Biostatistics I: Linear Regression

Multiple Linear Regression

Nicole S. Erler
Department of Biostatistics, Erasmus Medical Center
n.erler@erasmusmc.nl
@N_Erler

Erasmus MC
University Medical Center Rotterdam

The Multiple Linear Regression Model

Basic assumptions:

- **single** continuous **response** variable
- **multiple covariates** of mixed type (continuous or categorical)

The Multiple Linear Regression Model

Basic assumptions:

- **single** continuous **response** variable
- **multiple covariates** of mixed type (continuous or categorical)

The model is then formally written as:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}_{\substack{\text{additive linear systematic component} \\ \text{(linear predictor)}}} + \underbrace{\varepsilon_i}_{\substack{\text{error} \\ \text{terms}}}$$

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

The Multiple Linear Regression Model

Basic assumptions:

- **single** continuous **response** variable
- **multiple covariates** of mixed type (continuous or categorical)

The model is then formally written as:

$$y_i = \underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}}_{\substack{\text{additive linear systematic component} \\ \text{(linear predictor)}}} + \underbrace{\varepsilon_i}_{\substack{\text{error} \\ \text{terms}}}$$

$$\mathrm{E}(\varepsilon_i) = 0, \quad \mathrm{var}(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

- Extension of simple linear regression to multiple covariates.
- **Note:** Both are **univariate** models!

What Makes the Linear Model Linear?

A **linear** regression model is **linear in the regression coefficients** and the error term.

Linear

- $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\varepsilon}$
- $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1^2 + \beta_2 \log(\mathbf{x}_2) + \boldsymbol{\varepsilon}$
- $\log(\mathbf{y}) = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \boldsymbol{\varepsilon}$

Not linear

- $\mathbf{y} = \beta_0 + \exp(\beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2) + \boldsymbol{\varepsilon}$
- $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1 / (\beta_2 \mathbf{x}_2) + \boldsymbol{\varepsilon}$
- $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x}_1^{\beta_2} + \boldsymbol{\varepsilon}$

Example: Child Growth

Our data might look like this:

height	age	sex	race
112	6.53	boy	caucasian
108	4.76	girl	caucasian
117	6.33	boy	asian
114	5.34	boy	other
100	2.95	girl	caucasian

How would our regression model look like,

Example: Child Growth

Our data might look like this:

height	age	sex	race
112	6.53	boy	caucasian
108	4.76	girl	caucasian
117	6.33	boy	asian
114	5.34	boy	other
100	2.95	girl	caucasian

How would our regression model look like,

$$\mathbf{height}_i = \beta_0 + \beta_1 \mathbf{age}_i + \beta_2 \mathbf{sex}_i + \beta_3 \mathbf{race}_i + \varepsilon_i?$$

Coefficients of Continuous Covariates

In the model

$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{race}_i + \varepsilon_i$$

β_1 describes the change in the expected height when **age is increased by one unit** and all **other covariates are held constant**.

Coefficients of Continuous Covariates

In the model

$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{race}_i + \varepsilon_i$$

β_1 describes the change in the expected height when **age is increased by one unit** and all **other covariates are held constant**.

$$\text{height}_{age} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{race}$$

$$\text{height}_{age+1} = \beta_0 + \beta_1 (\text{age} + 1) + \beta_2 \text{sex} + \beta_3 \text{race}$$

Coefficients of Continuous Covariates

In the model

$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{race}_i + \varepsilon_i$$

β_1 describes the change in the expected height when **age is increased by one unit** and all **other covariates are held constant**.

$$\text{height}_{age} = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \text{race}$$

$$\text{height}_{age+1} = \beta_0 + \beta_1 (\text{age} + 1) + \beta_2 \text{sex} + \beta_3 \text{race}$$

$$\text{height}_{age+1} - \text{height}_{age} = \beta_1 (\text{age} + 1) - \beta_1 \text{age} \quad = \beta_1$$

Categories as Numeric Values

We could use the following coding:

- **sex**:
  "boy" = 0, "girl" = 1
- **race**:
  "caucasian" = 0, "asian" = 1, "other" = 2

height	age	sex	race
112	6.53	0	0
108	4.76	1	0
117	6.33	0	1
114	5.34	0	2
100	2.95	1	0

Categories as Numeric Values

We could use the following coding:

- **sex**:
  "boy" = 0, "girl" = 1
- **race**:
  "caucasian" = 0, "asian" = 1, "other" = 2

height	age	sex	race
112	6.53	0	0
108	4.76	1	0
117	6.33	0	1
114	5.34	0	2
100	2.95	1	0

This results in the linear predictors:

$$\text{boy (sex} = 0\text{):} \quad \beta_0 + \beta_1 \textbf{age} + \beta_3 \textbf{race}$$

$$\text{girl (sex} = 1\text{):} \quad \beta_0 + \beta_1 \textbf{age} + \beta_2 + \beta_3 \textbf{race}$$

Categories as Numeric Values

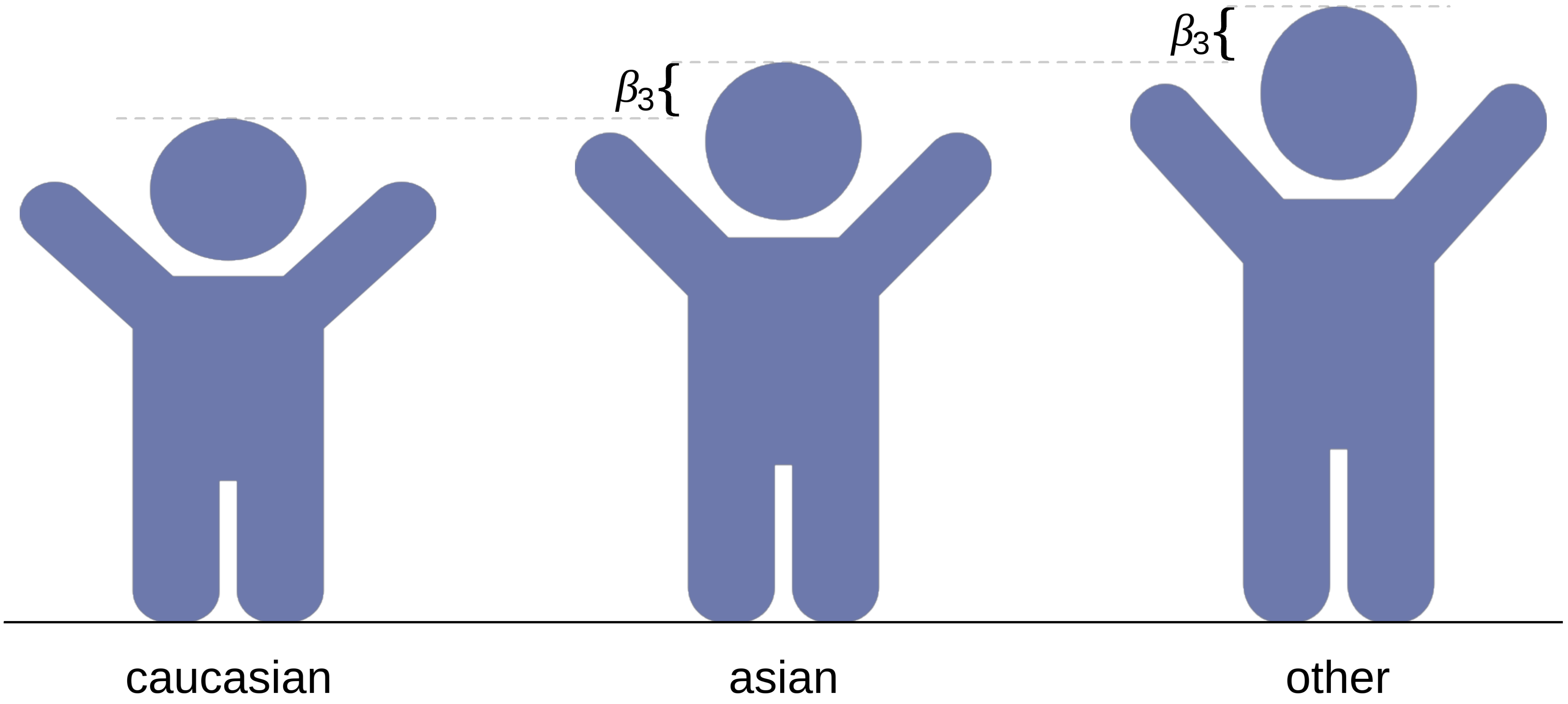
What would this look like for the **effect of race**?

