Statistical Computing
Week 6: 8. November 2019
Regression 1

Regression

- Do variables W and X influence the value of variable Y?
- Can you express Y as a function of W and X, even though Y is observed with variability?
- Can you predict what the value of Y will be for given values of W and X?

Examples

- Rent depends on size of flat/house (floor area)
- A car's fuel consumption depends on its speed
- Income depends on years spent in education

In machine learning:
A major subtopic is *supervised learning*.

Supervised learning models all have an *outcome variable* Y which we want to fit to the available data and obtain predictions for given values of the *predictor variables*.

There are two major types of supervised learning.
Classification: Y is nominal, ordinal or discrete.
Regression: Y is continuous.

Regression is a major topic in machine learning.
Eg. An artificial neural network is a regression model if Y is continuous.

In each of the examples on Slide 1, there is one **predictor variable**, X,* which has an effect on the **outcome variable**, Y. †

Regression function

Each observation of the independent variable y_i is a realisation from the regression function $y = f(x)$, which is a function of the predictor variable x_i. The value of y_i is observed "with error".

Possible reasons for the "error" term are:

- natural variation
- imprecise measurement
- other unobserved variables
- ...

---

*also: independent variable, Ger. Einflussgröße= "influence variable"

†also: dependent variable, Ger. Zielgröße="target variable"

Regression function

Regression function examples

Simple linear	:	$f(x) = b_0 + b_1 x$
Quadratic	:	$f(x) = b_0 + b_1 x + b_2 x^2$
Multiple linear	:	$f(u, v, w) = b_0 + b_1 u + b_2 v + b_3 w$

In the multiple linear regression we have three variables u, v and w that all have a linear influence on y.

Simple linear Regression

The data are observed in pairs: $x_1, y_1, \quad x_2, y_2, \quad \ldots, \quad x_n, y_n.$

There is a *true function* for the relationship between x and y: $f(x) = b_0 + b_1 x$.

Unfortunately the true values of the coefficients b_0 and b_1 are unknown.

The data we observe contains an error term ϵ_i: $y_i = b_0 + b_1 x_i + \epsilon_i$.

The true error term ϵ_i is also unknown.

We use the observed data to estimate the true coefficients $\widehat{b}_0$ and $\widehat{b}_1$.

Once we have the coefficient estimates, then we can calculate the
fitted value $\widehat{y}_i = \widehat{b}_0 + \widehat{b}_1 x_i$

and the **residual** $\widehat{\epsilon}_i = y_i - \widehat{y}_i$
for each observation $i = 1, 2, \ldots, n$.

The fitted values are points that lie on the regression line for each of the observed x_i values

Regression parameters:

b_0: true regression **intercept** coefficient (unknown)
b_1: true regression **gradient** coefficient (unknown)
$\widehat{b}_0$: estimated intercept coefficient
$\widehat{b}_1$: estimated gradient coefficient

Estimated values are written with a “hat”.

Data Example

Grasshoppers chirp at a rate which depends on the temperature.

