LAB 05

AML

30/10/2020

```
setwd("/Users/andrea/Desktop/UEA/Classes/Econometrics/Data")
library(GGally)
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
library(data.table)
```

Exercise 1

```
load("car.test.RData")
load("car.train.RData")
```

Use the data to build a model explaining the price of used cars

```
colnames(car.train)
```

```
##  [1] "Price"     "Age"       "KM"        "FuelType"  "HP"        "MetColor"
##  [7] "Automatic" "CC"        "Doors"     "Weight"
```

```
car.train <- data.table(car.train)
```

Manipulate the car.train object to add squared variables

```
car.train <- car.train[, Age2:=Age^2] # create new var. age^2
car.train <- car.train[, HP2:=HP^2] # create new var. hp^2
car.train <- car.train[, KM2:=KM^2] # create new var. km^2
car.train <- car.train[, Weight2:=Weight^2] # create new var. weight^2
```

Regress price against all other variables

```
out1 <- lm(Price ~ . , data = car.train)
summary(step(out1)) # select variables to include into the model using AIC
```

```
## Start:  AIC=12191.59
## Price ~ Age + KM + FuelType + HP + MetColor + Automatic + CC +
##     Doors + Weight + Age2 + HP2 + KM2 + Weight2
##
##             Df Sum of Sq        RSS   AIC
## - HP         1    101512 1155629993 12190
## - MetColor   1    245096 1155773577 12190
## - Doors      1    362635 1155891116 12190
## <none>                   1155528481 12192
## - KM         1   4804869 1160333350 12193
## - KM2        1   5614378 1161142858 12194
## - Automatic  1   6156850 1161685330 12194
## - FuelType   2  11347088 1166875569 12196
## - HP2        1  10217061 1165745542 12197
## - Weight2    1  14628686 1170157166 12200
## - CC         1  16922292 1172450773 12202
## - Weight     1  24374273 1179902754 12208
## - Age2       1 150757761 1306286241 12295
## - Age        1 502301715 1657830196 12501
##
## Step:  AIC=12189.66
## Price ~ Age + KM + FuelType + MetColor + Automatic + CC + Doors +
##     Weight + Age2 + HP2 + KM2 + Weight2
##
##             Df Sum of Sq        RSS   AIC
## - MetColor   1    234907 1155864900 12188
## - Doors      1    352299 1155982291 12188
## <none>                   1155629993 12190
## - KM         1   4987681 1160617674 12191
## - KM2        1   5521759 1161151752 12192
## - Automatic  1   6391686 1162021679 12192
## - Weight2    1  15009576 1170639568 12199
## - Weight     1  24683893 1180313886 12206
## - FuelType   2  29650278 1185280270 12208
## - CC         1  51344325 1206974318 12225
## - HP2        1 132673567 1288303559 12281
## - Age2       1 150830232 1306460225 12293
## - Age        1 504733611 1660363604 12500
##
## Step:  AIC=12187.84
## Price ~ Age + KM + FuelType + Automatic + CC + Doors + Weight +
##     Age2 + HP2 + KM2 + Weight2
##
##             Df Sum of Sq        RSS   AIC
## - Doors      1    321577 1156186477 12186
## <none>                   1155864900 12188
## - KM         1   4925321 1160790221 12190
## - KM2        1   5645757 1161510657 12190
## - Automatic  1   6301657 1162166557 12190
## - Weight2    1  14874551 1170739451 12197
## - Weight     1  24522735 1180387635 12204
## - FuelType   2  29498524 1185363424 12206
## - CC         1  51112748 1206977648 12223
## - HP2        1 132570109 1288435009 12279
## - Age2       1 151806190 1307671090 12292
## - Age        1 507773611 1663638511 12500
##
## Step:  AIC=12186.08
## Price ~ Age + KM + FuelType + Automatic + CC + Weight + Age2 +
##     HP2 + KM2 + Weight2
##
##             Df Sum of Sq        RSS   AIC
## <none>                   1156186477 12186
## - KM         1   5037578 1161224055 12188
## - KM2        1   5596264 1161782741 12188
## - Automatic  1   7365407 1163551884 12190
## - Weight2    1  15564105 1171750582 12196
## - Weight     1  26981390 1183167867 12204
## - FuelType   2  32649423 1188835900 12206
## - CC         1  51164720 1207351197 12221
## - HP2        1 145162799 1301349276 12286
## - Age2       1 156182032 1312368509 12293
## - Age        1 522164399 1678350876 12505
```

```
##
## Call:
## lm(formula = Price ~ Age + KM + FuelType + Automatic + CC + Weight +
##     Age2 + HP2 + KM2 + Weight2, data = car.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6620.3  -690.0    -7.1   664.0  5663.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.767e+04  7.576e+03  -2.333 0.019889 *
## Age            -2.478e+02  1.265e+01 -19.593  < 2e-16 ***
## KM             -7.586e-03  3.942e-03  -1.924 0.054632 .
## FuelTypeDiesel  2.477e+03  5.362e+02   4.620 4.44e-06 ***
## FuelTypePetrol  9.218e+02  4.250e+02   2.169 0.030343 *
## Automatic       4.143e+02  1.780e+02   2.327 0.020200 *
## CC             -3.380e+00  5.512e-01  -6.133 1.32e-09 ***
## Weight          5.637e+01  1.266e+01   4.454 9.56e-06 ***
## Age2            1.241e+00  1.158e-01  10.715  < 2e-16 ***
## HP2             2.327e-01  2.252e-02  10.331  < 2e-16 ***
## KM2            -4.005e-08  1.974e-08  -2.028 0.042835 *
## Weight2        -1.735e-02  5.129e-03  -3.383 0.000751 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1166 on 850 degrees of freedom
## Multiple R-squared:  0.8912, Adjusted R-squared:  0.8898
## F-statistic: 633.2 on 11 and 850 DF,  p-value: < 2.2e-16
```

