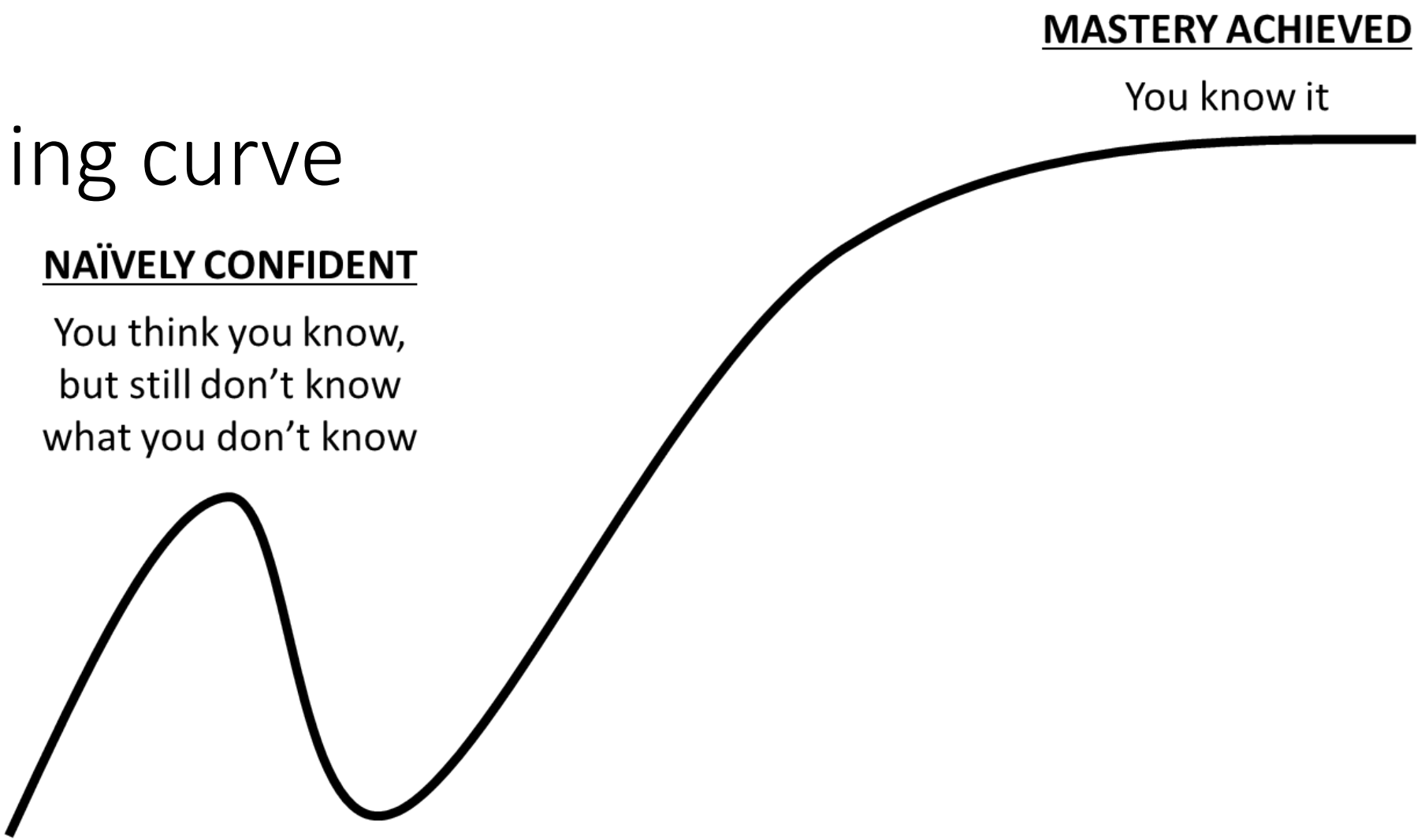
Statistics with Spa  ows II

Many models, matrices, and magic

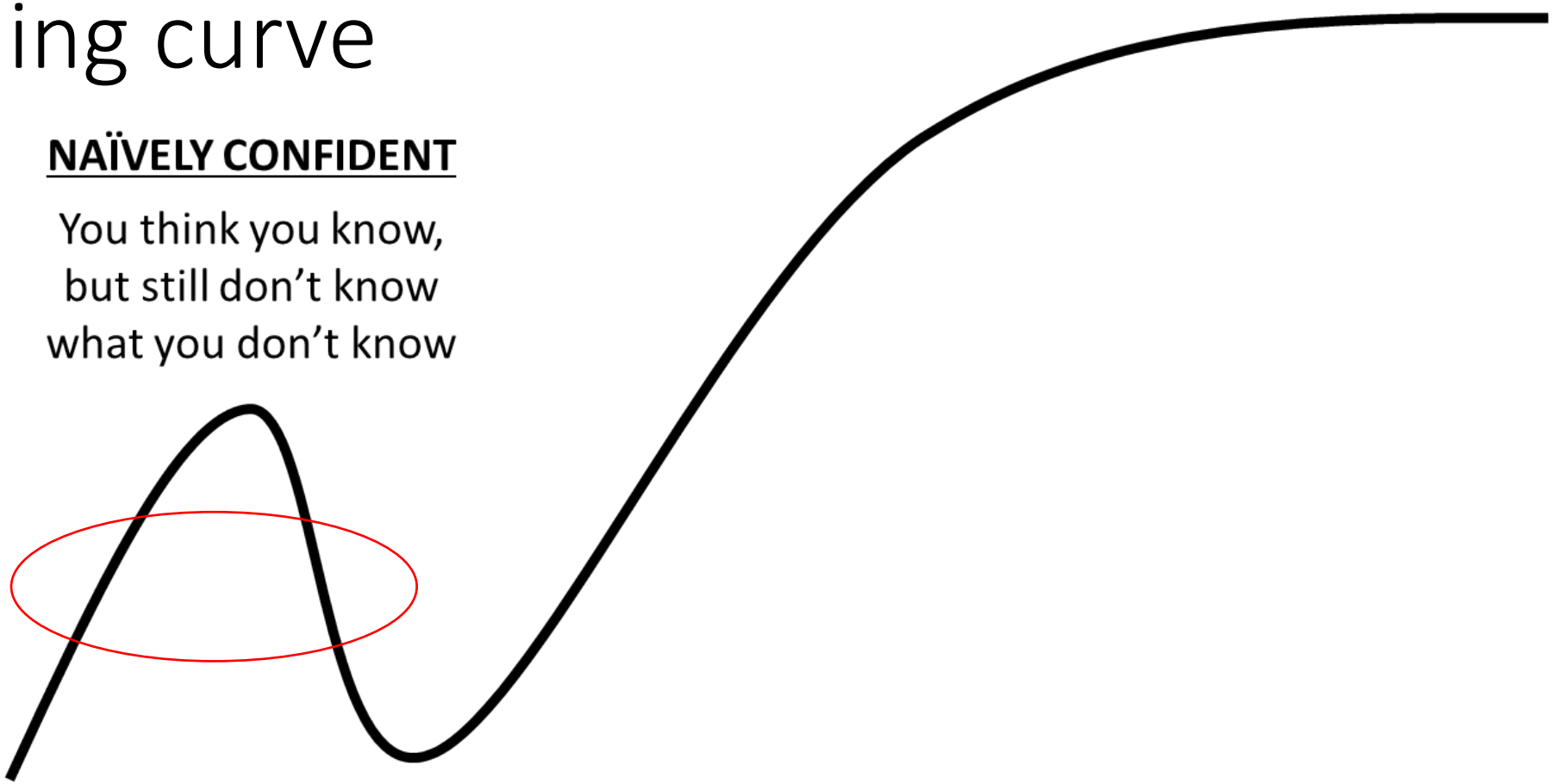
Julia Schroeder

Julia.schroeder@imperial.ac.uk

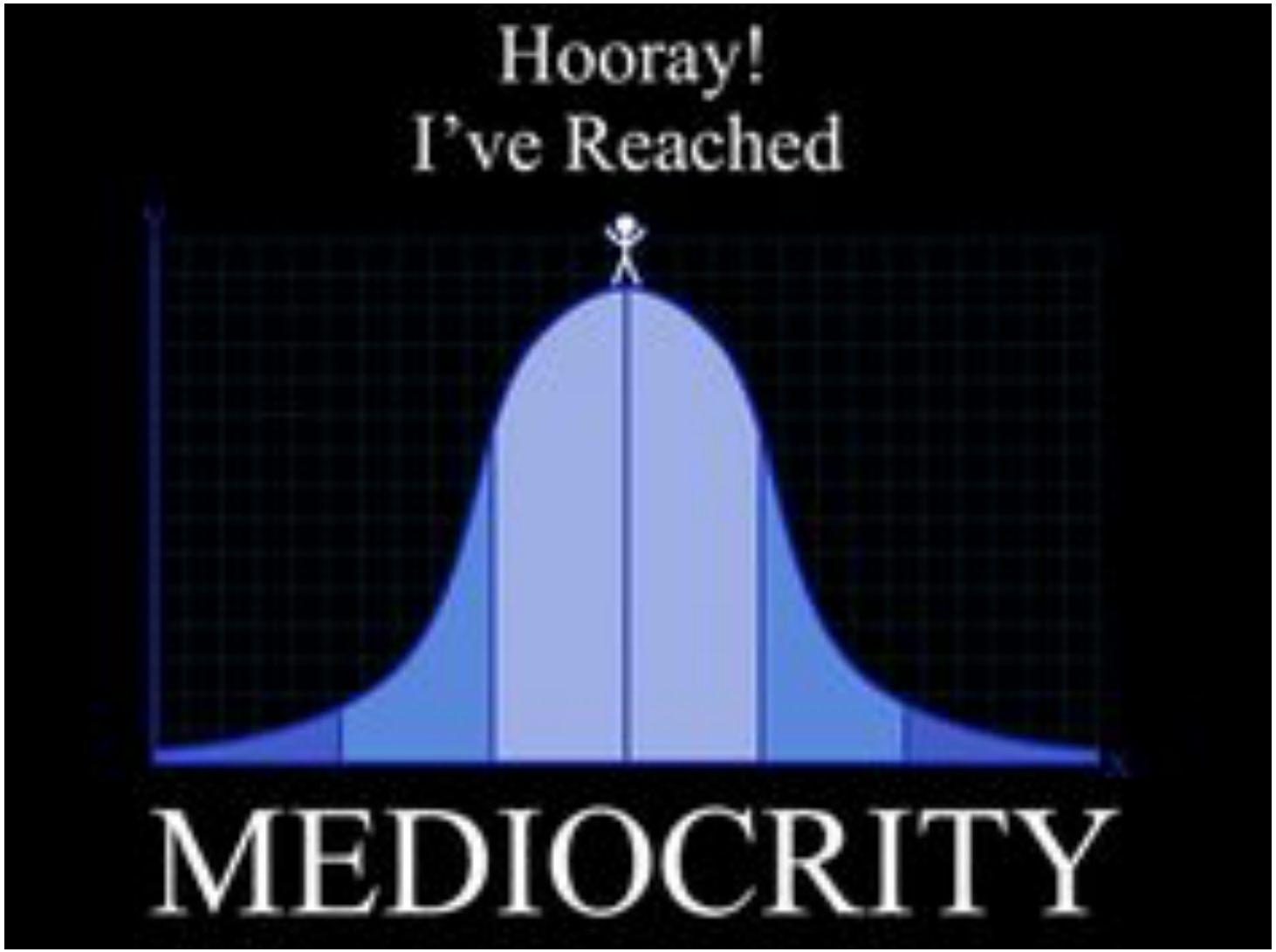
Learning curve



Learning curve



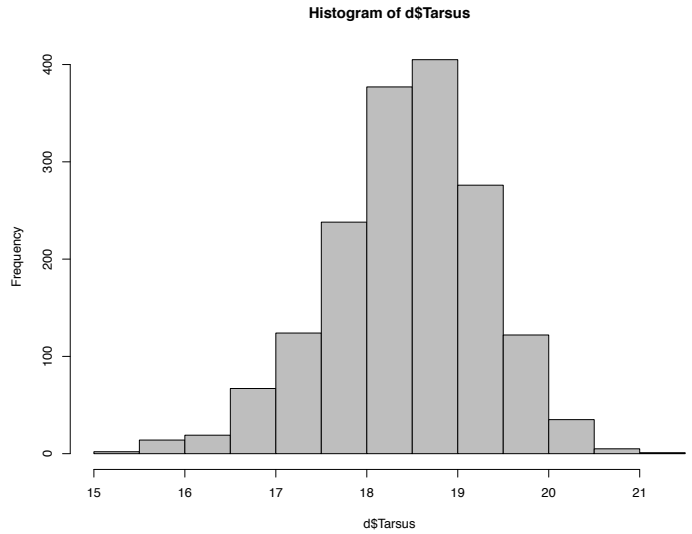
Learning curve



I never make the same mistake twice. I make it like five or six time, you know, just to be sure.

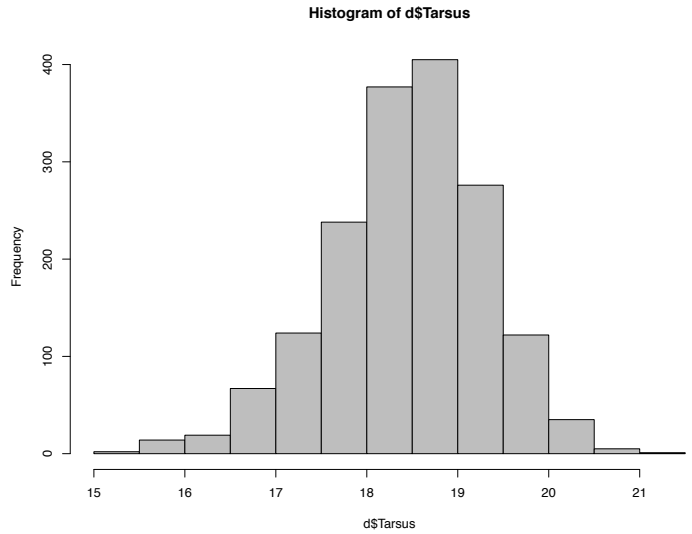
someecards
user card

Describing data distributions



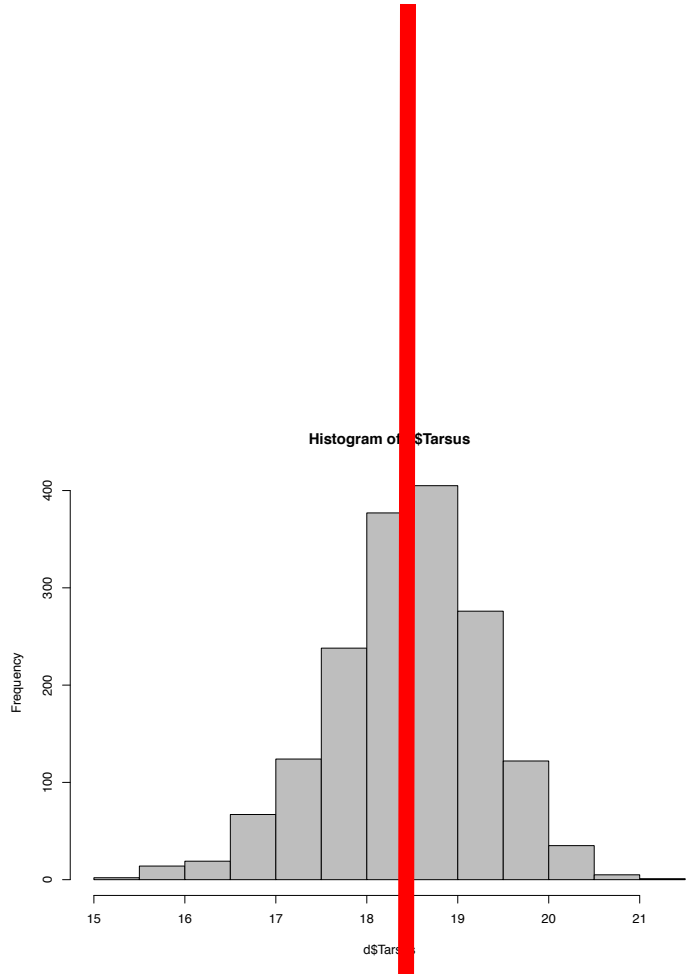
Describing data distributions

- Centrality



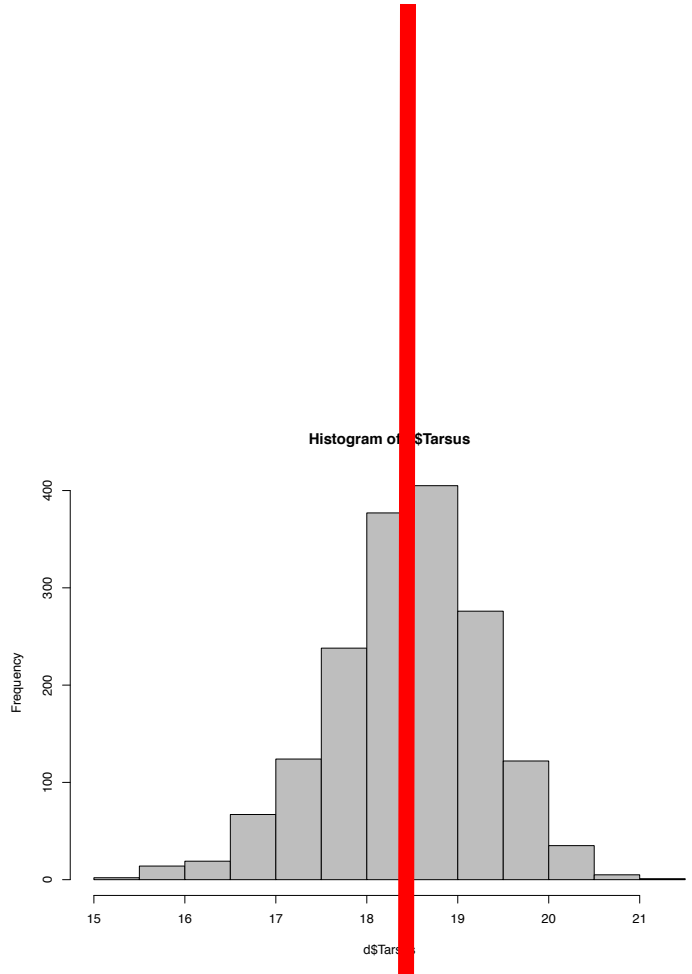
Describing data distributions

- Centrality



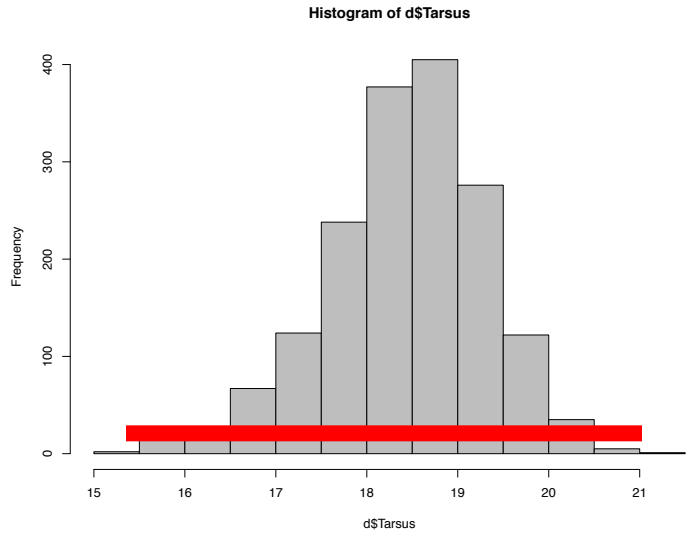
Describing data distributions

- Centrality
  - Mean $\overline{\mu}$
  - Mode
  - Median



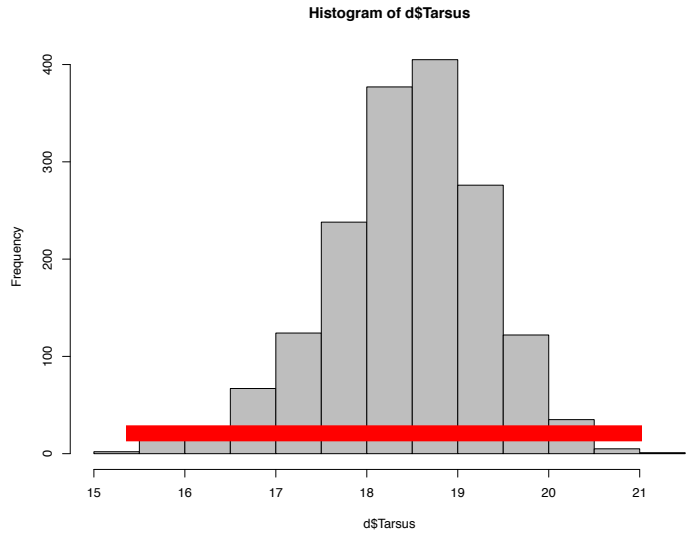
Describing data distributions

- Centrality
  - Mean $\overline{\mu}$
  - Mode
  - Median
- Spread



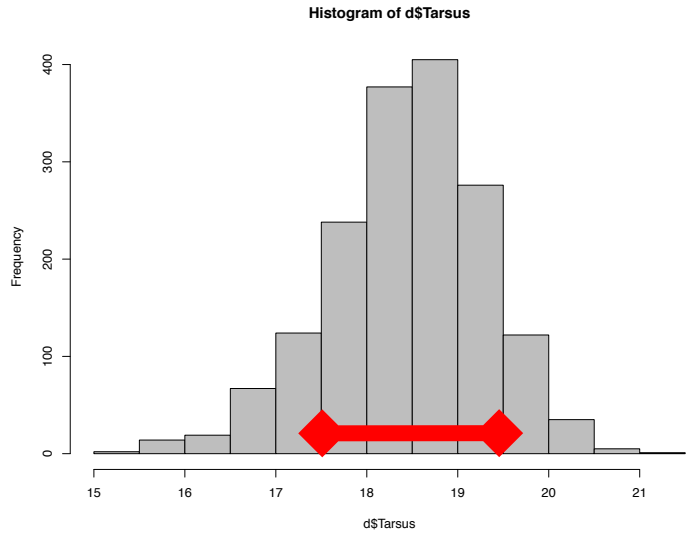
Describing data distributions

- Centrality
  - Mean $\overline{\mu}$
  - Mode
  - Median
- Spread
  - Range (min, max)



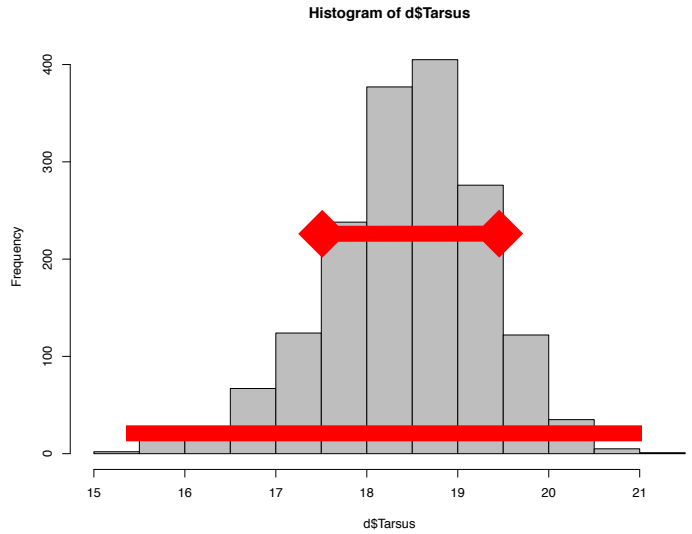
Describing data distributions

- Centrality
  - Mean $\overline{\mu}$
  - Mode
  - Median
- Spread
  - Range (min, max)
  - Standard deviation σ $-$ 68.2%



Describing data distributions

- Centrality
  - Mean $\overline{\mu}$
  - Mode
  - Median
- Spread
  - Range (min, max)
  - Standard deviation $\sigma - 68.2\%$
  - Variance σ^2



Adding variances

$$\sigma^2_{Tarsus+Wing} = \sigma^2_{Tarsus} + \sigma^2_{Wing} + 2COV_{Wing,Tarsus}$$

Adding variances

$$\sigma^2_{Tarsus+Wing} = \sigma^2_{Tarsus} + \sigma^2_{Wing} + 2COV_{Wing,Tarsus}$$

$$\sigma^2_{Tarsus+Wing} = \sigma^2_{Tarsus} + \sigma^2_{Wing}$$

If *Tarsus* and *Wing* are independent variables

Multiplying variances

$$10^2 \, \sigma^2_{Tarsus} = \sigma^2_{10Tarsus}$$

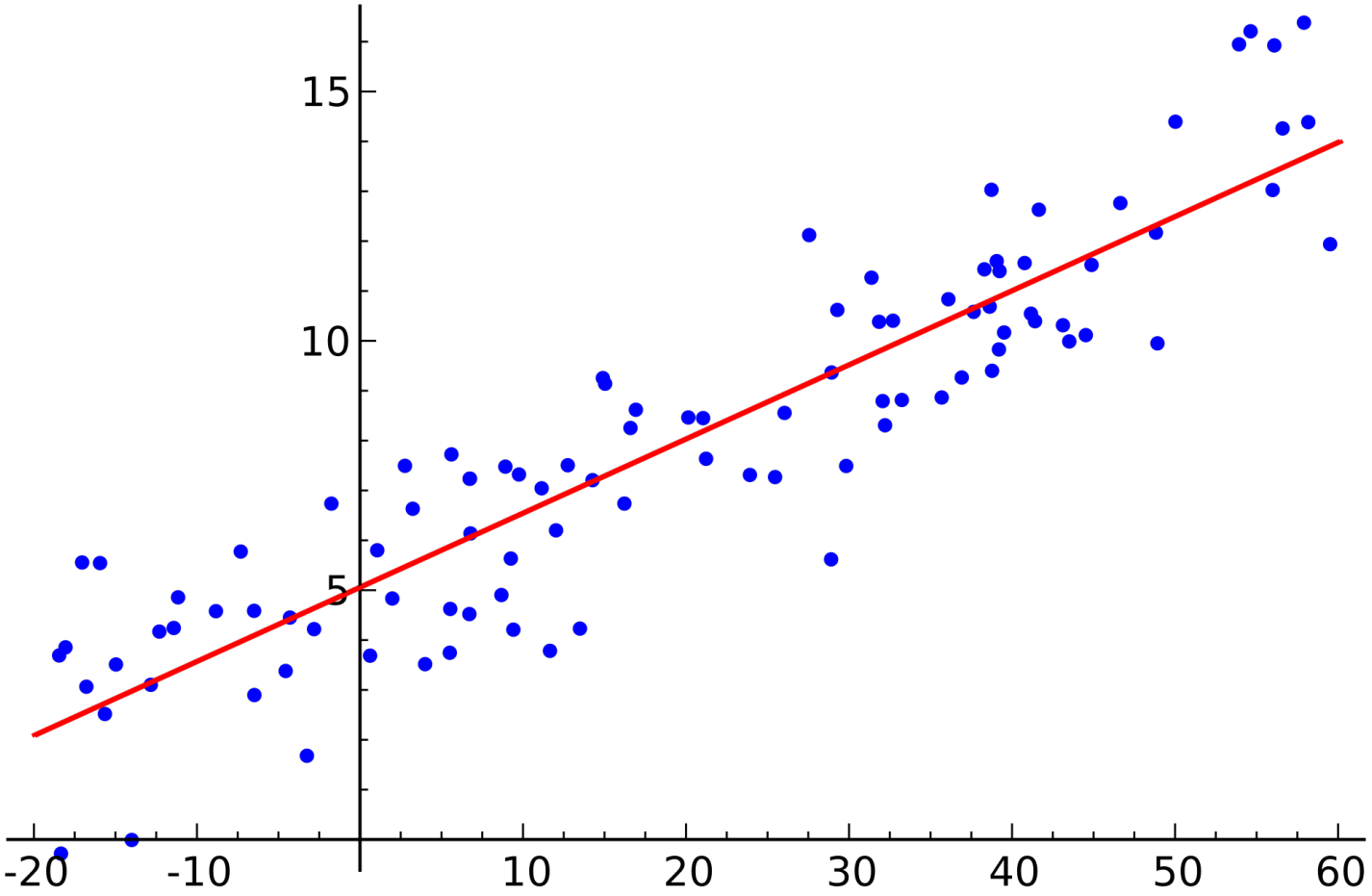
Linear models

- Basic building block of statistics
- Super useful
- Versatile
- Expandable

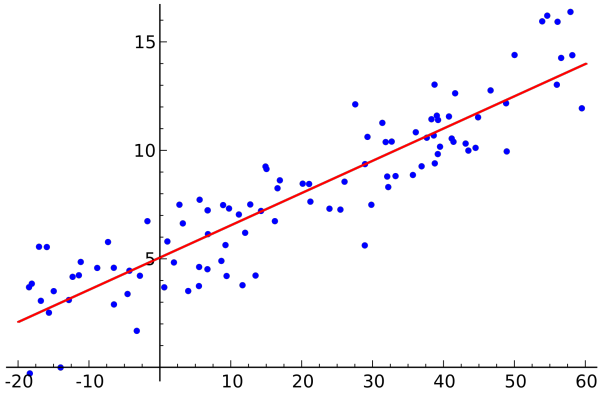
Fitting models to data

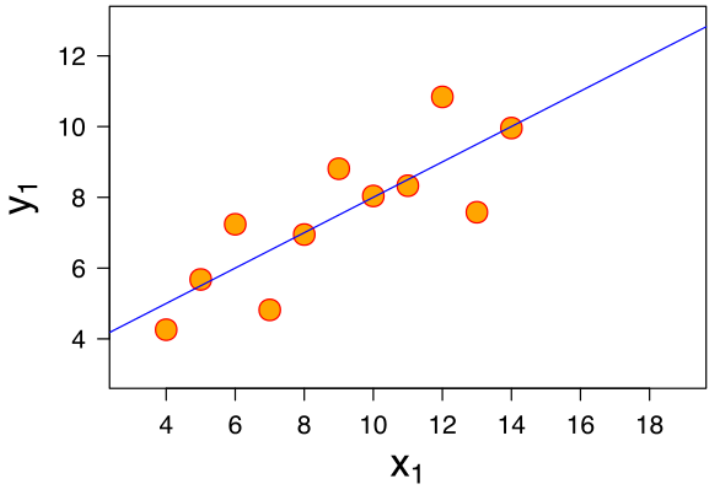
?

