Your name: ____

UCLA Anderson School of Management

Final Exam 2017, MGMT 237M.1 – Data Analytics and Machine Learning

Solutions

Prof. Lars A. Lochstoer

1. Consider the following panel regression:

$$e_{i,t} = \mu_t + \delta_1 bm_{i,t-1} + \delta_2 prof_{i,t-1} + \varepsilon_{i,t},$$

where $i = 1, ..., N$ refers to firm and $t = 1, ..., T$ refers to time. The left hand side variable, $e_{i,t}$, is log return on equity. The right hand side variables are log book-to-market and profitability. The time-varying intercept is a *time fixed effect.*

   (a) Assume that the cross-sectional average bm and $prof$ equals zero for each time t.

      i. Give the analytical expresson for the regession estimate of the time fixed effect at each time t. That is, $\hat{\mu}_t = ?$

In matrix notation,

$$\beta_{FE} = (X'MX)^{-1}(X'MY)$$

where

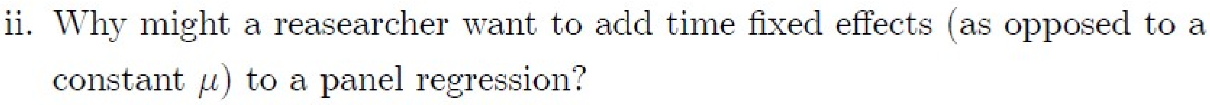
$$\beta_{FE} = \begin{bmatrix} \delta_{1,FE} \\ \delta_{2,FE} \end{bmatrix}$$

$$X = \begin{bmatrix} bm \\ prof \end{bmatrix}$$

Y is the matrix of dependent variables and M is the well-known partialling out matrix. Note that M is idempotent. Once we solve for β_{FE}, we can find $\hat{\mu}_t$ by using the following formula:

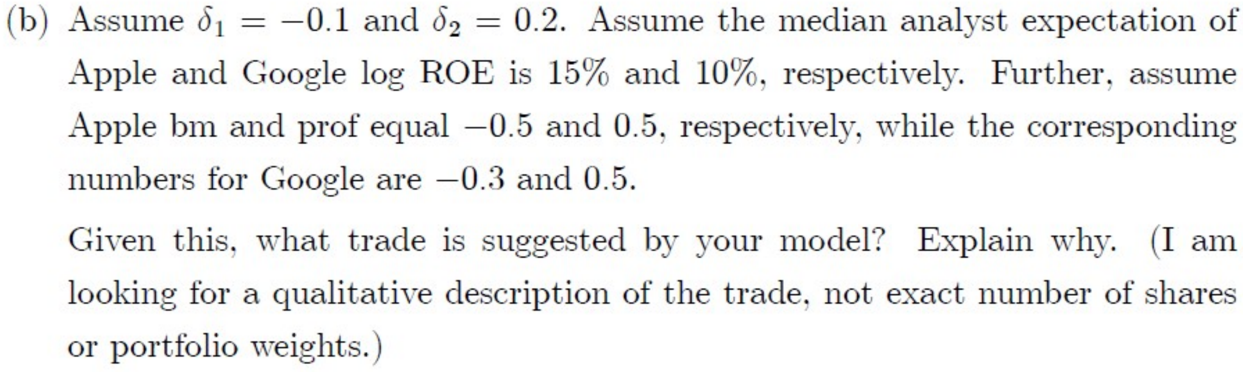
$$\hat{\mu}_t = \bar{y}_t - \bar{x}_t \beta_{FE}$$

where $\bar{y}_t$ and $\bar{x}_t$ are time t cross-sectional averages. Given that $\bar{x}_t = 0$, $\hat{\mu}_t = \bar{y}_t$.



ii. Why might a reasearcher want to add time fixed effects (as opposed to a constant μ) to a panel regression?

Time fixed effects capture the influence of aggregate time series trends. In other words, we use time fixed effects to remove the variation between time periods and focus on the "within" variation (variation within each time period). A useful example to think about is a regression of income on education: nominal wages increase over time (aggregate time series trend) and this increase tells us nothing about the relation between income and education, we are interested in how education influences income within each time period.



