IEOR E4525
Martin Haugh

Due: Monday 1 Feb 2016

Assignment 1

You do not need to submit anything for Questions 1 and 3.
You only need to do one of Questions 4 and 5.

1. *(Implications of Big Data)*

   The McKinsey Global Institute recently published a report titled "Big data: The next frontier for innovation, competition, and productivity". The report is available here.

   Read the **executive summary** of the report.

2. *(EDA with the Spam Filtering Data Set)*

   The csv file `spam.csv` contains a data set for emails that were categorized as spam or not spam. The documentation for this data set is in the file `spam-info.pdf`. All the files referred to here are on CourseWorks. (The data is courtesy of Di Cook at Iowa State via Chris Volensky at AT&T Labs.)

   (a) Look at the documentation. What is the variable of interest, i.e. the dependent variable?

   (b) For each of the independent variables, report something about it. Specifically, you should report on each variable's relationship with the response, i.e dependent, variable. Pay special attention to variable type (binary, ordinal, real) when doing this. Your comments should contain at least some tables and graphs.

   (c) Investigate the variable 'spampct'.

      i. How many missing values does it have?

      ii. Compare graphically the distribution for time.of.day for the cases where spampct is missing against the distribution of time.of.day when spampct is present. Do you see any differences?

      iii. Plot a scatter plot of time of day vs. spampct. How many unique points (x,y coordinates) are plotted? Explain a technique you might use to deal with the overplotting.

   (You may find the `R` commands in the file `eda_example.R` useful.)

3. *(Fitting Regression Models in `R`)*

   Work through the material of Section 3, "Lab on Linear Regression", in ISLR by James et al. This lab will show you how to fit linear regression models in R. You will first

need to install the `ISLR` package in `R`. (This package contains all the data-sets that are used in the ISLR book except for the `Boston` data-set which is part of the pre-installed `MASS` package. I also recommend using `RStudio` as your IDE for `R`.) The `R` code for all the labs in the book can be downloaded from here.

4. *(Exploring the Relationship Between Overfitting and Noise)*

   Do Exercise 13 in Chapter 3 of ISLR. (You are free to use whatever software you like but of course `R` is probably most convenient.)

5. *(The Effect of Missing Data)*

   In this exercise we will recreate the example in the paper, *"Missing Data: Our View of the State of the Art"* (2002) by Schafer and Graham that can be downloaded from CourseWorks with this assignment. (You do not need to read the paper!) In particular we will use linear regression to emphasize that ignoring missing data can lead to very misleading results. This is why it is important to understand the **mechanism** of missing data and whether or not missing data is **ignorable** or not.

   The basic set up is as follows. Patients come in for a blood pressure measurement in Month 1 (X) and Month 2 (Y). The blood pressure measurements (X, Y) for each patient are drawn IID from a bivariate Normal distribution with $\mu_1 = \mu_2 = 125$ and $\sigma_1 = \sigma_2 = 25$ and correlation $\rho = 0.6$. There are $N = 50$ patients in the study.

   In this problem we will investigate the impact of missing data when we try to estimate Month 2 data from Month 1 data and vice versa. We will consider three different types of missing data.

   (i) Missing Completely At Random (MCAR): In this case, a randomly chosen $\alpha = 73\%$ fraction of the patients do not come for Month 2 measurement, and therefore the second data point, Y, is missing for each of these patients. (This particular value of α was chosen so that all three data-sets have approximately equal number of samples.)

   (ii) Missing At Random (MAR): In this case, any patient who had a reading $X \leq 140$ in Month 1 does not show up for his Month 2 reading, Y. In this case, the missing Y data is random but it depends on the X data from the first month.

   (iii) Missing **Not** At Random (MNAR): In this case, any patient who had a Month 2 Y reading ≤ 140 is assumed to be missing. In this case, the missing Y data is **not** random, and it depends on the Y variable itself.

   The MCAR-MAR-MNAR terminology is somewhat unfortunate. MCAR is self-explanatory: the event that a particular data-point is missing is independent of all other data. Missing data is MAR if the distribution of *missingness* depends on the observed data but not on non-observed, i.e. missing, data. Finally, missing data is MNAR if the distribution of missingness depends on the unobserved data.

(a) Do the following steps using the software of your choice.

Step 1. Generate $N = 50$ data points from the bivariate distribution.

Step 2. Regress Y on X with an intercept term. Compute the coefficient corresponding to X.

