Lab07 - linear models

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)
```

```
## Warning: package 'mosaic' was built under R version 3.5.2
## Loading required package: dplyr
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:gridExtra':
##
##     combine
## The following objects are masked from 'package:stats':
##
##     filter, lag
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
## Loading required package: lattice
## Loading required package: ggformula
## Warning: package 'ggformula' was built under R version 3.5.2
## Loading required package: ggstance
##
## Attaching package: 'ggstance'
## The following objects are masked from 'package:ggplot2':
##
##     geom_errorbarh, GeomErrorbarh
##
## New to ggformula?  Try the tutorials:
##  learnr::run_tutorial("introduction", package = "ggformula")
##  learnr::run_tutorial("refining", package = "ggformula")
## Loading required package: mosaicData
## Loading required package: Matrix
```

```
##
## The 'mosaic' package masks several functions from core packages in order to add
## additional features.  The original behavior of these functions should not be affected by this.
##
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':
##
##     mean

## The following objects are masked from 'package:dplyr':
##
##     count, do, tally

## The following object is masked from 'package:ggplot2':
##
##     stat

## The following objects are masked from 'package:stats':
##
##     binom.test, cor, cor.test, cov, fivenum, IQR, median,
##     prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':
##
##     max, mean, min, prod, range, sample, sum
```

```
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 ---------------------------------------------------------------------------- t
## v tibble  2.0.1     v purrr   0.3.0
```

```
## v tidyr   0.8.2     v stringr 1.3.1
## v tibble  2.0.1     v forcats 0.3.0
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## -- Conflicts ------------------------------------------------------------------------------- tidyverse
## 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 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?

