# Part II

Aarushi Verma (*Coordinator*) Deekshita Saikia (*Programmer*)
Mohammad Anas (*Writer*) Tego Chang (*Checker*)
Sydney Donati-Leach (*Presenter*)

## Summary

In this analysis, we explore the effectiveness of the National Support Work (NSW) Demonstration program on the wages of disadvantaged male workers. The focus is mainly to compare males who participated in the program versus the ones who did not participate and determine whether the participants are more likely to be employed in 1978.

We also explore demographic factors that are likely to increase the chances earning non-zero wages for these workers. Exploratory data analysis is carried out on the dataset, and a logistic regresison model is fit using stepwise selection. We observe that people who participated in the program are more likely to earn non-zero wages as compared to people who did not participate in the training program. The age of the male workers was a predictor of interest as its effect on the odds of earning non-zero wages was different for the males who participated as compared to the males who did not participate in training.

## Data

The dataset consists of observations for 614 male workers, with 11 variables. We used the $re78$ variable available in the dataset to create a binary variable $re78Bi\_F$, which indicates whether the person was earning non-zero wages in 1978. We use this factor variable as the response variable in our analysis. We ensure that the independent variables have the correct data type before proceeding with the analysis. The variables *treat*, *black*, *hispan*, *married* and *nodegree* were converted to factors, while *age* and *educ* were used as numeric variables in our model.

## Exploratory Data Analysis

We start our exploratory analysis by observing the relationships of our predictor variables against our response variable, $re78\_F$. Given that the *treat* predictor is the main variable of interest, we explore it first. We see that the probability for earning non-zero wages is approximately the same for everyone regardless of whether they participated in the training program, as can be observed in the table below.

Table 1: Conditional Probabilities

| | 0 | 1 |
|---|---|---|
| NonZero W | 0.77 | 0.76 |
| Zero W | 0.23 | 0.24 |

We also test the association of our response variable with the categorical predictors using Chi-squared test of association. We notice that black people are less likely to earn non-zero wages as compared to non-black people. The Chi-squared test shows us that the predictors *hispan*, *married* and *nodegree* have no association with the our response variable. We also generate box plots to explore the effects of our continuous predictors

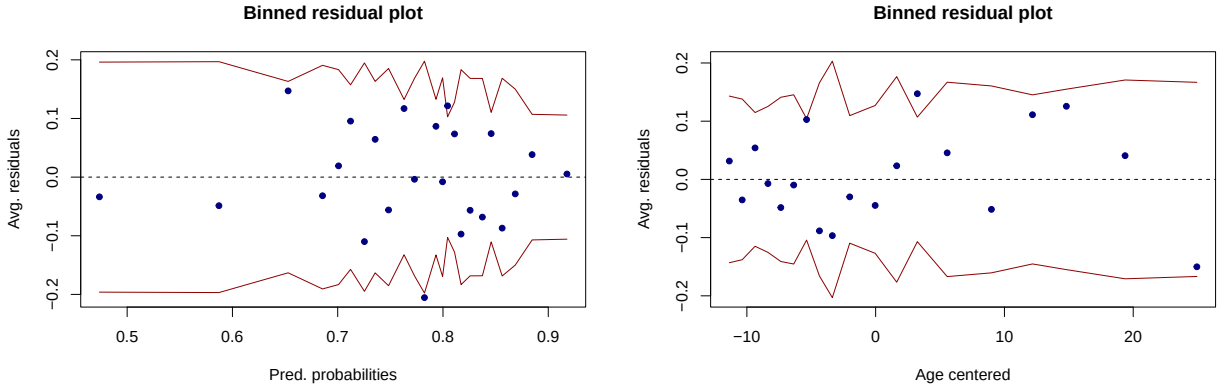
on our response variable. The box plots indicate that educated people are more likely to earn non-zero wages. We also note that younger people are more likely to earn non-zero wages.

We move on to explore whether the effect of the predictors on our response variable may be affected by other predictors. The box plot below indicate that for people who did not participate in the NSW training, younger people are more likely to earn non-zero wages, while the trend seems to be the opposite for the people who participated in the training program.


