

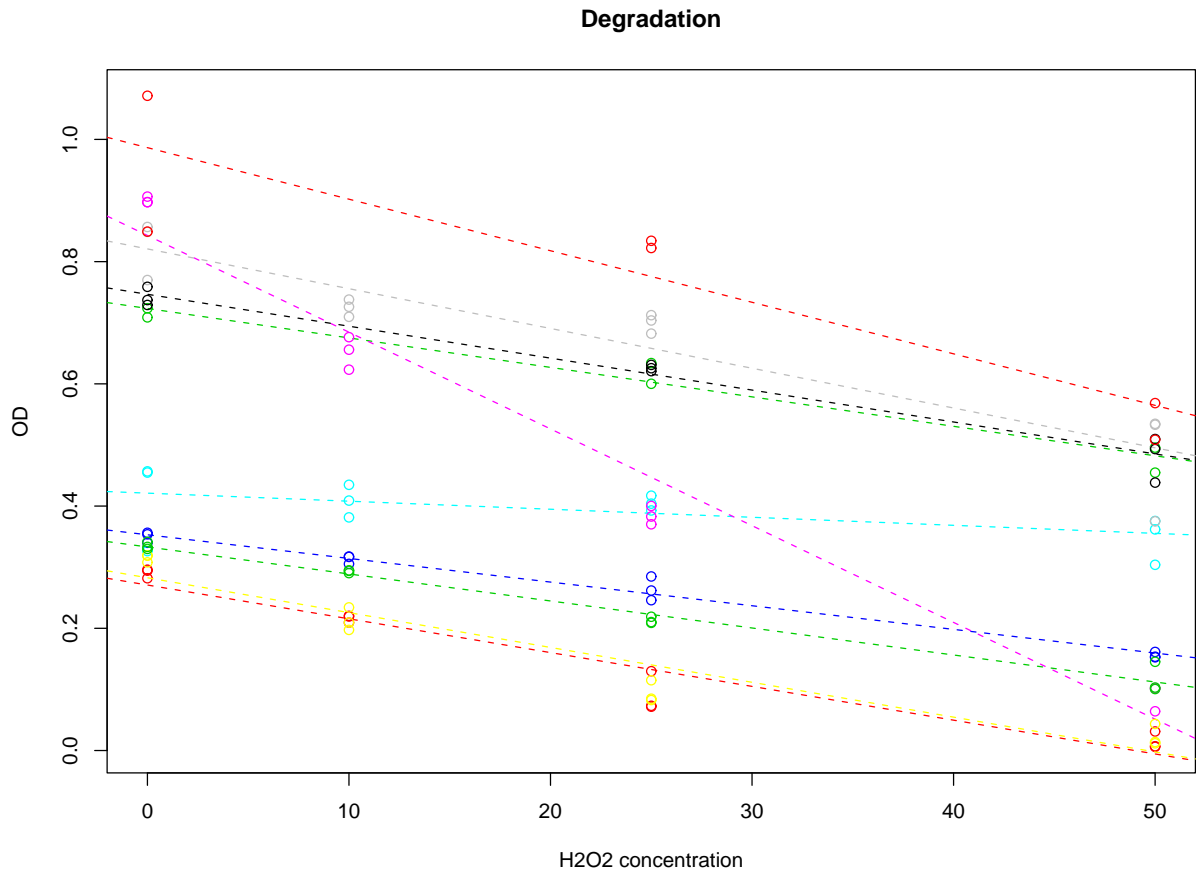
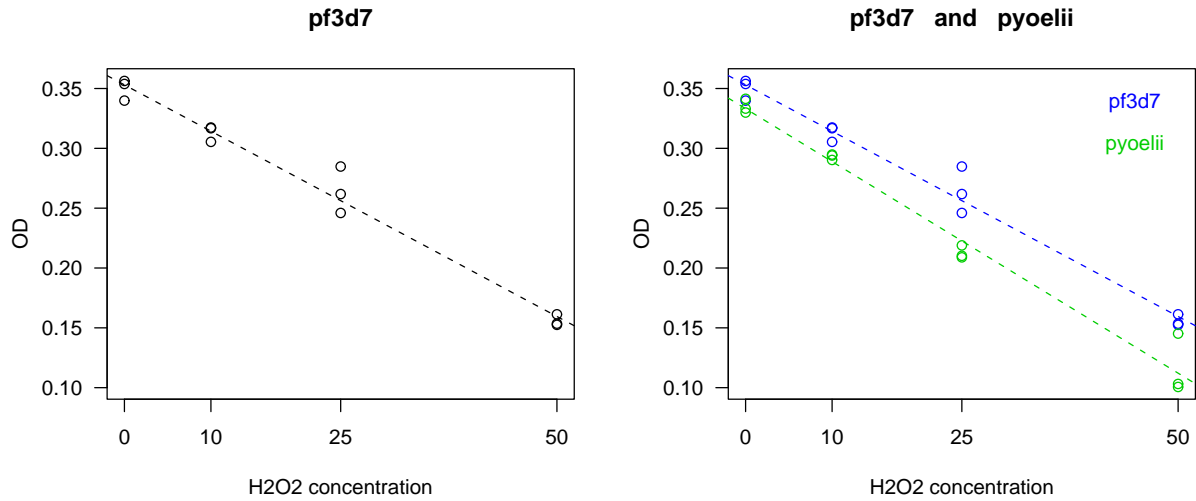
This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.



