STATS 250 Lab 13
Linear Regression Inference

Nick Seewald
nseewald@umich.edu

Week of 11/30/2020

Reminders 💡

Your tasks for the week running Friday 11/30 - Friday 12/4

Task	Due Date	Submission
Lab 13	Friday 12/4 8:00AM ET	Canvas
Homework 10	Monday 12/7 8:00AM ET	course.work

WE'RE IN THE HOME STRETCH! YOU CAN DO IT!

Homework 9 Comments

Question 1f:

The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey.

Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

Homework 9 Comments

Question 1f:

> Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

- p-value was ~0.387: not a lot of evidence against the null!
- Confidence intervals should support the conclusion of the hypothesis test
  - Null value is *reasonable*: it should be in the CI.
- You need to use statistical reasoning (e.g., p-values, hypothesis test results, etc.)

Homework 9 Comments

Question 2:

> The automotive engineer asks you, the statistics expert, to help him produce a 98% confidence interval for the population mean difference in decibel level when the car is being powered off (original part minus new part). You'll want to check any conditions, show your computations, and interpret the 98% confidence interval in context.

Checking the independence condition:

- <10% of the population is a **workaround** - we really need a random sample
- In a paired test, we don't have *between*-measurement independence, but we still need *within*-unit independence!

Homework 9 Comments

Question 7 💍:

> **Diamonds.** Conduct a hypothesis test to evaluate if there is a difference between the average standardized prices of 0.99 and 1 carat diamonds. Make sure to state your hypotheses clearly, show your computations, and interpret your results in context of the data.

- This is a **two-sided test**! "if there is a difference" does not imply a direction!
  - You need to **double** the output of `pt()` for a two-sided test.
- If you have questions, ask the instructional team, **not** Google.
  - See page 16 of your 4.1-4.3 notes for degrees of freedom.

Today's Goals

Statistics

- Learn about how to make inference for linear regression parameters
- Learn about conditions needed for valid inference in regression

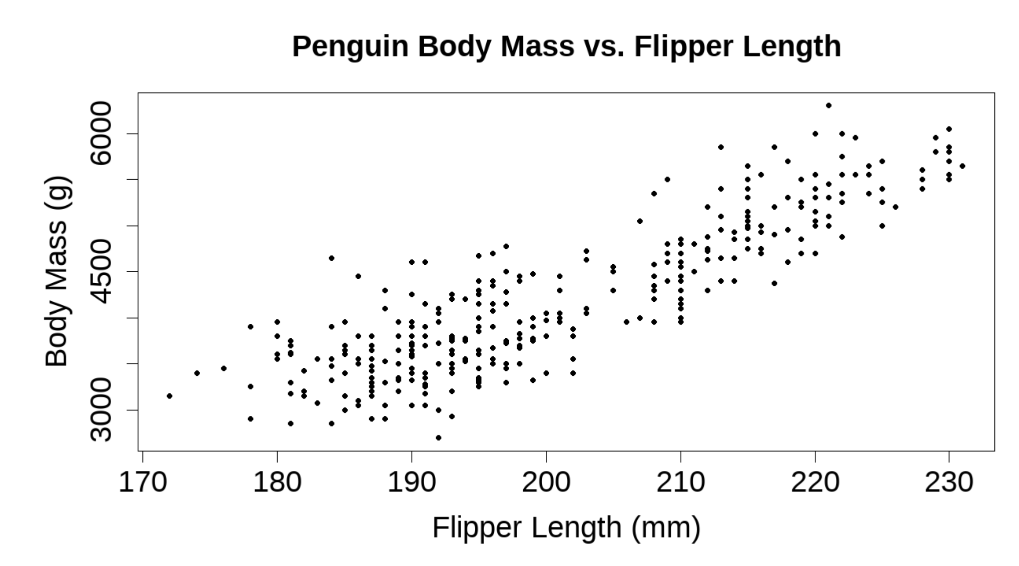
R

- Learn how to interpret output from `lm()` to make inference in regression
- Learn how to use R to check conditions for valid inference in regression

Scatterplots

```r
# Line ~44
penguins <- read.csv("https://raw.githubuserco
```

```r
plot(body_mass_g ~ flipper_length_mm,
     data = penguins,
     pch = 20,
     ylab = "Body Mass (g)",
     xlab = "Flipper Length (mm)",
     main = "Penguin Body Mass vs. Flipper Len
```



Formula notation!