We also observed that for non-black people, the training was more effective. Therefore, we include the interaction for the *black* and *treat* predictors in our model. Based on the results from our exploratory analysis, we saw that the interaction affects of *age* and *educ*, *hispan* and *treat*, and *re*74 and *treat* might be worth investigating as well.

## Model Building

We start by fitting a model that includes only the main effects except the *nodegree* variable. We exclude this variable from our model as it captures very similar information to the *educ* variable. The results of the model seem counter-intuitive as our main variable of interest, *treat* is statistically insignificant. We also note that the real earnings of a person in 1974, *re*74, the centered age variable, $age_c$ and *black* predictors have a significant affect on the odds on earning a non-zero wage in 1978. The residual deviance of our model is 634.95 which suggests that model is a better fit than the null model. To assess the model further, we observe binned plots of the residuals against the fitted values and the continuous predictors. We check for randomness in these plots to ensure that our model satisfies the independence of errors assumption and to investigate whether any transformations of the continuous predictors are required. The binned plot for residuals against fitted values seem fairly random, except for a couple of bins on the left area of the plot. The binned plots against continuous predictors look random for *educ* and *re*74, indicating that we do not need any transformations for these two variables. However, for *age_c*, we see a polynomial trend that has not been captured by our model.




To improve our model's fit, we start adding interactions between the effects to our model. We adopt a stepwise variable selection approach combined with Chi-squared tests to see which predictors and interaction effects amongst the predictors improve the fit of our model. To do this we specify a base/null model, where we only include the variables of interest; the treatment variable *treat*, demographic variables (*black*, *hispan* and *age_c*) and the interaction of *treat* with all the demographic variables. On the other hand, our full model contains all the variables and interactions in the null model, the interactions that were found to be interesting in our exploratory analysis, and the *married* and *educ* effects. We then perform stepwise model selection, using AIC as well as BIC as the criteria for variable selection. Both iterations of the logistic regression models shortlist *treat*, *age_c*, *re*74 and the interaction between *treat* and *age_c* variables in the stepwise selection process. However, as AIC tends to be more lenient as compared to BIC, we observe that *black*, *hispan*, interaction between *treat* and *hispan*, and the interaction between *treat* and *black* variables are also selected in the model using AIC for variable selection. The results of the model containing predictor variables using AIC are shown below.

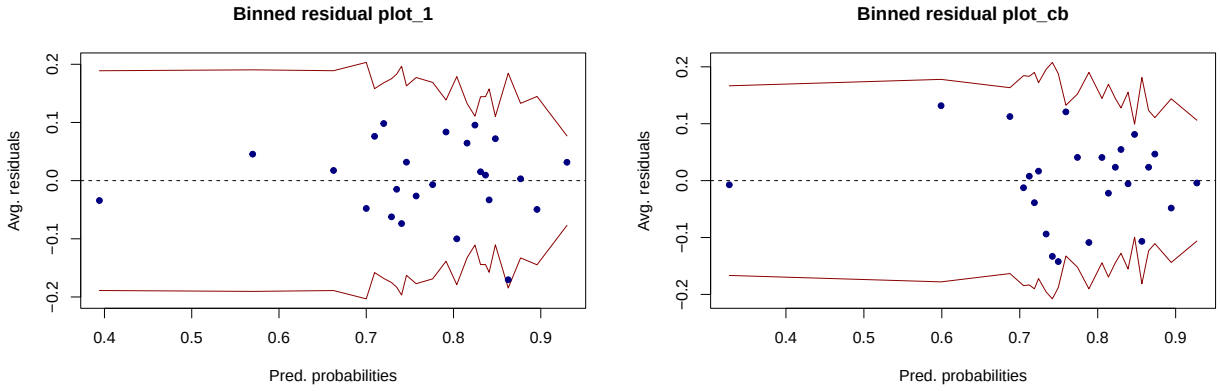
| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | 0.98 | 0.18 | 5.31 | 0.00 |
| treat1 | 0.52 | 0.29 | 1.78 | 0.08 |
| age_c | -0.06 | 0.01 | -4.95 | 0.00 |
| black1 | -0.48 | 0.27 | -1.79 | 0.07 |
| hispan1 | 0.05 | 0.36 | 0.13 | 0.90 |
| re74 | 0.00 | 0.00 | 3.89 | 0.00 |
| treat1:hispan1 | 15.09 | 722.55 | 0.02 | 0.98 |
| treat1:age_c | 0.08 | 0.03 | 2.86 | 0.00 |
| treat1:re74 | -0.00 | 0.00 | -1.68 | 0.09 |

