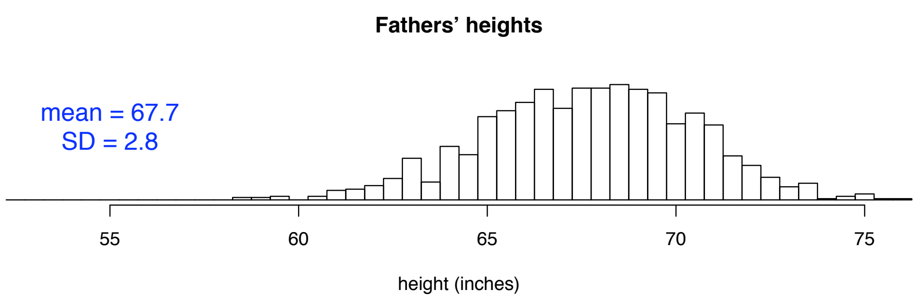
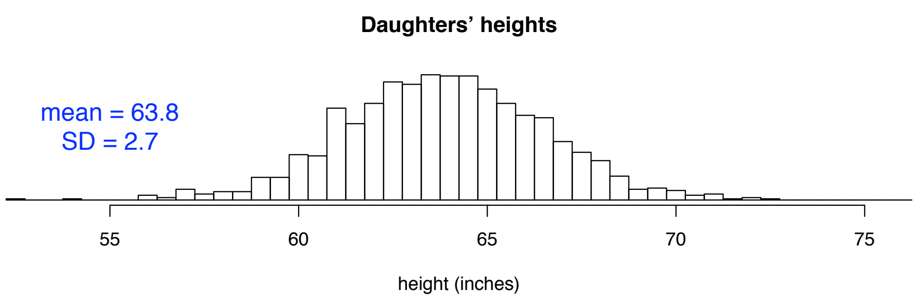
Correlation and Regression

Fathers' and daughters' heights

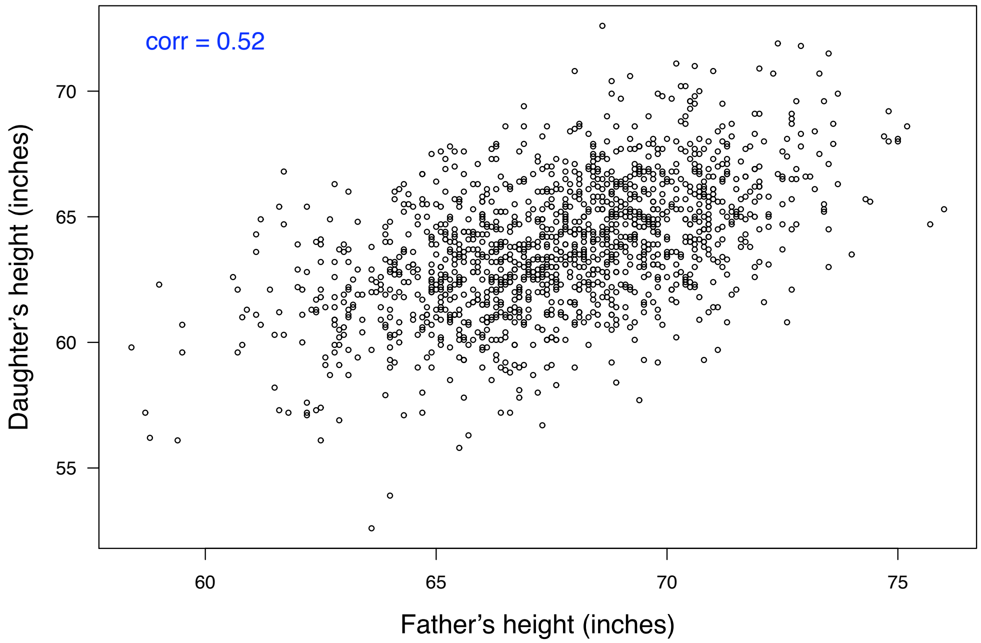




Reference: Pearson and Lee (1906) Biometrika 2:357-462

1376 pairs

Fathers' and daughters' heights



Reference: Pearson and Lee (1906) Biometrika 2:357-462 — 1376 pairs

Covariance and correlation

Let X and Y be random variables with

$$\mu_X = E(X),\ \mu_Y = E(Y),\ \sigma_X = SD(X),\ \sigma_Y = SD(Y)$$

For example, sample a father/daughter pair and let

X = the father's height and Y = the daughter's height.

Covariance

$$\text{cov}(X,Y) = E\{(X - \mu_X)\,(Y - \mu_Y)\}$$

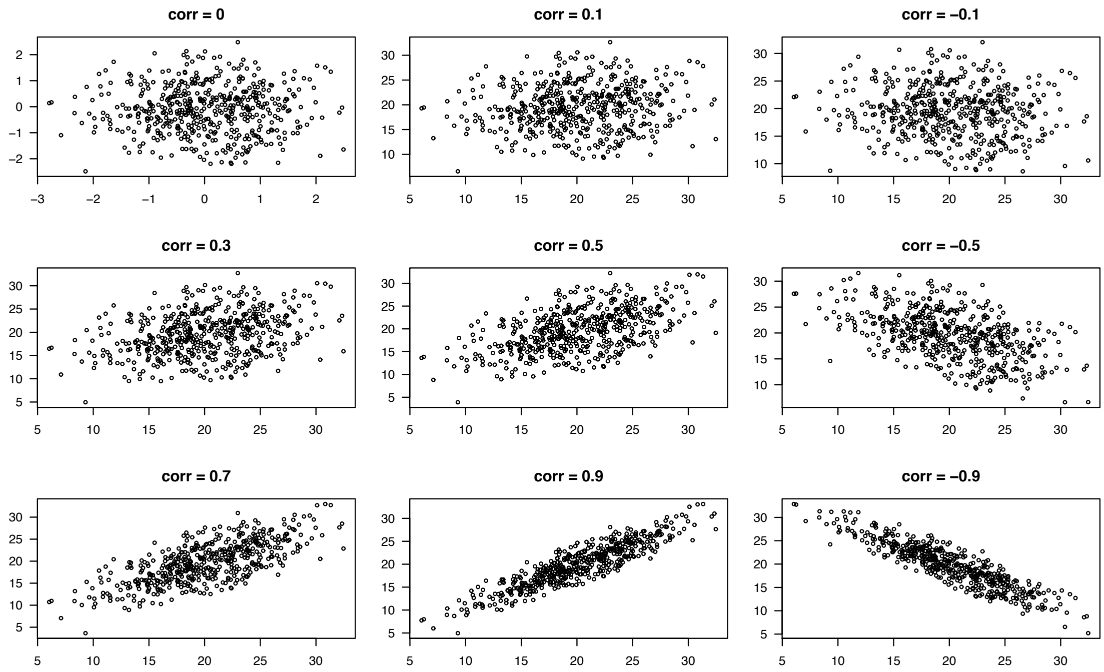
$\longrightarrow$ cov(X,Y) can be any real number

Correlation

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$\longrightarrow$ $-1 \leq \text{cor}(X,Y) \leq 1$

Examples



Estimated correlation

Consider n pairs of data: $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$

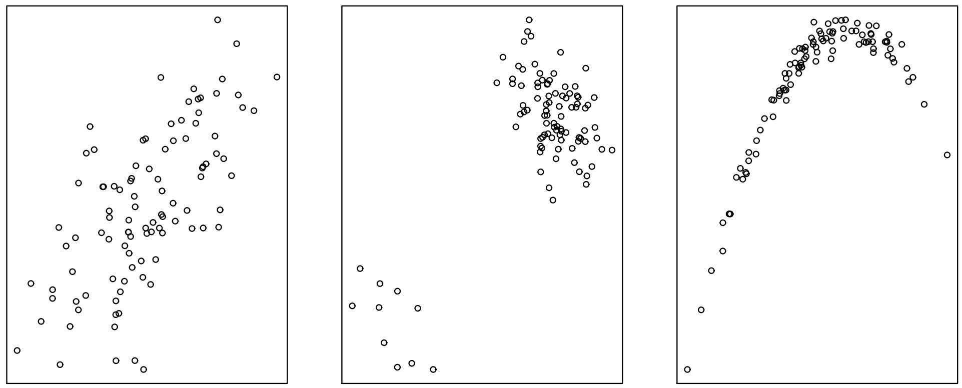
We consider these as independent draws from some bivariate distribution.

We estimate the correlation in the underlying distribution by:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \ \sum_i (y_i - \bar{y})^2}}$$

This is sometimes called the correlation coefficient.

Correlation measures linear association



$\longrightarrow$ All three plots have correlation ≈ 0.7!

Correlation versus regression

$\longrightarrow$ Covariance / correlation:

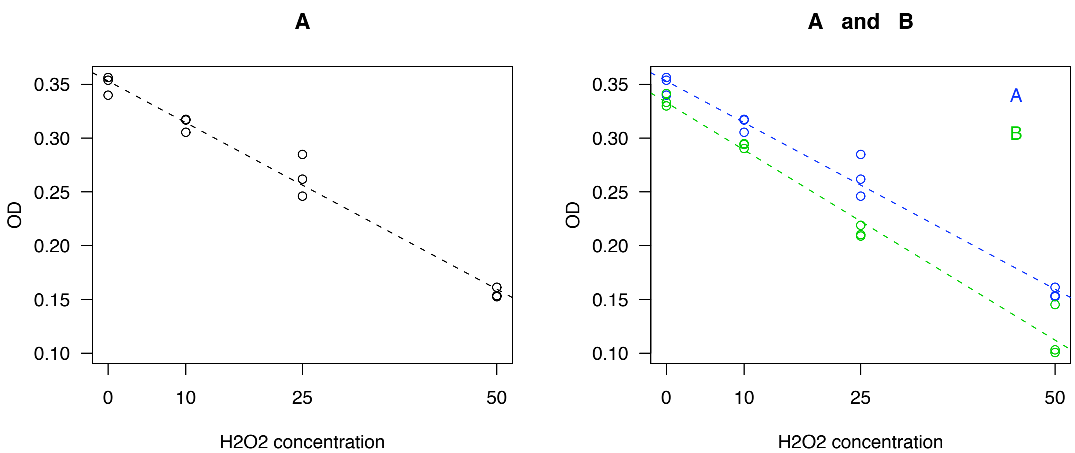
- Quantifies how two random variables X and Y co-vary.
- There is typically no particular order between the two random variables (e. g. , fathers' versus daughters' height).

