COMPSCI- 5590-0015

ECONOMETRICS OF DATA SCIENCE
HOMEWORK-3

Name: Mohammad Shaik

Student id: 16319967

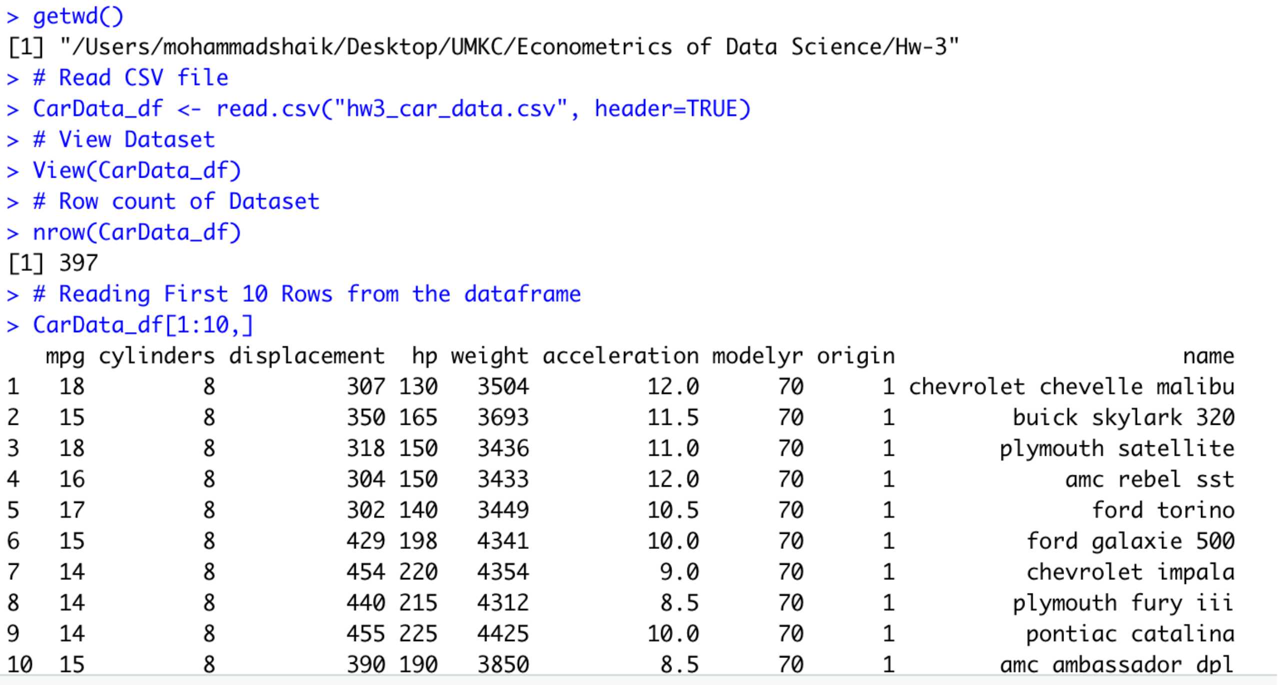
I. Estimate a logistic regression model with foreign as a function of selected variables.

1.Generate summary statistics for these variables (including min, max, mean, standard deviation, median, 25th and 75th quartiles) and interpret the results.
Ans:

To read the data, first I need to check whether the dataset located in the current working directory. Use getwd() function which returns the current Working Directory in fetching dataset.

