

STAT 707 Homework 1

Daniel Lupercio

3/1/2021

```
library(matlib)
library(GGally)
library(lubridate)
library(tidyverse)
library(ggpubr)
library(faraway)
```

6.27. In a small-scale regression study, the following data were obtained:

```
num_data <- matrix(c(7,33,42,4,41,33,16,7,75,3,49,28,21,5,91,8,31,55), nrow = 6, ncol = 3, byrow = TRUE)

##      [,1]  [,2]  [,3]
## [1,]    7   33   42
## [2,]    4   41   33
## [3,]   16    7   75
## [4,]    3   49   28
## [5,]   21    5   91
## [6,]    8   31   55

data1 <- as.data.frame(num_data)
```

We begin by isolating our matrices

```
Y <- matrix(matrix(num_data[,3],nrow = 6, ncol = 1), nrow = 6, ncol = 1); Y

##      [,1]
## [1,]   42
## [2,]   33
## [3,]   75
## [4,]   28
## [5,]   91
## [6,]   55

X <- matrix(c(1,7,33,1,4,41,1,16,7,1,3,49,1,21,5,1,8,31),nrow = 6, ncol = 3, byrow = T); X
```

```

##      [,1] [,2] [,3]
## [1,]    1    7   33
## [2,]    1    4   41
## [3,]    1   16    7
## [4,]    1    3   49
## [5,]    1   21    5
## [6,]    1    8   31

```

```
Xtran <- t(X); Xtran
```

```

##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    1    1    1
## [2,]    7    4   16    3   21    8
## [3,]   33   41    7   49    5   31

```

Assume that the regression model (6.1) with independent normal error terms is appropriate.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}$$

Using matrix methods, obtain

(a) b

$$b = (X'X)^{-1}X'Y$$

```
b <- inv(Xtran %*% X) %*% Xtran %*% Y; b
```

```

##      [,1]
## [1,] 33.9321020
## [2,] 2.7847707
## [3,] -0.2643979

```

Here, $\beta_0, \beta_1, \beta_2$ are 33.932, 2.2784, and -0.264, respectively.

(b) e

The fitted values are represented by $\hat{Y} = Xb$ and the residual terms by

$$e = Y - \hat{Y} = Y - Xb$$

```
e <- Y - X %*% b; e
```

```

##      [,1]
## [1,] -2.70036663
## [2,] -1.23087135
## [3,] -1.63764825
## [4,] -1.33091751
## [5,] -0.09029763
## [6,]  6.98606687

```

(c) \mathbf{H}

Define the matrix \mathbf{H} as follows $\mathbf{H} = X(X'X)^{-1}X'$

```
H <- X %*% inv(Xtran %*% X) %*% Xtran; H
```

```
##          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
## [1,]  0.23143639  0.25168006  0.21178834  0.1488734 -0.05475455  0.21099418
## [2,]  0.25168006  0.31240977  0.09437951  0.2662835 -0.14787196  0.22314063
## [3,]  0.21178834  0.09437951  0.70442097 -0.3191731  0.10446756  0.20412257
## [4,]  0.14887339  0.26628346 -0.31917314  0.6142637  0.14143589  0.14834214
## [5,] -0.05475455 -0.14787196  0.10446756  0.1414359  0.94040059  0.01632796
## [6,]  0.21099418  0.22314063  0.20412257  0.1483421  0.01632796  0.19708945
```

(d) SSR

Let \mathbf{J} denote a $n \times n$ matrix of 1's. In this exercise, we would have a 6×6 matrix.

```
J <- matrix(1, nrow = 6, ncol = 6)
```

$$\text{SSR} = bX'Y - \left(\frac{1}{n}\right)Y'JY = Y'[H - \frac{1}{n}J]Y, \quad (df = p-1)$$

```
t(Y) %*% (H - J*(1/6)) %*% Y
```

```
##          [,1]
## [1,] 3010.107
```

Thus, our residual sum of squares is approximately equal to 3010.107.

(e) $s^2\{\mathbf{b}\}$

$s^2b = MSE(X'X)^{-1}$ where $MSE = \frac{SSE}{n-p}$ and $SSE = Y'Y - b'X'Y$

First lets find our SSE.

```
SSE <- t(Y) %*% Y - t(b) %*% t(X) %*% Y; SSE
```

```
##          [,1]
## [1,] 61.89313
```

Now lets find our MSE value.

```
#n = 6 and p = 3
MSE <- SSE/(6-3); MSE
```

```
##          [,1]
## [1,] 20.63104
```

The estimated variance-covariance matrix s^2b now is equal to

```
s_squared_b <- 20.631 * inv(t(X) %*% X); s_squared_b
```

```
## [,1]      [,2]      [,3]
## [1,] 713.39022 -34.0595669 -13.5553968
## [2,] -34.05957   1.6568335   0.6421941
## [3,] -13.55540   0.6421941   0.2617044
```

(f) \hat{Y}_h when $X_{h1} = 10, X_{h2} = 30$

Our new matrix of values X_h

```
X_h <- matrix(c(1,10,30), nrow = 3, ncol = 1); X_h
```

```
## [,1]
## [1,] 1
## [2,] 10
## [3,] 30
```

The estimated mean response corresponding to X_h , denoted by \hat{Y}_h , is $\hat{Y}_h = X'_h b$

```
t(X_h) %*% b
```

```
## [,1]
## [1,] 53.84787
```

Thus \hat{Y}_h is equal to 53.848

(g) $s^2(\hat{Y}_h)$ when $X_{h1} = 10, X_{h2} = 30$

The estimated variance $s^2(\hat{Y}_h)$ is given by:

$$s^2(\hat{Y}_h) = MSE(X'_h(X'X)^{-1}X_h) = X'_h s^2(b) X_h$$

```
s_squared_yhat <- t(X_h) %*% s_squared_b %*% X_h; s_squared_yhat
```

```
## [,1]
## [1,] 5.408904
```

The estimated variance, $s^2(\hat{Y}_h) \approx 5.409$.

