Part Two

Summary

This report evaluates the relationship between the odd ratio of non-zero wages and its predictors. It primarily focuses on the influence of training and different demographic groups on the non-zero wages outcome. Logistic regression was used to produce a model that concluded workers who received training tend to have higher odd ratio of non-zero outcomes than those who did not. Moreover, the demographic groups prove to be statistically significant to this model and have effects to the resulting odd-ratio. This report will detail several interactions between the predictors that contribute significantly to the non-zero outcome.

Introduction

Using the same data from Part One, the likelihood of workers who received training earning non-zero wages more than those who did not receive training is explored in this report. The effects of receiving training on the odds of having non-zero wages are further analysed, as well as various interactions between the predictors. Moreover, the effects of the association between different demographic groups and training on the likelihood of workers receiving non-zero wages are also evaluated in this report.

Data

For this part of the report, the likelihood of workers who received non-zero wages and zero wages is determined by their wages in 1978 (re78). This will be referred to as the wage response variable, and is based upon whether re78 is zero or non-zero. As it is a binary variable, a summary table has been drawn in Appendix 1.1 to show the summary statistics of the data provided.

As shown in the summary statistics, there are more participants earning wages than people not earning wages in 1978. The wage_fact (wage factored) shows that 143 of the participants have zero wages and 471 participants are receiving wages in 1978. There is also an unbalanced distribution in races, where there are only 72 Hispanics compared to the 299 Black participants and participants of other races. Moreover, there are more than double the number of participants who did not receive treatments (429) compared to those who did receive training (185).

Conditional probabilities of wage given the predictors were explored. The table below shows the conditional probability of wage given the treat variable - the likelihood for both zero and non zero outcomes given that the participant received treatment or not are very similar. This result is also similar for married and nodegree where the outcome probability between zero and non zero groups between the binary predictors are roughly the same. For the different demographic groups, the likelihood for zero wage given that the participant is Hispanic is lower than both Black and participants of other races. However, the likelihood for non-zero wages given that participant is Hispanic is higher than for Black participants and other participants. (Appendix 1.2)

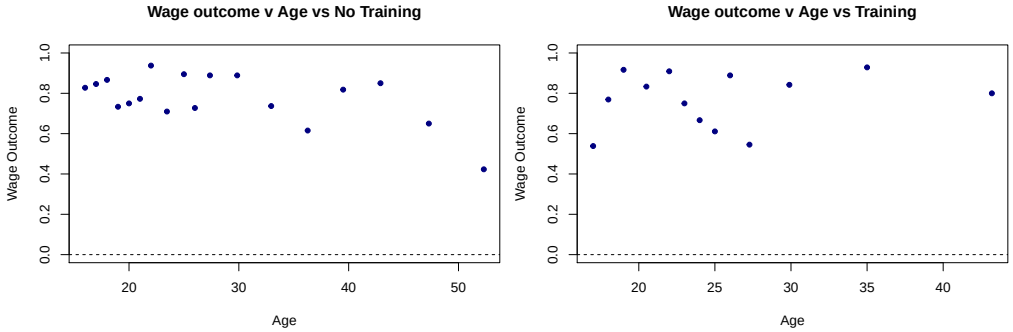
Table 1: Wage given treat

	treat0	treat1
wage0	0.23	0.24
wage1	0.77	0.76

Interactions between the predictor variables are also assessed, where the conditional probability table of non-zero wages between treat and race is shown in Appendix 1.6. The likelihood of Hispanics getting a non-zero wage given training is 100%, and the likelihood of Hispanics getting a non-zero wage given no training is 80.3%. This trend is similar for Black participants and participants of other races where the likelihood of

getting a non-zero wage is greater if they underwent training. It is also worth noting again that the there is not enough data for Hispanic participants compared to other and Black participants. Interactions between degree and train vs non-zero wage outcome can be shown in Appendix 1.4.

The following binned plots are shown to evaluate the interactions between age v train and educ v train on the non-zero wage outcome. The non-zero wage outcome v age v train show two different trends, where the training v age interactions shows a decrease in outcome around 22 - 25 year in age before increasing again at 30 years old. This is different compared to age v no training plot that shows a decrease in outcome at 30 years old before increasing at 36 years old, and then decreasing at around 42 years old. For educ v train and age v no-train on the non-zero wage outcome (Appendix 1.5), it is difficult to pinpoint a trend due to the lack of data points in the train v educ v wage outcome plot. Hence, the interactions of predictors on the response variable will be further analysed in this report.