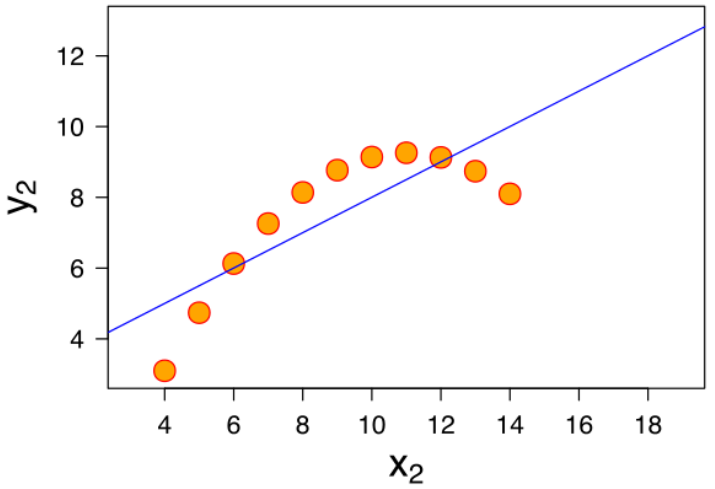
Fitting models to data

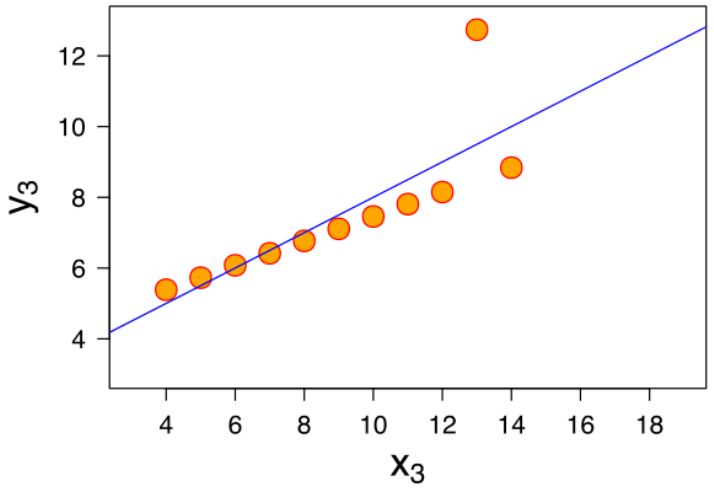


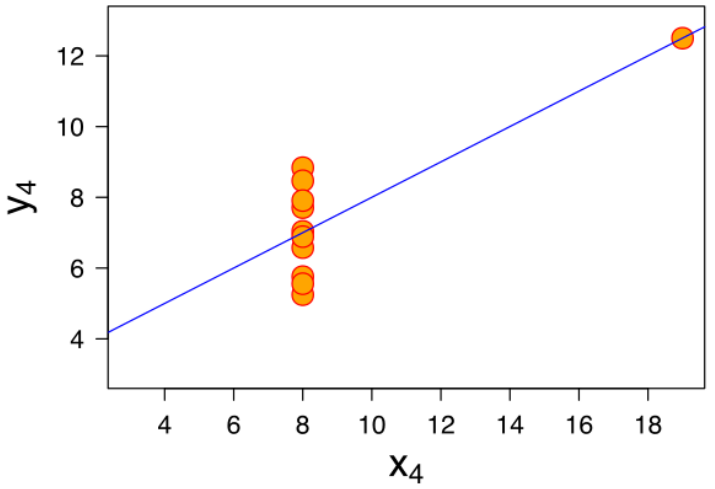
Fitting models to data









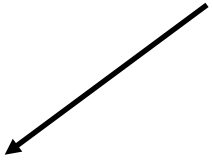


Linear models

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

Linear models

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

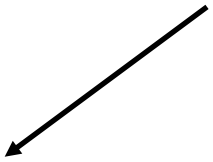


y
5
3
6
10
4

Data. Response variable,
e.g. sparrow body mass.

Linear models

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



y	i
5	1
3	2
6	3
10	4
4	5

Data. Response variable. Observation 1, 2, 3, etc.
e.g. sparrow body mass.

Linear models

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

y	i
5	1
3	2
6	3
10	4
4	5

Data. Response variable. Observation 1, 2, 3, etc.
e.g. sparrow body mass.

x	i
3	1
1	2
4	3
8	4
2	5

Data. Explanatory variable.
e.g. sparrow tarsus length.

Linear models

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

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models

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

$b_1 = ?$

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models

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

b_0 = ?

b_1 = ?

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models

- *Vectors* vs **scalars**

$$y_i = \boldsymbol{b_0} + \boldsymbol{b_1} x_i + \varepsilon_i$$

y	i
5	1
3	2
6	3
10	4
4	5

$b_0 = ?$

$b_1 = ?$

x	i
3	1
1	2
4	3
8	4
2	5

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models

$$y_i = \boldsymbol{b_0} + \boldsymbol{b_1} x_i + \boldsymbol{\varepsilon_i}$$

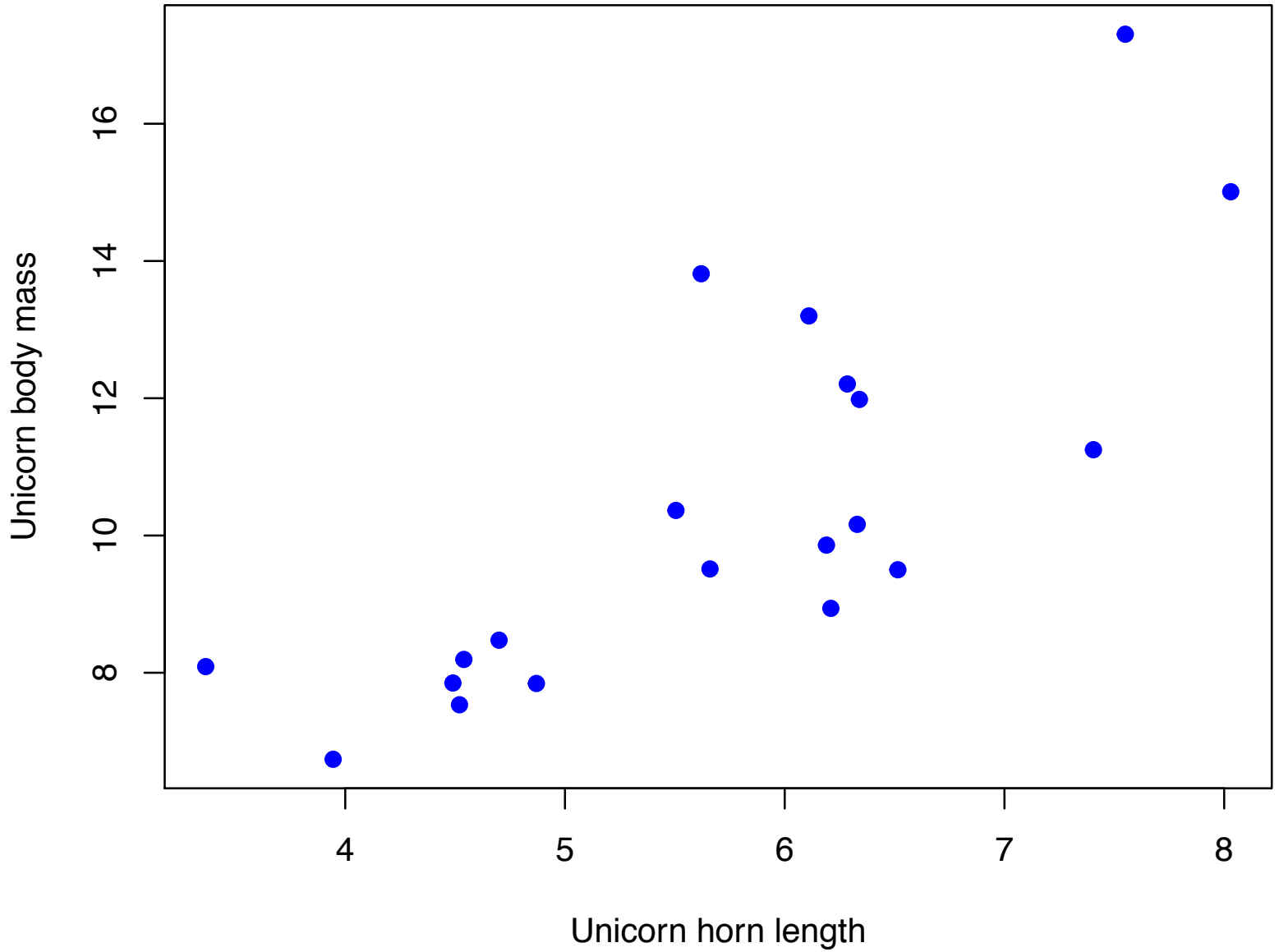
- Find solution: these parameter estimates (scalars) that minimise the left-over error residuals (vector)

Linear models - terminology

- Unicorn dataset – Runicorns.tx

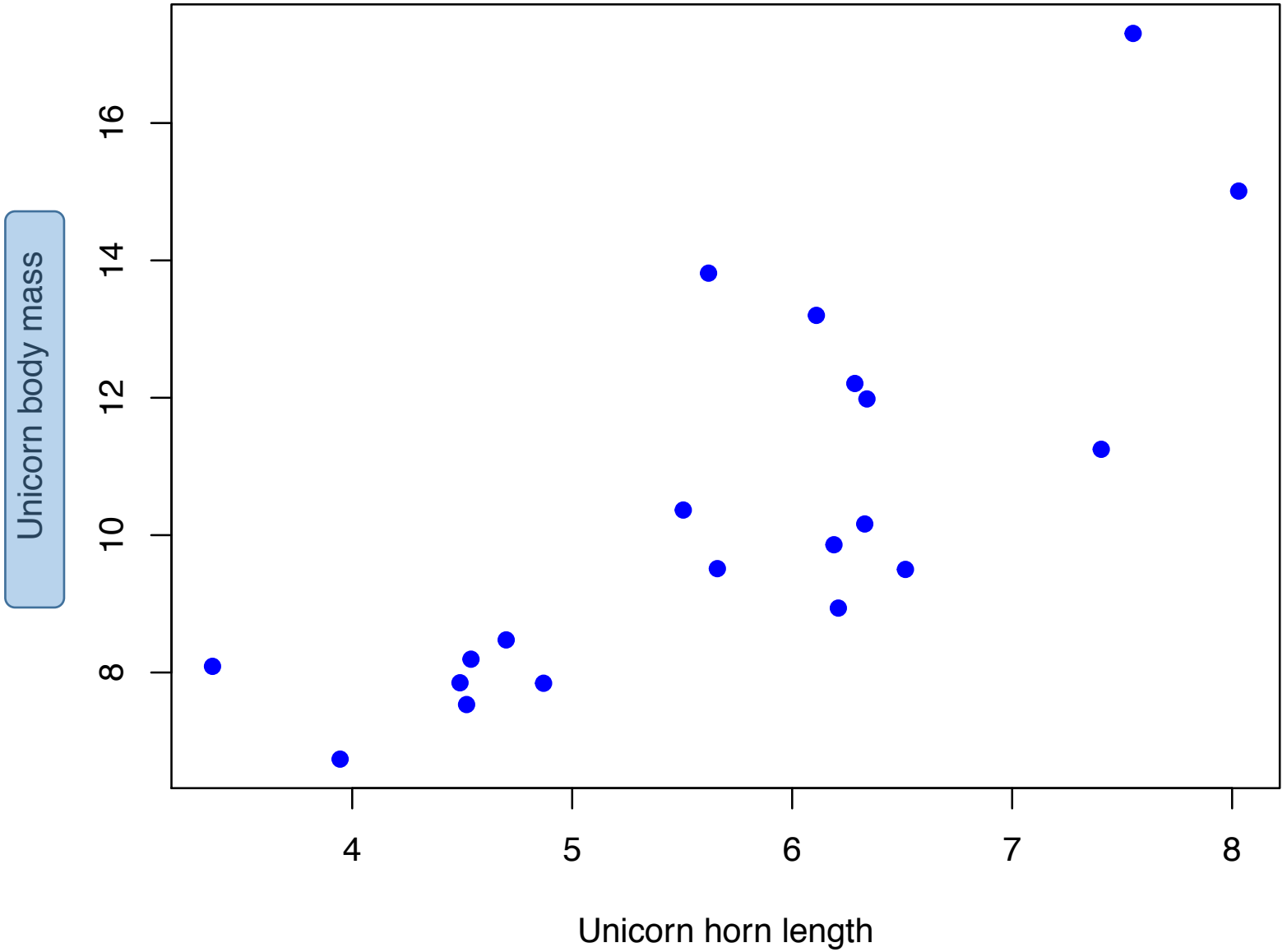
```
> head(uni)
                    Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941 Spring     0
4                Diane_Gumbo Female 10.365228      5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253 Autumn     1
>
```

Linear models - terminology



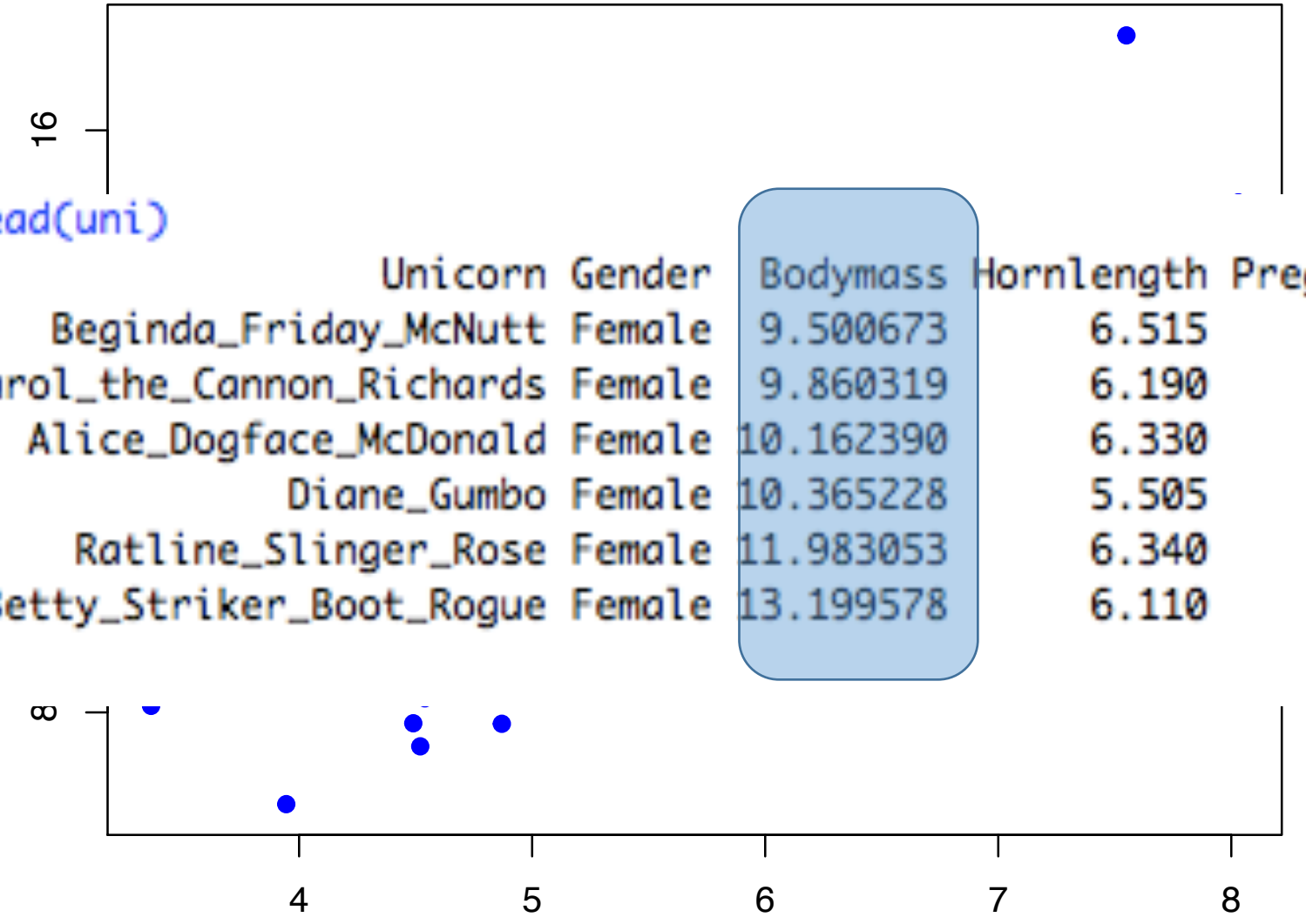
Linear models - terminology

Y = response variable, body mass



Linear models - terminology

Y = response variable, body mass

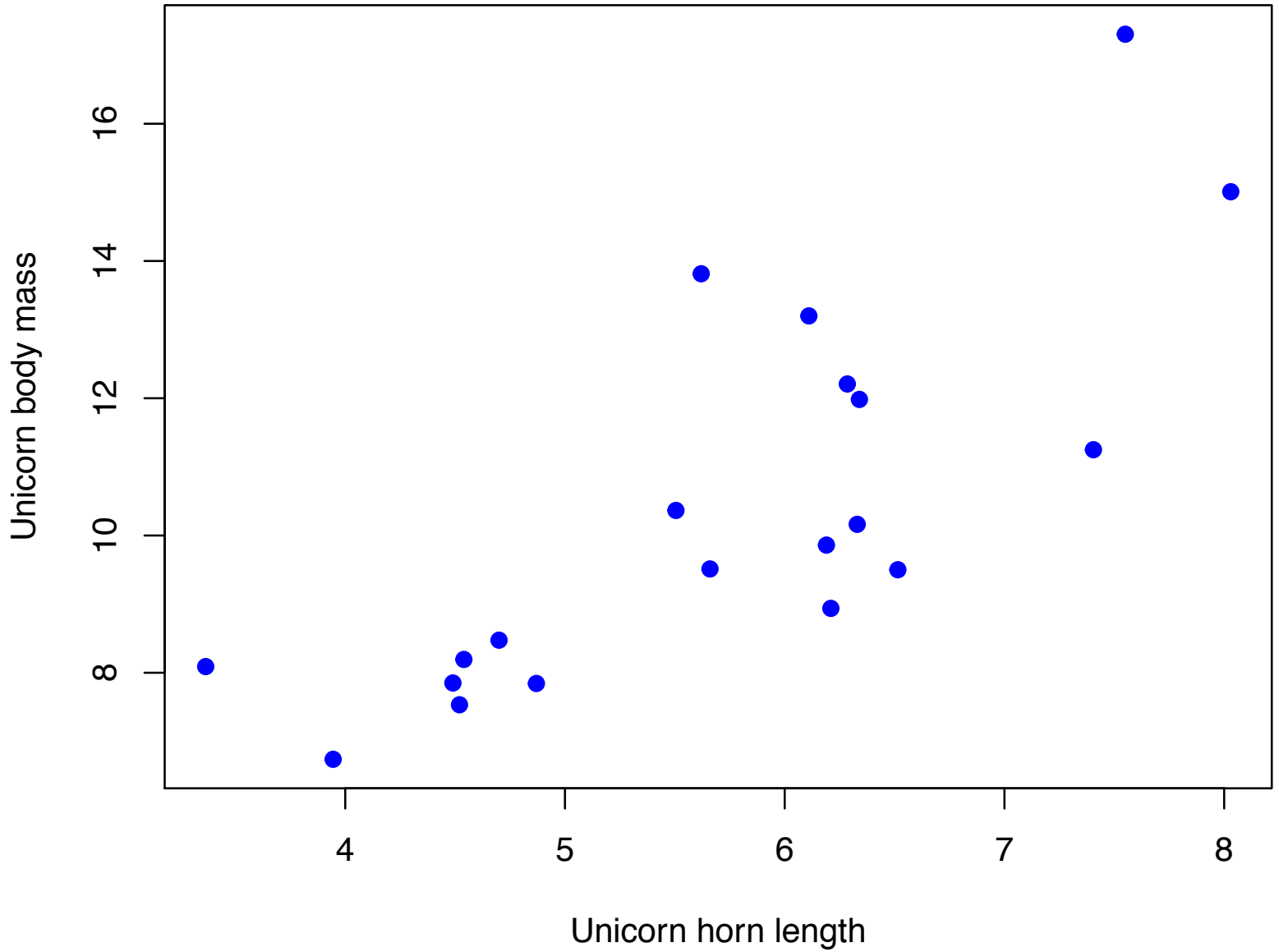



```
> head(uni)
                    Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4                Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

Linear models - terminology

Y = response variable, body mass

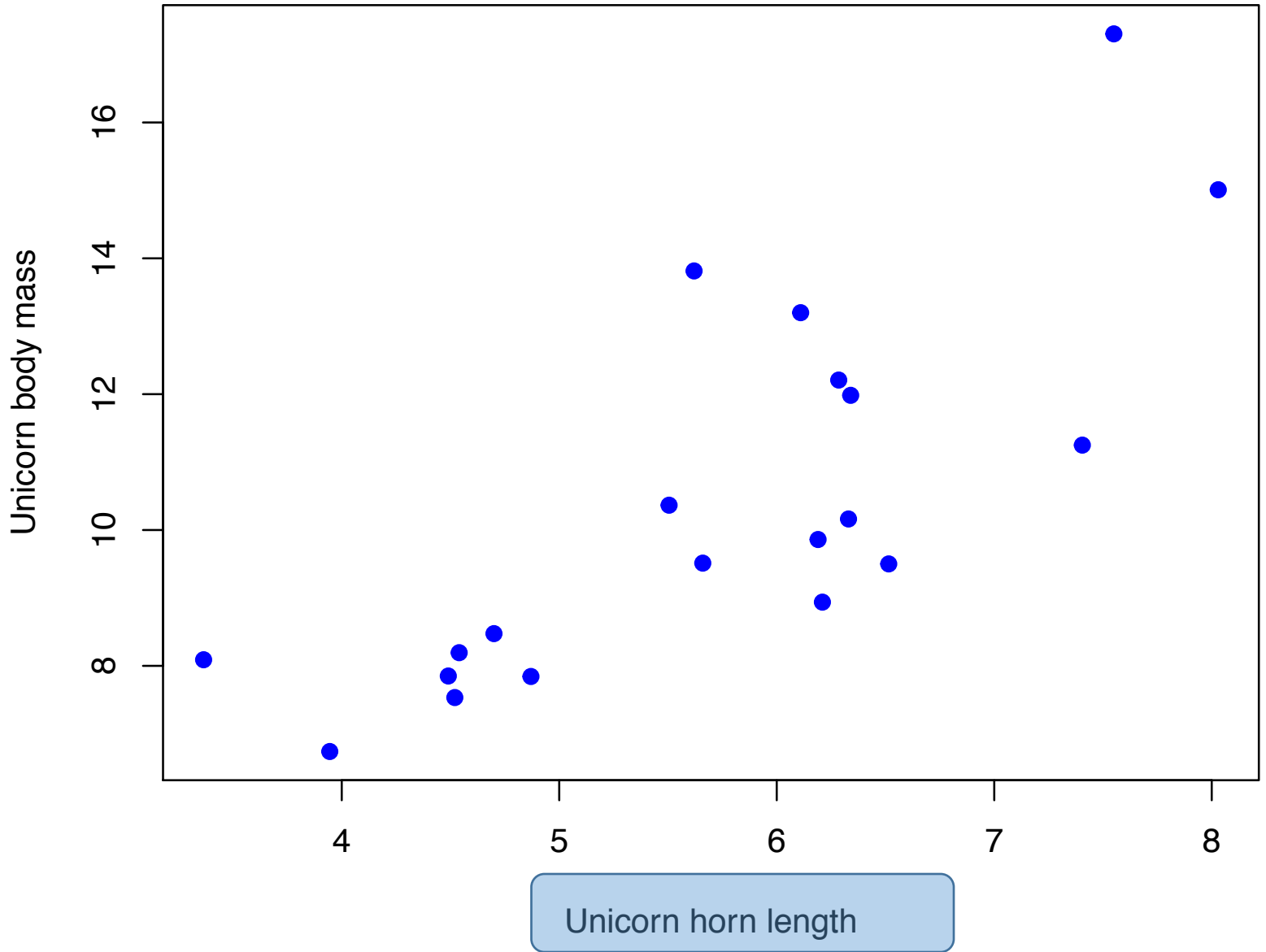
X = explanatory variable, horn length



Linear models - terminology

Y = response variable, body mass

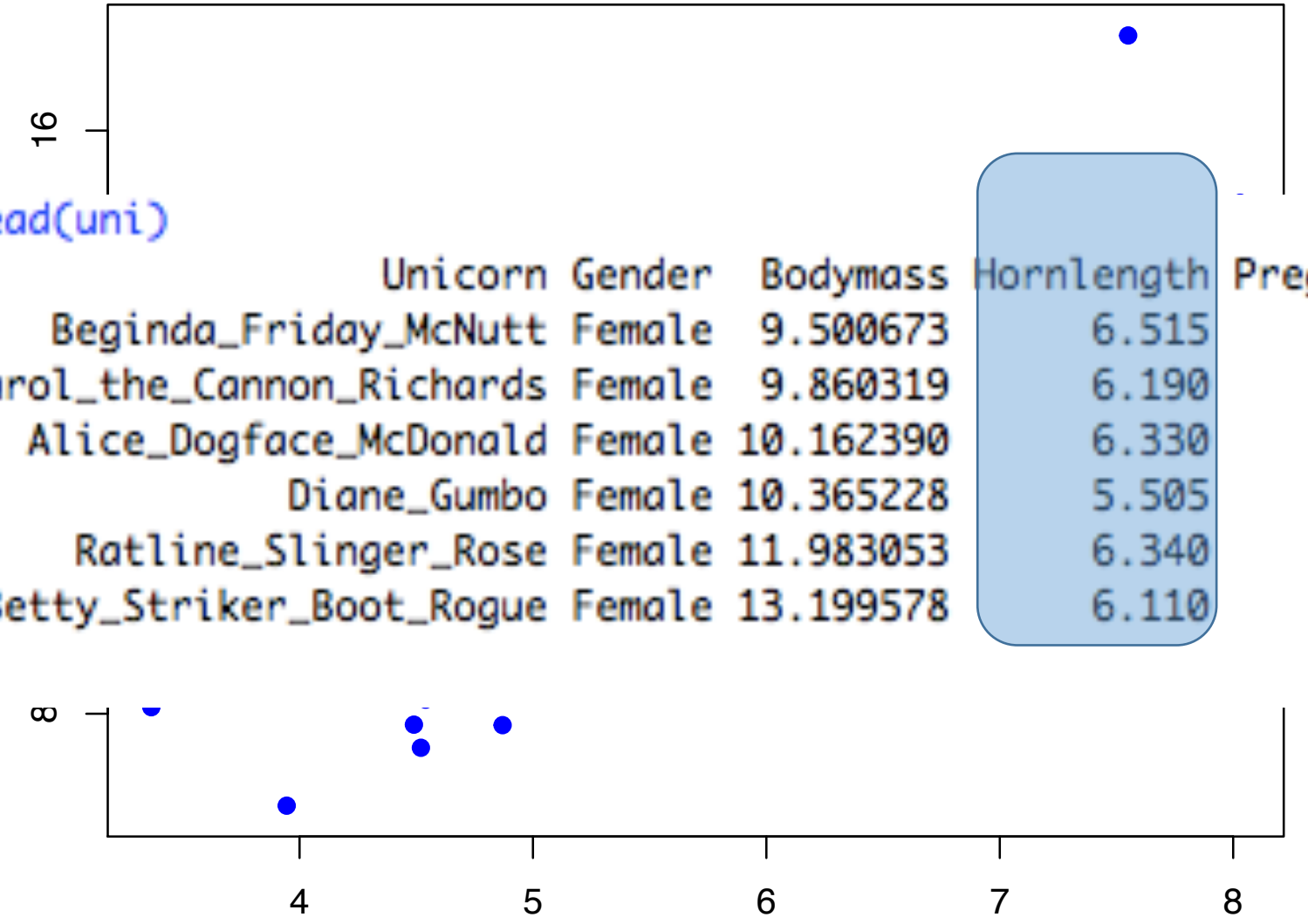
X = explanatory variable, horn length



Linear models - terminology

Y = response variable, body mass

X = explanatory variable, horn length




