Final

APPLIED LINEAR STATISTICAL MODELS
26:960:577, Fall 2020
Instructor: Mert Gurbuzbalaban
This version: December 9, 2020

Instructions: Please submit your answers through Canvas and write your name and your project partner's 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 ask them in class on December 10th (as this is the last week of classes and office hours).
Deadline: *December 16th, Tuesday night, 11:59pm.*
Attention: *No late submissions will be accepted. Make-up final exams are not available.*
General Information: Please write the name of the students in your group to the first page in your submission with a description of how the workload was shared between the two students participating in the final project.The first two questions are about linear regression, the third question is about discrete predictors (a.k.a. factors), the fourth question is about predicting the default rate of customers that own credit cards with logistic regression. 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) Check the constant variance assumption for the errors.
(c) Check the independentness of the errors assumption.
(d) Check the normality assumption
(e) Is nonlinearity a problem?
(f) Check for outliers, compute and plot Cook's distance
(g) Check for influential points.

(h) 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 book 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 *house.zip* from blackboard. 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 $HOUST \sim GDP + CPI + quarter$.

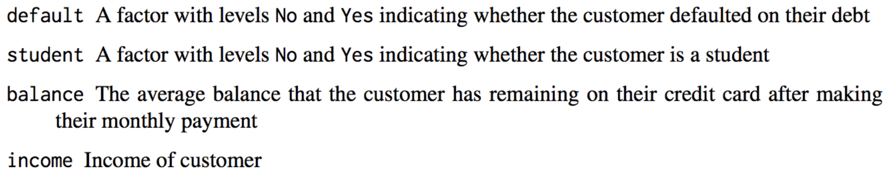
(b) Do you think the data needs some cleaning? If so, clean the data.

(c) Is there a seasonal effect you observe in data? Show necessary steps and explanation. This is an open-ended question and you are free to use any tool that you find appropriate.

(d) Do a pair-wise comparison for different quarters. Which quarter do you think is the best one to buy a house? Show necessary steps and explanation. Use any statistical test or tool that you think is appropriate, this is an open-ended question and there is no one way of answering the question.

(e) Add population to the first model, do the steps (b) and (c) again.

(4) Read the `train.csv` and `test.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 contains 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 other variables `balance` and `income` as the predictor. Make sure that variables in your model are 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) In your model, 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 to default?

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

(g) Now, let's add the binary variable `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.

(h) 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.

(5) These days, there are a lot of discussions about what should the healthcare system look like in US. For a scientific discussion, one should need to have a model of demand in the healthcare system. In this question, we will work on the the dataset *dvisit* which is about modeling the demand for doctor visits in terms of explanatory variables such as age, income, existence of health insurance and others. To load this dataset, in R, we type in the commands:

```
install.packages('faraway')
library(faraway)
data(dvisits)
```

which downloads the library 'faraway', loads this library and then pulls up the dataset dvisits from this library. The information about this dataset from the Faraway package can be found at the following document:

`https://cran.r-project.org/web/packages/faraway/faraway.pdf`

The following exercise is about fitting a model to data and checking diagnostics of it, making sure that our model is right. **Hints:** In class, we provided solutions in R to a similar problem but for a different dataset. I will also give many hints in the class for doing the homework, we will go over the homework questions together.

(a) Using the dvisits dataset, fit a model with the "hospdays" as the response and other variables as potential predictors. Make sure that variables in your model are significant. Note that there is no single perfect model for this dataset, you can do your best for the fit. We can accept any model as long as your variables are reasonably significant and you can justify the variables in your model in words about why/how they should be predictive. We will accept all the models as long as they do not have any serious flaws in them, so feel free to be creative and do not be afraid about playing with variables. 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 constant variance assumption for the errors.

(c) Check the normality assumption.

(d) Are the errors correlated?

(e) Check for leverage points, outliers, influential points.

(f) Check the structure of the relationship btw the predictors and the response

Our PhD students (and Masters students who are interested in doing academic research) can also check out the following research article about this dataset to get more information about economics of healtcare and potential research topics in this direction. However, this material is completely optional, not required for this class (but is provided for students interested in research).

`Cameron A, Trivedi P, Milne F and Piggot J (1988) A Microeconometric model of the demand for health care and health insurance in Australia, Review of Economic Studies 55, 85-106.`

Click on this link to access this paper

(6) The following data provides the Covid-19 cases per state since January:

`https://covidtracking.com/api/v1/states/daily.csv`

The purpose is to predict "the total number of cases in US per day" with linear regression. Please use the data till the end of September for training and the rest for testing. Perform diagnostics on your model and show that your model is a good model. This is a harder question (such as the optional homework HW3), so a "perfect model" may not exist; the purpose is to do "our best".