Homework Assignment #3

Math 437 - Modern Data Analysis

Due March 20, 2023

```
library(ggplot2)
```

Instructions

You should submit either two or three files:

1. You should write your solutions to the Simulation and Applied Problems in this R Markdown file and submit the (.Rmd) file.
2. You should knit the final solution file to pdf and submit the pdf. If you are having trouble getting code chunks to run, add `eval = FALSE` to the chunks that do not run. If you are having trouble getting R Studio to play nice with your LaTeX distribution, I will begrudgingly accept an HTML file instead.
3. Solutions to the Key Terms and Conceptual Problems can be submitted in a separate Word or pdf file or included in the same files as your solutions to the Simulation and Applied Problems.

This homework assignment is worth a total of **45 points**.

Key Terms (7 pts)

Read Chapter 3 of Introduction to Statistical Learning, Second Edition. Based on your reading, answer the following questions.

1. What is the difference between the *population regression line* and the *least-squares regression line*? When do we use the notation β_1 vs. $\hat{\beta}_1$?

Answer: The difference between the population regression line and the least squares regression line is that the population is the theoretical best fitting line for a model and uses the β_1 notation to be the true slope of the population. In practice, this value is unknown and thus we give it our best estimate which we denote $\hat{\beta}_1$.

2. What is a *residual*? Why are the residuals important in finding the $\hat{\beta}_j$?

Answer: A residual is the vertical distance between the observed value and the regression line which represents the difference between an observed response value and its corresponding predicted value. The sum of those squared residuals is then minimized to determine the least squares regression line which in turn helps us find $\hat{\beta}_j$.

3. Write a sentence to interpret what β_3 means in the model given by Equation (3.20). (Make sure to use appropriate units!)

Answer:β_3 is the estimate for how much we expect sales to be affected for every \$1,000 increase in spending on newspaper advertising, if we hold TV and radio advertising constant.

4. Give two statistics that measure the fit (or lack thereof) of a multiple linear regression model. For each statistic, as it increases, does the F-statistic for an ANOVA test with $H_0 : \beta_1 = \ldots = \beta_p = 0$ increase or decrease?

Answer: Two statistics that measure fit of a multiple linear regression model include the RSE (residual standard error) and R^2 values. As RSE increases we are seeing a worse fit for a model meaning its ANOVA F-statistic will decrease (RSE measures lack of fit so larger values imply larger lack of fit). As R^2 increases though we are seeing a greater amount of variance in the model be explained by X implying that as R^2 increases the F-statistic will increase as there is now greater evidence that at least one of the betas are not equal to another.

5. Explain how to turn a factor variable into one or more *dummy variables* (indicator variables).

Answer: To turn a factor variable with k classes into a dummy variables we must choose a baseline class which we will assign the value 0 to and create new variables for the remaining k - 1 classes that are assigned the value 1 (or we can use -1 and 1 rather than 0 and 1).

6. Refer to the model whose coefficient estimates are given in Table 3.8. If the *baseline* (reference level) changed to West, what would be the equation of the new least-squares regression plane?

$balance = 518.50 - 6.19(South) + 12.50(East)$

7. What is the difference between a *main effect* and an *interaction effect*?

Answer: The difference between a main effect and an interaction effect is that a main effect is the direct relationship between some variable X_1 and the response, Y, whereas the interaction effect is the relationship between the synergy between two predictors X_1 and X_2 and how their own interaction is related to the response, Y.

8. What kind of patterns should you look for in the *residual plot* to identify non-linear relationships between the predictors and the response? What kind of patterns should you look for to identify heteroscedasticity?

Answer: Residual plots would show systematic patterns such as curves if there was a non-linear relationship between the predictors and response. A fan or funnel shape in the residuals versus fitted values plot would indicate heteroscedasticity.

9. With what kind of data would you expect to see correlation of error terms when fitting a linear regression model? Why?

It is often time series data in which we would see correlation of errors when fitting a linear regression model. This is because there is often correlated data since our collection is from the same place just at different time points.

10. Give a rule of thumb for guessing that a point has an *outlier* residual. Give a rule of thumb for guessing that a point is a *high leverage* point.

Answer: A general rule of thumb for determining whether or not a point is an outlier is calculating their studentized residual and if the $|studentized.residual| > 3$ the point is a potential outlier. For high leverage points, if the leverage statistic is far greater than $\frac{p+1}{n}$ this is indication of a high leverage point.

11. Explain why creating a scatterplot matrix of the response variable y and all predictor variables is insufficient for detecting outliers and high leverage points.

A scatterplot matrix of the response variable and all predictor variables is insufficient because for it is sometimes difficult to determine if a point is an outlier simply by looking at the plot and for high leverage points it is possible to have an observation that is within the normal range of values for individual predictors but unusual for the full set of predictors and when there is more than two predictors, all the dimesions of the data cannot be plotted simultaneously.

12. Can *collinearity* be suspected based on inspecting the scatterplot matrix and correlation matrix? What about *multicollinearity*? Explain why/why not.

Answer: Collinearity can be suspected based on the scatterplot and correlation matrix. Multicollinearity on the other hand is sometimes not as easily sniffed out because it is possible for collinearity to exist between 3+

variables even if there isn't high correlation between any two pairs, and in that situation we would turn to the variance inflation factor.

13. What is the *variance inflation factor*? When/why do we use it? How is it computed?

The variance inflation factor of β_j is a measure of how much the variance of β_j is affected by multicollinearity. We use it to determine whether or not multicollinearity exists and how significant it is for each predictor. It is computed by calculating the ratio of variance from the model fitted about the full model to the variance from the model fitted just around $\hat{\beta}_j$.

14. Briefly explain how *k-nearest neighbors regression* works. What are its advantages and disadvantages compared to linear regression?

KNN regression works by identifying the K training observations closest to a given prediction point, x_0 and estimating $f(x_0)$ by averaging all the responses of those K observations. The advantages of KNN regression are that it is a non-parametric method and therefore works better than linear regression when there is a non-linear relationship between X and Y. However, when there is more than one predictor and a small number of observations per predictor, the "curse of dimensionality" begins to take effect since as the number of predictors, hence dimensions, grows, the "neighbors" of x_0 become increasingly spaced out leading to poor estimates of the response. Linear regression is also more easily interpreted.

Conceptual Problems

Conceptual Problem 1 (3 pts)

Textbook Exercise 3.7.3.

$ starting salary = 50 + 20(GPA)+.07(IQ) +35(College?) + .01(GPA:IQ) - 10(GPA:College?) $

(a): iv) For a fixed GPA and IQ someone who went to college will have a higher predicted starting salary than high school graduates provided their GPA is high enough. This is true and not ii since the interaction term decreases the overall starting salary, so long as the interaction between GPA and our factor is a value less than 3.5.

(b) $ starting salary = 50 + 20(4.0)+.07(110) +35(1) + .01(4.0()110) - 10(4.0)(1) = 137.1$ The predicted starting salary would be 137, 100 dollars.

(c) Just because we have a small interaction term this does not mean that the interaction between our variables is insignificant. The coefficient is based on the interaction as well as the standard deviations of both x and y. Thus the scaling of GPA and IQ can influence the size of the coefficient regardless of whether or not the coefficient is significant making the entire statement false.

