Kevin Yang
Department of Statistics & Data Science, Carnegie Mellon University
kevinyan@andrew.cmu.edu

Predicting Per Capita Income by County

Abstract

This study aims to answer several questions from social scientists related to the per capita income of counties in the United States. The dataset used here comes from Kuter et al. (2005) in *Applied Linear Statistical Models, Fifth Edition* and contains various stats from the 440 largest counties in the United States. Through the use of transformations, correlation plots, and various variable selection methods, a model was created to predict per capita income using seven of the thirteen variables present in the dataset, as well as finding collinearity among the variables, and how crimes and crime rate affect per capita income separately. All in all, the model is good but because of the small size of the dataset and because the dataset excludes all of the smaller counties in the United States, further research is needed.

Introduction

Every county in the United States faces different economic, health, and social well-being situations based on many different factors such as their geography, their demographics, and their infrastructure. These factors combined help determine if a country is “good” or “bad” to the general public and can further influence the county’s appeal if people want to visit or settle there. For this study, a county’s average income per capita will be the variable used to determine a county’s overall quality of life, the higher the better. The given dataset for this study will be used to answer four questions: first to see if any variables are related to each other, if the per capita income is related to crime rate in different regions of the United States, what’s the best combination of variables that can be used to predict the average income per capita for any given county, and if missing counties or states from the dataset makes a difference in the study.

Data

The data for this study comes from Kuter et al. (2005) from *Applied Linear Statistical Models, Fifth Edition*. The dataset contains information from the 440 most populous counties in the United States in 1990, each with an id, name, state, region, 12 other continuous variables, and the county’s average income per capita. Below is a table detailing each variable:

Variable	Description (All in 1990)
id	A given id for each county
county	County name
state	State the county is in

land.area	County land area
pop	County population
pop.18_34	Percent of county population between 18-34
pop.65_plus	Percent of county population above 65
doctors	Number of doctors in county
hosp.beds	Number of hospital beds in county
crimes	Number of serious crimes in county
pct.hs.grad	Percent of county population that completed high school
pct.bach.deg	Percent of county population that got a bachelor's degree
pct.below.pov	Percent of county population below poverty line
pct.unemp	Percent of county population unemployed
per.cap.income	County's income per capita (Response variable)
tot.income	County's total income
region	Region of the United States the county resides (NC, NE, W, S)

Preliminary statistics:

Continuous Variables:

	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

Categorical Variables:

	NC	NE	S	W
Freq	108	103	152	77

Methods

To start, histograms for every continuous variable were created to see the normality of each variable's distributions. In order for the analysis to run smoothly, each variable should be as normal as possible. As such, any variable with a noticeable skew was log transformed to pull outlier points closer to the rest of the data

For the first question regarding variables relating to other variables in the dataset, a correlation plot was created to visualize and quickly identify any highly correlated variables. These pairs were noted down as they have a high chance of being removed when the regression model was being created later on.

For the second question regarding crime and per capita income by region, multiple linear models were created. A new variable called crime rate was created using crimes and dividing it by the population amount. Six linear models were created, three using total crimes and three using crime rate. Within each of those groups, the three models consisted of crimes/crime rate on it's own, crimes/crime rate and region with no interaction variable, and crimes/crime rate with an interaction variable. Afterwards, summary tables, AIC values, and residual plots were made to compare the models to see if any model was particularly better than the rest.

To make the full regression model, three methods were used: VIF, all subsets, stepwise regression, and LASSO. The variables used in these methods are the transformed variables created at the start of the study. Since per capita income is a continuous variable, most of the categorical variables and the id column were removed from the model, this includes the county name, and the county state. Region is the only categorical variable being considered because it only has four levels, each with a fair amount of data for each level. Each variables' variance inflation factor (VIF) was calculated, and any variable with a VIF greater than 10 was removed from the model. Next, the three variable selection methods were completed, first without interaction variables, then with region interactions afterwards. The variables that were chosen from all three methods were compared to one another to see if a definitive model can be made. Then a final summary table and residual plots will be created to check if all of the linear model assumptions are satisfied.

To answer the fourth question, no analysis was conducted and discussion points were made in the discussion section of this paper.

Results

Starting with the distributions of some of the variables, the histograms created in table A revealed several of the variables being right skewed by outliers. As such, those variables were log transformed, specifically: crimes, doctors, hosp.beds, land.area, pop, tot.income, and per.cap.income (Table B).

With the correlation plot in table C, there are apparent strong correlations between pop, crimes, hosp.beds, doctors, total income, moderately strong correlations between per.cap.income, pct.bach.deg, and pct.hs.grad, and a strong negative correlation between pct.below.pov and pct.hs.grad.

