STATISTICAL REPORT

ON

MULTI-LINEAR AND LOGISTIC REGRESSION

Submitted By,

Sumanth Bijadi Sridhar Rao

x18140181[Cohort B]

MSc in Data Analytics, 2018/19

Table of Contents

Multi-Linear Regression

Multiple linear regression is simply an advanced version of simple linear regression where we try to predict values one variable based on other variables commonly known as dependent and independent variables respectively. Also, this form of regression helps identify the individual contributions of independent variables associated with the overall fit of the model chosen.

DATA SOURCE

Eurostat is the source from which I have chosen the dataset to perform this particular regression upon. This dataset gives information on expenditure over labor in an industry. The link to the above mentioned dataset is provided below.

http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=lc_n08struc_r1&lang=en

OBJECTIVE

The main objective is to find the correlation between the Compensation attribute and rest of the underlying attributes in the chosen dataset. The variance in compensation of employees with respect to taxes, expenditures, training cost and subsidies is to be determined while we are at it.

ASSUMPTIONS

- Assumption #1: The dependent variable is such that it can be measured on a continuous scale. This assumption is true in case of this particular dataset as the dependent variable is measured in terms of percentage and is observed to be continuous in nature. The dependent variable is Total Compensation as shown in the figure below.

	TotalCompensation	TrainingCost	Expenditure	Subsidies	ZPR_1	ZRE_1	SRE_1	COO_1	LEV_1
1	98.49	.77	.49	.34	-.24325	.26583	.27274	.00098	.00461
2	98.48	.78	.50	.34	-.26765	.28009	.28744	.00110	.00503
3	98.47	.78	.49	.35	-.25099	.25510	.26182	.00091	.00518
4	98.49	.79	.45	.36	-.23217	.25524	.26230	.00096	.00766
5	98.33	.85	.51	.42	-.31845	.19275	.19854	.00060	.01195
6	98.32	.85	.51	.42	-.31845	.18369	.18920	.00054	.01195
7	98.32	.85	.51	.42	-.31845	.18369	.18920	.00054	.01195
8	101.06	.32	.21	1.59	1.99458	.45533	.68045	.14276	.50678
9	99.61	.19	.39	.27	.78320	.29948	.31664	.00296	.06002
10	98.91	.58	.65	.17	-.18236	.58812	.61179	.00769	.03044
11	98.97	.45	1.22	1.61	1.11314	-.59564	-.81459	.14438	.41988
12	99.55	.42	.18	.21	.45397	.55976	.58338	.00733	.03387
13	100.09	.16	.15	.49	1.21629	.32040	.34179	.00403	.07579
14	95.15	2.12	.10	.13	-2.59262	-.51463	-1.03592	.81879	.70775
15	99.50	.26	.22	.00	.50116	.46937	.50338	.00951	.08512
16	99.67	.17	.15	.02	.72754	.40702	.44092	.00843	.10240
17	96.92	1.30	2.20	.58	-2.07946	.59842	1.05492	.58635	.63275
18	97.30	.64	.34	.43	.17744	-1.21427	-1.24940	.02291	.00999
19	99.14	.62	.27	.04	-.13138	.74776	.78072	.01373	.03719
20	96.80	.85	1.18	.35	-.83344	-.70112	-.76239	.02651	.10882
21	95.05	.28	.39	.11	.46300	-3.52549	-3.73219	.42032	.06225
22	99.35	.32	.35	.03	.33834	.48909	.51963	.00869	.06865

- Assumption #2: It is required to have two or more independent variables, continuous or otherwise. This assumption is also legit as there exist more than two continuous independent attributes within the dataset namely, Training Cost, Expenditure and Subsidies.

- Assumption #3: Independence of residuals is a must for a multi linear regression model.

  The above assumption is tested and result from Durbin-Watson is considered as proof for the assumption made. We see that the Durbin-Watson score here is 2.112 which is between the usual limits of 0 and 4. Rule of thumb followed here is that test statistic values in the range 1.5 to 2.5 are relatively normal. Hence, residuals are independent as assumed.

Model Summary[b]