(b) Assume $\delta_1 = -0.1$ and $\delta_2 = 0.2$. Assume the median analyst expectation of Apple and Google log ROE is 15% and 10%, respectively. Further, assume Apple bm and prof equal -0.5 and 0.5, respectively, while the corresponding numbers for Google are -0.3 and 0.5.

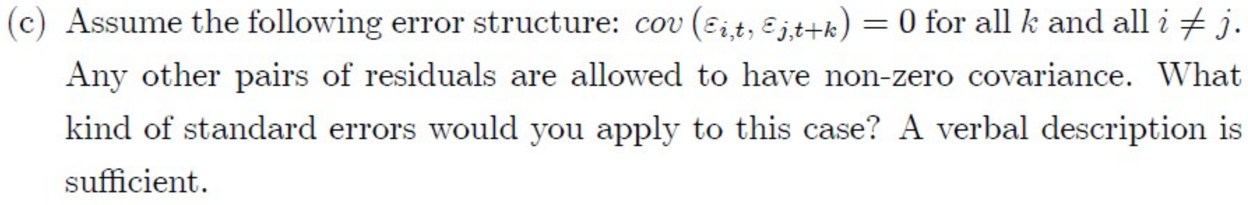
Given this, what trade is suggested by your model? Explain why. (I am looking for a qualitative description of the trade, not exact number of shares or portfolio weights.)

In order to answer this question, we need to demean the independent variables cross-sectionally:

$$E_{t-1}\left[\tilde{e_{i,t}}\right] = \delta_1 \tilde{bm}_{i,t-1} + \delta_2 \tilde{prof}_{i,t-1}$$

Based on that formula, the expected time t $\tilde{ROE}$ of Apple if 1% and the $\tilde{ROE}$ of Google is -1%.

The predicted Apple $R\tilde{O}E$ is 2.5% and the predicted Google $R\tilde{O}E$ is -2.5%. Based on these data points, Apple will fall short of meeting analysts' expectations and Google will exceed analysts' expectations. The trading strategy is long Google and short Apple.



(c) Assume the following error structure: $cov\left(\varepsilon_{i,t}, \varepsilon_{j,t+k}\right) = 0$ for all k and all $i \neq j$. Any other pairs of residuals are allowed to have non-zero covariance. What kind of standard errors would you apply to this case? A verbal description is sufficient.

Given the possibility of both cross-sectional and time series correlation between the residuals being non-zero, it is advisable to cluster standard errors by both time and individual firm.

2. Assume your research associate has deviced a stock-specific trading signal that she believes is a positive predictor of future cross-sectional stock returns. You want to see if the trading signal has marginal value above and beyond what is given by the firm log book-to-market ratios using Fama-MacBeth regressions.

   (a) Give the Fama-MacBeth regressions you will run. Use clear notation. (6 pts)

   Run the cross-sectional regression $R^e_{i,t} = \lambda_{0,t} + \lambda_{1,t} z_{i,t-1} + \varepsilon_{i,t}$ across stocks $i = 1, ..., N$ at each time t, for a total of T cross-sectional regressions.

   (b) Using matrix algebra, give an expression for the portfolio weights at each time t for the portfolio that replicates the estimated Fama-Macbeth coefficient on the trading signal from the previous question. Take care the define any matrices used. (6 pts)

   The portfolio weights are (from the class notes):

$$w_{i,t} = \frac{1}{N} \frac{z_{i,t} - E_N [z_{i,t}]}{var_N (z_{i,t})}$$

   where the subscript N denotes a moment taken across stocks (cross-sectional mean and variance.

   The Fama-MacBeth regression coefficient of interest $(\lambda_{1,t+1})$ is then:

$$\lambda_{1,t+1} = \sum_{i=1}^{N} w_{i,t} R^e_{i,t+1},$$

   which is a portfolio return.

   (c) Explain, using equations, how the t-statistic for the coefficients in the Fama-MacBeth are related to the Sharpe ratios on the coefficient-replicating portfolios (6 pts)

   The test statistic

$$\tau = \frac{\hat{\lambda}_1}{\sqrt{var\left(\hat{\lambda}_1\right)}},$$

where

$$\begin{aligned}\hat{\lambda}_1 &= \sum_{t=1}^{T} \hat{\lambda}_{1,t} \\ var\left(\hat{\lambda}_1\right) &= \frac{1}{T} var\left(\hat{\lambda}_{1,t}\right)\end{aligned}$$

where the variance of $\hat{\lambda}_{1,t}$ is the variance of the time-series of estimated Fama-MacBeth coefficients, $\hat{\lambda}_{1,t}$.

