Assignment 3: LR Report – Unemployment Model

Predict 411

Section 56

Winter Quarter

****

School of Continuing Studies

Northwestern University

****

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

****

Program Analyst

Wooddale Church

6630 Shady Oak Road

Eden Prairie, MN 55344

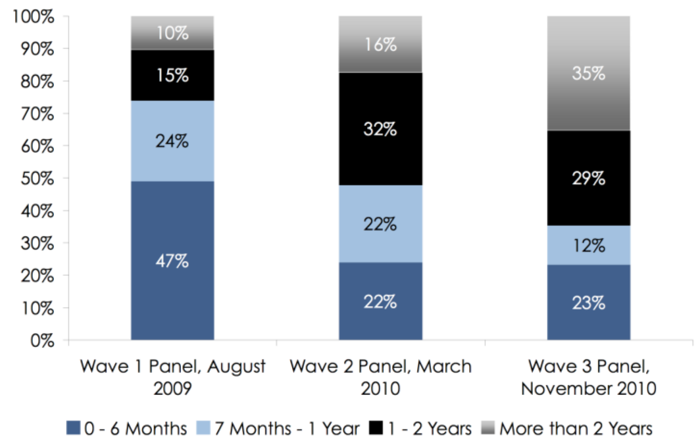
Executive Summary

Unemployment benefits are a safety net that prove invaluable when received. Ascertaining who receives these benefits is a hotly debated issue, which requires an astute skill set to dissect. In this study 4,877 individuals were studied, of which 3,335 received unemployment benefits. The response variable was a binary outcome of either 1-received benefits or 0-no benefits. In the exploratory data analysis (EDA), there were 19 variables. 12 variables were also binary and 5 were continuous. Three variables proved to be the most predictive and statistically sound for receiving unemployment benefits and were slack, married, and age. Through exploring the variables, some were highly predictive but lacked statistical strength, while others were statistically significant but lacked predictive power. The major conclusion drawn from this EDA is that the unemployment system in application is not reaching its intended audience for receiving unemployment benefits. The system is rewarding individuals that were fired, while neglecting individuals with kin. As a result of these findings, I recommend a stringent review of the qualifying factors used to discern the administration of unemployment benefits.

Introduction

From 2007 – 2009, the Great Recession claimed 8 million jobs and left over 14 million individuals unemployed (VanHorn & Zukin 2010). In a survey conducted by the Heldrich Center in 2010, 73 percent of surveyed Americans stated they had “firsthand experience” with the recession. The graphic below, created by Rutgers University, displays the length of unemployment throughout The

FIGURE 1. DURATION OF UNEMPLOYMENT



Great Recession, and the total percent of individuals that fall into each prospective category.

As a result of losing a job, one can apply for unemployment benefits which are provided through the state (about.com). Eligibility is based on how much money was earned during employment and the reason for unemployment. Ineligibility for unemployment benefits included:

"

- Quit without good cause
- Fired for misconduct
- Resigned because of illness (check on disability benefits)
- Left to get married
- Self-employed
- Involved in a labor dispute
- Attending school

" – about.

The objective of this report is to explore the relationship between the binary response variable y, did or did not receive unemployment benefits, and a list of potential explanatory variables by conducting a thorough analysis using Logistic regression. Listed below is my first-take on the relationship between the dependent variable and the independent variables before analyzing the data.

Explanatory Variable	**Projected Relationship with Y**
Age is the age of the subject	Older individuals will have a greater chance of receiving unemployment
Age2 is the square of Age divided by 10	This variable will mirror Age but have a better distribution
Tenure is the years of tenure at the last job	The longer the tenure the greater chance of receiving unemployment
Slack is an indicator variable that equals 1 if the subject was fired because of poor performance X	The result of being fired will be no unemployment, an inverse relationship with receiving unemployment
Abol is an indicator variable that equals 1 if the subject's position was eliminated X	As a result of the position being eliminated, there will be greater chance for receiving unemployment
Seasonal is an indicator variable that equals 1 if the subject was a temporary worker	Temporary workers will not receive unemployment

NWHITE is an indicator variable that equals 1 if the subject’s age is non-white	Non-white subjects will have a greater chance of receiving unemployment
School12 is an indicator variable that equals 1 if the subject has more than 12 years of education	More than 12 years of education will not affect unemployment benefits
Male is an indicator variable that equals 1 if the subject is male	If male, the relationship with Y will be negative
SMSA is an indicator variable that equals 1 if the subject lives in a SMSA	SMSA is undefined, therefore it is illogical to explain
Married is an indicator variable that equals 1 if the subject is married	Married (1) will have a positive correlation with Y
DKIDS is an indicator variable that equals 1 if the subject has kids	DKIDS (1) will have a positive correlation with Y
DYKIDS is an indicator variable that equals 1 if the subject has young kids	DYKIDS will not have a correlation with Y
YRDISP records the year when the job was lost-here, 1982=1 and 1991=10	YRDISP will not have a strong correlation in either direction based on its arbitrary status
RR is the replacement rate that is the ratio of the benefits received versus the last recorded weekly earnings	RR will not correlate with Y based on the binary structure of Y
RR2 is the square of RR	Same as RR
Head is an indicator variable that equals 1 if the subject is the head of the household	Head will have a positive correlation with Y based on sole income status
StateUR is the state unemployment rate	As the unemployment rate increases Y will decrease
StateMB is the maximum benefits available for a given state	As StateMB increases so will Y

The preamble in the United States Constitution states that the government’s task is to ensure domestic justice. Through analyzing this study, government officials can better assess the efficacy of specific public policy in meeting the needs of its people. As a result of this study, economists will better understand demographic information that correlates with unemployment percentages which in turn directly impacts the economy as a whole.

Analysis

In order to meet the objective of exploring the relationship between the dependent variable and independent variables, an exploratory data analysis must be conducted. This EDA will use specific techniques intended to work with data that has a binary response variable. I will be utilizing the EDA paradigm and structure put forth by Bruce Ratner found in his book *Statistical and Machine-Learning Data Mining*:

> Problem/Objective: Explore the relationship between Y. The response variable takes a value of 1 if the unemployed worker received unemployment benefits, and there are 19 independent variables. Management has stipulated that a Logistic regression model be built for this data set.
>
> Data: The data has been aggregated and has been supplied from management. There are no missing values. Please note that there is a binary response variable and binary explanatory variables.
>
> Analysis: In SAS, a PROC MEANS statement with a class statement will be used to assess the predictive accuracy of individual attributes with the response variable. Highly predictive attributes will be used in the Logistic regression model.
>
> Model: After assessing the data, a model will be used. Management has stipulated that a Logistic regression model be built for this data set.
>
> Results/Interpretation: Once the model has been created, an assessment of the model adequacy will be conducted to discern how well the variables predict Y.

