Multiple Linear Regression – Performance Assessment

Steven Schindler

Predictive Modeling – D208

Straw, Eric; PhD

College of I.T., Western Governors University

Table of Contents

Part I: Research Question

A1.

For my research question I have chosen to work with the Churn dataset. The research question is as follows: “What factors contribute to high averages of Outage_sec_perweek?” Outage_sec_perweek is a continuous variable that’s defined as average number of seconds per week of system outages in the customer’s neighborhood.

A2.

The goal of this analysis is to see which explanatory variables will have a high influence on Outage_sec_perweek. There are 6 potential variables that could contribute to high Outage_sec_perweek 4 quantitative variables and 2 categorical variables. The variables are yearly_equip_failure, bandwith_GB_year, Population, Tenure, Area, and InternetService.

For yearly_equip_failure which is the number of times in a year the customers equipment failed, is there a relationship between high equipment failure and high outage? Bandwith_GB_year is average amount of data used in a year by the customer. Is there a relationship between high data use and high outage? Population is the population within a mile of the customer. Is there higher number of outages in higher populations? Area is a type of rural, urban and suburban, does what area a customer live in contribute to higher outages? Tenure is the number of months a customer has stayed with the provider; does longer tenure mean more

outages? Finally, InternetService is the customers internet provider. Does the provider contribute to higher outages?

Part II: Method Justification

B1.

One assumption of multiple linear regression is that the relationship between independent variables and the dependent variable is linear. The next assumption is that residuals should be normally distributed with a mean of zero. The third assumption is that there is no Multicollinearity which is when two independent variables are highly correlated. Finally, that observations are selected independently and randomly from the population. (Statistics Solutions. (n.d.))

B2.

I used both Python and R in this performance assessment. I used Python to clean the data then I used R for the regression model. Python has many useful libraries that make it ideal to clean data such as pandas, numpy and stats. These libraries make it easy to find missing values, replace trailing and leading spaces as well as easily calculate Z-scores and detect outliers. Python also has easy readability of it's code.

I used R for the regression model because it too has many useful libraries such as pyler,dpyler,ggplot2 and car. R also has many built in functions that make it ideal working with data to get statistics. It provides summary statistics and with ggplot2 graphs are made easily to visualize the data. It also provides simple functions such as the lm function to calculate linear regression easier.

B3.

Outage_per_week is a quantitative continuous variable therefore it is appropriate to use linear regression to find what contributes to high outages. The use of more than one explanatory variable makes it appropriate to use multiple linear regression to predict the relationships between Outage_per_week and the rest of the explanatory variables.

Part III: Data Preparation

C1.

The goal for cleaning the data is to locate and impute missing values such as NAN on the variables for my research question of Outage_sec_perweek, yearly_equip_failure, bandwith_GB_year, Population ,Tenure, Area, and InternetService. Also, to check and treat for duplicates and outliers on the aforementioned variables.

The first thing I do is import the libraries in Python and then read the csv file into a dataframe called df. Then, I create a new dataframe with only the 6 variables that I need to answer my research question as well as the CaseOrder variable for a count. Next, I count the number of missing values and see that it is zero in each column. For duplicates I use a for loop to go through the categorical values and strip them of leading and trailing spaces.

I used Z-scores to find the number of outliers with Population, Outage_sec_perweek, and yearly_equip_failure all having outliers of 219, 28, and 94 respectively. Bandwith_GB_year and Tenure had no outliers. However, I did not remove outliers because when I did, I found that most of the outliers were larger numbers than the mean and I wish to run my regression model on all numbers so I left the outliers in. However, Population cannot be zero, so I removed

