SYSTEMS 302
LECTURE 15

• CONFIDENCE INTERVALS IN SIMPLE LINEAR REGRESSION
  • Confidence Intervals for Betas
  • Confidence Intervals for Conditional Means
  • Prediction Intervals for Individual Values

• EXTENSIONS TO MULTIPLE REGRESSION

• For next time:
  • Devore, Sections 13.4, 8.1
USED-CAR ADVERTISING PROBLEM

An axiom of business is that advertising increases revenues. The question is how much. Suppose we have both revenue, $R_i$, and advertising expenditure data, $A_i$ for 25 used-car dealerships, $i = 1, \ldots, 25$, and consider the linear model:

$$R_i = \beta_0 + \beta_1 A_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

where $\beta_1$ reflects the expected revenue gain for each new dollar of advertising, and where $\beta_0$ is the expected revenue from all other sources.

**Q1.** Can we place confidence bounds on the expected revenue gain from each dollar of advertising?

**Q2.** Can we place confidence bounds on the expected revenue from all sources?
ESTIMATING THE INTERCEPT IN SIMPLE REGRESSION

For any simple linear regression with data \((x_i, y_i), i = 1, \ldots, n\), recall from the normal equations of least squares that

\[
\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} = \frac{1}{n} \sum_{i=1}^{n} Y_i - \bar{x} \sum_{i=1}^{n} w_i Y_i
\]

\[
= \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right) Y_i
\]

where \(w_i = (x_i - \bar{x}) / \sum_{j=1}^{n} (x_j - \bar{x})^2\). So \(\hat{\beta}_0 \sim \text{Normal}\). Also

\[
E\left(\hat{\beta}_0\right) = E\left(\bar{Y} - \hat{\beta}_1 \bar{x}\right) = E\left(\bar{Y}\right) - \bar{x} E\left(\hat{\beta}_1\right)
\]

\[
= (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} = \beta_0
\]

So \(\hat{\beta}_0\) is a linear unbiased estimator of \(\beta_0\), with variance

\[
\text{var}\left(\hat{\beta}_0\right) = \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right)^2 \text{var}(Y_i)
\]

\[
= \sigma^2 \sum_{i=1}^{n} \left( \frac{1}{n} - \bar{x} w_i \right)^2
\]

\[
= \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]
ESTIMATING RESIDUAL STANDARD DEVIATIONS

For any simple linear regression with data \((x_i, y_i), i = 1,..,n\), if we denote the regression residuals by

\[
(1) \quad \hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i), \quad i = 1,..,n
\]

then the natural (unbiased) sample estimate of the residual variance, \(\sigma^2\), is given by

\[
(2) \quad s_n^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2
\]

(\text{where “2” in the denominator again denotes the degrees of freedom lost in fitting the two regression parameters, } \beta_0 \text{ and } \beta_1\). Hence, the associated estimate of the residual standard deviation, \(\sigma\), (called root-mean-square error) is given by

\[
 s_n = \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2}
\]
The least-squares estimates \((\hat{\beta}_0, \hat{\beta}_1)\) for the parameters \((\beta_0, \beta_1)\) in a simple linear regression model with data \((x_i, y_i), i = 1, \ldots, n\) have standardization with respective \(t\)-distributions:

\[
\frac{\hat{\beta}_0 - \beta_0}{s_n \sqrt{\sum_{i=1}^{n} x_i^2 / n \sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim T_{n-2}
\]

and

\[
\frac{\hat{\beta}_1 - \beta_1}{s_n \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \sim T_{n-2}
\]
CONFIDENCE INTERVALS FOR BETA ESTIMATES

For any simple linear regression with data \((x_i, y_i), i = 1, \ldots, n\), if we denote the standard error of \(\hat{\beta}_0\) by

\[
S_{\hat{\beta}_0} = s_n \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} x_i\right)^2}
\]

(1)

and similarly denote the standard error of \(\hat{\beta}_1\) by

\[
S_{\hat{\beta}_1} = \frac{s_n}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

(2)

then the (two-sided) \(100(1 - \alpha)\%\) confidence intervals for \(\beta_0\) and \(\beta_1\) are given by

\[
\left[\hat{\beta}_i \pm t_{\alpha/2, n-2} s_{\hat{\beta}_i}\right], \quad i = 0, 1
\]
ADVERTISING EXAMPLE

In the used-car advertising example with \( \alpha = .05 \), a 95% confidence interval on \( \beta_1 \) (expected revenue generated by each dollar of advertising) is given by

\[
\left[ \hat{\beta}_1 \pm t_{.025,n-2} s_{\hat{\beta}_1} \right]
\]

where in this case \( n = 25, \hat{\beta}_1 = 19.14, s_{\hat{\beta}_1} = 1.53 \), and

\[
t_{.025,n-2} = t_{.025,23} = 2.069 \quad [\text{Table A5}]
\]

imply that

\[
\left[ \hat{\beta}_1 \pm t_{.025,n-2} s_{\hat{\beta}_1} \right] = [19.14 \pm (2.069)(1.53)]
\]

\[
= [15.98, 22.31]
\]

\rightarrow \text{which is precisely the result in JMP.}
CONFIDENCE INTERVALS FOR CONDITIONAL MEANS