Exercise 6.28

You have been asked to evaluate two alternative models for predicting the number of active physicians (Y) in a CDI. Proposed model I includes as predictor variables total population (X_1), land area (X_2), and total personal income (X_3). Proposed model II includes as predictor variables population density (X_1 , the total population divided by land area), percent of population greater than 65 years old (X_2), and total personal income (X_3)

```

CD1 <- read.csv("/Users/daniel421/Desktop/R/STAT707/STAT_707_GLM2/CDI_Data.csv", header = FALSE)

names(CD1) <- c("ID", "county", "state", "land_area", "total_pop", "percent_pop_18_34", "percent_pop_65", "n")

CDI <- CD1
rm(CD1)

```

a. Prepare a boxplot for each of the predictor variables. What noteworthy information is provided by your plots?

```

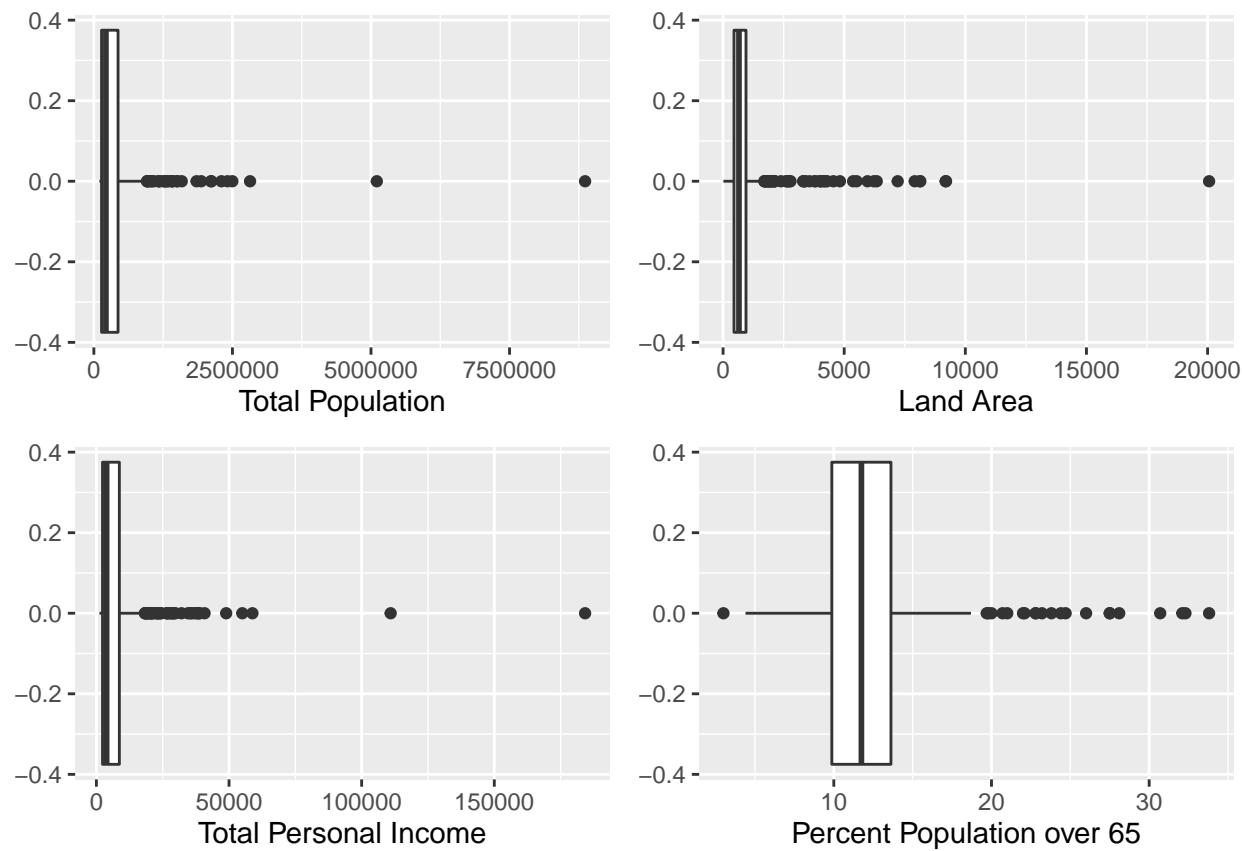
a1 <- ggplot(data = CDI, mapping = aes(x = total_pop)) + geom_boxplot() + labs(x = "Total Population");
a2 <- ggplot(data = CDI, mapping = aes(x = land_area)) + geom_boxplot() + labs(x = "Land Area");
a3 <- ggplot(data = CDI, mapping = aes(x = total_income)) + geom_boxplot() + labs(x = "Total Personal Income");
a4 <- ggplot(data = CDI, mapping = aes(x = percent_pop_65)) + geom_boxplot() + labs(x = "Percent Population over 65")

```

```

figure <- ggarrange(a1, a2, a3, a4,
                      ncol = 2, nrow = 2); figure

