Stat 3301: Homework 5

Due by date and time specified on Carmen

Nathan Johnson

Setup:

```
knitr::opts_chunk$set(echo = TRUE)
library(alr4)
library(tidyverse)
```

Instructions

- Replace "FirstName LastName (name.n)" above with your information.
- Provide your solutions below in the spaces marked "Solution:".
- Include any R code that you use to answer the questions; if a numeric answer is required, show how you calculated it in R. I have set the global option `echo = TRUE` to make sure the R code is displayed.
- Knit this document to HTML and upload both the HTML file and your completed Rmd file to Carmen
- Make sure your solutions are clean and easy-to-read by
  - formatting all plots to be appropriately sized, with appropriate axis labels.
  - only including R code that is necessary to answer the questions below.
  - only including R output that is necessary to answer the questions below (avoiding lengthy output).
  - providing short written answers explaining your work, and writing in complete sentences.
- Data files mentioned below are from the `alr4` package unless specified otherwise.

Concepts & Application In this assignment, you will

- use plots of data to assess whether log transforms are useful.
- see an example of the "log rule" in practice.
- fit regression models with the predictor and/or response in log scale.
- interpret $\hat{\beta}_1$ when the predictor and/or response are in log scale.
- construct confidence intervals when the predictor is in log scale.
- make a plot of a fitted model when the predictor is in log scale.

Question 1 The data set `ufcwc` in the `alr4` package contains data on a sample of western cedar trees that was taken in 1991 in the Upper Flat Creek stand of the University of Idaho Experimental Forest. Each case in the data set represents a tree. In this question we will look at the relationship between the variables:

Variable	Description
`Height`	height of the tree in decimeters
`Dbh`	diameter of the tree in mm measured at 137 cm above the ground

We will be interested in using tree diameter (`Dbh`) to predict tree height (`Height`).

(a) For the linear regression model $Height_i = \beta_0 + \beta_1 \log Dbh_i + e_i$, report the numeric values of estimated parameters $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}^2$.

(b) Use the fitted model to compute "by hand" (i.e., don't use the `predict` function) a 95% confidence interval for average tree height for trees that have a diameter of 500 mm.

(c) Make a scatterplot with `log(Dbh)` on the x-axis and `Height` on the y-axis, and add the estimated linear regression line to the plot.

(d) Make a scatterplot with `Dbh` (on its original scale) on the x-axis and `Height` on the y-axis. Use your fitted model from part (a) to add a (curved) line to the plot that represents the estimated average tree height as a function of tree diameter, $\hat{E}(\texttt{Height} \mid \texttt{Dbh})$. Add dashed (curved) lines to the plot that represent point-wise, 95% confidence intervals for the mean.

(e) Using the estimated slope coefficient, $\hat{\beta}_1$, write a sentence that quantifies what our model says about how average tree height depends on tree diameter (on its original scale).

Solution to Question 1 Your answers go here.

```
library(alr4)
```

A.

```
X = ufcwc$Dbh
Y = ufcwc$Height
SlogXY = sum((log10(X) - mean(log10(X)))*(Y))
SlogXX = sum((log10(X) - mean(log10(X)))^2)

B1 = SlogXY / SlogXX
B0 = mean(Y) - (B1 * mean(log10(X)))
lm.ufc = lm(Height ~ log10(Dbh), data = ufcwc)

s2 = sum((Y - (B0 + B1*log10(X)))^2)/(length(X)-2)
B1
```

```
## [1] 275.2032
```

```
B0
```

```
## [1] -463.3144
```

```
s2
```

```
## [1] 1111.185
```

$\hat{\beta}_0 = -463.3, \hat{\beta}_1 = 275.2, \hat{\sigma}^2 = 1111.19$

b. Use $\hat{\beta}_0 \pm t(\alpha/2, n-2) * se(\hat{\beta}_0 \mid X)$

```
value = B0 + B1 * log10(500)
alpha = 0.05
n = length(ufcwc$Height)
mean = mean(ufcwc$Dbh)
se = sqrt(s2) * sqrt((1/n) + (log10(500) - mean(log10(ufcwc$Dbh)))/SlogXX)
lwr = value - pt(alpha/2, n - 2, lower.tail = FALSE) * se
uwr = value + pt(alpha/2, n - 2, lower.tail = FALSE) * se
lwr
```

```
## [1] 276.5818
```

