Allan Swartz
6/3/2015

# Final Report

This report seeks to investigate certain factors that may help to explain wage gaps among the United States population. Originally the response variable was simply going to be **wage**, but due to the skew detected from the resulting histogram of the values, instead **log(wage)** will be the response variable. While there are a number of predictor variables already included from the study, additional interaction terms will be added and explored. The number of potential interactions is monumental, however only a certain few will be included so as to focus on implications to having these explanatory terms appear in the regression model. A current widely discussed topic revolves around women in the workplace and their wage comparisons to men. Along with the female variable, two new interactions are added including **nonwhite** with **female** and **married** with **female**. The former interaction separates white paid females from nonwhite females to see if race also plays a role to how much a woman gets paid. Businesses also seem to care about whether a woman is married or not – married indicating a greater likelihood of having children – so the latter interaction looks at this. **Education** and the three **regions** will be added as interactions to see if each one might value education differently to determine what one's wage should be. **Experience** will also be interacted with **job types** as well as **business types** to see if how long one has been in the workforce has an effect on wage depending on what kind of job that person has and what kind of business their company conducts. In total there will be 15 interactions added to the model.

The preliminary model with all 37 variables (or 38 parameters) has n=526 observations and is overall significant (F=17.64, p-value <0.0001) at a 95% confidence level. Assumptions were then checked, particularly to see if the error terms follow normality. Figure 1 shows the normal probability plot which appears to follow the normal line, however the Shapiro-Wilk test says otherwise. Considering the large amount of observations taken in this data set, the central limit theorem should hold. Linearity of the quantitative variables is another assumption that was checked by look at their residual plots against the respective predictor variable. For the most part they appear random with no obvious pattern, so linearity of these variables seems to hold. Some outliers were detected in these residual plots. Using the DFFITS method, with a threshold of $2\sqrt{\frac{p}{n}}$, only about 8.37% of the data actually appeared as outliers

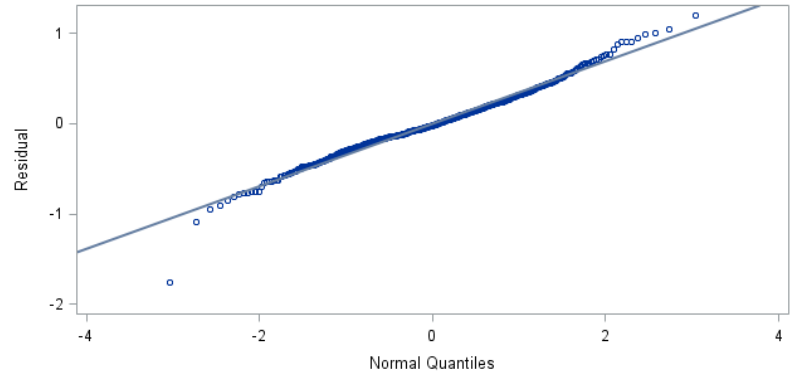


***Figure 1*** Q-Q plot of residuals. It appears to follow normality, however the Shapiro-Wilk value is 0.9847 with a p-value of <0.0001. $\alpha = 0.05$

according to this criteria. The influence of these observations should be offset by the fact that the data set is so large. The assumption of constant variance was checked by plotting the residuals against the fitted values as shown in Figure 2. This plot seems to show that the constant variance assumption is held, and therefore no remedial measures need to be taken based on the error variance. It also includes residuals plotted against time, which does not seem to violate independence.

