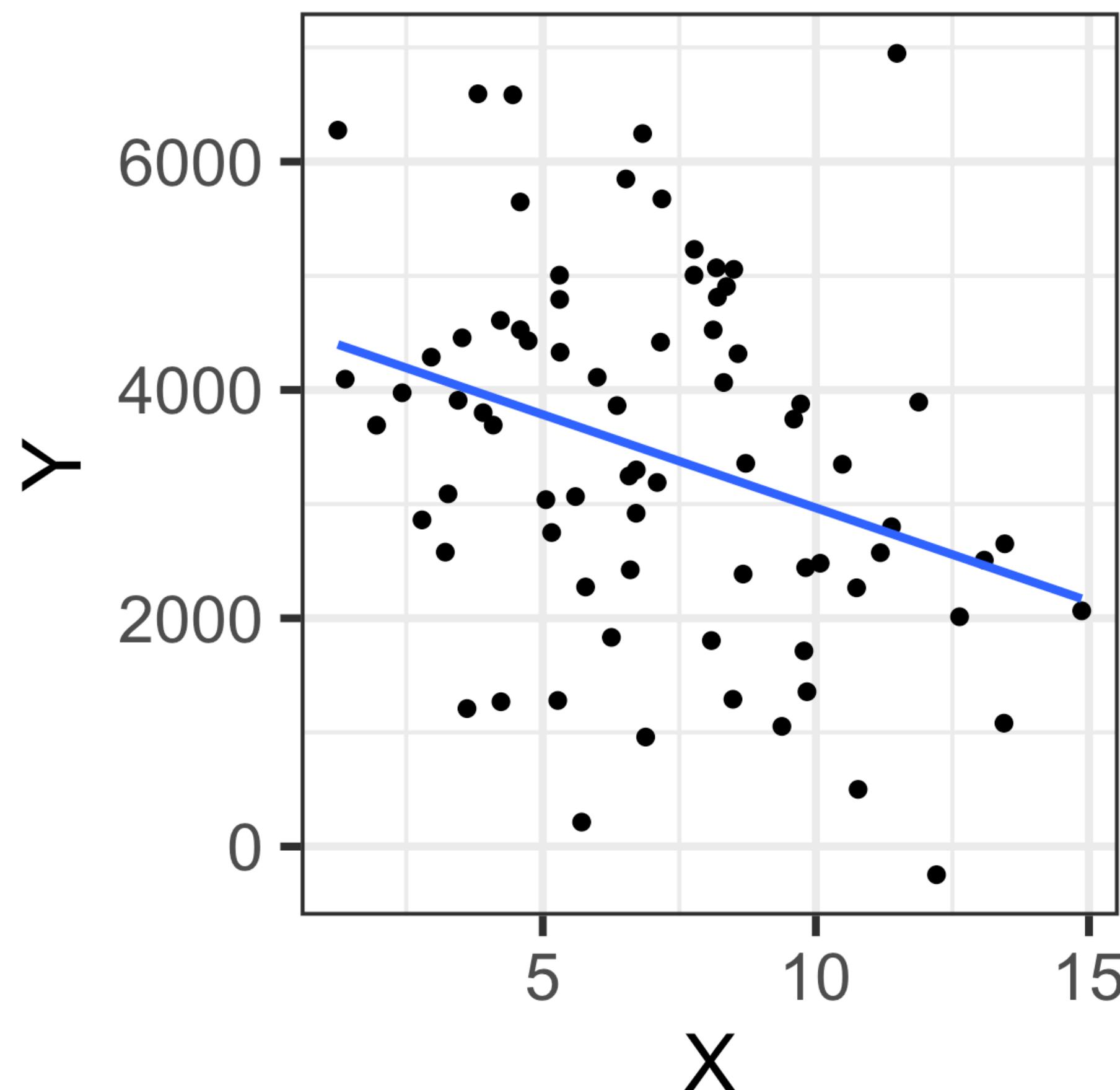


Lecture 14

11.18.21

Refresher Quiz



```
> summary(lm(Y ~ X))
```

Residuals:

	Min	1Q	Median	3Q	Max
	-3453.9	-992.9	-128.6	1022.9	4224.6

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4601.95	442.18	10.407	4.52e-16 ***
ca\$x	-163.51	56.52	-2.893	0.00503 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1536 on 73 degrees of freedom

Multiple R-squared: 0.1028, Adjusted R-squared: 0.09055

F-statistic: 8.368 on 1 and 73 DF, p-value: 0.00503

Given these data, what can you tell me about the relationship between X and Y? Be sure to be as specific as possible.

Testing the hypothesis $H_0 : \beta_1 = 0$

```
> summary(lm(happiness ~ income, data = income_data))
```

Call:

```
lm(formula = happiness ~ income, data = income_data)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.5033	-0.8873	0.0459	0.9218	3.1279

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.34291	0.16505	8.136	3.3e-15 ***
income	0.22765	0.01727	13.182	< 2e-16 ***

Signif. codes:	0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1			

Residual standard error: 1.234 on 496 degrees of freedom

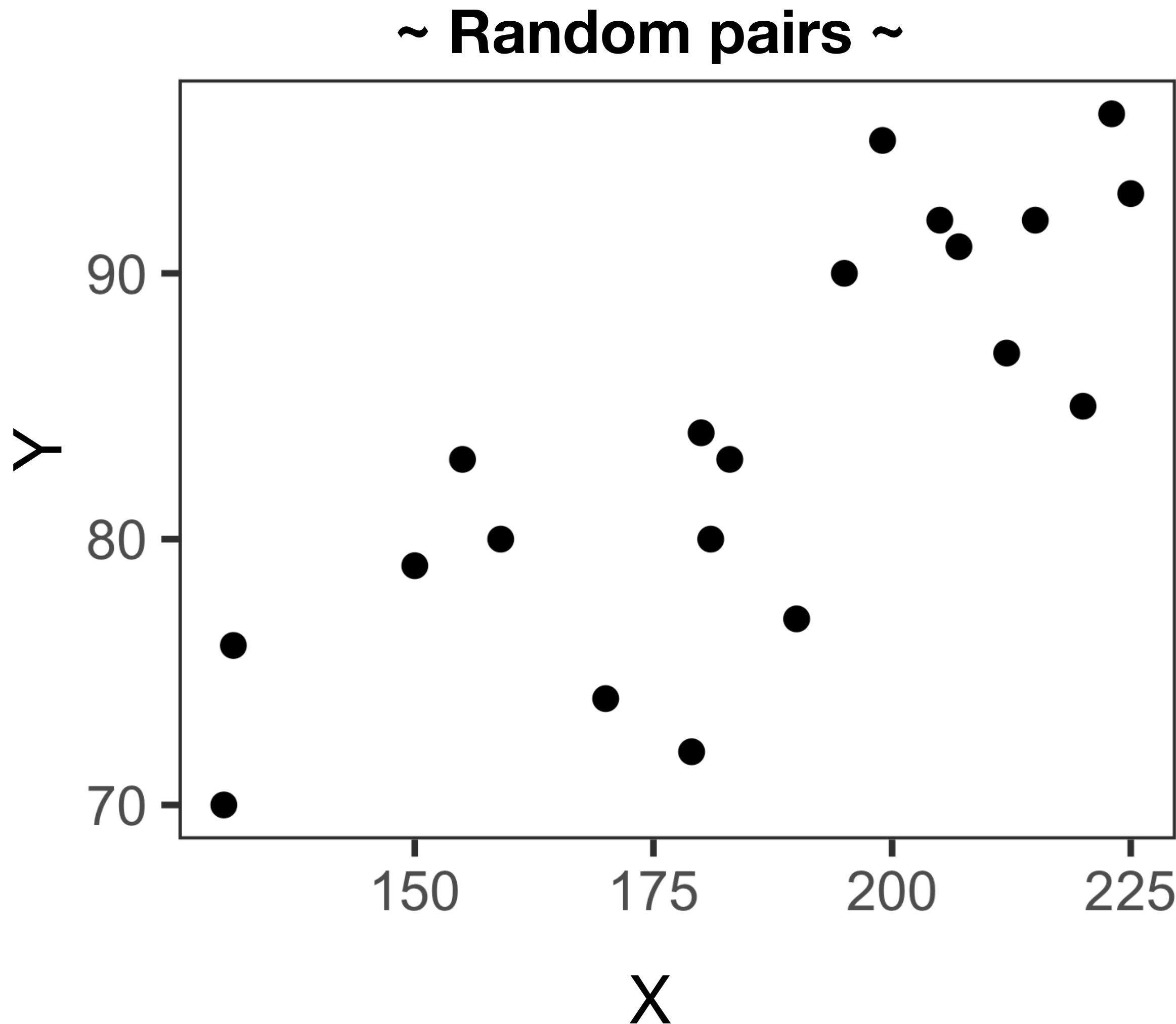
Multiple R-squared: 0.2595, Adjusted R-squared: 0.258

F-statistic: 173.8 on 1 and 496 DF, p-value: < 2.2e-16

Conditions for inference with regression

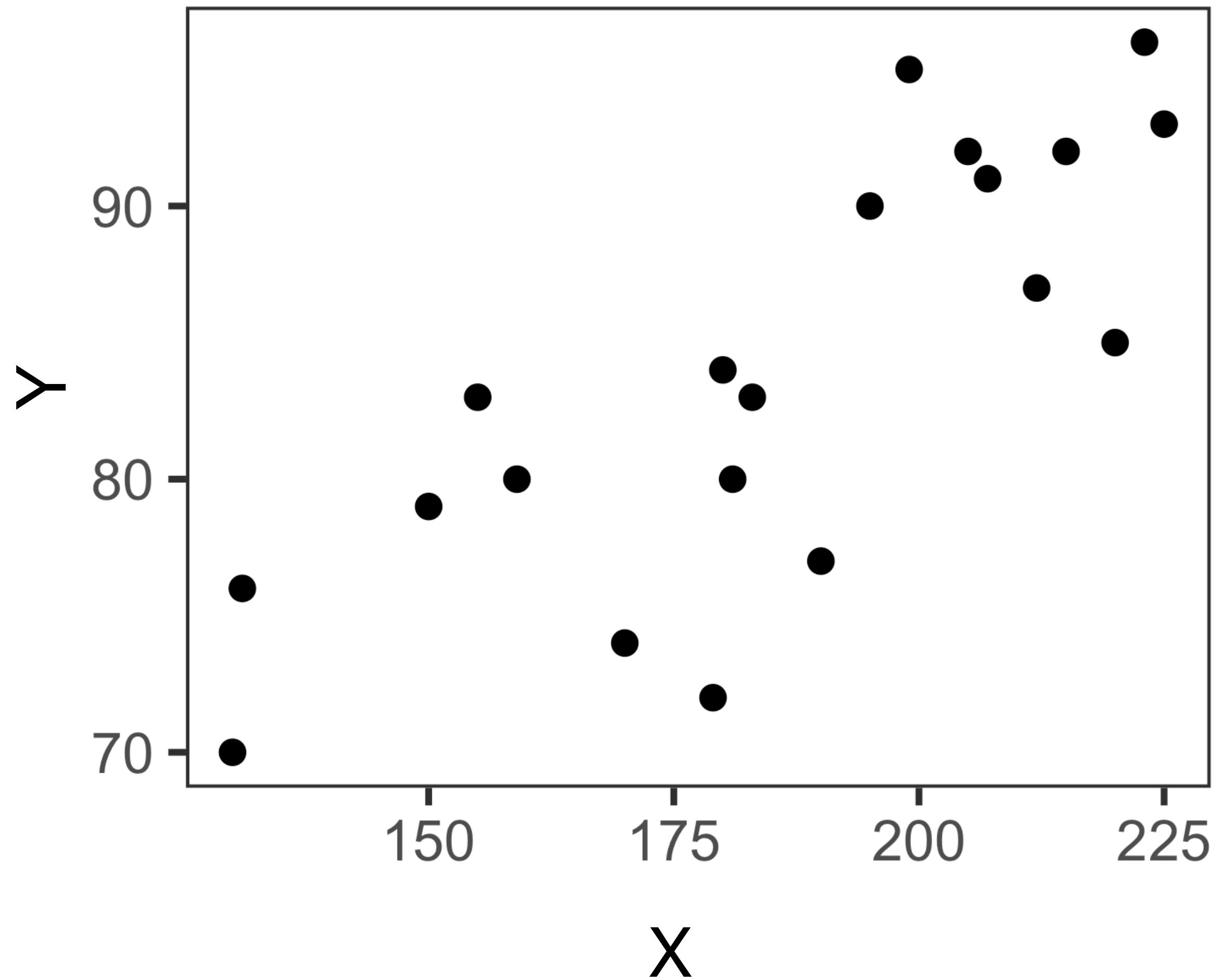
- **Random sampling:** avoid pairing, blocking, or hierarchical structure
 - Each (X, Y) is a random pair from the large population **OR** for a set X value, each Y value is a random observation from its population
