Medical Statistics with R

Dr. Gulser Caliskan
Prof. Giuseppe Verlato

Unit of Epidemiology and Medical Statistics
Department of Diagnostics and Public Health
University of Verona, Italy

LESSON 5 INDEX

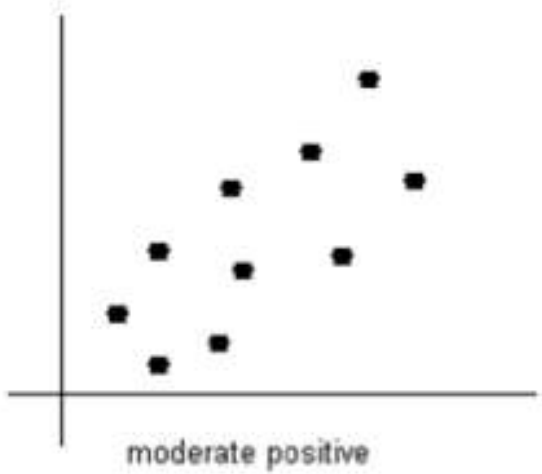
CORRELATION

A correlation exists between two variables when the values of one variable are somehow associated with the values of the other variable.

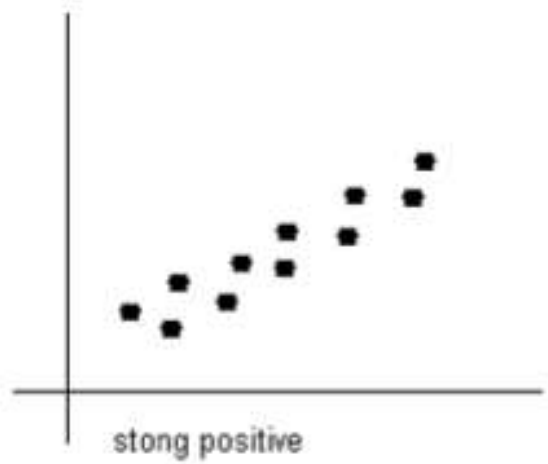
When you see a pattern in the data you say there is a correlation in the data. Though this lesson is only dealing with linear patterns, patterns can be exponential, logarithmic, or periodic.

To see this pattern, you can draw a scatter plot of the data. Remember to read graphs from left to right, the same as you read words. If the graph goes up the correlation is **positive** and if the graph goes down the correlation is **negative**.

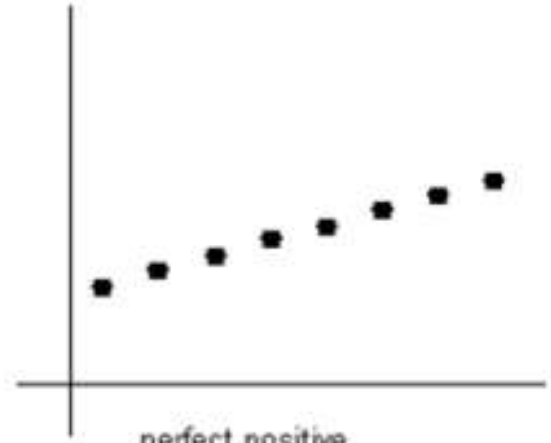
The words “**weak**”, “**moderate**”, and “**strong**” are used to describe the strength of the relationship between the two variables.