Model

Model Building

Given that the response variable is a binary variable, the relationship between non-zero outcome and its predictors is analysed using a logistic regression, as shown below:

$$wage_i|x_i \sim Bernoulli(\pi_i) log(\frac{\pi_i}{1-\pi_i}) = x_i\beta$$

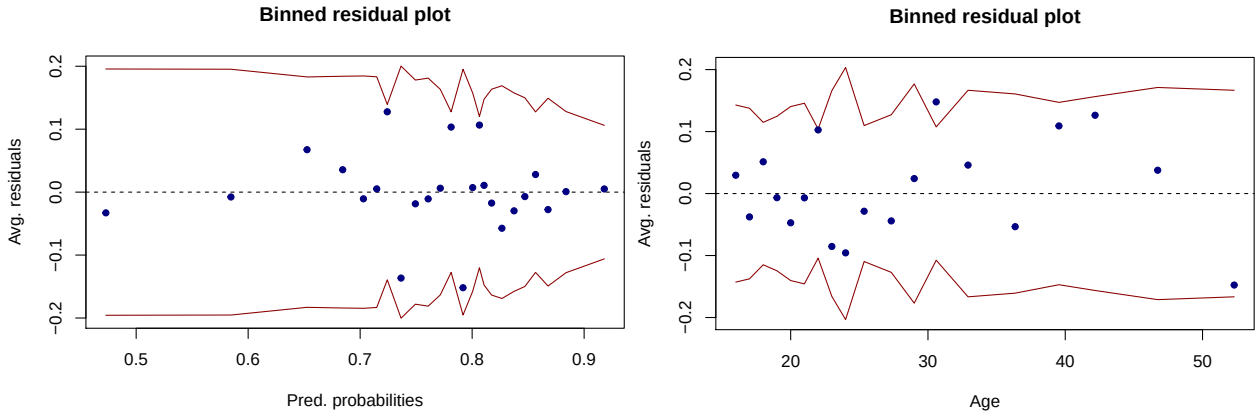
where $wage_i$ is the binary response variable indicating whether workers received a non-zero wage or not and x_i includes all predictor variables.

Model 1

For model 1, a model will be fitted for the wage response variable and all its predictors to see main effects. The x_i for model 1 includes the following predictors: treat, age, educ, race, married, nodegree, and re74. See Appendix 2.1 for the full summary table.

The residual deviance (634.79) is slightly lower than the null deviance (666.50), which tells us that the predictors are better than the worst model. Compared to the baseline(male participant of age ~27 years old and race as Black), the p-values for age, race(Other) and re74 show that they are statistically significant where the null hypothesis can be rejected. This suggests that with every unit increase in re74 whilst keeping the rest of the predictors constant, the odds ratio of a non-zero wage is expected to increase by $e^{6.011e-5}$. It also suggests that as the participant gets older by one year, odds ratio of a non-zero wage is expected to increase by $e^{-3.887e-2}$. If race(other) increases per unit, the odds ratio of a non-zero wage is expected to increase by $e^{-5.155e-1}$. The rest of the predictors have high values and suggest that they are significantly insignificant. However, we will continue to explore the relationship between the response and its predictor variables further in this report.

A residual binned plot was plotted for the non-zero wage probabilities v average residuals. This shows a pretty random dispersion of points, with all the points inside the 95% confidence band.



A binned plot for age v residuals is plotted above. Only 1 point lies outside the 95% confidence band and the rest of the points are dispersed quite randomly inside the 95% band.The rest of the predictors v residuals graph do not show sufficient insight as there are not enough data points in the graphs.

