Data Analysis in Astronomy and Physics

Lecture 10: Regression

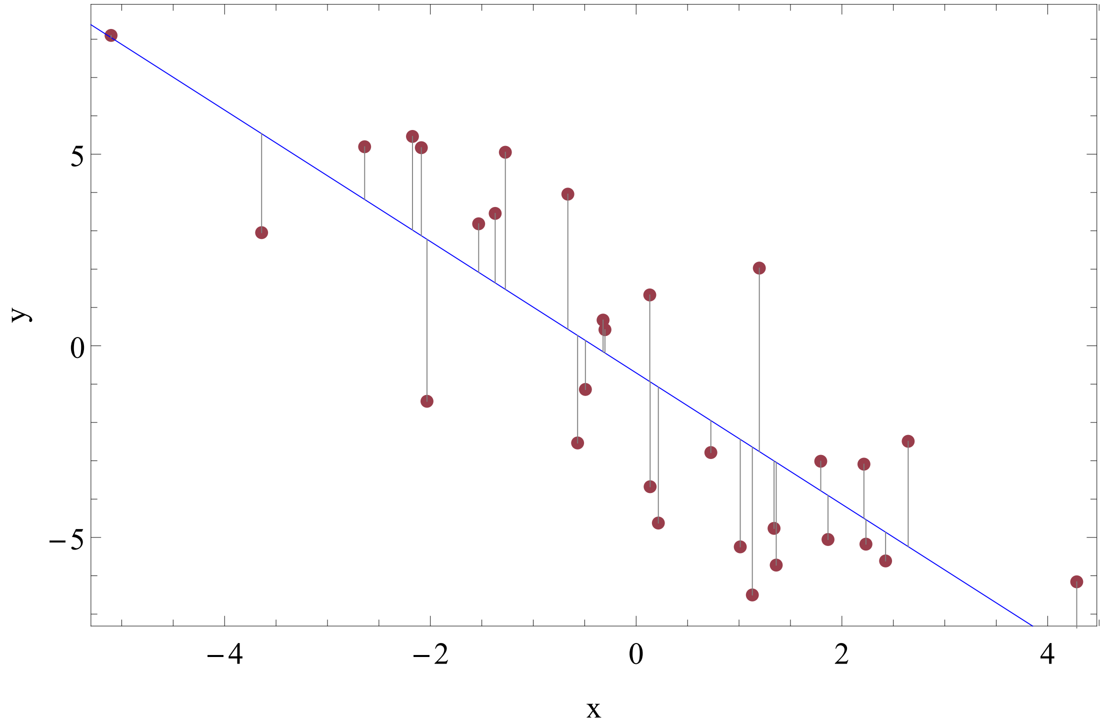
M. Röllig

Residuals

The residual (or fitting deviation) of an observed value is the difference between the observed value and the estimated function value:

- leftovers from the model fit
- data = fit + residual

Out[ ]= residual: $e_i = y_i - \hat{y}_i$

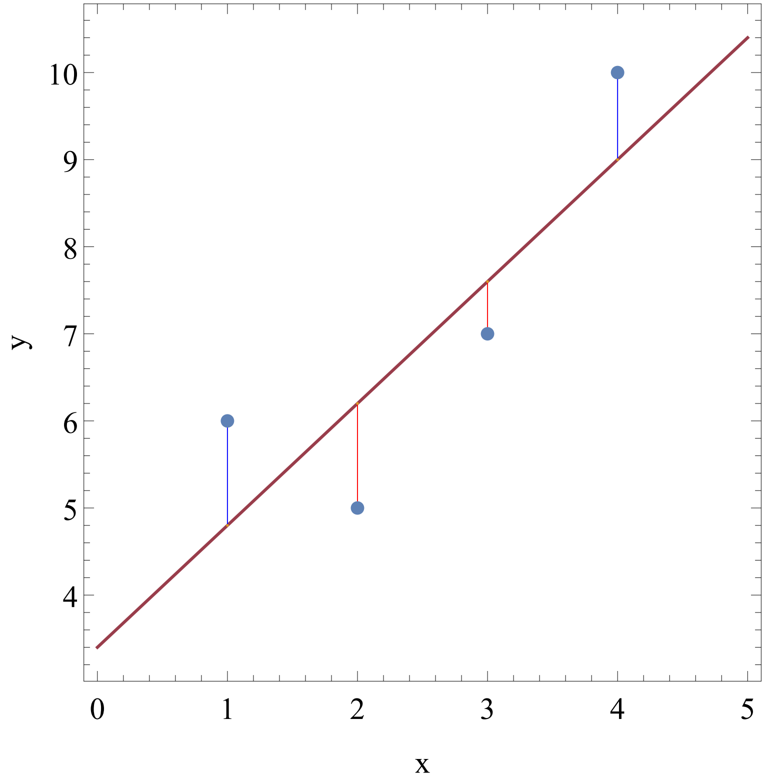


Simple example

As a result of an experiment, four (x, y) data points were obtained, (1, 6), (2, 5), (3, 7), and (4, 10) (shown in red in the picture on the right). We hope to find a line $y = \beta_1 + \beta_2 x$ that best fits these four points. In other words, we would like to find the numbers β_1 and β_2 that approximately solve the **overdetermined** linear system

Simple example

Out[•]=



$\beta_1 + 1\beta_2 = 6$
$\beta_1 + 2\beta_2 = 5$
$\beta_1 + 3\beta_2 = 7$
$\beta_1 + 4\beta_2 = 10$

The "error", at each point, between the curve fit and the data is the difference between the right- and left-hand sides of the equations above. The least squares approach to solving this problem is to try to make as small as possible the sum of the squares of these errors; that is, to find the minimum of the function

$$\begin{aligned} S(\beta_1, \beta_2) &= [6 - (\beta_1 + 1\,\beta_2)]^2 + [5 - (\beta_1 + 2\,\beta_2)]^2 + [7 - (\beta_1 + 3\,\beta_2)]^2 + [10 - (\beta_1 + 4\,\beta_2)]^2 \\ &= 4\,\beta_1^2 + 30\,\beta_2^2 + 20\,\beta_1\,\beta_2 - 56\,\beta_1 - 154\,\beta_2 + 210 \end{aligned}$$

Simple example

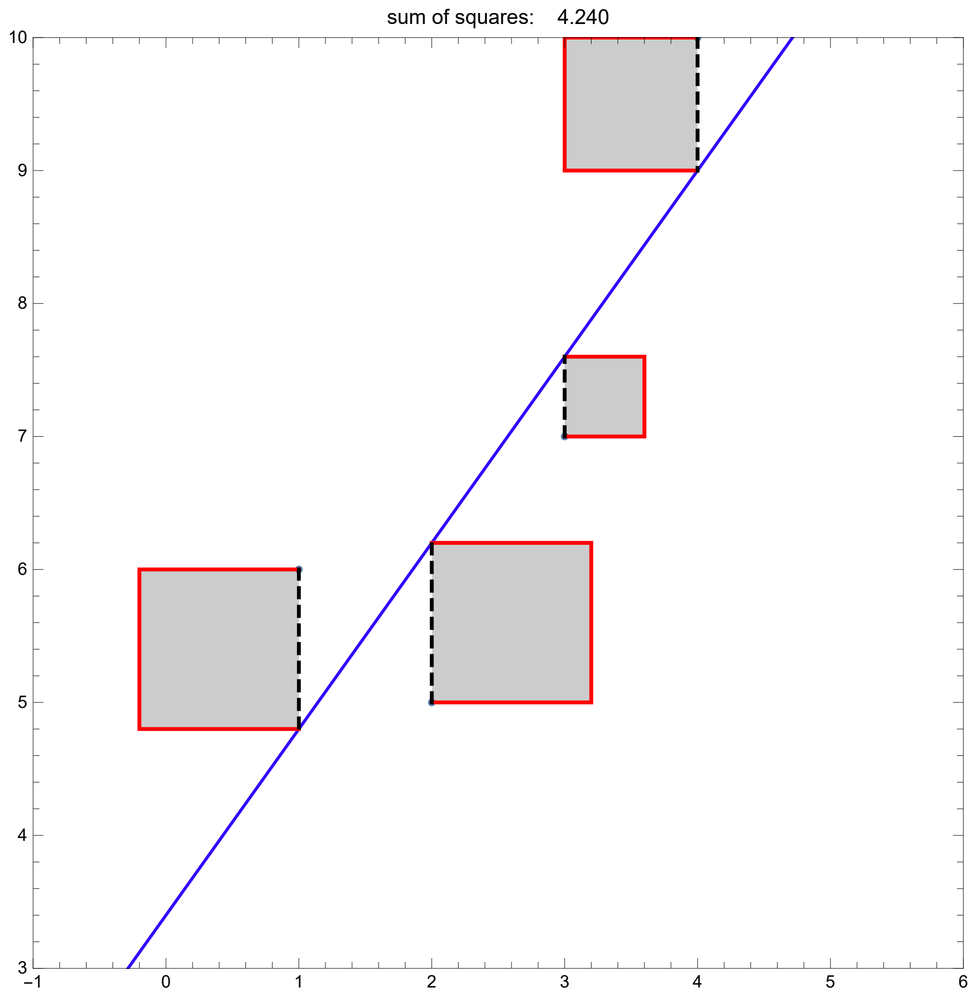
The minimum is determined by calculating the partial derivatives of $S(\beta_1, \beta_2)$ with respect to β_1 and β_2 and setting them to zero

$$\frac{\partial S}{\partial \beta_1} = 0 = 8\,\beta_1 + 20\,\beta_2 - 56$$

$$\frac{\partial S}{\partial \beta_2} = 0 = 20\,\beta_1 + 60\,\beta_2 - 154$$

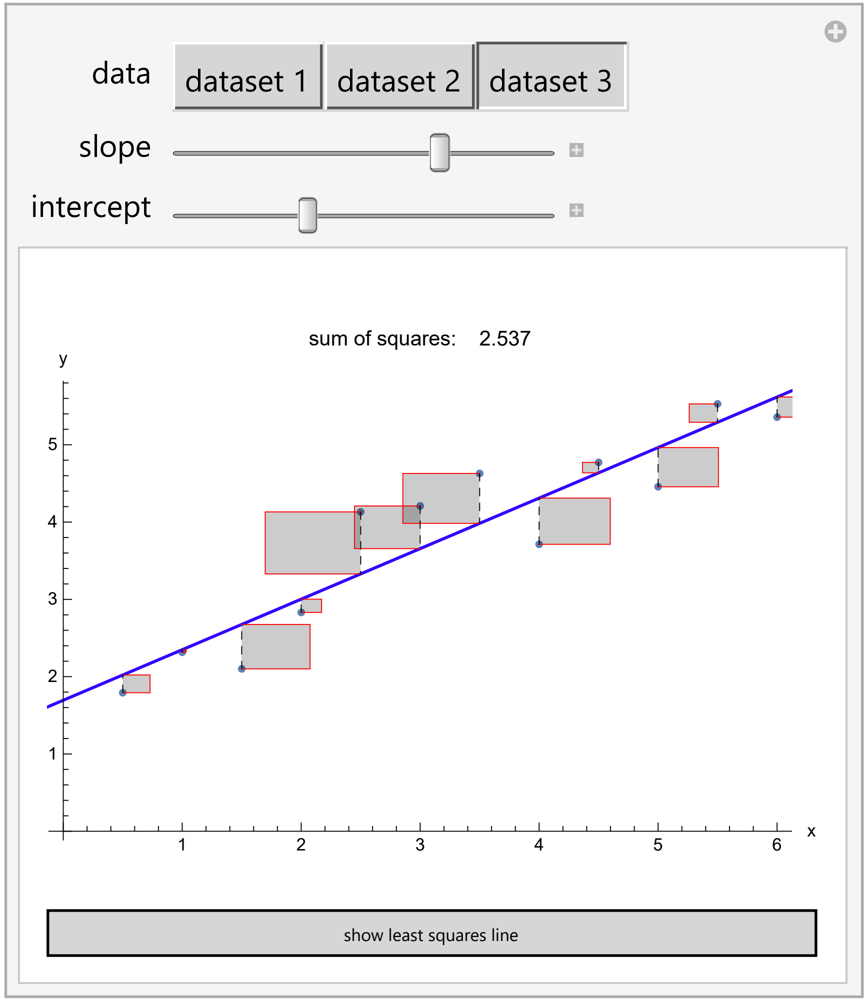
Simple example

This results in a system of two equations in two unknowns, called the **normal equations**, which give, when solved: $\boldsymbol{\beta_1 = 3.4}$ and $\boldsymbol{\beta_2 = 1.4}$ and the equation $\boldsymbol{y = 3.4 + 1.4\,x}$ of the line of the best fit. The residuals, that is, the discrepancies between the *y* values from the experiment and the *y* values calculated using the line of best fit are then found to be 1.1, -1.3, -0.7, and 0.9. The minimum value of the sum of squares of the residuals is $S(3.5,1.4) = 1.1^2 + (-1.3)^2 + (-0.7)^2 + 0.9^2 = 4.2$.



Least squares line

Out[ ]=



Simple example - Using a quadratic model

Importantly, in "linear least squares", we are not restricted to using a line as the model as in the above example. For instance, we could have chosen the restricted quadratic model $\hat{y} = \beta_1 x^2$. This model is still linear in the β_1 parameter, so we can still perform the same analysis, constructing a system of equations from the data points:

$$6 = \beta_1 (1)^2$$
$$5 = \beta_1 (2)^2$$
$$7 = \beta_1 (3)^2$$
$$10 = \beta_1 (4)^2$$

The partial derivatives with respect to the parameters (this time there is only one) are again computed and set to 0:

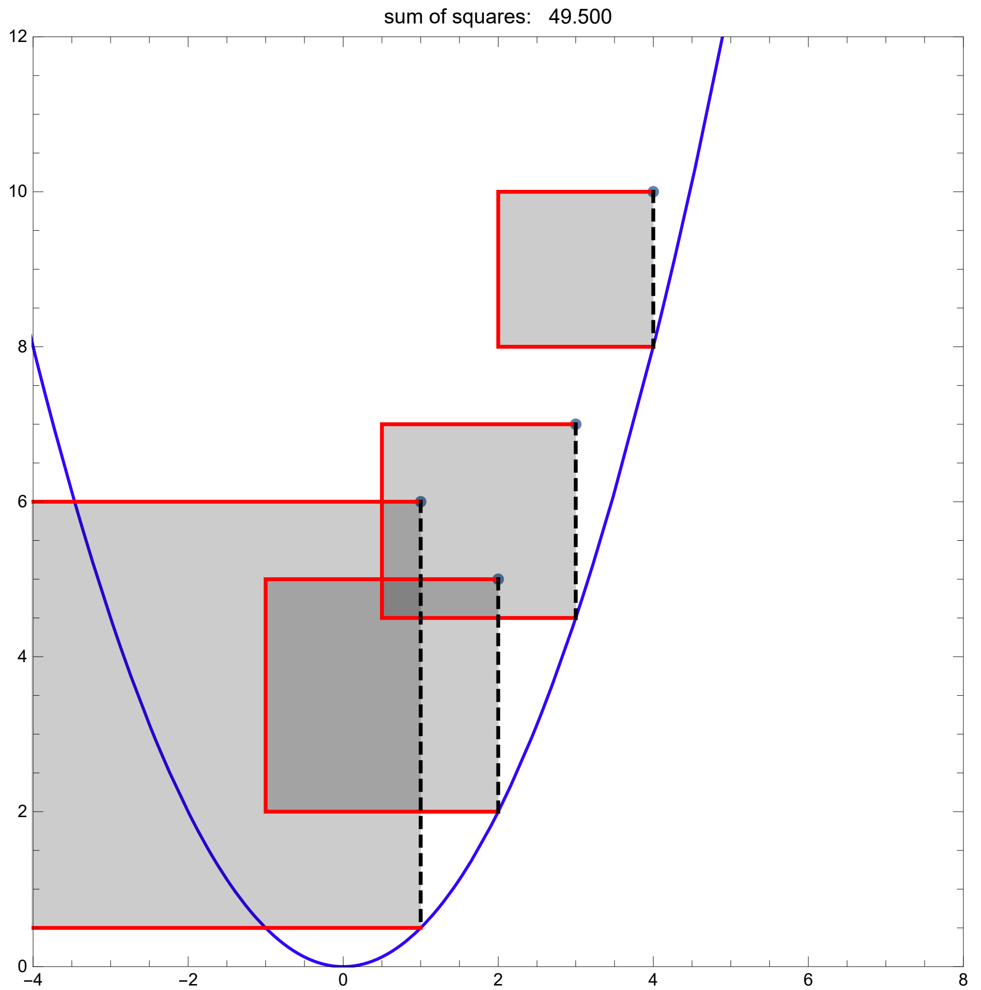
$$\frac{\partial S}{\partial \beta_1} = 0 = 708 \beta_1 - 498$$

