How Social, Health, and Economic Factors Affect Average Income in the United States

Wei-Yu Tseng, Department of Statistics and Data Science

weiyut@stat.cmu.edu

Abstract

This paper will investigate the relationship between several social, health, and economic factors and how they affect the per capita income. Using exploratory data analysis conducted by R programming language, we compare per capita income to 12 variables across different categories such as. In addition, we take other potentially explanatory variables into consideration. We find that the per capita income is highly correlated to land area, percent of population aged 18 to 34, the number of active physicians, the percent of bachelor's degrees, the percent below poverty level, and the percent unemployment of a county. However, the analysis only captured the less than 15% of the counties data in the United States, and the data is outdated. We probably need additional up to date data to draw a better conclusion.

1 Introduction

Income inequality is always one of the top public concerns in the United States. The average income across different areas (county) seem to vary over a large range. Some social scientists argue that differences between counties' economic, health and social well-being lead to the issue. In this report, we will investigate the relationship between per capita income and several social, health, and economic factors.

In particular we will examine the following research topics:

1. Explore relationships between several social, health, and economic factors; and
2. Illustrate the effect of serious crimes on per capita income across different geographical regions; and
3. Develop regression models to predict average income per person (per capita income) in 1990(in dollars) from several social, health, and economic factors; and
4. Discuss the impact of the sample size and missing values.

2 Data

The data provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. The data were obtained from Kutner et al. (2005)[1] and are given in the file cdi.dat.

The variables in the data set are shown in Table 1.

Variable Number	Variable	Variable Name	Description
1	id	Identification number	1–440
2	county	County	County name
3	state	State	Two-letter state abbreviation
4	land.area	Land area	Land area (square miles)
5	pop	Total population	Estimated 1990 population
6	pop.18_34	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	pop.65_plus	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	doctors	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	hosp.beds	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	crimes	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	pct.hs.grad	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	pct.bach.deg	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	pct.below.pov	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	pct.unemp	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	per.cap.income	Per capita income	Per-capita income (i.e., average income per person) of 1990 CDI population (in dollars)
16	tot.income	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	region	Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

[1] Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005) Applied Linear Statistical Models, Fifth Edition. NY: McGraw-Hill/Irwin.

Table 1: Variable Definitions for the cdi.dat data set.

Summary statistics for the 13 quantitative variables are given in Table 2. **Skew** section and further EDA (see A.3) show that most variables are substantially skewed right, except Percent high school graduates (pct.hs.grad).

	n	mean	sd	median	min	max	range	skew	kurtosis
land.area	440	1041.41	1549.92	656.50	15.0	20062.0	20047.0	6.21	57.09
pop	440	393010.92	601987.02	217280.50	100043.0	8863164.0	8763121.0	8.13	96.81
pop.18_34	440	28.57	4.19	28.10	16.4	49.7	33.3	1.22	4.02
pop.65_plus	440	12.17	3.99	11.75	3.0	33.8	30.8	1.92	6.74
doctors	440	988.00	1789.75	401.00	39.0	23677.0	23638.0	6.56	66.81
hosp.beds	440	1458.63	2289.13	755.00	92.0	27700.0	27608.0	6.01	53.36
crimes	440	27111.62	58237.51	11820.50	563.0	688936.0	688373.0	7.89	78.32
pct.hs.grad	440	77.56	7.02	77.70	46.6	92.9	46.3	-0.73	1.55
pct.bach.deg	440	21.08	7.65	19.70	8.1	52.3	44.2	0.96	1.16
pct.below.pov	440	8.72	4.66	7.90	1.4	36.3	34.9	1.68	5.76
pct.unemp	440	6.60	2.34	6.20	2.2	21.3	19.1	1.78	6.22
per.cap.income	440	18561.48	4059.19	17759.00	8899.0	37541.0	28642.0	1.22	2.42
tot.income	440	7869.27	12884.32	3857.00	1141.0	184230.0	183089.0	7.68	88.10

Table 2: Summary Statistics for 13 continuous variables.

	unique values
county	373
state	48
region	4

Table 3: The number of unique values for categorical variables.

Summary statistics for 3 categorial variables are given in Table 3, 4, and 5. Table 3 and 4 shows that there are 48 states, California (CA) has the most counties involved in this data with 34 counties while West Virginia (WV) has the least counties included, with only one county recorded in the data.

	CA	FL	PA	TX	OH	NY	MI	NC	NJ	IL	IN	MA	SC	WI	MD	WA
Freq	34	29	29	28	24	22	18	18	18	17	14	11	11	11	10	10

	CO	GA	LA	VA	CT	MO	TN	AL	MN	OR	AZ	ME	KS	NH	OK	UT
Freq	9	9	9	9	8	8	8	7	7	6	5	5	4	4	4	4

	HI	KY	MS	NE	RI	AR	DE	NM	NV	DC	ID	MT	ND	SD	VT	WV
Freq	3	3	3	3	3	2	2	2	2	1	1	1	1	1	1	1

Table 4: Summary Statistics for variable State.

Counties in different states may share names, Table 3 shows that although the data was collected from 440 different counties, there are only 373 different names.

	NC	NE	S	W
Freq	108	103	152	77

Table 5: Summary Statistics for variable Geographic region.

Table 5 shows that 152 counties are in the South US, 108 counties are in North-Central US, 103 counties are in Northeast US, and 77 counties are in West US.

3 Methods

1. **Explore relationships between several social, health, and economic factors**

To find the relationships between several social, health, and economic factors, we use pairwise scatterplot and correlation coefficient (see A.3) to help us get some insights from the data.

2. Illustrate the effect serious crimes on per capita income across different geographical regions

To illustrate the effect of serious crimes on per capita income and whether it varies across different geographical regions, we considered regression models of per capita income on predictors Total serious crimes and Geographical region (see B.1.1) as well as similar model with an additional predictor being the interaction of two variables (see B.1.2). In additional, we redo the experiment again but replace the predictor Total number of serious crimes by per capita serious crimes (see B.2) to see whether the result changes. We chose our final models based on a summary of each regression analysis, an examination of residual diagnostic plots and the analysis of covariance between models.

3. Develop regression models to predict average income per person (per capita income) in 1990(in dollars) from several social, health, and economic factors.

To develop a regression model to predict per capita income, we considered multiple regression models and LASSO models using logarithmic and power transformations of variables on some of the continuous variables and an additional predictor per capita serious crime (see C). We chose our final model based on a summary of each regression analysis, an examination of residual diagnostic plots, and some reasonable understanding of each variable (see C.4).

4. Discuss the impact of the sample size and missing values.

To discuss the impact of missing data and sample size, we would prefer using external information to make comments.

4 Results

1. Explore relationships between several social, health, and economic factors

From the pairwise scatter plots and correlation coefficient (see A.3), we discovered that:

1. Total population(pop) is strongly and postively related to the Number of active physicians(doctors), the number of hospital beds(hosp.beds), Total serious crimes(crimes), and Total personal income(tot.income).
2. Percent of population aged 18–34 (pop.18_34) has a strong negative relationship with Percent of population aged 65 or older(pop.65_plus) and a mediocre postive relationship with the Percent bachelor's degrees (pct.bach.deg).
3. The Number of active physicians(doctors) is highly correlated to the Number of hospital bed(hosp.beds), Total serious crimes(crimes), and Total income (tot.income).
4. The Percent high school graduates(pct.hs.grad) is positively related to the Percent bachelor's degrees (pct.bach.deg) and the per capita income(per.cap.income), while negatively related to the Percent below poverty level(pct.below.pov) and the Percent unemployment(pct.unemp).

2. Illustrate the effect of serious crimes on per capita income across different geographical regions

We considered multiple regression with categorical variable Geographical Region and continuous variable Total serious crimes for section 1, per capita serious crimes for section 2.

Section 1 – Total serious crimes

Both models we tested have similar diagnostics plots (see B.1.3), so we applied analysis of covariance to help us select the final model, our final model is the model without interaction effect:

$$\text{Per capita income} = 18110 + 0.0089 \cdot \text{Total serious crimes} + 2286 \cdot I_{region=Northeast} - 860.6 \cdot I_{region=South} - 142.8 \cdot I_{region=West} + \varepsilon \quad (1)$$

Table 6 gives the full table of estimated coefficients and standard errors for model (1).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.811e+04	3.784e+02	47.846	< 2e-16
crimes	8.915e-03	3.188e-03	2.797	0.00539
regionNE	2.286e+03	5.325e+02	4.293	2.17e-05
regionS	-8.606e+02	4.868e+02	-1.768	0.07782
regionW	-1.428e+02	5.796e+02	-0.246	0.80548

Table 6: Estimated coefficients and standard errors for model (1).

Table 7 gives the full table of analysis of covariance to compare interaction for section 1.

```
Model 1: per.cap.income ~ crimes + region + crimes:region
Model 2: per.cap.income ~ crimes + region
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    432 6438799739
2    435 6501791845 -3 -62992106 1.4088 0.2396
```

Table 7: Analysis of covariance comparison for section 1.

Section 2 – per capita serious crimes

Similar to section 1, both models we tested have similar diagnostics plots (see B.2.3), so we applied analysis of covariance to help us select the final model, our final model is still the model without interaction effect:

$$\text{Per capita income} = 18006.04 + 5773.2 \cdot \text{per capita serious crimes} + 2354.7 \cdot I_{region=Northeast} - 927.45 \cdot I_{region=South} - 34.92 \cdot I_{region=West} + \varepsilon \quad (2)$$

Table 8 gives the full table of estimated coefficients and standard errors for model (2)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	18006.04	537.04	33.528	< 2e-16
per.cap.crimes	5773.20	7520.41	0.768	0.4431
regionNE	2354.70	541.97	4.345	1.74e-05
regionS	-927.45	512.31	-1.810	0.0709
regionW	-34.92	586.03	-0.060	0.9525

Table 8: Estimated coefficients and standard errors for model (2).

Table 9 gives the full table of analysis of covariance to compare interaction for section 2.

```
Model 1: per.cap.income ~ per.cap.crimes + region + per.cap.crimes:region
Model 2: per.cap.income ~ per.cap.crimes + region
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    432 6607856753
2    435 6609753963 -3  -1897210 0.0413 0.9888
```

Table 9: Analysis of covariance comparison for section 2.

3. Develop regression models to predict average income per person (per capita income) in 1990(in dollars) from several social, health, and economic factors.

We considered multiple regressions using logarithmic and power transformations (see C.4.2) and Lasso regression using similar transformation as multiple regressions (see C.3.2 and C.4.3). Both approaches produced models with decent R^2 values and acceptable diagnostics plots (see C.4.2 and C.4.3).

Logarithmic and Power Transformations

Among models with logarithmic and power transformations, the models with the best residual diagnostic plots (see C.4.2), R^2_{adj}, and statistically significant predictors was the following model, shown with estimated regression coefficients:

$$\text{Per capita income} = 15786.56 - 800.3 \cdot log(\text{Land Area}) - 4.85 \cdot (\text{Percent of population aged } 18\text{–}34)^2 + 1121.76 \cdot \log(\text{Number of active physicians}) + 319.35 \cdot \text{Percent bachelor's degrees} - 350.42 \cdot \text{Percent below poverty level} + 314.39 \cdot \text{Percent unemployment} + \varepsilon \quad (3)$$

Table 10 gives the full table of estimated coefficients and standard errors for model (3)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15786.5564	927.8113	17.015	< 2e-16
log(land.area)	-800.2993	99.6280	-8.033	9.11e-15
I(pop.18_34^2)	-4.8520	0.3764	-12.891	< 2e-16
log(doctors)	1021.7628	85.8343	11.904	< 2e-16
pct.bach.deg	319.3505	16.9048	18.891	< 2e-16
pct.below.pov	-350.4243	22.2883	-15.722	< 2e-16
pct.unemp	314.3908	45.0838	6.973	1.16e-11

Table 10: Estimated coefficients and standard errors for model (3).

Lasso Regression Model

The Lasso method to help us select the variables to keep, and then fit a multiple regression based on the selection result of Lasso, shown here with estimated regression coefficients,

$$\begin{aligned}\text{Per capita income} = {} & 12144.72 - 4.89 \cdot (\text{Percent of population aged } 18\text{–}34)^2 + 953.14 \\ & \cdot \log(\text{Number of active physicians}) + 338.53 \cdot \text{Percent bachelor's degrees} - 338.79 \\ & \cdot \text{Percent below poverty level} + 260.33 \cdot \text{Percent unemployment} + 351.81 \\ & \cdot \log(\text{per capita crimes}) + 435.47 \cdot I_{region\,=\,Northeast} - 356.89 \cdot I_{region\,=\,South} - 1142.3 \\ & \cdot I_{region\,=\,West} + \varepsilon \end{aligned} \quad (4)$$

Table 11 gives the full table of estimated coefficients and standard errors for Lasso-based model (4)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	12144.7160	1347.8255	9.011	< 2e-16
I(pop.18_34^2)	-4.8927	0.3961	-12.352	< 2e-16
log(doctors)	953.1379	104.7391	9.100	< 2e-16
pct.bach.deg	338.5334	17.9307	18.880	< 2e-16
pct.below.pov	-338.7894	26.5331	-12.769	< 2e-16
pct.unemp	260.3254	50.0095	5.206	3.00e-07
log(per.cap.crimes)	351.8090	249.7461	1.409	0.160
regionNE	435.4662	273.2831	1.593	0.112
regionS	-356.8871	253.2255	-1.409	0.159
regionW	-1142.2991	284.6198	-4.013	7.06e-05

Table 11: Estimated coefficients and standard errors for Lasso-based model (4).