$\longrightarrow$ Regression

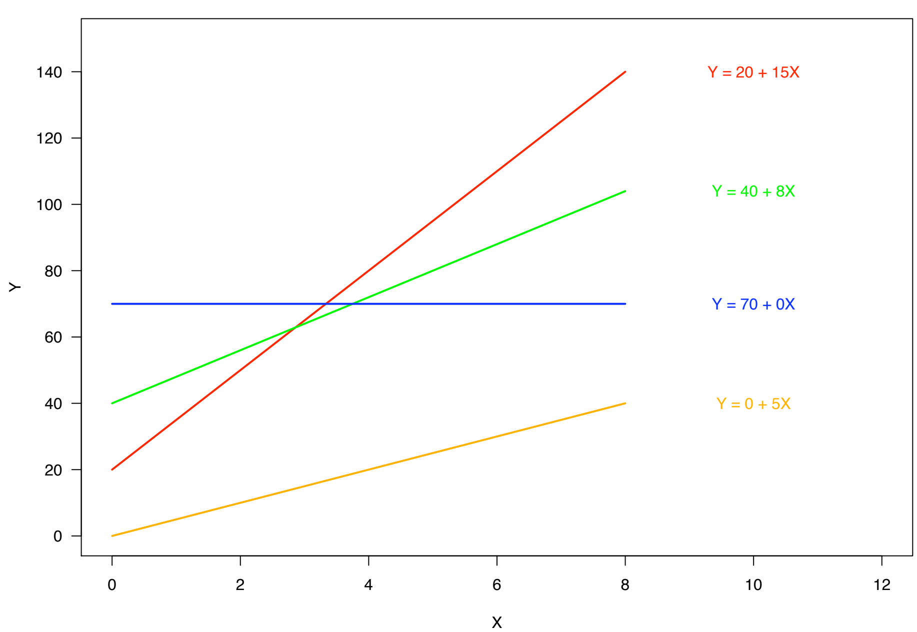
- Assesses the relationship between predictor X and response Y: we model $E[Y|X]$.
- The values for the predictor are often deliberately chosen, and are therefore not random quantities.
- We typically assume that we observe the values for the predictor(s) without error.

Example

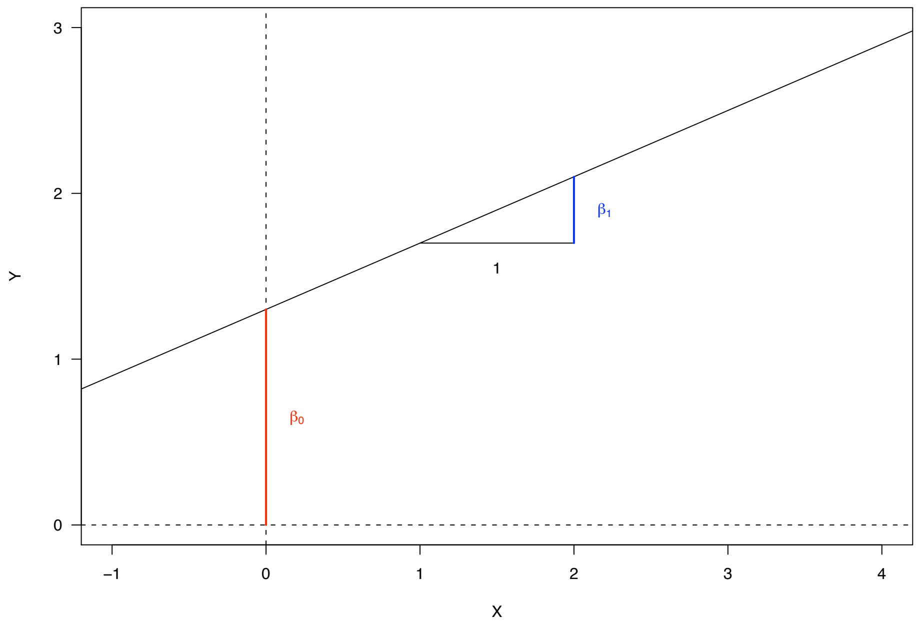
Measurements of degradation of heme with different concentrations of hydrogen peroxide (H_2O_2), for different types of heme.



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

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

This implies:

$$E[Y|X] = \beta_0 + \beta_1 X.$$

Interpretation:

For two subjects that differ by one unit in X, we expect the responses to differ by β_1.

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

Fitted values and residuals

We can write

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

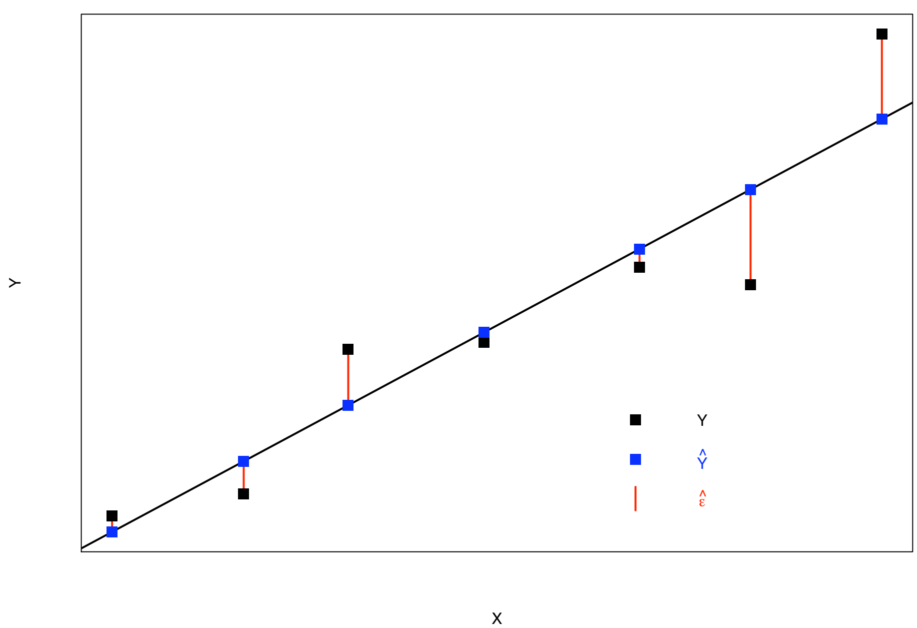
For a pair of estimates $(\hat{\beta}_0, \hat{\beta}_1)$ for the pair of parameters (β_0, β_1) we define the fitted values as

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

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.

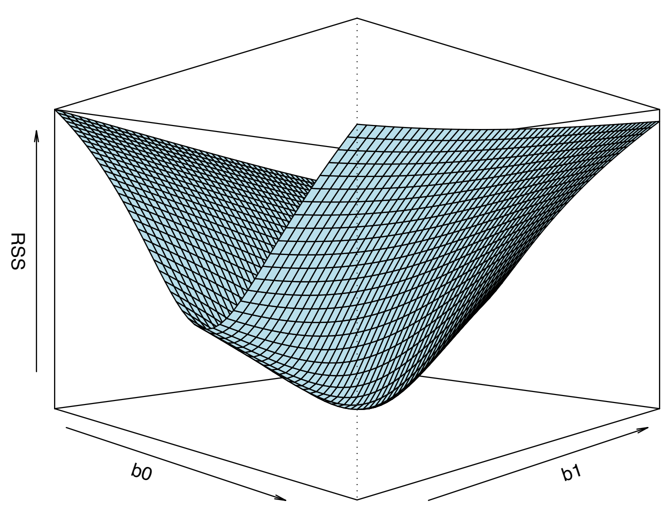
$$\text{RSS}(\beta_0, \beta_1) = \sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

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



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

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

is minimized by

$$\begin{aligned}
\hat{\beta}_1 &= \frac{SXY}{SXX} \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}$$

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

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

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

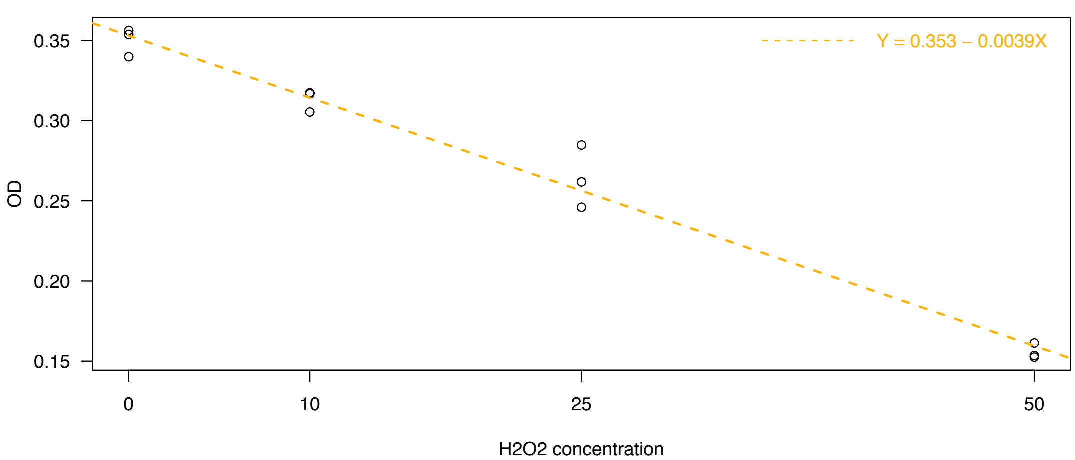
$\bar{x}$=21.25, $\bar{y}$=0.27, SXX=4256.25, SXY=– 16.48, RSS=0.0013.

Therefore

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

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

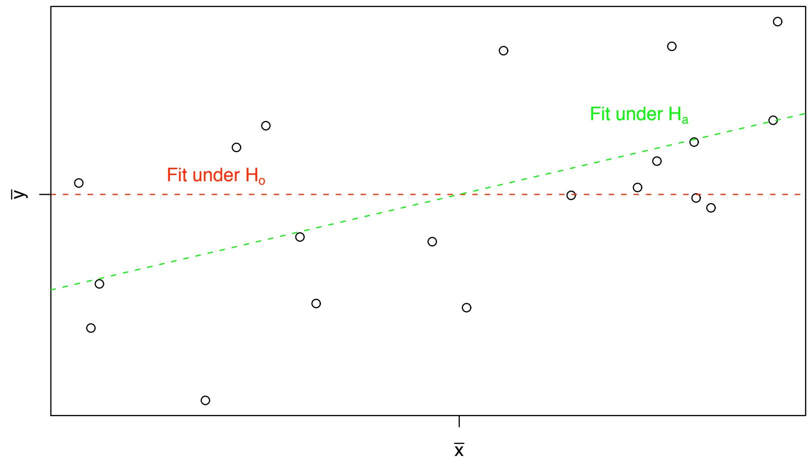
Example



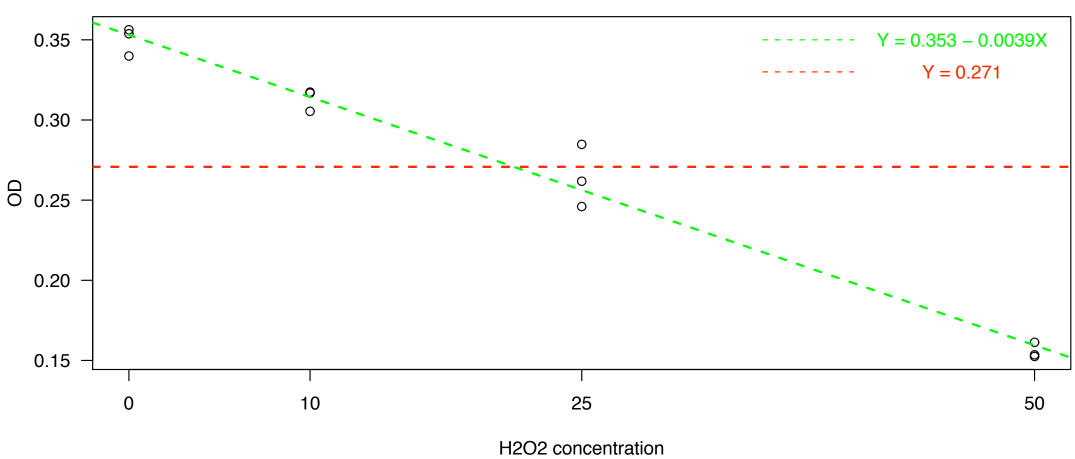
Comparing models

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

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



Example



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		

Example

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		

Parameter estimates

One can show that

$$E(\hat{\beta}_0) = \beta_0 \qquad E(\hat{\beta}_1) = \beta_1$$

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) \qquad \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SXX}$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\bar{x}}{SXX} \qquad \mathrm{Cor}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sqrt{\bar{x}^2 + SXX/n}}$$

⟶ Note: We're thinking of the x's as fixed.

