# Matrices and CDI modeling

```
# tinytex::install_tinytex()
```

```
setwd("~/Documents/Hunter College/Spring 2021/Stat 707/HW")
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

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

```
# Make Z matrix from textbook
data <- matrix(c(7,4,16,3,21,8,33,41,7,49,5,31,42,33,75,28,91,55), nrow = 3, ncol = 6, byrow = TRUE); da
```

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

Assume that regression model (6.1) with independent normal error terms is appropriate. Using matrix methods, obtain (a) b; (b) e; (c) H; (d) SSR; (e) $s^2\{b\}$; (f) $\hat{Y}_h$ when $X_{h1} = 10$, $X_{h2} = 30$; (g) $s^2\{\hat{Y}_h\}$ when $X_{h1} = 10$, $X_{h2} = 30$.

## (a) Estimated coefficients

$b = (X'X)^-1 * X'Y$

```
# (X'X)^-1 times X'Y
b <- X_tX_inv %*% X_tY ; b
```

```
##            [,1]
## [1,] 33.9321033
## [2,]  2.7847614
## [3,] -0.2644189
```

```
# Check using Linear model

# Transpose matrix
t_data <- t(data)
Z <- as.data.frame(t_data)

lm <- lm(V3 ~ V1 + V2, data = Z); lm
```

```
## 
## Call:
## lm(formula = V3 ~ V1 + V2, data = Z)
## 
## Coefficients:
## (Intercept)           V1           V2  
##     33.9321       2.7848      -0.2644
```

## (b) Vector of residuals

$e = Y - Xb$

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

```
##             [,1]
## [1,] -2.69960842
## [2,] -1.22997279
## [3,] -1.63735316
## [4,] -1.32985996
## [5,] -0.08999801
## [6,]  6.98679233
```

## (c) Hat matrix

$H = X(X'X)^-1 * X'$

```
H <- X %*% X_tX_inv %*% X_t; H
```

```
##             [,1]        [,2]        [,3]       [,4]        [,5]       [,6]
## [1,]  0.23143293  0.25167585  0.21178735  0.1488684 -0.05475543 0.21099091
## [2,]  0.25167585  0.31240459  0.09437844  0.2662773 -0.14787283 0.22313666
## [3,]  0.21178735  0.09437844  0.70442026 -0.3191744  0.10446672 0.20412159
## [4,]  0.14886839  0.26627729 -0.31917435  0.6142563  0.14143492 0.14833743
## [5,] -0.05475543 -0.14787283  0.10446672  0.1414349  0.94039955 0.01632707
## [6,]  0.21099091  0.22313666  0.20412159  0.1483374  0.01632707 0.19708635
```

## (d) Sum of squares of regression

$SSR = (b'X' - (1/n) * Y'J)Y$

```
# Multiply whole thing by Y
SSR <- a %*% Y; SSR
```

```
##          [,1]
## [1,] 3009.926
```

```
# Check using linear model
anova(lm)
```

```
## Analysis of Variance Table
##
## Response: V3
##           Df  Sum Sq Mean Sq  F value  Pr(>F)
## V1         1 3004.41 3004.41 145.2027 0.00123 **
## V2         1    5.51    5.51   0.2664 0.64140
## Residuals  3   62.07   20.69
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# sum squared V1 + V2 should add up to 3009.926
```

## (e) Variance-covariance matrix of estimated coefficients

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

```
# SSE = Y'Y - b'X'Y; (sum of squares of error)
SSE <- (Y_t %*% Y) - (b_t %*% X_t %*% Y); SSE
```

```
##          [,1]
## [1,] 62.07354
```

```
# SSE should equal the residuals sum sq of the anova test

# MSE = SSE/(n-p); (mean square error)
MSE <- SSE / (6-3); MSE
```

```
##           [,1]
## [1,] 20.69118
```

```
# MSE should equal the residuals mean sq of the anova test
```

```
# s^2{b}
s_sq_b <- 20.69118 * X_tX_inv; s_sq_b
```

```
##            [,1]        [,2]        [,3]
## [1,] 715.47116 -34.1589176 -13.5949375
## [2,] -34.15892   1.6616664   0.6440674
## [3,] -13.59494   0.6440674   0.2624678
```

### (f) Point estimate

$\hat{Y}_h = X_h' b$ when $X_{h1} = 10$, $X_{h2} = 30$

```
Yhat_h <- X_h_t %*% b; Yhat_h
```

```
##           [,1]
## [1,] 53.84715
```

### (g) Estimated variance

$s^2\{\hat{Y}_h\} = X_h' * s^2\{b\} * X_h$ when $X_{h1} = 10$, $X_{h2} = 30$