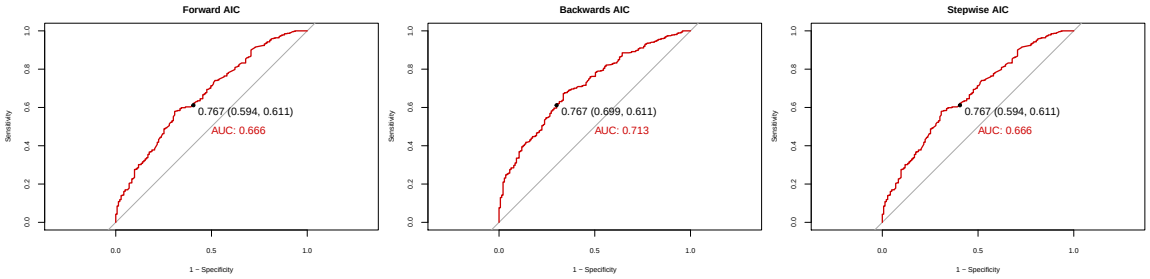
Model 2

For model 2, we added the interactions between the predictors in our logistic regression model.(Appendix 2.2 for summary)

As shown, race (Hispanic) shows that it is statistically significant due to its low p-value; thus, we can reject the null hypothesis. With every unit increase in raceHispanic, the odds ratio of a non-zero wage increases by $e^{1.001e1}$.. With every unit increase in re74 whilst keeping the rest of the predictors constant, the odds ratio of a non-zero wage is expected to increase by $e^{1.408e-4}$. As the interactions between raceHispanic and noDegree (contrasted to the baseline of raceBlack and degree) increases per unit, the non-zero wage odd ratio is expected to increase by $e^{-2.834e0}$. This is similar to the interactions between hispanic & nodegree and edu:raceHispanic, where with a unit increase, the non-zero odd ratio is expected to increase by $e^{-3.869e0}$ and $e^{-5.629e-1}$.

A f chi-squared test on model 1 and model 2 is carried out, giving a p-value of 0.048. This tells us that interactions are statistically significant. Therefore, we can confirm that interactions do influence the model and influence the relationship between non-zero outcome and its predictors.

Model Selection



Forward AIC, backwards AIC, and step-wise AIC model selection were performed on Model 2. Both forward AIC and step-wise AIC returned the following x_i predictors: treat, age, re74, race, treat∗age, re74∗race, and treat∗race, while backwards AIC returns the following x_i predictors: treat, age, educ, race, married, nodegree, re74, treat∗race, treat∗age, race∗re74, race∗married, race∗nodegree, educ∗race, age∗nodegree.

The ROC plots are shown below for these 3 selections, as well as the residual binned plots are shown in Appendix 2.3. The residual binned plots show relatively supportive graphs as most of the points lie within the 95% confidence band. On the other hand, the AUC for backwards AIC gives 71.3%, while forward AIC and backwards AIC give 66.7%. Ultimately, backwards AIC was chosen for our final model due to the higher AUC and the higher number of interactions included in the backwards AIC.

Final Model

The x_i for the final model includes the following predictors: treat, age, educ, race, married, nodegree, re74, treat∗race, treat∗age, race∗re74, race∗married, race∗nodegree, educ∗race, age∗nodegree

The residual deviance (600.6) is lower than the null deviance (666.5), which also tells us that the predictors are better than the worst model.

Table 2: Logistic Regression Model Output

	Odds Ratio	Std. Error	t-value	p-value	95% CI
(Intercept)	0.226	1.276	0.177	0.859	(-2.3, 2.73)
treat	-0.929	1.068	-0.87	0.384	(-2.93, 1.36)
age	-0.026	0.02	-1.322	0.186	(-0.06, 0.01)
educ	0.089	0.086	1.035	0.301	(-0.08, 0.26)
racehispanic	8.651	2.743	3.154	0.002	(3.61, 14.52)
raceother	0.753	1.571	0.479	0.632	(-2.32, 3.86)
married	0.144	0.341	0.423	0.672	(-0.53, 0.82)
nodegree	1.88	0.807	2.331	0.02	(0.3, 3.48)
re74	0	0	4.187	0	(0, 0)
treat:re74	0	0	-2.051	0.04	(0, 0)
treat:racehispanic	14.748	635.911	0.023	0.981	(239.08, 276.85)
treat:raceother	-0.813	0.843	-0.965	0.335	(-2.78, 0.67)
treat:age	0.081	0.029	2.821	0.005	(0.03, 0.14)
racehispanic:married	-2.685	0.937	-2.864	0.004	(-4.67, -0.94)
raceother:married	0.321	0.513	0.626	0.532	(-0.67, 1.35)
racehispanic:nodegree	-3.348	1.319	-2.539	0.011	(-6.14, -0.9)
raceother:nodegree	-0.578	0.647	-0.893	0.372	(-1.86, 0.68)
educ:racehispanic	-0.491	0.174	-2.827	0.005	(-0.86, -0.17)
educ:raceother	-0.073	0.119	-0.61	0.542	(-0.31, 0.16)
age:nodegree	-0.048	0.023	-2.071	0.038	(-0.09, 0)