Simple example - Using a quadratic model

and solved

$$\beta_1 = 0.703$$

leading to the best fit model $\hat{y} = .703\,x^2$



Linear least squares

Consider an overdetermined system

Out[ ]= $$\sum_{j=0}^{p} X_{ij}\,\beta_j = y_i \qquad (i = 1, 2, \ldots, n)$$

of n linear equations in p unknown coefficients $\beta_0, \beta_1, \ldots, \beta_p$ with $n > p$. This can be written in matrix form as

Out[ ]= $$X\beta = y\ , \qquad \text{where} \qquad X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

X is called design matrix. Such a system usually has no solution, so the goal is instead to find the coefficients β which fit the equations "best," in the sense of solving the quadratic minimization problem

Linear least squares

Out[ ]//TraditionalForm=

$$\hat{\beta} = \underset{\beta}{\arg\min}\, S(\beta)$$

where the objective function S is given by

Out[ ]//TraditionalForm=

$$\boxed{S(\beta) = \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{p} X_{\text{ij}}\, \beta_j \right|^2 = \| y - X\beta \|^2}$$

This minimization problem has a unique solution, provided that the n columns of the matrix X are linearly independent, given by solving the normal equations

Out[ ]//TraditionalForm=

$$\boxed{\left(X^{\mathsf{T}} X\right)\hat{\beta} = X^{\mathsf{T}}\, y}$$

Derivation of the normal equations

Define the i – th residual to be

Out[•]//TraditionalForm=

$$r_i = y_i - \sum_{j=0}^{p} X_{\mathrm{ij}} \beta_j$$

Then S can be rewritten

Out[•]//TraditionalForm=

$$S = \sum_{i=1}^{n} r_i^2$$

Derivation of the normal equations

S is minimized when its gradient vector is zero. (This follows by definition: if the gradient vector is not zero, there is a direction in which we can move to minimize it further.) The elements of the gradient vector are the partial derivatives of S with respect to the parameters:

Out[ ]=
$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_j} \qquad (j = 0,1,2,\ldots,p)$$

The derivatives are

Out[ ]//TraditionalForm=
$$\frac{\partial r_i}{\partial \beta_j} = -X_{ij}$$

Derivation of the normal equations

Substitution of the expressions for the residuals and the derivatives into the gradient equations gives

Out[•]= $$\frac{\partial S}{\partial \beta_j} = 2 \sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} X_{ij} \beta_j \right) (-X_{ij}) \qquad (j = 0, 1, 2, \ldots, p)$$

Derivation of the normal equations

Thus if $\hat{\beta}$ minimizes S, we have

Out[ ]= $$2\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} X_{ij}\,\hat{\beta}_j\right)(-X_{ij}) = 0 \qquad (j = 0,1,2,\ldots,p)$$

After some rearrangement, we obtain the normal equations:

Out[ ]= $$\sum_{i=1}^{n}\sum_{k=0}^{p} X_{ij}\,X_{ik}\,\hat{\beta}_k = \sum_{i=1}^{n} X_{ij}\,y_i \qquad (j = 0,1,2,\ldots,p)$$

The normal equations are written in matrix notation as

Out[ ]//TraditionalForm=

$$\boxed{\left(X^{\mathsf{T}} X\right)\widehat{\beta} = X^{\mathsf{T}} y}$$

where X^{T} is the matrix transpose of X. The solution of the normal equations yields the vector $\hat{\beta}$ of the optimal parameter values.

Example: $y = \beta_0 + \beta_1 x$

$$\hat{y}_i = \sum_{j=0}^{1} X_{ij}\,\beta_j = 1\,\beta_0 + X_{i1}\,\beta_1, \qquad \text{with } X_{i0} = 1 \text{ and } X_{i1} = x_i \text{ for all } i$$

$$r_i = y_i - \sum_{j=0}^{1} X_{ij}\,\beta_j = y_i - (1\,\beta_0 + X_{i1}\,\beta_1)$$

$$S = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} (y_i - (1\,\beta_0 + X_{i1}\,\beta_1))^2$$

Example: $y = \beta_0 + \beta_1 x$

We then write the normal equations

Out[ ]= $$\sum_{i=1}^{n} \sum_{k=0}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i \qquad (j = 0,1,2,\ldots,p)$$

$$j = 0: \sum_{i=1}^{n} \sum_{k=0}^{1} X_{i0} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i0} y_i$$

Example: $y = \beta_0 + \beta_1 x$

We then write the normal equations

Out[•]= $$\sum_{i=1}^{n} \sum_{k=0}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i \qquad (j = 0,1,2,\ldots,p)$$

$$j = 0 : \sum_{i=1}^{n} \sum_{k=0}^{1} X_{i0} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i0} y_i$$

$$: \sum_{i=1}^{n} 1 \left(1 \hat{\beta}_0 + X_{i1} \hat{\beta}_1\right) = \sum_{i=1}^{n} 1 y_i$$

Example: $y = \beta_0 + \beta_1 x$

We then write the normal equations

Out[ ]=
$$\sum_{i=1}^{n} \sum_{k=0}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i \qquad (j = 0,1,2,\ldots,p)$$

$$j = 0 : \sum_{i=1}^{n} \sum_{k=0}^{1} X_{i0} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i0} y_i$$

$$: \sum_{i=1}^{n} 1 \left(1 \hat{\beta}_0 + X_{i1} \hat{\beta}_1\right) = \sum_{i=1}^{n} 1 y_i$$

$$: n \hat{\beta}_0 + \sum_{i=1}^{n} x_i \hat{\beta}_1 = \sum_{i=1}^{n} y_i$$

Example: $y = \beta_0 + \beta_1 x$

We then write the normal equations

Out[ ]= $$\sum_{i=1}^{n}\sum_{k=0}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i \qquad (j = 0,1,2,\ldots,p)$$

$$j = 0 : \sum_{i=1}^{n}\sum_{k=0}^{1} X_{i0} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i0} y_i$$

$$: \sum_{i=1}^{n} 1 \left(1 \hat{\beta}_0 + X_{i1} \hat{\beta}_1\right) = \sum_{i=1}^{n} 1\, y_i$$

$$: n \hat{\beta}_0 + \sum_{i=1}^{n} x_i \hat{\beta}_1 = \sum_{i=1}^{n} y_i$$

$$: \hat{\beta}_0 + \frac{1}{n} \sum_{i=1}^{n} x_i \hat{\beta}_1 = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Example: $y = \beta_0 + \beta_1 x$

We then write the normal equations

Out[ ]= $$\sum_{i=1}^{n}\sum_{k=0}^{p} X_{ij}\, X_{ik}\, \hat{\beta}_k = \sum_{i=1}^{n} X_{ij}\, y_i \qquad (j = 0,1,2,\ldots,p)$$

$$j = 0 : \sum_{i=1}^{n}\sum_{k=0}^{1} X_{i0}\, X_{ik}\, \hat{\beta}_k = \sum_{i=1}^{n} X_{i0}\, y_i$$

$$: \sum_{i=1}^{n} 1\,\left(1\,\hat{\beta}_0 + X_{i1}\,\hat{\beta}_1\right) = \sum_{i=1}^{n} 1\, y_i$$

$$: n\,\hat{\beta}_0 + \sum_{i=1}^{n} x_i\,\hat{\beta}_1 = \sum_{i=1}^{n} y_i$$

$$: \hat{\beta}_0 + \frac{1}{n}\sum_{i=1}^{n} x_i\,\hat{\beta}_1 = \frac{1}{n}\sum_{i=1}^{n} y_i$$

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

Example: $y = \beta_0 + \beta_1 x$

$$j = 1 : \sum_{i=1}^{n} \sum_{k=0}^{1} X_{i1} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i1} y_i$$

Example: $y = \beta_0 + \beta_1 x$

$$j = 1 : \sum_{i=1}^{n} \sum_{k=0}^{1} X_{i1} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i1} y_i$$

$$: \sum_{i=1}^{n} X_{i1} \left(1 \hat{\beta}_0 + X_{i1} \hat{\beta}_1\right) = \sum_{i=1}^{n} X_{i1} y_i$$

Example: $y = \beta_0 + \beta_1 x$

$$j = 1 : \sum_{i=1}^{n} \sum_{k=0}^{1} X_{i1} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{i1} y_i$$

$$: \sum_{i=1}^{n} X_{i1} \left(1 \hat{\beta}_0 + X_{i1} \hat{\beta}_1\right) = \sum_{i=1}^{n} X_{i1} y_i$$

$$: \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Example: $y = \beta_0 + \beta_1 x$

Adding the result from $j = 0$:

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

Example: $y = \beta_0 + \beta_1 x$

We have

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

$$n\,\hat{\beta}_0 \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

Example: $y = \beta_0 + \beta_1 x$

Inserting the first into the second gives

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

$$n\,\hat{\beta}_0 \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

$$\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\,\hat{\beta}_1\right) \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

Example: $y = \beta_0 + \beta_1 x$

This gives

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

$$n\,\hat{\beta}_0 \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

$$\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\,\hat{\beta}_1\right) \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

$$\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \left(\sum_{i=1}^{n} x_i\right)^2 \hat{\beta}_1 + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

Example: $y = \beta_0 + \beta_1 x$

This gives

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

$$n\,\hat{\beta}_0 \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

$$\left(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i\,\hat{\beta}_1\right) \sum_{i=1}^{n} x_i + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

$$\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i - \left(\sum_{i=1}^{n} x_i\right)^2 \hat{\beta}_1 + n\,\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \sum_{i=1}^{n} x_i\, y_i$$

$$-\hat{\beta}_1 \left(\left(\sum_{i=1}^{n} x_i\right)^2 - n \sum_{i=1}^{n} x_i^2\right) = n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i$$

Example: $y = \beta_0 + \beta_1 x$

$$\hat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n} \frac{n}{n} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \frac{n}{n} \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - n \overline{y} \overline{x}}{\sum_{i=1}^{n} x_i^2 - n \left(\overline{x}\right)^2}$$

Example: $y = \beta_0 + \beta_1 x$

Out[ ]=
$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} = \frac{s_{xy}}{s_x^2}}$$

and

$$\boxed{\hat{\beta}_0 = \overline{y} - \overline{x}\,\hat{\beta}_1}$$

Remember that:

Out[ ]//TraditionalForm=
$$\boxed{r = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}} = \frac{\operatorname{cov}(x, y)}{s_x\, s_y}}$$

Example: $y = \beta_0 + \beta_1 x$

It follows that

Out[ ]=
$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \left(y_i - \overline{y}\right)}{\sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2} = \frac{\mathrm{cov}(x, y)}{s_x^2} = \frac{s_x\, s_y}{s_x^2} r = \frac{s_y}{s_x} r}$$

This gives us a simple equation to compute the linear regression in case of a least squares line:

Out[ ]= slope: $\boxed{\hat{\beta}_1 = \frac{s_y}{s_x} r}$ and intercept: $\boxed{\hat{\beta}_0 = \overline{y} - \overline{x}\, \hat{\beta}_1}$

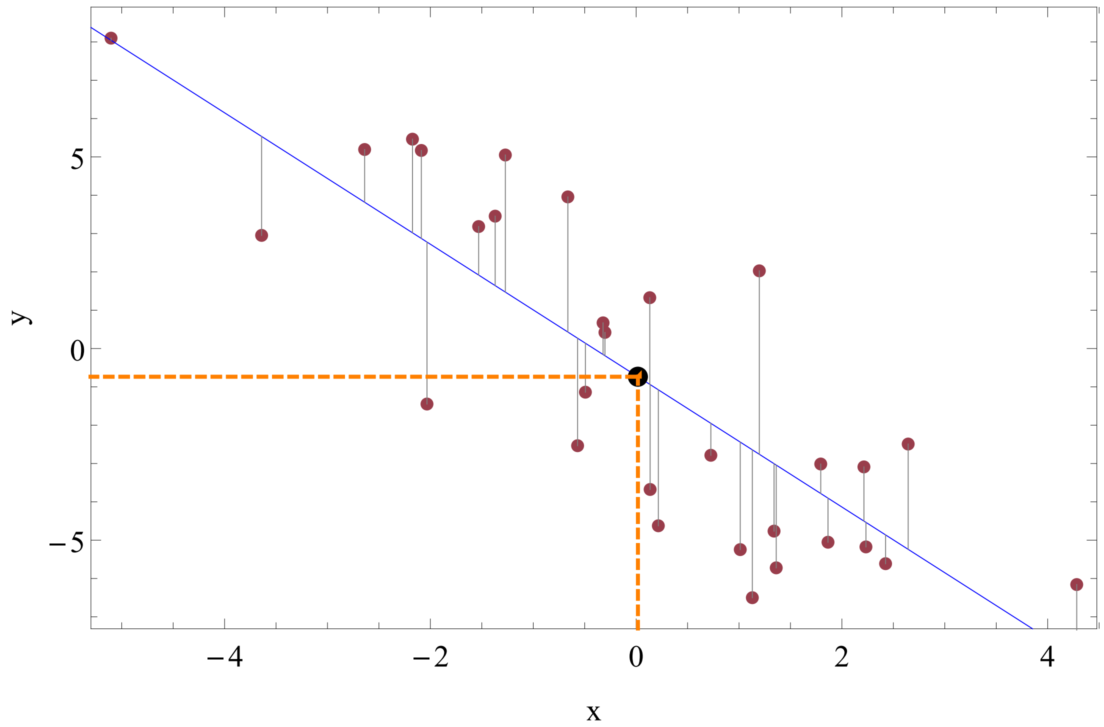
We only need to compute mean values $\overline{x}$ and $\overline{y}$, standard deviations s_x and s_y, and the correlation coefficient r.

