Parametrization of Factor Effects

Binary Predictor

\[ y_i = \# \text{ of bike trips on day } i \]
\[ x_i = \text{ indicator of rain on day } i-1, \ x_i \in \{0, 1\} \]

Linear Model

\[ E[y_i | x_i] = \beta_0 + \beta_1 x_i = \begin{cases} \beta_0 & \text{expected trips} = \mu_0 \\
\text{under no rain} & \\
\beta_0 + \beta_1 & \text{expected trips} = \mu_1 \\
\text{under rain} & \\
& \mu_1 - \mu_0 = \beta_0 + \beta_1 - \beta_0 \\
& \beta_1 = \text{difference in mean} = \text{"contrast" of } \mu_1, \mu_0. \\
\end{cases} \]

OLS estimation: given \( x, x \), obtain \( \{\hat{\beta}_0, \hat{\beta}_1\} \)

Guess:
\[ \hat{\mu}_0 = \text{mean}(y | x = 0) = \bar{y}_0 \]
\[ \hat{\mu}_1 = \text{mean}(y | x = 1) = \bar{y}_1 \]

Then:
\[ \begin{cases} \hat{\mu}_0 = \mu_0 \\
\hat{\mu}_1 = \mu_1 - \mu_0 \end{cases} \]
suggests
\[ \begin{cases} \hat{\beta}_0 = \hat{\mu}_0 = \bar{y}_0 \\
\hat{\beta}_1 = \hat{\mu}_1 - \hat{\mu}_0 = \bar{y}_1 - \bar{y}_0 \end{cases} \]

Are these the OLS estimates of \( (\beta_0, \beta_1) \)?
OLS Estimation: \( \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) = (X^TX)^{-1}X^T\gamma \)

\[
X = \begin{bmatrix} 1 & 1 & 1 \\
1 & 0 & 0 \\
1 & 0 & 1 \\
\end{bmatrix} \\
X^T\gamma = \begin{bmatrix} 1 & 1 & 1 \\
1 & 0 & 1 \\
\end{bmatrix} \begin{bmatrix} \gamma_1 \\
\gamma_2 \\
\gamma_3 \\
\end{bmatrix} = \begin{bmatrix} \Sigma \gamma_i \\
\Sigma \gamma_i; x_i \\
\end{bmatrix} = \begin{bmatrix} s \\
s_1 \\
\end{bmatrix}
\]

\[s_1 = \text{Sum} (\gamma \mid x = 1) \]

\[s = \text{Sum} (\gamma) \]

\[
(X^TX) = \begin{bmatrix} 1 & 1 & 1 \\
1 & 0 & 0 \\
1 & 0 & 1 \\
\end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\
1 & 0 & 0 \\
1 & 0 & 1 \\
\end{bmatrix} = \begin{bmatrix} n_0, n_1 \\
n_0, n \end{bmatrix}
\]

\[
(X^TX)^{-1} = \frac{1}{n_0, n_0} \begin{bmatrix} n_1, -n_1 \\
-n_1, n_1 \\
\end{bmatrix} = \frac{1}{n_0, (n-n_1)} \begin{bmatrix} n_1, -n_1 \\
n_0, n \end{bmatrix}
\]

\[
(\hat{\beta}_0, \hat{\beta}_1) = (X^TX)^{-1}X^T\gamma = \frac{1}{n_0, n_0} \begin{bmatrix} n_1, s-n, s_1 \\
ns_1-n, s \end{bmatrix}
\]

\[
\hat{\beta}_0 = \frac{1}{n_0, n_0} (n_1, s-n, s_1) = \frac{n_1}{n_0} (s-s_1) = s_0 \frac{s_0}{n_0}
\]

\[
\hat{\beta}_1 = \frac{1}{n_0, n_0} (ns_1-n, s) = \frac{1}{n_0, n_0} ((n-n_1)s_1 - n_1(s-s_1)) = \frac{nk}{n_0, n_0} s_1 - s_1 \frac{s_1}{n_0} = \bar{\gamma}_1 - \bar{\gamma}_0
\]
Alternative Parameterization