Between the six linear models made for the second question in tables D,E,F,G,H,and I, the models that contained total crimes tended to have more significant terms than with crime rate, and the model containing region terms with no interaction variables had the most significant terms with the highest R-squared value. Looking at the residual plots in table J for this particular model shows that it is random enough with no high influence points, but is a bit heavy tailed.

After calculating the VIF for each of the variables against per.cap.income, pop and tot.income were removed from the model since they had VIF greater than 100 (Table K). With the subsets methods, seven variables were selected with no region terms: land.area, pop.18_34, doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp (Table L). The linear model with these variables all resulted in significant terms and the residual plots were mostly ok. Adding in interaction terms with region resulted in some of the terms being significant meaning that region will likely be kept in the model (Table M).

For the stepAIC method, eight variables were selected in the best model (Table N). The variables were the same as the ones chosen in the subsets method except that pop.65_plus was added in. Including the region interaction terms also resulted in a model similar to the subsets method (Table O).

Lastly, with the LASSO method, six variables were selected in the best model (Table P). The variables were also similar to the ones from the subsets method except that pct.hs.grad was removed from the model. Similar to the stepAIC method, adding the region interaction terms should result in a model similar to the subsets method.

Discussion

The goal of this study was to answer four questions posed by the social scientists: Whether any of the variables in the 1990 county dataset are correlated with one another, If per capita income for a particular county can be better predicted using the total crimes, or the crime rate of the county, What the best combination of variables is to calculate a county's per capita income, and If the counties missing from the dataset made a difference in how the final model was structured.

From the correlation plot, it's clear that several of the variables were highly correlated with each other. The correlations typically came in groups of three and were all positively correlated with each other, one group which had population, total income, and per capita income, and the other group containing doctors, hospital beds, and crimes. The first group was correlated because per capita income was calculated dividing total income by population, and the other group was correlated because doctors and hospital beds are both related to the hospital environment and the serious crimes used in the dataset often send victims to the hospital as well. On the opposite end, *pct.below.pov*, and *pct.hs.grad* had a strong negative correlation with each other, likely meaning that someone who graduates high school has a lower chance of being in poverty in the future, which could be a study all on its own.

In comparing *total crimes* to *crime rate* to predict per capita income of a county, total crimes proved to be the better option. It doesn't seem to be an intuitive answer though since every county has a different population and usually proportions are used when that kind of variability exists. However, the best model for this also contains the *region* variable, meaning that the region of the United States the county resides in likely has a larger impact on per capita income, and total crimes is merely a good supplement to it. This is something that can be further studied as there are only 440 counties in the dataset used for this study and there are over 3000 counties in the United States.

When making the best model to predict per capita income, all three methods selected nearly the same variables to be in the final model. The only differences lie in the subsets method adding in pct.hs.grad, and the lasso method adding in pct.hs.grad and pop.65_plus in their models. Since the models were so similar to each other, the best model should be chosen based on the meaning of the variables such as the social, economic, and health factors and its implication. In that case the best model is likely the model chosen by the subsets method with land.area, pop.18_34, doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, and region interaction variables. From a social standpoint, each of these variables have a defendable reason as to why they belong in the model: land.area can measure population density which can affect per capita income, pop.18_34 is the age range where most people are earning income in their lives, the number of doctors can indicate the quality of care someone can get in the county which could mean higher incomes, pct.hs.grad as mentioned earlier is negatively correlated with pct.below.pov so the higher the percentage, the higher the income, pct.bach.deg is similar to pct.hs.grad, pct.unemp will reduce per capita income the higher it is. Pop.65_plus from the LASSO method wasn't included in this model since people older than 65 are usually retired and aren't working.

As for the question about missing counties in the dataset, it should be a bit worrying that they weren't considered in the model because the 440 counties used in this study are the 440 largest counties in the United States. These counties are likely not representative of the smaller counties with smaller populations, different age distributions, fewer medical resources, and fewer educational resources and instead might actually be outliers when compared to the 2500+ other counties not included in the dataset. This is definitely something that should be further researched, first by seeing if the subsets model from above can predict per capita income for a particular county, then by refitting the model to see how smaller counties influence the selected variables and their coefficients.

References

Kutner, M. H., Nachsheim, C. J., Neter, J., and Li, W. (2005), *Applied Linear Statistical Models* (Fifth ed.), NY: McGraw-Hill Irwin.

