Stat 230 HW 3

Name: Victor Huang

worked with: No one

Homework 3 is due **by 3pm Monday, Oct. 4**. Please complete the assignment in this Markdown document, filling in your answers and R code below. I didn't create answer and R chunk fields like I did with homework 1, but please fill in your answers and R code in the same manner as hw 1. Submit a hard copy of the **compiled pdf or word doc** either

- in class on Monday 9/20
- in drop-in office hours (Tuesday 9/21)
- in the paper holder outside my CMC 222 office door (hopefully it will be installed by then!)

Tips for using Markdown with homework sets:

- Work through a problem by putting your R code into R chunks in this .Rmd. Run the R code to make sure it works, then knit the .Rmd to verify they work in that environment.
  - Make sure you load your data in the .Rmd and include any needed `library` commands.
- Feel free to edit or delete questions, instructions, or code provided in this file when producing your homework solution.
- For your final document, you can change the output type from `html_document` to `word_document` or `pdf_document`. These two to output types are better formatted for printing.
  - on maize: you may need to allow for pop-ups from this site
- If you want to knit to pdf while running Rstudio from your computer (*not* from maize), you will need a LaTeX compiler installed on your computer. This could be MiKTeX, MacTeX (mac), or TinyTex. The latter is installed in R: first install the R package `tinytex`, then run the command `tinytex::install_tinytex()` to install this software.
  - If you are using maize, you don't need to install anything to knit to pdf!

Problem 1: Election Fraud: ch. 8 exercise 20 (a-c)

- For part (b), add the regression line and prediction bands from (b) to the plot created in (a).
- The data for this problem is `ex0820`
- for (a): To highlight the disputed election in your `ggplot`, add the layer `geom_point(data=filter(ex0820, Disputed == "yes"), color="red")` using the `filter` command from the `dplyr` package. You can even play with the `size` argument in `geom_point` to increase the size so the point stands out when printed in black and white.

From the functions below, we can see that a 49.3 percentage of machine-count would lead to a prediction of 50.8747 with a standard error of 9.864 on 20 degrees of freedom. The difference is 28.1253 percent from the predicted value.

