Linear Regression

Department of Computer Languages and Systems

Introduction

- Linear regressions allows you to model 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 model the relationship between 2 variables by fitting a linear function to our observed data (x_i, y_i):

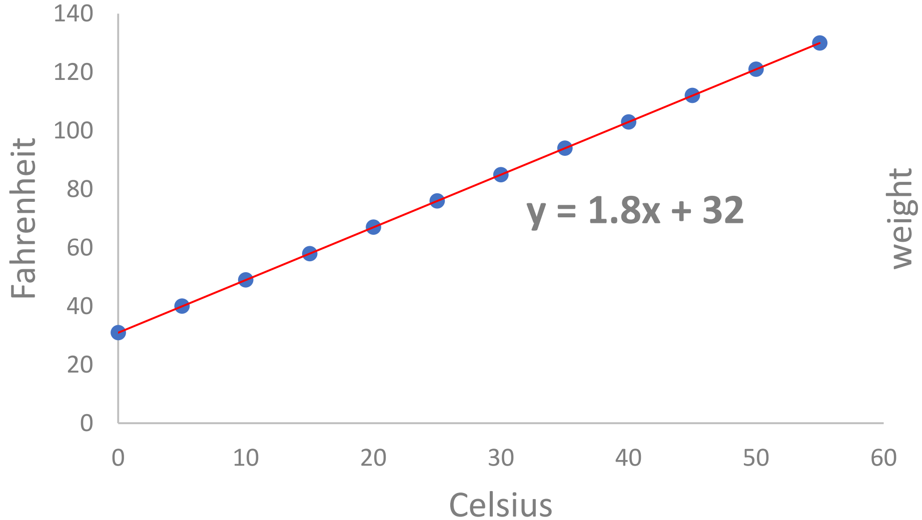
$$y = b_0 + b_1 x$$

  – This is a line where y is the dependent 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

Types of relationships

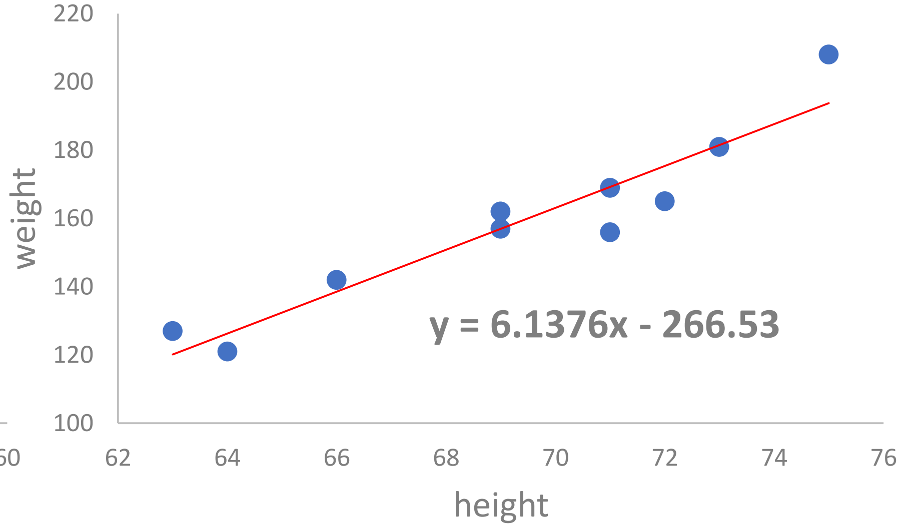
Deterministic relationship



the observed (x, y) data points fall directly on a line:

$$Fahr = 32 + 1.8\ Cels$$

Statistical relationship



the relationship between the variables is not perfect:

$$weight = 6.1376\ height - 266.53$$

Fundamentals of simple linear regression

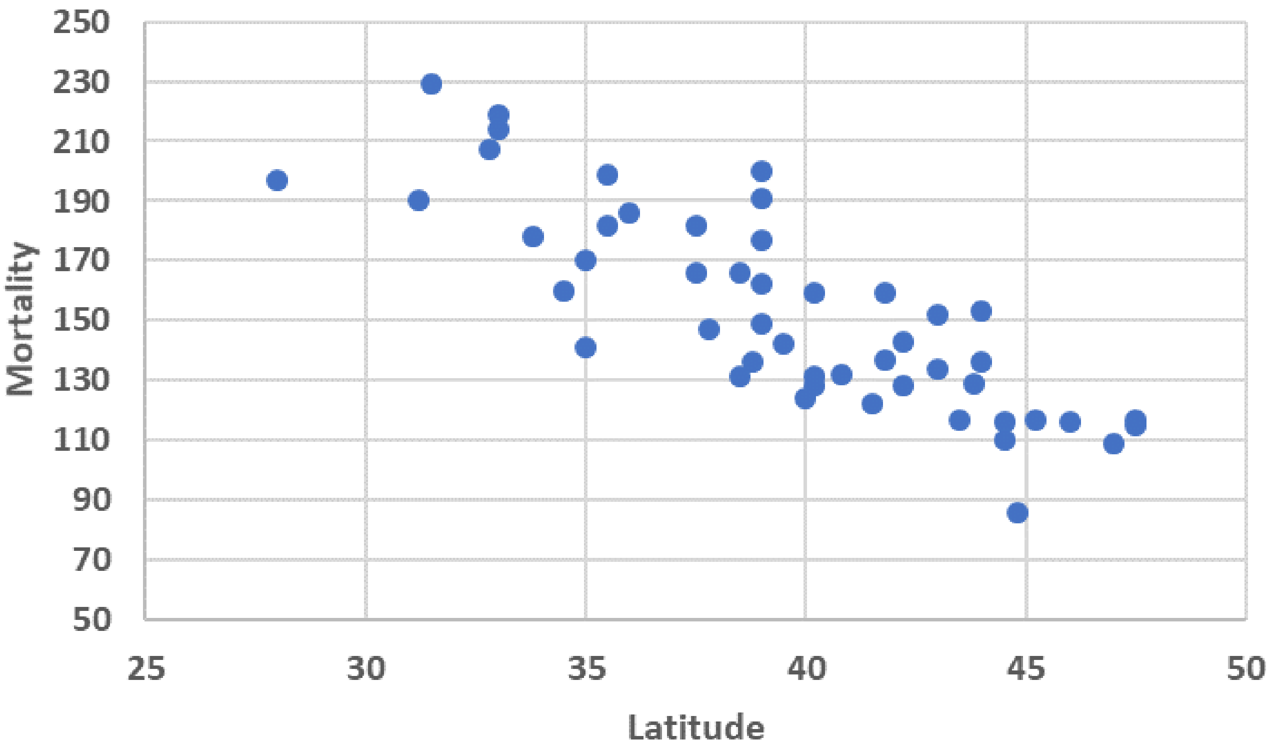
- Hypothesis (assumption):

  the response variable is a linear combination of parameters (regression coefficients) and the predictor variable