We observe that not all the predictors in the AIC model are significant at the 5% significant level. To check which iteration of the model fits better, we conduct a Chi-squared test and compared the residual deviance of these two model iterations. The p-value of this test turns out to be 0.012, indicating that at least one of the additional variables selected in our AIC model is significantly improving the fit of the model. We see in the results of our AIC model that the variables *black* and the interaction between *treat* and *re*74 have low p-values. Therefore, we fit a new model which included these two effects and all the variables from our BIC model. We compare this model to the BIC model using a Chi-squared test. The difference was statistically significant and we decided to keep these effects in our model. To check if the remaining effects in the AIC model (*hispan*, and its interaction with *treat* variable) would improve the fit of our model, we conduct another Chi-squared to compare the residual deviance between our new model and the AIC model. Based on the results of this test, which yields an insignificant p-value, we conclude that these effects do not lead to a statistically better fit.

The results of our model suggest that all the variables are significant at the 5% significance level except the interaction of *treat* and *re*74 variable. We also note that the residual deviance of this model falls to 626.03

proving that this is a better fit as compared to our baseline model that included only the main effects. However, when we look at the binned plot of residuals against our continuous predictors and predicted probabilities, we see that the polynomial trend in the *age_c* variable has still not been captured.

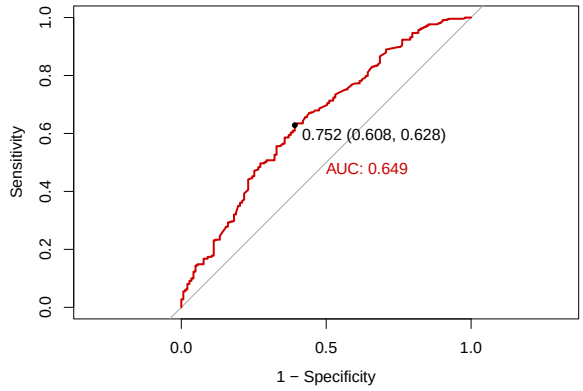
We observe that there is still a trend in the residuals plot of the *age_c* variable, which prompts us to investigate polynomial interactions of this effect in the model. We explore squared and cubic degree polynomials of *age_c* in our model. We also included the interactions of the *treat* variable with these transformations of *age_c*. We run the regression again with the additional variables to check if this improves our model fit. However, the residual deviance of this model turned out to be 620.32 which is not a significant improvement over our previous model, which is also confirmed by a Chi-squared test. We compared the binned residual plots of these models against the fitted values. These plots are shown below.




We see that there is not much difference in these plots and the bins on the far left of the plot are still present. Given that adding the squared and cubic terms of *age_c* in our model makes it harder to interpret the standard estimates and does not significantly improve the fit of our model, we decided not to include them in our model. Now that we have investigated the effects of including interactions and transformations to our model, our final model equation is given below.

$$log(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 treat_{i1} + \beta_2 age\_c_{i2} + \beta_3 black_{i3} + \beta_4 re74_{i4} + \beta_5 treat * age\_c_{i5} + \beta_6 treat * re74_{i6} \quad (1)$$

We now assess model performance by observing the RoC curve, as shown below.




We then leverage the confusion matrix to calculate the accuracy, sensitivity and specificity of our model. Classifying the outcomes using the probability threshold of 0.5 allows us to achieve an accuracy of 78% and

sensitivity of 98%. However, we see that our model does a poor job at predicting people who earned zero wages in 1978 (as can be seen from the low specificity of 11%). Therefore, to obtain a balance between specificity and sensitivity, we classify outcomes based on the probability threshold suggested by the ROC curve, which is 0.752. Using this threshold, we are able to achieve sensitivity rate of 63% and the specificity rate improves to 60%.