A VIF test was conducted (Appendix 2.5) to investigate the multicollinearity between the predictors. Treat predictor gave a 24.77, raceHispanic gave a 54.95, raceOther gave a 60.5, and nodegree gave a 14.74 VIF value. These categorical predictors resulted in VIF values of >= 10, which indicates high multicollinearity. Although this suggests high correlation for these predictors, it is not much of a concern as categorical predictors by default have high VIF values.

The race(Hispanic) gives a 0.0016 p-value, indicating that it is statistically significant and that we can reject the null hypothesis. As a unit of raceHispanic increases against the baseline (raceBlack), the odd ratio of non-zero wage is expected to increase by $e^{8.9e-2} = 1.09$. Although the p-value is high for other races, the odd ratio of non-zero wage is expected to increase by $e^{7.53e-1} = 2.12$ as raceOther increases per unit.

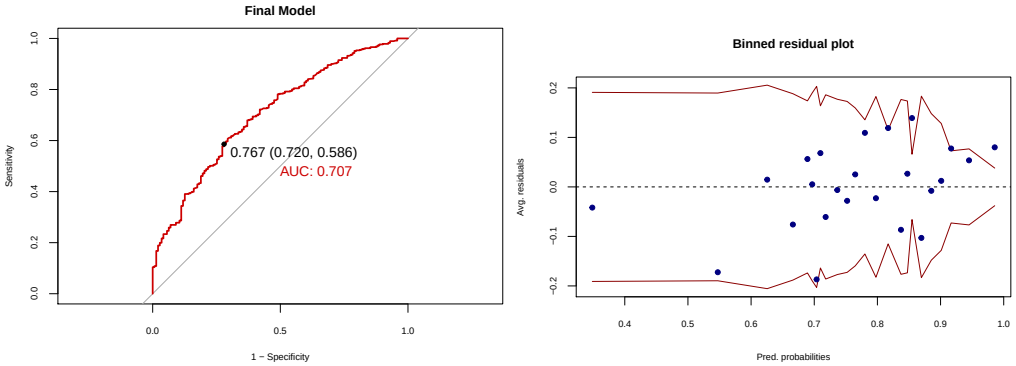
Several of the race interactions are shown to be statistically significant. The interaction between raceHispanic and married predictors give a 0.0048 p-value, suggesting that as this interaction increases per unit whilst keeping all the other predictors constant, the non-zero wage odd ratio is expected to increase by $e^{-2.69e0} = 0.068$. This is also the case for raceOther and no degree, where the odd ratio of non-zero wages is expected to

increase by $e^{-3.35} = 0.035$. Finally, the interaction between raceHispanic and education is also statistically significant, with a p-value of 0.0047, where the odd ratio of non-zero wages is expected to increase by $e^{-4.91e-1}$ =0.61 as the interaction unit increases.

This analysis shows that race is an important factor, where the effect of raceHispanic and most of its interactions contribute significantly to the overall model. To confirm this, a chi-squared test was performed on the final model with the race variable and interactions including race and on the final model without the race predictor and its interactions. The resulting p-value is 0.0035, which indicates that race is a contributing factor and is statistically significant to our model. However, it should be noted that even though race has significant effects on the odd-ratio of non-zero wages, the distribution of the different demographic groups is unbalanced as there isn't enough data on Hispanics.

Having no degree is also statistically significant with a low p-value of 0.02. It suggests that if a participant has no degree (against the baseline of having a degree), the odd ratio of non-zero wage is expected to increase by $e^{1.88} = 6.55$. The wage in 1974 predictor suggests that as re74 increases per unit, the odd ratio of non-zero wage is expected to increase by $e^{1.04e-4} = 1.00$. Re74 is statistically significant, and therefore we can reject the null hypothesis. The interaction of age and no degree (baseline is ~27 years old with a degree) is expected to increase the odd-ratio of the non-zero outcome by $e^{-4.8e-2} = 0.95$. It also has a low p-value of 0.038, proving to be statistically significant.

