Stat 6021: Addressing Guided Question Set for Module 4: Inference with Simple Linear Regression

Tom Lever

09/17/22

We will continue to use the dataset `penguins` from the `palmerpenguins` package. In the previous guided question set, we explored the linear relationship between body mass and flipper length of adult Gentoo penguins near Palmer Station, Antarctica.

1. Produce a scatterplot of body mass and flipper length for adult Gentoo penguins. Write the estimated linear regression equation.

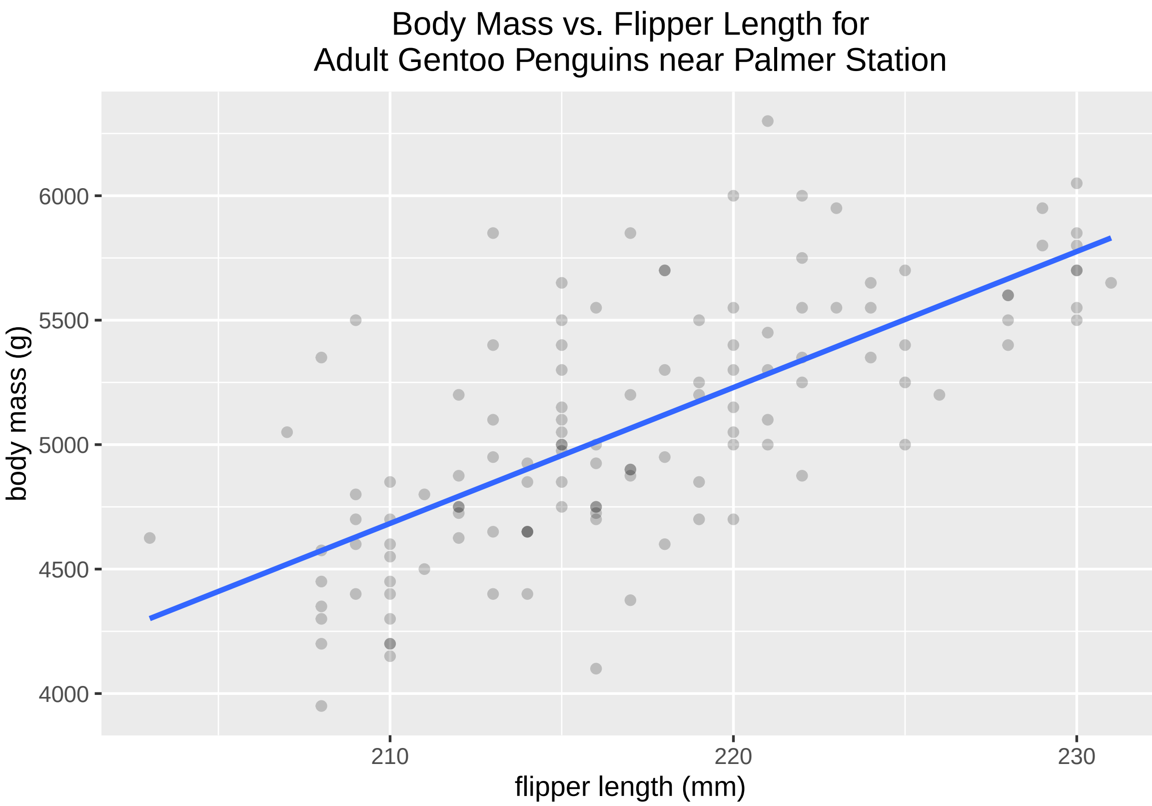
```
library(palmerpenguins)
library(dplyr)
library(ggplot2)
library(TomLeversRPackage)

species_flipper_length_and_body_mass <-
    palmerpenguins::penguins %>%
        select(species, flipper_length_mm, body_mass_g) %>%
        filter(!is.na(flipper_length_mm))
head(species_flipper_length_and_body_mass, n = 3)
```

```
## # A tibble: 3 x 3
##   species flipper_length_mm body_mass_g
##   <fct>               <int>       <int>
## 1 Adelie                181        3750
## 2 Adelie                186        3800
## 3 Adelie                195        3250
```

```
data_set <-
    species_flipper_length_and_body_mass %>% filter(species == "Gentoo")
ggplot(data_set, aes(x = flipper_length_mm, y = body_mass_g)) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "flipper length (mm)",
        y = "body mass (g)",
        title = paste(
            "Body Mass vs. Flipper Length for\n",
            "Adult Gentoo Penguins near Palmer Station",
            sep = ""
        )
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
```

```
    )
```




```
linear_model <- lm(body_mass_g ~ flipper_length_mm, data = data_set)
summarize_linear_model(linear_model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = data_set)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.18 -235.76  -51.93  170.75 1015.71
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6787.281   1092.552  -6.212 7.65e-09 ***
## flipper_length_mm    54.623      5.028  10.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.2 on 121 degrees of freedom
## Multiple R-squared:  0.4937, Adjusted R-squared:  0.4896
## F-statistic:   118 on 1 and 121 DF,  p-value: < 2.2e-16
##
## E(y | x) = B_0 + B_1 * x = -6787.281 + 54.623 * x
## Number of observations: 123
## Estimated variance of errors: 129744.04
## Multiple R:  0.702666524357519   Adjusted R:  0.699714227381436
```