Technical Appendix

Kevin Yang

10/17/2021

Table A: Untransformed Data

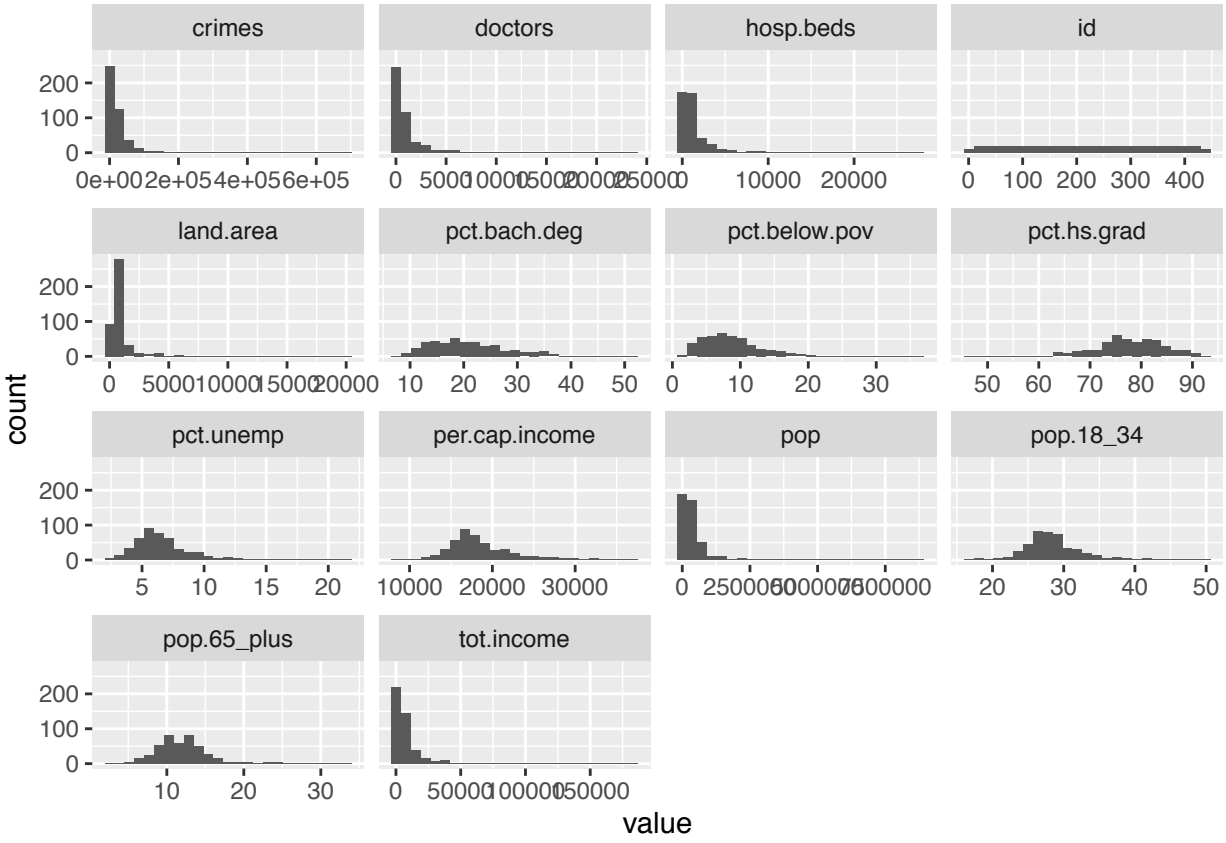

Table B: Transformed Data

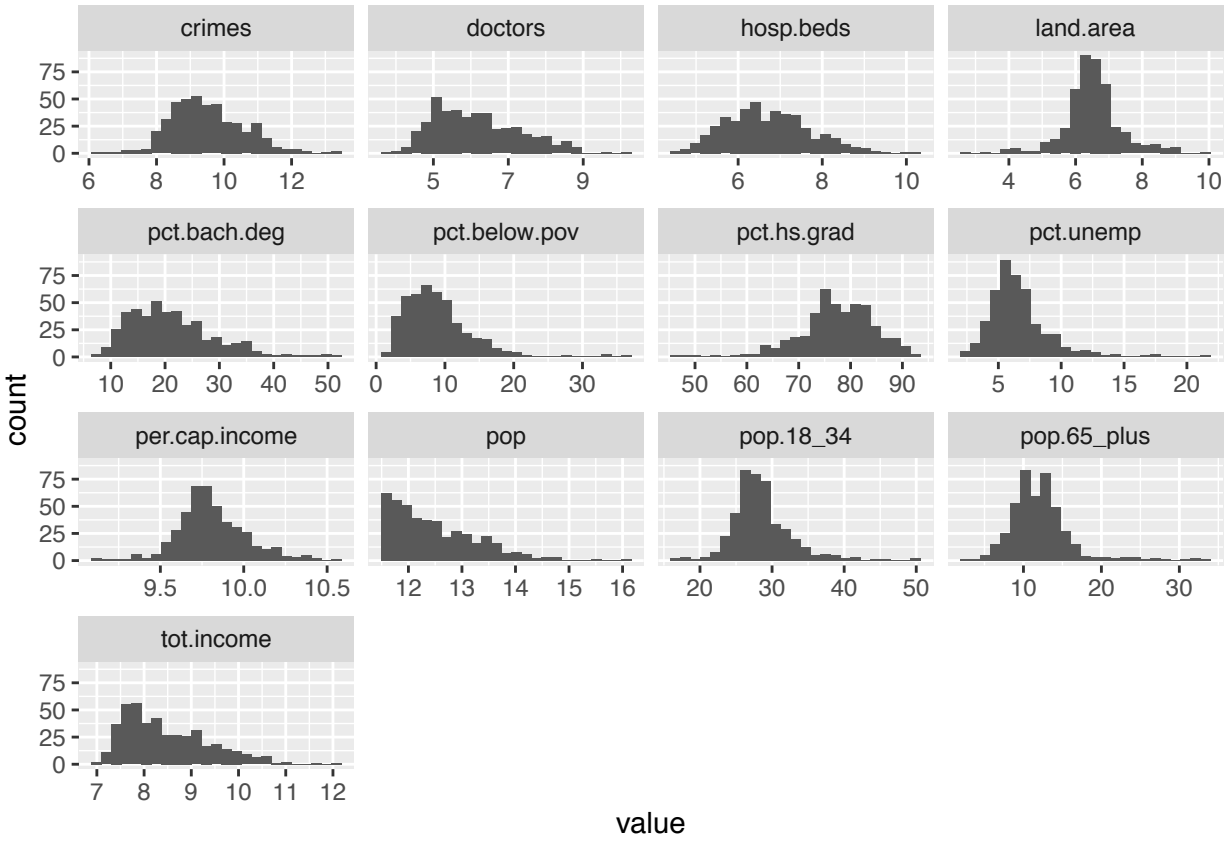

Table C: Correlation Plot

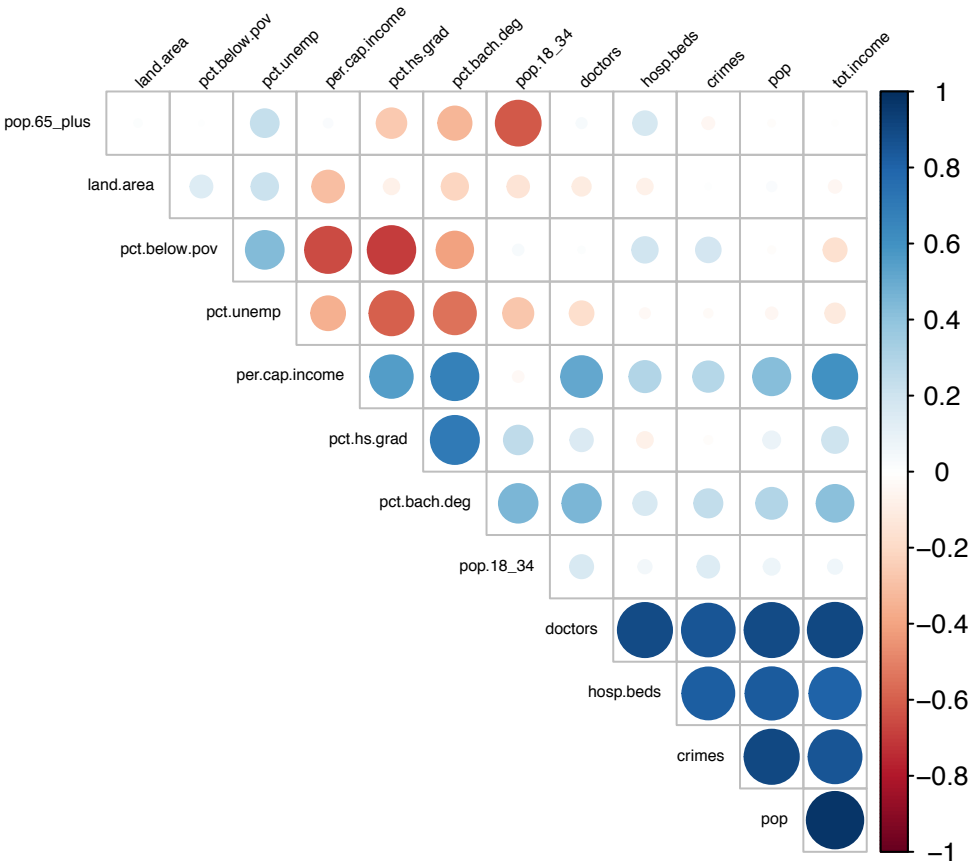

Table D: Total crimes with region interaction variables

```
#>
#> Call:
#> lm(formula = per.cap.income ~ crimes * region, data = x3)
#>
#> 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 ***
#> crimes           0.05064    0.01566   3.233  0.00132 **
#> 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
#> crimes:regionNE  0.03122    0.02311   1.351  0.17749
#> crimes:regionS   0.01211    0.02228   0.544  0.58696
#> crimes:regionW   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
```

