Name: Prentice Mui
Group: 8
Professor: Glosemeyer
Due Date: May 3, 2017

Stat 448 Final Report on Mroz Data

1 Introduction

1.1 Background

This document is a specific research and analysis into women employment and demographics from the year 1976. The data source being analyzed is called the Mroz data set found in the Ecdat data set in R. The data was initially sourced from Wooldridge's Econometric Analysis and can be found on this domain http://pages.stern.nyu.edu/~wgreene/Text/tables/tablelist5.htm. The cross sectional data was a survey conducted in 1976 asking 753 married women whether or not they had worked while also reporting on other demographic statistics relating to the women.

The data set contains a total of 18 variables with the main variable response being a categorical variable *work* with levels "yes" or "no" indicating whether the woman had worked or not in the year 1975. The variables' labels are listed in the table below.

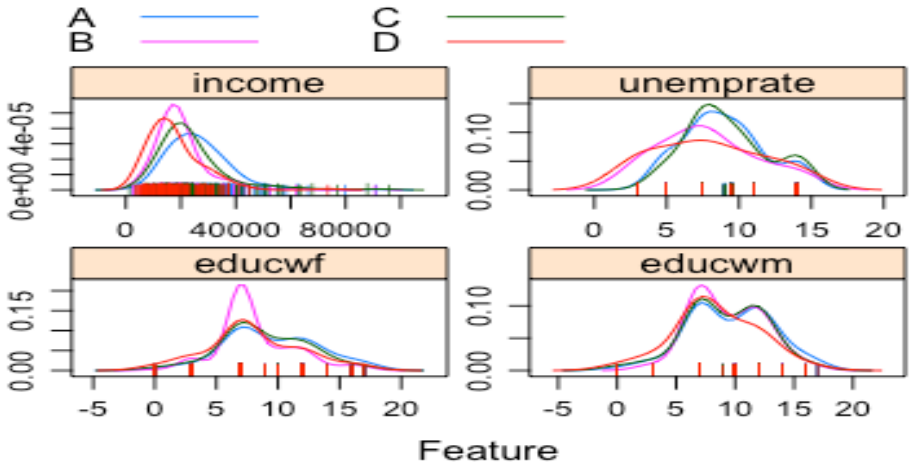
work	hoursw	child6	child618	agew	educw
hearnw	wagew	hoursh	ageh	educh	wageh
income	educwm	educwf	unemprate	city	experience

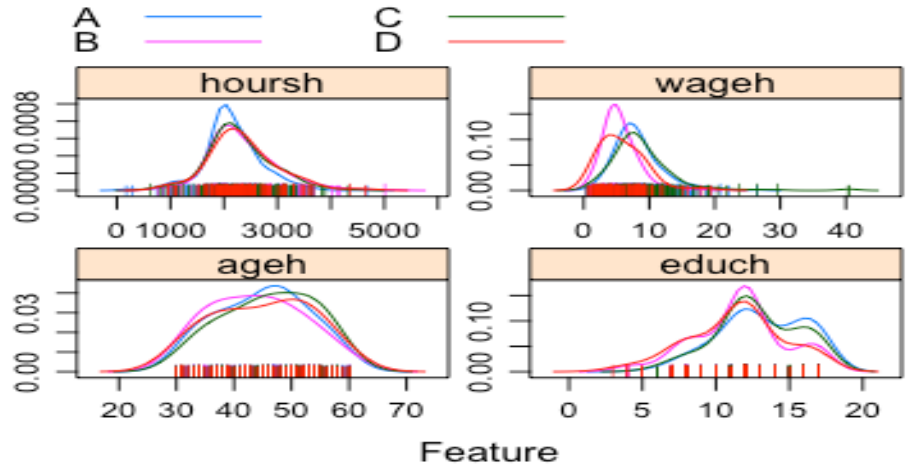
Here is the description of the variables in the table from left to right top to bottom. They represent the woman's hours worked in 1975, the number of children under the age of six in her household, the number children between six and eighteen years of age in her household, the wife's age, the woman's educational attainment in years, her wage in 1976, her husband's hours worked in 1975, her husband's age, her husband's educational attainment in years, her husband's wage in 1975, the family income in 1975, the woman's mother's educational attainment in years, the woman's father's educational attainment in years, the unemployment percentage rate in her residential county, whether or not she resides in a large city, and her total years of previous work experience.

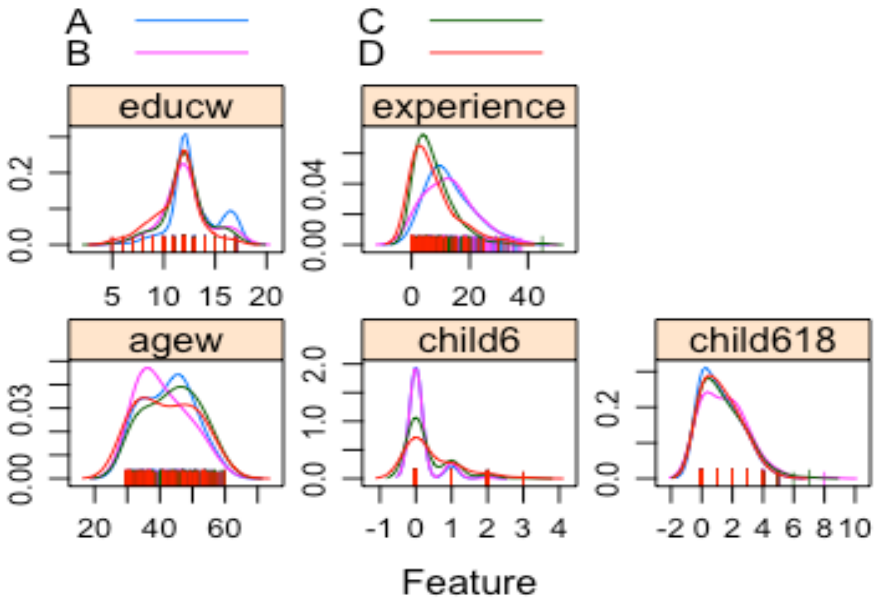
The goal of this analysis is to fit a discriminant model that hopefully accurately predicts whether a woman worked or not and whether she lived in the city. Therefore, the response that we are predicting will consist of four levels. To perform this analysis, a new variable called workcity will be coded and labeled as "A" for worked and lived in the city, "B" for worked but did not live in the city, "C" for did not work but lived in the city, and "D" neither worked nor lived in the city. The initial assumption is that various demographic characteristics of working, non-working, city, and non-city women should differ, therefore by finding out the different densities of these factors between the groups, a viable way to class the women should be discovered. After the model is fit and hopefully deemed valid, the model will be examined to see which factors were important in making that distinction and inferences will be made.

