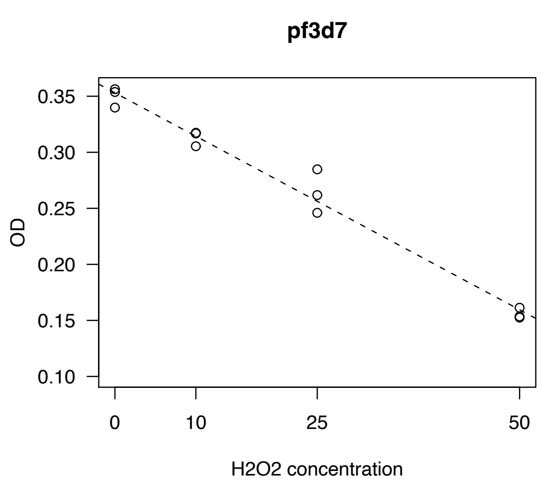
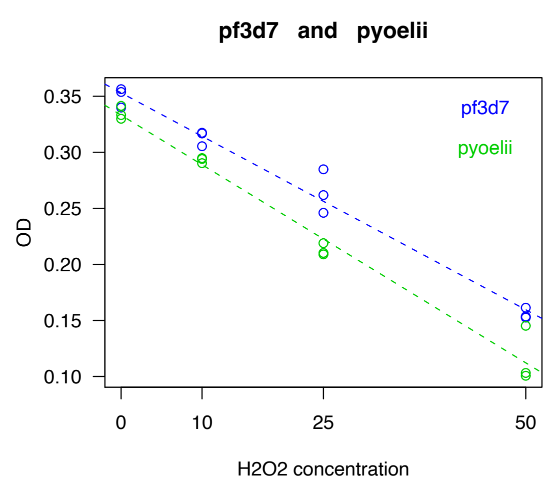
Linear Regression

