Assignment 1: The first on multiple regression

Barcelona Graduate School of Economics
Msc in Competition and Market Regulation

Course: Econometrics
Prof.: Javier Gómez Biscarri

Bendegúz István Biksi, Xia Mengzhen

January 1, 2016

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.

Our model of interest is:

$$CarPrice_i = \beta_0 + \beta_1 * Kilometers_i + \epsilon_i$$

Table 1: Regression table

	Car_Price
Kilometers	-0.0996***
	(-6.92)
Constant	25001.3***
	(15.53)
Observations	74
R^2	0.3997
$F_{(1,72)}$	47.934

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Consider the regression output above. One is able to see, that if we try to explain the variaton of the car prices, the variable kilometers is a reasonable good candidate for the explanation. The parameter of kilometers is 0.0996. It is statistically significant in every significance level, since it's t-statistic is higher than 3.9 (in absolute term). If a car has one more extra kilometer, than ceteris paribus, the price of the car will be lower by 0.0996 monetary unit.
The constant term is also statistically significant in every significance level as it's t-statistics is above 3.9. The meaning of the constant term is that: a car that has run 0 kilometers costs 25001.3, keeping everything the same. It is like an avarage price of a new car.
R^2 is 0.3997. Meaning that, the variation of kilometers explains almost 40% of the variation of the price. Another interpretation could be that the correlation between prices and kilometers are the square root of R^2 that is .6322 in absolute terms, because these two variables correlate negatively, so the correlation between price and kilometers is -0.6322.
The F-value here is simply the square of the t-value, because there is only one restriction. This value is way above the critical value of $F_{1,72}$, that is 3.84, so we can reject the $H_0 : \beta_{kilometers} = 0$. So again, kilometers have statistically significant explanatory power in determining a price of a car, in according to the t-test or the F-test.

b.)

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

	Car_Price	Std. Error	t	[95% Conf. Interval]
Kilometers	-0.0996***	.014389	-6.92	[-0.128,-0.0709]

Marginal effects; 95% confidence intervals in brackets

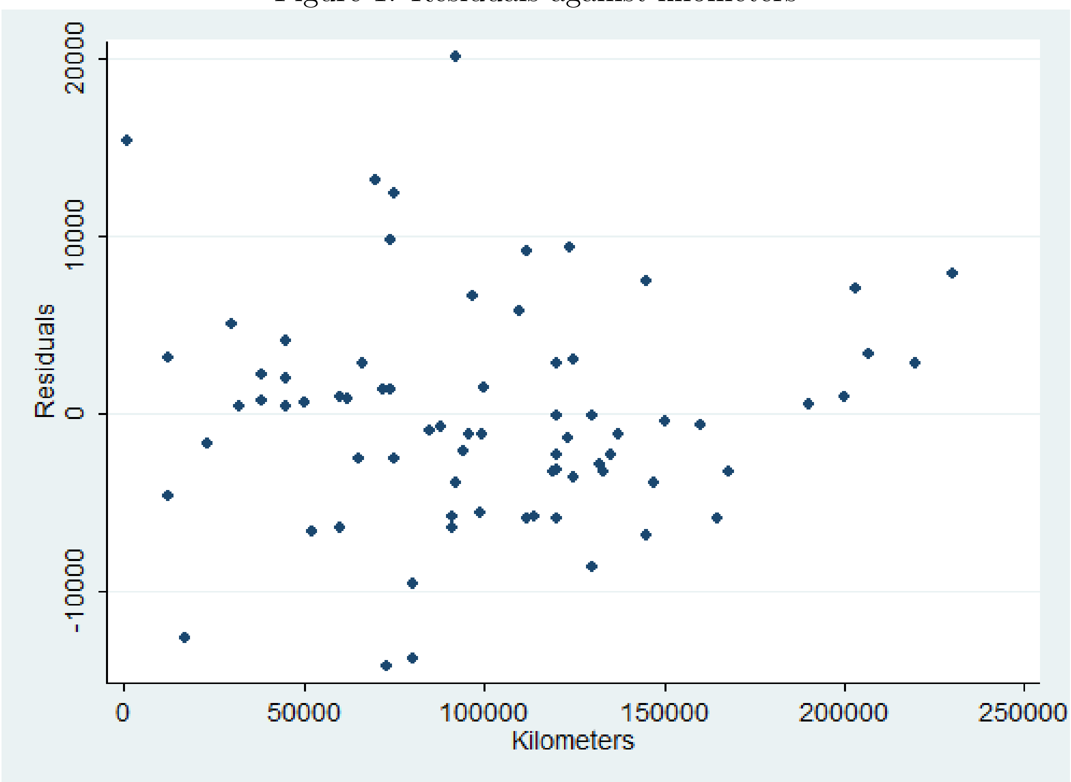
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As the coefficient of kilometers is -0.0996 and the standard error of the coefficient is .014389, one is able to compute a 95% confidence interval, that is $-0.0996 \pm 1.96 * 0.014389 = [-0.128, -0.0709]$. In the 95% of the cases the coefficient of kilometers would fall in the interval of $[-0.128, -0.0709]$. Therefore one extra kilometer would imply a change in the price with a value (in 95% of the cases) between -0.128 and -0.0709, ceteris paribus.