2. For adult Gentoo penguins near Palmer Station, what is the change in the predicted body mass (in grams) when flipper length increases by 1 mm? Also report the corresponding 95-percent confidence interval for the change in the predicted body mass when flipper length increases by 1 mm.

   A point estimate for the change in predicted body mass when flipper length increases by 1 mm is 54.623 g, based on the slope $\hat{\beta}_1 = 54.623 \, \frac{g}{mm}$ of linear model of body mass vs. flipper length for adult Gentoo penguins near Palmer Station.

   The standard error of the statistic / estimate slope $\hat{\beta}_1$

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{MS_{Res}}{S_{xx}}} = \sqrt{\frac{SS_{Res}/(n-2)}{\sum_{i=1}^{n}\left[(x_i - \bar{x})^2\right]}} = \sqrt{\frac{\sum_{i=1}^{n}\left[{e_i}^2\right]}{(n-2)\ \sum_{i=1}^{n}\left[(x_i - \bar{x})^2\right]}}$$

$$SE\left(\hat{\beta}_1\right) = \sqrt{\frac{\sum_{i=1}^{n}\left[(\hat{y}_i - \bar{y})^2\right]}{(n-2)\ \sum_{i=1}^{n}\left[(x_i - \bar{x})^2\right]}}$$

   The confidence interval for the change in the predicted body mass when flipper length increases by 1 mm is the confidence interval for the slope of the linear model and is

$$\left[\hat{\beta}_1 - t_{\alpha/2,\ df}\ SE\left(\hat{\beta}_1\right),\ \hat{\beta}_1 + t_{\alpha/2,\ df}\ SE\left(\hat{\beta}_1\right)\right]$$

```
slope = 54.623
confidence_level = 0.95
significance_level = 1 - confidence_level
number_of_observations = 123
number_of_parameters = 2 # x and y
degrees_of_freedom = number_of_observations - number_of_parameters
qt((1 + confidence_level)/2, degrees_of_freedom)
```

```
## [1] 1.979764
```

```
# quantile t_{\alpha/2, df} for which probability that test statistic is greater
# is half significance level
quantile <- qt(significance_level/2, degrees_of_freedom, lower.tail = FALSE)
quantile
```

```
## [1] 1.979764
```

```
standard_error_of_slope <- 5.028
slope - quantile * standard_error_of_slope
```

```
## [1] 44.66875
```

```
slope + quantile * standard_error_of_slope
```

```
## [1] 64.57725
```

```
confint(linear_model, level = confidence_level)
```

```
##                         2.5 %      97.5 %
## (Intercept)       -8950.27535 -4624.28587
## flipper_length_mm    44.66777    64.57724
```

   Since the confidence interval $\left[44.668 \ \frac{g}{mm},\ 64.577 \ \frac{g}{mm}\right]$ does not contain 0, we reject the null hypothesis $H_0 : \beta = 0 \ \frac{g}{mm}$. We have sufficient evidence to support the alternate hypothesis $H_1 : \beta \neq 0$.

