**Alysia J. Sturges**

**Econometrics**

**Summer 2013**

**Pension Regression Paper**

**Introduction**

With the rise of the feminist movement and the growing trend of women becoming heads of households, the question begs to be asked, do women also think about retirement as much as men? The theme song of the feminist movement being “Anything you can do I, I can do better, I can do anything better than you . .” asserting that they are equally matched with men in every facet, it’s no secret that, that assertion is not quite true. On average, men make more money than women do when the skills and education are held equal. Given the data set of the pension plans offered by companies in 1989, I wanted to test the theory that men and women are equal, at least when it comes to retirement preparation. Holding all things constant, I expected to see that men and women on average spend about the same amount of time preparing for retirement. What I ended up finding was that on average, men invest in their company’s pension plan approximately 6 years longer than women. Along with that both parties decrease the amount of money they actually invest in the pension plan as the wealth grows.

There were approximately 194 observations of men and women, 117 women to 77 men[[1]](#footnote-1), so I thought that the number would be in the woman’s favor. Of those 194 observations, women were more likely to have completed a bachelor’s degree and men were more likely to have obtained a Master’s degree.[[2]](#footnote-2) I also found that the amount of time spent in the pension plan was positively skewed.[[3]](#footnote-3) This could be due to people getting raises and investing in more stock, people could have switched companies a lot and there didn’t have time to invest in the pension, etc. Knowing that women on average do not spend as much time investing in their company’s pension plan, companies need to make it more attractive to women or get better literature out about their pension plane. The more a person pays into their pension, the better off a company should be and the higher likelihood that the person will stay with the company creating low turnover rates.

Since failing to accept the null hypothesis, I ran the Breusch-Pagan test and the White test to test for heteroskedasticity. I also estimated that “beta” coefficients to see if that would get me closer to the answer null hypothesis being accepted. Only at the beta level of the regression that the probability that males and females spend the same amount of time investing in pension plans What I found was that increasing the alpha coefficient against different levels made my beta coefficients smaller and closer to the mean. Before settling on my end equation I ran 2 other regressions to see if there was significance in the estimators. I ran histograms and scatterplots to see if heteroskedasticity existed. The model was homoskedastic.

**Model 1**

In my first regression analysis regressed pyears against female and wealth 89.[[4]](#footnote-4) Pyears was the number of years a person paid into their retirement plan, wealth89 being how much net worth a person had in thousands of dollars in 1989. Female was a binary dummy variable assuming that 1 is female otherwise zero for male. After running the regression I derived the following:

Holding wealth constant, men spend 16.46 year investing in their pension plan. That is approximately 9 more years than women. As wealth increases the amount of time decreases which as surprising because one would think as wealth increases so would the pension investment. The variables only accounted to 13.65% of the model[[5]](#footnote-5) which is not significant. The t-statistics of the model suggested that the female variable was significant but the wealth variable fell in the 95% CI. I interpreted that to mean that I should reject the model in favor of men spending more time investing in a pension plan versus their female counterpart.

**Model 2[[6]](#footnote-6)**

I reran the model this time adding the variable educ (for years of education) and switching wealth89 to the variable finc75 (dummy variable for men and women in the median income bracket of $35,000 to $50,000) where

The R2 in this regression was less than the initial R2 of the first model. I wanted to see the exact relationship between males and females in the median income bracket. The resulting equation being

Holding education and income constant, males seem to invest in the pension plan roughly 6 years longer than females. From model 1 to model 2 the amount of years investing in the pension plan decreased by 3 years (men investing roughly 9 more than women in model 1 and 6 in model 2). This was significant in that when specifying the income bracket, I was able to decrease the difference in the amount of years on average spent on investing getting the difference closer to zero. In the second model however, if I went off of the p-value alone, it seems that the only significant variable would be education and income because both are above .05. The model still failed to accept the hypothesis that on average men and women invested the same amount in a pension plan.

