Confounding and Interaction

- Interaction:
    - When the relationship between 2 variables of interest is different at different levels of another extraneous variable(s)
    - Interaction should be assessed before confounding is assessed.

- Confounding:
    - When the relationship between 2 variables of interest is *meaningfully* different when another extraneous variable(s) is included (vs. not included) in the model

    - An extraneous variable may be kept in a model in order to control for confounding.

- Interaction Modeling – 3 approaches
    - Include only interactions that are anticipated a-priori based on previous studies
        - Challenge: a-priori information about potential interactions may be limited

    - Include all possible interactions
        - For any higher-order interaction included, typically include the corresponding lower-order interactions
        - E.g., predicting the dependent variable (DV) based on 3 independent variables (IVs)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \beta_6 X_2 X_3 + \beta_7 X_1 X_2 X_3 + \varepsilon$$

        - Challenge: the number of terms may be too large for reliable estimates of all effects

    - Include only interactions with the DV and *primary* IV(s) – a compromise for reliability
        - E.g., suppose X_1 is the primary IV and X_2 & X_3 are of less interest

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3 + \varepsilon$$

- <u>Controlling for Confounding</u>

  Let $C_1,\ldots,C_p$ be p potential confounders and $\mathbf{C}^*$ & $\mathbf{C}^\#$ be different subsets of the C_i

  Let $\hat{\beta}_{1|Ci}$ be the value of the regression coefficient for X_1 after including C_i in the model

  Let $\hat{\beta}_1$ be the value of the regression coefficient for X_1 alone

  - Confounding exists if $\hat{\beta}_{1|C}$ is *substantially* different from $\hat{\beta}_1$

    r_{YX1} is *substantially* different from $r_{YX1|C}$

  - Identifying a minimal set of confounders to control for
    - Determine a baseline estimate for the relationship of interest $\hat{\beta}_{1|\mathbf{C}^*}$
    - Any subset of confounders $\mathbf{C}^\#$ where $\hat{\beta}_{1|\mathbf{C}\#} \approx \hat{\beta}_{1|\mathbf{C}^*}$ provides equivalent control
    - Choose a subset with equivalent control that leads to high precision

```
> EDUC2 <- dat$EDUC^2
> lm(LSTART ~ EDUC, data=dat)$coef        #(Intercept)        EDUC
                                          #      7.334       0.105
> lm(LSTART ~ EDUC+EDUC2, data=dat)$coef  #(Intercept)        EDUC       EDUC2
                                          #      9.715      -0.250       0.013
> lm(LSTART ~ EDUC+AGE, data=dat)$coef    #(Intercept)        EDUC         AGE
                                          #      7.049       0.113       0.005
> lm(LSTART ~ EDUC+JOBCAT, data=dat)$coef #(Intercept)        EDUC      JOBCAT
                                          #      7.675       0.049       0.183
  L.A <- lm(LSTART ~ AGE,    data=dat)
  L.J <- lm(LSTART ~ JOBCAT, data=dat)
  E.A <- lm(EDUC   ~ AGE,    data=dat)
  E.J <- lm(EDUC   ~ JOBCAT, data=dat)
cor(dat$LSTART, dat$EDUC) #[1] 0.7662246
cor(L.A$resid, E.A$resid) #[1] 0.7680637
cor(L.J$resid, E.J$resid) #[1] 0.5415896
```