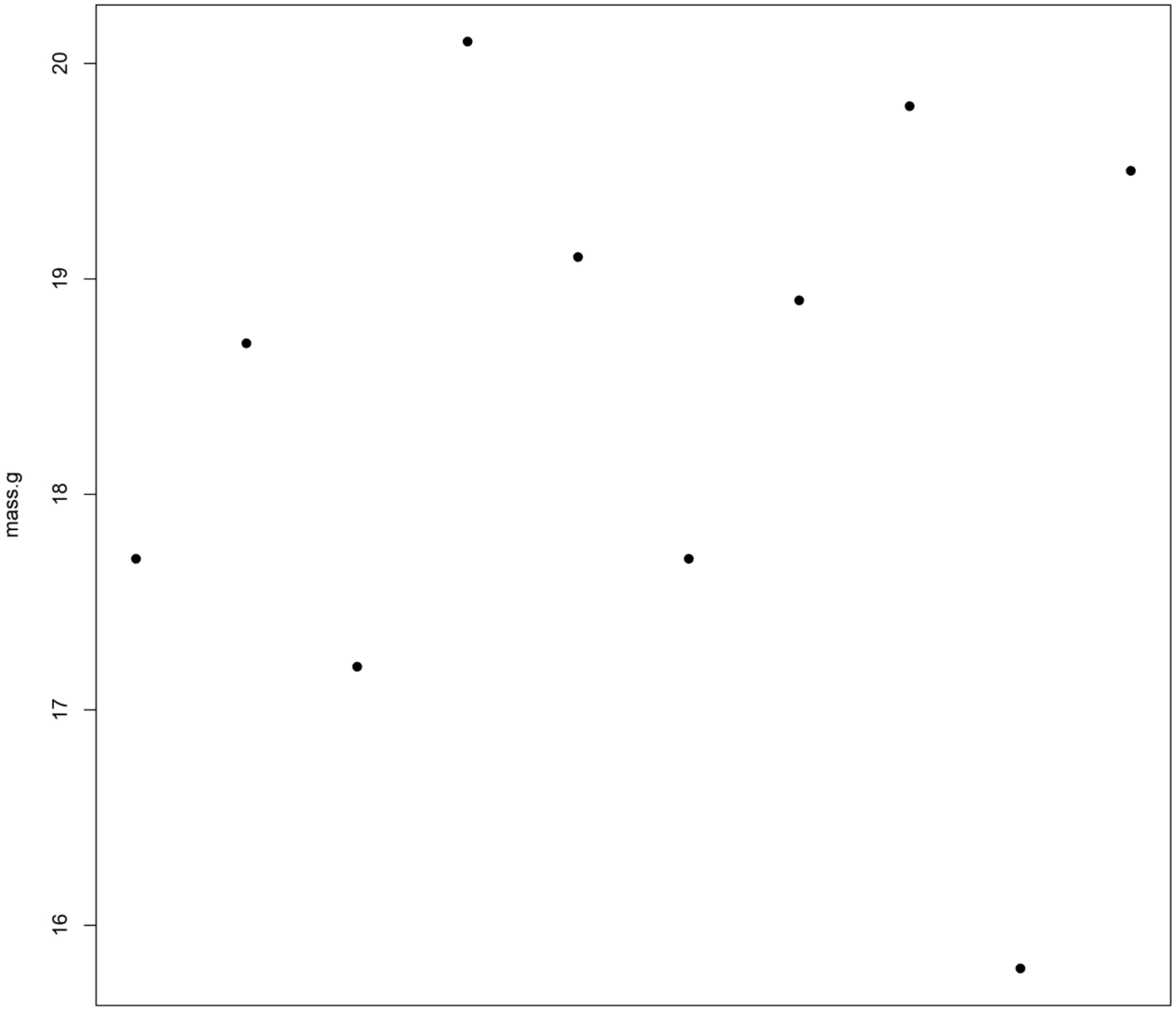
Regression part 1: Anatomy of the ANOVA

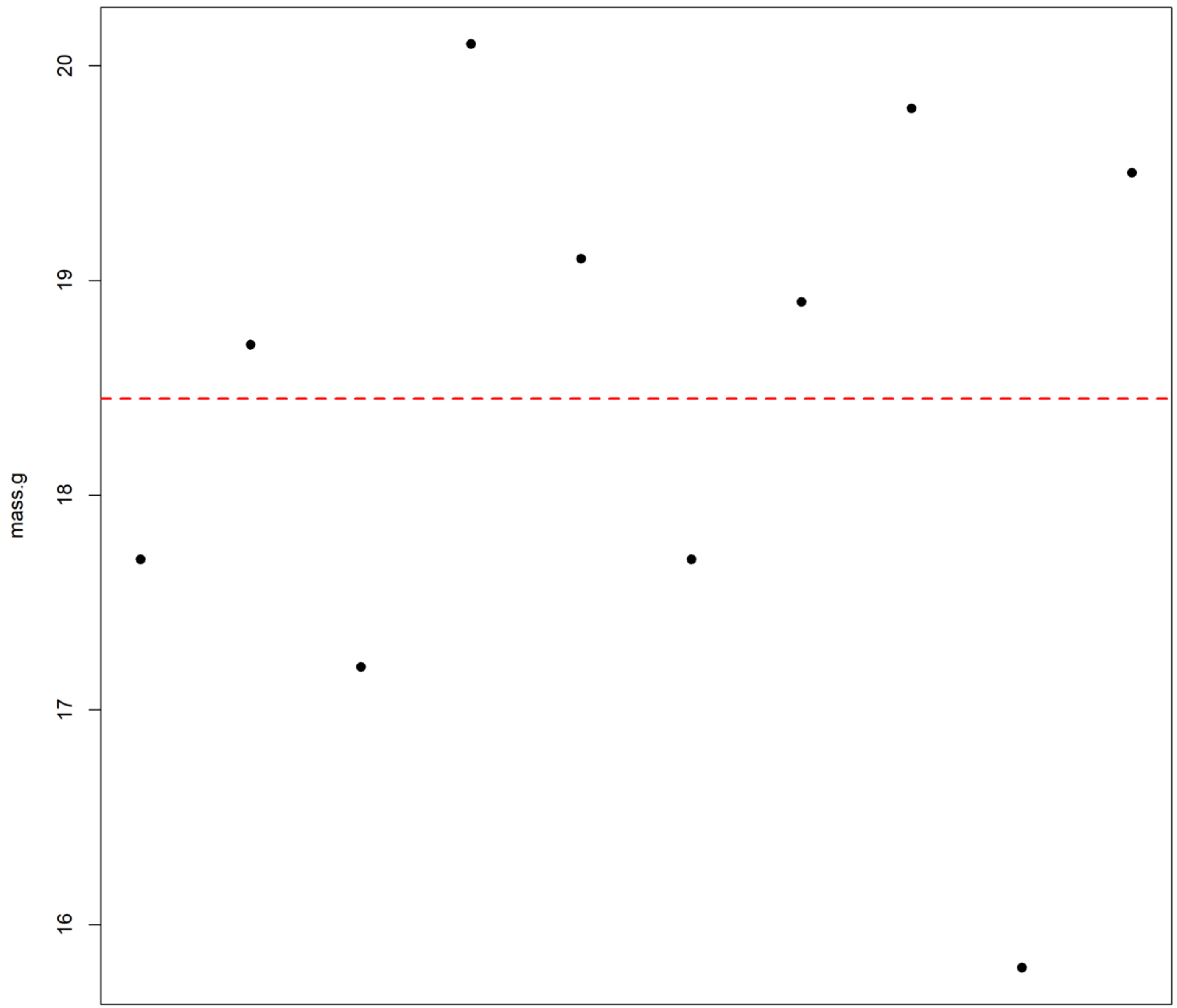
16 May 2022

> BILD 5 ->
Data analysis & design for biologists



How can we predict future fruit yields?

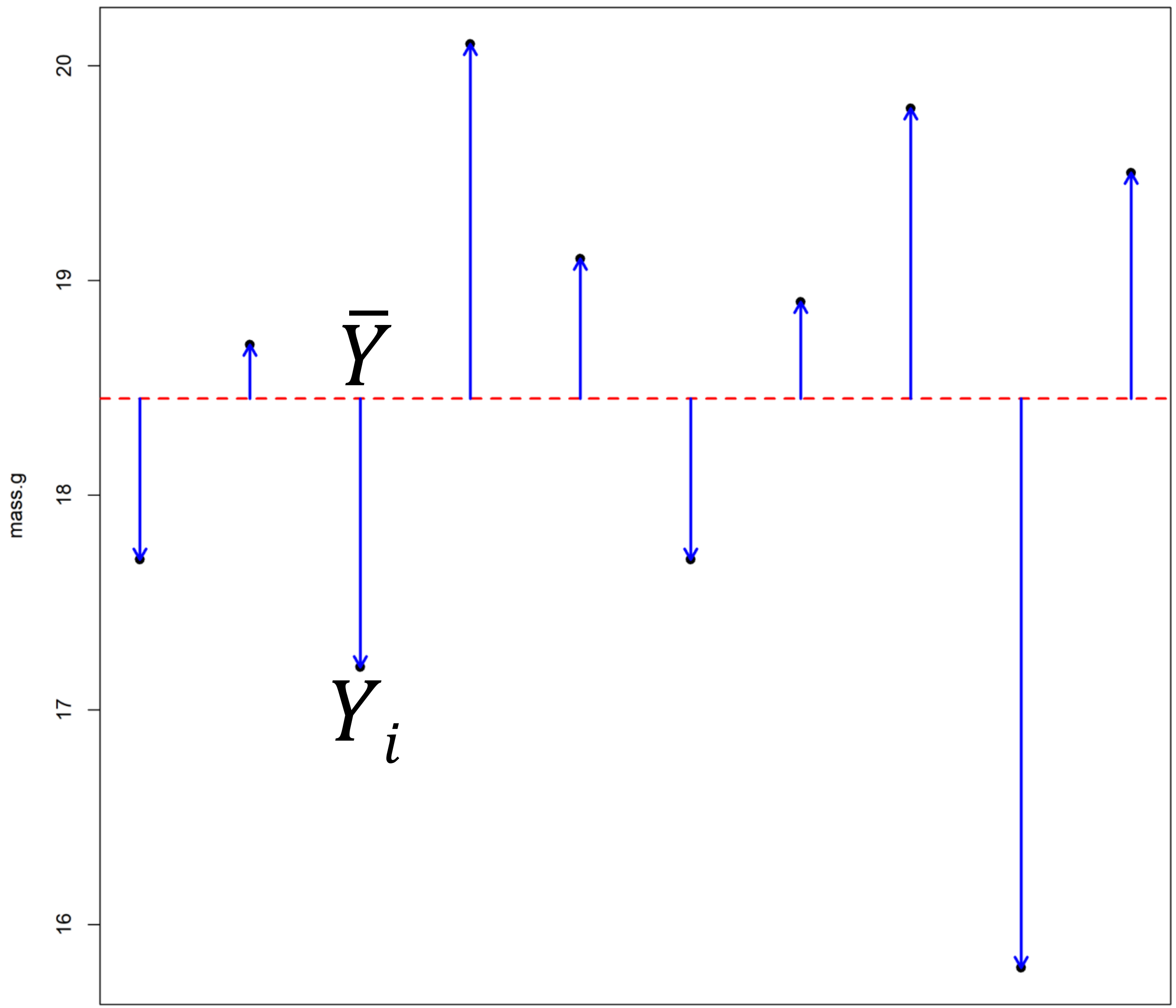
- Data from 10 individual plants
- Yield of fruit in grams
- What will our next yield be?



How can we predict future fruit yields?

$\bar{Y}$

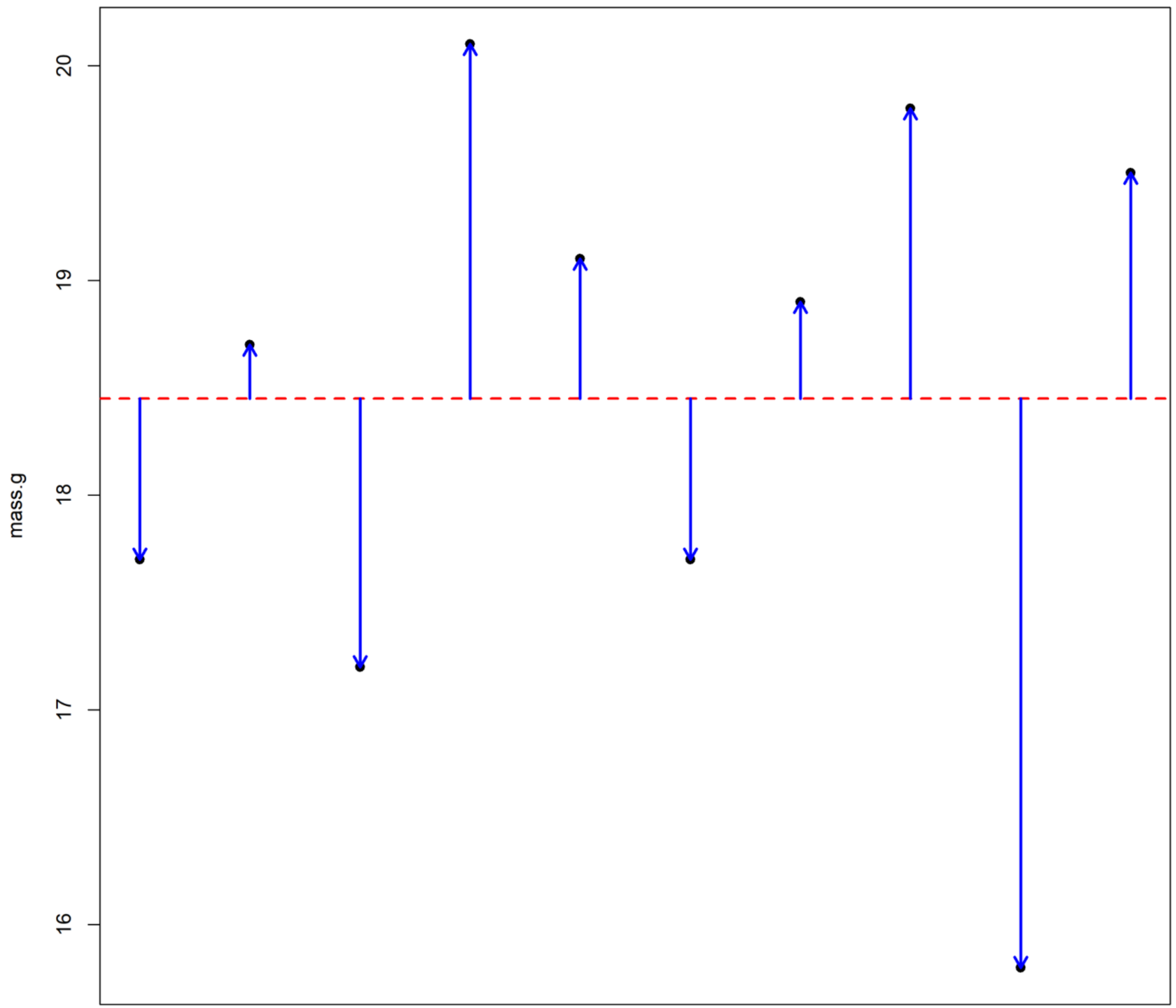
mean



How can we predict future fruit yields?

$$Y_i - \bar{Y}$$

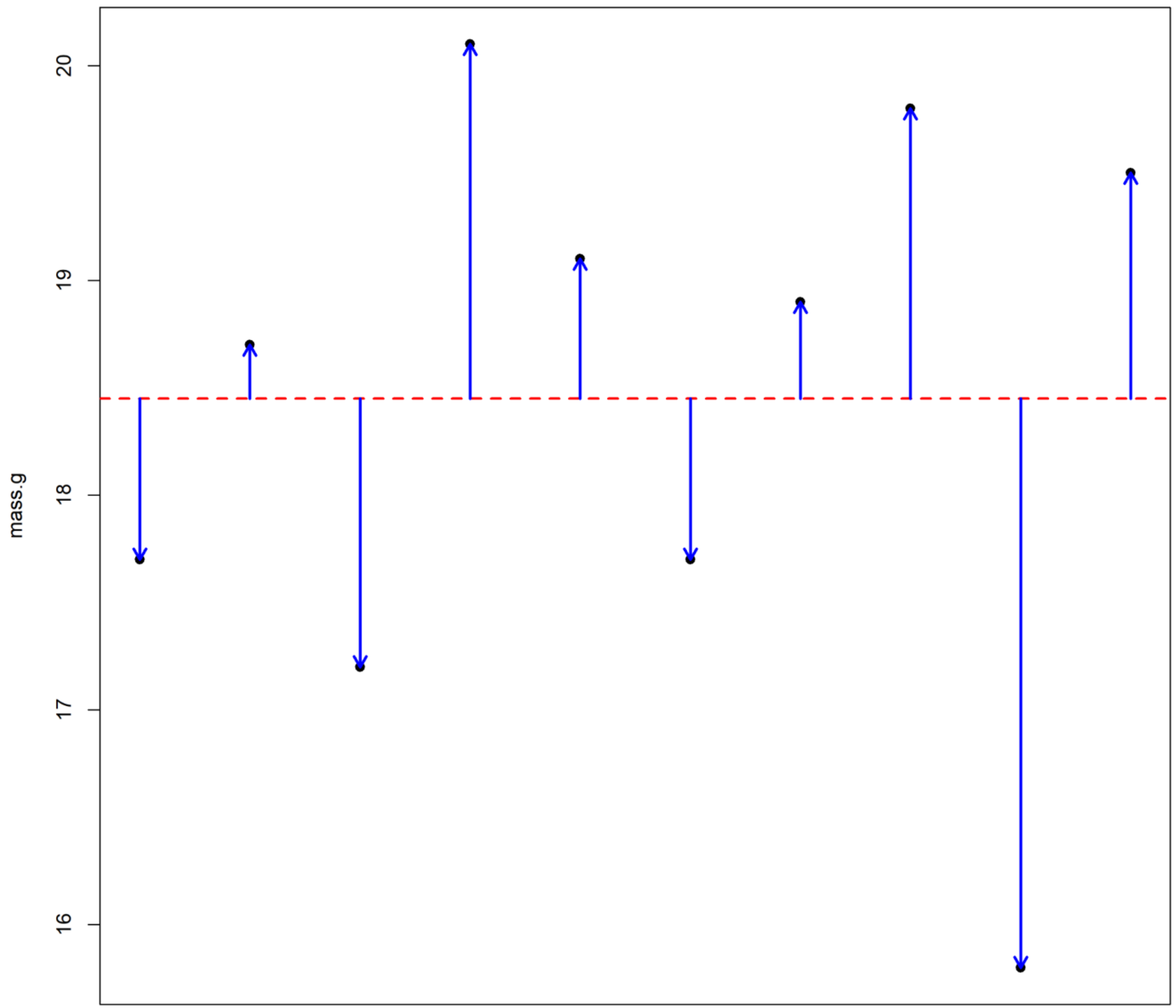
observation mean



How can we predict future fruit yields?