**Model 3: True Model [[7]](#footnote-7)**

I ran a third regression, this time including more dummy variables to first see if this would allow the R2 value to be significant. Second I ran this model to see if including more variables would allow me to accept my null hypothesis. I started with a simple equation

Where pyears is the number of years spent investing in a pension plan; female being a dummy variable where 1 is female otherwise 0 is male; ffinc50 is the dummy variable of the median income (between $35,000 and $50,000) multiplied by the gender female; ffinc100 being the dummy variable of a higher income (between $50,000 and $100,000) bracket; educ being the number of years of education; wealth89 being how much net worth a person had in thousands of dollars in 1989; finc100 representing a dummy variable specifying the person in the high income bracket; finc50 being a dummy variable specifying the person in the low income bracket; finc75 being the dummy variable for median income. Due to the fact that I needed a base level, I did not include my interaction term ffinc75. The final equation being

After running the first model, I was not at all shocked to see that education was positively valued because I assumed that the more education one achieved, the more likely one would get a better paying job that would allow them the extra income to save for retirement. Furthermore, I assumed that the more education that one achieved, the more aware of the benefits of savings that a person had available to them, thus allowing an informed decision about retirement savings options. After the first regression revealed that an increased wealth in wealth leads to a decrease in investing in the pension plan, I was not surprised to see the dummy variables for income to be negative. What was shocking was surprising was that holding all else constant, ffinc50 showed that if the woman is in the median bracket, they are likely to increase their investment of about 3 months. There is still a difference in the amount of investing. At the same income bracket, men spent roughly 9 more years investing than females. In looking at histograms, the income brackets men were more represented than females. There was a slight increase in female presence in the median income bracket, finc50.[[8]](#footnote-8)In all three of the graphs, men exceeded women in earnings.[[9]](#footnote-9) In order to see if there was any correlation between the variables, I ran a correlation matrix using stata. [[10]](#footnote-10) Looking at the matrix correlation there are positive correlations between education and income and wealth, however, negative correlation between incomes. It doesn’t appear to have significant correlation between any of the variables.

**Results**

Accepting the third model, all tests were performed based off of the regression

**Graphs[[11]](#footnote-11)**

In theory, histograms represent the underlying frequency distribution of the continuous data. The histogram for the PYears is positively skewed . [[12]](#footnote-12)The level of education is interesting because it has very few observations of people with less than a high school diploma (less than 12 years). [[13]](#footnote-13)In this case, one can assume that the people were likely to have obtained some secondary education and were able to make an informed decision about their investments. Also, it could have been assumed that the higher the education, the more money the person was expected to earn. There is a large number of people with just a high school diploma which is causing a spike in the graph at 12 then it drops dramatically with more of the results being around 18 (Master’s degree level complete assuming the standard 2 years after high school is an Associate’s degree, 2 years after that is a bachelors, 2 years after that a Master’s at 14, 16 and 18 years respectively). The shape of this curve would be more negatively skewed because there are more observations of people having gone on to finish the Master’s Degree. When I created a scatter plot of the residuals on the pension plan, they were linear and they were consistent and positive.[[14]](#footnote-14) When I graphed the predicted models, the histogram[[15]](#footnote-15) did reflect that the residuals were positively skewed. From looking at the scatter plots and the histograms I do not see heteroskedasticity. It does surprise me since most of my terms are binary dummy variables. It showed that both the scatter plots as well as the histogram for residuals backed up one another in interpretation and output.

**Testing the True Model**

In trying to prove my hypothesis, I ran a series of tests. I ran a Joint test,[[16]](#footnote-16) I ran the true model at the 1%, 5% and 10% alpha level using a stepwise regression in Stata. I performed the Robust regression, a beta coefficient test, The White Test and Breusch-Pagan Test. I also performed the Weighted Least Squared Test and Feasible Least Squared Test outlined by Wooldridge. The FGLS test was then rerun using the Stata FGLS.

I did not believe that my model is exhibiting heteroskedasticity, however, I still performed tests to ensure that heteroskedasticity was not present. The Robust Model[[17]](#footnote-17) was pretty close to my original model which is what I figured it would be. All of the coefficients exhibited the same positive and negative signs. Again, going back to the fact that many of my terms are binary dummy variables, I didn’t expect much heteroskedasticity. For the ordinal values and the non-ordinals values multiplied by the dummy variable, they did not seem to exhibit signs of heteroskedasticity. Both the White Test [[18]](#footnote-18)and the BP tests [[19]](#footnote-19)are dropping a significant amount of observances. In the model there are 191 observations with a df of 182 for the F-Test and 190 otherwise yet the White Tests and BP tests are significantly lower (see table). In looking at the BP test, the P-value is very small (.0019) thus according to this test the null hypothesis is rejected and heteroskedasticity is present. According to the White Test the P-value is .7374. This test would have us not reject the null hypothesis. An interesting note in the degrees of freedom: for the BP test the Chi2 is 1 whereas in the White’s test the Chi2 is 24. In this case in trying to prove that heteroskedasticity is not present I would reject the BP in favor of the White Test as a measure for heteroskedasticity. With the exception of female, educ and wealth89, all of my t-statistics fell within in the normal confidence interval. The F-Statistic [[20]](#footnote-20)was significantly low while the Residual Sum of Squares remained high. With the exception of female, educ and wealth89, all of the p-values were slightly above .05.

