**HUYNH DO**

**Module 6, Part 1/2**

Choose at least five variables from the GSS2018 data file to perform Multiple Linear.

1. **Specify which variable is the dependent and independent variable:**Multi linear regression method is used to examine how total family income is influenced by a set of following independent factors: Age, education level (highest year of school completed), marital status, and race.

The following variables are as followed:

* **Independent Variable (IV) :**
1. **Age**: The age (measured in years) of the participant and this is a continuous variable.
2. **Education Level:** The total number of years of formal education completed by the respondent. This is a continuous variable.
3. **Marital Status**: A categorical variable that records the respondent’s marital status such as: Married, divorced or separated.
4. **Race**: A categorical variable representing the respondent's racial group (e.g., White, Black, Other).
* **Dependent Variable (DV):**
1. **The Total Family Income**: This is measured on a scale and is considered a continuous variable.
2. **Using the variables that you’ve chosen, state the null and alternative hypotheses.**

With the above declared above variables, we can construct these 2 Hypotheses:

1. **Null Hypothesis (H₀):**
* There is no statistical relationship between one of those **IVs** with the dependent variable (total family income), meaning, all the regression coefficients are 0.
* **Mathematical Representation**:
	+ H0​:βage​ = βeducation​ = βmarital status​=βrace ​= 0
1. **Alternative Hypothesis (H₁):**
* At least one of the **IVs** has a statistically significant effect on the dependent variable (total family income). This means at least one of the regression coefficients is **NOT** 0
* **Mathematical Representation**:
	+ H1​:At least one βi​ <> 0
1. **Perform Multiple Linear Regression and interpret the results**
* **Descriptive**



The descriptive statistics offers a quick glance on how independent variable factors are related to family income status.

1. **Total Family Income:**
* Mean: The number 17.73, indicating that, on average, the total income family is $17,730.
* Standard Deviation: 5.940, showing the income may deviate mean which translates to [$17,730 +- $5,940].
1. **Age of Respondent:**
* Mean Age: 48.97 years, suggesting that people who participates the survey is almost 49 years old.
* Standard Deviation: 18.061, indicating a wide range of ages and considerable variability in the sample.
1. **Highest Year of School Completed:**
* Mean: The number 13.73 years, suggesting that, on average, people have completed approximately 13 to 14 years of education, which translating to someone either are currently attending or already completed a 2-year college degree.
* Standard Deviation: 2.974, showing moderate variability in the educational attainment of respondents. In other words, young adults either pursuit in higher education by transferring to university or staring to join the labor force right after graduating from high school.
1. **Marital Status:**
* Mean: 2.67, suggesting that, on average, respondents tend to report as divorce status.
* Standard Deviation: 1.689, reflecting a wide spread across marital spectrum.
1. **Race of Respondent:**
* Mean: 1.39, indicating that a majority of respondents belong to the WHITE race.
* Standard Deviation: 0.685, indicating some level of diversity but with clustering around a primary group.
1. **Descriptive Findings:**
* Income, age, and education levels show a wide range of values, with moderate variability while marital status and race have clustering tendencies around particular categories.
* **Variable List:**

****

* **Variables Entered**: All variables are entered in the regression model are listed.
* **Enter**: The Enter method was used for regression. This means that all independent variables listed were, unlike stepwise methods, entered into the model simultaneously. This also said in both table foot notes (**a & b bullet)**
* **Remove**: There are no variables are being excluded from the model.
* **Model Summary**

****

1. **Correlation Coefficient**: **R=** **0.536**

This number shows a positive relationship between the **IVs** and the independent total family income variable.

1. **Coefficient of Determination: R2 = 0.287**

The number 0.287 or **28.7**% of the dependent Total Family Income variable variance is explained by the model's IVs. However, a significant remaining 71.3% portion of the variance remains unexplained. This indicates that other factors included in the model might also influence family of the dependent Total Family Income variable.

1. **Adjusted R2** = **0.286**

The **R2 = 0.286** adjusts for the number of predictors in the model. It provides a more accurate measure of model fit for models with multiple predictors. In other words, there’s still room for improvement with additional predictors or alternative modeling approaches such as Occupation type, Employment Status, education Level of Spouse/Partner, etc…

1. **Std. Error of the Estimate**: **5.024**

This number provides an estimate of the typical distance that the observed values fall from the regression line. A smaller value indicates that the data points are closer to the fitted line, whereas a larger value suggests greater variability around the predicted values. Ideally, smaller values mean better, more precise predictions.

