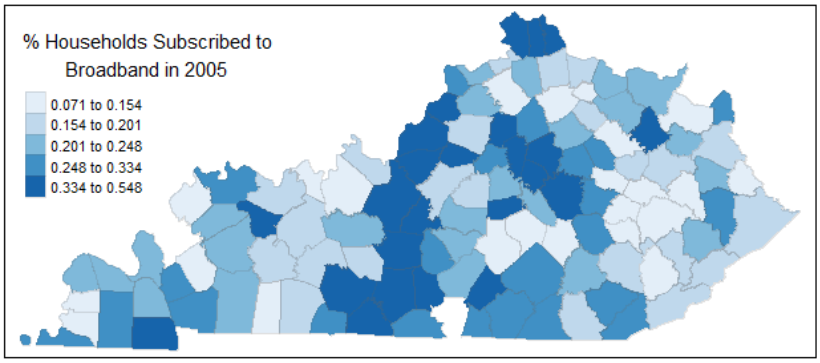
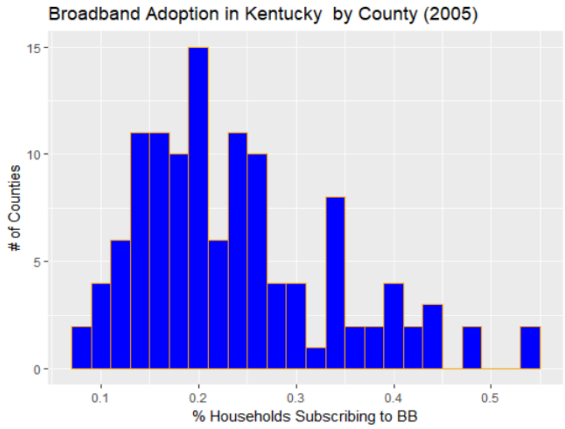
Leonardo A. Hughes
Broad Band Connection in Kentucky
Advanced Statistics in R

Abstract

The adoption of broadband internet provides increased access to information, higher quality goods, and creates measures in which we may further understand the actions of the consumer. This analysis seeks to identify which socioeconomic variables may contribute to the acquisition of higher quality internet providers in 2005 Kentucky. Employing linear regression on ad hoc data, this analysis seeks to identify the degree to which selected factors can predict the overall trends in broadband acquisition.





Measures

For this analysis, we will rely on four continuous variables in order to gage which factors contribute to broadband acquisition in Kentucky counties. The dependent variable will be the *adopt05* variable, which represents the percentage of the county population that acquired broadband internet in 2005. The independent variables in our model are the 2004 US census data for college degree holders and median age.

With bivariate and multivariate regression modeling, we will measure if any of our independent variables have a significant effect, or correlation, with the increasing trend of broadband internet users in Kentucky. A quick glance at the summary statistics for our variables shows the trends within each vector.

	adopt05	college	medage	unemp	geometry
nbr.val	1.170000e+02	1.170000e+02	1.170000e+02	1.170000e+02	4.692800e+05
nbr.null	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
nbr.na	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
min	7.100000e-02	4.100000e-02	3.930000e+01	9.000000e-03	3.348809e+06
max	5.480000e-01	4.720000e-01	4.930000e+01	9.200000e-02	6.018369e+06
range	4.770000e-01	4.310000e-01	1.000000e+01	8.300000e-02	2.669560e+06
sum	2.805600e+01	2.441400e+01	5.304800e+03	4.620000e+00	2.112939e+12
median	2.160000e-01	1.970000e-01	4.590000e+01	3.600000e-02	4.135951e+06
mean	2.397949e-01	2.086667e-01	4.534017e+01	3.948718e-02	4.502512e+06
SE.mean	9.445561e-03	7.558369e-03	2.102483e-01	1.868015e-03	1.201465e+03
CI.mean.0.95	1.870812e-02	1.497030e-02	4.164232e-01	3.699839e-03	2.354835e+03
var	1.043858e-02	6.684086e-03	5.171907e+00	4.082692e-04	6.774146e+11
std.dev	1.021694e-01	8.175626e-02	2.274183e+00	2.020567e-02	8.230520e+05
coef.var	4.260698e-01	3.918032e-01	5.015823e-02	5.117021e-01	1.827984e-01

In order to translate the R output into reader friendly terms, the last two integers tell user where to place the decimal point. For example, the mean value for college graduation can be interpreted as 20.8 rather than 2.08.

The summary statistics for college acquisition rates reveal that Kentucky's population does not attain high levels of education in an equally distributive manner. This may be consistent with the assumptions of the casual observer but it should be noted that the range of this variable is very high meaning several counties do in fact contain a very highly educated population, which can render outlier observations within our model.