`response variable ~ explanatory variable`

Recall: Linear Regression

```r
# Line 59
mod1 <- lm(body_mass_g ~ flipper_length_mm, data = penguins)
summary(mod1)
```

```

Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Regression Output (Lines 63-65)

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Equation of the regression line?

Regression Output (Lines 63-65)

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Equation of the regression line?

> $$\hat{y} = -5827.09 + 50.15x$$

Interpretation of b_1?

Regression Output (Lines 63-65)

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Equation of the regression line?

> $\hat{y} = -5827.09 + 50.15x$

Interpretation of b_1?

> We estimate that a one-millimeter longer flipper is associated with a **50.15**-gram **higher** body mass, on average, in the population of penguins represented by this sample.

Regression Diagnostics

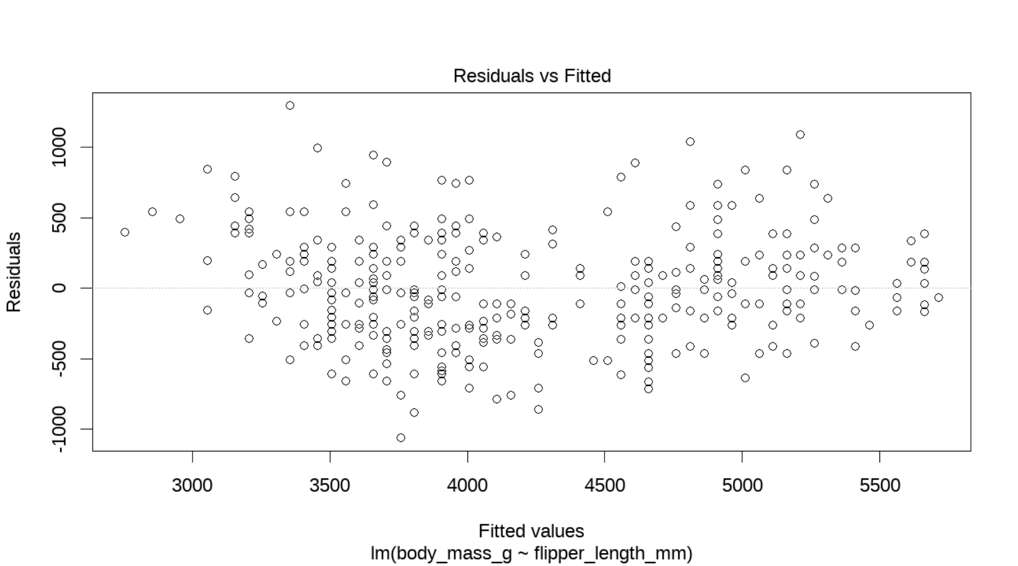
Checking Conditions

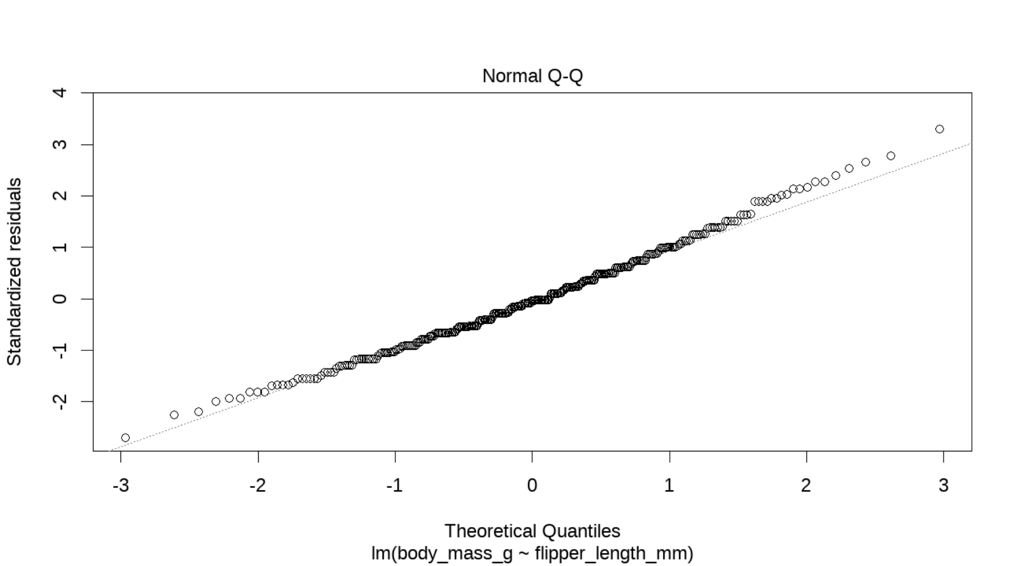
- **Linearity:** The relationship between the explanatory and response variables should be linear.
- **Independence:** The observations must be independent of one another. This does not mean that the response and explanatory variables are independent; rather, that the "individuals" from whom we collect information must be independent of each other.
- **Nearly Normal Residuals:** The residuals should come from a nearly-normal population of residuals.
- **Equal (constant) variability:** The variability of the residuals should not depend on where they are along the regression line.

Use the mnemonic "LINE"

Regression Diagnostics