all records where population is zero then uploaded my dataframe to a new csv file called clean_churn_file.csv. The code is below:

```
# import the numpy and pandas libraries
import numpy as np
import pandas as pd
import string as str
import matplotlib.pyplot as plt
from scipy import stats

#read in the churn data set as a pandas dataframe
df = pd.read_csv('churn_clean.csv')

#new dataframe for the variables that are needed and the CaseOrder variable for identification.
new_df = pd.DataFrame().assign(CaseOrder=df['CaseOrder'],
                               Outage_sec_perweek=df['Outage_sec_perweek'],
                               Yearly_equip_failure=df['Yearly_equip_failure'],
                               InternetService=df['InternetService'],
                               Bandwidth_GB_Year=df['Bandwidth_GB_Year'],
                               Population=df['Population'],
                               Area=df['Area'],
                               Tenure=df[‘Tenure’]
                              )

new_df.isna().sum()

#strips leading and trailing spaces
string_list = list(new_df.select_dtypes(include = {'object'}))
for i in string_list:
    new_df[i] = new_df[i].str.strip()
col_num_names =
['Population','Outage_sec_perweek','Yearly_equip_failure','Bandwidth_GB_Year',
’Tenure’]


#Find min max of the numeric datatypes except CaseOrder as well as number of outliers.
for i in col_num_names:
```

```
    #check for outliers using zscore
    new_df_count = new_df[(np.abs(stats.zscore(new_df[i])) < 3)].count()
    print(new_df[i].name,"number of outliers is ",10000 - new_df_count[i], ",\n")
    print(new_df[i].name,"min is ", new_df[i].min(),"max is ",new_df[i].max(),"\n")
#population cannot be zero.
count = (new_df['Population']==0).sum()
count

#less than 1 percent of the data so we can remove the zeros from population
new_df.drop(new_df[new_df['Population'] == 0].index, inplace = True)
new_df.to_csv('/Users/stevenschindler/Documents/R/D208/clean_churn_file.csv',
index=False)
```

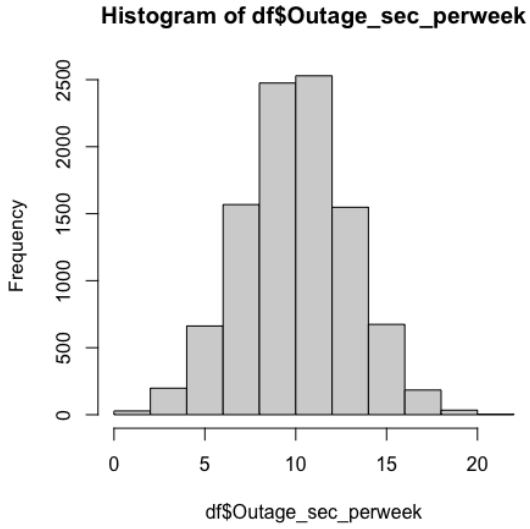
C2.

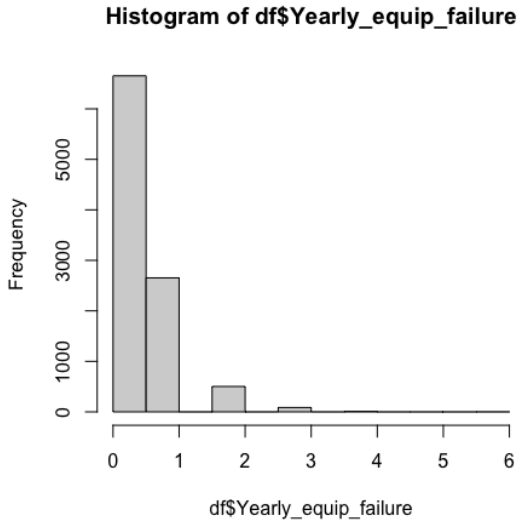
The mean and median for Outage_sec_perweek are 10.00427 and 10.01977 respectively. For Area the frequency for Rural, Suburban and Urban is 3298,3315 and 3290 respectively. Yearly_equip_failure has a mean of .399 and a median of 0. InternetService has a 3430 frequency for DSL, 4365 for Fiber Optic and 2108 for None. Population has a median of 2976 and a mean of 9852.Tenure has a median of 36.330 and a mean of 34.556. Finally, Bandwith_GB_year has a mean of 3395 and a median of 3343.6. The summary statistics for each variable can be seen in the picture below generated by R.