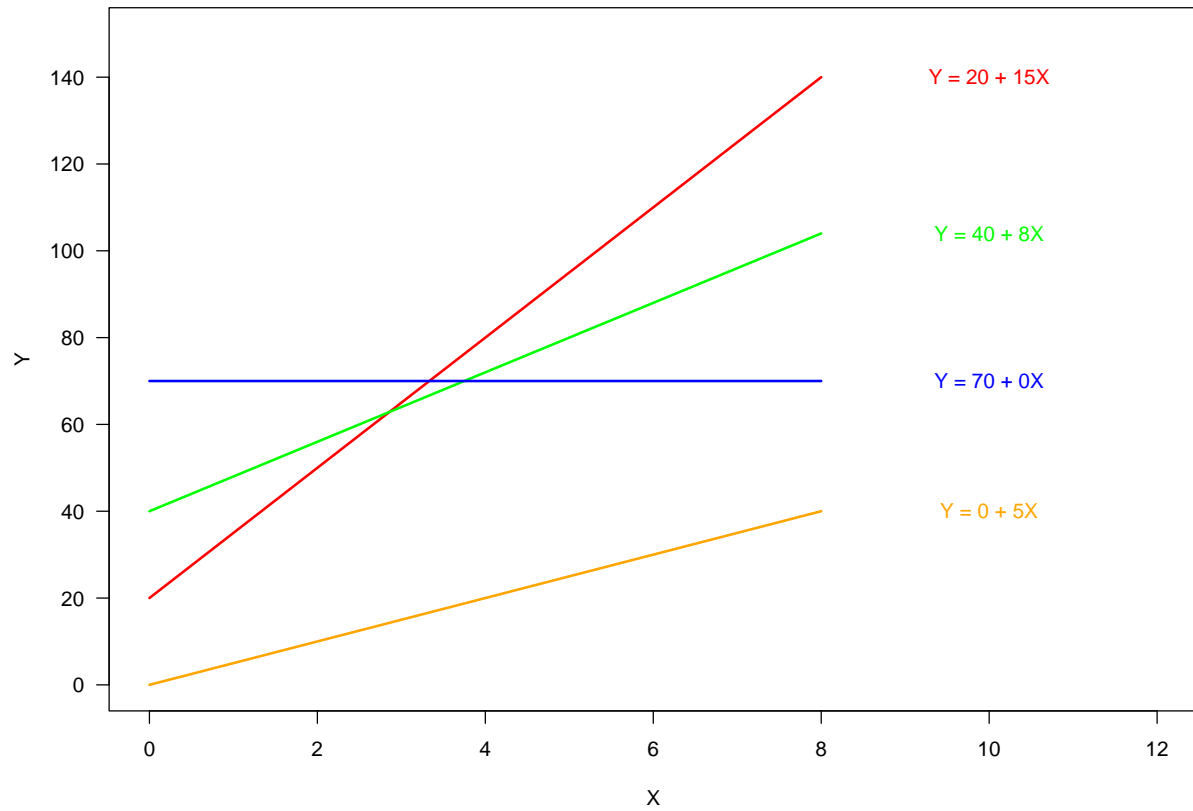
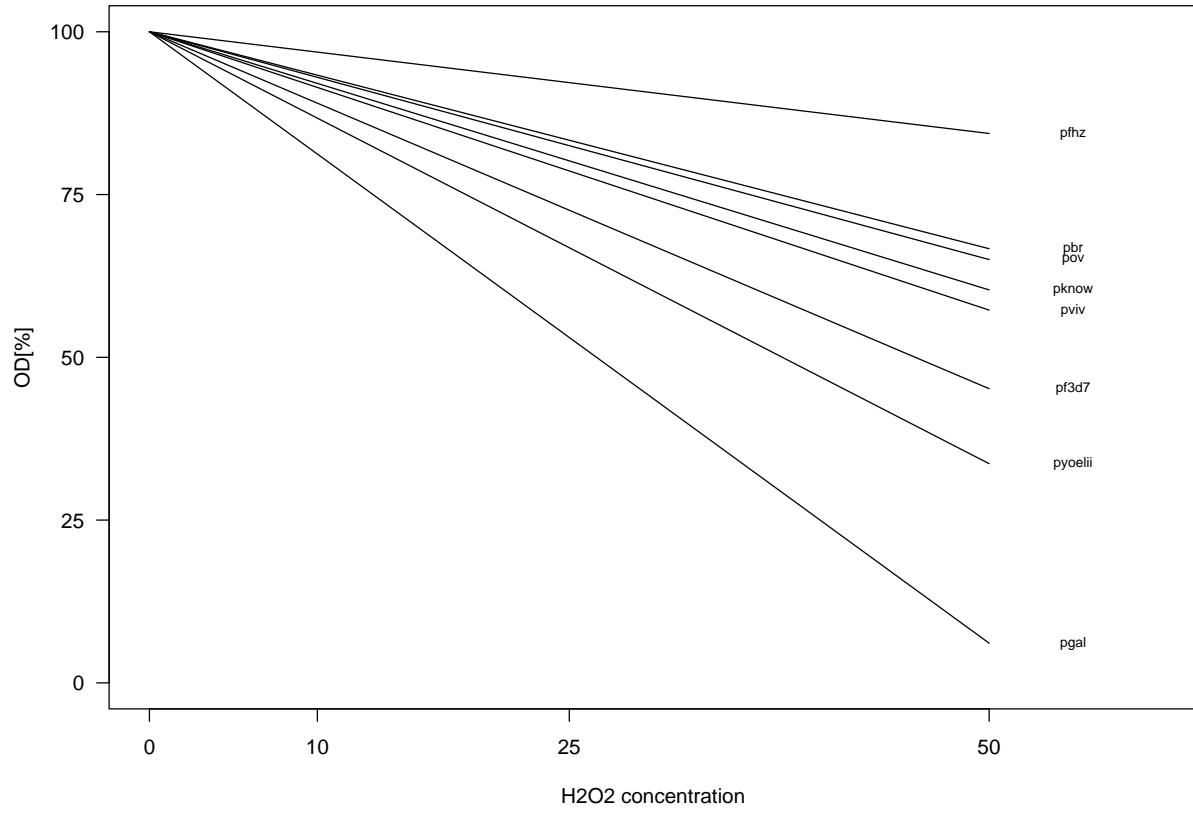
Copyright 2006, The Johns Hopkins University and Karl W. Broman. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

