Effects of Employment Sector, Hours Worked Per Week, Certifications, and Education on Salary Earnings

Marta Gonzalez

I decided to focus my issue on how employment sector, hours worked per week, certifications and education level relate to salary. Specifically, I wanted to analyze whether an individual’s education level, employment sector, hours worked per week, or certifications had any influence on or relationship to their salary when analyzed together as well as whether an individual’s hours worked per week had any influence on or relationship to their salary earnings when analyzed separately. I believe analyzing the relationship between these five variables is important because it can have financial, educational, and career implications on people working in the data field. This analysis could primarily affect anyone who is pursuing a career in a data related field and wants to learn how to maximize their salary earnings. However, it could also affect educational institutions, employers, or employment sectors looking to make their data related jobs or degrees more appealing to potential employees or students.

The population of respondents represented in my sample are all people with access to the survey on the Brent Ozar website. My particular sample included all survey respondents who identified as male or female.

My research question revolved around the theory that education level, employment sector, the number of hours an employee works per week, and the certifications held by employees generally tend to have an effect on the salary earnings of employees. My null hypotheses were that education level acquired had no effect on salary earnings, employment sector had no effect on salary earnings, hours worked per week had no effect on salary earnings, certifications obtained had no effect on salary earnings, and the interaction between education level, hours worked per week, certifications obtained, and employment sector had no effect on salary earnings.

My recommendation is that anyone in the data field trying to decide which sector to pursue should keep reading.

The data set I am using is a data set compiled by Brent Ozar[[1]](#footnote-1). The data set includes all Professional Salary Survey Results from the years 2017, 2018, and 2019. The survey used to create this data was publicly displayed on the website and anyone was able to take it. However, Ozar describes the data as coming from database administrators, analysts, architects, developers, and data scientists. Due to the dataset being from a range of years, the survey questions have evolved and new ones have been added each year. This has resulted in “Not Asked” being populated in certain columns for the years that did not provide a certain question. I used the Employment Sector variable which had a majority of substantially populated answers. I then made the decision not to include the “Education (K-12, college, university)”, “Student”, “Private business”, and “Non-profit” options as they contained frequencies too different from the rest of the categories. I also used the Education Level variable. I removed the “Doctorate/PhD” and “Bachelors” options as they contained frequencies too different from the rest of the categories. I did not remove any categories from the Certifications variable. There were also some NA values in the variables I used, so those had to be removed before beginning any analysis.

The sample was obtained through a public survey and any with the link could voluntarily respond. The resulting respondents amounted to 882 from 46 countries. The majority of the respondents from my cleaned dataset came from the State Government sector, had obtained an Associate’s degree and did not hold any certifications.

The target population for my sample is all data-related professionals from the year 2019. From the original dataset provided, I decided to use five variables, Employment Sector, Education Level, Hours Worked Per Week, Certifications, and Salary. The Salary variable was reported through fill in the blank questions where the respondents could put in whatever number they wanted (hopefully accurately). The Education Level and Employment Sector variables had answers only related to 5 education categories and 6 employment sector categories, so it was assumed that the survey provided the categories for level of education completed and respondents simply had to choose from the provided answers. The Certifications variable had 3 different categories, and it was assumed that respondents had to choose from the provided answers. The Education Level categories used were “Masters”, “Associates (2 years)”, and “None (no degree completed)”. The Employment Sector categories used were changed to “Federal Gov’t”, “State Gov’t”, and “Local Gov’t”. The Certifications categories used were changed to “None”, “Valid”, and “Expired”. In order to analyze the education categories in a frequency table, I adjusted the names to exclude the year associations and changed the ‘none’ category to “No degree”.

A screenshot of a cell phone

Description automatically generatedIn the Employment variable, the Education Level variable, and the Salary variable there were two columns of missing data, or data containing “NA”, that I removed before beginning the frequency table analysis. The largest percentage of respondents had received their associates degree at 40% of the sample. The lowest percentage of respondents had received their masters degree at 25%. The remaining 35% had not received any degree at all. The respondents in the local government and State government sectors were equal at 38%. The remaining 24% of respondents worked in the federal government. The largest percentage of respondents had not received any certifications at 45% of the sample. The lowest percentage of respondents had expired certifications at 27% and the remaining 28% had valid certifications. All of the frequencies among the three categories were relatively similar.

