Contents

MH3510 Regression Analysis Assignment

Diana (U1740430C)

February 16, 2020

Abstract

This assignment is about traffic monitoring for a section of road or highway. We are interested to predict the average annual daily traffic (aadt) for a section of road or highway by using linear regression. It is defined as the average, over a year, of the number of vehicles that pass through a particular section of a road each day.

1 Graphical Representation

We use four predictor variables to predict the response variable (aadt). The variables are as follows:

X_1: population of county in which road section is located–the second column of data
X_2: number of lanes in road section– the third column of data
X_3: width of road section (in feet)-the fourth column of data
Control (X4): two-category quality variable indicating whether or not there is control of access to road section (1=access control; 2=no access control).

We plot the response variable and its predictor variables in a scatter plot matrix below (Figure 1). In figure 1, aadt refers to the annual daily traffic or Y, population refers to X_1, num_of_lanes refers to the number of lanes or X_2, width_road refers to the width of road or X_3, and access_control refers to the control variable or X_4. We will use the variable's notation (X_i) and its name interchangeably.

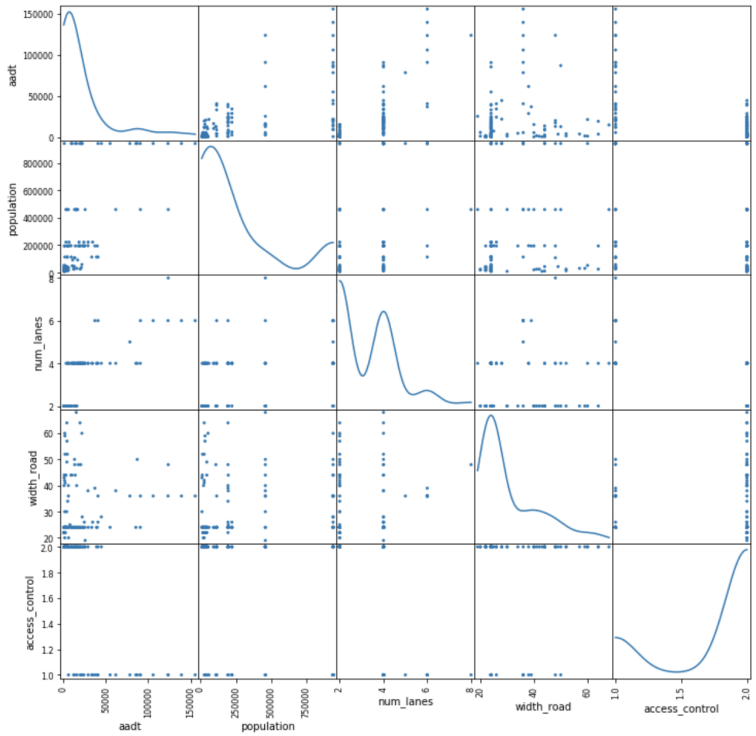

Figure 1: Scatter plot matrix.

Figure 1 shows the relationship between each variable. By analyzing the figure, it seems that aadt has relatively strong linear relationship with population, number of lanes, and access control. We could support our analysis by quantifying the correlation between each variable. The correlation is plotted in the covariance matrix figure below.

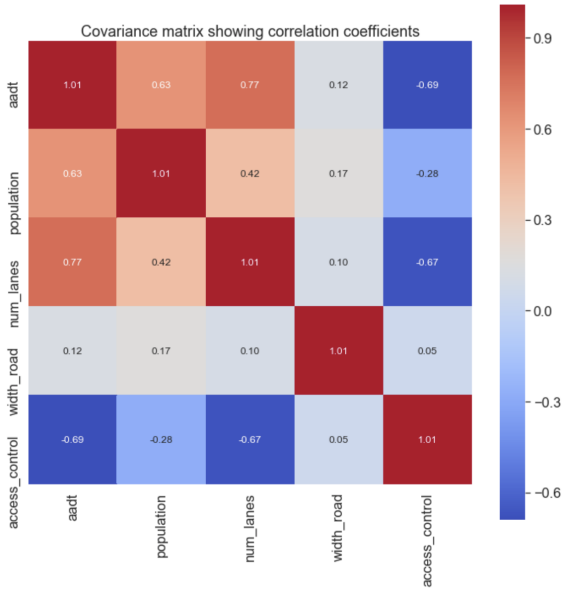

Figure 2: Covariance matrix.

