Team Project 1 - Linear Regression

Yuan Feng [yf115], Sebastián Soriano Pérez [ss1072], Vishaal Venkatesh [vv58], Abhiraj Vinnakota [agv9], Roderick Whang [rjw34]

10/2/2019

Summary

In this analysis, we've used a linear regression model having multiple predictors on a subset of the data used in the National Supported Work (NSW) Demonstration, to see if there is evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training.

We found out that their would be a mean increase of 5.287 in the value of the square rooted worker's salary post training as compared to those who did not received training. However, the likely range of increase was inbetween -2.404723 and 12.97928 which led us to conclude that there is no significant evidence to say that workers who received job training tend to earn higher wages than those who did not.

Introduction

The effect on-the-job training had on wages was an area of keen interest amongst social scientists in the 1970s. This was a time when the stock markets were booming and the National Supported Work (NSW) Demonstration in the United States wanted to see if the recent economic prosperity could be maintained byproviding citizens with on-the-job training to keep them competetive in the job market. Between '75 and '77, a collection of random men were selected and a random subset of these men were provided with on-the-job training. Their salary in '74 (before training) was recorded along with certain other details such as age, years of education, whether they had completed high school, marital status and their racial makeup. The annual wages of these gentlemen were continued to be recorded in the years '75 and '78.

In this analysis, we use a linear regression model with multiple predictors to answer the following questions:

a) Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training? Quantify the effect of the treatment, that is, receiving job training, on real annual earnings.

b) What is a likely range for the effect of training?

c) Is there any evidence that the effects differ by demographic groups?

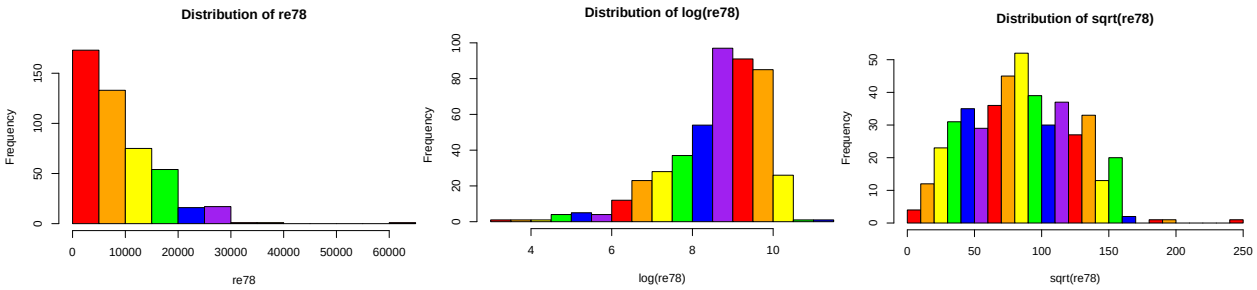
d) Are there other interesting associations with wages that are worth mentioning?

EDA

The data provided in the file "lalondedata.txt" has been used for this analysis. The data has 614 observations across a total of 10 variables in the dataset to start with.

The EDA of the variables is as follows: [For all the EDA plots refer to Appendix 1.1]

1. **re78**: When looking at the distribution of the response variable, we noticed a lot of zero values in the data. These values correspond to people who did not have a job and hence have a zero salary. As it does not make a lot of sense for us to try to predict an increase in salaries for people who do not even have jobs, we decided to remove this data.



On removing, we tried a couple of transformations, namely, log and sqrt, as the data of *re78* wasn't normally distributed. The log transformation though more normal than the original plot, was still skewed. The square root transformation worked better and gave us rather normal plot. Hence we decided to go with this transformation for the model.

2. **treat**: This is our predictor variable of choice. The mean of the response variable seems to be unaffected with the 'treatment' of training. However, the distribution of salaries is slightly more spread for people who haven't been trained. We believe the variable won't be significant in the model.

3. **age**: There seems to be a positive association of age with the response variable. This seems reasonable as the more senior you are, the more you tend to earn. This may be significant in the model.

4. **educ**: There seems to be a good positive association of *educ* with the response variable as well. We believe it's reasonable to believe that the more educated people tend to earn higher salaries. This may be a significant predictor as well.

5. **black**: Blacks seem to be earning a slightly lower salary compared to those who are not. The difference seems marginal though. It'll be interesting to see if the variable is significant.

6. **hispan** : There seems to not be any difference between the mean salary of those who are hispan and those who are not. We don't think this variable will be a significant predictor in the model.

7. **married**: Those who are married seem to be earning more than those who aren't. It's quite surprising that the marital status makes a difference in how much a person earns. It may be that married people are usually older and hence more senior and hence they are paid more. It will be interesting in checking for multicollinearity between age and married to see if our hypothesis is true. This variable may turn out to be significant in the model.

8. **nodegree**: It seems quite reasonable that people who did not graduate high school earn less on average as compared to those who did. We are surprised that the difference isn't more. No sign of the 'sheepskin' effect in our opinion.

9. **re74**: The *re74* seems to have a pretty good positive association with the response variable. It's natural to observe that people who earn higher salaries earlier also earn high salaries post the training.

10. **re75**: The *re75* variable has been excluded because it's unclear as to what it means. Some people might have not taken the training by 1975 or did. It's also unclear as to whether people were getting paid during training or not. Hence, the variable has been excluded.

Model selection

In order to obtain a final model for the response variable *sqrt(re78)* various methods for model selection were tested and interactions between predictors were considered as part of the full model. We used a forward, a backward and a stepwise approach in R to compare the best models each method found using AIC as the selection criteria. The model with the highest adjusted R-squared was the one obtained from the stepwise selection and it considered the variables *re74*, *educ*, *age*, *married*, *re74:age*, and *re74:married* with an adj. R-squared of 0.1658.

The multicollinearity of the model was analyzed with the VIF values of its predictor variables and the variables *re74* and *re74:age* had VIFs greater than 10.

For this reason, all of the numeric variables were centered and the multicollinearity problem was resolved.

Table 1: VIF Values for Both Models

	VIF without Centering	VIF for Centered Data
re74	14.562774	4.491449
educ	1.071224	1.071224
age	2.172280	1.392833
married	2.187221	1.432092
re74:age	16.424042	1.435037
re74:married	5.963144	4.075011

This model with centered variables was compared to a final model that added the variable *treat*, as it is the main interest of this model to find out if the training had any effect on the wages values. The new model has a slightly improved R-squared of 0.1672. An F-test to compare both models was executed with a p-value of 0.1786, thus it cannot be said the models are statistically different when comparing their RSS values.

Table 2: F-Test Results

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
464	605321.6	NA	NA	NA	NA
463	602957.9	1	2363.676	1.815022	0.1785659

Ultimately, we chose the model with *treat* as the final model. The final model is as follows:

$$\widehat{\sqrt{re78}}_i = \hat{\beta}_0 + \hat{\beta}_1 re74\_c_i + \hat{\beta}_2 educ\_c_i + \hat{\beta}_3 age\_c_i + \hat{\beta}_4 treat_i + \hat{\beta}_5 married_i + \hat{\beta}_6 re74\_c_i age\_c_i + \hat{\beta}_7 re74\_c_i married_i$$

Table 3: Coefficients and 95% Confidence Intervals

Coefficient	Estimate	Std. Error	t-Value	p-Value	2.5%	97.5%
$\hat{\beta}_0$	7.579×10^{1}	3.125	24.252	$< 2 \times 10^{-16}$	6.966076×10^{1}	8.190998×10^{1}
$\hat{\beta}_1$	-3.614×10^{-4}	5.230×10^{-4}	-0.691	0.489886	-1.386462×10^{-3}	6.636396×10^{-4}
$\hat{\beta}_2$	2.254	6.766×10^{-1}	3.332	0.000933	9.279744×10^{-1}	3.580003
$\hat{\beta}_3$	3.797×10^{-1}	2.172×10^{-1}	1.748	0.081181	-4.611471×10^{-2}	8.054472×10^{-1}
$\hat{\beta}_4$	5.287	3.925	1.347	0.178566	-2.404723	1.297928×10^{1}
$\hat{\beta}_5$	7.300	4.143	1.762	0.078719	-8.199235×10^{-1}	1.542034×10^{1}
$\hat{\beta}_6$	5.409×10^{-5}	2.724×10^{-5}	1.986	0.047650	7.024840×10^{-7}	1.074841×10^{-4}
$\hat{\beta}_7$	2.266×10^{-3}	6.175×10^{-4}	3.669	0.000272	1.055432×10^{-3}	3.476004×10^{-3}

We computed the VIF values of the predictors once again to make sure there are no colinearity issues in this model. No values greater than 5 were found so there is no concern of multicollinearity in the final model.

Table 4: VIFs for Final Model

	VIF
re74_c	4.562565
educ_c	1.071288
age_c	1.431420
treat	1.163622
married	1.514695
re74_c:age_c	1.473789
re74_c:married	4.092851

To verify the validity of the model, the assumptions of linearity, independence of errors, equal variance, and normality were analyzed with the plots that can be found in Appendix 1.2.
Note that the residual plots against each predictor seem to meet the linearity assumptions as the data points show no other patterns we should be worried about. The Residuals vs. Fitted values plot shows no discrenable pattern indicating independence of errors, however, shows marginal signs of heteroskedasticity as there is lesser variance in higher values. Finally, the Normal Q-Q plot indicates that the normality assumption is mostly met, expect for a little bit of skewness at both tails.

To interpret the coefficients in the scale of *re78* instead of *re78_sqrt*, that is undoing the square root transformation, we could calculate the partial derivatives of *re78* with respect to each one of the predictors in the following way:

$$\sqrt{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k \Longrightarrow y = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k)^2$$

$$\frac{\partial y}{\partial x_j} = 2\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\right) \frac{\partial\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k\right)}{\partial x_j} = 2\sqrt{y}\frac{\partial\sqrt{y}}{\partial x_j}$$

The partial derivatives of the form $\frac{\partial y}{\partial x_j}$ indicate what the rate of change in the response variable is with respect to the change on each predictor x_j, given any combination of values for the rest of the predictors. Our linear model for *re78_sqrt* becomes a polinomial model (2nd degree) when it is transformed for *re78*. The partial derivatives indicate the slopes of the curve of this new model with respect to each predictor (which would help us interpret the model in a similar way as the coefficients in a linear model do). As an example, the partial derivative for the first predictor is as follows:

$$\frac{\partial(re78)}{\partial(re74\_c)} = 2\sqrt{re78}\left(\beta_1 + \beta_6 age\_c + \beta_7 married\right)$$

Note that the interaction terms from the transformed model are taken into account by the partial derivatives with respect to the associated variables. To have a better understanding of this, we can compute the partial derivatives at the baseline values (where every predictor is equal to zero), and estimate how much a change in one unit (in any direction) of each predictor would impact the value of *re78*:

Table 5: re_78 Slopes at Baseline

Variable	re_78 Slope at Baseline
(intercept)	5744.1241*
re74_c	-0.0548
educ_c	341.6613
age_c	57.5549
treat	801.4035
married	1106.534

The intercept value indicates that a person with mean *re74*, mean *educ*, mean *age*, and *treat* and *married* equal to zero would recieve a salary of 5744.12 (*re78*). The rest of the solpe values indicate the rate of change at the baseline for a one unit change in the value of each variable. For instance: a person with baseline values on every predictor would see an average change of 57.55 in *re78* for a one unit change in *year* (if he was one year older he would earn 57.70 more for his wage in 1978, if he was one year younger he would earn 57.41 less). We could interpret the rest of the slopes at the baseline in a similar way. Notice that these values are not constant and they change as the predictors' values change too.

Results

The following are the results we obtained from the model for the questions of interest:

a) Is there evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training? Quantify the effect of the treatment, that is, receiving job training, on real annual earnings.

There is no significant evidence that workers who receive job training tend to earn higher wages than workers who do not receive job training since the coefficient of treat is not significant. With a p-value of 0.178566, we fail to reject the null hypothesis that the coefficient is equal to zero with a significance level of 0.05. However, according to our model the value of *re78_sqrt* is likely to increase by 5.287 on average when the worker receives training compared to those who do not. In terms of real annual earnings (*re78*), a baseline worker would see an increase of 829.36 in annual earnings if he received job training (with a positive one unit change in *treat* starting at the baseline, although the baseline *treat* slope which measures the change rate in both directions is 801.40 as we discussed before). We cannot be confident that this increase is not actually 0.

b) What is a likely range for the effect of training?

The coefficient for *treat* is not significant and the likely range of the effect includes 0: (-2.404723, 12.97928). This means we cannot be certain that the effect of training has any impact on the wages and it might as well be zero.

c) Is there any evidence that the effects differ by demographic groups?

There is no evidence that the effects of training differ by demographic group. Our model selection process eliminated the relevant variables. When including the variables *black* and *hispan*, as well as their interactions with treat (*treat:black*, *treat:hispan*), in our final model none of them were found to be significant (with p-values of 0.196086, 0.188362, 0.893265, and 0.296423, respectively). We cannot reject the null hypotheses that their coefficients are equal to zero.

d) Are there other interesting associations with wages that are worth mentioning?

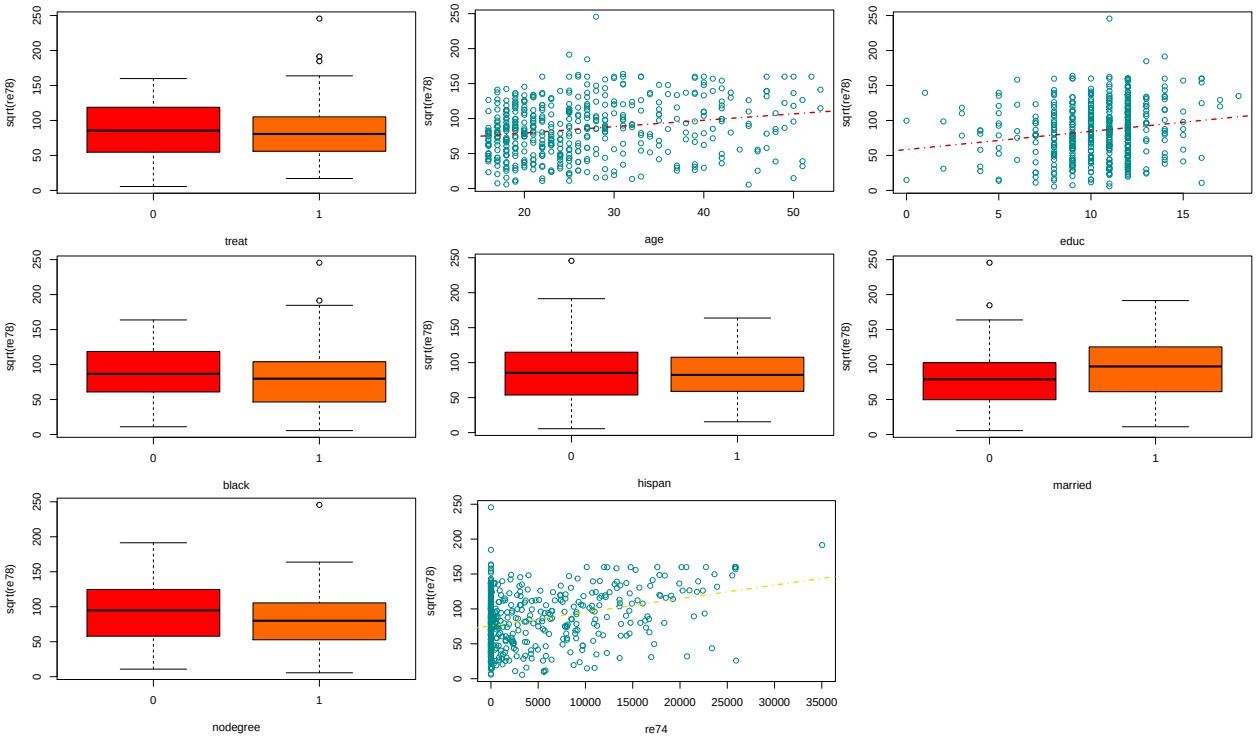
There were only three significant predictors: *educ_c*, *re74_c:agec*, and *re74_c:married* with a 0.05 significance value. A person earning 4951.43 in 1974, with 10.37 years of education, and age 26.7 would on average: earn 0.05 less in 1978 if his salary in 1974 was a unit higher; earn 346.74 more in 1978 if he studied for an extra year; earn 57.70 more in 1978 if he was one year older; earn 1159.82 more if he was married.

Conclusion & Limitations

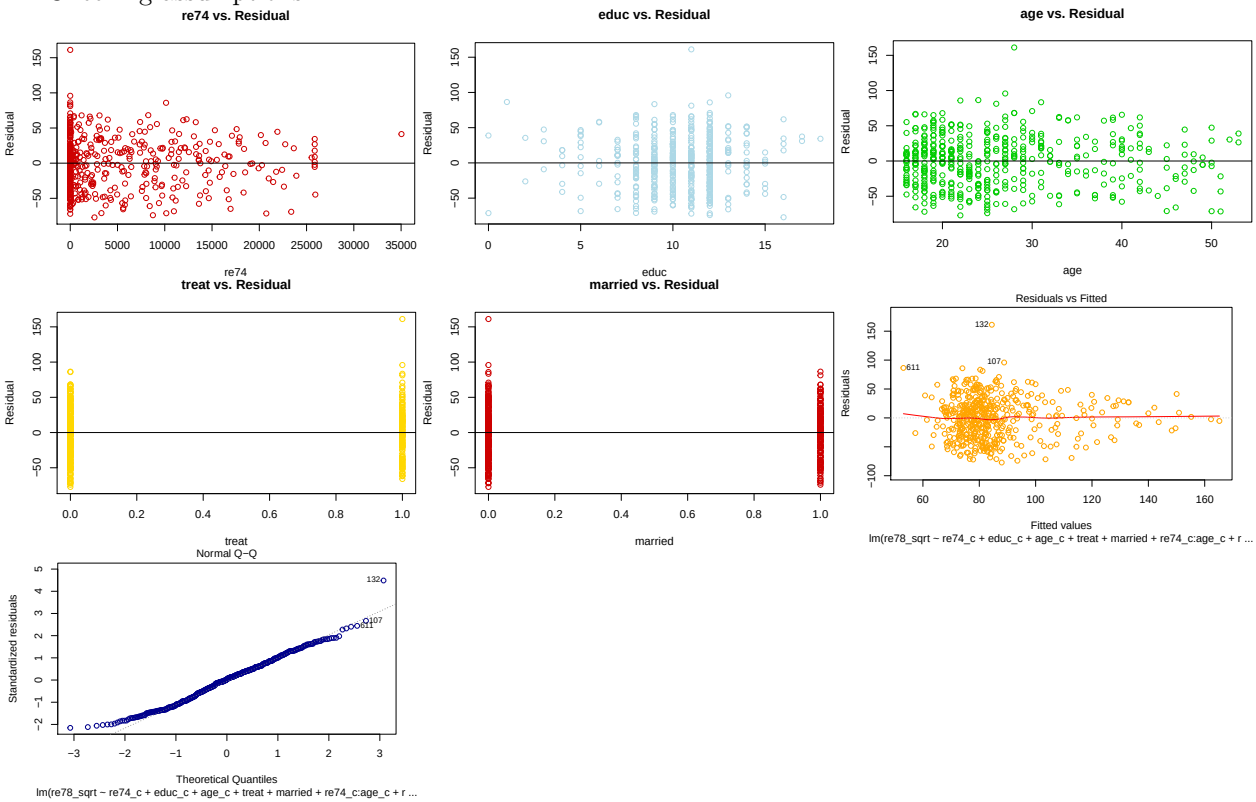
The adjusted R-squared value in the final model is 0.1672, meaning the results of the final model display a relatively low value of R squared, which means the percentage of right predictions is relatively low for this model. If we had access to more data or to other variables that may explain other factors evvefting the salary of workers we maybe able to better the variance in workers salaries.
That being said, we did find some interesting associations and have concluded successfully on the variable of interest, i.e, 'training of workers' with the model we have built.

Appendix

1.1 EDA



1.2 Checking assumptions



1.3 Models before and after centering the data

```r
lm_wages <- lm(formula = re78_sqrt ~ re74 + educ + age + treat + married + re74:age
               + re74:married, data = wages)
summary(lm_wages)
```

```
##
## Call:
## lm(formula = re78_sqrt ~ re74 + educ + age + treat + married +
##     re74:age + re74:married, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.161 -27.678   1.451  23.245 161.124
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.064e+01  9.662e+00   5.242 2.42e-07 ***
## re74         -1.842e-03  9.360e-04  -1.967 0.049726 *
## educ          2.254e+00  6.766e-01   3.332 0.000933 ***
## age           1.331e-01  2.734e-01   0.487 0.626484
## treat         5.287e+00  3.925e+00   1.347 0.178566
## married      -3.026e+00  5.110e+00  -0.592 0.554063
## re74:age      5.409e-05  2.724e-05   1.986 0.047650 *
## re74:married  2.266e-03  6.175e-04   3.669 0.000272 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.09 on 463 degrees of freedom
## Multiple R-squared:  0.1796, Adjusted R-squared:  0.1672
## F-statistic: 14.48 on 7 and 463 DF,  p-value: < 2.2e-16
```

```r
lm_wages <- lm(formula = re78_sqrt ~ re74_c + educ_c + age_c + treat + married
               + re74_c:age_c + re74_c:married, data = wages)
summary(lm_wages)
```

```
##
## Call:
## lm(formula = re78_sqrt ~ re74_c + educ_c + age_c + treat + married +
##     re74_c:age_c + re74_c:married, data = wages)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.161 -27.678   1.451  23.245 161.124
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     7.579e+01  3.125e+00  24.252  < 2e-16 ***
## re74_c         -3.614e-04  5.230e-04  -0.691 0.489886
## educ_c          2.254e+00  6.766e-01   3.332 0.000933 ***
## age_c           3.797e-01  2.172e-01   1.748 0.081181 .
## treat           5.287e+00  3.925e+00   1.347 0.178566
## married         7.300e+00  4.143e+00   1.762 0.078719 .
## re74_c:age_c    5.409e-05  2.724e-05   1.986 0.047650 *
## re74_c:married  2.266e-03  6.175e-04   3.669 0.000272 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 36.09 on 463 degrees of freedom
## Multiple R-squared:  0.1796, Adjusted R-squared:  0.1672
## F-statistic: 14.48 on 7 and 463 DF,  p-value: < 2.2e-16
```

```
confint.default(lm_wages)
```

```
##                        2.5 %       97.5 %
## (Intercept)     6.966076e+01 8.190998e+01
## re74_c         -1.386462e-03 6.636396e-04
## educ_c          9.279744e-01 3.580003e+00
## age_c          -4.611471e-02 8.054472e-01
## treat          -2.404723e+00 1.297928e+01
## married        -8.199235e-01 1.542034e+01
## re74_c:age_c    7.024840e-07 1.074841e-04
## re74_c:married  1.055432e-03 3.476004e-03
```