(d) Explain why adding industry dummies might improve the Sharpe ratios of your implied trading strategy. (6 pts)

Here, you simply include firms' book-to-market ratios in each Fama-MacBeth regression:

$$R^e_{i,t} = \lambda_{0,t} + \lambda_{1,t} z_{i,t-1} + \lambda_{2,t} bm_{i,t-1} + \varepsilon_{i,t}$$

We are still interested in $\hat{\lambda}_1$, defined as in the previous sub-question. The regression now effectively asks, what is the Sharpe ratio of a strategy based on the signal, holding firm book-to-market ratios constant. Let's take a simple example where the signal is book-to-market plus a little noise. Now, (ignoring estimation error) we would get $\lambda_1 = 0$, as holding book to market constant the signal would be a sort based only on noise. In the univariate regressions, however, the signal-based sort would be a sort on book-to-market (as well as a little noise) which would likely yield some excess return. In sum, we are estimating the marginal effect of the signal.

3. You want to predict whether the 6% out-of-the-money 1-month S&P500 put option will end up out of the money or not at expiration. Let x_t be a monthly binary variable that is 1 if the option ended up in the money at expiration (end of month t) and zero otherwise. In addition, you have the options' implied volatility, IV_{t-1}, the implied volatility skew, $Skew_{t-1}$, bid-ask, ba_{t-1}, and open interest, oi_{t-1}. These variables corresponding to the option maturing at time t are known at time $t-1$ (end of month $t-1$). You decide to use a logistic regression.

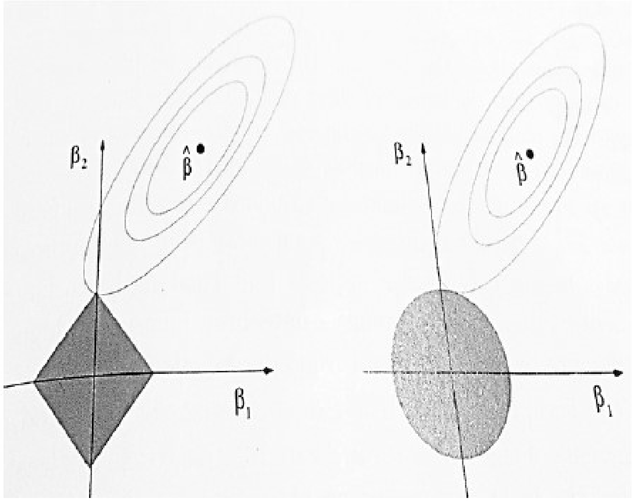
   (a) Give the logistic regression specification you will run. Take care to define all relevant variables. (10 pts)

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 IV_{t-1} + \beta_2 skew_{t-1} + \beta_3 ba_{t-1} + \beta_4 oi_{t-1}$$

where $p = Prob\,(x_t = 1)$

(c) You decide to add an Elastic Net constraint to the model.

   i. Explain clearly what an Elastic Net constraint is. (5 pts)
   The constraint is an L_1-constraint, which means the we put a constraint on the sum of the absolute values of each coefficient. This is compared to an L_2-constraint, where the constraint is on the sum of squared coefficient values, or OLS which has no constraint. The absolute value constraint produces kinks that make corner solutions with zero coefficients much more likely than in OLS or Ridge regressions, where the likelihood of an exactly zero coefficient is in fact zero. Also, recall the plot from the class notes for a visual depiction of this:



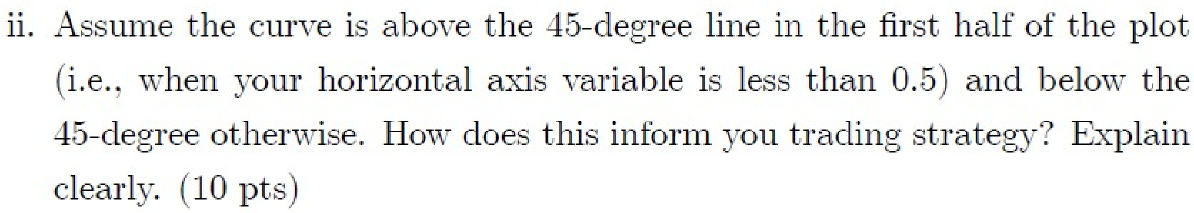
ii. Explain the K-fold cross-validation procedure often used to find the 'optimal' value of the constraint parameter 'lambda.' (7 pts)

Divide the data into K equal subsamples. Then, estimate the model based on K-1 subsamples and find mean squared error of model prediction on the K'th out-of-sample fold. Do this for all K-1 permutations. Choose the lambda that minimizes the average mean squared error across the K out of sample folds.