```

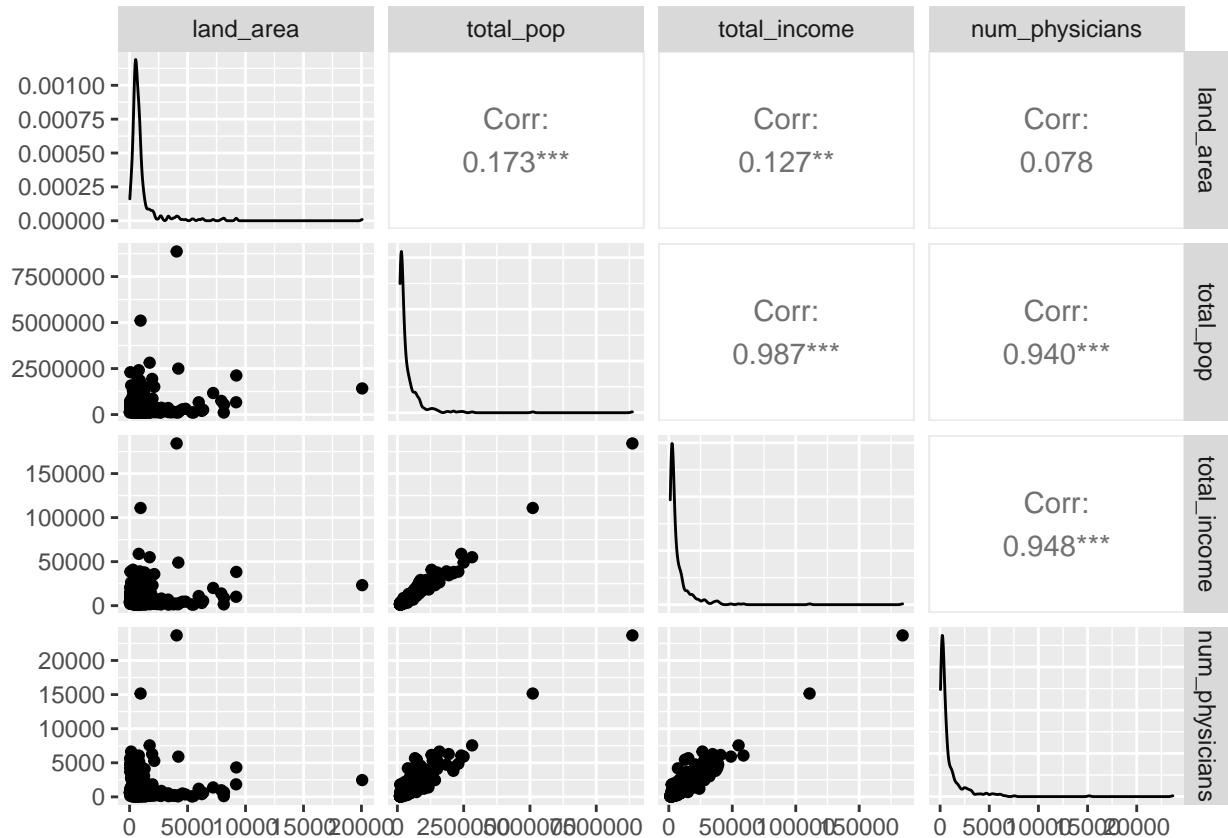


Three of the variables appear to be skewed to the right, many of the values fall out of range of the 4 quartiles.

b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

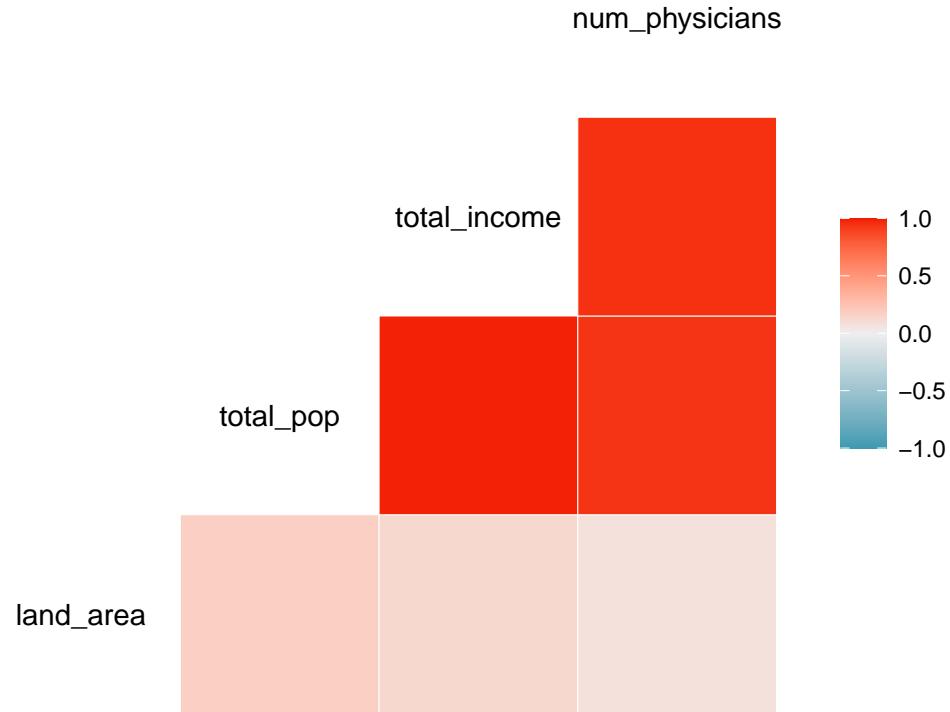
We will first begin with a combination of a scatter plot, density plot and the correlation values of the variables.

```
ggpairs(CDI, columns = c(4,5,16, 8))
```



We then look at a more visual correlation plot, detailing that the number of physicians, total personal income and the total population of the county are highly correlated with each other.

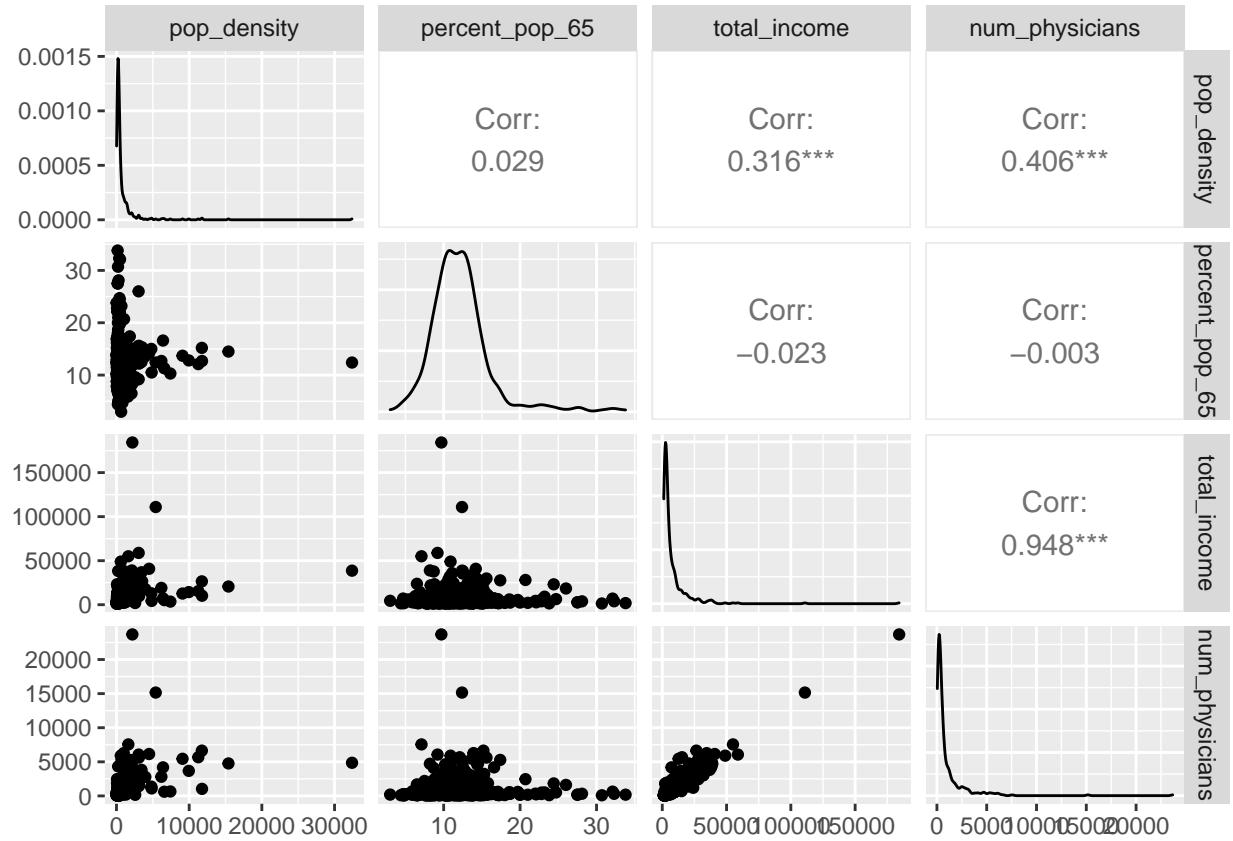
```
ggcorr(data = CDI[,c(4,5,16, 8)], method = c("everything","pearson"))
```