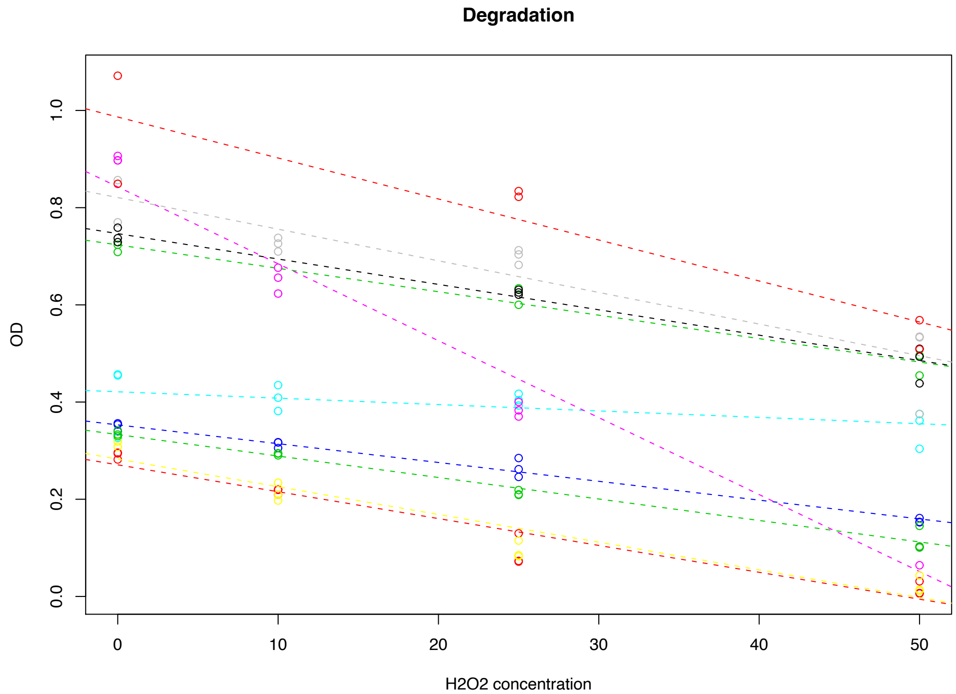
Example

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

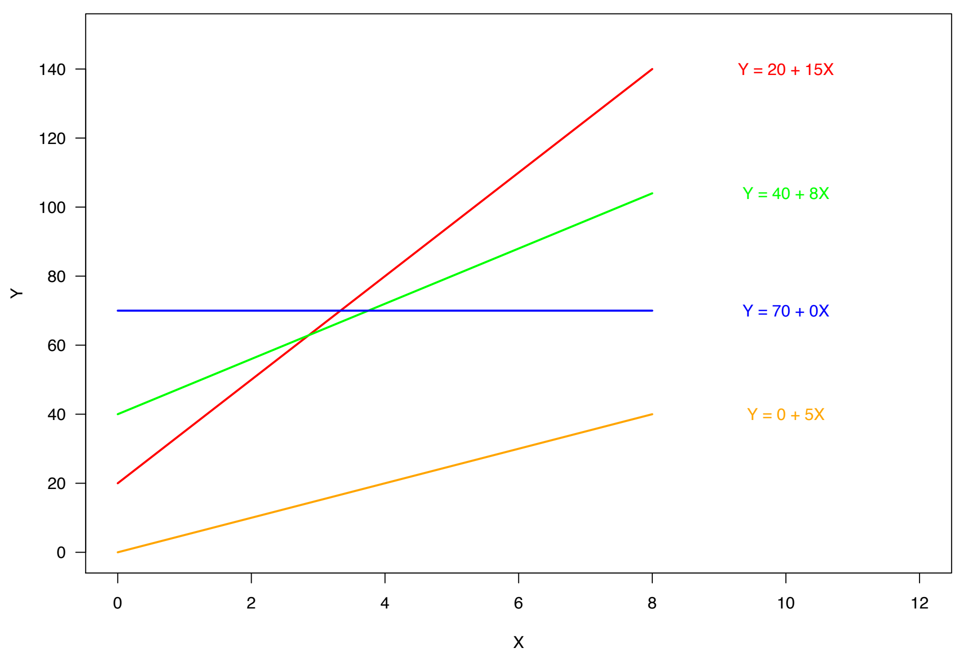




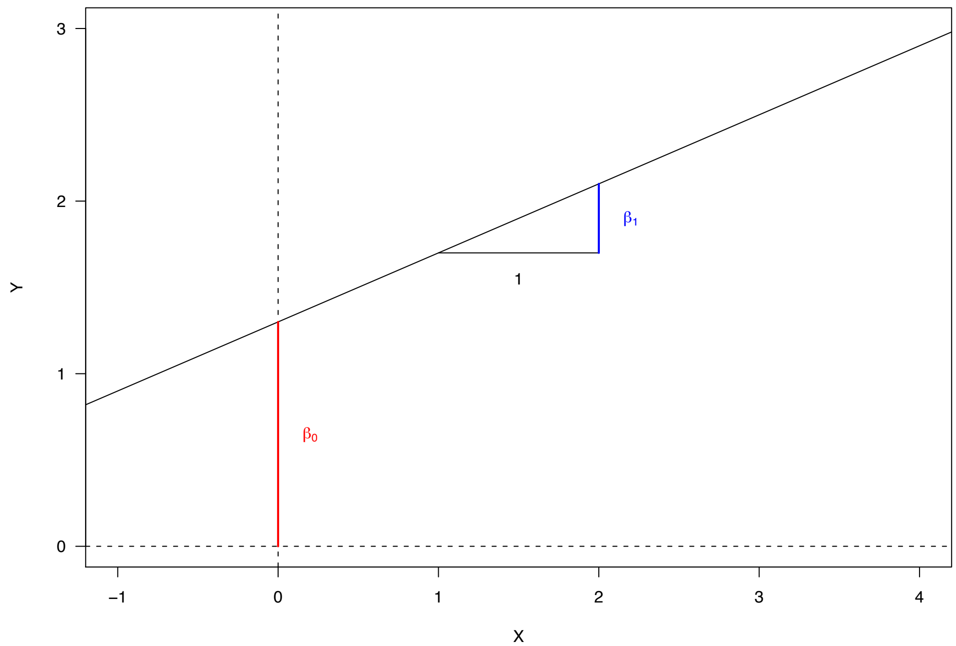
Example



Linear regression



Linear regression



The regression model

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

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

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

Fitted values and residuals

We can write

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

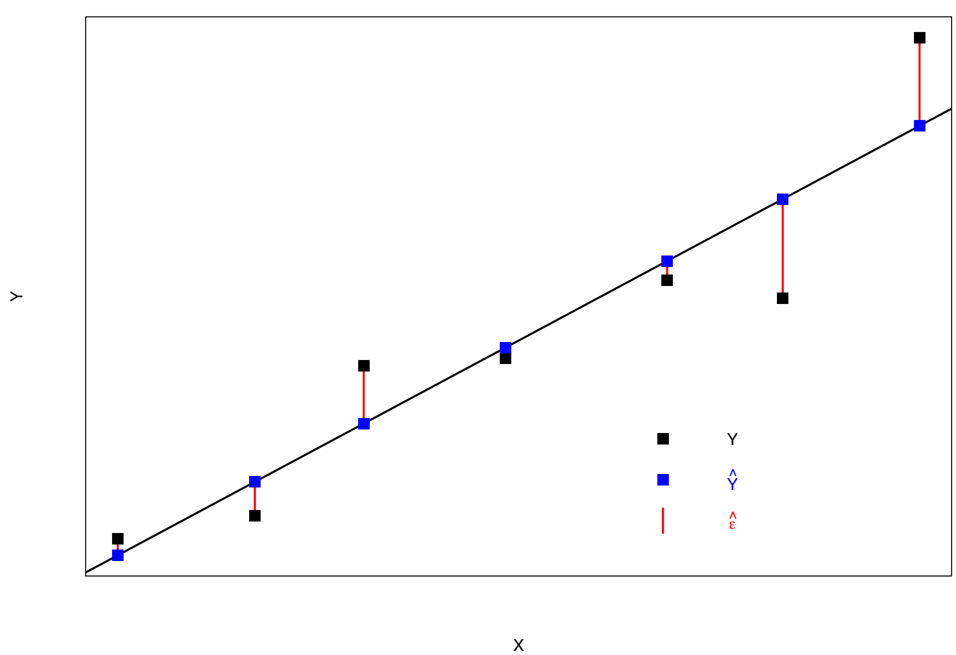
For a pair of estimates $(\hat{\beta}_0, \hat{\beta}_1)$ for the pair of parameters (β_0, β_1) we define the fitted values as

