OpenIntro online supplement

More inference for linear regression

Confidence interval for the slope of a regression line

Consider Figure 1 relating Elmhurst College aid and student family income. The equation for the least squares regression line along with the regression summary are shown in the following statistical software output:

```
The regression equation is

aid = 24.31933 - 0.04307 family_income

Predictor        Coef        SE Coef   T         P
Constant         24.31933    1.29145   18.831    < 2e-16
family_income    -0.04307    0.01081   -3.985    0.000229

S = 4.783    R-Sq = 24.86%    R-Sq(adj) = 23.29%
```

The low p-value in the family_income row, $P = 0.000229$, communicates that there is indeed correlation between family income and financial aid at this college, as we'd expect. The slope of the regression equation, $b_1 = -0.04307$ tells us that for each increase of $1,000 in family income, the model predicts that the corresponding financial aid award decreases by $43.07 on average.

● **Example 1** The average decrease of $43.07 is a point estimate b_1 based on a sample of $n = 50$ randomly selected students. Construct a 95% confidence interval (CI) to give an interval estimate of the genuine average decrease β_1, and interpret the meaning of the CI in context.

This CI is called a **t-interval for the slope of the regression line**. The quickest way to check the assumptions for inference is to look at the residual plot for the Elmhurst data. Since we see that data pairs are randomly scattered around the regression line in Figure 1, we can proceed.

We need a CI of this form:

$$\text{point estimate} \pm t^*_{df} \times \text{SE of estimate}$$

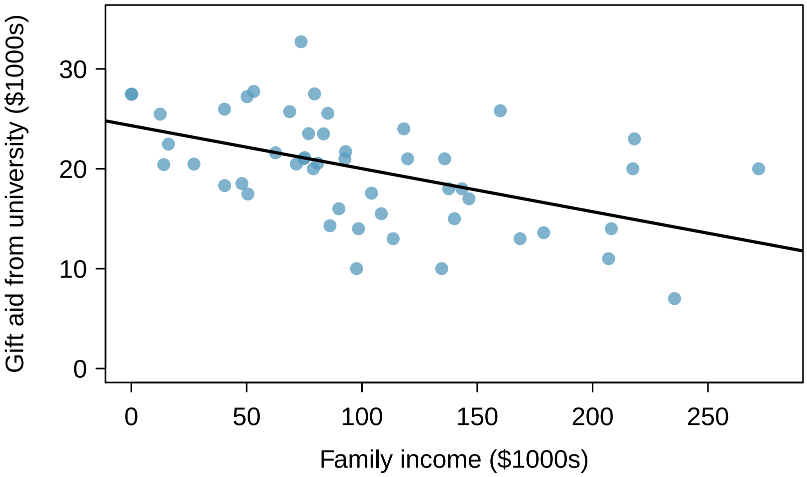

Figure 1: Elmhurst College gift aid and family income for $n = 50$ random first-year students.

with t^* determined by both our confidence level and the degrees of freedom of our regression model. Here, we have point estimate b_1 so the CI takes form:

$$b_1 \pm t^*_{df} \times SE_{b_1}$$

The regression summary provides the slope ($b_1 = -0.04307$) and its standard error ($SE_{b_1} = 0.01081$). With $n = 50$, we have $n-2 = 48$ degrees of freedom. Using either technology or a t-table, we find that with 95% confidence and $df = 48$, the multiplier for SE_{b_1} is $t^*_{48} = 2.01$. So we have the CI

$$-0.04307 \pm 2.01 \times 0.01081 \quad \rightarrow \quad -0.04307 \pm 0.0217.$$

Calculating the bounds for the CI, we have lower bound $-0.04307 - 0.0217 = -0.06477$ and upper bound $-0.04307 + 0.0217 = -0.02137$.

Now we can interpret the meaning of this CI in context. The lower bound -0.06477 represents an aid decrease of \$64.77 for each \$1,000 increase in family income. Now we carefully state our conclusion: We estimate with 95% confidence that for each \$1,000 increase in family income, the average decrease in financial aid is between \$21.37 and \$64.77.

