**Module 3 HW**

**DUE: Monday September 16th by 11:59 PM**

Professor Nick Williams Fall 2024

Economics 4010, University of Cincinnati

## **What are you being asked to do:**

Estimate and interpret bivariate regression models, using an AI model to generate the needed R code.

## Why is this important?

Learning how to use and interpret the bivariate regression model is the first step in mastering regression. You need to master regression, before we can reach our goal of estimating causal relationships in economics.

## Directions

* Turn in
  + Answers in a Word document through Canvas.
    - **IMPORTANT: Make sure you read and follow my directions in the companion Word document “Practicing Professionalism” that I distributed with the Module 2 HW**
    - In some circumstance, copying results from RStudio into Word is acceptable, but note that in many instances I ask you to interpret or explain. Below I make it clear when I want up to write an answer.
  + A copy of your R script through Canvas
    - This should NOT be a copy of your Console, but a copy of your R Script that I can run.
    - This MUST be in a separate text file, NOT a Word document.
* Grading
  + As stated in the syllabus, homework will be accepted one day late with a 25% penalty. Later homework will not be accepted.
  + As stated in the syllabus: Feel free to work with others. However, I recommend that you spend time working on the problems by yourself before working with others. **Every student must turn in their own, *unique* answers to the homework.**
  + The homework will be carefully graded out of 100 total points.
* Unlike the learning exercise and homework in Module 2, we will continue our efforts moving forward by utilizing an AI model to help with the R code. To that end, you should use some AI model (ChatGPT, Claude, CoPilot, etc.) to generate the needed R script to answer the questions below. One possible way to do this is with the following prompt:

I am an undergraduate student in econometrics. I have a homework that asks me to estimate bivariate regression models using R and then interpret the results.

Use the data set *county\_health\_subset\_2023.dta* to answer this question. The data frame contains information about health and related factors for almost all of the counties in the US. Variables include:

---------------------------------------------------------

Variable Label

------------- ------------------------------------------

statecode State FIPS Code

countycode County FIPS Code

fipscode 5-digit FIPS Code

county Name

income Median household income raw value, $

teenbirths Num births per 1000 female population, ages 15-19

uerate Unemployment rate

Packages that might be helpful in generating the results include

c("AER","car", "gmodels", "haven", "ivreg", "jtools", "pastecs", "plm", "psych", "readxl", "skedastic", "stargazer", "summarytools","tidyverse")

I will enter the explicit questions I need answered below. Please create the R code to allow me to answer them

* Typically 3 or 4 significant digits is adequate for our work. However, for this assignment ONLY make sure you set the significant digits to 8. This can be done by using the R command:

options("jtools-digits"=8)

## Question

Use the data set *county\_health\_subset\_2248.dta* to answer this question. The data frame contains information about health and related factors for almost all of the counties in the US.

* 1. Look at the descriptive statistics of the data. Make sure you understand the units of *teenbirths, income and uerate.*
  2. Create a histogram of *teenbirths.* 
     1. Turn in this graph by pasting it into your Word document.

A graph of a number of birth rate

Description automatically generated

* + 1. From the histogram, and the descriptive statistics in part a), how would you describe the distribution of *teenbirths*?

Its skewed to the left with an average of 23. This means that there are on average 23 births per 1000 women (age15-19) in the counties of US. The median is 21. Which means that there are probably some outliers on the higher end bringing the average up.

* 1. Estimate the equation below by ordinary least squares (OLS), and turn in your results by pasting them into your Word document. Make sure that you set the options("jtools-digits"=8) in your R program!
  2. For the regression in part c)
     1. Interpret the slope coefficient reported in the regression in part c). This means provide a numerical interpretation of what the coefficient means.

So this means for every teen birth income decreases by $-0.00045593 on average.

* + 1. Briefly discuss what from this regression says about the amount of the variation in birthweight that is explained by median income in a county.

R^2 = 0.422 which means 42.2% of the variability of Teen births can be explained by income, which is less than half. So there is a whole other 57% that is inexplicable via this model

* 1. Sometimes it is easier to interpret regressions (and more sensible) if we change the units of a variable.
     1. Create a new variable representing median income in thousands of dollars. Call this variable *income\_th.*
     2. Create a scatter plot with *teenbirths* on the y-axis and *income\_th* on the x-axis.Turn in this graph by pasting it into your Word document.

A graph showing a number of dots

Description automatically generated

* + 1. From this scatter, is there an observable relationship? Briefly explain.

Yeah the relationship seems more observable. It seems as though there is a negative slope between teen births and income, a relatively sharp downward at that once income was changed to thousands of dollars. I almost think it would be better explained by a exponential

* 1. Estimate the equation below by ordinary least squares (OLS), and turn in your results by pasting them into your Word document.

MODEL INFO:

*Observations:* 2927

*Dependent Variable:* teenbirths

*Type:* OLS linear regression

MODEL FIT:

*F*(1,2925) = 2138.35706280, *p* = 0.00000000

*R² =* 0.42232002

*Adj. R² =* 0.42212252

*Standard errors:OLS*

------------------------------------------------------------------------

Est. S.E. t val. p

----------------- ------------- ------------ -------------- ------------

(Intercept) 51.91932097 0.64428776 80.58405567 0.00000000

income\_th -0.45593434 0.00985967 -46.24237302 0.00000000

------------------------------------------------------------------------

* + 1. Interpret the slope coefficient reported in the regression in part f). Compare your results to those obtained in part c). This means provide a numerical interpretation of what the coefficient means. The slope coefficient is -0.45593434 which makes sense because we increased the scale of income.
    2. Compare the from the regressions in part c) with those of part f).

The R^2 value is the same, as none of the data has changed, we just changed the scale.

* 1. There are many reasons to suspect that the estimate of is biased and may not reflect a causal relationship. One likely problem is that *income\_th* may be endogenous in the regression in part (f). Briefly explain what this means and give an example of why this might be true.

This means that the variables are affected by another outside variable that we are not considering. For example, maybe in cheaper areas there are more teen pregnancies because it is cheaper to raise children vs in high cost areas (where salaries are also higher) there are less teen pregnancies.

* 1. Estimate the equation below by ordinary least squares (OLS),
     1. turn in your results by pasting them into your Word document.

MODEL INFO:

*Observations:* 2927

*Dependent Variable:* teenbirths

*Type:* OLS linear regression

MODEL FIT:

*F*(1,2925) = 300.34115795, *p* = 0.00000000

*R² =* 0.09311919

*Adj. R² =* 0.09280915

*Standard errors:OLS*

-----------------------------------------------------------------------

Est. S.E. t val. p

----------------- ------------- ------------ ------------- ------------

(Intercept) 12.41123540 0.64935605 19.11314352 0.00000000

uerate 2.93011186 0.16907398 17.33035366 0.00000000

-----------------------------------------------------------------------

* + 1. interpret the slope coefficient reported in the regression. This means provide a numerical interpretation of what the coefficient means.

The slope coefficient of 2.9 means that for every increase of 1% in unemployment rate, there are 2.9 more teen births per 1000 women.

* 1. For the regression in part h)
     1. What is the average of the predicted values for this model? (Remember that predicted values are another name for fitted values.) How does this compare to the mean of *teenbirths*?

The mean of teen births is 23 and the predicted value is 27, the predicted value is slightly above what the actual mean is

* + 1. What is the average of the residual values for this model?

1.387884e-15