$$\text{caucasian (race} = 0\text{):} \quad \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex}$$

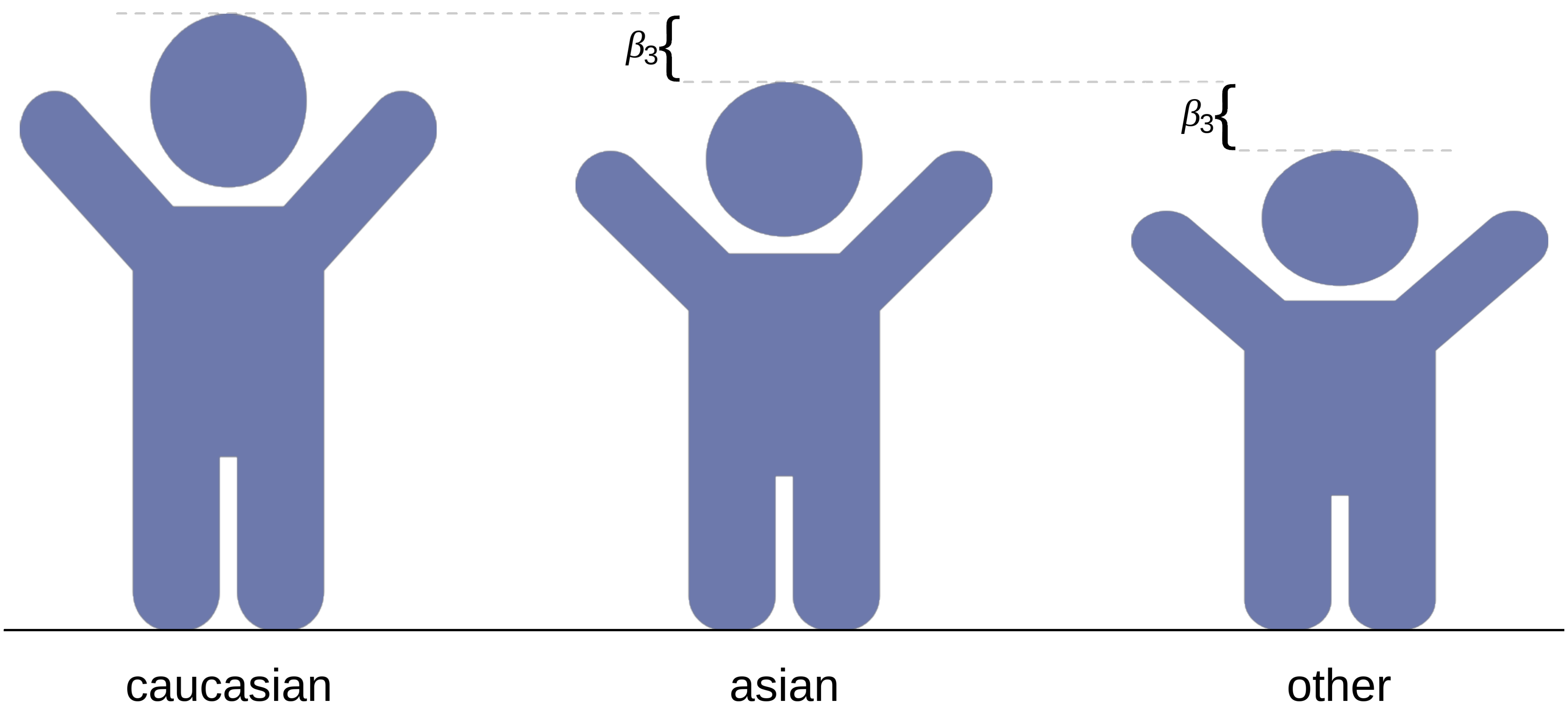
$$\text{asian (race} = 1\text{):} \quad \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3$$

$$\text{other (race} = 2\text{):} \quad \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + 2\beta_3$$

Categories as Numeric Values



Categories as Numeric Values



Categories as Numeric Values

To **avoid the link** between effects of different categories we need **additional parameters**.

In general: One parameter less than the number of categories.

Categories as Numeric Values

To **avoid the link** between effects of different categories we need **additional parameters**.

