STAD80 Analysis of Big Data **Assignment: 2**

Supervised Learning: Regression

Lecturer: Qiang Sun *Email: stad80.utsc@gmail.com*

You will have to submit '.rmd' R Markdown, '.tex' latex file (can be generated from R markdown) and '.pdf' file (can be generated from R Markdown). I should knit the '.rmd' file and compile '.tex' file without troubles. You can also use pythonnotebook + tex files for submission.

Question 1 (A Simple Linear Regression (30 points))**.** *Read in the* `housingprice.csv` *file using* `read.csv()` *function.*

(a) Rank the zipcodes by their average housing prices. What are the top 3 zipcodes whose average housing prices are most expensive? Create three boxplots of housing prices for these 3 zipcodes respectively.

(Hint: First convert the `zipcode` *column into factors. Then use* `tapply()` *and* `sort()` *functionals to compute the result.)*

(b) Visualize the relationship between `sqft_living` *and* `housing price` *by creating a scatter plot.*

The following questions continue from above questions. Load the training data `train.data.csv` *and testing data* `test.data.csv`*. We'll build our regression model on the training data and evaluate the model on the testing data.*

(c) Build a linear model on the training data using lm() by regessing the housing price on these variables: `bedrooms`*,* `bathrooms`*,* `sqft_living`*, and* `sqft_lot`*. What's the R^2 of the model on training data? What's the R^2 on testing data?*

(d) Add `zipcode` *in your linear model. What's the R^2 of the new model on the training data and testing data respectively?*

(e) The image below is Bill Gates' house. Load the file `fancyhouse.csv` *to obtain the features of the house. Guess the price of his house using your linear model. Do you think the predicted price is reasonable?*

Figure 1: Image fom Wikipedia Commons

(f) Suppose we have a linear regression problem with n training samples and d covariates. If $n > d + 1$, show that adding another covariate in the model never hurts R^2 over the training data.

(Hint: By definition, the ordinary least squares (OLS) estimator $\widehat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta} \in R^d} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$, where $\boldsymbol{Y}$ is the response and $\boldsymbol{X}$ is the design matrix. Denote the new design matrix with the additional covariate by $\boldsymbol{X}_1$. Then the OLS estimator for the new regression problem $\widehat{\boldsymbol{\beta}}_1 = \text{argmin}_{\boldsymbol{\beta} \in R^{d+1}} \|\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}\|_2^2$. Compare $\|\boldsymbol{Y} - \boldsymbol{X}_1\widehat{\boldsymbol{\beta}}_1\|_2^2$ and $\|\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2$).

Question 2 (Feature Engineering (20 points)). *Let's continue to improve the linear model we have. Instead of throwing only the raw data into the statistical model, we might want to use our intuition and domain expertise to extract more meaningful features from the raw data. This step is called feature engineering. Using meaningful features in the model is often crucial for successful data analysis.*

(a) Add another variable by multiplying the number of bedrooms by the number of bathrooms, which describes the combined benefit of having more bedrooms and bathrooms. Add this variable to the linear model we have in Question 6 (d). What's the R^2 of the new model on the training data and testing data respectively?

(Hint: You don't have to create a new column in the data frame. Try this trick in `lm()`*:* `lm( y ~ x1 + x2 + x1 * x2, data = your.data))`

(b) Using R^2 on the testing data as the metric for evaluating your model, propose another feature engineering that further improves the model you have in Question 2 (a).

(c) Polynomial regression is a general technique that allows you to add nonlinear features in your statistical model. Based on the model we have in Question 1 (d), add polynomial terms of the bedrooms and bathrooms variables of degrees 2 and 3 (no cross terms) in your model. Find out the R^2 of the new model on training data and testing data.

(Hint: You don't have to create a new data frame. Try this trick in `lm()`*.* `lm( y ~ poly(x1, degree) + x2 + x3, data = your.data))`

Question 3 (Wine Pricing (20 points)). *Wine pricing is a challenging job. Though wine is produced every year in a similar way, its price and quality varies between years. Since all the wines are meant to be aged, it is very hard to tell if wine will be good or not on the market in the future. Traditionally, wine companies solely rely on expert tasters to assess wine quality. However, in March 1990, Orley Ashenfelter, an economics professor in Princeton, claimed he could predict wine quality without actually tasting the wine; his method was surprisingly simple: just use ordinary linear regression! In fact, his method is so simple that Robert Parker, one of the most famous wine experts, once commented:*

"Arshenfelter is an absolute total sham".

We would see in this problem whether Ashenfelter is a sham or not through real data analytics.

