**Introduction**

Welcome to the assignment on baseball analysis!

This assignment has two problems related to regression. The assignment has a maximum of 75 available points. Problem A is well defined in the sense that steps to be followed are given, while Problem B is open-ended and is exploratory in nature. The problems are presented next.

* **Problem A:**Multiple Linear Regression for Baseball data: [25 points]

Data for 1987 salary of hitters in major league baseball (in USA), along with various career statistics is given in file “baseball\_hitters\_assignment2.csv". The variables present in the csv file are listed in Table 1. Do the following:

1. Identify rows with missing data points and eliminate them. [3 points]
2. Fit a Simple Linear Regression (SLR) model between response variable y (1987 salary) and x7 (no of years in the major leagues), i.e. a model of the form: y=b+ax7. [10 points]
   1. Obtain estimates of a, b and R2,R2adj for the model.
   2. Obtain 95% confidence intervals for parameters a and b.
   3. Are both a, b statistically significant with α = 0.05 (5% significance level) i.e. test the hypothesis H0: a = 0, H1: a 6= 0. Similarly for b.
   4. Plot the residuals versus x7,¯y, sample number, and the Q-Q plot. From these plots comment on the adequacy of the model.
3. Now fit a Multiple Linear Regression (MLR) model with y as the dependent variable, and x7,x27(i.e. years, years-squared) as independent variables. Thus, the model is of the form: y=a+bx7+cx27. [10 points]
   1. Repeat steps (1) to (4) of part (2), namely estimate the three regression coefficients, obtain R2,R2adj, obtain 95% confidence intervals on the three parameters, perform statistical significance test for the parameters, generate various plots of residuals.
4. Based on your analysis in (2) and (3), which of the two models do you think is relatively better and why? [2 points]

* **Problem B:**Building a "good" model to predict salaries of baseball hitters: [50 points]

In this part, we want to explore the use of all the given variables, and maybe some of their transformations to predict salary. This is an open-ended problem. Some tips/suggestions are given below:

* Perform data preprocessing, namely removal of rows with NA (missing) values, and replacement of a categorical variable (league at end of 1986: N or A) with indicator variable.
* Divide data into training and test sets (say 25% as test data). The model building is done on the training set and the performance of the test set can be used to compare different models.
* Perform exploratory data analysis (EDA):
  1. Plot histogram of salary and of log salary. You will see that log salary is more symmetric than salary. As a rule of thumb, it will be easier to build a model for a response variable which is symmetrically distributed. Thus, you can build a model for log salary instead of salary.
  2. Plot salary (or log of salary) versus all other variables. You can add additional variables based on patterns in these plots. For example, plot of log of salary versus years in the league appears to show quadratic behaviour. Thus, you can add years-squared as another possible regressor in the list of regressors you have. You can do this for other regressors as well. You can even try annualizing some variables. For example, you can create a new variable as number of runs (career) divided by no. of years in major leagues. This probably is a much better indicator of player performance than the total number of runs (career). Thus, the number of potential regressors can be significantly more than the given set of 17 in Table 1.

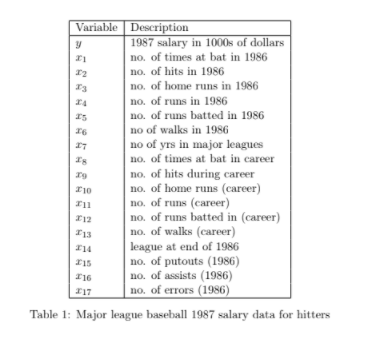


Table 1: Major league baseball 1987 salary data for hitters

* Based on EDA in above step, you can identify a set of promising variables. Build a multiple linear regression model using those variables.
* You can build several models as well and compare them using various metrics, such as R2,R2adj, residual plots behaviour, performance on test dataset, etc. Also, as far as possible, we would like all coefficients in the model to be statistically significant.
* Models can also be built in a “brute-force" manner once you have identified the set of all possible regressors. You can first build a model which has all the regressors. Now, you eliminate a regressor which is most insignificant (has highest p-value). Build a new model with remaining regressors. Again identify and remove the most insignificant regressor from this new model. Repeat these steps till you have a model which has all significant regressors. Note that every time you remove a regressor, you have to build the model again (i.e. identify coefficients again) for the remaining regressors.
* q-q plots can also be used to identify potential outliers. You can explore removal of one or a few outliers and rebuilding the models. You will see that R2, R2adj will increase.
* In the submission, show all the things you tried (as part of your jupyter notebook), and then report the best model as well.

**Dataset**

You can find the dataset below.

**Dataset**

**Download**

**Paper**

You can read more about the regression in baseball salary in the below paper.

**Paper**

**Download**

All the best for the assignment and the course ahead!

**Report an error**

**Instructions for Submission**

1. File Name: File name should contain assignment number and your email id, e.g., you  are submitting assignment 2 and your email id is xyz@abc.com, please name  submission file as assignment2\_xyz

2. Write your name and email id (full) in the very first block of the code, i.e., the topmost block in your jupyter notebook.

3. File format: After naming the file and executing the code, please convert it to an HTML file.  To convert a jupyter notebook to an HTML file, click on File on notebook, then  Download as and then click on HTML(.html).

4. For jupyter notebook on Colab follow:

* Download your file as .ipynb from Colab.
* Reupload the .ipynb file that you just downloaded to session storage.
* Your file is now uploaded to the current directory. Click on the file icon appearing on the left of the Colab screen and check if your file has been uploaded.
* Suppose the file name is “file\_name.ipynb”, run this command:-  “%% shell jupyter nbconvert --to html file\_name.ipynb” to convert it into .html format.
* Download the .html file.
* For more details refer to the following link which explains the procedure  nicely: https://stackoverflow.com/questions/53460051/convert-ipynb-notebook-to Html-in-google-colab

5. Before converting to HTML make sure all lines of codes are executed and expanded.  This way we will not have to run your code.

6. Before running a model or performing any analysis, please make sure to provide info about the dataset. We want to see the variables and their data types used in the model. For example, if you are going to fit say a linear regression model on a dataset df, simply write df.info() before executing your model.

7. And lastly, please provide all your suggestions and recommendations in one block at the bottom of your code, i.e., it should be your last block of code.

# Submission

This is an individual assignment.  For submissions obtained within one week after the deadline, there will be a 30% penalty. Submissions beyond one week after the deadline will not be accepted.

You have to submit everything in one Jupyter notebook.

Please make sure that you are not changing any of the file names that have been provided for download. The code that you are submitting should run at our end on the same files without any modification of the code.

**Regression**

**Submission details**

**Please Note**: This is an individual assignment. For submissions obtained within one week after the deadline, there will be a 30% penalty. Submissions beyond one week after the deadline will not be accepted.

You need to submit Jupyter notebook.

We strongly recommend you to submit atleast 30 minutes before your deadline

**Module Deadline**

**17 Apr '22**

**11:59 PM (IST)**

Deadline extension can not be requested as due date has crossed.

Upload your file & Save as draft

**Upload**

#### **Note: ONLY Feedback NO Grading post 24 Apr '22, 11:59 PM (IST)**