In general: One parameter less than the number of categories.

Most common coding choices:

Dummy coding

	$\text{race}^{(\text{asian})}$	$\text{race}^{(\text{other})}$
caucasian	0	0
asian	1	0
other	0	1

Effect coding

	$\text{race}^{(1)}$	$\text{race}^{(2)}$
caucasian	1	0
asian	0	1
other	-1	-1

Dummy Coding

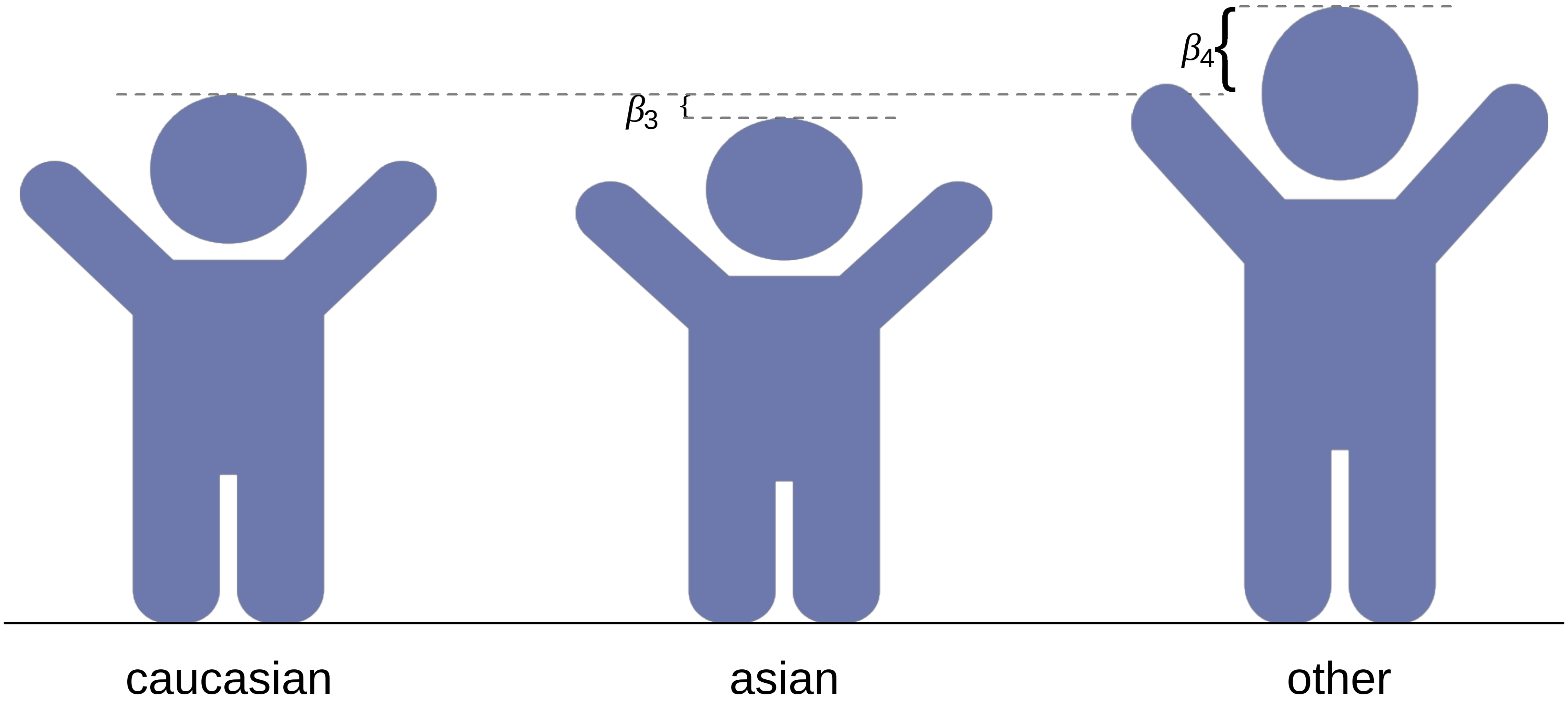
Using **dummy coding**, the model is:

$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{race}_i^{(asian)} + \beta_4 \text{race}_i^{(other)} + \varepsilon_i$$