$$t = \frac{15.7238}{9.864} = 1.594$$

$$p = 2 * (1 - pt(1.594, df = 20)) = 0.1266178$$

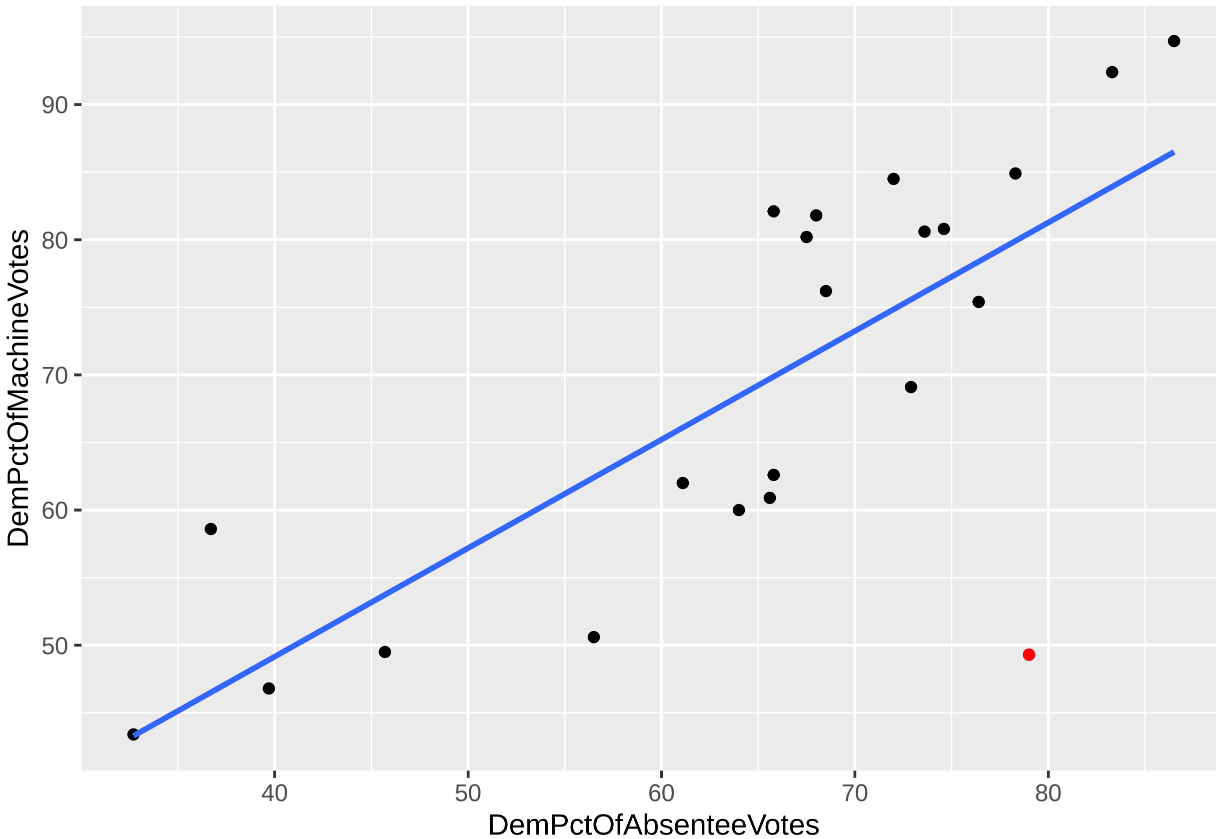
```
> 2*(1-pt(1.594, df = 20))
[1] 0.1266178
```

```
> 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)
> library(Sleuth3)
> ggplot(ex0820, aes(x =DemPctOfAbsenteeVotes, y = DemPctOfMachineVotes))+ geom_point() + geom_point(da
`geom_smooth()` using formula 'y ~ x'
```




```
> ex0820_lm <- lm(DemPctOfAbsenteeVotes ~ DemPctOfMachineVotes, data = ex0820)
> summary(ex0820_lm)

Call:
lm(formula = DemPctOfAbsenteeVotes ~ DemPctOfMachineVotes, data = ex0820)

Residuals:
    Min      1Q  Median      3Q     Max
-20.804  -5.382   1.220   5.258  28.127
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           15.7238     9.7878   1.606    0.124
DemPctOfMachineVotes   0.7130     0.1378   5.175  4.6e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.864 on 20 degrees of freedom
Multiple R-squared:  0.5725,	Adjusted R-squared:  0.5511
F-statistic: 26.78 on 1 and 20 DF,  p-value: 4.605e-05
```

Problem 2: Island Area and Species

Consider Conceptual Exercise 1 (pg.227) in chapter 8. (Its solution is at the end of the chapter.) They show in this example that a halving of area is associated with a 16% reduction in median number of species. What happens if we double area? Show all work, be specific (give an amount of change, explain what is changing (median? mean?), and explain your answer in context. Make sure to explain your answer in terms of the original scale of the variables (not on the log scale).

As shown below, doubling the area is associated with a 19% increase in median number of species.

$$\hat{\mu}(log(species)|log(area)) = 1.94 + 0.250log(area)$$
$$\hat{\mu}(species|area) = e^{1.94+0.250log(area)}$$
$$\hat{\mu}(species|area) = e^{1.94} \times area^{0.250}$$
$$PowerModel_{area} : 2^{0.250} = 1.19$$

Problem 3: Pollution

The data set `Pcb.csv` contains information on PCB (a hazardous industrial chemical) levels (ppm, parts per million) in various bodies of water for the years 1984 and 1985. Researchers would like to understand how levels of the pollutant varies from year to year. We will consider how to build a model for 1985 PCB levels based on the 1984 levels.

