Economics 167, Econometrics

Computer Lab 3

1. Open *STATA*.
2. Click *File*, then *Open*.
3. Locate the *STATA* dataset file *housing.dta* and open it.
4. The dataset is composed of twelve variables taken from "Semiparametric Estimation of a Hedonic Price Function", Anglin and Gençay (Journal of Applied Econometrics, Vol.11, 633-648, 1996). It contains the sales prices (*price*) of $n = 546$ houses, sold in July, August, and September 1987, in the city of Windsor, Canada, along with some important house features:
   - the lot size of the property in square feet (*lotsize*);
   - the number of bedrooms (*bedrooms*);
   - the number of full bathrooms (*bathrms*);
   - the number of garage places (*garagepl*);
   - the number of stories (*stories*);
   - a dummy for the presence of a driveway (*driveway*);
   - a dummy for the presence of a recreational room (*recroom*);
   - a dummy for the presence of a full basement (*fullbase*);
   - a dummy for the presence of central air conditioning (*airco*);
   - a dummy for being located in a preferred area (*prefarea*);
   - a dummy for using gas for hot water heating (*gashw*).
5. In this lab we will consider the relationship between house sale prices and house characteristics. We will estimate a price function, usually referred to as a *hedonic price function* in the economic literature. A hedonic price refers to the implicit price of a certain attribute as revealed by the sale price of a house, where a house is considered as a bundle of such attributes. In this context, the hedonic price function describes the expected price (or log price) of a house as a function of a number of characteristics.
6. As usual, we look at some summary statistics. Type *summarize* in the command

window to get:

Variable	Obs	Mean	Std. Dev.	Min	Max
price	546	68121.6	26702.67	25000	190000
lotsize	546	5150.266	2168.159	1650	16200
bedrooms	546	2.965201	.7373879	1	6
bathrms	546	1.285714	.5021579	1	4
stories	546	1.807692	.8682025	1	4
driveway	546	.8589744	.3483672	0	1
recroom	546	.1776557	.3825731	0	1
fullbase	546	.3498168	.4773493	0	1
gashw	546	.0457875	.2092157	0	1
airco	546	.3168498	.465675	0	1
garagepl	546	.6923077	.8613066	0	3
prefarea	546	.2344322	.4240319	0	1

7. We begin by considering a model that explains the logarithm of the sale price from the logarithm of the lot size, the numbers of bedrooms and bathrooms, and the presence of air conditioning. That is, we estimate the following linear regression model:

$$log\left(price_i\right) = \beta_0 + \beta_1 log\left(lotsize_i\right) + \beta_2 bedrooms_i + \beta_3 bathrms_i + \beta_4 airco_i + \varepsilon_i.$$

Note that we are taking the log of some variables and are keeping some other variables in levels. What is the meaning that should be given to, say, coefficient β_1? Since both variables *price* and *lotsize* are logged, then β_1 is an elasticity: *ceteris paribus*, it represents the percent change in price associated with a percent change in the size of the lot. *Ceteris paribus*, the coefficients β_2 and β_3 represent the percent change in price associated with a unit change in the number of bedrooms and bathrooms, respectively. Finally, β_4 represents the percent difference in price between a house with central air conditioning and a house without central AC, all the other factors kept constant.

Also note that, if we want to run the regression above in *STATA*, we first need to create the new variables *lprice=log(price)* and *llotsize=log(lotsize)*. We can do this by typing *gen lprice=log(price)* and *gen llotsize=log(lotsize)* in the command window. We can safely take the log of the two variables because, as we could also verify in the table of summary statistics, both are strictly positive. At this point we can run our regression, *regress lprice llotsize bedrooms bathrms airco*:

Source	SS	df	MS
Model	42.790971	4	10.6977427
Residual	32.6221992	541	.060299814
Total	75.4131702	545	.138372789