1.2 Descriptive Statistics and Visuals

Provided below are some tables and visuals that provide some initial insights into the nature of the data and serves as an initial scouting step to see if the analysis would be useful and if the goals and questions formed could be concluded.







The plots on the previous page are the density plots of thirteen variables that will be used to predict the employment and residential status of the woman in 1975. Notice that three variables are left out which include the woman's hours of work, her average hourly earnings in 1975, and her wage in 1976. The reason these variables are omitted is because they are too correlated with whether or not the woman worked and using those variables defeats the purpose of performing any analysis as any amount of wage or hours indicates that the woman worked.

Looking at the density plots for any noticeable differentiation between the for groups, it's a bit alarming that none of the variables really seem to discriminate between the classes. From an eye test, only the wage of the husband, experience, and number of children under 6 seems to make any distinction among the groups at all.

workcity=A

Variable	Mean	Std Dev
experience	13.1459854	7.8898462
hoursw	1289.58	777.5820579
child6	0.1423358	0.4080190
child618	1.2627737	1.2360286
agew	42.5474453	7.7113406
educw	12.9306569	2.2578171
hearnw	4.4735496	3.6706342
wagew	3.4387956	2.6343188
hoursh	2162.95	545.2312961
ageh	45.0036496	7.8667931
educh	13.1496350	2.8828563
wageh	8.1968646	3.6678783
income	26412.77	11616.07
educwm	9.6021898	3.5111746
educwf	9.4781022	3.6945503
unemprate	8.9598540	2.7695746

workcity=B

Variable	Mean	Std Dev
experience	12.8441558	8.3659205
hoursw	1326.69	775.9042502
child6	0.1363636	0.3627812
child618	1.5064935	1.4382576
agew	40.9480519	7.6566791
educw	12.1753247	2.2611615
hearnw	3.6512669	2.4708916
wagew	2.7358442	1.9773220
hoursh	2358.92	626.9769208
ageh	43.9090909	8.0742985
educh	11.6558442	3.0729260
wageh	5.4992468	2.6266355
income	20069.63	10652.47
educwm	9.3636364	2.9167494
educwf	8.1168831	3.0161830
unemprate	7.8084416	3.3373496

workcity=C

Variable	Mean	Std Dev
experience	7.5047619	6.5543350
hoursw	0	0
child6	0.3238095	0.5868615
child618	1.3952381	1.3906648
agew	43.8619048	8.2383731
educw	12.0714286	2.1116317
hearnw	0	0
wagew	0.1395238	0.6576453
hoursh	2288.24	592.9226375
ageh	46.2095238	7.9600621
educh	12.8428571	2.7597711
wageh	8.8992371	5.2992022
income	23863.60	13455.38
educwm	9.1857143	3.3516097
educwf	8.8666667	3.4794452
unemprate	9.1690476	2.8031500

workcity=D

Variable	Mean	Std Dev
experience	7.3826087	7.5679178
hoursw	0	0
child6	0.4434783	0.7156313
child618	1.2869565	1.2049985
agew	42.2260870	8.8099434
educw	11.2956522	2.2281056
hearnw	0	0
wagew	0	0
hoursh	2354.80	640.0997894
ageh	45.0347826	8.5048288
educh	11.4000000	3.2003289
wageh	5.8470948	3.4880829
income	17743.58	10203.12
educwm	8.3826087	3.4907964
educwf	8.0347826	3.8409850
unemprate	7.9173913	3.7498131

Above are four tables that gives means and standard deviations of each of the predictor variables relative to the class that they belong to. As evident, the variables that may help in classifying between the classes are experience, wage of the husband, unemployment rate, income, and children under six since these variables differ somewhat notably amongst pairs of classes; Though this indicates that it may only be able to differentiate work or city and not work and city. Again the variables of the hours that the woman worked, the woman's hourly earnings in 1975, and the woman's wage in 1976 will not be used specifically because they are directly correlated with work and essentially provide perfect discrimination for those levels. At this point, it seems that there are enough differences in the variables to possibly discriminate employment and residence though the prospect of finding a really well working model does not seem very likely.