Conceptual Problem 2 (4 pts total)

Part a (3 pts)

Consider multiple linear regression without an intercept. Suppose that x_1 and x_2 have correlation $r_{x_1x_2}$. Find a closed-form expression for $\hat{\beta}$, a vector containing the values of β_1 and β_2 when

$$RSS = \sum_{i=1}^{n}(y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2$$

is minimized. Express your final expression in terms of three standard deviations (s_{x_1}, s_{x_2}, s_y) and three correlations ($r_{x_1x_2}$, r_{x_1y}, and r_{x_2y}).

HINT 1: Remember that a function cannot be minimized unless the gradient (vector of partial derivatives) is the 0 vector. You do not need to prove that the point at which the gradient is 0 is a local minimum instead of a local maximum or saddle point.

HINT 2: It may be easiest to do a bunch of algebra to get the β_1 and β_2 coefficients outside the sum before taking partial derivatives with respect to β_j.

HINT 3: Because we are doing linear regression without an intercept, the formulas for sample variance and sample correlation simplify to:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2$$

and

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i}{s_x}\right)\left(\frac{y_i}{s_y}\right)$$

$$\hat{\beta} = \left(\frac{s_y(r_{x_1y} - r_{x_2y}r_{x_1x_2})}{s_{x_1}(1-r_{x_1x_2}^2)}, \frac{s_y(r_{x_2y} - r_{x_1y}r_{x_1x_2})}{s_{x_2}(1-r_{x_1x_2}^2)}\right)$$

Part b (1 pt)

In Math 338 you likely learned that in simple linear regression, the slope estimate $b_1 = \hat{\beta}_1 = r\frac{s_y}{s_x}$. Show that for $j = 1, 2$, if x_1 and x_2 are uncorrelated, then $\hat{\beta}_j = r_{x_jy}\frac{s_y}{s_{x_j}}$ $(j = 1, 2)$.

If x_1 and x_2 are uncorrelated, then $r_{x_1x_2} = 0$ so when we make this substitution in our $\hat{\beta}$ that we derived above, we obtain $\beta_1 = r_{x_1y}\frac{s_y}{s_{x_1}}$ and $\beta_2 = r_{x_2y}\frac{s_y}{s_{x_2}}$, which generalizes to $\hat{\beta}_j = r_{x_jy}\frac{s_y}{s_{x_j}}$ $(j = 1, 2)$. The actual algebra is in the attached pdf onenote file.

For 1 pt extra credit: Show that under the additional assumption that $\hat{\beta}_1 \neq 0$ and $\hat{\beta}_2 \neq 0$, the converse also holds; that is, if $\hat{\beta}_j = r_{x_jy}\frac{s_y}{s_{x_j}} \neq 0$ $(j = 1, 2)$, then $r_{x_1x_2} = 0$.

Simulation Problems

Simulation Problem 1 (5 pts)

Textbook Exercise 3.7.13 parts (a)-(f).

(a)

```
set.seed(1)
x <- rnorm(100,0,1)
```

(b)

```
eps <- rnorm(100, 0, 0.25)
```

(c)

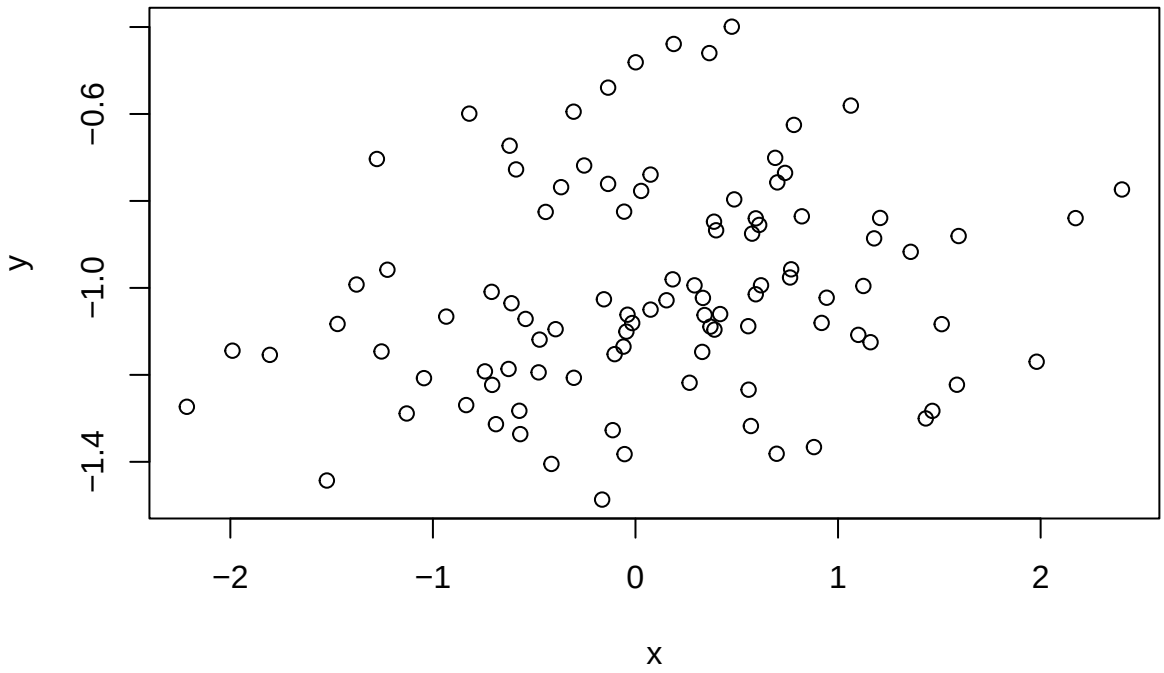
```
y = -1 + .05*x + eps
length(y)
```

```
## [1] 100
```

y has a length of 100 with $\beta_0 = -1$ and $\beta_1 = .05$

(d)

```
plot(x,y)
```



The relationship between the two variables is not too strong, this is due to the impact our variable eps had on the y.