   Since the confidence interval $\left[44.668 \ \frac{g}{mm},\ 64.577 \ \frac{g}{mm}\right]$ contains 50 $\frac{g}{mm}$, we fail to reject the null hypothesis $H_0 : \beta = 50 \ \frac{g}{mm}$.

3. Conduct a hypothesis test to determine whether or not there is a linear association between body mass and flipper length for adult Gentoo penguins near Palmer Station. State the hypothesis, p-value, and conclusion in context.

   Given a significance level 0.05, we test a null hypothesis $H_0 : \beta_1 = 0$ that the slope of a linear model of body mass vs. flipper length for adult Gentoo penguins near Palmer station is equal to 0. If we have sufficient evidence to reject the null hypothesis, we have sufficient evidence to support an alternate hypothesis $H_1 : \beta_1 = 0$ that the slope of the linear model is not equal to 0. We have sufficient evidence to reject the null hypothesis if the test statistic $t_0 = 10.863$ is greater than $t_{\alpha/2,df} = 1.980$. Since $t_0 > t_{\alpha/2,df}$, we reject the null hypothesis. We have sufficient evidence to support the alternate hypothesis.

   $$t_0 = \frac{\hat{\beta}_1^2}{MS_{Res}/S_{xx}} = \frac{\hat{\beta}_1^2 S_{xx}}{MS_{Res}} = \frac{\hat{\beta}_1 S_{xy}}{MS_{Res}} = \frac{MS_R}{MS_{Res}} = F_0$$

   We have sufficient evidence to reject the null hypothesis if the probability $p < 2.2 \times 10^{-16}$ is less than α. Since $p < \alpha$, we reject the null hypothesis. We have sufficient evidence to support the alternate hypothesis.

```
test_statistic <- 10.863
probability <- pt(test_statistic, degrees_of_freedom, lower.tail = FALSE)
probability
```

```
## [1] 6.656439e-20
```

```
test_null_hypothesis_involving_slope(linear_model, 0.05)
```

```
## Since probability 2.2e-16
## is less than significance level 0.05,
## we reject the null hypothesis.
## We have sufficient evidence to support the alternate hypothesis.
```

4. Are your results from parts 2 and 3 consistent?

   Yes; the results of constructing the confidence interval for the slope and testing the null hypothesis support rejecting the null hypothesis and supporting the alternate hypothesis.

5. Estimate the mean body mass for adult Gentoo penguins near Palmer Station with flipper lengths of $x_0 = 200\ mm$. Also report the 95-percent confidence interval for the mean body mass for Gentoo penguins with flipper lengths of x_0.

   A point estimate for the mean body mass for adult Gentoo penguins near Palmer Station with flipper lengths of $x_0 = 200\ mm$ is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \approx (-6787.281\ g) + \left(54.623\ \frac{g}{mm}\right)(200\ mm) = 4137.22\ g$.

