ACE 562
Fall 2005

Lecture 8: The Simple Linear Regression Model: R^2, Reporting the Results and Prediction

by
Professor Scott H. Irwin

Required Readings:

Griffiths, Hill and Judge. "Explaining Variation in the Dependent Variable," Section 8.1; "Reporting-Summarizing Results," Section 8.2; and "Predicting Expenditure," Section 7.3 in *Learning and Practicing Econometrics*

Overview

There are two major reasons for analyzing the linear statistical model,

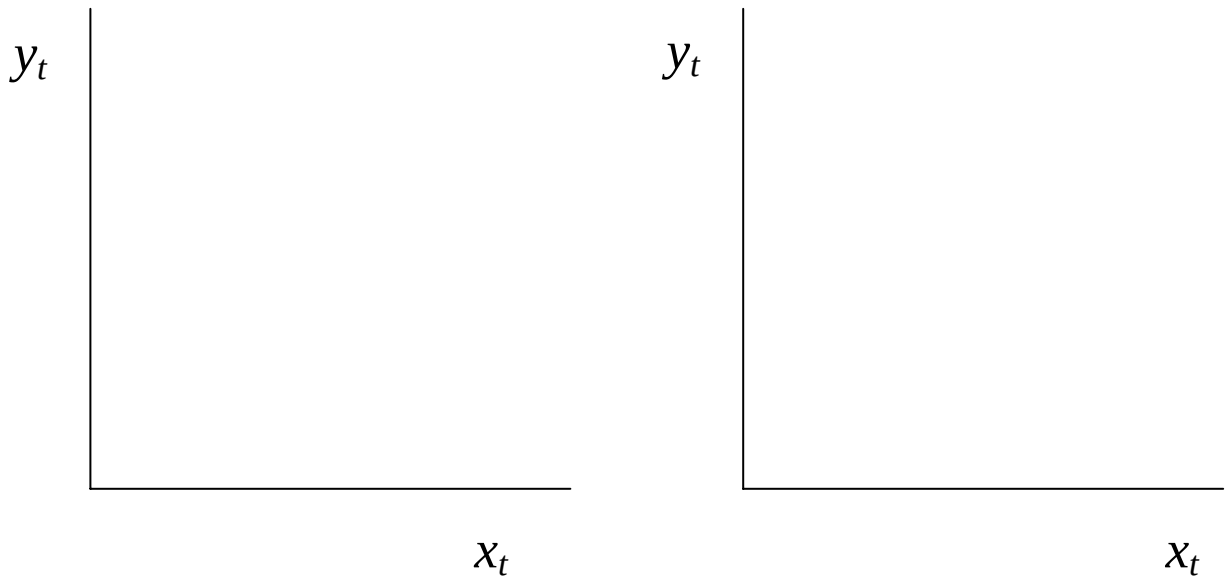
- Explain how y_t changes as x_t changes
- Predict y_t given x_t

Technique of LS guarantees that estimated line will be the "best-fitting," in the sense of having the smallest sum of squared errors

Despite being "best-fitting" the degree of fit can vary considerably for LS lines

- If the scatterplot of data are close to the line, we would say the LS line fits "well"
- If the scatterplot of data are not close to the line, we would say the LS line fits "poorly"

Suggests the need to quantify how well a LS line fits the data


y_t
x_t
y_t
x_t

Standard Error of Regression as a Measure of Fit

To begin, note that the estimated LS line yields a set of fitted values

$$\hat{y}_t = b_1 + b_2 x_t$$

The associated errors of fit are given by

$$\hat{e}_t = y_t - \hat{y}_t$$

It is natural to first consider the standard error as a measure of fit

- Provides a measure of the "typical" regression error

The formula to estimate the standard error of the regression is

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{\frac{\hat{e}_1^2 + \hat{e}_2^2 + \ldots + \hat{e}_T^2}{T-2}} = \sqrt{\frac{\sum_{t=1}^{T} \hat{e}_t^2}{T-2}}$$

Note that units of measurement for $\hat{\sigma}$ are always the same as for y_t

For the food expenditure problem, we found that

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{46.853} = 6.845$$

We interpreted this as saying that "the typical, or expected, error for the food expenditure LS regression line is $6.84 per week"

Some questions quickly arise

- Is $6.84 per week "big" or "small?"
- How does $6.84 per week compare to other LS lines?

