# Answering Questions of the CPS85 Dataset

**Sam Buechler // IMT 572 A**

## DATASET ANALYSIS AND RESEARCH QUESTION

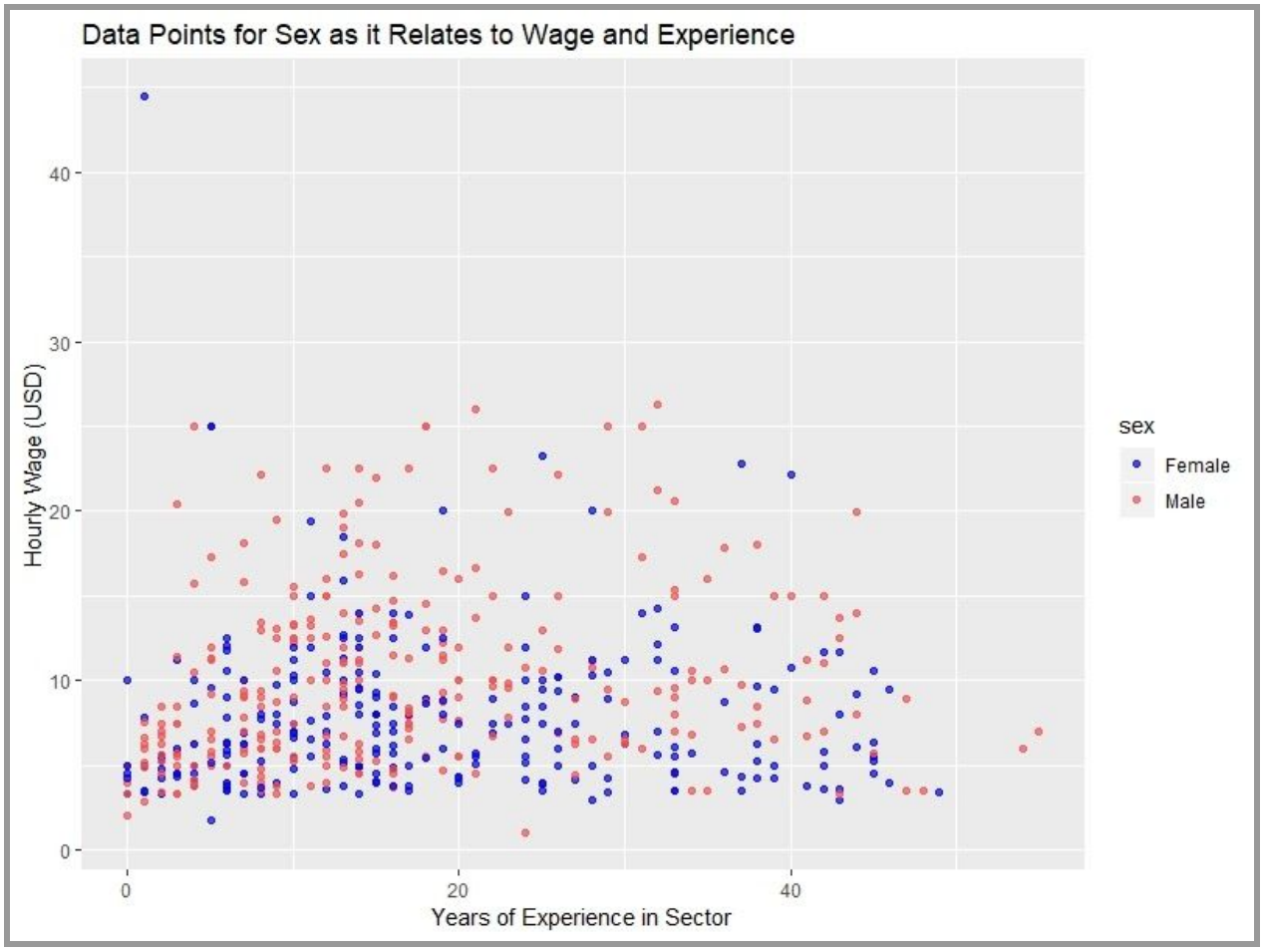
The dataset that I choose to examine was mosaicData's CPS85. I chose this set of data because the information that it represented aligns well with my interest in wage inequality and unionization of industries. CPS85 has 11 columns (fig. 1) and 534 rows of data; one row of data represents a working individual's individual and work identity factors.

