MAST7866: Linear Regression

Professor Jian Zhang

Introduction

How can we predict things? For example:

- Tomorrow's stock market price.
- The age of a viewer watching a given video on YouTube.
- The amount of prostate specific antigen (PSA) for a patient visiting a clinic.
- The temperature in this room tomorrow at 11am.

Statistical models

Statistical models are used to describe random variables.

For example, the raw marks of A-level mathematics students are normally distributed with mean μ and variance σ^2.

We assume that both the population and the sample follow the same model.

Introduction

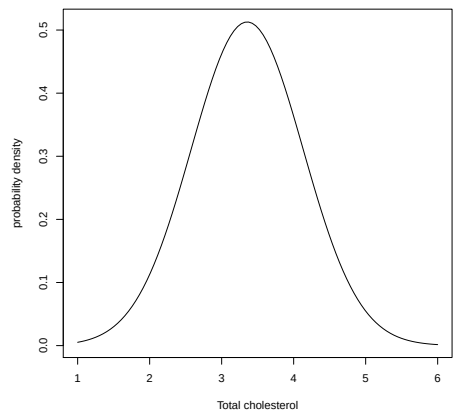
We are often interested in understanding and modelling the relationship between two variables.

In statistics, we are interested in models with parameters (which we consider to be fixed but unknown) and random elements.

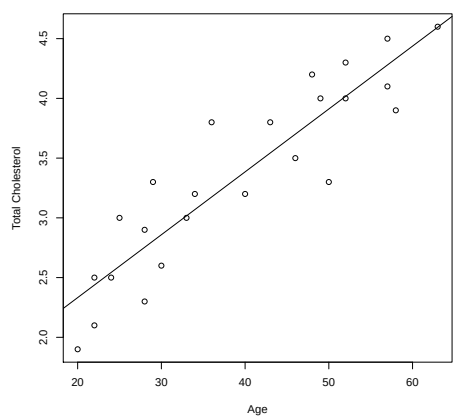
Illustrative Example

A doctor wishes to predict cholesterol for a new patient visiting a clinic. The doctor has measurements for previous patients. The sample mean is 3.35 and the sample standard deviation is 0.78.

If we assume that total cholesterol is normally distributed.



A scatterplot taking into account the relationship between age and total cholesterol.



Simple Linear Regression

Suppose that we observe a random sample of pairs of variables $(x_1,\ y_1), (x_2,\ y_2), \cdots, (x_n,\ y_n)$

y_i is the response variable, target variable or dependent variable.

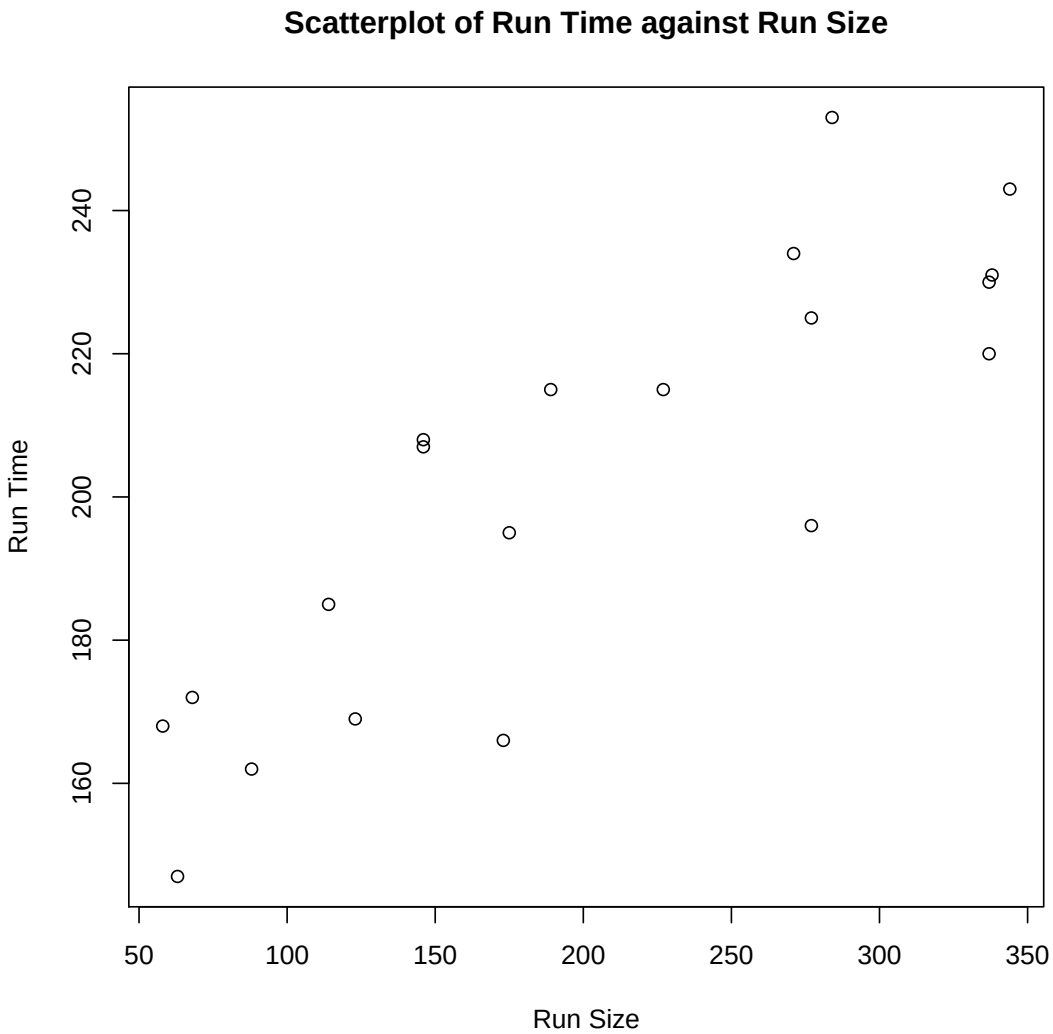
x_i is the covariate, regressor, independent variable or explanatory variable.

