

# Problem Set 7

Daniel Shapiro

10/26/2022

## Question 1 Background:

Download the dataset *quartet.dta* from the course website.

1a) Load the dataset. The file is in a format compatible with Stata, but not with base R. Google how to load .dta files in R.

```
dataset <- read_dta("quartet.dta")
```

1b) Regress each  $y$  on its corresponding  $x$  (e.g.,  $y_1$  on  $x_1$ ,  $y_2$  on  $x_2$ ) using the `lm()` command. Using `stargazer`, present the results in a nicely formatted table. Interpret the regression coefficients.

```
reg1 <- lm(`y1` ~ `x1`, data = dataset)
reg2 <- lm(`y2` ~ `x2`, data = dataset)
reg3 <- lm(`y3` ~ `x3`, data = dataset)
reg4 <- lm(`y4` ~ `x4`, data = dataset)

sum1 <- summary(reg1)
sum2 <- summary(reg2)
sum3 <- summary(reg3)
sum4 <- summary(reg4)

stargazer(reg1, reg2, reg3, reg4, type = "text")
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               y1      y2      y3      y4
##                               (1)      (2)      (3)      (4)
## -----
## x1                        0.500***
##                        (0.118)
##
## x2                        0.500***
##                        (0.118)
```

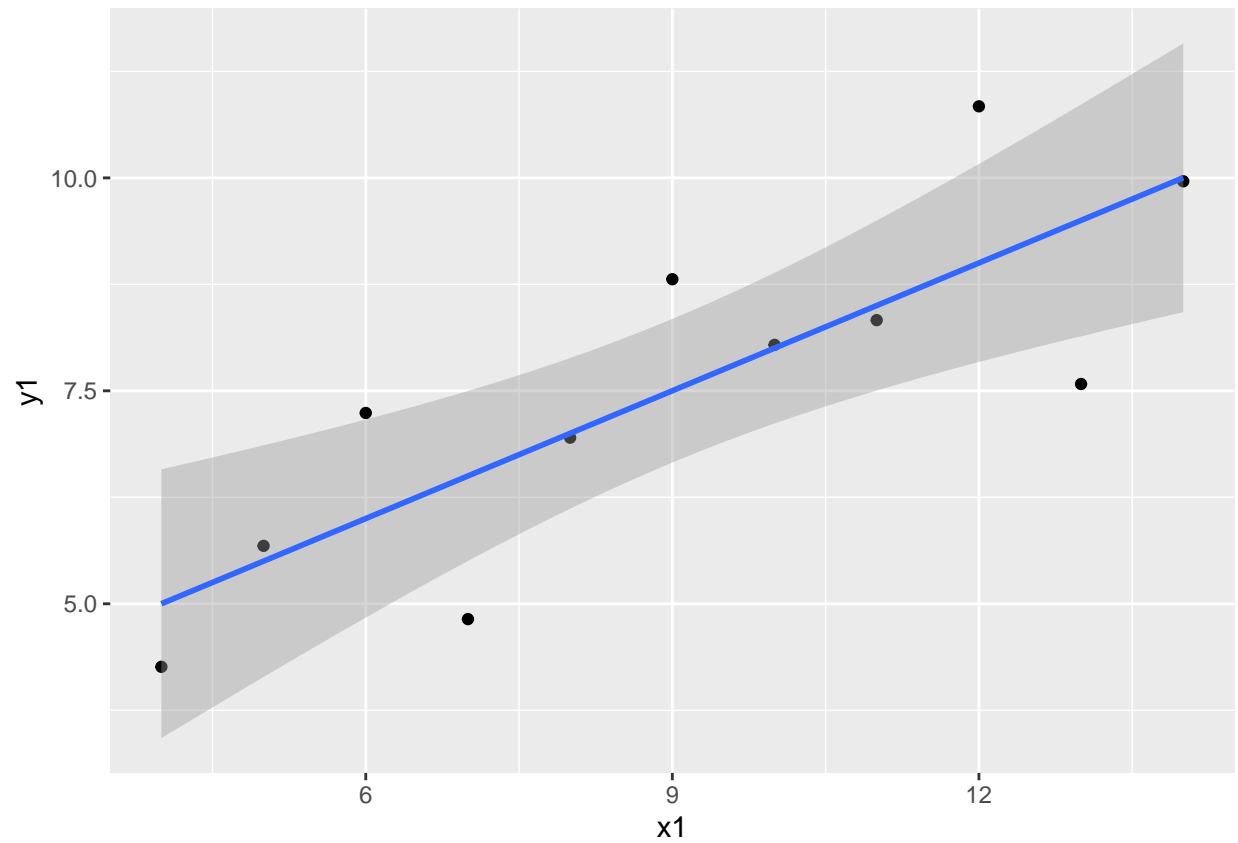
```
