Introductory Econometrics: Computer Problem Sheet 1

Sketches of Answers follow the Problem sheet*

I am assuming that you know the basics of Excel. For instance, I assume you know how to load in data and manipulate variables (e.g. takes log, squares, etc. of variables using the formula bar). Regressions can be run using the options which appear when you click on Tools then Data analysis. You will have a tutor in the computer lab to help you with these things. The purpose of this problem set is for you to gain familiarity with running regressions in Excel and interpreting results. I would encourage you to make sure you know the basic commands in your computer lab (if necessary, asking the tutor for help). Then spend some time by yourself experimenting with this data set.

This computer session uses data set HPRICE.XLS (available through MyPlace) which contains data on N=546 houses sold in Windsor, Canada. Our dependent variable, Y, is the sales price of the house in Canadian dollars. The explanatory variables included in this data set are:

- the lot size of the property (in square feet)
- the number of bedrooms
- the number of bathrooms
- the number of storeys (excluding the basement).
- A dummy variable = 1 if house has a driveway (=0 otherwise)
- A dummy variable = 1 if house has a recreation room (=0 otherwise)
- A dummy variable = 1 if house has a basement (=0 otherwise)
- A dummy variable = 1 if house has gas central heating (=0 otherwise)
- A dummy variable = 1 if house has air conditioning (=0 otherwise)
- The size of garage (number of cars it will hold)
- A dummy variable = 1 if house is in a desirable neighbourhood (=0 otherwise)

In this session, I want you to carry out an empirical analysis of this data set and discuss your empirical results as you would in an empirical project or dissertation *using only the methods associated with ordinary least squares*. The following questions should help structure this process:

i) Run a regression of the dependent variable on all the explanatory variables. How would you interpret the coefficient estimates? Does the interpretation of the dummy variables differ from the other explanatory variables?

ii) Are all the explanatory variables statistically significant? Why? If you find insignificant variables, omit them from the regression and repeat part i).

iii) Is there evidence of multicollinearity?

iv) What is the R^2? How would you interpret this number?

v) Now consider some extensions of the basic model. Generate a new explanatory variable which is the dummy for “desirable neighbourhood” times the “lot size” variable. Run a regression including all the explanatory variables described above plus this new explanatory variable. How do you interpret the coefficient on this new explanatory variable? Does inclusion of this new variable alter any of your results in parts i) through iv)?

vi) Generate a new variable which is “lot size” squared and include it in the regression described in part v). How do you interpret the coefficient on this new variable? Does inclusion of this new variable alter any of your results in parts i) through v)?

Answer to part i):

Here is Excel output for the regression:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.820441
R Square	0.673124
Adjusted R Square	0.66639
Standard Error	15423.19
Observations	546

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	11	2.62E+11	2.38E+10	99.96774	6.2E-122
Residual	534	1.27E+11	2.38E+08		
Total	545	3.89E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-4038.35	3409.471	-1.18445	0.236762	-10736	2659.27
size	3.546303	0.3503	10.12362	3.73E-22	2.858168	4.234438
bed	1832.003	1047	1.749764	0.080733	-224.741	3888.748
bath	14335.56	1489.921	9.621691	2.57E-20	11408.73	17262.38
stories	6556.946	925.2899	7.086369	4.37E-12	4739.291	8374.6
drive	6687.779	2045.246	3.269914	0.001145	2670.065	10705.49
rec	4511.284	1899.958	2.374413	0.017929	778.976	8243.592
basement	5452.386	1588.024	3.43344	0.000642	2332.846	8571.926
gas	12831.41	3217.597	3.987885	7.6E-05	6510.706	19152.11
air	12632.89	1555.021	8.123935	3.15E-15	9578.182	15687.6
garage	4244.829	840.5442	5.050096	6.07E-07	2593.65	5896.008
location	9369.513	1669.091	5.613544	3.19E-08	6090.724	12648.3

As an example of how to interpret coefficients, consider the coefficient for the first explanatory variable, lot size. It can be seen that $\hat{\beta}_1$=3.55. Below are some (very similar) ways of verbally stating what this value means.