A screenshot of a cell phone

Description automatically generatedI created a summary statistics chart for salary and hours worked per week (with outliers). and found the mean of Salary to be 84460.92208 while the median was slightly less at 82900, indication that the outliers had some effect on the mean. The mean for Hours Worked Per Week was 41.43377 while the median was 40. This indicated that the outliers had some effect on the mean. The standard deviation for Salary was 36535.587848 and the standard error was 1862.025973. The standard deviation for Hours Worked Per Week was 6.036038 and the standard error was .307625.

I created a scatter plot and three box plots to further examine my data. I filtered out the outliers in the Salary and Hours Worked Per Week variables to get a clear view of the data. The relationship between salary and hours worked per week appeared to be a positive, weak relationship. The amount of hours worked per week appeared to have a positive, weak strength effect on salary earnings. The model specification for the scatter plot was as follows: SalaryUSD A screenshot of a video game

Description automatically generated= 𝛽0 + 𝛽1(HoursWorkedPerWeek) + 𝛽2(EMP\_cat) + 𝛽3(EDUC\_cat) + 𝛽4(CERT\_cat). For the distribution of salary earnings by employment sector I found that the federal government plot was distributed much differently than the others and had the highest group mean of the three groups while that of local government had the lowest group mean. The medians of respondents who worked in local government and state government were very similar. I found that the distribution of the box plots of salary earnings by education level were all relatively distributed the same and the group means were also very similar. The plot for respondents who held an associate degree had a slightly larger median. The distribution for the expired certification plot was the most different from the other two plots and had the highest median and mean. The plot indicating the respondents with no certifications had the lowest median and mean.

A picture containing map, photo, man, game

Description automatically generatedA screenshot of a cell phone

Description automatically generatedI first conducted an analysis on the relationship between salary earnings and hours worked per week. The data reaches a minimum of 8 hours worked per week, so having 0 hours worked per week was not possible with this sample. As a result, the intercept is not interpretable. For each increase of 1 hour worked per week, there is an average increase of $503.9 in salary earnings. The p-value of .103 is greater than alpha (0.05), therefore we fail to reject null. This means that the coefficient for hours worked per week is not significantly different from zero and is not a significant predictor of salary earnings. As a result of our p-value being greater than alpha, the model is not an acceptable model for predicting salary earnings, and because r-squared is 0, r-squared is not significantly different from zero. We also know that this r-squared is not significantly different from zero because the F-test was not significant with the p value being .1029. The p-value is greater than alpha, therefore this model is not significantly better than a null or empty model.

The QQ Plot of Residuals was relatively normally distributed with some slight deviation in the lower tail and slight deviation in the upper tail.

A screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generatedA screenshot of a social media post

Description automatically generatedThe Residuals vs. Leverage plot showed a three possible influential outliers at points 222, 114, and 117. There does not seem to be any noticeable cone/funnel shape in the residuals vs. fitted plot, therefore I don't anticipate that the data has heteroscedasticity. As a result, it is assumed data is not violating the assumption of constant variance. The p-value for this assumption is also approximately .00158, so we know the data is not violating the assumption of constant variance. In this case there is a significant curve perhaps indicating a potential curvilinear relationship, so the data is violating the assumption of independent errors. After eliminating the outliers and refitting the model, there still appeared to be an influential outlier at point 294.

The second analysis I conducted was on the relationship between salary earnings and all four variables. For a respondent working in the federal government, their education level is predicted to be a Masters degree, they are predicted to not have certifications and their fitted/predicted salary A screenshot of a cell phone

