

Analysis on how Average Income per person Related to the County's Economic, Health and Social Well-being

Yucheng Wang
Department of Statistics and Data Science,
Carnegie Mellon University
yw6@andrew.cmu.edu

Abstract

We addressed the problem of what factors could have the largest effect on per-capita income in a county and learned how average income per person was related to other variables associated with the county's economic, health, and social well-being. We examined the CDI(county demographic information) data taken from Kutner et al.(2005)[\[1\]](#). From the correlation plot, it appeared that correlation exists between some of the variables in the dataset. We checked the effect of the numbers of crimes and per-capita crimes on the per-capita income, and by applying the linear regression model, we found that the number of crimes is not an influential factor. Also by building a proper linear regression model, we found that the number of doctors, the overall educational background, and the age composition of the population was the three most important factors affecting the per-capita income. In the end, we discussed the potential flaw of the dataset and found that the samples of the data might be biased since the 440 counties are not representative of all 3000 counties.

1 Introduction

Per-capita income has always been an important indicator of social development. In order to increase and balance people's average income in each county, it is crucial to understand the factors that may affect average income. How average income per person is related to other variables associated with the county's economic, health, and social well-being is an important question for us to answer. The purpose of this paper is, firstly, to discover the correlation between each pair of variables. Secondly, we will analyze the impact of the number of crimes or crime rate on per-capita income. Thirdly, we will find the best linear regression model predicting per-capita income from the other variables(region, crimes, and so on). Further, we will discuss the impact of states and counties that are not included in the data set. Our analysis methods are mainly from [\[2\]](#).

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	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	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor’s degrees	Percent of adult population (persons 25 years old or older) with bachelor’s degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	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)

Table 1: Variables explanation

2 Data

The dataset was taken from Kutner et al. (2005)¹. This dataset provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each row of the dataset has a unique id with a county name and state abbreviation and provides information on 14 variables (table 1) for each county, they are region, land.area, pop, pop.18_34, pop.65_plus, doctors, hosp.beds, crimes, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, per.cap.income and tot.income.

Firstly, table 1 is the explanation of each variables in the dataset, there are 17 variables in this dataset.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	SD
land.area	15.0	451.25	656.50	1041.41	946.75	20062.0	1549.92
pop	100043.0	139027.25	217280.50	393010.92	436064.50	8863164.0	601987.02
pop.18_34	16.4	26.20	28.10	28.57	30.02	49.7	4.19
pop.65_plus	3.0	9.88	11.75	12.17	13.62	33.8	3.99
doctors	39.0	182.75	401.00	988.00	1036.00	23677.0	1789.75
hosp.beds	92.0	390.75	755.00	1458.63	1575.75	27700.0	2289.13
crimes	563.0	6219.50	11820.50	27111.62	26279.50	688936.0	58237.51
pct.hs.grad	46.6	73.88	77.70	77.56	82.40	92.9	7.02
pct.bach.deg	8.1	15.28	19.70	21.08	25.33	52.3	7.65
pct.below.pov	1.4	5.30	7.90	8.72	10.90	36.3	4.66
pct.unemp	2.2	5.10	6.20	6.60	7.50	21.3	2.34
per.cap.income	8899.0	16118.25	17759.00	18561.48	20270.00	37541.0	4059.19
tot.income	1141.0	2311.00	3857.00	7869.27	8654.25	184230.0	12884.32

Table 2: Summary table of continuous variables

Then, the brief summary for each continuous variable could be found in table 2. From this table, we found a right-skewed pattern (the median was smaller than the mean) for some continuous variables (pop, land.area, doctors and etc.), which meant we might need some proper transformation in further analysis.

Table 3 is some of the counties and states pairs in this dataset. There are a total of 440 data points in the dataset, including 373 distinct counties and 48 distinct states, each data point could be represented by a distinct county & state pair.

According to table 4, around 36% of the counties are in the Southern region, around 25% of the counties are in the Northcentral region, around 23% of the counties are in the Northeast region, and around 18% of the counties are in the Western region.

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

Counties 1-110	Counties 111-220	Counties 221-330	Counties 331-440
Ada ID	Ector TX	Lycoming PA	Rockingham NH
Adams CO	El_Dorado CA	Macomb MI	Rockland NY
Aiken SC	El_Paso CO	Macon IL	Rowan NC
Alachua FL	El_Paso TX	Madison AL	Rutherford TN
Alamance NC	Elkhart IN	Madison IL	Sacramento CA
Alameda CA	Erie NY	Madison IN	Saginaw MI
Albany NY	Erie PA	Mahoning OH	Salt_Lake UT
Alexandria_City VA	Escambia FL	Manatee FL	San_Bernardino CA
Allegheny PA	Essex MA	Marathon WI	San_Diego CA
Allen IN	Essex NJ	Maricopa AZ	San_Francisco CA
Allen OH	Fairfax_County VA	Marin CA	San_Joaquin CA
Anderson SC	Fairfield CT	Marion FL	San_Luis_Obispo CA
Androscoggin ME	Fairfield OH	Marion IN	San_Mateo CA
Anne_Arundel MD	Fayette KY	Marion OR	Sangamon IL
Arapahoe CO	Fayette PA	Martin FL	Santa_Barbara CA
Arlington_County VA	Florence SC	Maui HI	Santa_Clara CA
Atlantic NJ	Forsyth NC	McHenry IL	Santa_Cruz CA
Baltimore MD	Fort_Bend TX	McLean IL	Sarasota FL

Table 3: First 18 rows of the counties & states pairs

	NC	NE	S	W
Freq	108	103	152	77

Table 4: Frequency of counties in each region

3 Method

In this paper, we used a correlation plot (figure 1) to find the correlation between different variables and analyzed the correlation thoroughly.

Then, we took log-transformation to some of the variables (land.area, pop, doctors, hosp.beds, crimes, tot.income, per.cap.income, pct.below.pov) in the dataset (see part 6.1).

After that, we performed regression analysis on the impact of the number of crimes and crime rate on per-capita income (see part 6.2) and got the most appropriate variables describing the crime that should be used in the following linear models.

After that, we performed regression analysis on all the variables (per-capita income was the dependent variable) to find out the most important factors that have the greatest impact on per capita income. We examined the VIF of the full model (figure 8) and removed two variables, which were population and total

income(see [6.3](#)). Then, we found the most suitable regression model by evaluating the AIC and BIC of the models. To be more specific, we applied stepAIC and stepBIC(see [6.3](#)) to reduce the full model. By comparing the result of stepAIC and the result of stepBIC, we obtained our optimal linear model and found the variables that have the greatest impact on per capita income. We also evaluated the validity of the models we used through diagnostic plots and the p-value of the estimated coefficients.

4 Result

4.1 Correlation Between Variables

We analyzed the correlation between each pair of variables in the dataset. According to the correlation plot([figure 1](#)), many variables are highly correlated(e.g. the variables doctors and population has a correlation around 0.8). We found that strong linear relationships exist between many variables, especially between population and number of doctors, number of beds in the hospital, number of crimes and total income; between total income and the number of beds in the hospital, and the number of doctors; as well as between percentage of high school graduation and percentage of the bachelor degree, percentage of the people below the poverty line and the percentage of unemployment.

The per-capita income is not so correlated with any of the variables, however slightly correlated with the percentage of high school graduation, the percentage of the bachelor's degree, the percentage of the people below the poverty line, the percentage of unemployment.

We found that the variable population might influence the result of our correlation plot seriously since it was correlated to too many variables and thus had a strong effect on the correlation between other variables.

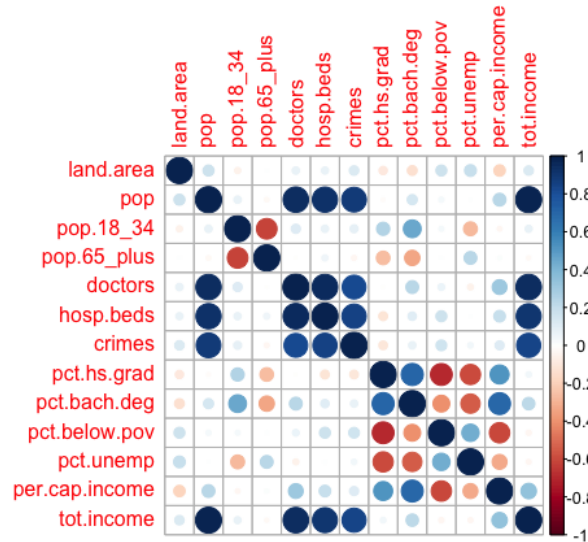


Figure 1: Correlation plot of selected variables

4.2 The Impact of Crime

Then, we analyzed the impact of the number of crimes and crime rate. We decided to take log-transformation for some of the continuous variables land.area, pop, doctors, hosp.beds, crimes, tot.income, per.cap.income in our future analysis(see part 6.1).

We applied F-test on the 6 models below and compared the AIC of the models(see part 6.2).

Model 1: $\log.\text{per.cap.income} \sim \log.\text{crimes}$

Model 2: $\log.\text{per.cap.income} \sim \log.\text{crimes} + \text{region}$

Model 3: $\log.\text{per.cap.income} \sim \log.\text{crimes} * \text{region}$

Model 4: $\log.\text{per.cap.income} \sim \log.\text{per.cap.crimes}$

Model 5: $\log.\text{per.cap.income} \sim \log.\text{per.cap.crimes} + \text{region}$

Model 6: $\log.\text{per.cap.income} \sim \log.\text{per.cap.crimes} * \text{region}$

We find out that the best model here is

$$\log.\text{per.cap.income} \sim \log.\text{crimes} + \text{region}.$$

Thus, we found that the log total number of crimes was the variable, which could properly describe crime in a specific county. According to the of figure 7 in part 6.2, we found that when the number of crimes increases by 1%, the per-capita income would increase by about 0.07%, thus we found that the number of crimes is not an important variable in predicting the per-capita income, since although it was statistically significant, its estimated coefficient is too low. Also, we found that region was an important variable in predicting the per-capita income here, the per-capita income was higher in NE(for more details see part 6.2).

4.3 Modeling on the most Influential Variables

After analyzing the effect of the number of crimes, we fitted a full model with all possible variables to find variables, which have the greatest impact on the per capita income. Then by VIF, we removed the variables total income and population from our model(see part 6.3). And by applying AIC and BIC criteria(see part 6.3), we found the best linear regression model(figure 12). The final model we chose is the model below, which is a good model that contained 7 variables, and all of them were significant.

$$\begin{aligned}\log.\text{per.cap.income} &= 10.22 - 0.014\text{pop.18.34} - 0.004\text{pct.hs.grad} + 0.015\text{pct.bach.deg} \\ &- 0.024\text{pct.below.pov} + 0.011\text{pct.unemp} - 0.036\log.\text{land.area} + 0.061\log.\text{doctors}\end{aligned}$$

The model told us that, for every 1% increase of the number of doctors, the per-capita income would increase 0.06%, which was reasonable since doctors were well-paid and could contribute the per capita income.

For every 1% increase of the land area, the per-capita income would decrease 0.036%, which was reasonable, since the rural area usually tends to be much larger than the city area and the per-capita income in rural may be much lower than that in the city area.

For every 1% increase of the population of 18 to 34 years old, we would tend to expect the per-capita income would multiply by the factor $\exp(-0.014) = 0.986$, which is reasonable, since the population of the age group 18 to 34 might have lower income than the other age group.

For every 1% increase of the population of the high school graduates, we would tend to expect the per-capita income would multiply by the factor $\exp(-0.0044) = 0.996$, which was reasonable, since it seemed that the change of population of the percentage of high school graduates was not an important factor in predicting the per-capita income since 0.996 was close to 1.

For every 1% increase of the population of the bachelor's degrees, we would tend to expect the per-capita income would multiply by the factor $\exp(0.015) = 1.015$, which is reasonable, since higher education often meant higher salary.

For every 1% increase of the population of the below poverty level, we would tend to expect the per-capita income would multiply by the factor $\exp(-0.024) = 0.976$, which is reasonable.

5 Discussion

5.1 Correlation Between Variables

According to the result (figure [1](#)), we find something reasonable as well as something surprising. The number of doctors and the number of hospital beds are highly positively correlated, which is reasonable because the two are a reflection of the local medical level. More doctors usually mean more beds in most cases. The total income is highly positively correlated with the population, which is also reasonable since more people often means more total income within a county. The positive correlation between the total number of crimes and the population can also be explained in this way.

The high correlation between the number of crimes and the number of doctors seemed to be surprising, but it might be because of the indirect effect of the population.

5.2 The Impact of Crime and region

According to the result, we found that for the impact of the number of crimes, if we ignored other variables, per-capita income is not so related to the number of crimes or the crime rate in different regions. It was showing that the crime rate itself was not an influential factor in predicting the per-capita income, for example, Los Angeles in CA was a place with high crimes number and high per-capita income.

Moreover, we found that the region was an important variable, the per capita income in different regions could vary significantly, a county in the region North East tends to have the highest per capita income. The per-capita income shows an unbalanced pattern in different regions.