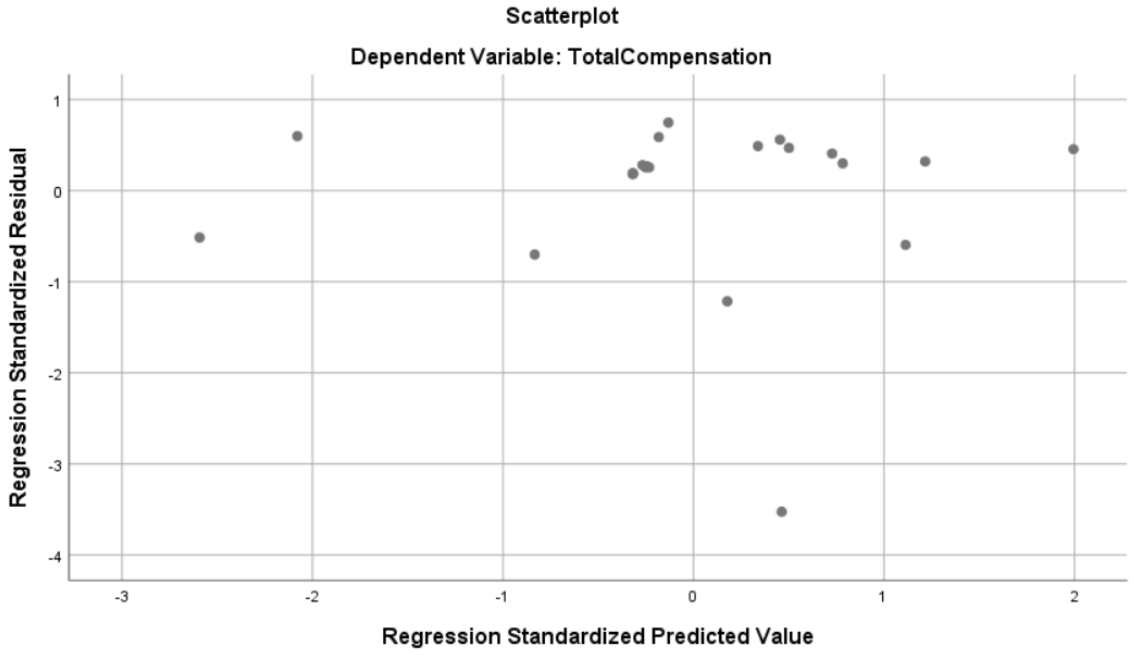
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.718[a]	.516	.435	1.10386	.516	6.394	3	18	.004	2.112

a. Predictors: (Constant), Subsidies, TrainingCost, Expenditure

b. Dependent Variable: TotalCompensation

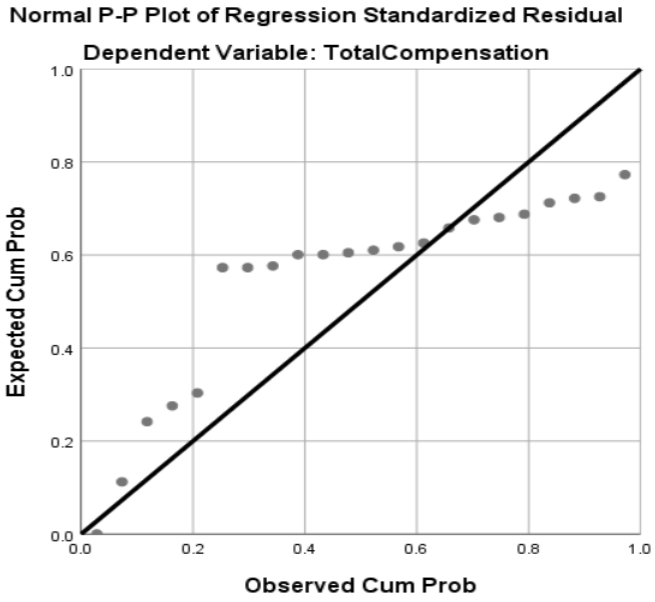
- Assumption #4: Relationship between dependent and every single independent variable is linear which is also true when they are collectively considered.

  Linearity check is carried out by examining the scatter plot obtained by plotting dependent variable against all the existing independent variables. To check for the assumption quoted above, it is important to note the points of existence of extreme values within the plot. In the figure below, we observe that the extreme values lie between -3 and -2 on the negative side and within 2 on the positive side. Therefore, based on the range given by Fidell and Tabachnick (Tabachnick, B.G. and Fidell, L.S. (2013) Using Multivariate Statistics. Pearson, Boston.), which is -3.3 to +3.3, it is safe to conclude that there exists a linear relationship.



- Assumption #5: Data ought to exhibit homoscedasticity wherein variance along the best fit line remains unchanged as we traverse further down the line.

  The above graph shows a P-P plot of standardized residual representing expected cumulative probability versus observed cumulative probability. Notice that the variance values along the best fit line lie not too close throughout but they do not lie too far from the line either. This demonstrates the homoscedastic nature of the data over which regression is performed.



- Assumption #6: Data is not multi-collinear, meaning no two independent variables are correlated.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	99.623	.502		198.284	.000					
	TrainingCost	-1.874	.579	-.564	-3.237	.005	-.651	-.607	-.531	.887	1.127
	Expenditure	-.700	.570	-.225	-1.230	.235	-.291	-.278	-.202	.802	1.246
	Subsidies	1.057	.609	.305	1.735	.100	.275	.378	.284	.870	1.149

a. Dependent Variable: TotalCompensation

From the Coefficients table above, it is clear that tolerance score for each of the independent variable is well above 0.1. Also, VIF is seen to be above 1 averaging around 1.174 which indicates a moderate correlation between predictors. This is not a very concerning score to worry about as a VIF score of about 5 to 10 indicates that correlation between predictors is high and may turn out to be problematic. Thus, assumption holds good for the chosen data.