##
## x3                                0.500***
##                                (0.118)
##
## x4                                0.500***
##                                (0.118)
##
## Constant                3.000**   3.001**   3.002**   3.002**
##                        (1.125)   (1.125)   (1.124)   (1.124)
##
## -----
## Observations                11         11         11         11
## R2                        0.667        0.666        0.666        0.667
## Adjusted R2                0.629        0.629        0.629        0.630
## Residual Std. Error (df = 9) 1.237        1.237        1.236        1.236
## F Statistic (df = 1; 9)    17.990*** 17.966*** 17.972*** 18.003***
## =====
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

This output looks very strange at first glance, given that each relationship has identical coefficients, but it actually ends up making sense once you look at the graphs produced below. Basically, the coefficients indicate that for each unit that  $x_1$  increases,  $y_1$  increases by around 0.5, and so on and so forth for  $x_2/y_2$ ,  $x_3/y_3$ , and  $x_4/y_4$ .

**1c) Using `ggplot()`, produce scatterplots of each  $y$  on its corresponding  $x$  and add both a linear regression line.**

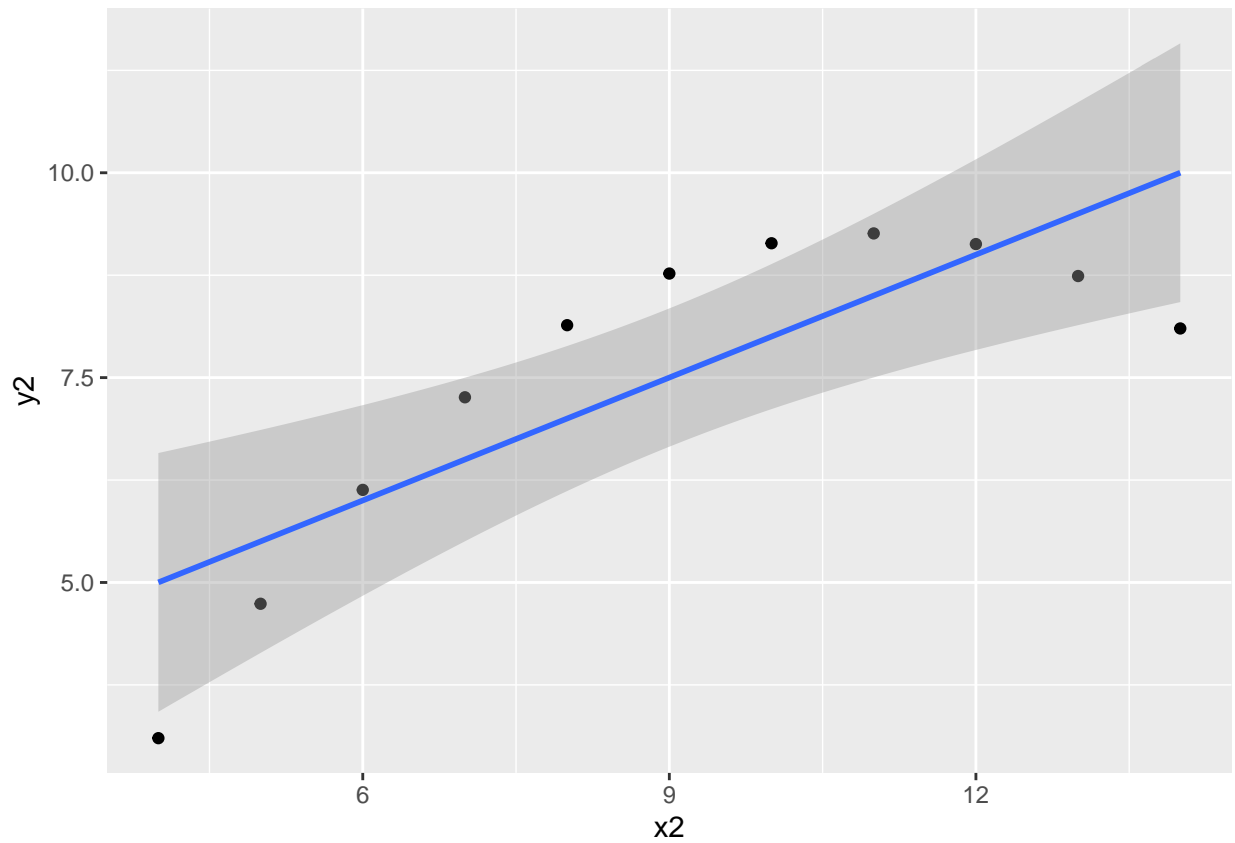
```
ggplot(dataset, aes(x = x1, y = y1)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