 - **Independence:** each observed pair (X, Y) must be considered independent of the others

Assume: random, independent samples

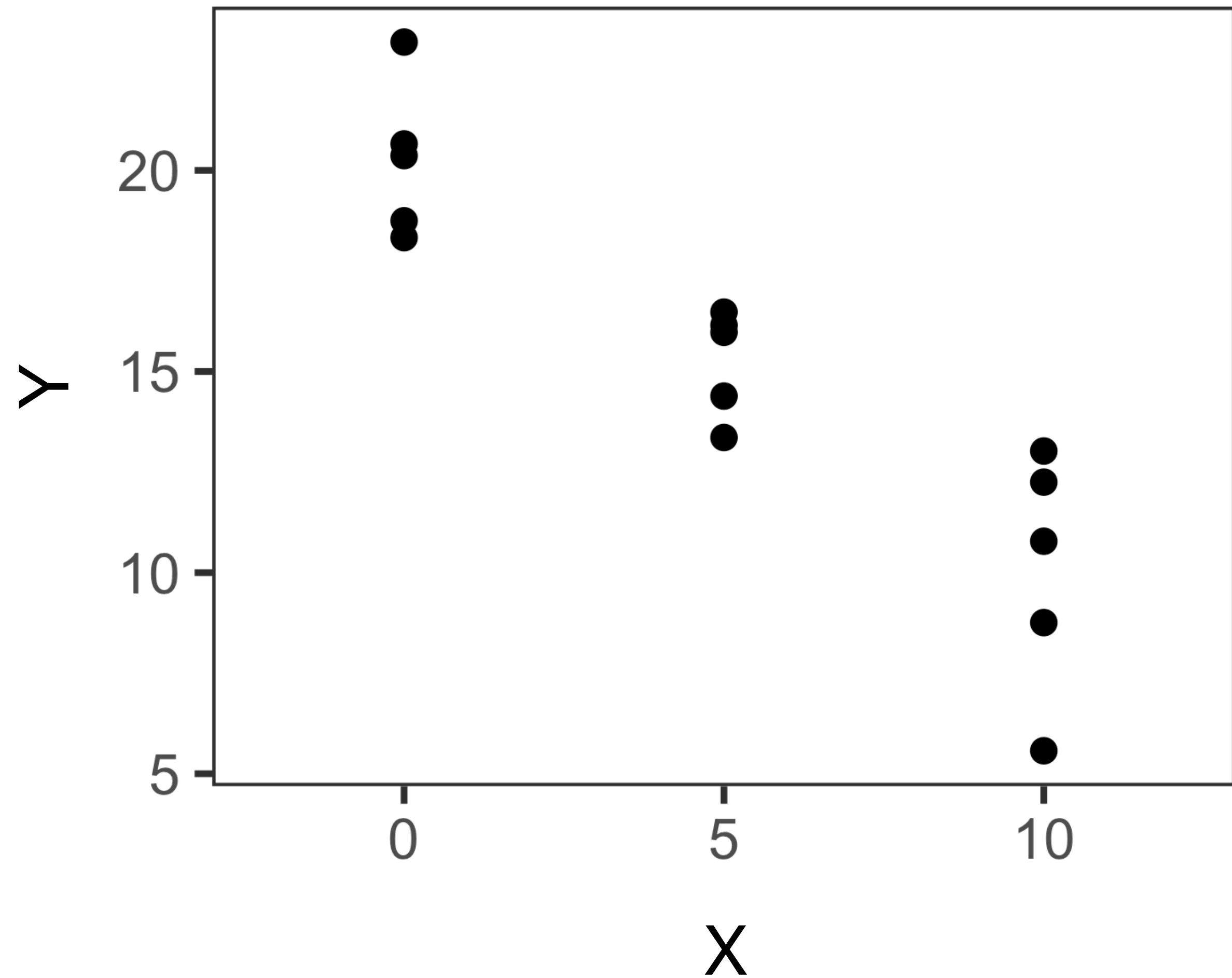


Assume: random, independent samples

~ Random pairs ~

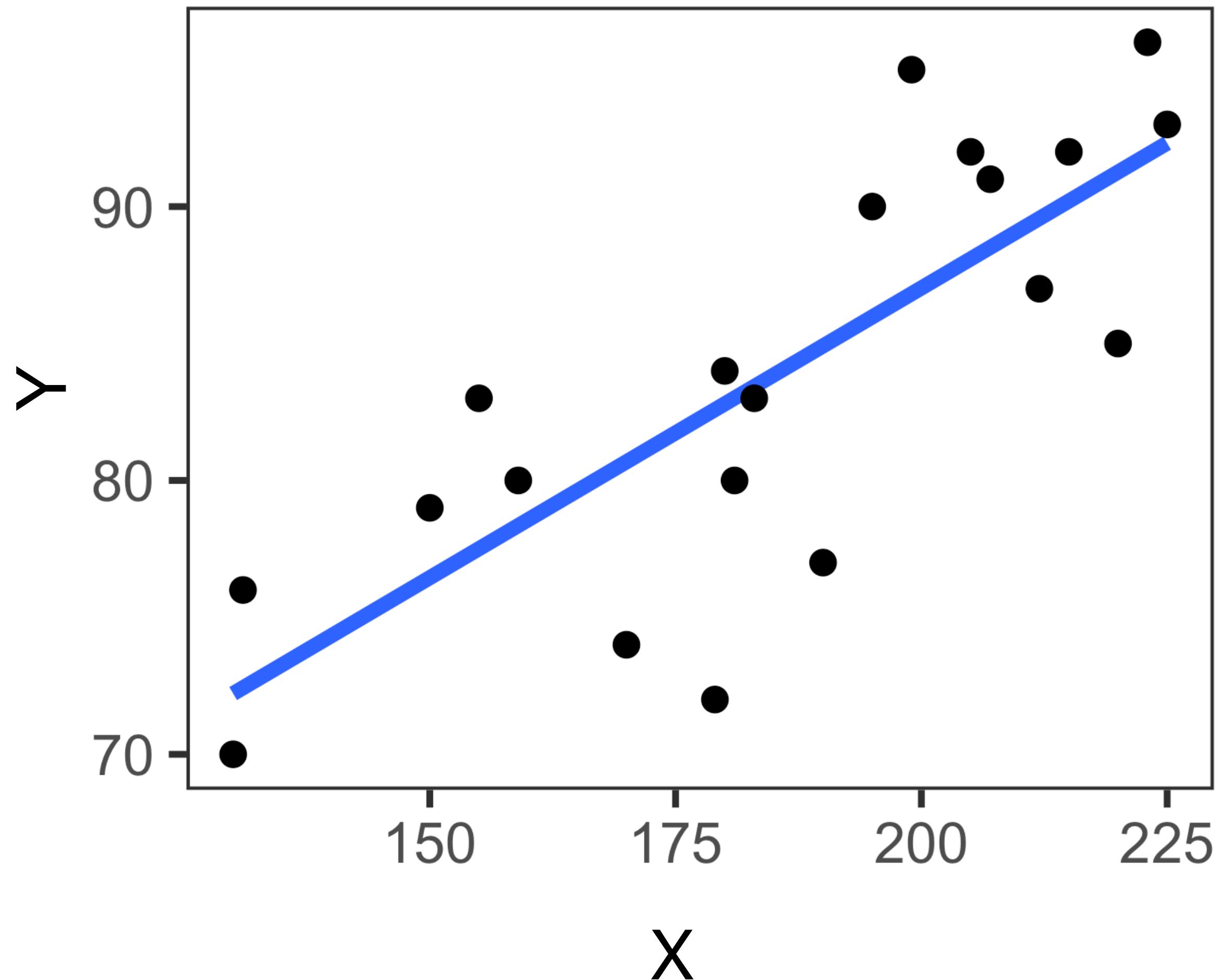


~ Random Y for each X ~

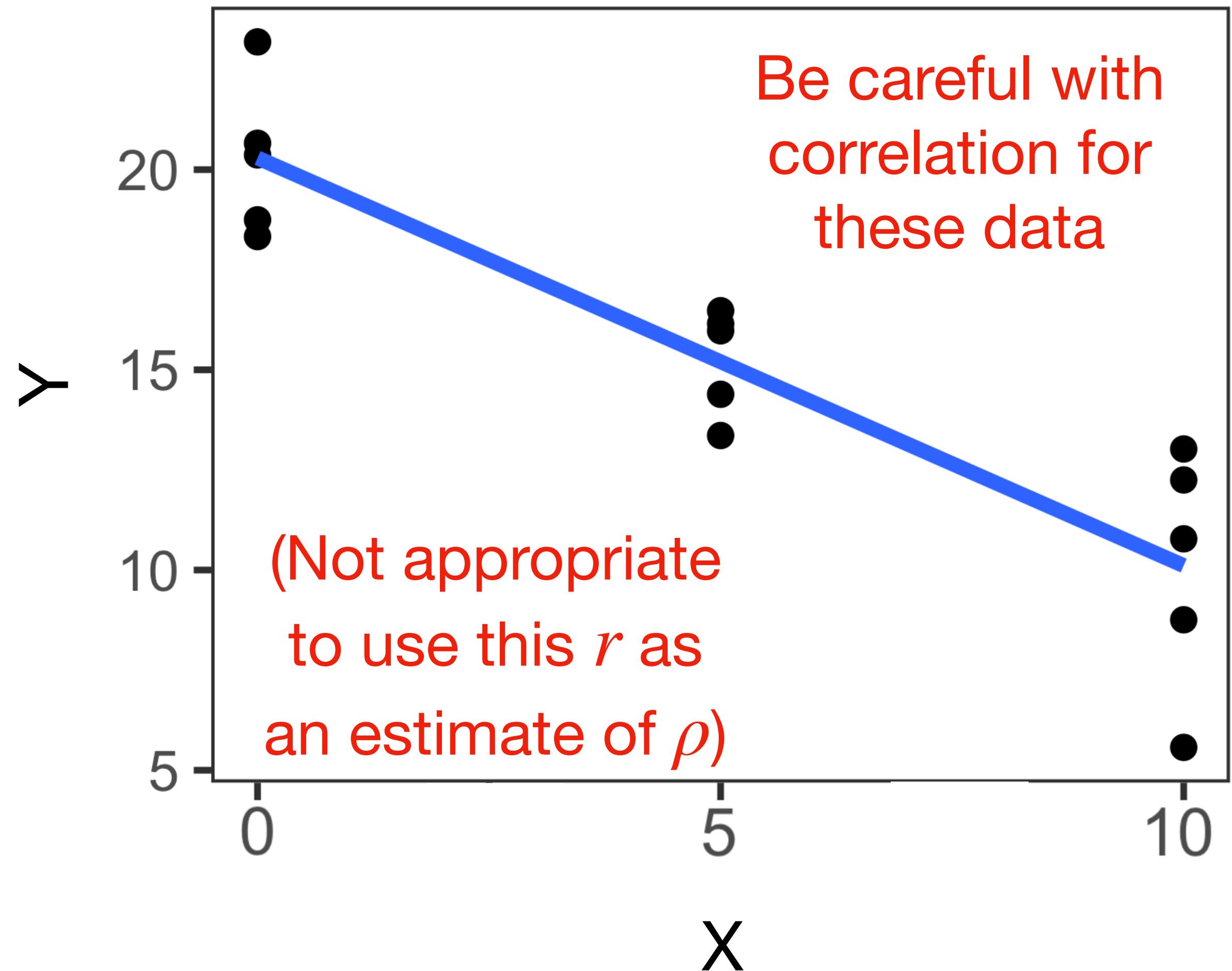


Assume: random, independent samples

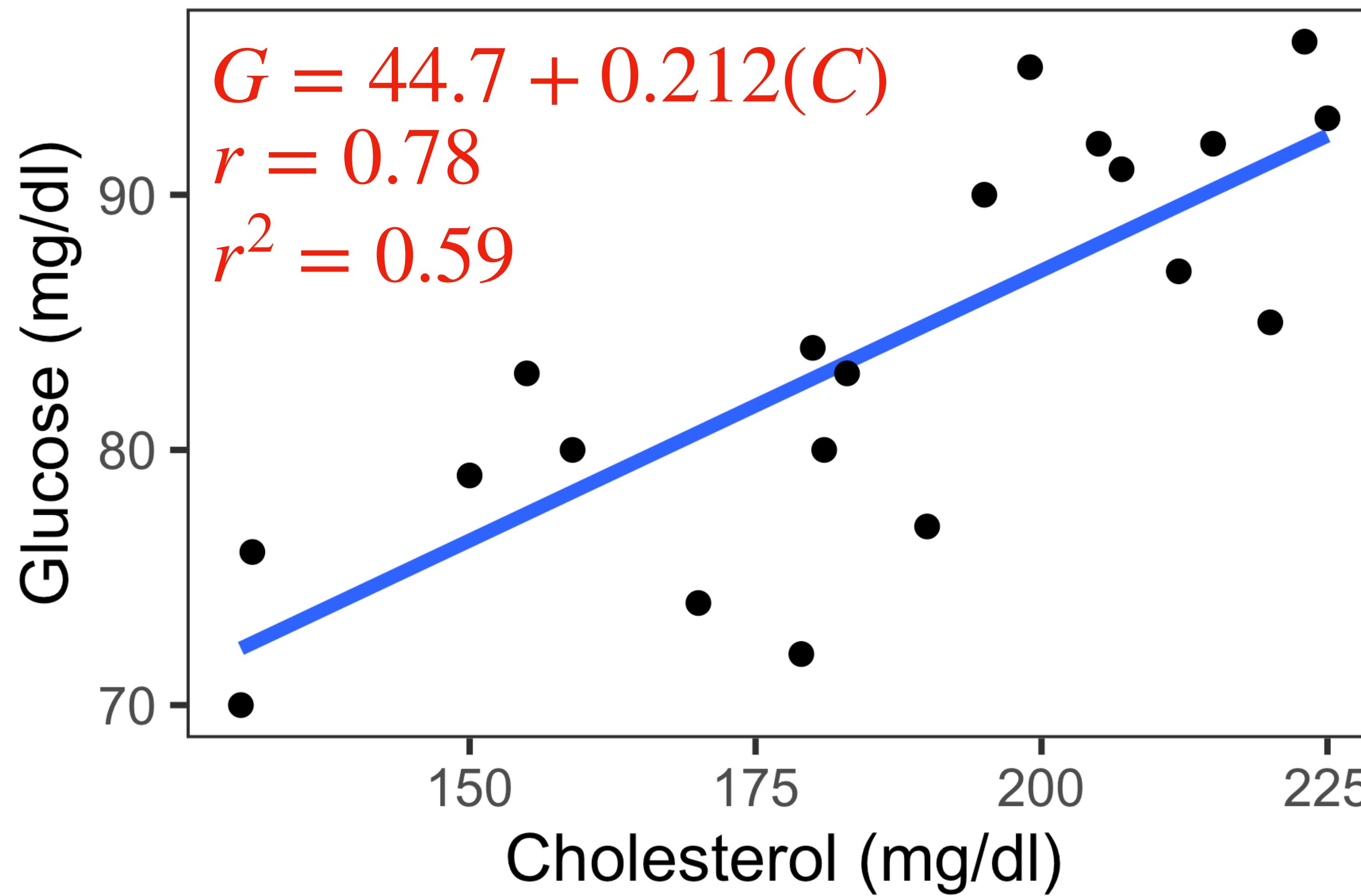
~ Random pairs ~



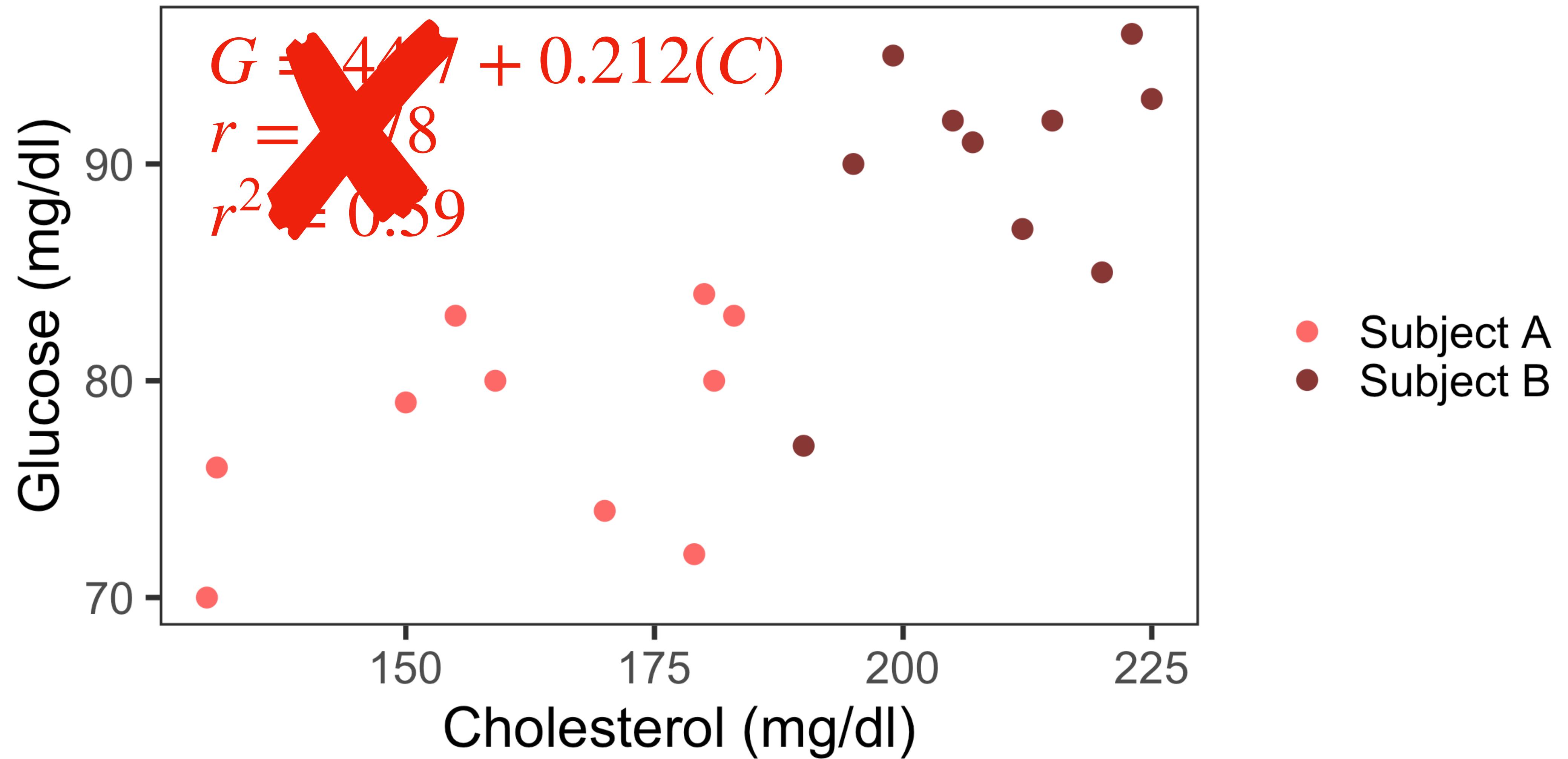
~ Random Y for each X ~



Assume: random, independent samples

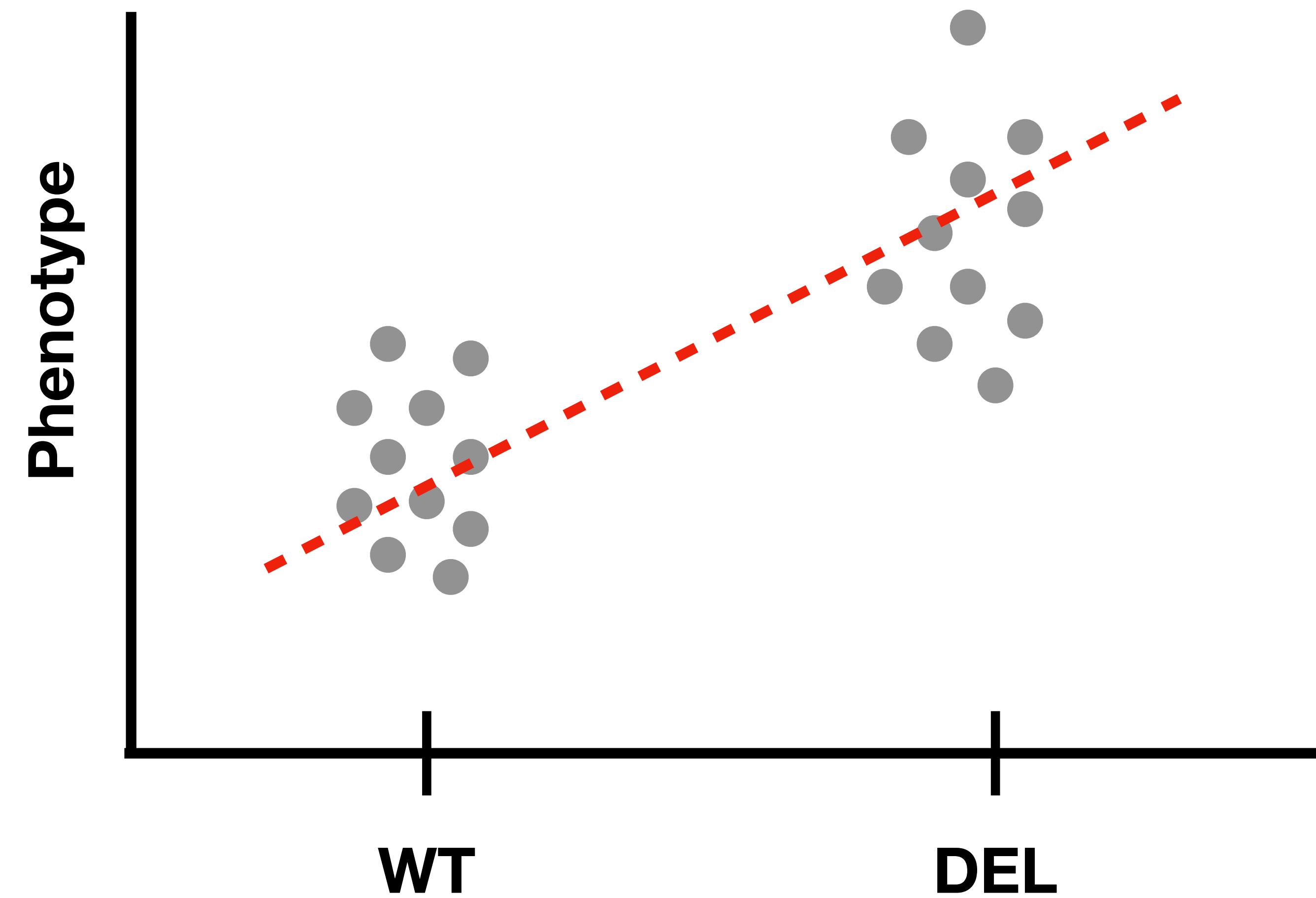


Assume: random, independent samples

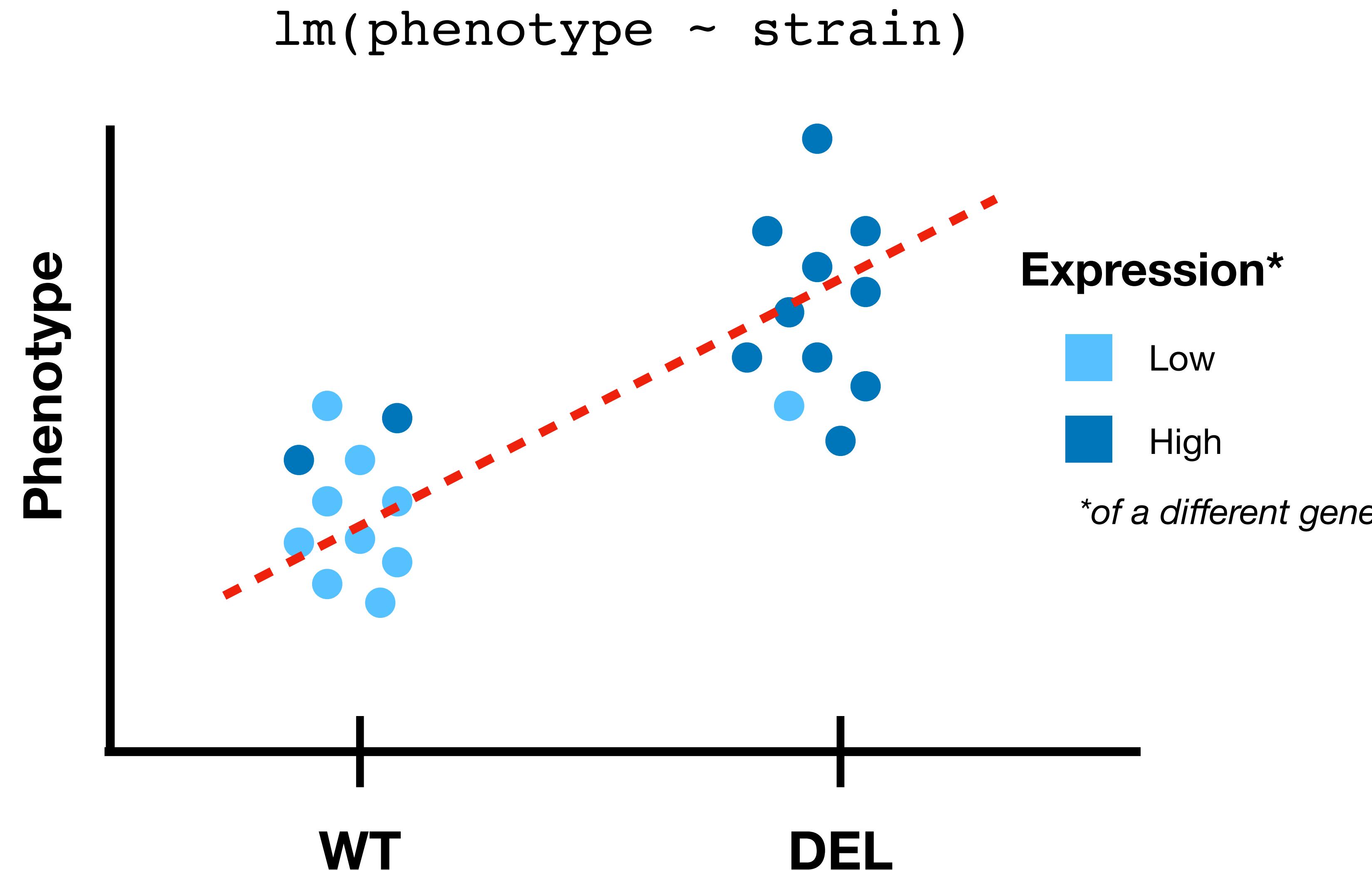


Using residual values to identify covariates

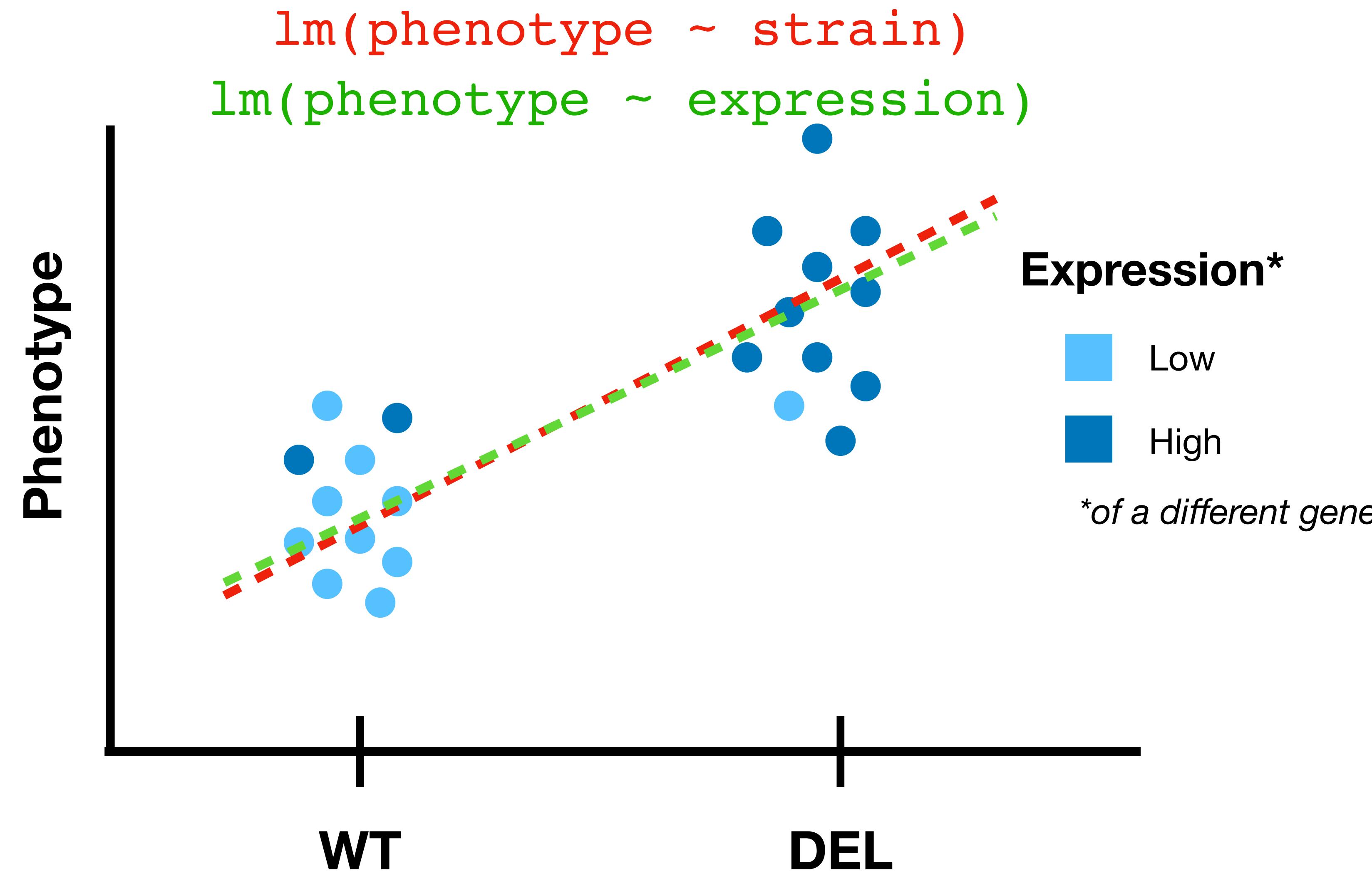
`lm(phenotype ~ strain)`



Using residual values to identify covariates



Using residual values to identify covariates

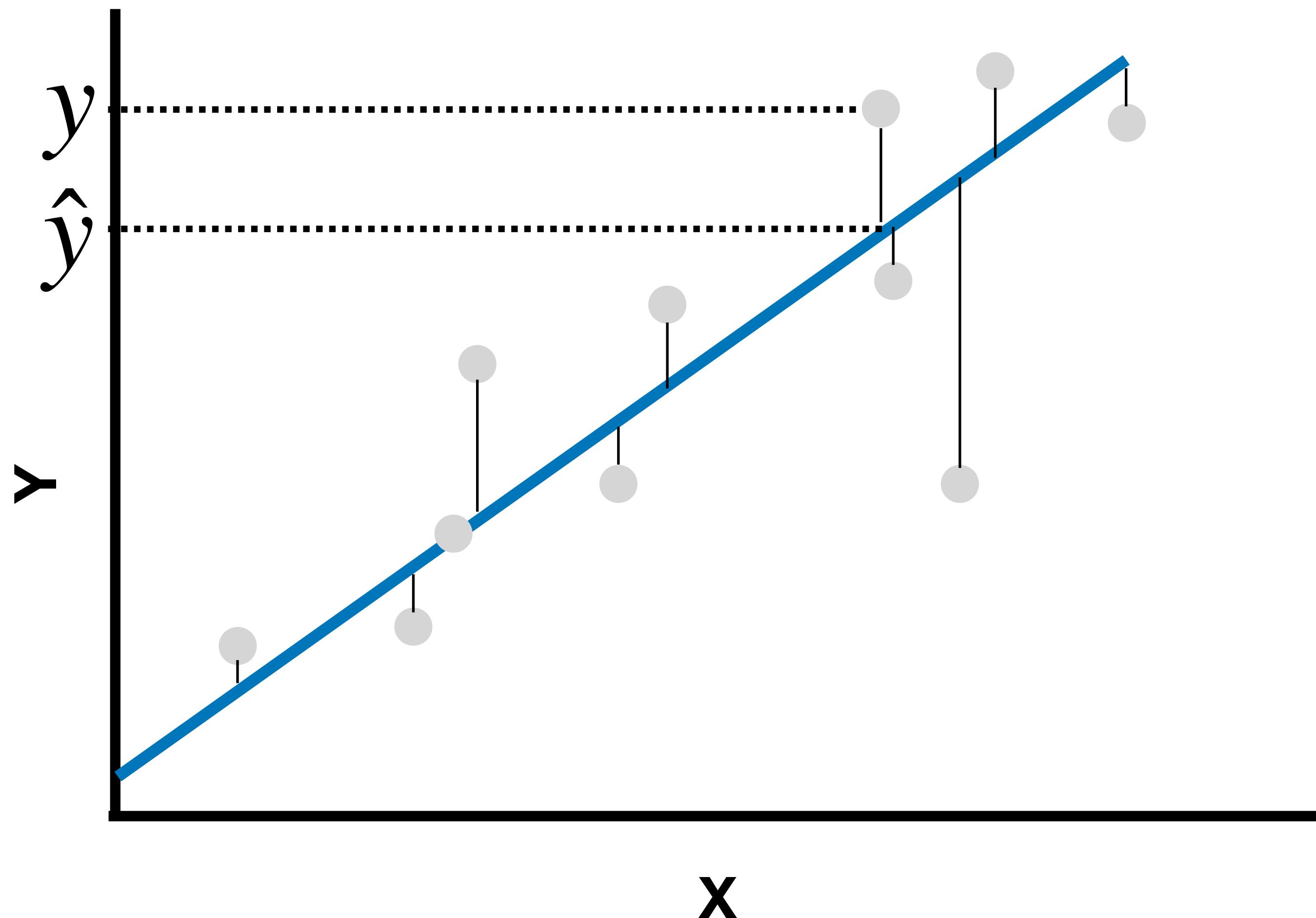


Conditions for inference with regression

- **Random sampling:** avoid pairing, blocking, or hierarchical structure
 - Each (X, Y) is a random pair from the large population **OR** for a set X value, each Y value is a random observation from its population
 - **Independence:** each observed pair (X, Y) must be considered independent of the others
- **Linearity:** there must be a linear relationship between X and Y
- **Normality:** The residuals must have a normal distribution

Residual plots to identify non-linearity

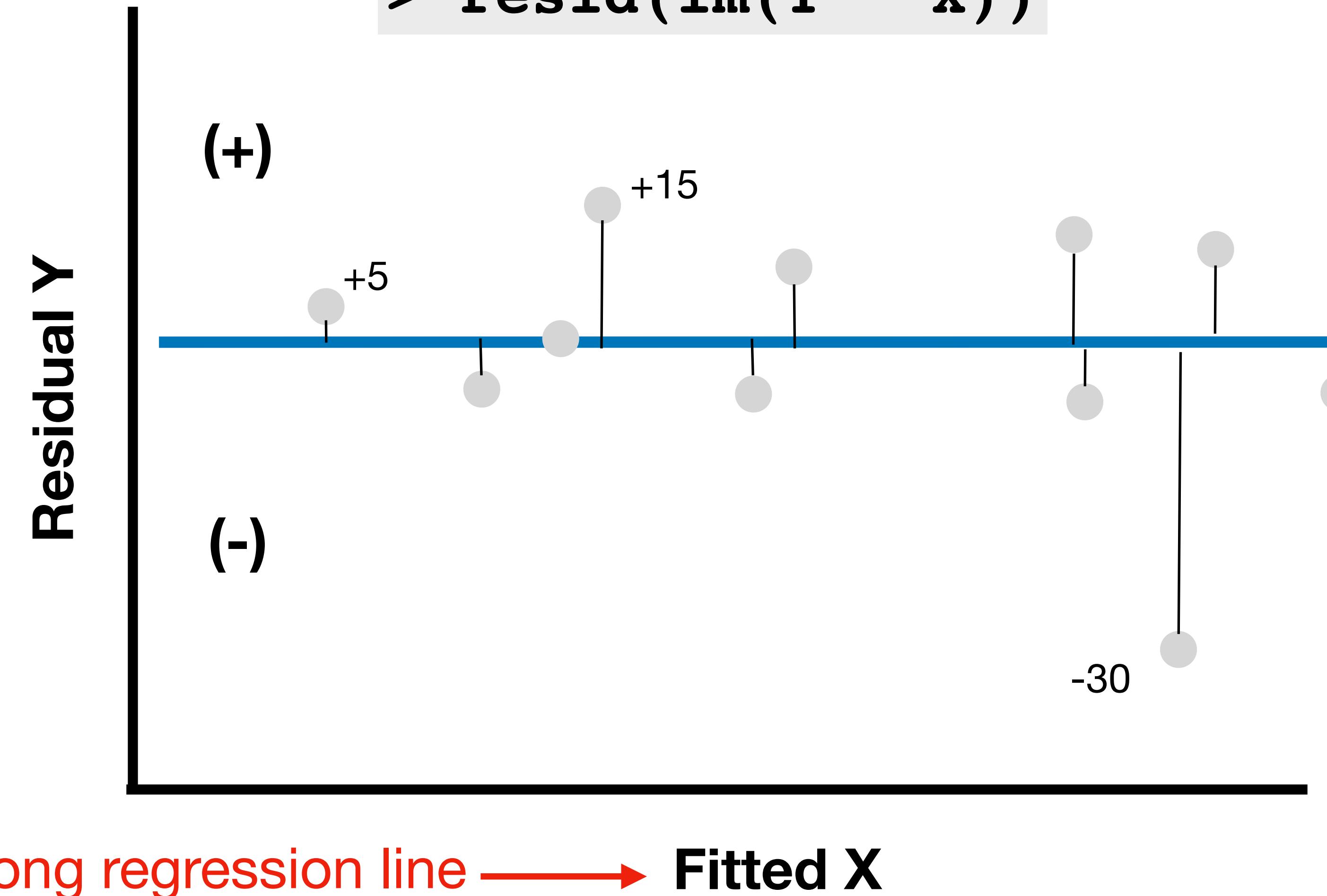
residual = observed - fitted



$$e_i = y_i - \hat{y}_i$$

Residual plots to identify non-linearity