Prediction interval for a response value

Suppose that a student named Cam is planning to attend Elmhurst College this coming fall. Cam would be curious about how much financial aid they are likely to receive based on their family's income of \$52,000. The linear regression model gives a point estimate of Cam's aid award:

$$\hat{y} = 24.31933 - 0.04307(52) = 22.07969 \text{ (in thousands)}$$

For Cam's family income $x^\star = 52$, or \$52,000, the regression model predicts an aid award of $\hat{y}^\star = 22.07969$, or about \$22,080. Here we've marked both the predictor and the response variables with a star to indicate our focus on one specific value of the predictor variable, a \$52,000 family income.

But we know that our regression model is based on a sample of $n = 50$ students, and that this estimate would be (at least a little) different if we had generated the regression model from a different sample of Elmhurst students.

Expecting such sample variation, and confident that our sample data meets the assumptions for inference, we can construct a confidence interval to give an interval estimate of Cam's expected aid. Here, the population parameter of interest is the unknown, specific aid award Cam should expect based on their family's income. We can calculate a 95% CI for that expected aid. This CI is called a **prediction interval for a response value**.

Like many intervals for measurements, this prediction interval takes the form:

$$\text{point estimate} \pm t^*_{df} \times \text{SE of estimate}$$

For a chosen predictor value $x^\star$, our point estimate is $\hat{y}^\star = b_0 + b_1 x^\star$. We find t^*_{df} using the t-distribution with $n - 2$ degrees of freedom. And the formula for the standard error of $\hat{y}^\star$ is

$$SE = \sqrt{s_e^2 + \frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2}$$

where s_e^2 is the variance of the residuals.[1]

This standard error formula is complicated, and a detailed description of its meaning is left to a more advanced statistics course. Here at the introductory statistics level, we should at least have some sense of why there are three (increasingly complicated) terms involved in this SE calculation.

- s_e^2 While estimating variation for a particular response $\hat{y}^\star$ to a chosen $x^\star$, this term represents the uncertainty associated with the residuals, the deviations between observed values and the line itself.

- $\frac{s_e^2}{n}$ Each particular response $\hat{y}^\star$ for each particular predictor $x^\star$ contributes its share to the total variance of the residuals—this share is represented by the term $\frac{s_e^2}{n}$, which is the average contribution to the variance of the residuals from one data pair.

- $(SE_{b_1})^2 \times (x^\star - \bar{x})^2$ The linear regression model is anchored at the point $(\bar{x}, \bar{y})$, and as we choose values further and from the mean $\bar{x}$, we should become less and confident in our predictions $\hat{y}$ for the response variable. This decreasing confidence in the accuracy of the model's predictions as we get further from the "middle" of the scatterplot is captured by the term $(SE_{b_1})^2 \times (x^\star - \bar{x})^2$, which increases in magnitude as our choice of $x^\star$ gets further from the predictor variable's mean $\bar{x}$.

The reasoning above is quite complicated, and understanding why these calculations make sense requires deep study. Don't be concerned if the explanations above aren't intuitive or even entirely understood. Let's return to Cam's family to illustrate the process.

● **Example 2** Construct a 95% prediction interval to estimate the amount of aid Elmhurst will award to Cam, whose family earns \$52,000 per year.

We need to construct a prediction interval for a response value. Here, the predicted response value is Cam's Elmhurst gift aid, $\hat{y}^\star$, considering their family's income of \$52,000, or $x^\star = 52$ for the linear regression model.

[1] Statistical software regression output usually includes the standard deviation of the residuals, often denoted as S. Square that value for s_e^2.

Whenever we construct a prediction interval, we should check that the conditions for fitting the model are met. Since the residuals for the Elmhurst aid vs. income data appear to have no pattern, we can assume that the residuals follow a normal distribution and proceed.

The 95% prediction interval is a CI for a measurement, so it takes the form:

$$\text{point estimate} \pm t^*_{df} \times \text{SE of estimate}$$

Substituting our point estimate $\hat{y}^\star$ and its SE we have:

$$\hat{y}^\star \pm t^*_{df} \times \sqrt{s_e^2 + \frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2}$$

In Example 1, we calculated the point estimate of Cam's gift aid: $\hat{y}^\star = b_0 + b_1 x^\star = 22.07969$. Using technology or a t-table, we also determined that with $df = 50 - 2 = 48$, a 95% confidence level requires $t^*_{48} = 2.011$ for its SE multiplier.