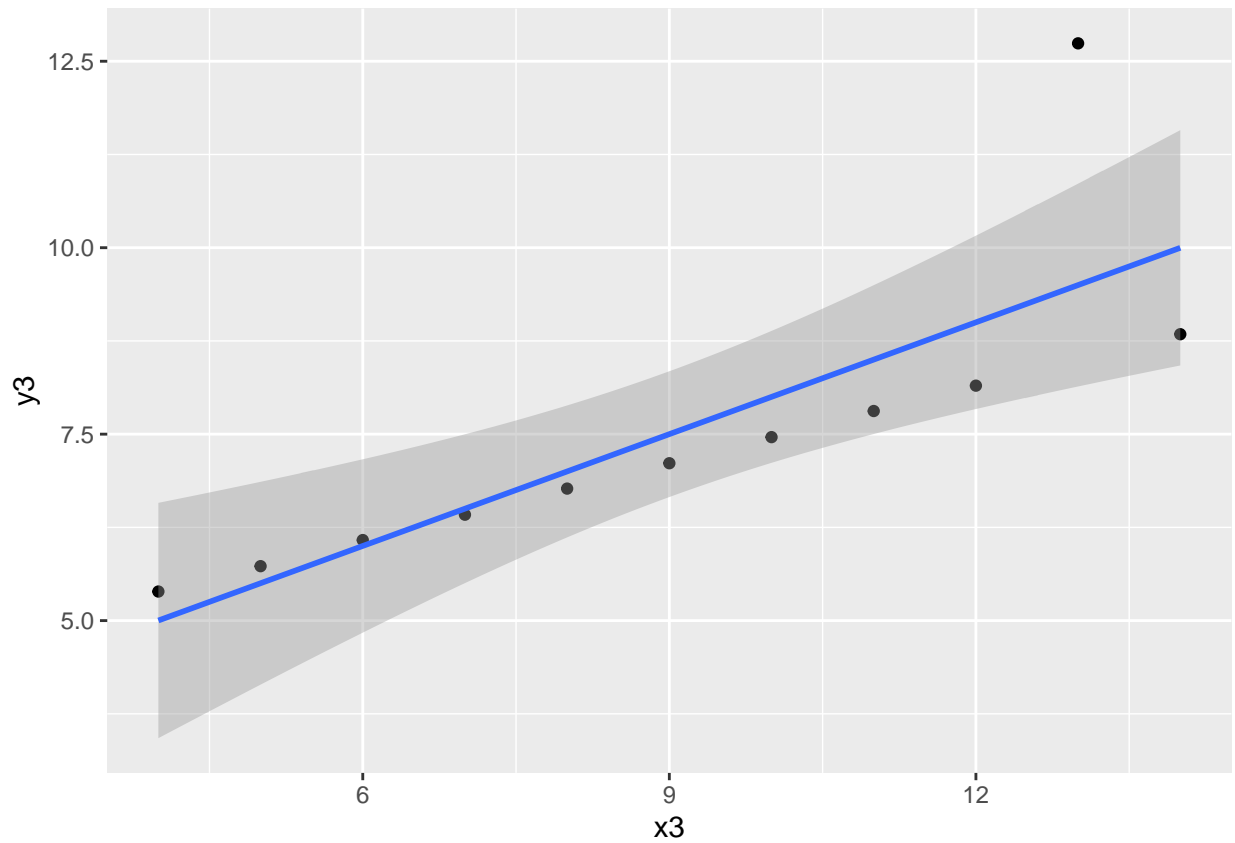
```
ggplot(dataset, aes(x = x2, y = y2)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



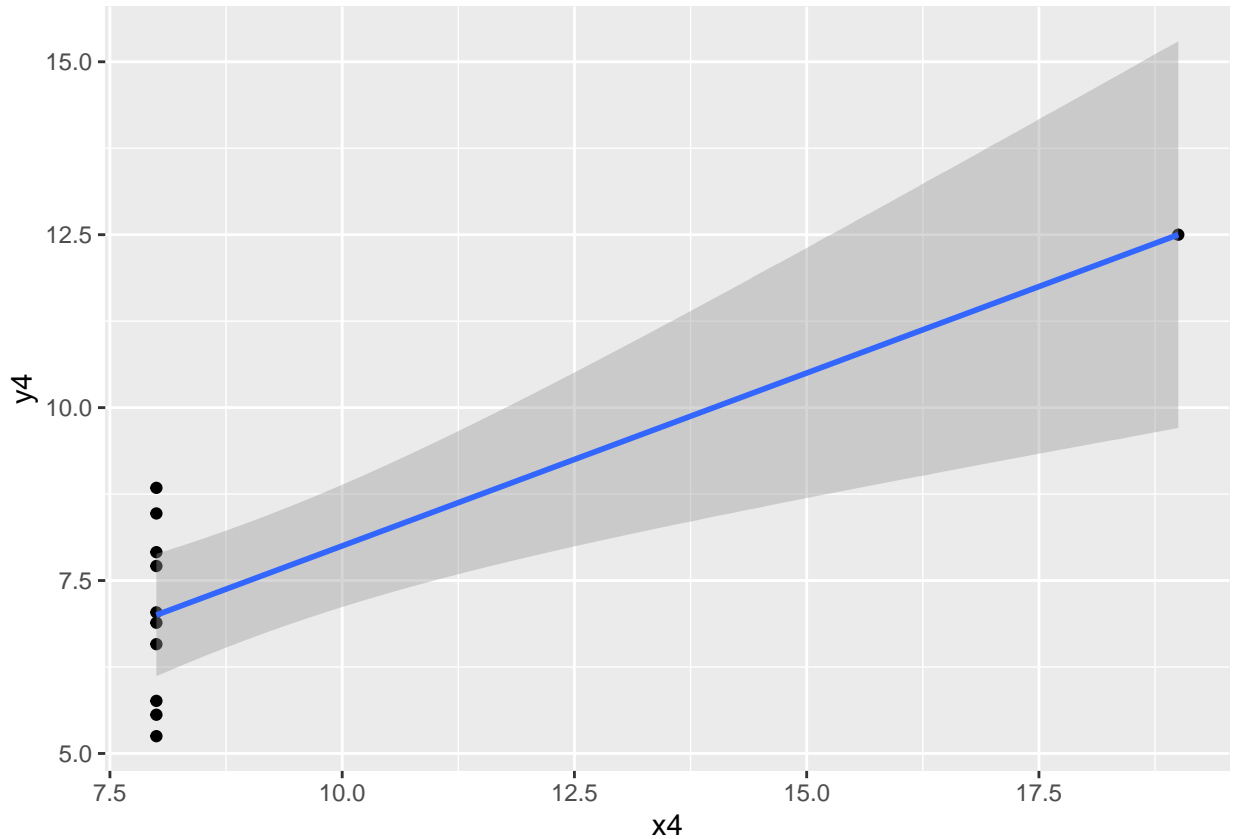
```
ggplot(dataset, aes(x = x3, y = y3)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



```
ggplot(dataset, aes(x = x4, y = y4)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



**1d) What can you conclude about the use of regression (or other fancy statistical techniques) from this example?**

Well, it's certainly not an infallible technique. These regressions technically all have the same coefficient, but the respective graphs of the variables' relationships look completely different from one another. In order to truly understand the nature of the data and the relationship between two variables, you need to do some additional analysis and remember that even if some relationships have the same coefficients, it doesn't necessarily mean that the relationship truly looks the same.

## Question 2 Background:

*Let's explore some data on expenditures for households in Pakistan. The data in `pakistan.dta` on the course website come from a nationally representative sample of households in Pakistan in 1995. Each of the  $n$  observations pertainst to a household  $i$ . Variable definitions are as follows (all monetary figures are in 1995 rupees):*

- `totexp`: Total monthly household expenditures (in rupees)
- `food`: Total monthly household expenditures on food (in rupees)
- `hhincome`: Total monthly household income (in thousands of rupees)
- `nfkids`: number of female children in the household
- `nmkids`: number of male kids in the household

- nfadult: number of female adults in the household
- nmadult: number of male adults in the household

2a) Engel's Law states that as households become richer, they spend a smaller percent of their total budget on food. Run a regression of the percent of total household expenditure on food (call it  $y$ ) on household income per capita (call it  $x$ ) using the `lm` command (see help file for how to structure your argument). Use `stargazer` to present your results in a nicely formatted table.