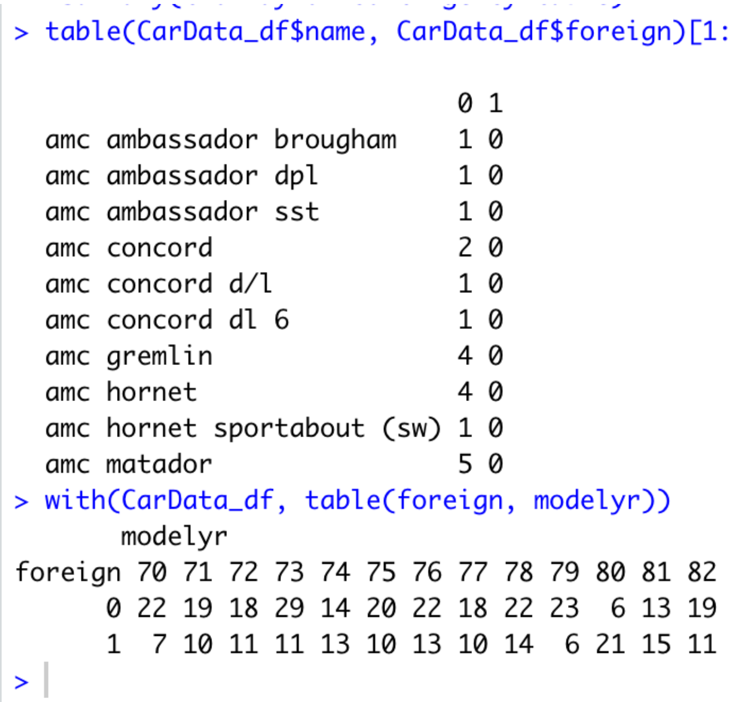
Note: It's not mandatory to use getwd() function, to find the dataset. Instead we can create the project folder having the dataset added manually.

Initially, I have read the "hw3_car_data.csv "dataset into the dataframe 'data' by using read.csv() function and read the first 10 Rows from the dataframe.



```
> getwd()
[1] "/Users/mohammadshaik/Desktop/UMKC/Econometrics of Data Science/Hw-3"
> # Read CSV file
> CarData_df <- read.csv("hw3_car_data.csv", header=TRUE)
> # View Dataset
> View(CarData_df)
> # Row count of Dataset
> nrow(CarData_df)
[1] 397
> # Reading First 10 Rows from the dataframe
> CarData_df[1:10,]
   mpg cylinders displacement  hp weight acceleration modelyr origin                      name
1   18         8          307 130   3504         12.0      70      1 chevrolet chevelle malibu
2   15         8          350 165   3693         11.5      70      1         buick skylark 320
3   18         8          318 150   3436         11.0      70      1        plymouth satellite
4   16         8          304 150   3433         12.0      70      1             amc rebel sst
5   17         8          302 140   3449         10.5      70      1               ford torino
6   15         8          429 198   4341         10.0      70      1          ford galaxie 500
7   14         8          454 220   4354          9.0      70      1          chevrolet impala
8   14         8          440 215   4312          8.5      70      1         plymouth fury iii
9   14         8          455 225   4425         10.0      70      1          pontiac catalina
10  15         8          390 190   3850          8.5      70      1         amc ambassador dpl
```

Also shown the distribution of foreign cars for the model years and frequency distribution of number of cars belonging to foreign or non-foreign.



```
> table(CarData_df$name, CarData_df$foreign)[1:10,]

                            0 1
  amc ambassador brougham   1 0
  amc ambassador dpl        1 0
  amc ambassador sst        1 0
  amc concord               2 0
  amc concord d/l           1 0
  amc concord dl 6          1 0
  amc gremlin               4 0
  amc hornet                4 0
  amc hornet sportabout (sw) 1 0
  amc matador               5 0
> with(CarData_df, table(foreign, modelyr))
       modelyr
foreign 70 71 72 73 74 75 76 77 78 79 80 81 82
      0 22 19 18 29 14 20 22 18 22 23  6 13 19
      1  7 10 11 11 13 10 13 10 14  6 21 15 11
> 
```

Basically, summary provides the statistics for each column of the entire dataset. Now to generate the summary statistics information for car data, I have used summary() function and then it returns the result as displayed in the console as in the below fig. Along with standard deviation by using apply().

