

## Regression Analysis II ... Inference in SLR

Every use of probability in data analysis is premised on some *model*. In Sessions 3 and 4 we saw that to support one- and two-sample inference methods, it is common to use models of "random sampling from a fixed (and often normal) population/universe." To go from the descriptive statistics of Session 6 to inference methods for simple linear regression, we must similarly adopt a probability model. The most convenient/common such model is the *normal simple linear regression model*. This can be described in several different ways.

**In words, the (normal) simple linear regression model is:**

The relationship between  $x$  and average  $y$  is linear (that is,  $\mu_{y|x} = \beta_0 + \beta_1 x$ ) and for a given  $x$ ,  $y$  is normally distributed with standard deviation  $\sigma$  (that doesn't depend upon  $x$ )

1

Pictorially, the (normal) simple linear regression model is:

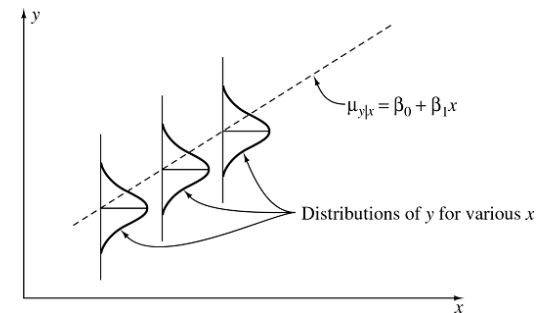


Figure 1: The Normal Simple Linear Regression Model

2

**In symbols, the (normal) simple linear regression model is:**

$$\begin{aligned} y &= \mu_{y|x} + \epsilon \\ &= \beta_0 + \beta_1 x + \epsilon \end{aligned}$$

In this equation,  $\mu_{y|x} = \beta_0 + \beta_1 x$  lets the average of the response  $y$  change with the predictor  $x$ , and  $\epsilon$  allows the observed response to deviate from the relationship between  $x$  and average  $y$  ... it is the difference between what is observed and the average response for that  $x$ .  $\epsilon$  is assumed to be normally distributed with mean 0 and standard deviation  $\sigma$  (that doesn't depend upon  $x$ ).

The normal simple linear regression model is not really that much different from what was used in Sessions 3 and 4, except that now the mean response is allowed to change with the value of the predictor variable. *When it is a sensible description of a real world scenario*, it supports a variety of useful methods of statistical inference with important real world implications.

3

**Point (Single Number) Estimates of Model Parameters  $\beta_0$ ,  $\beta_1$ , and  $\sigma$**

The values  $\beta_0, \beta_1, \sigma$  are "parameters" or characteristics of the SLR model that must be estimated from  $n$  data pairs  $(x_i, y_i)$ .

- $\beta_1$  is the rate of change of average  $y$  with respect to  $x$ . It is the increase in mean  $y$  that accompanies a 1 unit increase in  $x$ .
- $\beta_0$  is the  $y$  intercept of the linear relationship between  $x$  and average  $y$ .
- $\sigma$  measures the variability in  $y$  for any fixed value of  $x$ .

4

As estimates of the model parameters  $\beta_0$  and  $\beta_1$  we'll use the intercept and slope of the least squares line,  $b_0$  and  $b_1$  from Session 6. To estimate  $\sigma$  (that measures the spread in the  $y$  distribution for any given  $x$ ) we'll use

$$\begin{aligned} "s" &= \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} \\ &= \sqrt{\frac{SSE}{n - 2}} \end{aligned}$$

This is a kind of "square root of an average squared vertical distance from plotted  $(x, y)$  points to the least squares line." JMP calls this quantity the "root mean square error" for the regression analysis. We've (temporarily) put "s" in quote marks to alert the reader that this is NOT the sample standard deviation from Session 1, but rather a new kind of sample standard deviation crafted specifically for SLR.

*Example* In the Real Estate Example, applying the JMP results, we will estimate  $\beta_1$  as 16.0081,  $\beta_0$  as 1.9001, and  $\sigma$  as 3.53. The interpretation of the last of these is that one's best guess is that the standard deviation of  $y$  (price in \$1000) for any fixed home size is about 3.53 (\$1000) or \$3,530.

