Team Orange Project 1 Part 1

9/20/2020

Checker/Coordinator: Shangwen Yan; Presenter: Yi Feng
Programmer: Michael Tang; Writer: Jennie Sun

1. Summary

This analysis aims to find out whether or not job training for disadvantaged workers had an effect on their wages using a subset of the data from the National Supported Work (NSW) Demonstration containing only male participants from the 1970s. The analysis applies methods including but not limited to: preliminary screening using exploratory data analysis (EDA), model fitting using linear regression with multiple predictors, model selection using BIC and ANOVA test, and model validation using assumptions assessment, Cook's distance, and multicollinearity (VIF). Our final model shows that there is evidence suggesting that workers who receive job training tend to earn higher wages than workers who do not receive job training.

2. Introduction

This analysis discovers if there is any evidence that workers who receive job training(treat) tend to earn higher wages than workers who do not receive job training. It also quantifies the effect of the job training by providing an estimated range of real annual earnings. In addition, it provides evidence that shows if this effect differs by demographic groups, and points out other interesting associations with wages. Specifically, when a worker is trained, the older this workers gets, the more he earns in 1978 compared to 1974. There is also an association between wage difference and marital status, which seems to be the opposite trend compared to the current world.

3. Data

In order to demonstrate the effect of job training(treat), this analysis uses the difference between real annual earnings in 1978 and 1974 as the response variable (response variable = re78 - re74). It also reassembles years of education to 4 groups - elementary(0-5), middle(6-8), high(9-12), and college(13-18) - as a way to interpret the data better and to reduce the effect of erroneous/missing values and unbalanced data distribution based on education.

3.1 EDA

To start, we assess the normality of response variable by plotting a histogram showing the frequency distribution of wage difference between 1978 and 1974, which has a somewhat normal distribution.
First, we look at the EDAs for categorical variables:

- The boxplot of wage difference by treat shows a difference in the medium values and distribution between trained and untrained workers, suggesting that `treat` may be a significant variable to include when fitting the model.

- The boxplot of wage difference by education shows a difference in the medium values and distribution across different education levels, suggesting that `educ` may be a significant variable to include when fitting the model.

- The boxplot of wage difference by black shows a difference in the medium values and distribution between African Americans and other races, suggesting that `black` may be a significant variable to include when fitting the model.

- The boxplot of wage difference by hispanic shows a difference in the medium values and distribution between Hispanic ethnicity and other races, suggesting that `hispan` may be a significant variable to include when fitting the model.

- The boxplot of wage difference by marries shows a difference in the medium values and distribution between married and un-married, suggesting that `married` may be a significant variable to include when fitting the model.

- The boxplot of wage difference by nodegree shows a difference in the medium values and distribution between those who dropped out of high school and those who did not, suggesting that `nodegree` may be a significant variable to include when fitting the model.

Next, we look at the EDA for the continuous variable:

- The raw points on the scatterplots of wage difference by age shows a downward trend as age increases. When switched to the average counts, the downward trend becomes more prominent, suggesting that `age` may be a significant variable to include when fitting the model.

Through visual inspection, all predictor variables look significant.

3.2 Explore Interaction Terms

Because `nodegree` and `educ` contain similar information, there is no need to explore interactions between these two variables. Since we are interested in evidence that demonstrate a relationship between treat and wage difference, for now we only explore the potential interactions among wage difference, treat, and another predictors.
Same with EDA, we look at the interaction effect for categorical variables first:

- The boxplot of wage difference and treat by education shows a similar trend in the medium values and distribution across different education levels in both trained and untrained groups, suggesting that `treat`*`educ` may not be a significant interaction term to include when fitting the model.

- The boxplot of wage difference and treat by black shows a similar trend in the medium values and distribution between African Americans and other races in both trained and untrained groups, suggesting that `treat`*`black` may not be a significant interaction term to include when fitting the model.

- The boxplot of wage difference and treat by hispanic shows a similar trend in the medium values and distribution between Hispanic ethnicity and other races in both trained and untrained groups, suggesting that `treat`*`hispan` may not be a significant interaction term to include when fitting the model.

- The boxplot of wage difference and treat by marital status shows a similar trend in the medium values and distribution between married and un-married in both trained and untrained groups, suggesting that `treat`*`married` may not be a significant interaction term to include when fitting the model.

- The boxplot of wage difference and treat by nodegree shows a similar trend in the medium values and distribution between those who dropped out of high school and those who did not in both trained and untrained groups, suggesting that `treat*nodegree` may not be a significant interaction term to include when fitting the model.

Next, we look at the interaction effect for the continuous variable:

- The raw points on the scatterplots of wage difference and treat by age shows a downward trend in the untrained group but and upward trend in the trained group. When switched to the average wage difference and treat by age, the difference in trend becomes more prominent, suggesting that `treat*age` may be a significant interaction term to include when fitting the model.

Through visual inspection on the potential interaction effect, only the interaction between age and treat could be significant. EDA only depicts the visual representation of the data, the actual significance of each variable needs to be assessed by statistical tests. Specifically, if a predictor has been deemed significant through EDA but is dropped through model selection, we will perform ANOVA test to determine its significance.

4. Model

4.1 Model Validation

We center the age predictor variable so that it's most meaningful to interpret the intercept. First, we build a naive model with all main effects. From the summary output, treat1, age(centered), education(middle school), and married1 are significant predictor variables. Next, we assess the assumptions of linear regression.

- For normality, most points on the Normal Q-Q plot seem to fall on the 45 degree angle line, with some deviations at both ends, which could be caused by outliers in the dataset that need further investigations. Therefore, the normality assumption holds.
- In terms of independence and equal variance, most points on the residuals vs fitted values plot seem to be randomly distributed, so the independence assumption holds. The spread of points above and below the 0 residuals looks equal, and the red line tends to stay flat across fitted values. Point 132 seems to be far away from the rest of the data, which we will investigate later. Overall, the equal variance assumption holds.
- For linearity, we only assess residuals vs age(centered) as age is the only discrete/non-categorical predictor variable. Most points on the residuals vs age(centered) plot seem to be randomly and equally distributed above and below the 0.0 residuals line. Since no obvious pattern is identified, the linearity assumption holds.

Therefore, all linear regression assumptions are held. For the naive model, the model summary shows that treat, age_c, educ, married are statistically significant predictor variables. Even though other predictor variables don't show any statistical significance, it could be that the linear relationship between each one of these variable is not strong enough to be detected by this sample.

4.2 Model Selection

In this section, we perform model selection, with null_model only capturing the predictor that we or our client care about - treatment. And the full_model will include all main effects as well as the interactions between treat and other variables. If any variables that were previously determined to be statistically significant are removed through the model selection process, we will perform F-test on these terms to evaluate their significance. Because the data is not health related, we use BIC as our selection criterion as it is more strict at selecting variables and False Negative (FN) and False Positive (FP) don't matter much in this case. The results from all three model selection criteria - forward, backward, and stepwise - returned

the same model, which all use treat, age_c and treat:age_c as the predictors.

Based on these predictors, we create a new model(model_1), and evaluate the significance of the predictors that we previously deemed significant through visual inspection using ANOVA test.

- To start, we create a different model that includes educ(model_2) on top of the new model(model_1). The ANOVA test between model_1 and model_2 has a p-value of 0.05965. Since it is only slightly higher than 0.05 but below 0.1, it might still be worth keeping.
- Next, we create a different model that includes black(model_3) on top of the new model(model_1). The ANOVA test between model_1 and model_3 has a p-value of 0.5834. Since it is much higher than 0.05, we drop this predictor variable.
- Next, we create a different model that includes hispan(model_4) on top of the new model(model_1). The ANOVA test between model_1 and model_4 has a p-value of 0.4182. Since it is much higher than 0.05, we drop this predictor variable.
- Next, we create a different model that includes married(model_5) on top of the new model(model_1). The ANOVA test between model_1 and model_5 has a p-value of 0.0144. Since it is below 0.05, we keep this predictor variable.
- Next, we create a different model that includes nodegree(model_6) on top of the new model(model_1). The ANOVA test between model_1 and model_6 has a p-value of 0.7417 Since it is much higher than 0.05, we drop this predictor variable.
- From above, because only education and married have significant p-values, we create a different model that includes both educ and married(model_7) on top of the new model(model_1). The ANOVA test between model_1 and model_7 has a p-value of 0.01328. Since it is below 0.05, we keep both predictor variables. Therefore, the final model will include treat, age_c, educ, married and treat:age_c

4.3 Final Model

4.3.1 Model Assessment

The normality assumption still holds because most points on the Normal Q-Q plot still seem to fall on the 45 degree angle line. The independence and equal variance assumption still hold as the points do not seem follow an obvious pattern and the spread looks equal, except for point 132. The linearity assumption still holds as there is no obvious pattern either on the residuals vs age(centered) plot. In terms of potential outliers, all points are within the 0.5 cook's distance range and are below 0.1 leverage score on the standardized residuals vs leverage plot. Point 132 is a potential outlier beyond 6 standard residuals but with low cook's distance and leverage score. We take a look at Point 132. According to our final model output, education (high school) and age have negative coefficients. However, Point 132 also only has a high school degree and is one year older than the average age. Therefore, this outlier doesn't align with our model. This person may have extraordinary life experience, so his attributes cannot be captured by the model. Therefore, since it has low cook's distance and low leverage score, it has low impact on the model performance so is not excluded. In the end, we check for multicolinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. The VIF values of all the predictor variables are below 5, which suggests that there should be no multicolinearity issues.

4.3.2 Model Interpretation

Final Model Equation:

$\hat{wage\_diff} = 2472.10\hat{treat}1 - 146.85\hat{age_c} - 3203.14\hat{educmiddle} - 2437.88\hat{educhigh} - 832.38\hat{educcollege} - 1638.78\hat{married}1 + 235.07\hat{treat1:age_c} + 4530.30$

Except for education(college), all other predictor variables are either significant at p-value<0.05 significant level - treat1, age_c, educ(middle school), married, and treat1:age - or significant at p-value<0.10 significant level - educ(high school). Below are the interpretations of those significant predictor variables:

- treat1: Compared to treat0(non-trained workers), the average real annual earnings for treat1(trained workers) is expected to increase by $2472.10 from 1974 to 1978, holding other variables constant.
- age_c: With an additional year increase in worker's age, the average real annual earnings is expected to decrease by $146.85 from 1974 to 1978, holding other variables constant.
- educ(middle school): Compared to workers with up to 5 years of education (elementary school), the average real annual earnings for workers with 6-8 years of education (middle school) is expected to decrease by $3203.14 from 1974 to 1978, holding other variables constant.
- educ(high school): Compared to workers with up to 5 years of education (elementary school), the average real annual earnings for workers with 9-12 years of education (high school) is expected to decrease by $2437.88 from 1974 to 1978, holding other variables constant.
- married1: Compared to workers who are not married, the average real annual earnings for married workers is expected to decrease by $1638.78 from 1974 to 1978, holding other variables constant.
- treat1:age_c: With every year increase in age given the treatment, the average annual earnings is expected to increase by $235.07 from 1974 to 1978, holding other variables constant.
- intercept: For someone who is not trained, at an average age, with an elementary education level, and not married, we expect to see an average wage difference of $4530.30 between 1974 and 1978.

5. CONCLUSION

To conclude from the final model, there is evidence suggesting that workers who receive job training tend to earn higher wages than workers who do not receive job training. This is supported by the positive intercept of treat1 in the final model. As we interpreted above, compared to treat0(non-trained workers), the average real annual earnings for treat1(trained workers) is expected to increase by $2472.10 from 1974 to 1978, holding other variables constant.

According to the 95% confidence interval of treat1 from the final model, a likely range of the change in real annual earnings from 1974 to 1978 for treat1(trained workers) is between $1043.74 and $3900.46, with all other variables held constant.

Additionally, there is evidence to suggest that the effects of job training differ by demographic groups. This can be seen from the effect of the interaction term treat1:age_c. When controlling treatment (when a worker is trained), the older one gets, the more one will earn in 1978 compared to 1974.

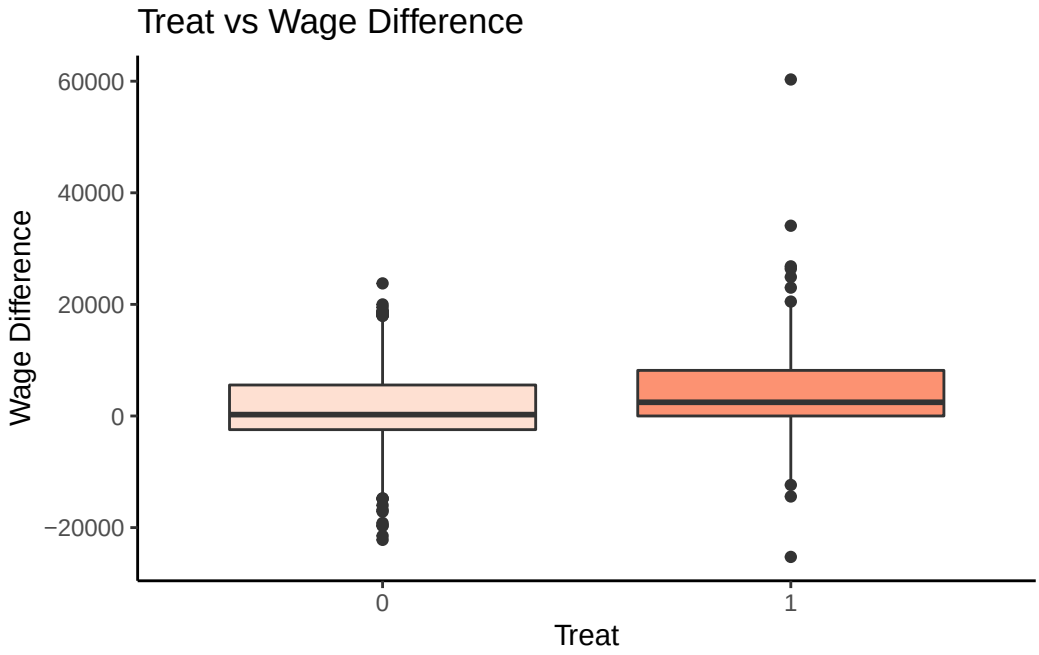
An interesting association with wages is the association between wage difference and marital status. Looking at the summary output of the final model, compared to workers who are not married, the average real annual earnings for married workers is expected to decrease by $1638.78 from 1974 to 1978, holding other variables constant. This seems to be a different trend compared to the situation in the current word, where married men tend to sit on the top of the wage ladder.

