Assignment 1

Tai Lo Yeung, Pan Yiming, Qiwei Liu

Barcelona Graduate School of Economics

Universitat Pompeu Fabra

Oct 26, 2020

A traditional regression exercise on price analysis

a

Estimate first a simple model where the only explanatory variable is kilometers. Interpret (i.e., explain the meaning and use of) the following output:
- Parameter values and t-statistics (significance of the two parameters).
- R2 of the model and the F-value.

By running code in R, we obtained:

```
Call:
lm(formula = Car_Price ~ Kilometers, data = data_problemset1)

Residuals:
     Min       1Q   Median       3Q      Max
-14122.8  -3466.4   -533.8   2912.6  20163.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.500e+04  1.610e+03  15.529  < 2e-16 ***
Kilometers  -9.962e-02  1.439e-02  -6.923 1.53e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6307 on 72 degrees of freedom
Multiple R-squared:  0.3997,Adjusted R-squared:  0.3913
F-statistic: 47.93 on 1 and 72 DF,  p-value: 1.529e-09
```

The formula is $Car\_Price = 25000 - 0.09962 * Kilometers$, which means that when kilometer is 0, a car worth 25000, and the price for used cars drop 0.09962 per kilometer.
Standard Error for the Intercept is $1.61 * 10^3$ and for Kilometers is $1.439 * 10^-2$. Basically, the standard error of a statistic is just the standard deviation of its sampling distribution.
t value or t-statistic measures the size of the difference relative to the variation in the sample data.
R2 or R-squared shows the proportion of explained variance by the model. In our case, it's 0.3997, that is there are 39.97% of total variance captured by the model.
F-statistic or F-value tests the overall significance of the regression model. In specific, it tests the null hypothesis that all of the coefficients in the model are equal to zero, p-value just simply shows us the possibility that the null hypothesis is true.

b

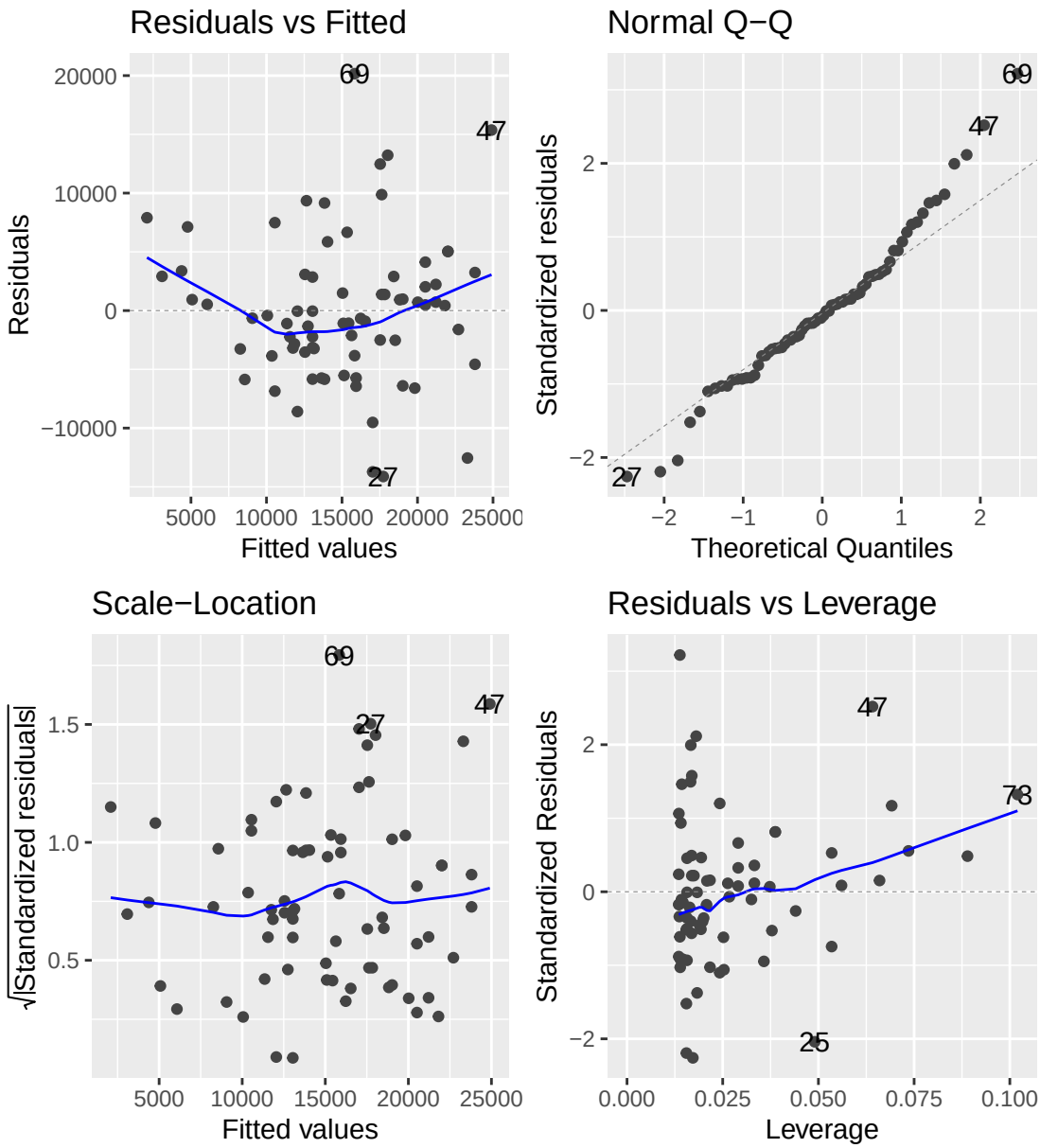
Construct and interpret a confidence interval for the marginal effect of kilometers.

```
                    2.5 %        97.5 %
(Intercept) 21791.8100858  2.821072e+04
Kilometers     -0.1283057 -7.093773e-02
```

