Team Project Logistic

Jaryl Ngan

30 September 2019

Summary

Introduction

The aim of this study is identify whether giving job training to workers has a positive effect in helping them attain a non-zero wage job after the training is complete. The workers started receiving job training in 1974 and finished their training in 1978. We will be using a logistic regression in this analysis. There are some other variables in the data, we will try to take those variables into account and determine if there is an association between receiving job training and wages in 1978. Finally, we will see if there are any other interesting associations which affect wages in 1978.

Data

First, we created the response variable and named it outcome. It is binary categorical variable which is 1 if the person has a positive wage in 1978 and 0 if the person has a zero wage in 1978.

A simple check was done to see if there was any missing data and unforunately, we do not have any, thus we do not need any methods to handle them.

The X column seems like a primary key, it seems hard to extract any informative features from it, thus we will not consider it.

We created a binary categorical variable called prevemp which is 1 if the worker had a non-zero wage in 1974 and 0 if the worker had a zero wage in 1974.

From the exploratory data analysis, we decided to create a new binary categorical variable from education called edu1 with 1 indicating that the individual had more than 8 years of education and 0 otherwise.

We also split the dataset into a train set and test set. We will train our model on 80% of the data and do an out sample test with the remaining 20% of the data to see if we have overfit our model.

Exploratory Data Analysis

First, we shall get an idea of the response variable, outcome. The number of workers who have a positive wage and those who have no wage using a table since our response variable is a binary variable. From the Table 1 in the appendix, we see that there are 471 people with positive wages and 143 people without wages. We therefore, have a somewhat unbalanced data set. This may give us the illusion of a good model, as a naive model which predicts everyone to have a positive wage will have a decently high accuracy, however this would not be a good model.

Table 1

```
table(empdata$outcome)
```

```
##
##   0   1
## 143 471
```

Next, we will use the joint probability and conditional probability tables to explore the other categorical variables

We explore to see whether there is a relationship between being married and having a positive wage , the conditional probability table seems to indicate that being married or not does not affect the probability of having a positive wage. This is verified by doing a Chi-Squared test which gives a p-value of 0.5756, thus we do not reject the null hypothesis that the variables are independent. The relevant tables can be found from Table 2.1-2.2 in the appendix.

Table 2.1 - 2.2

```
#married
table(empdata[,c("outcome","married")])/sum(table(empdata[,c("outcome","married")]))
```

```
##        married
## outcome          0          1
##       0 0.14169381 0.09120521
##       1 0.44299674 0.32410423
```

```
apply(table(empdata[,c("outcome","married")])/sum(table(empdata[,c("outcome","married")])),
      2,function(x) x/sum(x))
```

```
##        married
## outcome         0         1
##       0 0.2423398 0.2196078
##       1 0.7576602 0.7803922
```

```
chisq.test(table(empdata[,c("outcome","married")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(empdata[, c("outcome", "married")])
## X-squared = 0.31339, df = 1, p-value = 0.5756
```

We explore to see whether there is a relationship between being black and having a positive wage, the conditional probability table seems to indicate that being black does affect the probability of having a positive wage. This is verified by doing a Chi-Squared test which gives a p-value of 0.0201, thus we do have strong evidence to reject the null hypothesis that the variables are independent. The relevant tables can be found from Table 3.1-3.2 in the appendix.

Table 3.1-3.2

```
#black
table(empdata[,c("outcome","black")])/sum(table(empdata[,c("outcome","black")]))
```

```
##        black
## outcome         0         1
##       0 0.1205212 0.1123779
##       1 0.4837134 0.2833876
```

```r
apply(table(empdata[,c("outcome","black")])/sum(table(empdata[,c("outcome","black")])),
      2,function(x) x/sum(x))
```

```
##        black
## outcome         0         1
##       0 0.1994609 0.2839506
##       1 0.8005391 0.7160494
```

```r
chisq.test(table(empdata[,c("outcome","black")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(empdata[, c("outcome", "black")])
## X-squared = 5.4034, df = 1, p-value = 0.0201
```

We explore to see whether there is a relationship between being hispanic and having a positive wage, the conditional probability table seems to indicate that being hispanic does affect the probability of having a positive wage. However, the Chi-Squared test returns a p-value of 0.205. This means there is no evidence to reject the null hypothesis that the variables are independent. This is a somewhat surprising result. From the joint probability tables, we see that there is very little data for hispanics, thus the conclusions drawn from this variable may not be very accurate. The relevant tables can be found from Table 4.1-4.2 in the appendix.

Table 4.1-4.2

```r
#hisp
table(empdata[,c("outcome","hispan")])/sum(table(empdata[,c("outcome","hispan")]))
```

```
##        hispan
## outcome          0          1
##       0 0.21335505 0.01954397
##       1 0.66938111 0.09771987
```

```r
apply(table(empdata[,c("outcome","hispan")])/sum(table(empdata[,c("outcome","hispan")])),
      2,function(x) x/sum(x))
```

```
##        hispan
## outcome         0         1
##       0 0.2416974 0.1666667
##       1 0.7583026 0.8333333
```

```r
chisq.test(table(empdata[,c("outcome","hispan")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(empdata[, c("outcome", "hispan")])
## X-squared = 1.6048, df = 1, p-value = 0.2052
```

#treat

We explore to see whether there is a relationship between receiving job training and having a poistive wage , the conditional probability table seems to indicate that receiving job training does not affect the probability of having a positive wage. This is verified by doing a Chi-Squared test which gives a p-value of 0.769, thus we do not reject the null hypothesis that the variables are independent. The relevant tables can be found from Table 5.1-5.2 in the appendix.

Table 5.1-5.2

```
#treat
table(empdata[,c("outcome","treat")])
```

```
##        treat
## outcome   0   1
##       0  98  45
##       1 331 140
```

```
apply(table(empdata[,c("outcome","treat")])/sum(table(empdata[,c("outcome","treat")])),
      2,function(x) x/sum(x))
```

```
##        treat
## outcome         0         1
##       0 0.2284382 0.2432432
##       1 0.7715618 0.7567568
```

```
chisq.test(table(empdata[,c("outcome","treat")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(empdata[, c("outcome", "treat")])
## X-squared = 0.086541, df = 1, p-value = 0.7686
```

We explore to see whether there is a relationship between having a college degree and having a positive wage , the conditional probability table seems to indicate that having a college degree or not does not affect the probability of having a positive wage. This is verified by doing a Chi-Squared test which gives a p-value of 0.5053, thus we do not reject the null hypothesis that the variables are independent. The relevant tables can be found from Table 6.1-6.2 in the appendix.

Table 6.1-6.2

```
#nodegree
table(empdata[,c("outcome","nodegree")])
```

```
##        nodegree
## outcome   0   1
##       0  49  94
##       1 178 293
```

```
apply(table(empdata[,c("outcome","nodegree")])/sum(table(empdata[,c("outcome","nodegree")])),
      2,function(x) x/sum(x))
```

```
##         nodegree
## outcome         0         1
##       0 0.215859 0.2428941
##       1 0.784141 0.7571059
```

```
chisq.test(table(empdata[,c("outcome","nodegree")]))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(empdata[, c("outcome", "nodegree")])
## X-squared = 0.44379, df = 1, p-value = 0.5053
```

Now, we will check to see if there are any interesting association effect between getting job training and the demographic effects on getting a positive wage.