(e and f)

```
required_data <- as.data.frame(x,y)
```

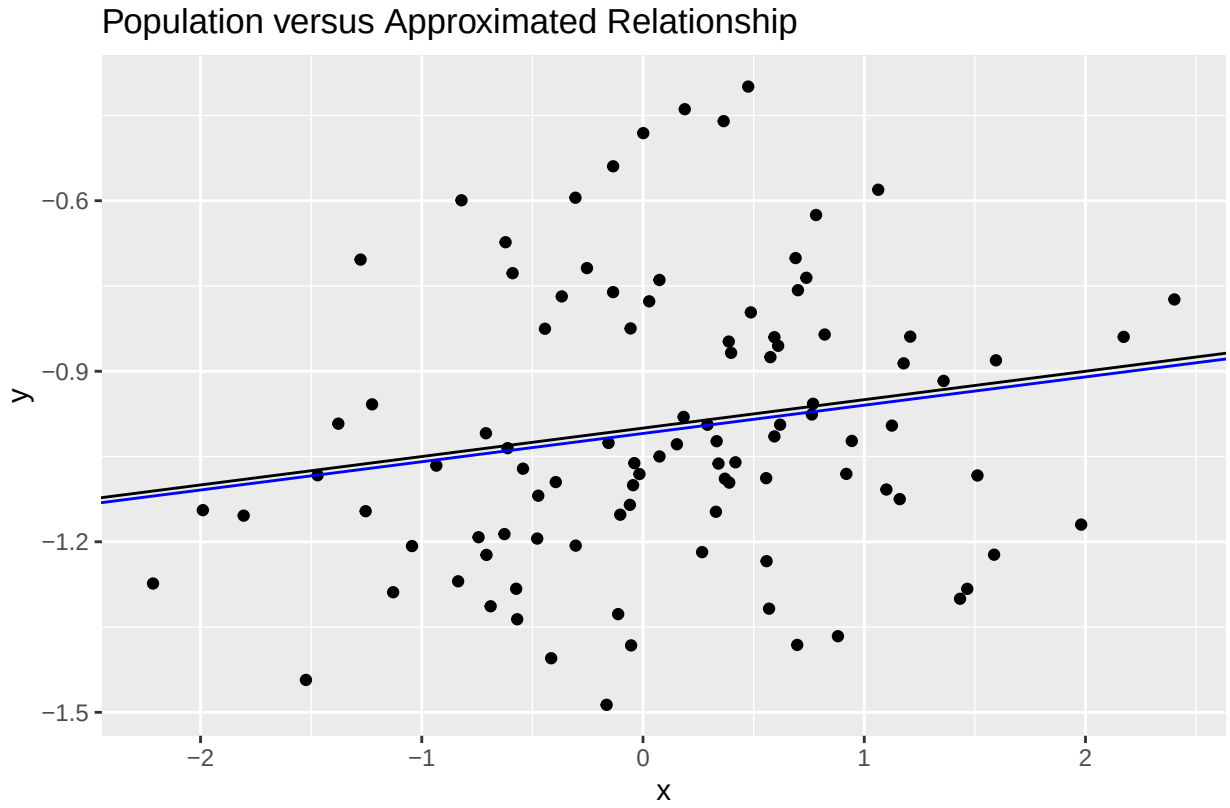
```
## Warning in as.data.frame.numeric(x, y): 'row.names' is not a character vector
## of length 100 -- omitting it. Will be an error!
```

```
our_model <- lm(y ~ x, data = required_data)
summary(our_model)
```

```
##
## Call:
## lm(formula = y ~ x, data = required_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.46921 -0.15344 -0.03487  0.13485  0.58654
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.00942    0.02425 -41.631   <2e-16 ***
## x            0.04973    0.02693   1.847   0.0678 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2407 on 98 degrees of freedom
## Multiple R-squared:  0.03363,    Adjusted R-squared:  0.02377
## F-statistic:  3.41 on 1 and 98 DF,  p-value: 0.06781
```

```
ggplot(data = required_data, aes(x=x,y=y))+
  geom_point()+
  geom_abline(slope=0.04973, intercept = -1.00942, color = "blue") +
```

```
  geom_abline(slope = 0.05, intercept = -1) +
  labs(caption = "The black line represents the true, 'simulated' values while the blue line represents
       title = "Population versus Approximated Relationship")
```



The black line represents the true, 'simulated' values while the blue line represents the linear model's approximations

(Note I was not sure how to make a custom legend in ggplot)

our $\hat{\beta}_0 = -1.00942$ and our $\hat{\beta}_1 = 0.04973$. These results are very close to the true $\beta_0 = -1$ and $\beta_1 = 0.05$ so our model did well in that regard despite the included variance. It should be noted though that our model does not do that well at capturing the data with both a low R-squared value as well as p-value larger than 0.05.

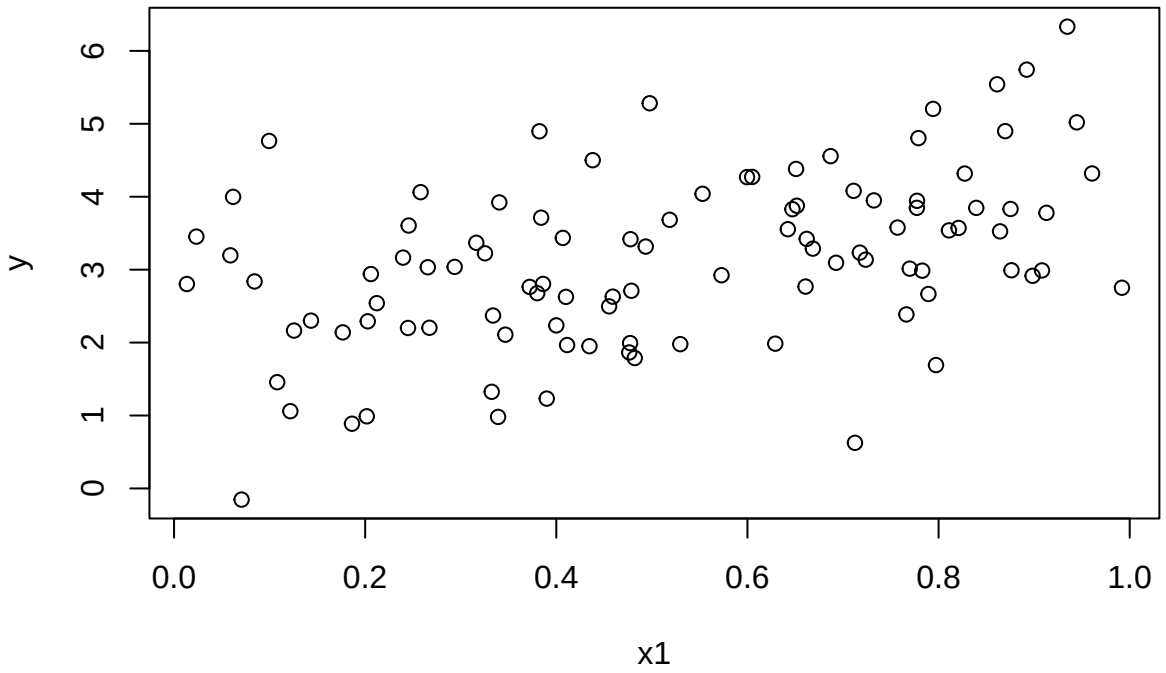
Simulation Problem 2 (4 pts total)

Part a (Code: 1 pt; Explanation: 2 pts)

Textbook Exercise 3.7.14 parts (a)-(c).

