Linear Regression

Matt Eldridge

19 October 2015

Outline

- What is regression analysis and what can you use it for?
- Linear regression (some theory!)
- Fitting linear models in R
- Non-linear relationships

Regression analysis

- Statistical method for modelling the relationship between 2 or more variables
- One of the variables is the **response** (or **dependent**) variable
- The other variables are the **explanatory** (or **independent**) variables
- Both response and explanatory variables are continuous, i.e. real numbers with decimal places (weights, intensities, growth rates)

Uses of regression analysis

1. Understanding the functional relationships between the dependent variable and the independent variables

2. Predicting or estimating the unknown value of the dependent variable for given values of the independent variables

Is regression analysis appropriate for your data?

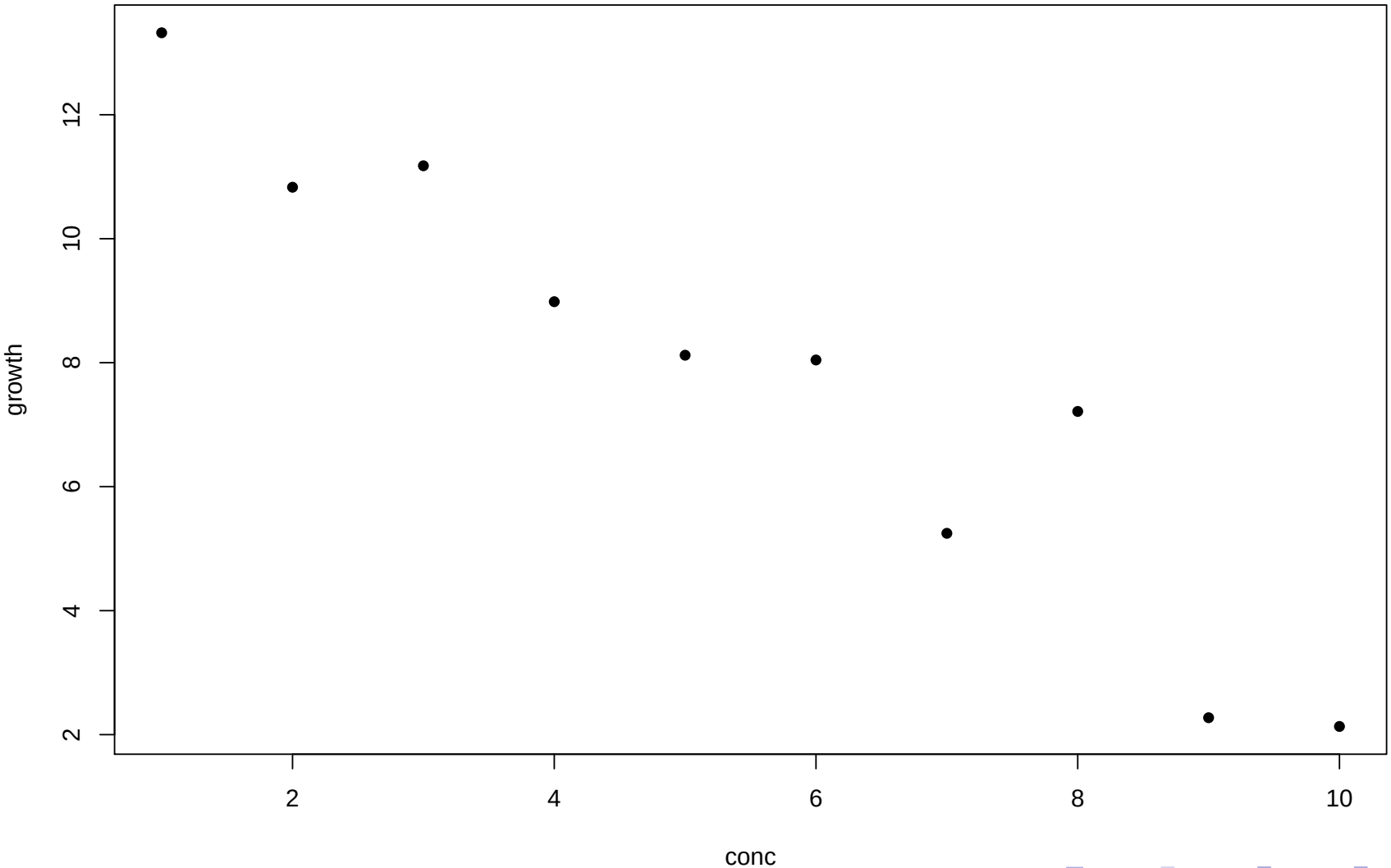
What is the most natural way of plotting your data?

- XY scatter plot $\Rightarrow$ regression
- Box plot $\Rightarrow$ ANOVA, t-test or non-parametric equivalent

Have you obtained measurements of some quantity at various conditions?

