**Statistical\_Analysis\_2 \_12310015**

**1. How would you design an experiment to uncover the causal effect of IQ on wage? Is this experiment feasible? Is it ethical?**

Wage/salary is a private data. Disclosure of wage/salary of employee by the employer is not possible as they are bound by non-disclosure agreement.

It is only possible if the employer is willing to understand the causal effect of IQ on wage, highly unlikely.

Judging employees on IQ is unethical as an employer.

**2. What kind of factors might be contained in u? Are these factors likely to be correlated with IQ?**

U might include the following factors,

1. Number of years of education/ highest degree
2. Number of years of working experience
3. Job profile and the salary with the profile
4. Communication skill of the person

Working Experience and job profile are not correlated to IQ in any way.

Educational years/ highest degree and Communication might have a relation with IQ.

**3. Will a simple regression of wage on IQ using the observational data uncover the causal effect of IQ on wage? Explain.**

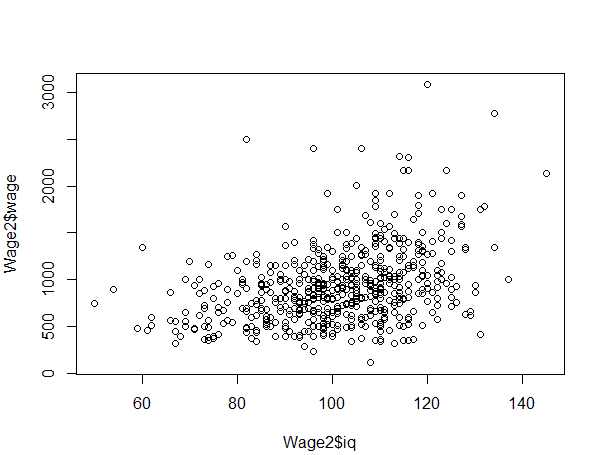
The causal effect of IQ and wage will be determined by simple regression of the collected data with respect to wage.

An iterative process of identifying factors that impact wage, regression modelling with the factors and hypothesis testing to validate the impact of the variables on wage. This needs to go on till we cannot eliminate any factor that impacts wage in any way or data for the factor cannot be collected in an experiment or in random trails.

**4. What are the average values of wage and IQ in the sample? What are their minimum and maximum values?**

|  |  |  |
| --- | --- | --- |
|  | Wage (in '000s) | IQ |
| Avg | 928.7 | 101.0045 |
| Max | 3078 | 145 |
| Min | 115 | 50 |

**5. Obtain a scatterplot of wage and IQ. What kind of relationship is indicated by this scatterplot in terms of (a) direction, (b) linearity, and (c) strength?**



1. Direction: The scatter shows a positive relation
2. Linearity: The points are not on a straight line. But a straight line can be approximated to fit the scatter
3. Strength: There is not too much rise in wage with rise in IQ, so a moderate strength

**6. What is the correlation coefficient between wage and IQ? Is this what you expected based on your characterization in part 5?**

> cor(Wage2$wage,Wage2$iq)

[1] 0.3725479

Positive value for correlation coefficient means a positive relationship between wage and IQ. The coefficient value being 0.3725479 shows a moderate strength.

The correlation was expected to be positive and the strength to be as evident.

**7. Estimate the model in (1) using OLS.**

**(a) Interpret the estimated intercept βˆ 0.**

**(b) Interpret the estimated coefficient βˆ 1.**

**wage = β0 + β1IQ + u, ---------(1)**

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -44.65407375 |
| iq | 9.636709833 |

1. The estimated intercept coefficient states that if IQ of an employee is zero, then the wage should be -44.65407375. This value does not make much sense as wage cannot be negative.
2. The estimated coefficient βˆ 1 states that if in an employee, the IQ increase/ decreases by 1 point, then the wage should also increase/ decrease by 9.636709833 given that all other conditions remain the same.

