

MATH 204: PRINCIPLES OF STATISTICS 2
WINTER 2026
REGRESSION AND SUBGROUPS

If we have a continuous predictor (a covariate) X_1 and a factor predictor X_2 that takes L levels, we may consider constructing models based on them as follows to carry out a subgroup analysis. In the following description, the notation $\mathbb{1}_{\{x\}}(t)$ is the **indicator function**, where

$$\mathbb{1}_{\{x\}}(t) = \begin{cases} 1 & x = t \\ 0 & x \neq t \end{cases}.$$

Using the usual notation we could consider

MODEL 0 : 1. This model specifies that there is no effect of either predictor. The mean model can be written simply

$$\mathbb{E}[Y|x] = \beta_0$$

that is, the expected response is **constant** across the data set. This model has only $p = 1$ parameter.

MODEL 1 : X_2 . This model specifies that there is only an effect of the factor predictor X_2 . Taking the first level of the factor as the baseline group, the mean model can be written

$$\mathbb{E}[Y|x] = \beta_0 + \sum_{l=1}^{L-1} \delta_{l0} \mathbb{1}_{\{x_2\}}(l+1) = \begin{cases} \beta_0 & x_2 = 1 \\ \beta_0 + \delta_{10} & x_2 = 2 \\ \beta_0 + \delta_{20} & x_2 = 3 \\ \vdots & \vdots \\ \beta_0 + \delta_{L-1,0} & x_2 = L \end{cases}$$

that is, there is a different expected value for each subgroup indicated by the value of X_2 . This model has a total of $p = L$ parameters, of which $k = L-1$ are associated with the effect of X_2 ; these parameters are

$$\delta_{10}, \delta_{20}, \dots, \delta_{L-1,0}.$$

MODEL 2 : X_1 . This model specifies that there is only an effect of the continuous predictor X_1 . The mean model can be written

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x_1$$

and we have a simple linear regression. This model has a total of $p = 2$ parameters, of which $k = 1$ is associated with the effect of X_1 .

MODEL 3 : $X_1 + X_2$. This model specifies that there is an **additive** effect of the continuous predictor X_1 and the factor predictor X_2 . The mean model can be written

$$\mathbb{E}[Y|x] = \beta_0 + \sum_{l=1}^{L-1} \delta_{l0} \mathbb{1}_{\{x_2\}}(l+1) + \beta_1 x_1 = \begin{cases} \beta_0 + \beta_1 x_1 & x_2 = 1 \\ (\beta_0 + \delta_{10}) + \beta_1 x_1 & x_2 = 2 \\ (\beta_0 + \delta_{20}) + \beta_1 x_1 & x_2 = 3 \\ \vdots & \vdots \\ (\beta_0 + \delta_{L-1,0}) + \beta_1 x_1 & x_2 = L \end{cases}$$

that is, there are L subgroups, and for each subgroup we have a straight line model with **different intercept** but the **same slope**; that is the regression lines in the different subgroups are **parallel**. This model has a total of $p = L + 1$ parameters, of which $k = L$ are associated with the combined effects of X_1 and X_2 ; these parameters are

$$\delta_{10}, \delta_{20}, \dots, \delta_{L-1,0} \quad \text{and} \quad \beta_1$$

MODEL 4 : $X_1 + X_2 + X_1 : X_2$. This model specifies that both predictors have an effect on the outcome, and that there is an **interaction**; for example, the effect of X_1 on the outcome is **different** for different levels of X_2 . The mean model can be written