After the full model was investigated, several model selection methods were used. Stepwise selection at an $\alpha = 0.01$ level yielded the variables **education**, **tenure**, **married**, **smsa**, **trade**, **services**, **professional job**, **service job**, **tenuresquared**, and **married female**. Since only two of the business type indicators appeared, a multiple regression coefficient test was conducted to see if the remaining indicators could be added to the model and at least one remain significant. Using the extra sum of squares method, all business types came out significant (F=18.031, with 3, 520 degrees of freedom, p-value <0.0001). Similarly job types was tested, since clerical services did not appear in the model, and all three variables also came out as significant (F=7.394, with 6, 517 degrees of freedom, p-value <0.0001). The final model under stepwise selection added in all business and job types. Next, the full model was fitted under Mallows Cp. The model chosen, with a Cp value of 11.9247 and 14 variables, included **eductation**, **experience**, **tenure**, **female**, **married**, **number of dependents**, **smsa**, **west region**, **trade**, **services**, **professional job**, **service job**, **experiencesquared**, and **married female**. Like in the stepwise selection, indicator values were all tested ($\alpha = 0.05$) for their coefficients in their respective groups if only a certain number of them appeared in the chosen model. Regions did not have significance (F=2.377, with 3, 511 degrees of freedom, p-value 0.0691) so they were entirely removed from the model. However business types and job types were tested (F=7.489, df=6, 513, p-value <0.0001 and F=11.6973, df=3, 509, p-value <0.0001 respectively) and came out significant. It turned out that both Mallows Cp and Stepwise selection yielded very similar models, however Mallows Cp includes **education** instead. Finally, adjusted $R^2$ selection was

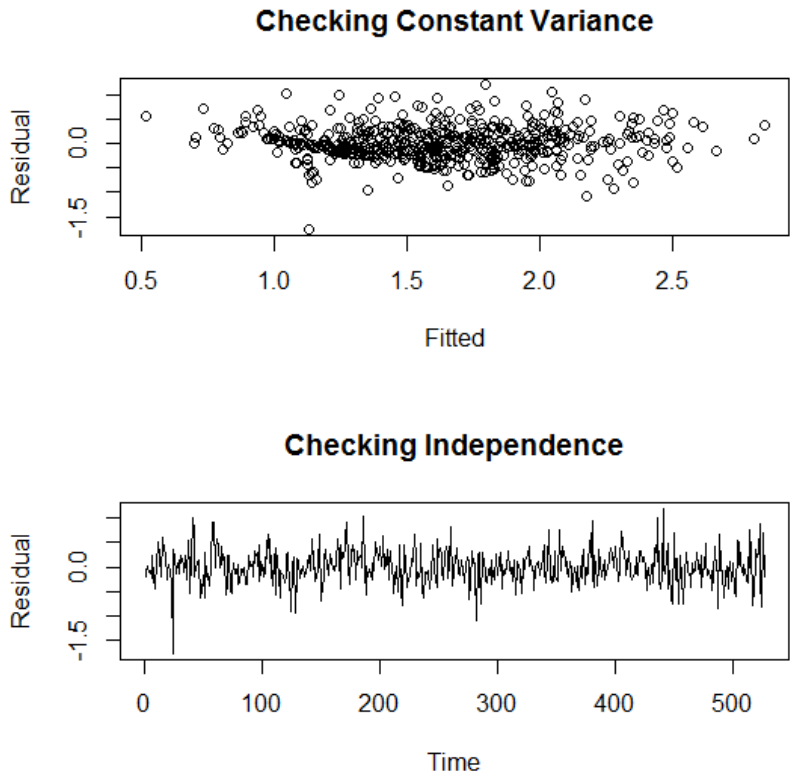


***Figure 2*** The top graph shows the residuals plotted against the fitted values. The second graph shows residuals plotted against time.

tested, however this method did not achieve the parsimony of the other two and consequently was not so interpretable. Furthermore, it seems appropriate to choose between one of the two methods that produces similar results. The final model chosen was that outlined by Mallows Cp, mostly due to the fact that it included **education** as a variable to be explored in the model.