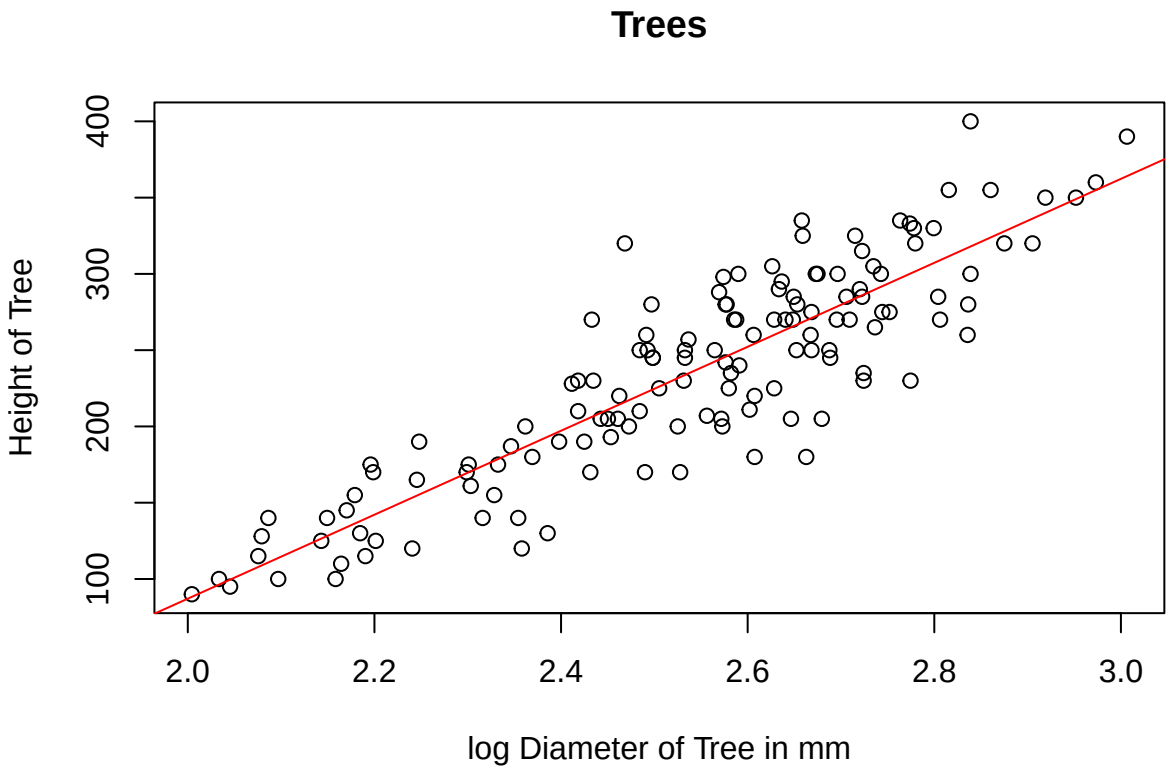
```
uwr
```

```
## [1] 282.3194
```

The average tree height will be between 276.58 and 282.32 decimeters with 95% confidence.