That is, we are 95% confident that the intercept value is in the interval [21791.8100858, 28210.72] and coefficient of Kilometers is in [-0.1283057, -0.07093773].

c

Graph the residuals against kilometers. Do you see any evidence of a nonlinear effect? Explain what could cause this apparent nonlinearity. Test statistically whether indeed the effect of kilometers is nonlinear. Interpret the conclusion of the test. How will you modify the model so that the nonlinearity is accounted for?



From what we see on the graph, there is no evidence of non-linearity. We will test it below:

Ramsey-Reset tests

1. fitted ys

```
RESET test

data:  p_k
RESET = 2.638, df1 = 1, df2 = 71, p-value = 0.1088
```

2. power of Xs

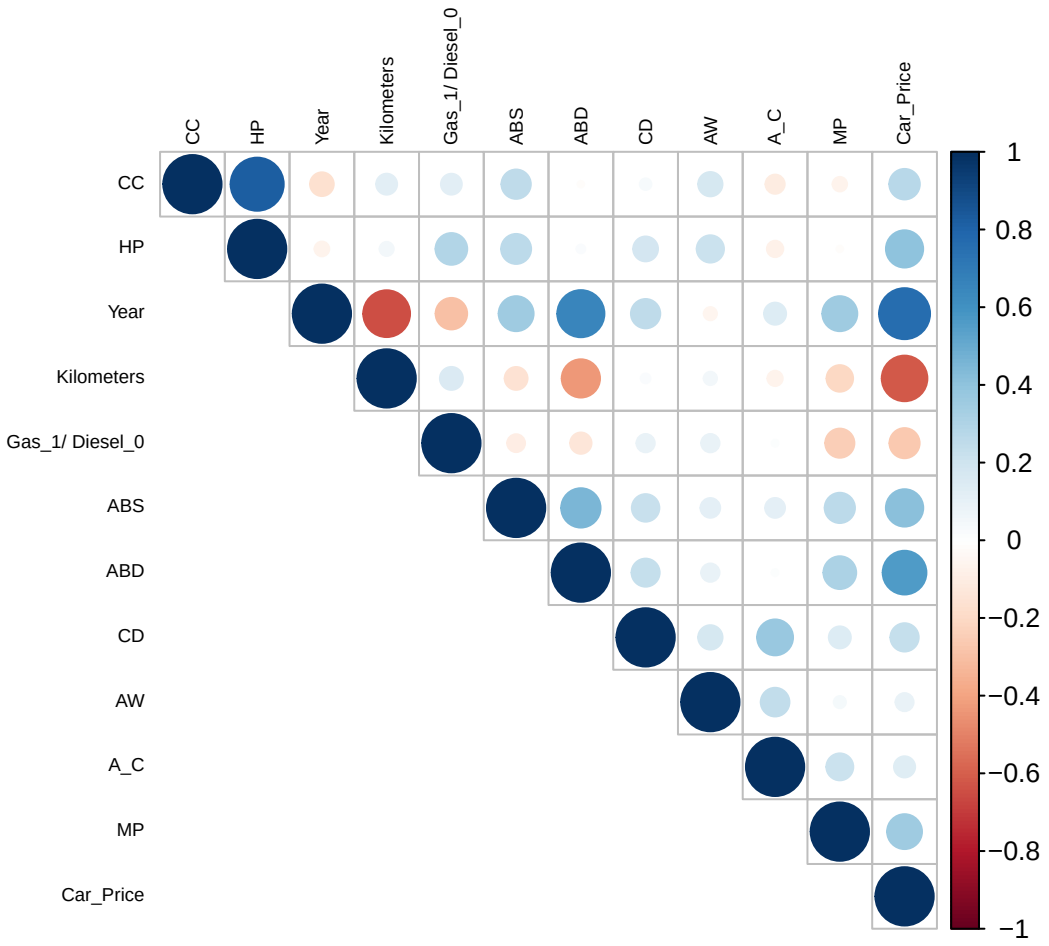
```
RESET test

data:  p_k
RESET = 3.9419, df1 = 1, df2 = 71, p-value = 0.05096
```

In both tests, we failed to reject the null hypothesis that there is no omitted variables in this model. So there is no statistical evidence against the linearity assumption here.
In the case where there is statistical evidence for nonlinearity, we can add new independent variables or simply take log transformation.

d

Select now a more complete model where you include other relevant variables (maybe a look at the correlation matrix - Stata command: correlate - can help you start the process).
- Interpret the parameters and t-stats of your final model of choice.
- Test whether the new additional variables have significant joint explanatory power (that is, do they add anything to the simpler model you estimated in c?).

First, we constructed the correlation matrix as:



Then we filtered the variables and choose only those with high correlation with car price and do the regression. Actually because ABD is highly correlated with year and ABS, we should drop it here but we will do the test first.

```
Call:
lm(formula = Car_Price ~ HP + Year + Kilometers + ABS + ABD +
    MP, data = data_problemset1)

Residuals:
    Min      1Q  Median      3Q     Max
-6926.9 -1980.3   -60.2  1900.7 13586.0

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.388e+06  3.440e+05  -6.941 2.92e-09 ***
HP           6.835e+01  8.287e+00   8.248 1.65e-11 ***
Year         1.199e+03  1.725e+02   6.954 2.78e-09 ***
Kilometers  -3.539e-02  1.113e-02  -3.179  0.00232 **
ABS         -4.892e+01  1.703e+03  -0.029  0.97718
ABD          8.079e+02  1.362e+03   0.593  0.55519
MP           2.112e+03  1.326e+03   1.593  0.11631
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3554 on 61 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.8302,Adjusted R-squared:  0.8135
F-statistic: 49.71 on 6 and 61 DF,  p-value: < 2.2e-16
```

Now we further drop those variables that are not significant in the regression.

```
Call:
lm(formula = Car_Price ~ HP + Year + Kilometers, data = data_problemset1)

Residuals:
    Min      1Q  Median      3Q     Max
-6502.7 -2038.5  -121.2  1693.2 13537.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.664e+06  2.714e+05  -9.817 8.55e-15 ***
HP           6.628e+01  7.365e+00   8.999 2.65e-13 ***
Year         1.339e+03  1.356e+02   9.875 6.74e-15 ***
Kilometers  -3.581e-02  1.056e-02  -3.390  0.00115 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3477 on 70 degrees of freedom
Multiple R-squared:  0.8226,Adjusted R-squared:  0.815
F-statistic: 108.2 on 3 and 70 DF,  p-value: < 2.2e-16
```

