Written answers for Introduction to Statistics - HW2

1. Model 1: sei vs. educ

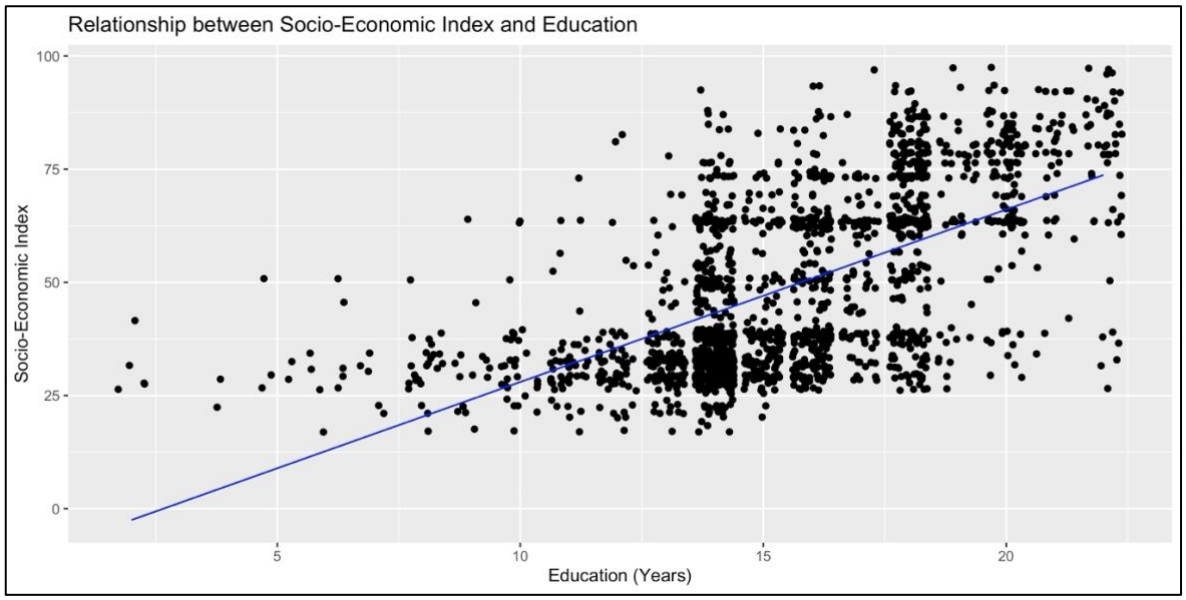
Select two variables, \sei" (socioeconomic status) and \educ" (years of education). Note that \socioeconomic index (SEI) of current occupation" scores summarize the differences in prestige between occupations, as assessed by the education required and the earnings provided. It is commonly conceptualized as the social standing or class of an individual or group.

1. Calculate Pearson's r correlation coefficient to show the relationship between the two variables and provide an interpretation of the coefficient.

   Ans: The Pearson's *r* correlation coefficient is 0.5957052 (ref R file). is a positive value and is between 5<*r*<=7. Therefore, we can say that while, the year of education is **positively correlated** with the socio-economic index, the strength of their relationship is **weak.**

2. Using the \ggplot2" package, create a scatter plot to visualize the relationship between \sei" (socioeconomic status) and \educ" (years of education), including a regression line based on a two-variable regression model.

   Ans: Please find below the scatter plot, which includes the **regression line (in blue)** that visually presents the relationship between socio-economic index (Y-axis) and years of education (X-axis). I used \ggplot package, removed the confidence interval (as it doesn't provide much value visually) and kept the **jittering to avoid any overlaps** with the values. This plot should also be downloaded as "scatter_sei_educ.jpeg" when you run the code.



3. Run a regression analysis with "\sei" as the dependent variable and "\educ" as the independent variable. Present the results in a table using the \Stargazer" package.

   Ans: Please find the results of my regression analysis between the socio-economic index (DV) and years of education (IV) in Table 1 using the "\Stargazer" package. This analysis will also be saved as "model1.txt" when you run the code.

	Dependent variable:
	Socio-Economic Index
Education (Years)	3.81***
	(0.12)
Constant	-10.13***
	(1.86)
Observations	1,906
R2	0.35
Adjusted R2	0.35
Residual Std. Error	15.67 (df = 1904)
F Statistic	1,047.32*** (df = 1; 1904)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 1

4. Write out the regression equation based on your regression coefficient, representing the relationship between those two variables.

