Multiple Regression

NGSchool 2022

Nuno Sepúlveda, 16.09.2022
nuno.Sepulveda@mini.pw.edu.pl
http://www.immune-stats.net

Objectives

Touch-base on multiple linear regression

Two covariates

More than two covariates

Estimation and hypothesis testing

Use R to conduct data analysis

Two covariates

Y and x_1

$+$

x_2

binary

categorical

quantitative

Two covariates

Y and x_1

$+$

x_2

binary categorical quantitative

Binary X_2 covariate

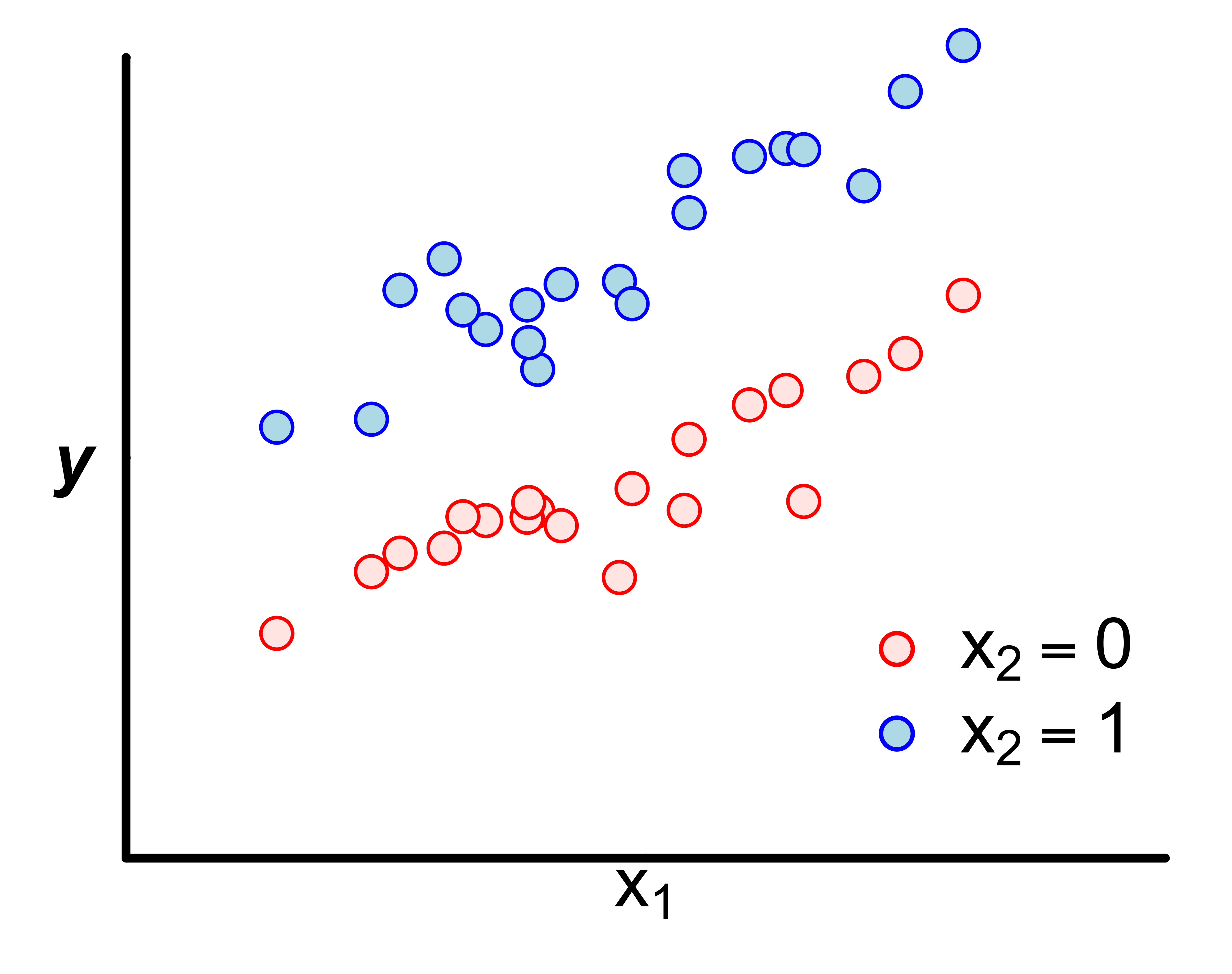
$$x_2 = 0,1$$

Male = 0, Female =1

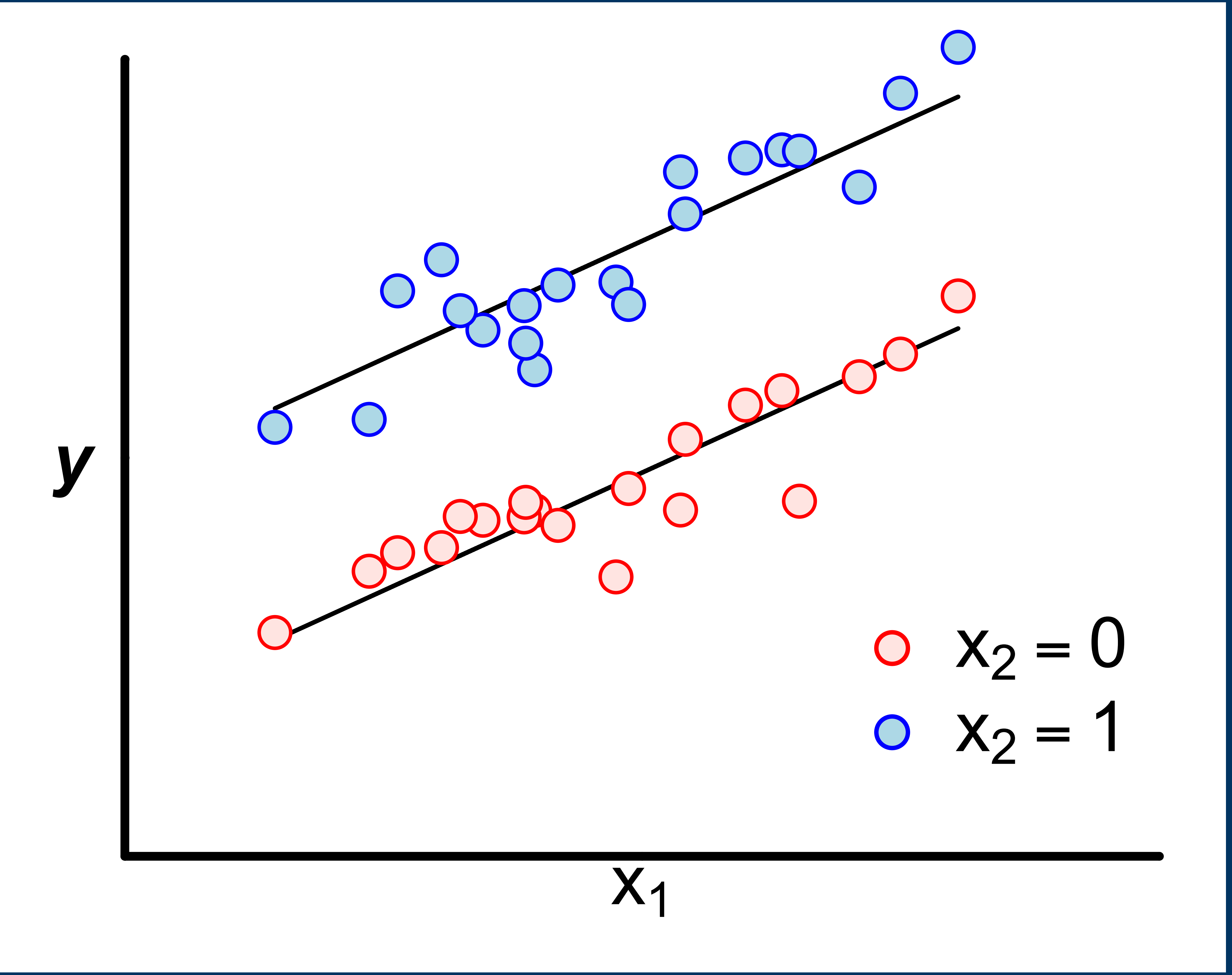
Healthy = 0, Sick =1

Placebo = 0, New Treatment =1

First data pattern



What does this model imply in terms of slope and intercept?



Model with main effects only

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$$

b_0 = intercept (overall mean)