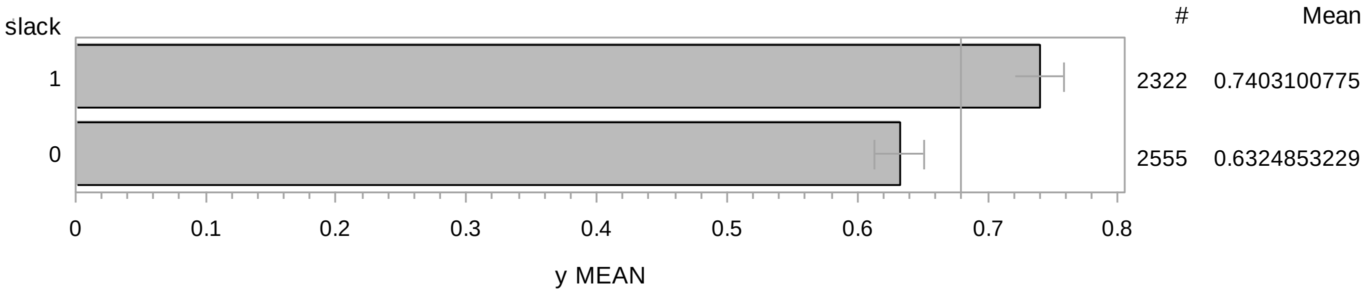
A properly executed EDA for management must reflect that the data was the driving force behind constructing the model. The steps outlined above are ordered such that the data drives building the best Logistic regression model, and the analyst's personal bias is mitigated.

Data

There are a total of 4,877 observations with 0 completely missing values per row. Management has requested to focus on 20 of the 22 variables in this data set. The response variable has a binary outcome and will be coded as a dummy variable. Given that the focus of the EDA is on individuals that received unemployment benefits, this response will be coded as 1 and 0 equals not receiving unemployment benefits. There are independent variables that are binary and their coding has been previously discussed above. Below you will find general descriptive statistics of the variables and their correlation with the response variable.

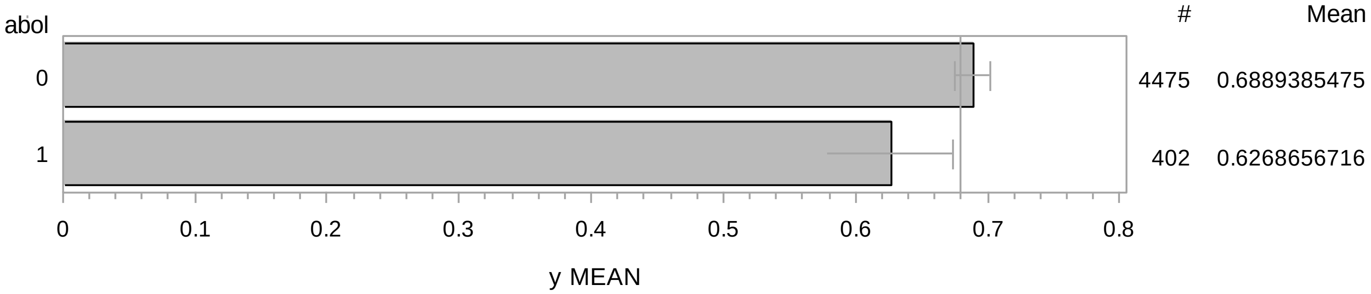
In this data set, there are 4,877 observations. Of the observations, 3,335 observations received unemployment benefits, which is a .6838 probability. When analyzing the variables, this probability is the bench march for which predictive probability will be compared against. I am looking for variables that have binary averages that are polar to this base probability.

Appendix 1 displays the proof for the descriptive statistics of the binary outcome Slack. I will use this logic for all the other binary outcome variables. Slack is an indicator variable that equals 1 if the subject was fired because of poor performance. Out of 4,877 observations, 2,322 were let go as a result of poor performance. Of the 2,322 observations that were fired for poor performance, 1,719 received unemployment benefits. Dividing 1719/2322 = .7403 which as a percentage equals 74%, and demonstrates the probability that one that was fired would receive unemployment benefits. In the data set, 2,555 individuals fell into this category. If an individual that was let go for a reason other than poor performance, the response is 0. Of the 2,555 individuals 63% (1615/2555) received unemployment benefits. On a personal note, this data reveals that if an individual was fired for poor performance (s)he has a greater probability of receiving unemployment benefits than an individual that was let go for reasons other than poor performance. From this insight, it can be said that the unemployment system is rewarding those that were fired for poor performance with a higher probability of receiving unemployment benefits. This is a very important insight that I would bring up to management. The horizontal bar chart below demonstrates all the above calculation, and in addition visually demonstrates the 95% confidence intervals.

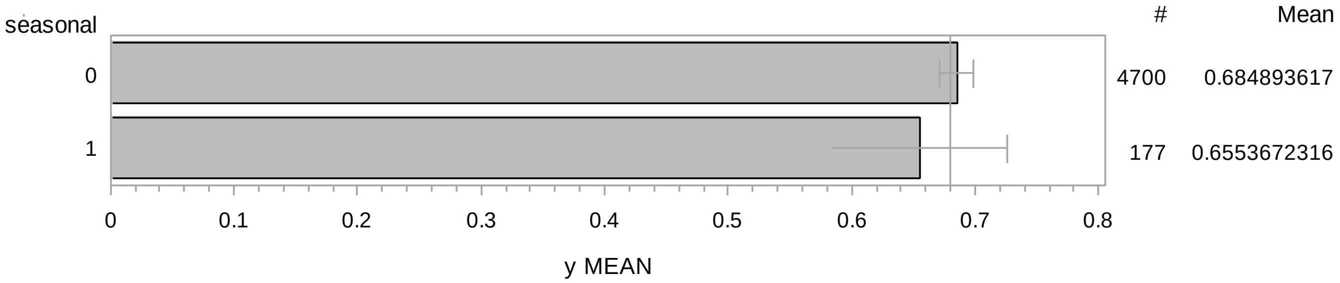


Given the calulations described above, I will descrive the other binary independent variables. This will be demonstrated via the horizontal bar chart. Please refer to Appendix 1 to see the proof for the calulations.

