Assignment 2

# Scenario

John is having an animated discussion with his colleague, Mary, about the *importance* of total Cholesterol level and its relationship with BMI. Mary emphasises that literature focuses on the level of LDL (low density lipoprotein) cholesterol level, *not* total cholesterol. John suggests analysing the HERS dataset that contains these variables to obtain a proper answer to this question. Consider providing your syntax and output. Round to three decimal places (unless otherwise stated) and assume a priori alpha for p<0.05.

In this exercise, you will be helping John and Mary to arrive at a reasonable conclusion.

***Paste your full syntax at the bottom of this assignment.***

# Data Prep

Open the HERS dataset. Find two variables “tchol1” labeled as “Year 1 total cholesterol (mg/dl)”, “LDL1”, labeled as” year 1 LDL cholesterol (mg/dl)”, and BMI (kg/m2).

Hint: Working with Datasets with Many Variables

Sometimes when working in datasets with many variables, it can be cumbersome to try to find the variables you need. There are a few things that can make your life easier.

**First**, if you need to keep all the variables: You can use the search box above the variables list to quickly jump to the variables you are looking for. Note, you can also search for parts of variable names.

Graphical user interface, text, application

Description automatically generated

**Second**, if you know ahead of time exactly which variables you will need and which you will not need, you can drop the variables you will not need. If you do this, be careful not to over-wright the original data file! You can either save as a new data set, or just keep the data prep syntax to configure the original dataset to your needs in future analyses.

# Questions

### Q1. Carry out a simple linear regression with BMI as the outcome variable and Cholesterol level as the predictor.

Report the **correlation coefficient** between BMI and cholesterol level (“Year 1 total cholesterol (mg/dl)”) Round your response to three decimal places.

(HINT: Remember that the Coefficient of determination (R2) is simply the square of the correlation coefficient in a simple regression). Enter the correlation coefficient in the box.

### Q2. What proportion of BMI is explained by cholesterol level? (Express as a percent, e.g., 0.129 would be expressed as 12.9)

### Q3. Now carry on a simple linear regression with BMI as the outcome variable and LDL level as the predictor.

Report the **correlations coefficient** between BMI and LDL level (use the LDL variable LDL1, labeled as year 1 LDL cholesterol (mg/dl)). Round your response to three decimal places. Enter the correlation coefficient in the box.

### Q4. What proportion of BMI is explained by LDL level? (Express as a percent, e.g., 0.129 would be expressed as 12.9)

### Q5. Assuming a priori alpha of p<0.05 which of the following statements is correct?

Select one

1. Cholesterol is a significant predictor of BMI, but LDL is not
2. LDL is a significant predictor of BMI, but Cholesterol is not
3. Neither LDL nor Cholesterol are statistically significant predictors of BMI
4. Both LDL and Cholesterol are statistically significant predictors of BMI
5. We do not have enough information to determine

### Q6. Based on the above correlation coefficient and coefficient of determination for the relationship between BMI and Cholesterol or LDL, who was correct (Mary or John) in their assertion that cholesterol versus ldl was a “more important” predictor of bmi, and why?

**Hint:** when reporting your answer, support any statements of fact with the appropriate statistical results.

### Q7. Now come at this from a different perspective: interpret the relationship between BMI and cholesterol level based on the regression coefficient.

For ***ten*** unit change in cholesterol level, how much change should we see in BMI? (round to two decimal places). Report the P value associated with the t-test in this relationship. State as a complete sentence with all critical information provided.

**Hint:** Sometimes a one unit change in the predictor variable is clinically meaningless (when is the last time you had a physician concerned over a 1-point change in your serum cholesterol?). So, it may be more clinically relevant to interpret the relationship between the change in the predictor (X) and the outcome (Y) in terms of a larger unit of change. E.g., to interpret the effect of a 100 unit change in X, simply multiply the coefficient by 100.

### Q8. Interpret the relationship between BMI and LDL level.

For **ten** unit change in LDL level, how much change should we see in BMI? (round to two decimal places). Report the P value associated with the t-test in this relationship.

## Q9. Generate a scatterplot between BMI and LDL and include a linear fit line on top of the scatterplot.

### Q9.1. Paste the graph below

### Q9.2. Paste the syntax for the above graph below

### Q10. In a simple linear regression, which estimate determines whether the predictor variable has a statistically significant relationship with the outcome variable?