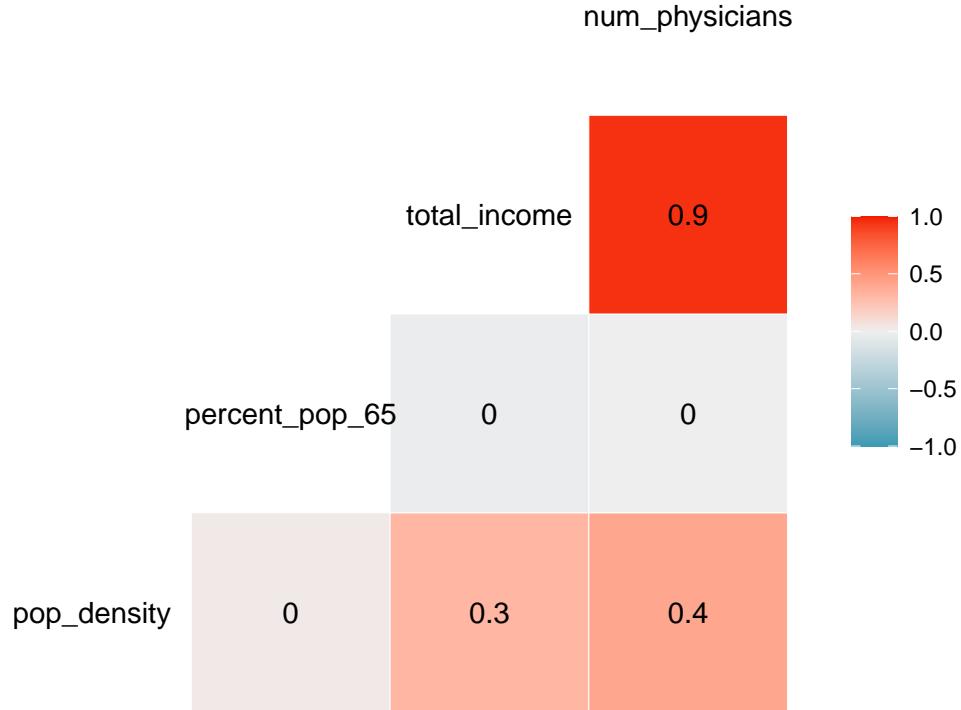
We will now look at the second model. A new variable “pop_density” will need to be created, to represent the total population divided by the land area.

```
CDI["pop_density"] <- CDI$total_pop/CDI$land_area
```

```
ggpairs(data = CDI[,c(18,7,16,8)])
```



```
ggcorr(data = CDI[,c(18,7,16,8)], method = c("everything","pearson"), label = TRUE)
```



We see that most of the variables are skewed to the right, with the “percent of population 65 or older”, being the only variable who appears to approach a normal distribution. Two pairs, “percent_pop_65” with “total_income” and “percent_pop_65” with “num_physicians” have low negative correlation values. The number of physicians again, appears to be highly correlated with total income.

c. For each proposed model, fit the first-order regression model (6.5) with three predictor variables.

For both models, we will fit the regression:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

using our indicated predicted variables.

Model 1

```
model_1 <- glm(num_physicians ~ total_pop + land_area + total_income, data = CDI)
```

$$\hat{Y}_1 = -13.3162 + 0.000836618X_1 - 0.065523X_2 + 0.094132X_3$$

Model 2

```
model_2 <- glm(num_physicians ~ pop_density + percent_pop_65 + total_income, data = CDI)
```

$$\hat{Y}_2 = -170.574 + 0.0961589X_1 + 6.33984X_2 + 0.126566X_3$$

- d. Calculate R^2 for each model. Is one model clearly preferable in terms of this measure?

```
model_1_R2 <- with(summary(model_1), 1 - deviance/null.deviance)
model_2_R2 <- with(summary(model_2), 1 - deviance/null.deviance)
```

The coefficient of multiple determination, R^2 for model 1 is approximately equal to 0.903, while the R^2 for model 2 is approximately equal to 0.912. You can't determine which model is more preferable in terms of their R^2 , other tests of normality should be conducted.

- e. For each model, obtain the residuals and plot them against \hat{Y} , each of the three predictor variables, and each of the two-factor interaction terms.