If a participant only received training while all the other predictors remained the same, the non-zero outcome odd ratio is expected to increase by $e^{-9.29e-01} = 0.4$. Although treat predictor itself does not have a low p-value to be statistically significant on its own, interactions between treat & re74 and treat & age are in fact statistically significant. Treat & re74 has a p-value of 0.040, where the odds of non-zero wages increases by $e^{-9.42e-5} = 1$ per unit increase (against the baseline of no treatment). Furthermore, as the interaction between treatment and age increases per unit, the odds of non-zero wages is expected to increase by $e^{8.07e-2}$ = 1.08. The low p-value of 0.0048 shows that this interaction is statistically significant. Performing on a 95% confidence prediction interval on the model gives $e^{-2.93e0}$, $e^{1.36e0}$. This tells us that the likely range of having treatment is (0.053, 3.91).



This residual binned plot shows that two of the data points lie outside of the 95% confidence band, with majority of the points lying randomly inside the band. This is overall a good residual binned plot. The accuracy of the final model is 61.72%, with a sensitivity of 58.6% and specificity of 72%. This gives a relatively moderate balance of sensitivity and specificity. Moreover, as shown, the AUC is 71.3%.

Conclusion

In conclusion, from the model assessment and EDA, workers who receive job training tend to be more likely to have positive wages than worked who do not receive training. The odd ratio of having non-zero wages is expected to increase by 0.4 if the participant received training. The likely range for the effect of training is (0.053, 3.91). There is sufficient evidence that the effects differ by demographic groups, particularly for Hispanic participants. This can be shown by the generally low p-values from raceHispanic predictors and its interaction with married, nodegree, and education. Despite this, there are several limitations to this report

that include the lack of data on Hispanic participants, the re75 predictor, and the accuracy of this model. Therefore, to improve this analysis, more and better quality data should be collected for the demographic groups as well as incorporating re75 should be considered to get a higher model accuracy.

Appendix

1.1

1.2

Table 3: Conditional table for wage given married status

	not_married0	married
wage0	0.2423398	0.2196078
wage1	0.7576602	0.7803922

Table 4: Conditional table for wage given degree

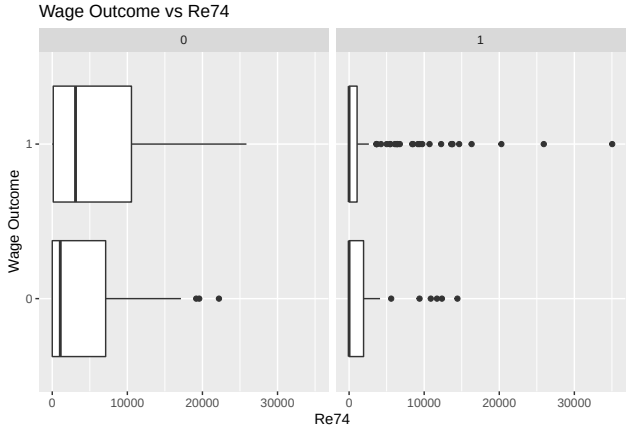
	nodegree	degree
wage0	0.215859	0.2428941
wage1	0.784141	0.7571059

Table 5: Conditional table for wage given race

	black	hispanic	other
wage0	0.2073579	0.1666667	0.2839506
wage1	0.7926421	0.8333333	0.7160494

```
##          educ
## wage_fact         0   1 2   3         4          5         6         7
##         0 0.3333333 0.5 0 0.4 0.3333333 0.08333333 0.5833333 0.3703704
##         1 0.6666667 0.5 1 0.6 0.6666667 0.91666667 0.4166667 0.6296296
##          educ
## wage_fact         8         9       10        11        12        13        14
##         0 0.3064516 0.1690141 0.183908 0.2315789 0.2165605 0.2592593 0.0952381
##         1 0.6935484 0.8309859 0.816092 0.7684211 0.7834395 0.7407407 0.9047619
##          educ
## wage_fact  15   16 17  18
##         0 0.3 0.25  0 0.5
##         1 0.7 0.75  1 0.5
```

