2016 Chicago Booth Big Data Midterm

- This is an **INDIVIDUAL** exam: you must work alone.
- The exam must be submitted before **8:30 am on Wednesday Feb 10**.
  This deadline is the same for all sections.
- You should submit a Google Doc or pdf document as a link via Google forms, exactly as we request for homework submission. We will post the form link on piazza.
- Each sub-question (i.e., X.i) is worth the same amount. Within each question, the sub-questions are roughly ordered by difficulty and the X.5 sub-questions are really hard!
- The exam is not easy! And it is long.
  You may still do fine even if you cannot finish, but you will want to start early.
- Coding questions should be asked on Piazza, but you must not give away any answers when you formulate your query. We will only answer questions on R coding – not statistics concepts. We will continue to answer conceptual questions if they pertain to homework solutions or lecture examples.
- You will be graded on your written answers and analysis. Please include and refer to plots to illustrate your answers.
- Please label your graphs and figures (e.g. "Figure 1: Relationship between A and B."), and do not forget to label the axes on these graphs.
- Do not hand in detailed code, but do submit meaningful R output to help us understand your answer (e.g., printed $\hat{\beta}$ values, etc). All but immediately relevant R output should be relegated to an appendix.
- Please be very clear: your answers need to be concise, precise, and easy to understand. Neatness in presentation is expected.

Light Beer purchases in the US

Data and code are on the course site. This includes `lightbeer-start.R`; much of the code you will need has already been written for you in this script.

The data are 73,115 observations light beer purchases in the US. These observations are taken from the Nielsen Home-Scan database.

`lightbeer.csv` contains:

- `household`: an identifier of the purchasing household
- `upc_description`: details on the purchased *item* (e.g., a 6-pack)
- `quantity`: the number of items purchased
- `beer_brand`: *Bud Light, Busch Light, Coors Light, Miller Lite,* or *Natural Light*.
- `beer_spend`: total dollar value of purchase.
- `beer_floz`: total volume of beer, in fluid ounces.
- `price_floz`: price per fl.oz. (i.e., `beer_spend`/`beer_floz`)
- `container_descr`: the type of container.
- `promotion` Whether the item was promoted (coupon or otherwise).
- `market`: Scantrack market (or state if rural)
- demographic data, including buyer type (married couple, male, or female), income, occupation, education, age, and whether the household has a dishwasher or a microwave.

We will consider `beer_spend`, `price_floz`, and `beer_brand` (and transformations thereof) as the responses of interest.

1 Linear models and false discovery control

Our starter script defines the basic regression `mle`.

1.1

Write the mathematical formula for the regression model behind `mle` and describe it in words. You don't need to list every demographic variable, but be sure to describe important interactions and the probability model that is implied by the deviance we've minimized. What is our estimate for the error variance?

1.2

Compare (i.e., with a plot) the fitted residuals and log volume. Based upon this plot alone, (a) describe the fit and (b) suggest one degree of freedom that could be added to your model to improve the fit (and see how it works).

1.3

Given the original `mle` fit, all other things equal, what is the change in price

- for Rural North Dakota relative to Chicago?
- for Bush Light relative to Natural Light?
- for Coors Light due to promotion, at average log volume?
- for a Coors Light 6 pack (12oz cans) due to promotion?

1.4

Plot the p-value distribution. What does it tell you about the underlying signal in this regression? What is the p-value cutoff for an FDR of 1%? What is the FDR at a p-value cutoff of 0.05?

1.5

Run a bootstrap for the sampling distribution of the coefficient on log(volume) – $\beta_{\texttt{logvol}}$ – *using the code and random seed in the starter script*. Compare the 80% confidence interval built from your bootstrap to the theoretical 80% confidence interval implied by `summary(mle)`. Give potential explanations for any discrepancy in the regions covered by these intervals.

2 Treatment Effects Estimation

For this question, we'll attempt to measure the treatment effect of pricing on sales (measured as `exp(logspend)`). In the starter script, I've built for you one large model matrix of demographic information and another containing beer information (brand, promotion status).

2.1

Fit a univariate regression for `logspend` on `logprice`, as well as a lasso with AICc selection for `logspend` onto `logprice` and demographic and beer information. Describe in words and compare each procedure's estimate for the sensitivity of sales to price.

2.2

Describe why you should be suspect of the estimates from each procedure in 2.1. How would you instead implement a randomized controlled trial to estimate price sensitivity, and why would your suspicions no longer apply in that setting?

2.3

Build a model to predict `logprice` from the controls. Describe the in-sample fit for this regression; what are the implications for your ability to measure the treatment effect of price?

2.4

Use the fitted values from 2.3 in a new lasso that estimates the treatment effect of `logprice` on `logspend`. How does your estimate compare from the naive lasso in 2.1, if at all? Why?

2.5

Repeat both the naive lasso and the treatment lasso for only transactions in Los Angeles by people making more than $200,000. You should use the same price regression from 2.3; why? Compare the estimated price effects to each other and to the full sample results.

3 Lasso Linear Regression and Model Selection

The starter script defines a large design of interactions between price and our other variables. This is combined with the beer and demographic information in the sparse matrix `x`.

3.1

Run a cross-validated lasso for `logspend` onto `x`. Describe the criteria used to choose models under `select="1se"` and `select="min"` rules. What are estimated out-of-sample R^2 for models fit under each rule?

3.2

Compare AICc, AIC, and BIC selection for this lasso to each other and to the CV choices. (You can exponentiate each IC divided by n to plot them on the same scale as the CV results).

3.3

What is the effect of price on sales of Bud Light for people earning $> \$200k$ in Los Angeles? Compare this to the treatment effect estimates in question 2. Why is it different and in what settings would we prefer the current model?

3.4

What is the average price sensitivity in our model? What is the price sensitivity for the median consumer (see starter script for `xbar_median`).

3.5

Bootstrap to estimate the sampling distribution for the AICc selected model degrees of freedom (note that you don't need to run CV in the bootstrap). *You can alter the bootstrap code that is written for you in the starter script.* Compare the results to your full sample AICc regression degrees of freedom, and attempt to describe the sources of any discrepancy.

4 Logistic Regression and Classification

4.1

Define the binary outcome variable `ybud` that is one for Bud Light and zero otherwise. Fit the lasso path for regression of `ybud` onto the demographic information in `xdemog` (no need for cross validation). Plot and describe (quickly) the fitted values.

4.2

What is the precise function that was minimized when fitting the AICc selected model here? What is the in-sample R^2 for this fit?

4.3

What is the sensitivity and specificity of a classification rule that considers every consumer with probability greater than 0.1 of buying Bud Light a potential Bud customer? Plot the ROC curve and add this point to the picture.

4.4

Looking only at the main effect coefficient (in starter script), how do the odds of a consumer choosing Bud Light change in Rural California relative to the intercept market? What markets are included in this intercept? How does the story change when you consider interaction between income and being in Rural California?

4.5

Use the starter script commands to run a multinomial logistic regression for brand choice onto the demographic info. Quickly describe the computing environment set up by the `makeCluster` command here (including the number of cores you are using). Consider the multinomial regression coefficients related to your answers in 4.4: what can you add to your story?

Bonus Freestyle Analysis

Provide additional analysis of the data. Points will handed out only for insightful use of data mining tools, not for scattershot application of techniques.

Do not spend too much effort here: it is worth only up to 1/2 of one of the questions.

USE A MAXIMUM OF TWO PAGES, INCLUDING PLOTS, TO ANSWER.