STA 108 Project 2

Finding the Best Multiple Linear Regression Model:

Predicting the Number of Active Physicians in a County with Multiple Predictors

https://www.google.com/url?sa=i&url=https%3A%2F%2Fpatientengagementhit.com%2Ffeatures%2Fpreventing-physician-burnout-from-impacting-the-patient-experience&psig=AOvVaw0wDk98EzYGvv_-4JlP-tnB&ust=1685843664838000&source=images&cd=vfe&ved=0CAQQjB1qFwoTCMCw5Zj_pf8CFQAAAAAdAAAAABAJ

Jasper Dong
Allison Peng

Amy Kim
STA 108

I) Introduction:

The dataset, CDI2, consists of numerical data that consists of 7 variables, with a sample size of 440 observations. Population (X_1) is the estimated total population, Income (X_2) is the total personal income in dollars, Physician is the number of professionally active non-federal physicians, Bed (X_3) is the total number of beds, cribs, and bassinets, Area (X_4) is the land area in square miles, Senior (X_5) is the percent of population aged 65 years old or older, Crime (X_6) is the total number of serious crimes, and. Our goal is to predict the number of active physicians in a county (Y) using a multiple linear regression model. With 6 variables available to predict Y, we will determine which variables are the most significant to build the best model. By comparing each model, we will determine which multiple regression model is the best to predict the number of physicians.

II) Summary:

We first conduct exploratory data analysis to inspect the individual data types of each variable, as well as the initial relationship between each predictor variable and the response variable. We observed five number summaries, means, and standard deviation values. We observe high standard deviations for each variable except for Senior, indicating that values are generally more spread out away from the mean. We also observe extremely high maximum data points compared to the third quartile of each variable, indicating the presence of possible outliers in our dataset.

We also analyze the relationship between each variable with one another by use of a correlation plot and a correlation matrix. From both the correlation plot and the correlation matrix, we see that there is a strong positive linear correlation between response Physician and predictors Population, Bed, Crime, and Income, indicating that these predictors could be the best to fit a multiple linear regression model. However, we also observe strong correlation between the predictor variables, indicating that there may be high multicollinearity present, which could make interpretation of coefficients used in the regression model more difficult.

III) Variable Selection:

We will be using base multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \qquad i = 1....n$$

We can use the extra sum of squares to determine the coefficients of partial determination to measure the effect of an added predictor variable in addition to other variables (in our case, the base model). Looking at the table with all of the coefficients of partial determination, we see that adding the variable Bed is the best for the multiple linear regression model, as the highest was $Y^2_{3|1,2} = 0.554$. Furthermore, a General Linear F Test to test the hypotheses H_0: □$_3$ = 0 vs. H_A: □$_3 \neq 0$, with a significance level of $\alpha = 0.0002$. Thus, we reject the null hypothesis and we conclude that the full model, or the model with the predictor Bed is a better fit.

IV) Model Comparison and Fit:

We are given two proposed models:

Proposed Model 1) $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \epsilon \qquad i = 1....n$

Proposed Model 2) $Y_i = \beta_0 + \beta_1 (X_1/X_4) + \beta_2 X_2 + \beta_3 X_3 + \epsilon \qquad i = 1....n$

Looking at the R^2, or coefficient of determination values for each model, we can see that Model 1 has a higher R^2 value of 0.955. Since our number of regressors is the same in both models, we do not need to use the R^2_{adj} value to compare. Thus, we will continue and inspect Model 1 further by performing diagnostics.

V) Model Diagnostics:

Using the Proposed Model 1, with Population, Income, and Bed as predictors for our best model, we performed model diagnostics to see if the assumptions of the Normal linear regression hold, as well as detecting the presence of outliers and high leverage points in our dataset. The assumptions we test for are:

1) Error terms are independent

2) Error terms are normally distributed

3) Error terms have constant variance (homoscedasticity)

When assessing independence of error terms, we created a residual index plot, and observed a pattern present in the plotted values: as index increases, errors become more centered around 0. This indicates that the error terms may not be independent of each other.

When assessing normality of error terms, we created a Normal Q-Q plot, and observed that plotted values do not follow the straight line. Although error terms seem to be symmetrical about the center, plotted values show greater deviations at the ends. This, in addition to the small Shapiro-Wilks p-value, suggests a non-Normal distribution.

