What factors influence the income per person? Is there any relevant factor which looks surprising?

Hongsheng Xie|hongshex@andrew.cmu.edu

Abstract

For most people around the world, income and salary is one of the most important issue because money can improve the quality of people's life. This report will use a dataset containing information on many aspects to explore which factors will play a great role in affecting income per person, especially the factors which will not be considered at the beginning. By doing Exploratory Data Analysis, multivariable regression, diagnosis visualization, and stepwise variable selection, several factors that affect income significantly are summed up. Crime, one surprising factor in the final variable set, is studied independently, and the reason why crime works is explained. However, even though the final conclusion looks reasonable, lack of data for most of counties reduces the credibility, and further research is needed.

Introduction

Some (fictional!) social scientists are interested in looking at the historical data, to learn how average income per person was related to other variables associated with the county's economic, health and social well-being. With the dataset containing variables including distribution of age, education background, crime situation, etc., a statistical analysis can be developed to generate a professional conclusion that what factors play a great role for per capital income.

Data

The data used to study factors affecting average income per person is taken from Kutneret al. (2005)1: It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The definitions of the variables are given in Table 1 on p. 2 of this document.

Methods

For question one, correlation graph is used to show the correlations between each pair of variables. Pairs with correlated coefficients are studied, and surprised correlations are explained with analysis. For question two, the dataset is firstly separated depending on regions, and then summary tables are generated to study the distribution of per.cap.income in all regions. After looking at the data characteristic in different region, we study whether crimes or per.cap.crime works better in the regression model with per.cap.income. Regression models with and without interaction are tested and the best models are chosen. Then compare the two best models using crime and per.cap.crime with AIC and BIC tests to check which model works better. For question three, because county, state and id are useless categorical variables, these variables are not considered. Besides, because per.cap.income is a deterministic function of population and total income, using these two variables will disturb other variables seriously. So, these two variables are also not considered. Then a multivariable regression model is fitted with logarithm transformation on the variables which skew extremely. Finally, stepwise multivariable selection is adopted and the variable set with lowest AIC is selected. For question four,

Results

From the correlation table, correlation for each pair of variables is shown. 1. Tot.Income. Tot.Income has a strong positive correlation with pop, doctor, hosp.beds, and crimes, and has a medium positive correlation with land. area, pct.bach.deg, and per.capital.income. A reasonable person would expect a strong correlation between population and total income, number of active physicians(doctor), hospital beds because more people always mean larger Gross Domestic Product, more physicians, more hospitals and hospital beds and then larger total income. However, he/she may not expect a correlation between total income and crime. Considering number of crimes is calculated by population multiplying criminal rate, crime may have a strong positive correlation with total income because the positive correlation between crime and population. 2. Pct.unemp. Pct.unemp has a strong positive correlation with pct.below.pov. It is convincing because unemployed people don't have enough income and are always below poverty level. Pct.unemp has a strong negative correlation with pct.hs.grad and pct.bach.deg. It can be explained that high school and bachelor graduated students have sufficient knowledge to help them find a job. Seeing most of the correlations for a pair of variables are reasonable, this report will then focus on the surprising ones. There are strong positive correlations between crimes and doctors and between crimes and hosp.beds. They are surprising because physician is a career with very low criminal rate in common sense. These surprising results appear because both of crimes and doctor have a strong correlation with population. High crimes always mean high population, and high population will then cause large number of active physicians. This logic also works for the correlation between hosp.beds and crimes.

After classifying all data by four regions, the summary tables show that the region with largest crimes median is W, the median for S is a bit less, and the medians for NC and NE are around 70% of median of W. However, means of crimes are quite different. Means of crimes in S, NC, and NE are similar, but the mean of crimes for W region is much larger. This means that the distribution graph for W, NC, NE region skews to the right and a small part of quite high values increase mean value significantly. Compared to other three regions, S region has a more normal distribution. The difference between mean and median is not quite big.