Suppose we set the correlation between the response variable and the predictor variable to be more than 0.6 or less than -0.6 for variables to have strong linear relationship. The correlation between aadt and population, number of lanes, and access control are 0.63, 0.77, and -0.69 respectively. Thus, the numbers have further supported our analysis that population, number of lanes, and access control might be suitable predictors for our linear regression model. On the other hand, the correlation between aadt and width of road is low, which is 0.12. This shows that width of road might not have a strong linear relationship with aadt, therefore, it might not be a suitable predictor to predict the average annual daily traffic. We shall validate these assumptions by building a linear regression model, analyze the model, and improve the model.

2 Modelling Multiple Linear Regression

We fit the data to the following multiple linear regression model:

$$y_i = \beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \beta_3 * x_{i,3} + \beta_4 * x_{i,4}$$

,or we can rewrite it in the form of:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4$$

We obtained the following Ordinary Least Square (OLS) Regression result:

OLS Regression Results

Dep. Variable:	aadt	**R-squared:**	0.753
Model:	OLS	**Adj. R-squared:**	0.744
Method:	Least Squares	**F-statistic:**	88.29
Date:	Mon, 21 Oct 2019	**Prob (F-statistic):**	2.84e-34
Time:	12:39:44	**Log-Likelihood:**	-1335.0
No. Observations:	121	**AIC:**	2680.
Df Residuals:	116	**BIC:**	2694.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.118e+04	1.16e+04	1.821	0.071	-1855.322	4.42e+04
population	0.0330	0.005	7.017	0.000	0.024	0.042
num_lanes	9157.9390	1530.642	5.983	0.000	6126.310	1.22e+04
width_road	100.2890	124.274	0.807	0.421	-145.852	346.430
access_control	-2.361e+04	4520.295	-5.223	0.000	-3.26e+04	-1.47e+04

Omnibus:	27.238	**Durbin-Watson:**	1.314
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	81.523
Skew:	0.762	**Prob(JB):**	1.98e-18
Kurtosis:	6.721	**Cond. No.**	3.75e+06

Figure 3: OLS Regression result.

Thus, we are able to substitute in the values we have obtained by fitting the data to our model:

$$y_i = 2.118 * 10^4 + 0.0330 * x_{i,1} + 9157.9390 * x_{i,2} + 100.2890 * x_{i,3} - 2.361 * 10^4 * x_{i,4}$$

3 Adequacy Checking

3.1 From the viewpoint of the fitted model

For this assignment, we will use the level of significance $\alpha = 0.05$.

3.1.1 t-tests

In order to check the significance of the fitted parameters, we conduct t-tests. Using the level of significance $\alpha = 0.05$, the rejection region is when

$$|t^*| > t_{116,0.025} = 1.981$$

1. $H_0 : \beta_0 = 0$.
   $H_1 : \beta_0 \neq 0$.
   The t-value for β_0 from the table above is 1.821, whereas its $Pr(> |t|)$ is 0.071. Since $|t^*| = 1.821 < 1.981 = t_{116,0.025}$, thus we do not reject H_0.

2. $H_0 : \beta_1 = 0$.
   $H_1 : \beta_1 \neq 0$.
   The t-value for β_1 from the table above is 7.017, whereas its $Pr(> |t|) < 0.005$. Since $|t^*| = 7.017 > 1.981 = t_{116,0.025}$, thus we reject H_0.
3. $H_0 : \beta_2 = 0$.
   $H_1 : \beta_2 \neq 0$.
   The t-value for β_2 from the table above is 5.983, whereas its $Pr(> |t|) < 0.005$. Since $|t^*| = 5.983 > 1.981 = t_{116,0.025}$, thus we reject H_0.
4. $H_0 : \beta_3 = 0$.
   $H_1 : \beta_2 \neq 0$.
   The t-value for β_3 from the table above is 0.807, whereas its $Pr(> |t|)$ is 0.421. Since $|t^*| = 0.807 < 1.981 = t_{116,0.025}$, thus we do not reject H_0.
5. $H_0 : \beta_4 = 0$.
   $H_1 : \beta_2 \neq 0$.
   The t-value for β_4 from the table above is -5.223, whereas its $Pr(> |t|) < 0.005$. Since $|t^*| = 5.223 > 1.981 = t_{116,0.025}$, thus we reject H_0.

By the t-tests, we can conclude that β_0 and β_3 are not statistically different from 0. Furthermore, we show that β_1, β_2, and β_4 are statistically different from 0.

