Problem Set 08: Inference for Regression Solutions

Professor Bradley Warner

June, 2022

Documentation:

We used all the resources available to instructors from the authors of Modern Dive.

Introduction

For this problem set you will apply statistical inference to a linear modeling and explore methods to check the required conditions. To start we will build a model using data from the `palmerpenguins` package. The `penguins` data contains size measurements for three penguin species observed on three islands in the Palmer Archipelago, Antarctica.

First we will start with our typical exploratory data analysis and then build our linear model. From there we will use our new skills to make inferences about our regression model and check the necessary conditions.

Setup

First load the necessary packages:

```
library(tidyverse)
library(moderndive)
library(infer)
library(palmerpenguins)
```

The data

```
pen <-penguins  %>%
  filter(!is.na(flipper_length_mm))
```

Take a moment to look at the data in the viewer. The dataset contains 8 variables. You can read more about the variables by typing ?penguins

For our lab we will focus on four variables, the explanatory variables include:

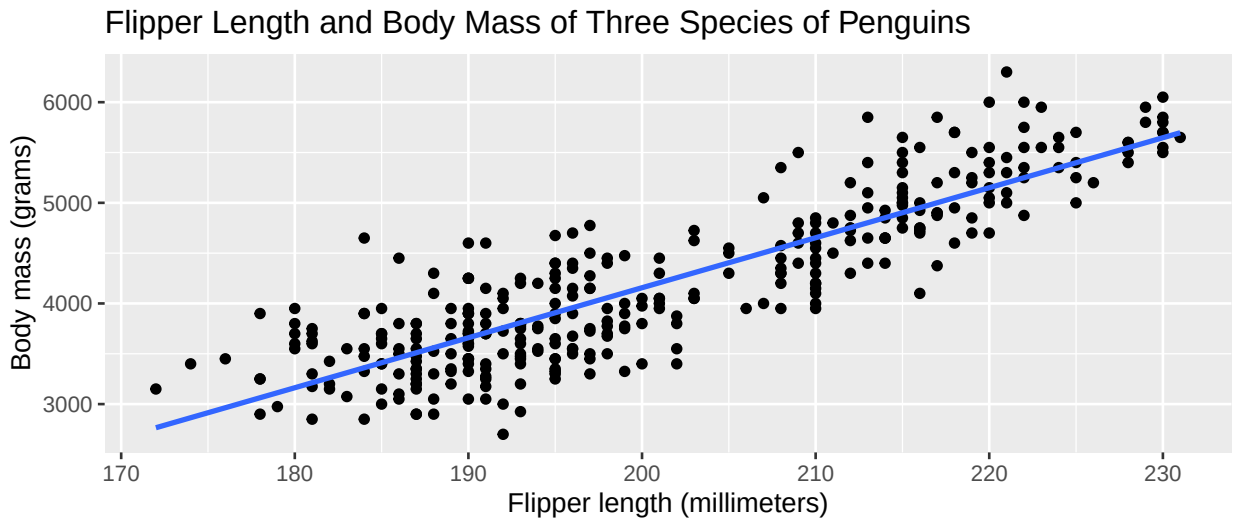
- `flipper_length_mm` - an integer denoting flipper length (millimeters)
- `bill_length_mm` - a number denoting bill length (millimeters)
- `species`- denotes penguin species (Adélie, Chinstrap and Gentoo)

The outcome variable `body_mass_g` is an integer denoting body mass (grams).

Visualization

We will start by investigating the relationship between 'flipper_length_mm' and 'body_mass_g'.

```
ggplot(data = pen, aes(y = body_mass_g, x = flipper_length_mm)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Flipper length (millimeters)",
       y = "Body mass (grams)",
       title = "Flipper Length and Body Mass of Three Species of Penguins")
```



Exercise 1

Does the relationship appear to be positive or negative? Does it look to be reasonably linear?