The final model, chosen under Mallows Cp, is significant (F=33.42, p-value <0.0001) at a 95% confidence level. Table 1 shows all of the least squares estimates for this model as well as their respective significant levels. A lack of fit test was performed to determine that this response surface is adequate, and no significant departure was detected (F=0.51, p-value 0.8575). Additionally, the PRESS score was obtained for model validation (71.3345) and when comparing with the MSE (PRESS/n=0.1356, MSE=0.13) they are very much the same. This demonstrates that the MSE is a reasonable indicator for the predictive ability of this model.

| Variable | Estimate | SE | t Value | Pr > \|t\| | Variable | Estimate | SE | t Value | Pr > \|t\| |
|---|---|---|---|---|---|---|---|---|---|
| **Intercept** | 0.7944 | 0.1126 | 7.05 | <.0001 | **smsa** | 0.1423 | 0.0372 | 3.83 | 0.0001 |
| **educ** | 0.0467 | 0.0077 | 6.05 | <.0001 | **construc** | -0.0613 | 0.0852 | -0.72 | 0.4719 |
| **exper** | 0.0256 | 0.0050 | 5.13 | <.0001 | **ndurman** | -0.1232 | 0.0626 | -1.97 | 0.0497 |
| **expersq** | -0.0005 | 0.0001 | -5.05 | <.0001 | **trcommpu** | -0.0985 | 0.0881 | -1.12 | 0.2639 |
| **tenure** | 0.0216 | 0.0063 | 3.43 | 0.0007 | **trade** | -0.3022 | 0.0531 | -5.69 | <.0001 |
| **tenursq** | -0.0004 | 0.0002 | -1.72 | 0.0864 | **services** | -0.2849 | 0.0676 | -4.22 | <.0001 |
| **female** | -0.1246 | 0.0531 | -2.34 | 0.0195 | **profocc** | 0.2181 | 0.0478 | 4.57 | <.0001 |
| **married** | 0.1841 | 0.0524 | 3.52 | 0.0005 | **clerocc** | 0.0461 | 0.0565 | 0.82 | 0.415 |
| **marriedfemale** | -0.2417 | 0.0672 | -3.6 | 0.0004 | **servocc** | -0.0962 | 0.0564 | -1.7 | 0.0889 |
| **numdep** | -0.0266 | 0.0141 | -1.89 | 0.0589 | **profserv** | -0.0945 | 0.0583 | -1.62 | 0.1057 |

***Table 1*** The least squares estimates, their standard errors, t values, and p-values. Indicators of similar types were kept in the model after determining if they would be significant, and main effects were included with interactions.

Many of the interactions that were added into the model seemed to drop out, with the exception of **married female**. Examining **female**, **married**, and their interaction variables seems to show that there is a disparity between a man's wage against a woman's wage. When all other variables are held constant, and when the indicator for **female** and **married female** are both 0, there is a positive trend in the estimated values implying that men who are married get paid more. When **married** and **married female** are both 0, a negative trend emerges. And even more surprising, when **female**, **married**, and **married female** are all 1, there is an greater negative estimate. This shows that not only does a female get less pay than men at a comparable experience and education level, but also that a female who is married gets even less than an unmarried one. Another conclusion that can be made is that, as expected, **education**, **experience**, and **tenure** all have a significant positive impact on wage. When all other variables are held constant, **education** seems to have the greatest impact compared to the other three. Business and job types for the most part do not signficantly explain the model, save for **trade**, **manufacturing**, and **services** on the business side and **professional** job types on the other. These are interpreted as that **trade**, **manufacturing**, and **service** business have a lower intercept while **professional** job types have a higher intercept.

While the regions did not make it into the final model, the indicator for a metropolitan area did. This seems to make a significant difference in determining wage, but it is most likely due to the lack of high paying jobs for equivalent education and experience in rural areas. At a 95% confidence level, number of dependents turns out to not be significant in determining wage. When both **experience** and **tenure** interaction with themselves, there is significant negative effect on the overall model. The interactions of **experience** on **businesses** and **jobs** was not included, indicating that experience does not seem to vary depending on the level of these indicators. Similarly, **education** does not seem to vary with different values of **regions**. **Nonwhite females** also was dropped during the model selection. Even though the majority of these interaction terms were not included after model selection, conclusions about wage differences particularly among women could be made. Overall it seems that the greatest estimated wage would be for a highly educated, highly experienced male in a professional capacity, living in a metropolitan area, married, and with tenure as long as all the values are within the scope of the model.