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?

Figure 1: Residuals against kilometers



There is no such evidence of nonlinear effect according to this graph. Nonlinearity could arrise from the fact that it does matter how many kilometers the car had already run, so the effect of one additional kilometer would not be the same for every values of kilometers. It seems logical to consider that a one unit change from 100 to 101 has a different effect on prices, than a one unit change from 10000 to 10001. To test wheter there is nonlinearity we can test that the model has or has not got omited variables.

Ramsey RESET test using powers of the fitted values of Car Price
Ho: model has no omitted variables
F(3, 69) = 1.73
Prob > F = 0.1692

It turns out from the omitted variable test that we can not reject the H_0 since the value of the F-statistic is smaller $1.73 < 2.74$, than the critical value of $F(3, 69)$. As we do not reject we take the null, so the model has no omitted variables, therefore there is no such an evidence for a nonlinear effect. To take into account a nonlinear effect we could generate the square of kilometers and estimate the following model:

$$CarPrice_i = \beta_0 + \beta_1 * Kilometers_i + \beta_2 * Kilometers_i^2 + \epsilon_i$$

If the parameter of $Kilometers^2$ is statistically significant, then there is evidence of a nonlinear effect.

Table 2: Dependent variable: Car Price

Variable	**Coefficient**	**(Std. Err.)**
$Kilometers$	-0.187**	(0.048)
$Kilometers^2$	0.000†	(0.000)
$Constant$	28711.713**	(2509.395)
N	74	
R^2	0.429	
$F_{(2,71)}$	26.655	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

From Table 2. it turns out that on 5% significance level or in 95% confidence interval, we can not reject the null hypothesis, that the effect of $Kilometers^2$ is 0. According to this, there is no statistically significant nonlinear effect of kilometers on prices on 5% significance level.

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?).

Our final model of choice is:

$$CarPrice_i = \beta_0 + \beta_1 * Kilometers_i + \beta_2 * HorsePower_i + \beta_3 * Year_i + \beta_4 * Gas_i + \epsilon_i$$

Table 3. Dependent variable: Car Price

Variables	Car_Price	t-stat.
Kilometers	-0.0372***	-3.92
	(0.00949)	
Horse Power	75.15***	10.84
	(6.934)	
Year	1,197***	9.48
	(126.3)	
Gas	-3,845***	-4.23
	(908.8)	
Constant	-2.380e+06***	-9.42
	(252,654)	
N	74	
R-squared	0.859	
F(4,69)	105.21	

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

All coefficient of the explanatory variables are statistacally significant even in 0.1% significance level, because every parameters' t-statistic is higher in absolute terms, than the critical value 3.9. $R^2 = .859$, so the variance of the explanatory variables explain the 85% of the variance of the dependent variable. The F-test is testing the null hypothesis, wether all the independent variables have joint significant explanatory power or not. Here the F-value is 105.21, that is way above the critical value of $F(4, 69) = 2.5$, so we can reject that every $\beta_i = 0, i = 1, ..4$.

To test wether the new additional variables have significant joint explanatory power compared to the simple on explanatory variable model, we can run a test with three null hypothesis.

Joint parameter test
$H_0 : \beta_{HorsePower} - \beta_{Year} = 0$
$\mathrm{H}_0 : \beta_{HorsePower} - \beta_{Gas} = 0$
$\mathrm{H}_0 : \beta_{HorsePower} = 0$
F(3, 69) = 75.02
Prob > F = 0.0000

As the F-value is 75.02, that is above $F(3, 69) = 2.74$ we can reject to the combined null hypothesis, so at least on of the parameters is not equal to 0.

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).

e.1. The price of the car goes down by 1200 euros per additional year old the car is.