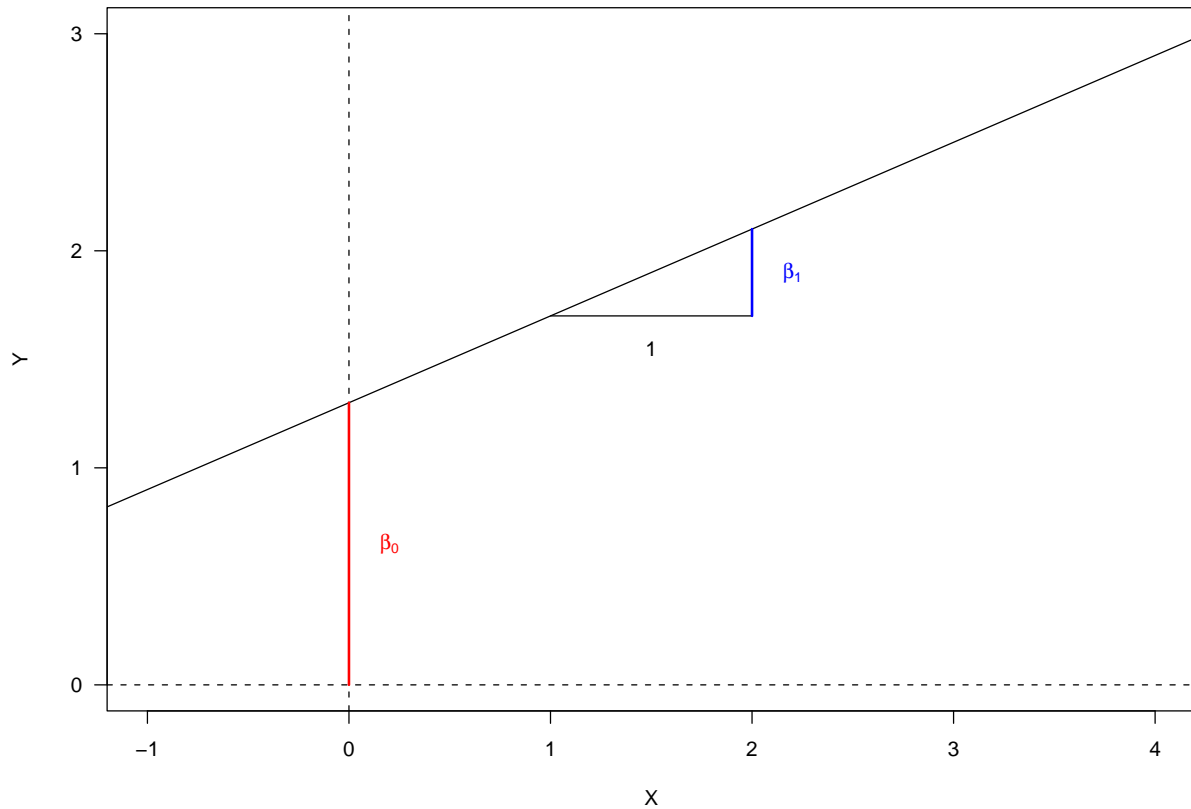
Example

Measurements of degradation of heme with different concentrations of hydrogen peroxide (H_2O_2), for different species of heme.



Degradation [%]





The regression model

Let X be the predictor and Y be the response. Assume we have n observations $(x_1, y_1), \dots, (x_n, y_n)$ from X and Y . The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \sim \text{iid } N(0, \sigma^2).$$

How do we estimate $\beta_0, \beta_1, \sigma^2$?