```
s_sq_Yhat_h <- X_h_t %*% s_sq_b %*% X_h; s_sq_Yhat_h
```

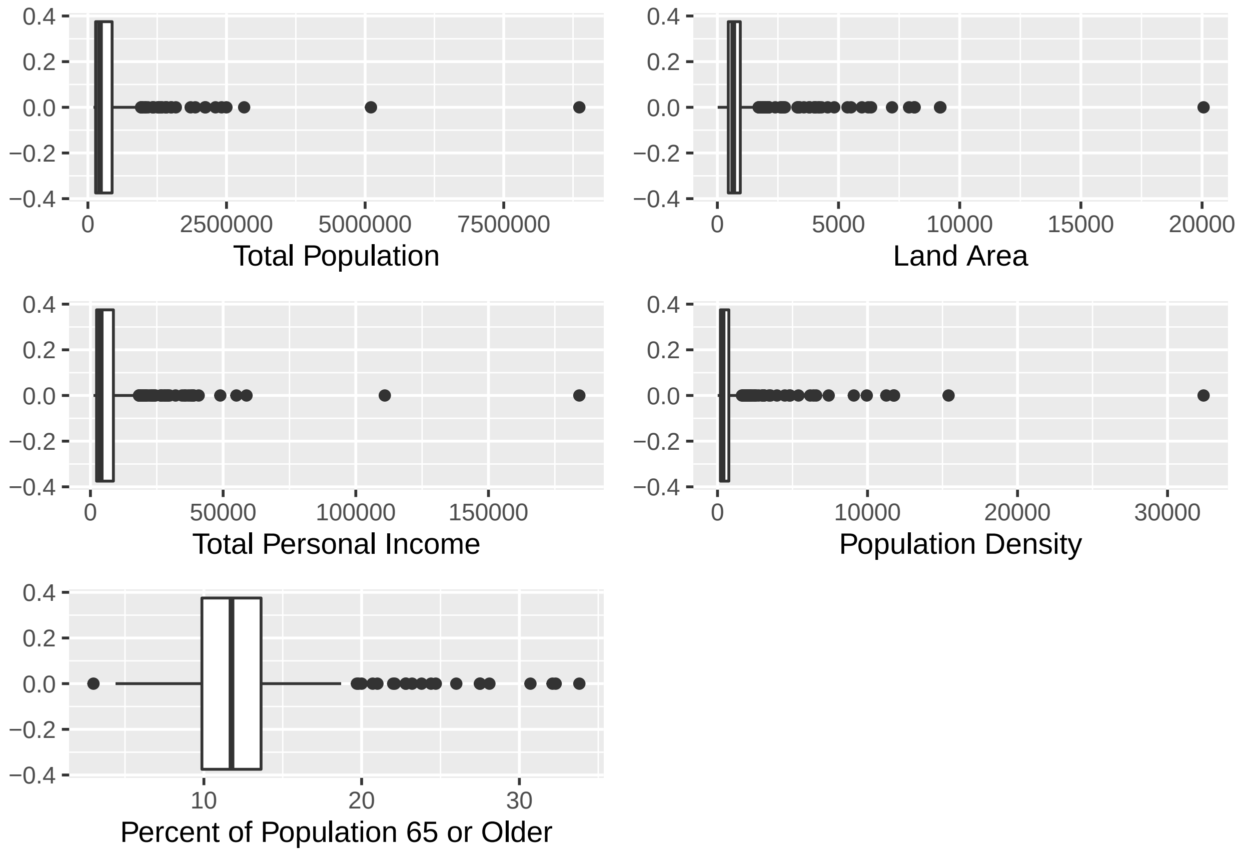
```
##          [,1]
## [1,] 5.42462
```

**6.28. Refer to the CDI data set in Appendix C.2. 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 (X1), land area (X2), and total personal income (X3). Proposed model II includes as predictor variables population density (X1, total population divided by land area), percent of population greater than 64 years old (X2), and total personal income (X3).**

```r
# import county demographic information (CDI)
CDI <- read.csv("CDI_Data.csv", header = F)

names(CDI) <- c("ID", "county", "state", "land_area", "total_pop", "precent_pop_18_34", "percent_pop_65

# create X1 variable, total population divided by land area
CDI$pop_density <- CDI$total_pop/CDI$land_area
```

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

