AAEC 4804/5804G, STAT 4804: Fundamentals of Econometrics

Your Name Here

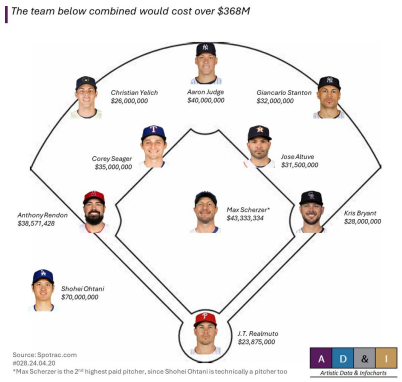
Spring 2025 – Homework #4

Instructions

This homework is intended to help you review the material covered in Lecture 5. This week we have a greater emphasis on real world application.

You are strongly encouraged to work with your classmates, but you must submit your own Solutions.

Question 1: Major League Baseball (MLB) Data



Gabe is an avid baseball fan. He is recently saw an article of the breakdown of salaries in the MLB. As he gleaned through the article, he noticed a bit of variability in the salary of the players and he wondered if this might be due to the position of the player. As you might notice from the image below, he might be on to something.

Gabe decided to do a little research and found a dataset (`mlb1` from the `wooldridge` package) on the salaries of players in the MLB, their position, and their performance. He decided to run a regression of the salary of the player on their position and performance. Gabe is interested in the following model:

$$\log(salary) = \beta_0 + beta_1 years + \beta_2 gamesyr + \beta_2 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop + \beta_{13} outfield + u$$

here, *catcher* will serve as the base category for the position of the player.

(i.) Run and report the model he suggested and interpret the results on the coefficients *year*, *gamesyr*, *shrtstop*, and *outfield*. **In cases where these coefficients are statistically significant, discuss the main takeaway from this finding. Also, ensure that your interpretations are in terms of salary and not in terms of the log of salary.**

(ii.) Gabe's two favorite players are both outfielders. He is curious to know if, controlling for other factors, the outfield position and first base players earn the same salary. Set up, conduct, and interpret the results of a hypothesis test to determine if this is true.

(ii.) Raisa, Gabe's friend, has just watched her first baseball game and now thinks she is an expert on the game. She thinks that the position of the player is not important in determining the salary of the player. She wants to test this hypothesis. Set up, conduct, and interpret the results of a hypothesis test to determine if this is true.

Question 2: Beauty is in the Eye of the Employer?

Katelyn and Michele are bitter rivals who are both interested in the same job. They are both very qualified for the job, but they are also very different. Michele has won a pageants and was also a Southern Belle. Suffice to say, she is an attractive person. Katelyn, on the other hand, is at best, average. Michele is wondering if her attractiveness will help her get the job, while Michele is wondering if her lack of attractiveness will hurt her chances of getting the job.

She recruited Kealyn and Huyoung, who have both taken this econometrics class, before to help her out. They came across a few papers, chief among them, the paper by Hamermesh and Biddle (1994) that found that attractive people earn more than unattractive people and were more likely to be employed and promoted than unattractive people. A modified version of their data is available in the `beauty` dataset in the `wooldridge` package.

(i) Using the `summarize()` function, calculate and report separately, the proportion of men and women that are classified as being above- and below-average looking. **Report these values in percentages. Ensure that you explicitly indicate the Gender of the respondents in your table.**

(ii) Kealyn and Huyoung now proceed by estimating the following regression:

$$\log(wage) = \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

(a) Katelyn argues that the model will suffer from perfect multicollinearity (or the dummy variable trap in this case). Explain why she is wrong and should consider taking this class next year.

(b) Run the regression separately for males and females and report the results using the `stargazer()` function with appropriately labelled models. Give the **precise** interpretations of the coefficients on the **independent variables in terms of wage**.

(iii) Michele's friends decided to conduct the hypothesis test: $H_0 : \beta_1 = 0$ against $H_1 : \beta_{belavg} < 0$. Explain in words what they are testing and conclude appropriately given your results.

Hint: If you recall, your regression model reports the `p-value` from the two-sided alternative hypothesis ($H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$). You could divide that pvalue by 2, or simply pass your t-stat and `df` to the `pt()` function and compute the one-sided value yourself. I would like you to do the latter.

(iv) Michele has seen the results above but wants to be sure that the results will stand up to scrutiny. She knows that other important variables have been omitted from the model and could bias our results. She suggested that you add the following: `educ`, `exper`, `expersq`, `union`, `goodhlth`, `black`, `married`, `south`, `bigcity`, `smllcity`, and `service` to the model.

- Estimate the models implied above and combine them with those prior into a single `stargazer` table. Present your four models ordered by gender. **Again, ensure your models are properly labelled. Use the `omit.stat = c("f", "ser")` argument to keep only the number of obs., R^2, and $\bar{R}^2$**
- Did adding the controls change the coefficients on `belavg` and `abvavg`? If so, how?

Note: Although you will be running the regression with the full set of results, I would like you to restrict your display to only the coefficients in original model. As we did in class, you can simply add a line item to your `stargazer` function to let the readers know that the first model does not have additional controls, but these do.

(v) Michele is now closer to being convinced that look do not truly matter in explaining wages. Although she might be deflated by your findings, she asks that you conduct a Chow test, at the 5% significance level, to test for equality in the slopes of the male and female regressions in **part (iv)**. To be explicit, your null is that the slopes of the `male` and `female` equations are the same/equal.

Hint: Use the SSR from the two separate regressions to compute SSR_u. The SSR_{res} will come from the model Michele proposed in part (iv) with the dummy variable `female` added.