Example: Production Runs

The data record the time taken for a production run (in minutes) and the number of items produced for 20 randomly selected orders.

We are interested in understanding how these two variables are related.

Example: Production Runs



Regression

It is assumed that the effect of x_i on y_i is linear and that there are noisy observations

$$y_i = \alpha + \beta x_i + e_i$$

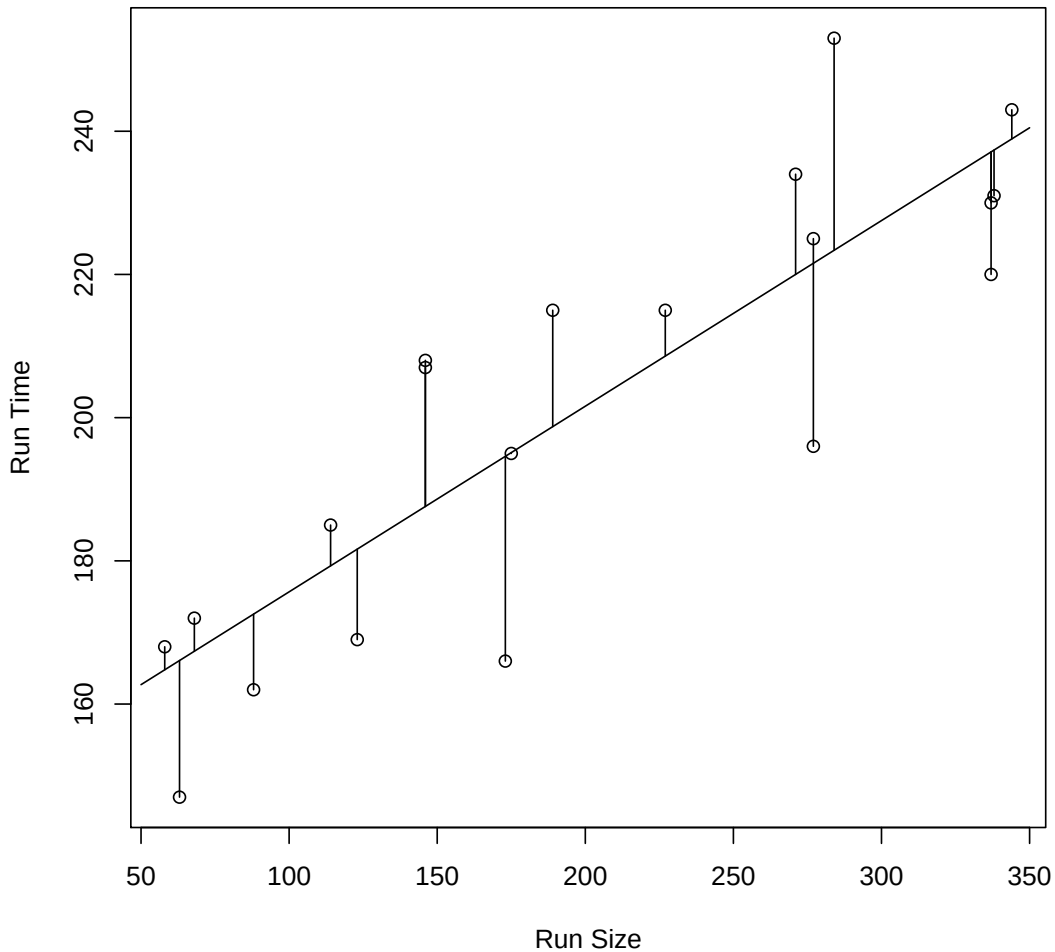
where

- y_i is called the response variable.
- x_i is called the explanatory variable (treated as non-random).
- e_i is called the error which is random with $\mathrm{E}(e_i) = 0$, $\mathrm{Var}(e_i) = \sigma^2$ and $\mathrm{Cov}(e_i,\ e_j) = 0,\ i \neq j$.
- α is called the intercept.
- β is called the slope.

This implies that the mean of y_i is $\alpha + \beta x_i$ and the variance of y_i is σ^2.

Estimating α and β: least squares

The model has three parameters α, β and σ^2 which we would like to estimate from the sample.