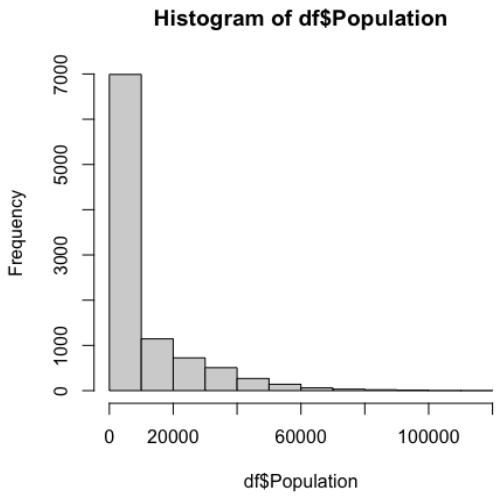
```
> summary(df$Outage_sec_perweek)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.09975  8.02424 10.01977 10.00427 11.97330 21.20723
> summary(df$Area)
   Length     Class      Mode
     9903 character character
> summary(df$Yearly_equip_failure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   0.399   1.000   6.000
> summary(df$InternetService)
   Length     Class      Mode
     9903 character character
> summary(df$Population)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2     775    2976    9852   13288  111850
> summary(df$Bandwidth_GB_Year)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  155.5  1237.8  3343.6  3395.0  5588.2  7159.0
> summary(df$Tenure)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   7.944  36.330  34.556  61.494  71.999
>
```

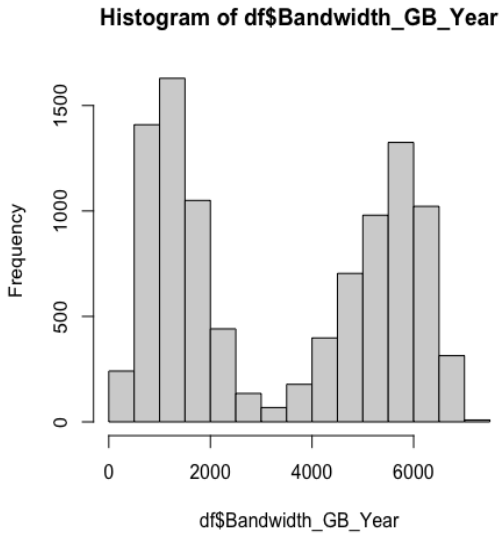
C3.

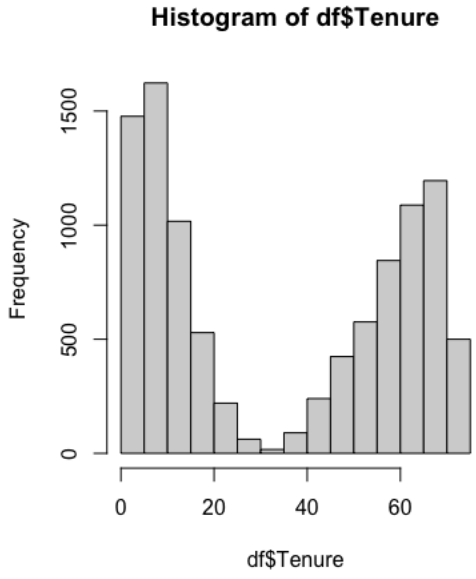
Below are the 7 univariate visualizations for all variables needed to for the research question the 5 numeric variables are given in histograms while the 2 categorical variables are given as bar plots. They are in the following order: Outage_sec_perweek, Yearly_equip_failure, Population, Bandwidth_GB_Year, Tenure, Area and InternetService.