Answer: The relationship appears both to be linear and positive

Create a linear regression model

```
pen_model <- lm(body_mass_g ~ flipper_length_mm, data = pen)
get_regression_table(pen_model)
```

```
## # A tibble: 2 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept          -5781.     306.       -18.9       0  -6382.   -5179.
## 2 flipper_length_mm     49.7      1.52      32.7       0     46.7     52.7
```

Exercise 2

Write a sentence interpreting the 95% confidence interval for β_1 given in the regression table.

Answer: We are 95% "confident" that the true slope is between 46.7 and 52.7 suggesting that there is a meaningful relationship between body mass and flipper length.

Exercise 3

Recall that the test statistic and p-value correspond to the hypothesis test:

$$H_0 : \beta_1 = 0$$
$$\text{vs } H_A : \beta_1 \neq 0$$

Write up the results & conclusions for this hypothesis test.

Answer: We reject the hypothesis that there is no relationship between flipper length and body mass in favor of the hypothesis that there is a relationship. That is to say, the evidence suggests there is a significant relationship, one that is positive.

You may remember that this hypothesis test is only valid if certain "conditions for inference for regression" are met. Let's take a closer look those conditions.

1. Linearity of relationship between variables
2. Independence of the residuals
3. Normality of the residuals
4. Equality of variance of the residuals

Linearity of relationship between variables

Exercise 4

Did you say that the relationship between `flipper_length_mm` and `body_mass_g` appears to be linear?

Answer: Yes, we saw the linear relationship in the scatterplot.

Independence of the residuals

The observations in our data must be independent of one another. In this data, we can not be sure this is case, for example, some of the penguins included may be related (siblings, parents). We are not given enough information to verify this condition has been met.

Normality of the residuals

The third condition is that the residuals should follow a Normal distribution centered 0. To check for normality, create a histogram.

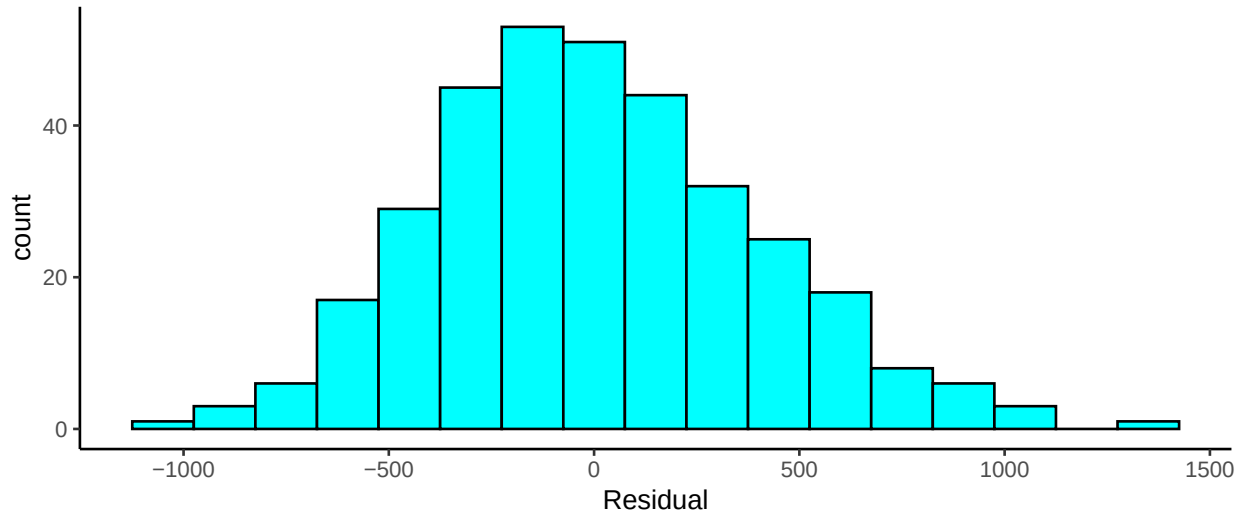
The code to get the residuals is given.

```
regression_points <- get_regression_points(pen_model)
```

