Homework 8 Solutions

Due: Thursday 4/16/20 by 8:30am

Rubric:

- Maximum of 3 points for 1.-2., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for mostly correct answers but at least one substantial issue;
  - 3 points for nearly/exactly correct.
- Maximum of 4 points for 3.-4., determined as follows:
  - 0 points for no solutions whatsoever or `R` output only;
  - 1 point for an honest effort but very few correct answers or `R` output only plus a figure;
  - 2 points for about half correct answers;
  - 3 points for mostly correct answers but at least one substantial issue;
  - 4 points for nearly/exactly correct.

1. Problem 6.8 from the `.pdf` version of the textbook. Requires use of the `brand_preference` data that has been posted on the Homework page.

(a)

We obtain an 99 percent confidence interval estimate of $(73.88, 80.67)$, which can be interpreted as a 99% confidence interval for the average degree of brand liking for a product with moisture content of 5 and sweetness of 4.

```
link <- url("http://maryclare.github.io/stat525/content/homework/brand_preference.RData")
load(link)
close(link)

# We can compute this by hand or using the predict function
# First, let's do it by hand
# Start by constructing the response vector and X matrix
Y <- data$Y
X1 <- data$X1
X2 <- data$X2
n <- length(Y)
X <- cbind(rep(1, n), X1, X2)

# Compute regression coefficient estimates and residuals
b <- solve(t(X)%*%X)%*%t(X)%*%Y
e <- Y - X%*%b

# Compute the estimated variance-covariance matrix of the
# estimated regression coefficients
ssr <- sum(e^2)
s.sq.b <- ssr*solve(t(X)%*%X)/(n - 3)

# Computed estimated average future response value and its
# standard error
```

```
X.h <- c(1, 5, 4)
Y.hat.h <- X.h%*%b
s.sq.hat.h <- t(X.h)%*%s.sq.b%*%X.h
alpha <- 0.01
low.99 <- Y.hat.h + qt(alpha/2, n - 3)*sqrt(s.sq.hat.h)
hig.99 <- Y.hat.h + qt(1 - alpha/2, n - 3)*sqrt(s.sq.hat.h)

# Alternatively, R can do this for us quickly
linmod <- lm(Y~X1 + X2, data = data)
pred <- predict(linmod,
                newdata = data.frame("X1" = 5, "X2" = 4),
                interval = "confidence",
                level = 0.99)
low.99 <- pred[2]
hig.99 <- pred[3]
```

(b)

We obtain an 99 percent prediction interval of $(68.48, 86.01)$.

```
# Again, we can compute this by hand or using the predict function
# First, let's do it by hand
s.sq.pred <- t(X.h)%*%s.sq.b%*%X.h + ssr/(n - 3)
alpha <- 0.01
pred.low.99 <- Y.hat.h + qt(alpha/2, n - 3)*sqrt(s.sq.pred)
pred.hig.99 <- Y.hat.h + qt(1 - alpha/2, n - 3)*sqrt(s.sq.pred)

# Alternatively, R can do this for us quickly
pred <- predict(linmod,
                newdata = data.frame("X1" = 5, "X2" = 4),
                interval = "prediction",
                level = 0.99)
pred.low.99 <- pred[2]
pred.hig.99 <- pred[3]
```

2. Problem 6.14 from the `.pdf` version of the textbook, parts (a)-(b). Requires use of the `grocery_retailer` data that has been posted on the Homework page.

(a)

We obtain an 95 percent prediction interval for the mean handling time of these shipments of $(4,232.45, 4,324.28)$.

```
link <- url("http://maryclare.github.io/stat525/content/homework/grocery_retailer.RData")
load(link)
close(link)

# Having done this by hand once,
# let's use R's predict function to  do this
linmod <- lm(Y ~ X1 + X2 + X3, data = data)
pred <- predict(linmod,
                newdata = data.frame("X1" = 282000,
                                     "X2" = 7.10,
                                     "X3" = 0),
                interval = "confidence",
                level = 0.95)
```

```
low.95 <- pred[2]
hig.95 <- pred[3]
```

(b)

We obtain an 99 percent prediction interval of for the total labor hours for the three shipments of $(3,986.63, 4,570.10)$.

```
# Alternatively, R can do this for us quickly
pred <- predict(linmod,
                newdata = data.frame("X1" = 282000,
                                     "X2" = 7.10,
                                     "X3" = 0),
                interval = "prediction",
                level = 0.95)
pred.low.95 <- pred[2]
pred.hig.95 <- pred[3]
```

3. Problem 6.28 from the `.pdf` version of the textbook. Requires use of the `CDI` data that has been posted on the Homework page. You will need to look to the textbook appendices for a description of the variables in this dataset. See page 1,367 of the `.pdf` version of the textbook.