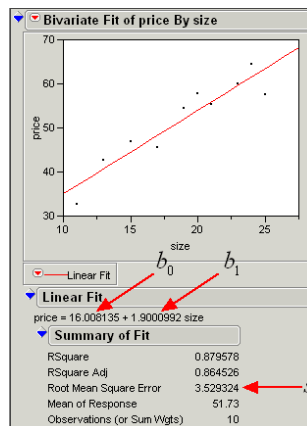


Figure 2: JMP Report for SLR of Price by Size

## Confidence and Prediction Limits

A beauty of adopting the SLR model is that (*when it is a sensible description of a real situation*) it supports the making of statistical inferences, and in particular the making of confidence (and prediction) limits. It, for example, allows us to quantify how precise  $b_0$ ,  $b_1$ , and  $s$  are as estimates of respectively  $\beta_0$ ,  $\beta_1$ , and  $\sigma$ .

Consider first the making of **confidence limits for  $\sigma$** . These can be based on a  $\chi^2$  distribution with  $df = n - 2$ . That is, one may use the interval

$$\left( s\sqrt{\frac{n-2}{U}}, s\sqrt{\frac{n-2}{L}} \right)$$

for  $U$  and  $L$  percentage points of the  $\chi^2$  distribution with  $df = n - 2$ .

*Example* In the Real Estate Example, since there are  $n = 10$  homes represented in the data set, a 95% confidence interval for  $\sigma$  is

$$\left( 3.53\sqrt{\frac{10-2}{17.535}}, 3.53\sqrt{\frac{10-2}{2.180}} \right)$$

that is,

$$(2.38, 6.76)$$

One's best guess at the standard deviation of price for a fixed home size is 3.53, but to be "95% sure" one must hedge that estimate down to 2.38 and up to 6.76.

Next, consider **confidence limits for  $\beta_1$** , the rate of change of mean  $y$  with respect to  $x$ , the change in mean  $y$  that accompanies a unit change in  $x$ . Confidence limits are

$$b_1 \pm tSE_{b_1}$$

9

where  $t$  is a percentage point of the  $t$  distribution with  $df = n - 2$  and  $SE_{b_1}$  is a standard error (estimated standard deviation) for  $b_1$ ,

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Notice that  $\sum (x_i - \bar{x})^2 = (n - 1) s_x^2$ . So the denominator of  $SE_{b_1}$  increases as the  $x$  values in the data set become more spread out. That is, the precision with which one knows the slope of the relationship between mean  $y$  and  $x$  increases with the variability in the  $x$ 's.

*Example* In the Real Estate Example, let's make 95% confidence limits for the increase in average price that accompanies a unit increase in size ( $\beta_1$ ). Limits are

$$b_1 \pm t \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

10

Since  $(n - 1) s_x^2 = 9(4.733)^2 = 201.6$ , and  $df = 10 - 2 = 8$ , these limits are

$$1.9001 \pm 2.306 \frac{3.53}{\sqrt{201.6}}$$

that is

$$1.9001 \pm .5732$$

The \$19.001/ft<sup>2</sup> value is (by 95% standards) good to within \$5.73.

JMP automatically prints out both  $SE_{b_1}$  and confidence limits for  $\beta_1$ . For the Real Estate Example, this looks like

11

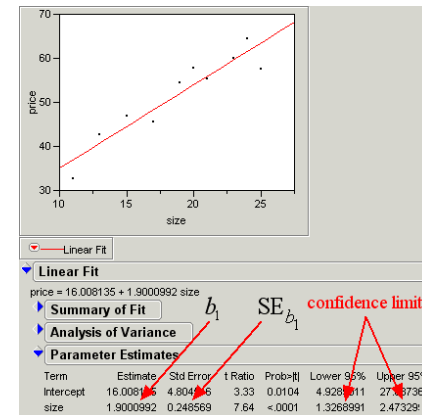


Figure 3:  $b_1, SE_{b_1}$ , and Confidence Limits for  $\beta_1$

12

*Exercise* For the small fake SLR data set, make 95% confidence limits for both  $\sigma$  and  $\beta_1$ .