The Elmhurst data regression summary provides the SD of the residuals s_e (denoted by `S` in our summary above) and the SE of the slope SE_{b_1}. both need to be squared for prediction interval calculations:

$$s_e^2 = (4.783)^2 \qquad (SE_{b_1})^2 = (0.01081)^2$$

Last, we need the mean family income from the Elmhurst data set, $\bar{x} = 101.779$.

With all of these values, we carefully calculate[2] Cam's aid prediction interval:

$$\begin{aligned}
&\text{point estimate} \pm t^*_{df} \times \text{SE of estimate}\\
&\qquad \hat{y}^\star \pm t^*_{df} \times \sqrt{s_e^2 + \frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2}\\
&22.0796 \pm 2.011 \times \sqrt{(4.783)^2 + \frac{(4.783)^2}{50} + (0.01081)^2 \times (52 - 101.779)^2}\\
&22.0796 \pm 2.011 \times 4.860\\
&22.0796 \pm 9.772
\end{aligned}$$

Which yields a 95% prediction interval of $(12.308, 31.851)$. So we can predict with 95% confidence that Cam's family can expect Elmhurst to offer between \$12,308 and \$31,851 in gift aid. Equivalently, we could predict with 95% confidence that Cam's family can expect Elmhurst to offer gift aid of about \$22,080 with a margin of error of \$9,772.

Cam's prediction interval has a huge margin of error. With a range of nearly \$20,000, this prediction interval may not help Cam much with their financial planning.

Confidence interval for the mean response value

We know from the Central Limit Theorem that mean values are more predictable than individual measurements. So if Cam and their family consider the *mean* amount of aid

[2] Calculations were performed using a spreadsheet to minimize rounding error. If you check these calculations "by hand," you will notice error in the third decimal place.

that Elmhurst offers families earning \$52,000 per year, then they can expect a smaller margin of error. Suppose that Cam would like to estimate the mean financial aid award for all prospective Elmhurst students whose families earn \$52,000 per year. In addition to the point estimate, we can also construct a **confidence interval for the mean response value** at a specific family income using the following formula:

$$\hat{y}^\star \pm t^*_{df} \times \sqrt{\frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2}$$

where $\hat{y}^\star = b_0 + b_1 x^\star$ is the predicted response for a chosen predictor value $x^\star$.

Wait. Doesn't that CI look familiar? A lot like the prediction interval for a specific response value? The CI we use to estimate mean response to a chosen $x^\star$ stated above uses the same point estimate $\hat{y}^\star$ as our prediction interval for a specific response value. That makes sense because a linear regression model predicts the mean response $\hat{y}$ for any chosen x-value, but we usually read that response as a specific prediction.

See the difference between the prediction interval for a specific value and the confidence interval for the mean response? The SE for the confidence interval for the mean response value lacks the term, s_e^2, that accounts for variation among all of the residuals. Since we're considering the *mean* response for one input, we do not need to include the variance of all the residuals s_e^2 here.

● **Example 3** Construct a 95% confidence interval for the mean amount of aid Elmhurst will award to students whose families earn \$52,000 per year.

The Elmhurst student aid data meets the conditions for inference. Remember that random scatter of data pairs around the regression line—no patterns there.

In our work for Example 2 above, we collected all of the values we need. So we can quickly calculate the 95% CI for the mean response:

$$\begin{aligned}
&\hat{y}^\star \pm t^*_{df} \times \sqrt{\frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2} \\
&22.0796 \pm 2.011 \times \sqrt{\frac{(4.783)^2}{50} + (0.01081)^2 \times (52 - 101.779)^2} \\
&22.0796 \pm 2.011 \times 0.864 \\
&22.0796 \pm 1.738
\end{aligned}$$

The 95% CI is $(20.342, 23.817)$. Again we have a point estimate of \$22,080, but the margin of error for the mean response is only \$1,738. Cam and their family can be 95% confident that the mean gift aid award for Elmhurst students whose families' incomes are \$52,000 is between \$20,342 and \$23,817.

Regression response CIs and extrapolation

Now that we can, for any chosen predictor value $x^\star$, estimate both a specific response to $x^\star$ with a prediction interval or the mean response to $x^\star$ with a CI, we can get a better sense of the utility of the regression model.

