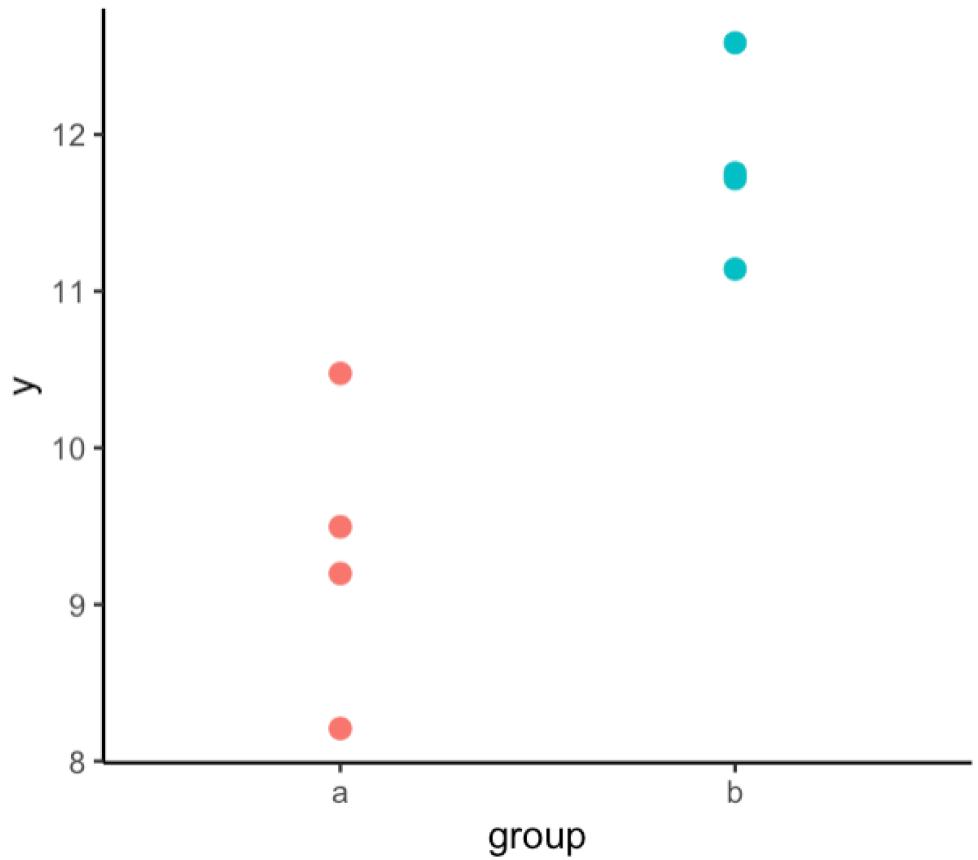
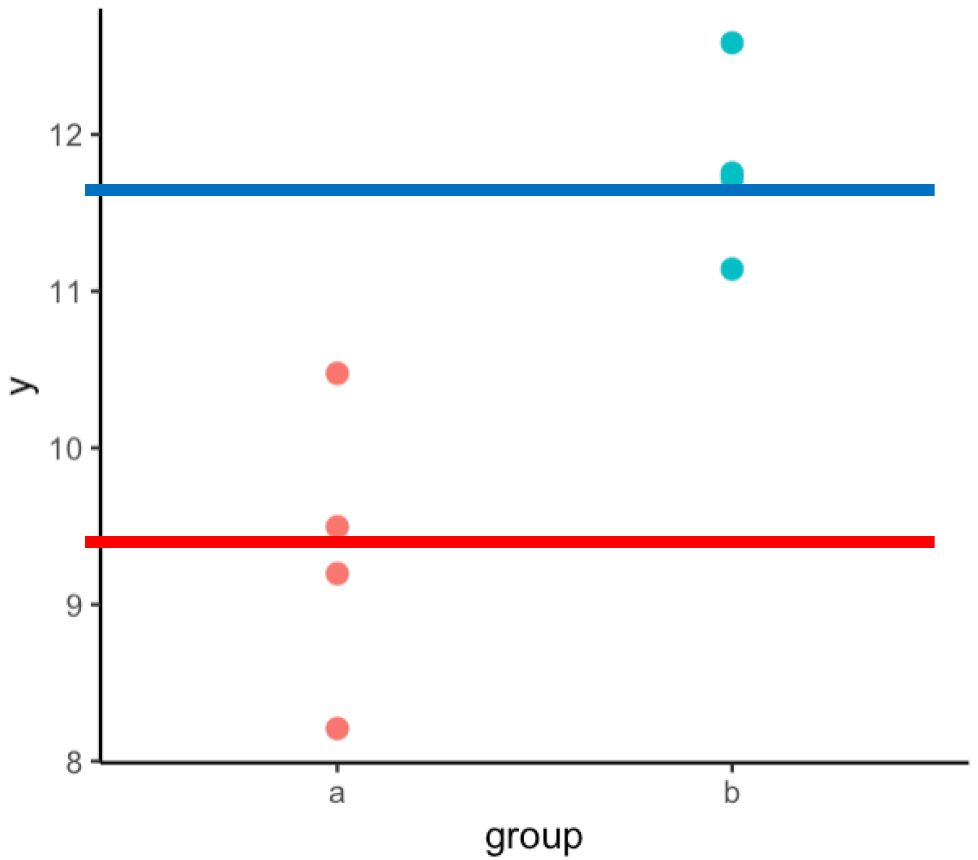
PLS 206 **Applied Multivariate Modeling in Agricultural and Environmental Sciences**

