ASSIGNMENT SUBMISSION FORM

This will be the first page of your assignment

Course Name : **Statistical Analysis - 2**
Assignment Title : **Individual Assignment**
Submitted by : **Unnati Khinvasara**

Student Name	PG ID
Unnati Khinvasara	12120097

ISB Honour Code

- I will represent myself in a truthful manner.
- I will not fabricate or plagiarise any information with regard to the curriculum.
- I will not seek, receive or obtain an unfair advantage over other students.
- I will not be a party to any violation of the ISB Honour Code.
- I will personally uphold and abide, in theory and practice, the values, purpose and rules of the ISB Honour Code.
- I will report all violations of the ISB Honour Code by members of the ISB community.
- I will respect the rights and property of all in the ISB community.
- I will abide by all the rules and regulations that are prescribed by ISB.

Note: Lack of awareness of the ISB Honour Code is never an excuse for a violation. Please go through the Honour Code in the student handbook, understand it completely. Please also pay attention to the following points:

- Please do not share your assignment with your fellow students under any circumstances if the Honour Code scheme prohibits it. The HCC considers both parties to be guilty of an Honour Code violation in such circumstances.
- If the assignment allows you to refer to external sources, please make sure that you cite all your sources. Any material that is taken verbatim from an external source (website, news article etc.) must be in quotations. A much better practice is to paraphrase the source material (it still must be cited).

(Please start writing your assignment below)

Data Background

We want to analyse the relationship between the wages and IQ of the employees of a firm.
We have a random sample of 935 employees. Variables available are as follows -

- 'Wage' contains the wages of the employees in thousands of rupees.
- 'IQ' contains their IQ scores.
- 'Educ' contains their number of years of education.

We proceed with the following model that explains the relationship between wages & IQ -

$$wage = \beta_0 + \beta_1 IQ + \epsilon \qquad \text{(Model 1)}$$

Q1 - What kind of factors might be contained in ϵ? Are these likely to be correlated with IQ?

ϵ refers to Error Term of the population data. It consists of all other variables which account for the difference in the estimated y-hat value vs the actual y value. In the given scenario, other factors like Gender, Age, Daily Hours Worked, Work Experience could also be a contributing factor that potentially impact Wages apart from IQ.

It is key to note that one of the basic assumptions of linear regressions is independence between variables. Thus, these factors contained in the error term (like Gender / Age / etc.) for the purpose of Linear Regression are assumed to be constant, i.e., for purpose of Model1, we assume no correlation between factors present in the ϵ and IQ.

However, in reality, these other factors could be correlated with IQ or be independent, for which we need to check for multicollinearity and/or perform variance inflation factor analysis, etc.

Q2 - Will a simple regression of wage on IQ uncover the causal effect of IQ on wage? Explain.

Simple regression of wage on IQ merely helps us understand the relationship between the two variables and derive the line of best fit to statistically derive the dependent variable (wage) based on independent variable (IQ). However, it does not imply relation of causality between the two. To uncover causality, we need to have domain knowledge and can also conduct hypothesis testing.

Simply put, any two random variables, for eg – Wages and Temperature, would also be able to give a line of regression. However, temperature (in most cases!) would not have any causal effect on wages.

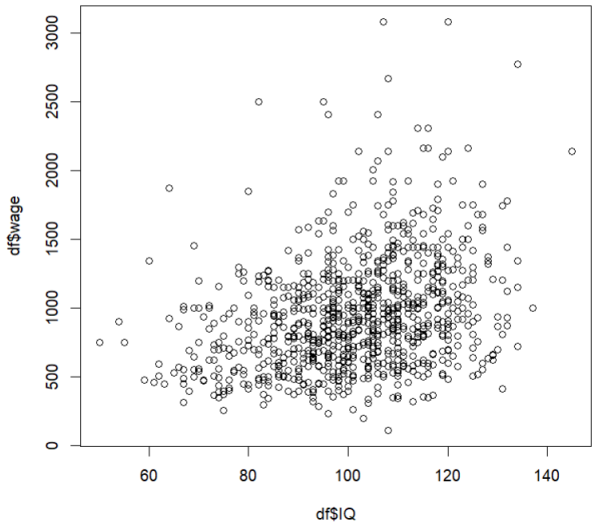
Q3 - What are the average values of wage and IQ in the sample? What are their minimum and maximum values?

	Wage	IQ
Average	957.9	101.30
Minimum Value	115	50
Maximum Value	3078	145

```
> summary(df)
      wage              IQ
 Min.   : 115.0   Min.   : 50.0
 1st Qu.: 669.0   1st Qu.: 92.0
 Median : 905.0   Median :102.0
 Mean   : 957.9   Mean   :101.3
 3rd Qu.:1160.0   3rd Qu.:112.0
 Max.   :3078.0   Max.   :145.0
```

Q4 - How would you characterize the relationship between wage and IQ in terms of (a) direction, (b) linearity, and (c) strength?