```r
plot(mod1, which = c(1, 2), add.smooth = FALSE, id.n = 0)
```



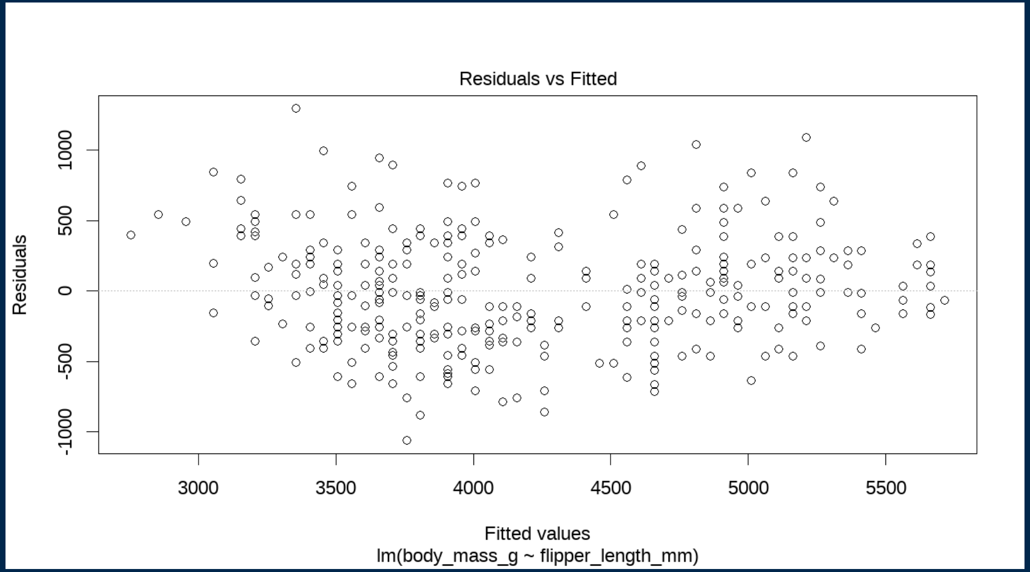


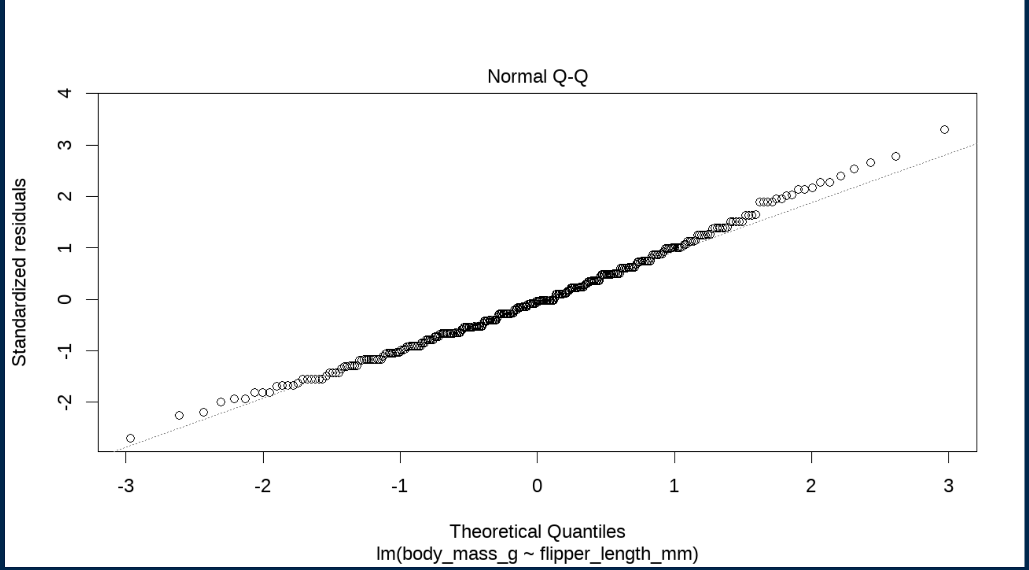
- Linearity
- Equal variance

- Nearly Normal

Regression Diagnostics