(d) You decide to use an ROC-curve to show the results and to help devise a trading strategy based on this model. Explain:

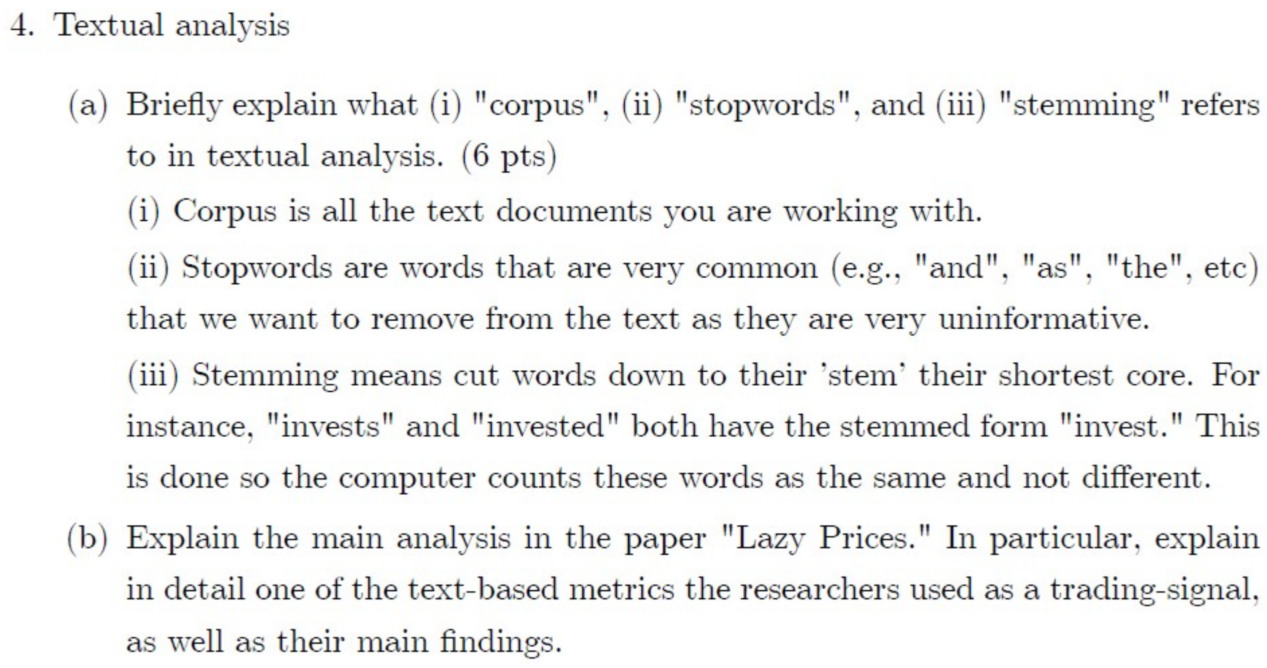
i. What an ROC curve is. In particular, what goes on each axis and how do you define these variables.

We use ROC curves to compare the performance of predictive models. An ROC has the true positive rate (also called sensitivity) on the y-axis and the false positive rate (100-specificity) on the x-axis. The true positive rate represents the proportion of observations with a certain characteristic (think of days during which the stock market is up) that are correctly identified to possess that characteristic. The false positive rate is the proportion of observations that do not possess the characteristic that are erroneously determined to possess the characteristic (think of days during which the stock market is down but our model states that the stock market is up).



ii. Assume the curve is above the 45-degree line in the first half of the plot (i.e., when your horizontal axis variable is less than 0.5) and below the 45-degree otherwise. How does this inform you trading strategy? Explain clearly. (10 pts)

Based on the scenario described in this part, we would conclude that our model works fine for some observations but fails for others. An example of such a scenario is the value effect related to small and big stocks. A reasonable approach would be to only use our predictive model for cases in which $x < 0.5$.



