The Relationship Between Education and Income

By

Yihan Wang 111698969

Xiyi Lin 113070130

Yechen Li 112248017

Introduction

In the previous project, we discussed the relationship between education and income by using the data from the National Longitudinal Surveys of Youth 1997 on the NLS Investigator, which the survey year is 2015. We conclude that both two linear regression models well demonstrate that the highest degree received positively affects the household income, along with other control variables such as sex, age, and region. However, since linear regression assumes a linear relationship between the input and output variables, it fails to fit complex data properly. Actually, in most real-life scenarios, the relationship between the variables of the dataset is not linear, and hence using linear regression is not enough. In the following research, we are going to use the nonlinear regressions, the interaction terms, and the instrumental variable to see whether they help the model to fit our data or not in more general cases.

Data Description

Our data is from the National Longitudinal Surveys of Youth 1997 on the NLS Investigator, which the survey year is 2015. It contains data on 5419 respondents, their income defined as the gross family income in the previous year, the highest degree they have obtained, such as high school, junior college, undergraduate, master, Ph.D., professional degree, etc. The integrated data also reflects the impact of age, gender, and region on household income.

Outline of the Econometric Model(s)

Regression model 1: linear

$$\text{hh\_income} = \beta0 + \beta1 * \text{highest\_degree} + u$$

Regression model 2: linear (more add-ons)

$$\text{hh\_income} = \beta0 + \beta1 * \text{highest\_degree} + \beta2 * \text{male} + \beta3 * \text{age} + \beta4 * \text{Northeast} + \beta5 * \text{North} + \beta6 * \text{South} + u$$

Regression model 3: polynomials (cubic)

$$\text{hh\_income} = \beta0 + \beta1 * \text{highest\_degree} + \beta2 * \text{highest\_degree}\hat{}2 + \beta3 * \text{highest\_degree}\hat{}3 + u$$

Regression model 4: log-transformation (not applicable)

$$\log(\text{hh\_income}) = \beta0 + \beta1 * \text{highest\_degree} + u$$

Regression model 5: interactions between independent variables (between the binary and continuous variables)

$$hh\_income = \beta0 + \beta1*highest\_degree + \beta2*male + \beta3*(highest\_degree \times male) + u$$

Regression model 6: interactions between one binary variable and non-linear transformation

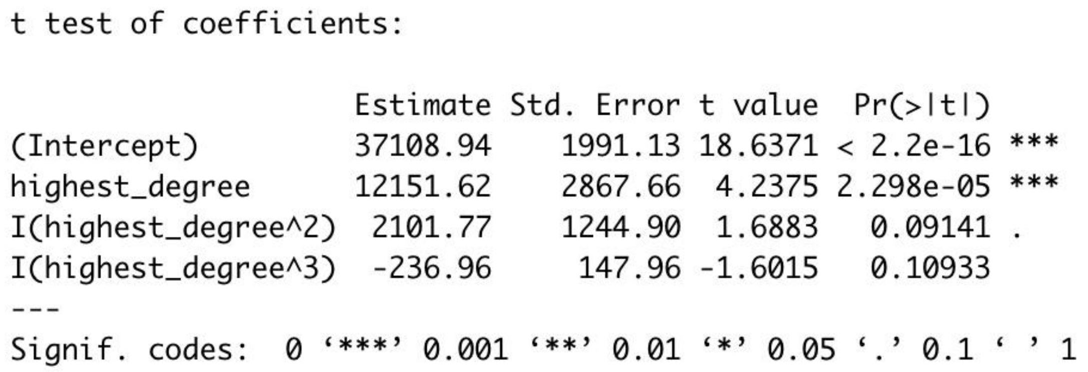
$$hh\_income = \beta0 + \beta1*highest\_degree + \beta2*highest\_degree\text{^}2 + \beta3*highest\_degree\text{^}3 + \beta4*male + \beta5*(highest\_degree \times male) + \beta6*(highest\_degree\text{^}2 \times male) + \beta7*(highest\_degree\text{^}3 \times male) + u$$

Relation with Chapter 6 (non-linear regressions)

In Chapter 6, we no longer limit ourselves to linear assumptions but extend our scope to the non-linear domain, where we study the non-linear transformation by two complementary approaches: polynomials in X and logarithmic transformations. In this project, we keep the original model of a single independent regressor highest_degree 【Model 1】 and then do the cubic specification 【Model 3】 and log-transformation 【Model 4】 to find out the better behaved one. What's more, we also try to use interactions between independent variables in both linear【Model 5】and non-linear models 【Model 6】 to get a better fit for our data.