With regression models using logarithm of crimes, logarithm of crimes and region, and interaction between logarithm of crimes and region, it shows that a regression model with logarithm of crimes and region has a lowest P value and does the best. Regression models using logarithm of per-capita crime, region, and their interaction are also tested, and it shows that a model with both logarithm of per-capita crime and regions, without interaction works the best. Comparing the two best models and using AIC and BIC test models, it shows that the model using total crime numbers performs better in both tests. With the coefficient estimates, the baseline per.cap.income in each region could be estimated. In the NC region, the baseline salary is $9,798.65. In the NE it is $10,829.18. In the S it is $8,955.29, and in the W it is $9,228.02.

Without the category variables which can hardly show the correlation with per.cap.income and the variables which will disturb the performance of other variables, the multivariable regression model shows that all variables in the model now are significant. The marginal model plots also fit quite well and prove the phenomenon. After stepwise multivariable selection, it shows that model with log.land.area,

pop.18_34, log.doctor, pct.hs.grad, pct.bach.deg, pct.below.pov, and pct.unemp has the lowest AIC and is the best model.

For the counties not in the dataset, research about the population of different counties is taken, and it shows that the mean and median value of population for all counties in United State are much smaller than those for the counties in the dataset. The lack of data might be caused by the small size of missing counties or difficulties in research. But it can also be caused by negligence of investigation which will cause an unreasonable conclusion. So, researchers should either replenish the data for missing states and counties or specify the target and purpose of this research.

References & Citations

Kutneret al. (2005)1
36-617 Applied regression analysis HW6 solution

Proj 1

Hongsheng Xie

10/18/2021

```
data1 <- read.table("C:/Users/danie/Desktop/CMU/36-617 applied regression analysis/HW6/cdi.dat",
header=TRUE)

library(ggplot2)
```

Q1. part(a)

```
continue_series <- c(1,4,5,6,7,8,9,10,11,12,13,14,15,16)
category_series <- c(2, 3, 17)
df <- data.frame("Min" = rep(0,17),
                 "1st Qu" = rep(0,17),
                 "Medidan" = rep(0,17),
                 "Mean" = rep(0,17),
                 "3rd Qu" = rep(0,17),
                 "Max" = rep(0,17))
for (i in continue_series) {
  s <- summary(data1[,i])
  df[i,1] <- s[1]
  df[i,2] <- s[2]
  df[i,3] <- s[3]
  df[i,4] <- s[4]
  df[i,5] <- s[5]
  df[i,6] <- s[6]
}
for (i in category_series) {
  df[i,] <- c(NA,NA,NA,NA,NA,NA)
}

table(data1[,17])
```

```
##
##  NC  NE   S   W
## 108 103 152  77
```

```
NC <- data1[data1$region == "NC",]
NE <- data1[data1$region == "NE",]
S <- data1[data1$region == "S",]
W <- data1[data1$region == "W",]
```

```
df_na <- data.frame("IS.NA" = rep(TRUE,17),
                    "NA amount" = rep(0,17))

for (i in c(1:17)) {
  l <- data1[,i]
  s <- sum(is.na(l))
  if (s > 0) {
    df_na[i,1] = TRUE
    df_na[i,2] = s
  } else {
    df_na[i,1] = FALSE
    df_na[i,2] = 0
  }
}

# There is no missing value for any column.
```

```
par(mfrow=c(2,2))
boxplot(data1$land.area,main="land.area")
hist(data1$land.area)
# Land area skews extremely to the right. Most of the land areas are quite small.

#boxplot(data1$pop, main = "pop")
#hist(data1$pop)
# Population skews extremely to the right. Populations in most of the counties are quite small.

boxplot(data1$pop.18_34, main = "pop.18_34")
hist(data1$pop.18_34)
```

```
# Most of the frequencies of 25-30 years old people in counties are in the interval [20,35], but
there are both some outliers larger than 35% and smaller than 20%.

boxplot(data1$pop.65_plus, main = "pop.65_plus")
hist(data1$pop.65_plus)
# Most of the frequencies of 65+ years old people in counties are in the interval [5,20], but th
ere are both some outliers larger than 20% and smaller than 5%.

boxplot(data1$doctors, main = "doctors")
hist(data1$doctors)
```

```
# The numbers of active physicians skews extremely to the right. The numbers in most of the coun
ties are quite small.

boxplot(data1$hosp.beds, main = "hosp.beds")
hist(data1$hosp.beds)
# Number of hospital beds skews extremely to the right. The numbers in most of the counties are
 quite small.

boxplot(data1$crimes, main = "crimes")
hist(data1$crimes)
```

```
# Total serious crimes skew extremely to the right. The numbers in most of the counties are rela
tively small.

boxplot(data1$pct.hs.grad, main = "pct.hs.grad")
hist(data1$pct.hs.grad)
# Most of the percents high school graduates are in the interval [60,90], but there are some out
liers smaller than 60%.

boxplot(data1$pct.bach.deg, main = "bach.deg")
hist(data1$pct.bach.deg)
```

```
# Most of the percents bachelor's degree are in the interval [10,40], but there are some outlier
s larger than 40%.

boxplot(data1$pct.below.pov, main = "pct.below.pov")
hist(data1$pct.below.pov)
# The percentages below poverty level skew a little to the right. Most the the percentages are i
n the interval [0,20], but there are some outliers larger than 25%.

boxplot(data1$pct.unemp, main = "pct.unemp")
hist(data1$pct.unemp)
```

```
# The percentages of unemployment skew a little to the right. Most the the percentages are in th
e interval [0,10], but there are some outliers larger than 15%.

boxplot(data1$per.cap.income, main = "per.cap.income")
hist(data1$per.cap.income)
# Average income per person skews to the right. It is reasonable because  the percentage of high
school graduate is higher than bachelor's degree.

boxplot(data1$tot.income, main = "tot.income")
hist(data1$tot.income)
```

```
# Total personal income skews extremely to the right. Total personal income in most of the count
ies are quite small because of the distribution of population. Besides, small county is always l
ess prosperous than big counties and cause a lower average income.

pairs(~land.area+pop+pop.18_34+pop.65_plus+doctors+hosp.beds+crimes+pct.hs.grad+pct.bach.deg+pc
t.below.pov+pct.unemp+per.cap.income+tot.income, data = data1)
```

```
# The scatter matrix shows that the variables about amount like land area, population, amount of
hospital beds are closely related because generally, more land area means more population and mo
re crimes, hospitals and facilities. The percentage variables mostly skew to the right, and this
phenomenon needs further study.
```