Sum of Squared Deviations

$$\sum (Y_i - \overline{Y})^2$$



How can we predict future fruit yields?

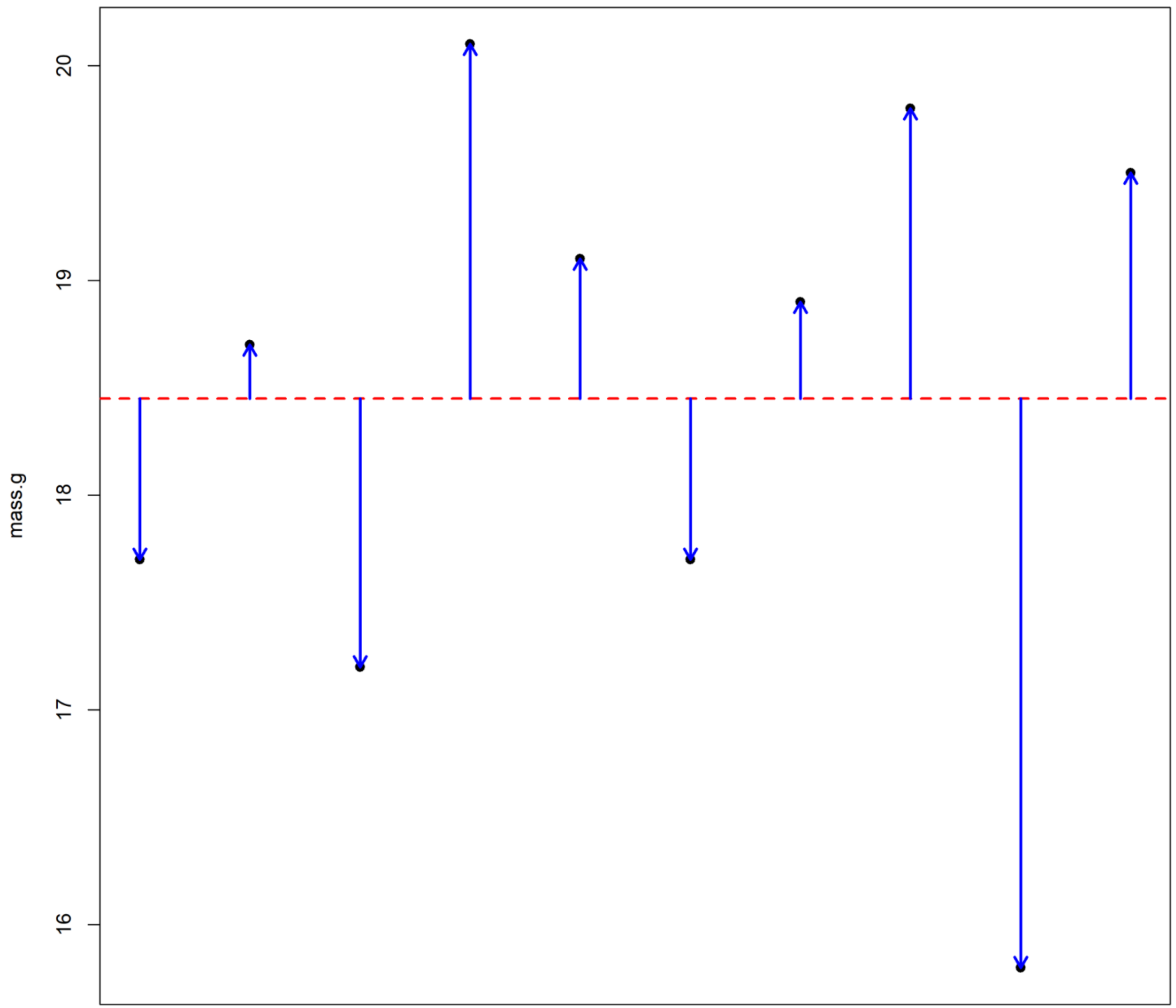
Sum of Squared Deviations

$$\sum (Y_i - \overline{Y})^2$$

1) $(17.7 - 18.45)^2$
2) $(18.7 - 18.45)^2$
3) $(17.2 - 18.45)^2$
4) $(20.1 - 18.45)^2$
5) $(19.1 - 18.45)^2$
6) $(17.7 - 18.45)^2$
7) $(18.9 - 18.45)^2$
8) $(19.8 - 18.45)^2$
9) $(15.8 - 18.45)^2$

\+ 10) $(19.5 - 18.45)^2$

Sum of Squared Deviations = 16.045

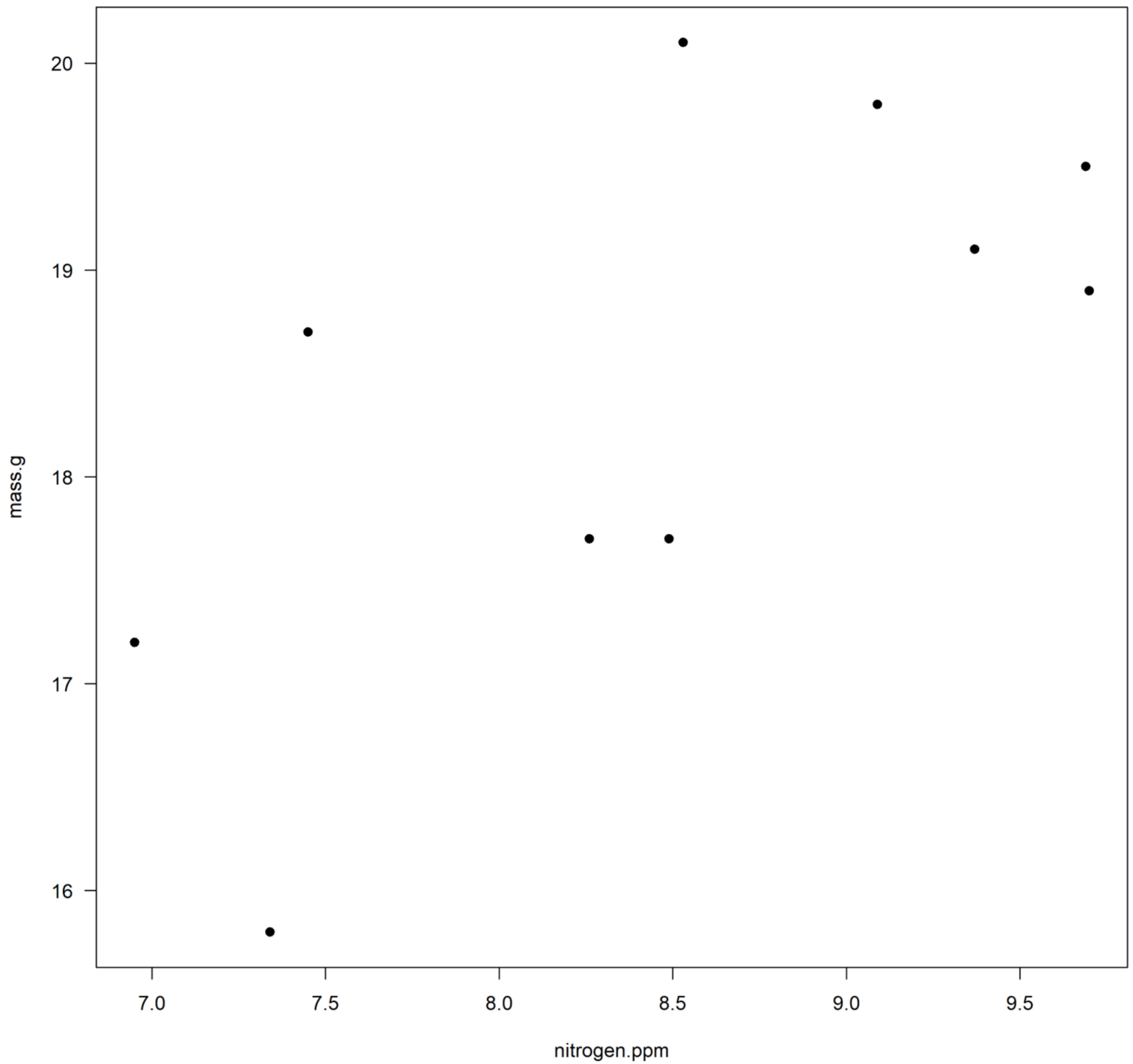


How can we predict future fruit yields?

Sum of squares

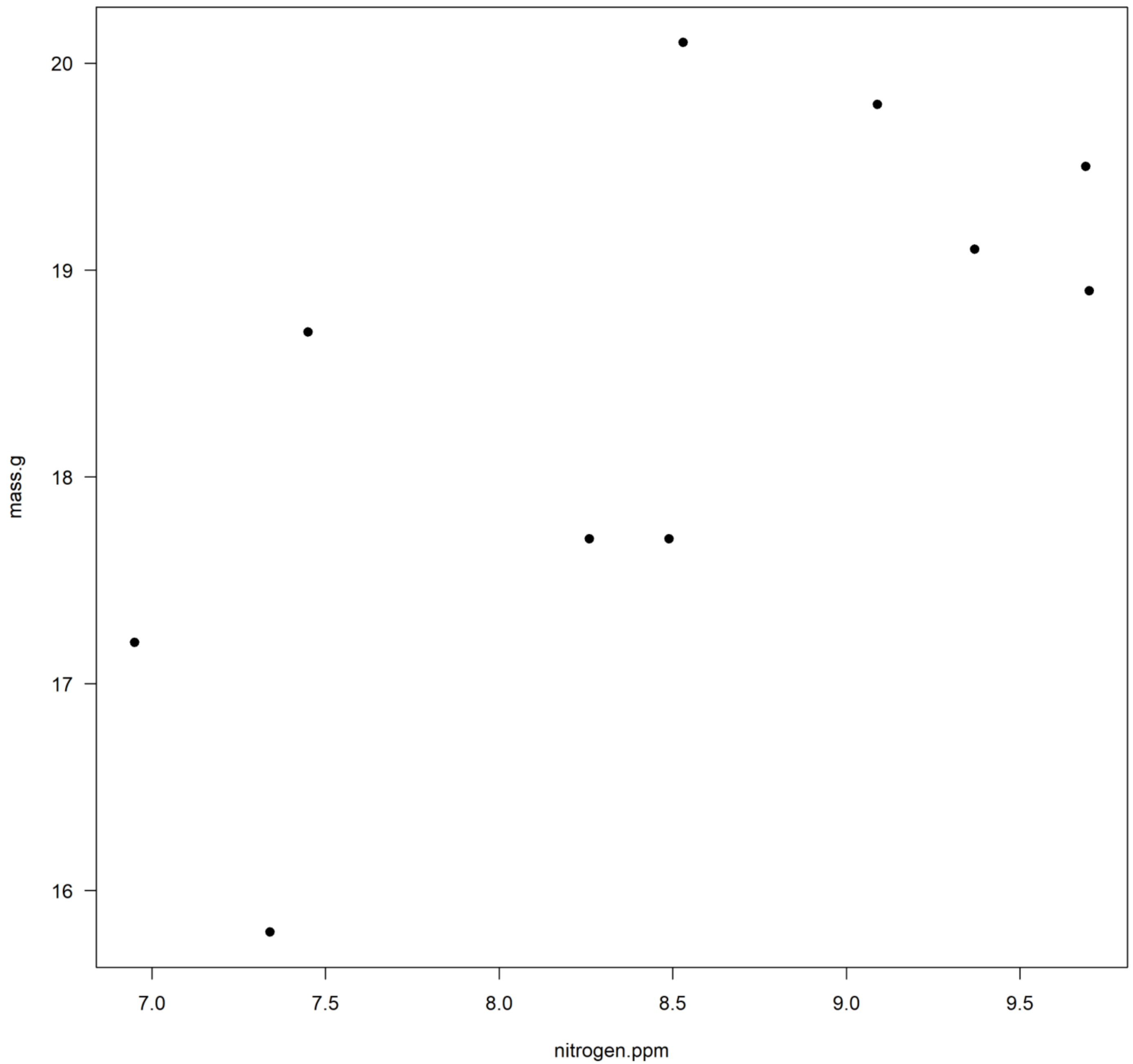
$$\text{Variance} = \frac{\sum(Y_i - \overline{Y})^2}{n-1}$$

Degrees of freedom



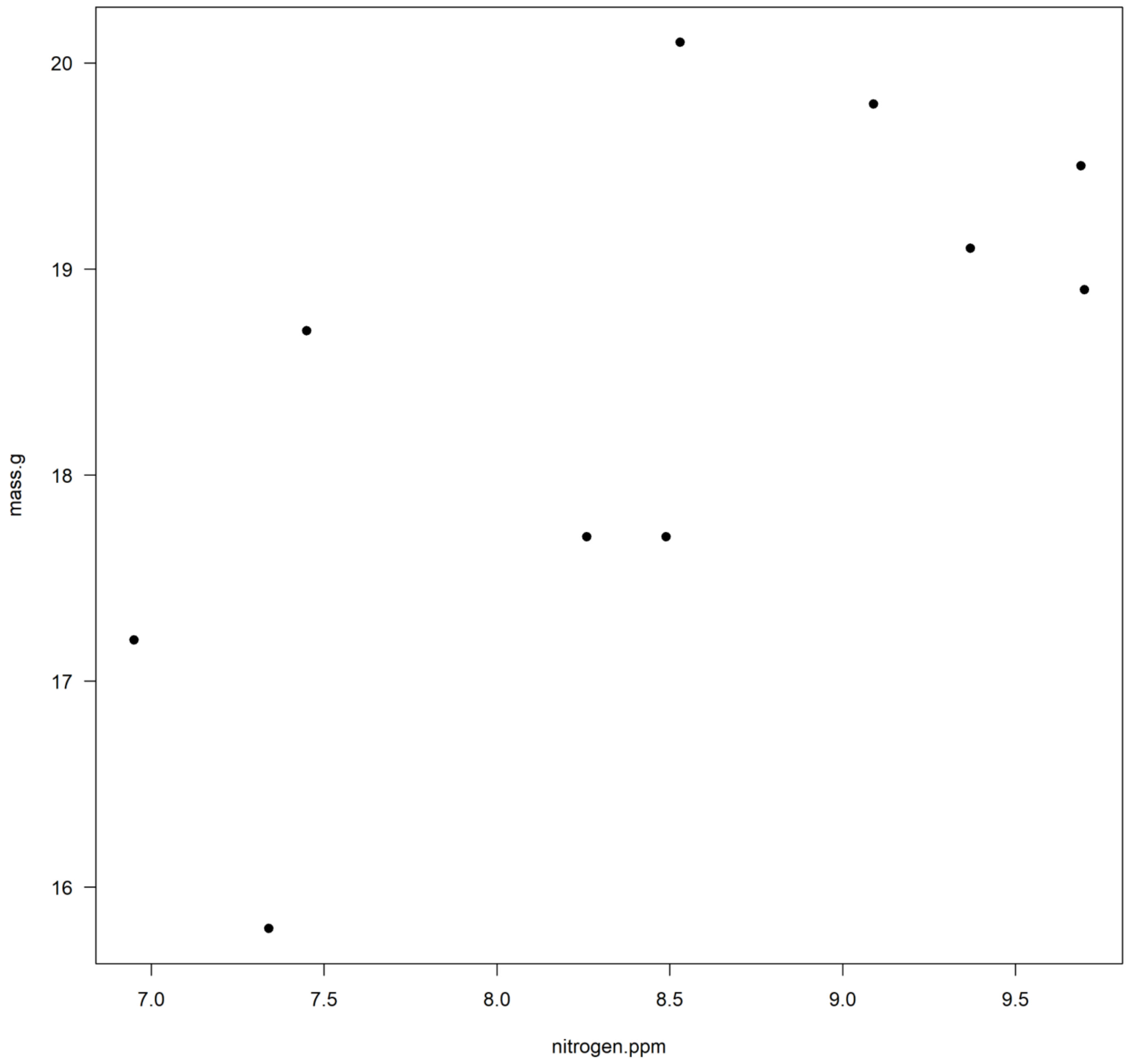
What if we know more about our system?