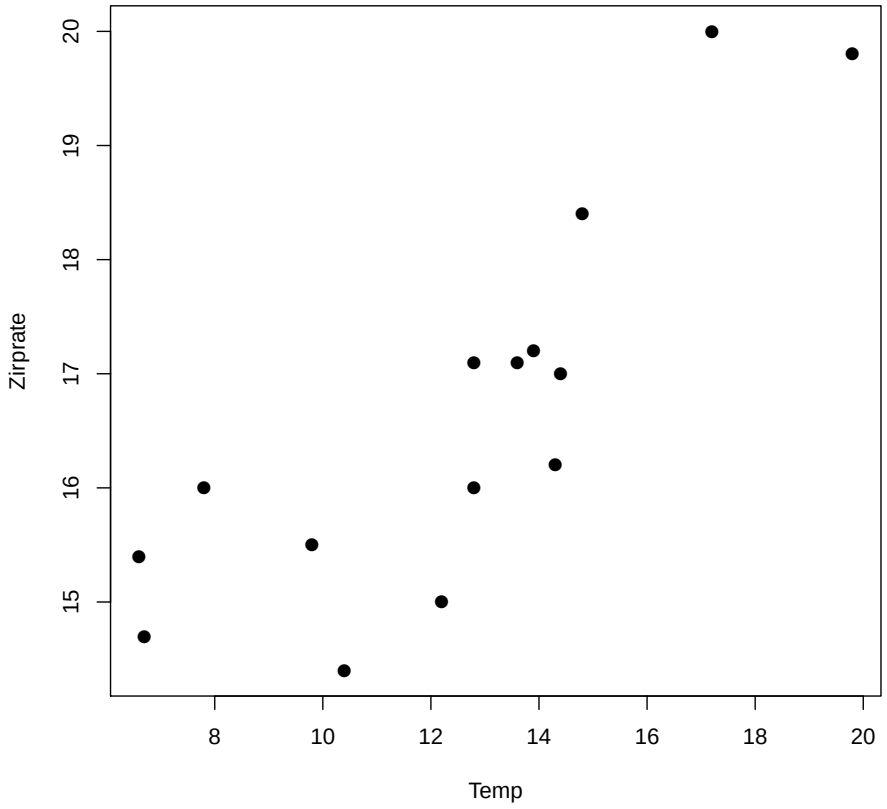
	1	2	3	4	5	6	7	8	9
Temperature	17.2	7.8	19.8	14.8	12.8	9.8	6.7	13.6	6.6
Chirp rate	20.0	16.0	19.8	18.4	17.1	15.5	14.7	17.1	15.4

	10	11	12	13	14	15
Temperature	14.3	12.2	13.9	12.8	14.4	10.4
Chirp rate	16.2	15.0	17.2	16.0	17.0	14.4

Predictor variable: X is temperature

Outcome variable: Y is chirp rate

Grasshopper Data: Chirp rate against temperature



The true regression function is $y_i = b_0 + b_1 x_i + \epsilon_i$

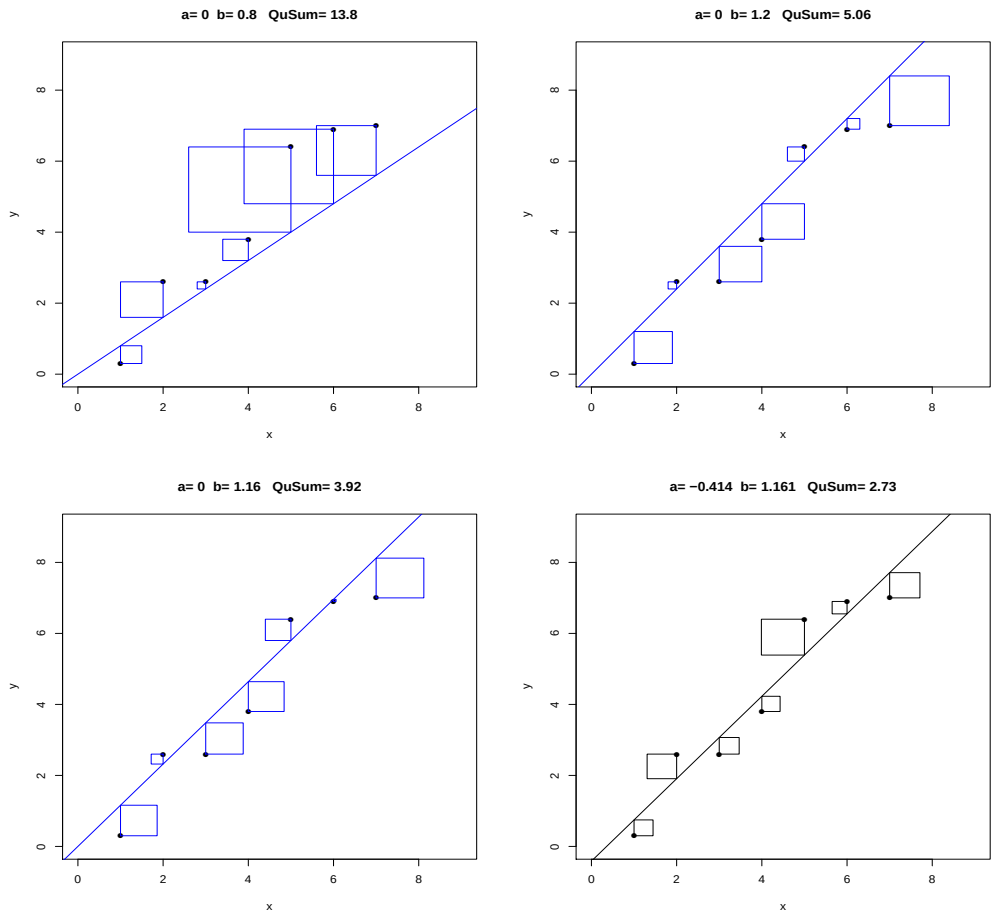
The chirp rate values don't all lie on a straight line.
ϵ_i is the error term corresponding to the natural variability in the chirp rates.