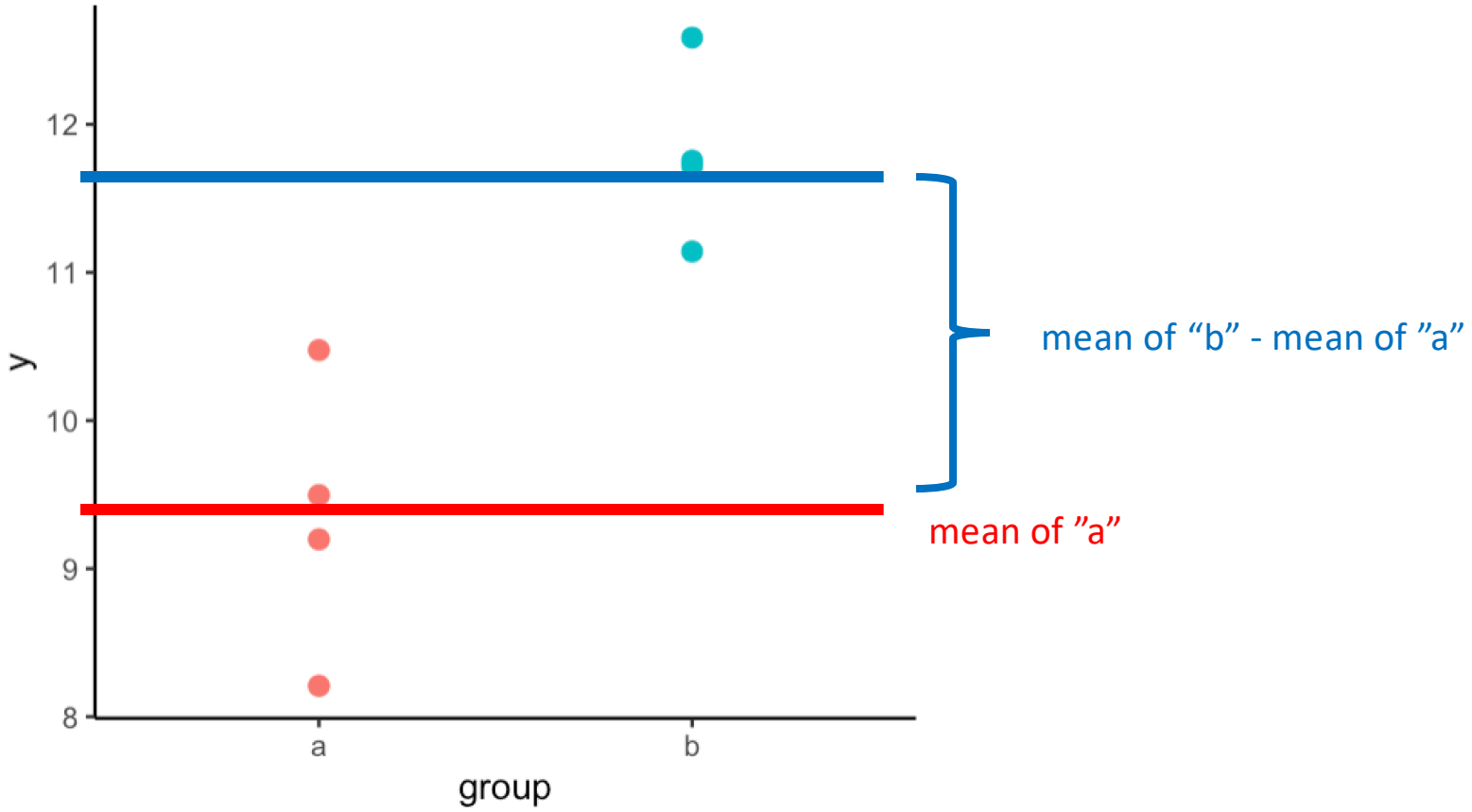
General Linear Models

General linear model

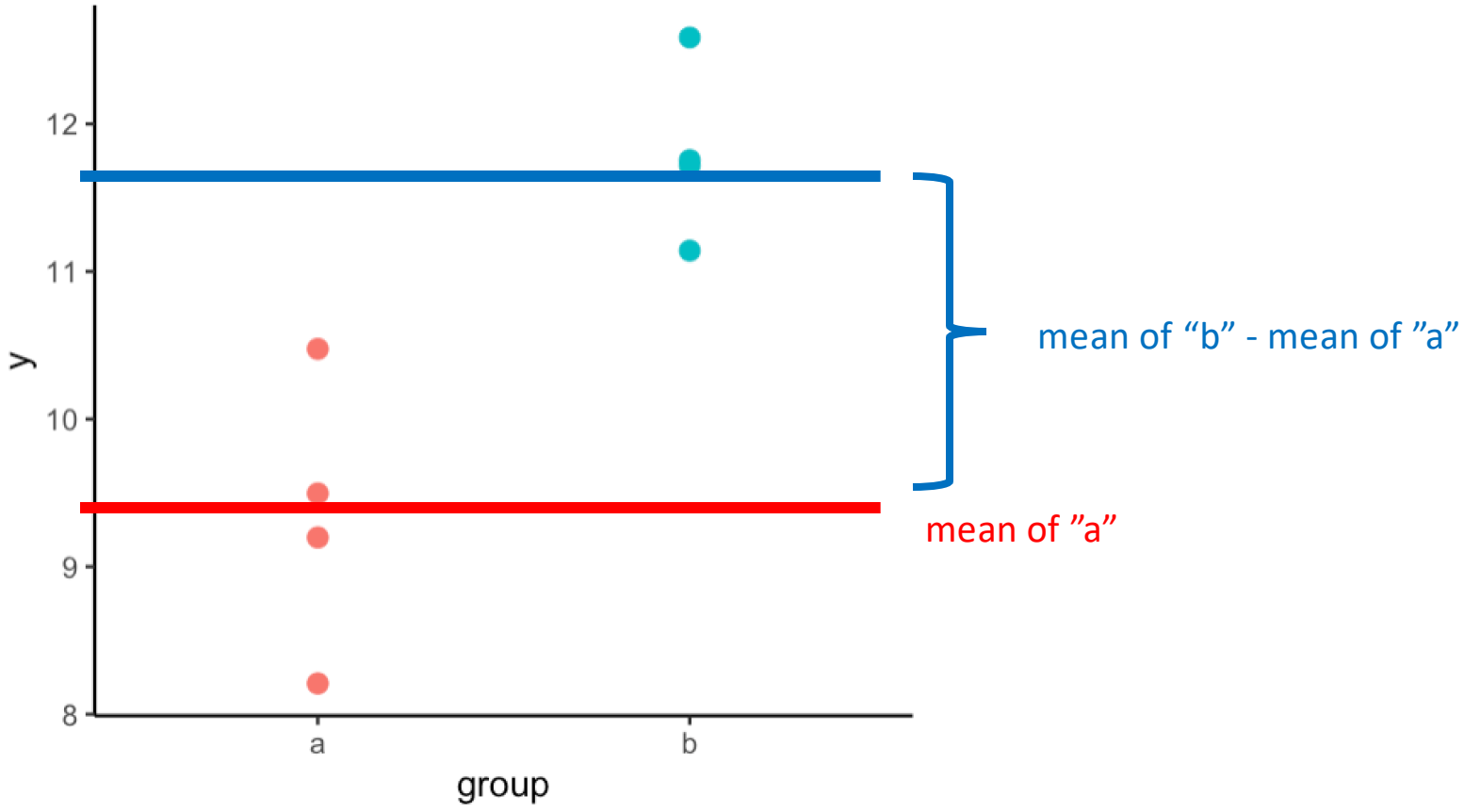
- Categorical predictors
- Design matrix
- Hypothesis testing
- Categorical and linear predictors
- Interactions


12
11
10
9
8
y
a
b
group


12
11
10
9
8
y
a
b
group


12
11
10
9
8
y
a
b
group
mean of “b” - mean of ”a”
mean of ”a”

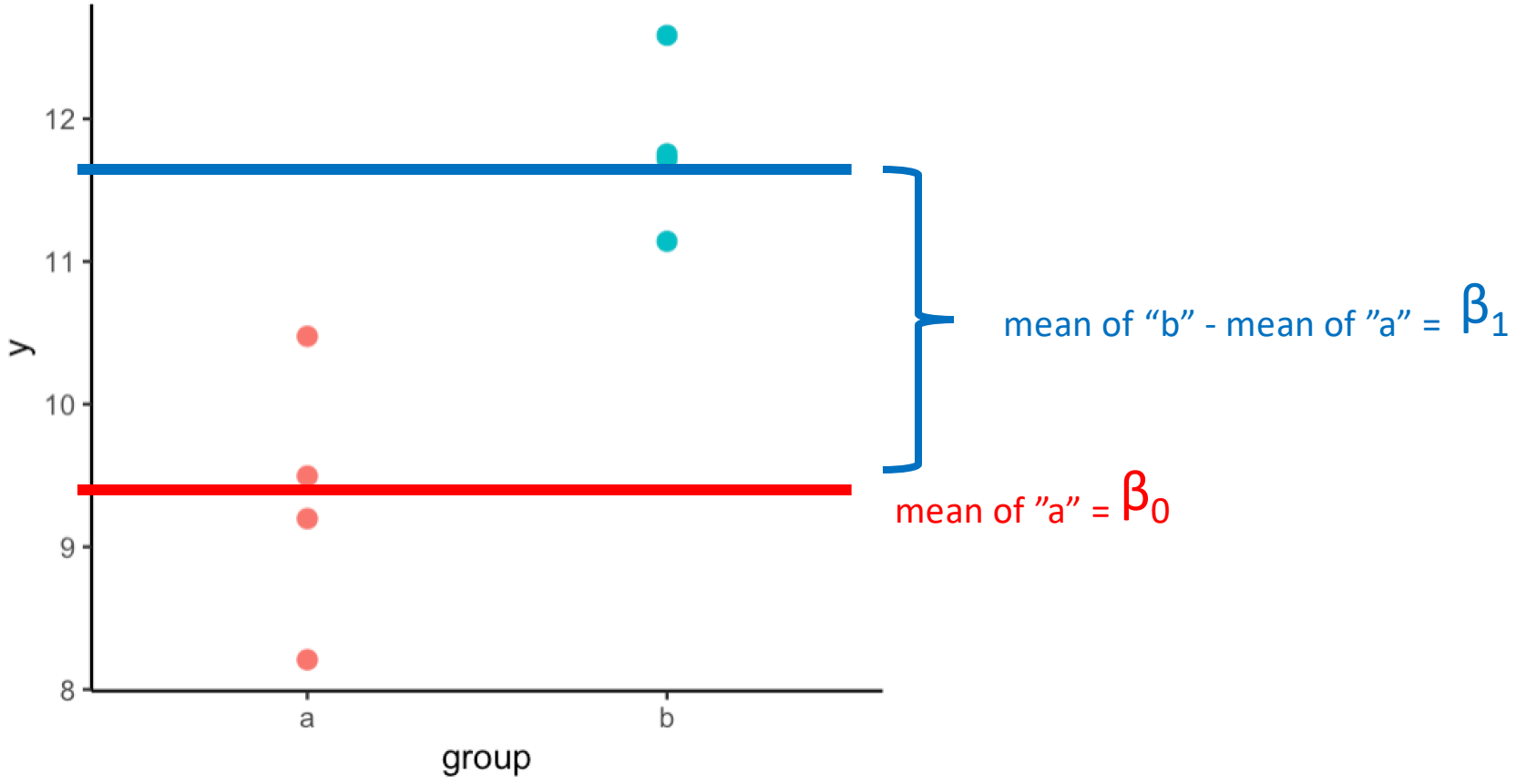
$$Y_i = \beta_{0\ (\text{mean of "a"})} + \beta_{1\ (\text{mean of "b" - mean of "a"})} * X_i + \varepsilon_i$$

