Title

Zach Ohl – zohl@andrew.cmu.edu

Abstract

The goal of this study is to examine how the average income per person in U.S. counties is related to average income other geographic, economic, and social variables. The data consist of observations for the 440 most populous counties in the country and 14 variables for each county describing information from the years 1990 and 1992. We approached this topic by searching for a linear regression model that predicts average income based on some combination of the 14 variables. The model that we found included the variables _____, as well as interactions between ____. The variables included in our model show that ___ are useful predictors of a county's average income per person, but caution must be taken before using this model to make predictions about the nearly 3000 other U.S. counties.

Introduction

Our goal is to discover how certain features of a county's economic, health and social well-being are related to the county's per capita income. The variables used in the study contain information on the county's geography, demographics, metrics of physical health, crime, education, and economic information. Using these variables, we have been given the following tasks:

1. List and describe any apparent relationships between variables. Explain the relationships in terms of the meanings of the variables, if possible.
2. Describe the predictive ability of a county's crime numbers and region of the country on its per-capita income.
3. Model per-capita income using a combination of the other variables. Choose a statistically valid model that reflects the meaning of the variables and can be interpreted by the social scientists who requested the study.
4. Describe the consequences of the missing states and counties from the data set.

Data

The data for this study comes from the textbook *Applied Linear Statistical Models*, by Kutner and others. their original source was the Geospatial and Statistical Data Center at the University of Virginia. The 1990-1992 dataset contains 440 observations, each representing a unique U.S. county. [An initial look at the counties might suggest that county is a categorical variable, since multiple observations have the same value of county. But these actually represent duplicates of the same county name in different states, so all 440 county observations are unique.] The values of 17 variables are included in the dataset but we will use 14 of them—13 quantitative and 1 categorical variable. Variable definitions are listed in Table 1.

Table 1: Variable definitions

	Variable	Definition
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)

Variables summaries are found in Table 1, Table 2, and Figure 1 of the technical appendix.

*NOTE: to rough draft reviewer, when I knit the markdown file, it puts some of the figures/tables towards the bottom of the document instead of outputting in the order of the code. I haven't figured out a fix yet.

Methods

1.

We will look at summaries of each variable, check for missing values, and examine the distributions of each variable alone and against other variables to look for patterns. The shapes of single-variable distributions will be used to suggest transformations of variables. The plots of paired variables will be used to look for high correlation between predictors and linear relationships among variables, especially between predictors and the response, per-capita income.

After identifying the relationships, we will examine the questions about whether the relationships are expected and can be explained using the meanings of the variables. The shapes of the distributions will also be used to suggest transformations for the models in research questions 2 and 3.

2.

To address this question, we will first find 3 models that predict per-capita income using region and crime (the raw crime variable). One model will use just crime, one will use crime and region, and one will use those two predictors plus their interactions. The crime variable will be log-transformed to deal with its right skew.

We'll then repeat the three models above, but with log(crimes) replaced by log(crimes)/population. We will compare the 6 total models using ANOVA tests and information criteria, while making sure modeling assumptions are met according to the residual diagnostics. If the models suggested by AIC and BIC measures disagree, we'll lean toward the BIC-suggested model because our goal is understanding the relationship between the variables, not making super accurate predictions.

Once a model is selected, we'll offer our best interpretation of the coefficients in order to answer the research question.

3.

Before fitting models, we will transform most, but not all, of the predictors using log, because of the skews shown in their distributions. We may consider a couple power transformations, especially on the one left-skewed variable, later in selection process once the variables have been narrowed down.

We will remove the variable for total income because it is too directly related to the response variable, per-capita income (total income/population). We will leave in population for now since it is only inversely related to the response variable. We'll also add a new variable, population density (population/land area). Based on prior knowledge of incomes in U.S. cities, I wouldn't expect population or land area to be a significant predictor of per capita income, but I could imagine this new variable being significant.

We'll then look for a model using all remaining variables except for state. We may look at state as a predictor later, but for now, a categorical variable that has 48 categories and is unlikely to be significant would only make the model selection messier. At first we'll only use variables on their own with no interactions. If any of the region indicator variables are selected, we'll follow the convention of keeping all four region indicators.

Then we will proceed to variable selection using all using the 'all subsets' method, starting with the 12 numerical variables and one categorical (region) variables. We will check the diagnostics and compare the subsets of variables suggested when using AIC vs using BIC as criteria. Again we will lean toward the BIC-suggested model unless there is a compelling reason not to. This is because the purpose of our model is understanding the predictors, not using them to make pinpoint-accurate predictions.