This leads to the following linear predictors:

$$\begin{aligned}
\text{caucasian:} \quad & \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 0 + \beta_4 0 && = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} \\
\text{asian:} \quad & \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 1 + \beta_4 0 && = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \\
\text{other:} \quad & \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 0 + \beta_4 1 && = \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_4
\end{aligned}$$

Dummy Coding



Effect Coding

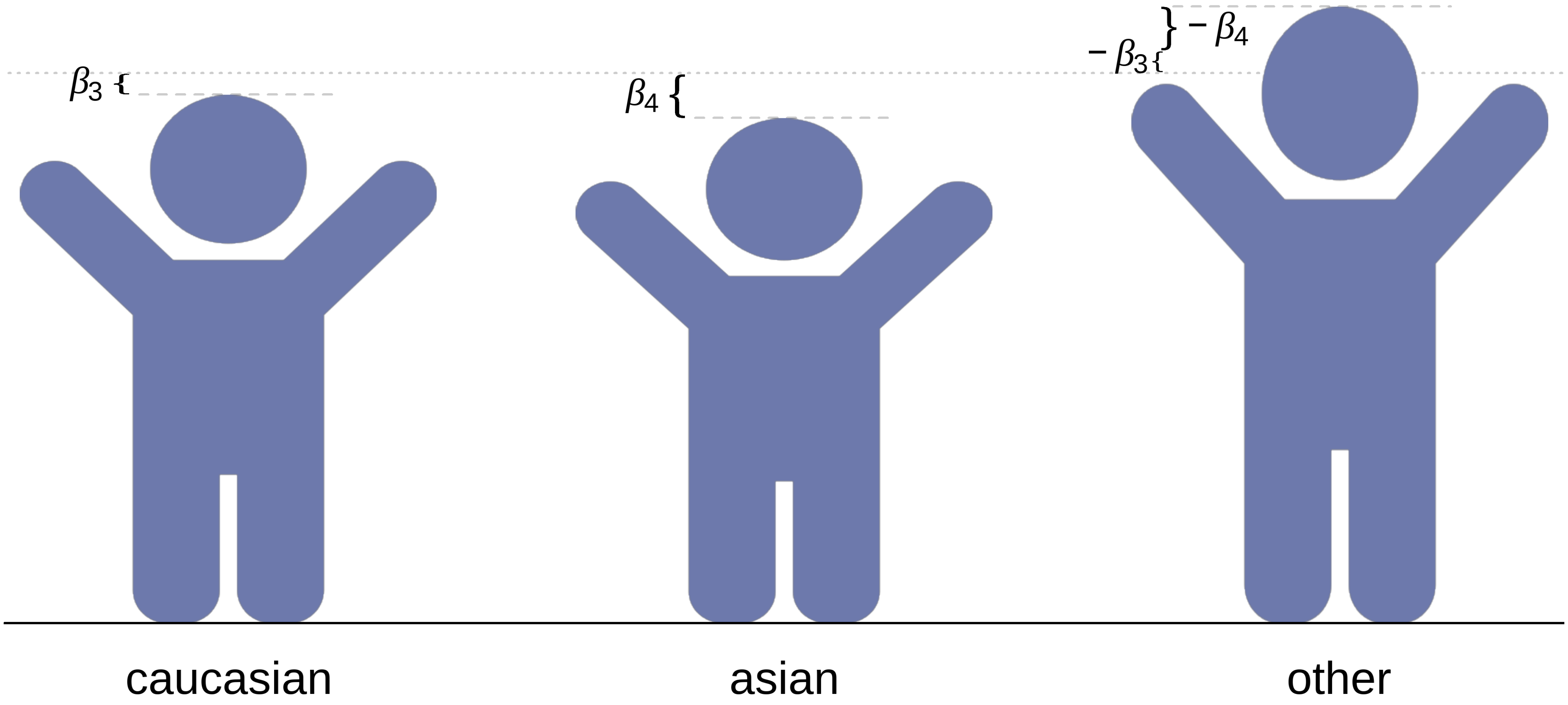
Using **effect coding**, the model is:

$$\text{height}_i = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 \text{race}_i^{(1)} + \beta_4 \text{race}_i^{(2)} + \varepsilon_i$$