```r
par(mfrow=c(2,2))

p1 <- ggplot(data = CDI, mapping = aes(x = total_pop)) + geom_boxplot() + labs(x = "Total Population")
p2 <- ggplot(data = CDI, mapping = aes(x = land_area)) + geom_boxplot() + labs(x = "Land Area")
p3 <- ggplot(data = CDI, mapping = aes(x = total_income)) + geom_boxplot() + labs(x = "Total Personal In
p4 <- ggplot(data = CDI, mapping = aes(x = pop_density)) + geom_boxplot() + labs(x = "Population Density
p5 <- ggplot(data = CDI, mapping = aes(x = percent_pop_65)) + geom_boxplot() + labs(x = "Percent of Popu

grid.arrange(p1, p2, p3, p4, p5, ncol = 2)
```

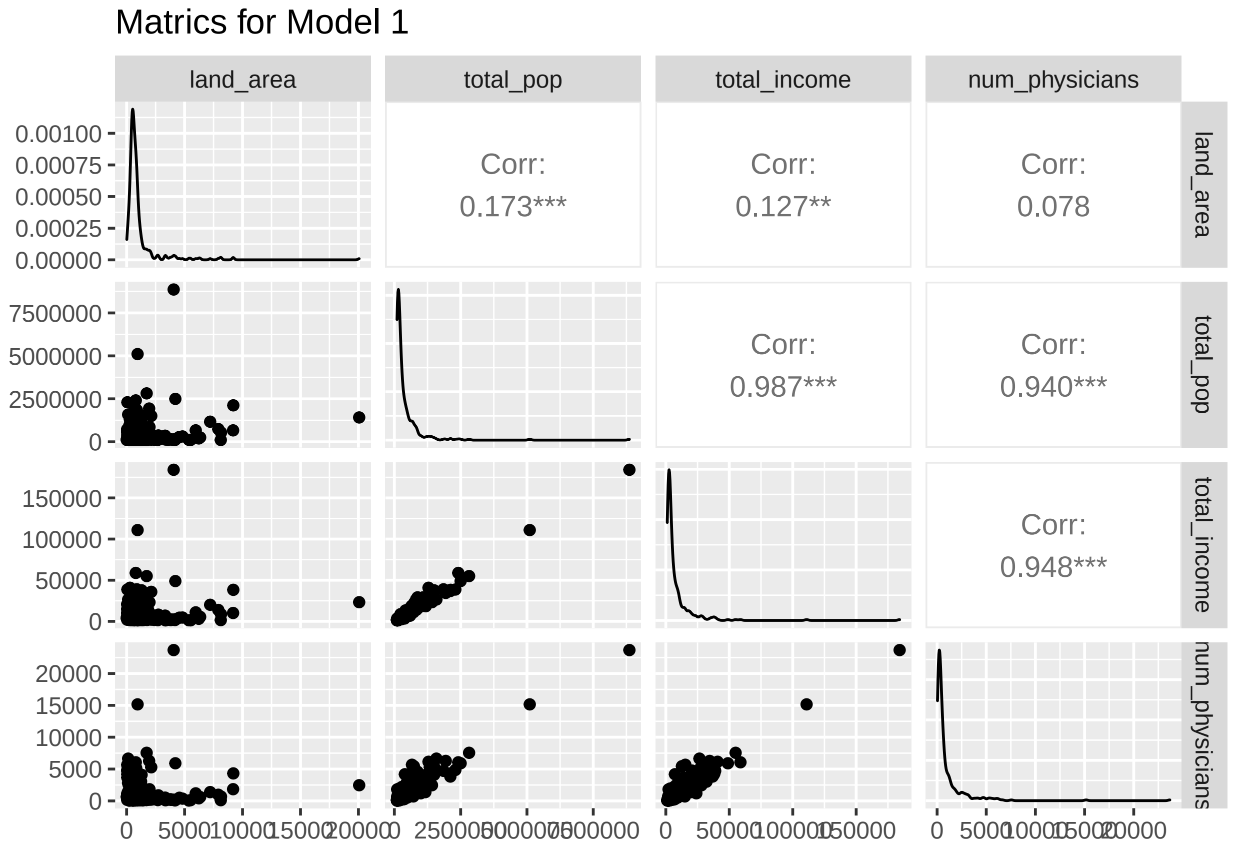


The boxplots above show us that total population, land area, total personal income, and population density are right skewed. Majority of the data for these predictor variables are concentrated to the left of the graph, with many outliers to the right. Only the predictor variable that has a more normal distribution of data is percent of population 65 or older.

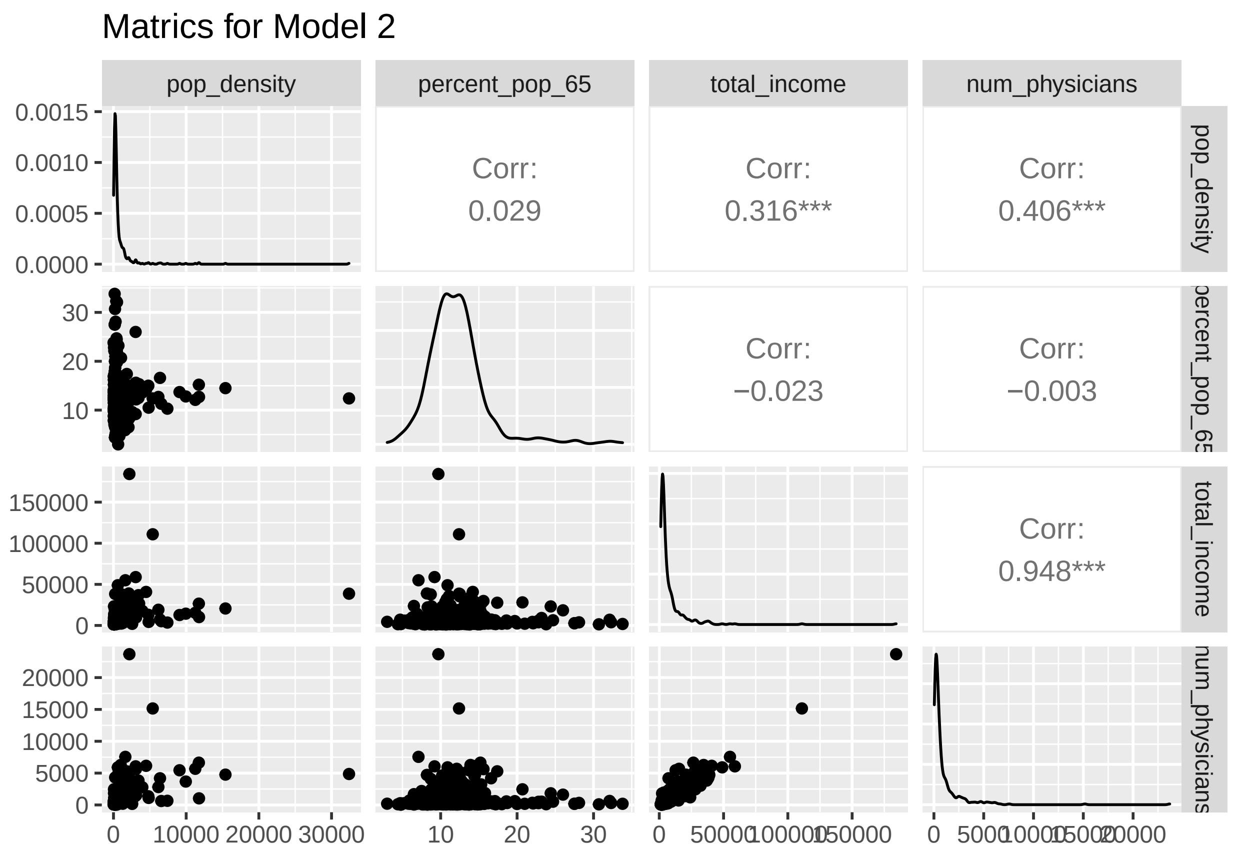
### b. Obtain the scatter plot matrix and the correlation matrix for each proposed model. Summarize the information provided.

Matrices for Model 1