The standard estimates of all the coefficients in our model were statistically significant except the interaction between the *treat* and *re*74 predictors. We exponentiate the standard estimates and their confidence intervals to interpret them on the odds scale. The model results are as shown below.

| | coeffecients_2.5 | coeffecients | coeffecients_97.5 | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|---|---|
| (Intercept) | 1.99 | 2.77 | 3.87 | 0.17 | 5.99 | 0.00 |
| age_c | 0.92 | 0.95 | 0.97 | 0.01 | -4.98 | 0.00 |
| re74 | 1.00 | 1.00 | 1.00 | 0.00 | 3.84 | 0.00 |
| treat1 | 1.08 | 1.89 | 3.31 | 0.29 | 2.24 | 0.03 |
| black1 | 0.34 | 0.55 | 0.90 | 0.25 | -2.40 | 0.02 |
| age_c:treat1 | 1.02 | 1.08 | 1.14 | 0.03 | 2.77 | 0.01 |
| re74:treat1 | 1.00 | 1.00 | 1.00 | 0.00 | -1.60 | 0.11 |

Despite having interactions, multicollinearity is not present in our model as the variance inflation factor of each of the variables is well within range (<10). A table indicating the VIFs of all the variables in our model is provided below.

| | x |
|---|---|
| age_c | 1.41 |
| re74 | 1.58 |
| treat1 | 1.83 |
| black1 | 1.59 |
| age_c:treat1 | 1.28 |
| re74:treat1 | 1.51 |

## Conclusion

To sum up our analysis, the people who participated in the program are more likely to earn non-zero wages in 1978 as compared to people who did not participate. Statistically speaking, the odds of earning positive wages after receiving training is 1.89 times the odds of a person who did not receive training for a 27 year old male who was unemployed in 1974. Age was found to an interesting variable in our analysis. For a person who did not receive training, a 1 year increase in age decreases the odds of earning positive wages by 5.5%. The association between *age_c* and *treat* was also an interesting association in the model. For people who participated in the training, 1 year increase in age leads to an increase in the odds of earning non-zero wages by 1.9%. We get this value by multiplying the standard estimates (odds scale) of *age_c* and the estimates of the interaction between *age_c* and *treat*. Non-b;ack people are also more likely to earn non-zero wages in 1978.

## Limitations

- Our model assumes that the length of the NSW training remained the same for all participants. However, in reality, participants joined the training program between March 1975 and July 1977, and the randomization over this 2-year period led to people with different characteristics joining the program.
- There is an uneven distribution in our response variable. The proportion of people earning non-zero wages is very high compared to the proportion of unemployed people. A potential solution to this problem can be under sampling the data of males earning non-zero wages.