Linear regression line: slope

Out[∘]= slope: $\boxed{b_1 = \frac{s_y}{s_x} R}$

s_x: SD of x
s_y: SD of y
R=cor(x,y)

Linear regression line: intercept

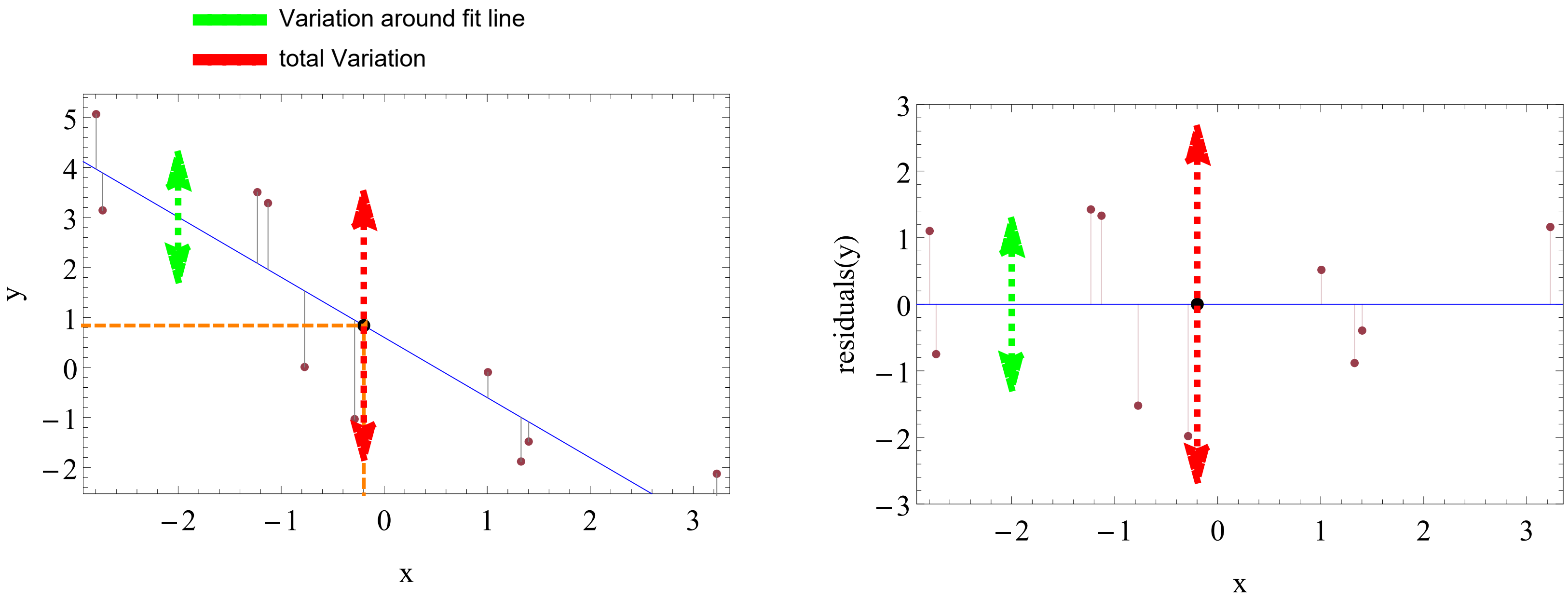


Recall from above that the least squares line always goes through $(\overline{x}, \overline{y})$, so instead of $y = b_0 + b_1 x$ we can write:

Out[∘]= intercept: $\boxed{b_0 = \overline{y} - b_1 \overline{x}}$

Strength of a fit - R^2 (R squared)

Out[ ]=

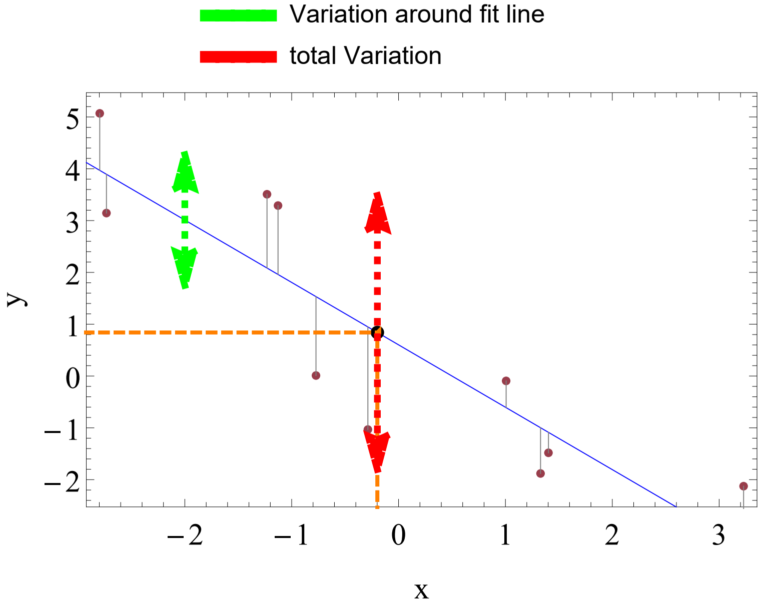


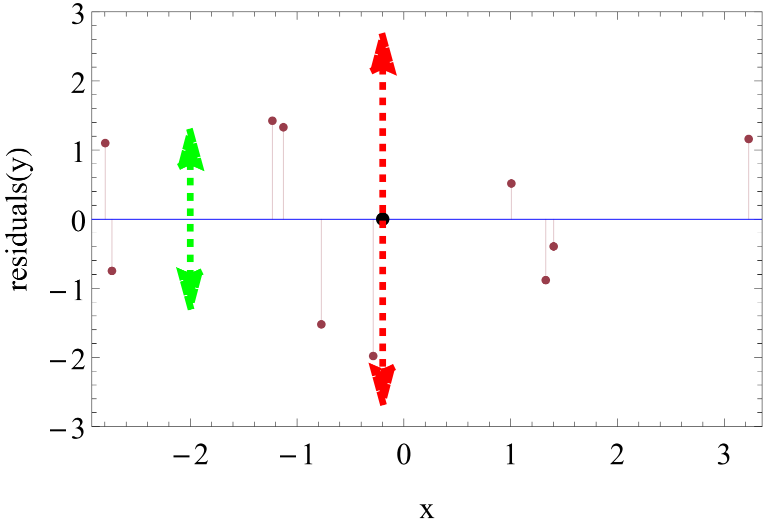
The least squares line fitting splits the data variation into two components:

- the variation of the residuals (not explained by the model)
  - the variation of the predicted values (predicted by the least squares line) and the mean value of the response variable (explained by the model).

Strength of a fit - R^2 (R squared)

Out[ ]=





Out[ ]=

$$\boxed{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}$$

with: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_p X_{ip}$

Strength of a fit - R^2 (R squared)

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y}\right)^2$$

$$= \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 + 2\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)\left(\hat{Y}_i - \overline{Y}\right) + \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2$$

Strength of a fit - R^2 (R squared)

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y}\right)^2$$

$$= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 + 2 \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) \left(\hat{Y}_i - \overline{Y}\right) + \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

if the residuals are $U_i = Y_i - \hat{Y}_i$, then

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right) \left(\hat{Y}_i - \overline{Y}\right) = \sum_{i=1}^{n} U_i \left(\hat{Y}_i - \overline{Y}\right) = \sum_{i=1}^{n} U_i \hat{Y}_i - \overline{Y} \sum_{i=1}^{n} U_i = \sum_{i=1}^{n} U_i \hat{Y}_i - \overline{Y}\, 0$$

i.e. the empirical mean value of the residuals is zero.

Strength of a fit - R^2 (R squared)

$$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i+\hat{Y}_i-\overline{Y}\right)^2$$

$$=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+2\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)\left(\hat{Y}_i-\overline{Y}\right)+\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2$$

if the residuals are $U_i = Y_i - \hat{Y}_i$, then

$$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)\left(\hat{Y}_i-\overline{Y}\right)=\sum_{i=1}^{n}U_i\left(\hat{Y}_i-\overline{Y}\right)=\sum_{i=1}^{n}U_i\,\hat{Y}_i-\overline{Y}\sum_{i=1}^{n}U_i=\sum_{i=1}^{n}U_i\,\hat{Y}_i-\overline{Y}\,0$$

$$\sum_{i=1}^{n}U_i\,\hat{Y}_i=\hat{\beta}_0\sum_{i=1}^{n}U_i+\hat{\beta}_1\sum_{i=1}^{n}U_i\,X_{i1}+\ldots+\hat{\beta}_p\sum_{i=1}^{n}U_i\,X_{1\,p}=\hat{\beta}_0\,0+\hat{\beta}_1\,0+\ldots+\hat{\beta}_p\,0=0$$

i.e. the estimated values $\hat{Y}_i$ and the residuals U_i are uncorrelated.

Strength of a fit - R^2 (R squared)

Remember, that in the derivation of the least squares model we saw that:

$$2\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} X_{ij}\,\hat{\beta}_j\right)(-X_{ij}) = 0$$

$$j = 0: \qquad -2\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} X_{ij}\,\hat{\beta}_j\right) = -2\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right) \Rightarrow \sum_{i=1}^{n} U_i = 0$$

Strength of a fit - R^2 (R squared)

Remember, that in the derivation of the least squares model we saw that:

$$2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{p}X_{ij}\,\hat{\beta}_j\right)\,(-X_{ij})=0$$

$$j=0: \qquad -2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{p}X_{ij}\,\hat{\beta}_j\right)=-2\sum_{i=1}^{n}\left(y_i-\hat{Y}_i\right)\Rightarrow\sum_{i=1}^{n}U_i=0$$

$$j>0: \qquad -2\sum_{i=1}^{n}\left(y_i-\sum_{j=0}^{p}X_{ij}\,\hat{\beta}_j\right)(-X_{ij})=-2\sum_{i=1}^{n}\left(y_i-\hat{Y}_i\right)\,(-X_{ij})\Rightarrow\sum_{i=1}^{n}U_i\,X_{ij}=0$$

Strength of a fit - R^2 (R squared)

We now define the coefficient of determination: R^2 or R squared

$$\boxed{SS_{tot} = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}$$ The total sum of squares (proportional to the variance of the data)

Out[∘]=

$$\boxed{SS_{reg} = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2}$$ The regression sum of squares, also called the explained sum of squares

$$\boxed{SS_{res} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2}$$ The sum of squares of residuals, also called the residual sum of squares

The most general definition of the coefficient of determination is

Out[∘]//TraditionalForm=

$$\boxed{R^2 = 1 - \frac{SS_{res}}{SS_{tot}}}$$

Strength of a fit - R^2 (R squared)

In case of just one dependent variable, R^2 equals the square of Pearsson's correlation coefficient:

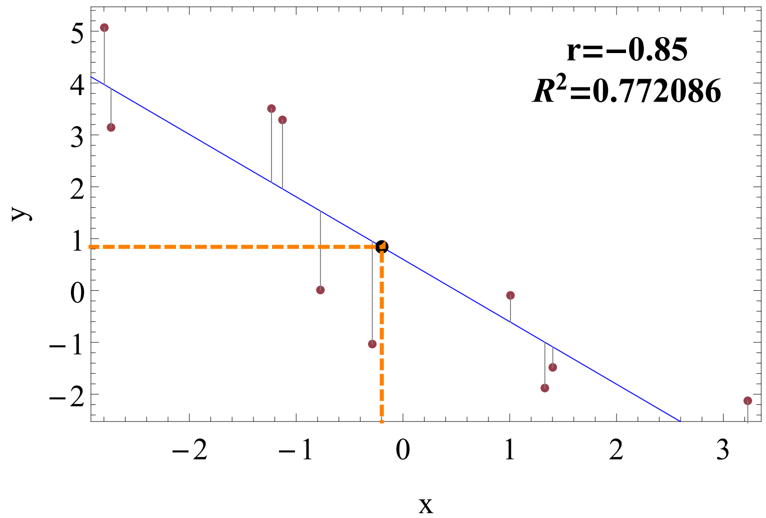
Out[ ]//TraditionalForm=

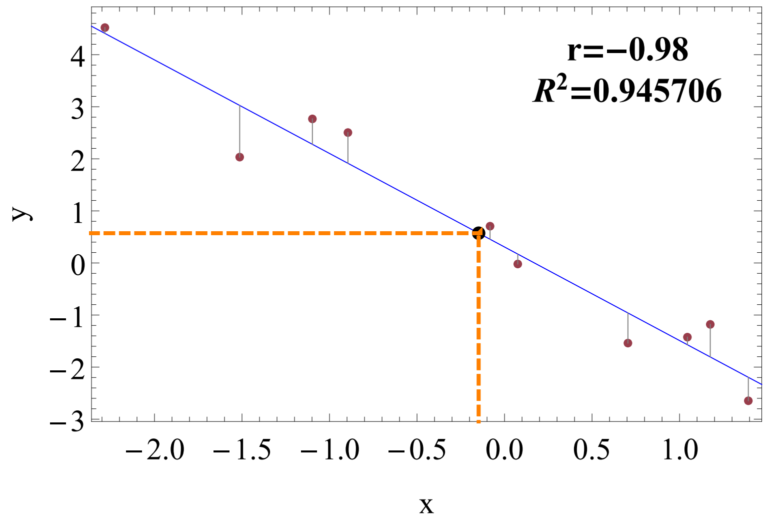
$$\boxed{R^2 = r_{\text{xy}}^2 = \frac{s_{\text{xy}}^2}{s_x^2\, s_y^2}}$$

The R^2 of a linear model describes the amount of variation in the response that is explained by the least squares line.

Strength of a fit - R^2 (R squared)

Out[ ]=



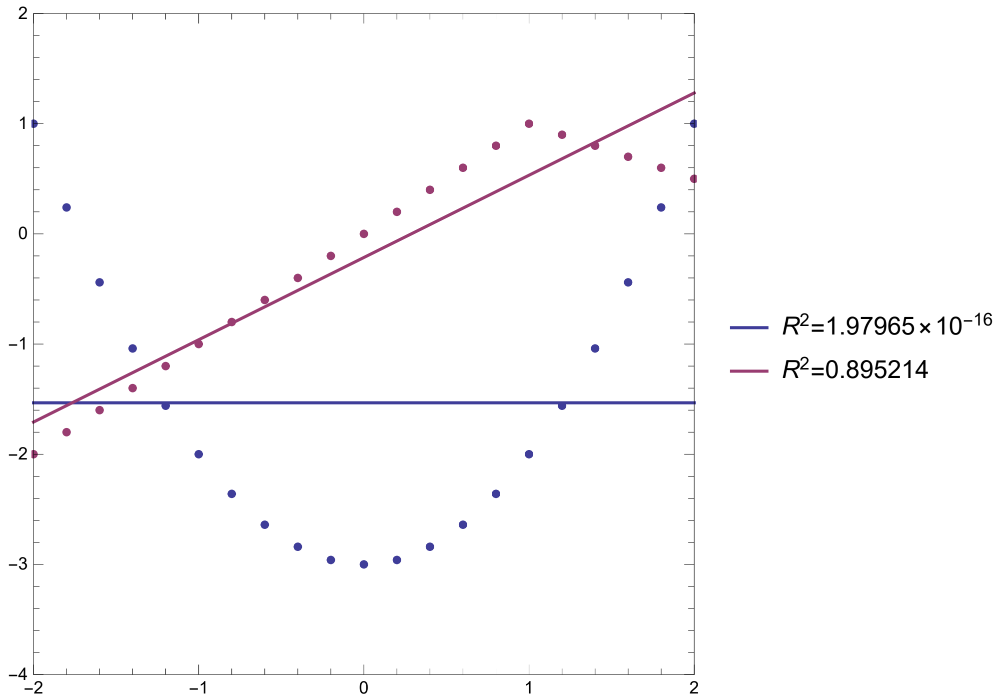


In the left figure, 77.2% of the data’s variability is explained by the model, i.e. the least squares line. In the right figure, the least squares fit accounts for 94.6% of the total data variability.

