12 Simple Linear Regression and Correlation

12.4 Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

Let x^* denote a specified value of the independent variable x.

Once the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have been calculated, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ can be regarded either as a point estimate of $\mu_{Y \cdot x^*}$ (the expected or true average value of Y when $x = x^*$) or as a prediction of the Y value that will result from a single observation made when $x = x^*$.

The point estimate or prediction by itself gives no information concerning how precisely $\mu_{Y \cdot x^*}$ has been estimated or Y has been predicted.

Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

This can be remedied by developing a CI for $\mu_{Y \cdot x^*}$ and a prediction interval (PI) for a single Y value.

Before we obtain sample data, both $\hat{\beta}_0$ and $\hat{\beta}_1$ are subject to sampling variability—that is, they are both statistics whose values will vary from sample to sample.

Suppose, for example, that $\beta_0 = 50$ and $\beta_1 = 2$.

Then a first sample of (x, y) pairs might give $\hat{\beta}_0 = 52.35$, $\hat{\beta}_1 = 1.895$; a second sample might result in $\hat{\beta}_0 = 46.52$, $\hat{\beta}_1 = 2.056$; and so on.

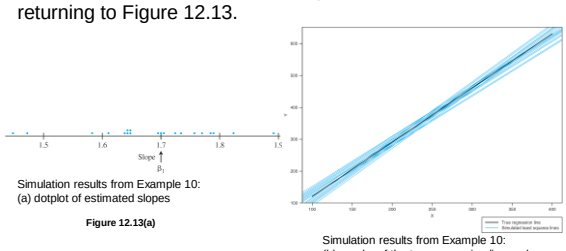
Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

It follows that $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$ itself varies in value from sample to sample, so it is a statistic. If the intercept and slope of the population line are the aforementioned values 50 and 2, respectively, and $x^* = 10$, then this statistic is trying to estimate the value 50 + 2(10) = 70.

The estimate from a first sample might be 52.35 + 1.895(10) = 71.30, from a second sample might be 46.52 + 2.056(10) = 67.08 , and so on.

Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

This variation in the value of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ can be visualized by returning to Figure 12.13.



Simulation results from Example 10:
(a) dotplot of estimated slopes

Figure 12.13(a)

Simulation results from Example 10:
(b) graphs of the true regression line and 20 least squares lines (from S-Plus)

Figure 12.13(b)

Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

Consider the value $x^* = 300$. The heights of the 20 pictured estimated regression lines above this value are all somewhat different from one another.

The same is true of the heights of the lines above the value $x^* = 350$. In fact, there appears to be more variation in the value of $\hat{\beta}_0 + \hat{\beta}_1(350)$ than in the value of $\hat{\beta}_0 + \hat{\beta}_1(300)$.

We shall see shortly that this is because 350 is further from $\bar{x} = 235.71$ (the "center of the data") than is 300. Methods for making inferences about β_1 were based on properties of the sampling distribution of the statistic $\hat{\beta}_1$.

Inferences Concerning $\mu_{Y \cdot x^*}$ and the Prediction of Future *Y* Values

In the same way, inferences about the mean *Y* value $\hat{\beta}_0 + \hat{\beta}_1 x^*$ are based on properties of the sampling distribution of the statistic $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Substitution of the expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ into $\hat{\beta}_0 + \hat{\beta}_1 x^*$ followed by some algebraic manipulation leads to the representation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ as a linear function of the *Yi*'s:

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] Y_i = \sum_{i=1}^{n} d_i Y_i$$

The coefficients $d_1, d_2, \ldots, d_n$ in this linear function involve the x_i's and x^*, all of which are fixed.

Inferences Concerning $\mu_{Y\cdot x^*}$ and the Prediction of Future *Y* Values

Application of the rules to this linear function gives the following properties.

Proposition

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1$ where x^* is some fixed value of x. Then

1. The mean value of $\hat{Y}$ is

$$E(\hat{Y}) = E(\hat{\beta}_0 + \hat{\beta}_1 x^*) = \mu_{\hat{\beta}_0+\hat{\beta}_1 x^*} = \beta_0 + \beta_1 x^*$$

Thus $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is an unbiased estimator for $\beta_0 + \beta_1 x^*$ (i.e., for $\mu_{Y\cdot x^*}$).

