Chapter 5

Inference in the Simple Regression Model: Interval Estimation, Hypothesis Testing, and Prediction

Assumptions of the Simple Linear Regression Model

SR1. $y_t = \beta_1 + \beta_2 x_t + e_t$

SR2. $E(e_t) = 0 \;\Leftrightarrow\; E(y_t) = \beta_1 + \beta_2 x_t$

SR3. $\text{var}(e_t) = \sigma^2 = \text{var}(y_t)$

SR4. $\text{cov}(e_i, e_j) = \text{cov}(y_i, y_j) = 0$

SR5. x_t is not random and takes at least two different values

SR6. $e_t \sim N(0, \sigma^2) \;\Leftrightarrow\; y_t \sim N[(\beta_1 + \beta_2 x_t), \sigma^2]$ (*optional*)

From Chapter 4

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2 \sum x_t^2}{T\sum(x_t - \bar{x})^2}\right)$$

$$b_2 \sim N\left(\beta_2, \frac{\sigma^2}{\sum(x_t - \bar{x})^2}\right)$$

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_t^2}{T-2}$$

This Chapter introduces additional tools of statistical inference: **Interval estimation, prediction, prediction intervals, hypothesis testing**.

5.1 Interval Estimation

5.1.1 The Theory

A standardized normal random variable is obtained from b_2 by subtracting its mean and dividing by its standard deviation:

$$Z = \frac{b_2 - \beta_2}{\sqrt{\operatorname{var}(b_2)}} \sim N(0,1) \qquad (5.1.1)$$

The standardized random variable Z is normally distributed with mean 0 and variance 1.

5.5.1a The Chi-Square Distribution

- Chi-square random variables arise when standard normal, $N(0,1)$, random variables are squared.

If $Z_1, Z_2, \ldots, Z_m$ denote m <u>independent</u> $N(0,1)$ random variables, then

$$V = Z_1^2 + Z_2^2 + \ldots + Z_m^2 \sim \chi^2_{(m)} \qquad (5.1.2)$$

- The notation $V \sim \chi^2_{(m)}$ is read as: the random variable V has a *chi-square distribution with m <u>degrees of freedom</u>*.

$$E[V] = E\left[\chi^2_{(m)}\right] = m$$

$$\text{var}[V] = \text{var}\left[\chi^2_{(m)}\right] = 2m \tag{5.1.3}$$

- *V* must be nonnegative, $v \geq 0$
- the distribution has a long tail, or is *skewed*, to the right.
- As the degrees of freedom *m* gets larger the distribution becomes more symmetric and "bell-shaped."
- As *m* gets large the chi-square distribution converges to, and essentially becomes, a normal distribution.

5.5.1b The Probability Distribution of $\hat{\sigma}^2$

- The random error term e_t has a normal distribution, $e_t \sim N(0, \sigma^2)$.
- Standardize the random variable by dividing by its standard deviation so that $e_t / \sigma \sim N(0,1)$.
- $(e_t / \sigma)^2 \sim \chi^2_{(1)}$.
- If all the random errors are independent then

$$\sum_t \left(\frac{e_t}{\sigma}\right)^2 = \left(\frac{e_1}{\sigma}\right)^2 + \left(\frac{e_2}{\sigma}\right)^2 + \cdots + \left(\frac{e_T}{\sigma}\right)^2 \sim \chi^2_{(T)} \qquad (5.1.4)$$

$$V = \frac{\sum_t \hat{e}_t^2}{\sigma^2} = \frac{(T-2)\hat{\sigma}^2}{\sigma^2} \tag{5.1.5}$$

- V does not have a $\chi^2_{(T)}$ distribution because the least squares residuals are *not* independent random variables.
- All T residuals $\hat{e}_t = y_t - b_1 - b_2 x_t$ depend on the least squares estimators b_1 and b_2. It can be shown that only T–2 of the least squares residuals are independent in the simple linear regression model.

$$V = \frac{(T-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(T-2)}$$

- We have *not* established the fact that the chi-square random variable V is statistically independent of the least squares estimators b_1 and b_2, but it is.