$x^\star$	$\hat{y}^\star \pm ME$	95% prediction interval
0	24.3 ± 10.0	$(14.4, 34.3)$
52	22.1 ± 9.8	$(12.3, 31.9)$
100	20.0 ± 9.7	$(10.3, 29.7)$
250	13.6 ± 10.2	$(3.3, 23.8)$
500	2.8 ± 13.0	$(-10.2, 15.8)$

Table 2: 95% prediction intervals for gift aid (in 1,000s of dollars) for students at varying levels of family income.

$x^\star$	$\hat{y}^\star \pm ME$	95 % CI for mean resp.
0	24.3 ± 2.6	$(21.7, 26.9)$
52	22.1 ± 1.7	$(20.3, 23.8)$
100	20.0 ± 1.4	$(18.7, 21.4)$
250	13.6 ± 3.5	$(10.1, 17.0)$
500	2.8 ± 8.8	$(-6.0, 11.5)$

Table 3: 95% CIs for mean gift aid (in 1,000s of dollars) awarded to all Elmhurst students whose families have the indicated income level.

Let's use both inference techniques to estimate both specific gift aid amounts and mean gift aid amounts for Elmhurst students with these family incomes:

$0 - the lowest possible family income

$52,000 - about the median US income (like Cam's family)

$100,000 - close to the mean family income for the sample of 50 Elmhurst students

$250,000 - near the maximum family income from the sample

$500,000 - an income significantly greater than any family's income in the sample

The results are summarized in Table 2 and Table 3. We can see that in both tables, the margin of error increases as the income level gets further from the mean income for the Elmhurst sample data, $101,800.

The further away our chosen incomes are from the mean, the larger our intervals become. For family income $500,000, both intervals include negative values, which make absolutely no sense in this situation. The $500,000 example illustrates why extrapolation can be treacherous. The Elmhurst aid vs. income regression model estimates the mean gift aid fairly well, but not for incomes which are far above those incomes in the 50-student sample.

We can get a visual sense of how these interval estimates change for different x-values from Figure 4. Looking at the Confidence and Prediction bands, it might seem like it's just a better idea to estimate the mean response rather than to predict a specific response. But we should remember that it's important to estimate using the interval that makes sense for the question at hand.

If Cam is curious about how much gift aid they can expect based on their family income, Cam should stick with the prediction interval. As we calculated in Example 2, Cam can be 95% confident that Elmhurst will award them between $12,308 and $31,851 in gift aid. But if Cam wants to consider the average gift aid awarded to all students with the same family income as Cam's family, then Cam should consider the confidence interval for mean response, which we calculated in Example 3, again with 95% confidence, to be between $20,342 and $23,817. Looking at Figure 4, can you see both these intervals?

