**The Impact of economic indicators and broadband availability**

**on the digital divide**

**Executive Summary**

The internet represents a fundamental change in the way Americans connect with each other, gather information, and conduct everyday life during the COVID-19 pandemic. A study has been conducted to evaluate the factors influencing on the digital divide. A tool was developed using multiple linear regression model to analyze the characteristics of the households without broadband internet based on economic indicators and broadband availability. To construct the model, a training set was used with a .80 proportion of the data and a hold-out test set of .20 of the proportion was used to generate prediction and validate the model. The final model was selected after stepwise variable selection process using Akaike Information Criterion (AIC), Mean Squared Error (MSE), and adjusted R-squared criteria. Diagnostic measurements were also built to check the quality of the model. The final model’s adjusted R-squared is .646 and the MSE is 0.045. Variables in the final model include population, no health insurance, poverty, SNAP (households receiving food stamps), and broadband availability. Poverty, SNAP, and no health insurance are positively related to the proportion of the households without internet. On the other hand, population and broadband availability are negatively related to the proportion of the households without internet. The proportion of individuals below the poverty line has the greatest impact on the rate of households without internet. The final model shows that a 1-unit increase in poverty increases the proportion of households without internet by 1.8% if all other variables are held the same. With a 1-unit increase in the poverty, the proportion of the households without internet with 95% confidence would increase from 1.5% to 2.1%. People with low-income levels are less likely to get broadband service at home. Knowing the relationship between economic inequality and digital divide would help governments determine internet access programs for low-income people.

**Introduction**

Thirty years after the debut of the World Wide Web, internet usage has grown rapidly for all Americans. Today, internet usage is ubiquitous. While many aspects of the digital gap have narrowed over time, the broadband adoption gap for homes still remains. The disparity in online access is apparent in the gap between school-age children who have access to high-speed internet at home and those who do not. In 2015, 35% of lower-income households with school-age children did not have a broadband internet connection at home [1]. It is important to understand broadband internet access for the US population especially in this epidemic situation.

The study was conducted using data from 3000 observations. The dataset contains the households without broadband internet, population, four economic indicators like unemployment rates, percent without health insurance, poverty, and percent who received SNAP and three indicators of broadband availability such as the number of broadband internet service providers, broadband cost, broadband availability, and the number of internet service providers with 25Mbps. Many questions have been proposed in favor of these issues including 1) What variables increase or decrease the proportion of households without internet? 2) How accurate is the model proposed in this analysis? 3) How can we implement the model? and 4) How can we improve the final model? This analysis answers questions by exploring the dataset shown in Table 1. To conclude the findings, this analysis also develops a model that uses multiple linear regression to help Americans understand the impact of economic indicators and broadband availability on the digital divide. For variable selection, AIC, MSE, and adjusted R-squared are used for the model selection. The multicollinearity issue is considered by checking Variable Inflation Factor (VIF). Any values greater than VIF 10 are considered as a strong multicollinearity effect. Influential points and outliers are identified to build a more accurate model. The whole data analysis and reporting is based on R version 3.6.3.

Table 1: Description of response variable (no\_internet) and predictors

|  |  |
| --- | --- |
| **Variable** | **Definition** |
| **No\_Internet** | Households without internet |
| **Population** | Population |
| **Unemp** | Unemployment |
| **Health\_ins** | Individuals without health insurance |
| **Poverty** | Individuals below the poverty line |
| **SNAP** | Households receiving food stamps |
| **Broad\_num** | Number of broadband internet service providers |
| **Broad\_avail** | Broadband internet available households |
| **Broad\_cost** | Broadband cost in month |
| **All25\_bbn** | Number of internet service providers with 25Mbps |

**Exploratory Data Analysis**

Before completing any advanced analysis, a look into the exploratory data analysis will help better describe and understand the dataset structure. During the exploratory analysis process, it is important to access all the significant relationships of each variable with the response variable. The dataset contains a total of 3000 observations with no missing values. A descriptive statistic table such as minimum, lower IQR, median, mean, upper IQR, and maximum for each variable included in the study are shown in Table A1 in the Appendix. One of the assumptions of a linear regression model is that the response variable needs to be normally distributed. To test the normality of the response variable, the Shapiro-Wilk test was applied to check the null hypothesis: the data are normally distributed. The response variable is not normally distributed since p-value is less than 2.2e-16. Graphically, the response variable is right skewed shown in Figure A1 in the Appendix. To investigate an appropriate transformation, a Box-Cox transformation method was used, and a log transformation of the response variable was proposed. Figure 1 depicts the normality of the response variable after log transformation which implies that the assumption is no longer violated. Figure A1 shows that the population is heavily right skewed. Normally distributed explanatory variables are not an assumption of a linear regression, however, using approximately normal distributed predictors can help strength of the model. Due to this fact, a log transformation was applied to the population to ensure a linear relationship with response variable. Histogram and QQplot were used to check the distribution of all variables and scatter plot was utilized to check the relationship between predictors and the response variable.

|  |  |
| --- | --- |
|  |  |

Figure 1: Histogram of Log transformed dependent variable (the proportion of the households without internet) and Log transformed independent variable population.

A correlation plot was created to identify the correlations between predictors and between predictors and response variable. Looking at the correlation plot in Figure 2 and Figure A2, poverty and SNAP have a high correlation of 0.82. High correlation between predictors means that there might be multicollinearity issues when including those variables in the model. This leads the model to have unstable and unreliable coefficients. For further investigation on the multicollinearity issue, VIF was checked to see if any variable with high multicollinearity should be removed from the model. The correlation plot also suggests that poverty may be one of the most influential predictors in the regression model and economic indicators have more impact on the model than broadband factors. Moreover, unemployment, poverty, no health insurance, and SNAP are positively correlated with the response variable while population, broadband number, broadband availability, broadband cost, and number of internet service providers with 25Mbps are negatively correlated with the response variable.

