Lab 6 - linear models

Solutions

Goals

The goal in this lab is to practice interpreting the coefficient estimates in simple linear regression models (linear models with one quantitative explanatory variable), conducting hypothesis tests, and finding confidence intervals for the coefficients.

Loading packages

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

```
library(readr)
library(ggplot2)
library(gridExtra)
library(mosaic)
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 displaysthe 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)
```

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

```
## v tibble  2.1.3     v purrr   0.3.2
## v tidyr   0.8.3     v stringr 1.4.0
## v tibble  2.1.3     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_conf
## 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 mosaic::stat()             masks ggplot2::stat()
## x mosaic::tally()            masks dplyr::tally()
```

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

```
## Parsed with column specification:
## cols(
##   pct_entire_margined = col_double(),
##   mean_annual_temp_C = col_double()
## )
```

```
head(leaf)
```

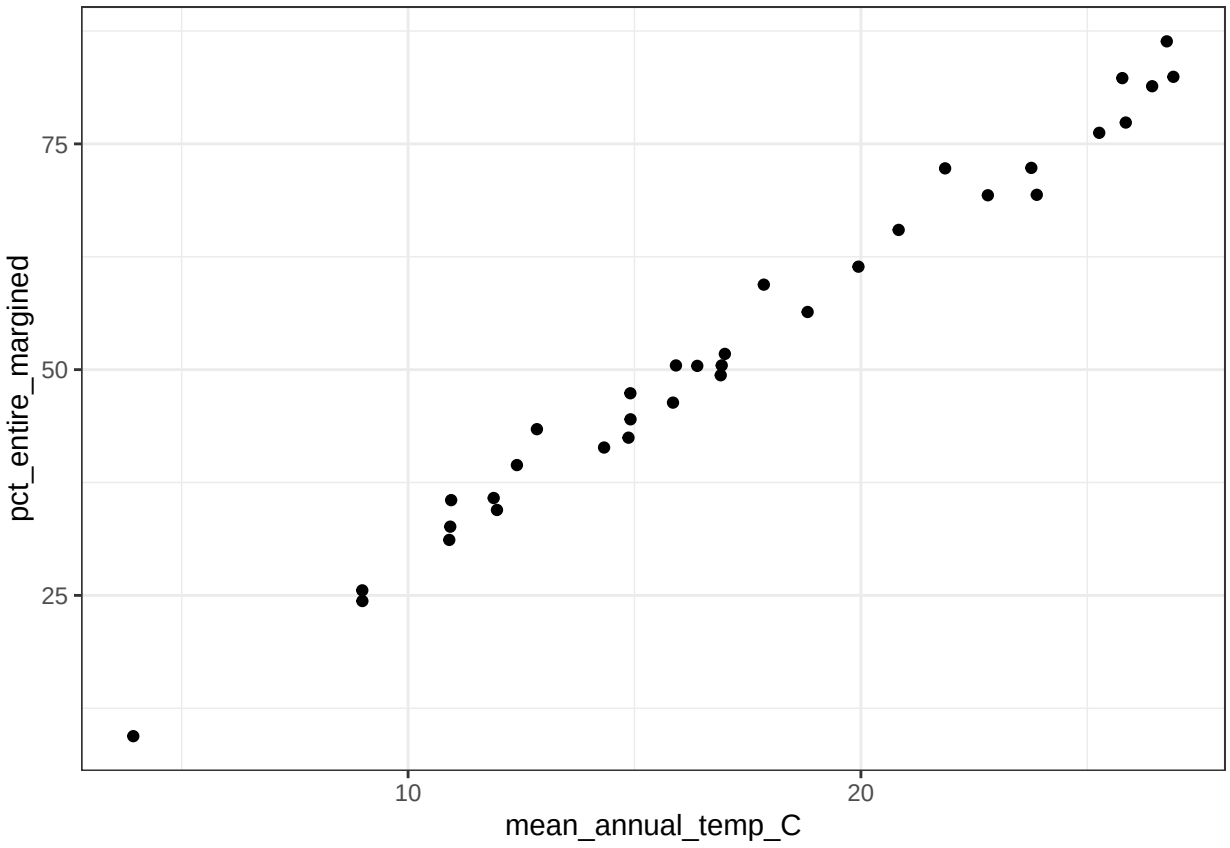
```
## # A tibble: 6 x 2
##   pct_entire_margined mean_annual_temp_C
##                 <dbl>              <dbl>
## 1         86.35674576        26.75519498
## 2         82.42964550        26.90082024
## 3         81.38752686        26.43200957
## 4         82.28502110        25.77290558
## 5       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

1. Which variable in the data set is the explanatory variable and which is the response?

The explanatory variable is mean annual temperature (in degrees Celsius) and the response variable is percent of leaves in the forest canopy that are "entire" (percent entire margined).

2. Make a scatter plot of the data with the explanatory variable on the horizontal axis and the response on the vertical axis.