Estimating α and β: least squares

We define the function

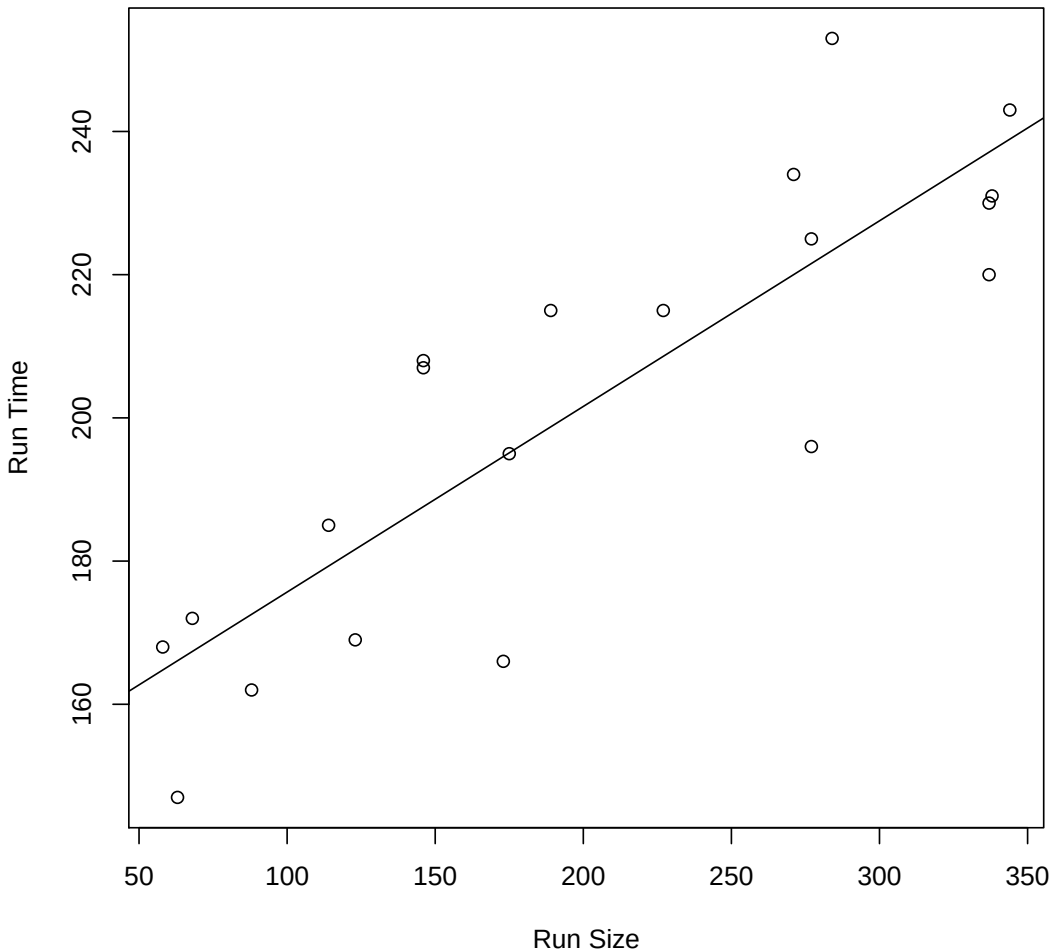
$$S(\alpha, \beta) = \left(y_1 - (\alpha + \beta x_1)\right)^2 + \left(y_2 - (\alpha + \beta x_2)\right)^2 + \cdots + \left(y_n - (\alpha + \beta x_n)\right)^2$$

The principle of least squares is that the minimizer $(\hat{\alpha}, \hat{\beta})$ of $S(\alpha, \beta)$ is a reasonable estimator of (α, β).

$\hat{\alpha}$ and $\hat{\beta}$ are called the least squares estimators (LSE).

Fitted regression line for production runs example

The fitted regression line is $\hat{\alpha} + \hat{\beta}x$ or, in this example,
RunTime $= 149.8 + 0.2592 \times$ RunSize



Interpretation of α and β

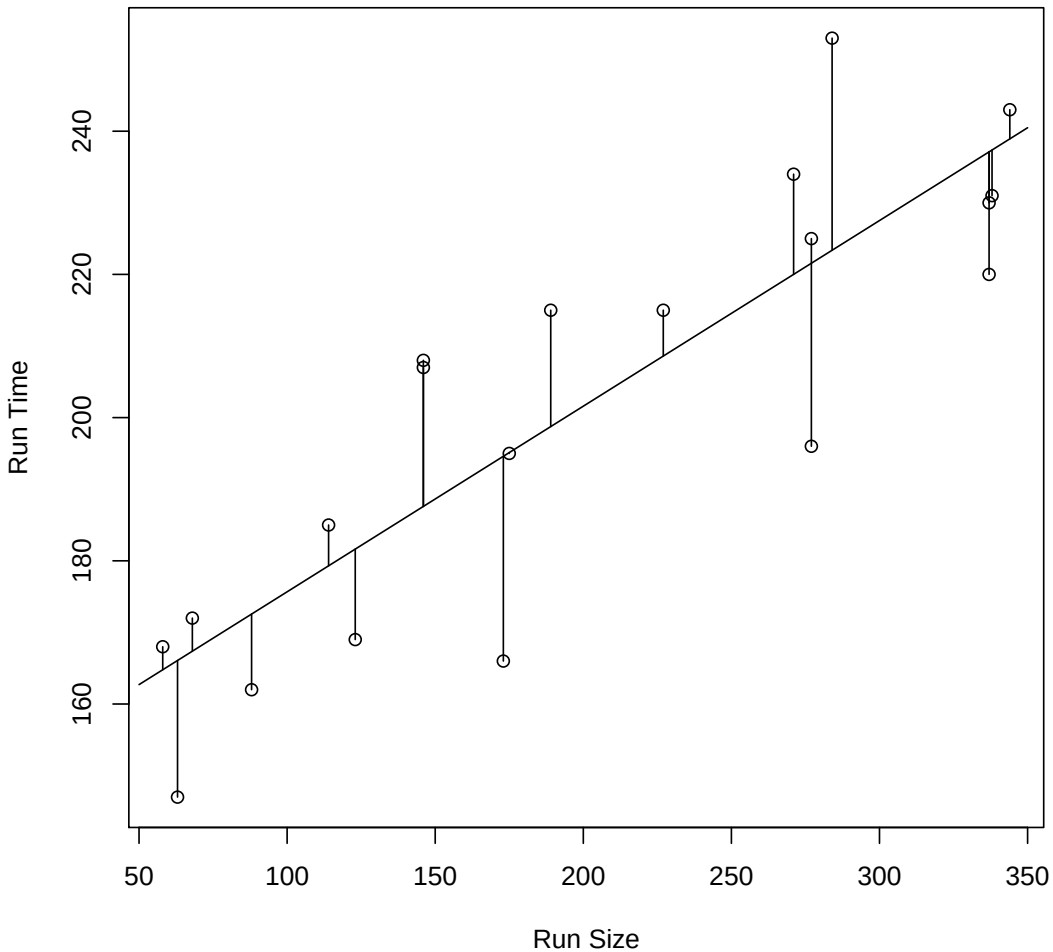
The parameter α represents the mean value of the response if $x = 0$.

