# The Effect that County Demographic Information has on Per Capita Income

Emily Zeng, emilyzen@andrew.cmu.edu

### **1. ABSTRACT**

Per capita income is an important measure to evaluate the standard of living in a population, specifically by looking at its relationship with certain variables associated with a county's economic, health, and social well-being. The data includes county demographic information for 440 of the most populous counties in the US and 14 variables pertaining to economics and other health/social well-being metrics. To answer the research questions presented, we use exploratory data analysis methods, linear regression models, variable selection, and model selection methods. We find for the different regions in the US, there is a significant difference in the average salary that an individual will make, and that to predict per capita income, interaction terms between region and some of the numeric variables were included to come up with the best model. Limitations of this study include the lack of second interaction terms, and model selection methods; implications of this study indicate that \_\_\_\_\_\_ (need to include implications).

### 2. INTRODUCTION

The 4 main research questions of this study are as follows: 1. What pairwise relationships between the variables exist? 2. How are crimes and region related to per capita income? 3. What is the best model that predicts per capita income? 4. Does it matter that there is missing data for counties and states in the data? The importance of being able to predict per capita income is because if we can predict average income for an individual, we can then determine the standard of living in different areas in the US. (probably need to elaborate a little more on why should we care)

#### 3. DATA

The data used in this study is from Kutner (Kutner 2005) that includes county demographic information for 440 counties information, where the variables are defined in *Table 1* (page 3) from the years 1990 - 1992. Looking at *Table 1*, we see that identification number is the same as the row number, which is not a very helpful variable. A summary table for the numerical variables is shown in *Table 2* (page 4). Region is a categorial variable, so a separate frequency table is shown in *Table 3* (page 4). We see that there are the most datapoints in the southern region of the US. After initial exploratory data analysis, we can see that the best predictors for per capita income are pct.below.pov, pct.hs.grad, and pct.bach.deg (*Figure 1*) (page 4). We also look at the relationships between per capita income, and all the other numerical variables in *Figure 2* (page 5). As mentioned earlier, it does seem like pct.below.pov, pct.hs.grad, pct.bach.deg, and maybe pct.unemp are the best predictors for per capita income. When looking at *Figure 3* (page 5), multiple of the numerical variables are right skewed, indicating that their distribution would be more normal if transformed via a log transformation. After taking log transformations of a select few variables, the skewedness look better in their histograms, as seen in *Figure 4* (page 6). We chose to transform these 6 variables by looking at the best 4 predictors for per capita income (as mentioned above) and the other 2 by looking at *Figure 2*. Lastly, we want to see if per capita income differs by region. In *Figure 5 (page 7)*, the boxplot shows that the mean of the northeast region (NE) has a much higher mean per capita income when compared to the other 3 regions.

#### 4. METHODS

For the first research question, we look at correlogram plots and scatterplots of per capita income vs all the numerical variables to determine what relationships each pair of variables had with each other.

For the second research question, we fit linear models using the base R lm() function. These models include crime, region, and/or interactions between the two variables. Additionally, since the question asks whether there is a difference in choosing a model when defining crime as "number of crimes" or as "per capita crime = (number of crimes) / population", we fit models that include per capita crime, region, and/or interactions between the two variables. We use ANOVA tests and BIC/AIC for model selection to determine which of the models is better in terms of statistical significance and information criteria. For the third research question, we use all subsets without the variable 'region' for variable selection to fit the best model that doesn't include interaction terms. Afterwards, we calculate the variance inflation factors to determine if there are multicollinearities between the predictor variables. We then add in interaction terms between region and all the other numeric variables, before choosing only the interaction terms that are statistically significant. Lastly, we compare the initial model all subsets chose without any interactions with the model that has some interaction terms between region and the other numerical variables using an ANOVA test.

For the fourth research question, we look at the missing data on states and counties and reference data found online about population density and land area to determine whether having no data on certain states and counties is a concern.

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 physi- cians	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, rob- bery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school grad- uates	Percent of adult population (persons 25 years old or older) who com- pleted 12 or more years of school
12	Percent bachelor's de- grees	Percent of adult population (persons 25 years old or older) with bach- elor'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 pop- ulation (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 Cen- sus, 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. Variable definitions in CDI 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
рор	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 statistics for numeric variables

	NC	NE	S	W
Freq	108	103	152	77

Table 3. Frequency table for region



Figure 1. Correlation heatmap of numeric all variables in cdi dataset



Figure 2. Scatterplots between per.cap.income vs all numeric variables



Figure 3. Histograms of all numeric variables



Figure 4. Histograms after transforming chosen variables



Figure 5. per.cap.income differences based on region (NC, NE, S, W)