R^2 is a statistic that will give some information about the goodness of fit of a model. In regression, the R^2 coefficient of determination is a statistical measure of how well the regression line approximates the real data points. An R^2 of 1 indicates that the regression line perfectly fits the data.

R^2 - Limits

The coefficient of determination indicates the quality of the linear approximation, but **not whether the the linear approximation is a suitable model**!



R^2 - Limits

Common misconceptions:

- A high R^2 allows reliable predictions (The trend switch in the red data above is not covered by the model).
- A high R^2 indicates that the model is a good approximation of the data (The red data shows differently.).
- A $R^2 \approx 0$ indicates that there is no dependence between the explanatory and the dependent variable (The blue data/line above shows differently).

Example: $y = \beta_0 + \beta_1 x + \beta_2 x^2$

This gives the following normal equations

Out[ ]//TraditionalForm=

$$5\beta0 + 10\beta2 = -5$$
$$10\beta1 = 0$$
$$10\beta0 + 34\beta2 = 4$$

In[ ]:=
```
MatrixForm[X]
```

Out[ ]//MatrixForm=

$$\begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}$$

with the solution:

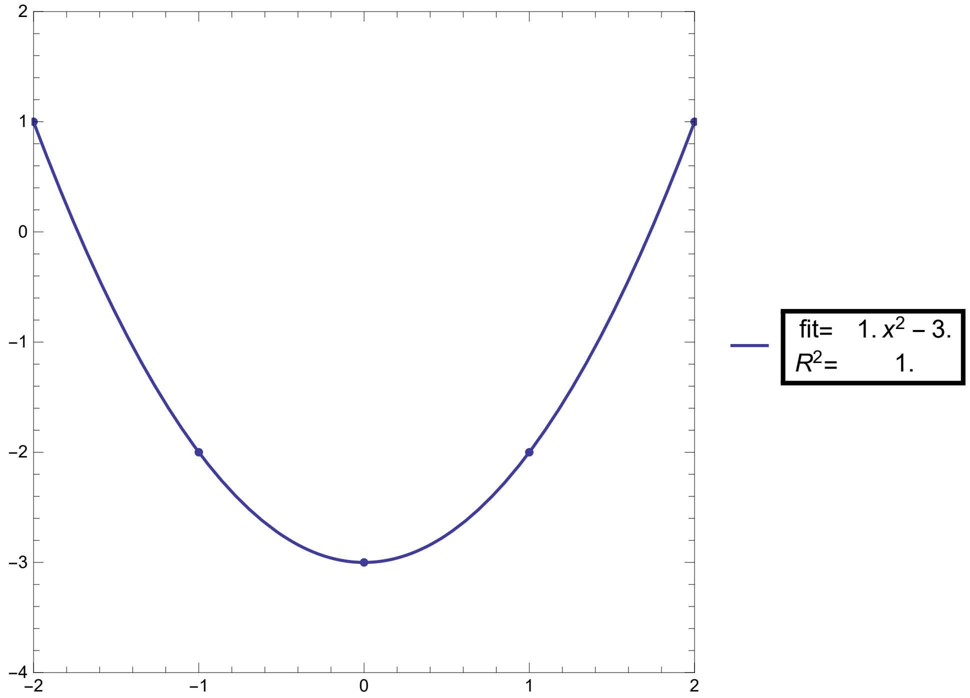
In[ ]:=
```
Solve[Thread[Flatten[(Transpose[X].X).β̂, 1] == Transpose[X].y], {β0, β1, β2}]
```

Out[ ]=
```
{{β0 → -3, β1 → 0, β2 → 1}}
```

Example: $y = \beta_0 + \beta_1 x + \beta_2 x^2$

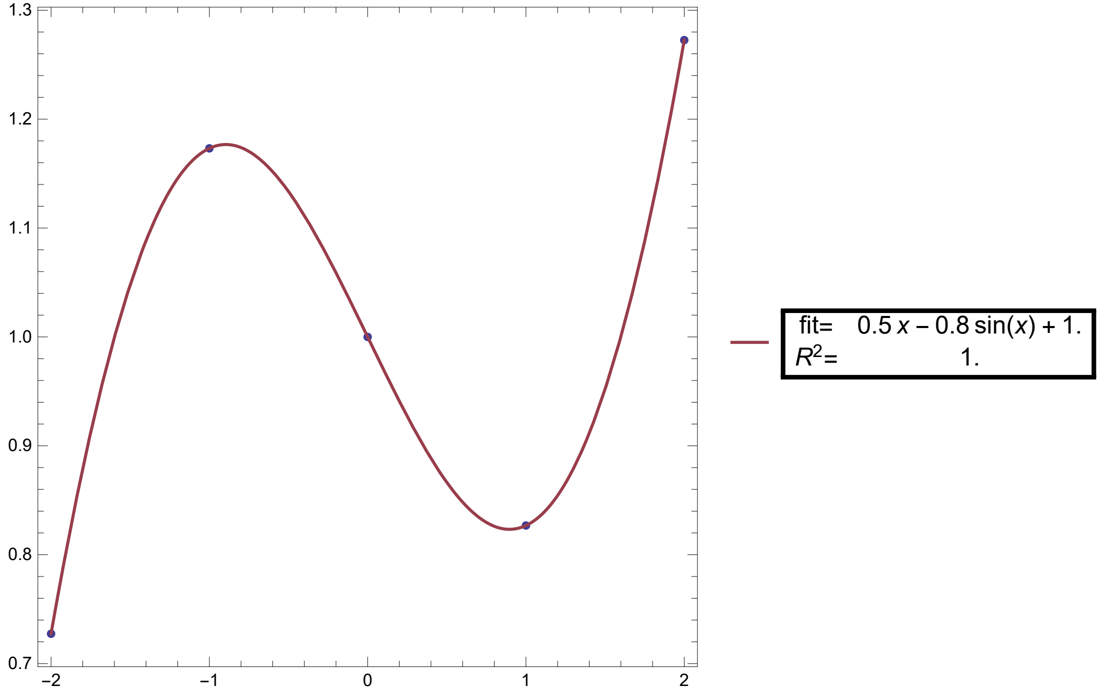
with the solution:

Out[ ]= {{β0 → -3, β1 → 0, β2 → 1}}



Example: $y = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

The model equations need to be linear in β not in x!



Out[ ]//TraditionalForm=

$$\boxed{\left(X^{\mathsf{T}} X\right)\hat{\beta} = X^{\mathsf{T}} y}$$

Example: $y = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

the data is {{-2, 0.73}, {-1, 1.17}, {0, 1.},{1, 0.83}, {2, 1.27}},

$$X = \begin{pmatrix} 1 & x_1 & \text{Sin}[x_1] \\ 1 & x_2 & \text{Sin}[x_2] \\ 1 & x_3 & \text{Sin}[x_3] \\ 1 & x_4 & \text{Sin}[x_4] \\ 1 & x_5 & \text{Sin}[x_5] \end{pmatrix} = \begin{pmatrix} 1 & -2 & -0.909297 \\ 1 & -1 & -0.841471 \\ 1 & 0 & 0 \\ 1 & 1 & 0.841471 \\ 1 & 2 & 0.909297 \end{pmatrix}, X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ -0.91 & -0.84 & 0 & 0.84 & 0.91 \end{pmatrix}, \hat{\beta} = \begin{pmatrix} \beta0 \\ \beta1 \\ \beta2 \end{pmatrix}, y = \begin{pmatrix} 0.73 \\ 1.17 \\ 1. \\ 0.83 \\ 1.27 \end{pmatrix}, x = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

Example: $y = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

This gives the following normal equations

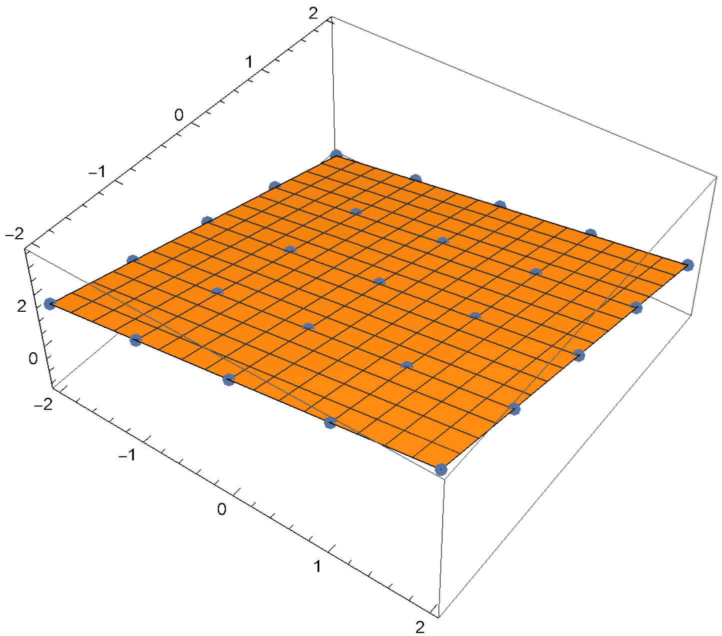
Out[ ]//TraditionalForm=

$$5\,\beta 0 = 5.$$
$$10\,\beta 1 + \beta 2\,(2\sin(1) + 4\sin(2)) = 0.74$$
$$\beta 1\,(2\sin(1) + 4\sin(2)) + \beta 2\left(2\sin^2(1) + 2\sin^2(2)\right) = 0.20492$$

with the solution:

Out[ ]= {{β0 → 1., β1 → 0.49348, β2 → -0.788476}}

Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$



fit=	$0.5\,x - 0.8\,y + 1.$
R^2=	1.

Out[ ]//TraditionalForm=

$$\boxed{(X^\top X)\hat{\beta} = X^\top y}$$

Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

the data is {{-2, -2, 1.6}, {-2, -1, 0.8}, {-2, 0, 0.}, {-2, 1, -0.8},…, {2, 2, 0.3999999999999999}},

$$X = \begin{pmatrix} 1 & x_{11} & x_{22} \\ 1 & x_{12} & x_{22} \\ \ldots & \ldots & \ldots \\ 1 & x_{15} & x_{22} \end{pmatrix} = \begin{pmatrix} 1 & -2 & -2 \\ 1 & -2 & -1 \\ \ldots & \ldots & \ldots \\ 1 & 2 & 2 \end{pmatrix}, X^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ -2 & -2 & \ldots & 2 \\ -2 & -1 & \ldots & 2 \end{pmatrix}, \hat{\beta} = \begin{pmatrix} \beta 0 \\ \beta 1 \\ \beta 2 \end{pmatrix}, y = \begin{pmatrix} 1.6 \\ 0.8 \\ \ldots \\ 0.399 \end{pmatrix}, x_1 = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}, x_2 = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

This gives the following normal equations

Out[ ]//TraditionalForm=

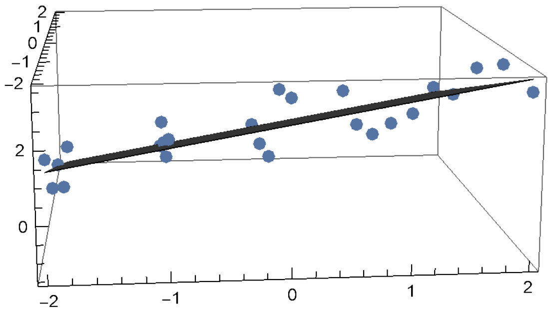
$$25\,\beta0 = 25.$$
$$50\,\beta1 = 25.$$
$$50\,\beta2 = -40.$$

with the solution:

Out[ ]= {{β0 → 1., β1 → 0.5, β2 → -0.8}}

Example: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ (+ random error)

Out[ ]=



fit=	$0.552808\,x - 0.716796\,y + 1.0347$
R^2=	0.84464

When we add a random error term to each data point, the coefficient of determination decreases. We also find slightly different values of β compared to the previous example. We might be tempted to add an additional explanatory variable to improve the fit:

Out[ ]//TraditionalForm=

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots$$

Adjusted R^2

The use of an adjusted R^2 (often written as $\overline{R}^2$ and pronounced “R bar squared”) is an attempt to take account of the phenomenon of the R^2 automatically and spuriously increasing when extra explanatory variables are added to the model. It is a modification that adjusts for the number of explanatory terms in a model relative to the number of data points.

The adjusted R^2 can be negative, and its value will always be less than or equal to that of R^2.

Unlike R^2, the adjusted R^2 increases when a new explanator is included only if the new explanator improves the R^2 more than would be expected by chance.

Adjusted R^2

The adjusted R^2 is defined as

Out[ ]//TraditionalForm=

$$\boxed{\bar{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-p-1} = R^2 - \left(1 - R^2\right) \frac{p}{n-p-1}}$$

where p is the total number of regressors in the model (not counting the constant term), and n is the sample size.

Derivation:

replace SS_{res} with $MSS_{res} = SS_{res} / (n - p)$

replace SS_{tot} with $MSS_{tot} = SS_{tot} / (n - 1)$

Standard error

Standard error SE_{data} of the original data:

Out[∘]//TraditionalForm=

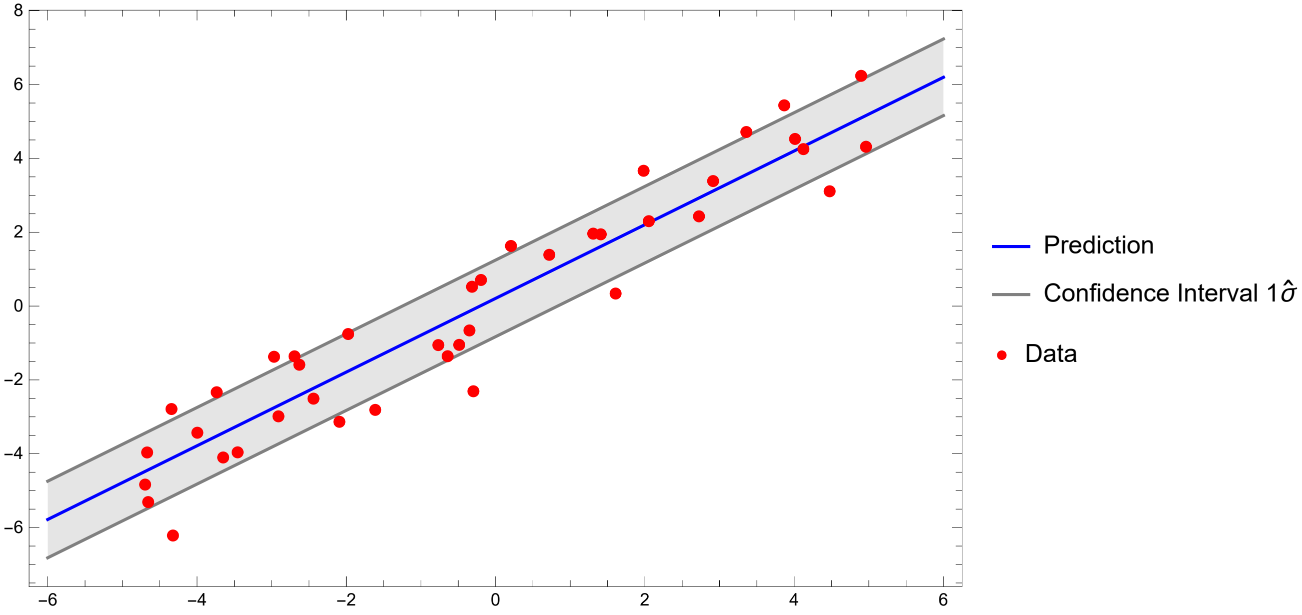
$$SE_{data} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