Suggests the need for a measure of fit that does not depend on the units of measurement of the dependent variable y_t

R^2, or the coefficient of determination, provides a pure, unit-less measure of fit

R^2 as a Measure of Fit

In the previous section we noted that the estimated regression errors are given by,

$$\hat{e}_t = y_t - \hat{y}_t$$

Re-arranging this expression, we can show that the value of y_t can be decomposed into two components,

$$y_t = \hat{y}_t + \hat{e}_t$$

To begin the derivation of R^2 it is helpful to subtract the mean of *y* from both sides of the equation

$$(y_t - \bar{y}) = (\hat{y}_t - \bar{y}) + \hat{e}_t$$

In words, this says,

Total deviation in y_t = component explained by x_t +
unexplained component

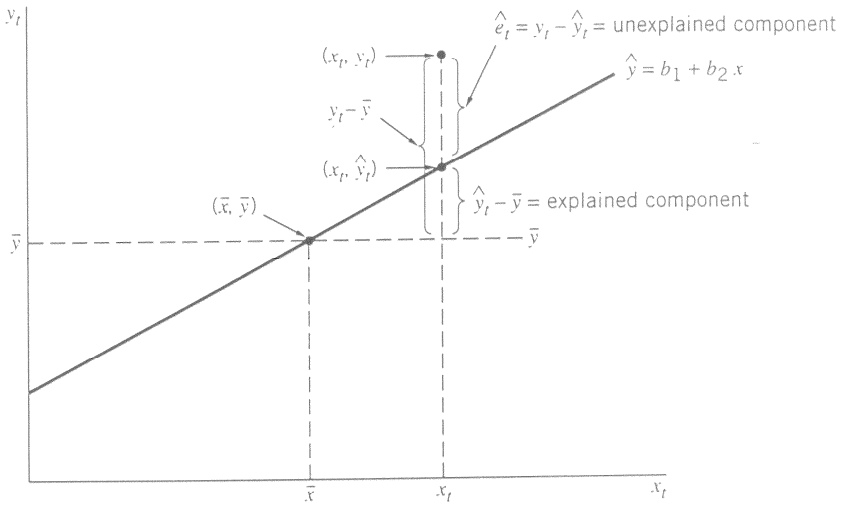


FIGURE 6.1 Explained and unexplained components of y_t

Hill, C., W. Griffiths, and G. Judge. *Undergraduate Econometrics.* John Wiley & Sons, Inc., New York, NY 1997.

Since we are interested in "variation" and not "deviation," let's square both sides of the previous equation,

$$(y_t - \bar{y})^2 = [(\hat{y}_t - \bar{y}) + \hat{e}_t]^2$$

Which can be expanded as follows,

$$(y_t - \bar{y})^2 = (\hat{y}_t - \bar{y})^2 + \hat{e}_t^2 + 2(\hat{y}_t - \bar{y})\hat{e}_t$$

Now, sum both sides of the previous equation,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T}\hat{e}_t^2 + 2\sum_{t=1}^{T}(\hat{y}_t - \bar{y})\hat{e}_t$$

Since,

$$2\sum_{t=1}^{T}(\hat{y}_t - \bar{y})\hat{e}_t = 2\sum_{t=1}^{T}(\hat{y}_t\hat{e}_t - \bar{y}\hat{e}_t)$$

$$= 2\sum_{t=1}^{T}\hat{y}_t\hat{e}_t - 2\bar{y}\sum_{t=1}^{T}\hat{e}_t$$

$$= 2\sum_{t=1}^{T}\hat{y}_t\hat{e}_t$$

$$= 2\sum_{t=1}^{T}(b_1 + b_2 x_t)\hat{e}_t$$

$$= 2b_1\sum_{t=1}^{T}\hat{e}_t + 2b_2\sum_{t=1}^{T}x_t\hat{e}_t$$

$$= 2b_2\sum_{t=1}^{T}x_t\hat{e}_t = 0$$

Hence, the earlier relationship reduces to,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T}\hat{e}_t^2$$

This is an important relationship, which shows the decomposition of total sample variation in y_t into explained and unexplained components

Now, define the following terms,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \text{Sum of Squares Total } (SST)$$

$$\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 = \text{Sum of Squares Regression } (SSR)$$

$$\sum_{t=1}^{T}\hat{e}_t^2 = \text{Sum of Squares Error } (SSE)$$

Hence,

$$SST = SSR + SSE$$

This decomposition is provided in the regression output of virtually all econometric packages

Usually labeled as the analysis of variance table

Table 6.1 **Analysis of Variance Table**

Source of Variation	DF	Sum of Squares	Mean Square
Explained	1	SSR	$SSR/1$
Unexplained	$T-2$	SSE	$SSE/(T-2)$ $[=\hat{\sigma}^2]$
Total	$T-1$	SST	

Hill, C., W. Griffiths, and G. Judge. *Undergraduate Econometrics.* John Wiley & Sons, Inc., New York, NY 1997.