## APPENDIX

### R Script

```
# Reading in libraries
library(xtable)
library(ggplot2)
library(rms)
library(arm)
library(e1071)
library(caret)
library(pROC)
library(stargazer)
# library(kable)

# Reading in dataset
setwd("C:\..\\..\\..\\..\\
      ..\\..\\..")
nsw <- read.csv("lalondedata.txt")

# Checking observations and dataset summary
head(nsw)
dim(nsw)
summary(nsw)
str(nsw)

nsw <- read.csv("lalondedata.txt",header=T,
                colClasses=c("factor","factor","numeric","numeric",
                             "factor", "factor", "factor", "factor",
                             "numeric", "numeric", "numeric"))

# Creating binary response
nsw$re78Bi <- 0
nsw$re78Bi[nsw$re78 > 0] <- 1

# Creating factor variable with labels
nsw$re78Bi_F <- factor(nsw$re78Bi,
                               levels=c(0,1),labels=c("Zero","Non-Zero"))

str(nsw)

### EDA
## boxplots for the numeric variables

# age vs re78Bi_F
ggplot(nsw,aes(x=re78Bi_F, y=age, fill=re78Bi_F)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Greens") +
  labs(title="Age vs Wage",
       x="Wage?",y="Age") +
  theme_classic() + theme(legend.position="none")

# educ vs re78Bi_F
```

```r
ggplot(nsw,aes(x=re78Bi_F, y=educ, fill=re78Bi_F)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Greens") +
  labs(title="Educ vs Wage",
       x="Wage?",y="Educ") +
  theme_classic() + theme(legend.position="none")
# Educ indeed has an effect

# re74 vs re78Bi_F
ggplot(nsw,aes(x=re78Bi_F, y=re74, fill=re78Bi_F)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Greens") +
  labs(title="Previous Wage vs Wage",
       x="Wage?",y="Previous Wage") +
  theme_classic() + theme(legend.position="none")
# Previous wage indeed has an effect

### EDA
## tables for the factor variables

# treat vs re78Bi_F
table(nsw[,c("re78Bi_F","treat")])
# table(nsw[,c("re78Bi_F","treat")])/sum(table(nsw[,c("re78Bi_F","treat")]))
apply(table(nsw[,c("re78Bi_F","treat")])/
        sum(table(nsw[,c("re78Bi_F","treat")])),
      2,function(x) x/sum(x))
tapply(nsw$re78Bi_F, nsw$treat, function(x) table(x)/sum(table(x)))
chisq.test(table(nsw[,c("re78Bi_F","treat")]))
# surprisingly have no effect!

# black vs re78Bi_F
table(nsw[,c("re78Bi_F","black")])
apply(table(nsw[,c("re78Bi_F","black")])/
        sum(table(nsw[,c("re78Bi_F","black")])),
      2,function(x) x/sum(x))
tapply(nsw$re78Bi_F, nsw$black, function(x) table(x)/sum(table(x)))
chisq.test(table(nsw[,c("re78Bi_F","black")]))
# black indeed has an effect

# hispan vs re78Bi_F
table(nsw[,c("re78Bi_F","hispan")])
apply(table(nsw[,c("re78Bi_F","hispan")])/
        sum(table(nsw[,c("re78Bi_F","hispan")])),
      2,function(x) x/sum(x))
tapply(nsw$re78Bi_F, nsw$hispan, function(x) table(x)/sum(table(x)))
chisq.test(table(nsw[,c("re78Bi_F","hispan")]))

# married vs re78Bi_F
table(nsw[,c("re78Bi_F","married")])
apply(table(nsw[,c("re78Bi_F","married")])/
        sum(table(nsw[,c("re78Bi_F","married")])),
      2,function(x) x/sum(x))
tapply(nsw$re78Bi_F, nsw$married, function(x) table(x)/sum(table(x)))
```

```r
chisq.test(table(nsw[,c("re78Bi_F","married")]))

# nodegree vs re78Bi_F
table(nsw[,c("re78Bi_F","nodegree")])
apply(table(nsw[,c("re78Bi_F","nodegree")])/
        sum(table(nsw[,c("re78Bi_F","nodegree")])),
      2,function(x) x/sum(x))
tapply(nsw$re78Bi_F, nsw$nodegree, function(x) table(x)/sum(table(x)))
chisq.test(table(nsw[,c("re78Bi_F","nodegree")]))

## binnedplots of continuous predictors versus re78Bi

par(mfrow=c(1,1))
# age vs re78Bi_F
binnedplot(y=nsw$re78Bi,nsw$age,xlab="Age",ylim=c(0,1),col.pts="navy",
           ylab ="Non-zero Wage?",main="Binned Age and Non-zero Wage cases",
           col.int="white")