mpg	cylinders	displacement	hp	weight	acceleration	modelyr	origin	foreign
Min. :9.00	Min. :3.000	Min. :68	Min. :46.0	Min. :1613	Min. : 8.00	Min. :70.00	Min. :1.000	Min. :0.0000
1st Qu.:17.50	1st Qu.:4.000	1st Qu.:98	1st Qu.:75.0	1st Qu.:2219	1st Qu.:13.80	1st Qu.:73.00	1st Qu.:1.000	1st Qu.:0.0000
Median :23.00	Median :4.000	Median :146	Median : 92.0	Median :2790	Median :15.50	Median :76.00	Median :1.000	Median :0.0000
Mean :23.52	Mean :5.453	Mean :193	Mean :104.1	Mean :2965	Mean :15.59	Mean :75.94	Mean :1.594	Mean :0.3829
3rd Qu.:29.00	3rd Qu.:8.000	3rd Qu.:262	3rd Qu.:125.0	3rd Qu.:3609	3rd Qu.:17.20	3rd Qu.:79.00	3rd Qu.:2.000	3rd Qu.:1.0000
Max. :46.60	Max. :8.000	Max. :455	Max. :230.0	Max. :5140	Max. :24.80	Max. :82.00	Max. :3.000	Max. :1.0000
S.dev :7.7893952	S.dev :1.7029297	S.dev :104.6902392	S.dev :38.4027282	S.dev :851.0898765	S.dev :2.7838445	S.dev :3.6840323	S.dev :0.8160446	S.dev :0.4867007

modyr70	**modyr71**	**modyr72**	**modyr73**	**modyr74**	**modyr75**	**modyr76**	**modyr77**	**modyr78**	**modyr79**	**modyr80**	**modyr81**	**modyr82**
Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.0000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000	Min. :0.00000
1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.0000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000
Median :0.00000	Median :0.00000	Median :0.00000	Median :0.0000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000	Median :0.00000
Mean :0.07305	Mean :0.07305	Mean :0.07305	Mean :0.1008	Mean :0.06801	Mean :0.07557	Mean :0.08816	Mean :0.07053	Mean :0.09068	Mean :0.07305	Mean :0.06801	Mean :0.07053	Mean :0.07557
3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.0000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000
Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.0000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000	Max. :1.00000
S.dev :0.2605434	S.dev :0.2605434	S.dev :0.2605434	S.dev :0.3013847	S.dev :0.2520809	S.dev :0.2646372	S.dev :0.2838870	S.dev:0.2563595	S.dev:0.2875160	S.dev:0.2605434	S.dev:0.2520809	S.dev:0.2563595	S.dev:0.2646372

From the above stats, there is a large discrepancy between the maximum and minimum values of most independent variables. ‘foreign’, and variables i.e., ‘displacement’, 'weight', 'horsepower' are terms used interchangeably. The mean and median in ‘mpg’, ‘cylinders’, ‘acceleration’, ‘modelyr’, ‘origin’ are nearly identical. Furthermore, the stats for model years are ranging from 0-1 as it has values only in 0-1 and 1980’s stats are mostly similar to each other.

2.State and justify your binary logit model explaining “foreign”.

Ans.

As given, the model built on this dataset is a Binary Logit Model by considering foreign variable as a function. Since the dataset is having more than two independent variables, the model to be build is either Multilinear Regression Model or Binary Logit Model, but the foreign attribute is a categorical variable having only values 0 or 1. Hence we conclude that Binary Logit Model need to build. Therefore, the Model build on the top of the car dataset with foreign as a function is as follows:

foreign = β0+ β1mpg + β2cylinders + β3displacement +β4hp+β5weight +β6acceleration +β7modelyr+β8origin +β9modyr70 + β10modyr71 +…………+β21 modyr82+u

From the above equation,

foreign is a Dependent variable

mpg, cylinders, displacement, hp, ..., modyr82 are Independent variables.

β0, β1, β2, β3, β4, ..., β21 are the coefficients.

u is a y-intercept.

