Chirag Singhvi

PY538 – Final Presentation write-up

Topic: Multiple Linear Regression

The aim of this presentation was to illustrate the technique of multiple linear regression, its uses and some of the critiques of this statistical modelling technique.

What is multiple linear regression?

It is an attempt to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to the observed data. The equation is of the form: $i = c + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 \,{}_{....}+ \varepsilon$ where i is the response variable and all the D_n are the explanatory variables. The β_n are the coefficients of regression or the parameters of the regression. The c term is the intercept on the response variable axis and is a constant. The ε is the residual term that represents the deviations of the observed values from their means which are normally distributed with mean 0 and variance ς.

Most regression packages use the least squares method to obtain the best fitting line. It is calculated by minimizing the sum of squares of the vertical deviations from each data point to the line. The sum of the residuals themselves must equal zero.

To learn more about the technique, we shall make use of an example:

We will plot the income of **employees with disabilities**.

Notes on the data

- From the Current Population Survey (CPS) in the United States
- 2010-2015
- 1,217,397 observations
- 487,219 observations after ‘cleaning’ the data
  - Data points dropped when subject was:
    - Unemployed
    - Not part of the labor force
    - Had zero income

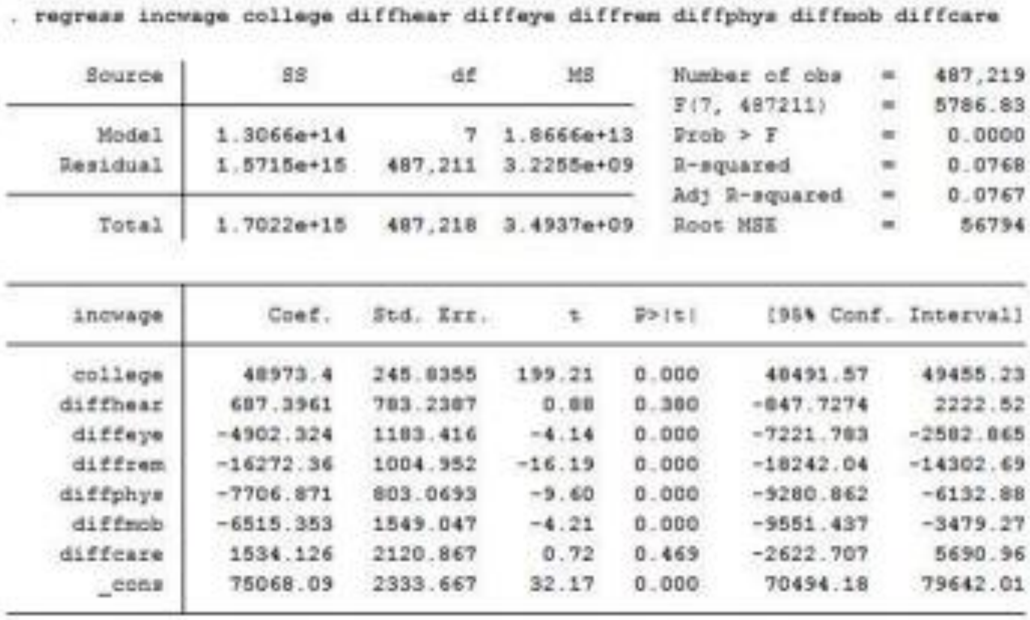
Variables

Variable Name	Variable Description
year	Survey year
educ	Educational attainment code
empstat	Employment Status
labforce	Labor force status
incwage	Wage and salary income
diffhear	Hearing difficulty
diffeye	Vision difficulty
diffrem	Difficulty remembering
diffphys	Physical difficulty
diffmob	Disability limiting mobility
diffcare	Personal care limitation

Regression Output:

Regression 1 –
College graduates

$$i = c + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 \; .....$$



```
. regress incwage college diffhear diffeye diffrem diffphys diffmob diffcare

      Source |       SS           df       MS      Number of obs   =   487,219
-------------+----------------------------------   F(7, 487211)    =   5786.83
       Model |  1.3066e+14         7  1.8666e+13   Prob > F        =    0.0000
    Residual |  1.5715e+15   487,211  3.2255e+09   R-squared       =    0.0768
-------------+----------------------------------   Adj R-squared   =    0.0767
       Total |  1.7022e+15   487,218  3.4937e+09   Root MSE        =     56794

------------------------------------------------------------------------------
     incwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |    48973.4   245.8355   199.21   0.000     48491.57    49455.23
    diffhear |   687.3961   783.2387     0.88   0.380    -847.7274     2222.52
     diffeye |  -4902.324   1183.416    -4.14   0.000    -7221.783   -2582.865
     diffrem |  -16272.36   1004.952   -16.19   0.000    -18242.04   -14302.69
    diffphys |  -7706.871   803.0693    -9.60   0.000    -9280.862    -6132.88
     diffmob |  -6515.353   1549.047    -4.21   0.000    -9551.437    -3479.27
    diffcare |   1534.126   2120.867     0.72   0.469    -2622.707     5690.96
       _cons |   75068.09   2333.667    32.17   0.000     70494.18    79642.01
------------------------------------------------------------------------------
```