The median age vector, in comparison to the college variable, is relatively uniform. For both variables, the standard error mean, standard deviation, and coefficient variance will be used within the regression models in order to calculate the slope values for the dependent and independent variables in testing for correlation.

Methods

This portion of the analysis will hone in on multiple regression, or the creation of a linear model using one or more predictive variables. The first model will introduce the adoption of broadband internet access as the dependent variable. This will remain the dependent variable for all the models as well. The x variable, otherwise known as the control, will be overall college acquisition by county. Lastly, a model incorporating the median age of Kentucky counties will be placed into the model in order to evaluate whether or not age has a correlational relationship with broadband adoption in the state.

The forthcoming regression output will provide statistics in how strongly our variables correlate. It is good practice to begin at the P value produced by the model in that this integer value represents whether or not a given independent variable provides more explanatory power than simply predicting dependent variable by its mean. Typically, P values represent the probability of achieving a value of T that if the independent variable has no effect on the dependent variable. The T value is the given by when the coefficient estimates are divided by the standard error.

The F statistic is produced when dividing the model sum of squares, or explained variance, by the residual sum of squares, or unexplained variance. It is the ratio that explains difference in means between the dependent and independent variable(s).

The Residual Standard Error is the average amount that observations deviate from a regression line. Inherent in all regression models exists an error in that independent variables cannot perfectly predict the dependent variable and we can therefore employ the residual standard error to quantify this deviation.

Results

Model 1: Broadband Adoption and College

The first model will test the assumption as to whether or not college attainment serves as a meaningful predictor to broadband access. The following regression output statistics are based on a bivariate case and therefore reading the output is somewhat straightforward.

First we should look at the probability of T. The T ratio is a value of the coefficient, or the estimated change in the independent variables per unit change in the dependent variable, divided by the standard error of the model. The Intercept value places broadband acquisition at 0.09% at the county level with a 0.7 unit increase per unit increase in college acquisition. It should be noted here that the college variable is a decimal percentage ranging from 0 to 1. The T test proves that the correlation between the dependent and independent variable exists within the threshold of statistical significance.

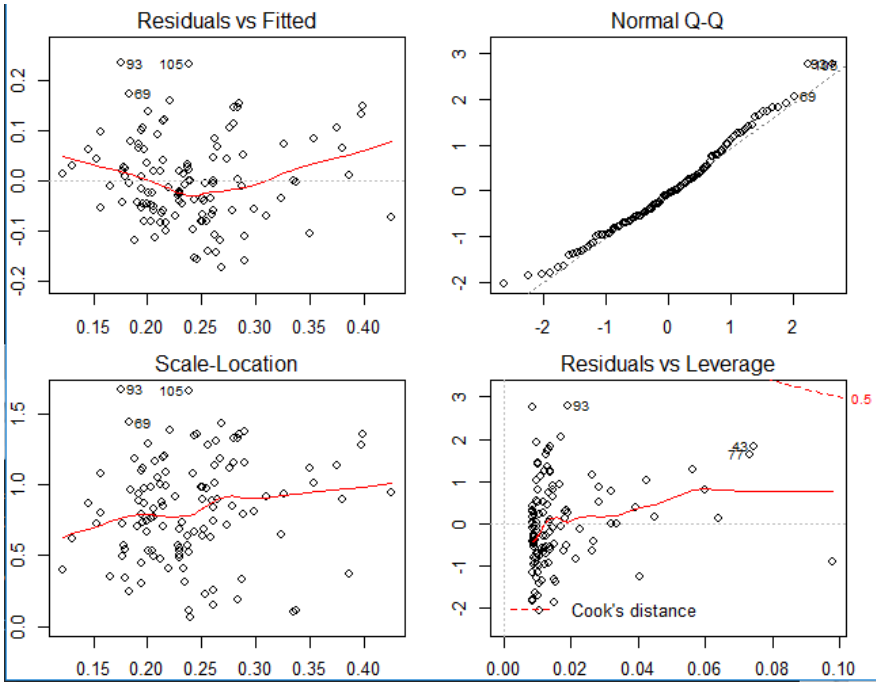
```
Call:
lm(formula = adopt05 ~ college, data = ky_corr_subset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.172227 -0.059473 -0.005473  0.052330  0.234118

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.09270    0.02155   4.301 3.58e-05 ***
college      0.70492    0.09622   7.326 3.53e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08473 on 115 degrees of freedom
Multiple R-squared:  0.3182,    Adjusted R-squared:  0.3123
F-statistic: 53.67 on 1 and 115 DF,  p-value: 3.526e-11
```

The following plots are the regression diagnostics for the linear regression model *adopt05* by college attainment. Residual plots place the residuals on the y axis while placing the predicted values onto the x. This test is performed in order to measure the difference between the observed values to the mean values predicted within the regression model.

