Statistics with Spa ows

Lecture 11-b

Julia Schroeder

Julia.schroeder@imperial.ac.uk

Outline

- Linear models – going big
- Categorical and continuous predictors

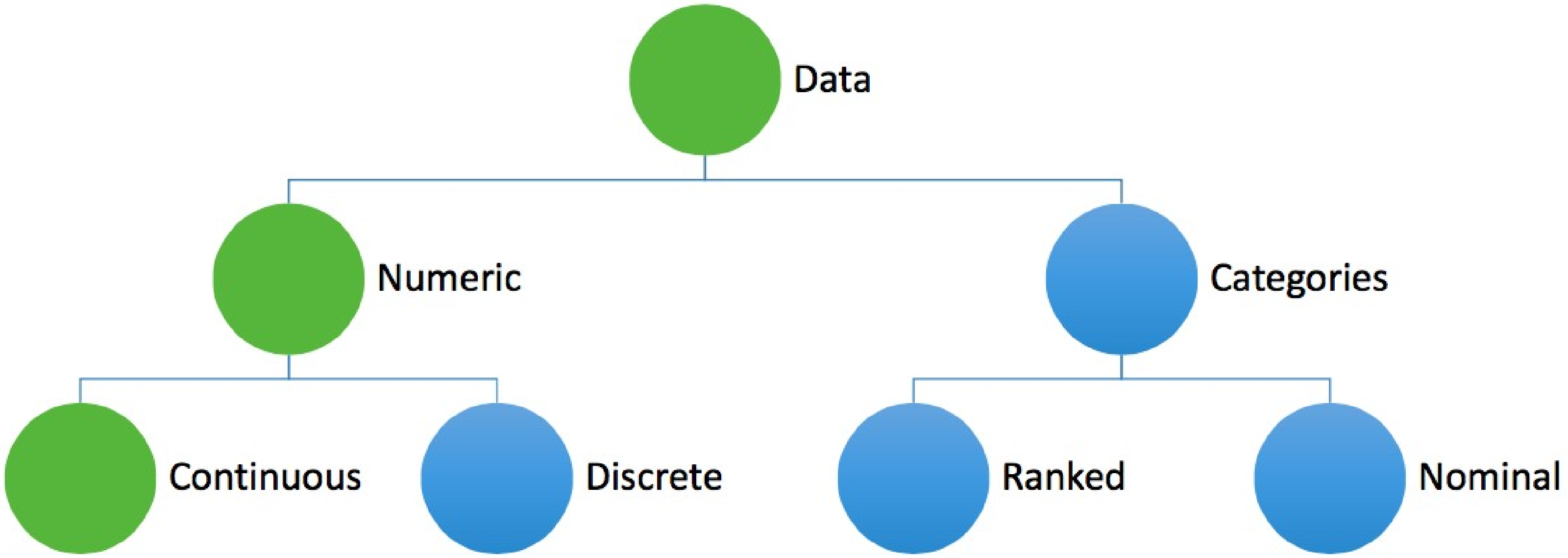
Multiple linear models

```
lm(response~explanatory)
```

Multiple linear models

```
lm(response~explanatory)
```

Data types



Multiple linear models

Response y:

Continuous

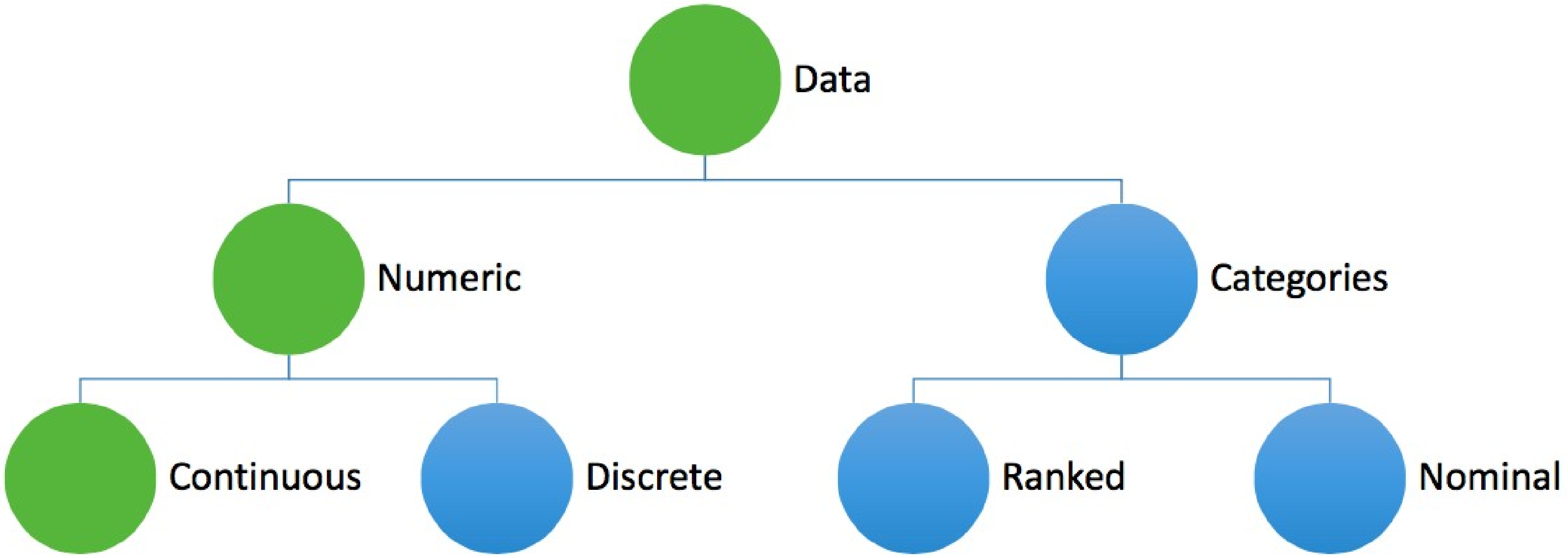
Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

```
lm(response~explanatory)
```

Data types



Multiple linear models

Response y:
Continuous

Explanatory x:
Continuous (tarsus, wing, mass)
Categorical (Sex, Year, Observer, BirdID)

```
lm(response~explanatory)
```

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

Multiple linear models

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

```
lm(response~explanatory)
```

$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$${}_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i2} + \varepsilon$$

Multiple linear models

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

`lm(response~explanatory)`

We can have more than one explanatory variable!

$${}_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i2} + \varepsilon$$

Multiple linear models

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

```
lm(response~explanatory)
```

We can have more than one explanatory variable!