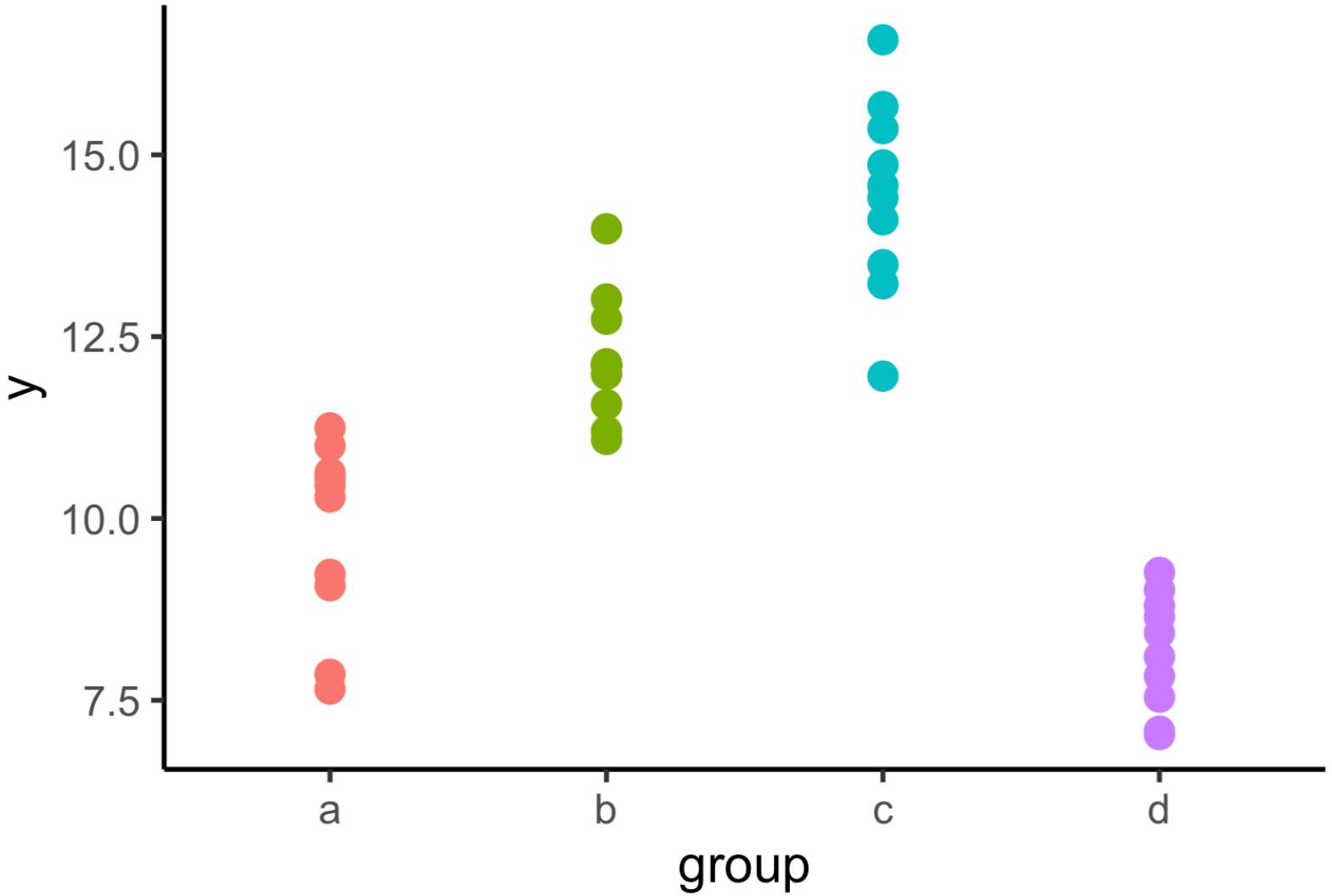


If group is “b” $X_i = 1$, if group is not “b” $X_i = 0$

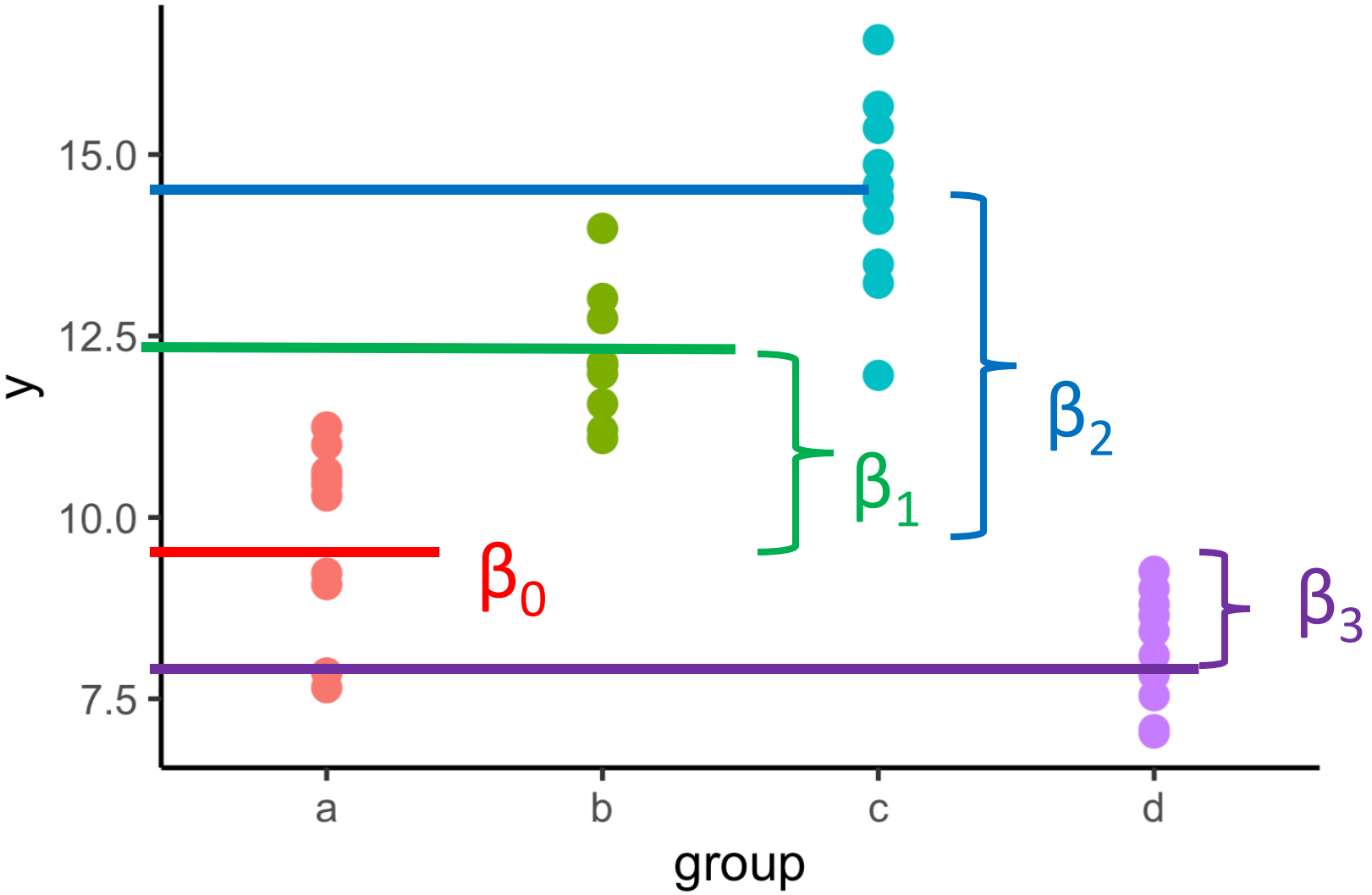
$$Y_i = \underbrace{\beta_{0\ (\text{mean of "a"})} + \beta_{1\ (\text{mean of "b" - mean of "a"})} * X_i}_{\text{Predicted value}} + \varepsilon_i$$