Take preferred model from previous step and run usual commands

```
out <- lm( Price ~ Age + Age2 + KM + KM2 + HP + HP2 + Weight + Weight2 + FuelType + Automatic + CC
           , data = car.train)
stargazer(out, type = "text")
```

```
##
## ===============================================
##                         Dependent variable:
##                     ---------------------------
##                                Price
## -----------------------------------------------
## Age                         -248.050***
##                               (12.706)
##
## Age2                          1.243***
##                               (0.116)
##
## KM                            -0.007*
##                               (0.004)
##
## KM2                         -0.00000**
##                              (0.00000)
##
## HP                            -6.844
##                              (27.792)
##
## HP2                          0.255***
##                              (0.093)
##
## Weight                       56.100***
##                              (12.709)
##
## Weight2                      -0.017***
##                              (0.005)
##
## FuelTypeDiesel             2,317.297***
##                              (842.180)
##
## FuelTypePetrol               918.760**
##                              (425.368)
##
## Automatic                    409.799**
##                              (179.082)
##
## CC                           -3.203***
##                              (0.907)
##
## Constant                   -17,348.070**
##                             (7,693.953)
##
## -----------------------------------------------
## Observations                    862
## R2                             0.891
## Adjusted R2                    0.890
## Residual Std. Error    1,166.929 (df = 849)
## F Statistic          579.810*** (df = 12; 849)
## ===============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

```
out2 <- lm( Price ~ MetColor + Doors + Age + Age2 + KM + KM2 + HP + HP2 + Weight + Weight2 + FuelType + Automatic
+ CC
            , data = car.train)
stargazer(out, out2, type = "text") #Can see these two variables add nothing to the model and reduce F-stat
```

```
##
## =========================================================================
##                                     Dependent variable:
##                     -----------------------------------------------------
##                                            Price
##                               (1)                        (2)
## -------------------------------------------------------------------------
## MetColor                                               36.492
##                                                       (86.095)
##
## Doors                                                  -26.311
##                                                       (51.033)
##
## Age                        -248.050***               -246.858***
##                              (12.706)                  (12.865)
##
## Age2                         1.243***                  1.233***
##                              (0.116)                   (0.117)
##
## KM                           -0.007*                   -0.007*
##                              (0.004)                   (0.004)
##
## KM2                        -0.00000**                -0.00000**
##                             (0.00000)                 (0.00000)
##
## HP                           -6.844                    -7.595
##                             (27.792)                  (27.843)
##
## HP2                         0.255***                  0.254***
##                             (0.093)                   (0.093)
##
## Weight                      56.100***                 59.215***
##                             (12.709)                  (14.009)
##
## Weight2                     -0.017***                 -0.018***
##                             (0.005)                   (0.006)
##
## FuelTypeDiesel            2,317.297***              2,266.370***
##                            (842.180)                 (848.706)
##
## FuelTypePetrol              918.760**                 940.408**
##                            (425.368)                 (428.126)
##
## Automatic                   409.799**                 391.723**
##                            (179.082)                 (184.395)
##
## CC                          -3.203***                 -3.200***
##                             (0.907)                   (0.908)
##
## Constant                  -17,348.070**             -19,255.510**
##                           (7,693.953)               (8,505.219)
##
## -------------------------------------------------------------------------
## Observations                  862                       862
## R2                           0.891                     0.891
## Adjusted R2                  0.890                     0.890
## Residual Std. Error   1,166.929 (df = 849)      1,168.016 (df = 847)
## F Statistic         579.810*** (df = 12; 849) 496.086*** (df = 14; 847)
## =========================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Apply data table to test data set and create squared variables

```
car.test <- data.table(car.test)
car.test <- car.test[, Age2:=Age^2]
car.test <- car.test[, HP2:=HP^2]
car.test <- car.test[, Weight2:=Weight^2]
car.test <- car.test[, KM2:=KM^2]
```