# seems no obvious trend, slightly went down

# educ vs re78Bi_F
binnedplot(y=nsw$re78Bi,nsw$educ,xlab="Educ",ylim=c(0,1),col.pts="navy",
           ylab ="Non-zero Wage?",main="Binned Educ and Non-zero Wage cases",
           col.int="white")
# roughly linearly incase trend as expected

# re74 vs re78Bi_F
binnedplot(y=nsw$re78Bi,nsw$re74,xlab="Previous Wage",
           ylim=c(0,1),col.pts="navy",
           ylab ="Non-zero Wage?",
           main="Binned Previous Wage and Non-zero Wage cases",
           col.int="white")
# roughly linearly incase trend as expected

## short brief of observations:
# shall include in the base model: educ, re74, treat (as question wants), black

## interaction
# age vs re78Bi_F by educ
table(nsw$age)
summary(nsw)
ggplot(nsw,aes(x=re78Bi_F, y=age, fill=re78Bi_F)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Age vs Non-zero Wages, by Educ",
       x="Non-zero Wages?",y="Age") +
  theme_classic() + theme(legend.position="none") +
  #scale_x_discrete(labels=c("0" = "No","1" = "Yes")) +
  facet_wrap( ~ educ)
# shall investigate the interaction

# re74 vs re78Bi_F by treat
ggplot(nsw,aes(x=re78Bi_F, y=re74, fill=re78Bi_F)) +
  geom_boxplot() + #coord_flip() +
```

```
  scale_fill_brewer(palette="Reds") +
  labs(title="Previous Wage vs Non-zero Wages, by treat",
       x="Non-zero Wages?",y="Previous Wage") +
  theme_classic() + theme(legend.position="none") +
  #scale_x_discrete(labels=c("0" = "No","1" = "Yes")) +
  facet_wrap( ~ treat)
# this can be included in the full model (try stay away from log to avoid loss
# observation)

# age vs re78Bi_F by treat
ggplot(nsw,aes(x=re78Bi_F, y=age, fill=re78Bi_F)) +
  geom_boxplot() + #coord_flip() +
  scale_fill_brewer(palette="Reds") +
  labs(title="Age vs Wage by Treat (training)",
       x="Wage?",y="Age") +
  theme_classic() + theme(legend.position="none") +
  facet_wrap(~treat)
# shall investigate the interaction

# educ vs re78Bi by black
par(mfcol=c(2,1))
table(nsw$educ)
#first plot for black = 0
binnedplot(nsw$educ[nsw$black=="0"], y=nsw$re78Bi[nsw$black=="0"],
           xlab = "Educ", ylab = "Non-zero Wage",
           main = "Binned Educ and Non-zero Wage cases (Black = 0)")

#next the plot for black = 1
binnedplot(nsw$educ[nsw$black=="1"], y=nsw$re78Bi[nsw$black=="1"],
           xlab = "Educ", ylab = "Non-zero Wage",
           main = "Binned Educ and Non-zero Wage cases (Black = 1)")
# shall investigate the interaction as linear increase trend disappear
# when black = 1

# educ_F vs re78Bi_F by black
nsw_black0 <- nsw[nsw$black=="0",]
nsw_black1 <- nsw[nsw$black=="1",]
table(nsw$black,nsw$educ)
apply(table(nsw_black0[,c("re78Bi_F", "educ")])/
        sum(table(nsw_black0[,c("re78Bi_F", "educ")])),
      2,function(x) x/sum(x))

apply(table(nsw_black1[,c("re78Bi_F", "educ")])/
        sum(table(nsw_black1[,c("re78Bi_F", "educ")])),
      2,function(x) x/sum(x))
# shall investigate the interaction

# educ_F vs re78Bi_F by hispan
nsw_hispan0 <- nsw[nsw$hispan=="0",]
nsw_hispan1 <- nsw[nsw$hispan=="1",]
table(nsw$hispan,nsw$educ)
apply(table(nsw_hispan0[,c("re78Bi_F", "educ")])/
        sum(table(nsw_hispan0[,c("re78Bi_F", "educ")])),
```

```
      2,function(x) x/sum(x))

apply(table(nsw_hispan1[,c("re78Bi_F", "educ")])/
        sum(table(nsw_hispan1[,c("re78Bi_F", "educ")])),
      2,function(x) x/sum(x))
# given that low observation in elementary and after high for hispan, we shall
# NOT investigate the interaction

# black vs re78Bi_F by treat
nsw_treat0 <- nsw[nsw$treat=="0",]
nsw_treat1 <- nsw[nsw$treat=="1",]
table(nsw$black,nsw$treat)
apply(table(nsw_treat0[,c("re78Bi_F", "black")])/
        sum(table(nsw_treat0[,c("re78Bi_F", "black")])),
      2,function(x) x/sum(x))

apply(table(nsw_treat1[,c("re78Bi_F", "black")])/