Description automatically generatedearnings are $86416.210. Holding all else equal, the average salary earnings for respondents working in local government is predicted to be $25693.621 less than that of federal government workers. Holding all else equal, the average salary earnings for respondents working in state government is predicted to be $20330.886 less than that of federal government workers. Holding all else equal, the average salary earnings for respondents with an associates degree is predicted to be $4888.846 more than that of respondents with a masters degree. Holding all else equal, the average salary earnings for respondents with no degree is predicted to be $4757.764 less than that of respondents with a masters degree. Holding all else equal, the average salary earnings for respondents with valid certifications is predicted to be $1387.354 more than that of respondents with no certifications. Holding all else equal, the average salary earnings for respondents with expired certifications is predicted to be $18009.197 more than that of respondents with no certifications. The p-value greater than .05 are for Hours Worked Per Week at .414, Associates Degree at .280, No degree at .310, and Valid Certifications at .746, so we have determined that these are not significant predictors of salary earnings. The p-values for Expired Certifications, State Government, Local Government, and all values held in the intercept are less than .001 and are determined to be significant predictors of salary earnings. For every increase of 1 hour worked per week, there is an average increase of $240.643 in salary earnings. R-squared is about 0.12, meaning that the model explains about 12% of the variance in salary earnings which is not substantial. We know that this r-squared is significantly different from zero because the F-test was significant with the p value being 2.048e-09. The p-value is less than alpha, therefore this model is significantly better than a null or empty model.

A picture containing photo

Description automatically generatedA picture containing map, photo, man, game

Description automatically generated The QQ Plot of Residuals shows a relatively normal distribution with slight variation in the lower tail and slight deviation in the upper tail.

A screenshot of a social media post

Description automatically generatedThe Residuals vs. Leverage plot does not show any influential outliers. There does not seem to be any noticeable cone/funnel shape, therefore I do not anticipate that there is heteroscedasticity. This data is not violating the assumption of constant variance. The errors appear to be independent as the residuals still appear relatively normal with no influential outliers, and no clear violation of constant variance is evident. The model is good, with slight significance and weak strength predictive power. In the test for multicollinearity, all VIF’s were around 1 so the data does not have multicollinearity.

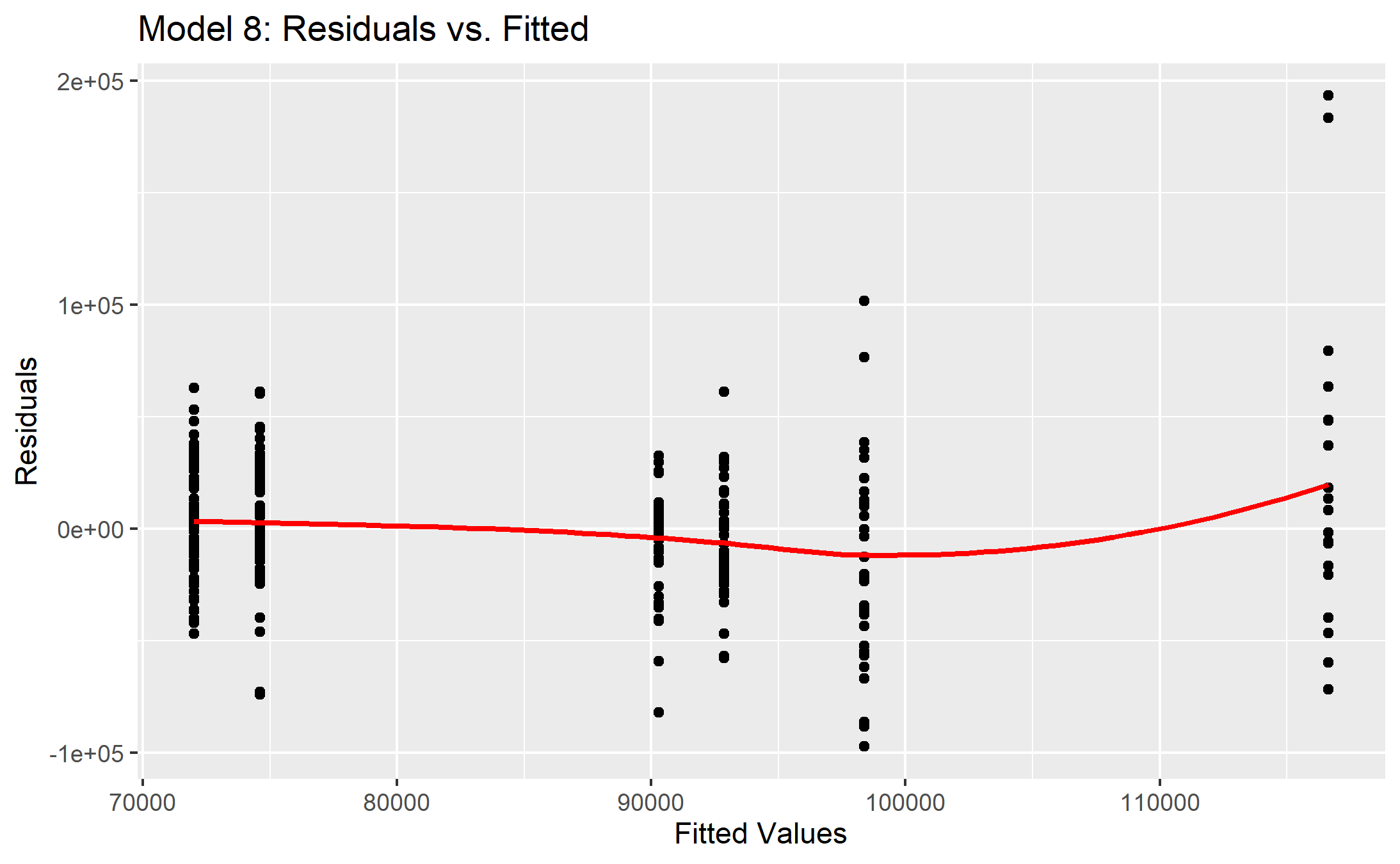
The p-value of 2.246e-09 is less than alpha = 0.05, therefore we reject null. This means that the larger model is preferred. I would retain the larger model to best predict salary earnings.