Table E: Total crimes with region, no interaction variables

```
#>
#> Call:
#> lm(formula = per.cap.income ~ crimes + region, data = x3)
#>
#> 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 ***
#> crimes       0.066695   0.008421   7.920 2.00e-14 ***
#> 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 .
#> ---
#> 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
```

Table F: Total crimes only

```
#>
#> Call:
#> lm(formula = per.cap.income ~ crimes, data = x3)
#>
#> 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 ***
#> 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
```

Table G: Crime rate and region with interaction variables

```
#>
#> Call:
```

```
#> lm(formula = per.cap.income ~ crimerate * region, data = x3)
#>
#> 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 ***
#> crimerate           0.03454    0.03327   1.038    0.300
#> 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
#> crimerate:regionNE  0.02924    0.05232   0.559    0.577
#> crimerate:regionS  -0.01104    0.05554  -0.199    0.843
#> crimerate:regionW   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
```

Table H: Crime rate and region, no interaction variables

```
#>
#> Call:
#> lm(formula = per.cap.income ~ crimerate + region, data = x3)
#>
#> 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 ***
#> crimerate    0.04243    0.02148   1.975  0.04885 *
#> 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
#> ---
#> 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
```

Table I: Crime rate only

```
#>
#> Call:
#> lm(formula = per.cap.income ~ crimerate, data = x3)
#>
```

```
#> 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 ***
#> crimerate   -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
```

Table J: Residual plots for best model (From table E)

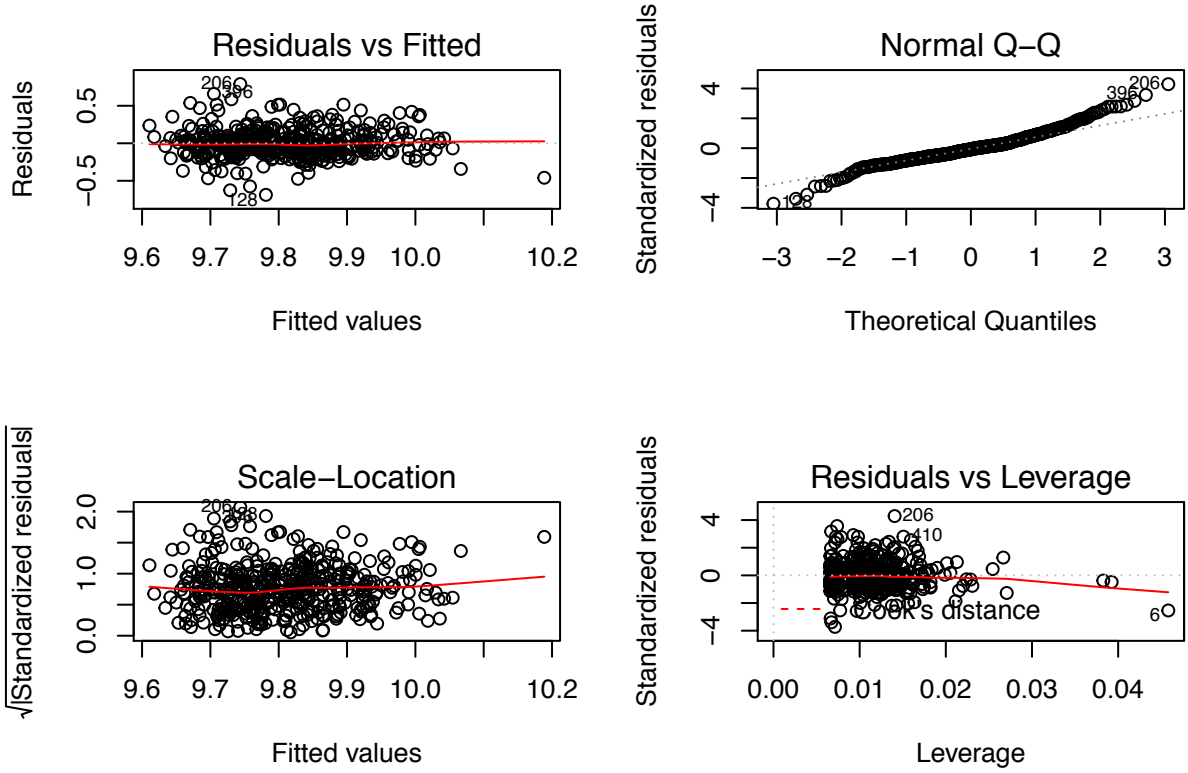

Table K: VIF values

```
#>     land.area           pop     pop.18_34   pop.65_plus       doctors
#>      1.348568    101.081007      2.723926      2.187009     17.278105
#>     hosp.beds        crimes   pct.hs.grad  pct.bach.deg pct.below.pov
#>      9.713256      7.433688      4.014452      6.288770      5.440728
#>     pct.unemp    tot.income
#>      1.957833    125.495194
```