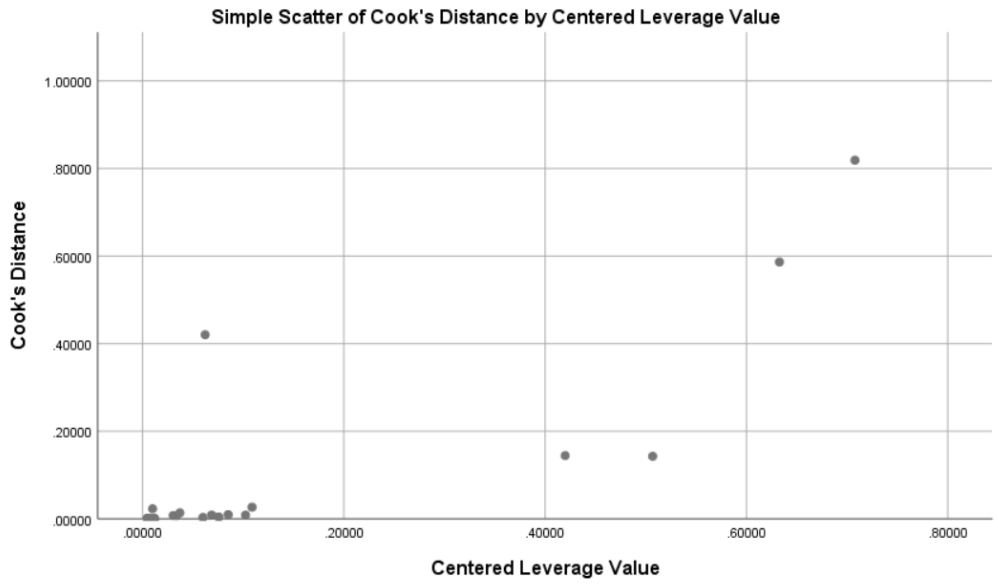
- Assumption #7: Observe the absence of significant outliers which in other words can be considered unusual values when performing multiple regression.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	95.7181	100.5574	98.4532	1.05496	22
Std. Predicted Value	-2.593	1.995	.000	1.000	22
Standard Error of Predicted Value	.247	.958	.416	.225	22
Adjusted Predicted Value	94.8672	100.1997	98.4709	1.08029	22
Residual	-3.89163	.82542	.00000	1.02197	22
Std. Residual	-3.525	.748	.000	.926	22
Stud. Residual	-3.732	1.055	-.005	1.025	22
Deleted Residual	-4.36135	2.05277	-.01769	1.34037	22
Stud. Deleted Residual	-7.627	1.058	-.188	1.765	22
Mahal. Distance	.097	14.863	2.864	4.529	22
Cook's Distance	.001	.819	.101	.219	22
Centered Leverage Value	.005	.708	.136	.216	22

a. Dependent Variable: TotalCompensation

To check for outliers, we look at the maximum and minimum Cook's distance from the residual statistics table above. The maximum value that the Cook's distance can take is 1 in order to consider the absence of significant outliers. Here, in this case, the minimum value is 0.001 and the maximum value is 0.819. Also, the scatter plot below shows that there are no major outliers when Cook's distance is plotted against Leverage value.



