Dummy variable encoding in linear regression

Miaoyan Wang

Department of Statistics
UW Madison

Multiple Linear Regression Model

The multiple linear regression (MLR) model for the data $(x_{i1}, x_{i2}, \dots, x_{i,p-1}, y_i)$ is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

for $i = 1, 2, \dots, n$, where

- Y_i is the ith observation of the **response variable**.
- X_{ik} is the ith observation of the kth **explanatory variable** for $k = 1, \dots, p-1$.
- ε_i is the ith **random error** term.
- The random errors follow a normal distribution with mean zero and variance σ^2 and are independent of each other.
- That is, $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$.

Example: $p = 3$

- Example: # of explanatory variables = 2.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad \varepsilon_i \sim \text{iid } N(0, \sigma^2),$$

  for $i = 1, \ldots, n$.
- Mean response:

$$\mathbb{E}(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}.$$

- Interpretation:
  - β_0: Intercept. The mean response $\mathbb{E}(Y)$ at $X_1 = X_2 = 0$.
  - β_1: Slope. The change in the mean response $\mathbb{E}(Y)$ per unit increase in X_1, when X_2 is held constant.
  - β_2: Slope. The change in the mean response $\mathbb{E}(Y)$ per unit increase in X_2, when X_1 is held constant.

Dummy variable

- The predictors in the linear model can be either continuous (e.g., age, height) or categorical (e.g., gender, group)
- For a categorical predictor that has p categories, define $p-1$ dummy variables:

$$X_{ik} = \begin{cases} 1 & \text{observation } i \text{ is in category } k \\ 0 & \text{otherwise} \end{cases}$$

  where $k = 1, \ldots, p-1$.
- Include dummy variables as predictors in the linear model.
- Example. Consider n i.i.d. observations from the following model:

$$Y = \beta_0 + \beta_1 \text{Age} + \beta_2 X + \boldsymbol{\varepsilon}, \quad \text{where } \varepsilon \sim i.i.d.\ N(0, \sigma^2),$$

  with $X = 1$ if male, $X = 0$ if female.
- What is the interpretation for β_0, β_1, and β_2?

Example with categorical variables

Consider the effect of education on hourly wages (Y). The education is classified into three categories:

Option in Survey (O)	Meaning (M)
1	College dropout
2	College
3	MS and above

Which model makes more sense?

- $Y = \beta_0 + \beta_1 O + \varepsilon$?
- $Y = \beta_0 + \beta_1 \mathbb{1}_{\text{college}} + \beta_2 \mathbb{1}_{\text{MS and above}} + \varepsilon$?
- $Y = \beta_0 + \beta_1 \mathbb{1}_{\text{college dropout}} + \beta_2 \mathbb{1}_{\text{college}} + \varepsilon$?

(In all cases, assume $\varepsilon \sim i.i.d. N(0, \sigma^2)$)

Example (Cont.)

- To include the eduction as predictor in a regression model, define 2 dummy variables X_1 and X_2:

Option in Survey (O)	Meaning (M)	X_1	X_2
1	College dropout	0	0
2	College	1	0
3	MS and above	0	1

- Baseline (all dummies 0): college dropout;
- $X_1 = 1$, if the highest degree is college, 0 otherwise;
- $X_2 = 1$, if degree with MS and above, 0 otherwise.

Include X_1 and X_2 as dummy variables in a regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \underbrace{\beta_3 X_3 + \ldots + \beta_p X_p}_{\text{other predictors, e.g., age}} + \varepsilon, \quad \varepsilon \sim i.i.d.\ N(0, \sigma^2).$$

Models in matrix form

- Response variable: $\boldsymbol{Y}_{n\times 1} = (Y_1, Y_2, \ldots, Y_n)'$.
- Design matrix:

$$\boldsymbol{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

- Random error: $\boldsymbol{\varepsilon}_{n\times 1} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$.
- Regression coefficients: $\boldsymbol{\beta}_{p\times 1} = (\beta_0, \beta_1, \ldots, \beta_{p-1})'$.
- The multiple linear regression model can be written as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}_{n\times 1}, \sigma^2 \boldsymbol{I}_{n\times n}).$$

Inference on the linear contrast

Recall the study that investigates the effect of education on hourly salary (Y):

Education	X_1	X_2
College dropout	0	0
College	1	0
MS and above	0	1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \quad \text{where} \quad \varepsilon \sim i.i.d.\ N(0, \sigma^2).$$

Suppose we are interested in testing:

- The mean salary for “MS and above” is the same as for “College”:
  $H_0 : \beta_1 = \beta_2 \longleftrightarrow H_0 : 0 * \beta_0 + 1 * \beta_1 - 1 * \beta_2 = 0$
- The mean salary for “College” is the same as for “College dropout”:
  $H_0 : \beta_1 = 0 \longleftrightarrow H_0 : 0 * \beta_0 + 1 * \beta_1 + 0 * \beta_2 = 0$
- Compared to college dropout, the mean salary increase for “MS and above” is twice as that for “College”:
  $\mathrm{H}_0 : \beta_2 = 2\beta_1 \longleftrightarrow H_0 : 0 * \beta_0 + 2 * \beta_1 - 1 * \beta_2 = 0$

Inference on the linear contrast

- All these hypothesis tests could be expressed as a linear contrast:

$$H_0 : c_0\beta_0 + c_1\beta_1 + c_2\beta_2 = 0 \quad \text{v.s.} \quad H_\alpha : c_0\beta_0 + c_1\beta_1 + c_2\beta_2 \neq 0,$$

  for a given vector $\boldsymbol{c} = (c_0, c_1, c_2)$. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$.
- What is the distribution of $\boldsymbol{c}'\hat{\boldsymbol{\beta}}$ under the null? Multivariate normal with

$$\mathbb{E}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \boldsymbol{c}'\boldsymbol{\beta}, \quad \text{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}) = \underline{\hspace{4cm}} = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{c}$$

- In case σ^2 is unknown, plug in the estimator $\hat{\sigma}^2$. (what is the form of $\hat{\sigma}^2$?)

$$\frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\widehat{\text{Var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})}}} \sim T_{n-3}$$