The parameter β represents the change in the mean value of the response when x increases by one unit.

For example, on average the run time increases by 0.26 minutes (16 seconds) if the run length increases by one item.

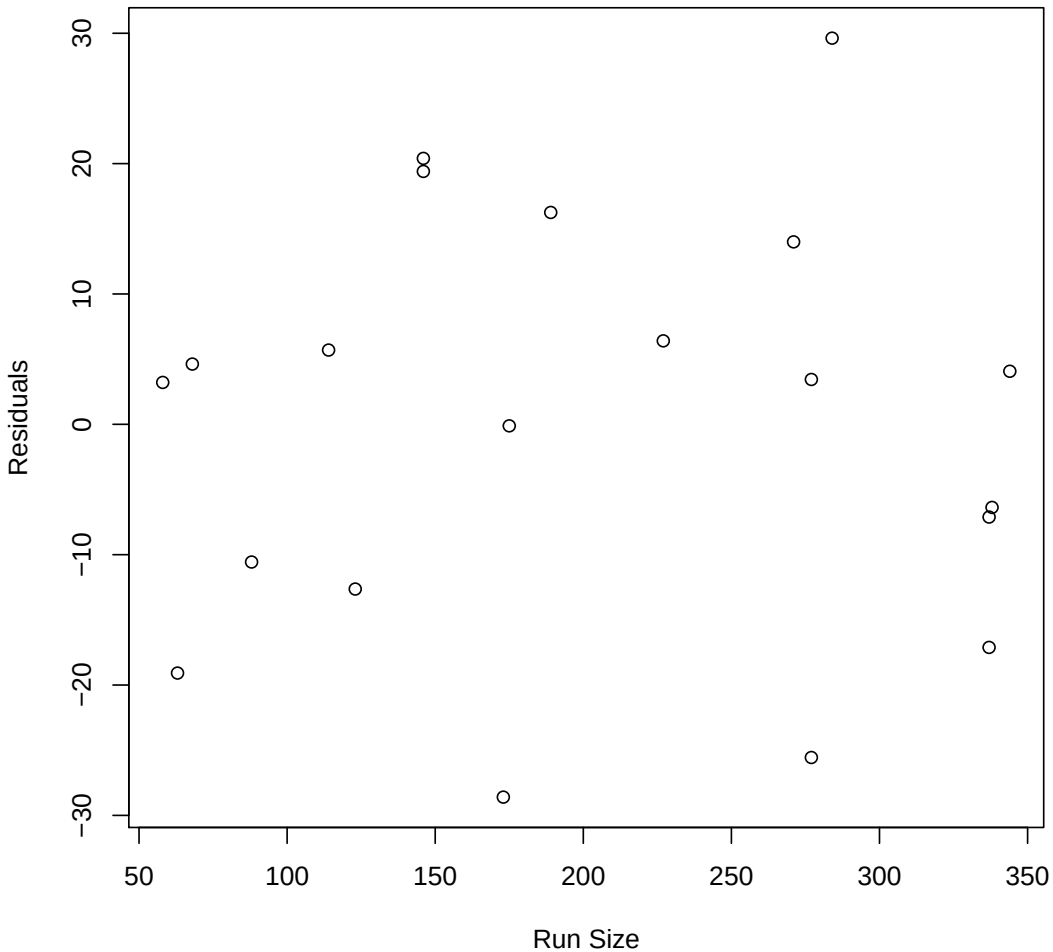
Estimating σ^2

The fitted values are $y_i = \hat{\alpha} + \hat{\beta} x_i$ and the fitted line is $y = \hat{\alpha} + \hat{\beta} x$.



Estimating σ^2: Residuals

The residual is defined to be $r_i = y_i - \hat{\alpha} - \hat{\beta}\, x_i$



Estimating σ^2

The Residual Sum of Squares (RSS) is defined to be

$$\begin{aligned} & \text{RSS} \\ = \; & r_1^2 + r_2^2 + \cdots + r_n^2 \\ = \; & \left(y_1 - \hat{\alpha} - \hat{\beta} x_1\right)^2 + \left(y_2 - \hat{\alpha} - \hat{\beta} x_2\right)^2 + \cdots + \left(y_n - \hat{\alpha} - \hat{\beta} x_n\right)^2 \end{aligned}$$

