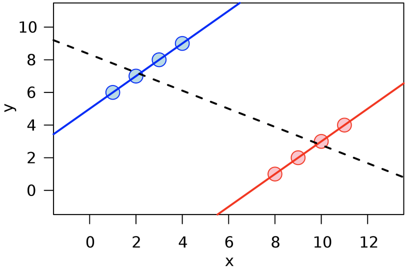
Confounding

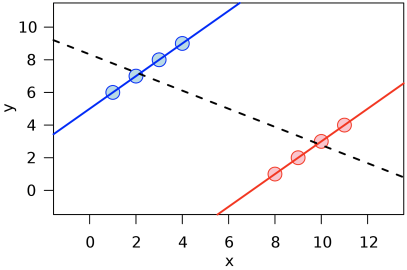
This slide discusses an remaining issue in Notes Nov 14 on Broken-stick regression that we do not have time to finish today.

- Pupose of Broken-stick regression: make the prediction continuous over covariate values.
  - This may be not always needed.
- Simpson's paradox: a trend appears in several groups of data but disappears or reverses when the groups are combined.



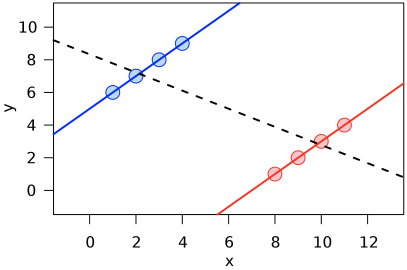
This slide discusses an remaining issue in Notes Nov 14 on Broken-stick regression that we do not have time to finish today.

- Pupose of Broken-stick regression: make the prediction continuous over covariate values.
  - This may be not always needed.
- Simpson's paradox: a trend appears in several groups of data but disappears or reverses when the groups are combined.



This slide discusses an remaining issue in Notes Nov 14 on Broken-stick regression that we do not have time to finish today.

- Pupose of Broken-stick regression: make the prediction continuous over covariate values.
  - This may be not always needed.
- Simpson's paradox: a trend appears in several groups of data but disappears or reverses when the groups are combined.



Confounder

- Confounder (Confounding variable):
  - A variable such as Z that is associated with both the dependent and independent variables in a regression model.
- Suppose

$$
\begin{aligned}
Y =& \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_Y \\
Z =& \gamma_0 + \gamma_1 X + \epsilon_Z
\end{aligned}
$$

- But we fit $Y \sim 1 + X$. Then

$$Y = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)X + \epsilon_Y + \beta_2\epsilon_Z$$

- Coefficient of X: $\beta_1 + \beta_2\gamma_1$ can differ from β_1 a lot.
  - Example: $\beta_1 = -2$ and $\beta_2\gamma_1 = 3$,
    - Regress $Y \sim X + Z$, coefficient of X is -2
    - Regress $Y \sim X$, coefficient of X is $-2 + 3 = 1$

Confounder

- Confounder (Confounding variable):
  - A variable such as Z that is associated with both the dependent and independent variables in a regression model.
- Suppose

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_Y$$
$$Z = \gamma_0 + \gamma_1 X + \epsilon_Z$$

- But we fit $Y \sim 1 + X$. Then

$$Y = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)X + \epsilon_Y + \beta_2\epsilon_Z$$

- Coefficient of X: $\beta_1 + \beta_2\gamma_1$ can differ from β_1 a lot.
  - Example: $\beta_1 = -2$ and $\beta_2\gamma_1 = 3$,
    - Regress $Y \sim X + Z$, coefficient of X is -2
    - Regress $Y \sim X$, coefficient of X is $-2 + 3 = 1$

Confounder

- Confounder (Confounding variable):
  - A variable such as Z that is associated with both the dependent and independent variables in a regression model.
- Suppose

$$\begin{aligned} Y =& \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_Y \\ Z =& \gamma_0 + \gamma_1 X + \epsilon_Z \end{aligned}$$

- But we fit $Y \sim 1 + X$. Then

$$Y = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)X + \epsilon_Y + \beta_2\epsilon_Z$$

- Coefficient of X: $\beta_1 + \beta_2\gamma_1$ can differ from β_1 a lot.
  - Example: $\beta_1 = -2$ and $\beta_2\gamma_1 = 3$,
    - Regress $Y \sim X + Z$, coefficient of X is -2
    - Regress $Y \sim X$, coefficient of X is $-2 + 3 = 1$

Confounder

- Confounder (Confounding variable):
  - A variable such as Z that is associated with both the dependent and independent variables in a regression model.
- Suppose

$$
\begin{aligned}
Y =& \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_Y \\
Z =& \gamma_0 + \gamma_1 X + \epsilon_Z
\end{aligned}
$$

- But we fit $Y \sim 1 + X$. Then

$$Y = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)X + \epsilon_Y + \beta_2\epsilon_Z$$

- Coefficient of X: $\beta_1 + \beta_2\gamma_1$ can differ from β_1 a lot.
  - Example: $\beta_1 = -2$ and $\beta_2\gamma_1 = 3$,
    - Regress $Y \sim X + Z$, coefficient of X is -2
    - Regress $Y \sim X$, coefficient of X is $-2 + 3 = 1$

Confounder

- Confounder (Confounding variable):
  - A variable such as Z that is associated with both the dependent and independent variables in a regression model.
- Suppose

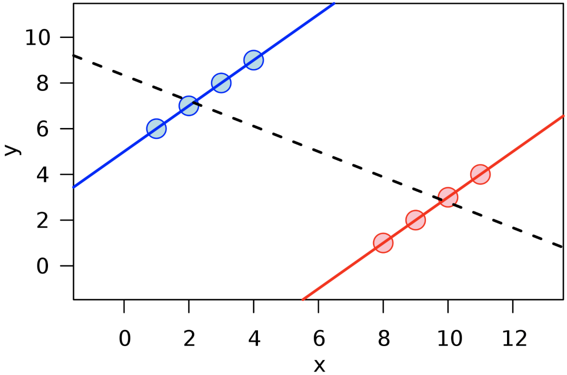
$$\begin{aligned} Y =& \beta_0 + \beta_1 X + \beta_2 Z + \epsilon_Y \\ Z =& \gamma_0 + \gamma_1 X + \epsilon_Z \end{aligned}$$

- But we fit $Y \sim 1 + X$. Then

$$Y = \beta_0 + \beta_2\gamma_0 + (\beta_1 + \beta_2\gamma_1)X + \epsilon_Y + \beta_2\epsilon_Z$$

- Coefficient of X: $\beta_1 + \beta_2\gamma_1$ can differ from β_1 a lot.
  - Example: $\beta_1 = -2$ and $\beta_2\gamma_1 = 3$,
    - Regress $Y \sim X + Z$, coefficient of X is -2
    - Regress $Y \sim X$, coefficient of X is $-2 + 3 = 1$

Simpson's Paradox



- Can be viewed as Z being a binary variable in the previous slide.
- Instead of using broken stick, one may want to find the confounder Z and put it in the regression for interpretation.