3.1.2 F statistic

We use F test to test the significance of the multiple linear regression. We test the hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$H_1 : \textit{at least one } \beta_i \neq 0$$

From the OLS Regression result, we obtained the value of the F-statistic with the degree of freedom (4, 116) is 88.29. In other words, $F = 88.29$. It is bigger than $F^{0.05}_{4,116} = 2.45$. Therefore, since $F > F^{0.05}_{4,116}$, we reject the null hypothesis.

Moreover, $p - value = 2.84 * 10^{-34} < 0.005$. The value of the F-statistic and the p-value show that the fitted parameters are significant to predict the response variable. In other words, there is a regression relation between the response variable Y and the predictor variables X_1, X_2, X_3, X_4.

3.1.3 R-squared

The value of the R-squared is 0.753 and the value of the adjusted R-squared is 0.744. Based on the given boundary of a good model's adjusted R-squared on the lecture slides ($0.6 < R_a < 0.95$), we can conclude that the fitted regression line is a good model because its adjusted R-squared is 0.744, which lies in the given boundary. This shows that 74.4% of the total variation is explained by the regression.

3.2 From the viewpoint of residuals

3.2.1 Normality checking

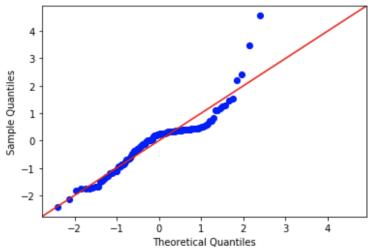

Figure 4: QQ-plot.

The QQ-plot shows that there is a little departure from the red line. This suggests that the error distribution is not normally distributed. More precisely, the distribution is skewed to the right.

3.2.2 Checking for time effects

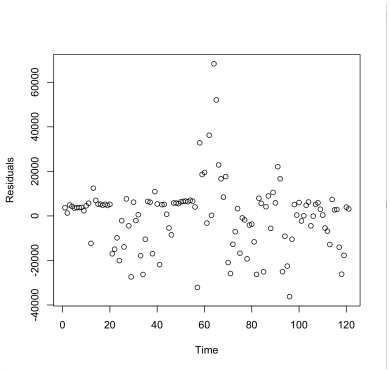

Figure 5: Residuals against time.

The graph depicts no pattern in the data when the residuals are plotted against time. It suggests that the error terms are independent over time.

3.2.3 Checking for the constancy of error variance

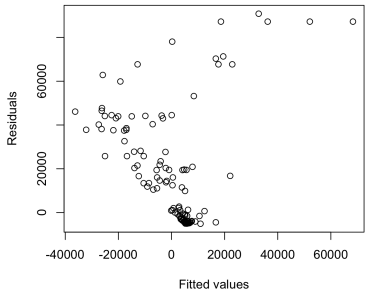

Figure 6: Residuals against fitted values.

By looking at the residual plot, it exhibits systematic pattern: as the values tend to be bigger, the residuals tend to be bigger as well. Furthermore, it seems that the residuals form a quadratic curve, which implies that there might be polynomial of degree two included in the residuals. It suggests that the variance is not constant and indicates the need of a curvilinear regression function.

In order to further support our analysis that the variance of the error terms is not constant, we can plot the residuals against each predictor variables and we obtained:

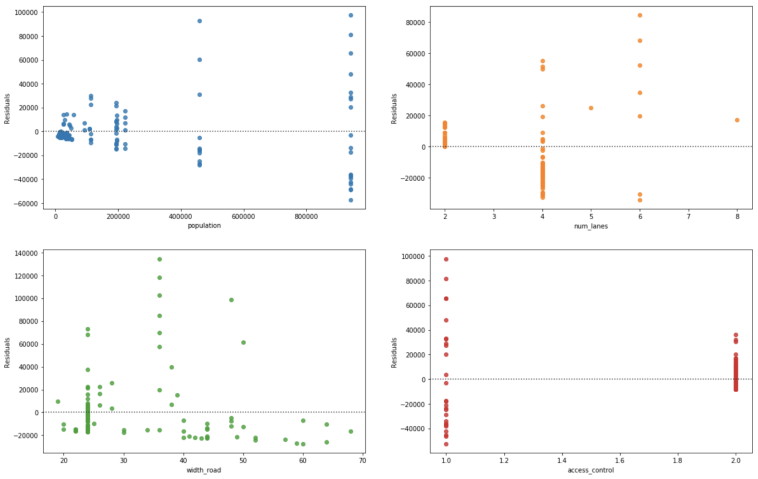

Figure 7: Residuals against predictor variables.