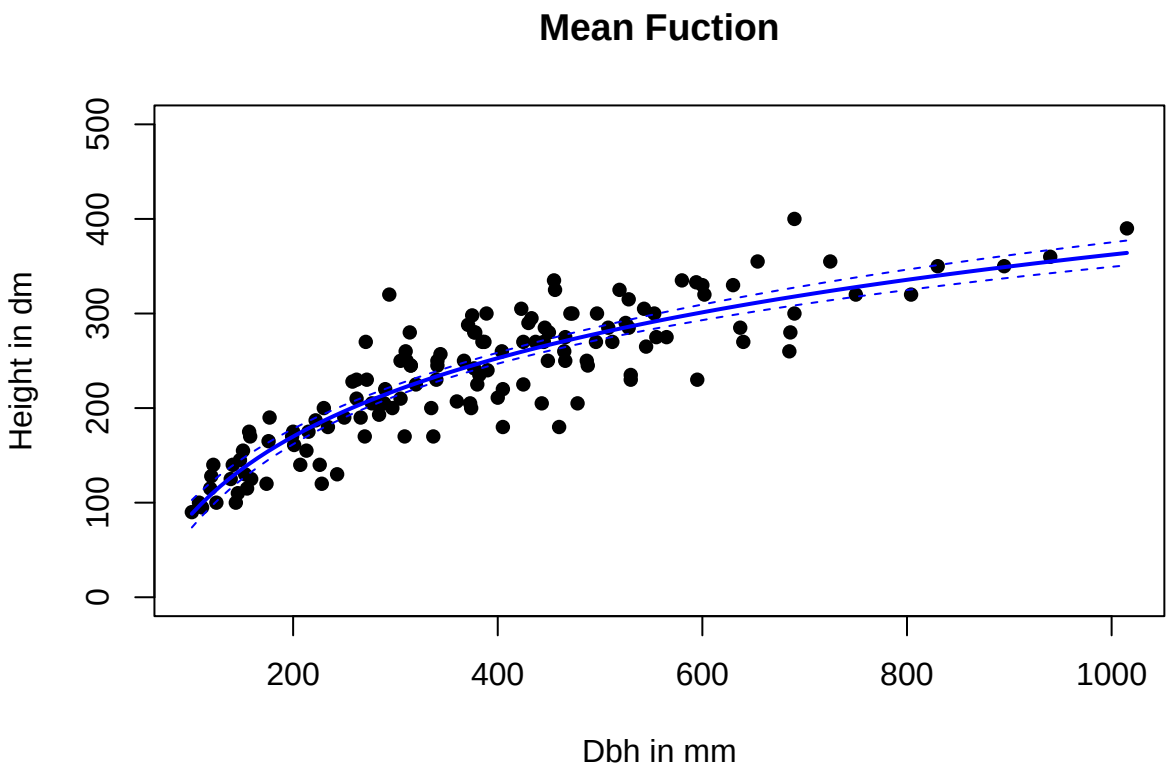
c.

```
plot(log10(ufcwc$Dbh), ufcwc$Height, xlab = "log Diameter of Tree in mm", ylab = "Height of Tree", main
abline(a = B0, b = B1, col = "Red")
```



d.

```
d.grid = data.frame(Dbh=seq(from=min(ufcwc$Dbh), to=max(ufcwc$Dbh), length.out=100))
ypred = predict(lm.ufc, newdata=d.grid, interval="confidence")
plot(ufcwc$Dbh, ufcwc$Height, pch=16, xlab="Dbh in mm", ylab="Height in dm",
main="Mean Fuction", ylim=c(0,500))
lines(d.grid[,1], ypred[,1], col="blue", lwd=2)
lines(d.grid[,1], ypred[,2], col="blue", lty=2); lines(d.grid[,1], ypred[,3], col="blue", lty=2)
```



e.

```
TenPercent = B1 * log10(1.1)
TenPercent
```

```
## [1] 11.3914
```

For every 10% larger the diameter of the tree is, there will be a 11.39 decimeter increase in height of the tree.

Question 2 The data set `AdRevenue.txt` on `Carmen` (Source: Sheather, 2009) contains data on the `n = 70` top magazines in the US in terms of gross advertising revenue in 2006. The variables in the data set include the magazine name, the parent company of the magazine, the total ad revenue per page (in thousands of dollars) and the magazine circulation (in millions). This problem concerns predicting gross advertising revenue per page in 2006 from circulation.

(a) Make four plots: (i) `AdRevenue` vs. `Circulation`, (ii) `AdRevenue` vs. `log(Circulation)`, (iii) `log(AdRevenue)` vs. `Circulation`, (iv) `log(AdRevenue)` vs. `log(Circulation)`. Explain why transforming both `AdRevenue` and `Circulation` to log scale is useful for building a linear regression model that uses circulation to predict ad revenue.