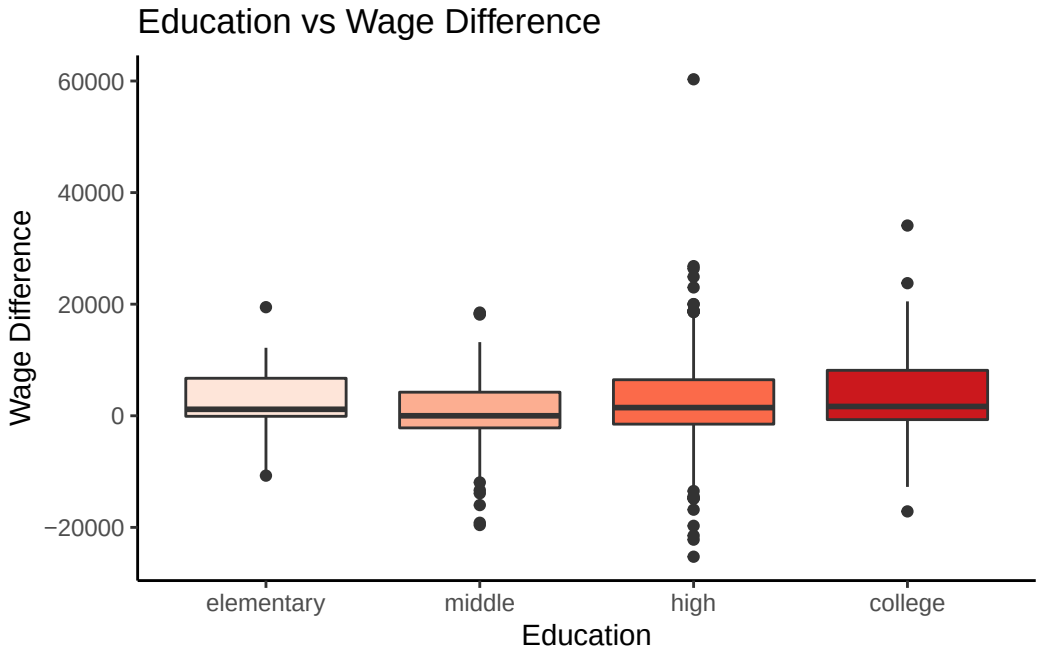
Another interesting finding associated with wages can be seen from the effect of the interaction term treat1:age_c. Without considering treatment, age_c is expected to have a negative effect on wage difference between 1978 and 1974, meaning that the older one gets, the less one will earn from 1974 to 1978. However, when controlling treatment (when a worker is trained), the older one gets, the more one will earn in 1978 compared to 1974. This may suggest that job training could be even more meaningful for people who are older. This also affirms that job training may potentially have a positive impact on wages.

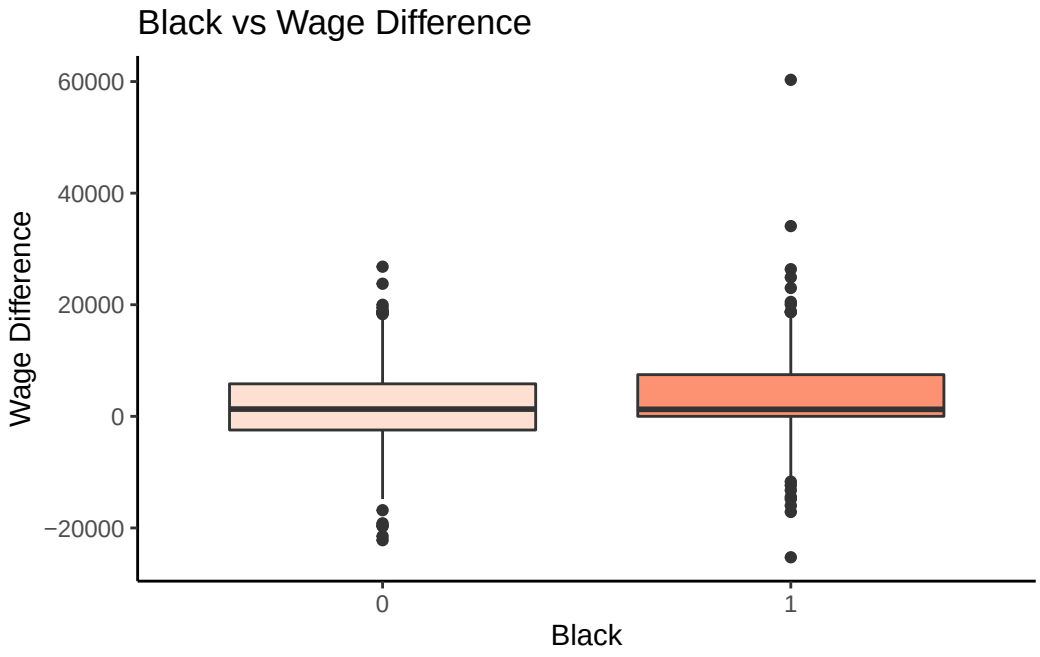
It is important to notice that there are some limitations of this analysis. On the normal q-q plot, there are still a few points on both ends that are deviated from the 45 degree angle line. We decide not to perform transformations, specifically log transformation, because of the negative values in our response variable (wage difference). Furthermore, this analysis has been conducted by only considering a subset of the data containing only male participants. Perhaps a more comprehensive analysis including information on female participants and other possible predictors will lead to a better model and understanding of the relationships in the data.

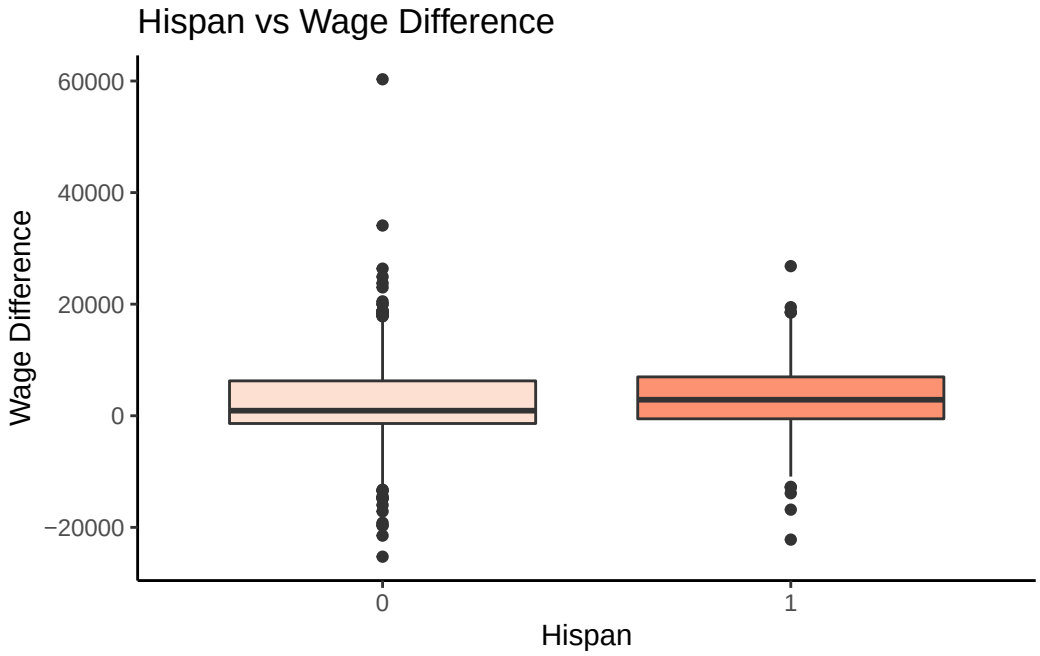
Appendex

EDA

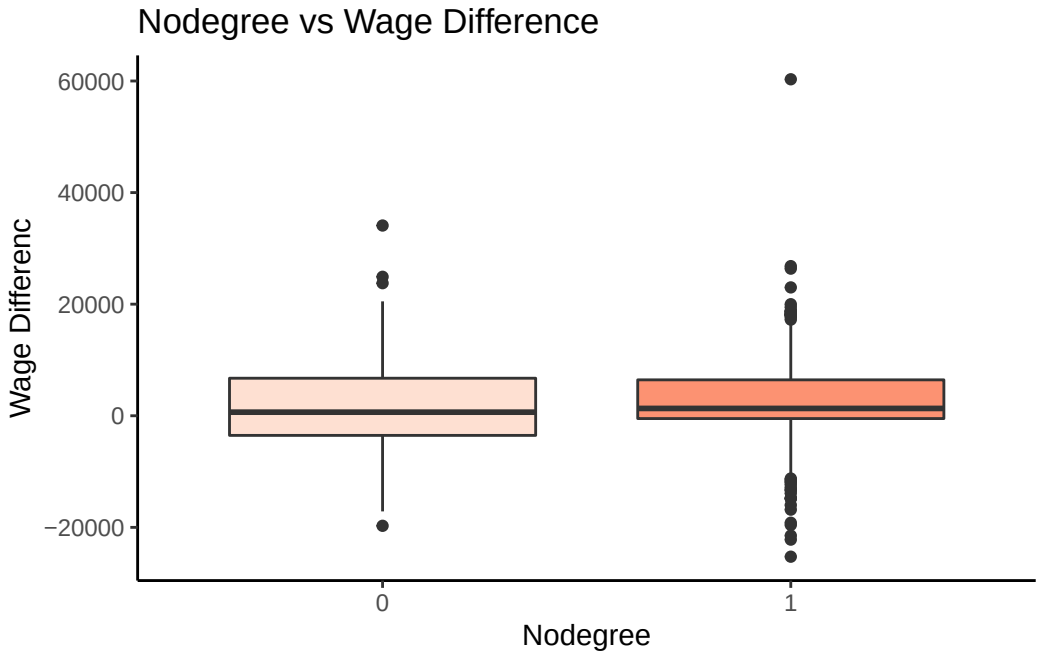


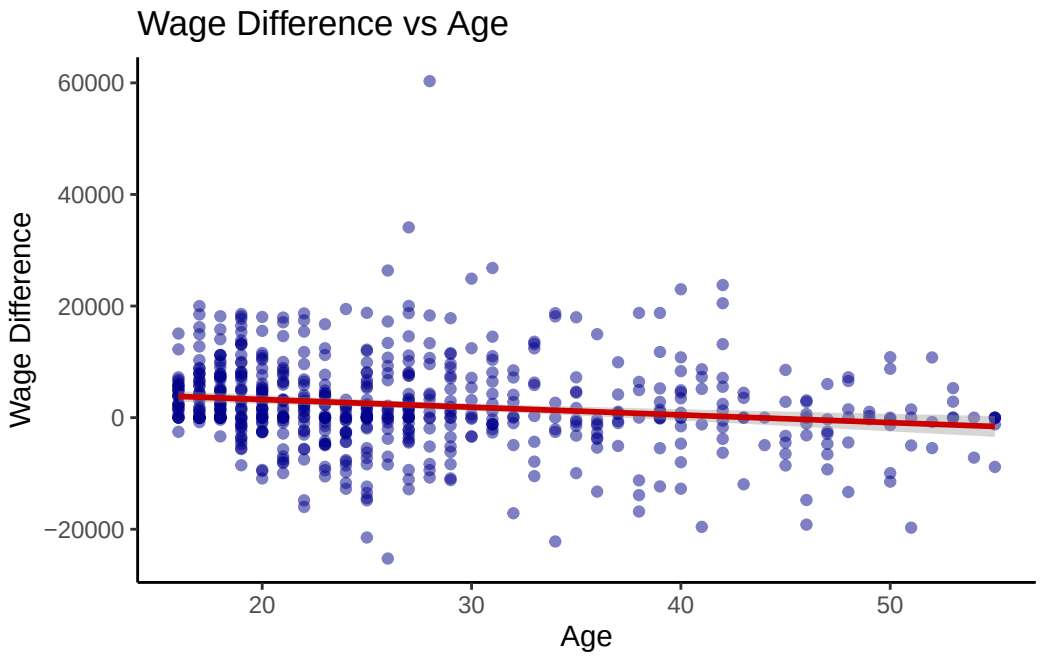

Education vs Wage Difference
60000
40000
20000
0
−20000
Wage Difference
elementary
middle
high
college
Education

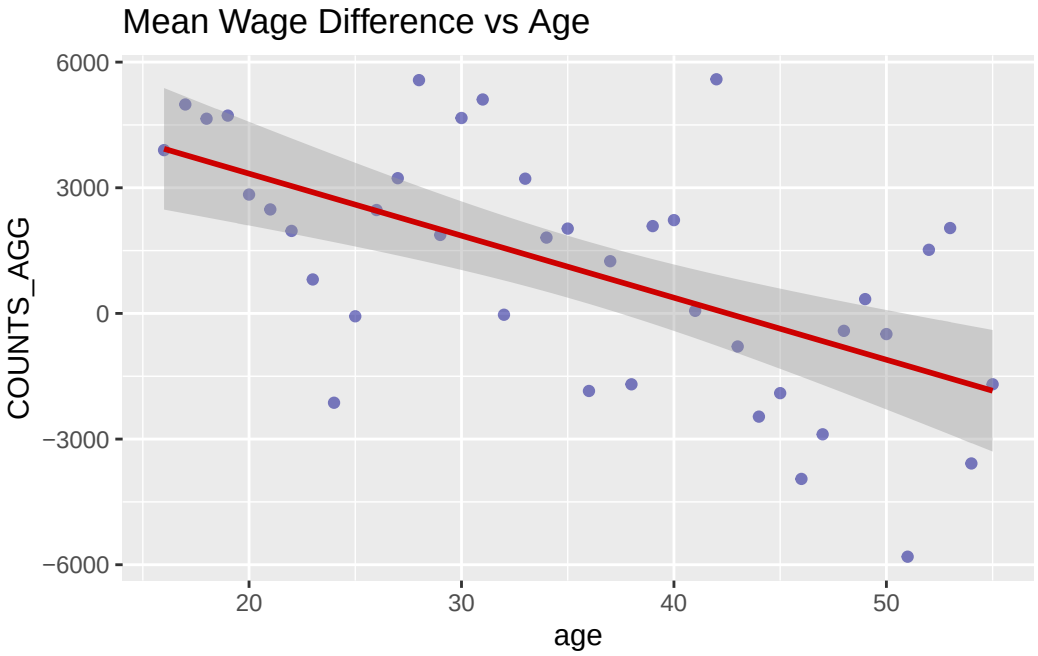

Black vs Wage Difference
60000
40000
20000
0
−20000
Wage Difference
0
1
Black


Hispan vs Wage Difference
60000
40000
20000
0
−20000
Wage Difference
0
1
Hispan

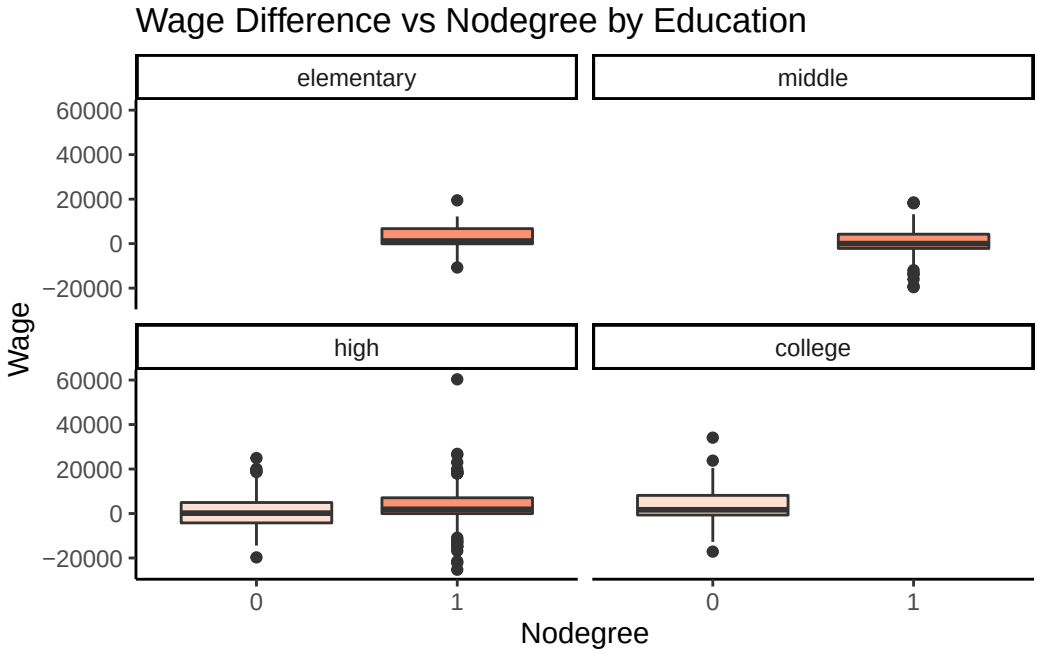

Married vs Wage Difference
60000
40000
20000
0
−20000
Wwage Difference
0
1
Married