The plots of residuals against population, number of lanes, and width of road show that the residuals tend to be bigger as the value of each predictor variable

increases. On the other hand, the plot of residuals against access control shows that the residuals tend to be smaller as the value increases. These lead to the same conclusion as our previous analysis: the variance of the residuals is not constant.

3.2.4 Checking for linearity

By analysing Figure 4, we also come to the conclusion that the assumption of linearity is violated. Furthermore, we have also seen that the error terms are not normally distributed and the variance of the error terms is not constant. Thus, there should be some remedies to the existing model in order for the model to provide a more accurate prediction and depicts our data more precisely. We expect the aforementioned problems on the existing model to vanish.

3.3 Remedy Measures

We will do remedy for the fitted regression line so that it will be an appropriate model for our data. We will try two methods, the first one is by analyzing our graphs and the second one is Box-Cox method

3.3.1 Taking square root of y

Previously, figure 6 indicates the need of a curvilinear regression function. This suggest that we can try the transformation $\sqrt{y}$. Thus, the transformed model is as follows:

$$\sqrt{y_i} = \beta_0 + \beta_1 * x_{i,1} + \beta_2 * x_{i,2} + \beta_3 * x_{i,3} + \beta_4 * x_{i,4}$$

,or we can rewrite it in the form of:

$$\sqrt{Y} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4$$

We fit our data to the this model and obtained the following result:

OLS Regression Results

Dep. Variable:	np.sqrt(aadt)	**R-squared:**	0.857
Model:	OLS	**Adj. R-squared:**	0.852
Method:	Least Squares	**F-statistic:**	173.5
Date:	Mon, 21 Oct 2019	**Prob (F-statistic):**	5.58e-48
Time:	12:39:46	**Log-Likelihood:**	-592.16
No. Observations:	121	**AIC:**	1194.
Df Residuals:	116	**BIC:**	1208.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	82.7181	25.091	3.297	0.001	33.023	132.413
population	9.505e-05	1.02e-05	9.361	0.000	7.49e-05	0.000
num_lanes	32.6470	3.301	9.889	0.000	26.108	39.186
width_road	0.1401	0.268	0.523	0.602	-0.391	0.671
access_control	-57.9981	9.750	-5.949	0.000	-77.309	-38.687

Omnibus:	3.586	**Durbin-Watson:**	1.583
Prob(Omnibus):	0.166	**Jarque-Bera (JB):**	3.633
Skew:	-0.169	**Prob(JB):**	0.163
Kurtosis:	3.779	**Cond. No.**	3.75e+06

Figure 8: OLS Regression result of the transformed data.

By transforming Y to $\sqrt{Y}$, we obtained a higher value for the adjusted R-squared, which is 85.2%. Furthermore, by conducting similar t-tests, we can see that β_3 is not statistically different from 0. We also plot the residuals in the QQ-plot and obtained a much better result: there is only a little departure from the red line. The new QQ-plot suggests that the residuals are approximately normally distributed.

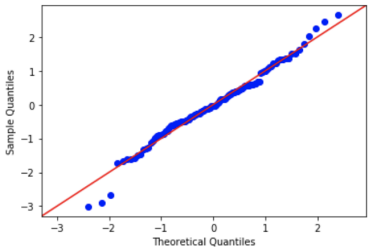

Figure 9: QQ-plot of the transformed data.

Next, we shall see the residual plot for the newly transformed Y and check whether the residuals' variance are constant or not.

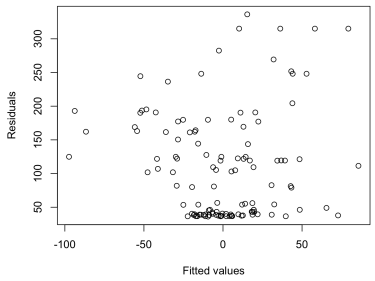

Figure 10: Residual against fitted values plot of the transformed data.

The plot above indicates that now, the residuals have constant variance.

3.3.2 Box-Cox Transformation

From the section about adequacy checking, we observed that the model is not linear, error terms are not normally distributed, and the variance of the residuals is not constant. Another way to overcome these violation is by transforming via Box-Cox transformation.

Assuming Y is a random variable from some distribution that may depend on the predictor variables and Y takes on only positive values, the Box-Cox transformation model is defined as:

$$Y^* = \frac{Y^\lambda - 1}{\lambda}, \;\; \lambda \neq 0$$

In Box-Cox tranformation, we must find $\hat{\lambda}$ that maximizes the likelihood estimator of λ. The graph below shows how the log-likelihood when λ varies.