We will first obtain the residuals of both models, and concatenate them to our CDI data frame.

```
CDI["model_1_resid"] <- residuals(model_1)
CDI["model_2_resid"] <- residuals(model_2)
```

Model 1 diagnostics

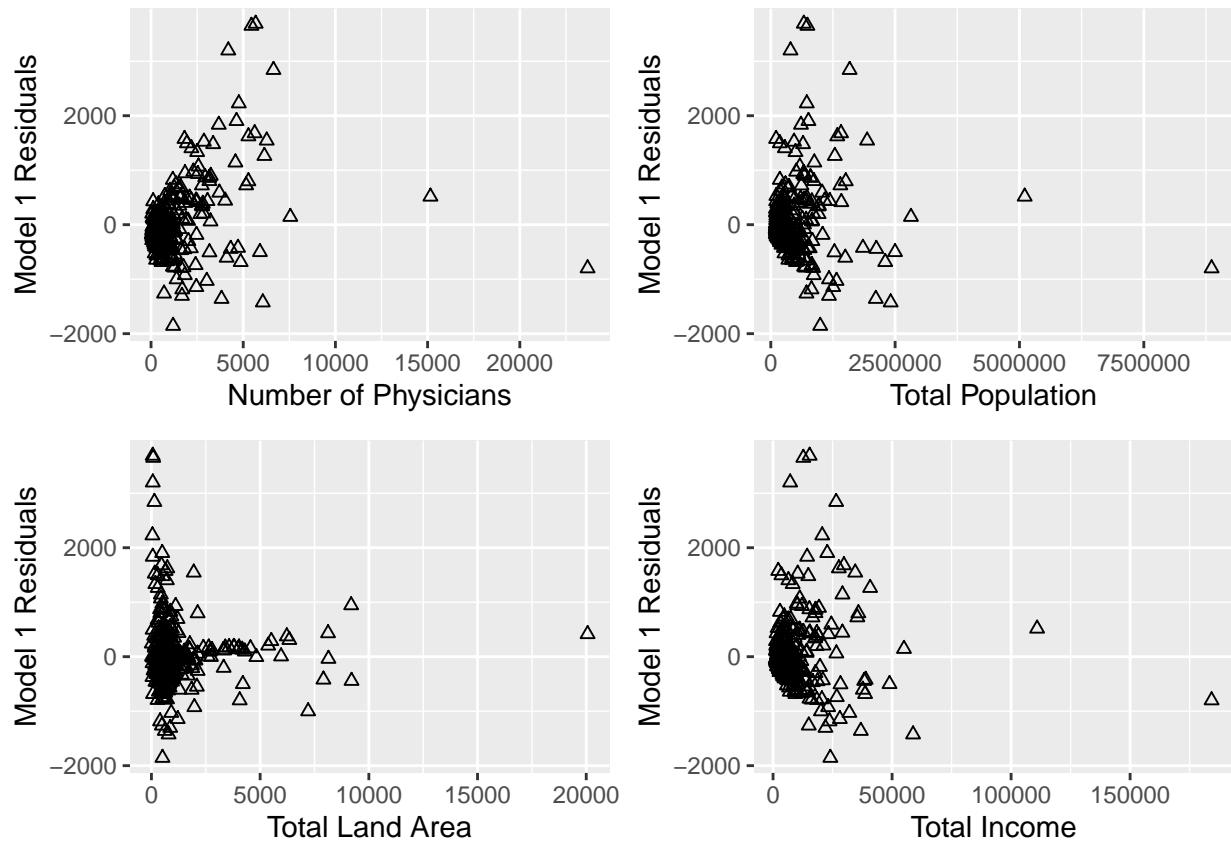
```
b1 <- ggplot(data = CDI, mapping = aes(x = num_physicians, y = model_1_resid)) + geom_point(shape = 24)

b2 <- ggplot(data = CDI, mapping = aes(x = total_pop, y = model_1_resid)) + geom_point(shape = 24) + la

b3 <- ggplot(data = CDI, mapping = aes(x = land_area, y = model_1_resid)) + geom_point(shape = 24) + la

b4 <- ggplot(data = CDI, mapping = aes(x = total_income, y = model_1_resid)) + geom_point(shape = 24) + la

figure2 <- ggarrange(b1,b2,b3,b4,
                      ncol = 2, nrow = 2); figure2
```



