Assignment 1. Statistics 3859A

Ian McLeod

October 5, 2017

Instructions:

- Complete this assignment using R/RStudio
- This document and its associated PDF is available from my webpage LINK HERE
- Compile your assignment as a PDF and upload to your Dropbox on OWL
- Due Date: October 23, 2017.

About tidyverse

Basically Tidyverse is a recent popular dialect of R that is becoming widely used in data science circles as for example in Kaggle postings. Tidyverse is documented the book *R for Data Science* by Grolemund & Wickham and the entire book is freely available online .

It is not the purpose of this course to teach much about tidyverse but I recommend it if you have time and interest to pursue it further. All basic coding that you will need for the assignments will be provided in the lectures and tutorials and you will just need to adapt it. To get started with this assignment, please review:

- 00TutorialSept12.Rmd
- 00LeadPollutionDataset.Rmd

that are available from my homepage

Further R scripts will be provided in upcoming tutorials and lectures.

In the Rmd file setup, I load the packages in tidyverse. These packages can be installed on your computer by entering the command **install.packages("tidyverse")**.

Overview of the Iowa Schools dataset

There are n=133 average test scores for schools in the K=6 largest cities. The test score offers a standardized measure of academic achievement. The purpose of the study is to investigate if there is a relationship between academic achievement, as measured by the test, and poverty. It is expected that students from economically disadvantaged backgrounds will do less well. Data on the average income in the school district was not available so a proxy variable for poverty was used. The percentage of students who received subsidized breakfasts and lunchs was available so this was used as the "Poverty" variable.

The dataset is available in CSV format on my webpage and can input to R using the function **read.csv()** as shown below. I use the tidverse function **glimpse** to verify and summarize the dataframe.

```
df <- read.csv("http://www.stats.uwo.ca/faculty/aim/2017/3859/data/Iowatest.csv")
glimpse(df)
```

```
## Observations: 133
## Variables: 4
## $ School  <fctr> Coralville, Hills, Hoover, Horn, Kirkwood, Lemme, Lin...
## $ Poverty <int> 20, 42, 10, 5, 34, 17, 3, 24, 21, 34, 24, 35, 4, 57, 2...
## $ Test    <int> 65, 35, 84, 83, 49, 69, 88, 63, 65, 58, 52, 61, 81, 43...
## $ City    <fctr> Iowa City, Iowa City, Iowa City, Iowa City, Iowa City...
```

Next a multi-panel visualzation of the data is presented that is comprised of scatterplots of **Test** vs **Poverty** for each **City**. Notice that the scales in each plot are *identical*. This means it is easy to make visual comparisons between the different panels. Multi-panel displays are an important and popular data visualization method in *Data Science* - much better, in most situations, than superimposing all panels together on one plot!

```
ggplot(df, mapping=aes(x=Poverty, y=Test)) +
 geom_point() +
 geom_smooth(se=FALSE, method="lm") +
 facet_wrap( ~City)
```

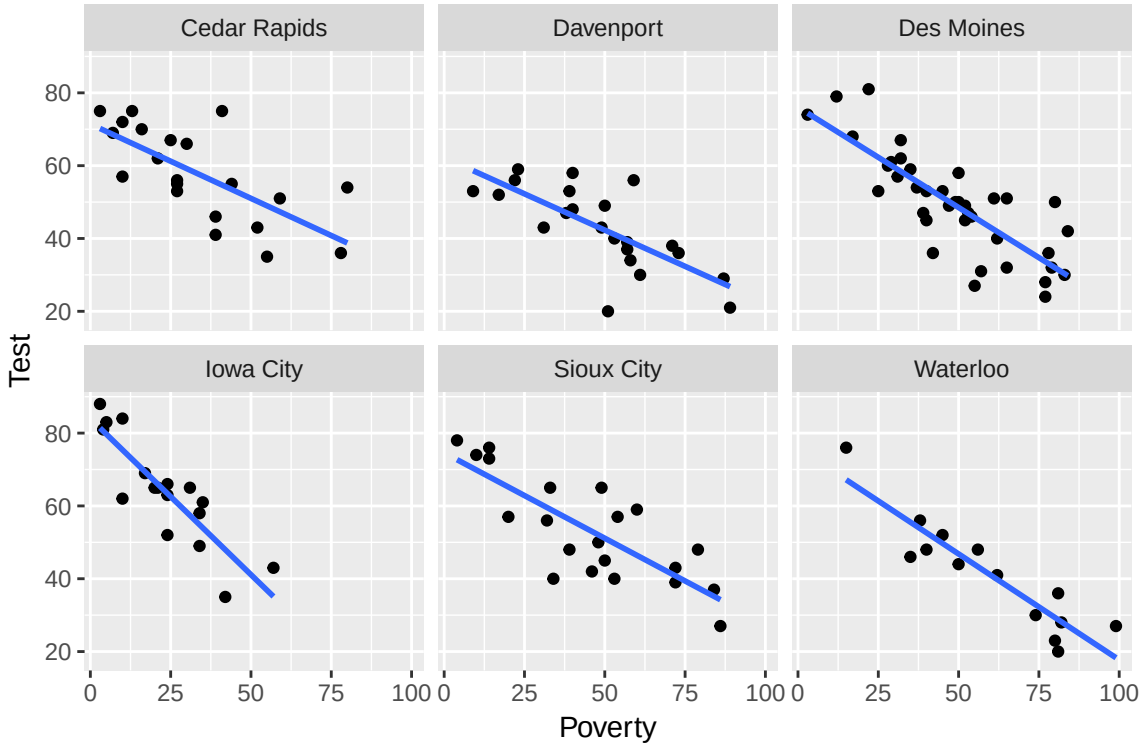

Figure 1: Iowa schools dataset with fitted least squares regression

Assignment Questions

Question 1.

Some of the output from an Excel regression analsis is shown below in Figure 2. Briefly explain what this output means and numerically verify it using suitable R functions discussed in class.

Question 2.

Fit a simple linear regression of **Test** on **Poverty** using the entire dataset. Comment of the fit. Perform the standard regression diagnostic checks and comment on any model indequacy. Also examine the residual dependency plots - you may use the R script presented in the SAT analysis script that is available on my webpage - see **SAT.Rmd** for script.

Question 3.

	A	B	C	D	E	F	G	H	I
1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0.8204564							
5	R Square	0.6731486							
6	Adjusted R Sc	0.6706536							
7	Standard Erro	8.7659537							
8	Observations	133							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	20731.47967	20731.48	269.7938	1.30856E-33			
13	Residual	131	10066.29477	76.84194					
14	Total	132	30797.77444						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
17	Intercept	74.605782	1.613250157	46.24564	5.26E-83	71.41438852	77.79718	71.4143885	77.7971758
18	Poverty	-0.5357754	0.032618712	-16.4254	1.31E-33	-0.600303021	-0.47125	-0.600303	-0.4712478

Figure 2: Simple Linear Regression Fit

It is thought that the slope parameter is the same for all Cities but the intercept parameter may differ. Using indicator variables, fit the model with a constant slope parameter but different intercept parameters for each city. Compare the diagnostic checks for this model with the previous model. Use the extra-sum-of-squares principle to perform a suitable test to determine if this model or the previous model is better.