```
plot(mod1, which = c(1, 2), add.smooth = FALSE, id.n = 0)
```





Take a minute to describe your thoughts about the conditions on line 90 of your lab document.

Regression Inference

Conditions seem okay! Let's make some inference.

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Q: At the population level, is there a relationship between penguin flipper length and body mass?

If not, then the slope of the "true" line should be zero.

$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

Regression Inference

Conditions seem okay! Let's make some inference.

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

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

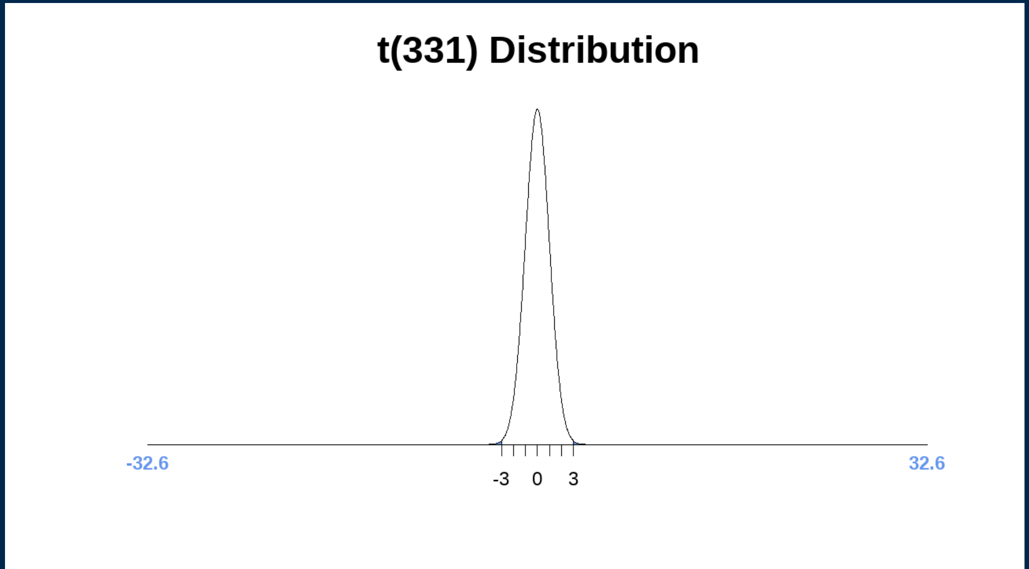
$$t = \frac{b_1 - b_{1,\text{null}}}{SE_{b_1}} = \frac{50.15 - 0}{1.54} = 32.56$$

Regression Inference

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Compute the p-value on line 116.