Parameter estimates

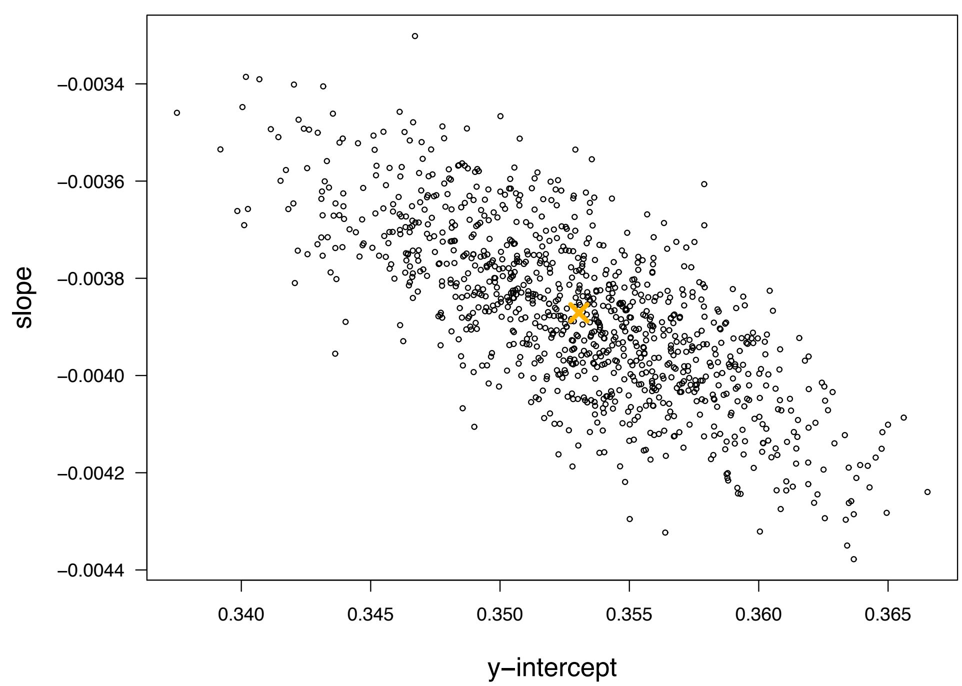
One can even show that the distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$ is a bivariate normal distribution!

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

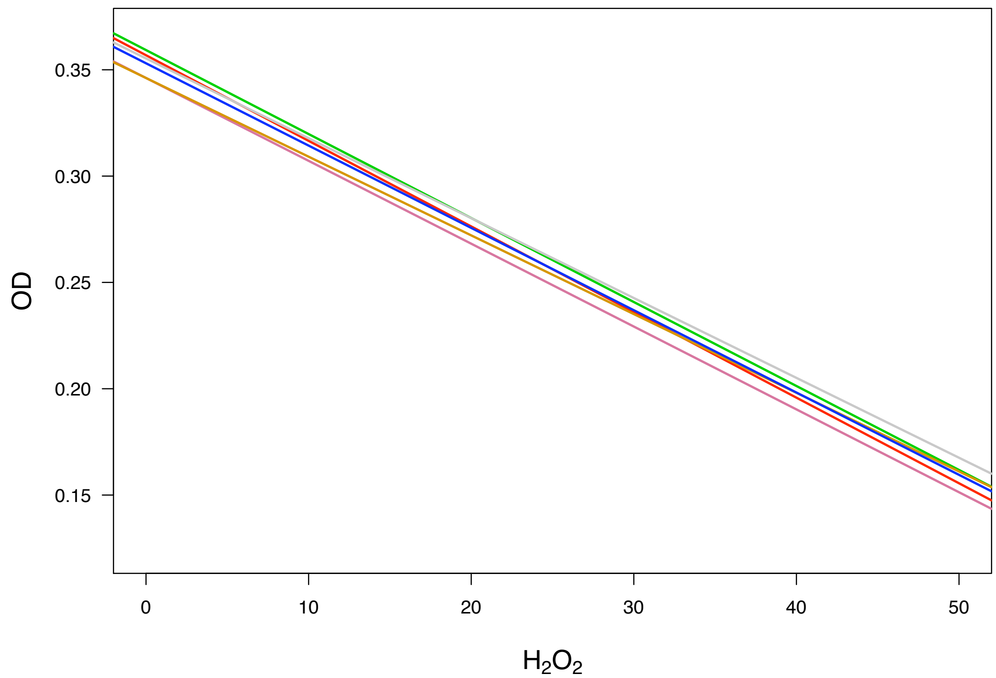
where

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

Simulation: coefficients



Possible outcomes



Confidence intervals

We know that

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

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

$\longrightarrow$ 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

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

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 example 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!

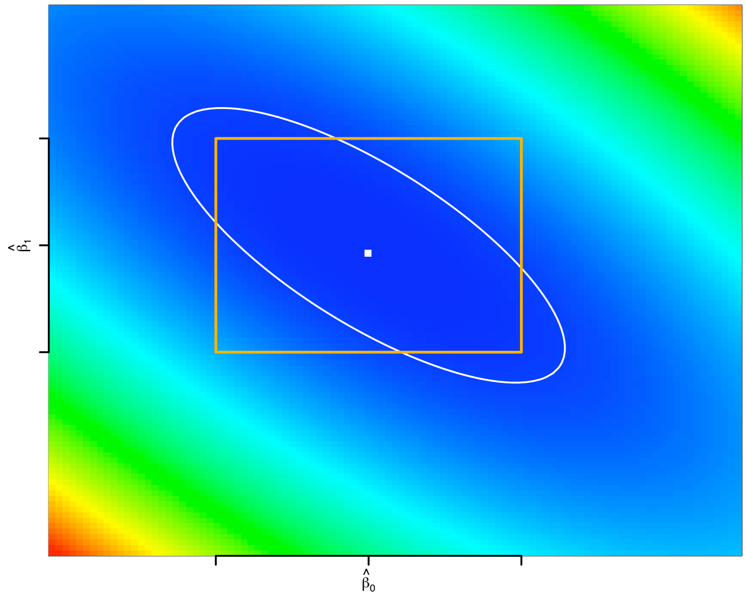
Joint confidence region

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

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

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

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

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

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

We also define

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

Coefficient of determination

We previously wrote

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

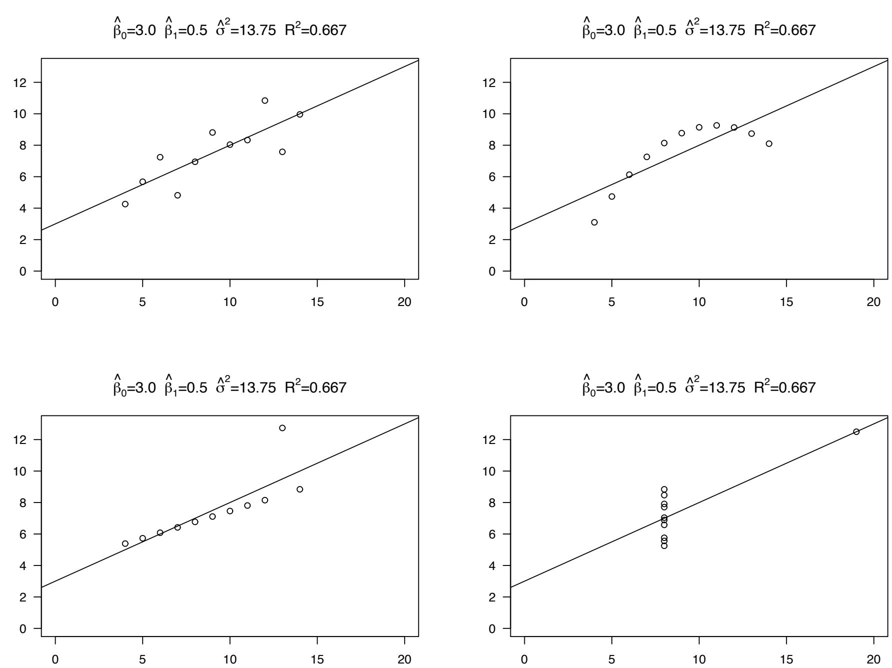
Define

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

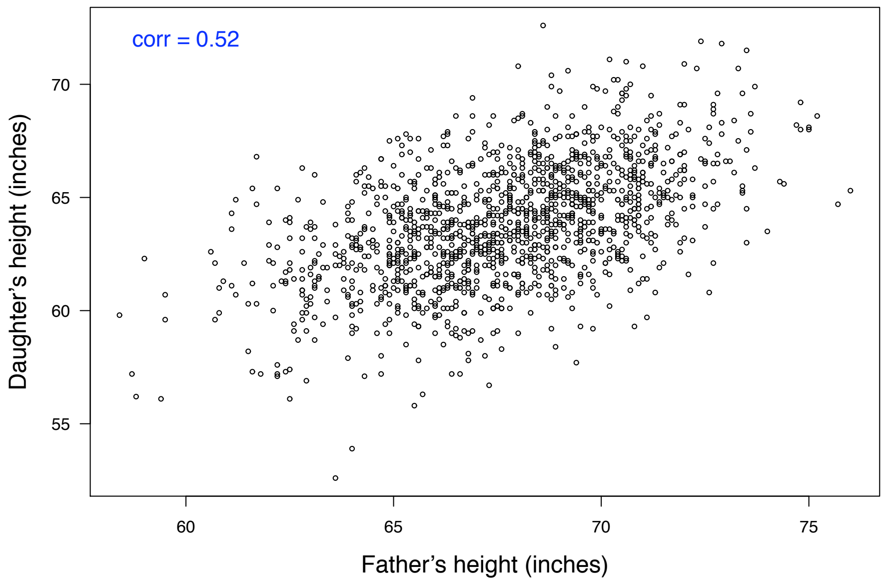
R^2 is often called the coefficient of determination. Notice that