A widespread use of the information in the analysis of variance table is to define a measure of the proportion of variation in y explained by x within the regression model

To obtain this measure, first divide the previous equation by SST to obtain the relationship in proportionate form,

$$\frac{\text{SST}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}} + \frac{\text{SSE}}{\text{SST}}$$

or,

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

Now, we can define,

$$R^2 = \frac{SSR}{SST}$$

Shows that R^2 measures the total sample variation in y_t explained by the variation in x_t

The formal term for R^2 is coefficient of determination

R^2 often is stated in percentage terms as follows,

$$R^2 = \frac{SSR}{SST} X\ 100$$

Note that by substituting into the SST equation in proportionate form, we obtain,

$$1 = R^2 + \frac{SSE}{SST}$$

or,

$$R^2 = 1 - \frac{SSE}{SST}$$

Two important limits can be placed on R^2,

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

- ***Why is R^2 non-negative?***
- ***What does it mean if R^2 is 0?***
- ***What does it mean if R^2 is 1?***

Correlation and R^2

The sample correlation coefficient for two random variables x and y is,

$$r_{x,y} = \frac{\hat{\text{cov}}(x, y)}{\sqrt{\hat{\text{var}}(x)\,\hat{\text{var}}(y)}}$$

There are two interesting relationships between $r_{x,y}$ and R^2 in the case of the simple linear regression model,

$$r_{x,y}^2 = R^2$$

$$r_{\hat{y},y}^2 = R^2$$

R^2 for the Food Expenditure Example

For the household food expenditure and income example, the relevant calculation is,

$$R^2 = \frac{826.6}{2607.0} = 0.317$$

Indicates we are able to explain 31.7% of the total variation in food expenditure by the variation in income

This leaves 68.3% of the variation unexplained, suggesting the "explanatory power" of the model is low

- Typical of cross-sectional data
- Regressions based on economic time-series data tend to have much higher $R^2 s$ due to shared time-trends of the variables

Table 8.1 Summary of Least Squares Results

Variable	Coefficient	Standard Error	t-Value (zero null)	p-Value
Constant	7.3832	4.0080	1.84	0.07330
Income	0.2323	0.0553	4.20	0.00016

Table 8.2 **Analysis of Variance Table**

	Source of Variation	Degrees of Freedom	Mean Square	Explained/ Unexplained
Explained	826.64	1	826.64	17.64
Unexplained	1780.4	38	46.85	
Total	2607.0	39	66.847	
		In General		
Explained	$\Sigma(\hat{y}_t - \bar{y})^2$	1	$\Sigma(\hat{y}_t - \bar{y})^2 / 1$	
Unexplained	$\Sigma(y_t - \hat{y})^2$	$T-2$	$\Sigma(y_t - \hat{y})^2 / (T-2) = \hat{\sigma}^2$	
Total	$\Sigma(y_t - \bar{y})^2$	$T-1$	$\Sigma(y_t - \bar{y})^2 / (T-1)$	

Griffiths, W.E., R.C. Hill and G.C. Judge. *Learning and Practicing Econometrics*. John Wiley & Sons, Inc., New York, NY, 1993.

Sample Regression Output from Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.563096017**
R Square	**0.317077125**
Adjusted R Square	**0.29910547**
Standard Error	**6.844922384**
Observations	**40**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**1**	**826.6352172**	**826.6352**	**17.64318**	**0.000155136**
Residual	**38**	**1780.412573**	**46.85296**		
Total	**39**	**2607.04779**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**7.383217543**	**4.008356335**	**1.841956**	**0.073296**	**-0.731275911**	**15.497711**
X Variable 1	**0.23225333**	**0.055293429**	**4.200378**	**0.000155**	**0.120317631**	**0.34418903**

Words of Wisdom from Peter Kennedy (p. 27)

In general, econometricians are interested in obtaining "good" parameter estimates where "good" is not defined in terms of R^2. Consequently, the measure R^2 is not of much importance in econometrics. Unfortunately, however, many practitioners act as though it is important, for reasons that are not entirely clear, as noted by Cramer (1987, p.253),

> "These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings."

Implications

It is a mistake to focus too closely on R^2 as a measure of econometric "success"

A low R^2 does not necessarily indicate the estimated parameters do not provide useful information

Table 5.2 Average Weekly Expenditure on Food and Average Weekly Income in Dollars for 40 Households of Size 3