(a)

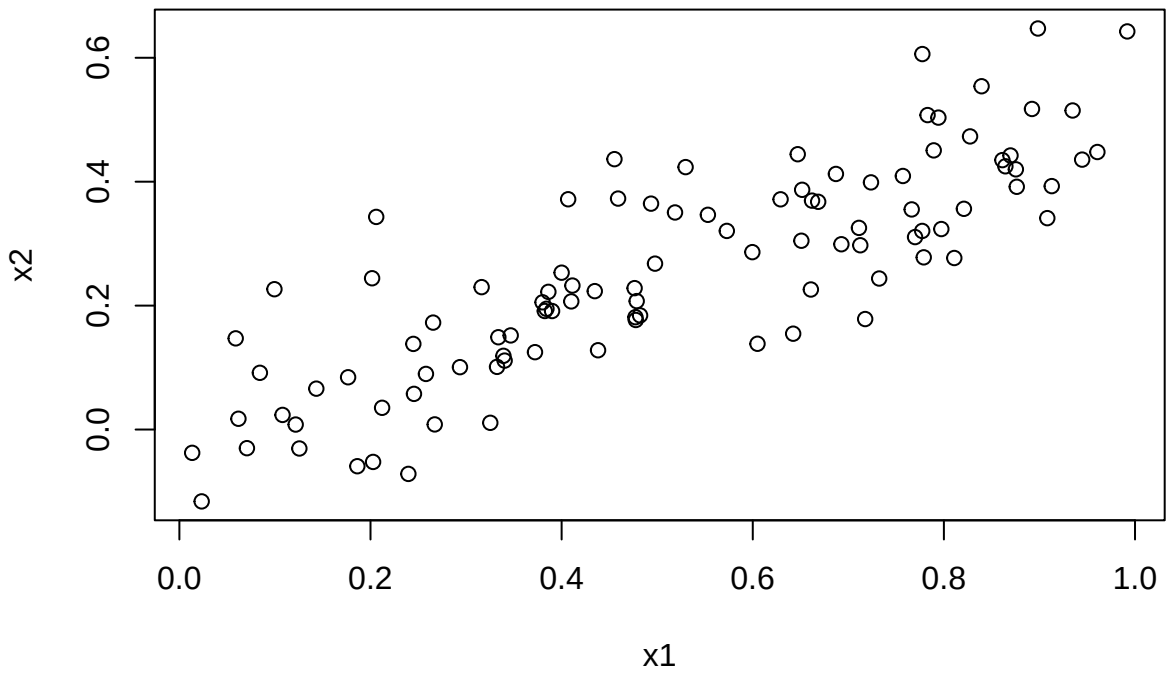
```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 *x1 + rnorm(100) / 10
y <- 2 + 2*x1 + 0.3*x2 + rnorm(100)

plot(x1,y)
```



$Y = (2 + rnorm(100)) + 2X_1 + 0.3X_2)$ where the regression coefficients $\beta_0 = 4.277879$, $\beta_1 = 2$ and $\beta_2 = 0.3$.

(b)

```
plot(x1,x2)
```




```
cor(x1,x2)
```

```
## [1] 0.8351212
```

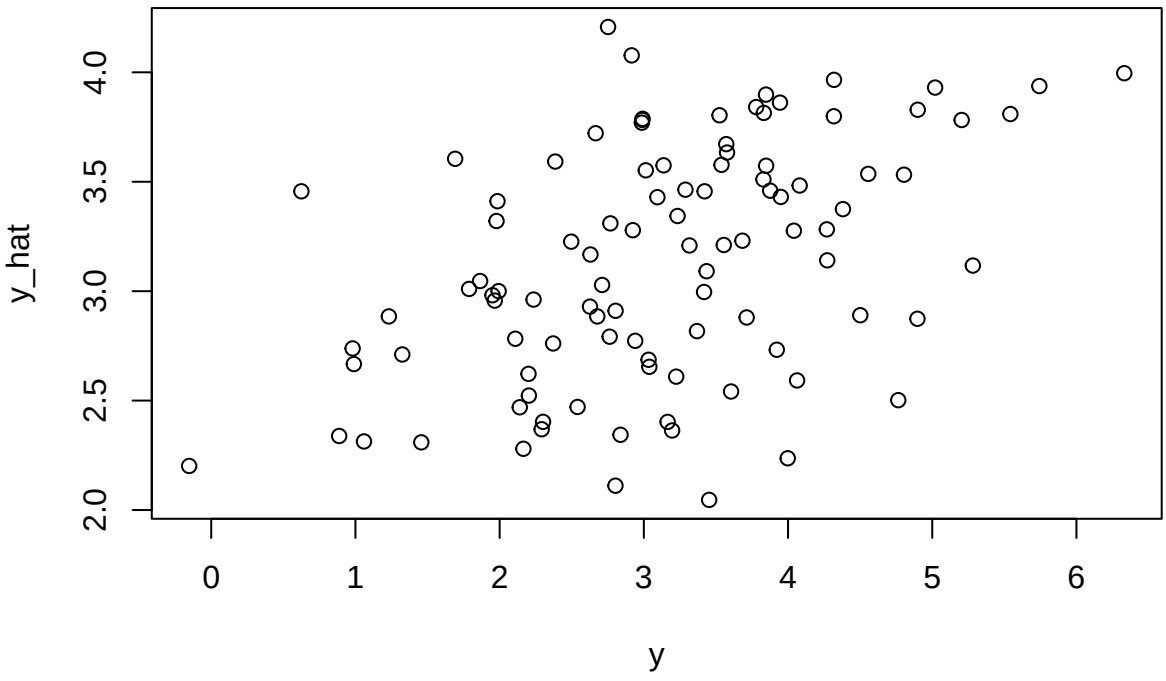
The correlation between x1 and x2 is 0.8351212.

(c)

```r
data <- data.frame(y,x1,x2)

linear_model <- lm(y ~ x1 + x2, data = data)
summary(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1            1.4396     0.7212   1.996   0.0487 *
## x2            1.0097     1.1337   0.891   0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

```r
y_hat = 2.13 + 1.44*x1 + 1.01*x2

plot(y, y_hat)
```



$\hat{y} = 2.13 + 1.44(x_1) + 1.01(x_2)$

The results show $\hat{\beta}_0 = 2.13$ which is lower than our original intercept, $\hat{\beta}_1 = 1.44$ which is 0.6 less than our original β_1, and $\hat{\beta}_2 = 1.01$ which is 0.71 higher than our original β_2. At an $\alpha = 0.05$ significance level you could reject the null hypothesis $H_0 : \beta_1 = 0$ but you could not reject the null hypothesis $H_0 : \beta_2 = 0$.

Part b (Code: 0.5 pts; Explanation: 0.5 pts)

Center `x1`, `x2`, and `y` to have mean 0 (the slopes should remain the same after centering the data). What are the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize MSE/RSS? Compare these values to the true β_1 and β_2 as well as the $\hat{\beta}_1$ and $\hat{\beta}_2$ obtained in 3.7.14c.

HINT: you should have derived a formula in Conceptual Problem 2, so all you have to do is find the right values based on your simulated data and plug them into the formula.

```
data <- data.frame(scale(data, scale = FALSE))

sy <- sd(data$y)
sx1 <- sd(data$x1)
sx2 <- sd(data$x2)

rx1y <- cor(x1, y)
rx2y <- cor(x2, y)
rx1x2 <- cor(x1,x2)

(betahat1 <- sy*(rx1y - rx2y*rx1x2)/(sx1*(1 - rx1x2^2)))
```

```
## [1] 1.439555
```

```
(betahat2 <- sy*(rx2y - rx1y*rx1x2)/(sx2*(1 - rx1x2^2)))
```

```
## [1] 1.009674
```

The results for the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ that minimize RSS are the same values given by the linear model in part c above, so their comparison holds and $\hat{\beta}_1$ is lower than our original β_1, and $\hat{\beta}_2$ is higher than our original β_2.

Applied Problems

Applied Problem 1 (7 pts total)

The code chunk below creates a `mpg_new` dataset from the `mpg` dataset in the ggplot2 package. Using this new dataset:

```
library(ggplot2)
library(dplyr)
mpg_new <- mpg %>%
  filter(year == 2008, class == "compact")
```

Part a (Code: 1 pt; Explanation: 1 pt)

Fit a simple linear regression model predicting highway gas mileage (`hwy`) from the fuel type `fl` (r = regular gas, p = premium gas). Interpret the `flr` coefficient in the summary. Is this coefficient significant (at the 5% significance level)?