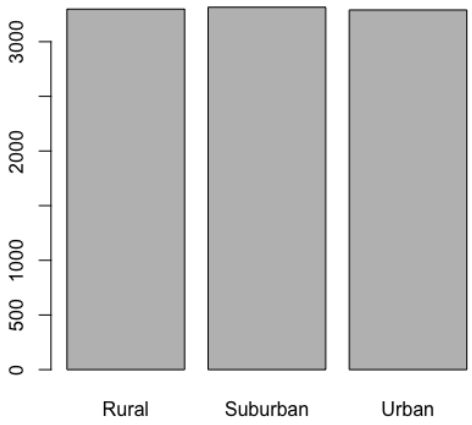


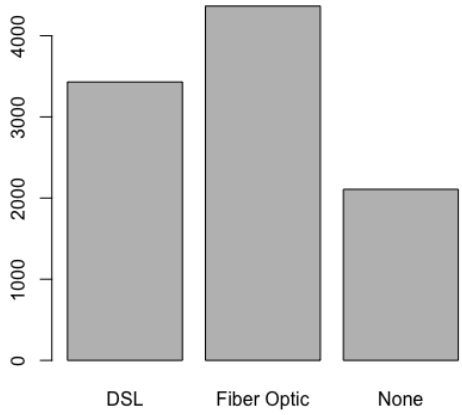



Histogram of df$Population
Frequency
0
1000
3000
5000
7000
0
20000
60000
100000
df$Population

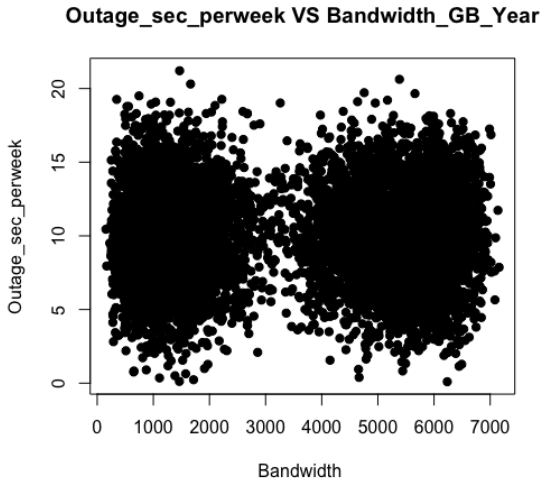

Histogram of df$Bandwidth_GB_Year
Frequency
0
500
1000
1500
0
2000
4000
6000
df$Bandwidth_GB_Year