When assessing constant variance of error terms, we created a plot of residuals vs. fitted values, and observed that residuals seem to cluster around smaller fitted values, before becoming less frequent as fitted values become larger. We also observe the presence of possible outliers in this plot, as there seem to be a few points that deviate from the mean. In addition, the small Fligner-Killeen Test p-value indicates there is not constant variance.

To detect for possible outliers, we looked for any studentized residuals greater than 3 and high leverage points. We found 12 possible outliers in our dataset with the studentized residuals and 55 possible high leverage points. We will be considering the studentized residuals as our outliers as it removes a lesser proportion of the dataset, at 2.727%.

We can also assess if the model has multicollinearity by looking at the VIFs, or variance inflation factors. The variables Population and Income have VIFs that are higher than 10, thus the Population and Income variable are likely correlated with other predictor variables in the model. A solution would be to remove the variables from the model.

Based on our diagnostics, we conclude that the assumptions of the Normal linear regression do not hold for our chosen model. However, we will continue to use this model for interpretation and prediction.

VI) Interpretation:

When estimated total population increases by 1 unit, we expect the number of physicians to decrease by -0.002 on average, holding all other predictor variables constant.

When total personal income increases by 1 unit, we expect the number of physicians to increase by 0.138 on average, holding all other predictor variables constant.

When total number of beds, cribs, and bassinets increases by 1 unit, we expect the number of physicians to increase by 0.487 on average, holding all other predictor variables constant.

We do not interpret our intercept of -89.105, as in reality, it would be impossible to have a negative number of physicians.

The R^2 value of 0.955 indicates that 95.5% is the proportionate reduction of total variation in Y associated with the use of the set of X variables, Population, Income, and Bed. The partial coefficient of determination $Y^2_{3|1,2} = 0.554$ indicates the proportion of decrease in SSE when the X_3 variable is added to the model with X_1 and X_2.

We are 95% confident that the estimated coefficient for population is between (-0.002, -0.001), the estimated coefficient for income is between (0.121, 0.155), and the estimated coefficient for bed is between (0.429, 0.544). Since all coefficients do not include 0, the estimates are significant.

VII) Prediction

$$\hat{Y}_i = -89.105 - 0.002X_1 + 0.138X_2 + 0.487X_5 \text{ , } X_1 = 394000, X_2 = 8500, X_5 = 300$$

Using the estimated coefficients, we found the predicted value for physicians to be 509.559, or 509 physicians.

VIII) Conclusion:

Based on our findings, we found the multiple linear regression model between response variable Physician and predictor variables Population, Income, and Bed to be our statistically best model, with its high R^2 compared to other models.

However, limitations of our model include strong multicollinearity, as well as the presence of outliers and high leverage points, that could make interpretation of our model difficult.

Tables/Plots

I) Data Preparation:

X_1 = Population

X_2 = Income

X_3 = Bed

X_4 = Area

X_5 = Senior

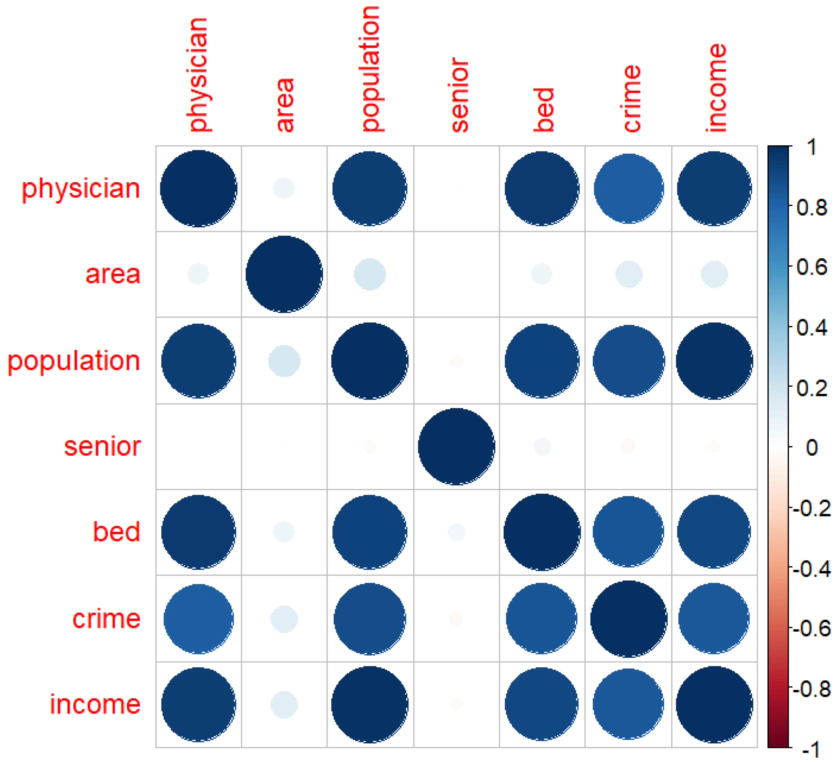
X_6 = Crime

i) Correlation matrix:

Numerical Matrix:

```
           physician  area  population  senior    bed  crime income
physician      1.000 0.078       0.940  -0.003  0.950  0.820  0.948
area           0.078 1.000       0.173   0.006  0.073  0.129  0.127
population     0.940 0.173       1.000  -0.029  0.924  0.886  0.987
senior        -0.003 0.006      -0.029   1.000  0.053 -0.035 -0.023
bed            0.950 0.073       0.924   0.053  1.000  0.857  0.902
crime          0.820 0.129       0.886  -0.035  0.857  1.000  0.843
income         0.948 0.127       0.987  -0.023  0.902  0.843  1.000
```

Visual Matrix:



ii) Numerical Summaries

a. Five Number Summary

	physician	area	population	senior	bed	crime	income
min	39.0	15.0	100043	3.000	92.0	563	1141
1st Qu.	182.8	451.2	139027	9.875	390.8	6220	2311
Median	401.0	656.5	217280	11.750	755.0	11820	3857
3rd Qu.	1036.0	946.8	436064	13.625	1575.8	26280	8654
Max	23677.0	20062.0	8863164	33.800	27700.0	688936	184230

b. Mean and Standard Deviations

	Physician	Area	Population	Senior	Bed	Crime	Income
Mean	988.0	1041.4	393011	12.170	1458.6	27112	7869
Standard Deviation	1789.75	1549.922	601987	3.993	2289.134	58237.51	12884.32

II) Variable Selection

i) Coefficients of Partial Determination

$Y^2_{3\|1,2}$	$Y^2_{4\|1,2}$	$Y^2_{5\|1,2}$	$Y^2_{6\|1,2}$
0.554	0.029	0.004	0.007

ii) Summary of R^2 Values

	Model 1	Model 2
R^2	0.955	0.912
R^2_{adj}	0.955	0.911

iii) Estimation of Coefficients

	Intercept	Population	Income	Bed
Model 1	-89.105	-0.002	0.138	0.487

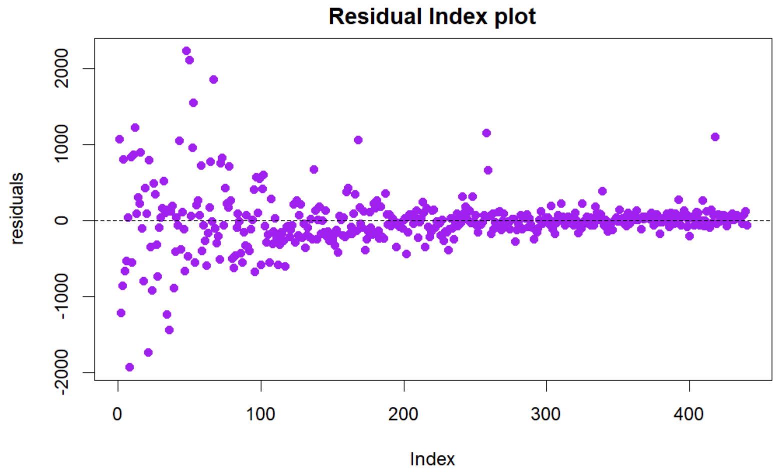
	Intercept	Population Density = Population/Area	Income	Senior
Model 2	-170.574	0.096	0.127	6.340