```
simple_LR_model <- lm(hwy ~ fl, data = mpg_new)
summary(simple_LR_model)
```

```
##
## Call:
## lm(formula = hwy ~ fl, data = mpg_new)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -3.3333 -2.6154 -0.1154  1.3846  6.6667
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.6154     0.7773  35.529   <2e-16 ***
## flr           2.7179     1.2152   2.237   0.0369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.802 on 20 degrees of freedom
## Multiple R-squared:  0.2001, Adjusted R-squared:  0.1601
## F-statistic: 5.002 on 1 and 20 DF,  p-value: 0.03687
```

the "flr" variable implies that having regular gas will increase our highway mpg by 2.7179 gallons and with a p-value of .0369 is considered significant.

Part b (Code: 1 pt; Explanation: 1 pt)

Fit a multiple linear regression model predicting highway gas mileage from the fuel type and the city gas mileage. Interpret the `flr` coefficient in the summary. Is this coefficient significant (at the 5% significance level)?

```
multiple_LR_model <- lm(hwy ~ fl+ cty, data = mpg_new)
summary(multiple_LR_model)
```

```
##
## Call:
## lm(formula = hwy ~ fl + cty, data = mpg_new)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.98631 -0.80535  0.03149  0.75531  3.01369
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.6983     2.4854   3.500   0.0024 **
## flr           0.5254     0.6764   0.777   0.4469
## cty           0.9644     0.1251   7.708  2.9e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.415 on 19 degrees of freedom
## Multiple R-squared:  0.8062, Adjusted R-squared:  0.7858
## F-statistic: 39.51 on 2 and 19 DF,  p-value: 1.7e-07
```

Having regular gas will increase our highway mpg by .5254 but with a p-value of .4469 cannot be considered significant at $\alpha = 0.05$

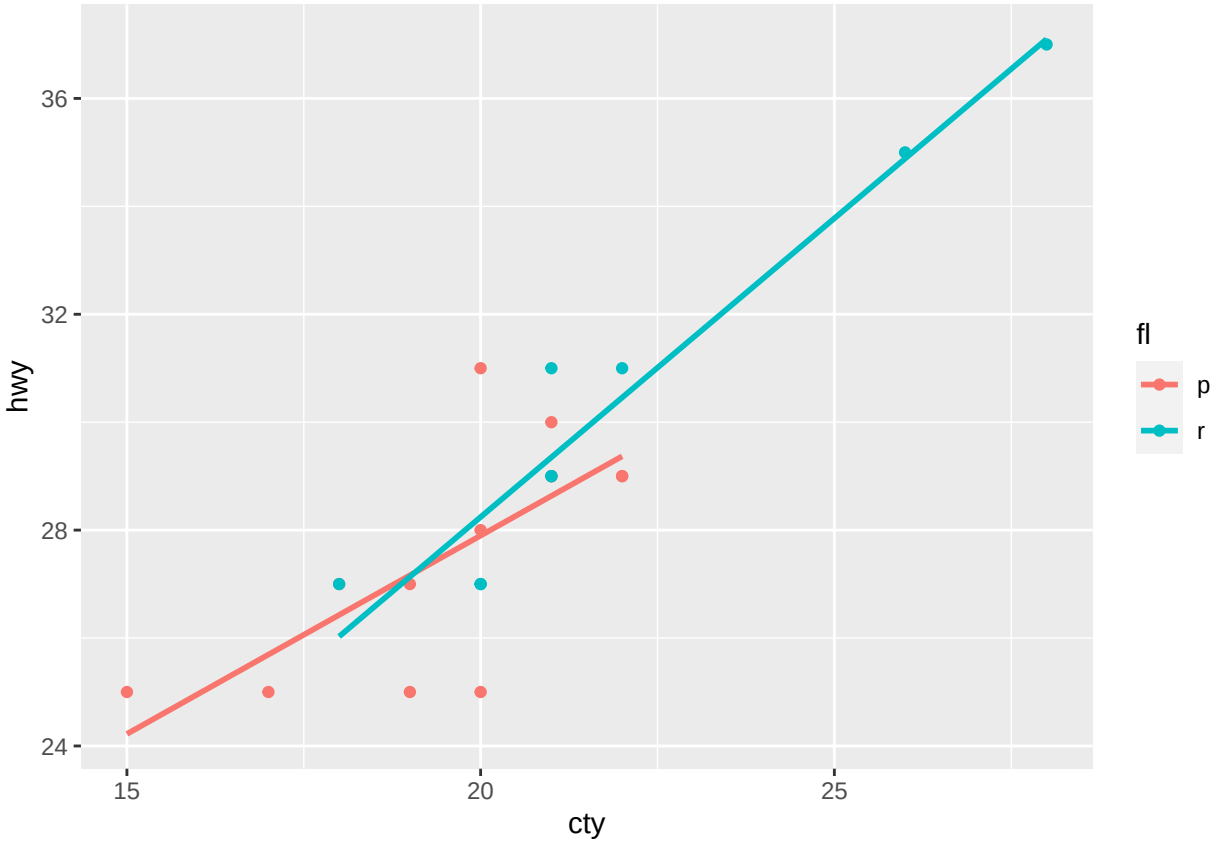
Part c (Code: 1 pt; Explanation: 1 pt)

Using the code in the Multiple Linear Regression Examples file as a guide, create an interaction plot showing city gas mileage on the x-axis and using `fl` as the `trace.factor`. What does the interaction plot reveal about the relationship between `cty` and `fl` that helps explain your results from part (b)?