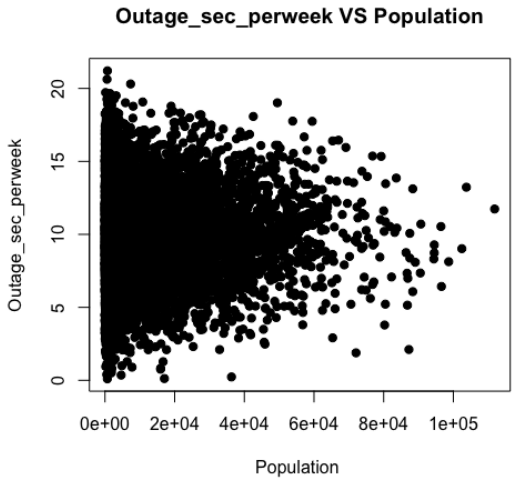

Histogram of df$Tenure
Frequency
0
500
1000
1500
0
20
40
60
df$Tenure

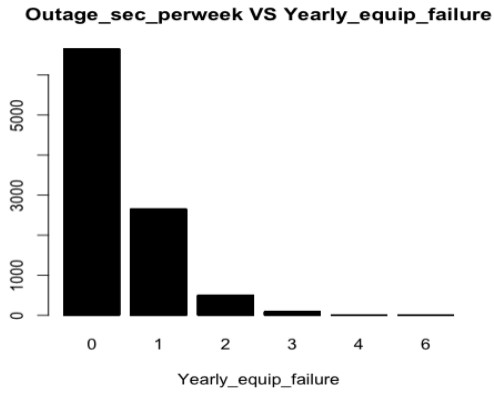


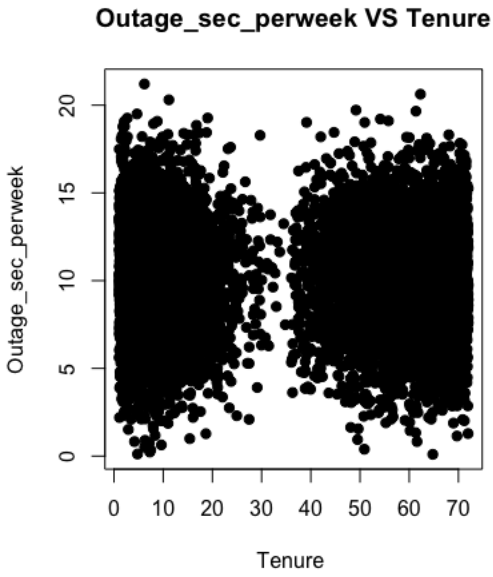


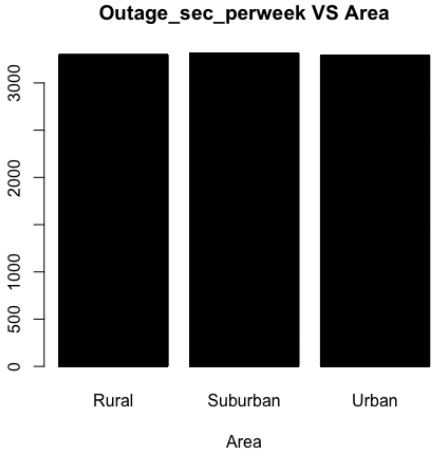
Next are the 6 bivariate graphs as each explanatory variable is paired with Outage_sec_perweek made in R.

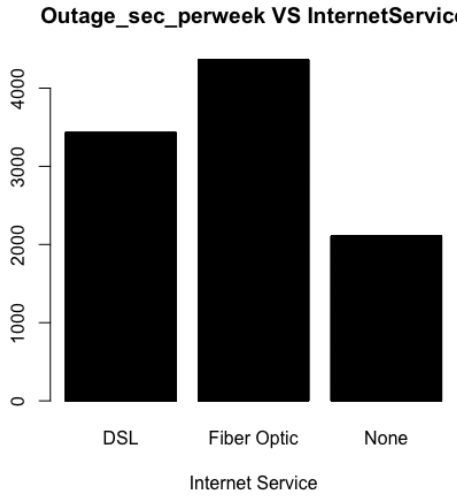

Outage_sec_perweek VS Bandwidth_GB_Year
Outage_sec_perweek
0
5
10
15
20
0
1000
2000
3000
4000
5000
6000
7000
Bandwidth


Outage_sec_perweek VS Population
Outage_sec_perweek
0
5
10
15
20
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Population


Outage_sec_perweek VS Yearly_equip_failure
0
1000
3000
5000
0
1
2
3
4
6
Yearly_equip_failure


Outage_sec_perweek VS Tenure
Outage_sec_perweek
0
5
10
15
20
0
10
20
30
40
50
60
70
Tenure


Outage_sec_perweek VS Area
0
500
1000
2000
3000
Rural
Suburban
Urban
Area


Outage_sec_perweek VS InternetService
0
1000
2000
3000
4000
DSL
Fiber Optic
None
Internet Service

C4.

My goals for data transformation are to transform the two categorical variables into a numeric representation. There is no specific order for InternetService or Area, so I used One hot encoding to transform the three values of each variable to columns of their own. With a 1 being the customer has that value and a 0 if the customer does not have that value. First, I factor InternetService and Area using the as.factor method on each variable then, I call the one_hot function of the dataframe as a table and assign it to a new dataframe new_df. The code for one hot encoding is as follows.

```
#factor InternetService and area
df$InternetService <- as.factor(df$InternetService)
df$Area <- as.factor(df$Area)
# One-Hot-Encoding

new_df <- one_hot(as.data.table(df))
head(new_df)
```

C5.

The new data file is called new_churn_clean.csv.

Part IV: Model Comparison and Analysis

D1.

My initial model is as follows $Y = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + B_5X_5 + B_6X_6 + B_7X_7 + B_8X_8$. With Y being Outage_sec_perweek and B_0 being the y-intercept. B_1 through B_8 are the coefficients of the following X-values X_1 = Yearly_equip_failure, X_2 = Population, X_3 = Bandwidth_GB_Year, X_4 =

InternetService_DSL, X_5 = InternetService_Fiber_Optic, X_6 = Area_Rural, and X_7 = Area_Urban, X_8 =Tenure. This gives us Y = 9.890 + 1.533e-02X_1 + 1.085e-06X_2 + 3.705e-04X_3 - 2.479e-01X_4 - 2.848e-02X_5 - 2.957e-02X_6 -2.347e-02X_7 - 3.023e-02X_8. I left out InternetService_None and Area_Suburban using k-1 to reduce multicollinearity among InternetService and Area respectively. Here is my model in R:

model <- lm(Outage_sec_perweek ~ Yearly_equip_failure + Population + Bandwidth_GB_Year + InternetService_DSL + InternetService_Fiber_Optic + Area_Rural + Area_Urban + Tenure, data = new_df)

D2.