Table L: Subsets method no interaction variables

```
#>  [1] -257.5260 -502.4302 -572.5538 -682.8532 -732.1894 -761.5908 -772.0715
#>  [8] -770.5990 -766.2235 -760.4131

#>   (Intercept)     land.area     pop.18_34       doctors   pct.hs.grad
#>  10.222495041  -0.035674062  -0.013900201   0.060676872  -0.004406396
#>  pct.bach.deg pct.below.pov     pct.unemp
#>   0.015385301  -0.024278371   0.010603691

#>
#> Call:
#> lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
#>     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp, data = x4)
#>
#> 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 ***
#> land.area     -0.0356741  0.0047767  -7.468 4.53e-13 ***
#> pop.18_34     -0.0139002  0.0011113 -12.508  < 2e-16 ***
#> doctors        0.0606769  0.0040183  15.100  < 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 ***
#> ---
#> 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
```

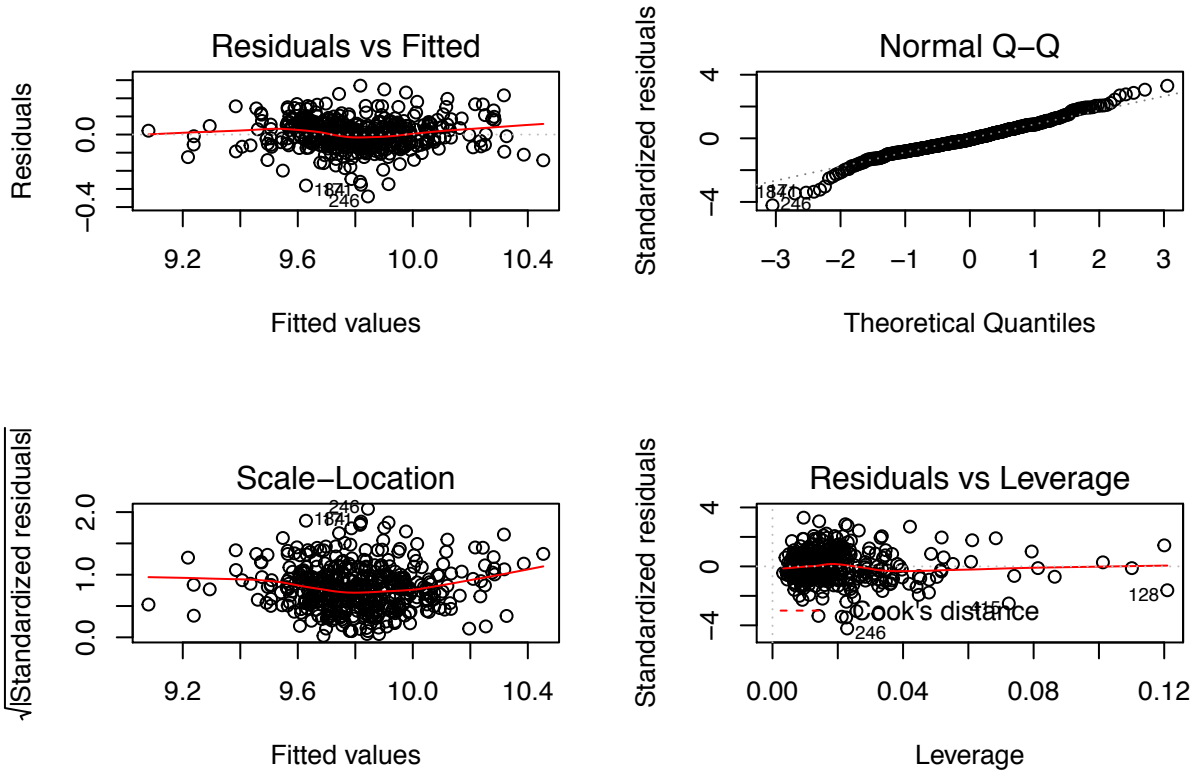

Table M: Region interaction variables added onto subsets model in Table L

```
#>
#> Call:
#> lm(formula = per.cap.income ~ (land.area + pop.18_34 + doctors +
#>     pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp) *
#>     region, data = x5)
#>
#> Residuals:
#>       Min        1Q    Median        3Q       Max
#> -0.250782 -0.042332 -0.002298  0.040559  0.313570
#>
#> Coefficients:
#>                          Estimate Std. Error t value Pr(>|t|)
#> (Intercept)            10.1244260  0.2826240  35.823  < 2e-16 ***
#> land.area              -0.0364187  0.0151355  -2.406 0.016564 *
#> pop.18_34              -0.0147940  0.0026043  -5.681 2.55e-08 ***
#> doctors                 0.0544169  0.0093221   5.837 1.08e-08 ***
#> pct.hs.grad            -0.0024773  0.0034110  -0.726 0.468088
#> pct.bach.deg            0.0140833  0.0029254   4.814 2.09e-06 ***
#> pct.below.pov          -0.0237085  0.0036234  -6.543 1.81e-10 ***
#> pct.unemp               0.0180393  0.0048923   3.687 0.000257 ***
#> regionNE                0.3243992  0.3577081   0.907 0.365004
#> regionS                -0.0345856  0.3131668  -0.110 0.912116
#> regionW                 1.5043946  0.4226868   3.559 0.000416 ***
#> land.area:regionNE     -0.0037179  0.0201435  -0.185 0.853656
#> land.area:regionS      -0.0047582  0.0174155  -0.273 0.784825
```

```
#> land.area:regionW       0.0151234  0.0181871   0.832 0.406154
#> pop.18_34:regionNE     -0.0024780  0.0036873  -0.672 0.501939
#> pop.18_34:regionS      -0.0008777  0.0030680  -0.286 0.774970
#> pop.18_34:regionW       0.0014122  0.0040925   0.345 0.730220
#> doctors:regionNE       -0.0046251  0.0132571  -0.349 0.727359
#> doctors:regionS         0.0043337  0.0114401   0.379 0.705019
#> doctors:regionW        -0.0034863  0.0131576  -0.265 0.791173
#> pct.hs.grad:regionNE   -0.0037529  0.0044150  -0.850 0.395813
#> pct.hs.grad:regionS     0.0021198  0.0037853   0.560 0.575790
#> pct.hs.grad:regionW    -0.0190188  0.0045881  -4.145 4.13e-05 ***
#> pct.bach.deg:regionNE   0.0069429  0.0040312   1.722 0.085776 .
#> pct.bach.deg:regionS   -0.0015774  0.0032000  -0.493 0.622328
