Chapter 3
Properties and Hypothesis Testing

3.1 Types of data

The regression techniques developed in previous chapters can be applied to three different kinds of data.

1. Cross-sectional data.
2. Time series data.
3. Panel data.

The first consists on observing various economic unit (e.g. firms, countries, households, individuals) at one point in time. For example, we observe the wages, experience and education of many individuals, only once and at all at the same time. The second consists on observing the same economic unit at different point in time. For example, we observe daily stock prices over many years. Finally, the third combines the characteristics of the first and the second. That is, we observe various economic units at repeated points in time. For example, we have information about the inflation, unemployment and GDP of a group of countries and over many years.

3.2 Assumptions of the model

When the regressors in our econometric model are non stochastic, we will make the following six assumptions.

1. The model is linear in the parameters and it is correctly specified.

$$Y = \beta_1 + \beta_2 X + u \tag{3.1}$$

$$Y = \beta_1 X^{\beta_2} + u \tag{3.2}$$

Equation 2.1 is linear in β, while Equation 2.2 is not.

2. There is some variation in the regressor in the sample. We need variation in the variable X to identify the relationship. Consider the OLS estimator for β_2:

$$b_2 = \frac{\sum_{i=0}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=0}^{n}(X_i - \bar{X})^2}. \tag{3.3}$$

   If there is no variation in X, then the denominator is zero and we cannot obtain b_2.
3. The expected value of the disturbance term is zero.

$$E(u_i) = 0 \quad \text{for all} \quad i. \tag{3.4}$$

   Some u_i will be negative, some will be positive, but on average they will be zero. If a constant is included in the model, the condition is satisfied automatically.
4. The disturbance term is homoscedastic.
   Homoscedasticity means that the variance of the error terms u_i is constant across all observations i. Hence, we can write:

$$\sigma_{u_i}^2 = \sigma_u^2 \quad \text{for all} \quad i. \tag{3.5}$$

   Because the error term has zero mean (from assumption 3), then the population variance of u_i is equal to:

$$E(u_i^2) = \sigma_u^2 \quad \text{for all} \quad i. \tag{3.6}$$

   σ_u^2 is a population parameter, therefore it is unknown and need to be estimated.
5. The values of the disturbance terms have independent distributions.

$$u_i \text{ is distributed independently of } u_j \text{ for all } j \neq i. \tag{3.7}$$

   This means that there is no *autocorrrelation* in the error term. This means that the population covariance between u_i and u_j is zero:

$$\sigma_{u_i u_j} = 0. \tag{3.8}$$

   With assumptions 1 through 5, we says that OLS coefficients are BLUE: Best Linear Unbiased Estimators. They are best, because they have the smallest variance across all unbiased estimators.
6. The disturbance term has a normal distribution.

$$u_i \sim N[0, \sigma_u^2] \quad \text{for all} \quad i. \tag{3.9}$$

   The error term is distributed normal with mean zero and variance σ_u^2. This assumption becomes useful at the time of performing t tests, F tests, and constructing confidence intervals for β_1 and β_2 using the regression results. The justification for this assumption depends on the *central limit theorem*. This one state that if a random variable is the composite result of the effects of a large number of other random variables (that are not necessarily normal), it will have an approximately normal distribution.

3.3 Unbiasedness of the coefficients

Recall that an estimator $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$. The expected value of the estimator is equal to the true population parameter. For the slope coefficient in the OLS regression we have:

$$\begin{aligned} b_2 &= \frac{\sum_{i=0}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=0}^{n}(X_i - \bar{X})^2} \\ &= \beta_2 + \frac{\sum_{i=0}^{n}(X_i - \bar{X})u_i}{\sum_{i=0}^{n}(X_i - \bar{X})^2} \\ &= \beta_2 + \sum_{i=1}^{n} a_i u_i \end{aligned} \tag{3.10}$$

where

$$a_i = \frac{(X_i - \bar{X})}{\sum_{i=0}^{n}(X_i - \bar{X})^2}. \tag{3.11}$$

Thus, this shows that b_2 is equal to its true value, β_2, plus a linear combination of the values of the error terms. If we take expectations of b_2 we have:

$$E(b_2) = E(\beta_2) + E\Big(\sum_{i=1}^{n} a_i u_i\Big) = \beta_2 + \sum_{i=1}^{n} E(a_i u_i) = \beta_2 + \sum_{i=1}^{n} a_i E(u_i) = \beta. \tag{3.12}$$

The term a_i goes out of the expectation because a_i is only a function of constant Xs. In addition, the last equality holds because $E(u_i) = 0$. Hence, b_2 is an unbiased estimator of β_2, $E(b_2) = \beta_2$.