- Mass in grams of our fruit yield
- Concentration in parts per million of nitrogen in the soil



What if we know more about our system?

- Fruit mass likely ***depends*** upon the nitrogen available to the plant.



What if we know more about our system?

- Fruit mass likely ***depends*** upon the nitrogen available to the plant.
- The starting nitrogen concentration is likely ***independent*** of the fruit on the plant.

General Linear Models

- Model how a dependent variable (Y) changes over an independent variable (X)

General Linear Models

- Model how a dependent variable (Y) changes over an independent variable(X)
- “regression”
- $Y = mX + b$

General Linear Models

dependent variable

independent variable

$$Y = mX + b$$

slope

Y intercept

General Linear Models

$$Y = mX + b$$

$$Y_i = \underbrace{\beta_1}_{\text{slope}} X_i + \underbrace{\beta_0}_{\text{Y intercept}}$$

Parameters

General Linear Models

$$Y = mX + b$$

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

$$Y_i = b_1 X_i + b_0 + \textit{error}$$

slope Y intercept

Statistics

General Linear Models

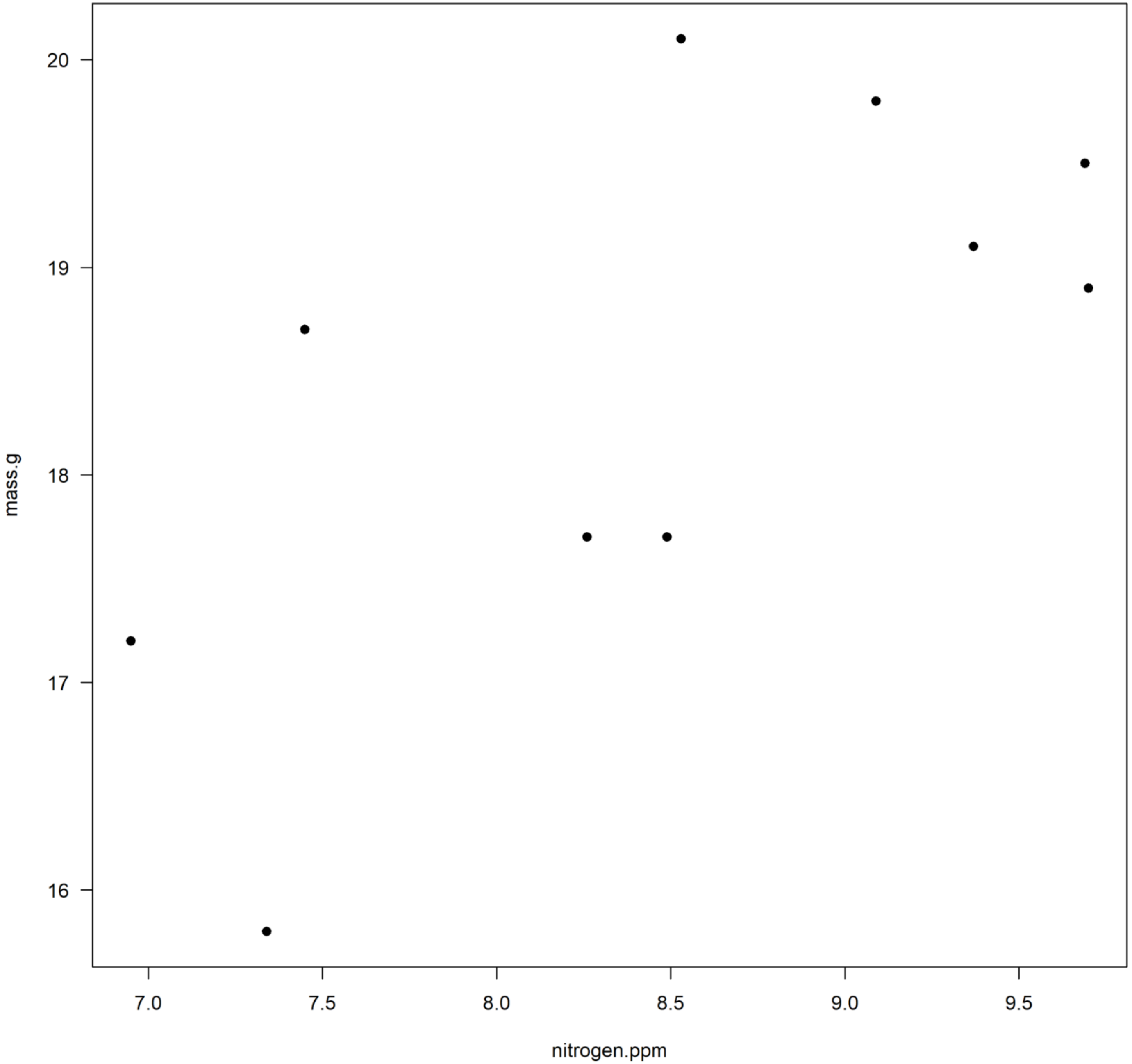
- Model how a dependent variable (Y) changes over an independent variable(X)
- “regression”
- $Y = mX + b$
- Does not have to be a straight line!

General Linear Models

- Model how a dependent variable (Y) changes over an independent variable(X)
- “regression”
- $Y = mX + b$
- Does not have to be a straight line!
- No parameter in the model is multiplied by another parameter
- $Y_i = b_1X_i + b_0 + \textit{error}$

Small Group Discussion - 3 min.

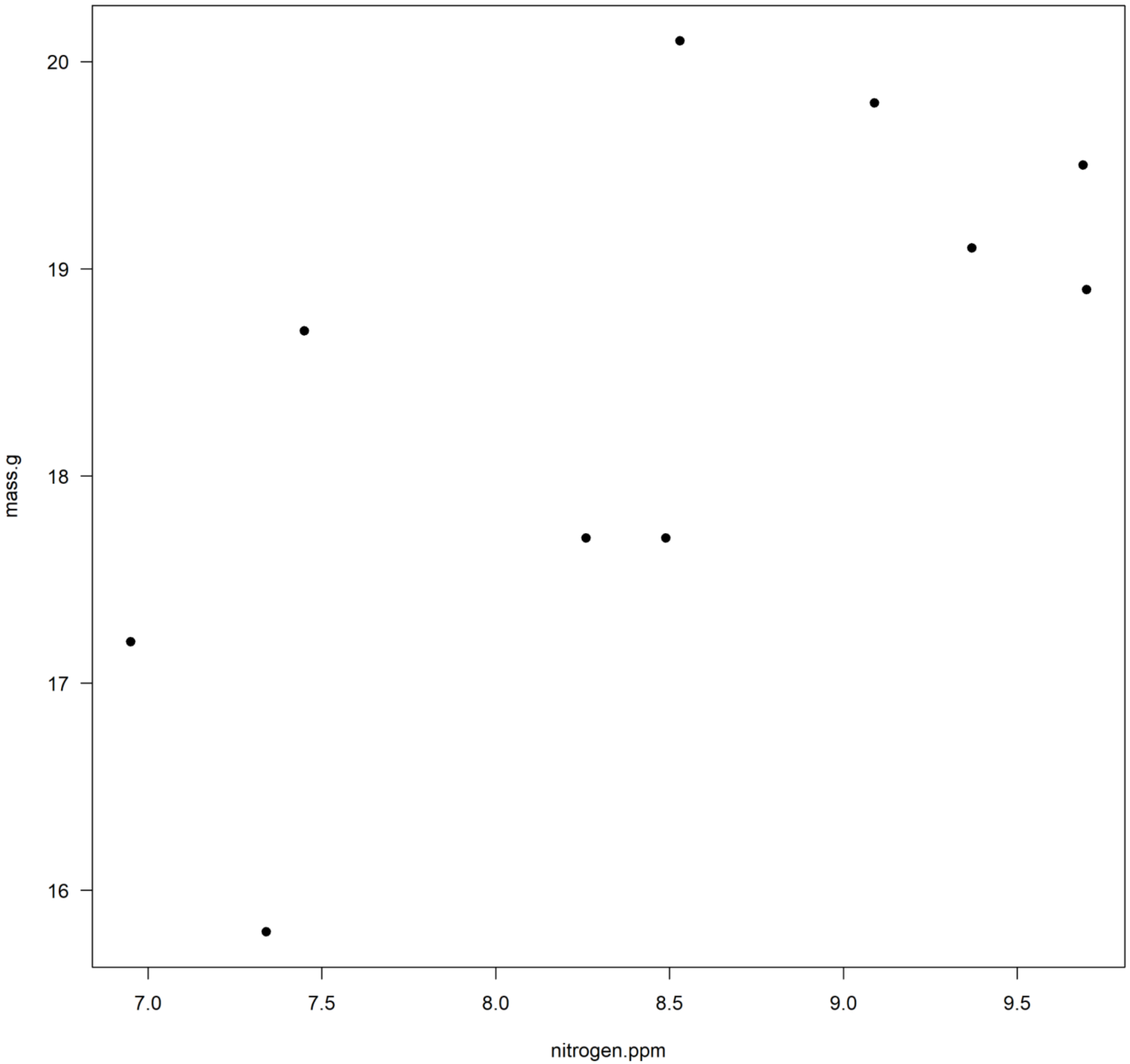
- How is the linear model different from a correlation model?
- Come up with a few examples of data that would be better modeled as a general linear model than as a correlation/covariation.



So you want to use a linear model?

How do we estimate β_1 and β_0?

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

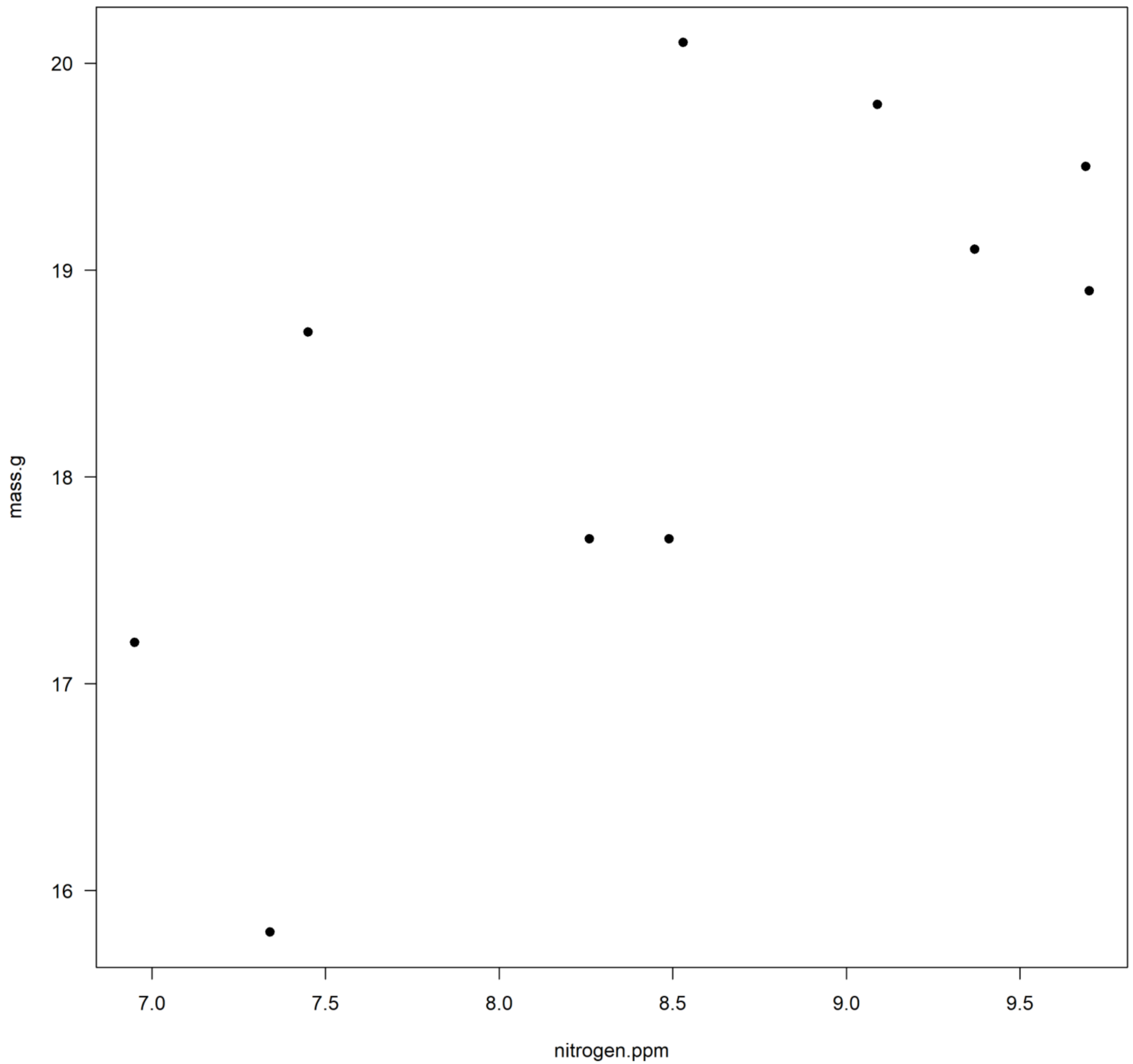


So you want to use a linear model?

How do we estimate β_1 and β_0?

Ordinary Least Squares

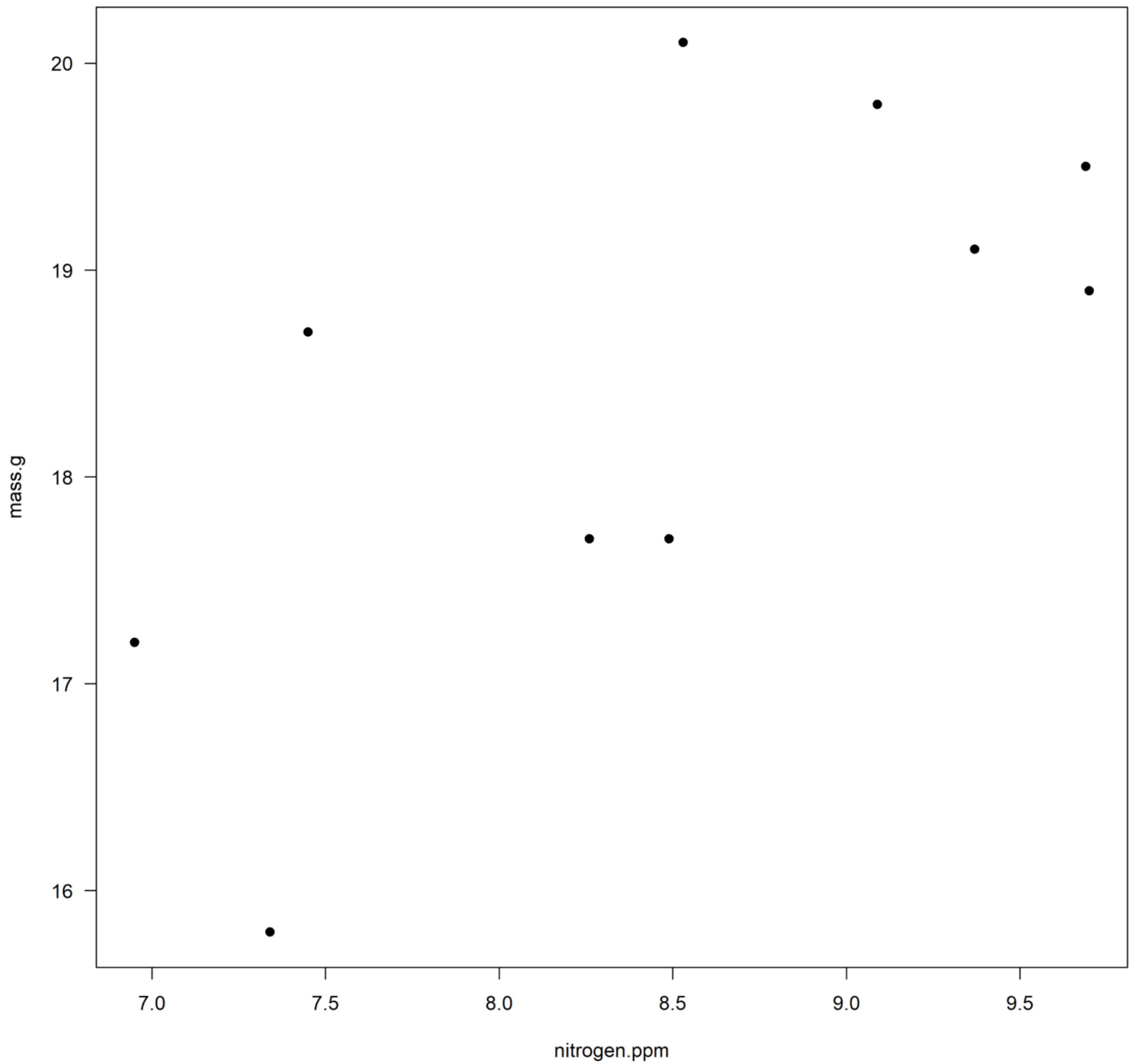
$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$



How do we estimate β_1 and β_0? Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