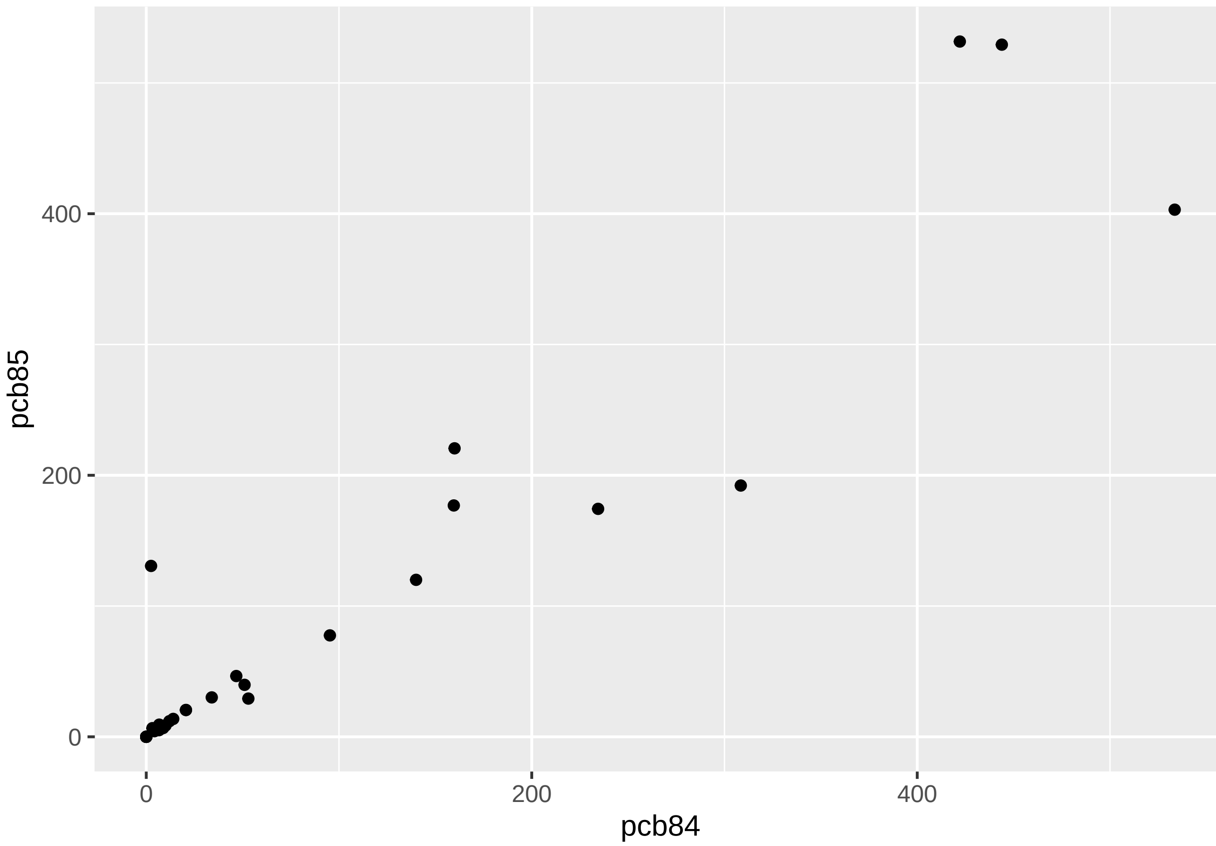
```
> pcb <- read.csv("http://math.carleton.edu/kstclair/data/Pcb.csv")
```

(3a)

Plot `pcb85` against `pcb84`. Notice the obvious outlier. Identify this point by site name and row number. Redraw this plot without the outlier. Explain why a SLR model is not appropriate for this data even with the outlier removed.

The outlier is Boston Harbor row 4. The SLR still doesn't work due to a lack of normality (shape of population response values is not described by a normal distribution)

```
> ggplot(pcb %>% slice(-4), aes(x = pcb84, y = pcb85)) + geom_point()
```



(3b)

Plot log of `pcb8` against log of `pcb84` (including the outlier from (a)). Explain why we could fit a SLR model to the logged versions of pcb.

