**HUYNH DO**

**Module 6, Part 2/2**

1. **State the purpose of the exercise:**The purpose of this exercise is to use Multiple Linear Regression approach to determine how the following Independent variables (**IVs**): "Highest Year of School Completed," "Age of Respondent," "Marital Status," and "Race of Respondent" influence "Total Family Income”.

With that having said, in this exercise, we addressed the following 3 main concerns:

1. Understand the individual and combined contributions and the weight of the statistical significance of each **IV** to "Total Family Income” variable.
2. Evaluate the assumptions of Multiple Linear Regression to ensure the validity of the model.
3. Identify any potential outliers or exception data points that might impact the validity of the outcomes.
4. **Key Findings:**
5. **Regression Summary**

* The coefficients for each predictor allowed us to interpret their relative contributions. For example, "Highest Year of School Completed" had a positive coefficient, suggesting that higher educational attainment is associated with increased family income. Conversely, variables like "Age of Respondent" and "Marital Status" had negative coefficients, indicating a potential decrease in income depending on age or marital status changes.

1. **Model Fit**

* The overall model showed statistical significance, which suggests that the predictors, when taken together, explain a meaningful portion of the variance in family income.
* However, despite the statistical significance, the fit of the model must be interpreted with caution, as there were signs of variability and some predictors had small effect sizes.

1. **Interpretation of Coefficients**:

* "Highest Year of School Completed": A higher number of years completed was associated with an increase in total family income. The unstandardized coefficient provided an estimate of the income increase per additional year of education.
* "Age of Respondent": The negative coefficient indicated that, on average, older respondents had slightly lower family income, potentially reflecting factors like retirement or changes in work capacity.
* "Marital Status" and "Race of Respondent": These variables also displayed significant effects on family income, with their negative coefficients suggesting specific subgroups may experience reduced income compared to the reference groups.

1. **Collinearity Diagnostics**:

* Collinearity statistics, including Variance Inflation Factors (VIF) and tolerance values, were within acceptable limits, indicating minimal multicollinearity among the predictors. This is important for the stability and interpretability of regression coefficients.
* The collinearity diagnostics table suggested some correlation among predictors but no severe issues found.

1. **Residual Analysis and Assumption Checking**:

* A horizontal regression line at y=0 indicated no systematic patterns in the residuals, suggesting that the linear model was appropriate for the data. However, a few outliers were identified with outstanding residual (y-axis) values, potentially influencing the model's estimates. These outliers were fully addressed in module 6A and they were just anomaly data.

1. **Conclusions:**

* All predictors (Independent variables) used in this model had statistically significant effects on total family income, though the magnitude and direction of these effects varied. For instance, education positively impacted income, while age and certain demographic factors had an opposite (negative) impact.
* Although there are a few outliers and variability in the residuals, but the model fit was overall proven reasonable. Also, the assumptions of Multiple Linear Regression were mostly met, with no major violations detected.