```
seconddata <- read_dta("pakistan.dta")

# First, I need to make a new data column with the percentage.

seconddata <- seconddata %>%
  mutate(pctfood = food/totexp * 100) %>%

# I divided my totexp by 1000 to make the results more usable.

  mutate(totexp = totexp/1000)

percentfood <- lm(pctfood ~ totexp, data = seconddata)

stargazer(percentfood, type = "text")
```

```
##
## =====
##                      Dependent variable:
##                      -----
##                      pctfood
## -----
## totexp                -0.715***
##                      (0.032)
##
## Constant              52.360***
##                      (0.290)
##
## -----
## Observations           4,720
## R2                     0.094
## Adjusted R2            0.094
## Residual Std. Error    16.671 (df = 4718)
## F Statistic            490.985*** (df = 1; 4718)
## =====
## Note:                  *p<0.1; **p<0.05; ***p<0.01
```

2b) Produce a scatterplot of  $y$  and  $x$ , with a superimposed linear regression line, using `ggplot()`. Also add a more flexible loess line, in a different color or line style.

```
ggplot(seconddata, aes(x = totexp, y = pctfood)) +
  geom_point() +
  geom_smooth(method = "lm") +
  ylim(0, 100) +
  stat_smooth(method = "loess", color = "red") +
  labs(title = "Relationship Between Total Expenses and Expenses on Food",
       subtitle = "Data from Pakistan, 1995",
       x = "Total Family Expenses (Divided by 1000, in 1995 rupees)",
       y = "Percentage Spent on Food")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```

```
## 'geom_smooth()' using formula 'y ~ x'
```

```
## Warning: Removed 56 rows containing missing values (geom_smooth).
```

```
## Warning: Removed 45 rows containing missing values (geom_smooth).
```



2c) Interpret the slope coefficient. Does Engel's Law seem to hold?

It does, the slope is negative. In this case, as income increases, the percentage spent on food (in general) decreases.



**2d) Interpret the intercept coefficient. Do you have a lot of confidence in this finding? Why or why not?**

First, I'm going to pull up the summary of the regression, so that we can see the intercept coefficient.

```
summary(percentfood)

##
## Call:
## lm(formula = pctfood ~ totexp, data = seconddata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -52.338 -11.470   0.083  11.234 196.738
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.35993    0.28982  180.67  <2e-16 ***
## totexp       -0.71495    0.03227  -22.16  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.67 on 4718 degrees of freedom
## Multiple R-squared:  0.09426,    Adjusted R-squared:  0.09407
## F-statistic:  491 on 1 and 4718 DF,  p-value: < 2.2e-16
```

I have no confidence in this finding, because it does not make logical sense. Essentially, the intercept tells us that when a household has an income of 0, they can be expected to spend 50% of their income on food. Needless to say, this makes no logical sense. We should not look at the intercept value as a key part of the model here.

