Exercise Session 3

The data file *collegetown* contains observations on 500 single-family houses sold in Baton Rouge, Louisiana, during 2009–2013. The data include sale price (in thousands of dollars), *PRICE*, and total interior area of the house in hundreds of square feet, *SQFT*.

1. Plot house price against house size in a scatter diagram

**gnuplot price sqft --output=display**

1. Estimate the linear regression model *PRICE* = β1 + β2*SQFT* + *e.* Interpret the estimates. Draw a sketch of the fitted line.



**If the size of the house increases by one unit, price increases by 13.4 thousand dollars**

1. Estimate the quadratic regression model *PRICE* = α1 + α2*SQFT*2 + *e*. Compute the marginal effect of an additional 100 square feet of living area in a home with 2000 square feet of living space.

**genr sqft2=sqft^2**

**ols price const sqft2**

**Take a derivative wrt sqft in *PRICE* = α1 + α2*SQFT*2 + *e:***

$$\frac{∂PRICE}{∂SQFT}=2\*α\_{2}\*sqft$$

**If sqft=2000 then**

$$\frac{∂PRICE}{∂SQFT}=2\*α\_{2}\*2000=4000\*0.18=720$$

**If you increase the size of the house, price will increase by 720 thousand dollars**

1. For the regressions in (b) and (c), compute the least squares residuals and plot them against *SQFT*. Do any of our assumptions appear violated?



**Assumptions don’t seem violated that error terms should not be correlated with the explanatory variable**

1. One basis for choosing between these two specifications is how well the data are fit by the model. Compare the sum of squared residuals (*SSR*) from the models in (b) and (c). Which model has a lower *SSR*? How does having a lower *SSR* indicate a “better-fitting” model?

**The second model has lower SSR. Lower SSR means that there is less variation unexplained in the model. SSR is tightly related with the goodness of fit measure in fact R2= 1-SSR/SST, therefore, larger SSR will deliver worse goodness of fit.**

Solutions are also available as script file *collegetown*