Gender Pay Gap: A Myth?

T. Arthur, J. Miao, L. Zarate-Hernandez
STAT-021 Final Project, Swarthmore College, April 29th, 2022

Introduction

We analyzed a hypothetical dataset that was used by Glassdoor to demonstrate how companies can assess gender pay gaps. Glassdoor oversimplified the analysis by conducting the multiple linear regression on population dataset, not checking if assumptions were met, and automatically treating the model with the most predictors as the best model. In our analysis, we wish to show the complexity of using multiple linear regression to determine gender pay gap.

Data

- Hypothetical dataset
- Variables: Base Pay, Gender, Age, Education, JobTitle, Bonus, Performance Evaluation(perfEval)
- Response variable: Gross Income (Base Pay + Bonus)

Methods

- Random sampling of 500 observations
- Best subset selection
- Compared three models on assumptions
- Compared models adjusted R^2 & residuals standard error
- Compared significance of coefficients

Model

$$GrossIncome = 60267.60 + 910.41(AGE) - 3819.46(EducationHighSchool) + 2506.05(EducationMasters) + 4838.96(EducationPhD) - 1927.17(JobTitleDriver + 5038.41(JobTitleFinancialAnalyst) - 4776(JobTitleGraphicDesigner) - 5141(JobTitleIT) + 28009.44(JobTitleManager) - 13848.92(JobTitleMarketingAssociate) - 1174.95(JobTitleSalesAssociate) 11978.94(JobTitleSoftwareEngineer) + 2200.24(JobTitleWarehouseAssociate)$$

Adjusted R^2: 0.5056

Main Takeaways

- No gender pay gap in this hypothetical company
- Limited time with respect to building more models
- Use of hypothetical data affected methods used
- Recognize that there is no perfect model to assess the existence of gender pay gap
- Transparency about conclusions as well as statistical method

Results & Discussion

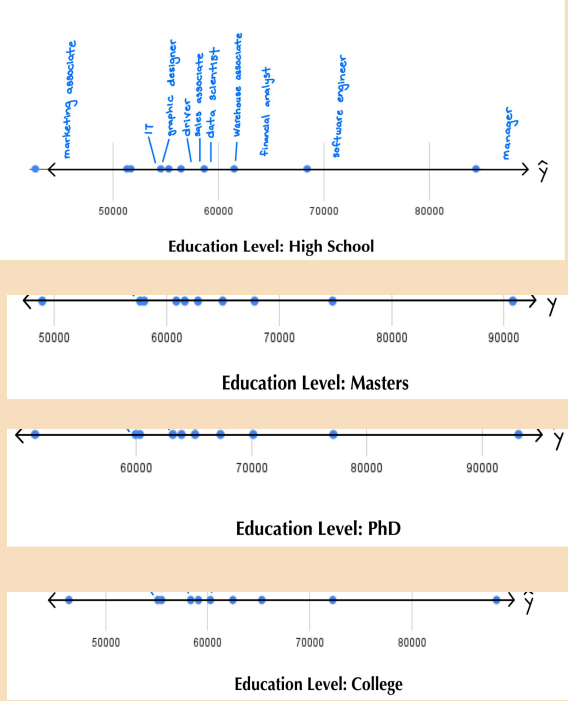


Figure 1: Line plots visualizing regression estimates. Gross income trend among the levels of job title remains constant through the different education levels.

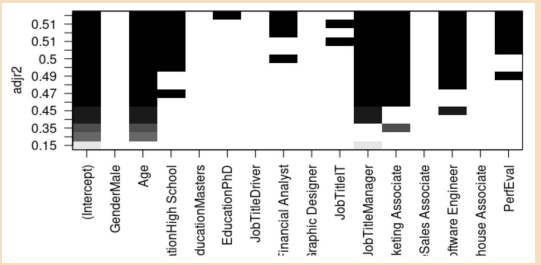


Figure 2: Best subset selection adjusted R^2 output. All models did not contain gender as predictor.