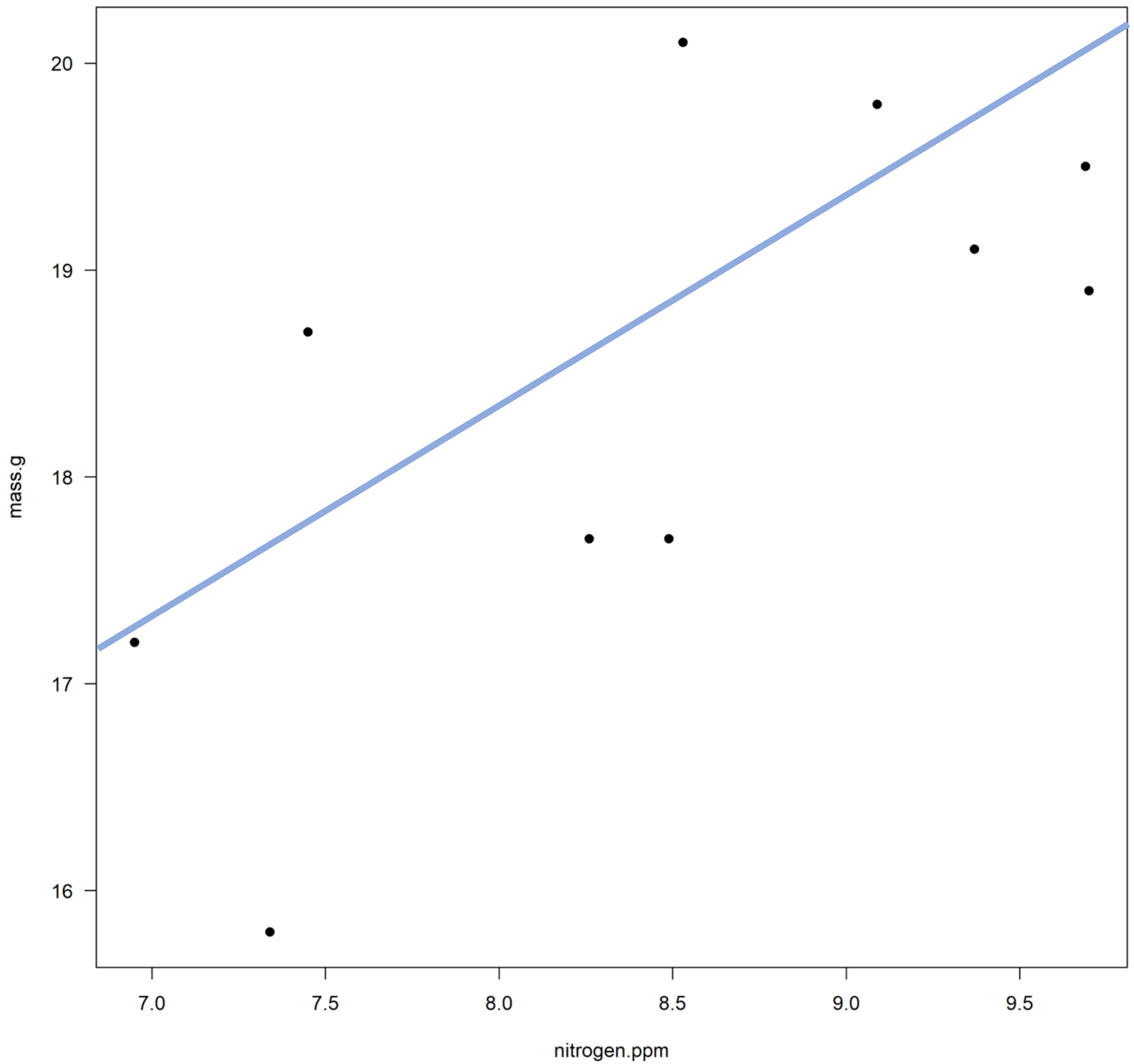
$$b_1 = \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{(X_i - \overline{X})(X_i - \overline{X})}$$



How do we estimate β_1 and β_0?
Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

$$b_1 = \frac{\cancel{(X_i - \overline{X})}(Y_i - \overline{Y})}{\cancel{(X_i - \overline{X})}(X_i - \overline{X})}$$



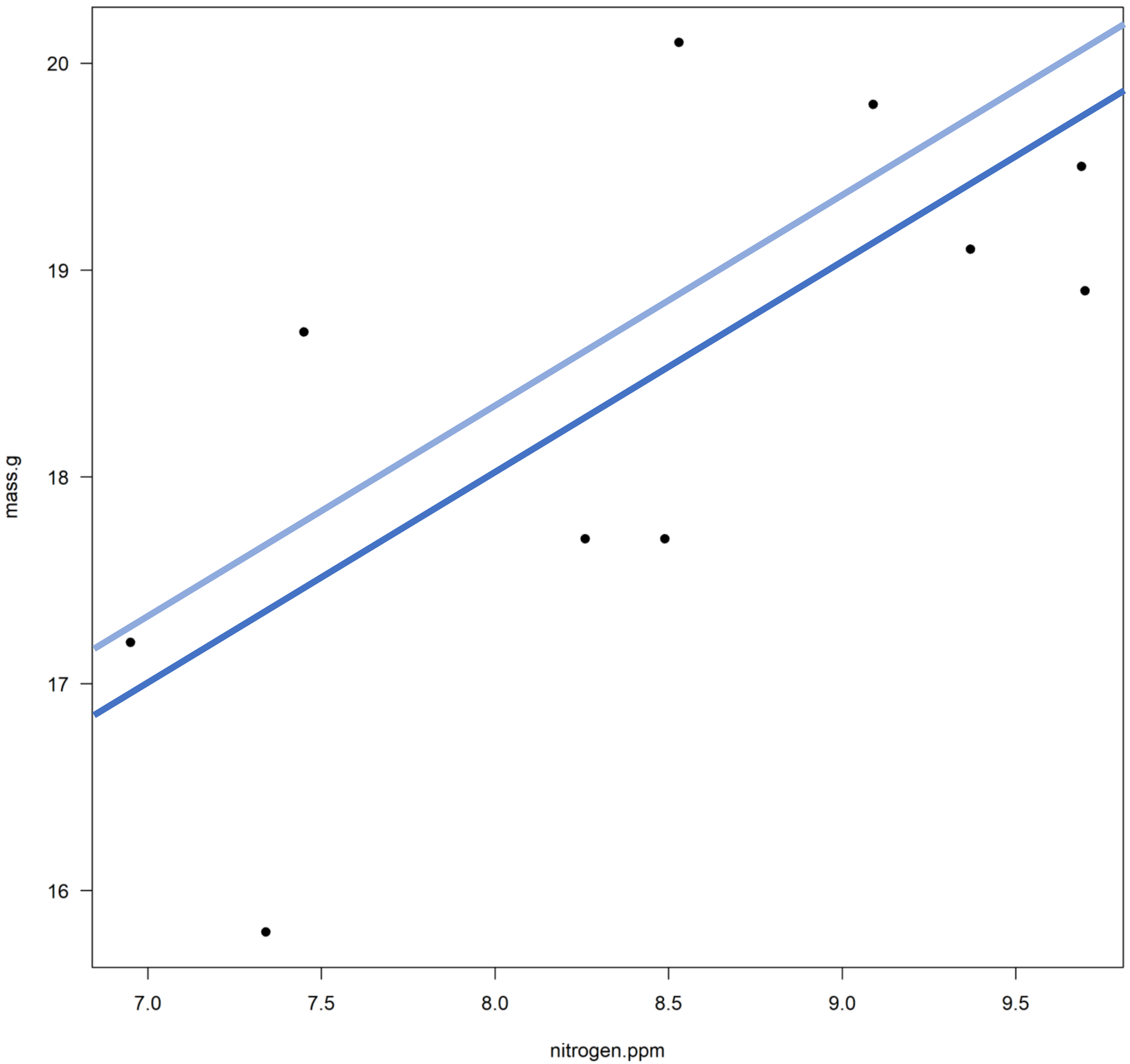
How do we estimate β_1 and β_0?
Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

Rise

$$b_1 = \frac{\cancel{(X_i - \overline{X})}(Y_i - \overline{Y})}{\cancel{(X_i - \overline{X})}(X_i - \overline{X})}$$

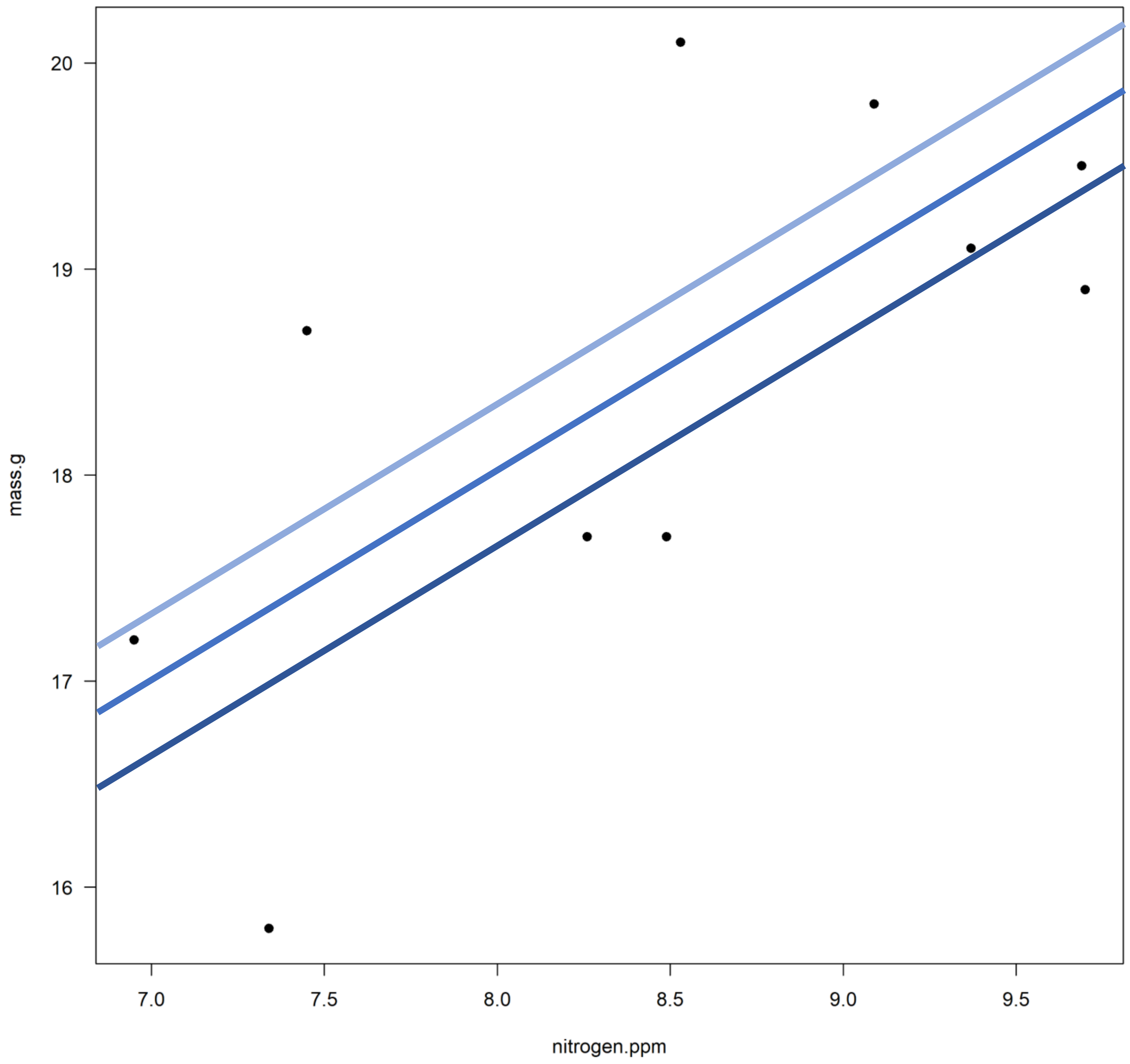
Run



How do we estimate β_1 and β_0?

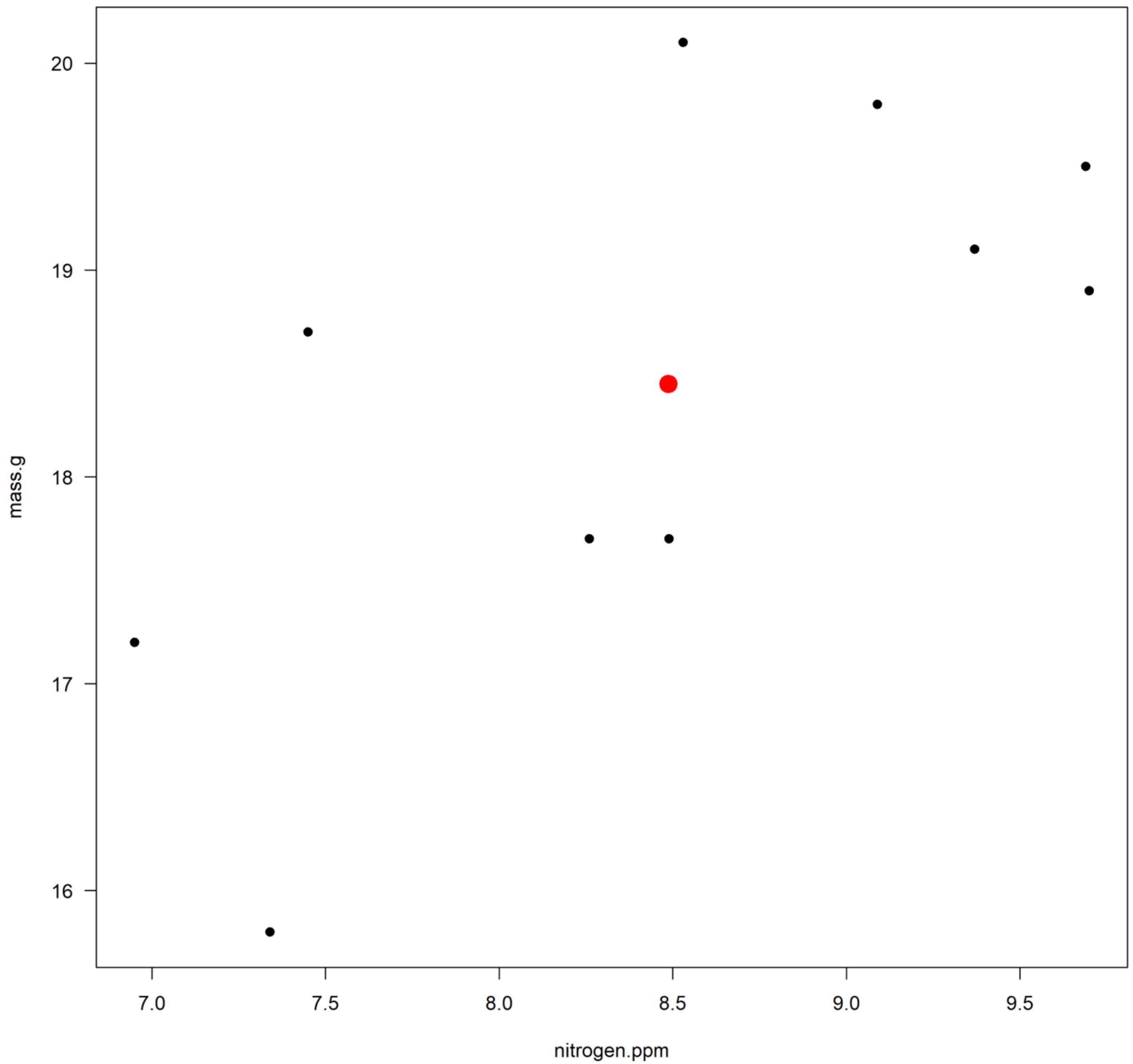
Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$



How do we estimate β_1 and β_0?
Ordinary Least Squares

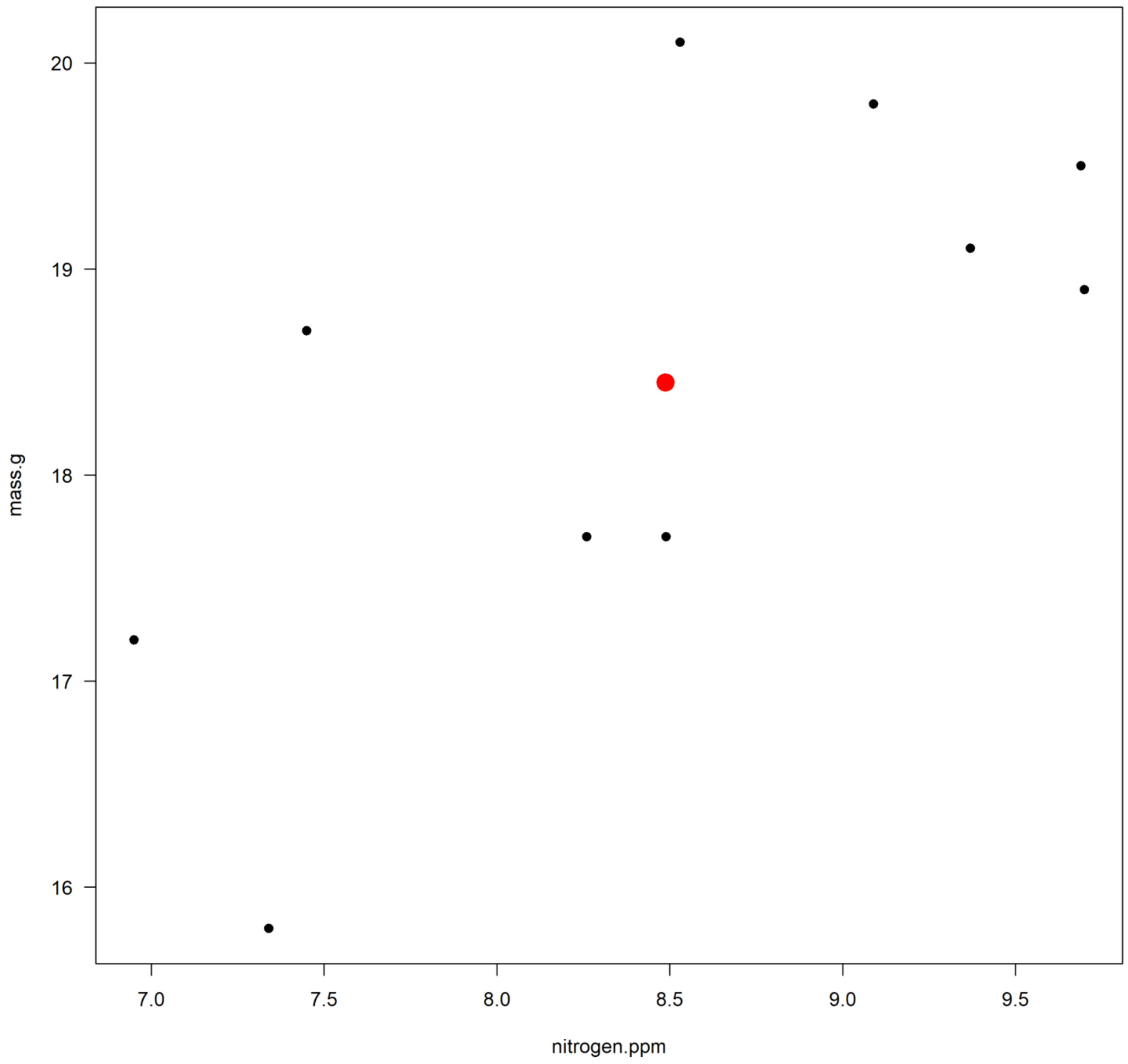
$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$



