HW1

Solutions

Details

Grading

20% of your grade on this assignment is for completion. A quick pass will be made to ensure that you've made a reasonable attempt at all problems.

Some of the problems will be graded more carefully for correctness. In grading these problems, an emphasis will be placed on full explanations of your thought process. You usually won't need to write more than a few sentences for any given problem, but you should write complete sentences! Understanding and explaining the reasons behind your decisions is more important than making the "correct" decision.

Solutions to all problems will be provided.

Collaboration

You are allowed to work with others on this assignment, but you must complete and submit your own write up. You should not copy large blocks of code or written text from another student.

Sources

You may refer to class notes, our textbook, Wikipedia, etc.. All sources you refer to must be cited in the space I have provided at the end of this problem set.

In particular, you may find the following resources to be valuable: * Courses assigned on DataCamp * Example R code from class * Cheat sheets and resources linked from [http://www.evanlray.com/stat340_f2019/resources.html]

Load Packages

The following R code loads packages needed in this assignment.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Problem 1: 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".

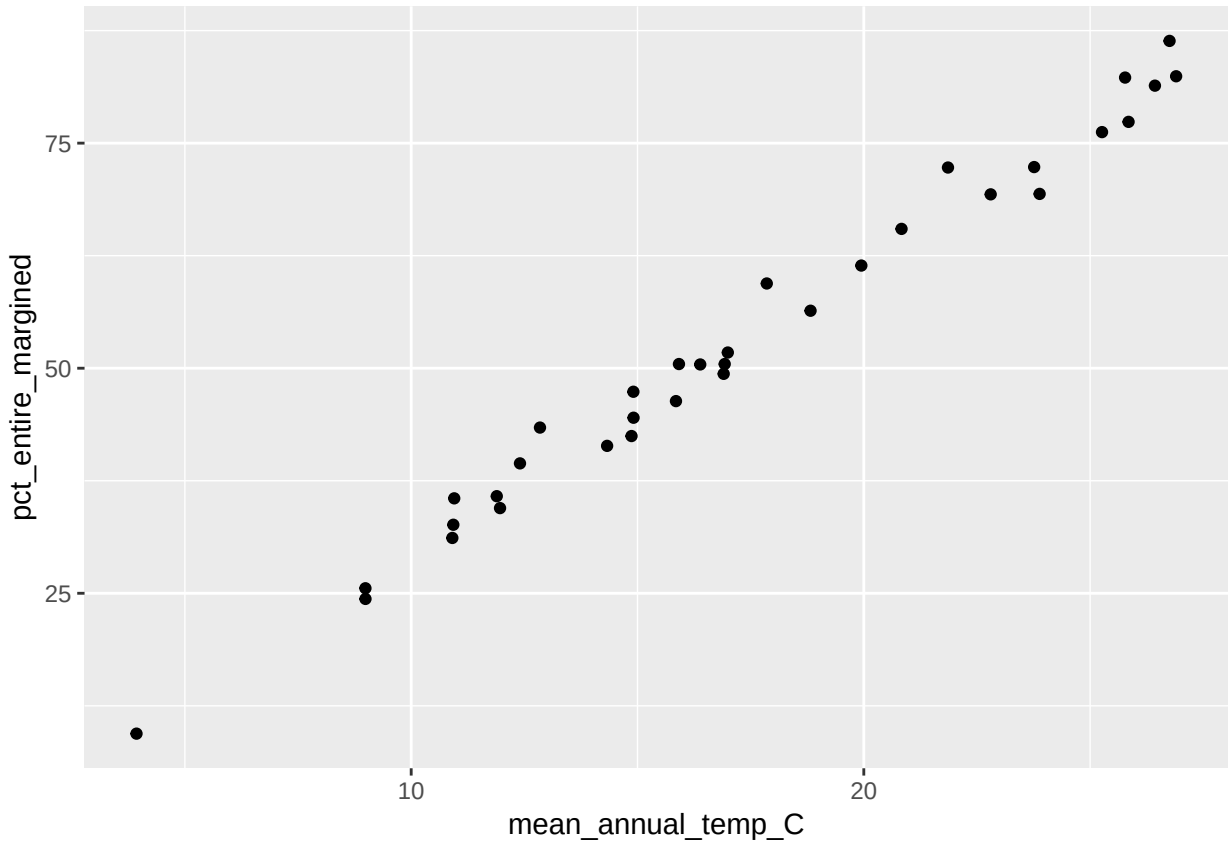
The data we will work with were extracted from that plot and are available in a spreadsheet at http://www.evanlray.com/data/misc/leaf_margins/leaf_margins.csv

(a) Read the data into an R data frame and make a plot with the mean annual temperature on the horizontal axis and the percent of leaves in the given location that are entire margined on the vertical axis.

```
leaf_margins <- 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()
## )
```

```
ggplot(data = leaf_margins, mapping = aes(x = mean_annual_temp_C, y = pct_entire_margined)) +
  geom_point()
```



(b) Fit a linear regression model to the data. Describe the interpretation of the estimated slope.

```
lm_fit <- lm(pct_entire_margined ~ mean_annual_temp_C, data = leaf_margins)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = pct_entire_margined ~ mean_annual_temp_C, data = leaf_margins)
##
## 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
```

We estimate that an increase in mean annual temperature of one degree C is associated with an increase of about 3.18 in the mean percentage of leaves that are entire-margined.

(c) Conduct a hypothesis test of the claim that there is no association between mean temperature and percent of leaves that are entire margined. For this test, clearly state your null and alternative hypotheses in terms of model parameters. State your conclusion in a complete sentence. You don't need to "reject" or "fail to reject" the null hypothesis; instead, interpret what the p-value for the test means in terms of strength of evidence against the null hypothesis. For example, a small p-value like 0.000002 indicates very strong evidence against the null hypothesis, while a large p-value like 0.2 indicates no evidence against the null hypothesis.