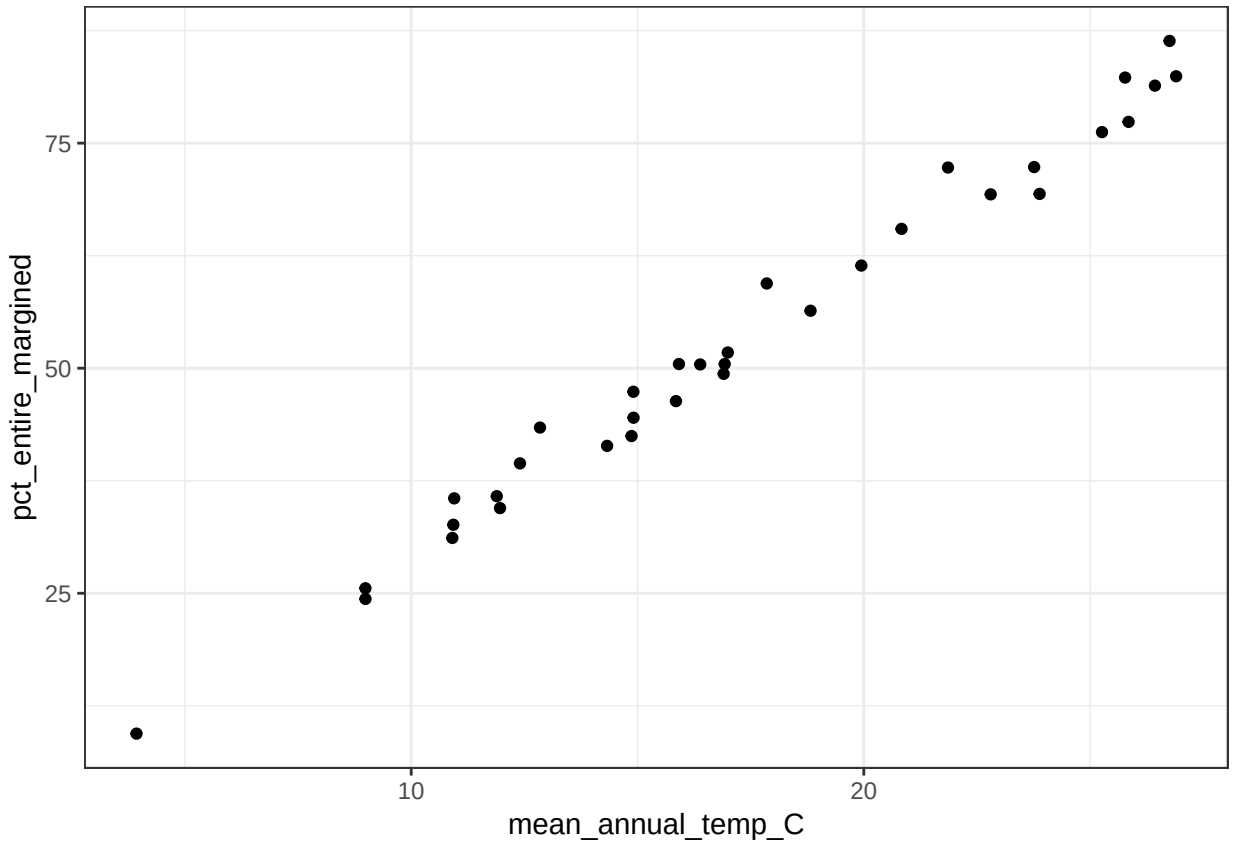
Explanatory: `mean_annual_temp_C`

Response: `pct_entire_margined`

We believe that the mean annual temperature in a given location may affect the percent of leaves in that location that are 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, mapping = 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.

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

```
##
## 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.

The model estimates that across all forests where the mean annual temperature is 0 degrees Celsius, the average percent of leaves that are entire margined is -2.2%. Note that this estimate is not reliable since 0 degrees is quite a bit lower than the smallest mean temperature in our data set. We cannot expect the linear relationship observed in our data to extend so far beyond the range of temperatures represented in the data set.

It is estimated that the mean percent of leaves that are entire margined increases by about 3.2 percentage points for each increase in the mean annual temperature in the forest of 1 degree Celsius.

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?

$H_0 : \beta_1 = 0$. A change in the mean temperature is not associated with a change in the mean percent of leaves that are entire margined.

$H_A : \beta_1 \neq 0$. A change in the mean temperature is associated with a change in the mean percent of leaves that are entire margined.

From the summary output above, the p-value for this test is less than 2.2×10^{-16}. The data provide extremely strong evidence against the null hypothesis of no association between mean annual temperature and percent of leaves that are entire margined.

Here is a calculation of the p-value by hand:

```
# find the test statistic
(3.181 - 0)/0.068
```

```
## [1] 46.77941
```

```
# look up the sample size, which is needed to find the degrees of freedom
nrow(leaf)
```

```
## [1] 34
```

```
pt(-46.779, df = 34 - 2) + pt(46.779, df = 34 - 2, lower.tail = FALSE)
```

```
## [1] 4.884615e-31
```

Our calculated p-value is about 4.885×10^{-31}. At first this might look different from the p-value R reported above – but note that the R output just said that the p-value was less than 2.2×10^{-16}. Basically, the p-value calculations are not precise enough to be reliable to 17 decimal places, so R just reports that the p-value is something less than 2.2×10^{-16}. This is consistent with our results.

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?

```
# automatic calculations
confint(model_fit)
```

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

```r
# calculation by hand
qt(0.975, df = 34 - 2)
```

```
## [1] 2.036933
```

```r
3.181 - 2.037 * 0.068
```

```
## [1] 3.042484
```

```r
3.181 + 2.037 * 0.068
```

```
## [1] 3.319516
```

I didn't ask for an interpretation of the interval as part of this lab, but for your reference, here is an interpretation: We are 95% confident that in the population of all forests in the world (or whatever population was sampled to obtain the forests used in this study), an increase of 1 degree in the mean annual temperature is associated with an increase of between 3.04 and 3.32 percentage points in the percentage of leaves that are entire margined.

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 that 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 on average.

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

From the linear model summary output in part 3, the p-value for this test is 0.0919. The data do not provide strong evidence against the null hypothesis that in forests where the average temperature is 0 degrees C, 0 percent of leaves are entire margined on average.

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.

```r
predict_data <- data.frame(
  mean_annual_temp_C = 20
)

predict(model_fit, newdata = predict_data, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 61.44647 60.54114 62.35181
```

We estimate that among forests where the mean annual temperature is 20 degrees C, the mean percent of leaves that are entire margined is about 61.45%. We are 95% confident that in forests where the mean annual temperature is 20 degrees C, the mean percent of leaves that are entire margined is between about 60.54% and 62.35%. For about 95% of samples, an interval calculated in this way would contain the mean percent of leaves that are entire margined in forests where the mean annual temperature is 20 degrees, if all model conditions were satisfied.

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.

