Problems on regression analysis (actual estimation of regression models)

Note that while you will not be asked to estimate regression models for the exam, the final project asks you to estimate them. I have posted these problems so that you can practice estimating models and compare the results that you obtain with the results that I obtain to make sure that you are estimating the models correctly.

General comment: What I have done for the regression models is use output from Stata rather than Excel because it is so much easier to understand (essentially, everything is labeled for you).

1. Consider the case study at the end of chapter 5 on Mrs. Smyth's pies (I know you have some of these answers already – the point of replication is to make sure that you can properly conduct the procedure). The data are posted on the website as MrsSmythsPies.txt. The variables included are:

Location	The city from which the observation came
Year-Qtr	The year and quarter for the observation
Sales	The quantity sold in that location and year-quarter
Price	The price at that location and year-quarter
Advertising	The advertising expenditures for that location and year-quarter
CompetitorPrice	Average competitor's price for that location and year-quarter
Income	Average household income for that location and year-quarter
Population	Population for that location and year-quarter
Time	Linear time trend variable (1 is for 2006-1, 2 is for 2006-2, ..., 8 is for 2007-4)

a Estimate the following linear regression model (this model is the one in the textbook):

$$\begin{aligned} Sales &= \beta_0 + \beta_1 Price + \beta_2 Advertising + \beta_3 CompetitorPrice + \beta_4 Income + \\ &\quad \beta_5 Population + \beta_6 Time \end{aligned}$$

Report the coefficient estimates and the standard errors for each variable.

Answer:

The coefficient estimates and standard errors for each variable are at the end of this file under Regression output for problem 1, 1. a.

b Use a two-tailed test for each individual regression coefficient using the null hypothesis that $\beta_i = 0$. Report the t-statistics and your findings on significance level. What is the interpretation of each estimated coefficient?

Answer:

Looking at the regression output for part a (you may want to print out the regression output so that you can look at it while reading these answers – I have mine open on a dual screen monitor setup while typing the answers), the estimated coefficients for price, advertising, and population are all significant at the 1% level. Competitor's price is significant at the 5% level, while the intercept (constant) is significant at the 10% level. Income and the linear time trend are not significant.

For price, a \$1 increase in price would lead to a decrease of $122,606$ in quantity sold, all other variables in the model held constant.

For advertising, a \$1 increase in advertising would lead to an increase of 5.84 in quantity sold, all other variables in the model held constant.

For competitor's price, a \$1 increase in price would lead to an increase of $29,867$ in quantity sold, all other variables in the model held constant.

For population, a 1 person increase in population would lead to an increase of 0.03 in quantity sold, all other variables in the model held constant.

The constant suggests that if all of the independent variables are zero, then $529,774$ units would be sold (but this result should not be taken too seriously because there are not many independent variables that are close to 0).

The coefficients on income and time are positive, suggesting higher incomes lead to more quantity sold and over time people are buying more, but I would not be too specific about interpreting these variables because they are not statistically different than zero.

Note that the magnitude of some of those changes is quite difference – a one unit change in population only leads to a 0.03 increase in quantity sold, whereas a $1 increase in competitor's price leads to an increase of 29,867 in quantity sold. However, these variables are on much different scales – a $1 price change is a large percentage change in price (the average competitor's price is $6.09) whereas a 1 person increase in population is basically meaningless (the average population is over 7 million people).

c Using an F-test, test for the significance of the regression using the null hypothesis that $\beta_1 = \beta_2 = ... = \beta_6 = 0$. What is the critical value of the F-distribution for the test you are conducting? (**Note**: I believe the reported F-value in the text is incorrect – it is reported as 45.16, but every program I use has either 46.15 or 46.16 depending on whether it is rounding or truncating after the second decimal place).