```
interact_plot <- ggplot(mpg_new, aes(x = cty, y = hwy)) +
  geom_point(aes(color = fl)) +  # color-code points
```

```
  geom_smooth(aes(color = fl), method = "lm", se = FALSE)  # add a regression line for each group
print(interact_plot)
```

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



This interaction plot reveals that the difference between premium and regular gas is substantial as premium gas is maxing out at approximately 29 whereas regular gas climbs all the way up to roughly 37 highway mpg.

Part d (Explanation: 1 pt)

Argue that the error terms in this model are not independent. (HINT: actually view the dataset and see what cars are in it.)

The error terms of this model are not independent as there are 22 entries in the data but only a total of 5 different car variations meaning there are many data points that are looking at the same car thus having dependent errors.

Applied Problem 2 (15 pts total)

Mediation analysis is a technique, often used in social science and health science, that combines causal modeling with multiple linear regression to estimate the effect of a potential confounding variable on the relationship between a predictor and response variable. The steps in mediation analysis are the following:

1. Propose a causal model explicitly naming the predictor variable X, response variable Y, and mediator M. The causal model proposes that the predictor variable causes changes in *both* the mediator and the response, and that the response variable is caused by changes in *both* the predictor and the mediator.
2. Fit a simple linear regression model predicting Y from X, $Y = c_0 + c_1 X$.
3. Fit a simple linear regression model predicting M from X, $M = b_0 + b_1 X$.
4. Fit a multiple linear regression model predicting Y from *both* M and X, $Y = c_0' + c_1' X + c_2' M$.

5. Compute the estimated *direct* effect of X on Y, c_1.
6. Compute the estimated *indirect* effect of X on Y through the mediator M. We can calculate this as either $c_1 - c_1'$ or $(b_1)(c_2')$.
7. Obtain bootstrap confidence intervals for the direct and indirect effects.

Part a (Code: 3 pts; Explanation: 6 pts)

Blake and colleagues (2020) asked 63 Australian women aged 18-40 to record a short video for a mock job interview.

The researchers proposed a causal model in which `sexual_motivation` (mediator M) mediates the relationship between `beautification` (predictor X) and `assertive_behavior` (response Y). These variables can be found in the `assertive_woman` dataset.

Perform an exploratory data analysis on this dataset. You should be able to answer the following questions:

- What does each variable in the dataset mean? How was it measured? (It would be a good idea to read the *Experiment 1 Procedure and materials* section of the linked paper.)

Beautification is a factor variable. Yes implies that they were instructed to bring an outfit that they found attractive and no implied they were requested to bring a comfortable outfit, or one they would wear to see family or friends.

Sexual motivation was a self assessment of how seductive they felt they looked, other terms such as promiscuous and flirtatious were used and the average score was taken.

Assertive behavior was measured based on 2 individual people scoring how assertive these women were on a scale from 1-7

- Is there any missing data? If so, is there an obvious explanation for any of the missingness?

```
library(readr)
neccessary_data <- read_csv("assertive_woman.csv")
```

```
## Rows: 63 Columns: 3
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): beautification
## dbl (2): sexual_motivation, assertive_behavior
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
sum(is.na(neccessary_data$beautification))
```

```
## [1] 0
```

```
sum(is.na(neccessary_data$sexual_motivation))
```

```
## [1] 0
```

```
sum(is.na(neccessary_data$assertive_behavior))
```

```
## [1] 0
```

```
sum((neccessary_data$beautification) == 0)
```

```
## [1] 0
```

```
sum((neccessary_data$sexual_motivation) == 0)
```

```
## [1] 0
```

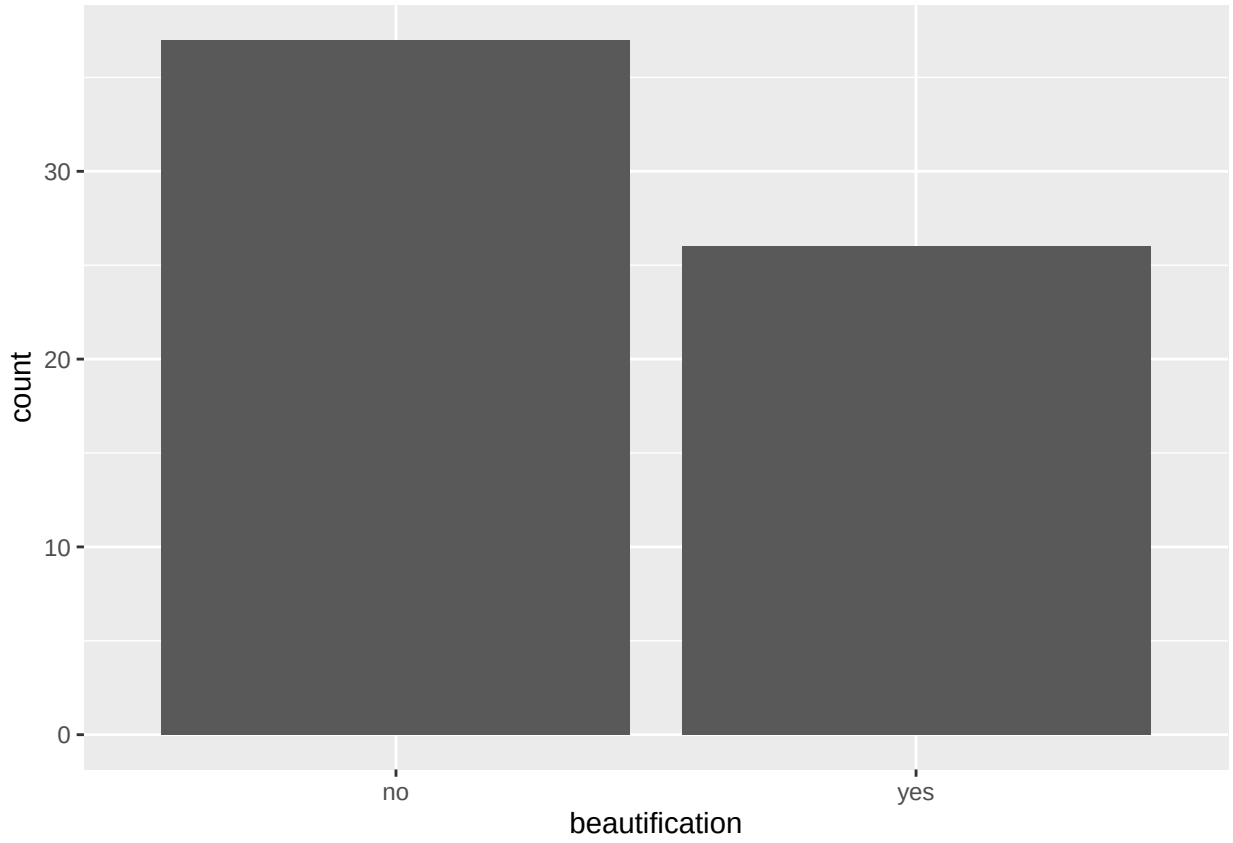
```
sum((neccessary_data$assertive_behavior) == 0)
```