### Question 3 Background:

*We're going to look at the relationship between age and income, using the dataset `vote.csv` from the course website.*

**3a) First, run a simple regression of income ( $y$ ) on age ( $x$ ) using the `lm` command. What do the results tell you? Briefly, how would you interpret the intercept and slope coefficients?**

```
vote <- read.csv("vote.csv")

aregression <- lm(income ~ age, data = vote)
summary(aregression)

##
## Call:
## lm(formula = income ~ age, data = vote)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -10.0149 -2.7149 0.6514 3.2151 6.3676
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.988661  0.294633  50.872  <2e-16 ***
## age         -0.051250  0.005637  -9.091  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.813 on 1498 degrees of freedom
## Multiple R-squared:  0.05229,    Adjusted R-squared:  0.05166
## F-statistic: 82.65 on 1 and 1498 DF,  p-value: < 2.2e-16
```

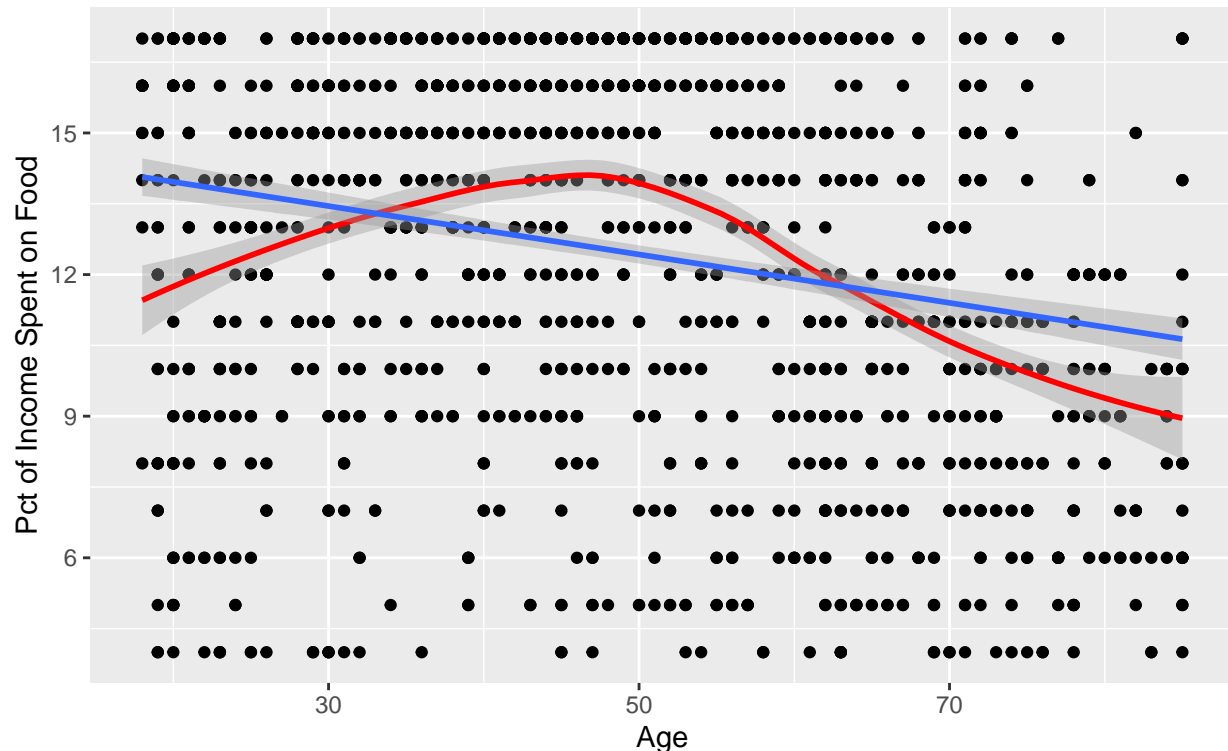
First, it's worth pointing out that the income variable, if I'm interpreting 3c) correctly, actually represents the percentage of a person's income that they spend on food. So we should analyze our results with this factor in mind. The intercept here really isn't helpful, because age is on the x-axis, meaning that this is the value when  $x = 0$ , or when age = 0. Presumably, no baby has an income, unless they are very precocious and Arkansas and South Carolina are ignoring their child labor laws. The slope coefficient (-0.051250) essentially means that for each additional year, people in this dataset tend to spend about .05% less of their income on food, with a high degree of significance.

**3b) Create a scatterplot of the data, where age is on the  $x$ -axis and income is on the  $y$ -axis. Add a line representing the regression line found in part (a), and a more flexibly-fit loess line. What does this tell you about the relationship between age and income?**

```
ggplot(vote, aes(x = age, y = income)) +
  geom_point() +
  geom_smooth(method = "loess", color = "red") +
  geom_smooth(method = "lm") +
  labs(x = "Age",
       y = "Pct of Income Spent on Food",
       title = "Relationship of Age and Percentage of Income Spent on Food",
       subtitle = "Arkansas and South Carolina, 2000")
```

```
## 'geom_smooth()' using formula 'y ~ x'
## 'geom_smooth()' using formula 'y ~ x'
```

## Relationship of Age and Percentage of Income Spent on Food Arkansas and South Carolina, 2000



Including the loess line helps us see a more accurate picture of the trends in this data. Here we see that on average, people don't spend too much of their income on food when they're 20, but that number gradually increases until they hit about 45, after which it steadily drops, most likely (in my opinion) because people are still making a decent amount of money but no longer are supporting/feeding children. The relationship between the age and income variables is definitely not linear. I mean, a linear regression line can be fit to it and we can see a general pattern, but it doesn't show us the "true nature" of the data as well as this loess line does.