5.3 Best model and the most Influential Variables

For the most influential variables, the result of our regression model in part [6.3](#) told us that the baseline per-capita income in a county is 27447\$, and the most influential variable is the number of doctors since

the doctor is a well-paid job, it is indeed an important indicator of per-capita income in a county. Also, the model result showed that education background is an important factor affecting per capita income, which is understandable that people with high educational backgrounds usually tend to have higher income. Another factor was the composition of the population, in most cases, the income of young adults was often much lower than that of middle-aged people, so a county that contains more young adults usually had lower per-capita income. The proportion of the population below the poverty line is a simple indicator of per-capita income. The greater the number of poor people, the lower the per-capita income. In sum, we found that the number of doctors, the overall educational background, and the age composition of the population was the three most important factors affecting the per-capita income.

One potential problem here was the sign of the coefficient of pct.unemp in this model, which is positive. We expected it to be negative since the unemployment rate might reasonably lower the per-capita income. We tried to add interaction terms in the model, however, we cannot change its sign. Thus, we might need more information about this variable.

5.4 Potential bias in the dataset, limitations, and future work

One thing worth thinking about is that our dataset includes a total of 440 of the 3000 counties, but not all counties. This is likely to have a huge deviation in our analysis results. First of all, it is easy to understand that incomplete samples may bias our analysis results. If the 440 samples are not collected randomly, our analysis might be a biased one, since the samples cannot represent the overall situation. To determine whether we should worry about the missing counties, we need to know whether our 440 samples could represent all 3000 counties(the population). According to the information we have, the 440 counties were the most populous in the United States and were not sampled randomly. Thus we can assert that the 440 samples of the 3000 counties are seriously biased(and have not been chosen randomly) and not representative for all 3000 counties. It means that the result and conclusion of our analysis might also be biased and may not be correct for each county in the United States.

Further, there are some possible improvements in our analysis. From the correlation plot([figure 1](#)), we found that some of the variables were correlated, which might because of the indirect effect of the variable population. This evidence told us that multicollinearity might exist in our data. We solved this problem by removing two variables population and total income. However, we might use a shrinkage regression model like lasso regression or ridge regression to analyze our data, by using such methods, the multicollinearity problem could be fixed and we could get a simple and explicit model. Moreover, we could apply principal component analysis before we do the regression analysis. By PCA, we could also solve the multicollinearity problem, and get a simpler model, however, the model might not be so explainable.

Also, the data we used in our analysis was an old one, which is around 30 years old. Thus our result might also be outdated, for improvement, more recent data should be used in future analysis.

In the end, although there was some limitation in our analysis, our model still provided a convincing insight into the per capita income in the United States. And also found how average income per person related to the county's economy, health, region, educational background, and social well-being.

References

- [1] Kutner, M.H., Nachsheim, C.J., J. Neter, Li, and W. *Applied Linear Statistical Models, Fifth Edition*. NY: McGrawHill/Irwin, 2005.
- [2] Simon J. Sheather. *A Modern Approach to Regression*. Springer, 2009.

6 Technical Appendix

6.1 Appendix A Transformation

We first visualized all the continuous variables by the histogram (figure 2). According to the figure 2, we found that land.area, pop, doctors, hosp.beds, crimes, tot.income, per.cap.income, pct.below.pov were right-skewed seriously, thus we took log-transformation on these variables.

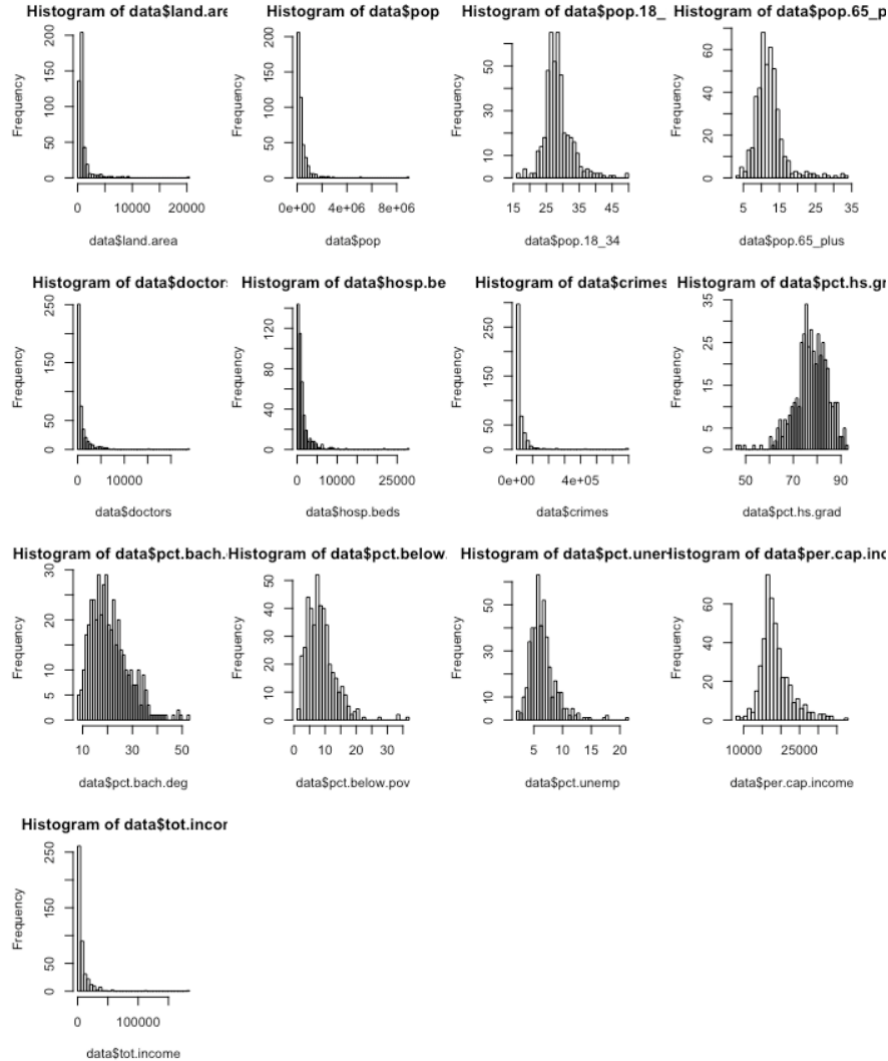


Figure 2: Histograms for continuous variables (before transformation)

The histograms of the variables after transformation is in figure 3. After transformation, we found almost all the continuous variables were unimodal and nearly normal distributed. We have to admit that transformation might lower the interpretation ability of the model, but for the overall performance of the model, we still chose to transform the variables by log transformation.

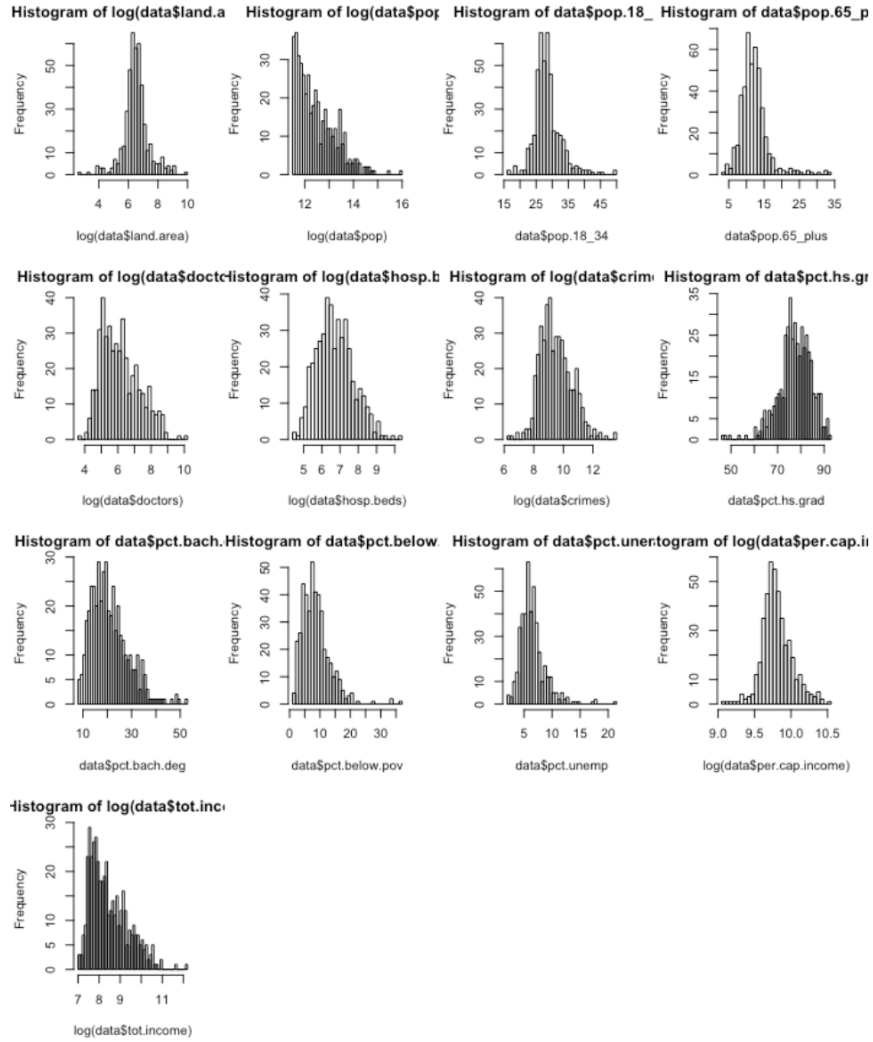


Figure 3: Histograms for continuous variables(after transformation)

6.2 Appendix B Analysis of the impact of crimes

In this section, we analysed the impact of the numbers of crimes of crime rate. We fitted three models with variables region and log.crimes as follow,

Model 1: $\log.\text{per.cap.income} \sim \log.\text{crimes}$
 Model 2: $\log.\text{per.cap.income} \sim \log.\text{crimes} + \text{region}$
 Model 3: $\log.\text{per.cap.income} \sim \log.\text{crimes} * \text{region}$

By apply F-test on the three model above (figure 4), we find that, we should choose model 2.

Then we defined a new variable $\log.\text{per.cap.crimes} = \log(\text{crimes}/\text{pop})$, which was per-capita crimes (we took log since crimes/pop right skewed seriously). After that we also fitted three models,

Model 4: $\log.\text{per.cap.income} \sim \log.\text{per.cap.crimes}$
 Model 5: $\log.\text{per.cap.income} \sim \log.\text{per.cap.crimes} + \text{region}$
 Model 6: $\log.\text{per.cap.income} \sim \log.\text{per.cap.crimes} * \text{region}$

Also by applied F-test on this three models (figure 5) we find that the best model within the three models was model 5.

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.crimes
Model 2: log.per.cap.income ~ region + log.crimes
Model 3: log.per.cap.income ~ region * log.crimes
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     438 17.271
2     435 14.949  3    2.32194 22.4823 1.523e-13 ***
3     432 14.872  3    0.07678  0.7434  0.5266
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: F-test for model 1,2,3

```
Analysis of Variance Table

Model 1: log.per.cap.income ~ log.per.cap.crimes
Model 2: log.per.cap.income ~ region + log.per.cap.crimes
Model 3: log.per.cap.income ~ region * log.per.cap.crimes
  Res.Df    RSS Df Sum of Sq    F    Pr(>F)
1     438 18.697
2     435 16.952  3    1.74465 14.8407 3.263e-09 ***
3     432 16.928  3    0.02408  0.2048  0.893
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 5: F-test for model 4,5,6

Since the R-squared of the two models are similar, we chose the model with lower AIC (figure 6), which was model 2 (figure 7). This result meant that the total number of crimes might be a better choice in fitting the per.cap.income. Thus we used this variable in our further analysis.

	df <dbl>	AIC <dbl>
model2	6	-227.4746
model5	6	-172.1347

Figure 6: AIC for model 2,5

According to the of figure 7, we found that when the number of crimes increases by 1%, the per-capita income would increase by about 0.07%, thus we found that the number of crimes is not an important variable in predicting the per-capita income, since although it was statistically significant, its estimated coefficient is too low. And in this model, region is an important factor in predicting the per-capita income, our baseline per-capita income in this model is $\exp(9.19) = 9799\$$ (in NC), in NE we have our per-capita income= $9799 \times \exp(0.1) = 10829\$$, in S we have our per-capita income= $9799 \times \exp(-0.087) = 8955\$$, in W we have our per-capita income= $9799 \times \exp(-0.055) = 9228\$$.