We can even mix continuous and factorial explanatory variables!

$${}_{i}=b_0+b_1 x_{i0}+b_2 x_{i1}+b_3 x_{i2}+\varepsilon$$

Let's try this

- b_0 = intercept

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + \varepsilon$$

- b_1 = estimates effect of continuous variable x_0
- b_2 = estimates effect of 2-level factor x_1

Let’s try this

- b_0 = intercept

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + \varepsilon$$

- b_1 = estimates effect of continuous variable x_0 (tarsus)
- b_2 = estimates effect of 2-level factor x_1 (sex)

Let’s try this

- b_0 = intercept

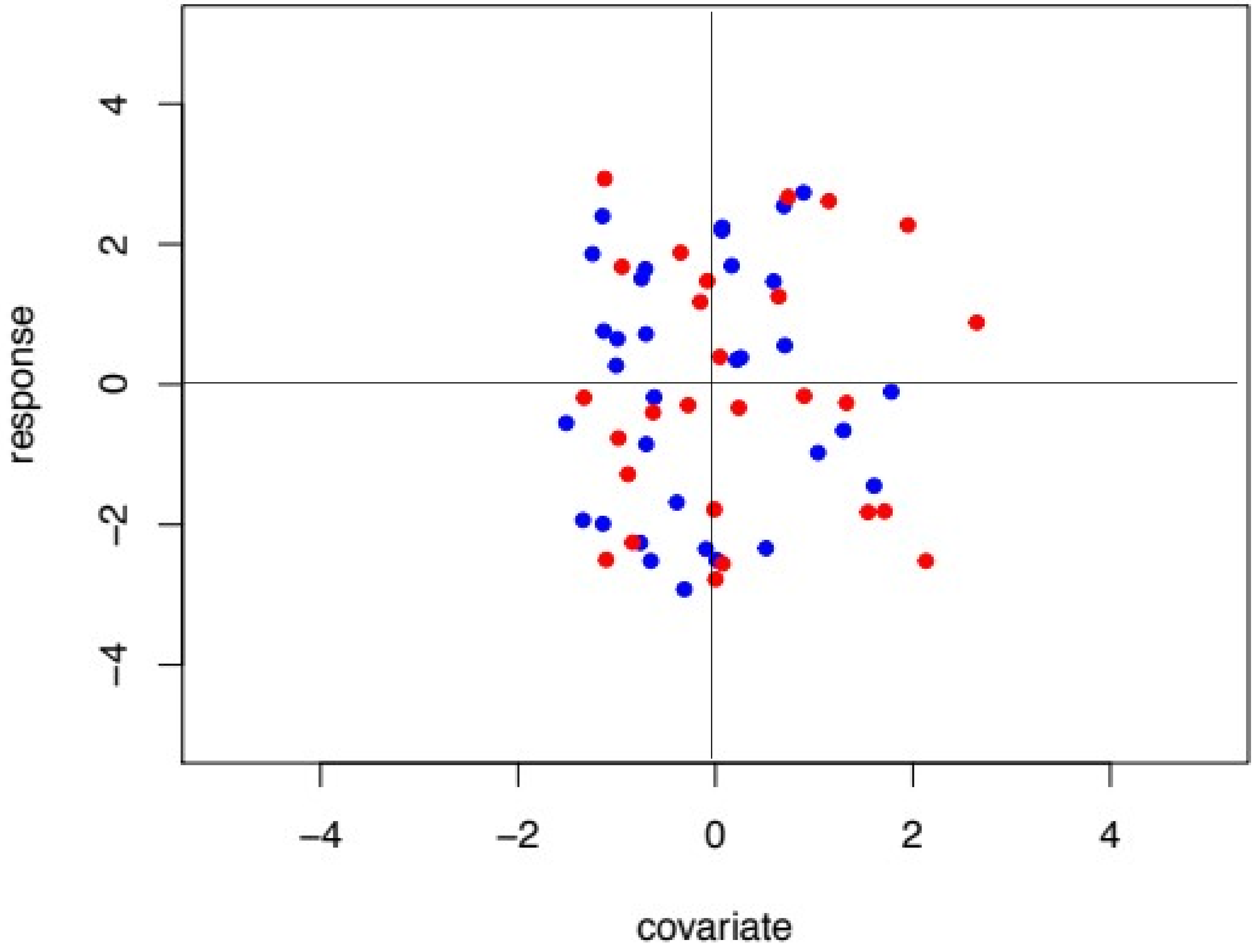
$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + \varepsilon$$

- b_1 = estimates effect of continuous variable x_0 (tarsus)
- b_2 = estimates effect of 2-level factor x_1 (sex)

- Sex will be re-coded internally. Females are 0 (and blue).