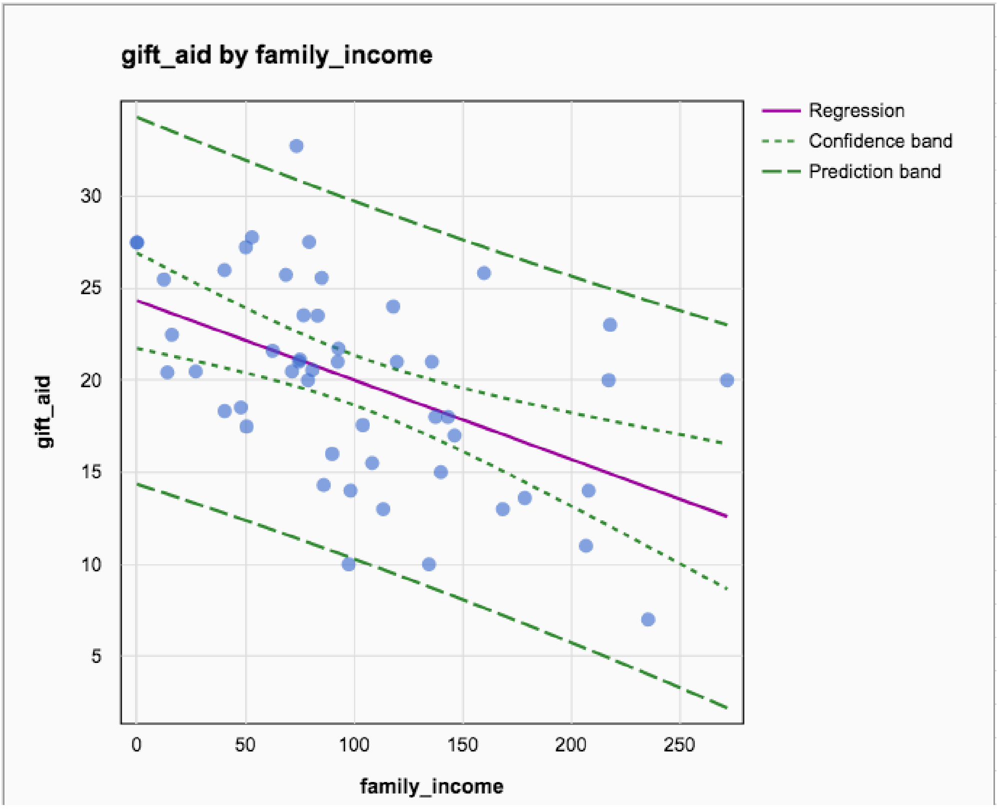

Figure 4: Elmhurst scatterplot with regression line, confidence band, and prediction band

Constructing a prediction interval for a response value

1. State the name of the CI being used.
   - Prediction interval for a response value.
2. Verify conditions.
   - The residual plot has no pattern.
3. Plug in the numbers and write the interval in the form:

   $$\text{point estimate } \pm t^\star \times \text{SE of estimate}$$

   Estimating a specific response value for chosen predictor value $x^\star$,

   - The point estimate is $\hat{y}^\star$.
   - $df = n - 2$
   - The critical value t^* can be found using technology or on the t-table at row $df = n - 2$.
   - $SE = \sqrt{s_e^2 + \frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2}$
4. Evaluate the CI and write in the form (_ , _).
5. Interpret the interval. A generic statement: "We are [XX]% confident that this interval contains the specific response $y^\star$ corresponding to $x^\star$."
6. State a meaningful conclusion to the original question, including context details as needed.

Constructing a confidence interval for the mean response value

1. State the name of the CI being used.
   - Confidence interval for the mean response value.

2. Verify conditions.
   - The residual plot has no pattern.

3. Plug in the numbers and write the interval in the form:

   $$\text{point estimate } \pm t^\star \times \text{SE of estimate}$$

   Estimating the mean response value for chosen predictor value $x^\star$,

   - The point estimate is $\hat{y}^\star$.
   - $df = n - 2$
   - The critical value t^* can be found using technology or on the t-table at row $df = n - 2$.
   - $SE = \sqrt{\frac{s_e^2}{n} + (SE_{b_1})^2 \times (x^\star - \bar{x})^2}$

4. Evaluate the CI and write in the form (_ , _).
5. Interpret the interval. A generic statement: “We are [XX]% confident that this interval contains the mean response value predicted by the regression model to the value $x^\star$.”
6. State a meaningful conclusion to the original question, including context details as needed.

Exercises

1 Murders and poverty, Part IV. Exercise 8.29 (*OpenIntro AHSS*) presents data examining the relationship between poverty and murder. Among 20 randomly selected metropolitan areas, $\bar{x} = 19.720\%$ is the mean percentage of people living in poverty. Regression output for predicting annual murders from percentage living in poverty is shown below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-29.901	7.789	-3.839	0.001
poverty%	2.559	0.390	6.562	0.000

$$s = 5.512 \qquad R^2 = 70.52\% \qquad R^2_{adj} = 68.89\%$$

Find a 95% prediction interval for the annual murders per million in a metropolitan area with

(a) 9% of the population living in poverty (like Washington, D.C.)[3]

(b) 15% of the population living in poverty (about the average poverty rate for US cities)

(c) 40% of the population living in poverty (like Detroit, Michigan)[4]

2 Murders and poverty, Part V. Find a 95% confidence interval for the mean annual murders per million in a metropolitan area with

(a) 9% of the population living in poverty (like Washington, D.C.)

(b) 15% of the population living in poverty (the average poverty rate for US cities)

(c) 40% of the population living in poverty (like Detroit, Michigan)

3 Murders and poverty, Part VI. In 2014, there were 105 murders in Washington, D.C., or 165.7 murders per million.[5] Washington, D.C., has a low poverty rate of 9%.

