2015 Chicago Booth Big Data Midterm

- This is an **INDIVIDUAL** exam: you must work alone.

- The exam must be submitted on chalk before **8:30 am on Friday May 8**.
  This deadline is the same for all sections. You should submit a pdf document on chalk.

- 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.

Washington DC Bikeshare Data

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

The data are 17379 observations of hourly counts from 2011 to 2012 for bike rides (rentals) from the Capital Bikeshare system in Washington DC. It was originally compiled by Fanaee and Gama in 'Event labeling combining ensemble detectors and background knowledge' (2013).

`bikeshare.csv` contains:

- `season`: 1:spring, 2:summer, 3:fall, 4:winter
- `yr`: year (0:2011, 1:2012)
- `mnth`: month (1 to 12)
- `hr`: hour (0 to 23)
- `holiday`: whether day is holiday or not
- `weekday`: day of the week, counting from 0:sunday.
- `notbizday`: if day is either weekend or holiday is 1, otherwise is 0.
- `weathersit`:
    1. Clear, Few clouds, Partly cloudy, Partly cloudy
    2. Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    3. Light Snow, Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    4. Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- `temp` Temperature, measured in standard deviations from average.
- `hum`: Humidity, measured in standard deviations from average.
- `windspeed`: Wind speed, measured in standard deviations from average.
- `dteday`: date
- `cnt`: count of total rental bikes

We will consider `cnt` (and transformations thereof) as the response of interest.

1 Models, Outliers, and False Discovery

I've aggregated the data to *daily* counts and run the simple regression `daylm`.

1.1

What are the in-sample SSE and R^2 for this regression?

1.2

Write out the mathematical formula for `daylm` and describe it in words. Make sure to describe the probability model that is implied by the deviance we've minimized. Do you have any criticisms of this model?

1.3

A standardized residual for response y and prediction $\hat{y}$ is $r_i = (y_i - \hat{y}_i)/\hat{\sigma}$, where $\hat{\sigma}$ is the estimated standard deviation of residuals $y - \hat{y}$. Calculate the standardized residuals for `daylm`.

Now, we'll call the *outlier p-value* $2 \times \mathrm{p}(Z < -|r_i|)$ where $Z \sim \mathrm{N}(0,1)$. In R, this is `2*pnorm(-abs(std_resids))`. Calculate these *p-values*. Describe what null hypothesis distribution they correspond to and why small values indicate a possible *outlier* day.

1.4

What is the p-value rejection region associated with a 5% False Discovery Rate here?
Which observations (days) are in this rejection region? Do you have any explanation for them?

1.5

Plot the p-value distribution. What does it tell you about the assumptions of the probability model we used for our regression?

2 Lasso Linear Regression and Model Selection

For this question, consider the `cv.gamlr` object I've fit as `fitlin`.

2.1

What is our response variable? Describe the columns of our model matrix. How has this model addressed the 'outlier detection' of question 1?

2.2

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 using these λ?

2.3

Compare AICc, AIC, and BIC selection to each other and to the CV rules.

2.4

Print the top three `dteday` effects by absolute value under your preferred selection rule, and describe the implied effect on `cnt`. Can you explain any of these?

2.5

Bootstrap to get an estimate of the sampling distribution for AICc and BIC selected lambdas. Compare these distributions and describe why they look like they do.

3 Logistic Regression and Classification

The managers of Capital Bikeshare have found that the system works smoothly until more than 500 bikes are rented in any one hour. At that point, it becomes necessary to insert extra bikes into the system and move them across stations to balance loads.

3.1

Define the binary outcome variable `overload` that is one if `cnt` > 500, zero otherwise. Fit and plot the lasso path for regression of `overload` onto the same model matrix used in Question 2 (no need for cross validation).

3.2

Summarize how hour-of-day effects the probability of an overload during business days. Consider a single hour with a strong effect and compare this to its effect in the regression of Q2.

3.3

Suppose that it costs you $200/hr in overtime pay if you have an overload (`cnt` > 500) with your usual number of staff. Staffing an extra driver to move the bikes costs only $100/hr and means you don't have to pay any overtime. At what probability for `overload` > 0 will you want to staff an extra driver?

3.4

Plot and describe the ROC curve for your AICc-optimal regression from 3.1. What is the sensitivity and specificity of your rule from 3.3 if applied with this regression?

3.5

Now, take the `test` sample and

- fit the regression path excluding this sample (e.g., on `mmbike[-test,]`).

- use the AICc-optimal model from this path to predict for the test set.

- plot the 'out-of-sample' ROC curve for these predictions.

Compare this curve to your ROC curve from 3.4 and describe what they imply about the quality of AICc selection for this regression.

4 Treatment Effects Estimation

For this question, we'll revisit the regression model of Q2 with the goal to infer the *independent* effect of humidity.

4.1

Based on 'naive' `fitlin` estimate, what is the effect of an extra standard deviation of humidity (`hum` increasing by one unit) on the count of bikes rented?

4.2

Predict humidity from a model matrix `x` that includes all our covariates except humidity. Describe how close the predicted values are to true humidity, and how this is relevant for our goal in this question.

4.3

Obtain an estimate for the treatment effect of humidity on bike rentals, and describe this estimate in plain words. How does it compare to your estimate in 4.1?

4.4

Extend the `fitlin` model from 4.1 (and Q2) to allow for the effect of humidity to depend upon the temperature. Describe the resulting relationship between humidity and ride count.

4.5

Finally, extend your causal model from 4.3 to measure the temperature-dependent effect of humidity. Describe the results and compare to 4.4. Hint: you need to control for the *interaction* between temperature and the portion of humidity that is predictable from the controls.

Bonus

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

Do not spend too much (or any!) effort here: it is worth little.