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

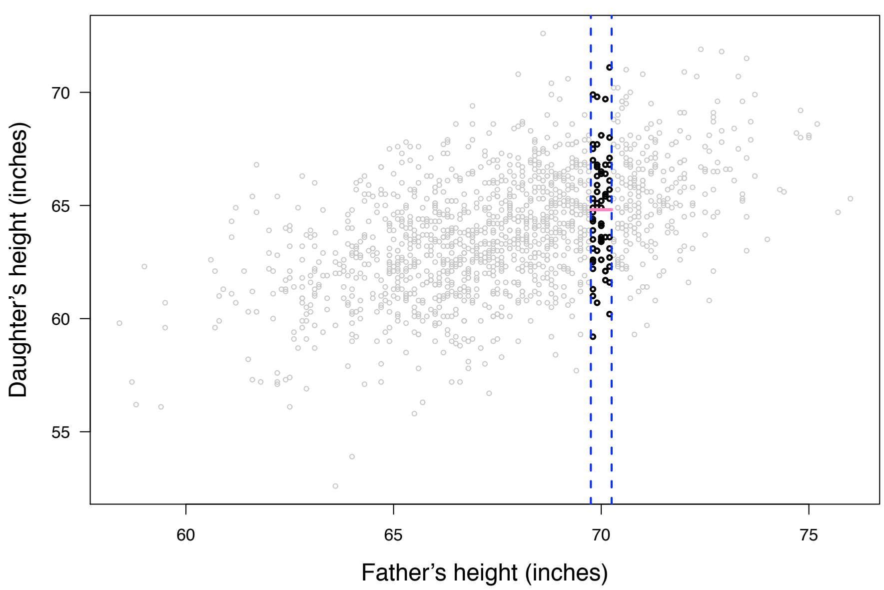
The Anscombe Data



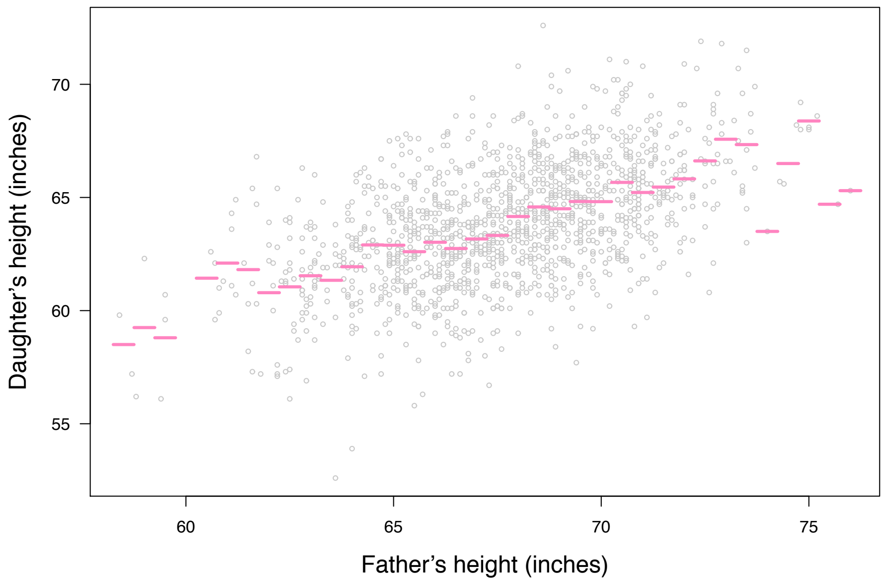
Fathers' and daughters' heights



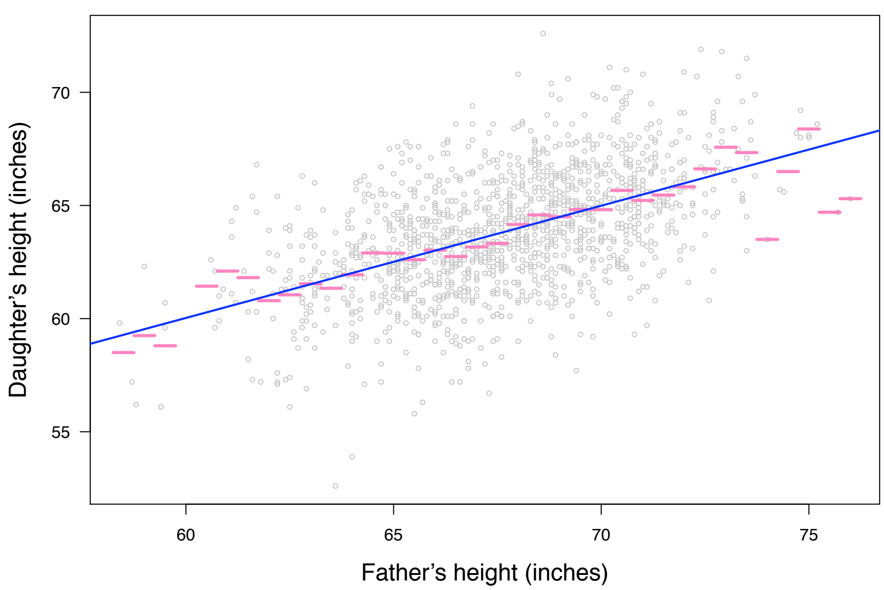
Linear regression



Linear regression

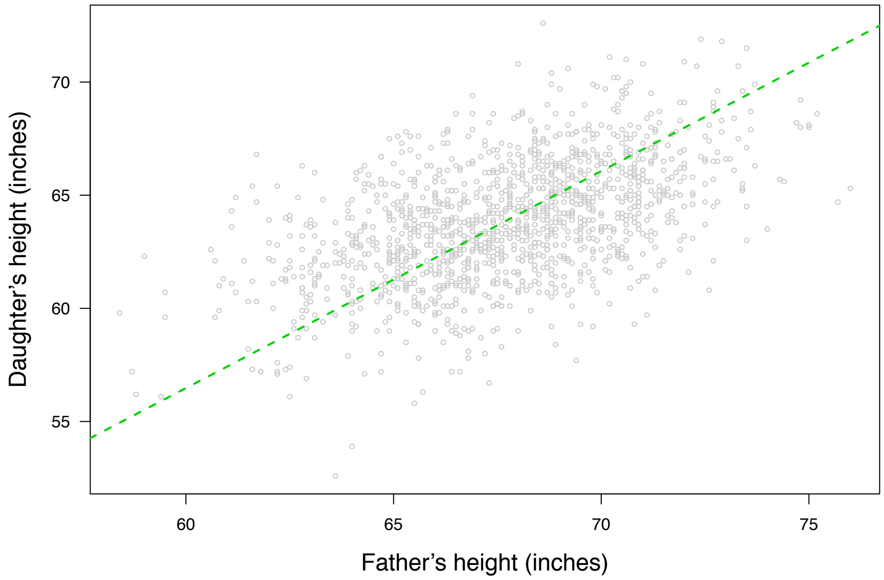


Regression line



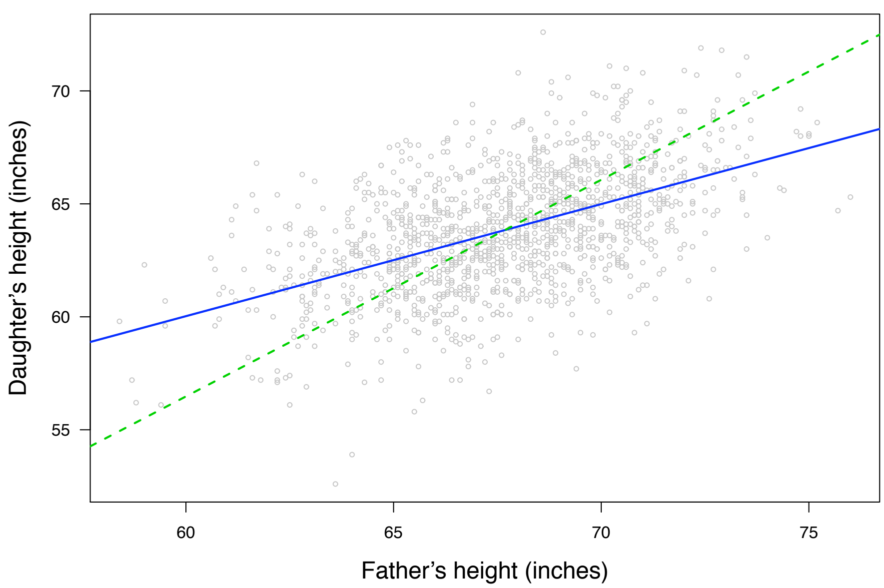
⟶ Slope = r × SD(Y) / SD(X)

SD line



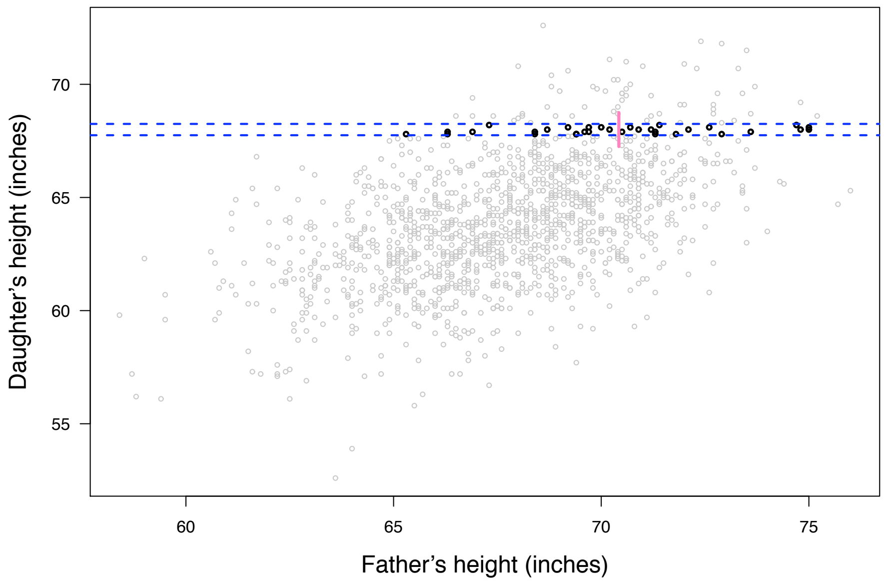
⟶ Slope = SD(Y) / SD(X)

SD line vs regression line

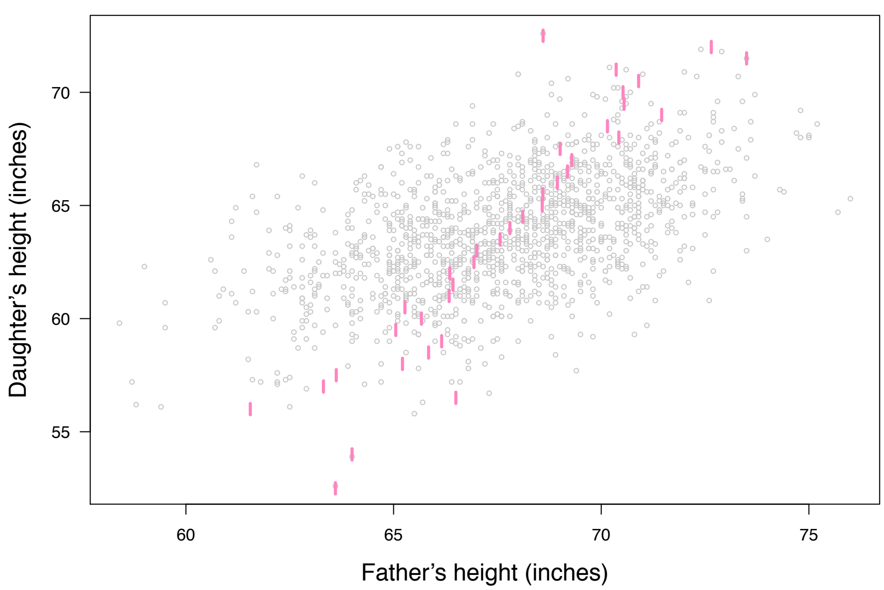


⟶ Both lines go through the point ($\bar{X}$, $\bar{Y}$).