Number of obs	= 546
F(4, 541)	= 177.41
Prob > F	= 0.0000
R-squared	= 0.5674
Adj R-squared	= 0.5642
Root MSE	= .24556

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llotsize	.4004218	.0278122	14.40	0.000	.3457886	.455055
bedrooms	.0776997	.0154859	5.02	0.000	.0472798	.1081195
bathrms	.2158305	.0229961	9.39	0.000	.1706578	.2610031
airco	.2116745	.0237213	8.92	0.000	.1650775	.2582716
_cons	7.093777	.231547	30.64	0.000	6.638935	7.548618

The four variables included in the regression do a fairly good job at explaining the variance of the house prices, as we can see from the R^2, in this case equal to 56.74%. All the coefficients are statistically significant (high t ratios and corresponding p-values virtually equal to zero). According to this estimated equation, a house that has central AC is expected to sell at a 21.17% higher price than a house without it, if both houses have the same number of bedrooms and bathrooms and the same lot size. A 1% larger lot increases the expected sale price by 0.4%. An additional bedroom or bathroom is expected to increase the sale price by 7.77% or 21.58%, respectively.

8. At this point we may wonder if the linear specification we have adopted is good enough, or if we should take into account the possibility that the relationship that we are trying to estimate between the dependent variable and the set of regressors is non-linear. An easy way to do this is to run a RESET test. After estimating an equation in *STATA*, we can type *estat ovtest* to run this test:

```
Ramsey RESET test using powers of the fitted values of lprice
       Ho:  model has no omitted variables
                 F(3, 538) =      0.56
                  Prob > F =      0.6408
```

Since the F statistic is low and the corresponding p-value pretty high, we cannot reject the null. In other words, we do not find evidence of non-linearity in the model.

By looking at the degrees of freedom associated with the test statistic, we can try to figure what *STATA* is doing. The first degree of freedom is 3, the second is 538. From class we know that, if we want to run a RESET test, we need to do the following:

- estimate the regression;
- take the fitted values of the dependent variable, in this case $\widehat{log\,(price_i)}$;
- run the auxiliary regression

$$\begin{aligned} log\,(price_i) \;\; &= \;\; \beta_0 + \beta_1 log\,(lotsize_i) + \beta_2 bedrooms_i + \beta_3 bathrms_i + \beta_4 airco_i \\ &\quad + \alpha_2 \widehat{log\,(price_i)}^2 + \alpha_3 \widehat{log\,(price_i)}^3 + ... + \alpha_Q \widehat{log\,(price_i)}^Q + \varepsilon_i; \end{aligned}$$

- run the F test:

$$\begin{cases} H_0 : \alpha_2 = \alpha_3 = ... = \alpha_Q = 0 \\ H_1 : \text{not } H_0 \end{cases},$$

  using an $F_{Q-1,n-K-Q+1}$ distribution to compute the critical values.

In this case, $Q-1 = 3 \Longrightarrow Q = 4$ and $n-K-Q+1 = 546-5-Q+1 = 538 \Longrightarrow Q = 4$. In practice, and by default, *STATA* sets $Q = 4$ to run the RESET test. Unfortunately, there is no way to change this option at the moment. However, we know how to run a RESET test "*manually*" and we can thus use an alternative approach to run the same test with a different Q. Following the four steps indicated above, after estimating the regression, we need to compute the fitted values of the dependent variable in the model. We can do this by typing *predict lprice_hat* (i.e., we are assigning the name *lprice_hat* to the fitted dependent variable). Then we can run the auxiliary test regression with $Q = 3$, for example. But we first need to generate $\widehat{log\,(price_i)}^2$ and $\widehat{log\,(price_i)}^3$. Type *gen lprice_hat2=lprice_hat^2* and *gen lprice_hat3=lprice_hat^3*.

