STA 108 Project I

County Demographic Information

A Report Presented to

Professor Jiming Jiang of the Department of Statistics

University of California, Davis

In Partial Fulfillment

of the Requirements for the

Completion of STA 108

---

by
Ian Xu
Jeffrey Ugochukwu
Yiheng Lu

February 10, 2020

Introduction

In this project we use a data set containing county demographic information for 440 of the most populated counties in the United States. We are interested in the relationship between the number of active physicians and total population, hospital beds, and total personal income respectively.

The data set contains county demographic information for 440 of the most populated counties in the United States. Generally the data pertains to the years 1990-1992. Some counties with incomplete data were deleted from the data set.

The goal is to find out if different nests for sparrows on Kent Island attracted different size sparrows. We are interested in the relationship between the size of the sparrows and the living environment they are demanding. Single Factor ANOVA is used as our approach to determine the relationship between different nests and the size of sparrows. To obtain results of the 3 pairwise comparisons between groups, we use the Tukey correction. And Semi-studentized residual, Shapiro-Wilks method and Brown-Forsythe Test are applied to find the fitting model for our dataset.

"This data set 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

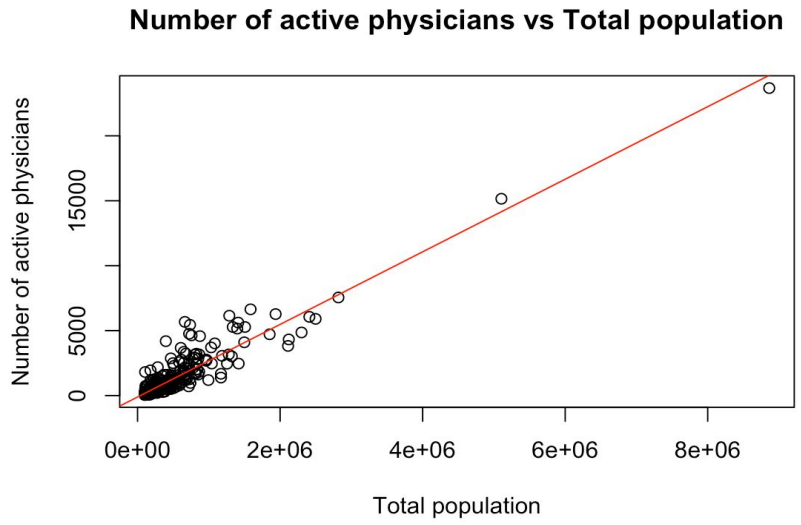
single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992."

I. Fitting regression models.

Three predictor variables

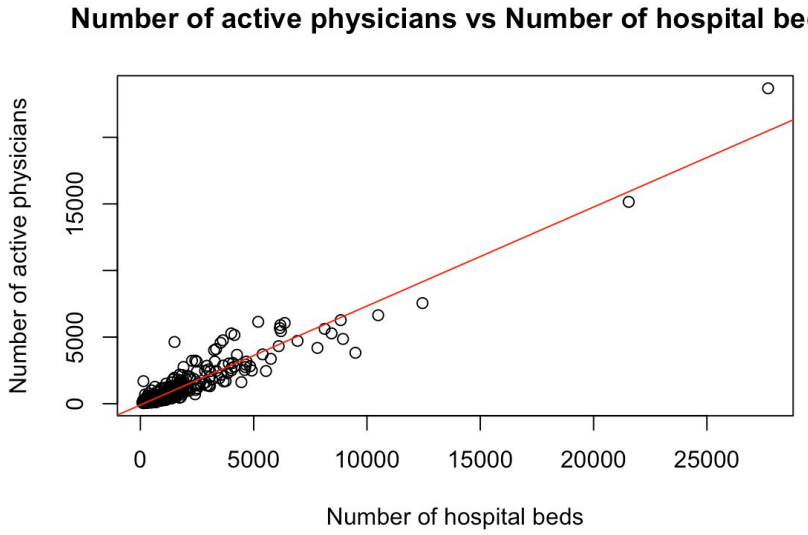
Total population vs number of active physician:

$\hat{Y}$ = -1.106e+02 + 2.795e-03X



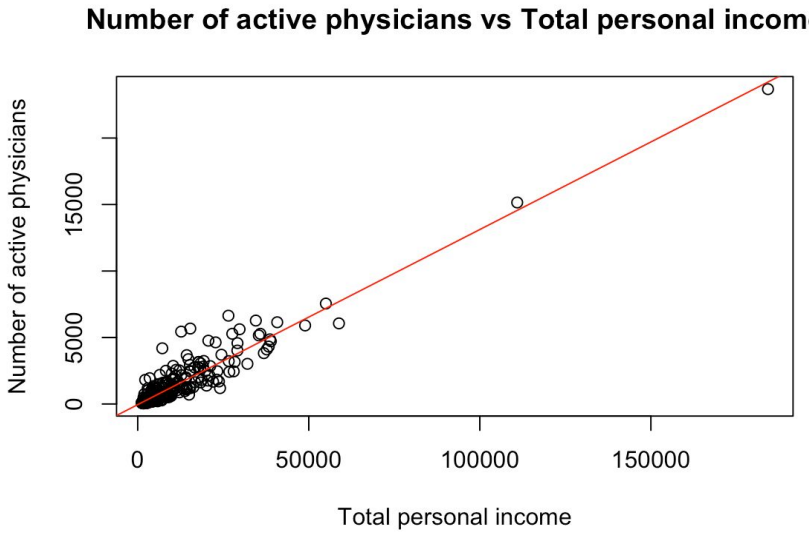
Number of hospital beds vs number of active physician:

$\hat{Y}$ = -95.93218 + 0.74312X



Total.personal.income vs number of active physician:

$\hat{Y} = -48.39485 + 0.13170X$



For all three predictor variables, the plots show that a linear regression appears to provide a good fit, but there are outliers.

MSE Table:

Predictor Variable	**MSE**
Total Population	372203.5
Number of Hospital Beds	310191.9
Total Personal Income	324539.4

The predictor variable "Number of Hospital Beds" with the smallest MSE leads to the smallest variability around the fitted regression line.

Four Region

Estimated Regression function:

Region 1(NE):$\hat{Y}$ = 9223.82 + 522.16X

Region 2(NC):$\hat{Y}$ = 13581.41 + 238.67X

Region 3(S):$\hat{Y}$ = 10529.79 + 330.61X

Region 4(W):$\hat{Y}$ = 8615.05 + 440.32X

The estimated regression functions appear to be similar numerically.

MSE Table:

Geographic Region	MSE
1(NE)	7335008
2(NC)	4411341
3(S)	7474349
4(W)	8214318

The NC region with the smallest MSE leads to the smallest variability around the fitted regression line.

II. Measuring linear association

R^2 Table

Predictor Variable	R^2
Total Population	0.8840674
Number of Hospital Beds	0.9033826

Total Personal Income	0.8989137

The number of hospital beds accounts for the largest reduction in the variability in the number of active physicians.

III. Inference about regression parameters

With regards to the inference of the regression parameters (Percent of Bachelor's Degrees vs Per Capita Income) for the 4 regions, the results for the 90% confidence intervals and ANOVA calculations per region are given below.

90% confidence intervals:

Geographic Region	90% Confidence intervals
1(NE)	(460.5177,583.80)
2(NC)	(193.4858,283.853)
3(S)	(285.7076,375.5158)
4(W)	(364.7585,515.8729)

ANOVA Tables:

Region 1(NE):

Source of variance	Degree of freedom	Sum of squares	Mean squares	F-statistic	P-value
Treatments	1	1450517671	1450517671	197.75	< 2.2e-16
Error	101	740835765	7335008		

Region 2(NC):

Source of variance	Degree of freedom	Sum of squares	Mean squares	F-statistic	P-value

Treatments	1	338907694	338907694	76.826	3.344e-14
Error	106	467602149	4411341		

Region 3(S):

Source of variance	Degree of freedom	Sum of squares	Mean squares	F-statistic	P-value
Treatments	1	1109873245	1109873245	148.49	< 2.2e-16
Error	150	1121152411	7474349		

Region 4(W):

Source of variance	Degree of freedom	Sum of squares	Mean squares	F-statistic	P-value
Treatments	1	773745787	773745787	94.195	6.856e-15
Error	75	616073841	8214318		