Orley Arshenfelter

Robert Parker

In this study, we will use a dataset `wine.csv`*. We start with a description of the covariates of this dataset . For wine in each year from* 1952 *to* 1978*, we collect the logarithm of its price in the* $1990 - 1991$ *wine auction, average growing season temperature (*`AGST`*), winter rain amount, harvest rain amount, age of wine and etc.*

Part I. Preliminary Analysis

Load `wine.csv`*. Give four scatter plots of* `Price` *v.s.* `AGST`*,* `Price` *v.s.* `WinterRain`*,* `Price` *v.s.* `HarvestRain` *and* `Price` *v.s.* `Age`*.* `Price` *should be on the y-axis. Which variable do you think is most correlated with* `Price`*? Justify your observation by calculating the Pearson's correlation.*

Part II. Marginal Regression Analysis.

Fit a marginal regression model: $\texttt{Price} \sim \texttt{AGST}$*. Report the fitted coefficient values and* R^2*.*

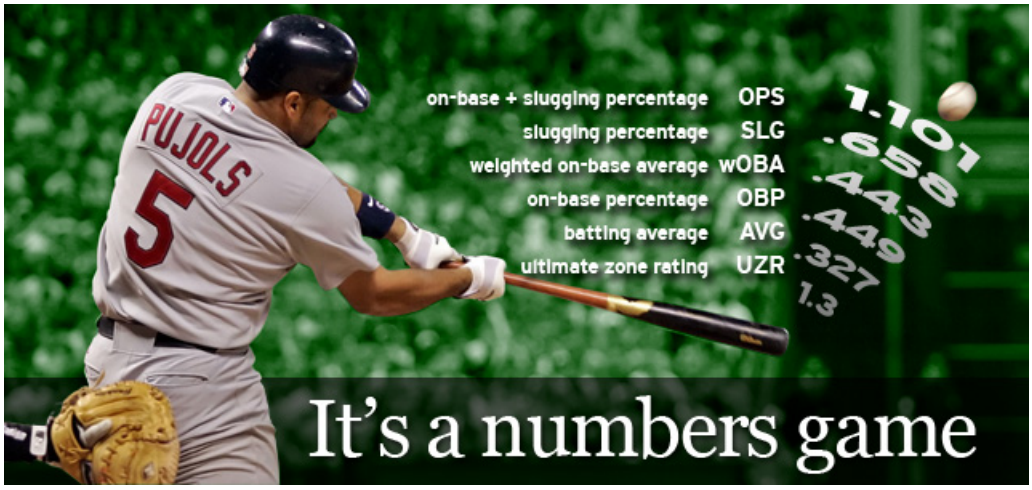
Part III. Multiple Regression Analysis.

Add `HarvestRain`*,* `Age`*,* `WinterRain` *and* `FrancePop` *to your model in Part II one by one. Report how* R^2 *on the training data changes as we add more and more covariates. For each model, also report your* R^2 *over the testing data provided in* `winetest.csv`*. Which model should we choose based on* R^2*? According to Prof. Ashenfelter's system in* 1990*, heavy rains in the winter followed by a hot summer improve wine quality, while rainfall before the harvest damages it. Is your model consistent with Prof. Ashenfelter's finding?*

The ending of the story is that regression beats Parker! Parker thought the wine in 1986 was "very good to sometimes exceptional", while Ashenfelter's model predicted that the wine in 1986 was mediocre and the wine produced in 1989 instead would be "the wine of the century". It turned out that the 1989 wine was sold for more than twice the price of 1986. Here we see the power of simple regression in wine price prediction; it can sometimes beat human experts.

Note: *More details about this story can be found in a fascinating article from 1990 New York Times:*

`http://www.nytimes.com/1990/03/04/us/wine-equation-puts-some-noses-out-of-joint.html`



Question 4 (Moneyball: The Analytics Edge in Sports (30 points))**.** *Baseball is one of the most symbolic sports in the US. Back into the 20th century, baseball teams relied on respected and experienced scouts to recruit players. However, as statistical analytics were introduced to this game, the rule changed. Michael Lewis's bestselling book* Moneyball: The Art of Winning an Unfair Game *tells the story of how Billy Beane, the general manager of the Oakland Athletics in 1998, effectively used statistical analytics to turn this losing team into a winning team with very stringent payroll budget. His great success was due to the discovery of significant features that create runs through statistical analysis.*

In particular, Billy found that on-base percentage (`OBP`*) and slugging percentage (*`SLG`*) have much better performance in measuring offensive capability than on batting average (*`BA`*), which has always been baseball's most famous and well-published statistic. In this problem, we would use the real data to verify Billy's claim, and see how regression analysis can reshape decision making procedures in baseball team management and lead to dramatic enhancement in team performance.*