Final Model

Both models (1) and (2) had similar R^2_{adj} values (0.8122 and 0.7962, respectively) and similarly good residual diagnostic plots that follow the assumption of residuals (see C.4.2 and C.4.3). Since both models have very similar residual diagnostics, we would prefer to make our final decision based on information criterion such as AIC and BIC.

Model	R^2_{adj}	AIC	AIC_c	BIC
Model 1 Multiple Regression Model	**0.8122**	**7833.55**	**7833.88**	**7866.24**
Model 2 Lasso-based Regression Model	0.7962	7917.27	7872.94	7872.32

Three information criterions and adjusted R^2 all suggest model (1), the multiple regression model, as our final model.

5 Discussion

1. Explore relationships between several social, health, and economic factors

According to the relationships we found (see A.3):

1. Relationships between total population and the Number of active physicians, the number of hospital beds, Total serious crimes, and Total personal income are understandable since the more people you have, it is also likely to have more crimes; Large in total population also means the county could be a metro area, therefore with higher income, and more hospitals and thus more physicians and hospital beds.

2. Relationships between the Percent of population aged 18-34 and Percent of population aged 65 or older, the Percent bachelor's degrees are reasonable since the higher percentage of young people in the county, the lower percentage of older people in the county; and bachelor's degree is also easier to obtain nowadays (for youngsters) compared to ancient times (for elders).
3. Relationships between the Number of active physicians and the Number of hospital bed, Total serious crimes(crimes), and Total income can be explained as the same reason in 1. As stated in 1., since the number of active physicians is heavily affected by the number of hospitals, and metro areas are likely to have more hospitals, the observed relationships here aren't for no reason.
4. Relationships between the Percent high school graduates the Percent bachelor's degrees and the per capita income, Percent below poverty level and the Percent unemployment are also justifiable since better education usually gets paid better and gets job easier.

2. Illustrate the effect of serious crimes on per capita income across different geographical regions

Among the models we considered (multiple linear regression model with and without interaction) for each part (using either Total serious crimes or per capita crimes as predictor), the results stand out the same:

$$\text{Per capita income} = 18110 + 0.0089 \cdot \text{Total serious crimes} + 2286 \cdot I_{region=Northeast} - 860.6 \cdot I_{region=South} - 142.8 \cdot I_{region=West} + \varepsilon \quad (1)$$

$$\text{Per capita income} = 18006.04 + 5773.2 \cdot \text{per capita serious crimes} + 2354.7 \cdot I_{region=Northeast} - 927.45 \cdot I_{region=South} - 34.92 \cdot I_{region=West} + \varepsilon \quad (2)$$

Both model (1) and model (2) have similar coefficients, and both without interactions, table 12 help us understand two models more:

With all other variables remain unchanged,

Condition	Model (1)	Model (2)
Total serious crimes	For every 1 more serious crime in total, the per capita income increases by 8900 dollars	N/A
per capita serious crimes	N/A	For every 1 case increase in per capita serious crimes, the per capita income increases by 5773.2 dollar
Geographical Region = Northeast	If the county located in northeast of the US, the per capita income increases by 2286 dollars	If the county located in northeast of the US, the per capita income increases by 2345.7 dollars

Geographical Region = South	If the county located in northeast of the US, the per capita income decreases by 860.6 dollars	If the county located in northeast of the US, the per capita income decreases by 924.45 dollars
Geographical Region = West	If the county located in northeast of the US, the per capita income decreases by 142.8 dollars	If the county located in northeast of the US, the per capita income decreases by 34.92 dollars

Table 12: Interpretation of model (1) and (2).

The per capita income does vary over Geographical regions, however, there is no evidence showing that the relationship between per capita income and crimes differ across regions. Even when we use the predictor per capita serious crimes instead of Total serious crimes to avoid the underlying bias that may cause by population factor, the result still held.

3. Develop regression models to predict average income per person (per capita income) in 1990(in dollars) from several social, health, and economic factors.

Among models we considered (log and power transformations in multiple regression, as well as multiple regression based on Lasso), we found that the model that can best represent the relationship between per capita income and several social, health, and economic factors is:

$$\begin{aligned}\text{Per capita income} = {} & 15786.56 - 800.3 \cdot log(\text{Land Area}) - 4.85 \cdot (\text{Percent of population aged } 18\text{–}34)^2 \\ & + 1121.76 \cdot \log(\text{Number of active physicians}) + 319.35 \cdot \text{Percent bachelor's degrees} - 350.42 \\ & \cdot \text{Percent below poverty level} + 314.39 \cdot \text{Percent unemployment} + \varepsilon \qquad (3)\end{aligned}$$

Variables log (Land Area right), $(\text{Percent of population aged } 18\text{–}34)^2$, log (Number of active physicians), Percent bachelor's degrees, Percent below poverty level, and , Percent unemployment are all highly significant predictors of per capita income; the variation in predicted per capita income accounts for $R^2 \cdot 100\% = 81.47\%$ of the variation in these variables.

The model can be interpreted as the following:

With all other variables remain unchanged,

1. Approximately every 1 percent increase in Land Area will decrease per capita income by 8 dollars.
2. Every n percent increase in Percent of population aged 18– 34, the per capita income will decrease by $4.85 \cdot n^2$ dollars.
3. Approximately every 1 percent increase in the Number of active physicians will increase per capita income by 11.22 dollars.
4. Every 1 percent increase in the Percent bachelor's degrees raises the per capita income by 319.35 dollars.
5. Every 1 percent increase in the Percent below poverty level, the per capita income will decrease by 350.42 dollars
6. Every 1 percent increase in the Percent unemployment increases the per capita income by 314.39 dollars

There is one predictor that may require further discussion: the Percent unemployment. An increase in unemployment rate raises the per capita income seems unreasonable. We suspected that this could be the result of multicollinearity, however, the Variance Inflation Factor of this model (see C.5) disagreed with this assumption.

Another important reminder is that the data is quite old, from 1990. The economy condition and formation 3 decades ago may not follow the principles we are familiar with in the 21st century. This limits the generalizability of the results to the present time.

It also might be useful to have more than the data from 440 counties, and 48 states, which we will further discuss in **part 4**.

Also note that, both predictors in **part 2** are not included in the model (3), this is also due to the reason of multicollinearity, as stated in **part 1**, many of those predictors are related to metro areas, the variance inflation factor already helped us filter them out.

4. Discuss the impact of the sample size and missing values.

The missing data and sample size indeed influenced the result. We only have 440 observations from 48 states (including Washington DC) in the dataset, while there are approximately 3000 counties from 51 states (including Washington DC) in the country. In addition, plenty of states in the dataset only have very few counties recorded, and most of those states have low income and do not have too many tier 1 or tier 2 cities. According to World Population Review [2], the top five richest states in the US (in terms of per capita GDP) are, New York (NY), Massachusetts (MA), Washington (WA), California (CA), and Connecticut (CT). These five states have contributed 85 observations in the dataset, which is nearly 20% of all data, which also indicate the data could likely be composed of the counties in metro areas, while the condition in metro areas may differ from countryside.

References

Sheather, S.J. (2009), A Modern Approach to Regression with R. New York: Springer Science + Business Media LLC.

[2] World Population Review, https://worldpopulationreview.com/state-rankings/richest-states-in-usa

Techincal Appendix

A

A.1 Make a table or tables showing appropriate summary statistics for each variable in thedata set. Note that summary statistics for continuous variables will be different fromthe summary statistics for categorical variables.

A.1.1 Summary: Numerical Variables

```
describe(cdi[,4:16]) %>%  select(-trimmed, -mad, -se, -vars ) %>%
  round(2) %>% kbl(booktabs=T,caption="Summary Statistics for 13 continuous variables") %>%
  kable_classic(latex_options = "HOLD_position")
```

Table 1: Summary Statistics for 13 continuous variables

	n	mean	sd	median	min	max	range	skew	kurtosis
land.area	440	1041.41	1549.92	656.50	15.0	20062.0	20047.0	6.21	57.09
pop	440	393010.92	601987.02	217280.50	100043.0	8863164.0	8763121.0	8.13	96.81
pop.18_34	440	28.57	4.19	28.10	16.4	49.7	33.3	1.22	4.02
pop.65_plus	440	12.17	3.99	11.75	3.0	33.8	30.8	1.92	6.74
doctors	440	988.00	1789.75	401.00	39.0	23677.0	23638.0	6.56	66.81
hosp.beds	440	1458.63	2289.13	755.00	92.0	27700.0	27608.0	6.01	53.36
crimes	440	27111.62	58237.51	11820.50	563.0	688936.0	688373.0	7.89	78.32
pct.hs.grad	440	77.56	7.02	77.70	46.6	92.9	46.3	-0.73	1.55
pct.bach.deg	440	21.08	7.65	19.70	8.1	52.3	44.2	0.96	1.16
pct.below.pov	440	8.72	4.66	7.90	1.4	36.3	34.9	1.68	5.76
pct.unemp	440	6.60	2.34	6.20	2.2	21.3	19.1	1.78	6.22
per.cap.income	440	18561.48	4059.19	17759.00	8899.0	37541.0	28642.0	1.22	2.42
tot.income	440	7869.27	12884.32	3857.00	1141.0	184230.0	183089.0	7.68	88.10

Most of the varibles here are right skewed (skewness >0) implying that we may need to apply transformation when fitting models.
In addition, there are several variables have extremely large kurtosis.

A.1.2 Summary: Categorical Variables

```
apply(cdi[,c(2,3,17)],2,function(x) {length(unique(x))}) %>%
  kbl(booktabs=T,col.names="unique values",caption="Unique values of categorical variables") %>%
  kable_classic(full_width=F,latex_options = "HOLD_position")
```

Table 2: Unique values of categorical variables

	unique values
county	373
state	48
region	4

```
tb <- rbind(with(cdi,table(state)%>% sort(TRUE)))  %>% as.data.frame()
tb3 = tb[,1:16]
tb4 = tb[,17:32]
tb5 = tb[,33:48]
row.names(tb3) <- "Freq"
row.names(tb4) <- "Freq"
row.names(tb5) <- "Freq"
tb3 %>% kbl(booktabs=T,caption="Counts of each State(1)") %>% kable_classic(full_width=F, latex_options
```

Table 3: Counts of each State(1)

	CA	FL	PA	TX	OH	NY	MI	NC	NJ	IL	IN	MA	SC	WI	MD	WA
Freq	34	29	29	28	24	22	18	18	18	17	14	11	11	11	10	10

```
tb4 %>% kbl(booktabs=T,caption="Counts of each State(2)") %>% kable_classic(full_width=F, latex_options
```

Table 4: Counts of each State(2)

	CO	GA	LA	VA	CT	MO	TN	AL	MN	OR	AZ	ME	KS	NH	OK	UT
Freq	9	9	9	9	8	8	8	7	7	6	5	5	4	4	4	4

```
tb5 %>% kbl(booktabs=T,caption="Counts of each State(3)") %>% kable_classic(full_width=F, latex_options
```

Table 5: Counts of each State(3)

	HI	KY	MS	NE	RI	AR	DE	NM	NV	DC	ID	MT	ND	SD	VT	WV
Freq	3	3	3	3	3	2	2	2	2	1	1	1	1	1	1	1

Table 2, 3, 4, and 5 show that the dataset contains data from 48 different states in the US, among all of states, California(CA) has the most counties involved in this data with 34 counties while West Virginia(WV) has the least counties included, with only one county recorded in the data.

Table 6: Counts of each region

	NC	NE	S	W
Freq	108	103	152	77

```
tb6 <- rbind(with(cdi,table(region)))
row.names(tb6) <- "Freq"
tb6 %>% kbl(booktabs=T,caption="Counts of each region") %>% kable_classic(full_width=F)
```

Table 6 shows that 152 counties are in the South US, 108 counties are in North Central US, 103 counties are in North East US, and 77 counties are in West US.

