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*.

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

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

b. Estimate the linear regression model $PRICE = \beta_1 + \beta_2 SQFT + e$. Interpret the estimates. Draw a sketch of the fitted line.

| | coefficient | std. error | t-ratio | p-value | |
|---|---|---|---|---|---|
| const | -115.424 | 13.0882 | -8.819 | 1.95e-017 | *** |
| sqft | 13.4029 | 0.449164 | 29.84 | 5.92e-113 | *** |

| | | | |
|---|---|---|---|
| Mean dependent var | 250.2369 | S.D. dependent var | 171.4765 |
| Sum squared resid | 5262847 | S.E. of regression | 102.8006 |
| R-squared | 0.641317 | Adjusted R-squared | 0.640596 |
| F(1, 498) | 890.4114 | P-value(F) | 5.9e-113 |
| Log-likelihood | -3024.863 | Akaike criterion | 6053.726 |
| Schwarz criterion | 6062.155 | Hannan-Quinn | 6057.033 |

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

c. Estimate the quadratic regression model $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$. 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 $PRICE = \alpha_1 + \alpha_2 SQFT^2 + e$:**

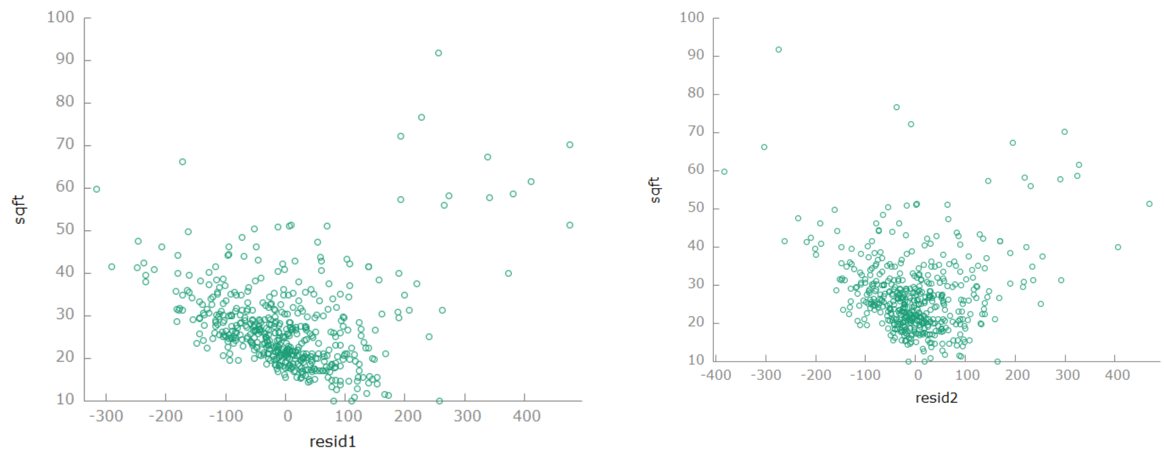
$$\frac{\partial PRICE}{\partial SQFT} = 2 * \alpha_2 * sqft$$

**If sqft=2000 then**

$$\frac{\partial PRICE}{\partial SQFT} = 2 * \alpha_2 * 2000 = 4000 * 0.18 = 720$$

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

d. 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**

e. 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 $R^2$= 1-SSR/SST, therefore, larger SSR will deliver worse goodness of fit.**

Solutions are also available as script file *collegetown*