3.4 Precision of the coefficients

We are also interested on how precise b_1 and b_2 are in estimating the population parameters β_1 and β_2. A measure of this precision are their population variances, given by:

$$\sigma_{b_1}^2 = \sigma_u^2 \Big(\frac{1}{n} + \frac{\bar{X}}{\sum_{i=0}^{n}(X_i - \bar{X})^2}\Big), \quad \text{and} \tag{3.13}$$

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=0}^{n}(X_i - \bar{X})^2} \tag{3.14}$$

One concern in the implementation of the above formulas is that σ_u^2 is an unknown population parameter and need to be estimated. A natural estimator for this regression variance is the variance of the regression errors. Because the population regression errors u_i are also unknown, we use the sample counterparts e_i and adjust for the corresponding degrees of freedom. Hence, we have:

$$S_u^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2. \tag{3.15}$$

This S_u^2 is the unbiased estimator of σ_u^2, and $n-2$ are the degrees of freedom. We subtract two from the sample size because we are estimating two parameters: the regression constant and one slope coefficient. Then, we use the following formulas to estimate the standard errors of b_1 and b_2:

$$S_{b_1} = \sqrt{S_u^2\Big(\frac{1}{n} + \frac{\bar{X}}{\sum_{i=0}^{n}(X_i-\bar{X})^2}\Big)}, \quad \text{and} \tag{3.16}$$

$$S_{b_2} = \sqrt{\frac{S_u^2}{\sum_{i=0}^{n}(X_i-\bar{X})^2}}. \tag{3.17}$$

3.5 The Gauss-Markov theorem

The *Gauss-Markov theorem* simply states that when assumptions 1 through 5 above are satisfied, the OLS estimators are Best Linear Unbiased Estimators (BLUE) of the regression parameters. Best refers to smallest variance.

3.6 Hypotheses testing

Hypothesis testing is simply a method of making decisions using data. It starts with the formulation of the null and the alternative hypotheses and then uses some test statistics to assess the truth of the null hypothesis.

3.6.1 Formulation of the null hypothesis

The formulation of the null hypothesis starts with a relationship in mind. For example, that the percentage rate of price inflation (p) depends on the percentage rate of wage inflation (w) following the linear equation:

$$p_i = \beta_1 + \beta_2 w_i + u_i \tag{3.18}$$

Then, you want to test the hypothesis that the price inflation is equal to the wage inflation. This is denoted by H_0 and it is know as the *null hypothesis*. In addition, we also define an alternative hypothesis, denoted by H_1 and represents the conclusion of the test if the null hypothesis is rejected. For our example the null and the alternative hypothesis are written as:

$$H_0 : \beta_2 = 1 \tag{3.19}$$

$$H_1 : \beta_2 \neq 1 \tag{3.20}$$

In general, the null and alternative hypotheses are:

$$H_0 : \beta_2 = \beta_2^0 \tag{3.21}$$

$$H_1 : \beta_2 \neq \beta_2^0. \tag{3.22}$$

3.6.2 *t*-tests

Recall that β_2 is unknown and that we have to use the estimate b_2. Then, the decision rule to reject the null hypothesis should compare the estimate b_2 with the hypothesized value β_2^0. Intuitively, if the values are far apart, then there is evidence against the null. This comparison should take into account the fact that b_2 is subject to some sampling variation (it is not the actual β_2). We will use the following statistic:

$$z = \frac{b_2 - \beta_2^0}{\sigma_{b_2}} \tag{3.23}$$

The numerator is just the distance between the regression estimate and the hypothesized value, with the denominator is the standard deviation of b_2, given by the square root of the expression in Equation 3.14. z is the number of standard deviations between b_2 and β_2. For a known σ_{b_2}, this one follows a normal distribution. However σ_{b_2} is unknown and we need to use the estimate of the standard error of b_2. This one is given by S_{b_2} and it is presented in Equation 3.17. Then we use the following t-statistic:

$$t = \frac{b_2 - \beta_2^0}{S_{b_2}} \tag{3.24}$$

To know if the deviations between b_2 and β_2^0 are significantly large, we compare this t-statistic with the critical values from the table t distribution with $n-2$ degrees of freedom. The null hypothesis is not rejected if the following condition is met:

$$-t_{n-2,\alpha/2} \leq \frac{b_2 - \beta_2^0}{S_{b_2}} \leq t_{n-2,\alpha/2} \tag{3.25}$$

Where $t_{n-2,\alpha/2}$ is just the notation of the critical value than comes from the t distribution with $n-2$ degrees of freedom and at significance level α. The *significance*