Answer:

Again, one nice thing about Stata is that many standard test statistics are already reported. The F-value is 46.16 (we can find this even if we only have the R^2, the number of restrictions, and the number of degrees of freedom in the model):

$$F = \frac{R^2 (n-k)}{(1-R^2)(k-1)} = \frac{0.8710 * 41}{0.1290 * 6} = 46.14$$

Note that this is slightly different from the value in the output (46.16) but that is due to rounding. The critical value for the $F_{6,41}$ at the 1% level is 3.29 (it is 2.34 and 1.93 for the 5% and 10% levels, respectively – I would check the 1% level first). Because $46.16 > 3.29$ we can reject the null hypothesis and conclude that at least one β_i is different from zero. Note that there is a line in the Stata output, "Prob > F = 0.0000," that tells us the exact significance level for the F-test (the same is true for the individual t-tests, under the column labeled $P > |t|$).[1]

d Rather than using a linear time trend variable, another method of introducing the time periods into the model is to use dummy variables for each time period. Create one dummy variable for each of the 8 time periods.

- Estimate the same model as in part **a**, but instead of $\beta_6 Time$ include ALL 8 dummy variables. Note that when you attempt to estimate this model you should either get an error message that states that you cannot estimate this model or that the software has chosen one of the dummy variables (or possibly the intercept) to exclude so that it can estimate the model. The point of this particular exercise is so that you will know what will happen if you fall into the "dummy variable trap" and include all of the dummy variables you have created for a particular qualitative or categorical variable.

Answer:

Looking at the regression output for 1. d. - all dummy variables included you will notice (1) a comment at the top that reads "note: FstQtr2006 omitted because of collinearity" and (2) no estimated coefficient for FstQtr2006 but instead the word "(omitted)". In this case, when all the dummy variables are included, Stata cannot calculate the estimates because it cannot invert the data matrix because it is not full column rank (there are linearly dependent columns). So Stata chooses one variable and omits it - in this case it happened to be FstQtr2006.

[1] When the significance values are listed as 0.0000 it is not really the exact significance level – in these cases it means "<0.0001" which people generally interpret as "highly significant."

- Estimate the same model as in part **a**, but instead of $\beta_6 Time$ include dummy variables for all year-quarter combinations except 2006-1. What test would you use to determine if the dummy variables are (individually) stastically significant? How would you interpret the estimated coefficients for any statistically significant dummy variables? How would you test if the entire group of 7 dummy variables that you included in this model have a significant impact (jointly) on the regression model?

Answer:

First, compare the output of this regression model with the one in 1. d. - all dummy variables included. Notice that other than having omitted FstQtr2006 all the estimates are identical – because FstQtr2006 was the variable omitted by Stata both models in part d result in the same output (had a different variable been excluded, say FthQtr2007, then the dummy variable coefficients would change but the slopes would be the same).

Second, notice that all the t-statistics for the dummy variables are less than 1.5, which means they are not statistically significant. What this means is that there is no difference between time periods. Had there been a difference between time periods, when interpreting the results we would say something such as "Compared to the first quarter 2006, sales in the second quarter 2007 were 25,155 higher."

e As mentioned in class, the income and population variables appear to be linear interpolations between U.S. Census dates (which occur every 10 years). Suppose that you wanted to control for "location" because you believe that there are important differences between cities/regions. Could dummy variables be used to control for location? Explain.

Answer:

We could use dummy variables rather than the income and population variables. If we believe there are differences in the cities, and we do not have good data on the variables in which we are interested, then a simple way to control for the city is to use dummy variables. I have estimated a model that excludes income, population, and the time trend but includes dummy variables for cities (the excluded city is Washington D.C.). The output is under 1. e. Notice that all of these dummy variables are significant at the 10% level (some at the 5% and 1% levels), which means that there is a significant difference between the Washington D.C. area and the other areas. For instance, the estimated coefficient of $-95,941.81$ for Atlanta means that about 96,000 less units sold in Atlanta than Washington D.C. for some reason. While we do not know the exact reason, we do know that when we control for price, advertising expenditures, and competitor's price there is some difference between Atlanta and Washington D.C. The other estimated coefficients for the dummy variables have a similar meaning – the key is to remember they are all compared to the excluded category, which is Washington D.C.