iv) Confidence Interval for Model 1

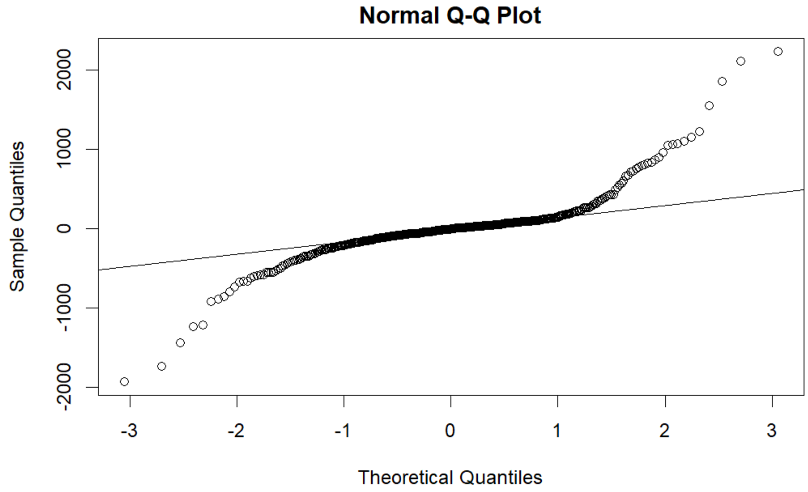
□hat$_1$ (Population)	(-0.002, -0.001)
□hat$_2$ (Income)	(0.121, 0.155)
□hat$_3$ (Bed)	(0.429, 0.544)

III) Model Diagnostics

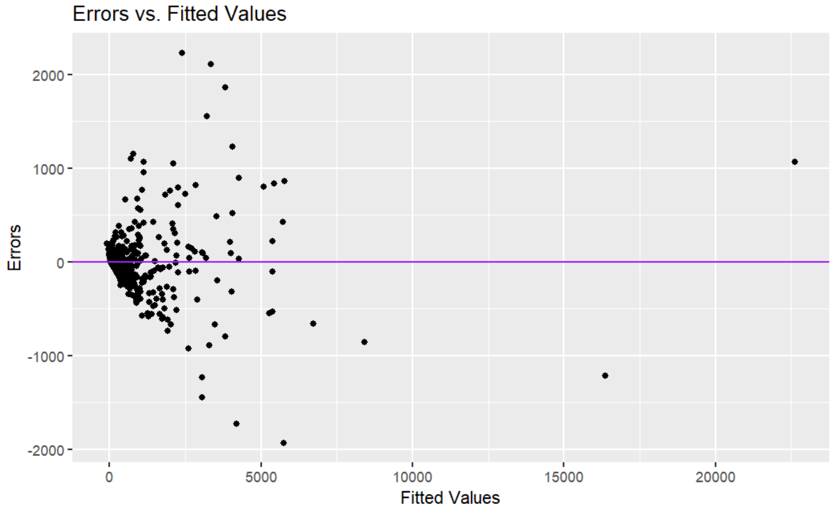
i) Assessing Independence



ii) Assessing Normality



iii) Assessing Constant Variance



iii) Hypothesis Tests for Constant Variance and Normality of Errors

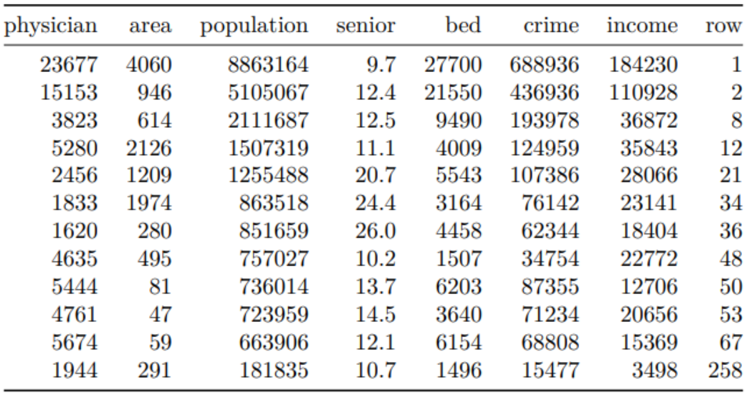
	Fligner-Killeen Test	Shapiro-Wilks Test
P-value	< 0.00000000000000022	< 0.00000000000000022