Analyses of Nonlinear Regression Models

- **Model 3 Table: Cubic non-linear regression** *R-squared = 0.141



```
t test of coefficients:

                     Estimate Std. Error t value  Pr(>|t|)    
(Intercept)          37108.94    1991.13 18.6371 < 2.2e-16 ***
highest_degree       12151.62    2867.66  4.2375 2.298e-05 ***
I(highest_degree^2)   2101.77    1244.90  1.6883   0.09141 .  
I(highest_degree^3)   -236.96     147.96 -1.6015   0.10933    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

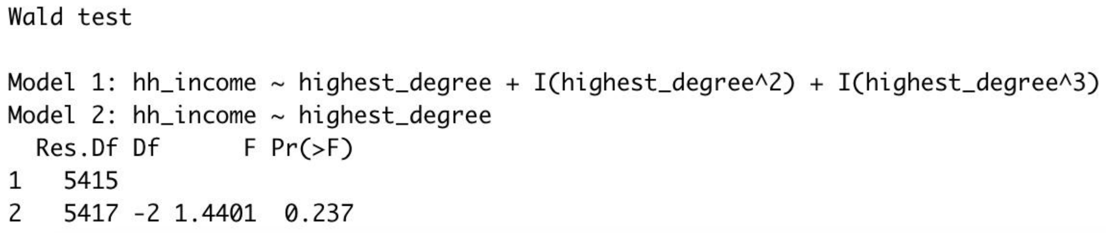
Under Model 3, say, the predicted change in hh_income for a change in highest_degree from 2 to 3, we can calculate the effect by:

Δhh_income hat = 37108.94+12151.62*3+2101.77*3^2-236.96*3^3 -(37108.94+12151.62*2+2101.77*2^2-236.96*2^3) = 18158.23,

which means the difference of hh_income between people who have the high school diploma and those who attend Associate or Junior college is $18158.23. Moreover, by performing the hypothesis testing:

H0: population coefficients on highest_degree^2 and highest_degree^3 = 0. (linearity)

vs. H1: at least one of these coefficients is nonzero. (non-linearity)

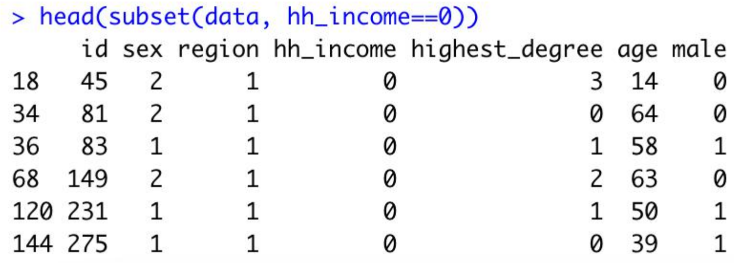


```
Wald test

Model 1: hh_income ~ highest_degree + I(highest_degree^2) + I(highest_degree^3)
Model 2: hh_income ~ highest_degree
  Res.Df Df      F Pr(>F)
1   5415                 
2   5417 -2 1.4401  0.237
```

The |t-value| = 1.6015 < 1.96, we cannot reject H0 at 5% significance level that the powers of highest_degree has no effect on TS. By using F-test to compare both models, p-value=0.237 is also larger than the significance level of 5%, assuring our decision is right.

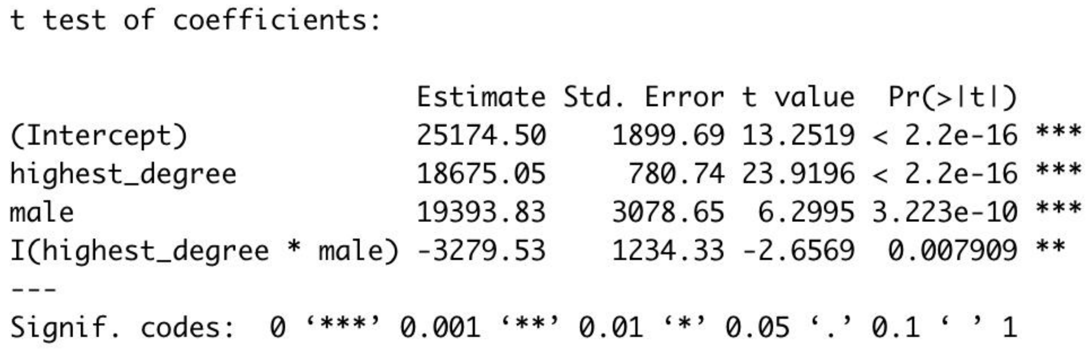
- **Model 4 Table: log-transformation (not applicable)**