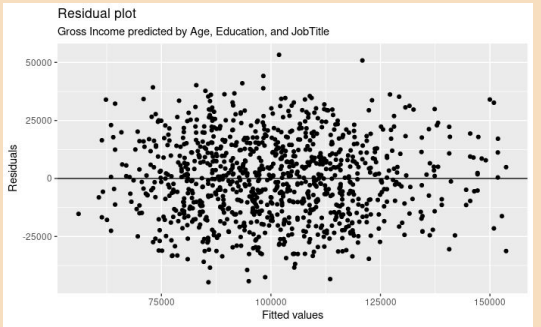


Figure 3: Residual vs fitted values plot. Assumption of linearity and constant variance met.

Future Work

- Further analysis: Standardized, studentized residuals, leverage, and cook's distance
- Additional variables: hours worked, race, and ethnicity
- Comparison of Glassdoor and our statistical method on real data

References

- Chamberlain, Andrew. 2017. "How to Analyze Your Gender Pay Gap: An Employer's Guide." Glassdoor

Sample Size Matters: Effect Size in Study Replications

Anika Rajamani, Abe Porschet, Koji Flynn-Do

April 29, 2022

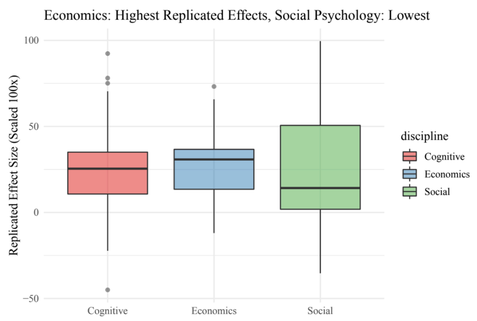
Hypothesis

Science as an institution is predicated on a foundation of replicability: from a frequentist perspective, it is the long-run ability to replicate the same empirical results a plurality of times that gives science explanatory power. Unfortunately, the "replication crisis" — in which results from studies in many domains (but especially psychology and medicine) have failed to be successfully replicated — has called the reliability of modern science into question.

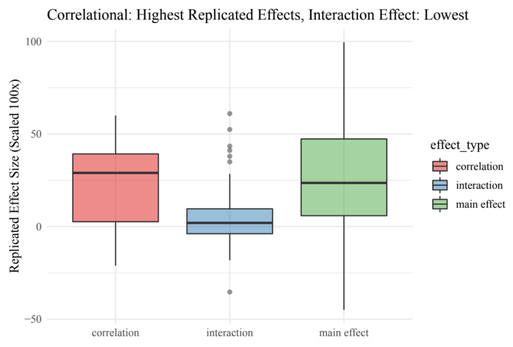
So, we ask: what should one expect a replication of said study to report as the replicated effect size? What factors does that depend upon? One hypothesis is that the replicated effect size is likely to be only weakly correlated with the original effect size, and that including variables such as the discipline (economics, social psychology, cognitive psychology), the sample size in the original, and others will improve prediction.

Dataset

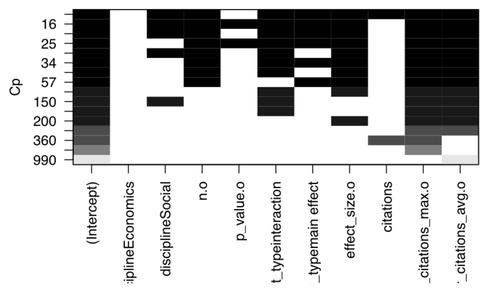
To investigate this question, we use the data set from Altmejd et al., "Predicting the Replicability of Social Science Lab Experiments," 2019. Crucially, we are not making causal claims in this report - our data are observational, not randomly sampled, and contain several limitations (discussed further in Section 3).

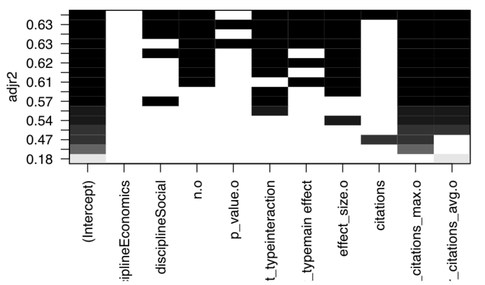


Dataset Cont.