**3c) Let's analyze the data using a different way of presenting the results. First, convert the age variable from a continuous variable into a categorical or factor variable, with the following bins: 18 – 35, 36 – 50, 51 – 65, 66 – 85. Produce histograms of the relationship between each age bin and the percent of income spent on food. What do your results suggest?**

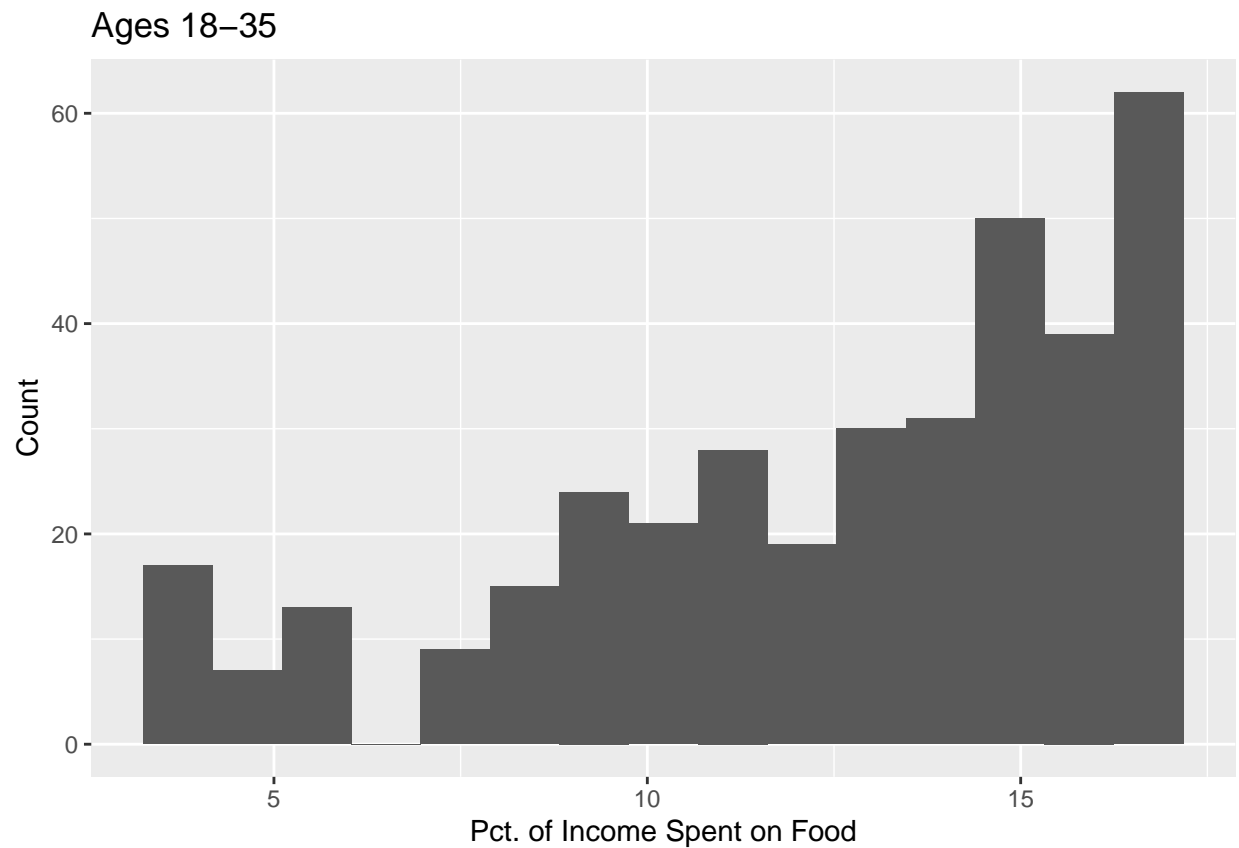
```
# Here, I can use a for loop. First, I have to create a new column.

vote3c <- vote %>%
  mutate(age_categorical = 1)

for(i in 1:1500){
  if(vote3c$age[i] %in% c(66:85)){vote3c$age_categorical[i] <- 4}
  else if(vote3c$age[i] %in% c(51:65)){vote3c$age_categorical[i] <- 3}
  else if(vote3c$age[i] %in% c(36:50)){vote3c$age_categorical[i] <- 2}
  else{vote3c$age_categorical[i] <- 1}
}
```