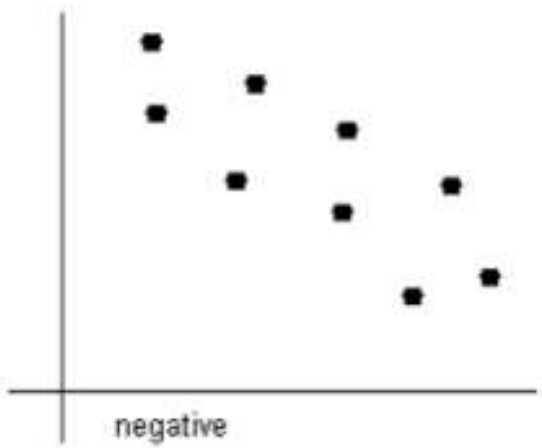
moderate positive



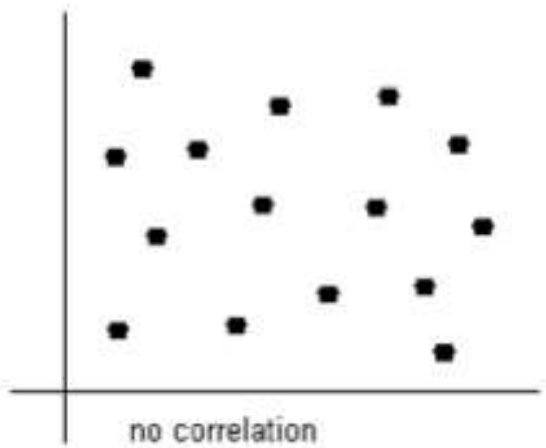
stong positive



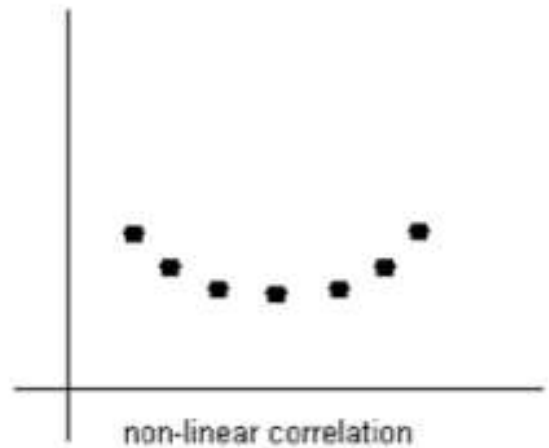
perfect positive
correlation



negative



no correlation



non-linear correlation

The linear correlation coefficient is a number that describes the strength of the linear relationship between the two variables. It is also called the Pearson correlation coefficient after Karl Pearson who developed it.

The symbol for the sample linear correlation coefficient is r. The symbol for the population correlation coefficient is **ρ** (Greek letter rho)