Abol: An indicator variable that equals 1 if the subject's position was eliminated. Visually it can be seen that with the 95% confidence interval the variables might have some overlap. As a result of the overlap, I would not want to use this variable in the model. From an unemployment program (UP) stand point, one would want the system to have empathy on individuals that lost their job as a result of it being eliminated. Yet, the UP is rewarding individuals that did not lose their job as a result of it being eliminated. In short, this is a problem that government administrators need to fix, although the implication is normative.

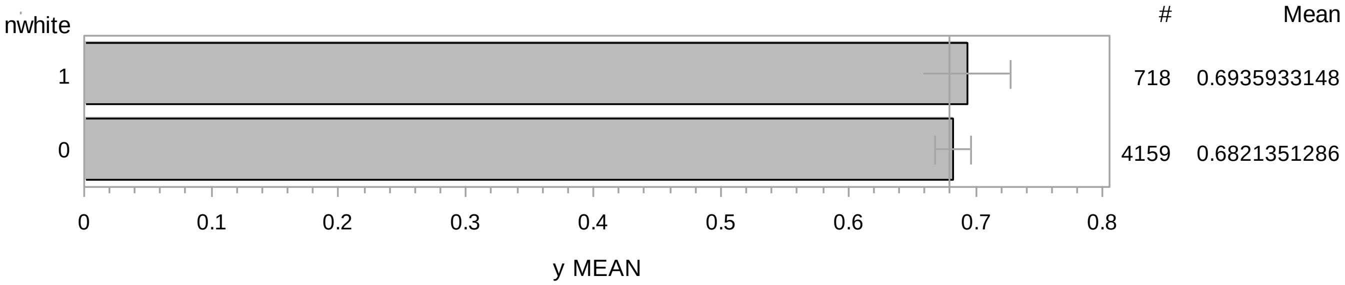


Seasonal: An indicator variable that equals 1 if the subject was a temporary worker. The UP is set up in a way that as an individual builds tenure it increase the amount and probability of receiving unemployment. This variable should be a no-brainer in that seasonal workers would not enter the UP. But, the results reveal a startling contrast from the UP theory. As demonstrated below, seasonal workers are about as likely to receive unemployment benefits as non-season workers. There is a huge disparity between the numbers of observations between this variable, which could skew the conclusions. This is not a strong variable to consider using in the model.

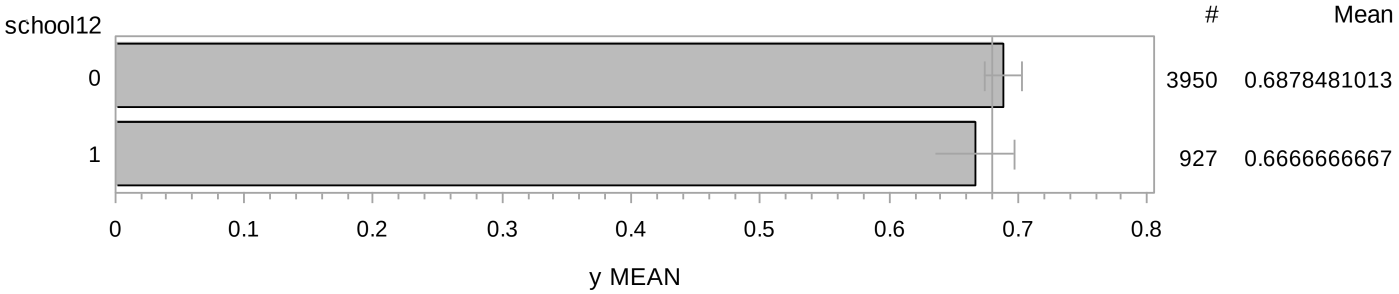


From this data, I am starting to see a trend in that the theory of the UP is starkly different from the actual execution of the UP.

Non-White: Indicator variable that equals 1 if the subject’s ethnicity is non-white. This variable also has a large disparity between the binary outcomes. Only 14 % of the total observations are non-white. This study is done in a specific geographical area. If the purpose of this study is to draw greater conclusions than just this specific area, the skews of each variables need to match up with the greater areas demographic percentages. The variable is not predictive one way or the other in regard to predicting the acceptance of unemployment benefits. I would not use this variable in the model.

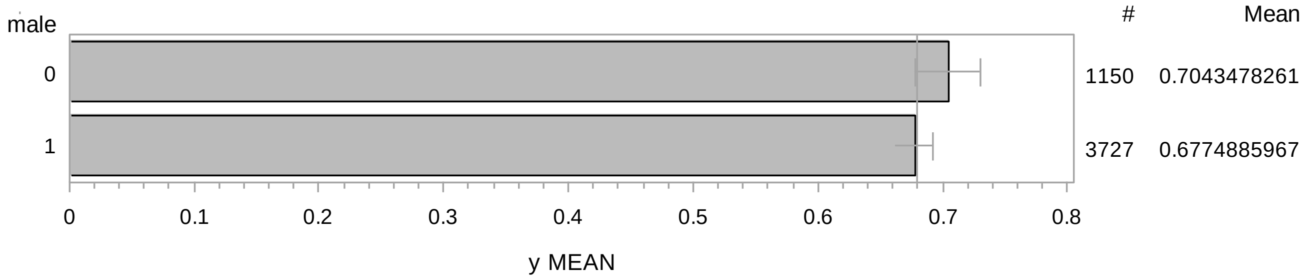


School 12: An indicator variable that equals 1 if the subject has more than 12 years of education. 19% of the observations had an education greater than 12 years. The initial analysis from the bar graph shows that higher education is not a factor in receiving unemployment benefits. But, an