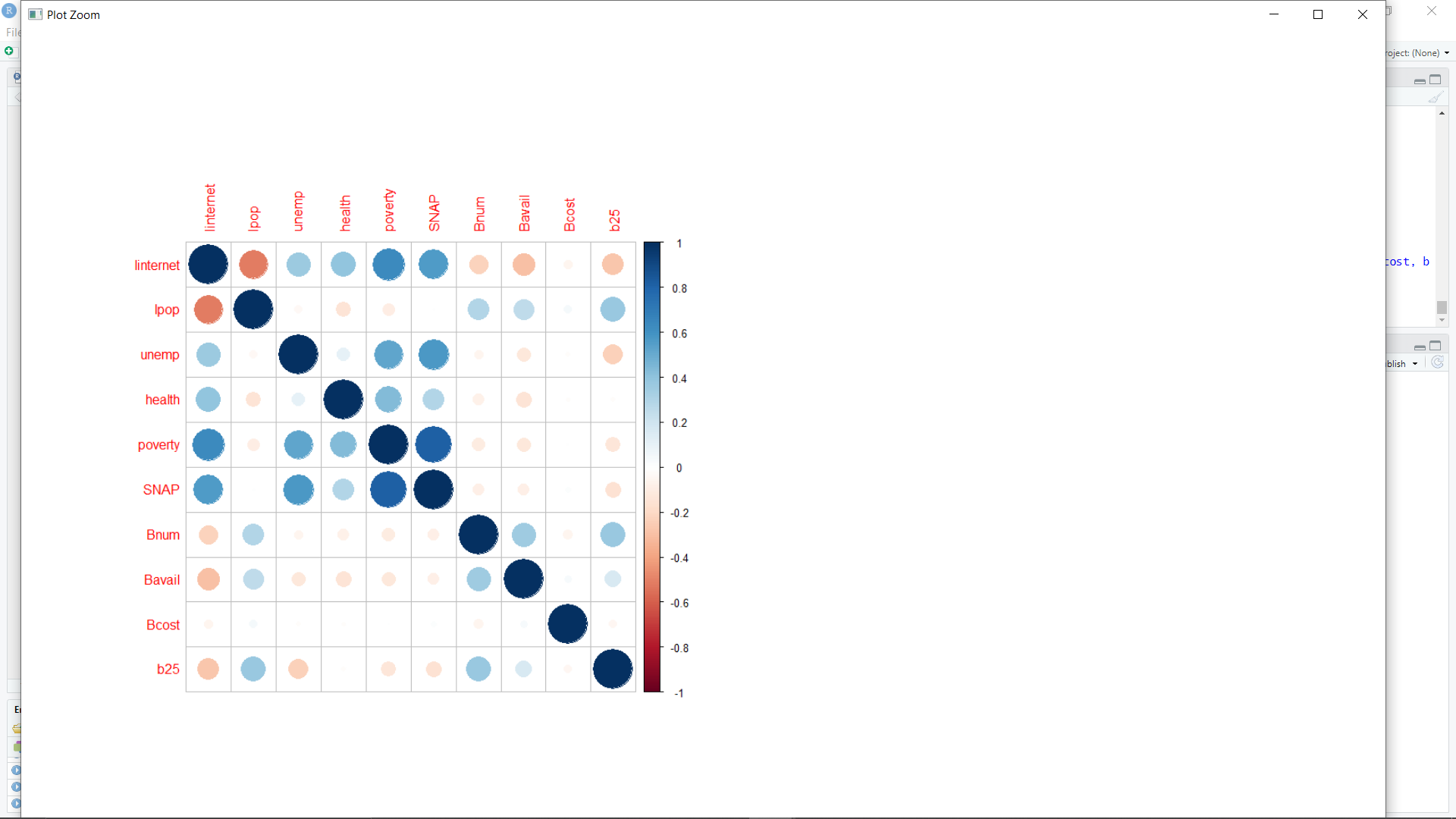


Figure 2: Correlation matrix plot with variables of linternet, lpop, unemp, health, poverty, SNAP, Bnum, Bavail, Bcost, b25.

The scatter plots of each variable against the log transformed response variable were examined to build the linear regression model. The scatter plot of logged population and broadband cost in Figure A3 appears that it may be quadratic, but if we cut the data in a cutoff point, it appears that we have two linear relationships. For logged population, there appears to be a cutoff point at approximately 9.8 and for broadband cost, there appears to have a cutoff point of 130. When a piecewise linear function of both variables is applied to the analysis, out of 3000 observations, only 1882 observations would be used in the analysis. As a result, the piecewise method was not applied to the analysis in order to have a good predictive power with a sufficient data in the model.

**Statistical Analyses**

Model selection is a crucial step in a data analysis. As mentioned earlier, a train set of .80 of the proportion of the data was used to train the model and a test set of .20 of the proportion of the data was used to test the model. After exploring the relationship between each variable and the response variable, a full model with predictors and previously detailed transformed predictor was built with multivariate linear regression. The full model includes log transformed population, unemployment, no health insurance, poverty, SNAP, broadband cost, broadband availability, broadband cost, and the number of internet service providers with 25Mbps. To select variables for the final model, a stepwise variable selection process was run using AIC, MSE, and adjusted R-squared. In this case, the ideal best model has smaller AIC and MSE, and higher adjusted R-squared. After a stepwise function is implemented, the results propose a model that eliminates unemployment and broadband number, broadband 25Mbps. In the stepwise selected model, the coefficient of broadband was almost zero, so the broadband cost was also removed. The adjusted model includes log transformed population, no health insurance, poverty, SNAP, broadband availability shown in Table 2. After building the three models, a model with interaction terms was also constructed using all combinations of variables from the previous model. A stepwise selection method was also applied to the interaction model to select the best variables using AIC. Looking at Table 2, adjusted stepwise model with interaction terms I and adjusted stepwise model with interaction terms II have slightly lower MSE, lower AIC, and higher adjusted R-squared than previous models. A large F-statistic corresponds to a statistically significant p-value. All F-statistic’s p-values are less than 2.2e-16. It means that all models are pretty good. However, the adjusted stepwise model with interaction terms II has too many predictors for the model to handle, and the coefficient without health insurance contradicts the EDA results. It was expected to be positive in EDA, but it was negative in the adjusted stepwise model with interaction terms II. Therefore, the adjusted stepwise model was chosen as the final model shown in Table 2. The final model includes five predictors such as population, no health insurance, poverty, SNAP, and broadband availability. The VIF test was implemented to check if the model has multicollinearity issues presented. All VIF values were less than 4, indicating no multicollinearity issues in the final model.