The formula for r is

$$r = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Where

$$SS_x = \sum (x - \bar{x})^2$$

$$SS_y = \sum (y - \bar{y})^2$$

$$SS_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

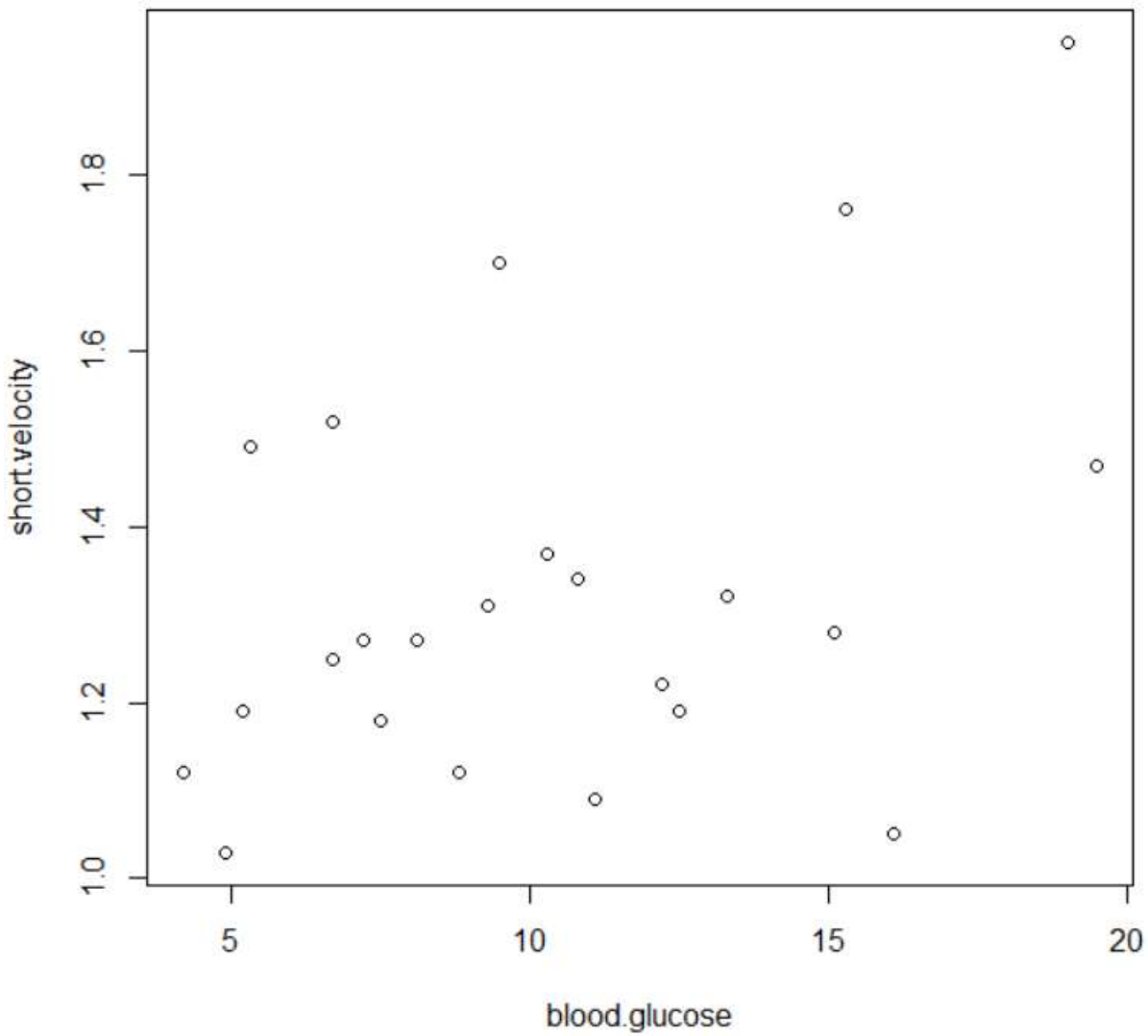
Interpretation of the correlation coefficient r is always between −1 and 1. r = −1 means there is a perfect negative linear correlation and r = 1 means there is a perfect positive correlation.

The closer r is to 1 or −1, the stronger the correlation. The closer r is to 0, the weaker the correlation.

CAREFUL: r = 0 does not mean there is no correlation. It just means there is **no linear correlation**. There might be a very strong curved pattern.

EXERCISE:

```
> library(ISwR)
> data(thuesen)
> attach(thuesen)
> thuesen
   blood.glucose short.velocity
1           15.3            1.76
2           10.8            1.34
3            8.1            1.27
4           19.5            1.47
5            7.2            1.27
6            5.3            1.49
7            9.3            1.31
8           11.1            1.09
9            7.5            1.18
10          12.2            1.22
11           6.7            1.25
12           5.2            1.19
13          19.0            1.95
14          15.1            1.28
15           6.7            1.52
16           8.6              NA
17           4.2            1.12
18          10.3            1.37
19          12.5            1.19
20          16.1            1.05
21          13.3            1.32
22           4.9            1.03
23           8.8            1.12
24           9.5            1.70
```


short.velocity
1.0
1.2
1.4
1.6
1.8
5
10
15
20
blood.glucose

```
> shapiro.test(blood.glucose)

        Shapiro-Wilk normality test

data:  blood.glucose
W = 0.94525, p-value = 0.2133

> shapiro.test(short.velocity)

        Shapiro-Wilk normality test

data:  short.velocity
W = 0.90033, p-value = 0.02568
```

```
> cor(blood.glucose,short.velocity)
[1] NA
> cor(blood.glucose,short.velocity,use="complete.obs")
[1] 0.4167546
> cor.test(blood.glucose,short.velocity)

        Pearson's product-moment correlation

data:  blood.glucose and short.velocity
t = 2.101, df = 21, p-value = 0.0479
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.005496682 0.707429479
sample estimates:
      cor
0.4167546
```

Spearman's rank Correlation Coefficient

As with the one- and two-sample problems, you may be interested in nonparametric variants. These have the advantage of not depending on the normal distribution and, indeed, being invariant to monotone transformations of the coordinates. The main disadvantage is that its interpretation is not quite clear.