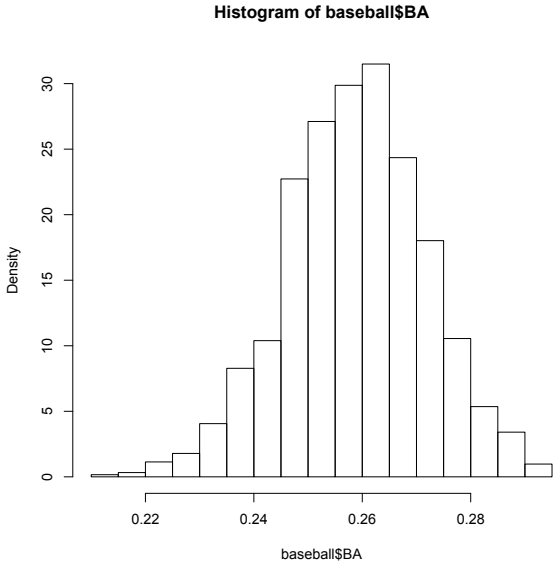
First of all, let's familiarize ourselves with some key terminologies.

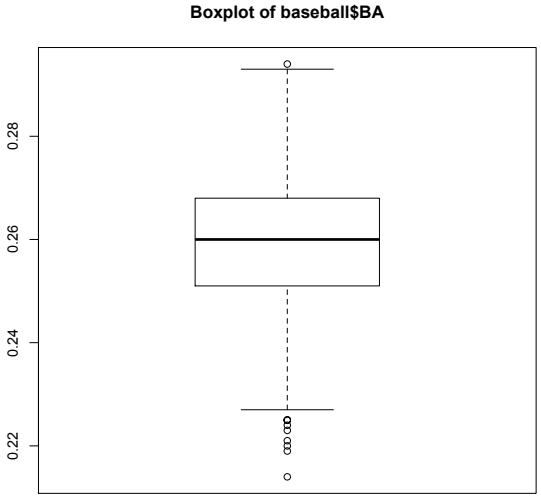
1. *Plate appearances: It is counted every time a player comes to bat regardless of the outcome of that time at the plate*
2. *At bat: It is counted only the times a player gets a hit or make an out.*
3. *Batting average (*`BA`*): It represents the percentage of at bats that result in hits for a particular baseball player.*
4. *On-base percentage (*`OBP`*). It is a measure of the number of times a player gets on base by hit, walk, or hit by pitch, expressed as a percentage of his total number of plate appearances.*
5. *Slugging percentage (*`SLG`*): It is calculated as total bases divided by at bats.*

In this problem, we would use the dataset baseball.csv *that has statistics of performance of all Major League Baseball (MLB) teams from 1962 to 2012. We list below meanings of important features in the dataset. You are encouraged to google the detailed information if you are not pretty sure about the terminologies here.*

1. `RS`*: runs scored*
2. `RA`*: runs allowed*
3. `W`*: the number of winning games*
4. `OBP`*: on-base percentage*
5. `SLG`*: slugging percentage*
6. `BA`*: batting average*
7. `Playoffs`*: whether making a playoff, 1 means yes and 0 means no*
8. `OOBP`*: opponent on-base percentage*
9. `OSLG`*: opponent slugging percentage*

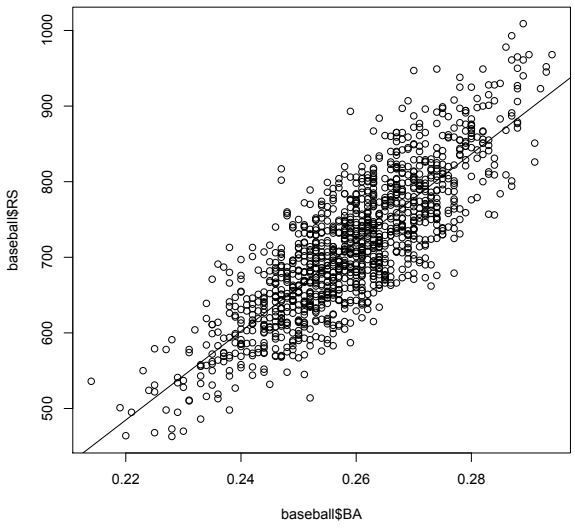
Part I. Preliminary Analysis. Load the dataset baseball.csv. *Plot histograms and boxplots for* `OBP`*,* `SLG`*,* `BA`*. Also report the mean and median for these three statistics. The purpose of doing this is to get an initial idea of how the data are distributed and detect potential outliers. Below we present the histogram and boxplot for* `BA` *for your reference.*