4. Textual analysis

(a) Briefly explain what (i) "corpus", (ii) "stopwords", and (iii) "stemming" refers to in textual analysis. (6 pts)

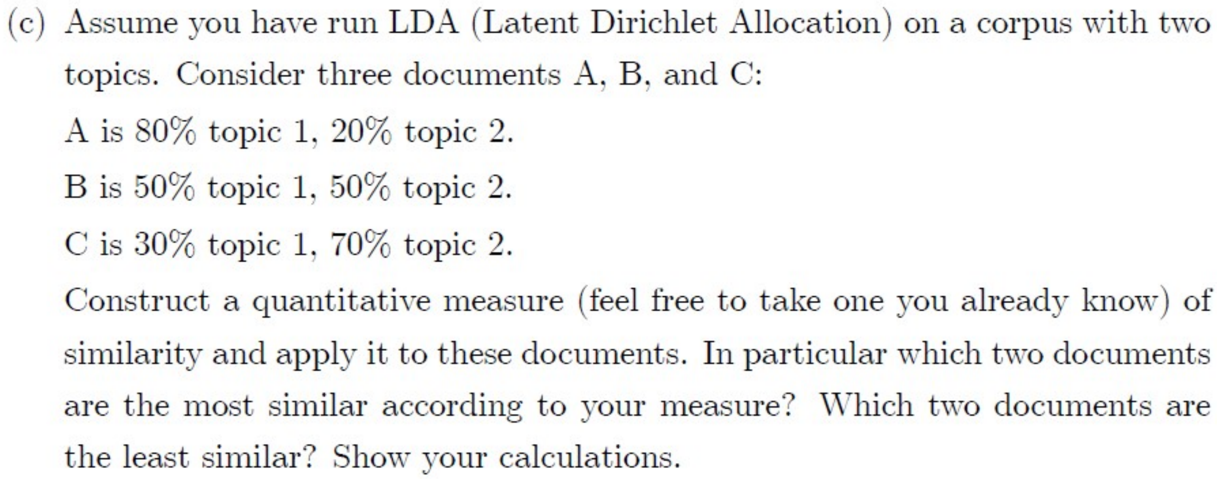
(i) Corpus is all the text documents you are working with.

(ii) Stopwords are words that are very common (e.g., "and", "as", "the", etc) that we want to remove from the text as they are very uninformative.

(iii) Stemming means cut words down to their 'stem' their shortest core. For instance, "invests" and "invested" both have the stemmed form "invest." This is done so the computer counts these words as the same and not different.

(b) Explain the main analysis in the paper "Lazy Prices." In particular, explain in detail one of the text-based metrics the researchers used as a trading-signal, as well as their main findings.

See Topic 8, slides 66 through 80 for a detailed description of the Cohen et al. (2015) paper.



(c) Assume you have run LDA (Latent Dirichlet Allocation) on a corpus with two topics. Consider three documents A, B, and C:

A is 80% topic 1, 20% topic 2.

B is 50% topic 1, 50% topic 2.

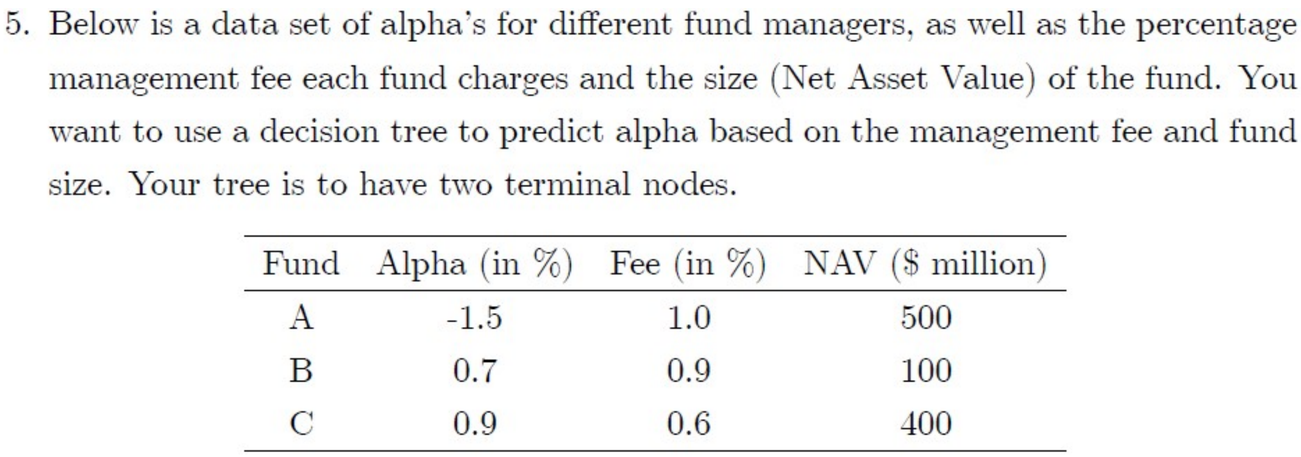
C is 30% topic 1, 70% topic 2.

