• Endogeneity
(1) weight 1-year ago
(2) current weight
(3) followed diet

$$\underbrace{(y_{i1} - y_{i0})}_{\text{difference in weight}} = \alpha + \underbrace{\beta d_i}_{\text{diet}} + \underbrace{\gamma y_{i0}}_{\text{last year weight}} + \underbrace{x_i'}_{\text{other variables}}\delta + \varepsilon_i$$

(a) why would diet be endogenous?
• Because it is likely that the diet variable is correlated with the error term. The source of endogeneity in this case is probably ommitted variable bias. It is very likely that there is a variable, not specified in the model that is associated with following the diet AND associated with the outcome.
In this case, the OLS estimates are inconsistent, meaning that it does not converge to the population parameter even in large samples.
An ommitted variable could simply be "overall desire to lose weight". Subjects that are really interested in losing weight could self-select into following the diet ($cov(x,\varepsilon)>0$) and also to adhere to other kinds of healthy lifestyle ($cov(\varepsilon, y)>0$). Therefore, in this case, we are overestimating the true effect of the diet

(b) $Z_i = \begin{cases} 1 & \text{advertisement} \\ 0 & \text{no advertisement} \end{cases}$

b.i) $1/n\, Z'\varepsilon \to 0$. this condition states that the correlation between the instrument and the error term converges to 0 as the sample size grows. Within the application, it simply requires that there is no ommitted variable that could be jointly correlated with the outcome and with advertisement. If that was the case, advertising would also be endogenous

b.ii) $1/n\, Z'X \to Q \neq 0$. this condition states that as the sample size grows, the instruments and the endogenous variables should be sufficiently correlated. Otherwise, it would be difficult to use Z as an instrument to estimate the causal effect of X. In the application, advertisement should be correlated with following the diet. Otherwise, the variable would not explain any of the variation in X (it would be mostly explained by the error term. We are back to square one)

(c)

c.i) It is possible to test this requirement statiscally given the available variables. We could simply rely on the outcome of a simple regression of the endogenous variables, the instrument, and the remaining exogenous variables. If the coefficient for the instrument is statiscally significant.

c.ii) Is also possible given the available variables. Because we could use a Sargan test, which requires that the number of instruments to be larger than the number of parameters in the OLS estimation. Besides advertisement, we could use the other exogenous variables as instruments, which could suffice as requirements for our purposes.

cd) $Z_i \perp\!\!\!\perp y_{i0}$ and x_i

$(y_{i1} - y_{i0}) = \alpha + \beta d_i + \eta_i$

$$\beta_{2SLS} = (Z'X)^{-1} Z'y$$

· assume a model with no intercept.

$$b_{2SLS} = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{\sum_i (z_i - \bar{z})(x_i - \bar{x})}$$

$$\cdot\ b_{2SLS} = \frac{\sum_i (z_i y_i - \bar{z} y_i - z_i \bar{y} + \bar{z}\bar{y})}{\sum (z_i x_i - \bar{z} x_i - z_i \bar{x} + \bar{z}\bar{x})}$$

$$= \frac{\sum z_i y_i - \bar{z} \sum_i y_i - \bar{y} \sum_i z_i + \sum_i \bar{z}\bar{y}}{\sum_i z_i x_i - \bar{z} \sum x_i - \bar{x} \sum_i z_i + \sum_i \bar{z}\bar{x}}$$

· We can rewrite the terms

i) $\sum_i z_i y_i = \bar{d} \Delta_1$ iii) $\sum_i z_i \bar{y} = \bar{y} \sum_i z_i = \Delta \bar{d}$

ii) $\bar{z} \sum y_i = \Delta \bar{d}$ iv) $\sum_i \bar{z}\bar{y} = \bar{z}\bar{y} \sum_i 1 = \Delta \bar{d}$

v) $\sum_i z_i x_i = \bar{d}^1$ vii) $\bar{x} \sum_i z_i = \bar{d}\bar{d}^1$

vi) $\bar{z} \sum x_i = \bar{d}\bar{d}^1$ viii) $\sum \bar{z}\bar{x} = \bar{d}\bar{d}^1$

$$b_{2SLS} = \frac{\bar{d}\Delta_1 - \Delta \bar{d}}{\bar{d}^1 - \bar{d}\bar{d}^1} = \frac{\bar{d}(\Delta_1 - \Delta)}{\bar{d}^1(\bar{d} - 1)} \ //$$