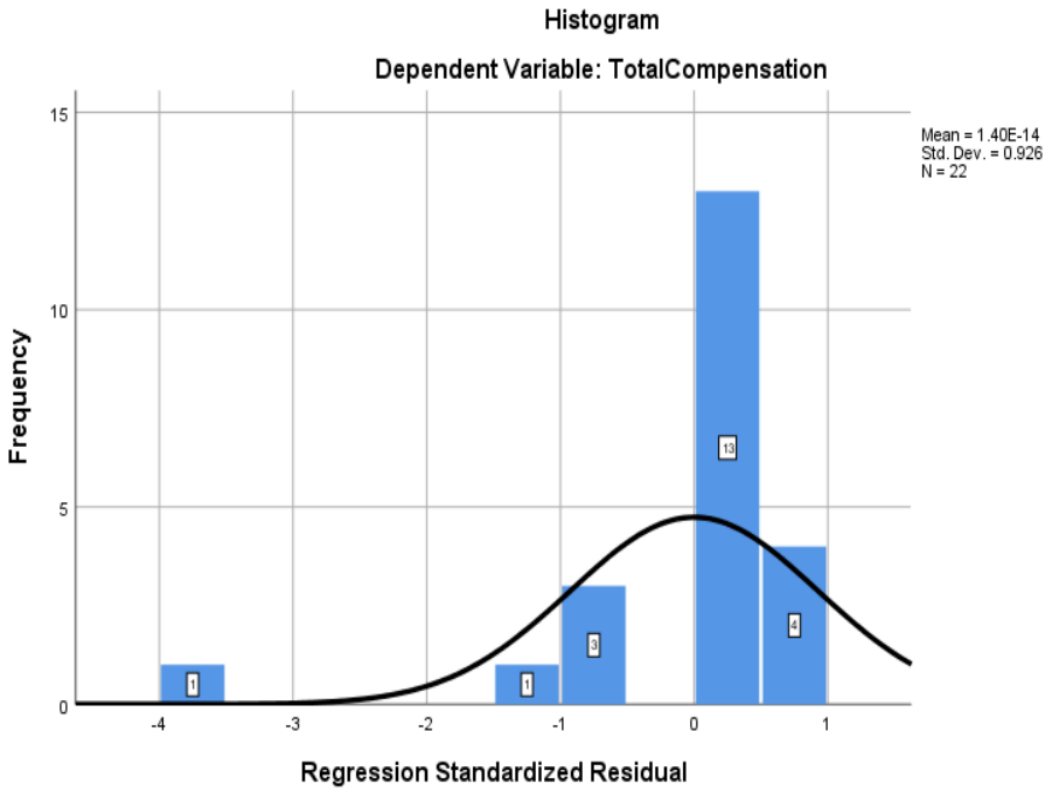
- Assumption #8: Residuals or errors are observed to be normally distributed.

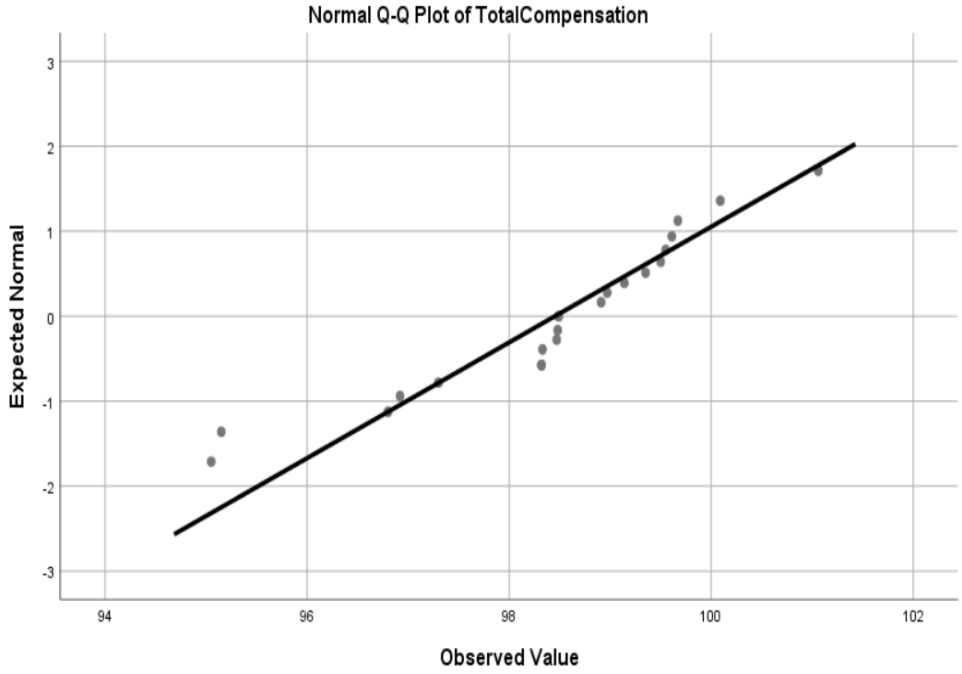
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
TotalCompensation	.237	22	.002	.914	22	.058

a. Lilliefors Significance Correction

Since the Shapiro-Wilk significance score from the normality test table is over the 0.005 benchmark which is .058 in this case, the assumption that the data is normally distributed is a correct one although we see skewness towards the negative side in the histogram below. Since the skewness is caused due to only one observation, it is negligible and can be considered to be almost normally distributed. Another way of testing for the same is by observing the Normal Q-Q plot. The data is said to be normal if all the values on that graph is not too dispersed from the best fit line.


Histogram
Dependent Variable: TotalCompensation
Mean = 1.40E-14
Std. Dev. = 0.926
N = 22
Frequency
Regression Standardized Residual


Normal Q-Q Plot of TotalCompensation
Expected Normal
Observed Value

ANALYSIS

Fitness of model

Let us determine how well our model fits. For this, it is important to look at the Model Summary table generated in SPSS during the regression process. The important attributes that have a significance in determining the fitness of model with respect to the data at hand are the R, R^2, adjusted R^2, and standard error of estimate.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.718[a]	.516	.435	1.10386	.516	6.394	3	18	.004	2.112

a. Predictors: (Constant), Subsidies, TrainingCost, Expenditure

b. Dependent Variable: TotalCompensation

Observe that the R, R^2, and Adjusted R^2 carry a value of 0.718, 0.516 and 0.435 respectively which are considered to be good scores. By observing the R score it is safe to say that the level of prediction as a result of the regression performed is of a good measure.