```
> head(subset(data, hh_income==0))
     id sex region hh_income highest_degree age male
18   45   2      1         0              3  14    0
34   81   2      1         0              0  64    0
36   83   1      1         0              1  58    1
68  149   2      1         0              2  63    0
120 231   1      1         0              1  50    1
144 275   1      1         0              0  39    1
```

Though it is a good way to do log-transformation alternatively, we cannot operate this method in this project because our hh_income data includes the value of zero (undefined in log), preventing us from doing such transformations.

- **Model 5 Table: interactions between binary and continuous variables** *R-squared = 0.1488



```
t test of coefficients:

                         Estimate Std. Error t value  Pr(>|t|)    
(Intercept)              25174.50    1899.69 13.2519 < 2.2e-16 ***
highest_degree           18675.05     780.74 23.9196 < 2.2e-16 ***
male                     19393.83    3078.65  6.2995 3.223e-10 ***
I(highest_degree * male) -3279.53    1234.33 -2.6569  0.007909 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The other aspect of fitting is to include the “interaction term” highest_degree*male as a regressor. We compare the various cases, where the effect of highest_degree depends on the dummy variable male, Δhh_income/Δhighest_degree = β1+β3*male:

①When male==0, β1 + β3*0 = β1, then

hh_income hat = 25174.50 + 18675.05*highest_degree

②When male==1, β1 + β3*1 = β1 + β3, then hh_income hat = 25174.50 + 18675.05*highest_degree + 19393.83 - 3279.53*highest_degree = 44568.33 + 15395.52*highest_degree

The two regression lines have different intercept and slopes. That is, the highest_degree is estimated to have a larger effect when the gender is female, given 18675.05>15395.52.

We can prove the observation above by doing the hypothesis testing:

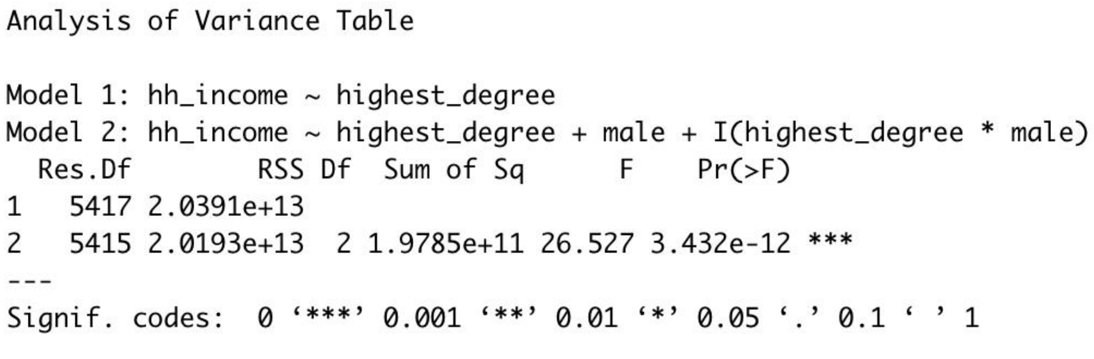
H0: the two regression lines have the same slope i.e. the coefficient on highest_degree*male is 0.

Since |t|=|-3279.53/1234.33|=2.656931>1.96, the coefficient is significantly different from 0. Thus, we reject H0 at a 5% significance level. Similarly, another hypothesis testing for intercepts:

H0: the two regression lines have the same intercept i.e. the coefficient on the male is 0.

Given |t|=|19393.83/3078.65|=6.299459>1.96, the coefficient is also significantly different from 0. Again, we reject H0 at a 5% significance level.

