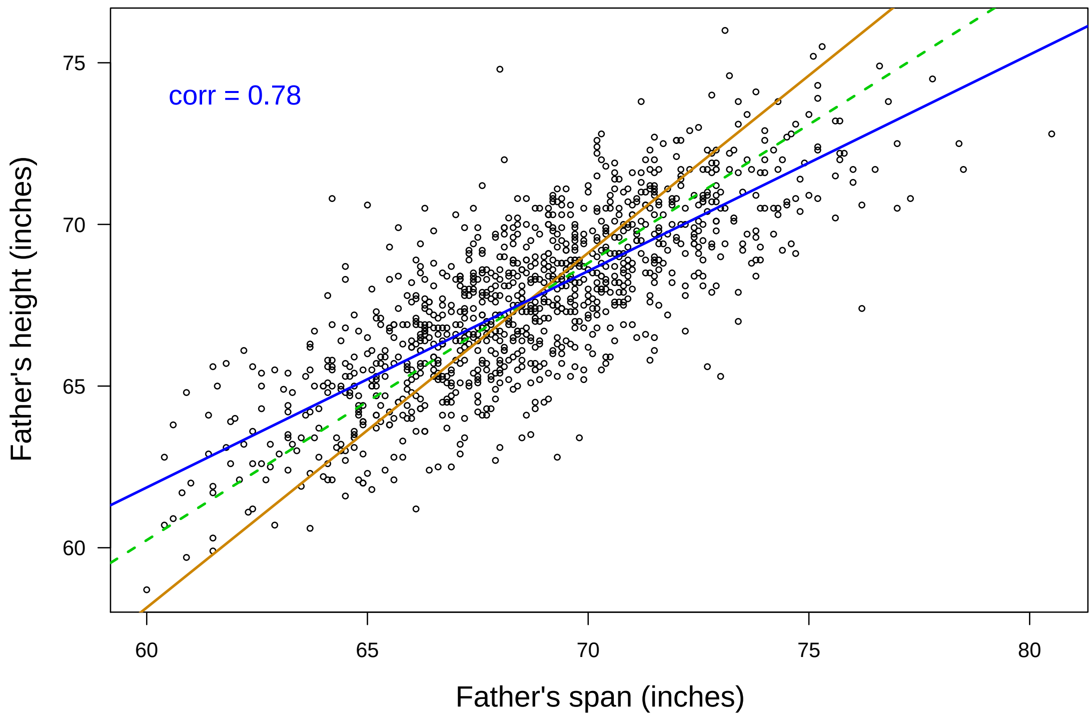
From last time . . .



The equations

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

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

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

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

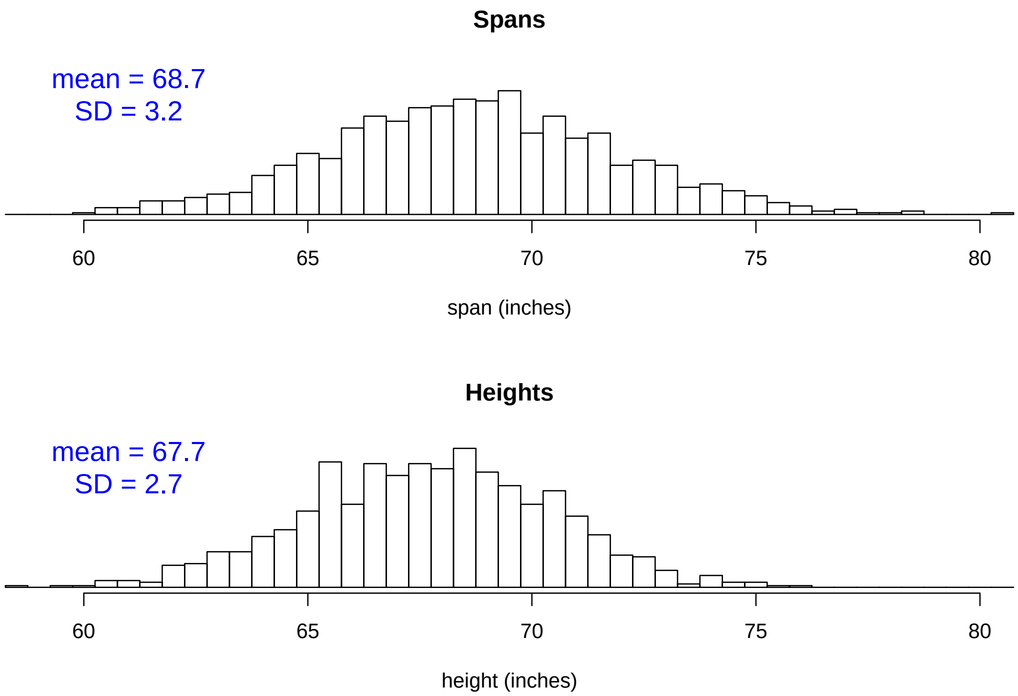
Regression of x on y (for predicting x from y)