A.2 Indicate where (in which variables) there is missing data (NA's), if any, how much there is (in each variable) and why it might be there.

```
apply(cdi,2,function(x) any(is.na(x)))
```

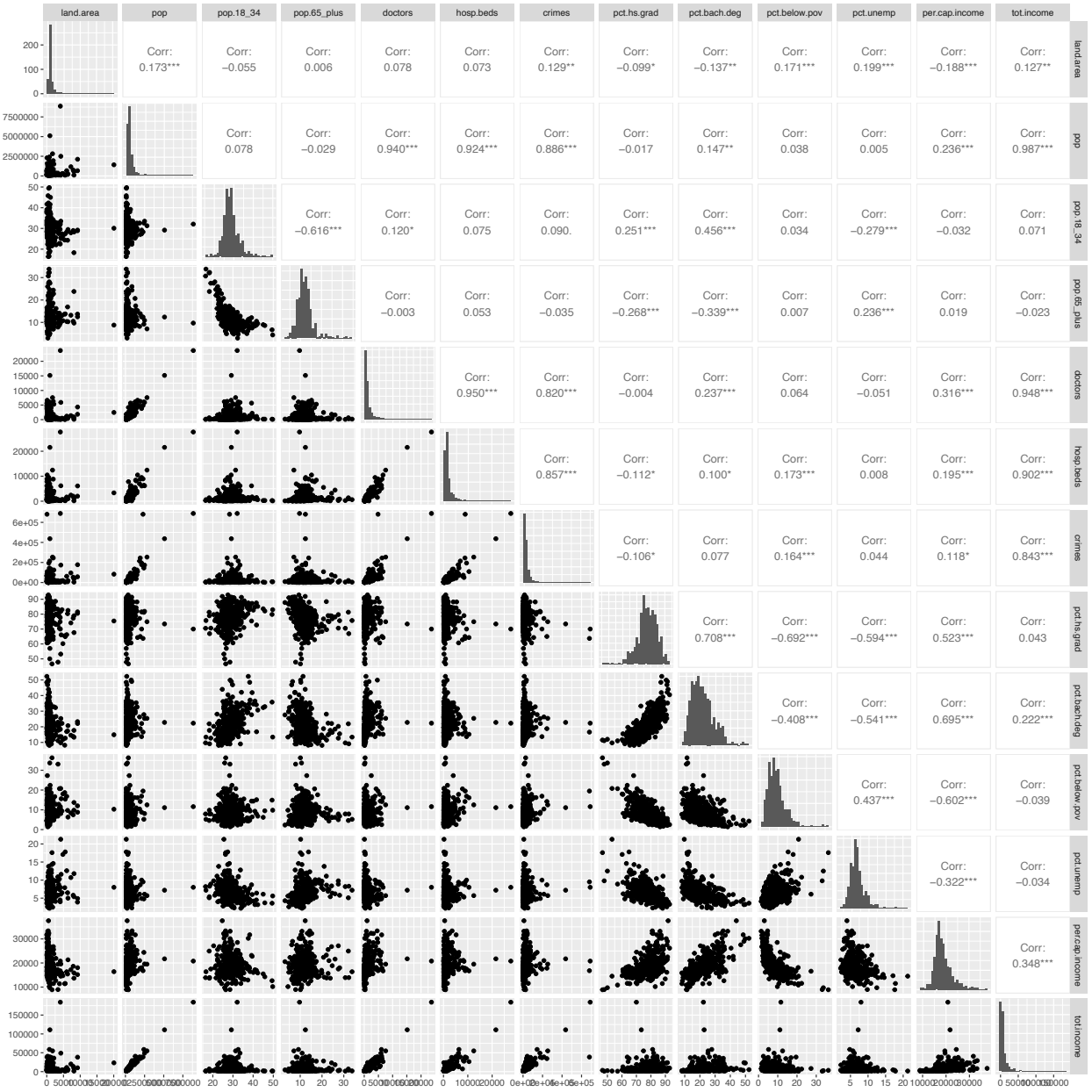
```
##            id        county         state     land.area           pop
##         FALSE         FALSE         FALSE         FALSE         FALSE
##     pop.18_34   pop.65_plus       doctors     hosp.beds        crimes
##         FALSE         FALSE         FALSE         FALSE         FALSE
##   pct.hs.grad  pct.bach.deg pct.below.pov    pct.unemp per.cap.income
##         FALSE         FALSE         FALSE         FALSE         FALSE
##    tot.income        region
##         FALSE         FALSE
```

There is no missing data (NA's) in the dataset cdi.

A.3 Make some appropriate descriptive EDA plots to illustrate any important features of the variables or possible important relationships among them

```
graphs = ggpairs(cdi[,4:16], diag = list(continuous = "barDiag"))
print(graphs, progress =F, warning=F)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



As mentioned earlier in **A.1.1**, the histograms show that most of the variables are right skewed, except for percent high school graduates(pct.hs.grad)), who's skewness is negative.
There are some apparent relationships between variables:
1.Total population(pop) is strongly and postively related to the number of active physicians(doctors), the number of hospital beds(hosp.beds), total serious crimes(crimes), and total personal income(tot.income). These relationships are understandable since the more people you have, it is also likely to have more crimes; Large in total population also means the county could be a metro area, therefore with higher income, more hospitals and thus more physicians and hospital beds.

2.Percent of population aged 18-34(pop.18_34) has a strong negative relationship with Percent of population aged 65 or older(pop.65_plus) and a mediocre postive relationship with the Percent bachelor's degrees(pct.bach.deg). These relationships are reasonable since the higher percentage of young people in the county, the lower percentage of older people in the county; and bachelor degree is also easier to obtain nowadays.

3.The number of active physicians(doctors) is highly correlated to the number of hospital bed(hosp.beds), total serious crimes(crimes), and total personal income (tot.income). As stated in 1., since the number of physicians is heavily effected by the number of hospitals, and metro areas are likley to have more hospitals, the observed relationships here aren't for no reason.

4.Same condition in 3 can be applied to the number of hospital bed(hosp.beds) and total serious crimes(crimes).

5.The percent high school graduates(pct.hs.grad) is postively related to the percent bachelor's degrees(pct.bach.deg) and the per capita income(per.cap.income), while negatively related to the percent below poverty level(pct.below.pov) and the percent unemployment(pct.unemp). These are also justifiable since better education usually gets paid better and gets job easier.

6.Similar argunments in 5 can be applied to the percent bachelor's degrees(pct.bach.deg), the percent below poverty level(pct.below.pov), and the percent unemployment(pct.unemp).

B

Build a regression model that predicts per-capita income from crime rate and region of the country. Should there be any interactions in the model? What does your model say about the relationship between per- capita income and crime rate? Do your answers change, depending on whether you use number of crimes, or "per-capita crime" = (number of crimes)/(population) as a crime rate measure? If so, which one best answers the question? Why? Show the fitted model results and explain your answer to these questions in terms of those results.

B.1 Total serious crimes vs Per capita income

B.1.1 Model - with interaction

```
lm1 = lm(per.cap.income ~ crimes + region + crimes:region, data = cdi)
lm1 %>% summary()
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + crimes:region,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8582.4 -2225.2  -676.2  1563.4 19504.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.800e+04  4.092e+02  43.995  < 2e-16 ***
## crimes           1.361e-02  7.882e-03   1.726   0.0851 .
## regionNE         2.573e+03  5.736e+02   4.487 9.28e-06 ***
## regionS         -1.056e+03  5.606e+02  -1.884   0.0602 .
## regionW         -5.654e+01  6.372e+02  -0.089   0.9293
## crimes:regionNE -1.272e-02  9.677e-03  -1.314   0.1895
## crimes:regionS   6.348e-03  1.136e-02   0.559   0.5765
## crimes:regionW  -4.295e-03  9.486e-03  -0.453   0.6509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3861 on 432 degrees of freedom
## Multiple R-squared:  0.1099, Adjusted R-squared:  0.09543
## F-statistic: 7.616 on 7 and 432 DF,  p-value: 1.122e-08
```

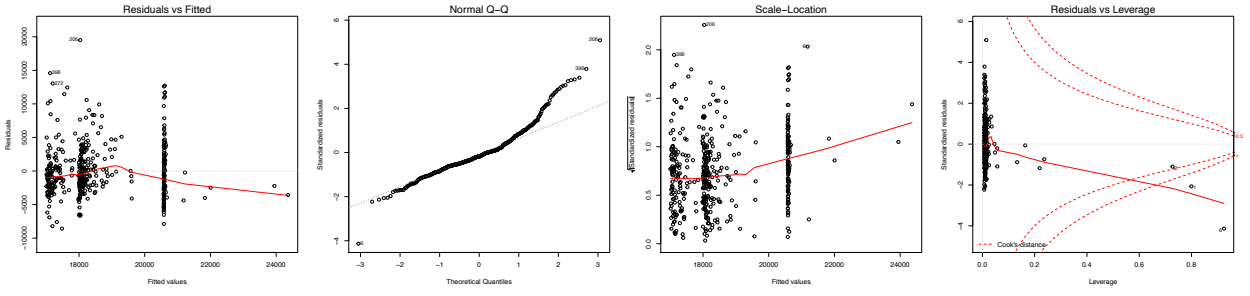
B.1.2 Model - without interaction

```
lm2 = lm(per.cap.income ~ crimes + region, data = cdi)
lm2 %>% summary()
```

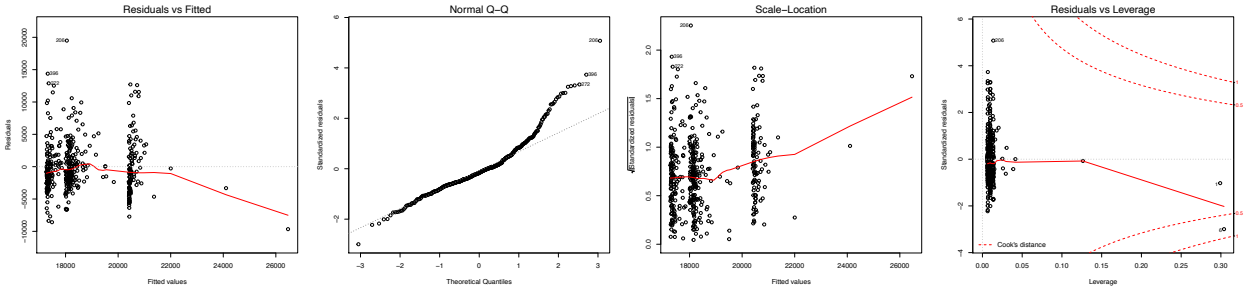
```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846  < 2e-16 ***
## crimes       8.915e-03  3.188e-03   2.797  0.00539 **
## regionNE     2.286e+03  5.325e+02   4.293 2.17e-05 ***
## regionS     -8.606e+02  4.868e+02  -1.768  0.07782 .
## regionW     -1.428e+02  5.796e+02  -0.246  0.80548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09288
## F-statistic: 12.24 on 4 and 435 DF,  p-value: 1.946e-09
```

B.1.3 Comparison

```
par(mfrow = c(1,4))
plot(lm1)
```




```
plot(lm2)
```



The diagnostics plots of both models show similar results, we can hardly judge by them. Therefore, the analysis of covariance is carried out here, with:
$H_0 : \beta_{interaction} = 0$
$H_1 : \beta_{interaction} \neq 0$

```
anova(lm1, lm2)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ crimes + region + crimes:region
## Model 2: per.cap.income ~ crimes + region
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    432 6438799739
## 2    435 6501791845 -3 -62992106 1.4088 0.2396
```

The p-value of the test is 0.2396 which isn't statistically significant. If we set our $\alpha = 0.05$, H_0 cannot be rejected, so we conculde the interaction term is not needed in this model.

B.2 Per capita serious crimes vs Per capita income

B.2.1 Model - with interaction

```
cdi$per.cap.crimes = cdi$crimes/cdi$pop
lm3 = lm(per.cap.income ~ per.cap.crimes + region + per.cap.crimes:region, data = cdi)
lm3 %>% summary()
```

```
##
## Call:
## lm(formula = per.cap.income ~ per.cap.crimes + region + per.cap.crimes:region,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8637.7 -2333.9  -629.5  1759.1 19515.6
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               18077.3      895.2  20.193   <2e-16 ***
## per.cap.crimes             4379.1    15893.5   0.276    0.783
## regionNE                   2329.0     1101.4   2.115    0.035 *
## regionS                   -1010.4     1323.8  -0.763    0.446
## regionW                    -670.0     1983.9  -0.338    0.736
```