```
lm(formula = log.per.cap.income ~ region + log.crimes, data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-0.68757 -0.10557 -0.01422  0.08905  0.78946

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.188431    0.079812  115.125  < 2e-16 ***
regionNE      0.104458    0.025531   4.091 5.11e-05 ***
regionS     -0.086983    0.023618  -3.683 0.00026 ***
regionW     -0.055280    0.028167  -1.963 0.05033 .
log.crimes    0.066695    0.008421   7.920 2.00e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1854 on 435 degrees of freedom
Multiple R-squared:  0.2032,    Adjusted R-squared:  0.1959
F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
```

Figure 7: Summary for model 2

6.3 Appendix C Regression analysis on the impact

In this part, we fitted our full regression model (figure 8) with all the variables except county, state, and id. According to the result (figure 8), we find that only log.pop, and log.tot.income were significant, it might be because of the multicollinearity or the deterministic relationship in some variables (total income, per-capita income, population). We first checked the VIF of this model (figure 9). According to the VIF, we found that the VIF of log.pop and log.tot.income was very high, also we have a deterministic relationship $\text{tot.income} = \text{pop} \times \text{per.cap.income}$ thus, we should remove log.tot.income and log.pop in our analysis. Thus our first reduced model (figure 10) here was much better since many of the variables in the model were significant. Also, our new VIF (figure 11) seemed to be much better too, although there are still two variables (log.hosp.beds, log.doctors) that has VIF larger than 10, but we chose not to remove any of them

since the multicollinearity is not so influential here.

```

Coefficients:
(Intercept)      1.382e+01  3.010e-04  45893.977  <2e-16 ***
pop.18_34        4.306e-07  2.037e-06    0.211    0.833
pop.65_plus      2.530e-07  1.921e-06    0.132    0.895
pct.hs.grad     -1.526e-07  1.592e-06   -0.096    0.924
pct.bach.deg      7.351e-07  1.743e-06    0.422    0.673
pct.below.pov    -6.753e-07  2.603e-06   -0.259    0.795
pct.unemp        3.180e-06  3.241e-06    0.981    0.327
regionNE        -5.810e-06  1.682e-05   -0.345    0.730
regionS         -2.812e-06  1.649e-05   -0.171    0.865
regionW         7.912e-06  2.048e-05    0.386    0.699
log.land.area    -6.637e-06  7.623e-06   -0.871    0.384
log.pop         -1.000e+00  6.536e-05 -15299.141  <2e-16 ***
log.doctors     -2.265e-05  1.931e-05   -1.173    0.242
log.hosp.beds    9.098e-06  1.724e-05    0.528    0.598
log.crimes       1.187e-05  1.510e-05    0.786    0.432
log.tot.income   1.000e+00  6.464e-05  15470.149  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0001072 on 424 degrees of freedom
Multiple R-squared:  1,    Adjusted R-squared:  1
F-statistic: 1.089e+08 on 15 and 424 DF,  p-value: < 2.2e-16

```

Figure 8: Full model

	GVIF	Df	GVIF ^{1/(2*Df)}
pop.18_34	2.786384	1	1.669246
pop.65_plus	2.247053	1	1.499017
pct.hs.grad	4.765290	1	2.182955
pct.bach.deg	6.798932	1	2.607476
pct.below.pov	5.614543	1	2.369503
pct.unemp	2.194333	1	1.481328
region	3.857097	3	1.252305
log.land.area	1.687575	1	1.299067
log.pop	102.006746	1	10.099839
log.doctors	18.660929	1	4.319830
log.hosp.beds	11.436936	1	3.381854
log.crimes	10.197704	1	3.193384
log.tot.income	128.688195	1	11.344082

Figure 9: Variance inflation factor of the full model

After that, we thought about finding the best model by applying some model selection criteria. Here we chose BIC and AIC criteria, and compare the two models obtained by the two criteria.

The model obtained by stepBIC is in figure [12](#) and the model obtained by stepAIC is in figure [13](#).

According to the diagnostic plots([figure 14](#) and [15](#)), we found that both models met the linear model assumptions, however, both contained some outliers like data points with index 246.

Comparing the two model results, we find that the model obtained by stepBIC contains fewer variables than the model obtained by stepAIC. Further, the Adjusted R-squared for each model was very close(0.8427

and 0.8459). Thus, we chose the model obtained by stepBIC to be our optimal model, since the model obtained by BIC was a simpler model(more explainable) with almost the same Adjusted R-squared with the model obtained by stepAIC.

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.353232   0.116270  89.045 < 2e-16 ***
pop.18_34    -0.015478   0.001307 -11.840 < 2e-16 ***
pop.65_plus  -0.002648   0.001392  -1.903 0.057700 .
pct.hs.grad  -0.005551   0.001172  -4.737 2.96e-06 ***
pct.bach.deg  0.016348   0.001057  15.468 < 2e-16 ***
pct.below.pov -0.024063   0.001413 -17.033 < 2e-16 ***
pct.unemp     0.008846   0.002380   3.717 0.000229 ***
regionNE     -0.003192   0.012683  -0.252 0.801414
regionS      -0.031864   0.012271  -2.597 0.009740 **
regionW      -0.014019   0.015422  -0.909 0.363849
log.land.area -0.035003   0.005420  -6.459 2.89e-10 ***
log.doctors   0.047877   0.013146   3.642 0.000304 ***
log.hosp.beds 0.008624   0.013050   0.661 0.509042
log.crimes    0.005828   0.008949   0.651 0.515255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08124 on 426 degrees of freedom
Multiple R-squared:  0.8502,    Adjusted R-squared:  0.8456
F-statistic: 185.9 on 13 and 426 DF,  p-value: < 2.2e-16

```

Figure 10: Reduced model 1

	GVIF	Df	GVIF ^{1/(2*Df)}
pop.18_34	1.997041	1	1.413167
pop.65_plus	2.053608	1	1.433041
pct.hs.grad	4.495620	1	2.120288
pct.bach.deg	4.353907	1	2.086602
pct.below.pov	2.878962	1	1.696750
pct.unemp	2.059659	1	1.435151
region	3.580885	3	1.236892
log.land.area	1.484666	1	1.218469
log.doctors	15.046783	1	3.879018
log.hosp.beds	11.403602	1	3.376922
log.crimes	6.238869	1	2.497773

Figure 11: VIF of reduced model 1

```

Call:
lm(formula = log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
    pct.below.pov + pct.unemp + log.land.area + log.doctors,
    data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-0.34147 -0.04886 -0.00538  0.04818  0.26969

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2224950  0.0931210  109.776 < 2e-16 ***
pop.18_34    -0.0139002  0.0011113  -12.508 < 2e-16 ***
pct.hs.grad  -0.0044064  0.0010823  -4.071 5.56e-05 ***
pct.bach.deg  0.0153853  0.0009246  16.641 < 2e-16 ***
pct.below.pov -0.0242784  0.0012583  -19.294 < 2e-16 ***
pct.unemp     0.0106037  0.0021771   4.871 1.56e-06 ***
log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
log.doctors   0.0606769  0.0040183  15.100 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.082 on 432 degrees of freedom
Multiple R-squared:  0.8452,    Adjusted R-squared:  0.8427
F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16

```

Figure 12: Reduced model obtain by step BIC

```

Residuals:
    Min       1Q   Median       3Q      Max
-0.34849 -0.04695 -0.00502  0.04524  0.28624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.3851173  0.1105475  93.943 < 2e-16 ***
pop.18_34    -0.0153941  0.0013021  -11.822 < 2e-16 ***
pop.65_plus  -0.0026499  0.0013137   -2.017 0.04430 *
pct.hs.grad  -0.0055059  0.0011696  -4.707 3.39e-06 ***
pct.bach.deg  0.0159212  0.0009688  16.434 < 2e-16 ***
pct.below.pov -0.0238604  0.0013529  -17.637 < 2e-16 ***
pct.unemp     0.0090479  0.0023017   3.931 9.86e-05 ***
regionNE     -0.0061091  0.0123398   -0.495 0.62080
regionS      -0.0311704  0.0114050   -2.733 0.00654 **
regionW      -0.0162724  0.0140361   -1.159 0.24697
log.land.area -0.0346133  0.0053943  -6.417 3.70e-10 ***
log.doctors   0.0608452  0.0041649  14.609 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08115 on 428 degrees of freedom
Multiple R-squared:  0.8498,    Adjusted R-squared:  0.8459
F-statistic: 220.1 on 11 and 428 DF,  p-value: < 2.2e-16

```

Figure 13: Reduced model obtain by step AIC

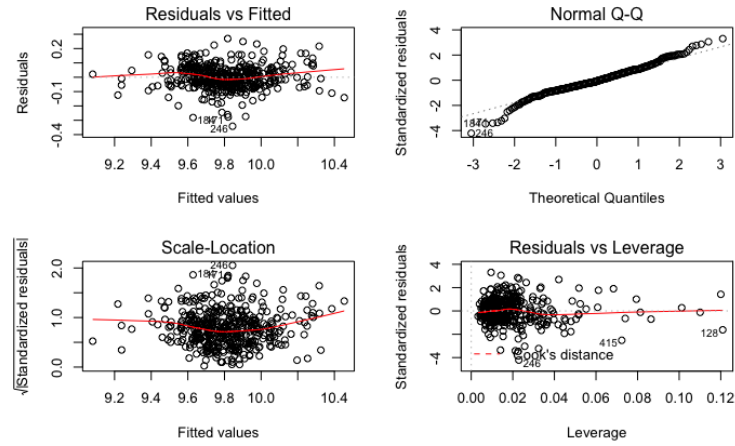


Figure 14: Diagnostic plot for model obtained by stepBIC

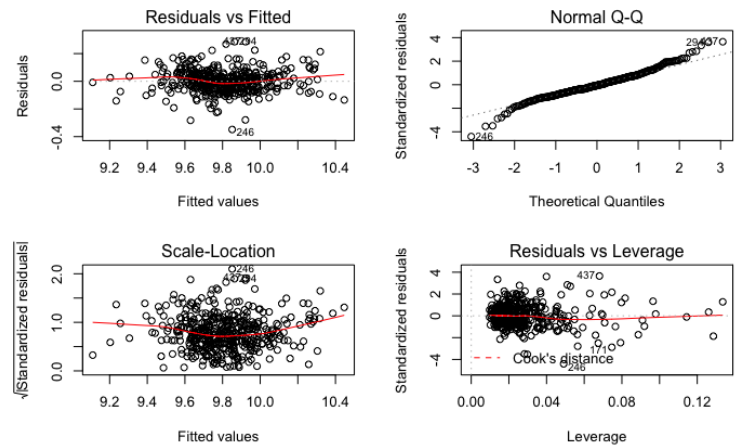


Figure 15: Diagnostic plot for model obtained by stepAIC

Technical Appendix Code

Yucheng Wang yw6

10/28/2021

```
library(car)

## Loading required package: carData

library(corrplot)

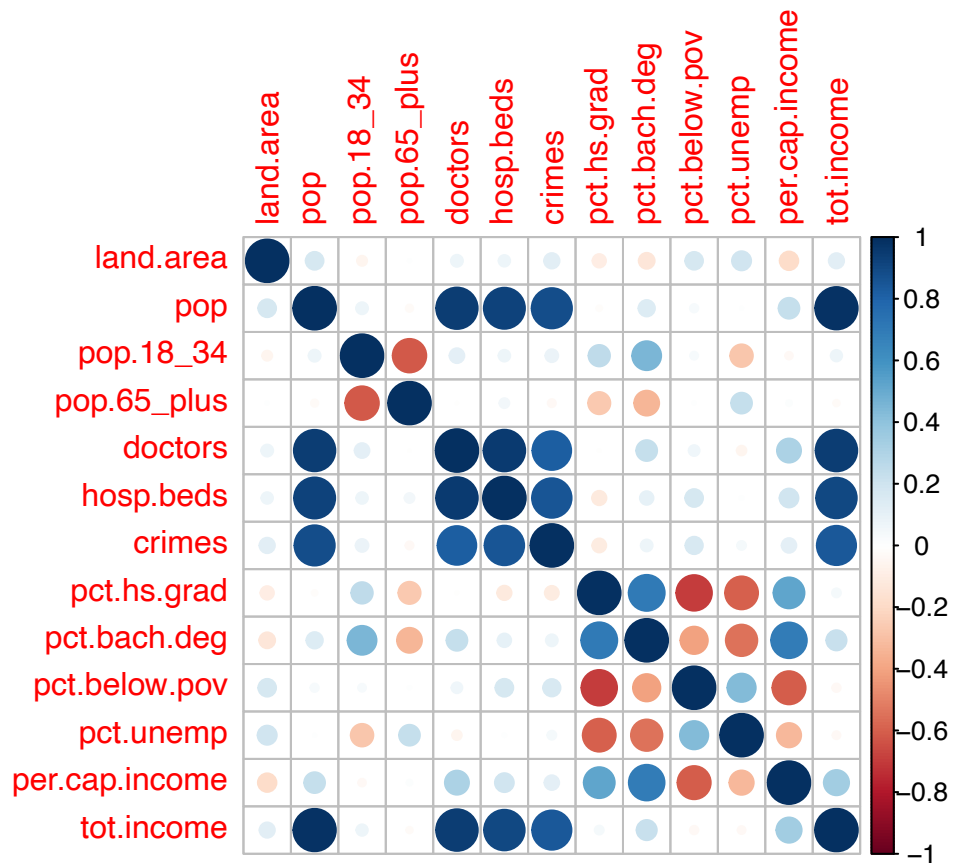
## corrplot 0.84 loaded

library(MASS)
data = read.csv("/Users/wyc/cdi.dat", sep=" ")
head(data)

##      id      county state land.area      pop pop.18_34 pop.65_plus doctors
## 1  1 Los_Angeles    CA      4060 8863164      32.1      9.7  23677
## 2  2      Cook     IL       946 5105067      29.2     12.4  15153
## 3  3      Harris    TX      1729 2818199      31.3      7.1   7553
## 4  4 San_Diego     CA      4205 2498016      33.5     10.9   5905
## 5  5      Orange    CA       790 2410556      32.6      9.2   6062
## 6  6      Kings    NY        71 2300664      28.3     12.4   4861
##      hosp.beds crimes pct.hs.grad pct.bach.deg pct.below.pov pct.unemp
## 1      27700 688936      70.0      22.3      11.6      8.0
## 2      21550 436936      73.4      22.8      11.1      7.2
## 3      12449 253526      74.9      25.4      12.5      5.7
## 4       6179 173821      81.9      25.3      8.1      6.1
## 5       6369 144524      81.2      27.8      5.2      4.8
## 6       8942 680966      63.7      16.6     19.5      9.5
##      per.cap.income tot.income region
## 1          20786      184230      W
## 2          21729      110928      NC
## 3          19517       55003      S
## 4          19588       48931      W
## 5          24400       58818      W
## 6          16803       38658      NE

data2 = data[, -c(1,2,3,17)]

corrplot(cor(data2))
```