2 Methods and Results

2.1 Initial Discriminant Analysis Model

Here is where the proc discrim model is actually applied to the model. Before proceeding further into the model, an initial exploratory analysis check was done to determine whether or not discriminant analysis is appropriate for this data. The output below is the manova test output of the discriminant analysis procedure. Since the four tests yield Pr > F values less than to .0001, it indicates that there are differences between at least some of the variables for the levels of workcity and discriminant analysis is a viable procedure to enact. Discriminant analysis

will be performed under proportional priors since the proportions of each class seem to differ and there is no reason to assume that the population proportions are equal.

Multivariate Statistics and F Approximations					
S=3 M=4.5 N=367.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.57512826	11.50	39	2183.2	<.0001
Pillai's Trace	0.48149365	10.87	39	2217	<.0001
Hotelling-Lawley Trace	0.64217589	12.12	39	1768.4	<.0001
Roy's Greatest Root	0.43403163	24.67	13	739	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

Discriminant analysis was fit to the data and the specific method of dealing with the covariance matrices was quadratic discriminant analysis(QDA) since the test of homogeneity within covariance matrices was rejected as seen below. This entails that the variance of the distributions of the variables differ between the classes so each distinct covariance matrix is used instead of a calculated pooled one.

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
552.118880	273	<.0001

After fitting the full main effects QDA, the following cross validation confusion/classification table was created. Cross validation was the method used to make classifications as cross validating allows for less bias in accuracy because the observation being predicted is not used in the model to predict.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.MROZ
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into workcity

From workcity	A	B	C	D	Total
A	162 59.12	49 17.88	46 16.79	17 6.20	274 100.00
B	66 42.86	52 33.77	17 11.04	19 12.34	154 100.00
C	53 25.24	26 12.38	98 46.67	33 15.71	210 100.00
D	16 13.91	28 24.35	38 33.04	33 28.70	115 100.00
Total	297 39.44	155 20.58	199 26.43	102 13.55	753 100.00
Priors	0.36388	0.20452	0.27888	0.15272	

Error Count Estimates for workcity

	A	B	C	D	Total
Rate	0.4088	0.6623	0.5333	0.7130	0.5418
Priors	0.3639	0.2045	0.2789	0.1527	

It's obvious that QDA was not very effective for classifying the employment status and residence of women in 1975 since the error rate was a poor .5418 meaning over half of the predictions were wrong. Looking at each individual class's error rate, we see that class D had the poorest accuracy while class A had comparatively the best so the model was best able to tell when a woman was a city worker. In the classification table we see that category B had the most error in being classed as A, so in terms of the data, non-city working women were more frequently incorrectly classed to be city working women. Meanwhile most of class D's erroneous classing was to class C meaning the model more frequently classed a non-working non-city working woman as just a non-working woman. Since this model worked so poorly, the next step is to perform variable selection in hopes of eliminating some insignificant variables that may be noise to model.

2.2 Variable Selected Model for Workcity

After variable selection was performed, the predictors, children between 6-18, age of the husband, education of the woman's mother, and education of the woman's father were deemed insignificant and removed from the full model. Afterwards the model was refit with the selected variables under the same parameters of QDA and proportional priors which yielded the cross validated classification and error tables below.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.MROZ
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into workcity

From workcity	A	B	C	D	Total
A	157 57.30	57 20.80	46 16.79	14 5.11	274 100.00
B	69 44.81	55 35.71	13 8.44	17 11.04	154 100.00
C	67 31.90	16 7.62	100 47.62	27 12.86	210 100.00
D	15 13.04	31 26.96	36 31.30	33 28.70	115 100.00
Total	308 40.90	159 21.12	195 25.90	91 12.08	753 100.00
Priors	0.36388	0.20452	0.27888	0.15272	

Error Count Estimates for workcity

	A	B	C	D	Total
Rate	0.4270	0.6429	0.5238	0.7130	0.5418
Priors	0.3639	0.2045	0.2789	0.1527	

Looking at the error table, the most shocking finding is that the overall error rate did not change. This is quite fascinating because in general, removing insignificant terms should improve accuracy at least ever so slightly but it did not occur here. Though the overall error did not change, the error rates for the individual classes did so slightly. In this reduced model, the error rate for class A worsened, the errors for classes B and C improved slightly, while class D's error was consistent. In a sense, this model could be preferred to the former one because it improved at predicting two classes for the price of one and each class was valued equally relative to the others. Looking more in depth into the classification table, classes B and C improved in accuracy only because they were now being misclassed as each other and D less

frequently though they are also now being misclassed more often into A. At this juncture, it seems that this may be best the model will get.

2.3 Modified Discriminant Model for Work

Because of the poor performance of the former model, the plan now is to simplify the goal a bit and have the model just predict whether or not the woman worked so there are only two levels of the response instead of four. The data set was centered on the work status of a woman anyway so it makes sense to revert back to this feature. It is of no use to use a model with an error rate of over 50% as based on guessing alone, the error rate is 75% for classing four response levels. Since the former model was negligibly improved over guessing, moving to a simpler analysis may yield a more useful model and interesting results.

Going through the same processes as before, the initial full model was check by the manova test to make sure differences existed across the classes, the covariances were deemed to be different so QDA is the method of learning, and priors are assumed to be proportional. From variable selection the predictors that were deduced to be insignificant were child618, ageh, educh, educwm, educwf, and unemprate. After removing those variables, a reduced QDA model was fit to the data to predict the response of the variable work with the results displayed below.

