# Part 5: Try it yourself:

**Looking at the results returned by ‘results\_model2 = model2.summary()’, and try to interpret the output.**

Table

Description automatically generated

**Questions:**

(1) What is the dependent variable and what are the independent variables in the model?

The dependent variable are charges, and the independent variables are age, male, children, smoker, and BMI.

(2) What does the Intercept of -1.205e+04 mean?

The intercept represents the expected value of charges when BMI, age, male, children, and smoker are set to zero.

(3) Which predictor variables are statistically significant in predicting charges?

Age, BMI, children, and smoker because they all have a p-value <0.10.

(4) What is the coefficient of children? Is this coefficient significant? What does the coefficient mean?

474.4111. Yes, it is because p is 0.000 which is less than 0.10. When the age, male BMI, and smoker are unchanged, and children increase the charges increase by 474.111.

(5) What are the four critical diagnostics for linear regression we talked about in the class? Are the estimates reliable based on results of these critical diagnostics?

Regression Diagnostic 1: Multi-collinearity issue. Based on the results of the correlation matrix, we didn’t see a high correlation between any two independent variables. This model does not have a multi-collinearity issue.

Regression Diagnostic 2: Heteroscedasticity in the error term. Based on the results of the white test, this model does not have a Heteroscedasticity issue, as the p-value of the test is less than 0.10.

Regression Diagnostic 3: this is not time-series data. Concern about this issue is minimal.

Regression Diagnostic 4: Model specification errors

1. The model does not include any “core” variables. No, many relevant factors are not included in the model. For example, drinking habits, because can influence both BMI and charges.
2. The model does not include superfluous variables. Yes this model does not
3. The functional form of the model is suitably chosen. Yes, as the dependent variable is a continuous variable, linear regression is suitable.
4. There are no errors in measurement in the regressed and the regressor. Not sure how variables are measured.
5. Outliers in the data if any are taken into account. According to the summary statistics, the concern about outliers is minimal in this model. (No extremely large or extremely small values for the dependent and independent variables.)
6. The probability distribution of the error term is well specified. We check whether the residuals are normally distributed. As the p-value is less than 0.00 evidence shows that the residuals are not normally distributed.

6) Based on the results, can we say “Bmi” has an “impact” on charges? Why?

No. The results show correlation only, as there could be variables that were omitted or neglected. Diet also can be a factor that can influence BMI and high charges, but it is not controlled by the model.