We can fit the SLR model now due to the fact that they haev a linear mean, constant Sd, narmality, and Independence.

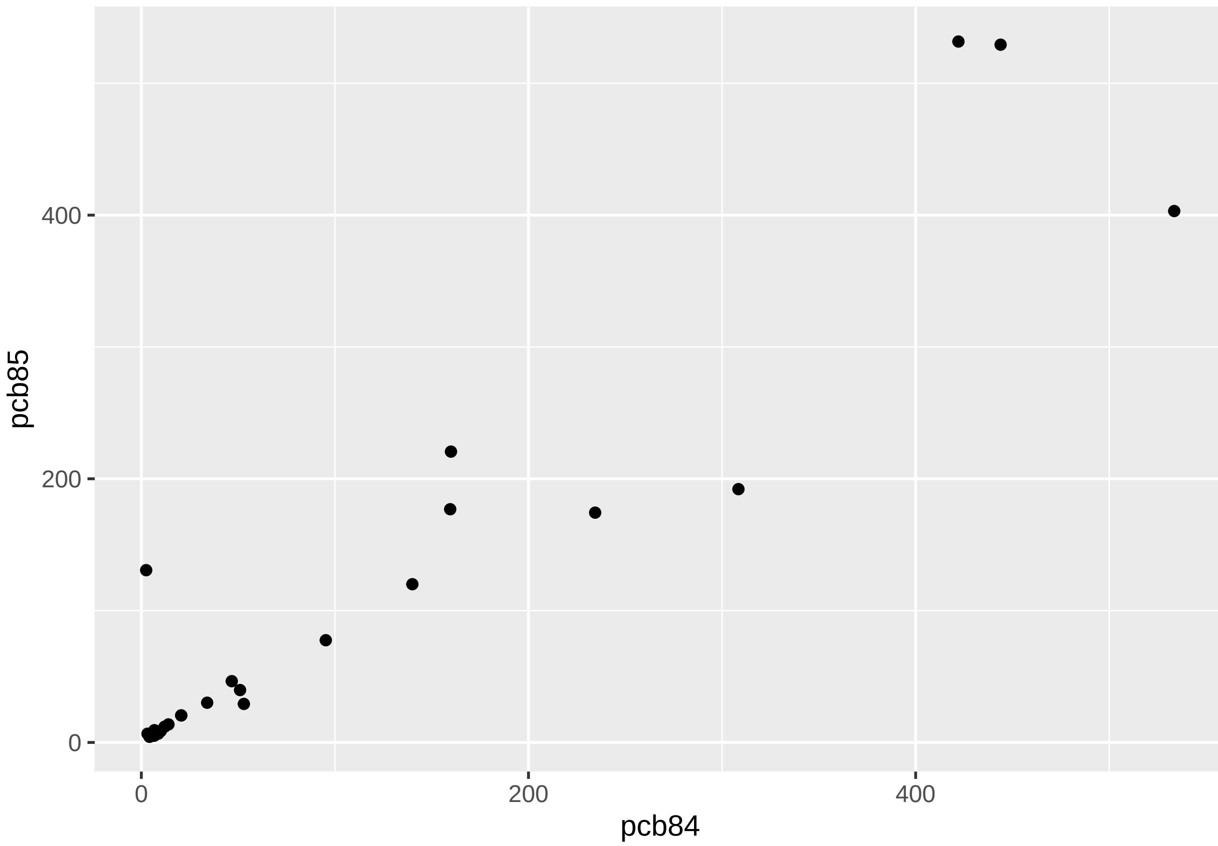
(3c)

Look at the values in the vector `log(pcb$pch85)`. Why are some values equal to `-Inf`? Give the names of the sites that contain no pcb in 1984 and 1985, then create a subsetted data set that excludes these sites. Use this new subsetted data set to complete parts (d) below.

