Assignment 8

Gus Lipkin

QMB 3200 ~ Advanced Quantitative Methods

Florida Polytechnic University

Source: politico.com

Introduction and Background

Assignment 8 was created as a way for students in QMB 3200 Advanced Quantitative Methods to practice multiple linear regression. The goal of the assignment was to find the two best models for the annual average daily traffic on Minnesota roads based on county population, the number of and width of lanes, and other variables. Because some variables are categorical, students had to separate them into dummy variables so that they could be used in creating their regressions. Students also had to check their models for normality, homoscedasticity, and linearity.

Data

The data used was supplied by the professor of the class and is called "MN DOT.txt" The file includes all tab-spaced columns and no column headers. Column names were specified in "Data Dictionary.pdf" To improve ease of use, column headers were added in Excel and saved as "Assignment 8 Data.xlsx" The file included the target variable, annual average daily traffic (aadt), three continuous variables, and four discrete variables. The three continuous are the county population (county_pop), number of lanes (lanes), and roadway width (width). The three discrete described the roadway access (access), functional class (class), truck restriction conditions (restrictions), and locale (locale).

Table 1: Continuous Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
aadt	121	19437.69	30237.37	201	155547
county_pop	121	263427.7	329470	7716	941411
lanes	121	3.099174	1.300032	2	8
width	121	31.12397	11.53514	19	68
lanewidth	121	11.62355	6.165567	4.75	32

The mean annual average daily traffic was just under 19.5k with a standard deviation of over 30.2k. The mean county population is 263.4k with a minimum of 7.7k and maximum of 941.4k. There is a standard deviation of 329.4k. The mean number of lanes is 3 with a minimum of 2 and a maximum of 8 and standard deviation of 1.30. The mean roadway width by itself is not necessarily as helpful as the lane width because roadway width is a function of the total number of lanes. To derive lane width, we simply divide the total roadway width by the number of lanes. The mean lane width is 11.62 with a standard deviation of 6.16 and minimum and maximum of 4.75 and 32, respectively.

Table 2: Discrete Summary Statistics

access	Freq.	Percent	Cum.		locale	Freq.	Percent	Cum.
0	94	77.69	77.69		1	64	52.89	52.89
1	27	22.31	100		2	39	32.23	85.12
Total	121	100			3	18	14.88	100
					Total	121	100	

restrictions	Freq.	Percent	Cum.
1	32	26.45	26.45
2	20	16.53	42.98
3	1	0.83	43.8
4	7	5.79	49.59
5	61	50.41	100
Total	121	100	

class	Freq.	Percent	Cum.
1	8	6.61	6.61
2	56	46.28	52.89
3	18	14.88	67.77
4	39	32.23	100
Total	121	100	

Of the 121 roads measured 94 are not access controlled which is 77.69 of all roads. 52.89% of roads are rural, 32.23% in urban areas with a population less than or equal to 50k, and 14.88% in urban areas with populations greater than 50k. 26.45% of roads have no truck restrictions, 16.53% have tonnage restrictions, 0.83 have time of day restrictions, 5.79% have tonnage and time of day restrictions, and 50.41% of roads do not allow trucks. Of the four roadway classes, 6.61% are rural interstate, 46.28% are rural non-interstate, 14.88% are urban interstate, and 32.23% are urban non-interstate.

Analysis and Discussion

Because we are trying to predict annual average daily traffic, a value which cannot be negative, we must first take its natural log so that for the final regression there is no way for the value to be negative. The same is also done with the county population. To get a good idea of which variables have the largest effect on lnaadt, we first do a multiple linear regression with all variables.

Table 3: Multiple Linear Regression of lnaadt with All Variables

regress lnaadt county_pop lanes width access rural_road interstate tonnage time no_trucks rural small lnPop lanewidth						
Source	SS	df	MS	Number of obs	121	
[redacted]	[redacted]	[redacted]	[redacted]	F(13, 107)	50.12	
Model	260.430433	13	20.0331102	Prob > F	0	
Residual	42.772173	107	0.399739935	R-squared	0.8589	
[redacted]	[redacted]	[redacted]	[redacted]	Adj R-squared	0.8418	
Total	303.202606	120	2.52668838	Root MSE	0.63225	
lnaadt	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
county_pop	-2.25E-07	3.78E-07	-0.6	0.552	-9.74E-07	5.23E-07
lanes	0.4926556	0.1485172	3.32	0.001	0.1982376	0.7870736
width	-0.0106935	0.0164382	-0.65	0.517	-0.0432803	0.0218933
access	1.221561	0.6827565	1.79	0.076	-0.1319238	2.575047
rural_road	-1.251343	0.4768788	-2.62	0.01	-2.1967	-0.3059863
interstate	-0.820191	0.7172451	-1.14	0.255	-2.242046	0.6016636

tonnage	-0.1228003	0.2836676	-0.43	0.666	-0.6851382	0.4395375
time	0.0013389	0.2695632	0	0.996	-0.5330386	0.5357164
no_trucks	-0.8571357	0.2911573	-2.94	0.004	-1.434321	-0.2799502
rural	0.3829349	0.4779691	0.8	0.425	-0.5645832	1.330453
small	0.2517579	0.2039545	1.23	0.22	-0.1525582	0.656074
lnPop	0.4010341	0.0982634	4.08	0	0.2062383	0.5958299
lanewidth	0.0302466	0.0379479	0.8	0.427	-0.0449807	0.105474
_cons	3.304334	1.260722	2.62	0.01	0.8050999	5.803568

While the total R^2 value of .8589 is quite good, it is impractical to use thirteen variables to predict lnaadt. The goal is to minimize the number of variables used while maintaining a high R^2 value. We can eliminate any variables with a P>t value greater than 0.05 because we can conclude that they do not have a significant effect on the regression.