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} +$$

Let’s try this



Sex: $y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + \varepsilon_i$

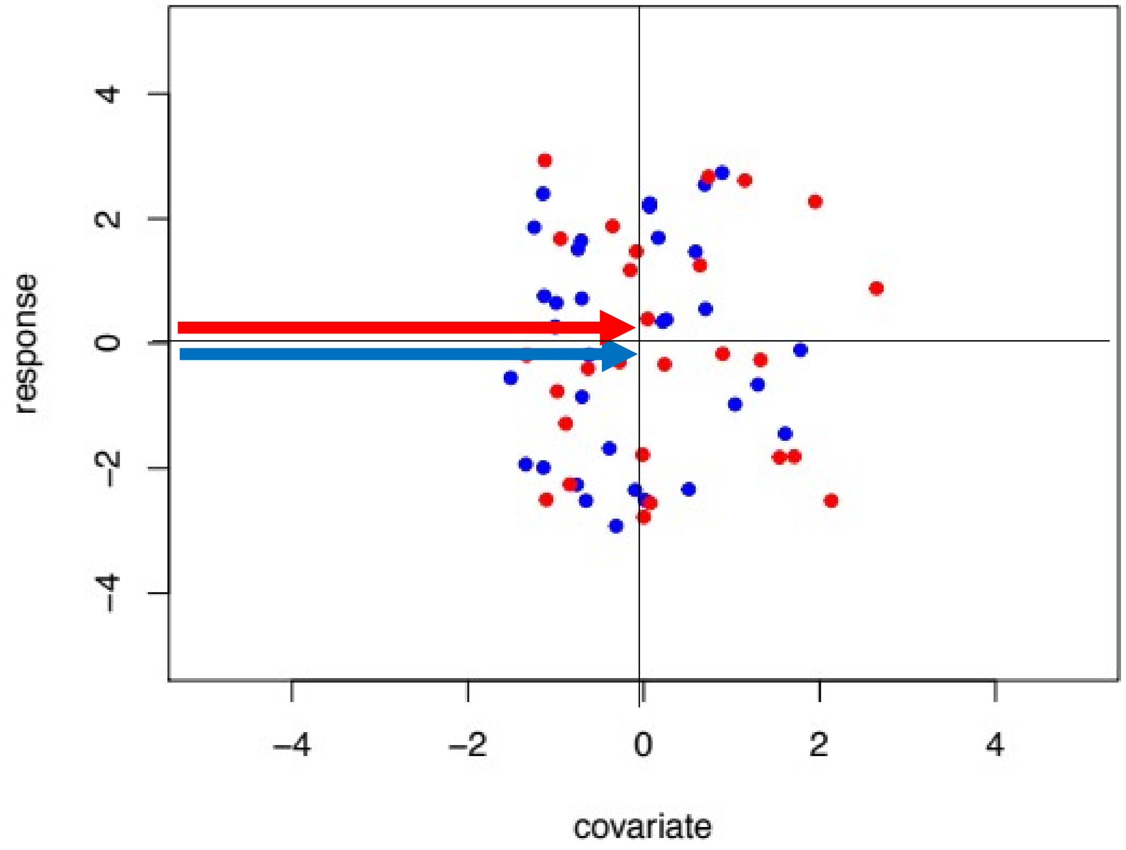
Female (x_{i2} = 0) $y_i = b_0 + b_1 x_{i0} + b_2 0 + ¿ \varepsilon_i b_0 + b_1 x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $y_i = b_0 + b_1 x_{i0} ¿ b_0 + b_1 b_2 x_{i0} * + b_2 1 + \varepsilon_i \varepsilon_i$

Female (x_{i2} = 0) $¿ 0 + 0 * x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $¿ 0 + 0 * x_{i0} + 0 + \varepsilon_i$

```
Call:
lm(formula = y ~ x + sx)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8702 -1.7023 -0.1178  1.5936  3.1370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02613    0.31286  -0.084    0.934
x            0.08222    0.23471   0.350    0.727
sx          -0.08598    0.47224  -0.182    0.856

Residual standard error: 1.784 on 57 degrees of freedom
Multiple R-squared:  0.00238,	Adjusted R-squared:  -0.03262
F-statistic: 0.06799 on 2 and 57 DF,  p-value: 0.9343
```



Sex:

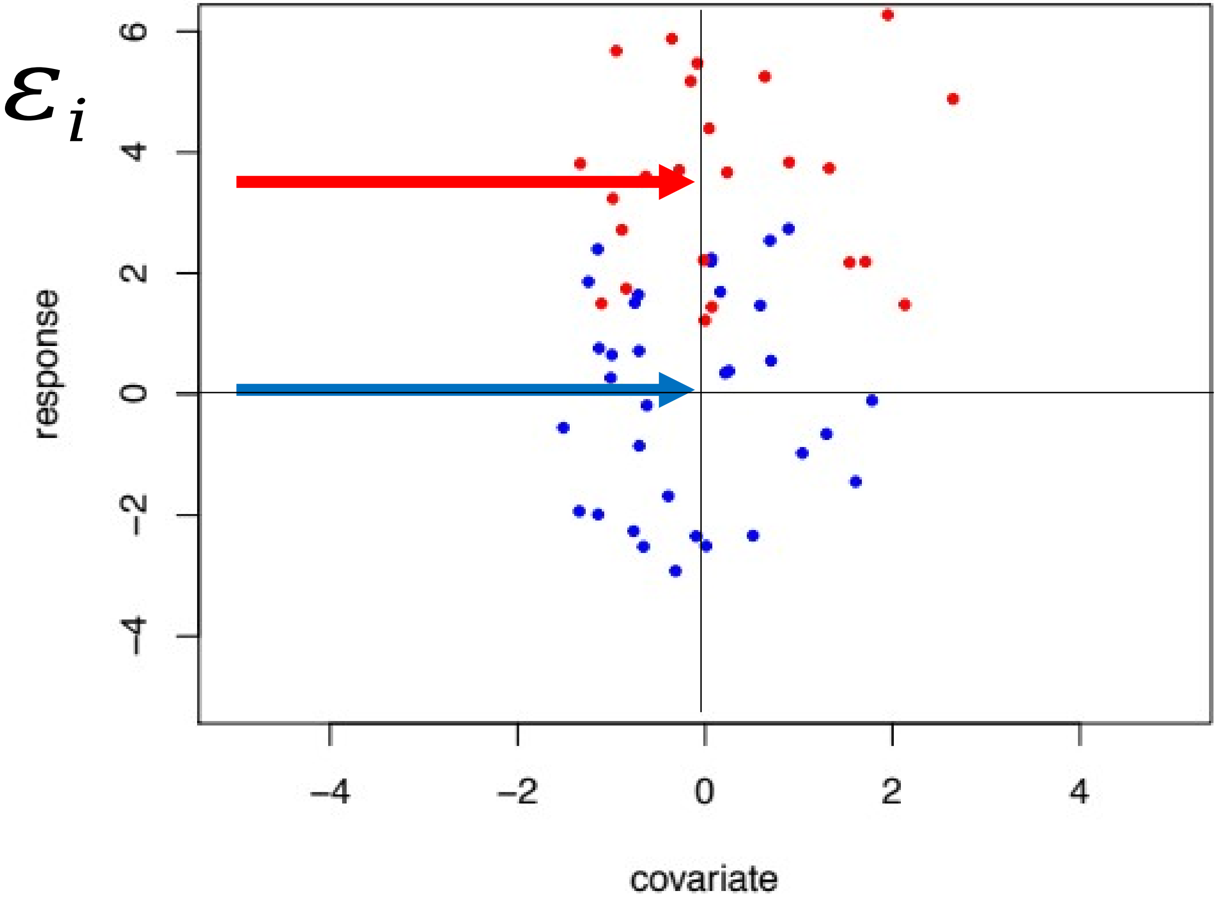
Female (x_{i2} = 0) $¿ b_0 + b_1 x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $¿ b_0 + b_1 x_{i0} + b_2 + \varepsilon_i$