Hint: you can use the `filter` command with the filtering arguments equal to `pcb84 > 0,  pcb85 >0`. You can similarly use filter to find the "0" pcb cases, but be sure to use `pch84 == 0` to ask which cases are equal to (`==`) 0.

The name of the sites are Pamilico Sound, Sapelo Sound, Tampa Bay, Mobile Bay, Round Island, Barataria Bay, San Antonio Bay, and Corpus Christi Bay

```
> ggplot(pcb %>% slice(-4, -12,-14,-16,-18,-19,-21,-22,-23), aes(x = pcb84, y = pcb85)) + geom_point()
```



(3d)

With the subsetted data `pcb_nonzero` from (c), fit the regression of log of pcb in 1985 on the log of pcb in 1984. Check and comment on model assumptions and identify two obvious outliers (by site name and row number from `pcb_nonzero`).

We get the following regression. The two obvious are Raritan Bay and San Diego Harbor (rows 9 and 24 respectively).

```
> lm(log(pcb85) ~ log(pcb84), data = pcb %>% slice(-4, -12,-14,-16,-18,-19,-21,-22,-23))

Call:
lm(formula = log(pcb85) ~ log(pcb84), data = pcb %>% slice(-4,
    -12, -14, -16, -18, -19, -21, -22, -23))

Coefficients:
(Intercept)   log(pcb84)
     0.6576       0.8340
```

(3e)

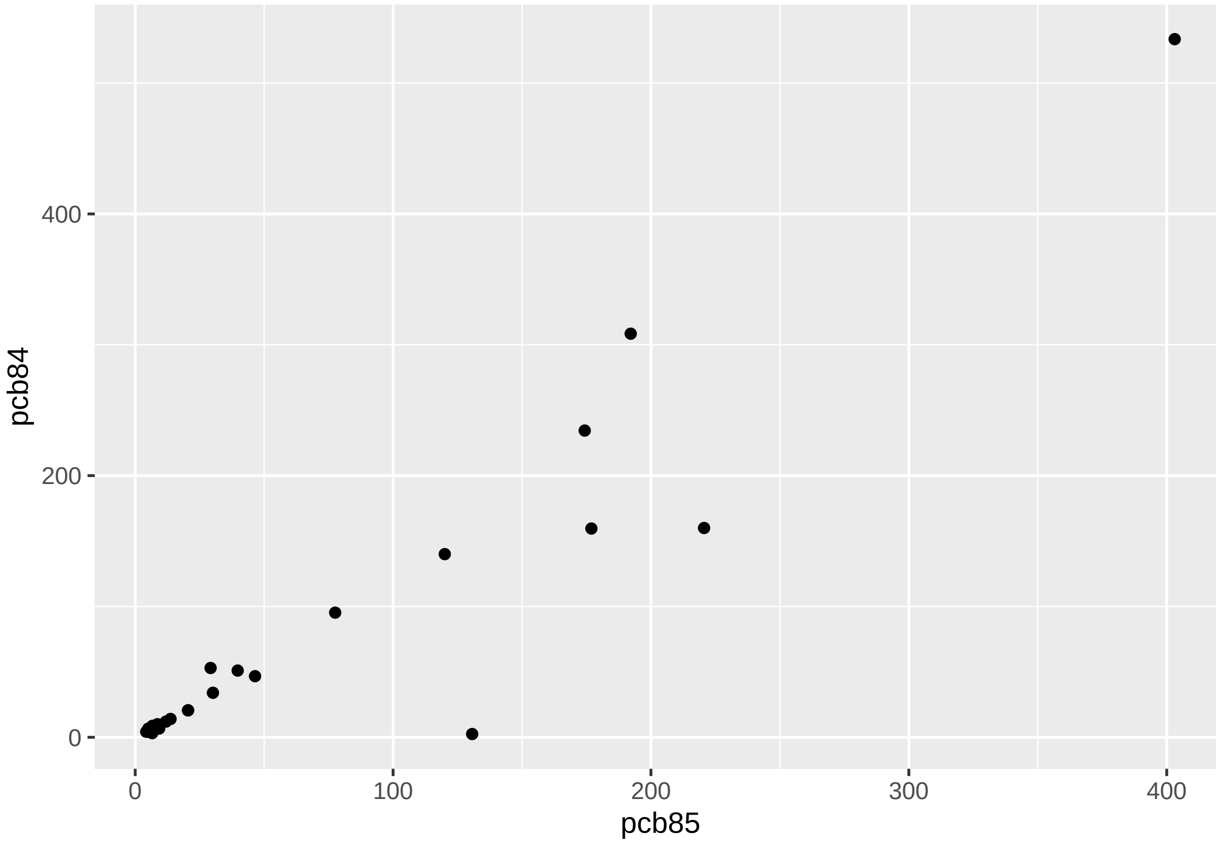
Create one more data frame, `pcb_nonzero2` that removes these two outliers. Then use this data to refit the model from (d) without the two outliers, verify that the residual plot looks better than in (d) and explain how they influence the estimated model slope and R-squared value.