part(b)

```
crime_rate <- data1$crimes / data1$pop
l <- lm(per.cap.income ~ crimes + region, data = data1)
l_inter <- lm(per.cap.income ~ crimes + region + region*crimes, data = data1)
summary(l_inter)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region + region * crimes,
##     data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8582.4 -2225.2  -676.2  1563.4 19504.7
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.800e+04  4.092e+02  43.995  < 2e-16 ***
## crimes           1.361e-02  7.882e-03   1.726   0.0851 .
## regionNE         2.573e+03  5.736e+02   4.487 9.28e-06 ***
## regionS         -1.056e+03  5.606e+02  -1.884   0.0602 .
## regionW         -5.654e+01  6.372e+02  -0.089   0.9293
## crimes:regionNE -1.272e-02  9.677e-03  -1.314   0.1895
## crimes:regionS   6.348e-03  1.136e-02   0.559   0.5765
## crimes:regionW  -4.295e-03  9.486e-03  -0.453   0.6509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3861 on 432 degrees of freedom
## Multiple R-squared:  0.1099, Adjusted R-squared:  0.09543
## F-statistic: 7.616 on 7 and 432 DF,  p-value: 1.122e-08
```

```
# The p-values for all the intersection parameters are not significant, so there might be no int
ersection between crime and region.

l_CR <- lm(data1$per.cap.income ~ crime_rate + data1$region)
l_CR_inter <- lm(data1$per.cap.income ~ crime_rate + data1$region + crime_rate:data1$region)
summary(l_CR_inter)
```

```
##
## Call:
## lm(formula = data1$per.cap.income ~ crime_rate + data1$region +
##     crime_rate:data1$region)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8637.7 -2333.9  -629.5  1759.1 19515.6
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 18077.3      895.2  20.193   <2e-16 ***
## crime_rate                   4379.1    15893.5   0.276    0.783
## data1$regionNE               2329.0     1101.4   2.115    0.035 *
## data1$regionS               -1010.4     1323.8  -0.763    0.446
## data1$regionW                -670.0     1983.9  -0.338    0.736
## crime_rate:data1$regionNE     288.4    20184.7   0.014    0.989
## crime_rate:data1$regionS     1558.9    20556.1   0.076    0.940
## crime_rate:data1$regionW    10655.5    32322.4   0.330    0.742
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3911 on 432 degrees of freedom
## Multiple R-squared:  0.08648,    Adjusted R-squared:  0.07168
## F-statistic: 5.842 on 7 and 432 DF,  p-value: 1.713e-06
```

```
# The p-values for all the intersection parameters are not significant, so there might be no int
ersection between crime rate and region.


summary(l_CR)
```

```
##
## Call:
## lm(formula = data1$per.cap.income ~ crime_rate + data1$region)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8634  -2300   -631   1710  19333
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18006.04     537.04  33.528  < 2e-16 ***
## crime_rate      5773.20    7520.41   0.768   0.4431
## data1$regionNE  2354.70     541.97   4.345 1.74e-05 ***
## data1$regionS   -927.45     512.31  -1.810   0.0709 .
## data1$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
```

```
# The model tells that the relationship between per.capital income and crime rate is not strong.
summary(l)
```

```
##
## Call:
## lm(formula = per.cap.income ~ crimes + region, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9661.0 -2260.7  -618.3  1650.0 19492.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.811e+04  3.784e+02  47.846  < 2e-16 ***
## crimes       8.915e-03  3.188e-03   2.797  0.00539 **
## regionNE     2.286e+03  5.325e+02   4.293 2.17e-05 ***
## regionS     -8.606e+02  4.868e+02  -1.768  0.07782 .
## regionW     -1.428e+02  5.796e+02  -0.246  0.80548
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3866 on 435 degrees of freedom
## Multiple R-squared:  0.1011, Adjusted R-squared:  0.09288
## F-statistic: 12.24 on 4 and 435 DF,  p-value: 1.946e-09
```

```
# The answers change when I use the number of crime in the model. The number of the crime answer
s the question the best because number of crime has a close correlation with per capital income.
```

