Data Analytics and Machine Learning | Prof. Lochstoer

Problem Set 1

Question 1 : On ggplot2 and regression planes

The classic dataset, `diamonds`, (you must load the `ggplot2` package to access this data) has about 50,000 prices of diamonds along with weight (`carat`) and quality of cut (`cut`).

1. Use ggplot2 to visualize the relationship between price and carat and cut. price in the dependent variable. Consider both the log() and sqrt() transformation of price.

2. Run a regression of your preferred specification. Perform residual diagnostics as you learned in 237Q,1. What do you conclude from your regression diagnostic plots of residuals vs. fitted and residuals vs. carat?

note: `cut` is a special type of variable called an ordered factor in R. For ease of interpretation, convert the ordered factor into a "regular" or non-ordinal factor.

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.3.3

data(diamonds)
cutf=as.character(diamonds$cut)
cutf=as.factor(cutf)
```

Use the StockRetAcct_insample.dta Stata dataset available at CCLE (Week 1) for the next two questions.

Question 2 : Nonlinear relations

A common concern is that the relationship between a predictive variable (X) and the outcome we are trying to predict (Y) is nonlinear. On the surface, this seems to invalidate linear regressions, such as the Fama-MacBeth regression. However, this is not generally the case. For instance, if Y = f(X) + noise, where f(.) is not linear in X, simply define a transformation of X as, generally, Z = a + bf(X). Now, it is clear that Y = a1 + b1*Z, for constants a, a1, b, and b1. In other words, one could include squared values of X in the regression, perhaps max(0,X), etc.

We will see this in action for the case of Issuance (lnIssue). This is the average amount of stock issuance in the last 36 months, normalized by market equity. Generally, firms that issue a lot of equity have low returns going forward.

a. Construct decile sorts (10 portfolios) as in the class notes, but now based on the issuance variable lnIssue. Give the average return to each decile portfolio, value-weighting stocks within each portfolio each year, equal-weighting across years.

b. Plot the average return to these 10 portfolios, similar to what we did in the Topic 1(e-f) notes. Discuss whether the pattern seems linear or not.

c. Since most of the 'action' is in the extreme portfolios, consider a model where expected returns to stocks is linear in a transformed issuance-characteristic that takes three values: -1 if the stock's issuance is in Decile 1, 1 if the stock's issuance is in decile 10, and 0 otherwise.

Create this transformed issuance variable and run a Fama-MacBeth regression with it. Report the results. What is the nature of the portfolio implied by the Fama-MacBeth regression? That is, what stocks do you go long, short, no position?

Question 3 : Double-sorts and functional forms

In the lecture notes we saw that the value spread is much larger for small stocks. Using this fact, I proposed a model where expected returns are linear in the book-to-market ratio as well as the interaction between book-to-market and size. In other words, holding size constant there is a linear relation between expected stock returns and book-to-market.

In this question, we will dig deeper into whether this is a reasonable assumption or not based on visual analysis.

a. Create independent quintile sorts based on book-to-market (lnBM) and size (lnME). That is create a quintile variable by year for book-to-market and then create a quintile variable by year for size.

b. For each size quintile, plot the average returns to the five book-to-market quintile portfolios. So, for size quintile 1, and book-to-market quintile 3, the stocks in this portfolio all have size quintile equal to 1 and book-to-market quintile equal to 3. Thus, I'm looking for five plots here, one for each size quintile.

Does the assumption of conditional linearity seem ok, or would you suggest a different model?