The SLR model also supports the making of **confidence limits for the average  $y$  at a given  $x$**  ( $\mu_{y|x} = \beta_0 + \beta_1 x$ ). (This could, for example, be the mean price for a home of size  $x = 11$  (100 ft<sup>2</sup>).) These are

$$\hat{y} \pm tSE_{\hat{\mu}}$$

In this formula,  $t$  is a percentage point for the  $t$  distribution for  $df = n - 2$ ,

$$\hat{y} = b_0 + b_1 x$$

and

$$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

13

Here,  $\bar{x}$  and  $\sum (x_i - \bar{x})^2$  come from the  $n$  data pairs in hand, while  $x$  represents the value of the predictor variable at which one wishes to estimate mean response (potentially a value different from any in the data set).

A problem different from estimating the mean response for a given  $x$  is that of predicting  $y_{\text{new}}$  at a particular value of  $x$ . **Prediction limits for  $y_{\text{new}}$  at a given  $x$**  are

$$\hat{y} \pm tSE_{\hat{y}}$$

In this formula,  $t$  and  $\hat{y} = b_0 + b_1 x$  are as above in confidence limits for  $\mu_{y|x}$  and

$$\begin{aligned} SE_{\hat{y}} &= s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \\ &= \sqrt{s^2 + (SE_{\hat{\mu}})^2} \end{aligned}$$

14

*Example* In the Real Estate Example, consider homes of size 2000 ft<sup>2</sup> (homes of size  $x = 20$ ). Let's make a 95% confidence interval for the mean price of such homes, and then a 95% prediction interval for the selling price of an additional home of this size.

First, the predicted selling price of a home of this size is

$$\begin{aligned} \hat{y} &= b_0 + b_1 x \\ &= 16.0081 + 1.9001(20) \\ &= 54.010 \end{aligned}$$

that is, the predicted selling price is \$54,010.

Then, confidence limits for the mean selling price of 2000 square foot homes are

$$\hat{y} \pm tSE_{\hat{\mu}}$$

15

that is,

$$\hat{y} \pm ts \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

Since for the 10 homes in the data set  $\bar{x} = 18.8$ , these limits are

$$54.010 \pm 2.306(3.53) \sqrt{\frac{1}{10} + \frac{(20 - 18.8)^2}{201.6}}$$

or

$$54.010 \pm 2.664$$

Prediction limits for an additional selling price for a home of 2000 square feet are then

$$\hat{y} \pm tSE_{\hat{y}}$$

16

or

$$\hat{y} \pm ts \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

which here is

$$54.010 \pm 2.306 (3.53) \sqrt{1 + \frac{1}{10} + \frac{(20 - 18.8)^2}{201.6}}$$

or

$$54.010 \pm 8.565$$

JMP will plot these confidence limits for mean response and prediction limits (so that they can simply be read from the plot). For the Real Estate Example these look like

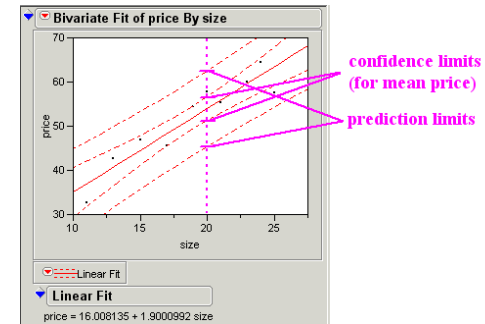


Figure 4: JMP 95% Confidence Limits for  $\mu_{y|x}$  and Prediction Limits for  $y_{new}$

Note (either by looking at the formulas for  $SE_{\hat{\mu}}$  and  $SE_{\hat{y}}$  or plots like the one above) that confidence intervals for  $\mu_{y|x}$  and prediction intervals for  $y_{new}$  are narrowest at the "center of the  $x$  data set," that is where  $x = \bar{x}$ . This makes sense ... one knows the most about home prices for homes that are typical in the data set used to make the intervals!