Fitted values and residuals

We can write

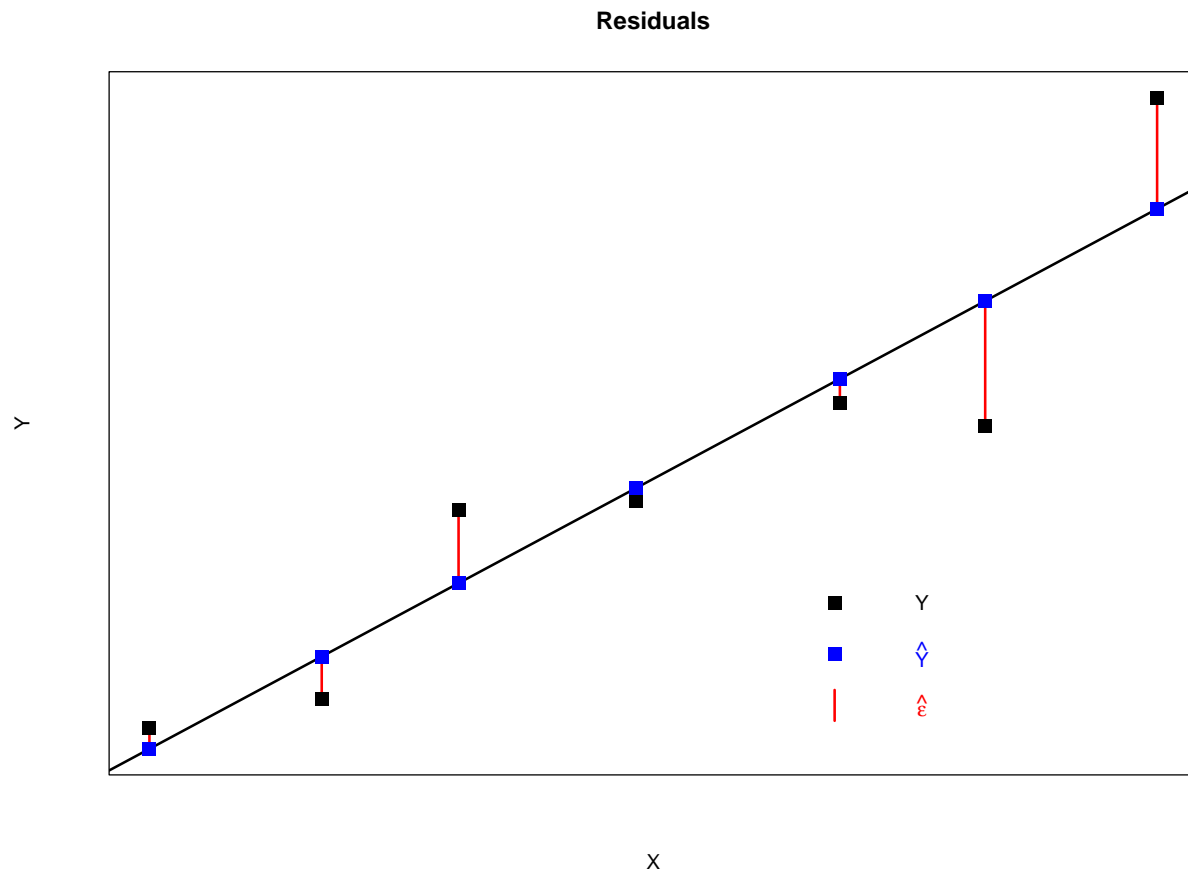
$$\epsilon_i = y_i - \beta_0 - \beta_1 x_i$$

For a pair of estimates $(\hat{\beta}_0, \hat{\beta}_1)$ for (β_0, β_1) we define the fitted values as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

The residuals are

$$\hat{\epsilon}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$



Residual sum of squares

For every pair of values for β_0 and β_1 we get a different value for the residual sum of squares.

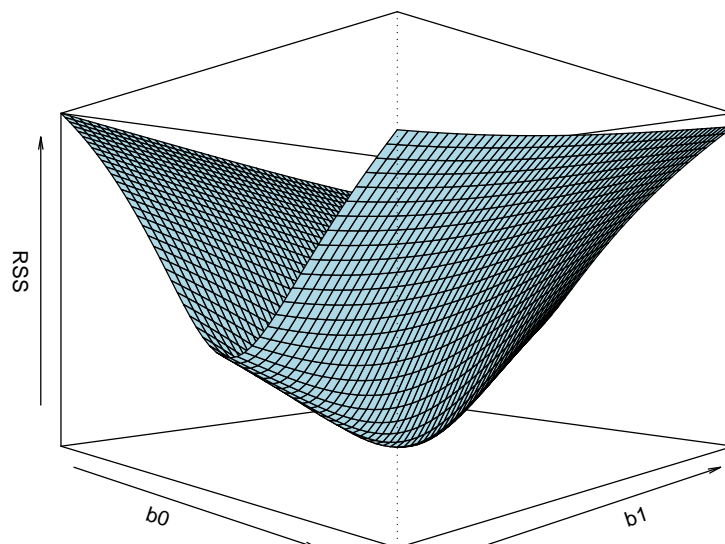
$$\text{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