The only notable interesting categorical interaction effect on getting a positive wage is between being black and receiving job treatment. As all the other demographic information do not seem to have an effect on getting a positive wage.

The joint probabilities of getting a positive wage and receiving job training for black people varies drastically from that of non-black people. This suggests that there is a possible interaction between being black and receiving job training on getting a positive wage, we will explore this later model building. The relevant tables can be found from Table 6.1 - 6.2 in the appendix.

Table 6.1 - 6.2

```
table(empdata[empdata$black == 1,c("outcome","treat")])/sum(table(empdata[empdata$black == 1,c("outcome
```

```
##         treat
## outcome         0         1
##       0 0.1069959 0.1769547
##       1 0.2510288 0.4650206
```

```
table(empdata[empdata$black == 1,c("outcome","treat")])/sum(table(empdata[empdata$black == 1,c("outcome
```

```
##         treat
## outcome         0         1
##       0 0.1069959 0.1769547
##       1 0.2510288 0.4650206
```

We will now explore the relationship between the continuous variables and the response variable. We will do so mainly using boxplots and binned plots. There are four continuous variables available, age, years of education, income in 1974 denoted as re74 and income in 1975 denoted as re75.

Figure 1.1 - 1.4

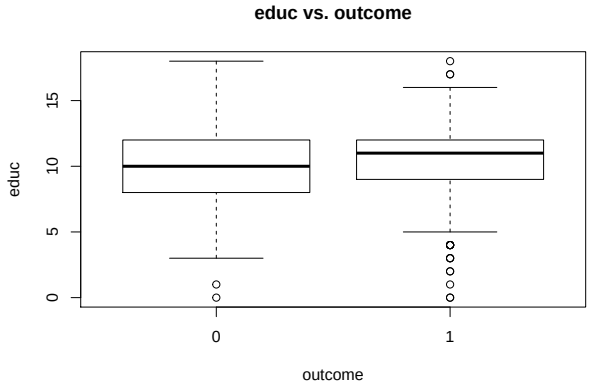


Figure 1.2

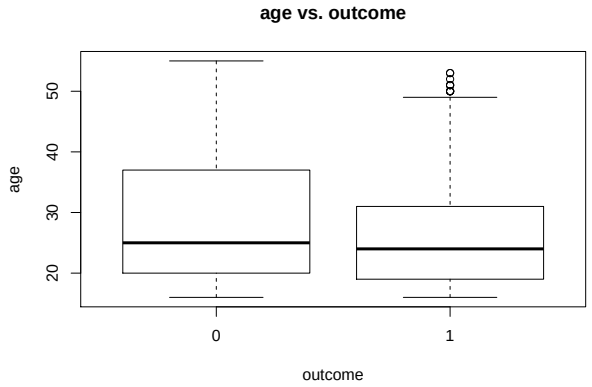


Figure 1.3

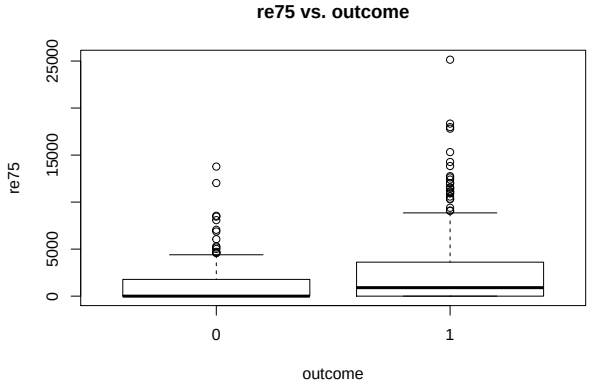
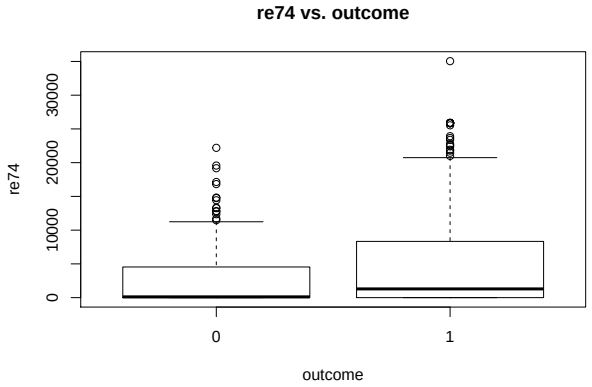


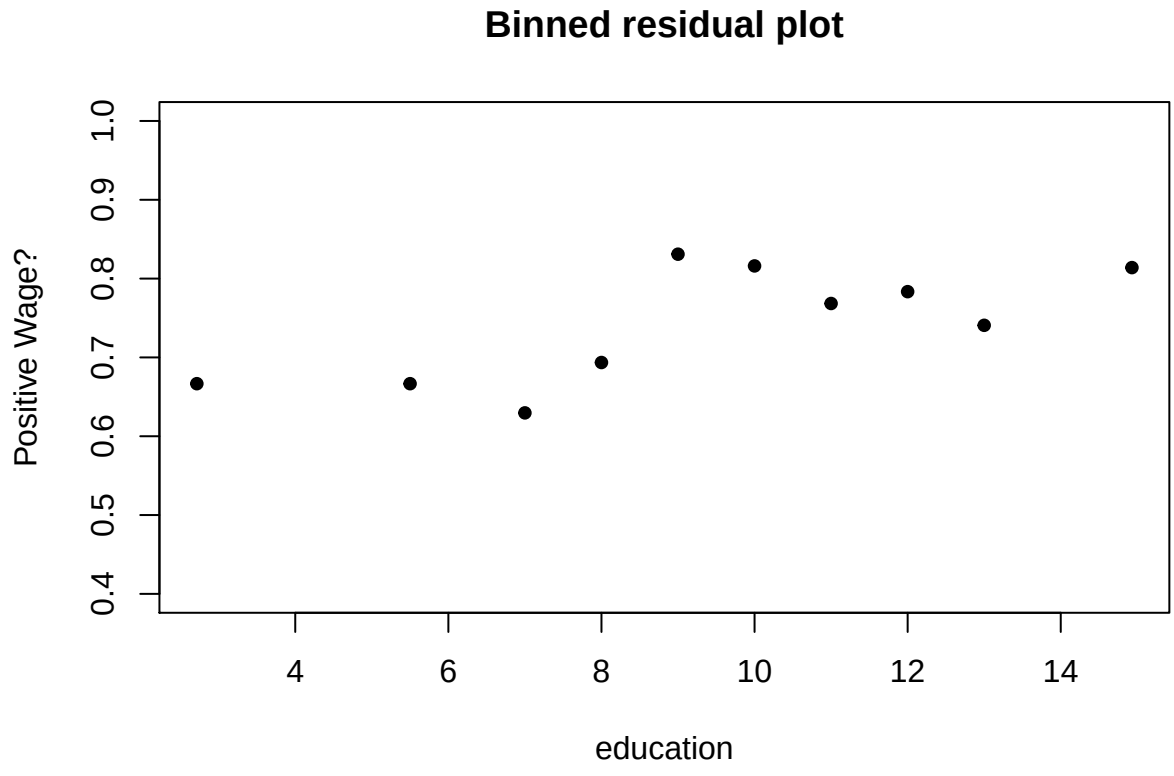
Figure 1.4



People who had a postive wage tended on average, to have more years of education. There was little difference between the mean age for those who had a positive wage and those who did not. People who had a positive wage tended to have on average, have a higher wage in 1974 and 1975. The boxplots of all the four continuous variables against the response variables can be seen in Figure 1.1 - 1.4 in the appendix.