A popular and simple choice is Spearman's rank correlation coefficient ρ. This is obtained quite simply by replacing the observations by their rank and computing the correlation. Under the null hypothesis of independence between the two variables the exact distribution of ρ can be calculated.

EXERCISE:

```
> cor.test(blood.glucose,short.velocity,method="spearman")

        Spearman's rank correlation rho

data:  blood.glucose and short.velocity
S = 1380.4, p-value = 0.1392
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.318002

Warning message:
In cor.test.default(blood.glucose, short.velocity, method = "spearman") :
  Cannot compute exact p-value with ties
```

REGRESSION:

Regression analysis is a set of statistical methods used for the estimation of relationships between a **dependent variable** and one or more **independent variables**. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them.

There are numerous types of regression models that you can use. This choice often depends on the kind of data you have for the dependent variable and the type of model that provides the best fit.

Types Of Regression

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Quantile Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Principal Components Regression (PCR)
- Partial Least Squares (PLS) Regression
- Support Vector Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Quasi Poisson Regression
- Cox Regression
- Tobit Regression

SIMPLE LINEAR REGRESSION ANALYSIS

We consider the modelling between the dependent and one independent variable. When there is only one independent variable in the linear regression model, the model is generally termed as a **simple linear regression** model.

When there are more than one independent variables in the model, then the linear model is termed as the **multiple linear regression** model.

Consider a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Y_i is the observed response or dependent variable for observation
- x_i is the observed predictor, regressor, explanatory variable, independent variable, covariate
- e_i is the error term

The terms $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are the parameters of the model. The parameter $\boldsymbol{\beta_0}$ is termed as an intercept term, and the parameter $\boldsymbol{\beta_1}$ is termed as the slope parameter. These parameters are usually called as **regression coefficients**.