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

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

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

Histograms



Error in prediction

Having no information about x,

Predict y as $\bar{y}$

Typical prediction error: SD(y)

For predicting height, SD(y) $\approx$ 2.73

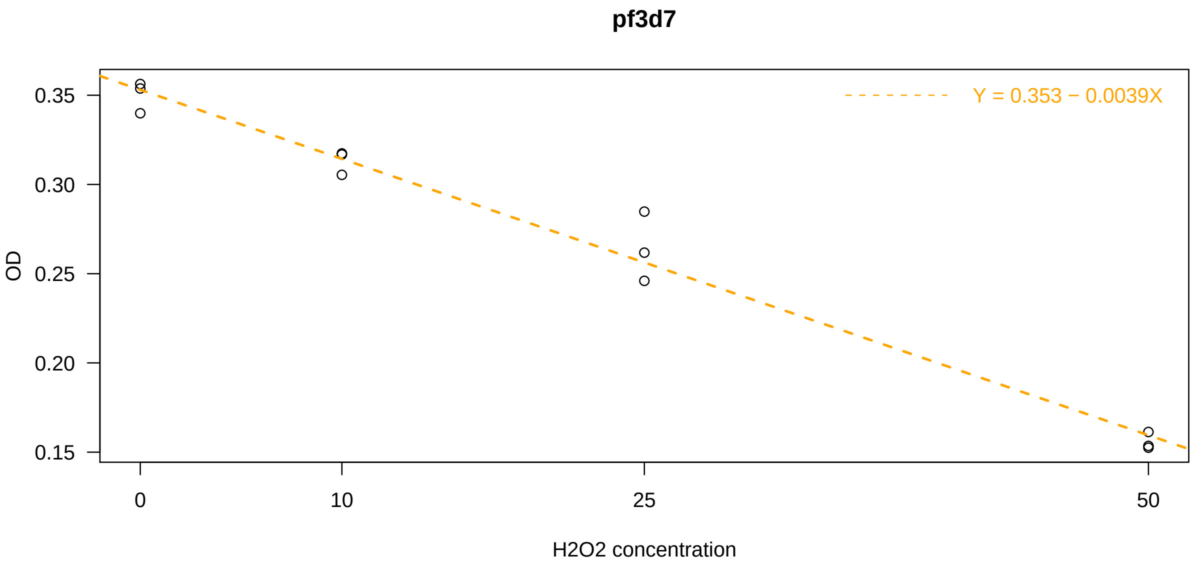
Having been told about x,

Predict y using the regression line: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Typical prediction error: SD(y) $\sqrt{1 - r^2}$

For predicting height from span, SD(y) $\sqrt{1 - r^2} \approx 1.71$

Back to that heme data. . .



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

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

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

Parameter estimates

We already know that

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

and in particular

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

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

Parameter estimates (2)

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.