Standard error of the estimate SE_{est}. With $r_i = y_i - \sum_{j=1}^{n} X_{ij}\,\beta_j$ it follows:

Out[∘]//TraditionalForm=

$$\hat{\sigma} = SE_{est} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} X_{ij}\,\beta_j\right)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} r_i^2}$$

Out[∘]=



Standard error

Note, that p is the number of explanatory variables **without the constant term β_0**! Standard error of the constant/absolute term $\hat{\beta}_0$:

Out[∘]//TraditionalForm=

$$\mathrm{SE}_{\hat{\beta}_0} = \mathrm{SE}_{\mathrm{est}} \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Standard error of the $\hat{\beta}_1$:

Out[∘]//TraditionalForm=

$$\mathrm{SE}_{\hat{\beta}_1} = \frac{\mathrm{SE}_{\mathrm{est}}}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Standard error

We found earlier that:

Out[ ]//TraditionalForm=

$$\boxed{\left(X^\mathsf{T} X\right)\hat{\beta} = X^\mathsf{T} y}$$

accordingly:

Out[ ]//TraditionalForm=

$$\boxed{\hat{\beta} = \left(X^\mathsf{T} X\right)^{-1}\left(X^\mathsf{T} y\right)}$$

The error standard deviation is estimated as

Out[ ]//TraditionalForm=

$$\hat{\sigma} = \mathrm{SE}_{\mathrm{est}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{p} X_{\mathrm{ij}}\,\beta_j\right)^2} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2}$$

Standard error

The variances of the $\hat{\beta}_i$ are the diagonal elements of the standard error matrix:

Out[ ]//TraditionalForm=

$$\boxed{\text{SE matrix} = \hat{\sigma}^2 \left(X^\mathsf{T} X\right)^{-1}}$$

Remember that we assume that the residuals are independently distributed according to $\mathcal{N}\left(0, \hat{\sigma}^2 I\right)$ with the identity matrix I. Also recall, that $\text{Var}(A.X) = A \times \text{Var}(X) \times A^T$, for some random vector X and some non-random matrix A

Out[ ]//TraditionalForm=

$$\text{SE matrix} = \left(X^\mathsf{T} X\right)^{-1}.X^\mathsf{T}\, \hat{\sigma}^2\, \mathrm{I}.X.\left(X^\mathsf{T} X\right)^{-1} = \hat{\sigma}^2 \left(X^\mathsf{T} X\right)^{-1}$$

Example: p=1

$$X^{T} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{21} & \ldots & X_{n1} \end{pmatrix}, \; X = \begin{pmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{pmatrix}$$

$$X^{T} X = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{21} & \ldots & X_{n1} \end{pmatrix} \begin{pmatrix} 1 & X_{11} \\ 1 & X_{21} \\ \vdots & \vdots \\ 1 & X_{n1} \end{pmatrix} = \begin{pmatrix} n & \sum X_{i1} \\ \sum X_{i1} & \sum X_{i1}^{2} \end{pmatrix}$$

$$\left(X^{T} X\right)^{-1} = \frac{1}{\det\left(X^{T} X\right)} \begin{pmatrix} \sum X_{i1}^{2} & -\sum X_{i1} \\ -\sum X_{i1} & n \end{pmatrix} = \frac{1}{n \sum X_{i1}^{2} - \left(\sum X_{i1}\right)^{2}} \begin{pmatrix} \sum X_{i1}^{2} & -\sum X_{i1} \\ -\sum X_{i1} & n \end{pmatrix} = \frac{1/(n-1)}{\operatorname{var}(X)} \begin{pmatrix} \sum X_{i1}^{2} \big/ n & -\overline{X} \\ -\overline{X} & 1 \end{pmatrix}$$

Example: p=1

Thus we can write the $SE_{\hat{\beta}_0}$, and $SE_{\hat{\beta}_1}$

$$SE_{\hat{\beta}_1} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2}\ \sqrt{\frac{1/(n-1)}{var(X)}}\ \sqrt{\sum X_{i1}^2 / n} = \frac{\hat{\sigma}\sqrt{\sum X_{i1}^2}}{\sigma_X \sqrt{n(n-1)}}$$

$$SE_{\hat{\beta}_0} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n} r_i^2}\ \sqrt{\frac{1/(n-1)}{var(X)}}\ \sqrt{\sum X_{i0}^2 / n} = \frac{\hat{\sigma}}{\sigma_X \sqrt{(n-1)}}$$

Example:

Y_i are the average maximum daily temperatures at $n = 1070$ weather stations in the U.S during March, 2001. The predictors are: latitude (X_i), longitude (X_2), and elevation (X_3).

Here is the fitted model:

Out[ ]//TraditionalForm=

$$\mathrm{E}(Y \mid X) = 101 - 2\, X_1 + 0.3\, X_2 - 0.003\, X_3$$

Average temperature decreases as latitude and elevation increase, but it increases as longitude increases. For example, when moving from Miami (latitude 25°) to Detroit (latitude 42°), an increase in latitude of 17°, according to the model average temperature decreases by $2 \cdot 17 = 34°$.

Example:

In the actual data, Miami's temperature was 83° and Detroit's temperature was 45°, so the actual difference was 38°. The sum of squares of the residuals is $\sum_i r_i^2 = 25\,301$, so the estimate of the standard deviation of ϵ is

Out[ ]//TraditionalForm=
$$\hat{\sigma} = \sqrt{\frac{25\,301}{1066}} \approx 4.9$$

The standard error matrix $\hat{\sigma}(X^T X)^{-1}$ is:

$$\begin{pmatrix} 2.4 & -3.2\times 10^{-2} & -1.3\times 10^{-2} & 2.1\times 10^{-4} \\ -3.2\times 10^{-2} & 7.9\times 10^{-4} & 3.3\times 10^{-5} & -2.1\times 10^{-6} \\ -1.3\times 10^{-2} & 3.3\times 10^{-5} & 1.3\times 10^{-4} & -1.8\times 10^{-6} \\ 2.1\times 10^{-4} & -2.1\times 10^{-6} & -1.8\times 10^{-6} & 1.2\times 10^{-7} \end{pmatrix}$$

The diagonal elements give the standard deviations of the parameter estimates, so $\mathrm{SD}(\hat{\beta}_0) = \sqrt{2.4} = 1.54919$, $\mathrm{SD}(\hat{\beta}_1) = \sqrt{7.9\times 10^{-4}} = 0.03$, etc.

Example: p=2

$$X^T = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{21} & \ldots & X_{n1} \\ X_{12} & X_{22} & \ldots & X_{n2} \end{pmatrix}, \; X = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ X_{11} & X_{21} & \ldots & X_{n1} \\ X_{12} & X_{22} & \ldots & X_{n2} \end{pmatrix} \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix} = \begin{pmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1} X_{i2} \\ \sum X_{i2} & \sum X_{i1} X_{i2} & \sum X_{i2}^2 \end{pmatrix}$$

$$\left(X^T X\right)^{-1} = \begin{pmatrix} \frac{\left(\sum_{i=1}^n X_{i1}^2\right) \sum_{i=1}^n X_{i2}^2 - \left(\sum_{i=1}^n X_{i1} X_{i2}\right)^2}{\det(a)} & \frac{\left(\sum_{i=1}^n X_{i2}\right) \sum_{i=1}^n X_{i1} X_{i2} - \left(\sum_{i=1}^n X_{i1}\right) \sum_{i=1}^n X_{i2}^2}{\det(a)} & \frac{\left(\sum_{i=1}^n X_{i1}\right) \sum_{i=1}^n X_{i1} X_{i2} - \left(\sum_{i=1}^n X_{i1}^2\right) \sum_{i=1}^n X_{i2}}{\det(a)} \\ \frac{\left(\sum_{i=1}^n X_{i2}\right) \sum_{i=1}^n X_{i1} X_{i2} - \left(\sum_{i=1}^n X_{i1}\right) \sum_{i=1}^n X_{i2}^2}{\det(a)} & \frac{n \sum_{i=1}^n X_{i2}^2 - \left(\sum_{i=1}^n X_{i2}\right)^2}{\det(a)} & \frac{\left(\sum_{i=1}^n X_{i1}\right) \sum_{i=1}^n X_{i2} - n \sum_{i=1}^n X_{i1} X_{i2}}{\det(a)} \\ \frac{\left(\sum_{i=1}^n X_{i1}\right) \sum_{i=1}^n X_{i1} X_{i2} - \left(\sum_{i=1}^n X_{i1}^2\right) \sum_{i=1}^n X_{i2}}{\det(a)} & \frac{\left(\sum_{i=1}^n X_{i1}\right) \sum_{i=1}^n X_{i2} - n \sum_{i=1}^n X_{i1} X_{i2}}{\det(a)} & \frac{n \sum_{i=1}^n X_{i1}^2 - \left(\sum_{i=1}^n X_{i1}\right)^2}{\det(a)} \end{pmatrix}$$

Example: p=2

with

$$\det(a) = 2\left(\sum_{i=1}^{n} X_{i1}\right)\left(\sum_{i=1}^{n} X_{i2}\right)\sum_{i=1}^{n} X_{i1} X_{i2} - n\left(\sum_{i=1}^{n} X_{i1} X_{i2}\right)^2 - \left(\sum_{i=1}^{n} X_{i1}\right)^2 \sum_{i=1}^{n} X_{i2}^2 + \left(\sum_{i=1}^{n} X_{i1}^2\right)\left(-\left(\sum_{i=1}^{n} X_{i2}\right)^2 + n\sum_{i=1}^{n} X_{i2}^2\right)$$

We see the obvious reason for using the matrix notation. The expressions quickly become impractically large. Further simplification of the matrix is left to the reader as an exercise.

Regression - p-value

One of the main goals of fitting a regression model is to determine which predictor variables are truly related to the response. This can be formulated as a set of hypothesis tests.

For each predictor variable X_i, we may test the null hypothesis $\beta_i = 0$ against the alternative $\beta_i \neq 0$. To obtain the p-value, first standardize the slope estimates:

$\hat{\beta}_1 / \mathrm{SD}(\hat{\beta}_1) = 2 / 0.03 = 71.1568 \approx -72$

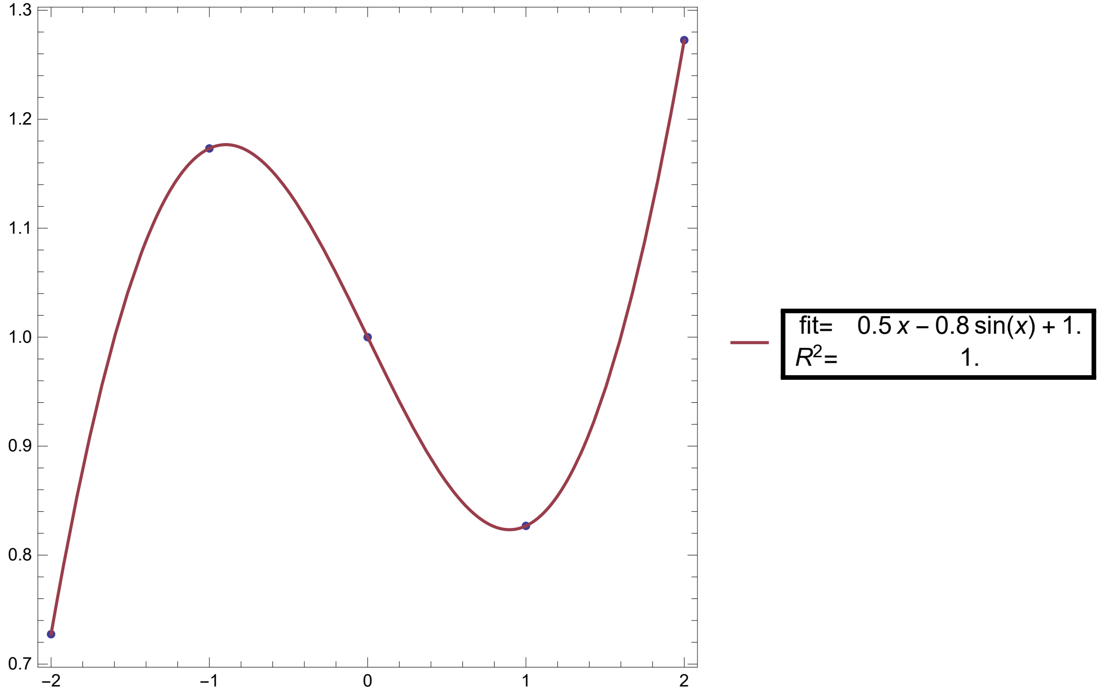
$\hat{\beta}_2 / \mathrm{SD}(\hat{\beta}_2) \approx 29$

$\hat{\beta}_3 / \mathrm{SD}(\hat{\beta}_3) \approx -9$

Then look up the result in a Z table. In this case the p-values are all extremely small, so all three predictors are significantly related to the response.

Example: $y = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

The model equations need to be linear in β not in x!