Table 4: First Reduced Multiple Linear Regression of lnaadt

Source	SS	df	MS	Number of obs	121	
[redacted]	[redacted]	[redacted]	[redacted]	F(4, 116)	158.26	
Model	2.56E+02	4	6.41E+01	Prob > F	0	
Residual	4.70E+01	116	4.05E-01	R-squared	0.8451	
[redacted]	[redacted]	[redacted]	[redacted]	Adj R-squared	0.8398	
Total	3.03E+02	120	2.53E+00	Root MSE	0.63623	
lnaadt	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]
lnPop	0.3312038	0.0578005	5.73	0	0.2167226	0.4456849
lanes	0.4851572	0.0635254	7.64	0	0.3593371	0.6109773
rural_road	-0.8880254	0.1713744	-5.18	0	-1.227454	-0.5485968
no_trucks	-0.8207183	0.1435158	-5.72	0	-1.104969	-0.5364671
_cons	4.368967	0.7319077	5.97	0	2.919331	5.818602

We now have an R^2 value of .8451 with only four variables, all of which have a P>t value less than 0.05. In an effort to further reduce the number of variables that we need to measure, we can drop the two variables with negative coefficients.

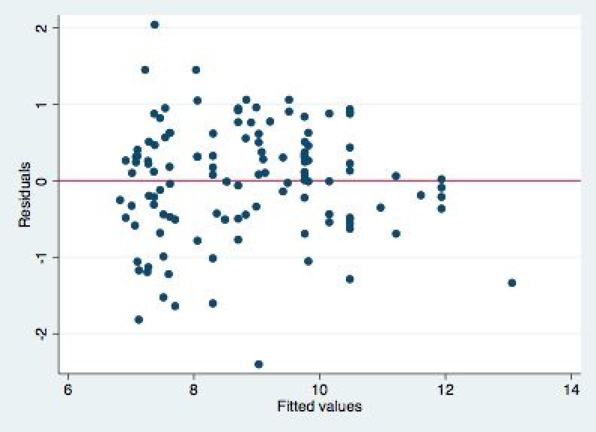
Table 5: Second Reduced Multiple Linear Regression of lnaadt

Source	SS	df	MS	Number of obs	121	
[redacted]	[redacted]	[redacted]	[redacted]	F(2, 118)	211.27	
Model	237.013419	2	118.506709	Prob > F	0	
Residual	66.189187	118	0.560925313	R-squared	0.7817	
[redacted]	[redacted]	[redacted]	[redacted]	Adj R-squared	0.778	
Total	303.202606	120	2.52668838	Root MSE	0.74895	
lnaadt	Coef.	Std. Err.	t	P>t	[95% Conf.	Interval]

lnPop	0.4590555	0.0529344	8.67	0	0.3542309	0.5638802
lanes	0.727169	0.0601381	12.09	0	0.6080793	0.8462587
_cons	1.258512	0.5502325	2.29	0.024	0.1689021	2.348122

An R^2 of 0.7817 is sufficient and requires only two variables instead of the four needed for the previous model. To make sure the model is accurate and not skewed and does not contain heteroskedasticity, we can plot the residuals vs fitted values and try and visually confirm that the data is not skewed.

Figure 1: Residuals vs Fitted Plot of Table 5



We can visually confirm that the data looks relatively evenly scattered across the line at y=0. We then move on to the residuals vs predictor plots with both lnPop and lanes.

Figure 2: Residuals vs Predictor (lnPop) Plot of Table 5

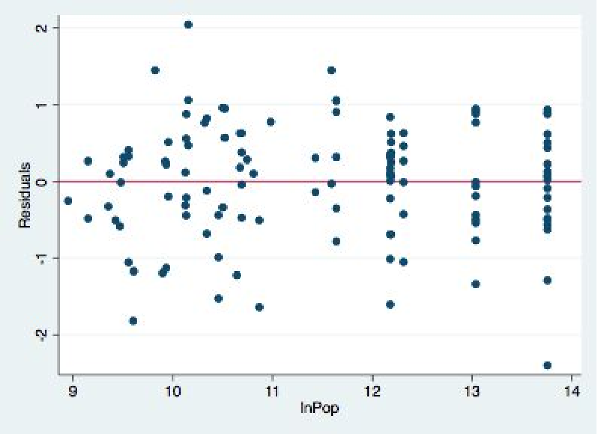
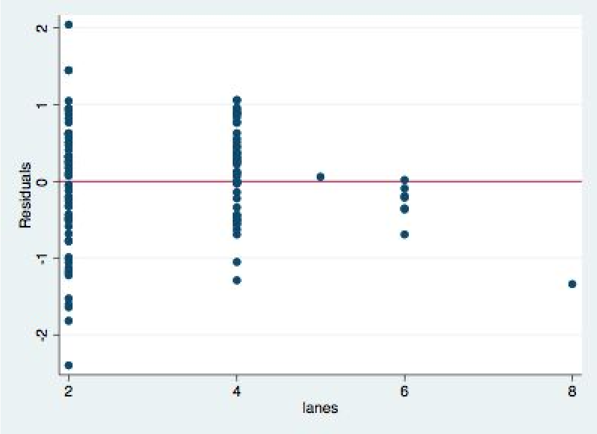