Finally, we obtained a model with all variables significant at a 0.01 level or above, the model is:
$Car\_Price = -2664000 + 66.28 * HP + 1339 * Year - 0.03581 * Kilometers$
The interpretation is similar with the one done in part a. All else equal, with one extra Horse Power bring up the price by 66.28, one late year the car is produced the price is higher by 1339, and one more Kilometers used leads to 0.03581 drop in price.
All variables' t-stat and p-value proved that the slope coefficients do significantly changed the predicted price of the car.

Now we perform the joint test for additional variables:

```
Linear hypothesis test

Hypothesis:
HP = 0
Year = 0
Kilometers = 0

Model 1: restricted model
Model 2: Car_Price ~ HP + Year + Kilometers

  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
1     73 4770633784
2     70  846321992  3 3924311792 108.19 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The p-value shows us that it is statistically significant at least one of the new explanatory variables we tested above in not equal to zero. That is, at least one of them affect the car price.

e

Once you have your final model, test the following hypotheses (in some cases, you may have to include variables that you had excluded in the final model):
1. The price of the car goes down by 1200 euros per additional year old the car is.
2. The number of CC and ABS have no impact on the price of the car.
3. The effect of having a CD player is the same as the difference between a gas and a diesel car.
4. Two cars that have the same characteristics except that one has 20000 more kilometers but it is of a design 4 years newer will sell for the same price.
5. There is no heteroskedasticity problem (White's test).
6. There is no heteroskedasticity related to YEAR or HP (Breusch-Pagan test).

1.

```
Linear hypothesis test

Hypothesis:
Year = 1200

Model 1: restricted model
Model 2: Car_Price ~ HP + Year + Kilometers

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     71 859072796
2     70 846321992  1  12750804 1.0546  0.308
```

$H_0 : \beta_{year} = 1200$, the p-value fail to reject the null hypothesis that the car price increases 1200 per year newer the car is.

2.

```
Linear hypothesis test

Hypothesis:
CC = 0
ABS = 0

Model 1: restricted model
Model 2: Car_Price ~ HP + Year + Kilometers + CC + ABS

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     70 846321992
2     68 828005045  2  18316947 0.7521 0.4752
```

$H_0 : \beta_{CC} = 0, \beta_{ABS} = 0$, the p-value fail to reject the null hypothesis that CC and ABS have no impact on the car price.

3.

```
Linear hypothesis test

Hypothesis:
CD - Gas = 0

Model 1: restricted model
Model 2: Price ~ HP + Y + Kilo + CD + Gas

  Res.Df       RSS Df Sum of Sq      F   Pr(>F)
1     69 776732303
2     68 668215059  1 108517244 11.043 0.001436 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$H_0 : \beta_{CD} = \beta_{Gas}$, the p-value this time reject the hypothesis that the effect of having a CD player is the same as the difference between a gas and a diesel car.

4.

```
Hypothesis:
- 4 Year - 20000 Kilometers = 0

Model 1: restricted model
Model 2: Car_Price ~ HP + Year + Kilometers

  Res.Df        RSS Df Sum of Sq     F   Pr(>F)
1     71 1378055701
2     70  846321992  1 531733709 43.98 5.82e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$H_0 : -4 * \beta_{Year} = 20000 * \beta_{Kilometers}$, the p-value also reject the hypothesis that the price of a car with 20000 kilometers record is equal to one 4 year younger.

5.

```
studentized Breusch-Pagan test

data:  p_p
BP = 8.0606, df = 2, p-value = 0.01777
```

$H_0 : \sigma_i = \sigma$, the p-value reject the homoskedasticity null hypothesis.

6.

```
> bptest(p_p, ~data_problemset1$Y)

studentized Breusch-Pagan test

data:  p_p
BP = 0.0037042, df = 1, p-value = 0.9515
```

```
> bptest(p_p, ~data_problemset1$HP)

studentized Breusch-Pagan test

data:  p_p
BP = 3.3104, df = 1, p-value = 0.06884
```

The p-value fail to reject the homoskedasticity related to HP or Year.

f

Estimate the model with and without robust standard errors. Do the robust errors –and, therefore, the conclusions on significance- differ much from regular OLS standard errors?