Then we'll repeat the above process, but with interactions added in. If the algorithm for all subsets method takes forever to complete with all the additional potential variables, we'll try a stepwise method instead. Similarly to before, if the interaction of a continuous variable and a region indicator variable is selected, we'll keep that continuous variable's interaction with all four region indicators.

If the winning models from both the interaction and non-interaction processes both meet modelling assumptions equally well, we'll choose between them by comparing adjusted r^2, AIC, and BIC.

Eventually we'll try a penalized regression method – with interactions and without – and compare the results with the results from above.

4.

To answer this question, we'll consider how the 440 counties in the dataset ended up there, what they have in common, and what the missing counties have in common. We'll use the limited information given, and if we have time, possibly look up more information on the included counties to check for any patterns.

Results

1.

The variables crimes, doctors, hosp.beds, land.area, pop, and total.inc are all heavily right skewed. Predictors pct.bach.deg, pct.below.pov, pct.unemp, pop.65plus all appear only slightly right-skewed, while pct.hs.grad has a slight left skew. The response, per.cap.income and the last numeric predictor, pop.18-34, both appear relatively symmetric.

Based on the scatter plots of the two-way distributions (Figure 2 in the technical appendix), there are apparent linear relations for pairs of variables you would expect, for instance, doctors vs hospital beds and population vs total income. Some interesting variables that have a relationships with per.cap.income include pop, doctors, pct.hs.grad, pct.bach.deg, pct.below.pov, pct.unemp, and tot.income. The relationship per.cap.income vs pc.bach.deg relationship looks very linear, while the pattern appears more curved for the plots of per.cap.income vs pct.hs.grad, pct.below.pov, and pc.unemp. The relationships between the response and the pop, doctors, and tot.income predictors just look vaguely positive—no real pattern is obvious.

The correlation plot (Tech. appendix, Figure 3) also illustrates these relationships.

2.

[This whole section from the homework needed redone but I didn't have time.
I ended up choosing the model based on raw crime numbers, with region included, but no interaction terms]

3.

[same as #2]

4.

[not sure what goes here. This whole answer seems to fit in the Discussion section.]

Discussion

1.

Most of the relationships we noticed are expected. [will elaborate later]

2.

According to the model, per-capital income and crime have a positive relationship. Since we know that the people earning the higher incomes tend to commit less crimes, the relationship is most likely due to other variables, such as population-density. Bigger cities are likely to have more crimes as well as higher average incomes.

The previous two answers do not change when per-capita crime used instead of total crimes. The coefficient on crime is still positive (about 0.459) and analysis of covariance still suggests that the interaction terms are unneeded ($F \approx 0.120$). However, the per-capita crime coefficient is less significant in that model and the crimes coefficient is significant in the model I chose, which the main reason I chose it.

Because higher per-capita income is not actually caused by crime, we are only interested in which variable predicts that income better: either total crimes or per-capita crimes. Because of lurking variables or other reasons, total crimes predicts per-capita income better than per-capita crimes. If we were using a predictor that was more likely to have a direct affect on income, like college education, it would make more sense to use the 'per-capita' version of both predictor and response for the sake of interpretability and consistency.

3.

[still needs filled in: strengths and weaknesses of model choices. interpret coefficients – both literal meaning in terms of 'slope' and the meaning of them being in the model. discuss limitations of study – this leads into #4]

4.

We have two important pieces of information about how the 440 counties in the dataset were chosen:

-They are the 440 most populous counties in the U.S. (with exceptions-see below).

-Any observations with missing values were deleted from the set.

Each of these facts are reasons that the sample of counties in the study is not random. There is plenty of reason to doubt that the most populous counties in the U.S. are representative of the rest of the 2600ish counties. The minimum county population in the dataset was over 10,000, but there are plenty of counties with only a few thousand people and even a few hundred people. One county in Hawaii has 86 people! We cannot assume that the models based on the largest counties would generalize to small or medium sized counties, or any actual random sample with a wide range of county populations.

Deleting any observations with missing values of a variable is another non-random method of choosing counties. It is possible that the observations with missing data tend to have something in common and that the data is missing for a reason. This might overlap with the last problem since smaller counties are less likely to keep complete records.

Because of the reasons above—yes, we should be worried about the missing counties.