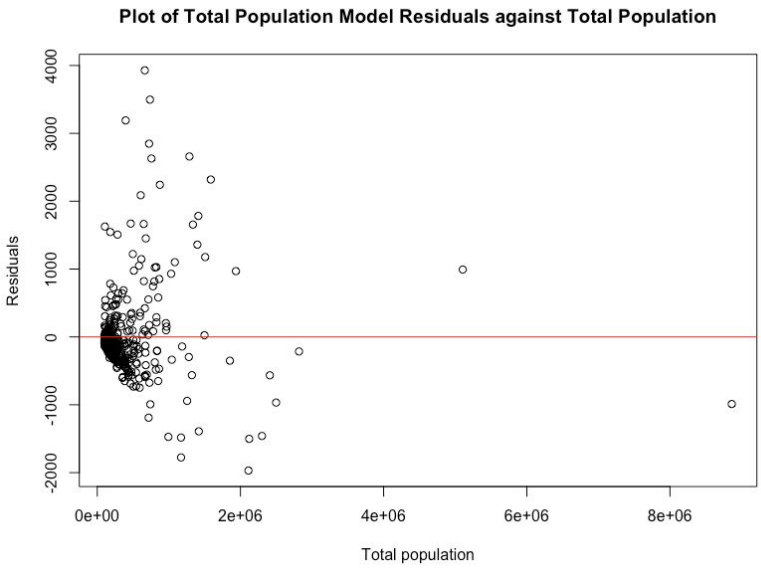
Through analyzing the 90% confidence intervals for β1 hat, each region, we see that there is no region in which all four regions overlap, but some of the regions overlap with each other. For instance, regions 3 and 4's confidence intervals are (285.7076, 375.5158) and (364.7585, 515.8729) respectively. These confidence intervals have an overlapping region—(364.7585, 375.5158). Because of this overlapping region, we cannot conclude that the difference between these two groups is statistically significant. Another instance of this is regions 1 and 4, which have confidence intervals of (460.5177, 583.80) and (364.7585, 515.8729) respectively. Here, we see an overlapping region of (460.5177, 515.8729), which means that we cannot conclude that the difference between these two groups is statistically significant. But since no other region combinations overlap with each other with respect to 90% confidence intervals, we can say that these other combinations are not statistically similar.

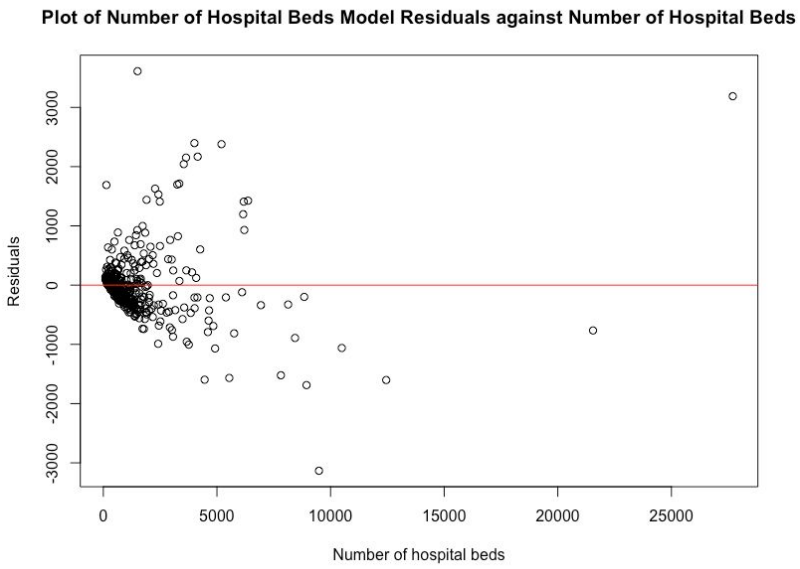
The ANOVA tables contain the F-Statistics and P-values for each region. Since the P-values, determined using the F-statistics, for each region is virtually 0, we can say that in