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.MROZ
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into work			
From work	no	yes	Total
no	201 61.85	124 38.15	325 100.00
yes	83 19.39	345 80.61	428 100.00
Total	284 37.72	469 62.28	753 100.00
Priors	0.43161	0.56839	

Error Count Estimates for work			
	no	yes	Total
Rate	0.3815	0.1939	0.2749
Priors	0.4316	0.5684	

Proving some sense of relief, this model works proves to work much better sporting an error rate of .2749. Note that this is still not exactly an ideal model as .2749 is not exactly an indicator of such but because this model is just much improved compared to the other credit has to at least be acknowledged to it. In terms of the actual class errors, the model seems to better at pinpointing working women rather than non-working as shown by the .1939 error for "yes" and the .3815 error for "no".

3 Conclusions

Unfortunately, sometimes statistical analysis and modeling may not be perfect and a machine learning task could prove to be unsuccessful as was seen in this case. The initial goal of classifying working woman and residence and making inferences based on the chosen predictors ended in failure. There will be no reason to further infer that model because it proved to not be effective in predicting the task anyway.

Luckily, there was some optimism to be discovered as the simple two class response work model did work considerably better. To that point, some inferences and interpretations based on that model will be conducted instead.

Based on the final reduced model, the variables that were present included, child6, agew, educw, hoursh, wageh, income, and experience. What is noticeable is that many of these variables either relate to the woman herself or capture wealth of the family. The variables removed were educh, educwm, educwf, and unemprate. The distinction of these variables is that they are education and economic factors that are not in the woman's control. Clearly this study was an observational study and not an experimental one so claiming causation is a stretch but there definitely seems to some correlation between a woman working and things that she can somewhat influence. In the appendix section, probability density graphs display how each of these factors and the interpretation was that the more children the woman had under the age of six the less likely it was that she worked. The older the woman was, the less likely she worked. The more education the woman had, the more likely she worked. The more her husband worked and earned, the less likely she worked. The more experience that the woman had, the more likely she worked.

These results do conform to common beliefs and stereotypes that, especially in the decades prior, woman did not hold as many high paying and high ranking positions so were less likely to work unless they had a college degree. Also it makes sense that a woman would work less if she has young kids and the family(husband) is making more money because she has more priorities to take care of in the household and the family can sustain on just the husband's income alone. Additionally, if a woman is older, it just means that she is likely nearing retirement. Lastly, if a woman has more experience, she is more likely to work probably because her experience constitutes as qualifications and requisite skills that jobs want in their candidates. So essentially, these factors seem to follow what would be somewhat commonplace social and familial knowledge. The seemingly most noteworthy point is that having a job is not really dependent on parents or familial education, a woman working has more to do with her skills, experience, responsibilities, and choice rather than factors that she cannot really control. These results seem very interesting and maybe further observational or experimental studies of socioeconomic behaviors could be studied further to see why or what causes woman or men for that matter to be able or willing to work vs otherwise.

A final reminder and reflection about the inferences and analysis performed. The reason for some shortcomings with these models may have to do with the nature of the data set itself. The data set is over 40 years old and has a relatively small sample size which could cause the data to be skewed or biased in some way. Furthermore, because of the age discrepancy of the data, the inferences made from it may not be easily translatable to modern era social sciences.

4 Appendices

4.1 Variable Selection for model predicting workcity

The STEPDISC Procedure
Stepwise Selection: Step 8

Statistics for Removal, DF = 1, 745			
Variable	Partial R-Square	F Value	Pr > F
child6	0.0810	65.67	<.0001
agew	0.0969	79.93	<.0001
educw	0.0250	19.07	<.0001
hoursh	0.0366	28.32	<.0001
wageh	0.0587	46.50	<.0001
income	0.0574	45.40	<.0001
experience	0.1208	102.35	<.0001

No variables can be removed.

Statistics for Entry, DF = 1, 744				
Variable	Partial R-Square	F Value	Pr > F	Tolerance
child618	0.0009	0.70	0.4047	0.3471
ageh	0.0005	0.36	0.5506	0.1996
educh	0.0008	0.57	0.4501	0.3248
educwm	0.0001	0.10	0.7509	0.3472
educwf	0.0000	0.00	0.9836	0.3466
unemprate	0.0003	0.21	0.6491	0.3442

No variables can be entered.

No further steps are possible.

The above charts display the final step in the discriminant stepwise selection method to choose the "best" variables to include in a reduced model. At this step, the variables that were deemed not significant were child618, ageh, educwm, and educwf. As seen in section 2.2, this modified model did not help in increasing the accuracy for predicting workcity.

4.2 Initial Discriminant Model for Work

The DISCRIM Procedure

Multivariate Statistics and Exact F Statistics					
S=1 M=5.5 N=368.5					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.69736942	24.67	13	739	<.0001
Pillai's Trace	0.30263058	24.67	13	739	<.0001
Hotelling-Lawley Trace	0.43396022	24.67	13	739	<.0001
Roy's Greatest Root	0.43396022	24.67	13	739	<.0001

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
263.109150	91	<.0001

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.MROZ
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into work			
From work	no	yes	Total
no	197 60.62	128 39.38	325 100.00
yes	92 21.50	336 78.50	428 100.00
Total	289 38.38	464 61.62	753 100.00
Priors	0.43161	0.56839	

Error Count Estimates for work			
	no	yes	Total
Rate	0.3938	0.2150	0.2922
Priors	0.4316	0.5684	