   Ans: The regression equation can be summarised as follows where the Socio-economic index and Years of Education are the dependent (Y) and independent (X) variables, respectively. $\boldsymbol{\alpha}$ is the y-intercept parameter, the slope parameter $\boldsymbol{\beta}$ is 3.81 and $\boldsymbol{\mu}$ indicates the random component of Y also known as residual. Residuals are the estimated error between the observed and the estimated of values of Y.

   $$Socio-Economic\ Index\ (Y) = \mathbf{-10.13}\ (\alpha) + \mathbf{3.81}\ Years\ of\ Education\ (X) + \boldsymbol{\mu} \quad \text{.....(I)}$$

5. Interpret the OLS regression results by explaining how each coefficient reflects the relationship and assessing the statistical significance. Also, interpret goodness-of- t measures to evaluate model accuracy.

   Ans: The Pearson's *r* correlation coefficient has already made clear that the variables educ and sei are positively related albeit the relationship is weak. To quantify the correlation further, we run a bivariate regression of these variables.

   The bivariate regression results summarised in equation (I) shows the value of y-intercept (α) as -10.13, meaning that when years of education equals 0, the average socio-economic index is -10.13. The negative intercept value indicates that the even when socio-economic index is 0, there is some value for the years of education. The slope parameter (β) which is the coefficient of years of education, is 3.81 points. This means that **with every unit change in years of education, the average socio-economic index will increase by 3.81 points.** In Table 1, the asterisk marks on the top of $\alpha\ and\ \beta$ indicates that the **p-value of $\boldsymbol{\alpha\ and \beta}$ are less than 0.01**(see note at end of Table 1) which corresponds

to 1% significance level. Thus, the coefficients explaining the relationship between the socio-economic index and education are statistically significant.

Goodness of Fit measures for Model 1:

1. *Residual Standard Error (RSE):*
   RSE gives the average distance of data points from the regression line and in Model 1, the average error is **15.67 points**. A lower RSE is better for the model.
2. *R^2:*
   0.35 which indicates that **35%** of the variance in the socio-economic index is explained by years of education.
3. *Adjusted R^2:*
   This statistic adjusts the R^2 for the number of parameters in the model by penalising it. For Model 1, R^2 and Adjusted R^2 are both **35%**, may be because it has only two variables.
4. *F- Statistic:*
   It is a hypothesis test of overall significance. It checks if at least one of the coefficients is non-zero. If the p-value for the F-test is less than 0.05 than the regression model fits the data better than the model with no independent variables. For Model 1, the value of the F-statistic is 1047.32 with a p-value <0.01 indicating that **the years of education can explain the variance in the dependent variable, the socio-economic index, than an intercept-only model**.

