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.

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.

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?

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).

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.
- 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?

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.

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.

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?

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?

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.

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?

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.