b_1, b_2 = main effects

$$\epsilon_i \rightsquigarrow N\left(\mu = 0, \sigma\right)$$

Binary X_2 covariate

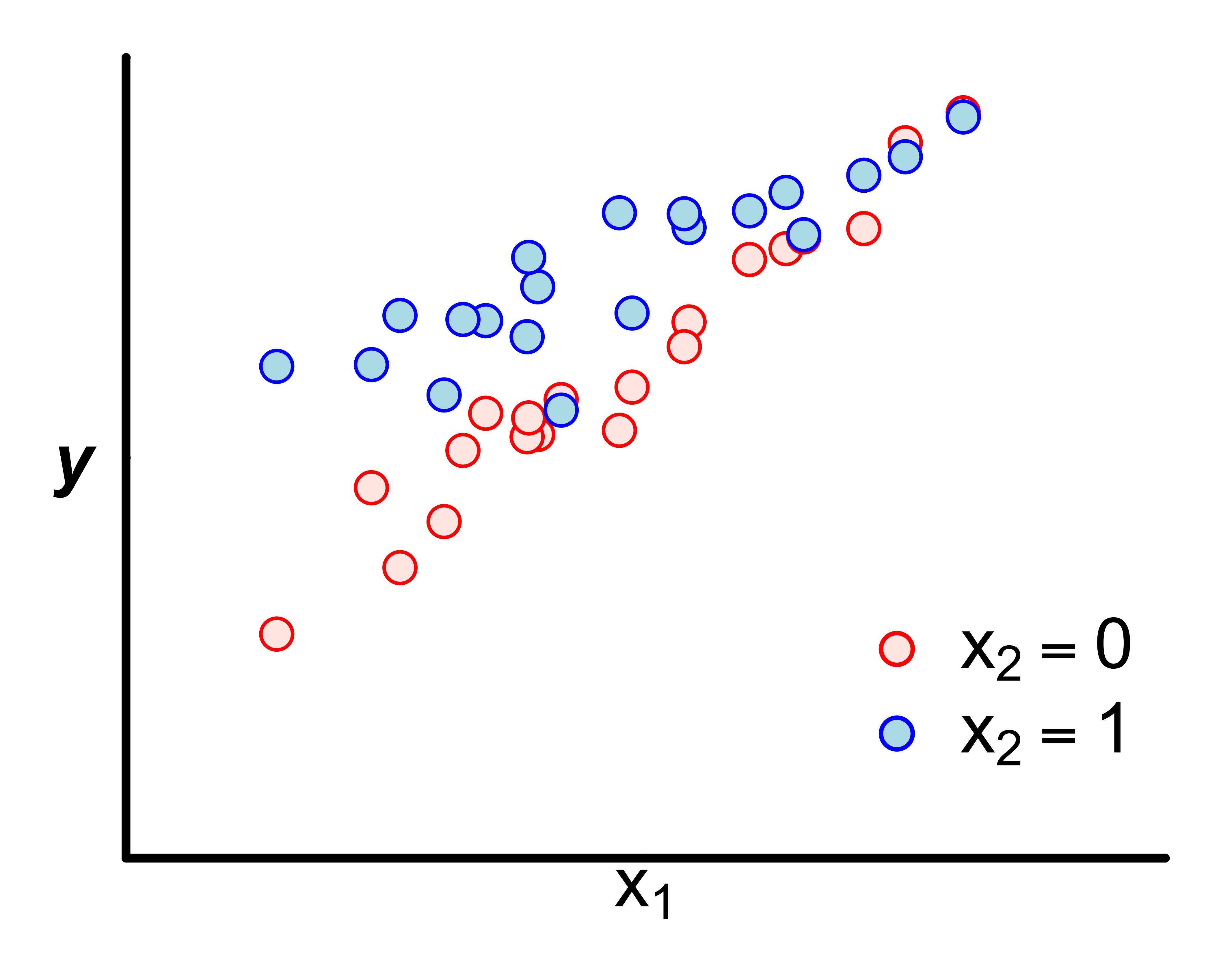
$$x_2 = 0,1$$

Male = 0, Female =1

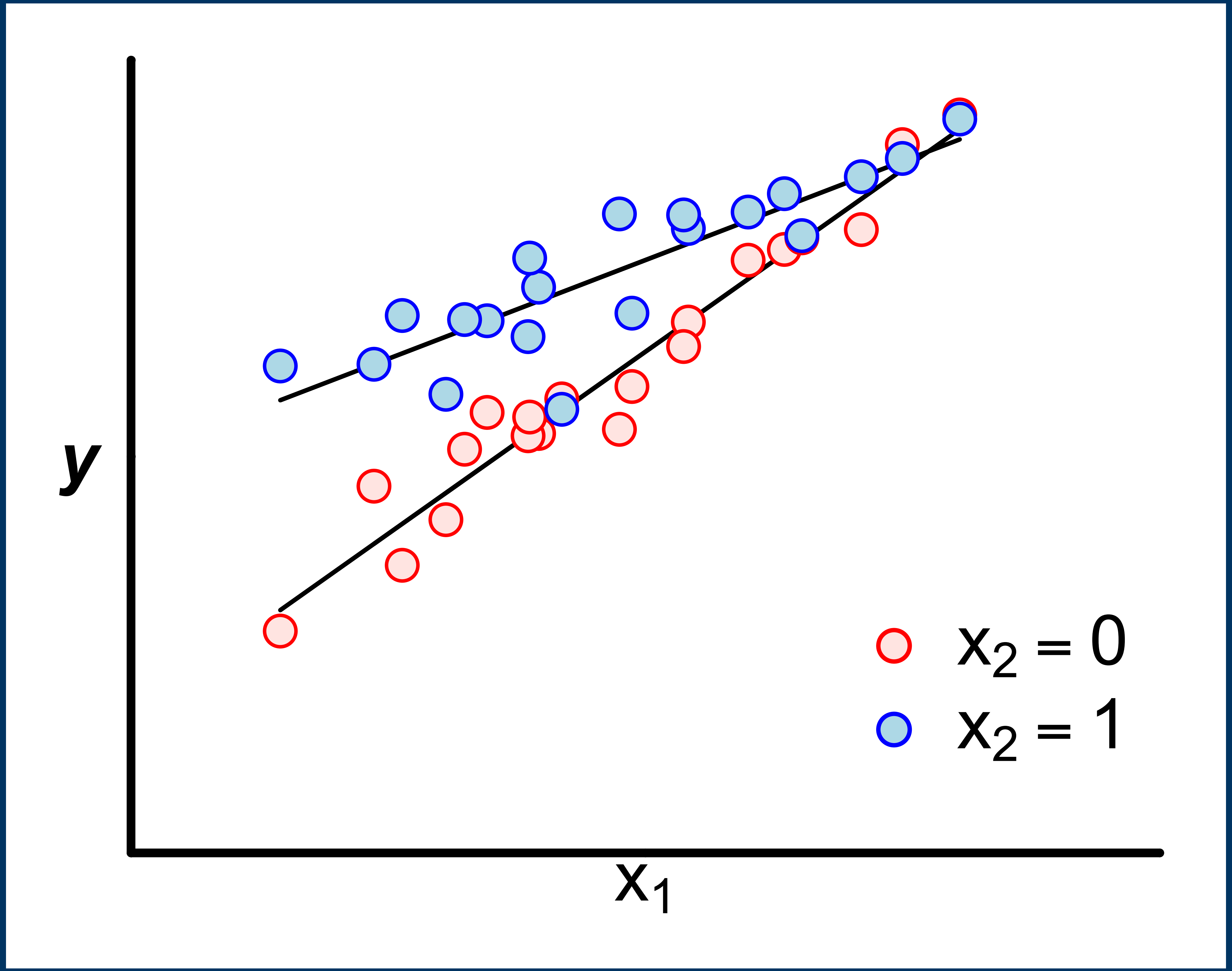
Healthy = 0, Sick =1

Placebo = 0, New Treatment =1

Second data pattern



What does this model imply in terms of slope and intercept?