**8. In terms of the model parameters, state the null hypothesis that IQ is not (linearly) associated with wage. State the alternative hypothesis that IQ is (linearly) associated with wage.**

**Ho**: wage is not linearly associated with IQ. Therefore, wage != β0 + β1\* IQ

It can also be said that β1 = 0

**Ha**: wage is linearly associated with IQ. Therefore, wage = β0 + β1\* IQ

It can also be said that β1 != 0

**9. Can you reject the null hypothesis in part 8 against the alternative hypothesis in part 8 at the 5% significance level? Explain. Suppose that instead of (1), we estimate the model**

**wage = β0 + β1IQ + β2educ + u. --------(2)**

For Model 1,

**wage = β0 + β1\*IQ + u**, ---------(1)

Regression analysis on Excel gives the following,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -44.65407375 | 82.49874259 | -0.54126975 | 0.588458197 |
| Iq | 9.636709833 | 0.807830779 | 11.92911942 | 1.60283E-30 |

Now, at a 5% significance level (0.05), we can reject Ho as p-value for IQ is less than 0.05

This means that IQ have a statistically significant linear relationship with Wage at 5% significance level.

For Model 2,

**wage = β0 + β1\*IQ + β2\*educ + u**. --------(2)

Regression analysis on Excel gives the following,

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -270.2085968 | 88.25415044 | -3.061709794 | 0.002267498 |
| Iq | 6.379808481 | 0.94336675 | 6.762808296 | 2.46167E-11 |
| Educ | 41.15278053 | 6.504257286 | 6.327052993 | 3.96989E-10 |

The p-values for both β1 and β2 are very low. They are both less than 0.05

This means that IQ and educ have a statistically significant linear relationship with Wage at 5% significance level.

Hence, we reject Ho.

**10. Does this model give a more convincing causal effect of IQ on wage? Why or why not?**

Model 2, **wage = β0 + β1\*IQ + β2\*educ + u,** gives a more convincing causal effect of IQ on wage as educational years is now taken into consideration in linear modelling. This reduces the bias in the model.

**11. What are the average, minimum, and maximum values of educ in the sample?**

|  |  |
| --- | --- |
|  | Edu |
| Avg | 13.47458 |
| Max | 18 |
| Min | 9 |

**12. Estimate the model in (2) using OLS.**

**(a) Interpret the estimated intercept βˆ 0. Does the intercept make sense?**

**(b) Interpret the estimated coefficients βˆ 1 and βˆ 2.**

|  |  |  |
| --- | --- | --- |
|  | *Coefficients* | *Standard Error* |
| Intercept | -270.2085968 | 88.25415044 |
| Iq | 6.379808481 | 0.94336675 |
| Educ | 41.15278053 | 6.504257286 |

The estimated intercept of βˆ 0 states that is the IQ of an employee is zero and the educational years is zero then the wage for the employee should be -270.2085968

This does not make any sense as IQ of a person cannot be zero and wage cannot be negative.

The estimated coefficient of βˆ 1 states that for an increase/decrease in IQ by 1 value, the wage should be increased/decreased by 6.379808481, given that there is no change in his/her educational years or any other conditions.

The estimate coefficient of βˆ 2 states that for an increase/decrease in educational years of 1 unit, the wage should be increased/ decreased by 41.15278053, given that there is no change in IQ or any other conditions.

**13. Regress educ on IQ and verify the omitted variable bias formula**

**β˜ 1 = βˆ 1 + βˆ 2 ˜δ1, where all notation is as in the lecture slides.**

E (wage| IQ,Educ) = β0 + β1\*IQ + β2\*Educ

β0 = -270.2085968

β1 = 6.379808481

β2 = 41.15278053

E (Educ| IQ) = δ0 + δ1\*Q

δ0 = 5.480906031

δ1 = 0.079141708

Then, E (wage| IQ) = (β0 + β2\* δ0) + (β1 + β2\* δ1)\* IQ

= -44.65407375 + 9.636709833\* IQ

> b0 = -270.2085968

> b1 = 6.379808481

> b2 = 41.15278053

> d0 = 5.480906031

> d1 = 0.079141708

>

> b1 + b2\*d1

[1] 9.63671

Hence, the formula is verified.

**14. What is the predicted wage of the first individual in the sample according to the estimate of model in (2)? Is this individual overpaid or underpaid?**

|  |  |  |
| --- | --- | --- |
| *Observation* | *Predicted wage* | *Residuals* |
| 1 | 855.2258092 | -155.2258092 |

The first observation in the dataset is having a wage of 700, but the model predicts a wage of 855.2258092. So, according to the model, there is a residual showing the person to be overpaid.