```
ggplot(data=leaf, aes(x=mean_annual_temp_C, y=pct_entire_margined)) +
  geom_point() +
  theme_bw()
```



3. Fit a linear model to this data set and print out a summary of the model fit.

```
lm_leaf <- lm(pct_entire_margined ~ mean_annual_temp_C, data=leaf)
summary(lm_leaf)
```

```
##
## Call:
## lm(formula = pct_entire_margined ~ mean_annual_temp_C, data = leaf)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4387 -1.4147 -0.8165  1.8490  4.9296
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -2.16513    1.24613  -1.737   0.0919 .
## mean_annual_temp_C  3.18058    0.06808  46.718   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.361 on 32 degrees of freedom
## Multiple R-squared:  0.9856, Adjusted R-squared:  0.9851
## F-statistic:  2183 on 1 and 32 DF,  p-value: < 2.2e-16
```

4. Interpret the estimated intercept and slope in context.

Intercept: For a mean annual temperature of 0 degrees Celsius, we expect -2.165 percent of the leaves in the canopy to be "entire".

Slope: For a 1 degree increase in mean annual temperature (in degrees Celsius), we expect an increase of 3.18 percent in percent of the leaves in the canopy that are "entire".

5. Conduct a hypothesis test of the claim that the average temperature in a given location has no effect on the percent of leaves in forests there that are entire margined. State your hypotheses in symbols and written sentences and interpret the p-value in terms of strength of evidence against the null hypothesis. Do you know how you could find the p-value for this test given the estimate and a standard error of the estimate?

Hypotheses:

$H_0 : \beta_1 = 0$. The average temperature has no effect on the percent of leaves in forests there that are entire margined.

$H_A : \beta_1 \neq 0$. The average temperature has an effect on the percent of leaves in forests there that are entire margined.

Interpetation:

There is very strong evidence (p-value < 2e-16) that there is a positive relationship (t=3.18) between the average temperature and the percent of the leaves in the canopy that are entire. If the average temperature has no effect on the percent of leaves in the forests there that are entire margined (under the null), then in similar studies (similar study population and sampling method) we would expect to get a result like the one we got here purely by chance in about 0% of these samples (since the p-value is so small in this case).

Find p-value given the estimate and a standard error of the estimate:

```
## Find test statistic
t_stat <- (3.18058-0)/0.06808

## Find P(|T| > t_stat) - this is the p-value
2*pt(t_stat, df=32, lower.tail = FALSE)
```

```
## [1] 5.088978e-31
```

6. Find a 95% confidence interval for the amount by which the average percent of leaves that are entire margined increases for each 1-degree increase in the average temperature. Do you know how you could find the p-value for this test given the estimate and a standard error of the estimate?

```
confint(lm_leaf)
```

```
##                        2.5 %    97.5 %
## (Intercept)        -4.703410 0.3731551
## mean_annual_temp_C  3.041905 3.3192557
```

This does not specifically ask for an interpretation, but let's write one anyway! We are 95% confident that the mean amount by which the average percent of leaves that are entire margined increases for each 1-degree increase in the average temperature is between 3.042 and 3.319 percent. For 95% of samples from a similar population (collected in a similar way), corresponding intervals constructed in this way would contain the

true mean amount by which the average percent of leaves that are entire margined increases for each 1-degree increase in the average temperature.

Find p-value given the estimate and a standard error of the estimate:

You would do this in the same way as you did for 5. $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$

7. Conduct a hypothesis test of the claim that on average, in forests where the average temperature is 0 degrees C, 0 percent of leaves are entire margined.

$H_0 : \beta_0 = 0$. In forests where the average temperature is 0 degrees C, 0 percent of leaves are entire margined.

$H_A : \beta_0 \neq 0$. In forests where the average temperature is 0 degrees C, the percent of leaves that are entire margined is something other than 0.

Interpretation: (You can continue using the summary output from 3.)

There is inconclusive evidence (p-value = 0.09) that in forests where the average temperature is 0 degrees, the percent of leaves that are entire margined is something other than 0. If in forests where the average temperature is 0 degrees, the percent of leaves that are entire margined is 0, then in similar studies (similar study population and sampling method) we would expect to get a result like the one we got here purely by chance in about 9% of these samples.

8. Find an estimate and a 95% confidence interval for the mean percent of leaves that are entire margined in forests where the mean annual temperature is 20 degrees C.