Lactoferrin data set

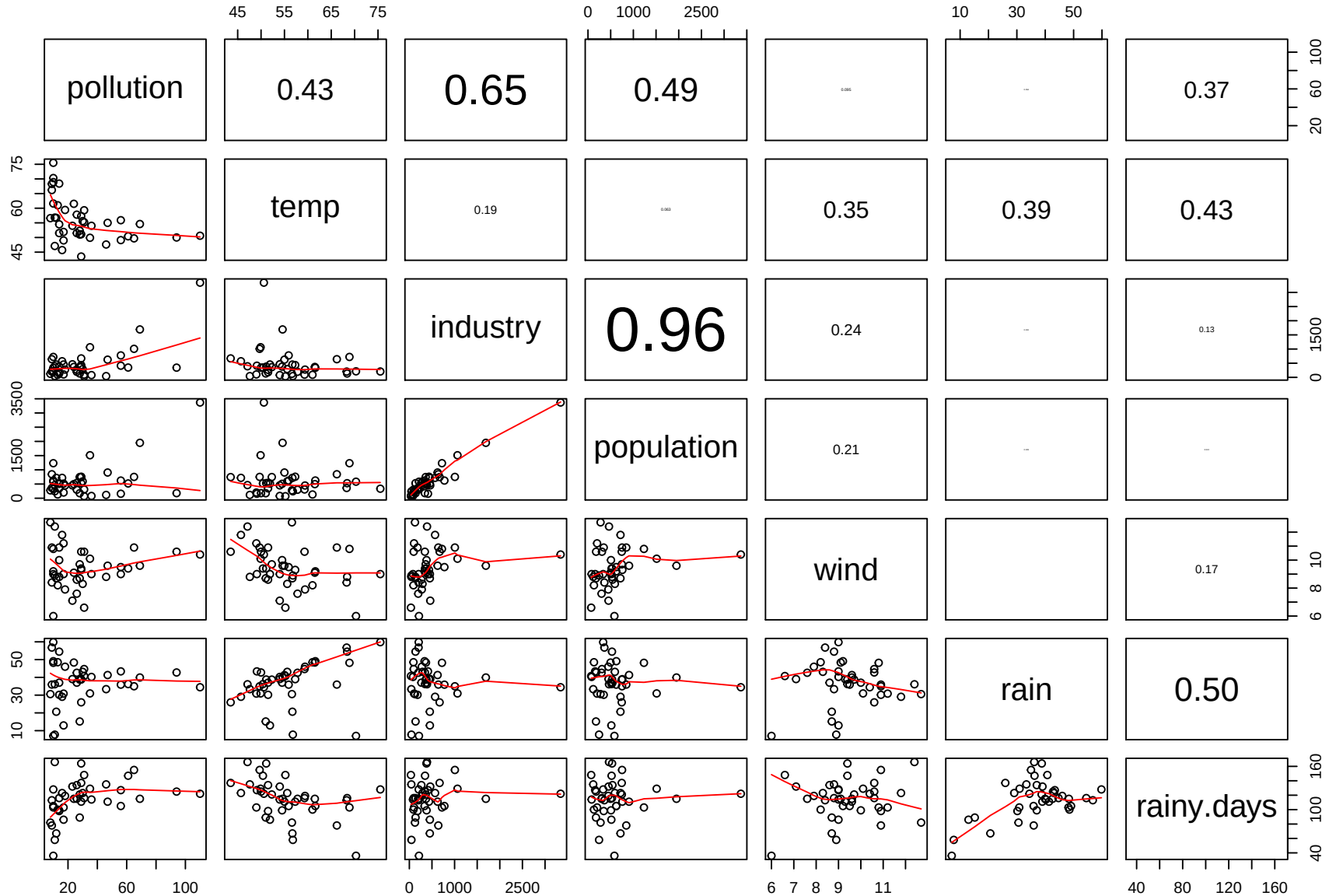
- Dose-response experiment where an *E. coli* strain was exposed to various concentrations of the growth inhibitor, lactoferrin



Types of regression analysis

- **Linear regression** – one response variable and one explanatory variable where the relationship can be described through a linear model
- **Multiple linear regression** – fits a linear model using multiple explanatory variables
- **Polynomial regression** – used to test for non-linearity in a relationship
- **Non-linear regression** – to fit a specified non-linear model to the data
- **Non-parametric regression** – used when there is no obvious functional form
- **Logistic regression** - when the response variable is a nominal (or categorical) variable

Pollution data set



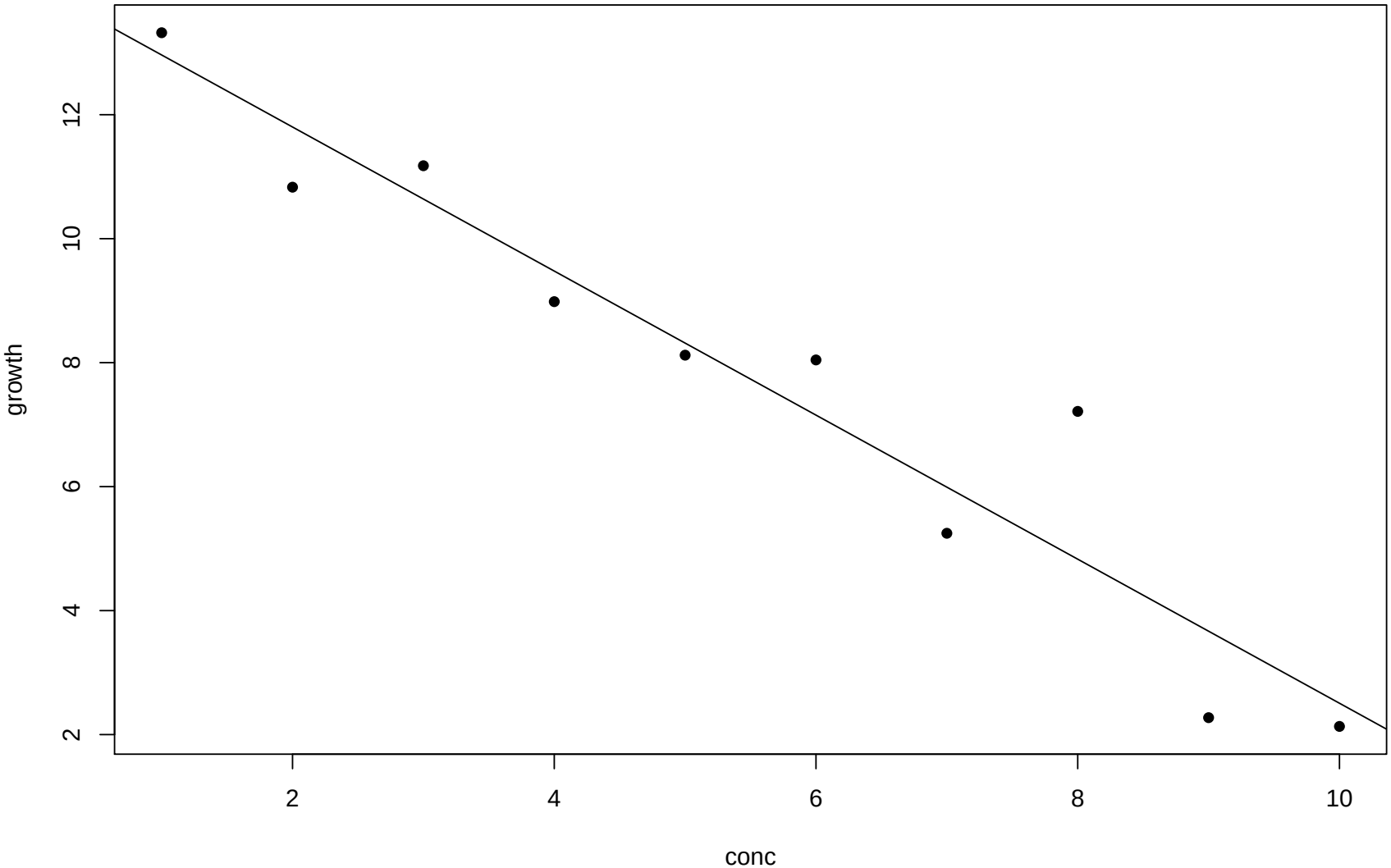
Linear regression

- Only one explanatory variable and one response variable
- Fits the simplest model of all, a straight line, to the data