```
> head(uni)
                    Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941 Spring     0
4                Diane_Gumbo Female 10.365228      5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253 Autumn     1
>
```

Linear models

- *Vectors* vs **scalars**

$$y_i = \boldsymbol{b_0} + \boldsymbol{b_1} x_i + \varepsilon_i$$

y	i
5	1
3	2
6	3
10	4
4	5

$b_0 = ?$

$b_1 = ?$

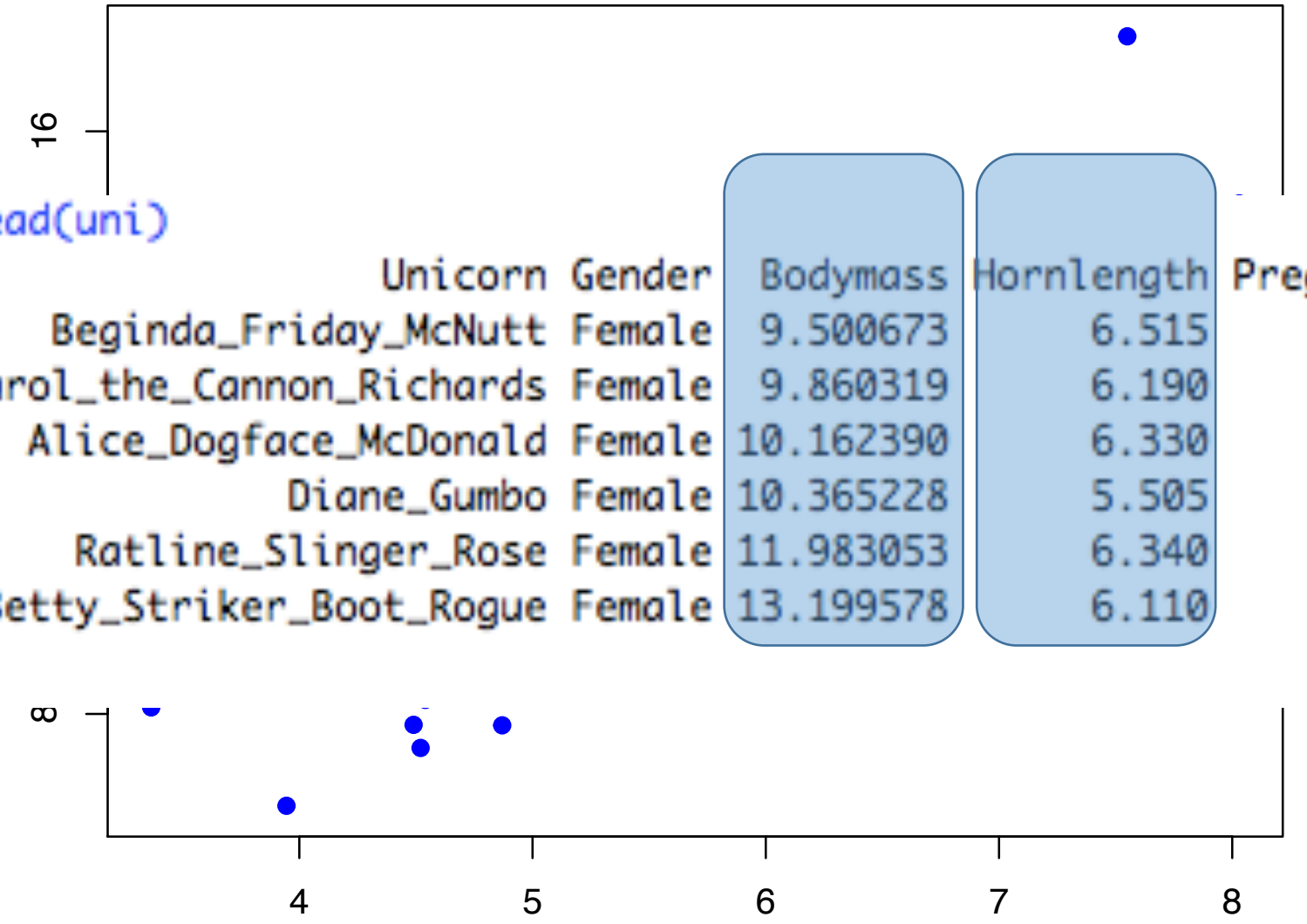
x	i
3	1
1	2
4	3
8	4
2	5

ε	i
?	1
?	2
?	3
?	4
?	5

Linear models - terminology

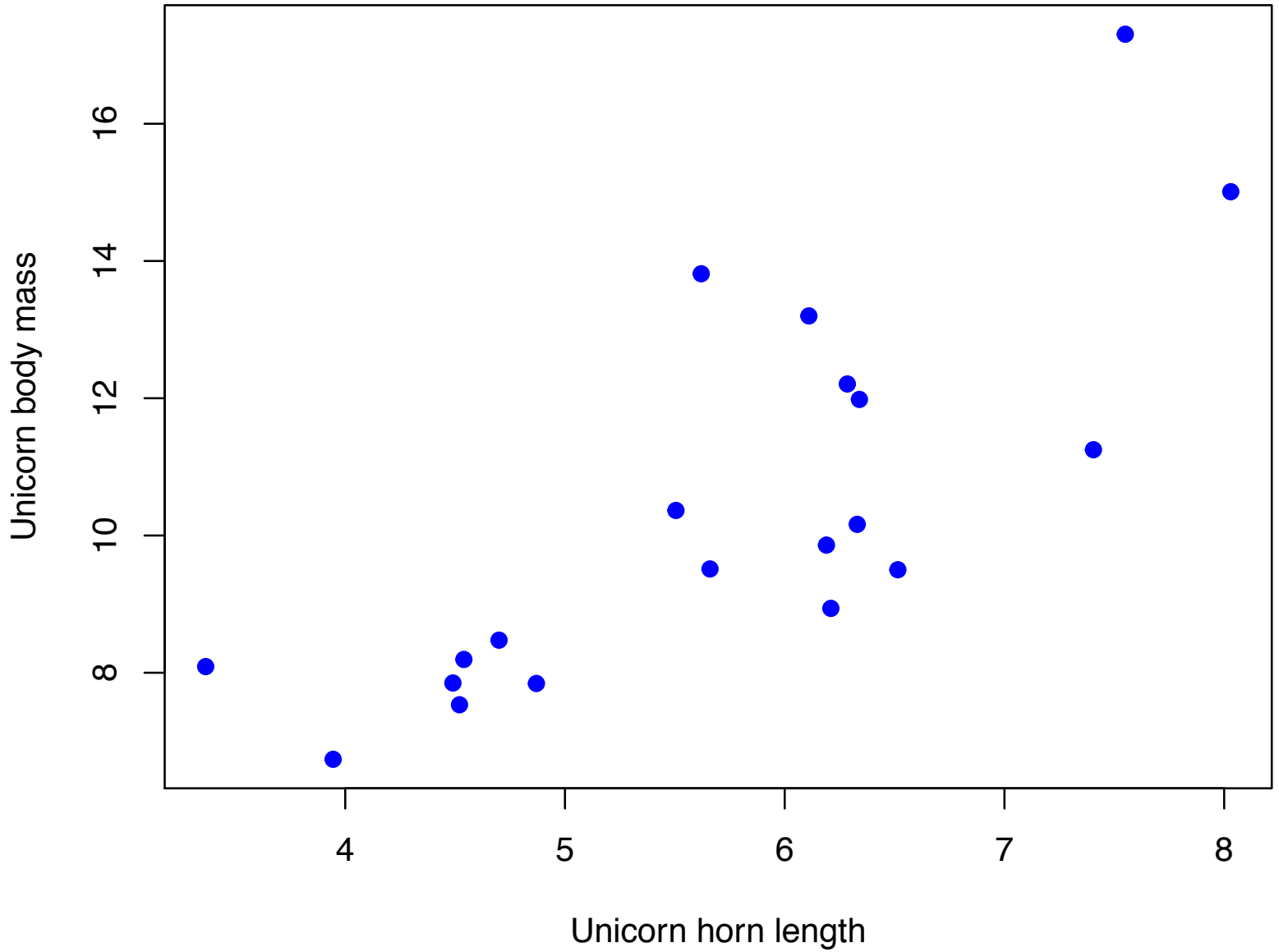
Y = response variable, body mass

X = explanatory variable, horn length




```
> head(uni)
                      Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4                Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

Linear models - terminology

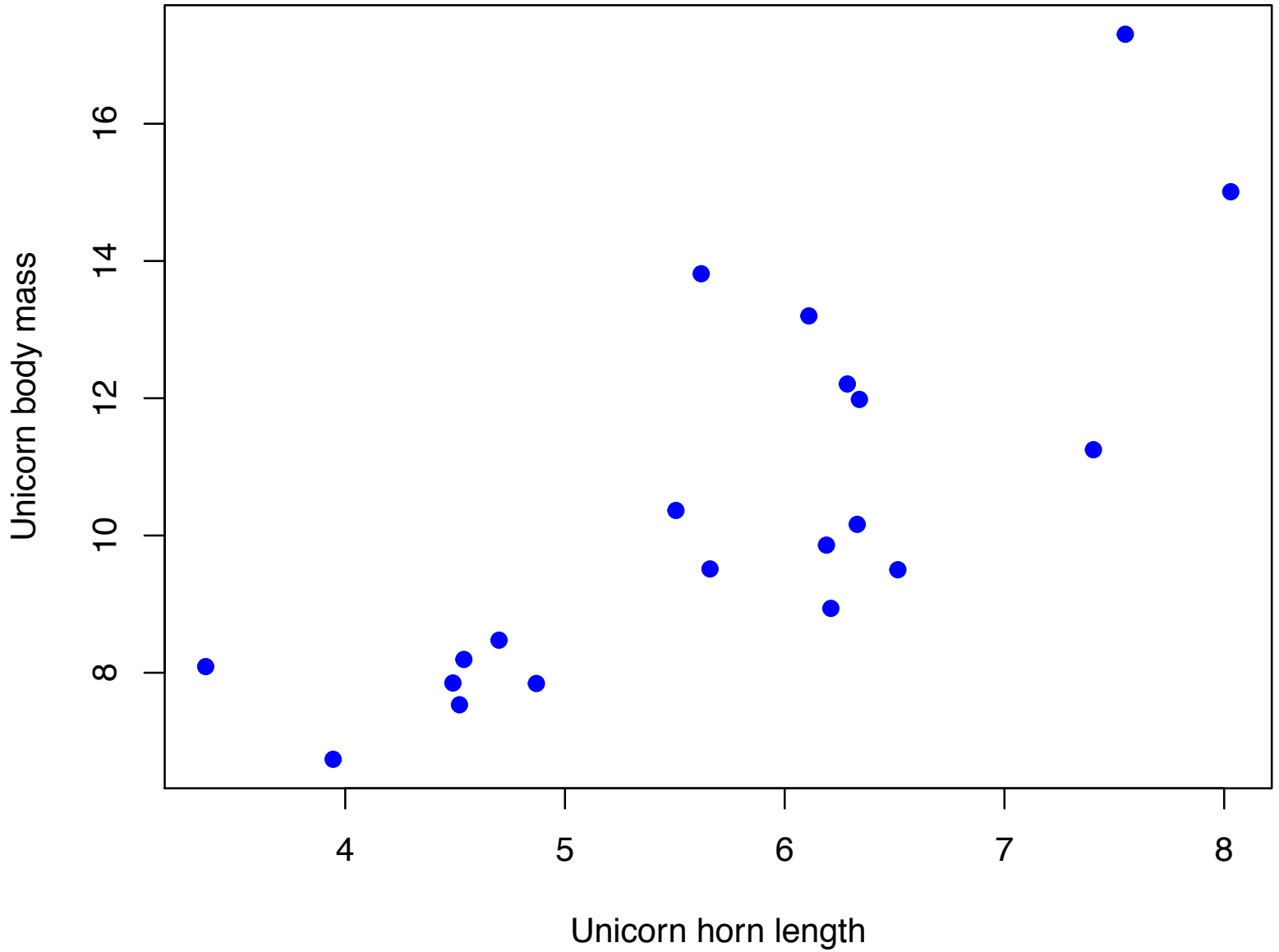


Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

Linear models - terminology



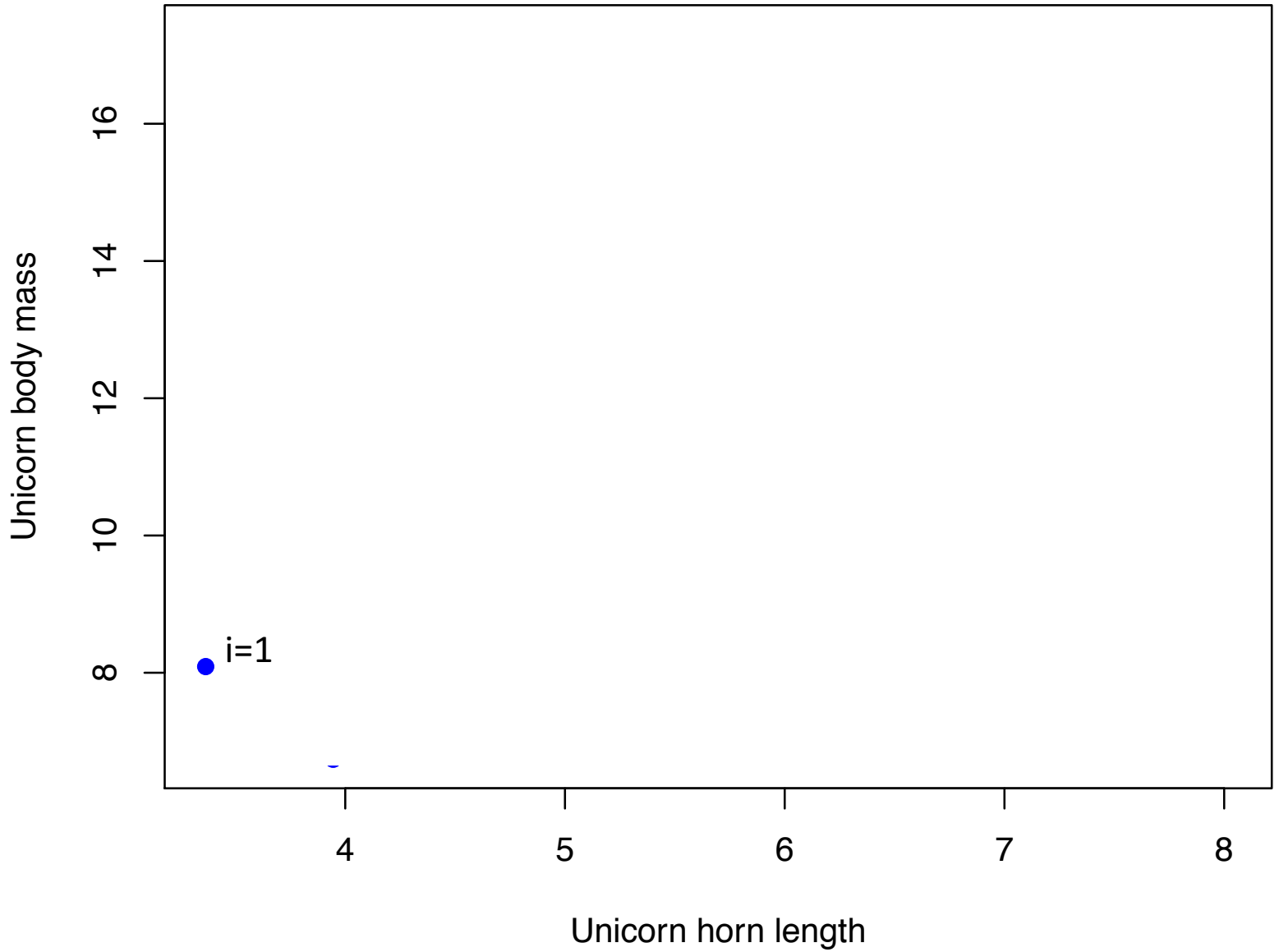
Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

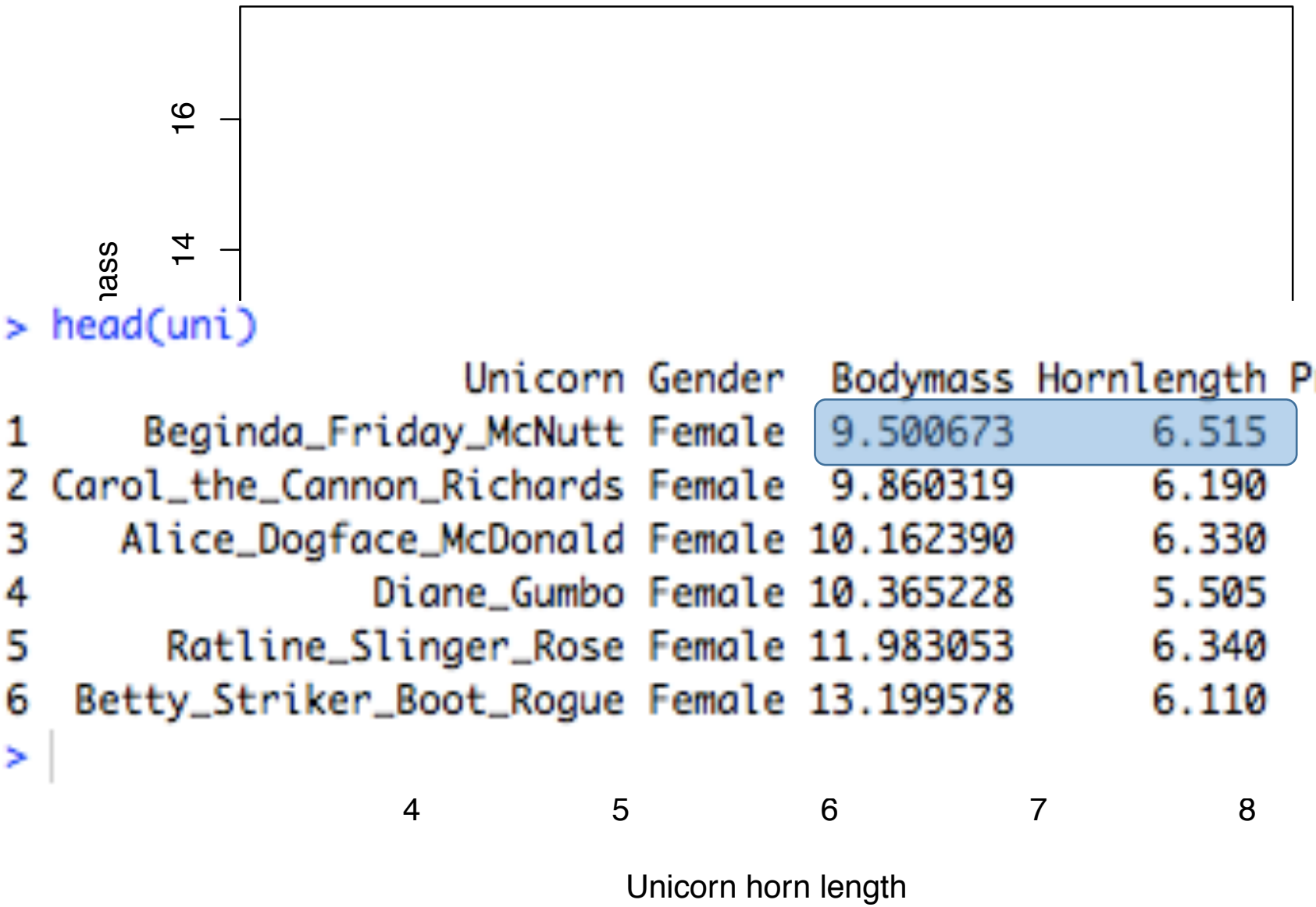
Linear models - terminology

Y = response variable, body mass

X = explanatory variable, horn length

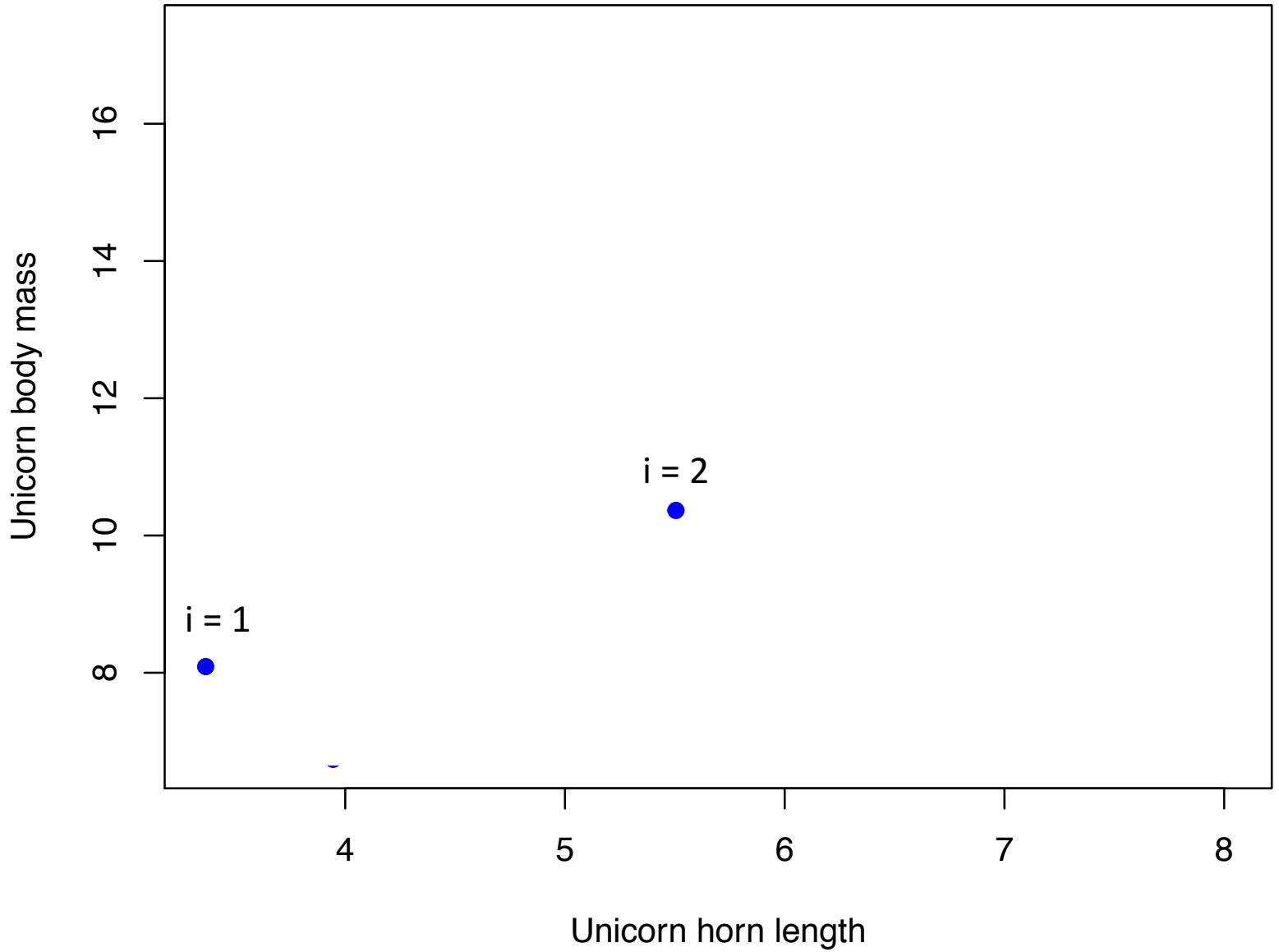
Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length




```
> head(uni)
                      Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4              Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

Linear models - terminology



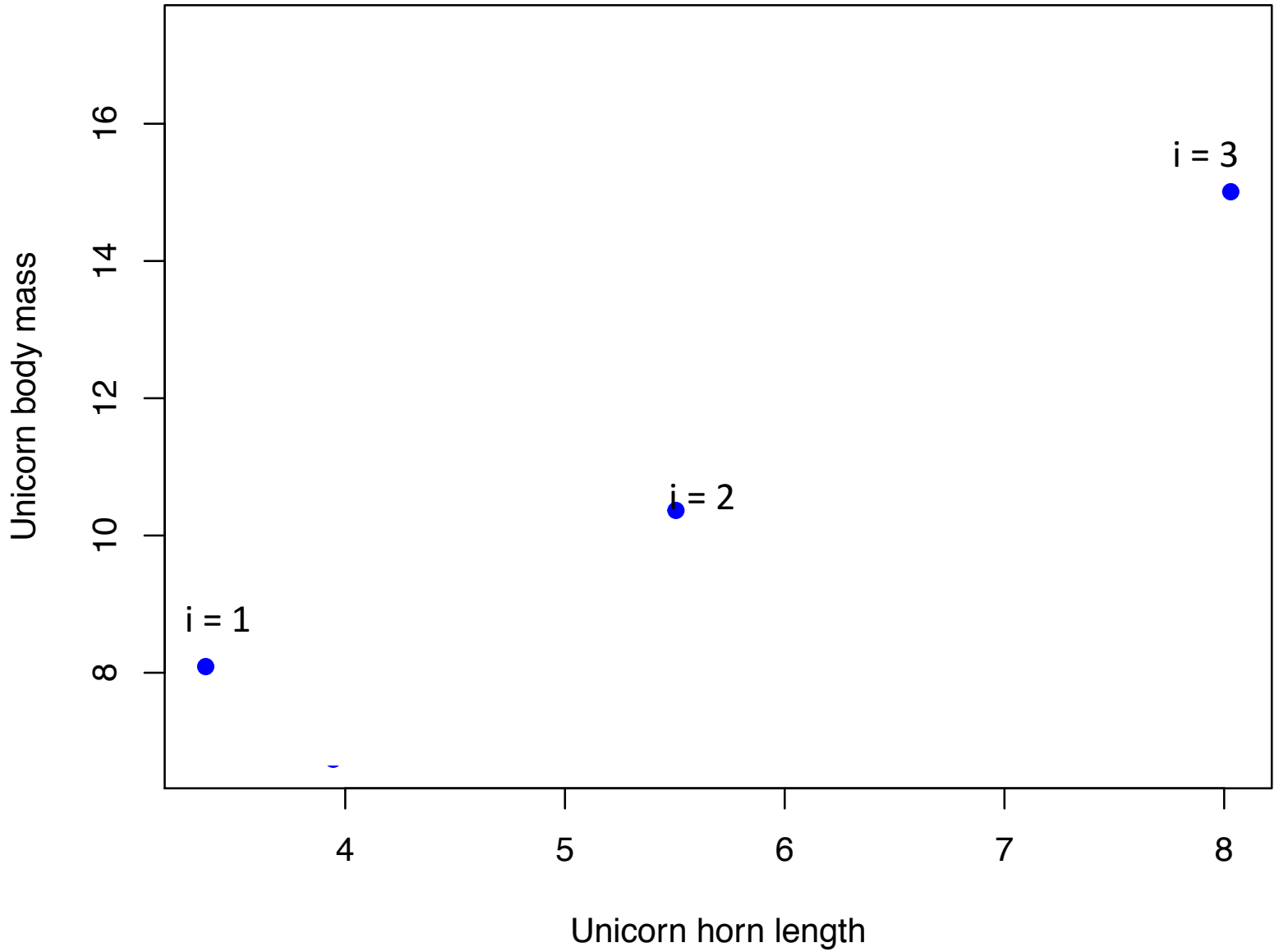
Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

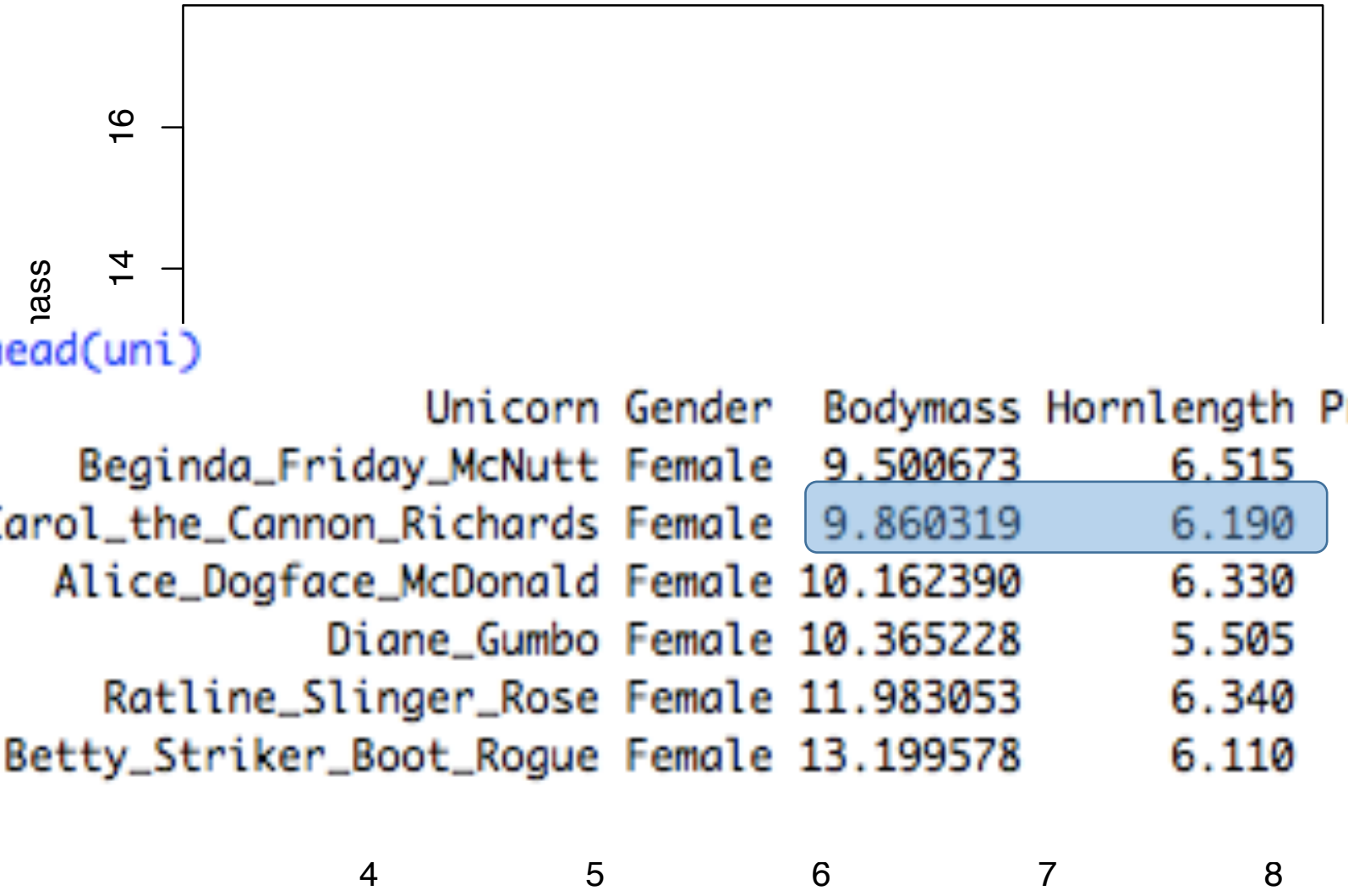
Linear models - terminology

Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length




```
> head(uni)
                      Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4              Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