Statistical Significance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23.372	3	7.791	6.394	.004[b]
	Residual	21.933	18	1.218		
	Total	45.305	21			

a. Dependent Variable: TotalCompensation

b. Predictors: (Constant), Subsidies, TrainingCost, Expenditure

The F- ratio in the above Anova table tests the goodness of fit for the data of the overall regression model.

$F(3, 18) = 6.394, p < 0.0005$

Since, F- ratio is not that of a large number, we can conclude that the variance among group means is moderate. A large F ratio is seen either when the null hypothesis turns out to be wrong or that the random sampling process gave values that are exorbitant in some groups than others (Laerd Statistics, 2018).

Estimated Model Coefficients

The equation generally used to calculate the prediction of dependent variable which in our case is the Total Compensation variable from independent variables such as Training Cost, Expenditure and Subsidies is given by:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	99.623	.502		198.284	.000					
	TrainingCost	-1.874	.579	-.564	-3.237	.005	-.651	-.607	-.531	.887	1.127
	Expenditure	-.700	.570	-.225	-1.230	.235	-.291	-.278	-.202	.802	1.246
	Subsidies	1.057	.609	.305	1.735	.100	.275	.378	.284	.870	1.149

a. Dependent Variable: TotalCompensation

PDV = 99.623 – (1.874 * Training Cost) – (.700 * Expenditure) – (1.057 * Subsidies)

Unstandardized coefficients help in determining by how much the value varies when we consider dependent variable with respect to a particular independent variable when all other independent variables are left untouched.

With that in mind, it can analyzed from the above coefficients table that, for an increase in each unit of Training Cost, there is a drop of 1.874 percent in the Total Compensation. Similarly, an increase in expenditure by one unit will result in 0.7 percent decrease in Total Compensation and likewise with respect to change in other independent variables.

Statistical Significance of the Independent Variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	99.623	.502		198.284	.000					
	TrainingCost	-1.874	.579	-.564	-3.237	.005	-.651	-.607	-.531	.887	1.127
	Expenditure	-.700	.570	-.225	-1.230	.235	-.291	-.278	-.202	.802	1.246
	Subsidies	1.057	.609	.305	1.735	.100	.275	.378	.284	.870	1.149

a. Dependent Variable: TotalCompensation

Test for statistical significance involves the process of making sure that the standardized or unstandardized coefficients are equal to zero in the population or sample being examined. It is usually safe to conclude that any given coefficient is statistically significantly different to zero if the p value is observed to be less than 0.05. The p value is given by 'Sig' column seen in the above table. It is noticeable how only one of the independent variable bares a 5% risk of being significantly different while the others carry a much higher percentage of risk of conclusion.

RESULT

A multilinear regression was performed to predict Total Compensation from Training Cost, Expenditure and Subsidies. Below are the significant findings:

1. The model is a good fit with an acceptable prediction level indicated by the observed R score equal to 0.718.

2. $F(3,18) = 6.394$ with a p-value ('Sig') below 0.005.

3. $R^2 = 0.516$

4. Only one variable among the three added to statistically significant prediction of $p < 0.05$.

LOGISTIC REGRESSION

Logistic Regression also known as logit regression is a method used to predict dichotomous outcomes.

DATA SOURCE

The data obtained is from data.gov.uk and the link to this particular dataset is given below.

url: https://data.gov.uk/dataset/8c395d81-1158-4014-8900-8d590000b4b1/vehicle-testing-outcomes-by-test-centre/datafile/321b5ae8-50b9-44fc-a72e-a059a89ce0ec/preview

The dataset contains information about tests performed on various vehicles and the result of the performed tests which is either pass or fail.

OBJECTIVE

To perform logistic regression on the vehicle test dataset to determine factors such as goodness of fit of model, accuracy of prediction, variance and efficiency.

ASSUMPTIONS

Assumption #1: The dependent variable is measured on a dichotomous scale.

This assumption holds good for the data considered in this case as the dependent variable varies between only two values: pass and fail given by Test Details attribute.

Assumption #2: There exist one or more independent variables, continuous or categorical within the sample.

This assumption is true since there are four independent variables, 3 of which are continuous while the other is categorical; nominal to be more precise.

The figure below provides proof for the above assumptions being true.