Let us plot the scatterplot taking IQ as x (independent variable) & Wages as y (dependent variable).



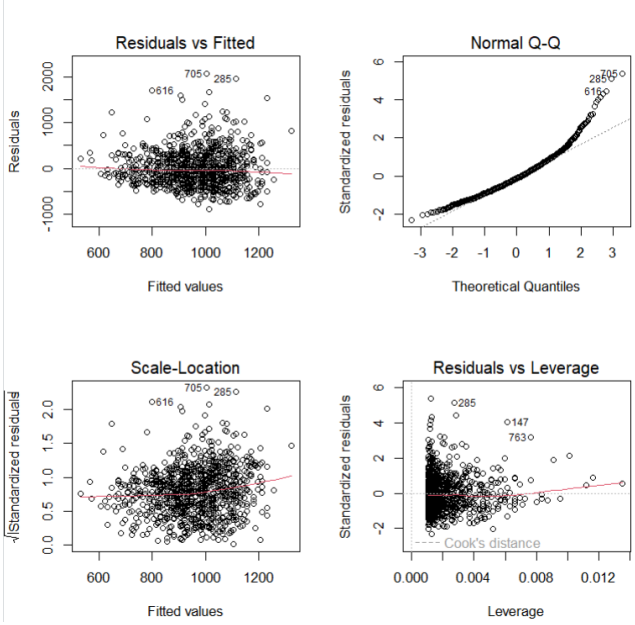
We can see the displacement of the sample values along both the axes from the scatterplot.

Direction – Since we can see the points going in the upward direction from left to right, we can assume a positive relationship between both the variables. To further check, the sign in the covariance and / or correlation coefficient should cross verify the assumption.

It should be noted that covariance is 1881.324, i.e. shows positive relationship. However, the number does not signify any magnitude.

```
> cov(df$wage,df$IQ)
[1] 1881.324
```

Linearity – Visually from the above scatterplot, we can see that points are moving left to right in a straight upward direction. From which we can infer fair linearity. However, we should check the linearity assumption by plotting a QQ plot or from other diagnostic plots for linear regression (plotted below).



We can see that the QQ plot has a fairly straight line barring the end extremes. Thus, we can deduce that residual values are more or less normally distributed.

Further, the residual vs fitted graph, the points are spread around the line, which is fairly straight. This also reinforces linearity assumption.

The scale location plot checks the assumption of homoscedasticity (residuals are spread randomly means homoscedastic)

Strength – From the main scatter plot, we can see points fairly closely clustered together, from which we can infer that the strength would be moderate between IQ and wages. To get the exact strength (magnitude), we could check the correlation coefficient between the variables.

Q5 - What is the correlation coefficient between wage and IQ? Is this what you expected based on your characterization in part 4?

```
> cor(df$IQ,df$wage)
[1] 0.3090878
```

The correlation coefficient between Wage and IQ is 0.3090878. This is roughly what we expected based on Q4 above since we estimated moderately positive linear relationship between the two based on the scatterplot above.

Q6 - Estimate the model in (1) using OLS.

(a) Interpret the estimated intercept $\beta 0$

(b) Interpret the estimated coefficient $\beta 1$. Does it have the sign that you would expect?

```
> summary(m1)

Call:
lm(formula = df$wage ~ df$IQ)

Residuals:
   Min     1Q Median     3Q    Max
-898.7 -256.5  -47.3  201.1 2072.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.9916    85.6415   1.366    0.172
df$IQ         8.3031     0.8364   9.927   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 384.8 on 933 degrees of freedom
Multiple R-squared:  0.09554,   Adjusted R-squared:  0.09457
F-statistic: 98.55 on 1 and 933 DF,  p-value: < 2.2e-16
```

Model 1 SLR Equation:

Wages = 116.9916 + 8.3031 * IQ + Error

$\beta 0$ = 116.9916

It is a constant value in the equation and also the average value for Y, when X is 0. Standard error is +-85.6415 in this coefficient on account of sampling variability.

$\beta 1$ = 8.3031

It is the slope for the independent variable of IQ. We can interpret it as the estimated increase in wages when one unit of IQ increases with a standard error margin of 0.8364 due to sampling variability. Further, it does have the sign we expect (i.e., +ve, since both covariation and correlation have been positive)

Q7 - In terms of the model parameters, state the null hypothesis that IQ is not (linearly) associated with wage. State the alternative hypothesis that a higher IQ is associated with a higher wage.