References

Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005) Applied Linear Statistical Models, Fifth Edition. NY: McGraw-Hill/Irwin. Original source: Geospatial and Statistical Data Center, University of Virginia.

Project 1 (working title) - Technical Appendix

Zach Ohl

Load packages:

```
library(tidyverse)
library(car)
library(ggplot2)
library(knitr)
library(kableExtra)
library(reshape2)
library(reshape)
library(grid)
library(gridExtra)
```

Read in data

Make sure no missing data. Visual inspections suggests no other types of NAs.

```
all(!is.na(income))
[1] TRUE
```

Table 1 shows five-number summaries and means/standard deviations of the numeric variables.

```
# getting kable errors during knit, so i'm using crappy table version
# for now apply(income_numeric,2,function(x) c(summary(x),SD=sd(x)))
# %>% as.data.frame %>% t() %>% round(digits=2) %>% kable(caption =
# 'Summary tables of the numeric variables') %>%
# kable_styling(latex_options = 'HOLD_position')
cat("Table 1: Summaries of numeric variables")
Table 1: Summaries of numeric variables
summary(income_numeric)
   land.area            pop             pop.18_34      pop.65_plus
 Min.   :   15.0   Min.   : 100043   Min.   :16.40   Min.   : 3.000
 1st Qu.:  451.2   1st Qu.: 139027   1st Qu.:26.20   1st Qu.: 9.875
 Median :  656.5   Median : 217280   Median :28.10   Median :11.750
 Mean   : 1041.4   Mean   : 393011   Mean   :28.57   Mean   :12.170
 3rd Qu.:  946.8   3rd Qu.: 436064   3rd Qu.:30.02   3rd Qu.:13.625
 Max.   :20062.0   Max.   :8863164   Max.   :49.70   Max.   :33.800
    doctors          hosp.beds           crimes        pct.hs.grad
 Min.   :   39.0   Min.   :   92.0   Min.   :   563   Min.   :46.60
 1st Qu.:  182.8   1st Qu.:  390.8   1st Qu.:  6220   1st Qu.:73.88
 Median :  401.0   Median :  755.0   Median : 11820   Median :77.70
 Mean   :  988.0   Mean   : 1458.6   Mean   : 27112   Mean   :77.56
 3rd Qu.: 1036.0   3rd Qu.: 1575.8   3rd Qu.: 26280   3rd Qu.:82.40
 Max.   :23677.0   Max.   :27700.0   Max.   :688936   Max.   :92.90
  pct.bach.deg   pct.below.pov      pct.unemp      per.cap.income
 Min.   : 8.10   Min.   : 1.400   Min.   : 2.200   Min.   : 8899
 1st Qu.:15.28   1st Qu.: 5.300   1st Qu.: 5.100   1st Qu.:16118
```

```
 Median :19.70   Median : 7.900   Median : 6.200   Median :17759
 Mean   :21.08   Mean   : 8.721   Mean   : 6.597   Mean   :18561
 3rd Qu.:25.32   3rd Qu.:10.900   3rd Qu.: 7.500   3rd Qu.:20270
 Max.   :52.30   Max.   :36.300   Max.   :21.300   Max.   :37541
   tot.income
 Min.   :  1141
 1st Qu.:  2311
 Median :  3857
 Mean   :  7869
 3rd Qu.:  8654
 Max.   :184230
```

```
# kable errors during knit. Using crappy table versions for now tmp
# <- rbind(with(income, table(region))) row.names(tmp) <- 'Freq' tmp
# %>% kbl(booktabs=T,caption=' ') %>% kable_classic(full_width=F)
cat("Table 2: Summaries of categorical variables State and Region")
Table 2: Summaries of categorical variables State and Region
table(income$state)

AL AR AZ CA CO CT DC DE FL GA HI ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC
 7  2  5 34  9  8  1  2 29  9  3  1 17 14  4  3  9 11 10  5 18  7  8  3  1 18
ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV
 1  3  4 18  2  2 22 24  4  6 29  3 11  1  8 28  4  9  1 10 11  1

table(income$region)

 NC  NE   S   W
108 103 152  77
```

Check for unique values of county to make sure it's not a categorical variable.

```
length(unique(income$county))  #373 'unique' counties
[1] 373
length(unique(paste(income$county, income$state)))  #actually 440 unique countys
[1] 440
```