For the dataset, I had built 5 models and check the significant level for each independent variables towards the foreign variable.

bin_logit_model1:
Considered all the independent variables

foreign = β0+ β1mpg + β2cylinders + β3displacement +β4hp+β5weight +β6acceleration +β7modelyr+β8origin +u

bin_logit_model2:

foreign = β0+ β1mpg + β2cylinders + β3displacement +β4hp+β5weight +β6acceleration+β7modelyr+β8origin +β9modyr70 + β10modyr71 +…………+β21 modyr82+u

bin_logit_model3

foreign = β0+ β1mpg + β2cylinders + β3displacement +β4hp+β5weight +β6acceleration + β9modyr80 + β10modyr81+β11 modyr82+u

bin_logit_model4:

foreign = β0+ β1mpg + β2cylinders + β3displacement +β4hp+β5weight +β6acceleration +β7modyr70 + β8modyr71 +…………+β19 modyr82+u

bin_logit_model5:

foreign = β0+ β1mpg + β2cylinders + β3displacement +β4weight +β5modyr70 + β6modyr71 +…………+β16modyr80+u

3.Estimate a binary logit model that includes binary variables for model year.

Ans:

The logistic regression model is build by calling the glm() function with parameters function string , family argument and data argument.

The string parameter(formula) we passed is the different variables as shown in the models and dataset we passed is filtered dataset and family is set to binomial

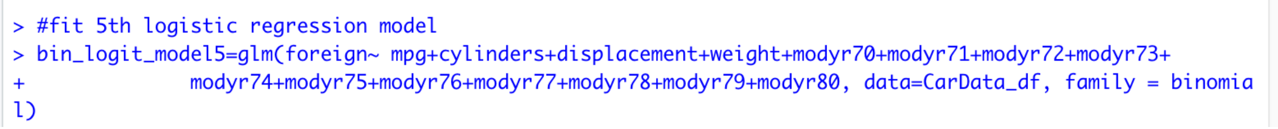
Finally, after building several models, the below model is the final model

The binary logit model equation:

bin_logit_model5:

$foreign = \beta0+ \beta1mpg + \beta2cylinders + \beta3displacement + \beta4weight + \beta5modyr70 + \beta6modyr71 + \ldots\ldots\ldots\ldots + \beta16modyr80 + u$

The following is the model build with the parameters as shown below:



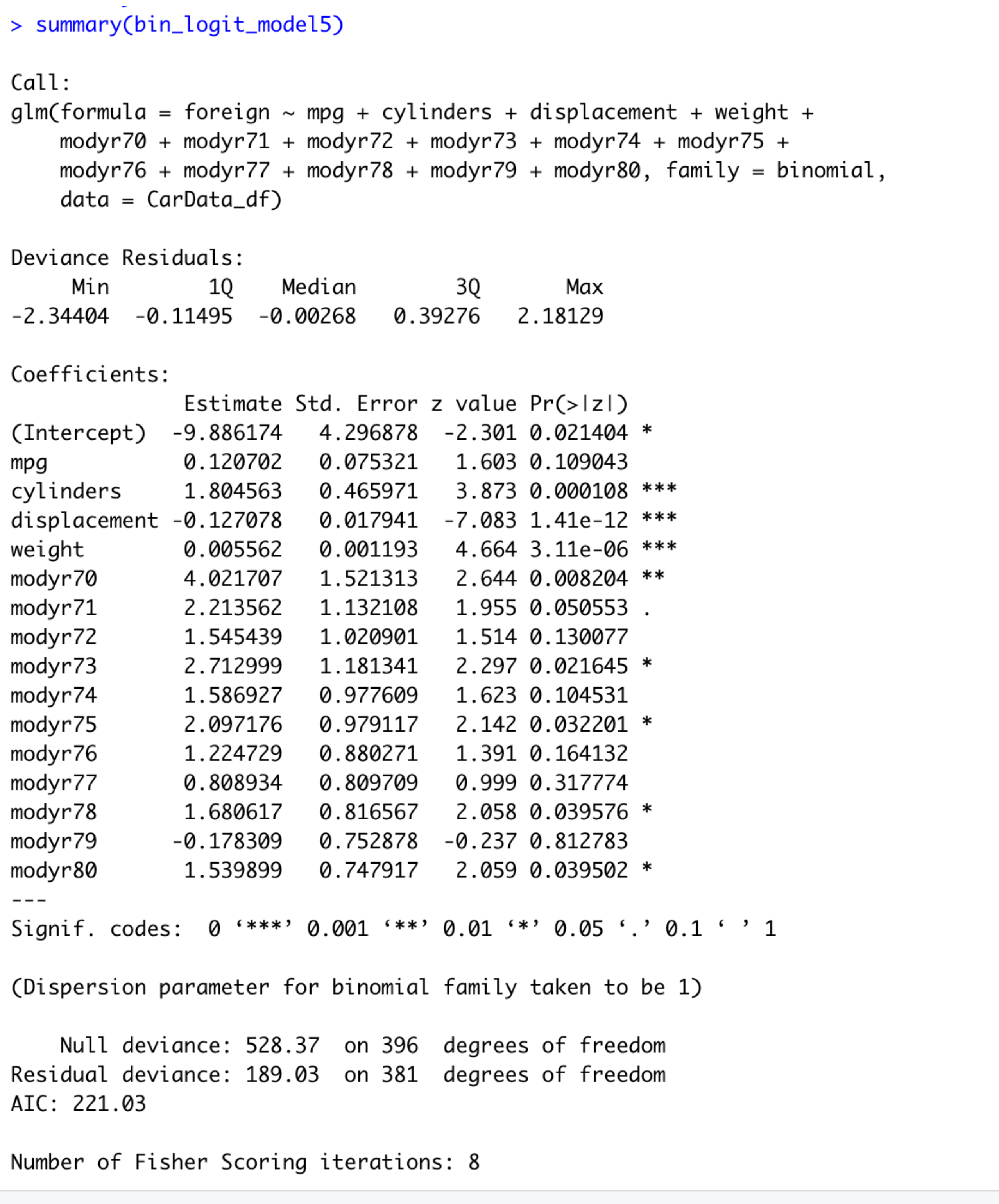
```
> #fit 5th logistic regression model
> bin_logit_model5=glm(foreign~ mpg+cylinders+displacement+weight+modyr70+modyr71+modyr72+modyr73+
+              modyr74+modyr75+modyr76+modyr77+modyr78+modyr79+modyr80, data=CarData_df, family = binomia
l)
```

4. Summarize your results in a table and interpret them (odds ratios are convenient).

Ans:

After building the model for bin_logit_model5 function explained above.

The summary() prints the summary statistics of the fitted model as shown below:



```
> summary(bin_logit_model5)

Call:
glm(formula = foreign ~ mpg + cylinders + displacement + weight +
    modyr70 + modyr71 + modyr72 + modyr73 + modyr74 + modyr75 +
    modyr76 + modyr77 + modyr78 + modyr79 + modyr80, family = binomial,
    data = CarData_df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.34404  -0.11495  -0.00268   0.39276   2.18129

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.886174   4.296878  -2.301 0.021404 *
mpg           0.120702   0.075321   1.603 0.109043
cylinders     1.804563   0.465971   3.873 0.000108 ***
displacement -0.127078   0.017941  -7.083 1.41e-12 ***
weight        0.005562   0.001193   4.664 3.11e-06 ***
modyr70       4.021707   1.521313   2.644 0.008204 **
modyr71       2.213562   1.132108   1.955 0.050553 .
modyr72       1.545439   1.020901   1.514 0.130077
modyr73       2.712999   1.181341   2.297 0.021645 *
modyr74       1.586927   0.977609   1.623 0.104531
modyr75       2.097176   0.979117   2.142 0.032201 *
modyr76       1.224729   0.880271   1.391 0.164132
modyr77       0.808934   0.809709   0.999 0.317774
modyr78       1.680617   0.816567   2.058 0.039576 *
modyr79      -0.178309   0.752878  -0.237 0.812783
modyr80       1.539899   0.747917   2.059 0.039502 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 528.37  on 396  degrees of freedom
Residual deviance: 189.03  on 381  degrees of freedom
AIC: 221.03

Number of Fisher Scoring iterations: 8
```