I reran the regression and did a stepwise regression at the 1%, [[21]](#footnote-21)5% [[22]](#footnote-22)and 10% levels. What I found was that at the 1% level everything was removed except for the dummy variable Female. At the 5% level, everything had been removed except for females. At the 10% level female, ffinc50, ffinc100, educ and wealth89 are included in the model. Only at the 10% level is the R2 anywhere near 15%. [[23]](#footnote-23)It keeps increasing as the alpha increases.

I reran the regression [[24]](#footnote-24)with beta coefficients [[25]](#footnote-25)and it really was no difference. The slope parameter was negative, which is still significantly less than one and holding all things constant would favor the male. If I were to input 0 for female, 1 for ffinc50 and 1 for education , the regression model moves closer to 0 zero which makes me believe that the only model that could allow for my hypothesis to be true that males and females tend to invest the same amount of time for retirement. It is only under this model that I could with a little shred of light state that the null is true. There are no significant amount of deviation, unfortunately that could prove the null to be true on all occasions. What could assist in the null being true is if there was more of a sample population. The larger the population, the more likely that the null is true. Furthermore, as we input 1 for female, the regression becomes The null is still being rejected due to the fact that men tend to put more into the pension plan than women certificates of deposit. The beta coefficients stated that for 1 unit positive standard deviation change in X is expected to result in a negative beta coefficient change in Y meaning that in the binary model for female, it would be a negative standard deviation in the pension. Mostly the beta coefficients favor the male and reinforce that as wealth increases for both male and female, there would be a negative standard deviation.

**FGLS and WLS**

Both the FGLS and WLS Tests were run under the Wooldridge steps in Stata[[26]](#footnote-26). In step 5 of the Wooldridge FGLS test, it was stated that model, after going through steps 1- 4, should use the equation 8.27 which stated that the

In order to obtain this I squared each of the original values and the divided each one by **.** For the WLS[[27]](#footnote-27), the assumed was given as var(u|x)=σ2age. Under the WLS model the t-stat for female now lies within the 95% CI. WeightWealth89[[28]](#footnote-28) remains significant. The confidence intervals for this model seem to have increased. Weighteduc is still significant. The standard errors have significantly decreased compared to the non-weighted least squares model. The confidence Intervals for the WLS seem to have become significantly larger making in increasingly hard for the t-stats to lie outside of them. That increase would make it increasingly hard to accept the null hypothesis. The R2 has increased to 23% roughly and so has the adjusted r squared. If I were to make weightfemale equal to zero for male and make weightffinc50 equal to 1 and weightfinc50 I would get 0.286 (-.04-0+1.00-.674) which is closer to zero and in favor of the males as opposed to the female.

I ran the FGLS using the stata command and the Wooldridge steps in the Introductory to Econometrics book. The Wooldridge Test differed enormously than the stat test.

1. The female coefficient for the Wooldridge test is significantly higher than that of the Stats FGLS tests. The Stata FGLS test is closer to the True Model.
2. The Confidence intervals for the Wooldridge test is significantly higher in range than the stata test. For example, the 95% confidence interval for the fglsffinc50 is between -497.6779 and 478.9545.

This over inflated model would have to be reviewed to see if the input steps were correct. I would fail to accept the Wooldridge model because the numbers are grossly overestimated. I would however, use the stata model that rejects the hypothesis because it is closer to the True Model.