iv) VIFs - Variance Inflation Factor

	Population	Income	Bed
Variance Inflation Factor - VIF	49.365	38.877	6.977

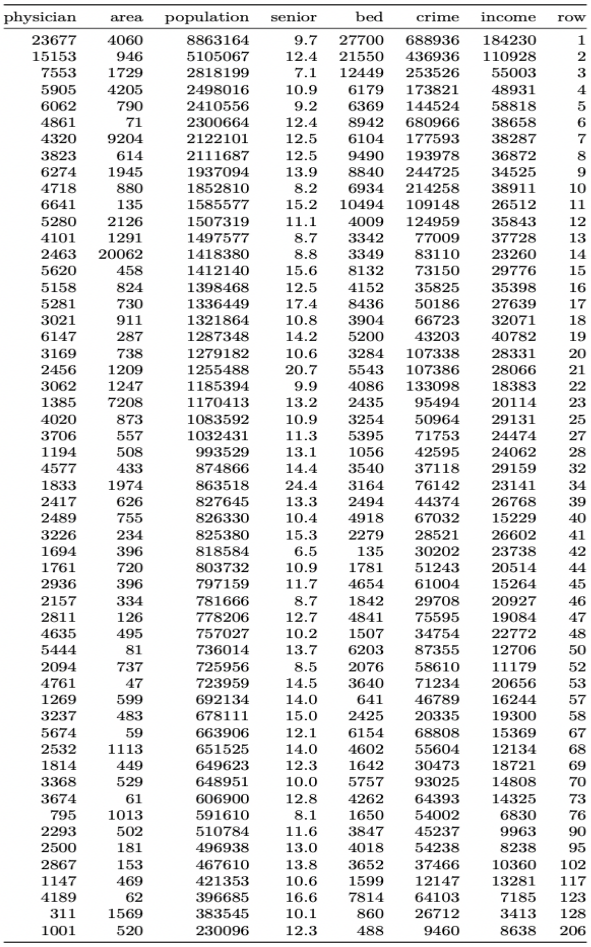
v) Outliers

i) Method: Studentized Residuals



physician	area	population	senior	bed	crime	income	row
23677	4060	8863164	9.7	27700	688936	184230	1
15153	946	5105067	12.4	21550	436936	110928	2
3823	614	2111687	12.5	9490	193978	36872	8
5280	2126	1507319	11.1	4009	124959	35843	12
2456	1209	1255488	20.7	5543	107386	28066	21
1833	1974	863518	24.4	3164	76142	23141	34
1620	280	851659	26.0	4458	62344	18404	36
4635	495	757027	10.2	1507	34754	22772	48
5444	81	736014	13.7	6203	87355	12706	50
4761	47	723959	14.5	3640	71234	20656	53
5674	59	663906	12.1	6154	68808	15369	67
1944	291	181835	10.7	1496	15477	3498	258