K-fold cross-validation method was used to evaluate the model performance on different subsets of the training data. The average prediction error rate in cross-validation was 0.046. In addition, MSE was 0.045 and error rate was 0.066 in the train set. The final model performs very good on the training dataset. A .20 proportion of data was used as a test set for model validation. After generating the prediction, the MSE was 0.047 and error rate was 0.068 in the test set. This indicates that the model is predicting well.

Table 2: Regression validation metrics including MSE, adjusted R-squared, F. statistics,

and AIC

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **MSE** | **Adj. R.** | **F. statistics** | **AIC** |
| **Full model**  (linternet~lpop+unemp+health+poverty+  SNAP+Bnum+Bavail+Bcost+b25) | 0.045 | 0.647 | 491 | -605 |
| **Stepwise model**  (linternet~lpop+health+poverty+SNAP+  Bavail+Bcost) | 0.045 | 0.648 | 736 | -609 |
| **Adjusted stepwise model**  (linternet~lpop+health+poverty+SNAP+  Bavail) | 0.045 | 0.646 | 879 | -602 |
| **Adjusted stepwise model**  **with interaction terms I**  (linternet ~ lpop+health+poverty+SNAP+  Bavail+lpop:health+lpop:poverty+  lpop:SNAP+lpop:Bavail+health:poverty+  health:SNAP+health:Bavail+poverty:SNAP+  poverty:Bavail) | 0.040 | 0.685 | 436 | -871 |
| **Adjusted stepwise model**  **with interaction terms II**  (linternet ~ lpop + health + poverty + SNAP +  Bavail + lpop:health + lpop:poverty +  lpop:SNAP + health:poverty + health:SNAP +  health:Bavail + poverty:SNAP) | 0.040 | 0.685 | 436 | -871 |

Table 3: Summary statistics table of the final model

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Estimate** | | **Std. Error** | | **P-value** | **95% C.I.** |
| **Intercept** | 3.9041 | 0.0359 | | 0.0000 | | ( 3.833, 3.975) |
| **Log population** | -0.1073 | 0.0031 | | 0.0000 | | (-0.113, -0.101) |
| **No health insurance** | 0.0071 | 0.0011 | | 0.0000 | | ( 0.005, 0.009) |
| **Poverty** | 0.0175 | 0.0013 | | 0.0000 | | ( 0.015, 0.021) |
| **SNAP** | 0.0157 | 0.0012 | | 0.0000 | | ( 0.013, 0.018) |
| **Bavail** | -0.0018 | 0.0002 | | 0.0000 | | (-0.002, -0.001) |

Table 3 presents the summary statistics for the final model. Table 2 suggests MSE and adjusted R-squared of the final model. The final model has an adjusted R-squared of 0.65 indicating that 65% of variation in the outcome variable can be explained by the model predator variables such as log transformed population, no health insurance, poverty, SNAP, and broadband availability. The final model shows that if the population increases by 10%, the households without internet would decrease by 1%. With a 10% increase in the population, the households without internet with 95% confidence would decrease from 1% to 1.1%. On the other hand, the rate of the households without internet increases 1.8% when poverty increases one unit. With a 1-unit increase in the poverty, the rate of households without internet with 95% confidence would increase from 1.5% to 2.1%. The rate of the households without internet increases 1.6% when SNAP increases one unit. With a 1-unit increase in SNAP the rate of households without internet with 95% confidence would increase from 1.3% to 1.8%. However, the impact of broadband availability on the households without internet is minimal. Two of the biggest impacts on the household broadband internet subscriptions are poverty and SNAP.

After selecting the final model, it is important to evaluate the robustness of the model. The assumption of multivariate linear regression is that random error of the model follows independent and identical normal distribution with zero mean and constant variance. The assumption was confirmed by studentized residual plot and normal Q-Q plot. As can be seen in Figure 3, there is no distinct pattern in the studentized residuals, and most observations are in -2.5 to 2.5 range. This means that the model shows a good linear relationship. The normal Q-Q plot indicates that the residuals are normally distributed even with slight deviations in the tail.