It does look better, The model becomes more linear and the R-squared value decreased.

```
> lm(log(pcb85) ~ log(pcb84), data = pcb %>% slice(-4,-12,-14,-16,-18,-19,-21,-22,-23,-9,-24))
```

```
Call:
lm(formula = log(pcb85) ~ log(pcb84), data = pcb %>% slice(-4,
    -12, -14, -16, -18, -19, -21, -22, -23, -9, -24))

Coefficients:
(Intercept)   log(pcb84)
     0.7816       0.7816
> ggplot(pcb %>% slice(-4, -12,-14,-16,-18,-19,-21,-22,-23,-9,-24), aes(x = pcb85, y = pcb84)) + geom_p
```



(3f)

Using the estimated model from (e) without the two outliers, interpret the slope of the equation and R-squared, in context. Make sure to interpret the slope effect on the original scale of both variables.

The slope shows that an increase in 1 unit of log(pcb84) is associated with a 0.7816 increase in log(pcb85). On the original scale. A doubling of pcb84 is associated with a 71% increase in the median level of pcb in 85.

Problem 4: Mammal Brain Weights

Consider Conceptual Exercise 4 (pg.261) in chapter 9. Note that they use the natural log [base-e] for each variable in this model:

$$\hat{\mu}(\log(brainwt) \mid x) = 0.8548 + 0.5751 \log(bodywt) + 0.4179 \log(gest) - 0.3101 \log(litter)$$

(4a)

Write down the expression for the (mean? median?) brain weight given body weight, gestation, and litter. Show all work and simplify as much as possible.

The first equation shows the mean of the log of the brain weight given the body weight, gestation, and litter

The second equation shows the median of the brain weight given the body weight, gestation, and litter

$$\hat{\mu}(\log(brainwt) \mid x) = 0.8548 + 0.5751 \log(bodywt) + 0.4179 \log(gest) - 0.3101 \log(litter)$$
$$\hat{med}(brainwt \mid x) = e^{0.8548+0.5751 \log(bodywt)+0.4179 \log(gest)-0.3101 \log(litter)}$$

$$\hat{med}(brainwt \mid x) = \frac{e^{0.8548} \times (bodywt)^{0.5751} \times gest^{0.4179}}{litter^{0.3101}}$$

(4b)

Interpret, in context and on the original scale, the effect of body size on brain weight. As usual, show all work and be specific - give an amount of the effect and be careful about what is changing (median? mean?). Make sure to explain your answer in terms of the original scale of the variables (not on the log scale). Show all work.