Model with main effects and interaction term

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{1i} x_{2i} + \epsilon_i$$

$b_0 =$ intercept

$b_1, b_2 =$ main effects

$b_3 =$ interaction effect

Two covariates

Y and x_1

$+$

x_2

- binary
- **categorical**
- quantitative

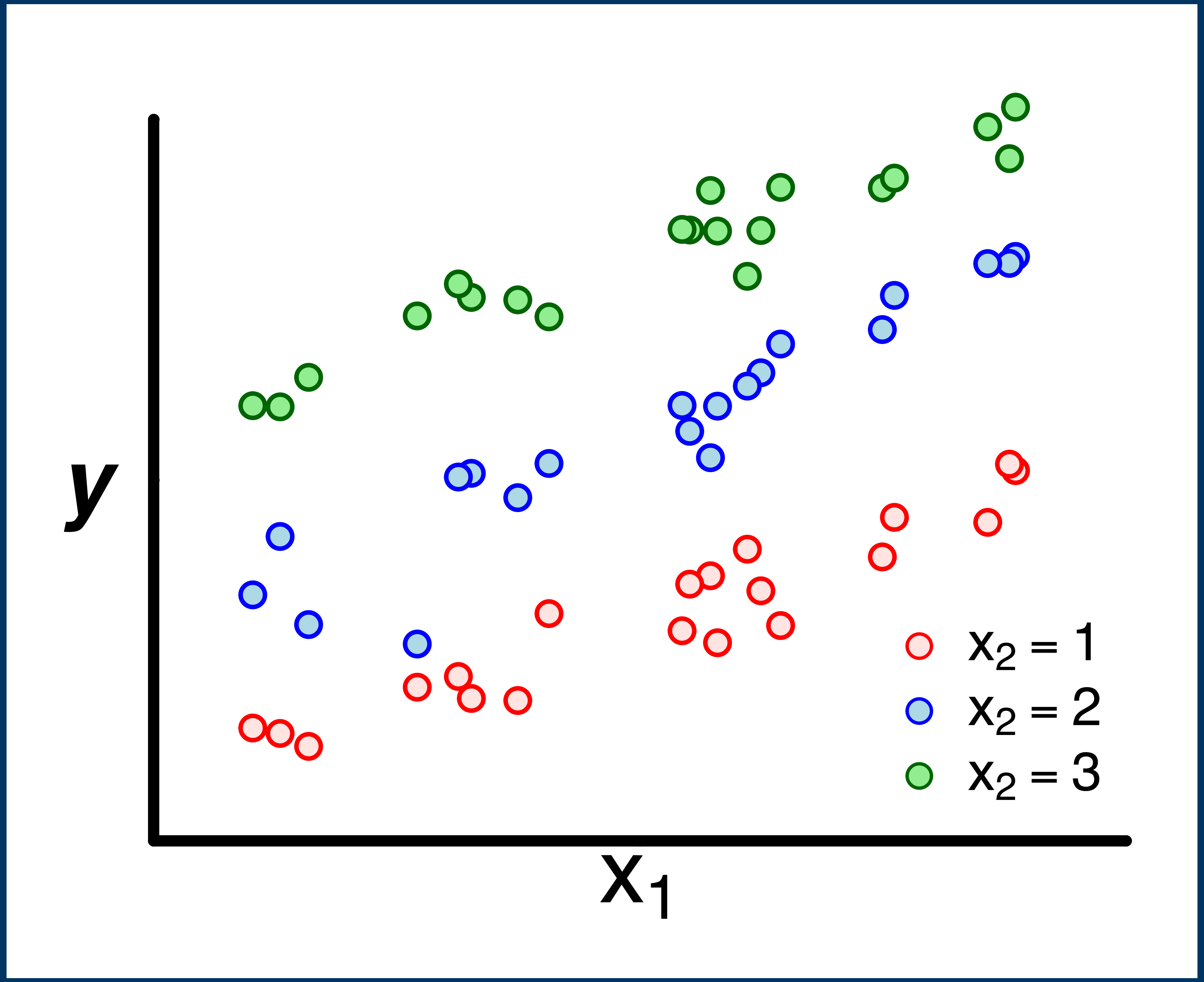
Categorical X_2 covariate

$$x_2 \in \{C_1, \ldots, C_k\}$$

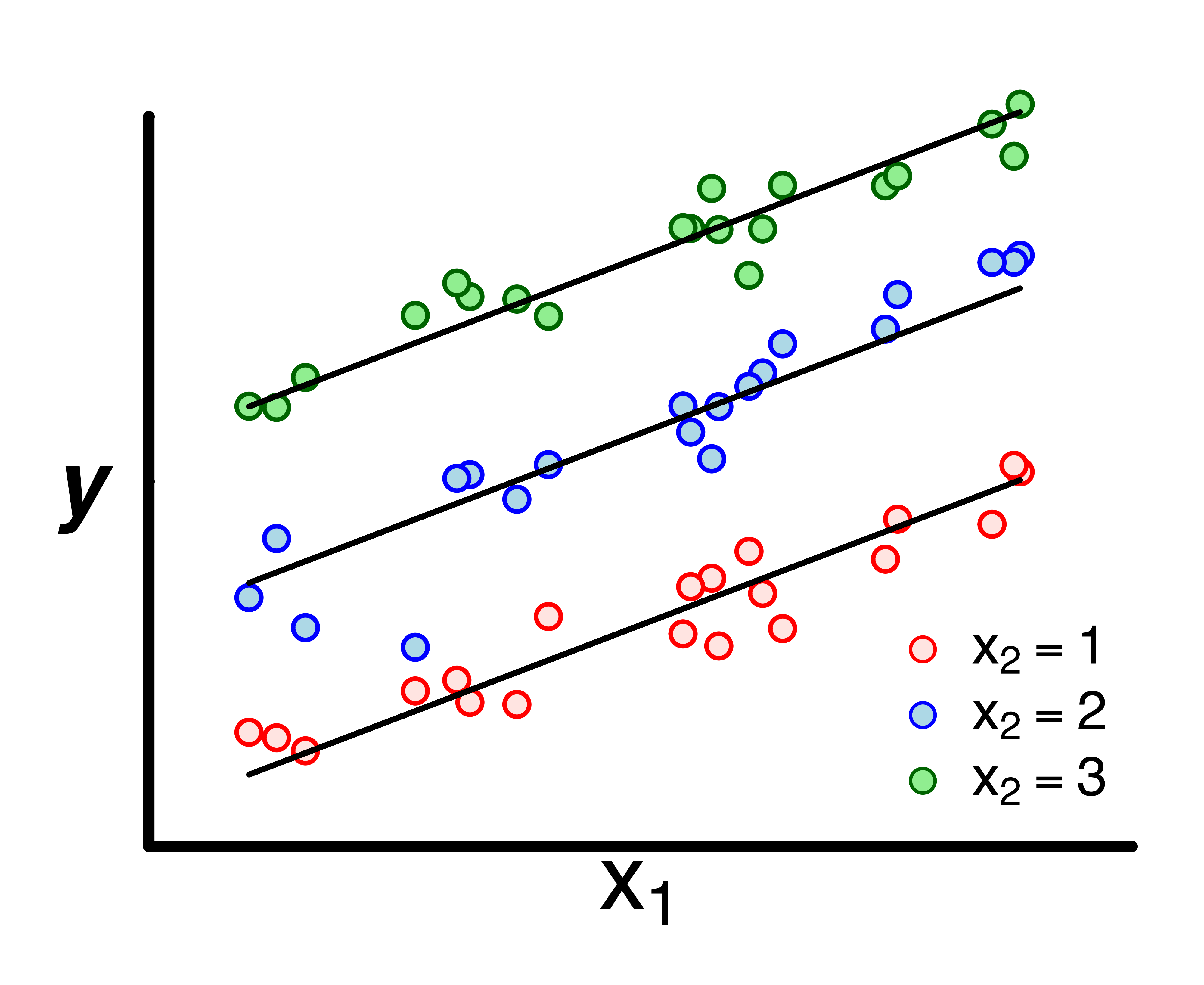
$$x_2 \in \{AA, AB, BB\}$$

$$x_2 \in \{\text{Placebo}, T_1, T_2, T_3\}$$

First data pattern



What does this model imply in terms of slope and intercept?



Model with main effects only

$$y_i = b_0 + b_1 x_{1i} + \sum_{l=2}^{k} b_{2l} x^*_{2li} + \epsilon_i$$

b_0 = intercept (overall mean)

$b_1, b_{22}, \ldots, b_{2k}$ = main effects

$$x^*_{2li} = \begin{cases} 1, & \text{if } x_{2i} = l \\ 0, & \text{otherwise} \end{cases}.$$