2. Consider the housing data in the Excel file "SassHousingData." The data is in the first Excel sheet and the variable definitions are in the second sheet. I realize many of these questions are the same as above, but they are the standard questions that are asked.

a Estimate the following model:

$$\begin{aligned} price &= \beta_0 + \beta_1 livingarea + \beta_2 lotarea + \beta_3 age + \beta_4 swimmingpool + \\ &\quad \beta_5 fireplace + \beta_6 spa \end{aligned}$$

Report the coefficient estimates and the standard errors for each variable.

Answer:

The estimates are in the Regression output for problem 2 at the end of the file.

b Use a two-tailed test for each individual regression coefficient using the null hypothesis that $\beta_i = 0$. Report the t-statistics and your findings on significance level. What is the interpretation of each estimated coefficient?

Answer:

For this model, the t-statistics for livingarea, lotarea, age, and swimmingpool are all greater than 2.57 (there are 980 observations, and 973 degrees of freedom, so we can use the shortcut critical values for the t-distribution) so they are all significant at the 1% level. The t-statistics for spa and the constant are all greater (in absolute value) than 1.96 so they are all significant at the 5% level (you can see that the t-statistics for both very close to 2.57 – the column $P > |t|$ shows the exact significance levels, which are 1.2% and 1.3%, respectively). The only variable that is not statistically significant is fireplace.

For livingarea, a 1 sq. foot increase will lead to an \$87.81 increase in sales price, all other variables in the model held constant.

For lotarea, a 1 sq. foot increase will lead to a \$5.13 increase in sales price, all other variables in the model held constant.

For age, a 1 year increase in age will lead to a \$1633.28 increase in sales price, all other variables in the model held constant.

For swimmingpool, having a swimming pool leads to a \$23,994 increase in sales price when compared to not having a swimming pool.

For spa, having a spa leads to a \$37,480.94 increase in sales price when compared to not having a spa.

For the intercept, sales price equals -\$25,568.89 when all other variables equal zero, but again this result is not very useful because our independent variables are typically not close to zero.

The fireplace variable is statistically insignificant.

The result on age seems a little counterintuitive – typically older homes sell for less, but there may be some features of the house (community it is located in, the type of house structure, etc.) that the age variable is capturing that we are unaware of (it may be that older homes were built using better materials than newer homes, therefore they sell for more).

c Using an F-test, test for the significance of the regression using the null hypothesis that $\beta_1 = \beta_2 = ... = \beta_6 = 0$. What is the critical value of the F-distribution for the test you are conducting?

Answer:

Again, Stata give the F-value, which is 115.52. Calculating it using R^2 we have:

$$F = \frac{0.416 * (973)}{0.584 * (6)} = 115.52$$

The critical value for the $F_{6,973}$ is 2.80, and because $115.52 > 2.80$ we can reject the null and conclude that at least one β_i is different than zero. As a rule, once you get past the $F_{6,6}$ critical value as long as your F-statistic is greater than 10 you will be able to reject the null hypothesis for this test.

d How would you specify your model if you wanted to see if having a swimming pool affected the slope of the line with respect to lotarea? Estimate this model.

Answer:

To see if swimming pool affected the slope of lotarea we would need to create an interaction term using swimmingpool and lotarea. To create this interaction term create a new column (I called mine LotSwimInt for "Lotarea-Swimmingpool-Interaction") and, for each observation, multiply lotarea by swimmingpool. The output for this model is in 2. c. Notice that the estimated coefficient for the interaction term is insignificant (t-statistic of 0.68) and that now the estimated coefficient is also insignificant (t-statistic of 1.54). This outcome happens sometimes – when adding a new variable, it may be correlated with an old variable and "soak up" (or "use up" or "take away") some of the old variable's explanatory power.

e At times we find that an independent variable has a positive impact on the dependent variable, but that this impact increases at a decreasing rate (for instance, the 900^{th} square foot of a home may have a positive impact on sales price, but this marginal impact may be lower than the 800^{th} square foot of a house). To capture this possibility we can introduce squared terms in the model. Estimate the following model:

$$\begin{aligned} price &= \beta_0 + \beta_1 livingarea + \beta_2 lotarea + \beta_3 age + \beta_4 livingarea^2 + \\ &\quad \beta_5 lotarea^2 + \beta_6 age^2 \end{aligned}$$

What is the effect of livingarea, lotarea, and age on the sales price of the house?