1. **Model Summary Findings**:
* Approximately 28.7% of the variability in total family income is explained by the model, while 71.3% remains unexplained.
* The model's predictive performance can likely be improved by exploring additional predictors or refining the existing ones.
* The Adjusted R Square is close to the R Square suggesting that the model’s complexity (number of predictors) is reasonably justified without overfitting.
* **ANNOVA**

****

* 1. **Sum of Squares (SS)**
* **Residual Sum of Squares (54014.620): This left over value after applied the model shows the variance in Total Family Income that is not clarified by the predictors IVS.**
* **Total Sum of Squares (75769.170): This is the total variability in Total Family Income, which is equal to the sum of the Regression and Residual sum of squares.**
	1. **Degrees of Freedom:**
* **Residual df = 2140: This represents the total number of observations minus the number of predictors minus 1 (N - k - 1).**
* **Total df (2144): This is N - 1.**
	1. **Mean Square (MS):**
* **Regression Mean Square (5438.637): Calculated as the Regression Sum of Squares divided by the Regression df.**
* **Residual Mean Square (25.240): Calculated as the Residual Sum of Squares divided by the Residual df.**
	1. **F-statistic**
* **This is the ratio of the Regression Mean Square to the Residual Mean Square. The F-statistic tests whether the regression model provides a better fit to the data than a model with no predictors (essentially, whether the predictors collectively have a statistically significant effect on the dependent variable).**
	1. **Significance Level (Sig.)**:
* The p-value is < 0.001 demonstrate that the regression model is statistically significant and that at least one of the predictors has a significant relationship with Total Family Income. Therefore, we can start **rejecting the null hypothesis**.
	1. **Findings:**
* **The model is statistically significant overall, meaning that the predictors (age, the highest year of school completed, marital status, and race) collectively explain significant variance in Total Family Income.**
* **Both large F-statistic (215.473) and extreme small p-value (<0.001) are strong evidence against the null hypothesis (that the predictors have no effect), thus, confirming that the regression model is meaningful.**
* **While the model is statistically significant, we should also evaluate the weight and significance of individual predictors, by using the coefficients table, to understand which specific variables contribute most to predicting income.**
* **Coefficients**

****

We will interpret 3 unstandardized model coefficients

1. **Constant (Intercept):**
* **Interpretation**: The Unstandardized Coefficient (B)=12.907 represents the anticipated value of the Total Family Income when all **IVs** = 0 (age, the highest year of school completed, marital status, and race). Whereas the intercept itself might not have important translation in this context due to categorical factors like marital status and race, it serves as a baseline value for the model.
1. **Highest Year of School Completed:**
* Unstandardized Coefficient (B): 0.767
* Standardized Coefficient (Beta): 0.377
* t-value: 20.368, p-value (<.001): This is highly significant.
* **Interpretation**: For each completed year of education, the Total Family Income is expected to increase by 0.767 or $767 (controlling for other variables). The positive coefficient indicates that higher education levels are associated with higher family income. Both strong t-statistic and small p-value again confirm that this strong relationship.
1. **Marital Status:**
* Unstandardized Coefficient (B): -1.245
* Standardized Coefficient (Beta): -0.352
* t-value: -17.887, p-value (<.001): This is highly significant.
* **Interpretation**: The negative coefficient suggests that certain marital statuses (e.g., divorced, separated, etc.) are likely to lower total family income. Specifically, any change in marital status is associated with a decrease of 1.245 or $1,245 in total family income.
1. **Race of Respondent**
* Unstandardized Coefficient (B): -0.543
* Standardized Coefficient (Beta): -0.062
* t-value: -3.288, p-value (.001): This is statistically significant.
* **Interpretation**: Certain racial categories cause negative impact the total family income (**lower**). Specifically, A unite change in racial category is associated with a decrease of 0.543 units in income or $543. While significant, the effect size (as shown by the standardized coefficient) is relatively small.
1. **Collinearity Statistics**
* **Tolerance**
* **Highest year of school completed: 0.970**
* **Age of respondent: 0.859**
* **Marital status: 0.858**
* **Race of respondent: 0.951**
* **VIF** **values**
* Highest year of school completed: 1.031
* Age of respondent: 1.165
* Marital status: 1.165
* Race of respondent: 1.051
* **Collinearity Statistic Findings**
* Low VIF suggests that the **IV is not** highly correlated each other (crosstalk), meaning it provides unique information to the model that is not redundant with other variables. In this case, **all of the VIF values are well below 10, indicating no significant multicollinearity concerns among the IVs.**
* **The Tolerance values are all above 0.1, which further supports the conclusion that multicollinearity does not compromise the model.**
* All predictors appear to be relatively independent, making the coefficients more reliable for interpretation.
1. **Coefficient Summary :**
* The highest year of school completed has the strongest positive effect on family income among the IVs (predictor), with:
1. Unstandardized Coefficient (B) = 0.767
2. Standardized Coefficient (Beta) = 0.377
* Race, Marital status and Age cause negative relationships with income (lower), indicating that certain marital statuses and older age may be associated with lower total family income.
* All predictors (**IVs**) have statistical low p-values, meaning they each contribute meaningfully to explaining variation in total family income. The magnitude and direction of these coefficients give insight into the strength and nature of these relationships.
* **Collinarity Diagnostics**