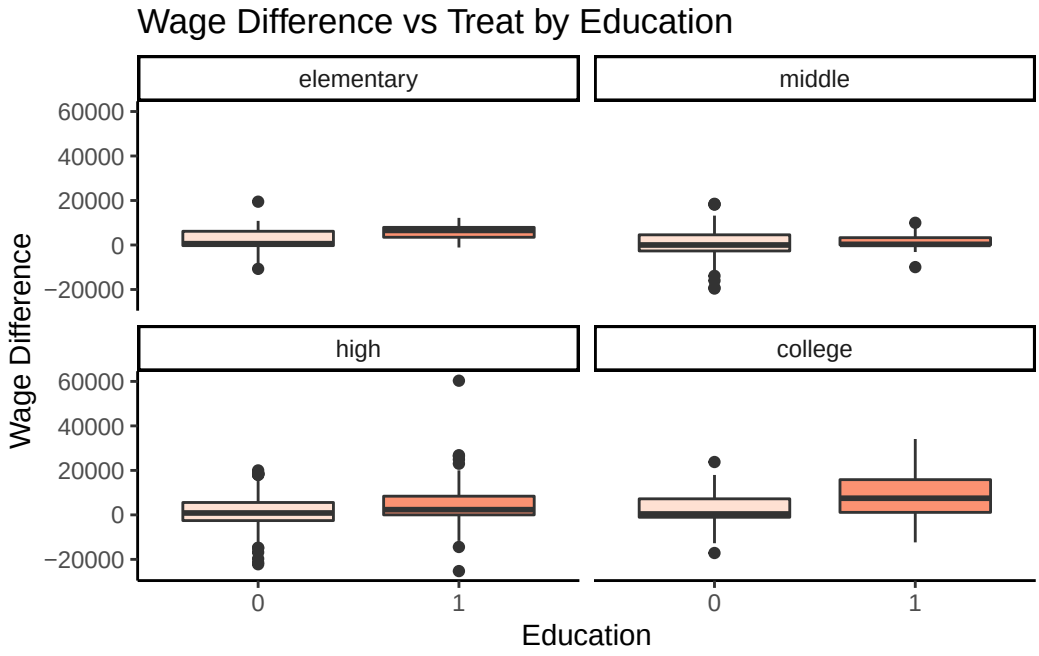

Nodegree vs Wage Difference
60000
40000
20000
0
−20000
Wage Differenc
0
1
Nodegree

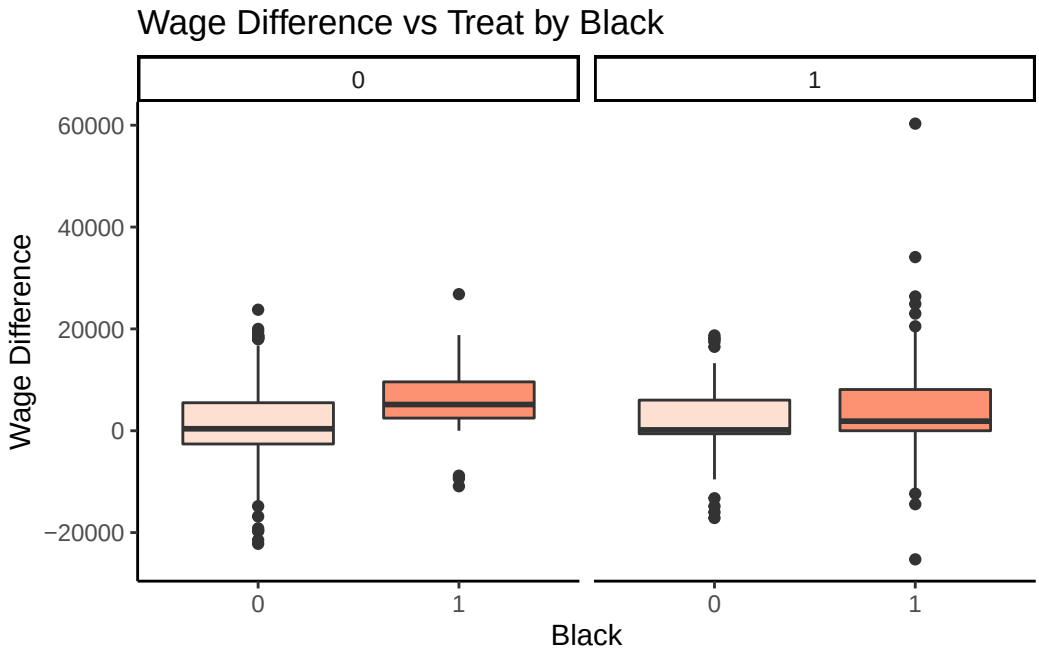

Wage Difference vs Age
60000
40000
20000
0
−20000
Wage Difference
20
30
40
50
Age

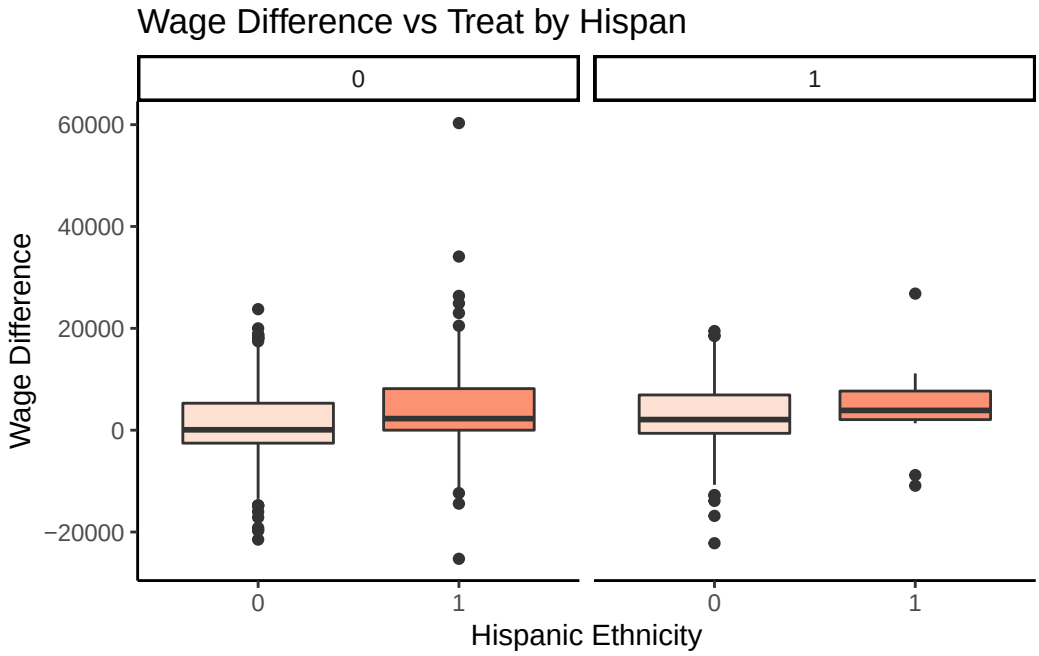


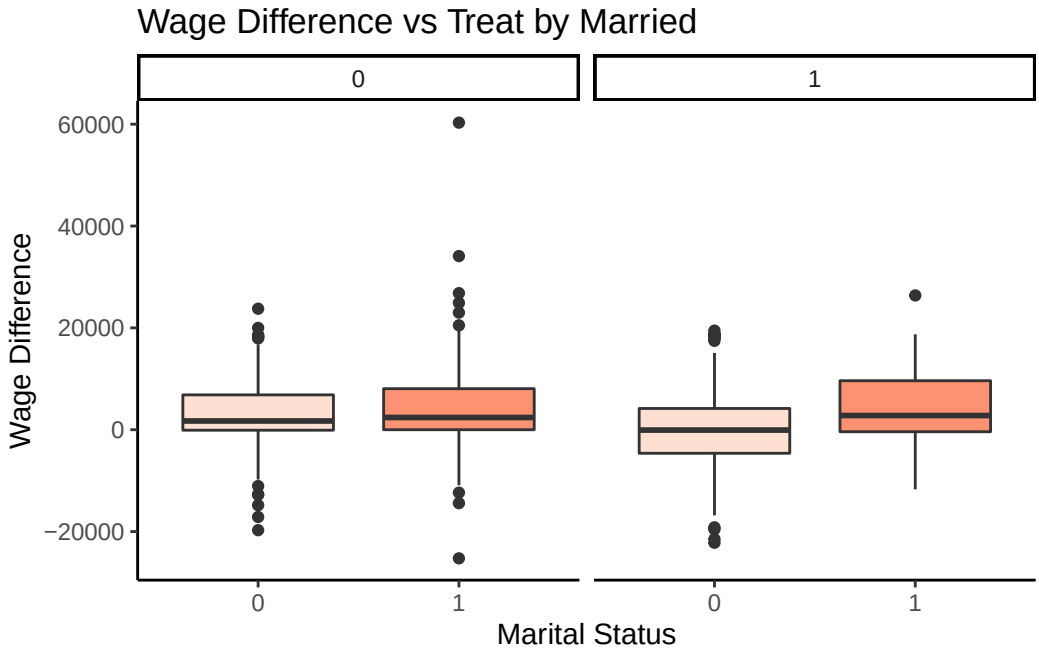
Explore Interactions

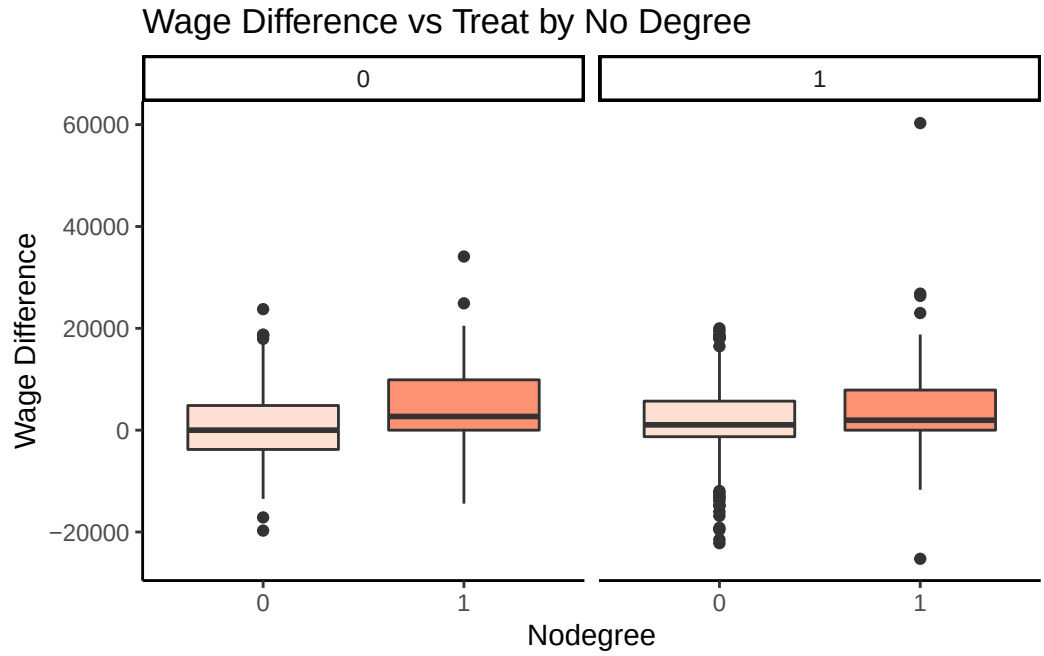


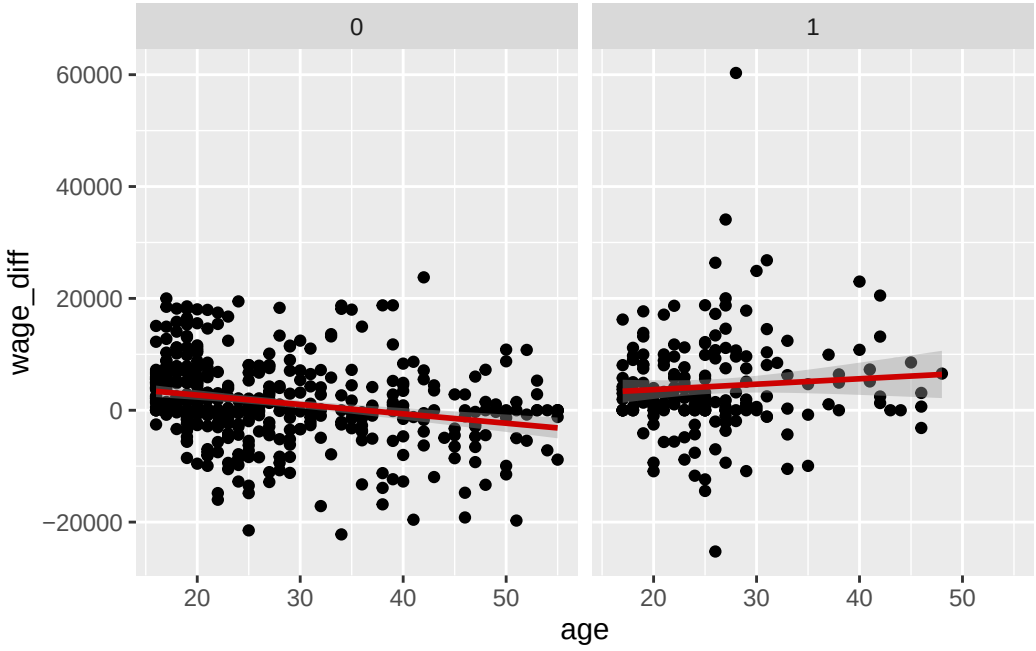

Wage Difference vs Treat by Education
elementary
middle
high
college
60000
40000
20000
0
−20000
Wage Difference
0
1
Education

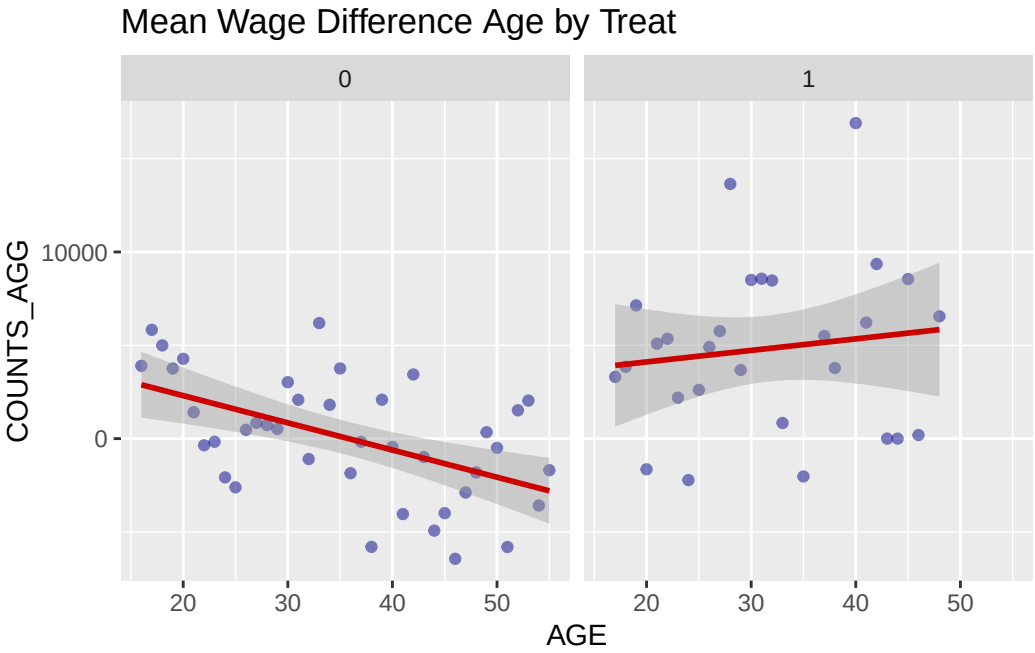

Wage Difference vs Treat by Black
0
1
60000
40000
20000
0
−20000
Wage Difference
Black