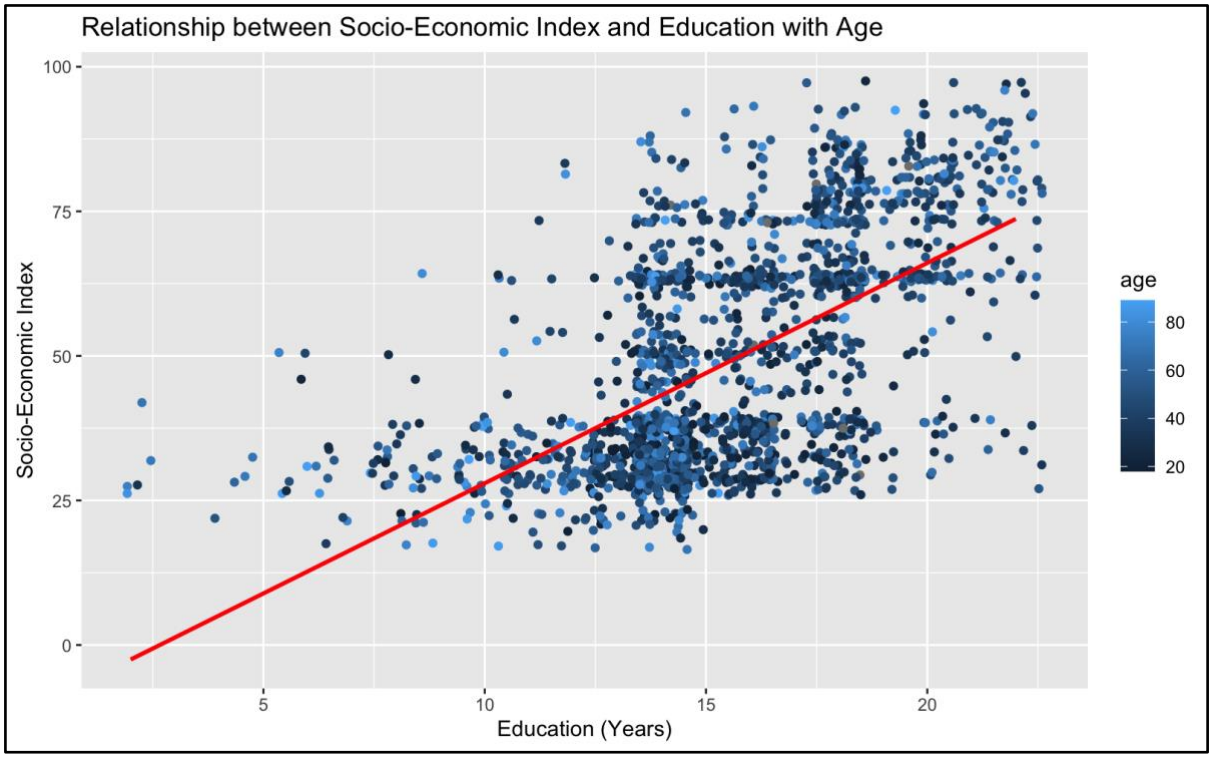
Therefore, the regression model is accurate and statistically significant, however, lets observe the changes when this relationship is controlled for other variables.

2. Model 2: sei with educ and age

Select three variables: \sei" (socioeconomic status), \educ" (years of education), and \age"(years of age).

1. Using the \ggplot2" package, create a scatter plot to visualize the relationship between the dependent variable \sei" and the independent variable \educ," with points coloured by \age." Additionally, include a regression line from a model involving \sei" and \educ."

   Ans: Please find below the scatter plot including the **regression line (in red)** that presents the relationship between the socio-economic status (Y-axis) and years of education (X-axis) with points coloured by variable age in **gradients of blue.** The sequential scale on the right shows that the darker the blue colour of the dot, lower the age of the participant and vice-versa.
   This plot should also be downloaded as "scatter_sei_educ_age.jpeg" when you run the code.



2. Run a multiple regression analysis with \sei" as the dependent variable and \educ" as the independent variable, controlling for \age". Present the results in a table using the \Stargazer" package.

   Ans: Please find the results of my multiple regression analysis between socio-economic index (DV), years of education (IV) and age (control variable) in Table 2 using the "\Stargazer" package. This analysis should also be downloaded as "model2.txt" when you run the code.

	Dependent variable:
	Socio-Economic Index
Education (Years)	3.85***
	(0.12)
Age	0.11***

	(0.02)
Constant	-16.03***
	(2.16)
---	---
Observations	1,898
R2	0.37
Adjusted R2	0.36
Residual Std. Error	15.53 (df = 1895)
F Statistic	546.16*** (df = 2; 1895)
==========	==========
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2

3. Write out the regression equation based on your regression coefficient, representing the relationship between those three variables.

   Ans: The regression equation, representing the relationship between the three variables, Socio-economic Index (Y), years of education (X) and Age (Z) is summarised in equation II. α_1 is the y-intercept parameter, the slope parameter β_1 for X is 3.81, the slope parameter β_2 for Z is 0.11 and μ indicates the residual.

$$Socio - Economic\ Index\ (Y) = -\mathbf{16.03}\ (\alpha 1) + \mathbf{3.85}\ Years\ of\ Education\ (X) + \mathbf{0.11}\ Age\ (Z) + \mu \quad \ldots\ldots \text{(II)}$$

4. Interpret the OLS regression results by explaining how each coefficient reflects the relationship and assessing the statistical significance. Also, interpret goodness-of- t measures to evaluate model accuracy.