```
## per.cap.crimes:regionNE    288.4    20184.7   0.014    0.989
## per.cap.crimes:regionS    1558.9    20556.1   0.076    0.940
## per.cap.crimes:regionW   10655.5    32322.4   0.330    0.742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3911 on 432 degrees of freedom
## Multiple R-squared:  0.08648,    Adjusted R-squared:  0.07168
## F-statistic: 5.842 on 7 and 432 DF,  p-value: 1.713e-06
```

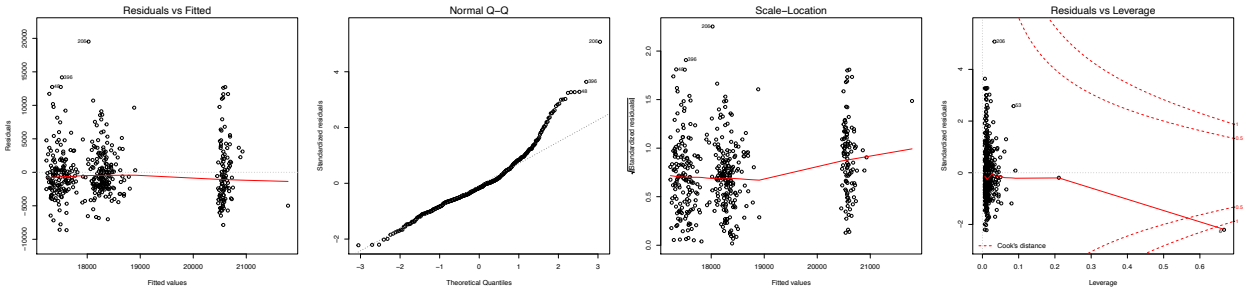
B.2.2 Model - without interaction

```
lm4 = lm(per.cap.income ~ per.cap.crimes  + region, data = cdi)
lm4 %>% summary()
```

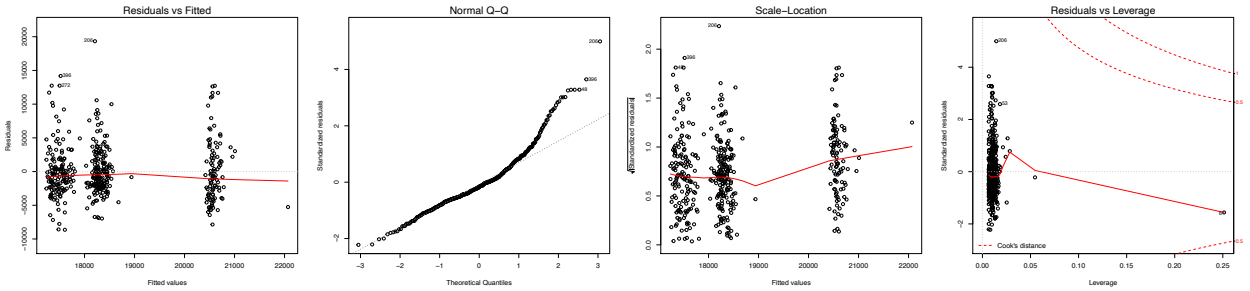
```
##
## Call:
## lm(formula = per.cap.income ~ per.cap.crimes + region, data = cdi)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8634  -2300   -631   1710  19333
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18006.04     537.04  33.528  < 2e-16 ***
## per.cap.crimes  5773.20    7520.41   0.768   0.4431
## regionNE        2354.70     541.97   4.345 1.74e-05 ***
## regionS         -927.45     512.31  -1.810   0.0709 .
## regionW          -34.92     586.03  -0.060   0.9525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3898 on 435 degrees of freedom
## Multiple R-squared:  0.08622,    Adjusted R-squared:  0.07782
## F-statistic: 10.26 on 4 and 435 DF,  p-value: 6.007e-08
```

B.2.3 Comparison

```
par(mfrow = c(1,4))
plot(lm3)
```



```r
plot(lm4)
```



Again, the diagnostics plots of both models show similar results, we can hardly judge by them. Therefore, the analysis of covariance is carried out here, with:
$H_0 : \beta_{interaction} = 0$
$H_1 : \beta_{interaction} \neq 0$

```r
anova(lm3, lm4)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ per.cap.crimes + region + per.cap.crimes:region
## Model 2: per.cap.income ~ per.cap.crimes + region
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    432 6607856753
## 2    435 6609753963 -3  -1897210 0.0413 0.9888
```

The p-value of the test is 0.9888 which isn't statistically significant. If we set our $\alpha = 0.05$, H_0 cannot be rejected, so we conculde the interaction term is not needed in this model.

Both predictors, total serious crimes and per captia serious crimes, suggest that we do not need the interaction term.

C

Make a table or tables showing appropriate summary statistics for each variable in thedata set. Note that summary statistics for continuous variables will be different fromthe summary statistics for categorical variables.

C.1 Original Model(11 variables: 10 numerical and 1 categorical)

The very first model to be tested is shown as below:

```
lm_origin = lm(per.cap.income ~ .-tot.income -pop -id -county - state - crimes, data = cdi)
```

The id variable has nothing to do with the prediction, while the per capita income $= \frac{Totalincome}{Population}$, we taken them out. In addition, different counties in different states represent completely different data, thus there is no point to include this variable, many states aslo only have one or few data, therefore, county and state variables are also precluded from the model. Also, we use per capita serious crimes in the model instead of total serious crimes to eliminate the underlying bias that may cause by the population.

Use VIF to check whether there exists multicollinearity:

```
rms::vif(lm_origin)
```

```
##      land.area      pop.18_34    pop.65_plus        doctors      hosp.beds
##       1.444885       2.011387       1.784483      15.820578      16.130874
##    pct.hs.grad   pct.bach.deg  pct.below.pov      pct.unemp       regionNE
##       4.588350       3.456760       2.929957       1.925453       1.854842
##        regionS        regionW per.cap.crimes
##       2.084975       2.055302       1.909048
```

The cutoff value here is decided to be 4, we will examine variables with VIF larger than 4.

C.1.1 The number of physicians(doctor) vs the number of hospital beds(hosp.beds)

It is understandable that the number of physicians(doctor) is collinear with the number of hospital beds(hosp.beds) since the number of hospital in the city/county decides the number of both, their pearson-correlation is also high,

```
cor(cdi$doctors, cdi$hosp.beds)
```

```
## [1] 0.9504644
```

we can almost be sure that either of two variables needs to be removed, since hosp.beds yields a higher VIF, we consider dropping it here.

C.1.2 The percent high school graduates(pct.hs.grad) vs the Percent bachelor's degree(pact.bach.deg)

This is also an easy pick since people need to graduate from high school first before attending to the college.

```
cor(cdi$pct.hs.grad, cdi$pct.bach.deg)
```

```
## [1] 0.7077867
```

The correaltion between two variables is approximately 0.708, which is relatively high. Because The percentage of high school graduates(pct.hs.grad) has a higher VIF, we drop it first.

C.1.3 Model after multicollinearity adjustment

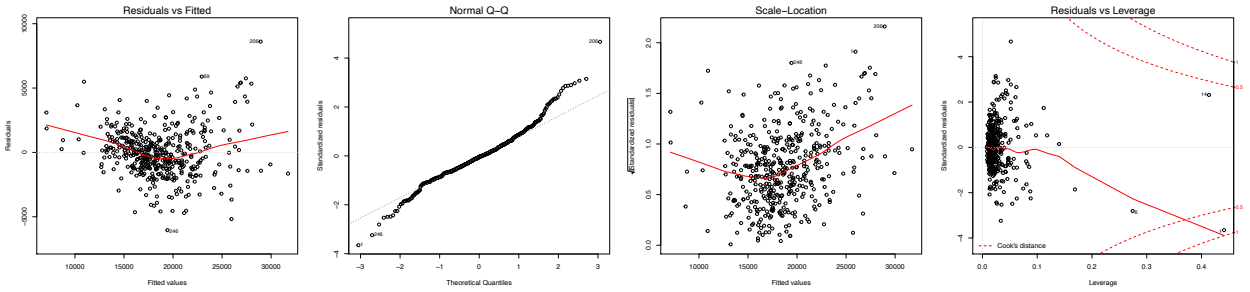
```
lm_adj_1 = lm(per.cap.income ~ .-tot.income -pop -id -county - state - crimes - pct.hs.grad - hosp.beds

lm_adj_1 %>% summary
```

```
##
## Call:
## lm(formula = per.cap.income ~ . - tot.income - pop - id - county -
##     state - crimes - pct.hs.grad - hosp.beds, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6041.3 -1126.1   -79.0   947.1  8610.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.019e+04  1.159e+03  17.425  < 2e-16 ***
## land.area      -1.261e-01  7.005e-02  -1.800  0.07256 .
## pop.18_34      -3.361e+02  3.060e+01 -10.983  < 2e-16 ***
## pop.65_plus     1.506e+00  3.014e+01   0.050  0.96017
## doctors         4.155e-01  5.654e-02   7.349 1.02e-12 ***
## pct.bach.deg    3.807e+02  1.730e+01  22.012  < 2e-16 ***
## pct.below.pov  -3.305e+02  2.811e+01 -11.759  < 2e-16 ***
## pct.unemp       2.359e+02  5.176e+01   4.559 6.73e-06 ***
## regionNE        7.910e+02  2.797e+02   2.828  0.00491 **
## regionS        -5.028e+02  2.575e+02  -1.952  0.05154 .
## regionW        -8.268e+02  3.282e+02  -2.519  0.01212 *
## per.cap.crimes  1.932e+04  4.331e+03   4.462 1.04e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1895 on 428 degrees of freedom
## Multiple R-squared:  0.7875, Adjusted R-squared:  0.782
## F-statistic: 144.2 on 11 and 428 DF,  p-value: < 2.2e-16
```

The R^2 seems decent right here, although there are still some predictors that are not statistically significant. This problem will be dealt later, we now want to check whether the residuals follow some vital assumptions such as no mean structure, constant variance, and normally distributed.

```
par(mfrow = c(1,4))
plot(lm_adj_1)
```



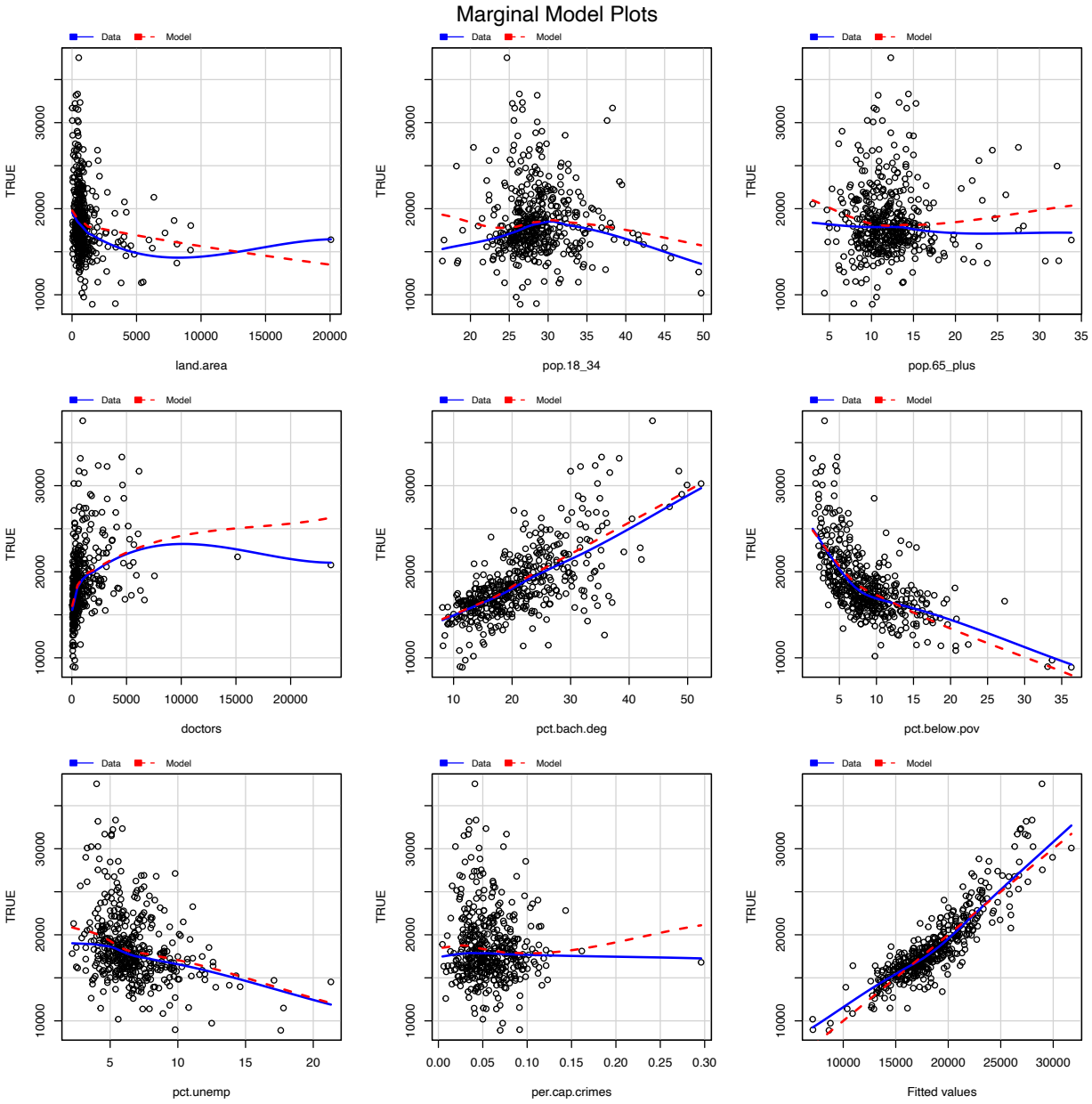
Apparently, none of the assumptions are held, and there seems to be a bad influential point in the model, these suggest we may need some transformation on predictors or the response.

C.2 Transformation

C.2.1 Marginal model plots