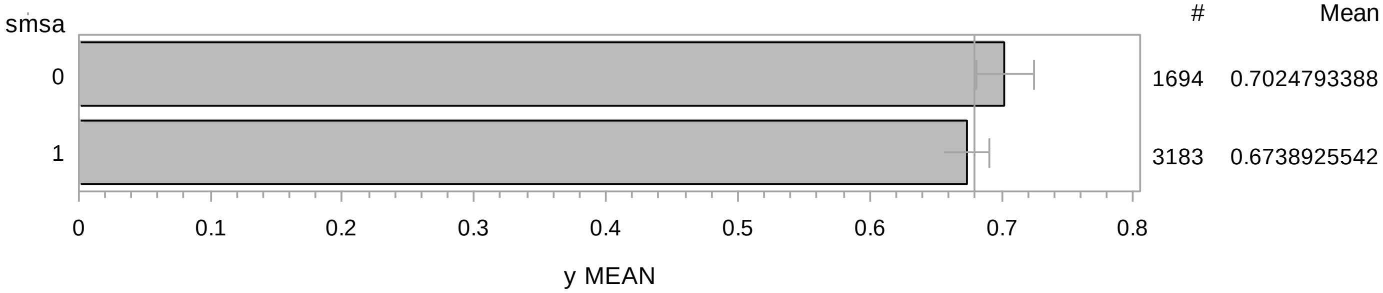


additional insight is that those with higher education were less inclined, as a percent, to be unemployed than the greater population of observations.

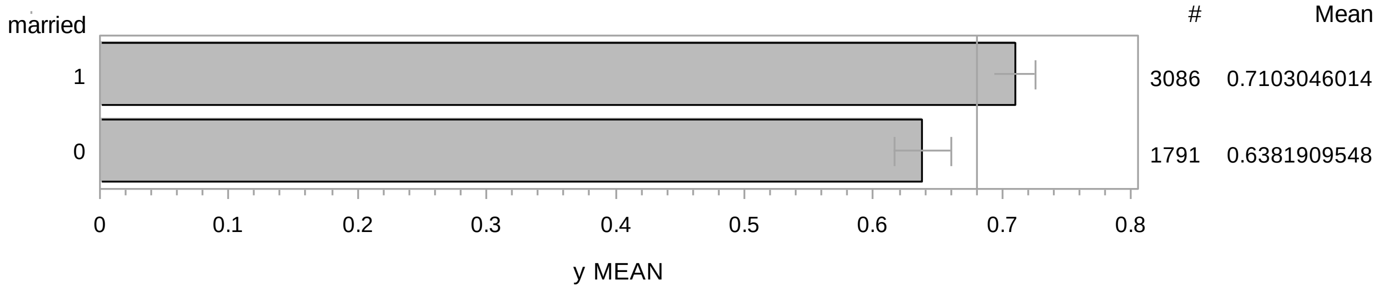
Male: An indicator variable that equals 1 if the subject is male. The distributions between both responses for this variable are better than the past three variables. There is a slight difference between being male verses non-male, but once the 95% confidence intervals (CI) are taken into consideration the variable becomes less predictive.



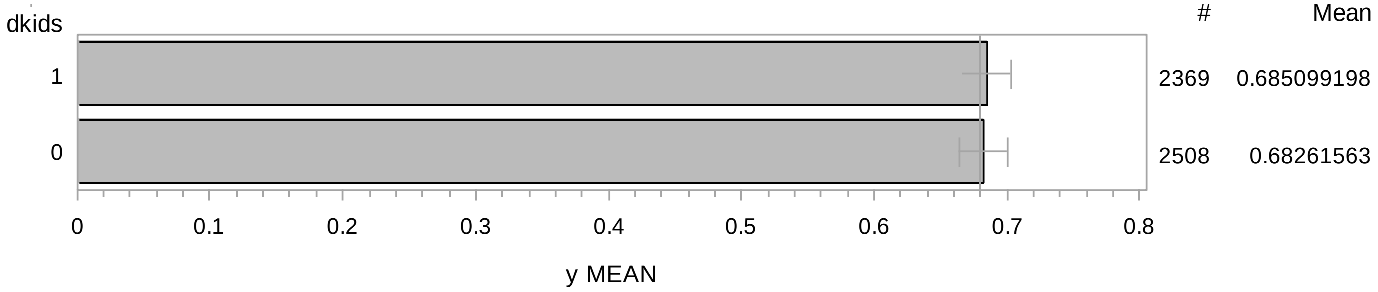
SMSA: An indicator variable that equals 1 if the subject lives in a SMSA. The distributions between both responses for this variable are rather equal. Like many of the variables in this model, there is no real distinguishing factor between either of the responses, and I would not use this variable in a model.



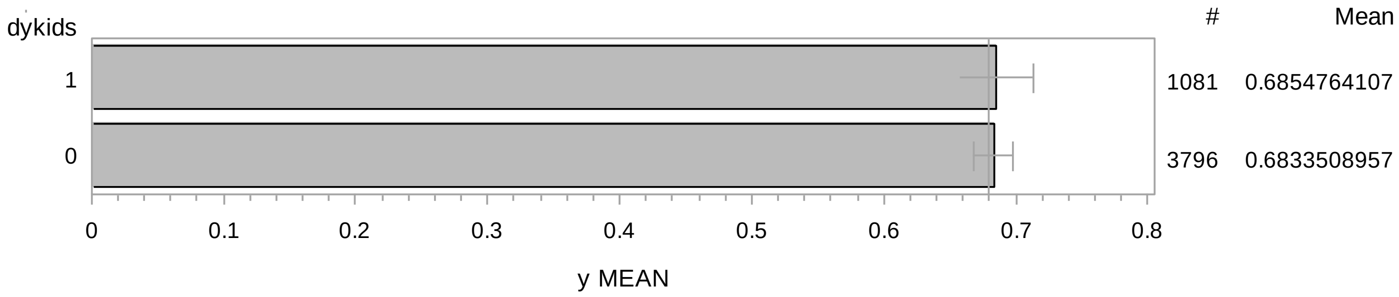
Married: An indicator variable that equals 1 if the subject is married. The distribution looks reasonable. There is s large enough gap between the outcomes for me to identify this variable as predictive, and warrant using it in the model.