We can look at RSS as a function of β_0 and β_1 . We try to minimize this function, i. e. we try to find

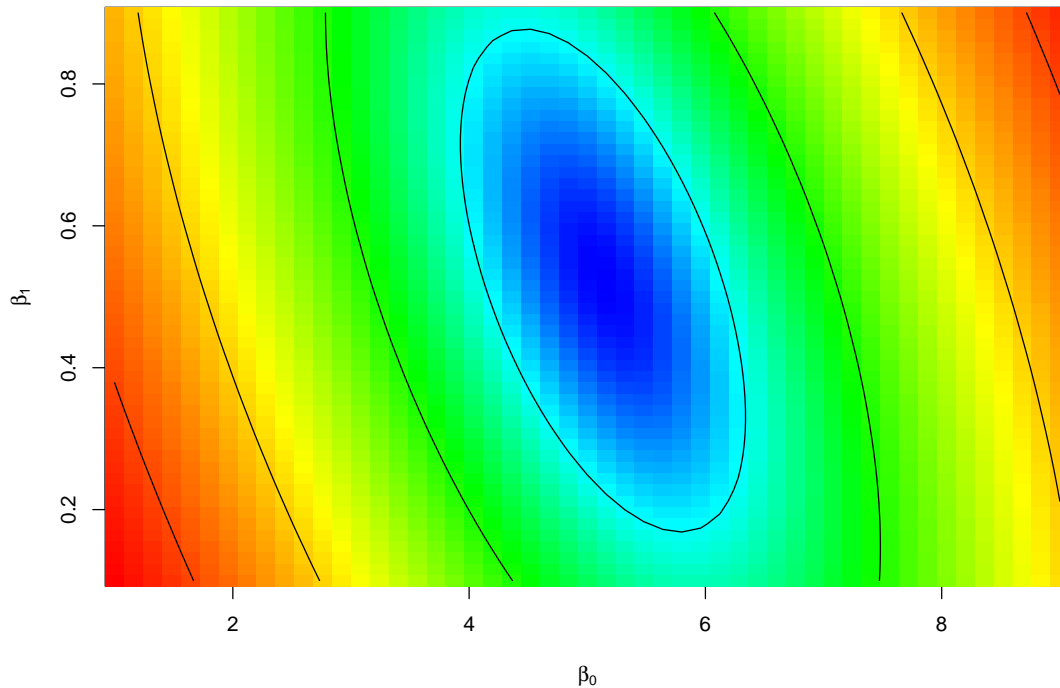
$$(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} \text{RSS}(\beta_0, \beta_1)$$

Hardly surprising, this method is called least squares estimation.

Residual sum of squares



Residual sum of squares



Notation

Assume we have n observations: $(x_1, y_1), \dots, (x_n, y_n)$.

$$\bar{x} = \frac{\sum_i x_i}{n}$$

$$\bar{y} = \frac{\sum_i y_i}{n}$$

$$SXX = \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n(\bar{x})^2$$

$$SYY = \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n(\bar{y})^2$$

$$SXY = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}$$

$$RSS = \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{\epsilon}_i^2$$

Parameter estimates

The function

$$\text{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

is minimized by

$$\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Useful to know

Using the parameter estimates, our best guess for any y given x is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Hence

$$\hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 \bar{x} = \bar{y}$$

That means every regression line goes through the point (\bar{x}, \bar{y}) .