Note also, that by setting  $x = 0$  in confidence limits for  $\mu_{y|x}$  one can make confidence limits for  $\beta_0$ . In practice this is typically done only when  $x = 0$  is inside the range of  $x$  values in an SLR data set. Where  $x = 0$  is an extrapolation outside the data set in hand, the value  $\beta_0$  is not usually of independent interest. In the case where one does take  $x = 0$ , the corresponding value of  $SE_{\hat{\mu}}$  can be termed  $SE_{b_0}$ .

*Exercise* For the small fake SLR data set, make 95% confidence limits for  $\mu_{y|x}$  and 95% prediction limits for  $y_{new}$  if  $x = 1$ .

## Hypothesis Testing in SLR ... $t$ Tests

The normal SLR model supports the testing of  $H_0 : \beta_1 = \#$  using the statistic

$$t = \frac{b_1 - \#}{SE_{b_1}} = \frac{b_1 - \#}{\frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}}$$

and the  $t$  distribution with  $df = n - 2$  to find  $p$ -values. Similarly, it is possible to test  $H_0 : \mu_{y|x} = \#$  using the statistic

$$t = \frac{\hat{y} - \#}{SE_{\hat{\mu}}} = \frac{\hat{y} - \#}{s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2}}}$$

and the  $t$  distribution with  $df = n - 2$  to find  $p$ -values.

Of these possibilities, only tests of  $H_0 : \beta_1 = 0$  are common. Notice that according to the SLR model

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

and therefore if  $\beta_1 = 0$ ,

$$\mu_{y|x} = \beta_0 + 0 \cdot x = \beta_0$$

and the mean response *doesn't change with  $x$* . So people usually interpret a test of  $H_0 : \beta_1 = 0$  in the SLR model as a way of investigating whether " $x$  has an influence or effect on  $y$ ." Note that when  $\beta_1 \neq 0$  the test statistic for  $H_0 : \beta_1 = 0$  becomes

$$t = \frac{b_1}{SE_{b_1}} = \frac{b_1}{\frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}}$$

*Example* In the Real Estate Example, one might ask whether there is statistically significant evidence that mean price changes with home size, in terms of

21

a test of  $H_0 : \beta_1 = 0$ . Here

$$\begin{aligned} t &= \frac{b_1}{SE_{b_1}} \\ &= \frac{b_1}{\frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}} \\ &= \frac{1.9001}{.248569} \\ &= 7.64 \end{aligned}$$

and the  $p$ -value for  $H_a : \beta_1 \neq 0$  is

$P$  (a  $t$  random variable with  $df = 8$  is  $> 7.64$  or is  $< -7.64$ )

which is less than .0001. This is tiny. We have enough data to see clearly that average price does indeed change with home size. (Notice that this was already evident from the earlier confidence interval for  $\beta_1$ .)

22

JMP produces *two-sided*  $p$ -values for this test directly on its report for SLR. (The  $p$ -value is for  $H_a : \beta_1 \neq 0$ .) Here this looks like

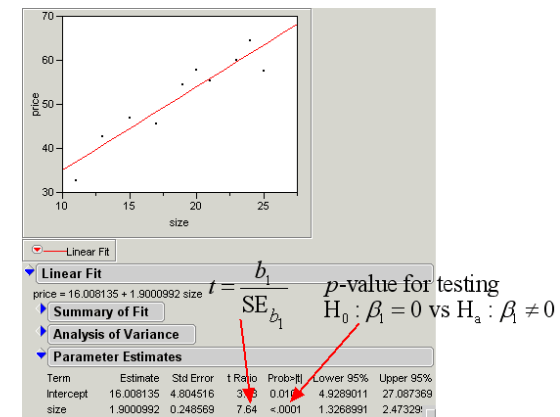


Figure 5: JMP Report and Testing  $H_0 : \beta_1 = 0$

23

24

*Exercise* For the small fake SLR data set, test  $H_0 : \beta_1 = 0$  versus  $H_a : \beta_1 \neq 0$ .