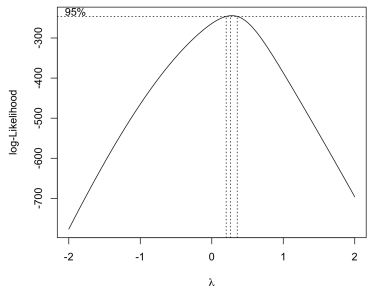

Figure 11: Box-Cox Log-likelihood against lambda.

And we tabulate λ and its corresponding log-likelihood in the table below. Notice that the values of log-likelihood are sorted in a descending manner.

	lambda	lik
[1,]	0.2626263	-244.3579
[2,]	0.3030303	-244.4070
[3,]	0.2222222	-245.3194
[4,]	0.3434343	-245.5623
[5,]	0.1818182	-247.1826
[6,]	0.3838384	-247.8787
[7,]	0.1414141	-249.8407
[8,]	0.4242424	-251.3987
[9,]	0.1010101	-253.1889
[10,]	0.4646465	-256.1089

Figure 12: Box-Cox Lambda Table.

We see that λ that maximizes log-likelihood is $\lambda = 0.2626263$. Thus, we try to fit our data to the following model:

$$\frac{Y^{\lambda} - 1}{\lambda} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_3 + \beta_4 * X_4$$

$$where\ \lambda = 0.2626263$$

And we obtained the following result:

OLS Regression Results

Dep. Variable:	np.true_divide((np.power(aadt, 0.2626263) - 1), 0.262626263)	**R-squared:**	0.837
Model:	OLS	**Adj. R-squared:**	0.831
Method:	Least Squares	**F-statistic:**	149.0
Date:	Thu, 24 Oct 2019	**Prob (F-statistic):**	9.69e-45
Time:	10:25:46	**Log-Likelihood:**	-406.03
No. Observations:	121	**AIC:**	822.1
Df Residuals:	116	**BIC:**	836.0
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	27.4608	5.388	5.096	0.000	16.788	38.133
population	1.832e-05	2.18e-06	8.401	0.000	1.4e-05	2.26e-05
num_lanes	7.1541	0.709	10.090	0.000	5.750	8.558
width_road	0.0214	0.058	0.371	0.711	-0.093	0.135
access_control	-9.5535	2.094	-4.563	0.000	-13.701	-5.406

Omnibus:	0.987	**Durbin-Watson:**	1.556
Prob(Omnibus):	0.610	**Jarque-Bera (JB):**	0.548
Skew:	-0.010	**Prob(JB):**	0.760
Kurtosis:	3.329	**Cond. No.**	3.75e+06

Figure 13: Box-Cox OLS Regression result.

The adjusted R-squared is lower than what we have obtained by transforming Y to $\sqrt{Y}$.

Furthermore, we again plot the residuals in the QQ-plot and observe that the blue dots approximately follow the red line. As compared to the QQ-plot of the $\sqrt{Y}$, this QQ-plot provides a better result. The new QQ-plot suggests that the residuals are almost normally distributed.

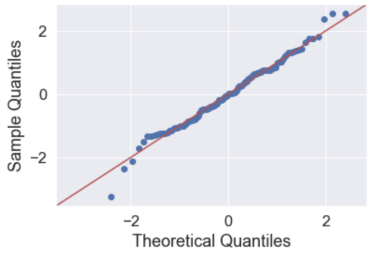

Figure 14: Box-Cox QQ-plot.

Next, we observe the residual plot for the newly transformed Y and check whether the residuals' variance are constant or not.

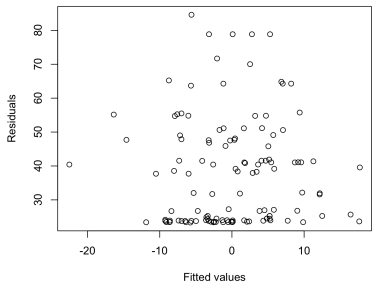

Figure 15: Box-Cox Residuals Plot.

There is no systematic found in this graph and therefore the variance of the error terms is constant.

3.3.3 Comparison between the methods

In both methods, we obtained approximately normally distributed residuals and constant error variance. However, Box-Cox method produced lower adjusted R-squared (0.831), while the square root method produced higher adjusted R-squared (0.852). This difference suggests us to choose square root method over the Box-Cox method for this data.