Next, we will look at the binned plots. Figure 2.1

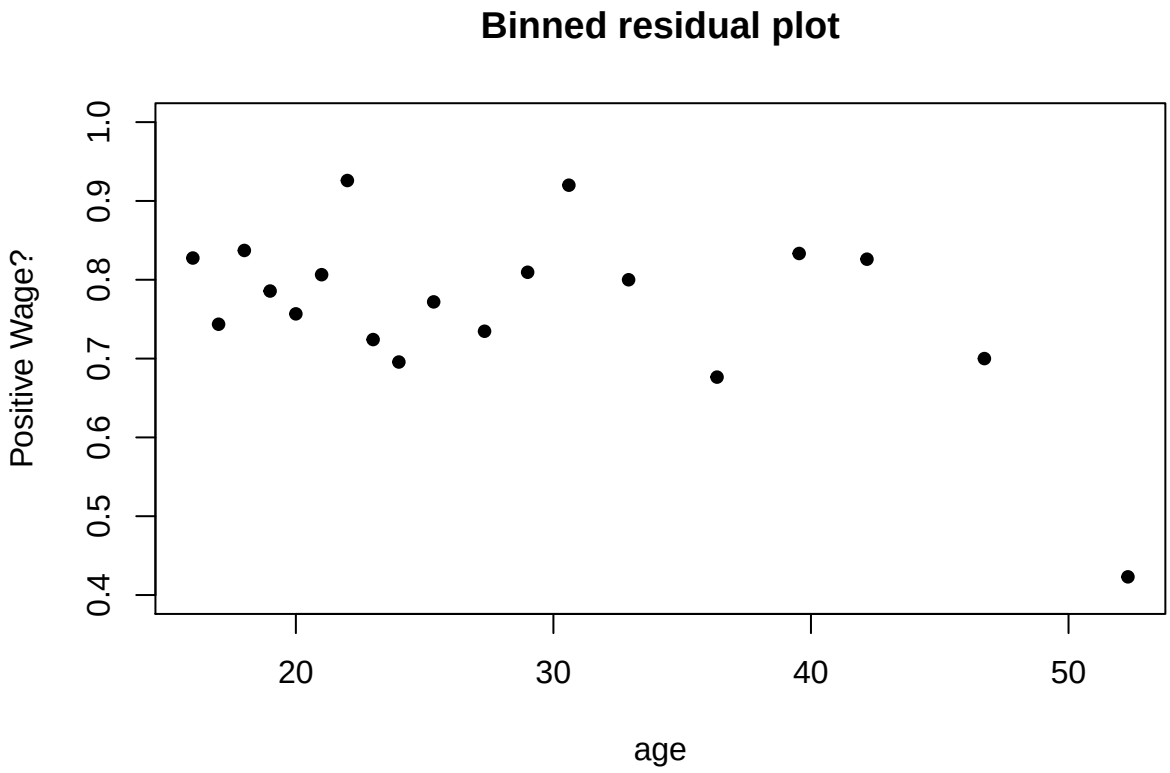
```
binnedplot(y=empdata$outcome, x=empdata$educ, xlab='education', ylab = 'Positive Wage?', ylim=c(0.4,1))
```



For years of education, there seems to be significant change in the probability of having a positive wage after 9 years of education. 9 years of education in the US system seems to be the time when they complete their junior high. We can consider creating a binary categorical variable that splits the people who have

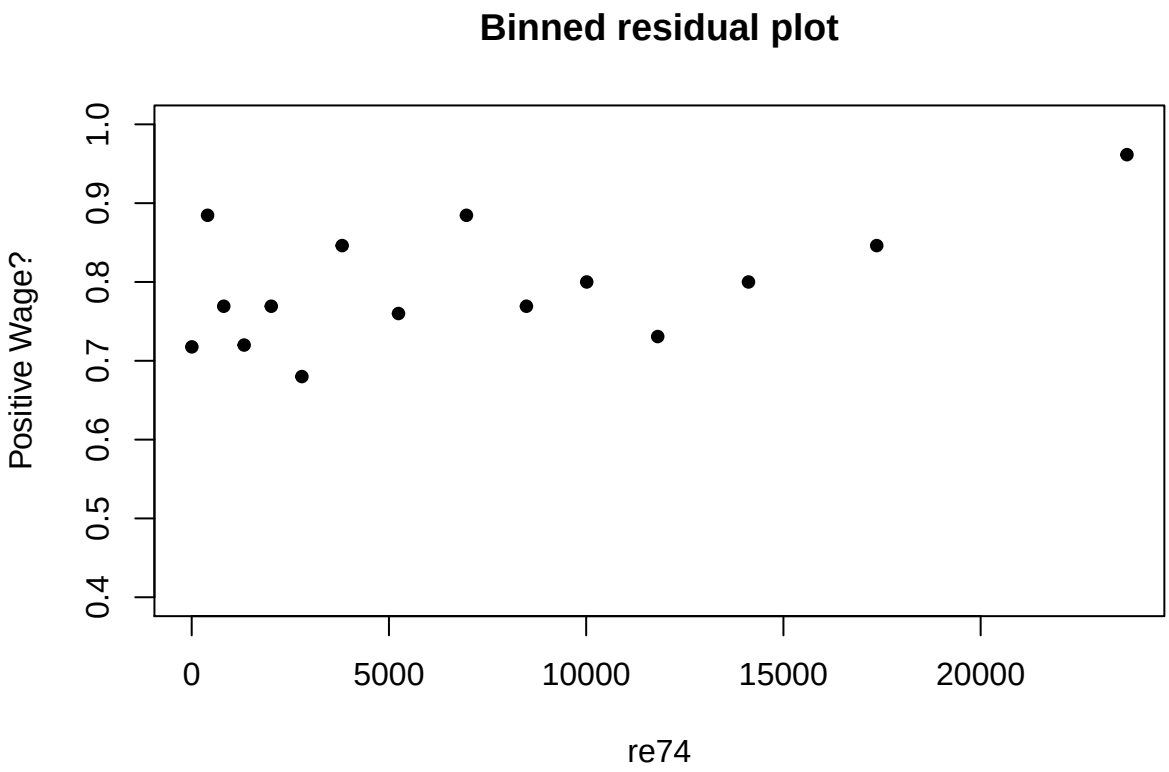
complete more than 8 years of education and those who have not. The binned plot is shown in Figure 2.1 in the appendix.

```
binnedplot(y=empdata$outcome, x=empdata$age, xlab='age', ylab = 'Positive Wage?', ylim=c(0.4,1))
```