- Preliminary assessment of the strength of the hypothesis:
  - regression plot (scatterplot)
  - linear correlation coefficient

Regression plot

Scatterplot: latitude vs mortality from skin cancer



Linear correlation coefficient

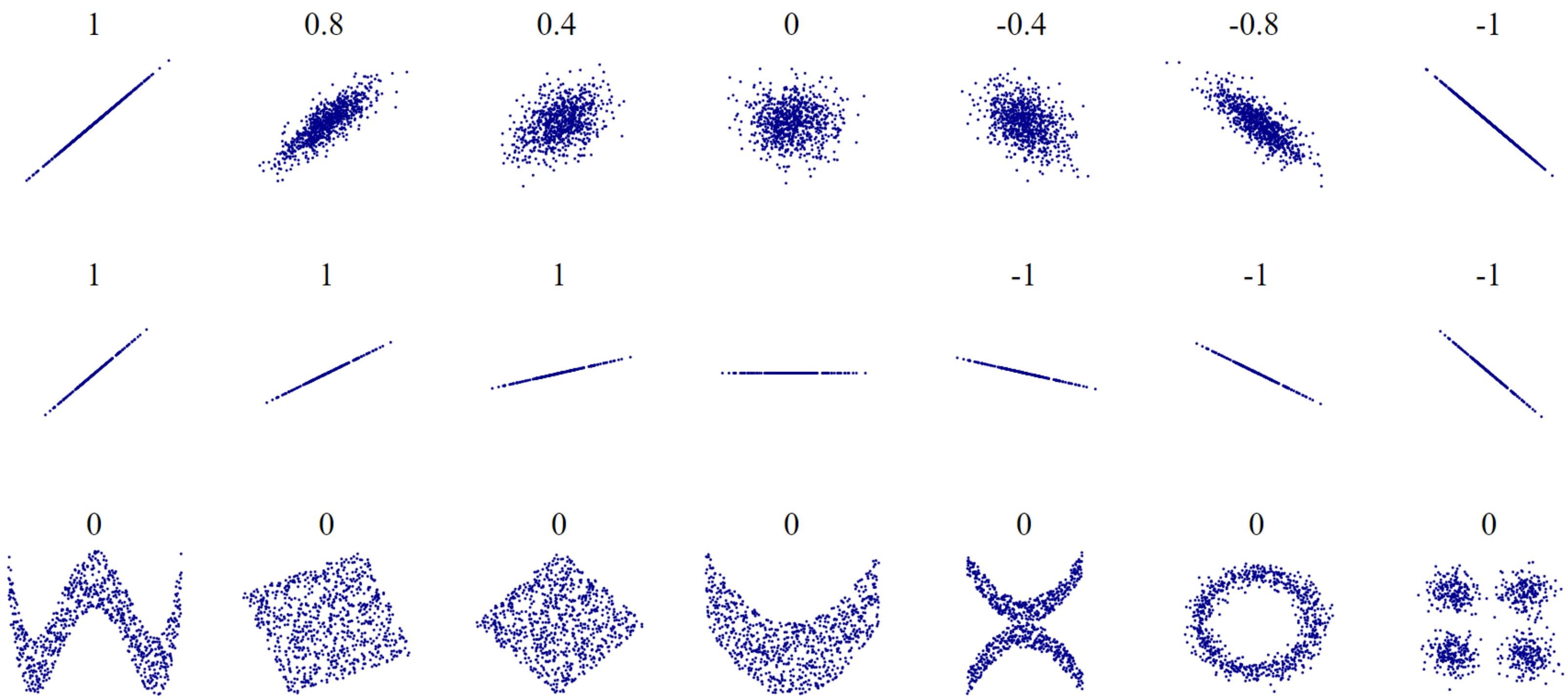
Pearson's correlation coefficient: a measure of linear correlation between two variables:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- 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

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**

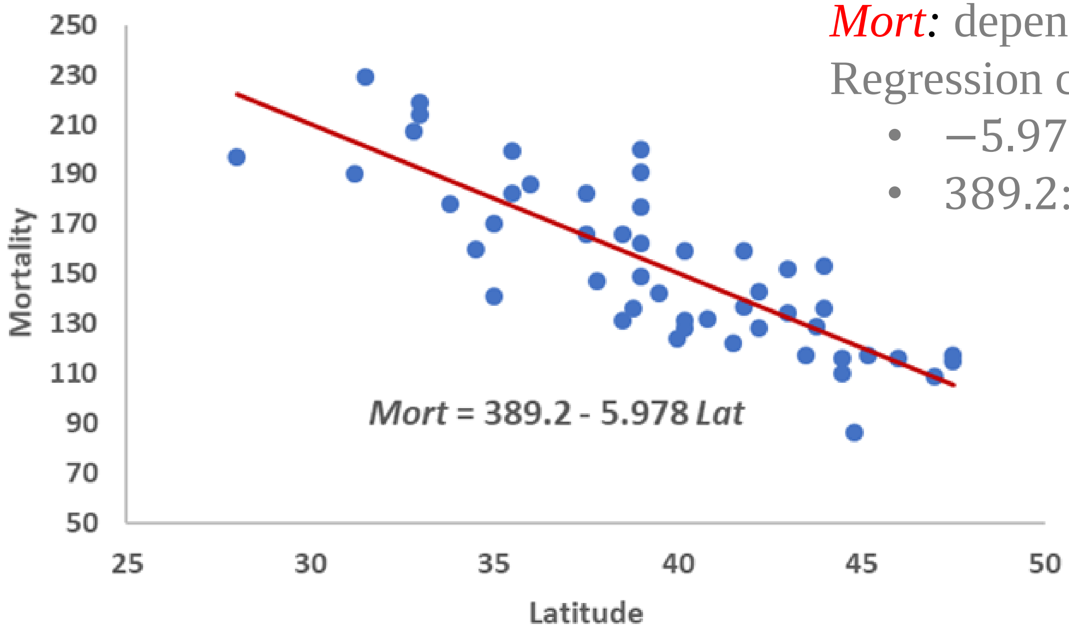
X	Y
0.5	0.2
0.4	0.4
0.5	0.6
0.4	0.8
0.5	1
r	**0.00**

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

An example

latitude predicts mortality from skin cancer

Regression equation



Linear function:

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

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

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

Looking for the “best fitting line"

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:

$$e_i = y_i - \widehat{y_i}$$

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

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

Looking for the “best fitting line"

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

$$\frac{\partial L}{\partial b_0} = 0, \quad \frac{\partial L}{\partial b_1} = 0$$

- and setting to 0, we obtain

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

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

Looking for the “best fitting line"

Because the formulas for b_0 and b_1 are derived using the least squares criterion, the resulting equation $\hat{y}_i = b_0 + b_1 x_i$ 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

