Homework Assignment 4

Panagiotis (Panos) Toulis– Chicago Booth
BUS41100 Applied Regression Analysis

1 Residual Surrealism

Load the data in `boo.csv` and run MLR of `y` on the remaining variables `x1, x2, x3, x4, x5, x6`.

(a) Using `summary` of the model which variables appear significant?

(b) Test the normality assumption through a histogram of the studentized residuals.

(c) Generate the residual plot, i.e., plot the fitted values on the x-axis and the residuals on the y-axis (use `pch=20` and `cex=0.6` for best results). Comment on the plot.

2 Infant Nutrition

Revisit the infant nutrition data from homework 3. In that homework, we fit three different models:

$$\mathbb{E}[\texttt{woh}|\texttt{age}] = \beta_0 + \beta_1 \texttt{age}$$
$$\mathbb{E}[\texttt{woh}|\texttt{age}] = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{age}^2$$
$$\mathbb{E}[\texttt{woh}|\texttt{age}] = \beta_0 + \beta_1 \texttt{age} + \beta_2 \mathbb{1}\{\texttt{age} > 7\} + \beta_3 \texttt{age} \times \mathbb{1}\{\texttt{age} > 7\}$$

Informally using residual plots (i.e. looking at patterns) we decided the third model was the best. Now use our diagnostic tools to re-examine the three models. Specifically, discuss which model has better diagnostic plots, and briefly explain why.

3 Price Elasticity of Cheese

This question considers sales volume as well as price and display activity for packages of "Borden Sliced Cheese". The data, available as `cheese.csv` on the course site, are taken from Rossi, Allenby, and McCulloch's *Bayesian Statistics and Marketing*. For each of 88 stores (`store`) in different US cities, we have repeated observations of the sales volume (`vol`, in terms of packages sold), unit price (`price`), and whether the product was advertised with an in-store display (`disp` = 1 for display).

(a) Ignoring `price`, do the in-store `disp`lays have an effect on log `sales`? Is there reason to suspect that your result is confounded by pricing strategies?

(b) A better question: is `price` elasticity for Borden cheese affected by the presence of in-store advertisement? Test this by running one single regression.

(c) Do you have a possible economic explanation for your results in **(b)**?

4 Market Diagnostics

We will work on model diagnostics for the CAPM model based on `mfunds.csv`.

(a) Run the CAPM model of `windsor` on `valmrkt` as we did in earlier problems. Plot the diagnostics and comment on the plots. Which assumption seems the least plausible?

(b) Does the normality plot indicate that our (studentized) residuals are heavier- or thinner-tailed compared to the standard normal?

(c) Execute the Shapiro-Wilk test. Do you accept or reject the normality hypothesis?

(d) Execute a test for heteroskedasticity. Then, calculate the robust standard error for the beta of the CAPM model. Is this very different from the regular standard error? How does this relate to the result from the heteroskedasticity test?

(e) Run the CAPM model on all other funds, one-by-one. For each model, calculate which datapoint has the highest Cook's distance measure. Is there agreement across assets? What does this tell you about the dependence between assets?

5 International Trade

Load the `trade.csv` data from (Rose and Engel, 2002) and answer the following questions.

(a) How many observations does the dataset have? Does each row (datapoint) correspond to one country or a pair of countries?

(b) Find the country with most trade entries and the country with the highest total trade volume.

(c) Run the regression of `lvalue` (log trade volume) on control variables (as in Week 4 slides) and present the results. Is being in a currency union (`cu=1`) beneficial to trade?

(d) Using the `vif` diagnostic (or other tools) , do you see 'excessive' multicollinearity in the data?

(e) The variables `lrgdp` and `lrgdpcc` are the total GPD and GDP per capita, respectively, of the trade pair. Argue why the regression above is essentially fitting the so-called 'gravity model' of trade.[1]

(f) Run a regression with clustered errors by both countries in the pair (`ccode1,ccode2`) as we did in class. Briefly explain why the standard error of `cu` is now larger. Is the coefficient of `cu` still significant?

(g) In the diagnostic plots of the fitted model from (c) that we saw in class, the leverage plot has a cluster of points with higher leverage than the rest (> 0.04). Try to identify what type of countries are in this cluster, and argue whether this is concerning or not.

[1] `https://en.wikipedia.org/wiki/Gravity_model_of_trade`