Linear Regression

Department of Computer Languages and Systems

Introduction

- Linear regressions tries to approximate a relationship between one dependent (response, outcome) and one or more independent (predictor, explanatory) variable(s):
  - Simple linear regression: it concerns the study of only one independent variable
  - Multiple linear regression: it concerns the study of two or more independent variables

Introduction

Purposes of regression analysis

- Explanatory: A regression analysis explains the relationship between the response and predictor variables

- Predictive: A regression model can give a point estimate of the response variable based on the value of the predictors

Simple linear regression

- We want to find the linear relationship between a dependent variable y and an independent variable x by fitting a linear function to our observed data (x_i, y_i):

$$y = b_0 + b_1 x + \varepsilon$$

  - This is a line where y is the variable we want to predict, x is the input variable we know and b_0 and b_1 are the regression coefficients that we need to estimate
  - b_0 is called the intercept (or bias) because it determines where the line intercepts the y-axis. The b_1 term is called the slope because it defines the slope of the line. ε is the residual error.

Multiple linear regression

- The equation of a multiple linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m + \varepsilon$$

where y is the dependent variable, x_i is the independent variable i, β_0 is the constant of the equation, β_i is the regression coefficient associated to the variable x_i, and ε is the residual error

Multiple linear regression (ii)

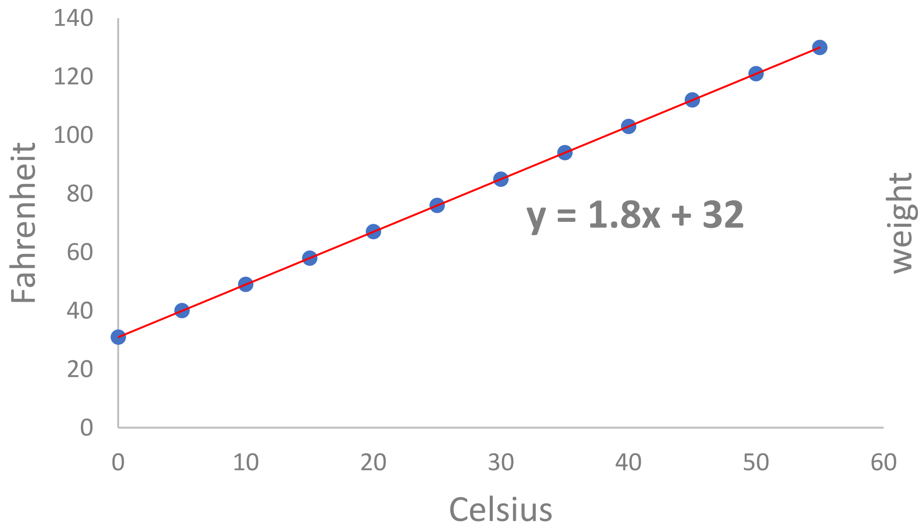
- If we have a sample with a total of n observations, the multiple linear regression model can be expressed in matrix form:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

that is, $Y = X\beta + \varepsilon$

Types of relationships

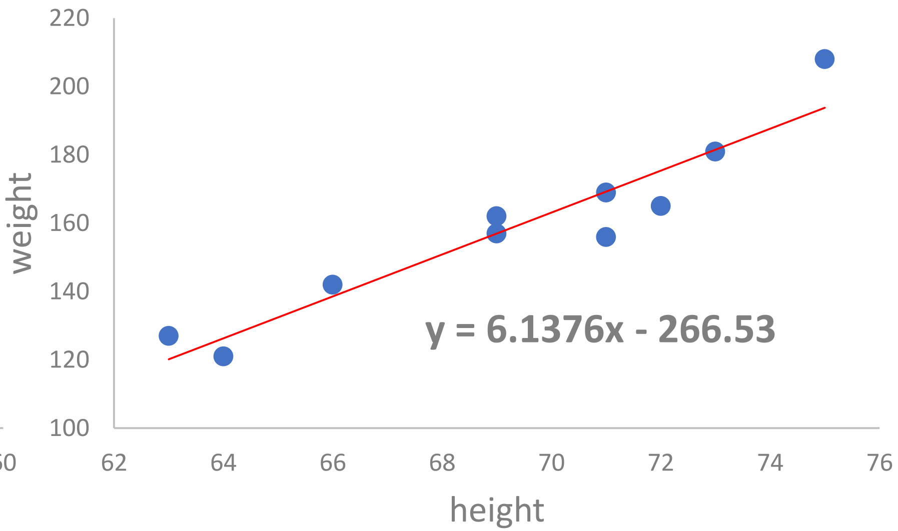
Deterministic relationship



when there is a mathematical formula that allows the values of one of the variables to be calculated from the values of the other

$$Fahr = 32 + 1.8\,Cels$$

Statistical relationship



when there is no mathematical expression that relates the variables exactly

$$weight = 6.1376\,height - 266.53$$

Fundamentals of simple linear regression

- Hypothesis (assumptions):
  - Linearity: the response variable is a linear combination of parameters (regression coefficients) and the predictor variable
  - Normality: the variables follow a symmetric and Gaussian distribution

- Preliminary assessment of the strength of the hypothesis:
  - Linearity: linear correlation coefficient
  - Normality: regression plot (scatterplot), histogram, Q-Q plot

Linear correlation coefficient

Pearson's correlation coefficient: a measure of linear correlation between two variables:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad -1 \leq r \leq 1$$

where is the sample covariance S_{xy} and S_x and S_y are the standard deviations of the variables x and y