        sum(table(nsw_treat1[,c("re78Bi_F", "black")])),
      2,function(x) x/sum(x))
# shall investigate the interaction

# treat vs re78Bi_F by black
nsw_black0 <- nsw[nsw$black=="0",]
nsw_black1 <- nsw[nsw$black=="1",]
table(nsw$black,nsw$treat)
apply(table(nsw_black0[,c("re78Bi_F", "treat")])/
        sum(table(nsw_black0[,c("re78Bi_F", "treat")])),
      2,function(x) x/sum(x))

apply(table(nsw_black1[,c("re78Bi_F", "treat")])/
        sum(table(nsw_black1[,c("re78Bi_F", "treat")])),
      2,function(x) x/sum(x))

# treat vs re78Bi_F by hispan
nsw_hispan0 <- nsw[nsw$hispan=="0",]
nsw_hispan1 <- nsw[nsw$hispan=="1",]
table(nsw$hispan,nsw$treat)
apply(table(nsw_hispan0[,c("re78Bi_F", "treat")])/
        sum(table(nsw_hispan0[,c("re78Bi_F", "treat")])),
      2,function(x) x/sum(x))

apply(table(nsw_hispan1[,c("re78Bi_F", "treat")])/
        sum(table(nsw_hispan1[,c("re78Bi_F", "treat")])),
      2,function(x) x/sum(x))
# shall investigate the interaction

# brief: interactions shall be considered:
# treat: hispan, treat:black, educ:black, re74:treat, age:educ, age:treat

nsw$age_c <- nsw$age - mean(nsw$age)

#Model 2 - Main effects (Numeric Education)
base_model <- glm(re78Bi_F ~ educ + black + hispan + treat + re74 + age_c +
```

```
                married, data = nsw, family = binomial) # FINAL BASE MODEL

summary(base_model)
# Model 1 - considers all the main effects
# (excluding No degree)and education as numeric)
# At 95% significance Black, Re74 and Age_c are significant
# AIC = 650.95

#Model Assessment

#save the raw residuals
rawresid1 <- residuals(base_model,"resp")

#binned residual plots - model
binnedplot(x=fitted(base_model),y=rawresid1,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")
#looks good

# For significant coeffs
#binned residual plots - centered Age
binnedplot(x=nsw$age_c,y=rawresid1,xlab="Age centered",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")
#no trend 2 points outside

#binned residual plots - Re74
binnedplot(x=nsw$re74,y=rawresid1,xlab="Re74",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")
#not as much of a trend 2 points outside

#binned residual plots - Educ
binnedplot(x=nsw$educ,y=rawresid1,xlab="Educ",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")


n <- nrow(nsw)

null_model <- glm(re78Bi_F~ treat + age_c +
                    black + hispan + treat:black + treat:hispan +
                    treat:age_c,data=nsw,family=binomial)

full_model <- glm(re78Bi_F ~ treat*black  + treat*hispan +
                    re74*treat + educ*black + age_c*treat
                  + age_c*educ + married, data = nsw, family = binomial)

AIC_stepwise <- step(null_model,
                     scope = formula(full_model),direction="both",trace=0)
summary(AIC_stepwise)
# Model call - re78Bi_F ~ treat + age_c + black + hispan + re74 + treat:hispan +
#   treat:age_c + treat:re74
```

```r
BIC_stepwise <- step(null_model,scope= formula(full_model),direction="both",
     trace=0,k = log(n))
summary(BIC_stepwise)
# Model call - re78Bi_F ~ treat + age_c + re74 + treat:age_c

# ~~~~~~~~ Work in progress ~~~~~~~~~~~~~
# Comparing the AIC and BIC models by an ANOVA test
anova(BIC_stepwise, AIC_stepwise, test = "Chisq")

# ~~~~~~~~~~~~~~~~~~~~AIC_Stepwise is our final model~~~~~~~~~~~~~~~~~~~~~~~~

# Compare models

# Comparing black and treat:re74
compare1 <- glm(re78Bi_F ~ treat + age_c  + re74 +black
                + treat:age_c + treat:re74, data = nsw, family = binomial)
anova(BIC_stepwise,compare1, test="Chisq")

anova(compare1, AIC_stepwise, test="Chisq")


Final_Model <- glm(formula = re78Bi_F ~ treat + age_c  + re74 + black
                + treat:age_c + treat:re74, family = binomial,
                   data = nsw)

rawresid4 <- residuals(Final_Model,"resp")