Out[ ]//TraditionalForm=

$$\boxed{\left(X^{\mathsf{T}} X\right)\hat{\beta} = X^{\mathsf{T}} y}$$

Example: $y = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

the data is {{-2, 0.73}, {-1, 1.17}, {0, 1.},{1, 0.83}, {2, 1.27}},

$$X = \begin{pmatrix} 1 & x_1 & \mathrm{Sin}[x_1] \\ 1 & x_2 & \mathrm{Sin}[x_2] \\ 1 & x_3 & \mathrm{Sin}[x_3] \\ 1 & x_4 & \mathrm{Sin}[x_4] \\ 1 & x_5 & \mathrm{Sin}[x_5] \end{pmatrix} = \begin{pmatrix} 1 & -2 & -0.909297 \\ 1 & -1 & -0.841471 \\ 1 & 0 & 0 \\ 1 & 1 & 0.841471 \\ 1 & 2 & 0.909297 \end{pmatrix}, X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 2 \\ -0.91 & -0.84 & 0 & 0.84 & 0.91 \end{pmatrix}, \hat{\beta} = \begin{pmatrix} \beta 0 \\ \beta 1 \\ \beta 2 \end{pmatrix}, y = \begin{pmatrix} 0.73 \\ 1.17 \\ 1. \\ 0.83 \\ 1.27 \end{pmatrix}, x = \begin{pmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}$$

Example: $y = \beta_0 + \beta_1 x + \beta_2 \sin(x)$

with the solution:

Out[ ]= $\{\{\beta 0 \to 1., \beta 1 \to 0.49348, \beta 2 \to -0.788476\}\}$

The standard error and the standard error matrix (compare with the LinearModelFit parameter table from *Mathematica*):

$$SE = 7.46826\times10^{-16}$$

$$\sqrt{SE^2 \left(X^T X\right)^{-1}} = \begin{pmatrix} 3.33991\times10^{-16} & 0. & 0. \\ 0. & 8.45673\times10^{-16} & 0. + 1.11329\times10^{-15}\,i \\ 0. & 0. + 1.11329\times10^{-15}\,i & 1.52633\times10^{-15} \end{pmatrix}$$

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	1.	3.68219×10^{-16}	2.71577×10^{15}	0.
xx	0.49348	9.3234×10^{-16}	5.29291×10^{14}	0.
Sin[xx]	−0.788476	1.68275×10^{-15}	-4.68563×10^{14}	0.

and the t-statistics:

Out[ ]= $\{2.99409\times10^{15}, 5.83535\times10^{14}, -5.16583\times10^{14}\}$

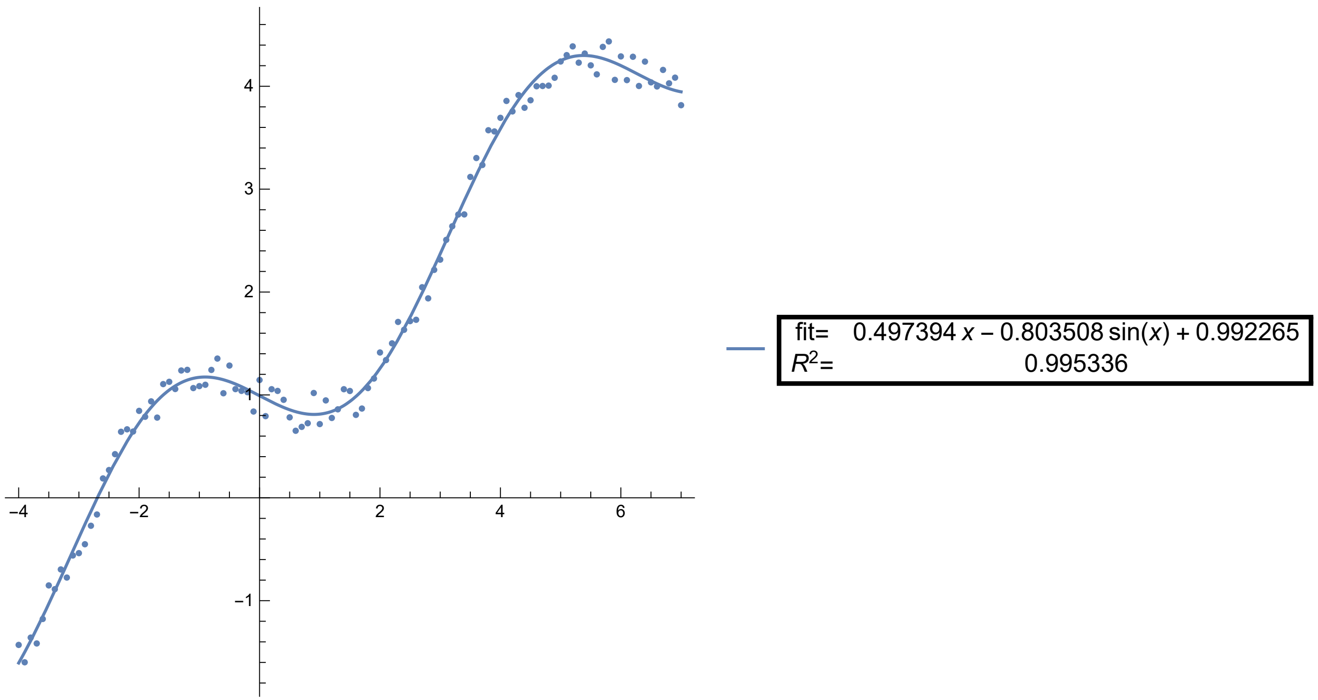
leading to p-values of essentially 0.

Example:

Using the same example as above, with a larger plot range and an additional error term

err

Out[ ]=



	Estimate	Standard Error	t-Statistic	P-Value
1	0.992265	0.0124711	79.5653	1.04192×10^{-97}
x	0.497394	0.00348642	142.666	7.93369×10^{-125}
Sin[x]	−0.803508	0.0168201	−47.7707	1.84109×10^{-74}

Regression with categorical explanatory variables

Example: poverty vs. region

explanatory variable: 0: east / 1: west

model: $\widehat{\texttt{poverty}} = 11.17 + 0.38\ \text{region} : \text{west}$

For eastern states plug in 0 for x : $\widehat{\texttt{poverty}} = 11.17 + 0.38 \times 0 = 11.17$ (reference level/ intercept)

For western states plug in 1 for x : $\widehat{\texttt{poverty}} = 11.17 + 0.38 \times 1 = 11.55$

Regression with categorical explanatory variables

Example: poverty vs. region

Now we use a new region variable with four levels: northeast, mid-west, west, south.

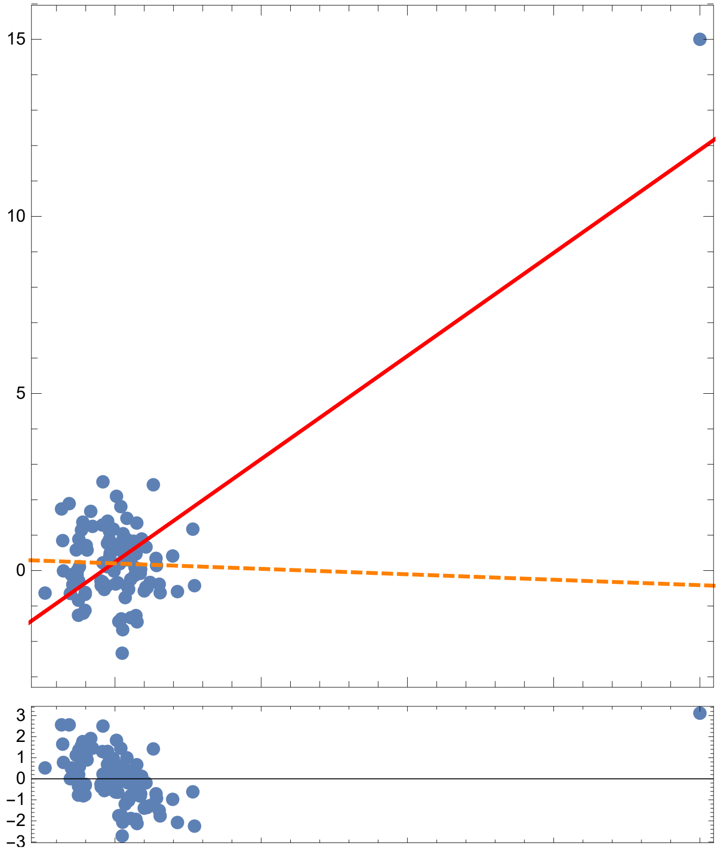
	Estimate	Std. Error	t value	$\mathcal{P}(>\|t\|)$
(Intercept)	9.5	0.87	10.94	0.
region4:midwest	0.03	1.15	0.02	0.98
region4:west	1.79	1.13	1.59	0.12
region4:south	4.16	1.07	3.87	0.

$$\widehat{\texttt{poverty}} = 9.5 + 0.03\,\text{reg4 : mw} + 1.79\,\text{reg4 : }w + 4.16\,\text{reg : }s$$

The predicted poverty rate for western states is:

$$\widehat{\texttt{poverty}} = 9.5 + 0.03\times 0 + 1.79\times 1 + 4.16\times 0 = 11.29$$

Outliers in regression



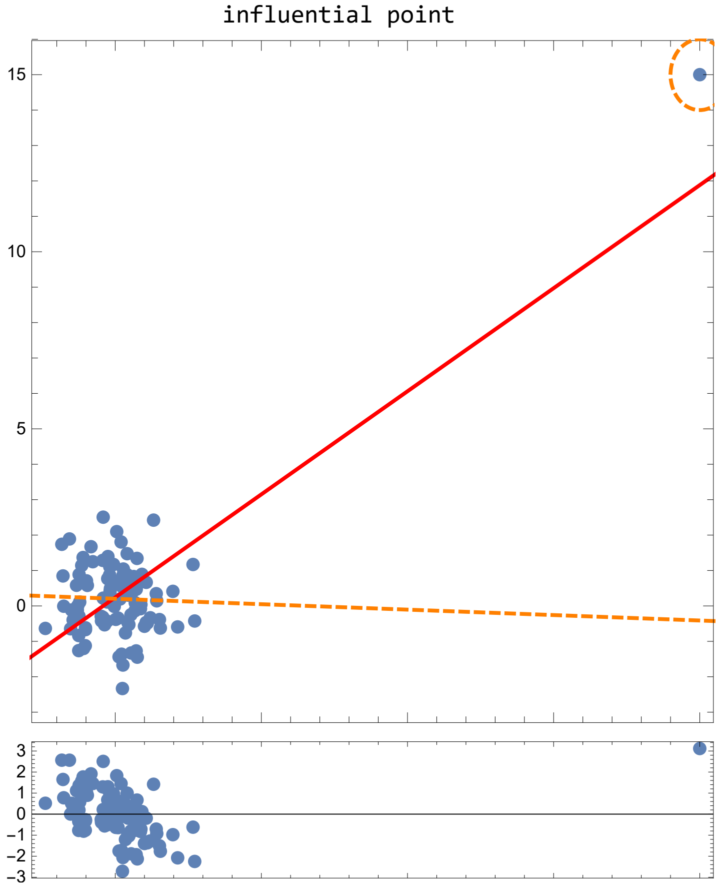
How does the outlier influence the least squares line?

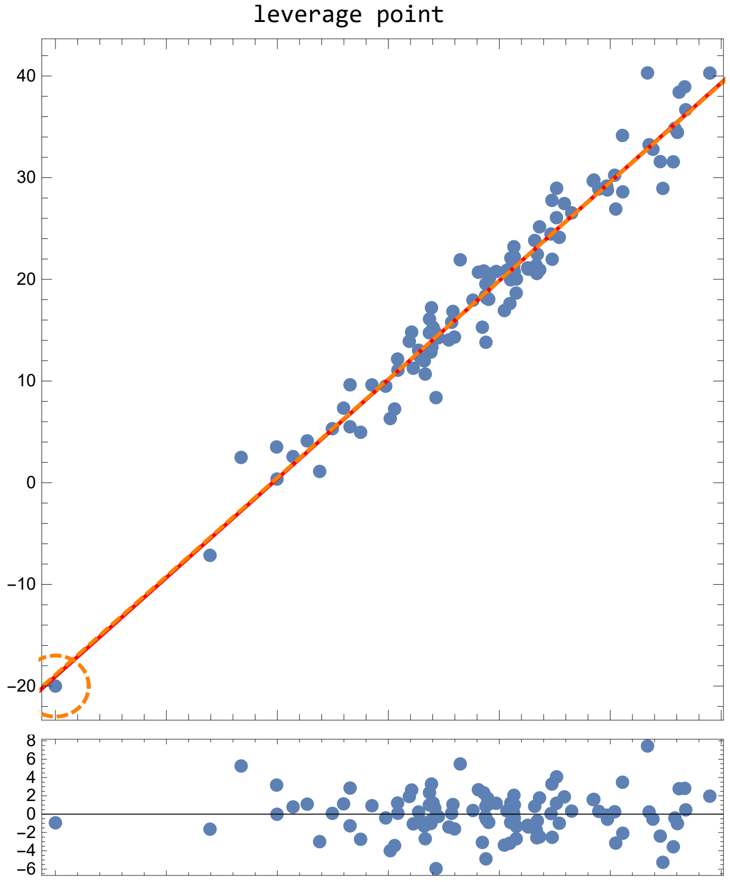
Without the outlier there is no relationship between x and y!

Outliers in regression

- outliers are points that fall away from the cloud of points.
- outliers that fall horizontally away from the center of the cloud but don’t influence the slope of the regression line are called **leverage points**
- outliers that actually influence the slope of the regression line are called **influential points**
- usually high leverage points
- to determine if a point is influential, visualize the regression line with and without the point, and ask: *Does the slope of the line change considerably?*