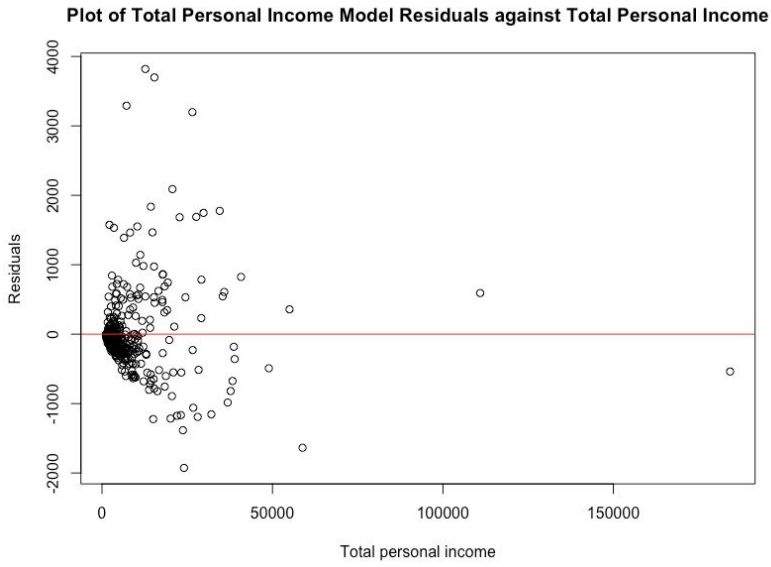
regions 1, 2, 3, and 4, there are linear relationships between individuals that have bachelor's degrees and their per capita income.

IV. Regression diagnostics

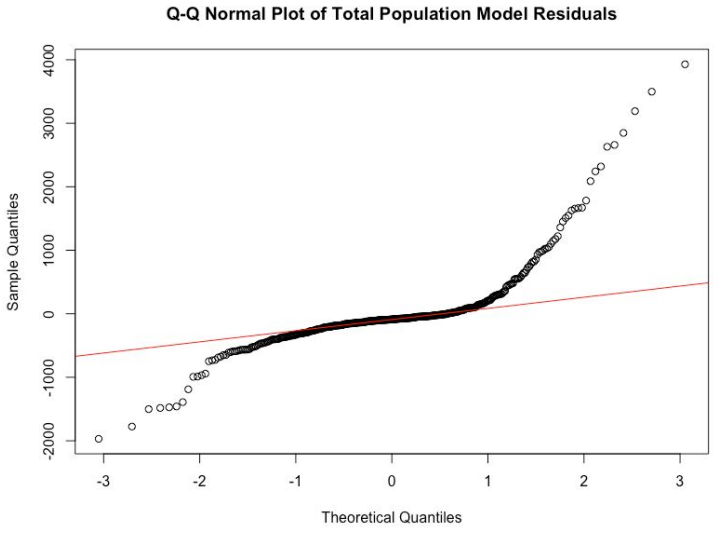
Residual plots:

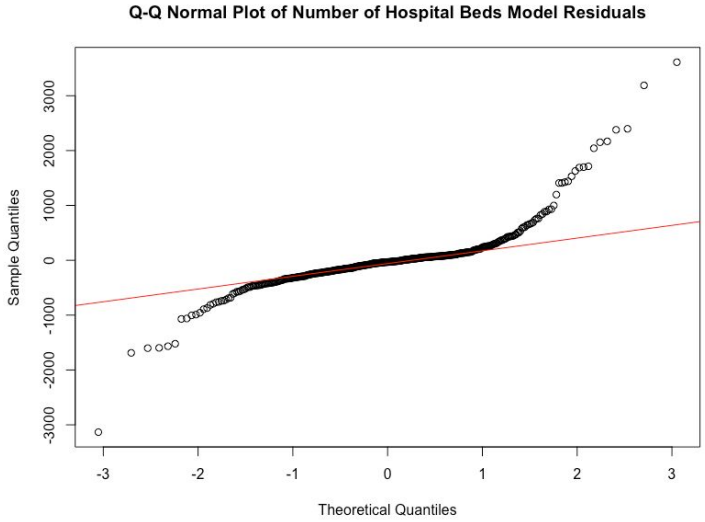


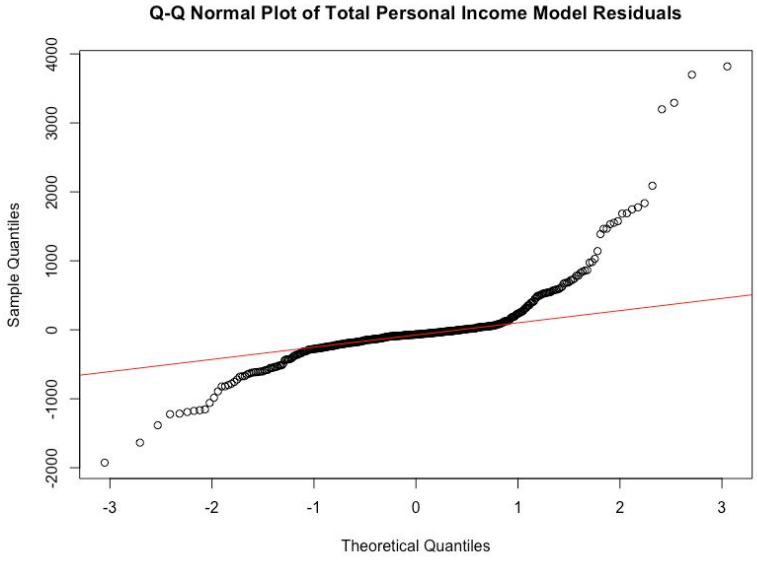




Q-Q normal plots:







After analyzing the scatterplots of the residuals for each model, it appears that each model suffers from heteroskedasticity—as the number of active physicians increases, the variances of the residuals also increase, which means the variances are not constant. Moreover, the residuals for the number of hospital beds and total personal income model seem to be a little skewed in the positive direction; whereas, the residuals in fit 2 is relatively symmetric about the zero line.

Through analyzing the Q-Q normal plots (normal probability plots), the residuals for each fit do not appear to have a standard normal distribution because of the heavy tails on both sides. It does, however, seem to have a normal distribution towards the center of the data.Additionally, it should be noted that the residuals with the most symmetric distribution is Fit 2. Fits 1 and 3 seem to have more extreme values in the upper quantiles.

While all three models suffer from heteroskedasticity and errors that are not normally distributed, the Number of Hospital Beds model is arguably the best because its residuals are the most even in positive and negative spread and most normally distributed of the three models. The appropriateness

of using this model is further reaffirmed when comparing the R^2 values of the three models. The Number of Hospital Beds model has the highest R^2 of each model at about 0.9, which means this model has the best goodness of fit of the three models.

V. Discussion

To explain the inference for the Percent of Bachelor's Degrees vs Per Capita Income simply, we noticed how each region has an interval that measures how close they are to getting the correct slope and within the range of those intervals, we see how some of those intervals overlap with each other and how some of them don't. Regions 3 and 4 overlap with each other since in their respective intervals (Region 3: 285.7076 to 375.5158; Region 4: 364.7585 to 515.8729) they both contain the range of 364.7585 to 375.5158. This means that since their intervals don't overlap, they won't have a very similar slope. Regions 1 and 2 don't overlap at all since they're not within the same range as each other (Region 1: 460.5177 to 583.80; Region 2: 193.4858 to 283.853). This would mean that it would be inconclusive to say whether or not their actual slope values are similar because it depends on where the location of their

perspective slope values lie in their intervals. When we look at the variation for all of the regions based on the Percent of Bachelor's Degrees vs Per Capita Income, we see that their values that determine how significant the relationship between the Percent of Bachelor’s Degrees vs Per Capita Income variables is, are greater than the chance that there isn’t a relationship between these factors in the data. This proves that there is a high possibility that there’s a strong linear relationship between the individuals that have a bachelor's degree and their per capita income in every region.

We use the plot of the residuals of each model against their predictor variable to examine how the residuals change as the value of the predictor variable changes. For each residual plot, we see that when the predictor variable’s value is small (close to zero), the spread of the residuals is small since they are close together. However, when the predictor variable’s value is large, the spread of the residuals becomes larger since they become more spread apart. Since the spread becomes larger as the predictor variable increases in value, the spread of spread of the residuals is not constant.