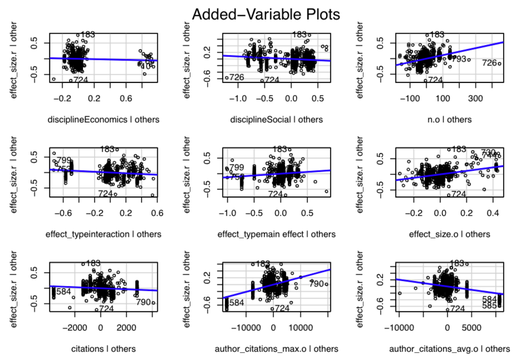
Fitting Model





We picked the variables to minimize Mallow's Cp and to maximize the adjusted r^2 values of the model. These are shown in the figures above.

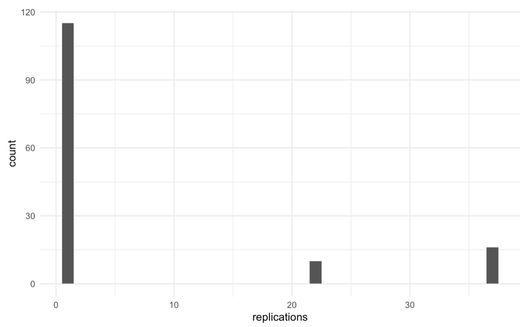
As shown in the next column, the added-variable plots suggest that all the predictors account for some variance in the complete model.



Limitations with the Dataset

Our dataset is limited in several ways; the population we can discuss is extremely narrow. Because of the problems with our data we can really only draw conclusions only about our specific study-replication pairs and not extrapolate.

There are large disparities in replication counts. As seen above, the 16 studies with 37 replications each are together responsible for roughly 64% of the overall study-replication pairs. The 10 studies with 22 replications each are together responsible for roughly 24% of the 4 overall pairs. That leaves just about 12% of the pairs coming from the other 115 studies. There is simply no way our data independently sampled. What selection process led to these 26 studies being replicated so often relative to the rest of the data? We cannot say for certain, but a guess is that some combination of surprising results, high-citation counts, and the relative difficulty of replicating these studies are some factors that contribute to their selection.



Limitations continued

Our data set had a few other small contributing problems, we had a few missing p-values and effect sizes, and one of the studies had 230,025 replications which is a wild outlier. We removed these and had a clearer picture of the data. Since the data set only included point estimates for the response variable and its explanatory counterpart effect size of original study, we have no idea of the uncertainty for the estimates which means we don't have enough information to use our data for inference. This severely limited our ability to make judgements

Conclusions

In this report, we analyzed a data set of experimental social science studies and replications, finding that, in addition to the effect size in the original study, the discipline (cognitive psychology, economics, or social psychology) and effect type (correlation, main effect, interaction effect) seem to matter substantially in this data set. "In this data set" is essential phrasing - there are many limitations of the data, including possibly dominating selection effects dictating which studies are replicated in the first place. Further research with more robust data could help determine if these relationships are an artifact of the data or if they correspond to really existing trends in social scientific experimental research. We emphasize the extremely narrow and contingent nature of this analysis and urge caution with any generalization.

References

Altmejd, Adam, Anna Dreber Almenberg, Eskil Forsell, Teck-Hua Ho, Juergen Huber, Taisuke Imai, Magnus Johannesson, Michael Kirchler, Gideon Nave, and Colin Camerer. 2017. "Predicting the Replicability of Social Science Lab Experiments," April. https://doi.org/10.17605/OSF.IO/4FN73.

Athletic Performance; How Can Athletes Reach the 1%?