\[ x_i = \begin{cases} +\frac{1}{2} & \text{if rain} \\ -\frac{1}{2} & \text{if no rain} \end{cases} \]

\[ E[Y | X] = \alpha_0 + \alpha_1 x = \alpha_0 + \alpha_1 \frac{1}{2} \quad \text{under rain} \]
\[ E[Y | X] = \alpha_0 - \alpha_1 \frac{1}{2} \quad \text{under no rain} \]

\[ m_1 + m_0 = \frac{\alpha_0 + \alpha_0}{2} = \frac{\alpha_0}{2} \]
\[ m_1 - m_0 = \frac{\alpha_1 + \alpha_1}{2} = \frac{\alpha_1}{2} \]

\[ X = \begin{pmatrix} 1 & \frac{1}{2} \\ 1 & -\frac{1}{2} \\ 1 & \frac{1}{2} \\ 1 & -\frac{1}{2} \end{pmatrix} \]
\[ X^T Y = \begin{pmatrix} S \\ \frac{1}{2} (S_1 - S_0) \end{pmatrix} \]
\[ (\hat{\alpha}_0, \hat{\alpha}_1) = (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y \]

Guess:
\[ \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} + \bar{Y}_o / 2 \\ \bar{Y}_1 - \bar{Y}_o \end{pmatrix} \]

Check:

Let \( x_1 = 1 \) if rain, 0 otherwise
\( x_2 = 1 \) if no rain, 0 otherwise

\[ E(Y | X_1, X_2) = \hat{\gamma}_0 + \hat{\gamma}_1 x_1 + \hat{\gamma}_2 x_2 \]

\[ X = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{pmatrix} \]
\[ \hat{\beta} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y \]

\[ (\hat{X}^T \hat{X}) = \begin{bmatrix} n_1 & n_1 & n_0 \\ n_1 & n_1 & 0 \\ n_0 & 0 & n_0 \end{bmatrix} \]

\[ r_1 = n_1 - n_2 \]

Model is overparameterized:

\[ \text{not invertible!} \]
Try instead \( x_i = \text{rain indicator} \), \( x_0 = \text{no rain indicator} \)

\[
E(x_i, x_0) = \begin{cases} 
0 & x_0 = 1, x_i = 0 \\
1 & x_0 = 1, x_i = 0 
\end{cases}
\]

\[
\begin{bmatrix} 
0 & 1 \\
1 & 0 \\
\vdots & \vdots \\
0 & 1 \\
\end{bmatrix}
\]

\[
x^T x = \begin{pmatrix} n_0 & 0 \\
0 & n_1 \end{pmatrix} \quad (x^T x)^{-1} = \begin{pmatrix} \frac{1}{n_0} & 0 \\
0 & \frac{1}{n_1} \end{pmatrix} \quad x^T y = \begin{pmatrix} n_0 \bar{y}_0 \\
0 & \bar{y}_1 \end{pmatrix}
\]

\[
(x^T x)^{-1} x^T y = \begin{pmatrix} \bar{y}_0 \\
\bar{y}_1 
\end{pmatrix}
\]

\[
E(y | x) = \beta_0 + \beta_1 x, \quad x \in \{ 0, 1 \}
\]

\[
E(y | x) = \alpha_0 + \alpha_1 x, \quad x \in \mathbb{R} - \frac{1}{2}, \frac{1}{2} \}
\]

\[
E(y | x) = \mu_0 x + \mu_1, \quad x \in \mathbb{R}, \quad x \in \{ 0, 1 \}, \quad x_0 = 0
\]

All three models imply the same mean model for \( y \):
Two unrelated means, under each of the two \( x \)-conditions.

"Set to zero" parameterization: \( E(y | x) = \beta_0 + \beta_1 x \)

\( \beta_1 \) = "effect" of \( x \): Under one of the conditions (no rain), this is 0. (R-default)

"Sum to zero" parameterization: \( E(y | x) = \alpha_0 + \alpha_1 x \)

Effect under rain = \( \frac{x_1}{2} \)
No rain = \( -\frac{x_1}{2} \)

"ANOVA" parameterization: \( E(y | x) = \alpha_1 x \)