5.1.1c The *t*-Distribution

- A "*t*" random variable (no upper case) is formed by dividing a standard normal, $Z \sim N(0,1)$, random variable by the square root of an *independent* chi-square random variable, $V \sim \chi^2_{(m)}$, that has been divided by its degrees of freedom, m.

If $Z \sim N(0,1)$ and $V \sim \chi^2_{(m)}$, and if Z and V are independent, then

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)} \tag{5.1.7}$$

- The *t*-distribution's shape is completely determined by the degrees of freedom parameter, m, and the distribution is symbolized by $t_{(m)}$.

- The t-distribution is less "peaked," and more spread out than the $N(0,1)$.
- The t-distribution is symmetric, with mean $E[t_{(m)}]=0$ and variance $\mathrm{var}[t_{(m)}]=m/(m-2)$.
- As the degrees of freedom parameter $m\to\infty$ the $t_{(m)}$ distribution approaches the standard normal $N(0,1)$.

5.1.1d A Key Result

$$t = \frac{Z}{\sqrt{V/(T-2)}} = \frac{\dfrac{b_2-\beta_2}{\sqrt{\dfrac{\sigma^2}{\sum(x_t-\bar{x})^2}}}}{\sqrt{\dfrac{(T-2)\hat{\sigma}^2}{\sigma^2}\Big/ (T-2)}} = \frac{b_2-\beta_2}{\sqrt{\dfrac{\hat{\sigma}^2}{\sum(x_t-\bar{x})^2}}} \tag{5.1.8}$$

$$= \frac{b_2-\beta_2}{\sqrt{\widehat{\mathrm{var}}(b_2)}} = \frac{b_2-\beta_2}{\mathrm{se}(b_2)}$$

5.1.2 Obtaining Interval Estimates

If assumptions SR1-SR6 of the simple linear regression model hold, then

$$t = \frac{b_k - \beta_k}{\text{se}(b_k)} \sim t_{(T-2)}, \quad k = 1, 2 \tag{5.1.9}$$

For k=2

$$t = \frac{b_2 - \beta_2}{\text{se}(b_2)} \sim t_{(T-2)} \tag{5.1.10}$$

where

$$\widehat{\text{var}}(b_2) = \frac{\hat{\sigma}^2}{\sum (x_t - \overline{x})^2} \quad \text{and} \quad \text{se}(b_2) = \sqrt{\text{var}(b_2)}$$

Find critical values t_c from a $t_{(m)}$ distribution such that

$$P(t \geq t_c) = P(t \leq -t_c) = \frac{\alpha}{2}$$

where α is a probability value often taken to be α=.01 or α=.05.

- Consequently we can make the probability statement

$$P(-t_c \leq t \leq t_c) = 1 - \alpha \qquad (5.1.11)$$

$$P[-t_c \leq \frac{b_2 - \beta_2}{\sqrt{\mathrm{se}(b_2)}} \leq t_c] = 1 - \alpha$$

$$P[b_2 - t_c \mathrm{se}(b_2) \leq \beta_2 \leq b_2 + t_c \mathrm{se}(b_2)] = 1 - \alpha \qquad (5.1.7)$$

5.1.3 The Repeated Sampling Context

Table 5.1 Least Squares Estimates from 10 Random Samples

n	b_1	se(b_1)	b_2	se(b_2)	$\hat{\sigma}^2$
1	51.1314	27.4260	0.1442	0.0378	2193.4597
2	61.2045	24.9177	0.1286	0.0344	1810.5972
3	40.7882	17.6670	0.1417	0.0244	910.1835
4	80.1396	23.8146	0.0886	0.0329	1653.8324
5	31.0110	22.8126	0.1669	0.0315	1517.5837
6	54.3099	26.9317	0.1086	0.0372	2115.1085
7	69.6749	19.2903	0.1003	0.0266	1085.1312
8	71.1541	26.1807	0.1009	0.0361	1998.7880
9	18.8290	22.4234	0.1758	0.0309	1466.2541
10	36.1433	23.5531	0.1626	0.0325	1617.7087

- The 95% confidence interval estimates for the parameters β_1 and β_2 are given in Table 5.2.