Let $\beta 1$ be the population slope intercept of the IQ on Wages.

H0 : $\beta 1 = 0$ (then IQ * $\beta 1 = 0$, hence wages would not be associated with IQ, but only on $\beta 0$ constant)

H1 : $\beta 1 > 0$ (since we have prior notion for direction of $\beta 1$ to be positive, we take up a one-sided alternate hypothesis)

Q8 - Can you reject the null hypothesis in part 7 against the alternative hypothesis in part7 at the 5% significance level?

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.9916    85.6415   1.366    0.172
df$IQ         8.3031     0.8364   9.927   <2e-16 ***
```

$t_{\beta 1}$ = (8.3031 – 0) / 0.8364 = 9.927
Since 9.927 > 1.646488 (i.e t-value at 5% LOS), we reject the null at 5% level.
β^1 is statistically greater than 0 at the 5% level.

Suppose that instead of (Model 1), we estimate the model

$$wage=\beta 0+\beta 1IQ+\beta 2educ+\epsilon \qquad \textbf{(Model 2)}$$

Q9 - What are the average, minimum, and maximum values of educ in the sample?

The values of educ in the sample are as follows –

Average	= 13.47
Minimum Value	= 9
Maximum Value	= 18

```
      educ
Min.   : 9.00
1st Qu.:12.00
Median :12.00
Mean   :13.47
3rd Qu.:16.00
Max.   :18.00
```

Q10 - Estimate the model (2) using OLS.

(a) Interpret the estimated intercept β0.Does the intercept make sense?

(b) Interpret the estimated coefficients β1 and β2

```
> summary(m2)

Call:
lm(formula = df$wage ~ df$IQ + df$educ)

Residuals:
    Min      1Q  Median      3Q     Max
-860.29 -251.00  -35.31  203.98 2110.38

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -128.8899    92.1823  -1.398    0.162
df$IQ          5.1380     0.9558   5.375 9.66e-08 ***
df$educ       42.0576     6.5498   6.421 2.15e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 376.7 on 932 degrees of freedom
Multiple R-squared:  0.1339,    Adjusted R-squared:  0.132
F-statistic: 72.02 on 2 and 932 DF,  p-value: < 2.2e-16
```

Model 2 SLR Equation:
Wages = -128.8899 + 5.1380 * IQ + 42.0576 * Education Years + Error

β0 = -128.8899

As seen above for model 1, this is a constant value in the equation and also the average value for Y, when X is 0, with a standard variance of +-92.1823. However, it is worthwhile to note that the intercept is negative, which does not make it invalid because it is merely a constant value. It just signifies the start of the best fit regression line on the Y Axis.

In this specific case, it can so happen that at a lower value of IQ and Education, the wage values aren't probable. However, as the variable number increases to a fairly larger value, we do get accurate predictions.
Thus, the overall relationship between the variables is important in a linear model, rather an a +ve or -ve intercept.

β1 = 5.1380

It is the estimate of the slope for variable of IQ. It simply means that everything else remaining same, if IQ increases by 1 unit, the corresponding increase in wages would be 5.138 (with the standard error margins as given).
Further, it is also interesting to note that between Model 1 and Model 2, the standard error for β1 has increased from 0.83 to 0.95. This can be on account of the high correlation (0.52) between the independent variables (IQ & Education).

β2 = 42.0576

This is the estimate of slope for education variable. Like β1 above, it simply means that everything else remaining same, if education increases by 1 year, the corresponding increase in wages would be 42.05 (before accounting for standard error)

Q11 - Regress educ on IQ and verify the omitted variable bias formula
$\tilde{\beta}1=\beta1+\beta2\tilde{\delta}1$, ***where all notation is as in the lecture slides.***

The omitted variable bias formula regressing educ on IQ is formulated as follows –

beta^~ is similar to Beta1^ + Beta2^ * delta1~
60.214 = 5.1380 + 42.0576 * 0.075256

If alternatively, we want to regress IQ on education, the following is the OMV formula –
8.3031 = 42.0576 + 5.1380 * 3.5338

Q12 - 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?

Model 2 SLR Equation:
Wages = -128.8899 + 5.1380 * IQ + 42.0576 * Education Years + Error

Details of first individual –

```
> head(df,1)
  wage IQ educ
1  769 93   12
```

Thus, predicted wage of the first individual is = -128.8899 + 5.1380 * 93 + 42.0576 * 12
= Rs. 853.6353 (in Rs. '000)

But as we can see, the actual reported wage of the person is only Rs. 769 (In Rs, '000), thus he seems to be an underpaid.