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

The residuals are

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

Residuals



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.

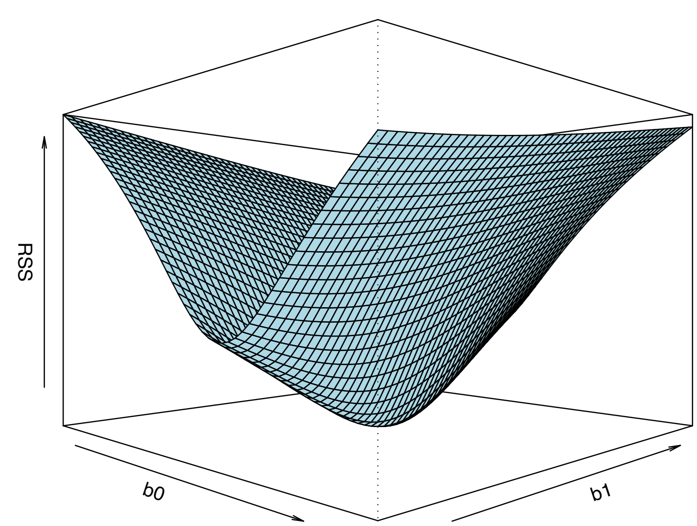
$$\mathrm{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

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

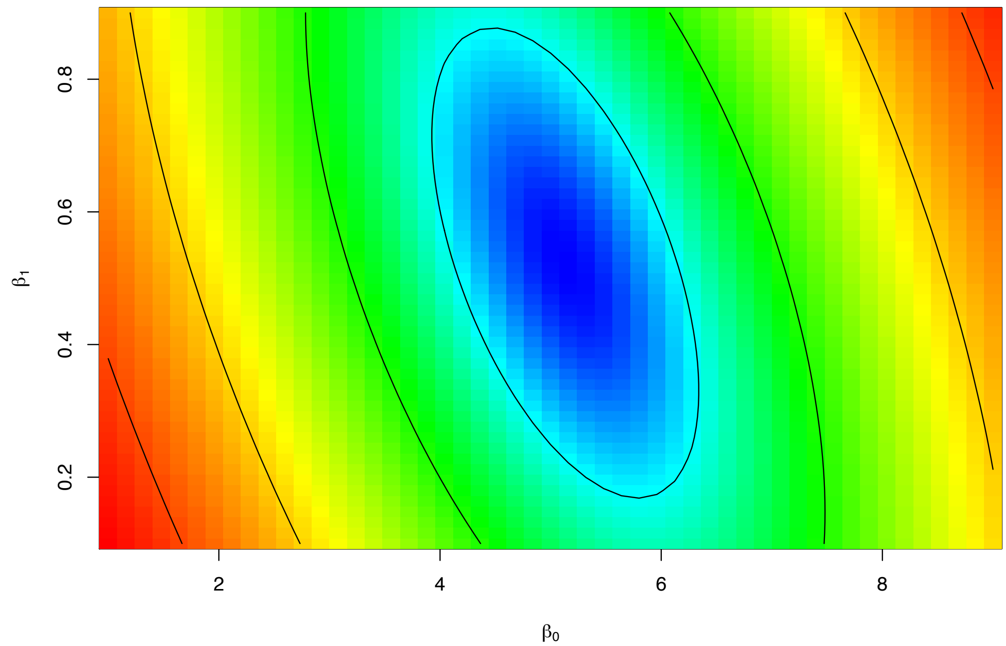
$$(\hat{\beta}_0, \hat{\beta}_1) = \min_{\beta_0, \beta_1} \mathrm{RSS}(\beta_0, \beta_1)$$

Hardly surprising, this method is called least squares estimation.