$$\begin{aligned} \mathbb{E}[Y|x] &= \beta_0 + \sum_{l=1}^{L-1} \delta_{l0} \mathbb{1}_{\{x_2\}}(l+1) + \beta_1 x_1 + \sum_{l=1}^{L-1} \delta_{l1} \mathbb{1}_{\{x_2\}}(l+1)x_1 \\ &= \begin{cases} \beta_0 + \beta_1 x_1 & x_2 = 1 \\ (\beta_0 + \delta_{10}) + (\beta_1 + \delta_{11})x_1 & x_2 = 2 \\ (\beta_0 + \delta_{20}) + (\beta_1 + \delta_{21})x_1 & x_2 = 3 \\ \vdots & \vdots \\ (\beta_0 + \delta_{L-1,0}) + (\beta_1 + \delta_{L-1,1})x_1 & x_2 = L \end{cases} \end{aligned}$$

that is, there are L subgroups, and for each subgroup we have a straight line model with **different intercept** and **different slope**. This model has a total of $p = 2L$ parameters, of which $k = 2L - 1$ are associated with the combined effects of X_1 and X_2 ; these parameters are

$$\delta_{10}, \delta_{20}, \dots, \delta_{L-1,0} \quad \text{and} \quad \beta_1, \delta_{11}, \delta_{21}, \dots, \delta_{L-1,1}$$

Note that for Model 4, we could write the model

$$\mathbb{E}[Y|x] = \beta_{l0} + \beta_{l1} x_i \quad x_2 = l$$

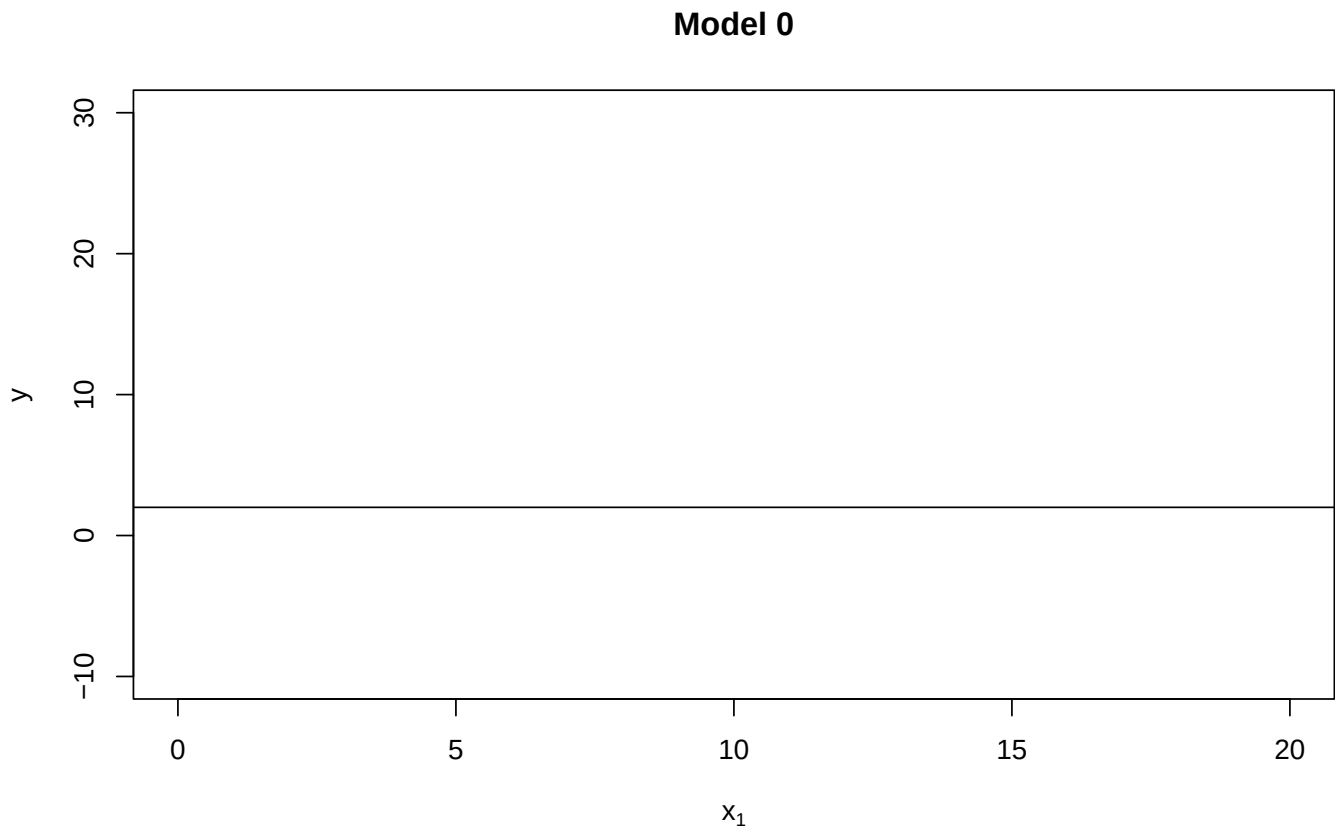
for $l = 1, 2, \dots, L$, that is, with a separate intercept and slope for each subgroup.