```
plotT(333 - 2, shadeValues = c(-32.56, 32.56),
      direction = "outside",
      xlim = c(-34, 34))
```

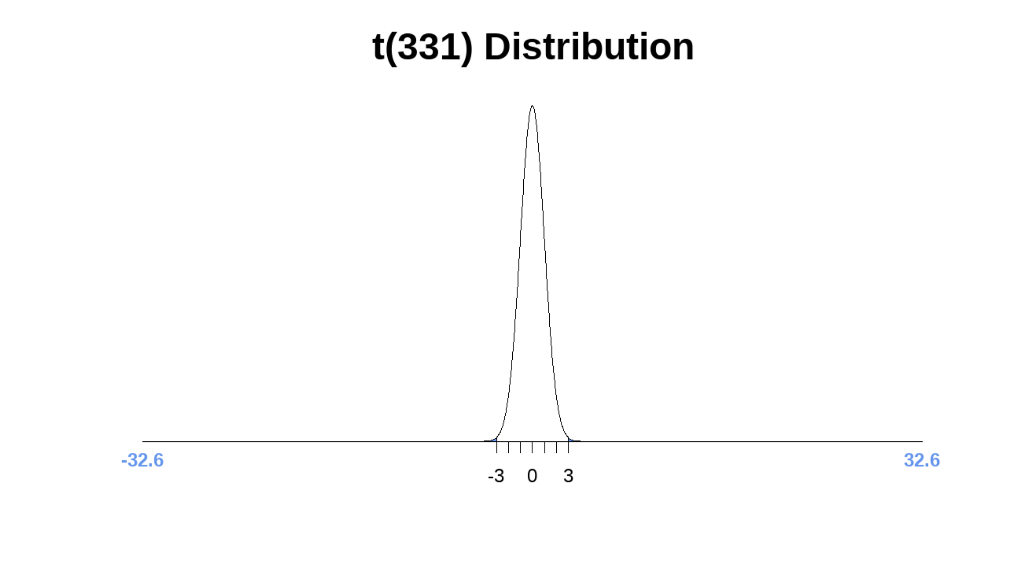


Regression Inference

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

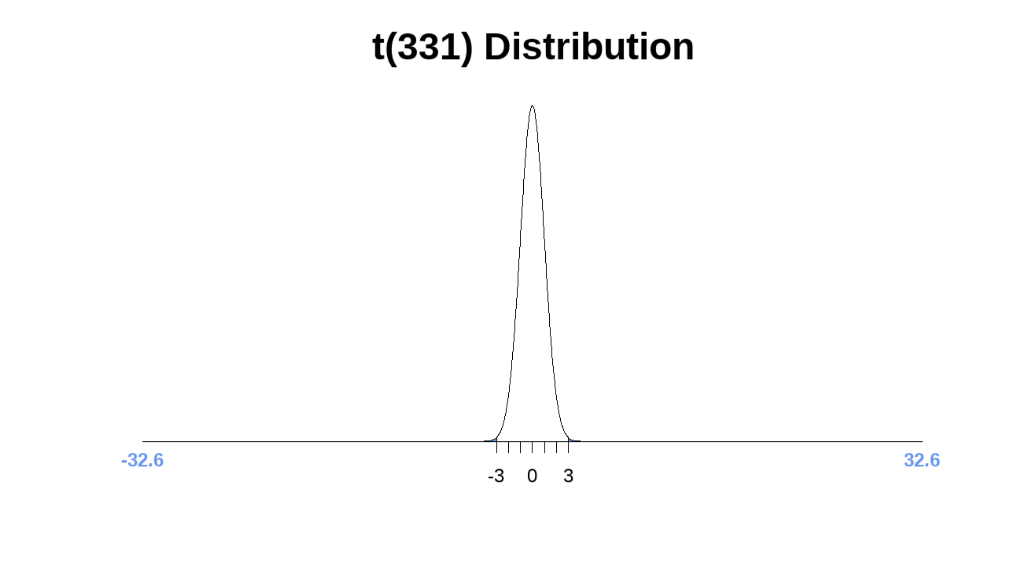
```
2 * pt(-32.6, df = 331)
```

```
[1] 2.336624e-105
```



Regression Inference

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```



```r
2 * pt(-32.6, df = 331)
```

```
[1] 2.336624e-105
```

This is **nonsense precision**. Do not, under any circumstances, report this p-value as-is. It is zero.

Regression Inference

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max 
-1057.33  -259.79   -12.24   242.97  1293.89 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -5872.09     310.29  -18.93   <2e-16 ***
flipper_length_mm    50.15       1.54   32.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.3 on 331 degrees of freedom
Multiple R-squared:  0.7621,	Adjusted R-squared:  0.7614 
F-statistic:  1060 on 1 and 331 DF,  p-value: < 2.2e-16
```

Important Note

Notice that the p-value column in the output is labeled `Pr(>|t|)`?

- This is a **two-sided p-value** for the test that the coefficient is equal to zero.

Confidence Intervals for Regression Parameters

Let's make a 95% confidence interval for each of β_0 and β_1.

$$b_1 \pm t^* \times \mathrm{SE}_{b_1}$$

```r
confint(mod1, level = .95)
```

```
                        2.5 %      97.5 %
(Intercept)       -6482.47224 -5261.71313
flipper_length_mm    47.12339    53.18314
```