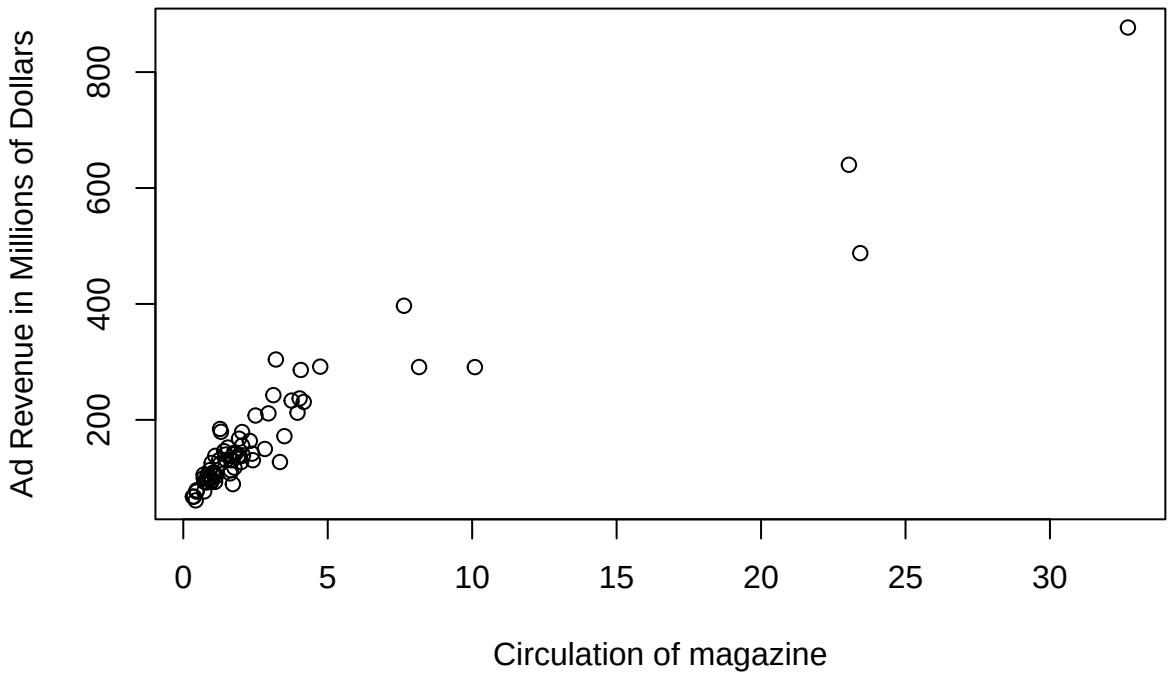
(b) Recall the book's definition of the "log rule" (p188): "If the values of a variable range over more than one order of magnitude and the variable is strictly positive, then replacing the variable by its logarithm is likely to be helpful" in linearizing relationships. Is that the case here? Say why or why not.

(c) A company is interested in quantifying how increases in circulation are associated with changes in ad revenue. To do this, they fit a linear regression model with `log(Circulation)` as the predictor and `log(AdRevenue)` as the response. Use the estimated parameter $\widehat{\beta}_1$ to provide, in plain English, a careful interpretation of how a change in the predictor variable `Circulation` (on its original scale) is associated with a change in average ad revenue (on its original scale).