1.3 RE74 v Outcome

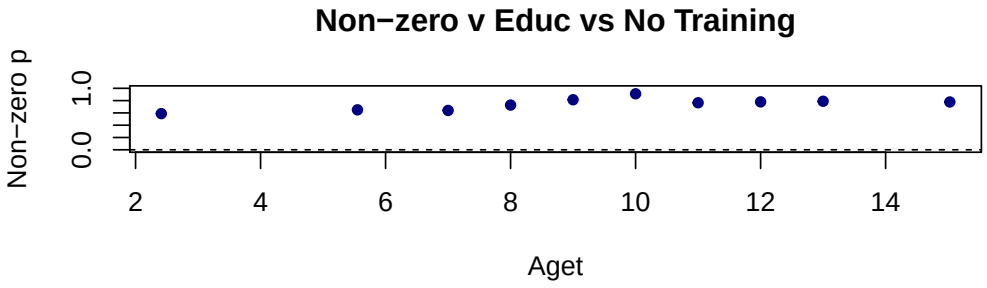


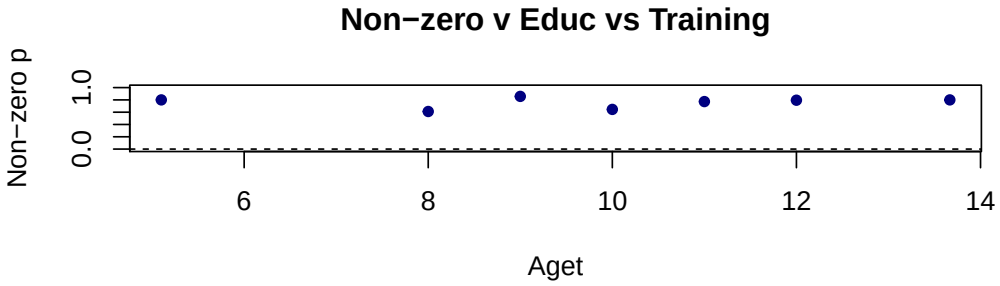
1.4

Table 6: Conditional probability of wage being non-zero given treat and education

treat0	treat1	nodegree
0.78	0.80	0
0.77	0.74	1

1.5





1.6 Conditional Probability

Table 7: Conditional probability of wage being non-zero given treat and race

treat0	treat1	race
0.79	0.89	black
0.80	1.00	hispanic
0.70	0.72	other

\

2.1

2.2

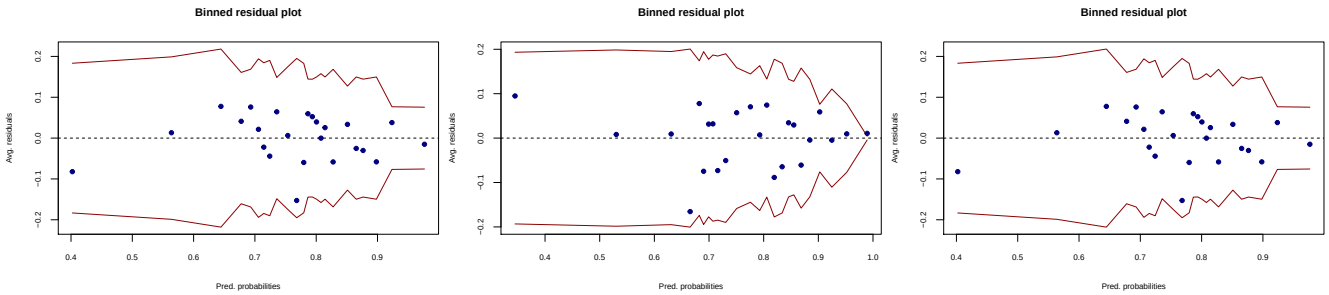
Step-wise AIC

Forward AIC

2.3

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.5296	0.8193	1.87	0.0619
treat	0.3892	0.2669	1.46	0.1448
age	-0.0389	0.0106	-3.66	0.0003
educ	0.0449	0.0526	0.85	0.3937
racehispanic	0.2848	0.3631	0.78	0.4329
raceother	-0.5155	0.2605	-1.98	0.0478
married	0.1031	0.2377	0.43	0.6644
nodegree	0.1194	0.2953	0.40	0.6860
re74	0.0001	0.0000	3.02	0.0026