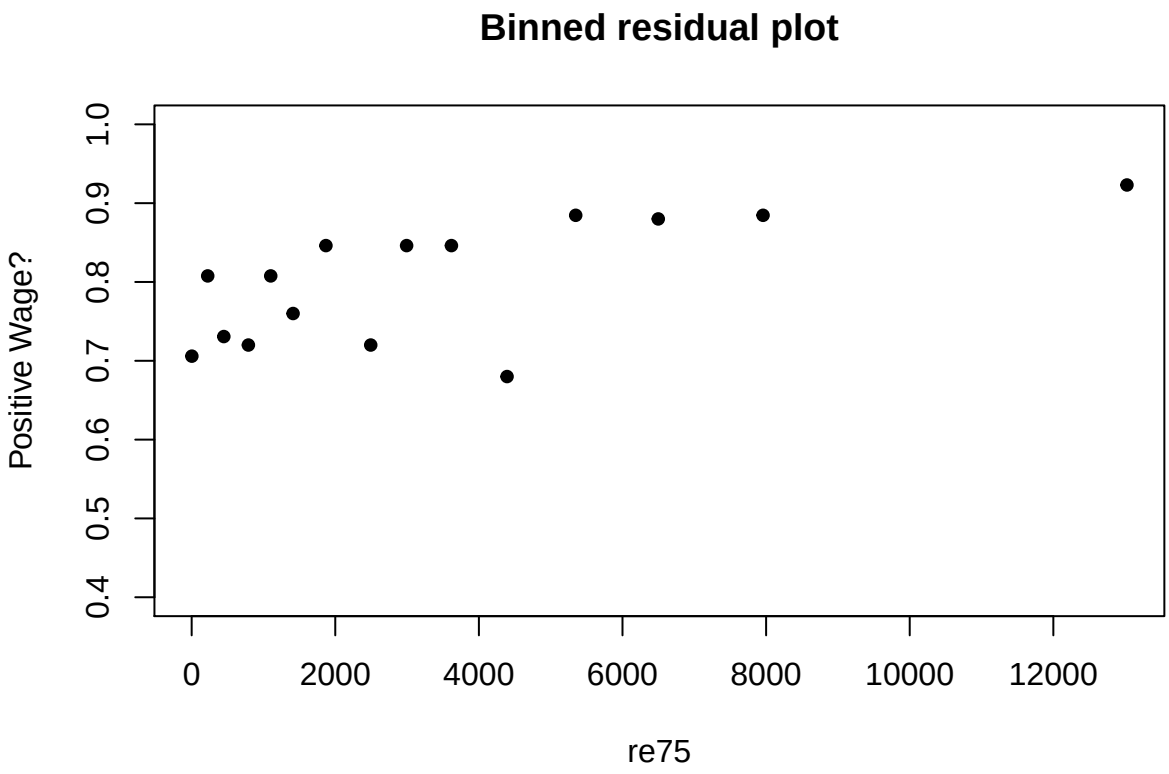
There seems to be a trignometric relationship for age, however, from the boxplots age does not seem like it will be a useful predictor, so we will not try to fix the relationship for now. The binned plot is shown in Figure 2.2 in the appendix.

```
binnedplot(y=empdata$outcome, x=empdata$re74, xlab='re74', ylab = 'Positive Wage?', ylim=c(0.4,1))
```



The binned plots of the response variable on re74 and re75 seem to show a linear relationship, this is acceptable for our case as we can use a inverse logit function that is approximately linear. The binned plots are shown in Figure 2.3-2.4 in the appendix.

```
binnedplot(y=empdata$outcome, x=empdata$re75, xlab='re75', ylab = 'Positive Wage?', ylim=c(0.4,1))
```



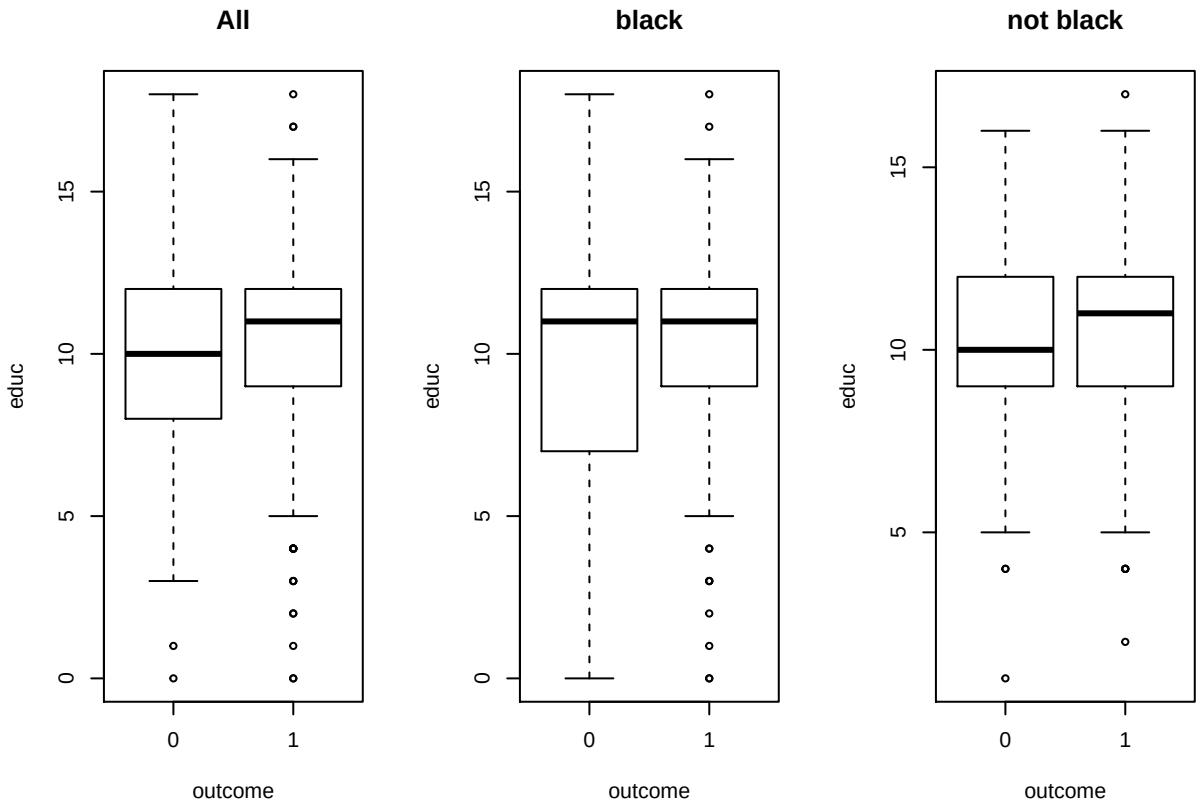
Following this, we will check for any other interesting interactions terms between categorical - continuous variables.