$$\epsilon_i \rightsquigarrow N\left(\mu = 0, \sigma\right)$$

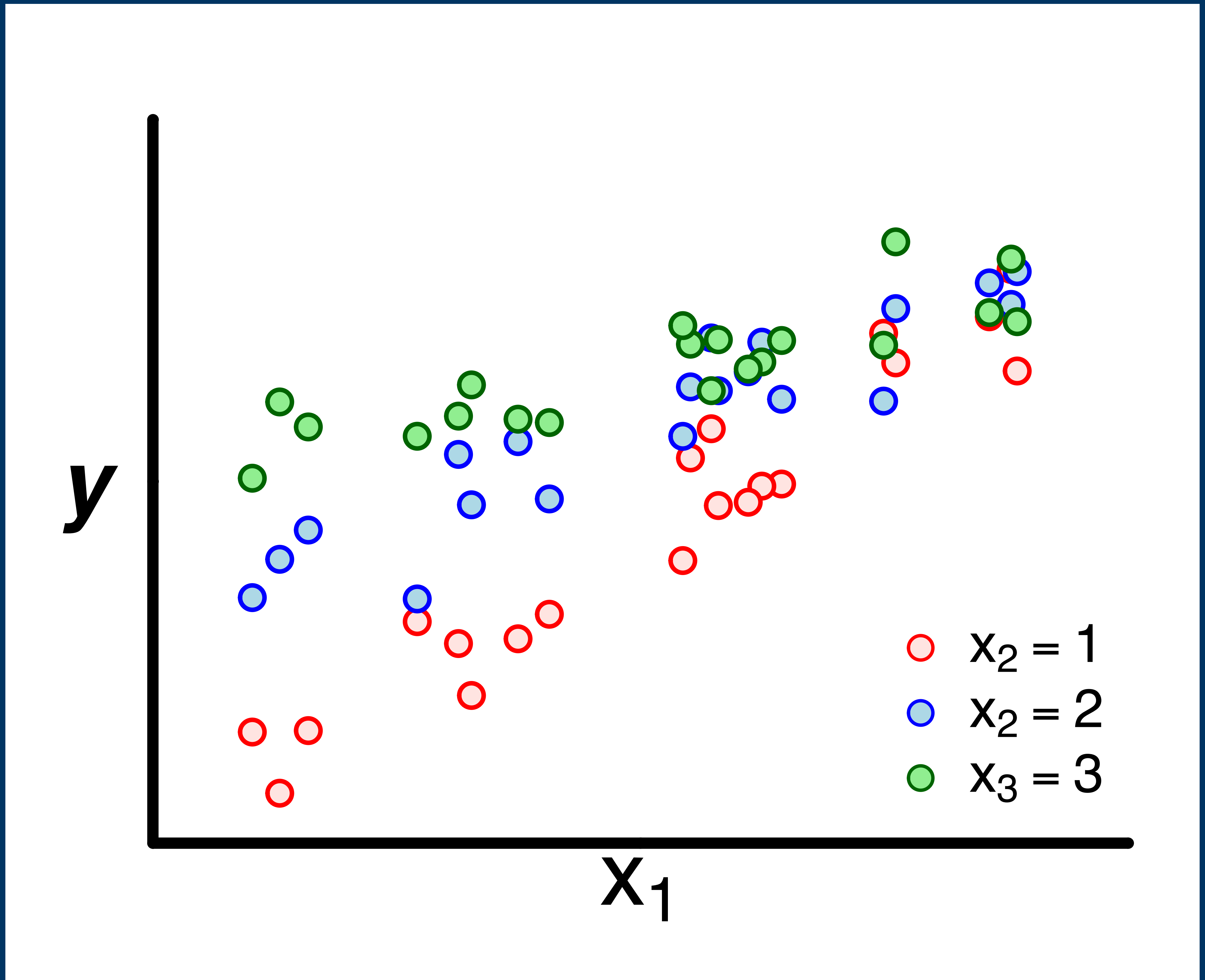
Categorical X_2 covariate

$$x_2 \in \{C_1, \ldots, C_k\}$$

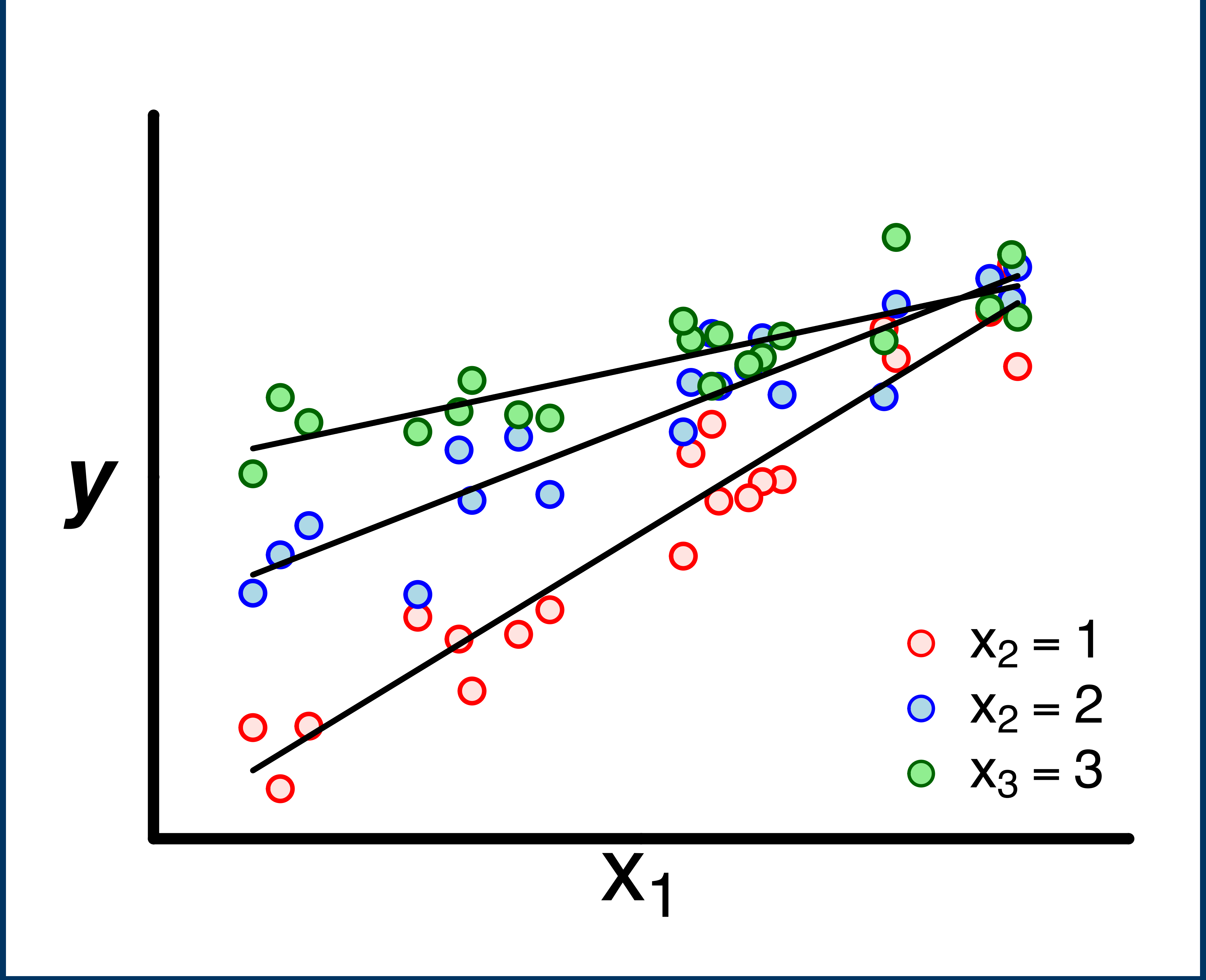
$$x_2 \in \{AA, AB, BB\}$$

$$x_2 \in \{\text{Placebo}, T_1, T_2, T_3\}$$

Second data pattern



Is this a good model?