Table 5.2 Interval Estimates from 10 Random Samples

n	$b_1 - t_c\text{se}(b_1)$	$b_1 + t_c\text{se}(b_1)$	$b_2 - t_c\text{se}(b_2)$	$b_2 + t_c\text{se}(b_2)$
1	-4.3897	106.6524	0.0676	0.2207
2	10.7612	111.6479	0.0590	0.1982
3	5.0233	76.5531	0.0923	0.1910
4	31.9294	128.3498	0.0221	0.1551
5	−15.1706	77.1926	0.1032	0.2306
6	−0.2105	108.8303	0.0334	0.1838
7	30.6237	108.7261	0.0464	0.1542
8	18.1541	124.1542	0.0278	0.1741
9	−26.5649	64.2229	0.1131	0.2384
10	−11.5374	83.8240	0.0968	0.2284

5.1.4 An Illustration

- For the food expenditure data

$$P[b_2 - 2.024\text{se}(b_2) \le \beta_2 \le b_2 + 2.024\text{se}(b_2)] = .95 \tag{5.1.14}$$

- The critical value t_c = 2.024, which is appropriate for α = .05 and 38 degrees of freedom
- It can be computed exactly with a software package.
- To construct an interval estimate for β_2 we use the least squares estimate b_2 = .1283 which has the standard error

$$\text{se}(b_2) = \sqrt{\hat{\text{var}}(b_2)} = \sqrt{0.0009326} = 0.0305$$

A "95% confidence interval estimate" for β_2:

$$b_2 \pm t_c \mathrm{se}(b_2) = .1283 \pm 2.024(.0305)=[.0666,.1900]$$

5.2 Hypothesis Testing

Components of Hypothesis Tests

1. A *null* hypothesis, H_0
2. An *alternative* hypothesis, H_1
3. A test *statistic*
4. A *rejection* region

5.2.1 The Null Hypothesis

The "null" hypothesis, which is denoted H_0 (*H-naught*), specifies a value for a parameter. The null hypothesis is stated $H_0 : \beta_2 = c$, where c is a constant, and is an important value in the context of a specific regression model.

5.2.2 The Alternative Hypothesis

For the null hypothesis H_0: $\beta_2 = c$ three possible alternative hypotheses are:

- H_1: $\beta_2 \neq c$.
- H_1: $\beta_2 > c$
- H_1: $\beta_2 < c$.

5.2.3 The Test Statistic

$$t = \frac{b_2 - \beta_2}{\mathrm{se}(b_2)} \sim t_{(T-2)} \tag{5.2.1}$$

If the null hypothesis H_0: $\beta_2 = c$ is ***true,*** then

$$t = \frac{b_2 - c}{\mathrm{se}(b_2)} \sim t_{(T-2)} \tag{5.2.2}$$

If the null hypothesis is ***not true***, then the t-statistic in equation 5.2.2 does ***not*** have a t-distribution with T–2 degrees of freedom

5.2.4 The Rejection Region

- The level of significance of the test α is frequently chosen to be .01, .05 or .10.
- The rejection region is determined by finding critical values t_c such that

$$P(t \geq t_c) = P(t \leq -t_c) = \alpha / 2 .$$

Rejection rule for a two-tailed test: If the value of the test statistic falls in the rejection region, either tail of the *t*-distribution, then we reject the null hypothesis and accept the alternative.

- Sample values of the test statistic in the central non-rejection area are *compatible with the null hypothesis* and are not taken as evidence *against* the null hypothesis being true.
- Finding a sample value of the test statistic in the non-rejection region <u>does not make the null hypothesis true!</u>

If the value of the test statistic falls between the critical values $-t_c$ and t_c, in the non-rejection region, then we **do not reject** the null hypothesis.

5.2.5 The Food Expenditure Example

Test the null hypothesis that $\beta_2 = .10$ against the alternative that $\beta_2 \neq .10$, in the food expenditure model.

Format for Testing Hypotheses

1. Determine the null and alternative hypotheses.
2. Specify the test statistic and its distribution if the null hypothesis is true.
3. Select α and determine the rejection region.
4. Calculate the sample value of the test statistic.
5. State your conclusion.