```
# Combined scatter plot matrix, density plot, and correlation matrix

# Matrices for Model 1
ggpairs(CDI, columns = c(4,5,16,8),  title = "Matrics for Model 1")
```



```
# Matrices for Model 2
ggpairs(CDI, columns = c(18,7,16,8), title = "Matrics for Model 2")
```



The pairwise scatter plot and correlation matrix allow us to see the relationship between any two variables from the CDI dataset. We see that for model 1, we see a highly positive linear correlation between total population and total income, total population and number of active physicians, and total income and number of active physicians. For model 2, we see a highly positive linear correlation between total income and number of active physicians. There are minor positive correlations between number of active physicians and population density as well as total income and population density.

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

```
# For model 1
m1_mod <- glm(num_physicians ~ total_pop + land_area + total_income, data = CDI); m1_mod
```

```
##
## Call:  glm(formula = num_physicians ~ total_pop + land_area + total_income,
##     data = CDI)
##
## Coefficients:
##  (Intercept)     total_pop     land_area  total_income
##   -1.332e+01     8.366e-04    -6.552e-02     9.413e-02
##
## Degrees of Freedom: 439 Total (i.e. Null);  436 Residual
## Null Deviance:       1.406e+09
## Residual Deviance: 136900000     AIC: 6824
```