```
mmps(lm_adj_1)
```

```
## Warning in mmps(lm_adj_1): Interactions and/or factors skipped
```



Marginal model plots show that we may need to apply transformations on Land Area(land.area), the percentage of age 18 to 34 (pop_18_34), the percentage of age 65 and older (pop_65_plus), the number of physicians(doctor), and per capita serious crimes(per.cap.crimes).

C.2.2 Log-transformation

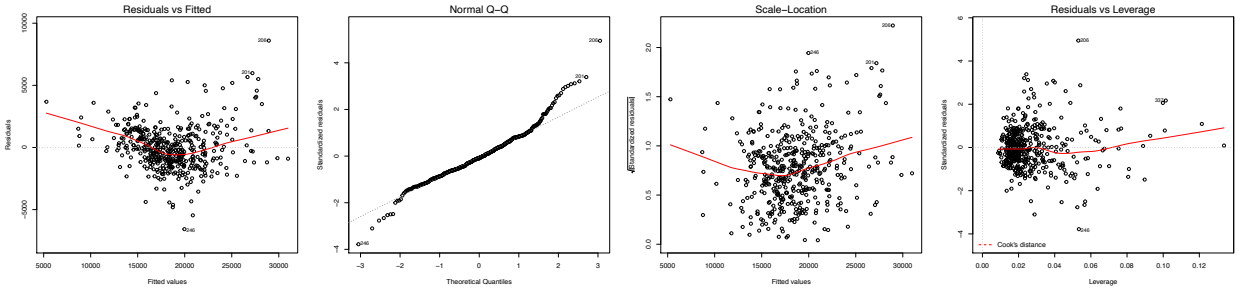
Since many of these variables are right-skewed, we try to apply log-transformation on them first, this includes: Land Area(land.area), the number of physicians(doctors), the percentage of age 18 to 34 (pop_18_34), the percentage of age 65 and older (pop_65_plus), and per capita serious crimes(per.cap.crimes).

```
lm_adj_1_log = lm(per.cap.income ~ log(land.area) + log(pop.18_34) + log(pop.65_plus) +
                                   log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                   log(per.cap.crimes) + region, data = cdi)
```

```
lm_adj_1_log %>% summary()
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + log(pop.18_34) +
##     log(pop.65_plus) + log(doctors) + pct.bach.deg + pct.below.pov +
##     pct.unemp + log(per.cap.crimes) + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6573.6 -1077.9  -165.4   952.7  8600.7
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         43761.04    3990.21  10.967  < 2e-16 ***
## log(land.area)       -747.82     119.04  -6.282 8.23e-10 ***
## log(pop.18_34)      -9217.26     906.83 -10.164  < 2e-16 ***
## log(pop.65_plus)     -323.40     417.33  -0.775   0.4388
## log(doctors)         1048.90     105.10   9.980  < 2e-16 ***
## pct.bach.deg          313.37      17.89  17.521  < 2e-16 ***
## pct.below.pov        -335.62      25.92 -12.949  < 2e-16 ***
## pct.unemp             275.78      48.98   5.631 3.25e-08 ***
## log(per.cap.crimes)   217.97     247.02   0.882   0.3781
## regionNE              474.63     272.86   1.739   0.0827 .
## regionS              -300.96     248.90  -1.209   0.2273
## regionW              -209.89     315.01  -0.666   0.5056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1787 on 428 degrees of freedom
## Multiple R-squared:  0.811,  Adjusted R-squared:  0.8061
## F-statistic:   167 on 11 and 428 DF,  p-value: < 2.2e-16
```

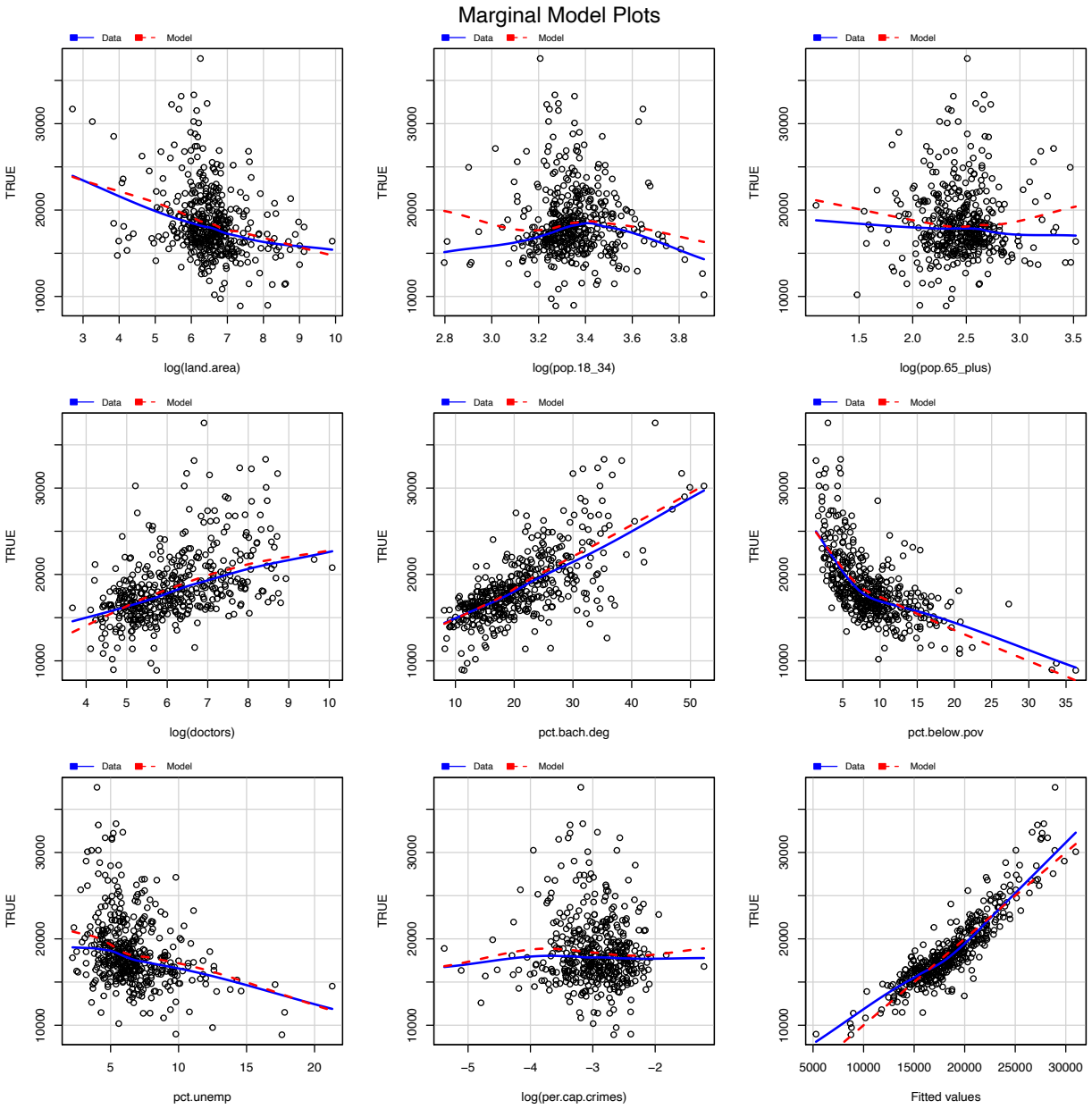
```
par(mfrow = c(1,4))
plot(lm_adj_1_log)
```



The result of log transformations seems tobe successful, given that not only R^2 increases, but also the diagnostics plots imporve, although there is still a slight curvature in the mean structure of residuals, the bad influential point is eliminated.
One thing we should notice is that the per capita serious crimes(per.cap.crimes) become insignificant in the new model.

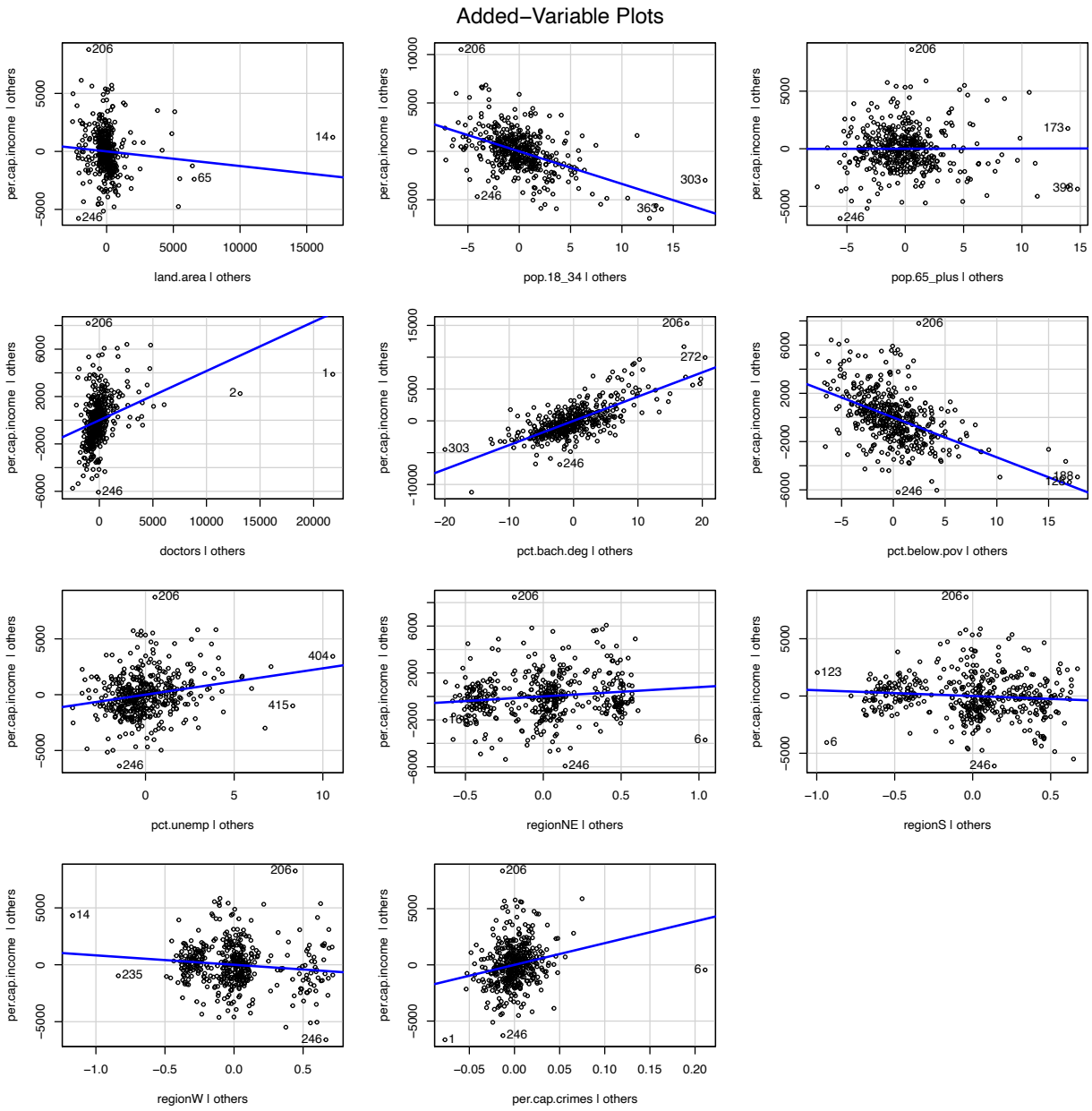
```
lm_adj_1_log %>% mmps()
```