Figure 3: Residuals vs Predictor (lanes) Plot of Table 5



Both charts of the residuals look to be even about the line at y=0 so we move to the last step of verifying the model, conducting a White's test.

Table 6: White's Test of the Best Model

White's test for Ho: homoskedasticity	
against Ha: unrestricted heteroskedasticity	

chi2(5)	10.08		
Prob > chi2	0.0731		
Cameron & Trivedi's decomposition of IM-test			
Source	chi2	df	p
Heteroskedasticity	10.08	5	0.0731
Skewness	7.28	2	0.0263
Kurtosis	0.7	1	0.4031
Total	18.05	8	0.0208

With a total p-value of 0.0208, we can conclude that the model is of lnaadt vs lanes and lnPop meets all requirements and is, indeed, an adequate model.

Conclusion

Due to the limited nature of the data given, the assignment was relatively straightforward. Being able to predict annual average daily traffic is useful because it allows road designers to forecast the traffic that a road may get once built. The best model we produced took into account all the variables given, along with some new ones derived from the originals and had an R^2 of 0.8589. By reducing the size of the model to only those variables which had a significant effect, we were able to find a model with an R^2 value of 0.8451. Further reducing the model to only two variables allowed for a final R^2 of 0.7817. We then analyzed the final model and found that it fit the criteria for an accurate model. Further versions of the model may include what kinds of roadways the road being measured connects to or whether or whether or not the roadway has traffic signals.

Appendix

RMD File

~~~
---
title: "R Notebook"
output: html_notebook
---

```{r}
library(tidyverse)
library(readxl)
```

```{r}
df <- read_xlsx("../Assignment 8/Assignment 8 Data.xlsx")
```



```{r}
#re-code access so no access is 0 instead of 2
df$access[df$access == 2] <- 0
```

```{r}
#re-code roadway functional class
#rural == 1, urban == 0
#interstate == 0, no interstate == 0
df$rural_road <- ifelse(df$class == 1 | df$class == 2, 1, 0)
df$interstate <- ifelse(df$class == 1 | df$class == 3, 1, 0)
```

```{r}
#re-code restrictions
#tonnage == 1, no tonnage restrictions == 0
#time == 1, no time restrictions == 0
#notrucks == 1, no truck restrictions == 0
df$tonnage <- ifelse(df$restrictions == 2 | df$restrictions == 4, 1, 0)
df$time <- ifelse(df$restrictions == 3 | df$restrictions == 4, 1, 0)
df$no_trucks <- ifelse(df$restrictions == 5, 1, 0)
```

```{r}
#determine that there are no rural counties with large populations
df %>%
  filter(county_pop > 50000 && locale == 1) %>%
  nrow()
~~~

```
#re-code urban sizes
#rural == 1, urban == 0
#small == 1, >50,000 == 0
df$rural <- ifelse(df$locale == 1, 1, 0)
df$small <- ifelse(df$locale == 1 | df$locale == 2, 1, 0)
``` 

```{r}
dfDummy <- df %>%
  select(aadt, county_pop, lanes, width, access, rural_road, interstate, tonnage, 
time, no_trucks, rural, small)
``` 

```{r}
#log transforms
dfDummy$lnAADT <-  log(dfDummy$aadt)
dfDummy$lnPop <- log(dfDummy$county_pop)
``` 

```{r}
dfDummy$laneWidth <- dfDummy$width / dfDummy$lanes
dfDummy
``` 

```{r}
write_excel_csv(dfDummy, "../Assignment 8/Assignment 8 Data Mod.csv")
``` 
```

Stata do-file

```
capture log close
log using gus_lipkin_Assignment_6, replace
import delimited "/Users/guslipkin/Documents/Fall2020/QMB 3200 ~ Advanced 
Quantitative Methods/Assignment 8/Assignment 8 Data Mod.csv"
clear all

summ aadt county_pop lanes width
tabulate access
tabulate restrictions
tabulate locale

regress lnaadt county_pop lanes width access rural_road interstate tonnage time 
no_trucks rural small lnPop laneWidth
regress lnaadt lnPop lanes rural_road no_trucks
regress lnaadt lanes lnPop
```
```

```
rvfplot, yline(0)
rvpplot lnPop, yline(0)
rvpplot lanes, yline(0)
estat imtest, white

log close
```