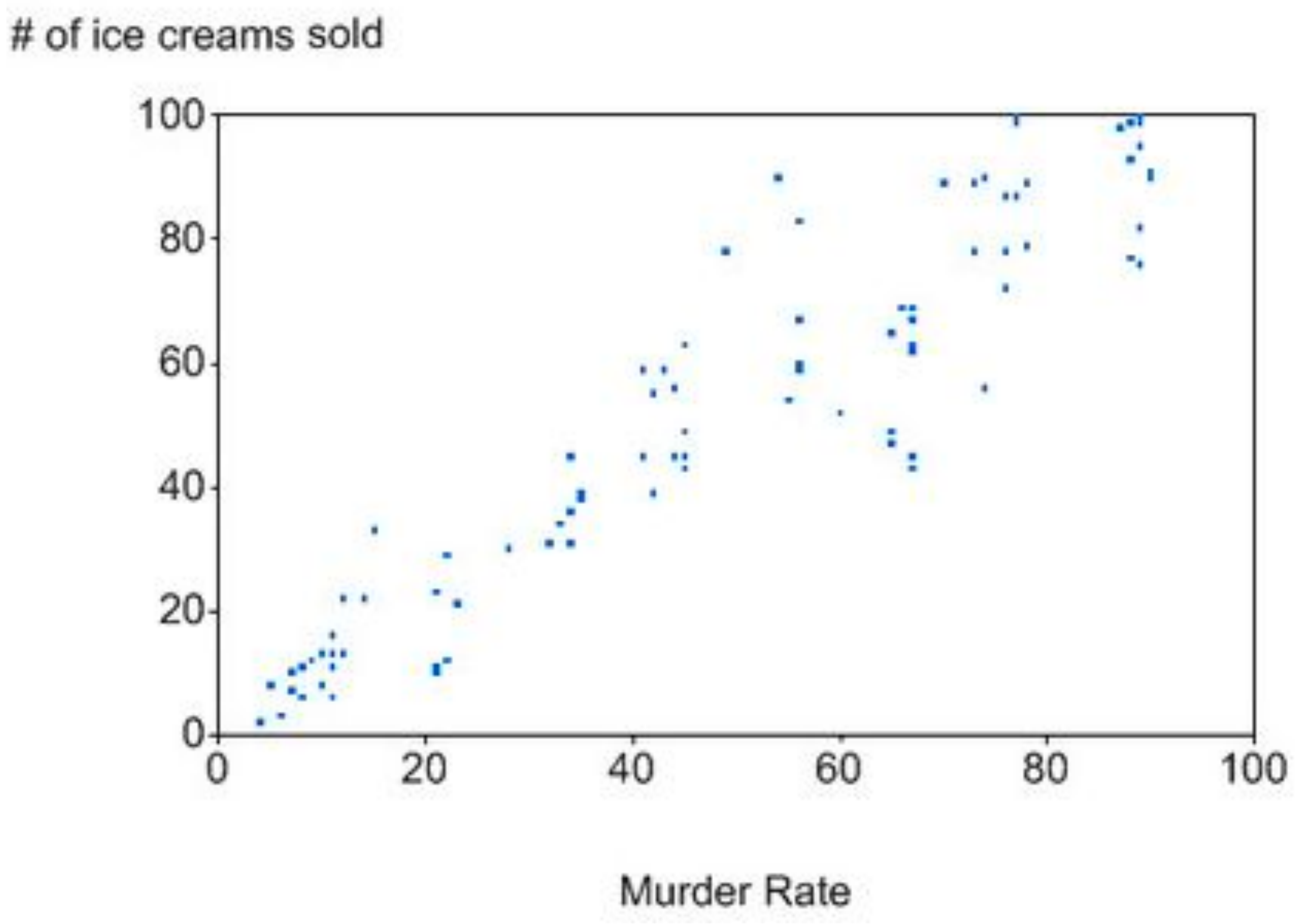
IDS 702

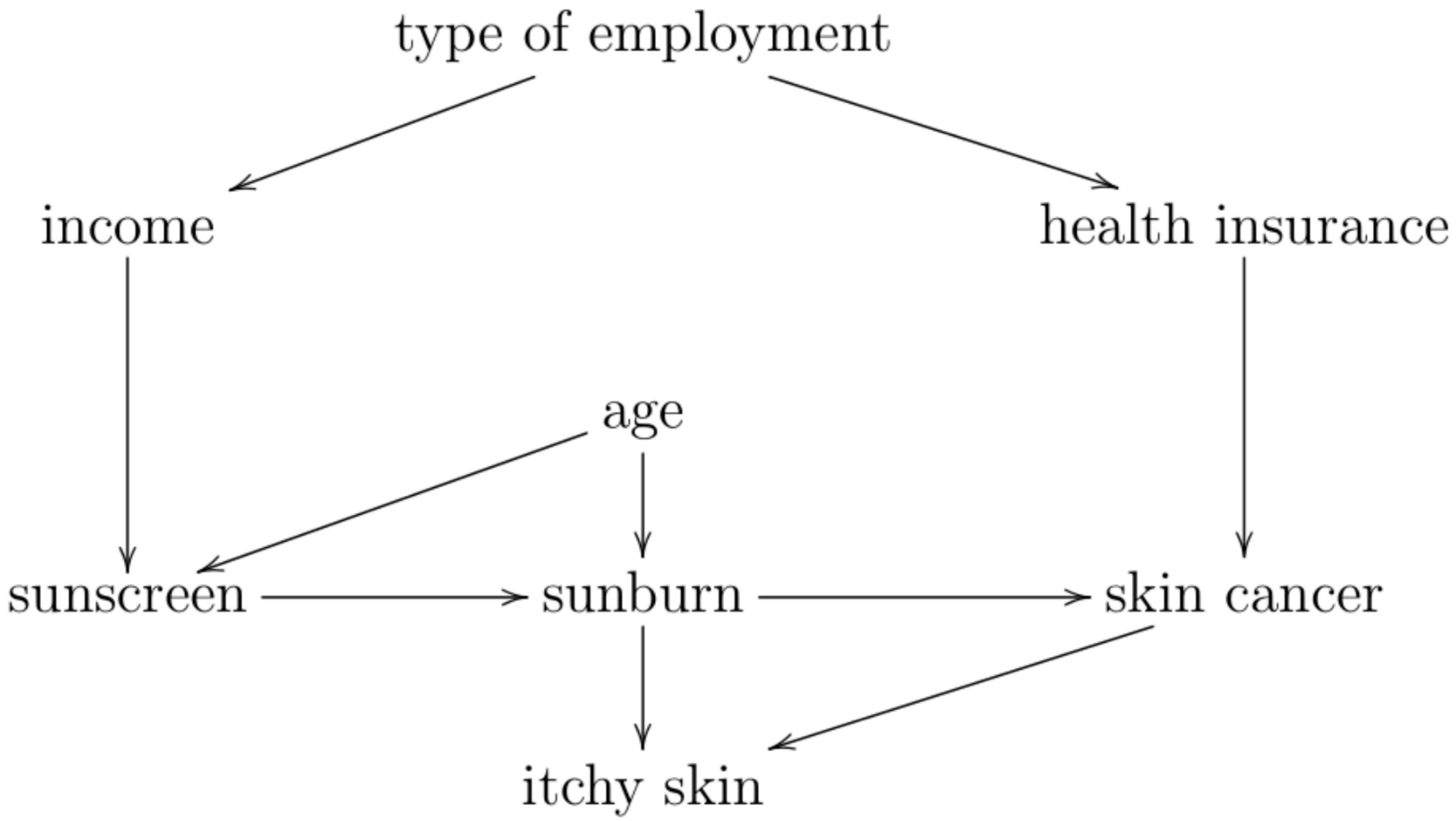
Introduction to multiple linear regression

Most relationships cannot be fully explained by two variables

- **Confounding variables** are related to both variables of interest and explain (at least) some of the relationship between them



Directed Acyclic Graph (DAG)



https://health.ucdavis.edu/ctsc/area/Resource_Library/documents/directed-acyclic-graphs20220209.pdf

Multiple Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i ; \epsilon_i \overset{\text{iid}}{\sim} N(0,\sigma^2), i = 1,...n$$

We can also write the model as:

$$y_i \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, \sigma^2)$$

$$E[Y \mid X_1 = x_1, \ldots, X_p = x_p] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Matrix representation

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon; \epsilon \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

Interpreting coefficient estimates

- Each estimated coefficient is the amount Y is expected to increase when the value of the corresponding predictor is increased by one unit, *holding the values of the other predictors constant*

- What if the predictor is not continuous? Find out in the next video!

Which variable is the strongest predictor of the outcome?

- The coefficient that has the strongest linear association with the outcome variable is the one with the largest absolute value of T (test statistic), which equals the coefficient estimate over the corresponding SE
- Note: T is NOT the size of the coefficient and the size of the coefficient is not indicative of strength of association with the outcome
- Coefficients are sensitive to the scale of the predictors, but T is not

Inference: F test for overall association

Is there a relationship between the predictors (taken together) and the response?

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Inference: individual coefficient estimates

$$H_0 : \beta_j = 0$$ (GIVEN all the other variables in the model)

$$T = \frac{estimate - Null}{SE} = \frac{\hat{\beta}_j - 0}{SE(\hat{\beta}_j)}$$

$$CI = \hat{\beta}_j \pm SE(\hat{\beta}_j)C_\alpha$$