Thereafter, we throwback to compare Model 5 with Model 1:

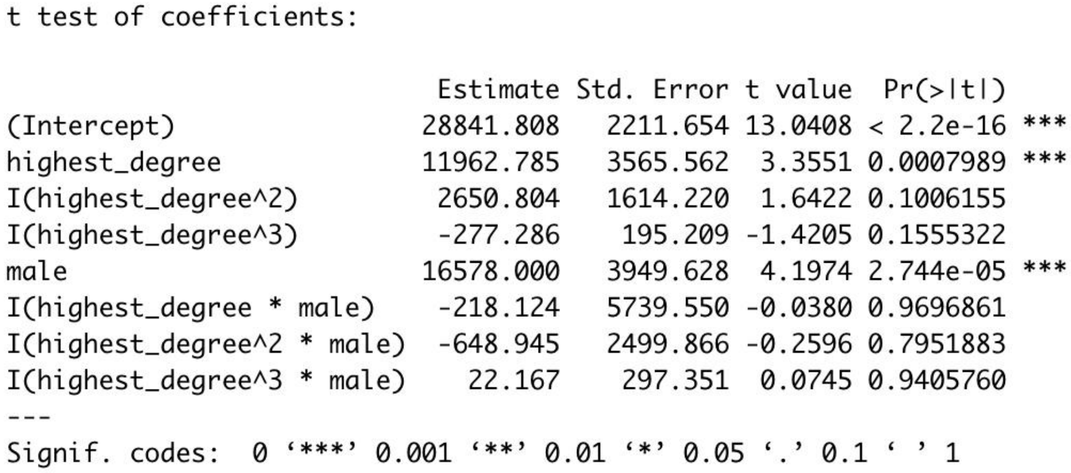


```
Analysis of Variance Table

Model 1: hh_income ~ highest_degree
Model 2: hh_income ~ highest_degree + male + I(highest_degree * male)
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1   5417 2.0391e+13                                   
2   5415 2.0193e+13  2 1.9785e+11 26.527 3.432e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Given F-statistic = 26.527 > 3 (q=2) with p-value = 3.432e-12 < 0.05, we have strong evdience to believe that Model 5 is significantly different from Model 1, and Model 5 has a better fit.

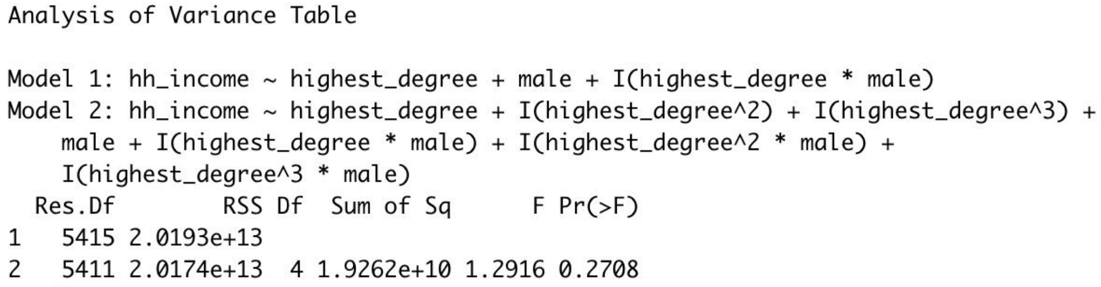
- **Model 6 Table: interactions between one binary and non-linear transformation** *R-squared = 0.149



```
t test of coefficients:

                            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)                28841.808   2211.654 13.0408 < 2.2e-16 ***
highest_degree             11962.785   3565.562  3.3551 0.0007989 ***
I(highest_degree^2)         2650.804   1614.220  1.6422 0.1006155    
I(highest_degree^3)         -277.286    195.209 -1.4205 0.1555322    
male                       16578.000   3949.628  4.1974 2.744e-05 ***
I(highest_degree * male)    -218.124   5739.550 -0.0380 0.9696861    
I(highest_degree^2 * male)  -648.945   2499.866 -0.2596 0.7951883    
I(highest_degree^3 * male)    22.167    297.351  0.0745 0.9405760    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Furthermore, we extend our research interest: are there non-linear interactions? Still, we will use a binary-continuous interaction specification but by adding highest_degree*male, highest_degree^2*male, and highest_degree^3*male while making the regression line non-linear. Compared to what we estimated before, R-squared increases a little bit than that of each model, showing that this regression line explains more of the variance of the dependent variable. To perform tests of joint hypotheses of Model 5 and Model 6 by using ANOVA table:



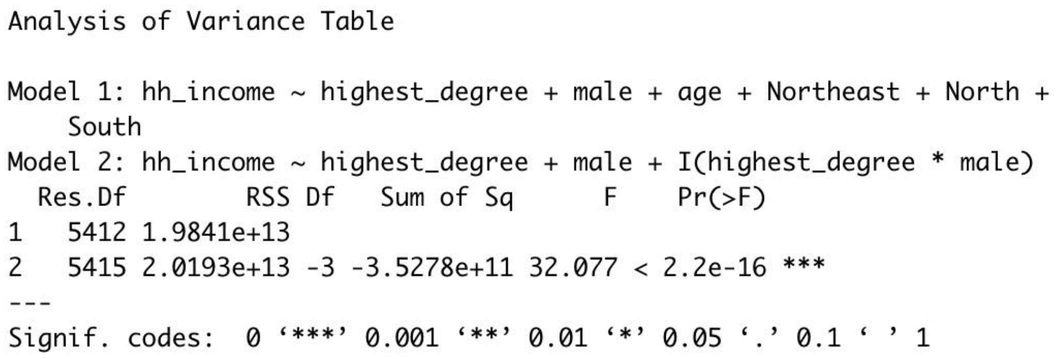
```
Analysis of Variance Table

Model 1: hh_income ~ highest_degree + male + I(highest_degree * male)
Model 2: hh_income ~ highest_degree + I(highest_degree^2) + I(highest_degree^3) + 
    male + I(highest_degree * male) + I(highest_degree^2 * male) + 
    I(highest_degree^3 * male)
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1   5415 2.0193e+13                            
2   5411 2.0174e+13  4 1.9262e+10 1.2916 0.2708
```

As we can see, F-statistic = 1.2916 < 2.37 (q=4) with p-value = 0.2708 >

0.05, then we conclude that Model 5 and Model 6 are not significantly different at a 5% significance level.

Summary of the results



```
Analysis of Variance Table

Model 1: hh_income ~ highest_degree + male + age + Northeast + North +
    South
Model 2: hh_income ~ highest_degree + male + I(highest_degree * male)
  Res.Df        RSS Df   Sum of Sq      F    Pr(>F)
1   5412 1.9841e+13
2   5415 2.0193e+13 -3 -3.5278e+11 32.077 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As shown above, we compare both fittest models in our two projects: the table shows that Model 5 and Model 2 are significantly different based on p-value=2.2e-16 < 0.05. We know the adjusted R-squared increases only if the new term improves the model more than it would be expected by chance, and it can also decrease with poor quality predictors. While the adjusted R-squared=0.1488 from Model 5 is still lower than that of Model 2 from our first project, 0.1632, we strongly believe that Model 2 is the fittest one.

Relation with Chapter 7 (instrumental variables)

In this section, we want to check the correlation coefficients between the residuals and the regressors. During the research, we may generate the omitted variable bias, simultaneous casualty bias, or errors-in-variables bias without consideration, making the least square assumption#1 not holds. But this problem can be solved by introducing instrumental variables regression to eliminate bias.

Endogeneity bias

To isolate the part of endogenous variable highest_degree that is not correlated with error terms so that we can estimate β1, we need to find an exogenous instrumental variable (IV), Z. In our project, it is very common to see that the “grant-in-aid received” has something to do with the highest_degree earned in reality, causing corr(Zi, Xi)≠0. Here comes the method of Two-Stage Least Squares (TSLS).

In the first stage, we want to find such a variable like “grant-in-aid received” as our exogenous IV, and then isolate the part of highest_degree that is uncorrelated with u by regressing highest_degree on IV using OLS and thus obtain the predicted values:

$$\textbf{highest\_degree} = \boldsymbol{\alpha 0 + \alpha 1 * \textbf{grant} + \textbf{vi}}$$

where the IV satisfies two conditions corr(Zi, ui)=0 and corr(Zi, Xi)≠0.

In the second stage, we plug the predicted values highest_degree_hat instead of highest_degree into Model 2 to perform the OLS regression to get the β1 hat TSLS.

Conclusion

Generally, in this project, we continue studying the relationship between income and education by using NSLY97 from 2015. After trying different regressions, our Model 2 with more regressors yields a better explanation than all new models in our second project. However, it is still a meaningful attempt throughout the whole process. Our discussion on IV helps us to conduct the potential research deeper in the future as well.