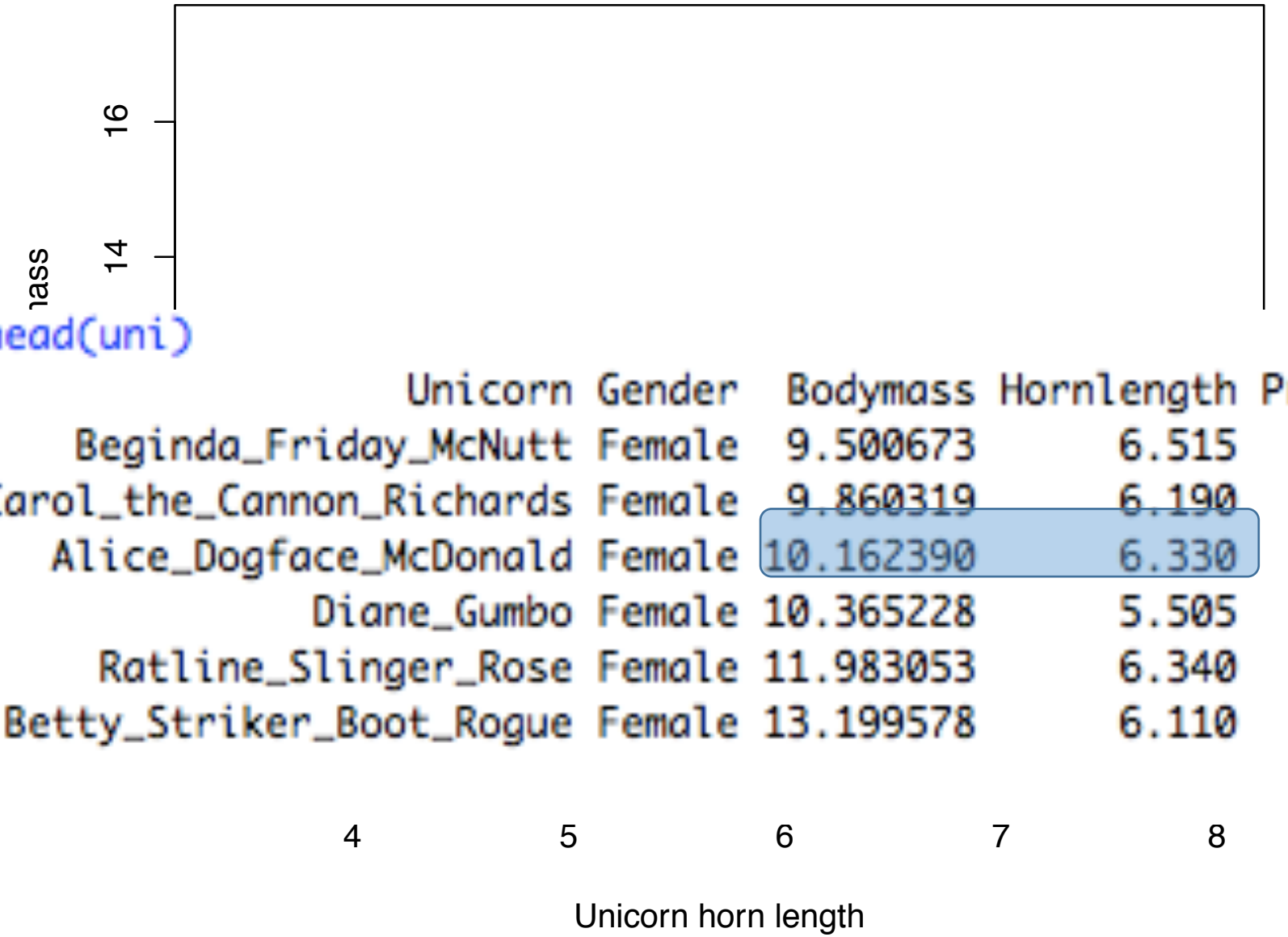
Linear models - terminology

Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length




```
> head(uni)
                      Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4              Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

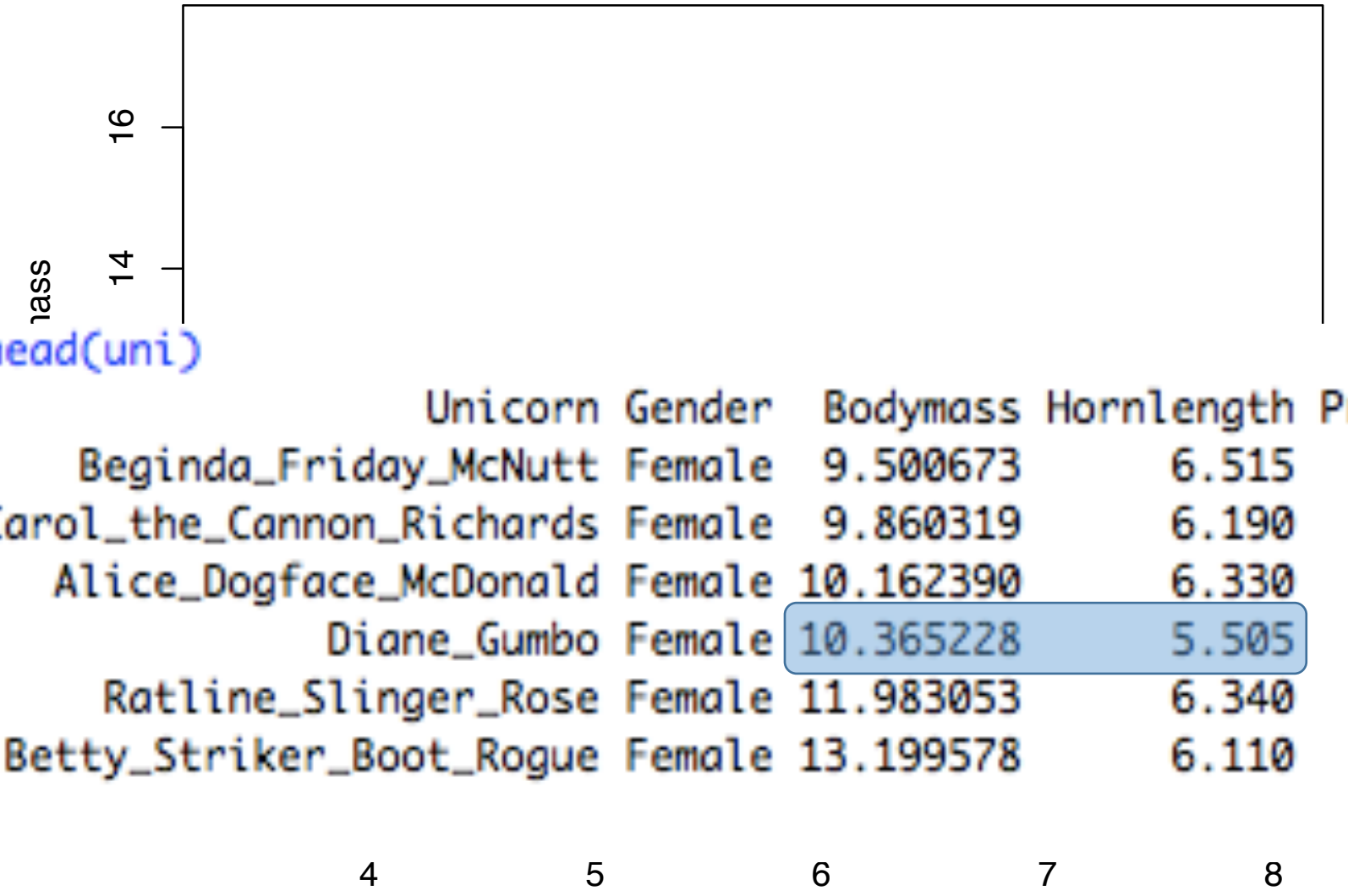
Linear models - terminology

Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length




```
> head(uni)
                      Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4              Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

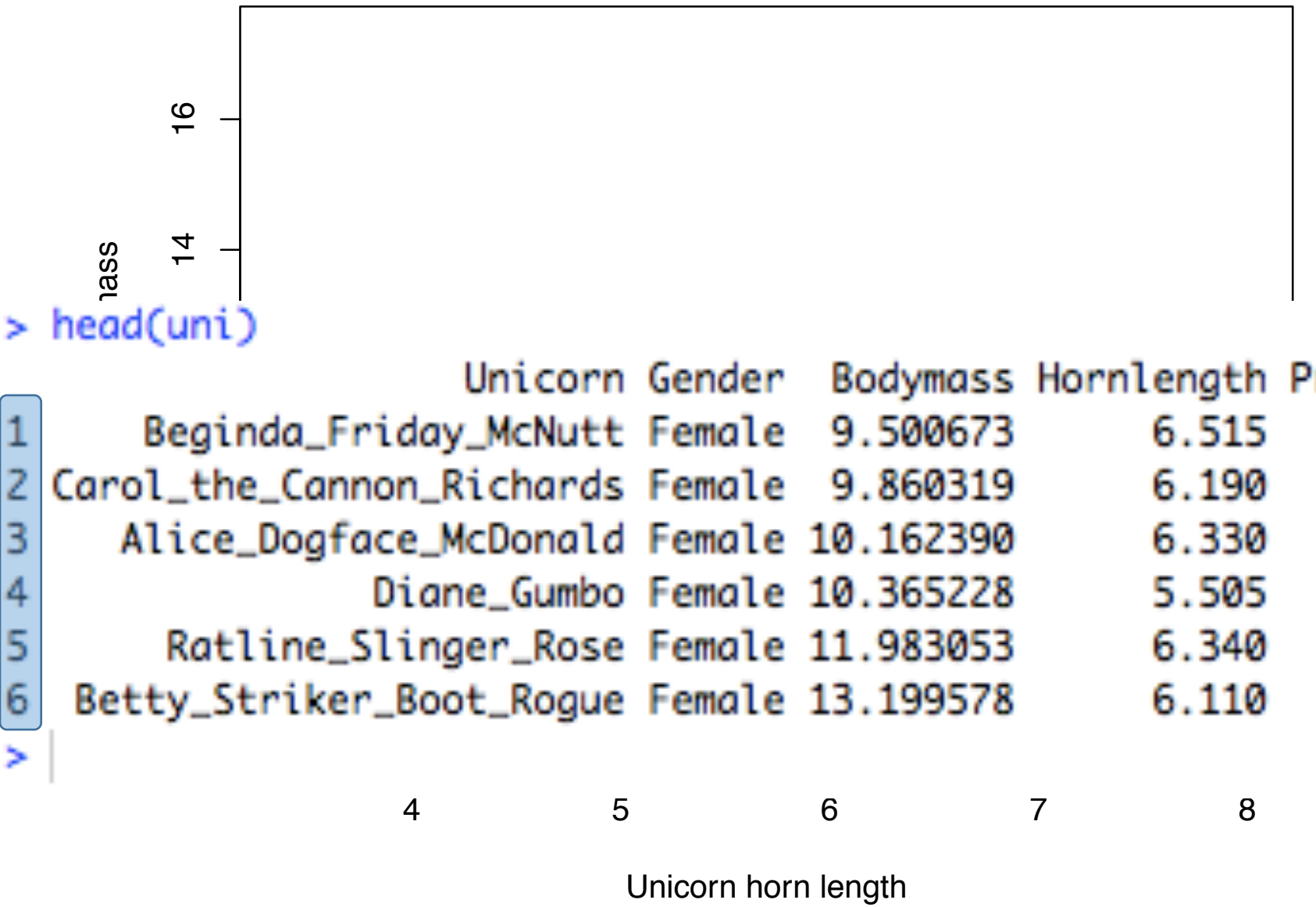
Linear models - terminology

Y = response variable, body mass

X = explanatory variable, horn length

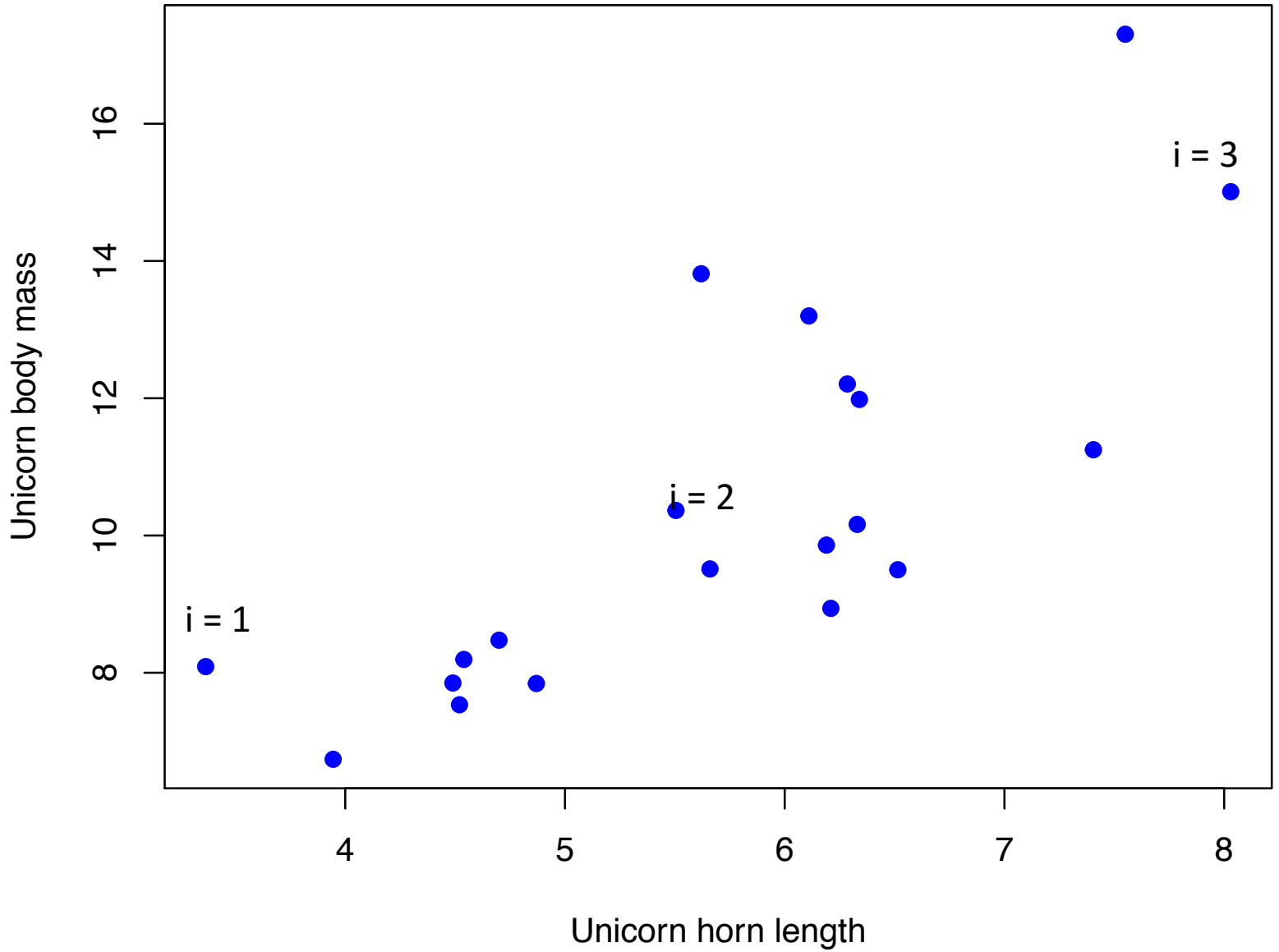
Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length




```
> head(uni)
                      Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673       6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319       6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390       6.330        0 8.190941 Spring     0
4              Diane_Gumbo Female 10.365228       5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053       6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578       6.110        0 1.084253 Autumn     1
>
```

Linear models - terminology



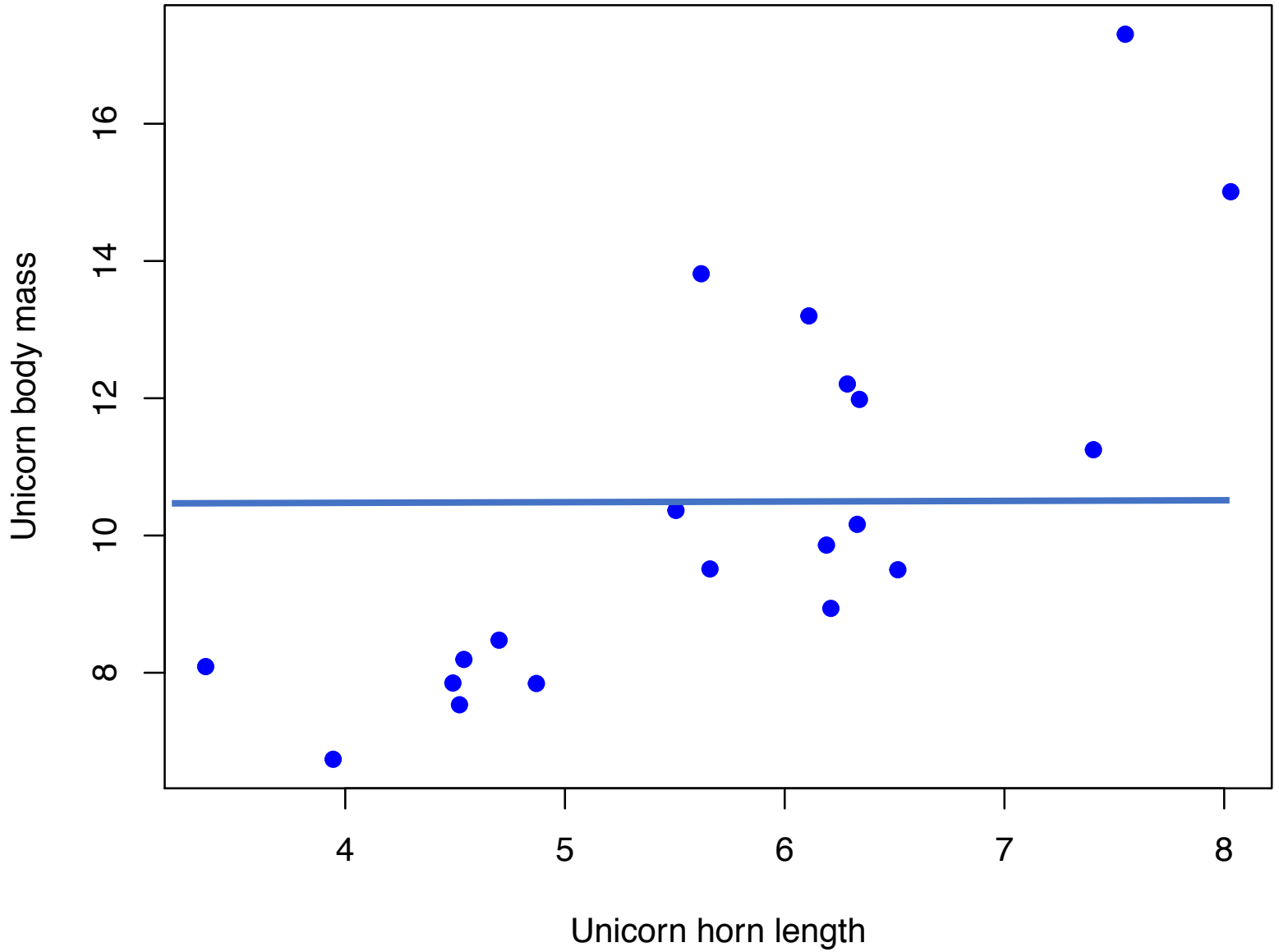
Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

Linear models - terminology



Y = response variable, body mass

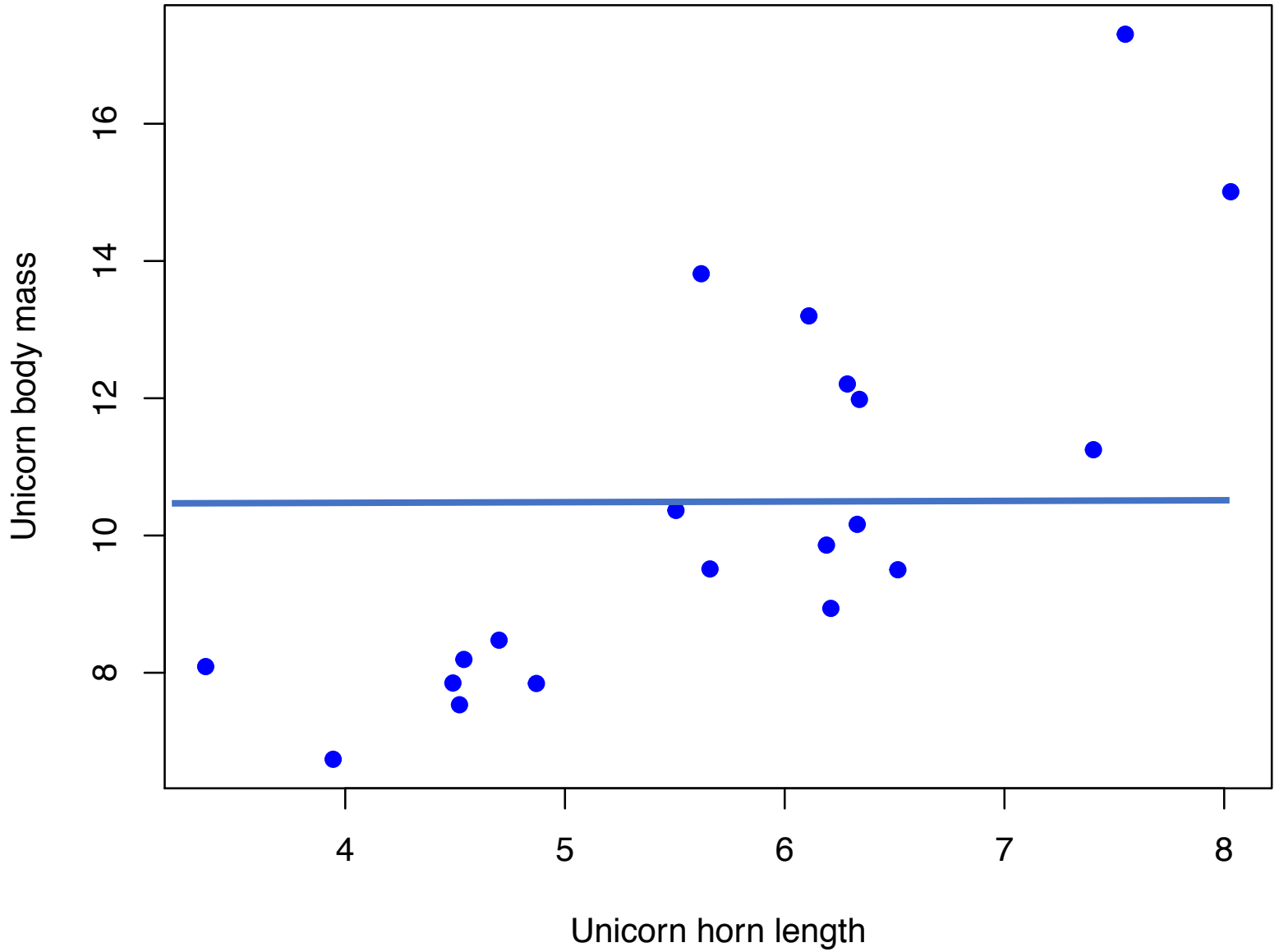
X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

```
> head(uni)
                    Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941 Spring     0
4               Diane_Gumbo Female 10.365228      5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253 Autumn     1
>
```

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

$b_0 = ?$

$b_1 = ?$

ε	i
?	1
?	2
?	3
?	4
?	5

```
> head(uni)
                    Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941 Spring     0
4                Diane_Gumbo Female 10.365228      5.505        0 1.584386 Autumn     1
5       Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253 Autumn     1
>
```

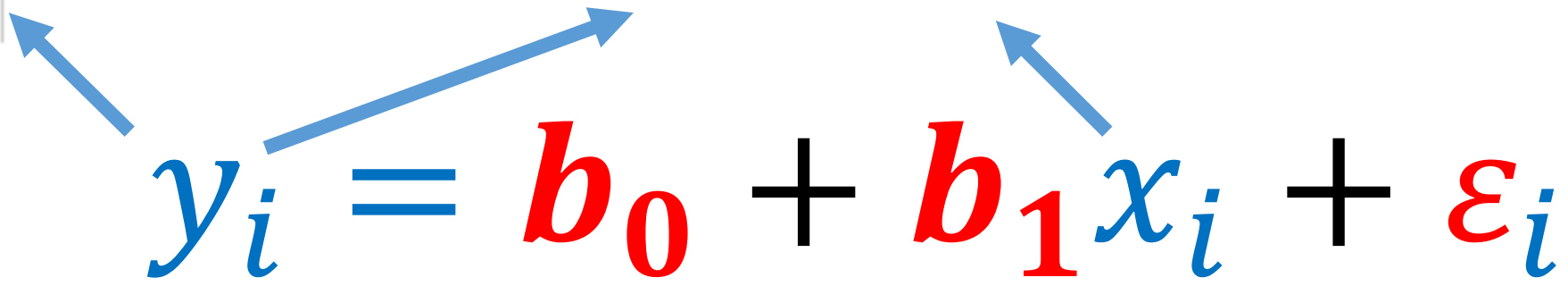
$$y_i = b_0 + b_1 x_i + \varepsilon_i$$

$b_0 = ?$

$b_1 = ?$

ε	i
?	1
?	2
?	3
?	4
?	5

```
> head(uni)
                     Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941 Spring     0
4               Diane_Gumbo Female 10.365228      5.505        0 1.584386 Autumn     1
5      Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253 Autumn     1
>
```



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

$b_0 = ?$

$b_1 = ?$

ε	i
?	1
?	2
?	3
?	4
?	5

```
> head(uni)
                    Unicorn Gender  Bodymass Hornlength Pregnant   Height Season Glizz
1     Beginda_Friday_McNutt Female  9.500673      6.515        0 3.758080 Spring     0
2 Carol_the_Cannon_Richards Female  9.860319      6.190        0 1.558938 Autumn     0
3    Alice_Dogface_McDonald Female 10.162390      6.330        0 8.190941 Spring     0
4                Diane_Gumbo Female 10.365228      5.505        0 1.584386 Autumn     1
5       Ratline_Slinger_Rose Female 11.983053      6.340        0 4.287208 Spring     1
6  Betty_Striker_Boot_Rogue Female 13.199578      6.110        0 1.084253 Autumn     1
>
```

ε

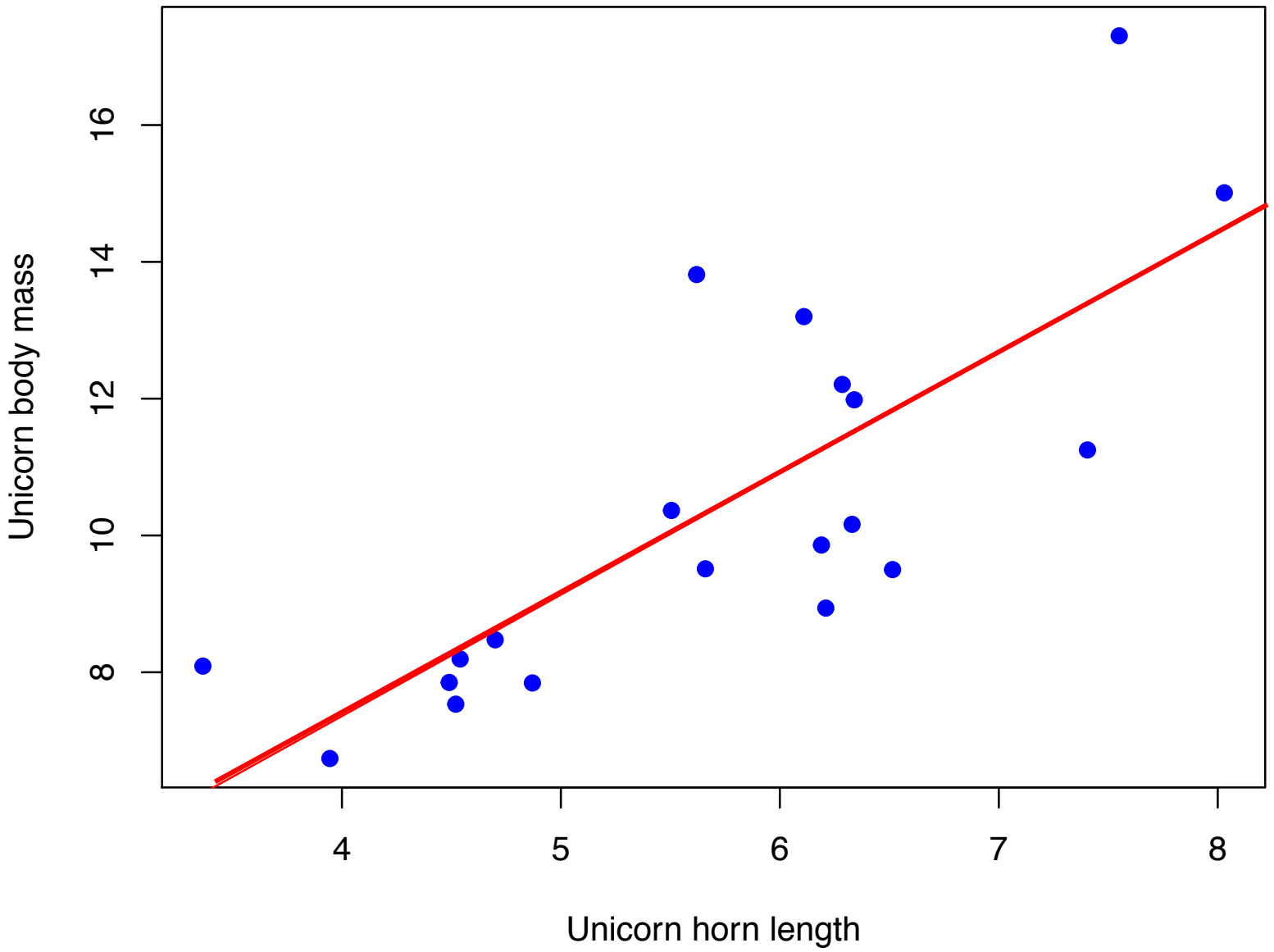
? ? ? ? ? ?