How do we estimate β_1 and β_0? Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

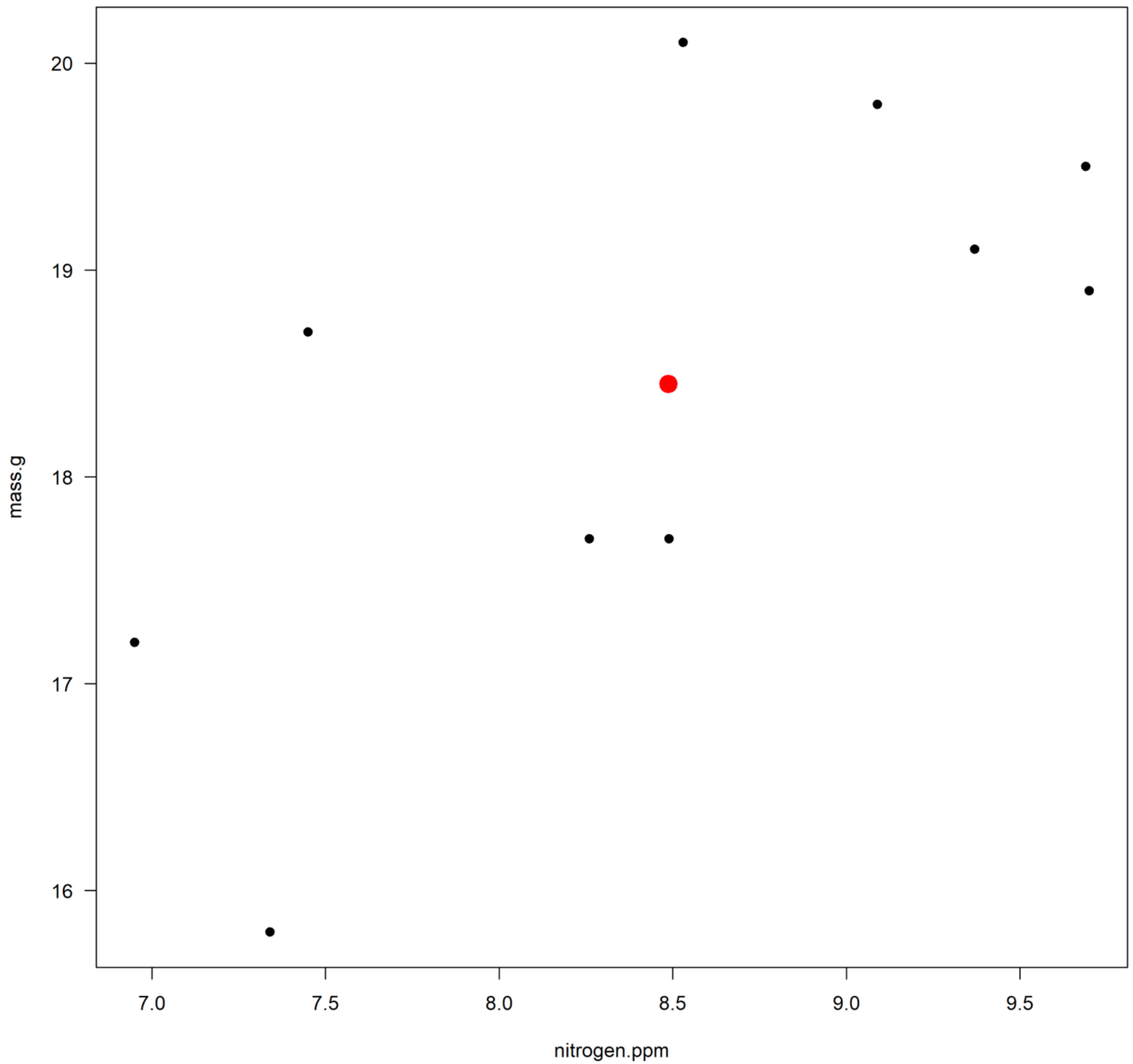
$$Y_i = b_1 \cdot X_i + b_0$$



How do we estimate β_1 and β_0?
Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

$$Y_i = b_1 \cdot X_i + b_0$$

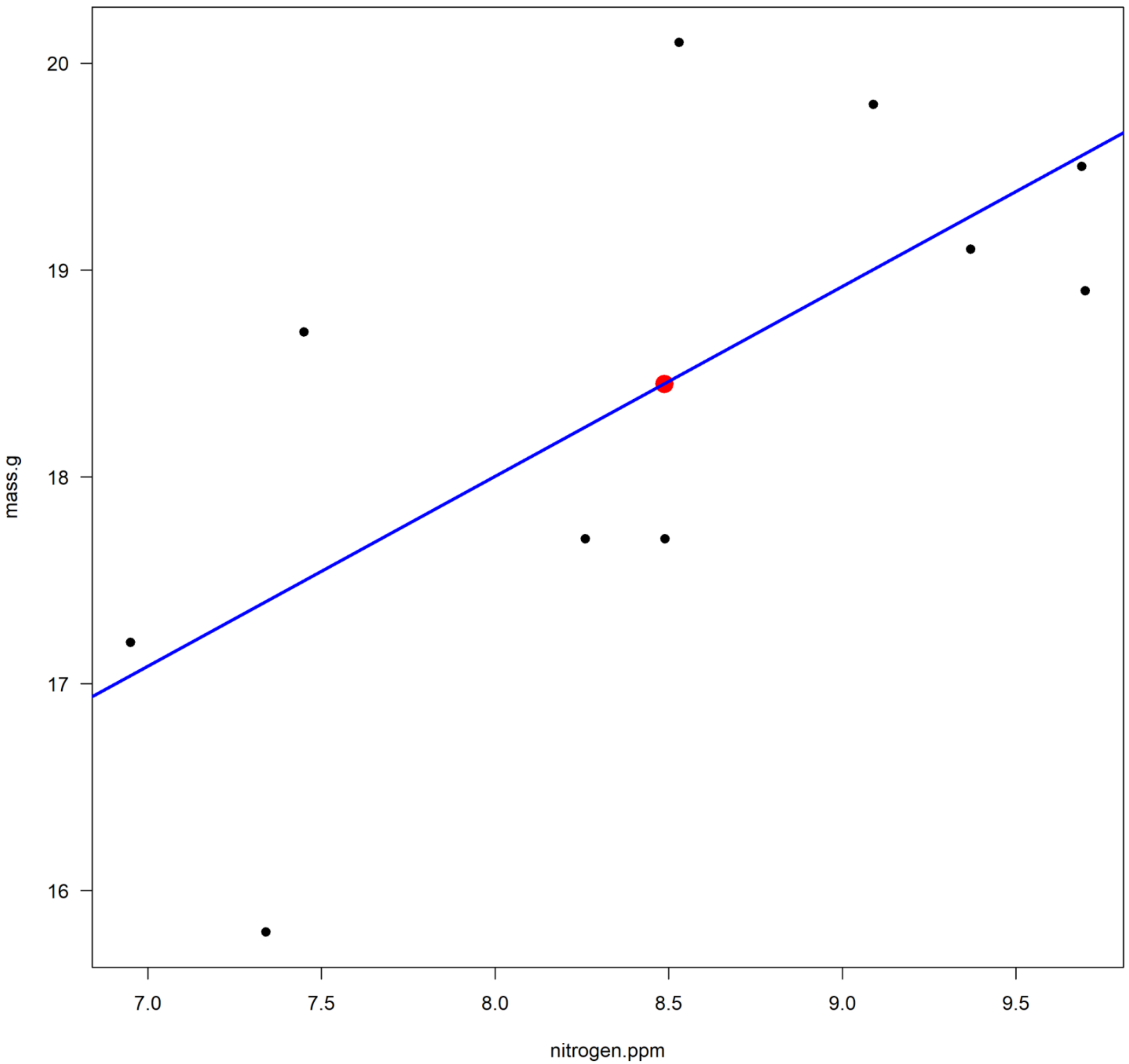


How do we estimate β_1 and β_0?
Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

$$Y_i = b_1 \cdot X_i + b_0$$

$$b_0 = \overline{Y} - b_1 \cdot \overline{X}$$



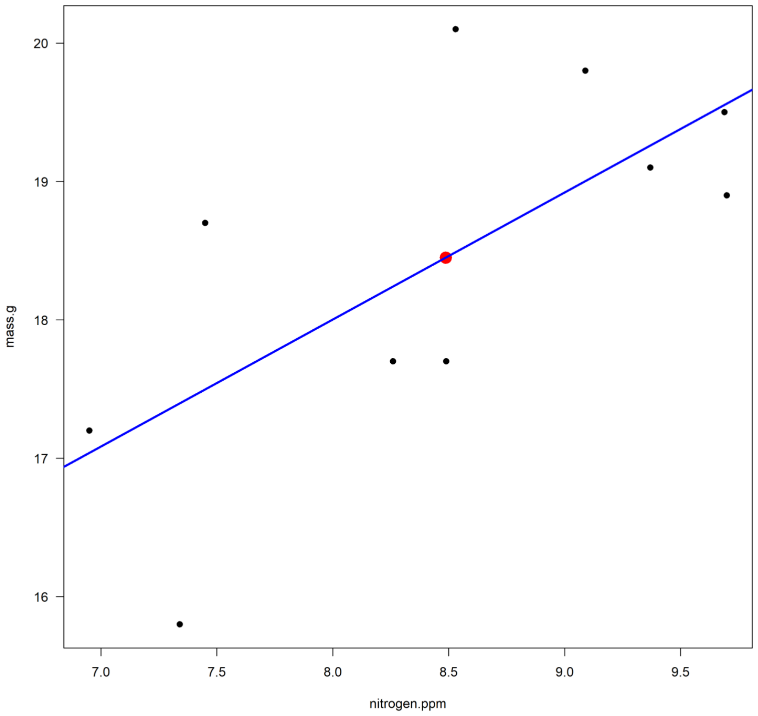
How do we estimate β_1 and β_0?

Ordinary Least Squares

$$\beta_1 \sim b_1 = \frac{cov(X,Y)}{var(X)}$$

$$\beta_0 \sim b_0 = \bar{Y} - b_1 \cdot \bar{X}$$

$$Y_i = b_1 X_i + b_0$$

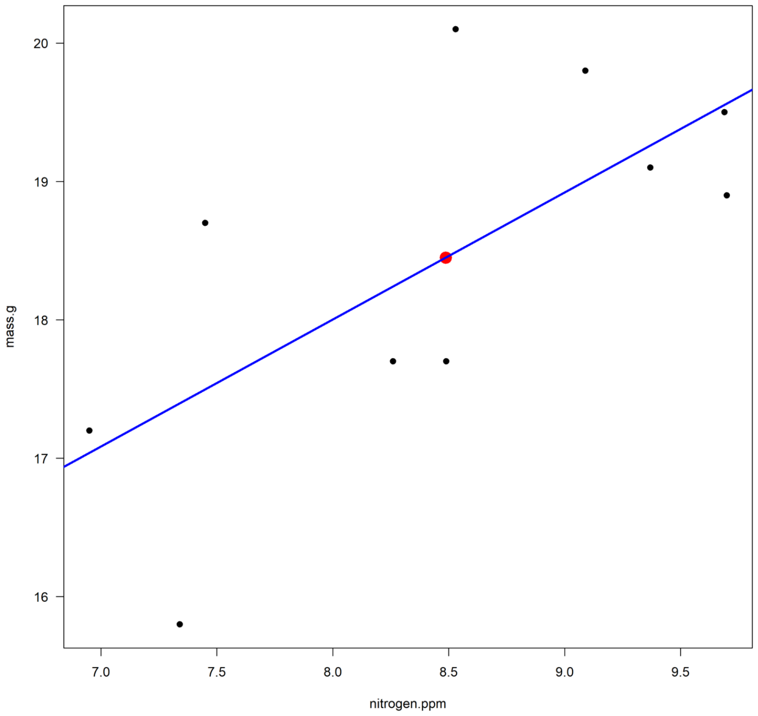


How do we estimate β_1 and β_0?

Ordinary Least Squares

$$Y_i = b_1 X_i + b_0$$

$$Y_i = 0.92 X_i + 10.66$$

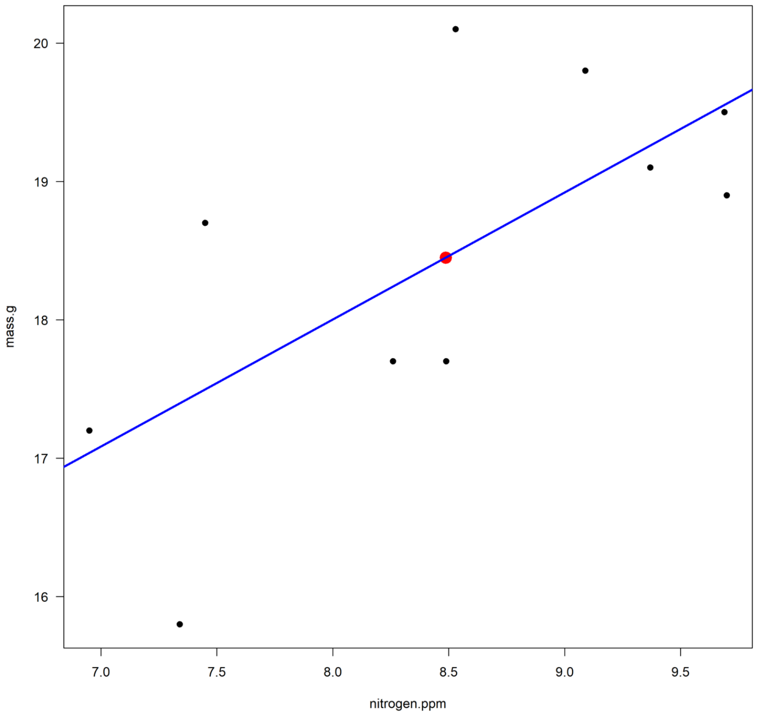


How do we estimate β_1 and β_0?
Ordinary Least Squares

$$Y_i = b_1 X_i + b_0$$

$$Y_i = 0.92 X_i + 10.66$$

$$\text{Expected Fruit Mass} = 0.92 \cdot (\text{Observed Nitrogen}) + 10.66$$



How do we estimate β_1 and β_0?