For any simple regression with data \((x_i, y_i), i = 1, \ldots, n,\) if we let, 
\[
\bar{x} = \frac{1}{n} \sum_{i} x_i,
\]
and for each \(x\) let \(\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x\) denote the estimated value of the conditional mean, 
\(\mu_{Y|x} = \beta_0 + \beta_1 \cdot x,\) then the standard error of this estimated value is given by

\[
s_{\hat{y}(x)} = s_n \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

For each \(\alpha \in (0, 1),\) the corresponding \(100(1 - \alpha)\%\) confidence interval on the true value of the conditional mean, \(\mu_{Y|x},\) is then given by

\[
\left[ \hat{y}(x) \pm t_{\alpha/2, n-2} s_{\hat{y}(x)} \right]
\]
ROD WEIGHT EXAMPLE

For an observed rough casting weight of $x = 2.72$ oz, determine a 95% confidence interval on the expected finished casing weight, $\mu_{y|x}$. Here the BLU estimator of this conditional mean is given by

(1) $\hat{y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = .308 + .642(2.72) = 2.054$

with associated standard error

(2) $s_{\hat{y}(x)} = s_n \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

To determine this value, observe that since $n = 25$, $\bar{x} = 2.64$, and since the root mean squared error is given [from JMP] by

(3) $s_n = \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 = .0113$

it only remains to calculate leverage, $\Sigma_i (x_i - \bar{x})^2$. But by the identity

(4) $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$
we see that

\[ \sum_{i=1}^{n} (x_i - \bar{x})^2 = (n - 1) s^2 = (24)(.0391)^2 \quad \text{[from JMP]} \]
\[ = .0367 \]

and thus that

\[ s_{Y(2.72)} = (.0113) \sqrt{\frac{1}{25} + \frac{(2.72 - 2.64)^2}{.0367}} = .0052 \]

Finally, since \( t_{0.025,n-2} = t_{0.025,23} = 2.069 \), it follows that the desired 95% confidence interval is given by

\[ 2.054 \pm (2.069)(.0052) = [2.043, 2.065] \]

These sharp bounds are a result of the small root-mean-squared error in the present case.
PREDICTION INTERVALS FOR INDIVIDUAL VALUES

For any simple regression with data \((x_i, y_i), i = 1, \ldots, n\), and \(\bar{x} = \frac{1}{n} \sum_i x_i\), if the random variable, \(\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x\), again denotes the estimated value of the conditional mean at \(x\), then for any new observed value, \(Y\), at \(x\) the standard error of the deviation, \(Y - \hat{Y}(x)\), is given by

\[
S_{Y-\hat{Y}(x)} = \sqrt{S_n^2 + S_{\hat{Y}(x)}^2}
\]

\[
= S_n \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

For each \(\alpha \in (0,1)\), the corresponding \(100(1 - \alpha)\%\) prediction interval (PI) for the realized value of \(Y\) is then given by

\[
\left[ \hat{Y}(x) \pm t_{\alpha/2, n-2} S_{Y-\hat{Y}(x)} \right]
\]
ROD WEIGHT EXAMPLE (Cont'd)

To obtain a 95% prediction interval for finished casting weight, \( Y(x) \) given rough casting weight, \( x = 2.72 \text{ oz} \), we only need to recalculate the relevant prediction standard error as follows:

\[
(2) \quad s_{Y-\hat{Y}(x)} = \sqrt{s_n^2 + s_{\hat{Y}(x)}^2} = s_n \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}
\]

Having already calculated the value, \( s_{\hat{Y}(2.72)} = 0.0052 \), it is seen that the first equality offers the simplest approach in this case, and yields

\[
(3) \quad s_{Y-\hat{Y}(2.72)} = \sqrt{(0.0113)^2 + (0.0052)^2} = 0.0124
\]

Finally, recalling that \( \hat{y}(2.72) = 2.054 \text{ and } t_{.025,23} = 2.069 \), it follows that the desired 95% prediction interval for this case is given by

\[
(5) \quad 2.054 \pm (2.069)(0.0124) = [2.028, 2.079]
\]

These wider bounds emphasize the key difference between individual and mean predictions.
CRITICAL CONFIDENCE EXAMPLE

Suppose that the minimal acceptable finished-casting weight for rods is $y = 2.02\text{ oz}$. How confident can one be that a rough casting weight of $x = 2.72\text{ oz}$ will yield an acceptable rod? Observe first that

$\text{(1)} \quad P \left( \frac{Y - \hat{Y}(x)}{S_{Y - \hat{Y}(x)}} \geq -t_{\alpha,n-2} \right) = 1 - \alpha$

$\Rightarrow P(Y \geq \hat{Y}(x) - t_{\alpha,n-2} S_{Y - \hat{Y}(x)}) = 1 - \alpha$

so that $[\hat{y}(x) - t_{\alpha,n-2} S_{Y - \hat{Y}(x)}, \infty)$ yields a $100(1 - \alpha)$% upper prediction interval for $Y$. Thus the relevant “knife edge” problem here is to find $\alpha$ so that

$\text{(2)} \quad \hat{y}(x) - t_{\alpha,n-2} S_{Y - \hat{Y}(x)} = 2.02$

$\Rightarrow t_{\alpha,n-2} = \frac{\hat{y}(x) - 2.02}{S_{Y - \hat{Y}(x)}} = \frac{2.054 - 2.02}{.0124} = 2.742$

$\Rightarrow \alpha = \alpha(2.742,23) < \alpha(2.7,23) = .006$

$\Rightarrow C^* > 100(1 - .006) = \boxed{99.4\%}$