3.4 Check for sequential dependence or Autocorrelation

We will use Durbin Watson test to check the possible sequential dependence. The test statistic is:

$$d = \sum_{u=2}^{n} (e_u - e_{u-1})^2 / \sum_{u=1}^{n} e_u^2$$

$H_0 : \rho = 0$
$H_1 : \rho > 0$
From the OLS Regression results above, $d = 1.583 < d_L$. Thus, we reject H_0 and conclude that there is positively serially correlated. It implies that there is little autocorrelation in the data.
After we have conducted adequacy check for the newly transformed model, we see that the model is now appropriate to predict the response variable.

3.5 Check for Multicollinearity with VIF

We aimed to check whether the predictor variables are correlated to each other, or in other words, multicollinearity is present in our data. A method that is widely accepted to detect multicollinearity is variance inflation factors or VIF (Kutner et al., 2011, p.408). We obtained the following result:

	VIF Factor	features
0	2.0	population
1	5.5	num_lanes
2	8.6	width_road
3	6.6	access_control

Figure 16: Variance Inflation Factors.

As suggested by Kutner (2011, p.409), if the largest VIF values among all X_i exceed 10, it is an indication that multicollinearity may be unduly influencing the least squares estimates. However, we observe in the VIF table above, the largest VIF is 8.6. Thus, multicollinearity may be present in our data but does not influence our least squares estimates severely. To conclude, we may ignore this and proceed with our transformed model.

4 F-test for Reduced Model and Full Model

Previously on section 3.1.1, we concluded that β_3 are not statistically significant from 0, thus it indicates that X_3 may not be a suitable predictor to predict aadt. This indicates that we shall reduce the model. The reduced model is denoted

by ω and the full model is denoted by Ω.

$$\omega : \sqrt{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta 4 X_4$$

$$\Omega : \sqrt{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta 4 X_4$$

By fitting the data to the reduced model, we obtained the following OLS regression result:

Dep. Variable:	np.sqrt(aadt)	**R-squared:**	0.856
Model:	OLS	**Adj. R-squared:**	0.853
Method:	Least Squares	**F-statistic:**	232.7
Date:	Mon, 21 Oct 2019	**Prob (F-statistic):**	3.84e-49
Time:	12:39:46	**Log-Likelihood:**	-592.30
No. Observations:	121	**AIC:**	1193.
Df Residuals:	117	**BIC:**	1204.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	84.8372	24.684	3.437	0.001	35.952	133.722
population	9.579e-05	1e-05	9.557	0.000	7.59e-05	0.000
num_lanes	32.8589	3.266	10.060	0.000	26.390	39.328
access_control	-57.2167	9.604	-5.957	0.000	-76.238	-38.196

Omnibus:	3.617	**Durbin-Watson:**	1.582
Prob(Omnibus):	0.164	**Jarque-Bera (JB):**	3.661
Skew:	-0.173	**Prob(JB):**	0.160
Kurtosis:	3.779	**Cond. No.**	3.72e+06

Figure 17: OLS Regression result of the reduced model.

Furthermore, we conduct F-test of ω against Ω:

$$H_0 : \beta_{q+1} = ... = \beta_{q+1} = 0$$

$$H_1 : not\ all\ \beta_j, q+1, ..., p\ equal\ zero.$$

The result is shown in the table below:

```
Analysis of Variance Table

Model 1: sqrt(y) ~ x1 + x2 + x4
Model 2: sqrt(y) ~ x1 + x2 + x3 + x4
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    117 126502
2    116 126205  1    297.08 0.2731 0.6023
```

Figure 18: F-test for reduced model against full model.

$F = 0.2731 < 3.92 = F_{1,116}^{0.05}$. Since $F < F_{1,116}^{0.05}$, we do not reject hypothesis. This indicates that using the reduced model is sufficient and therefore the reduced model is significant. Moreover, it supports our previous analysis that β_3 is not statistically different from 0 and X_3 does not exhibit a strong linear relationship with our response variable. Thus, we should not include X_3 in our model. To conclude, the final model that is appropriate for predicting aadt is given by:

$$\sqrt{Y} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \beta_4 * X_4$$

Following are the normal probability plot and the residual plot of our final model:

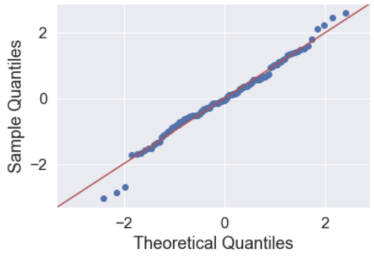

Figure 19: QQ-plot of the final model.