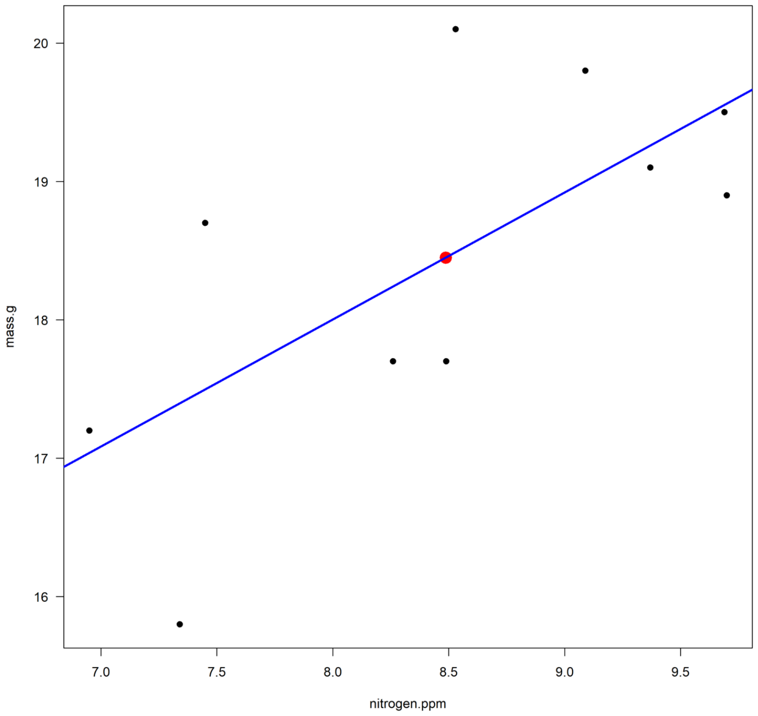
Ordinary Least Squares

$$Y_i = b_1 X_i + b_0$$

$$Y_i = 0.92 X_i + 10.66$$

Expected Fruit Mass = 0.92 · (Observed Nitrogen) + 10.66

Observed Nitrogen ↑ 1ppm : Expected Fruit Mass ↑ 0.9g



How do we estimate β_1 and β_0?
Ordinary Least Squares

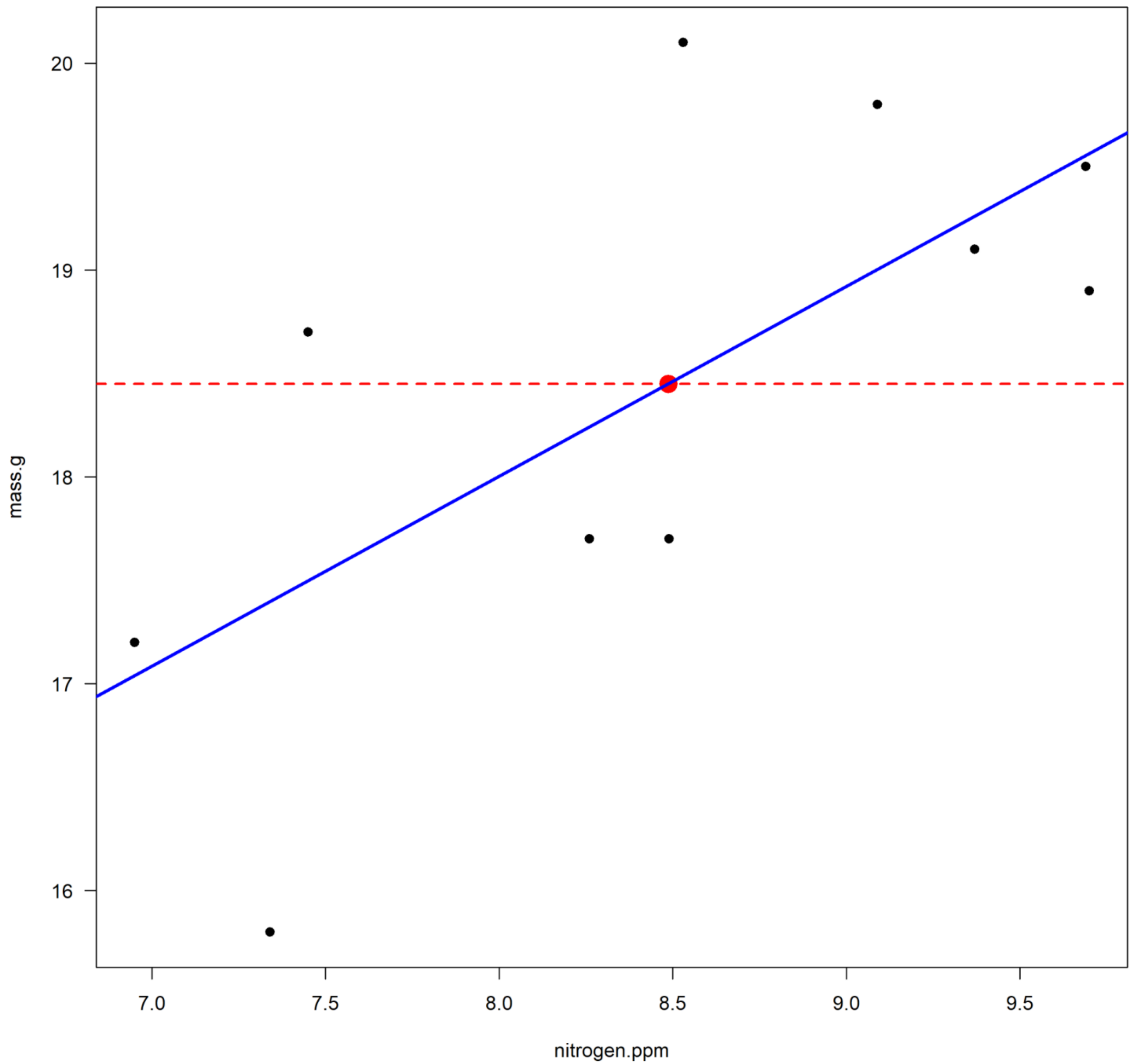
$$Y_i = b_1 X_i + b_0$$

$$Y_i = 0.92 X_i + 10.66$$

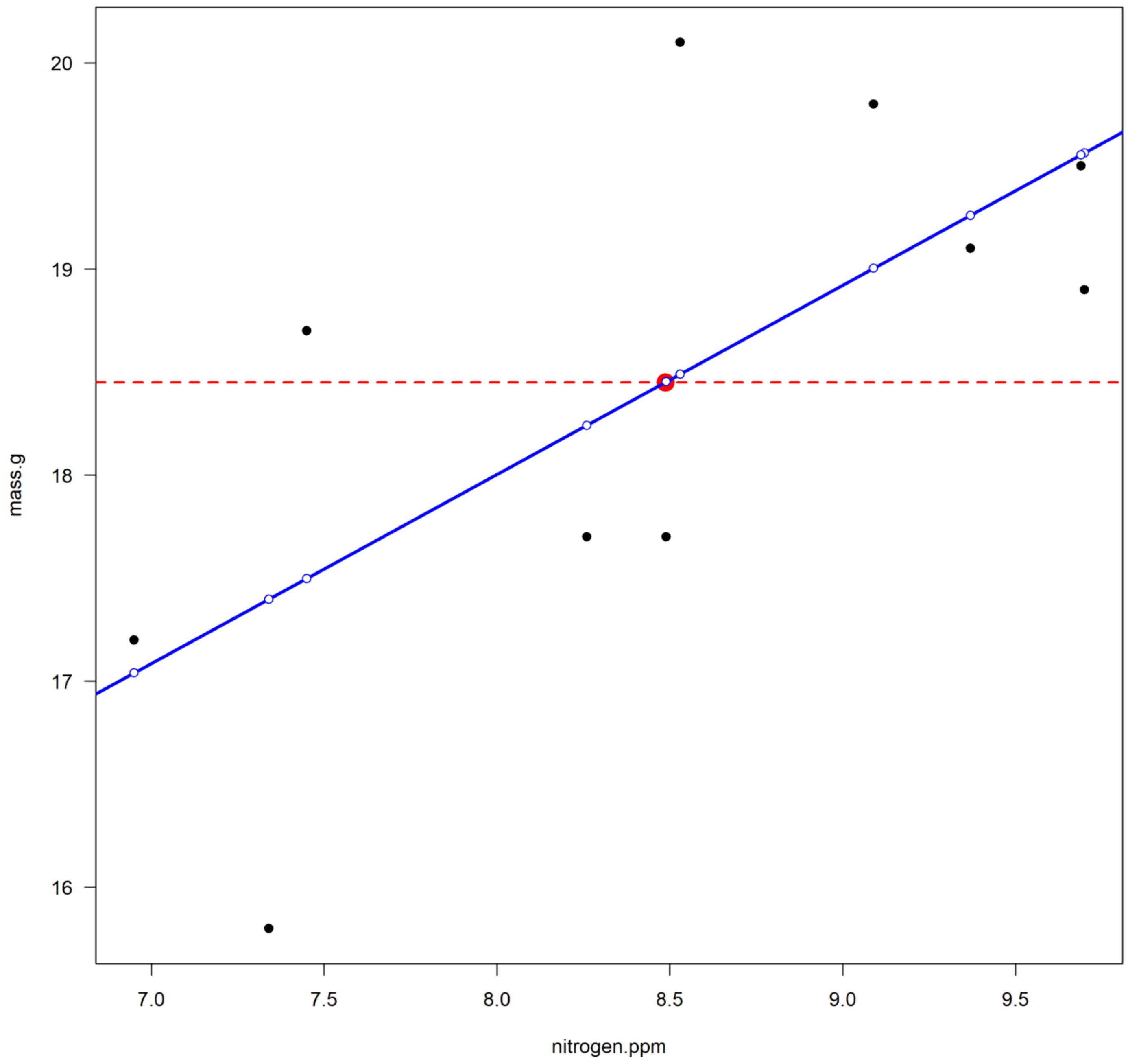
Expected Fruit Mass = 0.92 · (Observed Nitrogen) + 10.66

Observed Nitrogen = 0.0 : Expected Fruit Mass 10.66

Lets jump into R and see how this is done…



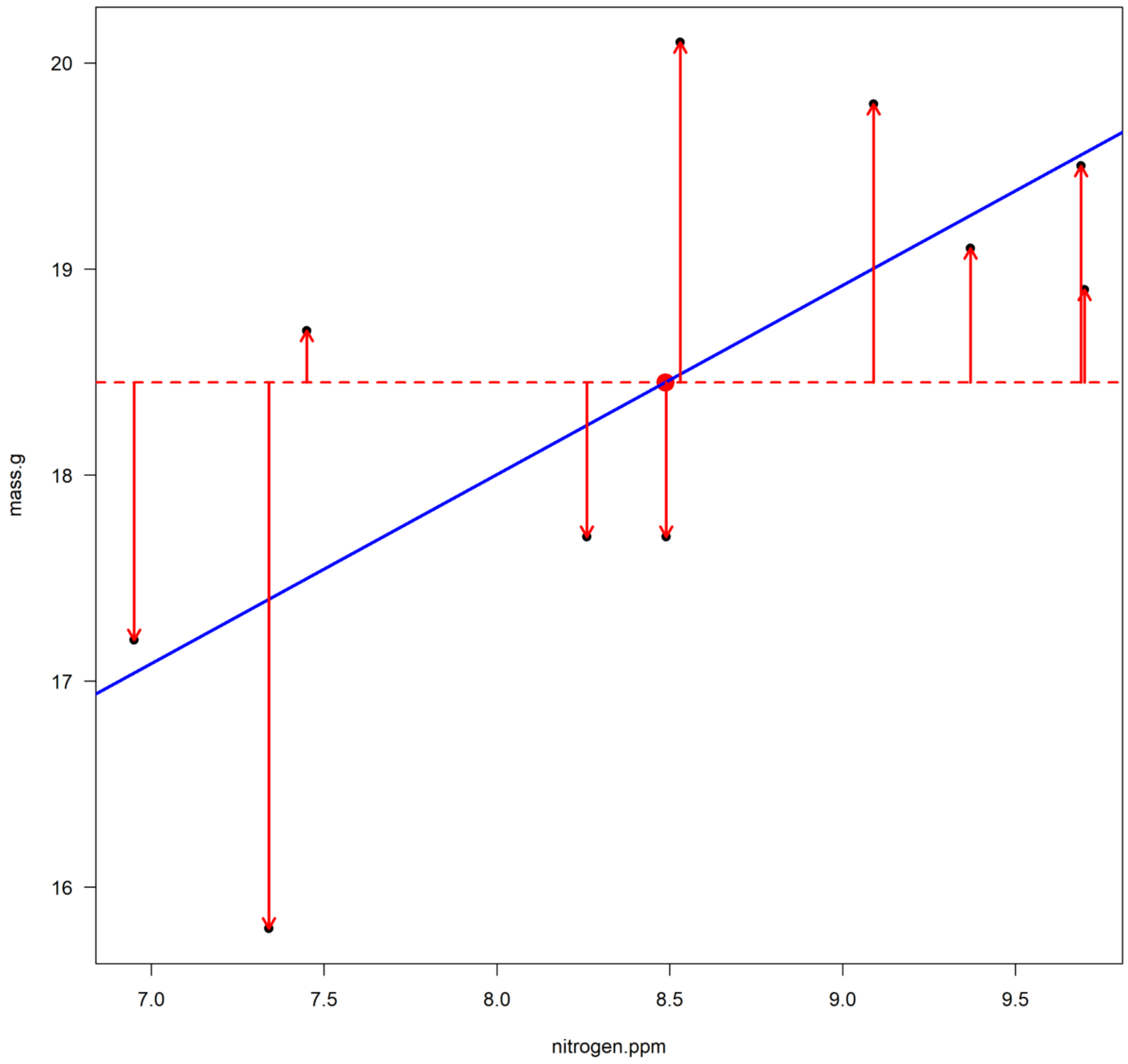
How much variation in Y does our model explain?



How much variation in Y does our model explain?

Predicted values
Fitted values $\hat{Y}_i$

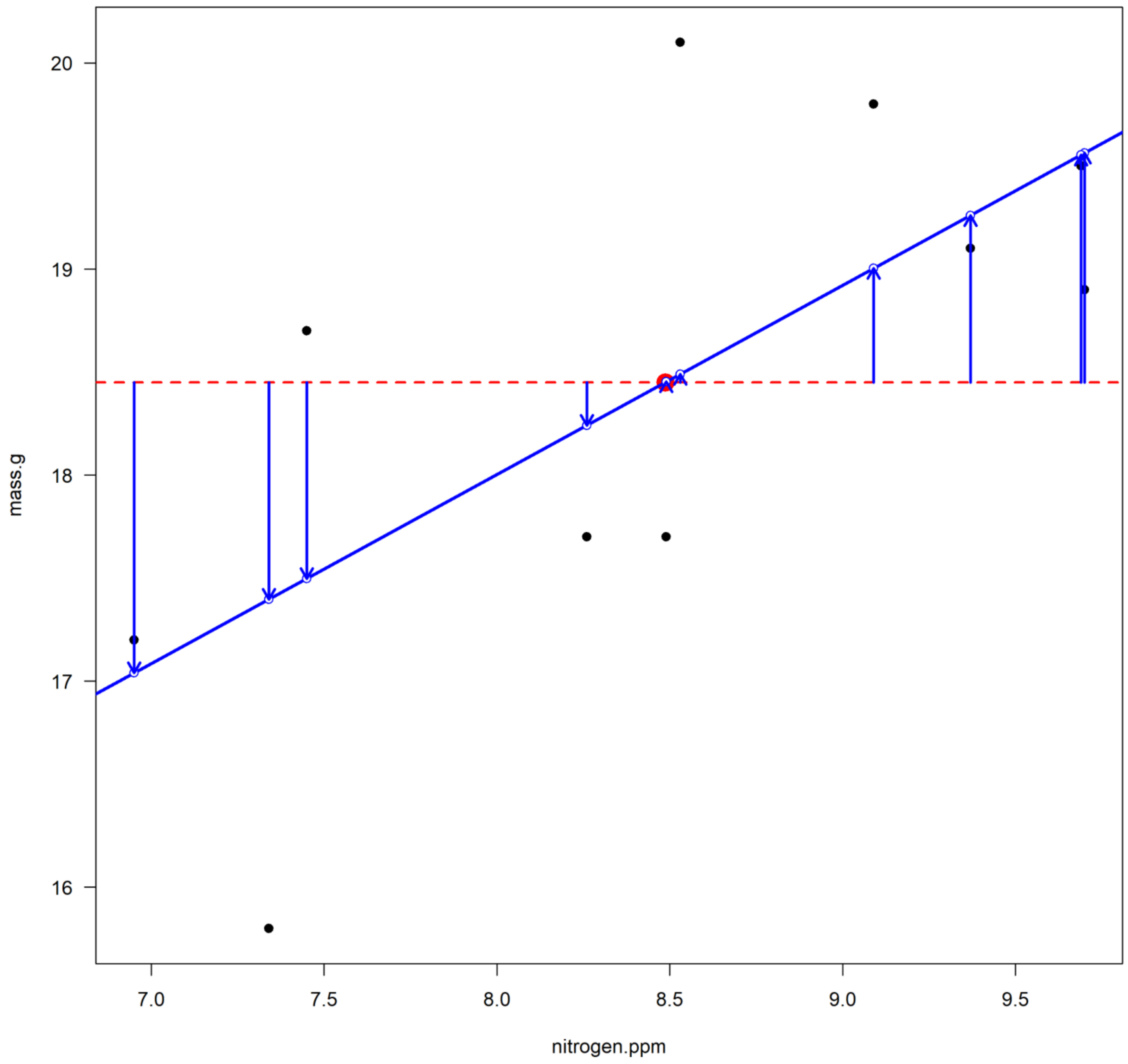
$$\hat{Y}_i = b_1 \cdot X_i + b_0$$



How much variation in Y does our model explain?