$$y = ax + b$$

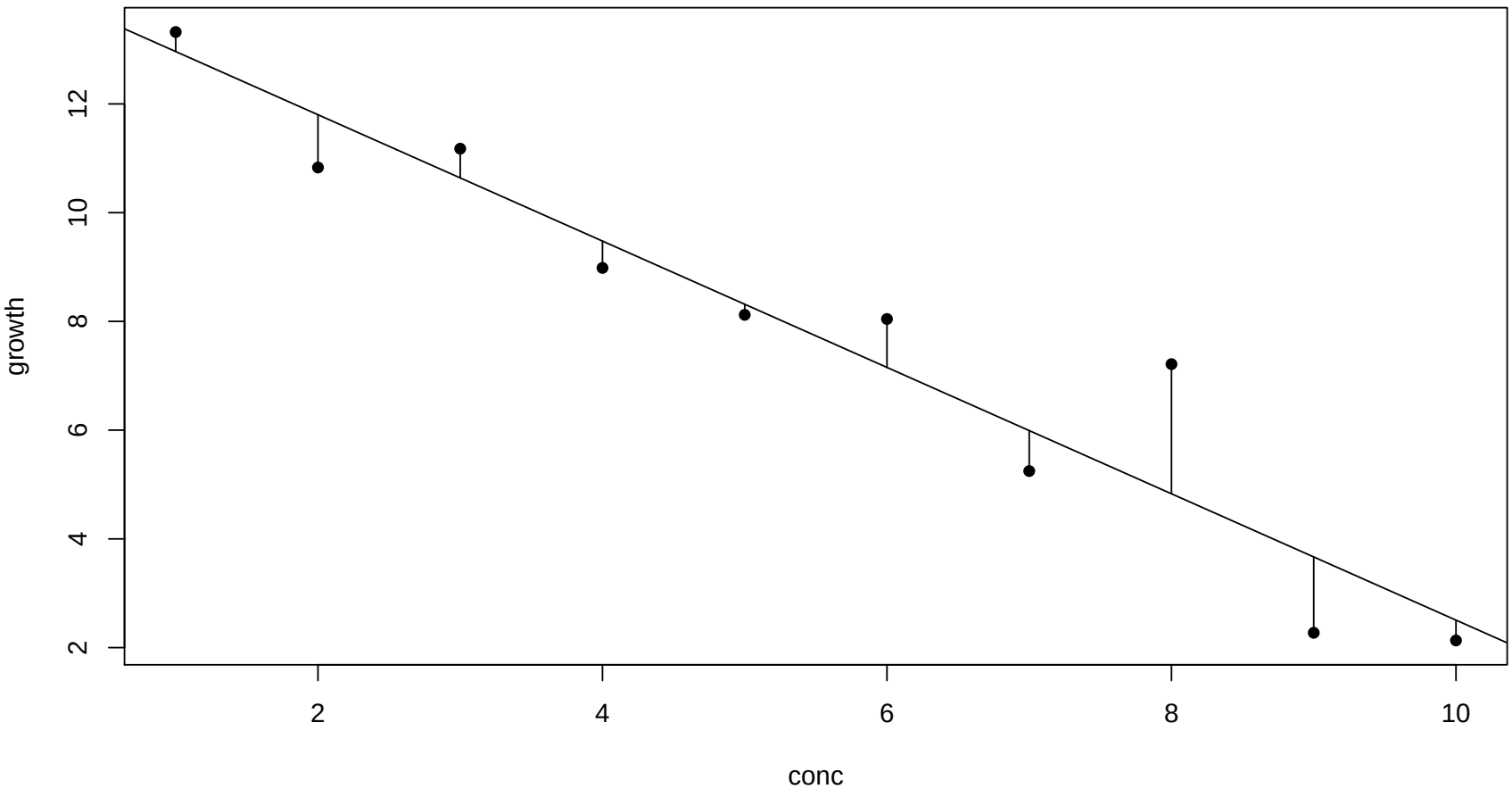
- a is the slope (or gradient) of the straight line
- b is the intercept, i.e. the value of y when x is 0.
- This is known as a **linear model**

Straight line fit for the lactoferrin data set



Residuals

- The residuals are the differences between the actual and the fitted values



- Choose the straight line that minimizes the square of these differences (this is known as 'least squares')

Fitting the linear model

The formal way to write the linear model is:

$$y_i = a + bx_i + \varepsilon_i$$

- y is the response variable, e.g. the value that was measured
- x is the explanatory variable, e.g. the condition that was varied
- i indicates the i-th observation
- ε is the error term which is assumed to be normally distributed

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Fitting the linear model

- Error sum of squares, SSE

$$SSE = \sum_i \varepsilon_i^2 = \sum_i (y_i - a - bx_i)^2$$

Mathematically, we set the derivative of this function with respect to the slope to zero ($dSSE/db = 0$), do the same for the derivative with respect to the intercept ($dSSE/da = 0$), and then solve the resulting simultaneous equations.

Linear regression in R

- The tilde symbol, $\sim$, is used in describing a model in R

$$growth \sim conc$$

- This can be read as 'growth is modelled as a function of concentration'
- Use the `lm` function to build the linear model

Reading the data in R

```
data <- read.csv("lactoferrin.csv")
data
```

```
##     conc growth
## 1      1  13.32
## 2      2  10.83
## 3      3  11.18
## 4      4   8.98
## 5      5   8.12
## 6      6   8.04
## 7      7   5.25
## 8      8   7.21
## 9      9   2.27
## 10    10   2.13
```

Alternatively, use the 'Import Dataset' function in RStudio.