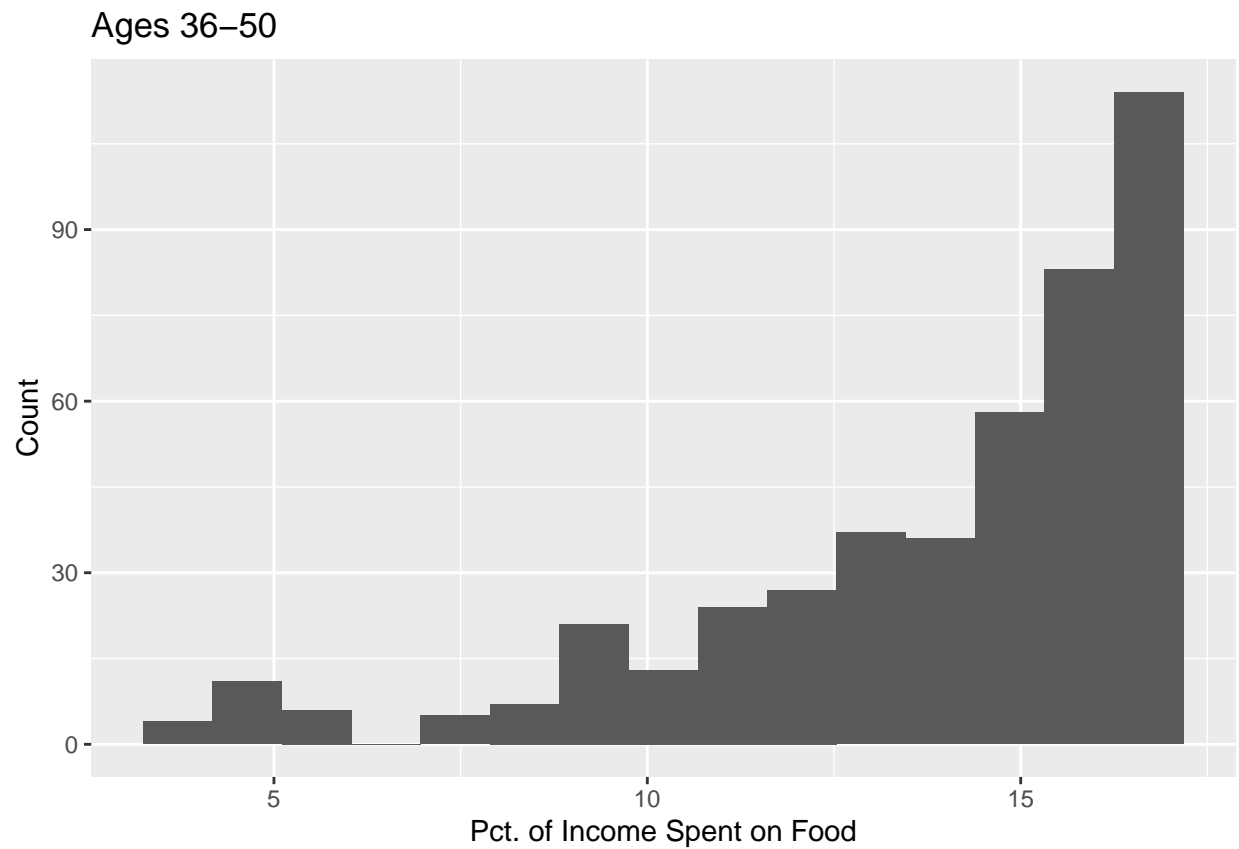
```
first_histdata <- vote3c %>%
  filter(age_categorical == 1)

ggplot(first_histdata, aes(x = income)) +
  geom_histogram(bins = 15) +
  labs(x = "Pct. of Income Spent on Food",
       y = "Count",
       title = "Ages 18-35")
```