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

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

Making predictions

Once we have obtained the "estimated regression coefficients" b_0 and b_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:

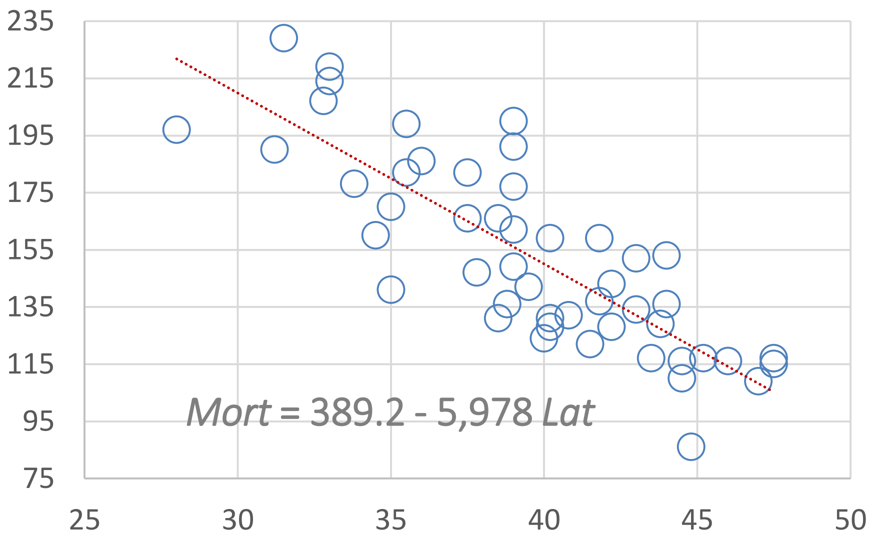
$$\hat{y}_i = 389.2 - (5.978 \times 38) = 132.2$$

Making predictions

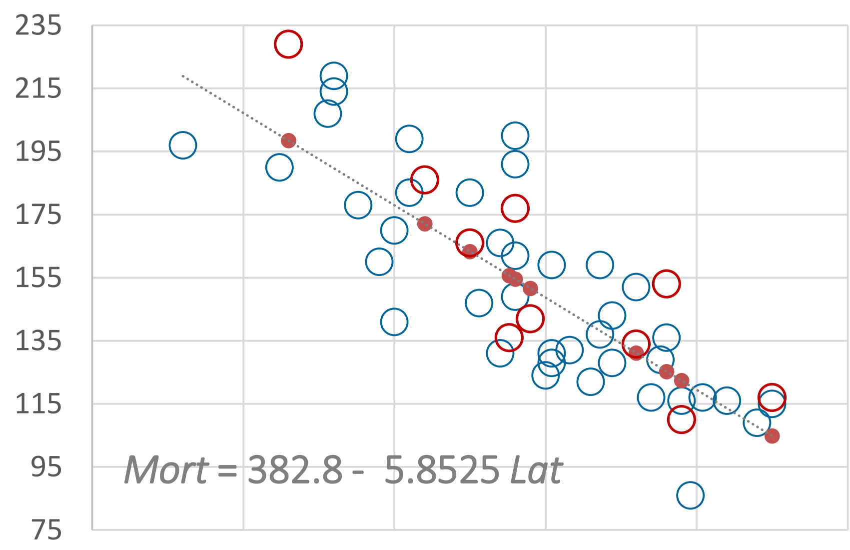
$$Mort = 389.2 - 5.978\ Lat$$

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



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 (○)

Interpreting the slope

Meaning:

the slope b_1 represents the expected mean change in the response variable for each unit of change in the predictor variable

An example: $Mort = 389.2 - 5.978\ Lat$

for each additional degree of latitude, the expected mean mortality from skin cancer is reduced by almost 6 people (per 10 million people)

$Lat_1 = 1 \rightarrow Mort_1 = 383.222$ $Lat_2 = 2 \rightarrow Mort_2 = 377.244$

$Mort_2 - Mort_1 = 377.244 - 383.222 = -5.978$

Interpreting the slope

- if $b_1 = 0$, then there is no relationship between the variables

$$y = b_0 + b_1 x = b_0 + 0 \cdot x = b_0$$

$$\Downarrow$$

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

Interpreting the intercept

Meaning:

the intercept b_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:

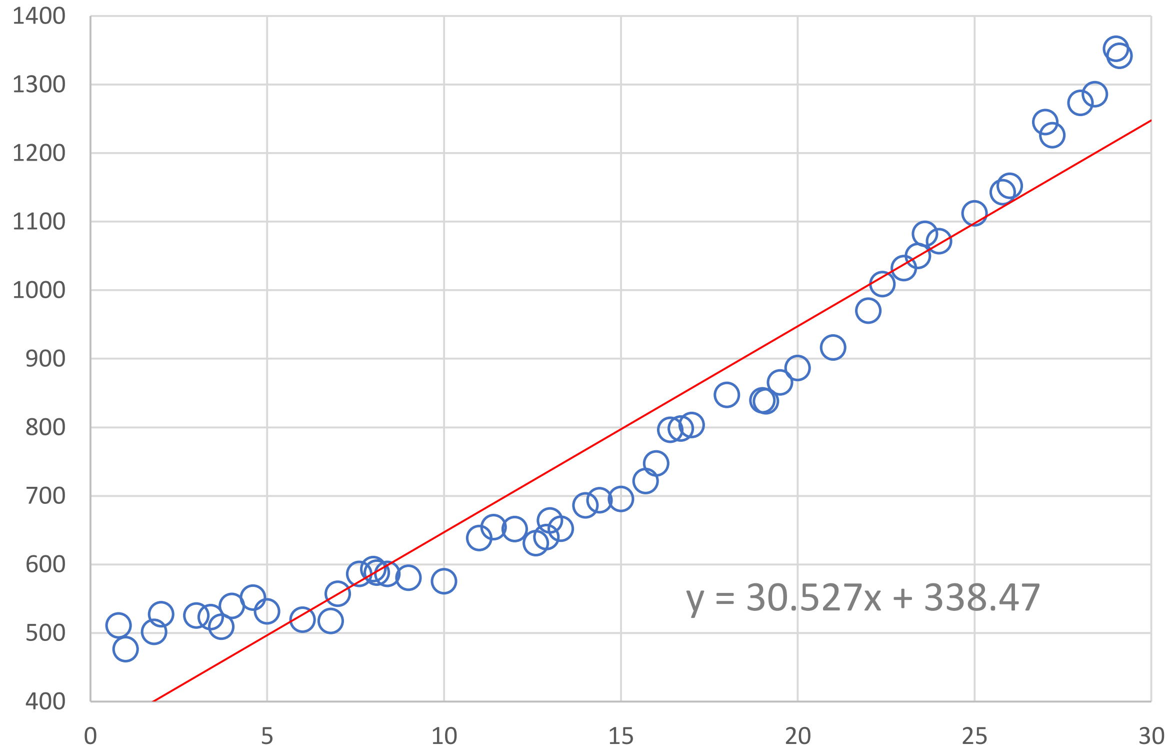
$$Weight = 6.1376\, Height - 266.53$$

a person who is 0 inches tall is predicted to weigh -266.53 pounds!