Observation Number	Household Expenditure on Food y_t	Household Income x_t	Observation Number	Household Expenditure on Food y_t	Household Income x_t
1	9.46	25.83	21	17.77	71.98
2	10.56	34.31	22	22.44	72.00
3	14.81	42.50	23	22.87	72.23
4	21.71	46.75	24	26.52	72.23
5	22.79	48.29	25	21.00	73.44
6	18.19	48.77	26	37.52	74.25
7	22.00	49.65	27	21.69	74.77
8	18.12	51.94	28	27.40	76.33
9	23.13	54.33	29	30.69	81.02
10	19.00	54.87	30	19.56	81.85
11	19.46	56.46	31	30.58	82.56
12	17.83	58.83	32	41.12	83.33
13	32.81	59.13	33	15.38	83.40
14	22.13	60.73	34	17.87	91.81
15	23.46	61.12	35	25.54	91.81
16	16.81	63.10	36	39.00	92.96
17	21.35	65.96	37	20.44	95.17
18	14.87	66.40	38	30.10	101.40
19	33.00	70.42	39	20.90	114.13
20	25.19	70.48	40	48.71	115.46

Griffiths, W.E., R.C. Hill and G.C. Judge. *Learning and Practicing Econometrics*. John Wiley & Sons, Inc., New York, NY, 1993.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.56
R Square	0.32
Adjusted R Square	0.30
Standard Error	6.84
Observations	40.00

ANOVA

	df	*SS*
Regression	1.00	826.64
Residual	38.00	1780.41
Total	39.00	2607.05

	Coefficients	*Standard Error*
Intercept	7.38	4.01
X Variable 1	0.23	0.06

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.26
R Square	0.07
Adjusted R Square	0.01
Standard Error	5.96
Observations	20.00

ANOVA

	df	*SS*
Regression	1.00	45.41
Residual	18.00	639.62
Total	19.00	685.02

	Coefficients	*Standard Error*
Intercept	9.17	12.94
X Variable 1	0.21	0.19

Reporting the Results of Regression Analysis

There are several, standard methods of reporting regression results

One common form is,

$$\hat{y}_t = 7.3832 + 0.2323 x_t \qquad R^2 = 0.317$$
$$(4.0080)\ \ (0.0553) \qquad\qquad (s.e.)$$

where *s.e.* stands for estimated standard error

Another is to replace the standard errors by *t*-statistics for a zero null,

$$\hat{y}_t = 7.3832 + 0.2323 x_t \qquad R^2 = 0.317$$
$$(1.84) \qquad (4.20) \qquad\qquad (t - stat.)$$

Finally, it has become commonplace in recent years to also report *p*-values in either reporting format,

$$\hat{y}_t = 7.3832 + 0.2323x_t \qquad R^2 = 0.317$$
$$(4.0080)\ \ (0.0553) \qquad (s.e.)$$
$$[0.07330]\ \ [0.00016] \qquad [p-value]$$

$$\hat{y}_t = 7.3832 + 0.2323x_t \qquad R^2 = 0.317$$
$$(1.84)\ \ (4.20) \qquad (t-stat.)$$
$$[0.07330]\ \ [0.00016] \qquad [p-value]$$

If results for a large number of regressions must be reported, a tabular format should be employed

The same basic information should be reported in the table

Prediction with Regression Models

Prediction is a subject of great practical importance

- Often given little treatment in textbooks
- We will cover the subject in detail

The terms prediction and forecast can be used interchangeably

In a regression setting, we want to predict the value of the dependent variable y_0 for a given value of the independent variable x_0

Example: What would be the level of food expenditure for a family that has a weekly income of $60?

- An example of cross-sectional prediction
- We will consider two approaches to answering this specific question

Case 1: β_1, β_2, x_0 and σ^2 Known

Assume a linear statistical model is the data generating process for food expenditure,

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

where, as before, y_t is food expenditure, x_t is income, and e_t and y_t are assumed to be *iid* with the following distributions,

$$e_t \sim N(0, \sigma^2) \qquad \text{and} \qquad y_t \sim N(\beta_1 + \beta_2 x_t, \sigma^2)$$

It is important to note that we are assuming that β_1 and β_2 are known

Since the statistical model holds for any observation, we can write the following version,

$$y_0 = \beta_1 + \beta_2 x_0 + e_0$$

where y_0 is the predicted value of food expenditure for a given value of income x_0

At this point, y_0 is not a prediction because the value of e_0 is unknown

The best we can do is to use the expected value of y_0 as our prediction,

$$E(y_0) = \hat{y}_0 = E\left[\beta_1 + \beta_2 x_0 + e_o\right] = \beta_1 + \beta_2 x_0$$

$\hat{y}_0$ is called the least squares predictor

Note that $\hat{y}_0$ can differ from y_0 because the future disturbance e_0 may differ from its implicit predictor, which is its mean value of 0

Hence, $\hat{y}_0$ is a random variable and we are interested in its properties in a repeated sampling context

It is conventional to examine the sampling properties of the forecast error, rather than the sampling properties of $\hat{y}_0$ directly

Forecast error

The forecast error is defined as the difference between the actual y_0 and the prediction $\hat{y}_0$

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

Note that the forecast error in this case is exactly equal to the error term in the statistical model