Residual sum of squares



Residual sum of squares



Notation

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

$$
\begin{aligned}
\bar{x} &= \frac{\sum_i x_i}{n} \\
\bar{y} &= \frac{\sum_i y_i}{n} \\
SXX &= \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n(\bar{x})^2 \\
SYY &= \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n(\bar{y})^2 \\
SXY &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y} \\
RSS &= \sum_i (y_i - \hat{y}_i)^2 = \sum_i \hat{\epsilon}_i^2
\end{aligned}
$$

Parameter estimates

The function

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

is minimized by

$$\hat{\beta}_1 = \frac{\mathrm{SXY}}{\mathrm{SXX}}$$

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

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 following property:

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

In particular, this implies

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

Example

H_2O_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

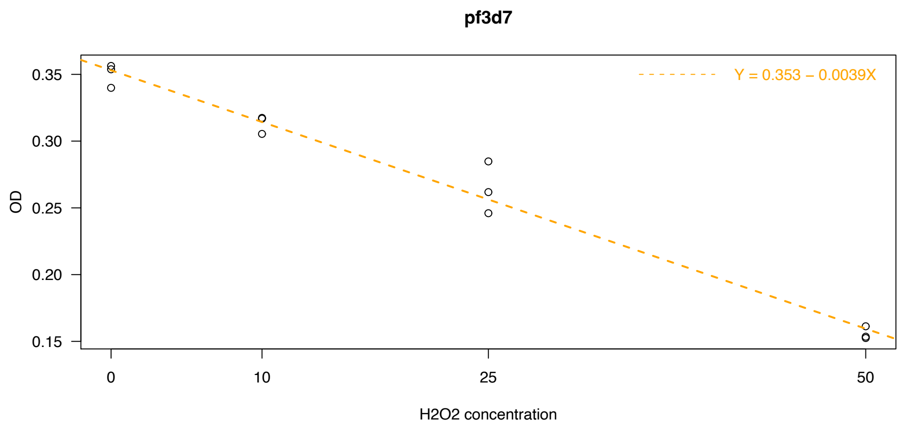
$\bar{x} = 21.25$, $\bar{y} = 0.27$, $\text{SXX} = 4256.25$, $\text{SXY} = -16.48$, $\text{RSS} = 0.0013$.

Therefore

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

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

Example



Interpretation

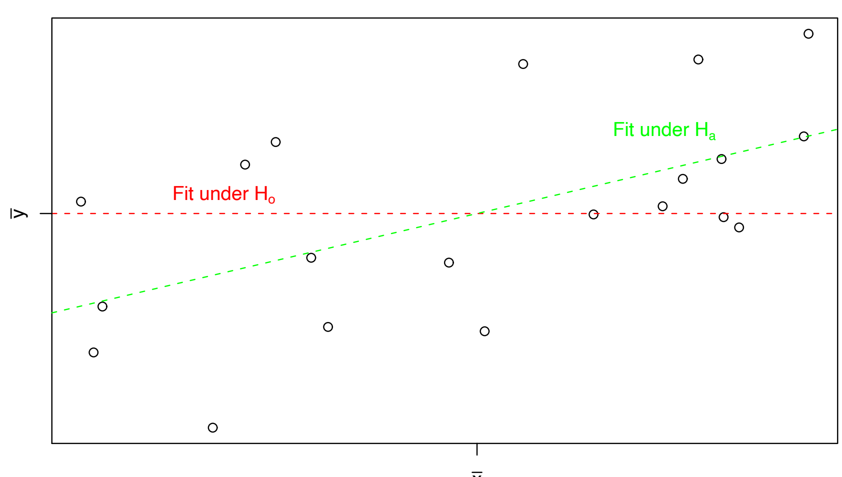
At zero concentration we expect the optical density to be 0.353.

Comparing two experiments that differ by one unit concentration, we expect the optical density to be 0.0039 lower in the experiment with the larger concentration.