Validation

Does the linear function fit the data well?

Is it suitable for the observed distribution?



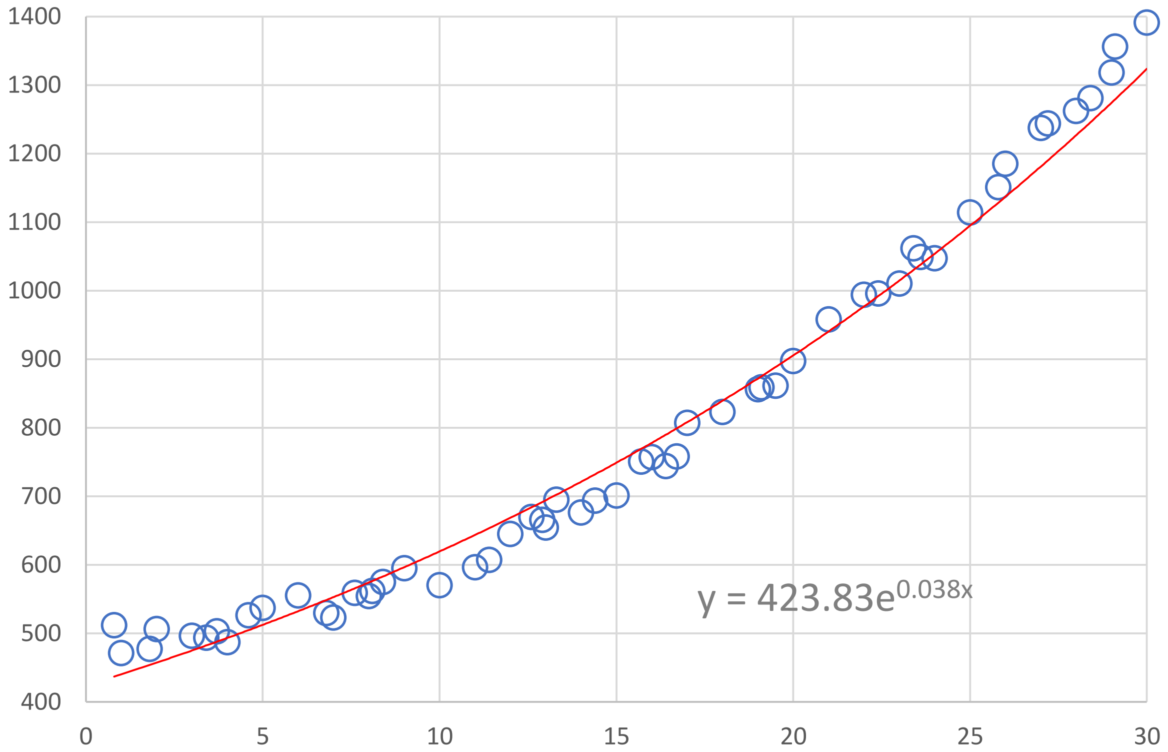
data generation:

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

where m is a random number between -30 and 30

Validation

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



Validation

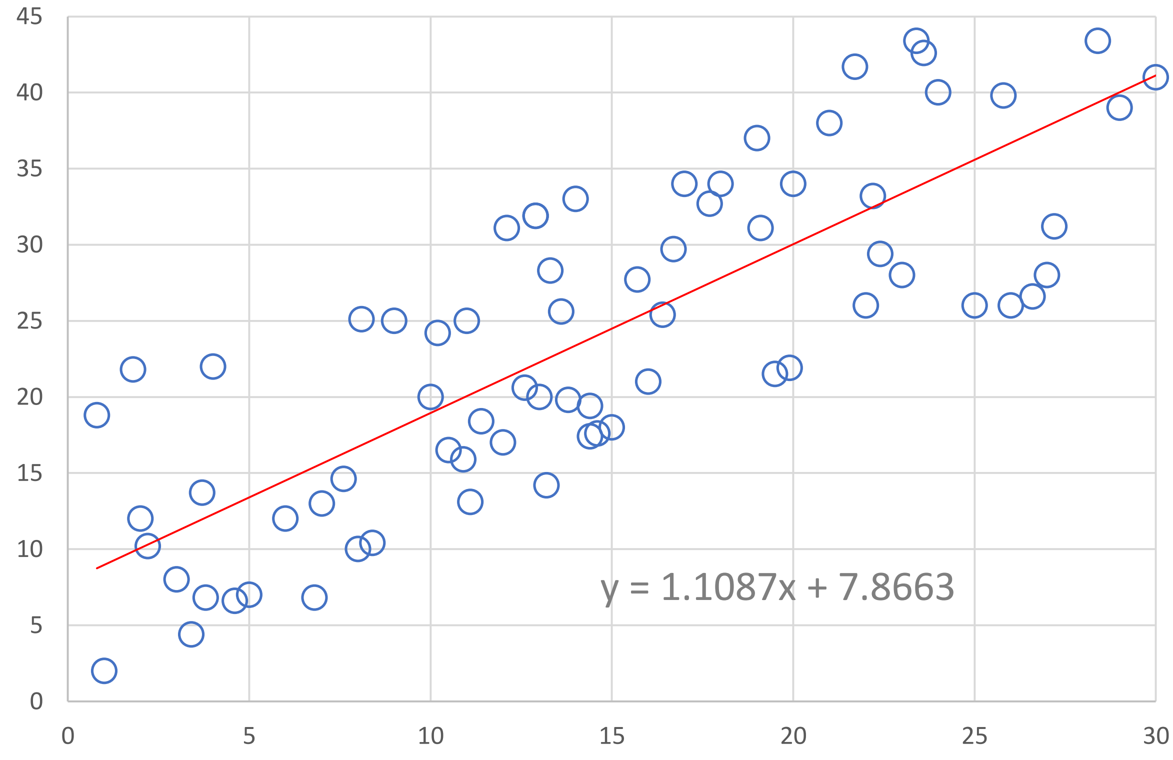
Does the **linear function** fit the data well?