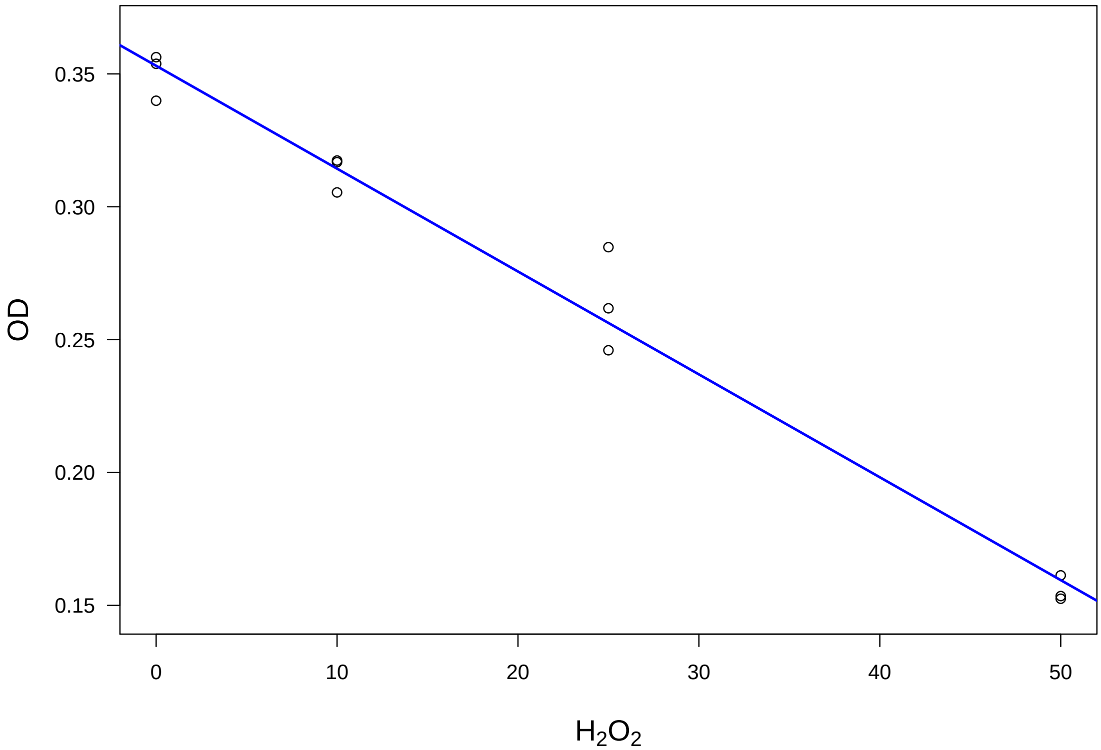
Parameter estimates (3)

