Problem Set 1
Machine Learning
PPHA 30545
Due: Monday, January 25

We collect notation. For the regression model

$$Y_i = \beta_0 + X_{i,1}\beta_1 + \cdots + X_{i,p}\beta_p + \epsilon_i,\ i = 1, ..., n, \tag{1}$$

we have that n is the number of observations, p is the number of covariates, $E[\epsilon_i] = 0$ and $V(\epsilon_i) = \sigma^2$.

Each question is equally weighted. You will be graded on the correctness of your code, but also on the clarity and quality of your plots and discussions.

1. *This exercise aims at developing an understanding of the conservative nature of the Bonferroni correction, and at practicing the –important– skill of communicating statistical results and concepts using plots. The main objective of this exercise is to produce and discuss the plot required in 1.a-1.b. Recall that for large n the t-statistic corresponding to the null hypothesis $H_{0,j} : \beta_j = \beta_j^0$ is distributed with a Gaussian null distribution, specifically it is distributed as the normal random variable*

$$z_j \sim N(0, 1),\ j = 1, ..., p. \tag{2}$$

1.a. Generate the following figure: for test statistics with null distribution (2), plot the probability of false rejection of the joint null (that the p individual null distributions are the true distributions) at critical level $\alpha = 0.95$ against the value of p, using the Bonferroni correction.

1.b For different values of ρ, generate the test statistics according to

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_p \end{pmatrix} \sim N(0, \Sigma),\ \Sigma = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

Overlay the "probability of false rejection versus p" relationships for different values of ρ atop the plot for 1.a. In other words, all "probability of false rejection versus p" curves should be in the same plot.

1.c *bonus question*: what is the biggest value of ρ that you can use such that Σ remains a covariance matrix?

2. *This exercise aims at familiarizing yourself with the concepts of model selection as multiple hypothesis testing and out-of-sample fit criteria.*

a. Download the data set Hitters from the ISLR library –the R CRAN library complementing the course's textbook.[1]

b. Divide the dataset into a training set of n_{train} observations and a test set of n_{test} observations. Obviously, $n_{\text{train}} + n_{\text{test}} = n$, but you need to choose n_{train} and n_{test}. Explain the tradeoff between having a big training/small test versus small training/big test, and why the values you chose are reasonable. Some R functions useful for this question are presented in section 6.5 of your textbook.

c. Using the training data, select the model made of the 7 coefficients with the smallest p-values according to the regression fit of the full model.

d. Using the training data, select the best model made of 7 coefficients according to the forward stepwise selection procedure (p. 247 for hints). You do not need to code the procedure yourself –use an R package!– but explain what this procedure does, and how it is different from the one in c.

e. Using training data, select the best model made of 7 coefficients according to the best subset procedure (p. 244 for hints).

f. Compute the sample mean squared error in the test set for each method fitted in c, d and e, and collect the results in a table. Discuss.

g. *bonus question*: Repeat exercises c-f for different sizes of the subset of coefficients, and present your results in an extended table or plot.

h. *bonus question*: For selecting larger subsets with the best subset selection method, compare the performance of the package **leaps** with that of **bestsubset**.[2] Consider adding interactions.

i. *bonus question*: Can you suggest a more efficient way to split and use the data as training and testing sets?

[1] The dataset will also be made available on Canvas for those who wish to use Python.

[2] Can be installed with the following commands

```
library(devtools)
install_github(repo="ryantibs/best-subset", subdir="bestsubset")
```

You will also need to install the Gurobi solver.

3. *Preliminary analysis of the bikeshare.csv dataset. This dataset will come back later when we have more advanced methods in our toolkit. But we can already proceed to a preliminary analysis of the data. Find the data description sheet for this dataset in the assignment release page on canvas.*[3]

a. The data has been aggregated to *daily* counts to run the simple regression daylm. What are the in-sample sum-of-squared errors and R^2 for this regression?

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

c. A standardized residual for response Y and fitted value $\hat{Y}$ is $r_i = (Y_i - \hat{Y}_i)/\hat{\sigma}$. Calculate the standardized residuals for daylm. Now, we'll call the outlier p-value $P(Z > |r_i|)$ where $Z \sim N(0,1)$. In R, this is 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.

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

4. *bonus question*: Consider the drawing on the second to last slide of the deck of Lecture 2. Produce the equivalent drawing to illustrate the omitted variable bias phenomenon.

hint: Think of $Y_i = X_{i,1}\beta_1 + X_{i,2}\beta_2 + \epsilon_i$ as the long regression, $Y_i = X_{i,1}\beta_1 + \epsilon_i$ as the short regression, and produce a drawing that has the projection of Y on the span of X_1 and X_2, the projection of Y on the span of X_1, and one other projection.

[3] This data set was put together for pedagogical purposes by Matt Taddy, who graciously shared it.