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

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

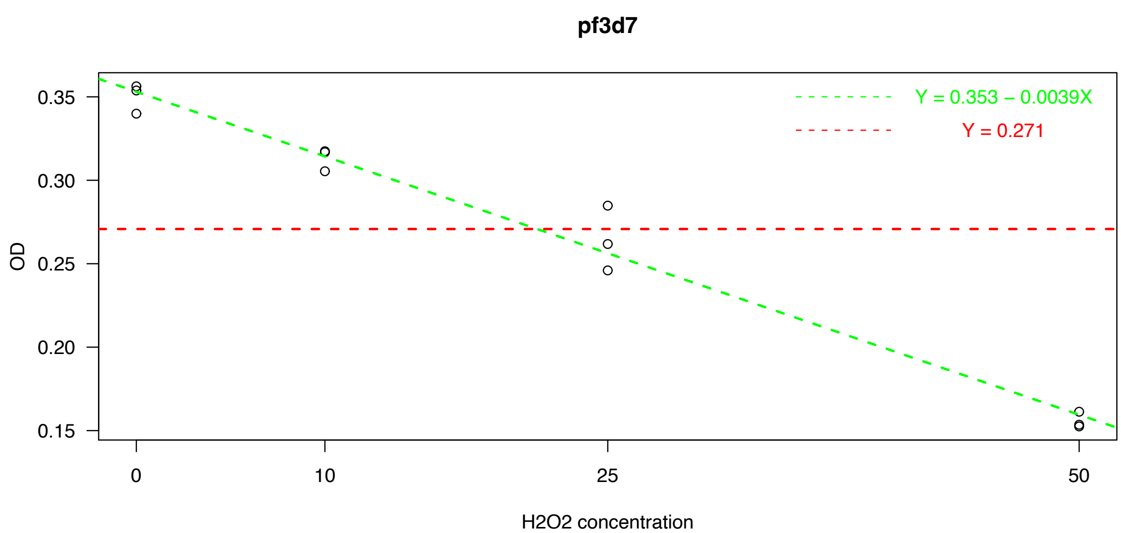
ANOVA

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

David Sullivan’s pf3d7 data

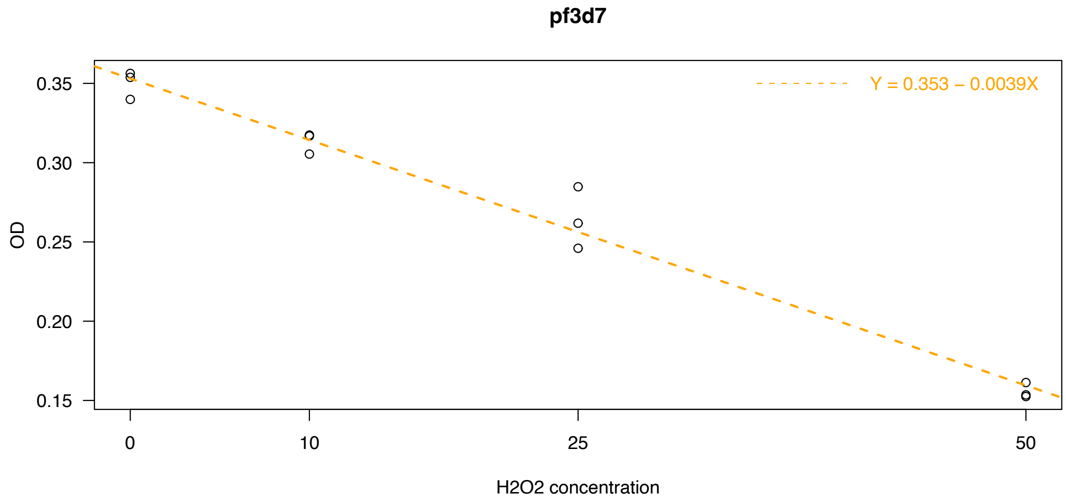
Source	**df**	**SS**	**MS**	**F**
regression on X	1	0.06378	0.06378	484.1
residuals for full model	10	0.00131	0.00013	
total	11	0.06509		

David Sullivan’s pf3d7 data



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

Parameter estimates



Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ where $\epsilon_i \sim$ iid Normal$(0, \sigma^2)$

Estimates: $\hat{\beta}_1 = \sum_i (x_i - \bar{x})(y_i - \bar{y}) / \sum_i (x_i - \bar{x})^2$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ $\quad \hat{\sigma} = \sqrt{\sum_i (y_i - \hat{y}_i)^2 / (n-2)}$

Parameter estimates

We already know that

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

and in particular

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

$\longrightarrow$ What about $\hat{\beta}_0$ and $\hat{\beta}_1$?

Parameter estimates

One can show that

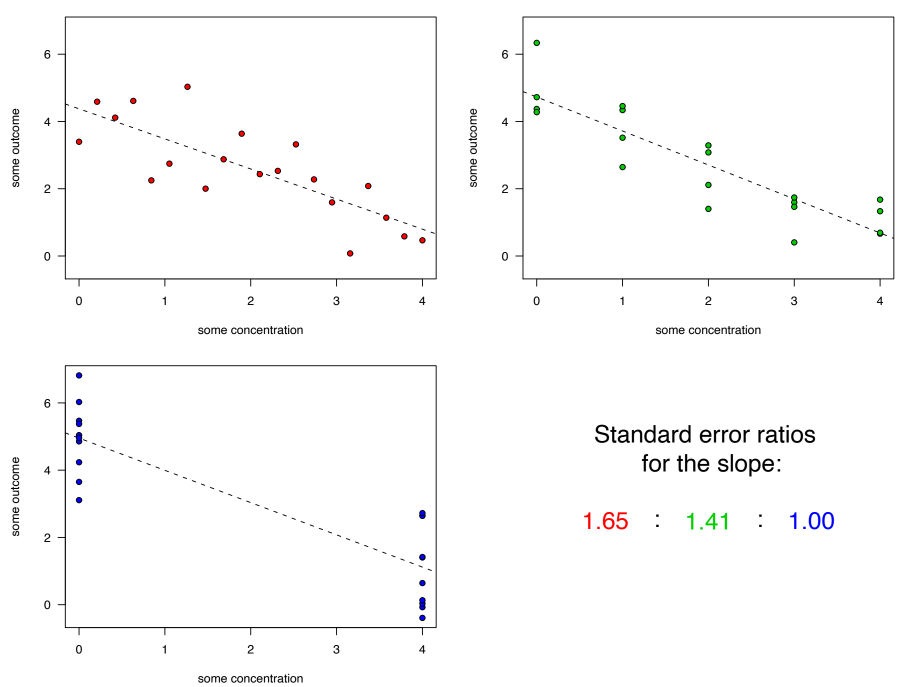
$$E(\hat\beta_0) = \beta_0 \qquad E(\hat\beta_1) = \beta_1$$