Model 2 Diagnostics

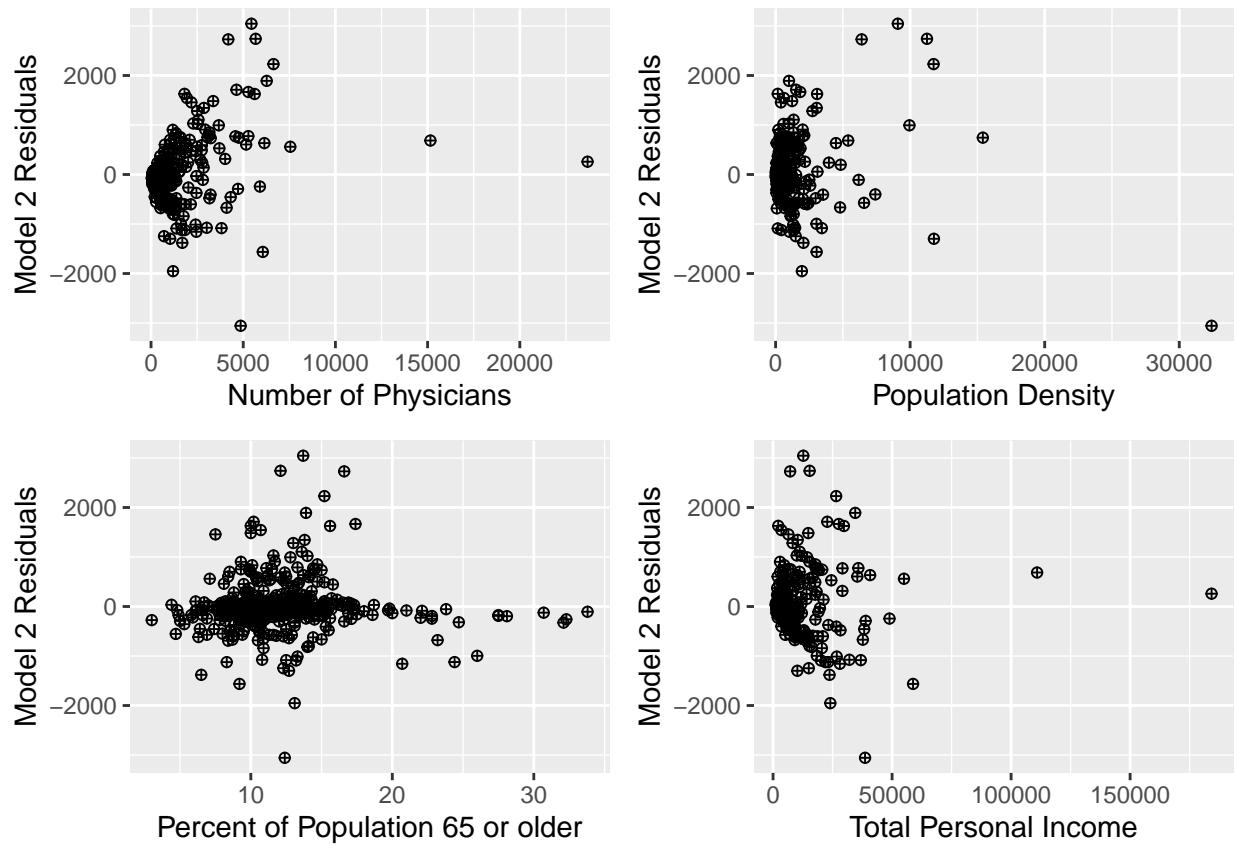
```
c1 <- ggplot(data = CDI, mapping = aes(x = num_physicians, y = model_2_resid)) + geom_point(shape = 10)

c2 <- ggplot(data = CDI, mapping = aes(x = pop_density, y = model_2_resid)) + geom_point(shape = 10) + ...

c3 <- ggplot(data = CDI, mapping = aes(x = percent_pop_65, y = model_2_resid)) + geom_point(shape = 10)

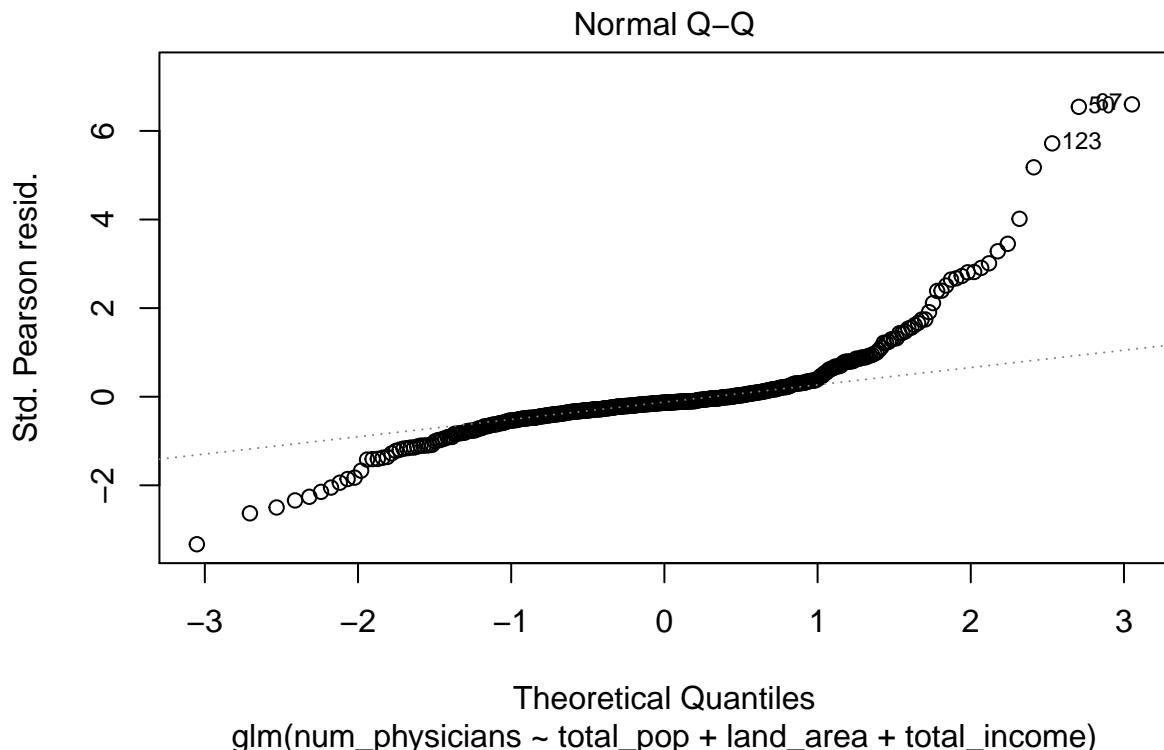
c4 <- ggplot(data = CDI, mapping = aes(x = total_income, y = model_2_resid)) + geom_point(shape = 10) + ...

figure3 <- ggarrange(c1,c2,c3,c4,
                      ncol = 2, nrow = 2); figure3
```



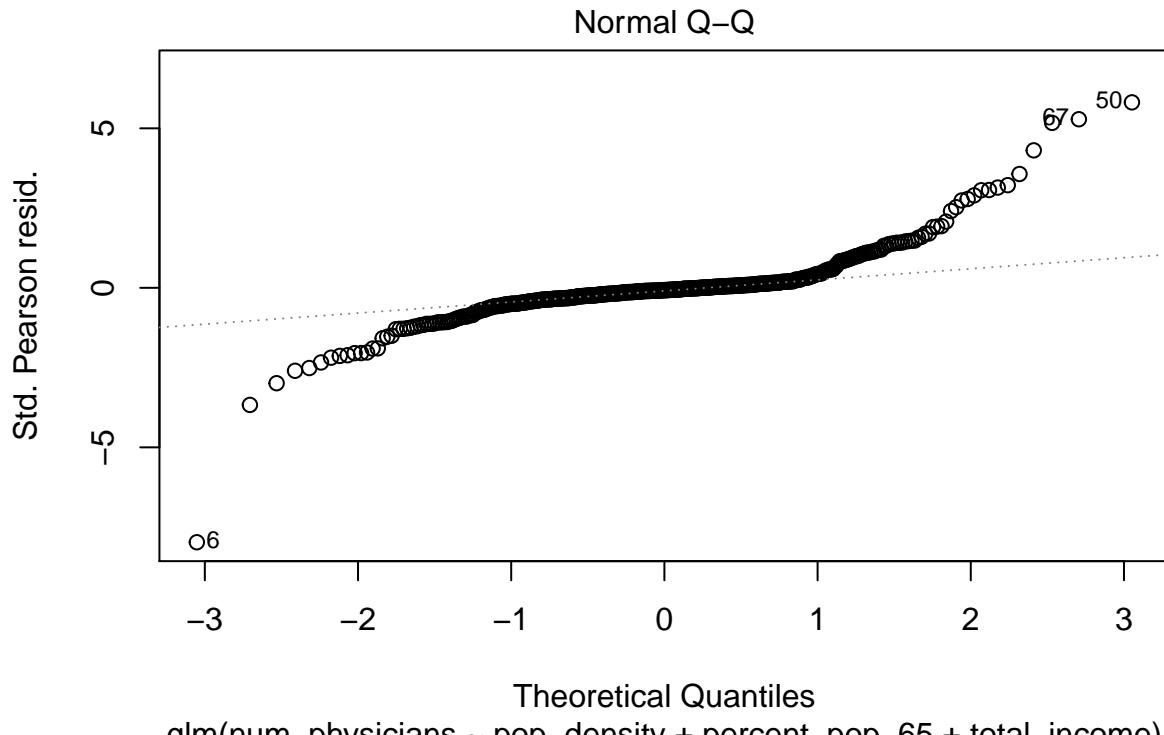
QQ-plot for Model 1

```
plot(model_1, which = 2)
```