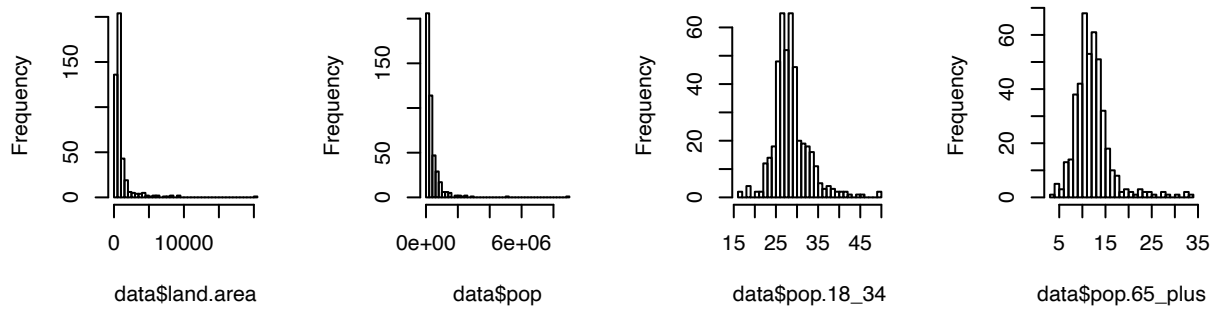


```

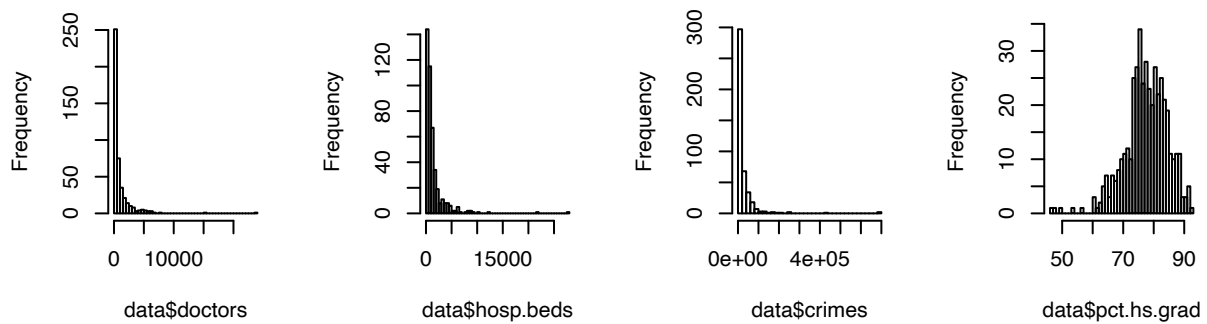
par(mfrow = c(2,4))
hist(data$land.area,breaks=30)
hist(data$pop,breaks=40)
hist(data$pop.18_34,breaks=40)
hist(data$pop.65_plus,breaks=40)
hist(data$doctors,breaks=40)
hist(data$hosp.beds,breaks=40)
hist(data$crimes,breaks=40)
hist(data$pct.hs.grad,breaks=40)

```

Histogram of data\$land.area Histogram of data\$pop Histogram of data\$pop.18_34 Histogram of data\$pop.65_plus

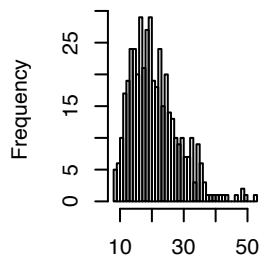


Histogram of data\$doctors Histogram of data\$hosp.beds Histogram of data\$crimes Histogram of data\$pct.hs.grad

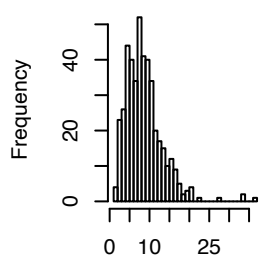


```
hist(data$pct.bach.deg,breaks=40)
hist(data$pct.below.pov,breaks=40)
hist(data$pct.unemp,breaks=40)
hist(data$per.cap.income,breaks=40)
hist(data$tot.income,breaks=40)
```

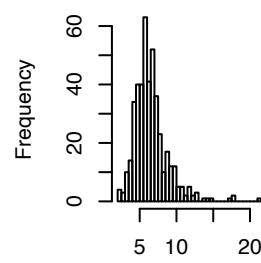
histogram of data\$pct.bach histogram of data\$pct.below.pov histogram of data\$pct.unemp histogram of data\$per.cap.income



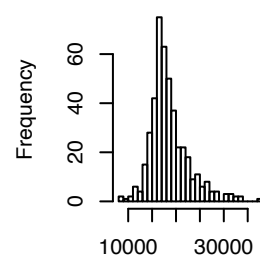
data\$pct.bach.deg



data\$pct.below.pov

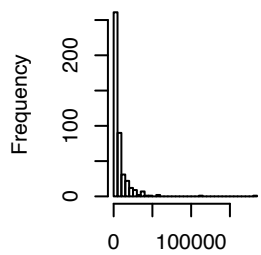


data\$pct.unemp



data\$per.cap.income

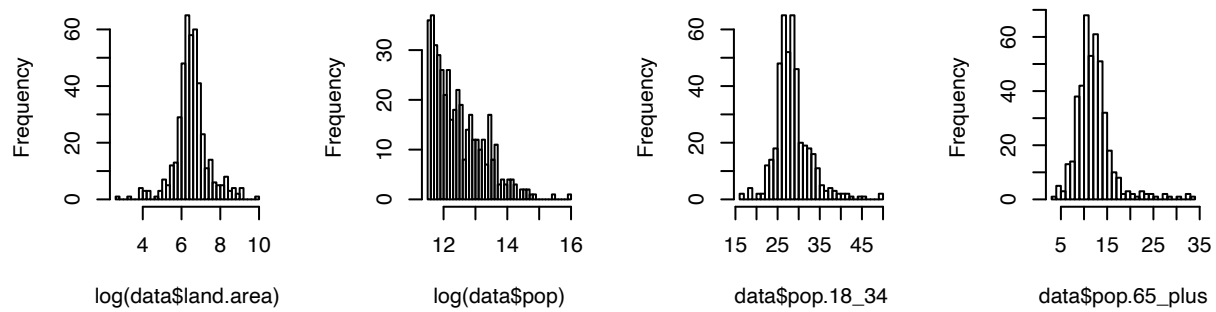
histogram of data\$tot.inc



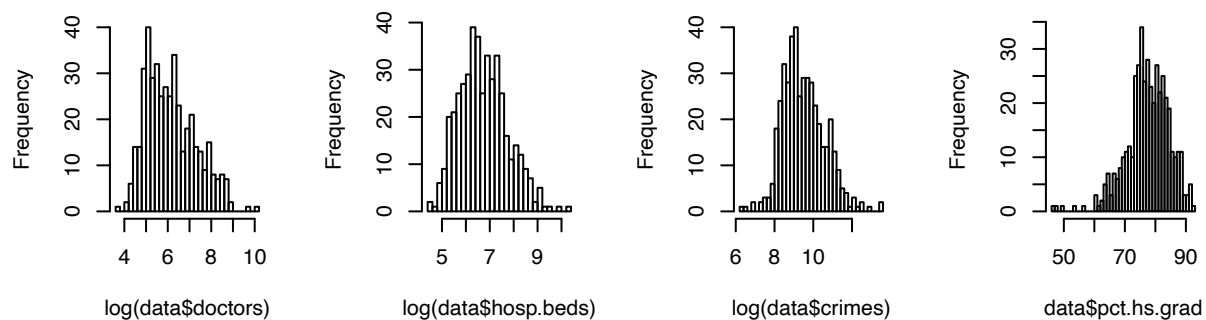
data\$tot.income

```
par(mfrow = c(2,4))
hist(log(data$land.area),breaks=30)
hist(log(data$pop),breaks=40)
hist(data$pop.18_34,breaks=40)
hist(data$pop.65_plus,breaks=40)
hist(log(data$doctors),breaks=40)
hist(log(data$hosp.beds),breaks=40)
hist(log(data$crimes),breaks=40)
hist(data$pct.hs.grad,breaks=40)
```

istogram of log(data\$land.area) Histogram of log(data\$pop) Histogram of data\$pop.18_34 histogram of data\$pop.65



istogram of log(data\$doctors) Histogram of log(data\$hosp.beds) Histogram of log(data\$crimes) Histogram of data\$pct.hs.

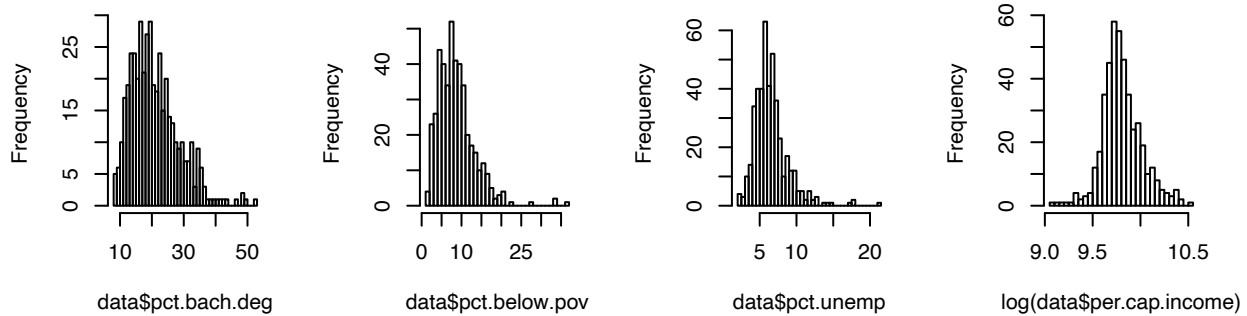


```
hist(data$pct.bach.deg,breaks=40)
hist(data$pct.below.pov,breaks=40)
hist(data$pct.unemp,breaks=40)
hist(log(data$per.cap.income),breaks=40)
hist(log(data$tot.income),breaks=40)
data$log.land.area=log(data$land.area)
data$log.pop=log(data$pop)
data$log.doctors=log(data$doctors)
data$log.hosp.beds=log(data$hosp.beds)
data$log.crimes=log(data$crimes)
data$log.per.cap.income=log(data$per.cap.income)
data$log.tot.income=log(data$tot.income)
data$log.per.cap.crimes=log(data$crimes/data$pop)
data = data[,-c(4,5,8,9,10,15,16)]
head(data)
```