```
## Warning in mmps(.): Interactions and/or factors skipped
```

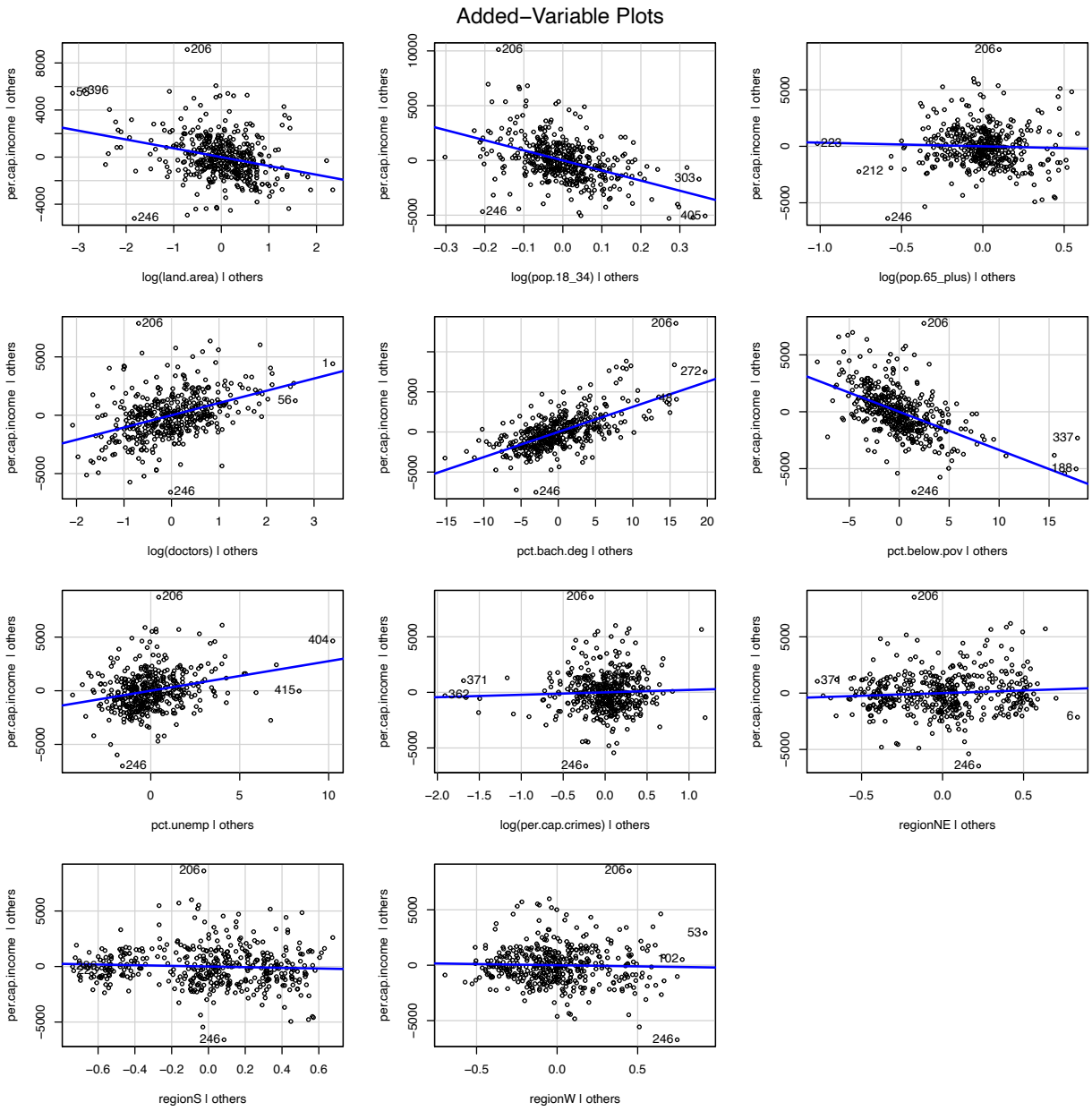


The marinal model plots now look mostly fine, although the percentage of age 18 to 34 (pop_18_34) and the percentage of age 65 and older (pop_65_plus) still don't fit well. However, since the percentage of age 65 and older (pop_65_plus) isn't statisitcally significant in either of models, we may consider dropping it.

```
avPlots(lm_adj_1, layout = c(4,3))
```




```
avPlots(lm_adj_1_log, layout = c(4,3))
```



The added variable plots of before transformation model and after transformation model both suggest that the percentage of age 65 and older (pop_65_plus) isn't contributing to the per capita income prediction. Therefore, we hope to drop this term. Also note that both models did not reach consensus on per capital serious crimes(per.cap.crimes), so we try to keep it for later discussion instead of dropping it.

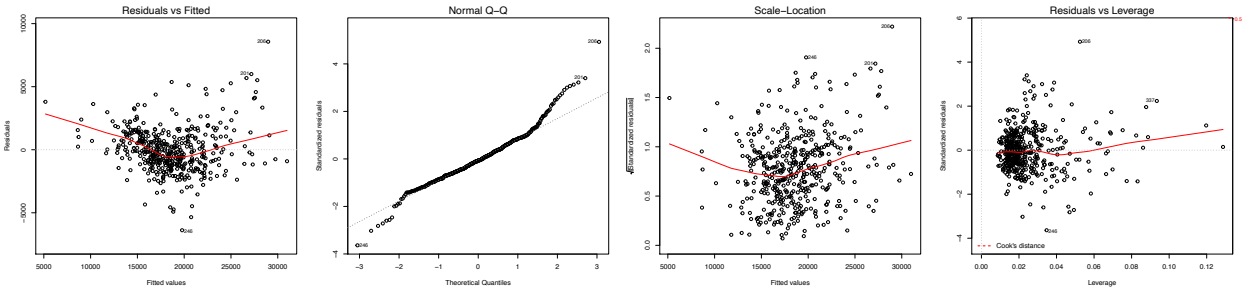
```
lm_adj_1_log_drop = lm(per.cap.income ~ log(land.area) + log(pop.18_34) +
                                        log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                        log(per.cap.crimes) + region, data = cdi)
```

```
lm_adj_1_log_drop %>% summary()
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + log(pop.18_34) +
```

```
##     log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
##     log(per.cap.crimes) + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6384.4 -1091.2  -126.7   974.3  8568.3
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         41599.12    2851.46  14.589  < 2e-16 ***
## log(land.area)       -743.70     118.87  -6.256 9.54e-10 ***
## log(pop.18_34)      -8797.54     726.97 -12.102  < 2e-16 ***
## log(doctors)         1027.67     101.42  10.133  < 2e-16 ***
## pct.bach.deg          316.04      17.54  18.018  < 2e-16 ***
## pct.below.pov        -336.63      25.87 -13.010  < 2e-16 ***
## pct.unemp             276.39      48.95   5.647 2.98e-08 ***
## log(per.cap.crimes)   210.26     246.70   0.852    0.395
## regionNE              429.65     266.49   1.612    0.108
## regionS              -292.40     248.54  -1.176    0.240
## regionW              -193.95     314.19  -0.617    0.537
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1786 on 429 degrees of freedom
## Multiple R-squared:  0.8107, Adjusted R-squared:  0.8063
## F-statistic: 183.8 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(1,4))
plot(lm_adj_1_log_drop)
```



After dropping the variable the percentage of age 65 and older (pop_65_plus), the model still looks valid, with similar diagnostics plots and R^2. We then use analysis of covariance to help justify the move. Where
$H_0: \ \beta_{log(pop65_plus)} = 0$
$H_1: \ \beta_{log(pop65_plus)} \neq 0$

```
anova(lm_adj_1_log_drop, lm_adj_1_log)
```

```
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(land.area) + log(pop.18_34) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + log(per.cap.crimes) +
##     region
## Model 2: per.cap.income ~ log(land.area) + log(pop.18_34) + log(pop.65_plus) +
```

```
##     log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
##     log(per.cap.crimes) + region
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    429 1368988665
## 2    428 1367070605  1   1918060 0.6005 0.4388
```

The p-value here is 0.4388, therefore, we cannot reject H_0. That is, dropping the term may be a great idea.

C.2.3 Transformation - the percentage of age 18 to 34 (pop_18_34)

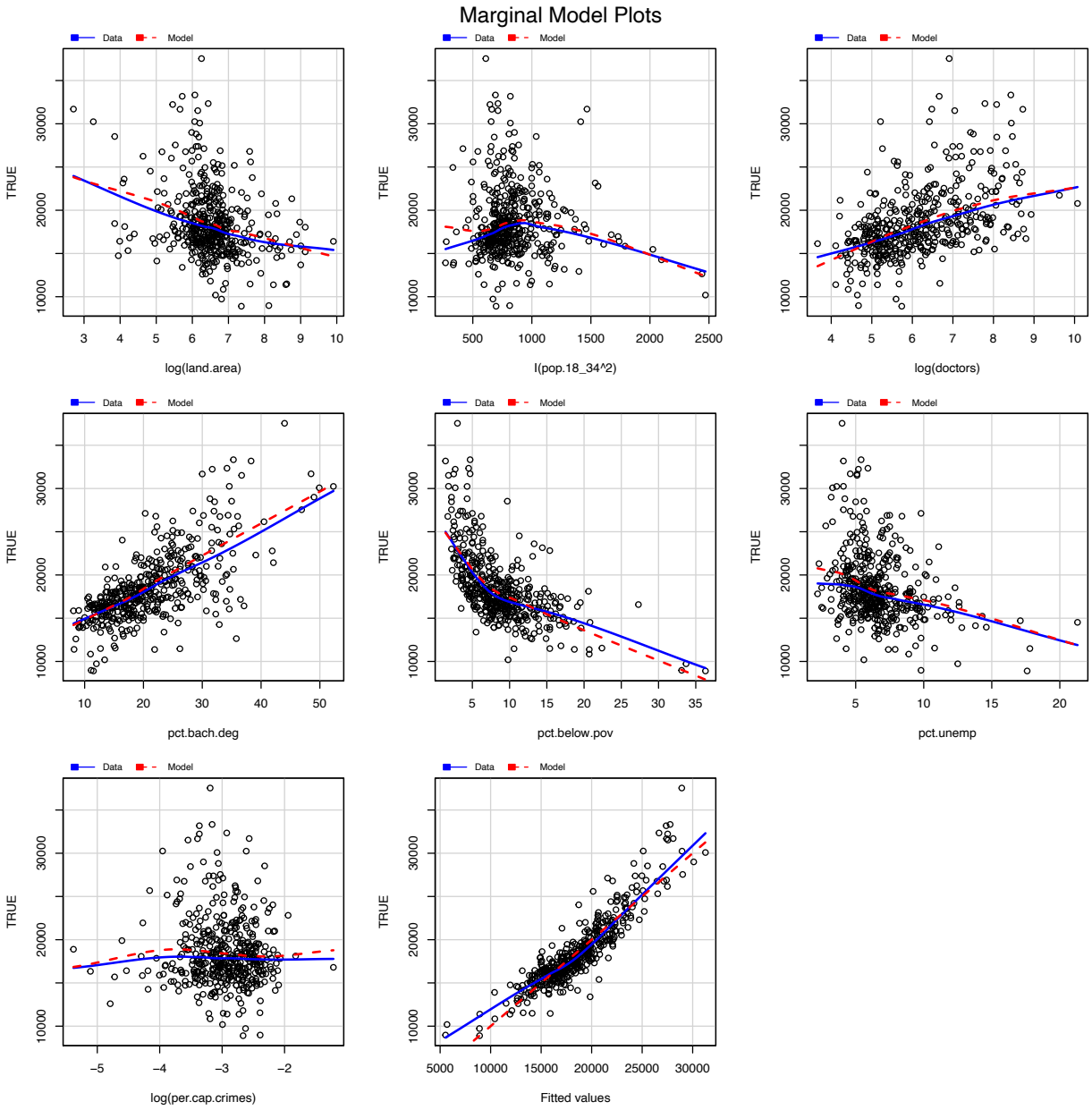
In previous section we could see that the marginal model plot of the variable the percentage of age 18 to 34 (pop_18_34) still isn't great, which we may need some adjustment on the variable.

The new transformation to be applied is 2-order power transformation, because there appears to be a slight curvature in the scatter plot.