```
> resid(lm(Y ~ x))
```



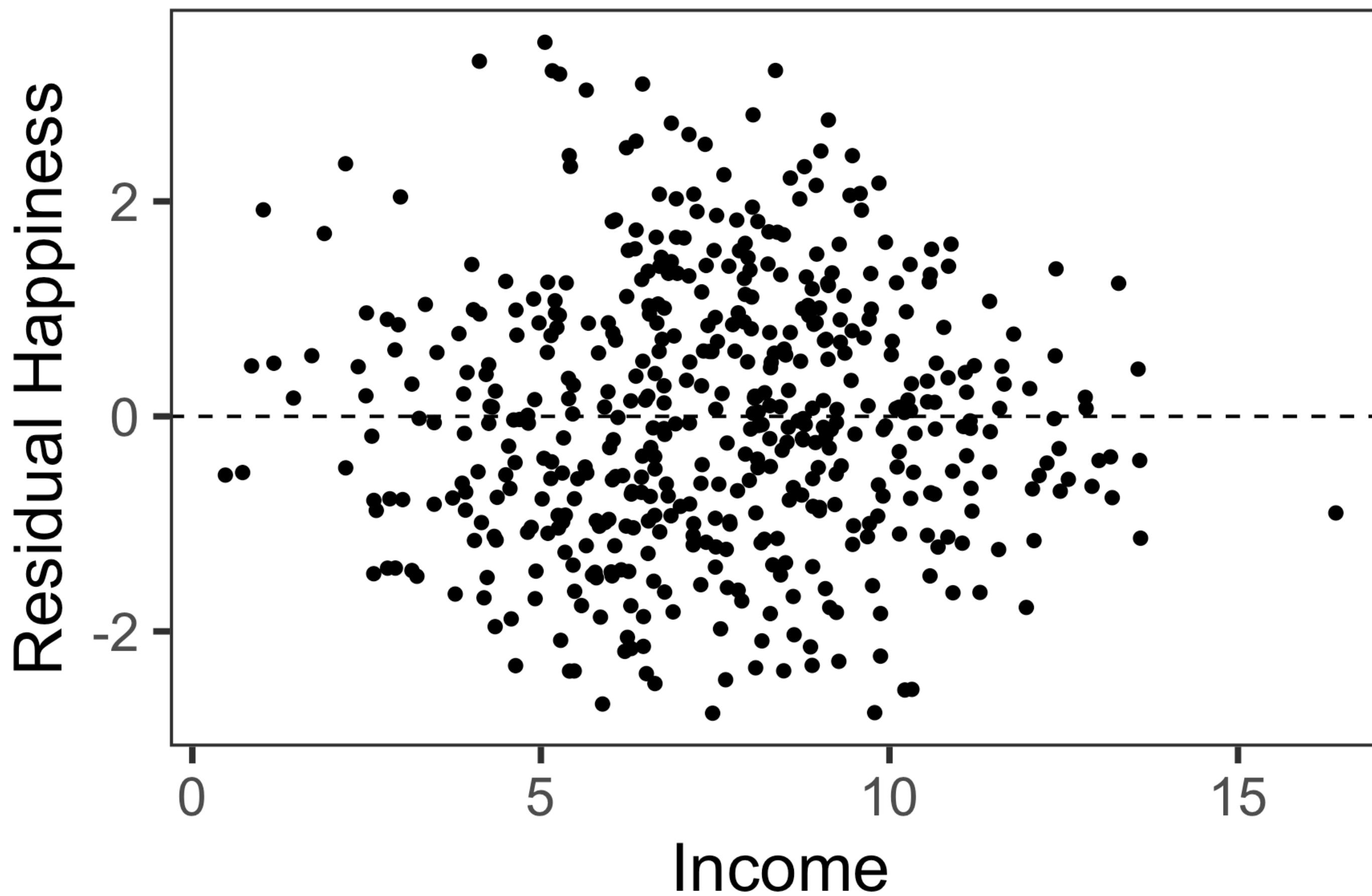
residual = observed - fitted

$$e_i = y_i - \hat{y}_i$$

- Mean of zero
- Equally positive as negative
- No remaining shape
- Low residual values

Residual plots to identify non-linearity

```
> resid(lm(Y ~ x))
```



residual = observed - fitted

$$e_i = y_i - \hat{y}_i$$

- Mean of zero
- Equally positive as negative
- No remaining shape
- Low residual values

Residual plots to identify non-linearity

```
> summary(lm(happiness ~ income, data = income_data))
```

Call:

```
lm(formula = happiness ~ income, data = income_data)
```

Residuals

Residuals:

Min	1Q	Median	3Q	Max
-3.5033	-0.8873	0.0459	0.9218	3.1279

Coefficients:

Can also see summary of residuals with `summary(lm(Y~X))`

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.34291	0.16505	8.136	3.3e-15 ***
income	0.22765	0.01727	13.182	< 2e-16 ***

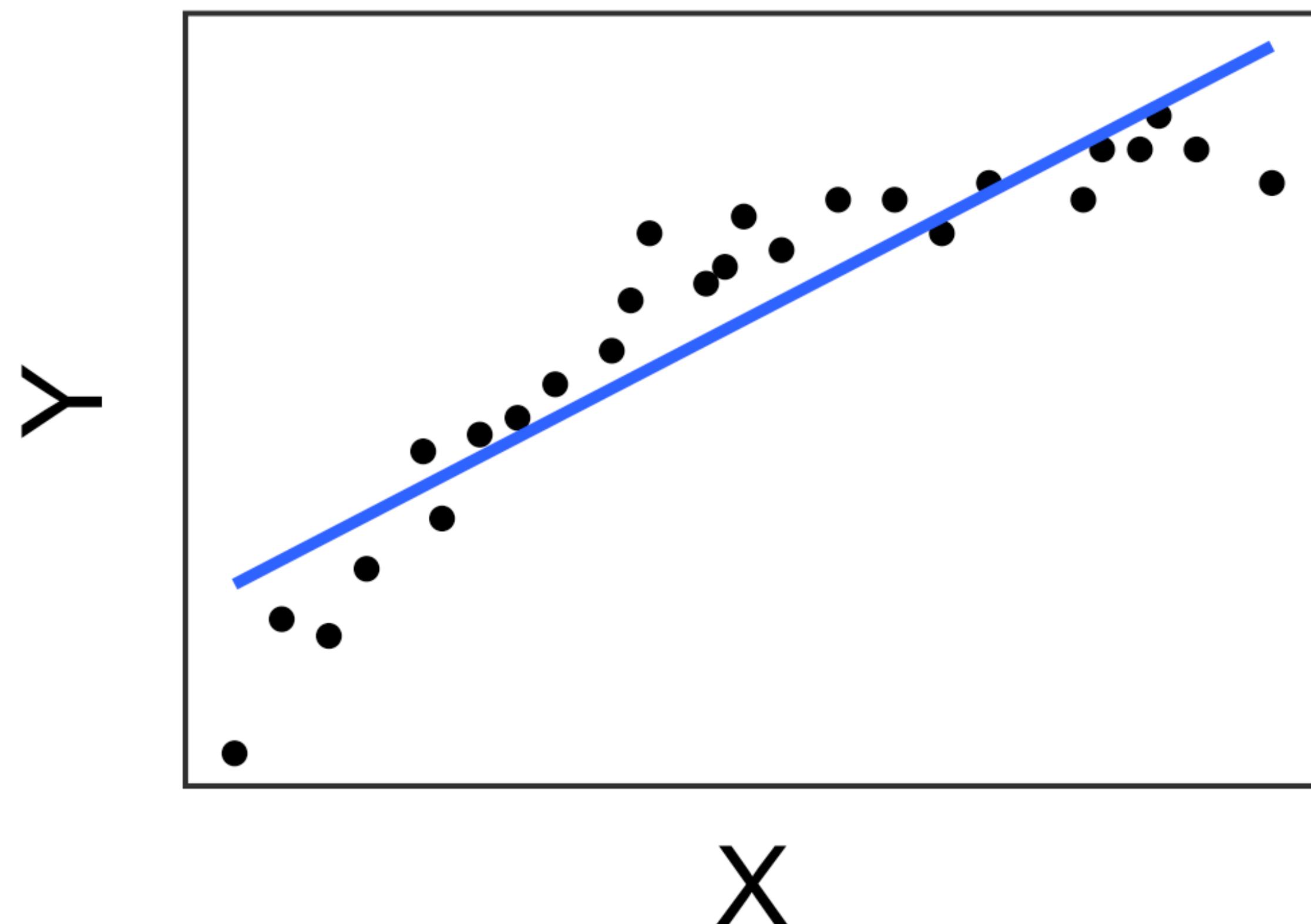
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.234 on 496 degrees of freedom

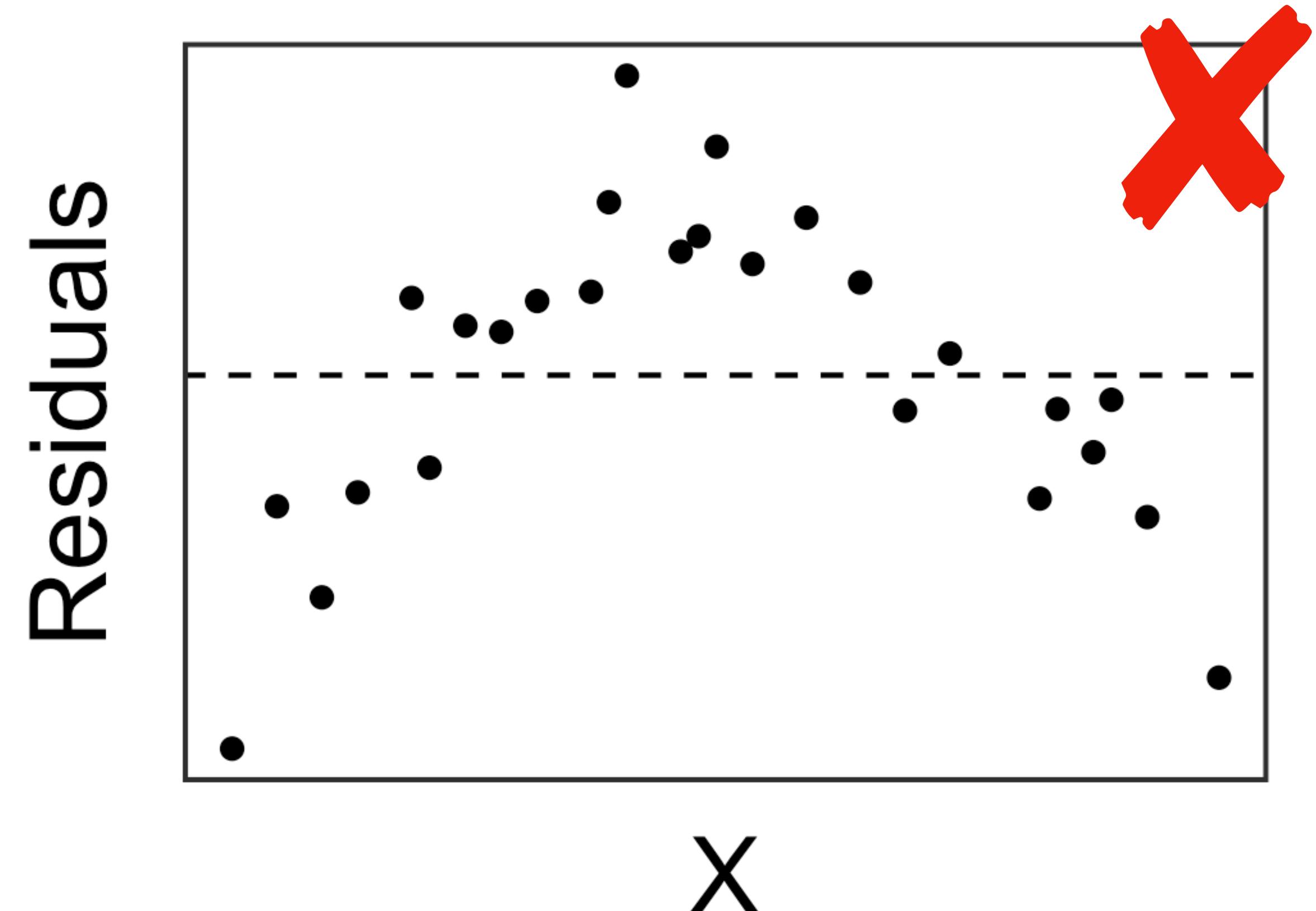
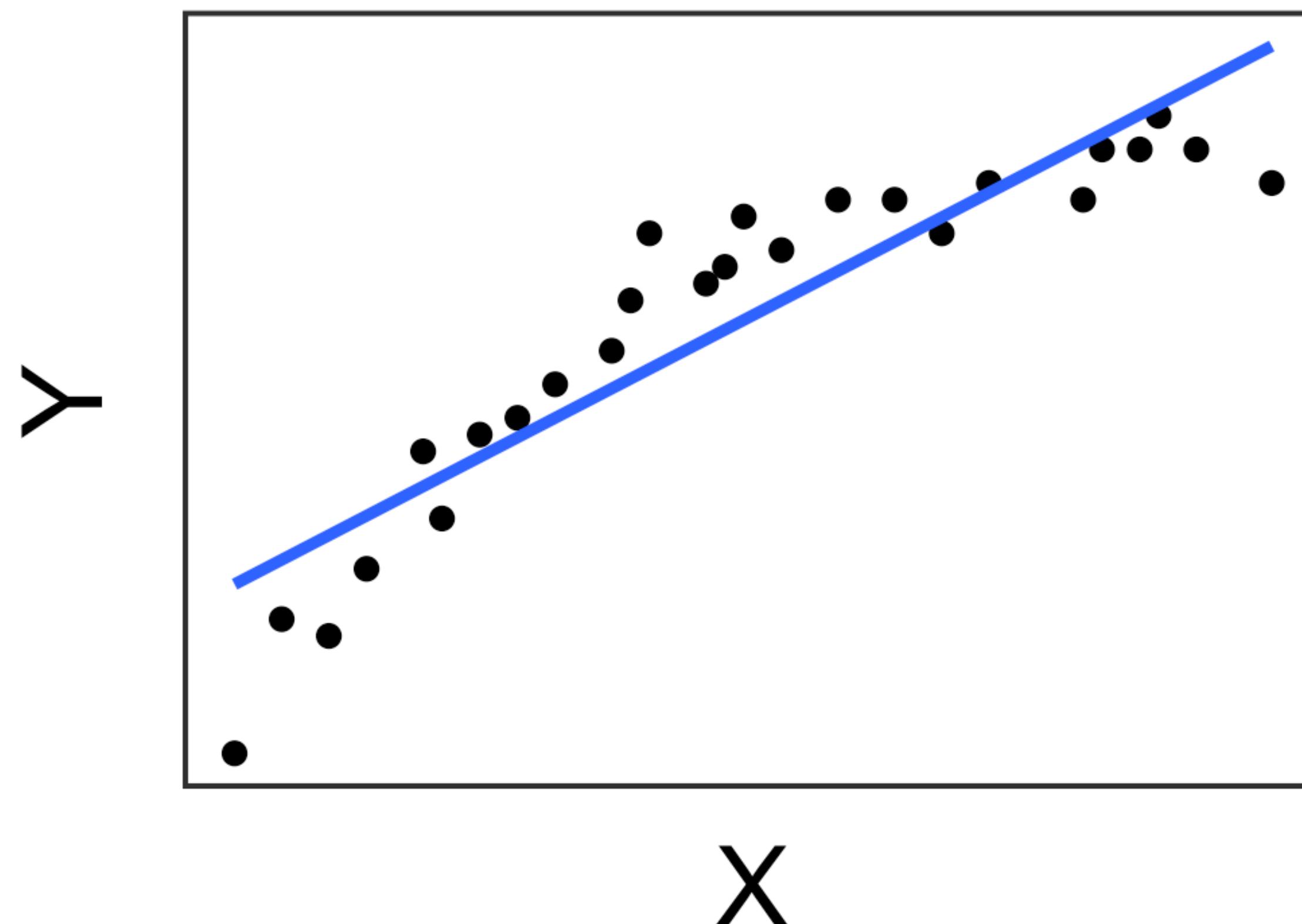
Multiple R-squared: 0.2595, Adjusted R-squared: 0.258

F-statistic: 173.8 on 1 and 496 DF, p-value: < 2.2e-16

Residual plots to identify non-linearity

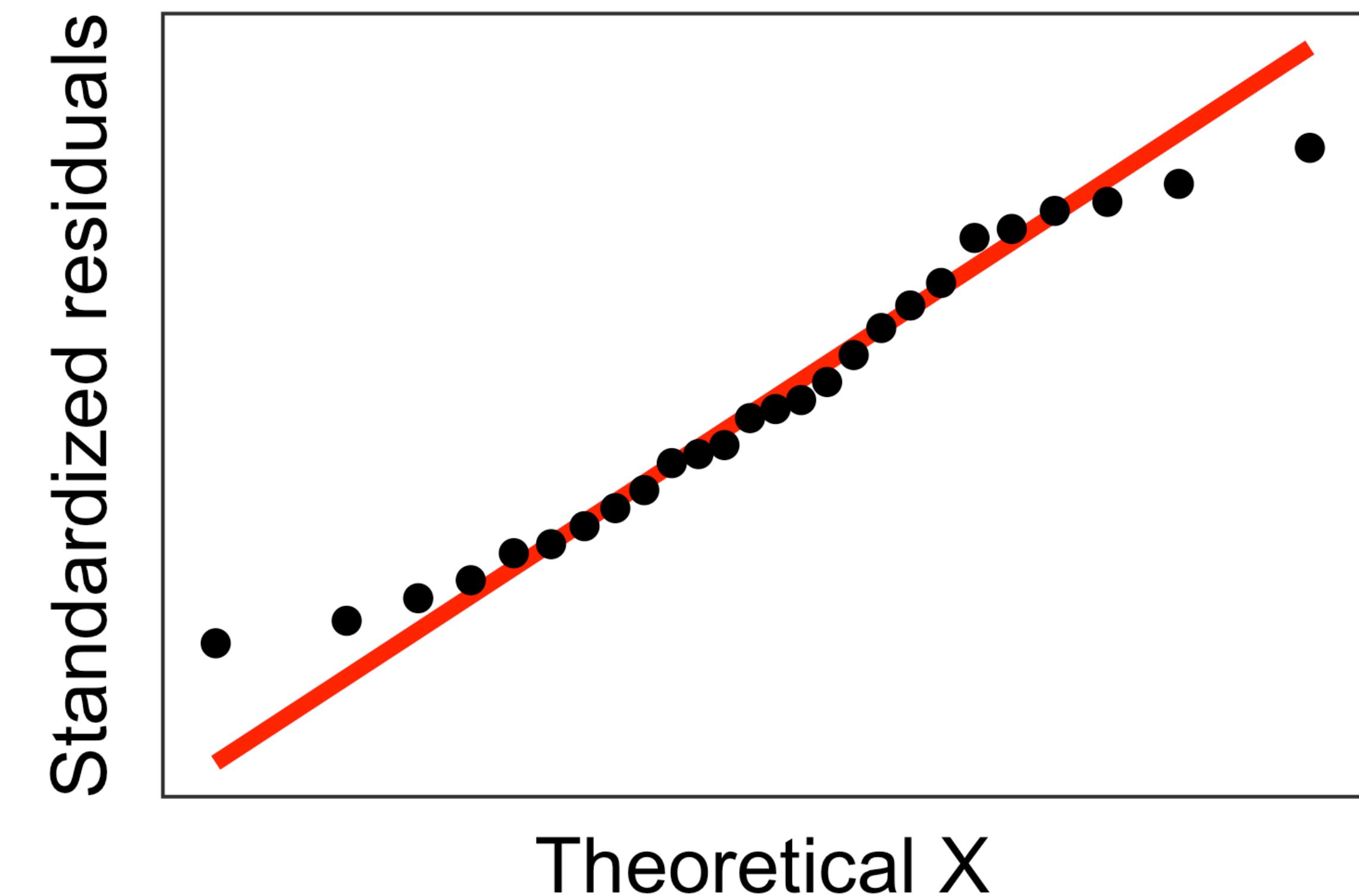
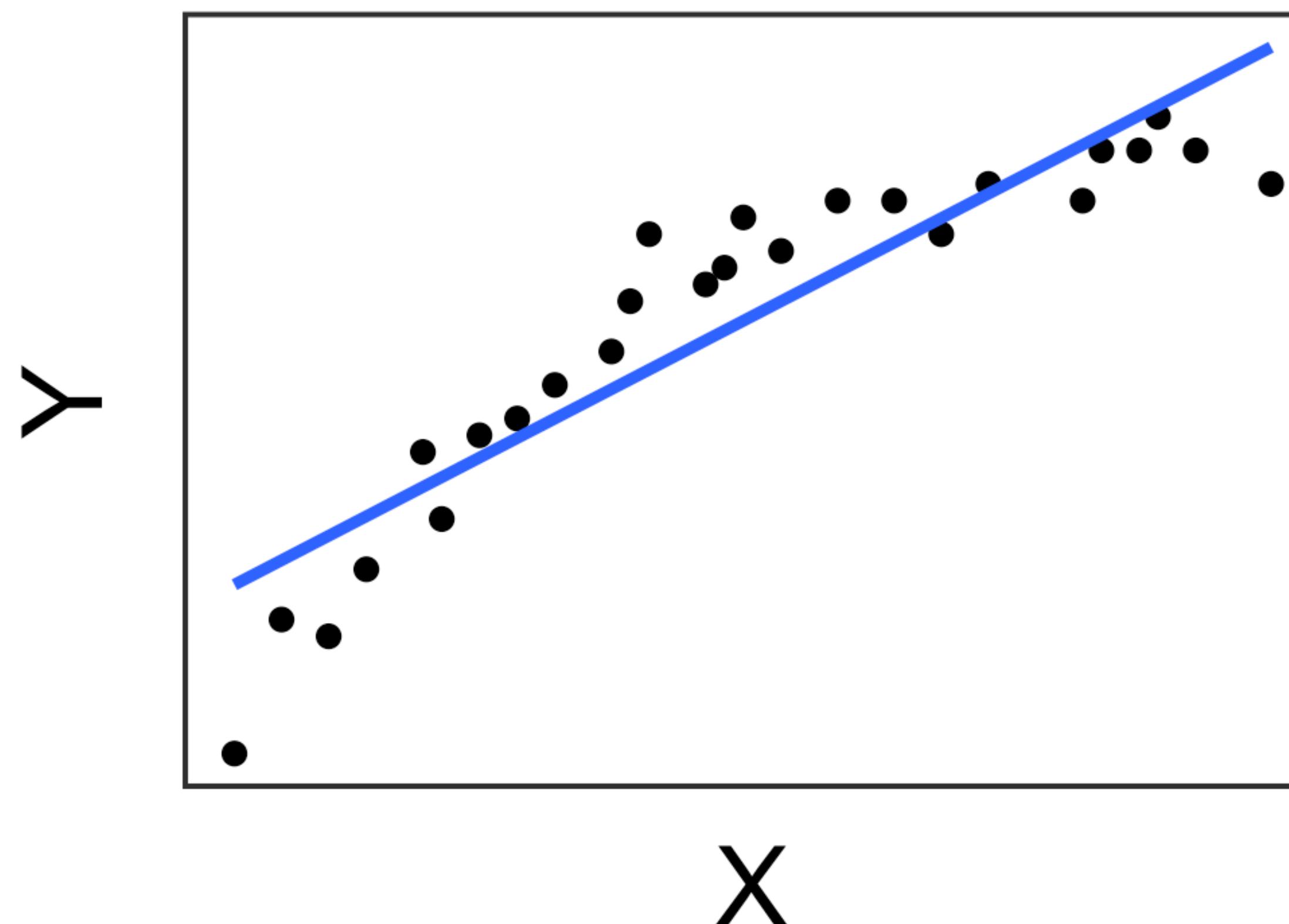


Residual plots to identify non-linearity



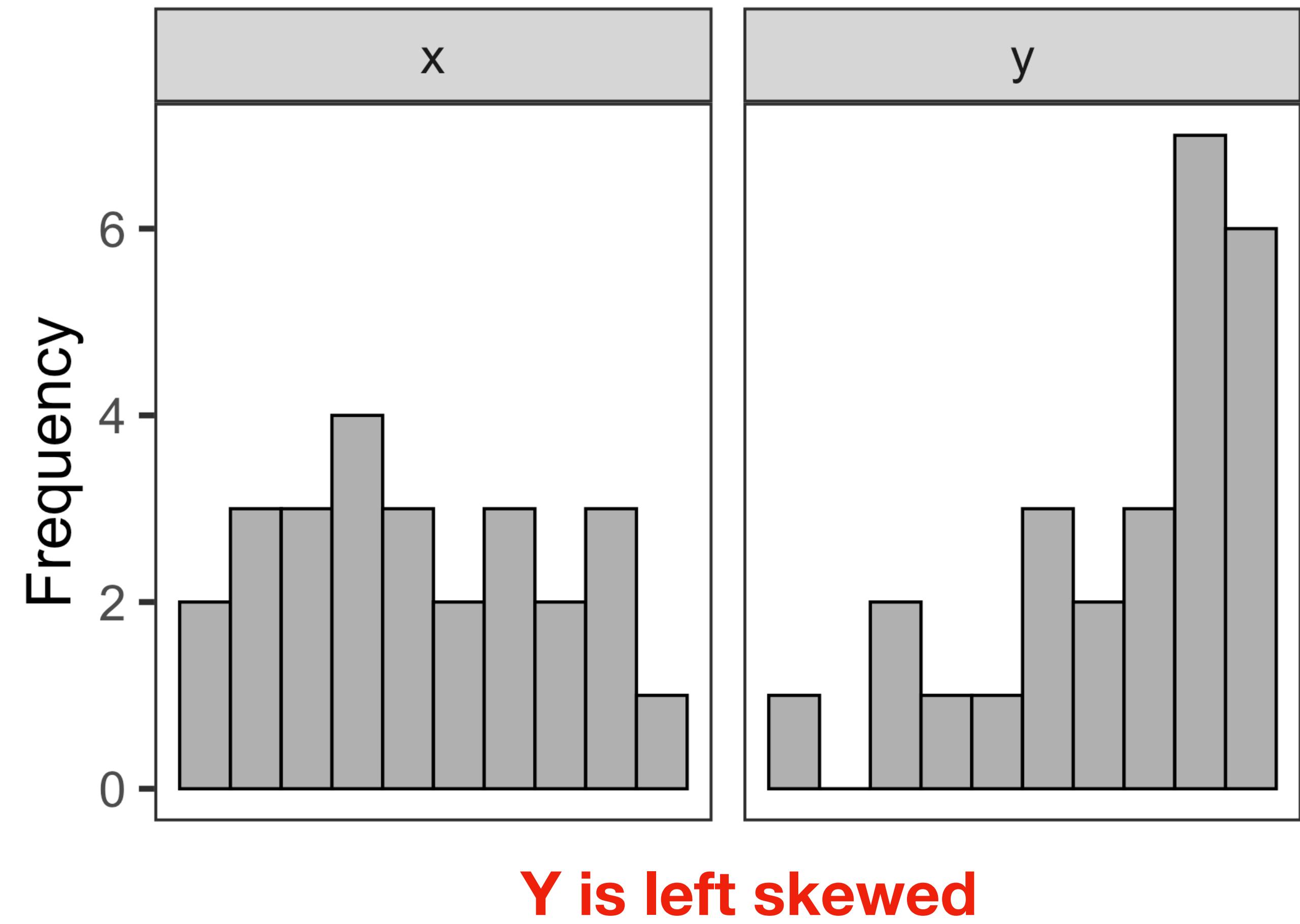
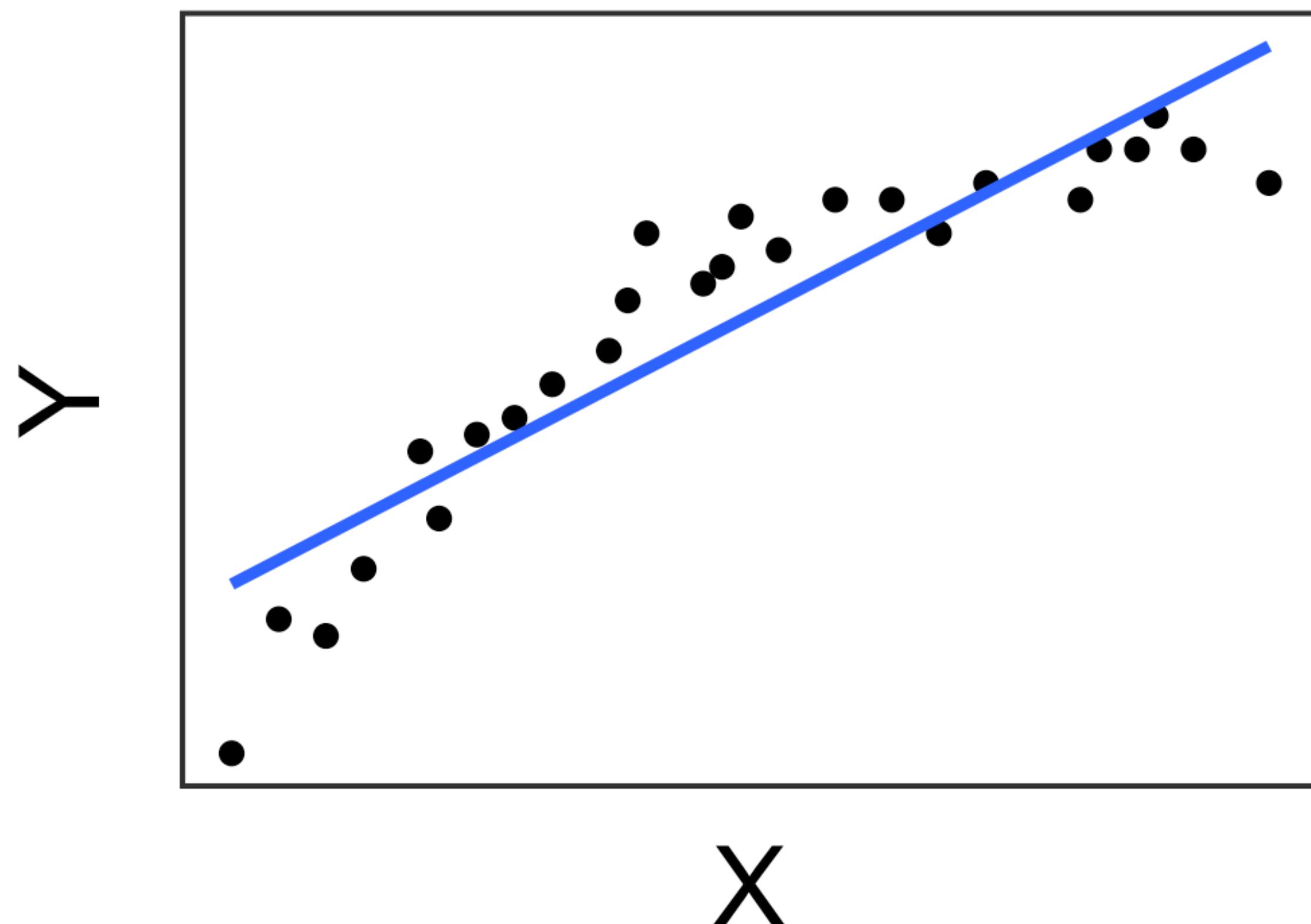
Data does not follow
a linear model

Q-Q plots to identify normality



**Residuals look
normally distributed**

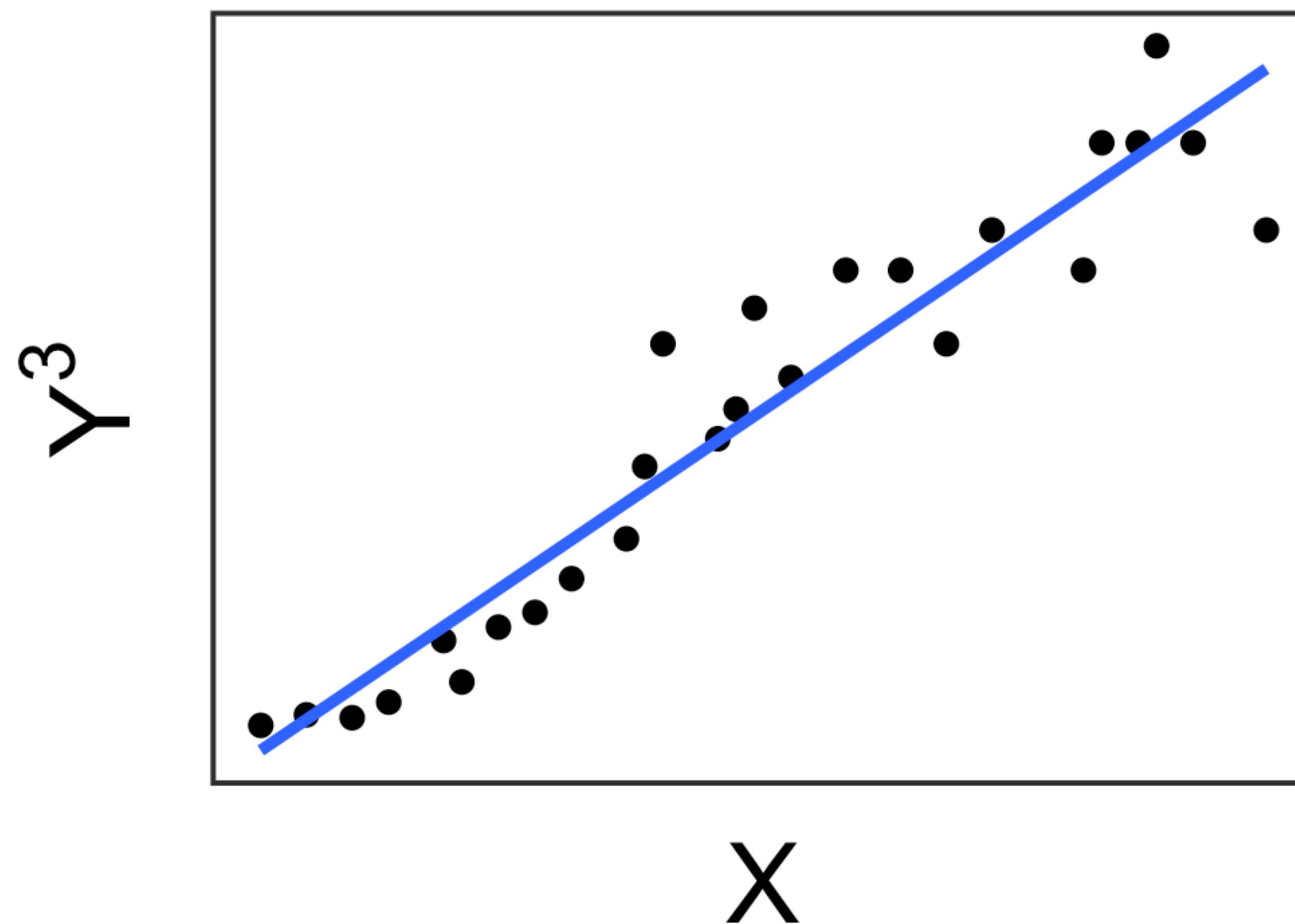
Residual plots to identify non-linearity



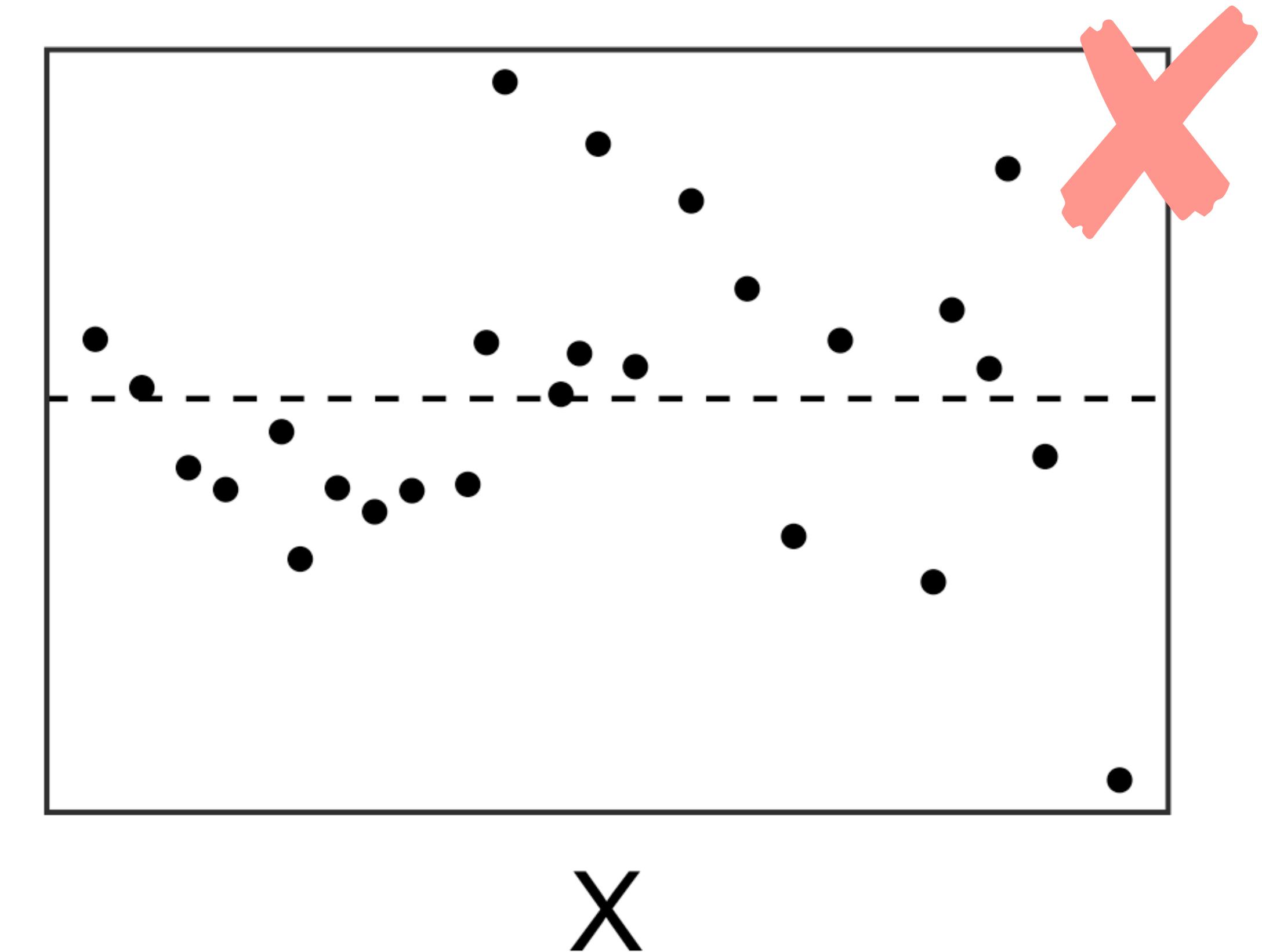
Y is left skewed

Could try Y^2

Residual plots to identify non-linearity

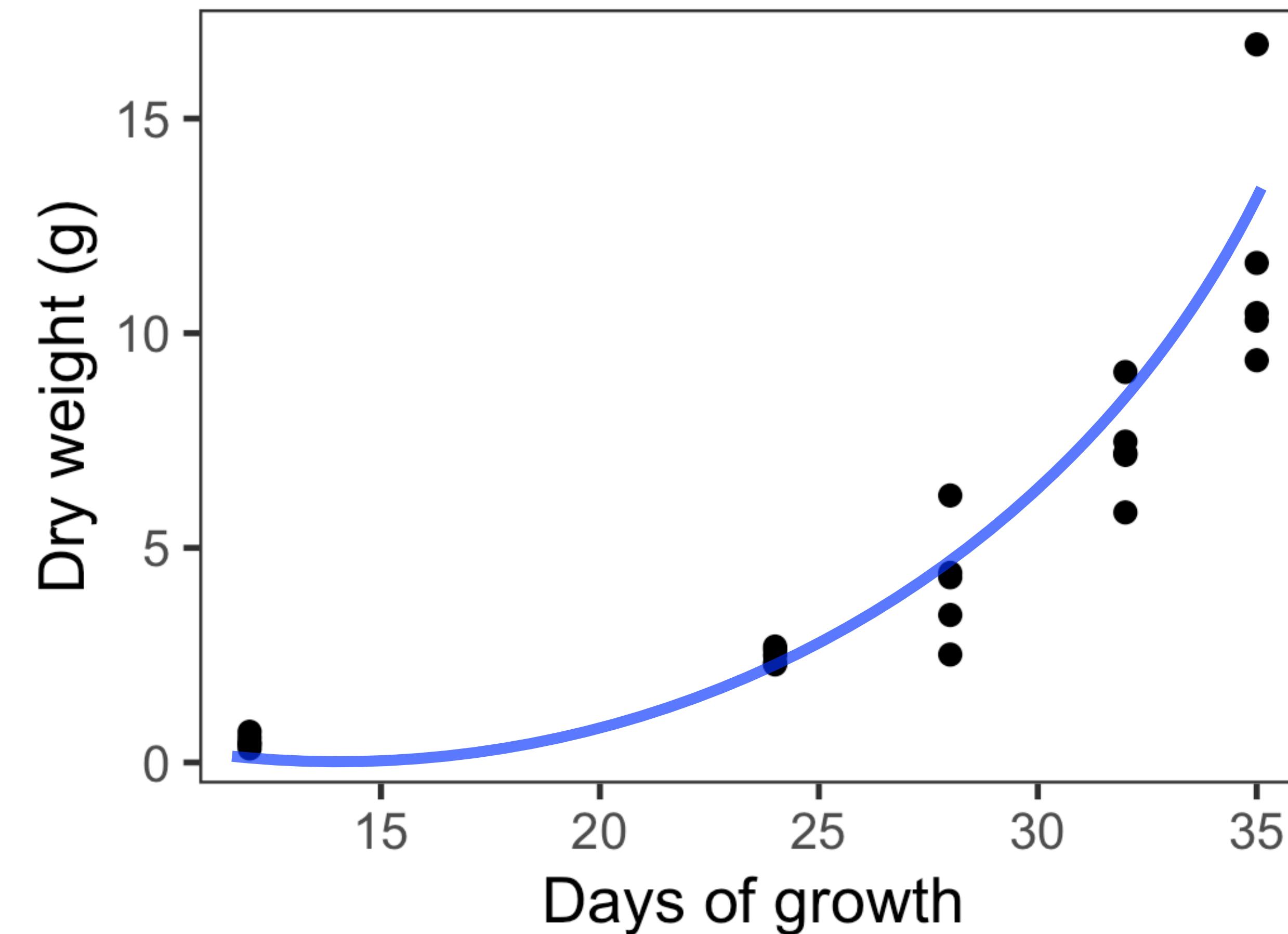


Residuals

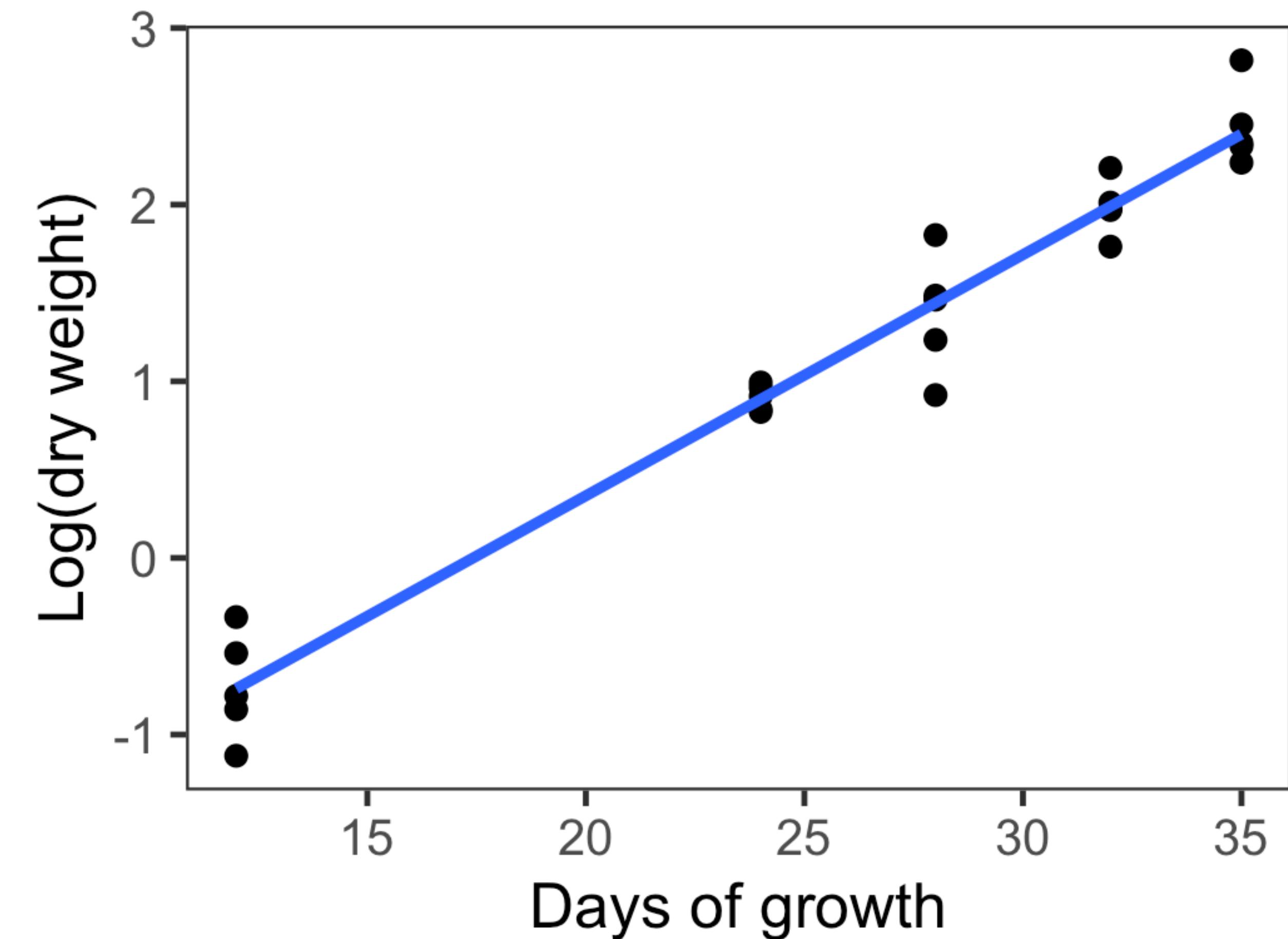
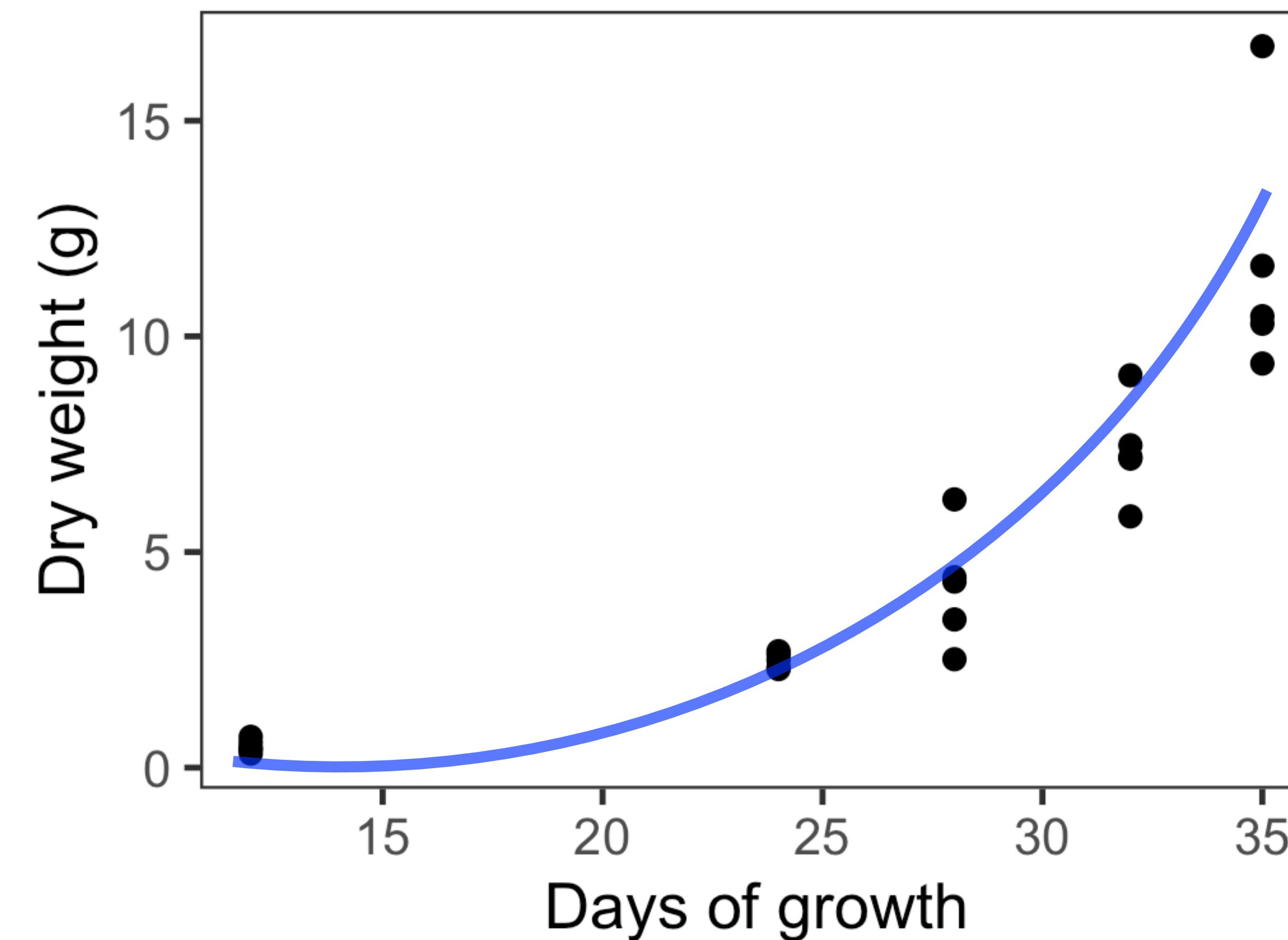


Not great, but much better!

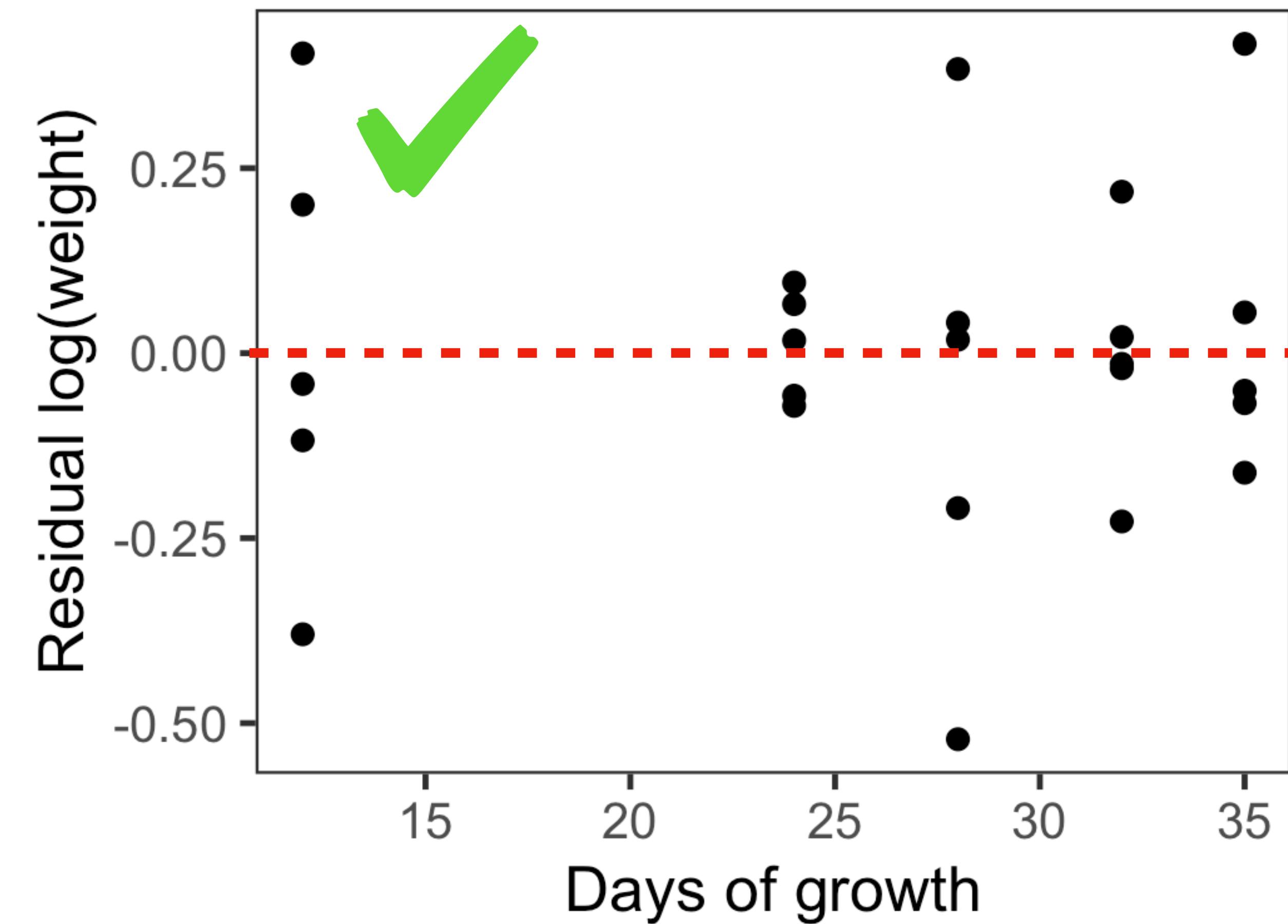
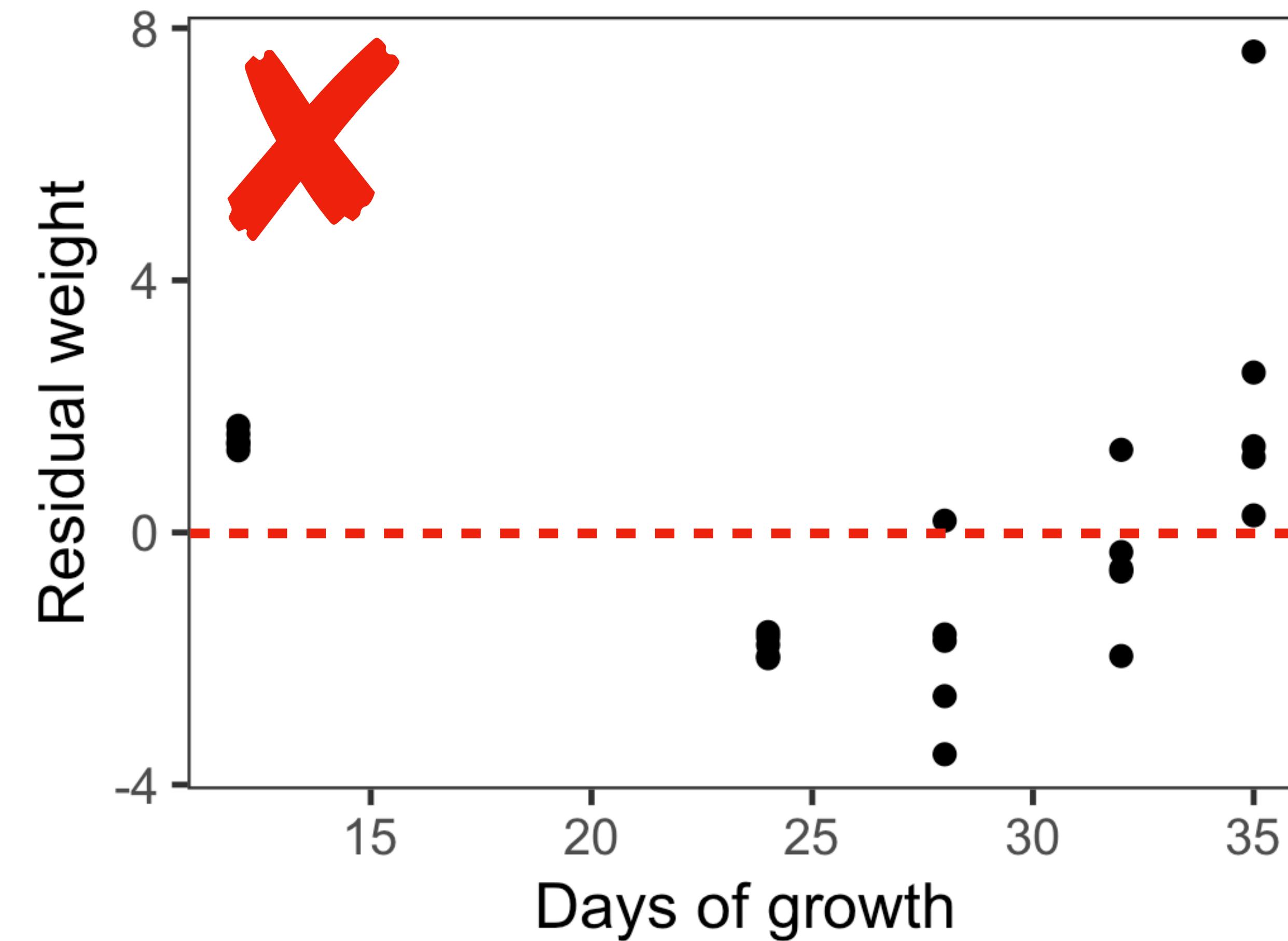
Transformations to obtain linearity



Transformations to obtain linearity



Transformations to obtain linearity



Transformations to obtain linearity

Great package for
tidy statistics!

broom::augment(model)

happiness	income	.fitted	.resid	.std.resid	.hat	.sigma	.cooksdi
2.3144890	3.862647	2.961527	-0.64703769	-0.90205658	0.002251376	0.7182356	9.180461e-04
3.4334898	4.979381	3.758680	-0.32519010	-0.45334273	0.002183070	0.7186765	2.248227e-04
4.5993734	4.923957	3.719116	0.88025693	1.22713119	0.002147257	0.7177335	1.620203e-03
2.7911138	3.214372	2.498771	0.29234235	0.40772807	0.003053610	0.7187050	2.545969e-04
5.5963983	7.196409	5.341251	0.25514736	0.35655407	0.006973369	0.7187334	4.463777e-04
2.4585559	3.729643	2.866585	-0.40802919	-0.56888086	0.002370292	0.7185909	3.844547e-04
3.1929918	4.674517	3.541060	-0.34806836	-0.48520140	0.002036760	0.7186549	2.402367e-04
1.9071368	4.498104	3.415132	-1.50799483	-2.10208970	0.002008681	0.7156164	4.446893e-03