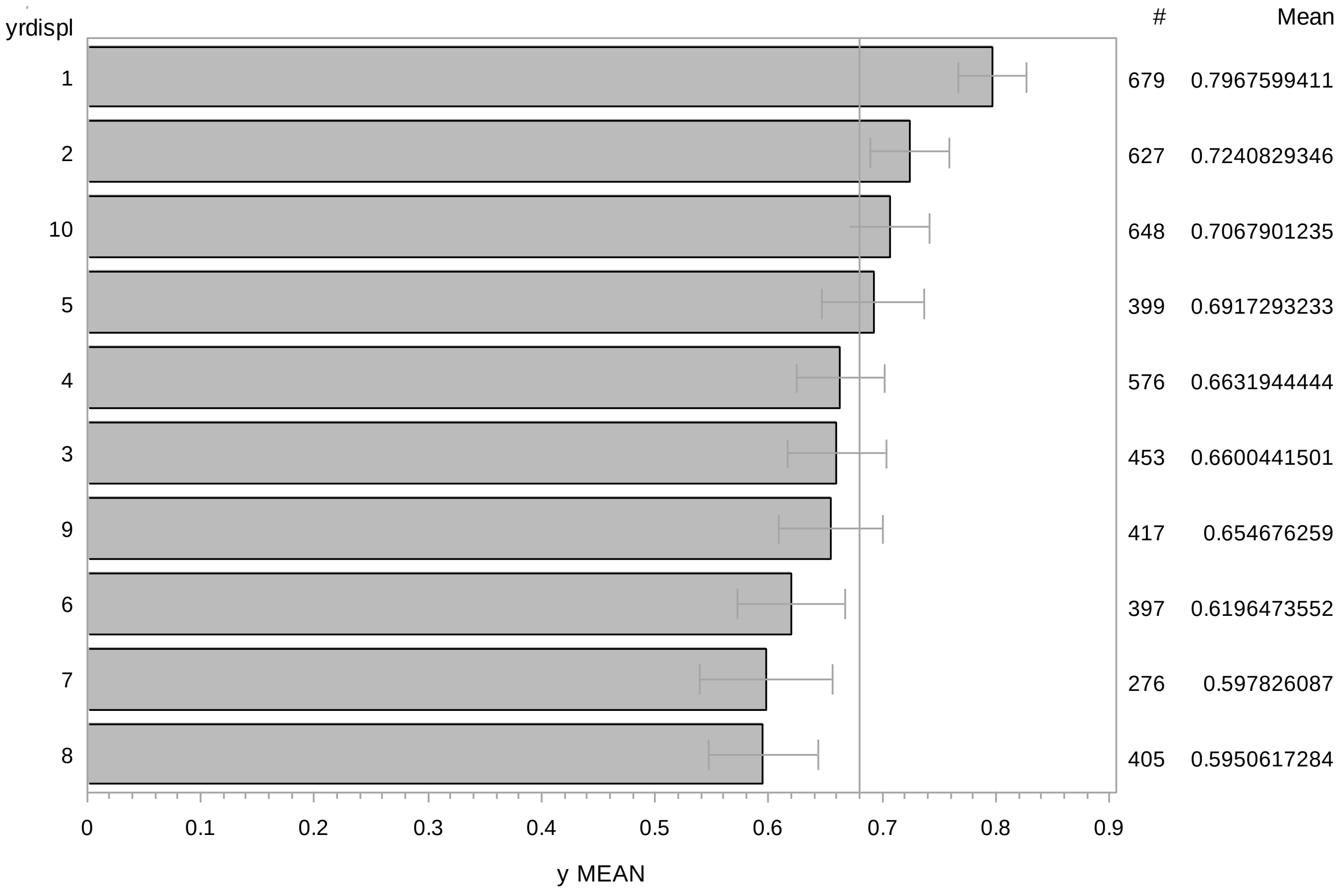
DKIDS: An indicator variable that equals 1 if the subject has kids. Notice how the distribution is almost even between the two outcomes. But, neither outcome is predictive and should not be used in the model. This is another example of how the UP is not reaching those in greater need.



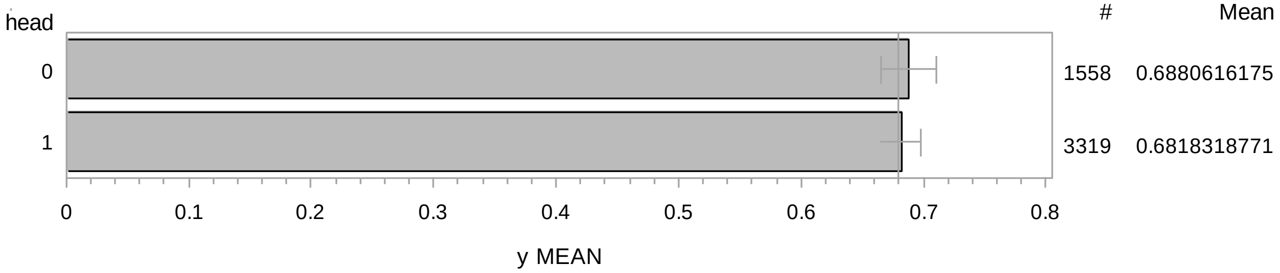
DYKIDS: An indicator variable that equals 1 if the subject has young kids. This variable has the same outcome as DKids, except for the fact that fewer observations had young kids.



<u>YRDISP:</u> Records the year when the job was lost-here, 1982=1 and 1991=10. The number of observations between each variable is a rather equal distribution. From the results, it would appear that as time increased the probability of receiving unemployment decreased. I would expect this variable to be correlated with other variables that are time sensitive.

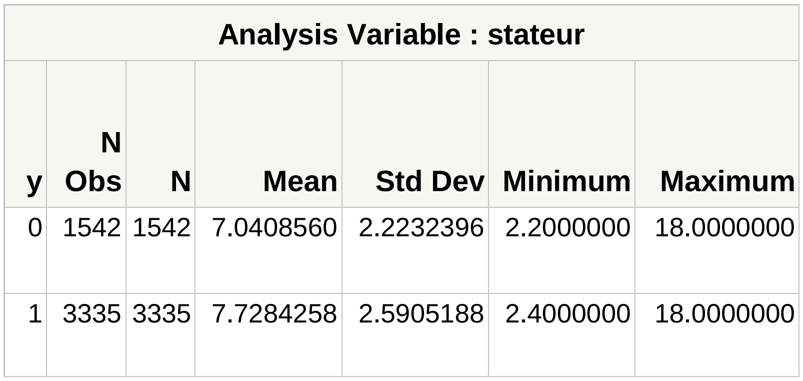


Head: An indicator variable that equals 1 if the subject is the head of the household. This variable has little predictive output and should not be used in the model.