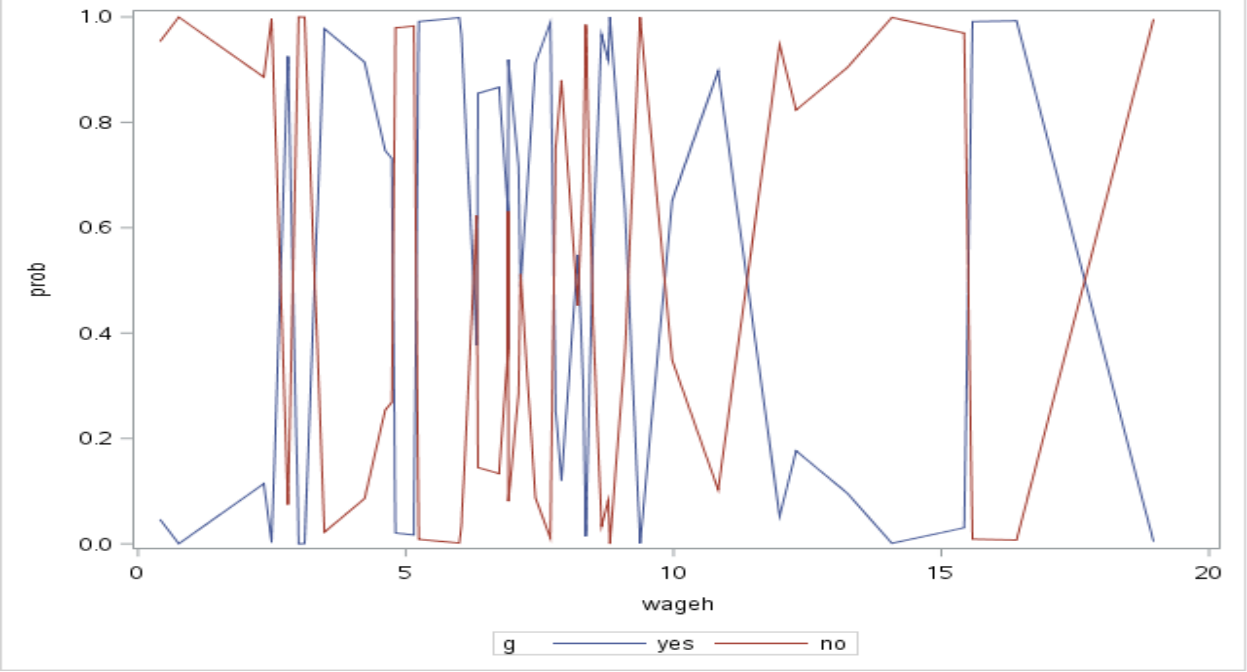
The three images below show the manova test, pooled covariance test, and the cross validated classification results respectively from left to right for the full discriminant model that predicted work. The manova result indicated that differences did exist and the model should use individual covariance matrices This model preceded the final reduced model presented in 2.3 and also shows improvement over the models in sections 2.1 and 2.2.

4.3 Kernel Density

Error Count Estimates for work			
	no	yes	Total
Rate	0.4677	0.1636	0.2948
Priors	0.4316	0.5684	

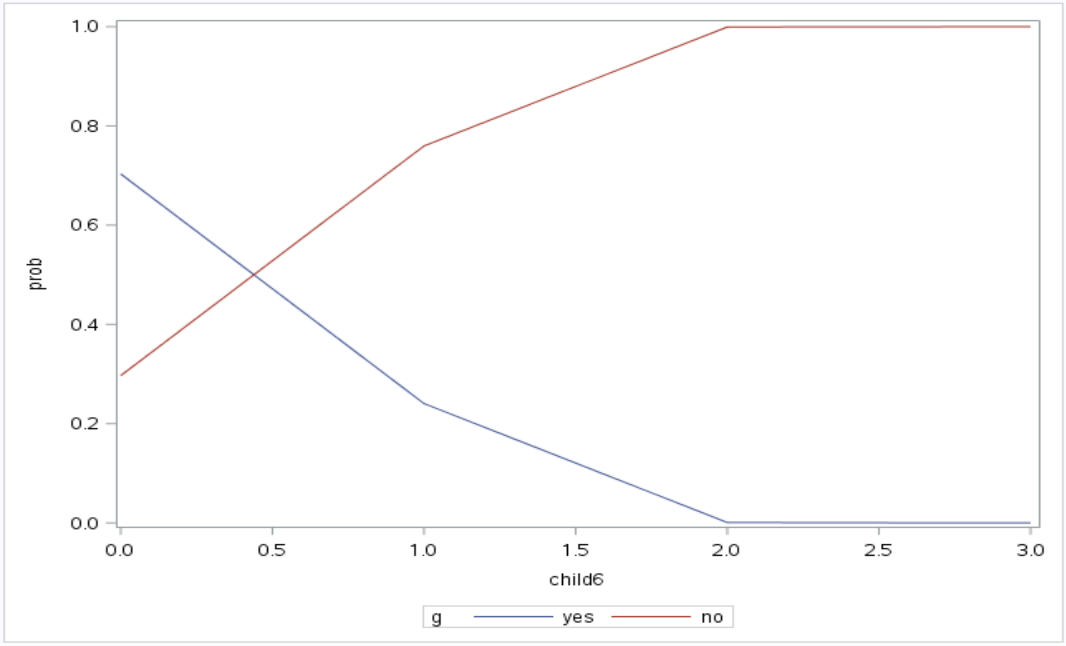
A kernel density estimation of the predictors was also tested on the selected variables but since the cross validated error proved to be higher than the QDA model, this method was not used in the final analysis and results.