QQ-plot for Model 2

```
plot(model_2, which = 2)
```



We will conduct the Shapiro-Wilk test for as our first test for normality. We will be testing the null hypothesis H_0 , that our residuals ε_i are normally distributed.

```
shapiro.test(CDI$model_1_resid)
```

Model I

```
##  
## Shapiro-Wilk normality test  
##  
## data: CDI$model_1_resid  
## W = 0.75754, p-value < 2.2e-16
```

```
shapiro.test(CDI$model_2_resid)
```

Model II

```
##
```

```

## Shapiro-Wilk normality test
##
## data: CDI$model_2_resid
## W = 0.80268, p-value < 2.2e-16

```

Our p-values < 0.05 indicates that we should reject the H_0 and conclude the residuals are not normally distributed.

We will conduct Kolmogorov-Smirnov Tests on both model residuals. Does ε_i and the number of physicians come from the same distribution? “A K-S Test quantifies a distance between the cumulative distribution function of the given reference distribution and the empirical distributions of given two samples, or between the empirical distribution of given two samples.”

```
ks.test(CDI$model_1_resid, CDI$num_physicians)
```

Model I

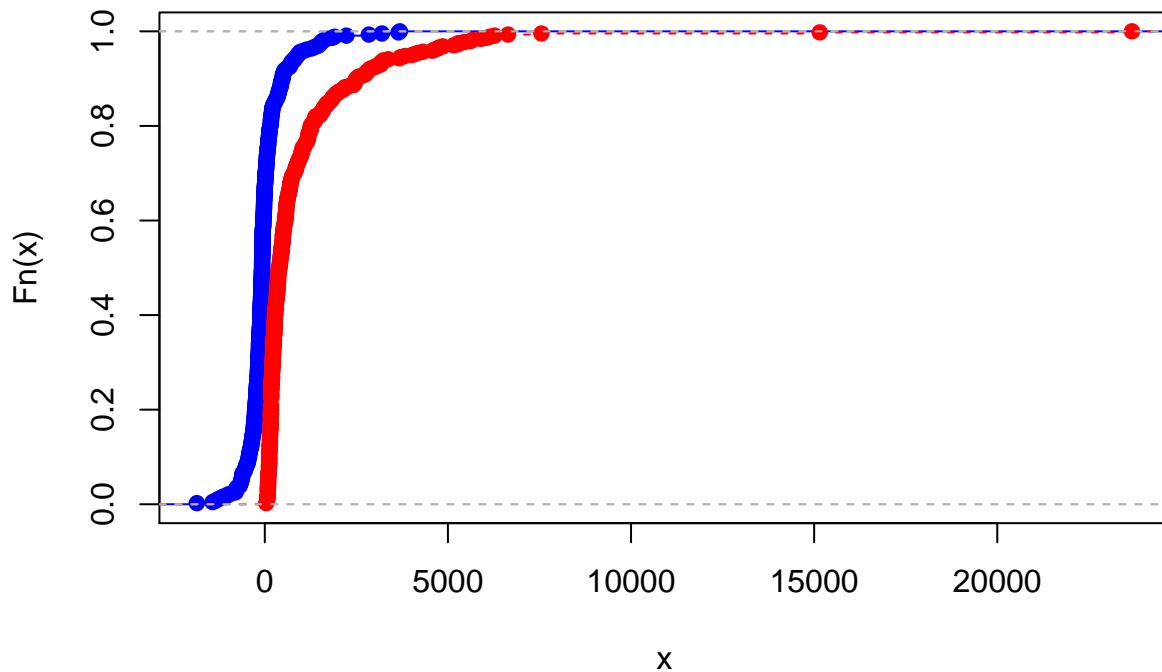
```

##
## Two-sample Kolmogorov-Smirnov test
##
## data: CDI$model_1_resid and CDI$num_physicians
## D = 0.73636, p-value < 2.2e-16
## alternative hypothesis: two-sided

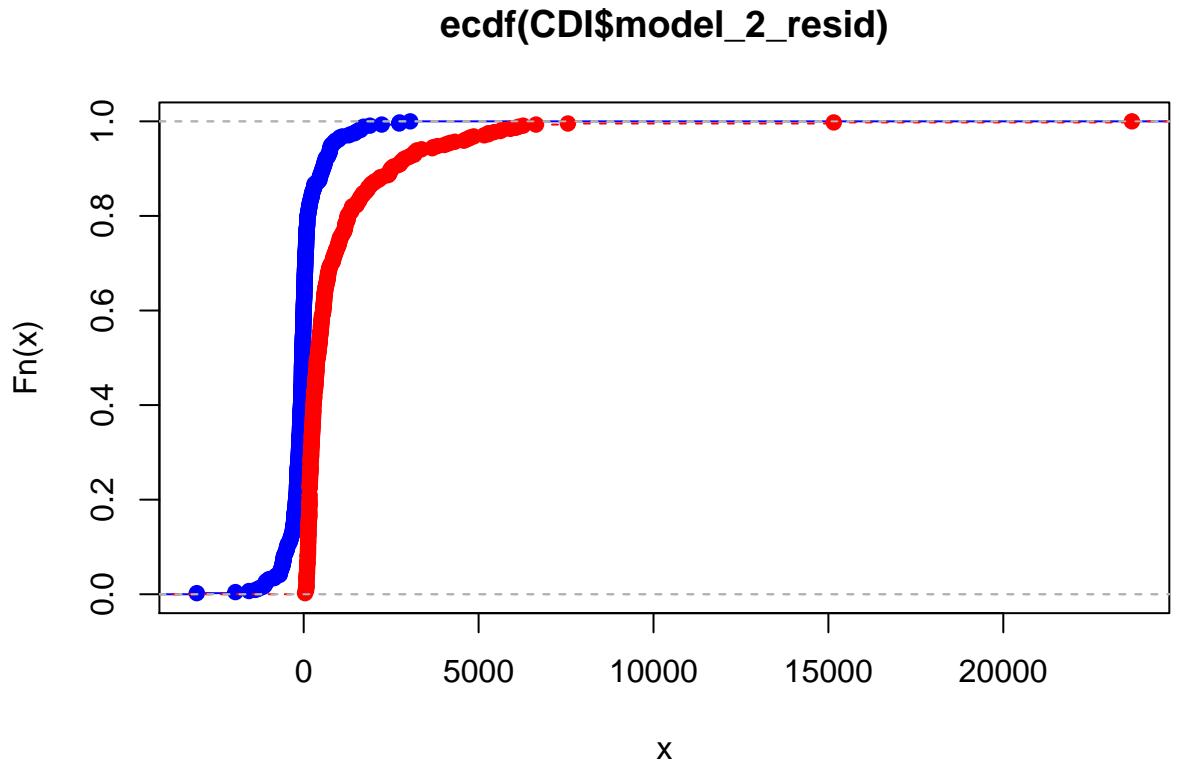
plot(ecdf(CDI$model_1_resid), xlim = range(c(CDI$model_1_resid,CDI$num)), col = "blue")
plot(ecdf(CDI$num_physicians), add = TRUE, col = "red", lty = "dashed")

