Keynes Le
Econ 140: Econometrics
Ryan Edwards
6 December 2024

To what extent do race and identity influence Major League Baseball player salaries, beyond differences in ability?

Motivated by our shared interest in sports and systematic inequalities, we explored MLB data from the 1992 season to see whether racial disparities were prevalent in player salaries, and how performance metrics influenced these outcomes, We hypothesized that black and Hispanic players would earn less on average than white players, with these gaps narrowing after the inclusion of performance metrics such as hits and home runs as control variables.

The data sourced from the *New York Times* (April 11, 1993) and the *Baseball Encyclopedia* (9th Edition), contains 353 observations and 47 variables for statistics in the 1992 MLB season. Key variables included log-transformed player salaries, city demographics (percentage of Black, Hispanic, and white residents, and number of hits and home runs. The data was collected by G. Mark Holmes for a term project at MSU and supplemented with city population figures from the Statistical Abstract of the United States. Players whose race or ethnicity could not be determined were excluded.

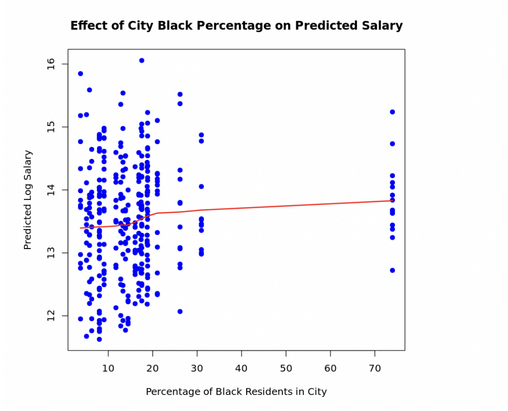
We ran one regression with log salary as the outcome variable and city racial demographics as treatment variables without controls and another regression with controls. Without controls, the percentage of Black residents showed a small but significant positive association with salary, while the Hispanic percentage was insignificant. This indicated that performance, rather than race, or identity, is the primary determinant of salaries. However, causality is limited by the observational nature of the data, as omitted variables such as team revenue and negotiation skills could skew our results. An RCT, which is normally ideal for establishing causal relationships, is not possible in this context.

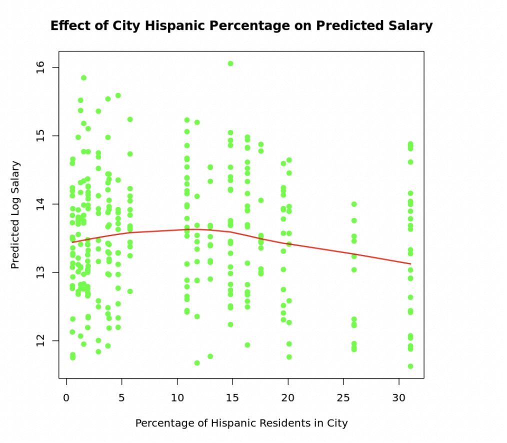
Our findings partially support our hypothesis. While Black and Hispanic players did earn less on average when controlling for performance, the percentage of Black residents in a city played an indirect role in influencing salary. The inclusion of hits and home runs as controls did however diminish explanatory gaps, which confirmed that performance is the strongest determinant of salary. Nonetheless, understanding limitations, such as the historical nature of the dataset (1992) and omitted variables as well as future research using more recent data is necessary to fully understand systemic inequalities and structural dynamics in MLD salaries.

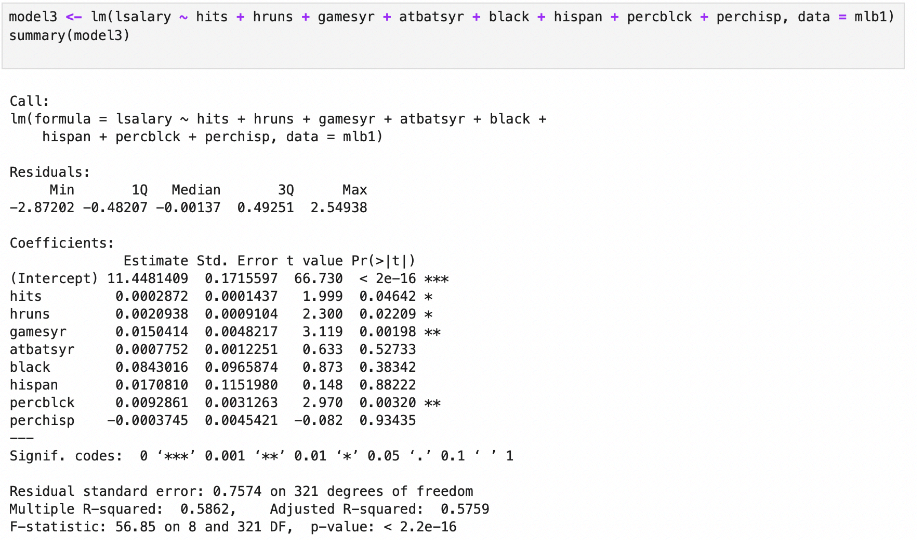
Works Cited, Graphs, and Regression Data

Wooldridge: 115 data sets from "Introductory econometrics. (n.d.).

https://cran.r-project.org/web/packages/wooldridge/wooldridge.pdf







```
model3 <- lm(lsalary ~ hits + hruns + gamesyr + atbatsyr + black + hispan + percblck + perchisp, data = mlb1)
summary(model3)
```

```
Call:
lm(formula = lsalary ~ hits + hruns + gamesyr + atbatsyr + black +
    hispan + percblck + perchisp, data = mlb1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.87202 -0.48207 -0.00137  0.49251  2.54938 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 11.4481409  0.1715597  66.730  < 2e-16 ***
hits         0.0002872  0.0001437   1.999  0.04642 *  
hruns        0.0020938  0.0009104   2.300  0.02209 *  
gamesyr      0.0150414  0.0048217   3.119  0.00198 ** 
atbatsyr     0.0007752  0.0012251   0.633  0.52733    
black        0.0843016  0.0965874   0.873  0.38342    
hispan       0.0170810  0.1151980   0.148  0.88222    
percblck     0.0092861  0.0031263   2.970  0.00320 ** 
perchisp    -0.0003745  0.0045421  -0.082  0.93435    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7574 on 321 degrees of freedom
Multiple R-squared:  0.5862,	Adjusted R-squared:  0.5759 
F-statistic: 56.85 on 8 and 321 DF,  p-value: < 2.2e-16
```