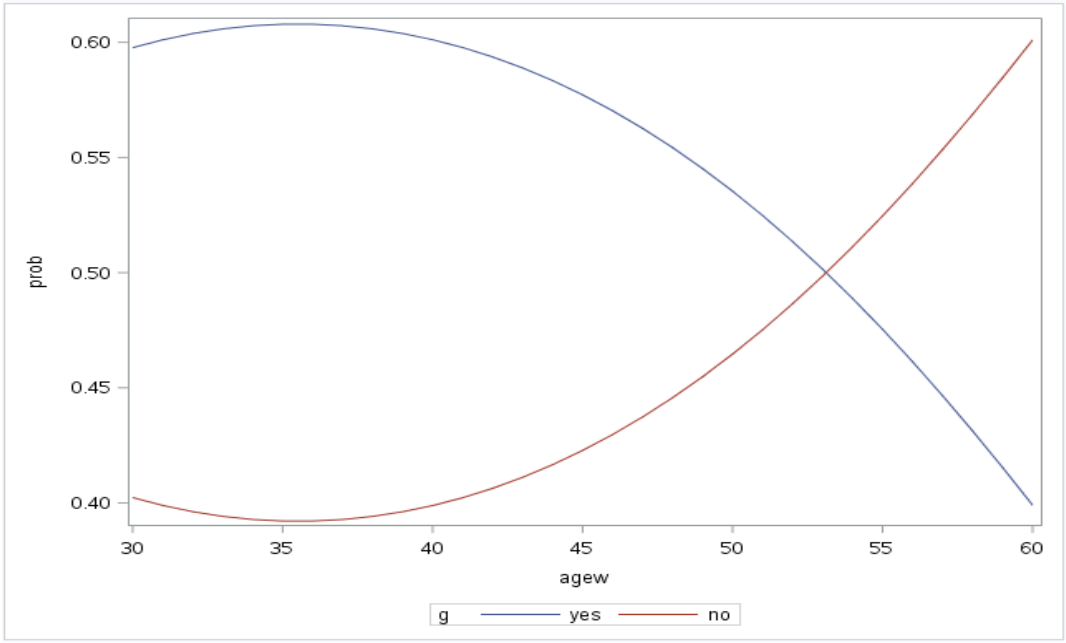
4.4 Multivariate Posterior Probabilities

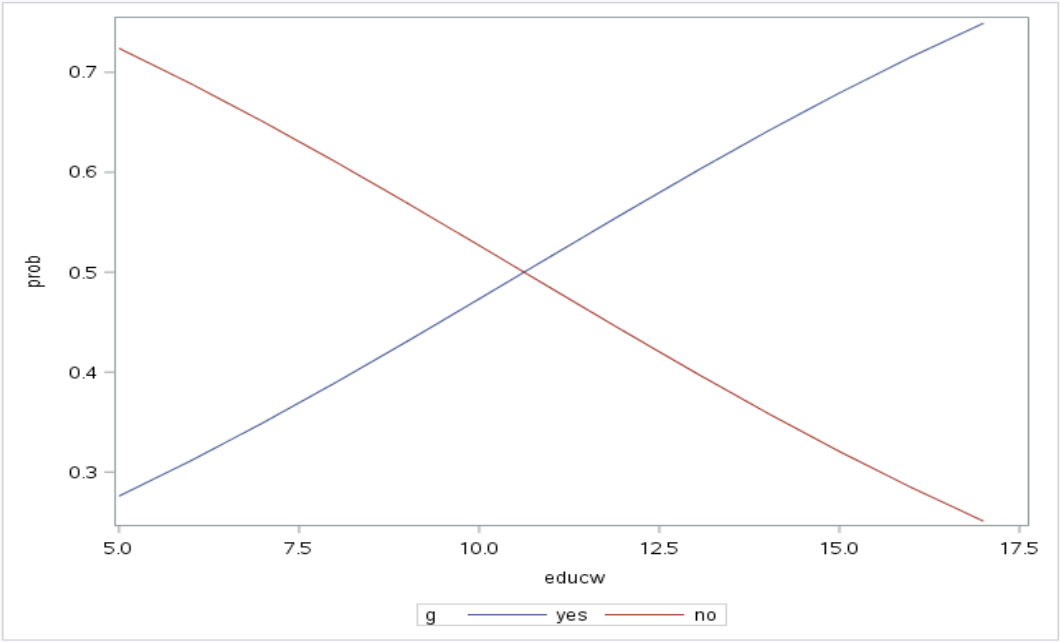


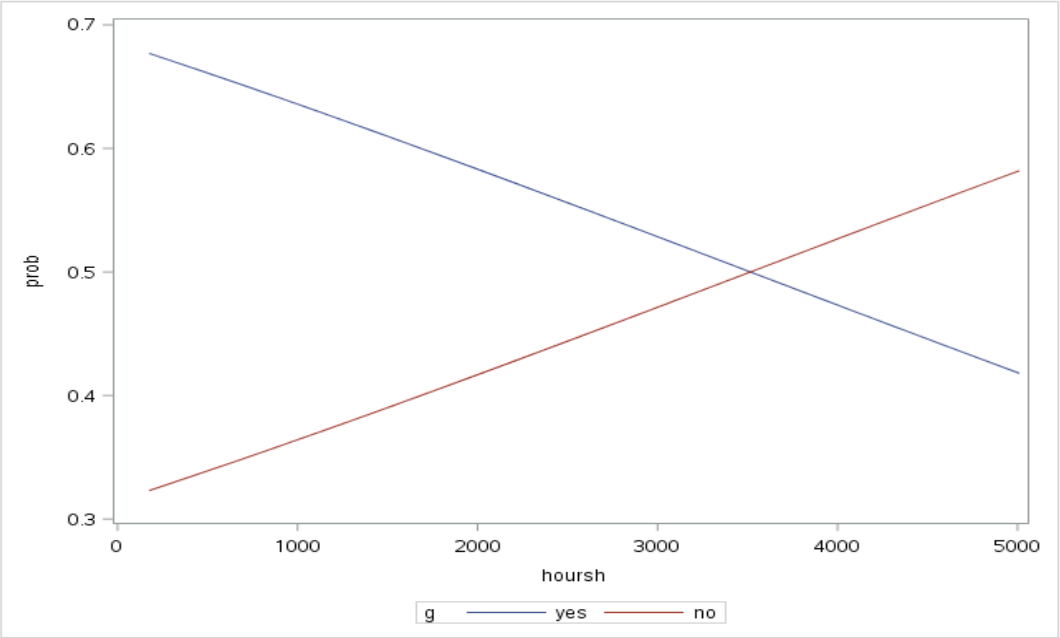
This graph represents the posterior probability for the reduced work model based on the wageh varying while all other variables were held to a constant. The constants were determined as the modes of the data set. Since holding variables constant requires the specific numeric values to exist in the data set many of the observations did not have the specific combinations of the variables so only a few observations were used to create this one variable posterior probability relationship. Obviously the plot does not look sharp at all.