To test for multicollinearity, we can use variance inflation factor which measures how much one independent variable is influenced by another independent variable. Running the VIF function in R on the model we can see from the screenshot below that Bandwidth_GB_Year and Tenure have a strong correlation because their VIFs are both over 100. We then run backward steps elimination to reduce the initial model. Running the function ols_step_backward_p, which does backward step elimination using the P value, we see that Area_Urban, Area_Rural, Yearly_equip_failure, InternetService_Fiber_Optic and Population can be removed from the model.

Vif:

```
> vif(model)
      Yearly_equip_failure                Population         Bandwidth_GB_Year        InternetService_DSL
                  1.000388                  1.000350                114.405141                  2.663693
InternetService_Fiber_Optic                Area_Rural                Area_Urban                    Tenure
                  1.717325                  1.330858                  1.330863                113.181967
```

Backward Step Elimination: (R-Squared Academy Ltd. (n.d.))

```
                                Elimination Summary
-------------------------------------------------------------------------------------------
        Variable                                Adj.
Step             Removed          R-Square    R-Square     C(p)        AIC         RMSE
-------------------------------------------------------------------------------------------
   1    Area_Urban                    8e-04       1e-04     7.1029    49697.2853    2.9736
   2    Area_Rural                    8e-04       2e-04     5.1823    49695.3649    2.9734
   3    Yearly_equip_failure          8e-04       3e-04     3.2903    49693.4729    2.9733
   4    InternetService_Fiber_Optic   8e-04       4e-04     1.4206    49691.6033    2.9732
   5    Population                    8e-04       5e-04    -0.3024    49689.8806    2.9730
-------------------------------------------------------------------------------------------
>
```

D3.

The reduced model is now in the form of $Y = B_0 + B_1X_1 + B_2X_2$. Where X_1 = Bandwidth_GB_Year and X_2 = InternetService_DSL. Which gives us $Y = 1.002e+01 + 4.444e-06X_1 - 7.614e-02X_2$. The new model in R is *new_model <- lm(formula = Outage_sec_perweek ~ Bandwidth_GB_Year + InternetService_DSL, data = new_df)*. The summary for the first model is:

```
Call:
lm(formula = Outage_sec_perweek ~ Yearly_equip_failure + Population +
    Bandwidth_GB_Year + InternetService_DSL + InternetService_Fiber_Optic +
    Area_Rural + Area_Urban + Tenure, data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9307 -1.9815  0.0097  1.9538 11.1918

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 9.890e+00  1.097e-01  90.132   <2e-16 ***
Yearly_equip_failure        1.533e-02  4.696e-02   0.326   0.7442
Population                  1.085e-06  2.065e-06   0.525   0.5995
Bandwidth_GB_Year           3.705e-04  1.462e-04   2.534   0.0113 *
InternetService_DSL        -2.479e-01  1.025e-01  -2.419   0.0156 *
InternetService_Fiber_Optic -2.848e-02  7.887e-02  -0.361   0.7180
Area_Rural                 -2.957e-02  7.314e-02  -0.404   0.6860
Area_Urban                 -2.347e-02  7.319e-02  -0.321   0.7484
Tenure                     -3.023e-02  1.202e-02  -2.514   0.0120 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.974 on 9894 degrees of freedom
Multiple R-squared:  0.0008545, Adjusted R-squared:  4.667e-05
F-statistic: 1.058 on 8 and 9894 DF,  p-value: 0.3898
```