```
lm_adj_1_log_drop_power = lm(per.cap.income ~ log(land.area) + I(pop.18_34^2) +
                                              log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                              log(per.cap.crimes) + region, data = cdi)
```

```
lm_adj_1_log_drop_power %>% mmps
```

```
## Warning in mmps(.): Interactions and/or factors skipped
```

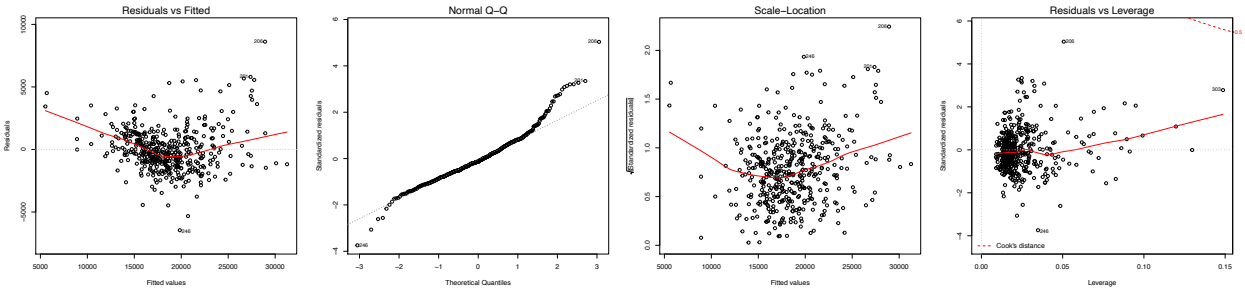



```
par(mfrow = c(1,4))
summary(lm_adj_1_log_drop_power)
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + I(pop.18_34^2) +
##     log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
##     log(per.cap.crimes) + region, data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6456.3 -1065.8  -227.5   917.6  8619.2
##
```

```
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)         16133.4716  1441.1005  11.195  < 2e-16 ***
## log(land.area)       -730.0429   116.8671  -6.247 1.01e-09 ***
## I(pop.18_34^2)         -4.8958     0.3797 -12.895  < 2e-16 ***
## log(doctors)          965.3426   100.4135   9.614  < 2e-16 ***
## pct.bach.deg          323.5774    17.3529  18.647  < 2e-16 ***
## pct.below.pov        -332.0577    25.4553 -13.045  < 2e-16 ***
## pct.unemp             281.4239    48.0540   5.856 9.42e-09 ***
## log(per.cap.crimes)   144.3994   241.6781   0.597    0.550
## regionNE              394.5796   262.0290   1.506    0.133
## regionS              -217.3272   243.7476  -0.892    0.373
## regionW              -234.6244   309.0958  -0.759    0.448
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1756 on 429 degrees of freedom
## Multiple R-squared:  0.817,  Adjusted R-squared:  0.8128
## F-statistic: 191.6 on 10 and 429 DF,  p-value: < 2.2e-16
```

```
plot(lm_adj_1_log_drop_power)
```



The marginal model plots look fine after the transformation, so is the R^2. However, there is a slightly increased curvature in the mean structure of residuals , which may require further discussion later.

C.3 Model Selection

We have some basic idea that our model may somehow include predictor log(Land Area), *the percentage of age* $18-34^2$, log(the number of phyisicians), the percentage of bachelor degree, the percentage of population below poverty, the percentage of unemployment and maybe the per capita serious crimes and factor region.
We will use some stepwise methods and LASSO to determine whether to keep the last two predictors

C.3.1 Stepwise methods

C.3.1.1 Backward AIC

```
backAIC <- step(lm_adj_1_log_drop_power,direction="backward", data=cdi)
```

```
## Start:  AIC=6585.35
## per.cap.income ~ log(land.area) + I(pop.18_34^2) + log(doctors) +
```

```
##     pct.bach.deg + pct.below.pov + pct.unemp + log(per.cap.crimes) +
##     region
##
##                        Df  Sum of Sq        RSS    AIC
## - log(per.cap.crimes)  1    1101265 1324507543 6583.7
## - region               3   15804207 1339210485 6584.6
## <none>                              1323406278 6585.3
## - pct.unemp            1  105803362 1429209640 6617.2
## - log(land.area)       1  120378207 1443784485 6621.7
## - log(doctors)         1  285111210 1608517488 6669.2
## - I(pop.18_34^2)       1  512925687 1836331965 6727.5
## - pct.below.pov        1  524934906 1848341184 6730.3
## - pct.bach.deg         1 1072622424 2396028702 6844.5
##
## Step:  AIC=6583.71
## per.cap.income ~ log(land.area) + I(pop.18_34^2) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + region
##
##                   Df  Sum of Sq        RSS    AIC
## - region           3   15655398 1340162941 6582.9
## <none>                          1324507543 6583.7
## - pct.unemp        1  107398237 1431905780 6616.0
## - log(land.area)   1  125939644 1450447187 6621.7
## - log(doctors)     1  401761103 1726268646 6698.3
## - I(pop.18_34^2)   1  514495717 1839003260 6726.1
## - pct.below.pov    1  541414542 1865922085 6732.5
## - pct.bach.deg     1 1075148458 2399656001 6843.2
##
## Step:  AIC=6582.88
## per.cap.income ~ log(land.area) + I(pop.18_34^2) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
##                   Df  Sum of Sq        RSS    AIC
## <none>                          1340162941 6582.9
## - pct.unemp        1  150510839 1490673780 6627.7
## - log(land.area)   1  199715582 1539878523 6642.0
## - log(doctors)     1  438579326 1778742267 6705.5
## - I(pop.18_34^2)   1  514310650 1854473591 6723.8
## - pct.below.pov    1  765074995 2105237936 6779.6
## - pct.bach.deg     1 1104554723 2444717664 6845.4
```

C.3.1.2 Backward BIC

```
n = dim(cdi)[1]
backBIC <- step(lm_adj_1_log_drop_power,direction="backward", data=cdi, k = log(n))
```

```
## Start:  AIC=6630.3
## per.cap.income ~ log(land.area) + I(pop.18_34^2) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + log(per.cap.crimes) +
##     region
##
##                        Df  Sum of Sq        RSS    AIC
## - region                3   15804207 1339210485 6617.3
```

```
## - log(per.cap.crimes)  1    1101265 1324507543 6624.6
## <none>                              1323406278 6630.3
## - pct.unemp            1  105803362 1429209640 6658.1
## - log(land.area)       1  120378207 1443784485 6662.5
## - log(doctors)         1  285111210 1608517488 6710.1
## - I(pop.18_34^2)       1  512925687 1836331965 6768.3
## - pct.below.pov        1  524934906 1848341184 6771.2
## - pct.bach.deg         1 1072622424 2396028702 6885.4
##
## Step:  AIC=6617.27
## per.cap.income ~ log(land.area) + I(pop.18_34^2) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp + log(per.cap.crimes)
##
##                        Df  Sum of Sq        RSS    AIC
## - log(per.cap.crimes)  1     952456 1340162941 6611.5
## <none>                              1339210485 6617.3
## - pct.unemp            1  146668835 1485879320 6656.9
## - log(land.area)       1  200329489 1539539974 6672.5
## - log(doctors)         1  384101315 1723311799 6722.1
## - I(pop.18_34^2)       1  506082021 1845292505 6752.2
## - pct.below.pov        1  612337189 1951547674 6776.9
## - pct.bach.deg         1 1101883185 2441093669 6875.3
##
## Step:  AIC=6611.49
## per.cap.income ~ log(land.area) + I(pop.18_34^2) + log(doctors) +
##     pct.bach.deg + pct.below.pov + pct.unemp
##
##                   Df  Sum of Sq        RSS    AIC
## <none>                          1340162941 6611.5
## - pct.unemp        1  150510839 1490673780 6652.2
## - log(land.area)   1  199715582 1539878523 6666.5
## - log(doctors)     1  438579326 1778742267 6730.0
## - I(pop.18_34^2)   1  514310650 1854473591 6748.3
## - pct.below.pov    1  765074995 2105237936 6804.1
## - pct.bach.deg     1 1104554723 2444717664 6869.9
```

C.3.1.3 Forward AIC

```
mint <- lm(per.cap.income~1,data=cdi)
forwardAIC <- step(mint,scope=list(lower=~1,
                                   upper=~log(land.area) + I(pop.18_34^2) +
                                             log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                             log(per.cap.crimes) + region),
                   direction="forward", data = cdi)
```

```
## Start:  AIC=7312.69
## per.cap.income ~ 1
##
##                        Df  Sum of Sq        RSS    AIC
## + pct.bach.deg          1 3497562202 3735858256 7024.0
## + pct.below.pov         1 2619026410 4614394048 7116.9
## + log(doctors)          1 1802674458 5430745999 7188.6
## + pct.unemp             1  750662754 6482757704 7266.5
```

```
## + log(land.area)       1  678306626 6555113832 7271.4
## + region               3  614711894 6618708564 7279.6
## <none>                              7233420458 7312.7
## + log(per.cap.crimes)  1   32524815 7200895643 7312.7
## + I(pop.18_34^2)       1   18692245 7214728213 7313.6
##
## Step:  AIC=7023.97
## per.cap.income ~ pct.bach.deg
##
##                       Df  Sum of Sq        RSS    AIC
## + I(pop.18_34^2)       1 1153543836 2582314420 6863.5
## + pct.below.pov        1  876387334 2859470923 6908.3
## + region               3  529246738 3206611519 6962.8
## + log(doctors)         1  314532819 3421325437 6987.3
## + log(land.area)       1  189900740 3545957517 7003.0
## + log(per.cap.crimes)  1   86165996 3649692260 7015.7
## + pct.unemp            1   29796055 3706062201 7022.4
## <none>                              3735858256 7024.0
##
## Step:  AIC=6863.48
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2)
##
##                       Df Sum of Sq        RSS    AIC
## + pct.below.pov        1 458583482 2123730939 6779.5
## + region               3 374723689 2207590732 6800.5
## + log(land.area)       1 239261873 2343052548 6822.7
## + log(doctors)         1 221458704 2360855716 6826.0
## + pct.unemp            1  16452726 2565861694 6862.7
## + log(per.cap.crimes)  1  15610196 2566704225 6862.8
## <none>                             2582314420 6863.5
##
## Step:  AIC=6779.45
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov
##
##                       Df Sum of Sq        RSS    AIC
## + log(doctors)         1 465281677 1658449262 6672.6
## + log(land.area)       1 188140065 1935590874 6740.6
## + region               3 171406270 1952324669 6748.4
## + pct.unemp            1 122574509 2001156430 6755.3
## + log(per.cap.crimes)  1  49006033 2074724906 6771.2
## <none>                             2123730939 6779.5
##
## Step:  AIC=6672.64
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov +
##     log(doctors)
##
##                       Df Sum of Sq        RSS    AIC
## + log(land.area)       1 167775482 1490673780 6627.7
## + pct.unemp            1 118570739 1539878523 6642.0
## + region               3 114408451 1544040811 6647.2
## <none>                             1658449262 6672.6
## + log(per.cap.crimes)  1   2876256 1655573006 6673.9
##
## Step:  AIC=6627.72
```

```
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov +
##     log(doctors) + log(land.area)
##
##                        Df Sum of Sq        RSS    AIC
## + pct.unemp             1 150510839 1340162941 6582.9
## + region                3  58768000 1431905780 6616.0
## <none>                               1490673780 6627.7
## + log(per.cap.crimes)   1   4794461 1485879320 6628.3
##
## Step:  AIC=6582.88
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov +
##     log(doctors) + log(land.area) + pct.unemp
##
##                        Df Sum of Sq        RSS    AIC
## <none>                               1340162941 6582.9
## + region                3  15655398 1324507543 6583.7
## + log(per.cap.crimes)   1    952456 1339210485 6584.6
```

C.3.1.4 Forward BIC

```
forwardBIC <- step(mint,scope=list(lower=~1,
                                   upper=~log(land.area) + I(pop.18_34^2) +
                                             log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                             log(per.cap.crimes) + region),
                   direction="forward", data = cdi, k = log(n))
```

```
## Start:  AIC=7316.78
## per.cap.income ~ 1
##
##                        Df  Sum of Sq        RSS    AIC
## + pct.bach.deg          1 3497562202 3735858256 7032.1
## + pct.below.pov         1 2619026410 4614394048 7125.1
## + log(doctors)          1 1802674458 5430745999 7196.7
## + pct.unemp             1  750662754 6482757704 7274.7
## + log(land.area)        1  678306626 6555113832 7279.5
## + region                3  614711894 6618708564 7296.0
## <none>                               7233420458 7316.8
## + log(per.cap.crimes)   1   32524815 7200895643 7320.9
## + I(pop.18_34^2)        1   18692245 7214728213 7321.7
##
## Step:  AIC=7032.14
## per.cap.income ~ pct.bach.deg
##
##                        Df  Sum of Sq        RSS    AIC
## + I(pop.18_34^2)        1 1153543836 2582314420 6875.7
## + pct.below.pov         1  876387334 2859470923 6920.6
## + region                3  529246738 3206611519 6983.2
## + log(doctors)          1  314532819 3421325437 6999.5
## + log(land.area)        1  189900740 3545957517 7015.3
## + log(per.cap.crimes)   1   86165996 3649692260 7028.0
## <none>                               3735858256 7032.1
## + pct.unemp             1   29796055 3706062201 7034.7
##
```