We estimate

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-2}.$$

In RStudio

```
> fit_production <- lm(RunTime ~ RunSize, data = production)
> summary(fit_production)
Call:
lm(formula = RunTime ~ RunSize, data = production)

Residuals:
    Min      1Q  Median      3Q     Max
-28.597 -11.079   3.329   8.302  29.627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-squared:  0.7302,Adjusted R-squared:  0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```

Residuals and Model checking

The linear regression model makes certain assumptions:

- The mean of y_i is linear, *i.e.* $\alpha + \beta x_i$.
- The errors are normally distributed.
- The variance of the errors is the same for all values of x.

The validity of all our inferences depends on these assumptions being correct.

Three plots will be used to check if the data and the model agree:

- Residuals versus fitted values.
- QQ-plot of the standardized residuals.
- Square root of standardized residuals versus fitted values.

Fitted values and residuals

The fitted values $\hat{y}_i$ are

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, \qquad i = 1, 2, \ldots, n.$$

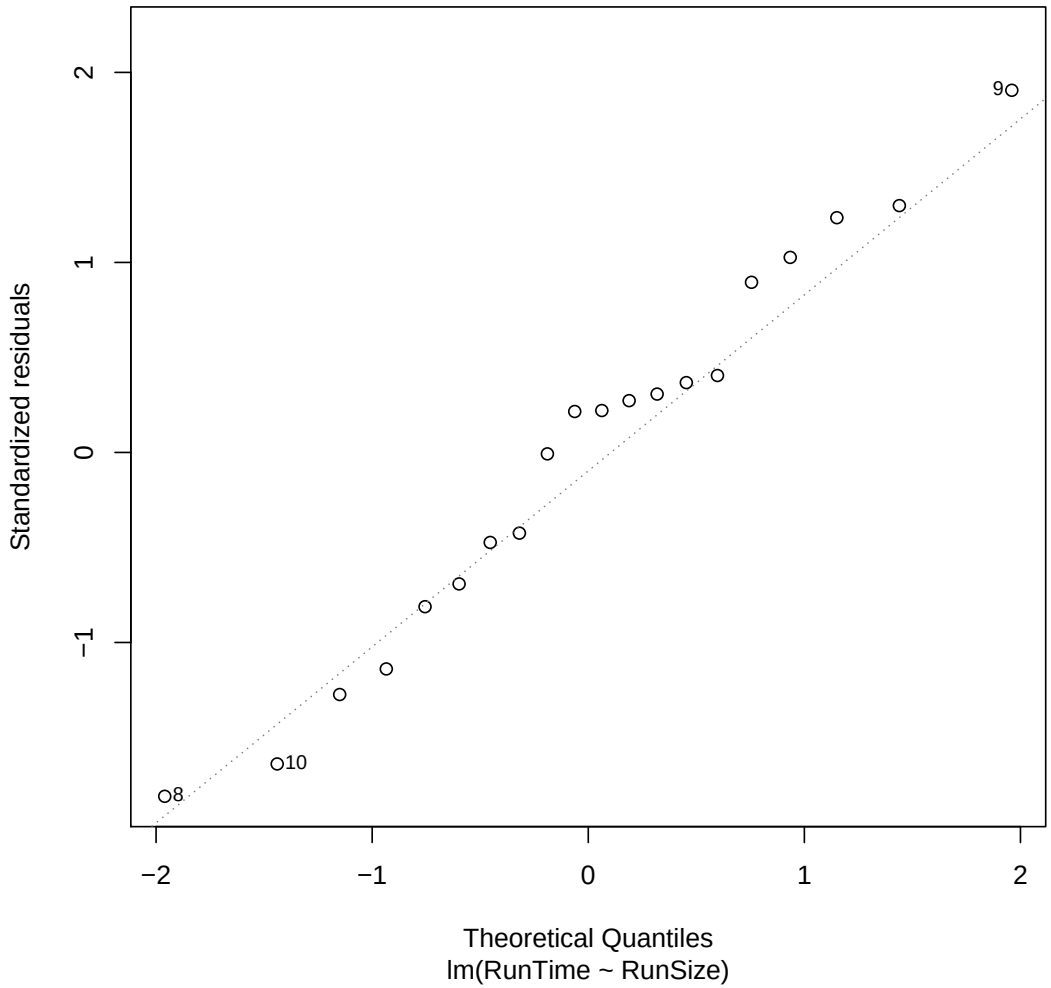
The residuals r_i are

$$\begin{aligned} r_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\alpha} - \hat{\beta} x_i, \qquad i = 1, 2, \ldots, n. \end{aligned}$$

In practice, we use standardized residuals.

Checking normality: QQ-plot of standardized residuals

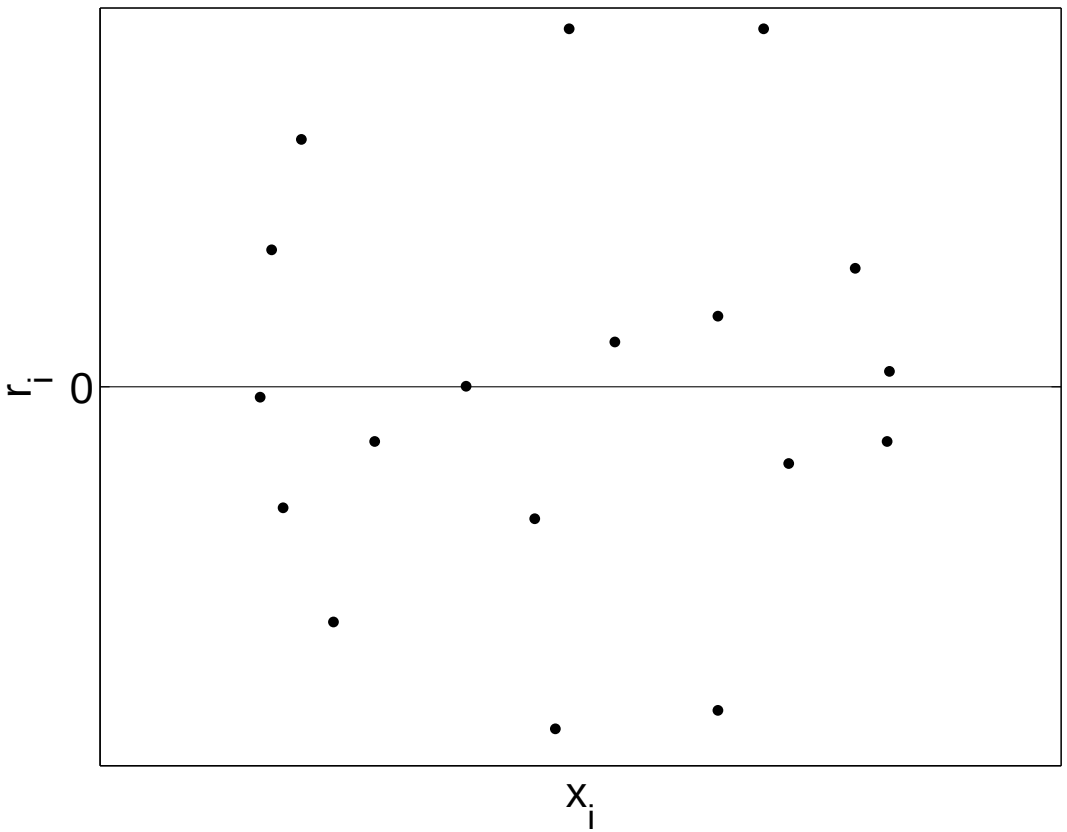
This plot shows the residuals in increasing order against what we expect from the normal distribution.



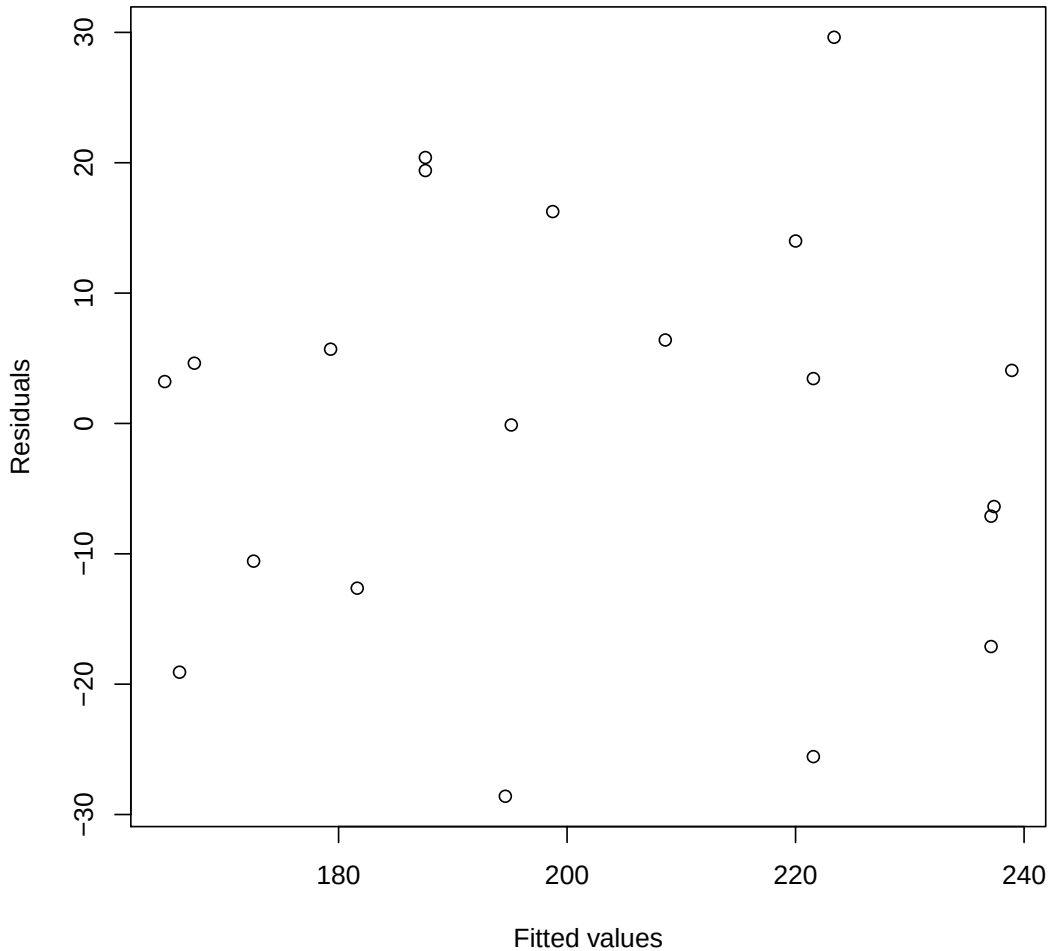
Checking linearity: residuals versus fitted values

If the model is appropriate and the assumptions hold, there should be no pattern in the residuals.

A typical plot of residuals against explanatory variables in this case.

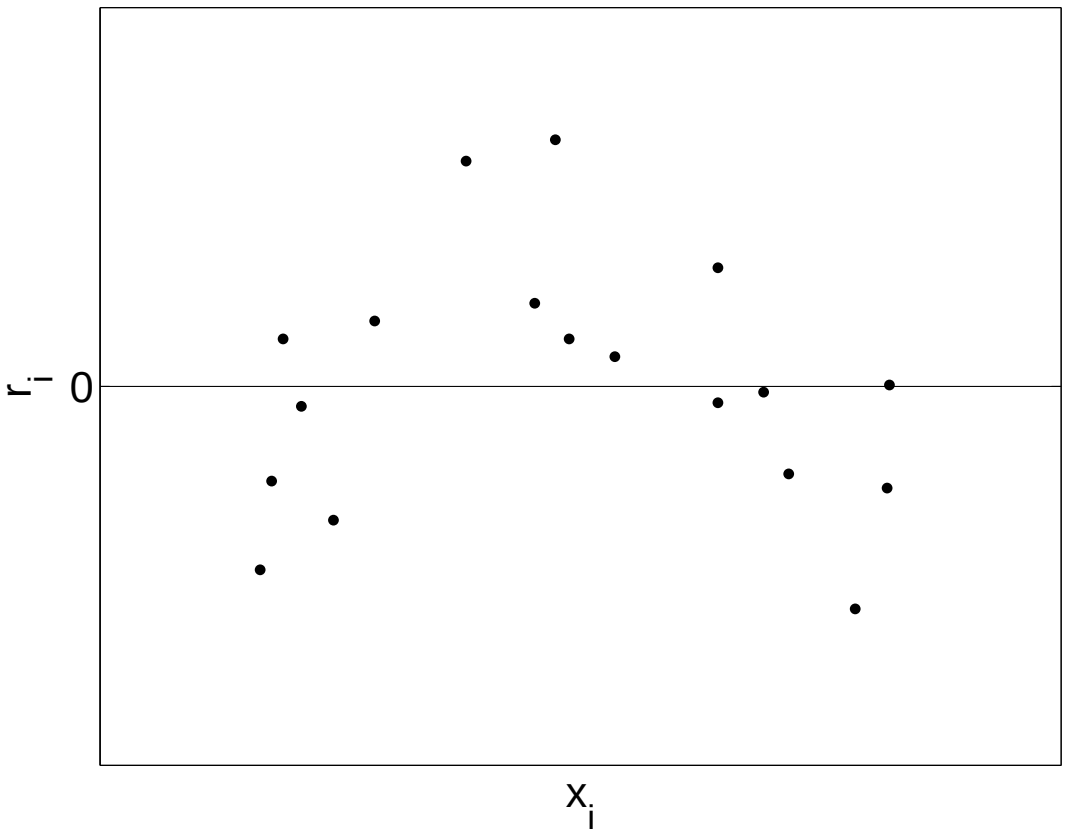


Residuals for production run data



A departure from linearity

The underlying relationship between x and y may not be linear



Example: Predicting wine prices using critic scores

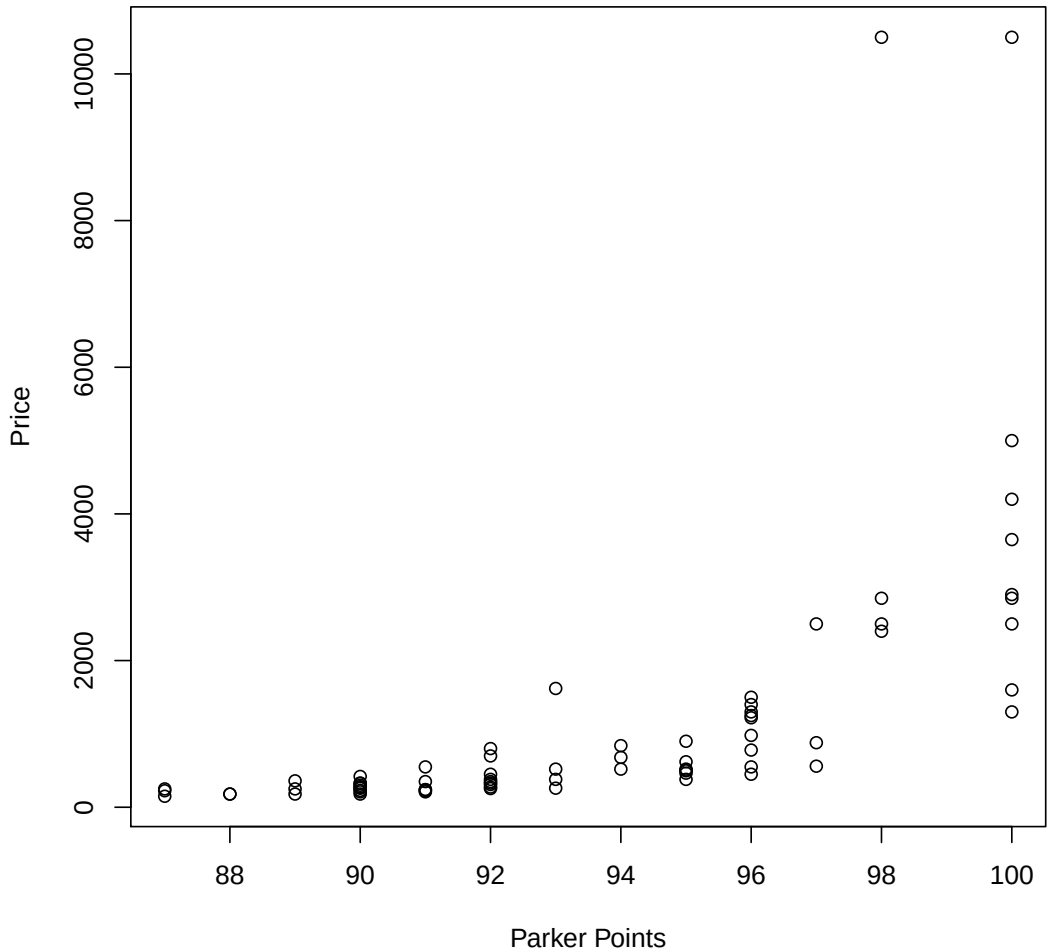