Figure 3.1

```
par(mfrow=c(1,3))
boxplot(educ~outcome, data=empdata, main='All',main='educ vs. outcome')
```

```
## Warning in bxp(list(stats = structure(c(3, 8, 10, 12, 18, 5, 9, 11, 12, :
## Duplicated argument main = "educ vs. outcome" is disregarded
```

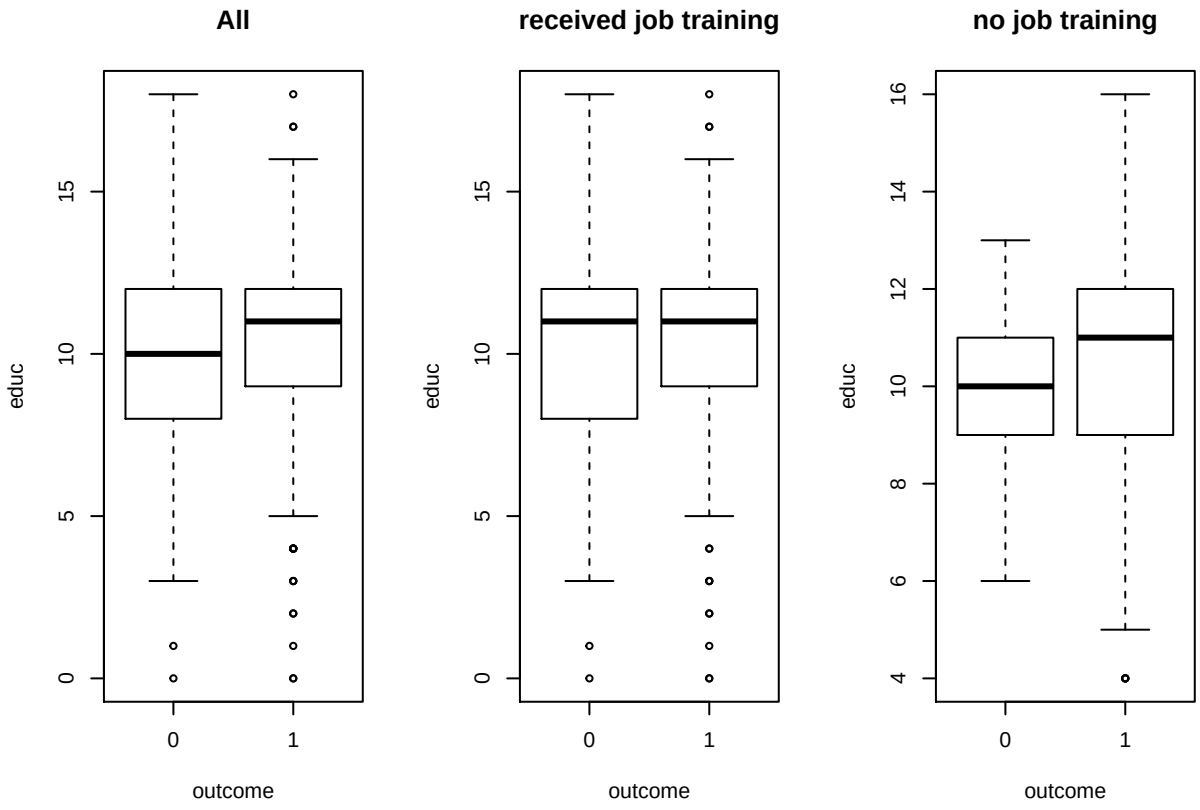
```
boxplot(educ~outcome, subset=(black==0),data=empdata, main='black')
boxplot(educ~outcome, subset=(black==1),data=empdata, main='not black')
```



The variables black and education both seem to be useful predictors for the response variable, thus we will experiment to see if there is any indication of an interaction between them. The mean and the distribution between black and non-black people seem the same, which does not suggest there is an interaction term. The plot can be seen in Figure 3.1 of the appendix.

Figure 3.2

```
par(mfrow=c(1,3))
boxplot(educ~outcome, data=empdata, main='All')
boxplot(educ~outcome, subset=(treat==0),data=empdata, main='received job training')
boxplot(educ~outcome, subset=(treat==1),data=empdata, main='no job training')
```



One of the main research questions is about the effectiveness of job training. Thus, although job training did not seem to be a strong predictor variable, we will check if it has any interaction terms. There seems to be some indication of an interaction between them as the mean and distribution of the years of education between people who received job training and those who did not receiving training differs.

By the similar reasoning as above, re74 and black, re75 and black, re75 and treat, re75 and treat seem to have possible interaction effects. We will explore this further in model building. The plots can be seen from Figure 3.3-3.6

Figure 3.3

```
par(mfrow=c(1,3))
boxplot(re74~outcome, data=empdata, main='re74 vs. outcome')
boxplot(re74~outcome, subset=(treat==0),data=empdata, main='black')
boxplot(re74~outcome, subset=(treat==1),data=empdata, main='not black')
```

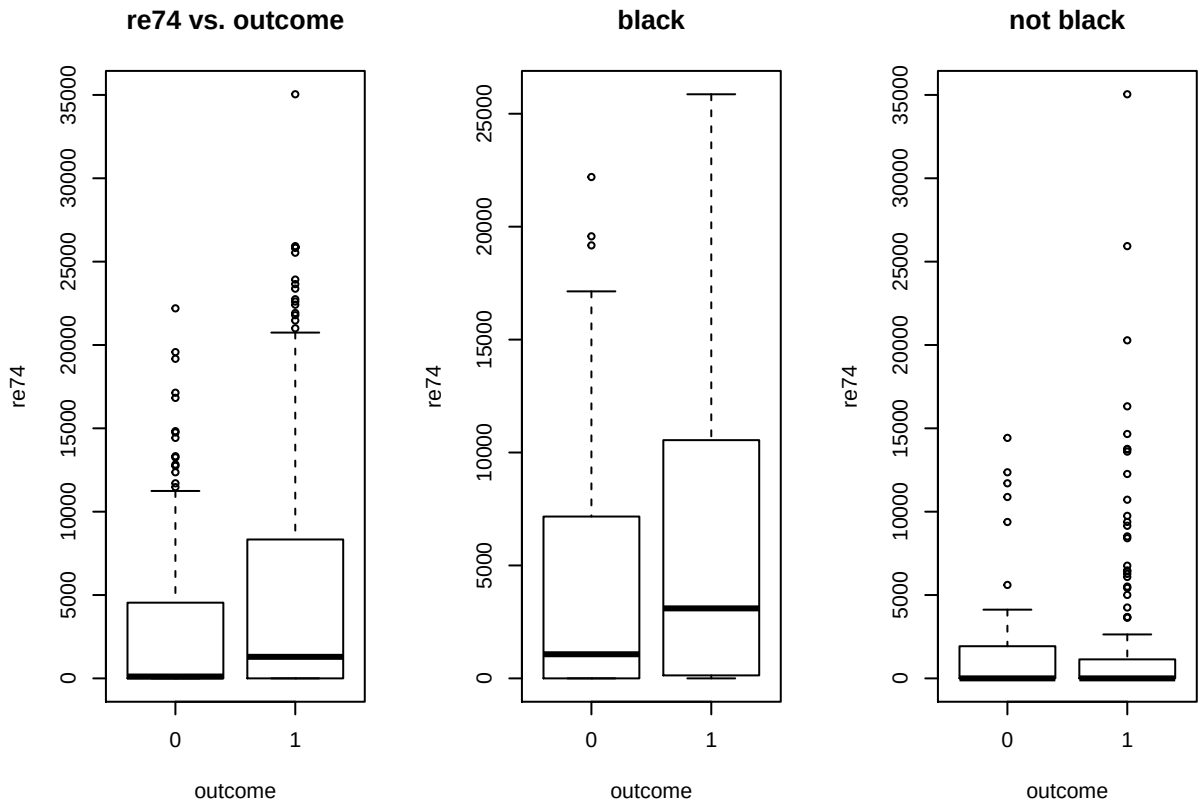


Figure 3.4

```
par(mfrow=c(1,3))
boxplot(re75~outcome, data=empdata, main='age vs. outcome')
boxplot(re75~outcome, subset=(treat==0),data=empdata, main='treat')
boxplot(re75~outcome, subset=(treat==1),data=empdata, main='no treat')
```

