

Stat 431, Fall 2003  
Second Project (“Simple” and/or polynomial regression)

Use this page as the **cover page** for your project. Staple it to additional pages with answers and JMP output as necessary.

You may work individually or collaboratively in groups of up to 3 students per group. Each “collaborative” of 1 to 3 students should hand in one project write-up.

THE ASSIGNMENT WILL BE DUE **Nov. 25**.

(Different groups should work independently.)

List names here:

**NAMES:**

1. \_\_\_\_\_

2. \_\_\_\_\_

3. \_\_\_\_\_

The data for this project is located on the class website:  
[www-stat.wharton.upenn.edu/~lzhao](http://www-stat.wharton.upenn.edu/~lzhao) .

## Description of the Data

The data for this project contains information about major league baseball players (except pitchers) who were on major league team rosters for the full major league season in both 1986 and 1987. Included is information about their salary in 1987, their performance in 1986, and their career performance up through 1986. (You don't need to understand anything about baseball to do a good job on this assignment, though such knowledge might make the assignment more fun.) All such players whose 1987 salaries were publicly reported are included.

This data was gathered to see whether standard performance indicators over previous years could be reliably used to predict a player's current salary. If so, then this information could be used in that way in salary negotiations and arbitration.

You will be expected to use JMP to answer the following questions. These will help you develop a simple linear regression prediction equation for salary (or for Logsalary), and then to describe how well that equation works.

The relevant variables are:

**salary** = player's salary in 1987; **LN(salary)** = Natural log of **salary**; **Years** = years played in the major leagues prior to 1987;

**runs'86** = runs scored in 1986; **hr'86** = home runs hit in 1986; **rbi'86** = runs batted in during 1986; **hits'86** = base hits during 1986; **ab'86** = times at bat during 1986; **walks'86** = number of bases-on-balls during 1986;

**runs,career** = runs scored during the player's entire major league career; similarly for other performance variables;

**runscr/yr** = career runs scored divided by years played; similarly for other performance variables;

And three variables relating to fielding performance in 1986 – **put outs, assists, and errors.**

**Goal of the Project:** The end goal of this project is to find a *single* x-variable to use in a linear or polynomial regression to predict salary (or log salary). You want to choose your x-variable so as to produce the best possible overall prediction. You will then use your prediction equation to predict the 1987 salary of Pete O'Brien (player #264 in the data). [Later in the semester we will re-analyze this data via a multiple regression, and will then be able to use more than a single x-variable in our equation.]

One of the first things you need to do in analyzing this data is to decide whether to try to predict salary or its logarithm. (Or perhaps some other easy to use transformation of salary.) Bear in mind that you will eventually be using a regression technique designed for situations where salary is well modeled by a linear or polynomial equation of one of the other variables and also has normal and homoscedastic residuals.

1. Make histograms of both salary and log salary. (Print out and submit these plots along with others requested in the following questions.) Do you see any potential outliers on these plots that seem to need further investigation? Do either of these plots follow approximately a normal distribution? (The standard assumptions for regression analyses don't require that they do, only that their residuals be normal. Experience has shown however that in most situations data which is itself closer to being normal stands a better chance of also having normal residuals.)

Now choose a few of the possible explanatory variables as possible x-variables and make scatterplots of salary (=y) versus x and of Log(salary) as y versus those same x variables. Draw the simple regression lines and examine the pattern of relationship and residuals. [There are very many possible x-variables to choose from. Unless you find differently, I suggest beginning here with some or all of **Years; hits, career; runs, career; and runscr/yr.**]

2. Which choice of x and y-variables seems to come closest to satisfying the standard assumptions in the simple regression plots you have made? Choose that **y-variable** as the target of your analysis for future questions (unless further analyses lead you to reexamine your choice and change your mind). Submit at least one pair of scatterplots that exhibit why you chose your y-variable, and briefly explain your reasoning.

3. Do any of these plots indicate that you may want to also transform one/some x-variable(s)? If so, you may want to use that/those transformed variable(s) in place of the original one(s) in the following questions. (If so, submit the printout of that plot and explain.)

4. Which single x-variable or transform thereof is the best **linear** predictor of your y-variable? Submit table(s) or plot(s) that support your choice.

Carry out a simple regression analysis of your chosen x and y variables. Submit the relevant computer printout(s) and answer the following questions for your chosen analysis:

5. What is the regression equation?
6. How good a (linear) predictor is your chosen  $x$ ? - i.e., what is R-squared, and what does that mean?
7. Do you see any special influential points or potential outliers on your scatterplot? If so how have they affected your analysis?
8. Do the standard assumptions for linear regression analysis appear to be reasonably well satisfied? If not, why not? And what are the consequences? (And, can you think of any other way to analyze your chosen  $x$ - $y$  pair of variables that would avoid or alleviate the difficulty?)