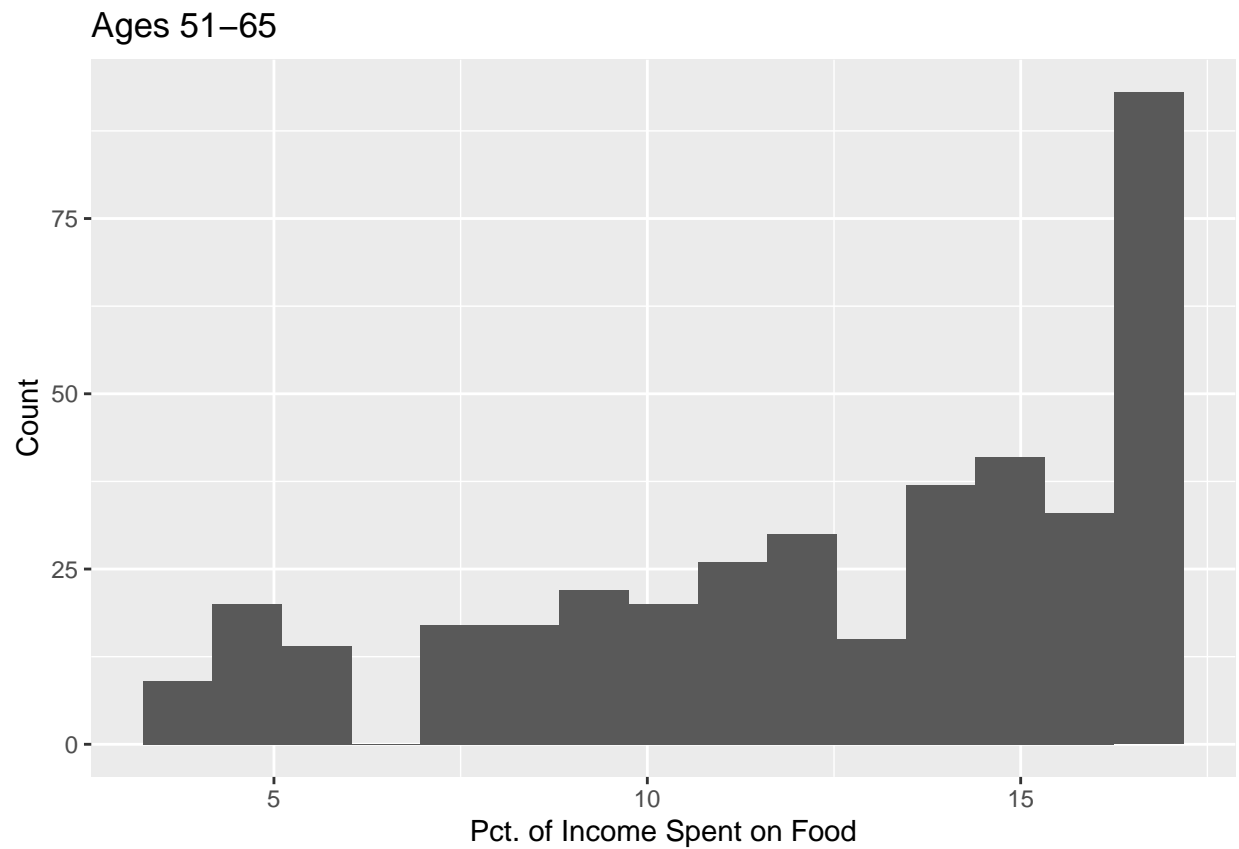


```
second_histdata <- vote3c %>%
  filter(age_categorical == 2)

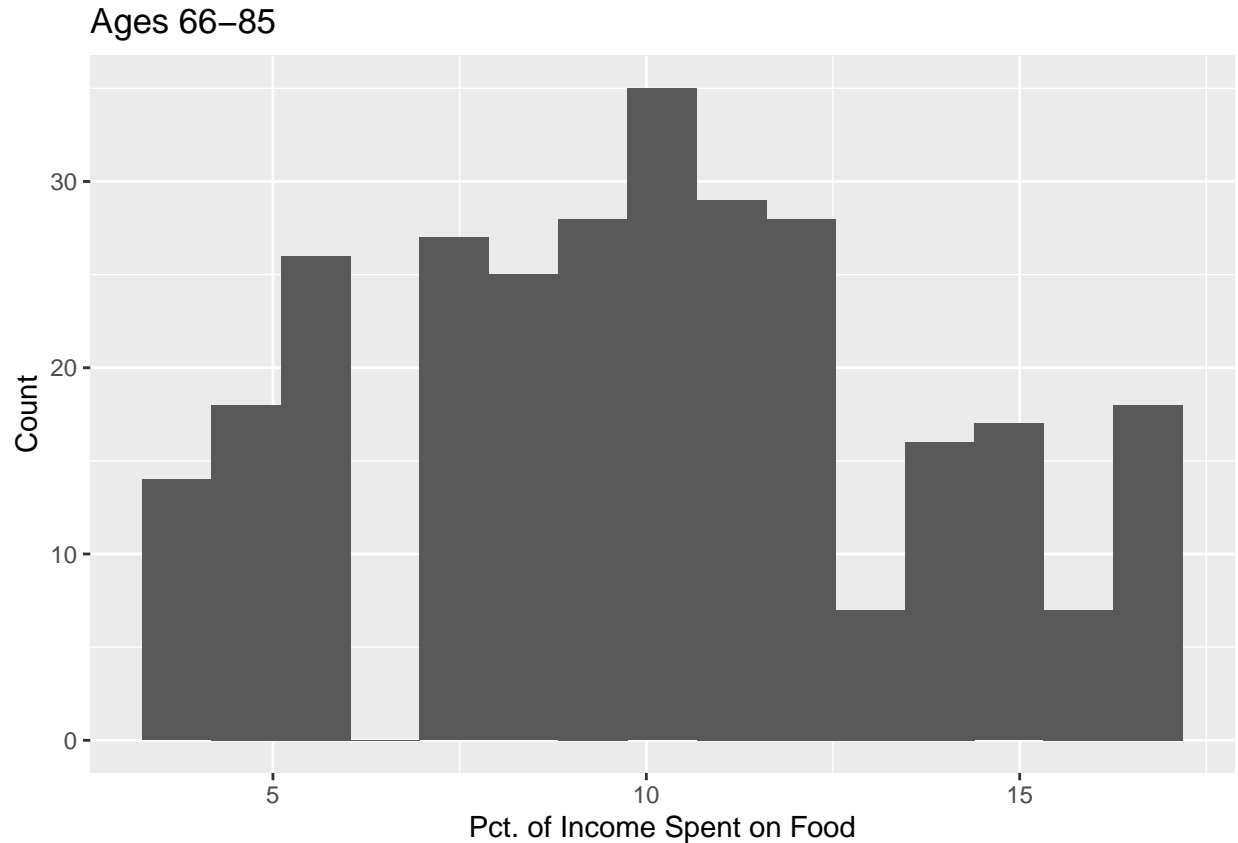
ggplot(second_histdata, aes(x = income)) +
  geom_histogram(bins = 15) +
  labs(x = "Pct. of Income Spent on Food",
       y = "Count",
       title = "Ages 36-50")
```



```
third_histdata <- vote3c %>%  
  filter(age_categorical == 3)  
  
ggplot(third_histdata, aes(x = income)) +  
  geom_histogram(bins = 15) +  
  labs(x = "Pct. of Income Spent on Food",  
       y = "Count",  
       title = "Ages 51-65")
```



```
fourth_histdata <- vote3c %>%  
  filter(age_categorical == 4)  
  
ggplot(fourth_histdata, aes(x = income)) +  
  geom_histogram(bins = 15) +  
  labs(x = "Pct. of Income Spent on Food",  
       y = "Count",  
       title = "Ages 66-85")
```



This is some interesting data. Generally speaking, people from ages 36-65, the graphs look similar – the quantity of people rises gradually per percentage until it gets to a certain point until it spikes upward. The spike is more dramatic for these two age groups than for people between 18-35, although that age group does have a relatively similar-looking graph to the next two age groups. For people from 66-85, though, the graph looks completely different. The pattern looks more centered around 10% than 15% like in the preceding graphs, and there are results to both the left and right of the maximum. It seems that in general, people ages 66 and above have either different spending habits on food in general, different earning habits, different life situations, etc. that leads for the histogram for this population to look different than the others.