Wage Difference vs Treat by Hispan
0
1
60000
40000
20000
0
−20000
Wage Difference
0
1
0
1
Hispanic Ethnicity


Wage Difference vs Treat by Married
0
1
60000
40000
20000
0
−20000
Wage Difference
0
1
0
1
Marital Status


Wage Difference vs Treat by No Degree
0
1
60000
40000
20000
0
−20000
Wage Difference
0
1
0
1
Nodegree


0
1
60000
40000
20000
0
−20000
wage_diff
20
30
40
50
20
30
40
50
age



Model Validation

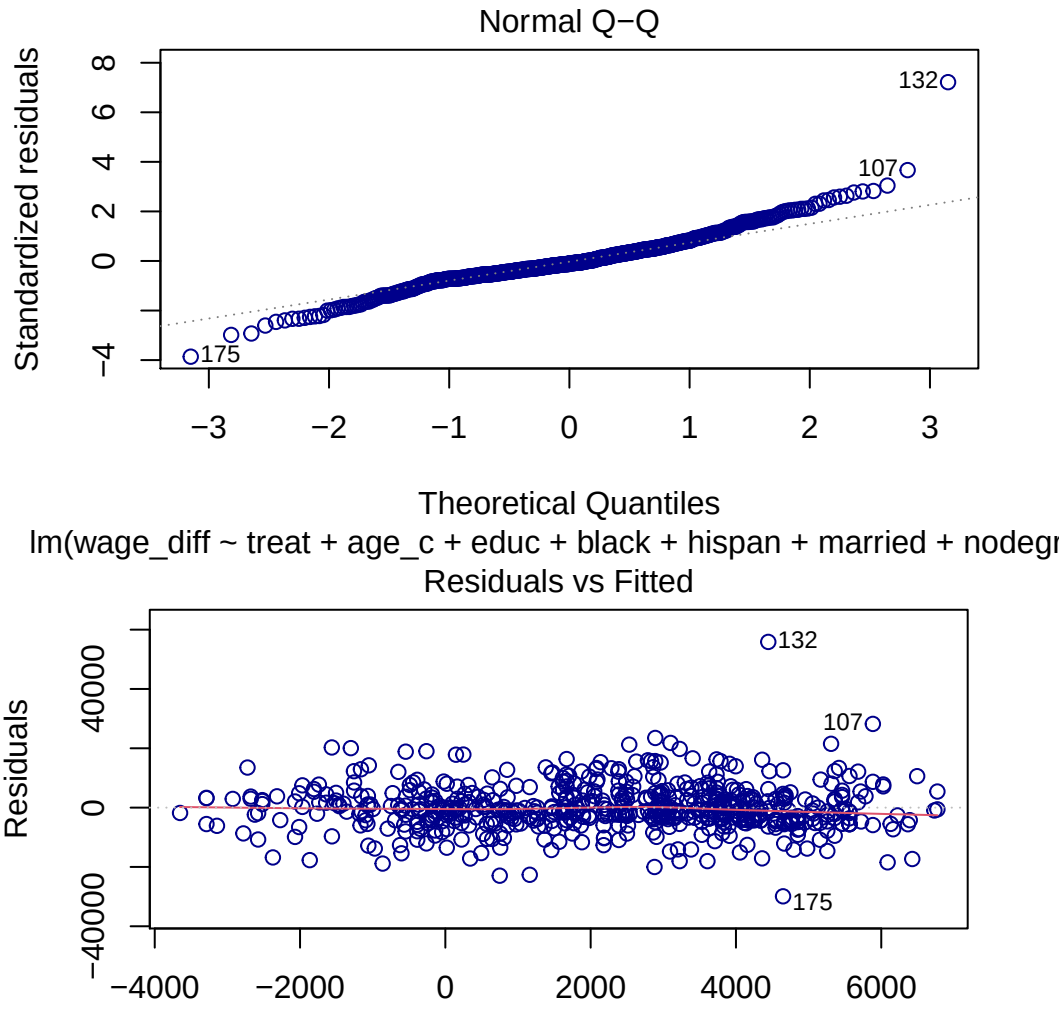
Build a naive model with all main effects:

Table 1: Naive Model Summary

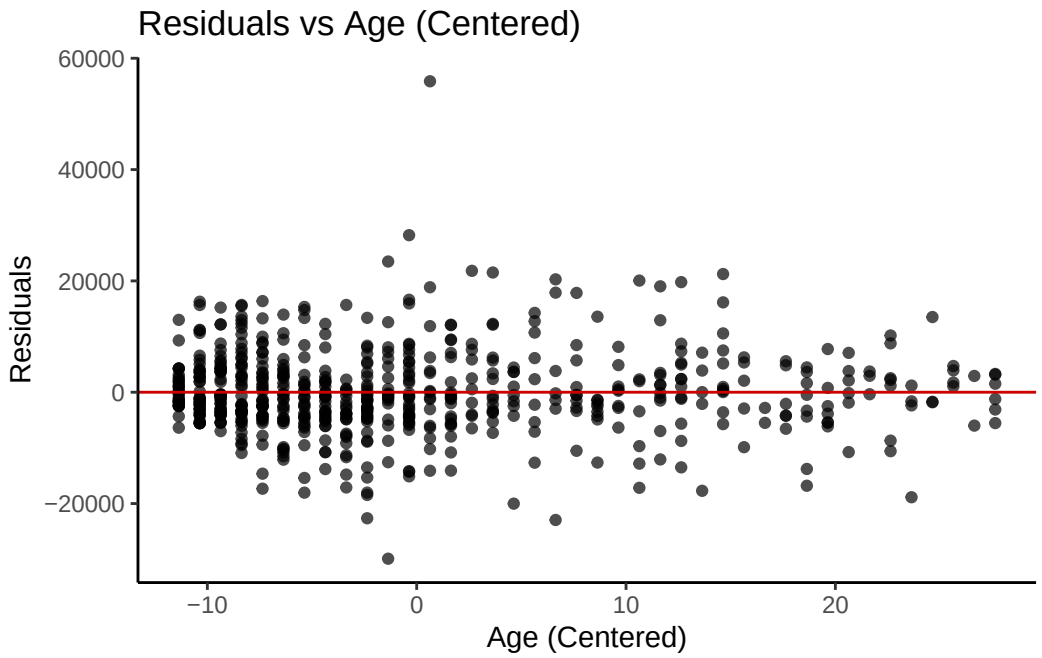
term	estimate	std.error	statistic	p.value
(Intercept)	3560.9417	1741.9307	2.0442	0.0414
treat1	2497.3597	871.8618	2.8644	0.0043
age_c	-101.5772	35.7363	-2.8424	0.0046
educmiddle	-3425.2494	1590.7118	-2.1533	0.0317
educhigh	-2022.8684	1512.1451	-1.3377	0.1815
educcollege	457.1590	1888.3043	0.2421	0.8088
black1	-668.0617	861.3603	-0.7756	0.4383
hispan1	501.0934	1045.2115	0.4794	0.6318
married1	-1758.4304	732.7225	-2.3999	0.0167
nodegree1	1142.1762	808.5269	1.4127	0.1583

Residual standard error: 7771 on 604 degrees of freedom
Multiple R-squared: 0.07802, Adjusted R-squared: 0.06428
F-statistic: 5.679 on 9 and 604 DF, p-value: 1.429e-07

Assess normality assumption:



Assess linearity:



Model Selection

Use BIC as selection criterion
Forward selection:

```
n <- nrow(data_1)
model_forward <- step(null_model, scope = formula(full_model),
                      direction = 'forward', trace = 0, k = log(n))
model_forward$call
```

```
lm(formula = wage_diff ~ treat + age_c + treat:age_c, data = data_1)
```

Stepwise selection:

```
model_stepwise <- step(null_model, scope = formula(full_model),
                       direction="both", trace=0, k = log(n))
model_stepwise$call
```

```
lm(formula = wage_diff ~ treat + age_c + treat:age_c, data = data_1)
```

Backward selection:

```
model_backward <- step(full_model, direction="backward", trace=0, k = log(n))
model_backward$call
```

```
lm(formula = wage_diff ~ treat + age_c + treat:age_c, data = data_1)
```

Create a new model:

```
model_1 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c, data = data_1)
```

Create a different model that includes educ:

```
model_2 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + educ,
              data = data_1)
```

Perform ANOVA test:

```
model_1 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c, data = data_1)
model_2 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + educ,
              data = data_1)
anova(model_1, model_2)
```

```
Analysis of Variance Table

Model 1: wage_diff ~ treat + age_c + treat:age_c
Model 2: wage_diff ~ treat + age_c + treat:age_c + educ
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    610 3.6992e+10
2    607 3.6543e+10  3 449077766 2.4865 0.05965 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Create a different model that includes black:

```
model_3 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + black,
              data = data_1)
anova(model_1, model_3)
```

```
Analysis of Variance Table

Model 1: wage_diff ~ treat + age_c + treat:age_c
Model 2: wage_diff ~ treat + age_c + treat:age_c + black
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    610 3.6992e+10
2    609 3.6974e+10  1  18284190 0.3012 0.5834
```

Create a different model that includes hispan:

```
model_4 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + hispan,
              data = data_1)
anova(model_1, model_4)
```

```
Analysis of Variance Table

Model 1: wage_diff ~ treat + age_c + treat:age_c
Model 2: wage_diff ~ treat + age_c + treat:age_c + hispan
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    610 3.6992e+10
2    609 3.6953e+10  1  39820144 0.6563 0.4182
```

Create a different model that includes married:

```
model_5 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + married,
              data = data_1)
anova(model_1, model_5)
```

```
Analysis of Variance Table

Model 1: wage_diff ~ treat + age_c + treat:age_c
Model 2: wage_diff ~ treat + age_c + treat:age_c + married
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    610 3.6992e+10
2    609 3.6630e+10  1 362278153 6.0231 0.0144 *
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Create a different model that includes nodegree:

```
model_6 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + nodegree,
              data = data_1)
anova(model_1, model_6)
```

```
Analysis of Variance Table

Model 1: wage_diff ~ treat + age_c + treat:age_c
Model 2: wage_diff ~ treat + age_c + treat:age_c + nodegree
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    610 3.6992e+10
2    609 3.6986e+10  1   6602864 0.1087 0.7417
```

Create a different model that includes educ and married:

```
model_7 <- lm(formula = wage_diff ~ treat + age_c + treat:age_c + educ + married,
              data = data_1)
anova(model_1, model_7)
```

```
Analysis of Variance Table

Model 1: wage_diff ~ treat + age_c + treat:age_c
Model 2: wage_diff ~ treat + age_c + treat:age_c + educ + married
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    610 3.6992e+10
2    606 3.6231e+10  4 761263846 3.1832 0.01328 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model assessment on the final model

```
model_final = lm(formula = wage_diff ~ treat + age_c + educ + married + treat:age_c,
                 data = data_1)
summary_final = tidy(model_final)
kable(summary_final, format='markdown', booktabs = T, linesep = "",
      escape = F, caption = "Final Model Summary", digits=4)
```

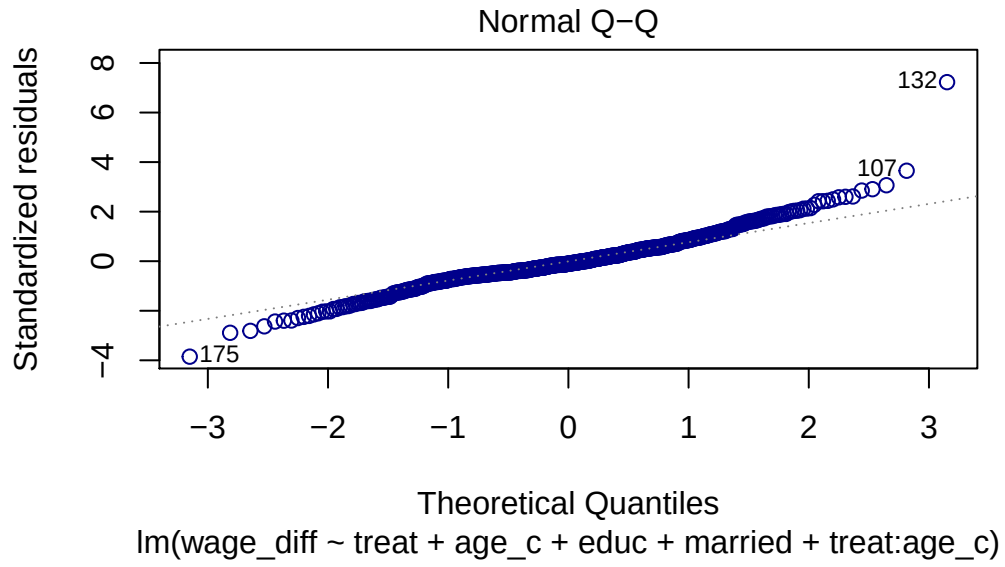
Table 2: Final Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	4530.3047	1444.5334	3.1362	0.0018
treat1	2472.0969	727.3139	3.3989	0.0007
age_c	-146.8542	37.9063	-3.8741	0.0001
educmiddle	-3203.1445	1572.7132	-2.0367	0.0421
educhigh	-2437.8822	1447.5537	-1.6841	0.0927
educcollege	-832.3836	1659.5960	-0.5016	0.6162
married1	-1638.7772	717.1629	-2.2851	0.0227
treat1:age_c	235.0668	87.3893	2.6899	0.0073

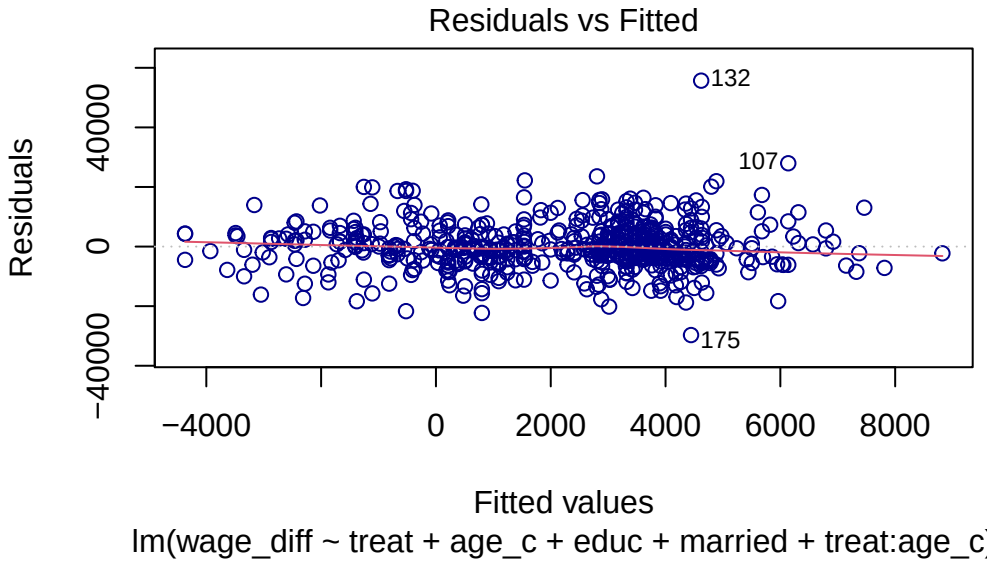
Residual standard error: 7732 on 606 degrees of freedom
Multiple R-squared: 0.08417, Adjusted R-squared: 0.07359
F-statistic: 7.956 on 7 and 606 DF, p-value: 2.954e-09

Assess normality assumption:

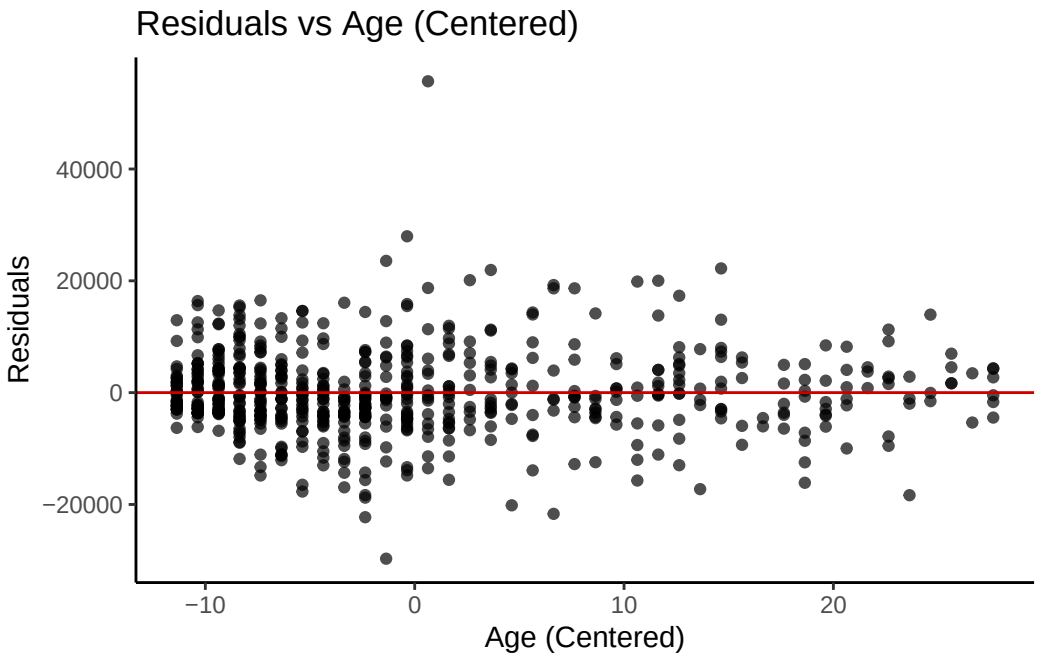
```
plot(model_final, which = 2, col = c('blue4'))
```