Define car.test$y_hat which uses model from first data set to predict values for second data set

```
car.test$y_hat <- predict(out1, newdata=car.test)
colnames(car.test)
```

```
##  [1] "Price"     "Age"       "KM"        "FuelType"  "HP"        "MetColor"
##  [7] "Automatic" "CC"        "Doors"     "Weight"    "Age2"      "HP2"
## [13] "Weight2"   "KM2"       "y_hat"
```

```
head(car.test)
```

```
##    Price Age     KM FuelType  HP MetColor Automatic   CC Doors Weight Age2
## 1: 11290  49  80320   Petrol 110        1         1 1600     3   1070 2401
## 2: 15950  19  51884   Petrol  97        1         0 1400     3   1100  361
## 3:  8500  80 100458   Petrol 110        0         0 1600     5   1085 6400
## 4:  8900  67  54847   Petrol 110        0         0 1600     3   1050 4489
## 5: 15950  28  29206   Petrol  97        1         0 1400     5   1110  784
## 6: 15950  30  67660   Petrol 110        1         0 1600     3   1105  900
##      HP2 Weight2         KM2     y_hat
## 1: 12100 1144900  6451302400 11514.079
## 2:  9409 1210000  2691949456 17013.095
## 3: 12100 1177225 10091809764  8309.281
## 4: 12100 1102500  3008193409  9141.654
## 5:  9409 1232100   852990436 15691.126
## 6: 12100 1221025  4577875600 14807.865
```

Our model does a reasonable job of predicting price.

Use the RMSE to compare the performance of your model in carTrain.RData and carTest.RData

```
#install.packages("Metrics")
library(Metrics)
car.train$y_hat <- predict(out1, newdata=car.train)
head(car.train)
```

```
##    Price Age    KM FuelType HP MetColor Automatic   CC Doors Weight Age2  HP2
## 1: 13750  23 72937   Diesel 90        1         0 2000     3   1165  529 8100
## 2: 13950  24 41711   Diesel 90        1         0 2000     3   1165  576 8100
## 3: 13750  30 38500   Diesel 90        0         0 2000     3   1170  900 8100
## 4: 12950  32 61000   Diesel 90        0         0 2000     3   1170 1024 8100
## 5: 16900  27 94612   Diesel 90        1         0 2000     3   1245  729 8100
## 6: 18600  30 75889   Diesel 90        1         0 2000     3   1245  900 8100
##           KM2 Weight2    y_hat
## 1: 5319805969 1357225 16242.52
## 2: 1739807521 1357225 16432.39
## 3: 1482250000 1368900 15430.64
## 4: 3721000000 1368900 14830.84
## 5: 8951430544 1550025 16391.52
## 6: 5759140321 1550025 16131.60
```

Define residual and residual^2.

```
car.train$u_hat <- car.train$y_hat - car.train$Price
car.train$u_hat2 <- (car.train$u_hat)^2
head(car.train)
```

```
##    Price Age    KM FuelType HP MetColor Automatic   CC Doors Weight Age2  HP2
## 1: 13750  23 72937   Diesel 90        1         0 2000     3   1165  529 8100
## 2: 13950  24 41711   Diesel 90        1         0 2000     3   1165  576 8100
## 3: 13750  30 38500   Diesel 90        0         0 2000     3   1170  900 8100
## 4: 12950  32 61000   Diesel 90        0         0 2000     3   1170 1024 8100
## 5: 16900  27 94612   Diesel 90        1         0 2000     3   1245  729 8100
## 6: 18600  30 75889   Diesel 90        1         0 2000     3   1245  900 8100
##           KM2 Weight2    y_hat      u_hat    u_hat2
## 1: 5319805969 1357225 16242.52  2492.5194 6212653.1
## 2: 1739807521 1357225 16432.39  2482.3940 6162280.1
## 3: 1482250000 1368900 15430.64  1680.6424 2824558.9
## 4: 3721000000 1368900 14830.84  1880.8411 3537563.1
## 5: 8951430544 1550025 16391.52  -508.4822  258554.2
## 6: 5759140321 1550025 16131.60 -2468.3975 6092986.2
```