The null hypothesis is $H_0 : \beta_1 = 0$. The alternative hypothesis is $H_A : \beta_1 \neq 0$. The p-value for the test is less than 2×10^{-16}, providing extremely strong evidence against the null hypothesis.

(d) Find and interpret a 95% confidence interval for the slope.

```
confint(lm_fit)
```

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

We are 95% confident that that an increase of 1 degree Celsius in mean annual temperature is associated with an increase in mean percent of leaves that are entire-margined that is between 3.04 and 3.32.

Not required since I didn't specifically ask for it: For 95% of samples we might take, a 95% confidence interval calculated using this procedure would contain the increase in mean percent of leaves that are entire-margined that is associated with an increase in mean annual temperature of 1 degree Celsius.

(e) State the model you have fit in matrix form.

$Y = X\beta + \varepsilon$, where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_2 \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

and each ε_i is independently distributed as Normal$(0, \sigma^2)$

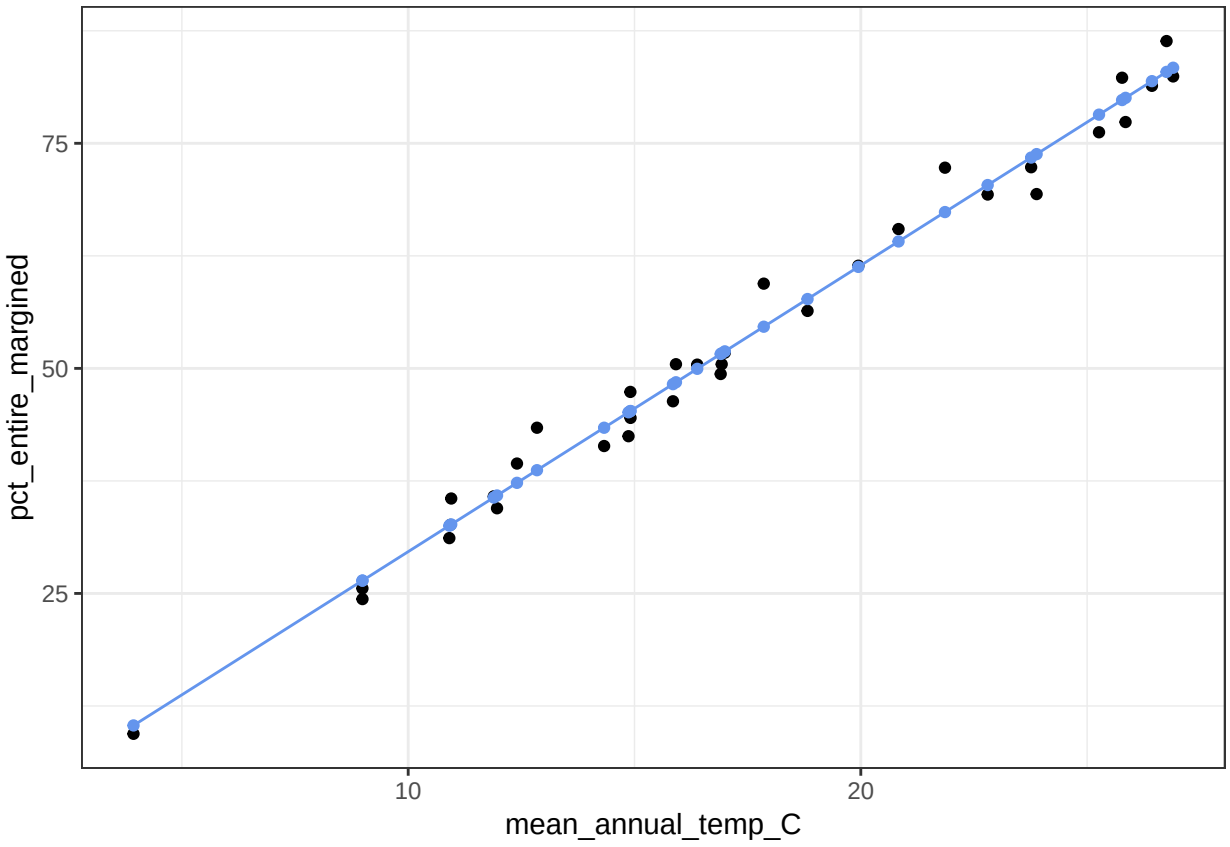
(f) Extract the design matrix from your model fit and use it to find the fitted values for the regression. Make a scatter plot showing the original data as well as the fitted values.

```r
X <- model.matrix(lm_fit)
y <- matrix(leaf_margins$pct_entire_margined)

y_hat <- X %*% solve( t(X) %*% X) %*% t(X) %*% y

leaf_margins <- leaf_margins %>%
  mutate(
    y_hat = y_hat
  )

ggplot(data = leaf_margins) +
  geom_point(mapping = aes(x = mean_annual_temp_C, y = pct_entire_margined)) +
  geom_point(mapping = aes(x = mean_annual_temp_C, y = y_hat), color = "cornflowerblue") +
  geom_line(mapping = aes(x = mean_annual_temp_C, y = y_hat), color = "cornflowerblue") +
  theme_bw()
```



Collaboration and Sources

If you worked with any other students on this assignment, please list their names here.

If you referred to any sources (including our text book), please list them here. No need to get into formal citation formats, just list the name of the book(s) you used or provide a link to any online resources you used.