#### 5. **RESULTS**

Our first research question deals with the issue of what specific relationships are there between the variables in our CDI dataset. When we look at the data one pair of variables at a time, we see that total income is related to population, doctors, hospital beds, and crimes, population is related to doctors, hospital beds, and crimes, and per capita income is related to percent hs grad, pct bach deg, and total income (page 11 of the appendix). These are eveballed results of which variables are highly correlated with each other, but we can be more meticulous. We can show the variables that have strong correlations with each (define strong correlation as r > 0.4) in *Table 4*. (NEED TO FIX ALIGNMENT IN TABLE) If we look at the relationships between the numeric variables, many of them make sense. For example, it makes sense that doctors and number of hospital beds will have some sort of relationship since if there aren't enough doctors, there might not be as many hospital beds. Similarly, whether one has a bachelor's degree will influence one's income since having a bachelor's degree will allow one to make more money. One relationship that is surprising is doctors and crimes, with a very strong positive relationship. (AM I SUPPOSED TO EXPLAIN EVERY SINGLE RELATIONSHIP B/T THE VARS) (SHOULD I BE DOING REGION WITH EVERY NUMERIC VARIABLE)

Variable 1	Variable 2	Correlation (r)
рор	tot.income	0.9867476
doctors	hosp.beds	0.9504644
doctors	tot.income	0.9481106
рор	doctors	0.9402486
рор	hosp.beds	0.9237384
hosp.beds	tot.income	0.9020615
рор	crimes	0.8863318
hosp.beds	crimes	0.8568499
crimes	tot.income	0.8430980
doctors	crimes	0.8204595
pct.hs.grad	pct.bach.deg	0.7077867
pct.bach.deg	per.cap.income	0.6953619
pct.hs.grad	pct.below.pov	-0.6917505
pop.18_34	pop.65_plus	-0.6163096
pct.below.pov	per.cap.income	-0.6017250
pct.hs.grad	pct.unemp	-0.5935958
pct.bach.deg	pct.unemp	-0.5409069
pct.hs.grad	per.cap.income	0.5229961
pop.18_34	pct.bach.deg	0.4560970
pct.pelow.pov	pct.unemp	0.4369472
pet.bacil.ueg	pct.below.pov	-0.4084238

Table 4. Strong correlations between numeric variables

Our second research question asks the question of do crime and region influence per capita income, and whether this relationship is different when we define "crime" in 2 different ways. According to our linear regression model (page 19 of appendix) and taking into account interactions between crime and region, we come to the final model of

$$per \ capita \ income = crimes + region \tag{2.1}.$$

In the end, defining "crime" as the number of crimes instead of as "per capita crime", i.e. (number of crimes) / population, does make a difference. However, model 2.1 that uses crimes is better in terms of AIC and BIC. (probably need to insert a table comparing r square, anova, and aic/bic)

Our third research question asks the question of what is the best model to predict per capita income, when taking into account all of the variables in the CDI dataset. Our final model comes down to (page 24 of appendix)

Per capita income = log(land.area) + pop.18\_34 + log(doctors) + pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp)+ pct.hs.grad\*region + pct.bach.deg\*region + + log(pct.below.pov)\*region (NEED TO PUT INTO LATEX FORM) (3.1)

We look at the model diagnostics plots for model 3.1 (*Figure 6* and page 27 of appendix). The residuals are roughly centered around 0 and have constant variance. There seems to be a right tail in the qq plot. There doesn't seem to be any points that are outliers and/or highly influential. Despite the diagnostic plots not looking perfect, this is a tradeoff we are willing to make. (SHOULD I HAVE SUMMARY OF MODEL OR COEFFICIENTS??)



Figure 6. Diagnostic plots for model chosen by regsubsets including some interaction terms between region and other numeric variables

Our fourth question asks the question of whether it'd be an issue that there are missing county and state data. We see that three states are missing out 51 (this number counts and includes DC as the 51th state), Alaska, Iowa, and Wyoming (pages 28-29 in appendix) (States101.com). Out of the 3000 US counties, our data has 440 counties, but only 378 unique counties represented.

# 6. **DISCUSSION**

Our analyses and statistical methods all aim to answer the 4 research questions that were presented in the Introduction.

The first question is answered by looking at correlations between all of the numerical variables. The second is answered by building a model that includes crime and region to predict per capita income. The third is answered by also building the best model that includes potentially all numeric variables and region to predict per capita income. The last is answered by looking at the missing state data and making inferences about county data.

It is always important to interpret our final models we are presenting. Specifically, for our final model that answers research question 2, the interpretation is as follows. In the US, for every 1 unit of per capita income increase, there is a ~1% increase in crime. This increase is statistically significant. The different regions of the US have an effect on per capita income. We conclude this because for each region (NC, NE, S, and W), the baseline salaries are \$18,106, \$2,286+\$18,106 = \$20,392, -\$860+\$18,106 = \$17,246, and -\$142.83+\$18,106 = \$17,963. All of the salaries in each region are different, and all differ from the baseline salary in the NC region, which is \$18,106. Overall, the amount of money one makes in each region does differ, but it doesn't seem like it affects the crime rate (should probably explain why).

(Maybe insert a table that has baseline salaries for each region)

# (NEED TO HAVE INTERPRETATION FOR SUBSET MODEL)

The fourth research question poses the issue of having missing data for both states and counties. The interpretation is as follows. There are 48/51 states being represented in the data. The three missing states are Alaska, Iowa, and Wyoming. Alaska has the lowest population density in the entirety of the 50 states, with land area of ~ 86%. Iowa has a relatively small population density, with a land area of ~99%. Wyoming has an even smaller population density, also with a land area of ~99%. Approximately 96% of the data in terms of states is being represented in this sample of 48/51 states, which seems like a pretty good representation of the 51 states. Additionally, since the population density in these missed states is so small, relative to the other states in the US, missing these three states' data seems okay, as there are 48 other states that makes up for the missing data (States101.com).