Note that:

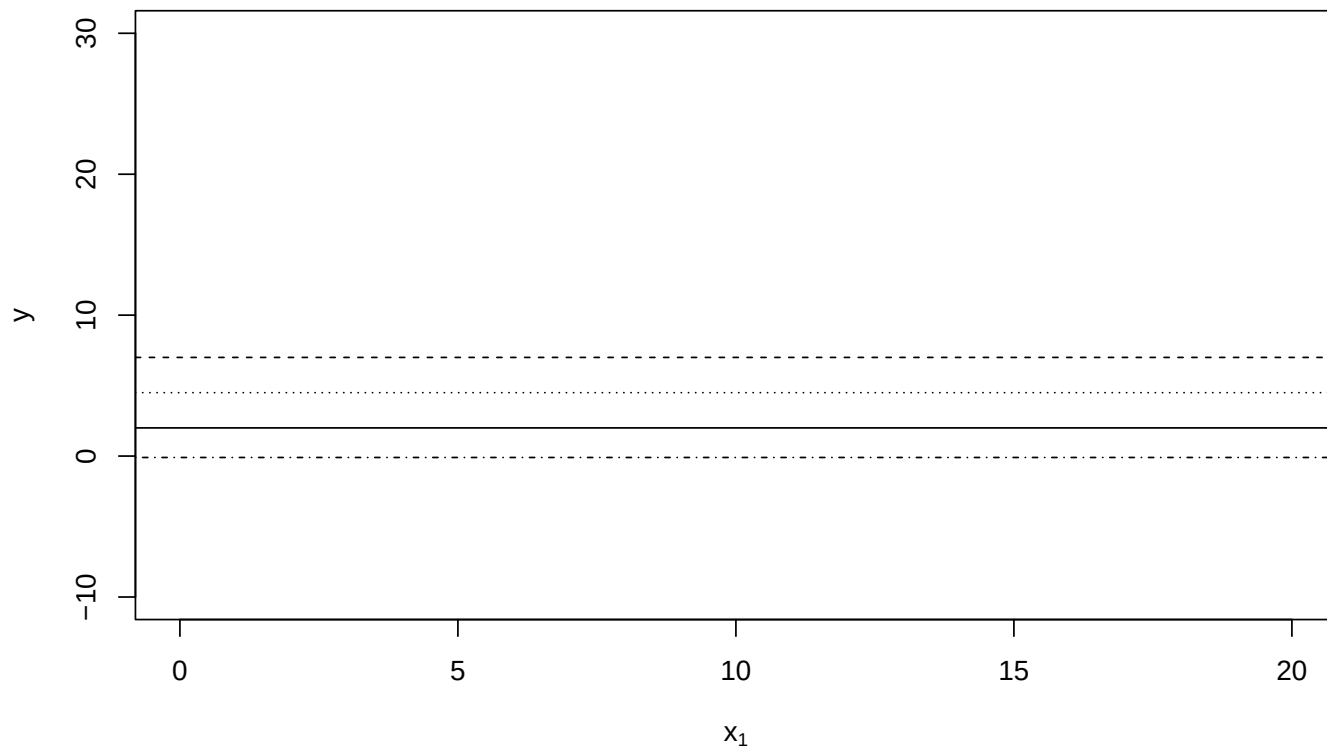
- All of the models are examples of the linear model, that is, the parameters of the mean model always appear in linear form.
- The parameters $\delta_{10}, \dots, \delta_{L-1,0}$ and $\delta_{11}, \dots, \delta_{L-1,1}$ are **contrast** parameters that quantify how much the intercepts and slopes for the subgroups differ from the baseline group.
- We can obtain all of the models by putting restrictions on the parameters of Model 4.