Exercise 5

Add code for the histogram below.

```
ggplot(regression_points, aes(x = residual)) +
  geom_histogram(binwidth = 150, color = "black", fill="cyan") +
  labs(x = "Residual") +
  theme_classic()
```

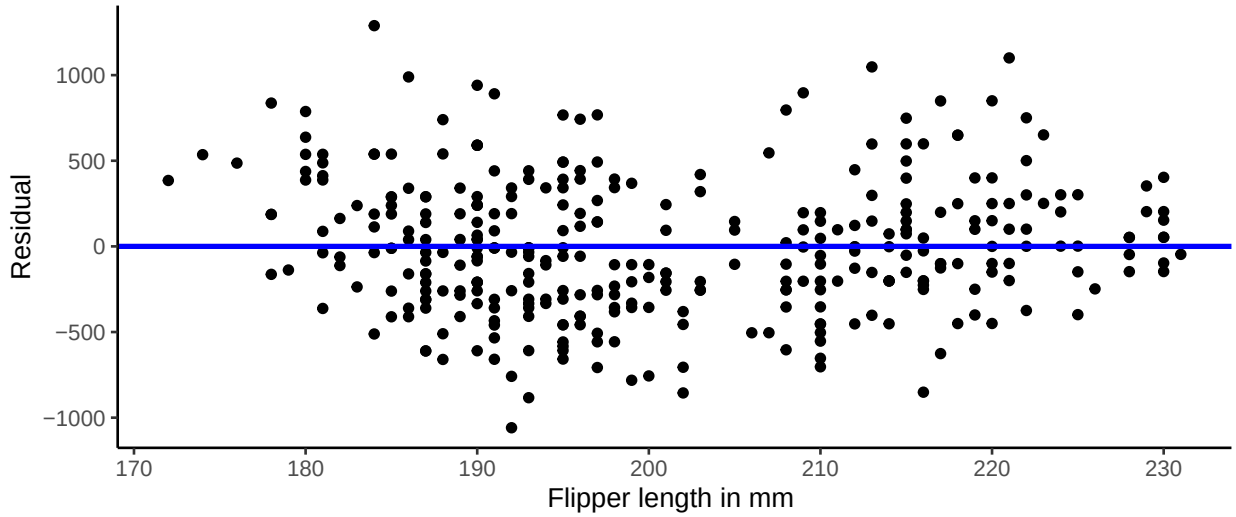


Answer:Yes, the distribution of residuals appears to be approximately normal and centered at 0.

Equality of variance of the residuals

The final condition says that the residual should exhibit equal variance across all of the values of the explanatory variable.

```
ggplot(regression_points, aes(x = flipper_length_mm, y = residual)) +
  geom_point() +
  labs(x = "Flipper length in mm ", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  theme_classic()
```



Exercise 6

Answer: For the most part, the residuals appear to have equal variance across all levels of flipper length.

It is also acceptable to state:

There appears to be some values of flipper length where were residual values are smaller.

