Regression Confidence Intervals

Daniel Turek

University of Otago

- Given a set of $n = 22$ data points, we calculate the following:
  - $\bar{x} = 8$, $\bar{y} = 16$
  - $\sum(x_i - \bar{x})^2 = 100$
  - $\sum(y_i - \hat{y})^2 = 80$
  - $\sum(x_i - \bar{x})(y_i - \bar{y}) = 150$
- We are conveniently given the t-value: $t_{20} = 2.10$

- Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$
- Write the equation of the best-fit linear regression line
- Find the 95% Confidence Interval for $\hat{\beta}_1$
- Find the best estimate of $\hat{y}$, for a new data point: $x = 2$
- Calculate the 95% Confidence Interval for this prediction, $\hat{y}$

Introduction

- We have looked at ANOVA to test the fit of the model.
- It is also possible to get an idea of the fit of the model, by calculating a 95% confidence interval for the slope of the model.

Example

- If β_1 is the <u>true slope</u> of the regression line then the standard error of β_1 is :

$$\sigma_{\beta_1} = \frac{\sigma_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Example

- Where the variance of the errors σ_e^2 is estimated using the formula:

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}$$

- NOTE: s_e^2 is just MSE
- The divisor is (n-2) rather than (n-1) because we are estimating β_0 and β_1 in $\hat{y}$.

Example

- Therefore, the 95% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{n-2} \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Example

- Easiest way to explain this concept is through an example, so let us return to our height correlation example.
- Using ANOVA we found that the regression effect dominated the residual effect.
- Running a regression we get the line of least squares as:

$$\hat{y} = 107.996 + 0.366x$$

- Where $\hat{y}$ is the height of the child and x is the height of the father and:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 23485.35$$

Example

x_i	y_i	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
172.0	174.0	3.054	9.329
200.0	190.0	8.807	77.556
172.0	158.0	-12.946	167.590
172.0	175.0	4.054	16.438
165.0	155.0	-13.384	179.126
$\vdots$	$\vdots$	$\vdots$	$\vdots$
195.0	176.0	5.054	25.546
176.0	178.0	7.420	55.061
			31742

- Therefore

$$s_e^2 = \frac{31742}{309-2} = 103.39$$

- This is of course the residual mean square

ANOVA example

Source of variation	SS	DF	MS	F
Regression	3146	1	3146	30.429
Residual (Error)	31742	307	**103.39**	
Total	34888	5		

Example

- The standard error of the slope is estimated to be:

$$\frac{s_e}{\sqrt{\Sigma(x_i - \bar{x})^2}} = \frac{\sqrt{103.39}}{\sqrt{23485.35}} = 0.0663$$

- The t-value with 307 df is close to 1.96

Example

- The 95% confidence interval is:

$$\beta_1 \pm t_{307} \times s_{\beta_1}$$
$$0.366 \pm 1.96(0.0663)$$
$$0.366 \pm 0.1300$$

- The confidence interval is

$$0.236 < \beta_1 < 0.496$$

- We can conclude the slope of the regression line is wholly positive, so there is evidence as the father's height increases, the child's height increases.

Hypothesis Test.

- Apart from ANOVA and confidence intervals, we can also use a hypothesis test to check whether or not the relationship is important.
- The general format for the hypothesis test is:

  H_0 $\beta_1 = 0$ (The slope is zero, so y and x are not linearly related.)

  H_A $\beta_1 \neq 0$ (The slope is not zero, so y and x are linearly related.)

Hypothesis Test.

- We need to calculate a test statistic for this hypothesis test. Remember the general form is :

$$t = \frac{\text{Sample statistic - Null value}}{\text{Standard error}}$$

- The t here is short for test statistic, it is **NOT** related to Students t-distribution
- In the case of test statistic for a slope this becomes:

$$t = \frac{\beta_1 - 0}{s.e._{\beta_1}}$$

Hypothesis Test.