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3231	0.3607	6.44	0.0000
treat	-0.9671	1.0552	-0.92	0.3594
age	-0.0554	0.0113	-4.92	0.0000
re74	0.0001	0.0000	4.12	0.0000
racehispanic	0.5069	0.4762	1.06	0.2871
raceother	-0.0409	0.3279	-0.12	0.9008
treat:age	0.0787	0.0276	2.85	0.0043
re74:racehispanic	-0.0001	0.0001	-1.52	0.1292
re74:raceother	-0.0001	0.0000	-2.30	0.0215
treat:racehispanic	14.1200	722.0313	0.02	0.9844
treat:raceother	-0.9900	0.8329	-1.19	0.2346



2.4 Confusion Matrix for 1) Forward AIC 2) Backward AIC 3) Stepwise AIC

Table 8: Prediction table for forward AIC

	0	1
0	85	183
1	58	288

Table 9: Accuracy for forwards AIC

	x
Accuracy	0.61

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.3231	0.3607	6.44	0.0000
treat	-0.9671	1.0552	-0.92	0.3594
age	-0.0554	0.0113	-4.92	0.0000
re74	0.0001	0.0000	4.12	0.0000
racehispanic	0.5069	0.4762	1.06	0.2871
raceother	-0.0409	0.3279	-0.12	0.9008
treat:age	0.0787	0.0276	2.85	0.0043
re74:racehispanic	-0.0001	0.0001	-1.52	0.1292
re74:raceother	-0.0001	0.0000	-2.30	0.0215
treat:racehispanic	14.1200	722.0313	0.02	0.9844
treat:raceother	-0.9900	0.8329	-1.19	0.2346

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0879	1.2837	0.07	0.9454
treat	-0.9711	1.0725	-0.91	0.3652
age	-0.0255	0.0198	-1.29	0.1979
educ	0.0943	0.0862	1.09	0.2740
racehispanic	9.5044	3.0618	3.10	0.0019
raceother	0.8750	1.5727	0.56	0.5780
married	0.0699	0.3487	0.20	0.8411
nodegree	1.9262	0.8094	2.38	0.0173
re74	0.0001	0.0000	3.99	0.0001
treat:racehispanic	14.7477	622.9618	0.02	0.9811
treat:raceother	-0.9763	0.8459	-1.15	0.2484
treat:age	0.0792	0.0285	2.78	0.0054
racehispanic:re74	0.0000	0.0001	0.32	0.7466
raceother:re74	-0.0001	0.0000	-2.48	0.0131
racehispanic:married	-2.9654	1.1413	-2.60	0.0094
raceother:married	0.5206	0.5279	0.99	0.3241
racehispanic:nodegree	-3.8357	1.5394	-2.49	0.0127
raceother:nodegree	-0.5997	0.6488	-0.92	0.3553
educ:racehispanic	-0.5367	0.1878	-2.86	0.0043
educ:raceother	-0.0583	0.1183	-0.49	0.6222
age:nodegree	-0.0477	0.0233	-2.04	0.0409

Table 10: Sensitivity and Specificity for forwards AIC

	x
Sensitivity	0.61
Specificity	0.59

Table 11: Prediction table for backwards AIC

	0	1
0	100	183
1	43	288

Table 12: Accuracy for backwards AIC

	x
Accuracy	0.63

Table 13: Sensitivity and Specificity for backwards AIC

	x
Sensitivity	0.61
Specificity	0.70

Table 14: Prediction table for stepwise AIC

	0	1
0	85	183
1	58	288

Table 15: Accuracy for stepwise AIC

	x
Accuracy	0.61

Table 16: Sensitivity and Specificity for stepwise AIC

	x
Sensitivity	0.61
Specificity	0.59

2.5 VIF

Table 17: VIF for final model

	x
treat	24.77
age	3.93
educ	5.20
racehispanic	54.95
raceother	60.46
married	2.71
nodegree	14.74
re74	1.85
treat:re74	1.62
treat:racehispanic	1.00
treat:raceother	15.00
treat:age	12.72

	x
racehispanic:married	4.58
raceother:married	2.11
racehispanic:nodegree	10.58
raceother:nodegree	9.13
educ:racehispanic	19.45
educ:raceother	39.53
age:nodegree	13.34