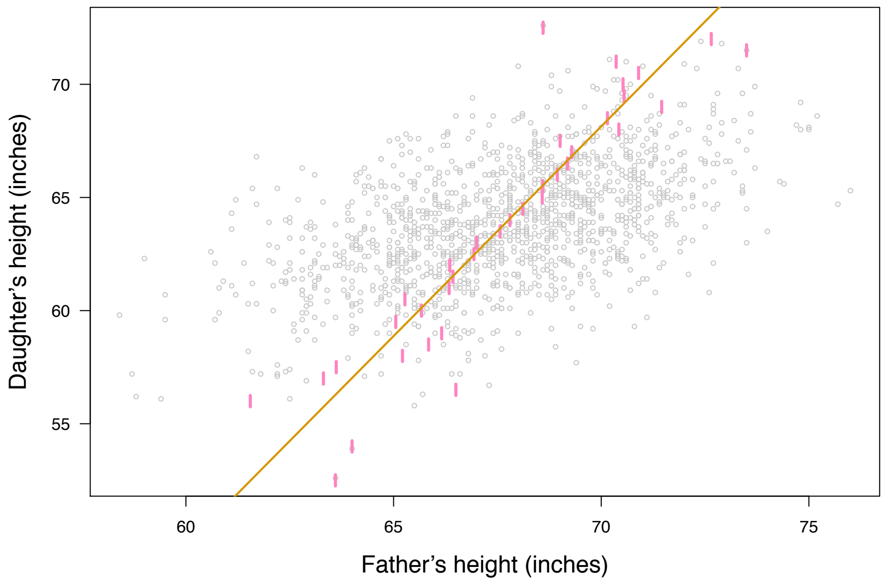
Predicting father's ht from daughter's ht



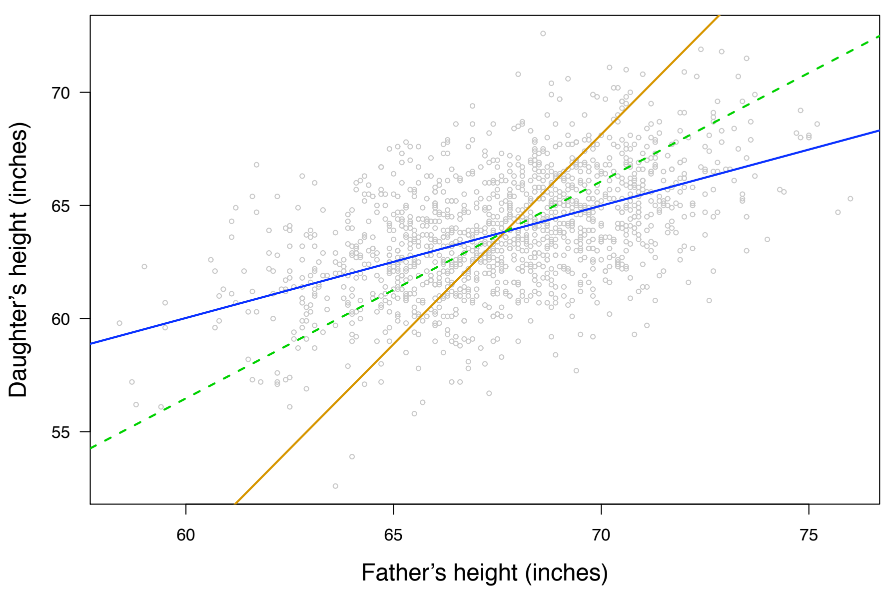
Predicting father's ht from daughter's ht



Predicting father's ht from daughter's ht



There are two regression lines!



The equations

Regression of y on x (for predicting y from x)

Slope = $r \frac{SD(y)}{SD(x)}$ Goes through the point $(\bar{x}, \bar{y})$

$\hat{y} - \bar{y} = r \frac{SD(y)}{SD(x)} (x - \bar{x})$

$\longrightarrow \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ where $\hat{\beta}_1 = r \frac{SD(y)}{SD(x)}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

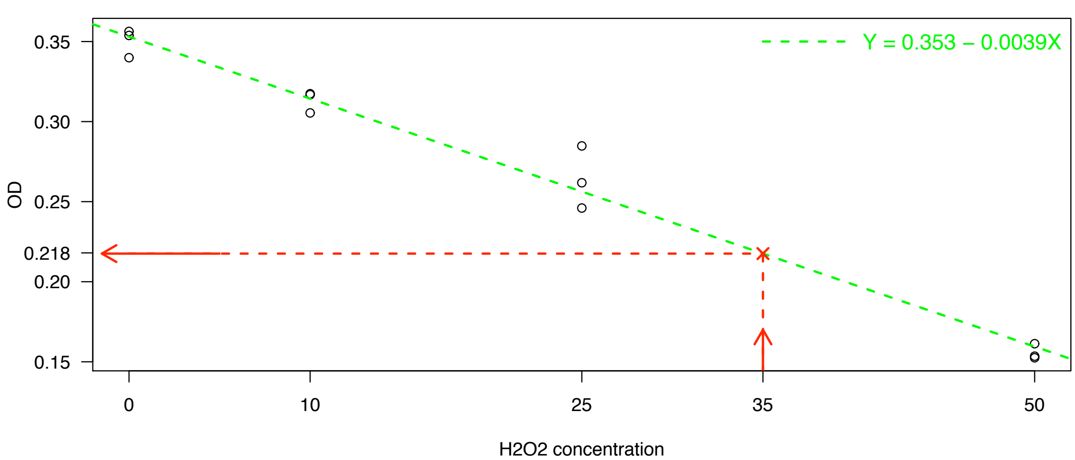
Regression of x on y (for predicting x from y)

Slope = $r \frac{SD(x)}{SD(y)}$ Goes through the point $(\bar{y}, \bar{x})$

$\hat{x} - \bar{x} = r \frac{SD(x)}{SD(y)} (y - \bar{y})$

$\longrightarrow \quad \hat{x} = \hat{\beta}^\star_0 + \hat{\beta}^\star_1 y$ where $\hat{\beta}^\star_1 = r \frac{SD(x)}{SD(y)}$ and $\hat{\beta}^\star_0 = \bar{x} - \hat{\beta}^\star_1 \bar{y}$

Estimating the mean response



$\longrightarrow$ We can use the regression results to predict the expected response for a new concentration of hydrogen peroxide. But what is its variability?

Variability of the mean response

Let $\hat{y}$ be the predicted mean for some x, i. e.

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

Then

$$\begin{aligned} E(\hat{y}) &= \beta_0 + \beta_1 x \\ \text{var}(\hat{y}) &= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right) \end{aligned}$$

where $y = \beta_0 + \beta_1 x$ is the true mean response.

Why?

$$\begin{aligned} E(\hat{y}) &= E(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= E(\hat{\beta}_0) + x\,E(\hat{\beta}_1) \\ &= \beta_0 + x\,\beta_1 \end{aligned}$$

$$\begin{aligned} \text{var}(\hat{y}) &= \text{var}(\hat{\beta}_0 + \hat{\beta}_1 x) \\ &= \text{var}(\hat{\beta}_0) + \text{var}(\hat{\beta}_1 x) + 2\,\text{cov}(\hat{\beta}_0, \hat{\beta}_1 x) \\ &= \text{var}(\hat{\beta}_0) + x^2\,\text{var}(\hat{\beta}_1) + 2\,x\,\text{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SXX} \right) + \sigma^2 \left(\frac{x^2}{SXX} \right) - \frac{2\,x\,\bar{x}\,\sigma^2}{SXX} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right] \end{aligned}$$

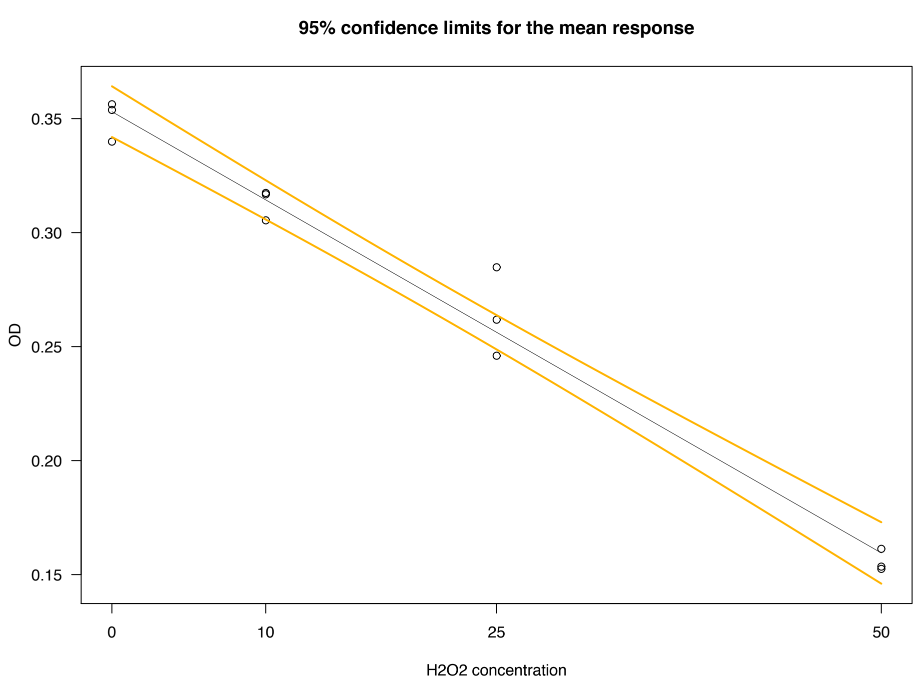
Confidence intervals

Hence

$$\hat{y} \pm t_{(1-\frac{\alpha}{2}),n-2} \times \hat{\sigma} \times \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SXX}}$$

is a $(1-\alpha)\times 100\%$ confidence interval for the mean response given x.

Confidence limits



Prediction

Now assume that we want to calculate an interval for the predicted response $y^\star$ for a value of x.