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

$b_0 = ?$

$b_1 = ?$

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

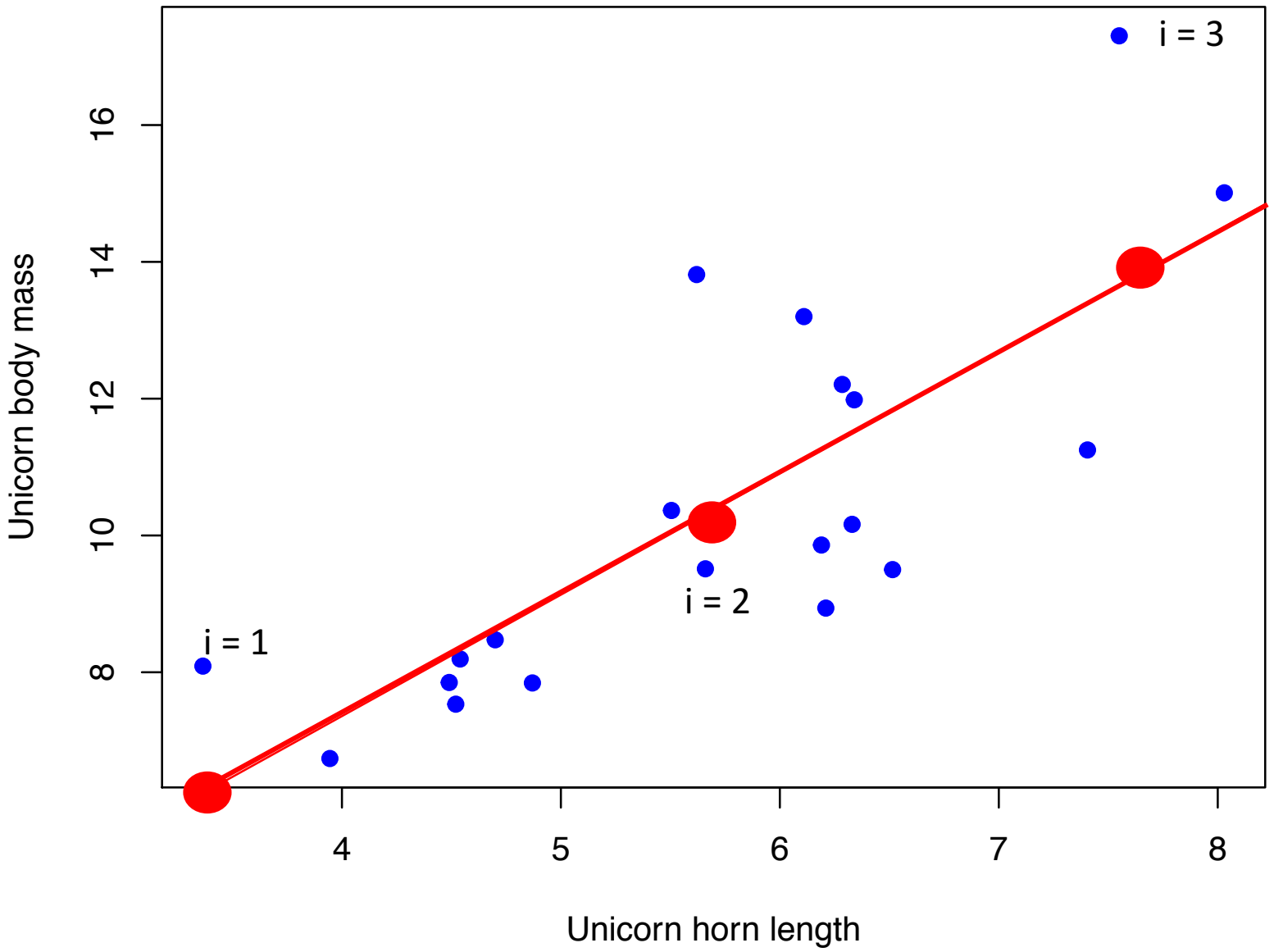
X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

b_0 and b_1, together with X_i, describe $\widehat{Y_i}$

$\widehat{Y_i}$ = PREDICTED

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

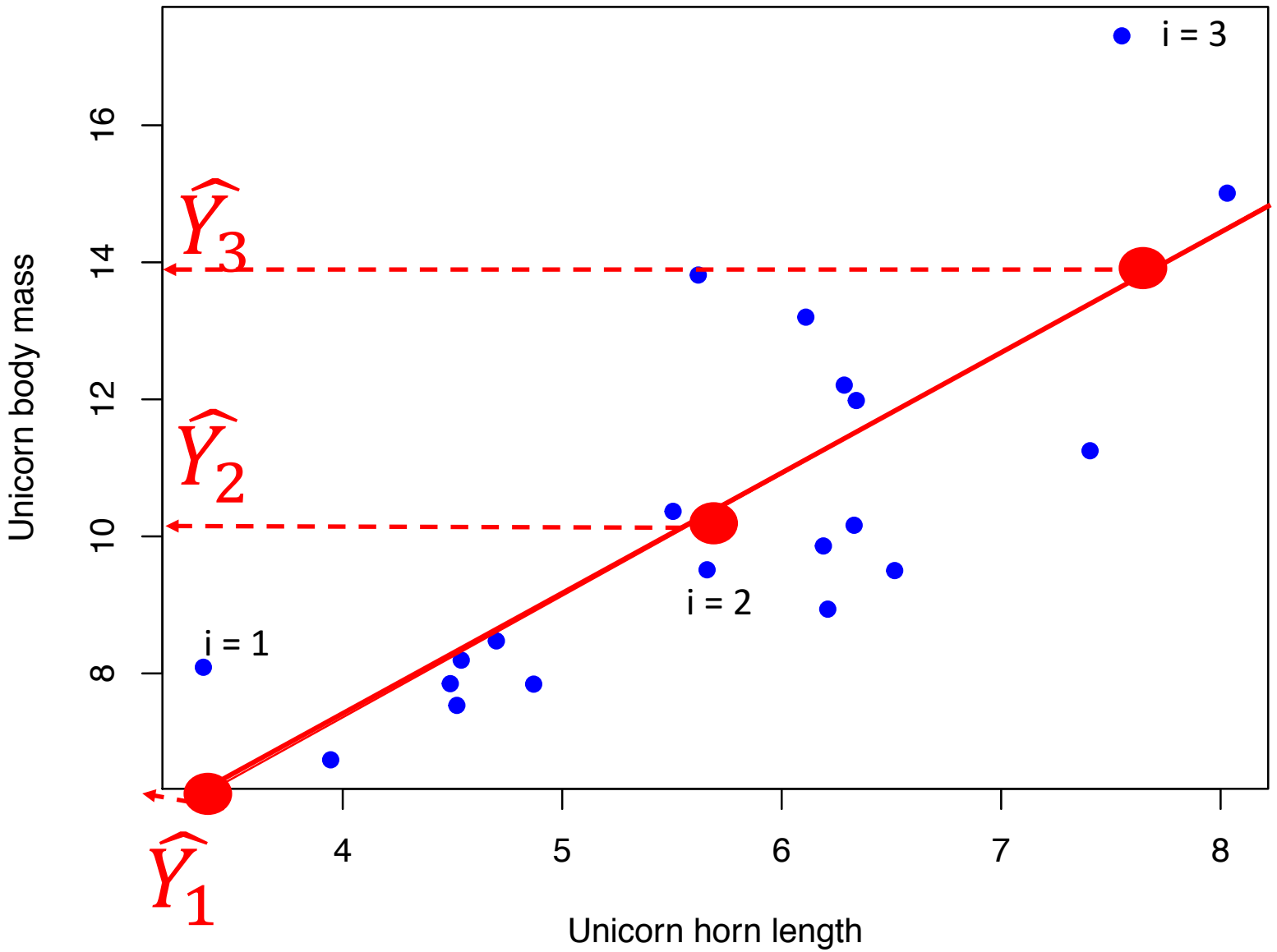
X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

b_0 and b_1, together with X_i, describe $\widehat{Y_i}$

$\widehat{Y_i}$ = PREDICTED

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

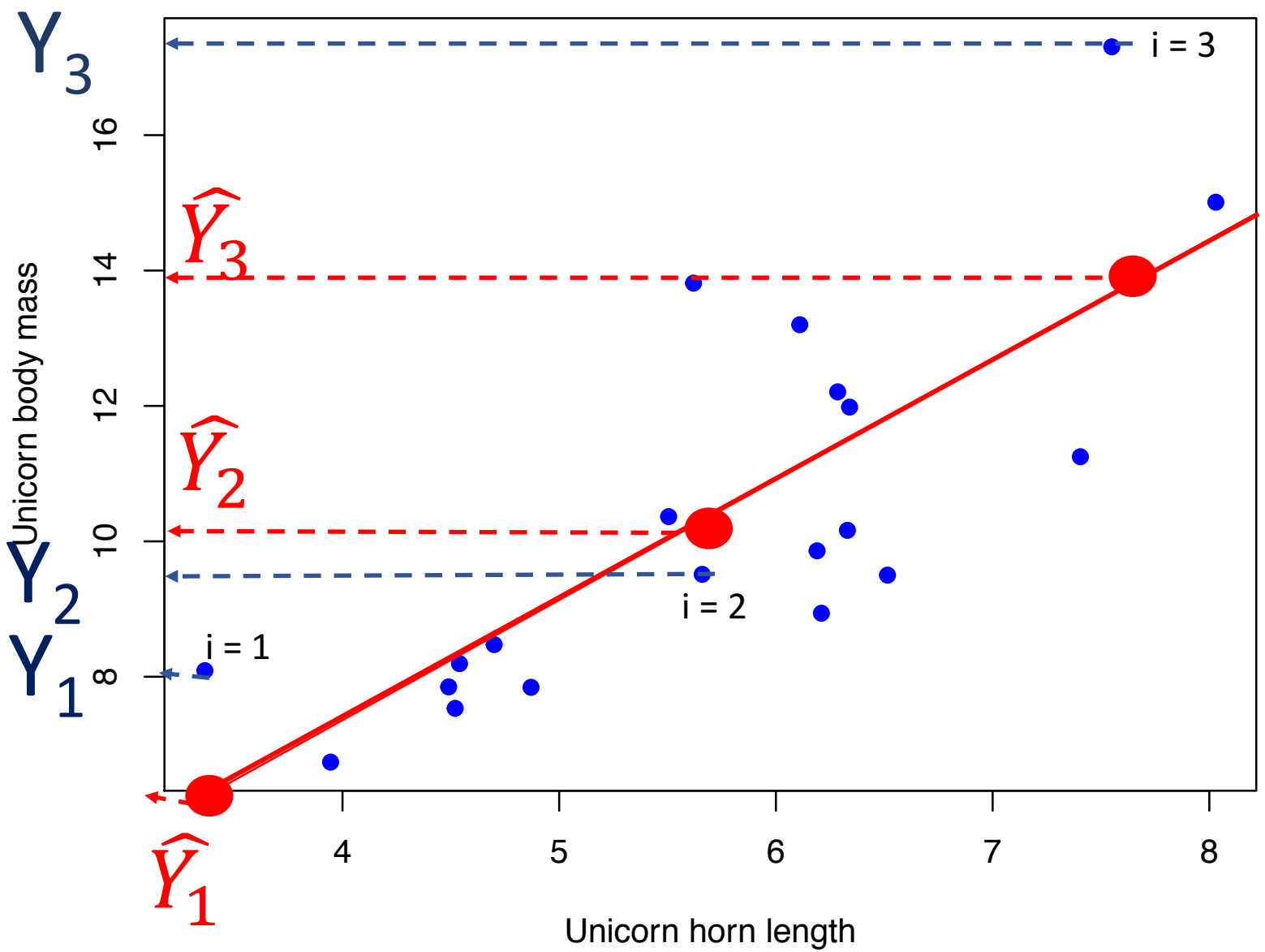
X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

b_0 and b_1, together with X_i, describe $\widehat{Y_i}$

$\widehat{Y_i}$ = PREDICTED

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

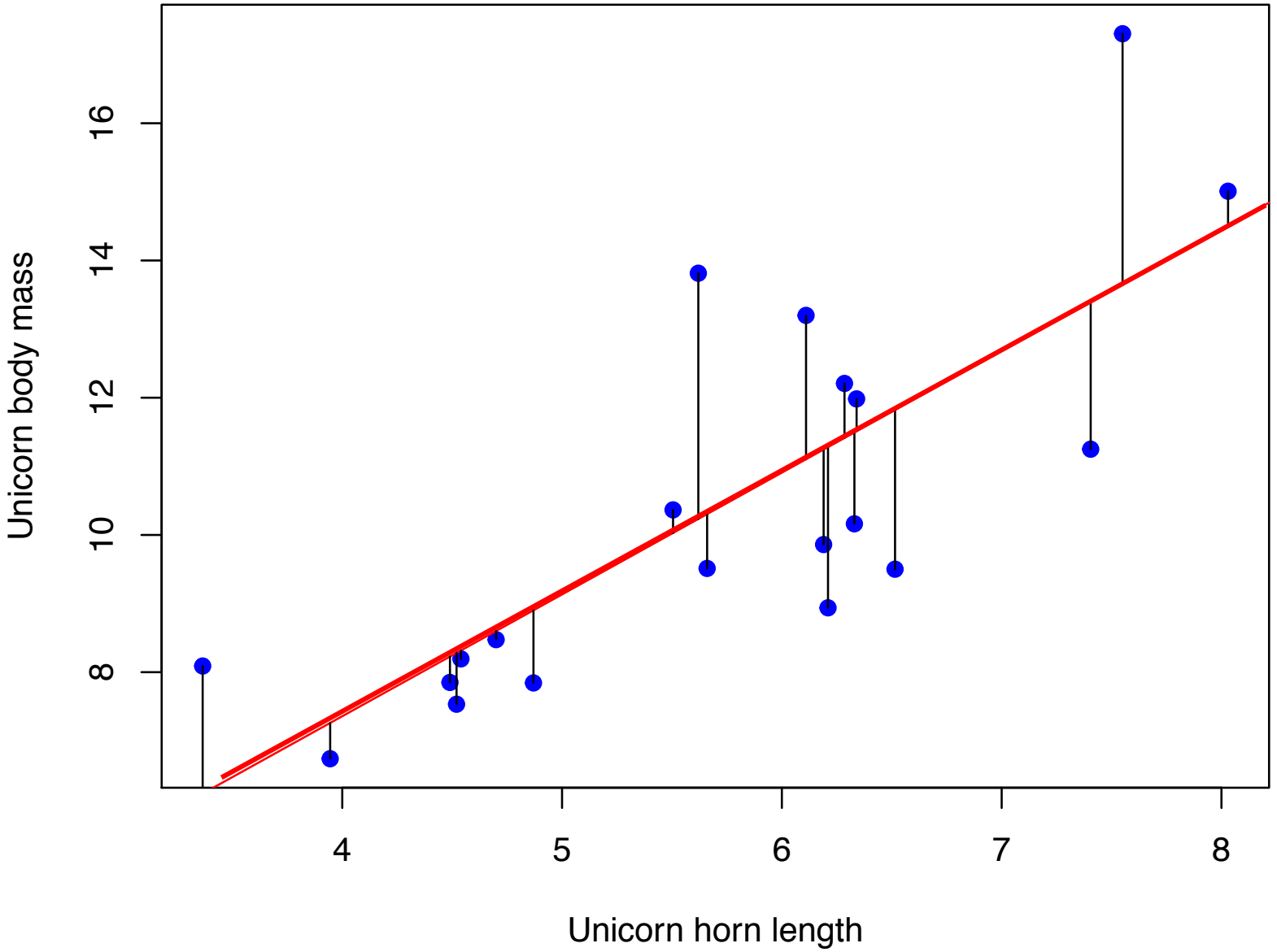
X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

b_0 and b_1, together with X_i, describe $\hat{Y}_i$

$\hat{Y}_i$ = PREDICTED

Linear models - terminology



Y = response variable, body mass

X = explanatory variable, horn length

Y_i = OBSERVATIONS of body mass

X_i = OBSERVATIONS of horn length

$\overline{Y_i}$= EXPECTED

b_0 and b_1, together with X_i, describe $\widehat{Y_i}$

$\widehat{Y_i}$ = PREDICTED

ε_i = ERROR, residuals

Linear model predictors

- Continuous

Linear model predictors

- Continuous
- Categorical

Linear model predictors

- Continuous
- Categorical

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

y	i
5	1
3	2
6	3
10	4
4	5

$b_0 = ?$

$b_1 = ?$

x	i
blue	1
red	2
blue	3
blue	4
yellow	5

ε	i
?	1
?	2
?	3
?	4
?	5

Linear model predictors

- Continuous
- Categorical

- → GENERAL linear models

Linear model predictors

- Continuous
- Categorical

- → GENERAL linear models

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \cdots + \varepsilon_i$$

Linear model predictors

- Continuous
- Categorical

- → GENERAL linear models

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \cdots + \varepsilon_i$$

Linear model predictors

- Continuous
- Categorical
- Interactions between variables

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

Linear model predictors

- Continuous
- Categorical
- Interactions between variables
- Squared terms

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i1}^2 + b_3 x_{i2} + b_4 x_{i1} x_{i2} + \varepsilon_i$$

Process

1. Outliers?

Process

1. Outliers?
2. Homogeneity of variances?

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation

Process

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation

Process – common problems

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation

What to include?

What to include?

What to include?

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?

$Body\ mass \sim Hornlength$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?

$$Body\ mass \sim Hornlength$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?

$$Body\ mass \sim Hornlength$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?
  - Food availability (time of day, day of season, annual variation in food = year)

$$Body\ mass \sim Hornlength$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?
  - Food availability (time of day, day of season, annual variation in food = year)
  - Age?

$$Body\ mass \sim Hornlength$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?
  - Food availability (time of day, day of season, annual variation in food = year)
  - Age?
  - Sex?

$$Body\ mass \sim Hornlength$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?
  - Food availability (time of day, day of season, annual variation in food = year)
  - Age?
  - Sex?
  - Reproductive status (pregnant/not)

$$Body\ mass \sim Hornlength$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?
  - Food availability (time of day, day of season, annual variation in food = year)
  - Age?
  - Sex?
  - Reproductive status (pregnant/not)
  - Wearing jewels/not?

$$Body\ mass \sim Horn\ length$$
$$Body\ mass_i = \beta_0 + \beta_1 Horn\ length_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Does the horn of the unicorns help them to eat more so they are heavier?
  - Food availability (time of day, day of season, annual variation in food = year)
  - Age?
  - Sex?
  - Reproductive status (pregnant/not)
  - Wearing jewels/not?
  - Size?

What to include? **HYPOTHESIS**

- Maybe heavier unicorns are able to grow larger horns?

$$Horn\ length \sim Body\ mass$$

What to include? **HYPOTHESIS**

- Maybe heavier unicorns are able to grow larger horns?

$$Horn\ length \sim Body\ mass$$

$$Horn\ length_i = \beta_0 + \beta_1 Hornlength_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Maybe heavier unicorns are able to grow larger horns?

$$Horn\ length \sim Body\ mass$$

$$Horn\ length_i = \beta_0 + \beta_1 Hornlength_i + \varepsilon_i$$

What to include? **HYPOTHESIS**

- Maybe heavier unicorns are able to grow larger horns?
  - Age?
  - Sex?
  - Size?
  - Nutrients in diet?
  - Ecology (shorter horns in denser forests, longer horns on grasslands)
  - Predator presence

Process – common problems

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation

How to decide which model is good?

?

Different ways to compare models

- Information criterion (AIC, BIC, DIC)
- Log-likelihood

Model selection

- Common sense!
- Information criterion (AIC, BIC, DIC)
- Likelihood ratio test
- Step wise deletion

Model selection

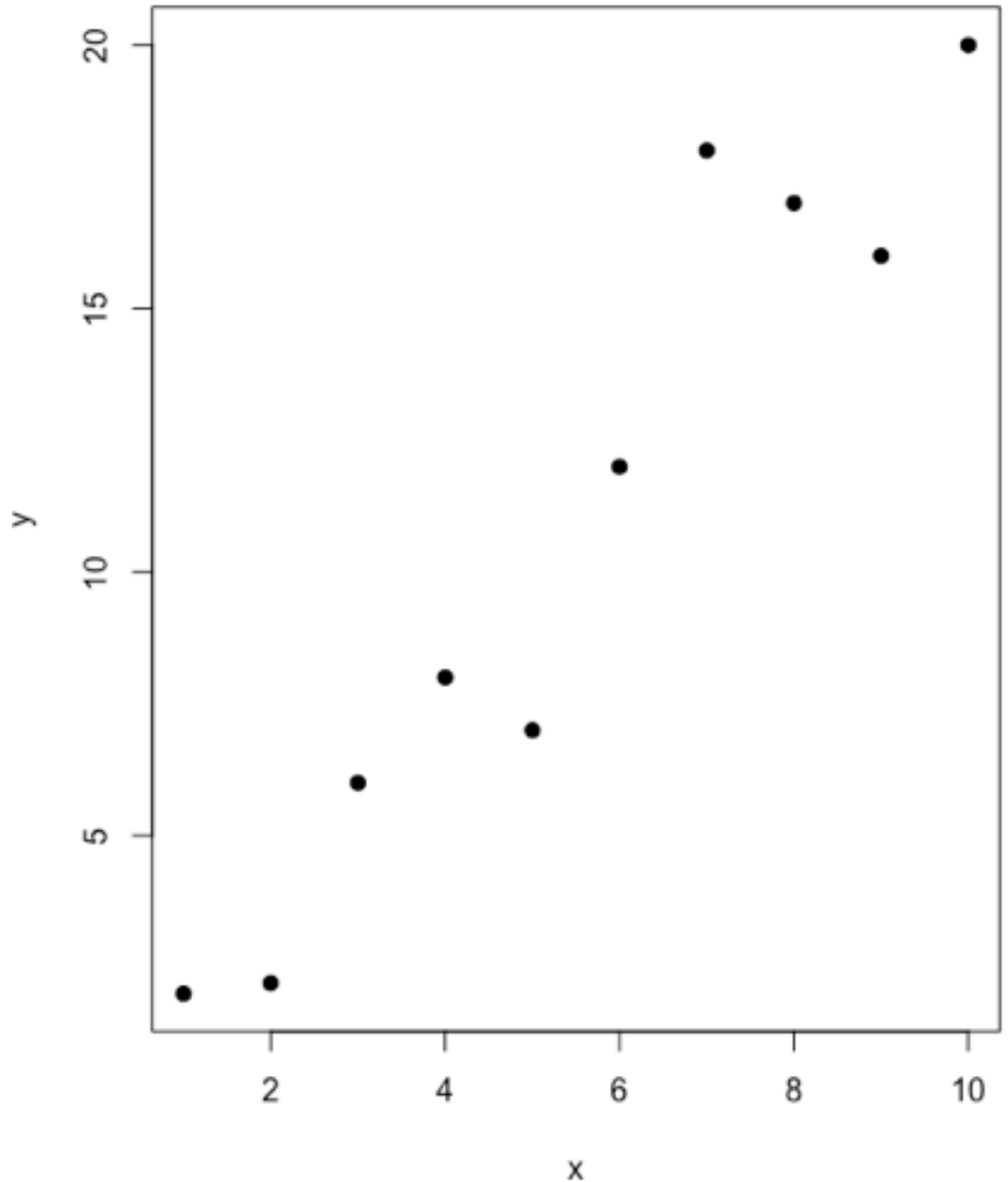
- Is less important than you think

Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

Model selection

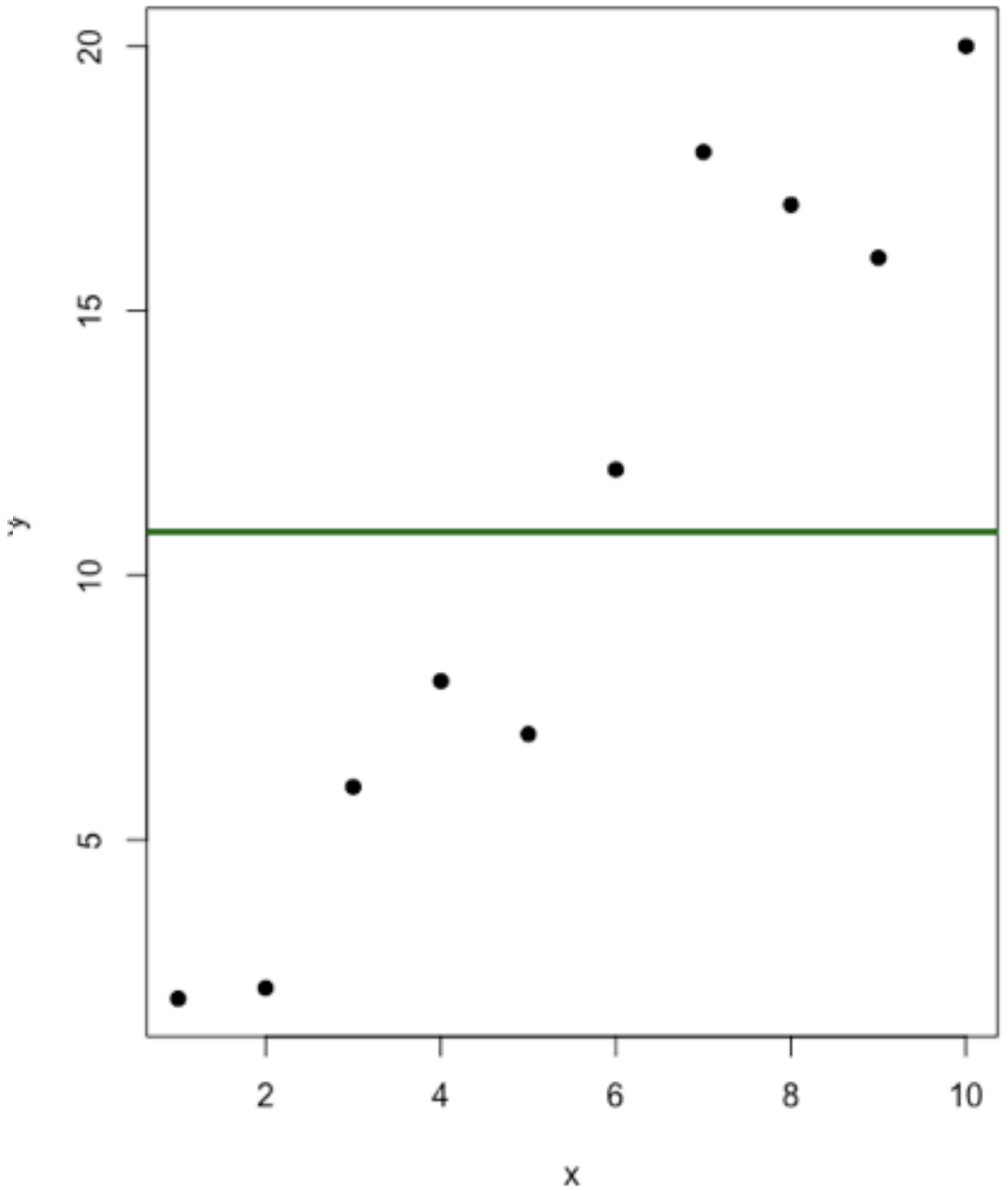
- Is less important than you think
- Only gives information about goodness of fit, not biology!



Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

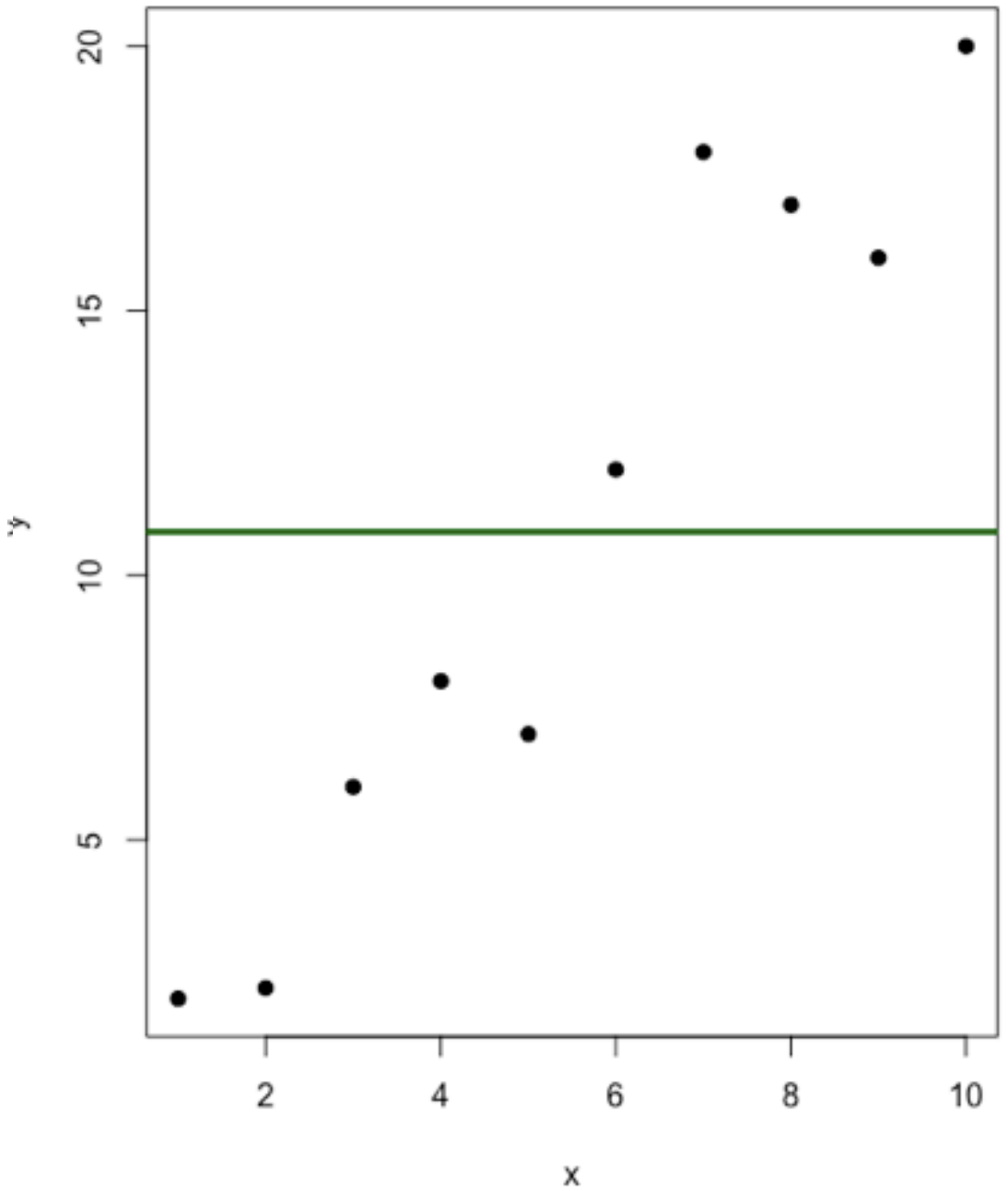
$$y_i = \beta_0$$



Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

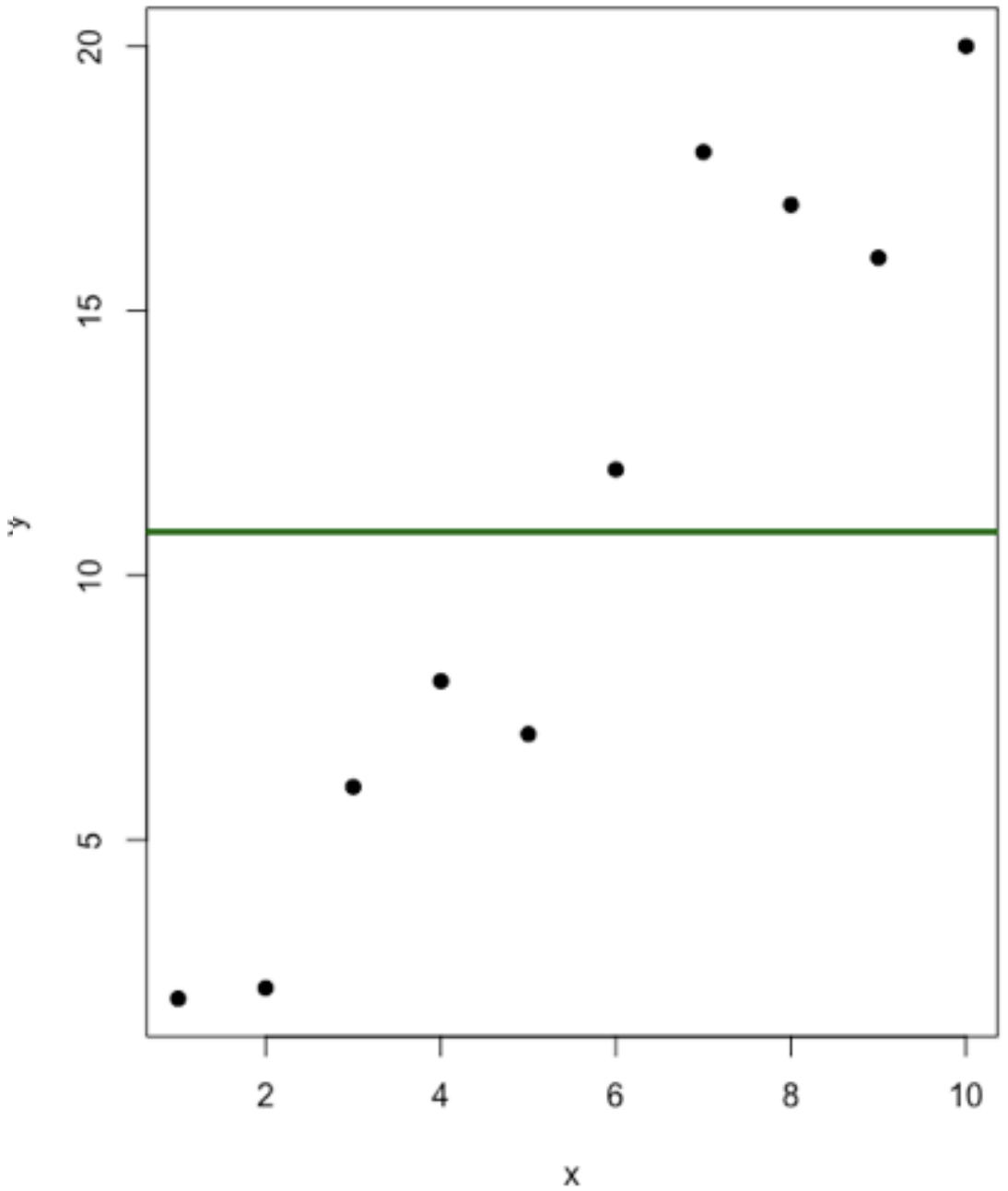
$y_i = \beta_0,\ df = n\text{-}1$



Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

$y_i = \beta_0,\ df = n\text{-}1 = 10$

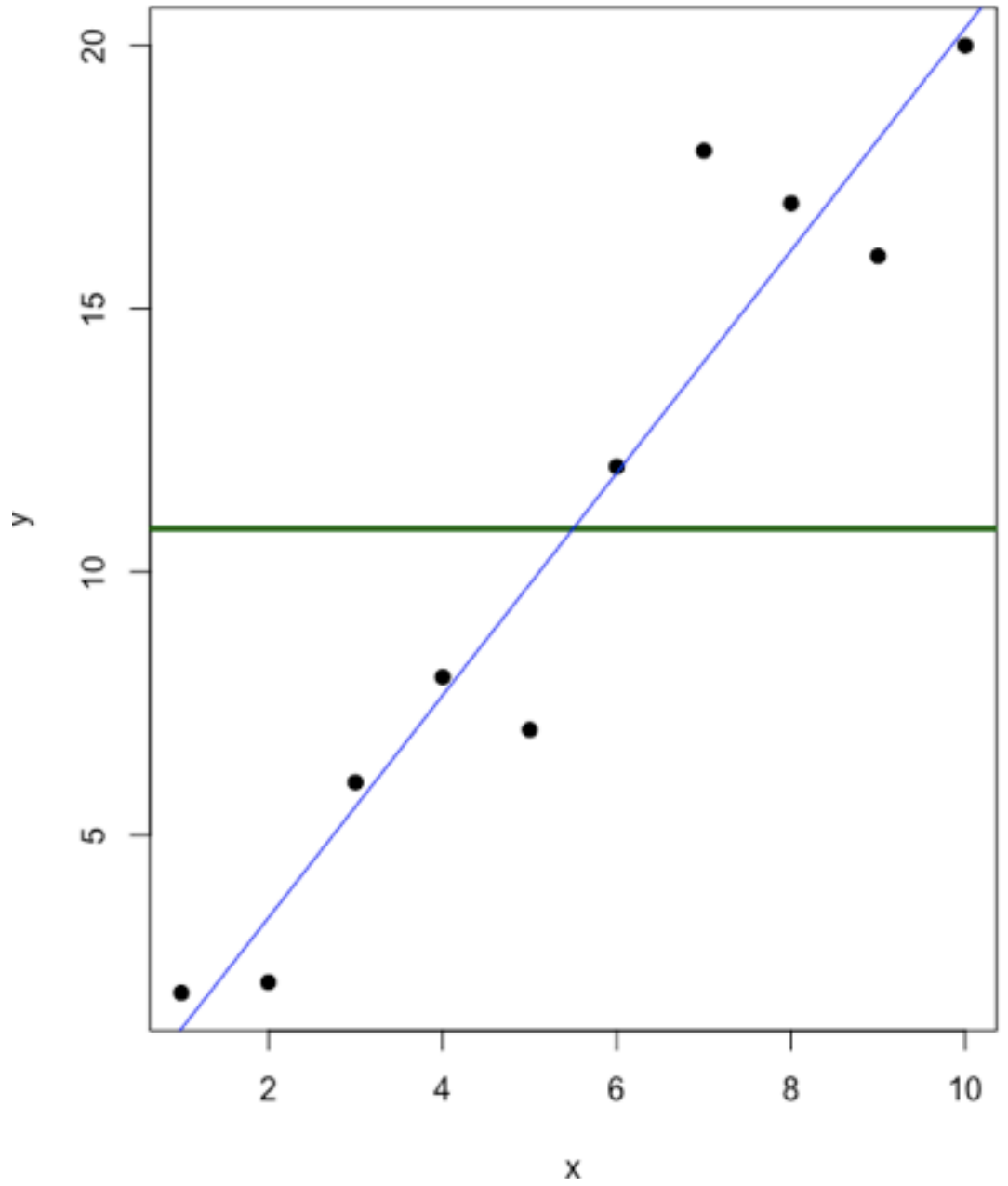


Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

$y_i = \beta_0,\ df = n\text{-}1 = 10$

$y_i = \beta_0 + \beta_1 x_i$

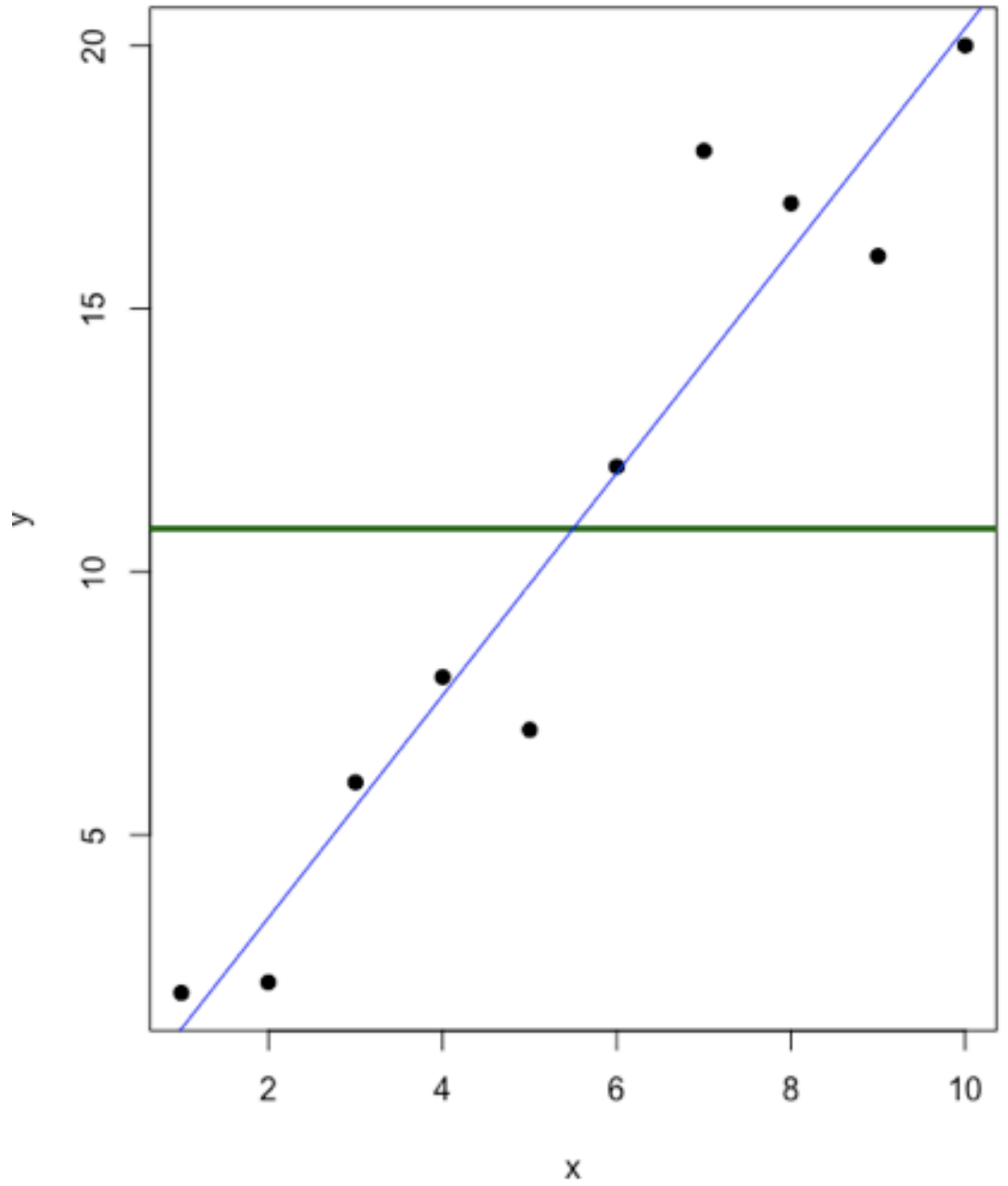


Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i, df = n\text{-}2 = 8$



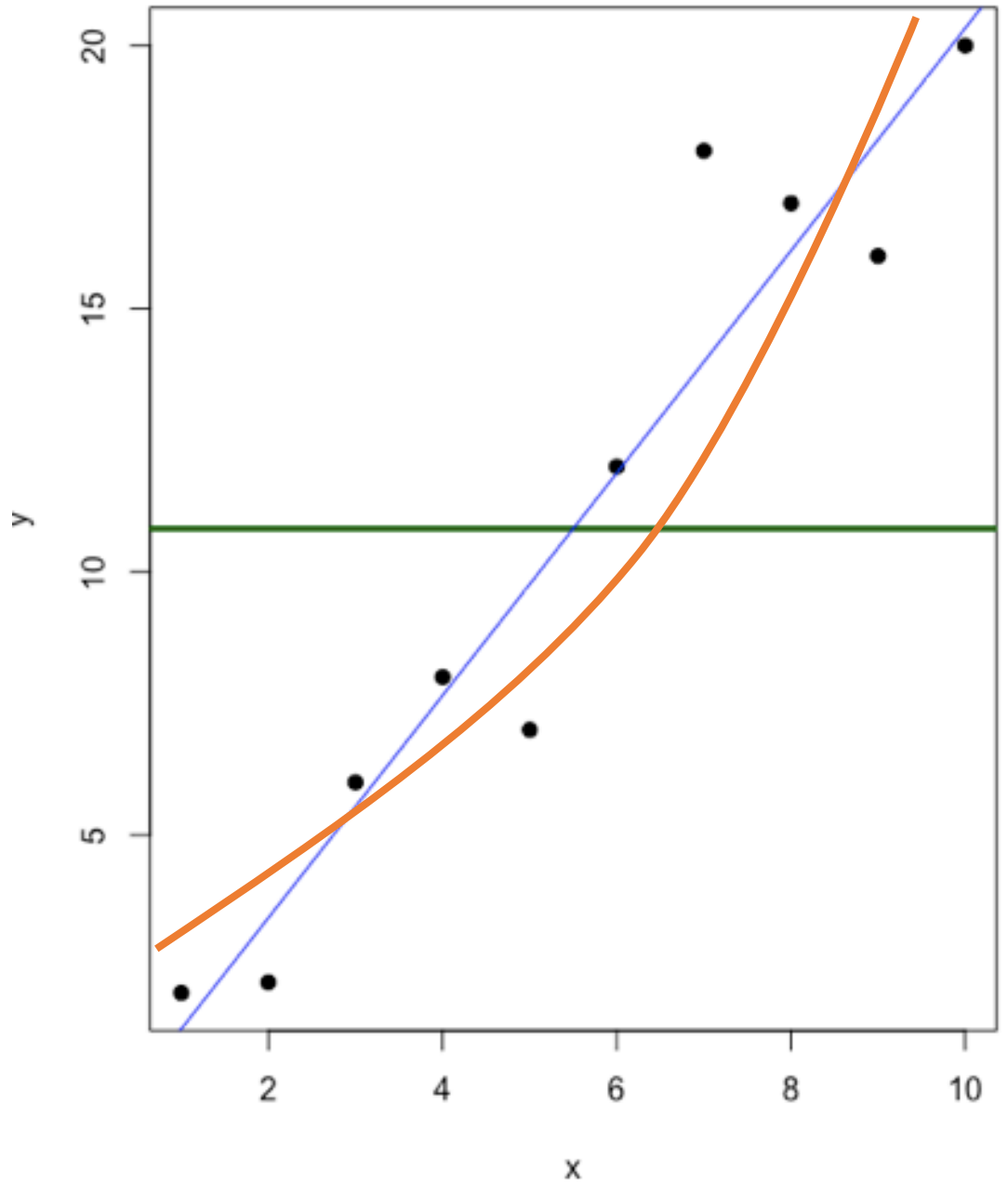
Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

$$y_i = \beta_0,\ df = n\text{-}1 = 9$$

$$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2$$



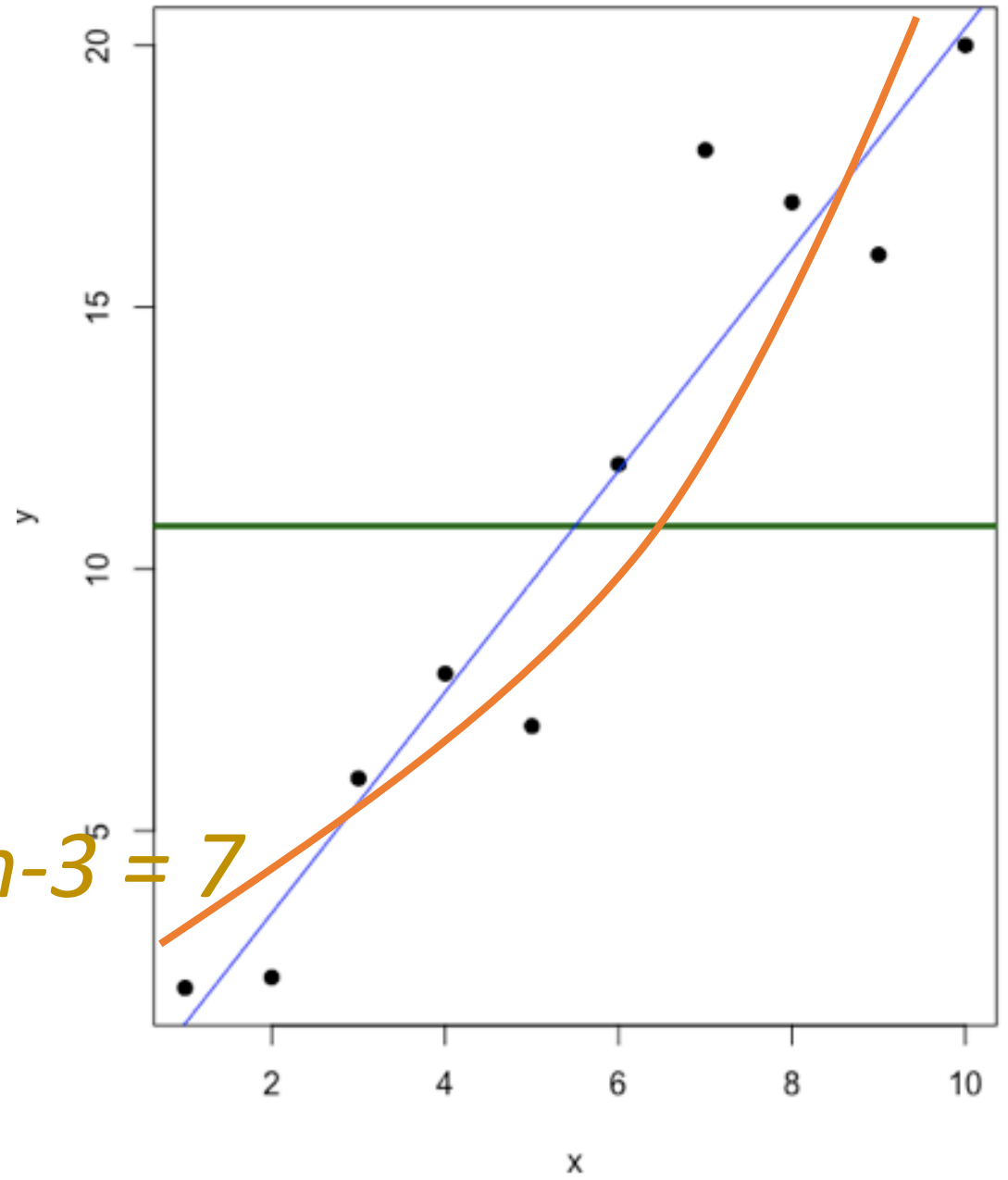
Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = n\text{-}3 = 7$



Model selection

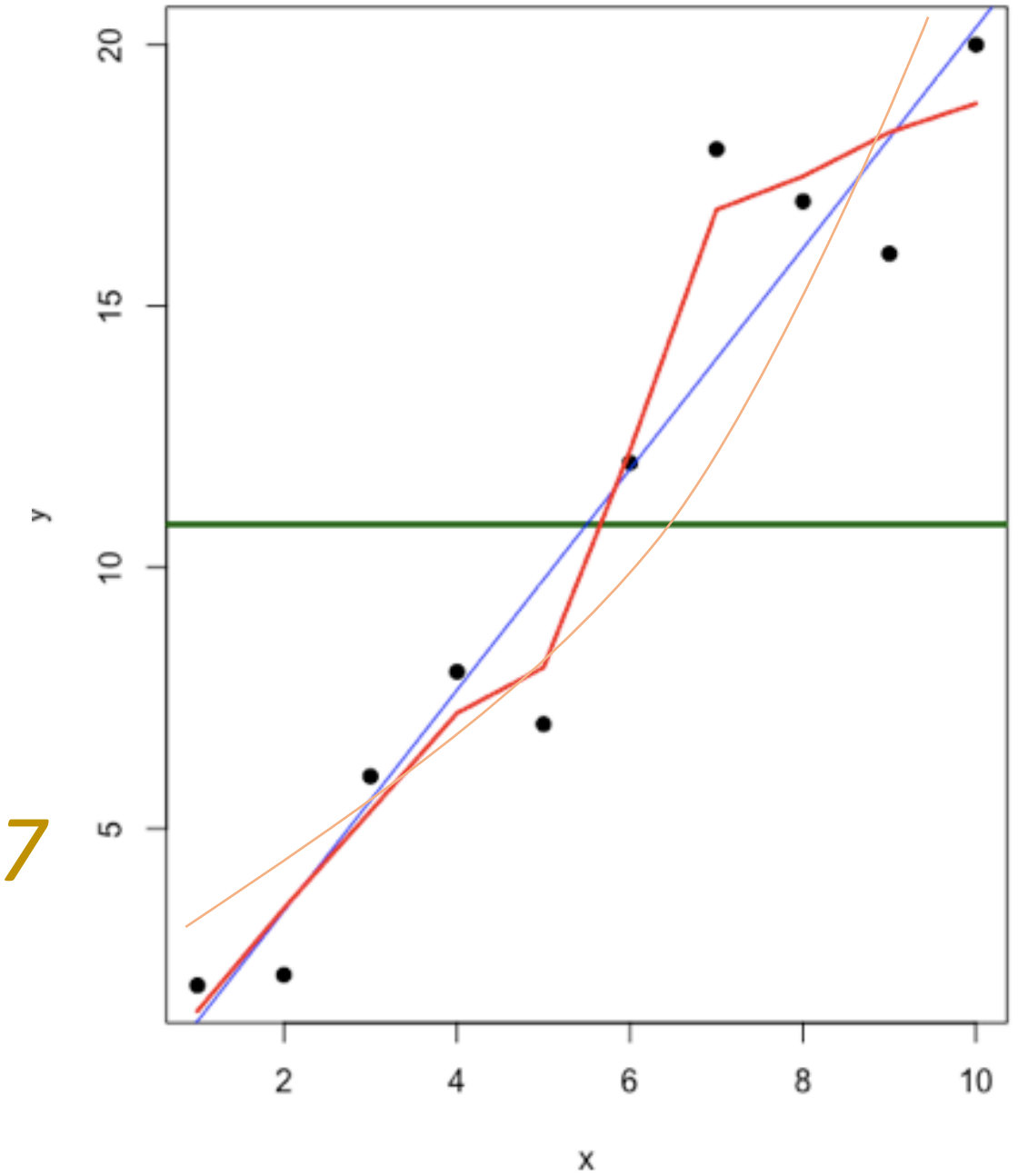
- Is less important than you think
- Only gives information about goodness of fit, not biology!

$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$



Model selection

- Is less important than you think
- Only gives information about goodness of fit, not biology!

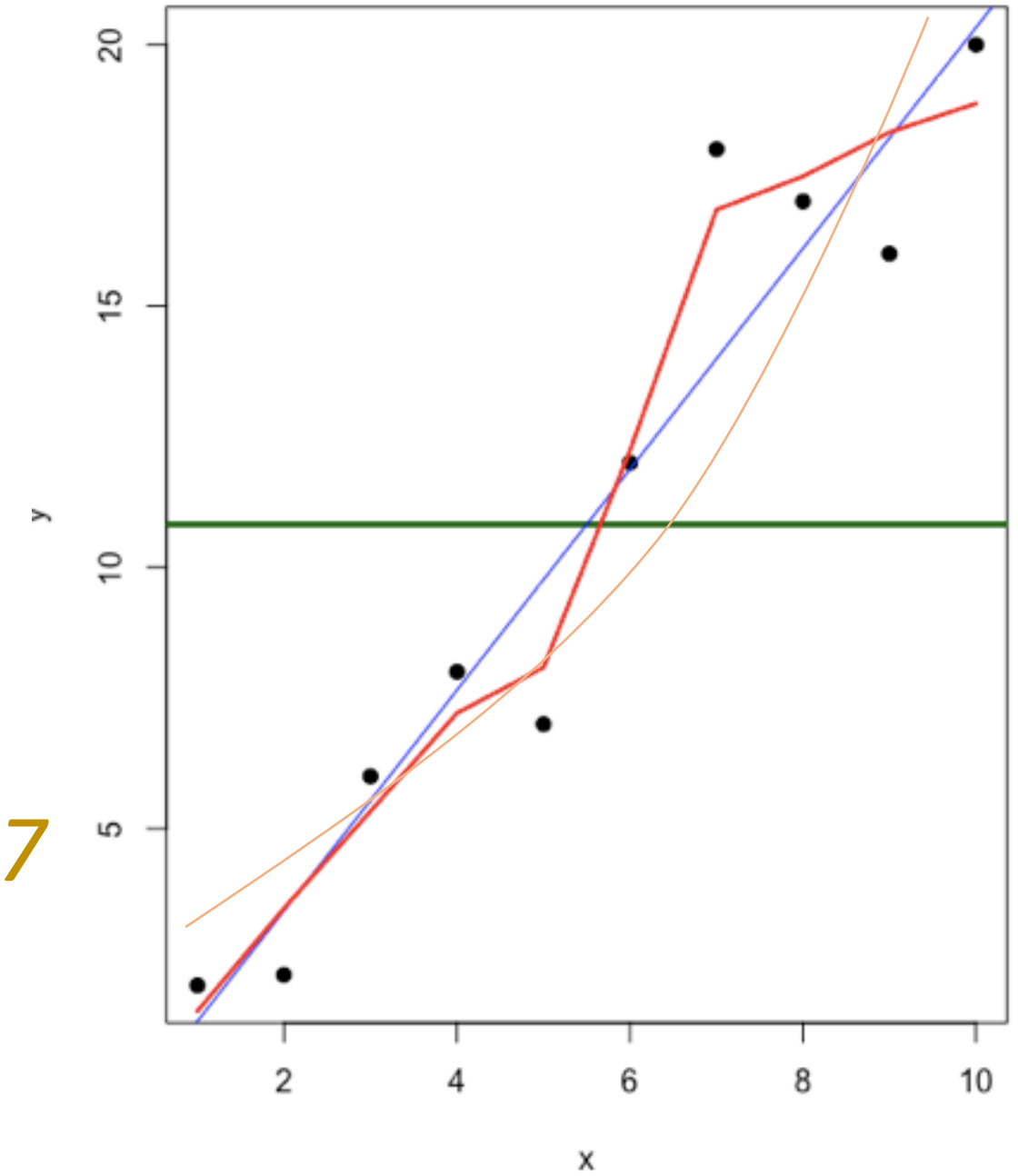
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Top – down: increasing fit

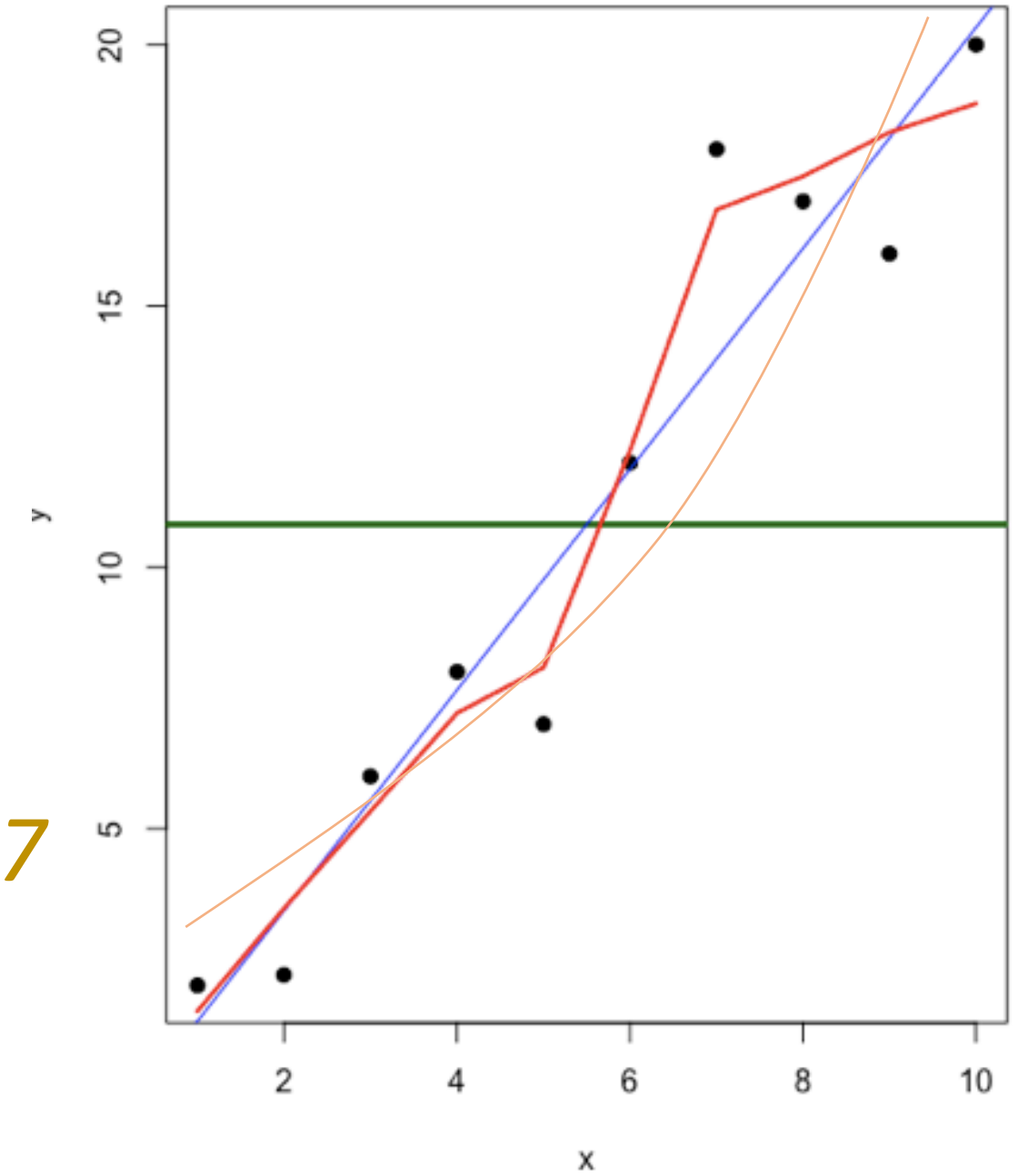
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Top – down: increasing fit

Bottom – up: loss of *df*s

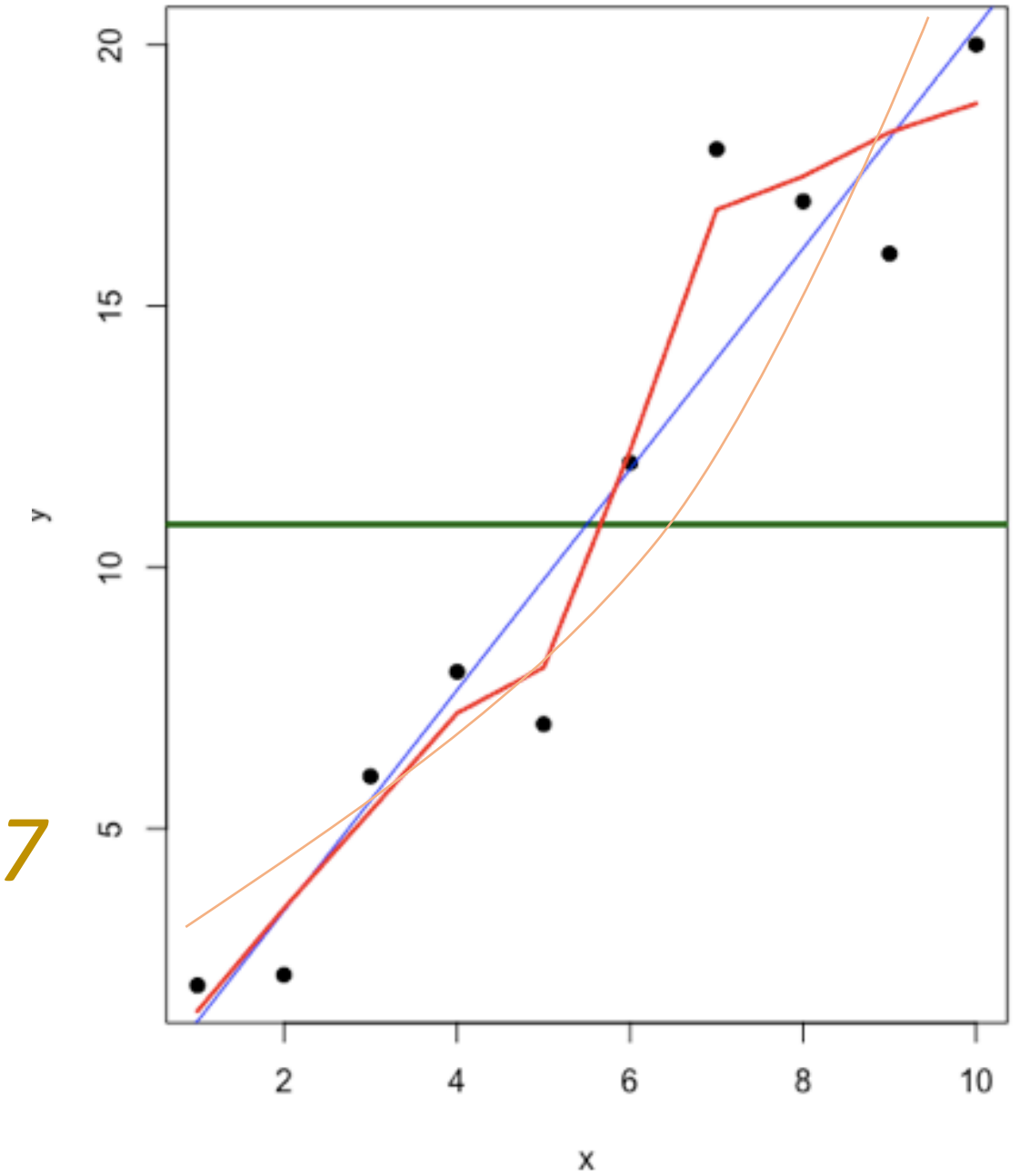
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Top – down: increasing fit