(a) Use the regression model to calculate the residual for Washington, D.C., in 2014. Describe the meaning of the residual.

(b) Does Washington, D.C.'s prediction interval from Exercise 0.1(a) capture its genuine number of murders per million from 2014?

(c) Does the confidence interval for the mean annual murders per million for metropolitan areas with 9% poverty from Exercise 0.2 capture Washington, D.C.'s genuine number of murders per million from 2014?

(d) Why do you think that this linear regression model succeeds or fails to capture the actual data from Washington, D.C. in 2014?

4 Cats, Part III. Exercise 8.30 (*OpenIntro AHSS*) presents regression output from a model for predicting the heart weight (in g) of cats from their body weight (in kg). The model is based on data from a sample of 144 domestic cats with mean weight $\bar{x} = 2.724$ kg. The regression output is provided below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.357	0.692	-0.515	0.607
body wt	4.034	0.250	16.119	0.000

$$s = 1.452 \qquad R^2 = 64.66\% \qquad R^2_{adj} = 64.41\%$$

Find a 95% prediction interval for the heart weight (in g) of a cat if you know that its weight is

(a) 5.5 kg (the weight of this writer's cat, Dale)

[3] http://www.huffingtonpost.com/2011/10/25/cities-poverty-rates-lowest-census_n_1031285.html

[4] http://www.cbsnews.com/media/americas-11-poorest-cities/

[5] http://wtop.com/local/2016/03/d-c-baltimore-city-among-top-murder-capitals-u-s/

(b) 9 kg (the weight of a very large, healthy cat)

(c) 21 kg (the weight of the heaviest known cat)

5 Cats, Part IV. Refer again to the regression output for the model that predicts cats' heart weights (in g) from the weights (in kg) provided above in Exercise 0.4.

Find a 95% confidence interval for the mean heart weight (in g) of cats that weigh

(a) 5.5 kg (the weight of this writer's cat, Dale)

(b) 9 kg (the weight of a very large, healthy cat)

(c) 21 kg (the weight of the heaviest known cat)

6 Cats, Part V. Consider the prediction intervals from Exercise 0.4 and the confidence intervals from Exercise 0.5. Should we trust that all of those intervals are accurate? (If you don't know where to start, do a little online research to find out what range of cats' heart weights are reasonable.)

Solutions

Inference for linear regression

1

(a) $SE = \sqrt{(5.512)^2 + \frac{(5.512)^2}{20} + (0.390)^2 \times (9 - 20.570)^2} = 7.03$

The prediction interval is -6.87 ± 14.763, or $(-21.63, 7.89)$. With 95% confidence, the regression model predicts fewer than 7.9 murders per million for a metropolitan area with 9% of its people living in poverty. (We haven't stated the lower bound for this interval, since "-22 murders per million" makes no sense.)

(b) $SE = \sqrt{(5.512)^2 + \frac{(5.512)^2}{20} + (0.390)^2 \times (15 - 20.570)^2} = 5.94$

The prediction interval is 8.48 ± 12.481, or $(-4.00, 20.96)$. When a metropolitan area has 15% of its people living in poverty, we can say with 95% confidence, the regression model predicts 8.48 murders per million with a margin of error of 12.48 murders per million. (Notice the lower bound will be negative here, too.)

(c) $SE = \sqrt{(5.512)^2 + \frac{(5.512)^2}{20} + (0.390)^2 \times (40 - 20.570)^2} = 9.72$

The 95% prediction interval is 72.46 ± 20.419, or $(52.04, 92.88)$. With 95% confidence, the regression model predicts between 52.0 and 92.9 murders per million for a metropolitan area with 40% of its people living in poverty.

2

(a) $SE = \sqrt{\frac{(5.512)^2}{20} + (0.390)^2 \times (9 - 20.570)^2} = 4.36$

The confidence interval is -6.87 ± 9.157, or $(-16.03, 2.29)$. With 95% confidence, the regression model predicts that the mean number of murders per million for metropolitan areas with 9% of people living in poverty will be fewer than 2.3 murders per million.