Effect coding will lead to the following linear predictors:

$$\begin{aligned}
\text{caucasian: } & \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 1 + \beta_4 0 &&= \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 \\
\text{asian: } & \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 0 + \beta_4 1 &&= \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_4 \\
\text{other: } & \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} + \beta_3 (-1) + \beta_4 (-1) &&= \beta_0 + \beta_1 \text{age} + \beta_2 \text{sex} - \beta_3 - \beta_4
\end{aligned}$$

Effect Coding



Interpretation of the Intercept

Dummy coding:

$$\beta_0 + \beta_1 \mathbf{age} + \beta_2 \mathbf{sex} + \beta_3 \mathbf{race}_i^{(asian)} + \beta_4 \mathbf{race}_i^{(other)}$$

In dummy coding, the intercept β_0 is the expected outcome when **all covariate values are zero**, i.e., for a caucasian (race$^{(\text{asian})}$ = race$^{(\text{other})}$ = 0) boy (sex = 0) with zero years of age.

Interpretation of the Intercept

Dummy coding:

$$\beta_0 + \beta_1 \mathbf{age} + \beta_2 \mathbf{sex} + \beta_3 \mathbf{race}_i^{(asian)} + \beta_4 \mathbf{race}_i^{(other)}$$

In dummy coding, the intercept β_0 is the expected outcome when **all covariate values are zero**, i.e., for a caucasian (race$^{\text{(asian)}}$ = race$^{\text{(other)}}$ = 0) boy (sex = 0) with zero years of age.

Effect coding:

$$\beta_0 + \beta_1 \mathbf{age} + \beta_2 \mathbf{sex} + \beta_3 \mathbf{race}^{(1)} + \beta_4 \mathbf{race}^{(2)}$$

With effect coding there is no scenario where all effects are zero.

Interpretation of the Intercept

In **effect coding** the intercept represents the **average expected response over all categories** (when all other covariates are zero).

$$
\begin{aligned}
\text{height}_{cauc.} &= \beta_0 + \beta_3 \\
\text{height}_{asian} &= \beta_0 + \beta_4 \\
\text{height}_{other} &= \beta_0 - \beta_3 - \beta_4
\end{aligned}
$$

Interpretation of the Intercept

In **effect coding** the intercept represents the **average expected response over all categories** (when all other covariates are zero).

$$
\begin{aligned}
\text{height}_{cauc.} &= \beta_0 + \beta_3 \\
\text{height}_{asian} &= \beta_0 + \beta_4 \\
\text{height}_{other} &= \beta_0 - \beta_3 - \beta_4
\end{aligned}
$$

$$
\begin{aligned}
\frac{\text{height}_{cauc.} + \text{height}_{asian} + \text{height}_{other}}{3} &= \frac{\beta_0 + \beta_3 + \beta_0 + \beta_4 + \beta_0 - \beta_3 - \beta_4}{3} \\
&= \frac{3\beta_0}{3} = \beta_0
\end{aligned}
$$

Multiple Linear Regression in Matrix Notation

The basic model of **multiple linear regression in matrix notation** is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Multiple Linear Regression in Matrix Notation

The basic model of **multiple linear regression in matrix notation** is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Sidenote: Tansposing Vectors and Matrices

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \Rightarrow \quad \mathbf{y}^\top = (y_1, \ldots, y_n)$$

Sidenote: Tansposing Vectors and Matrices

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad \Rightarrow \quad \mathbf{y}^\top = (y_1, \ldots, y_n)$$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \ldots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \ldots & x_{np} \end{pmatrix} \quad \Rightarrow \quad \mathbf{X}^\top = \begin{pmatrix} 1 & \ldots & 1 \\ x_{11} & \ldots & x_{n1} \\ \vdots & & \vdots \\ x_{1p} & \ldots & x_{np} \end{pmatrix}$$

Estimation via OLS

Ordinary Least Squares (OLS) Estimator

$$\sum_{i=1}^{n} \underbrace{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})}_{\hat{\varepsilon}_i}^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

Estimation via OLS

Ordinary Least Squares (OLS) Estimator

$$\sum_{i=1}^{n} \underbrace{(y_i - \mathbf{x}_i^\top \boldsymbol{\beta})}_{\hat{\varepsilon}_i}{}^2 \longrightarrow \min_{\boldsymbol{\beta}}$$

The **least squares principle** in matrix notation

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \longrightarrow \min_{\boldsymbol{\beta}}$$