```
plot(model_final, which = 1, col = c('blue4'))
```

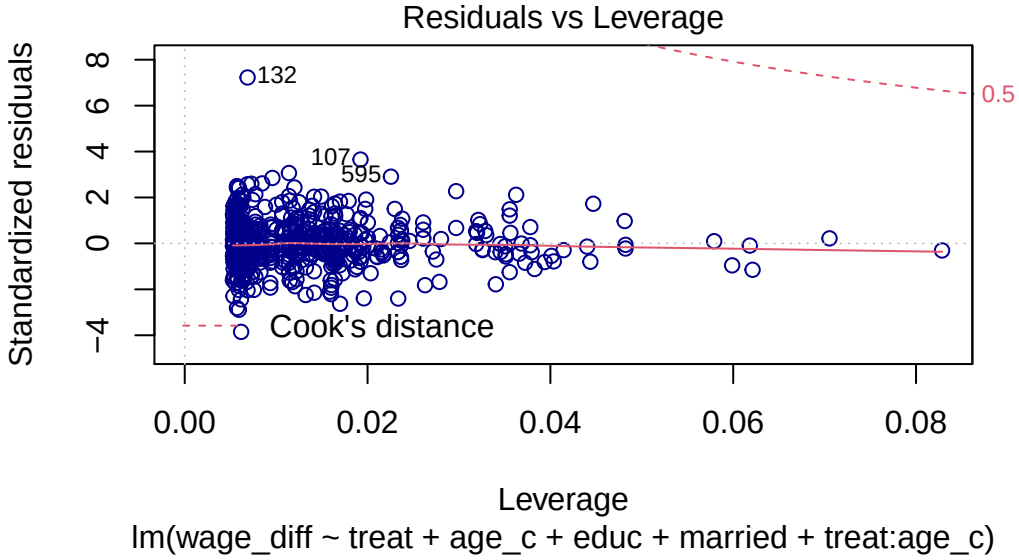



```
ggplot(data_1,aes(x=age_c, y=model_final$residual)) +
  geom_point(alpha = .7) +
  geom_hline(yintercept=0,col="red3") +
  theme_classic() +
  labs(title="Residuals vs Age (Centered)",x="Age (Centered)",y="Residuals")
```



Check for outlier(s):

```
plot(model_final, which = 5, col = c('blue4'))
```



Outlier 132:

```
kable(lalonde_data[132,])
```

	X	treat	age	educ	black	hispan	married	nodegree	re74	re75	re78	wage_diff
132	NSW132	1	28	high	1	0	0	1	0	1284.079	60307.93	60307.93

Check for multicolinearity:

```
kable(vif(model_final))
```

	GVIF	Df	GVIF^(1/(2*Df))
treat	1.143655	1	1.069418
age_c	1.438438	1	1.199349
educ	1.118629	3	1.018860
married	1.282604	1	1.132521
treat:age_c	1.242738	1	1.114781

95% CI:

```
kable(confint(model_final), digits=4)
```

	2.5 %	97.5 %
(Intercept)	1693.4053	7367.2041
treat1	1043.7351	3900.4586
age_c	-221.2978	-72.4106
educmiddle	-6291.7744	-114.5145
educhigh	-5280.7131	404.9488
educcollege	-4091.6414	2426.8742
married1	-3047.2036	-230.3508
treat1:age_c	63.4441	406.6895

Team Orange Project 1 Part 2

1. Summary

This analysis aims to find out whether or not job training for disadvantaged workers had an effect on whether or not they can have positive wages using a subset of the data from the National Supported Work (NSW) containing only male participants from the 1970s. The analysis applies methods including but not limited to: preliminary screening using exploratory data analysis (EDA), model fitting using logistic regression with multiple predictors, model selection using BIC and ANOVA test and Chi-square test, model validation using binned residual plots, as well as model performance assessment using ROC curve and confusion matrix. Our final model shows that there is no evidence suggesting that workers who receive job training tend to be more likely to have positive wages than workers who do not receive job training.

2. Introduction

This analysis aims to explore if there is any evidence that workers who receive job training(`treat`) tend to be more likely to have positive wages than workers who do not receive job training. Our model also quantifies the effect of the job training by providing a 95% confidence interval for odds. In addition, it investigates if this effect differs by demographic groups, and points out other interesting associations with odds of having positive wages. Specifically, when a worker was trained, the older this worker was, the higher the odds of having a positive wage in 1978. Besides, this analysis also finds that black men were disproportionately disadvantaged in the job market, compared with men in other ethnicity groups who had identical attributes in age, training experience, and wage back in 1974.

3. Data

In order to quantify the effect of job training(`treat`) on the odds of having positive wages, this analysis uses `employed78_n` as the predictor variable, which is encoded by whether or not the person had a positive wage in 1978. It also reassembles years of education to 4 groups: elementary(0-5), middle school(6-8), high school(9-12), and college and beyond(13-18) - as a way to better interpret the data and to reduce the effect of erroneous/missing values as well as imbalanced data distribution based on education. In addition, this analysis also uses a mean-centered version of the age variable in order to make intercept interpretation more comprehensible. Unlike `age`, `re74` is interpretable when the value is 0, which means that this person had no wage in 1974. Thus, we do not intend to center `re74`.

3.1 EDA

To begin with, we assess the distribution of response variable by plotting a table of people who had positive wages and zero wages in 1978, finding that the number of people who were having positive wages were about three times as many as those who were not.

First, we look at the EDA for the continuous variables: `age` and `re74`:

- The boxplot of `age` by `treat` shows a difference in the median values and distribution between trained and untrained workers, suggesting that `age` may be a significant variable to include when fitting the model.
- The boxplot of `re74` by `treat` shows a difference in the median values and distribution between trained and untrained workers, suggesting that `re74` may be an importnat variable to include when fitting the model.

Next, we look at the EDA for the discrete variables:

- The table of `employed78_n` and `treat` shows no signifinant difference in the odds of having positive wages.
- The table of `employed78_n` and `edcu` implies that people who had a high school degree were most likely to have positive wages compared to people with other levels of education. However, this can be due to insufficient data for people who had education level other than high school, and this factor should be taken into consideration in model-fitting.

- The table of `employed78_n` and `black` shows that black men were less likely to have positive wages than men in other ethnicity groups.

- The table of `employed78_n` and `hispan` shows that hispanic men had a higher odds of having positive wages than non-hispanic men. Nevertheless, an important fact to note is that non-hispanic actually includes black men who were noticeably under-represented in the job market. Thus, the seemingly high odds of employment of `hispan` might be a result of inflated overall unemployement rate that all other ethnicity groups combined.
- The table of `employed78_n` and `married` shows that married men were more likely to have positive wages.
- The table of `employed78_n` and `no-degree` shows that high school dropouts were less likely to have positive wages than those who had education level of high shcool or above.

Through visual inspection, all predictor variables are worth further statistical investigation.

3.2 Explore Interaction Terms

Because `nodegree` and `educ` contain similar information, there is no need to explore interactions between these two variables. Since we are interested in evidence that demonstrate a relationship between treat and employment, we then focus on the potential interactions among employment, treat, and other predictors.

Same with EDA, we look at the interaction effect for the continuous variables first:

- The boxplot of employment status and treat by age shows different age-employement patterns by treatment group in the median values, suggesting that `treat:age` may not be a significant interaction term to include when fitting the model.

Next, we look at the interaction effect for the discrete variables:

- The contingency table between treat and employment by different education level shows that :
  1. College graduates had higher probablity of employement when trained compared with those who did not receive the traning.
  2. High school graduates had lower odds when trained compared with those who were not trained.
  3. Middle school graduates had lower odds when trained compared with those who did not receive job traning.
  4. Elementary school graduates who were trained all had positive wages, which could be attributed to the insuffficiency of data about this group.
- The contingency table between treat and employment by black and non-black shows that :
  1. Trained black had a slightly higher odds of having positive wages than black that were not trained.
  2. The marginal increase of employment odds for trained non-black was much higher than those non-black who were not trained.
- The contingency table between treat and employement by hispanic and non-hispanic shows that :
  1. All trained hispanic men ended up getting hired.
  2. Trained non-hispanic had decreased odds of having positive wages, which can be a reflection of an under-represented black community.
- The contingency table between treat and employement by marital status shows that employment odds were higher for those who were married, and lower odds for those who were not.
- The contingency table between treat and employement by no-degree status shows that employment odds were higher for trained workers with a highshcool degree or above, while the odds were lower for trained workers without a highschool degree.

Through visual inspection on the potential interaction effect, the interaction between `age` and `treat` is most likely to be significant. However, EDA only offers a visual representation of the data, and the actual significance of each variable needs to be assessed by statistical tests. Specifically, if a predictor has been deemed significant through EDA but is dropped through model selection, we will perform ANOVA test to determine its statistical significance.

4. Model

4.1 Model Validation

We first center the age predictor variable so that it's most meaningful to interpret the intercept. Then we build a naive model with all main effects including `age_c`, `educ`, `black`, `hispan`, `married`, `re74` and `nodegree`. From the summary output, age(centered), re74, edumiddle, and black are significant predictor variables. Next, we moved to model assessment using binned residual plots which map model residuals against predicted values and against continuous predictor variable.

The residual plot between residuals and predicted values implies a trend that is somewhat quadratic, and this is worth taking notes when we get to our final model assessment. The residual plot between residuals and centered age does not have any clear pattern, where points are randomly distributed both above and below zero.

As we get to model validation using confusion matrix with a threshold level of 0.5, we found that the model has a 0.78 accuracy but staggeringly different sensitivity and specificity: 0.98 and 0.10, respectively. In order to minimize the penalty on sensitivty, we decide to set the threshold at the mean value of `employed78_n` instead, and obtained an accuracy of 0.61, sensitivity and specificity of 0.61 and 0.60, respectively.

When examined using ROC, this naive model got an AUC value of 0.6558.

Overall, the naive model summary shows that `age_c`, `re74`,`edumiddle`, and `black` are statistically significant predictor variables. Even though other predictor variables don't show any statistical significance, it could be that the linear relationship between each one of these variable is not strong enough to be detected by this sample.

4.2 Model Selection

In this section, we perform model selection, with null_model only capturing the predictor that we or our client care about - treatment. And the full_model will include all main effects as well as all interactions between treat and other variables. If any variables that were previously deemed statistically significant are removed through the model selection process, we will perform ANOVA with Chi-square test on these terms to evaluate their significance. Because the data is not health related, we use BIC as our selection criterion as it is more strict at selecting variables, and False Negative (FN) as well as False Positive (FP) don't matter much in this case.

The results from backward and forward selection returned the same model, both using `age_c`, `treat`, `re74`, and `treat:age_c` as predictors. On the other hand, stepwise model selection return two predictors: `age_c` and `re74`. The next step of model selection is then to test whether or not `treat` and `treat:age_c` are signifncant predictors. We do so by running an ANOVA test using Chi-square between the stepwise model and two other models (treatagemodel and treatmodel) which only differ from the stepwise model by one term: `treat` or `treat:age_c`.

The test between `treatmodel`and `BIC_stepwise_model`show that `treat` is not sifnicant with p-value = 0.79. The deviance test also shows that `treat` is not a valuable predictor for positive wage, given that deviance value only decreased by less than 1 unit with its presence in the new model. However, since it is our variable of interest, we decide to keep `treat` as part of our analysis.
The test between `treatagemodel`and `BIC_stepwise_model` show that `treat:age_c` is sifnicant with p-value = 0.0079. As a result, this interaction term is worth keeping in the final model.

However, given that BIC is only a criterion which does not determine the significance of a predictor. The next thing is to test predictor variables not recommended by BIC one at a time and compare the new model with our `BIC_forward_model` which works as a tentative baseline model. After testing variables besides the ones in the tentative model, we find that among `black`, `hispan`, `married`,`nodegree`, only the new model with `black` is significant in the ANOVA test. Thus, we will also include `black` into our tentative model despite the fact that BIC method filtered out this variable.
The last thing before finalizing the model is to test all interaction terms and look for ones that are statistically significant.

- Firstly, we create a different model that includes the interaction between `treat` and `educ` on top of the BIC_forward_model. The ANOVA test between these two has a p-value of 0.066, showing that this interaction term is not statistically sifnificant.
- Next, we create a different model that includes the interaction between `treat` and `black` on top of the BIC_forward_model. The ANOVA test between these two has a p-value of 0.12, showing that this interaction term is not statistically significant.
- Next, we create a different model that includes the interaction between `treat` and `married` on top of the BIC_forward_model. The ANOVA test between these two has a p-value of 0.47, showing that this interaction term is not statistically significant.
- Next, we create a different model that includes the interaction between `treat`and `nodegree` on top of the BIC_forward_model. The ANOVA test between these two has a p-value of 0.90, showing that this interaction term is not statistically significant.
- Next, we create a different model that includes the interaction between `treat`and `re74` on top of the BIC_forward_model. The ANOVA test between these two has a p-value of 0.13, showing that this interaction term is not statistically significant.
- Based on the information above, as none of these interaction terms are statistically significant, we decide to not include any one of these interaction terms.

Therefore, the final model will include `treat`, `age_c`, `re74`, `black` and `treat:age_c`.

4.3 Final Model

4.3.1 Model Assessment

After fitting our final model, we use binned residual plots to first assess the overall model fitting, and then examine residuals against the two continuous variables in the model : `age_c` and `re75`. All three graphs turn out to have a random distribution of points, and the quadratic trend that appears in the naive model does not exist anymore. Besides, no outliers are spotted. Thus we can be confident that our model is valid.

Using confusion matrix, we find that at the mean employmed78_n level, our model has an accuracy of 60%, sensitivity at 0.61 and specificity at 0.60. The ROC curve shows that the final model has an AUC value of 0.64.

In the end, we check for multicolinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of collinearity. The VIF values of all the predictor variables are below 2, which suggests that there should be no multicolinearity issues.

An important thing to note is that our final model in fact does not have a better performance compared with the naive model in terms of accuracy, sensitivity, specifificy, and AUC. It can be due to the fact that the naive model is overfitted to the data whereas the final model aims to generalize a trend from the data based on fewer predictors. As a result, we decide to move on with the final model which has higher external validity.

4.3.2 Model Interpretation

Final Model Equation:

$$logit(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}) = 1.084 + 0.5027 * \hat{treat}1 - 0.05374\hat{age}_c - 0.00007079 * \hat{re74} - 0.6217\hat{black}1 - 0.07340\hat{treat1:age_c}$$

Apart from `treat1` being the only predictor which has a p-value of 0.062, all other predictor variables (including the intercept) are significant at p-value < 0.05 significance level. After converting the coefficients to odds scale, the interpretations of all predictor variables are as follows:

- intercept: For a male who is not black, untrained, at an average age, with an elementary education level, the odds for this person to have positive wages are expected to be 2.96.
- treat: Holding all other variables constant, a perosn who received training was expected to have increased odds of having positive wages in 1978 by a factor 1.65.
- re74: Holding all other variables constant, every unit increase in the person's wage in 1974 is expected to increase the odds of this person having positive wages in 1978 by a factor of 1.000071.
- age_c: Holding all other variables constant, every unit increase in the person's age is expected to decrease the odds of this person having positive wages in 1978 by a factor of 0.95.
- black1: Holding all other variables constant, the odds of having positive wages for a man of black ethnicity are expected to decrease by a factor of 0.54.
- treat1:age_c: With every year increase in age given the treatment, the odds of having positive wages are expected to increase by a factor of 1.076 on top of the effect that treat has, holding other variables constant. Similarly, compared with a person who was not trained, the odds of having positve wages for a trained person at the same age are expected to increase by 1.076 on top of the effect that age has on positive wages odds.

5. Conclusion

To conclude from the final model, there is no evidence that supports the argument that workers who received job training tend to be more likely to have positive wages than those who did not. As we interpreted above, compared to `treat0`(non-trained workers), the odds of having non-zero wages for treat1(trained workers) are expected to increase by a factor of 1.65 from 1974 to 1978, holding other variables constant. According to the 95% confidence interval of treat1 from the final model, a likely range of the multiplicative effect that `treat` has on having positive wages in 1978 is between 0.98 and 2.82.In addition, there is evidence to suggest that the effects of job training differ by demographic groups. This can be seen from the both the main effect of `age` and the interaction effect between `treat1:age_c`. Looking at `age` alone, the older the man was, the lower the odds were for him to have positive wages in 1978. However, when controlling for treatment (i.e., for trained workers), the older one was, the more likely one would have positive wages in 1978. Thus, we may conclude that older men mgiht benefit more from receiving the training than younger people in terms of having non-zero wages.

An interesting finding regarding having non-zero wages is that it is highly correlated with one's wage in 1974. This is also valid intuitively, given that people who had higher wages in 1974 were more likely to be competitive in the job market, and thus it was more likely for them to have positive wages in 1978.

Another important finding can be derived from the multiplicative effect that `black` has on employment. Based on the summary output from the final model, it is noticeable that holding all other variables constant, being a member of the black community is expected to decrease this person's odds of having positive wages in 1978 by 0.54, with a 95% confidence interval from 0.33 to 0.87. This model output reflects a situation where black men were significantly disadvantaged in the job market, even with identical attributes in other categories such as age and traning status. This result opens up a window for observing racial inequality in the labor market, as well as setting a standard to which we can compare recent data to examine progress or improvement on such inequalities.

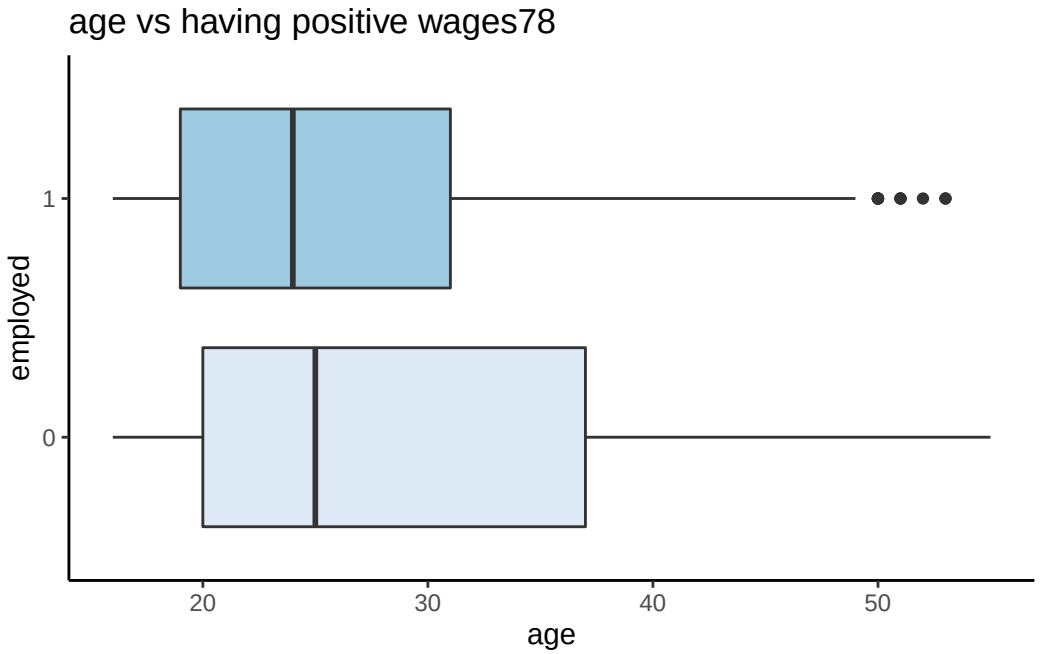
It is also important to note that there are some limitations in our model and model analysis. First, our model is based on a subset of data where only male participants were included. Perhaps a more comprehensive analysis including information on female participants and other possible predictors will lead to a better model and understanding of the relationships in the data. Besides, the classification of `black` and `hispan` in our data are dichotomoou, which simply divides people into either belonging to one ethnicity group or not. This practice makes it easier to have inflated positive wage odds for Hispanic people, given that non-hispanic group actually includes African American who are disproportionately disadvantaged in the job market.

Appendix

EDA

EDA with continuous variables