In terms of county, it's a bit harder to determine whether it is an issue that only about 12% of the counties data is being represented. There only 378 unique counties out the 3000 total counties in the US. This is an issue that should be further investigated; it would be nice to know how the data was collected. For now, it's better to be safe and say that it is an issue

that so many counties are missing in the data. Unless the method in which the data for counties is disclosed, there is no concrete evidence that 378 counties is a good representative sample for the 3000 counties in the United States.

There are several limitations that this study suffers from. Firstly, there is no justification as to why state and county were not included in the model that answered question 3. State and county could be good predictors of per capita income, but they were completely left out from the models. Another limitation could be that linear model assumptions were potentially not met. This would cause any linear regression model to be invalid, as assumptions must be met before doing linear regression. As mentioned in research question four, there is missing data for counties and states. (probably need to improve/clarify limitations that are more specific to my model)

Future work can be done to determine how the data was collected, so that there can be further investigation on whether it is an issue that counties has so much missing data. Implications arise when we think about the differences in per capita income when it comes to the different regions of the US. The lower per capita income in a region, the lower the standard of living is. According to our data analyses (refer to interpretation of model 2 that answers question 2), the south and west regions of the US have the lowest per capita income. There could be two explanations for this: 1) the South is not as technologically advanced, or at least during the 90's, as the rest of the country, and 2) the western region of the US has a lot of land area, maybe with an emphasis of agriculture and horticulture, which isn't as high paying as other industry type jobs.

It might be a good idea to redo this study on newer data to see just how much the west and southern regions of the US have expanded. Additionally, it might be a good idea to focus resources on the southern and western regions, whether that's technological, economical, socially, etc., then their standard of living could be improved to be on par with the rest of the nation.

# 7. REFERENCES

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

Sheather, S. J. (2009), "A Modern Approach to Regression with R," Springer eBooks.

"U.S. States Populations, Land Area, and Population Density," *States101.com* [online]. Available at *https://www.states101.com/populations.* 

Williams, C. (2020), "How to Create a Correlation Matrix with Too Many Variables in R," *Towards Data Science.* 

# 8. TECHNICAL APPENDIX

library(glmnet)
library(MASS)
library(leaps)
library(car)
library(dplyr)

library(ggplot2)
library(stats4)
library(car)
library(mctest)
library(gtsummary)
library(kableExtra)
library(tidyr)
library(reshape2)

#### question 1

```
cdi <- read.table("../data/cdi.dat")
cdi_edit <- cdi[,-c(1,2,3,17)] ## remove id, state, county, and region</pre>
```

	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
рор	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 statistics for numeric variables

apply(cdi, 2, function(x) any(is.na(x))) ## doesn't seem to have any NA's in the data

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

	NC	NE	S	W
Freq	108	103	152	77

Table 3: Frequency table for region

cdi\_actual <- cdi[,-c(1,2,3)] ## remove id, state, and county</pre>

```
## histograms of all numeric vars - probably need to fix the x axes, numbers
are squished
ggplot(gather(cdi_edit), aes(value)) +
```

geom\_histogram(bins = 30) +

facet\_wrap(~key, scales = 'free\_x')





## crimes, doctors, hosp beds, land area, pop, and total income need to be lo
g transformed

```
## distribution of region
ggplot(cdi_actual, aes(x=region)) +
   geom_bar(fill='lightblue') + labs(x = "Region") ## most data is in southern
   region
```



Figure 2: Distribution of Region

## want to check correlation between predictors and lin relationship between
per.cap.income and all other predictors

```
corgraph <- function(df) {
  cormat <- cor(df)
  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="gold",mid="white",high="navy")
}</pre>
```

corgraph(cdi\_edit)



Figure 3: Correlation heatmap of numeric all variables in cdi dataset

high correlations: tot.income and pop, tot.income and doctors, tot.income and hosp.beds, tot.income and crimes, pop and doctors, pop and hosp.beds, pop and crimes, per.cap.income and pct.hs.grad, pct.bach.deg, tot.income issues with multicollinearity

# ## looking at the relationships b/t numeric vars where r > 0.4 and a correlogram

```
corr_simple <- function(data=df,sig=0.4){</pre>
  #convert data to numeric in order to run correlations
  #convert to factor first to keep the integrity of the data - each value wil
L become a number rather than turn into NA
  df cor <- data %>% mutate if(is.character, as.factor)
  df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
  #run a correlation and drop the insignificant ones
  corr <- cor(df cor)</pre>
  #prepare to drop duplicates and correlations of 1
  corr[lower.tri(corr,diag=TRUE)] <- NA</pre>
  #drop perfect correlations
  corr[corr == 1] <- NA</pre>
  #turn into a 3-column table
  corr <- as.data.frame(as.table(corr))</pre>
  #remove the NA values from above
  corr <- na.omit(corr)</pre>
  #select significant values
  corr <- subset(corr, abs(Freq) > sig)
```

```
#sort by highest correlation
corr <- corr[order(-abs(corr$Freq)),]
#print table
print(corr)
#turn corr back into matrix in order to plot with corrplot
mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")
#plot correlations visually
corrplot(mtx_corr, is.corr=FALSE, tl.col="black", na.label=" ")
}
```

```
corr_simple(cdi_edit)
```