#> pct.bach.deg:regionW    0.0071026  0.0036374   1.953 0.051541 .
#> pct.below.pov:regionNE -0.0014134  0.0050896  -0.278 0.781381
#> pct.below.pov:regionS   0.0072764  0.0040739   1.786 0.074827 .
#> pct.below.pov:regionW  -0.0161639  0.0054271  -2.978 0.003071 **
#> pct.unemp:regionNE     -0.0083596  0.0073758  -1.133 0.257720
#> pct.unemp:regionS      -0.0249396  0.0065867  -3.786 0.000176 ***
#> pct.unemp:regionW      -0.0201466  0.0067713  -2.975 0.003101 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 0.0759 on 408 degrees of freedom
#> Multiple R-squared:  0.8747, Adjusted R-squared:  0.8652
#> F-statistic: 91.91 on 31 and 408 DF,  p-value: < 2.2e-16
```

Table N: StepAIC with no interaction variables

```
#>
#> Call:
#> lm(formula = per.cap.income ~ land.area + pop.18_34 + pop.65_plus +
#>     doctors + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp,
#>     data = x4)
#>
#> Coefficients:
#>   (Intercept)      land.area      pop.18_34    pop.65_plus        doctors
#>     10.315967      -0.036493      -0.015349      -0.002766       0.062605
#>   pct.hs.grad   pct.bach.deg  pct.below.pov      pct.unemp
#>     -0.004658       0.015215      -0.024614       0.010769
```

Table O: StepAIC with region interaction variables

```
#>
#> Call:
#> lm(formula = per.cap.income ~ land.area + pop.18_34 + doctors +
#>     crimes + pct.hs.grad + pct.bach.deg + pct.below.pov + pct.unemp +
#>     region + doctors:region + crimes:region + pct.hs.grad:region +
#>     pct.bach.deg:region + pct.below.pov:region + pct.unemp:region,
#>     data = x5)
#>
#> Coefficients:
#>            (Intercept)               land.area               pop.18_34
#>             10.1212121             -0.0324537             -0.0153759
```

```
#>                doctors                  crimes             pct.hs.grad
#>              0.0412157               0.0131113              -0.0031715
#>           pct.bach.deg           pct.below.pov              pct.unemp
#>              0.0149138              -0.0233414               0.0160990
#>               regionNE                 regionS                 regionW
#>              0.0005355              -0.0904471               1.8843762
#>       doctors:regionNE         doctors:regionS         doctors:regionW
#>             -0.0249320               0.0161981               0.0664384
#>        crimes:regionNE          crimes:regionS          crimes:regionW
#>              0.0287435              -0.0113999              -0.0704979
#>   pct.hs.grad:regionNE     pct.hs.grad:regionS     pct.hs.grad:regionW
#>             -0.0020914               0.0026168              -0.0184737
#>  pct.bach.deg:regionNE    pct.bach.deg:regionS    pct.bach.deg:regionW
#>              0.0057137              -0.0021509               0.0045162
#> pct.below.pov:regionNE   pct.below.pov:regionS   pct.below.pov:regionW
#>             -0.0034259               0.0066183              -0.0150228
#>     pct.unemp:regionNE       pct.unemp:regionS       pct.unemp:regionW
#>             -0.0070316              -0.0231696              -0.0174992
```