In my final model, I removed all insignificant predictors. This included the Hours Worked Per Week variable, and the Education Variable as well as the Valid Certification category of the Certifications variable. While respondents holding a Master’s degree was a significant predictor of salary earnings, none of the other categories were, making it impossible to include education as one of the predictors of salary earnings as the model fitting required at least two levels in every variable.

A screenshot of a cell phone

Description automatically generated For a respondent working in the federal government, they are predicted to not have certifications and their fitted/predicted salary earnings are $98367.70. Holding all else equal, the average salary earnings for respondents working in local government is predicted to be $26353.98 less than that of federal government workers. Holding all else equal, the average salary earnings for respondents working in state government is predicted to be $23766.21 less than that of federal government workers. Holding all else equal, the average salary earnings for respondents with expired certifications is predicted to be $18269.67 more than that of respondents with no certifications. The p-values for all of the variable categories are less than .001 and are determined to be significant predictors of salary earnings. R-squared is about 0.12, meaning that the model explains about 12% of the variance in salary earnings which is not substantial. We know that this r-squared is significantly different from zero because the F-test was significant with the p value being 1.608e-08. The p-value is less than alpha, therefore this model is significantly better than a null or empty model.

A screenshot of a social media post

Description automatically generatedA picture containing photo

Description automatically generated The QQ Plot of Residuals shows a relatively normal distribution with slight variation in the lower tail and slight deviation in the upper tail. The Residuals vs. Leverage plot does not show any influential outliers. There does not seem to be any noticeable cone/funnel shape, therefore I do not anticipate that there is heteroscedasticity. This data is not violating the assumption of constant variance. The errors appear to be independent as the residuals still appear relatively normal with no influential outliers, and no clear violation of constant variance is evident. The model is good, with slight significance and weak strength predictive power. In the test for multicollinearity, all VIF’s were around 1 so the data does not have multicollinearity.

I was able to reject my null hypothesis that there was no effect of employment sector, education level, hours worked per week, and certifications combined on salary earnings. I was not able to reject the null hypothesis that hours worked per week had no effect on salary earnings. These findings could have significant implications on whether or not employees attempt to acquire certifications as well as which employment sector data-related employees decide to pursue as their salaries will either benefit or not.

This analysis could be further improved in the future by providing more equal frequencies for the dataset. This analysis could also potentially be improved by testing more possible predictors of salary earnings related to this field.

1. <https://www.brentozar.com/archive/2019/01/the-2019-data-professional-salary-survey-results/> [↑](#footnote-ref-1)