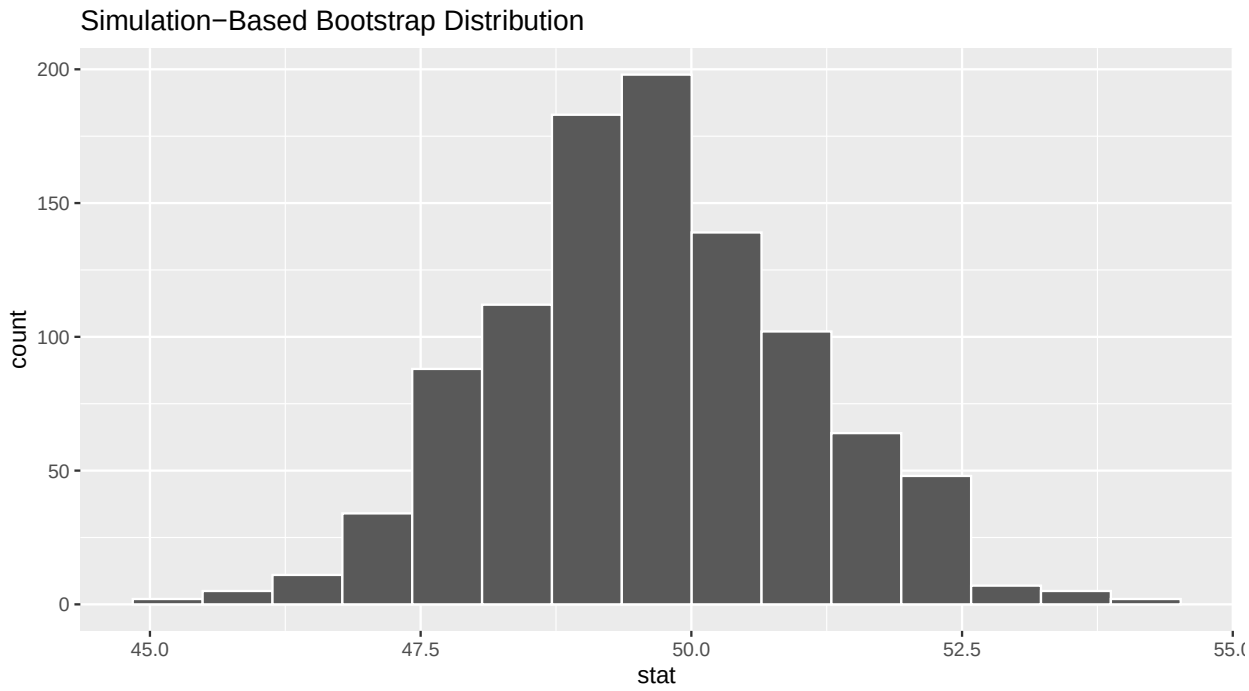
Exercise 7

Find a 95% confidence interval.

Now let's circle back and take a second look at the confidence intervals. Using this bootstrap distribution, we'll construct the 95% confidence interval using the percentile method and (if appropriate) the standard error method as well. We can compare our results to the results from R (which uses mathematical formula to construct confidence intervals.)

Step 1: Calculate the bootstrap statistic and Visualize the bootstrap distribution

```
set.seed(126)
bootstrap_distn_slope <- pen %>%
  specify(formula = body_mass_g ~ flipper_length_mm) %>%
  generate(reps = 1000, type = "bootstrap") %>%
  calculate(stat = "slope")
visualize(bootstrap_distn_slope)
```



Step 2: Calculate CI from the a bootstap resample

Find a 95% CI using percentile method

```
bootstrap_distn_slope %>%
  get_ci(type="percentile")
```

```
## # A tibble: 1 x 2
##   lower_ci upper_ci
##      <dbl>    <dbl>
## 1     47.0     52.3
```

Part 2

For the next part you will check the conditions for regression inference for a new model. This model will have `bill_length_mm` and `species` as explanatory variables, and we'll use the parallel slopes model. Let's fit the parallel slopes model

```
# Fit regression model:
pen_parallel <- lm(body_mass_g ~ bill_length_mm + species, data = pen)

# Get regression table:
get_regression_table(pen_parallel)
```

```
## # A tibble: 4 x 7
##   term              estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept            154.     269.       0.572   0.568   -375.      683.
```

```
## 2 bill_length_mm         91.4      6.89    13.3     0         77.9     105.
## 3 species: Chinstrap   -886.      88.2    -10.0     0      -1059.     -712.
## 4 species: Gentoo       579.      75.4      7.68    0        430.      727.
```

```
# Get regression points:
regression_points_par <- get_regression_points(pen_parallel)
```