Continuous Variables:

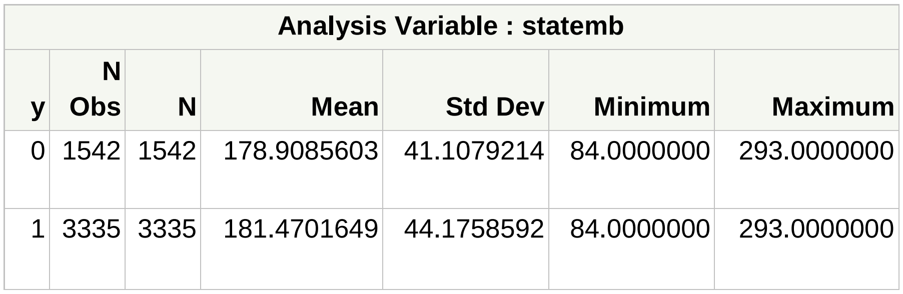
StateUR: Is the state unemployment rate. When analyzing a continuous variable with a binary response variable, one needs to analyze the mean between the binary outcomes. My assumption for this variable was stated as the unemployment rate increases, Y will decrease. Thus, I would expect response 1 to have a lower mean than 0. The data reflects a slightly higher mean for 1, which is the opposite of what I expected. However, both outcomes have a very similar mean along with similar standard deviations, which leads me to believe that StateUR is a poor predictor for receiving unemployment benefits.



Analysis Variable : stateur

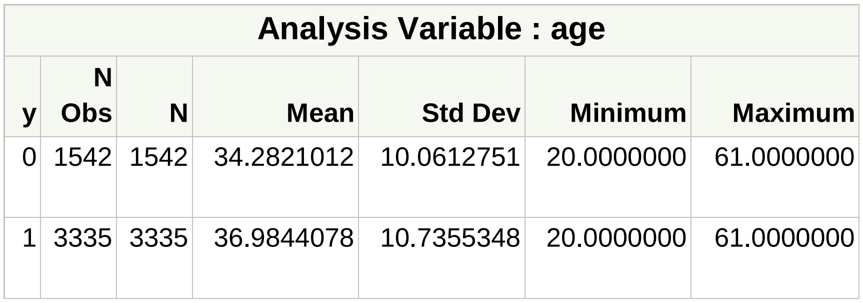
y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	7.0408560	2.2232396	2.2000000	18.0000000
1	3335	3335	7.7284258	2.5905188	2.4000000	18.0000000

StateMB: The maximum benefits available for a given state. My assumption is as StateMB increases so will Y. In order to see this demonstrated, I would have expected the mean for binary outcome 1 to be much higher than binary outcome 0. In reality, the data reflects a similar mean between both variables along with large standard deviations that result in no real discernment between either variable.



Analysis Variable : statemb						
y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	178.9085603	41.1079214	84.0000000	293.0000000
1	3335	3335	181.4701649	44.1758592	84.0000000	293.0000000

Age: The age of the subject. The ages between each outcome do not differ that greatly, which results in a poor prediction of Y. There is about a 2 year increase between receiving unemployment benefits, but this is not large enough to warrant using this variable as a strong predictor variable.



Analysis Variable : age						
y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	34.2821012	10.0612751	20.0000000	61.0000000
1	3335	3335	36.9844078	10.7355348	20.0000000	61.0000000

When the variable is squared, it generates the same output with the exception of all variables being squared. In addition, the variable is divided by ten. This produces a bigger gap between the variables, which leads me to include this variable as strong predictor variable.

Analysis Variable : age2

y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	127.6426070	77.0177510	40.0000000	372.1000000
1	3335	3335	148.3063568	86.0000614	40.0000000	372.1000000

Tenure: The years of tenure at the last job. This variable demonstrates that as tenure increases so does the likelihood of receiving unemployment benefits. But, the standard deviations muddy the waters of clarity such that this variable is not that predictive.

Analysis Variable : tenure

y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	4.6757458	5.6278644	1.0000000	41.0000000
1	3335	3335	6.1211394	6.5098112	1.0000000	40.0000000

Replacement Rate: The replacement rate of the benefits received versus the last recorded weekly earnings. This variable demonstrates very little difference between the binary outcomes and is not a predictive variable.

Analysis Variable : rr

y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	0.4436315	0.1069243	0.0386100	0.6690648
1	3335	3335	0.4359320	0.1057829	0.0782609	0.6911765

When the variable is squared it demonstrates the same output as rr, but the figures are larger.

Analysis Variable : rr2						
y	N Obs	N	Mean	Std Dev	Minimum	Maximum
0	1542	1542	0.2082343	0.0856315	0.0014907	0.4476477
1	3335	3335	0.2012234	0.0841274	0.0061248	0.4777250

Results

From the data analysis, management has encouraged to start with a Logistic Regression model. Through using Proc Logistic in SAS, the following analysis will be a brief overview of the data. In order to better understand the SAS output, I am comparing the full model (FM) to a model that I would prefer based on my initial EDA. My model (MM) has the following variables: slack, abol, male, married, yrdispl, age, tenure, and age 2. The goal of comparing these two models is to better understand the outputs and how they evaluate a model.

Model Information		Description
Data Set	WORK.WEEK3	This is the data set.
Response Variable	y	The data is compared with this variable.
Number of Response Levels	2	This is a binary response.
Model	binary logit	
Optimization Technique	Fisher's scoring	This is the method used to maximize the likelihood function.

		Description
Number of Observations Read	4877	This is the total data set.
Number of Observations Used	4877	All the observations were used.

MM Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	6088.056	5872.561
SC	6094.548	5930.992
-2 Log L	6086.056	5854.561

FM Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	6088.056	5786.393
SC	6094.548	5916.239
-2 Log L	6086.056	5746.393

Analyzing the inferential statistics AIC, SC, and -2 Log L are informal methods to assess the model fit. All of these statistics can be used to compare different sets of variables. Higher values for these statistics mean a worse fit to the data. It can clearly be seen that model FM has lower values for all three statistics. Of the three statistics, the -2 Log Logarithm is the most important. This statistic is the maximized value of the logarithm which is derived from the likelihood function times -2 (Allison 2012). The statistic is impacted by the number of observations, and naturally produces a better score for models that have more covariates. Like in Ordinary Least Squares, AIC and BIC are used to penalize models that have more variables. Although FM has more variables, it still has lower scores than the model with fewer variables.