```
## [1] 0
```

There is no missing data in our data frame however the paper did talk about a few pieces of missing data that were removed as a result of incomplete results.

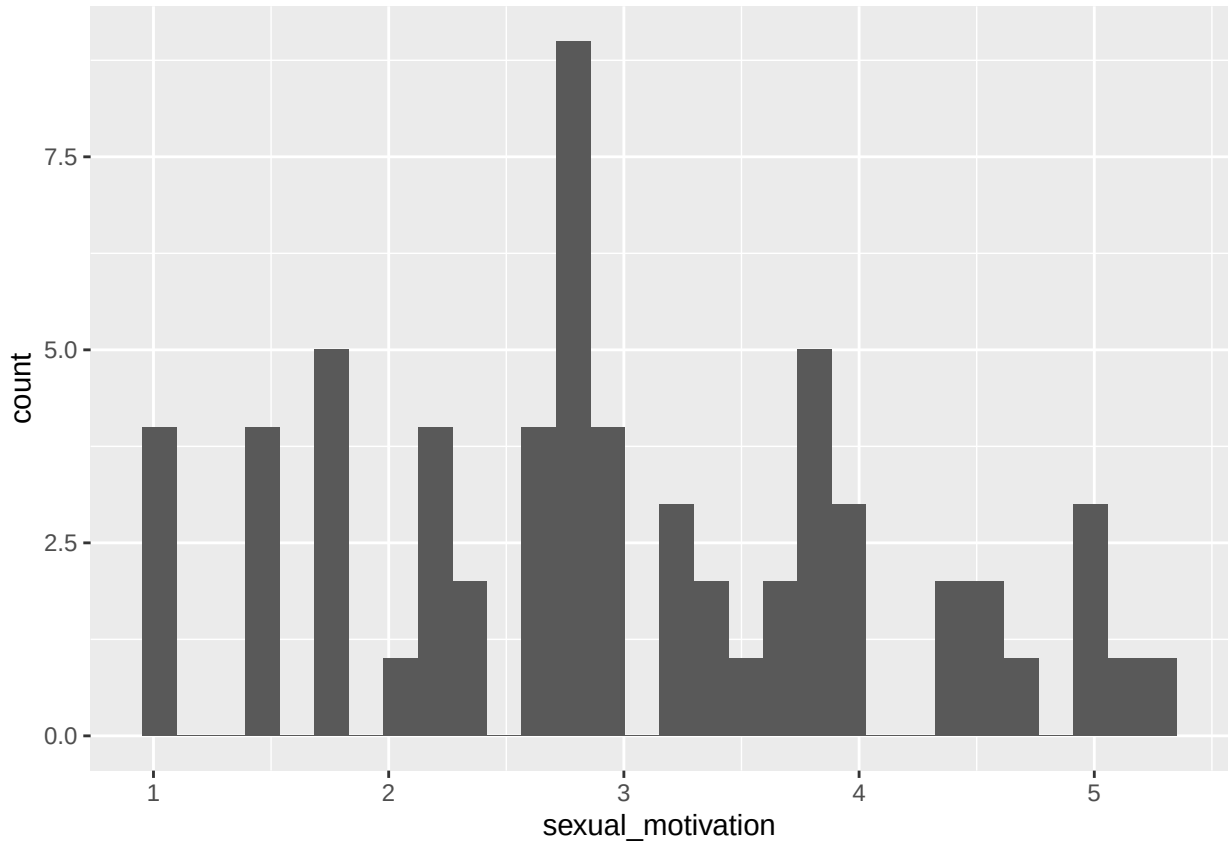
- What is the (univariate) distribution of each variable?

```
neccessary_data %>%
  ggplot(aes(x=beautification)) +
  geom_bar()
```



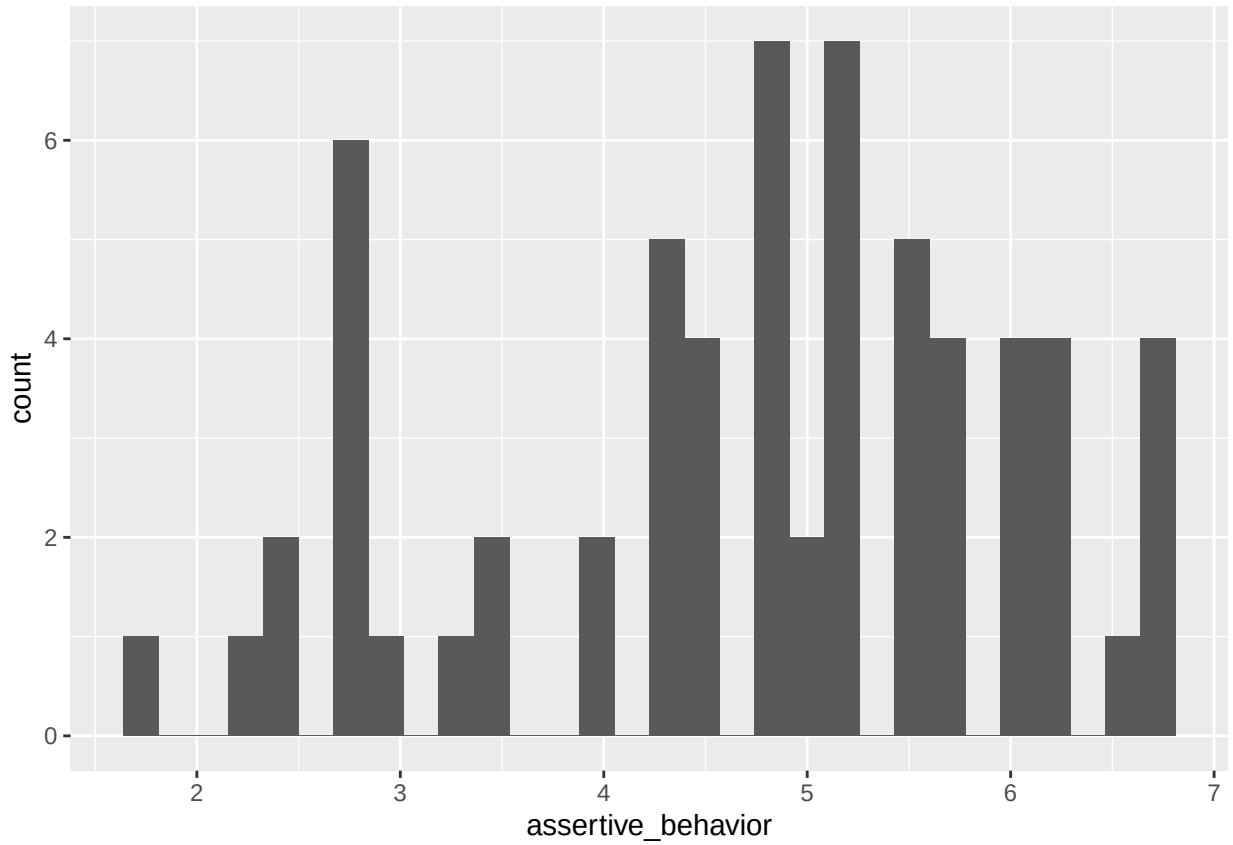