Variance estimates

As variance estimate we use

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-2}$$

This quantity is called the **residual mean square**. It has the property

$$(n-2) \times \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$$

In particular, this implies

$$E(\hat{\sigma}^2) = \sigma^2$$

Example

H ₂ O ₂ concentration			
0	10	25	50
0.3399	0.3168	0.2460	0.1535
0.3563	0.3054	0.2618	0.1613
0.3538	0.3174	0.2848	0.1525

We get

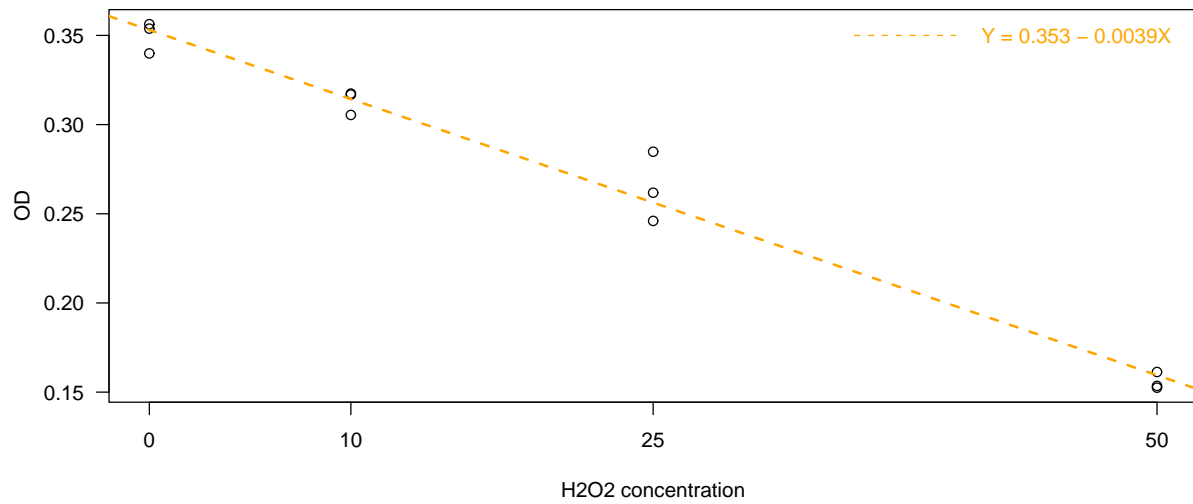
$$\bar{x} = 21.25, \quad \bar{y} = 0.27, \quad \text{SXX} = 4256.25, \quad \text{SXY} = -16.48, \quad \text{RSS} = 0.0013.$$

Therefore

$$\hat{\beta}_1 = \frac{-16.48}{4256.25} = -0.0039, \quad \hat{\beta}_0 = 0.27 - (-0.0039) \times 21.25 = 0.353,$$

$$\hat{\sigma} = \sqrt{\frac{0.0013}{12-2}} = 0.0115.$$

pf3d7

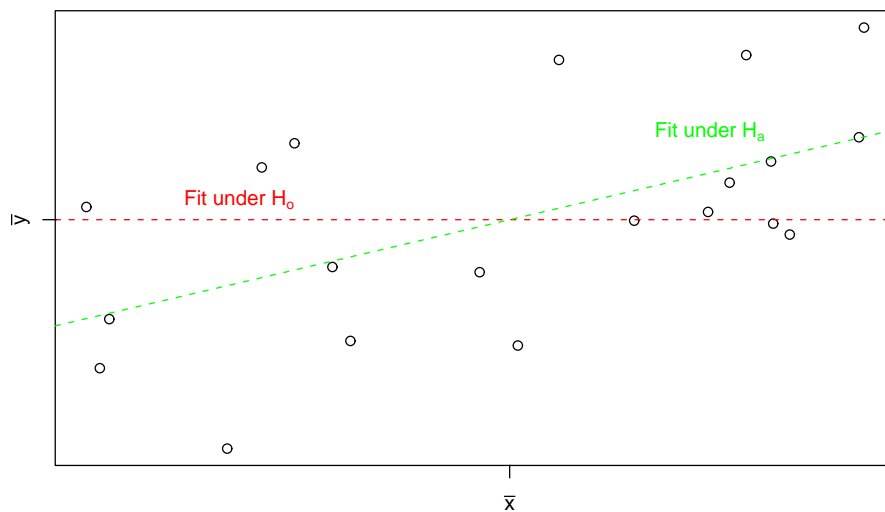


The R function `lm()` does all these calculations for you. And more!