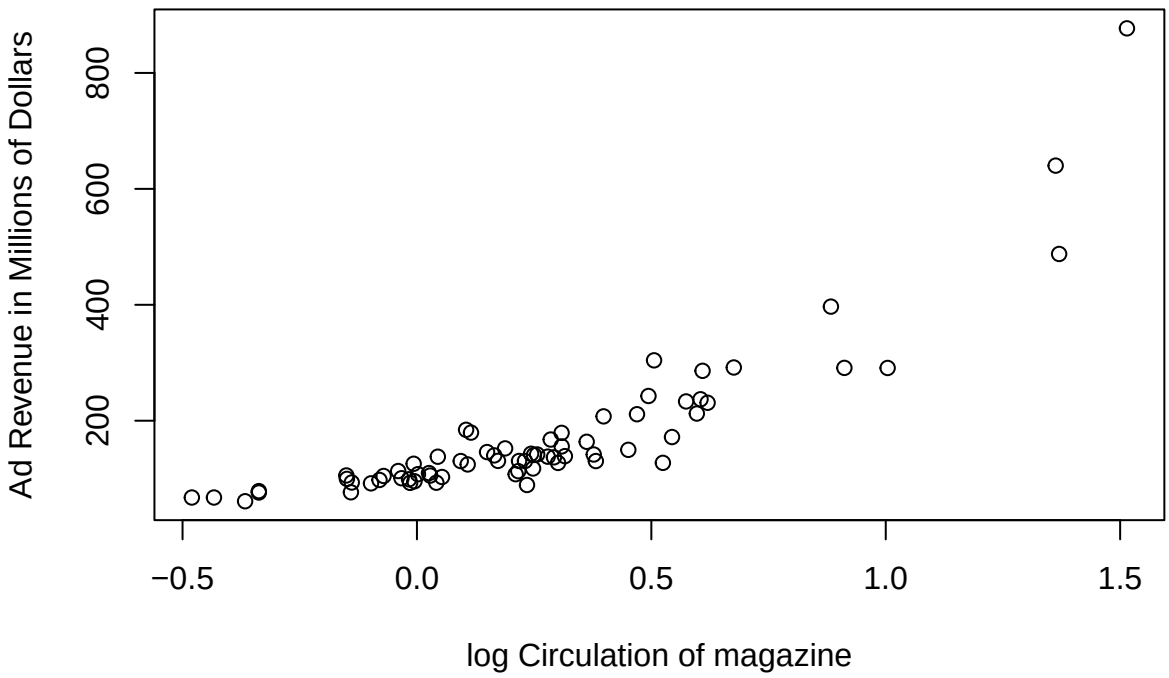
Solution to Question 2 Your answers go here.

a.

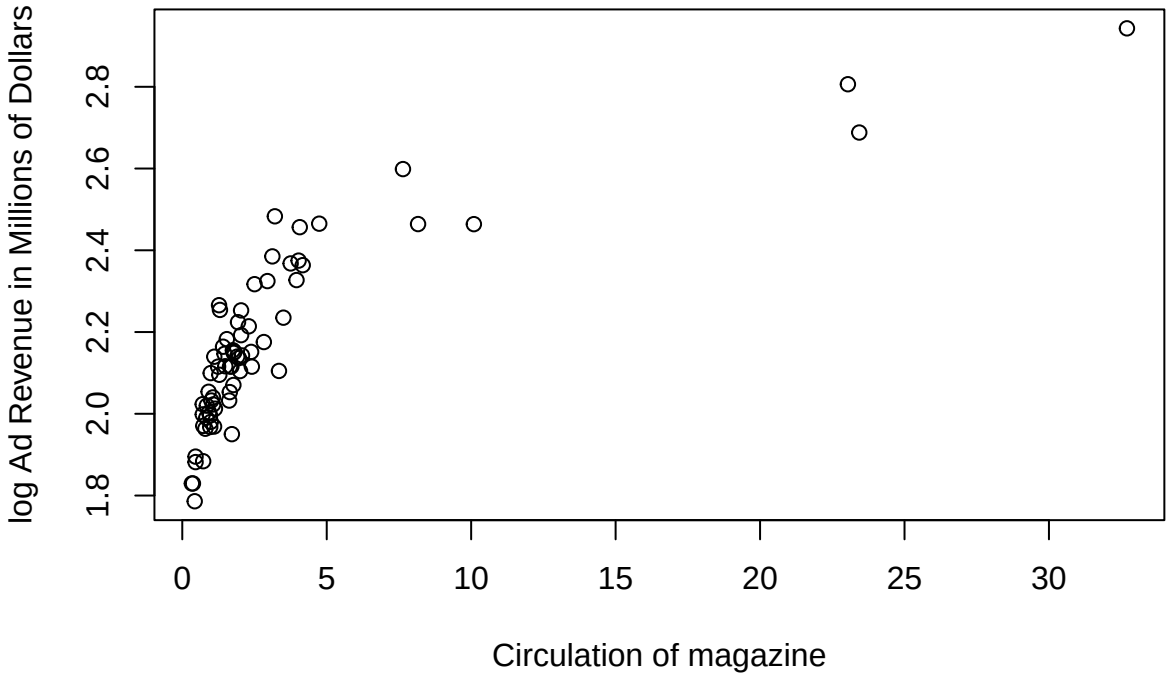
```
rev = read.csv('AdRevenue-1.txt')
plot(rev$Circulation, rev$AdRevenue, xlab = "Circulation of magazine", ylab = "Ad Revenue in Millions o
```