(b) $SE = \sqrt{\frac{(5.512)^2}{20} + (0.390)^2 \times (15 - 20.570)^2} = 2.22$

The 95% confidence interval is 8.48 ± 4.654, or $(3.83, 13.14)$. With 95% confidence, the regression model predicts that the mean number of murders per million for metropolitan areas with 15% of people living in poverty will be between 3.8 and 13.1 murders per million.

(c) $SE = \sqrt{\frac{(5.512)^2}{20} + (0.390)^2 \times (40 - 20.570)^2} = 8.00$

The 95% confidence interval is 72.46 ± 16.817, or $(55.64, 89.28)$. With 95% confidence, the regression model predicts that the mean number of murders per million for metropolitan areas with 40% of people living in poverty will be between 55.6 and 89.3 murders per million.

3

(a) The residual calculation for Washington, D.C. is

$$residual = 165.7 - (-6.87) = 172.57.$$

That means that there were about 173 more murders per million in Washington, D.C., in 2014 than this regression model predicts.

(b) Washington, D.C.'s prediction interval, $(-21.63, 7.89)$, does not capture its genuine number of murders, about 166 per million, from 2014. (It's not even close.)

(c) The confidence interval for the mean annual murders per million for metropolitan areas with 9% poverty, $(-16.03, 2.29)$, does not capture Washington, D.C.'s 173 murders per million from 2014.

(d) The regression model uses many cities to capture a general trend. In 2014, Washington, D.C., was outstanding both for its low poverty rate and its high number of murders per million. This kind of outlier won't be predicted by a linear regression model.

4

(a) $SE = \sqrt{(1.452)^2 + \frac{(1.452)^2}{144} + (0.25)^2 \times (5.5 - 2.724)^2} = 1.61$

The prediction interval is 21.83 ± 3.191, or $(18.64, 25.02)$. With 95% confidence, the regression model predicts that a 5.5 kg (body weight) cat's heart will weigh between 18.6 g and 25.0 g.

(b) $SE = \sqrt{(1.452)^2 + \frac{(1.452)^2}{144} + (0.25)^2 \times (9 - 2.724)^2} = 2.14$

The 95% prediction interval is 35.95 ± 4.233, or $(31.72, 40.18)$. With 95% confidence, the regression model predicts that a 9 kg (body weight) cat's heart will weigh between 31.7 g and 40.2 g.

(c) $SE = \sqrt{(1.452)^2 + \frac{(1.452)^2}{144} + (0.25)^2 \times (21 - 2.724)^2} = 4.80$

The 95% prediction interval is 84.36 ± 9.481, or $(74.88, 93.84)$. With 95% confidence, the regression model predicts that a 21 kg (body weight) cat's heart will weigh between 74.9 g and 93.8 g.

5

(a) $SE = \sqrt{\frac{(1.452)^2}{144} + (0.25)^2 \times (5.5 - 2.724)^2} = 0.70$

The confidence interval is 21.83 ± 1.393, or $(20.44, 23.22)$. With 95% confidence, the regression model predicts that the mean heart weight of cats with body weight 5.5 kg is between 20.4 g and 23.2 g.

(b) $SE = \sqrt{\frac{(1.452)^2}{144} + (0.25)^2 \times (9 - 2.724)^2} = 1.57$

The 95% confidence interval is 35.95 ± 3.111, or $(32.84, 39.06)$. With 95% confidence, the regression model predicts that the mean heart weight of cats with body weight 9 kg is between 32.8 g and 39.1 g.

(c) $SE = \sqrt{\frac{(1.452)^2}{144} + (0.25)^2 \times (21 - 2.724)^2} = 4.57$

The 95% confidence interval is 84.36 ± 9.036, or $(75.32, 93.39)$. With 95% confidence, the regression model predicts that the mean heart weight of cats with body weight 5.5 kg is between 75.3 g and 93.4 g.

6 An online search for "cat heart weight" reveals that healthy cat hearts typically weigh less than 20 g (and less than 40 g for cats with heart conditions). That means that many of these intervals are outside the realm of possibility. Since our choices for cat body weight are outside the weights in the data set, this is no surprise. Extrapolating using a linear regression model to x-values outside the data set usually gives bad predictions.

References

[These still need details and formatting.]

Mosteller et al, *Beginning Statistics with Data Analysis*, etc.

authors, *Dictionary/Outline of Basic Statistics*, etc.