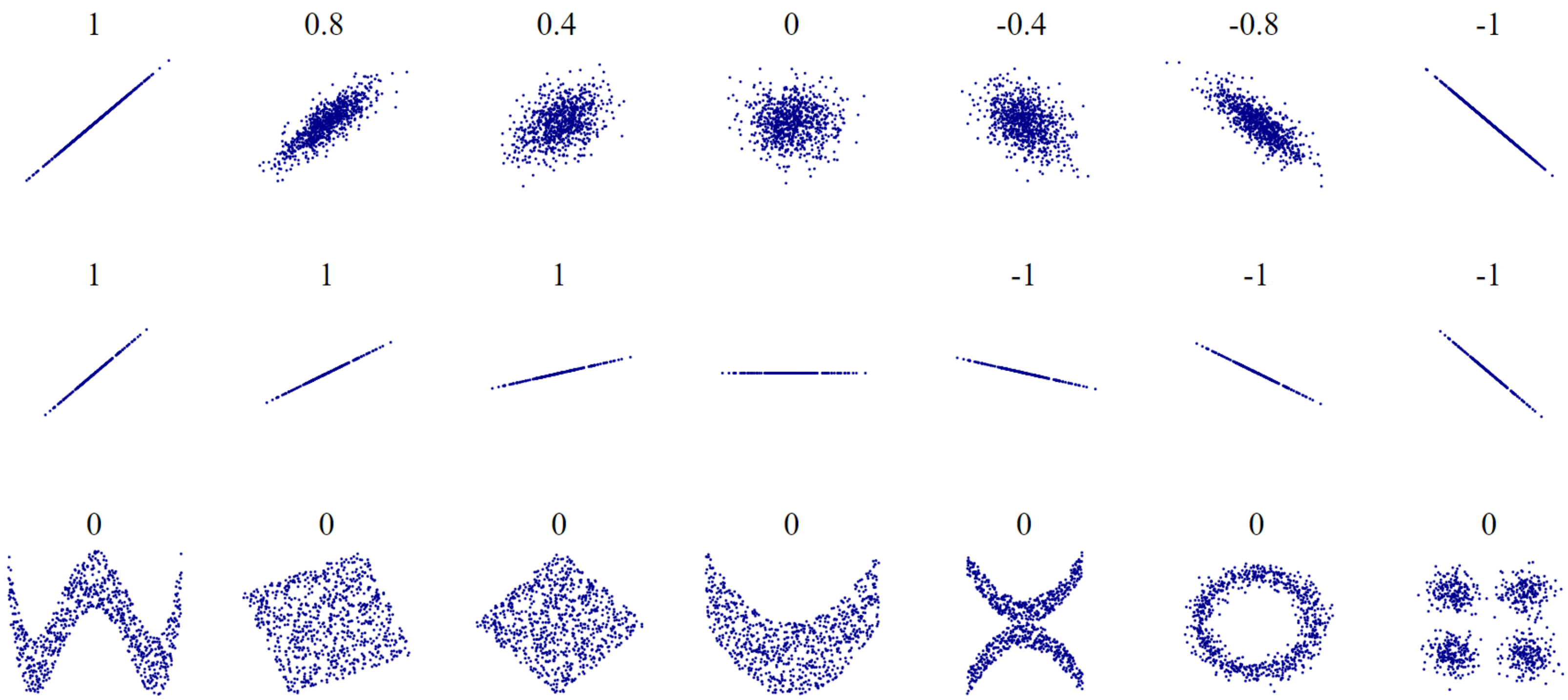
Pearson's correlation coefficient

- If $\boldsymbol{r = -1}$, then there is a perfect negative linear relationship between x and y
- If $\boldsymbol{r = 1}$, then there is a perfect positive linear relationship between x and y
- If $\boldsymbol{r = 0}$, then there is no linear relationship between x and y

All other values of r tell us that the relationship between x and y is not perfect. A reasonable rule is to say that:

- the relationship is **weak** if $\boldsymbol{0 < |r| < 0.5}$
- the relationship is **strong** if $\boldsymbol{0.8 < |r| < 1}$

Pearson's correlation coefficient



By DenisBoigelot, https://commons.wikimedia.org/w/index.php?curid=15165296

Pearson’s correlation coefficient

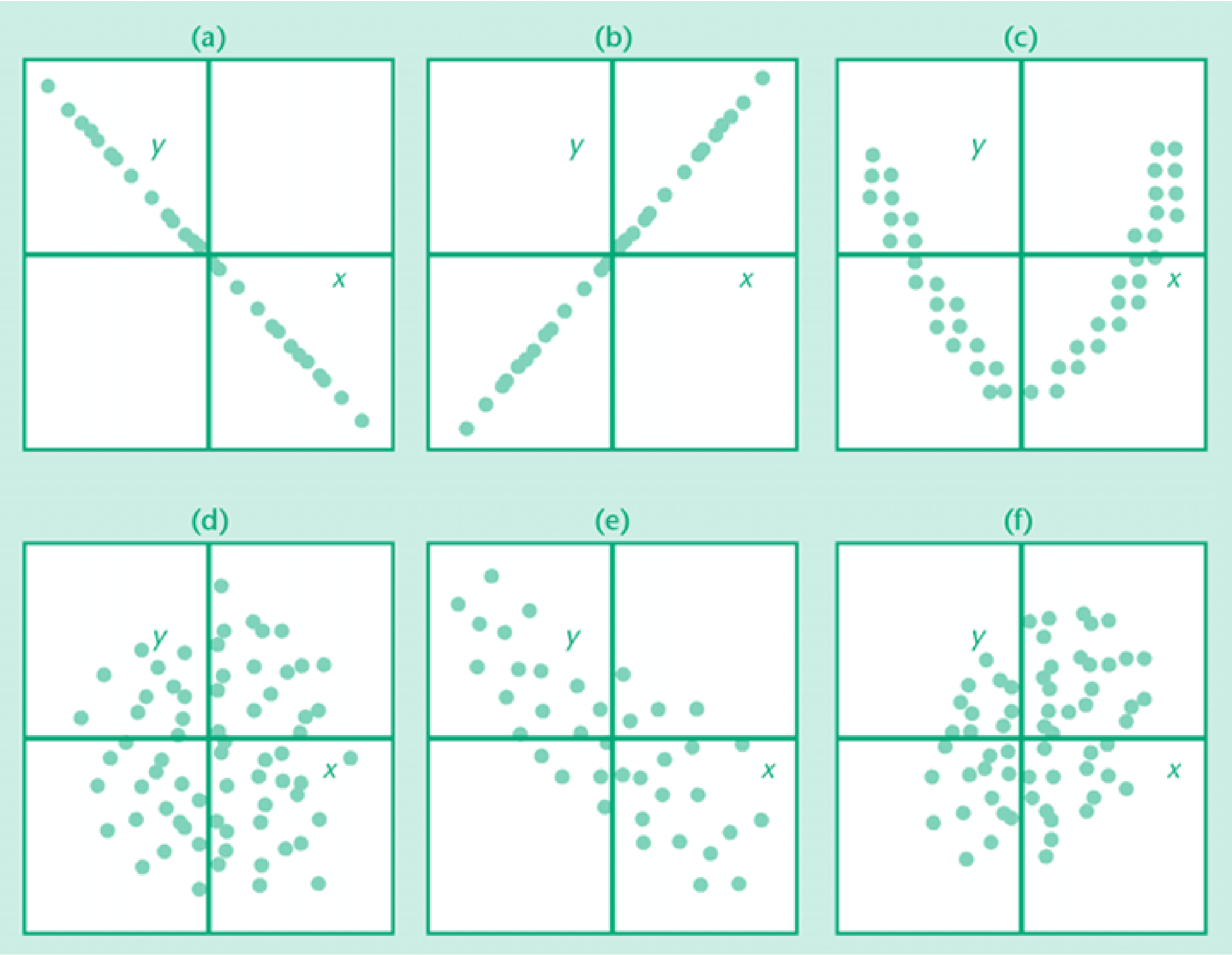
X	Y
0.1	0.2
0.2	0.4
0.3	0.6
0.4	0.8
0.5	1
r	**1.00**

X	Y
0.5	0.2
0.4	0.4
0.3	0.6
0.2	0.8
0.1	1
r	**-1.00**

X	Y
0.5	0.2
0.4	0.4
0.5	0.6
0.4	0.8
0.5	1
r	**0.00**

X	Y
1	0.2
100	0.4
1000	0.6
10000	0.8
100000	1
r	**0.76**

Regression plot (scatterplot)