Liam Maniscalco and Will Vespole

STAT021 - April 29, 2022

Introduction

- **Driving question:** *How do different variables drive or inhibit athletic performance?*
- While there have been studies that have analyzed the impact of individual factors on athletic performance (i.e. sleep on injury risk), we were interested in testing how the confluence of several variables would impact athletic performance
- By testing multiple variables, we hoped to be able to measure their individual impact on athletic performance and also analyze their collective impact

Methods

- **Observational units** = 50 Swarthmore Lacrosse players
- **5 predictor variables:** Sleep (measured in hours per night), Nutrition (self-measured on a scale of 1-10), Workouts (per week, in addition to practices), Stress (self-measured on a scale of 1-10), and Class (senior, junior, sophomore, or freshman)
- **2 response variables:** ShuttleChange (change in 300 yd shuttle time from January to April) and BenchChange (change in bench press one rep max from January to April)
- **Model used:** Multiple Linear Regression
- **Primary outcome measured:** Whether the predictor variables had a statistically significant impact on the response variables

Key Infographics

First MLR:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.1533     7.6483   1.066    0.293
Sleep        -0.9483     0.9021  -1.051    0.299
Nutrition    -0.1995     0.5419  -0.368    0.715
Workouts     -0.3805     0.5539  -0.687    0.496
Stress       -0.1398     0.4436  -0.315    0.754
ClassJr.      1.1112     2.3244   0.478    0.635
ClassSo.      1.5374     2.1794   0.705    0.484
ClassSr.     -0.1358     2.2282  -0.061    0.952

Residual standard error: 5.516 on 42 degrees of freedom
Multiple R-squared:  0.09268,   Adjusted R-squared:  -0.05854
F-statistic: 0.6129 on 7 and 42 DF,  p-value: 0.742


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.0209    24.8912  -1.166   0.2502
Sleep         7.0750     2.9360   2.410   0.0204 *
Nutrition    -3.9888     1.7637  -2.262   0.0290 *
Workouts      2.5635     1.8026   1.422   0.1624
Stress       -0.2783     1.4437  -0.193   0.8481
ClassJr.      2.9418     7.5647   0.389   0.6993
ClassSo.      0.6647     7.0928   0.094   0.9258
ClassSr.     -7.7524     7.2517  -1.069   0.2912
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.95 on 42 degrees of freedom
Multiple R-squared:  0.2356,    Adjusted R-squared:  0.1082
F-statistic:  1.85 on 7 and 42 DF,  p-value: 0.1029
```

Revised MLR:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.1218     5.4989   1.477    0.146
Sleep        -0.9271     0.8493  -1.092    0.281
Nutrition    -0.2864     0.4967  -0.577    0.567
Workouts     -0.3087     0.5042  -0.612    0.543

Residual standard error: 5.314 on 46 degrees of freedom
Multiple R-squared:  0.07773,   Adjusted R-squared:  0.01758
F-statistic: 1.292 on 3 and 46 DF,  p-value: 0.2883


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -28.660     18.145  -1.579   0.1211
Sleep          7.039      2.803   2.512   0.0156 *
Nutrition     -4.376      1.639  -2.670   0.0105 *
Workouts       2.594      1.664   1.559   0.1259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 17.54 on 46 degrees of freedom
Multiple R-squared:  0.2012,    Adjusted R-squared:  0.1491
F-statistic: 3.863 on 3 and 46 DF,  p-value: 0.01514
```

Results

- Multiple linear regression was used to test if Sleep, Nutrition, and Workouts significantly predicted **Change in Shuttle times**
- Fitted Regression: $y = 8.1218 - 0.9271X_1 - 0.2864X_2 - 0.3087X_3 + e$, where X_1 is Sleep, X_2 is Nutrition, and X_3 is Workouts.
- No coefficients were statistically significant ($p > 0.05$) and test was not significant ($p > 0.05$)
- Multiple linear regression was used to test if Sleep, Nutrition, and Workouts significantly predicted **Change in Bench Press**
- Fitted Regression: $y = -28.660 + 7.039X_1 - 4.376X_2 - .594X_3 + e$, where X_1 is Sleep, X_2 is Nutrition, and X_3 is Workouts.
- Both Sleep and Nutrition were statistically significant, with p-value = 0.0156 and p-value = 0.0105, respectively. The overall test was significant (p-value = 0.1259).

Conclusions

- The **Sleep** and **Nutrition** variables each had significant effects on the change in **Bench Press**, and this overall model was statistically significant. The was not true for the Shuttle model, as no coefficients or the model were significant.
- Workouts and Stress are more difficult to standardize and could have ambiguous effects from person to person.
- Future projects would benefit from more controlled environment and longer test period.