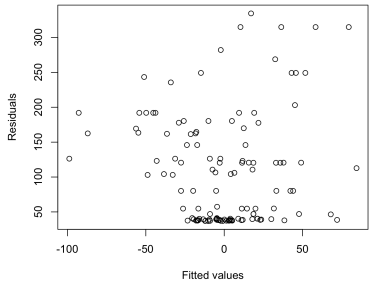

Figure 20: Residuals against fitted values for final model plot.

5 Prediction

We predict aadt from the data given in NTULearn by fitting the following data into our regression model:
$X_1 = 50000$, $X_2 = 3$, $X_3 = 60$, $X_4 = 2$

1. The 95% confidence interval is given by:

```
     fit      lwr      upr
73.77008 65.62839 81.91176
```

Figure 21: Confidence Interval.

Notice that Y in our regression model is $\sqrt{Y}$, thus the 95% confidence interval is given by:

fit	lwr	upr
5442.02323	4307.08557	6709.53643

Thus, with confidence coefficient 0.95, we estimate that the average annual daily traffic for a section of road or highway are between 4307 and 6710. We provide this confidence interval by taking into consideration population, number of lanes, and access control. It is when population equals to 50000, number of lanes equals to 3, and there is no access control.

2. The 95% prediction interval is given by:

```
     fit      lwr      upr
73.77008 8.142269 139.3979
```

Figure 22: Prediction Interval.

Notice that Y in our regression model is $\sqrt{Y}$, thus the 95% prediction interval is given by:

fit	lwr	upr
5442.02323	66.29654	19431.77452

Thus, with confidence coefficient 0.95, we estimate that the annual daily traffic for a section of road or highway are between 66 and 19432. We provide this confidence interval by taking into consideration population, number of lanes, and access control. It is when population equals to 50000, number of lanes equals to 3, and there is no access control.

6 Appendix

This assignment is written in Python and R programming language. We use R programming language in section 5 only.
Firstly, we all import important libraries and modules first.

```
#import all libraries
import pandas as pd
import numpy as np
import seaborn as sns
```

```
import matplotlib.pyplot as plt
from scipy import stats
import statsmodels.api as sm
import statsmodels.formula.api as smf
%matplotlib inline
from sklearn.linear_model import LinearRegression
from pandas.plotting import scatter_matrix
from sklearn.metrics import mean_squared_error
```

Next, we import the data and modify the data frame so that the data frame is suitable to be fitted into our regression model.

```
df = pd.read_fwf('aadt.txt', header = None)
df.columns = ["aadt", "population", "num_lanes", \
        "width_road", "access_control", "others_1", \
        "others_2", "others_3"]
del df['others_1']
del df['others_2']
del df['others_2']
df.head()
```

The output is:

	aadt	population	num_lanes	width_road	access_control
0	1616	13404	2	52	2
1	1329	52314	2	60	2
2	3933	30982	2	57	2
3	3786	25207	2	64	2
4	465	20594	2	40	2

Figure 23: df.head() output.

The code for figure 1:

```
scatter_matrix(df, alpha=1, figsize=(12, 12), \
            diagonal='kde')
```

The code for figure 2:

```
from sklearn.preprocessing import StandardScaler
stdsc = StandardScaler()
X_std = stdsc.fit_transform(df[cols].iloc[:,range(0,5)] \
        .values)
cov_mat = np.cov(X_std.T)
plt.figure(figsize=(10,10))
```

```
sns.set(font_scale=1.5)
hm = sns.heatmap(cov_mat,
                 cbar=True,
                 annot=True,
                 square=True,
                 fmt='.2f',
                 annot_kws={'size': 12},
                 cmap='coolwarm',
                 yticklabels=cols,
                 xticklabels=cols)
plt.title('Covariance matrix showing correlation \
           coefficients', size = 18)
plt.tight_layout()
plt.show()
```

The code for figure 3:

```
model = smf.ols("aadt ~ population + num_lanes \
       + width_road + access_control", data = df).fit()
model.summary()
```

Multiple Linear Regression model:

```
x = df[['population', 'num_lanes', 'width_road', \
    'access_control']]
y = df['aadt']
lm = LinearRegression()
lm.fit(x,y)
lm.intercept_
lm.coef_
yhat = lm.predict(x)
```

The code for figure 4:

```
res = model.resid
fig = sm.qqplot(res, fit = True, line = '45')
plt.show()
```