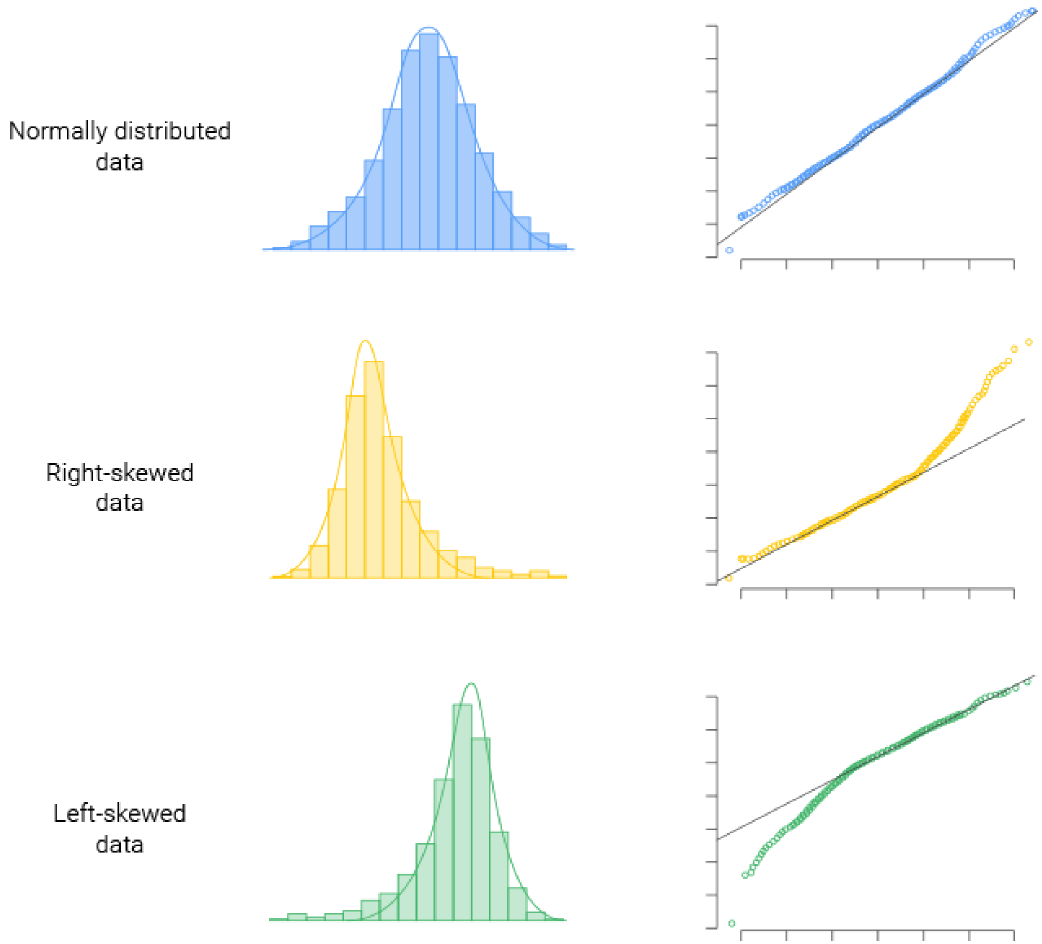


Regression plot (scatterplot)

- (a) and (b): the points fit perfectly on the straight line, so that we have a linear relationship between the two variables:
  - (a): with negative slope, which indicates that as X increases, Y becomes smaller and smaller.
  - (b) with positive slope.
- (c): it is possible to ensure the existence of a strong relationship between the two variables, but it is not a linear relationship.
- (d): the points are completely dispersed, so that there is no type of relationship between the variables.
- (e) and (f): there is some kind of linear relationship between the two variables:
  - (e): a type of linear dependence with a negative slope.
  - (f): linear relationship with positive slope, but not as strong as the previous case.

Q-Q plot

It plots the cumulative distribution functions. If they come from the same distribution (in this case, the normal distribution), the points should fall approximately on a straight line



Parameter estimation

After making the scatter diagram and observing a possible linear relationship between the two variables, the next step will be to find the equation of the line that best fits the cloud of points: the regression line.

The equation of the regression line will be defined by determining the values of the intercept (b_0) and slope (b_1):

$$y = b_0 + b_1 x$$

Parameter estimation

When we use $\widehat{y_i} = b_0 + b_1 x_i$ to predict the actual response y_i, we make a prediction error (or residual error) of size:

$$e_i = y_i - \widehat{y_i} = y_i - (b_0 + b_1 x_i)$$

The "best fitting line" will be the one that minimizes differences between observed and predicted data (ordinary least squares criterion):

$$L = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

Parameter estimation

We have to calculate b_0 and b_1 for the equation of the line that minimizes the sum of the squared prediction errors:

by applying derivatives with respect to b_0 and b_1, and setting them equal to 0

$$\frac{\partial L}{\partial b_0} = 0 \text{ and } \frac{\partial L}{\partial b_1} = 0$$

we obtain

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \big/ {n-1}}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \big/ {n-1}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

S_{xy} is the covariance of observations (x_i, y_i)

S_x^2 is the variance of observation x_i

Parameter estimation

Because the formulas for b_0 and b_1 are derived using the least squares criterion, the resulting equation $\widehat{y_i} = b_0 + b_1 x_i$ is referred to as the least squares regression line (or least squares line)

Note that the least squares line passes through the point $(\bar{x}, \bar{y})$, since when $x = \bar{x}$, then

$$y = b_0 + b_1\bar{x} = \bar{y} - b_1\bar{x} + b_1\bar{x} = \bar{y}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Parameter estimation

From now on, we will write the regression line as follows:

$$\hat{y} = \beta_0 + \beta_1 x$$

where the parameters of the line β_0 and β_1 are given by:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

Interpreting the slope

Meaning:

the slope β_1 represents the expected mean change in the response variable for each unit of change in the predictor variable

An example: $$weight = 0.979009\ height - 96.1121$$

A positive slope value indicating that weight increases with height at a rate of 0.979 kg per centimeter

$height_1 = 1 \rightarrow weight_1 = -95.133091$ $height_2 = 2 \rightarrow weight_2 = -94.154082$

$$weight_2 - weight_1 = -94.154082 - (-95.133091) = 0.979009$$

Interpreting the slope

- if $\beta_1 = 0$, then there is no relationship between the variables

$$\hat{y} = \beta_0 + \beta_1 x = \beta_0 + 0 \cdot x = \beta_0$$

$$\Downarrow$$

(horizontal “no relationship” line in the regression plot)

Interpreting the intercept

Meaning:

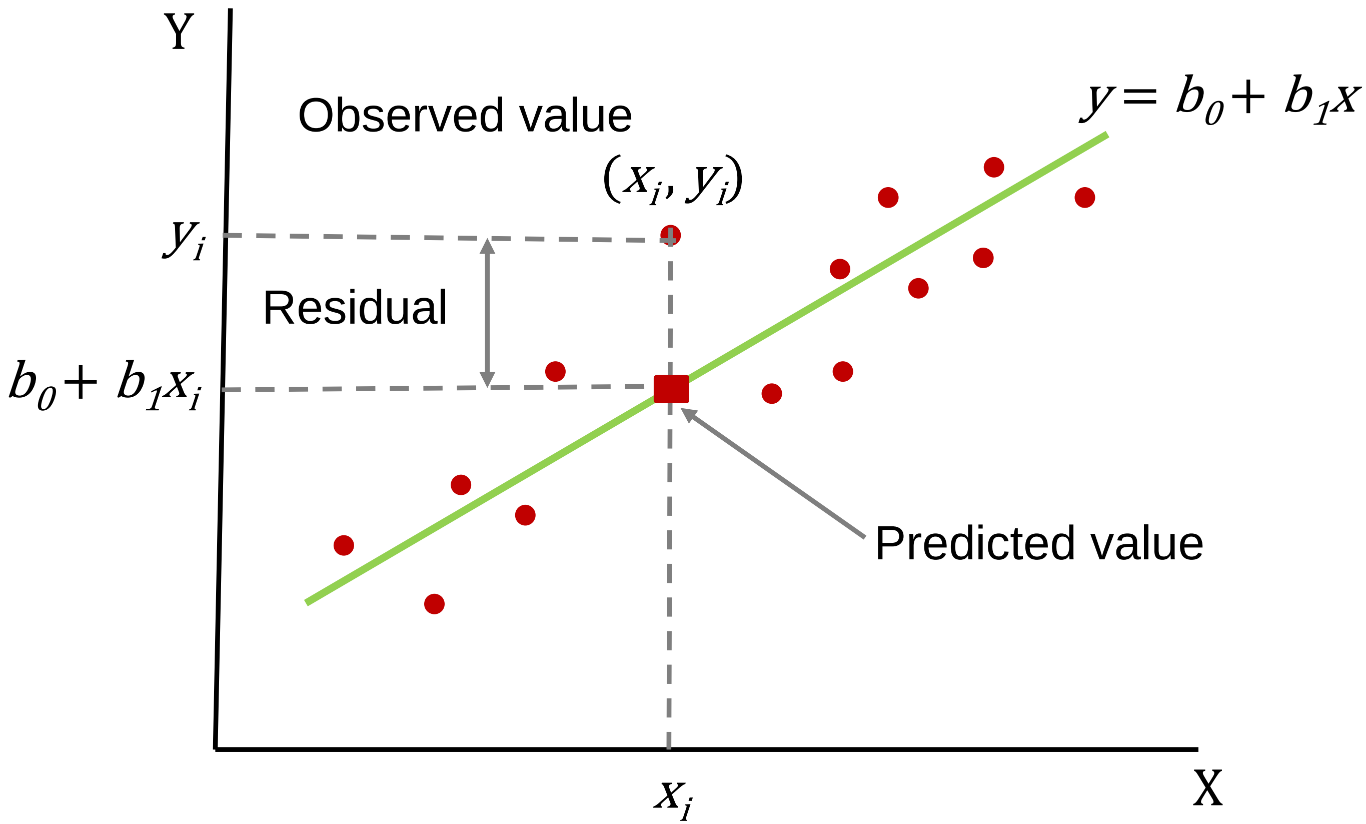
the intercept β_0 only makes sense when the predictor variable can equals 0, Then, it is simply the expected value of the response variable at that value

An example where the intercept has no intrinsic meaning:

$$weight = 0.979009\ height - 96.1121$$

a person who is 0 cm tall is predicted to weigh -96.1121 kg!