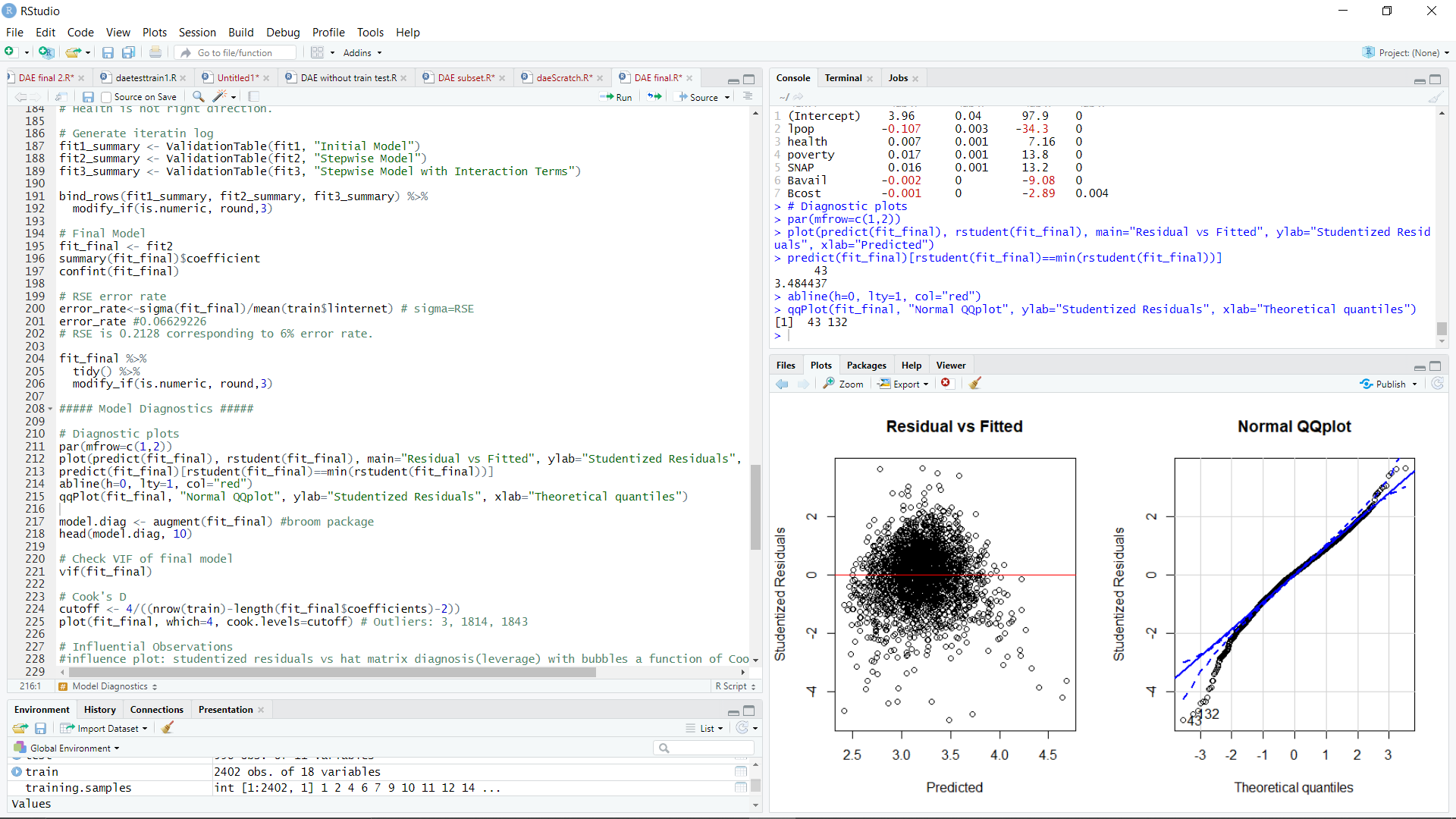


Figure 3: Residual vs Fitted plot and Normal QQ plot

Figure 4 shows that there are some observations in the cook’s distance and influence plot that did not make the cutoff point including observation 3, 10, 1814, and 1843. Observation 3, Chambers County in Alabama, has the highest broadband cost ($299.95). Observation 10, Dekalb County in Alabama, also has high broadband cost ($187.47). The broadband cost of observation 3 and 10 make sense because the broadband installation cost is expensive due to the local nature. Observation 1814, Buffalo County in South Dakota, has the highest rate of no health insurance. It is also high in poverty and SNAP. Observation 1843, Oglala Lakota County in South Dakota, has the highest SNAP rate. It is also high in poverty, and no health insurance rate. Therefore, those observations were not removed from the data because there was not good reason to remove them and those data reflects the characteristics of the region well.

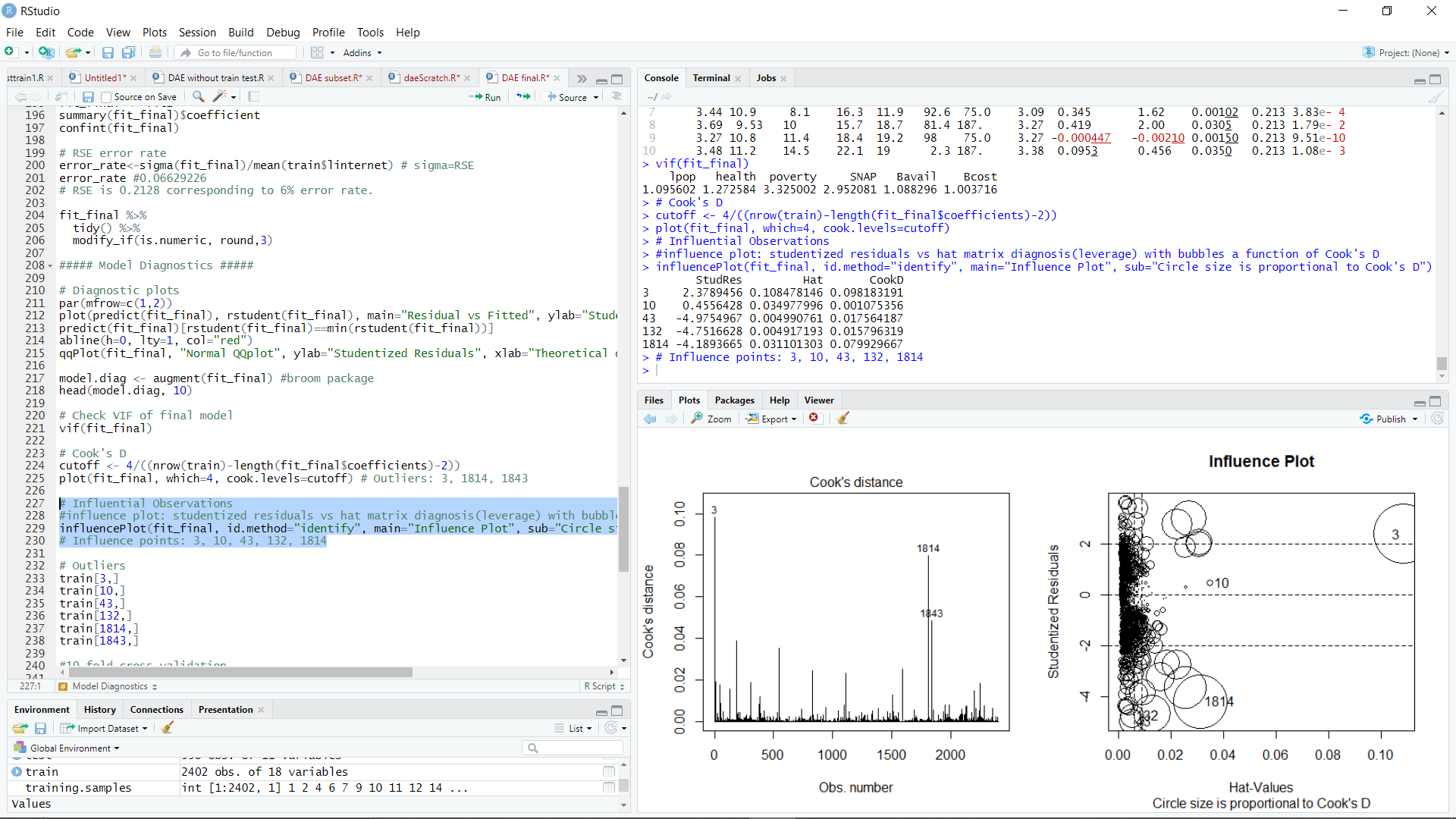


Figure 4: Cook’s distance plot and influence plot

**Conclusion**

The goal of this analysis is to understand the relationship between the household without broadband internet and some of the factors suggested by the dataset such as population, unemployment, individuals without health insurance, individuals below poverty line, households receiving food stamps, number of broadband internet service providers, broadband internet available households, broadband cost in month, and number of internet service providers with 25Mbps. A multivariate linear regression model was applied to estimate the relationship. The final model was built with adjusted MSE of 0.045, R-squared of .65 and AIC of -602 using 5 predictors including log transformed population, individuals without health insurance, poverty, SNAP, and broadband availability. As a result, the two most important factors in the outcome variable were the proportion of individuals below poverty line, and the second important factor was the proportion of the households receiving food stamps. This study shows that households with low-income families have low rates of broadband internet service.