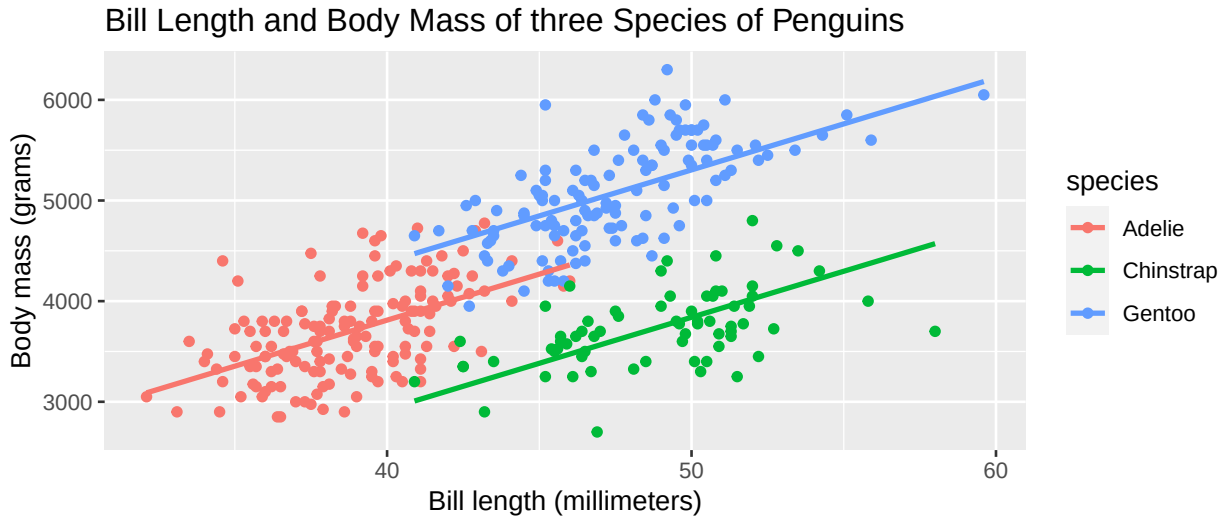
Let us once again inspect the conditions necessary for inference with regression.

1. Linearity of relationship between variables
2. Independence of the residuals
3. Normality of the residuals
4. Equality of variance of the residuals

Exercise 8

Check for Linearity of relationship between variables

```
ggplot(data = pen, aes(y = body_mass_g, x = bill_length_mm, color=species)) +
  geom_point() +
  geom_parallel_slopes(method = "lm", se = FALSE) +
  labs(x = "Bill length (millimeters)",
       y = "Body mass (grams)",
       title = "Bill Length and Body Mass of three Species of Penguins")
```



Answer: The linearity assumption appears to be reasonable.