```
ggplot(lab,aes(x=employed78, y=age, fill=employed78)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="age vs having positive wages78",
       x="employed",y="age") +
  theme_classic() + theme(legend.position="none")
```




```
ggplot(lab,aes(x=employed78, y=re74, fill=employed78)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="re74 vs having positive wages78",
       x="employed",y="anual total income 1974") +
  theme_classic() + theme(legend.position="none")
```

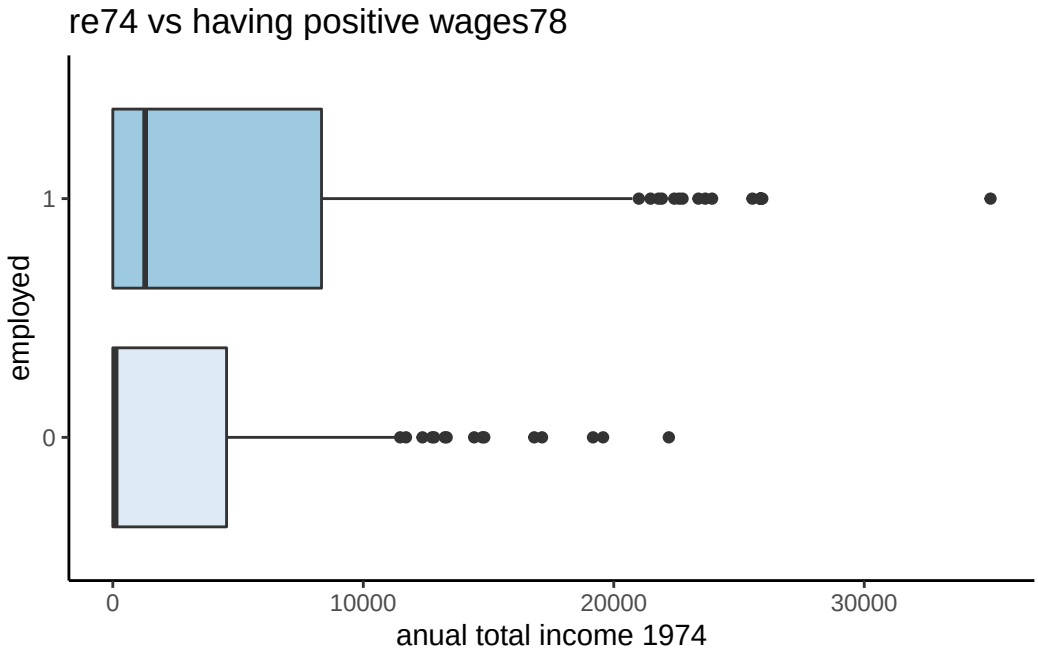

re74 vs having positive wages78
employed
1
0
0
10000
20000
30000
anual total income 1974

Table 1: Employed78 vs. Treat

	0	1
0	0.2284382	0.2432432
1	0.7715618	0.7567568

Table 2: Employed78 vs. Education

	elementary	middle school	high school	college and beyond
0	0.2424242	0.3564356	0.204878	0.2142857
1	0.7575758	0.6435644	0.795122	0.7857143

EDA with discrete variables

Employed vs Treat:

```
kable(apply(table(lab[,c("employed78","treat")])/sum(table(lab[,c("employed78","treat")])),
      2,function(x) x/sum(x)),caption = 'Employed78 vs. Treat')
```

Employed vs Education:

```
kable(apply(table(lab[,c("employed78","educ")])/sum(table(lab[,c("employed78","educ")])),
      2,function(x) x/sum(x)),caption = 'Employed78 vs. Education')
```

Employed vs Black:

```
kable(apply(table(lab[,c("employed78","black")])/sum(table(lab[,c("employed78","black")])),
      2,function(x) x/sum(x)), caption = 'Employed78 vs. Black')
```

Employed vs. Hispanic:

```
kable(apply(table(lab[,c("employed78","hispan")])/sum(table(lab[,c("employed78","hispan")])),
      2,function(x) x/sum(x)),caption = 'Employed78 vs. Hispanic')
```

Employed vs Marrid:

```
kable(apply(table(lab[,c("employed78","married")])/sum(table(lab[,c("employed78","married")])),
      2,function(x) x/sum(x)),caption = 'Employed78 vs. Marital Status')
```

Employed vs NoDegree:

```
kable(apply(table(lab[,c("employed78","nodegree")])/sum(table(lab[,c("employed78","nodegree")])),
      2,function(x) x/sum(x)),caption = 'Employed78 vs. NoDegree')
```

Table 3: Employed78 vs. Black

	0	1
0	0.1994609	0.2839506
1	0.8005391	0.7160494

Table 4: Employed78 vs. Hispanic

	0	1
0	0.2416974	0.1666667
1	0.7583026	0.8333333

Table 5: Employed78 vs. Marital Status

	0	1
0	0.2423398	0.2196078
1	0.7576602	0.7803922

Interaction

```
ggplot(lab,aes(x=employed78, y=age, fill=employed78)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="age vs having positive wages78",
       x="employed",y="age") +
  theme_classic() + theme(legend.position="none") + facet_wrap(~treat)
```

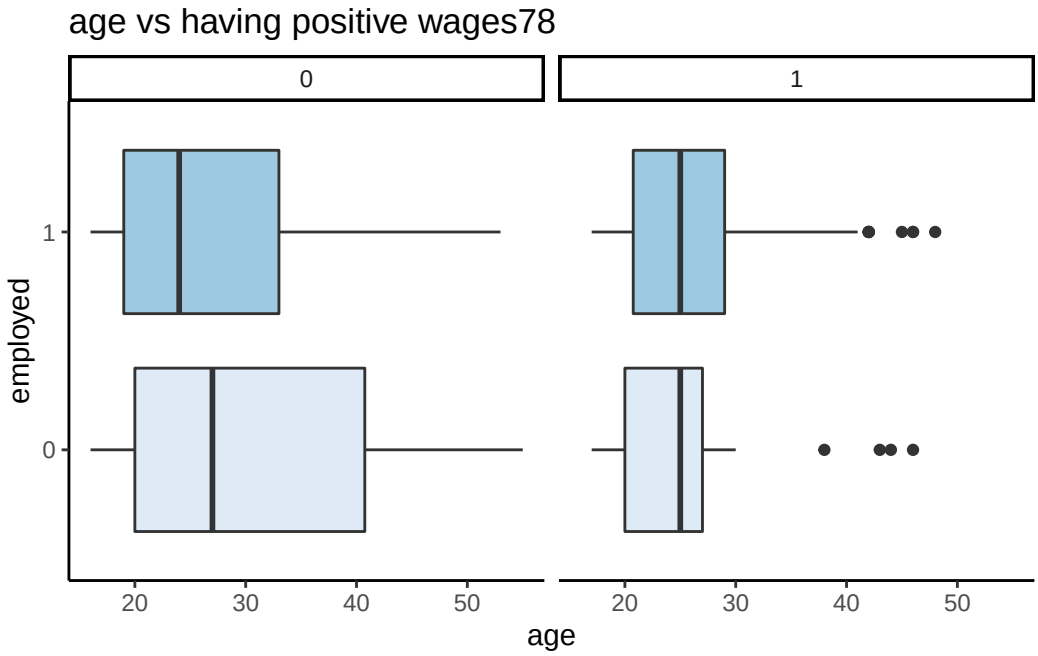


Table 6: Employed78 vs. NoDegree

	0	1
0	0.215859	0.2428941
1	0.784141	0.7571059