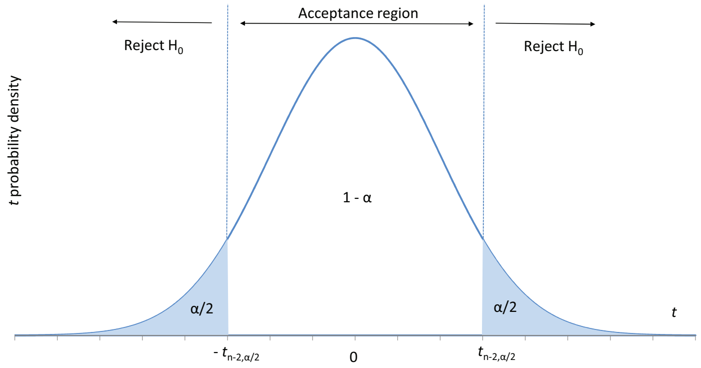


Fig. 3.1 Acceptance region for the *t*-test.

level is the probability that we reject the null hypothesis when in fact it is true. The rejection regions are illustrated in Figure 3.1.

3.6.3 Confidence intervals

The *confidence interval* indicates the reliability of an estimate. The confidence interval for the population parameter β_2 can be derived from Equation 3.25 in the following way:

$$\begin{aligned} 1-\alpha &= P\Big(-t_{n-2,\alpha/2} \leq \tfrac{b_2-\beta_2}{S_{b_2}} \leq t_{n-2,\alpha/2}\Big) \qquad (3.26)\\ 1-\alpha &= P\Big(-t_{n-2,\alpha/2}\cdot S_{b_2} \leq b_2-\beta_2 \leq t_{n-2,\alpha/2}\cdot S_{b_2}\Big)\\ 1-\alpha &= P\Big(b_2-t_{n-2,\alpha/2}\cdot S_{b_2} \leq \quad \beta_2 \quad \leq b_2+t_{n-2,\alpha/2}\cdot S_{b_2}\Big) \end{aligned}$$

The meaning of the above equation is that the population parameter β_2 will be between the lower confidence limit $b_2 - t_{n-2,\alpha/2} \cdot S_{b_2}$ and the upper confidence limit $b_2 + t_{n-2,\alpha/2} \cdot S_{b_2}$ with probability $(1-\alpha)$ or $100 \times (1-\alpha)\%$. The *p values* provide an alternative approach to reporting the significance of regression coefficients or when carrying out more general hypothesis testing. As you can see from Equation 3.25 and Figure 3.1, different significance levels α can yield a different conclusion in the rejection or not of the null hypothesis. The p value of a hypothesis test represent the minimum significance level at which the null is rejected. Then, when the p value is below the significance level α we reject the null.

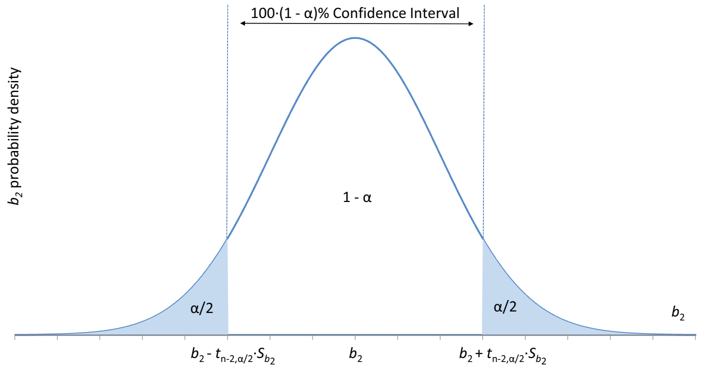


Fig. 3.2 Confidence interval for β_2.

3.6.4 *F test*

A useful tool if we want to test if there is no relationship between X and Y if the F test. In the simple linear regression model with only one slope coefficient, the null and the alternative in an F test are:

$$H_0 : \beta_2 = 0 \tag{3.27}$$

$$H_1 : \beta_2 \neq 0. \tag{3.28}$$

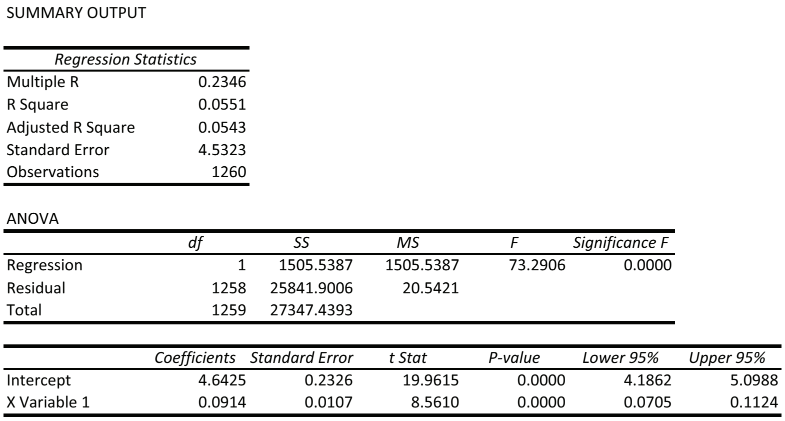
This test is build on the idea of testing how good is the regression model in explaining the variation in Y. In Equation 2.15 we already separated the variation of Y into its ‘explained’ and ‘unexplained’ components. These are:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{3.29}$$