Interpreting the residual error



An example

latitude predicts mortality from skin cancer

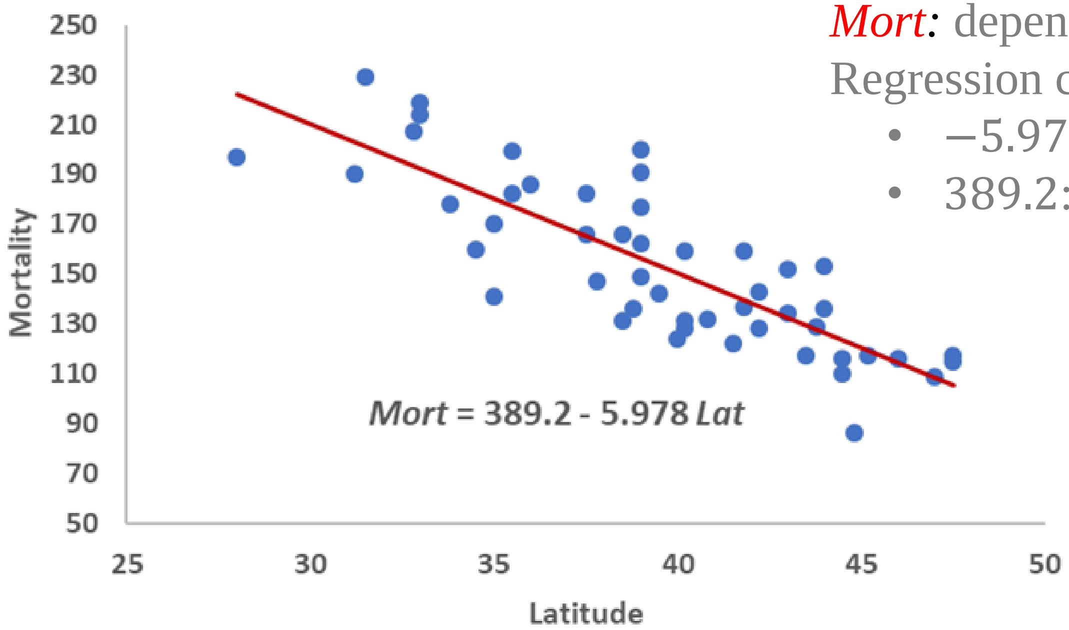
Regression line: an example

Linear function:

$$\boldsymbol{Mort} = \mathbf{389.2} - \mathbf{5.978}\ \boldsymbol{Lat}$$

Lat: independent (predictor) variable
Mort: dependent (response) variable
Regression coefficients:

- -5.978: slope of the line
- 389.2: intercept of the line



Making predictions: an example

Once we have obtained the "estimated regression coefficients" β_0 and β_1, we can predict future responses

- a common use of the estimated regression line:

$$\hat{y}_i = 389.2 - 5.978x_i$$

- predict (mean) mortality of a state at 38 degrees north latitude:

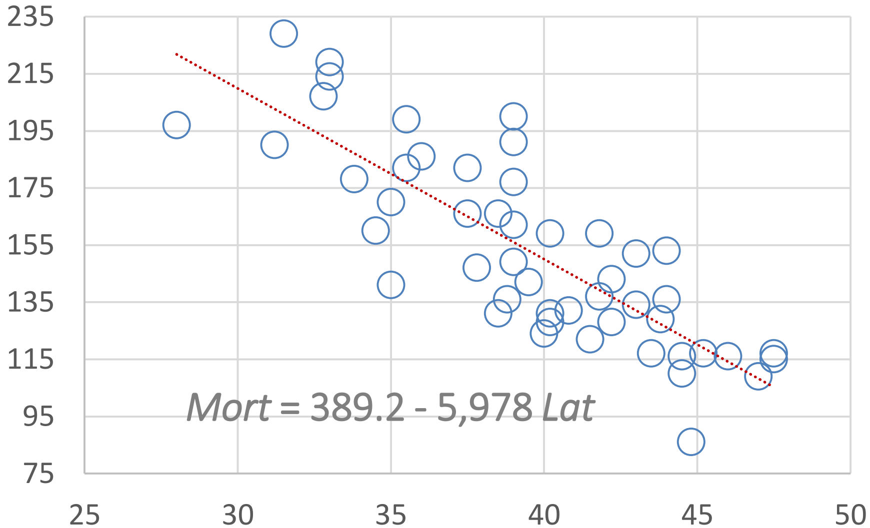
$$\hat{y}_i = 389.2 - (5.978 \times 38) = 132.2$$

Making predictions: an example

$$Mort = 389.2 - 5.978\ Lat$$

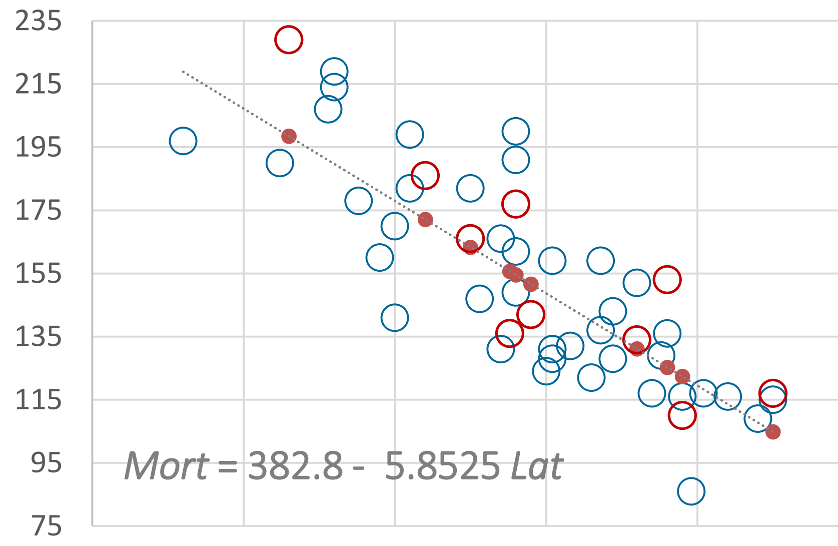
State	latitude (predictor var.)	mortality (response var.)	mortality' (prediction)	residual error
Florida	28,0	197	221,8	-24,8
Texas	31,5	229	200,9	28,1
California	37,5	182	165,0	17,0
Washington, DC	39,0	177	156,1	21,0
New York	43,0	152	132,2	19,8
South Dakota	44,8	86	121,4	-35,4
Minnesota	46,0	116	114,2	1,8

Making predictions: an example



regression equation considering the data of all available states

regression equation considering the first 39 states in alphabetical order + prediction (•) over the remaining 10 states + observed data (○)



Residual error: an example

residual error (or prediction error):