We consider the relationship between the score that a wine critic, Robert Parker, gives to a wine and the price of that wine. The data are the price and rating for 72 wines from the 2000 vintage in Bordeaux.

The variables are

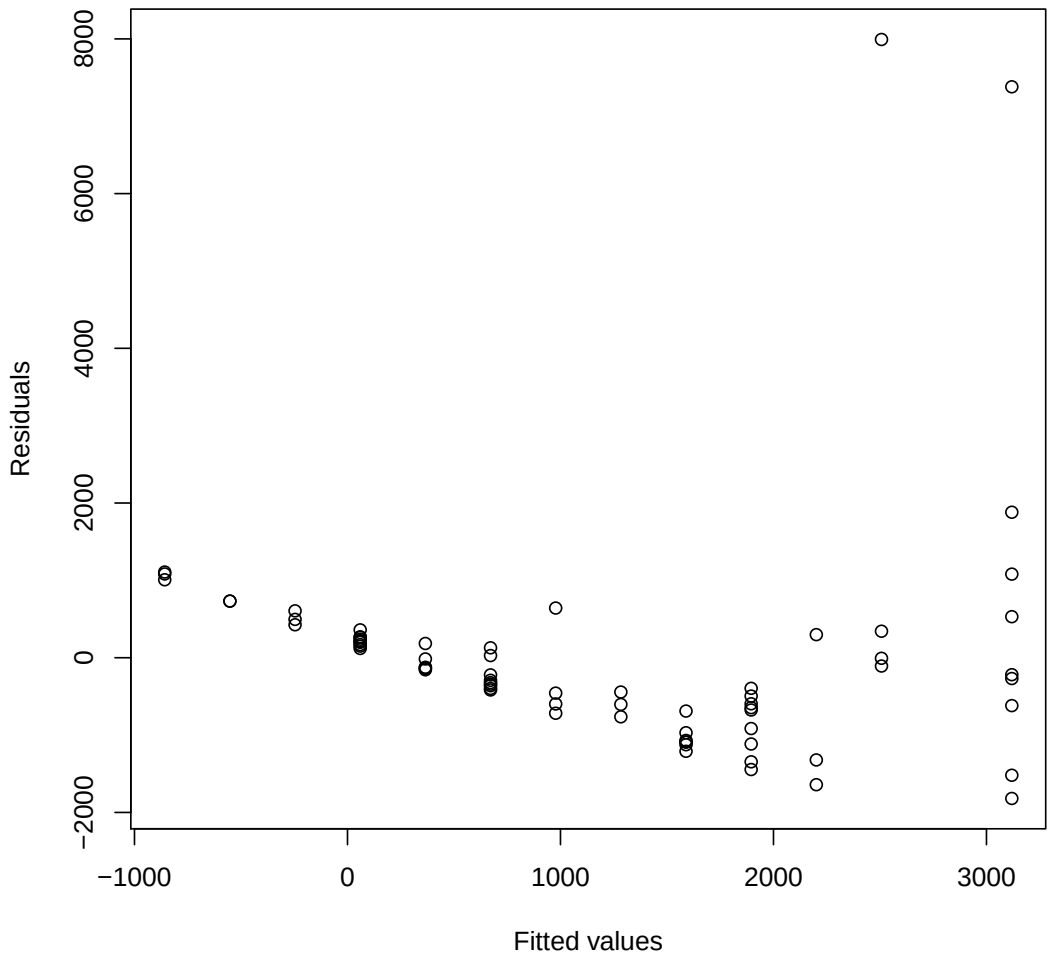
- Y – the prices (in pounds) on the wholesale brokers' auction market per dozen bottles, duty paid but excluding delivery and VAT in London in September 2003.
- X – Robert Parker uses a 100-point rating system with wines given a whole number score between 50 and 100 as follows:

96–100 points	Extraordinary
90–95 points	Outstanding
80–89 points	Above average to very good
70–79 points	Average
50–69 points	Below average to poor

Example: Predicting wine prices using critic scores



Residuals for wine prices example



Clearly, there is a U-shaped pattern to the residuals

Understanding RStudio output for the lm function

In addition to the estimates, the `lm` function also provides information to help us understand:

- The estimation variability, *i.e.* how close is our estimate to the true value?
- How useful is the variable for predicting the response?