Female (x_{i2} = 0) $¿ 0 + 0 * x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $¿ 0 + 0 * x_{i0} + 3.91 + \varepsilon_i$

```
Call:
lm(formula = y ~ x + sx)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8702 -1.7023 -0.1178  1.5936  3.1370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02613    0.31286  -0.084    0.934    
x            0.08222    0.23471   0.350    0.727    
sx           3.91402    0.47224   8.288 2.29e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.784 on 57 degrees of freedom
Multiple R-squared:  0.5608,	Adjusted R-squared:  0.5454 
F-statistic:  36.4 on 2 and 57 DF,  p-value: 6.524e-11
```



Sex:

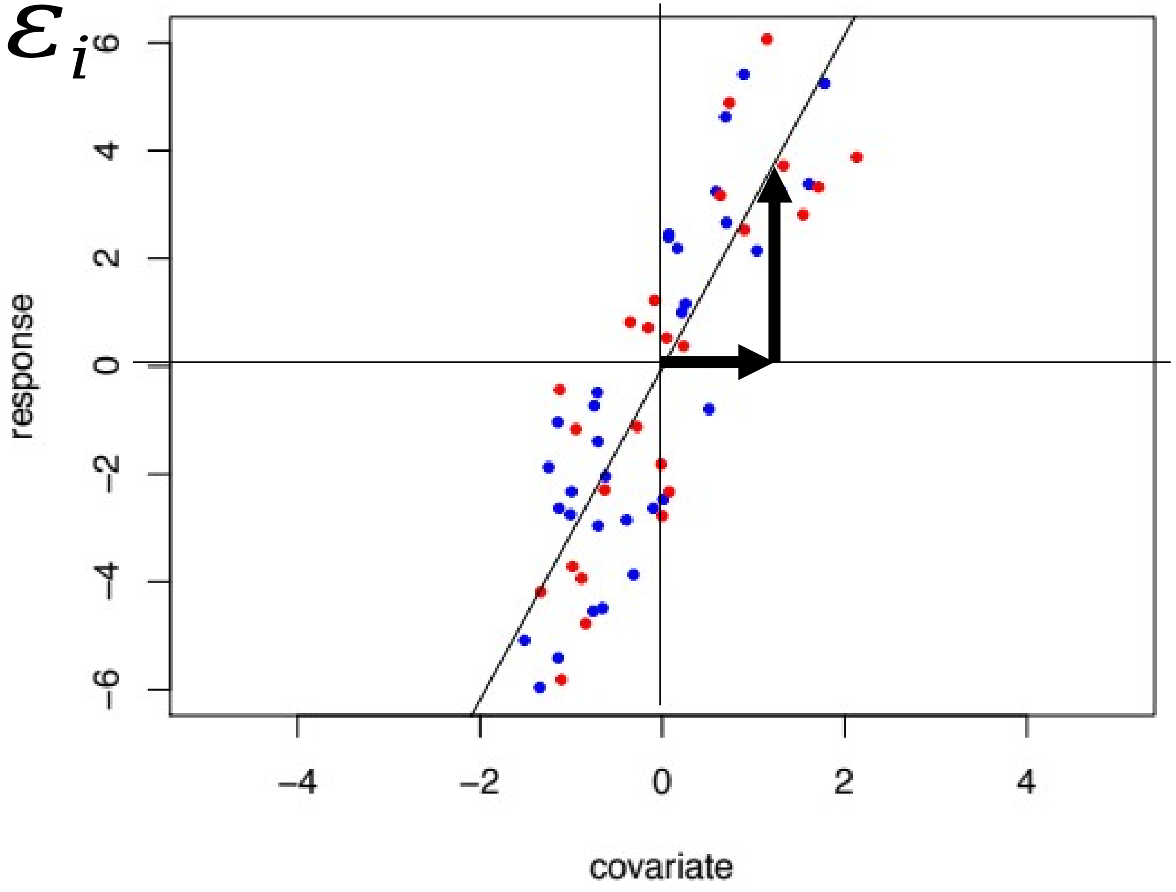
Female (x_{i2} = 0) $¿\, b_0 + b_1 x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $¿\, b_0 + b_1 x_{i0} + b_2 + \varepsilon_i$

Female (x_{i2} = 0) $¿\, 0 + 3.08 * x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $¿\, 0 + 3.08 * x_{i0} + 0 + \varepsilon_i$

```
Call:
lm(formula = y ~ x + sx)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8702 -1.7023 -0.1178  1.5936  3.1370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02613    0.31286  -0.084    0.934
x            3.08222    0.23471  13.132   <2e-16 ***
sx          -0.08598    0.47224  -0.182    0.856
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.784 on 57 degrees of freedom
Multiple R-squared:  0.7579,    Adjusted R-squared:  0.7494
F-statistic: 89.24 on 2 and 57 DF,  p-value: < 2.2e-16
```



Sex:

Female (x_{i2} = 0) $¿ b_0 + b_1 x_{i0} + \varepsilon_i$

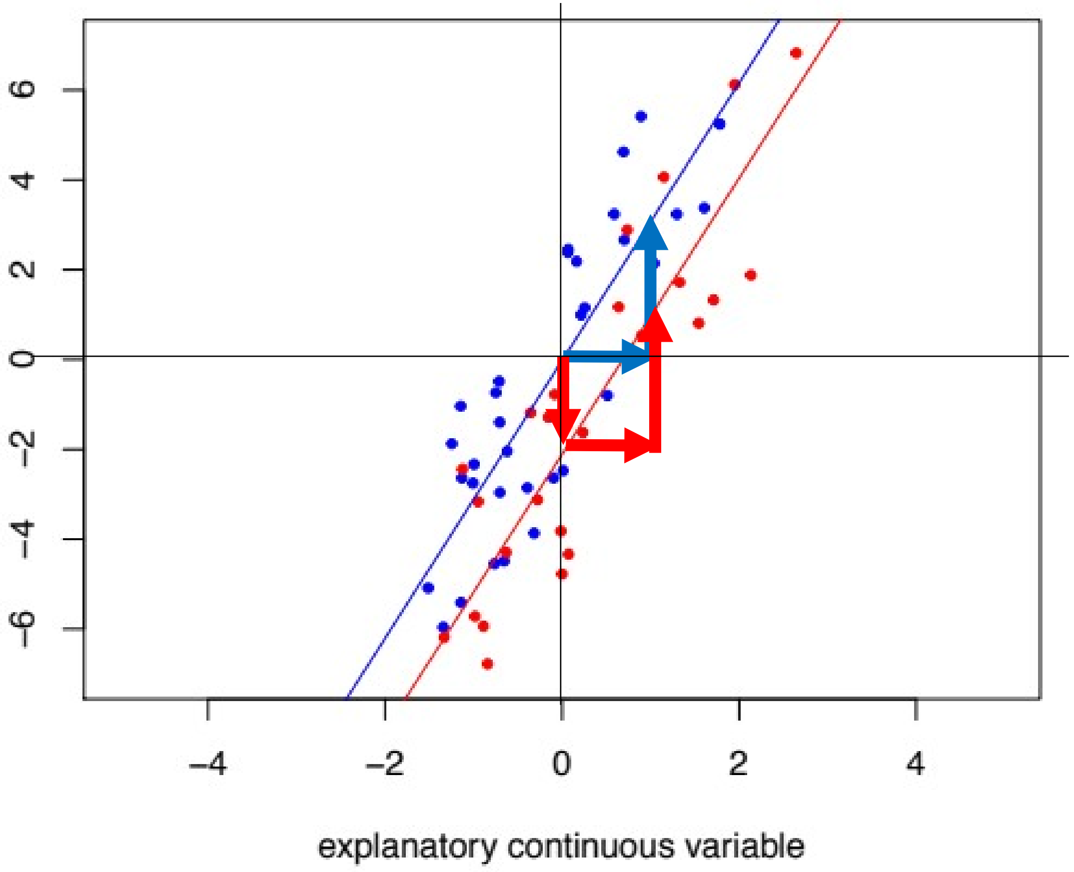
Male (x_{i2} = 1) $¿ b_0 + b_1 x_{i0} + b_2 + \varepsilon_i$

Female (x_{i2} = 0) $¿ 0 + 3.08 x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $¿ 0 + 3.08 x_{i0} - 2.09 + \varepsilon_i$

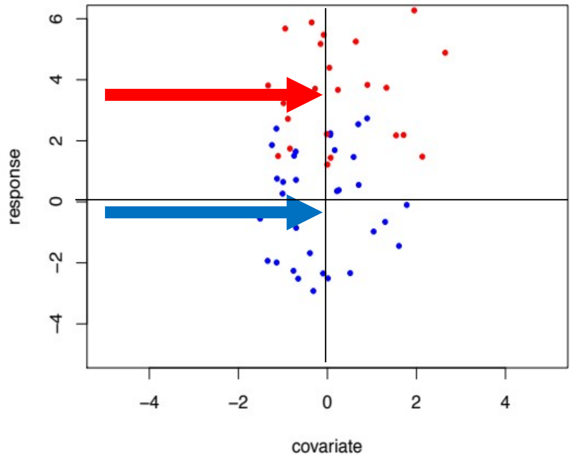
$¿ -2.09 + 3.08 x_{i0} + \varepsilon_i$

```
Call:
lm(formula = y ~ x + sx)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8702 -1.7023 -0.1178  1.5936  3.1370 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02613    0.31286  -0.084    0.934    
x            3.08222    0.23471  13.132  < 2e-16 ***
sx          -2.08598    0.47224  -4.417 4.53e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.784 on 57 degrees of freedom
Multiple R-squared:  0.7553,	Adjusted R-squared:  0.7467 
F-statistic: 87.96 on 2 and 57 DF,  p-value: < 2.2e-16
```

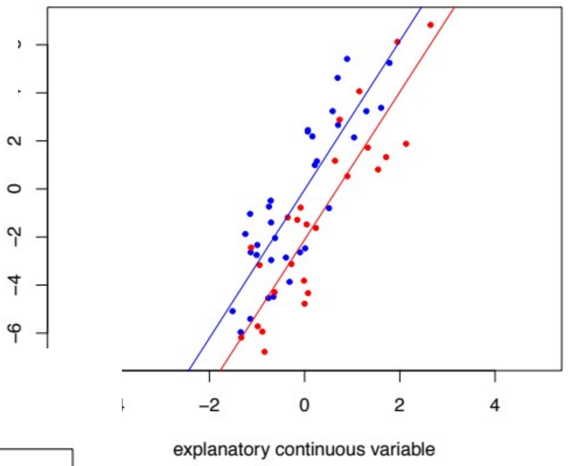


Ok

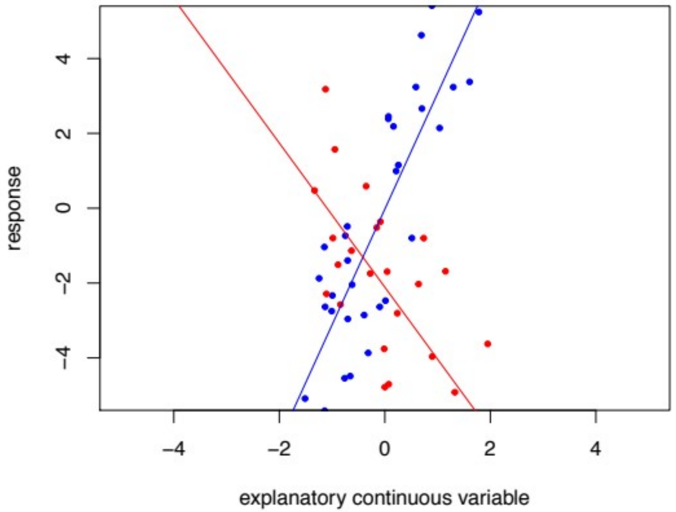
- We can estimate two different intercepts



- We can estimate one slope and two intercepts



- How can we estimate a separate slope for each sex?



$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} +$$

Interactions of terms:

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i0} x_{i1} + \varepsilon_i$$

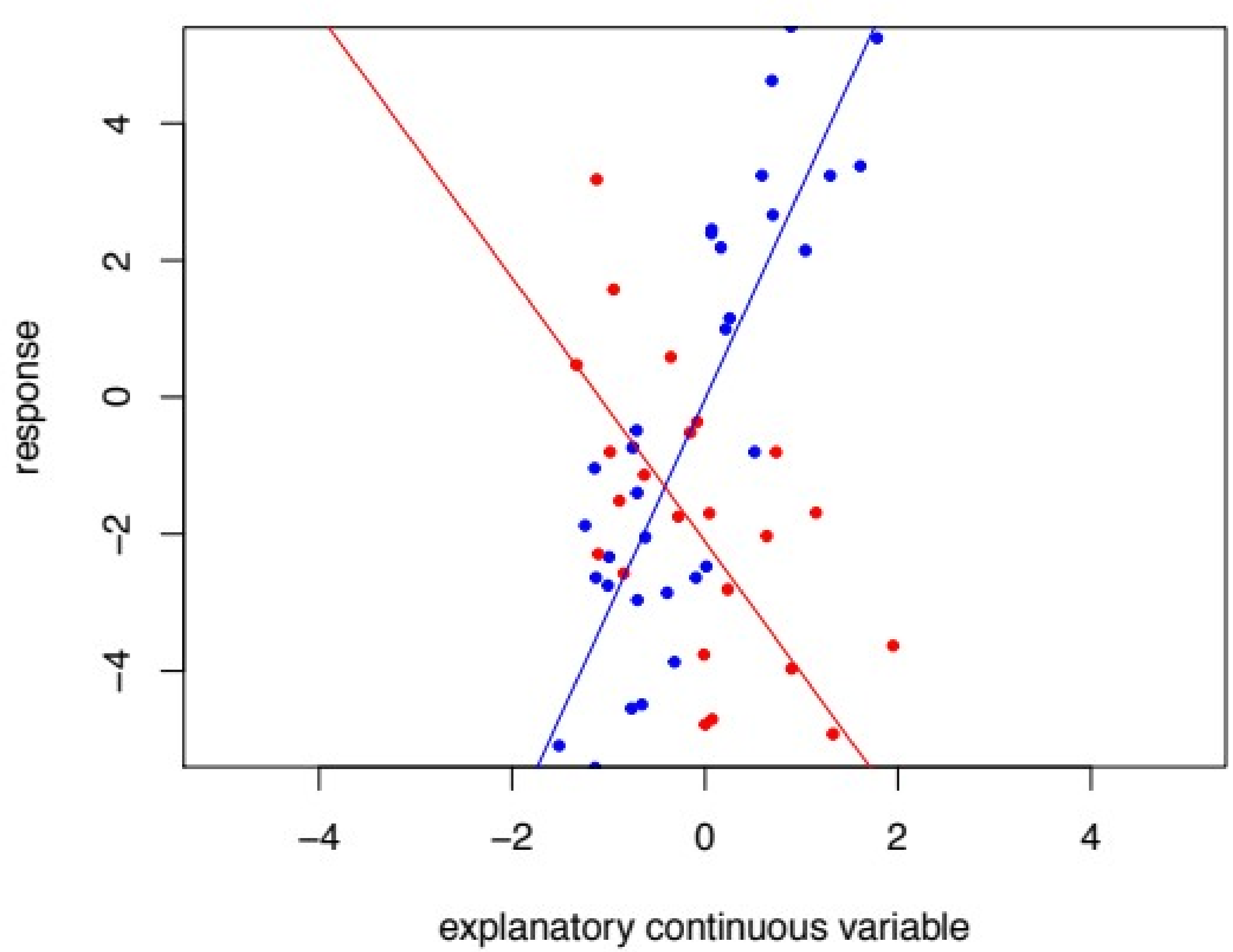
- interaction between sex and tarsus
- one more parameter estimate
- one more degree of freedom
- but not more variables

Sex:

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i0} x_{i1} + \varepsilon$$

Female (x_{i2} = 0) $b_0 + b_1 x_{i0} + b_2 0 + b_3 x_{i0} 0 + \varepsilon_i \text{¿} b_0 + b_1 x_{i0} + \varepsilon_i$

Male (x_{i2} = 1) $b_0 + b_1 x_{i0} + b_2 + b_3 x_{i0} + \varepsilon_i \text{ ¿} b_0 + b_2 + (b_1 + b_3) x_{i0} + \varepsilon$



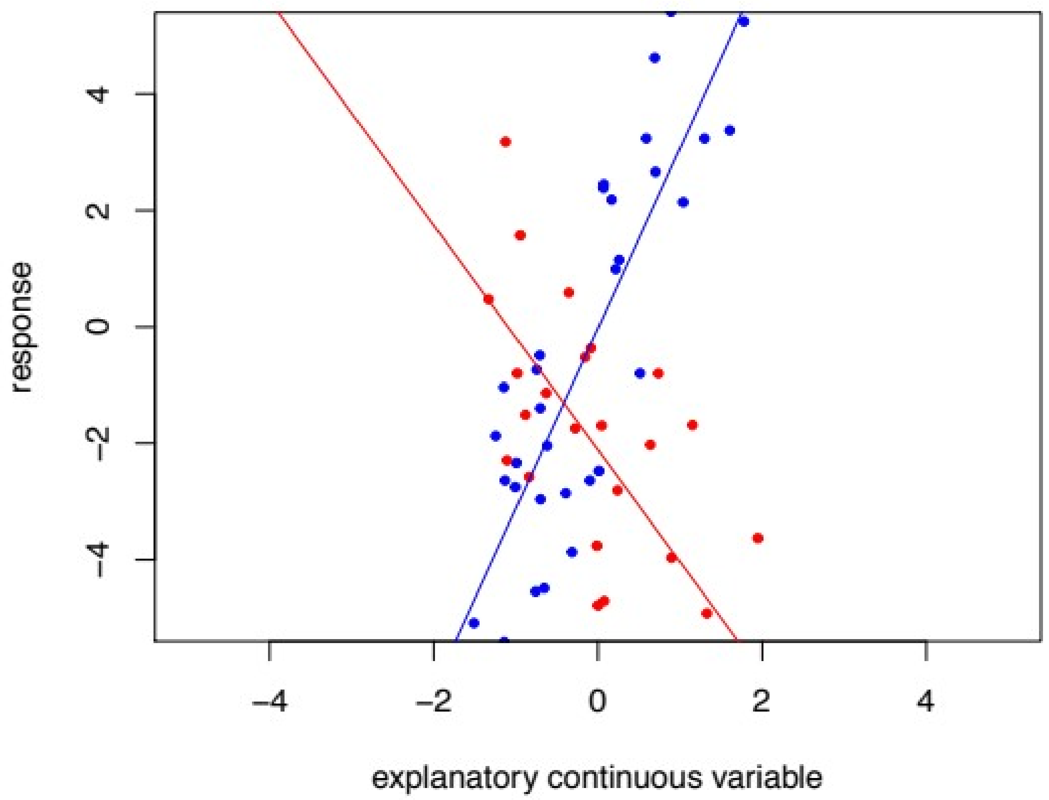
$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3$$

Female ($x_{i2} = 0$) ¿ $b_0 + b_1 x_{i0}$ ¿ $0 + 3.09 x_{i0}$

Male ($x_{i2} = 1$) ¿ $b_0 + b_2 + (b_1 + b_3) x_{i0}$ ¿ $(0 - 2.09) + (3.09 - 5.02) x_{i0}$

¿ $-2.09 + 1.93 x_{i0}$

```
Call:
lm(formula = y ~ x * sx)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8687 -1.7008 -0.1129  1.5931  3.1264 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02453    0.31856  -0.077    0.939    
x            3.09216    0.35685   8.665 6.30e-12 ***
sx          -2.08574    0.47648  -4.377 5.30e-05 ***
x:sx        -5.01776    0.47700 -10.519 7.07e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.8 on 56 degrees of freedom
Multiple R-squared:  0.7008,    Adjusted R-squared:  0.6848 
F-statistic: 43.73 on 3 and 56 DF,  p-value: 1.079e-14
```



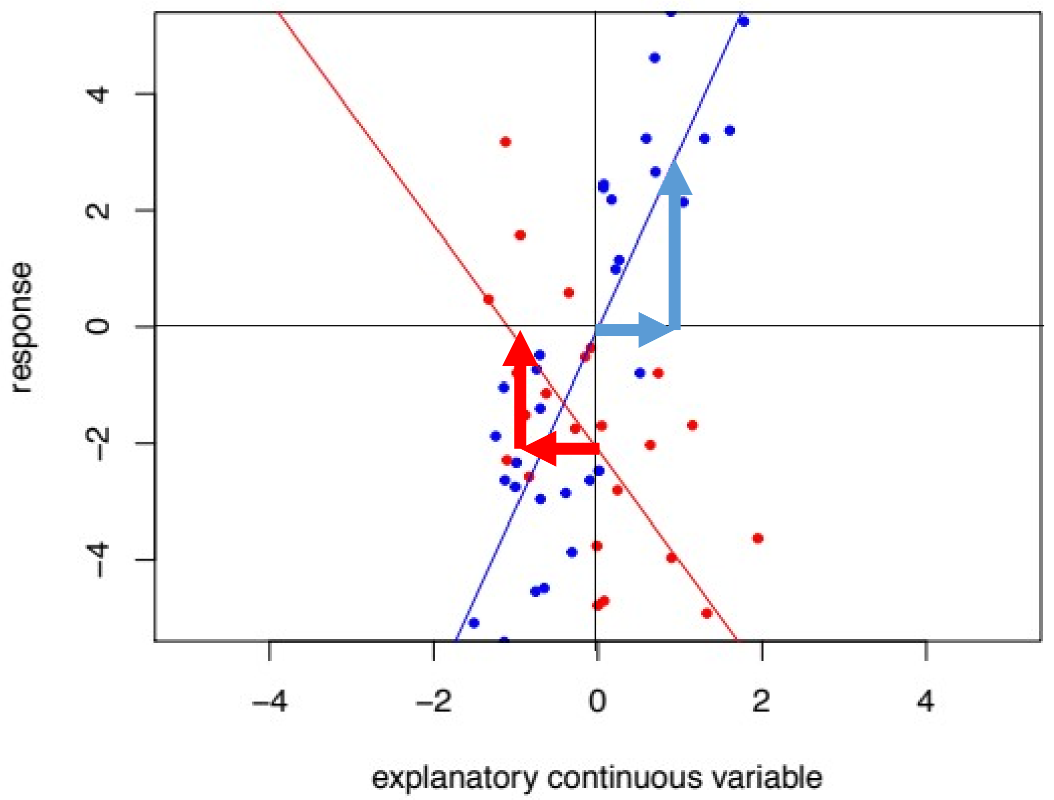
$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3$$

Female (x_{i2} = 0) $¿\, b_0 + b_1 x_{i0}$ $¿\, \mathbf{0+3.09\,x_{i0}}$

Male (x_{i2} = 1) $¿\, b_0 + b_2 + (b_1 + b_3)\, x_{i0}$ $¿\, (0 - 2.09) + (3.09 - 5.02)\, x_{i0}$

$¿\, \mathbf{-2.09 + 1.93\,x_{i0}}$

```
Call:
lm(formula = y ~ x * sx)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8687 -1.7008 -0.1129  1.5931  3.1264

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02453    0.31856  -0.077    0.939
x            3.09216    0.35685   8.665 6.30e-12 ***
sx          -2.08574    0.47648  -4.377 5.30e-05 ***
x:sx        -5.01776    0.47700 -10.519 7.07e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.8 on 56 degrees of freedom
Multiple R-squared:  0.7008,    Adjusted R-squared:  0.6848
F-statistic: 43.73 on 3 and 56 DF,  p-value: 1.079e-14
```