| ***Fig. 1****- Description of columns & values in CPS85 dataset* | |
|---|---|
| **CONTINUOUS VARIABLES** | |
| **wage** | Describes the hourly wage of the individual in USD. |
| **educ** | Describes the number of years of formal education for the individual. |
| **exper** | Describes the years of experience an individual has in the "sector" column |
| **age** | Describes the individual's age at time of data collection. |
| **CATEGORICAL VARIABLES** | |
| **race** | Marked by White (W) or Non-White (NW), referring to an individual's race. |
| **sex** | Marked by Male (M) or Female (F), describing an individual's sex. |
| **hispanic** | Marked by Non-Hispanic (NH) or Hispanic (H), referring to whether the individual is hispanic or not. |
| **south** | Marked by South (S) or Not-South (NS), what these refer to is unclear. |
| **married** | Marked by "Married" or "Single," describing an individual's marital status. |
| **union** | Marked by "Not" or "Union," that describes whether the individual's wage is controlled by union bargaining. |
| **sector** | Marked by "clerical," "const," "manag," manuf," "other," "prof," "sales," and "service," describing which sector the individual is working in at time of data collection. |

While I began examining the data with an interest in how Union association affected wages and which sectors saw more union associations than others, I was ultimately drawn to

examining if common pay gaps across marginalizations played into the wages throughout the data set. I ended up with the research question: *What effect does race, sex, and experience have on wage?* I, again, choose to utilize these specific variables because they pertain to current research that interests me and the variables themselves showed more concrete relationships to each other than others in the data set. With wages being the outcome, and experience being the predictor, race and sex are both defined as confounders - there are several societal implications on the amount of experience someone has within a field (women and people of color haven't been able to work in certain sectors for as long as men have and even once they legally were able to, societal pressures and bias often prevented them from joining), and as we'll see through the coefficients, each of these variables have an effect on wages as well.

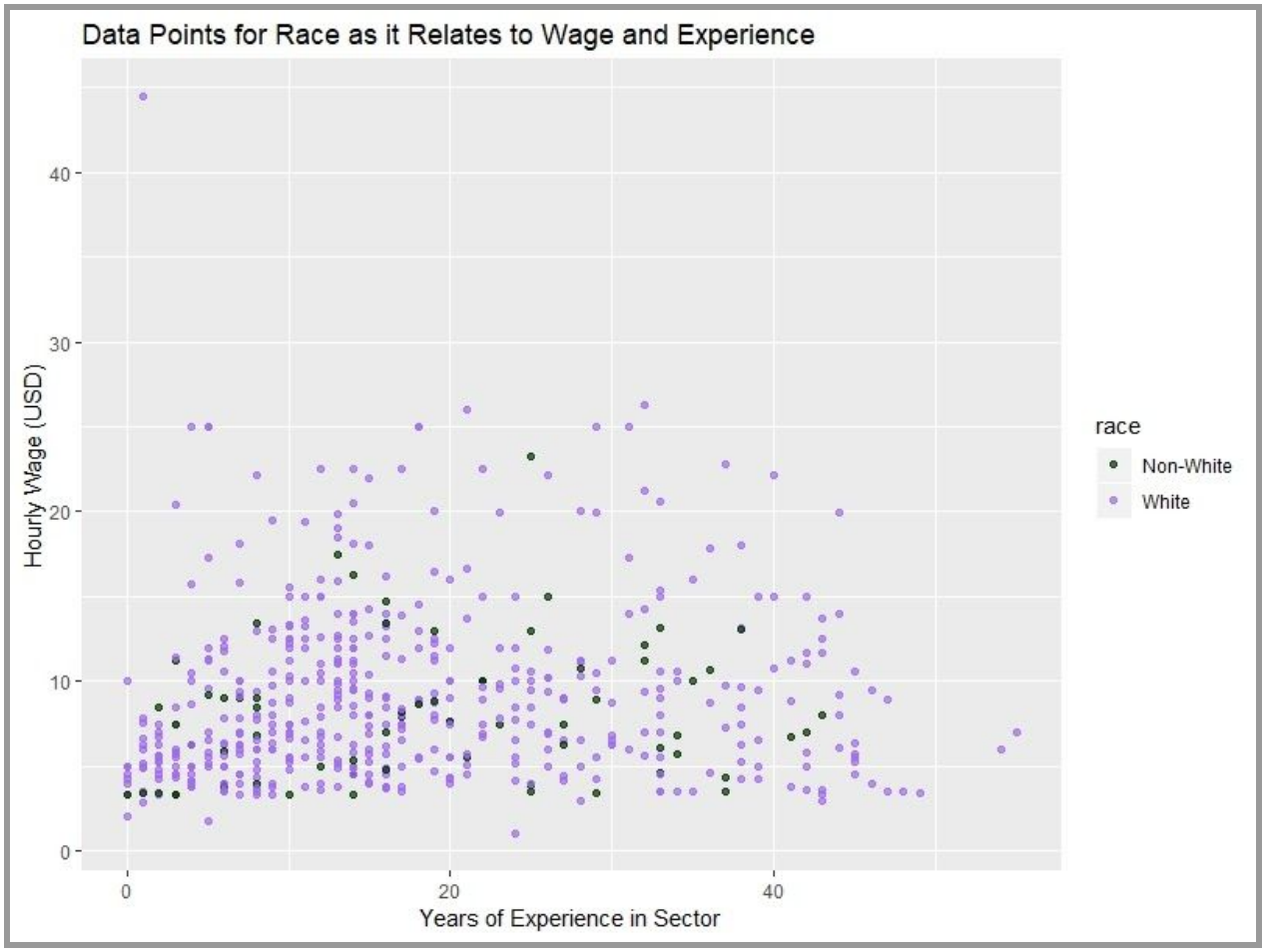
***Fig.2*** *- Data points for sex as it relates to wage and experience*