- So for the height example this becomes

$$t = \frac{\beta_1 - 0}{s.e._{\beta_1}}$$
$$t = \frac{0.366 - 0}{0.0663}$$
$$t = 5.5204$$

- The p-value for this can be found using RCmdr with the commands

Distributions > Continuous distribution > Normal distribution > Normal probabilities

- the p-value associated with this is 1.6911×10^{-18} which is very significant, giving us the same conclusion that we came up with from the confidence intervals.

Prediction Interval

- We can find the predicted value ($\hat{y}_i$) of a value x_i by substituting it into our regression equation.
- For example say we wanted to predict the height of a child whose father was 175cm tall, we would just use our regression equation.

$$\begin{aligned}\hat{y} &= 107.996 + 0.366x \\ \hat{y}_{175} &= 107.996 + 0.366 \times 175 \\ \hat{y}_{175} &= 172.046\end{aligned}$$

- But what is the error associated with this prediction?

Standard error for a prediction

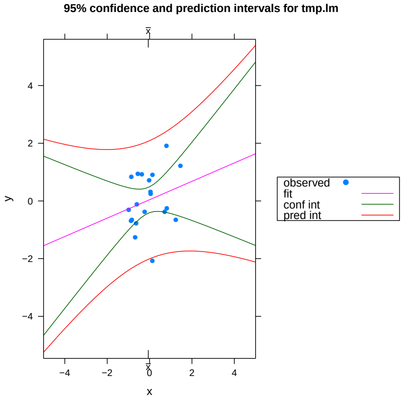
- At value X $= x_k$ say the estimated standard error of the prediction is

$$s_{\hat{y}} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

- where s_e is the residual standard deviation.
- ~~When n is large $s_{\hat{y}}$ will be small and the prediction interval will be approximately $\hat{y} \pm t \times s_e$~~

Note

- The further we get away from the mean x value the wider our prediction interval becomes.



Standard error for a prediction

- For our height example $s_e = \sqrt{103.39} = 10.168$
- If make the prediction for a father of 175cm tall then :

$$s_{\hat{y}} = s_e\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$$

$$s_{\hat{y}} = 10.168\sqrt{1 + \frac{1}{309} + \frac{(175 - 178.493)^2}{23485.35}}$$

$$s_{\hat{y}} = 10.1871$$

Prediction Interval

- We use the same t-value that was used in the confidence of slope. ($t_{307} = 1.96$)

$$\hat{y}_{175} \pm t_{307} \times s_{\hat{y}}$$
$$172.046 \pm 1.96(10.1871)$$
$$172.046 \pm 19.9667$$

- So the prediction interval is

$$152.0793 < \hat{y}_{175} < 192.0127$$

Prediction Interval

- **Important Note**: While the process for constucting a confidence interval and a prediction interval is identical. There is a conceptual difference.
- A confidence interval estimates an unknown population parameter.
- A prediction interval instead estimates the potential data value for an individual.

Summary

Summary. I encourage you to write this down:

Standard error of the residuals: $s_e = \sqrt{\text{MSE}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}$

Standard error of the slope: $\text{se}(\hat{\beta}_1) = \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

Confidence interval for slope: $\hat{\beta}_1 \pm t_{n-2} \times \text{se}(\hat{\beta}_1)$

Standard error for a prediction: $\text{se}(\hat{y}) = s_e\sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}}$

Confidence interval for a prediction: $\hat{y} \pm t_{n-2} \times \text{se}(\hat{y})$

- Given a set of $n = 22$ data points, we calculate the following:
  - $\bar{x} = 8$, $\bar{y} = 16$
  - $\sum(x_i - \bar{x})^2 = 100$
  - $\sum(y_i - \hat{y})^2 = 80$
  - $\sum(x_i - \bar{x})(y_i - \bar{y}) = 150$
- We are conveniently given the t-value: $t_{20} = 2.10$

- Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$
- Write the equation of the best-fit linear regression line
- Find the 95% Confidence Interval for $\hat{\beta}_1$
- Find the best estimate of $\hat{y}$, for a new data point: $x = 2$
- Calculate the 95% Confidence Interval for this prediction, $\hat{y}$