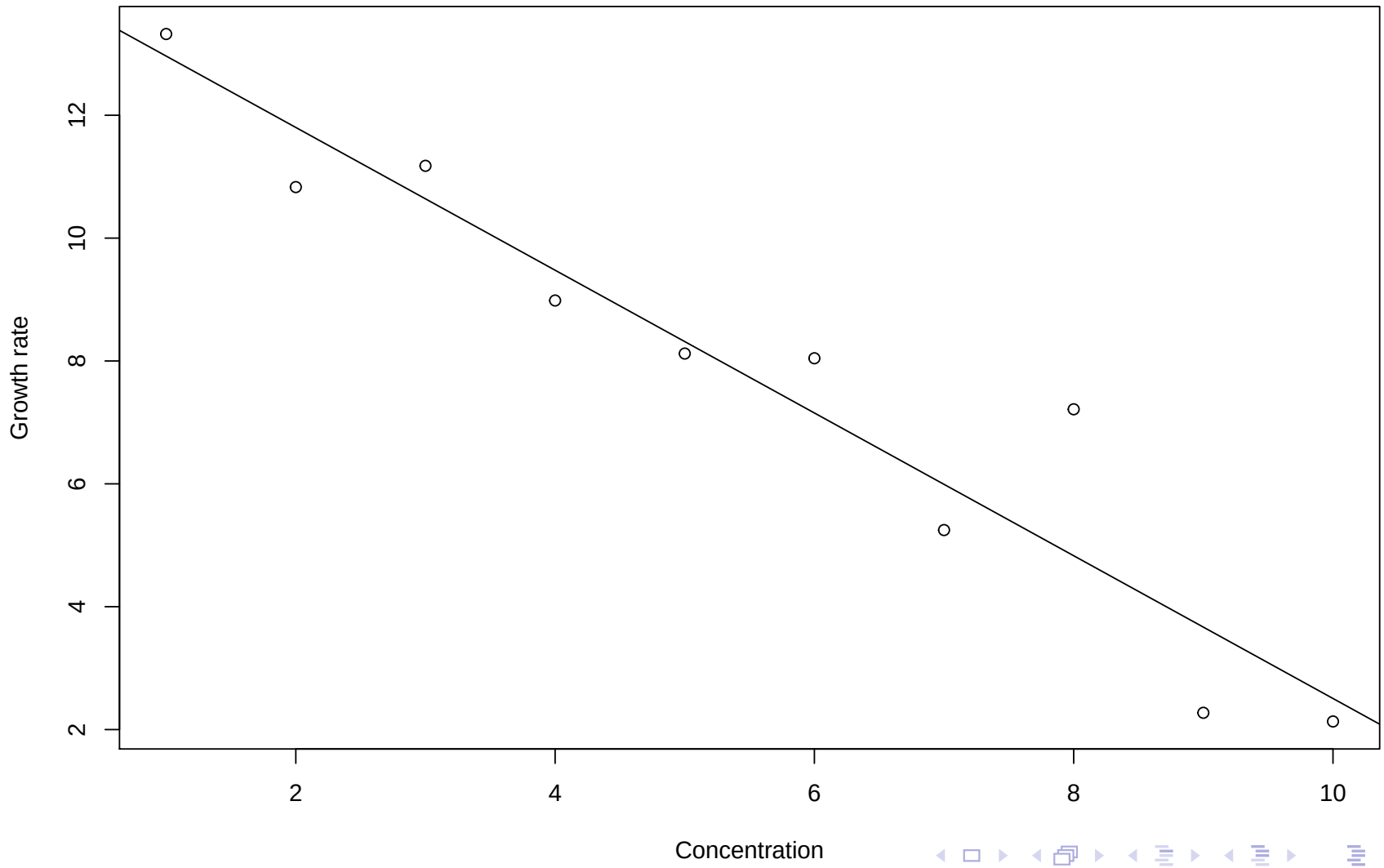
Fitting the linear model

```
model <- lm(growth ~ conc, data = data)
model
```

```
##
## Call:
## lm(formula = growth ~ conc, data = data)
##
## Coefficients:
## (Intercept)         conc
##       14.12        -1.16
```

Plotting the line of best fit

```
plot(data, xlab="Concentration", ylab="Growth rate")
abline(model)
```



Useful functions

```
coefficients(model)
```

```
## (Intercept)        conc
##       14.12      -1.16
```

```
fitted(model)
```

```
##     1     2     3     4     5     6     7     8     9
## 12.96 11.80 10.64  9.48  8.32  7.15  5.99  4.83  3.67
```

```
residuals(model)
```

```
##      1      2      3      4      5      6      7      8
##  0.359 -0.971  0.537 -0.493 -0.195  0.890 -0.744  2.384
```

Standard errors for the model coefficients

- Need to know how reliable are our estimates of the regression parameters (slope and intercept)
- Depends on the error variance, s^2

$$\text{variance}, s^2 = \frac{\text{sum of squares}}{\text{degrees of freedom}}$$

- Perform an analysis of variance.

Error variance in regression

- Total variation in y, represented by the total sum of squares of y, SSY is:

$$SSY = \sum_i (y_i - \bar{y})^2$$

- $\bar{y}$ is the mean value of y
- SSY can be partitioned into separate components for the variation that is explained by the model, SSR, and the unexplained variation that is the error sum of squares, SSE

$$SSY = SSR + SSE$$

Error variance in regression

$$SSY = SSR + SSE = \sum_i (y_i - \bar{y})^2$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

$$SSR = \sum_i (\hat{y}_i - \bar{y}_i)^2$$

- $\hat{y}$ are the fitted values of the response variable

$$\hat{y}_i = a + bx_i$$

ANOVA table for regression

Source	**Sum of squares**	**d.f.**	**Mean squares**	F **ratio**
Regression	$SSR = 111.38$	1	111.38	82.84
Error	$SSE = 10.76$	8	$s^2 = 1.34$	
Total	$SSY = 122.14$	9		

- The **degrees of freedom** (d.f.) depend on how many parameters have been estimated from the data in calculating the sum of squares
- For $SSY = \sum_i (y_i - \bar{y})^2$, one parameter is fixed, the mean value of y, so we have $n - 1$ degrees of freedom.
- For $SSE = \sum_i (y_i - a - bx_i)^2$, we need to know a and b, so we have $n - 2$ degrees of freedom.