Answer:

This model is in part 2. e. I'll pick age and think about how it enters the model – I will take the partial derivative with respect to age:

$$\frac{\partial price}{\partial age} = \beta_3 + 2\beta_6 age$$

A similar partial derivative can be found for livingarea and lotarea. Notice that in this model the effect of age on sales price now depends on the age level itself, so an age of 1 and an age of 40 will have different impacts.

In the estimated model, only lotarea and lotarea2 both have significant estimates so I will work with that variable. The coefficient for lotarea is 9.395189 and the coefficient for lotarea2 is -0.0000969. Lotarea has a minimum of 1296, a mean of 7411, and a maximum of 50,529. The table below shows the impact of each of these different values of lotarea on sales price:

Lot area	$\Delta SalesPrice$
1296	9.27
7411	8.68
50529	4.50

We can see that as the lot size increases the impact on sales price is still positive but begins decreasing. Moving from 1295 to 1296 square feet is more valuable than moving from 50528 to 50529 square feet.

f You should have estimated 3 models for question 2 – one in part **a**, one in part **e**, and one in part **d**. Now estimate a basic model:

$$price = \beta_0 + \beta_1 livingarea + \beta_2 lotarea + \beta_3 age$$

Each of the other models has a different set of independent variables that has been excluded. For each of those models, conduct an F-test to see if the excluded variables are jointly insignificant.

Answer:

Keep in mind that the F-value for this type of test is calculated as:

$$\frac{\left(R^2_{UR} - R^2_R\right)/q}{\left(1 - R^2_{UR}\right)/\left(n-k\right)} \,\tilde{}\, F_{q,n-k}$$

The results of this estimation are reported in 2. f. The main item we need from this estimation is the R^2, which is equal to 0.4028. I am going to use $R^2_{model\ f}$ to represent the R^2 for different models (so that would be for model f). Keep in mind that $R^2_f = 0.4028$ and this model is the RESTRICTED model.

In model 2. a. we had 6 indepedent variables (not including the intercept) and in model 2. f. we had 3 (not including the intercept). So there are 3 restrictions in this model (we have imposed that the coefficient estimates of these three excluded variables are all equal to zero). For model a, we have $R_a^2 = 0.4160$. The $n-k$ comes from the UNrestricted model, which is model a. The F-value is:

$$\begin{aligned} F &= \frac{(.416 - .4028)/3}{(1-.416)/(973)} \\ F &= \frac{.0132 * 973}{.584 * 3} \\ F &= 7.33 \end{aligned}$$

The critical value for the $F_{3,973}$ at the 1% level is 3.78. So we can conclude that at least one of the estimated coefficients for swimmingpool, fireplace, or spa is statistically different than zero.

In model 2. c. we had 7 independent variables (not including the intercept) so there are 4 restrictions. The $R_c^2 = .4163$. Our F-value is (note that $n-k$ in this model is 972, reflecting the additional independent variable):

$$\begin{aligned} F &= \frac{(.4163 - .4028)/4}{(1-.4163)/972} \\ F &= \frac{.0135 * 972}{.5837 * 4} \\ F &= 5.62 \end{aligned}$$

The critical value for the $F_{4,972}$ at the 1% level is 3.32. Because $5.62 > 3.32$ we can reject the null and conclude that at least one of the excluded variables has an estimated coefficient that is significantly different than zero.

Finally we look at model 2. e. Model 2. e. was the model with the squared terms. There were 6 independent variables (not including the intercept) so there are 3 restrictions. We have $R_e^2 = 0.4100$. Our F-value is:

$$\begin{aligned} F &= \frac{(.41 - .4028)/3}{(1-.41)/973} \\ F &= \frac{.0072 * 973}{.59 * 3} \\ F &= 3.96 \end{aligned}$$

We know from above that the critical value for $F_{3,973}$ at the 1% level is 3.78. Because $3.96 > 3.78$ we can reject the null and conclude that at least one of the excluded squared terms has an estimated coefficient that is statistically different than zero.

