The University of Kansas

Department of Economics

Final Project **Spring 2020**
Econ 526 - Introduction to Econometrics **Instructor: Caio Vigo Pereira**

The file *mlb1_final_project.RData* (also available in *csv*, *xlsx* and *dta* formats) contains a dataset with a random sample with salary information and career statistics for players in the Major League Baseball (MLB). The dataset consists of the following variables (variable's name and description):

```
salary      1993 season salary measured in dollars
teamsal     team payroll measured in dollars
years       years in major leagues
games       career games played
atbats      career at bats
runs        career runs scored
hits        career hits
doubles     career doubles
triples     career triples
hruns       career home runs
hispan      =1 if hispanic
yrsallst    years as all-star
pcinc       city per capita income
```

Analyze your data, run the OLS regressions and answer the questions below.

1. Print out the **descriptive statistics** of your dataset. (*in R, use 'stargazer' command*)
   (a) What is the sample size?
   (b) What is the maximum number of years a player has been playing in MLB?
   (c) What was the minimum salary of a MLB player?
   (d) What is the (sample) average of the team payroll?

2. Plot the histogram of *salary* and *games* using *breaks* or *bins = 30*. Don't forget to add a title and label your axes.

3. (**Model 1**) Consider the following econometric model:

$$\text{salary} = \beta_0 + \beta_1 \text{games} + u \tag{1}$$

   Run this regression and print out the **output of your regression** (*in R, use 'stargazer' command*).

4. Write the **OLS regression function** with the estimates for the parameters from model (1) above and the standard errors under them.

5. Make a scatter plot with *games* in the horizontal axis and *salary* in the vertical axis. Plot the SRF in green with a 90% confidence interval. Don't forget to add a title and label your axes.

6. Based on the graph above, what characteristic of the errors in the population do you believe might be showing in this sample? Plot the diagnostic residual plots.

7. (**Model 2**) Consider the following econometric model:

$$\text{salary} = \beta_0 + \beta_1 \text{games} + \beta_2 \text{pcinc} + \beta_3 \text{teamsal} + \beta_4 \text{yrsallst} + u \tag{2}$$

   Run this regression and print out the **output of your regression** (*in R, use 'stargazer' command*).

8. Based on your regression from model (2) above, what is the estimated effect in your dependent variable for a player who has two more years as all-star, holding number of games, city per capita income and team payroll constant?

9. Based on your regression from model (2) above, what percentage of the variation in salary is explained by *games*, *pcinc*, *teamsal* and *yrsallst*?

10. For the regression from model 2, print out the 99% confidence interval. Explain which **independent variable(s)** are(is) statistically significant based on this confidence interval.

11. (**Model 3**) Consider the following econometric model:

$$\log(\text{salary}) = \beta_0 + \beta_1 \text{hruns} + \beta_2 \text{pcinc} + u \tag{3}$$

   Run this regression and print out the **output of your regression** (*in R, use 'stargazer' command*).

12. Based on your regression from model (3) above, what is the estimated effect in your dependent variable for a player with five more home runs holding city per capita income constant?

13. Based on your regression from model (3) above, which independent variable(s) is(are) statistically significant at 5% significance level? What about 1% significance level?