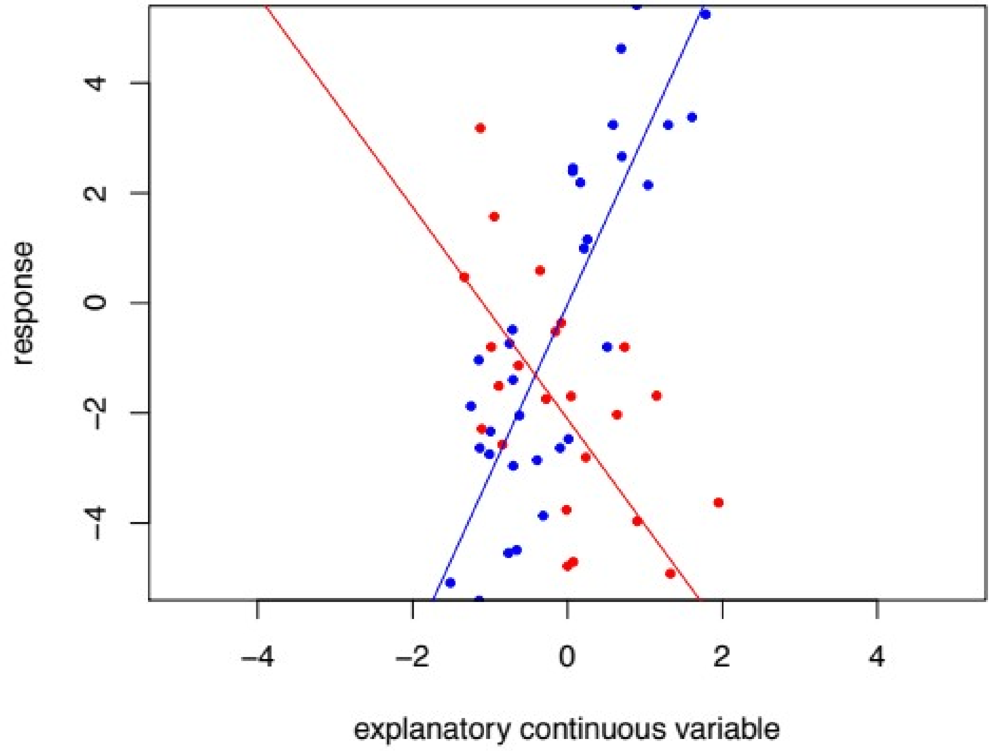


$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i0} x_{i1} + \varepsilon$$

Let's try this

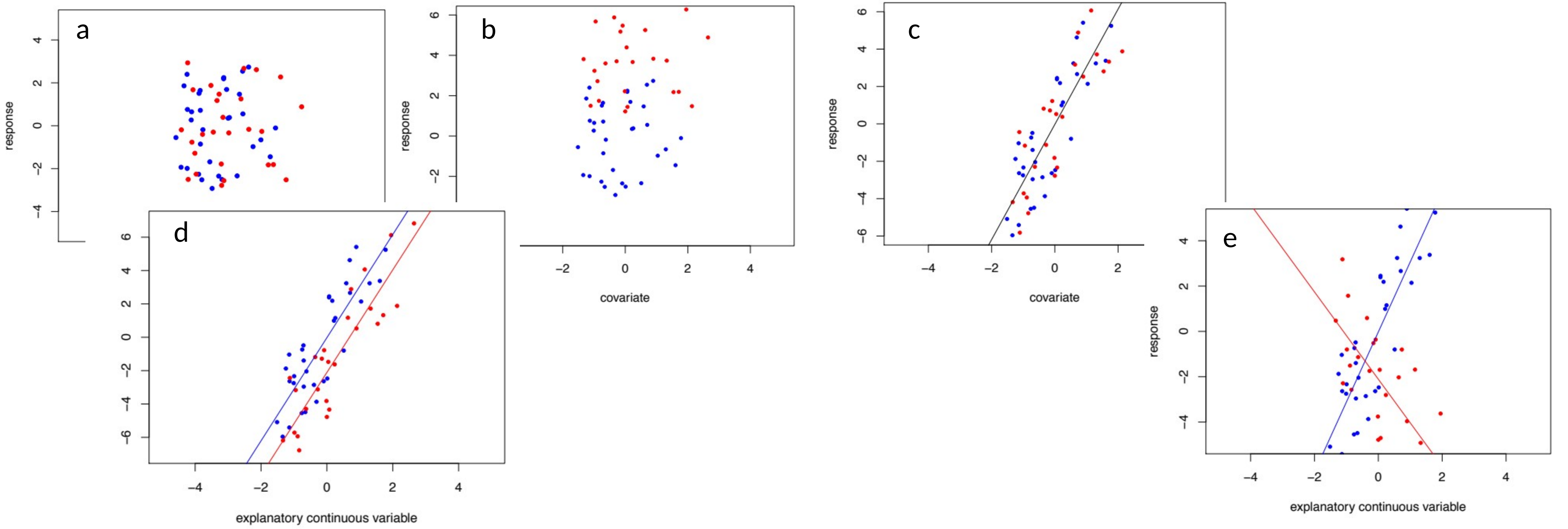
- Intercept: b_0 reference group, b_2: *difference* of group 1 to reference
- Slope: b_1 slope of reference group, b3 *difference* of slope of group 1 to reference
- Do not interpret estimates by themselves!

```
Call:
lm(formula = y ~ x * sx)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8687 -1.7008 -0.1129  1.5931  3.1264 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.02453    0.31856  -0.077    0.939    
x            3.09216    0.35685   8.665 6.30e-12 ***
sx          -2.08574    0.47648  -4.377 5.30e-05 ***
x:sx        -5.01776    0.47700 -10.519 7.07e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.8 on 56 degrees of freedom
Multiple R-squared:  0.7008,	Adjusted R-squared:  0.6848 
F-statistic: 43.73 on 3 and 56 DF,  p-value: 1.079e-14
```



$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i0} x_{i1} + \varepsilon$$

plot	a	b	c	d	e
Intercept					
b_1 (sex)	0	+	0	+	+
b_2 (tarsus)	0	0	+	+	+
b_3 (tarsus x sex)	0	0	0	0	-



What else?

- Interactions between continuous variables and categorical predictors with more than 2 levels
- Multiple continuous predictors
- Interactions between categorical predictors
- → hand outs!

- Avoid (at all costs):
- Interactions between 2 continuous predictors
- 3- or more-way interactions

Take home: categorical predictors:

- Interpreting categorical x continuous predictor interactions
- When interaction terms are present – never interpret the estimates in isolation
- You will need to do some math's to interpret your results