Inferences Concerning $\mu_{Y\cdot x^*}$ and the Prediction of Future *Y* Values

2. The variance of $\hat{Y}$ is

$$V(\hat{Y}) = \sigma_{\hat{Y}}^2 = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum x_i^2 - (\sum x_i)^2/n}\right] = \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]$$

And the standard deviation $\sigma_{\hat{Y}}$ is the square root of this expression. The estimated standard deviation of $\hat{\beta}_0 + \hat{\beta}_1 x^*$, denoted by $s_{\hat{Y}}$ or $s_{\hat{\beta}_0+\hat{\beta}_1 x^*}$, results from replacing σ by its estimate s:

$$s_{\hat{Y}} = s_{\hat{\beta}_0+\hat{\beta}_1 x^*} = s\sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

3. $\hat{Y}$ has a normal distribution.

Inferences Concerning $\mu_{Y\cdot x^*}$ and the Prediction of Future *Y* Values

The variance of $\hat{\beta}_0 + \hat{\beta}_1 x^*$ is smallest when $x^* = \bar{x}$ and increases as x^* moves away from $\bar{x}$ in either direction.

Thus the estimator of $\mu_{Y\cdot x^*}$ is more precise when x^* is near the center of the x_i's than when it is far from the values at which observations have been made. This will imply that both the CI and PI are narrower for an x^* near $\bar{x}$ than for an x^* far from $\bar{x}$.

Most statistical computer packages will provide both $\hat{\beta}_0 + \hat{\beta}_1 x^*$ and $s_{\hat{\beta}_0+\hat{\beta}_1 x^*}$ for any specified x^* upon request.

Inferences Concerning $\mu_{Y\cdot x^*}$

Inferences Concerning $\mu_{Y \cdot x^*}$

Just as inferential procedures for β_1 were based on the t variable obtained by standardizing β_1, a t variable obtained by standardizing $\hat{\beta}_0 + \hat{\beta}_1 x^*$ leads to a CI and test procedures here.

Theorem

The variable

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}} \qquad \textbf{(12.5)}$$

has a t distribution with $n - 2$ df.

13

Inferences Concerning $\mu_{Y \cdot x^*}$

A probability statement involving this standardized variable can now be manipulated to yield a confidence interval for $\mu_{Y \cdot x^*}$

A 100(1 – α)% **CI for** $\mu_{Y \cdot x^*}$, the expected value of Y when $x = x^*$, is

$$\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2, n-2} \cdot s_{\hat{\beta}_0 + \hat{\beta}_1 x^*} = \hat{y} \pm t_{\alpha/2, n-2} \cdot s_{\hat{Y}} \qquad \textbf{(12.6)}$$

14

Inferences Concerning $\mu_{Y \cdot x^*}$

This CI is centered at the point estimate for $\mu_{Y \cdot x^*}$ and extends out to each side by an amount that depends on the confidence level and on the extent of variability in the estimator on which the point estimate is based.

15

Example 13

Corrosion of steel reinforcing bars is the most important durability problem for reinforced concrete structures.

Carbonation of concrete results from a chemical reaction that lowers the pH value by enough to initiate corrosion of the rebar.

Representative data on x = carbonation depth (mm) and y = strength (MPa) for a sample of core specimens taken from a particular building follows (read from a plot in the article “The Carbonation of Concrete Structures in the Tropical Environment of Singapore,” *Magazine of Concrete Res.*, 1996: 293–300).

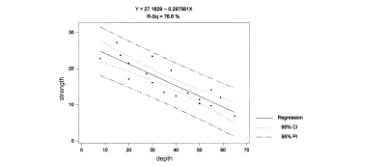
16

Example 13

cont'd

A scatter plot of the data (see Figure 12.17) gives strong support for use of the simple linear regression model.

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4
x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8



Minitab scatter plot with confidence intervals and prediction intervals for the data of Example 12.13

Figure 12.17

17

Example 13

cont'd

Relevant quantities are as follows:

$\sum x_i = 659.0$ $\quad \sum x_i^2 = 28{,}967.50$ $\quad \bar{x} = 36.6111$ $\quad S_{xx} = 4840.7778$

$\sum y_i = 293.2$ $\quad \sum x_i y_i = 9293.95$ $\quad \sum y_i^2 = 5335.76$

$\hat{\beta}_1 = -.297561$ $\quad \hat{\beta}_0 = 27.182936$ $\quad \text{SSE} = 131.2402$

$r^2 = .766$ $\quad s = 2.8640$

Let's now calculate a confidence interval, using a 95% confidence level, for the mean strength for all core specimens having a carbonation depth of 45 mm—that is, a confidence interval for $\hat{\beta}_0 + \hat{\beta}_1 (45)$.