All of these results in part f should have been "expected" because in each case we were including one or more independent variables that had estimated coefficients that were individually statistically different than zero. This result is true despite the very small increase in R^2, as the R^2 with just livingarea, lotarea, and age was 0.4028, and no other model increased R^2 beyond 0.4163. These results suggests some drawbacks in looking just at differences in R^2, which is one reason why we use $\overline{R}^2$. All of the output for the regression models reports $\overline{R}^2$ and we can see that the respective $\overline{R}^2$ for the models with more independent variables are greater than the $\overline{R}^2$ for the restricted model. You may want to go to the Mrs. Smyth's Pie question and see if the dummy variables for time are jointly significant.

```
*****************************************Regression output for problem
1******************************************************

1. a.

. regress sales price advertisingexpenditures competitorsprice income population
time

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  6,    41) =   46.16
       Model |  1.2649e+12     6  2.1082e+11           Prob > F      =  0.0000
    Residual |  1.8727e+11    41  4.5676e+09           R-squared     =  0.8710
-------------+------------------------------           Adj R-squared =  0.8522
       Total |  1.4522e+12    47  3.0898e+10           Root MSE      =   67584

------------------------------------------------------------------------------------
-----
                  sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf.
Interval]
------------------------+-----------------------------------------------------------
-----
                  price |  -122606.8   16422.38    -7.47   0.000    -155772.5
-89441.16
advertisingexpenditures |   5.837648   1.650494     3.54   0.001     2.504408
9.170887
       competitorsprice |   29866.59   13449.22     2.22   0.032     2705.344
57027.83
                 income |   2.042729   3.762305     0.54   0.590      -5.5554
9.640858
             population |    .030258   .0039448     7.67   0.000     .0222913
.0382247
                   time |   2815.493   4539.242     0.62   0.539    -6351.694
11982.68
                  _cons |   529773.7   271330.9     1.95   0.058    -18190.08
1077738
------------------------------------------------------------------------------------
-----

1. d. - all dummy variables included

. regress sales price advertisingexpenditures competitorsprice income population
FstQtr2006 SecQtr
> 2006 ThdQtr2006 FthQtr2006 FstQtr2007 SecQtr2007 ThdQtr2007 FthQtr2007
note: FstQtr2006 omitted because of collinearity

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F( 12,    35) =   21.41
       Model |  1.2781e+12    12  1.0651e+11           Prob > F      =  0.0000
    Residual |  1.7408e+11    35  4.9737e+09           R-squared     =  0.8801
-------------+------------------------------           Adj R-squared =  0.8390
       Total |  1.4522e+12    47  3.0898e+10           Root MSE      =   70524

------------------------------------------------------------------------------------
-----
                  sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf.
Interval]
------------------------+-----------------------------------------------------------
-----
                  price |  -117647.2   17582.39    -6.69   0.000    -153341.4
-81953.05
advertisingexpenditures |   5.747695   1.735591     3.31   0.002     2.224257
9.271133
       competitorsprice |   32732.68   14917.46     2.19   0.035     2448.631
```

```
63016.73
               income |   2.318146   3.962822     0.58   0.562    -5.726811
10.3631
           population |   .0302853   .0041292     7.33   0.000     .0219025
.038668
          FstQtr2006 |          0  (omitted)
          SecQtr2006 |     200.33   41393.98     0.00   0.996    -83833.91
84234.57
          ThdQtr2006 |   38834.52   41660.74     0.93   0.358    -45741.28
123410.3
          FthQtr2006 |   10388.41   42586.65     0.24   0.809    -76067.09
96843.92
          FstQtr2007 |   53294.51   41491.83     1.28   0.207    -30938.39
137527.4
          SecQtr2007 |    25155.2   43337.54     0.58   0.565    -62824.69
113135.1
          ThdQtr2007 |   6857.452   42760.14     0.16   0.874    -79950.24
93665.14
          FthQtr2007 |   26626.84   42142.61     0.63   0.532     -58927.2
112180.9
                _cons |   459198.9   294219.1     1.56   0.128    -138097.5
1056495
------------------------------------------------------------------------------------
-----

.

1. d. - exclude First-Qtr 2006

. regress sales price advertisingexpenditures competitorsprice income population 
SecQtr2006 ThdQtr
> 2006 FthQtr2006 FstQtr2007 SecQtr2007 ThdQtr2007 FthQtr2007

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F( 12,    35) =   21.41
       Model |  1.2781e+12    12  1.0651e+11           Prob > F      =  0.0000
    Residual |  1.7408e+11    35  4.9737e+09           R-squared     =  0.8801
-------------+------------------------------           Adj R-squared =  0.8390
       Total |  1.4522e+12    47  3.0898e+10           Root MSE      =   70524

------------------------------------------------------------------------------------
-----
                sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf. 
Interval]
------------------------+-----------------------------------------------------------
-----
                price |  -117647.2   17582.39    -6.69   0.000    -153341.4
-81953.05
advertisingexpenditures |   5.747695   1.735591     3.31   0.002     2.224257
9.271133
     competitorsprice |   32732.68   14917.46     2.19   0.035     2448.631
63016.73
               income |   2.318146   3.962822     0.58   0.562    -5.726811
10.3631
           population |   .0302853   .0041292     7.33   0.000     .0219025
.038668
          SecQtr2006 |     200.33   41393.98     0.00   0.996    -83833.91
84234.57
          ThdQtr2006 |   38834.52   41660.74     0.93   0.358    -45741.28
123410.3
          FthQtr2006 |   10388.41   42586.65     0.24   0.809    -76067.09
96843.92
          FstQtr2007 |   53294.51   41491.83     1.28   0.207    -30938.39
```

```
137527.4
            SecQtr2007 |    25155.2   43337.54     0.58   0.565    -62824.69
113135.1
            ThdQtr2007 |   6857.452   42760.14     0.16   0.874    -79950.24
93665.14
            FthQtr2007 |   26626.84   42142.61     0.63   0.532     -58927.2
112180.9
                 _cons |   459198.9   294219.1     1.56   0.128    -138097.5
1056495
------------------------------------------------------------------------------------
-----

.


1. e.

. regress sales price advertisingexpenditures competitorsprice Atlanta Chicago
Dallas LA Minneapol
> is

      Source |       SS       df       MS              Number of obs =      48
-------------+------------------------------           F(  8,    39) =   41.26
       Model |  1.2987e+12     8  1.6234e+11           Prob > F      =  0.0000
    Residual |  1.5345e+11    39  3.9347e+09           R-squared     =  0.8943
-------------+------------------------------           Adj R-squared =  0.8727
       Total |  1.4522e+12    47  3.0898e+10           Root MSE      =   62727

------------------------------------------------------------------------------------
-----
                 sales |      Coef.   Std. Err.      t    P>|t|     [95% Conf.
Interval]
-----------------------+------------------------------------------------------------
-----
                 price |  -112972.3   16211.89    -6.97   0.000    -145763.9
-80180.61
advertisingexpenditures |   4.779049   1.636488     2.92   0.006      1.46894
8.089158
      competitorsprice |    29883.8   12610.32     2.37   0.023     4377.018
55390.58
               Atlanta |  -95941.81    34989.1    -2.74   0.009    -166713.9
-25169.68
               Chicago |   54943.86   31921.07     1.72   0.093      -9622.6
119510.3
                Dallas |  -72195.29   34229.92    -2.11   0.041    -141431.8
-2958.753
                    LA |   218670.4   46955.06     4.66   0.000     123694.8
313645.9
           Minneapolis |  -95901.05   39234.78    -2.44   0.019    -175260.9
-16541.22
                 _cons |   829700.1   127341.4     6.52   0.000     572127.8
1087272
------------------------------------------------------------------------------------
-----

.


************************************************Regression output for problem
2************************************************

2. a.
```

```
. regress price livingarea lotarea age swimmingpool fireplace spa

      Source |       SS       df       MS              Number of obs =     980
-------------+------------------------------           F(  6,   973) =  115.52
       Model |  2.8406e+12     6  4.7343e+11           Prob > F      =  0.0000
    Residual |  3.9875e+12   973  4.0982e+09           R-squared     =  0.4160
-------------+------------------------------           Adj R-squared =  0.4124
       Total |  6.8281e+12   979  6.9746e+09           Root MSE      =   64017

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  livingarea |   87.81351   5.480863    16.02   0.000     77.05784    98.56919
     lotarea |   5.128513   .5785581     8.86   0.000     3.993148    6.263878
         age |   1633.282   174.4171     9.36   0.000     1291.005    1975.559
swimmingpool |      23994   6257.319     3.83   0.000      11714.6    36273.39
   fireplace |  -6314.106   5730.753    -1.10   0.271    -17560.17    4931.952
         spa |   37480.94   14887.58     2.52   0.012      8265.49     66696.4
       _cons |  -25568.89   10327.93    -2.48   0.013    -45836.46   -5301.307
------------------------------------------------------------------------------
```