Further studies can investigate regression models with nonlinear relationships between predictors and dependent variables. This can produce a more accurate regression model, but it can be more difficult to interpret. Moreover, as the proportion of Americans who use their smartphones as their primary online access at home increases, further studies may include individuals who only use smartphones to investigate the dependence of home broadband and smartphone at various income levels. Today, roughly 1 in 5 US adults are “smartphone-only” internet users, but they do not have traditional home broadband service. It can be an interesting topic for further study.

**References**

Monica Anderson, Madhumitha Kumar, Digital divide persists even as lower income Americans make gains in tech adoption, Pew Research Center, 2019. [1]

Alboukadel Kassambara, Practical guide to principal component methods in R, 2017.

Michael H. Kutner, Christopher J. Nachtsheim, John Neter, William Li, Applied Linear Statistical Models, McGraw-Hill Irwin, 2004.

**Appendix A**

TableA1. Summary statistics table of all variables

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Min** | **Lower IQR** | **Median** | **Mean** | **Upper IQR** | **Max** |
| **No\_internet** | 4.20 | 20.30 | 25.30 | 26.31 | 31.40 | 74.10 |
| **Population** | 272 | 12072 | 27296 | 108765 | 71980 | 10039107 |
| **Unemp** | 1.40 | 3.00 | 3.70 | 3.986 | 4.600 | 18.300 |
| **Health\_ins** | 1.70 | 6.20 | 9.10 | 9.915 | 12.500 | 42.400 |
| **Poverty** | 2.30 | 11.00 | 14.70 | 15.52 | 19.00 | 49.70 |
| **SNAP** | 0.00 | 8.68 | 12.40 | 13.20 | 16.70 | 53.50 |
| **Broad\_num** | 1.00 | 2.75 | 4.00 | 4.21 | 6.00 | 15.00 |
| **Broad\_avail** | 0.10 | 63.30 | 81.45 | 75.01 | 93.80 | 100.00 |
| **Broad\_cost** | 25.00 | 59.99 | 65.15 | 67.35 | 70.74 | 299.95 |
| **All25\_bbn** | 0.55 | 4.73 | 5.89 | 6.16 | 7.44 | 14.79 |

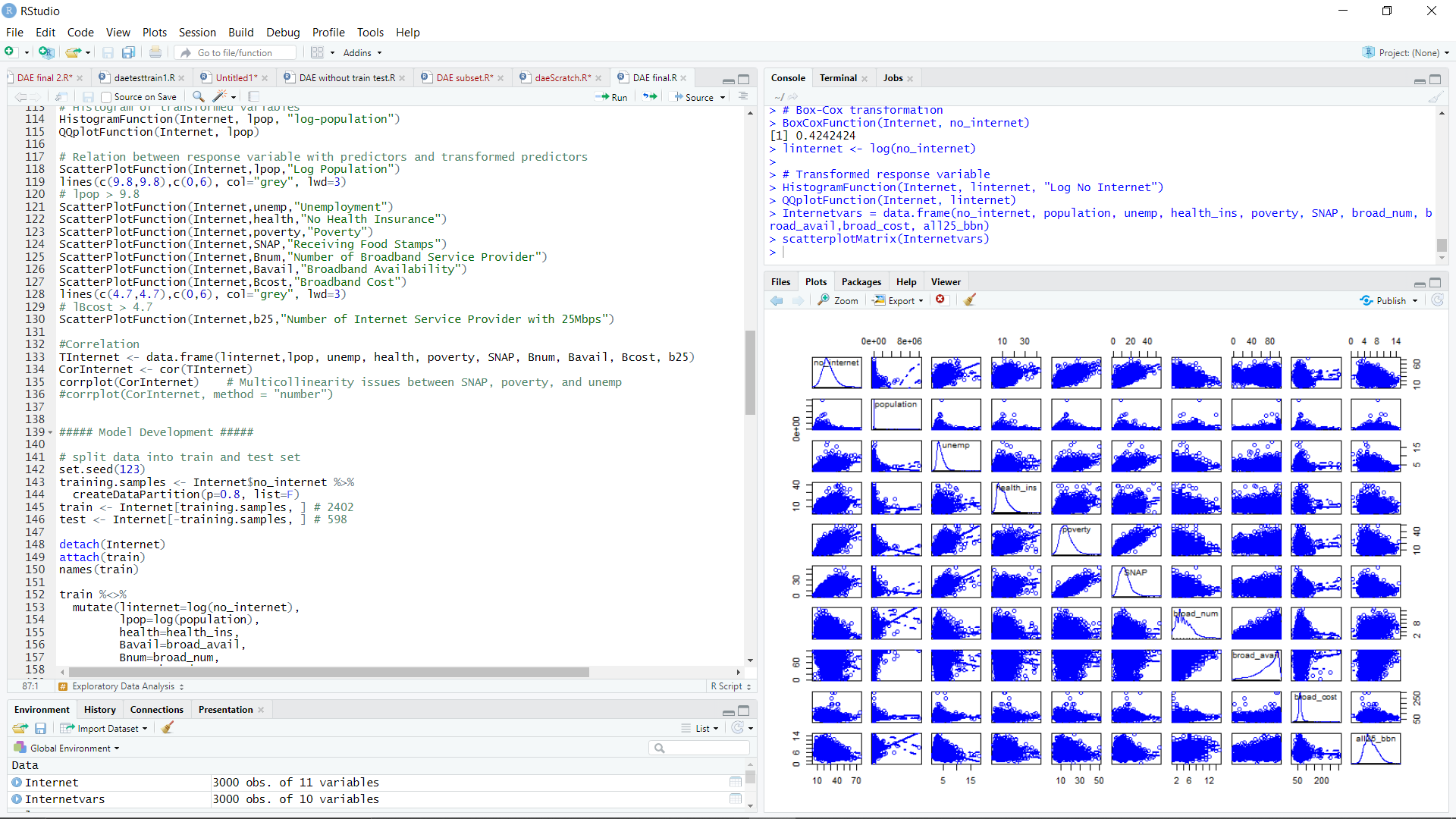


Figure A1: Scatterplot matrix shows all scatter plots of variables (no\_internet, population, unemp, health\_ins, poverty, SNAP, broad\_num, broad\_avail,broad\_cost, all25\_bbn)