##		Var1	Var2	Freq
##	158	рор	tot.income	0.9867476
##	70	doctors	hosp.beds	0.9504644
##	161	doctors	tot.income	0.9481106
##	54	рор	doctors	0.9402486
##	67	рор	hosp.beds	0.9237384
##	162	hosp.beds	tot.income	0.9020615
##	80	рор	crimes	0.8863318
##	84	hosp.beds	crimes	0.8568499
##	163	crimes	tot.income	0.8430980
##	83	doctors	crimes	0.8204595
##	112	pct.hs.grad	pct.bach.deg	0.7077867
##	152	pct.bach.deg	per.cap.income	0.6953619
##	125	pct.hs.grad	<pre>pct.below.pov</pre>	-0.6917505
##	42	pop.18_34	pop.65_plus	-0.6163096
##	153	<pre>pct.below.pov</pre>	per.cap.income	-0.6017250
##	138	pct.hs.grad	pct.unemp	-0.5935958
##	139	pct.bach.deg	pct.unemp	-0.5409069
##	151	pct.hs.grad	per.cap.income	0.5229961
##	107	pop.18_34	pct.bach.deg	0.4560970
##	140	<pre>pct.below.pov</pre>	pct.unemp	0.4369472
##	126	<pre>pct.bach.deg</pre>	<pre>pct.below.pov</pre>	-0.4084238



Figure 4: Correlogram of numeric variables, colored circles show high enough correlations

```
## scatterplots between all numeric vars and per.cap.income
cdi_edit %>%
 gather(-per.cap.income, key = "var", value = "value") %>%
ggplot(aes(x = value, y = per.cap.income)) +
  geom_point() +
  facet_wrap(~ var, scales = "free") +
  theme_bw()
```



*Figure 5: Scatterplots between per.cap.income vs all numeric variables* 

best predictors for per.cap.income: pos pct.bach.deg, neg pct.below.pov, pos pc.hs.grad, neg pct.unemp (some need transformations b/c not completely lin relationship)

```
## difference between region and per.cap.income using boxplot --> looks like
ne has significantly higher mean of per.cap.income
ggplot(cdi_actual,aes(x=region,y=per.cap.income)) +
geom_boxplot(notch=F)
```



*Figure 6: per.cap.income differences based on region (NC, NE, S, W)* 

```
## histograms of transformed vars
par(mfrow=c(3,2))
hist(log(cdi_actual$pct.bach.deg))
hist(log(cdi_actual$pct.below.pov))
hist(log(cdi$pct.hs.grad)) ## somehow worse - not going to transform
hist(log(cdi$pct.unemp))
hist(log(cdi_actual$doctors))
hist(log(cdi_actual$land.area))
```



Figure 7: Histograms of chosen transformed variables

## first 4 chosen from the best predictors in scatterplots against per cap in come, the other 2 are identified from looking at histograms of all other nume ric vars

need to add correlation plot with transformed vars and new scatterplots against per cap income vs transformed vars

```
question 2
## create models
mod1 <- lm(per.cap.income ~ crimes, data = cdi_actual)</pre>
mod2 <- lm(per.cap.income ~ crimes + region, data = cdi_actual)</pre>
mod3 <- lm(per.cap.income ~ crimes*region, data = cdi_actual)</pre>
summary(mod2) ## crimes and ne significant
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = cdi_actual)
##
## Residuals:
##
       Min
                 10 Median
                                 3Q
                                         Max
## -9661.0 -2260.7 -618.3 1650.0 19492.6
##
## Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
##
```