par(mfcol=c(1,1))

binnedplot(x=fitted(Final_Model),y=rawresid4,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot_1",col.pts="navy")

# Real Annual Earnings in 1974
binnedplot(x=nsw$re74,y=rawresid4,xlab="re74 Earnings",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")

# Centered Age
binnedplot(x=nsw$age_c,y=rawresid4,xlab="Age centered",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")

#Transformations with the age predictor

nsw$age_c_sq <- (nsw$age_c)^2
nsw$age_c_cb <- (nsw$age_c)^3

Final_Model_1_sq <- glm(formula = re78Bi_F ~ treat + age_c  + age_c_sq +
                           re74 + black
                + treat:age_c + treat:age_c_sq + treat:re74,
                family = binomial, data = nsw)
summary(Final_Model_1_sq)
```

```
Final_Model_1_cb <- glm(formula = re78Bi_F ~ treat + age_c  + age_c_sq +
                          age_c_cb + re74 + black
                + treat:age_c + treat:age_c_sq + treat:age_c_cb +
                  treat:re74, family = binomial, data = nsw)
summary(Final_Model_1_cb)

rawresid5 <- residuals(Final_Model_1_sq,"resp")

binnedplot(x=fitted(Final_Model_1_sq),y=rawresid5,xlab="Pred. probabilities",
          col.int="red4",ylab="Avg. residuals",
          main="Binned residual plot_sq",col.pts="navy")

binnedplot(x=nsw$age_c,y=rawresid5,xlab="Age",
          col.int="red4",ylab="Avg. residuals",
          main="Binned residual plot",col.pts="navy")
binnedplot(x=nsw$age_c_sq,y=rawresid5,xlab="Age squared",
          col.int="red4",ylab="Avg. residuals",
          main="Binned residual plot",col.pts="navy")

rawresid6 <- residuals(Final_Model_1_cb,"resp")
binnedplot(x=fitted(Final_Model_1_cb),y=rawresid6,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot_cb",col.pts="navy")
binnedplot(x=nsw$age_c,y=rawresid6,xlab="Age",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")
binnedplot(x=nsw$age_c_sq,y=rawresid6,xlab="Age squared",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")
binnedplot(x=nsw$age_c_cb,y=rawresid6,xlab="Age Cubed",
           col.int="red4",ylab="Avg. residuals",
           main="Binned residual plot",col.pts="navy")


Final_Model <- glm(formula = re78Bi_F ~ age_c + re74 + treat  + black
                + treat:age_c + treat:re74, family = binomial,
                   data = nsw)

summary(Final_Model)

exp(Final_Model$coefficients)
exp(confint.default(Final_Model))

# confusion matrix
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(Final_Model) >= 0.5,
                                             "1","0")),
                            as.factor(nsw$re78Bi),positive = "1")
Conf_mat$table
Conf_mat$overall["Accuracy"];
Conf_mat$byClass[c("Sensitivity","Specificity")]
#True positive rate and True negative rate

table(nsw$re78Bi_F)
```

```r
#ROC curve...
roc(nsw$re78Bi,fitted(Final_Model),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3")

#let's repeat with the marginal percentage in the data
Conf_mat_FM1 <- confusionMatrix(as.factor(ifelse(fitted(Final_Model) >= 0.752,
                                                 "1","0")),
                                as.factor(nsw$re78Bi),positive = "1")
Conf_mat_FM1$table
Conf_mat_FM1$overall["Accuracy"];
Conf_mat_FM1$byClass[c("Sensitivity","Specificity")]

tb2 <- data.frame(
  row_title = c("Threshold = 0.5", "Threshold = 0.752 (Best)"),
  col1 = c(Conf_mat$overall[1], Conf_mat_FM1$overall[1]),
  col2 = c(Conf_mat$byClass[1], Conf_mat_FM1$byClass[1]),
  col3 = c(Conf_mat$byClass[2], Conf_mat_FM1$byClass[2]),
  col4 = c(0.65, 0.65)
  )
colnames(tb2) <- c("", "Accuracy", "Sensitivity", "Specificity", "AUC")
xtable(tb2, digits = 2)

# Checking multicollinearity

library(rms)
vif(Final_Model)

# All variables well within range
```