Root mean squared error

```
rmse.train <- sqrt(sum(car.train$u_hat2)/nrow(car.train))
rmse.train
```

```
## [1] 1157.808
```

Alternatively use function rmse:

```
rmse.train.2 <- rmse(car.train$Price, car.train$y_hat)
rmse.train.2
```

```
## [1] 1157.808
```

On the other data set, define residuals

```
car.test$u_hat <- car.test$y_hat - car.test$Price
car.test$u_hat2 <- (car.test$u_hat)^2
head(car.test)
```

```
##    Price Age     KM FuelType  HP MetColor Automatic   CC Doors Weight Age2
## 1: 11290  49  80320   Petrol 110        1         1 1600     3   1070 2401
## 2: 15950  19  51884   Petrol  97        1         0 1400     3   1100  361
## 3:  8500  80 100458   Petrol 110        0         0 1600     5   1085 6400
## 4:  8900  67  54847   Petrol 110        0         0 1600     3   1050 4489
## 5: 15950  28  29206   Petrol  97        1         0 1400     5   1110  784
## 6: 15950  30  67660   Petrol 110        1         0 1600     3   1105  900
##      HP2 Weight2         KM2     y_hat      u_hat     u_hat2
## 1: 12100 1144900  6451302400 11514.079   224.0795   50211.60
## 2:  9409 1210000  2691949456 17013.095  1063.0951 1130171.14
## 3: 12100 1177225 10091809764  8309.281  -190.7189   36373.72
## 4: 12100 1102500  3008193409  9141.654   241.6545   58396.88
## 5:  9409 1232100   852990436 15691.126  -258.8741   67015.81
## 6: 12100 1221025  4577875600 14807.865 -1142.1354 1304473.28
```

```
rmse.test <- rmse(car.test$Price, car.test$y_hat)
rmse.test
```

```
## [1] 1312.619
```

Model 1 - Compound depreciation

```
summary(out6a <- lm( log(Price) ~ Age, data=car.train))
```

```
##
## Call:
## lm(formula = log(Price) ~ Age, data = car.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.83085 -0.07564  0.00464  0.09004  0.45757
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 10.0066369  0.0149195  670.71   <2e-16 ***
## Age         -0.0138416  0.0002534  -54.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1378 on 860 degrees of freedom
## Multiple R-squared:  0.7762, Adjusted R-squared:  0.7759
## F-statistic:  2983 on 1 and 860 DF,  p-value: < 2.2e-16
```

```
coeffs.out6a <- coefficients(out6a)
logValue <- coeffs.out6a[1]
Value <- exp(logValue)
delta <- 1 - exp(coeffs.out6a[2])
Value
```

```
## (Intercept)
##    22173.14
```

```
delta
```

```
##        Age
## 0.01374627
```

Model 2 - Linear depreciation

```
summary(out6b <- lm( Price ~ Age, data=car.train))
```

```
##
## Call:
## lm(formula = Price ~ Age, data = car.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6234.2  -942.0    68.6   832.5 11888.0
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20056.425    179.011  112.04   <2e-16 ***
## Age          -167.361      3.041  -55.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1653 on 860 degrees of freedom
## Multiple R-squared:  0.7789, Adjusted R-squared:  0.7786
## F-statistic:  3029 on 1 and 860 DF,  p-value: < 2.2e-16
```

```
coeffs.out6b <- coefficients(out6b)
Value2 <- coeffs.out6b[1]

alpha <- coeffs.out6b[2]/Value2

Value2
```

```
## (Intercept)
##    20056.42
```

```
alpha
```

```
##          Age
## -0.008344515
```

Which model is better?

Compare R^2. However, remember that the first model is log y so must convert to y before getting R^2.

Model 1

```
summary(out6a)$sigma
```

```
## [1] 0.1377964
```

```
gamma <- exp((summary(out6a)$sigma)^2/2)
car.train$logyhat <- predict(out6a, newdata= car.train)
car.train$yhat <- gamma*exp(car.train$logyhat)
cor(car.train$yhat, car.train$Price)^2
```

```
## [1] 0.8116621
```

Model 2

```
summary(out6b)$r.squared
```

```
## [1] 0.7788599
```