Bottom – up: loss of *df*s

Trade-off explanatory power with complexity

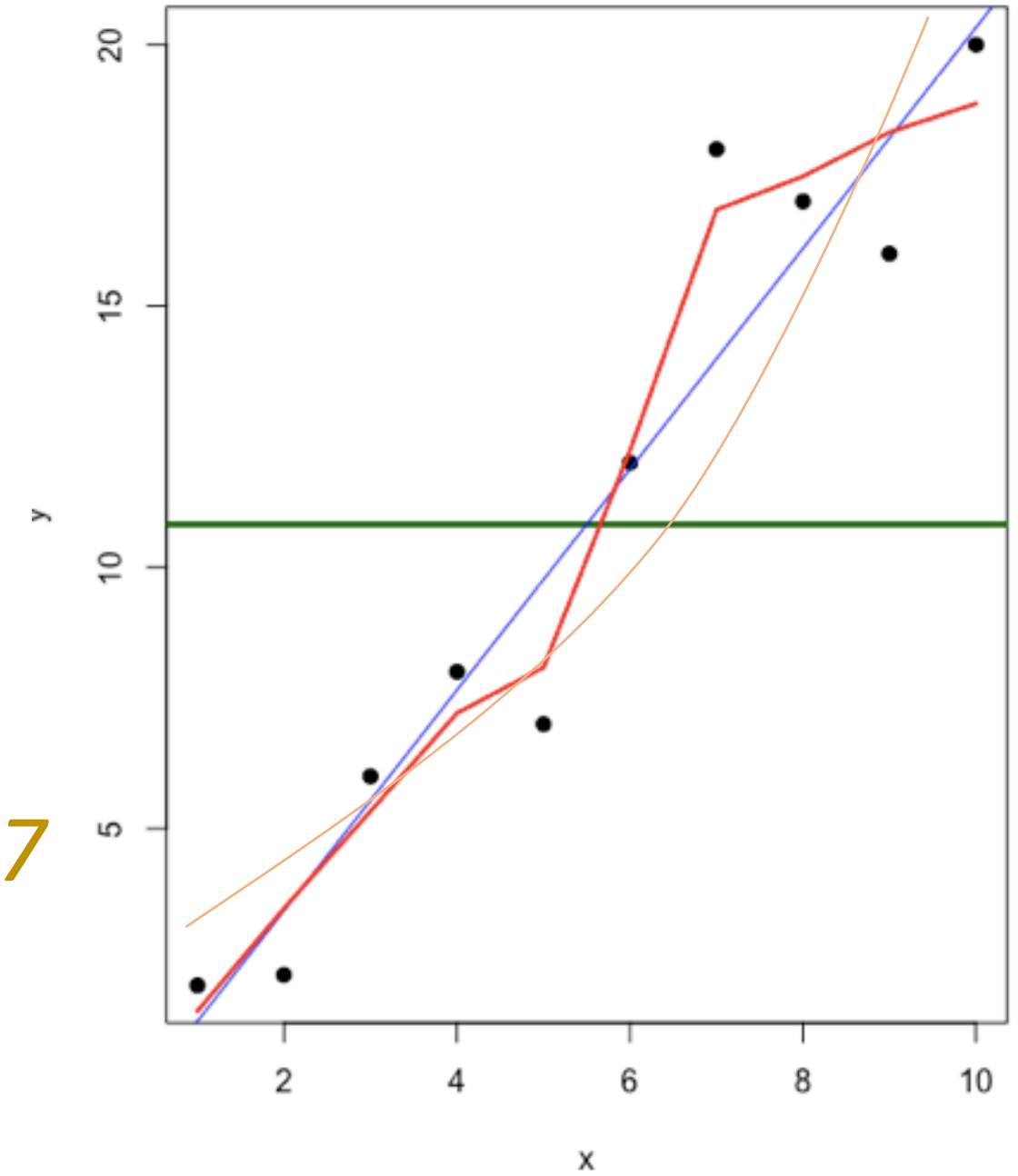
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Which model to select?

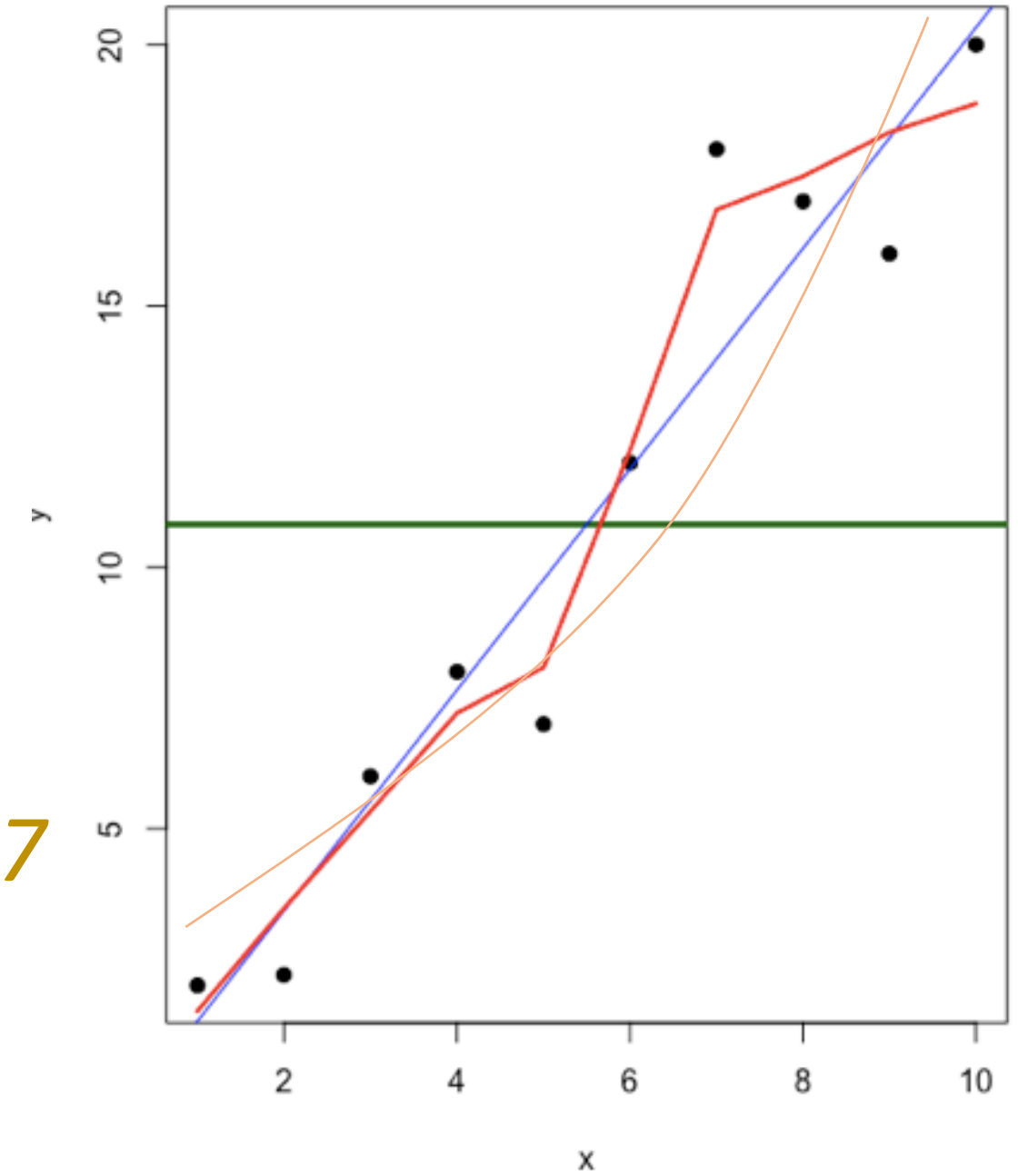
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Which model to select?

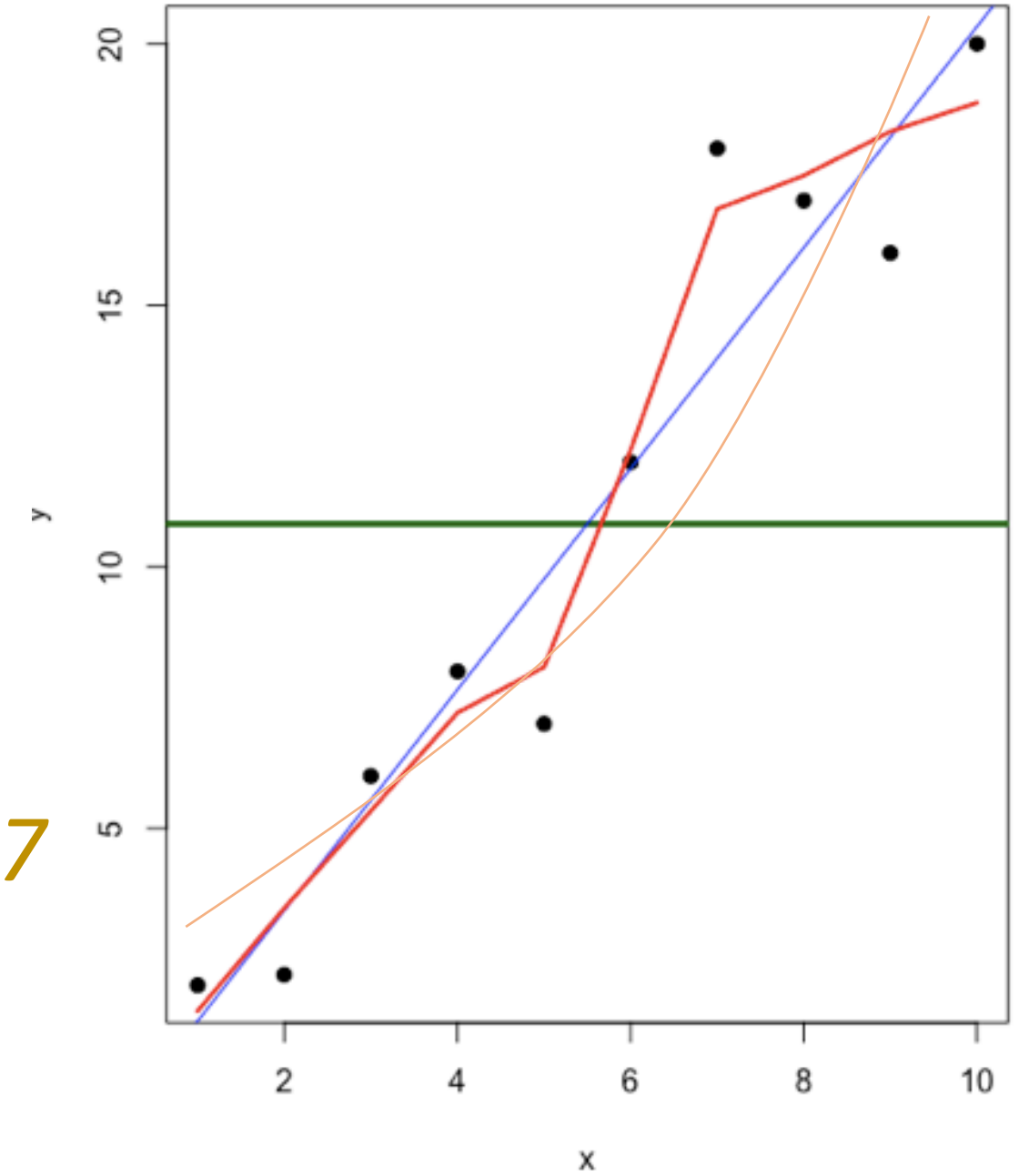
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Which model to select?

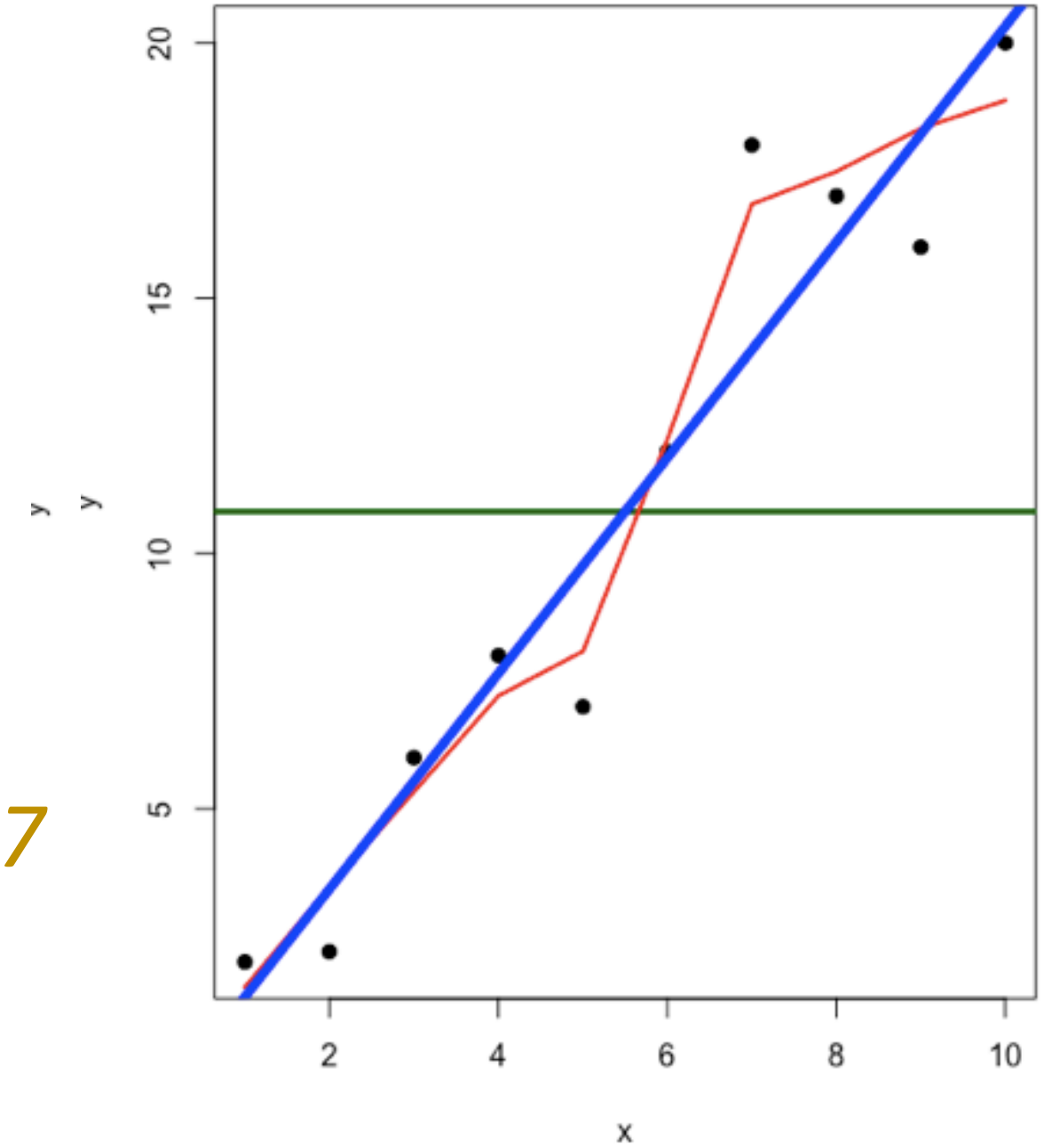
$y_i = \beta_0,\ df = n\text{-}1 = 9$

$\boldsymbol{y_i = \beta_0 + \beta_1 x_i,\ df = n\text{-}2 = 8}$

$y_i = \beta_0 + \beta_1 x_i + \beta_1 x_i^2,\ df = 7$

$y_i = LOESS\ f,\ df = n\text{-}5 = 5$

$y_i = y_i,\ df = n\text{-}10 = 0$



Model selection

Which model to select?

- Sparrow parental care in relationship to offspring number

Model selection

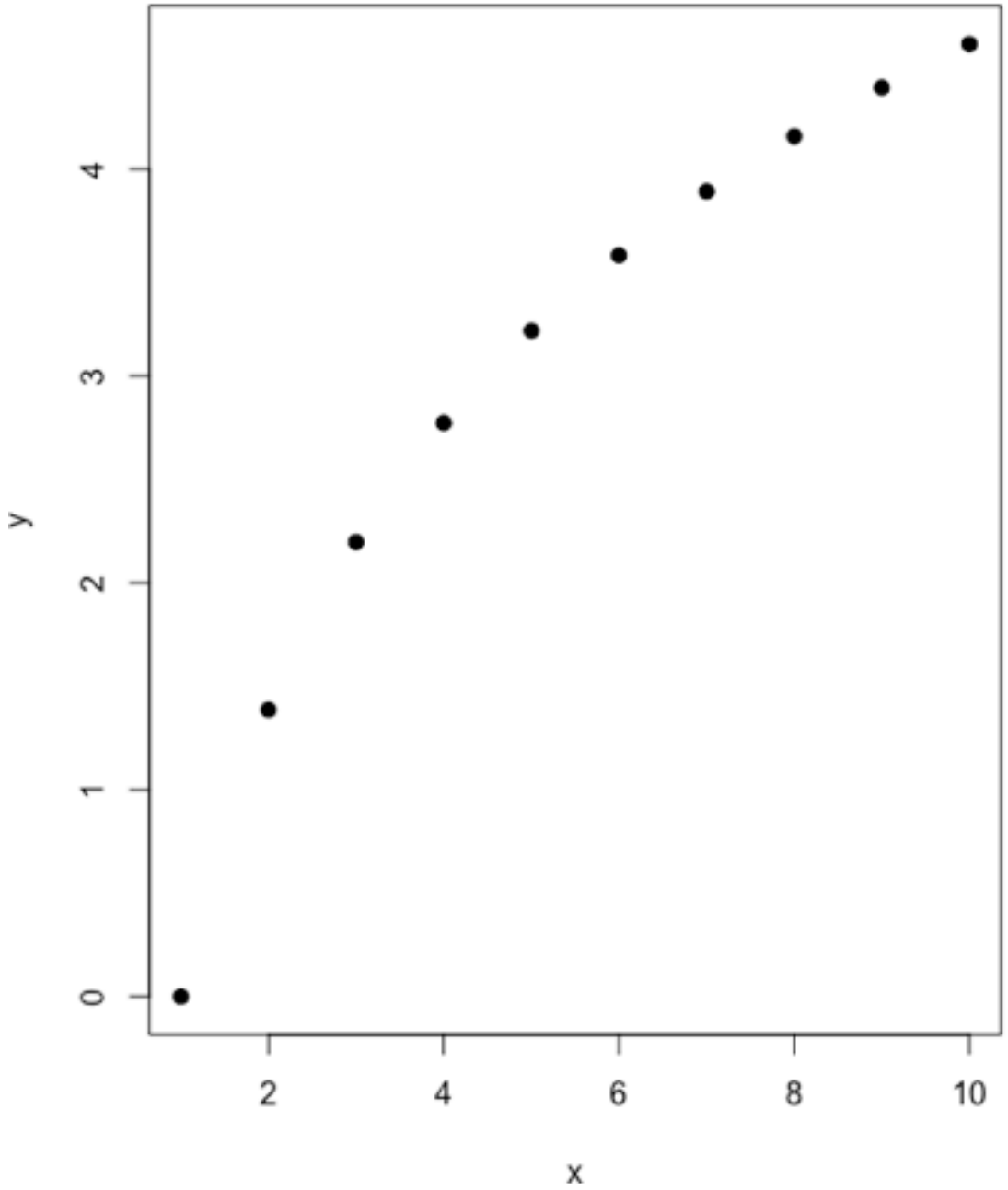
Which model to select?

- Sparrow parental care in relationship to offspring number
- Expect a positive slope

Model selection

Which model to select?

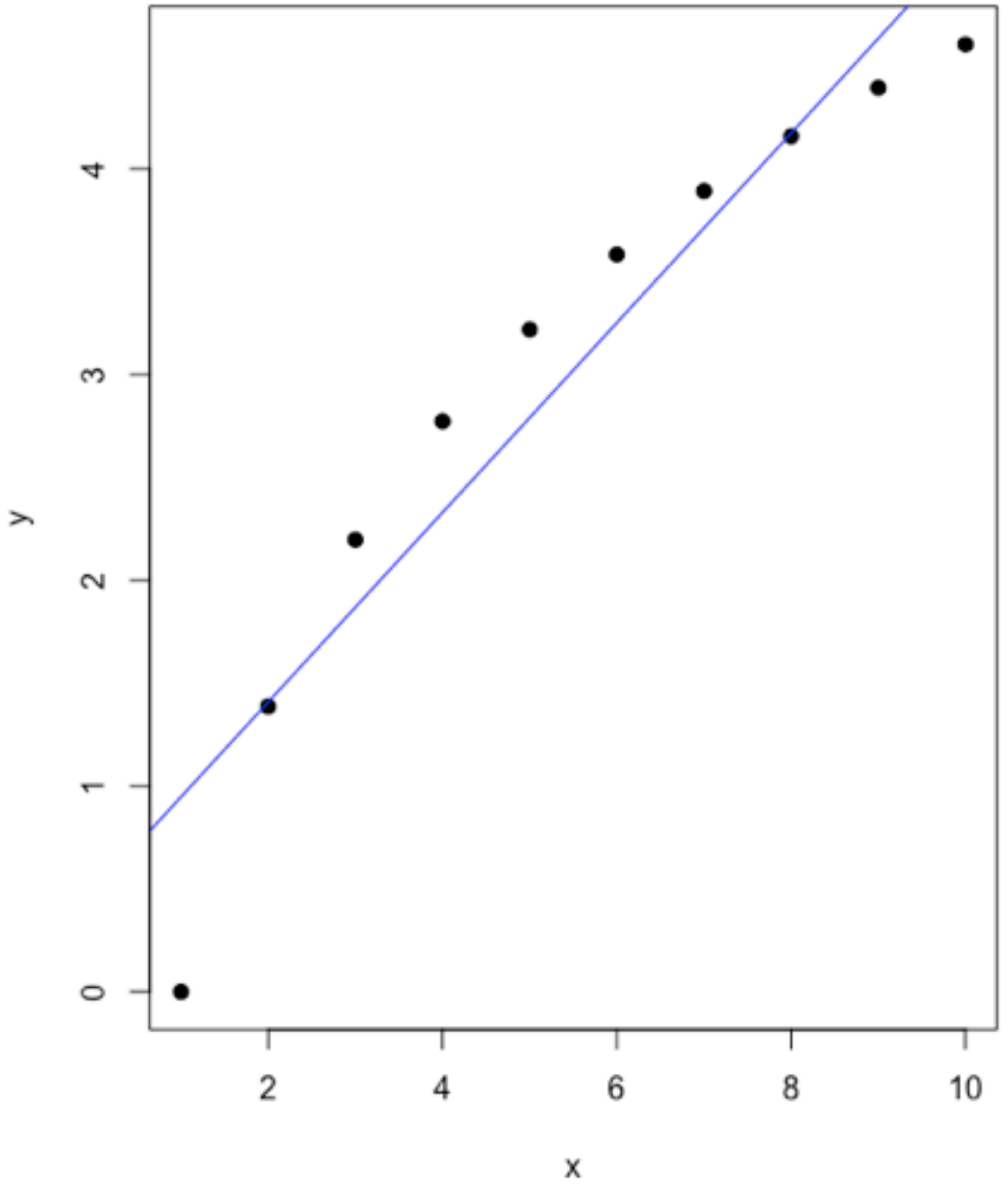
- Sparrow parental care in relationship to offspring number
- Expect a positive slope



Model selection

Which model to select?

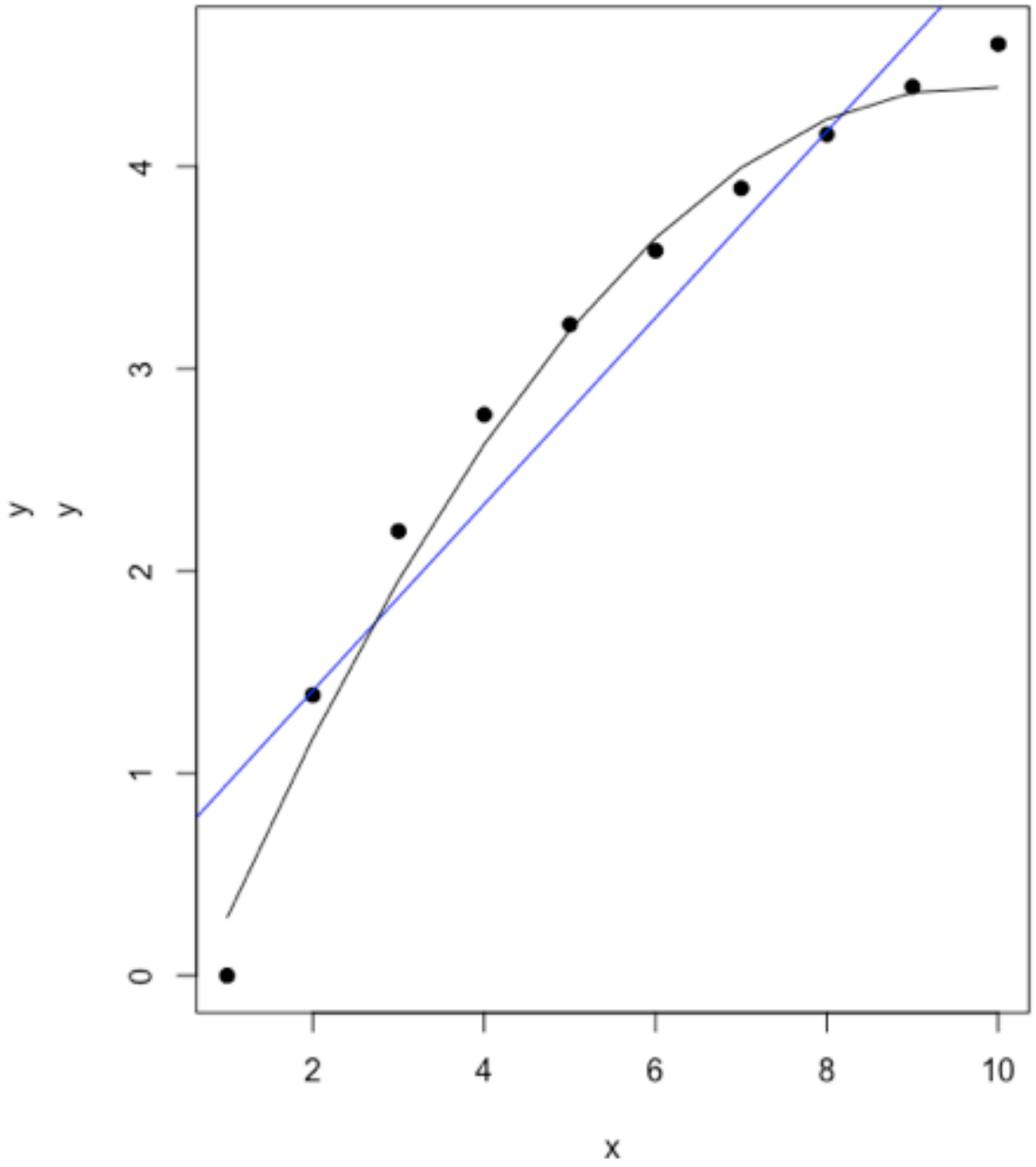
- Sparrow parental care in relationship to offspring number
- Expect a positive slope



Model selection

Which model to select?