There are two sources of uncertainty:

(a) the mean response

(b) the natural variation σ^2

The variance of $\hat{y}^\star$ is

$$\text{var}(\hat{y}^\star) = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX} \right)$$

Prediction intervals

Hence

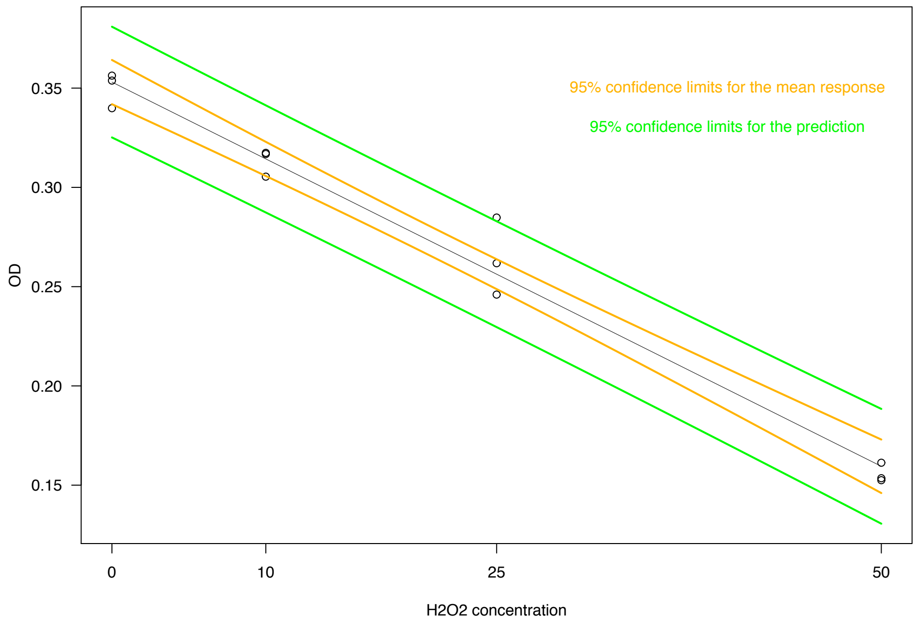
$$\hat{y}^\star \pm t_{(1-\frac{\alpha}{2}), n-2} \times \hat{\sigma} \times \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{SXX}}$$

is a $(1-\alpha)\times 100\%$ prediction interval for the predicted response given x.

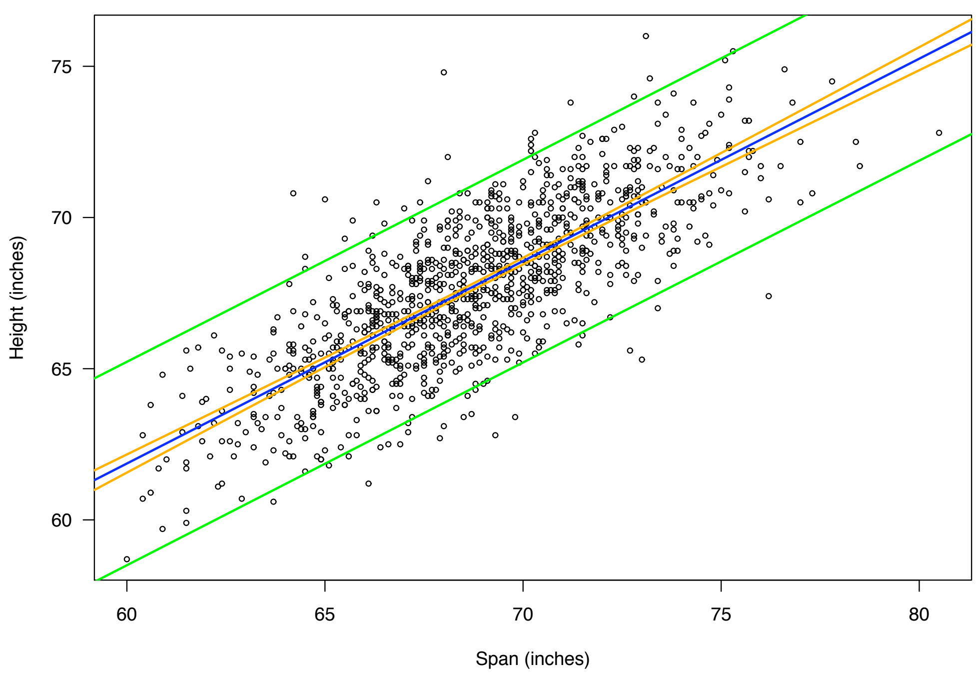
$\longrightarrow$ When n is very large, we get roughly

$$\hat{y}^\star \pm t_{(1-\frac{\alpha}{2}), n-2} \times \hat{\sigma}$$

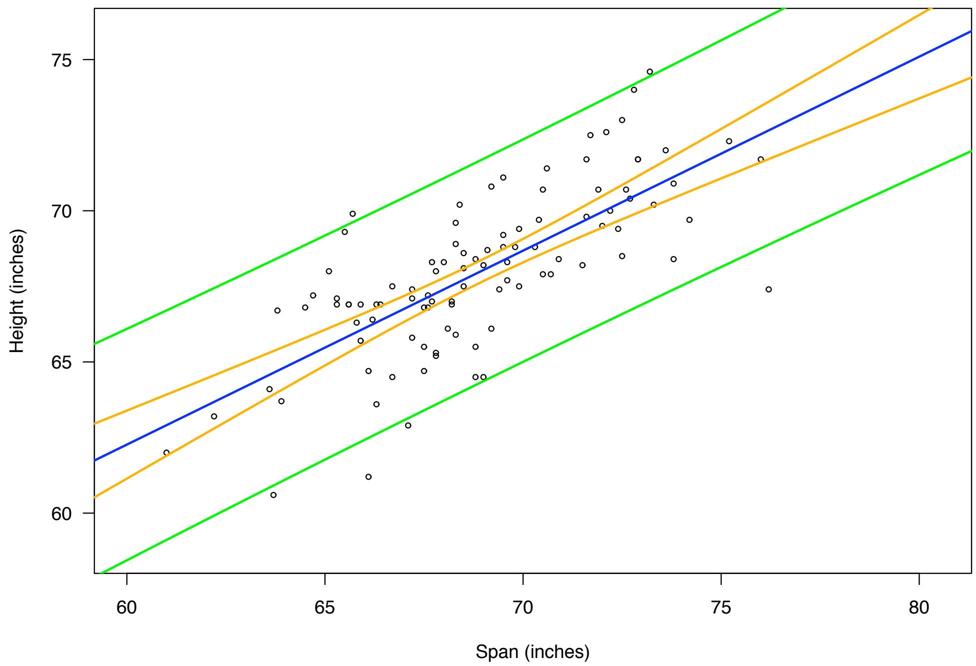
Prediction intervals



Span and height



With just 100 individuals



Regression for calibration

That prediction interval is for the case that the x's are known without error while

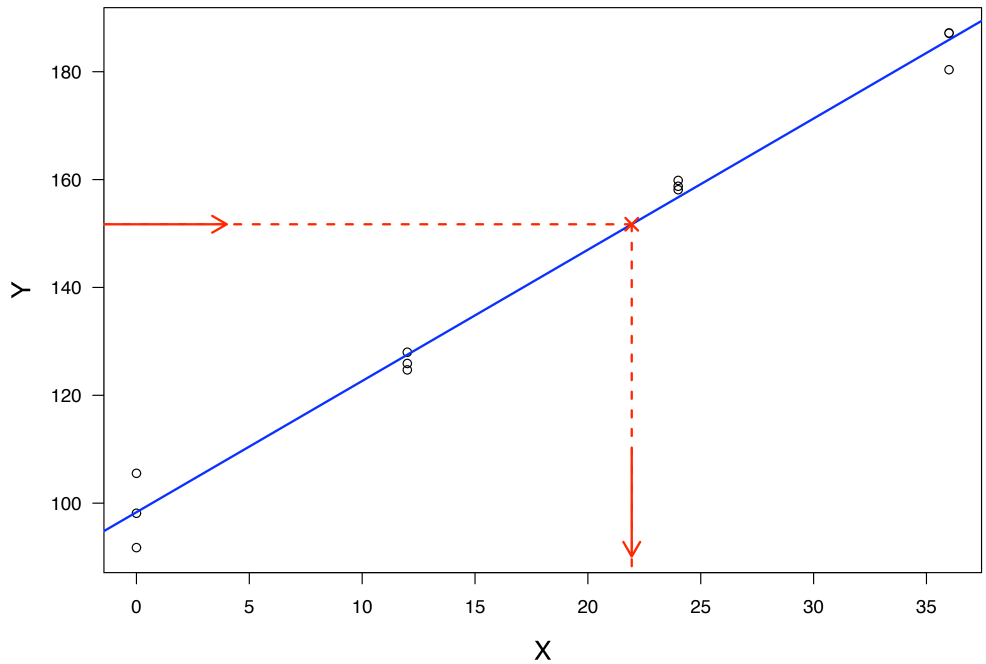
$$y = \beta_0 + \beta_1 x + \epsilon \quad \text{where } \epsilon = \text{error}$$

⟶ Another common situation:

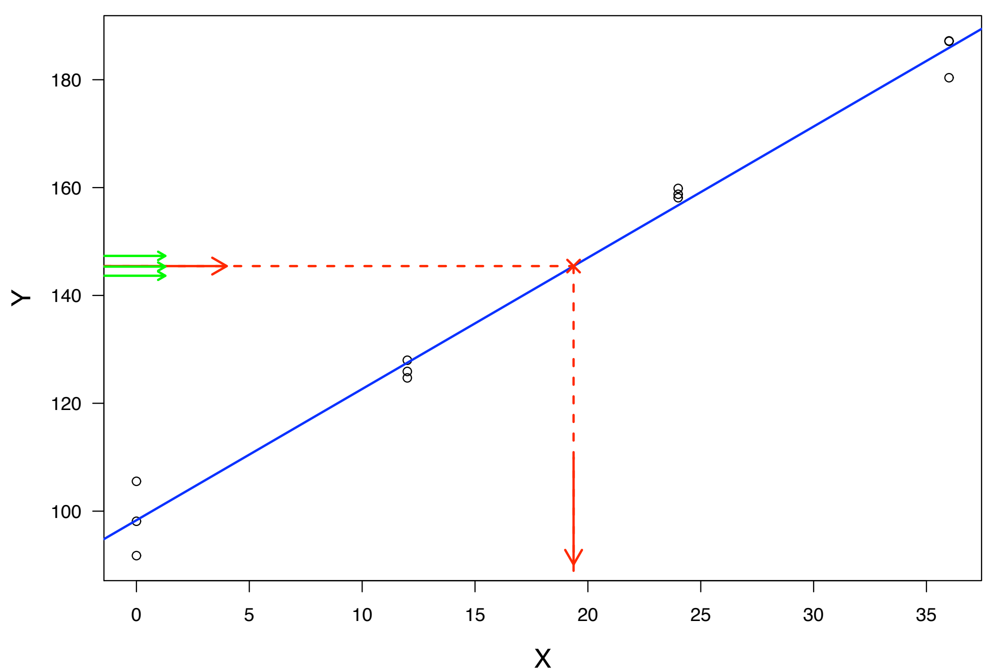
- ○ We have a number of pairs (x,y) to get a calibration line/curve.
- ○ x's basically without error; y's have measurement error.
- ○ We obtain a new value, $y^\star$, and want to estimate the corresponding $x^\star$:

$$y^\star = \beta_0 + \beta_1 x^\star + \epsilon$$

Example



Another example



Regression for calibration

⟶ Data: (x_i, y_i) for $i = 1, \ldots, n$

with $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $\epsilon_i \sim$ iid Normal$(0, \sigma)$

$y_j^\star$ for $j = 1, \ldots, m$

with $y_j^\star = \beta_0 + \beta_1 x^\star + \epsilon_j^\star$, $\epsilon_j^\star \sim$ iid Normal$(0, \sigma)$ for some $x^\star$

⟶ Goal:
Estimate $x^\star$ and give a 95% confidence interval.

⟶ The estimate:

Obtain $\hat{\beta}_0$ and $\hat{\beta}_1$ by regressing the y_i on the x_i.

Let $\hat{x}^\star = (\bar{y}^\star - \hat{\beta}_0)/\hat{\beta}_1$ where $\bar{y}^\star = \sum_j y_j^\star / m$

95% CI for $\hat{x}^\star$

Let T denote the 97.5th percentile of the t distr'n with n–2 d.f.