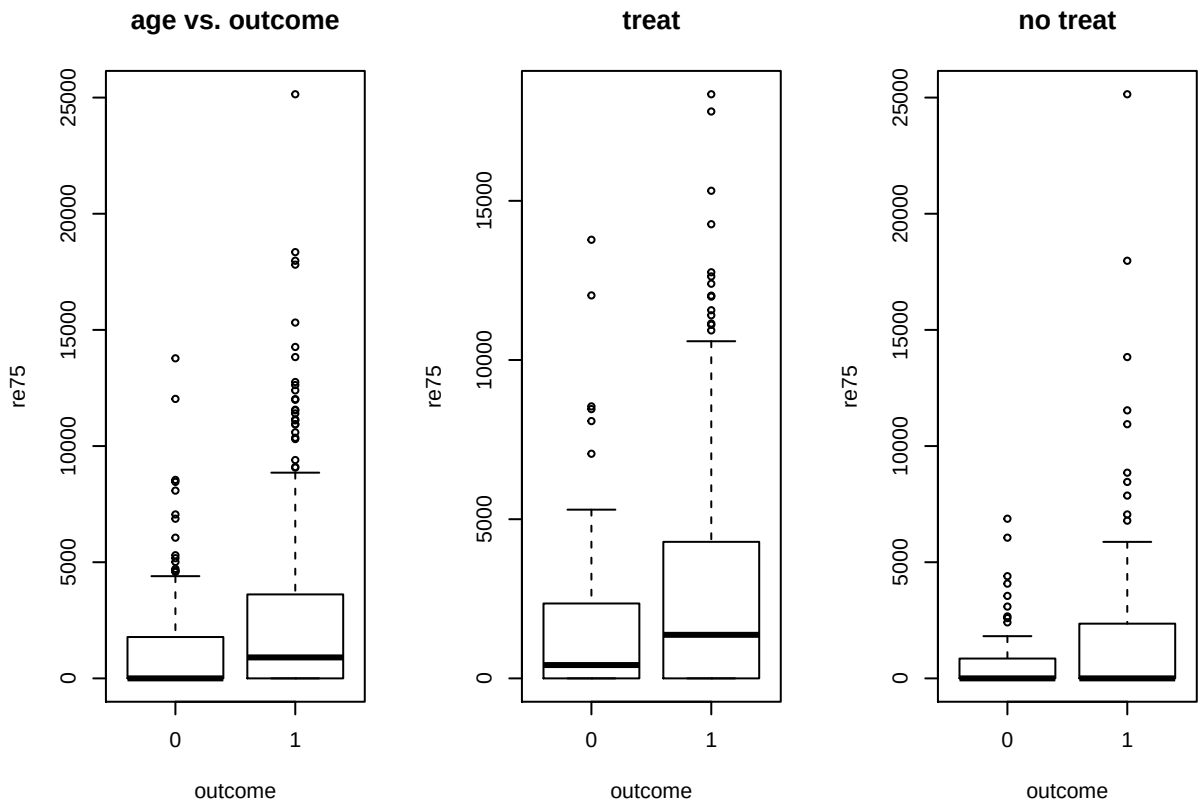


Figure 3.5

```
par(mfrow=c(1,3))
boxplot(re74~outcome, data=empdata, main='re74 vs. outcome')
boxplot(re74~outcome, subset=(black==0),data=empdata, main='black')
boxplot(re74~outcome, subset=(black==1),data=empdata, main='not black')
```