	TestDetails	AllFullTests	CarsFullTests	HGVFullTests	Type.Of.Car
1	1	33677	27355	1555	1
2	0	8806	7103	399	2
3	1	54137	44698	1208	3
4	0	12438	10321	351	1
5	1	76720	64757	1629	1
6	0	13013	10829	285	2
7	1	46363	39473	948	2
8	0	11304	9478	342	3
9	1	42358	34650	1377	3
10	0	11124	8906	505	1
11	1	47934	37862	2291	1
12	0	11317	9547	310	3
13	1	28952	25229	426	3
14	0	7581	6401	198	2
15	1	32009	26005	865	2
16	0	8630	7113	323	3
17	1	30922	26240	845	2
18	0	8101	7051	150	2
19	1	49238	42452	1038	1
20	0	10470	8944	343	3
21	1	52343	44940	795	1
22	0	13179	10866	457	3
23	1	57484	44548	2819	2
24	0	12497	9992	543	1
25	1	50992	42619	1795	1
26	0	14818	11835	803	3
27	1	82381	73994	546	3
28	0	18308	16144	267	1
29	1	31727	26367	794	3
30	0	7568	6017	350	2
31	1	717237	601189	18931	2
32	0	169154	140547	5626	2

Assumption #3: Dependent variable has mutually exclusive and exhaustive categories while having independent observations. The assumption is true as we can observe from the above figure of the chosen data.

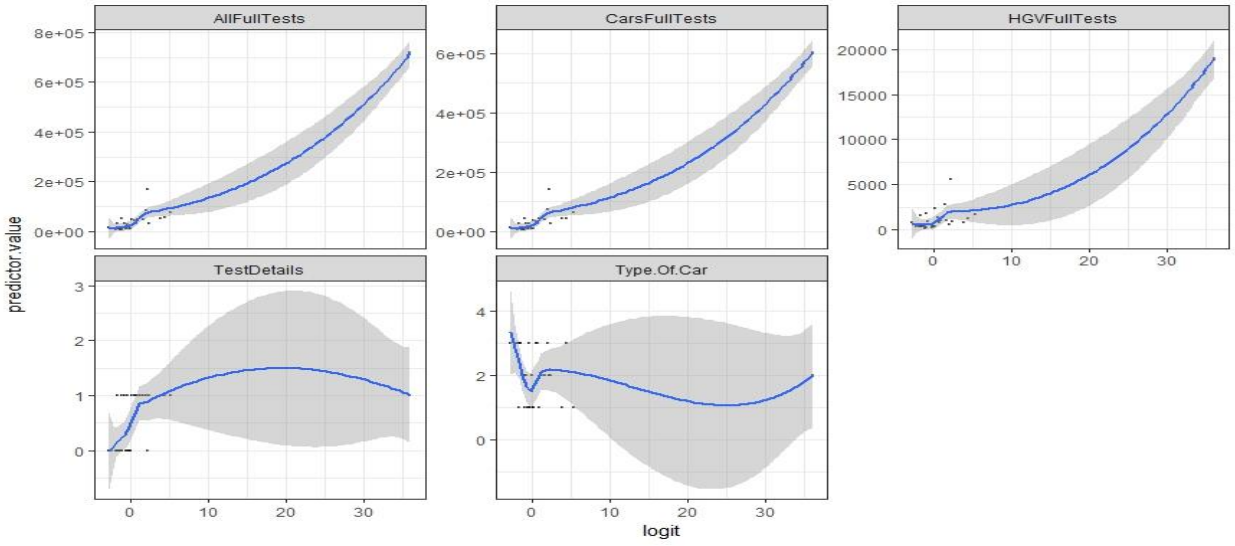
Assumption #4: There is a linear relationship between any continuous independent variables and the logit transformation of the dependent variable.

In order to test this assumption, the following R code is used to plot graphs of independent variables against predictor values.

```
library(tidyverse)

library(broom)
```

```
theme_set(theme_classic())
# Select only numeric predictors
data <- blr %>%
  dplyr::select_if(is.numeric)
predictor <- colnames(data)
# Binding the logit and tidying the data for plot
data <- data %>%
  mutate(logit = log(p/(1-p))) %>%
  gather(key = "predictor", value = "predictor.value", -logit)
#plotting linear graphs
ggplot(data, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

From the graphs below, it can be observed that independent variables, AllFullTests, CarsFullTests and HGVFullTests exhibit a linear behavior when plotted against logit transformation of the dependent variable TestDetails.



ANALYSIS

Execution of Output on RStudio:

Goodness of Fit Test:

The model is tested for goodness of fit using the glm function in R as:

```
library('aod')
library('ggplot2')
library('pscl')
library('ROCR')
logregression <- glm(TestDetails ~ AllFullTests + CarsFullTests
+ HGVFullTests + Type.Of.Car, family = binomial, data = blr)
```

Running the above code and extracting its summary gives us information such as Deviance Residuals, coefficients and significance codes. Since the regression being performed is binomial in nature, the dispersion parameter for family is considered to be 1.

Below shown is the output from RStudio giving the summary of the glm function which helped determine the extent to which model is fit.