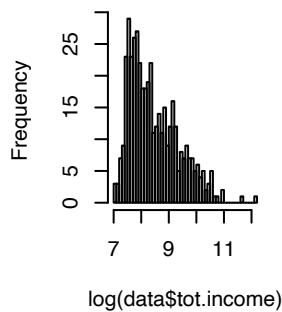
```
##   id      county state pop.18_34 pop.65_plus pct.hs.grad pct.bach.deg
## 1  1 Los_Angeles  CA    32.1      9.7      70.0      22.3
## 2  2      Cook    IL    29.2     12.4      73.4      22.8
## 3  3      Harris  TX    31.3      7.1      74.9      25.4
## 4  4 San_Diego   CA    33.5     10.9      81.9      25.3
## 5  5      Orange  CA    32.6      9.2      81.2      27.8
## 6  6      Kings  NY    28.3     12.4      63.7      16.6
##   pct.below.pov pct.unemp region log.land.area log.pop log.doctors
## 1          11.6      8.0     W      8.308938 15.99741 10.072259
## 2          11.1      7.2    NC      6.852243 15.44574  9.625954
## 3          12.5      5.7     S      7.455298 14.85161  8.929700
## 4           8.1      6.1     W      8.344030 14.73101  8.683555
## 5           5.2      4.8     W      6.672033 14.69537  8.709795
```

```
## 6      19.5      9.5      NE      4.262680 14.64871      8.488999
## log.hosp.beds log.crimes log.per.cap.income log.tot.income log.per.cap.crimes
## 1      10.229188 13.44290      9.942035      12.12394      -2.554511
## 2      9.978131 12.98754      9.986403      11.61664      -2.458202
## 3      9.429396 12.44322      9.879041      10.91514      -2.408387
## 4      8.728912 12.06578      9.882672      10.79817      -2.665226
## 5      8.759198 11.88120      10.102338      10.98220      -2.814167
## 6      9.098515 13.43127      9.729313      10.56251      -1.217441
```

histogram of data\$pct.bach histogram of data\$pct.below.pov histogram of data\$pct.unemp histogram of log(data\$per.cap.income)



histogram of log(data\$tot.income)

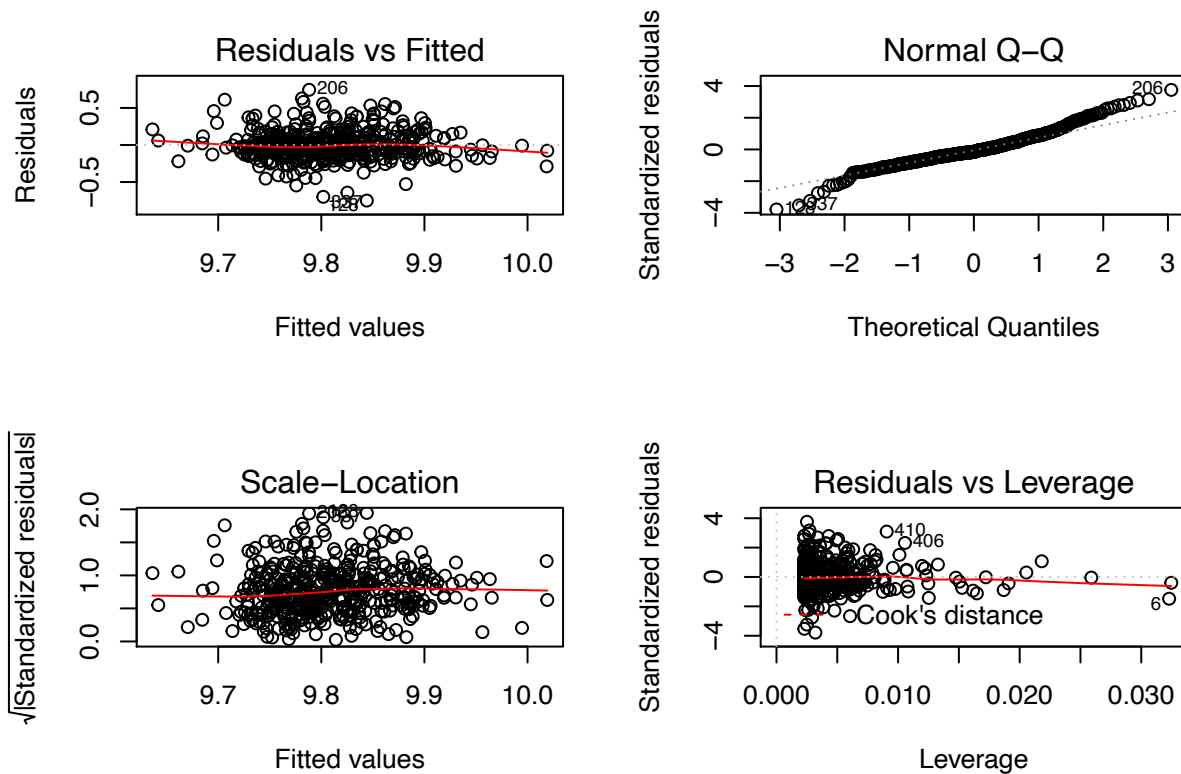


```
par(mfrow=c(2,2))
model1 = lm(log.per.cap.income ~ log.crimes, data)
summary(model1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.crimes, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.75042 -0.11569 -0.02976  0.09597  0.74498
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.295146   0.083764  110.97  < 2e-16 ***
## log.crimes    0.053858   0.008758   6.15 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1986 on 438 degrees of freedom
```

```
## Multiple R-squared:  0.07948,    Adjusted R-squared:  0.07738
## F-statistic: 37.82 on 1 and 438 DF,  p-value: 1.752e-09
```

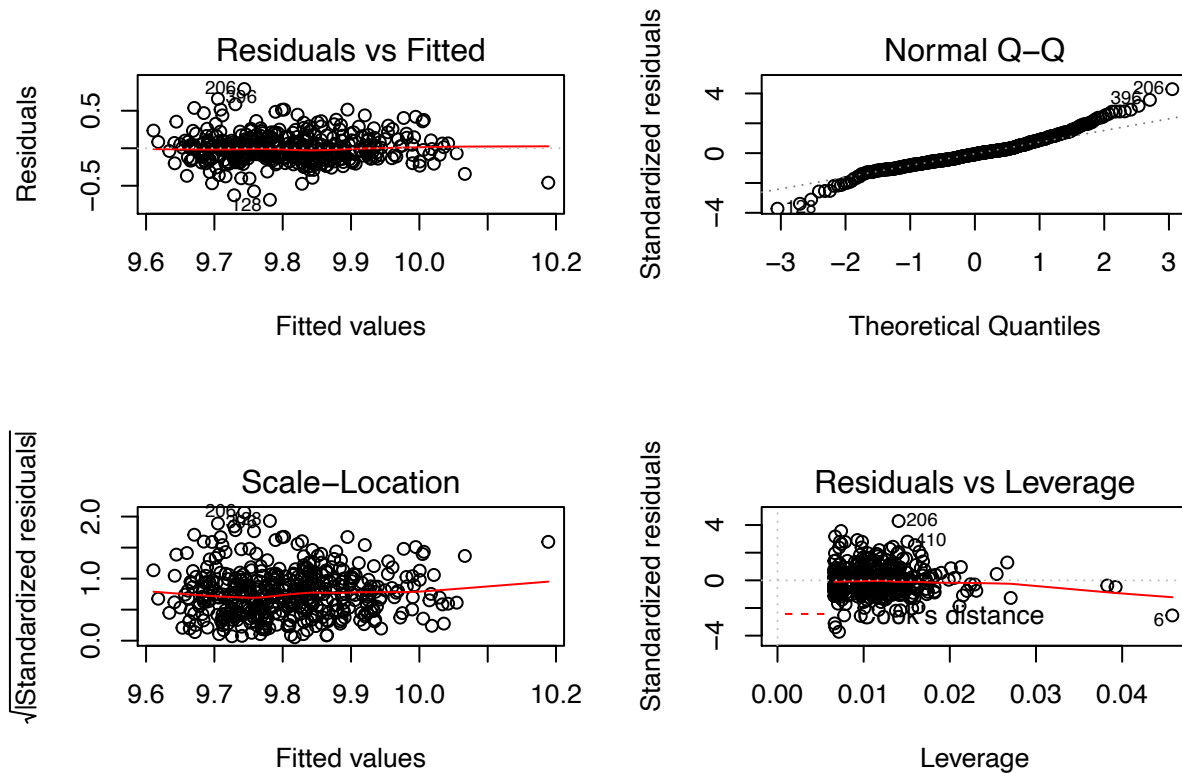
```
plot(model1)
```



```
par(mfrow=c(2,2))
model2 = lm(log.per.cap.income~ region+ log.crimes,data)
summary(model2)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ region + log.crimes, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68757 -0.10557 -0.01422  0.08905  0.78946
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.188431   0.079812 115.125 < 2e-16 ***
## regionNE     0.104458   0.025531   4.091 5.11e-05 ***
## regionS     -0.086983   0.023618  -3.683 0.00026 ***
## regionW     -0.055280   0.028167  -1.963 0.05033 .
## log.crimes   0.066695   0.008421   7.920 2.00e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1854 on 435 degrees of freedom
## Multiple R-squared:  0.2032, Adjusted R-squared:  0.1959
## F-statistic: 27.74 on 4 and 435 DF,  p-value: < 2.2e-16
```

```
plot(model2)
```

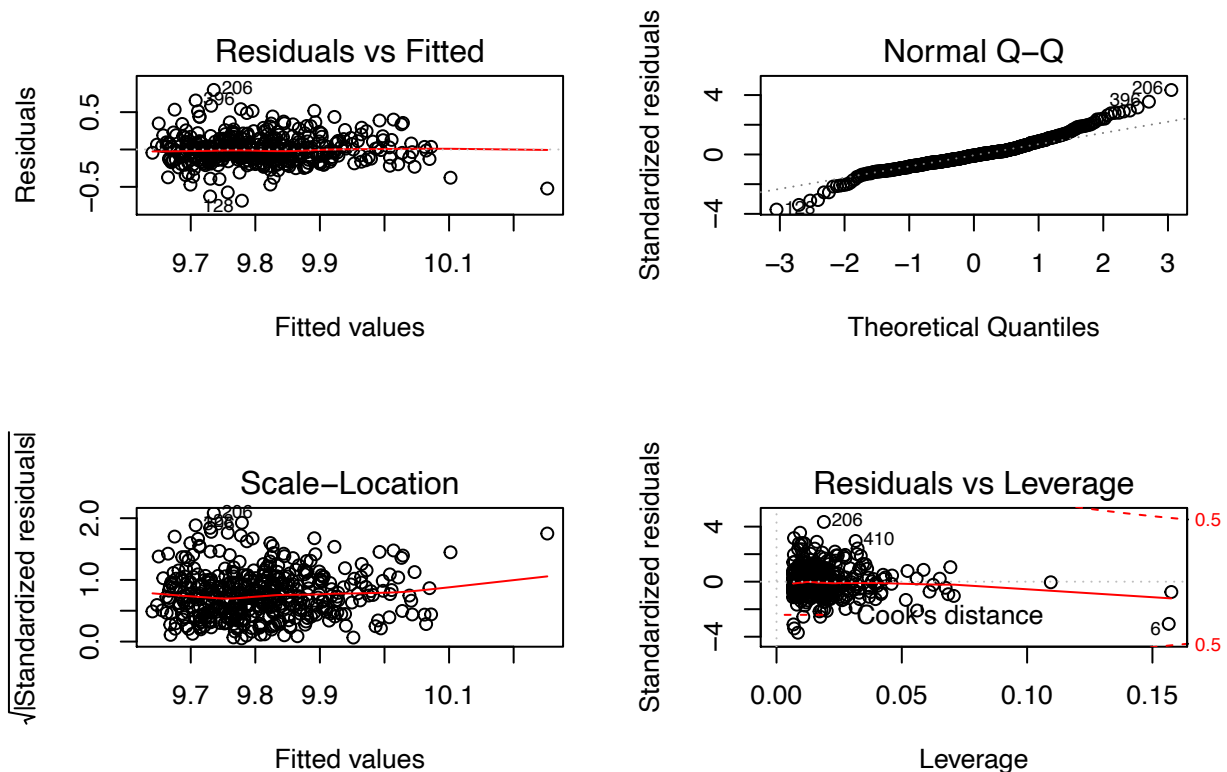


```
par(mfrow=c(2,2))
model3 = lm(log.per.cap.income~ region* log.crimes,data)
summary(model3)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ region * log.crimes, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68552 -0.10418 -0.01444  0.08302  0.79755
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.33677    0.14579   64.044 < 2e-16 ***
## regionNE       -0.18407    0.21515   -0.856  0.39272
## regionS        -0.19717    0.21211   -0.930  0.35312
## regionW        -0.31439    0.24465   -1.285  0.19947
## log.crimes      0.05064    0.01566    3.233  0.00132 **
## regionNE:log.crimes 0.03122    0.02311    1.351  0.17749
## regionS:log.crimes 0.01211    0.02228    0.544  0.58696
## regionW:log.crimes 0.02727    0.02523    1.081  0.28028
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1855 on 432 degrees of freedom
## Multiple R-squared:  0.2073, Adjusted R-squared:  0.1945
```

```
## F-statistic: 16.14 on 7 and 432 DF, p-value: < 2.2e-16
```

```
plot(model3)
```



```
anova(model1,model2,model3)
```

