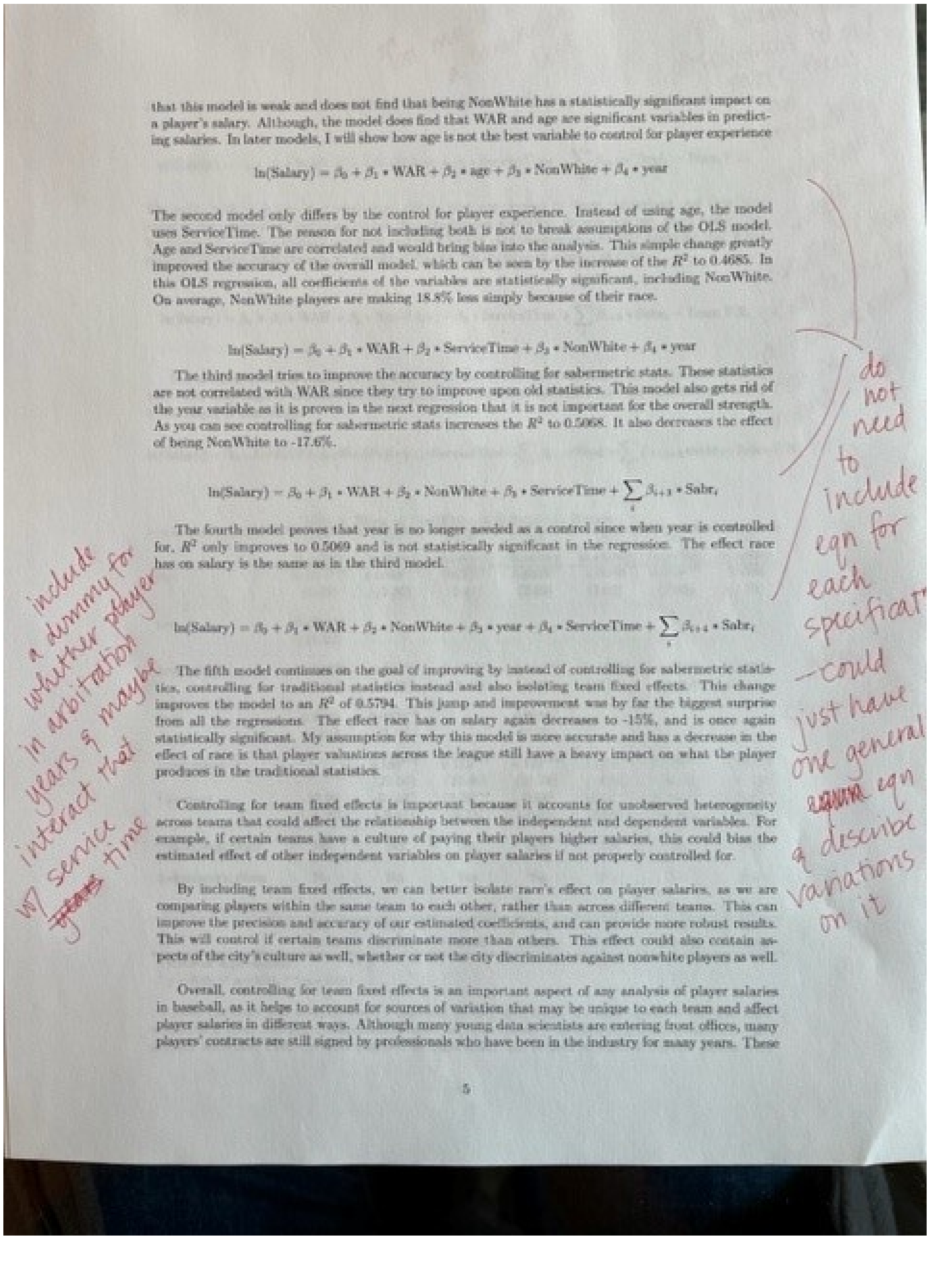

that this model is weak and does not find that being NonWhite has a statistically significant impact on a player's salary. Although, the model does find that WAR and age are significant variables in predicting salaries. In later models, I will show how age is not the best variable to control for player experience

$$\ln(\text{Salary}) = \beta_0 + \beta_1 * \text{WAR} + \beta_2 * \text{age} + \beta_3 * \text{NonWhite} + \beta_4 * \text{year}$$

The second model only differs by the control for player experience. Instead of using age, the model uses ServiceTime. The reason for not including both is not to break assumptions of the OLS model. Age and ServiceTime are correlated and would bring bias into the analysis. This simple change greatly improved the accuracy of the overall model. which can be seen by the increase of the R^2 to 0.4685. In this OLS regression, all coefficients of the variables are statistically significant, including NonWhite. On average, NonWhite players are making 18.8% less simply because of their race.

$$\ln(\text{Salary}) = \beta_0 + \beta_1 * \text{WAR} + \beta_2 * \text{ServiceTime} + \beta_3 * \text{NonWhite} + \beta_4 * \text{year}$$

The third model tries to improve the accuracy by controlling for sabermetric stats. These statistics are not correlated with WAR since they try to improve upon old statistics. This model also gets rid of the year variable as it is proven in the next regression that it is not important for the overall strength. As you can see controlling for sabermetric stats increases the R^2 to 0.5068. It also decreases the effect of being NonWhite to -17.6%.

$$\ln(\text{Salary}) = \beta_0 + \beta_1 * \text{WAR} + \beta_2 * \text{NonWhite} + \beta_3 * \text{ServiceTime} + \sum_i \beta_{i+3} * \text{Sabr}_i$$

The fourth model proves that year is no longer needed as a control since when year is controlled for, R^2 only improves to 0.5069 and is not statistically significant in the regression. The effect race has on salary is the same as in the third model.

$$\ln(\text{Salary}) = \beta_0 + \beta_1 * \text{WAR} + \beta_2 * \text{NonWhite} + \beta_3 * \text{year} + \beta_4 * \text{ServiceTime} + \sum_i \beta_{i+4} * \text{Sabr}_i$$

The fifth model continues on the goal of improving by instead of controlling for sabermetric statistics, controlling for traditional statistics instead and also isolating team fixed effects. This change improves the model to an R^2 of 0.5794. This jump and improvement was by far the biggest surprise from all the regressions. The effect race has on salary again decreases to -15%, and is once again statistically significant. My assumption for why this model is more accurate and has a decrease in the effect of race is that player valuations across the league still have a heavy impact on what the player produces in the traditional statistics.

Controlling for team fixed effects is important because it accounts for unobserved heterogeneity across teams that could affect the relationship between the independent and dependent variables. For example, if certain teams have a culture of paying their players higher salaries, this could bias the estimated effect of other independent variables on player salaries if not properly controlled for.

By including team fixed effects, we can better isolate race's effect on player salaries, as we are comparing players within the same team to each other, rather than across different teams. This can improve the precision and accuracy of our estimated coefficients, and can provide more robust results. This will control if certain teams discriminate more than others. This effect could also contain aspects of the city's culture as well, whether or not the city discriminates against nonwhite players as well.

Overall, controlling for team fixed effects is an important aspect of any analysis of player salaries in baseball, as it helps to account for sources of variation that may be unique to each team and affect player salaries in different ways. Although many young data scientists are entering front offices, many players' contracts are still signed by professionals who have been in the industry for many years. These

do not need to include eqn for each specificat'n — could just have one general eqn & describe variations on it

include a dummy for whether player in arbitration years & maybe interact that w/ service ~~years~~ time