4.6 Potential Issues

Categorical predictors ~~take~~ on finitely many unordered values, sometimes called factors.

$\boxed{\text{Ex}}$ If x_i = male/female has only two levels.

Need a convention to numerically code x, using a dummy variable, or indicator variable [or in ML, one-hot encoding]. E.g.,

$$x_i := \begin{cases} 1 & i\text{th person female} \\ 0 & \text{" " male} \end{cases}$$

Then

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & i\text{th person female} \\ \beta_0 + \varepsilon_i & \text{" " male} \end{cases}$$

where

$$\beta_0 = \text{Avg response for males}$$

$$\beta_0 + \beta_1 = \text{'' '' '' females}$$

$$\beta_1 = \text{Change in avg resp. for females vs. males}$$

Note we could have coded

$$x_i = \begin{cases} 1 & i\text{th person female} \\ -1 & \text{'' '' male} \end{cases}$$

$$\Rightarrow \quad Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & i\text{th person female} \\ \beta_0 - \beta_1 + \varepsilon_i & \text{'' '' male} \end{cases}$$

so $\beta_0 =$ Average overall response

$\beta_1 =$ difference from the average response for m/f.

Important: predictions will not change with different encodings, but interpretations will!

More than 2 levels requires multiple dummy variables

$\boxed{\text{Ex}}$ $Y =$ life expectancy in country

$X \in \{$ Africa, OECD, other $\}$ (3 levels)

Could use

$$x_{i1} = \begin{cases} 1 & \text{ith country in OECD} \\ 0 & \text{'' '' not '' ''} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{ith country in other} \\ 0 & \text{" " not " "} \end{cases}$$

$$\Rightarrow y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i & \text{ith country OECD} \\ \beta_0 + \beta_2 + \varepsilon_i & \text{" " other} \\ \beta_0 + \varepsilon_i & \text{" " Africa} \end{cases}$$

$\Rightarrow$ β_0 = Avg life exp for countries in Africa

β_1 = Change in l.e. for a country in OECD over Africa

β_2 = " " " " " " in other " "

$\boxed{\text{Bad idea!}}$

$$x_i = \begin{cases} 0 & \text{ith country in Africa} \\ 1 & \text{" " " other} \\ 2 & \text{" " " OECD} \end{cases}$$

$$y = \beta_0 + \beta_1 x + \varepsilon = \begin{cases} \beta_0 + \varepsilon & \text{Afr} \\ \beta_0 + \beta_1 + \varepsilon & \text{other} \\ \beta_0 + 2\beta_1 + \varepsilon & \text{OECD} \end{cases}$$

Beyond additivity + linearity

Additivity: Effect of a predictor on Y is independent of values of other predictors. To overcome, use interactions:

$\boxed{\text{Ex}}$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon = \begin{cases} \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 + \varepsilon & m \\ = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon & \\ \beta_0 + \beta_1 x_1 + \varepsilon & F \end{cases}$$

Y = salary x_1 = yrs since degree $x_2 = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$

β_1 = Avg raise/yr for females

$(\beta_1 + \beta_3)$ = " " " " males

β_0 = Starting salary for females

$(\beta_0 + \beta_2)$ = " " " males

Linearity: Y depends linearly on X. Polynomial regression is an easy way to overcome:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

or

$$\log Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\Rightarrow \quad Y = e^{\beta_0} e^{\beta_1 X} e^{\varepsilon}$$

1) Y depends exponentially on X

2) errors are multiplicative

Degrees of freedom

$X \quad n \times (p+1)$

The model

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \quad \text{with } p \text{ features}$$

uses $(p+1)$ degrees of freedom. If H is the hat matrix, then

$$\operatorname{tr} H = \sum_{i=1}^{n} H_{ii} = \operatorname{tr}\left(X (X^T X)^{-1} X^T\right)$$

$$= \operatorname{tr}\left(X^T X (X^T X)^{-1}\right) = \operatorname{tr}(I_{p+1}) = \underline{p+1}$$

For any linear predictor $\hat{Y} = MY$, define $\operatorname{tr}(M) =$ degrees of freedom