Total sum of squares
SS Total

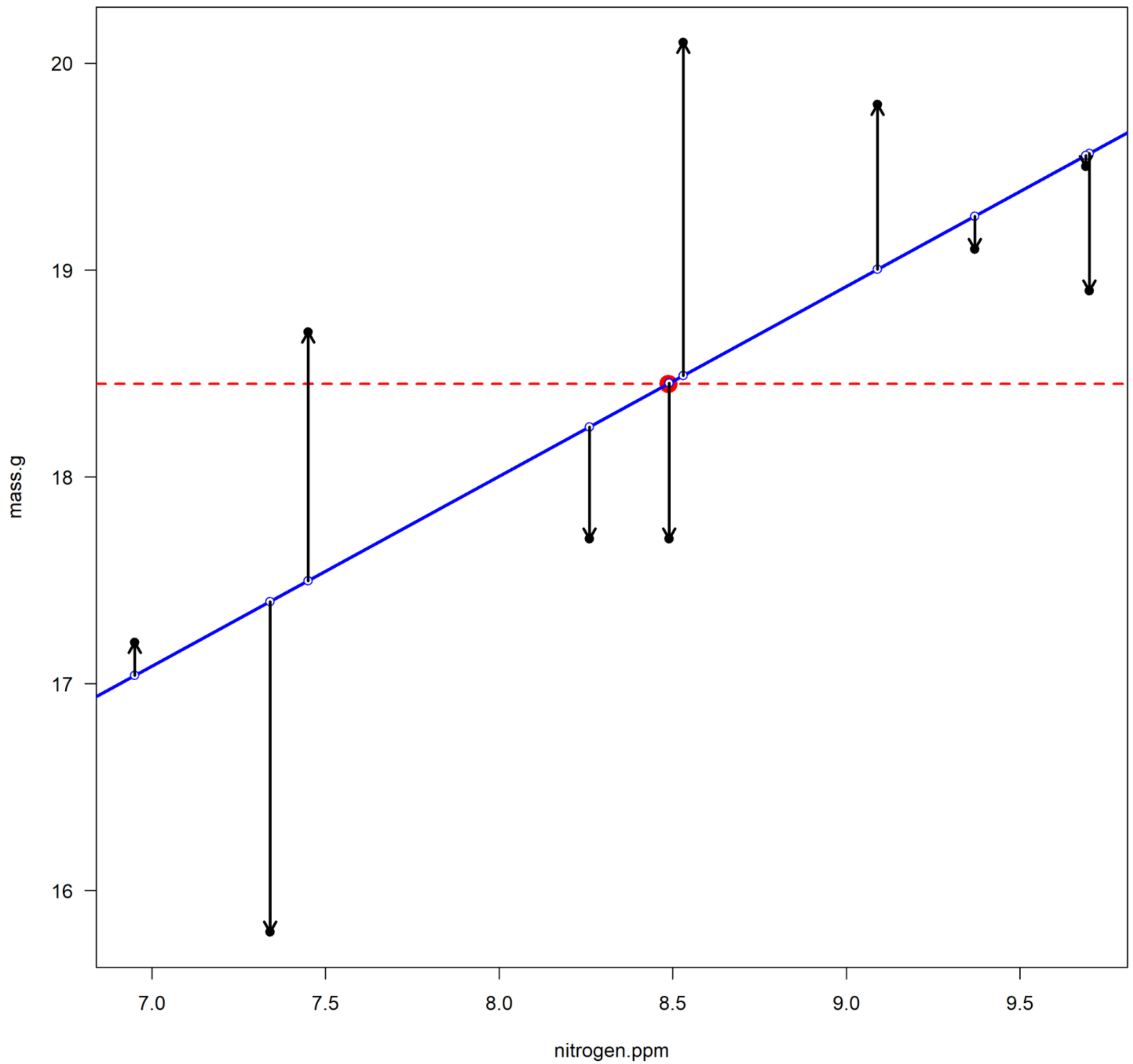
$$\sum (Y_i - \overline{Y})^2$$



How much variation in Y does our model explain?

Model sum of squares
SS Regression

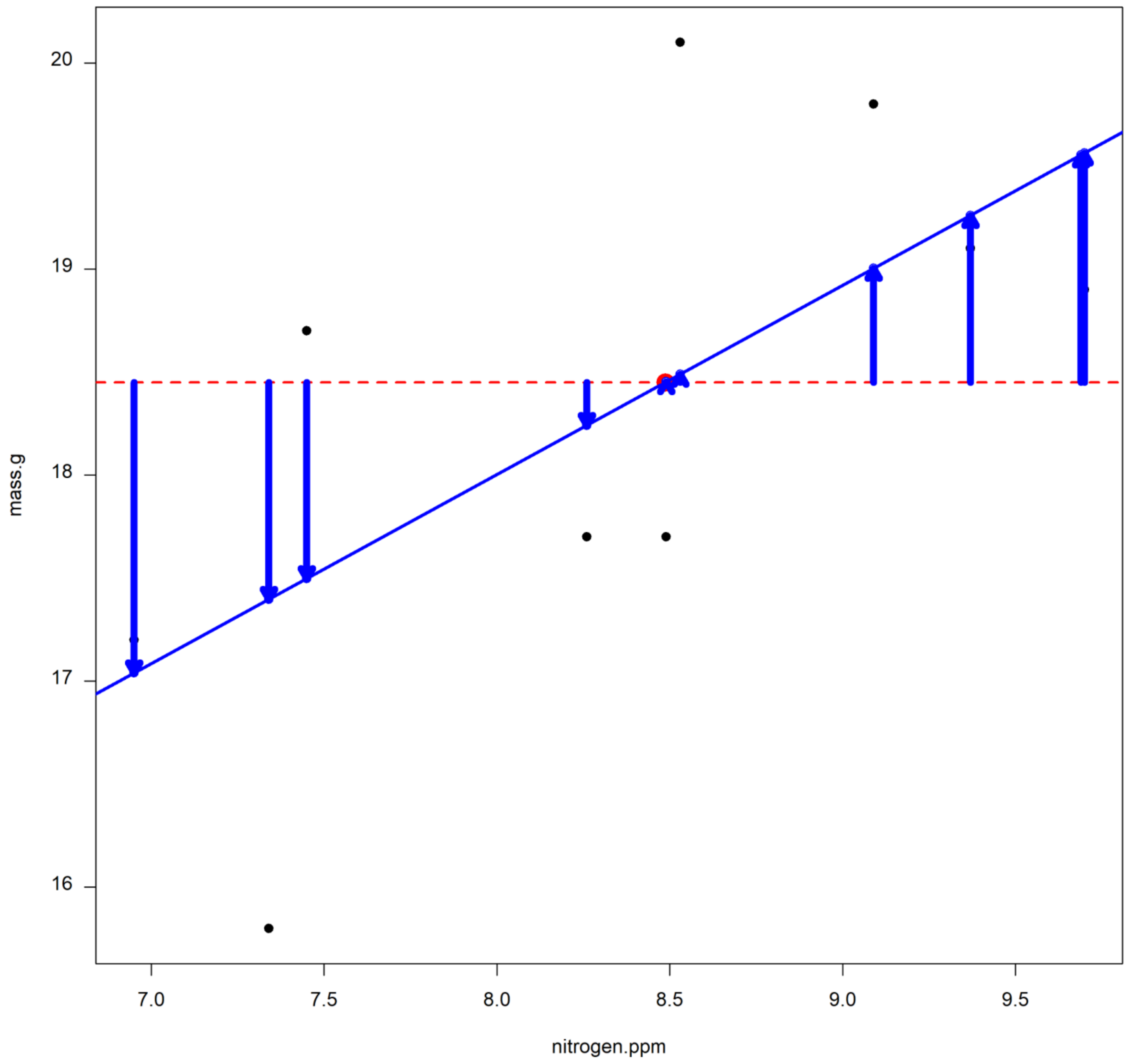
$$\sum(\hat{Y}_i - \bar{Y})^2$$



How much variation in Y does our model explain?

Residual sum of squares
SS Error

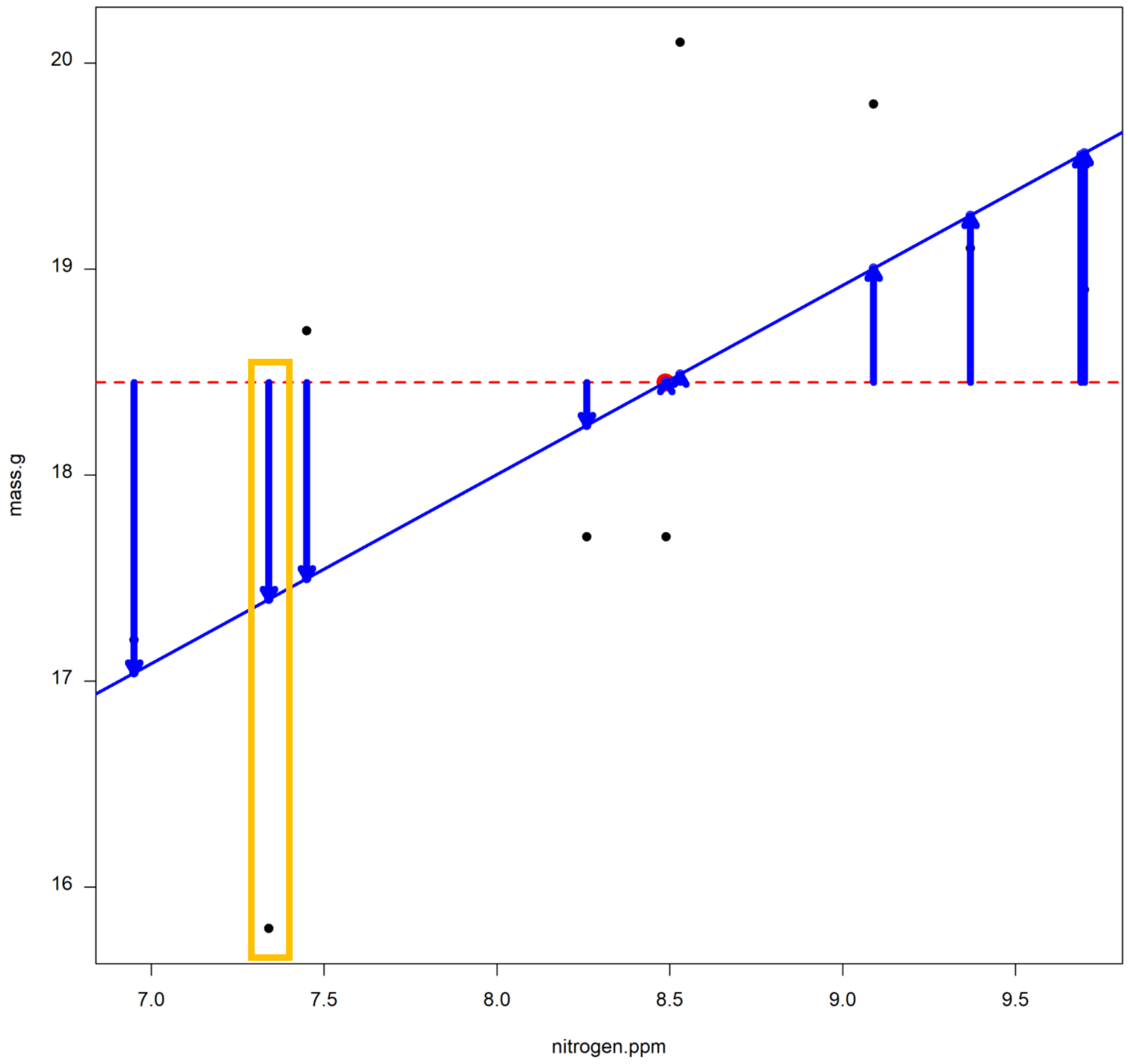
$$\sum (Y_i - \hat{Y}_i)^2$$



How much variation in Y does our model explain?

Model sum of squares
SS Regression

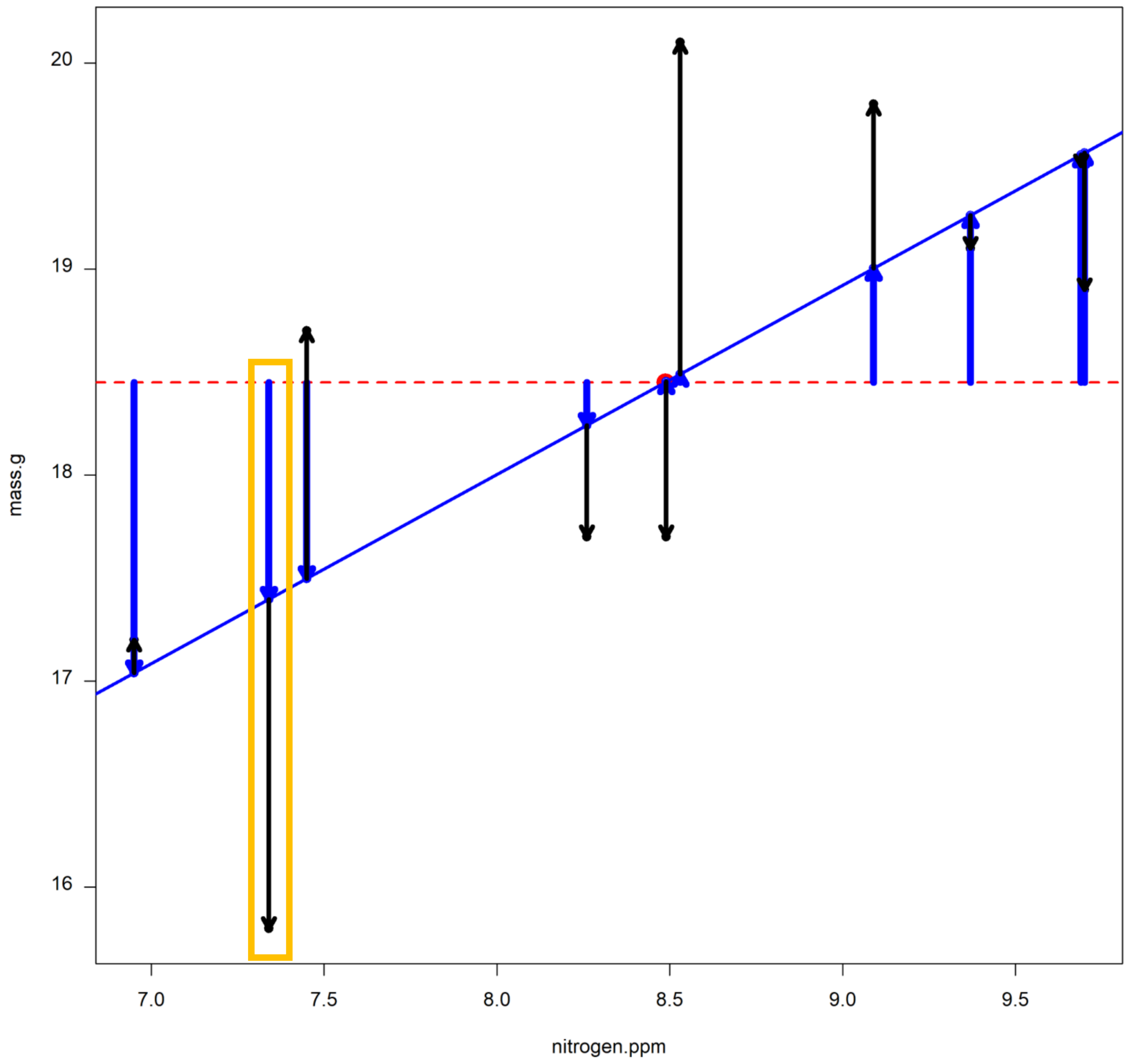
$$\sum(\hat{Y}_i - \overline{Y})^2$$



How much variation in Y does our model explain?

Model sum of squares
SS Regression

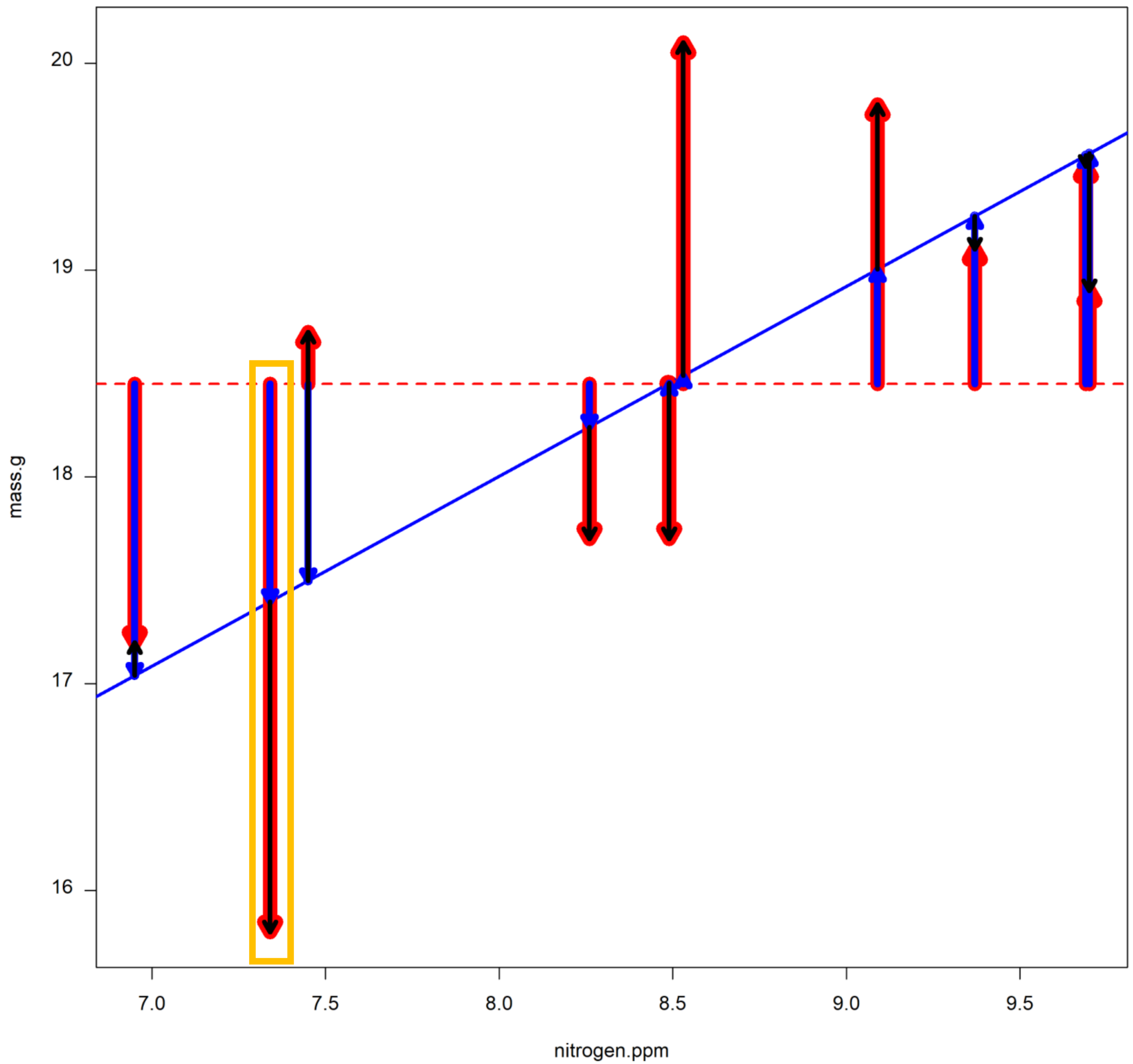
$$\sum(\hat{Y}_i - \overline{Y})^2$$



How much variation in Y does our model explain?