Applying to Food Expenditure example,

1. The null hypothesis is H_0: $\beta_2 = .10$. The alternative hypothesis is H_1: $\beta_2 \neq .10$.
2. The test statistic $t = \dfrac{b_2 - .10}{\mathrm{se}(b_2)} \sim t_{(T-2)}$ *if the null hypothesis is true.*
3. Let us select α=.05. The critical value t_c is 2.024 for a t-distribution with $(T-2) = 38$ degrees of freedom.
4. Using the data in Table 3.1, the least squares estimate of β_2 is $b_2 = .1283$, with standard error $\mathrm{se}(b_2)$=0.0305. The value of the test statistic is $t = \dfrac{.1283 - .10}{.0305} = .93$.
5. Conclusion: Since t=.93 < t_c=2.024 we *do not reject* the null hypothesis

5.2.6 Type I and Type II Errors

We make a correct decision if:

- The null hypothesis is *false* and we decide to *reject* it.
- The null hypothesis is *true* and we decide *not* to reject it.

Our decision is incorrect if:

- The null hypothesis is *true* and we decide to *reject* it (a Type I error)
- The null hypothesis is *false* and we decide *not* to reject it (a Type II error)

Facts about the probability of a Type II error:

- The probability of a Type II error varies inversely with the level of significance of the test, α,
- The closer the true value of the parameter is to the hypothesized parameter value the larger is the probability of a Type II error
- The larger the sample size T, the lower the probability of a Type II error, given a level of Type I error α.
- The test based on the *t*-distribution that we have described is a very good test.

5.2.7 The *p*-Value of a Hypothesis Test

The *p*-value of a test is calculated by finding the probability that the *t*-distribution can take a value greater than or equal to the absolute value of the *sample value of the test statistic.*

Rejection rule for a two-tailed test: When the ***p-value*** of a hypothesis test is *smaller* than the chosen value of α, then the test procedure leads to *rejection* of the null hypothesis.

- If the *p*-value is greater than α we do not reject the null hypothesis.
- In the food expenditure example the *p*-value for the test of H_0: $\beta_2 = .10$ against H_1: $\beta_2 \neq .10$ is
- p=.3601, which is the area in the tails of the $t_{(38)}$ distribution where $|t| \geq .9263$

5.2.8 Tests of Significance

- In the food expenditure model one important null hypothesis is H_0: $\beta_2 = 0$.
- The general alternative hypothesis is H_1: $\beta_2 \neq 0$.
- Rejecting the null hypothesis implies that there is a "statistically significant" relationship between y and x.

5.2.8a A Significance Test in the Food Expenditure Model

1. The null hypothesis is H_0: $\beta_2 = 0$. The alternative hypothesis is H_1: $\beta_2 \neq 0$.
2. The test statistic $t = \dfrac{b_2}{\mathrm{se}(b_2)} \sim t_{(T-2)}$ *if the null hypothesis is true.*
3. Let us select α=.05. The critical value t_c is 2.024 for a t-distribution with $(T–2)$ = 38 degrees of freedom.

4. The least squares estimate of β_2 is b_2 = .1283, with standard error se(b_2)=0.0305. The value of the test statistic is $t = .1283/.0305 = 4.20$.
5. Conclusion: Since t=4.20 > t_c=2.024 we *reject* the null hypothesis and accept the alternative, that there is a relationship between weekly income and weekly food expenditure

The p-value for this hypothesis test is p=.000155, which is the area in the tails of the $t_{(38)}$ distribution where $|t|\geq 4.20$. Since $p \leq \alpha$ we reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 \neq 0$, and thus that a "statistically significant" relationship exists between y and x.

Remark: "Statistically significant" does not, however, necessarily imply "economically significant."

- For example, suppose the CEO of the supermarket chain plans a certain course of action *if* $\beta_2 \neq 0$.
- Furthermore suppose a large sample is collected from which we obtain the estimate b_2 = .0001 with se(b_2) = .00001, yielding the *t*-statistic t = 10.0.
- We would reject the null hypothesis that $\beta_2 = 0$ and accept the alternative that $\beta_2 \neq 0$. Here b_2 = .0001 is statistically different from zero.
- However, .0001 may not be "economically" different from 0, and the CEO may decide not to proceed with her plans.