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

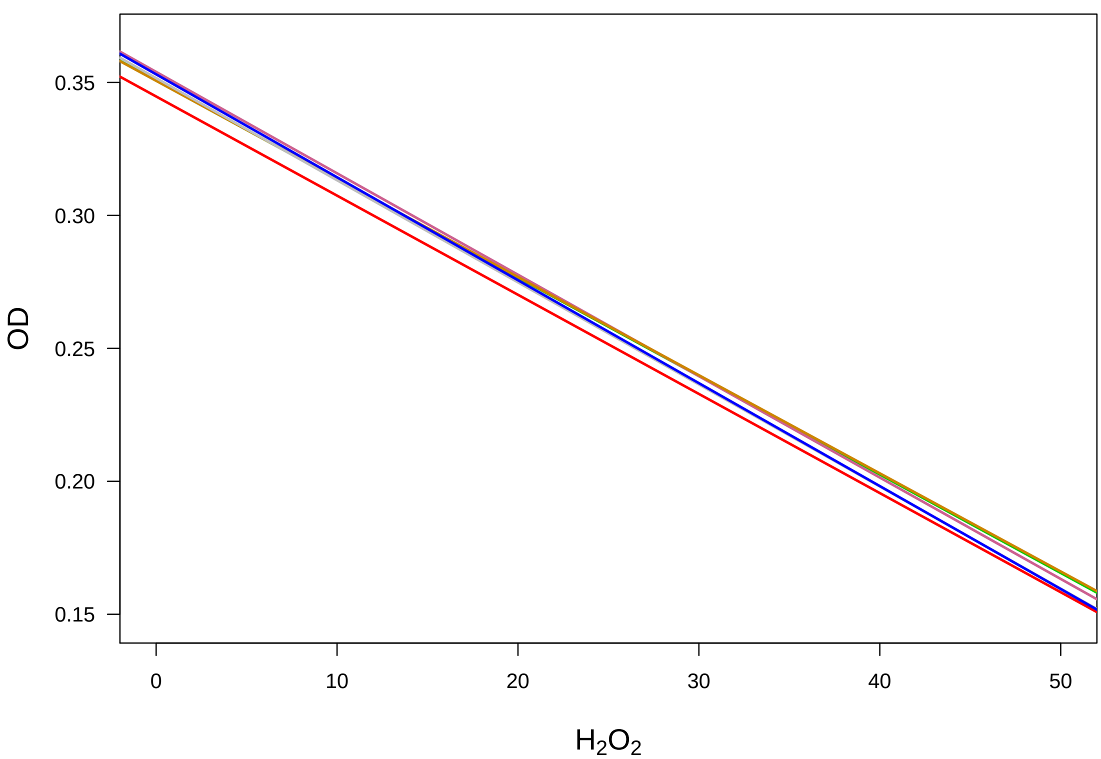
$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} \sim N(\beta, \Sigma)$$

where

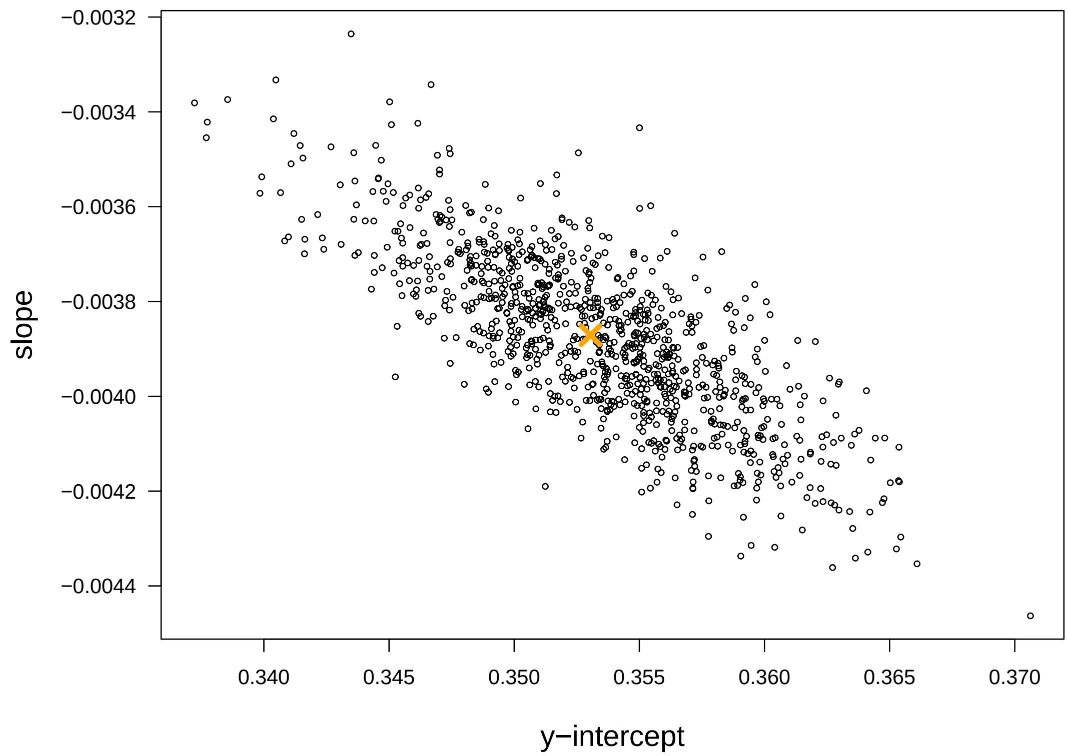
$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \sigma^2 \begin{pmatrix} \frac{1}{n} + \frac{\bar{x}^2}{SXX} & \frac{-\bar{x}}{SXX} \\ \frac{-\bar{x}}{SXX} & \frac{1}{SXX} \end{pmatrix}$$