Model sum of squares
SS Regression $\sum(\hat{Y}_i - \overline{Y})^2$

Residual sum of squares
SS Error $\sum(Y_i - \hat{Y}_i)^2$



How much variation in Y does our model explain?

Model sum of squares
SS Regression

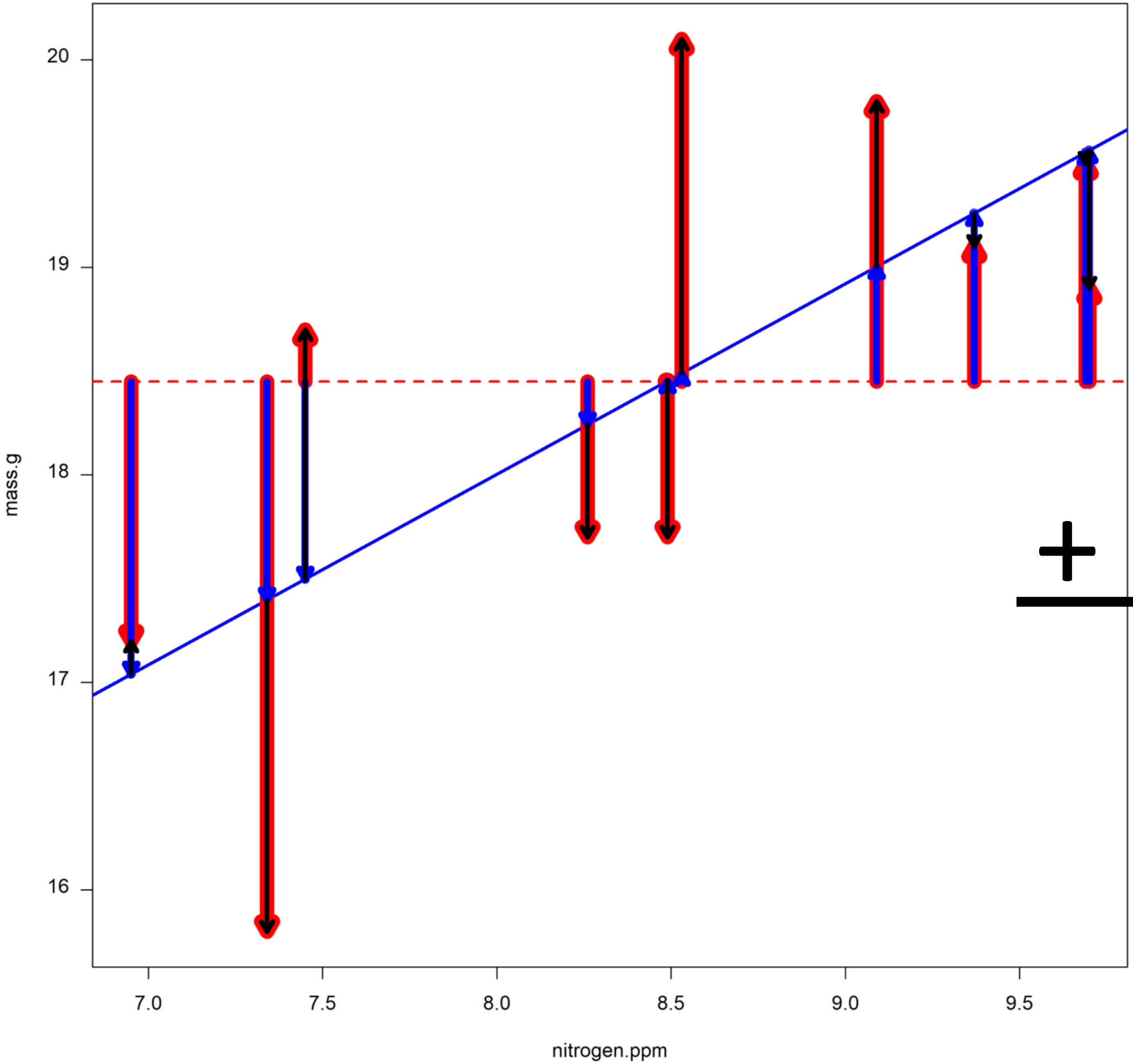
$$\sum(\hat{Y}_i - \bar{Y})^2$$

Residual sum of squares
SS Error

$$\sum(Y_i - \hat{Y}_i)^2$$

Total sum of squares
SS Total

$$\sum(Y_i - \bar{Y})^2$$



How much variation in Y does our model explain?

Model sum of squares
SS Regression

$$\sum(\hat{Y}_i - \bar{Y})^2$$

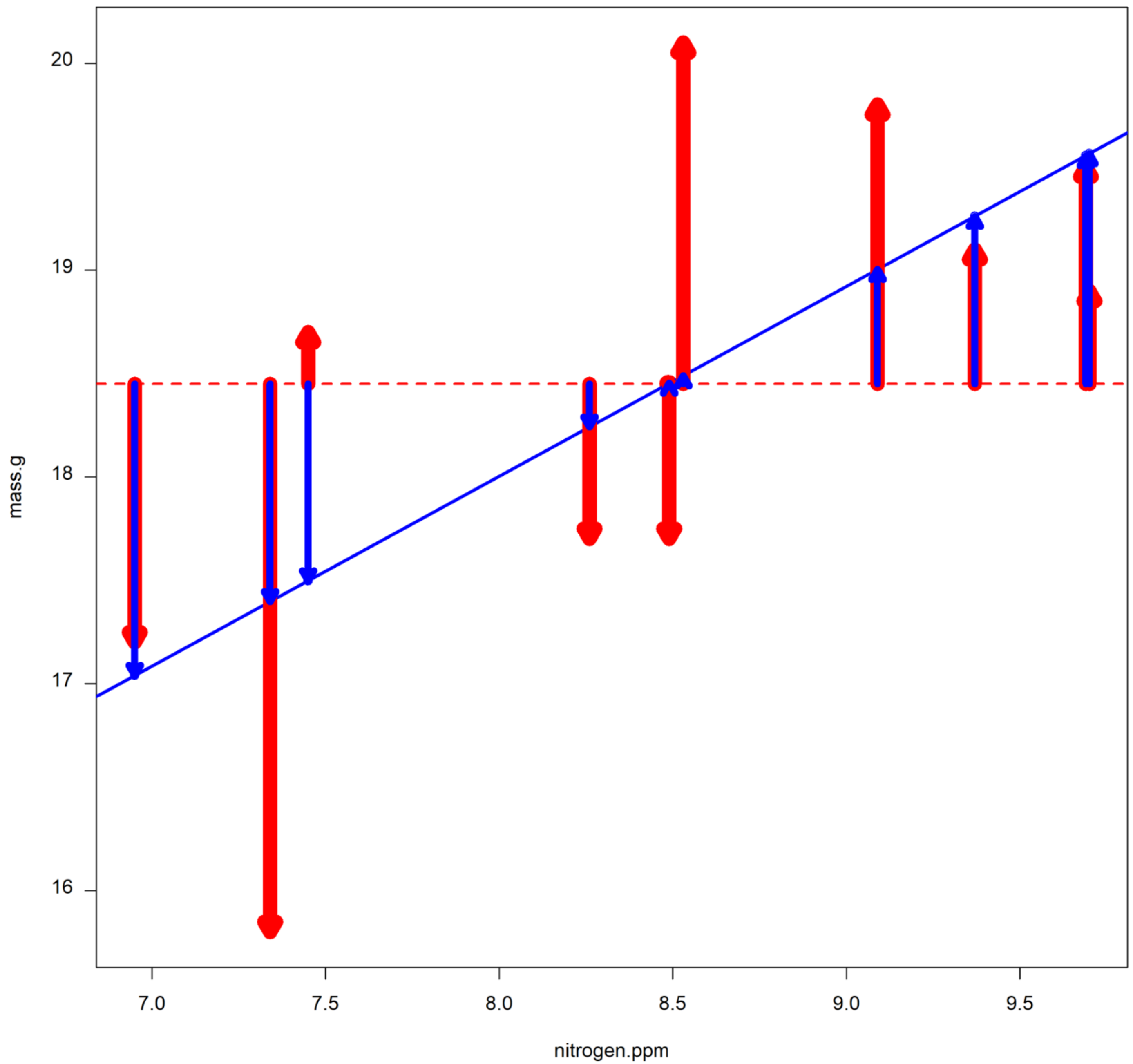
Residual sum of squares
SS Error

$$\sum(Y_i - \hat{Y}_i)^2$$

+

Total sum of squares
SS Total

$$\sum(Y_i - \bar{Y})^2$$



How much variation in Y does our model explain?

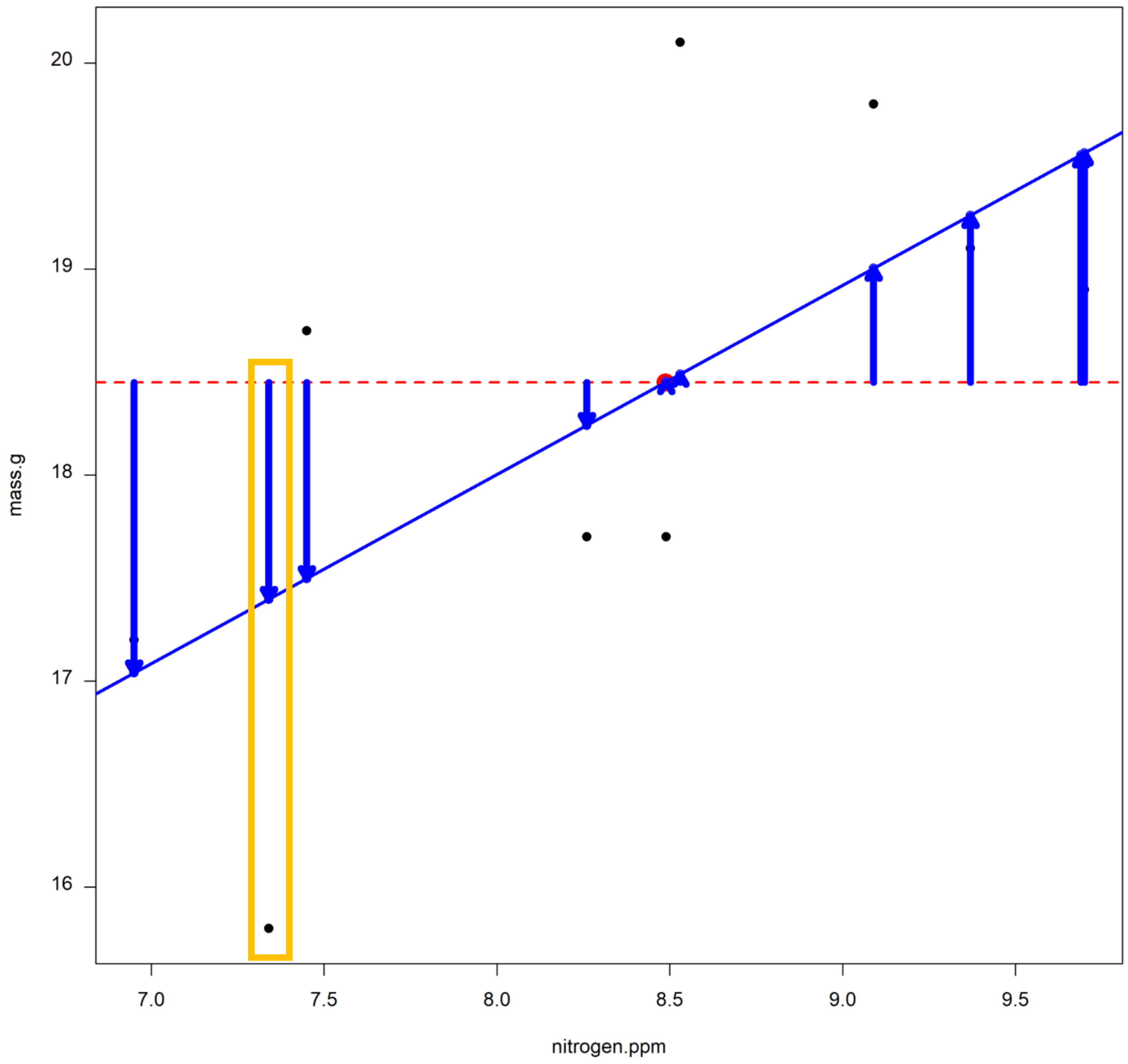
Coefficient of determination

R^2

Model sum of squares

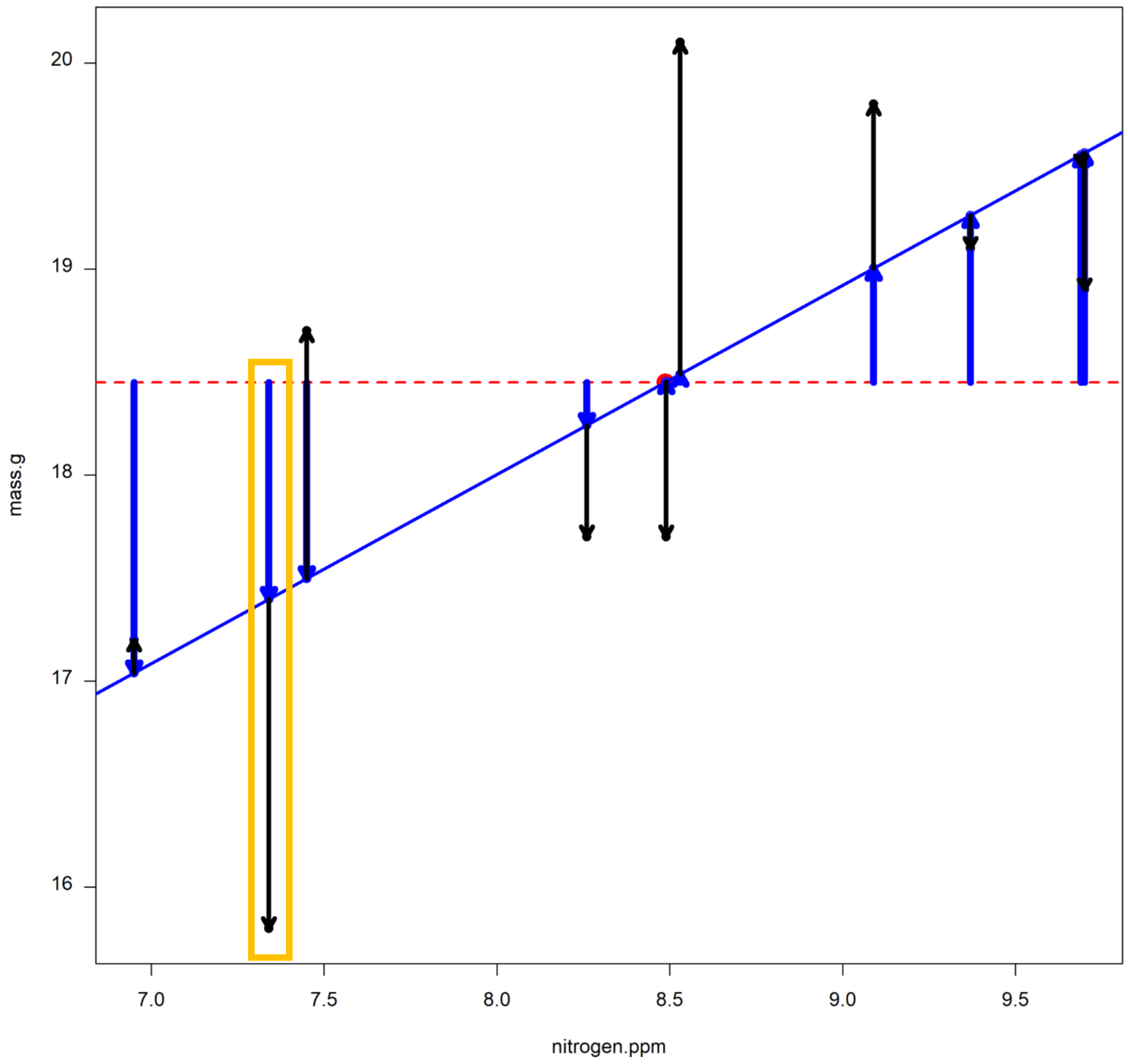
$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}$$

Total sum of squares



How much variation in Y does our model explain on average?

$$\text{Mean Square} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{df}$$



If the model is “good”, it should account for more variation on average than the average variation the error accounts for

$$F = \frac{\dfrac{\sum(\hat{Y}_i - \bar{Y})^2}{df}}{\dfrac{\sum(Y_i - \hat{Y}_i)^2}{df}}$$