2.9424499	3.121631	2.432570	0.50987997	0.71118370	0.003214180	0.7184589	8.154585e-04
3.7379416	4.639914	3.516360	0.22158191	0.30888007	0.002027982	0.7187563	9.693831e-05
3.1754061	4.632840	3.511309	-0.33590329	-0.46824107	0.002026383	0.7186666	2.225930e-04
2.0090465	2.773179	2.183836	-0.17478978	-0.24388421	0.003919933	0.7187824	1.170366e-04
5.9518141	7.119479	5.286336	0.66547825	0.92983921	0.006697419	0.7181987	2.914819e-03
5.9605473	7.466653	5.534158	0.42638937	0.59616497	0.008005262	0.7185679	1.434066e-03

Transformations to obtain linearity

Great package for
tidy statistics!

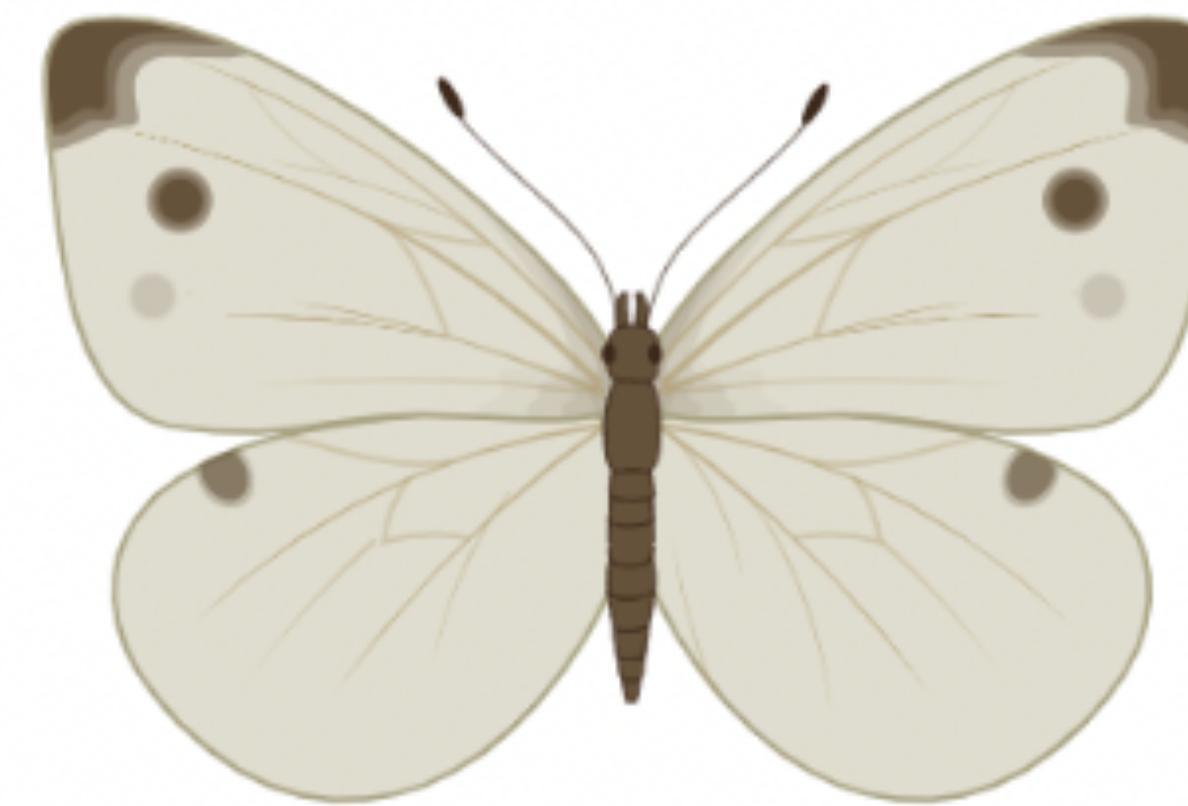
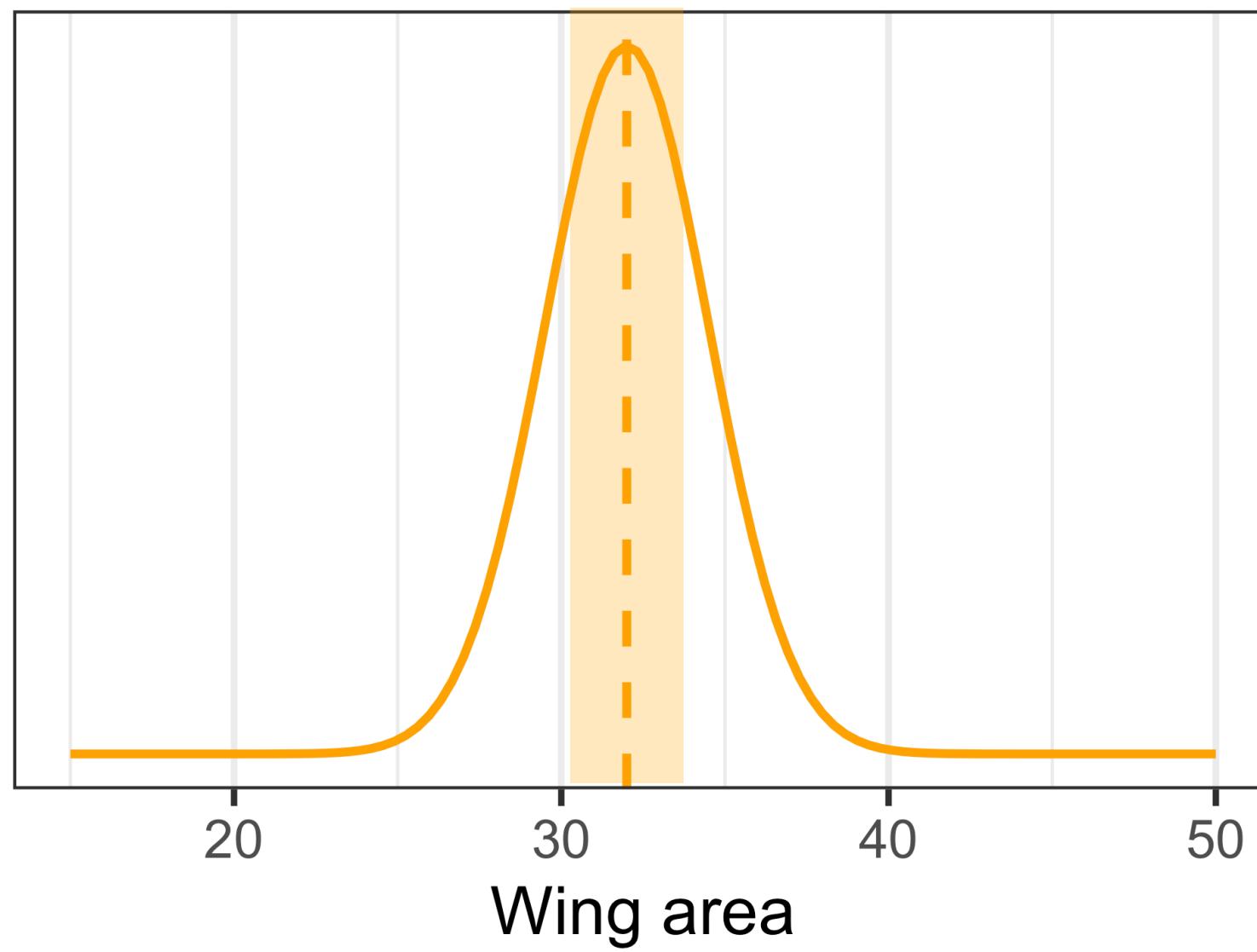
broom::augment(model)

happiness	income	.fitted	.resid	.std.resid	.hat	.sigma	.cooksdi
2.3144890	3.862647	2.961527	-0.64703769	-0.90205658	0.002251376	0.7182356	9.180461e-04
3.4334898	4.979381	3.758680	-0.32519010	-0.45334273	0.002183070	0.7186765	2.248227e-04
4.5993734	4.923957	3.719116	0.88025693	1.22713119	0.002147257	0.7177335	1.620203e-03
2.7911138	3.214372	2.498771	0.29234235	0.40772807	0.003053610	0.7187050	2.545969e-04
5.5963983	7.196409	5.341251	0.25514736	0.35655407	0.006973369	0.7187334	4.463777e-04
2.4585559	3.729643	2.866585	-0.40802919	-0.56888086	0.002370292	0.7185909	3.844547e-04
3.1929918	4.674517	3.541060	-0.34806836	-0.48520140	0.002036760	0.7186549	2.402367e-04
1.9071368	4.498104	3.415132	-1.50799483	-2.10208970	0.002008681	0.7156164	4.446893e-03
2.9424499	3.121631	2.432570	0.50987997	0.71118370	0.003214180	0.7184589	8.154585e-04
3.7379416	4.639914	3.516360	0.22158191	0.30888007	0.002027982	0.7187563	9.693831e-05
3.1754061	4.632840	3.511309	-0.33590329	-0.46824107	0.002026383	0.7186666	2.225930e-04
2.0090465	2.773179	2.183836	-0.17478978	-0.24388421	0.003919933	0.7187824	1.170366e-04
5.9518141	7.119479	5.286336	0.66547825	0.92983921	0.006697419	0.7181987	2.914819e-03
5.9605473	7.466653	5.534158	0.42638937	0.59616497	0.008005262	0.7185679	1.434066e-03

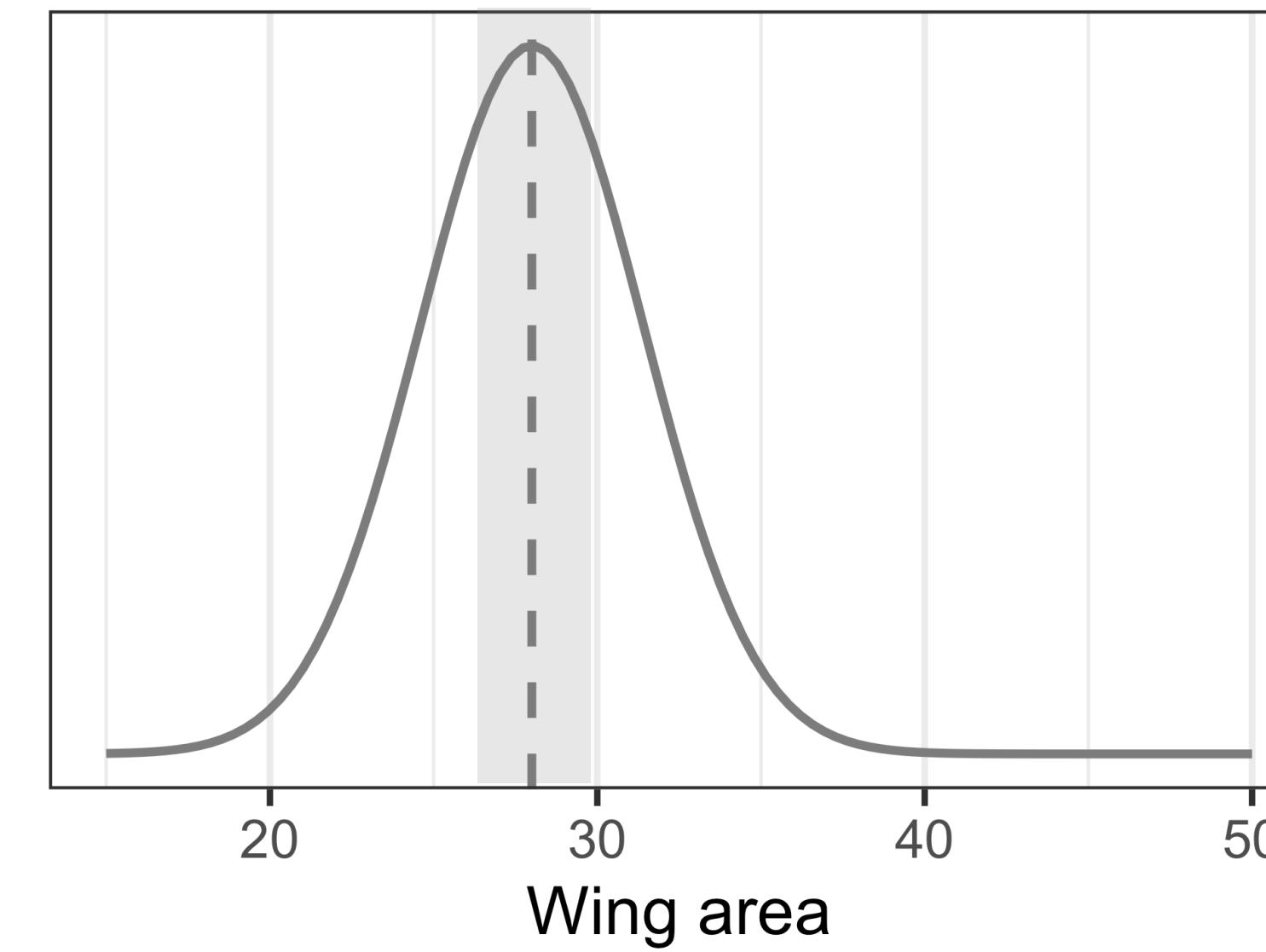
Regression and the *t*-test



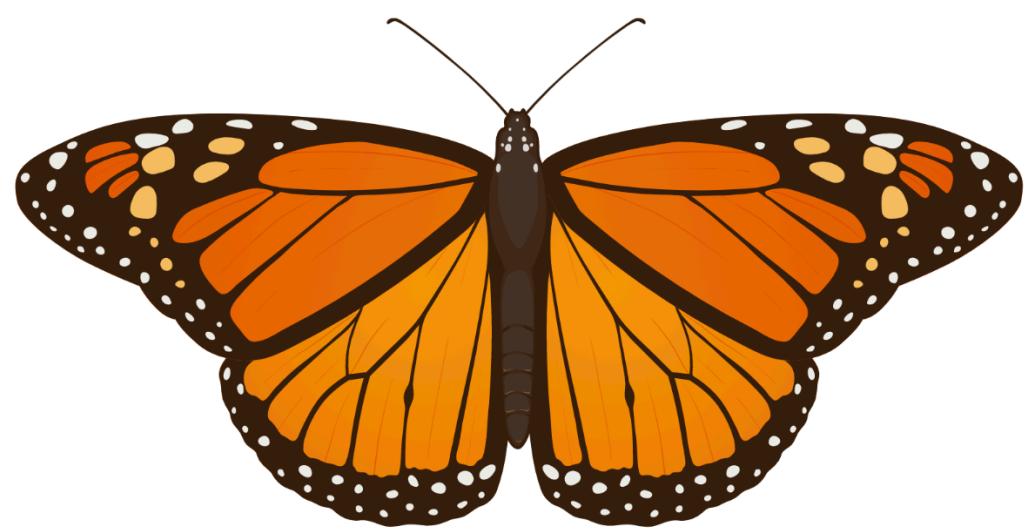
$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$



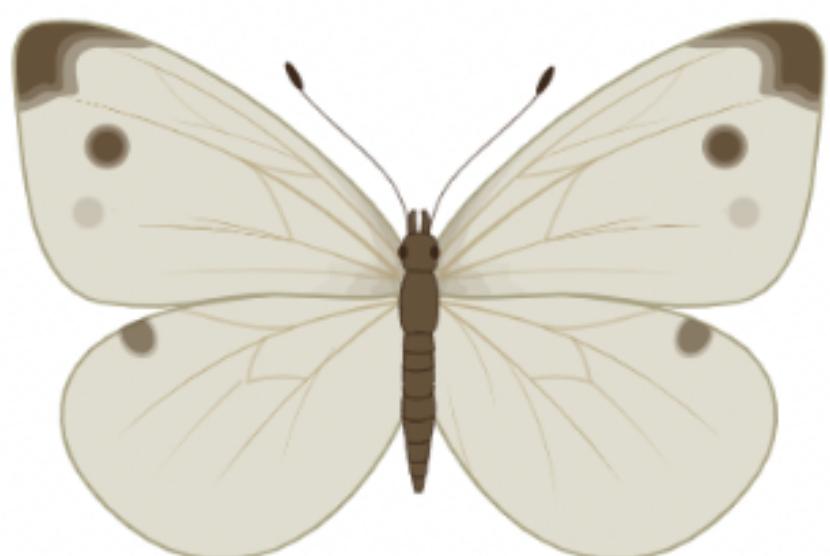
$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$



Regression and the *t*-test



$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$



$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$

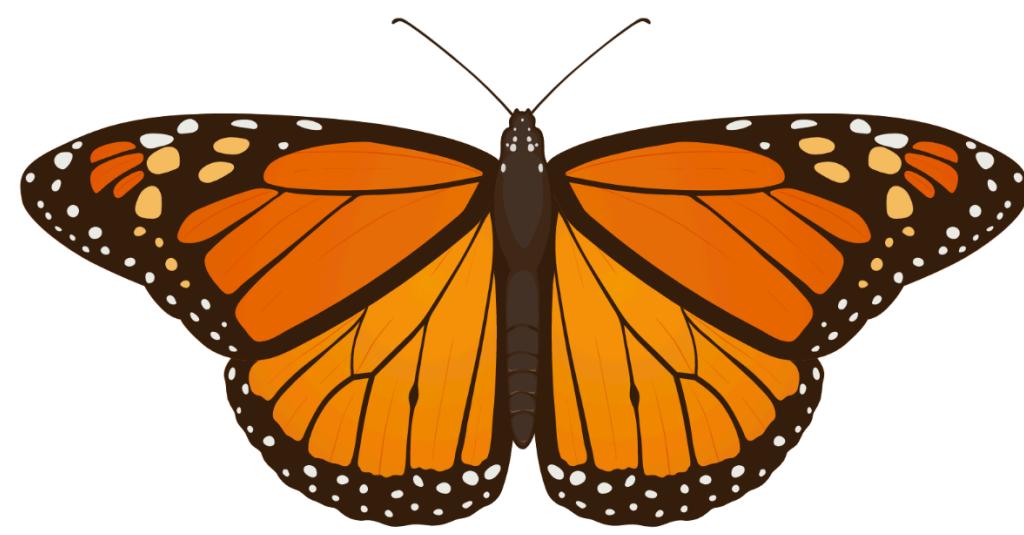
```
# make populations for butterfly  
y1 <- rnorm(14, 32, 2.5)  
y2 <- rnorm(12, 28, 3.4)  
  