METHODOLOGY

After finally settling on my research question, I choose to answer the question through a multivariate linear regression. I choose this method because I was more interested in which factors contributed to changes in wages as well as the way in which they affected those wages. Equally, when running an initial plot on both sex and race as it pertained to wages and experience (fig. 2 & 3) *without* a regression line, a linear relationship was clearly evident as described during our regression lectures. These plots also show some other important relationships that I'll discuss when reviewing the weaknesses of my method.

***Fig.3*** - *Data points for race as it relates to wage and experience*



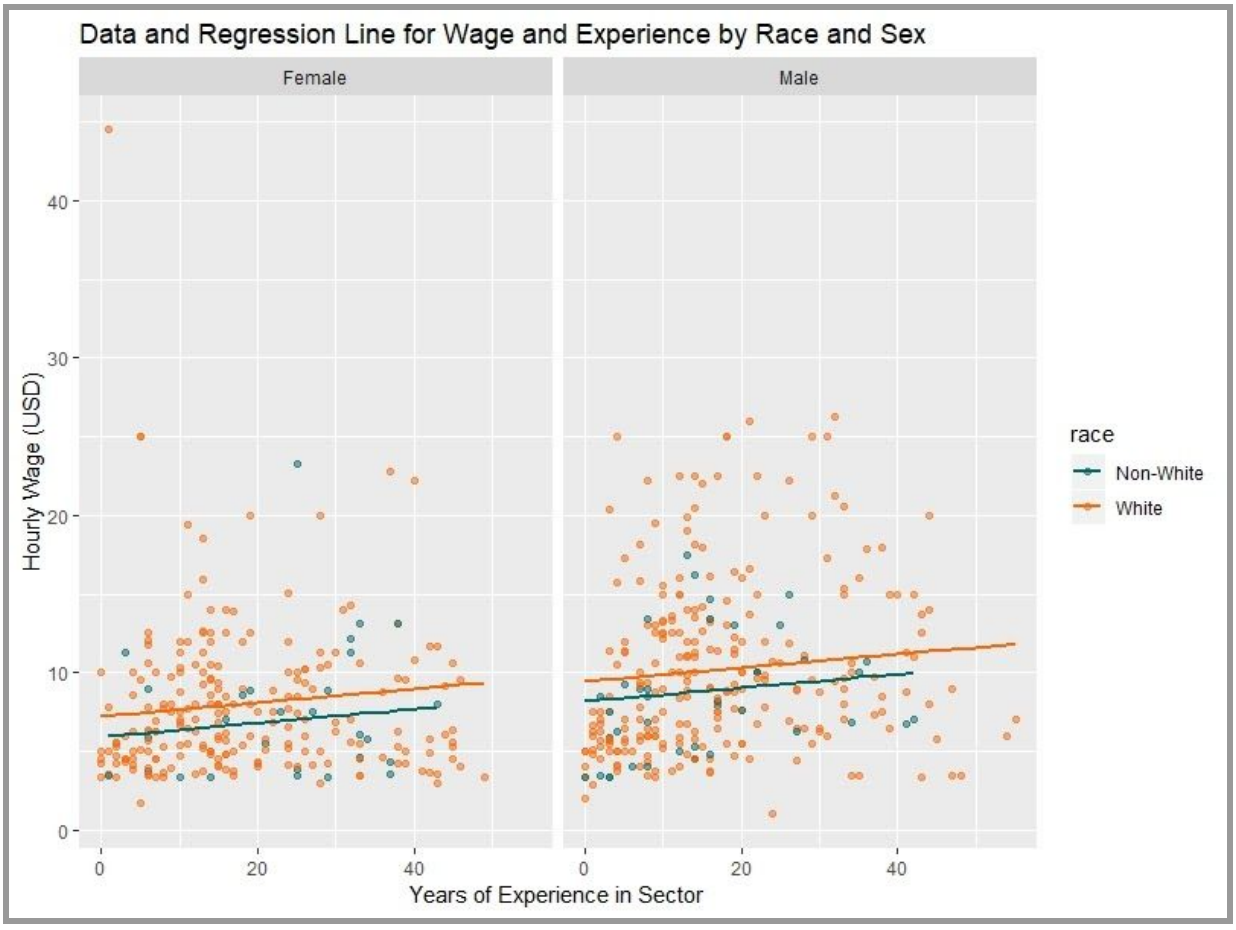
MULTIVARIATE REGRESSION OUTPUT

**Full Multivariate Regression:** $wages = \beta_0 + \beta_1 experience + \beta_2 race + \beta_3 sex + \varepsilon$

The intercept for this regression pulled in *sexF* and *raceNW* into its coefficient (5.93926) leaving us with remaining coefficients *exper* (0.04388), *sexM* (2.22396) and *raceW* (1.25686). The p-values of all coefficients are statistically relevant (Intercept: 1.16e-14, *exper*: 0.0127 *sexM*: 4.60e-07, *raceW:* 0.550). All of the Std. Errors prove that the coefficients are certain, (Intercept: 10.74743, *exper*: 0.01755, *sexM*: 0.43556, *raceW:* 0.65365). Based on this output, we can infer the following description:

> A non-white female worker is expected to earn an hourly wage of **$5.94**, a non-white male worker is expected to earn an hourly wage of ($5.94 + $2.22) = **$8.16**, a white female worker is expected to earn an hourly wage of ($5.94 + $1.26) = **$7.20**, and a white man is expected to earn an hourly wage of ($5.94 + $2.22 + $1.26) = **$9.42**. Each additional year of experience is expected to raise that rate by **$0.04**.

