Figure 13.10. SIMPLE LINEAR REGRESSION: Inference for Relationships

Program 25 in: *Against All Odds: Inside Statistics*

Association between two *categorical* variables is displayed in a *two-way table*, and the null hypothesis that no association is present is tested by the chi-squared test discussed in Program 24; association between two *quantitative* variables is displayed by a *scatterplot.* If the association is *linear,* it is described by a *regression line* and its strength is measured by the correlation coefficient. This program discusses inference about straight-line relationships. You may want to review least squares regression from Programs 8 and 9 before studying this program. Before using the inference methods in this program, check that the data show a straight-line relationship and look for *outliers* and *influential observations.*

Confidence intervals and tests always make statements about attributes of the *respondent population* from which our data are a *sample.* To this point, we have learned methods for inference about population *averages* and population *proportions.* To talk about inference in the linear regression setting, we must first say what attributes we will draw conclusions about. The statistical model for *simple linear regression* says that, in the respondent population, the average of the response variable **Y** depends on the explanatory variable **X** in a straight-line fashion. The notation $\mu_Y(\mathrm{x})$ stands for *the mean of the random variable Y representing the response variable* **Y** *when the explanatory variable* **X** *has the value* x. The model for the population says, first, that $\mu_Y = \beta_0 + \beta_1\mathrm{x}$; the slope β_0 and intercept β_1 of this population regression line are unknown parameters we want to draw conclusions about.

When we have n observations on the variables **X** and **Y**, we estimate the unknown β_1 and β_0 by the slope b_1 and intercept b_0 of the *least squares regression line*; *i.e.*, we let the least squares line we fit to the data estimate the unknown population regression line. Notice that, to emphasize the *b*s are *estimates* of the βs, we use different notation from Program 8: the least squares line is now $\hat{\mathrm{y}} = b_0 + b_1\mathrm{x}$, instead of $\mathrm{y} = a + b\mathrm{x}$.

Estimating β_1 and β_0 by b_1 and b_0 is much like using the sample average $\overline{\mathrm{y}}$ to estimate a model parameter μ representing a respondent population average. In both cases, we need information about the *distribution* of the population in order to move on to confidence intervals or tests of significance. The rest of the statistical model for regression says that the random variable Y_j representing the response y_j for the value x_j of the explanatory variable has a *normal* distribution, and that the *standard deviation* σ of Y_j is the *same* for *all* values of x_j; by contrast, the *mean* of Y_j *does* change with x_j – it is $\beta_0 + \beta_1\mathrm{x}_j$.

The standard deviation σ describes how *variable* the response is. Variation is not measured around a *fixed* centre but about the population *regression line*, because the average of **Y** *changes* with **X**. To give a test or confidence interval, we must estimate σ. Recall the *estimated residuals* $\hat{\mathrm{r}}_\mathrm{j}$ = *observed* y – *predicted* y = $\mathrm{y_j} - \hat{\mathrm{y}}_\mathrm{j}$. The square of the sample standard deviation (*i.e.*, the sample variance) of Y_j about the least squares line is as shown at the right; this variance (MSE or s^2) is our estimate of σ^2 (so *s* estimates σ). The divisor n – 2 in the expression at the right is the *degrees of freedom* of s^2 and of *s*.

$$\mathrm{MSE} \equiv s^2 = \frac{\Sigma\hat{\mathrm{r}}_\mathrm{j}^2}{\mathrm{n}-2}$$

The preliminary calculations in a regression problem are: first, calculate the least squares line $\hat{\mathrm{y}} = b_0 + b_1\mathrm{x}$; second, calculate the estimated residuals from this line and the standard deviation *s* (or obtain MSE from the ANOVA table). These calculations are quite lengthy, so that you should use a statistical calculator or software whenever possible; in exercises, you may be given this basic information.

Several kinds of inference are possible in the regression setting; this program presents the two most important: *inference about the slope* β_1 of the population line, and *prediction* of the response for a given **X**. Both kinds of inference use *t* procedures and the *t* distribution with n – 2 degrees of freedom. Although the formulae are more complicated, the ideas are similar to *t* procedures for the model parameter μ representing a respondent population average. The similarity is due to the fact that the estimated slope b_1 and the predicted response $\hat{\mathrm{y}} = b_0 + b_1\mathrm{x}$ both have *normal* distributions. You can find detailed formulae in the Course Materials (or the Text). All make use of the basic calculations of b_1, b_0 and *s*.

Because the slope β_1 is the change in the *average* of **Y** when **X** changes by 1, it is the most important parameter that describes the relationship between **Y** and **X**. The $100(1-\alpha)\%$ confidence interval for β_1 is as given in the *upper* expression at the right. Throughout this Program, ${}_{\alpha}t^*_{\mathrm{n}-2}$ is the upper $\alpha/2$ critical value of the $t_{\mathrm{n}-2}$ distribution; $\hat{s.d.}(b_1)$ is the estimated standard deviation (or standard error) of the statistic b_1. To test the null hypothesis H_o: $\beta_1 = 0$, use the *t*-statistic given at the *lower* right; H_o states that there is *no* linear relationship between **Y** and **X**, or that straight-line dependence on **X** is of no value in *predicting* (or 'explaining') **Y**. Inference about the intercept b_0 is similar but is less often important.

$$b_1 \pm {}_{\alpha}t^*_{\mathrm{n}-2} \times \hat{s.d.}(b_1)$$

$$\frac{b_1}{\hat{s.d.}(b_1)} \sim t_{\mathrm{n}-2}$$

To predict the *mean response* $\mu_Y(\mathrm{x} = \mathrm{x}^*)$ for a given value x* of the explanatory variable **X**, we use the value $\hat{\mu} = b_0 + b_1\mathrm{x}^*$ from the least squares line when **X** is x*. The confidence interval for the mean response is as shown at the right.

$$\hat{\mu} \pm {}_{\alpha}t^*_{\mathrm{n}-2} \times \hat{s.d.}(\hat{\mu})$$

Instead of predicting the mean (long-term average) response, we may wish to predict the *individual* response **Y** on a particular (future) occasion when **X** has the value x*. Again use the predicted value from the least squares line: $\hat{\mathrm{y}} = b_0 + b_1\mathrm{x}^*$. There

(*continued overleaf*)

is more uncertainty in predicting a *single* Y than in predicting the mean response μ_Y when x* is the value of **X**. So the *prediction interval* for Y is *wider* than the confidence interval for μ_Y; it has the form shown at the right below. In these intervals, the *point estimates* $\hat{\mu}$ and $\hat{y}$ are the *same* value – the notation is just to remind us what we are using the least squares line to predict: an *average* response and an *individual* response. The estimated standard deviation (or standard error) $\widehat{s.d.}(\hat{y})$ is similar to $\widehat{s.d.}(\hat{\mu})$ but has an *extra* term that reflects the additional uncertainty in predicting an *individual* observation.

$$\hat{\mu} \pm {}_{\alpha}t^{*}_{n-2} \times \widehat{s.d.}(\hat{y})$$