References

- https://link.springer.com/article/10.1007/s40279-014-0147-0
- https://pubmed.ncbi.nlm.nih.gov/32005349/
- https://www.jstor.org/stable/48514109?searchText=nutrition+athletic+performance&searchUri=%2Faction%2FdoBasicSearch%3FQuery%3Dnutrition%2Bathletic%2Bperformance%26so%3Drel%26efgs%3DeyJjdHkiOlsiYW05MWNtNWhiOT09Il0sImRpc2MiOltdfQ%253D%253D&ab_segments=0%2Fbasic_search_gsv2%2Fcontrol&refreqid=fastly-default%3Afa256c86a94714aee0ba7c851e79e4dc&seq=4

Statically Check(ed) Mate: Predicting the Length of a Chess Game

Malavika Eby, Sumin Byun, Alex Li, and Mackenzie Tucker

Swarthmore College / April 29, 2022

INTRODUCTION

Chess is a notoriously intricate game that, despite being created long ago, has successfully evolved with the internet. One of the prominent examples of this is Lichess, a virtual space for chess players around the world to play each other. Lichess ranks players allowing them to have a reference for the abilities of their opponent. Our goal was to look to see if these ratings had a relationship to the length of the game. Along with other variables, we attempted to create a multiplier linear regression model that predicted the total number of turns in a chess game on Lichess.

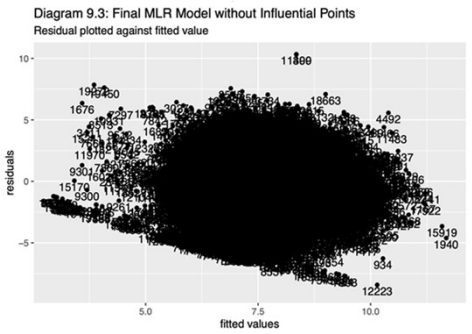
RESEARCH QUESTION

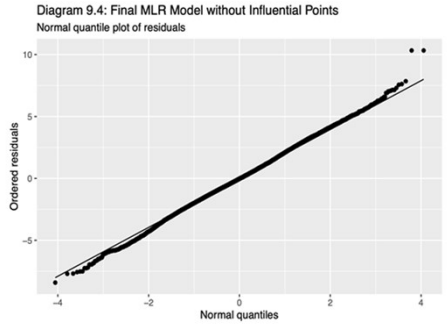
To what extent do the absolute differences in ratings of two chess players have a statistical significance on predicting the number of turns to end a Lichess chess game in a larger statistics model?

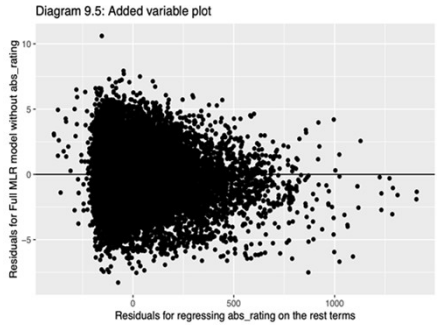
BACKGROUND AND METHODOLOGY

- In order to predict the total number of terms, we designed a multiple linear regression model in R Studio using the dataset.
- The data is a collection of separate games from Lichess API, and it has 20, 058 individual cases, where each case is an observational unit.
- Our main variable of interest was abs_rating, a numerical variable of the absolute difference between the player's ratings.
- Ratings are determined by a rating system called Glicko 2
  - It creates 95% confidence intervals for the player's score.

ASSUMPTION CHARTS







OUR MODEL

sqrt(turns) = if_mate - winner + duration + time_control + time_added + avg_rating + abs_rating + opening_name

USING OUR MODEL

From our research question, we completed a hypothesis test based on statistical significance of abs_rating.

The null hypothesis states that the absolute value of difference in rating doesn't have statistical significance in predicting turns.

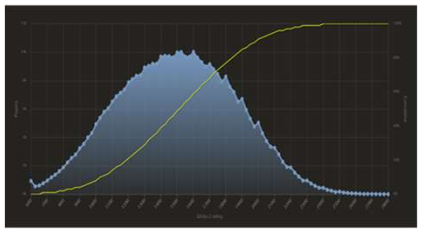
$$H_0: \beta_{abs\_rating} = 0$$

The alternative hypothesis states that the absolute value of difference in rating has statistical significance in predicting turns.

$$H_A: \beta_{abs\_rating} \neq 0$$

MORE INFORMATION ABOUT RATINGS

This is the distribution ratings of rapid chess games the week of April 25th. It further shows that the variable of interest (abs_rating) appears to be approximately Normal.