Model with main effects and interaction terms

$$y_i = b_0 + b_1 x_{1i} + \sum_{l=2}^{k} b_{2l} x^*_{2li} + \sum_{l=2}^{k} b_{3l} x_{1i} x^*_{2li} + \epsilon_i$$

b_0 = intercept (overall mean)

$b_1, b_{22}, \ldots, b_{2k}$ = main effects

$b_{32}, \ldots, b_{3k}$ = interaction effects

$\epsilon_i \rightsquigarrow N\left(\mu = 0, \sigma\right)$

$$x^*_{2li} = \begin{cases} 1, & \text{if } x_{2i} = l \\ 0, & \text{otherwise} \end{cases}.$$

Two covariates

Y and x_1

$+$

x_2

binary

categorical

quantitative

Quantitative X_2 covariate

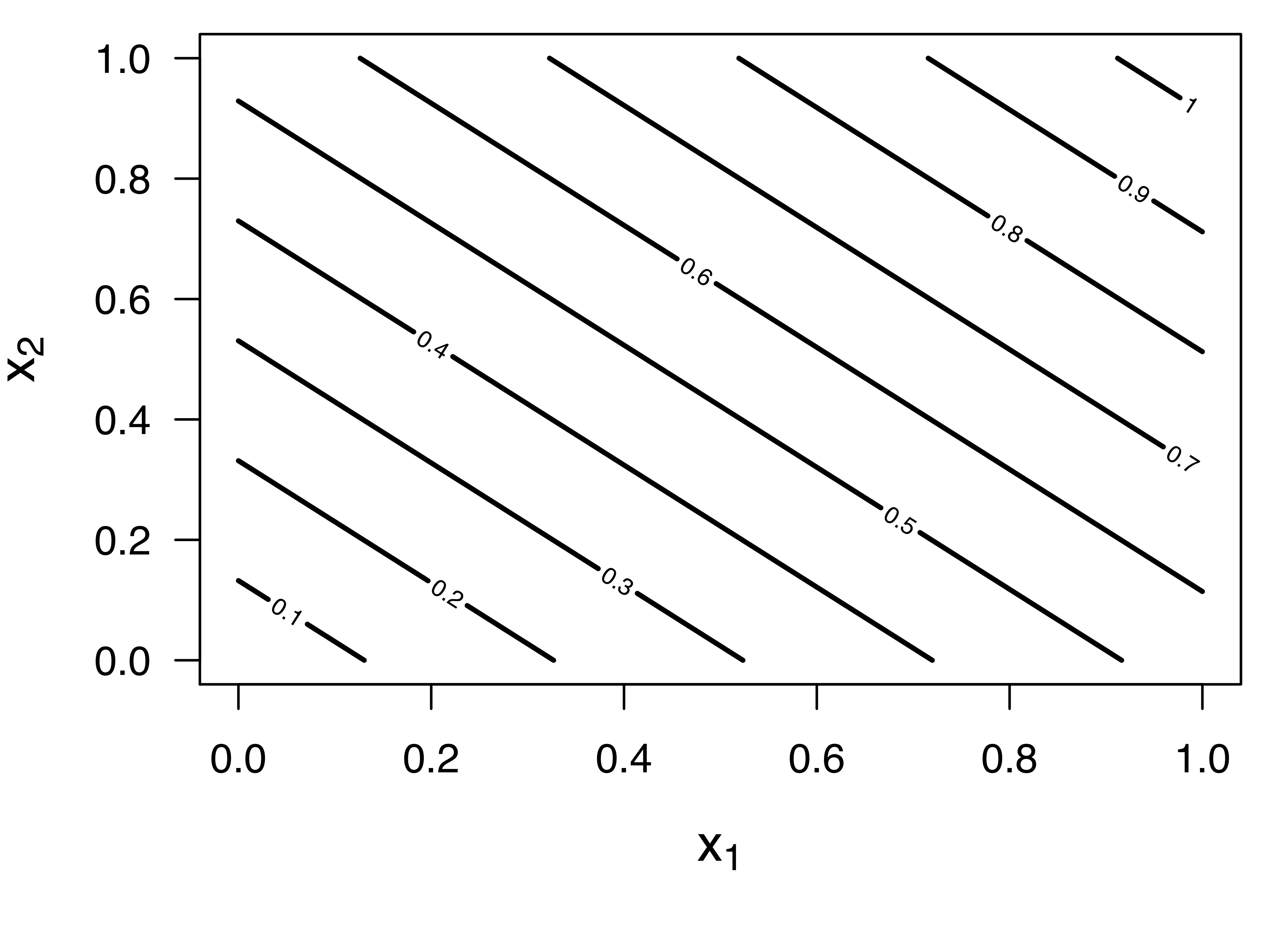
$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$$