```
> summary(logregression)

Call:
glm(formula = TestDetails ~ AllFullTests + CarsFullTests + HGVFullTests +
    Type.Of.Car, family = binomial, data = blr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1539  -0.7350  -0.1762   0.7176   1.9147

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.067558   1.343789  -0.050   0.9599
AllFullTests   0.002429   0.001282   1.895   0.0581 .
CarsFullTests -0.002594   0.001399  -1.855   0.0636 .
HGVFullTests  -0.007576   0.003816  -1.985   0.0471 *
Type.Of.Car   -0.626160   0.612892  -1.022   0.3069
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 44.361  on 31  degrees of freedom
Residual deviance: 29.771  on 27  degrees of freedom
AIC: 39.771

Number of Fisher Scoring iterations: 7
```

Also, the McFadden score can be used to determine statistical fitness of the model. The McFadden score can be obtained in R by using pR2 function as follows:

```
pR2(logregression)
```

```
> pR2(logregression)
        llh      llhNull           G2     McFadden         r2ML         r2CU
-14.8854612 -22.1807098  14.5904972    0.3289006    0.3661564    0.4882085
```

Variance Test:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	28.176[a]	.397	.529

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

This table contains methods of variance. The values associated are referred to as pseudo R^2 values. The variation that the data presents is from 39.7% to 52.9% based on Cox & Snell R Square and Nagelkerke R Square attributes respectively.

Prediction of Category:

The cut value chosen during the prediction process is .500 as seen in the table below. This means that a particular case is classified under a dependent category if the probability of that category being true is above .500.

Classification Table[a]

Observed			Predicted TestDetails 0	Predicted TestDetails 1	Percentage Correct
Step 1	TestDetails	0	13	3	81.3
		1	3	13	81.3
	Overall Percentage				81.3

a. The cut value is .500

Here, the TestDetails can bare the categorical values 0 and 1 which indicate either pass or fail scenario and it is evident that the percentage of the categorical value being correct during prediction is 81.3% for both categories.

Variables in Equation:

Wald Test is chosen in order to determine the variables part of the equation which depicts the contribution of each independent variable with respect to the model and the significance of that model.

The R code used to undertake this test is as given below:

```
wald.test(b = coef(logregression), Sigma = vcov(logregression), Terms = 3:4)
```

```
> wald.test(b = coef(logregression), Sigma = vcov(logregression), Terms = 3:4)
Wald test:
----------

Chi-squared test:
X2 = 4.0, df = 2, P(> X2) = 0.14
```

The following are observed:

1. X^2 (at df = 2) = 4.0
2. Significance / p is observed to be 0.14

Since, the Wald Test performed in this fashion on RStudio cannot show the significance of each independent variable individually rather than collectively, anova function in R is used. For the sake of obtaining extra details, the following table was obtained from SPSS.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	AllFullTests	.003	.001	4.567	1	.033	1.003	1.000	1.006
	CarsFullTests	-.003	.001	4.419	1	.036	.997	.994	1.000
	HGVFullTests	-.009	.004	4.878	1	.027	.991	.984	.999
	TypeOfCar			2.405	2	.300			
	TypeOfCar(1)	1.253	1.205	1.082	1	.298	3.503	.330	37.177
	TypeOfCar(2)	-.727	1.286	.320	1	.572	.483	.039	6.007
	Constant	-1.653	.960	2.961	1	.085	.192		

a. Variable(s) entered on step 1: AllFullTests, CarsFullTests, HGVFullTests, TypeOfCar.

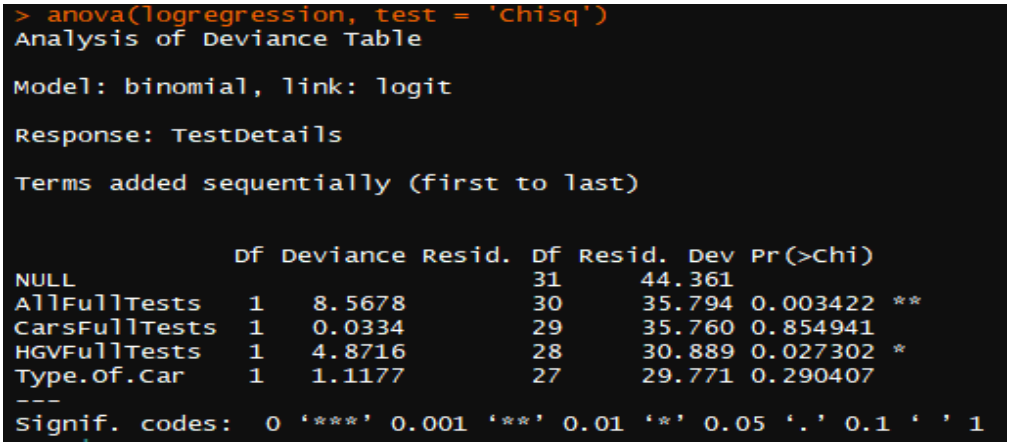
The following can be concluded by observing the above table:

- All independent variable don't add significantly to the model. To be clearer, three of the continuous variables exhibit significance below 0.05 while the categorical independent variable exhibits a significance of over 0.25, different for each category.
- Category 1 under Type of Car variable observes more success in fact 3.503 times more than it observes failure with respect to all the continuous independent variables.

The deviance is tested by checking the analysis of variance (ANOVA), which in this case is done on RStudio using the function anova:

```
anova(logregression, test = 'Chisq')
```

The output after executing the above code is shown below.



The residual deviance is observed to be less as the values do not vary by a large amount as given by Resid.Dev attribute in the above figure.

Model Efficiency:

In order to assess the efficiency of the model, the predictive capability of the model should be established. For this, execution of following code in R carried out.

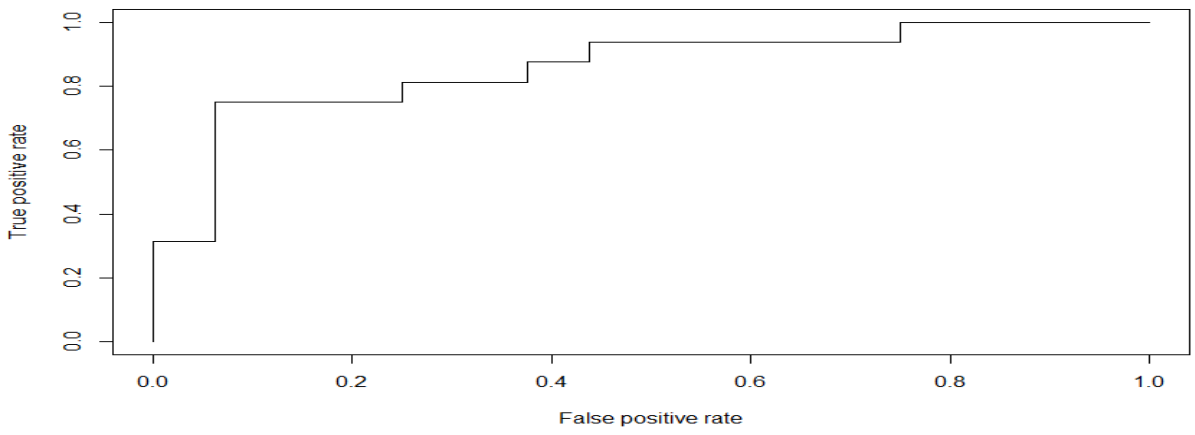
```
fitted.results <- predict(logregression, newdata = subset(blr,
select = c("TestDetails", "AllFullTests", "CarsFullTests",
"HGVFullTests", "Type.Of.Car")), type = "response")
fitted.results <- ifelse(fitted.results > 0.5,1,0)

misClasificError <- mean(fitted.results != blr$TestDetails)
print(paste('Accuracy', 1- misClasificError))
p <- predict(logregression, newdata = subset(blr, select =
c("TestDetails", "AllFullTests", "CarsFullTests",
"HGVFullTests", "Type.Of.Car")), type = "response")
pr <- prediction(p, blr$TestDetails)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

The figure below shows the results produced by the execution of this code:

```
> print(paste('Accuracy', 1- misClasificError))
[1] "Accuracy 0.78125"
```

Accuracy, hence obtained gives the predictive capability of the model. 0.78125 is considered to be a good score.

The auc function from the above code gives the area under the Receiver Operating Characteristic which is seen to be 0.859375 and the plot(prf) statement above gives the graph representing the ROC curve by plotting true positives versus false positives. The graph is as shown below.



RESULT

By performing logistic regression on the chosen data as such, the following were determined.

1. McFadden score = 0.3289
2. Cox & Snell R Square score = 0.397
3. Nagelkerke R Square score = 0.529
4. Accuracy of prediction = 0.78125

REFERENCES

1. Laerd Statistics, https://statistics.laerd.com/spss-tutorials/binomial-logistic-regression-using-spss-statistics.php
2. Tabachnick, B.G. and Fidell, L.S. (2013) “Using Multivariate Statistics”, Pearson, Boston.
3. STHDA, http://www.sthda.com/english/articles/36-classification-methods-essentials/148-logistic-regression-assumptions-and-diagnostics-in-r/
4. Statistics Solutions, https://www.statisticssolutions.com/assumptions-of-logistic-regression/
5. Data Science Plus, https://datascienceplus.com/perform-logistic-regression-in-r/
6. Institute for Digital Research and Education, https://stats.idre.ucla.edu/r/dae/logit-regression/