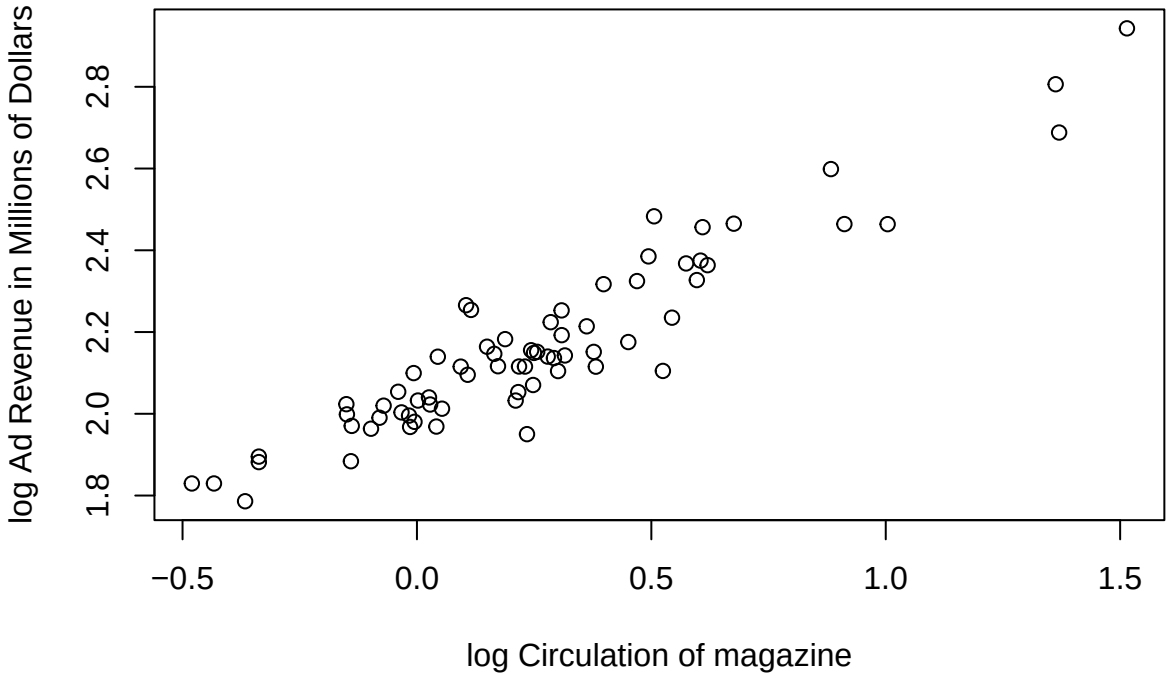

```
plot(log10(rev$Circulation), rev$AdRevenue, xlab = "log Circulation of magazine", ylab = "Ad Revenue in
```




```
plot(rev$Circulation, log10(rev$AdRevenue), xlab = "Circulation of magazine", ylab = "log Ad Revenue in
```




```
plot(log10(rev$Circulation), log10(rev$AdRevenue), xlab = "log Circulation of magazine", ylab = "log Ad
```



Plots i, ii, iii are all not suitable for linear modeling. It is easy to see just from visuals that there is not a linear relation between these models. But plot iv does appear to be linear.

b. Yes, it is the case here. While said rule applies to plot ii and plot iii, it unfortunately did not seem to create the linear model we were hoping for. But plot iv does create what appears to be a linear model.

c.

```
SlogXlogYrev = sum((log10(rev$Circulation) - mean(log10(rev$Circulation)))*(log10(rev$AdRevenue))) / (l
SlogXXrev = sum((log10(rev$Circulation) - mean(log10(rev$Circulation)))^2) / (length(rev$Circulation) -

B1rev = SlogXlogYrev / SlogXXrev
B1rev
```

```
## [1] 0.528758
```

```
interpret = exp((log10(1.1))*B1rev) - 1
interpret
```

```
## [1] 0.02212799
```

For a 10% increase in circulation, there is a 2.21% increase in ad revenue.