Solutions to F-Test and Prediction In-Class Exercise

1. **Do each of the models presented have significant explanatory power? How do you know?**

   Yes. The F-statistic for both models has a p-value of less than .05 (see the ANOVA table for each model), so we can reject the null hypothesis that all the coefficients are really 0. At least one is not 0, and the model is better than the average at explaining sale price (this is a shortcut to F-Test#1).

   We can also work this out by formally performing an F-test #1:

   Null hypothesis: All the coefficients are equal to zero.
   Alternative Hypothesis: At least one of the coefficients not equal to zero

   Decision rule: If the calculated F-statistic is larger than the critical value of F then we reject the null hypothesis.

   For model 1: $F \geq F_c$ ($F_c = F_{(2,22)} = 3.44$ at the 5% level from the table) so we reject the null hypothesis. We get the F from the ANOVA table (F=296.94). 2 is the number of regressors and numerator degrees of freedom. 22 is n-k-1, the degrees of freedom for the denominator. Thus, All the coefficients are not zero, the equation provides better information than just the mean monthly rent, and the R^2 is significant.

   For model 2: $F \geq F_c$ ($F_c = F_{(5,19)} = 2.74$ at the 5% level) so we reject the null hypothesis. F in this example is 190.84 from the ANOVA table. 5 is the number of regressors and the numerator degrees of freedom. 19 is n-k-1, the degrees of freedom for the denominator. All the coefficients are not zero, the equation provides better information than just the mean monthly rent, and the R^2 is significant.

   I'll calculate the F statistic for model 1. Note that the F-statistics appear on the regression output, as does the p-value associated with seeing an F-statistic (Significance F) of that size if the null hypothesis were true.

   For Model 1:
   From regression output:
   R^2=.98(from MODEL SUMMARY)
   ESS=1031823493278.99 (from ANOVA-Regression)
   RSS=42046454035 (from ANOVA-Residual)
   k=2 (number of regressors)
   n=25 (from ANOVA: total df+1)

   F calculated two ways:

$$F = \frac{\frac{\text{ESS}}{k}}{\frac{\text{RSS}}{n-k-1}} = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}}$$

$$F = \frac{\frac{\text{ESS}}{k}}{\frac{\text{RSS}}{n-k-1}} = \frac{\frac{1031823493278.99}{2}}{\frac{42026454035}{25-2-1}} = \frac{515911746639.50}{1911202456} = 269.94$$

$$F = \frac{\frac{R^2}{k}}{\frac{1-R^2}{n-k-1}} = \frac{\frac{.98}{2}}{\frac{1-.98}{25-2-1}} = \frac{0.48042293}{0.001779734} = 269.94$$

2. **Does the number of parking areas in combination with the age of the structure and lot size add to the explanatory power of the model?** (Compare model 1 with model 2) (HINT: remember that on Excel printout, the ESS and RSS are in the ANOVA table. The ESS (explained sum of squares) is called "regression" and RSS (residual sum of squares) is called residual.)

Hypothesis: (are a subset of coefficients really equal to zero?)

Restricted equation is Model 1	H_0: The coefficients for lot size, age, and number of on-site parking spaces are all equal to zero.
Unrestricted equation is Model 2	H_a: The coefficients for lot size, age, and number of on-site parking spaces are not zero.

Find test (F) statistic:

Pick one way. I'll demonstrate both here.

$$F_{(M,n-k-1)} = \frac{\frac{\text{RSS}_R - \text{RSS}_{UR}}{M}}{\frac{\text{RSS}_{UR}}{n-k-1}} \qquad F_{(M,n-k-1)} = \frac{\frac{R^2_{UR} - R^2_R}{M}}{\frac{1-R^2_{UR}}{n-k-1}}$$

M=3 (the number of additional regressors)

k=5 (number of regressors in unrestricted model)

UR is the model with all the regressors (model 2).

R is the model with 3 of the regressors omitted (model 1).

$$F_{(M,n-k-1)} = \frac{\frac{\text{RSS}_R - \text{RSS}_{UR}}{M}}{\frac{\text{RSS}_{UR}}{n-k-1}} = \frac{\frac{42046454035-20965196398}{3}}{\frac{20965196398}{25-5-1}} = \frac{\frac{21081257637}{3}}{\frac{20965196398}{19}} = \frac{7027085879}{1103431389} = 6.37$$

$$F_{(M,n-k-1)} = \frac{\frac{R^2_{UR} - R^2_R}{M}}{\frac{1-R^2_{UR}}{n-k-1}} = \frac{\frac{.98-.96}{3}}{\frac{1-..98}{25-5-1}} = \frac{\frac{.02}{3}}{\frac{.02}{19}} = \frac{.0067}{.00105} = 6.38$$

The results are slightly different, likely due to rounding.

Compare to Critical Value
Critical $F_{(3,19)}$=3.13 at a 5% level (from table); So, $F>F_c$ so we can reject the null hypothesis (in this case the null equation). The coefficients for lot size, age, and number of parking spaces are not zero—together they add to the explanatory power of the model. The increase in the R-square is real. A model with fewer explanatory variables would not be just as good.

3. **Given the print out you've seen so far and the tests that you've conducted, what variables do you think should remain in the model? Is there other information you'd like before you settle on a model?**

   This is a judgement call. You might say that lot size, age, and number of parking spaces should remain all remain, because of the test above. Alternatively, you might think that the non-significant parameters should be dropped. Yet, theory might dictate that all are important to modeling price. You need to think carefully and make an argument for keep or dropping variables.

4. **Based on Model 2, tell your boss the predicted sale values for the two buildings. Here are the characteristics:**

   Use the estimated parameters for model 2 and substitute the values associated with each factor to come up with a predicted sales price for these two buildings.

Characteristic	Building 1	Building 2
No. of Apartments	14	30
Age of Structure	85	15
Lot Size	11,745	21,000
No. of On-Site Parking Spaces	0	30
Gross Building Area	23,704	30,000
Predicted values	$457,989	$771,603

Model 2:
$$\text{Sales Price} = \hat{\beta}_0 + \hat{\beta}_1(\text{No. of Apartments}) + \hat{\beta}_2(\text{Gross Blg Area}) + \hat{\beta}_3(\text{Lot Size}) + \hat{\beta}_4(\text{Age of Structure}) + \hat{\beta}_5(\text{No. On-site Parking Spaces})$$

Building 1
=92787.86855+4140.419335(14)+15.54412915(23,704)+0.961907855(11,745)
-853.184365(85)+2605.868571(0)
=$457,989

Building 2
=92787.86855+4140.419335(30)+15.54412915(30,000)+0.961907855(21,000)
-853.184365(15)+2605.868571(30)
=$771,603

How do you interpret the coefficients in this model? Are they in the directions you'd expect them to be?