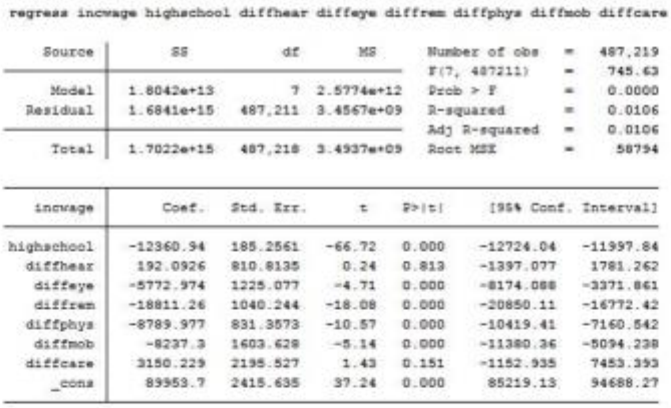
This is the regression output using STATA. Here *i* stands for the income of an individual and the D_n stand for different disabilities. The β_n coefficients can be found in the “Coef.” column. This data implies that one can expect a reduction in income by -4902.324 dollars because of the presence of a vision difficulty (diffeye variable). One might be (rather should be) surprised to see positive coefficients for diffhear and diffcare. Presence of difficulties should not be increasing someone’s income. To explain this, we take a look at the subsequent columns on this data table. The ‘t’ and the ‘p>|t|’ columns are descriptors of hypothesis testing. This technique tests whether the coefficient is significant. It sets the null hypothesis to be that the coefficient is equal to zero. The alternative hypothesis would of course be β =! 0 (not equal to). The t-statistic is the parameter estimate divided by its standard deviation. The p>|t| column checks the probability of the calculate t-statistic value to be greater than a set value for the t(n-p-1) distribution table, where n is the population figure and p the number of variables. If the p value is greater than 0.05 the null hypothesis cannot be rejected. In this case, we can see that for most disabilities, the p value is exactly 0. This entails that these coefficients have a zero chance of being zero. On the other hand, diffhear and diffcare have quite high p values, meaning that the coefficients might as well be insignificant. A look at the confidence interval column (last on the right) suggests that we would find, with 95 percent confidence, that the actual coefficient value will be within the two extremes. We can see that the values for coefficients for diffhear and diffcare can be negative after all as both the minima are negative.

The overall quality of a regression can be signified by the coefficient of determination R^2. This value, which is a fraction between 0 and 1, is the square of the correlation between the predicted income figures and the actual income figures. The value indicates the extent to which the dependent variable is predictable. In this case, the r-squared value of 0.07 is very low, indicating that this model will be an abysmal predictor of an individual's income. This makes sense as presence of disabilities and level of education are far from the only predictors of a person's wages.

This regression was run only for employees with a college degree, a similar one was run for employees with only a high-school degree. The data is given below.