The summary for the new model is:

```
> summary(new_model)

Call:
lm(formula = Outage_sec_perweek ~ Bandwidth_GB_Year + InternetService_DSL,
    data = new_df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8674 -1.9779  0.0126  1.9685 11.2613

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         1.002e+01  5.777e-02 173.374   <2e-16 ***
Bandwidth_GB_Year   4.444e-06  1.375e-05   0.323    0.747
InternetService_DSL -7.614e-02  6.315e-02  -1.206    0.228
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.974 on 9900 degrees of freedom
Multiple R-squared:  0.0001508, Adjusted R-squared:  -5.119e-05
F-statistic: 0.7466 on 2 and 9900 DF,  p-value: 0.474
```

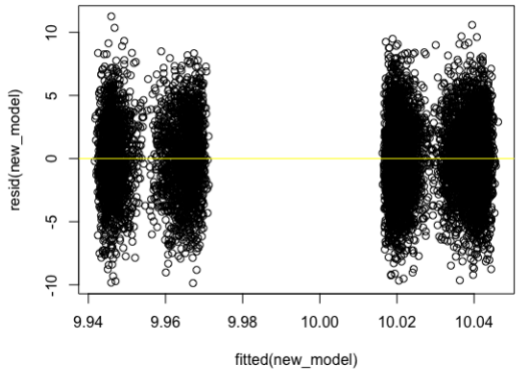
E1.

Comparing the two models the AIC and BIC of the first model is 49699.18 and 49771.19 respectively, while the AIC and BIC of the second model is 49694.16 and 49722.96. So according to this metric the second model is a better fit but not by much. If you look at the P-values in the screenshots above, we actually see that it increases thus making it less statistically significant assuming a significance value of .05.

E2.

```
> AIC(model)
[1] 49699.18
> AIC(new_model)
[1] 49694.16
> BIC(model)
[1] 49771.19
> BIC(new_model)
[1] 49722.96
```

Above is the output for both AIC and BIC (ProjectPro. (n.d.).) on each model performed in R. From the second screenshot in D3 we can see that the reduced model's residual standard error is 2.974 on 9900 degrees of freedom. Below is a screenshot of the residual plot.(Statology. (n.d.))



E3.

Below is my entire code in R which I've also included as a separate attachment called D208_MLR.R.

```
library(plyr)
library(dplyr)
library(ggplot2)
library(car)
library(mltools)
library(data.table)
library(olsrr)
getwd()
setwd('/Users/stevenschindler/Documents/R/D208')
df <- read.csv('clean_churn_file.csv', header = TRUE)
head(df)

summary(df$Outage_sec_perweek)
summary(df$Area)
summary(df$Yearly_equip_failure)
summary(df$InternetService)
summary(df$Population)
summary(df$Bandwidth_GB_Year)
summary(df$Tenure)


hist(df$Outage_sec_perweek)
hist(df$Yearly_equip_failure)
hist(df$Population)
hist(df$Bandwidth_GB_Year)
hist(df$Tenure)



#factor InternetService and area
df$InternetService <- as.factor(df$InternetService)
df$Area <- as.factor(df$Area)

Area <- table(df$Area)
InternetService <- table(df$InternetService)
```

```
barplot(Area)
barplot(InternetService)

# bivariate graphs
dsl_vs_out <- table(df$Outage_sec_perweek, df$InternetService)
barplot(dsl_vs_out, main="Outage_sec_perweek VS InternetService",
      xlab="Internet Service")
yearly_vs_out <- table(df$Outage_sec_perweek, df$Yearly_equip_failure)
barplot(yearly_vs_out, main="Outage_sec_perweek VS Yearly_equip_failure",
      xlab="Yearly_equip_failure")
area_vs_out <- table(df$Outage_sec_perweek, df$Area)
barplot(area_vs_out, main="Outage_sec_perweek VS Area",
      xlab="Area")

plot(df$Bandwidth_GB_Year, df$Outage_sec_perweek,
main="Outage_sec_perweek VS Bandwidth_GB_Year",
   xlab="Bandwidth ", ylab="Outage_sec_perweek", pch=19)

plot(df$Population, df$Outage_sec_perweek, main="Outage_sec_perweek VS
Population",
   xlab="Population ", ylab="Outage_sec_perweek", pch=19)

plot(df$Tenure, df$Outage_sec_perweek, main="Outage_sec_perweek VS
Tenure",
   xlab="Tenure ", ylab="Outage_sec_perweek", pch=19)

# One-Hot-Encoding

new_df <- one_hot(as.data.table(df)) (Data Tricks. (n.d.).)
head(new_df)

write.csv(new_df, "new_churn_clean.csv", row.names=FALSE)

new_df <- new_df %>% rename("InternetService_Fiber_Optic" =
"InternetService_Fiber Optic")

model <- lm(formula = Outage_sec_perweek ~ Yearly_equip_failure + Population
+
         Bandwidth_GB_Year + InternetService_DSL +
InternetService_Fiber_Optic +
```