Table P: LASSO method

```
#>   lambda.1se   lambda.min
#> 0.0064883132 0.0005775994
```

```
#> 11 x 1 sparse Matrix of class "dgCMatrix"
#>                          1
#> (Intercept)    9.878369962
#> land.area     -0.032063002
#> pop.18_34     -0.011810866
#> pop.65_plus    .
#> doctors        0.059230219
#> hosp.beds      .
#> crimes         .
#> pct.hs.grad    .
#> pct.bach.deg   0.011645778
#> pct.below.pov -0.019928341
#> pct.unemp      0.005894554
```

Code Appendix

```
knitr::opts_chunk$set(comment = "#>", tidy.opts = list(width.cutoff = 70),
    tidy = TRUE)
set.seed(1645)
library(tidyverse)
library(car)
library(leaps)
library(MASS)
library(glmnet)
library(kableExtra)
setwd("~/Documents/College/Semester 9/Applied Linear Modeling/ALM HW6")
x <- read.table("cdi.dat")
cdinumeric <- x[, -c(1, 2, 3, 17)]  ## get rid of id, county, state and (for now) region
apply(cdinumeric, 2, function(x) c(summary(x), SD = sd(x))) %>%
```

```
    as.data.frame %>%
    t() %>%
    round(digits = 2) %>%
    kbl(booktabs = T, caption = " ") %>%
    kable_classic()

tmp <- rbind(with(x, table(region)))
row.names(tmp) <- "Freq"
knitr::kable(tmp)
ggplot(gather(x[, c(1, 4:16)]), aes(value)) + geom_histogram(bins = 25) +
    facet_wrap(~key, scales = "free_x")  # Reference (1) below
x2 <- x[4:16]
x2[, 7] <- log(x2[, 7])
x2[, 5] <- log(x2[, 5])
x2[, 6] <- log(x2[, 6])
x2[, 1] <- log(x2[, 1])
x2[, 2] <- log(x2[, 2])
x2[, 13] <- log(x2[, 13])
x2[, 12] <- log(x2[, 12])
ggplot(gather(x2), aes(value)) + geom_histogram(bins = 25) + facet_wrap(~key,
    scales = "free_x")
corx <- cor(x2, method = "pearson")
corrplot::corrplot(corx, type = "upper", order = "hclust", tl.col = "black",
    tl.srt = 45, diag = F, tl.cex = 0.5)
x3 <- x %>%
    mutate(crimerate = crimes/pop)
x3 <- x3[, 4:18]
x3[, 7] <- log(x3[, 7])
x3[, 5] <- log(x3[, 5])
x3[, 6] <- log(x3[, 6])
x3[, 1] <- log(x3[, 1])
x3[, 2] <- log(x3[, 2])
x3[, 13] <- log(x3[, 13])
x3[, 12] <- log(x3[, 12])
x3[, 15] <- log(x3[, 15])
y <- lm(per.cap.income ~ crimes * region, data = x3)
y2 <- lm(per.cap.income ~ crimes + region, data = x3)
y3 <- lm(per.cap.income ~ crimes, data = x3)
y4 <- lm(per.cap.income ~ crimerate * region, data = x3)
y5 <- lm(per.cap.income ~ crimerate + region, data = x3)
y6 <- lm(per.cap.income ~ crimerate, data = x3)
summary(y)
summary(y2)
summary(y3)
summary(y4)
summary(y5)
summary(y6)
par(mfrow = c(2, 2))
plot(y2)
all <- lm(per.cap.income ~ ., data = x2)
vif(all)
x4 <- x3[, -c(2, 13, 14, 15)]
superset <- regsubsets(per.cap.income ~ ., data = x4, nvmax = 11)
```

```
s <- summary(superset)
s$bic  # Best model at 7
coef(superset, 7)
summary(lm(per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp, data = x4))
par(mfrow = c(2, 2))
plot(lm(per.cap.income ~ land.area + pop.18_34 + doctors + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp, data = x4))
x5 <- x3[, -c(2, 13, 15)]
summary(lm(per.cap.income ~ (land.area + pop.18_34 + doctors + pct.hs.grad +
    pct.bach.deg + pct.below.pov + pct.unemp) * region, data = x5))
aic2 <- stepAIC(lm(per.cap.income ~ ., data = x4), direction = "both",
    k = 2, trace = 0)
aic2
aic3 <- stepAIC(lm(per.cap.income ~ . * region, data = x5), direction = "both",
    k = 2, trace = 0)
aic3
set <- cv.glmnet(as.matrix(x4[, -11]), as.matrix(x4[, 11]))
c(lambda.1se = set$lambda.1se, lambda.min = set$lambda.min)
coef(set, s = set$lambda.1se)
```