$$\text{Var}(\hat\beta_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\text{SXX}}\right) \qquad \text{Var}(\hat\beta_1) = \frac{\sigma^2}{\text{SXX}}$$

$$\text{Cov}(\hat\beta_0, \hat\beta_1) = -\sigma^2 \frac{\bar{x}}{\text{SXX}} \qquad \text{Cor}(\hat\beta_0, \hat\beta_1) = \frac{-\bar{x}}{\sqrt{\bar{x}^2 + \text{SXX}/n}}$$

$\longrightarrow$ Note: We're thinking of the x's as fixed.

Experimental design



Standard error ratios for the slope:

1.65 : 1.41 : 1.00

Parameter estimates

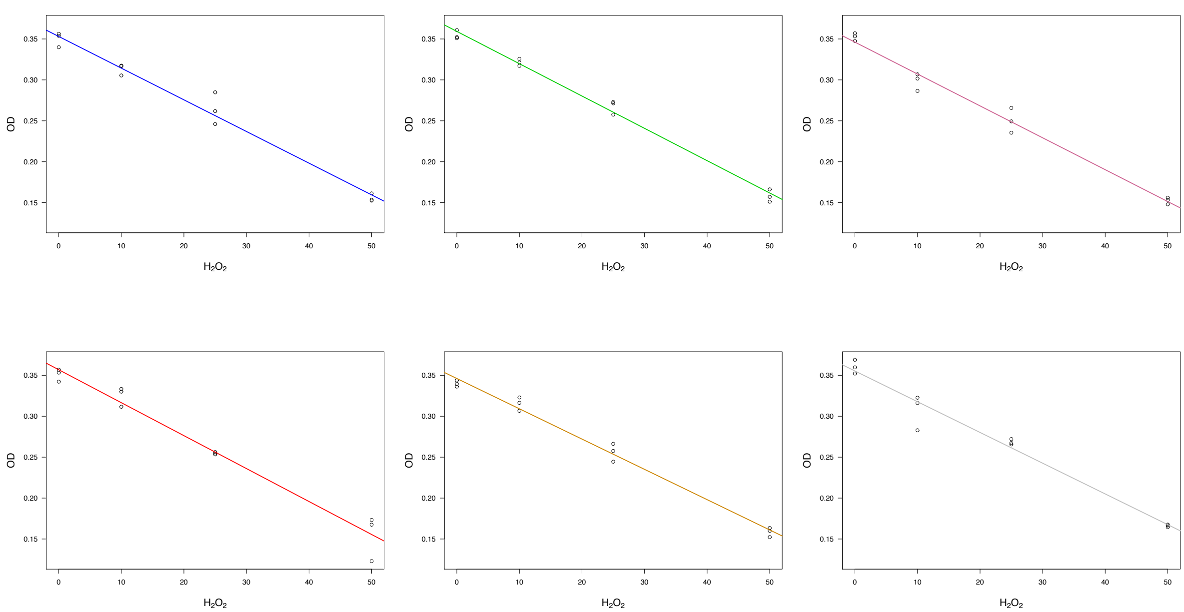
One can even show that the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is a bivariate normal distribution!