In regression, the dependent variable, or the adoption of broadband internet in Kentucky, can be explained by the independent variable and the remaining error. Therefore, the mean value of broadband adoption is a function of the college graduation rates. In reading the diagnostic plots, it is important to note that in plotting the residuals, or errors, the observer should expect to see no discernible pattern in the error distributions. This would lend itself to the notion that there remains explanatory power in the error, and that in turn, linear regression may not be the most suitable model to use in gaging correlation between the two variables.

The Normal Q-Q, or Normal Quantile Quantile plot in regression diagnostics will display the residuals in a double quantile graph in order to check whether the error is normally distributed. Since the error provided through linear regression should be random, also referred to as heteroskedastic, using the scatter plot to display the error values should reveal an ascending line from the negative value quantiles to the positive quantiles (bottom left to top right). Should there be any values that deviate significantly from the residual mean then it is shows that it is less likely that these residual values came from normal distributions. In the Normal QQ plots provided below, we see that the residual points do in fact lie close to the normal distribution line despite there being several points at the tails that deviate.

These deviations from the mean are also referred to as outliers. In regression diagnostics, a common tool in testing whether or not certain observations have a proportionately larger effect on the model. Using Cook’s Distance, we observe the influence of the fitted values and whether or not these outliers may distort the output results of the regression analysis. In the bivariate case for Model 1, it appears that all observations lie within the parabola and we may deduce that there are no outlier error data points that distort the regression results.



Model 2: Broadband Adoption by College and Median Age

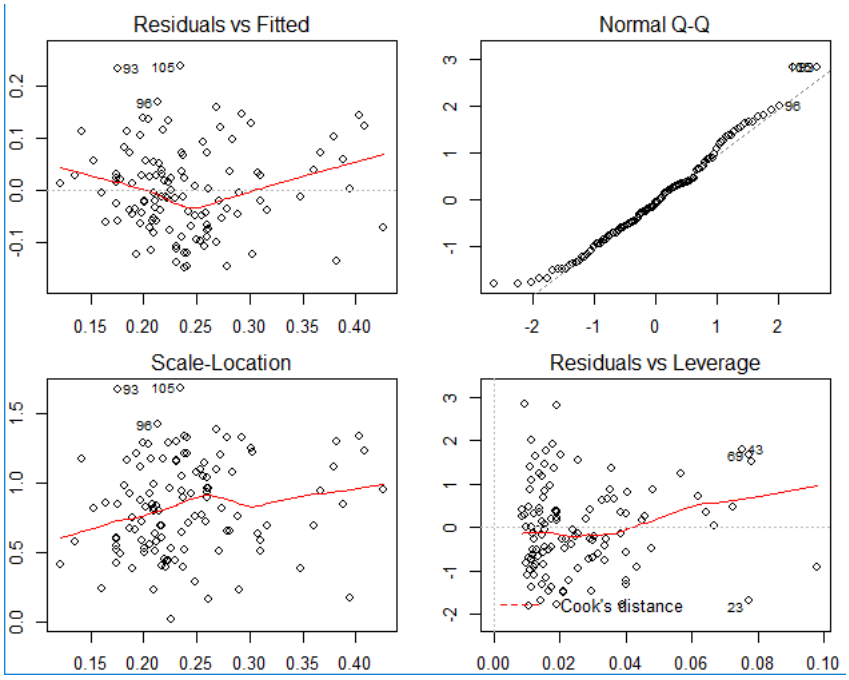
```
Call:
lm(formula = adopt05 ~ college + medage, data = ky_corr_subset)

Residuals:
      Min        1Q    Median        3Q       Max
-0.149017 -0.058523 -0.004803  0.051985  0.236858

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.427573   0.174304   2.453   0.0157 *
college      0.635859   0.101568   6.260 6.97e-09 ***
medage      -0.007068   0.003651  -1.936   0.0554 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08373 on 114 degrees of freedom
Multiple R-squared:  0.3399,    Adjusted R-squared:  0.3283
F-statistic: 29.35 on 2 and 114 DF,  p-value: 5.227e-11
```

The multivariate linear regression model employs college graduation and median age into a model which measures their respective correlation with the dependent variable. A key distinction in interpreting multiple regression output is that the coefficient value of the independent variables tells how much the dependent variable will change with a unit increase by the *independent* variable, controlling all other independent variables.

The coefficient values in multiple regression, similar to the bivariate case, show the size of the impact the independent variables have on the dependent variable. Including median age in the model has in fact increased the fitness of the model's overall predictive power on broadband acquisition, but using this variable by itself as a predictor for broadband adoption would be inappropriate in that it does not reach the required 95% certainty threshold, described by the P value and again tested by dividing the coefficient by its standard error.