Numeric variable distributions:

```
ggplot(gather(income_numeric), aes(value)) + geom_histogram(bins = 12) +
    facet_wrap(~key, scales = "free_x") + theme(strip.text = element_text(size = 14,
    color = "red"))
```

```
# NOTE: this will be replaced with smaller group of only relevant
# pairwise plots that aren't tiny
pairs(income_numeric[-c(1, 2, 6, 14), ])  #removed 2 population outliers, 1 crime outlier, and 1 land o
```

Define new transformed vars:

Because of the skew. take log tranformations of:
Here are the distributions of the variable with ____ transformed:

To summarize two-way relationship among variables, look at this correlation graph, which shows higher positive correlations as blue and negative correlations as red.
Correlation graph: (graph doesn't work)

```
corgraph <- function(df) {
    cormat <- cor(df)
```

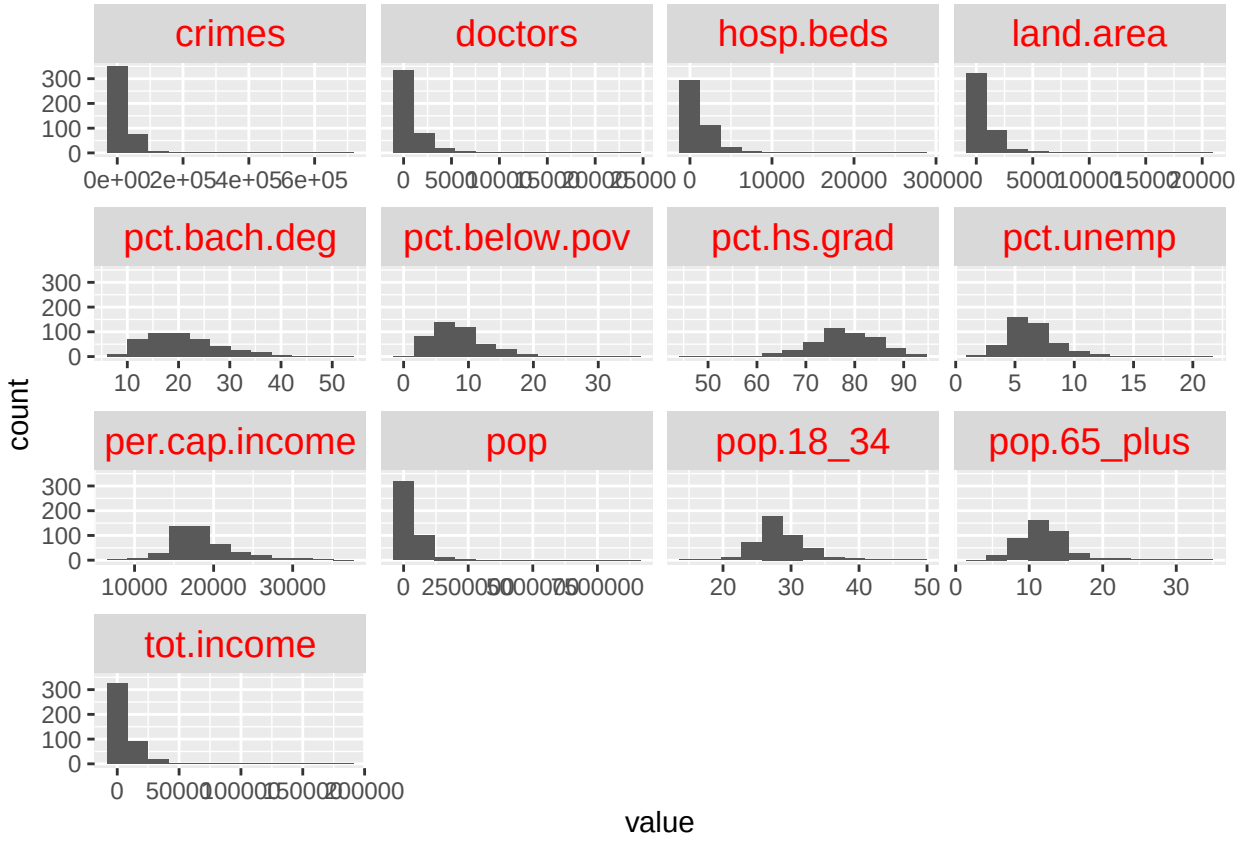

Figure 1: Histograms of the numeric variables

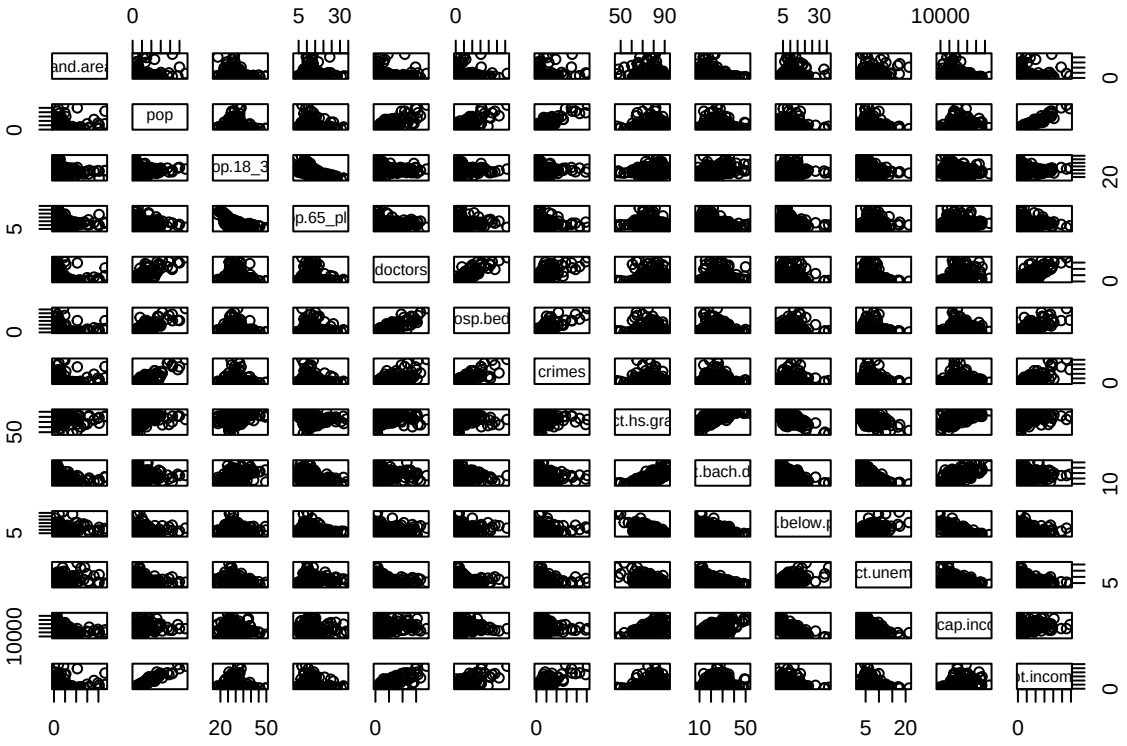

Figure 2: Pairs plots for numeric variables

```
    melted_cormat <- melt(cormat)  ## need library(reshape2) for this...
    ggplot(data = melted_cormat, aes(x = Var1, y = Var2, fill = value)) +
        geom_tile() + theme(axis.text.x = element_text(angle = 45, vjust = 0.9,
        hjust = 1)) + scale_fill_gradient2(low = "red", mid = "white",
        high = "blue")
}
corgraph(income_numeric)