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim \mathrm{N}(\beta, \Sigma)$$

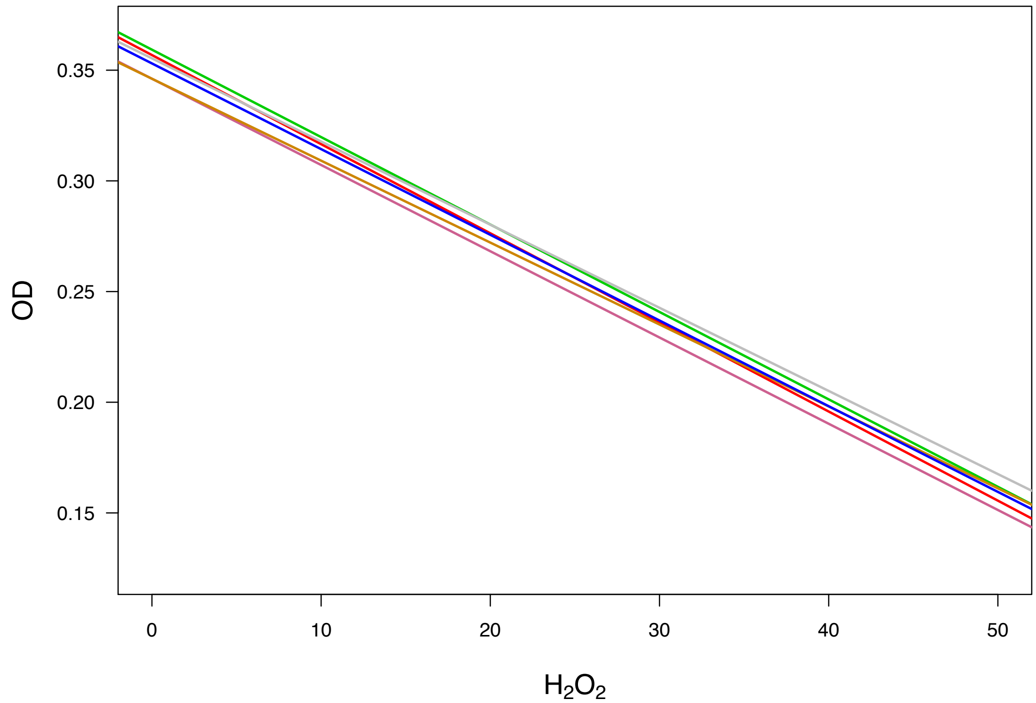
where

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{SXX} & \frac{-\bar{x}}{SXX} \\ \frac{-\bar{x}}{SXX} & \frac{1}{SXX} \end{pmatrix}$$

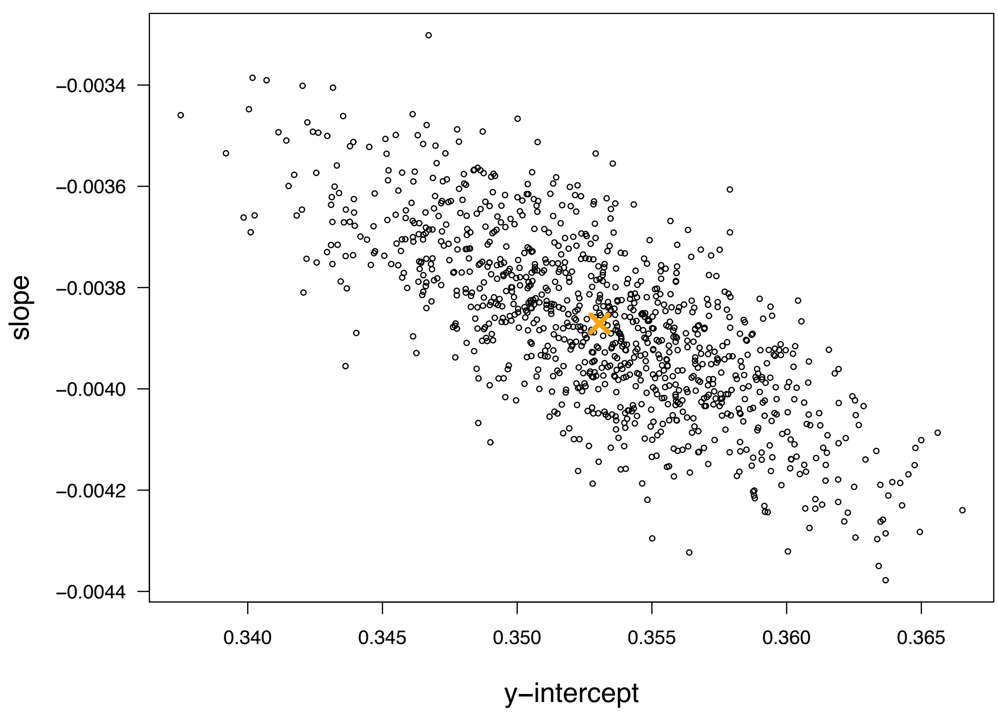
Possible outcomes



Possible outcomes



Simulation: coefficients



Confidence intervals

We know that

$$\hat{\beta}_0 \sim \mathrm{N}\left(\beta_0,\ \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{SXX}\right)\right)$$

$$\hat{\beta}_1 \sim \mathrm{N}\left(\beta_1,\ \frac{\sigma^2}{SXX}\right)$$

$\longrightarrow$ We can use those distributions for hypothesis testing and to construct confidence intervals!