```
predict_df <- data.frame(
  mean_annual_temp_C <- 20
)

predict(lm_leaf, predict_df, interval="confidence", se.fit = TRUE)
```

```
## $fit
##        fit      lwr      upr
## 1 61.44647 60.54114 62.35181
##
## $se.fit
## [1] 0.4444587
##
## $df
## [1] 32
##
## $residual.scale
## [1] 2.361419
```

Once again, this doesn't ask for us to interpret the 95% confidence interval, but let's do it anyway.

We are 95% confident that the mean percent of leaves that are entire margined in forests where the mean annual temperature is 20 degrees C is between 60.541 and 62.352 percent. For 95% of samples we would expect the corresponding intervals for forests where the mean annual of temperature is 20 degrees C to contain the true mean percent of leaves that are entire margined.

9. Find a set of 3 Bonferroni-adjusted confidence intervals for the mean percent of leaves that are entire margined in forests where the mean annual temperature is 15 degrees C, 20 degrees C, and 25 degrees C. Use a family-wise confidence level of 95%.

```
predict_df <- data.frame(
  mean_annual_temp_C <- c(15,20,25)
)

predict(lm_leaf, predict_df, interval="confidence", se.fit = TRUE, level=1-0.05/3)
```

```
## $fit
##        fit      lwr      upr
## 1 45.54357 44.44597 46.64118
## 2 61.44647 60.32358 62.56937
## 3 77.34938 75.67718 79.02158
##
## $se.fit
##         1         2         3
## 0.4344482 0.4444587 0.6618835
##
## $df
## [1] 32
##
## $residual.scale
## [1] 2.361419
```

We are 95% confident that the mean percent of leaves that are entire margined in forests where the mean annual temperature is 15 degrees C is between 44.446 and 46.641 percent, for forests where the mean annual temperature is 20 degrees C is between 60.323 and 62.569 percent, and for forests where the mean annual temperature is 25 degrees C is between 75.677 and 79.022 percent. For 95% of samples like the one in this study, the corresponding intervals at each of these three mean temperatures constructed in this way will simultaneously contain their respective means.

10. Create a scatterplot of the data with the estimated line overlaid on top, and lines showing the bounds of Scheffe-based confidence intervals for the means at each value of X in the range of the data.

```
intervals <- predict(lm_leaf, interval="prediction") %>%
  as.data.frame()
```

```
## Warning in predict.lm(lm_leaf, interval = "prediction"): predictions on current data refer to _futur
```

```
head(intervals)
```

```
##        fit      lwr      upr
## 1 82.93191 77.87893 87.98490
## 2 83.39509 78.33683 88.45334
## 3 81.90400 76.86244 86.94555
## 4 79.80766 74.78826 84.82706
## 5 80.05030 75.02842 85.07219
## 6 78.18582 73.18248 83.18916
```

```r
leaf <- leaf %>%
  bind_cols(
    intervals
  )
head(intervals)
```

```
##        fit      lwr      upr
## 1 82.93191 77.87893 87.98490
## 2 83.39509 78.33683 88.45334
## 3 81.90400 76.86244 86.94555
## 4 79.80766 74.78826 84.82706
## 5 80.05030 75.02842 85.07219
## 6 78.18582 73.18248 83.18916
```

```r
ggplot(data=leaf, aes(x=mean_annual_temp_C, y=pct_entire_margined)) +
  geom_point() +
  geom_smooth(method="lm", se = FALSE) +
  geom_line(aes(y=lwr), linetype=2) +
  geom_line(aes(y=upr), linetype=2) +
  theme_bw()
```