```
> summary(fit_production)

Call:
lm(formula = RunTime ~ RunSize)

Residuals:
    Min      1Q  Median      3Q     Max
-28.597 -11.079   3.329   8.302  29.627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-squared:  0.7302,Adjusted R-squared:  0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```

Estimation variability

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	149.74770	8.32815	17.98	6.00e-13	***
RunSize	0.25924	0.03714	6.98	1.61e-06	***

Standard errors provide us with a measure of how close our estimate is to the true parameter values.

Estimation variability

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
```

t-value is Estimate/Standard error.

Estimation variability

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	149.74770	8.32815	17.98	6.00e-13	***
RunSize	0.25924	0.03714	6.98	1.61e-06	***

In regression problems, it is natural to be interested in whether or not $\beta = 0$ which is equivalent to x having no effect on y.

We can test the null hypothesis $H_0 : \beta = 0$ against the alternative hypothesis $H_1 : \beta \neq 0$ with a t-test.

If $\beta = 0$, then t-value follows a t_{n-2} distribution. The output gives the p-value of this test.

Estimation variability

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	149.74770	8.32815	17.98	6.00e-13	***
RunSize	0.25924	0.03714	6.98	1.61e-06	***

The `lm` function provides a "star-rating" for the t-statistic with the interpretation:

- . as "some evidence" (Pr between 0.05 and 0.1).
- * as "fairly strong evidence" (Pr between 0.01 and 0.05).
- ** as "strong evidence" (Pr between 0.001 and 0.01).
- *** as "very strong evidence" (Pr smaller than 0.001).

Therefore, we have very strong evidence that β (the effect of RunSize) is different to zero or we have strong evidence that RunTime depends on RunSize.

Measuring the strength of relationship: R^2 and $\bar{R}^2$

We may be interested in understanding the strength of relationship between x and y.

The coefficient of determination or multiple correlation coefficient is

$$R^2 = \frac{\text{RegSS}}{\text{TSS}} = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

- $0 < R^2 < 1$.
- A larger R^2 implies a better linear fit.

Measuring the strength of relationship: R^2 and $\bar{R}^2$

An alternative measurement of the strength of relationship is the adjusted coefficient of determination

$$\bar{R}^2 = 1 - \frac{\text{RSS}/(n-2)}{\text{TSS}/(n-1)} = 1 - \frac{\text{RMS}}{\text{TMS}}$$

where TMS is total mean square.

We will return to the difference between R^2 and $\bar{R}^2$ when we look at multiple regression.

```
> summary(fit_production)

Call:
lm(formula = RunTime ~ RunSize)

Residuals:
    Min      1Q  Median      3Q     Max
-28.597 -11.079   3.329   8.302  29.627

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 149.74770    8.32815   17.98 6.00e-13 ***
RunSize       0.25924    0.03714    6.98 1.61e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.25 on 18 degrees of freedom
Multiple R-squared:  0.7302,        Adjusted R-squared:  0.7152
F-statistic: 48.72 on 1 and 18 DF,  p-value: 1.615e-06
```