From the model summary, we can see that the coefficients cylinders, displacement, weight, are highly significant at 0.1% level of significance level. Therefore, the Null-Hypothesis can be rejected where the model can bring the insights from these variables more accurately. Similarly, modyr73, modyr75, modyr78, modyr80 are also highly significant at 1% level of significance level and displacement, foreign are significant at 5% level of significance level. But also the other independent variables contribute towards the target variable as per model build. Moreover, these independent variables as per the alternative hypothesis is that they are jointly significant to each other.

Odds ratio:

	x
(Intercept)	5.08732106982388E-05
mpg	1.12828909149835
cylinders	6.07731631451136
displacement	0.88066505807778
weight	1.00557754554781
modyr70	55.7962599411742
modyr71	9.14824561749666
modyr72	4.69003028652147
modyr73	15.0744226038043
modyr74	4.88870424899546
modyr75	8.14313815301791
modyr76	3.40324283321631
modyr77	2.24551230395788
modyr78	5.36886704737011
modyr79	0.83668345530853
modyr80	4.66411834192839

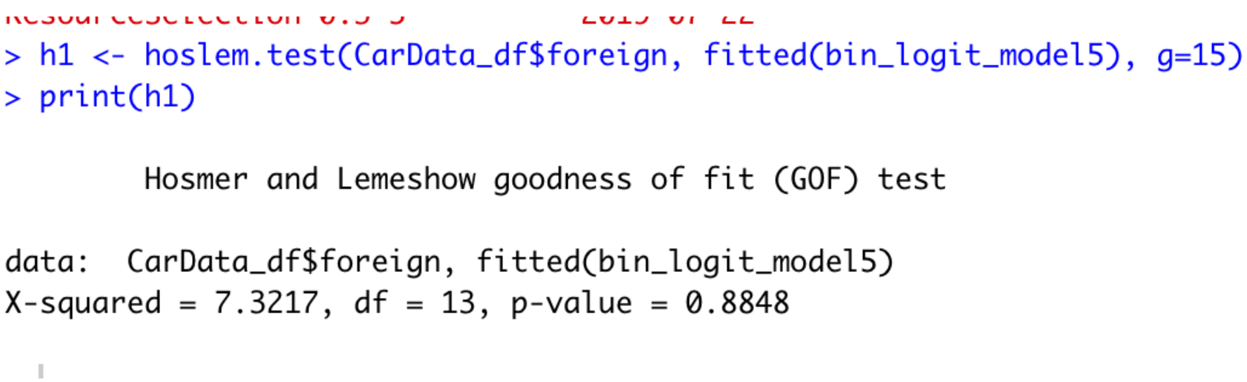
As per the coefficient estimates, few of the variables are important and significant predictors of whether the vehicle is of foreign origin evidenced by the variables comparatively high z-value and low p-value.

The odds ratio is also statistically significant as because the independent variable having p values for the calculated coefficients are mostly greater than 5%. Let me explain with one scenario how each independent variable is accounting for the dependent variable. For example, consider mpg variable, 1 unit increase in coefficient value of cylinders is associated with the 1.2 times increase in the odds function of the being the vehicle foreign w.r.t non-foreign. Hence, the variables displacement, weight, modyr80 are more related to foreign variable when compared to other independent variables when considered jointly.