Based on the above relationship, we can examine some important sampling properties of the forecast error

Mean forecast error

We can examine the expected value of the forecast error that we should expect in repeated sampling,

$$E(f) = E(y_0 - \hat{y}_0) = E(e_0) = 0$$

This shows that the least square predictor is an unbiased linear predictor

On average, in the repeated sampling sense, the predicted food expenditure will equal the actual value

Variance of the forecast error

While the least squares prediction is unbiased, it may still be wide of the mark for any particular prediction

The "reliability" of the prediction in repeated sampling is measured by the variance of the prediction

$$\text{var}(f) = E(y_0 - \hat{y}_0)^2 = E(e_0^2) = \sigma^2$$

Shows that the variance of the forecast error is exactly equal to the variance of the regression error term (also assumed to be known)

Standard error of the forecast

$$se(f) = \sqrt{\text{var}(f)} = \sqrt{\sigma^2} = \sigma$$

95% confidence interval for forecast

We can construct a standard normal random variable as follows,

$$Z_f = \frac{y_0 - \hat{y}_0}{\sqrt{\text{var}(f)}} = \frac{f}{\sigma} \sim N(0,1)$$

Since Z_f is a standard, normal random variable, we can write,

$$P[-1.96 \le Z_f \le 1.96] = 0.95$$

Substituting for Z_f,

$$P[-1.96 \le \frac{y_0 - \hat{y}_0}{\sigma} \le 1.96] = 0.95$$

Multiply the inequality in the brackets by σ,

$$P[-1.96\sigma \le y_0 - \hat{y}_0 \le 1.96\sigma] = 0.95$$

Now, add $\hat{y}_0$ to each term,

$$P[-\hat{y}_0 - 1.96\sigma \le -y_0 \le -\hat{y}_0 + 1.96\sigma] = 0.95$$

Hence, the 95 percent confidence interval for y_0 is,

$$\hat{y}_0 \pm 1.96\sigma$$

Interpretation: In repeated sampling, we expect 95% of interval predictions to contain the realized y_0

We can generalize to any prediction confidence level, $1-\alpha$, as follows,

$$P[\hat{y}_0 - Z_{\alpha/2}\sigma \leq y_0 \leq \hat{y}_0 + Z_{\alpha/2}\sigma] = 1-\alpha$$

and,

$$\hat{y}_0 \pm Z_{\alpha/2}\sigma$$

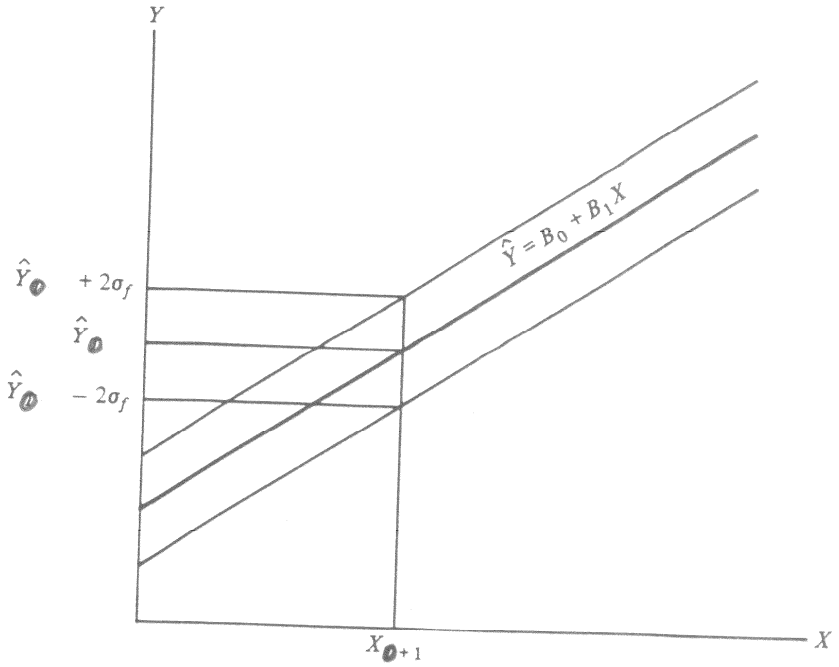


Figure 18.2 Standard Error of Forecast when β_0, β_1, and X_0 Are Known.
(and Confidence Interval)

Johnson, A.C., M.B. Johnson, and R.C. Buse. *Econometrics: Basic and Applied.* Macmillan Publishing Co., 1989.