0.35
0.30
0.25
0.20
0.15
OD
0
10
20
30
40
50
H_2O_2



0.35
0.30
0.25
0.20
0.15
OD
0
10
20
30
40
50
H_2O_2



Confidence intervals

We know that

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

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

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 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

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

For the squared test statistic we get

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

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

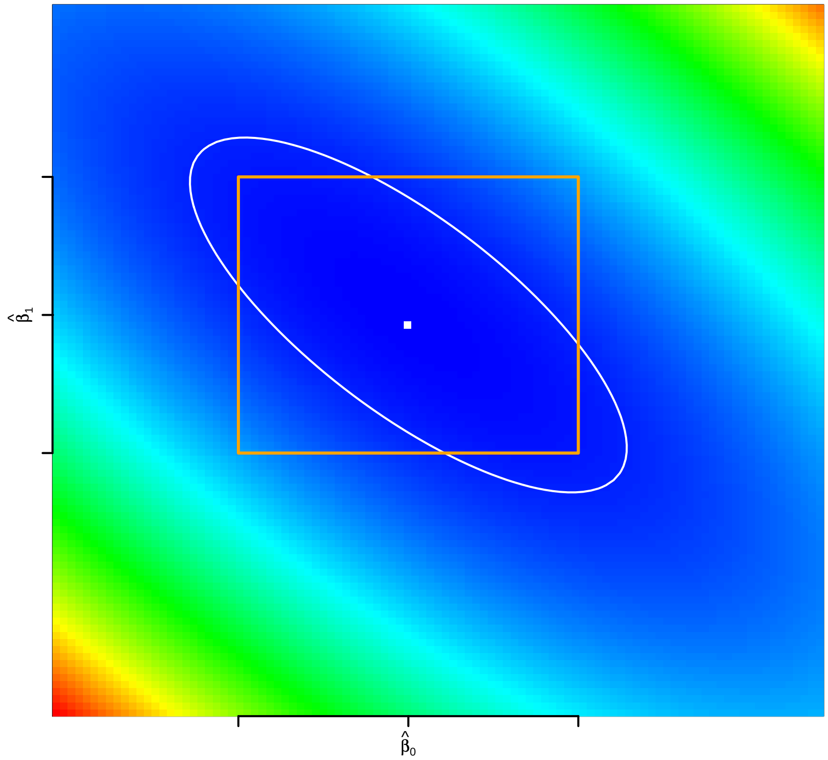
Joint confidence region

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

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

where

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



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

In the previous lecture we wrote

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

Define

$$\text{R}^2 = \frac{\text{SS}_{\text{reg}}}{\text{SYY}} = 1 - \frac{\text{RSS}}{\text{SYY}}$$

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

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

Back to the heme data

The scientist was actually interested in the slopes when one re-scales the y-axis so that the y-intercept is at 1.

$y = \beta_0 + \beta_1 x + \epsilon$ becomes $y/\beta_0 = 1 + (\beta_1/\beta_0)x + \epsilon'$

So we're really interested in β_1/β_0.

