Contributors: Aashish Singh, Alexander Saucedo, Prinu Mathew, Nyckeisha' Sam, Qingwei Zhang

MSCA 31007: Statistical Analysis

Mondays 6-9pm

Group Assignment 1 – Homework

Autumn 2022

#\*DP02\_0065P -> propbac (Bachelor's degree)

#\*DP03\_0062 -> medhhinc (Median household income)

#\*DP03\_0096P -> propcov (Health insurance coverage)

#\*DP03\_0128P -> proppov (PERCENTAGE OF FAMILIES AND PEOPLE WHOSE INCOME IN THE PAST 12 MONTHS IS BELOW THE POVERTY LEVEL)

#\*DP04\_0047P -> proprent (Renter-occupied)

#\*DP05\_0001 -> totpop (Total population)

#\*DP05\_0018 -> medage (Median age)

1. **Create a linear model object in which median household income explains baccalaureate attainment rates at the tract level, using the lm() command. Summarize the model object using the summary() command.**

**H0 (null hypothesis)**: Baccalaureate attainment rates doesn’t depend on median household income

Radj2 (Adjusted R-squared) = 1 – [(1 - R2) \* (n-1)]

n-k-1

R2 = R squared of the model

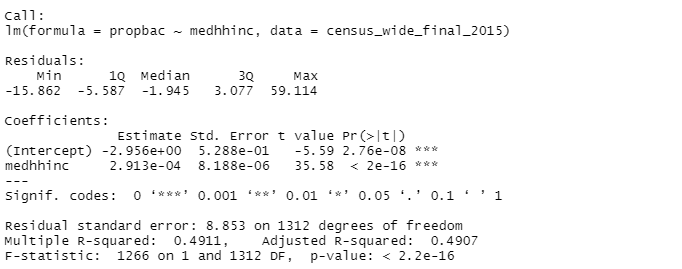
N = Observation

K = Independent variables

As ‘k’ increases, Adjusted R square will always fall unless there is a significant increase in the R square value. Basically, what the formula does is, the more decent or correct independent variable you add (i.e., the independent variable that fit the model which caused the R squared to increase), the Adjusted R square will increase. But if you add incorrect independent variable, the R square might go up, but the Adjusted R square will always fall

So Adjusted R square will help you to decide whether the regression model is fit w.r.t to independent variables that you choose

*Linear regression summary showing the relationship between medhhinc (*independent *variable) & propbac (*dependent *variable) for 2015 population from Cook County, IL*



df = 1312 because we have two estimated parameters: slope and intercept

Since **p-value = ~0.0000** that corresponds to **t = 35.58** is less than some threshold (e.g., α = .05) then we reject the null hypothesis and conclude that there is a statistically significant relationship between the *medhhinc* variable and the *propbac* variable

Here **R-squared is 0.4911**, **Adjusted R-squared is 0.4907**. Adjusted R-squared is slightly lower than R-square which indicate regression model fits w.r.t to *medhhinc variable* that we choose