MM Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	231.4948	8	<.0001
Score	224.5101	8	<.0001
Wald	214.6478	8	<.0001

FM Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	339.6629	19	<.0001
Score	326.3187	19	<.0001
Wald	305.0800	19	<.0001

The Global Null hypothesis tests that all the explanatory variables have coefficients equal to zero. It can be seen that both models have at least one coefficient that does not equal zero. Both models also have a significant p-value. FM has much higher scores, and it should be noted that this model was considered the best. As a result of these preliminary findings, I am going to investigate only model FM from this point.

Interpreting coefficients for Logistic Regression (LR) is not as straight forward as in Ordinary Least Squares Regression. In LR a logit coefficient of .5 can be interpreted as .5 log odds increase for every 1-unit increase in the explanatory variable, assuming all the other coefficients are held constant (Allison). It is really hard to conceptualize a .5 log-odds increase, which can be explained by the fact that LR captures a nonlinear relationship. From the Maximum Likelihood Estimates output, what can be gleaned is the statistical significance as well as the sign of the estimate. A positive or negative sign indicates the direction of the relationship (uoregon.edu). In addition to assessing the sign of a coefficient, analyzing the statistical significance is important. P-values assess the probability that your sample results are chance or extreme given that the null hypothesis is true. As the p-value increases, the probability increases that the sample estimate is based on pure chance. Lower p-values are an indicator of a statistically solid coefficient. The log-odds ratio is a far better output for understanding the coefficients, and will be explained in the next paragraph.

I prefer to assess variables at the 99% level, which equals a p-value of .01 or less.

FM Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Description of Variable
Intercept	1	-2.8005	0.6042	21.4861	<.0001	NA
rr	1	3.0681	1.8682	2.6969	0.1005	This variable is not statistically significant.
rr2	1	-4.8906	2.3335	4.3924	0.0361	This variable is not statistically significant.
age	1	0.0677	0.0239	8.0169	0.0046	Statistically significant Positive relationship
age2	1	-0.00597	0.00304	3.8585	0.0495	This variable is not statistically significant
tenure	1	0.0312	0.00664	22.1189	<.0001	Statistically significant Positive relationship
slack	1	0.6248	0.0706	78.2397	<.0001	Statistically significant Positive relationship The large estimate is indicative of an influential predictor.
abol	1	-0.0362	0.1178	0.0943	0.7588	This variable is not statistically significant
seasonal	1	0.2709	0.1712	2.5042	0.1135	This variable is not statistically significant
head	1	-0.2107	0.0812	6.7276	0.0095	Statistically significant Negative relationship
married	1	0.2423	0.0794	9.3075	0.0023	Statistically significant Positive relationship The relatively large estimate is indicative of an influential predictor.
dkids	1	-0.1579	0.0862	3.3552	0.0670	This variable is not statistically significant
dykids	1	0.2059	0.0975	4.4601	0.0347	This variable is not statistically significant
smsa	1	-0.1704	0.0698	5.9598	0.0146	This variable is not statistically significant
nwhite	1	0.0741	0.0930	0.6349	0.4255	This variable is not statistically significant
yrdispl	1	-0.0637	0.0150	18.0409	<.0001	Statistically significant Negative relationship
school12	1	-0.0653	0.0824	0.6270	0.4285	This variable is not statistically significant
male	1	-0.1798	0.0875	4.2204	0.0399	This variable is not statistically significant

FM Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Description of Variable
stateur	1	0.0956	0.0159	36.1127	<.0001	Statistically significant Positive relationship
statemb	1	0.00603	0.00101	35.6782	<.0001	Statistically significant Positive relationship

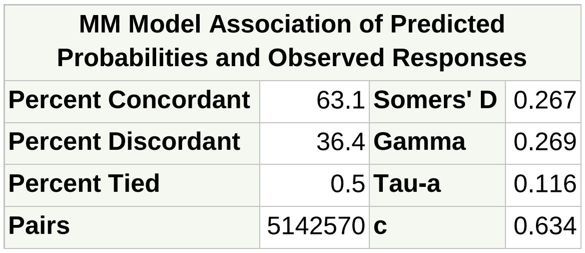
Of the 19 variables, 8 proved to be statistically significant and warrant further investigation of influence on the response variable.

The odds ratio is simply computed by taking the natural log of e^estimate. In addition to the odds, this calculation is also adjusted since it controls for other variables. Once the odds ratio has been calculated, the output is easier to understand. For example, the odds ratio for married can be calculated as 2.7182 (e) ^ .2423 = 1.2741. This can be interpreted as an individual that is marries has 1.27 times the odds of receiving unemployment benefits than non-married individuals. I personally prefer probability to odds. This can be translated from the following calculation, probability = 1.2741/(1 + 1.2741) = .5602. Therefore, a married person has a 56% greater chance of receiving unemployment benefits than an unmarried person. Below is a break down for each coefficient and the corresponding odds ratio, in addition I calculated the probability. See appendix 2 for the proof done in excel. Please note, the statistical validation was interpreted in the preceding section, and this section is simply interpreting the odds ratio.

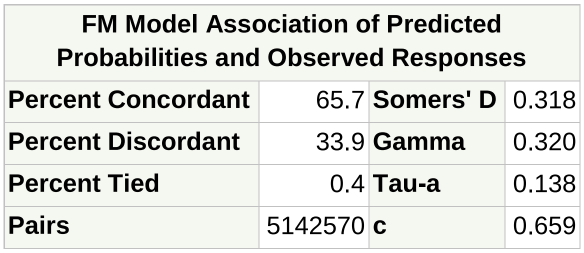
Odds Ratio Estimates					
Effect	Point Estimate	Probability	95% Wald Confidence Limits		Description
rr	21.500	0.95555556	0.552	836.914	This variable has strong odds, but is not statistically significant.
rr2	0.008	0.00793651	<0.001	0.728	There is little predictive significance with this variable.
age	1.070	0.51690821	1.021	1.121	As age increases, the likelihood of receiving benefits increases, but not by much.
age2	0.994	0.49849549	0.988	1.000	There is less of a chance with Age transformed.
tenure	1.032	0.50787402	1.018	1.045	Tenure does really matter one way or the other.
slack	1.868	0.65132497	1.626	2.145	In this study, being a slacker increases your probability of receiving benefits.
abol	0.964	0.49083503	0.766	1.215	You have less chance of receiving benefits if your job was eliminated.
seasonal	1.311	0.56728689	0.937	1.834	Being a seasonal worker increases probability of benefits, but there were just a few observations for this variable.
head	0.810	0.44751381	0.691	0.950	Being the head of a household negatively affects your chances of benefits.
married	1.274	0.56024626	1.090	1.489	Marriage is one of the few variables that increase your chances of receiving benefits.