## (Intercept) 1.811e+04 3.784e+02 47.846 < 2e-16 \*\*\* ## crimes 8.915e-03 3.188e-03 2.797 0.00539 \*\* ## regionNE 2.286e+03 5.325e+02 4.293 2.17e-05 \*\*\* ## regionS -8.606e+02 4.868e+02 -1.768 0.07782 . ## regionW -1.428e+02 5.796e+02 -0.246 0.80548 ## ---## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Residual standard error: 3866 on 435 degrees of freedom ## Multiple R-squared: 0.1011, Adjusted R-squared: 0.09288 ## F-statistic: 12.24 on 4 and 435 DF, p-value: 1.946e-09 anova(mod1, mod2, mod3) ## mod2 is the best which is just crimes + region ## Analysis of Variance Table ## ## Model 1: per.cap.income ~ crimes ## Model 2: per.cap.income ~ crimes + region ## Model 3: per.cap.income ~ crimes \* region ## Res.Df RSS Df Sum of Sq F Pr(>F) ## 1 438 7133487504 ## 2 435 6501791845 3 631695660 14.1275 8.444e-09 \*\*\* ## 3 432 6438799739 3 62992106 1.4088 0.2396 ## ---0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## Signif. codes: pcc <- cdi\_actual\$crimes / cdi\_actual\$pop ## create new var for per capita cr</pre> ime moda <- lm(per.cap.income ~ pcc, data = cdi\_actual)</pre> modb <-  $lm(per.cap.income \sim pcc + region, data = cdi actual)$ modc <- lm(per.cap.income ~ pcc\*region, data = cdi\_actual)</pre> summary(modb) ## only ne significant ## ## Call: ## lm(formula = per.cap.income ~ pcc + region, data = cdi actual) ## ## Residuals: ## Min 10 Median 3Q Max -8634 -2300 -631 ## 1710 19333 ## ## Coefficients: ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 18006.04 537.04 33.528 < 2e-16 \*\*\* ## pcc 5773.20 7520.41 0.768 0.4431 ## regionNE 2354.70 541.97 4.345 1.74e-05 \*\*\* ## regionS -927.45 512.31 -1.810 0.0709 . ## regionW -34.92 586.03 -0.060 0.9525 ## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Residual standard error: 3898 on 435 degrees of freedom ## Multiple R-squared: 0.08622, Adjusted R-squared: 0.07782 ## F-statistic: 10.26 on 4 and 435 DF, p-value: 6.007e-08 anova(moda, modb, modc) ## second model does best again ## Analysis of Variance Table ## ## Model 1: per.cap.income ~ pcc ## Model 2: per.cap.income ~ pcc + region ## Model 3: per.cap.income ~ pcc \* region ## Res.Df RSS Df Sum of Sq F Pr(>F) ## 1 438 7186843542 ## 2 435 6609753963 3 577089580 12.5761 6.753e-08 \*\*\* ## 3 432 6607856753 3 1897210 0.0413 0.9888 ## ---0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## Signif. codes: ## need to compare mod2 and modb BIC(mod2, modb) ## mod2 is smaller bic ## df BIC ## mod2 6 8548.957 ## modb 6 8556.203 AIC(mod2, modb) ## same with aic ## df AIC ## mod2 6 8524.436 ## modb 6 8531.682 round(coef(summary(mod2)),2) ## Estimate Std. Error t value Pr(>|t|) ## (Intercept) 18106.91 378.44 47.85 0.00 ## crimes 0.01 0.00 2.80 0.01 ## regionNE 2286.04 532.47 4.29 0.00 ## regionS -860.56 486.83 -1.77 0.08 ## regionW -142.83 579.62 -0.25 0.81

in the end, it does matter if we use crimes or per capita crimes (crimes/population). the better model of the 2 using aic/bic as a measure shows that model 2 with just crimes and region as the predictor variables for per capita income instead of using per capita crimes.

interpretation of mod2: in the us, for every 1 unit of per capita income increase, there is a  $\sim 1\%$  increase in crime. this increase is statistically significant. different regions of the us has an effect on per capita income. for each region (nc, ne, s, and w) the baseline salaries are "18,106", 2,286+18,106 = 20,392, -860+18106 = 17246, and -142.83+18106 = 17963. all of the salaries in each region are different, and all differ from the baseline salary in the

nc region, which is 18106. overall, the amount of money one makes in each region does differ, but it doesn't seem like it actually affect crime rate itself.

```
## residual plots for all 6 models
oldmar <- par()$mar
par(mfrow=c(6,4))
par(mar=c(2,2,2,2))</pre>
```

Vesiduals vs Fitted         Normal Q-Q           Image: second	Scale-LocatidResiduals vs Levera 응 패배 아 패배 아이
Vesiduals vs Fitted         Normal Q-Q           Image: space of the space o	Scale-Location esiduals vs Levera
Vesiduals vs Fitted         Normal Q-Q           Image: second	Scale-Location esiduals vs Levera
Vesiduals vs Fitted         Normal Q-Q           ■         •         ■           16000         18500         -3         -1         1         3	Scale-LocatidResiduals vs Levera ::::::::::::::::::::::::::::::::::::
Vesiduals vs Fitted         Normal Q-Q           Image: second	Scale-Locatic Residuals vs Levera
Vesiduals vs Fitted         Normal Q-Q           Image: second	Scale-Locatid Residuals vs Levera

Figure 8: Residual plots for all 6 models mentioned above (mod1, mod2, mod3, moda, modb, modc)

par(mar=oldmar)
## need to fix margins, they're way too narrow and can't really see much and
fix figure number