Residual

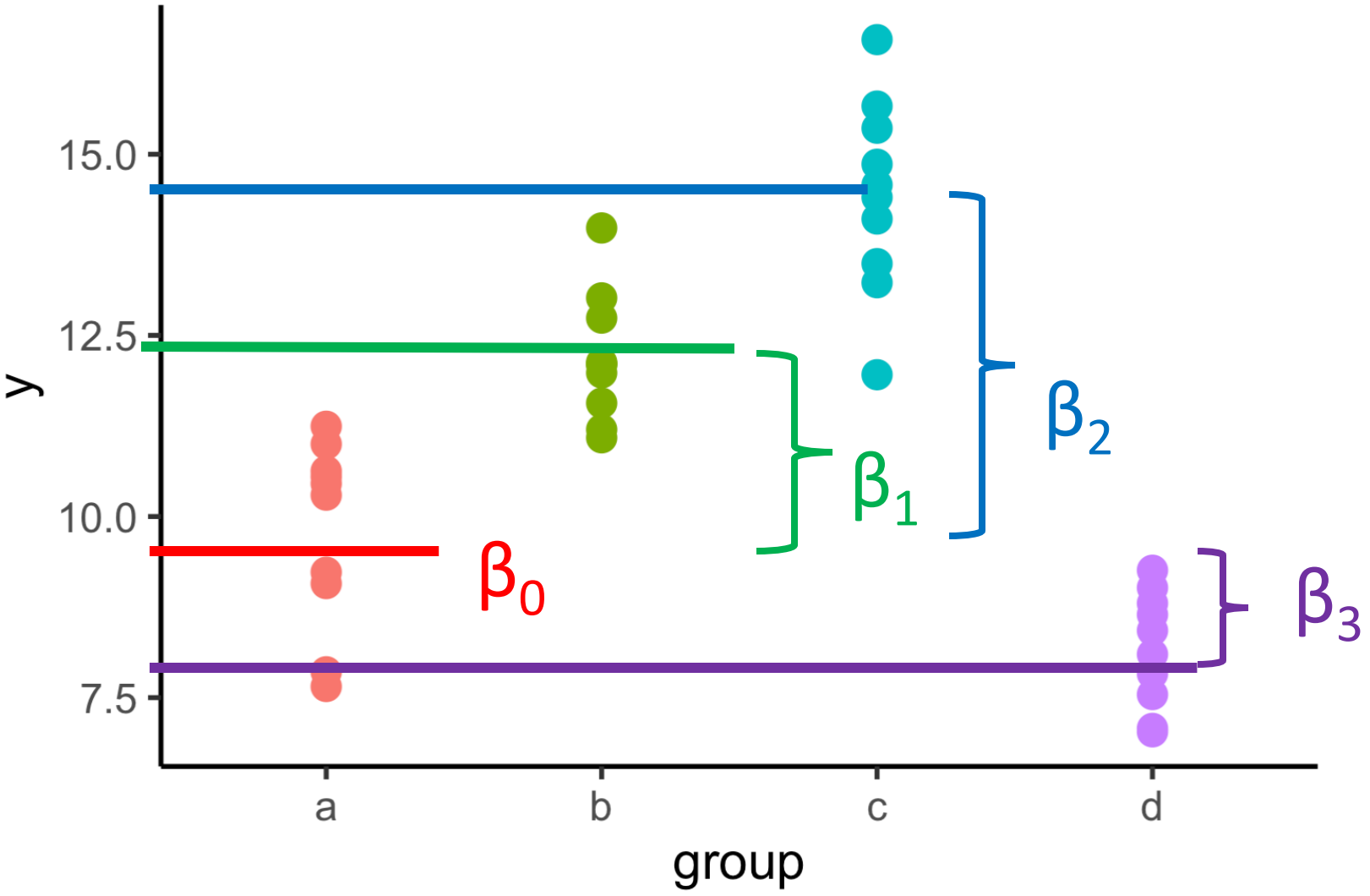



15.0
12.5
10.0
7.5
y
a
b
c
d
group

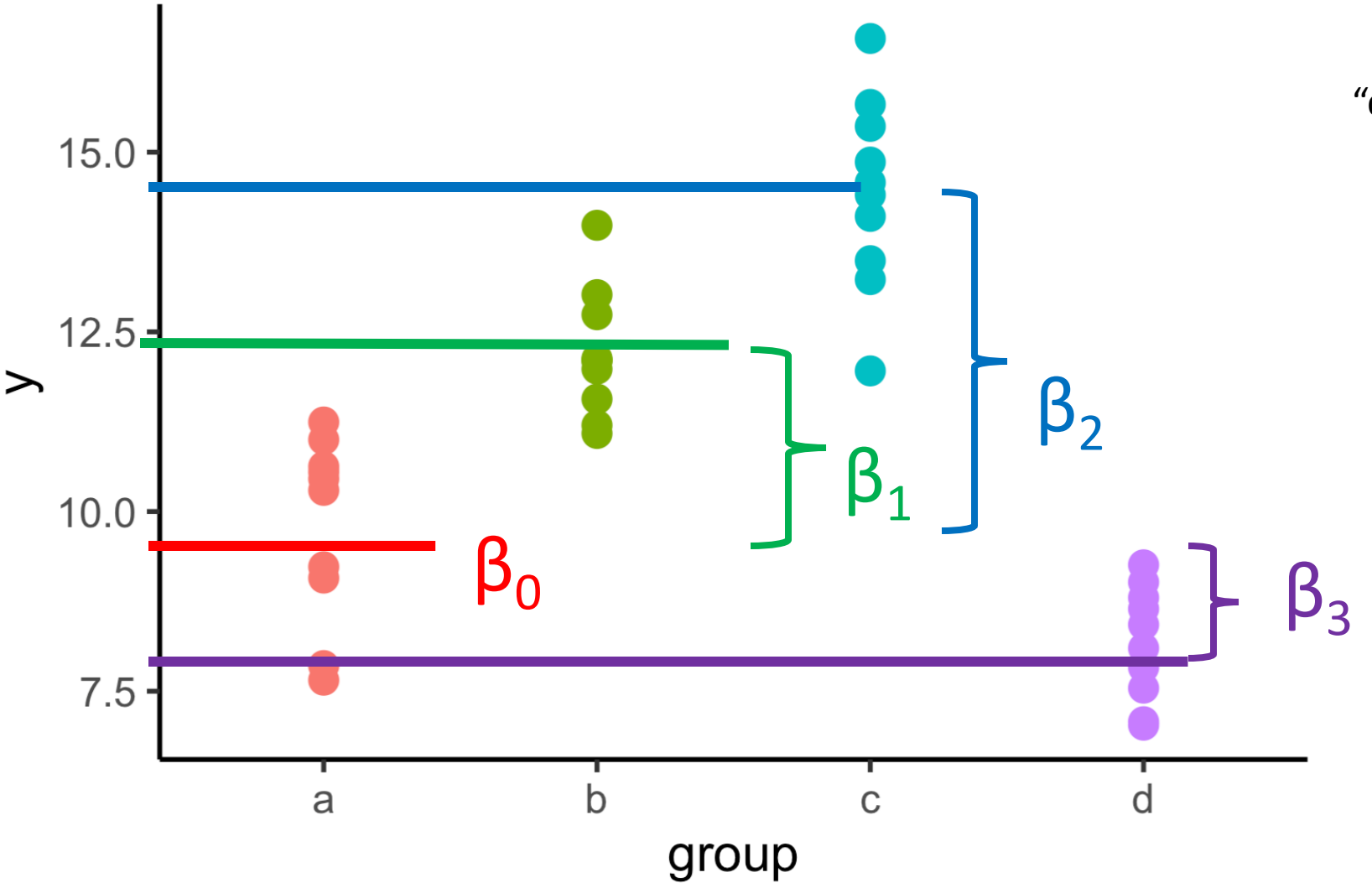
$$Y_i = \beta^* X_i + \varepsilon_i$$



$$Y_i = \beta^* X_i + \varepsilon_i \quad \dashrightarrow \quad Y_i = \beta_0 + \beta_1 * X_{i,b} + \beta_2 * X_{i,c} + \beta_3 * X_{i,d} + \varepsilon_i$$



$$Y_i = \beta^* X_i + \varepsilon_i \quad \dashrightarrow \quad Y_i = \beta_0 + \beta_1 * X_{i,b} + \beta_2 * X_{i,c} + \beta_3 * X_{i,d} + \varepsilon_i$$



"data$group"	Intercept	X_b	X_c	X_d
a	1	0	0	0
a	1	0	0	0
a	1	0	0	0
b	1	1	0	0
b	1	1	0	0
b	1	1	0	0
c	1	0	1	0
c	1	0	1	0
c	1	0	1	0
d	1	0	0	1
d	1	0	0	1
d	1	0	0	1