# calculate t test with t.test  
t.test(y1, y2)  
  
Welch Two Sample t-test
```

```
data: y1 and y2  
t = 4.2051, df = 23.351, p-value = 0.0003288  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
 1.646552 4.829972  
sample estimates:  
mean of x mean of y  
31.37680 28.13854
```

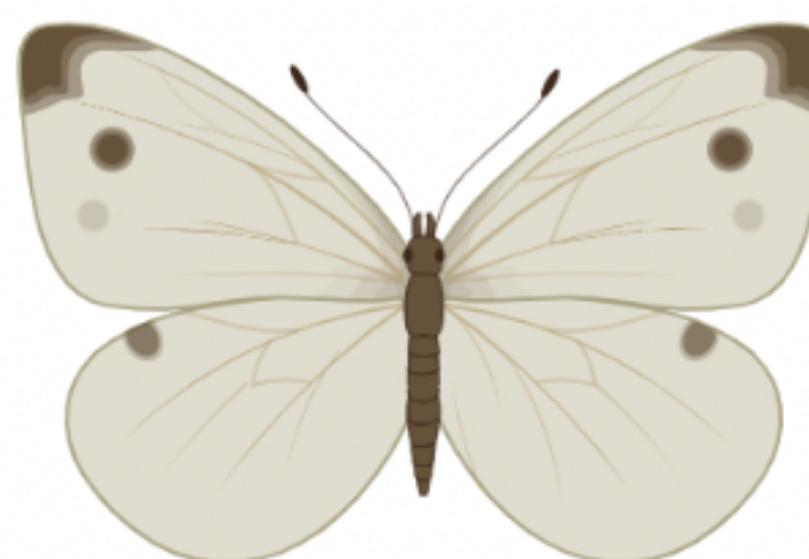
$$H_0 : \mu_1 = \mu_2$$

$$H_A : \mu_1 \neq \mu_2$$

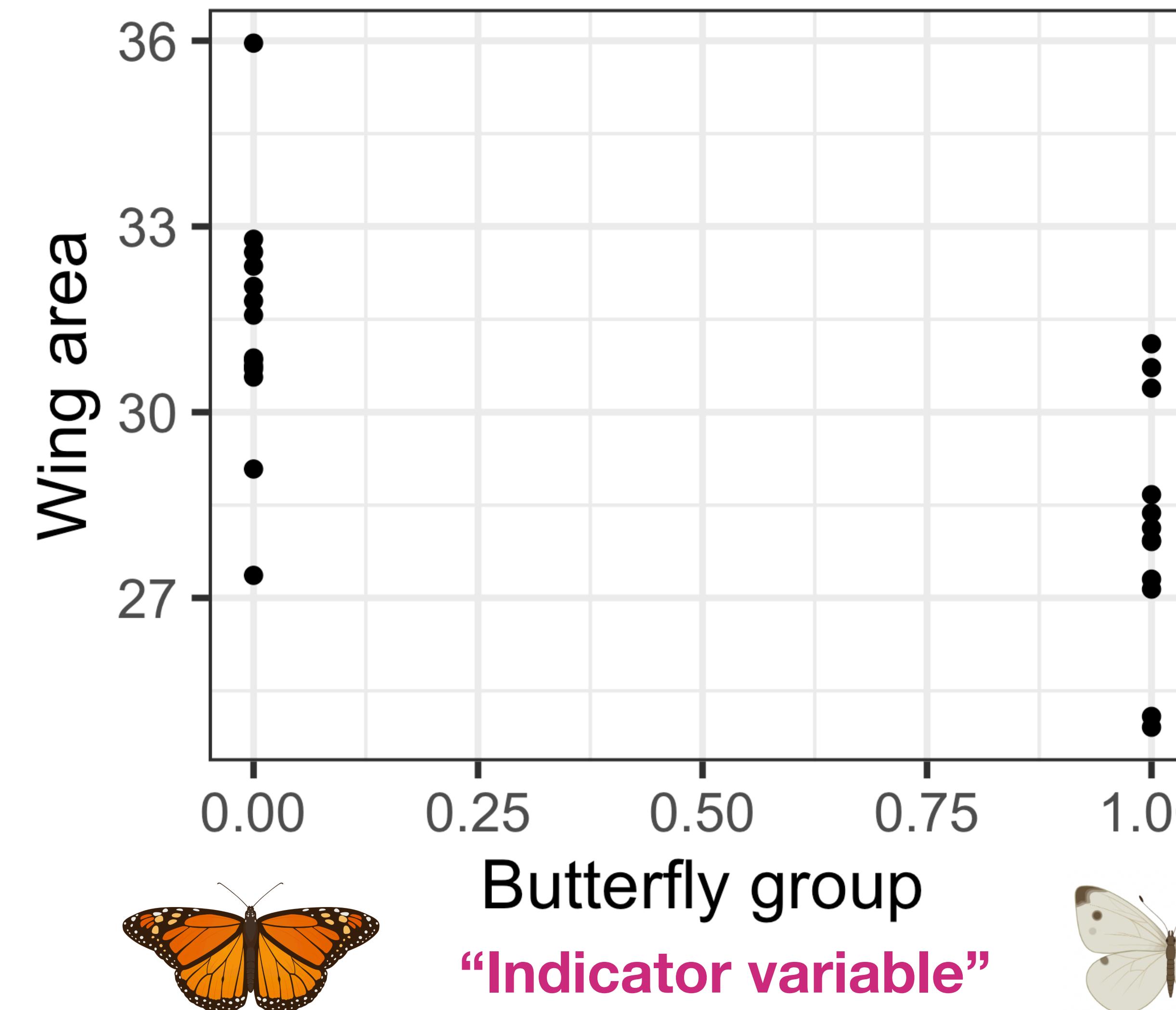
Regression and the *t*-test



$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$



$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$

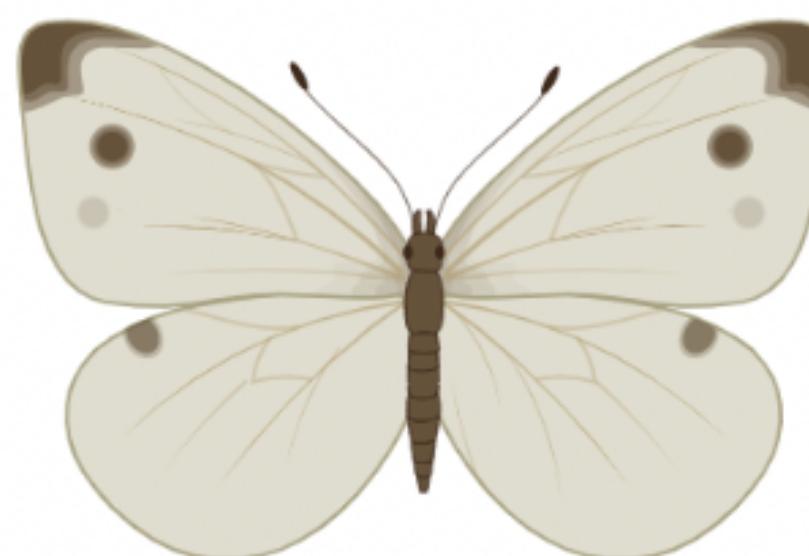


butterfly	area
0	32.79386
1	27.14261
0	31.56760
1	24.91155
0	31.79648
1	30.38884
0	30.69628
1	27.91059

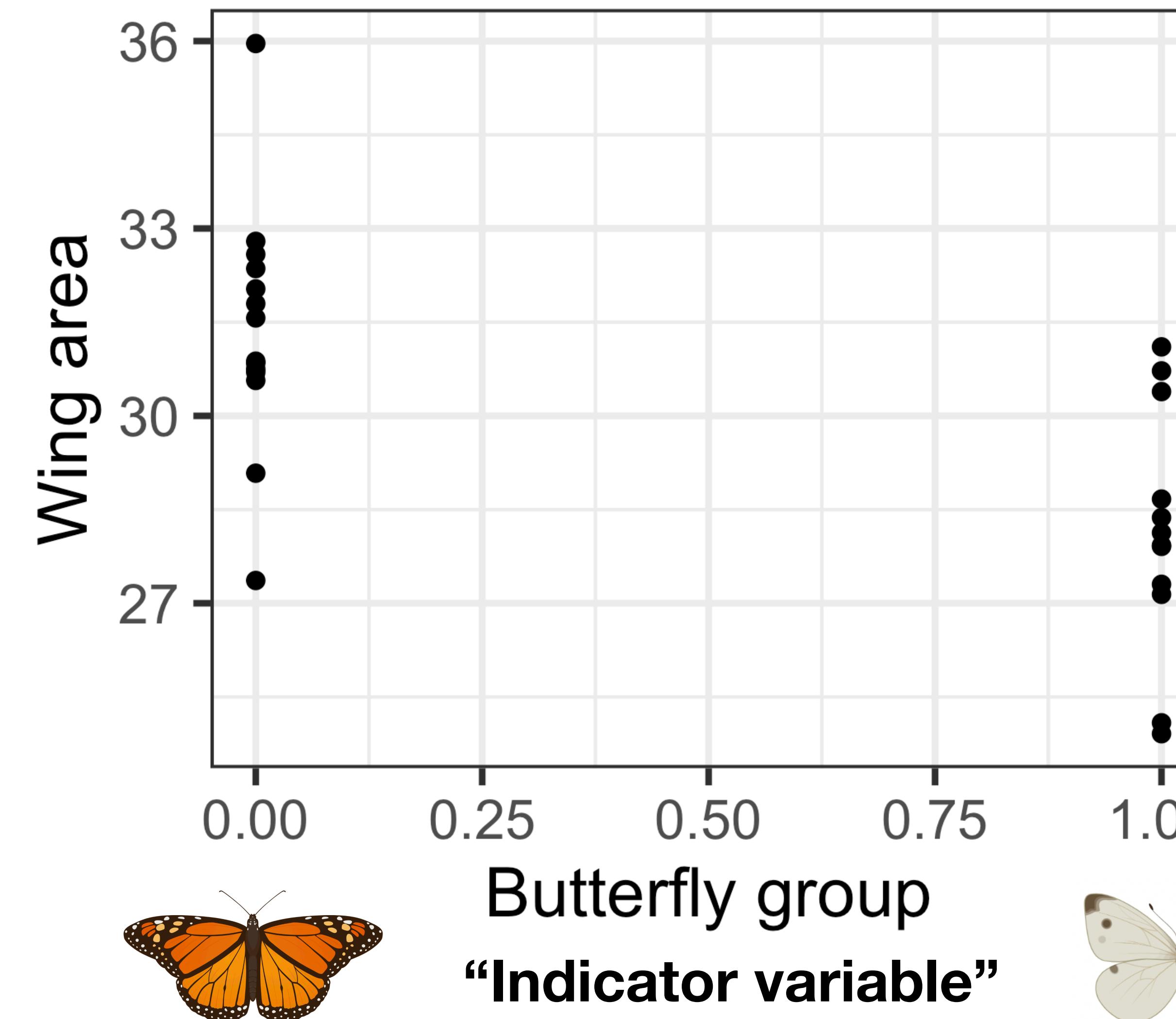
Regression and the *t*-test



$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$

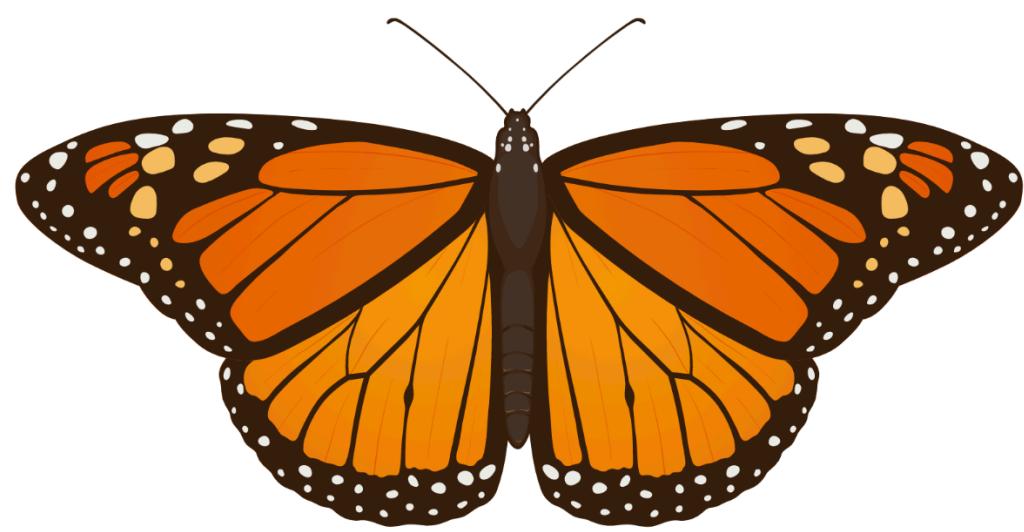


$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$

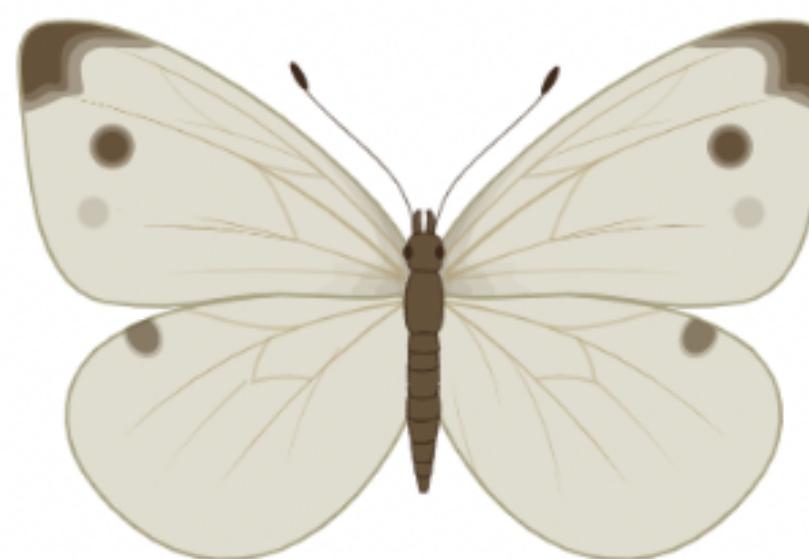


$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

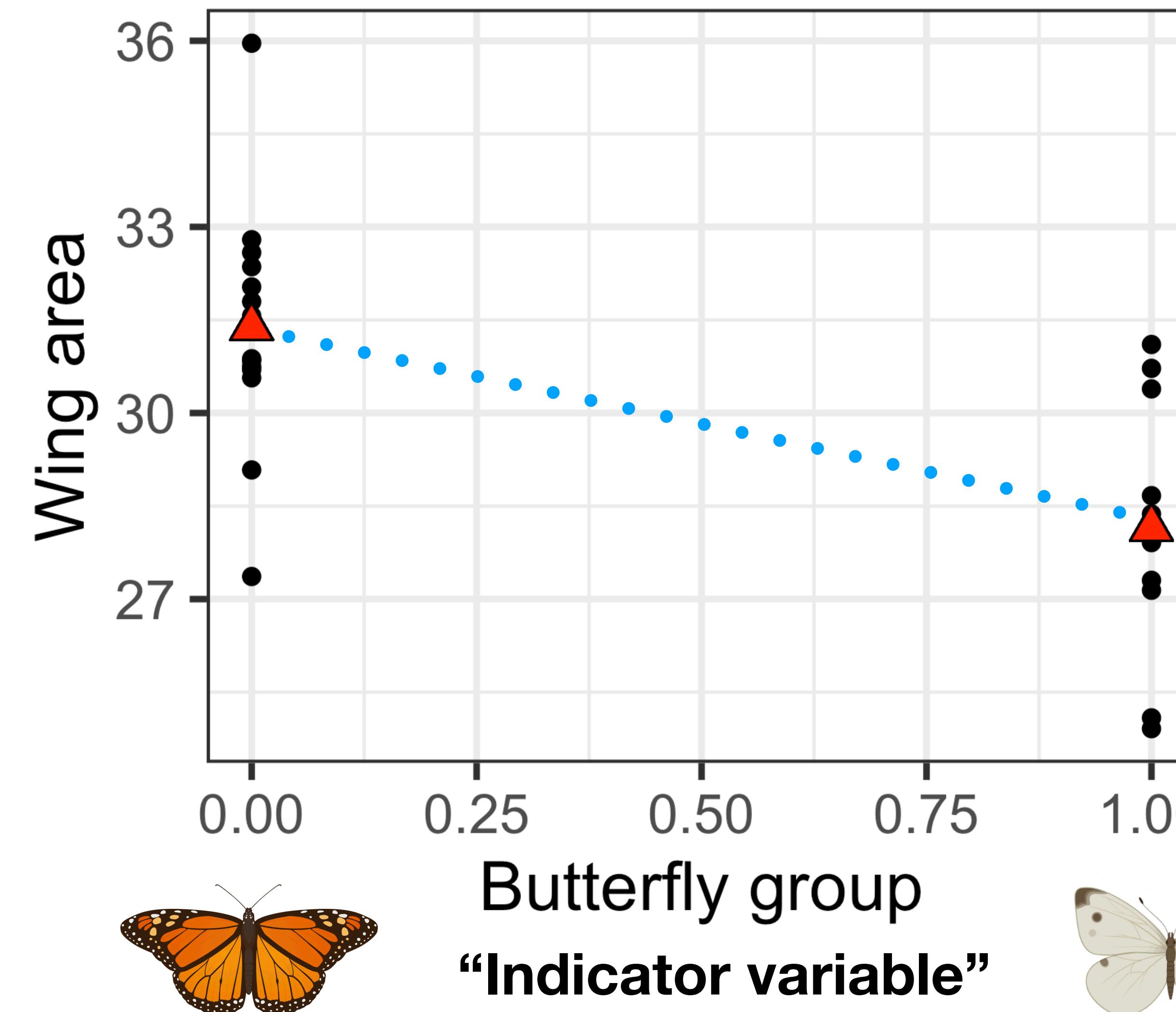
Regression and the *t*-test



$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$



$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$



$$H_0 : \beta_1 = 0$$

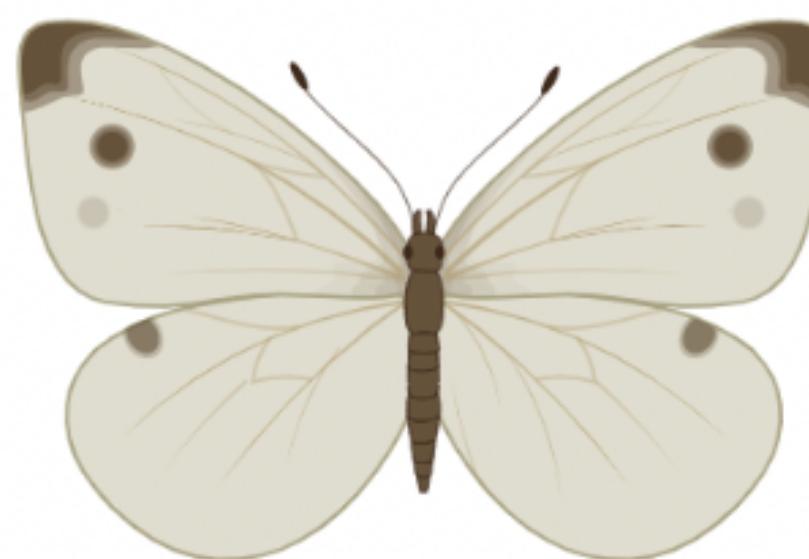
$$H_A : \beta_1 \neq 0$$

Difference between group means = β_1 !
 $(H_0 : \mu_1 - \mu_2 = 0)$

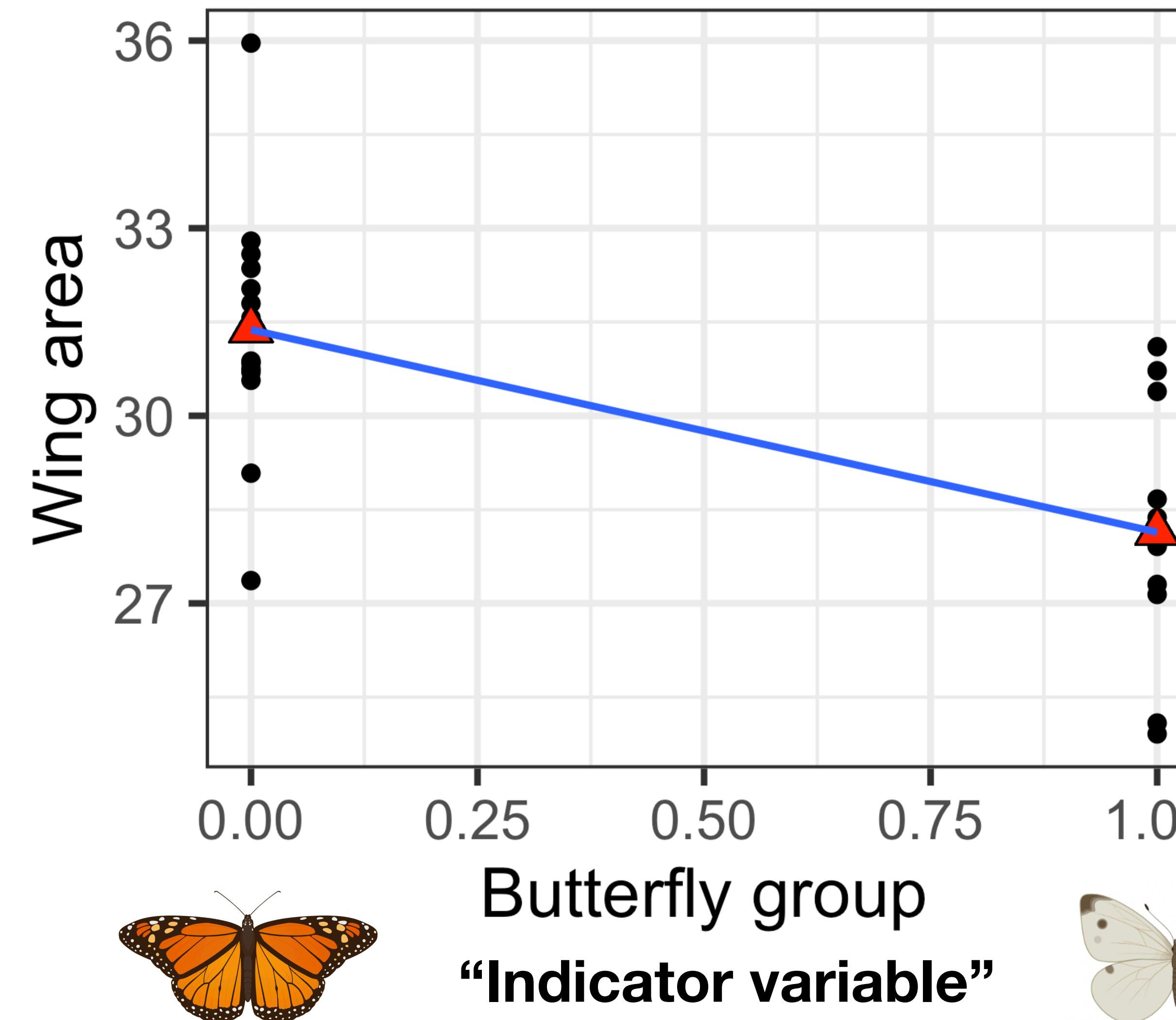
Regression and the *t*-test



$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$



$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$



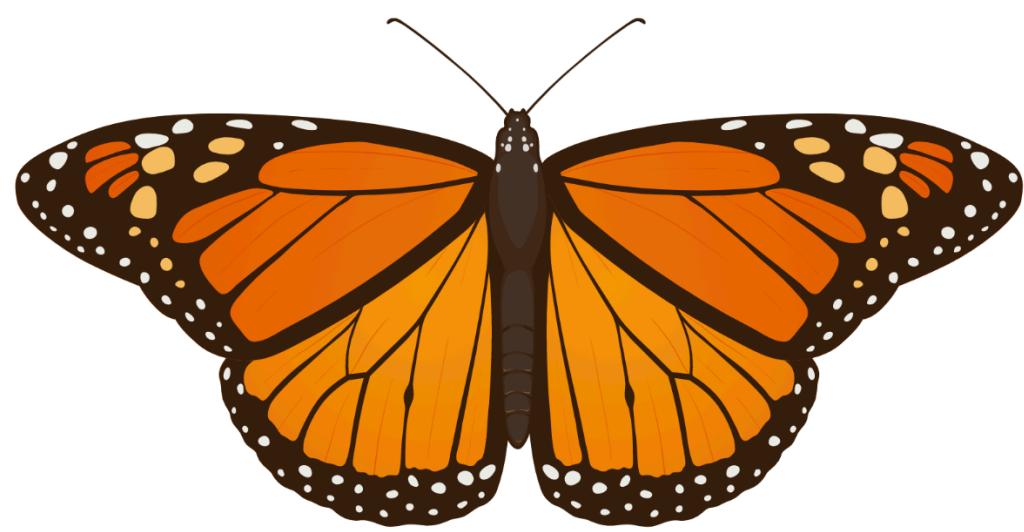
$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 \neq 0$$

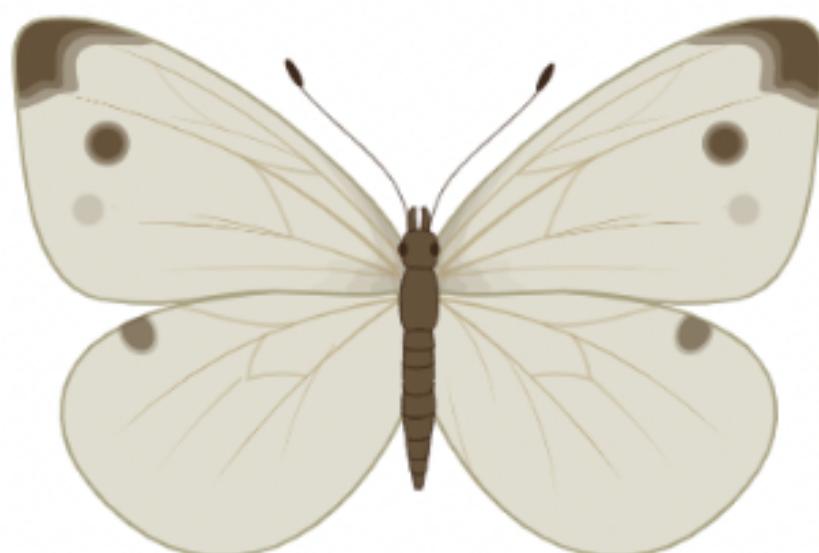
Difference between group means = β_1 !

$$(H_0 : \mu_1 - \mu_2 = 0)$$

Regression and the t -test



$$\bar{y}_1 = 32$$
$$s_1 = 2.5$$



$$\bar{y}_2 = 28$$
$$s_2 = 3.4$$

all:
m(formula = area ~ butterfly, data = butterfly)

Residuals:

Min	1Q	Median	3Q	Max
-4.0121	-0.7773	-0.1108	0.9000	4.5842

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	31.3768	0.5229	60.003	< 2e-16 ***
butterfly	-3.2383	0.7697	-4.207	0.000312 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.957 on 24 degrees of freedom

Multiple R-squared: 0.4245, Adjusted R-squared: 0.4005

F-statistic: 17.7 on 1 and 24 DF, p-value: 0.0003119

$$H_0 : \beta_1 = 0$$

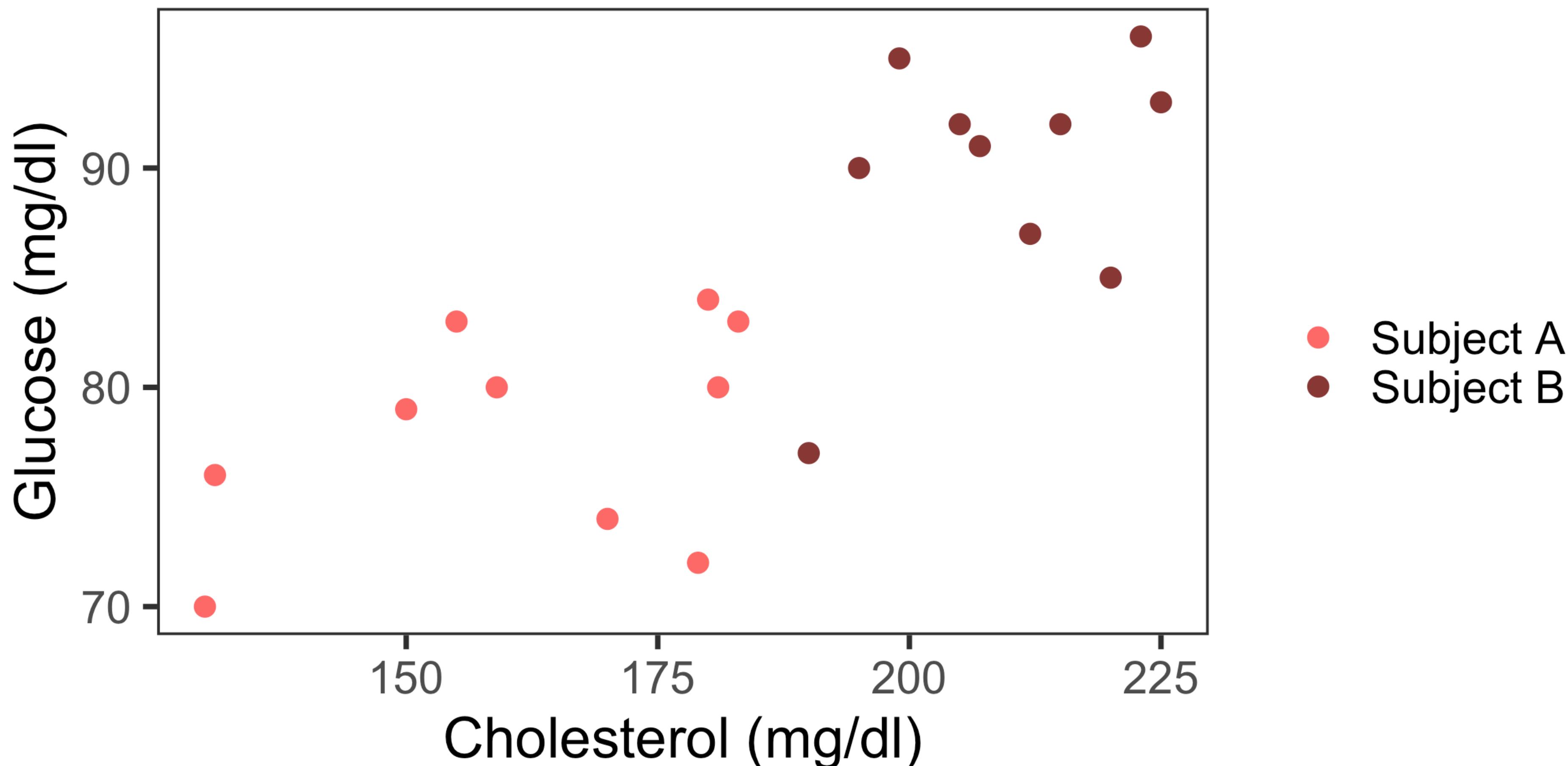
$$H_A : \beta_1 \neq 0$$

Difference between group means = β_1 !

$$(H_0 : \mu_1 - \mu_2 = 0)$$

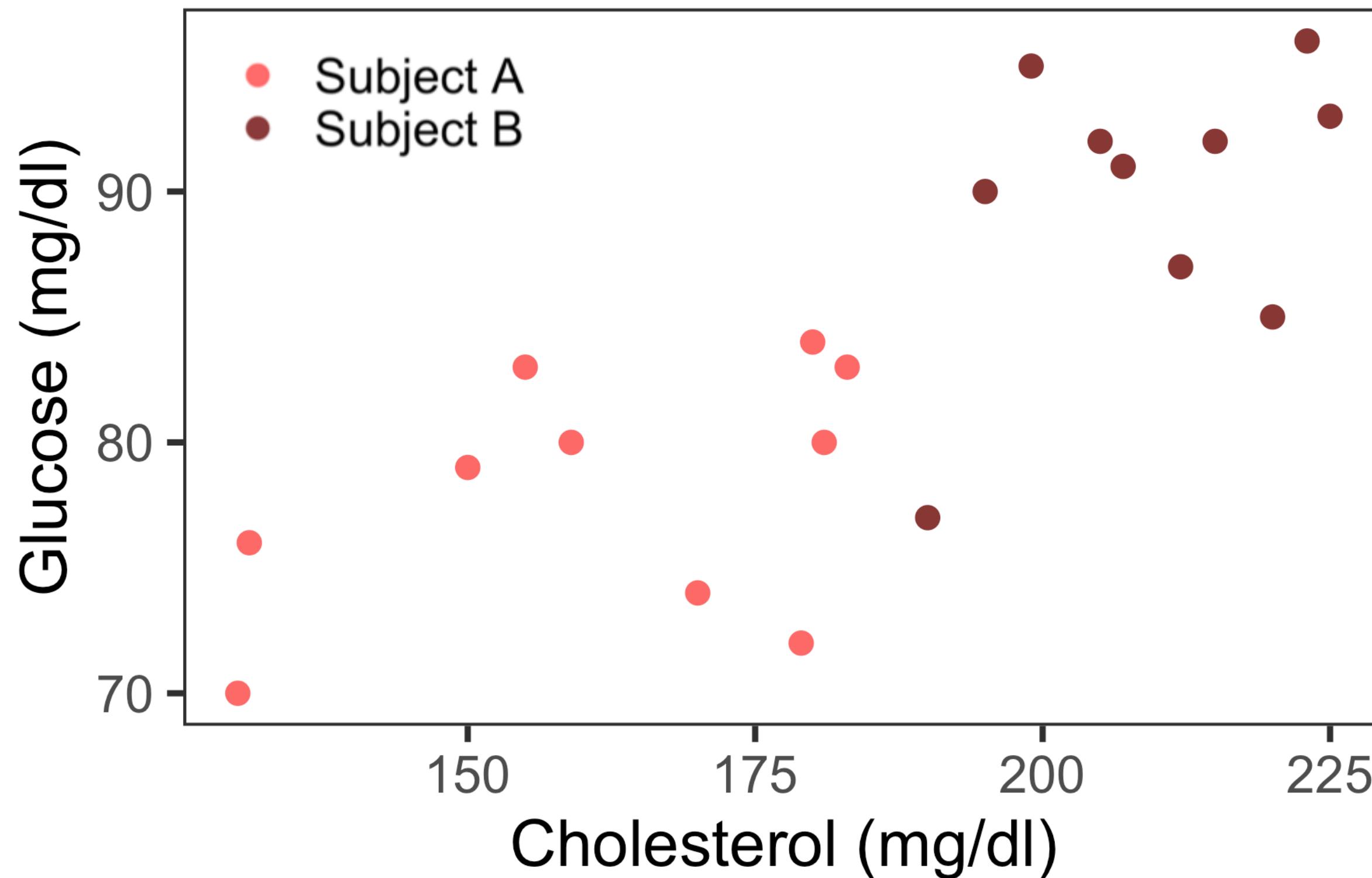
Regression and the *t*-test

Is there a relationship between cholesterol and glucose?



Regression and the *t*-test

Is there a relationship between cholesterol and glucose?



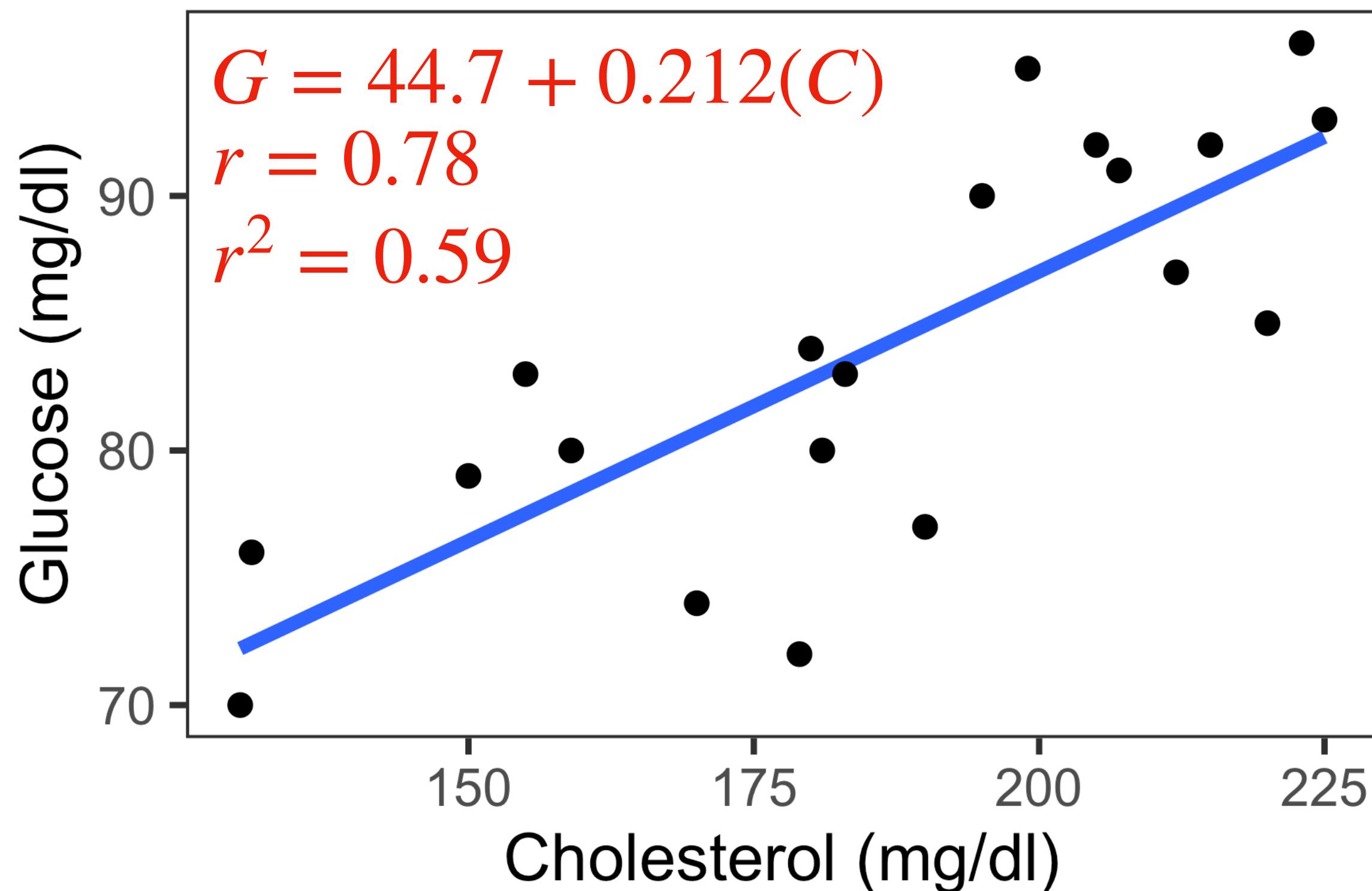
	Glucose (mg/dl)	
	Subject A	Subject B
Mean	78.1	89.8
SD	4.89	5.59
n	10	10

```
> t.test(subjectA$G, subjectB$G)
```

$t = -4.9814$,
 $df = 17.681$,
 $p\text{-value} = 0.0001018$

Regression and the *t*-test

Is there a relationship between cholesterol and glucose?

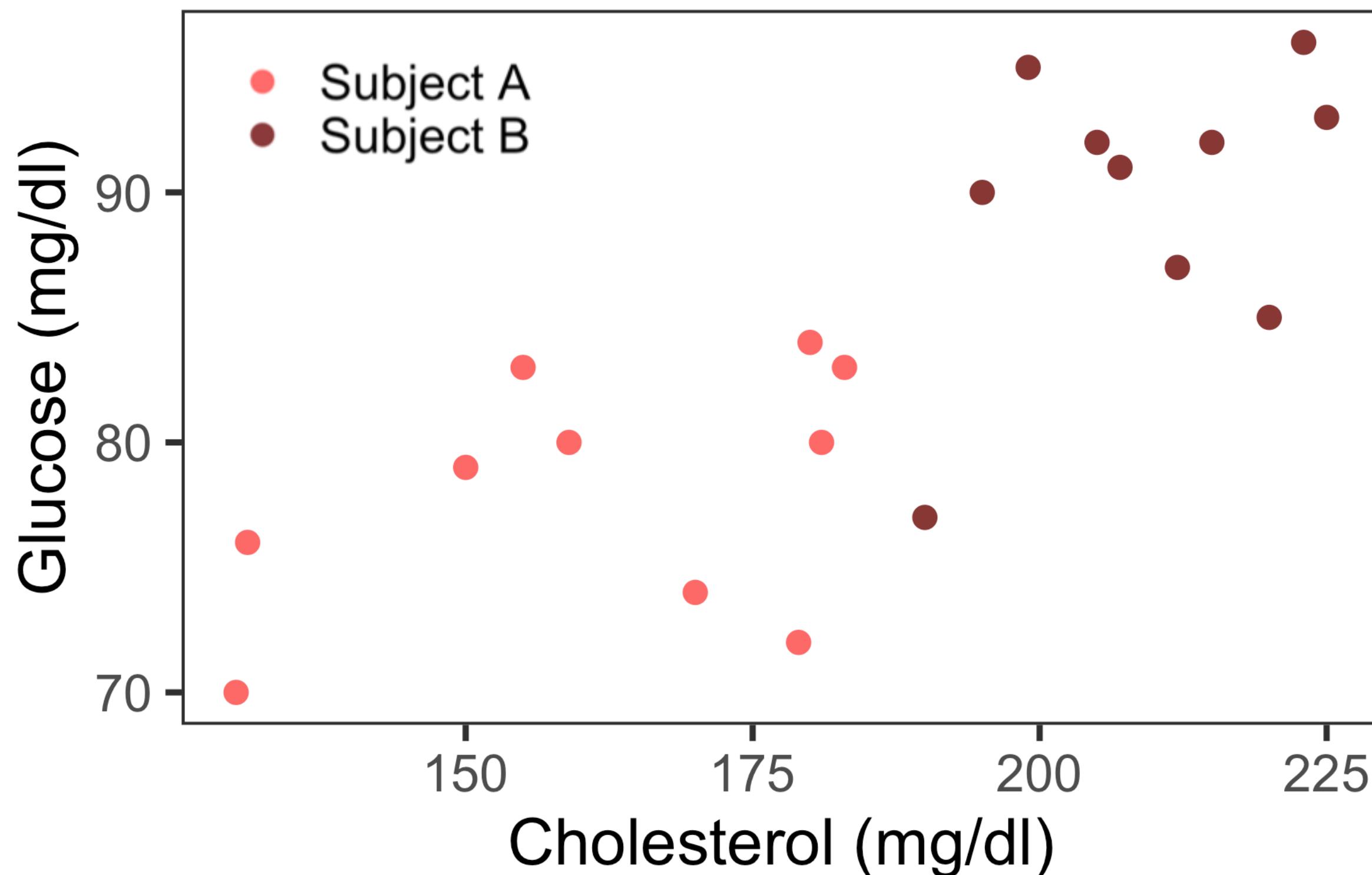


```
> lm(glucose ~ cholesterol)
```

$t = 5.330,$
 $df = 18,$
p-value = 4.57e-05

Multiple regression

Is there a relationship between cholesterol and glucose?