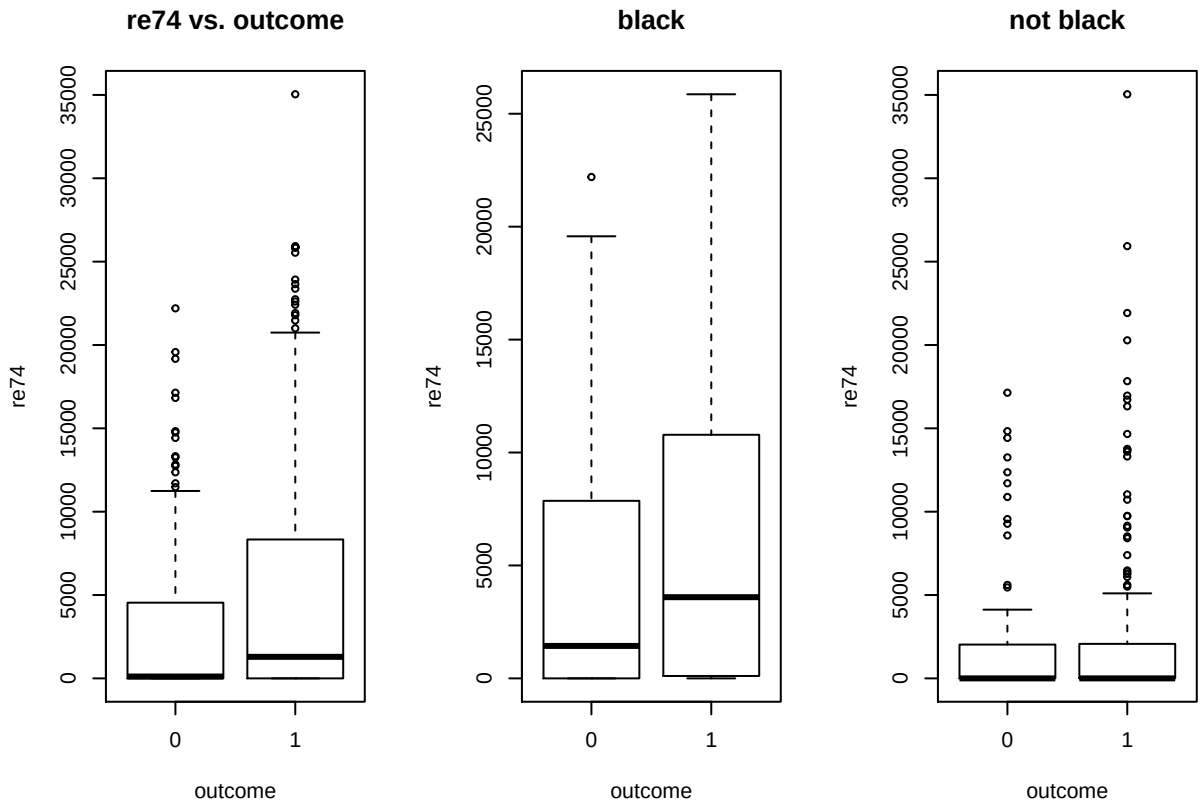
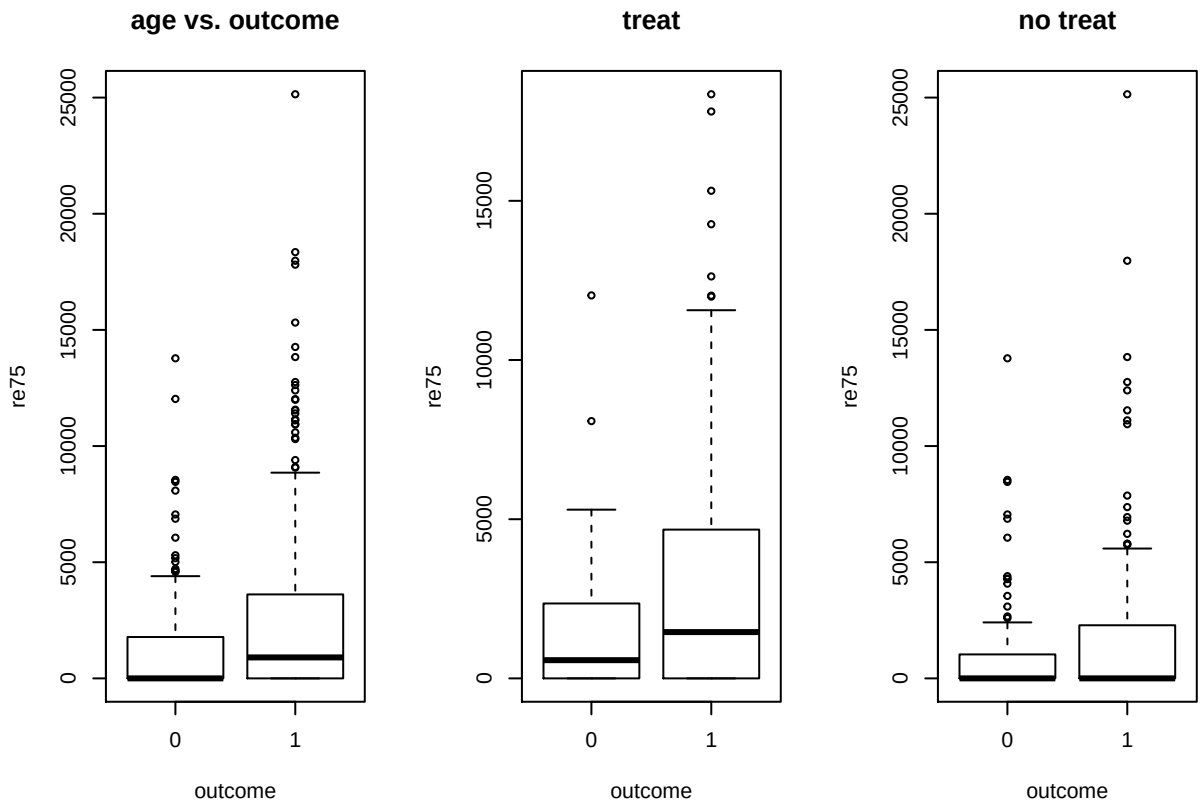


Figure 3.6

```
par(mfrow=c(1,3))
boxplot(re75~outcome, data=empdata, main='age vs. outcome')
boxplot(re75~outcome, subset=(black==0),data=empdata, main='treat')
boxplot(re75~outcome, subset=(black==1),data=empdata, main='no treat')
```



Model

We started by creating a model with all the predictors (*age*, *educ1*, *black*, *hispan*, *married*, *re74*, *75*) and then used step wise AIC method to eleminate variables and identify the most influeincing variables. According to this method the influencing variables turned out to be (*age*, *educ1*, *black*, *re74*, *75*) The p-value of these variables suggested that there is strong relationship between the above mentioned variables and *outcome*.

The AIC model did not include the treatment variable, but since we were specifically interested to know if treatment had any effect on the odds of having non-zero wages we included this variable in our analysis.

1. Model 1

From the EDA there seemed to be interaction effect between the variables *treat* and *age*. With variables recognized by AIC method and *treat* as the basic model, we expanded the model by adding an interaction term between the variables *treat* and *age*. The nested F-test confirmed that addition of the interaction term is significant.

2. Model 2

We were also interested to know if the ethnicity had any particular effect on odds of having non-zero wages. We found an interesting interaction between *hispan* and *educ1* variables. Hence in the next model we included *hispan* and the interaction term. Both the newly included terms had low p-values and hence were retained them in the model.

3. Model 3

EDA indicated some intercation between *re74* and *black* variables. We inlcuded this interaction term in this model. However with the additon of the ethnicity variable the significance of the race variable (*black*) reduced. Hence we did not include *hispan* and the related interaction term in this model. Since this model and *model 2* are not nested we did not compare them. A comaprison with the basic AIC model confirmed significant difference due to the addition of the interaction term.

The following table summarizes the accuracy, sensitivity, specifcity and Area Under the Curve (AUC) for the basic AIC model and the three models enlisted above.

Model	Acc	Sensitivity	Specificity	AUC
model 1	0.5651466	0.5244161	0.6993007	65.2%
model 2	0.5716612	0.5286624	0.7132867	66.2%
model 3	0.5114007	0.4140127	0.8321678	67.7%
model AIC	0.6237785	0.6433121	0.5594406	65.9%