The power model for the bodyweight shows that doubling the bodyweight would lead to an increase of 49% of the brain weight. The power model for the gestation period shows that doubling the gestation period would lead to an increase of 34% of the brain weight. The power model for the litter size shows that doubling the litter size would lead to a decrease of 19% of the brain weight. Looking at the original equation, we can see that $\beta_1 = 0.5751, \beta_2 = 0.4179, \beta_3 = -0.3101$. An increase in log of bodyweight, gestation period, and litter size by 1 unit is associated with an estimated 0.5751, 0.4179, and -0.3101 unit increase in the mean of the log of brain size, while all the other variables are held constant.

$$\hat{med}(brainwt \mid x) = \frac{e^{0.8548} \times (2 \times bodywt)^{0.5751} \times gest^{0.4179}}{litter^{0.3101}}$$

$$PowerModel_{bodyweight} = 2^{0.5751} = 1.49$$
$$PowerModel_{gestation} = 2^{0.4179} = 1.34$$
$$PowerModel_{litter} = 2^{-0.3101} = 0.81$$

Problem 5: Perch weights

Consider estimating the weight (g) of a perch (fish) given its width (cm) and length (cm). The estimated mean weight given width and length is given by the function

$$\hat{\mu}(\text{Weight} \mid \text{Width, Length}) = 113.9349 - 3.4827\text{Length} - 94.6309\text{Width} + 5.2412\text{Length} \times \text{Width}$$

(5a)

A one cm increase in length has what effect on mean weight?

A one cm increase in length is associated with a mean increase in weight by -3.4827 + 5.2412Width

$$\hat{\mu}(\text{Weight} \mid \text{Width, Length}) = 113.9349 - 3.4827\text{Length} - 94.6309\text{Width} + 5.2412\text{Length} \times \text{Width}$$
$$\hat{\mu}(\text{Weight} \mid \text{Width, Length + 1}) = 113.9349 - 3.4827(\text{Length + 1}) - 94.6309\text{Width} + 5.2412(\text{Length + 1}) \times \text{Width}$$

$$\hat{\mu}(\text{Weight} \mid \text{Width}, \text{Length} + 1) = 113.9349 - 3.4827\text{Length} - 3.4827 - 94.6309\text{Width} + 5.2412\text{Length} \times \text{Width} + 5.2412\text{Width}$$

$$\hat{\mu}(\text{Weight} \mid \text{Width}, \text{Length} + 1) - \hat{\mu}(\text{Weight} \mid \text{Width}, \text{Length}) = -3.4827 + 5.2412\text{Width}$$

(5b)
For perch that are 6 cm wide, a one cm increase in length has what effect on mean weight?

Using our previous equation, we can see that id width is held constant at 6cm, a one centimeter increase in length is associarted with a mean increase in weight by 27.9645 grams.

$$\hat{\mu}(\text{Weight} \mid \text{Width}, \text{Length} + 1) - \hat{\mu}(\text{Weight} \mid \text{Width}, \text{Length}) = -3.4827 + 5.2412(6) = 27.9645$$