5.2.8b Reading Computer Output

Dependent Variable: FOODEXP				
Method: Least Squares				
Sample: 1 40				
Included observations: 40				
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	40.76756	22.13865	1.841465	0.0734
INCOME	0.128289	0.030539	4.200777	0.0002

Figure 5.7 EViews Regression Output

5.2.9 A Relationship Between Hypothesis Testing and Interval Estimation

- There is an *algebraic* relationship between two-tailed hypothesis tests and confidence interval estimates that is sometimes useful.
- Suppose that we are testing the null hypothesis $H_0 : \beta_k = c$ against the alternative $H_1 : \beta_k \neq c$.
- If we *fail to reject* the null hypothesis at the α level of significance, then the value c will fall *within* a $(1-\alpha)\times 100\%$ confidence interval estimate of β_k.
- Conversely, if we reject the null hypothesis, then c will fall *outside* the $(1-\alpha)\times 100\%$ confidence interval estimate of β_k.
- This algebraic relationship is true because we fail to reject the null hypothesis when $-t_c \leq t \leq t_c$, or when

$$-t_c \le \frac{b_k - c}{\mathrm{se}(b_k)} \le t_c$$

$$b_k - t_c \mathrm{se}(b_k) \le c \le b_k + t_c \mathrm{se}(b_k)$$

5.2.10 One-Tailed Tests

- One-tailed tests are used to test H_0: $\beta_k = c$ against the alternative H_1: $\beta_k > c$, or H_1: $\beta_k < c$
- To test H_0: $\beta_k = c$ against the alternative H_1: $\beta_k > c$ we select the rejection region to be values of the test statistic t that support the *alternative*

- We define the rejection region to be values of t greater than a critical value t_c, from a t-distribution with T–2 degrees of freedom, such that $P(t \geq t_c) = \alpha$, where α is the level of significance of the test.
- The decision rule for this one-tailed test is, "Reject H_0: $\beta_k = c$ and accept the alternative H_1: $\beta_k > c$ if $t \geq t_c$." If $t < t_c$ then we do not reject the null hypothesis.
- Computation of the p-value is similarly confined to one tail of the distribution

In the food expenditure example test H_0: $\beta_2 = 0$ against the alternative H_1: $\beta_2 > 0$.

1. The null hypothesis is H_0: $\beta_2 = 0$. The alternative hypothesis is H_1: $\beta_2 > 0$.
2. The test statistic $t = \dfrac{b_2}{\mathrm{se}(b_2)} \sim t_{(T-2)}$ *if the null hypothesis is true.*
3. For the level of significance α=.05 the critical value t_c is 1.686 for a t-distribution with T–2=38 degrees of freedom
4. The least squares estimate of β_2 is b_2 = .1283, with standard error se(b_2)=0.0305.

   Exactly as in the two-tailed test the value of the test statistic is $t = \dfrac{.1283}{.0305} = 4.20$.
5. Conclusion: Since t=4.20 > t_c=1.686 we *reject* the null hypothesis and accept the alternative, that there is a positive relationship between weekly income and weekly food expenditure.

5.2.11 A Comment on Stating Null and Alternative Hypotheses

- The null hypothesis is usually stated in such a way that if our theory is correct, then we will reject the null hypothesis
- We set up the null hypothesis that there is *no* relation between the variables, H_0: $\beta_2 = 0$. In the alternative hypothesis we put the conjecture that we would like to establish, H_1: $\beta_2 > 0$.
- It is important to set up the null and alternative hypotheses *before* you carry out the regression analysis.

5.3 The Least Squares Predictor

We want to predict for a given value of the explanatory variable x_0 the value of the dependent variable y_0, which is given by

$$y_0 = \beta_1 + \beta_2 x_0 + e_0 \tag{5.3.1}$$

where e_0 is a random error. This random error has mean $E(e_0)=0$ and variance $\text{var}(e_0)=\sigma^2$. We also assume that $\text{cov}(e_0, e_t)=0$.