- Sparrow parental care in relationship to offspring number
- Expect a positive slope
- But diminishing

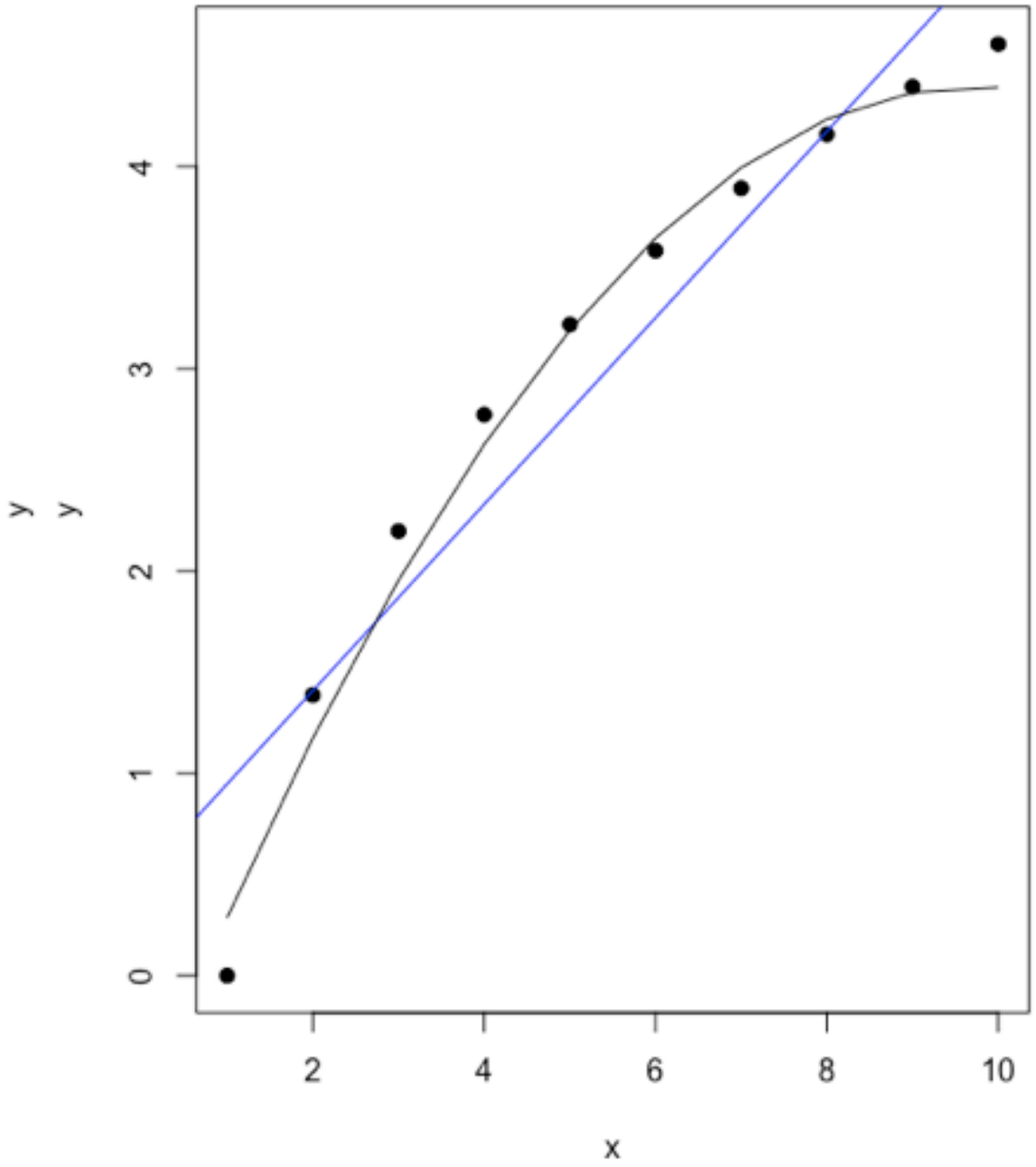


Model selection

Which model to select?

- Sparrow parental care in relationship to offspring number
- Expect a positive slope
- But diminishing

- Biology (and common sense) dictates a squared term



Model selection

- Step-wise deletion

Model selection

- Step-wise deletion
- Used to be go-to method but now not any longer suggested

Model selection

- Step-wise deletion
- Used to be go-to method but now not any longer suggested
- Make full model with all parameters and interactions

Model selection

- Step-wise deletion
- Used to be go-to method but now not any longer suggested
- Make full model with all variable and interactions
- Delete least significant variable – interactions first!

Model selection

- Step-wise deletion
- Used to be go-to method but now not any longer suggested
- Make full model with all variable and interactions
- Delete least significant variable – interactions first!
- **NEVER have interaction in but not the main effects!!!!!**

Model selection

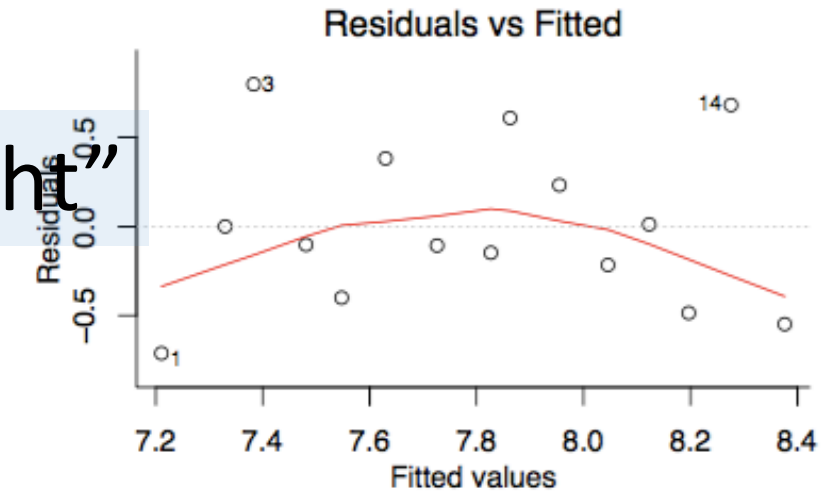
- Step-wise deletion
- Used to be go-to method but now not any longer suggested
- Make full model with all variable and interactions
- Delete least significant variable – interactions first!
- **NEVER have interaction in but not the main effects**
- Continue until all effects are significant, or needed otherwise

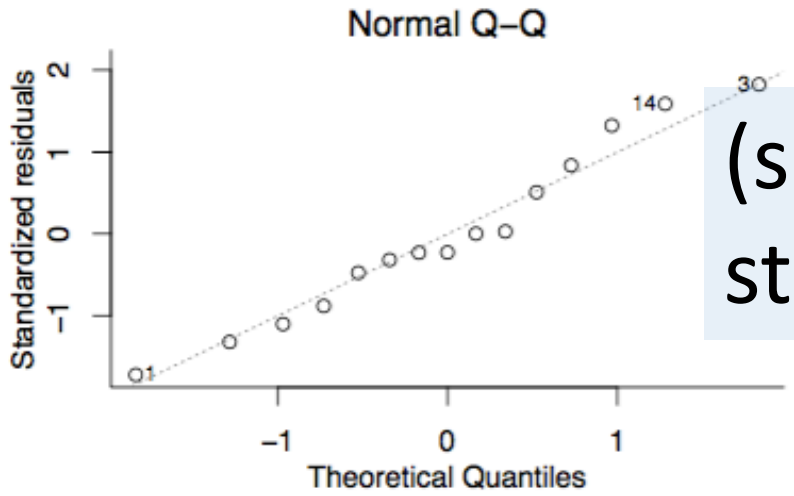
Process – common problems

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation

Diagnostic plots

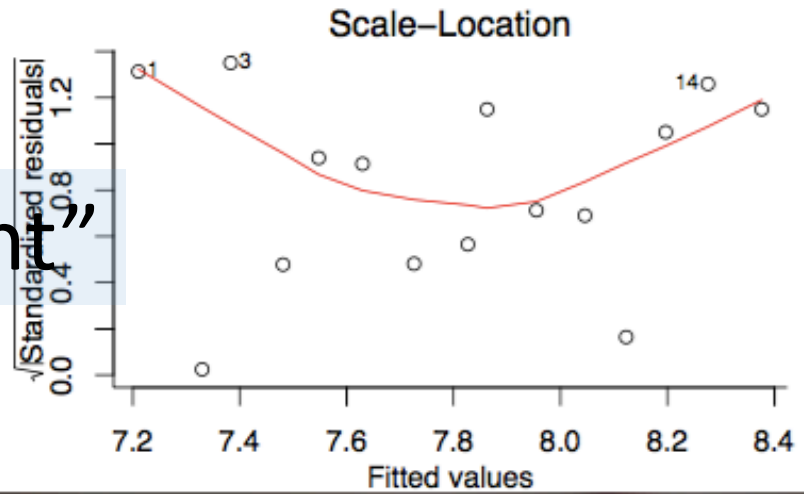
```
> mod <- lm(y ~ x, data=myData)
> plot(mod)
```

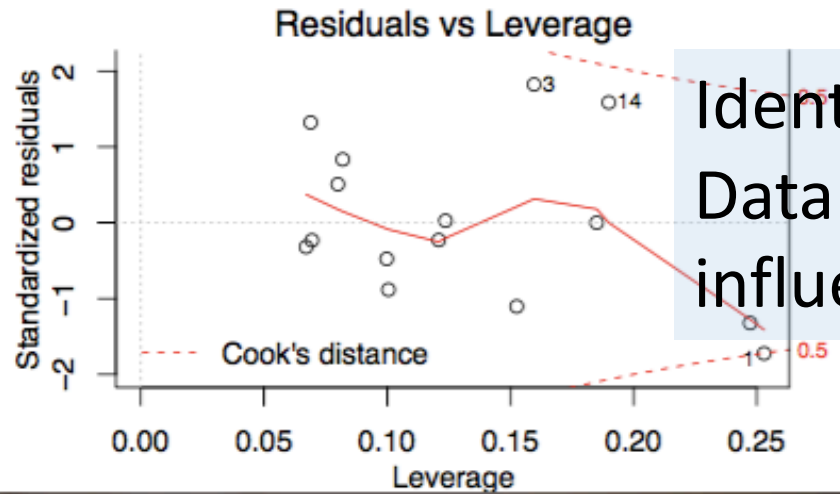




“starry night”

(somewhat) straight line



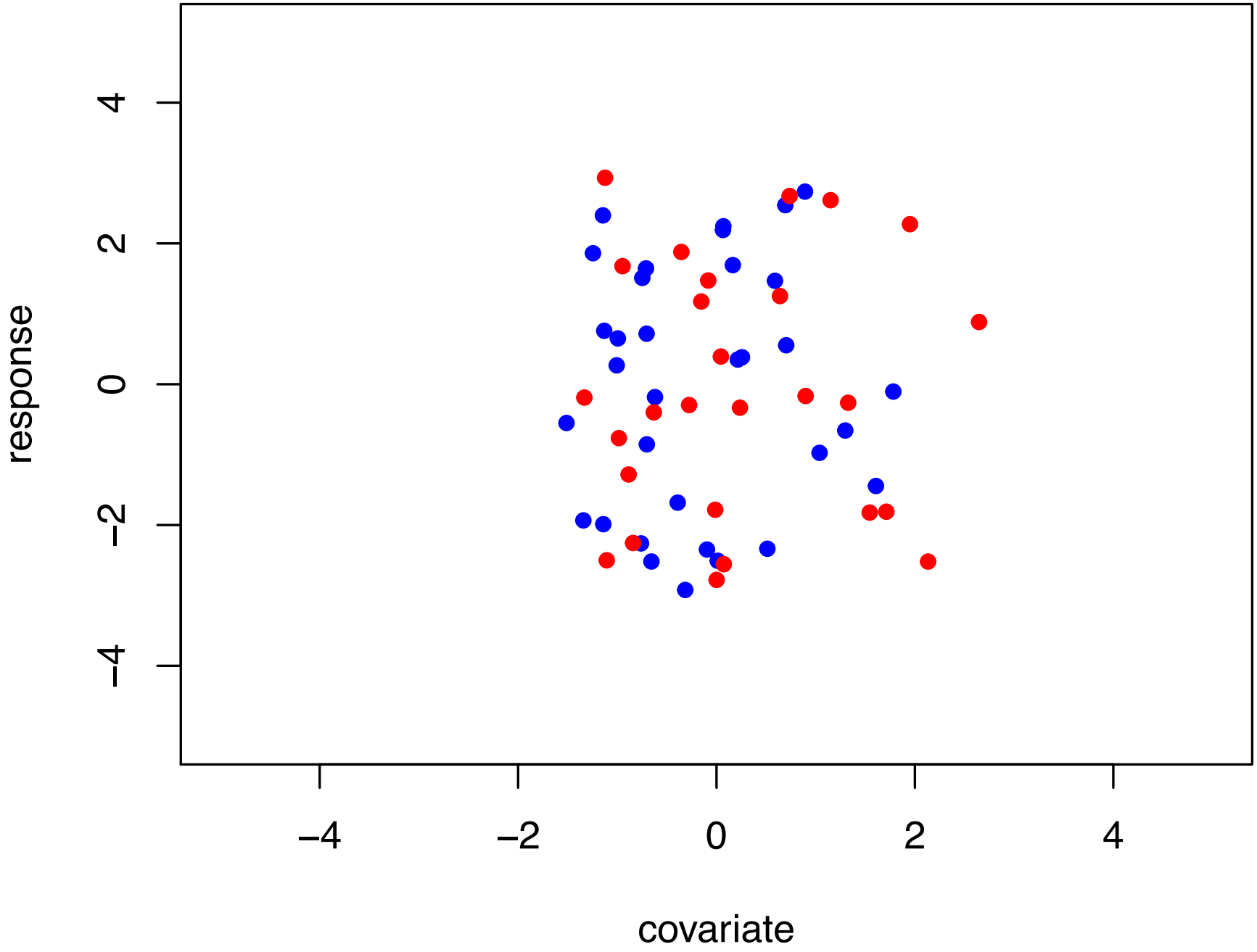


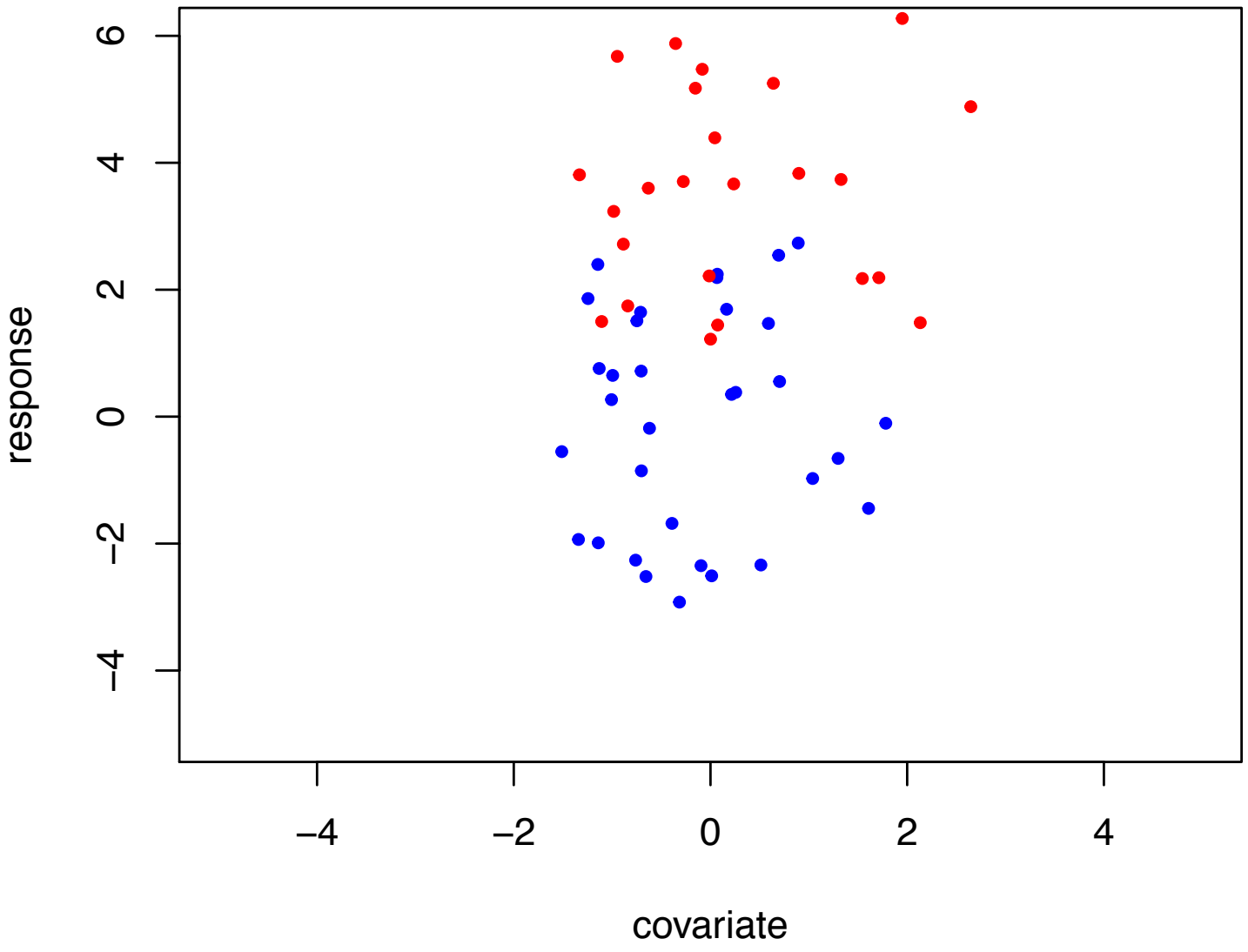
“starry night”

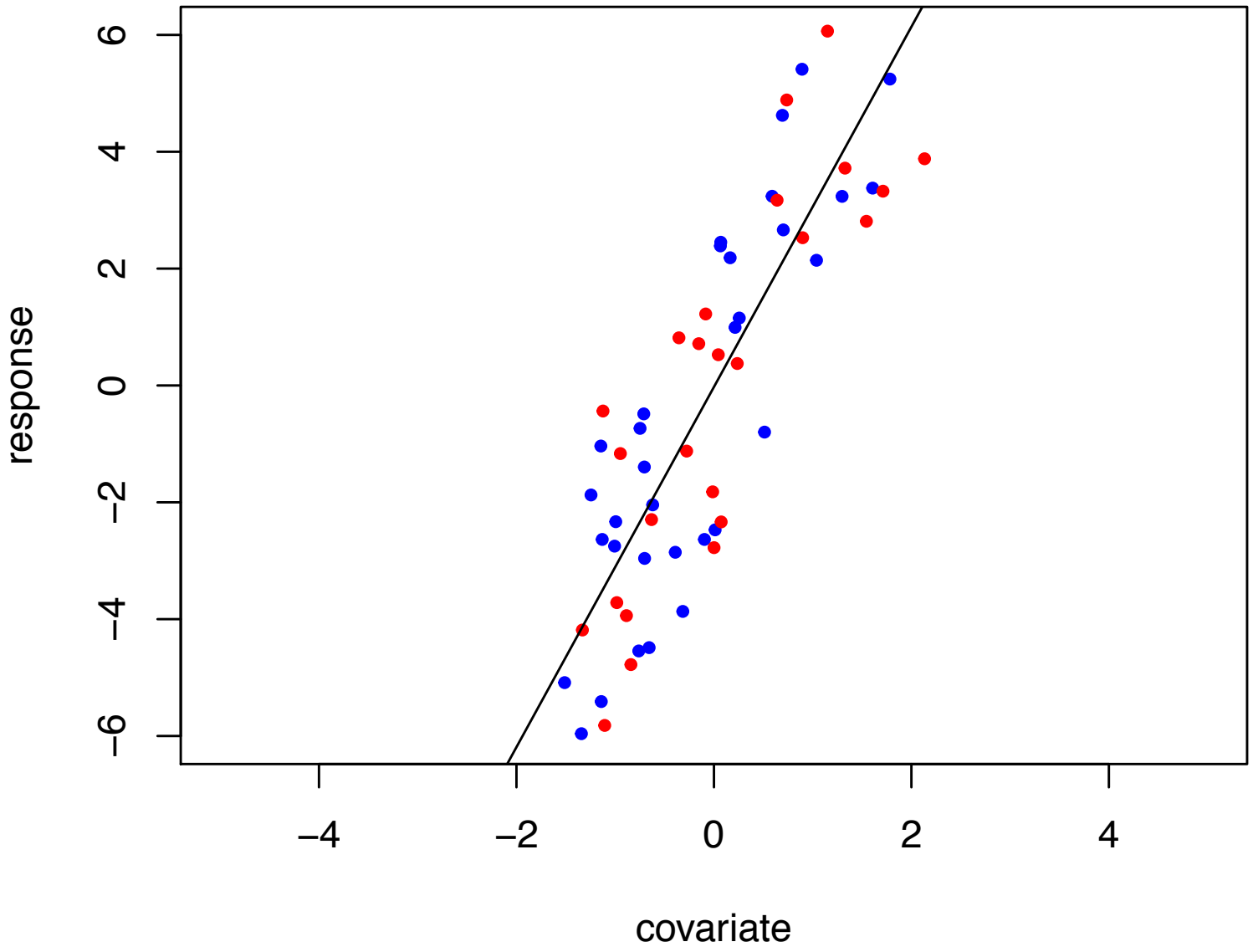
Identify outliers
Data with lots of influence

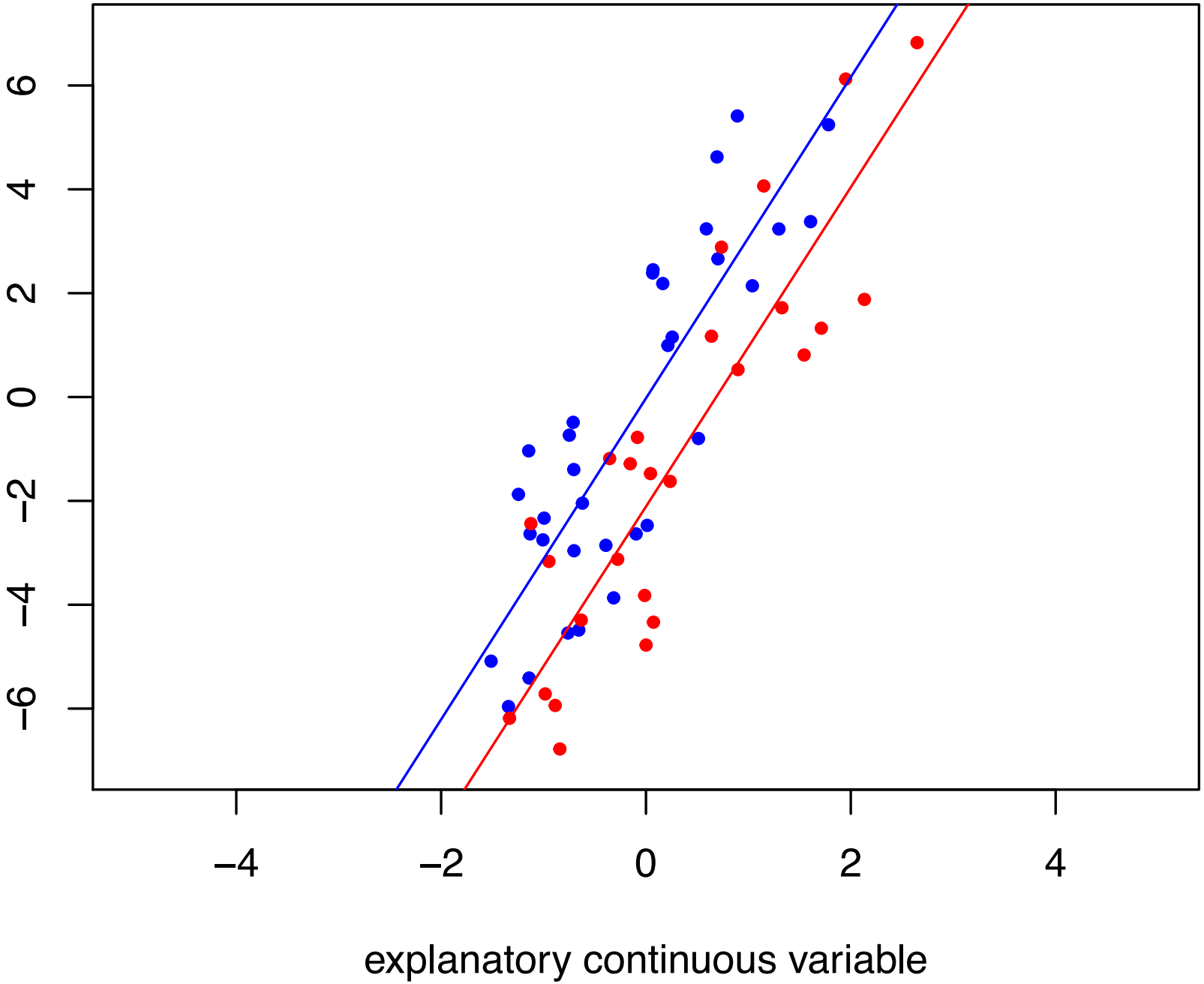
Process – common problems

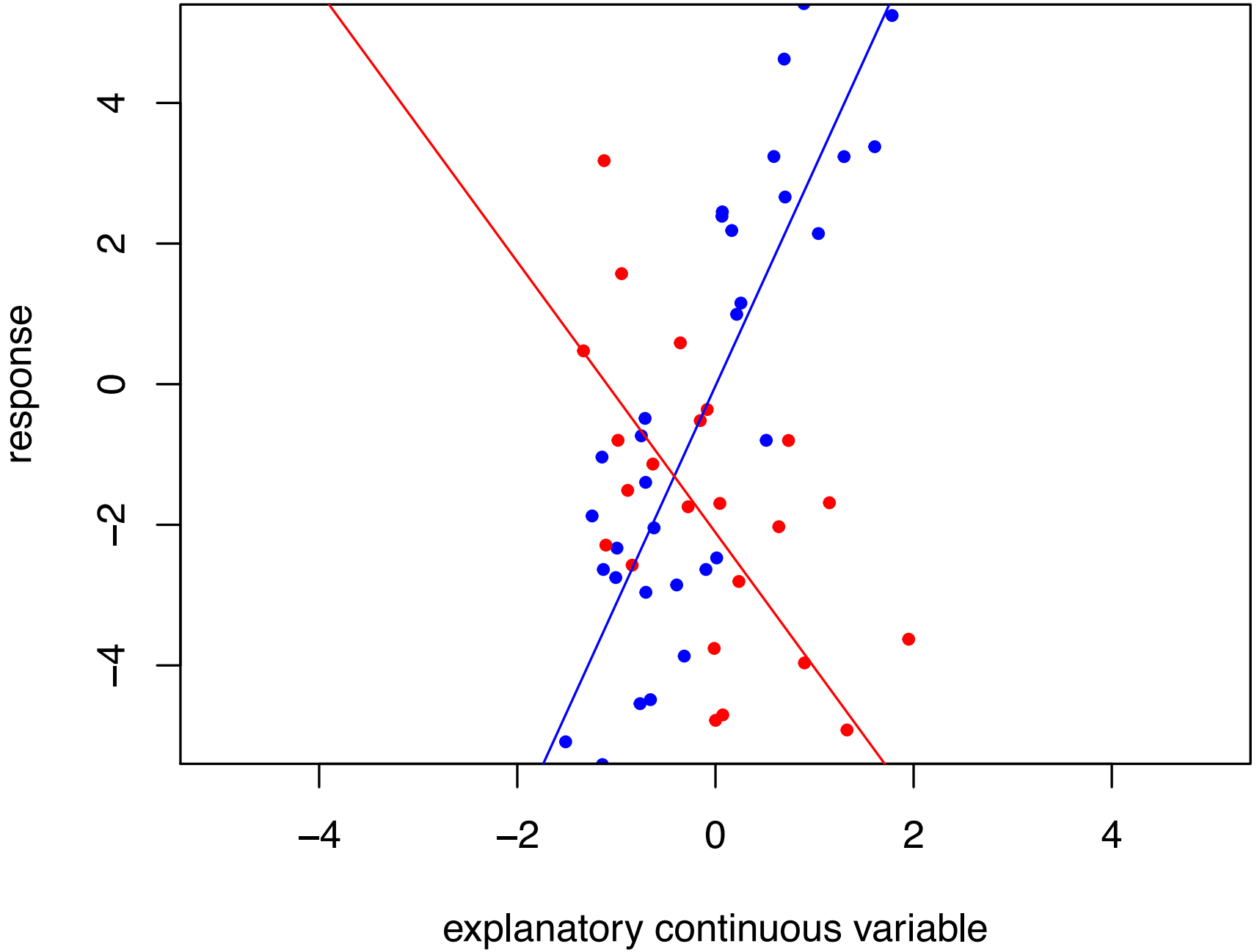
1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation


response
4
2
0
−2
−4
−4
−2
0
2
4
covariate


response
6
4
2
0
−2
−4
−4
−2
0
2
4
covariate

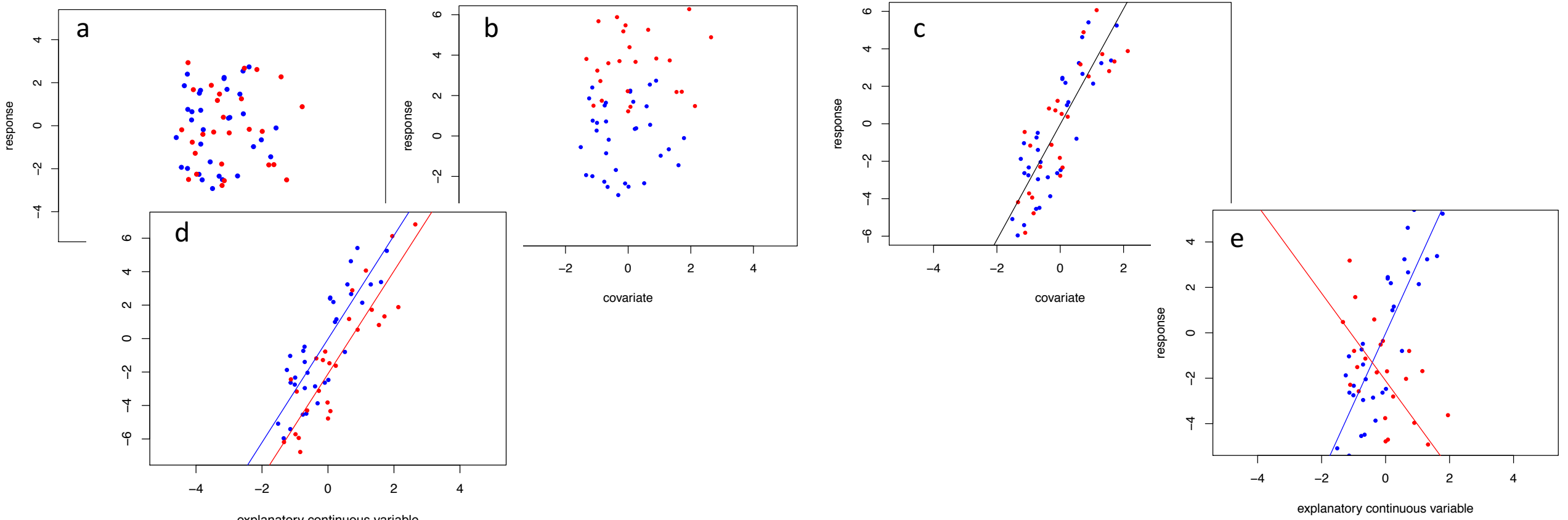

response
6
4
2
0
−2
−4
−6
−4
−2
0
2
4
covariate


6
4
2
0
−2
−4
−6
−4
−2
0
2
4
explanatory continuous variable


response
4
2
0
−2
−4
−4
−2
0
2
4
explanatory continuous variable

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

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	-



Process – common problems

1. Outliers?
2. Homogeneity of variances?
3. Normal distributed?
4. Zero-inflation?
5. Collinearity among covariates?
6. Plot data
7. Which covariates, fixed factors, and interactions?
8. Maximal model
9. Model selection
10. Make a decision
11. Model validation
12. Interpretation

- Remember units!
- Effect sizes!

Sparrows!

Unicorns 1.0!

Grasslands!

Unicorns 2.0