1. "An extra square foot of lot size will tend to add another $3.55 on to the price of a house, ceteris paribus."
2. "If we consider houses with the same number of bedrooms, bathrooms and storeys, an extra square foot of lot size will tend to add another $3.55 onto the price of the house."
3. "If we compare houses with the same number of bedrooms, bathrooms and storeys, those with larger lots tend to be worth more. In particular, an extra square foot of lot size is associated with an increased price of $3.55."

It is worth expanding on the motivation for the latter two expressions. We cannot simply say that "houses with bigger lots are worth more" since this is not the case (e.g. some nice houses on small lots will be worth more than poor houses on large lots). However, we can say that "if we consider houses that vary in lot size, but are

comparable in other respects, those with larger lots tend to be worth more". The two expressions above explicitly incorporate the qualification "but are comparable in other respects".

Dummy variables can be interpreted as affecting the intercept of the regression line. For instance, the coefficient on the basement dummy is 5452.39. So you can say that houses with a basement have a regression line with an intercept 5452.39 higher than those without a basement. Alternatively, you can say that, controlling for all other house characteristics, those with a basement tend to be worth $5452.39 more than those without.

Answer to part ii).

> Look of at the P-values (labelled as P>|t| in Excel. P-values of less than .01 indicate significance at the 1% level, P-values of less than .05 indicate significance at the 5% level, etc.

Using the 5% level of significance all of the explanatory variables except the number of bedrooms are significant (since their p-vales are less than .05). So you could re-run this regression excluding number of bedrooms (I will not put the results for the new regression here – they are not very different from those in the table above).

Note that the p-value on number of bedrooms is .08, so this explanatory variable is significant at the 10% level of significance.

Note that the intercept is not significant. Usually people just include an intercept in every regression they result and do not worry about its significance (but if you want you could delete the intercept).

Answer to part iii).

Calculate the correlation matrix for the explanatory variables. You can see that none of these correlations is near 1 or -1 (e.g. the highest ones are around .3. none anywhere near .9 or -.9). This provides informal evidence that multicollinearity is not a problem.

Answer to part iv)

In the regression above, the R-squared is .82. So 82% of the variation of house prices can be explained by the explanatory variables.

Answers to part v) and vi)
I will not provide detailed answers to these questions. These are meant just to get students experimenting with extending the model to allow for interactions or nonlinearities. Statistical things relating to the new explanatory variables are the same as for any explanatory variable (e.g. if its p-value is less than .05 then it is significant). My choices of what new explanatory variables to add are merely illustrative. In a real empirical application, you would experiment extensively with many different alternatives, running many different regressions with different explanatory variables

(e.g. trying interactions between different variables, squared terms for different variables, omitting explanatory variables which are insignificant, etc. etc.).

Interactions of dummies with a regular explanatory variable allow for you to have different values for the slope coefficient on the regular explanatory variable depending on whether the dummy is 0 or 1. So the slope coefficient for “lot size” can be one thing in desirable neighbourhoods and something different in bad neighbourhoods. This allows for features like: “in desirable locations, lot size has a big effect on price. In undesirable locations, lot size does not matter”.

Adding a squared term allow for a quadratic relationship between an explanatory variable and the dependent variable. Your exact results will depend on which regression you are adding the quadratic term to (e.g. is it the original one in part i), or one with insignificant variables omitted, or the regression of part iv) with the interaction term included, etc.). In the regression I ran I found lotsize-squared to be significant and negative. Along with the coefficient on lotsize being significant and positive, this indicates a quadratic relationship. That is the effect of lotsize on house price increases and then levels off (and eventually starts dropping again – but this latter probably occurs at very high house prices outside the range of our data).