Let $g = T \,/\, [|\hat{\beta}_1| \,/\, (\hat{\sigma}/\sqrt{SXX})] \quad = \quad (T\,\hat{\sigma}) \,/\, (|\hat{\beta}_1| \sqrt{SXX})$

⟶ If $g \geq 1$, we would fail to reject $H_0 : \beta_1 = 0$!

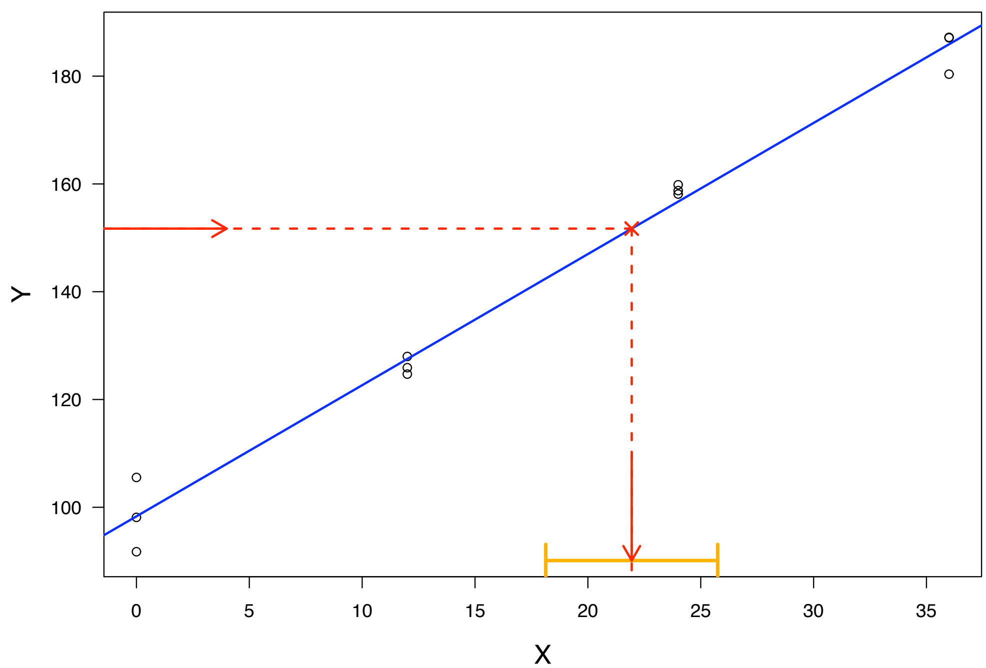
In this case, the 95% CI for $\hat{x}^\star$ is $(-\infty, \infty)$.

⟶ If $g < 1$, our 95% CI is the following:

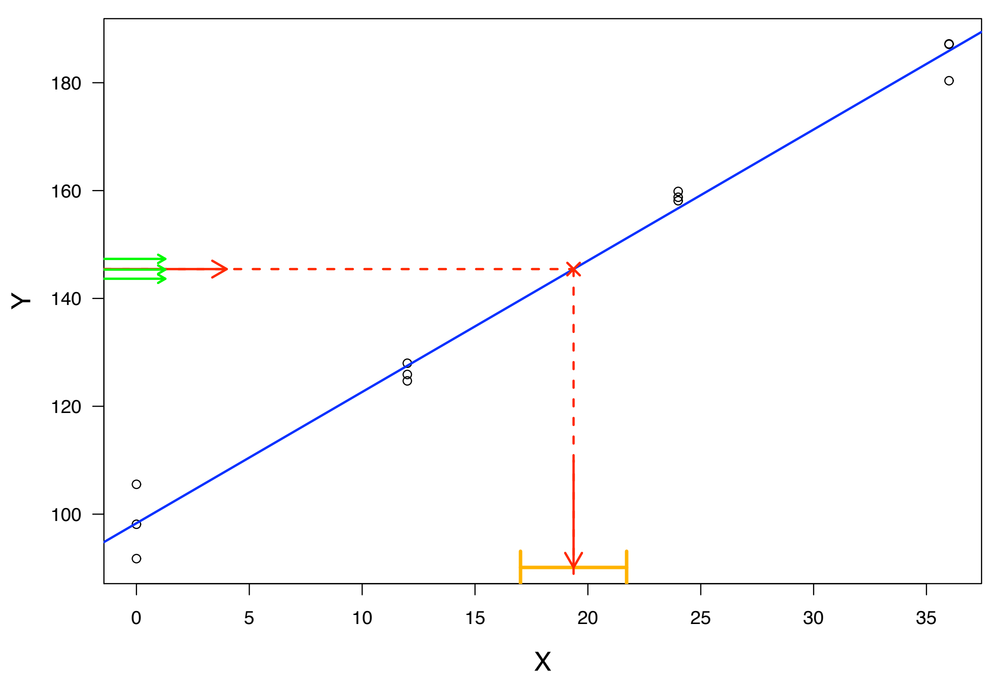
$$\hat{x}^\star \pm \frac{(\hat{x}^\star - \bar{x})\, g^2 + (T\,\hat{\sigma} \,/\, |\hat{\beta}_1|)\sqrt{(\hat{x}^\star - \bar{x})^2/SXX + (1 - g^2)\,(\frac{1}{m} + \frac{1}{n})}}{1 - g^2}$$

For very large n, this reduces to approximately $\hat{x}^\star \pm (T\,\hat{\sigma}) \,/\, (|\hat{\beta}_1| \sqrt{m})$