We'd estimate that by $\hat{\beta}_1/\hat{\beta}_0$, but what about its standard error?

First-order Taylor expansion

Consider $f(x,y) = x/y$.

A first-order Taylor expansion to approximate the function would be

$$f(x,y) \approx f(x_0,y_0) + (x - x_0)\left.\frac{\partial f}{\partial x}\right|_{(x_0,y_0)} + (y - y_0)\left.\frac{\partial f}{\partial y}\right|_{(x_0,y_0)}$$

Since $\partial f/\partial x = 1/y$ and $\partial f/\partial y = -x/y^2$, we obtain the following:

$$\begin{aligned} x/y &\approx x_0/y_0 + (x - x_0)/y_0 - (y - y_0)x_0/y_0^2 \\ &= (x_0/y_0)[1 + (x - x_0)/x_0 + (y - y_0)/y_0] \end{aligned}$$

How do we use this?

We use the first-order Taylor expansion of $\hat{\beta}_1/\hat{\beta}_0$ around β_1 and β_0.

Variance of a ratio

Remember that β_1 and β_0 are fixed, while $\hat{\beta}_1$ and $\hat{\beta}_0$ are random.

Add the fact that var(X+Y) = var(X) + var(Y) + 2 cov(X,Y)

$$\begin{aligned} \text{var}\{\hat{\beta}_1/\hat{\beta}_0\} &\approx \text{var}\{(\beta_1/\beta_0)[1 + (\hat{\beta}_1 - \beta_1)/\beta_1 + (\hat{\beta}_0 - \beta_0)/\beta_0]\} \\ &= (\beta_1/\beta_0)^2\{\text{var}(\hat{\beta}_1)/\beta_1^2 + \text{var}(\hat{\beta}_0)/\beta_0^2 + 2\,\text{cov}(\hat{\beta}_1,\hat{\beta}_0)/(\beta_1\beta_0)\} \end{aligned}$$

We then replace β_1 and β_0 in this formula with our estimates of them, $\hat{\beta}_1$ and $\hat{\beta}_0$. Further, we replace the variances and covariance with our estimates.

$$\widehat{\text{var}}\{\hat{\beta}_1/\hat{\beta}_0\} = (\hat{\beta}_1/\hat{\beta}_0)^2\{\widehat{\text{var}}(\hat{\beta}_1)/\hat{\beta}_1^2 + \widehat{\text{var}}(\hat{\beta}_0)/\hat{\beta}_0^2 + 2\,\widehat{\text{cov}}(\hat{\beta}_1,\hat{\beta}_0)/(\hat{\beta}_1\hat{\beta}_0)\}$$

The estimated SE is then

$$\widehat{\text{SE}}\{\hat{\beta}_1/\hat{\beta}_0\} = |\hat{\beta}_1/\hat{\beta}_0|\sqrt{[\widehat{\text{SE}}(\hat{\beta}_1)/\hat{\beta}_1]^2 + [\widehat{\text{SE}}(\hat{\beta}_0)/\hat{\beta}_0]^2 + 2\,\widehat{\text{cov}}(\hat{\beta}_1,\hat{\beta}_0)/(\hat{\beta}_1\hat{\beta}_0)}$$

Results

pf3d7:

$\hat{\beta}_0 = 0.353(0.005)$ $\hat{\beta}_1 = -0.0039(0.0002)$ $\hat{\text{cov}}(\hat{\beta}_1, \hat{\beta}_0) = -6.6 \times 10^7$

$\hat{\beta}_1/\hat{\beta}_0 \times 100$ = –1.10 (SE = 0.07).

	estimate	SE
bhem	-2.04	0.32
pgalnoel	-2.02	0.35
pgal	-1.88	0.17
pyoelii	-1.33	0.09
pf3d7	-1.10	0.07
pviv	-0.86	0.26
pknow	-0.79	0.14
pov	-0.70	0.07
pbr	-0.67	0.08
pfhz	-0.31	0.17