   The standard error of the statistic / estimate mean body mass for adult Gentoo penguins near Palmer station with flipper lengths of $x_0 = 200\ mm$ $E(\hat{y} \mid x_0)$

   $$SE\left\{E(\hat{y} \mid x_0)\right\} = \sqrt{MS_{Res}\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$

   The confidence interval for the mean body mass for adult Gentoo penguins with flipper lengths of $x_0 = 200\ mm$ is

   $$\left[E(\hat{y} \mid x_0) - t_{\alpha/2,\ df}\ SE\left\{E(\hat{y} \mid x_0)\right\},\ E(\hat{y} \mid x_0) + t_{\alpha/2,\ df}\ SE\left\{E(\hat{y} \mid x_0)\right\}\right]$$

```
predict(
    linear_model,
    data.frame(flipper_length_mm = 200),
    level = 0.95,
```

```
    interval = "confidence"
)
```

```
##       fit      lwr      upr
## 1 4137.22 3954.446 4319.993
```

6. Report the 95 percent prediction interval for the body mass of an adult Gentoo penguin with flipper length of $x_0 = 200\ mm$.

   The standard error of the prediction body mass for adult Gentoo penguins near Palmer station with flipper lengths of $x_0 = 200\ mm\ E\left(\hat{y} \mid x_0\right)$

$$SE\left\{\hat{y} \mid x_0\right\} = \sqrt{MS_{Res}\left[1 + \frac{1}{n} + \frac{\left(x_0 - \bar{x}\right)^2}{S_{xx}}\right]}$$

   The prediction interval for the body mass for adult Gentoo penguins with flipper lengths of $x_0 = 200\ mm$ is

$$\left[E\left(\hat{y} \mid x_0\right) - t_{\alpha/2,\ df}\ SE\left\{\hat{y} \mid x_0\right\},\ E\left(\hat{y} \mid x_0\right) + t_{\alpha/2,\ df}\ SE\left\{\hat{y} \mid x_0\right\}\right]$$

```
predict(
    linear_model,
    data.frame(flipper_length_mm = 200),
    level = 0.95,
    interval = "prediction"
)
```

```
##       fit      lwr      upr
## 1 4137.22 3401.121 4873.319
```

7. A researcher hypothesizes that for adult Gentoo penguins near Palmer Station, the predicted body mass increases by more than 50 g for each additional millimeter in flipper length. Conduct an appropriate hypothesis test. What are the null and alternate hypotheses, test statistic, and conclusion?

   Given a significance level $\alpha = 0.05$ and assumed linear-model slope $\beta_{10} = 50\ \frac{g}{mm}$, we test a null hypothesis $H_0 : \beta_1 \leq \beta_{10}$ that the predicted body mass increases by a mass less than or equal to 50 g for each additional millimeter in flipper length, and that the slope of the linear model of body mass vs. flipper length is less than or equal to β_{10}. If we have sufficient evidence to reject the null hypothesis, we have sufficient evidence to support an alterate hypothesis $H_1 : \beta_1 > \beta_{10}$ that the slope of the linear model of body mass vs. flipper length is greater than β_{10}. We have sufficient evidence to support an alternate hypothesis $H_1 : \beta_1 > \beta_{10}$ if the test statistic t_0 is greater than $t_{\alpha/2,\ df} = 1.980$, where

$$t_0 = \frac{\hat{\beta}_1 - \beta_{10}}{SE\left(\hat{\beta}_1\right)} = \frac{\left(54.623\ \frac{g}{mm}\right) - \left(50\ \frac{g}{mm}\right)}{\left(5.028\ \frac{g}{mm}\right)} = 0.919$$

   Since $t_0 = 0.919$ is less than $t_{\alpha/2,\ df} = 1.980$, we have insufficient evidence to reject that the null hypothesis that the predicted body mass increases by a mass less than or equal to 50 g for each additional millimeter in flipper length, and that the slope of the linear model of body mass vs. flipper length is less than or equal to β_{10}.

   We have sufficient evidence to reject the null hypothesis if the probability p that a test statistic is greater than the magnitude of t_0, assuming the null hypothesis is true, is less than α.

```
pt(0.919, degrees_of_freedom, lower.tail = FALSE)
```

```
## [1] 0.179962
```

   Since $p = 0.180$ is greater than α, we fail to reject the null hypothesis.