The code for figure 5 and 6 is written in R programming language:

```
raw <- read.table('Desktop/anaconda_files/aadt.txt',
         header=FALSE)
df <- data.frame(y=raw$V1,x1=raw$V2,x2=raw$V3,x3=raw$V4,
         x4=raw$V5)
mlr2 <- lm(y ~ x1+x2+x3+x4, data=df)
plot(residuals(mlr2),ylab='Residuals',xlab='Time')
plot(residuals(mlr2),fitted(mlr2),ylab='Residuals'
     ,xlab='Fitted values')
```

The code for figure 7:

```
# residual plot
fig, axs = plt.subplots(2,2, figsize =(20,20))

sns.residplot(df['population'], df['aadt'], \
                ax = axs[0,0])
sns.residplot(df['num_lanes'], df['aadt'], \
                ax = axs[0,1])
sns.residplot(df['width_road'], df['aadt'], \
                ax = axs[1,0])
sns.residplot(df['access_control'], df['aadt'], \
                ax = axs[1,1])

axs[0,0].set(ylabel = 'Residuals')
axs[0,1].set(ylabel = 'Residuals')
axs[1,0].set(ylabel = 'Residuals')
axs[1,1].set(ylabel = 'Residuals')

plt.show()
```

The code for figure 8:

```
tr_model = smf.ols("np.sqrt(aadt) ~ population + \
            num_lanes + width_road + access_control", \
            data = df).fit()
tr_model.summary()
```

The code for figure 9:

```
res = tr_model.resid
fig = sm.qqplot(res, fit = True, line = '45')
plt.show()
```

The code for figure 10 is written in R programming language:

```
mlr3 <- lm(sqrt(y) ~ x1+x2+x3+x4, data=df)
plot(residuals(mlr3),fitted(mlr3),ylab='Residuals'
      ,xlab='Fitted values')
```

The code for figure 11 and 12 is written in R programming language:

```
library(MASS)
b <- boxcox(y ~ x1+x2+x3+x4, data = df)
lambda <- b$x
lik <- b$y
bc <- cbind(lambda, lik)
sorted_bc <- bc[order(-lik),]
head(sorted_bc, n = 10)
```

The code for figure 13:

```
b_tr_model = smf.ols("np.true_divide((np.power(aadt, \
             0.2626263)-1), 0.262626263) ~ population \
             + num_lanes + width_road + access_control", \
             data = df).fit()
b_tr_model.summary()
```

The code for figure 14:

```
res = b_tr_model.resid
fig = sm.qqplot(res, fit = True, line = '45')
plt.show()
```

The code for figure 15 is written in R programming language:

```
mlr4 <- lm((y^0.2626263 - 1)/0.2626263 ~ x1+x2+x3+x4,
            data=df)
plot(residuals(mlr4),fitted(mlr4),ylab='Residuals'
     ,xlab='Fitted values')
```

The code for figure 16:

```
from patsy import dmatrices
from statsmodels.stats.outliers_influence \
        import variance_inflation_factor
vif = pd.DataFrame()
vif["VIF Factor"] = [variance_inflation_factor \
                     (x.values, i) for i in \
                     range(x.shape[1])]
vif["features"] = x.columns
vif.round(1)
```

The code for figure 17:

```
reduced_model = smf.ols("np.sqrt(aadt) ~ population \
                + num_lanes + access_control", \
                data = df).fit()
reduced_model.summary()
```

The code for figure 18 is written in R programming language:

```
mlr1 <- lm(sqrt(y) ~ x1+x2+x4, data=df)
mlr <- lm(sqrt(y) ~ x1+x2+x3+x4, data=df)
anova(mlr1,mlr)
```

The code for figure 19:

```
res = reduced_model.resid
fig = sm.qqplot(res, fit = True, line = '45')
plt.show()
```

The code for figure 20 is written in R programming language:

```
plot(residuals(mlr1),fitted(mlr1),ylab='Residuals'
     ,xlab='Fitted values')
```

The code for figure 21 and 22 is written in R programming language:

```
con <- data.frame(x1=50000,x2=3,x4=2)
predict(mlr1,con,interval='confidence',level=0.95)
predict(mlr1,con,interval='prediction',level=0.95)
```

7 References

Kutner, Neter, Nachtsheim, Wasserman. (2005). *Applied Linear Statistical Models*. New York: McGraw-Hill Education.

https://www.rdocumentation.org/packages/EnvStats/versions/2.3.1/topics/boxcox