```
Call:
lm(formula = Car_Price ~ HP + Year + Kilometers, data = data_problemset1)

Residuals:
    Min      1Q  Median      3Q     Max
-6502.7 -2038.5  -121.2  1693.2 13537.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.664e+06  2.714e+05  -9.817 8.55e-15 ***
HP           6.628e+01  7.365e+00   8.999 2.65e-13 ***
Year         1.339e+03  1.356e+02   9.875 6.74e-15 ***
Kilometers  -3.581e-02  1.056e-02  -3.390  0.00115 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We already done without robust standard errors as shown above, now we estimate the model with robust standard errors:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -2.6640e+06  3.1547e+05 -8.4445 2.774e-12 ***
HP           6.6277e+01  8.5286e+00  7.7712 4.820e-11 ***
Year         1.3393e+03  1.5778e+02  8.4885 2.301e-12 ***
Kilometers  -3.5806e-02  1.1686e-02 -3.0640    0.0031 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We simply noticed that the conclusion hasn't changed whether with robust standard errors or not. There are unnoticeable changes in estimate coefficients, a slight increase in standard errors, and the significance drops. Thus we still reject the hypothesis that coefficients are equal to zero.

g

You work for a used-car sales company. Your boss asks you to write a short report (a maximum of half a page) on which features (other than kilometers) could be modified or added to the cars in store that would increase the selling price. Your report should talk about the statistical evidence you have found in your analysis, but in terms that your boss can understand and put into practice.

Based on our research, there are two main factors that might vary the price of cars other than kilometers. One of which is the Year the cars are manufactured, statistically speaking, a year newer leads to approximately 1339 euros raise in price. The other factor is Horse Power, per plus in horse power will result as 66 euros more in selling price. So in our further business, we should focus more on newly manufactured vehicles with bigger horse power, these cars are more likely to sale in a higher price.

code

```
library(sandwich)
library(ggplot2)
library(lmtest)
library(zoo)
library(corrplot)
library(car)

# a
p_k <- lm(Car_Price ~ Kilometers, data = data_problemset1)
summary(p_k)

# b
confint(p_k)

# c
library(ggfortify)
autoplot(p_k)

resettest(p_k, power = 3, type = 'fitted')
resettest(p_k, power = 3, type = 'regressor')

# d
cor_mat <- as.matrix(cor(data_problemset1, use = 'na.or.complete'))
corrplot(cor_mat, type = 'upper', tl.pos = 'lt', tl.cex = 0.6, tl.col = 'black')

p_a <- lm(Car_Price ~ HP + Year + Kilometers + ABS + ABD + MP, data = data_problemset1)
summary(p_a)
```

```
p_p <- lm(Car_Price ~ HP + Year + Kilometers, data = data_problemset1)
summary(p_p)

zero <- c('HP = 0', 'Year = 0', 'Kilometers = 0')
linearHypothesis(p_p, zero)

# e
H_1 <- c('Year = 1200')
linearHypothesis(p_p, H_1)

H_2 <- c('CC = 0', 'ABS = 0')
p_2 <- lm(Car_Price ~ HP + Year + Kilometers + CC + ABS, data = data_problemset1)
linearHypothesis(p_2, H_2)

H_3 <- c('CD = Gas')
colnames(data_problemset1) <- c('CC', 'HP', 'Y', 'Kilo', 'Gas', 'ABS', 'ABD', 'CD', 'AW',
                                'AC', 'MP', 'Price')
p_3 <- lm(Price ~ HP + Y + Kilo + CD + Gas, data = data_problemset1)
linearHypothesis(p_3, H_3)

H_4 <- c('-4 * Year = 20000 * Kilometers')
linearHypothesis(p_p, H_4)

bptest(p_p, ~fitted(p_p) + I(fitted(p_p)^2))

bptest(p_p, ~data_problemset1$Y)

bptest(p_p, ~data_problemset1$HP)

# f
coeftest(p_p, vcovHC(p_p, type = 'HC0'))
```