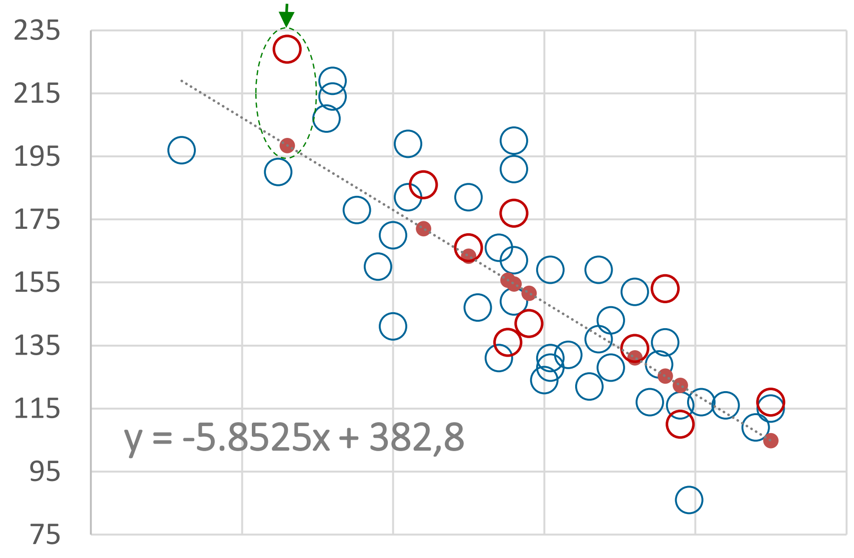
$$e_i = y_i - \hat{y}_i$$

residual error for Texas (prediction error)

$$x_i = 31{,}5, \;\; y_i = 229$$

$$\hat{y}_i = 198.45$$

$$e_i = 229 - 198.45 = 30.55$$



Validation: the quality of the fit

given a linear function inferred from observed data **(a sample)** ...

- is there a good fit to the observed data?
  - residual errors $\rightarrow$ residual plot
  - coefficient of determination (or *R*-squared value or R^2)
  - observed data vs. predicted data

- is the inferred model adequate for the general problem?
  - hypothesis test for the population correlation coefficient

Validation: the quality of the fit

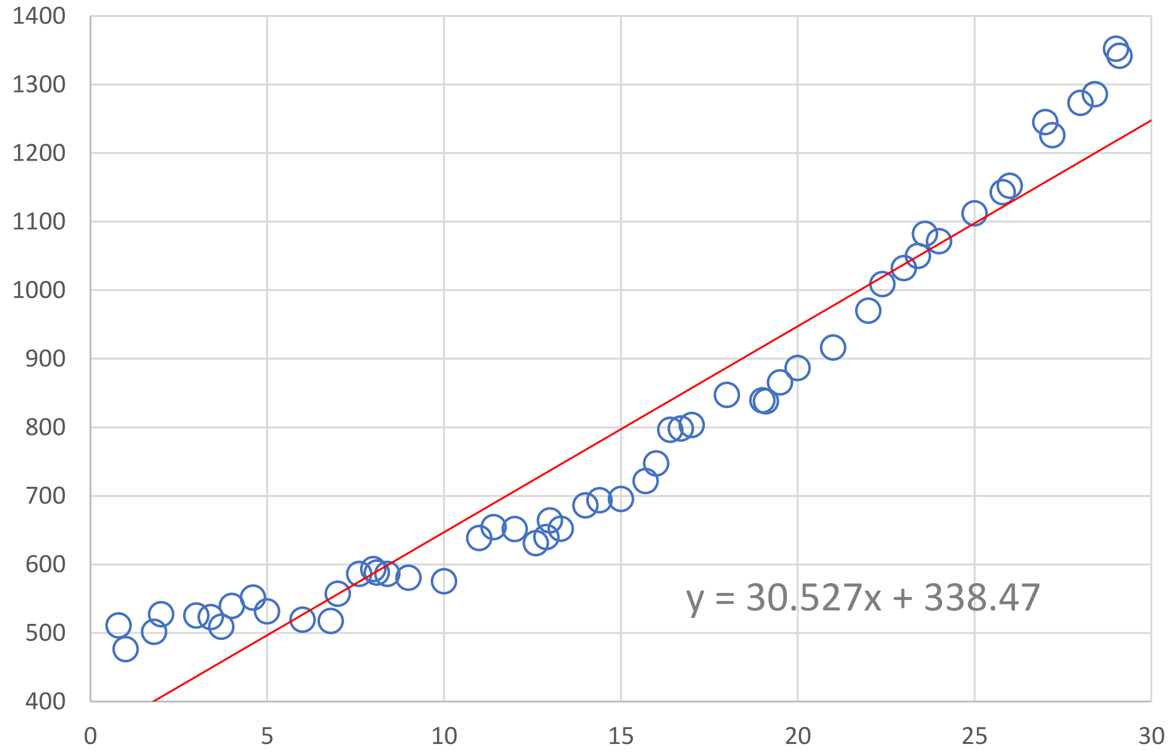
Does the linear function fit the data well?

Is it suitable for the observed distribution?

data generation:

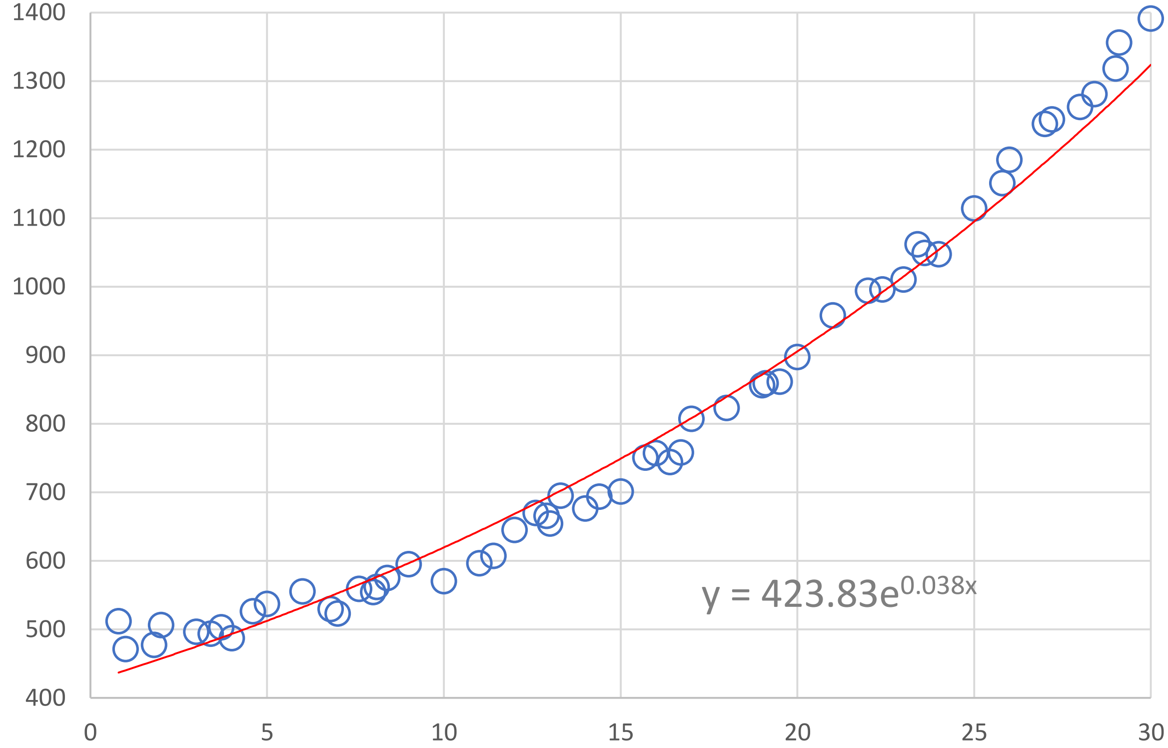
$$y = x^2 + 500 + m$$

where m is a random number between -30 and 30



Validation: the quality of the fit

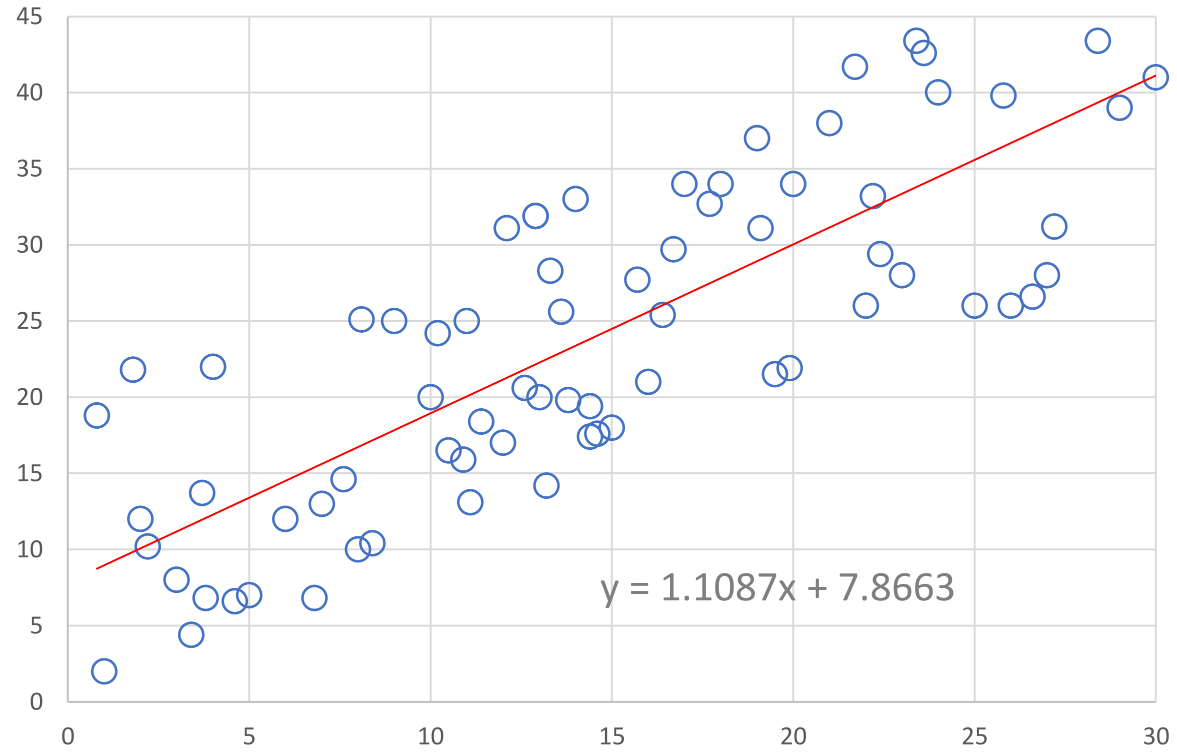
exponential function, more appropriate and better fitted than the linear function



Validation: the quality of the fit

Does the linear function fit the data well?

Is it suitable for the observed distribution?



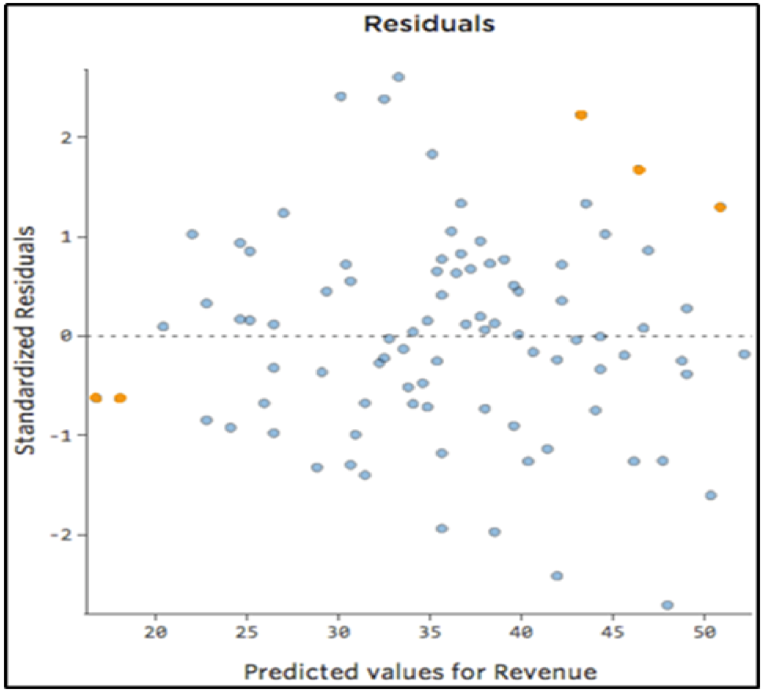
data generation:

$$y = x + 10 + m$$

where m is a random number between -10 and 10

Residual plot

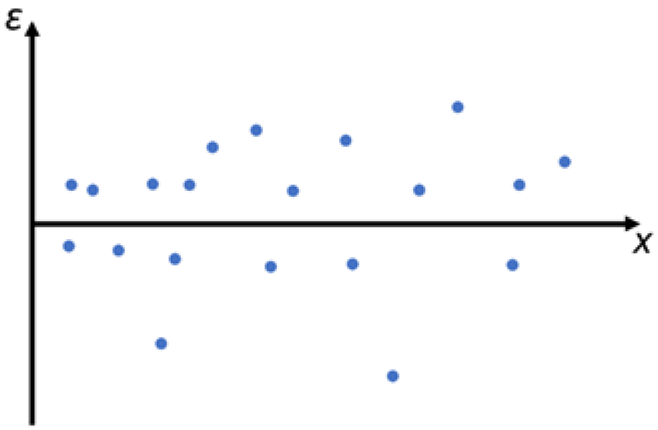
residual errors can be analysed using residual plots: the residual values (e_i) on the y-axis and the predicted values ($\hat{y}_i$) on the x-axis

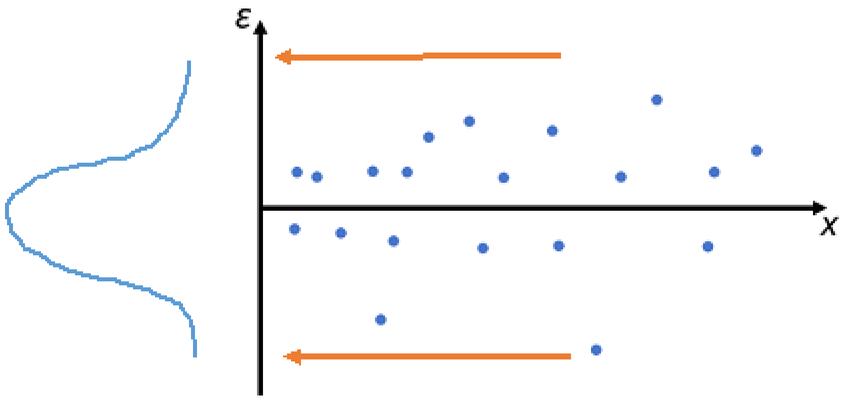


If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate

Residual plot

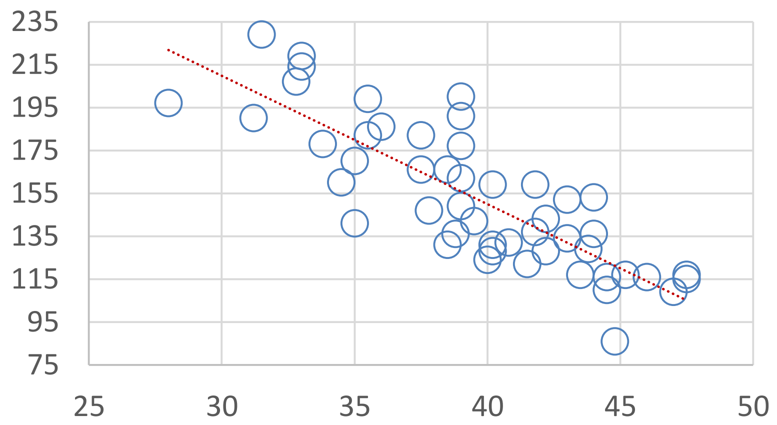
If we project all the residual values onto the y-axis, we end up with a normally distributed curve. This satisfies the assumption that the residuals of a regression model are independent and normally distributed