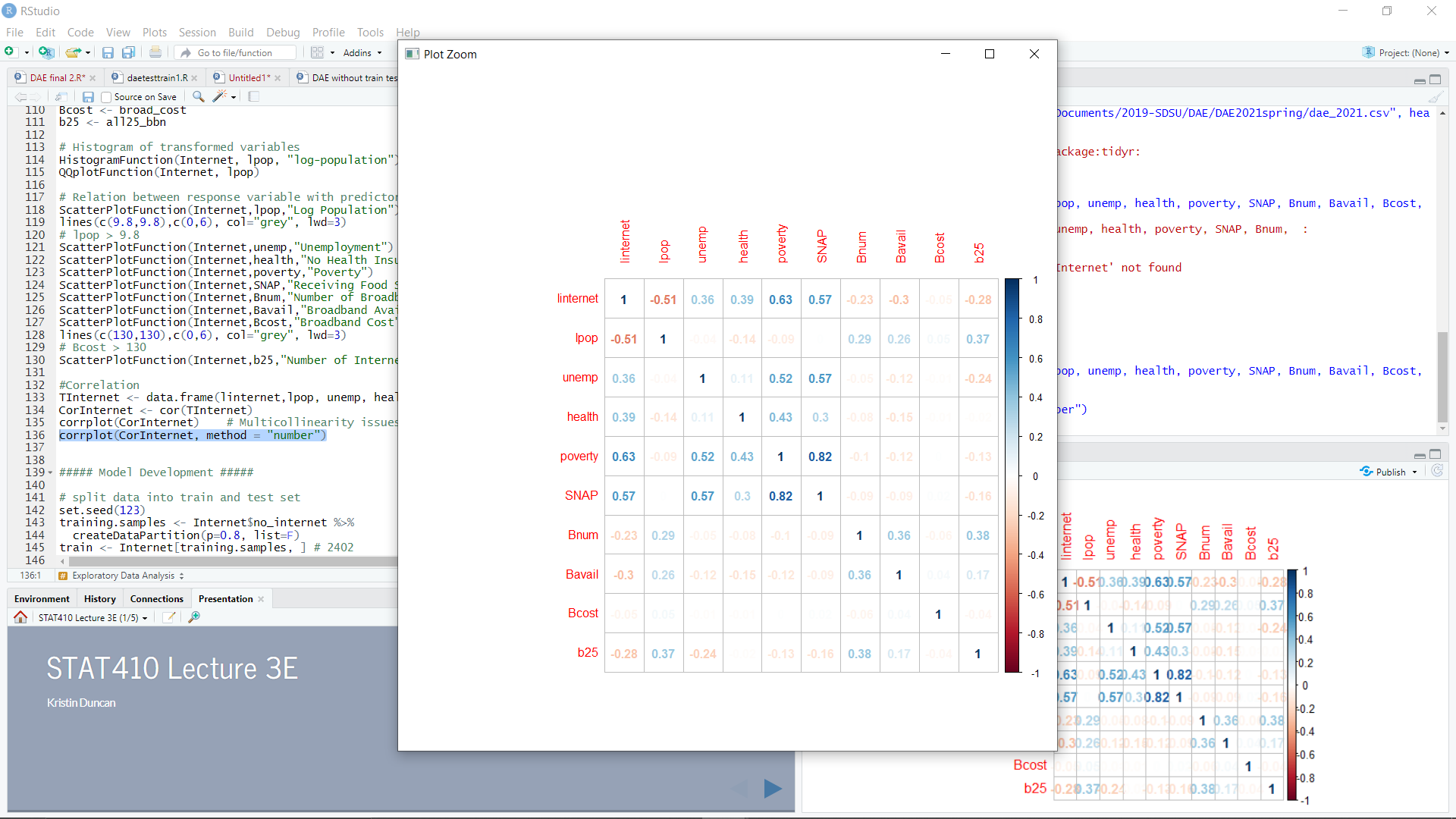


Figure A2: Correlation matrix plot with numbers

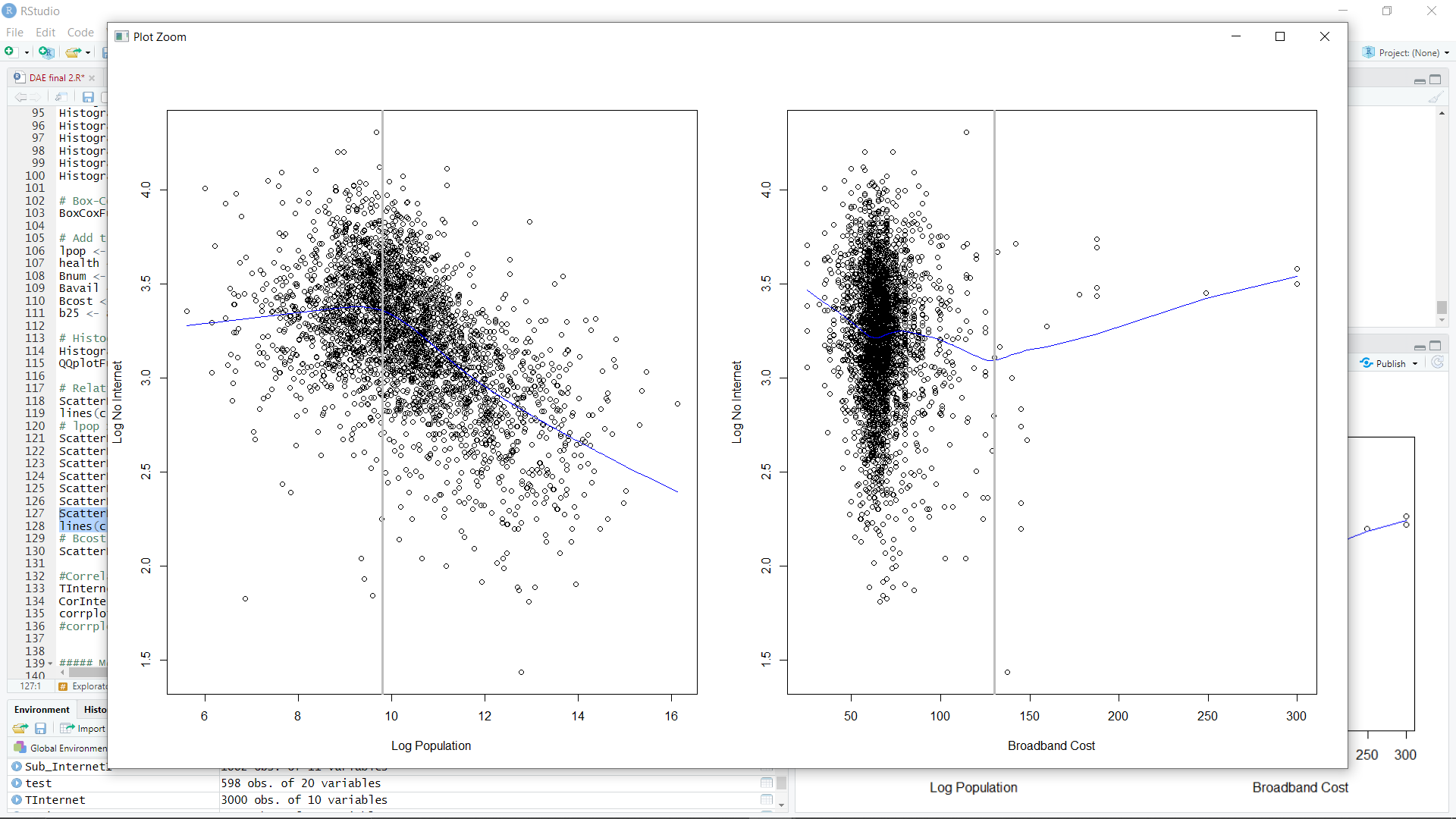


Figure A3: The scatter plot of log transformed population and broadband cost with log transformed response variable. It shows that cut offline of 9.8 for log transformed population and 130 for the broadband cost.

**Appendix B**

##### Packages #####

library(MASS); library(corrplot); library(car); library(leaps)

library(rsample); library(Metrics); library(ggplot2); library(boot)

##### Functions #####

HistogramFunction <- function(df, xvar, xlab){

ggplot(df, aes(x=xvar)) +

geom\_histogram(aes(y=stat(density)), bins=30, colour="black", fill="white")+

geom\_density(size=1, color="blue")+

geom\_vline(aes(xintercept=mean(xvar)), color="blue",

linetype="dashed")+

labs(title="",x=xlab, y = "Density")+

theme\_bw()

}

BoxCoxFunction <- function(df, xvar){

bc = boxcox(xvar~1)

lambda = bc$x[which.max(bc$y)]

return(lambda)

}

QQplotFunction <- function(df, xvar){

ggplot(df, aes(sample=xvar))+

stat\_qq(size=2, alpha=.4)+

stat\_qq\_line(col="blue", size=1)+

theme\_bw()

}

ScatterPlotFunction <- function(df, xvar, xlabname){

plot(x=xvar, y=linternet, xlab=xlabname, ylab="Log No Internet", data=df )

lines(lowess(xvar, linternet), col="blue")

}

ValidationTable <- function(fit, model\_type){

mod <- fit

fit\_summary <- tibble(Model=model\_type,

"Number of Features"=length((coef(mod) %>% names())[-1]),

MSE=mean(mod$residuals^2),

RMSE=sqrt(MSE),

Adj.R.squared=summary(mod)$adj.r.squared,

F.statistics=summary(mod)$fstatistic[[1]],

AIC=AIC(mod))

return(fit\_summary)

}

##### Load data #####

Internet<-read.csv("C://~dae\_2021.csv", header=T)

sum(is.na(Internet)) # no missing value

dim(Internet) # 3000\*11

