

Statistics with Spa OWS

Lecture 15

Julia Schroeder

Julia.schroeder@imperial.ac.uk

Outline

- Linear models – going big

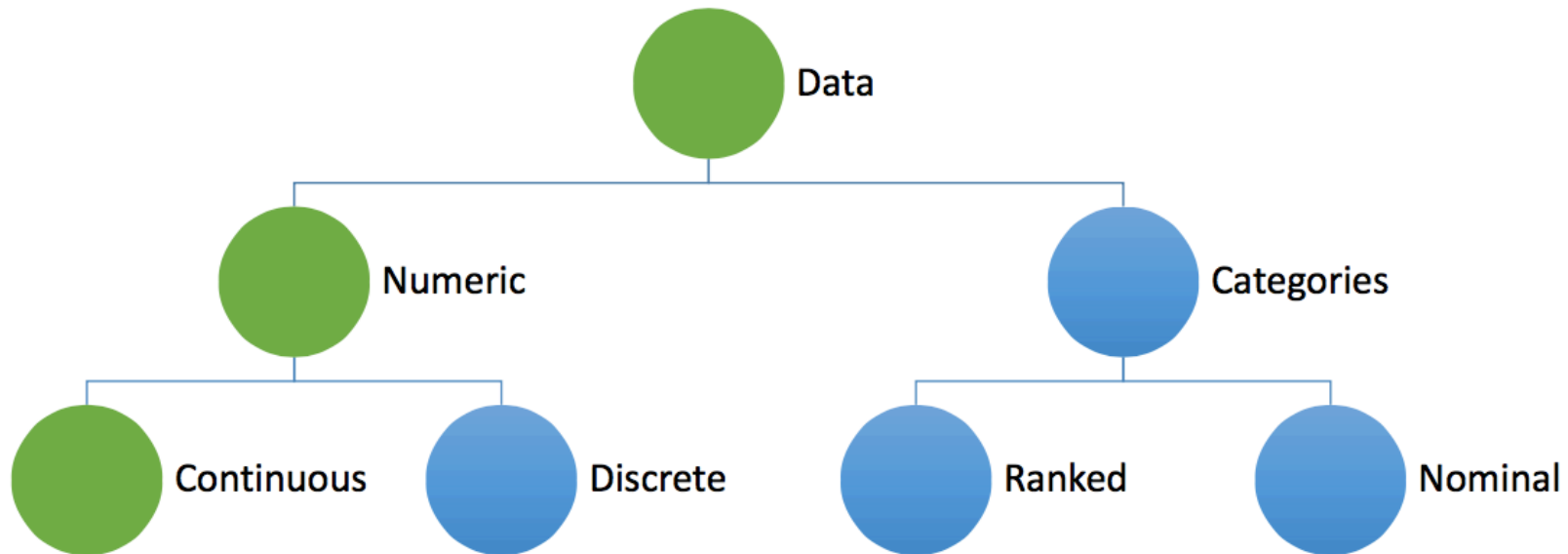
Multiple linear models

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

Multiple linear models

`lm(response~explanatory)`

Data types



Multiple linear models

`lm(response~explanatory)`

Response y:

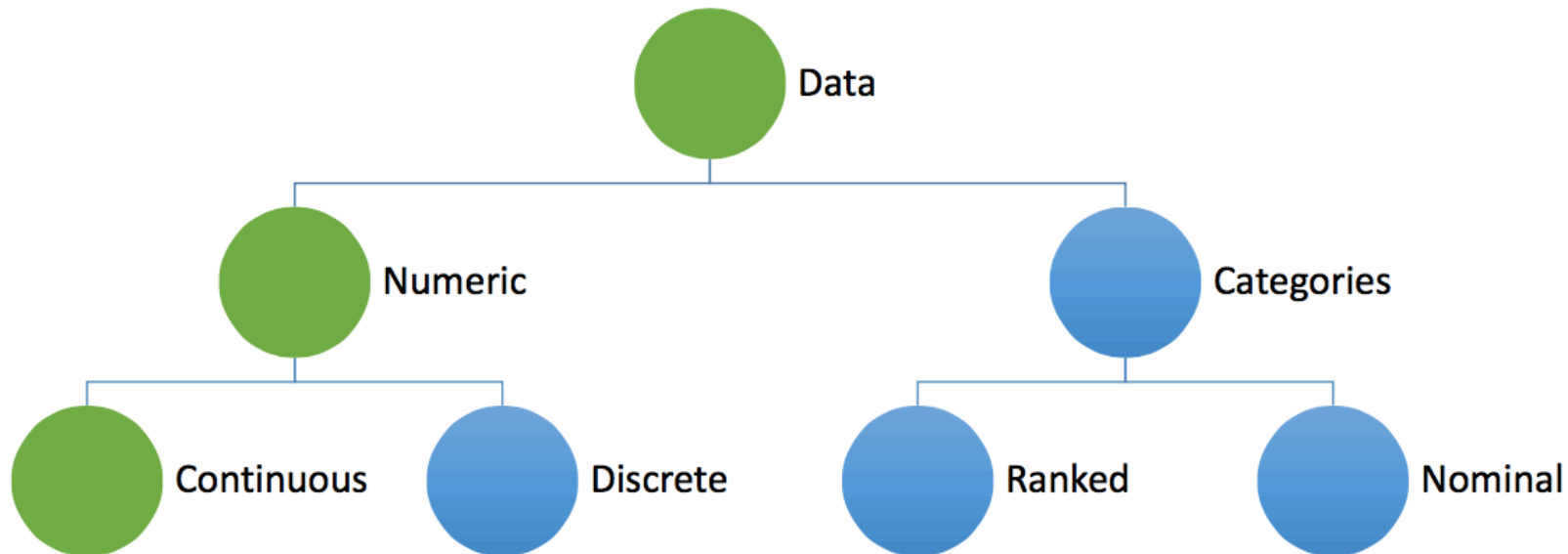
Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

Data types



Multiple linear models

`lm(response~explanatory)`

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

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

Multiple linear models

`lm(response~explanatory)`

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

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

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i2} + \varepsilon_i$$

Multiple linear models

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

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

We can have more than one explanatory variable!

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i2} + \varepsilon_i$$

Multiple linear models

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

Response y:

Continuous

Explanatory x:

Continuous (tarsus, wing, mass)

Categorical (Sex, Year, Observer, BirdID)

We can have more than one explanatory variable!

We can even mix continuous and factorial explanatory variables!

$$y_i = b_0 + b_1 x_{i0} + b_2 x_{i1} + b_3 x_{i2} + \varepsilon_i$$

Let's try this

- b_0 = intercept

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

- 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

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

- 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

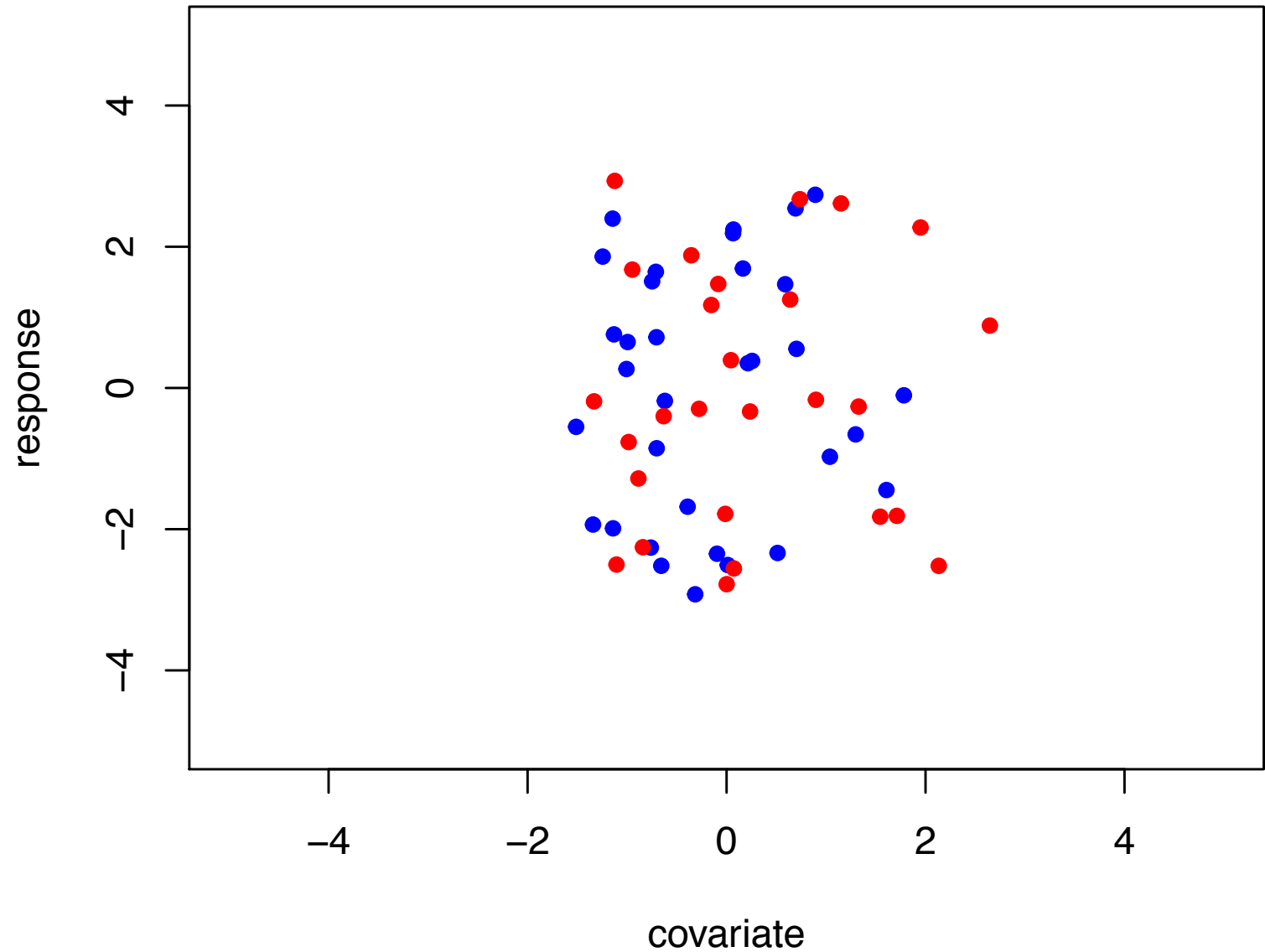
- b_0 = intercept

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

- 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.

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

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

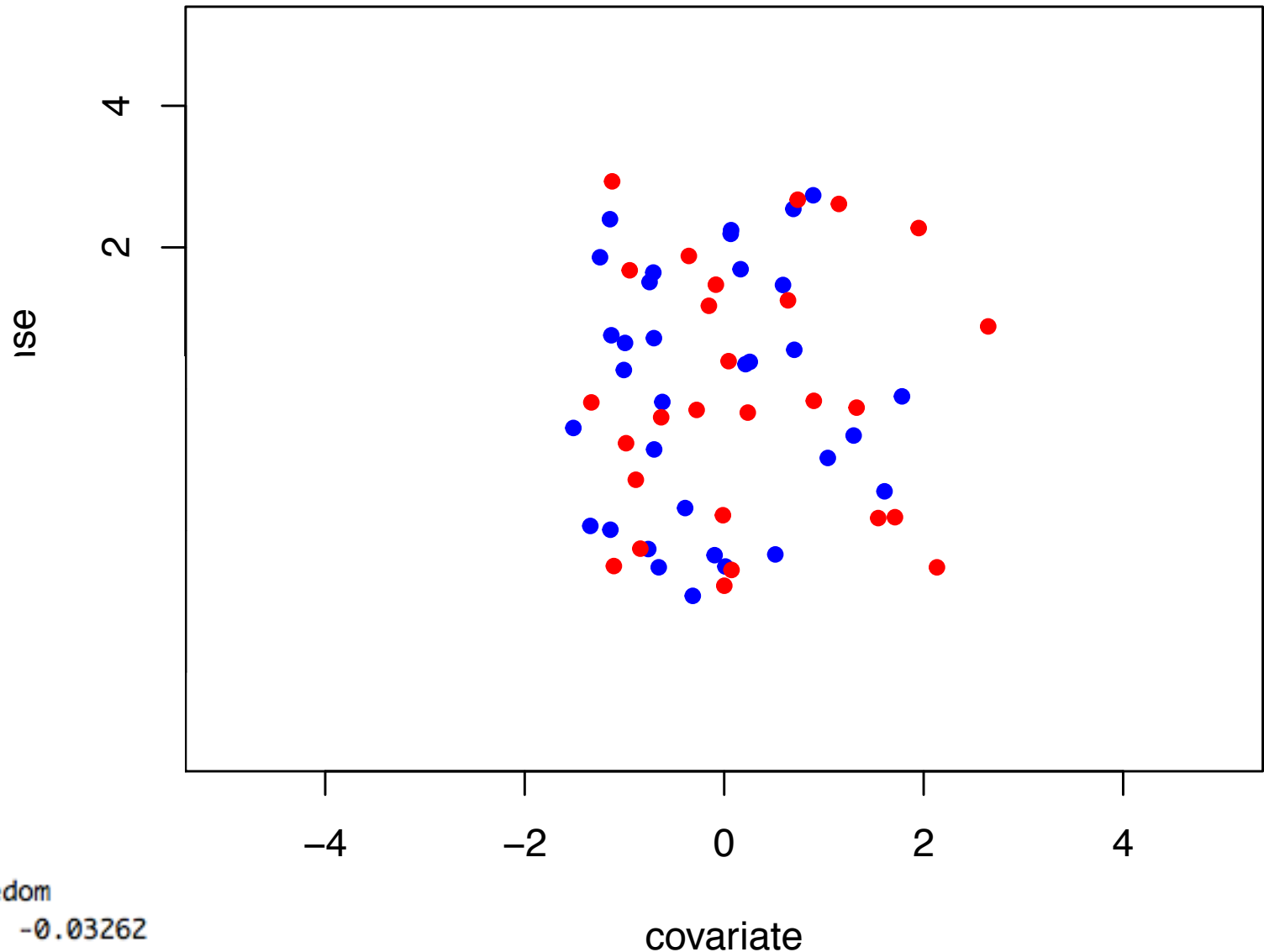
Residuals:

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

Coefficients:

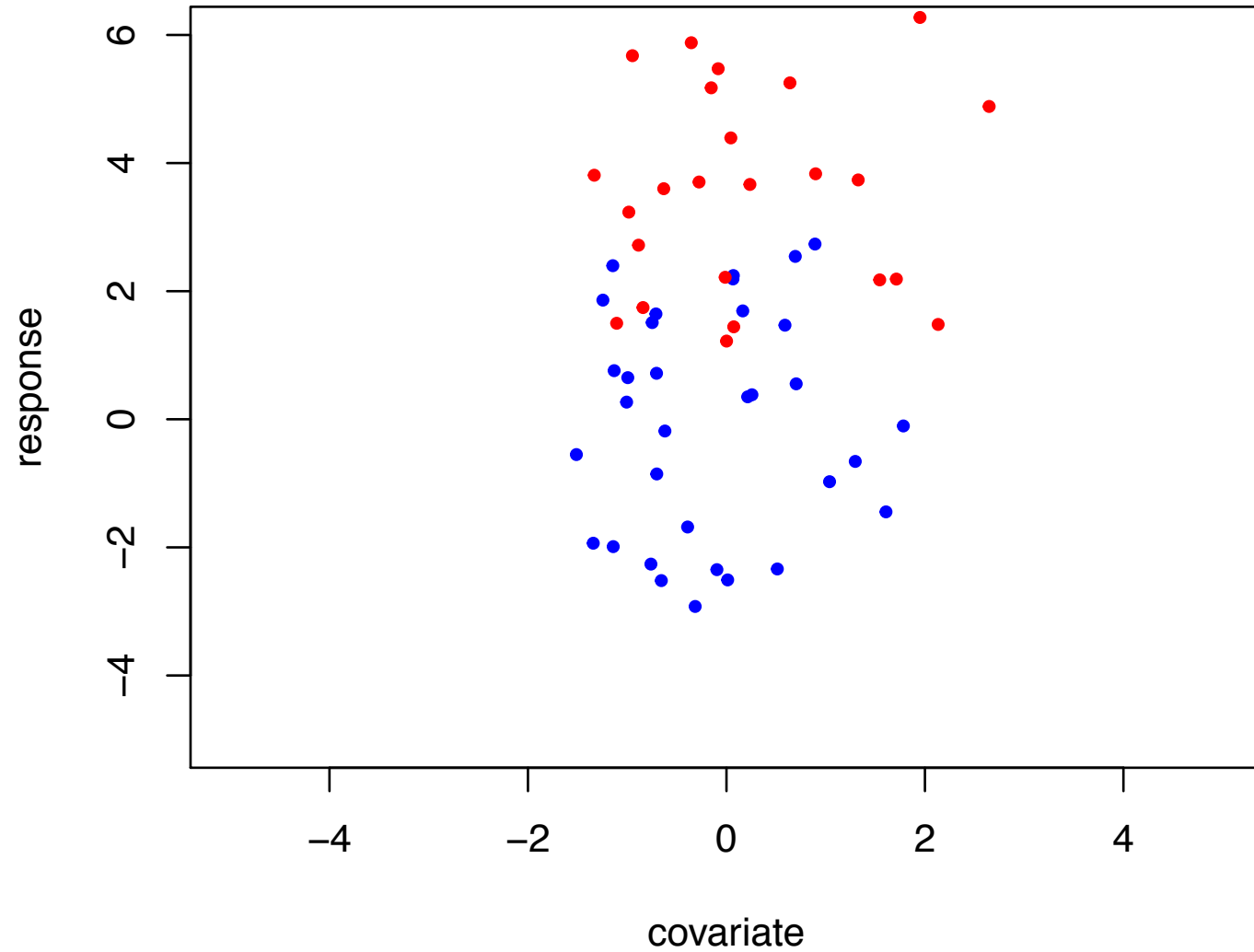
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.02613	0.31286	-0.084	0.934
x	0.08222	0.23471	0.350	0.727
sx	-0.08598	0.47224	-0.182	0.856

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



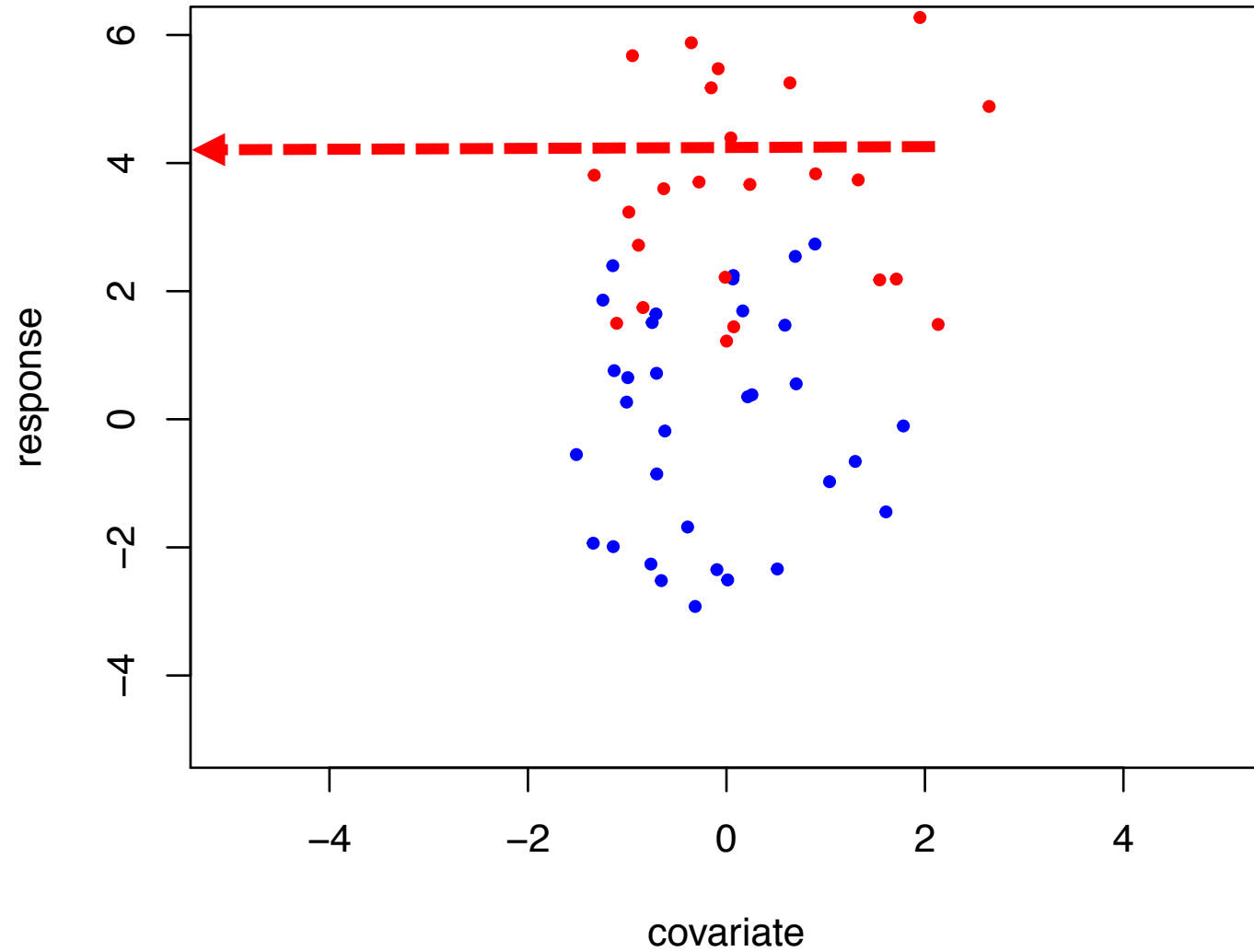
$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



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

Let's try this



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

Residuals:

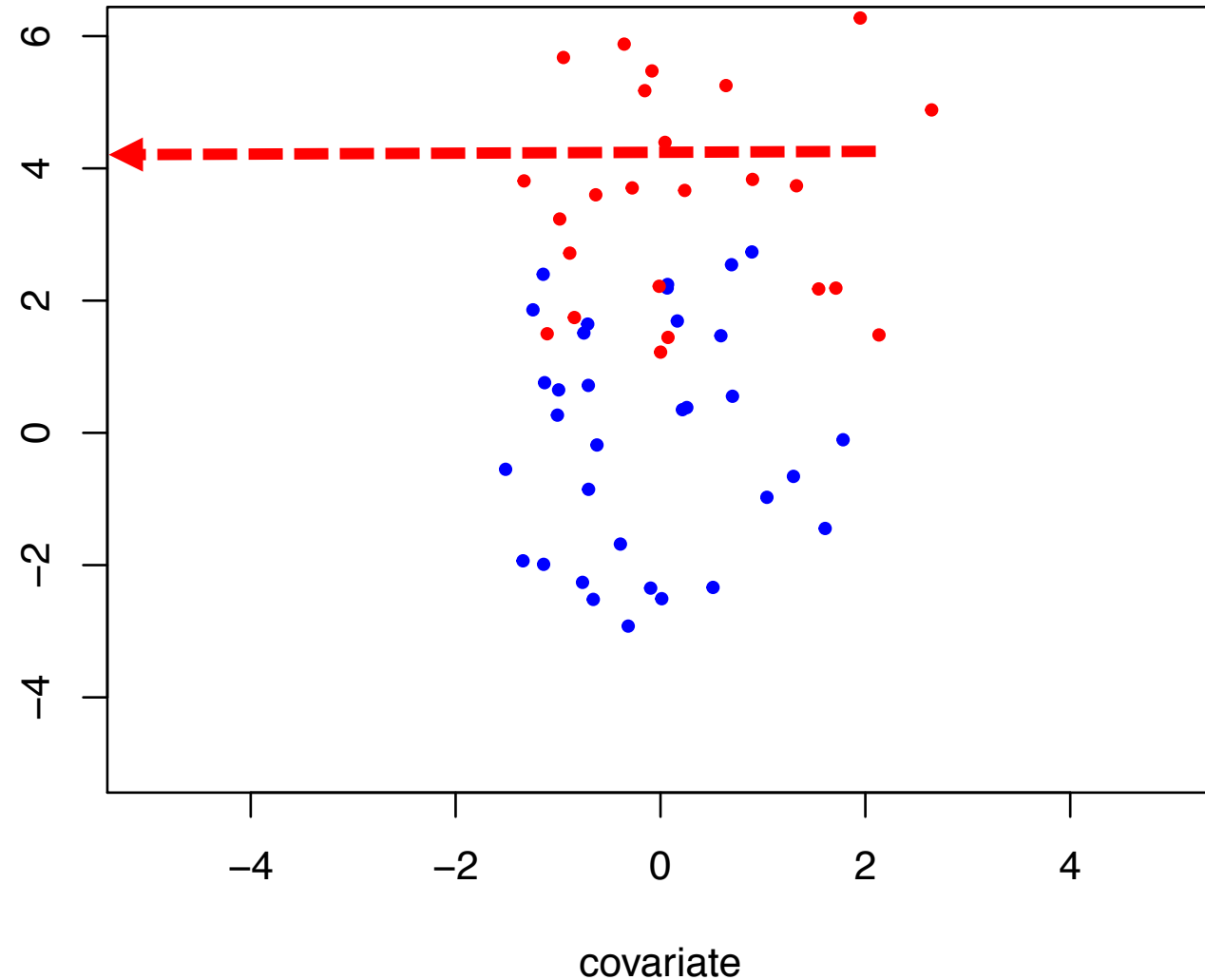
Min	1Q	Median	3Q	Max
-2.8702	-1.7023	-0.1178	1.5936	3.1370

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 ***

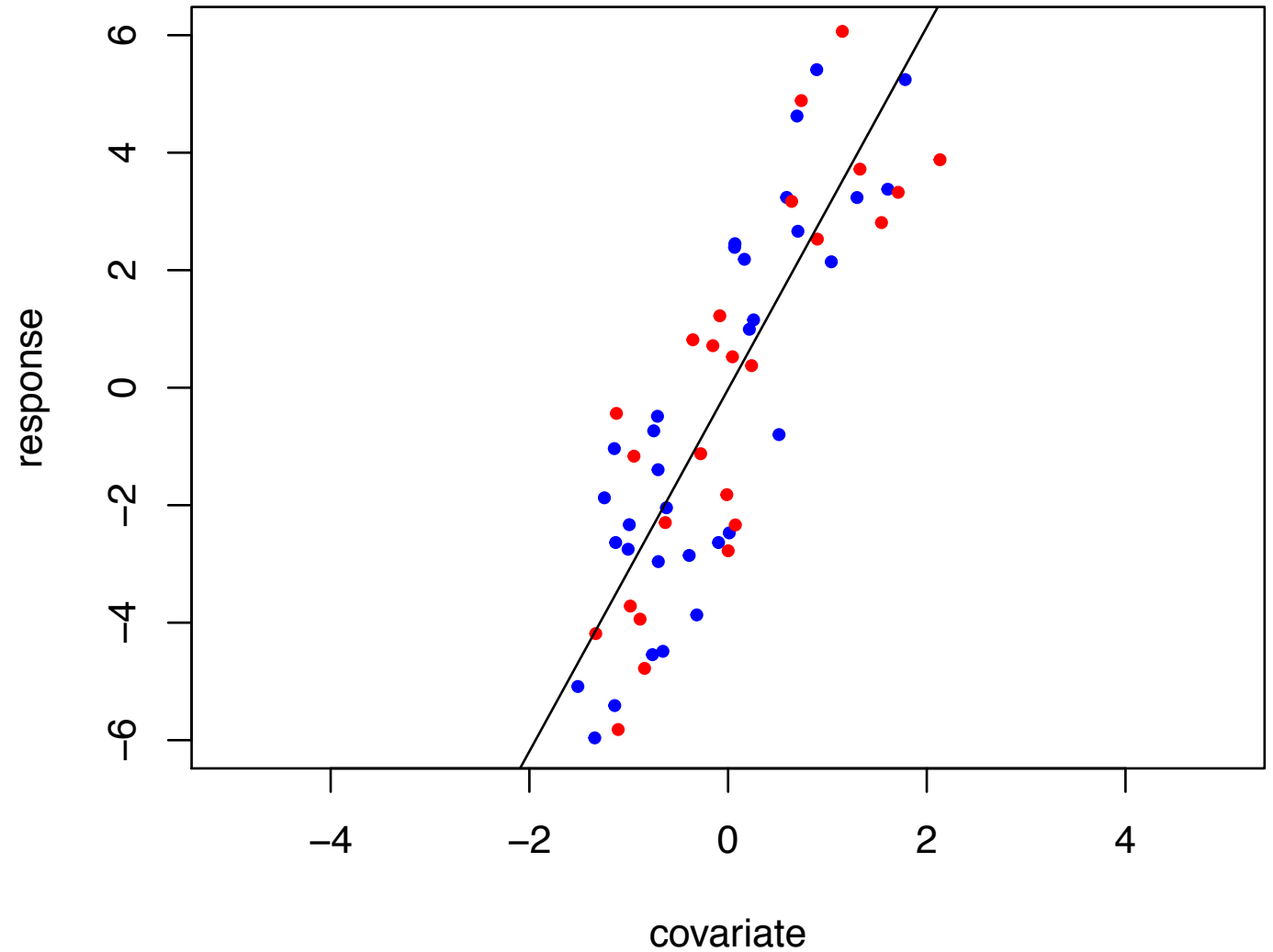
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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



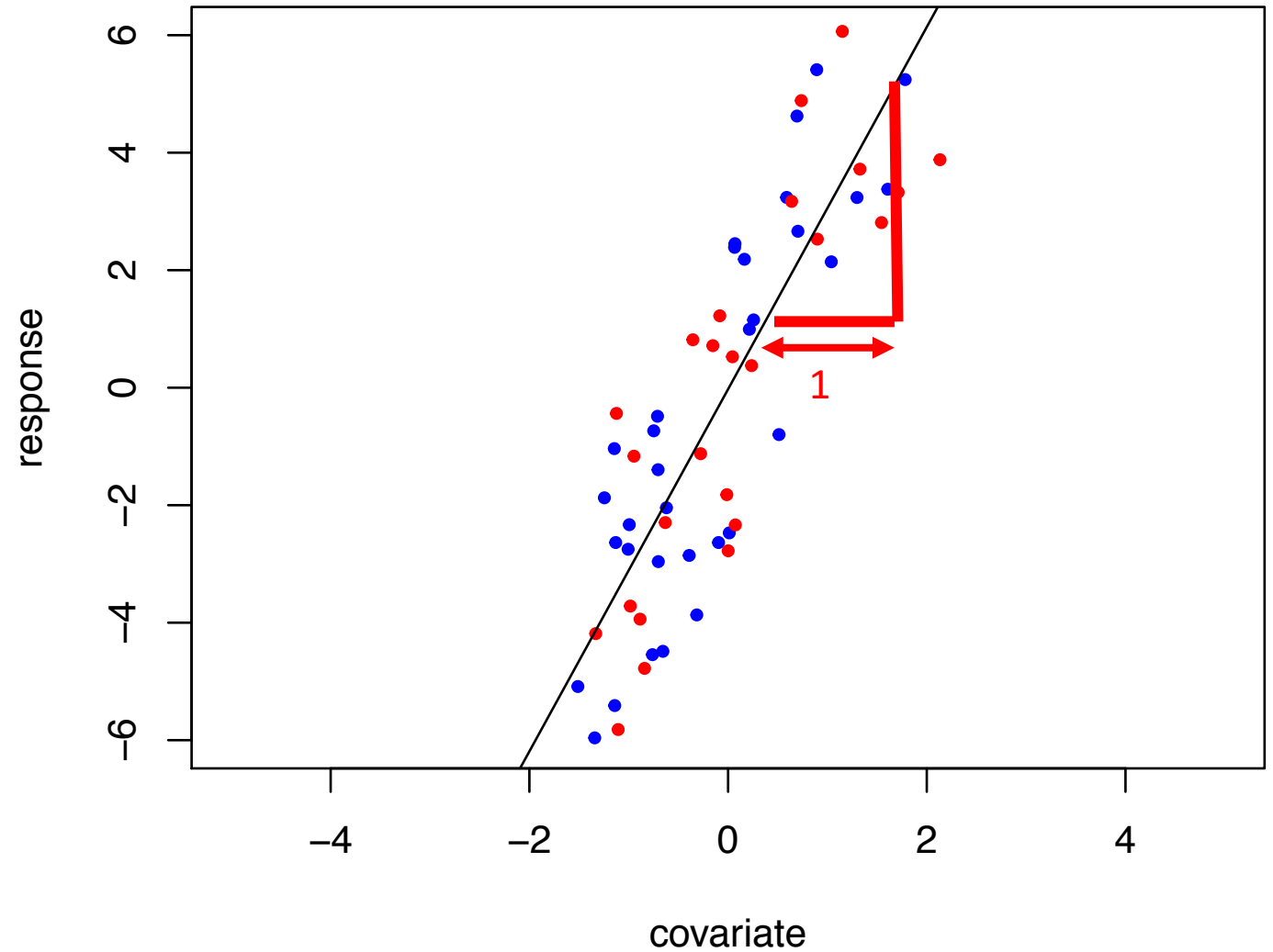
$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



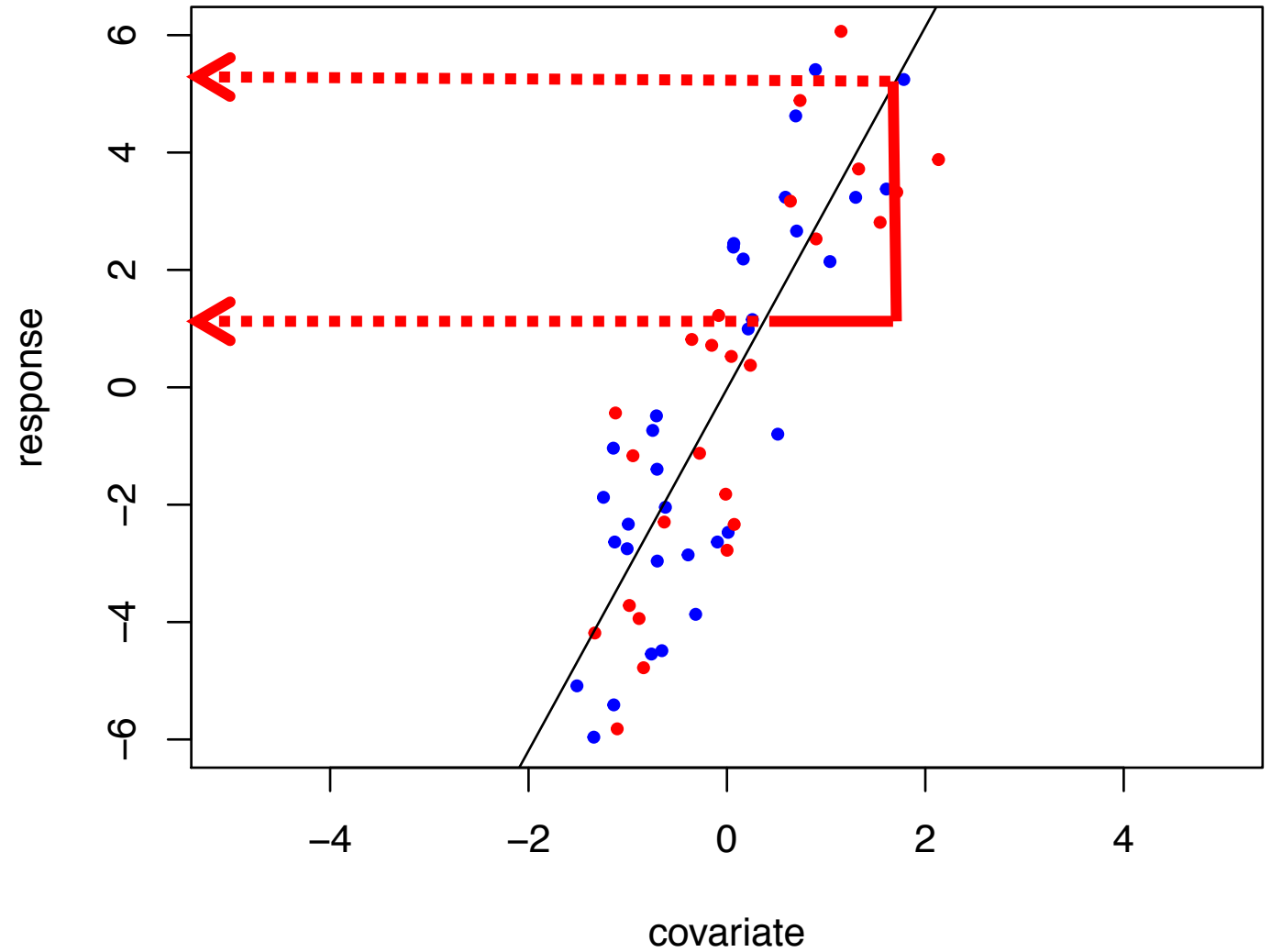
$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



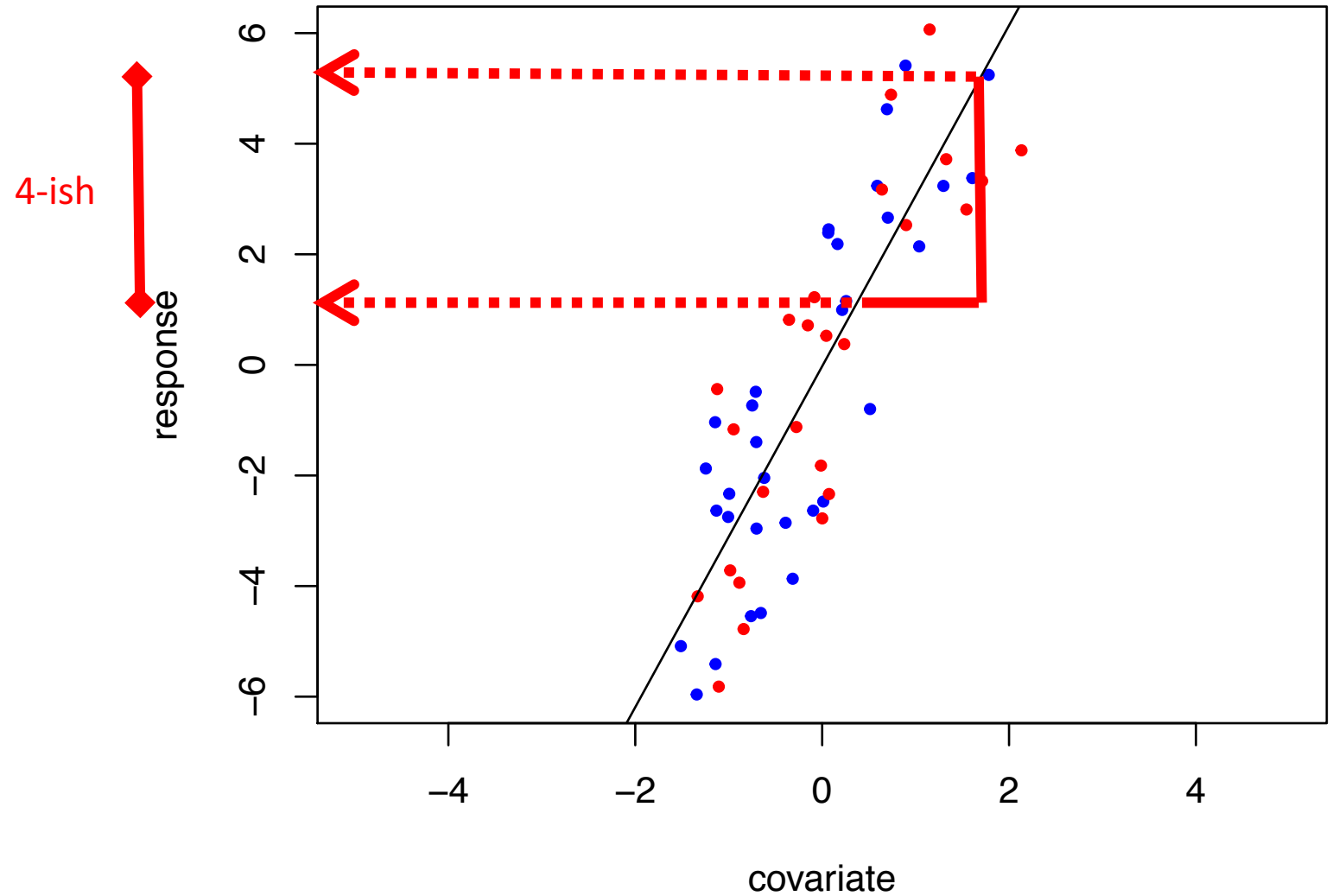
$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this



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

Let's try this



4-ish

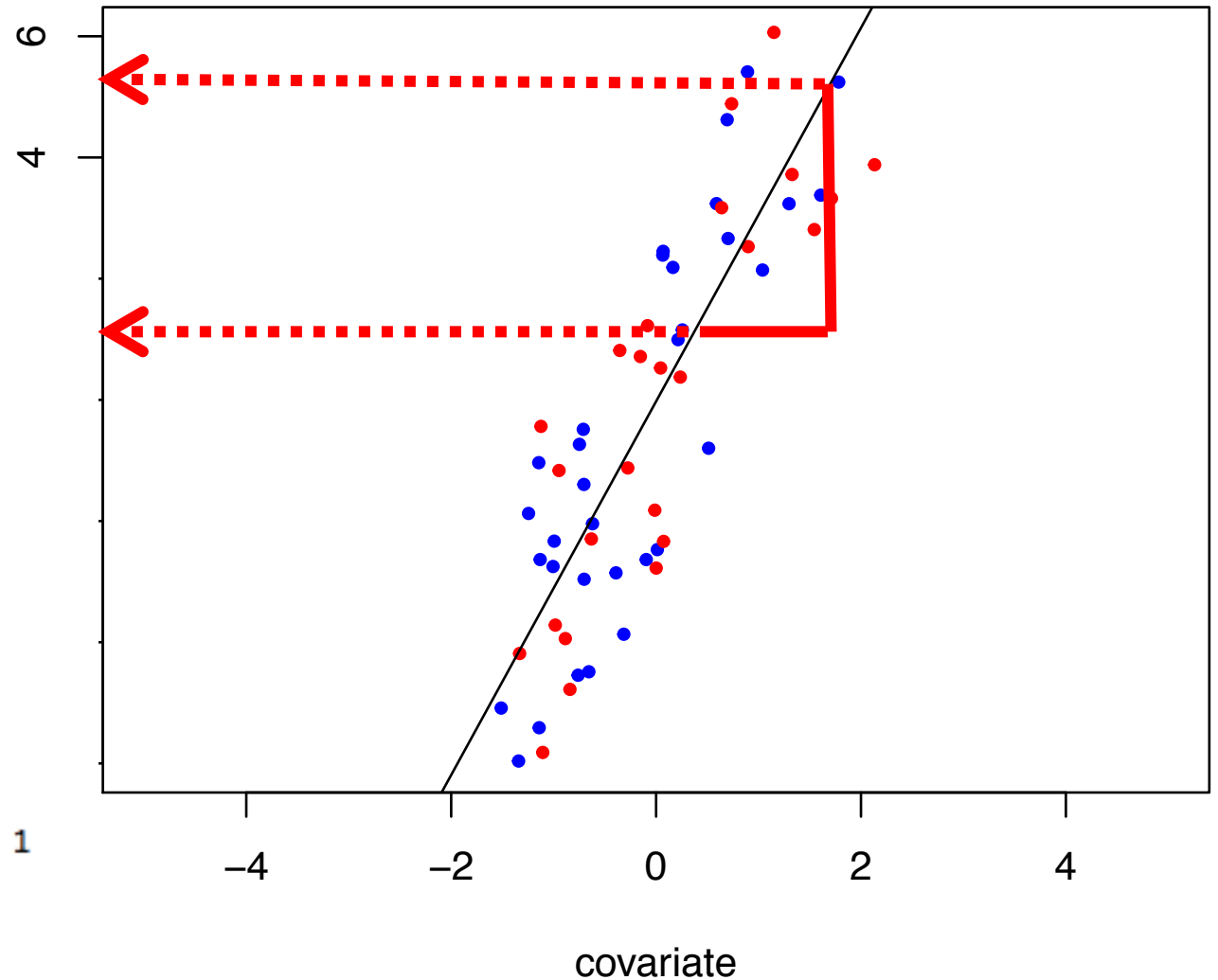


```
Call:
lm(formula = y ~ x + sx)

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

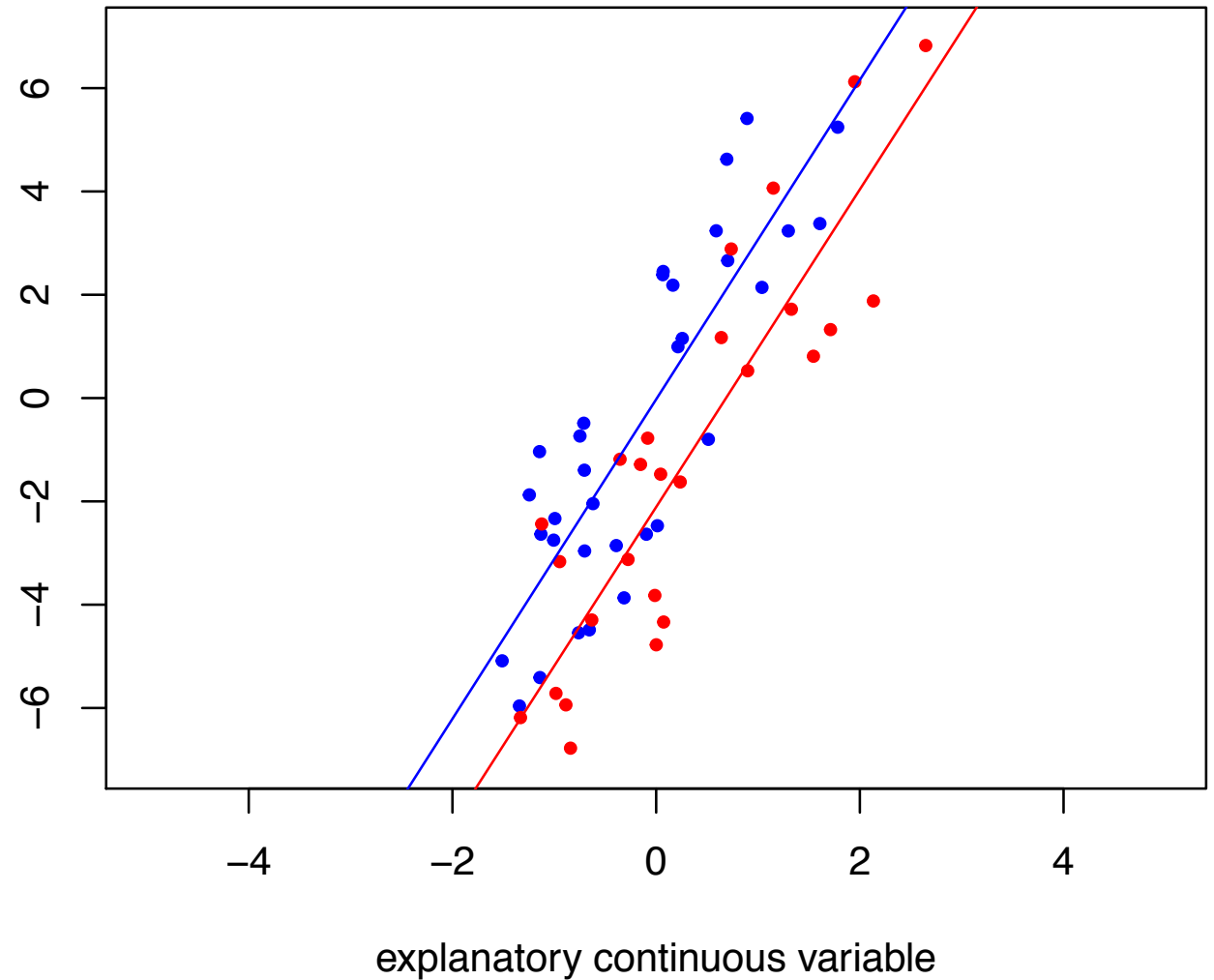
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02613    0.31286  -0.084   0.934
x             3.08222    0.23471  13.132 <2e-16 ***
sx           -0.08598    0.47224  -0.182   0.856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
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
```



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

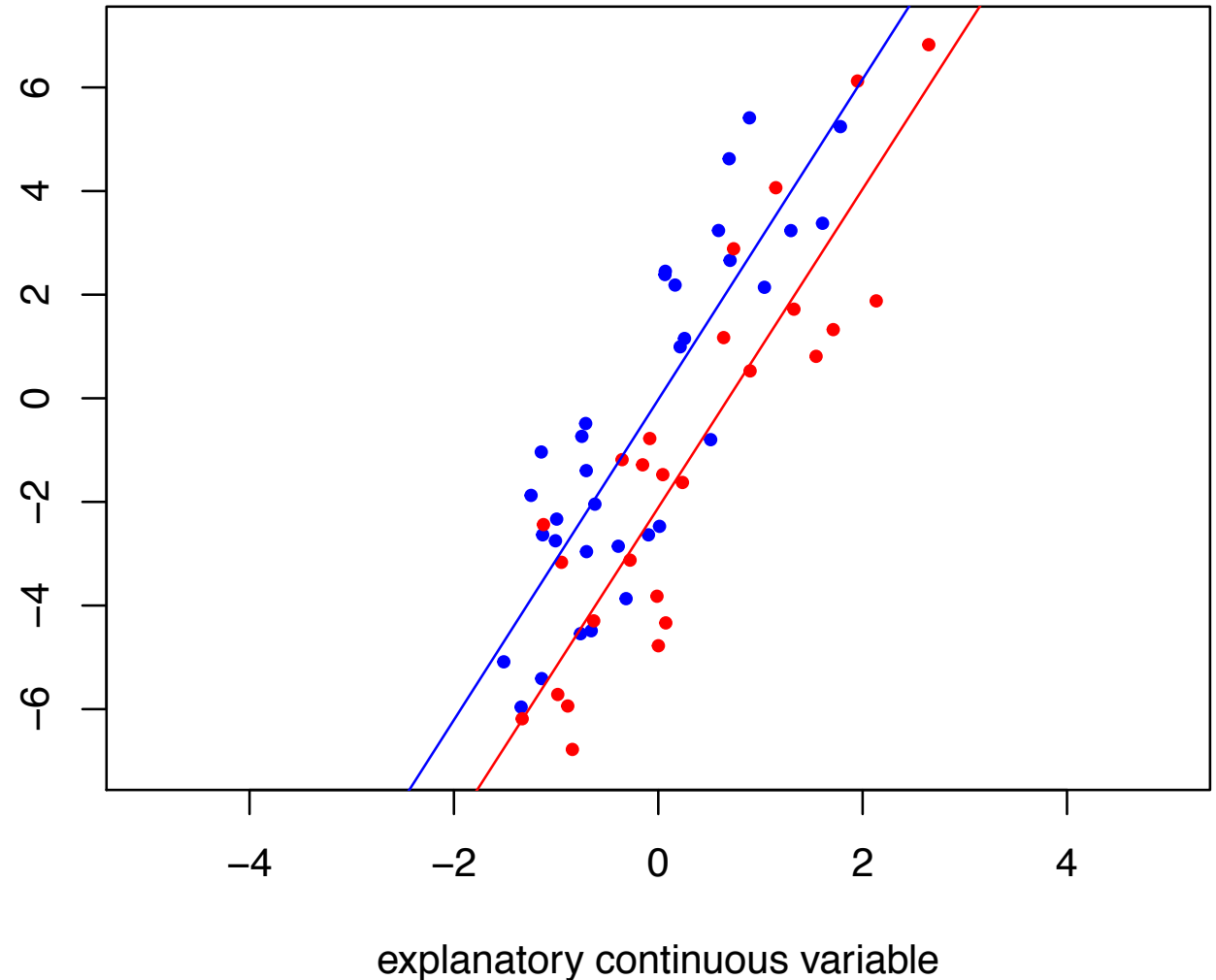
Let's try this



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

- We really need two slopes and two intercepts, one for each sex...



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

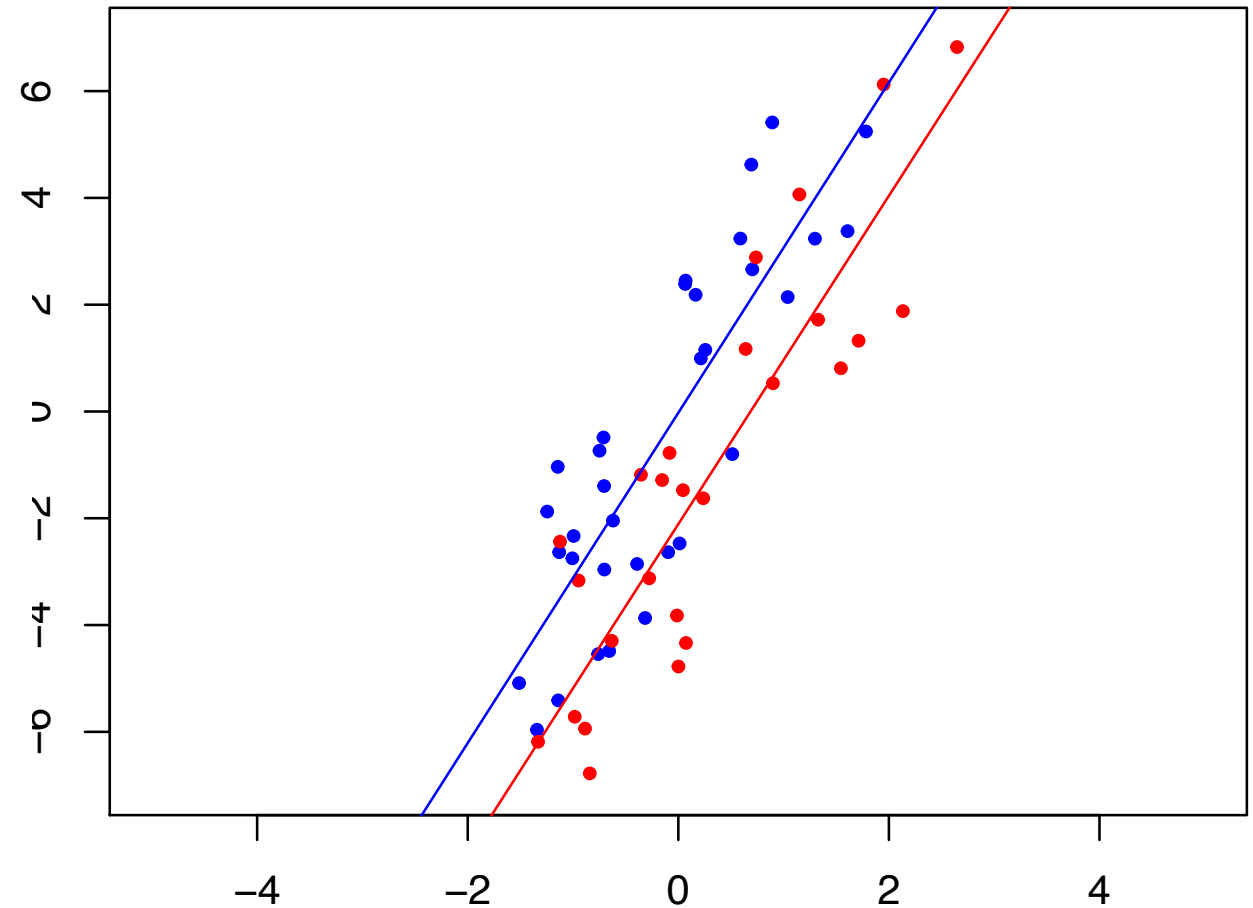
Let's try this

```
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
```



explanatory continuous variable

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

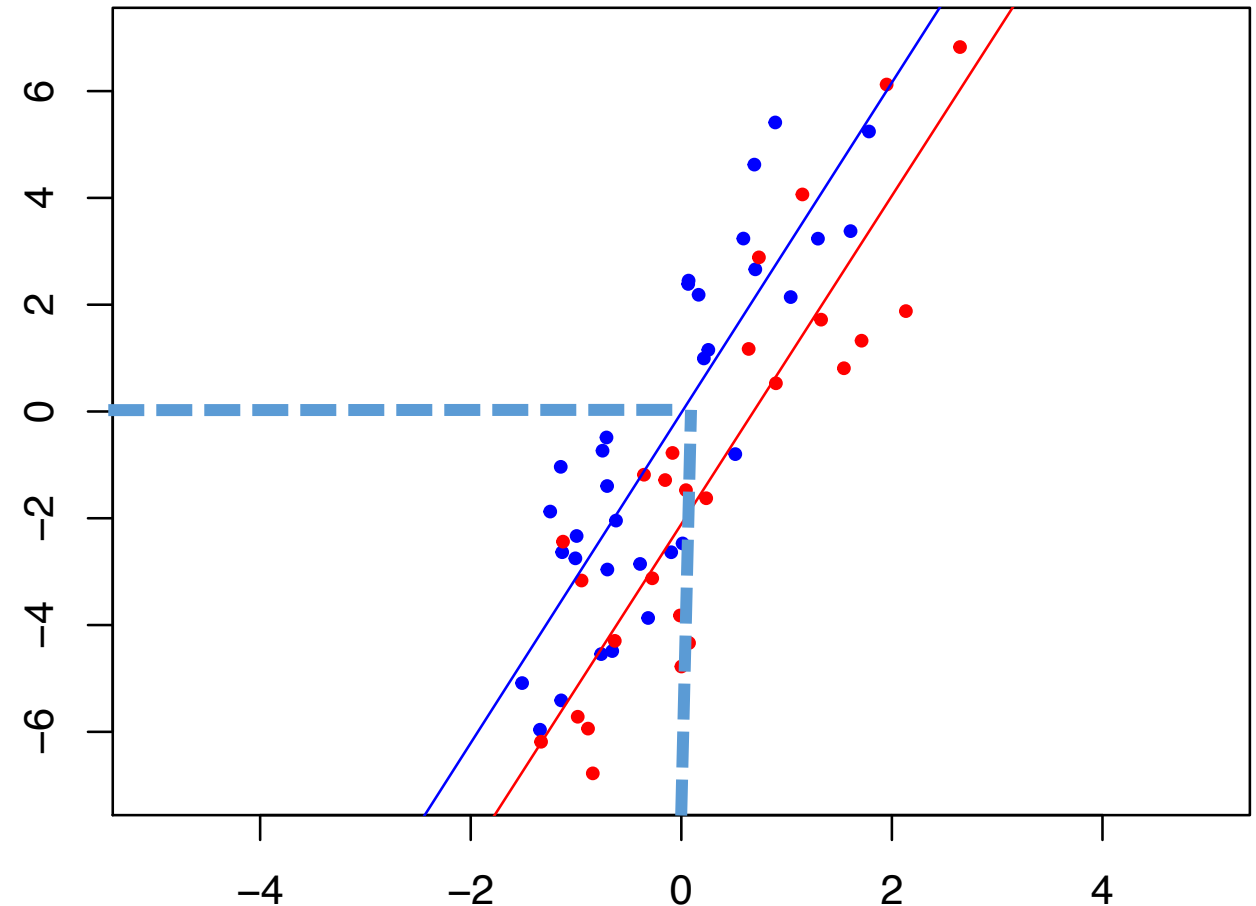
- Intercept: reference group

```
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
```



explanatory continuous variable

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

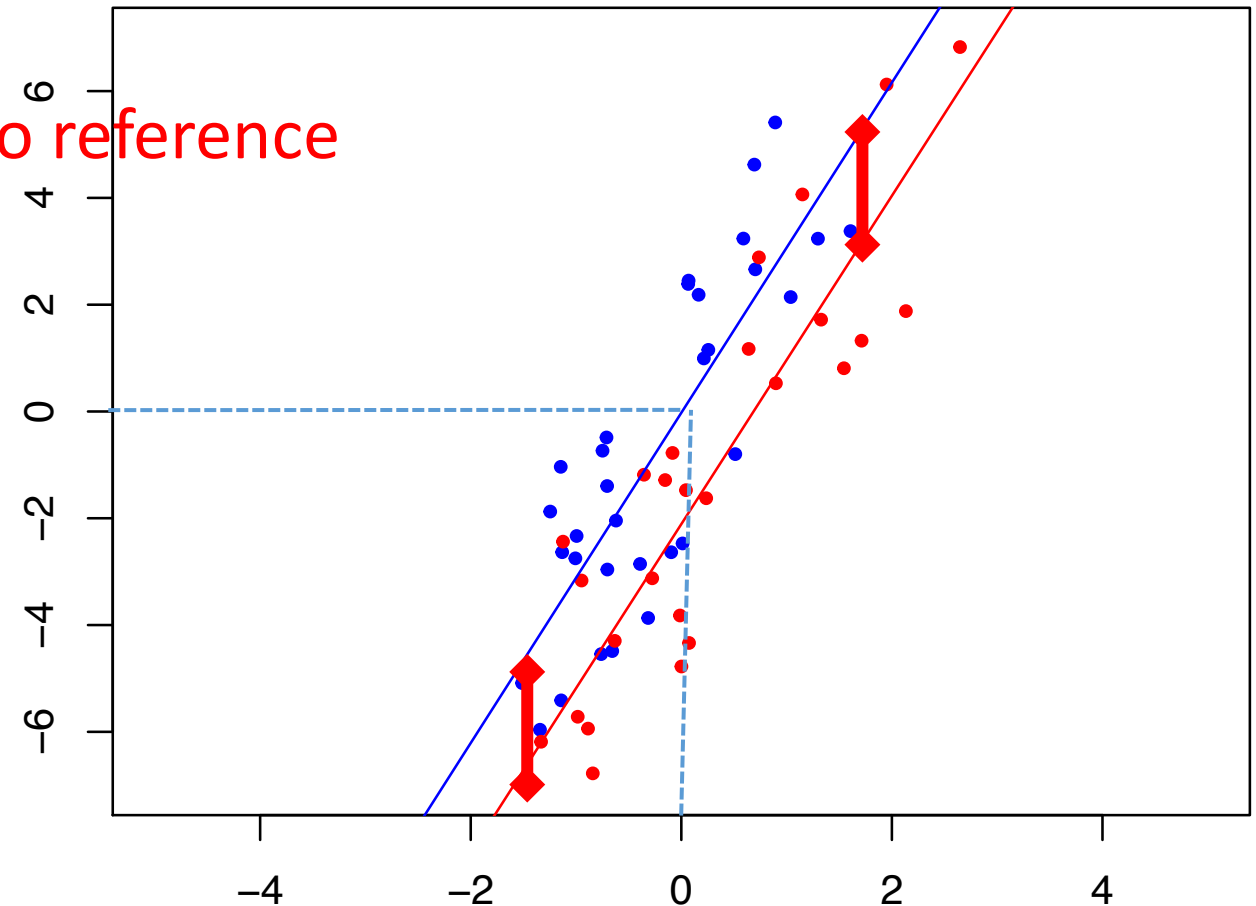
- Intercept: reference group
- b_2 (sx): *absolute difference of group 1 to reference*

```
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
```



explanatory continuous variable

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

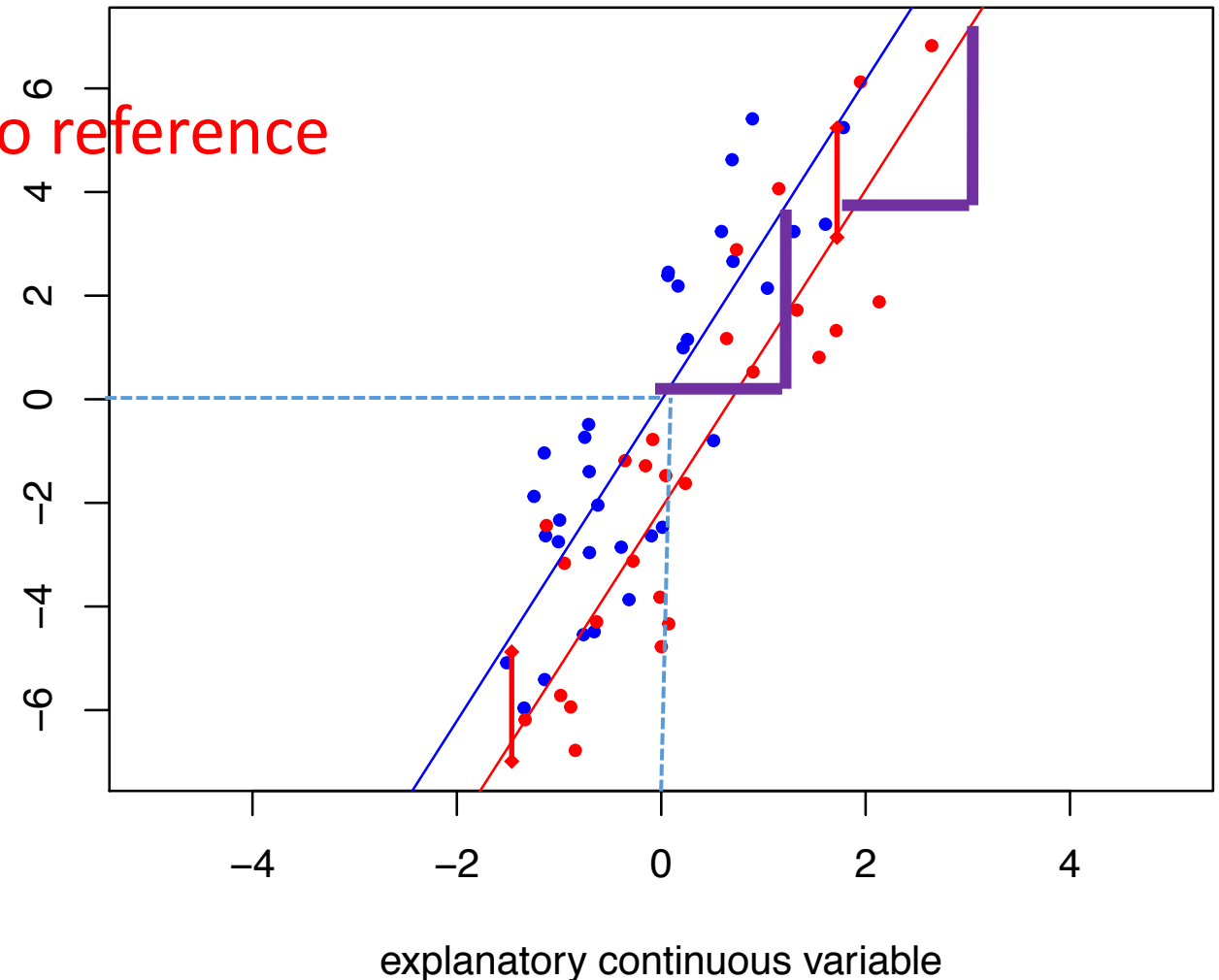
- Intercept: reference group
- b_2 (sx): *absolute difference* of group 1 to reference
- b_1 (x) slope, equal for both groups

```
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
```



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

- interaction between sex and tarsus

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

Let's try this

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

- interaction between sex and tarsus
- one more parameter estimate

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + \varepsilon_i$$

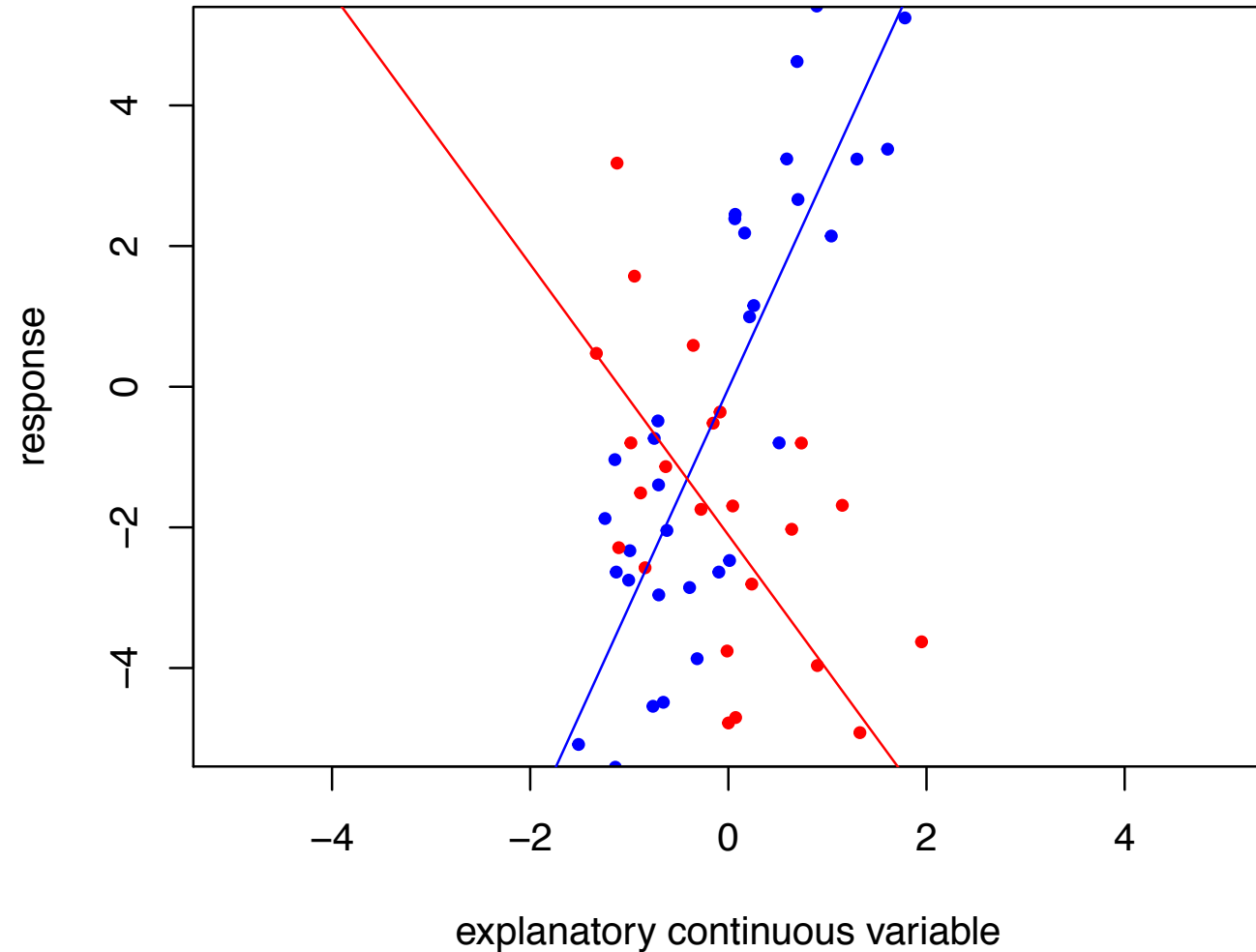
Let's try this

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

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

$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

Let's try this



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

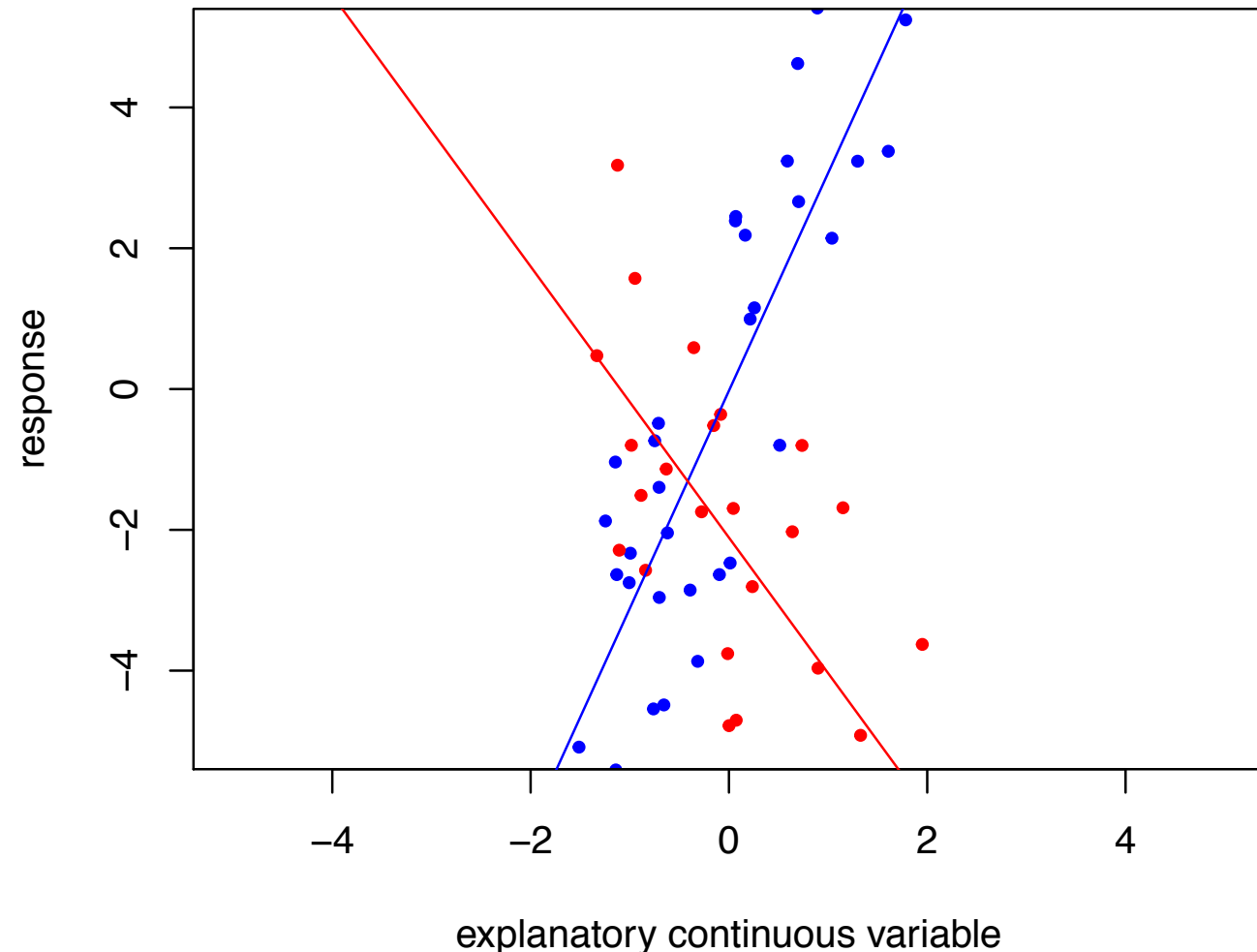
Let's try this

```
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
```



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

Let's try this

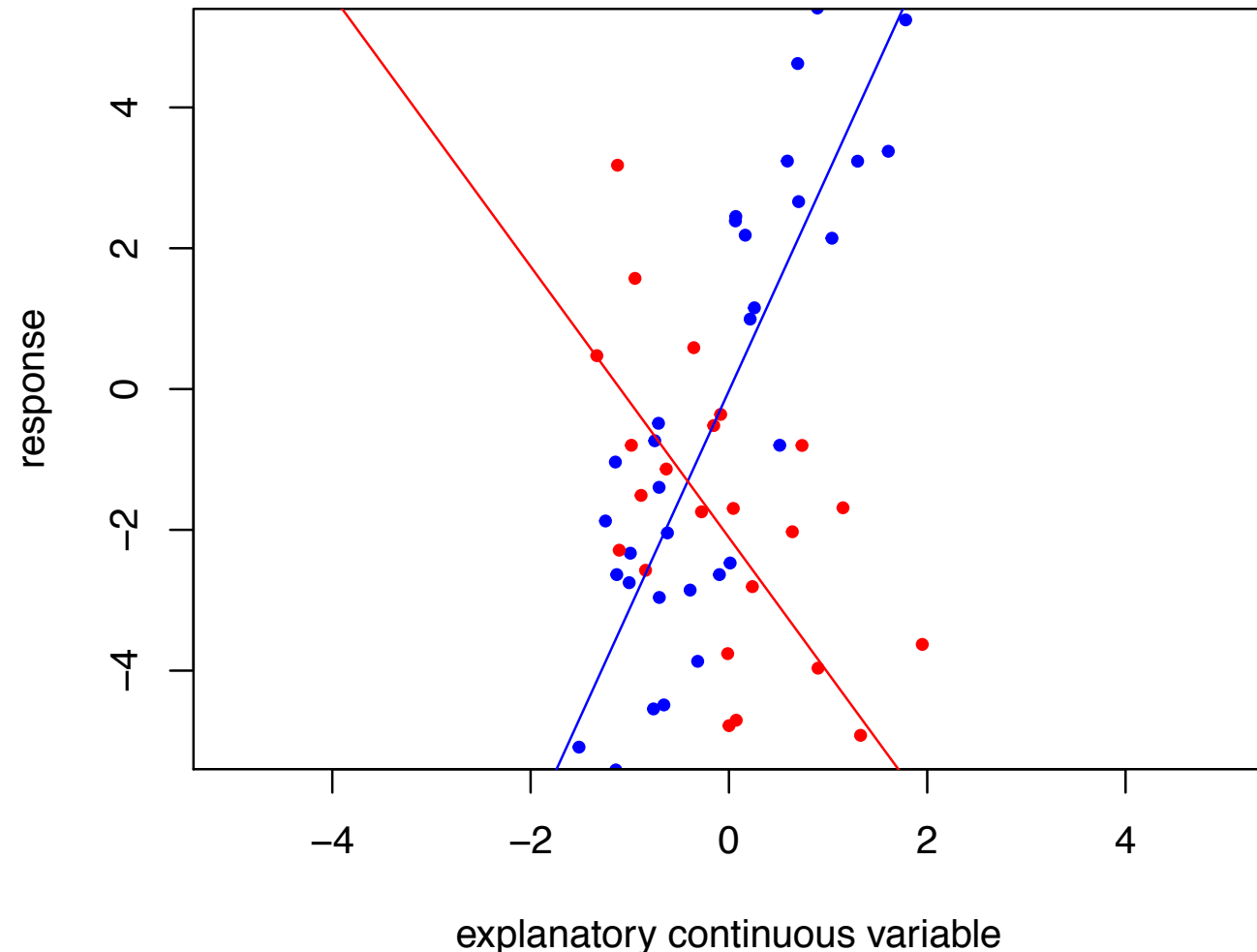
- Intercept: reference group, b_2 : *absolute difference* of group 1 to reference

```
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
```



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

Let's try this

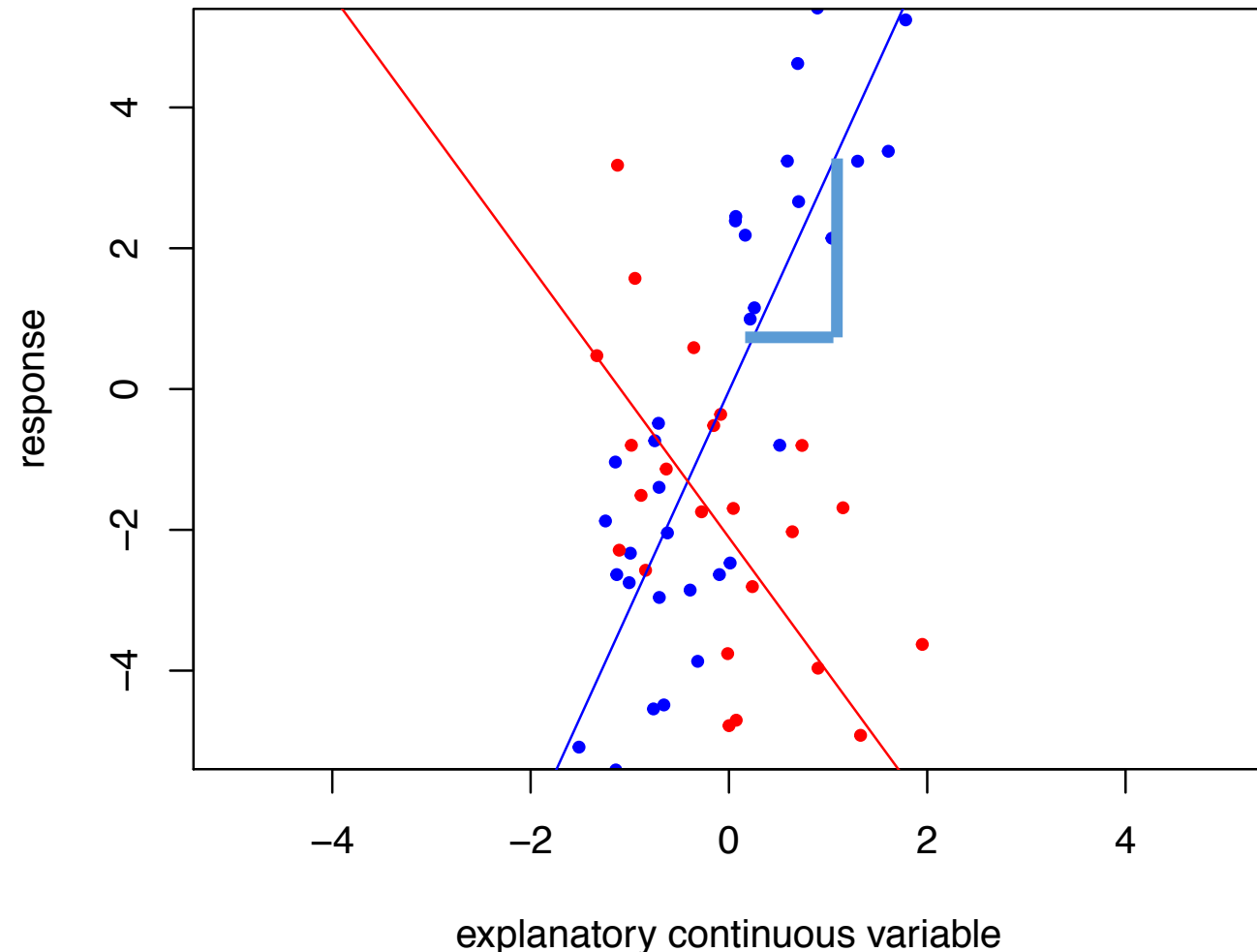
- Intercept: reference group, b_2 : *absolute difference* of group 1 to reference
- b_1 slope of reference group

```
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
```



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

Let's try this

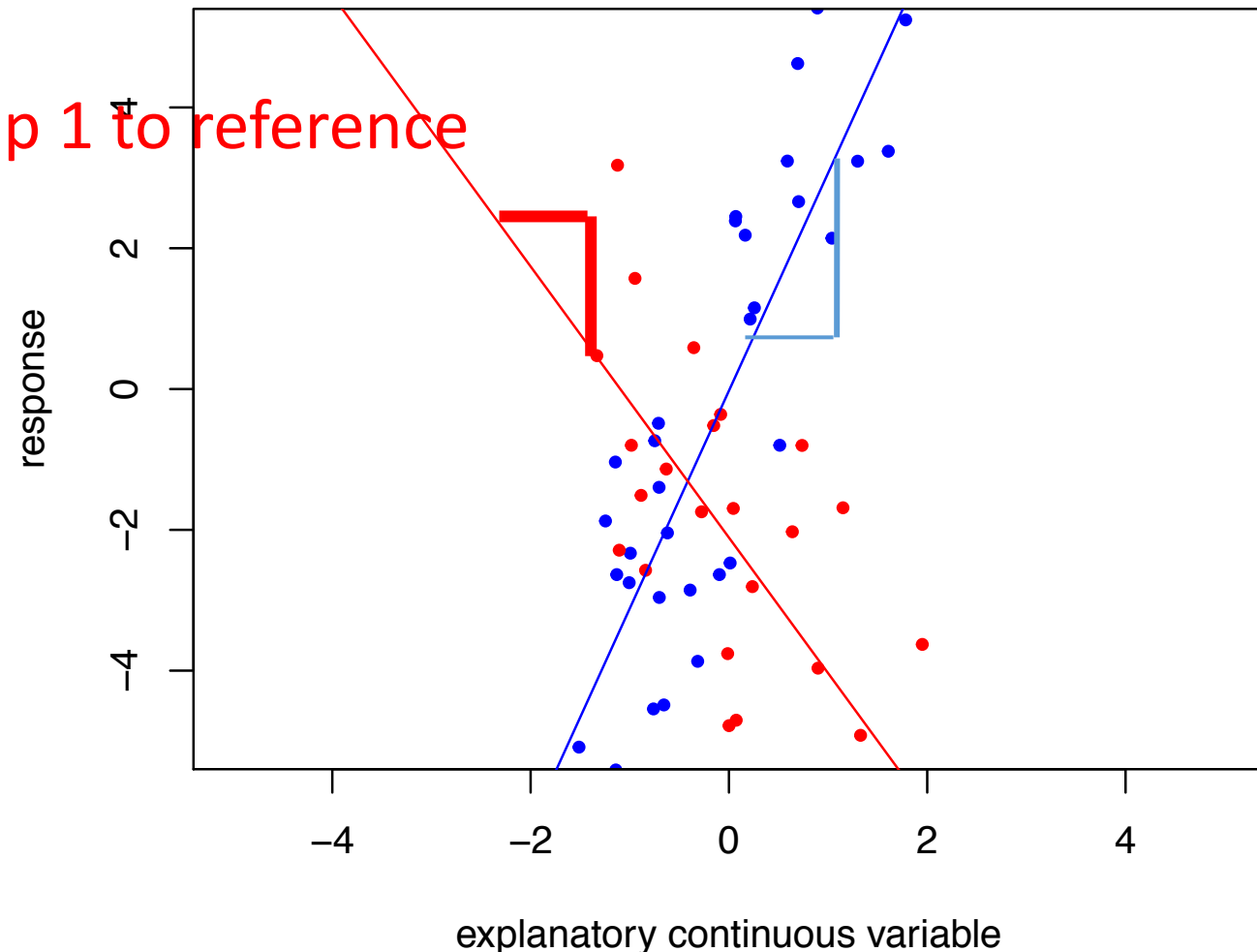
- Intercept: reference group, b_2 : *absolute difference* of group 1 to reference
- b_1 slope of reference group
- b_3 *absolute difference* of slope of group 1 to reference

```
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
```



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

Let's try this

- Intercept: reference group, b_2 : *absolute difference* of group 1 to reference
- b_1 slope of reference group
- b_3 *absolute difference* of slope of group 1 to reference

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

Residuals:

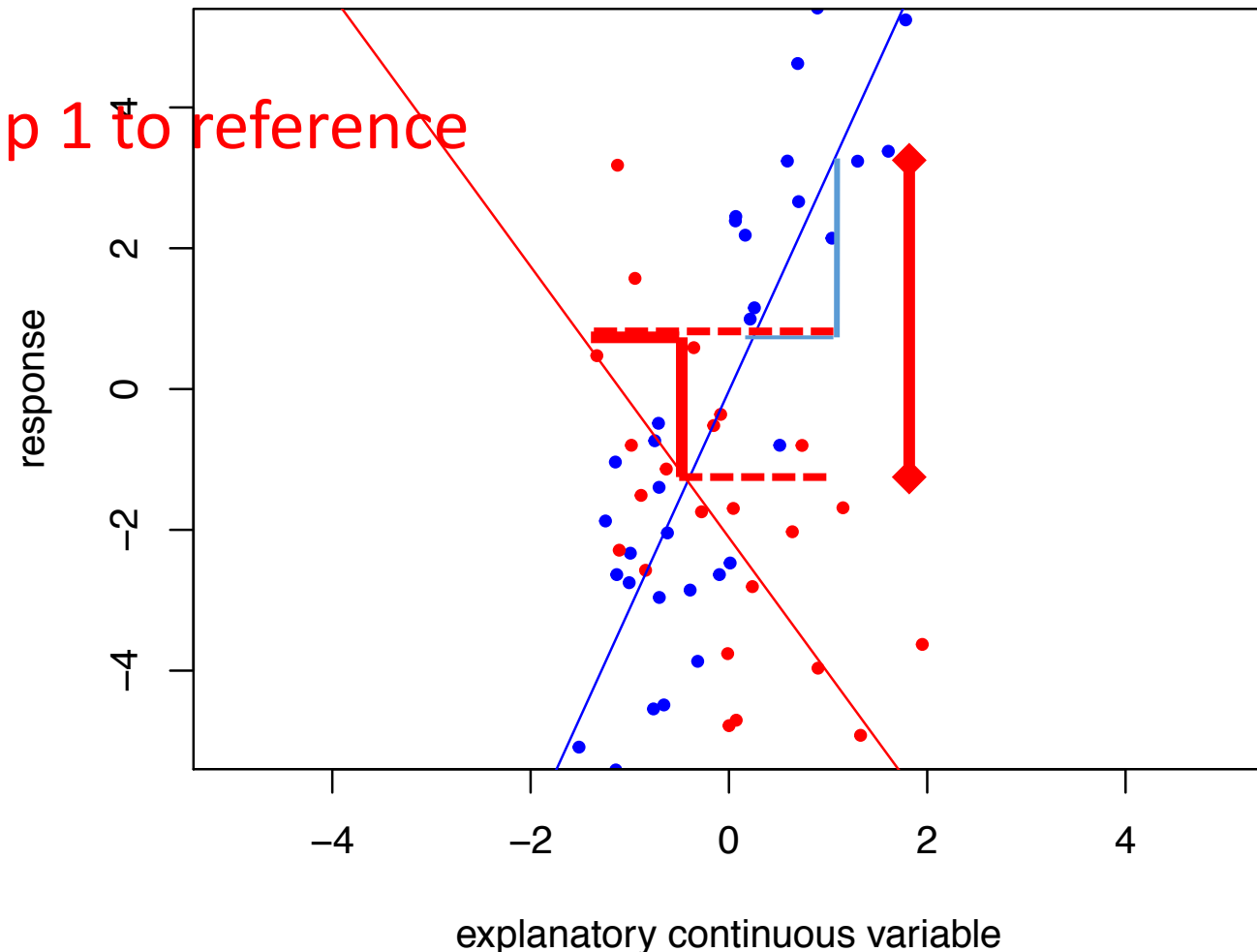
	Min	1Q	Median	3Q	Max
	-2.8687	-1.7008	-0.1129	1.5931	3.1264

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 ***

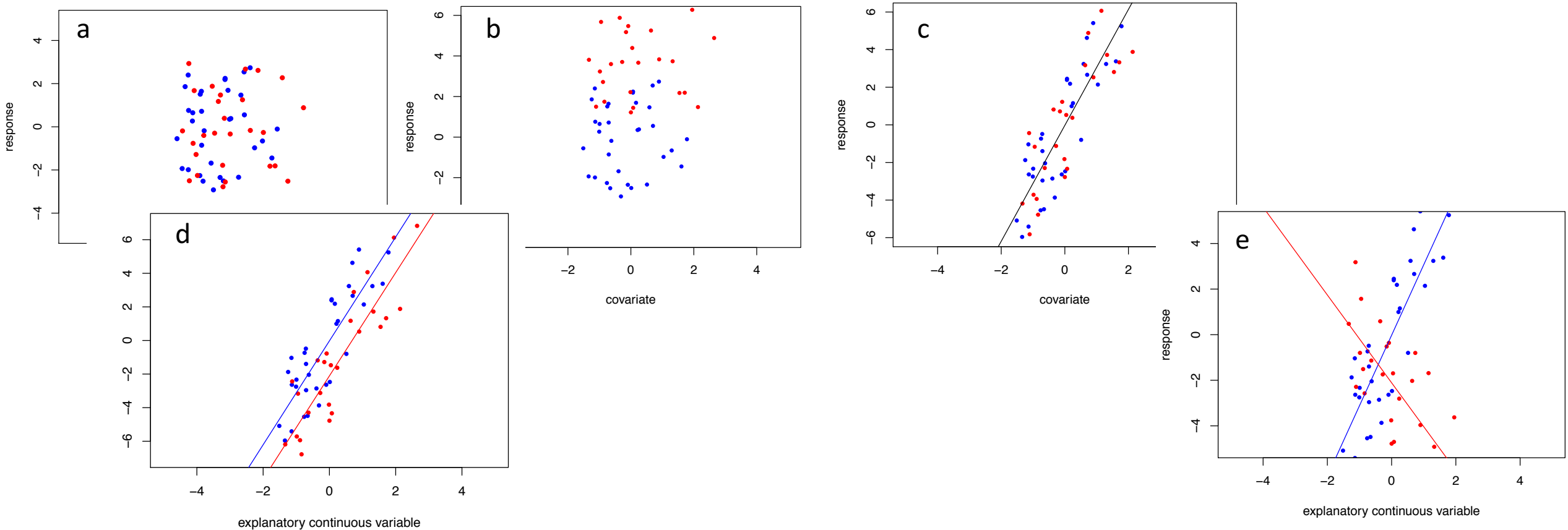
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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



$$y_i = b_0 + b_1x_{i0} + b_2x_{i1} + b_3x_{i0}x_{i1} + \varepsilon_i$$

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	-



Interpreting

- Interpret effect as *“if all other predictors are kept constant”*

Interpreting

- Interpret effect as *“if all other predictors are kept constant”*
- This is impossible for interaction effects

Linear models

- multivariate models
- two way analysis of variance
- multiple regression
- analysis of covariance (ANCOVA)
- interaction terms
- factorial analysis of variance

Linear models

- Response:
- continuous



- Explanatory (fixed)
- Covariates (continuous)
- Factors (categorical)

Learning aim

- Running linear models with more than one explanatory variable
- Interactions
- Interpretation

Do it now – HO 15!

- Run a model where you test whether the relationship between bill length (response) and tarsus differs between sexes in sparrows
- Interpret the results
- Write down model results on whiteboard
- Sketch model, including regression line(s) on whiteboard