2. c.

```
. regress price livingarea lotarea age swimmingpool fireplace spa lotswimint

      Source |       SS       df       MS              Number of obs =     980
-------------+------------------------------           F(  7,   972) =   99.03
       Model |  2.8425e+12     7  4.0607e+11           Prob > F      =  0.0000
    Residual |  3.9856e+12   972  4.1004e+09           R-squared     =  0.4163
-------------+------------------------------           Adj R-squared =  0.4121
       Total |  6.8281e+12   979  6.9746e+09           Root MSE      =   64035

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  livingarea |    88.6785   5.628357    15.76   0.000     77.63337    99.72363
     lotarea |   4.739783   .8139303     5.82   0.000      3.14252    6.337046
         age |   1639.126   174.6774     9.38   0.000     1296.338    1981.914
swimmingpool |   17519.39   11403.87     1.54   0.125    -4859.655    39898.43
   fireplace |  -6281.543    5732.54    -1.10   0.273    -17531.12    4968.038
         spa |   37303.83   14893.98     2.50   0.012     8075.765    66531.89
  lotswimint |   .7441167   1.095578     0.68   0.497    -1.405855    2.894088
       _cons |  -24292.53   10500.31    -2.31   0.021    -44898.43   -3686.639
------------------------------------------------------------------------------
```