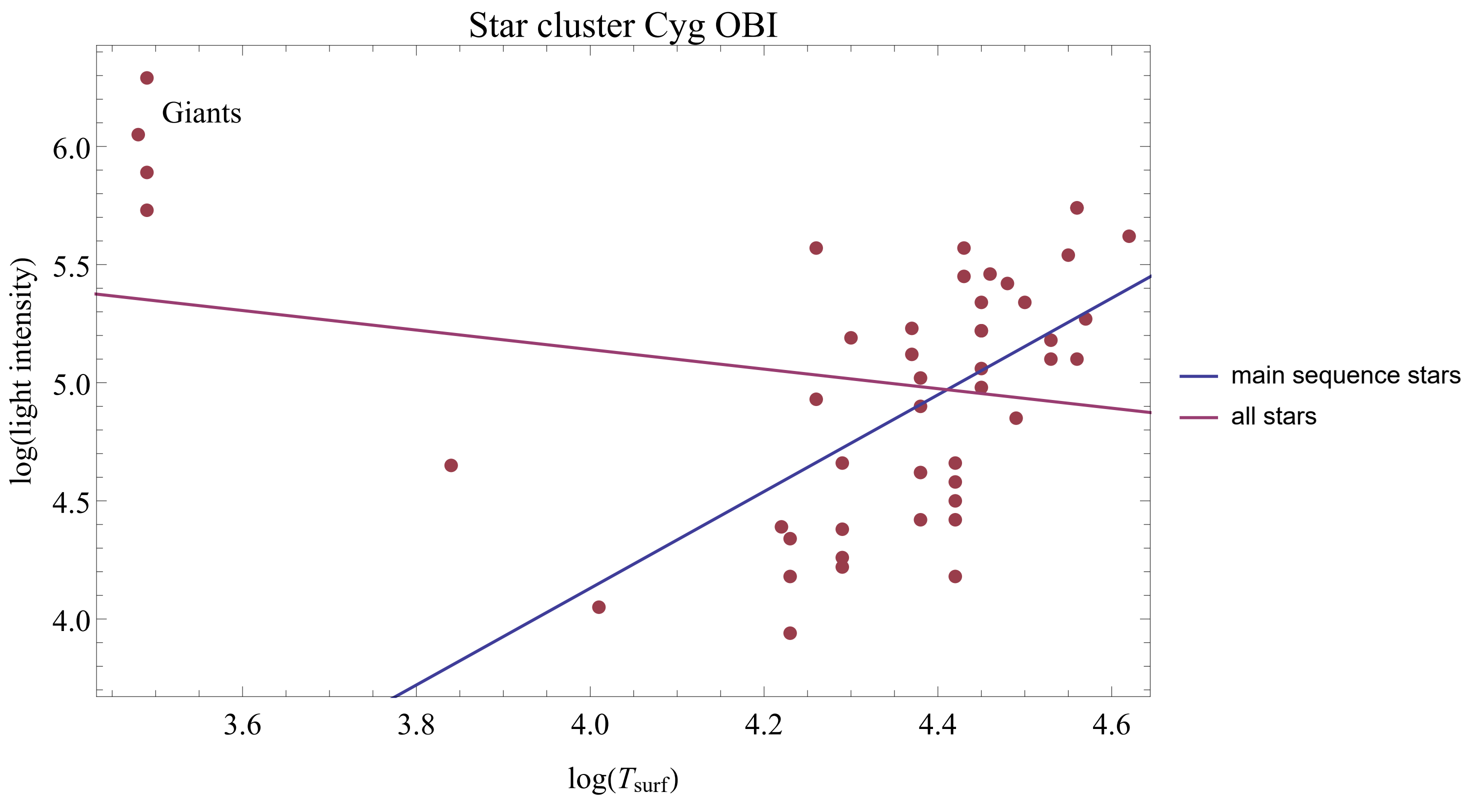
Outliers in regression





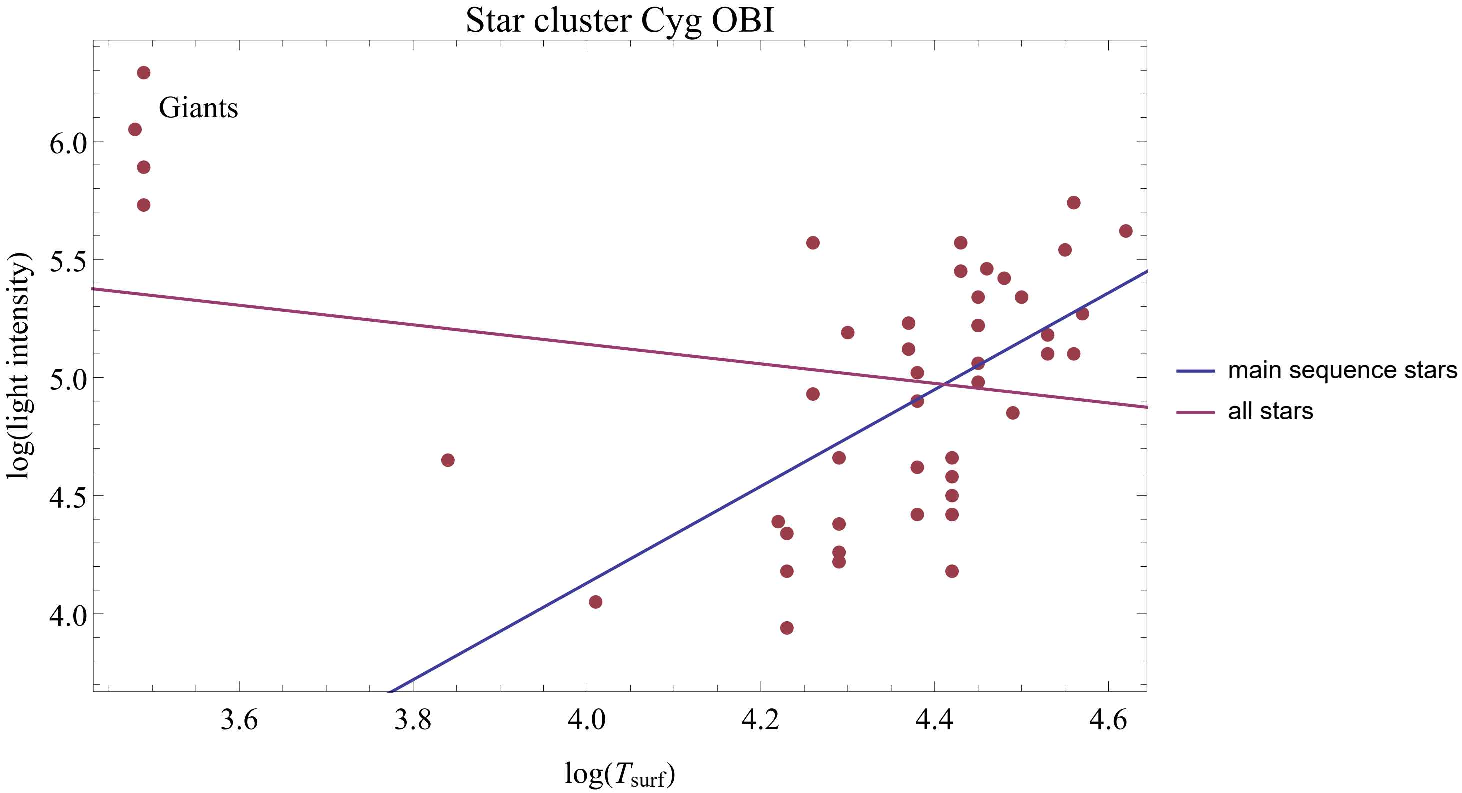
Influential points: Cyg OB1

Out[ ]=



Inference for linear regression

Out[ ]=



	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

linear model: log(intensity) = −4.05652 + 2.04666 log(T_{surf})

R squared= 0.366564

- hypothesis testing for significance of predictor
- confidence interval for slope

Testing for the slope

Is the explanatory variable a significant predictor of the response variable?

	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

$H_0: \beta_1 = 0$ The explanatory variable is not a significant predictor of the response variable, i.e. no relationship → slope is 0

$H_A: \beta_1 \neq 0$ The explanatory variable is a significant predictor of the response variable, i.e. no relationship → slope is different than 0

Use a t-statistic in inference for regression

Out[∘]= $$T = \frac{\text{point estimate} - \text{null value}}{\text{SE}} \implies \boxed{T = \frac{b_1 + 0}{\text{SE}_{b_1}} \qquad \text{df} = n - 2}$$

Testing for the slope

df = $n - 2$: Lose 1 dof for each parameter estimated. In linear regression we estimate 2 parameters: β_0 and β_1. Each p-value is the **two-sided *p*-value** for the t-statistic and can be used to assess whether the parameter estimate is statistically significantly different from 0.

	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

Out[ ]= $T = \frac{2.047 + 0}{0.42} = 4.87$ df = 43 − 2 = 41 p-value = $\mathcal{P}(|T|>4.87) = 0.0000169706$

Confidence interval

Calculate the 95% confidence interval for the slope of the relationship between log(T_{surf}) and log(intensity).

	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

df = 43 − 2 = 41, ⇒ t^*_{41}=-2.0195409704413745

CI = 2.04666 ± 2.06 × 0.420174 =(1.18,2.91)

We are 95% confident, that the slope for the linear regression line between log(T_{surf}) and log(intensity) is between 1.18 and 2.91.

Variability partitioning

- So far: t-test as a way to evaluate the strength of evidence for a hypothesis test for the slope of relationship between x and y.
- Alternative: consider the variability in y explained by x compared to the unexplained variability.
- Partitioning the variability in y to explained and unexplained variability requires analysis of variances (ANOVA)

Variability partitioning

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

	DF	SS	MS	F-Statistic	P-Value
x	1	3.90722	3.90722	23.7264	0.0000169706
Error	41	6.75182	0.164679		
Total	42	10.659			

sum of squares

total variability in y : $SS_{tot} = \sum (y - \overline{y})^2 = 10.659$

unexplained variability in y (residuals) : $SS_{res} = \sum \left(y - \hat{y}\right)^2 = \sum e_i^2 = 6.75182$

explained variability in y : $SS_{reg} = 10.659 - 6.75182 = 3.90772$

Variability partitioning

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

	DF	SS	MS	F-Statistic	P-Value
x	1	3.90722	3.90722	23.7264	0.0000169706
Error	41	6.75182	0.164679		
Total	42	10.659			

degrees of freedom

total degrees of freedom : $df_{tot} = 43 - 1 = 42$

regression degrees of freedom : $df_{reg} = 1$ (only 1 predictor)

residual degrees of freedom : $df_{res} = 42 - 1 = 41$

Variability partitioning

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	−4.05652	1.84414	−2.19968	0.0335238
x	2.04666	0.420174	4.87097	0.0000169706

	DF	SS	MS	F-Statistic	P-Value
x	1	3.90722	3.90722	23.7264	0.0000169706
Error	41	6.75182	0.164679		
Total	42	10.659			

mean squares

MS regression : $MS_{reg} = \frac{SS_{reg}}{df_{reg}} = \frac{3.90722}{1} = 3.90722$

MS residuals : $MS_{res} = \frac{SS_{res}}{df_{res}} = \frac{6.75182}{41} = 0.164679$

F statistic

ratio of explained to unexplained variability : $F_{(1,41)} = \frac{MS_{reg}}{MS_{res}} = 23.7264$

Variability partitioning

Inference

$H_o : \beta_1 = 0,\ H_A : \beta_1 \neq 0 \implies p - \text{value} = 0.0000169706$

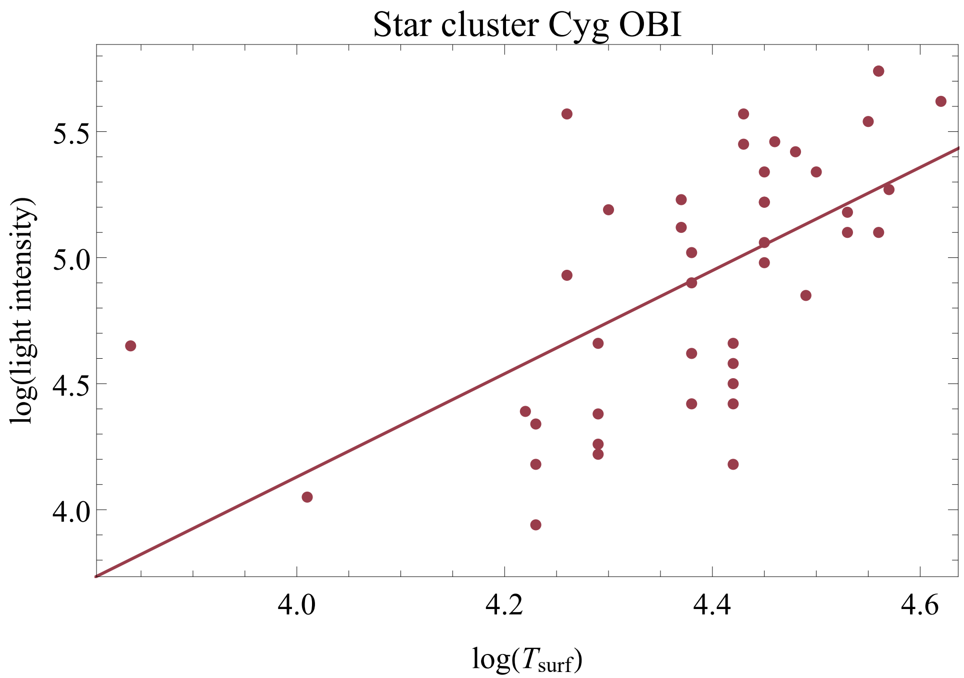
The data provide convincing evidence that the slope is significantly different than 0, i.e. the explanatory variable is a significant predictor of the response variable.

R^2 revisited

- R^2 is the proportion of variability in y explained by the model:
  - large ⟶ linear relationship between x and y exists
  - small ⟶ evidence provided by the data may not be convincing
- Two ways to calculate R^2 :
  - using correlation: square of the correlation coefficient
  - from the definition: proportion of explained to total variability

R^2 revisited

Out[ ]=



	DF	SS	MS	F-Statistic	P-Value
x	1	3.90722	3.90722	23.7264	0.0000169706
Error	41	6.75182	0.164679		
Total	42	10.659			

R= 0.605445

R squared= 0.366564

- R^2 = square of correlation coefficient = $0.605445^2 = 0.366564$
- $R^2 = \frac{\text{explained variability}}{\text{total variability}} = \frac{SS_{reg}}{SS_{tot}} = \frac{3.90722}{10.659}$=0.366564

Model selection

We will use the following data in this section:
We will consider ebay auctions of a video game called Mario Kart for the Nintendo Wii. The outcome variable of interest is the total price of an auction, which is the highest bid plus the shipping cost. We will try to determine how total price is related to each characteristic in an auction while simultaneously controlling for other variables.

price final auction price plus shipping costs, in US dollars
cond_new a coded two-level categorical variable, which takes value 1 when the game is new and 0 if the game is used
stock_photo a coded two-level categorical variable, which takes value 1 if the primary photo used in the auction was a stock photo and 0 if the photo was unique to that auction
duration the length of the auction, in days, taking values from 1 to 10
wheels the number of Wii wheels included with the auction (a Wii wheel is a plastic racing wheel that holds the Wii controller and is an optional but helpful accessory for playing Mario Kart

Model selection

Out[ ]//TableForm=

	totalPr	cond	stockPhoto	duration	wheels
1	51.55	1	1	3	1
2	37.04	0	1	7	1
3	45.5	1	0	3	1
4	44	1	1	3	1
5	71	1	1	1	2
6	45	1	1	3	0
7	37.02	0	1	1	0
8	53.99	1	1	1	2
9	47	0	1	3	1
10	50	0	0	7	1

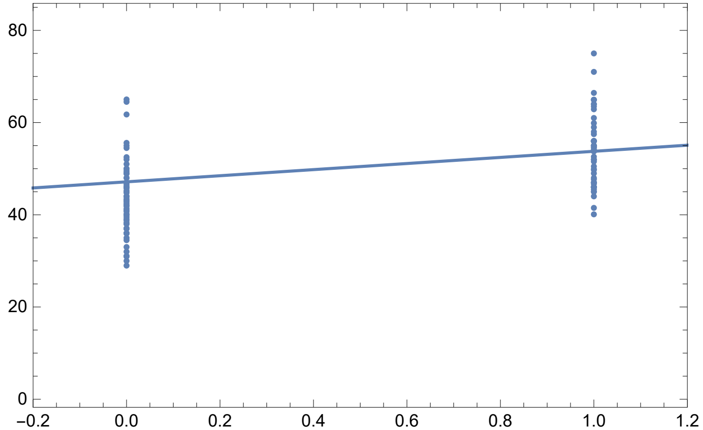
Example: A single-variable model for the Mario Kart data

price = $b_0 + b_1$ cond_new (0:used, 1:new)

Out[ ]= 42.8711 + 10.8996 x

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	42.8711	0.813981	52.6685	1.01937×10^{-93}
x	10.8996	1.25834	8.66188	1.0557×10^{-14}



The model predicts an extra \$10.90 for new games versus used ones.

Example: A multi-variable model for the Mario Kart data

price = $b_0 + b_1$ cond_new + b_2 stock_phot + b_3 duration + b_4 wheels

Out[ ]= 36.211 + 5.13056 x1 + 1.08031 x2 - 0.0268075 x3 + 7.28518 x4

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.211	1.51401	23.9173	1.43475×10^{-50}
x1	5.13056	1.05112	4.88103	2.91185×10^{-6}
x2	1.08031	1.05682	1.02222	0.30849
x3	−0.0268075	0.190412	−0.140787	0.888247
x4	7.28518	0.554693	13.1337	5.88816×10^{-26}

We first use R^2 to determine the amount of variability in the response that was explained by the model:

Out[ ]//TraditionalForm=

$$R^2 = 1 - \frac{\text{variability in residuals}}{\text{variability in the outcome}} = 1 - \frac{\mathrm{Var}(e_i)}{\mathrm{Var}(y_i)}$$

Example: A multi-variable model for the Mario Kart data

The regular R^2 is actually a biased estimate of the amount of variability explained by the model. To get a better estimate, we use the adjusted R^2.

Out[ ]//TraditionalForm=

$$R^2_{\text{adj}} = 1 - \frac{\text{Var}(e_i)/(\text{n}-\text{k}-1)}{\text{Var}(y_i)/(\text{n}-1)} = 1 - \frac{\text{Var}(e_i)}{\text{Var}(y_i)} \times \frac{n-1}{n-k-1}$$

where n is the number of cases used to fit the model and k is the number of predictor variables in the model. Because k is never negative, the adjusted R^2 will be smaller - often times just a little smaller - than the unadjusted R^2.

Example:Model selection

The best model is not always the most complicated. Sometimes including variables that are not evidently important can actually reduce the accuracy of predictions. In this section we discuss model selection strategies, which will help us eliminate from the model variables that are less important.

In this section, and in practice, the model that includes all available explanatory variables is often referred to as the **full model**. Our goal is to assess whether the full model is the best model. If it isn't, we want to identify a smaller model that is preferable.