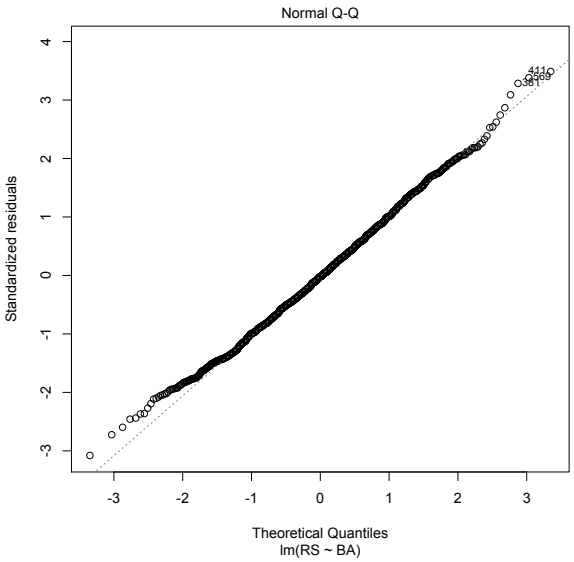




The mean and median of `BA` *are* 0.259 *and* 0.26*, meaning that the distribution is not skewed at all. You can also verify this from the boxplot and histogram. Analyze the other two quantities similarly.*

Part II. Marginal Regression Analysis. As characterized before, `BA`*,* `OBP` *and* `SLG` *are important quantities in determining the team performance, which is usually quantified by* `RS`*. Marginally regress* `RS` *on* `BA`*,* `OBP`*,* `SLG` *respectively and see how they are correlated. Give the scatter plot of the data and the fitted line. Report the coefficient values and* R^2*. Also give the QQ-plot of the fitted residuals, which is to verify that the residuals are not skewly distributed and the model is reasonable. Below we show the regression result using the model* $\texttt{RS} \sim \texttt{BA}$*. The intercept and slope are* -805.51 *and* 5864.84 *respectively, and* $R^2 = .6839$*.*





Traditionally, `BA` *is thought to be most responsible for* `RS` *since it removes contribution of lucky scoring like walk or hit by pitch and honestly reflects how well the batter hits the ball. Compare* R^2 *you obtained for* `BA`*,* `SLG` *and* `OBP`*. Is that consistent with the intuition?*

Part III. Multiple Regression Analysis.

The plots above only tell us part of the story. We now need to examine a regression output to see how `BA`*,* `SLG` *and* `OBP` *relate together in predicting our target variable* `RS`*. Fit the model* $\texttt{RS} \sim \texttt{BA} + \texttt{SLG} + \texttt{OBP}$*. Report the estimated coefficients for these covariates (along with their significance). Check the model by giving QQ plots of the residuals. Is the fitting result consistent with that in Part II, especially the fitted coefficient of* `BA`*? Summarize and justify your findings. Also fit the model* $\texttt{RS} \sim \texttt{BA} + \texttt{SLG}$*. Compare* R^2 *of the two models. Which model do you prefer?*

Part IV. Back to 2001 and Reshape the Baseball World.

Through the marginal regression analysis in Part II, we witness how `OBP`*,* `SLG` *and* `BA` *have a strong positive correlation with* `RS`*. However, of those three statistics, the traditional and most often used* `BA` *had the lowest correlation. Furthermore, when we examined how these variables work together in predicting* `RS` *in Part III, we noticed that* `OBP` *and* `SLG` *alone could do just as good a job predicting* `RS` *than a model that included* `BA`*. This suggests that team managers would best work on focusing on his team's* `OBP` *and* `SLG` *over* `BA` *to improve his team's run output for the season. Also in determining the payroll distribution, batters with higher* `OBP` *should be given higher salaries. This is the reason why Billy Beans could hire excellent players with low salaries since at that time* `OBP` *is not valued!*

Suppose we were data analytics for Oakland Athletics back in 2001, and our job was to predict how many games we would win in 2002. Based on historical players statistics, we estimated that in 2002 $\texttt{OBP} = .349$*,* $\texttt{SLG} = .430$*,* $\texttt{OOBP} = .307$ *and* $\texttt{OSLG} = .373$ *for Oakland Athletics. Add a column* $\texttt{RD} = \texttt{RS} - \texttt{RA}$*. Fit the models* $\texttt{W} \sim \texttt{RD}$*,* $\texttt{RS} \sim \texttt{OBP} + \texttt{SLG}$ *and* $\texttt{RA} \sim \texttt{OOBP} + \texttt{OSLG}$ *using the data before (not including)* 2002*, and try to predict how many games Oakland would win in 2002 by combining these three models. Check whether your prediction is accurate by looking at our dataset.*

Acknowledgement

The origin of the dataset `housingprice.csv` is from the Coursera open course Machine Learning Foundations: A Case Study Approach by Prof. Carlos Guestrin and Prof. Emily Fox. The open course also inspired the linear

regression part of this assignment.

We greatly appreciate the MIT online course *Analytics in Edge*, which provides interesting stories on power of statistical analytics in real applications and the relevant datasets.