****

**Interpretation**:

* This collinearity diagnostic indicates that while there is a collinearity with Dimension 5 =**17.674** is considered high value. However, the condition index does not exceed a critical threshold 30. Thus, we can safely say that the collinearity in this model may not be severe enough to perform further corrective actions.
* **Residual Statistics**

****

1. **Predicted Value and Interpretations**:
* Minimum = 5.75 and Maximum = 25.41 indicate that the model predicts income with this range. In another word, the model predicts that people are making from the range of $5,750 to $25,410.
* Mean: 17.73 or $17,730 is what model predicts, on average, how much a person is making.
* Standard Deviation: The number 3.185 shows the spread of predicted values around the mean.
1. **Residual Interpretations:**
* Minimum = -20.747 and Maximum = 15.067 show the range of prediction errors, with negative values indicating under-prediction and positive values indicating over-prediction by the model.
* Mean: 0.000 suggests that, on average, the model does not systematically over- or under-predict
* Standard Deviation: The 5.019 reflects the average magnitude of prediction errors, consistent with the overall error term from the model.
1. **Standardized Predicted Value**:
* Minimum: -3.761
* Maximum: 2.410
* Mean: 0.000
* Standard Deviation: 1.000
* **Interpretation**: The standardized predicted values are the predicted values of the dependent variable (total family income) converted into standard units (z-scores). These values have a mean of 0 and a standard deviation of 1, providing a way to understand the relative positioning of predictions compared to the average.
1. **Standardized Residual**:
* **Minimum = -4.130 and Maximum = 2.999 are beyond ±3 is (a typical range is ±2) considered potential outliers larger-than-expected residuals data points with either lager data point or the model’s algorithm was struggled to explain the correlation among the IV variables. In this case, a visual scatterplots will help us to visualize how far apart from the regression line.**
* **Mean: 0.000**
* **Standard Deviation: 0.999**
1. **Residual Statistic Findings:**
* **The residual statistics suggest that the model's predictions are centered around the observed values (mean residual close to zero), with some variability in prediction errors (standard deviation of residuals = 5.019).**
* **There may be a few outliers, as indicated by standardized residuals extending beyond ±3.**
* **Overall, the spread of predicted and residual values suggests a reasonably good fit but highlights areas where the model's predictions may deviate more substantially from actual values. Further examination of potential outliers or influential points may be warranted to assess their impact on the model.**
1. **What Hypotheses should we rejects?**

While the collinearity diagnostics demonstrates a mild collinearity among predictors, it does not significantly meddled with the significance of the centrality of the coefficients. More importantly, all I**Vs** have **Sig. < 0.001,** thus, we can confidently **reject** the **null hypotheses** for each predictor, concluding that all predictors in this model significantly contribute to explaining total family income.

1. **Make sure to assess the test assumption**

To assess the assumptions of the regression analysis, here’s a breakdown of the main assumptions:

1. **Linearity of the Relationship**
* **Assumption**: There should be a linear relationship between the independent variables and the dependent variable.
* **Assessment**: Check the scatterplots between each independent variable and the dependent variable. If the points roughly form a straight line, this indicates a linear relationship.
1. **Homoscedasticity (Constant Variance of Errors)**
* **Assumption**: The variance of residuals should be constant for all values of the independent variables.
* **Assessment**: We can visually assess residuals vs. fitted values plot. If the points are randomly scattered around zero with no apparent pattern, this suggests homoscedasticity. If there is a clear pattern (e.g., a funnel shape), it indicates heteroscedasticity, violating the assumption.
1. **Multicollinearity**
* **Assumption**: Independent variables should not be correlated with one another.
* **Assessment**: Review the Variance Inflation Factor (VIF) values and Tolerance values. VIF values below 10 and tolerance values above 0.1 generally indicate that multicollinearity is not a major concern.