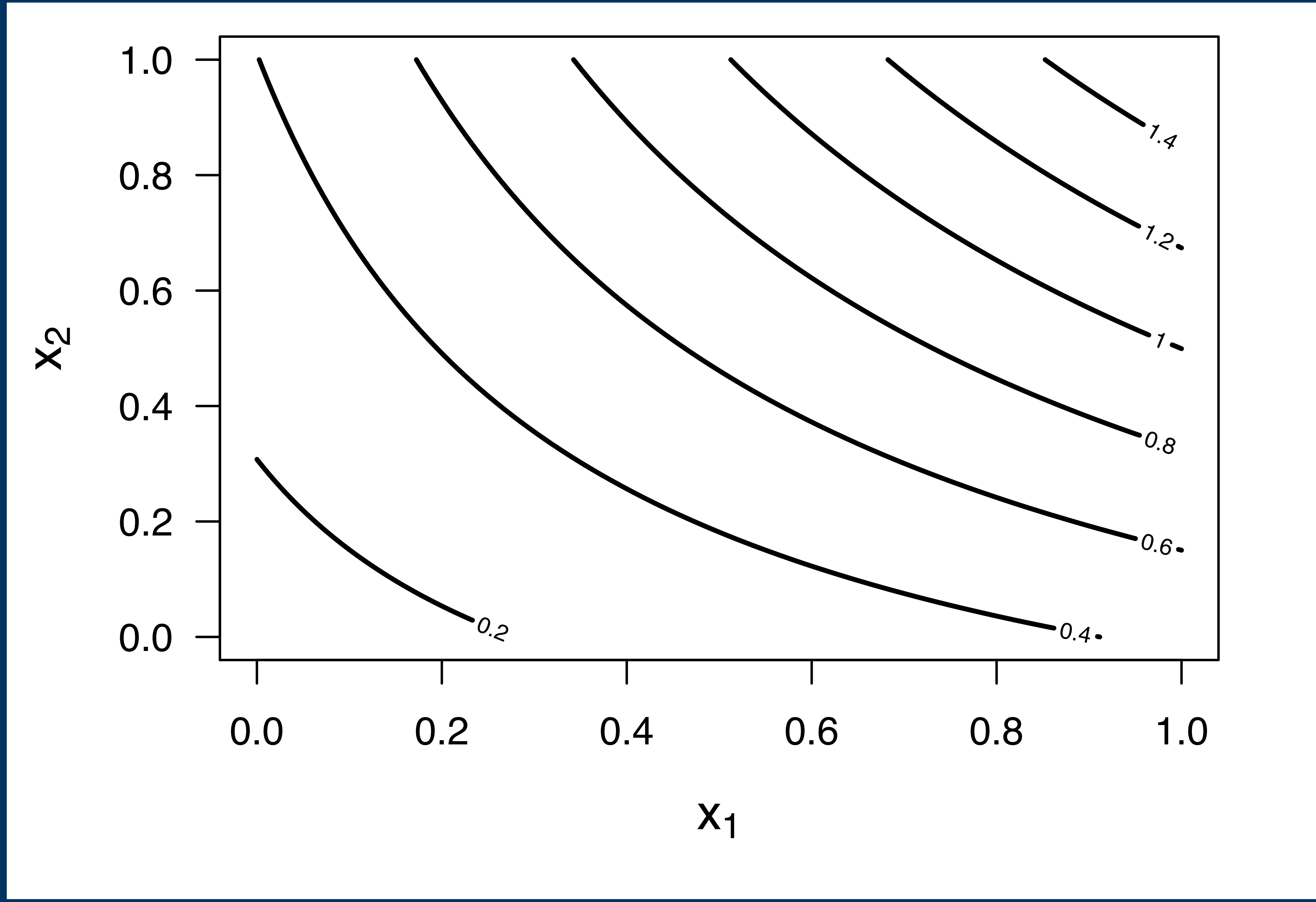
$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{1i} x_{2i} + \epsilon_i$$

Response Surfaces

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \epsilon_i$$

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{1i} x_{2i} + \epsilon_i$$





Multiple linear regression (p covariates)

$$y_i = b_0 + b_{1i}x_{1i} + b_2x_{2i} + \cdots + b_px_{pi} + \epsilon_i$$

$$\epsilon_i \rightsquigarrow N\left(\mu = 0,\sigma\right)$$

$a =$ overall mean of Y in the absence of any covariate effect

$b_i =$ slope concerning covariate x_i when the other covariates are fixed

Multiple linear regression (matrix form)

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \epsilon$$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

$$\epsilon_i \rightsquigarrow N\left(\mu = 0, \sigma\right), i = 1, \ldots, n$$

Estimation
Ordinary least squares

$$\hat{\mathbf{b}} = \underset{\hat{b}_0, \hat{b}_1, \ldots \hat{b}_p}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi}$$

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y}$$

Inference of interest

$H_0 : a = 0$ versus $H_1 : a \neq 0$

$$t = \frac{\hat{a}}{se\left(\hat{a}\right)} \,|\, H_0 \rightsquigarrow N\left(\mu = 0, \sigma = 1\right)$$

p-value < 0.05, reject H_0

p-value ≥ 0.05, not reject H_0

$H_0 : b_k = 0$ versus $H_1 : b_k \neq 0$

$$t = \frac{\hat{b}_k}{se\left(\hat{b}_k\right)} \,|\, H_0 \rightsquigarrow N\left(\mu = 0, \sigma = 1\right)$$

$k = 1, \ldots, p$

0.05 is the **significance level of the test**

Accuracy

Parsimony

Principles of model selection

Multicollinearity

Interpretability

Generalisation

Forward selection

"Empty" Model

Add covariate

Add covariate

Add covariate

⋮

Stop procedure

Increased accuracy compensates increased model complexity

Increased accuracy does not compensate increased model complexity

Backward elimination

"All covariates" Model

Remove covariate

Remove covariate

Remove covariate

Stop procedure

Decreased model complexity does not have an impact on model accuracy

Decreased model complexity has an impact on model accuracy

Stepwise regression

"Empty" Model

Add covariate 1

Add covariate 2
Remove covariate 1

Add covariate 3
Remove covariates 1, 2

⋮

Stop procedure

Increased accuracy compensates increased model complexity

Increased accuracy does not compensate increased model complexity

Stepwise regression

Advantages	Disadvantages
Remove multicolinearity	Overestimation of the number of predictors
Easy automation	Inflated type I errors
Speed	Unstable to slight changes in the data