Finally type *regress lprice llotsize bedrooms bathrms airco lprice_hat2 lprice_hat3*:

```
      Source |       SS       df       MS              Number of obs =     546
-------------+------------------------------           F(  6,   539) =  118.27
       Model |  42.8590469     6  7.14317448           Prob > F      =  0.0000
    Residual |  32.5541233   539  .06039726           R-squared     =  0.5683
-------------+------------------------------           Adj R-squared =  0.5635
       Total |  75.4131702   545  .138372789           Root MSE      =  .24576

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |  -33.38085   35.81353    -0.93   0.352    -103.7321    36.97036
    bedrooms |  -6.477424   6.949758    -0.93   0.352    -20.12935    7.174506
     bathrms |  -17.99997   19.30605    -0.93   0.352    -55.92429    19.92435
       airco |  -17.65349    18.9385    -0.93   0.352    -54.85581    19.54883
 lprice_hat2 |   7.474941   7.984481     0.94   0.350    -8.209574    23.15946
 lprice_hat3 |   -.220558   .2374835    -0.93   0.353    -.6870646    .2459486
       _cons |  -274.2659   300.7338    -0.91   0.362    -865.0199    316.4881
------------------------------------------------------------------------------
```

This is just a test regression. It is not intended to produce any meaningful results. We should only use it to run the F test needed for the RESET test. We should test the null that the coefficients associated to the variables *lprice_hat2* and *lprice_hat3* are jointly equal to zero. We type *test (lprice_hat2=lprice_hat3=0)* and get

```
 ( 1)  lprice_hat2 - lprice_hat3 = 0
 ( 2)  lprice_hat2 = 0

       F(  2,   539) =    0.56
            Prob > F =    0.5695
```

Not even in this case are we able to reject the null that the model has no omitted variables (or that it should be based on a different functional form).

9. *STATA* automatically computes and reports the R^2 and the adjusted R^2, $\overline{R}^2$, of a model in the regression output. If we want to compute the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) of the model, we should type *estat ic* (of course, only after estimating the first regression again, otherwise the information criteria would be calculated for the auxiliary regression we have used for running the RESET test):

```
-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    546   -234.2995   -5.528551      5      21.0571     42.5702
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

These four criteria can be used for comparing non-nested models with the same dependent variable. However, with the exception of the R^2, they can also be used as a comparison device for nested models. Let us try to include all the other characteristics on the right-hand side of the linear model. So, we estimate a new equation by typing *regress lprice llotsize bedrooms bathrms airco driveway recroom fullbase gashw* *garagepl prefarea stories*:

Source	SS	df	MS
Model	51.7748825	11	4.7068075
Residual	23.6382877	534	.044266456
Total	75.4131702	545	.138372789

Number of obs	=	546
F(11, 534)	=	106.33
Prob > F	=	0.0000
R-squared	=	0.6865
Adj R-squared	=	0.6801
Root MSE	=	.2104

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llotsize	.3031258	.0266931	11.36	0.000	.2506895	.3555622
bedrooms	.034399	.0142741	2.41	0.016	.0063588	.0624392
bathrms	.1657644	.0203286	8.15	0.000	.1258306	.2056981
airco	.1664238	.0213386	7.80	0.000	.1245059	.2083417
driveway	.110202	.0282261	3.90	0.000	.0547542	.1656498
recroom	.0579739	.0260528	2.23	0.026	.0067953	.1091524
fullbase	.1044881	.0216916	4.82	0.000	.0618768	.1470994
gashw	.1790231	.0438933	4.08	0.000	.0927984	.2652477
garagepl	.0479543	.0114765	4.18	0.000	.0254097	.070499
prefarea	.131851	.0226692	5.82	0.000	.0873192	.1763827
stories	.0916851	.0126144	7.27	0.000	.0669051	.116465
_cons	7.745093	.2163352	35.80	0.000	7.32012	8.170065

As expected, the R^2 is bigger in this model than in the first we estimated (68.65% vs 56.74%), since it includes the same regressors contained in the first plus a few more. So, the first model is nested in the second one – i.e., we can obtain the first model by imposing zero-restrictions on the coefficients associated with the regressors that are not in common. As such, the R^2 is not a reliable indicator, if we want to figure out whether the second model is better than the first. But we can test whether the increase in the R^2 we observe is significant by testing the null that the coefficients associated with the extra regressors (with respect to the first model) are jointly equal to zero. The t statistics in the last regression suggest that all the model parameters are statistically significant, if taken individually. To run the joint test of significance that we need, we should impose 7 zero-restrictions in the model by typing *test(driveway=recroom=fullbase=gashw=garagepl=prefarea=stories=0)*:

```
( 1)  driveway - recroom = 0
( 2)  driveway - fullbase = 0
( 3)  driveway - gashw = 0
( 4)  driveway - garagepl = 0
( 5)  driveway - prefarea = 0
( 6)  driveway - stories = 0
( 7)  driveway = 0

      F(  7,   534) =   28.99
           Prob > F =    0.0000
```

Given the value of the corresponding F statistic, we reject the null and conclude that the increase in the R^2 is indeed significant. The bigger $\overline{R}^2$ that we observe in the second model (68.01% vs 56.42%) leads us to the same conclusion, that the second model should be preferred to the first one according to this criterion. Finally, we may want to compare the other two information criteria of the two models. Type again *estat ic*:

Model	Obs	ll(null)	ll(model)	df	AIC	BIC
.	546	-234.2995	82.41164	12	-140.8233	-89.19186

Note: N=Obs used in calculating BIC; see **[R] BIC note**

and notice that both the AIC and the BIC are smaller in the second model. All this evidence should tell us that, most probably, the second model does a better job at explaining the variance of the dependent variable. Moreover, the model coefficients are all positive, which makes economic and intuitive sense. If we want to quickly look at possible misspecification problems, we can run another RESET test by typing *estat ovtest*:

```
Ramsey RESET test using powers of the fitted values of lprice
      Ho:  model has no omitted variables
                F(3, 531) =      0.36
                 Prob > F =      0.7804
```

which suggests that we probably do not have issues of nonlinearities in our specification.

10. To conclude, let us assume the following situation. After estimating the first model, we want to figure out whether an alternative non-nested model performs better or worse. Specifically, we are interested in the two alternative model specifications

$$\begin{aligned} log\left(price_i\right) &= \beta_0 + \beta_1 log\left(lotsize_i\right) + \beta_2 bedrooms_i + \beta_3 bathrms_i \qquad \text{(A)} \\ &\quad + \beta_4 airco_i + \varepsilon_i \\ log\left(price_i\right) &= \delta_0 + \delta_1 log\left(lotsize_i\right) + \delta_2 bedrooms_i + \delta_3 prefarea_i \qquad \text{(B)} \\ &\quad + \delta_4 recroom_i + \delta_5 garagepl_i + \nu_i. \end{aligned}$$

The two models are non-nested. To test whether model (A) performs better than model (B), we can again look at the corresponding R^2's, $\overline{R}^2$'s, AIC's, and BIC's. Let us type *regress lprice llotsize bedrooms prefarea recroom garagepl* to estimate model (B) and then *estat ic*:

```
      Source |       SS       df       MS              Number of obs =     546
-------------+------------------------------           F(  5,   540) =  108.22
       Model |  37.7449728     5  7.54899456           Prob > F      =  0.0000
    Residual |  37.6681974   540  .069755921           R-squared     =  0.5005
-------------+------------------------------           Adj R-squared =  0.4959
       Total |  75.4131702   545  .138372789           Root MSE      =  .26411