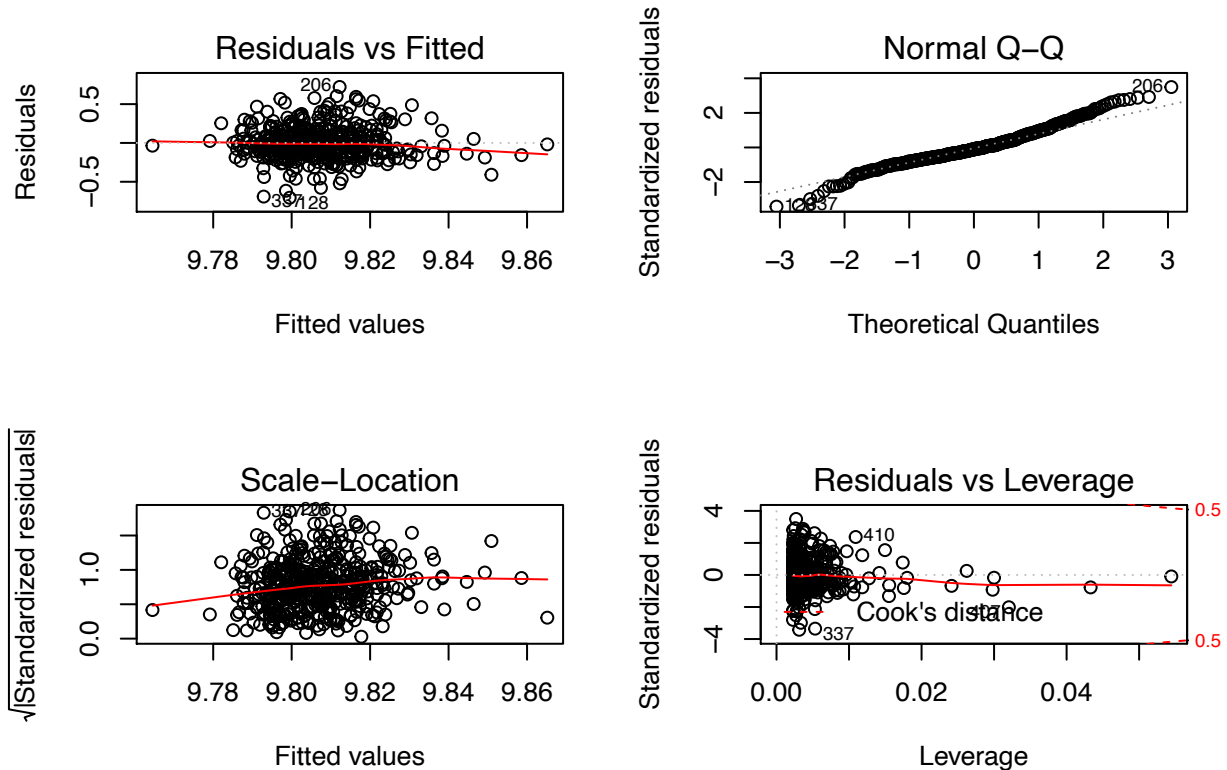
```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.crimes
## Model 2: log.per.cap.income ~ region + log.crimes
## Model 3: log.per.cap.income ~ region * log.crimes
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1     438 17.271
## 2     435 14.949   3    2.32194 22.4823 1.523e-13 ***
## 3     432 14.872   3    0.07678  0.7434  0.5266
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow=c(2,2))
model4 = lm(log.per.cap.income ~ log.per.cap.crimes,data)
summary(model4)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ log.per.cap.crimes, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7058 -0.1242 -0.0221  0.1066  0.7210
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.73510    0.05908 164.765  <2e-16 ***
## log.per.cap.crimes -0.02417    0.01959  -1.233    0.218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2066 on 438 degrees of freedom
## Multiple R-squared:  0.003461, Adjusted R-squared:  0.001186
## F-statistic: 1.521 on 1 and 438 DF, p-value: 0.2181
```

```
plot(model4)
```

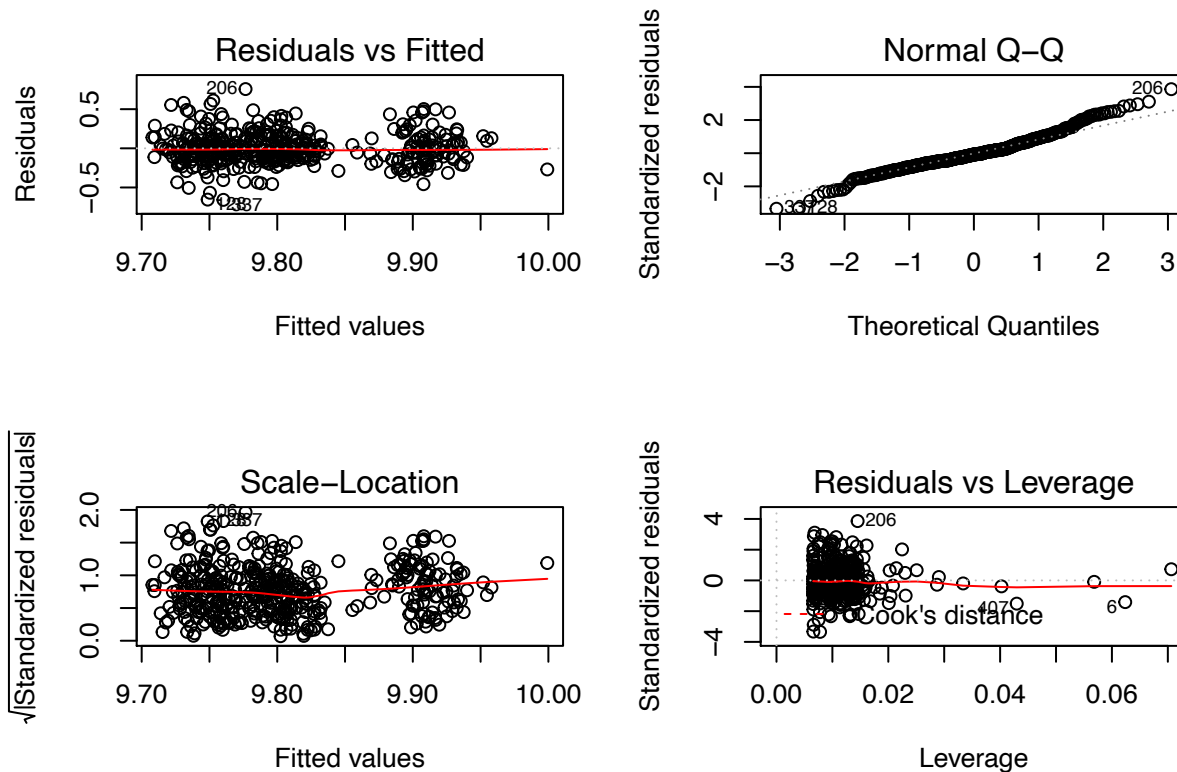


```
par(mfrow=c(2,2))
model5 = lm(log.per.cap.income ~ region + log.per.cap.crimes, data)
summary(model5)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ region + log.per.cap.crimes,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65832 -0.11431 -0.01548  0.10838  0.75657
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      9.93628    0.06934 143.303  < 2e-16 ***
## regionNE          0.11457    0.02760   4.151 3.99e-05 ***
## regionS         -0.07456    0.02624  -2.841  0.00471 **
```

```
## regionW          -0.02426    0.03002   -0.808   0.41952
## log.per.cap.crimes 0.04243    0.02148    1.975   0.04885 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1974 on 435 degrees of freedom
## Multiple R-squared:  0.09645,    Adjusted R-squared:  0.08814
## F-statistic: 11.61 on 4 and 435 DF,  p-value: 5.776e-09
```

```
plot(model5)
```

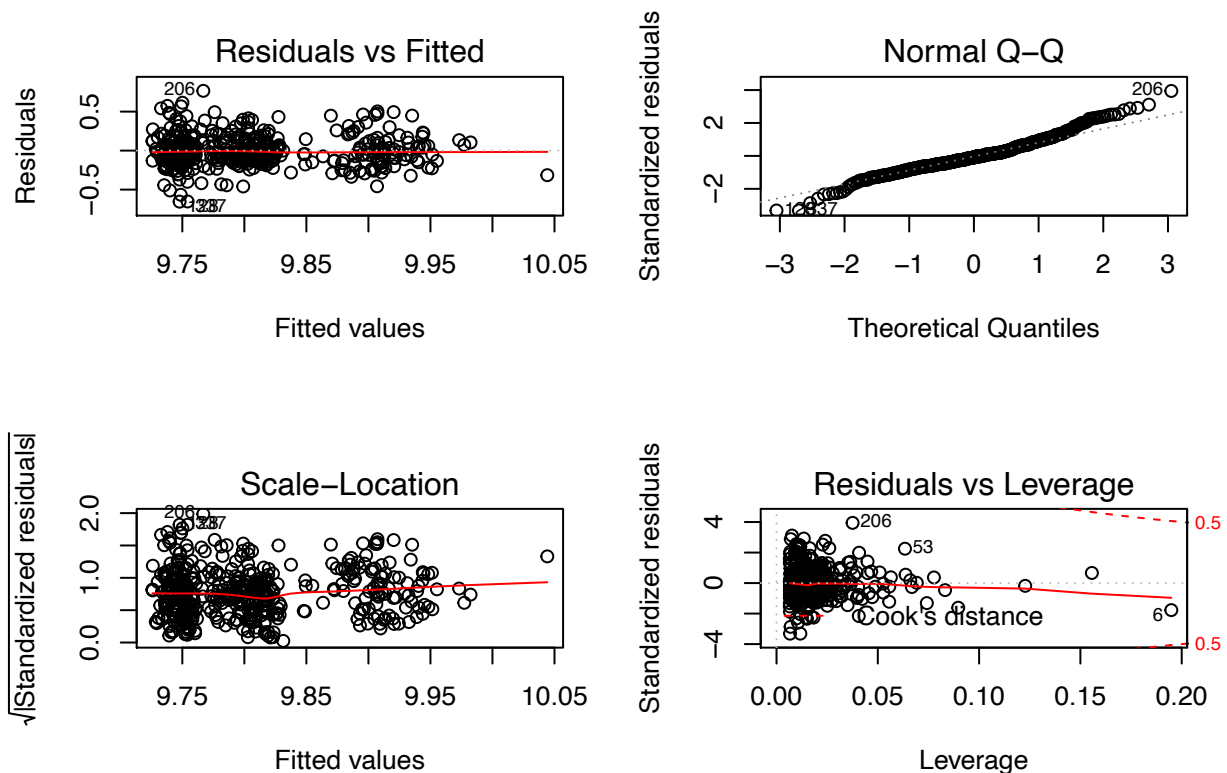


```
par(mfrow=c(2,2))
model6 = lm(log.per.cap.income~ region* log.per.cap.crimes,data)
summary(model6)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ region * log.per.cap.crimes,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.65410 -0.11829 -0.01708  0.10399  0.76628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.91177    0.10503   94.367  <2e-16 ***
## regionNE        0.21007    0.17165    1.224   0.222
## regionS       -0.10137    0.16072   -0.631   0.529
## regionW         0.07689    0.26753    0.287   0.774
```

```
## log.per.cap.crimes      0.03454    0.03327    1.038    0.300
## regionNE:log.per.cap.crimes 0.02924    0.05232    0.559    0.577
## regionS:log.per.cap.crimes -0.01104    0.05554   -0.199    0.843
## regionW:log.per.cap.crimes  0.03495    0.09268    0.377    0.706
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.198 on 432 degrees of freedom
## Multiple R-squared:  0.09773,    Adjusted R-squared:  0.08311
## F-statistic: 6.685 on 7 and 432 DF,  p-value: 1.575e-07
```

```
plot(model6)
```



```
anova(model4,model5,model6)
```

```
## Analysis of Variance Table
##
## Model 1: log.per.cap.income ~ log.per.cap.crimes
## Model 2: log.per.cap.income ~ region + log.per.cap.crimes
## Model 3: log.per.cap.income ~ region * log.per.cap.crimes
##   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
## 1      438 18.697
## 2      435 16.952   3    1.74465 14.8407 3.263e-09 ***
## 3      432 16.928   3    0.02408  0.2048    0.893
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

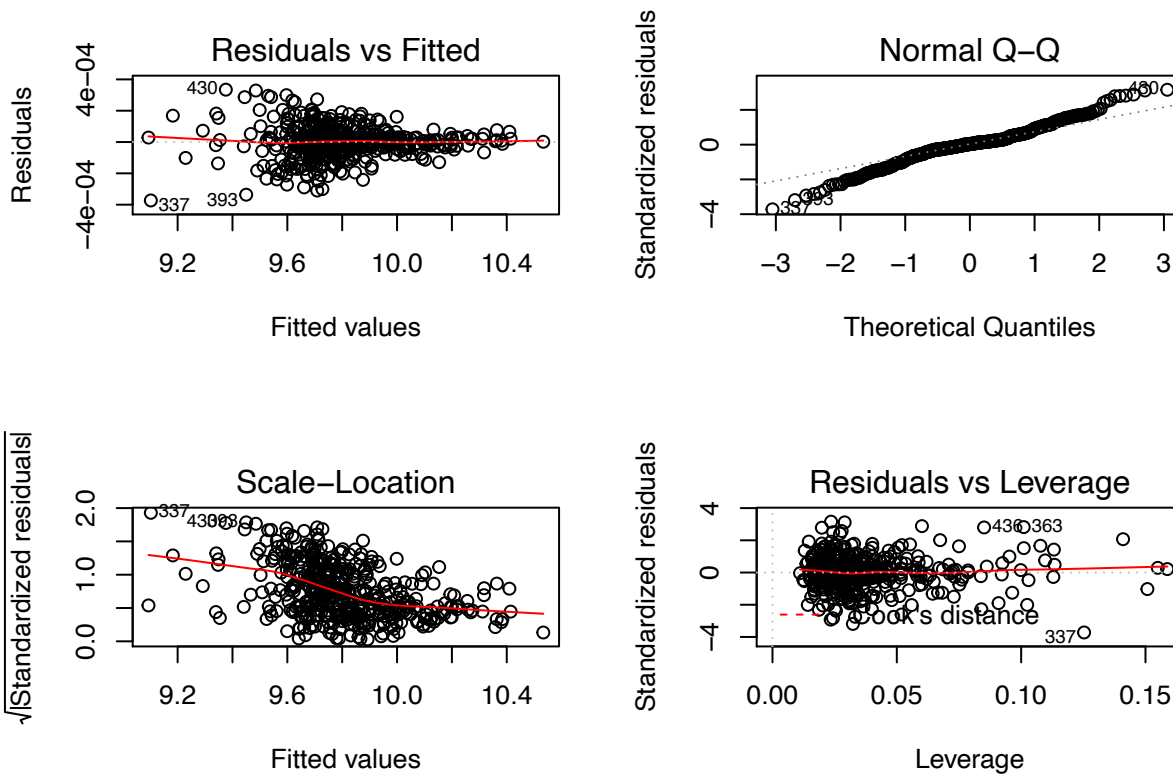
```
AIC(model12,model15)
```