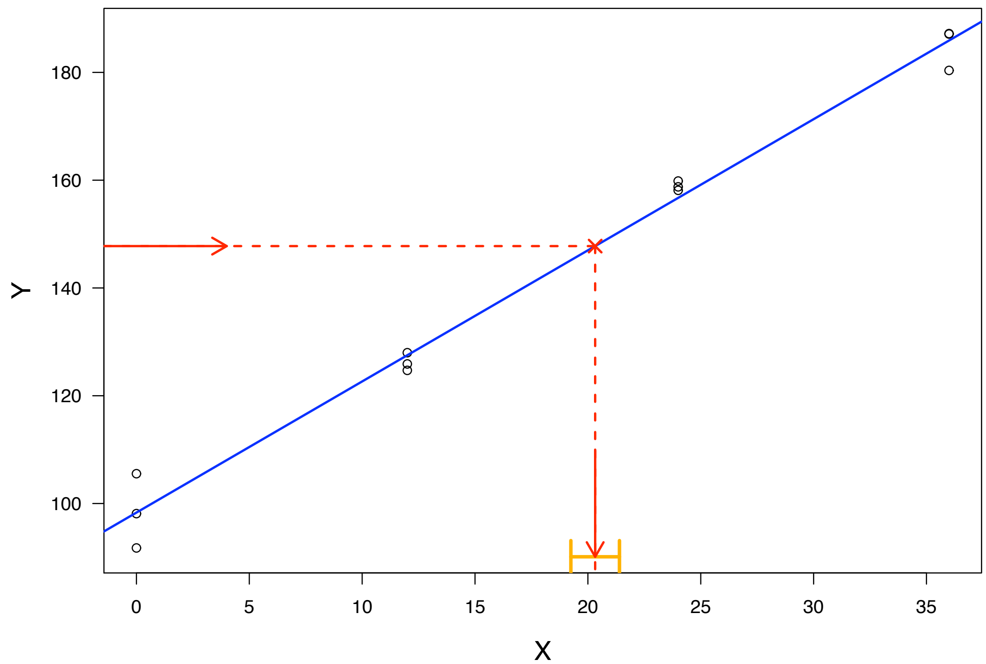
Example



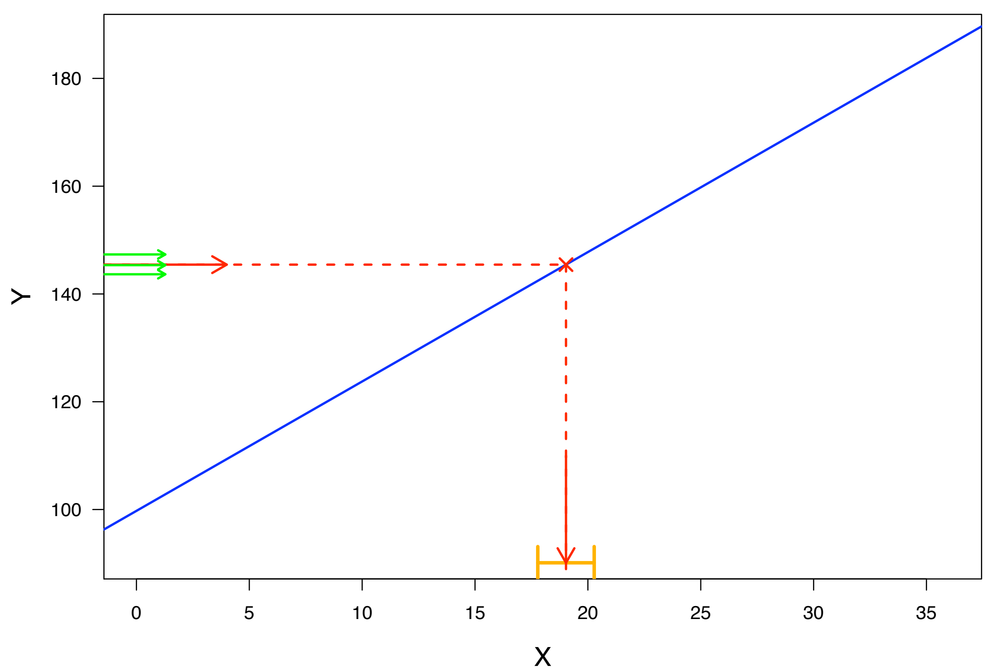
Another example



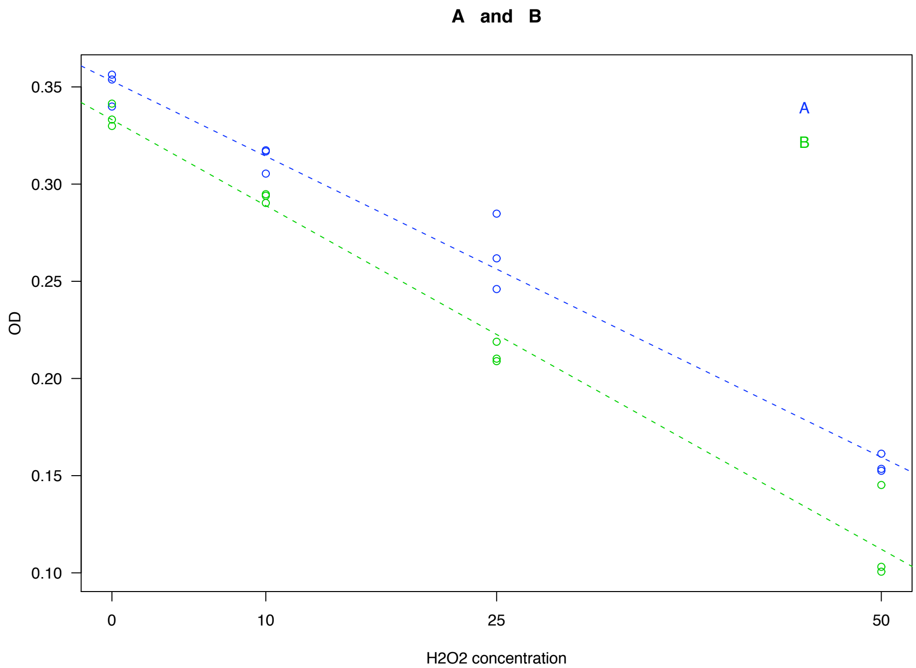
Infinite m



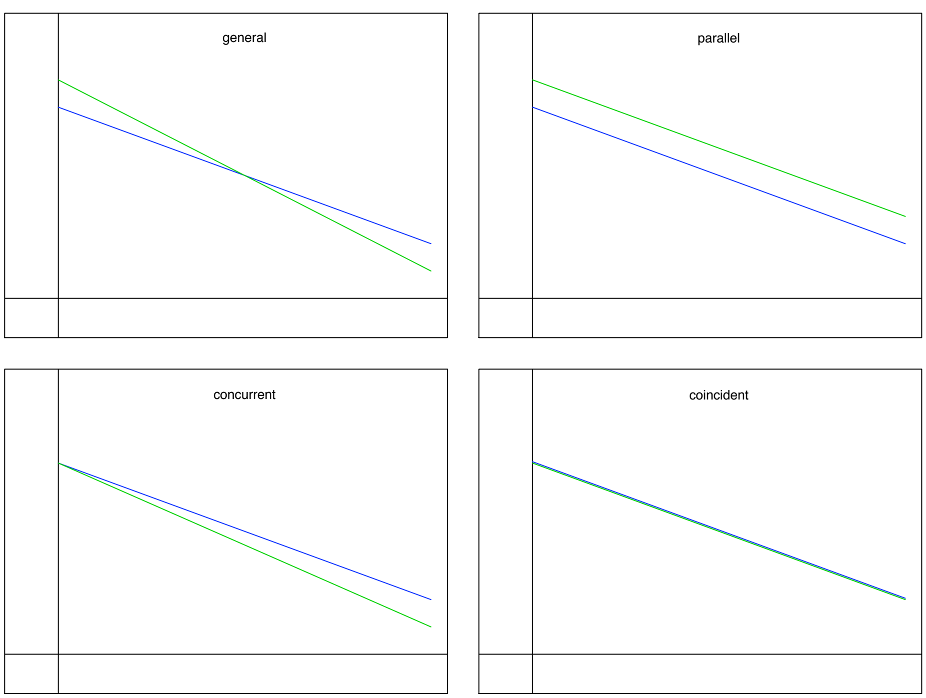
Infinite n



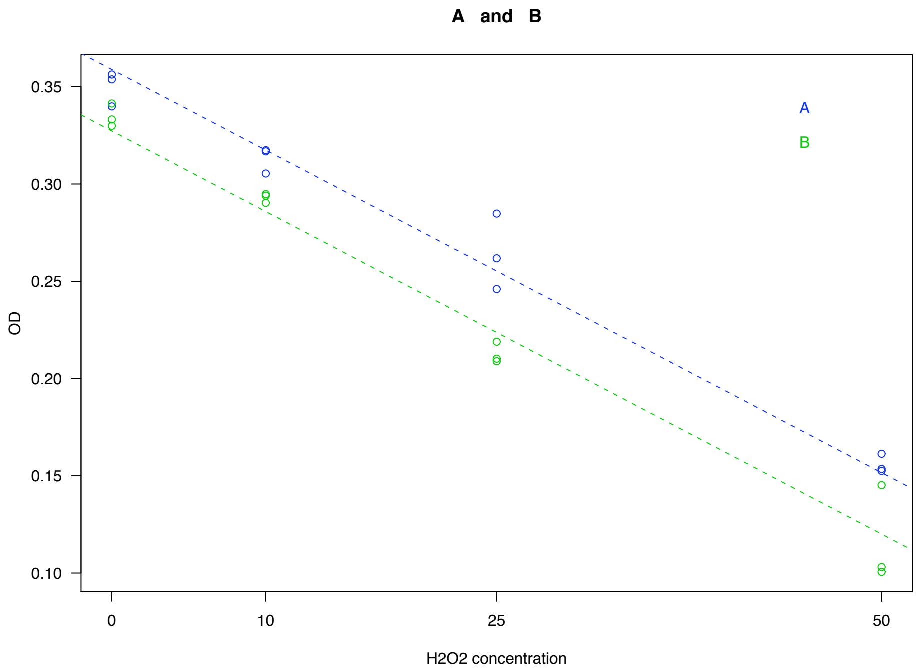
Multiple linear regression



Multiple linear regression



Multiple linear regression



More than one predictor

#	Y	X_1	X_2
1	0.3399	0	0
2	0.3563	0	0
3	0.3538	0	0
4	0.3168	10	0
5	0.3054	10	0
6	0.3174	10	0
7	0.2460	25	0
8	0.2618	25	0
9	0.2848	25	0
10	0.1535	50	0
11	0.1613	50	0
12	0.1525	50	0
13	0.3332	0	1
14	0.3414	0	1
15	0.3299	0	1
16	0.2940	10	1
17	0.2948	10	1
18	0.2903	10	1
19	0.2089	25	1
20	0.2189	25	1
21	0.2102	25	1
22	0.1006	50	1
23	0.1031	50	1
24	0.1452	50	1

The model with two parallel lines can be described as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

In other words (or, equations):

$$Y = \begin{cases} \beta_0 + \beta_1 X_1 + \epsilon & \text{if } X_2 = 0 \\ (\beta_0 + \beta_2) + \beta_1 X_1 + \epsilon & \text{if } X_2 = 1 \end{cases}$$

Multiple linear regression

A multiple linear regression model has the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

The predictors (the X's) can be categorical or numerical.

Often, all predictors are numerical or all are categorical.

And actually, categorical variables are converted into a group of numerical ones.

Interpretation

Let X_1 be the age of a subject (in years).

$$E[Y] = \beta_0 + \beta_1 X_1$$

$\longrightarrow$ Comparing two subjects who differ by one year in age, we expect the responses to differ by β_1.

$\longrightarrow$ Comparing two subjects who differ by five years in age, we expect the responses to differ by $5\beta_1$.

Interpretation

Let X_1 be the age of a subject (in years), and let X_2 be an indicator for the treatment arm (0/1).

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

⟶ Comparing two subjects from the same treatment arm who differ by one year in age, we expect the responses to differ by β_1.

⟶ Comparing two subjects of the same age from the two different treatment arms (X_2=1 versus X_2=0), we expect the responses to differ by β_2.

Interpretation

Let X_1 be the age of a subject (in years), and let X_2 be an indicator for the treatment arm (0/1).

$$E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

⟶ $E[Y] = \beta_0 + \beta_1 X_1$ (if X_2=0)

⟶ $E[Y] = \beta_0 + \beta_1 X_1 + \beta_2 + \beta_3 X_1 = \beta_0 + \beta_2 + (\beta_1 + \beta_3) X_1$ (if X_2=1)

⟶ Comparing two subjects who differ by one year in age, we expect the responses to differ by β_1 if they are in the control arm (X_2=0), and expect the responses to differ by $\beta_1 + \beta_3$ if they are in the treatment arm (X_2=1).

Estimation

We have the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim \text{ iid Normal}(0, \sigma^2)$$

$\longrightarrow$ We estimate the β's by the values for which

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

is minimized where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$ (aka "least squares").

$\longrightarrow$ We estimate σ by $\quad \hat{\sigma} = \sqrt{\dfrac{\text{RSS}}{n - (k+1)}}$

FYI

Calculation of the $\hat{\beta}$'s (and their SEs and correlations) is not that complicated, but without matrix algebra, the formulas are nasty.

Here is what you need to know:

- The SEs of the $\hat{\beta}$'s involve σ and the x's.
- The $\hat{\beta}$'s are normally distributed.
- Obtain confidence intervals for the β's using $\hat{\beta} \pm t \times \widehat{\text{SE}}(\hat{\beta})$
  where t is a quantile of t dist'n with n–(k+1) d.f.
- Test $H_0 : \beta = 0$ using $|\hat{\beta}|/\widehat{\text{SE}}(\hat{\beta})$
  Compare this to a t distribution with n–(k+1) d.f.

The example: a full model

$x_1 = [H_2O_2]$.

x_2 = 0 or 1, indicating type of heme.

y = the OD measurement.

The model: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon$

i.e.,

$$y = \begin{cases} \beta_0 + \beta_1 X_1 + \epsilon & \text{if } X_2 = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 + \epsilon & \text{if } X_2 = 1 \end{cases}$$

$\beta_2 = 0 \longrightarrow$ Same intercepts.

$\beta_3 = 0 \longrightarrow$ Same slopes.

$\beta_2 = \beta_3 = 0 \longrightarrow$ Same lines.

Results

```
Coefficients:
              Estimate Std. Error t value   Pr(>|t|)
(Intercept)    0.35305    0.00544    64.9   < 2e-16
x1            -0.00387    0.00019   -20.2   8.86e-15
x2            -0.01992    0.00769    -2.6     0.0175
x1:x2         -0.00055    0.00027    -2.0     0.0563

Residual standard error: 0.0125 on 20 degrees of freedom
Multiple R-Squared:  0.98,Adjusted R-squared: 0.977
F-statistic: 326.4 on 3 and 20 DF,  p-value: < 2.2e-16
```

Testing many parameters

We have the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim \text{ iid Normal}(0, \sigma^2)$$

We seek to test $\quad H_0 : \beta_{r+1} = \cdots = \beta_k = 0.$

In other words, do we really have just:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_r x_{ir} + \epsilon_i, \quad \epsilon_i \sim \text{ iid Normal}(0, \sigma^2)$$

?

What to do...

1. Fit the "full" model (with all k x's).
2. Calculate the residual sum of squares, RSS_{full}.
3. Fit the "reduced" model (with only r x's).
4. Calculate the residual sum of squares, RSS_{red}.
5. Calculate $F = \frac{(RSS_{red} - RSS_{full})/(df_{red} - df_{full})}{RSS_{full}/df_{full}}$.

   where $df_{red} = n - r - 1$ and $df_{full} = n - k - 1$).
6. Under H_0, $F \sim F(df_{red} - df_{full}, df_{full})$.

In particular. . .

Assume the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \quad \epsilon_i \sim \text{ iid Normal}(0, \sigma^2)$$

We seek to test $H_0 : \beta_1 = \cdots = \beta_k = 0$ (i.e., none of the x's are related to y).

$\longrightarrow$ Full model: All the x's

$\longrightarrow$ Reduced model: $y = \beta_0 + \epsilon$ $\quad RSS_{red} = \sum_i (y_i - \bar{y})^2$

$\longrightarrow$ $F = [(\sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{y}_i)^2)/k] \,/\, [\sum_i (y_i - \hat{y}_i)^2/(n-k-1)]$

Compare this to a $F(k, n-k-1)$ dist'n.

The example

To test $\beta_2 = \beta_3 = 0$

```
Analysis of Variance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x1:x2

     Res.Df       RSS   Df  Sum of Sq       F     Pr(>F)
  1      22   0.00975
  2      20   0.00312    2    0.00663   21.22    1.1e-05
```