question 3
cdi\_new <- cdi\_actual[,-c(2,13)] ## remove pop and total income because corre
Lated with per capita income
cdi\_no\_reg <- cdi\_new[,-c(12)] ## remove region</pre>

```
all.subset <- regsubsets(per.cap.income ~ log(land.area) + pop.18_34 + pop.65
_plus + log(doctors) + log(hosp.beds) + log(crimes) + pct.hs.grad + pct.bach.
deg + log(pct.below.pov) + log(pct.unemp), data = cdi_no_reg)</pre>
```

```
## let reg subsets find the best model for us
plot(all.subset)
```

```
which.min(summary(all.subset)$bic) ## model with minimum bic has 7 vars
## [1] 7
cbind(coef(all.subset,which.min(summary(all.subset)$bic))) ## coefficients of
the best reg subsets model that has 7 vars
##
                             [,1]
                     27091.20331
## (Intercept)
## log(land.area)
                       -623.02904
## pop.18_34
                       -249.38687
## log(doctors)
                      1119.67817
## pct.hs.grad
                       -72.82166
## pct.bach.deg
                       298.14366
## log(pct.below.pov) -3909.99239
## log(pct.unemp)
                       1678.42683
all.subset.mod <- lm(per.cap.income ~ log(land.area) + pop.18 34 + log(doctor
s) + pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp), data =
cdi no reg) ## general model chosen by reqsubsets
coef(summary(all.subset.mod)) ## all predictors are significant
##
                        Estimate Std. Error
                                                t value
                                                            Pr(>|t|)
## (Intercept)
                     27091.20331 1970.82790 13.746103 6.360451e-36
## log(land.area)
                                   94.76104 -6.574739 1.408797e-10
                       -623.02904
## pop.18_34
                       -249.38687 23.00306 -10.841463 2.189522e-24
## log(doctors)
                      1119.67817
                                   81.71519 13.702204 9.697278e-36
## pct.hs.grad
                       -72.82166
                                   19.58113 -3.718971 2.263959e-04
## pct.bach.deg
                       298.14366
                                   19.33180 15.422446 4.571538e-43
## log(pct.below.pov) -3909.99239 218.30290 -17.910858 4.673793e-54
## log(pct.unemp)
                      1678.42683 315.66424
                                              5.317127 1.694403e-07
```

best all reg subset model has 7 vars: land area, pop 18\_34, doctors, pct hs grad, pct bach deg, pct below pov, and pct unemp

if you look at the signs of the coefficients, we see several of them have the wrong sign (wrong direction of relationship with per cap income) - pct unemp and pct hs grad signs are wrong (look at original scatterplots w/ relationships b/t per capital income and the predictor vars)

vif(	all.subset.mod) ## none	e are above 5 actu	ıally	
## d	log(land.area)	pop.18_34	log(doctors)	pct.hs.gra
u ## 0	1.116567	1.520927	1.430151	3.08770

##	<pre>pct.bach.deg</pre>	<pre>log(pct.below.pov)</pre>	<pre>log(pct.unemp)</pre>
##	3.583155	2.194639	1.735595

# par(mfrow = c(2,2)) plot(all.subset.mod)



Figure 9: Diagnostic plots for model chosen by regsubsets

mmps(all.subset.mod) ## look at marginal model plots - blue lines line up wel
l with red dashed lines, probably didn't miss any transformationa/interaction
s



Figure 10: Marginal model plots for model chosen by regsubsets

all.subset.final.reg <- lm(per.cap.income ~ log(land.area) + pop.18\_34 + log(
doctors) + pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp) +
log(land.area)\*region + pop.18\_34\*region + log(doctors)\*region + pct.hs.grad
\*region + pct.bach.deg\*region + log(pct.below.pov)\*region + log(pct.unemp)\*re
gion, data = cdi\_new)</pre>

```
summary(all.subset.final.reg) ## pop.18_34, doctors, pct.bach.deg, pct.below.
pov, pct.unemp, regionw sign, landarea * regionne, pct.hs.grad*regionw, pct.b
elow.pov*regionw all sign interactions
```

```
##
## Call:
## lm(formula = per.cap.income ~ log(land.area) + pop.18_34 + log(doctors) +
##
       pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp) +
       log(land.area) * region + pop.18_34 * region + log(doctors) *
##
       region + pct.hs.grad * region + pct.bach.deg * region + log(pct.below.
##
pov) *
##
       region + log(pct.unemp) * region, data = cdi_new)
##
## Residuals:
       Min
                    Median
##
                10
                                3Q
                                        Max
## -4574.2 -918.8
                     -69.0
                             768.7
                                    6376.3
##
## Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
##
```