$H_0 : \beta_{Year} = 1200$
F(1, 69) = 0.00
Prob > F = 0.9838

We can not reject the null hypothesis because the F-value is above it's critical value, so one extra year, ceteris paribus effect can be 1200 monetary unit. Here one should not forget that one extra year means that the car was designed one year later, so the car is one year younger if we add one to the year variable. So the contrary is also true, minus one year means one additional year to the car's age. So there is statistical evidence with 95% confidence interval, that one extra year to the car's age, means 1200 monetary unit decrease in it the price.

e.2. The number of CC and ABS have no impact on the price of the car.

Here we have to estimate a new model that is:

$$CarPrice_i = \beta_0 + \beta_1 * Kilometers_i + \beta_2 * HorsePower_i + \beta_3 * Year_i + \beta_4 * Gas_i + \beta_5 * CC_i + \beta_6 * ABS_i + \epsilon_i$$

$$H_0 : \beta_{CC} - \beta_{ABS} = 0$$
$$\mathrm{H}_0 : \beta_{CC} = 0$$
$$\mathrm{F}(\ 2,\ 67) = 0.01$$
$$\mathrm{Prob} > \mathrm{F} = 0.9854$$

We can not reject the combined null hypothesis, so the effect of the variables ABS and CC on car prices statistically significantly 0.

e.3. The effect of having a CD player is the same as the difference between a gas and a diesel car.

The estimated model is:

$$CarPrice_i = \beta_0 + \beta_1 * Kilometers_i + \beta_2 * HorsePower_i + \beta_3 * Year_i + \beta_4 * Gas_i + \beta_5 * CD_i + \epsilon_i$$

$$H_0 : \beta_{CD} - \beta_{Gas} = 0$$
$$\mathrm{F}(\ 1,\ 68) = 11.04$$
$$\mathrm{Prob} > \mathrm{F} = 0.0014$$

As the p-value is smaller than: .05, we can reject the null hypothesis, so the effect of having a CD player is not the same as the difference between a gas and a diesel car.

e.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.

Here the estimated model is in d.)

$$H_0 : -20000 * \beta_{Kilometers} - 4 * \beta_{Year} = 0$$
$$\mathrm{F}(\ 1,\ 69) = 39.51$$
$$\mathrm{Prob} > \mathrm{F} = 0.0000$$

As the p-value of the test is smaller than .05, we can reject the null hypothesis.

e.5. There is no heteroskedasticity problem (White's test).

White's test for:

$$H_0 : \text{homoskedasticity}$$
$$H_a : \text{unrestricted heteroskedasticity}$$
$$chi^2(13) = 37.16$$
$$\text{Prob} > \text{chi}^2 = 0.0004$$

As tha p-value is smaller, than .05, we can reject the null hypothesis, so heteroscedasticity occurs with 95% confidence interval.

e.6. There is no heteroskedasticity related to YEAR or HP (Breusch-Pagan test).

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity:

$$H_o\text{: Constant variance}$$
$$\text{Variables: Year HP}$$
$$chi^2(2) = 12.79$$
$$\text{Prob} > \text{chi}^2 = 0.0017$$

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?

	(No Rob. Err.)	(Yes Rob. Err.)
Variables	Car_Price	Car_Price
Kilometers	-0.0372***	-0.0372***
	(0.00949)	(0.0106)
HP	75.15***	75.15***
	(6.934)	(9.168)
Year	1,197***	1,197***
	(126.3)	(147.4)
Gas_1Diesel_0	-3,845***	-3,845***
	(908.8)	(912.9)
Constant	-2.380e+06***	-2.380e+06***
	(252,654)	(294,849)
Observations	74	74
R-squared	0.859	0.859

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The first model is without robust standard errors and the second model is with the robust standard errors. The errors are higher in the second model as expected, but the conclusions on significance differ not much compared to the regular OLS.

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.

According to the analysis above, we found that horsepower, year, and cars running diesel have significant effect on the cars' prices. Meanwhile, we have found that the CC and ABS has no impact on car prices.

With high probability of this situation will happen, that one **add**itional **horsepower increase** the **price** of a car, keeping everything the same, with at least 56, maximum 93, on average with 75 monetary units.

Younger cars can be sold for **higher price**s. On avarage if a car is one year younger, compared to a similar but one year older car, the younger car's price is higher with 1200 monetary units. One year less means at least 900, maximum 1500 monetary units.

Cars running on **diesel** have **higher price**s compared to a similar but a gas user car. On avarage a diesel car's price is higher with 3845 monetary units. The difference between the price of a gas and a diesel user car can be 2024 up to 5666 in most of the cases, keeping every other caracterstics the same.

That is to say, more horsepower, higher the price; in contrast, the older the car is, the lower the price. As for the gas and diesel, we found that cars running with diesel have higher prices.