Case 2: β_1, β_2 and σ^2 Estimated

We start with the same assumption that a linear statistical model is the data generating process for food expenditure,

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

Again, since the statistical model holds for any observation, we can write the following version,

$$y_0 = \beta_1 + \beta_2 x_0 + e_0$$

where y_0 is the predicted value of food expenditure for a given value of income x_0

However, we now make the more realistic assumption that β_1 and β_2 must be estimated

In this case, the best we can do is: 1) replace β_1 and β_2 with the estimators b_1 and b_2, and 2) replace the unknown error with its expected value of zero

The least squares predictor is then,

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

Note that $\hat{y}_0$ can now differ from y_0 for two reasons

1. The future disturbance e_0 may differ from its implicit predictor, which is its mean value of 0

2. The estimators b_1 and b_2 are likely to produce estimates that differ from the true population parameters β_1 and β_2

Hence, $\hat{y}_0$ is a random variable and we are interested in its properties in a repeated sampling context

Again, it is conventional to examine the sampling properties of the forecast error, rather than the sampling properties of $\hat{y}_0$ directly

Forecast error

The forecast error is defined as the difference between the actual y_0 and the predictor $\hat{y}_0$

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

or,

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

Note that the forecast error in this case does not simply equal the regression error term

- The forecast error is now a function of three random variables, b_1, b_2, and e_t

Based on the above relationship, we can again examine the sampling properties of the forecast error

Mean forecast error

The expected value of the forecast error that we should expect in repeated sampling is,

$$\begin{aligned} E(f) = E(y_0 - \hat{y}_0) &= E[(\beta_1 - b_1) + (\beta_2 - b_2)x_0 + e_0] \\ &= [\beta_1 - E(b_1)] + [\beta_2 - E(b_2)]x_0 + E(e_0) \\ &= [\beta_1 - \beta_1] + [\beta_2 - \beta_2]x_0 + E(e_0) \\ &= E(e_0) = 0 \end{aligned}$$

This shows that even when the parameters have to be estimated the least squares predictor is unbiased

$\Rightarrow$ On average, in the repeated sampling sense, the predicted food expenditure will equal the actual value

Variance of the forecast error

While the least squares predictor is unbiased, it may still be wide of the mark for any particular prediction

The "reliability" of the predictor is measured by the variance of the forecast error

$$\text{var}(f) = E(y_0 - \hat{y}_0)^2 = E[(\beta_1 - b_1) + (\beta_2 - b_2)x_0 + e_0)]^2$$

Expanding the square,

$$\text{var}(f) = E[(\beta_1 - b_1)^2 + ((\beta_2 - b_2)x_0)^2 + e_0^2$$
$$+2(\beta_1 - b_1)(\beta_2 - b_2)x_0 + 2(\beta_2 - b_2)x_0 e_0 + (\beta_1 - b_1)e_0]$$

Take the expectations through to each term,

$$\text{var}(f) = E[(\beta_1 - b_1)^2] + E[((\beta_2 - b_2)x_0)^2] + E[e_0^2]$$
$$+2E[(\beta_1 - b_1)(\beta_2 - b_2)x_0] + 2E[(\beta_2 - b_2)x_0 e_0] + E[(\beta_1 - b_1)e_0]$$

Which reduces to,

$$\text{var}(f) = E[(\beta_1 - b_1)^2] + E[((\beta_2 - b_2)x_0)^2] + E[e_0^2]$$
$$+2E[(\beta_1 - b_1)(\beta_2 - b_2)x_0]$$

Now change the notation,

$$\text{var}(f) = \text{var}(b_1) + \text{var}(b_2)x_0^2 + 2\,\text{cov}(b_1, b_2)x_0 + \sigma^2$$

The next step is to substitute the definitions of $\text{var}(b_1)$, $\text{var}(b_2)$, and $\text{cov}(b_1, b_2)$ that we derived earlier,

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

$$+ 2\sigma^2\left[\frac{-\bar{x}}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\right]x_0 + \sigma^2$$

After some fairly tedious algebra, this can be reduced to,

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

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

Key points:

- Since term in brackets must be positive, forecast error variance is larger than variance of the regression
- Reflects fact that forecast error is influenced not only by the regression error, but also that parameters must now be estimated.
- The greater the distance between the mean of x and x_0, the greater the variance of the forecast error
- In other words, the more distant is the observation for the independent variable from its mean, the more uncertain is the prediction
- All else constant, the larger the sample, the smaller the variance of the forecast error.