ii) Method: High Leverage Points



physician	area	population	senior	bed	crime	income	row
23677	4060	8863164	9.7	27700	688936	184230	1
15153	946	5105067	12.4	21550	436936	110928	2
7553	1729	2818199	7.1	12449	253526	55003	3
5905	4205	2498016	10.9	6179	173821	48931	4
6062	790	2410556	9.2	6369	144524	58818	5
4861	71	2300664	12.4	8942	680966	38658	6
4320	9204	2122101	12.5	6104	177593	38287	7
3823	614	2111687	12.5	9490	193978	36872	8
6274	1945	1937094	13.9	8840	244725	34525	9
4718	880	1852810	8.2	6934	214258	38911	10
6641	135	1585577	15.2	10494	109148	26512	11
5280	2126	1507319	11.1	4009	124959	35843	12
4101	1291	1497577	8.7	3342	77009	37728	13
2463	20062	1418380	8.8	3349	83110	23260	14
5620	458	1412140	15.6	8132	73150	29776	15
5158	824	1398468	12.5	4152	35825	35398	16
5281	730	1336449	17.4	8436	50186	27639	17
3021	911	1321864	10.8	3904	66723	32071	18
6147	287	1287348	14.2	5200	43203	40782	19
3169	738	1279182	10.6	3284	107338	28331	20
2456	1209	1255488	20.7	5543	107386	28066	21
3062	1247	1185394	9.9	4086	133098	18383	22
1385	7208	1170413	13.2	2435	95494	20114	23
4020	873	1083592	10.9	3254	50964	29131	25
3706	557	1032431	11.3	5395	71753	24474	27
1194	508	993529	13.1	1056	42595	24062	28
4577	433	874866	14.4	3540	37118	29159	32
1833	1974	863518	24.4	3164	76142	23141	34
2417	626	827645	13.3	2494	44374	26768	39
2489	755	826330	10.4	4918	67032	15229	40
3226	234	825380	15.3	2279	28521	26602	41
1694	396	818584	6.5	135	30202	23738	42
1761	720	803732	10.9	1781	51243	20514	44
2936	396	797159	11.7	4654	61004	15264	45
2157	334	781666	8.7	1842	29708	20927	46
2811	126	778206	12.7	4841	75595	19084	47
4635	495	757027	10.2	1507	34754	22772	48
5444	81	736014	13.7	6203	87355	12706	50
2094	737	725956	8.5	2076	58610	11179	52
4761	47	723959	14.5	3640	71234	20656	53
1269	599	692134	14.0	641	46789	16244	57
3237	483	678111	15.0	2425	20335	19300	58
5674	59	663906	12.1	6154	68808	15369	67
2532	1113	651525	14.0	4602	55604	12134	68
1814	449	649623	12.3	1642	30473	18721	69
3368	529	648951	10.0	5757	93025	14808	70
3674	61	606900	12.8	4262	64393	14325	73
795	1013	591610	8.1	1650	54002	6830	76
2293	502	510784	11.6	3847	45237	9963	90
2500	181	496938	13.0	4018	54238	8238	95
2867	153	467610	13.8	3652	37466	10360	102
1147	469	421353	10.6	1599	12147	13281	117
4189	62	396685	16.6	7814	64103	7185	123
311	1569	383545	10.1	860	26712	3413	128
1001	520	230096	12.3	488	9460	8638	206