```
ggplot(lab,aes(x=employed78, y=re74, fill=employed78)) +
  geom_boxplot() + coord_flip() +
  scale_fill_brewer(palette="Blues") +
  labs(title="treat vs having positive wages78",
       x="employed",y="age") +
  theme_classic() + theme(legend.position="none") + facet_wrap(~treat)
```

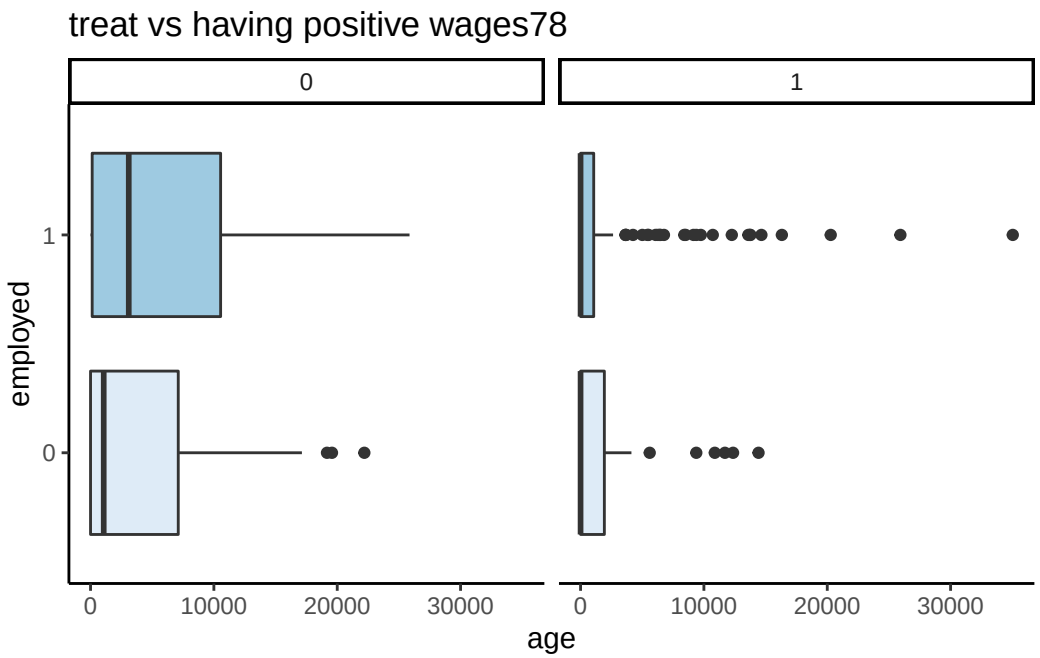


Table 7: Treat vs. College Degree

	0	1
0	0.2181818	0.2
1	0.7818182	0.8

Table 8: Treat vs. Highschool Degree

	0	1
0	0.1902985	0.2323944
1	0.8097015	0.7676056

```
college_lab <- subset(lab, educ == "college and beyond")
highschool_lab <- subset(lab, educ == "high school")
middle_lab <- subset(lab, educ == "middle school")
elementry <- subset(lab, educ == "elementary")
```

Employed78 vs Treat by Educ:

```
kable(apply(table(college_lab[,c("employed78","treat")])/sum(table(college_lab[,
      c("employed78","treat")])), 2,function(x) x/sum(x)),
      caption = "Treat vs. College Degree")
```

```
kable(apply(table(highschool_lab[,c("employed78","treat")])/sum(table(highschool_lab[,
      c("employed78","treat")])), 2,function(x) x/sum(x)),
      caption = "Treat vs. Highschool Degree")
```

```
kable(apply(table(middle_lab[,c("employed78","treat")])/sum(table(middle_lab[,
      c("employed78","treat")])), 2,function(x) x/sum(x)),
      caption = "Treat vs. Middleschool Degree")
```

```
kable(apply(table(elementry[,c("employed78","treat")])/sum(table(elementry[,
      c("employed78","treat")])),2,function(x) x/sum(x)),
      caption = "Treat vs. Elementary School Degree")
```

Employed78 vs Treat by Black:

```
black_lab <- subset(lab, black == 1)
noblack_lab <- subset(lab, black == 0)

kable(apply(table(black_lab[,c("employed78","treat")])/sum(table(black_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),
      caption = "Treat vs. Black")
```

```
kable(apply(table(noblack_lab[,c("employed78","treat")])/sum(table(noblack_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),
      caption = "Treat vs. NonBlack")
```

Table 9: Treat vs. Middleschool Degree

	0	1
0	0.3375	0.4285714
1	0.6625	0.5714286

Table 10: Treat vs. Elementary School Degree

	0	1
0	0.3076923	0
1	0.6923077	1

Table 11: Treat vs. Black

	0	1
0	0.2988506	0.275641
1	0.7011494	0.724359

Employed78 vs Treat by Hispan:

```
hisp_lab <- subset(lab, hispan == 1)
nohisp_lab <- subset(lab, hispan == 0)

kable(apply(table(hisp_lab[,c("employed78","treat")])/sum(table(hisp_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),
      caption = "Treat vs. NonHispanic")
```

```
kable(apply(table(nohisp_lab[,c("employed78","treat")])/sum(table(nohisp_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),
      caption = "Treat vs. NonHispanic")
```

Employed78 vs Treat by Married:

```
married_lab <- subset(lab, married == 1)
nomarried_lab <- subset(lab, married == 0)

kable(apply(table(married_lab[,c("employed78","treat")])/sum(table(married_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),caption = "Treat vs. Married")
```

```
kable(apply(table(nomarried_lab[,c("employed78","treat")])/sum(table(nomarried_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),caption = "Treat vs. NotMarried")
```

Employed78 vs Treat by Nodegree:

```
degree_lab <- subset(lab, nodegree == 0)
nodegree_lab <- subset(lab, nodegree == 1)

kable(apply(table(degree_lab[,c("employed78","treat")])/sum(table(degree_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)), caption = "Treat vs. Degree")
```

```
kable(apply(table(nodegree_lab[,c("employed78","treat")])/sum(table(nodegree_lab[,
      c("employed78","treat")])),2,function(x) x/sum(x)),  caption = "Treat vs. Nodegree")
```

Table 12: Treat vs. NonBlack

	0	1
0	0.2105263	0.0689655
1	0.7894737	0.9310345

Table 13: Treat vs. NonHispanic

	0	1
0	0.1967213	0
1	0.8032787	1

Table 14: Treat vs. NonHispanic

	0	1
0	0.2336957	0.2586207
1	0.7663043	0.7413793

Table 15: Treat vs. Married

	0	1
0	0.2318182	0.1428571
1	0.7681818	0.8571429

Table 16: Treat vs. NotMarried

	0	1
0	0.2248804	0.2666667
1	0.7751196	0.7333333

Table 17: Treat vs. Degree

	0	1
0	0.2196532	0.2037037
1	0.7803468	0.7962963

Table 18: Treat vs. Nodegree

	0	1
0	0.234375	0.259542
1	0.765625	0.740458

Naive Model

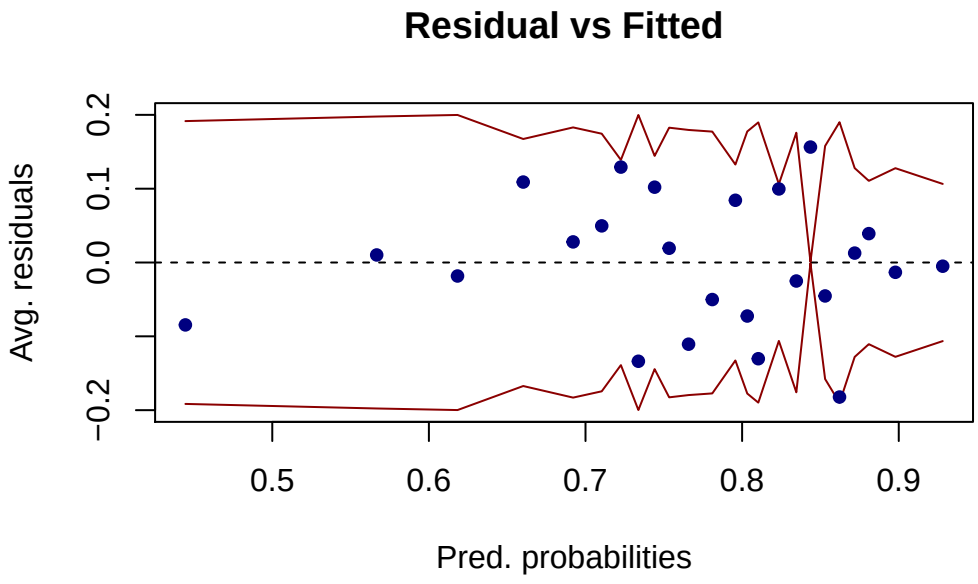
```
lab$age_c <- lab$age - mean(lab$age)
labreg <- glm(employed78 ~ age_c + re74 + treat + educ + black + hispan + married + nodegree,
              data = lab, family = binomial)
summary = tidy(labreg)
kable(summary,format='markdown', booktabs = T, linesep = "", escape = F, caption = "Naive Model Summary
```

Table 19: Naive Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	1.3053	0.5551	2.3513	0.0187
age_c	-0.0412	0.0109	-3.7744	0.0002
re74	0.0001	0.0000	3.1306	0.0017
treat1	0.3928	0.2684	1.4635	0.1433
educmiddle school	-1.0729	0.4933	-2.1750	0.0296
educhigh school	-0.2807	0.4826	-0.5817	0.5608
educcollege and beyond	-0.1679	0.6025	-0.2787	0.7805
black1	-0.6127	0.2652	-2.3106	0.0209
hispan1	0.1981	0.3617	0.5477	0.5839
married1	0.1137	0.2405	0.4728	0.6363
nodegree1	0.2059	0.2641	0.7798	0.4355

```
rawresid1 <- residuals(labreg,"resp")
```

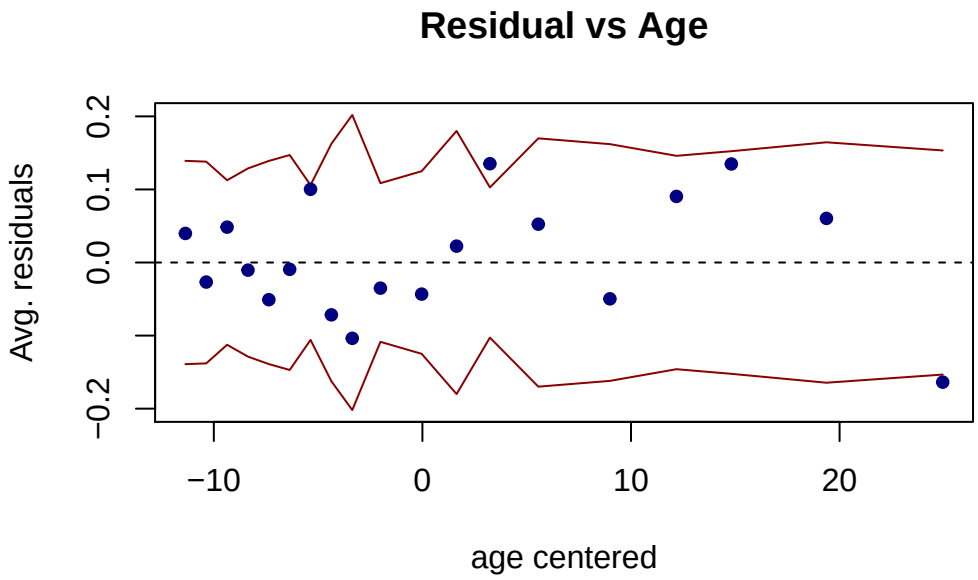
```
binnedplot(x=fitted(labreg),y=rawresid1,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Residual vs Fitted",col.pts="navy")
```



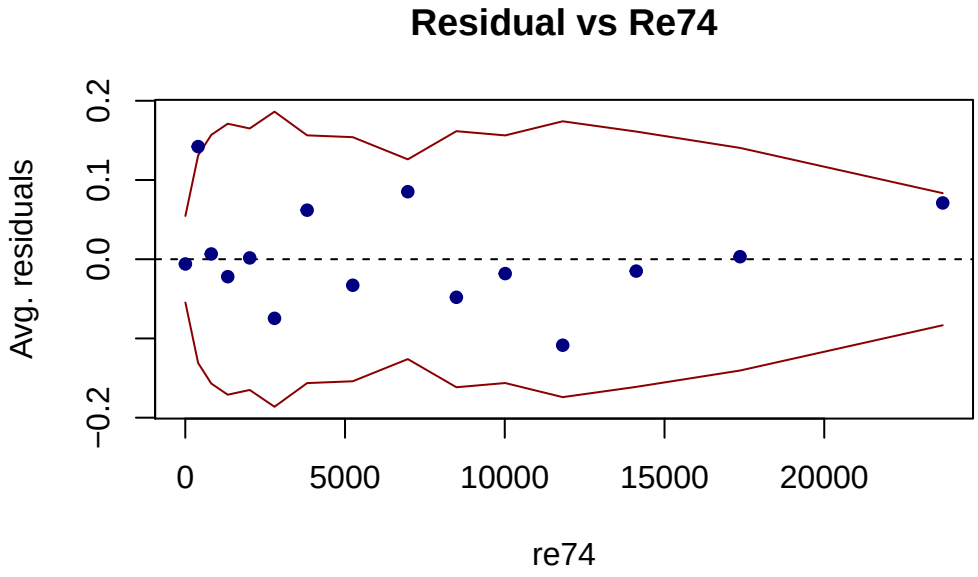

```
binnedplot(x=lab$age_c,y=rawresid1,xlab="age centered",
           col.int="red4",ylab="Avg. residuals",main="Residual vs Age",col.pts="navy")
```

Table 20: Confusion Matrix

	0	1
0	14	5
1	129	466




```
binnedplot(x=lab$re74,y=rawresid1,xlab="re74",
           col.int="red4",ylab="Avg. residuals",main="Residual vs Re74",col.pts="navy")
```



Using 0.5 threshold:

```
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(labreg) >= 0.5, "1","0")),
                            as.factor(lab$employed78),positive = "1")
kable(Conf_mat$table, caption='Confusion Matrix')
```

Table 21: Confusion Matrix

	0	1
0	86	185
1	57	286

```
kable(Conf_mat$overall["Accuracy"])
```

	x
Accuracy	0.781759

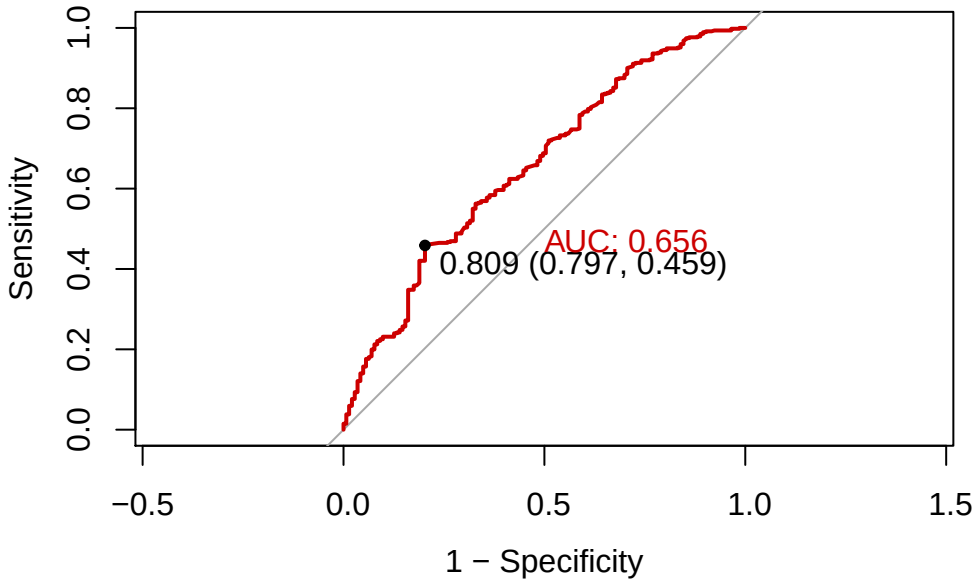
```
kable(Conf_mat$byClass[c("Sensitivity","Specificity")])
```

	x
Sensitivity	0.9893843
Specificity	0.0979021

Change to mean threshold:

```
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(labreg) >= mean(lab$employed78_n), "1","0")),
                            as.factor(lab$employed78_n),positive = "1")

invisible(roc(lab$employed78_n,fitted(labreg),plot=T,print.thres="best",legacy.axes=T,
    print.auc =T,col="red3"))
```




```
kable(Conf_mat$table, caption='Confusion Matrix')
```

```
kable(Conf_mat$overall["Accuracy"])
```

	x
Accuracy	0.6058632

```
kable(Conf_mat$byClass[c("Sensitivity","Specificity")])
```

	x
Sensitivity	0.6072187
Specificity	0.6013986

Model Selection

A full model with all interactions

```
labreg1 <- glm(employed78 ~ age_c + treat + educ+ re74 + black + hispan + married + nodegree
      + age_c:treat + treat:educ + treat:black + treat:hispan + treat:married + treat:nodegree +treat:re
summary = tidy(labreg1)
kable(summary)
```

term	estimate	std.error	statistic	p.value
(Intercept)	0.9128952	0.6405072	1.4252693	0.1540794
age_c	-0.0526200	0.0124605	-4.2229515	0.0000241
treat1	16.8311163	896.9062628	0.0187657	0.9850280
educmiddle school	-0.7600775	0.5363661	-1.4170873	0.1564574
educhigh school	-0.0113887	0.5417260	-0.0210230	0.9832273
educcollege and beyond	0.0375375	0.6911360	0.0543127	0.9566860
re74	0.0000908	0.0000237	3.8280610	0.0001292
black1	-0.4502408	0.3065970	-1.4685101	0.1419657
hispan1	0.0374476	0.3753438	0.0997688	0.9205279
married1	-0.0492965	0.2806885	-0.1756270	0.8605870
nodegree1	0.3692817	0.3318184	1.1129029	0.2657501
age_c:treat1	0.0526480	0.0300251	1.7534643	0.0795224
treat1:educmiddle school	-15.7306684	896.9059000	-0.0175388	0.9860068
treat1:educhigh school	-15.5196304	896.9058122	-0.0173035	0.9861945
treat1:educcollege and beyond	-15.4437756	896.9062200	-0.0172189	0.9862620
treat1:black1	-0.7118601	0.8421086	-0.8453305	0.3979263
treat1:hispan1	14.5091790	703.2625042	0.0206312	0.9835398
treat1:married1	0.7923568	0.6328474	1.2520504	0.2105515
treat1:nodegree1	-0.5659468	0.5819949	-0.9724258	0.3308388
treat1:re74	-0.0000935	0.0000464	-2.0172964	0.0436646

```
n <- nrow(lab)
null_model <- glm(employed78~treat,data=lab,family=binomial)
```

Using stepwise BIC:

```
BIC_stepwise_model <- step(null_model,scope=formula(labreg1),direction="both",
     trace=0,k = log(n))
BIC_stepwise_model$call
```

```
glm(formula = employed78 ~ age_c + re74, family = binomial, data = lab)
```

Using forward BIC:

```
BIC_forward_model <- step(null_model,scope=formula(labreg1),direction="forward",
                          trace=0,k = log(n))
BIC_forward_model$call
```

```
glm(formula = employed78 ~ treat + age_c + re74 + treat:age_c,
    family = binomial, data = lab)
```

Using backward BIC:

```
BIC_backward_model <- step(labreg1,scope=formula(labreg1),direction="backward",
                         trace=0,k = log(n))
BIC_backward_model$call
```

```
glm(formula = employed78 ~ age_c + treat + re74 + age_c:treat,
    family = binomial, data = lab)
```

Test of treat:age is significant:

```
treatagemodel=glm(formula = employed78 ~ treat:age_c + age_c + re74,
                  family = binomial, data = lab)
kable(anova(BIC_stepwise_model, treatagemodel, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
611	641.9121	NA	NA	NA
610	634.8502	1	7.06189	0.0078741

Test of treat is significant:

```
treatmodel=glm(formula = employed78 ~ treat + age_c + re74,
                  family = binomial, data = lab)
kable(anova(BIC_stepwise_model, treatmodel, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
611	641.9121	NA	NA	NA
610	641.8409	1	0.0711982	0.7896002

Test if black is significant:

```
blackmodel=glm(formula = employed78 ~ treat + age_c + re74 + black + treat:age_c,
                  family = binomial, data = lab)
kable(anova(BIC_forward_model, blackmodel, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
609	634.4948	NA	NA	NA
608	628.2804	1	6.21445	0.0126712

Test if hispan is significant:

```
hispanmodel=glm(formula = employed78 ~ treat + age_c + re74 + hispan + treat:age_c,
                  family = binomial, data = lab)
kable(anova(BIC_forward_model, hispanmodel, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
609	634.4948	NA	NA	NA
608	632.7500	1	1.744801	0.1865317

Test if married is significant:

```
marriedmodel=glm(formula = employed78 ~ treat + age_c + re74 + married + treat:age_c,
                  family = binomial, data = lab)
kable(anova(BIC_forward_model, marriedmodel, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
609	634.4948	NA	NA	NA
608	634.0568	1	0.4380025	0.5080881

Test if nodegree is significant:

```
nodegreemodel=glm(formula = employed78 ~ treat + age_c + re74 + nodegree + treat:age_c,
                    family = binomial, data = lab)
kable(anova(BIC_forward_model, nodegreemodel, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
609	634.4948	NA	NA	NA
608	634.4863	1	0.0085568	0.9262986

Add black to the model:

```
labreg2 = glm(formula = employed78 ~ treat + age_c + re74 + black + treat*age_c,
              family = binomial, data = lab)
```

Test if any other interaction is significant:

treat:educ

```
inter1model = glm(employed78 ~ treat + age_c + black+ re74 + treat*age_c  + treat*educ ,
                  family = binomial,
                  data = lab)
kable(anova(inter1model, labreg2, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
602	616.4396	NA	NA	NA
608	628.2804	-6	-11.84076	0.0656169

treat:black

```
inter2model = glm(employed78 ~ treat + age_c + black+ re74 + treat*age_c+ treat*black ,
                  family = binomial, data = lab)
kable(anova(inter2model, labreg2, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
607	625.1877	NA	NA	NA
608	628.2804	-1	-3.092707	0.0786439

treat:hispan

```
inter3model = glm(employed78 ~ treat + age_c + black+ re74 +treat*age_c+ treat*hispan ,
                  family = binomial, data = lab)
kable(anova(inter3model, labreg2, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
606	624.1020	NA	NA	NA
608	628.2804	-2	-4.178379	0.1237874

treat:married

```
inter4model = glm(employed78 ~ treat + age_c + black+ re74 + treat*age_c+ treat*married ,
                  family = binomial, data = lab)
kable(anova(inter4model, labreg2, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
606	626.7840	NA	NA	NA
608	628.2804	-2	-1.496391	0.4732198

treat:nodegree

```
inter5model = glm(employed78 ~ treat + age_c + black+ re74 + treat*age_c + treat*nodegree ,
                  family = binomial, data = lab)
kable(anova(inter5model, labreg2, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
606	628.0777	NA	NA	NA
608	628.2804	-2	-0.2026749	0.9036281

treat:re74

```
inter6model = glm(employed78 ~ treat + age_c + black+ re74 + treat*age_c + treat*re74 ,
                  family = binomial, data = lab)
kable(anova(inter6model, labreg2, test= "Chisq"))
```

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
607	626.0325	NA	NA	NA
608	628.2804	-1	-2.247927	0.1337935

Final Model

Model Summary

Table 22: Final Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	1.0840	0.1662	6.5216	0.0000
treat1	0.5027	0.2693	1.8666	0.0620
age_c	-0.0537	0.0111	-4.8445	0.0000
re74	0.0001	0.0000	3.6224	0.0003
black1	-0.6217	0.2493	-2.4938	0.0126
treat1:age_c	0.0734	0.0272	2.7004	0.0069

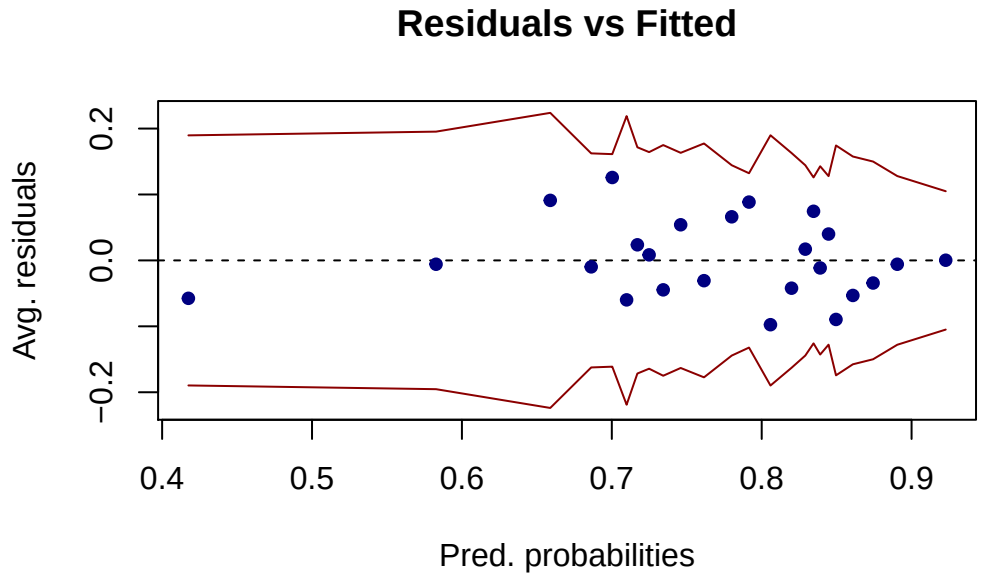
Confidence Interval

Table 23: Final Model Confidence Interval

	2.5 %	97.5 %
(Intercept)	0.7636	1.4162
treat1	-0.0204	1.0374
age_c	-0.0758	-0.0322
re74	0.0000	0.0001
black1	-1.1122	-0.1332
treat1:age_c	0.0215	0.1287

Model Validation

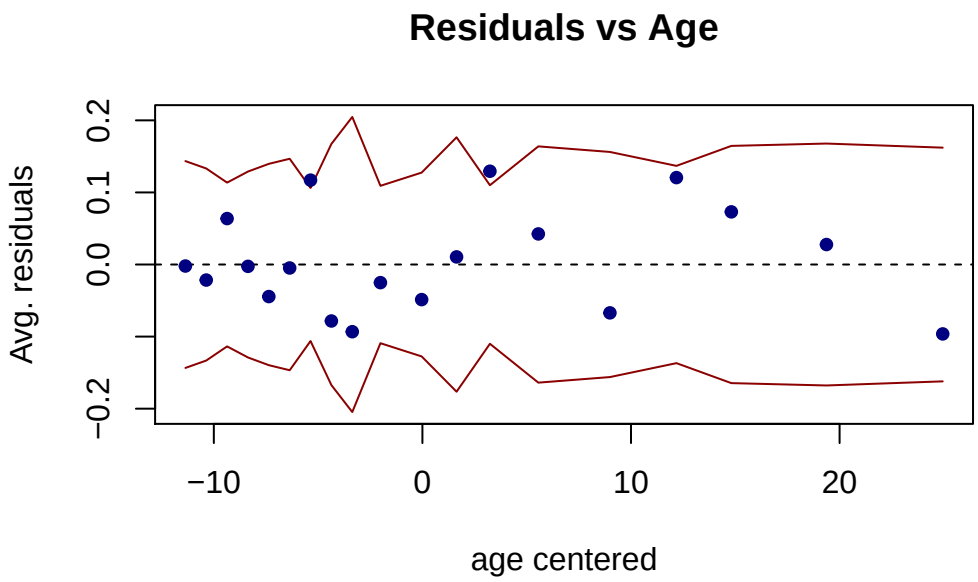
```
rawresid5 <- residuals(labreg2,"resp")
binnedplot(x=fitted(labreg2),y=rawresid5,xlab="Pred. probabilities",
           col.int="red4",ylab="Avg. residuals",main="Residuals vs Fitted",col.pts="navy")
```



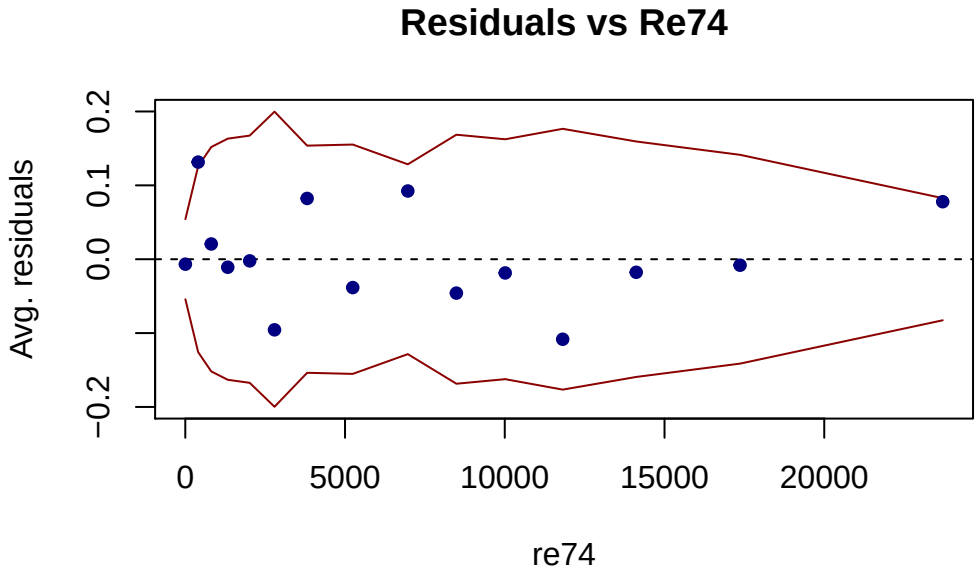

```
binnedplot(x=lab$age_c,y=rawresid5,xlab="age centered",
           col.int="red4",ylab="Avg. residuals",main="Residuals vs Age",col.pts="navy")
```

Table 24: Confusion Matrix

	0	1
0	87	187
1	56	284




```
binnedplot(x=lab$re74,y=rawresid5,xlab="re74",
           col.int="red4",ylab="Avg. residuals",main="Residuals vs Re74",col.pts="navy")
```



Model Assessment

Model performance:

```
Conf_mat <- confusionMatrix(as.factor(ifelse(fitted(labreg2) >= mean(lab$employed78_n), "1","0")),
                            as.factor(lab$employed78_n),positive = "1")
kable(Conf_mat$table,caption = 'Confusion Matrix')
```

Table 25: VIF

	x
treat1	1.630625
age_c	1.382674
re74	1.237373
black1	1.581915
treat1:age_c	1.269887

```
kable(Conf_mat$overall["Accuracy"])
```

	x
Accuracy	0.6042345

```
kable(Conf_mat$byClass[c("Sensitivity","Specificity")])
```

	x
Sensitivity	0.6029724
Specificity	0.6083916

Check for multicolinearity:

```
kable(vif(labreg2),caption='VIF')
```

ROC: