Studying Difference in Salaries of Males and Females

Shivangi Sinha

24th March 2019

INTRODUCTION

The National Longitudinal Study of Youth data, has data on annual incomes,intelligence test scores and years of education for males and females. Using a subset of this data containing 2584 observations, we will be answering our **Question of Interest - Is there any evidence that the mean salary for males exceeds the mean salary for females with the same years of education and AFQT scores? And by how much?**

Subject - The subject identification number

Gender - The gender of the subject - males or females who were between the ages of 14 and 22

AFQT - This is the percentile score on the AFQT intelligence test measured in 1981

Educ - Years of education completed by 2006

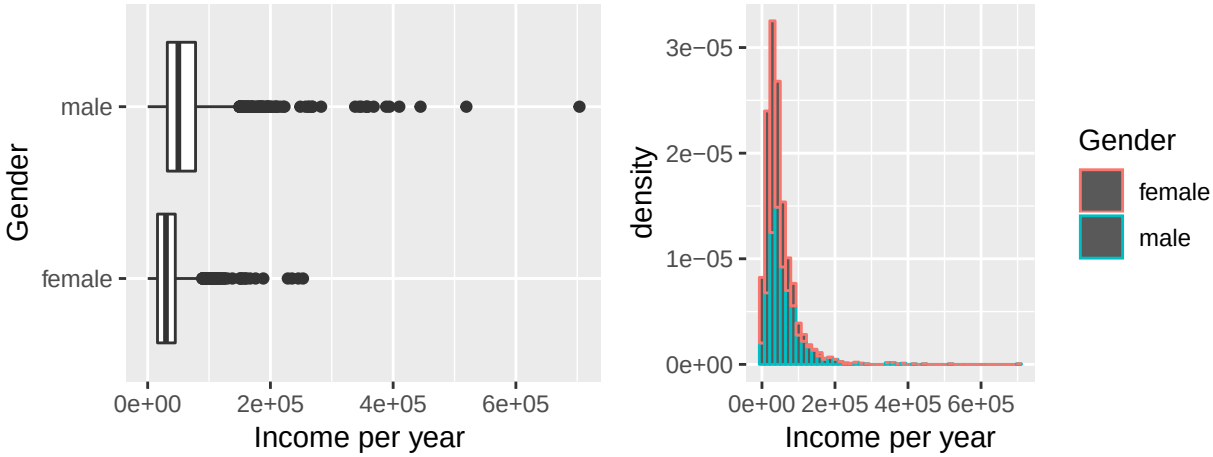
Numerical Summary

The number of observation in our sample is 2584. The variable "Gender" is a categorical variables while "AFQT","Educ" and "Income2005" are the quantitative variables.

Below is the numerical summary of explanatory variables.

	AFQT scores	Years of Education	Income
Mean	54.44	13.89	49417
Minimum	0	6	63
Maximum	100	20	703637

Graphical Summary



From the box plot, we can see that median salary for male is higher than that of female. On an average, male earns more salary than female. From the histogram we can infer that income distribution is right skewed.

METHODS

Assumptions for Inference

- We are assuming that salary is linearly related to years of education,gender and intelligence score.
- All the observations are independent to each other.
- The error term follows normal distribution with mean 0 and constant variance σ^2

In order to answer our question of interest, we will be fitting a regression model with income as response and the variables will be education and intelligence score. Apart from those coninuous variable, we will be using an indicator variable for the gender as our explanatory variable. This variable will take value 1 if the observed salary belongs to male and zero if it belongs to female. In order to for our assumption to be maintained and we are fitting a transformed value of Income. The transformation fucntion is log Here is the model to be fitted :

$$\log Y_{(Income)i} = \beta_0 + \beta_1 1(X_i = Male) + \beta_2 X_{(AFQT)i} + \beta_1 X_{(Educ)i} + \epsilon_i, i = 0, 1...2584$$

The fitted values will be calculated using this formula :

$$\hat{y}_{i(Income)} = \exp \hat{\beta}_0 \exp{(\hat{\beta} 1_{i(X=Male)})} exp(\hat{\beta} X_{i(AFQT)}) exp(\hat{\beta} X_{i(Educ)})$$

* Y_i is the Income but the model fitted is using transformed value of Income.

- X_i's are the explanatory variables corresponding to AFQT scores, years of education and the indicator variable(Taking value of 1 when response is male)
- β's are the parameters that are to be estimated.
- ϵ_i is the error terms

Using this model we will be a conducting F test to check if there is any effect of gender on the income using the reduced model. Null hypothesis: $\beta_1 = 0$ Alternative hypothesis : $\beta_1 \neq 0$ F statistic

$$F = \frac{(RSS_{(\beta_1=0)} - RSS_{fullmodel})/(4-3)}{RSS_{fullmodel}/(2584-4)} \sim F(4-3, 2584-4)$$

- $RSS_{(\beta_1=0)}$ is the residual sum of square of the smaller model when $\beta_1 = 0$
- $RSS_{fullmodel}$ is the residual sum of square of the full model when $\beta_1 \neq 0$

This test should give us the evidence if there is difference between the mean salary when gender is male and female. In order to find how much this difference is we will be using confidence interval of β_1 which is as follows :

$$e^{\hat{\beta}_1 \pm t_{(1-\alpha/2)} \times \sqrt{Var(\hat{\beta}_1)}}$$

RESULTS

After fitting the above model, we derived the following estimates after backtransforming-

	Estimates	Standard Error
Intercept	6192	0.103
Indicator Variable of Gender	1.867	0.034
AFQT	1.006	0.001
Education Years	1.08	0.008

Residual Standard Error	R-Square	Adjusted R-Square
0.8661	0.2101	0.2092

value	numdf	dendf
228.7	3	2580

Conducting a F test between these two models to check in β associated with Gender is zero.

$$Model1: \quad \log_{Y_{(Income)i}} = \beta_0 + \beta_1 1_{(X=Male)i} + \beta_2 X_{(AFQT)i} + \beta_1 X_{(Educ)i} + \epsilon_i, i = 0, 1...2579$$
$$Model2: \quad \log_{Y_{(Income)i}} = \beta_0 + \beta_1 X_{(AFQT)i} + \beta_2 X_{(Educ)i} + \epsilon_i, i = 0, 1...2579$$

Table 5: Analysis of Variance Table

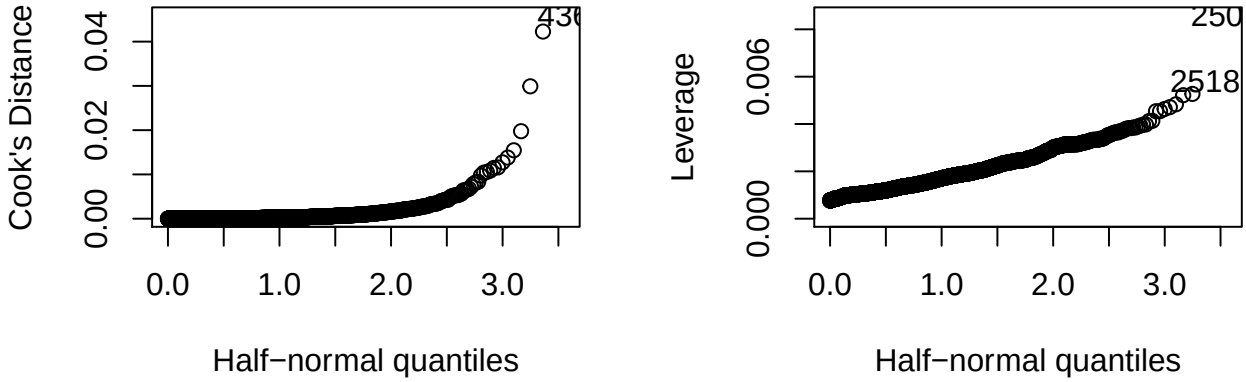
Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
2580	1935	NA	NA	NA	NA
2581	2186	-1	-250.5	333.9	2.986e-70

Since the p value is less than 0.05. We reject the null hypothesis of $\beta_1 = 0$

The confidence interval of β_1 is as follows after back tranforming the exponential value is as follows:

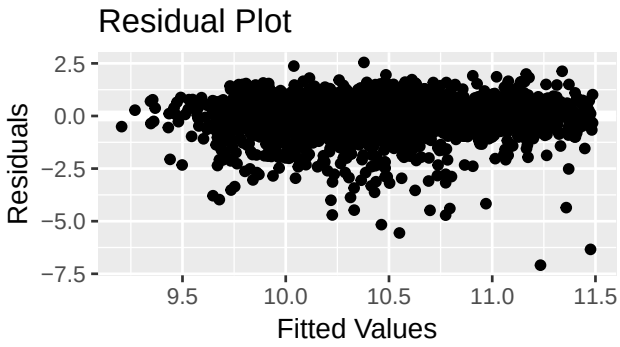
	Confidence Interval
Lower Bound	1.746
Upper Bound	1.997

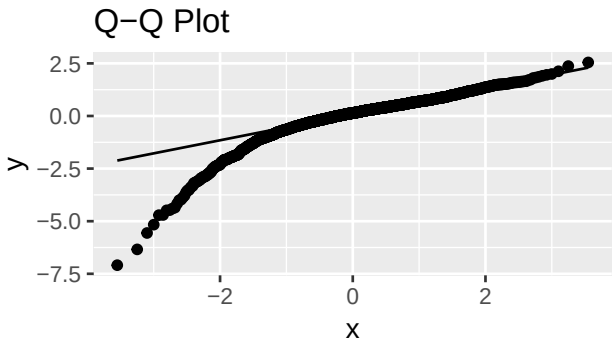
Checking for unusual values using half normal plot



Accoring to this observations pertaining to subject number 11972 and 11937 have high leverage and observations with subject number 1046 and 2980 are influential points.

Checking for Assumptions





Accoring to the above residual plot, we the see that the variance is clustered around the zero and it is contant over the values . The Q-Q plot shows that the the transormed variable of income may only be approximately normal since the starting values fall through but as the number of observations grows, it can be asymtotically normal due to Central Limit Theorum.

CONCLUSIONS

- Estimates : Since there were problems with the assumptions of errors in the data,we used the tranformed model in order to answer the question of interest.Accoring to our regression model estimates, it is estimated that in mean effect of gender with salaries of is 1.8(approx). Thus, we can say that we have the evidence that salary of males exceeds by females by salary multipied by 1.8. With 95 % confidence level, we can the say that this difference in salary lies between 1.7 and 1.9 with same level of intelligence scores and years of education.
- Reduced Model for Regression : Our estimated f statistic is 333.9 which give a small p-value. Thus, we have confidencing evidence that there may be an associated between gender and salaries.
- Rsquare : The adjusted R^2 is 0.2092 which is low for a linear model. Thus, I think we should try to get a bigger sample to get more accurate results.