Analysis of variance in R

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: growth
##           Df Sum Sq Mean Sq F value  Pr(>F)
## conc       1  111.4   111.4    82.8 1.7e-05 ***
## Residuals  8   10.8     1.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Summarizing the model

```
summary(model)
```

```
##
## Call:
## lm(formula = growth ~ conc, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.395 -0.681 -0.284  0.493  2.384
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.125      0.792    17.8  1.0e-07 ***
## conc          -1.162      0.128    -9.1  1.7e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 1.16 on 8 degrees of freedom
```

Confidence intervals for the model coefficients

```
confint(model, level = 0.95)
```

```
##                2.5 % 97.5 %
## (Intercept) 12.30 15.952
## conc        -1.46 -0.868
```

$$\text{confidence interval} = t\text{−value} \times \text{standard error}$$

$$CI_{95\%} = t_{(\alpha=0.025, \text{d.f.}=8)} \times \text{s.e.}$$

Measuring the degree of fit

- Output from `summary` function includes a value for r^2, a measure of the degree of fit
- r^2 is the fraction of the total variation in the response variable that is explained by the regression

$$r^2 = \frac{SSR}{SSY}$$

- r^2 varies from 0, when the regression explains none of the variation, to 1, when the regression explains all the variation
- r is Pearson's product-moment correlation coefficient

Correlation coefficient

```
cor.test(data$growth, data$conc)
```

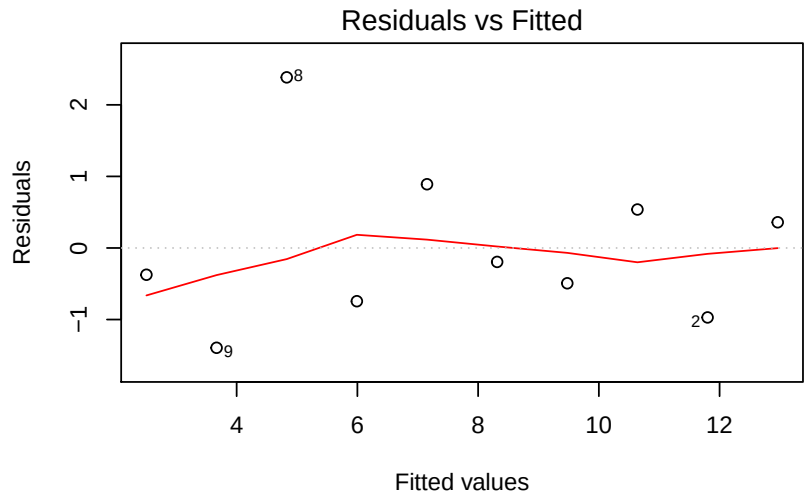
```
##
##  Pearson's product-moment correlation
##
## data:  data$growth and data$conc
## t = -9, df = 8, p-value = 2e-05
## alternative hypothesis: true correlation is not equal to
## 95 percent confidence interval:
##  -0.990 -0.816
## sample estimates:
##    cor
## -0.955
```

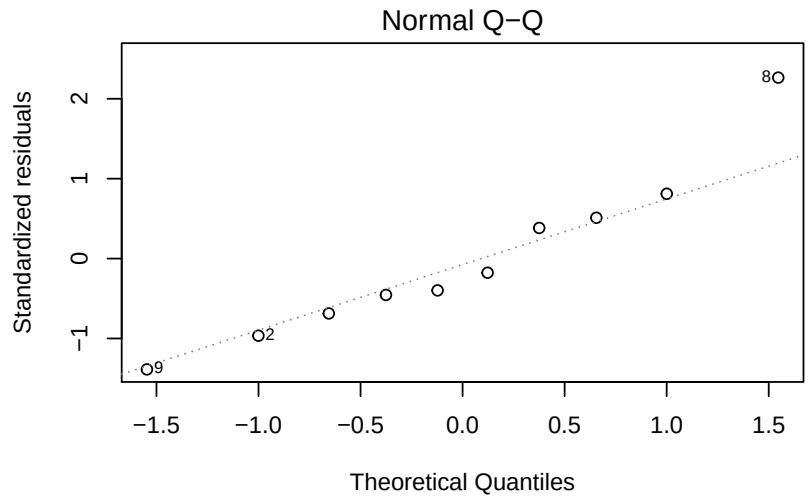
Model assumptions

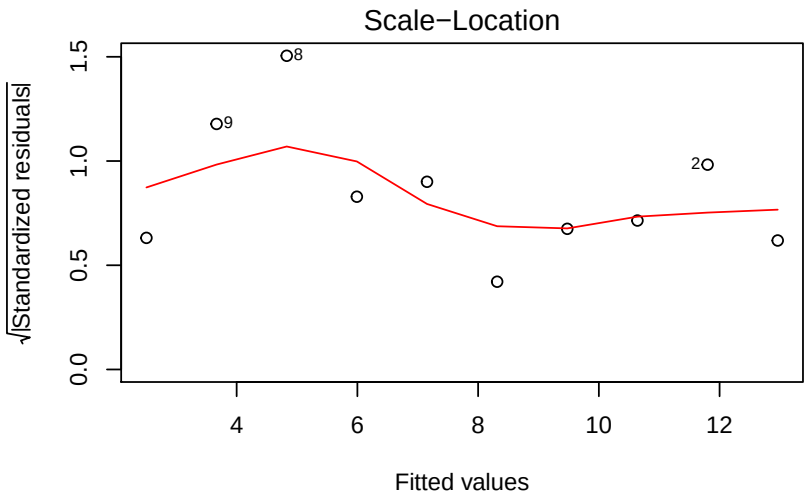
- **Linear relationship** between explanatory and response variable
- **Constant variance (homoscedasticity)** – variance of the errors is constant across the range of values for the explanatory variable
- **Independence of errors** – the errors in the response variables are uncorrelated with other
- **Normality of errors** – the residuals follow a normal distribution

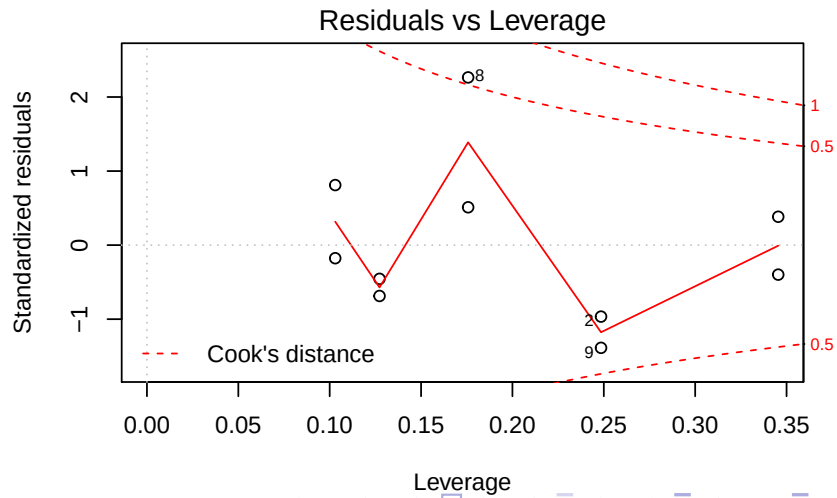
Checking the model assumptions

```
plot(model)
```