```
## Step:  AIC=6875.74
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2)
##
##                        Df Sum of Sq        RSS    AIC
## + pct.below.pov         1 458583482 2123730939 6795.8
## + region                3 374723689 2207590732 6825.0
## + log(land.area)        1 239261873 2343052548 6839.0
## + log(doctors)          1 221458704 2360855716 6842.4
## <none>                              2582314420 6875.7
## + pct.unemp             1  16452726 2565861694 6879.0
## + log(per.cap.crimes)   1  15610196 2566704225 6879.2
##
## Step:  AIC=6795.8
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov
##
##                        Df Sum of Sq        RSS    AIC
## + log(doctors)          1 465281677 1658449262 6693.1
## + log(land.area)        1 188140065 1935590874 6761.1
## + pct.unemp             1 122574509 2001156430 6775.7
## + region                3 171406270 1952324669 6777.0
## + log(per.cap.crimes)   1  49006033 2074724906 6791.6
## <none>                              2123730939 6795.8
##
## Step:  AIC=6693.08
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov +
##     log(doctors)
##
##                        Df Sum of Sq        RSS    AIC
## + log(land.area)        1 167775482 1490673780 6652.2
## + pct.unemp             1 118570739 1539878523 6666.5
## + region                3 114408451 1544040811 6679.9
## <none>                              1658449262 6693.1
## + log(per.cap.crimes)   1   2876256 1655573006 6698.4
##
## Step:  AIC=6652.24
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov +
##     log(doctors) + log(land.area)
##
##                        Df Sum of Sq        RSS    AIC
## + pct.unemp             1 150510839 1340162941 6611.5
## <none>                              1490673780 6652.2
## + region                3  58768000 1431905780 6652.8
## + log(per.cap.crimes)   1   4794461 1485879320 6656.9
##
## Step:  AIC=6611.49
## per.cap.income ~ pct.bach.deg + I(pop.18_34^2) + pct.below.pov +
##     log(doctors) + log(land.area) + pct.unemp
##
##                        Df Sum of Sq        RSS    AIC
## <none>                              1340162941 6611.5
## + log(per.cap.crimes)   1    952456 1339210485 6617.3
## + region                3  15655398 1324507543 6624.6
```

Unsurprisingly, all stepwise method suggest us to drop region factor and log(per capita serious crimes).

C.3.2 LASSO

```
X <- cdi %>% mutate(
                log_land.area = log(land.area),
                pop.18_34_squared = pop.18_34^2,
                log_doctors = log(doctors),
                log_per.cap.crimes = log(per.cap.income)
            ) %>%
            select(
              log_land.area,
              pop.18_34_squared,
              log_doctors,
              pct.bach.deg,
              pct.below.pov,
              pct.unemp,
              log_per.cap.crimes,
              region
            )
X <- fastDummies::dummy_cols(X) %>% select(-region)
```

```
LASSO_cv = cv.glmnet(as.matrix(X), cdi$per.cap.income, alpha = 1)
LASSO = glmnet(as.matrix(X), cdi$per.cap.income, alpha = 1, lambda = LASSO_cv$lambda.min)
LASSO$beta
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                s0
## log_land.area          .
## pop.18_34_squared     -0.2668745
## log_doctors         -172.6366600
## pct.bach.deg          46.4688456
## pct.below.pov         68.1513389
## pct.unemp             45.4738490
## log_per.cap.crimes 19727.4226813
## region_NC           -116.5305155
## region_NE            108.2296434
## region_S               .
## region_W               .
```

However, LASSO decides to keep both per capita serious crime and region but drops the predictor land area.

C.4 Final Selection

After the analysis, our final selection comes down to 3 candidates,
$1^{st} Model$: predictor incldues log(Land Area), *the percentage of* $age 18-34^2$, log(the number of phyisicians), the percentage of bachelor degree, the percentage of population below poverty, the percentage of unemployment, the log(per capita serious crimes) and factor region.

$2^{nd} Model$: The decision of stepwise selection, predictor incldues log(Land Area), *the percentage of* $age 18-34^2$, log(the number of phyisicians), the percentage of bachelor degree, the percentage of population below poverty, and the percentage of unemployment.

$3^{rd} Model$: The LASSO selection, predictor incldues *the percentage of* $age 18-34^2$, log(the number of phyisicians), the percentage of bachelor degree, the percentage of population below poverty, the percentage of unemployment, the log(per capita serious crimes) and factor region.

C.4.1 Model 1

```
M1 = lm(per.cap.income ~ log(land.area) + I(pop.18_34)^2 +
                                          log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                          log(per.cap.crimes) + region, data = cdi)
paste("AIC: ", AIC(M1) %>%round(2))
```

```
## [1] "AIC:  7837.3"
```

```
n = length(M1$residuals)
npar = length(M1$coefficients) +1
paste("AIC_c: ", (AIC(M1)+2*npar*(npar+1)/(n-npar-1)) %>%round(2))
```

```
## [1] "AIC_c:  7838.03"
```

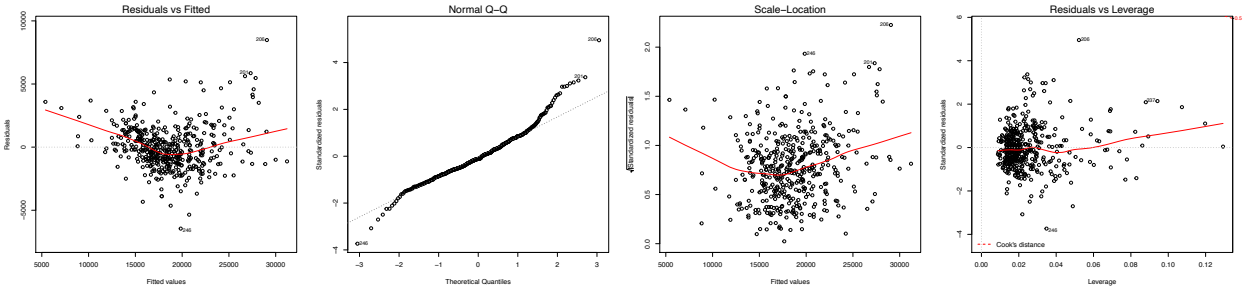
```
paste("BIC: ",BIC(M1) %>%round(2))
```

```
## [1] "BIC:  7886.34"
```

```
paste("R-squared: ",summary(M1)$adj.r.squared %>%round(4))
```

```
## [1] "R-squared:  0.8122"
```

```
par(mfrow = c(1,4))
plot(M1)
```



C.4.2 Model 2

```
M2 = lm(per.cap.income ~ log(land.area) + I(pop.18_34)^2 +
                                          log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp, data =
paste("AIC: ", AIC(M2) %>%round(2))
```

```
## [1] "AIC:  7835.29"
```

```
n = length(M2$residuals)
npar = length(M2$coefficients) +1
paste("AIC_c: ", (AIC(M2)+2*npar*(npar+1)/(n-npar-1)) %>%round(2))
```

```
## [1] "AIC_c:  7835.63"
```

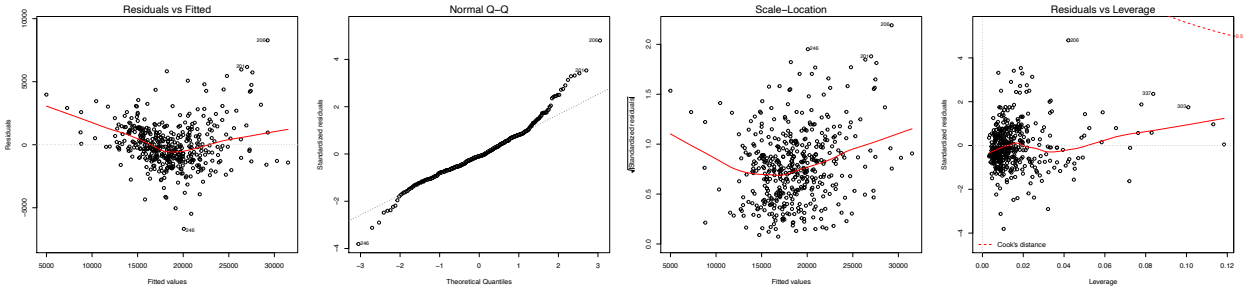
```
paste("BIC: ",BIC(M2) %>%round(2))
```

```
## [1] "BIC:  7867.99"
```

```
paste("R-squared: ",summary(M2)$adj.r.squared %>%round(4))
```

```
## [1] "R-squared:  0.8114"
```

```
par(mfrow = c(1,4))
plot(M2)
```



C.4.3 Model 3

```
M3 = lm(per.cap.income ~ I(pop.18_34)^2 + log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp +
                                        log(per.cap.crimes) + region, data = cdi)
paste("AIC: ", AIC(M3) %>%round(2))
```

```
## [1] "AIC:  7874.24"
```

```
n = length(M3$residuals)
npar = length(M3$coefficients) +1
paste("AIC_c: ", (AIC(M3)+2*npar*(npar+1)/(n-npar-1)) %>%round(2))
```

```
## [1] "AIC_c:  7874.86"
```

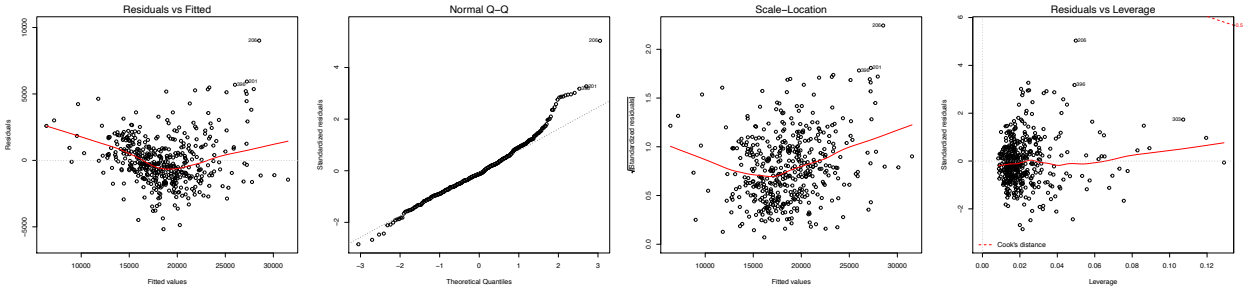
```
paste("BIC: ",BIC(M3) %>%round(2))
```

```
## [1] "BIC:  7919.2"
```

```
paste("R-squared: ",summary(M3)$adj.r.squared %>%round(4))
```

```
## [1] "R-squared:  0.7953"
```

```
par(mfrow = c(1,4))
plot(M3)
```



The diagnostics plots look similar for 3 models; although there are mean structures among residuals, these structures are acceptable; the normal Q-Q plots look fine despite some tails in either sides; the $\sqrt{Standardized\ Residuals}$ vs fitted values plots show also follow the constant variance assumption; there are no bad influential points in the model.

Therefore, our decision will be soley made on AIC, AIC_c, BIC and R^2_{adj}.

C.5 Decision

Our final selection is model 2, who has the lower AIC, AIC_c, BIC and a similar R^2_{adj} to Model 1.

```
M2 %>% summary
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + I(pop.18_34)^2 +
##     log(doctors) + pct.bach.deg + pct.below.pov + pct.unemp,
##     data = cdi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6676.3 -1054.7  -163.4   951.7  8283.0
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    20291.10    1081.32  18.765  < 2e-16 ***
## log(land.area)  -804.17      99.85  -8.054 7.83e-15 ***
## I(pop.18_34)    -305.46      23.87 -12.798  < 2e-16 ***
## log(doctors)    1059.00      85.60  12.372  < 2e-16 ***
## pct.bach.deg     318.79      16.94  18.823  < 2e-16 ***
## pct.below.pov   -350.68      22.34 -15.699  < 2e-16 ***
## pct.unemp        312.45      45.19   6.914 1.70e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1763 on 433 degrees of freedom
## Multiple R-squared:  0.814,  Adjusted R-squared:  0.8114
## F-statistic: 315.8 on 6 and 433 DF,  p-value: < 2.2e-16
```

The model is,
$Per\ capita\ income = 20291.1 - 804.17 \cdot log(LandArea) - 305.46 \cdot (the\ percent\ of\ population\ aged\ 18-34)^2 +$
$1059.9 \cdot log(the\ number\ of\ physicians) + 318.79 \cdot the\ percent\ bachelor's\ degree - 350.68 \cdot the\ percent\ population\ below\ poverty\ le$
$312.45 \cdot the\ percent\ unemployment$

```r
rms::vif(M2)
```

```
## log(land.area)   I(pop.18_34)   log(doctors)   pct.bach.deg  pct.below.pov
##       1.070246       1.413609       1.354909       2.374313       1.528688
##      pct.unemp
##       1.577237
```