Final

APPLIED LINEAR STATISTICAL MODELS
26:960:577, Fall 2017
Instructor: Mert Gurbuzbalaban

This version: December 14, 2017

Instructions: Please submit your answers through blackboard and write your name to the submissions. Late submissions will not be accepted. All the answers should be justified properly for getting any credit. For questions about the final, please do not send the instructor an email; but post a question to the blackboard instead. This way, all the class members can follow the questions and answers simultaneously.
Deadline: *December 20th, Wednesday, 23:50pm.*
Attention: *No late submissions will be accepted. Make-up final exams are not available.*

General Information: The first two questions are about linear regression (predicting stock prices in finance and modeling the taste of cheese products), the third question is about ANOVA models and their application to house prices. Finally, the last question is about predicting the default rate of customers that own credit cards with logistic regression. The second question is directly taken from the Faraway, Linear Models in R book; the problems are very similar to the homework questions and problems we solved in class. I would encourage you to go over the class notes to have a look at the relevant R code.

Questions:

(1) The file `stockdata.csv` uploaded to the blackboard is a dataset that contains the price of a stock in the last 100 days as the response and the following variables as predictors:
- vol: Volatility of the stock
- cap.to.gdp: The ratio of the market cap to GDP
- q.ratio: The ratio of market cap to net worth
- gaap: Shiller Cape index
- avg.allocation: Average investor equity allocation of the stock

Fit a model to explain price in terms of the predictors. Perform regression diagnostics to answer the following question. Display any plots that are relavant and explain your reasoning. Suggest possible improvements if there are any.

(a) Fit a model to explain price in terms of the predictors. Which variables are important, can any of the variables be removed ? Please use F-tests to justify.
(b) Construct confidence intervals using permutation tests.
(c) Check the constant variance assumption for the errors.
(d) Check the independentness of the errors assumption.
(e) Check the normality assumption
(f) Is nonlinearity a problem?
(g) Check for outliers, compute and plot Cook's distance
(h) Check for influential points.

(i) The return at time t is defined as

$$r(t) = p(t+1)/p(t) - 1$$

where p is the price data for day t. Are the returns normally distributed? Please justify your answer using Q-Q plots and normality tests.

(2) Repeat the same question from (a) to (h) on the `cheddar` dataset (except part (i)) from the Faraway package by fitting a model with `taste` as the response and the other three variables as predictors. Answer the questions posed in the first question.

(3) The problem is to discover relation between US new house construction starts data (HOUST) and macro economic indicators: GDP, CPI and Population (POP). Please download the relevant data from the *house.zip* file provided as an attachment. The description for this data can be found in *https://fred.stlouisfed.org/*.
   (a) Data preparation: combine all data into an R dataframe object, and construct dummy or factor variable for 4 quarters. First model is $HOUST \sim GDP + CPI + quarter$.
   (b) Use one-way ANOVA to determine whether there's a seasonal effect. Show necessary steps and explanation.
   (c) Do pair-wise comparison for different levels. In particular, construct %90 confidence intervals for the pairwise differences. Hint: Please see the lecture notes in the slides.
   (d) Add population to the first model, do the steps (b) and (c) again.

(4) Read the `train-default.csv` and `test-default.csv` files in R which contains training and test data containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt. These datasets contain the following information/variables:

`default`	A factor with levels `No` and `Yes` indicating whether the customer defaulted on their debt
`student`	A factor with levels `No` and `Yes` indicating whether the customer is a student
`balance`	The average balance that the customer has remaining on their credit card after making their monthly payment
`income`	Income of customer

Hints: In class, we provided solutions in R to a similar problem but for a different dataset.
   (a) Fit a logistic regression model with the `default` as the response and the variable `balance` as the predictor. Make sure that predictor variable in your model is significant. Perform regression diagnostics on this model to answer the following questions. Display any plots that are relevant.
   (b) Why is your model a good/reasonable model? Check the AIC and pseudo-R^2 values.
   (c) Give an interpretation of the regression coefficients (in words).
   (d) Form the confusion matrix over the test data. What percentage of the time, are your predictions correct?

(e) Now, let's add the variables `income` and `student` to the model. Fit a logistic regression model of the form "default balance + income + student", in other words, regress the variable `default` to all the other predictors with logistic regression. Repeat steps (a) to (d).

(f) In your model in question (e), what is the estimated probabilty of default for a *student* with a credit card balance of $\$2,000$ and an income of $\$40,000$? What is the probabilty of the default for a *non-student* with the same credit card balance and income?

(g) Are the variables `student` and `balance` are correlated? If yes, why do you think this is the case? If no, please explain.

(h) (Extra Credit) Does the data say that it is more likely for a student to default compared to a non-student for different values of income level? Please comment. In other words, if you were the credit card company, would you prefer students as customers or non-students as customers with the same income level?