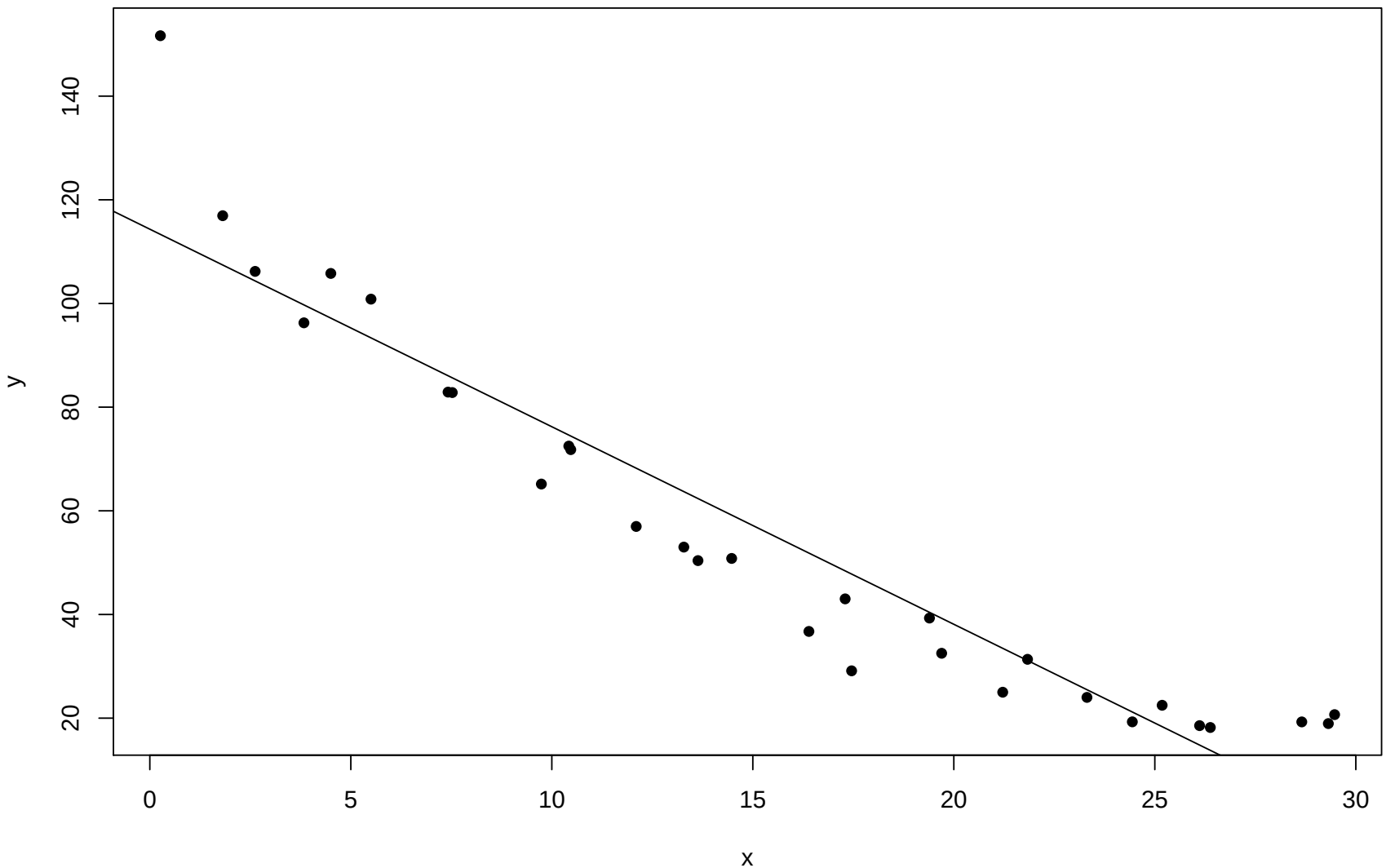




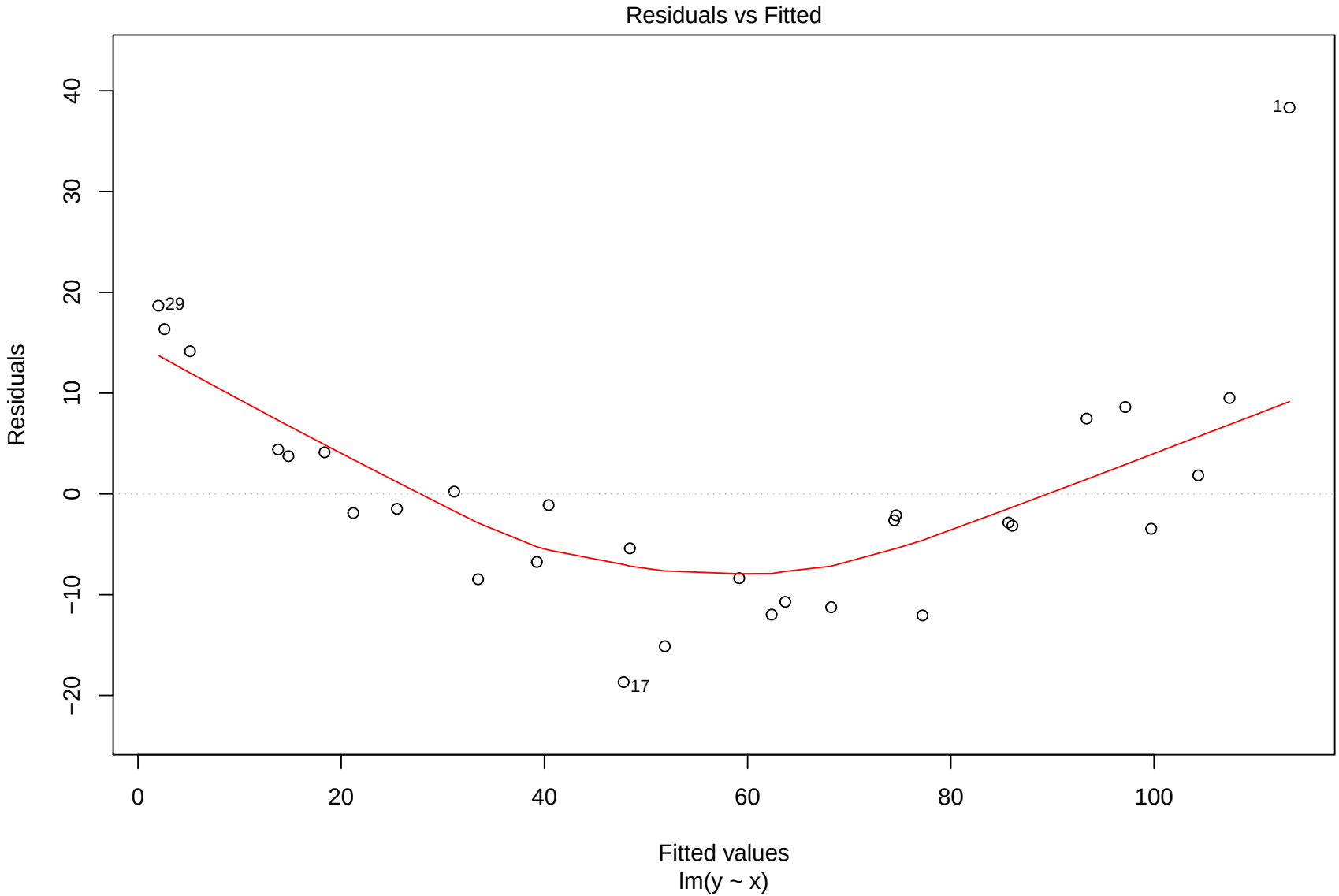




Radioactive decay data set



Radioactive decay data set



Assessing for non-linearity

Polynomial regression can be used to check for non-linearity in the relationship between the explanatory and response variable