**Conclusion**

The gaps in 1989 on women and men’s earnings have definitely not changed. It would be interesting to see how much the gap between men and women’s income have changed from then until now. Was still holds true is that women need to be more conscious about retirement just as much as men do. What I didn’t explore was the effect of marriage on the equation. Married women could allow their spouse to pay into their plan because it yields better savings. Another possible explanation for the gap could be the amount of maternity leave a woman was able to take. A pregnant married woman could have the option to stop working all together after the baby comes which would affect the longevity of the pension. If age could have been included, the maybe this analysis could have been run as a quadratic equation on age looking at the minimum average time a person started paying into the pension and the max time they paid into the pension.

I ran three regressions, the first being pyears on female and wealth. The second being pyears on female, educ and finc75. The third being pyears on female, ffinc50, ffinc100, educ, wealth, finc100, finc50, finc75. Due to the fact that were more females than males in observation, I expected the results to be a little higher than males because we have more females than males. In all 3 regressions the null hypothesis was to be rejected due to the fact that men in every earnings bracket out saved the women the women. In looking at my true regression model, I think that I may have included too many dummy variables that may be multicollinear, causing the model to be void. Knowing which dummy variable and interaction variables to include and exclude was a bit ambiguous. Possibly including the dummy variable ffinc75 may have some significance.

In all 3 models the R2 does not exceed 15%. I ran it with different dummy variables. I removed some dummy variables and all of the remained low. I concluded that the variables used result in me failing to accept the null hypothesis, thus there are many other factors that can account for the number of years in the pension plan. It was only at the beta level of the regression that the probability that males and females spend the same amount of time investing in pension plans. With an increase in population, the null could at some point be true.

**Tables and Models**

**Graph 1**



**Graph 2**

**Table 1**

**Model 1**



**Model 2**

****

**Model 3: True Model**



**Correlation Matrix**

**Graph 3**



**Graph 4**



**Graph 5**



**Regression @ 1%**



**Regression @ 5%**



**Regression @ 10%**



**Beta Model**



**Robust Model**



**Residuals Graph**

**Scatter Residuals**



**White Test**



**Breusch-Pagan Test**



**Joint test**



**Wooldridge Weighted Least Squares Test**

**Wooldridge FGLS Test**

****

**Stata FGLS Test**



1. Table 1 [↑](#footnote-ref-1)
2. Allowing 12 meaning high school complete and no further schooling, 14 an Associate’s degree, 16 a bachelor’s degree and 18 a Master’s degree. [↑](#footnote-ref-2)
3. Graph 1 [↑](#footnote-ref-3)
4. [↑](#footnote-ref-4)
5. Model 1 [↑](#footnote-ref-5)
6. Model 2 [↑](#footnote-ref-6)
7. Model 3 [↑](#footnote-ref-7)
8. Graph 5 [↑](#footnote-ref-8)
9. Graph 3, Graph 4, Graph 5 [↑](#footnote-ref-9)
10. Correlation Matrix [↑](#footnote-ref-10)
11. These graphs are obtained at the end of the paper. [↑](#footnote-ref-11)
12. Graph 1 [↑](#footnote-ref-12)
13. Graph 2 [↑](#footnote-ref-13)
14. Scatter Residuals [↑](#footnote-ref-14)
15. Residuals Graph [↑](#footnote-ref-15)
16. Joint Test [↑](#footnote-ref-16)
17. See Robust Model [↑](#footnote-ref-17)
18. White Test [↑](#footnote-ref-18)
19. See Breusch-Pagan Test [↑](#footnote-ref-19)
20. See Breusch-Pagan Test [↑](#footnote-ref-20)
21. See Regression @ 1% [↑](#footnote-ref-21)
22. See Regression @ 5% [↑](#footnote-ref-22)
23. See Regression @ 10% [↑](#footnote-ref-23)
24. Each time the regression was ran, I started with the original regression except in the case of the Wooldridge WLS and FGLS tests. In those cases I followed the steps in the book. [↑](#footnote-ref-24)
25. See Beta Model [↑](#footnote-ref-25)
26. To see the exact steps for both the WLS and FGLS refer to Wooldridge, Introductory Econometrics: A modern Approach 4e page 292 for the WLS and page 283 for the GLS [↑](#footnote-ref-26)
27. All of the variables in this model have been named weight(Var) because of step 5 in the Wooldridge book. Each were regressed using the given by **=** var(u|x)=σ2age [↑](#footnote-ref-27)
28. See Wooldridge Weighted Least Squares Test [↑](#footnote-ref-28)