18

Example 13

cont'd

The interval is centered at

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1(45) = 27.18 - .2976(45) = 13.79$$

The estimated standard deviation of the statistic $\hat{Y}$ is

$$s_{\hat{Y}} = 2.8640\sqrt{\frac{1}{18} + \frac{(45 - 36.6111)^2}{4840.7778}} = .7582$$

The 16 df t critical value for a 95% confidence level is 2.120, from which we determine the desired interval to be

$$13.79 \pm (2.120)(.7582) = 13.79 \pm 1.61 = (12.18, 15.40)$$

19

Example 13

cont'd

The narrowness of this interval suggests that we have reasonably precise information about the mean value being estimated.

Remember that if we recalculated this interval for sample after sample, in the long run about 95% of the calculated intervals would include $\hat{\beta}_0 + \hat{\beta}_1(45)$.

We can only hope that this mean value lies in the single interval that we have calculated.

20

Example 13

cont'd

Figure 12.18 shows Minitab output resulting from a request to fit the simple linear regression model and calculate confidence intervals for the mean value of strength at depths of 45 mm and 35 mm.

```
The regression equation is strength = 27.2 - 0.298 depth

Predictor       Coef         Stdev       t-ratio        P
 Constant      27.183         1.651         16.46    0.000
   depth     -0.29756       0.04116         -7.23    0.000
s = 2.864     R-sq = 76.6%     R-sq(adj) = 75.1%

Analysis of Variance

SOURCE          DF          SS          MS          F          P
Regression       1      428.62      428.62      52.25      0.000
     Error      16      131.24        8.20
     Total      17      559.86

   Fit     Stdev.Fit            95.0% C.I.            95.0% P.I.
13.793         0.758      (12.185, 15.401)       (7.510, 20.075)

   Fit     Stdev.Fit            95.0% C.I.            95.0% P.I.
16.768         0.678      (15.330, 18.207)      (10.527, 23.009)
```

Minitab regression output for the data of Example 12.13

Figure 12.18

Example 13

cont'd

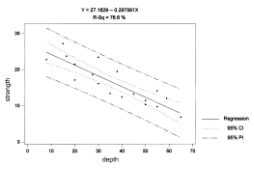
The intervals are at the bottom of the output; note that the second interval is narrower than the first, because 35 is much closer to $\bar{x}$ than is 45.

Example 13

cont'd

Figure 12.17 shows (1) curves corresponding to the confidence limits for each different x value and (2) prediction limits, to be discussed shortly. Notice how the curves get farther and farther apart as x moves away from $\bar{x}$.

x	8.0	15.0	16.5	20.0	20.0	27.5	30.0	30.0	35.0
y	22.8	27.2	23.7	17.1	21.5	18.6	16.1	23.4	13.4
x	38.0	40.0	45.0	50.0	50.0	55.0	55.0	59.0	65.0
y	19.5	12.4	13.2	11.4	10.3	14.1	9.7	12.0	6.8



Minitab scatter plot with confidence intervals and prediction intervals for the data of Example 13

Figure 12.17

Inferences Concerning $\mu_{Y \cdot x*}$

In some situations, a CI is desired not just for a single x value but for two or more x values.

Suppose an investigator wishes a CI both for $\mu_{Y \cdot v}$ and for $\mu_{Y \cdot w}$, where v and w are two different values of the independent variable.

It is tempting to compute the interval (12.6) first for $x = v$ and then for $x = w$.

Inferences Concerning $\mu_{Y \cdot x^*}$

Tests of hypotheses about $\hat{\beta}_0 + \hat{\beta}_1 x^*$ are based on the test statistic T obtained by replacing $\hat{\beta}_0 + \hat{\beta}_1 x^*$ in the numerator of (12.5) by the null value μ_0.

For example H_0: $\beta_0 + \beta_1(45) = 15$ in Example 13 says that when carbonation depth is 45 expected (i.e., true average) strength is 15.

The test statistic value is then $t = [\hat{\beta}_0 + \hat{\beta}_1(45) - 15]/s_{\hat{\beta}_0 + \hat{\beta}_1(45)}$, and the test is upper-, lower-, or two-tailed according to the inequality in H_a.

25

A Prediction Interval for a Future Value of *Y*

26

A Prediction Interval for a Future Value of *Y*

Rather than calculate an interval estimate for $\mu_{Y \cdot x^*}$, an investigator may wish to obtain an interval of plausible values for the value of Y associated with some future observation when the independent variable has value x^*.

Consider, for example, relating vocabulary size y to age of a child x. The CI (12.6) with $x^* = 6$ would provide an estimate of true average vocabulary size for all 6-year-old children.

Alternatively, we might wish an interval of plausible values for the vocabulary size of a particular 6-year-old child.

27

A Prediction Interval for a Future Value of *Y*

A CI refers to a parameter, or population characteristic, whose value is fixed but unknown to us.