   Ans: The multi-variable regression results summarised in equation (II) has a value of y-intercept (α) as -16.03, meaning that when years of education equals 0, the average socio-economic index is -16.03. The negative increase in α_1 can be intuitively visualised as the regression line getting steeper. The slope parameter (β_1) which is the coefficient on years of education is 3.85 points. This means that with every unit change in years of education, the average socio-economic index will be **increased by 3.81** points **keeping the age constant.** In multi-variate regression, its important to note that the value of each slope parameter is correlated while keeping the other variables constant. Similarly, the slope parameter (β_2) which is the coefficient for control variable age is 0.11 points. This indicates that with every unit change in age, there will be **an increase of 0.11 points in average socio-economic index** while keeping other variables constant.

   Like Table 1, in the Table 2, we note that the p-values for α_1, β_1 *and* β_2 are less than 0.01 which corresponds to 1% significance level. Thus, the coefficients explaining the relationship between socio-economic index and education with age are statistically significant.

   Goodness of Fit measures for Model 2:

   1. *Residual Standard Error (RSE):*
      Here, **RSE is 15.53 points which is lower than Model 1's RSE (15.67 points).** Lower the RSE, better fitted the regression line.
   2. *R^2:*

0.37 which indicates that **37% of variance in socio-economic index is explained by years of education and age.** This means the variable age increases in causal explainability of the relationship vs Model 1.

3. *Adjusted R^2:*
   Given that this is a 3-variable model, the Adjusted R^2 adjusts the R^2 for the number of parameters and provides with a value of 0.36 that is **36%**, an increase from Model 1.
4. *F- Statistic:*
   For Model 2, the value **of F-statistic is 546.16 with p-value <0.01** indicating that the year of education with age can explain the variance in the dependent variable, socio-economic index, then an intercept-only model.

Therefore, the regression model 2 is accurate and statistically significant as well! However, with a lower RSE and an increased Adjusted R^2 than Model 1, we can say that it explains the variance in the socio-economic index better.

3. Model 3: sei with educ, age, and tvhours

Select four variables: \sei" (socioeconomic status), \educ" (years of education), \age" (years of age), and \tvhours" (hours spent watching TV).

1. Run a multiple regression analysis with \sei" as the dependent variable and \educ" as the independent variable, while controlling for both \age" and \tvhours". Present the results in a table using the \Stargazer" package.

   Ans: Please find the results of my multiple regression analysis between socio-economic index (DV), years of education (IV) while controlling for age and TV hours presented in Table 3 using "\Stargazer" package. This analysis should also be downloaded as "model3.txt" when you run the code.

	Dependent variable:
	Socio-Economic Index
Education (Years)	3.70***
	(0.15)
Age	0.13***
	(0.03)
TV Hours	-0.90***
	(0.17)
Constant	-12.37***
	(2.78)
Observations	1,245
R2	0.38
Adjusted R2	0.38
Residual Std. Error	15.28 (df = 1241)
F Statistic	251.55*** (df = 3; 1241)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3

2. Write out the regression equation based on your regression coefficients, representing the relationship between those four variables.

   Ans: The regression equation, representing the relationship between the four variables, Socio-economic Index (Y), years of education (X), Age (Z_1) and TV Hours (Z_2) is summarised in equation III. $\hat{\alpha}_1$ is the y-intercept parameter, the slope parameter $\hat{\beta}_1$ for X is 3.81, the slope parameter $\hat{\beta}_2$ for Z_1 is 0.13, the slope parameter $\hat{\beta}_3$ for Z_2 *is* -0.90 and μ indicates the residual.

$$Socio - Economic\ Index\ (Y) = -\mathbf{12.37}\ (\alpha) + \mathbf{3.70}\ Years\ of\ Education\ (X) + \mathbf{0.13}\ Age\ (Z1) + (-\mathbf{0.90})\ TV\ Hours\ (Z2) + \boldsymbol{\mu}$$

…... (III)

3. Interpret the OLS regression results by explaining how each coefficients reflects the relationship and assessing the statistical significance. Also, interpret goodness-of- t measures to evaluate model accuracy.

Ans: The multi-variable regression results summarised in equation (III) has a value of y-intercept $\hat{\alpha}_1$ as -12.37, meaning that when years of education equals 0, the average socio-economic index is -12.37. The slope parameter ($\hat{\beta}_1$) which is the coefficient on years of education = 3.70 points. This means that with every unit change in years of education, the average socio-economic index will be increased by 3.70 points **keeping the age and tvhours constant.** Similarly, the slope parameter ($\hat{\beta}_2$) which is the coefficient for control variable age is 0.13 points. This indicates that with every unit change in age, there will be an increase of 0.11 points in average socio-economic index while keeping other variables constant. And the slope parameter ($\hat{\beta}_3$) which is the coefficient for control variable TV hours is -0.90 points. This indicates that **with every unit change in TV hours**, there will be a **decrease of 0.90 points in average socio-economic index**, keeping other variables constant.