```r
1 - 0.05/3
```

```
## [1] 0.9833333
```

The above calculation shows that to have a familywise confidence level of 95%, if we use the Bonferroni adjustment then each interval will have an individual confidence level of 98.3%.

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

predict(model_fit, newdata = predict_data, interval = "confidence", level = 0.983)
```

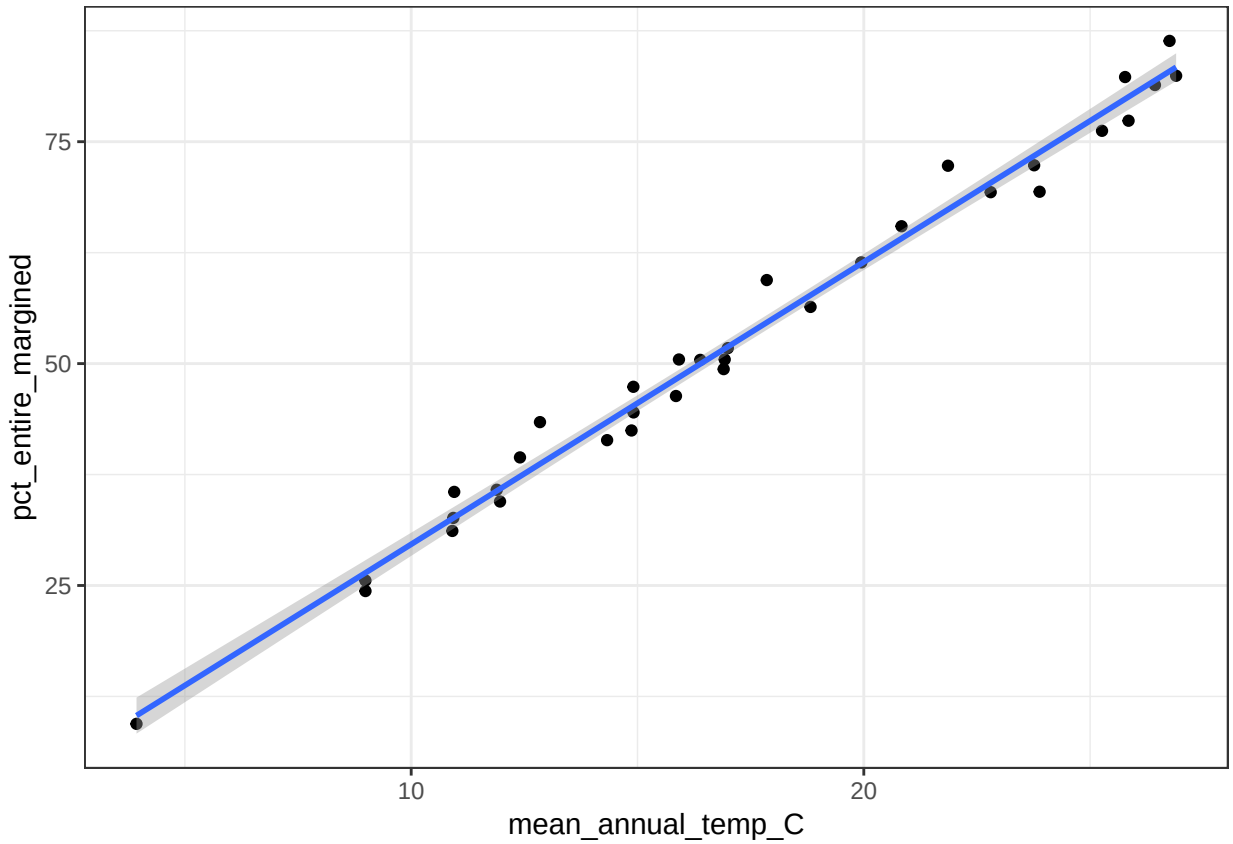
```
##        fit      lwr      upr
## 1 45.54357 44.44962 46.63753
## 2 61.44647 60.32731 62.56564
## 3 77.34938 75.68273 79.01602
```

Again, I didn't ask for an interpretation, but for your reference here is one:

We are 95% confident that in forests where the mean annual temperature is 15 degrees C the mean percent of leaves that are entire margined is between about 44.45% and 46.64%, in forests where the mean annual temperature is 20 degrees C the mean percent of leaves that are entire margined is between about 60.33% and 62.57%, and in forests where the mean annual temperature is 25 degrees C the mean percent of leaves that are entire margined is between about 75.68% and 79.02%. For about 95% of samples, all three of the confidence intervals calculated in this way would simultaneously contain the means they are estimating.

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

```
ggplot(data = leaf, mapping = aes(x = mean_annual_temp_C, y = pct_entire_margined)) +
  geom_point() +
  geom_smooth(method = "lm") +
  theme_bw()
```


75
50
25
pct_entire_margined
10
20
mean_annual_temp_C