Regression 2 – High
school graduates

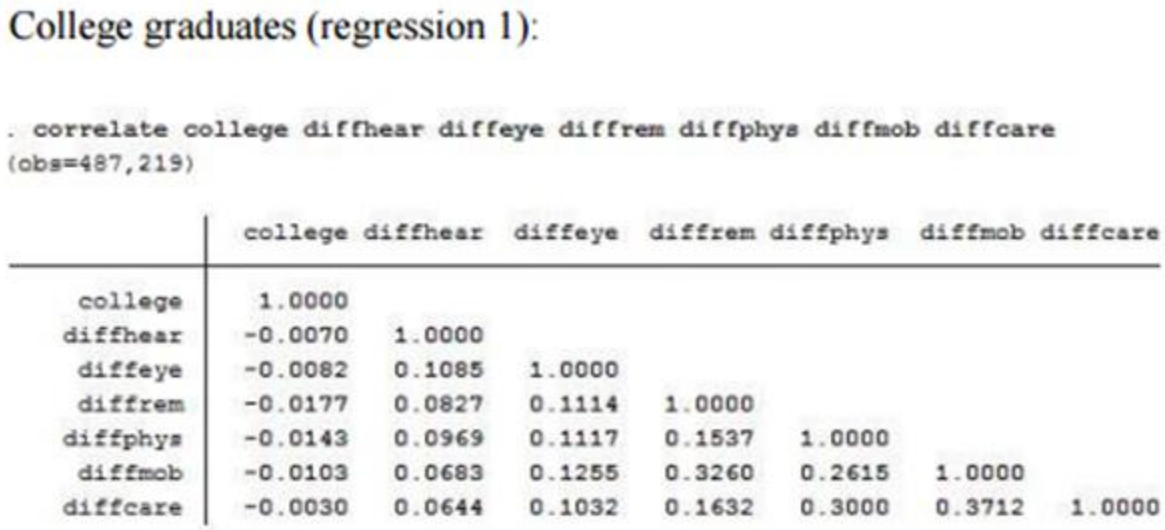
$$i = c + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 \ .....$$



```
regress incwage highschool diffhear diffeye diffrem diffphys diffmob diffcare

      Source |       SS           df       MS      Number of obs   =   487,219
-------------+----------------------------------   F(7, 487211)    =    745.63
       Model |  1.8042e+13         7  2.5774e+12   Prob > F        =    0.0000
    Residual |  1.6841e+15   487,211  3.4567e+09   R-squared       =    0.0106
-------------+----------------------------------   Adj R-squared   =    0.0106
       Total |  1.7022e+15   487,218  3.4937e+09   Root MSE        =     58794

------------------------------------------------------------------------------
     incwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  highschool |  -12360.94   185.2561   -66.72   0.000    -12724.04   -11997.84
    diffhear |   192.0926   810.8135     0.24   0.813    -1397.077    1781.262
     diffeye |  -5772.974   1225.077    -4.71   0.000    -8174.088   -3371.861
     diffrem |  -18811.26   1040.244   -18.08   0.000    -20850.11   -16772.42
    diffphys |  -8789.977   831.3573   -10.57   0.000    -10419.41   -7160.542
     diffmob |    -8237.3   1603.628    -5.14   0.000    -11380.36   -5094.238
    diffcare |   3150.229   2195.527     1.43   0.151    -1152.935    7453.393
       _cons |    89953.7   2415.635    37.24   0.000     85219.13    94688.27
------------------------------------------------------------------------------
```

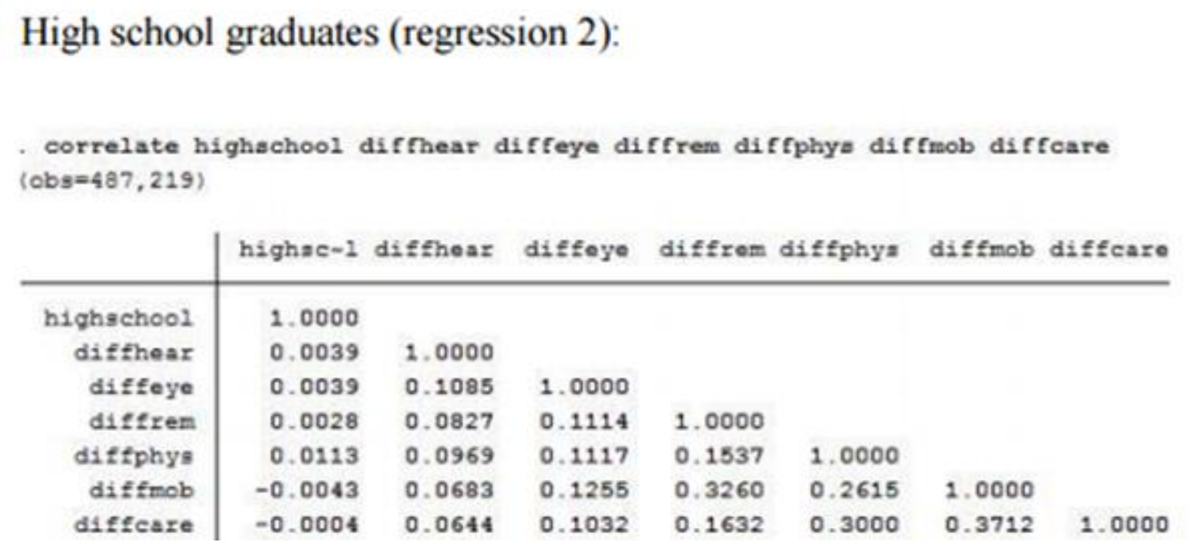
Collinearity:

A good regression must be checked for collinearity between variables. The tables below show that the disability variables are quite unrelated, that is, the presence of one disability has no bearing on the presence of another.