Area_Rural + Area_Urban + Tenure, data = new_df)

summary(model)
vif(model)
ols_step_backward_p(model) (R-Squared Academy Ltd., n.d.)

new_model <- lm(formula = Outage_sec_perweek ~ Bandwidth_GB_Year + InternetService_DSL, data = new_df)
summary(new_model)
vif(new_model)

AIC(model) (ProjectPro. (n.d.).)
AIC(new_model) (ProjectPro. (n.d.).)
BIC(model)
BIC(new_model)

plot(fitted(new_model),resid(new_model)) (Statology. (n.d.).)
abline(0,0,col = "yellow")

Part V: Data Summary and Implications

F1.

As previously stated in D3 the reduced model equation is Y = 1.002e+01 + 4.444e-06X_1 - 7.614e-02X_2. The y intercept of 10.02 means when Bandwidth and InternetService_DSL both equal 0 then Outage_sec_perweek = 10.02. The coefficient for X_1 is 0.000004444 meaning that for every use of Bandwidth all things being equal the Outage will increase by .000004444 seconds. The coefficient for X_2 is -0.07614 meaning for every use of InternetService_DSL all things being equal the Outage will decrease by .07614 seconds.

My model has a P-value of .474 which is greater than an assumed significance level of .05 meaning my model is not statistically significant nor is it

practically significant. It cannot reasonably answer what factors contribute to high outages in a 1-mile radius of the customer.

One of the big limitations of my analysis is regression is not resistant to outliers, and from my data preparation I did not remove all outliers in the dependent variable in order to include high values of outages. Another limitation is that I only included variables I intuitively thought would have an impact on Outage_sec_perweek. Perhaps another variable that one would not intuitively think would have an impact on outages does have a strong correlation. A limitation of the backwards step elimination in the first model I had P-value of .3898 and, in the new model after backwards step elimination I had P-value of .474, indicating that my first model may have had slightly better statistical significance.

F2.

The research question cannot be definitively answered to what cause high churn other than bad luck. We know that the more Bandwidth the customer uses per year the seconds of outages may increase by an insignificant factor of .000004444. We also know that using DSL may decrease your chances of having outages by a factor of .07614, but since correlation is not causation we don't know for sure if using lower Bandwidth or having DSL internet will reduce the number of outages.

A recommended course of action for the company since the model is not practically significant is an analysis could be run on larger number of outages in each state to see if a particular location may be correlated. But a cursory look into outages greater than 18 seconds in excel shows that there are 25 unique states that have those outage values. A more prudent cost-effective action would be to do nothing and accept outages will happen and the large numbers are outliers that cannot be easily predicted.

References

H.

R-Squared Academy Ltd. (n.d.). OLS step backward p-value selection. Retrieved April 07, 2023, from https://olsrr.rsquaredacademy.com/reference/ols_step_backward_p.html

Data Tricks. (n.d.). One-hot encoding in R: Three simple methods. Retrieved April 07, 2023, from https://datatricks.co.uk/one-hot-encoding-in-r-three-simple-methods

I.

Statistics Solutions. (n.d.). Assumptions of multiple linear regression. Retrieved April 07, 2023, from https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/

Statology. (n.d.). How to create a residual plot in R. Retrieved April 09, 2023, from https://www.statology.org/residual-plot-r/

ProjectPro. (n.d.). Evaluate time series models - AIC. Retrieved April 09, 2023, from https://www.projectpro.io/recipes/evaluate-time-series-models-aic