head(Internet,10); names(Internet)

###### Summary Statistic #####

summary(Internet); str(Internet)

attach(Internet)

###### Exploratory Data Analysis #####

# Response variable

HistogramFunction(Internet, no\_internet, "no\_internet")

QQplotFunction(Internet, no\_internet)

shapiro.test(Internet$no\_internet)

#The null hypothesis for this test is that the data are normally distributed.

# Box-Cox transformation

BoxCoxFunction(Internet, no\_internet)

linternet <- log(no\_internet)

# Transformed response variable

HistogramFunction(Internet, linternet, "Log No Internet")

QQplotFunction(Internet, linternet)

# Graphics for explorations of relationships

Internetvars = data.frame(no\_internet, population, unemp, health\_ins, poverty, SNAP, broad\_num, broad\_avail,broad\_cost, all25\_bbn)

scatterplotMatrix(Internetvars) # in CARS package

# Independent variable

# Histogram of non-transformed variables

HistogramFunction(Internet, population, "Population") # right skewed

HistogramFunction(Internet, unemp, "unemp")

HistogramFunction(Internet, health\_ins, "health\_insurance")

HistogramFunction(Internet, poverty, "poverty")

HistogramFunction(Internet, SNAP, "SNAP")

HistogramFunction(Internet, broad\_num, "broad\_num")

HistogramFunction(Internet, broad\_avail, "broad\_avail")

HistogramFunction(Internet, broad\_cost, "broad\_cost")

HistogramFunction(Internet, all25\_bbn, "all25\_bbn")

# Box-Cox transformation

BoxCoxFunction(Internet, population)

# Add transformed variables

lpop <- log(population)

health <- health\_ins

Bnum <- broad\_num

Bavail <- broad\_avail

Bcost <- broad\_cost

b25 <- all25\_bbn

# Histogram of transformed variables

HistogramFunction(Internet, lpop, "log-population")

QQplotFunction(Internet, lpop)