We use the Q-Q (Quantile-Quantile) normal plot to examine how the residuals are distributed. We expect most of the

residual values to be close to zero, and less and less residuals further away from zero. This sort of distribution looks like a mexican sombrero hat. When we plot the Q-Q normal graphs, we sequence the residuals in ascending order. We then calculate the quantiles in a theoretical distribution (in this case a normal distribution, which looks like a mexican sombrero hat) and we plot them against each other. If the residuals are also normally distributed like the theoretical distribution, we should see a straight line. Here, we see that in each model, the residuals are not normally distributed because we don't see a straight line.

VI. R Appendix

```
fit=lm(V8~V5, data=CDI)
summary(fit)
fit2=lm(V8~V16, data=CDI)
summary(fit2)
fit3=lm(V8~V9, data=CDI)
summary(fit3)

anova(fit)
anova(fit2)
anova(fit3)

plot(CDI$V5, CDI$V8, xlab = 'Total population', ylab = 'Number of
active physicians', main = 'Number of active physicians vs total
population')
abline(fit, col = 'red')
plot(CDI$V16, CDI$V8, xlab = 'Total personal income', ylab = 'Number
of active physicians', main = 'Number of active physicians vs total
personal income')
abline(fit2, col = 'red')
```

```
plot(CDI$V9, CDI$V8, xlab = 'Number of hospital beds', ylab = 'Number
of active physicians', main = 'Number of active physicians vs Number
of hospital beds')
abline(fit3, col = 'red')

summary(fit)$r.squared
summary(fit2)$r.squared
summary(fit3)$r.squared

confint(Region_1_Fit, level=0.9)
anova(Region_1_Fit)
confint(Region_2_Fit, level=0.9)
anova(Region_2_Fit)
confint(Region_3_Fit, level=0.9)
anova(Region_3_Fit)
confint(Region_4_Fit, level=0.9)
anova(Region_4_Fit)


Ian's code data
CDI = read.csv("/Users/Ian/Documents/STA 108/CDI.csv")
fit = lm(Number.of.active.physicians~Total.population, data = CDI)
summary(fit)
fit2 = lm(Number.of.active.physicians~Number.of.hospital.beds, data =
CDI)
summary(fit2)
fit3 = lm(Number.of.active.physicians~Total.personal.income,
data=CDI)
summary(fit3)

plot(CDI$Total.population, CDI$Number.of.active.physician, xlab =
'Total population', ylab = 'Number of active physicians', main =
'Plot of Total Population vs. Number of Active Physicians')
plot(CDI$Number.of.hospital.beds, CDI$Number.of.active.physician,
xlab = 'Number of hospital beds', ylab = 'Number of active
physicians', main = 'Plot of total population vs number of hospital
beds')
plot(CDI$Total.personal.income, CDI$Number.of.active.physician, xlab
= 'Total personal income', ylab = 'Number of active physicians', main
= 'Plot of total population vs Total personal income')

residualsfit1 = fit$residuals

residualsfit2 = fit2$residuals
residualsfit3 = fit3$residuals
```

```
plot(residualsfit1~CDI$Total.population, xlab = "Total population",
ylab = "Residuals", main = 'Plot of Total Population Model Residuals
against Total Population')
abline(h=0, col = 'red')
plot(residualsfit2~CDI$Number.of.hospital.beds, xlab = "Number of
hospital beds", ylab = 'Residuals', main = 'Plot of Number of
Hospital Beds Model Residuals against Number of Hospital Beds')
abline(h=0, col = 'red')
plot(residualsfit3~CDI$Total.personal.income, xlab = "Total personal
income", ylab = 'Residuals', main = 'Plot of Total Personal Income
Model Residuals against Total Personal Income')
abline(h=0, col = 'red')

qqnorm(residualsfit1, main = "Q-Q Normal Plot of Total Population
Model Residuals")
qqline(residualsfit1, col='red')
qqnorm(residualsfit2, main = "Q-Q Normal Plot of Number of Hospital
Beds Model Residuals")
qqline(residualsfit2, col='red')
qqnorm(residualsfit3, main = "Q-Q Normal Plot of Total Personal
Income Model Residuals")
qqline(residualsfit3, col='red')
```