Again, in Table 3, the **p-value of $\hat{\alpha}_1$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ is less than 0.01**(check asterisks) which corresponds to 1% significance level. Thus, the coefficients explaining the relationship between socio-economic index, education, age and TV hours is statistically significant.

<u>*Goodness of Fit measures for Model 3:*</u>

1. *Residual Standard Error (RSE):*
   Here, the RSE is **15.28 points** which is lower than Model 2 (15.53 points) as well as Model 1's RSE (15.67 points).
2. *R^2:*
   0.38 which indicates that **38%** of variance in socio-economic index is explained by years of education with age and TV hours. This means the variable TV hours increases in causal explainability of the relationship vs Model 2 which had only years of education and age.
3. *Adjusted R^2:*
   Given that this is a 4-variable regression model, the Adjusted R^2 statistic adjusts the R^2 for the number of parameters and provides with a value of 0.38 that is **38%**, same as R^2 however, an increase from Model 2 and 1.
4. *F- Statistic:*
   For Model 3, the value of **F-statistic is 251.55 with p-value <0.01** indicating that the year of education with age AND TV hours can explain the variance in the dependent variable, socio-economic index, than an intercept-only model.

Therefore, the regression model 3 is accurate and statistically significant. However, with a lower RSE and an increased Adjusted R^2 than Model 1 and Model 2, we can say that it explains the variance in the socio-economic index more accurately so far.

4. Model Comparison

1. Descriptive Statistics:

Use the \Stargazer" package to create a professional descriptive statistics table for all variables in Model 3. In doing so, relabel the variables and include a table title \Descriptive Statistics."

Ans: Please find the descriptive statistics table for all variables in Model 3 with table title using "\Stargazer" package in Table 4. This should also be downloaded as "descriptive.txt" when you run the code.

Descriptive Statistics

Statistic	N	Mean	St.Dev	Min	Max
Socio-Economic Index	1,911	48.76	19.51	17.10	97.20
Education (Years)	2,018	15.43	3.08	2	22
Age	2,013	47.71	17.35	18	89
TV Hours	1,324	2.98	2.66	0	24

Table 4

2. Combined Regression Table:

Use the \Stargazer" package to generate a clear, well-organized regression table that includes all l three models (Models 1, 2, and 3). Ensure that the x- and y-axes. As well as all variables, are properly labelled, and title the table \Regression Results for Socioeconomic Status (GSS 2010)."

Ans: Please find the combined regression table that includes all three models (Models 1, 2, and 3) as Table 5. This adds a title and labels and is made using "\Stargazer" package. This should also be downloaded as "model1_model2_model3.txt" when you run the code.

Regression Results for Socioeconomic Status (GSS 2010)

	Dependent variable:		
	Socio-Economic Index		
	(1)	(2)	(3)
Education (Years)	3.81***	3.85***	3.70***
	(0.12)	(0.12)	(0.15)
Age		0.11***	0.13***
		(0.02)	(0.03)
TV Hours			-0.90***
			(0.17)
Constant	-10.13***	-16.03***	-12.37***
	(1.86)	(2.16)	(2.78)

Observations	1,906	1,898	1,245
R2	0.35	0.37	0.38
Adjusted R2	0.35	0.36	0.38
Residual Std. Error	15.67 (df = 1904)	15.53 (df = 1895)	15.28 (df = 1241)
F Statistic	1,047.32*** (df = 1; 1904)	546.16*** (df = 2; 1895)	251.55*** (df = 3; 1241)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5