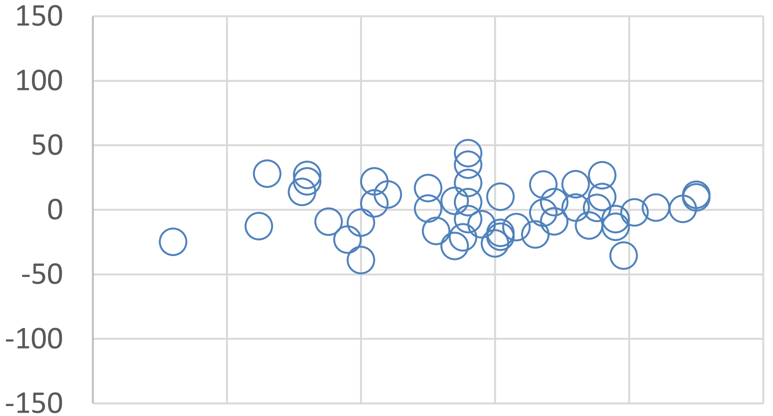
Residual plot

regression plot



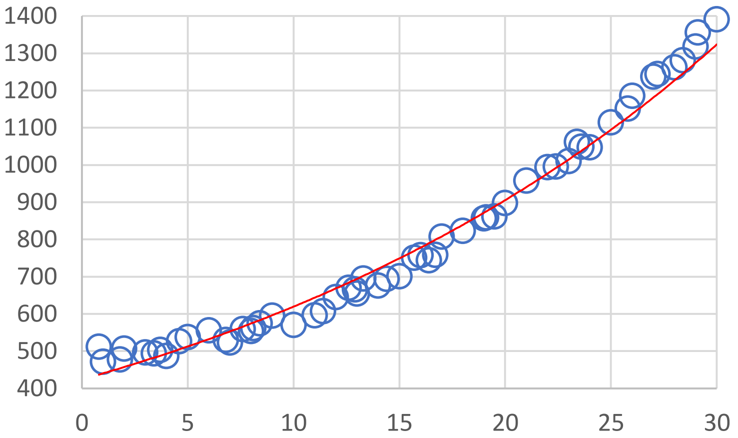
residual plot

$error \sim N(0, \sigma^2)$



Residual plot

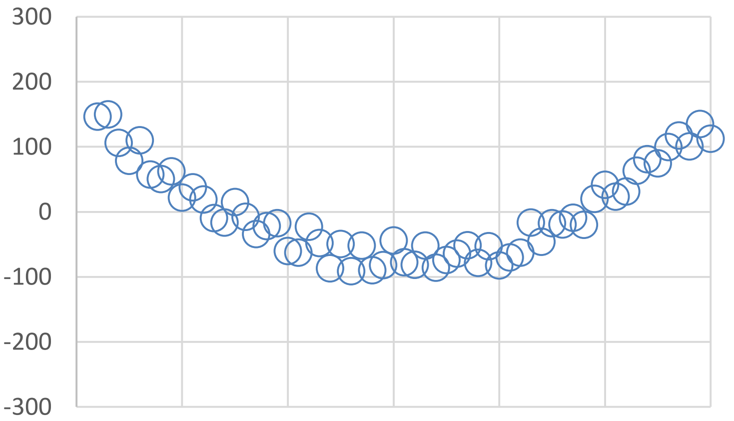
regression plot



residual plot

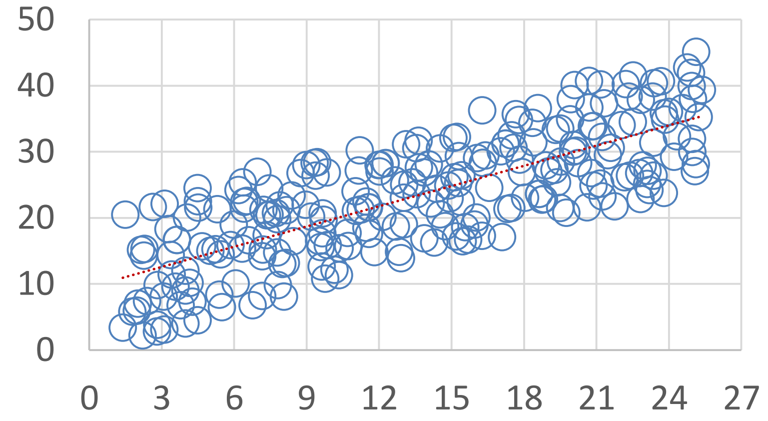
non-random error

$error \neq 0$



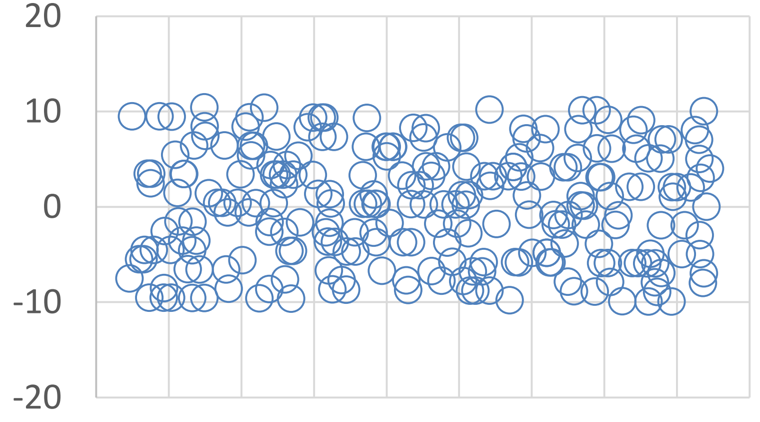
Residual plot

regression plot



residual plot

- random error
- random deviations
- $error \approx 0$



Coefficient of determination (R^2)

given ...

$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, *regression sum of squares*

it quantifies how far the estimated regression line, $\hat{y}_i$, is from the sample mean $\bar{y}$ (horizontal "no relationship" line)

$SSE = \sum_{i=1}^{n}(e_i)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, *error sum of squares*

it quantifies how much the data points, y_i, vary around the estimated regression line, $\hat{y}_i$

$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$, *total sum of squares*

it quantifies how much the data points, y_i, vary around their mean, $\bar{y}$

Coefficient of determination (R^2)

assuming that SST = SSR + SSE ...

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- R^2 is a proportion, its value ranges between 0 and 1
- R^2 measures the proportion of variation in the dependent variable explained by the independent variable
- R^2 indicates how close the data is to the regression line: the closer it is to 1, the better the fit
- R^2 does not indicate whether the regression model is adequate; you can get small values with a good model, and vice versa

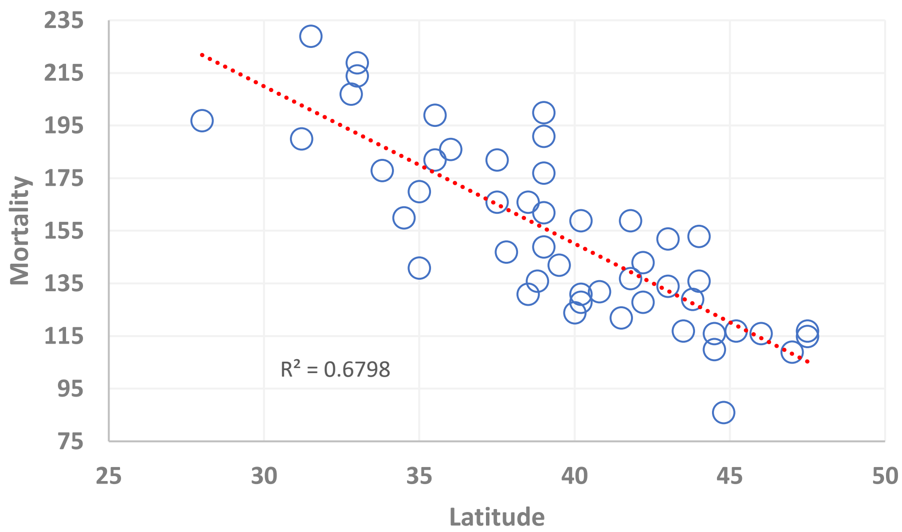
Coefficient of determination (R^2)