Statistical inference

We want to test: $H_0 : \beta_1 = \beta_1^\star$ versus $H_a : \beta_1 \neq \beta_1^\star$ (generally, $\beta_1^\star$ is 0.)

We use

$$t = \frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)} \sim t_{n-2} \quad \text{where} \quad se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{SXX}}$$

Also,

$$\left[\hat{\beta}_1 - t_{(1-\frac{\alpha}{2}),n-2} \times se(\hat{\beta}_1)\ ,\ \hat{\beta}_1 + t_{(1-\frac{\alpha}{2}),n-2} \times se(\hat{\beta}_1)\right]$$

is a $(1-\alpha)\times 100\%$ confidence interval for β_1.

Results

The calculations in the test $H_0 : \beta_0 = \beta_0^*$ versus $H_a : \beta_0 \neq \beta_0^*$ are analogous, except that we have to use

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \times \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX}\right)}$$

For the pf3d7 data we get the 95% confidence intervals

(0.342 , 0.364)	for the intercept
(– 0.0043 , – 0.0035)	for the slope

Testing whether the intercept (slope) is equal to zero, we obtain 70.7 (– 22.0) as test statistic. This corresponds to a p-value of 7.8×10^{-15} (8.4×10^{-10}).

Now how about that

Testing for the slope being equal to zero, we use

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)}$$

For the squared test statistic we get

$$t^2 = \left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)}\right)^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2/SXX} = \frac{\hat{\beta}_1^2 \times SXX}{\hat{\sigma}^2} = \frac{(SYY - RSS)/1}{RSS/n-2} = \frac{MS_{reg}}{MSE} = F$$

$\longrightarrow$ The squared t statistic is the same as the F statistic from the ANOVA!

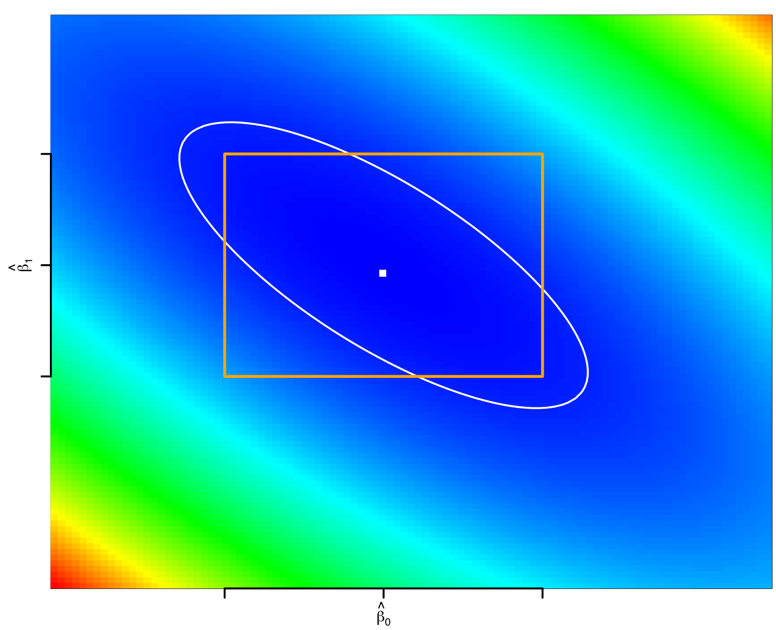
Joint confidence region

A 95% joint confidence region for the two parameters is the set of all values (β_0, β_1) that fulfill

$$\frac{\begin{pmatrix}\Delta\beta_0\\ \Delta\beta_1\end{pmatrix}^T \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix} \begin{pmatrix}\Delta\beta_0\\ \Delta\beta_1\end{pmatrix}}{2\hat{\sigma}^2} \quad \leq \quad F_{(0.95),2,n-2}$$

where $\Delta\beta_0 = \beta_0 - \hat{\beta}_0$ and $\Delta\beta_1 = \beta_1 - \hat{\beta}_1$.

Joint confidence region



Notation

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

We previously defined

$$\begin{aligned} SXX &= \sum_i (x_i - \bar{x})^2 = \sum_i x_i^2 - n(\bar{x})^2 \\ SYY &= \sum_i (y_i - \bar{y})^2 = \sum_i y_i^2 - n(\bar{y})^2 \\ SXY &= \sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y} \end{aligned}$$

We also define

$$r_{XY} = \frac{SXY}{\sqrt{SXX}\sqrt{SYY}}$$ (called the sample correlation)

Coefficient of determination

We previously wrote

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

Define

$$R^2 = \frac{SS_{reg}}{SYY} = 1 - \frac{RSS}{SYY}$$

R^2 is often called the coefficient of determination. Notice that

$$R^2 = \frac{SS_{reg}}{SYY} = \frac{(SXY)^2}{SXX \times SYY} = r_{XY}^2$$

The Anscombe Data