Example: $L = 4$. Here are plots of the mean models for Model 0 to Model 4

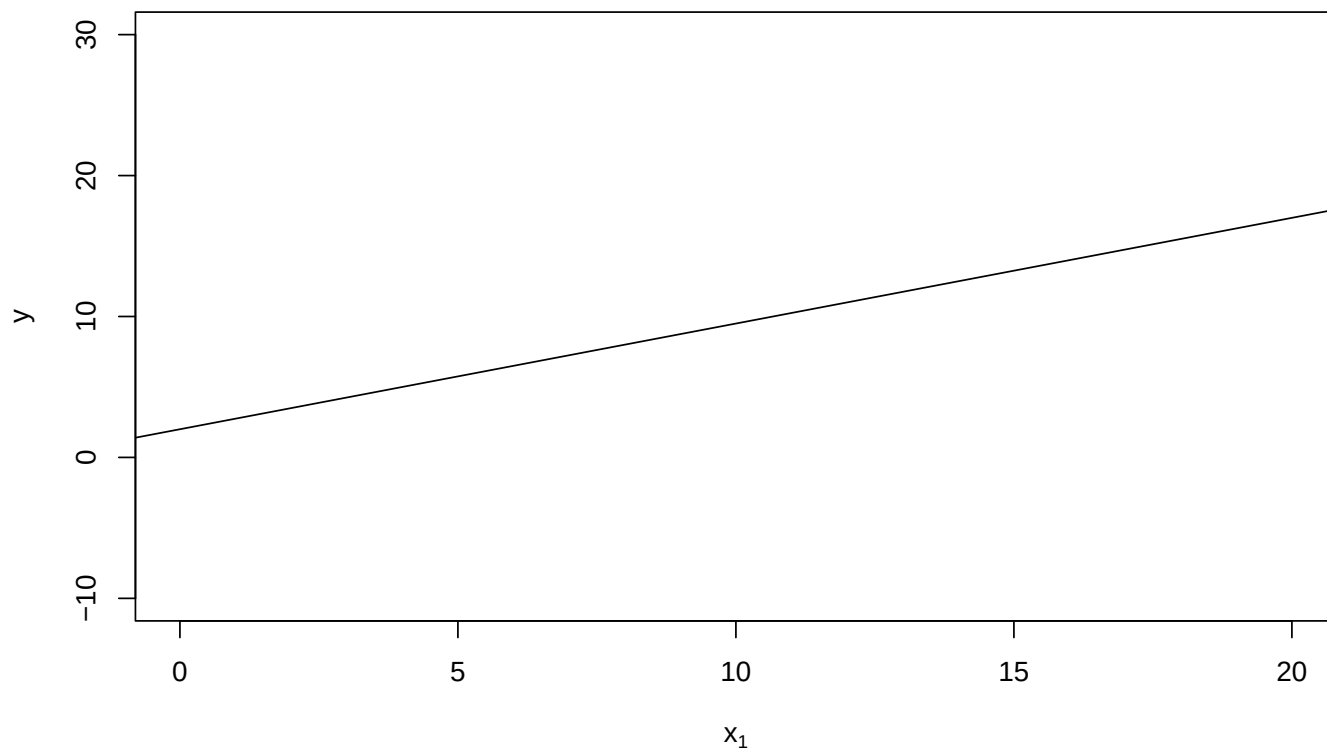
```
> L<-4
> be0<-2.0
> be1<-0.75
> be1vec<-c(-0.25,1.0,0.2)
> be2vec<-c(5,2.5,-2.1)
> par(mar=c(4,4,4,0))
> plot(0,0,type='n',xlab=expression(x[1]),ylab='y',
+      xlim=range(0,20),ylim=range(-10,30))
> abline(be0,0)
> title('Model 0')
```