- if $R^2 = 1$, all of the data points fall perfectly on the regression line. The response variable can be perfectly explained without error by the predictor variable
  - The residuals are 0 and so is the sum of their squares: $SSR = SST$
- if $R^2 = 0$, the estimated regression line is perfectly horizontal. The response variable cannot be explained by the predictor variable at all
  - the sum of residuals is maximum and we have $SSE = SST$

Coefficient of determination (R^2)

interpretation of R^2

$R^2 \times 100$ percent of the variance in y is 'explained by' the variation in the predictor variable x



68% of the variance in skin cancer mortality is due to or explained by latitude

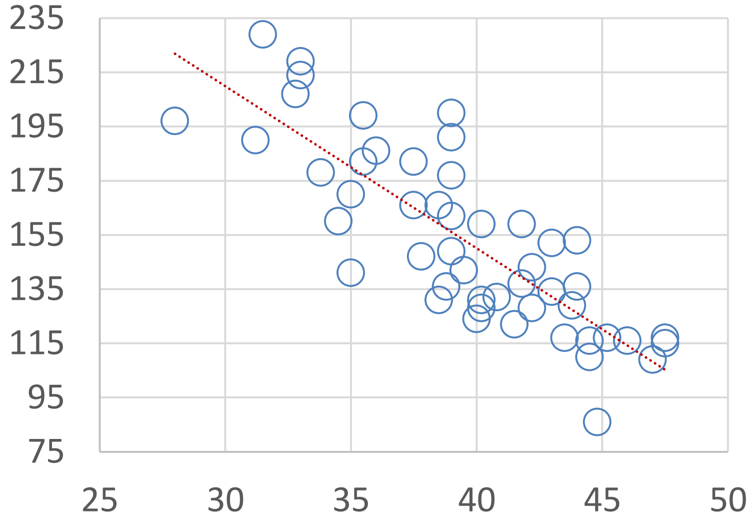
Coefficient of determination (R^2)

Relationship between R^2 and r:

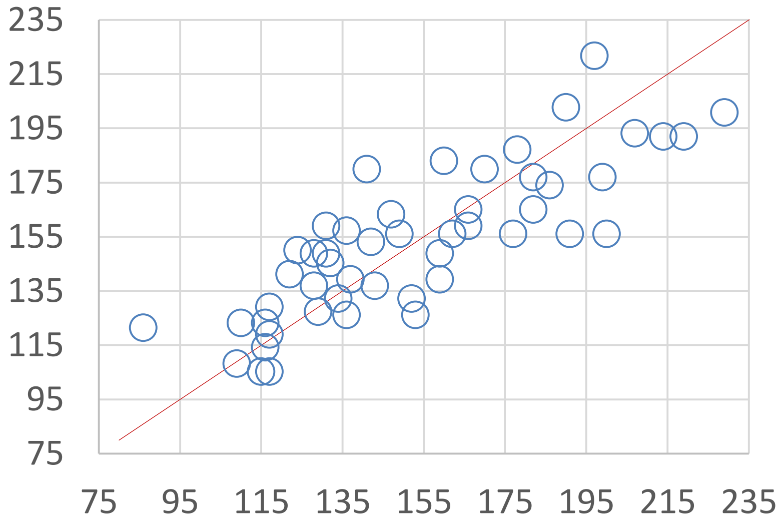
- in simple linear regression, $\boldsymbol{R^2 = r^2}$
  - this relationship helps us understand why we have considered a value of $r = 0.5$ to be weak. This value will represent $R^2 = 0.25$, that is, the regression model only explains 25% of the variability of the observations!
  - r gives us more information than R^2, since the sign of r tells us whether the relationship is positive or negative. With the value of r we can always calculate the value of R^2, but conversely the value of the sign will always remain indeterminate unless we know the slope of the line

Observed data vs. predicted data

regression plot



observed data (x-axis)
vs.
predictions (y-axis)



Hypothesis testing

the correlation coefficient r and the coefficient of determination R^2 summarize the strength of a linear relationship in samples only

- if we obtained a different sample of observations (x_i, y_i), we could obtain different r and R^2 values and different regression lines $\rightarrow$ potentially different conclusions

we have to draw conclusions about populations, not just samples

- so, we have to conduct a **hypothesis test (*t*-test) to see if the population slope $\boldsymbol{\beta_1}$ is significant**

Note that the intercept β_0 determines the average value of the variable Y for a value of X equal to zero. Since it does not always have a realistic interpretation in the context of the problem, we only make statistical inference about the slope

Hypothesis testing

t-test allows validating the linear relationship between the predictor variable and the response variable

$H_0: \beta_1 = 0$, the null hypothesis

$H_a: \beta_1 \neq 0$, the alternative hypothesis

intuition

if $\beta_1 = 0$, there is not a linear relationship between x and y

if $\beta_1 \neq 0$, there is a significant linear relationship between the variables

objective

to reject the null hypothesis (i.e., the variable x has an influence on the variable y and therefore, there is a linear relationship between the two variables)

Hypothesis testing

Steps for hypothesis testing:

1. specify the null and alternative hypotheses (see previous slide)
2. set a significance level α (typical values 0.01, 0.05)
3. construct a statistic T to test the null hypothesis H_0
4. define a decision rule to reject, or not, the null hypothesis H_0

Hypothesis testing

Steps for hypothesis testing:

1. specify the null and alternative hypotheses (see previous slide)
2. set a significance level α (typical values 0.01, 0.05)
3. construct a statistic T to test the null hypothesis H_0
4. define a decision rule to reject, or not, the null hypothesis H_0

Hypothesis testing

Steps for hypothesis testing:

1. specify the null and alternative hypotheses (see previous slide)
2. set a significance level α (typical values 0.01, 0.05)
3. construct a statistic T to test the null hypothesis H_0

$$T = \frac{\beta_1}{SE(\beta_1)}$$

where β is the estimated coefficient of the population slope, and

$$SE(\beta_1) = \sqrt{\frac{MSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

is the standard error of the estimated coefficient of the population slope

Hypothesis testing

Steps for hypothesis testing:

1. specify the null and alternative hypotheses (see previous slide)
2. set a significance level α (typical values 0.01, 0.05)
3. construct a statistic T to test the null hypothesis H_0
4. define a decision rule to reject, or not, the null hypothesis H_0
   - T follows a Student's t-distribution with $n-2$ degrees of freedom, where n is the number of data points (– 2 because we have two parameters, β_0 and β_1)
   - we calculate the p-value:

$$P(|t_{n-2}| > T) = 2P(t_{n-2} > T)$$

   - we reject the null hypothesis H_0 if $p\text{-value} \leq \alpha$

Hypothesis testing

interpreting the result of the hypothesis test

- the p-value indicates how likely is it to get such an extreme T value if the null hypothesis H_0 is true
- if $p\text{-value} \leq \alpha$ means that there is sufficient evidence at the level α to conclude that there is a linear relationship in the population between the predictor and response variables $\rightarrow$ we reject the null hypothesis H_0
- rejecting H_0 entails accepting H_a $\rightarrow$ there is a significant linear relationship between the variables
- given T and $n-2$, the p-value is obtained from the Student's t-distribution tables or from [some web sites]