Design Matrices

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \end{bmatrix}$$

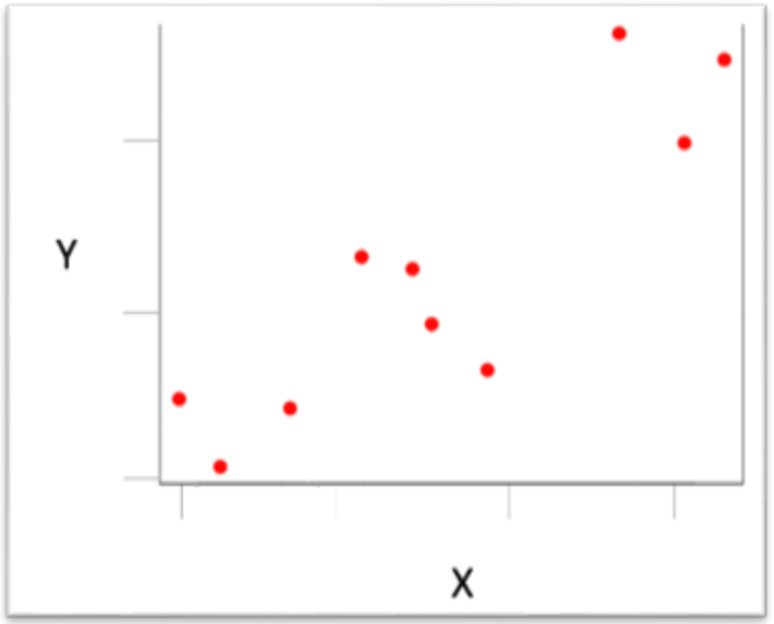
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$



$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$

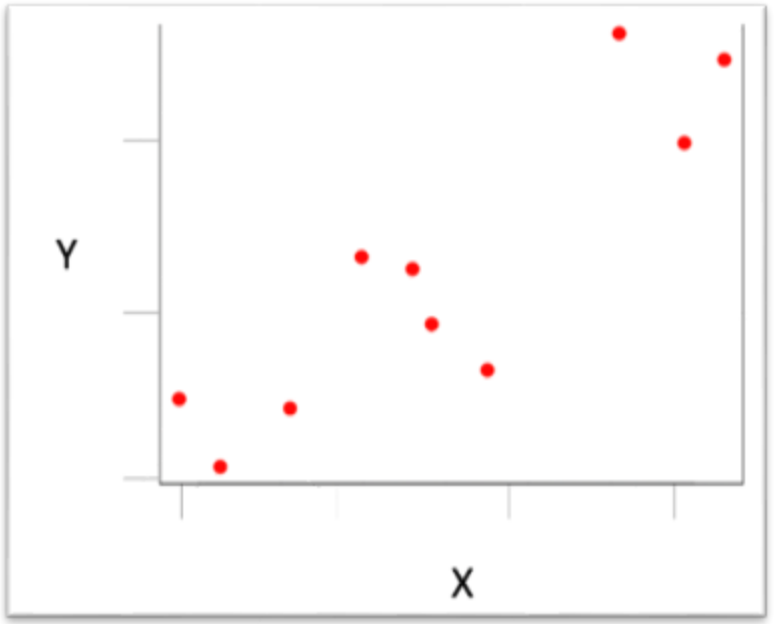
Predicted values

intercept

Slope of variable x

$Y =$

	x	
$\beta_0 +$	1.21489482	$* \beta_1$
$\beta_0 +$	0.11942794	$* \beta_1$
$\beta_0 +$	-2.24123708	$* \beta_1$
$\beta_0 +$	-3.10076793	$* \beta_1$
$\beta_0 +$	-0.11480229	$* \beta_1$
$\beta_0 +$	1.66663629	$* \beta_1$
$\beta_0 +$	-0.08880993	$* \beta_1$



$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$

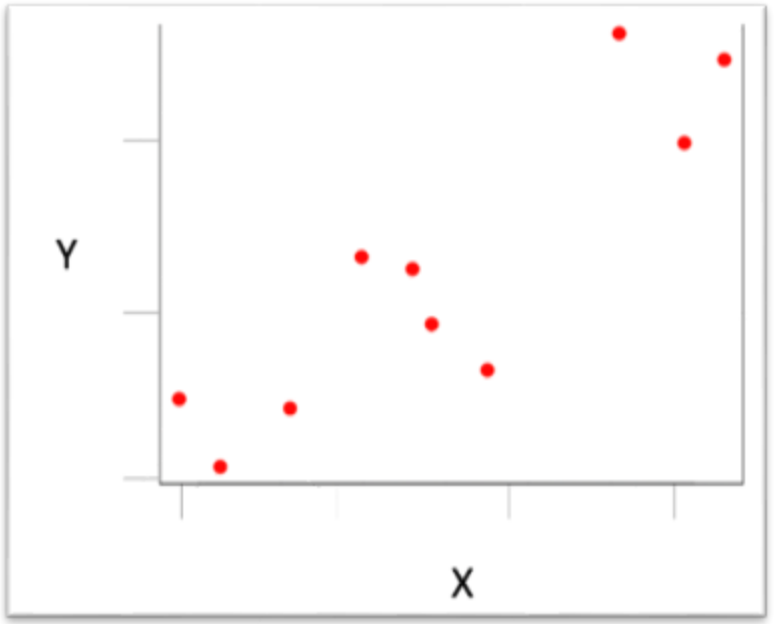
Predicted values

artificial "Intercept" variable

Slope of variable x

$$Y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} * \beta_0 + \begin{bmatrix} x \\ 1.21489482 \\ 0.11942794 \\ -2.24123708 \\ -3.10076793 \\ -0.11480229 \\ 1.66663629 \\ -0.08880993 \\ \end{bmatrix} * \beta_1$$

intercept



$$Y_i = \beta_0 + \beta_1 * X_i + \varepsilon_i$$

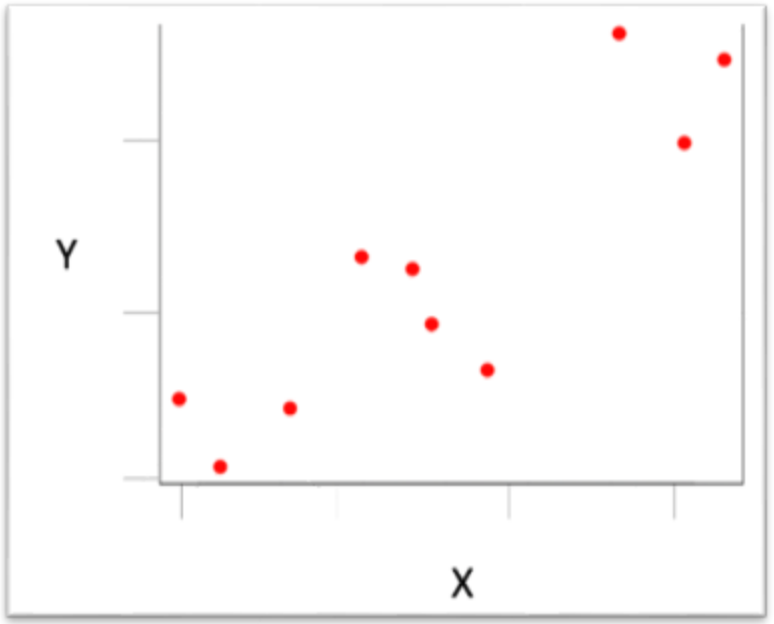
Predicted values

artificial "Intercept" variable

Slope of variable x

$Y =$

intercept		x	
1	$* \beta_0 +$	1.21489482	$* \beta_1$
1	$* \beta_0 +$	0.11942794	$* \beta_1$
1	$* \beta_0 +$	-2.24123708	$* \beta_1$
1	$* \beta_0 +$	-3.10076793	$* \beta_1$
1	$* \beta_0 +$	-0.11480229	$* \beta_1$
1	$* \beta_0 +$	1.66663629	$* \beta_1$
1	$* \beta_0 +$	-0.08880993	$* \beta_1$