```r
# For model 2
m2_mod <- glm(num_physicians ~ pop_density + percent_pop_65 + total_income, data = CDI); m2_mod
```

```
##
## Call:  glm(formula = num_physicians ~ pop_density + percent_pop_65 +
##     total_income, data = CDI)
##
## Coefficients:
##    (Intercept)     pop_density  percent_pop_65    total_income
##     -170.57422         0.09616         6.33984         0.12657
##
## Degrees of Freedom: 439 Total (i.e. Null);  436 Residual
## Null Deviance:       1.406e+09
## Residual Deviance: 124100000     AIC: 6781
```

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

```r
m1_r_sq <- with(summary(m1_mod), 1 - deviance/null.deviance); m1_r_sq
```

```
## [1] 0.9026432
```

```r
m2_r_sq <- with(summary(m2_mod), 1 - deviance/null.deviance); m2_r_sq
```

```
## [1] 0.9117491
```

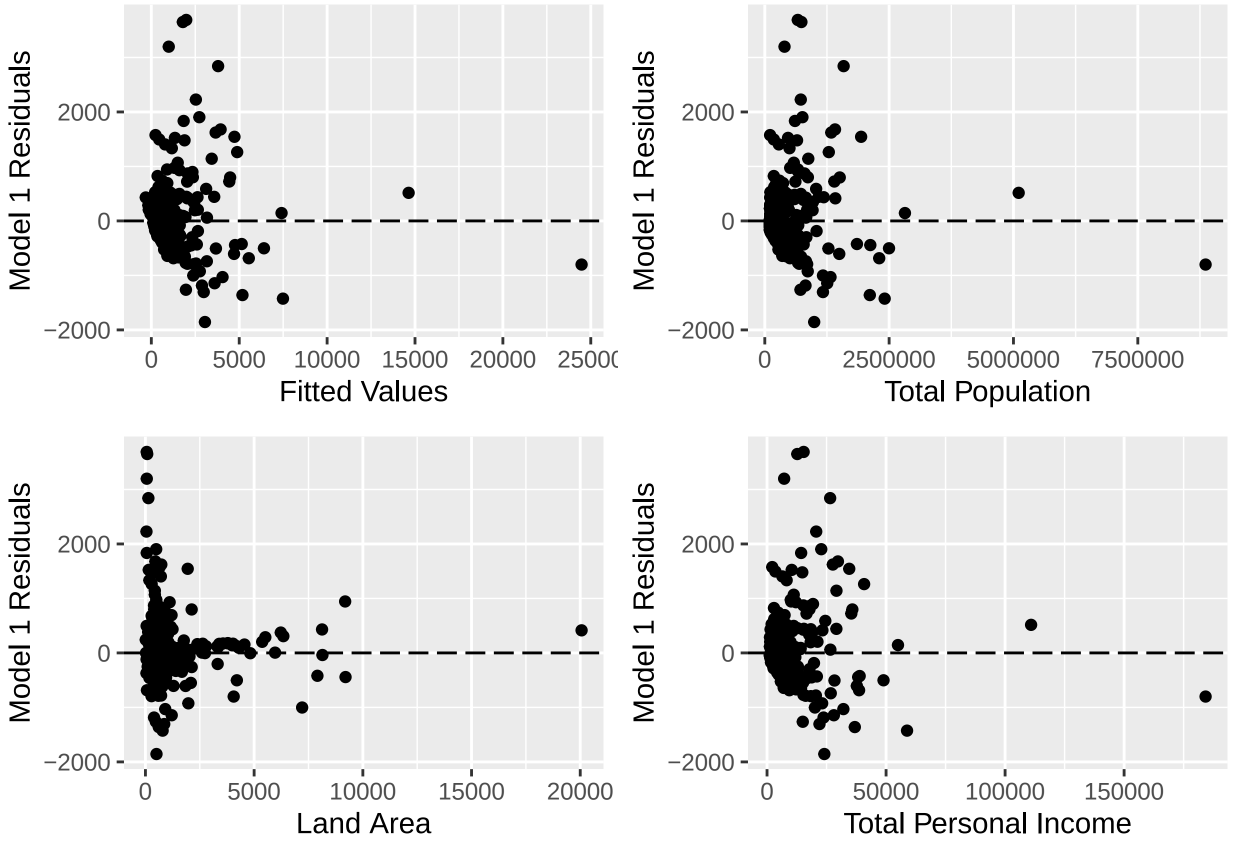
There is clearly no preferable model in terms of $R^2$ because both models have similar values. The data from both models are very close to the fitted regression line.