- Add a quadratic term, x^2, to the model

$$y = a + bx + cx^2$$

This is still a linear model even though the relationship is non-linear

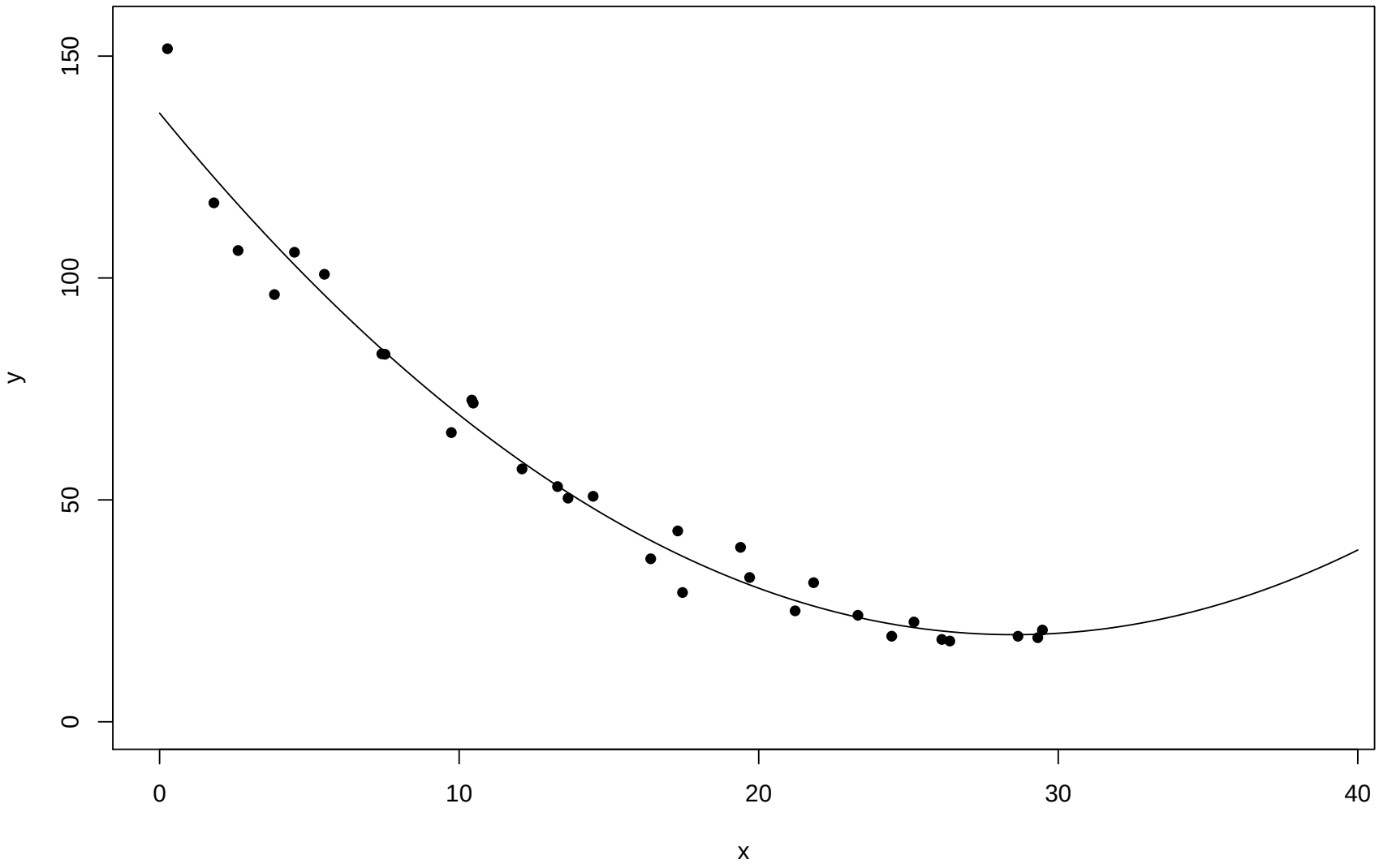
- Check for significance of the additional term

```
model <- lm(y ~ x + I(x^2), data = data)
summary(model)
```