```
> lm(glucose~cholesterol+subject)
```

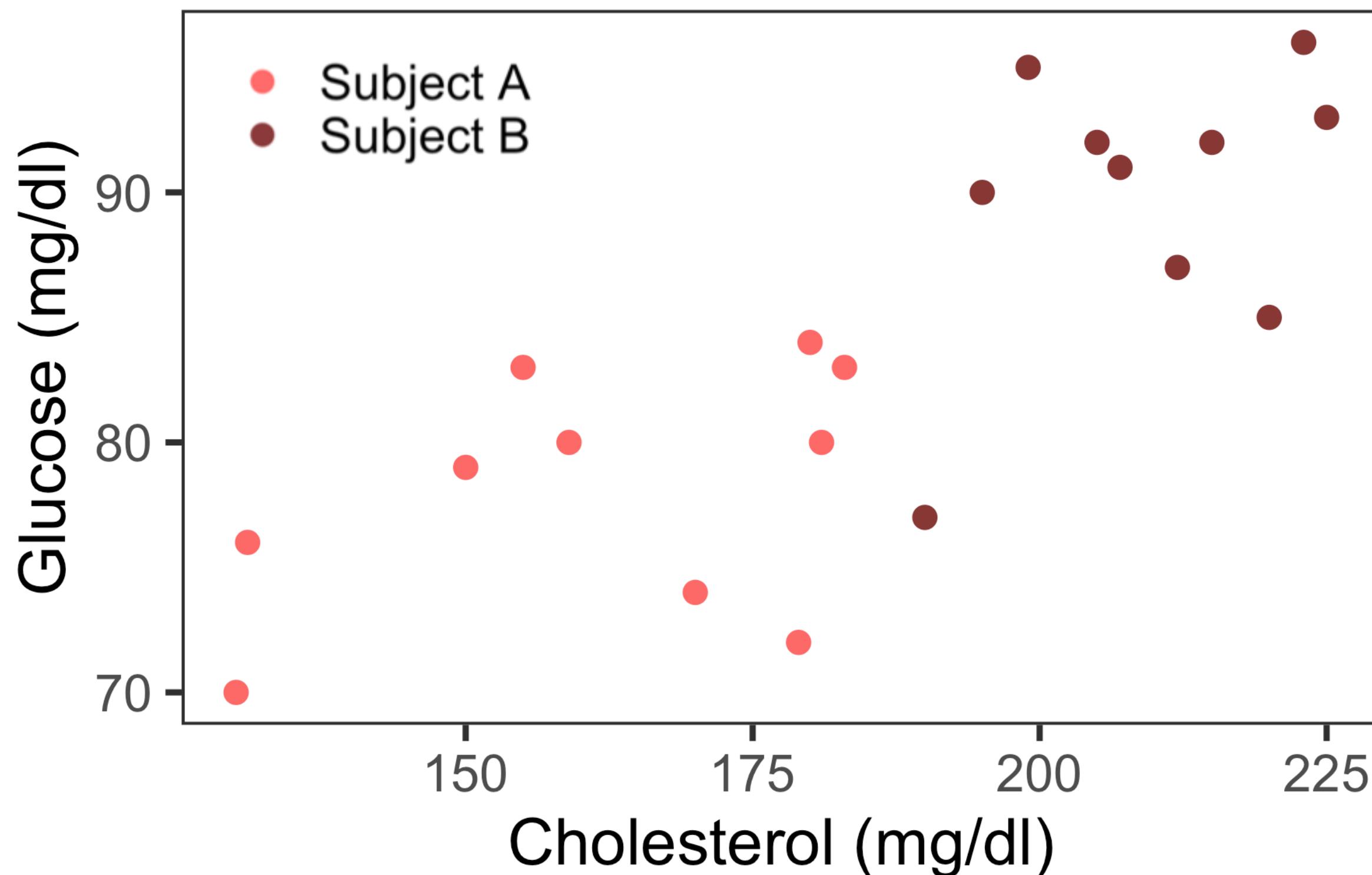
Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	56.91808	11.40891	4.989	0.000112	***
cholesterol	0.13091	0.06985	1.874	0.078206	.
subjectB	5.50776	3.96963	1.387	0.183218	

When subject and cholesterol are 0,
what is glucose? (Doesn't mean
anything biologically in our data)

Multiple regression

Is there a relationship between cholesterol and glucose?



```
> lm(glucose~cholesterol+subject)
```

Coefficients:

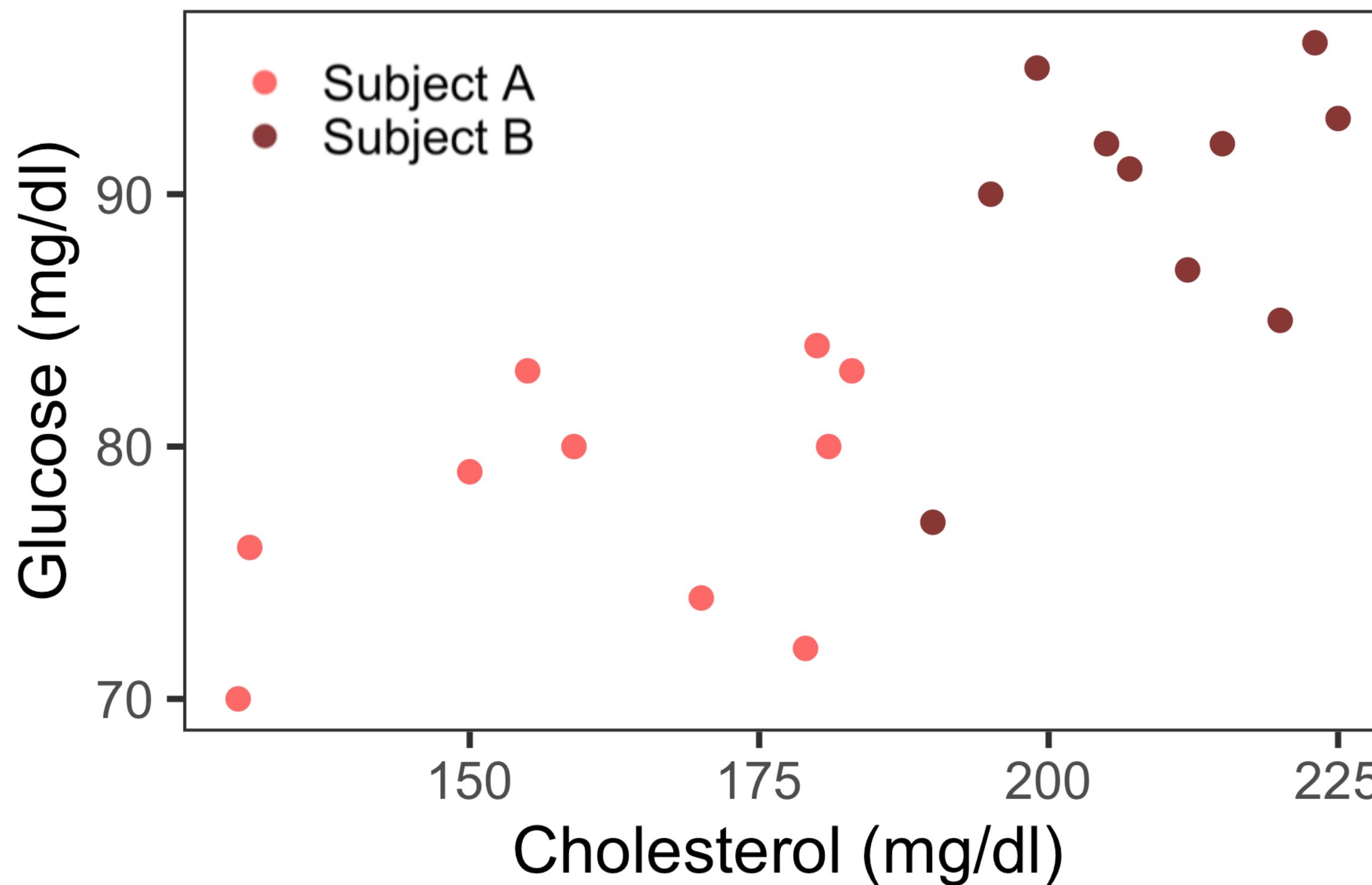
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	56.91808	11.40891	4.989	0.000112	***
cholesterol	0.13091	0.06985	1.874	0.078206	.
subjectB	5.50776	3.96963	1.387	0.183218	

When cholesterol increases by one unit, glucose increases by 0.13 units

(β_1 for cholesterol)

Multiple regression

Is there a relationship between cholesterol and glucose?



```
> lm(glucose~cholesterol+subject)
```

Coefficients:

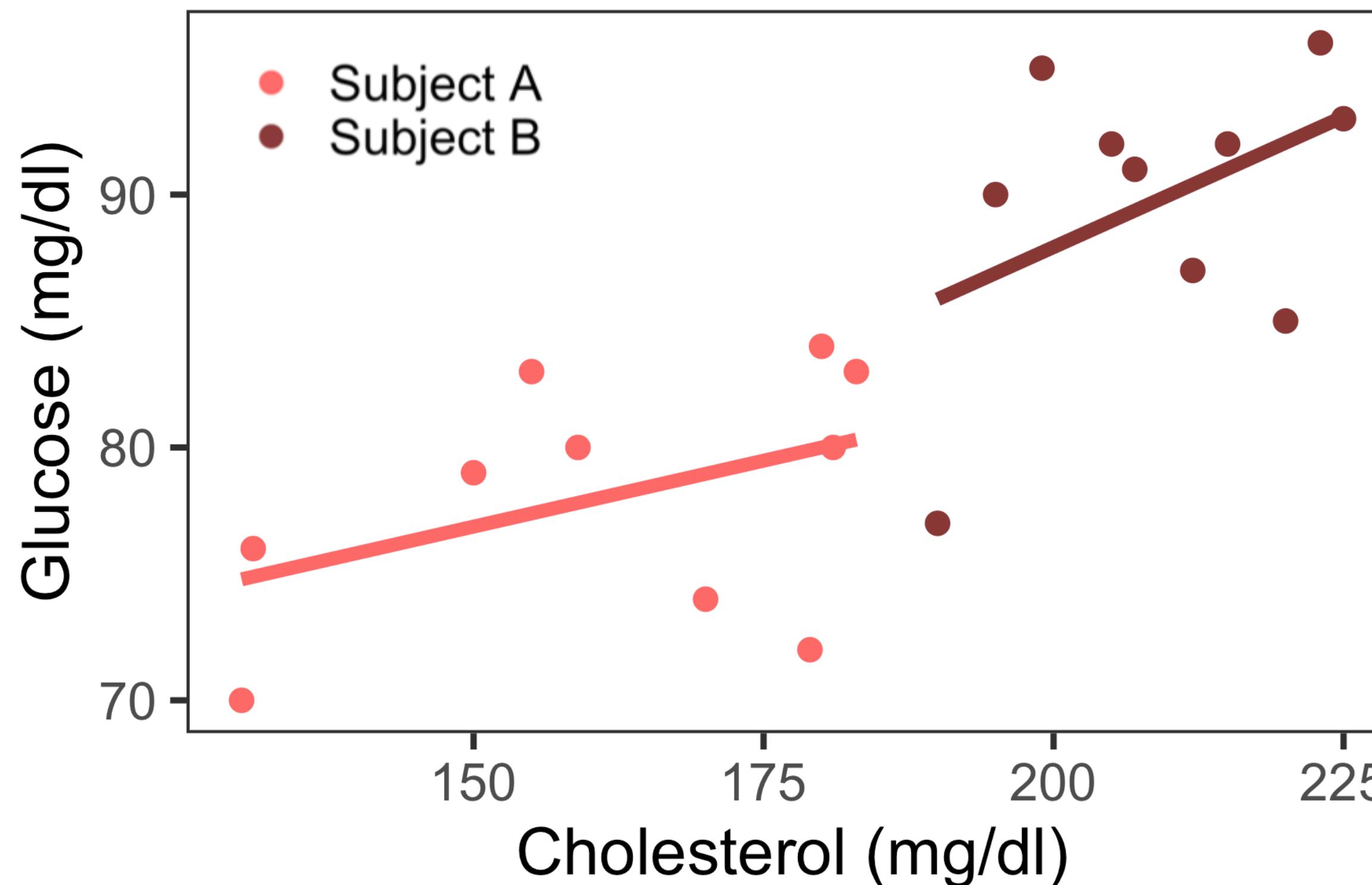
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	56.91808	11.40891	4.989	0.000112	***
cholesterol	0.13091	0.06985	1.874	0.078206	.
subjectB	5.50776	3.96963	1.387	0.183218	

When subject increases by one unit (i.e. from A to B), glucose increases by 5.5 units.

(β_1 for subject)

Multiple regression

Is there a relationship between cholesterol and glucose?



```
> lm(glucose~cholesterol+subject)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	56.91808	11.40891	4.989	0.000112 ***
cholesterol	0.13091	0.06985	1.874	0.078206 .
subjectB	5.50776	3.96963	1.387	0.183218

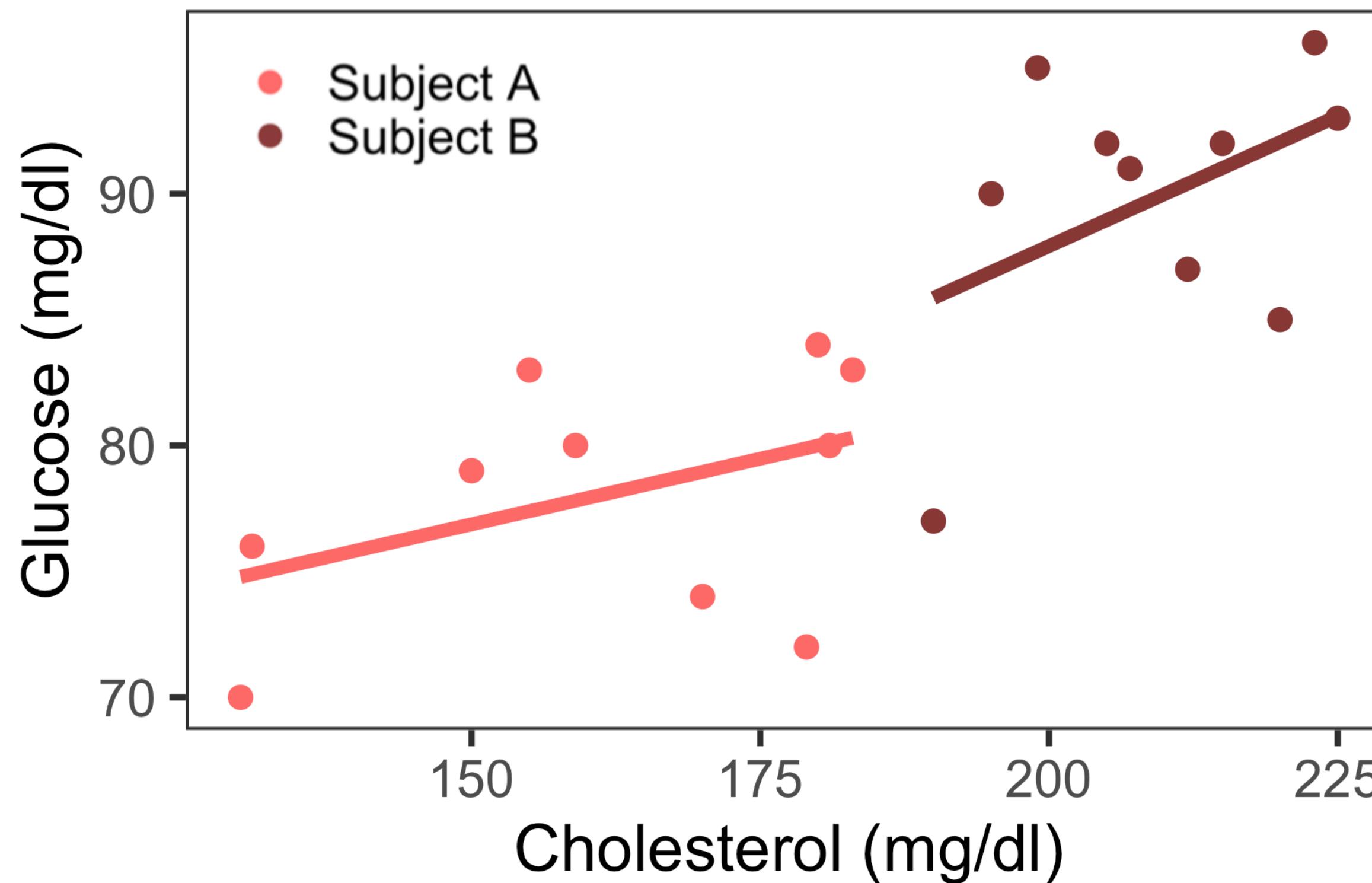
When subject increases by one unit (i.e. from A to B), glucose increases by 5.5 units.

(β_1 for subject)

$$G = 56.92 + 0.131(C) + 5.5(S)$$

Regression and the *t*-test

Is there a relationship between cholesterol and glucose (when controlling for subject)?



> lm(glucose ~ cholesterol)

t = 1.396,

b1 = 0.2,

p-value = 0.20

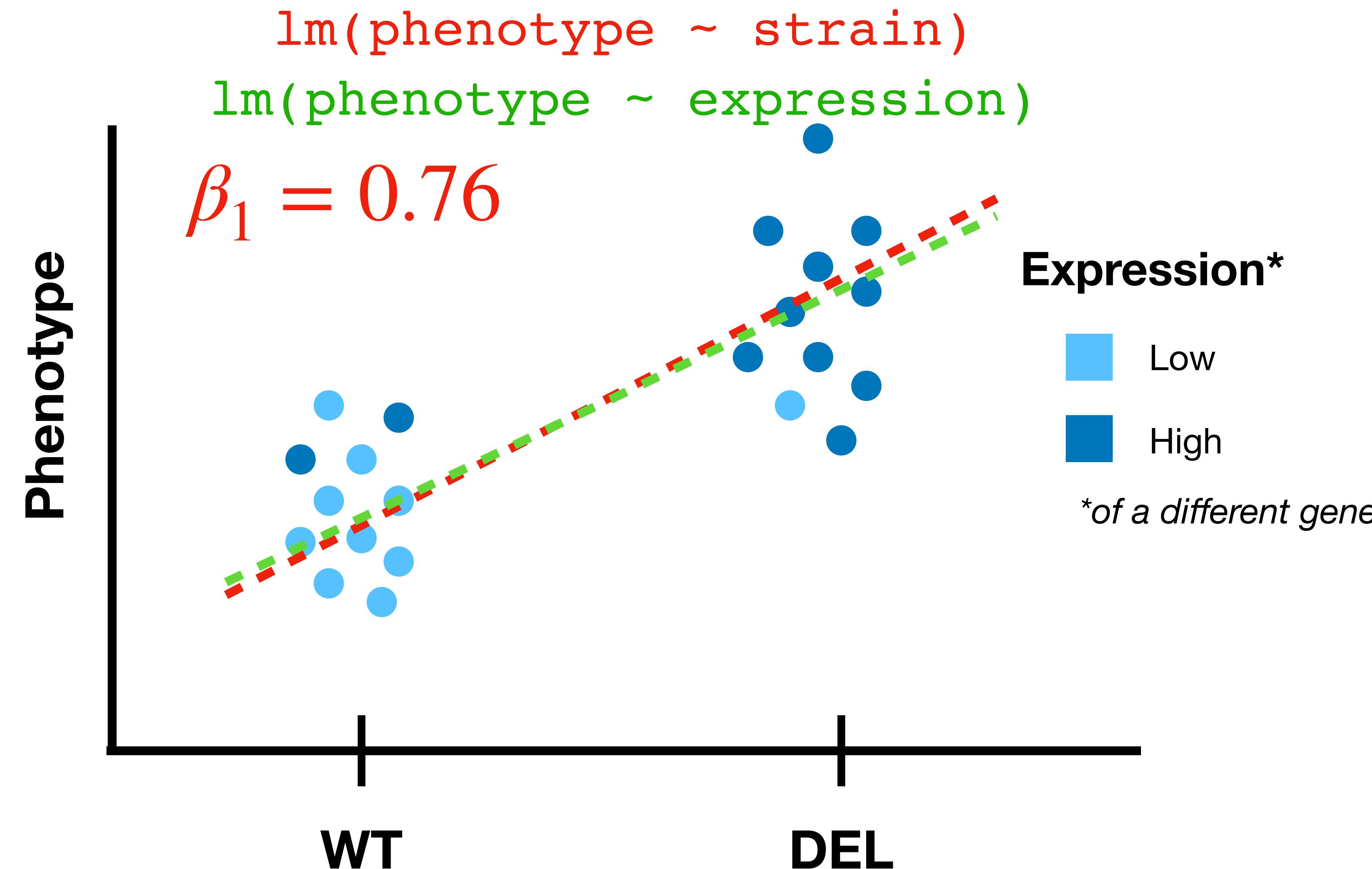
> lm(glucose ~ cholesterol)