##	(Intercept)	24348.03	6074.50	4.008 7.27e-05 '	***
##	log(land.area)	-360.22	305.79	-1.178 0.239490	
##	pop.18_34	-269.27	54.96	-4.899 1.39e-06 '	***
##	log(doctors)	908.78	186.08	4.884 1.50e-06 '	***
##	pct.hs.grad	-53.71	69.50	-0.773 0.440035	
##	pct.bach.deg	278.07	59.39	4.682 3.87e-06 '	***
##	<pre>log(pct.below.pov)</pre>	-3059.37	519.69	-5.887 8.22e-09 '	***
##	log(pct.unemp)	1848.92	607.63	3.043 0.002495 *	**
##	regionNE	9757.79	7627.71	1.279 0.201534	
##	regionS	834.28	6744.81	0.124 0.901620	
##	regionW	28193.60	8481.74	3.324 0.000967 *	***
##	<pre>log(land.area):regionNE</pre>	-100.35	405.23	-0.248 0.804538	
##	<pre>log(land.area):regionS</pre>	-430.46	351.28	-1.225 0.221131	
##	log(land.area):regionW	-110.36	366.69	-0.301 0.763601	
##	<pre>pop.18_34:regionNE</pre>	-119.91	77.46	-1.548 0.122400	
##	pop.18_34:regionS	13.20	64.17	0.206 0.837099	
##	pop.18_34:regionW	21.70	84.70	0.256 0.797958	
##	log(doctors):regionNE	66.93	268.02	0.250 0.802939	
##	log(doctors):regionS	53.44	231.45	0.231 0.817506	
##	log(doctors):regionW	153.63	258.08	0.595 0.551983	
##	<pre>pct.hs.grad:regionNE</pre>	-114.61	89.15	-1.286 0.199298	
##	<pre>pct.hs.grad:regionS</pre>	52.29	75.43	0.693 0.488555	
##	<pre>pct.hs.grad:regionW</pre>	-281.34	85.10	-3.306 0.001030	**
##	<pre>pct.bach.deg:regionNE</pre>	195.92	83.63	2.343 0.019622 '	*
##	<pre>pct.bach.deg:regionS</pre>	-31.34	65.24	-0.480 0.631230	
##	<pre>pct.bach.deg:regionW</pre>	125.04	73.82	1.694 0.091063	•
##	<pre>log(pct.below.pov):regionNE</pre>	-503.89	727.31	-0.693 0.488818	
##	<pre>log(pct.below.pov):regionS</pre>	103.85	634.03	0.164 0.869974	
##	<pre>log(pct.below.pov):regionW</pre>	-3384.68	874.56	-3.870 0.000127	***
##	<pre>log(pct.unemp):regionNE</pre>	-280.17	1023.02	-0.274 0.784329	
##	<pre>log(pct.unemp):regionS</pre>	-1592.38	865.77	-1.839 0.066601	•
##	<pre>log(pct.unemp):regionW</pre>	-1311.92	931.74	-1.408 0.159884	
##					
##	Signif. codes: 0 '***' 0.00	01 '**' 0.01	'*' 0.05	'.' 0.1 ' ' 1	
##	5				
##	Residual standard error: 152	26 on 408 de	grees of f	freedom	
##	Multiple R-squared: 0.8686	, Adjusted R	-squared:	0.8586	
##	F-statistic: 87.02 on 31 and	408 DF, p	-value: <	2.2e-16	

should keep: region b/c w sign, pct.hs.grad:region, log(pct.below.pov):region, pct.bach.deg:region drop: log(doctors):region, pop.18\_34:region, log(land.area):region, log(pct.unemp):region

all.subset.final.final <- lm(per.cap.income ~ log(land.area) + pop.18\_34 + lo
g(doctors) + pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp)
+ pct.hs.grad\*region + pct.bach.deg\*region + log(pct.below.pov)\*region, data
= cdi\_new) ## dropped interactions w/ region that were insignificant</pre>

summary(all.subset.final.final)

## ## Call: ## lm(formula = per.cap.income ~ log(land.area) + pop.18\_34 + log(doctors) + pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp) + ## pct.hs.grad \* region + pct.bach.deg \* region + log(pct.below.pov) \* ## ## region, data = cdi\_new) ## ## Residuals: ## Min 10 Median 3Q Max -4864.3 -820.6 -45.7 790.6 5909.6 ## ## ## Coefficients: ## Estimate Std. Error t value Pr(>|t|)## (Intercept) 27410.36 5194.61 5.277 2.11e-07 \*\*\* ## log(land.area) -599.99 108.95 -5.507 6.37e-08 \*\*\* ## pop.18\_34 -268.54 22.36 -12.009 < 2e-16 \*\*\* ## log(doctors) 1002.78 81.37 12.324 < 2e-16 \*\*\* 61.69 -0.814 0.415892 ## pct.hs.grad -50.24 ## pct.bach.deg 239.39 40.17 5.959 5.38e-09 \*\*\* 467.64 -6.624 1.07e-10 \*\*\* ## log(pct.below.pov) -3097.55 ## log(pct.unemp) 970.89 334.14 2.906 0.003859 \*\* ## regionNE 5514.53 6089.45 0.906 0.365673 ## regionS -5288.50 5528.07 -0.957 0.339288 ## regionW 25229.79 6443.16 3.916 0.000105 \*\*\* ## pct.hs.grad:regionNE -98.74 75.43 -1.309 0.191222 ## pct.hs.grad:regionS 55.14 67.90 0.812 0.417234 72.91 -3.768 0.000188 \*\*\* ## pct.hs.grad:regionW -274.71 3.421 0.000684 \*\*\* ## pct.bach.deg:regionNE 171.77 50.20 42.83 ## pct.bach.deg:regionS 23.58 0.550 0.582313 3.345 0.000896 \*\*\* ## pct.bach.deg:regionW 170.87 51.08 ## log(pct.below.pov):regionNE 633.22 -1.153 0.249655 -729.96 ## log(pct.below.pov):regionS 551.05 0.153 0.878090 84.58 ## log(pct.below.pov):regionW -3313.47 828.58 -3.999 7.52e-05 \*\*\* ## ---## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 ## ## Residual standard error: 1530 on 420 degrees of freedom ## Multiple R-squared: 0.8641, Adjusted R-squared: 0.858 ## F-statistic: 140.6 on 19 and 420 DF, p-value: < 2.2e-16 vif(all.subset.final.final) ## 5 greater than 5 GVIF Df GVIF<sup>(1/(2\*Df))</sup> ## ## log(land.area) 1.692031e+00 1 1.300781 ## pop.18 34 1.647587e+00 1 1.283584 ## log(doctors) 1.625573e+00 1 1.274980 1 ## pct.hs.grad 3.513328e+01 5.927333 ## pct.bach.deg 1.773780e+01 1 4.211627 ## log(pct.below.pov) 1.154453e+01 1 3.397724 ## log(pct.unemp) 2.229315e+00 1 1.493089

```
## region2.736822e+08325.480488## pct.hs.grad:region1.586555e+08323.267113## pct.bach.deg:region2.064872e+0435.237799## log(pct.below.pov):region8.945037e+0436.687498
```