$$TSS = ESS + RSS. \tag{3.30}$$

The total sum of squares (TSS) is the summation of the explained sum of squares (ESS) and the residual sum of squares (RSS). Then, the F *statistic* for goodness of fit of a regression is written as the explained sum of squares, per explanatory variable, divided by the residual sum of squares, per remaining degrees of freedom:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} \tag{3.31}$$



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.2346
R Square	0.0551
Adjusted R Square	0.0543
Standard Error	4.5323
Observations	1260

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1505.5387	1505.5387	73.2906	0.0000
Residual	1258	25841.9006	20.5421		
Total	1259	27347.4393			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	4.6425	0.2326	19.9615	0.0000	4.1862	5.0988
X Variable 1	0.0914	0.0107	8.5610	0.0000	0.0705	0.1124

Fig. 3.3 Regression output in MS Excel.

where k is the total number of coefficients we are estimating, hence $(k-1)$ is the number of slope coefficients. That is, the total number of parameters we are estimating minus the constant parameter. If we divide the numerator and the denominator by TSS, then the F statistics can be written in terms of the R^2 as follows:

$$F = \frac{(ESS/TSS)/(k-1)}{(RSS/TSS)/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \tag{3.32}$$

If this F statistic is greater that the critical value from the table F distribution with $(k-1)$ and $(n-k)$ degrees of freedom, $F_{k-1,n-k}$, we reject the null hypothesis and conclude that the regression model does not significantly explain the variation in variable Y. For the simple regression model with only one slope coefficient, $k=2$, we have:

$$F = \frac{R^2}{(1-R^2)/(n-2)}. \tag{3.33}$$

If this F statistic $> F_{1,n-2}$ we reject the null hypothesis presented in Equation 3.28.

3.7 Computer output

The computer regression output is very similar across different statistical packages. Figure 3.3 shows the output using MS Excel for the estimation of the following simple regression model:

$$wage_{=}\beta_1 + \beta_2 exper_i + u_i \tag{3.34}$$

To obtain the regression estimated coefficients we use Equations 2.4 and 2.5:

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = 0.091 \tag{3.35}$$

$$b_1 = \bar{Y} - b_2\bar{X} = 4.642 \tag{3.36}$$

The total sum of squares, estimates sum of squares, and residual sum of squares are obtained using 2.15 and 2.15:

$$TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = 27347.439 \tag{3.37}$$

$$ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 = 1505.539 \tag{3.38}$$

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 25841.901 \tag{3.39}$$

The regression R^2 comes from Equation 2.18:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} = 0.055 \tag{3.40}$$

From the square root of Equation 3.15:

$$S_u = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} = 4.532 \tag{3.41}$$

Then, the standard errors of the coefficients are computer using Equations 3.17 and 3.17:

$$S_{b_1} = \sqrt{S_u^2\left(\frac{1}{n} + \frac{\bar{X}}{\sum_{i=0}^{n}(X_i - \bar{X})^2}\right)} = 0.233 \tag{3.42}$$

$$S_{b_2} = \sqrt{\frac{S_u^2}{\sum_{i=0}^{n}(X_i - \bar{X})^2}} = 0.011 \tag{3.43}$$

The F statistic uses Equation 3.32:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = 73.291 \tag{3.44}$$

The *t* statistics use Equation 3.24:

$$t = \frac{b_1}{S_{b_1}} = 19.961 \tag{3.45}$$

$$t = \frac{b_2}{S_{b_2}} = 8.561 \tag{3.46}$$

Finally, for the 95% upper and lower confidence levels, we use Equation 3.26:

$$b_1 - t_{n-2,\alpha/2} \cdot S_{b_1} = 4.186 \tag{3.47}$$

$$b_1 + t_{n-2,\alpha/2} \cdot S_{b_1} = 5.099 \tag{3.48}$$

$$b_2 - t_{n-2,\alpha/2} \cdot S_{b_2} = 0.071 \tag{3.49}$$

$$b_2 + t_{n-2,\alpha/2} \cdot S_{b_2} = 0.112 \tag{3.50}$$