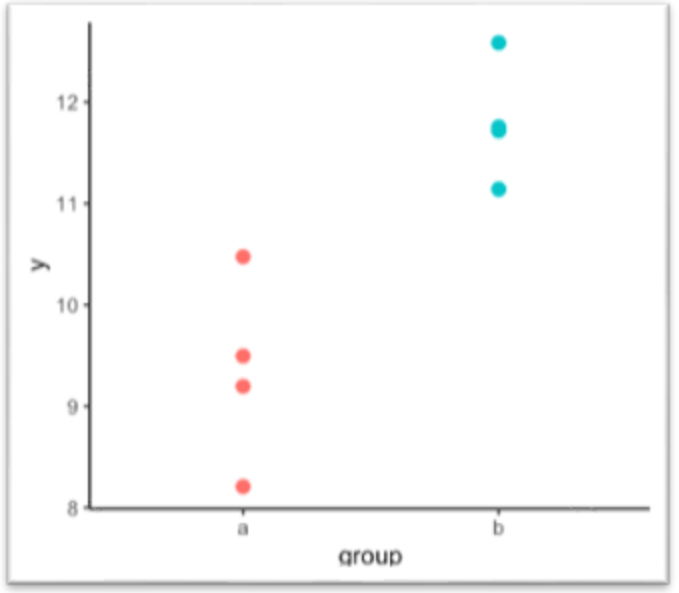


Design matrix

(Intercept)	x
1	1.21489482
1	0.11942794
1	-2.24123708
1	-3.10076793
1	-0.11480229
1	1.66663629
1	-0.08880993

$$Y_i = \underbrace{\beta_{0\ (\text{mean of "a"})} + \beta_{1\ (\text{mean of "b" - mean of "a"})} * X}_{\text{Predicted values}} + \varepsilon_i$$

Residual



artificial "Intercept" variable

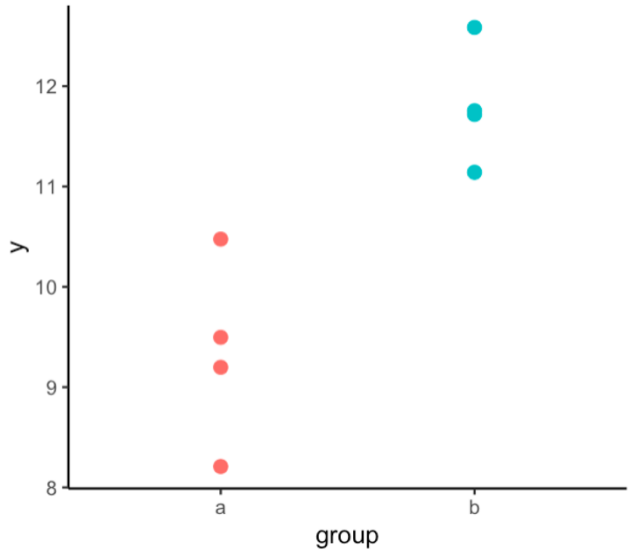
Artificial 'is it group "b"' variable?

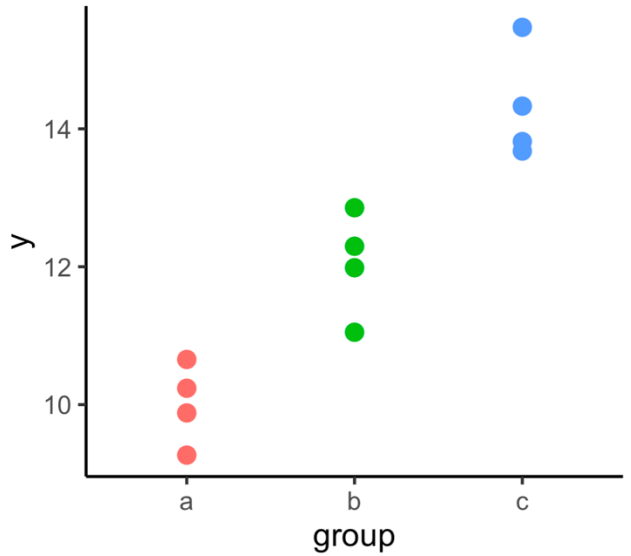
intercept

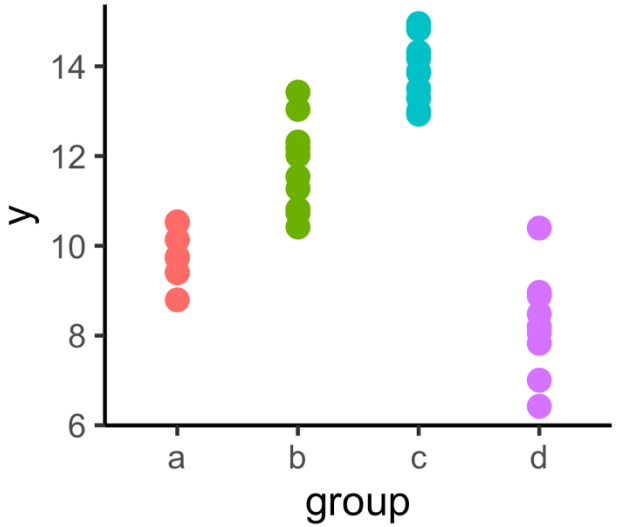
$$Y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} * \beta_0 + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} * \beta_1$$

Design matrix

(Intercept)	groupb
1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1







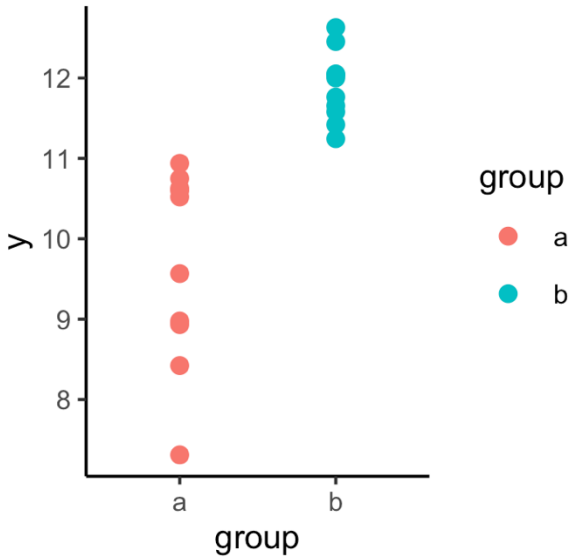
(Intercept)	groupb
1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1

(Intercept)	groupb	groupc
1	0	0
1	0	0
1	0	0
1	0	0
1	1	0
1	1	0
1	1	0
1	1	0
1	0	1
1	0	1
1	0	1
1	0	1

(Intercept)	groupb	groupc	groupd
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	0	0	0
1	1	0	0
1	1	0	0

Design matrices

Understanding how linear models deal with categorical variables (through design matrix) is important for interpreting model results



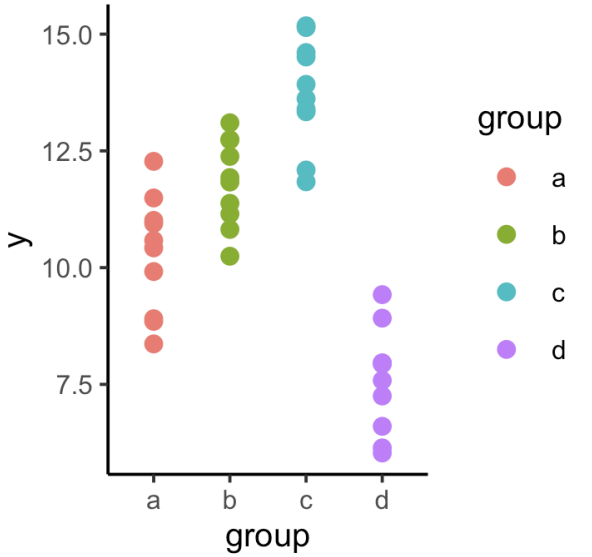

```
> model<-lm(y~group, data)
> summary(model)

Call:
lm(formula = y ~ group, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35372 -0.50797  0.00955  0.77309  1.27465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.6645     0.2903  33.289  < 2e-16 ***
groupb        2.2193     0.4106   5.405 3.89e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9181 on 18 degrees of freedom
Multiple R-squared:  0.6188,    Adjusted R-squared:  0.5976
F-statistic: 29.22 on 1 and 18 DF,  p-value: 3.893e-05
```

Mean of group a

Difference in mean for **group b**

** ***Same as*** t.test(y~group, data, var.equal = T)

More levels…




```
> model<-lm(y~group, data)
> summary(model)

Call:
lm(formula = y ~ group, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9260 -0.7554  0.1215  0.7709  1.9988

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.2771     0.3545  28.990  < 2e-16 ***
groupb        1.5542     0.5013   3.100  0.00375 **
groupc        3.4873     0.5013   6.956 3.75e-08 ***
groupd       -2.6970     0.5013  -5.380 4.69e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.121 on 36 degrees of freedom
Multiple R-squared:  0.819,     Adjusted R-squared:  0.804
F-statistic: 54.31 on 3 and 36 DF,  p-value: 1.927e-13
```

