Problem 1

Part I

In order to construct the portfolio, I created a generalized linear model with Lasso, where the companies that had a span of 1342 days on the stock market. Using the PRC data for these subset of companies, I constructed the generalized linear model using the SP 500 level as the response variable. In order to contruct the porfolio, I decided that the best number of different stocks to realistically have was 20, so I looked for lambda's with the length closest to 20 and used that lambda the get the coefficients of each company. I omitted any company with the coefficient of 0. These were the following companies: "AMERICAN_ELECTRIC_POWER_CO_INC", "AMERIPRISE_FINANCIAL_INC", "AUTODESK_INC", "BOEING_CO", "CAPITAL_ONE_FINANCIAL_CORP", "CINCINNATI_FINANCIAL_CORP", "DENTSPLY_INTERNATIONAL_INC_NEW", "HARLEY_DAVIDSON_INC","HONEYWELL_INTERNATIONAL_INC","ILLINOIS_TOOL_WORKS_INC","INTERNATIONAL_PAPER_CO", "MACERICH_CO","MOLEX_INC","NEWELL_RUBBERMAID_INC", "NEWS_CORP", "OMNICOM_GROUP_INC "PATTERSON_COMPANIES_INC", "TOTAL_SYSTEM_SERVICES_INC", and "WILLIAMS_COS"

Part II

The changing portfolios are not very stable, when comparing a following 60 day designed portfolio with the previous, the maximum number of stocks that are retained are 5 given that the portfolio being searched for are the ones closest to 20 different stocks. This occurs only twice out of the the 22 60-day intervals. Adding Penalties of lower valued coefficients to the newer companies may help to lower the amount of chnages

Part III

The difference in issue will be the way that PRC is accounted for. Instead of taking the absolute value, we will impose a subset of the SP data, such that all PRC values that are not greater than 0

Part IV

Part I

When I switched to the DailyReturns, the following companies were picked:"A_E_S_CORP","AMERISOURCEBERGEN_CORP","APOLLO_GROUP_INC", "BAKER_HUGHES_INC""CAPITAL_ONE_FINANCIAL_CORP","C SON_WORLDWIDE_INC","COACH_INC","DU_PONT_E_I_DE_NEMOURS___CO","FLIR_SYSTEMS_INC","INTERPUBLIC_GROUP_COS_INC", "K_L_A_TENCOR_CORP”,"MARRIOTT_INTERNATIONAL_INC_NEW", "MASTERCARD_INC", "MCKESSON_H_B_O_C_INC", "NABORS_INDUSTRIES_LTD", "PRICELINE_COM_INC", "SAFEWAY_INC","SNAP_ON_INC" "SPRINT_NEXTEL_CORP","WESTERN_UNION_CO". The difference was the construction of the text matrix, which was changed to DailyReturns from PRC.

Part II

The changing portfolios still are not very stable, when comparing a following 60 day designed portfolio with the previous, the maximum number of stocks that are retained are 4 instead of 5 given that the portfolio being searched for are the ones closest to 20 different stocks. This also occurs only twice out of the the 22 60-day intervals.

Part V

In order to do this part the ideal situation I attempted was to create a standard generalized linear model using Biblo, Askhi, DailyReturns, VOL, OPENPRC, and Numtrd to predict the SP Level. Then using the Predict SP 500 level for each stock, i would impliment another generalized linear model with Lasso. I wouldlook for lambda's with the length closest to 20 (see 1a for my explanation) and used that lambda the get the coefficients of each company. Those companies would be part of my stock portfolio.

Problem 2

Part A

$\delta_2(x) = x^T \sum^{-1} \mu_2 - \frac{1}{2} x_2^T \sum \mu_2 + log(\pi_2)$

$\delta_1(x) = x^T \sum^{-1} \mu_1 - \frac{1}{2} x_1^T \sum \mu_1 + log(\pi_1)$

Assume $\delta_2(x) > \delta_1(x)$ if LDA rule classifies to class 2

$x^T \sum^{-1} \mu_2 - \frac{1}{2} x_2^T \sum \mu_2 + log(\pi_2) > x^T \sum^{-1} \mu_1 - \frac{1}{2} x_1^T \sum \mu_1 + log(\pi_1) \implies x^T \sum^{-1} (\mu_1 - \mu_2) > \frac{1}{2} x_2^T \sum \mu_2 - log(\pi_2) - \frac{1}{2} x_1^T \sum \mu_1 + log(\pi_1) \implies x^T \sum^{-1} (\mu_1 - \mu_2) > \frac{1}{2} x_2^T \sum \mu_2 - log(\frac{N_2}{N}) - \frac{1}{2} x_1^T \sum \mu_1 + log(\frac{N_1}{N})$

Part B

$\sum_{i=1}^{N} (y_i - \beta_0 - \beta^T X)^2 \implies RSS = (Y - \beta_0 - X\beta)^T (Y - \beta_0 - X\beta)$

Part C

Given: $\hat{\sum}_\beta = (\boldsymbol{\mu_2 \mu_1})(\boldsymbol{\mu_2 \mu_1})^{\mathbf{T}} \implies \hat{\sum_\beta} \hat{\beta} = (\boldsymbol{\mu_2 \mu_1})(\boldsymbol{\mu_2 \mu_1})^{\mathbf{T}} \hat{\beta}$

Problem 3

Accurracy	Training	Test
LDA	0.7533333	0.6851852
Logistic Regression	0.75	0.691358

In order to analyze the data, I used the variables discussed in section to determine which factors to use. If I had attempted it, I would have used variable selection to determine the appropriate variables. Thus, the variables I use were the following, "tobacco", "ldl", "famhist", "age" to determine chd. The training and test set were created by randomly selecting 300 observations and the remaining 162 observations were placed in the test set. In order to determine the goodness of the classifers, I used the accurracy. Based on the accurracy of both the LDA and Logistic Regression, the Logistic Regression seems to be a better classifier.