Construct a quantitative measure (feel free to take one you already know) of similarity and apply it to these documents. In particular which two documents are the most similar according to your measure? Which two documents are the least similar? Show your calculations.

We cannot answer this question without making an additional assumption. I assume that the number of words contained within both topics 1 and 2 is similar to the length of each of the documents. Armed with the assumption, I propose the following measure:

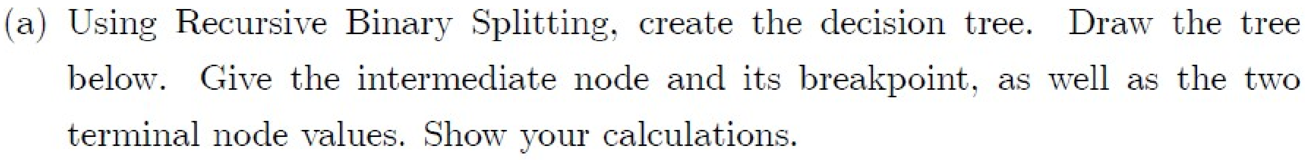
$$Pr\,(topic\,1_i)\,Pr\,(topic\,1_j) + Pr\,(topic\,2_i)\,Pr\,(topic\,2_j)$$

Based on this measure, A and B (0.5) and B and C (0.5) are tied for most similar and A and C (0.38) are the least similar. Any measure based on sound logic can be used instead of the one presented here.



5. Below is a data set of alpha's for different fund managers, as well as the percentage management fee each fund charges and the size (Net Asset Value) of the fund. You want to use a decision tree to predict alpha based on the management fee and fund size. Your tree is to have two terminal nodes.

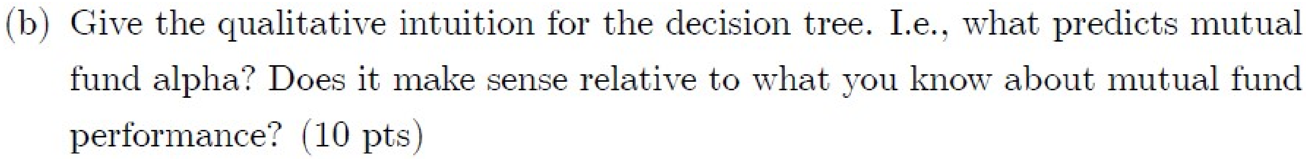
Fund	Alpha (in %)	Fee (in %)	NAV ($ million)
A	-1.5	1.0	500
B	0.7	0.9	100
C	0.9	0.6	400



(a) Using Recursive Binary Splitting, create the decision tree. Draw the tree below. Give the intermediate node and its breakpoint, as well as the two terminal node values. Show your calculations.