The least squares predictor of y_0,

$$\hat{y}_0 = b_1 + b_2 x_0 \tag{5.3.2}$$

The *forecast error* is

$$f = \hat{y}_0 - y_0 = b_1 + b_2 x_0 - (\beta_1 + \beta_2 x_0 + e_0)$$

$$= (b_1 - \beta_1) + (b_2 - \beta_2) x_0 - e_0 \qquad (5.3.3)$$

The expected value of f is:

$$E(f) = E(\hat{y}_0 - y_0) = E(b_1 - \beta_1) + E(b_2 - \beta_2) x_0 - E(e_0)$$

$$= 0 + 0 - 0 = 0 \qquad (5.3.4)$$

It can be shown that

$$\text{var}(f) = \text{var}(\hat{y}_0 - y_0) = \sigma^2 \left[1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum (x_t - \bar{x})^2} \right] \tag{5.3.5}$$

The forecast error variance is estimated by replacing σ^2 by its estimator $\hat{\sigma}^2$,

$$\hat{\text{var}}(f) = \hat{\sigma}^2 \left[1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum (x_t - \bar{x})^2} \right] \tag{5.3.6}$$

The square root of the estimated variance is the *standard error of the forecast*,

$$se(f) = \sqrt{\hat{\text{var}}(f)} \tag{5.3.7}$$

Consequently we can construct a standard normal random variable as

$$\frac{f}{\sqrt{\text{var}(f)}} \sim N(0,1) \tag{5.3.8}$$

Then

$$\frac{f}{\sqrt{\widehat{\text{var}}(f)}} = \frac{f}{\text{se}(f)} \sim t_{(T-2)}, \tag{5.3.9}$$

If t_c is a critical value from the $t_{(T-2)}$ distribution such that $P(t \geq t_c) = \alpha/2$, then

$$P(-t_c \leq t \leq t_c) = 1 - \alpha \tag{5.3.10}$$

Then

$$P[-t_c \le \frac{\hat{y}_0 - y_0}{\text{se}(f)} \le t_c] = 1 - \alpha$$

Simplify this expression to obtain

$$P[\hat{y}_0 - t_c\text{se}(f) \le y_0 \le y_0 + t_c\text{se}(f)] = 1 - \alpha \qquad (5.3.11)$$

A (1-α)×100% confidence interval, or prediction interval, for y_0 is

$$\hat{y}_0 \pm t_c\text{se}(f) \qquad (5.3.12)$$

- Equation 5.3.5 implies that, the farther x_0 is from the sample mean $\bar{x}$, the larger the variance of the prediction error
- Since the forecast variance increases the farther x_0 is from the sample mean of $\bar{x}$, the confidence bands increase in width as $|x_0 - \bar{x}|$ increases.

5.3.1 Prediction in the Food Expenditure Model

The predicted the weekly expenditure on food for a household with x_0 = \$750 weekly income is

$$\hat{y}_0 = b_1 + b_2 x_0 = 40.7676 + .1283(750) = 136.98$$

The estimated variance of the forecast error is

$$\hat{\text{var}}(f) = \hat{\sigma}^2 \left[1 + \frac{1}{T} + \frac{(x_0 - \bar{x})^2}{\sum (x_t - \bar{x})^2} \right] = 1429.2456 \left[1 + \frac{1}{40} + \frac{(750 - 698)^2}{1532463} \right] = 1467.4986$$

The standard error of the forecast is then

$$\text{se}(f) = \sqrt{\hat{\text{var}}(f)} = \sqrt{1467.4986} = 38.3079$$

The 95% confidence interval for y_0 is

$$\hat{y}_0 \pm t_c \text{se}(f) = 136.98 \pm 2.024(38.3079)$$

$$[59.44 \text{ to } 214.52]$$

- Our prediction interval suggests that a household with \$750 weekly income will spend somewhere between \$59.44 and \$214.52 on food.
- Such a wide interval means that our point prediction, \$136.98, is not reliable.
- We might be able to improve it by measuring the effect that factors other than income might have.