Is it **suitable** for the observed distribution?

data generation:

$$y = x + 10 + m$$

where m is a random number between -10 and 10



Validation

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

- is there a good fit to the observed data?
  - residual errors
  - 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

Residual errors

residual error (or prediction error):

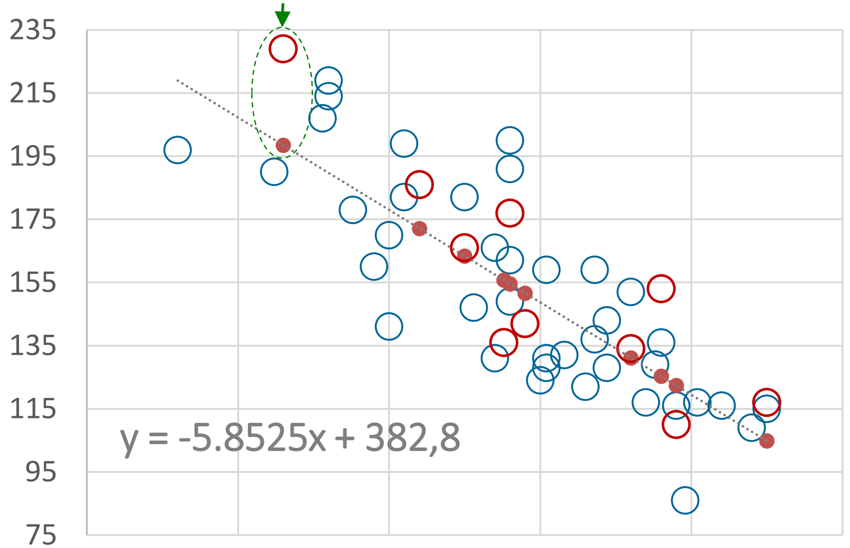
$$e_i = y_i - \hat{y}_i$$

residual error for Texas (prediction error)

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

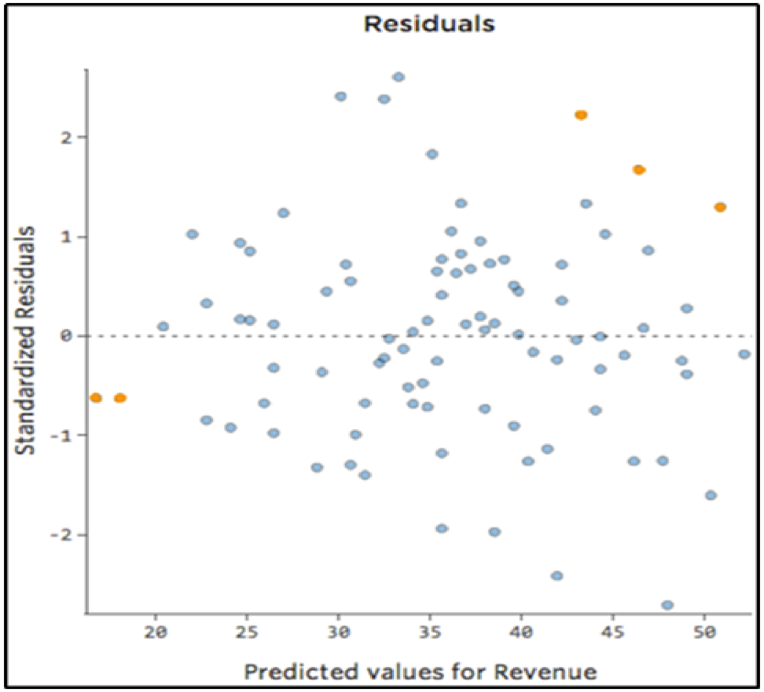
$\hat{y}_i = 198.45$

$e_i = 229 - 198.45 = 30.55$



Residual plot

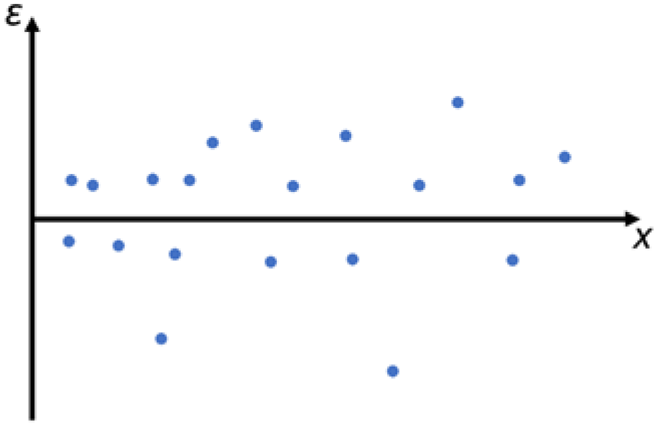
residual errors can be analysed using residual plots: the residual values on the y-axis and the predicted values on the x-axis

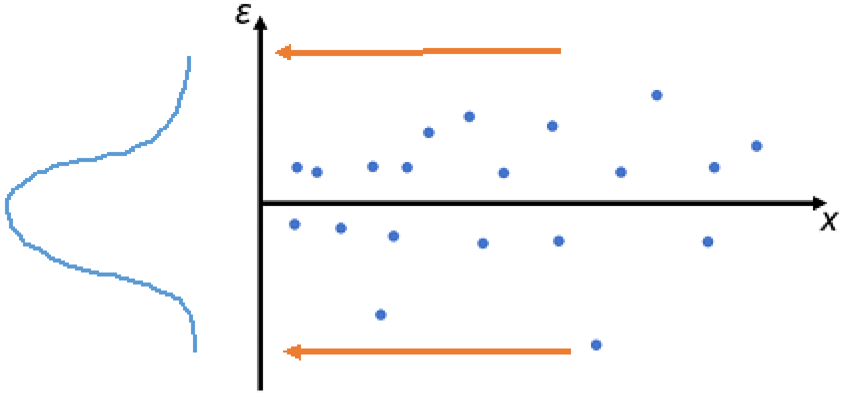


If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate

Residual plot

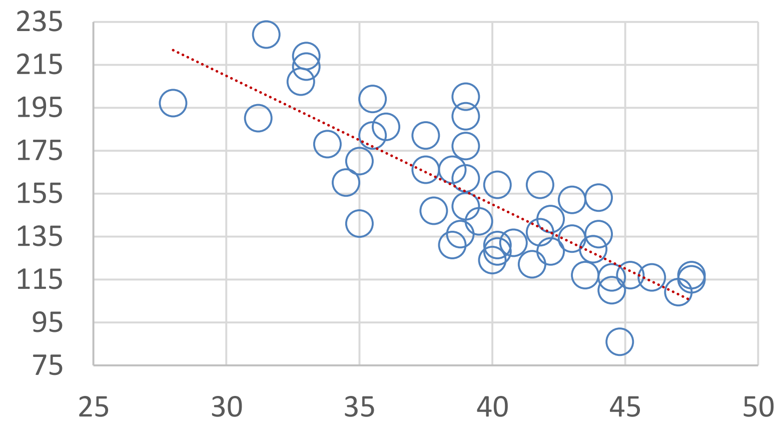
If we project all the residual values onto the y-axis, we end up with a normally distributed curve. This satisfies the assumption that the residuals of a regression model are independent and normally distributed