The true values of b_0 and b_1 are unknown. We have to *estimate* them.

The regression line is *"the line which best fits the data"*

When the total distance between the data points and the regression line is small, then the line fits the data well.

Our approach is to find the best line (the values of $\widehat{b_0}$ and $\widehat{b_1}$) using the method of least squares minimisation ...



QuSum = Sum of Squares

The squares in the diagrams are the values $\widehat{\epsilon}_i^2$ for a given $\widehat{b}_0$ and $\widehat{b}_1$. We want to minimise the total area of the squares, i.e. minimise

$$\widehat{\epsilon}_1^2 + \widehat{\epsilon}_2^2 + \cdots + \widehat{\epsilon}_n^2 = \sum_{i=1}^{n} \widehat{\epsilon}_i^2.$$

The values of $\widehat{b}_0$ and $\widehat{b}_1$ which minimise the sum of squares are our *best* intercept and gradient.

In this example:

the *best* intercept is $\widehat{b}_0 = -0.414$ and
the *best* gradient is $\widehat{b}_1 = 1.161$.

Method

$$\text{Minimise} \quad RSS = \sum_{i=1}^{n} \widehat{\epsilon}_i{}^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

(RSS = Residual Sum of Squares)

Computing the Coefficients

We don't need to explicitly minimise the residual sum of squares by trial and error.
The best estimates can be calculated directly using two formulae:

Linear coefficient (gradient)

$$\widehat{b}_1 = \frac{s_{xy}}{s_x^2} = \frac{\text{Covariance}}{\text{Variance X}}$$

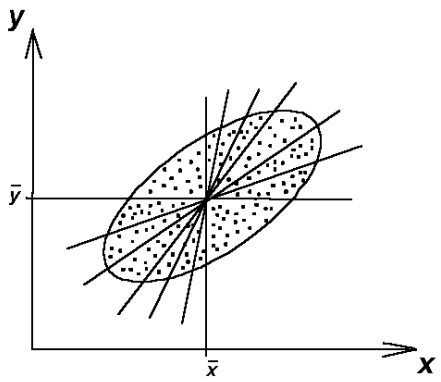
Intercept coefficient

$$\widehat{b}_0 = \overline{y} - \widehat{b}_1 \overline{x}$$

A consequence of the two formulae above is that the sum of the residuals is zero, as is the mean of the residuals.

$$\sum_{i=1}^{n}(y_i - \widehat{y}_i) = \sum_{i=1}^{n}\widehat{\epsilon}_i = 0 \quad \text{or} \quad \overline{\widehat{\epsilon}} = \frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_i = 0$$

Graphical interpretation: the regression line passes through $(\overline{x}, \overline{y})$.