```

.

Also, the correlation matrix for the variables after the transformations to correct skew are shown below.

From looking at a matrix of every possible pairs, most of the pairs didn't show any significant pattern or correlation, as suggested by the heat plot above. Of the correlations that were apparent, most appeared linear, which could signal good potential for the model, as well as collinearity.

```
income_num_region <- data.frame(income_numeric, region = income$region)

scatter.builder <- function(df, yvar = "per.cap.income") {
    result <- NULL
    y.index <- grep(yvar, names(df))
    for (xvar in names(df)[-y.index]) {
        d <- data.frame(xx = df[, xvar], yy = df[, y.index])
        if (mode(df[, xvar]) == "numeric") {
            p <- ggplot(d, aes(x = xx, y = yy)) + geom_point() + ggtitle("") +
                xlab(xvar) + ylab(yvar)
        } else {
            p <- ggplot(d, aes(x = xx, y = yy)) + geom_boxplot(notch = F) +
                ggtitle("") + xlab(xvar) + ylab(yvar)
        }
        result <- c(result, list(p))
    }
    return(result)
}
```

```
grid.arrange(grobs = scatter.builder(income_num_region))
```

The scatter plots confirm the results the correlations heat plot. The region box plots show overlapping IQRs of each region, but there are two regions with noticeably higher per-capita incomes than the other two.