# Linear Regression

*Jim Turman*

*November 3, 2016*

## Introduction

This report is a Linear Regression predicting the price of a Toyota Corolla given certain variables.

### Descriptive Statistics

Below are the descriptive statistics for the data set describing the Price, Age, KM on the car, Fuel Type, HorsePower (HP), Metallic Paint(MetColor), Transmission type (Automatic), Cubic Centimeters of engine displacement (CC), Doors and Weight. There are 10 variables in all and the table below shows their descriptive statistics: n is the number of records, sd is the standard deviation, trimmed is the trimmed mean exluding the upper and lower 5% of data, mad is the mean absolute deviation, se is the standard error. All other measures are as they are represented in the table. In the file Plots that comes with this file are the graphs that represent the data.

```
##           vars    n     mean       sd  median  trimmed      mad  min
## Price        1 1436 10730.82  3626.96  9900.0 10160.59  2446.29 4350
## Age          2 1436    55.95    18.60    61.0    57.93    17.79    1
## KM           3 1436 68533.26 37506.45 63389.5 65249.58 32517.12    1
## FuelType     4 1436     0.90     0.33     1.0     0.99     0.00    0
## HP           5 1436   101.50    14.98   110.0   102.96     0.00   69
## MetColor     6 1436     0.67     0.47     1.0     0.72     0.00    0
## Automatic    7 1436     0.06     0.23     0.0     0.00     0.00    0
## CC           8 1436  1566.83   187.18  1600.0  1548.00     0.00 1300
## Doors        9 1436     4.03     0.95     4.0     4.04     1.48    2
## Weight      10 1436  1072.46    52.64  1070.0  1066.13    37.06 1000
##              max  range  skew kurtosis     se
## Price      32500  28150  1.70     3.71  95.71
## Age           80     79 -0.82    -0.08   0.49
## KM        243000 242999  1.01     1.67 989.76
## FuelType       2      2 -1.72     4.29   0.01
## HP           192    123  0.95     8.79   0.40
## MetColor       1      1 -0.75    -1.45   0.01
## Automatic      1      1  3.87    12.99   0.01
## CC          2000    700  0.61     0.45   4.94
## Doors          5      3 -0.08    -1.87   0.03
## Weight      1615    615  3.10    19.26   1.39
```

### Linear Regression Analysis

Below is the beginning linear regression with all variables included in the model. If the P value shown is less than .05 then the variable is significant to the model. The hypothesis test is that at least one of the variables contributes significantly to the model. As shown below, MetColor, Automatic and Doors variables have a p value greater than .05 suggesting that they do not significantly contribute to the model.

```
##
## Call:
## lm(formula = Price ~ ., data = toy.r)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11209.6   -748.0      8.9    735.9   6374.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.358e+03  1.154e+03  -2.043   0.0412 *
## Age         -1.226e+02  2.589e+00 -47.336  < 2e-16 ***
## KM          -1.567e-02  1.285e-03 -12.190  < 2e-16 ***
## FuelType    -1.555e+03  2.497e+02  -6.225 6.31e-10 ***
## HP           5.279e+01  4.084e+00  12.926  < 2e-16 ***
## MetColor     5.563e+01  7.501e+01   0.742   0.4584
## Automatic    2.905e+02  1.560e+02   1.863   0.0627 .
## CC          -3.446e+00  4.024e-01  -8.565  < 2e-16 ***
## Doors       -2.535e+01  3.910e+01  -0.648   0.5169
## Weight       2.099e+01  1.096e+00  19.151  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1317 on 1426 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8681
## F-statistic:  1051 on 9 and 1426 DF,  p-value: < 2.2e-16
```