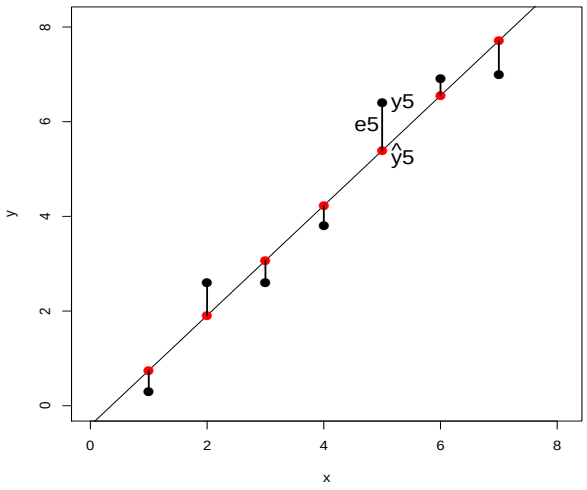


For every x_i, there is a point on the regression line. The y-coordinate of this point is called the **fitted value** $\widehat{y}_i$.

$$\widehat{y}_i = \widehat{f}(x_i) = \widehat{b}_0 + \widehat{b}_1 x_i$$

The difference between observed value y_i and the fitted value $\widehat{y}_i$ is called **residual**.

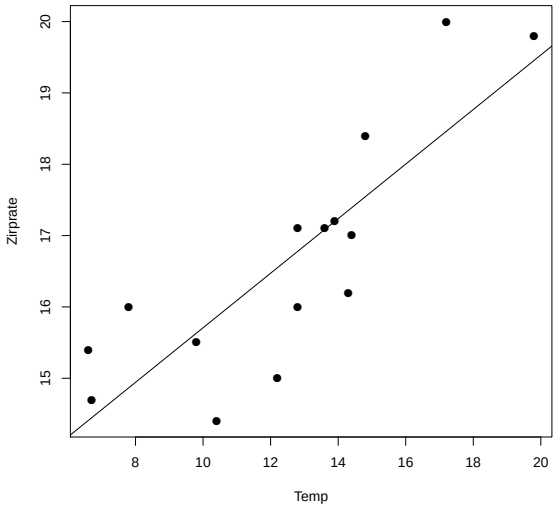
$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$



A **predicted value** is similar to a fitted value. In theory any real number x can be used in the regression formula, even if it is not a value in the data set.

For any $x \in \mathbb{R}$ the predicted value is $\widehat{f}(x) = \widehat{b}_0 + \widehat{b}_1 x$.

Grasshopper Example



Mean temp	$\overline{x} = 12.47$
Mean chirp	$\overline{y} = 16.65$
Var. temp.	$s_x^2 = 13.80$
Covariance	$s_{xy} = 5.278$
Gradient	$\widehat{b}_1 = s_{xy}/s_x^2$
	$= 5.278/13.80$
	$= 0.3825$
Intercept	$\widehat{b}_0 = \bar{y} - b\bar{x}$
	$= 16.65 - 0.3825 \cdot 12.47$
	$= 11.88$

The regression line has the form: $\widehat{f}(x) = 11.88 + 0.3825x$

The fitted value and the residual for $x_{11} = 12.2$ are:

The predicted values for $10^\circ C$ and $20^\circ C$ are:

Regression in R

Output from the grasshopper regression model

```
> lm.obj<-lm(formula = chirp ~ temp, data = Grasshoppers)
> summary(lm.obj)
Call: lm(formula = chirp ~ temp, data = Grasshoppers)

Residuals:
     Min       1Q   Median       3Q      Max
-1.54879 -0.58426  0.01574  0.60056  1.53880

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.882   0.908      13.084 7.36e-09 ***
temp         0.382   0.069       5.466 0.000108 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
Residual standard error: 0.9725 on 13 degrees of freedom
Multiple R-squared: 0.6968, Adjusted R-squared: 0.6735
F-statistic: 29.88 on 1 and 13 DF, p-value: 0.0001081
```


Comments

- `lm()` stands for linear model: linear regression is a type of linear model.
- The output of the function is stored in an object `lm.obj`
- This object contains lots of stored information about the regression model.
- `summary(lm.obj)` notices that `lm.obj` has class `"lm"` and so tailors the output accordingly.
- The *p*-values in the `Pr(>|t|)` column are important for deciding whether *x* has an influence on *y* and for choosing which predictor variables are important when multiple variables are available (details next week).

- Now a short **Workshop** covering what you have learnt today.
- **Next week** we will continue with the topic regression.