3. Regression with endogenous variable in log

```
. reg lprice bdrms sqrft lotsize

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   46.13
       Model |  4.98917377     3  1.66305792            Prob > F      =  0.0000
    Residual |  3.02842975    84  .036052735            R-squared     =  0.6223
-------------+------------------------------           Adj R-squared =  0.6088
       Total |  8.01760352    87  .092156362            Root MSE      =  .18988

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bdrms |   .0252388   .0285928     0.88   0.380    -.0316211    .0820987
       sqrft |   .0003641    .000042     8.67   0.000     .0002806    .0004477
     lotsize |   5.60e-06   2.04e-06     2.75   0.007     1.55e-06    9.65e-06
       _cons |   4.759375   .0935361    50.88   0.000     4.573369    4.945382
------------------------------------------------------------------------------
```

3.1. Prediction of average price for different values of bdrms sqrft lotsize
Assume residuals are normal

$\log y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ gives $E(\log y \mid x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

and $E(y \mid x) = e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k} e^{\sigma_u^2/2}$

For prediction, we use: $\hat{y} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_k x_k} e^{\hat{\sigma}_u^2/2}$

```
. reg lprice bdrms0 sqrft0 lotsize0

      Source |       SS           df       MS      Number of obs   =        88
-------------+----------------------------------   F(3, 84)        =     46.13
       Model |  4.98917377         3  1.66305792   Prob > F        =    0.0000
    Residual |  3.02842975        84  .036052735   R-squared       =    0.6223
-------------+----------------------------------   Adj R-squared   =    0.6088
       Total |  8.01760352        87  .092156362   Root MSE        =    .18988

------------------------------------------------------------------------------
      lprice |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bdrms0 |   .0252388   .0285928     0.88   0.380    -.0316211    .0820987
      sqrft0 |   .0003641    .000042     8.67   0.000     .0002806    .0004477
    lotsize0 |   5.60e-06   2.04e-06     2.75   0.007     1.55e-06    9.65e-06
       _cons |   5.431684   .0278272   195.19   0.000     5.376346    5.487022
------------------------------------------------------------------------------
```

For houses with bdrms=3, sqrt=1500, and lotsize=9000

$\hat{y} = e^{5.43} e^{.036/2} = 228.5 * 1.018 = 232.6$ thousand $

3.1. Choosing between regression on price or on log(price) : Use $\left[\text{cor}(y,\hat{y})\right]^2$

1) First do the regression on lprice (above) and get the prediction for all prices based on the previous method: $\hat{y} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1+\ldots+\hat{\beta}_k x_k} e^{\hat{\sigma}^2/2}$

```
. qui reg lprice bdrms sqrft lotsize
. predict lpricehat
. g pricehat=exp(lpricehat)*exp(.036052735/2)

. sum price pricehat

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        88     293.546    102.7134        111        725
    pricehat |        88    293.8179    84.32516   199.3371   608.2882


. correl price pricehat
(obs=88)

             |    price pricehat
-------------+------------------
       price |   1.0000
    pricehat |   0.8372   1.0000
```

R2 = (0.8372)² = .70090384

2) Now compare with the R2 obtained in the linear regression in section 1: `0.6724`

Recall that R2 in the linear model is also equal to $\left[\text{cor}(y,\hat{y})\right]^2$

```
. reg price bdrms sqrft lotsize

      Source |       SS       df       MS              Number of obs =      88
-------------+------------------------------           F(  3,    84) =   57.46
       Model |  617130.701     3  205710.234           Prob > F      =  0.0000
    Residual |  300723.805    84   3580.0453           R-squared     =  0.6724
-------------+------------------------------           Adj R-squared =  0.6607
       Total |  917854.506    87  10550.0518           Root MSE      =  59.833

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       bdrms |   13.85252   9.010145     1.54   0.128    -4.065141    31.77018
       sqrft |   .1227782   .0132374     9.28   0.000     .0964541    .1491022
     lotsize |   .0020677   .0006421     3.22   0.002     .0007908    .0033446
       _cons |  -21.77031   29.47504    -0.74   0.462    -80.38466    36.84405
------------------------------------------------------------------------------
```

What do you conclude?

4. Summary on choosing the variables and or functional form

For the *x* variables

a. Testing for the relevance of single variable *x*: use a t-test on its parameter

b. Testing for the relevance of several variables *x*: use an F-test for the joint test of their parameters equal to 0

c. Testing for 2 specifications that do not have the same *x* variables (for example log(distance) vs distance and distance^2): use the adjusted R-squared to choose the best specification

For the *y* variable:

Choosing between a model on log(price) or on price: get predictions and compare the square of the correlation between predictions and observations

```
. reg bwght cigs parity faminc motheduc fatheduc

      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  5,  1185) =    9.55
       Model |  18705.5567     5  3741.11135           Prob > F      =  0.0000
    Residual |  464041.135  1185  391.595895           R-squared     =  0.0387
-------------+------------------------------           Adj R-squared =  0.0347
       Total |  482746.692  1190  405.669489           Root MSE      =  19.789

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------:
        cigs |  -.5959362   .1103479    -5.40   0.000    -.8124352   -.3794373
      parity |   1.787603   .6594055     2.71   0.007     .4938709    3.081336
      faminc |   .0560414   .0365616     1.53   0.126    -.0156913    .1277742
    motheduc |  -.3704503   .3198551    -1.16   0.247    -.9979957    .2570951
    fatheduc |   .4723944   .2826433     1.67   0.095    -.0821426    1.026931
       _cons |   114.5243   3.728453    30.72   0.000     107.2092    121.8394
------------------------------------------------------------------------------

. reg bwght cigs parity faminc

      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  3,  1187) =   14.95
       Model |  17579.8997     3  5859.96658           Prob > F      =  0.0000
    Residual |  465166.792  1187  391.884408           R-squared     =  0.0364
-------------+------------------------------           Adj R-squared =  0.0340
       Total |  482746.692  1190  405.669489           Root MSE      =  19.796

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5978519   .1087701    -5.50   0.000    -.8112549   -.3844489
      parity |   1.832274   .6575402     2.79   0.005     .5422035    3.122345
      faminc |   .0670618   .0323938     2.07   0.039     .0035063    .1306173
       _cons |   115.4699   1.655898    69.73   0.000     112.2211    118.7187
------------------------------------------------------------------------------

. reg  bwght cigs parity fatheduc motheduc

      Source |       SS       df       MS              Number of obs =    1191
-------------+------------------------------           F(  4,  1186) =   11.34
       Model |  17785.5192     4   4446.3798           Prob > F      =  0.0000
    Residual |  464961.173  1186  392.041461           R-squared     =  0.0368
-------------+------------------------------           Adj R-squared =  0.0336
       Total |  482746.692  1190  405.669489           Root MSE      =   19.8

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.6059168   .1102182    -5.50   0.000    -.8221613   -.3896724
      parity |   1.765005   .6596156     2.68   0.008     .4708616    3.059149
    fatheduc |   .5795404   .2740186     2.11   0.035     .0419251    1.117156
    motheduc |  -.2764183   .3140955    -0.88   0.379     -.892663    .3398265
       _cons |   113.7365   3.694953    30.78   0.000     106.4871    120.9858
------------------------------------------------------------------------------
```