From the confection table:
 

* Highest year of school completed: VIF = 1.031
* Age of respondent: VIF = 1.165
* Marital status: VIF = 1.165
* Race of respondent: VIF = 1.051

These numbers suggest no significant multicollinearity.

1. **Using a scatterplot, visually display the results of the Multiple Linear Regression and explain your findings.**

****

1. Interpret the scatter plot of the standardized predicted value (x) versus the standardized residual (y).
* **Key Observation:**

**The plots show 3 main regions: middle (very high concentrated), upper and lower (they both have very low density). Overall, all plots seem to be distributed evenly above and under the horizontal regression line starting from y=0 and they all head downward along the x-axis.**

* **Interpretations:**
* **Horizontal Regression Line at Y=0**: This line indicates where residuals equal zero, serving as a baseline for residual variance and showing perfect predictions (points on this line) versus over- or under-predictions (points above or below it).
* The scatter plot's dense clustering above and below the regression line indicates some variance around the predictions. If the density of residuals is relatively symmetric around the line, it could mean model predicts well on average, but with variability.
* Downward patterns among the residuals indicate some form of heteroscedasticity, implying that residual variance changes across the range of predicted values. Since all plots are pointing downward consistently, we can translate these findings as followed
* **Negative Correlation**: There is a negative correlation between the predictors (IVs) and the dependent variable, meaning higher values in the predictors may contribute to lower total family income.
* **Model Fit**: Since all patterns consistently follow the same downward direction, it indicates that the regression model captures a meaningful trend in the data.
* Exception or Potential Outliers: There are a few plots on both upper and lower regions that seem to be out of range and far apart from the regression line, this indicates that the model struggles to accurately explain the true value between data points. For example, this explanation between Highest Year of School Completed" and "Age of Respondent," would become difficult to fully explain (unless we add another factor) to correlate, therefore, it becomes outlier.
* **Conclusion:**

While the model appears to have predictive capability, further diagnostics (like exploring residual patterns, other plots, or adding variables) could enhance its performance and stability. Additional predictors might include other socio-economic factors or interactions to explain total family income more comprehensively.

1. **Final Conclusion**:
* **Significant Predictors**
* **Highest year of school completed**: This predictor is positively associated with total family income, as evidenced by its positive unstandardized coefficient (B = 0.767) and high statistical significance (p < 0.001). This suggests that higher educational attainment is linked to increased family income.
* **Age of respondent**: This variable has a small but negative effect on total family income (B = -0.035, p < 0.001), indicating that with increasing age, there is a slight decrease in predicted family income.
* **Marital status**: It shows a strong negative effect (B = -1.245, p < 0.001), suggesting that changes in marital status are significantly associated with lower family income.
* **Race of respondent**: While statistically significant (p = 0.001), the effect size is smaller (B = -0.543), indicating a relatively moderate influence on family income.
* **Model Fit and Collinearity Diagnostics**:
* The VIF values are all below 10 (ranging from 1.031 to 1.165), suggesting no multicollinearity concerns among the predictors.
* Collinearity diagnostics did not indicate serious collinearity issues based on tolerance values being above 0.1.
* **Interpretation of Model Coefficients:**
* The positive coefficient for the highest year of school completed implies that each additional year of education contributes to a notable increase in predicted total family income.
* The negative coefficient for age indicates that, controlling for other factors; older respondents have slightly lower predicted family incomes, which could relate to various socio-economic factors.
* The negative impact of marital status suggests complexities, such as differences between single, divorced, or widowed individuals, on income.
* The negative coefficient for race suggests disparities that may need further exploration for their socio-economic causes**.**
* **Conclusion**:

The regression model shows that as individuals achieve higher levels of education, their income is likely to increase, while age, marital status, and race have varying degrees of negative associations. The model appears statistically significant, and multicollinearity does not pose a major problem. However, there is room to explore interactions, non-linear relationships, or additional predictors to improve model fit further and address any potential residual biases.