Data Analytics for Business

Homework 1 – Part 1

Linear Models

In this homework, we will perform linear regressions on a real dataset from a European Toyota car dealer on the sales records of used cars (Toyota Corolla). We would like to construct a reasonable linear regression model for the relationship between the sales prices of used cars and various explanatory variables (such as age, mileage, horsepower). We are interested to see what factors affect the sales price of a used car and by how much.

Download the data file "**UsedCars.csv**," which contains the following fields. Complete the tasks outlined below.

Data Description

Column Name	Variable Description
Id	ID number of each used car
Model	Model name of each used car
Price	The price (**in Euros**) at which each used car was sold
Age	Age **(in months**) of each used car as of August 2004
KM	Accumulated kilometers on odometer
HP	Horsepower
Metallic	Metallic color? (Yes = 1, No = 0)
Automatic	Automatic transmission? (Yes = 1, No = 0)
CC	Cylinder volume (**in cubic centimeters**)
Doors	Number of doors
Gears	Number of gears
Weight	Weight (**in kilograms**)

Task 1: Import Data and Run a Linear Regression

1. Import that data into R. Run a linear regression of **Price** on all the available explanatory variables (i.e., **Age, KM, HP, Metallic, Automatic, CC, Doors, Gears, Weight**). Use the **summary()** function to show the regression results. (**Note**: R is case sensitive, so be careful with the variable names.)

2. Calculate the **fitted values** of the response variable, and calculate the **residuals**. Co-list the original y values, fitted $\hat{y}$ values, and the residuals together for the first 10 observations. Check if the residuals equal $y - \hat{y}$.

Task 2: *t*-Statistic and *p*-Value

3. Re-produce the *t*-statistics for all $\hat{\beta}_j$, using the defining formula $\boldsymbol{t(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$. Co-list your calculated *t*-statistics along with the *t*-statistics generated from **summary()** of the regression results. They should be exactly the same.

4. Determine the **critical value** (or cutoff) of the *t*-statistic for a β estimate to be considered as significant at **95% confidence level**. You need to first determine the degree of freedom of your model (**Hint**: you can simply retrieve the value of **df.residual** from the regression result.) Then you need to find the corresponding percentile of the *t* distribution (with that degree of freedom). (**Hint**: use **qt()** function to find a certain percentile of a *t* distribution.)

5. Calculate the *p*-value for each $\hat{\beta}_j$ using the defining formula $\boldsymbol{p} = \mathbf{2} \cdot \mathbf{Pr}(\boldsymbol{t} < -|\boldsymbol{tstat}|)$. (**Hint**: use **pt()** function for the *cdf* of *t* distribution.) Co-list your calculated *p*-values along with the *p*-values generated from **summary()** of the regression results. They should be exactly the same.

6. Which explanatory variables have significant effects on the outcome, that is, which β estimates are significantly different from zero? You can find the answers either by comparing the *t*-statistics (obtained in Step 3) to the critical values (obtained in Step 4) or by comparing the *p*-values (obtained in Step 5) to (1-confidence level), as we discussed in class. The conclusions should be the same. (**Note**: use **95% confidence level**.)

Task 3: R^2 and VIF

7. Calculate the R-squared of the regression you have performed, using the defining formula $\boldsymbol{R^2 = \frac{SSE}{SST}}$. Compare your calculated value with the R-squared value calculated by the routine. (**Hint**: you can retrieve **r.squared** from the **summary()** output.) Again, they should be the same.

8. Install and load the package "**car**". Use the **vif()** function included in the package to calculate the **variance inflation factor (VIF)** for the β estimators. Examine the VIF values and discuss if there is any sign of multicollinearity among independent variables.

9. Re-produce the VIF for the coefficient of **Weight** (which has the largest VIF value), following these two steps:

    i. Regress **Weight** on all the other independent variables, and obtain the R-squared

    ii. Calculate the VIF using the defining formula $\boldsymbol{VIF(\hat{\beta}_j) = \frac{1}{1-R_j^2}}$.

Task 4: Model Comparison

10. Remove from the model the independent variables which are NOT significant according to your conclusion in Step 6. Run a new linear regression of **Price** on the remaining independent variables. Use the **summary()** function to show the regression results.

11. Retrieve and compare the R-squared and Adjusted R-Squared from the two models (the full regression with all independent variables in Step 1 versus the new model with only the independent variables that are significant in Step 10). Discuss your findings with regard to the relative magnitudes of the R-squared and the Adjusted R-Squared from the two models and what they imply.

12. Use the results from the model with a better fit to discuss the effects of certain independent variables on the dependent variable: Holding everything else equal, how much the sales price would decrease if a car were **one year older**? What if a car accumulated **10,000 more kilometers**?