Below are the confidence levels for the given variables.

```
##                      2.5 %        97.5 %
## (Intercept) -4.622093e+03   -94.3881168
## Age         -1.276522e+02  -117.4931773
## KM          -1.818931e-02    -0.0131468
## FuelType    -2.044477e+03 -1064.7406133
## HP           4.477963e+01    60.8020798
## MetColor    -9.151219e+01   202.7764608
## Automatic   -1.539937e+01   596.4856405
## CC          -4.235881e+00    -2.6571199
## Doors       -1.020607e+02    51.3541562
## Weight       1.884418e+01    23.1450328
```

Below is a correlation matrix showing the extent to which the variables are correlated with Price. A value close to 1 or -1 means a perfect positive or perfect negative correlation respectively. A value of 0 suggests no correlation. As shown below, FuelType, MetColor, Automatic, CC and Doors are not shown to have a strong correlation in predicting price.

```
##                Price         Age          KM    FuelType          HP
## Price      1.00000000 -0.87659050 -0.56996016 -0.06333851  0.31498983
## Age       -0.87659050  1.00000000  0.50567218  0.09199862 -0.15662202
## KM        -0.56996016  0.50567218  1.00000000 -0.32932709 -0.33353795
## FuelType  -0.06333851  0.09199862 -0.32932709  1.00000000  0.51807787
## HP         0.31498983 -0.15662202 -0.33353795  0.51807787  1.00000000
## MetColor   0.10890475 -0.10814958 -0.08050293  0.01842624  0.05871170
```

```
## Automatic  0.03308069  0.03171677 -0.08185408  0.07933825  0.01314403
## CC         0.16506697 -0.13318154  0.30215036 -0.70457577  0.05088370
## Doors      0.18532555 -0.14835921 -0.03619661 -0.02064540  0.09242450
## Weight     0.58119759 -0.47025318 -0.02859846 -0.51303478  0.08961406
##              MetColor   Automatic          CC       Doors      Weight
## Price      0.10890475  0.03308069  0.16506697  0.18532555  0.58119759
## Age       -0.10814958  0.03171677 -0.13318154 -0.14835921 -0.47025318
## KM        -0.08050293 -0.08185408  0.30215036 -0.03619661 -0.02859846
## FuelType   0.01842624  0.07933825 -0.70457577 -0.02064540 -0.51303478
## HP         0.05871170  0.01314403  0.05088370  0.09242450  0.08961406
## MetColor   1.00000000 -0.01933545  0.03492137  0.08524283  0.05792883
## Automatic -0.01933545  1.00000000 -0.06932134 -0.02765382  0.05724851
## CC         0.03492137 -0.06932134  1.00000000  0.12676764  0.65144958
## Doors      0.08524283 -0.02765382  0.12676764  1.00000000  0.30261764
## Weight     0.05792883  0.05724851  0.65144958  0.30261764  1.00000000
```

This next test is used to find the best combination of variables, which variables are needed and which are dropped, to have the most explanatory power in predicting the price of a car. The higher the R-SQ (adj) value the better the model is for predicting price. As show below the best model (highest R-Sq Adjusted) is the 7th iteration. That model has the variables Age, KM, FuelType, HP, Automatic, CC and Weight.

```
##   (Intercept) Age KM FuelType HP MetColor Automatic CC Doors Weight
## 1           1   1  0        0  0        0         0  0     0      0
## 2           1   1  0        0  0        0         0  0     0      1
## 3           1   1  1        0  0        0         0  0     0      1
## 4           1   1  1        0  1        0         0  0     0      1
## 5           1   1  1        0  1        0         0  1     0      1
## 6           1   1  1        1  1        0         0  1     0      1
## 7           1   1  1        1  1        0         1  1     0      1
## 8           1   1  1        1  1        1         1  1     0      1
## 9           1   1  1        1  1        1         1  1     1      1
##        R-Sq R-Sq (adj)          Cp
## 1 0.7684109  0.7682494 1088.286745
## 2 0.8050716  0.8047995  691.324054
## 3 0.8481042  0.8477860  225.017609
## 4 0.8617759  0.8613895   78.235067
## 5 0.8650265  0.8645546   44.859806
## 6 0.8685488  0.8679968    8.528615
## 7 0.8688811  0.8682383    6.912083
## 8 0.8689263  0.8681914    8.420365
## 9 0.8689649  0.8681379   10.000000
```