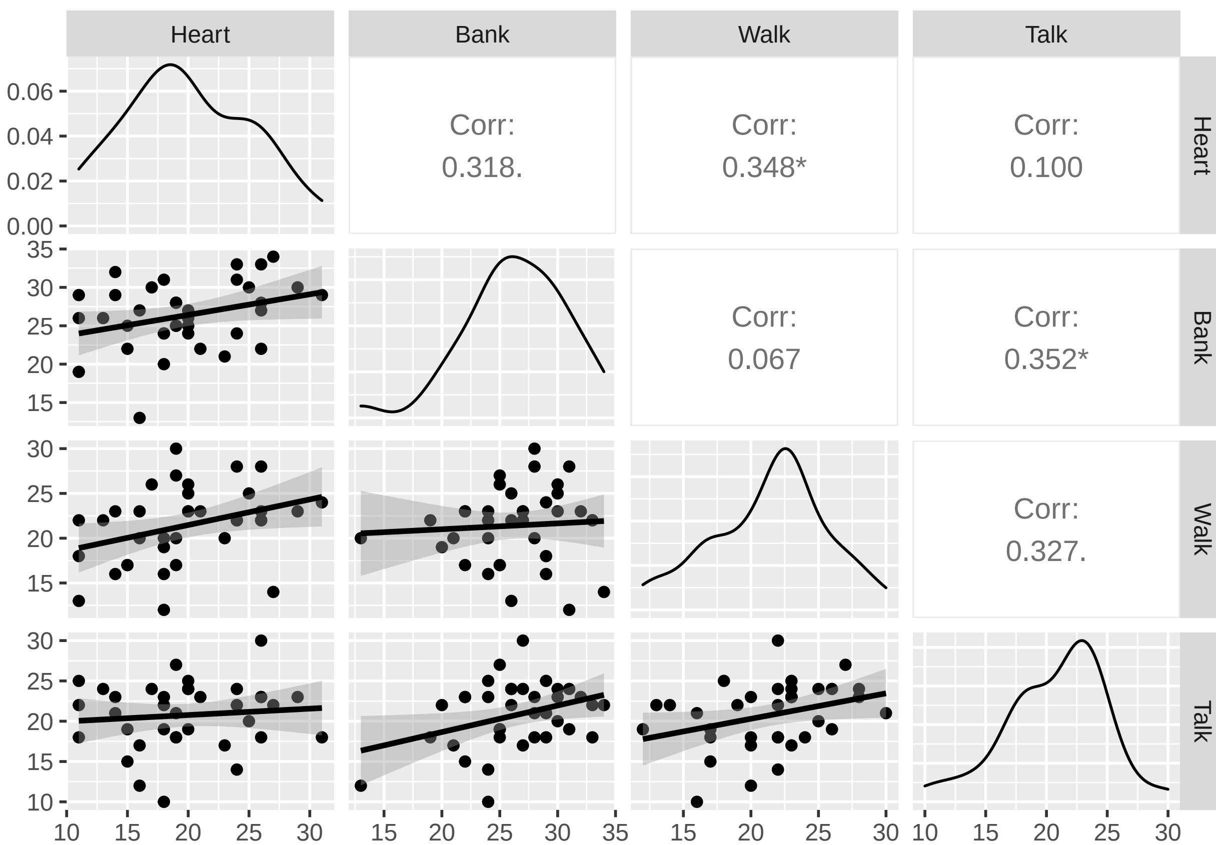
Problem 6: Pace of life and heart disease: ch. 9 exercise 14

Take a look at the data described in ch. 9 exercise 14. The data is in the Sleuth data frame `ex0914`. Answer the questions below about this data set:

(6a) Draw the scatterplot matrix of the four variables in this data set. Describe how `Heart` is related to the covariates `Bank`, `Walk` and `Talk`

Heart is directly related to the Bank, Walk, and Talk covariates with respective corresponding values of 0.318, 0.348, and 0.100.

```
> library(GGally)
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
> library(Sleuth3)
> ggpairs(ex0914, columns = c("Heart", "Bank", "Walk", "Talk"), lower = list(continuous = "smooth", se =
```



(6b) Fit the regression of `Heart` on `Bank`, `Walk` and `Talk` and write down the fitted (estimated) mean function.

The fitted (estimated) mean function is as follow:

$$\hat{\mu}(Heart|Bank, Walk, Talk) = 3.1787 + 0.4052Bank + 0.4516Walk - 0.1796Talk$$

```
> lm(formula = Heart ~ Bank + Walk + Talk, data = ex0914)

Call:
lm(formula = Heart ~ Bank + Walk + Talk, data = ex0914)

Coefficients:
(Intercept)         Bank         Walk         Talk
     3.1787       0.4052       0.4516      -0.1796
```

(6c) Interpret (in context!) the effects of each of the three predictors bank, walk, and talk on the response heart.

Looking at our fitted mean function we can see that a unit increase in Bank clerk speed is associated with a 0.4052 increase in heart disease rate, a unit increase in pedestrian walk speed is associated with a 0.4519 increase in heart disease rate, and a unit increase in postal clerk talking speed is associated with a 0.1796 decrease in heart disease rate.