### e. For each model, obtain the residuals and plot them against Y, and each of the three predictor variables. Also prepare a normal probability plot for each of the two fitted models. Interpret your plots and state your findings. Is one model clearly preferable in terms of appropriateness?

Notes: for part (e), except replace the normal probability plot, with the following (i) Produce a QQ-plot (ii) Perform 2 tests for normality to determine if the residuals are normally distributed. Can you identify 2 tests for normality?

Residuals for Model 1

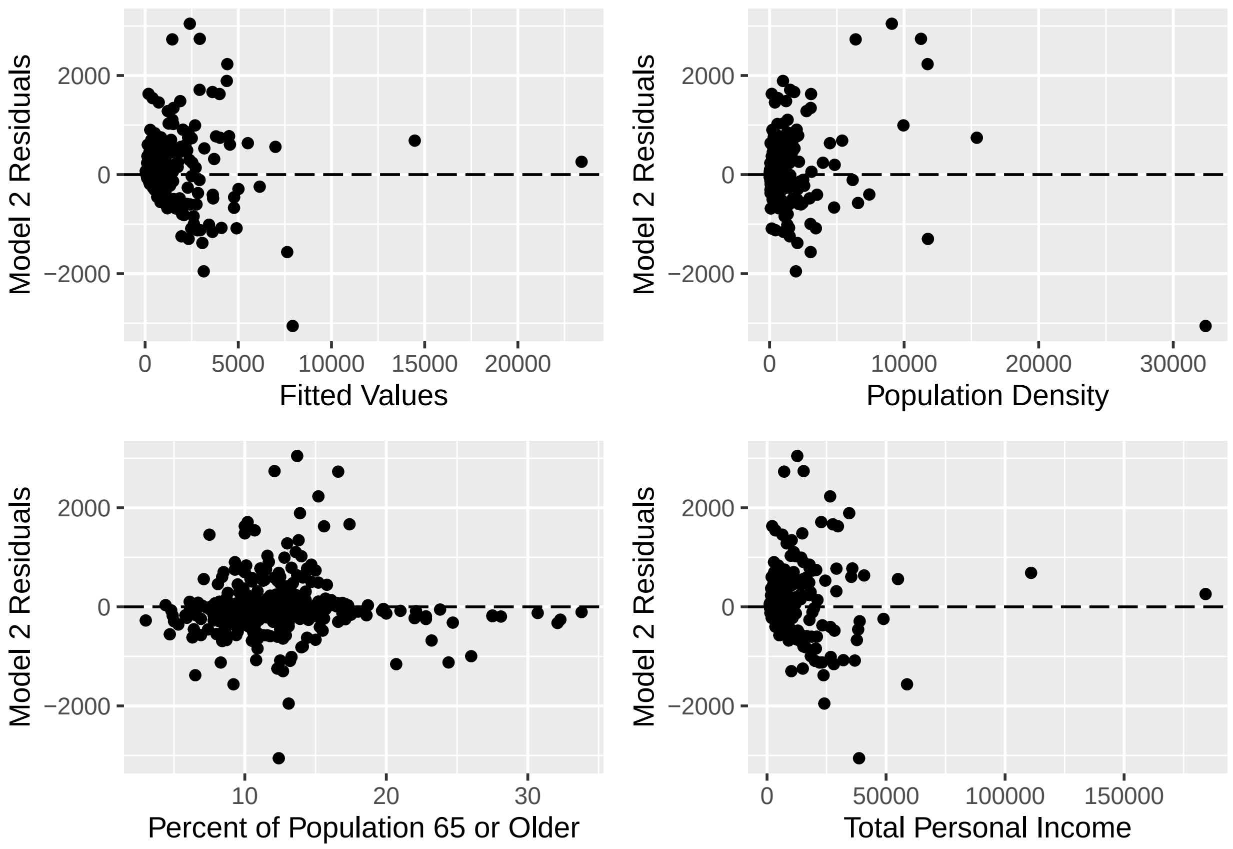
```r
CDI$m1_res <- residuals(m1_mod)
CDI$m1_pred <- predict(m1_mod)

par(mfrow=c(2,2))

p1 <- CDI %>%
  ggplot(aes(x = m1_pred, y = m1_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Fitted Values", y = "Model 1 Residuals")

p2 <- CDI %>%
  ggplot(aes(x = total_pop, y = m1_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Total Population", y = "Model 1 Residuals")

p3 <- CDI %>%
  ggplot(aes(x = land_area, y = m1_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Land Area", y = "Model 1 Residuals")

p4 <- CDI %>%
  ggplot(aes(x = total_income, y = m1_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Total Personal Income", y = "Model 1 Residuals")