## diagnostic plots for the final model chosen via reg subsets and with some
interaction terms w/ region
par(mfrow=c(2,2))

plot(all.subset.final.final)



Figure 11: Diagnostic plots for model chosen by regsubsets including some interaction terms between region and other numeric variables

diagnostic plots look decent... can't be perfect though

```
anova(all.subset.mod, all.subset.final.final) ## the 2nd model with some inte
raction terms with region better than model with no region
## Analysis of Variance Table
##
## Model 1: per.cap.income ~ log(land.area) + pop.18_34 + log(doctors) +
## pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp)
## Model 2: per.cap.income ~ log(land.area) + pop.18_34 + log(doctors) +
## pct.hs.grad + pct.bach.deg + log(pct.below.pov) + log(pct.unemp)
## pct.hs.grad * region + pct.bach.deg * region + log(pct.below.pov) *
## pct.hs.grad * region + pct.bach.deg * region + log(pct.below.pov) *
## region
```

## Res.Df RSS Df Sum of Sq F Pr(>F)432 1158947355 ## 1 420 982926692 12 176020663 6.2677 3.206e-10 \*\*\* ## 2 ## ---## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 BIC(all.subset.mod, all.subset.final.final) ## almost the same but first mode *L w/o interaction is better b/c bic favors smaller models* df BIC ## ## all.subset.mod 9 7808.408 ## all.subset.final.final 21 7808.967 AIC(all.subset.mod, all.subset.final.final) ## aic clearly favors model 2 wit h some region interaction terms b/c larger model ## df AIC ## all.subset.mod 9 7771.627 ## all.subset.final.final 21 7723.145 round(coef(summary(all.subset.final.final)), 2) ## Estimate Std. Error t value Pr(>|t|) 5.28 ## (Intercept) 27410.36 5194.61 0.00 ## log(land.area) -599.99108.95 -5.51 0.00 ## pop.18\_34 -268.54 22.36 -12.01 0.00 ## log(doctors) 1002.78 81.37 12.32 0.00 ## pct.hs.grad -50.24 61.69 -0.81 0.42 ## pct.bach.deg 239.39 40.17 5.96 0.00 ## log(pct.below.pov) -3097.55 467.64 -6.62 0.00 970.89 334.14 2.91 0.00 ## log(pct.unemp) ## regionNE 5514.53 6089.45 0.91 0.37 -5288.50 5528.07 -0.96 0.34 ## regionS ## regionW 25229.79 6443.16 3.92 0.00 ## pct.hs.grad:regionNE -98.74 75.43 -1.31 0.19 ## pct.hs.grad:regionS 55.14 67.90 0.81 0.42 72.91 ## pct.hs.grad:regionW -274.71 -3.77 0.00 171.77 50.20 3.42 0.00 ## pct.bach.deg:regionNE ## pct.bach.deg:regionS 23.58 42.83 0.55 0.58 170.87 51.08 3.35 0.00 ## pct.bach.deg:regionW ## log(pct.below.pov):regionNE -729.96 633.22 -1.15 0.25 ## log(pct.below.pov):regionS 84.58 551.05 0.15 0.88 ## log(pct.below.pov):regionW -3313.47 828.58 -4.00 0.00 \*\* need to try lasso at least, and then do model selection b/t the 2 \*\*

```
question 4
sort(unique(cdi$state))
## missing iowa, alaska, and wyoming
## [1] "AL" "AR" "AZ" "CA" "CO" "CT" "DC" "DE" "FL" "GA" "HI" "ID" "IL" "IN"
"KS" "KY" "LA" "MA"
```

[19] "MD" "ME" "MI" "MN" "MO" "MS" "MT" "NC" "ND" "NE" "NH" "NJ" "NM" "NV" "N Y" "OH" "OK" "OR"

[37] "PA" "RI" "SC" "SD" "TN" "TX" "UT" "VA" "VT" "WA" "WI" "WV"

there are 48/51 states (includes dc) being represented in the data. the 3 missing states are Alaska, iowa, and Wyoming. Alaska has the lowest population density in the entirety of the 50 states, with ~ 86% land area. Iowa has a relatively small population density, with a ~99% land area. Wyoming has an even smaller population density, and is also ~99% land area. ~96% of the data in terms of states is being represented, which I think is a good sample size for the 51 states (including dc). Additionally, since the population density in these missed states is so small, relative to the other states in the us, missing these 3 states' data seems okay, as there are 48 other states (including dc) that makes up for the missing data.

in terms of county, it's a bit harder to determine whether it is ok that only 10% of the counties data is being represented. There only 378 unique counties out 3000 total counties in the us. This is an issue that should be further investigated; it would be nice to know how the data was collected. For now, I would say it's an issue that so many counties are missing in the data, and unless the method in which the data for counties is disclosed, there is no concrete evidence showing that 378 counties is a good representative sample for the 3000 counties in the united states.