Mean of **group a**

Difference in mean between **group b and group a**

Difference in mean between **group c and group a**

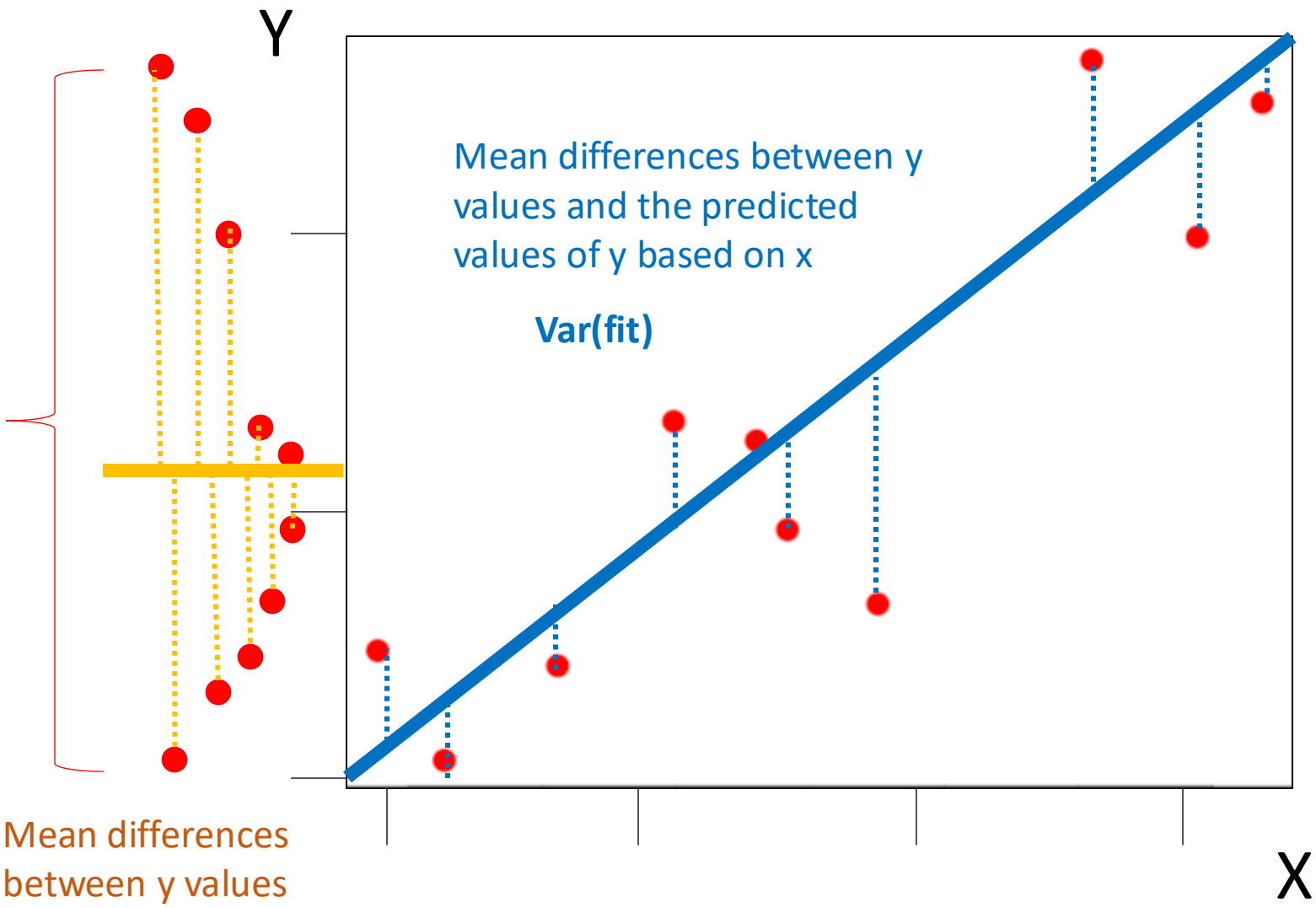
Difference in mean between **group d and group a**

***** compare all groups with*** TukeyHSD(aov(model))

R^2 and hypothesis testing

Var = Sum of Squares / (n-1)

$$R^2 = \frac{\text{var(mean)} - \text{var(fit)}}{\text{var(mean)}}$$

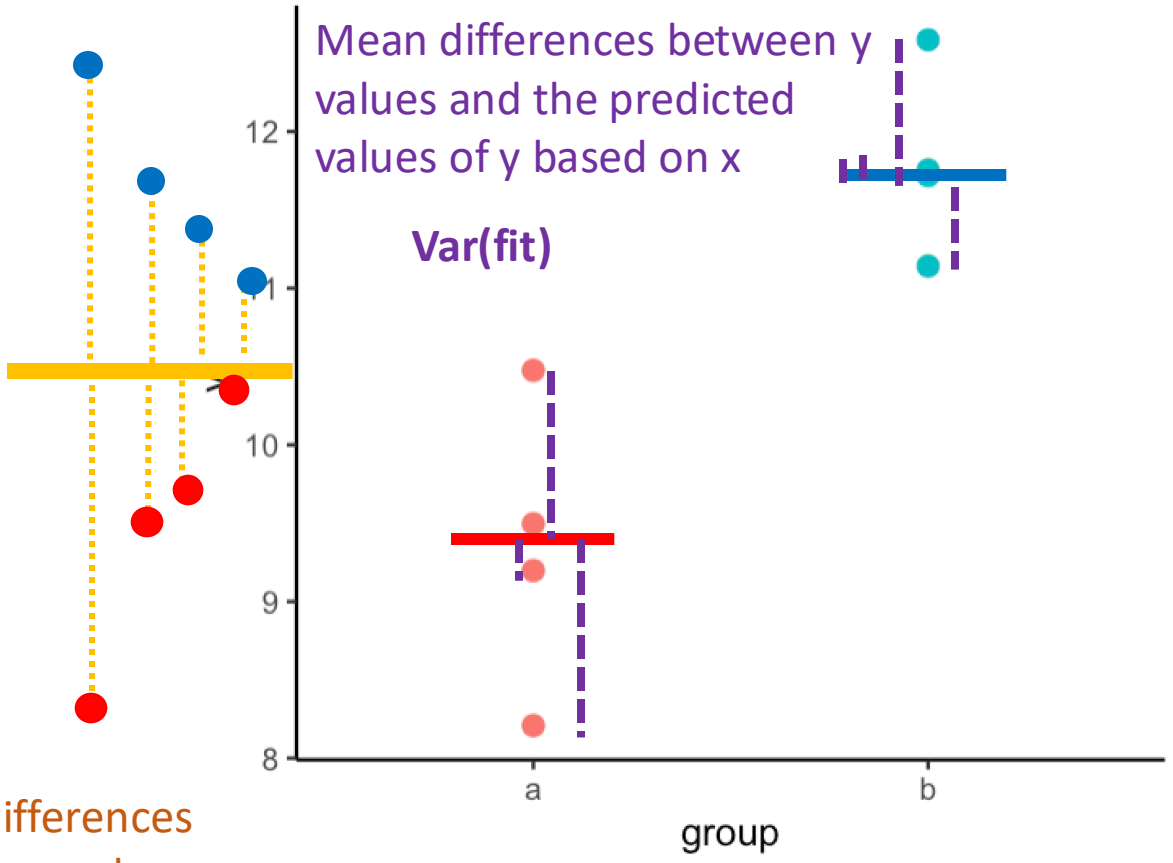


*__Fitted__ values is a synonym for __predicted__ values of y based on x

$$R^2 = \frac{\text{var(mean)} - \text{var(fit)}}{\text{var(mean)}}$$

test statistic

Does the fitted model explain more variance than expected compared to the simple model (variance around mean?