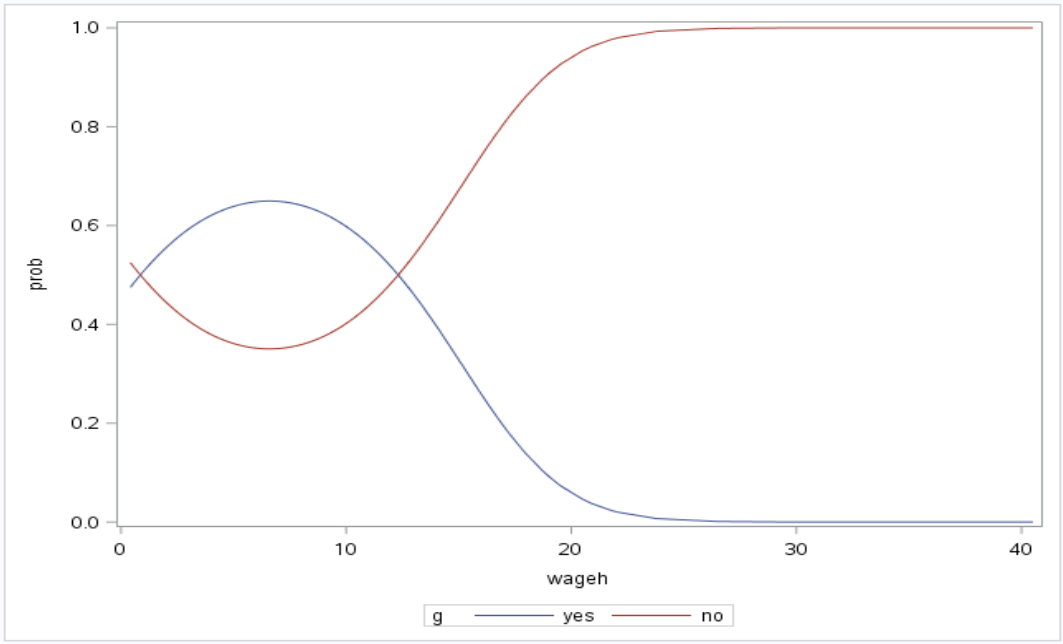
4.5 Univariate Posterior Probabilities

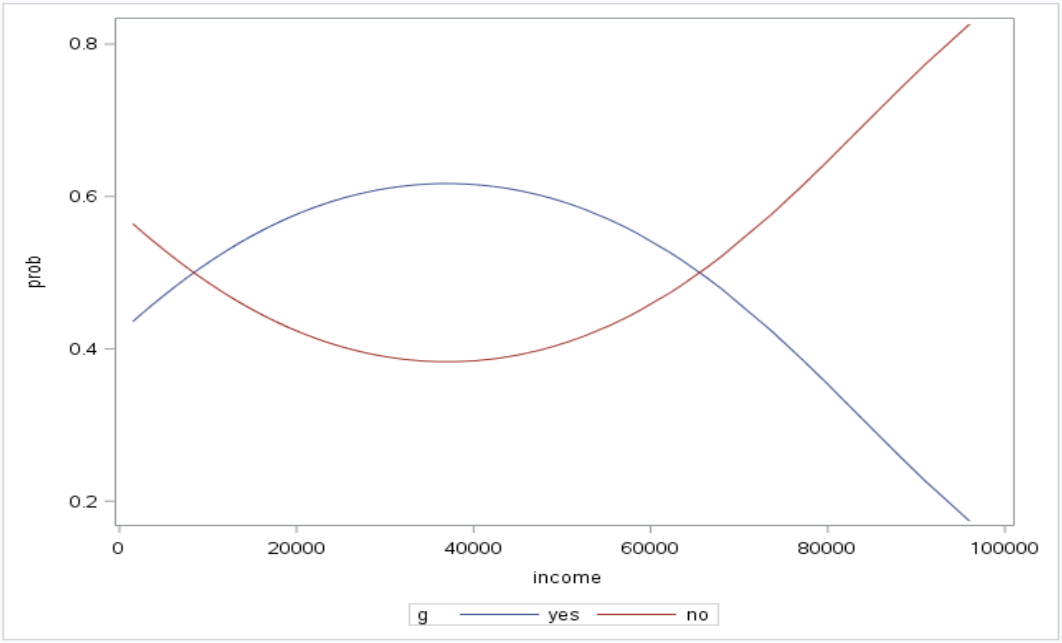

1.0
0.8
0.6
0.4
0.2
0.0
prob
0.0
0.5
1.0
1.5
2.0
2.5
3.0
child6
g
yes
no

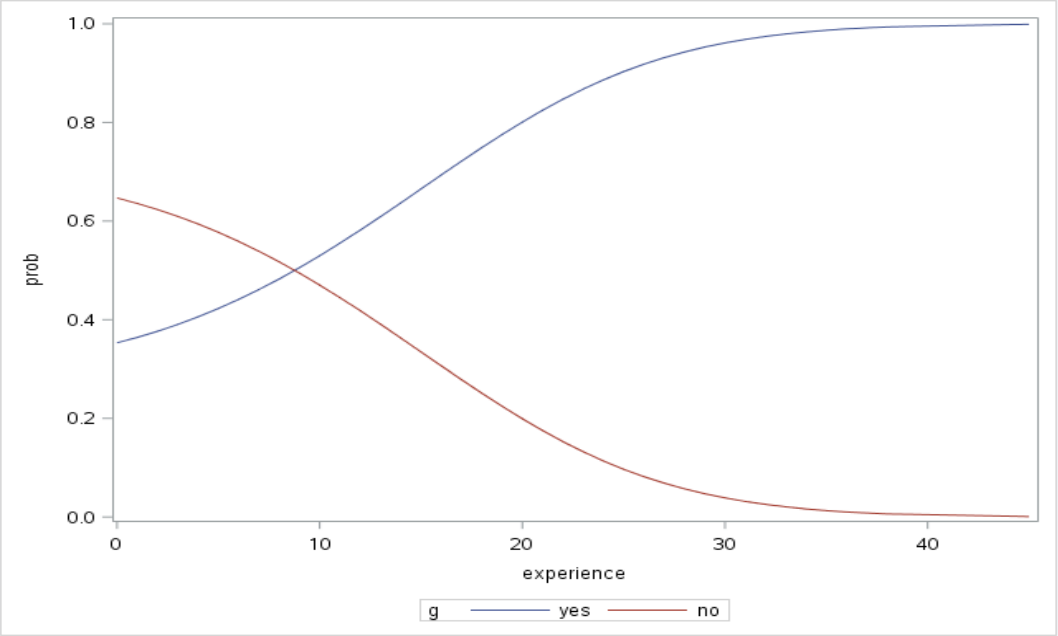

0.60
0.55
0.50
0.45
0.40
prob
30
35
40
45
50
55
60
agew
g
yes
no


0.7
0.6
0.5
0.4
0.3
prob
5.0
7.5
10.0
12.5
15.0
17.5
educw
g
yes
no


0.7
0.6
0.5
0.4
0.3
prob
0
1000
2000
3000
4000
5000
hoursh
g
yes
no


1.0
0.8
0.6
0.4
0.2
0.0
prob
0
10
20
30
40
wageh
g
yes
no


0.8
0.6
0.4
0.2
prob
0
20000
40000
60000
80000
100000
income
g
yes
no



Because of issue with multi-dimensional graphs, modified posterior probability density graphs were created to examine the effect of each individual variable on the probability of work. To do this, repeated discriminant analysis were done based on one predictor and the seven variables and their influence on the probability of work are plotted above and on the preceding three pages. The sole purpose of this was to understand a one to one relationship between that predictor and the probability of working. Clearly nothing else can be interpreted of it since the full model has multivariate interactions involved.

4.6 Possible Issues with Data

Below is the correlation plot of the variables used in the final model in section 2.3. The purpose of creating this was to understand possibly why the models were not performing well. In the previous section, we deduced the relationships of the seven variables with and probability of work. The only two that were positively correlated with the increased probability