Standard error of the forecast

In the definition of the variance of the forecast error, the variance of the regression, σ^2, is assumed to be known

This is rarely, if ever, likely to be true in practice

Replace σ^2 by its estimator $\hat{\sigma}^2$ and derive the estimated forecast error variance

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

The estimated standard error of the forecast is then,

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

95% confidence interval for prediction

Previously, we constructed a standard normal random variable as follows,

$$Z_f = \frac{y_0 - \hat{y}_0}{\sqrt{\text{var}(f)}} \sim N(0,1)$$

But we now must replace $\text{var}(f)$ with its estimate $\hat{\text{var}}(f)$, which results in a *t*-distributed random variable

$$t_f = \frac{y_0 - \hat{y}_0}{\sqrt{\hat{\text{var}}(f)}} = \frac{y_0 - \hat{y}_0}{\hat{se}(f)} \sim t_{\alpha/2, T-2}$$

From a *t*-table, we know that when $T = 40$ and $\alpha = 0.05$, the associated critical values are +2.024 and -2.024

Since t_f is a *t*-distributed random variable, we can write,

$$P[-2.024 \le t_f \le 2.024] = 0.95$$

Substituting for t_f,

$$P[-2.024 \le \frac{y_0 - \hat{y}_0}{\hat{se}(f)} \le 2.024] = 0.95$$

Multiply the inequality in the brackets by $\hat{se}(f)$,

$$P[-2.024 \cdot \hat{se}(f) \le y_0 - \hat{y}_0 \le 2.024 \cdot \hat{se}(f)] = 0.95$$

Now, add $\hat{y}_0$ to each term,

$$P[\hat{y}_0 - 2.024 \cdot \hat{se}(f) \le y_0 \le \hat{y}_0 + 2.024 \cdot \hat{se}(f)] = 0.95$$

Hence, the 95 percent confidence interval for y_0 is,

$$\hat{y}_0 \pm 2.024 \cdot \hat{se}(f)$$

We can generalize to any prediction confidence level, $1-\alpha$, as follows,

$$P[\hat{y}_0 - t_{\alpha/2,T-2} \cdot \hat{se}(f) \le y_0 \le \hat{y}_0 + t_{\alpha/2,T-2} \cdot \hat{se}(f)] = 1-\alpha$$

and,

$$\hat{y}_0 \pm t_{\alpha/2,T-2} \cdot \hat{se}(f)$$

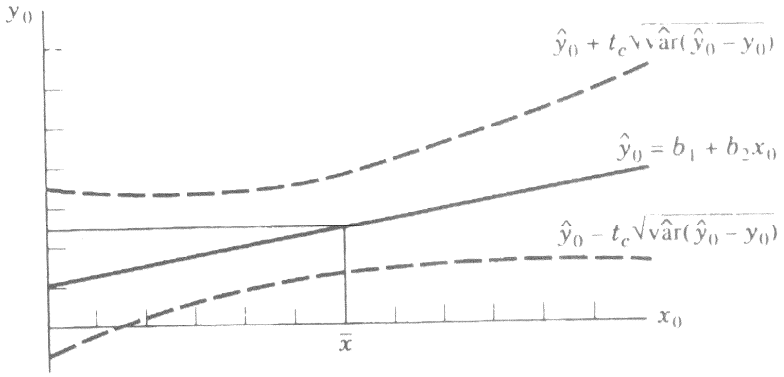


Figure 7.1 Point and interval prediction.

Griffiths, W.E., R.C. Hill and G.C. Judge. *Learning and Practicing Econometrics*. John Wiley & Sons, Inc., New York, NY, 1993.

Prediction Intervals in the Food Expenditure Example

Earlier, we generated the following estimates of the food expenditure-income relationship for a sample of 40 households,