ASSUMPTIONS

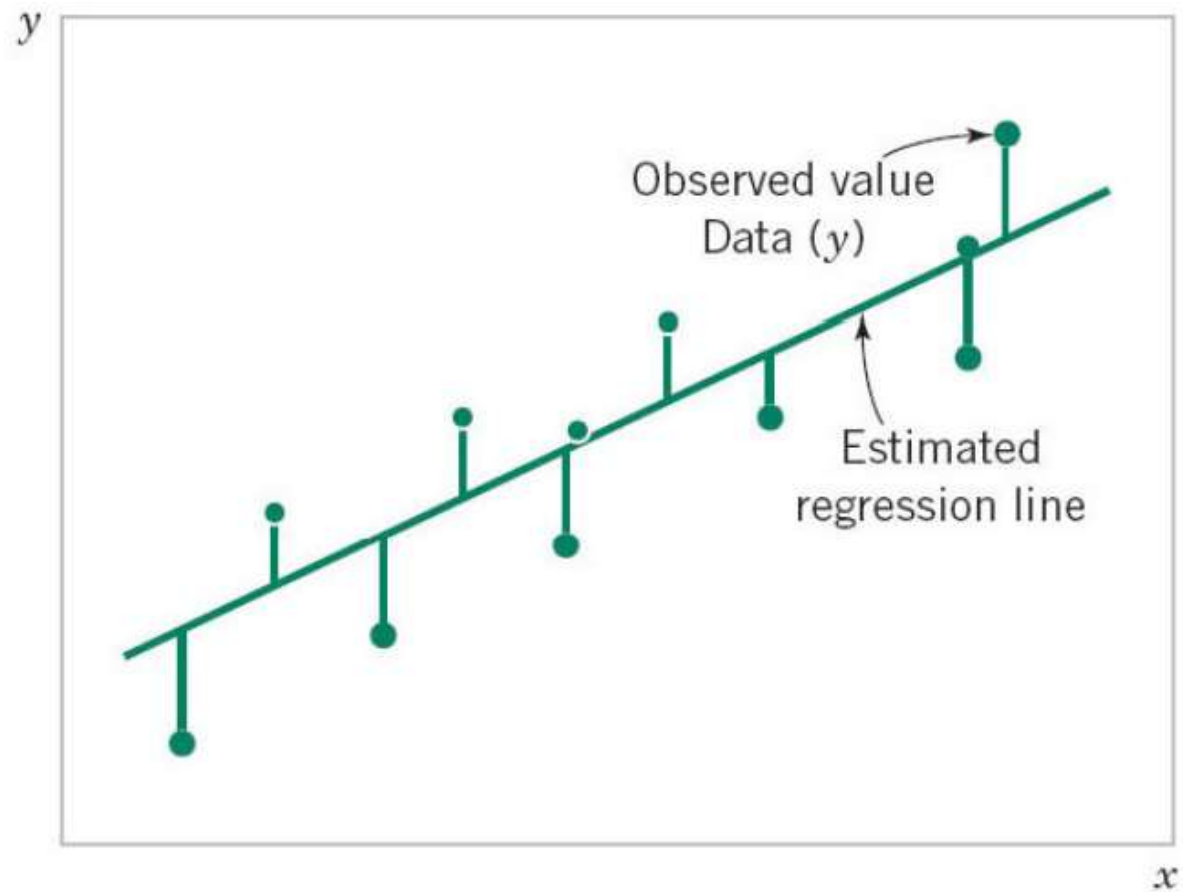
- Linear Relationship Exists Between Y And X
- Independent Errors
- Constant Variance Of Errors
- Normally Distributed Errors

ESTIMATION

We wish to use the sample data to estimate the population parameters: the slope $\beta 1$ and the intercept $\beta 0$.

Least Squares Estimation

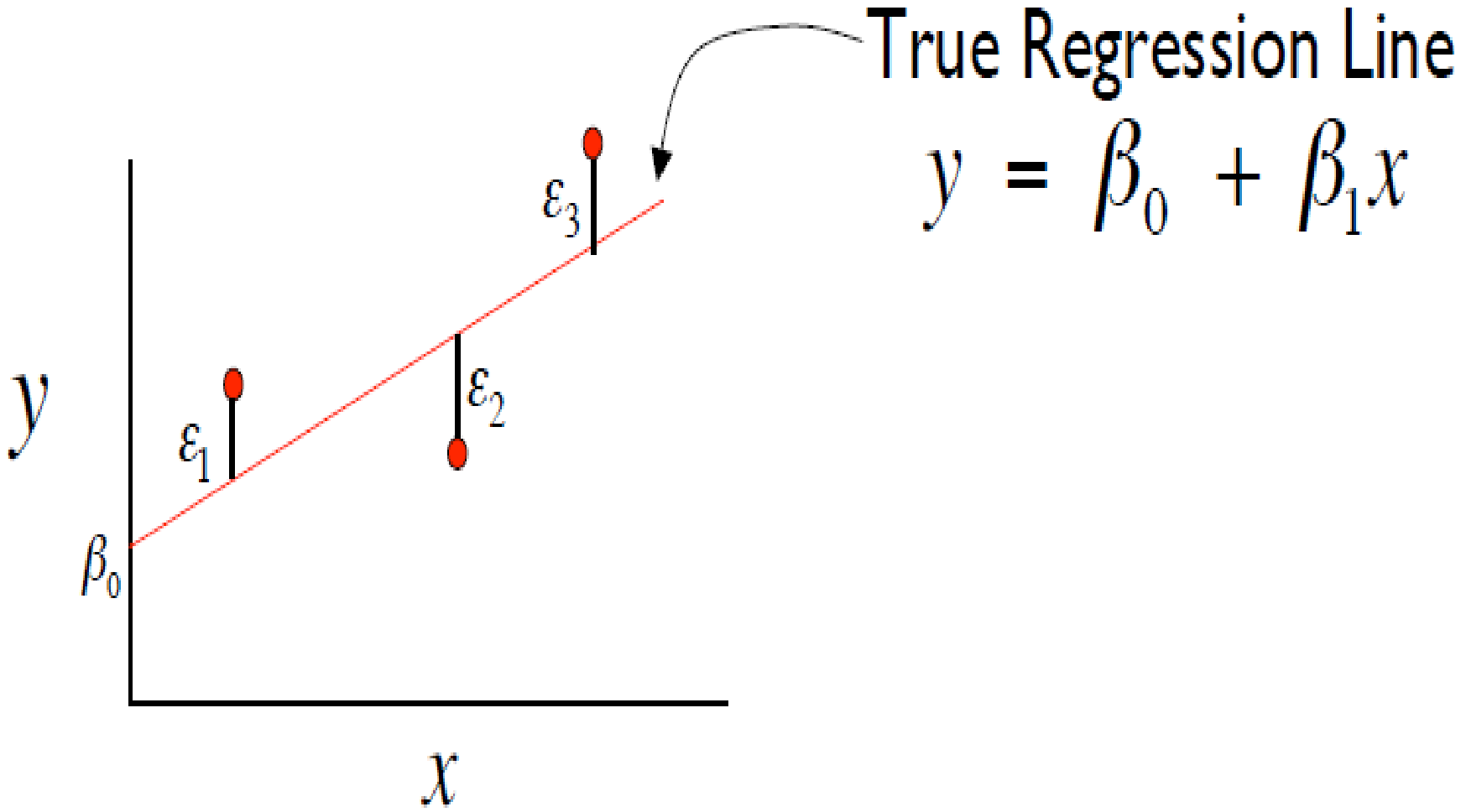
To choose the ‘best fitting line’ using least squares estimation, we minimize the sum of the squared vertical distances of each point to the fitted line.