of work were educw and experience. This means that these variables should be negatively correlated the other five. However, looking at the correlation plot this is not necessarily the case. For example, educw is somewhat noticeably positively correlated with child6, hoursh, wageh, and income, when the plots above suggest that the relationship should be reversed. Because of this is likely why the predicted models were not of great fit and why the probability plots of multiple dimensions were very chaotic. There may just be excessive noise and randomness amongst the observations that prevent any clear discriminant boundaries from being drawn and thereby yielding poorer fitting models.

Pearson Correlation Coefficients, N = 753 Prob > |r| under H0: Rho=0							
	child6	**agew**	**educw**	**hoursh**	**wageh**	**income**	**experience**
child6	1.00000	-0.43395 <.0001	0.10869 0.0028	0.02429 0.5057	0.03238 0.3749	-0.02778 0.4465	-0.19404 <.0001
agew	-0.43395 <.0001	1.00000	-0.12022 0.0009	-0.08437 0.0206	0.02701 0.4592	0.05244 0.1505	0.33402 <.0001
educw	0.10869 0.0028	-0.12022 0.0009	1.00000	0.07892 0.0304	0.28494 <.0001	0.36127 <.0001	0.06626 0.0692
hoursh	0.02429 0.5057	-0.08437 0.0206	0.07892 0.0304	1.00000	-0.23602 <.0001	0.12814 0.0004	-0.09937 0.0064
wageh	0.03238 0.3749	0.02701 0.4592	0.28494 <.0001	-0.23602 <.0001	1.00000	0.72502 <.0001	-0.10331 0.0045
income	-0.02778 0.4465	0.05244 0.1505	0.36127 <.0001	0.12814 0.0004	0.72502 <.0001	1.00000	-0.02770 0.4478
experience	-0.19404 <.0001	0.33402 <.0001	0.06626 0.0692	-0.09937 0.0064	-0.10331 0.0045	-0.02770 0.4478	1.00000