Odds Ratio Estimates					
Effect	Point Estimate	Probability	95% Wald Confidence Limits		Description
dkids	0.854	0.46062567	0.721	1.011	Having kids decreases your chances of benefits. This would be a point to show management.
dykids	1.229	0.55136833	1.015	1.487	Having young kids increases your chances of benefits.
smsa	0.843	0.4574064	0.736	0.967	The location detracts from the probability of receiving benefits.
nwhite	1.077	0.51853635	0.898	1.292	Being non-white gives a slight edge over white individuals.
yrdispl	0.938	0.48400413	0.911	0.966	As the years increase, your chances of receiving benefits decrease.
school12	0.937	0.48373774	0.797	1.101	Being educated decreases your chances of benefits.
male	0.835	0.45504087	0.704	0.992	Non-males have a greater chance of receiving benefits than white people.
stateur	1.100	0.52380952	1.067	1.135	As the state unemployment rate increases, so does a person’s chance of benefits.
statemb	1.006	0.50149551	1.004	1.008	State max benefits do not play that a major factor in receiving benefits.

As a final analysis, reviewing Association of Predicted Probabilities and Observed Responses output is helpful. I will compare the MM model and FM to explain the metrics.



MM Model Association of Predicted Probabilities and Observed Responses			
Percent Concordant	63.1	**Somers' D**	0.267
Percent Discordant	36.4	**Gamma**	0.269
Percent Tied	0.5	**Tau-a**	0.116
Pairs	5142570	**c**	0.634



FM Model Association of Predicted Probabilities and Observed Responses			
Percent Concordant	65.7	**Somers' D**	0.318
Percent Discordant	33.9	**Gamma**	0.320
Percent Tied	0.4	**Tau-a**	0.138
Pairs	5142570	**c**	0.659

After assessing the goodness of fit, it is desirable to analyze the statistics that measure the predictive power of specific variables. The approaches used in SAS from the PROC LOGISTIC command are ordinal measures of association that produce model-free measures of predictive power. The percent concordant mean is interpreted as a pair of observations with different responses, and the observation with the lower ordered response value has a lower predicted mean score than the observation with the higher ordered response value (UCLA.edu). Percent discordant is the opposite of concordant, and if there is a tie it is displayed in the third row. Somer's, Gamma, Tau-a, and C are all calculations based on the percent calculations. Higher scores relate to greater predictive power. The C calculation is desirable in that it relates to the ROC curve, and the Tau-a is most closely associated to the R squared of linear regression. Comparing the two models the full model is better, and the C value will directly relate to the ROC curve.

As a whole, this model is significant. In my opinion, there are variables that should be removed in order to adhere to the principle of parsimony. The greatest revelation from this study is the disparity between unemployment theory and application. My background is in public policy and implementation. The results from this study reveal that changes need to be made to the implementation of unemployment benefits such that it realigns with the overall goal of the program.

Future Work

Further recommendations on how this study can be improved upon are the following:

- Break up the continuous variables similar to variable yrsdispl.
    - Doing this will help see the incremental changes as the response variables demonstrates its linearity.
- Include variables that are more predictive.
    - Realistically, this is hard to do, but I would spend time trying to find other variables that were a factor in receiving unemployment benefits.
- Include another data set for a different region.
- Include data from a different time period.
    - Perhaps data on The Great Recession.

Through this initial EDA, coupled with the future work recommendations, delineating variables that increase/decrease receiving unemployment greatly benefits individuals and the government.

References

Ajmani, V. (2009). *Applied Econometrics Using the SASSystem*. Hoboken: John Wiley & Sons.

Works Cited

Aaron, . "Logistic Regression." *University of Oregon*. N.p., n.d. Web. 26 Jan. 2013. <http://pages.uoregon.edu/aarong/teaching/G4075_Outline/node16.html>.

Allison, Paul David. *Logistic regression using SAS theory and application, second edition*. 2nd ed. Cary, N.C.: SAS Institute, 2012. Print.

Doyle, Alison. "Unemployment Benefits." *Job Search, Interview & Employment Advice from About.com*. N.p., n.d. Web. 22 Jan. 2013. <http://jobsearch.about.com/cs/unemployment/a/unemployment.htm>.

"Logistic Regression." *Statistics Solutions*. UCLA, n.d. Web. 26 Jan. 2013. <http://www.ats.ucla.edu/stat/stata/dae/zinb.htm>.

Van Horn, Carl, and Cliff Zukin. "Unemployed Workers and the Great Recession." *Work Trends Reports, 2009-2010* 1.1 (2010): 1-19. *John J. Heldrich Center for Workforce Development Rutgers University*. Web. 22 Jan. 2013.

"The Anguish of Unemployment." *Rutgers*. Version 1. Rutgers State University, 1 Sept. 2009. Web. 22 Jan. 2013. <http://www.heldrich.rutgers.edu/sites/default/files/content/Heldrich_Work_Trends_Anguish_Unemployment.pdf>.

Appendix 1 (Emailed to you on 1/26/13)

Appendix 2

Effect	Point Estimate	Probablity	95% Wald Confidence Limits	
rr	21.5	0.95555556	0.552	836.914
rr2	0.008	0.00793651	<0.001	0.728
age	1.07	0.51690821	1.021	1.121
age2	0.994	0.49849549	0.988	1
tenure	1.032	0.50787402	1.018	1.045
slack	1.868	0.65132497	1.626	2.145
abol	0.964	0.49083503	0.766	1.215
seasonal	1.311	0.56728689	0.937	1.834
head	0.81	0.44751381	0.691	0.95
married	1.274	0.56024626	1.09	1.489
dkids	0.854	0.46062567	0.721	1.011
dykids	1.229	0.55136833	1.015	1.487
smsa	0.843	0.4574064	0.736	0.967
nwhite	1.077	0.51853635	0.898	1.292
yrdispl	0.938	0.48400413	0.911	0.966
school12	0.937	0.48373774	0.797	1.101
male	0.835	0.45504087	0.704	0.992
stateur	1.1	0.52380952	1.067	1.135
statemb	1.006	0.50149551	1.004	1.008