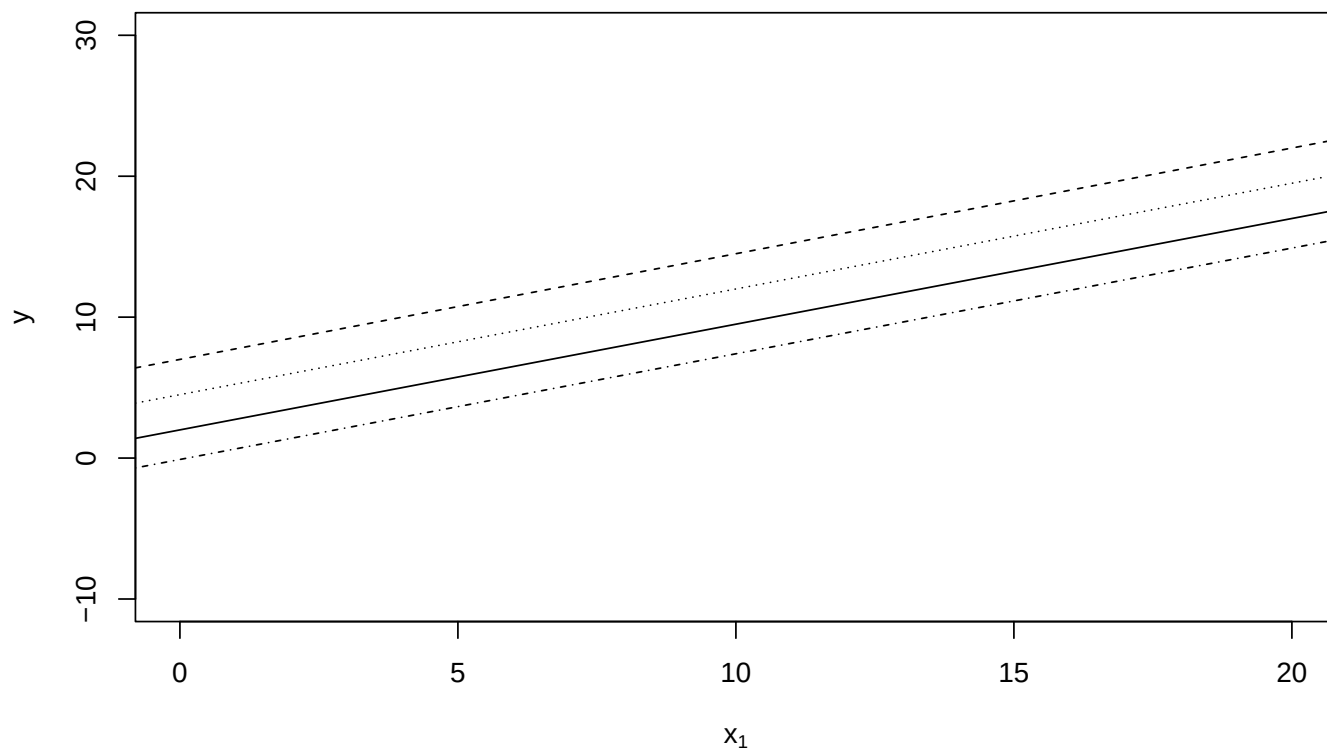
Model 1



Model 2



Model 3



Model 4