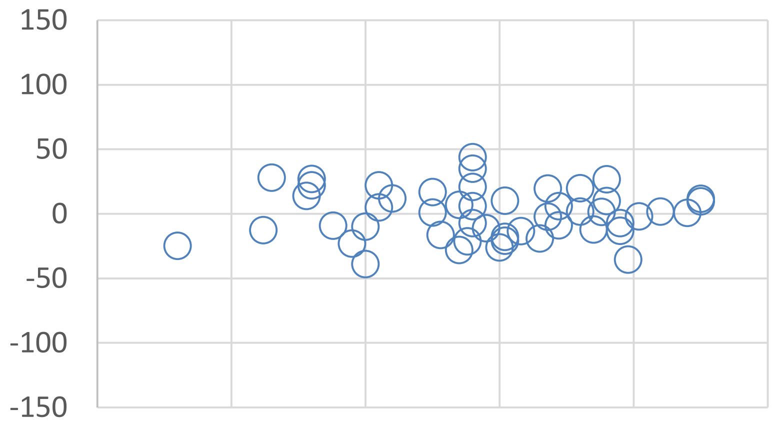
Residual plot

regression plot



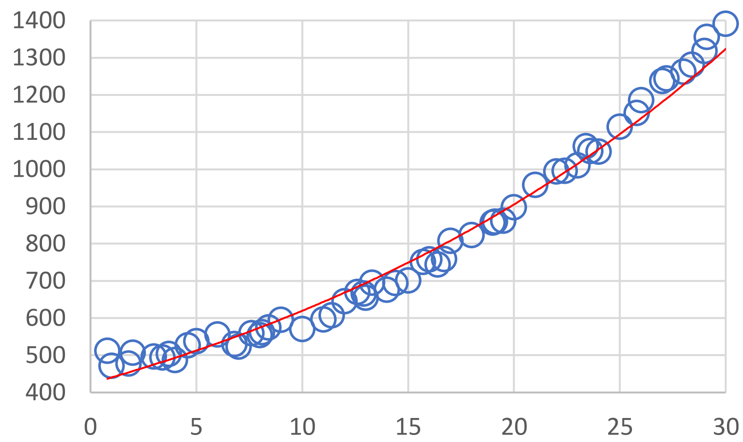
residual plot

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



Residual plot

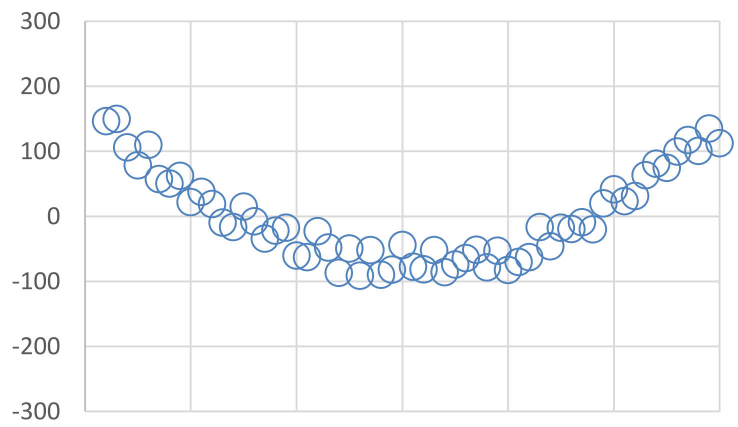
regression plot



residual plot

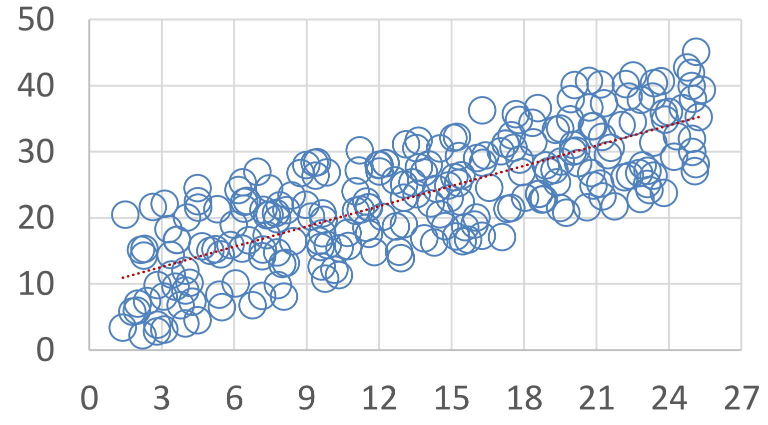
non-random error

$error \neq 0$



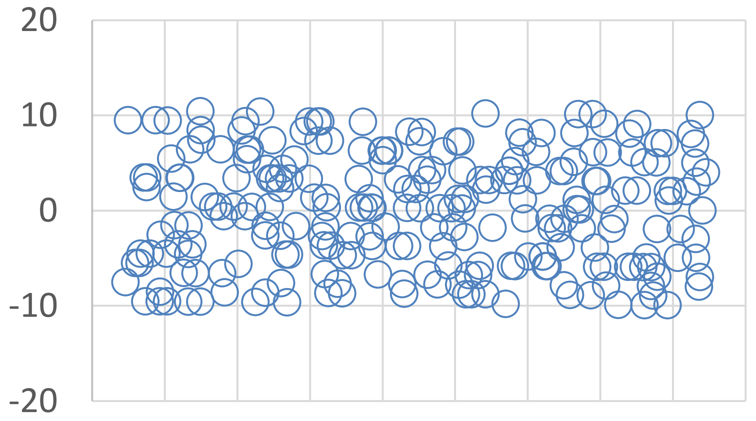
Residual plot

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 ...

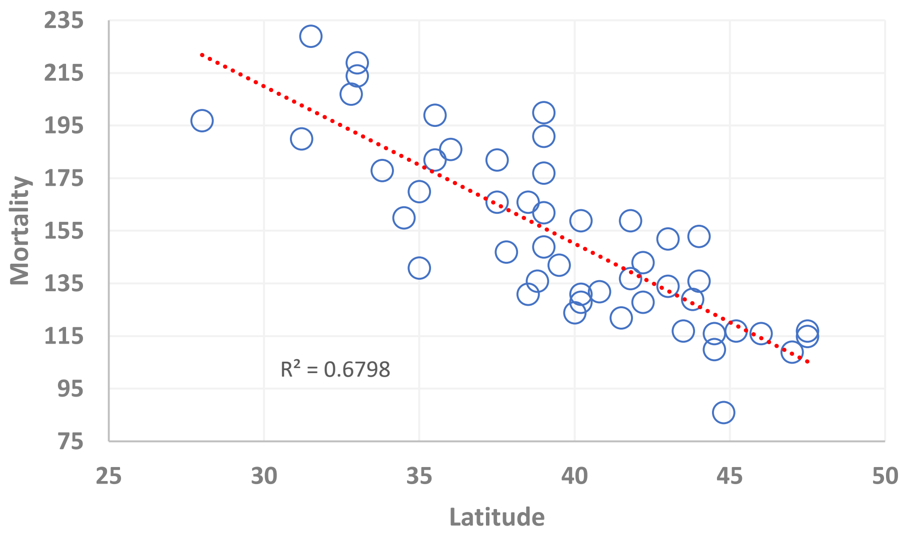
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