part(c,fig.show="hide") #Transformation

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

hist(data1$land.area)
data1$land.area <- log(data1$land.area)

hist(data1$pop)
data1$pop <- log(data1$pop)

hist(data1$pop.18_34)
data1$pop.18_34 <- log(data1$pop.18_34)

hist(data1$pop.65_plus)

hist(data1$doctors)
data1$doctors <- log(data1$doctors)

hist(data1$hosp.beds)
data1$doctors <- log(data1$hosp.beds)

hist(data1$crimes)
data1$crimes <- log(data1$crimes)

hist(data1$pct.hs.grad)

hist(data1$pct.bach.deg)

hist(data1$pct.below.pov)

hist(data1$pct.unemp)
data1$pct.unemp <- log(data1$pct.unemp)

hist(data1$per.cap.income)

hist(data1$tot.income)
data1$tot.income <- log(data1$tot.income)
```

#Interaction

```
l <- lm(per.cap.income~land.area+pop+pop.18_34+pop.65_plus+doctors+hosp.beds+crimes+pct.hs.grad+
pct.bach.deg+pct.below.pov+pct.unemp+tot.income, data = data1)
summary(l)
# After lots of try on different interactions, we find all of the interactions are not significa
nt.
```

#Variables Selection

```
backAIC <- step(l,direction="backward", data=data1[,continue_series])

om1 <- lm(per.cap.income~land.area, data = data1)
n <- length(om1$residuals)
backBIC <- step(l,direction="backward", data = data1[,continue_series], k=log(n))

# The best models selected by AIC backward and BIC backward are different. One has variable hos
p.beds, and one does not have. So we look at the regression model with hosp.beds and find it is
significant in 95% confidence. We decide to keep it.
l <- lm(per.cap.income ~ pop + pop.65_plus + doctors + hosp.beds + crimes + pct.hs.grad + pct.ba
ch.deg + pct.below.pov + tot.income, data = data1)
summary(l)

par(mfrow=c(2,2))
plot(l)

# The diagnosis plots show that the regression line on residual is not a straight line. However
it shows that the simple interactions are not significant, maybe there are some more complicated
interaction affect the diagnosis plots. But it is hard to explain such a complex model, so only
transformation with logarithm is enough.
```