Step 3. Remove some data points for Y using the MCAR model. Regress Y on X with an intercept term. Compute the coefficient corresponding to X.

Step 4. Remove some of the data points using the MAR model. Regress Y on X with an intercept term. Compute the coefficient corresponding to X.

Step 5. Remove some of the data points using the MNAR model. Regress Y on X with an intercept term. Compute the coefficient corresponding to X.

Step 6. Plot the best fit line for all points in X for each of the four conditions above.

Step 7. Comment on the best fit lines for each of the regressions.

(b) Compute the average estimate for β_1, i.e. the coefficient of X, in each of steps 2 to 5, by repeating the above procedure $T = 100$ times. Comment on what you observe.

(c) Compare your results to those in Table 2 of Schafer and Graham.

The takeaway here is that we can obtain valid results when we ignore data that is MCAR. Problems can arise when we ignore missing data that is only MAR but not MCAR. (MCAR is a subset of MAR.) Even more severe problems can arise when we ignore missing data that is MNAR. In general for missing data that we believe is MCAR or (especially) MNAR we need to account for the missingness in our inference.

A standard modern approach (to account for MCAR and (especially) MNAR data) is to model the joint distribution of missingness and the complete data, and then integrate out the missing data to obtain an appropriate likelihood which is then used to estimate unknown parameters. An alternative approach is to use a Bayesian model to impute multiple values of the missing data and then apply standard techniques to the resulting multiple complete data-sets. The results then need to be combined together in an appropriate manner. (It is also possible to impute a single value for each missing data-point and obtain good point estimates of unknown parameters – if a good imputation method is used. However, the analysis with the resulting imputed data-set can underestimate uncertainty in the parameter estimates and therefore lead to poor confidence intervals, hypothesis tests etc. Why?)

We will not concern ourselves with handling MCAR or MNAR data in this course but it's important to be aware of these issues. A good dose of common-sense and domain specific knowledge should help you understand whether or not you can safely ignore missing data.

6. *(Bias-Variance Trade-Off in a Very Simple Example)*

Suppose the true model is given by $y(x) = ax$, i.e. a linear model with no noise, and we want to fit a model $h(x) = c$, i.e. just an intercept regression. We can estimate the model after seeing some (random) training data, $\mathcal{D}$. Let $h_D(x)$ denote the model estimator based on $\mathcal{D}$. Then the error of this estimator is given by

$$\mathbb{E}_{(x,y)}\mathbb{E}_{\mathcal{D}}\left(y - h_{\mathcal{D}}(x)\right)^2$$

where:

(i) the outer expectation $\mathbb{E}_{(x,y)}$ is over a new test sample, (x, y)

(ii) the inner expectation $\mathbb{E}_{\mathcal{D}}$ is over the training data, $\mathcal{D}$.

By completing the squares, it is easy to see that

$$\mathbb{E}_{(x,y)}\mathbb{E}_{\mathcal{D}}\left(y - h_{\mathcal{D}}(x)\right)^2 = \underbrace{\mathbb{E}_{(x,y)}\left(y - \mathbb{E}_{\mathcal{D}}h_{\mathcal{D}}(x)\right)^2}_{\text{bias}^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left(h_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}h_{\mathcal{D}}(x)\right)^2}_{\text{variance}}$$

(a) Given a data set $\mathcal{D} = \{(y_i = ax_i, x_i), i = 1, \ldots, N\}$, compute the least squares estimator of $h_{\mathcal{D}}(x)$.

(b) Suppose the training data $x_i \sim \mathcal{N}(\mu, \sigma^2)$ IID, and the test data $x \sim \mathcal{N}(\mu, \sigma^2)$.

(i) Compute $\mathbb{E}_{\mathcal{D}}[h_{\mathcal{D}}(x)]$ where $h_{\mathcal{D}}(x)$ denotes the least squares estimator from part (a).

(ii) Compute the bias$^2 = \mathbb{E}_{(x,y)}\left(y - \mathbb{E}_{\mathcal{D}}h_{\mathcal{D}}(x)\right)^2$

(iii) Compute the variance $= \mathbb{E}_{\mathcal{D}}\left(h_{\mathcal{D}}(x) - \mathbb{E}_{\mathcal{D}}h_{\mathcal{D}}(x)\right)^2$

(c) Consider a new estimator, $h_{\mathcal{D}} := \frac{\beta}{n}\sum_{i=1}^{n} y_i$. Compute the value of β that minimizes the expected squared error, i.e. the sum of bias^2 and variance.