t = 1.350,

b1 = 0.1,

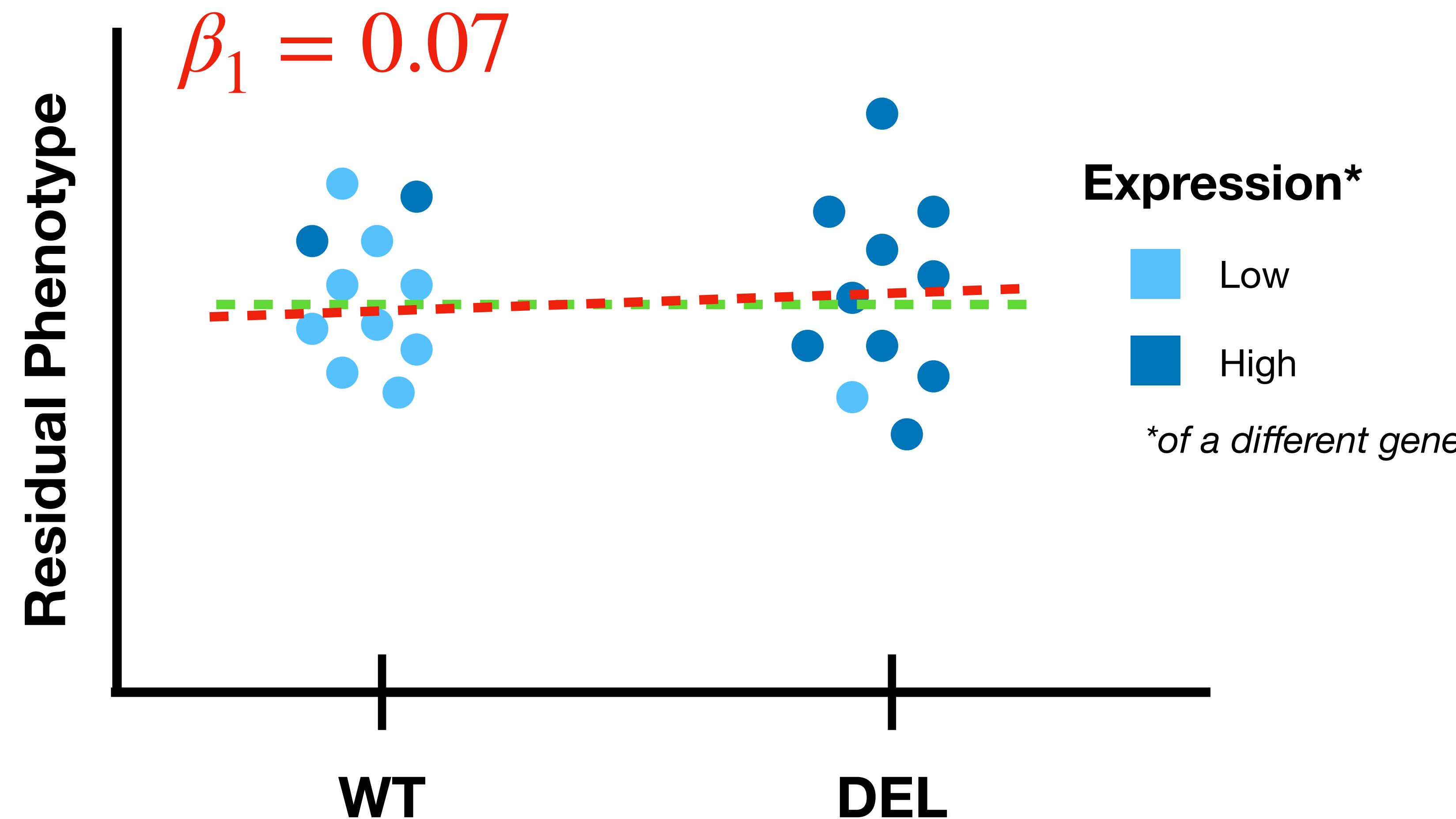
p-value = 0.21411

Using residual values to identify covariates

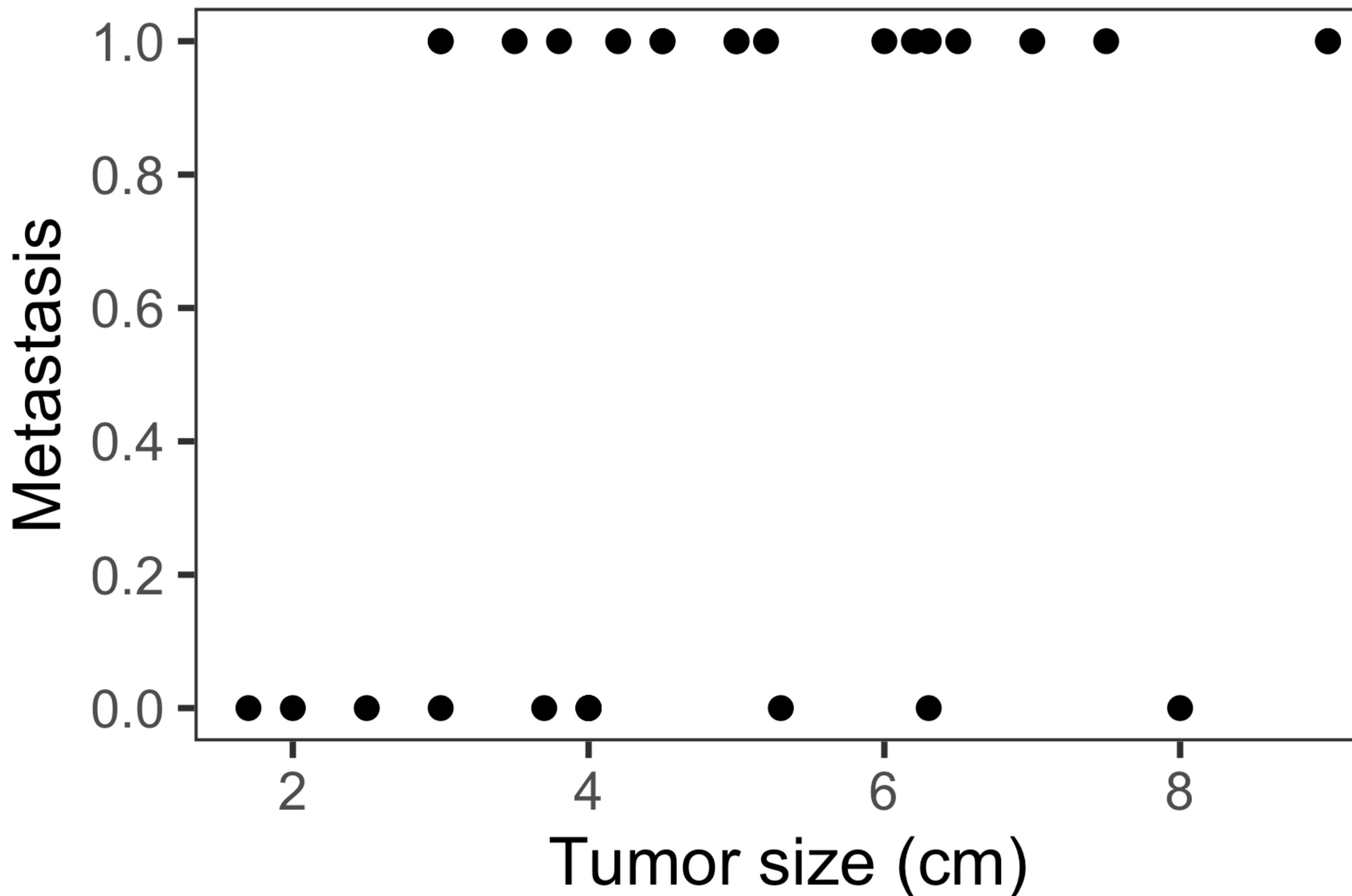


Using residual values to identify covariates

`lm(phenotype ~ strain + expression)`



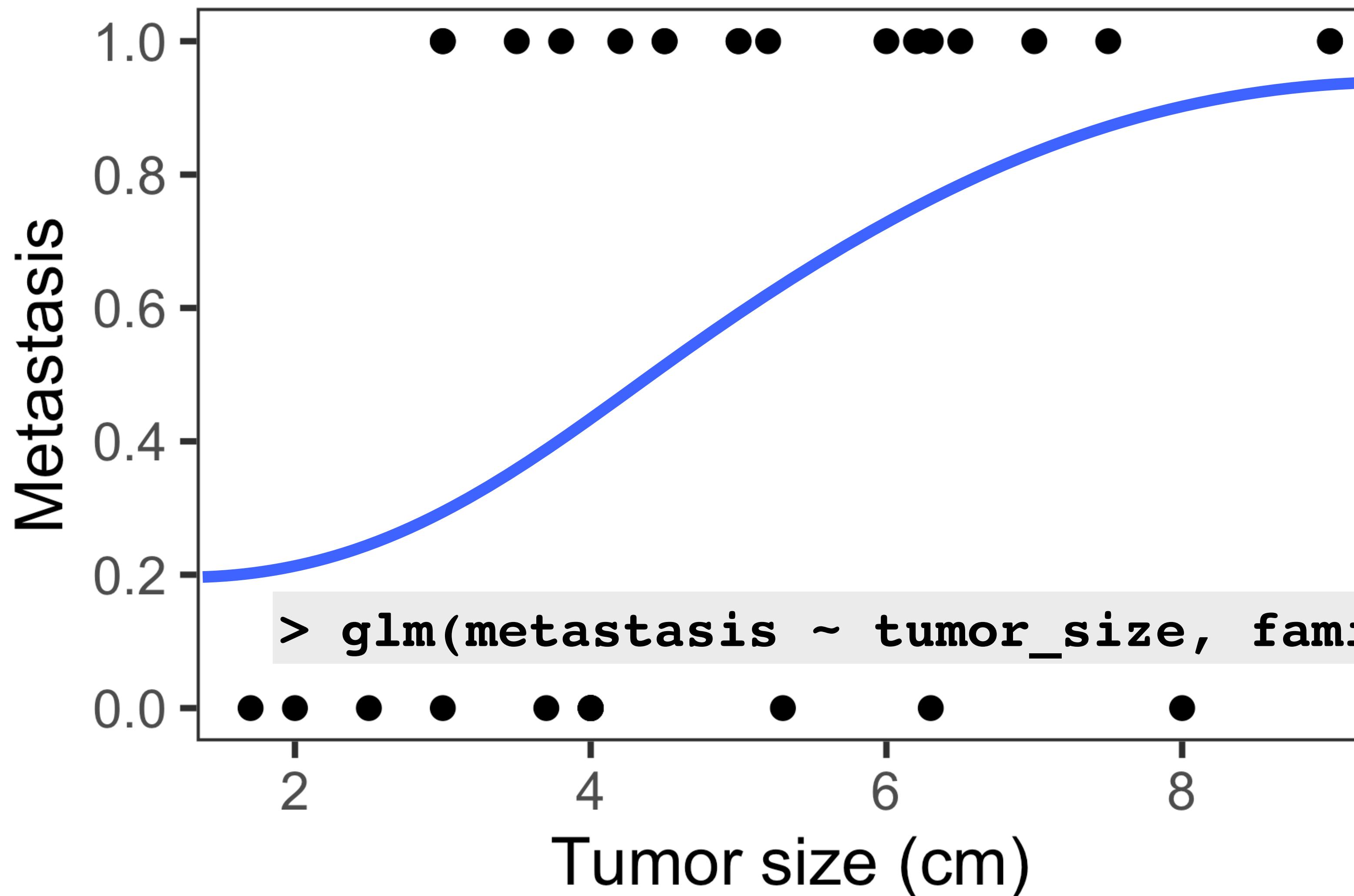
Introduction to logistic regression



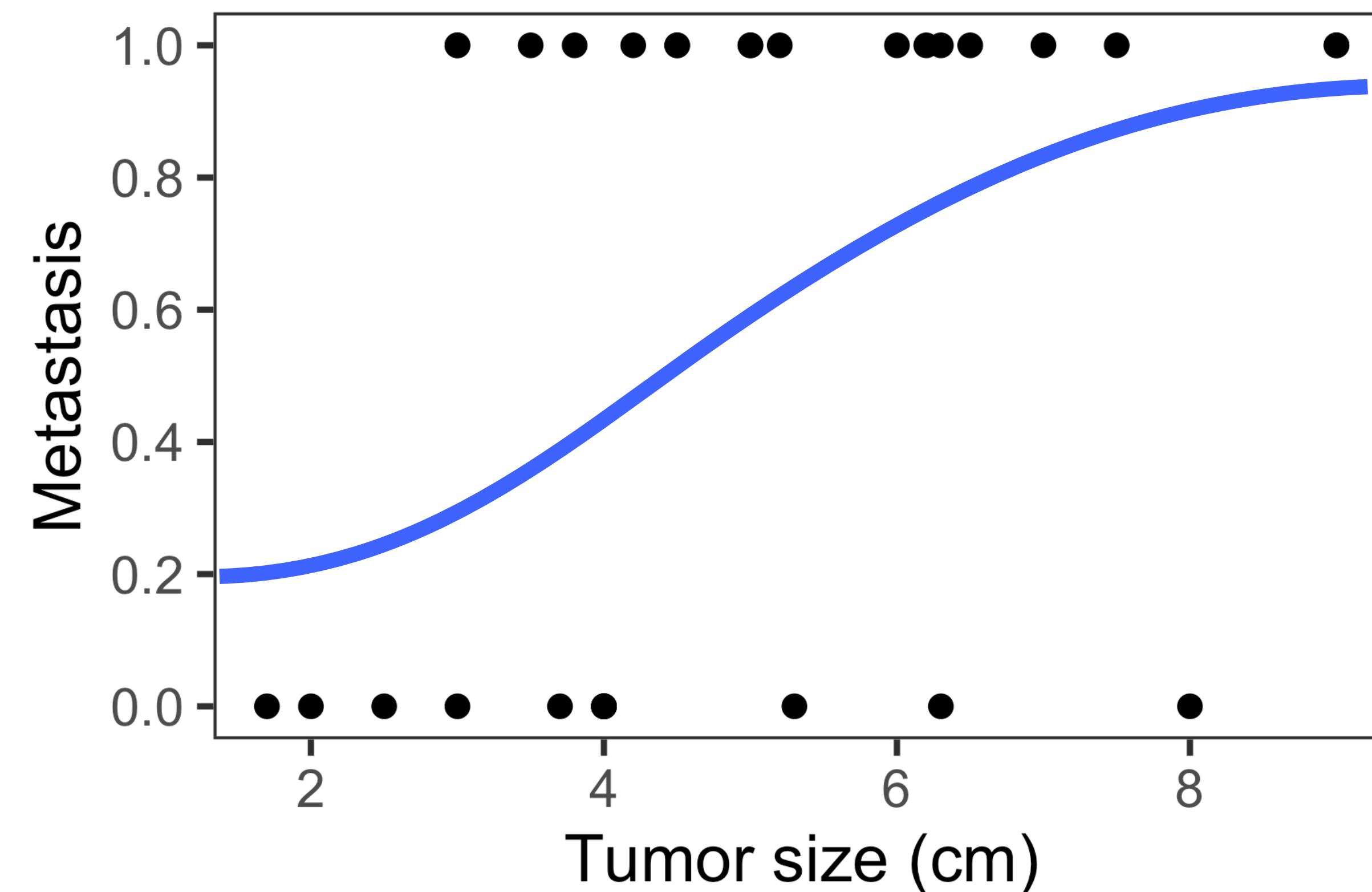
Introduction to logistic regression

- Model relationship between X and Y by fitting a response curve that is always between 0 and 1
- Probability that $Y = 1$ (metastasis) for a given value of X (tumor size)
- **Linear function does not remain between 0 and 1**
- Logistic regression will have an “S” shape
- No error term: modeling the probability of an event directly
- `glm(Y ~ x, family = "binomial")`

Introduction to logistic regression



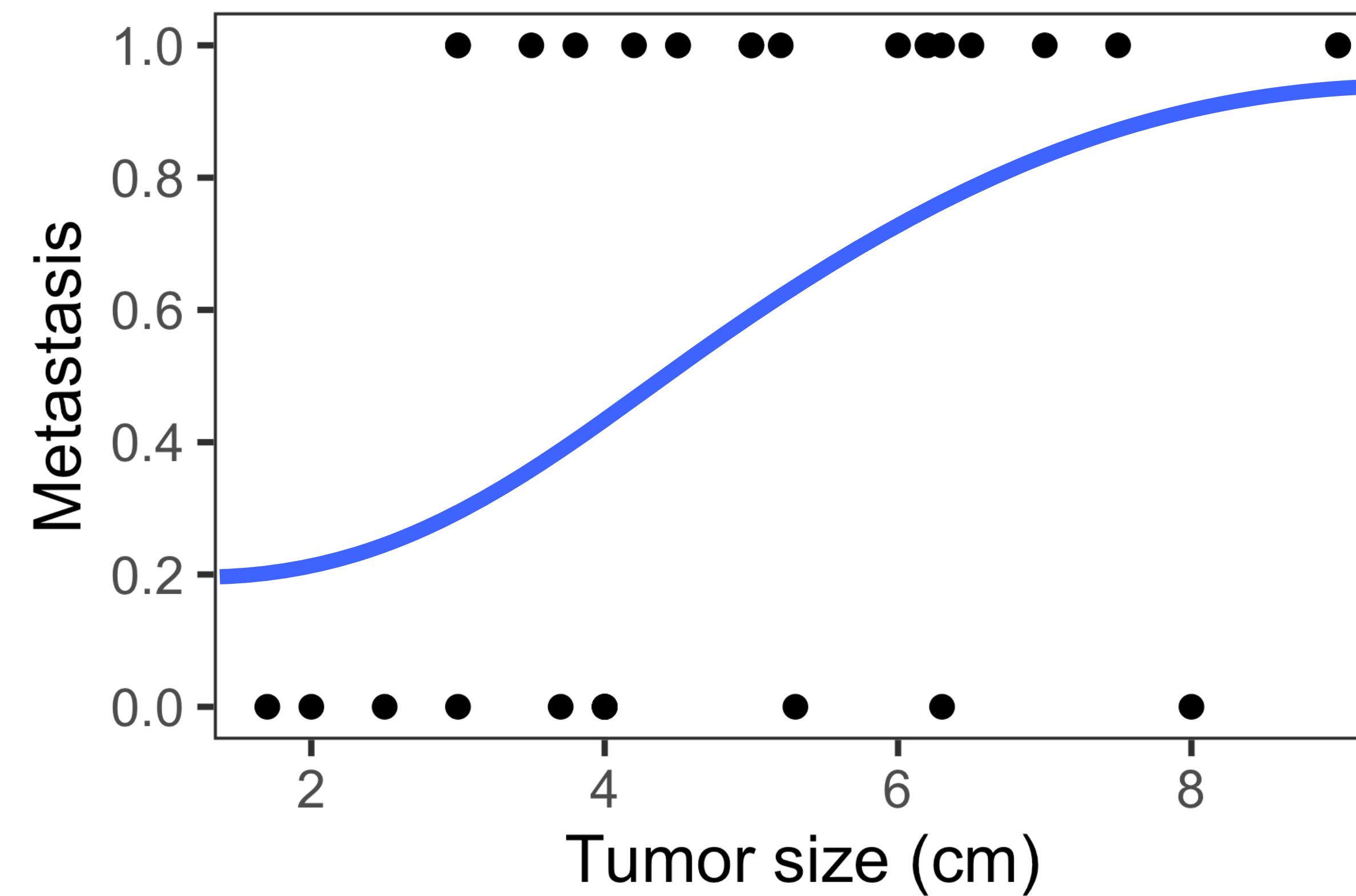
Introduction to logistic regression



Size range	Points with Y = 1	Points with Y = 0	Fraction Y = 1	Proportion Y = 1
(1.5, 3.0]				
(3.0, 4.5]				
(4.5, 6.0]				
(6.0, 7.5]				
(7.5, 9.0]				

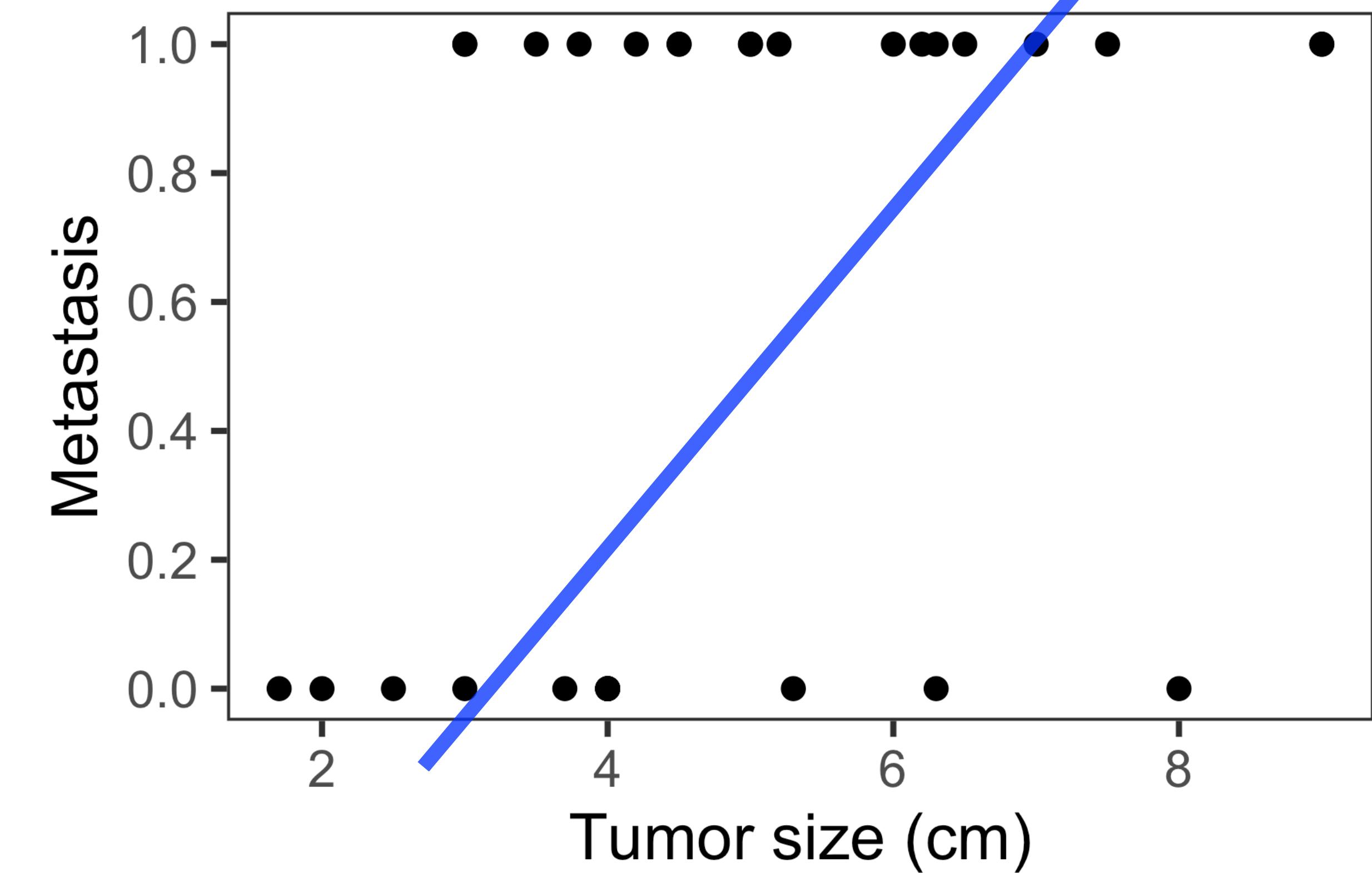
Introduction to logistic regression

Logistic regression



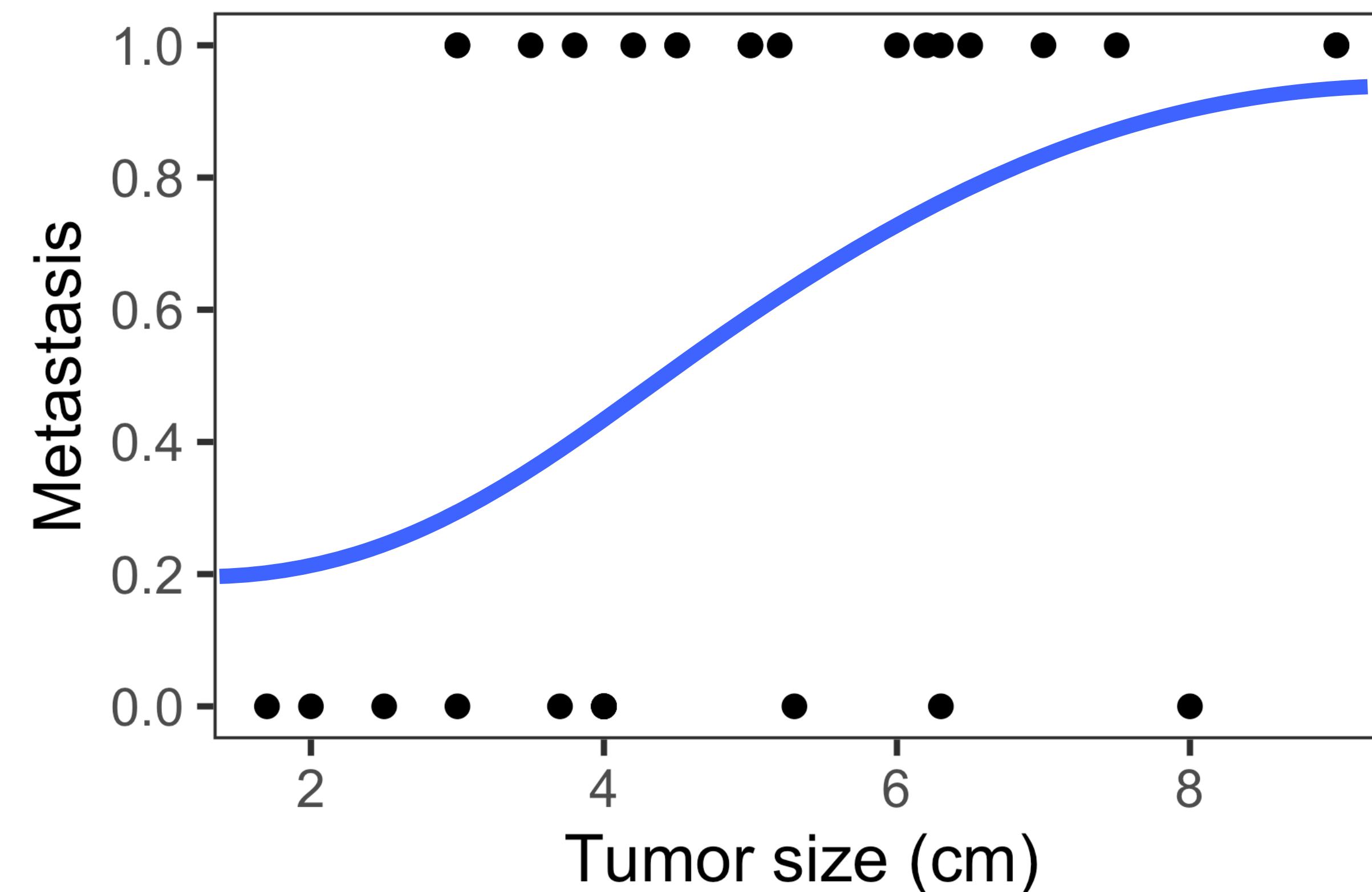
Predicted Y within 0-1 range

Linear regression



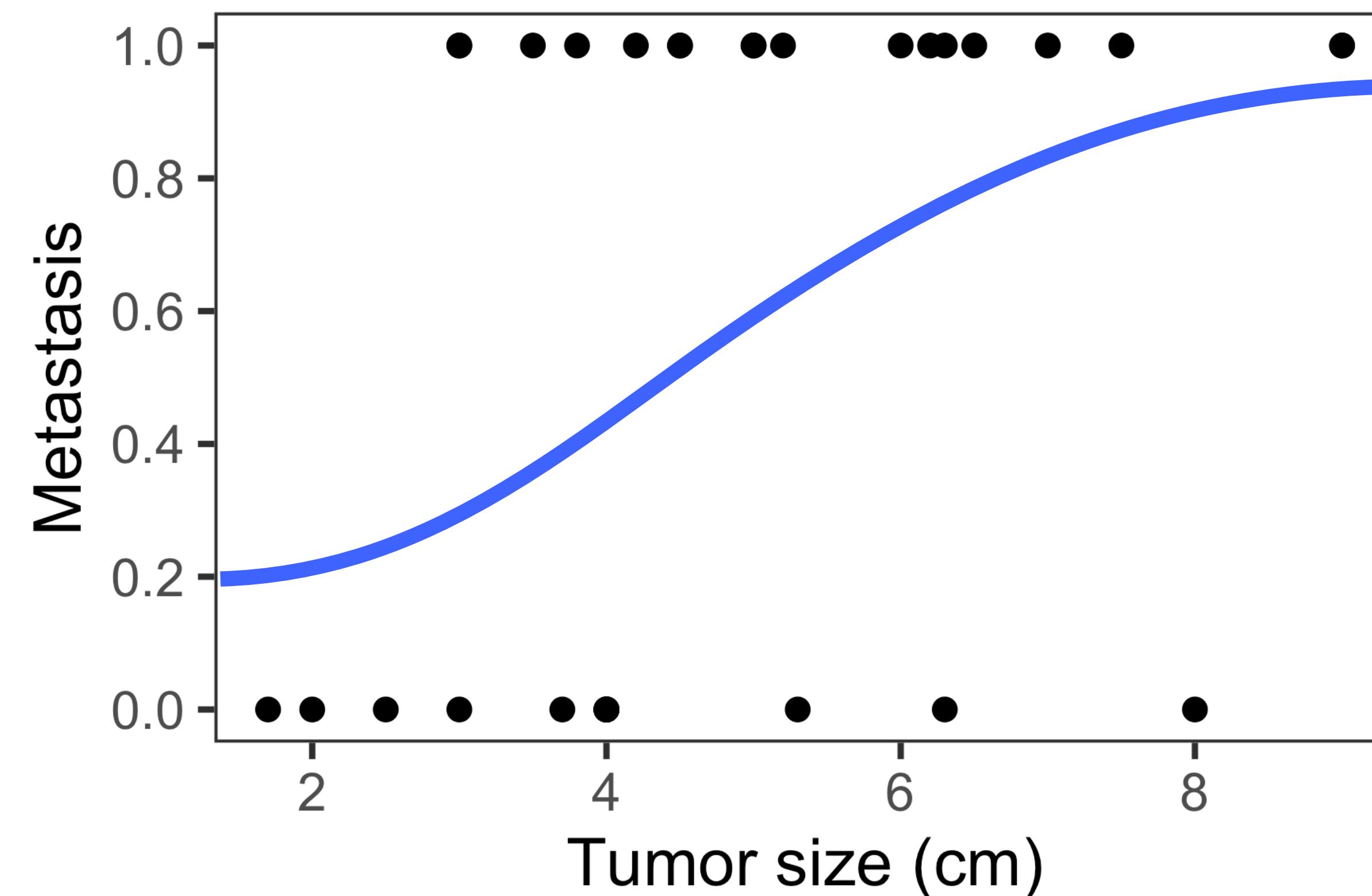
Predicted Y can exceed 0-1 range

Introduction to logistic regression



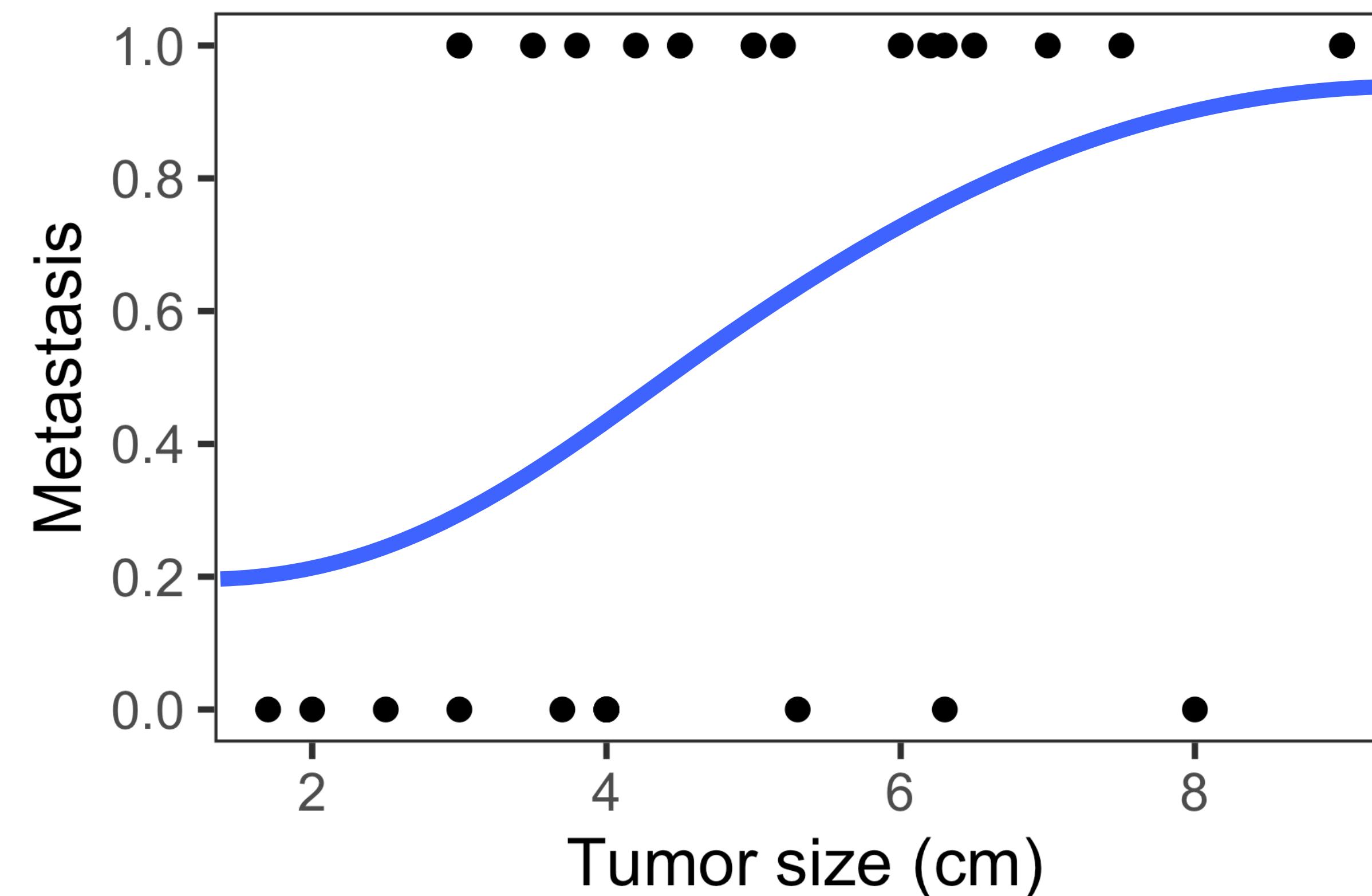
Size range	Points with Y = 1	Points with Y = 0	Fraction Y = 1	Proportion Y = 1
(1.5, 3.0]	2	4		
(3.0, 4.5]	5	6		
(4.5, 6.0]	4	1		
(6.0, 7.5]	5	1		
(7.5, 9.0]	2	1		

Introduction to logistic regression



Size range	Points with Y = 1	Points with Y = 0	Fraction Y = 1	Proportion Y = 1
(1.5, 3.0]	2	4	2/6	
(3.0, 4.5]	5	6	5/11	
(4.5, 6.0]	4	1	4/5	
(6.0, 7.5]	5	1	5/6	
(7.5, 9.0]	2	1	2/3	

Introduction to logistic regression



Size range	Points with Y = 1	Points with Y = 0	Fraction Y = 1	Proportion Y = 1
(1.5, 3.0]	2	4	2/6	0.33
(3.0, 4.5]	5	6	5/11	0.45
(4.5, 6.0]	4	1	4/5	0.80
(6.0, 7.5]	5	1	5/6	0.83
(7.5, 9.0]	2	1	2/3	0.67

Introduction to logistic regression

```
> glm(metastasis ~ tumor_size, family = "binomial")
```

Call:

```
glm(formula = Metastasis ~ Size, family = "binomial", data = metastasis)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0657	-1.1288	0.5657	0.9844	1.4185

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.0858	1.2256	-1.702	0.0888 .
Size	0.5117	0.2561	1.998	0.0457 *

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 42.165 on 30 degrees of freedom

Residual deviance: 37.002 on 29 degrees of freedom

AIC: 41.002

Number of Fisher Scoring iterations: 4

Real life examples of logistic models

Spam email detection

Extracts info like sender,
typos, key phrases and
produces probability

Credit card fraud

Extracts info like when/where
purchase, how much, produces
probability of fraud

Tumor prediction

Like the example we had!

Marketing

Will this user like this particular
advertisement?