$$b_1 = 7.3832 \quad b_2 = 0.2323 \quad \hat{\sigma}^2 = 46.853$$

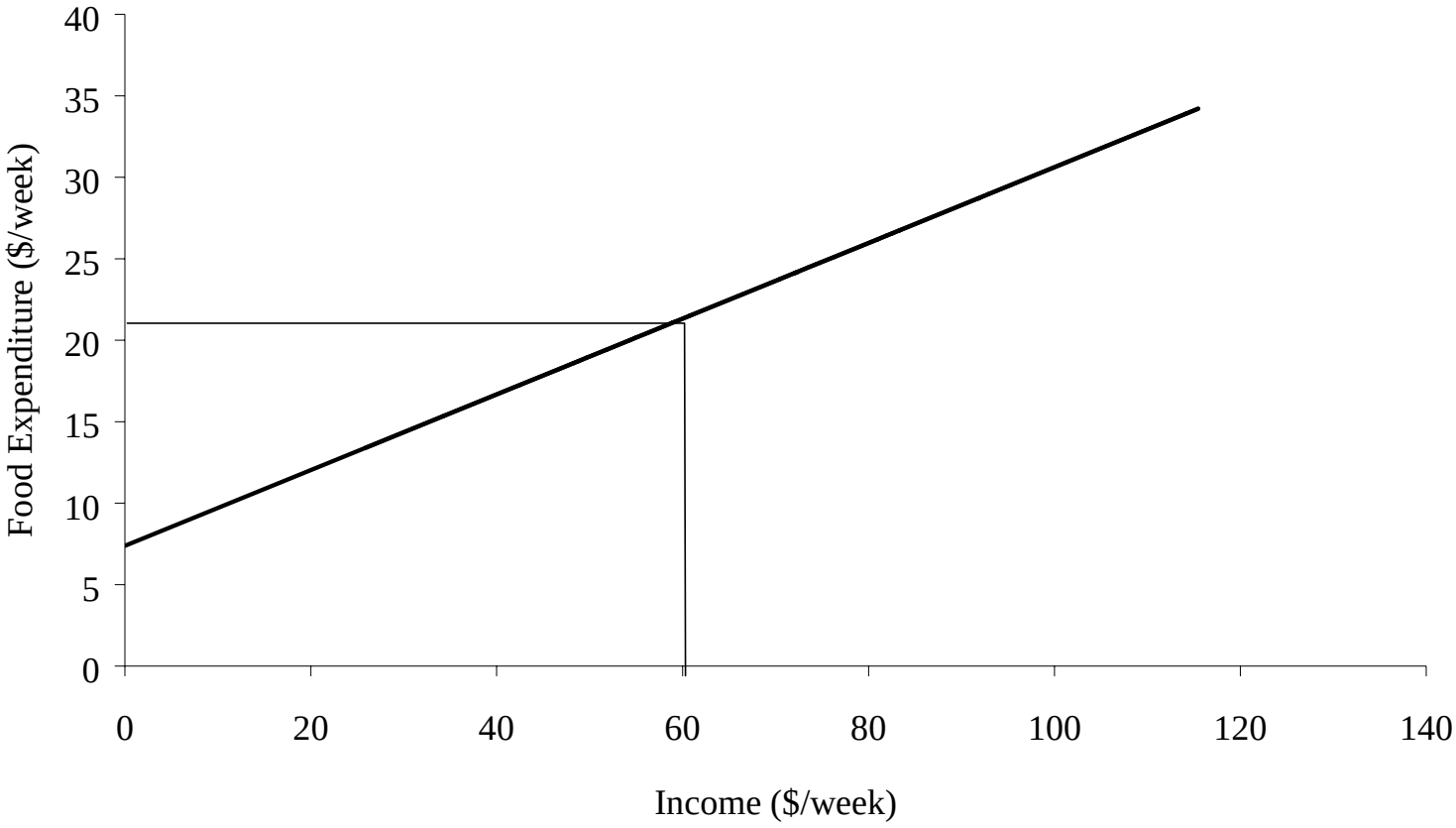
Based on these estimates the LS predictor is,

$$\hat{y}_0 = 7.3832 + 0.2323 x_0$$

If we set x_0 to \$60, then the prediction of household expenditure is,

$$\hat{y}_0 = 7.3832 + 0.2323(60) = 21.32$$

Prediction in the Food Expenditure Example



The corresponding estimate of the variance of the forecast error is,

$$\hat{\text{var}}(f) = 46.853\left[1 + \frac{1}{40} + \frac{(60 - 69.8)^2}{15,324.6}\right] = 48.318$$

The estimated standard error of the forecast is,

$$\hat{se}(f) = \sqrt{48.318} = 6.9511$$

The critical value for a t-distribution with $\alpha = 0.05$ and 38 degrees of freedom is 2.024, and hence, the 95% CI for the prediction is,

$$21.32 \pm 2.024 \cdot 6.9511$$

or,

$$\left(7.25 \leq \hat{y}_0 \leq 35.39\right)$$

Interpretation Guidelines:

In repeated sampling, we expect 95% of interval predictions to contain the realized y_0

If we use the interval predictor to compute a "large" number of interval predictions like $21.32 \pm 2.024 \cdot 6.9511$, 95% of these intervals will contain the realized y_0

It is incorrect to state,

> *"Given that income is $60 per week, there is a 0.95 probability that the realized y_0 will be between $7.25 and $35.39."*

Remember that our confidence is in the predictor not the particular prediction

Compromise language,

> *"Given that income is $60 per week, we are 95% confident that the interval between $7.25 and $35.39 will contain the realized* y_0*."*

> *"Given that income is $60 per week, we are 95% confident that the interval between $7.25 and $35.39 will contain the realized food expenditure per week."*

where "confident" is understood to apply to the prediction interval estimator in repeated sampling not the $7.25 to $35.39 interval estimate

Summary

- Our prediction interval suggests that a household with $60 in weekly income will spend somewhere between $7.25 to $35.39 on food
- Such a wide interval means that our point prediction, $21.32, is not reliable
- We might be able to improve our prediction by measuring the effect of factors other than income

Forecast CI's in the Food Expenditure Example