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    llotsize |   .3881969   .0315983    12.29   0.000     .3261262    .4502676
    bedrooms |   .1303206   .0156237     8.34   0.000     .0996299    .1610113
    prefarea |   .1666083   .0275966     6.04   0.000     .1123985    .2208181
     recroom |   .1399483   .0303646     4.61   0.000      .080301    .1995956
    garagepl |   .0698037   .0141512     4.93   0.000     .0420055    .0976018
       _cons |   7.273568    .262447    27.71   0.000     6.758026     7.78911
------------------------------------------------------------------------------

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    546   -234.2995   -44.79227      6     101.5845    127.4003
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Note that $R_A^2 = 56.74\% > R_B^2 = 50.05\%$, $\overline{R}_A^2 = 56.42\% > \overline{R}_B^2 = 49.59\%$, $AIC_A = 21.06 < AIC_B = 101.58$, $BIC_A = 42.57 < BIC_B = 127.40$. All four criteria suggest that we should favor model (A) over model (B). To be really sure of this, we can run an encompassing test to compare the two models. Take all the regressors in the two alternative specifications, (A) and (B), and run the regression

$$\begin{aligned} log\left(price_i\right) &= \gamma_0 + \gamma_1 log\left(lotsize_i\right) + \gamma_2 bedrooms_i + \gamma_3 bathrms_i \\ &\quad + \gamma_4 airco_i + \gamma_5 prefarea_i + \gamma_6 recroom_i + \gamma_7 garagepl_i + \eta_i, \end{aligned}$$

by typing *regress lprice llotsize bedrooms bathrms airco prefarea recroom garagepl:*

Source	SS	df	MS
Model	47.1172937	7	6.73104196
Residual	28.2958765	538	.052594566
Total	75.4131702	545	.138372789

Number of obs	=	546
F(7, 538)	=	127.98
Prob > F	=	0.0000
R-squared	=	0.6248
Adj R-squared	=	0.6199
Root MSE	=	.22934

lprice	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
llotsize	.3164688	.0280224	11.29	0.000	.261422	.3715157
bedrooms	.069939	.0144934	4.83	0.000	.0414685	.0984096
bathrms	.2025992	.021614	9.37	0.000	.160141	.2450574
airco	.1918874	.0222809	8.61	0.000	.1481192	.2356556
prefarea	.1589885	.0239885	6.63	0.000	.111866	.2061111
recroom	.0999257	.0265365	3.77	0.000	.047798	.1520535
garagepl	.0526689	.0123569	4.26	0.000	.0283953	.0769426
_cons	7.759381	.2319676	33.45	0.000	7.303708	8.215054

Then run the two statistical tests

$$\begin{cases} H_0 : \gamma_5 = \gamma_6 = \gamma_7 = 0 \\ H_1 : \text{not } H_0 \end{cases} ;$$

$$\begin{cases} H_0 : \gamma_3 = \gamma_4 = 0 \\ H_1 : \text{not } H_0 \end{cases} .$$

In both cases, we can use an F test (or a Wald test) to test the two null hypotheses. If we reject the null in the first test, that is evidence against model (A). If we reject the null in the second test, that is evidence against model (B). Type *test(prefarea=recroom=garagepl=0)* to run the first test, *test(bathrms=airco=0)* to run the second:

```
( 1)  prefarea - recroom = 0
( 2)  prefarea - garagepl = 0
( 3)  prefarea = 0

      F(  3,   538) =   27.42
           Prob > F =    0.0000

( 1)  bathrms - airco = 0
( 2)  bathrms = 0

      F(  2,   538) =   89.10
           Prob > F =    0.0000
```

We reject the null in both cases. In other words, we find statistical evidence against both models at the same time. What shall we do, then? If we really want to choose on of these two models, we should look at the R^2's, $\overline{R}^2$'s, AIC's, and BIC's as we did above and select one of the two specifications on the basis of the information we get from these criteria. Alternatively, we should seriously think about a completely different specification for the dependent variable. For example, we may want to use the model we estimated earlier with all the characteristics on the right-hand side of the equation. As we could see, that model exhibits good statistical properties and makes reasonable economic sense.