(a)

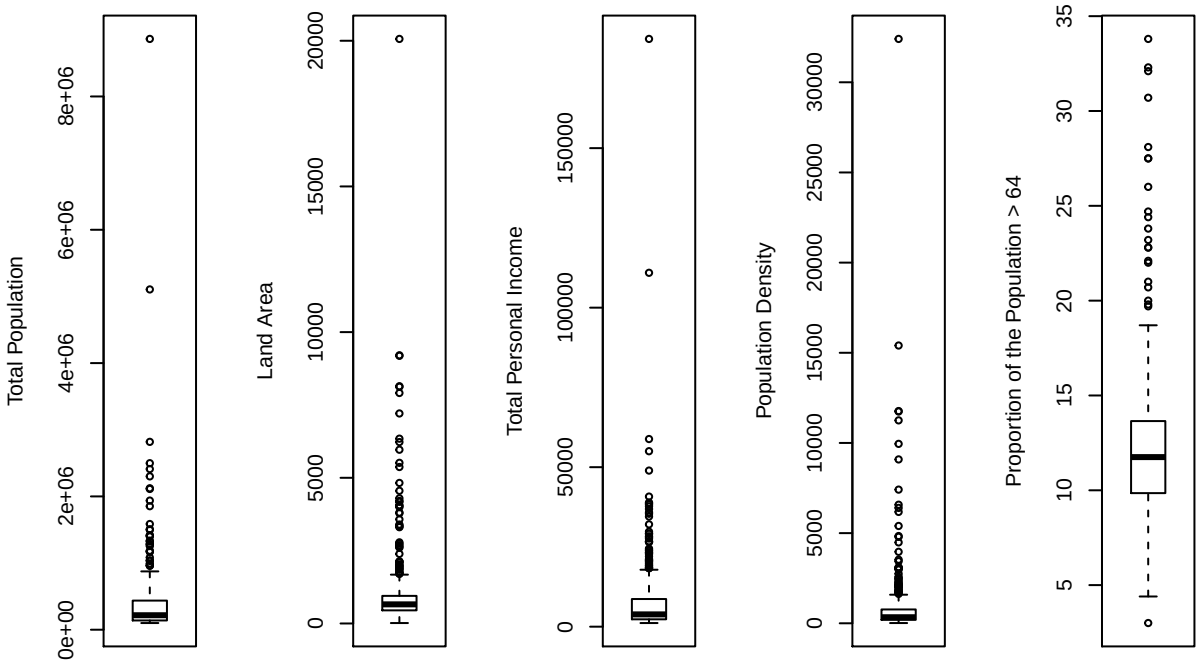
```
link <- url("http://maryclare.github.io/stat525/content/homework/CDI.RData")
load(link)
close(link)

# X8: Number of active physicians
# X5: Total population
# X4: Land area
# X16: Total personal income
# X18=X5/X4: Population density
# X7: Proportion of population over 64 years old

# Let's construct a new variable for population density
data$X18 <- data$X5/data$X4
```

In retrospect, I wish I had not assigned this part of this problem. Stem-and-leaf plots are *not* a good way to visualize predictor values. Boxplots or histograms would be a much more effective. I am just going to put boxplots here, and give people full credit for attempting stem-and-leaf plots.

```
par(mfrow = c(1, 5))
boxplot(data$X5, ylab = "Total Population")
boxplot(data$X4, ylab = "Land Area")
boxplot(data$X16, ylab = "Total Personal Income")
boxplot(data$X18, ylab = "Population Density")
boxplot(data$X7, ylab = "Proportion of the Population > 64")
```