# Relation between response variable with predictors and transformed predictors

ScatterPlotFunction(Internet,lpop,"Log Population")

lines(c(9.8,9.8),c(0,6), col="grey", lwd=3)

# lpop > 9.8

ScatterPlotFunction(Internet,unemp,"Unemployment")

ScatterPlotFunction(Internet,health,"No Health Insurance")

ScatterPlotFunction(Internet,poverty,"Poverty")

ScatterPlotFunction(Internet,SNAP,"Receiving Food Stamps")

ScatterPlotFunction(Internet,Bnum,"Number of Broadband Service Provider")

ScatterPlotFunction(Internet,Bavail,"Broadband Availability")

ScatterPlotFunction(Internet,Bcost,"Broadband Cost")

lines(c(130,130),c(0,6), col="grey", lwd=3)

# Bcost > 130

ScatterPlotFunction(Internet,b25,"Number of Internet Service Provider with 25Mbps")

#Correlation

TInternet <- data.frame(linternet,lpop, unemp, health, poverty, SNAP, Bnum, Bavail, Bcost, b25)

CorInternet <- cor(TInternet)

corrplot(CorInternet) # Multicollinearity issues between SNAP, poverty, and unemp

corrplot(CorInternet, method = "number")

##### Model Development #####

# split data into train and test set

set.seed(123)

training.samples <- Internet$no\_internet %>%

createDataPartition(p=0.8, list=F)

train <- Internet[training.samples, ] # 2402

test <- Internet[-training.samples, ] # 598

detach(Internet); attach(train)

train %<>%

mutate(linternet=log(no\_internet),

lpop=log(population),

health=health\_ins,

Bavail=broad\_avail,

Bnum=broad\_num,

Bcost=broad\_cost,

b25=all25\_bbn)

# Regression model/ Model selection

fit <- lm(linternet~lpop+unemp+health+poverty+SNAP+Bnum+Bavail+Bcost+b25, data=train)

summary(fit1)

stepAIC(fit1, direction = "both")

# Suggest removal of unemp, Bnum, and b25

fit1 <- lm(linternet~lpop+health+poverty+SNAP+Bavail+Bcost, data=train)

summary(fit1)

stepAIC(fit1, direction = "both")

# Suggest removal of Bcost

fit2 <- lm(linternet~lpop+health+poverty+SNAP+Bavail, data=train)

summary(fit2)

stepAIC(fit2, direction = "both")

# Interaction

fit3 <- lm( linternet ~ lpop+health+poverty+SNAP+Bavail+

lpop:health+lpop:poverty+lpop:SNAP+lpop:Bavail+

health:poverty+health:SNAP+health:Bavail+

poverty:SNAP+poverty:Bavail, data=train )

summary(fit3)

stepAIC(fit3, direction = "both")

# Health and Bavail are not right direction.

fit4 <- lm(linternet ~ lpop + health + poverty + SNAP + Bavail +

lpop:health + lpop:poverty + lpop:SNAP + health:poverty +

health:SNAP + health:Bavail + poverty:SNAP, data = train)

summary(fit4)

stepAIC(fit4, direction = "both")

# Health is still not right direction.

# Generate iteratin log

fit\_summary <- ValidationTable(fit, "Full Model")

fit1\_summary <- ValidationTable(fit1, "Stepwise Model")

fit2\_summary <- ValidationTable(fit2, "Stepwise Model with adjust")

fit3\_summary <- ValidationTable(fit4, "Stepwise Model with Interaction Terms I")

fit4\_summary <- ValidationTable(fit4, "Stepwise Model with Interaction Terms II")

bind\_rows(fit\_summary, fit1\_summary, fit2\_summary, fit3\_summary, fit4\_summary) %>%

modify\_if(is.numeric, round,3)

# Final Model

fit\_final <- fit2

summary(fit\_final)$coefficient

confint(fit\_final)

# Multicollinearity

car::vif(fit\_final) # no multicollinearity issue ( all VIF values are less than 4)

#MSE

MSE = mean(fit\_final$residuals^2)

# Error rate

error\_rate<-sigma(fit\_final)/mean(train$linternet)

error\_rate #0.06629

fit\_final %>%

tidy() %>%

modify\_if(is.numeric, round,3)

##### Model Diagnostics #####

# Diagnostic plots

par(mfrow=c(1,2))

plot(fit\_final)

plot(predict(fit\_final), rstudent(fit\_final), main="Residual vs Fitted", ylab="Studentized Residuals", xlab="Predicted")

predict(fit\_final)[rstudent(fit\_final)==min(rstudent(fit\_final))]

abline(h=0, lty=1, col="red")

qqPlot(fit\_final, "Normal QQplot", ylab="Studentized Residuals", xlab="Theoretical quantiles")

model.diag <- augment(fit\_final) #broom package

head(model.diag, 10)

# Cook's D

cutoff <- 4/((nrow(train)-length(fit\_final$coefficients)-2))

plot(fit\_final, which=4, cook.levels=cutoff) # Outliers: 3, 1814, 1843

# Influential Observations

#influence plot: studentized residuals vs hat matrix diagnosis(leverage) with bubbles a function of Cook's D

influencePlot(fit\_final, id.method="identify", main="Influence Plot", sub="Circle size is proportional to Cook's D")

# Influence points: 3, 10, 1814

# Outliers

train[3,]; train[10,]; train[1814,];train[1843,]

#10 fold cross validation

mod1 <- glm(linternet~lpop+health+poverty+SNAP+Bavail+Bcost, data=train)

cv.err <- cv.glm(train, mod1, K=10) #boot package

cv.err$delta #0.04557628 0.04555303

detach(train); attach(test)

#Predict

test %<>%

mutate(linternet=log(no\_internet),

lpop=log(population),

health=health\_ins,

Bavail=broad\_avail,

Bnum=broad\_num,

Bcost=broad\_cost,

b25=all25\_bbn)

test %<>%

mutate(internet\_preds=predict(fit\_final, newdata=.))

test %<>%

mutate(internet\_error=linternet-internet\_preds)

#MSE and R-squared

predictions <- predict(fit\_final, newdata=test)

data.frame(R2=R2(predictions, test$linternet),

MSE=mse(predictions, test$linternet))

# Prediction error rate

P\_error\_rate=rmse(predictions, test$linternet)/mean(test$linternet)

P\_error\_rate #0.06779982