Example:Model selection

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.211	1.51401	23.9173	1.43475×10^{-50}
x1	5.13056	1.05112	4.88103	2.91185×10^{-6}
x2	1.08031	1.05682	1.02222	0.30849
x3	−0.0268075	0.190412	−0.140787	0.888247
x4	7.28518	0.554693	13.1337	5.88816×10^{-26}

R^2_{adj}=0.710762 dof=136

The table provides a summary of the regression output for the full model for the auction data. The last column of the table lists p-values that can be used to assess hypotheses of the following form:

$H_0 : \beta_i = 0$ when the other explanatory variables are included in the model.
$H_A : \beta_i \neq 0$ when the other explanatory variables are included in the model.

Example:Model selection

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.211	1.51401	23.9173	1.43475×10^{-50}
x1	5.13056	1.05112	4.88103	2.91185×10^{-6}
x2	1.08031	1.05682	1.02222	0.30849
x3	−0.0268075	0.190412	−0.140787	0.888247
x4	7.28518	0.554693	13.1337	5.88816×10^{-26}

R^2_{adj}=0.710762 dof=136

- The coefficient x1 (cond_new) has a test statistic to T=4.88 and a p-value for its corresponding hypotheses ($H_0 : \beta_1 = 0$, $H_A : \beta_1 \neq 0$) of 0.
  Interpretation: If we keep all the other variables in the model and add no others, then there is strong evidence that a game’s condition (new or used) has a real relationship with the total auction price.
- Is there strong evidence that using a stock photo (x2) is related to the total auction price?
  The test statistic for x2 is 1.02 and its p-value is 0.31. Keeping the other predictors as they are, there is no strong evidence, that using a photo in an auction is related to the total price. We might consider removing the variable x2 from our model.

Example: Two model selection strategy (p-value based)

The **backward-elimination strategy** starts with the model that includes all potential predictor variables. Variables are eliminated one-at-a-time from the model until only variables with statistically significant p-values remain. The strategy within each elimination step is to drop the variable with the largest p-value, refit the model, and reassess the inclusion of all variables.

Out[∘]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.211	1.51401	23.9173	1.43475×10^{-50}
cond	5.13056	1.05112	4.88103	2.91185×10^{-6}
stockPhoto	1.08031	1.05682	1.02222	0.30849
duration	−0.0268075	0.190412	−0.140787	0.888247
wheels	7.28518	0.554693	13.1337	5.88816×10^{-26}

R^2_{adj}=0.710762   dof=136

1. removing the duration dependence

Out[∘]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.0483	0.974534	36.9902	3.56293×10^{-73}
cond	5.17628	0.996116	5.19647	7.20834×10^{-7}
stockPhoto	1.11772	1.0192	1.09667	0.274712
wheels	7.29836	0.544779	13.3969	1.11044×10^{-26}

R^2_{adj}=0.712832   dof=137

Example: Two model selection strategy (p-value based)

2. removing the stockPhoto dependence

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.7849	0.706557	52.0623	1.33414×10^{-92}
cond	5.58483	0.924509	6.04086	1.34618×10^{-8}
wheels	7.23284	0.541891	13.3474	1.29451×10^{-26}

R^2_{adj}=0.71241 dof=138

The two remaining predictors have statistically significant coefficients with p-values of about zero. We stop the elimination. Our final model:

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ = 36.79+5.58 cond + 7.23 wheels

Example: Two model selection strategy (R^2_{adj} based)

Instead of using the p-value as indicator for elimination we use R^2_{adj}:

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.211	1.51401	23.9173	1.43475×10^{-50}
cond	5.13056	1.05112	4.88103	2.91185×10^{-6}
stockPhoto	1.08031	1.05682	1.02222	0.30849
duration	−0.0268075	0.190412	−0.140787	0.888247
wheels	7.28518	0.554693	13.1337	5.88816×10^{-26}

R^2_{adj}=0.710762 dof=136

Example: Two model selection strategy (R^2_{adj} based)

We now remove one variable and identify the fit with the highest R^2_{adj} (the bottom left model with cond, stockPhoto, and wheels dependence is the best.)

	Estimate	Standard Error	t-Statistic	P-Value
1	37.2959	1.61755	23.057	5.27337×10^{-49}
stockPhoto	2.45541	1.10016	2.23186	0.027248
duration	−0.313954	0.195601	−1.60507	0.11078
wheels	8.25583	0.559295	14.7611	4.17451×10^{-30}

R^2_{adj}=0.662575 dof=137

	Estimate	Standard Error	t-Statistic	P-Value
1	37.175	1.18459	31.3822	2.02148×10^{-64}
cond	5.417	1.01325	5.34614	3.65996×10^{-7}
duration	−0.075751	0.184324	−0.410966	0.68174
wheels	7.2018	0.548753	13.1239	5.46696×10^{-26}

R^2_{adj}=0.710667 dof=137

Out[ ]=

	Estimate	Standard Error	t-Statistic	P-Value
1	36.0483	0.974534	36.9902	3.56293×10^{-73}
cond	5.17628	0.996116	5.19647	7.20834×10^{-7}
stockPhoto	1.11772	1.0192	1.09667	0.274712
wheels	7.29836	0.544779	13.3969	1.11044×10^{-26}

R^2_{adj}=0.712832 dof=137

	Estimate	Standard Error	t-Statistic	P-Value
1	45.6134	2.00188	22.7853	1.91706×10^{-48}
cond	10.0798	1.47246	6.84555	2.33695×10^{-10}
stockPhoto	−0.960692	1.56863	−0.612442	0.541261
duration	−0.448809	0.281634	−1.59359	0.113332

R^2_{adj}=0.348699 dof=137

Example: Two model selection strategy (R^2_{adj} based)

Then we repeat. This time, if we remove any further variable the R^2_{adj} will always decrease. Accordingly, we stop further elimination..

	Estimate	Standard Error	t-Statistic	P-Value
1	35.3144	1.05118	33.5952	2.65832×10^{-68}
stockPhoto	3.09847	1.03046	3.00689	0.00313702
wheels	8.53844	0.53388	15.9932	3.17871×10^{-33}

R^2_{adj}=0.658721 dof=138

	Estimate	Standard Error	t-Statistic	P-Value
1	36.7849	0.706557	52.0623	1.33414×10^{-92}
cond	5.58483	0.924509	6.04086	1.34618×10^{-8}
wheels	7.23284	0.541891	13.3474	1.29451×10^{-26}

R^2_{adj}=0.71241 dof=138

Out[∘]=

	Estimate	Standard Error	t-Statistic	P-Value
1	43.1026	1.24183	34.7089	4.7346×10^{-70}
cond	11.022	1.35606	8.12796	2.22674×10^{-13}
stockPhoto	−0.379595	1.53414	−0.247432	0.804942

R^2_{adj}=0.341433 dof=138

Example: Two model selection strategy (R^2_{adj} based)

Our final model using the R^2_{adj} criterion is:

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 = 36.0483 + 5.17628\ \text{cond} + 1.11772\ \text{stockPhoto} + 7.29836\ \text{wheels}$

The **forward-selection strategy** is the reverse of the backward-elimination technique. Instead of eliminating variables one-at-a-time, we add variables one-at-a-time until we cannot find any variables that present strong evidence of their importance in the model.

Both strategies might result in different “optimal” models. If so, select the model with the higher R^2_{adj}.

Conditions/assumptions for regression

Regression models using the model:

Out[ ]//TraditionalForm=

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

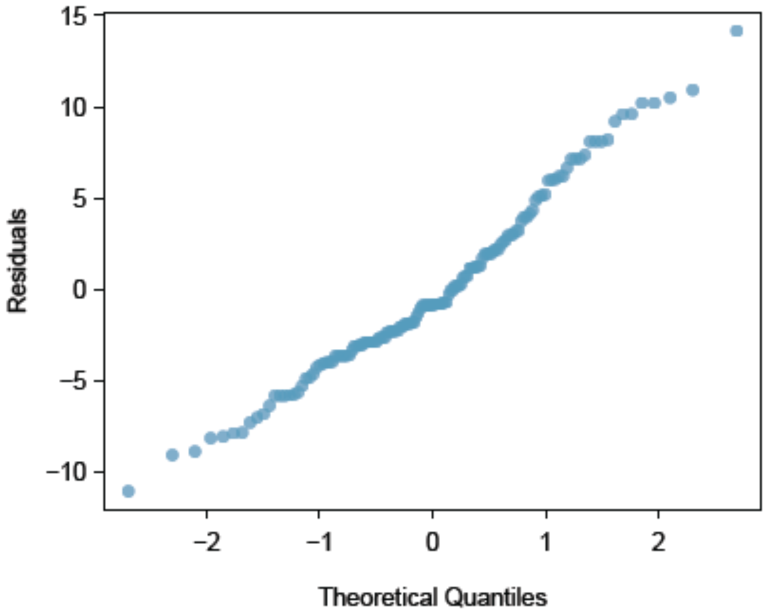
depends on the following assumptions:

- the residuals of the model are nearly normal
- the variability of the residuals is nearly constant
- the residuals are independent
- each variable is linearly related to the outcome.

We will show some graphical methods to assess the validity of these conditions:

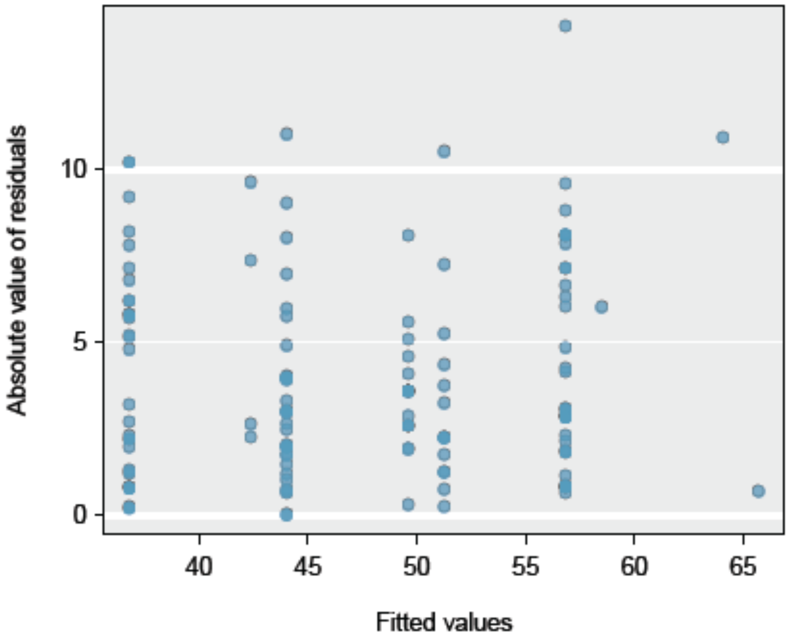
Conditions/assumptions for regression

- A normal probability plot of the residuals can be used to check for problematic outliers.



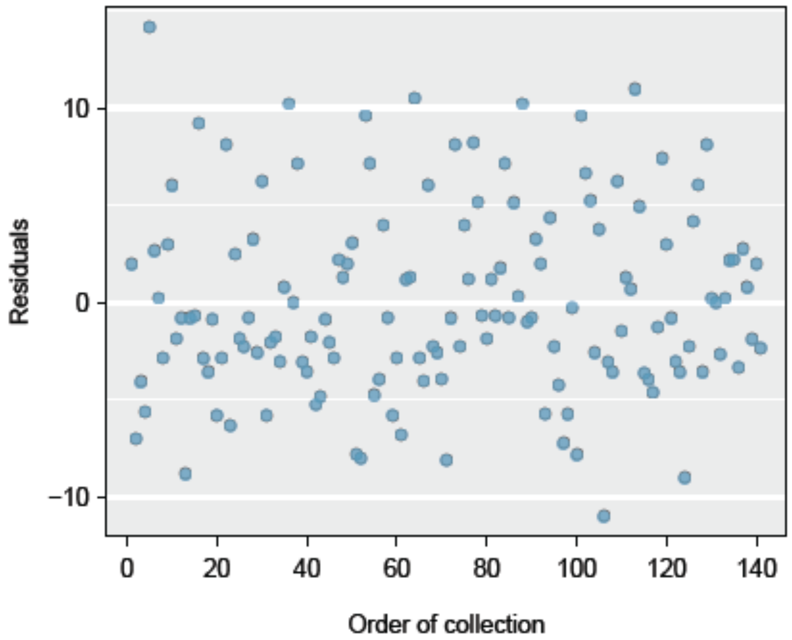
Conditions/assumptions for regression

- Comparing the absolute value of the residuals against the fitted values ($\hat{y}_i$) is helpful in identifying deviations from the constant variance assumption.



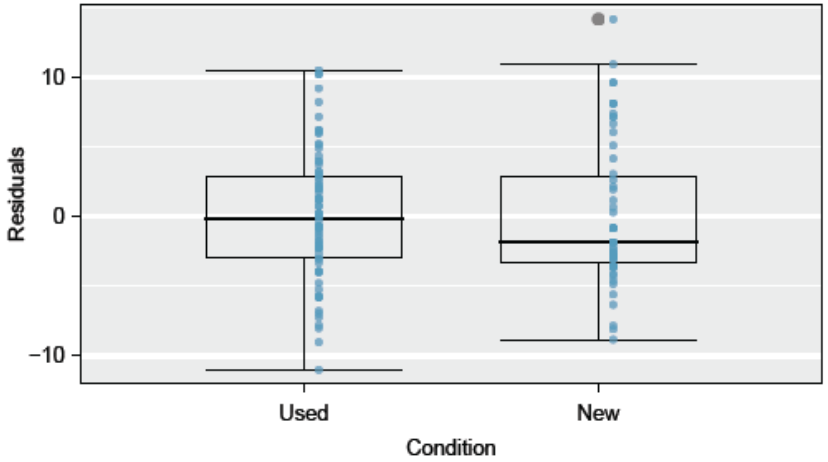
Conditions/assumptions for regression

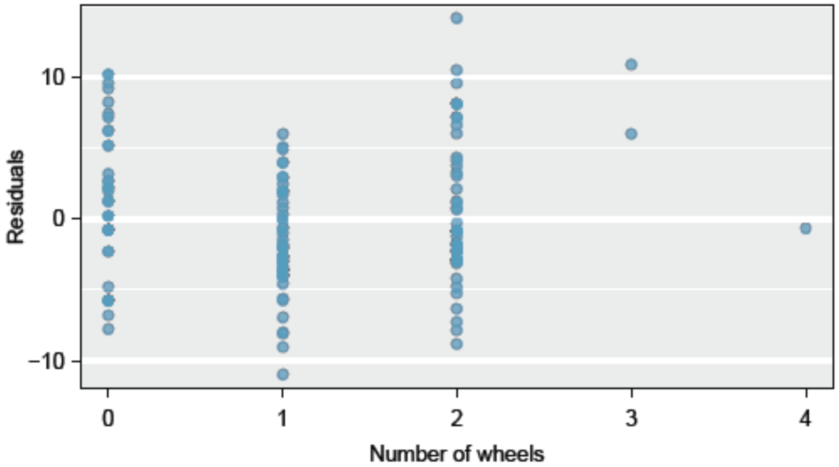
- Plotting residuals in the order that their corresponding observations were collected helps identify connections between successive observations. If it seems that consecutive observations tend to be close to each other, this indicates the independence assumption of the observations would fail.



Conditions/assumptions for regression

- Residuals against each predictor variable to identify residual trends.





Init