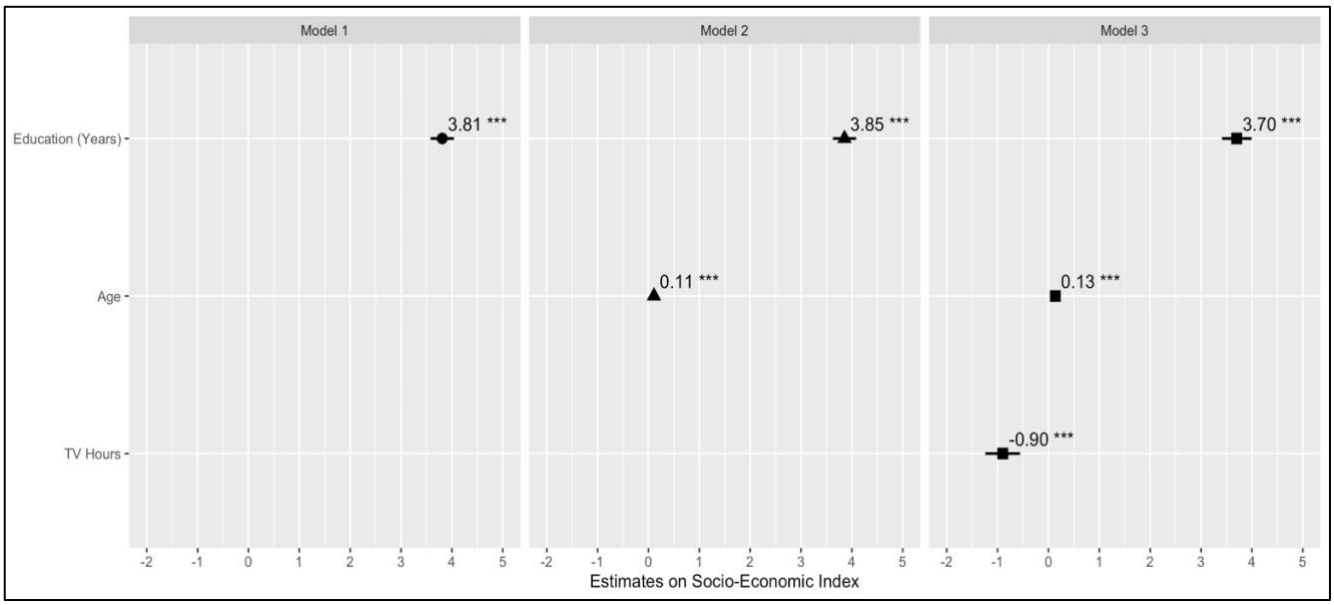
3. Model Comparisons:
Compare three regression models predicting socioeconomic status (\sei") based on education (\educ"). Analyse the differences in performance, coefficients, and goodness- of- t across the models. State which model you prefer and provide a rationale for your choice.

Ans: Let's refer to Table 5, which has all the regression results. All three models and their coefficients are statistically significant at a significance level of 1%, as shown by the F-statistic. But the question is, **is years of education the only cause for the dependent variable, the Socio-economic Index?**

We know from Model 1 that with one year added to education, the average socio-economic index increases by 3.81 points (β). However, the effect of years of education changes when we add for control variables age in Model 2 and age and tvhours in Model 3.

$$\beta \neq \beta_1 \neq \hat{\beta}_1$$

Additionally, β_2, (coefficient of age in Model 2) ≠ 0. And $\hat{\beta}_2, \hat{\beta}_3$ (coefficients/effects of age and tvhours in Model 3) ≠ 0. $\beta_2, \hat{\beta}_2, \hat{\beta}_3$ are all statistically significant at 1%. This indicates that β_1 in Model 1 is **biased** as it incorporates, in a non-transparent way, the effects of other variables on the dependent variable. The same tendency can be seen in Model 2. The comparison plot below shows all the coefficients for Model 1, Model 2, and Model 3 side by side for better visualisation. This plot should also be downloaded as "Q4_Plot.jpeg" when you run the code.



Regarding the Goodness-of-fit measures of all three models, the residual standard error has continuously decreased from Model 1 (15.67) to Model 2 (15.53) through Model 3 (15.28), while the Adjusted R^2 has shown an increase from 35% (Model 1) to 36% in Model 2 and 38% in Model 3. This signifies that the

best-fitted regression line and the highest explanation of the variance in the dependent variable, the socio-economic index, cannot be explained correctly only by the effects of education in isolation. It must be controlled for age and TV hours for better causal explainability given that F-test is significant for all coefficients at 1%.

From the interpretation of coefficients and goodness-of-fit measures across the three models, I conclude that Model 3 has the highest causal explainability about the variance (38%) in the dependent variable, socio-economic index, with the lowest residual standard error (15.28) and non-zero statistically significant coefficients for all the three variables. Therefore, I will prefer Model 3 over Model 2 and Model 1.