Mean differences between y values and the mean of y

Var(mean)

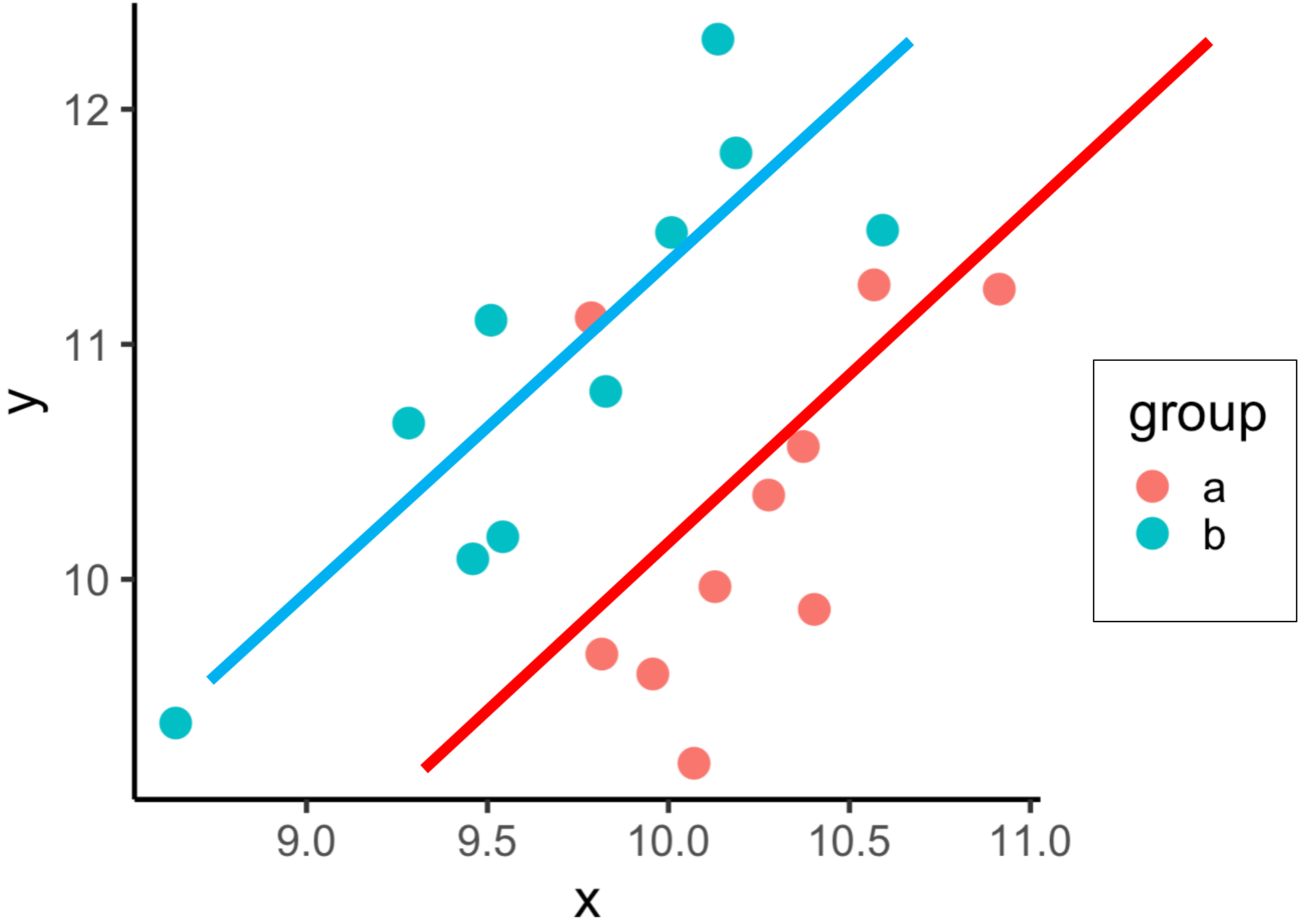
Combining groups and categorical variables

Categorical and continuous variables

	Estimate
(Intercept)	-2.1984
groupb	1.9734
x	1.2470

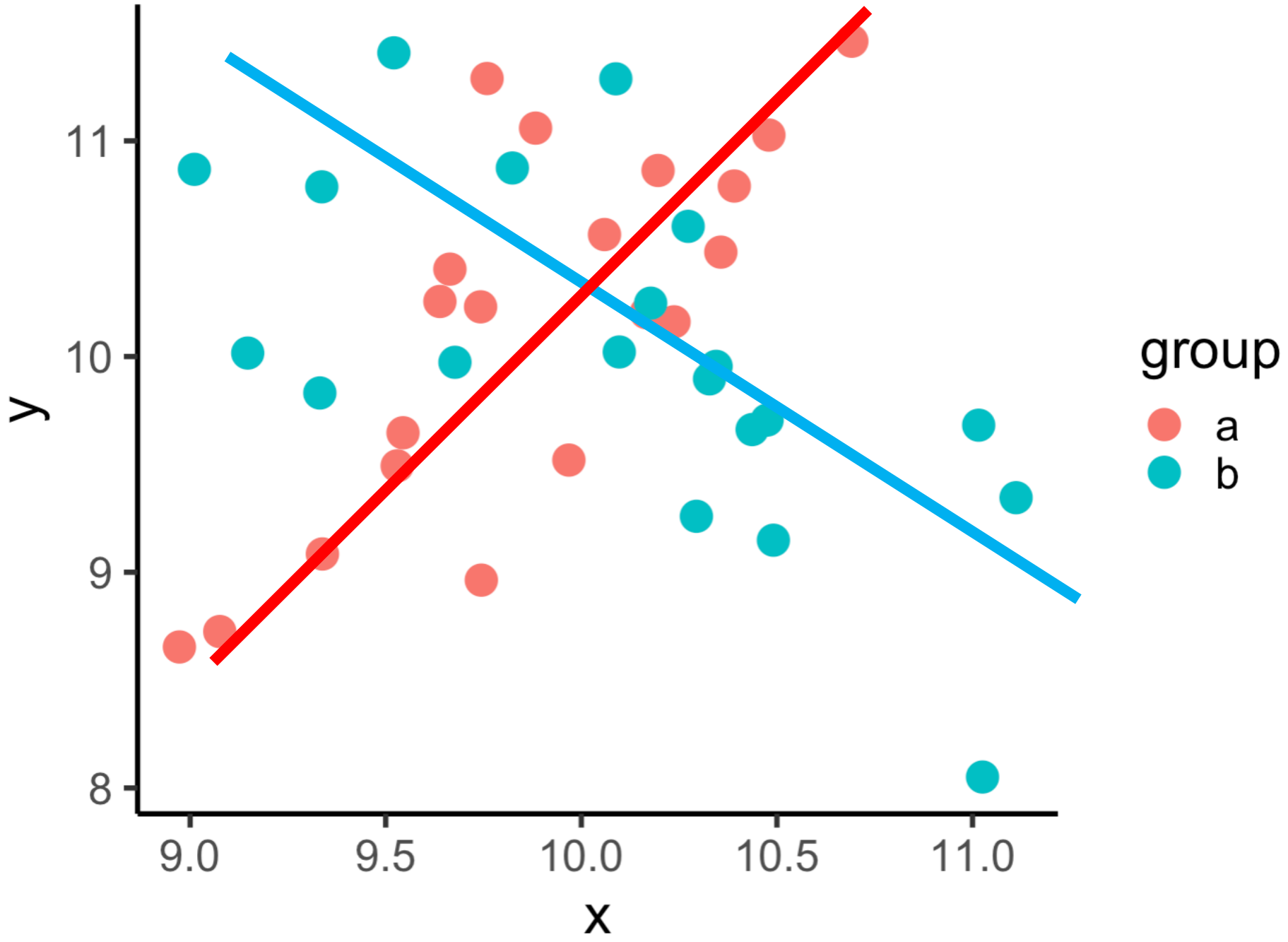
Even though we have groups in the data…

- (Intercept) is not the mean of group a, it is the intercept of group a

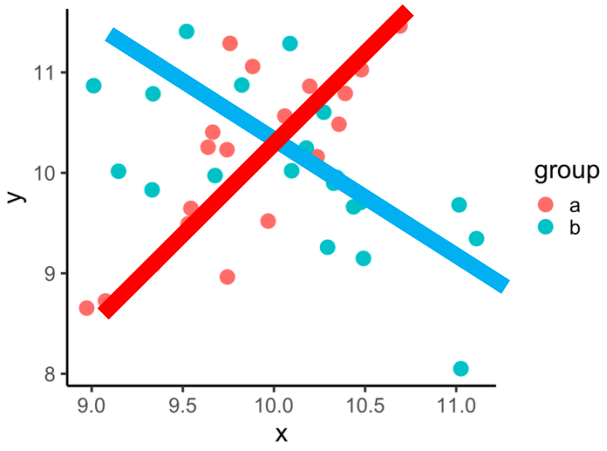


Interactions:
Challenges of interpreting "main" effects

	Estimate
(Intercept)	2.0290
groupb	18.1515
x	0.8159
groupb:x	-1.8374



Model summaries with interactions: interpretations




```
> model<-lm(y~group*x, data)
> summary(model)

Call:
lm(formula = y ~ group * x, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42722 -0.21549 -0.03956  0.18202  0.61148

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2038     0.1311  -1.554    0.129
groupb        2.2741     0.1914  11.879 5.15e-14 ***
x             1.0721     0.1077   9.958 6.96e-12 ***
groupb:x     -2.0597     0.1637 -12.586 9.54e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2854 on 36 degrees of freedom
Multiple R-squared:  0.8216,    Adjusted R-squared:  0.8067
F-statistic: 55.25 on 3 and 36 DF,  p-value: 1.497e-13
```

Intercept of group a

Difference in intercept for **group b**

Slope of group a

Difference in slope for **group b**