```
##           df      AIC
## model12  6 -227.4746
```

```
## model5 6 -172.1347
model_full=lm(log.per.cap.income~.-county-state-log.per.cap.crimes-id,data=data)
summary(model_full)

##
## Call:
## lm(formula = log.per.cap.income ~ . - county - state - log.per.cap.crimes -
##      id, data = data)
##
## Residuals:
##      Min        1Q      Median        3Q       Max
## -3.730e-04 -4.586e-05  3.740e-06  5.420e-05  3.343e-04
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)   1.382e+01  3.010e-04  45893.977  <2e-16 ***
## pop.18_34      4.306e-07  2.037e-06    0.211    0.833
## pop.65_plus    2.530e-07  1.921e-06    0.132    0.895
## pct.hs.grad   -1.526e-07  1.592e-06   -0.096    0.924
## pct.bach.deg    7.351e-07  1.743e-06    0.422    0.673
## pct.below.pov  -6.753e-07  2.603e-06   -0.259    0.795
## pct.unemp       3.180e-06  3.241e-06    0.981    0.327
## regionNE      -5.810e-06  1.682e-05   -0.345    0.730
## regionS       -2.812e-06  1.649e-05   -0.171    0.865
## regionW        7.912e-06  2.048e-05    0.386    0.699
## log.land.area  -6.637e-06  7.623e-06   -0.871    0.384
## log.pop       -1.000e+00  6.536e-05 -15299.141  <2e-16 ***
## log.doctors   -2.265e-05  1.931e-05   -1.173    0.242
## log.hosp.beds  9.098e-06  1.724e-05    0.528    0.598
## log.crimes     1.187e-05  1.510e-05    0.786    0.432
## log.tot.income 1.000e+00  6.464e-05  15470.149  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0001072 on 424 degrees of freedom
## Multiple R-squared:  1, Adjusted R-squared:  1
## F-statistic: 1.089e+08 on 15 and 424 DF, p-value: < 2.2e-16

par(mfrow = c(2,2))
plot(model_full)
```



```
vif(model_full)
```

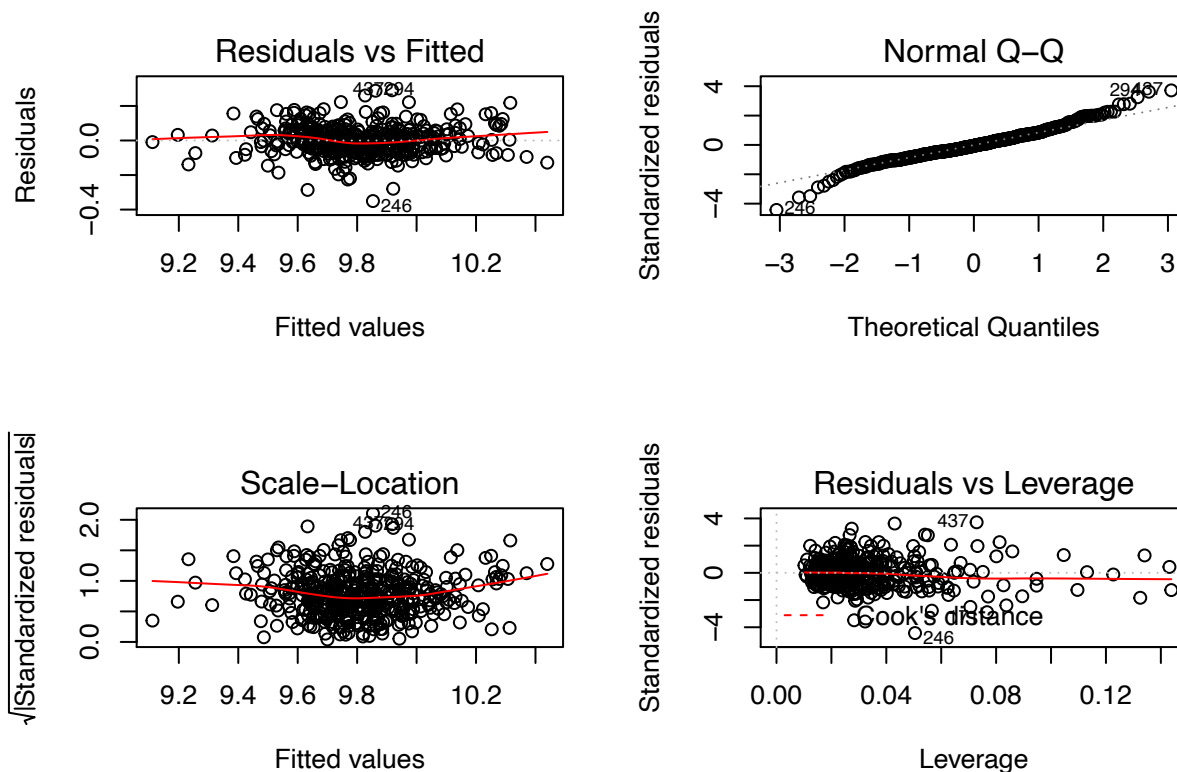
```
##              GVIF Df GVIF^(1/(2*Df))
## pop.18_34      2.786384  1      1.669246
## pop.65_plus    2.247053  1      1.499017
## pct.hs.grad    4.765290  1      2.182955
## pct.bach.deg   6.798932  1      2.607476
## pct.below.pov  5.614543  1      2.369503
## pct.unemp      2.194333  1      1.481328
## region         3.857097  3      1.252305
## log.land.area  1.687575  1      1.299067
## log.pop        102.006746  1     10.099839
## log.doctors    18.660929  1      4.319830
## log.hosp.beds  11.436936  1      3.381854
## log.crimes     10.197704  1      3.193384
## log.tot.income 128.688195  1     11.344082
```

```
model_red1=lm(log.per.cap.income~.-county-state-
              log.per.cap.crimes-id-log.tot.income-log.pop,data=data)
summary(model_red1)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ . - county - state - log.per.cap.crimes -
##     id - log.tot.income - log.pop, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35060 -0.04671 -0.00502  0.04525  0.29066
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.353232   0.116270  89.045 < 2e-16 ***
## pop.18_34    -0.015478   0.001307 -11.840 < 2e-16 ***
## pop.65_plus  -0.002648   0.001392  -1.903 0.057700 .
## pct.hs.grad  -0.005551   0.001172  -4.737 2.96e-06 ***
## pct.bach.deg   0.016348   0.001057  15.468 < 2e-16 ***
## pct.below.pov -0.024063   0.001413 -17.033 < 2e-16 ***
## pct.unemp      0.008846   0.002380   3.717 0.000229 ***
## regionNE      -0.003192   0.012683  -0.252 0.801414
## regionS       -0.031864   0.012271  -2.597 0.009740 **
## regionW       -0.014019   0.015422  -0.909 0.363849
## log.land.area -0.035003   0.005420  -6.459 2.89e-10 ***
## log.doctors    0.047877   0.013146   3.642 0.000304 ***
## log.hosp.beds  0.008624   0.013050   0.661 0.509042
## log.crimes     0.005828   0.008949   0.651 0.515255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08124 on 426 degrees of freedom
## Multiple R-squared:  0.8502, Adjusted R-squared:  0.8456
## F-statistic: 185.9 on 13 and 426 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model_red1)
```



```
vif(model_red1)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## pop.18_34    1.997041  1      1.413167
```

```
## pop.65_plus      2.053608  1      1.433041
## pct.hs.grad      4.495620  1      2.120288
## pct.bach.deg      4.353907  1      2.086602
## pct.below.pov     2.878962  1      1.696750
## pct.unemp         2.059659  1      1.435151
## region            3.580885  3      1.236892
## log.land.area     1.484666  1      1.218469
## log.doctors       15.046783  1      3.879018
## log.hosp.beds     11.403602  1      3.376922
## log.crimes        6.238869  1      2.497773
```

```
stepAIC(model_red1)
```

```
## Start:  AIC=-2195.37
## log.per.cap.income ~ (id + county + state + pop.18_34 + pop.65_plus +
##   pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##   region + log.land.area + log.pop + log.doctors + log.hosp.beds +
##   log.crimes + log.tot.income + log.per.cap.crimes) - county -
##   state - log.per.cap.crimes - id - log.tot.income - log.pop
##
##           Df Sum of Sq  RSS    AIC
## - log.crimes      1    0.00280 2.8142 -2196.9
## - log.hosp.beds    1    0.00288 2.8142 -2196.9
## <none>                        2.8114 -2195.4
## - pop.65_plus      1    0.02390 2.8353 -2193.6
## - region            3    0.05045 2.8618 -2193.5
## - log.doctors       1    0.08753 2.8989 -2183.9
## - pct.unemp         1    0.09117 2.9025 -2183.3
## - pct.hs.grad       1    0.14808 2.9594 -2174.8
## - log.land.area     1    0.27528 3.0866 -2156.3
## - pop.18_34         1    0.92508 3.7364 -2072.2
## - pct.bach.deg       1    1.57891 4.3903 -2001.2
## - pct.below.pov     1    1.91471 4.7261 -1968.8
##
## Step:  AIC=-2196.93
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##   pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
##   log.doctors + log.hosp.beds
##
##           Df Sum of Sq  RSS    AIC
## - log.hosp.beds    1    0.00463 2.8188 -2198.2
## <none>                        2.8142 -2196.9
## - region            3    0.04944 2.8636 -2195.3
## - pop.65_plus       1    0.03036 2.8445 -2194.2
## - pct.unemp         1    0.10475 2.9189 -2182.8
## - log.doctors       1    0.12320 2.9374 -2180.1
## - pct.hs.grad       1    0.14788 2.9620 -2176.4
## - log.land.area     1    0.27525 3.0894 -2157.9
## - pop.18_34         1    0.92508 3.7392 -2073.9
## - pct.bach.deg       1    1.59250 4.4067 -2001.6
## - pct.below.pov     1    1.95727 4.7714 -1966.6
##
## Step:  AIC=-2198.21
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##   pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
```

```

##      log.doctors
##
##              Df Sum of Sq    RSS      AIC
## <none>                2.8188 -2198.2
## - pop.65_plus      1   0.02680 2.8456 -2196.0
## - region            3   0.05597 2.8748 -2195.6
## - pct.unemp         1   0.10177 2.9206 -2184.6
## - pct.hs.grad       1   0.14595 2.9647 -2178.0
## - log.land.area     1   0.27116 3.0900 -2159.8
## - pop.18_34         1   0.92048 3.7393 -2075.9
## - log.doctors       1   1.40559 4.2244 -2022.2
## - pct.bach.deg      1   1.77875 4.5975 -1985.0
## - pct.below.pov     1   2.04855 4.8673 -1959.9
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##      pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
##      log.doctors, data = data)
##
## Coefficients:
##      (Intercept)      pop.18_34      pop.65_plus      pct.hs.grad      pct.bach.deg
##      10.385117      -0.015394      -0.002650      -0.005506      0.015921
##      pct.below.pov      pct.unemp      regionNE      regionS      regionW
##      -0.023860      0.009048      -0.006109      -0.031170      -0.016272
##      log.land.area      log.doctors
##      -0.034613      0.060845
model_red2 = step(model_red1,direction="backward",k=log(440))