```

ecdf(CDI\$model_1_resid)



```
plot(ecdf(CDI$model_2_resid), xlim = range(c(CDI$model_2_resid, CDI$num)), col = "blue")
plot(ecdf(CDI$num_physicians), add = TRUE, col = "red", lty = "dashed")
```



Model II

```
ks.test(CDI$model_2_resid, CDI$num_physicians)

##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  CDI$model_2_resid and CDI$num_physicians
## D = 0.74091, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

This test for both model residuals gives us a distance ≤ 0.73 . Our p-values are again less than our $\alpha = 0.05$, we fail our test for normality. We state with 95% confidence that the model residuals do not fit a normal distribution.

```
#knitr::purl("HW_1_Lupercio_Daniel.Rmd", "HW_1_Lupercio_Daniel.R", documentation = 2)
```

Homework 1

Saturday, February 27, 2021 8:02 PM

Exercise 6.22 For each of the following regression models, indicate whether it is a general linear regression model. If not, state whether it can be expressed in the form (6.7) by a suitable transformation:

$$(6.7) Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i$$

a. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 \log_{10} X_{i2} + \beta_3 X_{i1}^2 + \epsilon_i$

Yes, this is expressed in the form of (6.7).

b. $Y_i = \epsilon_i e^{(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)}$

From the start, this is not expressed in the form of (6.7).

Suppose $\log_e Y_i = \log(\epsilon_i \cdot e^{(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2)})$

$$\log_e Y_i = \log_e (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2) + \log_e (\epsilon_i)$$

Let $Y'_i = \log_e Y_i$ & $\epsilon'_i = \log_e (\epsilon_i)$

$$Y'_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^2 + \epsilon'_i$$

c. $Y_i = \log_{10} (\beta_1 X_{i1}) + \beta_2 X_{i2} + \epsilon_i$

This is already expressed in terms of (6.7). β_0 does not need to be present.

d. $Y_i = \beta_0 e^{(\beta_1 X_{i1})} + \epsilon_i = \log_e(\beta_0) + \beta_1 X_{i1} + \epsilon'_i$

This is expressed in terms of (6.7) and can not be modified by \log_e .

e. $Y_i = [1 + e^{(\beta_0 + \beta_1 X_{i1} + \epsilon_i)}]^{-1} = \frac{1}{[1 + e^{(\beta_0 + \beta_1 X_{i1} + \epsilon_i)}]}$

This not expressed in terms of (6.7).

$$\begin{aligned} \log_e(Y_i) &= Y'_i & Y'_i &= \log(1) - \log_e(1 + e^{(\beta_0 + \beta_1 X_{i1} + \epsilon_i)}) \\ &= 0 - (\log_e(1) + \log_e e^{(\beta_0 + \beta_1 X_{i1} + \epsilon_i)}) \\ &= 0 - 0 - (\beta_0 + \beta_1 X_{i1} + \epsilon_i) \\ &= -(\beta_0 + \beta_1 X_{i1} + \epsilon_i) \quad \checkmark \end{aligned}$$

6.24 Consider the multiple regression model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + \epsilon_i \quad i=1, \dots, n$$

where ϵ_i are indep. $N(0, \sigma^2)$

a. State the least squares criterion and derive the least squares normal equations.

normal equations (1.9) $\sum Y_i = n b_0 + b_1 \sum X_i$ sum of squared residuals: $\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$

$$\frac{\partial Q}{\partial \beta_0} = \sum \frac{\partial}{\partial \beta_0} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2})^2 = -2 \sum (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))$$

$$\frac{\partial Q}{\partial \beta_1} = \sum \frac{\partial}{\partial \beta_1} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))^2 = -2 \sum X_{i1} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))$$

$$\frac{\partial Q}{\partial \beta_2} = \sum \frac{\partial}{\partial \beta_2} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))^2 = -2 \sum X_{i1}^2 (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))$$

$$\frac{\partial Q}{\partial \beta_3} = \sum \frac{\partial}{\partial \beta_3} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))^2 = -2 \sum X_{i2} (Y_i - (\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2}))$$

$$\frac{\partial Q}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2})^2$$

$$\left[-\frac{1}{\sigma^2} \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2}) \right] = 0 \quad (i)$$

$$\sum X_{i1} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2}) = 0 \quad (ii)$$

$$\sum X_{i1}^2 (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2}) = 0 \quad (iii)$$

$$\sum X_{i2} (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2}) = 0 \quad (iv)$$

$$-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2})^2 = 0$$

$$\left[\frac{1}{2\sigma^4} \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2}) \right] = \frac{n}{2\sigma^2} \quad (v)$$

$$\frac{\sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i1}^2 - \beta_3 X_{i2})}{n} = \hat{\sigma}^2 \quad \blacksquare$$