College graduates (regression 1):

```
. correlate college diffhear diffeye diffrem diffphys diffmob diffcare
(obs=487,219)

             |  college diffhear  diffeye  diffrem diffphys  diffmob diffcare
-------------+---------------------------------------------------------------
     college |   1.0000
    diffhear |  -0.0070   1.0000
     diffeye |  -0.0082   0.1085   1.0000
     diffrem |  -0.0177   0.0827   0.1114   1.0000
    diffphys |  -0.0143   0.0969   0.1117   0.1537   1.0000
     diffmob |  -0.0103   0.0683   0.1255   0.3260   0.2615   1.0000
    diffcare |  -0.0030   0.0644   0.1032   0.1632   0.3000   0.3712   1.0000
```



High school graduates (regression 2):

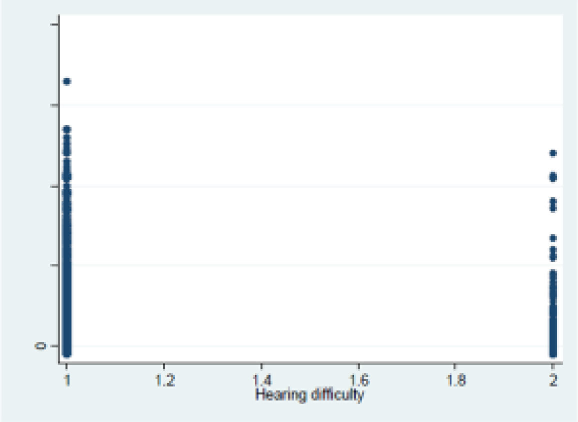
```
. correlate highschool diffhear diffeye diffrem diffphys diffmob diffcare
(obs=487,219)

             | highsc~l diffhear  diffeye  diffrem diffphys  diffmob diffcare
-------------+---------------------------------------------------------------
  highschool |   1.0000
    diffhear |   0.0039   1.0000
     diffeye |   0.0039   0.1085   1.0000
     diffrem |   0.0028   0.0827   0.1114   1.0000
    diffphys |   0.0113   0.0969   0.1117   0.1537   1.0000
     diffmob |  -0.0043   0.0683   0.1255   0.3260   0.2615   1.0000
    diffcare |  -0.0004   0.0644   0.1032   0.1632   0.3000   0.3712   1.0000
```

Heteroscedasticity:

Heteroscedasticity (also spelled heteroskedasticity) refers to the circumstance in which the variability of a variable is unequal across the range of values of a second variable that predicts it. A scatterplot of these variables will often create a cone-like shape, as the scatter (or variability) of the dependent variable widens or narrows as the value of the independent variable increases. The inverse of heteroscedasticity is homoscedasticity, which indicates that a dependent variable's variability is equal across values of an independent variable.

We conduct the Breusch-Pagan test for heteroscedasticity on this data to find out that the regression indeed is heteroscedastic. Below is the plot of the residuals. Normally this plot would show a distinctive pattern due to the heteroscedastic nature of the variance but in this case since the explanatory variable can have only two values (yes or no), the plot shows different y values at only 2 x values. The residual plots for the rest of the disability variables would be similar.

Residual plot



Conclusion:

To summarize, multiple linear regression proved to be a good technique to obtain estimates of coefficients for the disabilities but was a terrible predictor of income. This is simply due to numerous omitted variables that would explain an individual's income. These could include the geographical location, the nature of the college degree, other skills, race, gender etc. The regression could also do with more quantitative variables instead of qualitative ones. These measures would improve the R^2 value of the regression and make it a better model to predict an individual's income, but as it stands it is a reasonably good model to measure the effect of a disability on someone's income.

My presentation also included a research paper by three Iranian researchers who used multiple linear regression to predict the volume in trade by the price of a security. The paper by Gharehchopogh, Bonab & Khaze can be found here.[1]

[1] https://www.researchgate.net/publication/262639062_A_Linear_Regression_Approach_to_Prediction_of_Stock_Market_Trading_Volume_A_Case_Study