R Appendix

```r
knitr::opts_chunk$set(echo = FALSE, comment = NA)
options(scipen = 999) #Remove the scientific notation
#### LOADING IN DATASET ####
library(readr)
CDI2 <- read_csv("CDI2.csv")
#### SUMMARY ####
  # Correlation matrix
library(corrplot)
round(cor(CDI2),3)
corrplot(cor(CDI2))
  # Numerical Summaries
summary(CDI2)
lapply(CDI2, sd)
#### MODEL COMPARISON & FIT ####
model_1 = lm(physician ~ population + income + bed, data = CDI2)
CDI2$pdensity = CDI2$population/CDI2$area
model_2 = lm(physician ~ pdensity + senior + income, data = CDI2)
summary(model_1)
summary(model_2)
#### DIAGNOSTICS ####
  # Assessing Independence
plot(model_1$residuals,main = "Residual Index plot",xlab = "Index",ylab = "residuals",pch = 19, col = "purple"
abline(h = 0, lty = 2)
  # Assessing Normality
    # Normal Q-Q Plot
qqnorm(model_1$residuals)
qqline(model_1$residuals)
    # Shapiro-Wilks Test
the.SWtest = shapiro.test(model_1$residuals)
the.SWtest
  # Assessing Constant Variance
    # Plotting Errors vs. Fitted Values
library(ggplot2)
CDI2$ei = model_1$residuals
CDI2$yhat = model_1$fitted.values
qplot(yhat, ei, data = CDI2) +  ggtitle("Errors vs. Fitted Values") + xlab("Fitted Values") +
  ylab("Errors") + geom_hline(yintercept = 0,col = "purple")
    # Formal Testing
Group = rep("Lower",nrow(CDI2))
Group[CDI2$physician < median(CDI2$physician)] = "Upper"
Group = as.factor(Group)
CDI2$Group = Group
the.FKtest= fligner.test(CDI2$ei, CDI2$Group)
#### OUTLIERS ####
  # Leverage
p = 4
h = hatvalues(model_1)
n = 440
leverage = which(h > (p+1)/n)
  # Studentized Residuals
sei = rstudent(model_1)
outliers = which(abs(sei) > 3)
  # Table of outliers and leverage points
CDI2 <- read_csv("CDI2.csv")
outlier_table = CDI2[outliers,]
outlier_table$row = outliers
leverage_table = CDI2[leverage,]
```

```r
leverage_table$row = leverage
knitr::kable(outlier_table)
knitr::kable(leverage_table)
```

```r
# find the best variables to include using the partial R^2
CDI2 <- read_csv("Downloads/CDI2.csv")
base_model <- lm(physician ~ population + income, data = CDI2)
model1 <- lm(physician ~ area + population + senior + bed + crime + income, data = CDI2)
model2 <- lm(physician ~ population + income + senior, data = CDI2)
model3 <- lm(physician ~ population + income + crime, data = CDI2)
model4 <- lm(physician ~ population + income + bed, data = CDI2)
model5 <- lm(physician ~ population + income + area, data = CDI2)

anova(base_model)

ybar = mean(CDI2$physician)
SSTO = sum((CDI2$physician - ybar)^2)
SSE = 140967081

# partial senior
SSR_senior_population_income = sum((fitted(model2) - ybar)^2)
SSR_population_income = sum((fitted(base_model) - ybar)^2)

# partial crime
SSR_crime_population_income = sum((fitted(model3) - ybar)^2)

# partial bed
SSR_bed_population_income = sum((fitted(model4) - ybar)^2)

# partial area
SSR_area_population_income = sum((fitted(model5) - ybar)^2)


# partial senior given population and income
#ssr(senior|population,income)/sse(population,income)
#ssr(senior|population,income) = ssr(senior,population,income)-ssr(population,income)
round((SSR_senior_population_income - SSR_population_income)/SSE,3)

# partial crime given popualtio and income
#ssr(crime|population,income)/sse(population,income)
#ssr(crime|population,income) = ssr(crime,population,income) - ssr(population,income)
round((SSR_crime_population_income - SSR_population_income)/SSE,3)

# partial bed given population and income
#ssr(bed|population,income)/sse(population,income)
#ssr(bed|population,income) = ssr(bed,population,income) - ssr(population,income)
round((SSR_bed_population_income - SSR_population_income)/SSE,3)

# partial area given pouplation and income
#ssr(area|population,income)/sse(population,income)
#ssr(area|population,income) = ssr(area,population,income) - ssr(population,income)
round((SSR_area_population_income - SSR_population_income)/SSE,3)




#
# we want to check if the model with bed is better than the base model
```

```
reduced_model <- lm(physician ~ population + income, data = CDI2)
full_model <- lm(physician ~ population + income + bed, data = CDI2)

# general linear f test
anova(reduced_model)
sse_reduced = 140967081
dfr = 437
anova(full_model)
sse_full = 62896949
dff = 436
CDI2
fstat = ((sse_reduced - sse_full)/(dfr - dff))/(sse_full/dff)
rejection_region <- qf(0.95, dfr-dff, 440-dfr)
pf(fstat, dfr-dff, 440-dfr, lower.tail = FALSE)



# model fitting
population_density = CDI2$population/CDI2$area
model_1 <- lm(physician ~ population + income + bed, data = CDI2)
model_2 <- lm(physician ~ population_density + income + senior, data = CDI2)

plot(model_1)
summary(model_1)
summary(model_2)
round(model_1$coefficients,3)
round(model_2$coefficients,3)

intercept = model_1$coefficients[1]
population = model_1$coefficients[2]
income = model_1$coefficients[3]
bed = model_1$coefficients[4]

# confidence intervals for each estimated coefficient
round(bed + qt(1-(0.05/2),440-3)*(0.0292), 3)
round(bed - qt(1-(0.05/2),440-3)*(0.0292), 3)
round(population + qt(1-(0.05/2),440-3)*(0.0002116),3)
round(population - qt(1-(0.05/2),440-3)*(0.0002116),3)
round(income + qt(1-(0.05/2),440-3)*(0.008773),3)
round(income - qt(1-(0.05/2),440-3)*(0.008773),3)

# prediction
Yhat = intercept + bed*(300) + population*(394000) + income*(8500)
Yhat

# find the VIFs for Model 1
library(caTools)
library(car)
vif(model_1)
```