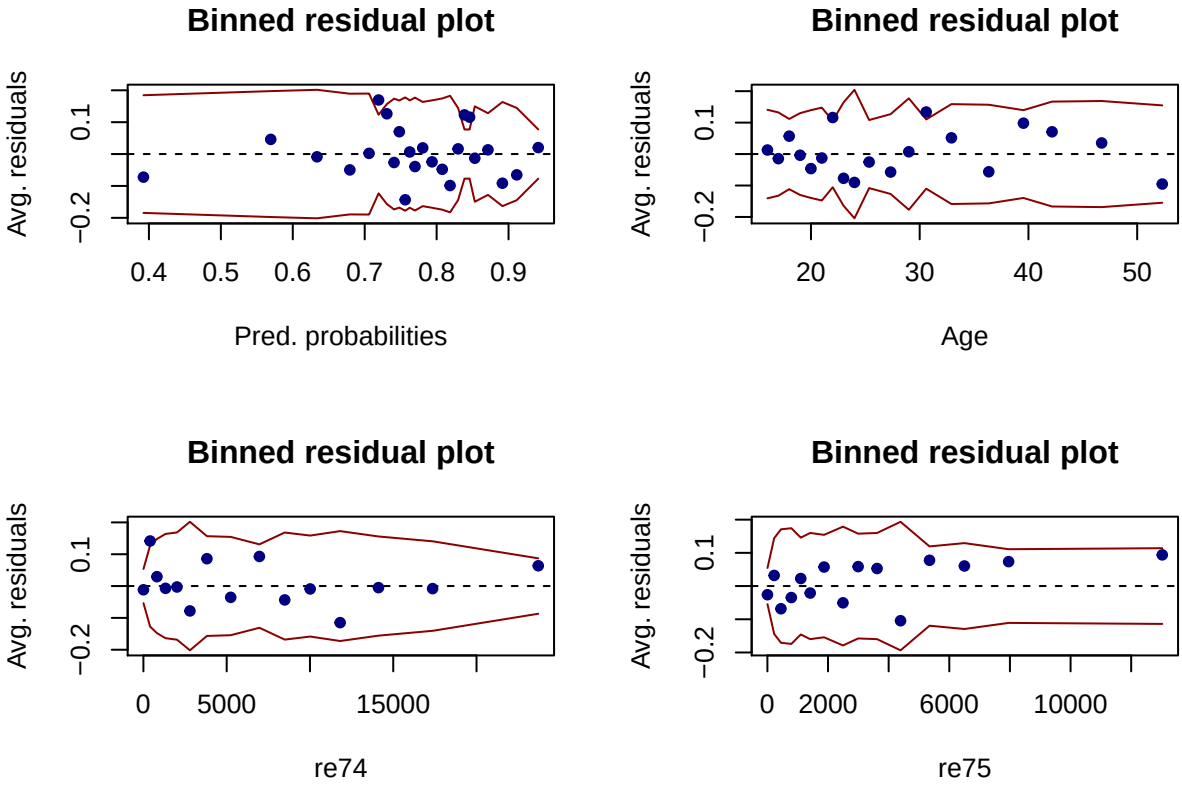
From the table we concluded that *model 1* had the highest accuracy, however, the balance between sensitivity and specificity seemed poor for this model. High sensitivity can be achieved by high threshold value as well, but since the outcome variable is skewed towards 1, correct modelling to get better specificity is imporant. For the second model as well the specificity was not very high, and the sensitivity-specificity difference was significant. Model 3 amongst all the four models had the best balance between sensitivity and specificity. As compared to the AIC model it had higher accuracy as well as higher AUC. For all the above mentioned reasons we selected model 4 as the final model.

The values of the weight estimates, standard error, z value and p-value for the final model are summarized in the table below.

term	estimate	std.error	statistic	p.value
(Intercept)	1.1068599	0.2142582	5.166011	0.0000002
educ1	0.4219009	0.2331017	1.809943	0.0703046
age_c	-0.0514293	0.0115105	-4.468044	0.0000079
re74_c	0.0000896	0.0000247	3.631884	0.0002814
treat	0.4384026	0.2684940	1.632821	0.1025067
black1	-0.7421387	0.2525436	-2.938656	0.0032964
age_c:treat	0.0732961	0.0272283	2.691910	0.0071044
re74_c:black1	-0.0000685	0.0000381	-1.797141	0.0723132

For the final model, we analyzed the binned residual plots (figures in Appendix)

1. From the residuals vs. fitted values there was no evident pattern in the plot, only one point was outside the confidence interval.
2. The residuals vs. age and residuals vs. *re74* plot did not have any specific pattern that we should have accounted for. One point lied outside the confidence interval for both the plots respectively
3. The mean residuals for *treat*, *educ1* and *black* also did not have any specific patterns.



Results

1. Question 1: *Quantify the effect of the treatment, that is, receiving job training, on the odds of having non-zero wages.*

Answer: From the final model, we conclude that increase in odds ratio for an individual given job training vs. an individual not given job training is 58% (with *age*, *re74*, *educ1* and *black* held constant). For individuals given job training, an incease in age by one year would increase odds of getting job by 7.3%.

2. Question 2: *What is a likely range for the effect of training?*

Answer: Confidence range for the odds ratio of getting a job for treatment is (-7%, 172%).

3. Question 3: *Is there any evidence that the effects differ by demographic groups?*

Answer:

- The final model does not include *hispan* and *married* variable. We included these variables one-by-one and did a nested F-test to check the significance of both the variables. The p-value of both the tests were high and hence we concluded that the variables *hispan* and *married* do not have a significant effect on odds on getting job.
- There is a negative effect of increase in age on odds of getting a job. For an increase in age by one year the odds of getting a job decrease by 5%.

- For black individuals the odds of getting job is 48% less than non-black individuals.

4. *Are there other interesting associations with positive wages that are worth mentioning?*

Answer:

- For a $1000 higher wage in 1974 the odds of getting a job increase by 6%. Whereas for a $1000 higher wage in 1975 the odds of getting a job increase by 12%.

We also did train-test split to check how well the model performs out of sample. We did 80%-20% split and trained the model with training data and checked the accuracy, specificity and sensitivity on test data. The table below summarized the performance.

```
set.seed(23)
sample <- sample.int(n = nrow(empdata), size = floor(.8*nrow(empdata)), replace = F)
empdata_train <- empdata[sample, ]
empdata_test  <- empdata[-sample, ]

model_4_t <- glm(formula = outcome ~ educ1 + age + re74 + treat + treat:age + black + re74:black + re75
summary(model_4)
```

```
##
## Call:
## glm(formula = outcome ~ educ1 + age_c + re74_c + treat + treat:age_c +
##     black + black:re74_c, family = binomial, data = empdata[c(-1,
##     -11)])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3785   0.3512   0.6054   0.7512   1.5696
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.107e+00  2.143e-01   5.166 2.39e-07 ***
## educ1          4.219e-01  2.331e-01   1.810 0.070305 .
## age_c         -5.143e-02  1.151e-02  -4.468 7.89e-06 ***
## re74_c         8.958e-05  2.467e-05   3.632 0.000281 ***
## treat          4.384e-01  2.685e-01   1.633 0.102507
## black1        -7.421e-01  2.525e-01  -2.939 0.003296 **
## age_c:treat    7.330e-02  2.723e-02   2.692 0.007104 **
## re74_c:black1 -6.849e-05  3.811e-05  -1.797 0.072313 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 666.50  on 613  degrees of freedom
## Residual deviance: 622.07  on 606  degrees of freedom
## AIC: 638.07
##
## Number of Fisher Scoring iterations: 4
```

```
res_test_4 <- empdata_test$outcome - predict(model_4_t, newdata = empdata_test, type="response")

Conf_mat_train_4 <- confusionMatrix(as.factor(ifelse(fitted(model_4_t) >= 0.77, "1","0")),
                                    as.factor(empdata_train$outcome),positive = "1")
Conf_mat_test_4 <- confusionMatrix(as.factor(ifelse(predict(model_4_t, newdata = empdata_test,
                   type="response")>=0.77,"1","0")), as.factor(empdata_test$outcome), positive = "1")

test_train_table <- data.frame("Data"=c("Train","Test"), "Accuracy"=c(Conf_mat_train_4$overall["Accuracy
                   Conf_mat_test_4$overall["Accuracy"]), "Sensitivity"=c(Conf_mat_train_4$byClass["Sensi
                   Conf_mat_test_4$byClass["Sensitivity"]), "Specificity"=c(Conf_mat_train_4$byClass["Spe
                   Conf_mat_test_4$byClass["Specificity"]))

kable(test_train_table) %>% kable_styling(position = 'center')
```

Data	Accuracy	Sensitivity	Specificity
Train	0.6374745	0.6370757	0.6388889
Test	0.5853659	0.5454545	0.6857143

Conclusion