grid.arrange(p1, p2, p3, p4, ncol = 2)
```



In a residual analysis, we want the data points to follow the dotted line. However, all of our residual plots for model 1 against the fitted values and predictors are skewed to the right. However, we should take into consideration that our original data was also skewed.

Residuals for Model 2

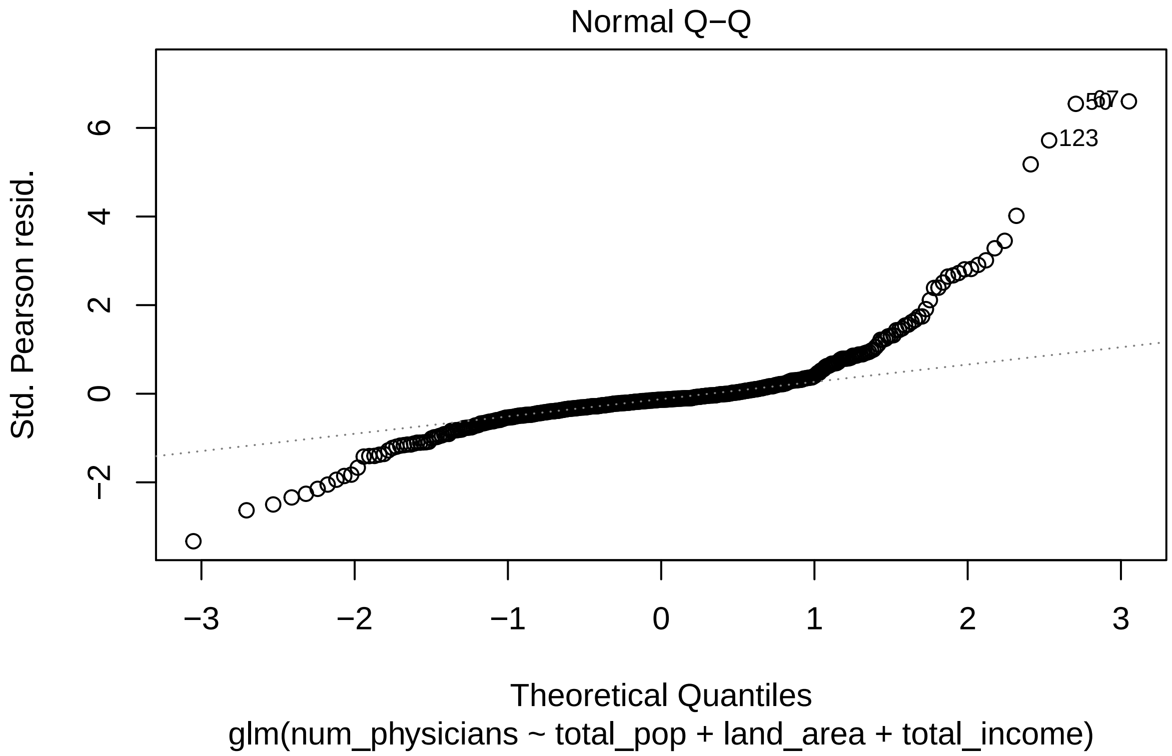
```r
CDI$m2_res <- residuals(m2_mod)
CDI$m2_pred <- predict(m2_mod)

par(mfrow=c(2,2))

p1 <- CDI %>%
  ggplot(aes(x = m2_pred, y = m2_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Fitted Values", y = "Model 2 Residuals")

p2 <- CDI %>%
  ggplot(aes(x = pop_density, y = m2_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Population Density", y = "Model 2 Residuals")


p3 <- CDI %>%
  ggplot(aes(x = percent_pop_65, y = m2_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Percent of Population 65 or Older", y = "Model 2 Residuals")

p4 <- CDI %>%
  ggplot(aes(x = total_income, y = m2_res)) +
  geom_point() +
  geom_hline(yintercept = 0, linetype = "longdash") +
  labs(x = "Total Personal Income", y = "Model 2 Residuals")

grid.arrange(p1, p2, p3, p4, ncol = 2)
```



Our residual plots for model 2 against the fitted values, population density, and total personal income are similar to the residual plots in model 1. However, the residuals seem uncorrelated with the predictor percent of population 65 or older because the data points are evenly distributed.

Normality Plots for Model 1