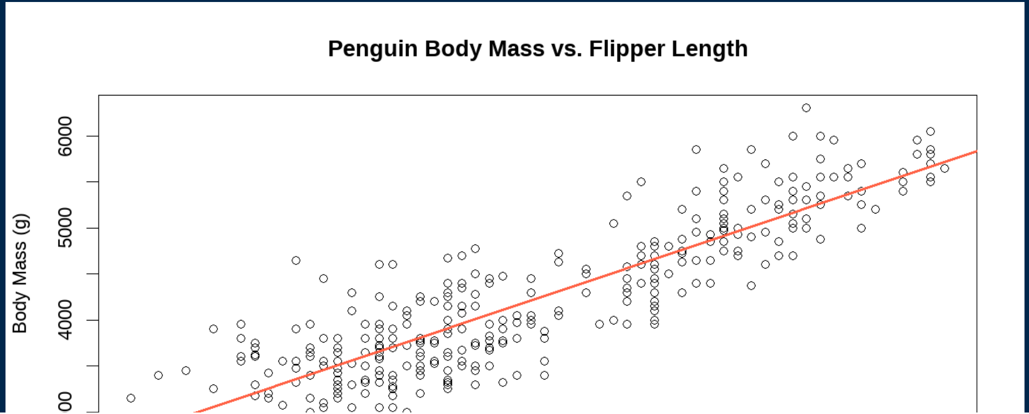
Exercise: Verify these CI's are correct using the regression output.

Plotting a Regression Line

Our estimated regression line is

$$y_i = -5827.09 + 50.15x$$

```
plot(body_mass_g ~ flipper_length_mm,
     data = penguins,
     ylab = "Body Mass (g)",
     xlab = "Flipper Length (mm)",
     main = "Penguin Body Mass vs. Flipper Leng
abline(mod1, col = "tomato", lwd = 2)
```



Code Cheat Sheet 💻

`lm(formula, data)`

- `formula` is a symbolic description of the model you want to fit: recall the syntax is `response ~ explanatory`.
- `data` is a data frame which contains the variables used in `formula`.

Code Cheat Sheet 💻

`pt(q, df, lower.tail = TRUE)`

- q is the x-axis value you want to find an area related to
- `df` is the degrees of freedom of the t distribution
- `lower.tail` determines whether `pt()` finds the area to the left or right of q. If `lower.tail = TRUE` (the default), it shades to the left. If `lower.tail = FALSE`, it shades to the right.

Code Cheat Sheet 💻

`qt(q, df, lower.tail = TRUE)`

- p is the probability or area under the curve you want to find an x-axis value for
- `df` is the degrees of freedom of the t distribution
- `lower.tail` determines whether `pt()` finds the area to the left or right of q. If `lower.tail = TRUE` (the default), it shades to the left. If `lower.tail = FALSE`, it shades to the right.

Code Cheat Sheet 💻

`plotT()`

- `df` refers to the degrees of freedom of the distribution to plot. You must provide this value.
- `shadeValues` is a vector of up to 2 numbers that define the region you want to shade
- `direction` can be one of `less`, `greater`, `outside`, or `inside`, and controls the direction of shading between `shadeValues`. Must be `less` or `greater` if `shadeValues` has only one element; `outside` or `inside` if two
- `col.shade` controls the color of the shaded region, defaults to `"cornflowerblue"`
- `...` lets you specify other graphical parameters to control the appearance of the normal curve (e.g., `lwd`, `lty`, `col`, etc.)

Code Cheat Sheet 💻

`plot(model, which, add.smooth, id.n)` for `lm()` output

- `model` is the regression model (an `lm` object)
- `which` controls which diagnostic plots you want to see. We're typically interested in just the first 2, so we'll set this to `c(1, 2)`.
- `add.smooth` controls whether or not to add a "smoother" to the residual plot. *SET THIS TO FALSE*.
- `id.n` controls the number of the most unusual points to identify in the plots. This is generally not helpful and confusing: *SET THIS TO `0`*.

Code Cheat Sheet 💻

`confint(object, level)`

- `object` is a fitted regression model (an `lm` object)
- `level` is the required confidence level, must be between 0 and 1.

Lab Project ⌨

Your tasks

- Complete the "Try It!" and "Dive Deeper" portions of the lab assignment by copy/pasting and modifying appropriate code from earlier in the document.

How to get help

- Piazza!
- Email your lab instructor (not stats250-miller@umich.edu)

Recap and Reminders 💡

What we learned:

- Checking conditions for valid linear regression inference
- Using R to quickly perform hypothesis tests for regression parameters
- Using R to create confidence intervals for regression parameters

Task	Due Date	Submission
Lab 13	Friday 12/4 8:00AM ET	Canvas
Homework 10	Monday 12/7 8:00AM ET	course.work