Stat 343: MLE for Simple Linear Regression

Introduction

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".

I pulled the data out of that plot and put them into the data set that the code below reads in and plots:

```r
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
```

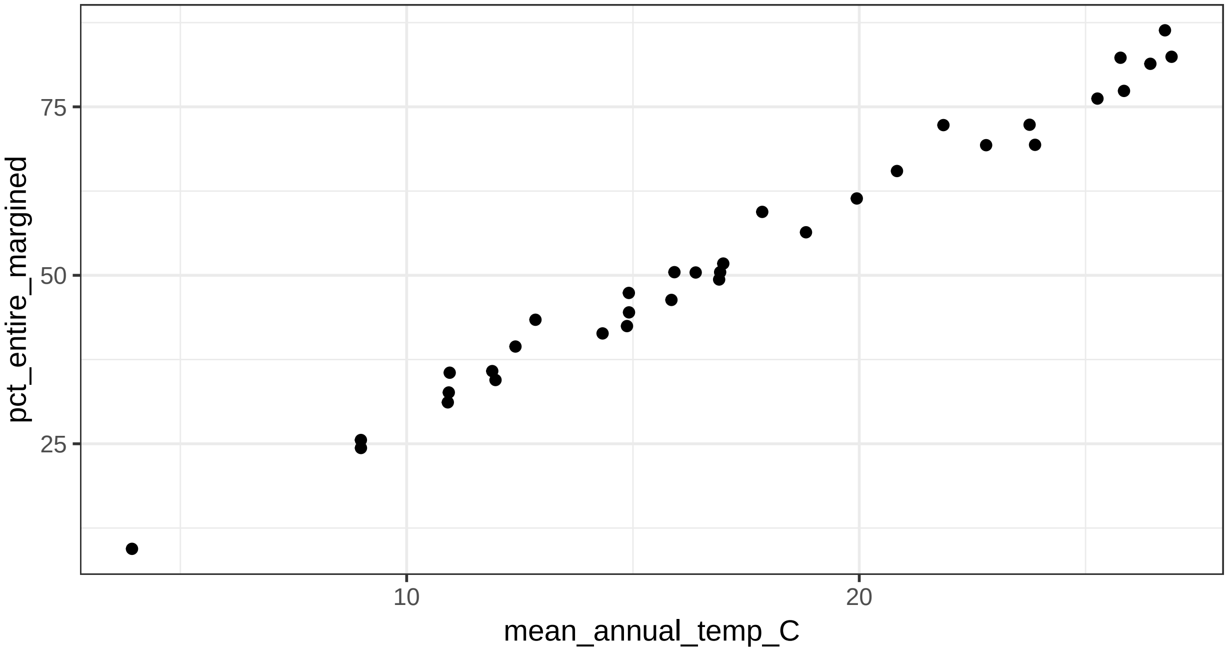
```r
library(ggplot2)

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()
## )
```

```r
head(leaf)
```

```
## # A tibble: 6 x 2
##   pct_entire_margined mean_annual_temp_C
##                 <dbl>              <dbl>
## 1                86.4               26.8
## 2                82.4               26.9
## 3                81.4               26.4
## 4                82.3               25.8
## 5                77.4               25.8
## 6                76.2               25.3
```

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



Let's consider a model for the percent of leaves in the forest canopy as a function of the mean annual temperature in that forest in degrees Celsius.

Notation and Model Statement

We define the following random variables:

Y_i = percent of leaves in forest number i that are entire margined, $i = 1, \ldots, n$.

X_i = mean annual temperature in forest number i in degrees Celsius, $i = 1, \ldots, n$.

We specify our model as follows:

$$\begin{aligned} (Y_i | X_i = x_i) &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i &\overset{\text{iid}}{\sim} \text{Normal}(0, \sigma^2) \end{aligned}$$

If we condition on the value x_i, then $\beta_0 + \beta_1 x_i$ is just a constant. Therefore, we could equivalently state this model as follows:

$$Y_i | X_i = x_i \overset{\text{iid}}{\sim} \text{Normal}(\beta_0 + \beta_1 x_i, \sigma^2)$$

This is a model for the conditional distribution of the percent of leaves that are entire margined in a particular forest given the mean annual temperature in that forest.

The model has three parameters β_0, β_1, and σ^2. For today, let's pretend that σ^2 is a known constant, and focus on estimation of β_0 and β_1.

1. Write down the probability density function for $Y_i|X_i = x_i$.

2. Write down the joint pdf for $Y_1, \ldots, Y_n|X_1 = x_1, \ldots, X_n = x_n$.

3. Find the log-likelihood function $\ell(\beta_0, \beta_1, \sigma^2|x_1, y_1, \ldots, x_n, y_n)$

4. Find a critical point of the log-likelihood function. This will involve taking the partial derivatives with respect to each of β_0 and β_1 and setting the results equal to 0 (remember, we're pretending σ is a known constant; if we were trying to estimate that too, we would need to also differentiate with respect to σ). You will have a system of 2 equations to solve.

You should get answers of the form:

$$\begin{aligned}
\beta_0 &= \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{\left(\sum_{i=1}^n x_i^2\right)\left(\sum_{i=1}^n y_i\right) - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n x_i y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} \\
\beta_1 &= \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}
\end{aligned}$$

To formally identify this critical point as a maximum, you'd have to also verify that the Hessian was negative definite; I won't ask us to do that in this class.

5. If you have extra time: In RStudio, find the maximum likelihood estimates for β_0 and β_1. Confirm that your answers match those from the `lm` function. Add a plot of the estimated regression line to the scatterplot (consider using `geom_abline`).

6. If you have even more extra time: Argue that maximizing the log-likelihood function from part 3 is equivalent to minimizing the sum of squared errors $SSE = \sum_{i=1}^n \{y_i - (\beta_0 + \beta_2 x_i)\}^2$. Thus, for this model, maximum likelihood is equivalent to estimating the coefficients by ordinary least squares.