- since R^2 is a proportion, its value ranges between 0 and 1
- R^2 indicates how close the data is to the regression line
- 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
- if $R^2 = 0$, the estimated regression line is perfectly horizontal. The response variable cannot be explained by the predictor variable at all

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

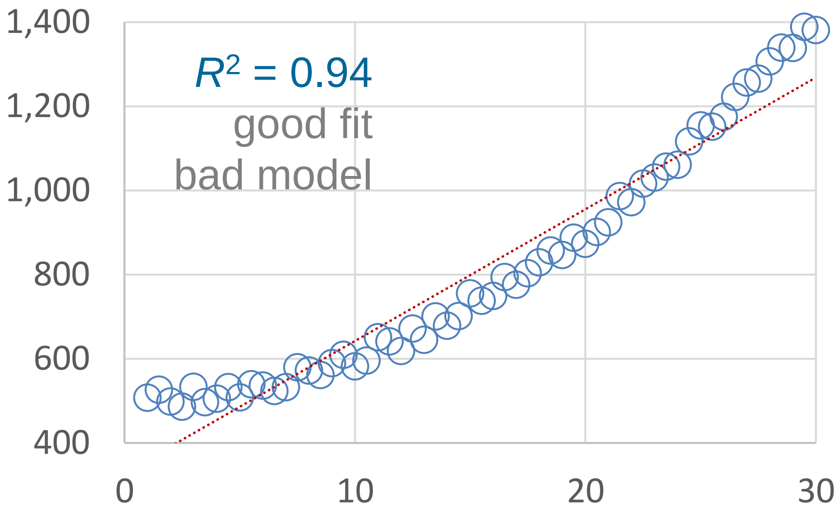


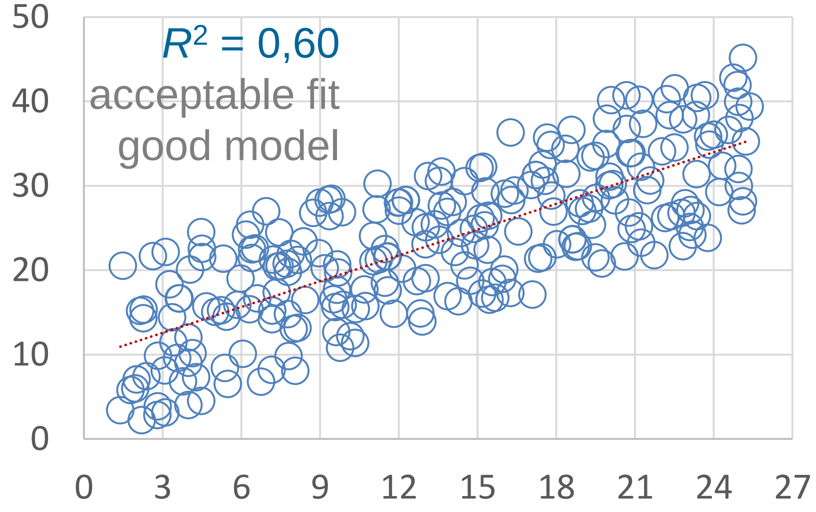
68% of the variance in skin cancer mortality is due to or explained by latitude

Coefficient of determination (R^2)

note that …

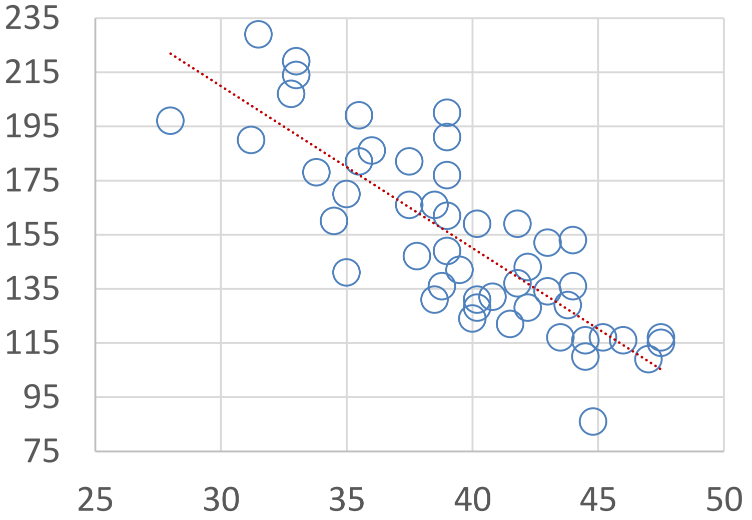
- in general, the larger the value of R^2, the better the fit
- R^2 does NOT indicate whether the regression model is adequate; you can get small values with a good model, and vice versa



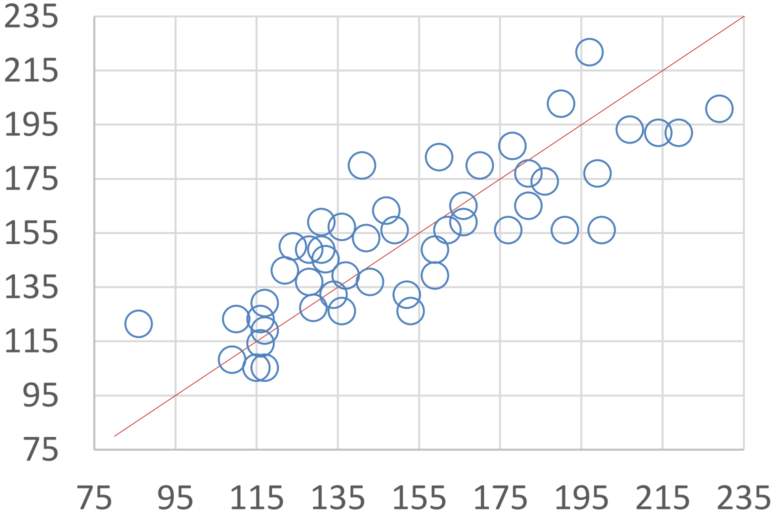


Observed data vs. predicted data

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, we could obtain different correlations and different R^2 values $\rightarrow$ 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**

Hypothesis testing

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

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

$H_a \colon \beta_1 \neq 0$, 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., verify that it is very improbable that $\beta_1 = 0$ in the population (not only in the observed sample)

Hypothesis testing

Steps for hypothesis testing:

1. to specify the null and alternative hypotheses (see previous slide)
2. to construct a statistic to test the null hypothesis H_0
3. to 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
2. construct a statistic to test the null hypothesis H_0

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

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

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

is the standard error of the estimated coefficient

Hypothesis testing

Steps for hypothesis testing:

1. specify the null and alternative hypotheses
2. construct a statistic to test the null hypothesis
3. to define a decision rule to reject, or not, the null hypothesis H_0
   - *T* follows a *t*-distribution with $n-2$ degrees of freedom, where n is the number of data points (– 2 parameters for simple linear regression)
   - we calculate the *p*-value:

     $$P(t > T) + P(t < -T) = 1 - P(T \leq t \leq -T)$$

   - we reject the null hypothesis H_0 if the *p*-value is smaller than the significance level α (e.g., 0.01, 0.05)

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*-value < α 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 *t*-distribution tables or from some web sites

Hypothesis testing (example 1)

$n = 170 \quad \alpha = 0.01$

$SE(\beta) = 0.014976 \qquad T = 60.84459$

$\mathrm{P}(x \le T) = 1 \rightarrow p\text{–value} = 0$

We reject H_0

Hypothesis testing (example 2)

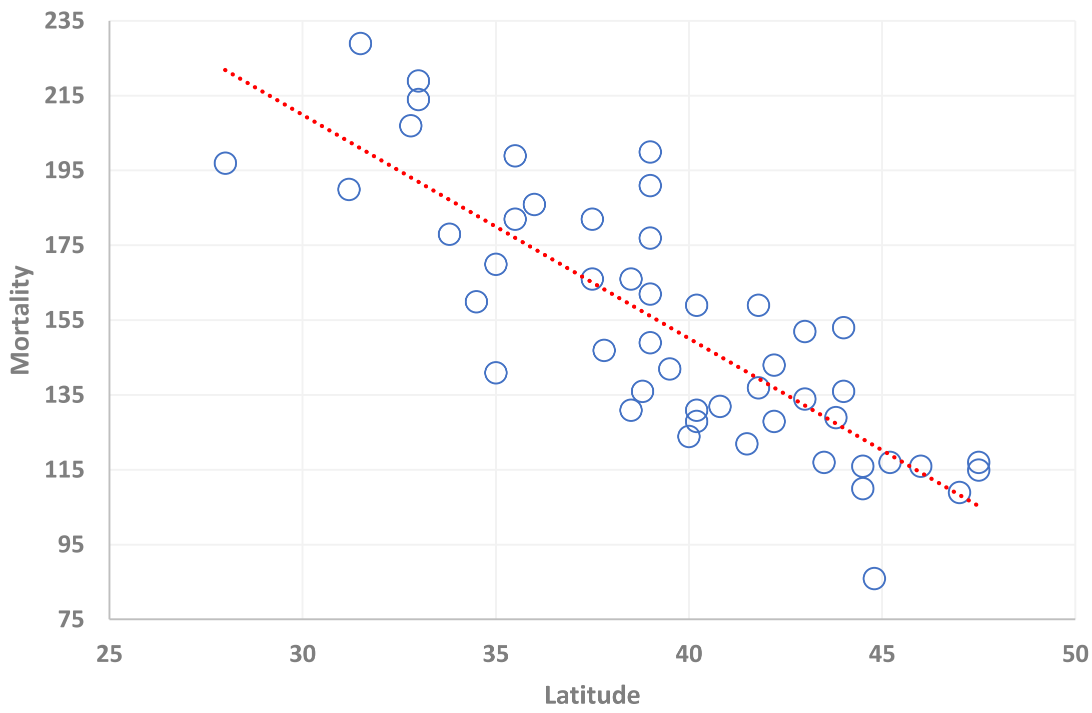
$T = -9.99$

$n - 2 = 47$

$\alpha = 0.01$

p–value $= 0$

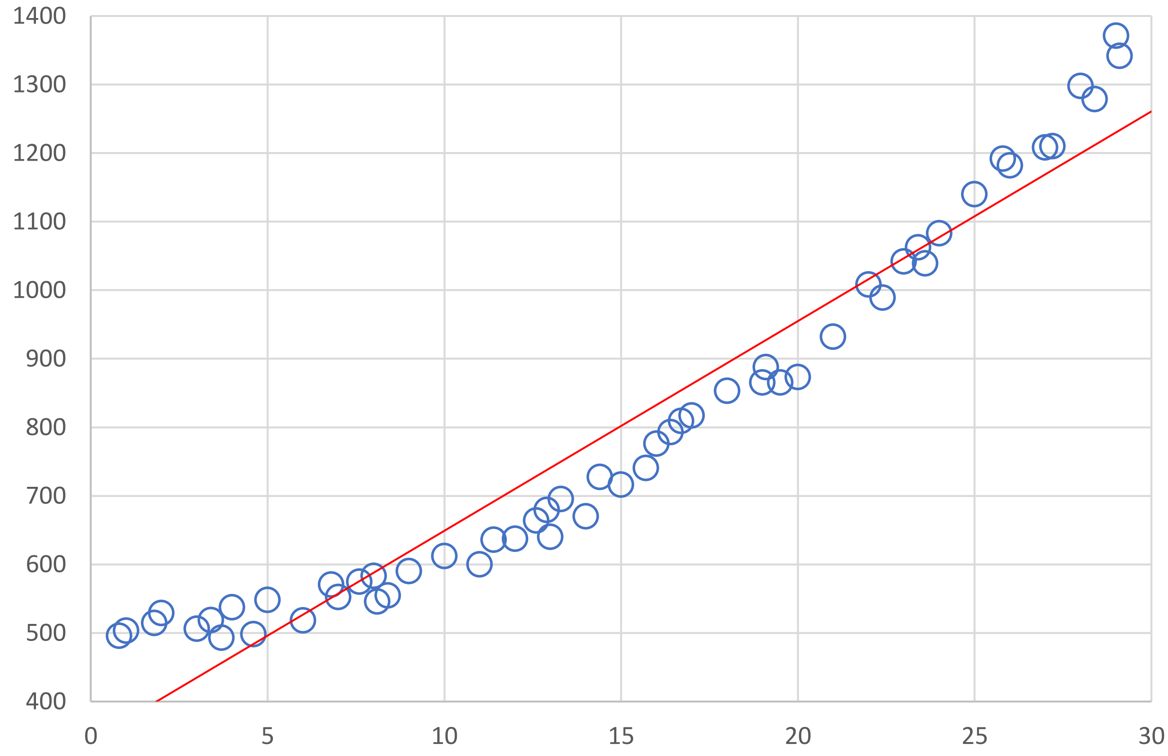
We reject H_0



Hypothesis testing (example 3)

$T = 29.03$
$n - 2 = 57$
$\alpha = 0.01$
p–value $= 0$

We reject H_0



Hypothesis testing (example 4)

Data:

$x \sim N(\mu = 5, \sigma = 2)$
$y \sim N(\mu = 5, \sigma = 2)$

$T = -0.05$
$n - 2 = 198$
$\alpha = 0.01$
p–value $= 0.96$

We do not reject H_0