Daniel O'Brien
DSC 424 Homework #1

2.
a.
```
     [,1]
[1,]  126
```

b.
```
    [,1]
[1,]   -6
[2,]  -18
[3,]    0
```
c.
```
     [,1]
[1,]  111
[2,]  864
[3,]  396
```
d.
```
    [,1] [,2] [,3]
[1,]  -49  -39    0
[2,]   11   59   40
[3,]   50    8   92
```
e.
```
    [,1] [,2] [,3]
[1,]   51   61    0
[2,]   11  -41   30
[3,]  -50   -2   -8
```
f.
```
    [,1] [,2]
[1,]    4   11
[2,]   11   83
```
g.
```
         [,1]       [,2]
[1,]  0.3933649 -0.05213270
[2,] -0.0521327  0.01895735
```
h.
```
    [,1]
[1,]   14
[2,]   14
```
i.
```
        [,1]
[1,]  4.777251
[2,] -0.464455
```

j.
[1] 211

Code for problem #2:

```
Z = matrix( c( 1, 1, 1, 1, 1, 0, 1, 9), nrow=4, ncol=2)
Y = matrix( c( 6, 0, 8, 0), nrow=4, ncol=1)
M = matrix( c( 1, 11, 0, 11, 09, 3, 0, 35, 42), nrow=3, ncol=3)
N = matrix( c( -50, 0, 50, -50, 50, 5, 0, 5, 50), nrow=3, ncol=3)
v = matrix( c( 45, 6, 9), nrow=3, ncol=1)
w = matrix( c( 2, 6, 0), nrow=3, ncol=1)
scalar = t(v)%*%w
scalar
product = -3*w
product
product = M %*% v
product
sum = M + N
sum
diff = M - N
diff
t(Z)%*%Z
solve((t(Z)%*%Z))
t(Z)%*%Y
beta = solve((t(Z)%*%Z)) %*% (t(Z)) %*% Y
beta
det(t(Z)%*%Z)
```

3.

The type of regression analysis that I decided to write about is the Poisson Regression. The study I chose focuses on whether participation in sports in early adolescents protects against a decline in physical activity during adolescence. The study concentrated on 7th grade students, age 12-13, and every 3 months for 5 years, the physical activity of these students were documented. This study was created because it asserts that physical activity improves the quality of life and longevity, although in industrialized countries, the majority of people are physically inactive. The belief that inactivity in youth leads to inactivity in adults motivated researchers to identify practices that will sustain physical activity in adolescents and adults. (Bélanger 2009)

Poisson regression was used to determine if there are any differences in the decline of physical activity from 7th grade to 11th grade with students who did and did not engage in physical activity in the 7th grade. Researchers created 3 separate models to analyze physical activity in 3 separate situations, physical activities in schools, physical activities in the community, and 'any' physical activity. I think that the Poisson model makes sense for this

study because documenting the number of physical activity sessions for each participant over a certain amount of time would be a count variable. Students filled out a questionnaire marking which days they engaged in physical activity, and which physical activities they participated in, this data was used to count the number of days the students participated in rigorous physical activity. Additionally, researchers could use that count variable to create ratios of the amount of physical activity over a certain amount of time.

The study found that there were no meaningful differences in gender of participants. Predictably, consistent engagement in physical activity increases in those students who engaged in organized physical activity compared to students who participated in physical activity outside of school and community organized physical activity. The study concluded that, although participation in school and community organized physical activity helps maintain higher levels of physical activity in secondary school, it does not protect against declining physical activity levels in adolescents. I thought that this study was pretty interesting, as a teacher and former adolescent myself, I think that physical activity is important, but it does tend to decline with age in general. I know that when I left school organized physical activity after high school, I saw a decline in my own rigorous physical activity.

Bélanger, M., Gray-Donald, K., O'Loughlin, J., Paradis, G., Hutcheon, J., Maximove, K., Hanely, J.(2009). Participation in organized sports does not slow declines in physical activity during adolescence. *International Journal of Behavioral Nutrition and Physical Activity, 6 (22).* https://link.springer.com/content/pdf/10.1186/1479-5868-6-22.pdf

4.

The journal article I chose, is about how different ethical theories impact the ways individuals and businesses handle big data. The, 'Big Data' in this article refers to the amount of data collected recently, according the Herschel (2017), "90% of the data in the world today has been created in the last two years alone" (p. 31). There are four theories taken into consideration: Kantianism, Utilitarianism, Social Contract Theory and Virtue Theory. According to the article, the use of these theories makes it possible to evaluate moral decision making when it comes to big data and determine what was intended versus what actually happens. (Herschel 2017).

Kantianism, originating from Immanuel Kant, focuses on what people should do, rather than what people actually do. It's implication to big data are conflicting, because everyone is their own person with their own set of beliefs and opinions on whether their information should be shared, and the way that big data is collected and shared do not often take into account the individuals feelings or opinions on the matter. Utilitarianism focuses on the idea of doing things for the greater good of society, this can lead to both positive and negative feelings about the ways that big data is collected and shared. For medical studies, it is easy to see that a greater purpose is being served, however if a marketing company is collected data to tailor advertisements to consumers, the same argument can be made, but might be less convincing. Social Contract Theory originates from the idea of developing an agreement or contract

between the individual and the society in which they live. Big data does not work seamlessly with Social Contract Theory, because often time data is collected, used and shared without the direct consent, or even awareness from the individual. Virtue ethics focuses on virtue or moral character rather than regulations and rules. Similar to Utilitarianism, if an argument can be made that big data is being collected for a cause that will benefit society, such as medical research would reflect in the positive intentions of those who collected the data and performed research with said data.

With an infinite number of the conflicting opinions, viewpoints and personal philosophies it cannot be said whether the collection of big data is right or wrong. However, the presence of moral theories to drive debate and exchange ideas is a positive thing. Companies, governments and individuals will continue to debate the process of collecting, using and sharing data for years to come, and a correct answer or even a set of guidelines may never be reached. In a subject matter of data, where correct answers, solutions and correlations are often arriving upon, it is interesting to have a debate that may never be answered.

Herschel, R., Miori, V. M. (2017). Ethics & Big Data. *Technology in Society, 49,* 31-36. https://www.sciencedirect.com/science/article/pii/S0160791X16301373

5.

a.

First, I checked for missing values in the dataset, our data set did not have any missing values, the next thing I checked about my model was how the data is distributed. I found that the dependent variable, expenses, was not evenly distributed, and therefore I transformation needed to take place. After performing a log transformation, the expenses values appeared to be much more evenly distributed. After that I checked for linearity in our dependent variable. There appeared to be linearity issues with the dependent variable, as the values on the residuals vs fitted plot formed a pattern. To remedy this, I attempted to perform different transformations in my dependent variable, however, all other transformations appeared to make the situation worse, so I kept the log transformation. There also appeared to be issues with the normality, which appeared to be off because of the shape of the line, that could be due to outliers. I ran the cook's distance plot and found 3 influential points observation #103, #431 and #1028. I removed those observations and reran the plots. After this, this problem persisted. The homoscedasticity assumption also did not appear to be met because of the shape on the scale-location plot. This also can be remedied by a transformation, however the transformations I performed did not seem to improve the homoscedasticity situation. Unable to fully resolve these issues, I moved onto the next topic, multicollinearity.

	age	gender_num	bmi	children	smoker_num	region_num	expenses
age	1.000000000	-0.020808830	0.107691641	0.05699222	-0.025210462	0.004315282	0.534392134
gender_num	-0.020808830	1.000000000	0.044778943	0.01558858	0.076184817	0.004613890	0.009489706
bmi	0.107691641	0.044778943	1.000000000	0.01558886	0.002361582	0.153158467	0.119418854
children	0.056992224	0.015588577	0.015588856	1.00000000	0.016583386	0.010604420	0.133338943
smoker_num	-0.025210462	0.076184817	0.002361582	0.01658339	1.000000000	-0.002154800	0.663460060
region_num	0.004315282	0.004613890	0.153158467	0.01060442	-0.002154800	1.000000000	-0.043530622
expenses	0.534392134	0.009489706	0.119418854	0.13333894	0.663460060	-0.043530622	1.000000000

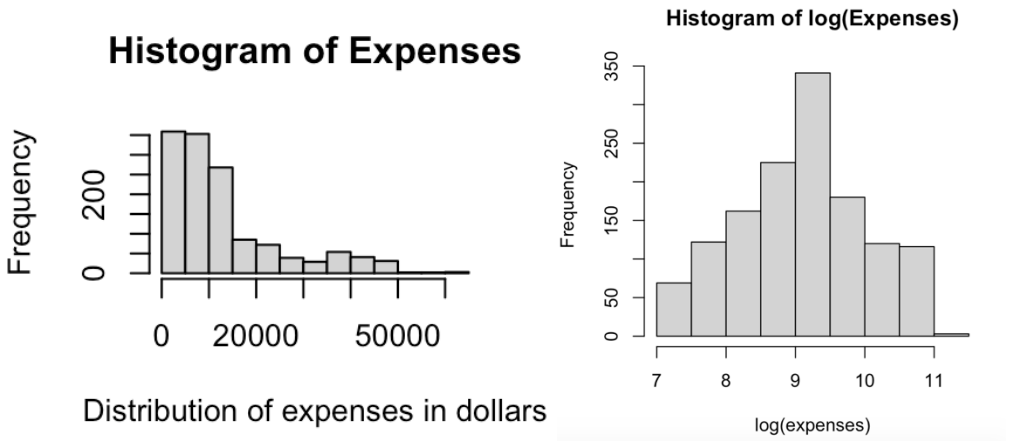
```
> vif(model1)
      age gender_num       bmi   children smoker_num region_num
 1.015411   1.008888  1.040583   1.002481   1.006468   1.025925
```

We can check for multicollinearity by first examining the correlation matrix. We have 6 independent variables and 1 dependent variable making this method manageable. The strongest correlation in the matrix between independent variables is 0.153 between region_num and bmi, which is not high enough to be of concern. After running the VIF for our variables, we can confirm that there are not any concerns related to multicollinearity, because the VIF values are well below 10. There were. No missing values in the dataset, so no rows needed to be removed. The distribution of the dependent variable was not evenly distributed and carried a skew to the right. This leads me to believe that a transformation may need to be performed.

b.

i. I ran the backward selection model to identify the strongest model. I chose to run backward selection, because backward selection begins with all of the variables and removes them if they are not significant predictors, and since it starts with all of the variables, they will all be considered for the model. Our backward selection model chose age, gender_num, bmi, children, smoker_num, region_num. All of the variables appear to be significant predictors with p-values much smaller. Gender_num has the largest p-value of the remaining variable of 0.00209, still indicating that this variable has an impact on our dependent variable.

c. For the visual representation, I have included a histogram of the dependent variable expenses. The graph does not show a symmetric distribution, which indicates that a transformation may be necessary to accurately represent the information. After producing a histogram of the dependent variable with a log transformation applied, the information appears to be much more evenly distributed. This indicates that a log transformation was necessary. We can see from the first histogram that the expenses values are skewed to the right. This tells us that most expense values are on the left side of the graph, the vast majority of expense values are below $20,000. After performing the log transformation, we can see that the log of expense values are much more evenly distributed.



Code for problem 5:

```
setwd("/Users/danielobrien/desktop")
insurance_dataset <- read.csv(file="insurance_dataset.csv", header=TRUE, sep=",")

#Check Sample Size and Number of Variables
dim(insurance_dataset)

#See structure of data
str(insurance_dataset)

#Check for missing values
sum(is.na(insurance_dataset))

#show names of variables
names(insurance_dataset)

#Show Column Numbers
library(psych)
describe(insurance_dataset)

#Show for first 6 rows of data
head(insurance_dataset)

#Show last 6 rows of data
tail(insurance_dataset)

#Remove text variables
insurance_dataset2=subset(insurance_dataset, select = -c(sex, smoker, region))
```

```
#Show for first 6 rows of data
head(insurance_dataset2)

#Histogram of Expenses
hist((insurance_dataset2$expenses), main ="Histogram of Expenses",
    xlab = "Distribution of expenses",
    ylab = "Frequency")

#Histogram of log(expenses)
hist((log(insurance_dataset2$expenses +1)), main = "Histogram of log(Expenses)",
    xlab = "log(expenses)",
    ylab = "Frequency")

#Check for Multicollinearity with Correlations
M<-cor(insurance_dataset2, method="spearman")
M

#Create initial linear regression model
model1 <- lm(log(expenses) ~ ., data = insurance_dataset2)
model1

plot(model1)

#Check VIF
install.packages("car")
library(car)
vif(model1)

model1 <- lm(log(expenses) ~ ., data=insurance_dataset2)
model1

#Run The Rest of the Assumptions
par(mfrow=c(2,2))  # set 2 rows and 2 column plot layout
model2 <- lm(log(expenses) ~ ., data=insurance_dataset2)  # linear model
plot(model2)

plot(model2, 4)

insurance_dataset3 <- insurance_dataset2[-c(103, 431, 1028), ]

par(mfrow=c(2,2))  # set 2 rows and 2 column plot layout
model3 <- lm(log(expenses) ~ ., data=insurance_dataset3)  # linear model
plot(model3)
```

```
plot(model3, 4)

#Check Model Summary
summary(model2)

#Creating Automatic Models
null = lm(log(expenses) ~ 1, insurance_dataset3)
null

full = lm(log(expenses) ~ ., insurance_dataset3)
full

#Backward Regression
insurance_Backward = step(full, direction="backward")
summary(insurance_Backward)

summary(model2)
```