```r
# QQ-plot
plot(m1_mod,which = 2)
```



```
# Shapiro-Wilk test
shapiro.test(CDI$m1_res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CDI$m1_res
## W = 0.75754, p-value < 2.2e-16
```

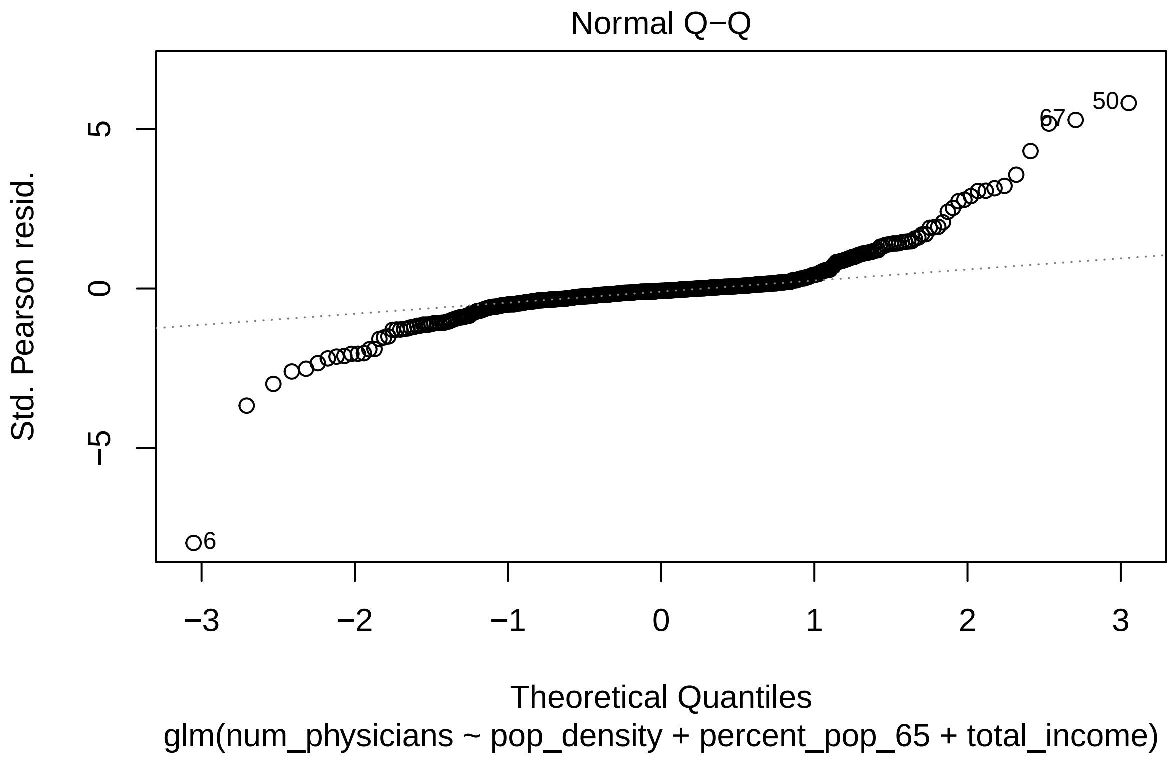
```
# Kolmogorov-Smirnov test
ks.test(CDI$m1_res,CDI$num_physicians)
```

```
## Warning in ks.test(CDI$m1_res, CDI$num_physicians): p-value will be approximate
## in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  CDI$m1_res and CDI$num_physicians
## D = 0.73636, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Normality Plots for Model 2

```
# QQ-plot
plot(m2_mod,which = 2)
```



```
# Shapiro-Wilk test
shapiro.test(CDI$m2_res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CDI$m2_res
## W = 0.80268, p-value < 2.2e-16
```

```
# Kolmogorov-Smirnov test
ks.test(CDI$m2_res,CDI$num_physicians)
```

```
## Warning in ks.test(CDI$m2_res, CDI$num_physicians): p-value will be approximate
## in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  CDI$m2_res and CDI$num_physicians
## D = 0.74091, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

The normal QQ plots assume that the errors are normally distributed, and if they are it should follow a normal distribution dotted line. The normal QQ plots for both model 1 and 2 deviate from the dotted line at both tails, suggesting some skewness. However, model 1 deviates a lot quicker from normal distribution on the right tail compared to that of model 2.

The Shapiro-Wilk test examines if a model is normally distributed. It overlaps a normal curve over the observed distribution and computes the percentage of which our model overlaps with it. The null hypothesis is that the data are normally distributed. Since both model 1 and model 2 have a p-value of less than the alpha level of 0.05, we can reject the null hypothesis.

The Kolmogorov-Smirnov (K-S) test examines if scores are likely to follow a distribution, in this case a normal distribution. The D statistic measures absolute max distance between the cumulative distribution function of the observed and normal curves. The closer D is to 0 the more likely it is that the two samples were drawn from the same distribution. For both model 1 and 2, we get a D statistic of approximately 0.74. And given that they both also have a p-value of less than 0.05, we can reject the null hypothesis that either model is normally distributed.