DISCUSSION AND CONCLUSIONS

- We removed four influential points to help our model meet the assumptions of completing a regression analysis.
- By transforming the response variable (square rooting it, we were able to further ensure assumptions were met.
- We got a coefficient of determination (adjusted) of 0.11, meaning a lot of the error in the model is not accounted for.
- We found that ratings are statistically significant in predicting the number of moves.

FURTHER APPLICATIONS

- We would like to focus more effort on a Poisson model because our response is a count variable.
- There is potential to broaden our conclusion by taking data sets from other online chess websites, as well as in-person games.
- Given the chance, we would like to ensure the randomness assumption by taking our own random samples of this data set.

REFERENCES

ACKNOWLEDGEMENTS

We would like to thank Professor Thornton for her guidance in preparing us all semester to complete this project. We also appreciate her additional help during the actual project creation process.

Predicting Life Expectancy

Which of the following impacts life expectancy the most?

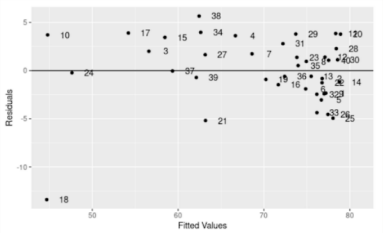
Infant mortality rates
New HIV/AIDS infections
Income level (GNI per capita)
Healthcare expenditure (% GDP)
Hepatitis B vaccination rates
Geographical region

PREDICTIONS

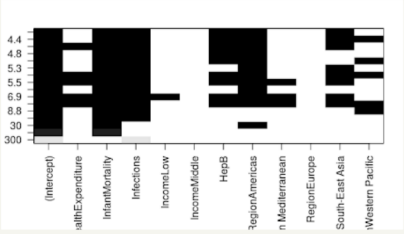
Initially, we expected infant mortality rates (per 1000 births) and income level (GNI per capita) to have the strongest impact on life expectancy.

FITTING THE MODEL

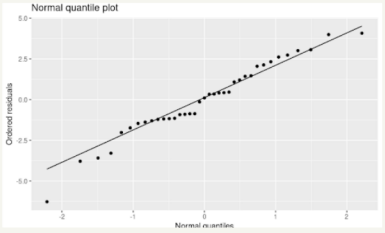
We used Mallow's Cp to find which of our predictors had the strongest impact on life expectancy. From there we were able to build residual and quantile plots, looking for outlying data points.



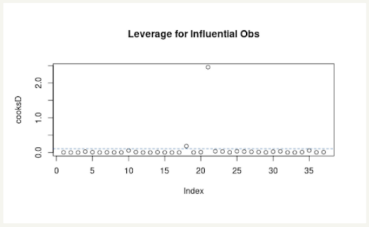
Residual plot of infant mortality rates & HIV infections in relation to life expectancy



Mallow's Cp model showing the impact of each predictor on the response.



Normality of residuals seems to be met. Kiribati was a consistent outlier (lower left corner)



Kiribati also had a high Cook's distance of 2.43, making it extremely influential

Most of our data was normally distributed, with a few outliers. Linearity and equal variance assumptions were met so we were able to draw conclusions. Our model had an adjusted R-squared value of 0.91.

RESULTS

Infant mortality rates and new HIV/AIDS infections (per 1000 uninfected) had the strongest impact on predicting life expectancy. An explanation for this could be how life expectancy is calculated.

CONCLUSION

Global access to healthcare is shaped by a variety of factors. Though life expectancy is only directly impacted by some of those factors, it is important to consder the ways all those factors interact in order to expand healthcare access.

Stat 021 Spring 2022 April 29
Jake Bohman, Isabelle Courtney, Yasharbek Sabitov, Raya Tuffaha

References
Data collected from WHO GHO
Life expectancy calculations from Britannica