Comparing models

We want to test whether $\beta_1 = 0$:

$$H_0 : y_i = \beta_0 + \epsilon_i \quad \text{versus} \quad H_a : y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$



Sum of squares

Under H_a :

$$RSS = \sum_i (y_i - \hat{y}_i)^2 = SYY - \frac{(SXY)^2}{SXX} = SYY - \hat{\beta}_1^2 \times SXX$$

Under H_0 :

$$\sum_i (y_i - \hat{\beta}_0)^2 = \sum_i (y_i - \bar{y})^2 = SYY$$

Hence

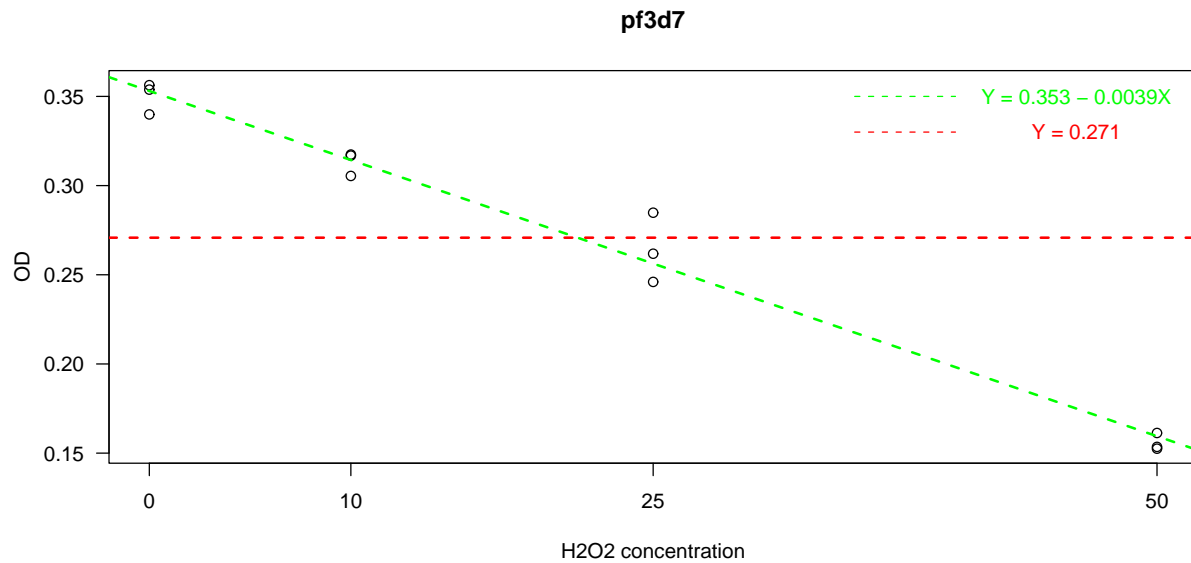
$$SS_{\text{reg}} = SYY - RSS = \frac{(SXY)^2}{SXX}$$

ANOVA

Source	df	SS	MS	F
regression on X	1	SS_{reg}	$MS_{\text{reg}} = \frac{SS_{\text{reg}}}{1}$	$\frac{MS_{\text{reg}}}{MSE}$
residuals for full model	$n - 2$	RSS	$MSE = \frac{RSS}{n - 2}$	
total	$n - 1$	SYY		

David Sullivan's pf3d7 data

Source	df	SS	MS	F
regression on X	1	0.06378	0.06378	484.1
residuals for full model	10	0.00131	0.00013	
total	11	0.06509		



Remember: The R function `lm()` does the calculations for you!