In contrast, a future value of Y is not a parameter but instead a random variable; for this reason we refer to an interval of plausible values for a future Y as a **prediction interval** rather than a confidence interval.

The error of estimation is $\beta_0 + \beta_1 x^* - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, a difference between a fixed (but unknown) quantity and a random variable.

28

A Prediction Interval for a Future Value of *Y*

The error of prediction is $Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)$, a difference between two random variables. There is thus more uncertainty in prediction than in estimation, so a PI will be wider than a CI. Because the future value Y is independent of the observed Y_i's,

$$\begin{aligned} V[Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)] &= \text{variance of prediction error} \\ &= V(Y) + V(\hat{\beta}_0 + \hat{\beta}_1 x^*) \\ &= \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \end{aligned}$$

29

A Prediction Interval for a Future Value of *Y*

Furthermore, because $E(Y) = \beta_0 + \beta_1 x^*$ and $\hat{\beta}_0 + \hat{\beta}_1 x^* = \beta_0 + \beta_1 x^*$, the expected value of the prediction error is $E(Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)) = 0$.

It can then be shown that the standardized variable

$$T = \frac{Y - (\hat{\beta}_0 + \hat{\beta}_1 x^*)}{S\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}}$$

has a t distribution with $n - 2$ df.

30

A Prediction Interval for a Future Value of *Y*

Substituting this T into the probability statement $P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-1}) = 1 - \alpha$ and manipulating to isolate Y between the two inequalities yields the following interval.

A 100(1 – α)% **PI for a future *Y* observation to be made when *x* = *x**** is

$$\begin{aligned} &\hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}} \\ &= \hat{\beta}_0 + \hat{\beta}_1 x^* \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{\beta}_0 + \hat{\beta}_1 x^*}} \qquad \textbf{(12.7)} \\ &= \hat{y} \pm t_{\alpha/2,n-2} \cdot \sqrt{s^2 + s^2_{\hat{Y}}} \end{aligned}$$

31

A Prediction Interval for a Future Value of *Y*

The interpretation of the prediction level 100(1 – α)% is analogous to that of previous confidence levels—if (12.7) is used repeatedly, in the long run the resulting intervals will actually contain the observed y values 100(1 – α)% of the time.

Notice that the 1 underneath the initial square root symbol makes the PI (12.7) wider than the CI (12.6), though the intervals are both centered at $\hat{\beta}_0 + \hat{\beta}_1 x^*$.

Also, as $n \to \infty$, the width of the CI approaches 0, whereas the width of the PI does not (because even with perfect knowledge of β_0 and β_1, there will still be uncertainty in prediction).

32

Example 14

Let's return to the carbonation depth-strength data of Example 13 and calculate a 95% PI for a strength value that would result from selecting a single core specimen whose depth is 45 mm. Relevant quantities from that example are

$\hat{y} = 13.79$ $s_{\hat{Y}} = .7582$ $s = 2.8640$

For a prediction level of 95% based on $n - 2 = 16$ df, the t critical value is 2.120, exactly what we previously used for a 95% confidence level.

33

Example 14

cont'd

The prediction interval is then

$$13.79 \pm (2.120)\sqrt{(2.8640)^2 + (.7582)^2} = 13.79 \pm (2.120)(2.963)$$
$$= 13.79 \pm 6.28$$
$$= (7.51, 20.07)$$

Plausible values for a single observation on strength when depth is 45 mm are (at the 95% prediction level) between 7.51 MPa and 20.07 MPa.

The 95% confidence interval for mean strength when depth is 45 was (12.18, 15.40). The prediction interval is much wider than this because of the extra $(2.8640)^2$ under the square root.

34

Example 14

cont'd

Figure 12.18, the Minitab output in Example 13, shows this Interval as well as the confidence interval.

```
The regression equation is strength = 27.2 - 0.298 depth

Predictor       Coef         Stdev      t-ratio       P
 Constant      27.183        1.651        16.46   0.000
   depth     -0.29756      0.04116        -7.23   0.000
s = 2.864     R-sq = 76.6%     R-sq(adj) = 75.1%

Analysis of Variance

SOURCE          DF          SS          MS          F          P
Regression       1      428.62      428.62      52.25      0.000
     Error      16      131.24        8.20
     Total      17      559.86

   Fit     Stdev.Fit           95.0% C.I.           95.0% P.I.
13.793         0.758     (12.185, 15.401)      (7.510, 20.075)

   Fit     Stdev.Fit           95.0% C.I.           95.0% P.I.
16.768         0.678     (15.330, 18.207)     (10.527, 23.009)
```

Minitab regression output for the data of Example 13

Figure 12.18

35