5. Discuss goodness-of-fit measures.

Ans:

Goodness of fit is a statistical test that determines how well the sample data fit the population. Simply put, it makes assumptions about whether the sample is biased, or whether it represents the expected data in a real population.
For the model we build by considering the variables, I check by using hoslem.test(). hoslem.test() which is Hosmer and Lemeshow goodness of fit.To built the test, first the package "ResourceSelection" is installed and imported the library. The test will have the parameters data, model built, number of groups dividing the data to fit the model as 'g'.
For the test, I passed the parameters the dataset I considered "CarData_df" and the model I built "bin_logit_model5" and no. of groups as 15.



```
ResourceSelection 0.3-5          2019-07-22
> h1 <- hoslem.test(CarData_df$foreign, fitted(bin_logit_model5), g=15)
> print(h1)

	Hosmer and Lemeshow goodness of fit (GOF) test

data:  CarData_df$foreign, fitted(bin_logit_model5)
X-squared = 7.3217, df = 13, p-value = 0.8848
```

From the above test, we can conclude that the p-value is greater than alpha value i.e 0.05.
Hence it fails to reject the null hypothesis. Although the model built is only significant for few variables, but overall when you considered as the data as a model, it is not significant.

II. Summarize the main ideas in “On the relevance of irrelevant alternatives,” by Benson et al. (2016).

There are various econometric model which can be built on the datasets based upon the attribute features and their behavior. The author discuss more on the multinomial logit model which comes with an fundamental model assumption called independence of irrelevant alternatives(IIA). Thereby implementing various statistical test on various datasets and determining what behavior is violated from one other. Also, few experiments have been implemented on both synthetic and real datasets showing how nested logit model is recovered by using traditional model.

The Multinomial logit model is very strict mathematical model and implies restrictions how people can change the behavior namely choice of abilities like characteristics and available choices. For example, if you want to buy a car from one other or choosing restaurants Japanese and Italian, the relative probability doesn’t depend upon the characteristics and choices due to the strict mathematical structure. Simply, the alternative is independent of the decision of the other two alternatives considering as a pair. When you consider one pair of alternatives, the relative probabilities or ratio depend only on the chosen alternative attributes and not on the third alternative attributes. Moreover, IIA property allows to add or remove additional attributes without effecting the model parameters. Hence if the IIA property is violating the Multinomial Logit model, then to overcome the author proposed Nested Logit model. This can be solved by using math and building the models like logit model.

Nested logit model violates the relative probability of the choosing the restaurant when new Japanese restaurant was included. It divides the both the restaurants by the binary tree thereby including both the Japanese restaurants in one node not reflecting the Italian restaurant. The author proposed four different observations to statistical hypothesis tests when you use the subsets of the data and decides the best possible translation to this observation. Simultaneous binomial (SB) uses binomial distribution interpreting a sample set, Multiple sample binomial (MSB) uses single choice set to other sets. Aggregated multiple sample binomial (AMSB) aggregates the pairs of each sets and the last translation model Choice set binomial (CSB) uses two binomial distributions.

The author collected datasets for the experiments to evaluate how each model have the impact on the dataset and ran the statistical tests by taking dataset as a pair irrelevant to each other and measured the frequency at a significance level of $\alpha = 0.05$. For example, the four different tests have been implemented on Japanese Cuisine and Lastfmgenre datasets which exhibits extreme violations of IIA. Further, the nested logit model will be recovered by implementing the algorithms. Firstly, powerful tree oracle is the first algorithm used which explains how it recovered in a quadratic time taking the pairs in a single query with the parameters equal, higher and lower. A quadratic-time algorithm, assuming the tree oracle is available and by taking large inputs irrespective of the size performing worse than the best algorithms. For the complexity analysis, it requires quadratic number of operations as well as queries. Similarly, a quadratic-time lower bound has been implemented. When considered real world datasets not having the sufficient samples to implement in the oracle, Greedy algorithm comes which is more effective on the parse. Hence, it finds all siblings of node, merges these nodes into new leaf nodes, and continues recursively.

By using recovery algorithms, few experiments have been done by implementing on the synthetic data. There are two requirements where the first, pairwise preference data means that for items i and j, knowledge of the probability of choosing each item if it is the only option available. Multiway preference data selects from more than two observations. Similarly, recovered nested logit trees by implementing on real world data. Data minimization has changed the way the recovery algorithms are performed, and the recovered tree shows an interesting combination of results.