```
neccessary_data %>%
  ggplot(aes(x = sexual_motivation)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
neccessary_data %>%
  ggplot(aes(x=assertive_behavior)) +
  geom_histogram()
```

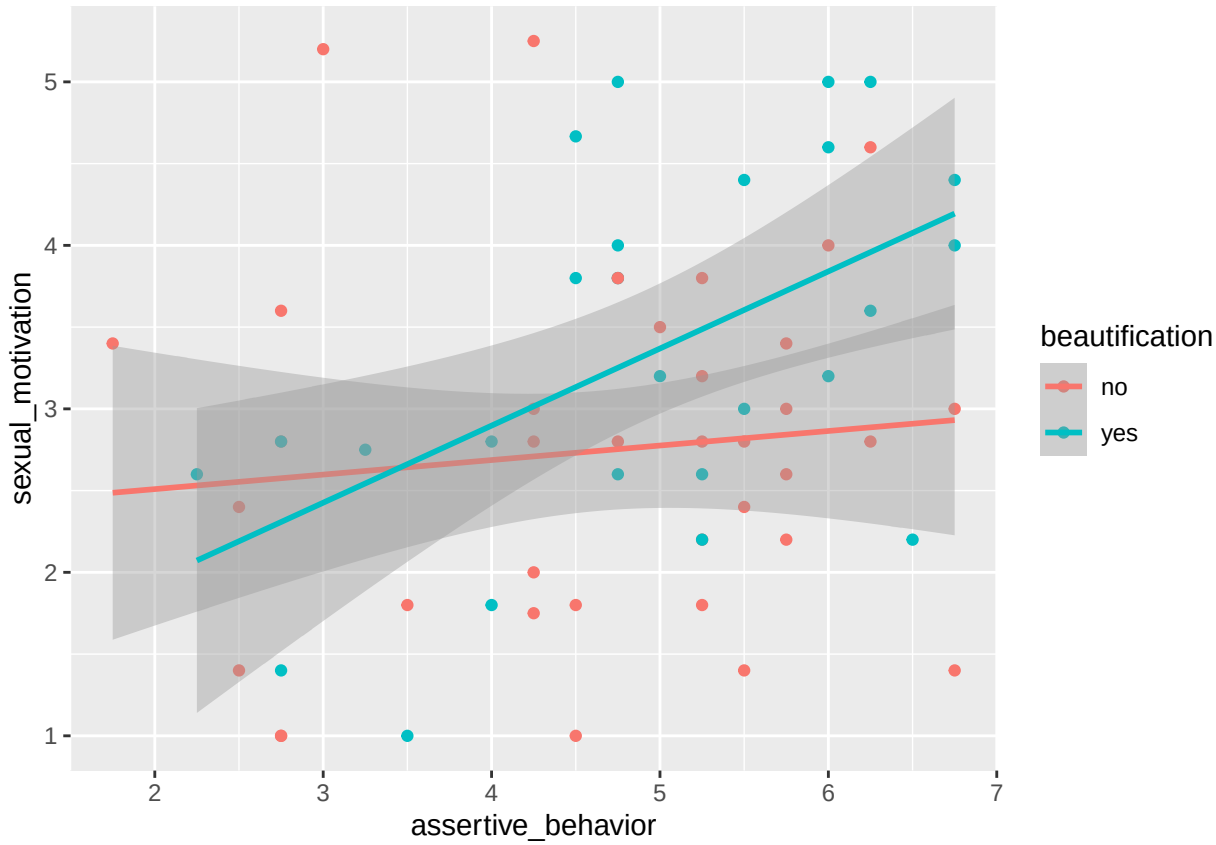
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



- Which pairs of variables appear to be related? Can you create a plot that shows potential relationships between all three variables?

```
ggplot(neccessary_data, aes(x = assertive_behavior, y = sexual_motivation)) +
  geom_point(aes(color = beautification)) +
  geom_smooth(aes(color = beautification), method = "lm", se = TRUE)
```

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



- Do any women appear to be particularly unusual (i.e., are there any outliers or women outside the range of sensible values for one or more variables)?

```
summary(neccessary_data$sexual_motivation)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.200   2.800   2.983   3.800   5.250
```

```
lb = 2.2-1.5*(3.8-2.2)
ub= 3.8+1.5*(3.8-2.2)
sum(neccessary_data$sexual_motivation > ub)
```

```
## [1] 0
```

```
sum(neccessary_data$sexual_motivation < lb)
```

```
## [1] 0
```

```
summary(neccessary_data$assertive_behavior)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.750   4.125   4.750   4.746   5.750   6.750
```

```
lb = 4.125-1.5*(5.750-4.125)
ub= 5.750+1.5*(5.750-4.125)
sum(neccessary_data$assertive_behavior > ub)
```

```
## [1] 0
```

```
sum(neccessary_data$assertive_behavior < lb)
```

```
## [1] 0
```

Based on our above results there do not appear to be any woman outside of the reasonable range in our sample.

```
summary(neccessary_data$assertive_behavior)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.750   4.125   4.750   4.746   5.750   6.750
```

Part b (Code: 2 pts)

Create the three regression models (as defined in the problem introduction) in R and obtain point estimates of the direct and indirect effect of `beautification` on `assertive_behavior`.

```
library(boot)
# 1. Propose a causal model explicitly naming the predictor variable X, response variable Y, and mediat

#x = beautification
#y = assertive behavior
#m = sexual motivation

# 2. Fit a simple linear regression model predicting Y from X, $Y = c_0 + c_1 X$.
first_lm <-lm(assertive_behavior ~ beautification, data = neccessary_data)
# 3. Fit a simple linear regression model predicting M from X, $M = b_0 + b_1 X$.
second_lm <-lm(sexual_motivation ~ beautification,
               data = neccessary_data)
# 4. Fit a multiple linear regression model predicting Y from *both* M and X, $Y = c_0' + c_1'X + c_2'M
third_lm <- lm(assertive_behavior ~ beautification + sexual_motivation,
               data = neccessary_data)
# 5. Compute the estimated *direct* effect of X on Y, $c_1$.
print(first_lm$coefficients["beautificationyes"])
```

```
## beautificationyes
##          0.268711
```

```
direct_effect = 0.268711
# 6. Compute the estimated *indirect* effect of X on Y through the mediator M. We can calculate this as
c_1.1 = 0.07803

indirect_effect = direct_effect - c_1.1
print(indirect_effect)
```

```
## [1] 0.190681
```

```
# 7. Obtain bootstrap confidence intervals for the direct and indirect effects.
```

direct_effect = 0.268711

indirect_effect= 0.190681

Part c (Code: 2 pts)

Create a function, `assertive_indirect_effect`, that takes in a data frame `df`, runs the three regression models from Part b using the data frame `df`, and returns the indirect effect.

Then, obtain a bootstrap 95% confidence interval for the indirect effect. It is probably easiest to use the `boot` and `boot.ci` functions in the `boot` package, but you can code your own bootstrapping if you really want to. The researchers used 95% BCa confidence intervals, but you do not need to replicate their work.

```r
assertive_indirect_effect <- function(df, indices){
  d <- df[indices,]
  first_lm <-lm(assertive_behavior ~ beautification,
                data = d)
  second_lm <-lm(sexual_motivation ~ beautification,
                 data = d)
  third_lm <- lm(assertive_behavior ~ beautification + sexual_motivation,
                 data = d)
  b_1 <- second_lm$coefficients["beautificationyes"]
  c_2 <- third_lm$coefficients["sexual_motivation"]
  ind_eff <- b_1*c_2
  return(ind_eff)
}

boot_assert <- boot(data = neccessary_data, statistic = assertive_indirect_effect, R = 1000)

boot.ci(boot_assert, type = "bca")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_assert, type = "bca")
##
## Intervals :
## Level       BCa
## 95%   ( 0.0009,  0.6041 )
## Calculations and Intervals on Original Scale
```

Part d (Explanation: 2 pts)

Based on your bootstrap confidence interval, do you find sufficient evidence to reject H_0 : there is no indirect effect of beautification on assertiveness as mediated via sexual motivation in favor of H_a : there is such an indirect effect? Explain your reasoning.

Since our bootstrapped confidence interval does not contain the value 0, we can therefore reject H0 since a score of 0 would imply no indirect effect.