## Start:  AIC=-2138.15
## log.per.cap.income ~ (id + county + state + pop.18_34 + pop.65_plus +
##      pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##      region + log.land.area + log.pop + log.doctors + log.hosp.beds +
##      log.crimes + log.tot.income + log.per.cap.crimes) - county -
##      state - log.per.cap.crimes - id - log.tot.income - log.pop
##
##              Df Sum of Sq    RSS      AIC
## - region            3   0.05045 2.8618 -2148.6
## - log.crimes        1   0.00280 2.8142 -2143.8
## - log.hosp.beds     1   0.00288 2.8142 -2143.8
## - pop.65_plus       1   0.02390 2.8353 -2140.5
## <none>                2.8114 -2138.2
## - log.doctors       1   0.08753 2.8989 -2130.8
## - pct.unemp         1   0.09117 2.9025 -2130.2
## - pct.hs.grad       1   0.14808 2.9594 -2121.7
## - log.land.area     1   0.27528 3.0866 -2103.1
## - pop.18_34         1   0.92508 3.7364 -2019.1
## - pct.bach.deg      1   1.57891 4.3903 -1948.1
## - pct.below.pov     1   1.91471 4.7261 -1915.7
##
## Step:  AIC=-2148.59
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##      pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##      log.doctors + log.hosp.beds + log.crimes

```

```

##
##           Df Sum of Sq   RSS   AIC
## - log.crimes      1    0.00180 2.8636 -2154.4
## - log.hosp.beds    1    0.01216 2.8740 -2152.8
## - pop.65_plus      1    0.03884 2.9006 -2148.7
## <none>                                2.8618 -2148.6
## - log.doctors      1    0.11565 2.9775 -2137.2
## - pct.hs.grad      1    0.12699 2.9888 -2135.6
## - pct.unemp        1    0.17289 3.0347 -2128.9
## - log.land.area    1    0.36392 3.2257 -2102.0
## - pop.18_34        1    0.94423 3.8060 -2029.2
## - pct.bach.deg     1    1.56251 4.4243 -1963.0
## - pct.below.pov    1    2.44318 5.3050 -1883.1
##
## Step:   AIC=-2154.4
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##      pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##      log.doctors + log.hosp.beds
##
##           Df Sum of Sq   RSS   AIC
## - log.hosp.beds    1    0.01116 2.8748 -2158.8
## - pop.65_plus      1    0.03709 2.9007 -2154.8
## <none>                                2.8636 -2154.4
## - pct.hs.grad      1    0.12662 2.9902 -2141.4
## - log.doctors      1    0.12889 2.9925 -2141.1
## - pct.unemp        1    0.17123 3.0348 -2134.9
## - log.land.area    1    0.37492 3.2385 -2106.3
## - pop.18_34        1    0.94270 3.8063 -2035.3
## - pct.bach.deg     1    1.59514 4.4587 -1965.7
## - pct.below.pov    1    2.47345 5.3371 -1886.5
##
## Step:   AIC=-2158.77
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##      pct.bach.deg + pct.below.pov + pct.unemp + log.land.area +
##      log.doctors
##
##           Df Sum of Sq   RSS   AIC
## - pop.65_plus      1    0.03031 2.9051 -2160.2
## <none>                                2.8748 -2158.8
## - pct.hs.grad      1    0.12309 2.9978 -2146.4
## - pct.unemp        1    0.16432 3.0391 -2140.4
## - log.land.area    1    0.38995 3.2647 -2108.9
## - pop.18_34        1    0.93157 3.8063 -2041.3
## - log.doctors      1    1.55295 4.4277 -1974.8
## - pct.bach.deg     1    1.80755 4.6823 -1950.2
## - pct.below.pov    1    2.53302 5.4078 -1886.8
##
## Step:   AIC=-2160.25
## log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##      pct.below.pov + pct.unemp + log.land.area + log.doctors
##
##           Df Sum of Sq   RSS   AIC
## <none>                                2.9051 -2160.2
## - pct.hs.grad      1    0.11147 3.0165 -2149.8

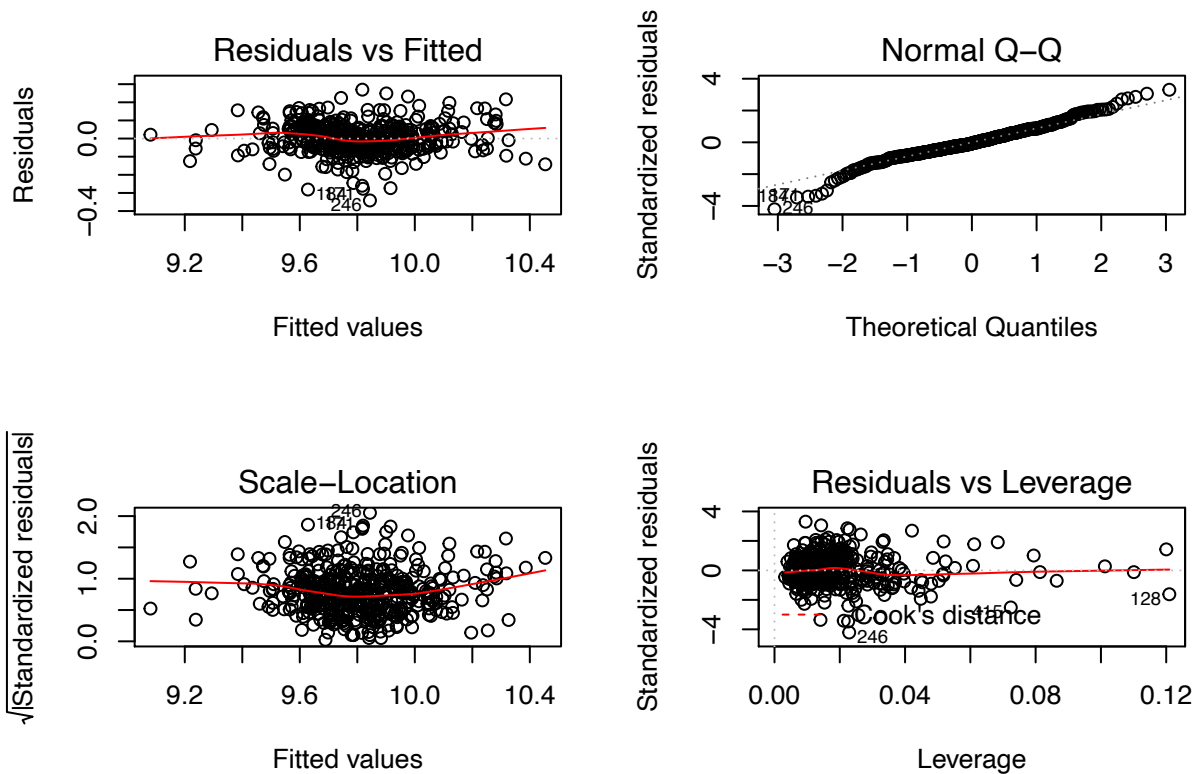
```

```
## - pct.unemp      1    0.15952 3.0646 -2142.8
## - log.land.area  1    0.37507 3.2801 -2112.9
## - pop.18_34      1    1.05209 3.9572 -2030.3
## - log.doctors    1    1.53330 4.4384 -1979.8
## - pct.bach.deg   1    1.86219 4.7673 -1948.4
## - pct.below.pov  1    2.50333 5.4084 -1892.9

summary(model_red2)

##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pct.hs.grad + pct.bach.deg +
##      pct.below.pov + pct.unemp + log.land.area + log.doctors,
##      data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.34147 -0.04886 -0.00538  0.04818  0.26969
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2224950  0.0931210 109.776 < 2e-16 ***
## pop.18_34    -0.0139002  0.0011113 -12.508 < 2e-16 ***
## pct.hs.grad  -0.0044064  0.0010823  -4.071 5.56e-05 ***
## pct.bach.deg   0.0153853  0.0009246  16.641 < 2e-16 ***
## pct.below.pov -0.0242784  0.0012583 -19.294 < 2e-16 ***
## pct.unemp      0.0106037  0.0021771   4.871 1.56e-06 ***
## log.land.area -0.0356741  0.0047767  -7.468 4.53e-13 ***
## log.doctors    0.0606769  0.0040183  15.100 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.082 on 432 degrees of freedom
## Multiple R-squared:  0.8452, Adjusted R-squared:  0.8427
## F-statistic: 336.9 on 7 and 432 DF,  p-value: < 2.2e-16

par(mfrow = c(2,2))
plot(model_red2)
```



```
model_red3 = stepAIC(model_red1)
```

```
## Start: AIC=-2195.37
## log.per.cap.income ~ (id + county + state + pop.18_34 + pop.65_plus +
##   pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
##   region + log.land.area + log.pop + log.doctors + log.hosp.beds +
##   log.crimes + log.tot.income + log.per.cap.crimes) - county -
##   state - log.per.cap.crimes - id - log.tot.income - log.pop
##
##           Df Sum of Sq  RSS   AIC
## - log.crimes    1    0.00280 2.8142 -2196.9
## - log.hosp.beds  1    0.00288 2.8142 -2196.9
## <none>                 2.8114 -2195.4
## - pop.65_plus    1    0.02390 2.8353 -2193.6
## - region         3    0.05045 2.8618 -2193.5
## - log.doctors    1    0.08753 2.8989 -2183.9
## - pct.unemp      1    0.09117 2.9025 -2183.3
## - pct.hs.grad    1    0.14808 2.9594 -2174.8
## - log.land.area  1    0.27528 3.0866 -2156.3
## - pop.18_34      1    0.92508 3.7364 -2072.2
## - pct.bach.deg   1    1.57891 4.3903 -2001.2
## - pct.below.pov  1    1.91471 4.7261 -1968.8
##
## Step: AIC=-2196.93
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
##   pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
##   log.doctors + log.hosp.beds
##
##           Df Sum of Sq  RSS   AIC
```

```
## - log.hosp.beds 1 0.00463 2.8188 -2198.2
## <none> 2.8142 -2196.9
## - region 3 0.04944 2.8636 -2195.3
## - pop.65_plus 1 0.03036 2.8445 -2194.2
## - pct.unemp 1 0.10475 2.9189 -2182.8
## - log.doctors 1 0.12320 2.9374 -2180.1
## - pct.hs.grad 1 0.14788 2.9620 -2176.4
## - log.land.area 1 0.27525 3.0894 -2157.9
## - pop.18_34 1 0.92508 3.7392 -2073.9
## - pct.bach.deg 1 1.59250 4.4067 -2001.6
## - pct.below.pov 1 1.95727 4.7714 -1966.6
##
## Step: AIC=-2198.21
## log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
## pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
## log.doctors
##
## Df Sum of Sq RSS AIC
## <none> 2.8188 -2198.2
## - pop.65_plus 1 0.02680 2.8456 -2196.0
## - region 3 0.05597 2.8748 -2195.6
## - pct.unemp 1 0.10177 2.9206 -2184.6
## - pct.hs.grad 1 0.14595 2.9647 -2178.0
## - log.land.area 1 0.27116 3.0900 -2159.8
## - pop.18_34 1 0.92048 3.7393 -2075.9
## - log.doctors 1 1.40559 4.2244 -2022.2
## - pct.bach.deg 1 1.77875 4.5975 -1985.0
## - pct.below.pov 1 2.04855 4.8673 -1959.9
```

```
summary(model_red3)
```

```
##
## Call:
## lm(formula = log.per.cap.income ~ pop.18_34 + pop.65_plus + pct.hs.grad +
## pct.bach.deg + pct.below.pov + pct.unemp + region + log.land.area +
## log.doctors, data = data)
##
## Residuals:
## Min 1Q Median 3Q Max
## -0.34849 -0.04695 -0.00502 0.04524 0.28624
##
## Coefficients:
## Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.3851173 0.1105475 93.943 < 2e-16 ***
## pop.18_34 -0.0153941 0.0013021 -11.822 < 2e-16 ***
## pop.65_plus -0.0026499 0.0013137 -2.017 0.04430 *
## pct.hs.grad -0.0055059 0.0011696 -4.707 3.39e-06 ***
## pct.bach.deg 0.0159212 0.0009688 16.434 < 2e-16 ***
## pct.below.pov -0.0238604 0.0013529 -17.637 < 2e-16 ***
## pct.unemp 0.0090479 0.0023017 3.931 9.86e-05 ***
## regionNE -0.0061091 0.0123398 -0.495 0.62080
## regionS -0.0311704 0.0114050 -2.733 0.00654 **
## regionW -0.0162724 0.0140361 -1.159 0.24697
## log.land.area -0.0346133 0.0053943 -6.417 3.70e-10 ***
## log.doctors 0.0608452 0.0041649 14.609 < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08115 on 428 degrees of freedom
## Multiple R-squared:  0.8498, Adjusted R-squared:  0.8459
## F-statistic: 220.1 on 11 and 428 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(model_red3)
```