2. e.

```
. regress price livingarea lotarea age livingarea2 lotarea2 age2

      Source |       SS       df       MS              Number of obs =     980
-------------+------------------------------           F(  6,   973) =  112.69
       Model |  2.7995e+12     6  4.6658e+11           Prob > F      =  0.0000
    Residual |  4.0287e+12   973  4.1405e+09           R-squared     =  0.4100
-------------+------------------------------           Adj R-squared =  0.4064
       Total |  6.8281e+12   979  6.9746e+09           Root MSE      =   64346

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  livingarea |   112.3081   23.63763     4.75   0.000     65.92146    158.6947
     lotarea |   9.395189   1.629497     5.77   0.000     6.197456    12.59292
```

age	2301.548	516.5536	4.46	0.000	1287.861	3315.236
livingarea2	-.007027	.0065974	-1.07	0.287	-.0199739	.0059198
lotarea2	-.0000969	.0000371	-2.61	0.009	-.0001697	-.000024
age2	-11.48054	7.452112	-1.54	0.124	-26.1046	3.14352
_cons	-78967.69	22738.96	-3.47	0.001	-123590.7	-34344.64

2. f.

```
. regress price livingarea lotarea age

      Source |       SS       df       MS              Number of obs =     980
-------------+------------------------------           F(  3,   976) =  219.45
       Model |  2.7505e+12     3  9.1683e+11           Prob > F      =  0.0000
    Residual |  4.0776e+12   976  4.1779e+09           R-squared     =  0.4028
-------------+------------------------------           Adj R-squared =  0.4010
       Total |  6.8281e+12   979  6.9746e+09           Root MSE      =   64637
```

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
livingarea	91.89326	5.333044	17.23	0.000	81.42771	102.3588
lotarea	5.410206	.5786326	9.35	0.000	4.274699	6.545713
age	1610.845	175.9701	9.15	0.000	1265.522	1956.168
_cons	-34761.97	9707.871	-3.58	0.000	-53812.67	-15711.27