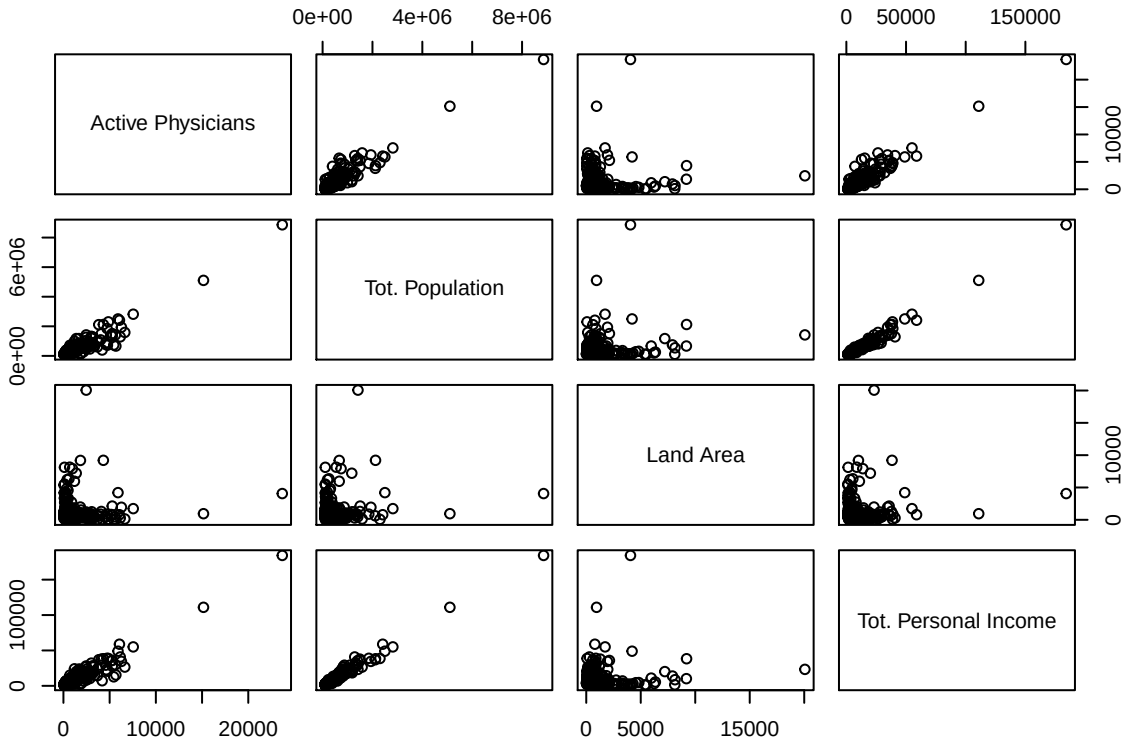


All of the predictors are skewed to the right, meaning that each predictor has several outlying values that are much larger than the rest.

(b)

The scatterplot matrix and correlation matrix for the first model are given first, followed by the scatterplot matrix and correlation matrix for the second model.

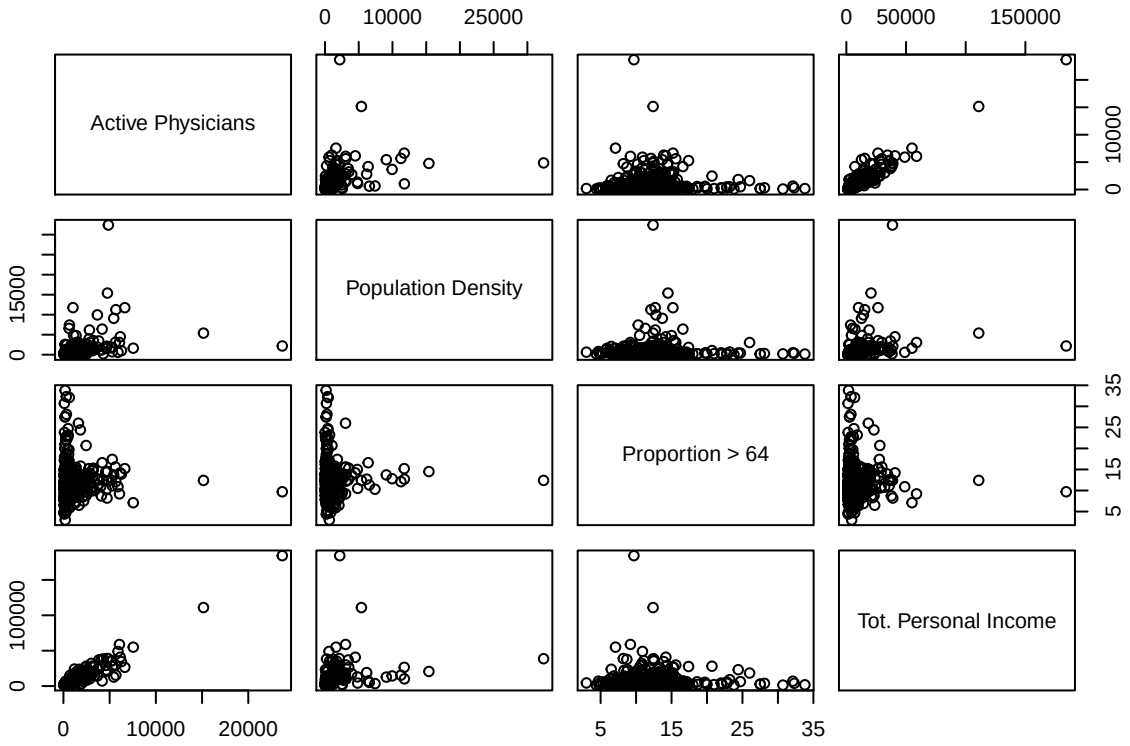
```
pairs(data[, c("X8", "X5", "X4", "X16")],
      labels = c("Active Physicians", "Tot. Population", "Land Area", "Tot. Personal Income"))
```



```
cor(data[, c("X8", "X5", "X4", "X16")])
```

	Active Physicians	Tot. Population	Land Area	Tot. Personal Income
Active Physicians	1.00	0.94	0.08	0.95
Tot. Population	0.94	1.00	0.17	0.99
Land Area	0.08	0.17	1.00	0.13
Tot. Personal Income	0.95	0.99	0.13	1.00

```
pairs(data[, c("X8", "X18", "X7", "X16")],
      labels = c("Active Physicians", "Population Density", "Proportion > 64", "Tot. Personal Income"))
```




```
cor(data[, c("X8", "X18", "X7", "X16")])
```

	Active Physicians	Population Density	Proportion > 64	Tot. Personal Income
Active Physicians	1.00	0.41	0.00	0.95
Population Density	0.41	1.00	0.03	0.32
Proportion > 64	0.00	0.03	1.00	-0.02
Tot. Personal Income	0.95	0.32	-0.02	1.00

With respect to the first model, we observe that the number of active physicians, has clear approximately linear relationship with the total population and total personal income. The number of active physicians is positively associated with both total population and total personal income. However, we also observe that total population and total personal income are strongly correlated with each other. Land area does not have a clear linear relationship with the response or the other predictors.

With respect to the second model, we observe that the number of active physicians has less clear approximately linear relationship with population density and no clear linear relationship with the proportion of persons over age 64. The number of active physicians is positively associated population density and total personal

income. We also observe that population density and total personal income are much less strongly with each other than total population and total personal income were.

(c)

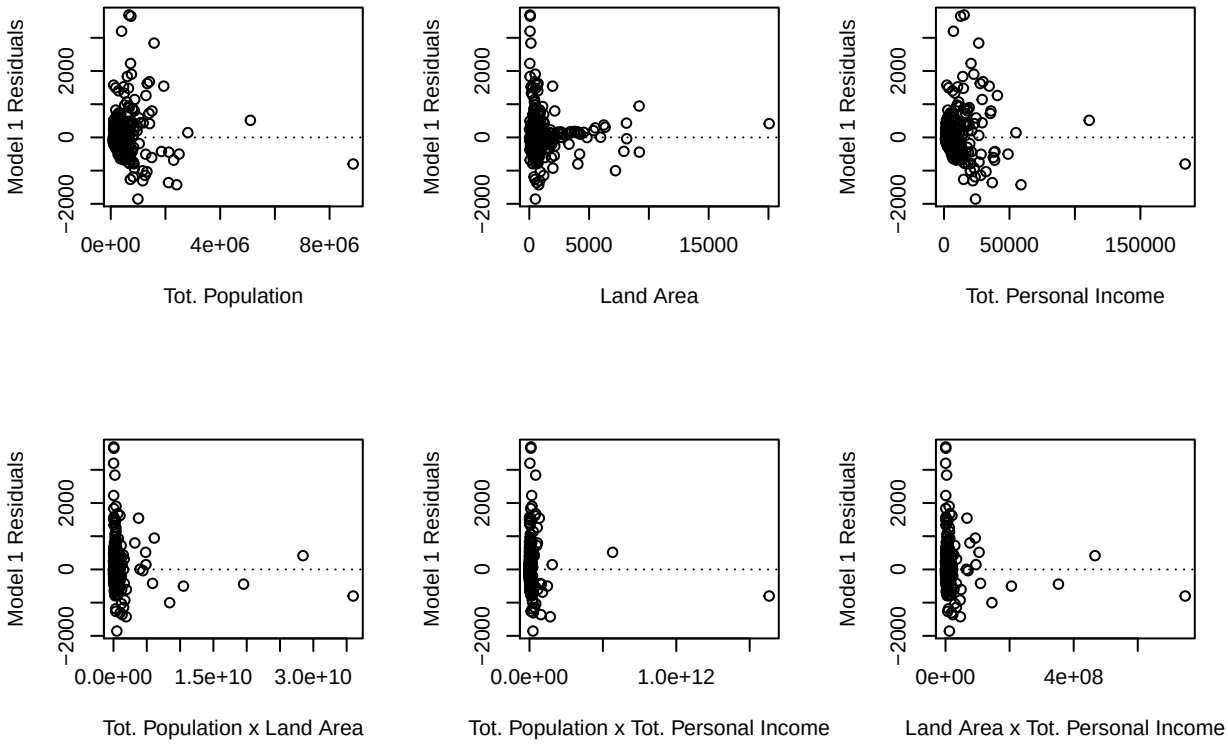
```
linmod1 <- lm(X8 ~ X5 + X4 + X16, data = data)
linmod2 <- lm(X8 ~ X18 + X7 + X16, data = data)
```

(d)

The R^2 of the first model is 0.9026, whereas the R^2 of the second model is 0.0.9117. In terms of R^2, the second model is preferable because it the predictors in the second model explain slightly more of the overall variation in the response.

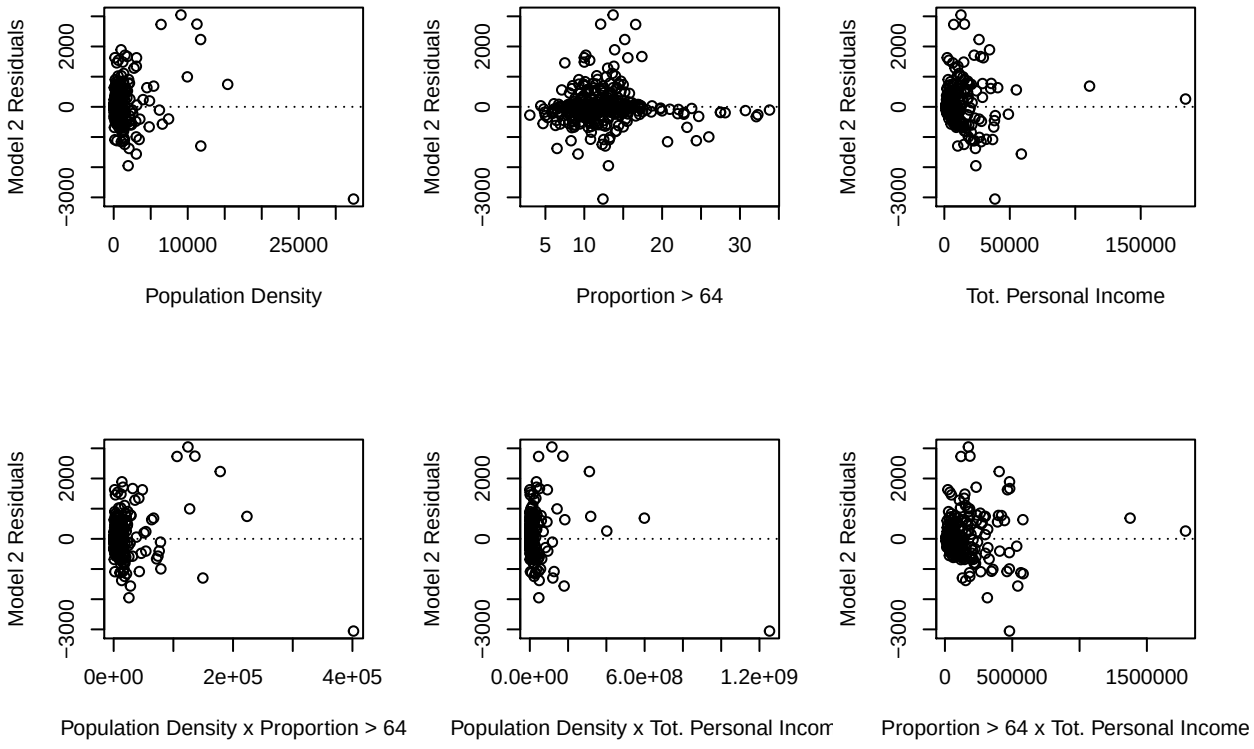
(e)

```
par(mfrow = c(2, 3))
plot(data$X5, linmod1$residuals,
     ylab = "Model 1 Residuals",
     xlab = "Tot. Population")
abline(h = 0, lty = 3)
plot(data$X4, linmod1$residuals,
     ylab = "Model 1 Residuals",
     xlab = "Land Area")
abline(h = 0, lty = 3)
plot(data$X16, linmod1$residuals,
     ylab = "Model 1 Residuals",
     xlab = "Tot. Personal Income")
abline(h = 0, lty = 3)
plot(exp(log(data$X5) + log(data$X4)), linmod1$residuals,
     ylab = "Model 1 Residuals",
     xlab = "Tot. Population x Land Area")
abline(h = 0, lty = 3)
plot(exp(log(data$X5) + log(data$X16)), linmod1$residuals,
     ylab = "Model 1 Residuals",
     xlab = "Tot. Population x Tot. Personal Income")
abline(h = 0, lty = 3)
plot(data$X4*data$X16, linmod1$residuals,
     ylab = "Model 1 Residuals",
     xlab = "Land Area x Tot. Personal Income")
abline(h = 0, lty = 3)
```



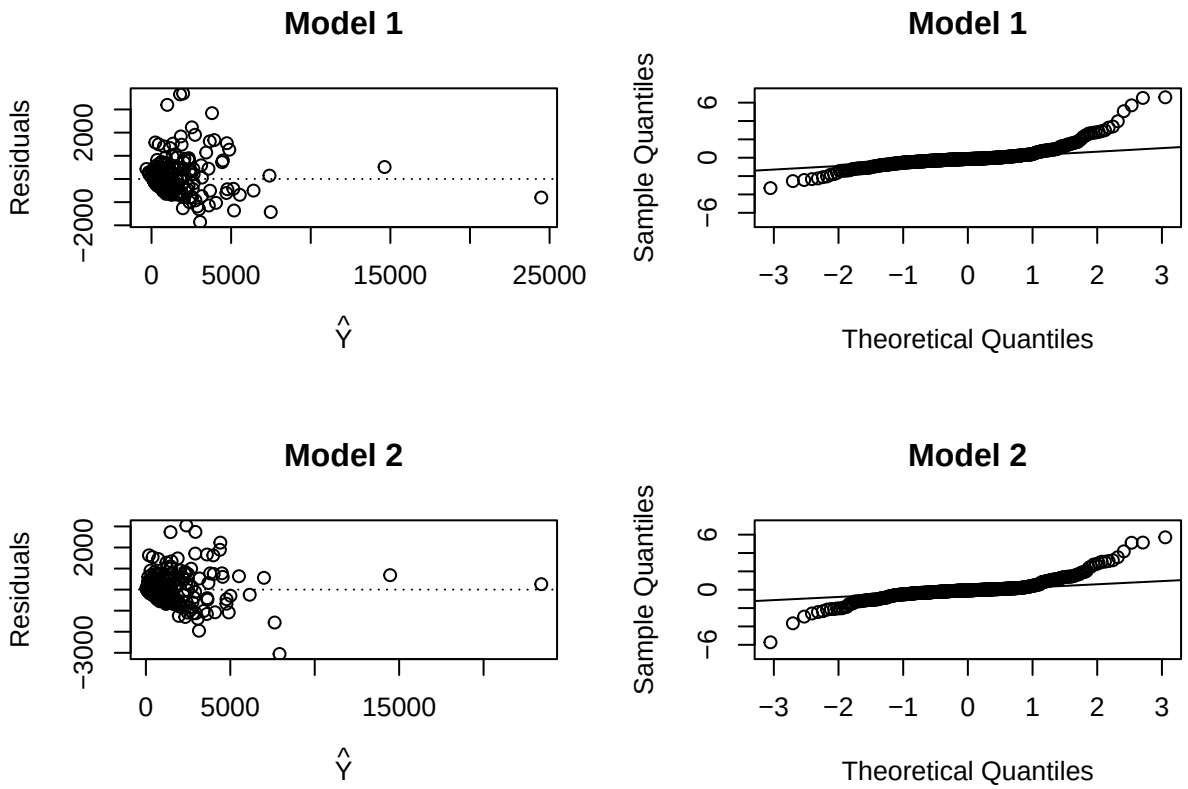
The residuals from the first model do not show any clear trends that suggest that any predictors need to be transformed, or that interaction terms need to be added.

```
par(mfrow = c(2, 3))
plot(data$X18, linmod2$residuals,
     ylab = "Model 2 Residuals",
     xlab = "Population Density")
abline(h = 0, lty = 3)
plot(data$X7, linmod2$residuals,
     ylab = "Model 2 Residuals",
     xlab = "Proportion > 64")
abline(h = 0, lty = 3)
plot(data$X16, linmod2$residuals,
     ylab = "Model 2 Residuals",
     xlab = "Tot. Personal Income")
abline(h = 0, lty = 3)
plot(data$X18*data$X7, linmod2$residuals,
     ylab = "Model 2 Residuals",
     xlab = "Population Density x Proportion > 64")
abline(h = 0, lty = 3)
plot(data$X18*data$X16, linmod2$residuals,
     ylab = "Model 2 Residuals",
     xlab = "Population Density x Tot. Personal Income")
abline(h = 0, lty = 3)
plot(data$X7*data$X16, linmod2$residuals,
     ylab = "Model 2 Residuals",
     xlab = "Proportion > 64 x Tot. Personal Income")
abline(h = 0, lty = 3)
```



For the most part, the residuals from the second model do not show any clear trends that suggest that any predictors need to be transformed, or that interaction terms need to be added. However, one exception is the residuals from the second model as a function of the proportion of the population over age 64, which are systematically negative when the proportion over age 64 exceeds 20. This suggests that it the relationship between the number of active physicians and the proportion over age 64 is not linear.

```
par(mfrow = c(2, 2))
plot(linmod1$fitted.values, linmod1$residuals,
     xlab = expression(hat(Y)), ylab = "Residuals",
     main = "Model 1")
abline(h = 0, lty = 3)
qqnorm(linmod1$residuals/summary(linmod1)$sigma,
       main = "Model 1", ylim = c(-7, 7))
qqline(linmod1$residuals/summary(linmod1)$sigma)
plot(linmod2$fitted.values, linmod2$residuals,
     xlab = expression(hat(Y)), ylab = "Residuals",
     main = "Model 2")
abline(h = 0, lty = 3)
qqnorm(linmod2$residuals/summary(linmod2)$sigma,
       main = "Model 2", ylim = c(-7, 7))
qqline(linmod2$residuals/summary(linmod2)$sigma)
```



Residuals from both models suggest that outlying values of the response that correspond to CDI's with very high numbers of active physicians may have a strong influence on the results, because the corresponding residuals are very close to zero. The residuals from both models deviate from what we would expect to observe if the errors were independent and identically distributed according to a normal distribution. The residuals from the second model appear to deviate more, which suggests that the second model may be less appropriate for this data even though it has a slightly higher R^2 because there is more evidence that the linear model assumptions may be violated for the second model, as opposed to the first.

4. Problem 6.29 from the `.pdf` version of the textbook. Requires use of the `CDI` data that has been posted on the Homework page. You will need to look to the textbook appendices for a description of the variables in this dataset. See page 1,367 of the `.pdf` version of the textbook. Hint: you can use the `subset` argument of the `lm` function to fit a regression model to a subset of the data. See below for an example.

(a)

```
# X10: Number of serious crimes
# X18: Population Density
# X15: Per Capita Personal Income
# X11: Proportion of High School Graduates
# X17: Geographic Region
linmod1 <- lm(X10 ~ X18 + X15 + X11,
              data = data,
              subset = X17 == 1)
linmod2 <- lm(X10 ~ X18 + X15 + X11,
              data = data,
              subset = X17 == 2)
linmod3 <- lm(X10 ~ X18 + X15 + X11,
              data = data,
              subset = X17 == 3)
```

```
linmod4 <- lm(X10 ~ X18 + X15 + X11,
              data = data,
              subset = X17 == 4)
```

Region	Regression Function
NE	$E\{Y\} = -64466.23 + 17.38X_1 - 1.41X_2 + 1182.58X_3$
NC	$E\{Y\} = -4163.27 + 33.62X_1 + 0.10X_2 - 2.76X_3$
S	$E\{Y\} = 38862.67 + 5.54X_1 + 1.96X_2 - 670.88X_3$
W	$E\{Y\} = 129323.42 + 5.72X_1 + 4.34X_2 - 2159.92X_3$

(b)

The estimated regression functions are not very similar across regions, with the exception of the sign of the estimate of b_1, which corresponds to the average change in the number of serious crimes when the population density increases by one unit, holding per capita personal income and the percent high school graduates constant. In all four regions, a one unit increase in population density is associated with an *increase* in the number of serious crimes, holding per capita personal income and the percent high school graduates constant. However, the magnitude of the increase differs substantially across the four regions. Whether or not the the average number of serious crimes tends to increase or decrease when per capita income increases by one unit and population density and the percent high school graduates are held constant depends on the region. Similarly, whether or not the the average number of serious crimes tends to increase or decrease when the percent high school graduates increases by one unit and population density and the per capita income are held constant depends on the region.

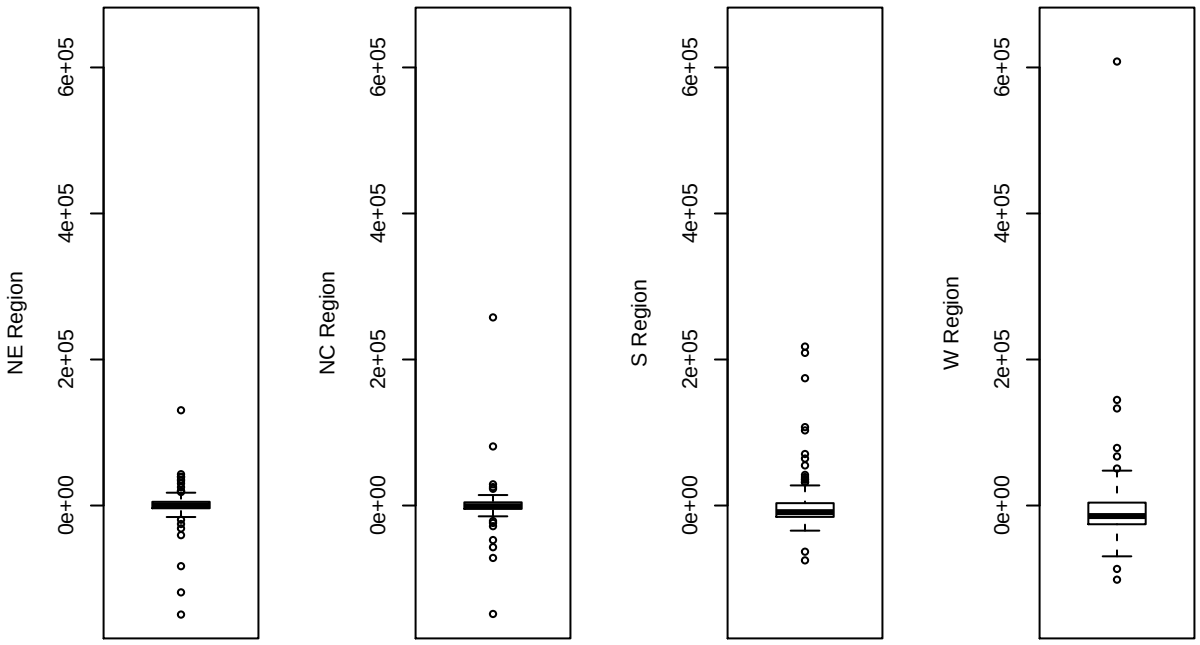
(c)

Region	MSE	R^2
NE	7.8739748×10^8	0.84
NC	1.0876238×10^9	0.53
S	1.3671082×10^9	0.09
W	6.6945914×10^9	0.09

The measures are not similar for all four regions. The model explains much more of the observed variability in the number of serious crimes in the NE and NC regions, especially in the NE region. This is reflected in the smaller MSE values and larger R^2 values in the NE and NC regions. Based on the very large MSE's and very large R^2 values in the S and W regions, we can conclude that the model explains very little of the observed variability in the number of serious crimes in the S and W regions.

(d)

```
par(mfrow = c(1, 4))
boxplot(linmod1$residuals, ylab = "NE Region",
        ylim = c(-150000, 650000))
boxplot(linmod2$residuals, ylab = "NC Region",
        ylim = c(-150000, 650000))
boxplot(linmod3$residuals, ylab = "S Region",
        ylim = c(-150000, 650000))
boxplot(linmod4$residuals, ylab = "W Region",
        ylim = c(-150000, 650000))
```



The boxplots of the residuals reflect the same trends we observed in (c) - the model produces relatively more and larger outlying positive residuals when applied to the S and W regions versus the N and NC regions. This means that the model tends to produce more estimates of the number of serious crimes that are far too low in the S and W regions as compared to the N and NC regions.