The optimal model for the given data and variables in predicting price for Toyota Corolla's is shown below, note that the Adjusted R-squared value is 0.8682 meaning we can say with 86.82% confidence that our model is accurate.

```
##
## Call:
## lm(formula = Price ~ ., data = toy.2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11106.1   -747.7      0.3    728.3   6371.6
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.208e+03  1.134e+03  -1.947   0.0518 .
## Age         -1.227e+02  2.584e+00 -47.498  < 2e-16 ***
## KM          -1.571e-02  1.284e-03 -12.240  < 2e-16 ***
## FuelType    -1.567e+03  2.472e+02  -6.339 3.09e-10 ***
## HP           5.288e+01  4.075e+00  12.977  < 2e-16 ***
## Automatic    2.957e+02  1.554e+02   1.902   0.0573 .
## CC          -3.435e+00  4.016e-01  -8.554  < 2e-16 ***
## Weight       2.079e+01  1.048e+00  19.839  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1317 on 1428 degrees of freedom
## Multiple R-squared:  0.8689, Adjusted R-squared:  0.8682
## F-statistic:  1352 on 7 and 1428 DF,  p-value: < 2.2e-16
```

## Predicting the Price

Because we live in the real world where most of the time we do not have the optimal data to feed into our optimal model we must adjust and make the best prediction possible with the given data, this case is no different. We have been asked to predict the price of a car given that it is 12 months old, uses petrol, has 185 horsepower, has metallic paint, has a standard transmission, a 2000 CC engine and 4 doors. Note that these variables are different from the optimal solution above. Because of this data we have to create a new linear regression model for the factors we are given. The new model is shown below and has an adjusted R-squared value of .8211. With this model we are able to predict with 82.11% confidence that the price of that car will be $22,772.63.

```
##
## Call:
## lm(formula = Price ~ ., data = toy.3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7743.0  -917.8    -2.5   845.8 10889.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18013.1737   653.1245  27.580  < 2e-16 ***
## Age          -158.7395     2.2938 -69.203  < 2e-16 ***
## FuelType    -2983.6858   281.3298 -10.606  < 2e-16 ***
## HP             80.1797     4.4713  17.932  < 2e-16 ***
## MetColor       69.5683    87.2992   0.797    0.426
## Automatic    1059.8909   177.2514   5.980 2.82e-09 ***
## CC             -3.0006     0.4238  -7.080 2.26e-12 ***
## Doors         186.6019    43.5539   4.284 1.95e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1534 on 1428 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8211
## F-statistic: 941.8 on 7 and 1428 DF,  p-value: < 2.2e-16
```

```
##        1
## 22772.63
```

The confidence intervals below show the Mean Error, Root Mean Square Error and Mean Absolute Percent Error in that order.

```
##                     2.5 %       97.5 %
## (Intercept) 16731.987223 19294.360128
## Age          -163.239073  -154.239859
## FuelType    -3535.549866 -2431.821782
## HP             71.408754    88.950692
## MetColor     -101.680116   240.816723
## Automatic     712.189858  1407.591937
## CC             -3.832028    -2.169183
## Doors         101.165454   272.038356
```

```
## [1] 1.575524
```

```
## [1] 1542.013
```

```
## [1] 11.12214
```

The Mean Error shows the fit of a data point to the line from the model. The Root Mean Square Error is the square root of the Mean Squared Error and is the distance on average of a data point from the fitted line measured vertically. The Mean Absolute Percent Error shows the percentage that the model is off from actual values. In this case the Mean Error is 1.575524. The Root Mean Square Error is 1542.013 and the Mean Absolute Error Percentage is 11.12% which means that the model predicting the cost of a Toyota developed earlier based on the data for the car available is with 11.12% of the actual price.