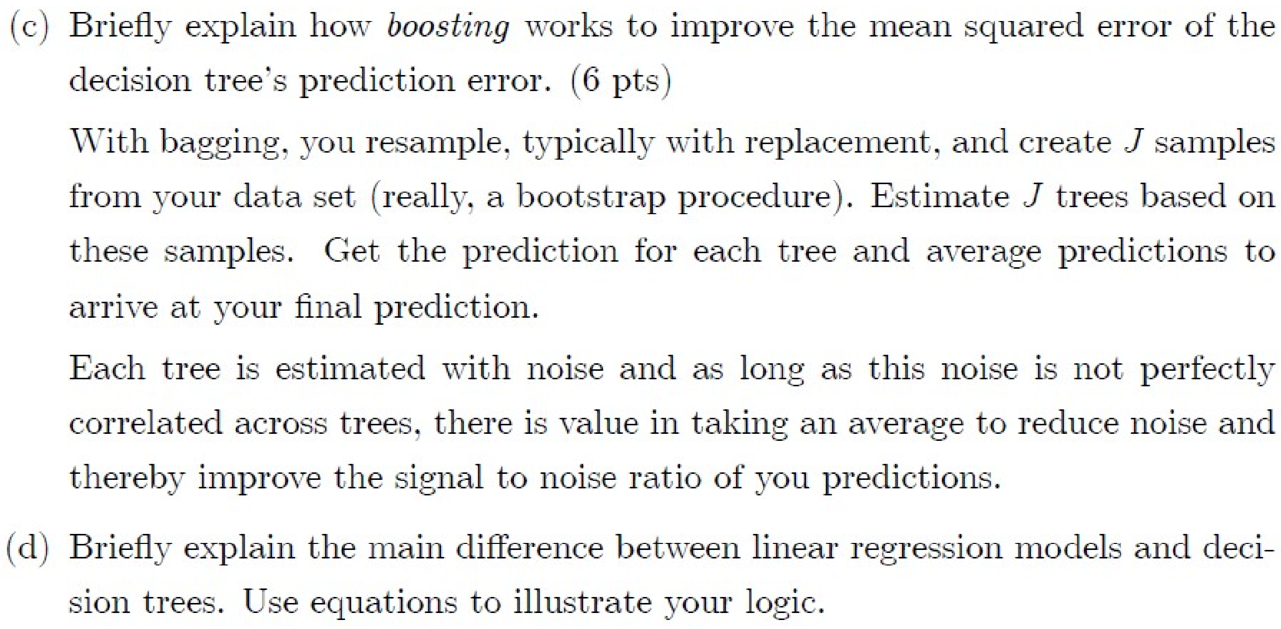
Fee$< 1 \Rightarrow \hat{\alpha} = 0.8$

Fee$\geq 1 \Rightarrow \hat{\alpha} = -1.5$



(b) Give the qualitative intuition for the decision tree. I.e., what predicts mutual fund alpha? Does it make sense relative to what you know about mutual fund performance? (10 pts)

Both size and management fees predict firm performance. Higher fees translate into lower alpha, as expected, given that management fees eat into fund's performance. Also, bigger funds (higher NAV) underperform relative to smaller funds (lower NAV). The second fact is a well-known puzzle in the finance academic literature. Note that in our case NAV and fee breakpoints provide the same information.



(c) Briefly explain how *boosting* works to improve the mean squared error of the decision tree's prediction error. (6 pts)

With bagging, you resample, typically with replacement, and create J samples from your data set (really, a bootstrap procedure). Estimate J trees based on these samples. Get the prediction for each tree and average predictions to arrive at your final prediction.

Each tree is estimated with noise and as long as this noise is not perfectly correlated across trees, there is value in taking an average to reduce noise and thereby improve the signal to noise ratio of you predictions.

(d) Briefly explain the main difference between linear regression models and decision trees. Use equations to illustrate your logic.

The two models also differ in functional form assumptions:

- Linear regression:

$$f(X) = \beta_0 + \sum_{j=1}^{P} X_j \beta_j$$

- Decisions trees:

$$f(X) = \sum_{m=1}^{M} \beta_m \mathbb{I}\left\{X \in R_m\right\}$$

The intuition is that with decision trees dummy variables allow for non-linear 'buckets' and the definition of the dummy variables is endogenous to the procedure.

Both approaches use the same optimization logic to obtain the coefficients of interest: we minimize the sum of squared errors. However, the minimization problems take on a different form:

- Linear regression :

$$min_{\tilde{\beta}} \left\{ \left(y - X\tilde{\beta}\right)' \left(y - X\tilde{\beta}\right) \right\}$$

- Decision trees:

$$min_{\{R_j\}_j} \sum_{j=1}^{J} \sum_{i \in R_j} \left(y - \hat{y}_{R_j}\right)^2$$