***Fig.4****- Data and regression line for wage and experience by race and sex*



The full multivariate regression is then plotted (fig. 4) utilizing facets and color by the categorical variables utilized in the regression. At this time I was unsure if one or the other would be preferred in terms of faceting v. coloring especially since their visually wasn't much difference between the two options.

WEAKNESSES: SCIENTIFIC, MORAL, AND SOCIETAL IMPLICATIONS

When reviewing these regressions there's a noticeable correlation of inequities between race, gender, and their intersection as it relates to wages and experience level. While the regression showed several of the results that I wanted, there were several weaknesses to my analysis. The first of those being the presence of sample population bias - the sample population was pretty evenly distributed across sex, but there was a much higher percentage of white participants than non-white participants which skews the data, and probably created the (barely) statistically insignificant p-value in the *raceW* coefficient.

There's further issue here with omitted variable bias as well - my exclusion of the *hispanic* variable (which was omitted because the percentage of non-hispanic participants was drastically smaller than non-white) definitely affects the results of this regression. Scientifically, the implications of this are the same as that of any incorrect or biased research - it creates a precedent that could take ages to undo because of the nature of scientific research. Of course, given the presence of a p-value (barely) higher than 0.05, this data wouldn't get very far regardless.

The moral and societal implications of this are vast, if incorrect data were presented on this topic, it could invariably discredit the struggles of certain communities and could potentially give motive for a variety of legal and societal changes. Thankfully, this research question has

been explored multiple times with greater datasets and, instead of discrediting entire communities, has been used to bring light to an injustice (although, unfortunately, it hasn't really brought about legal and societal changes, but the expectation is that it could).