This finding is rather surprising in that it reveals that the age of a population does not play a significant role in the usage of higher speed internet. In comparing the multivariate model to the bivariate regression model, we see only

an approximately 3% point increase in the R squared value, the statistic which measures the overall fitness of the model in predicting dependent variable. Despite their being a little added value to the predictive power of our model, running a regression diagnostics tests on our multiple regression model may show new dimensions to the types of residual errors that may be found in more complex models.



Both the Normal Quantile Quantile and Residuals Vs. Leverage diagnostics tests show there to be admissible degrees of influence in residual error deviation from the mean as well as an existence of outlier leverage, respectively. However, the Residual vs Fitted scatterplot shows that there is a degree of violation in the assumption of a linear correlation between the dependent variable and two independent variables in respect to the residual error produced in linear regression. The Scale-Location plot, like the Residuals vs. Fitted, should have points be seemingly randomly dispersed if the distribution of the residual error is normal, and bereft of any explanatory power in regression. By using the square root of the residuals verses fitted or predicted values, we see that heteroscedasticity is not violated and we can be sure that linear regression is an appropriate mechanism to gage the predictive power between the variables.

Conclusion

Linear Regression and regression diagnostics, applied to the case study of broadband internet acquisition in Kentucky in 2005, tests the linearity assumption that one observed demographic characteristic can be correlated to a second or more characteristic in a one-to-one relational expression.

The output of the regression model shows the predictive power of college attainment on broadband acquisition is statistically significant. The R squared values for the model ranged between 31% and 34% revealing that the slope of the regression line explained more of the variation in broadband internet usage than the null hypothesis would suggest, which posits its predictions based on the mean value of the dependent variable.

Further studies into the topic of broadband acquisition could rely on geographic variables such as dollars spent on infrastructure costs on the county level. A second variable, although potentially problematic due to the risk of multicollinearity, could be an adjusted variable for expendable household income. Such a measure would be able to assess the nature of highspeed internet as a collective good in regions of opaque market influences.

Appendix

```
#Kentucky BroadBand Regression Analysis


rm(list = ls())


install.packages("tmap")
install.packages("sf")
install.packages("tidyverse")



library(GISTools)
library(tmap)
library(sf)
library(tidyverse)

#Set the work space to the D Drive
setwd("D:/Adv Stat Urban Analysis/Multi_Regression")

#import an excel sheet
KY_bb <- read.csv("KY.bband.csv", header=TRUE)

#importing a shapefile into an R spatial dataframe
KY_counties <- st_read("D:/Adv Stat Urban Analysis/Multi_Regression/ky_counties.shp")

#should connect the broadband info from the excel sheet into the ky_counties spatial dataframe
KY_df <- inner_join(Ky_counties, KY_bb, by=c("NAMELSAD10"="ck_name"))

#lets make a histogram
kybb_hist <- ggplot(KY_bb, aes(x=adopt05)) + geom_histogram(binwidth=.02,
                                                    color="orange",
                                                    fill="blue") +
```

```
                                        labs(title="Broadband Adoption in Kentucky  by County (2005)",
                                        x="% Households Subscribing to BB",
                                        y="# of Counties")
kybb_hist

#lets make a box plot
kybb_box <- ggplot(KY_bb, aes(x = "", y = adopt05)) + geom_boxplot(color="orange",
                                                    fill="blue") +
                                                   labs(title="Boxplot of Broadband Adoption in Kentucky by County
(2005)",
                                                      y="% Households Subscribing to Broadband")
kybb_box

#Create the choropleth map based on
tm_shape(KY_df) + tm_borders(alpha = .2) + tm_polygons(col = 'adopt05', palette = 'Blues',
                                    title =  '% Households Subscribed to
      Broadband in 2005',
                                    style = 'quantile')

#descriptives

str(KY_df)
head(KY_df)
names(KY_df)

names(KY_bb)

stat.desc(KY_df[c("adopt05","college","medage","unemp")])

#does it correlate?

ky_corr_subset <- as.data.frame(KY_df)
```

```
summary(ky_corr_subset)

#Scatterplot matrices
pairs(~adopt05+college+medage, data=ky_corr_subset, main="Scatterplot Matrix of Potential Factors for
    BroadBand Adoption in KY")

#fit models
mod1 <- lm(adopt05~college, data = ky_corr_subset)
mod2 <- lm(adopt05~college+medage, data = ky_corr_subset)
mod3 <- lm(adopt05~unemp, data = ky_corr_subset)



#view results
summary(mod1)
summary(mod2)
summary(mod3)

#regression diagnostics
par(mfrow=c(2,2))
par(mar=c(2,2,2,2))
plot(mod1)
plot(mod2)
plot(mod3)
par(mfrow(c(1,1)))
```