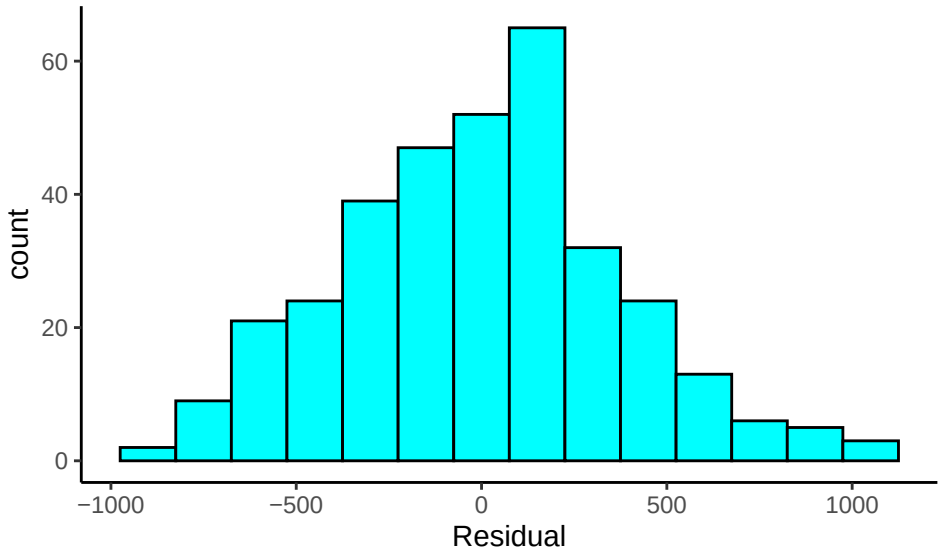
Check for Independence of the residuals

This is the same as the first model that we looked it. The observations in our data must be independent of one another. In this data, we can not be sure this is case, for example, some of the penguins included may be related (siblings, parents). We are not given enough information to verify this condition has been met.

Exercise 9

Check Normality of the residuals (and they should be centered at 0.)

```
#Add code for the histogram:
ggplot(regression_points_par, aes(x = residual)) +
  geom_histogram(binwidth = 150, color = "black", fill = "cyan") +
  labs(x = "Residual") +
  theme_classic()
```



Answer: The assumption of normality of residuals appear to be normal

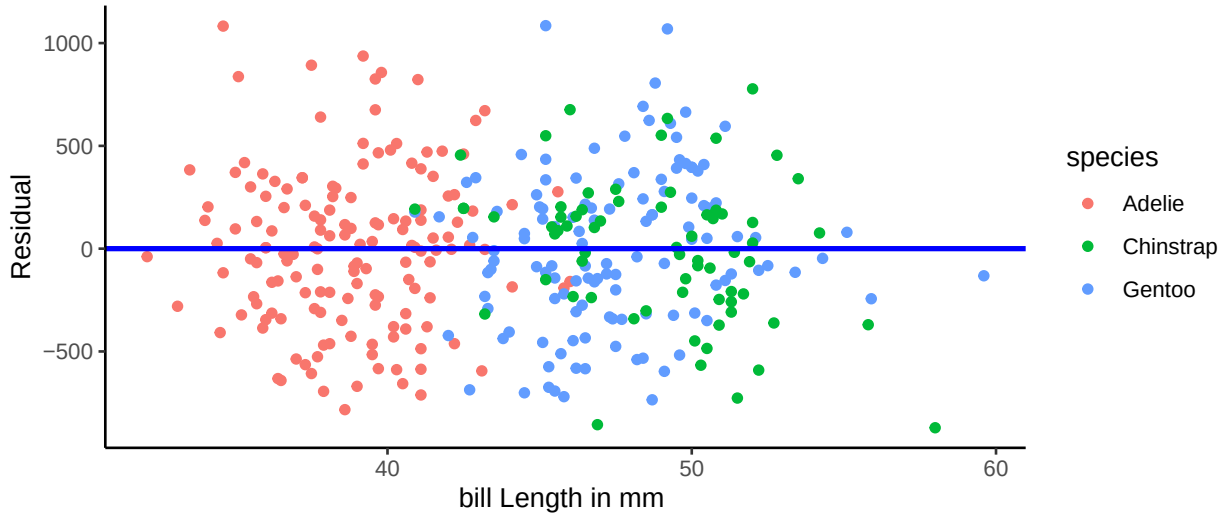
Exercise 10

Check for equality of variance of the residuals

To check this condition we can create a scatterplot that has our explanatory variable, `bill_length_mm`, on the x-axis and our residuals on the y-axis.

9a)

```
#Add code to check this condition:
ggplot(regression_points_par, aes(x = bill_length_mm, y = residual, color =species)) +
  geom_point() +
  labs(x = "bill Length in mm ", y = "Residual") +
  geom_hline(yintercept = 0, col = "blue", size = 1) +
  theme_classic()
```



Answer: Yes, the residuals appear to have equal variance across all x.

Documenting software

- File creation date: 2022-06-21
- R version 4.1.3 (2022-03-10)
- `tidyverse` package version: 1.3.1
- `infer` package version: 1.0.0
- `moderndive` package version: 0.5.4
- `palmerpenguins` package version: 0.1.0