Lab 6: Simple Linear Regression, Confidence Intervals versus Prediction Intervals

STAT 242: Intermediate Statistics

Goals

The goal in this lab is to practice finding confidence intervals for the coefficients, and prediction intervals for new observations. We will compare the widths for prediction versus confidence intervals. We will continue with the leaf margin data from before.

Loading packages

Here are some packages with functionality you may need for this lab. Run this code chunk now.

```
library(readr)
```

```
## Warning: package 'readr' was built under R version 4.0.5
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(gridExtra)
library(mosaic)
```

```
## Warning: package 'mosaic' was built under R version 4.0.3
```

```
library(dplyr)
options("pillar.sigfig" = 10) # print 10 significant digits in summarize output
```

Leaf Margins

For a variety of reasons, scientists are interested in the relationship between the climate of a region and characteristics of the plants and animals that live there. For example, this could inform thinking about the impacts of climate change on natural resources, and could be used by paleontologists to learn about historical climatological conditions from the fossil record.

In 1979, the US Geological service published a report discussing a variety of characteristics of forests throughout the world and discussed connections to the climates in those different regions (J. A. Wolfe, 1979, Temperature parameters of humid to mesic forests of eastern Asia and relation to forests of other regions of the Northern Hemisphere and Australasia, USGS Professional Paper, 1106). One part of this report discussed the connection between the temperature of a region and the shapes of tree leaves in the forests in that

region. Generally, leaves can be described as either "serrated" (having a rough edge like a saw blade) or "entire" (having a smooth edge) - see the picture here: https://en.wikibooks.org/wiki/Historical_Geology/Leaf_shape_and_temperature. One plot in the report displays the relationship between the mean annual temperature in a forested region (in degrees Celsius) and the percent of leaves in the forest canopy that are "entire".

The following R code reads in the data:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine()           masks gridExtra::combine()
## x mosaic::count()            masks dplyr::count()
## x purrr::cross()             masks mosaic::cross()
## x mosaic::do()               masks dplyr::do()
## x tidyr::expand()            masks Matrix::expand()
## x dplyr::filter()            masks stats::filter()
## x ggstance::geom_errorbarh() masks ggplot2::geom_errorbarh()
## x dplyr::lag()               masks stats::lag()
## x tidyr::pack()              masks Matrix::pack()
## x mosaic::stat()             masks ggplot2::stat()
## x mosaic::tally()            masks dplyr::tally()
## x tidyr::unpack()            masks Matrix::unpack()
```

```
leaf <- read_csv("http://www.evanlray.com/data/misc/leaf_margins/leaf_margins.csv")
```

```
## Rows: 34 Columns: 2
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (2): pct_entire_margined, mean_annual_temp_C
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
head(leaf)
```

```
## # A tibble: 6 x 2
##   pct_entire_margined mean_annual_temp_C
##                 <dbl>              <dbl>
## 1         86.35674576        26.75519498
## 2         82.42964550        26.90082024
## 3         81.38752686        26.43200957
## 4         82.28502110        25.77290558
## 5         77.36406594        25.84919343
## 6         76.22703233        25.26298548
```

1. Fit a linear model to this data set and print out a summary of the model fit. (We did this in the last lab, so you can copy and paste from there).

```r
linear_fit <- lm(pct_entire_margined ~ mean_annual_temp_C, data=leaf)
```

2. Find a 95% confidence interval for the mean percent of leaves that are entire margined in forests where the mean annual temperature is 17 degrees C. Do this using the predict function.

```r
predict_df <- data.frame(
  mean_annual_temp_C = c(17)
)

conf_17 <- predict(linear_fit, newdata = predict_df, interval = "confidence")
conf_17
```

```
##        fit     lwr      upr
## 1 51.90473 51.0787 52.73077
```

3. Interpret your interval in the context of the problem; include a statement of what it means to be 95% confident.

We are 95% confident that the mean percent of leaves that are entire margined in forests where the mean annual temperature is 17 degrees Celsius is between 51.079 and 52.731 percent. For 95% of samples like the one in this study, the corresponding confidence intervals will contain the true mean percent of leaves that are entire margined in forests where the mean annual temperature is 17 degrees C.

4. Find a 95% prediction interval for the percent of leaves that are entire margined in a forest that was not in our data set before, and that has a mean annual temperature of 17 degrees C. Do this using the predict function.

```r
pred_17 <- predict(linear_fit, newdata = predict_df, interval = "prediction")
pred_17
```

```
##        fit      lwr     upr
## 1 51.90473 47.02427 56.7852
```

5. Interpret your interval in the context of the problem. Be sure to interpret a prediction interval.

We are 95% confident that the percent of leaves that are entire margined in a forest with a mean annual temperature of 17 degrees C is between 51.905, 47.024, 56.785 and 51.905, 47.024, 56.785 percent. For 95% of

samples and 95% of forests with a mean annual temperature of 17 degrees Celsius like the one in this study, the corresponding prediction interval will contain the true percentage of leaves that are entire margined.

6. What is the difference between your intervals in (2) and (4)? Explain briefly.

(2) is a confidence interval - it only accounts for variation around the mean at x=17. (4) is a prediction interval - it accounts for both the variation around the mean and the variability in entire margined leaves among individual forests with a mean annual temperature of 17 degrees C.