Assessing for non-linearity

```
##
## Call:
## lm(formula = y ~ x + I(x^2), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.403  -2.267  -0.429   2.797  16.669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 137.1447     3.2076   42.76  < 2e-16 ***
## x            -8.2430     0.4803  -17.16  4.8e-16 ***
## I(x^2)        0.1446     0.0152    9.53  4.0e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 5.78 on 27 degrees of freedom
## Multiple R-squared:  0.976,  Adjusted R-squared:  0.974
```

Quadratic regression model



Fitting an exponential function

- An exponential decay function might be a better fit for our radioactive decay data

$$y = ae^{-bx}$$

- Some models can be **linearized** by **transforming** the exploratory or response variable
- In this case, take logarithms

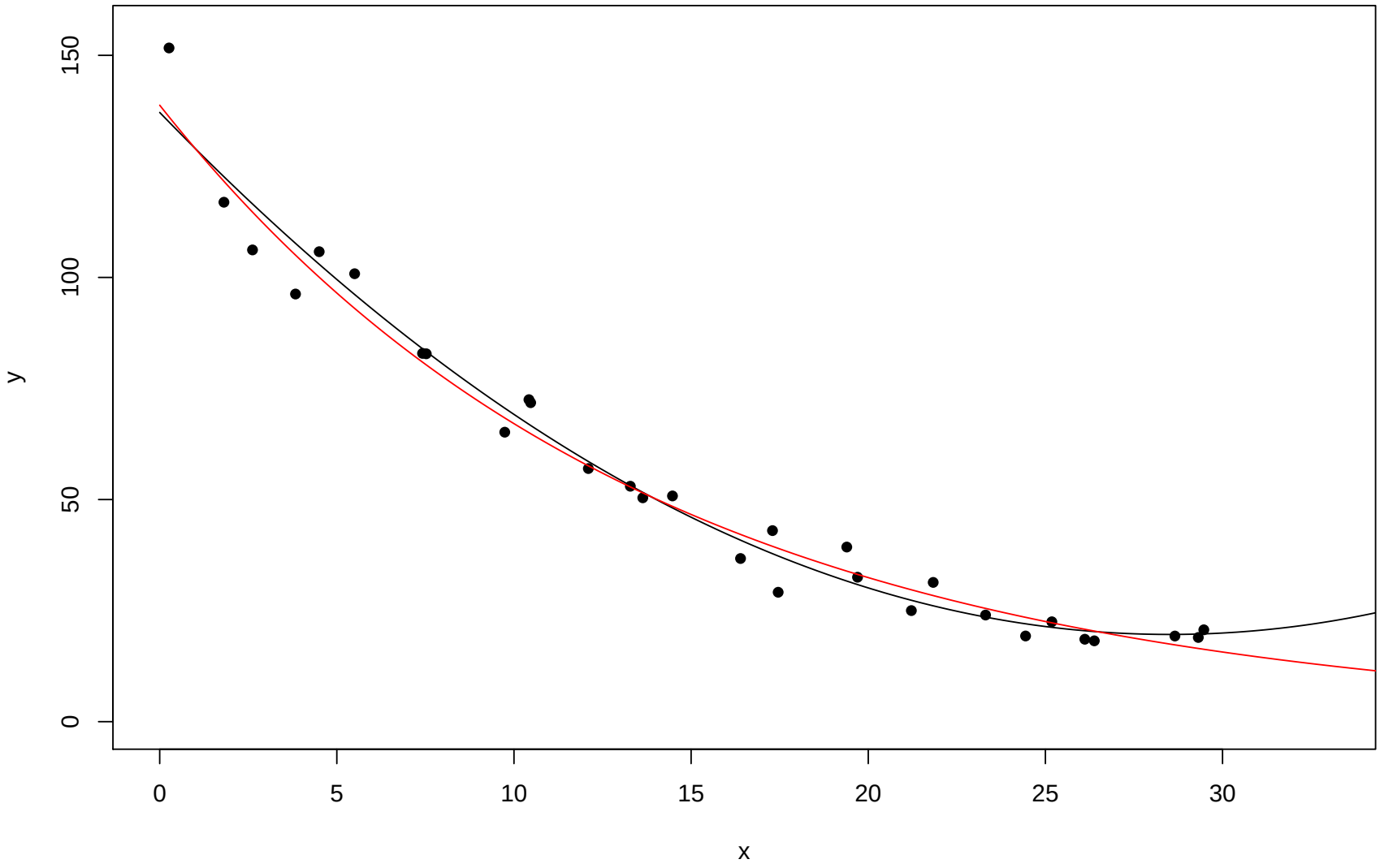
$$\ln(y) = \ln(a) - bx$$

```
model <- lm(log(y) ~ x, data = data)
summary(model)
```

Fitting an exponential function

```
##
## Call:
## lm(formula = log(y) ~ x, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.29218 -0.07294  0.00668  0.09593  0.23848
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.93285    0.04391   112.3   <2e-16 ***
## x           -0.07269    0.00247   -29.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.119 on 28 degrees of freedom
## Multiple R-squared:  0.969,  Adjusted R-squared:  0.968
## F-statistic:  866 on 1 and 28 DF,  p-value: <2e-16
```

Fitted models for the radioactive decay data



Multiple regression

- When there are 2 or more explanatory variables

$$y = a + bx_1 + cx_2 + dx_3 + ...$$

```
model <- lm(y ~ x1 + x2 + x3)
```

- Can include interaction terms

$$y = a + bx_1 + cx_2 + dx_1x_2$$

```
model <- lm(y ~ x1 + x2 + x1:x2)
```

- Care needed to avoid overfitting

Non-linear regression

- If a specific mechanistic model lends itself to the data that takes the form of a non-linear equation (one that cannot be linearized by transformation), e.g.

$$y = a - be^{-cx}$$

- Use the `nls` library in R
- Specify the model explicitly
- Need to provide initial guesses for the parameters

```
library(nls)
model <- nls(y ~ a - b * exp(-c * x),
             start = list(a = 50, b = 100, c = 0.05))
```

Summary

- Regression analysis models the relationships between explanatory variable(s) and a response variable
- Linear regression involves fitting a straight line of best fit by minimizing the sum of squares of the residuals ('least squares')
- The model fit can be assessed using diagnostic plots (residuals vs fitted, QQ-plot, etc.)
- Add quadratic terms to the linear model to check for non-linearity in the relationship
- Some non-linear functions, e.g. exponential decays, can be linearized by transformation, otherwise can use non-linear regression

Michael J. Crawley *'Statistics: An Introduction using R'*, Second Edition (Wiley, 2014)