum_{i=1}^n w_i u_i] \\
&= \mathrm{E}[\beta_2] + \mathrm{E}[\textstyle\sum_{i=1}^n w_i u_i] \\
&= \beta_2 + \textstyle\sum_{i=1}^n \mathrm{E}[w_i u_i] && \text{given independence over } i \\
&= \beta_2 + \textstyle\sum_{i=1}^n w_i \mathrm{E}[u_i] && \text{since } w_i \text{ depends on nonstochastic } x \\
&= \beta_2, && \text{since } \mathrm{E}[u_i] = 0 \text{ by ass. 3}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{V}[b_2] &= \mathrm{E}[(b_2 - \beta_2)^2] \\
&= \mathrm{E}[(\textstyle\sum_{i=1}^n w_i u_i)^2] && \text{since } b_2 = \beta_2 + \textstyle\sum_{i=1}^n w_i u_i \\
&= \mathrm{E}[(\textstyle\sum_{i=1}^n w_i u_i)(\sum_{j=1}^n w_j u_j)] && \text{squaring the sum} \\
&= \mathrm{E}[\textstyle\sum_{i=1}^n \sum_{j=1}^n w_i w_j u_i u_j] \\
&= \textstyle\sum_{i=1}^n \sum_{j=1}^n w_i w_j \mathrm{E}[u_i u_j] \\
&= \textstyle\sum_{i=1}^n w_i^2 \mathrm{E}[u_i^2] && \text{if } \mathrm{E}[u_i u_j] = 0, i \neq j \text{ (ass. 4)} \\
&= \textstyle\sum_{i=1}^n w_i^2 \sigma_u^2 && \text{if } \mathrm{E}[u_i^2] = \sigma_u^2 \text{ (ass. 5)} \\
&= \sigma_u^2 \textstyle\sum_{i=1}^n w_i^2 \\
&= \sigma_u^2 / \textstyle\sum_{i=1}^n (x_i - \bar{x})^2 && \text{since } \textstyle\sum_{i=1}^n w_i^2 = 1 / \sum_{i=1}^n (x_i - \bar{x})^2.
\end{aligned}$$

The least line uses

$$\begin{aligned}
\textstyle\sum_{i=1}^n w_i^2 &= \textstyle\sum_{i=1}^n \left[\frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^2 \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}. && \text{simplifying.}
\end{aligned}$$