y
Observed value
Data (y)
Estimated
regression line
x

We let ‘hats’ denote predicted values or estimates of parameters, so we have:

This vertical distance of a point from the fitted line is called a **residual**. The residual for observation *i* is denoted e_i an

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


True Regression Line
$y = \beta_0 + \beta_1 x$
y
x
β_0
ε_1
ε_2
ε_3

Estimate of the slope:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

Estimate of the Y-intercept:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Residuals Are Useful!

- They allow us to calculate the error sum of squares (SS_{res}):

$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- whereas the term $\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ describes the proportion of variability explained by the regression,

$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Analysis of variance for testing $H_0 : \beta_1 = 0$

Source of variation	Sum of squares	Degrees of freedom	Mean square	F
Regression	SS_{reg}	1	MS_{reg}	MS_{reg} / MSE
Residual	SS_{res}	$n-2$	MSE	
Total	s_{yy}	$n-1$		

We can also frame this in our now familiar ANOVA framework

- partition total variation into two components: $\boldsymbol{SS_{res}}$ (unexplained
- variation) and $\boldsymbol{SS_{reg}}$ (variation explained by linear model)

Hypothesis Test Of Individual Regression Coefficients

- Hypothesis tests for $\hat{\beta}_1$ can be done by simple t-test:

$$H_0 : \hat{\beta}_1 = 0$$
$$H_1 : \hat{\beta}_1 \neq 0$$

$$T = \frac{\hat{\beta}_1 - \beta_1}{se(\beta_1)}$$

Critical value : $\boldsymbol{t_{\alpha/2, n-(k-1)}}$

Confidence intervals are equally easy to obtain: $\boldsymbol{\hat{\beta}_1 \pm t_{\alpha/2, n-(k-1)} . se(\hat{\beta}_1)}$

GOODNESS OF FIT OF REGRESSION

It can be noted that a fitted model can be said to be good when residuals are small. Since SS_{res} is based on residuals, so a measure of the quality of a fitted model can be based on SS_{res}.

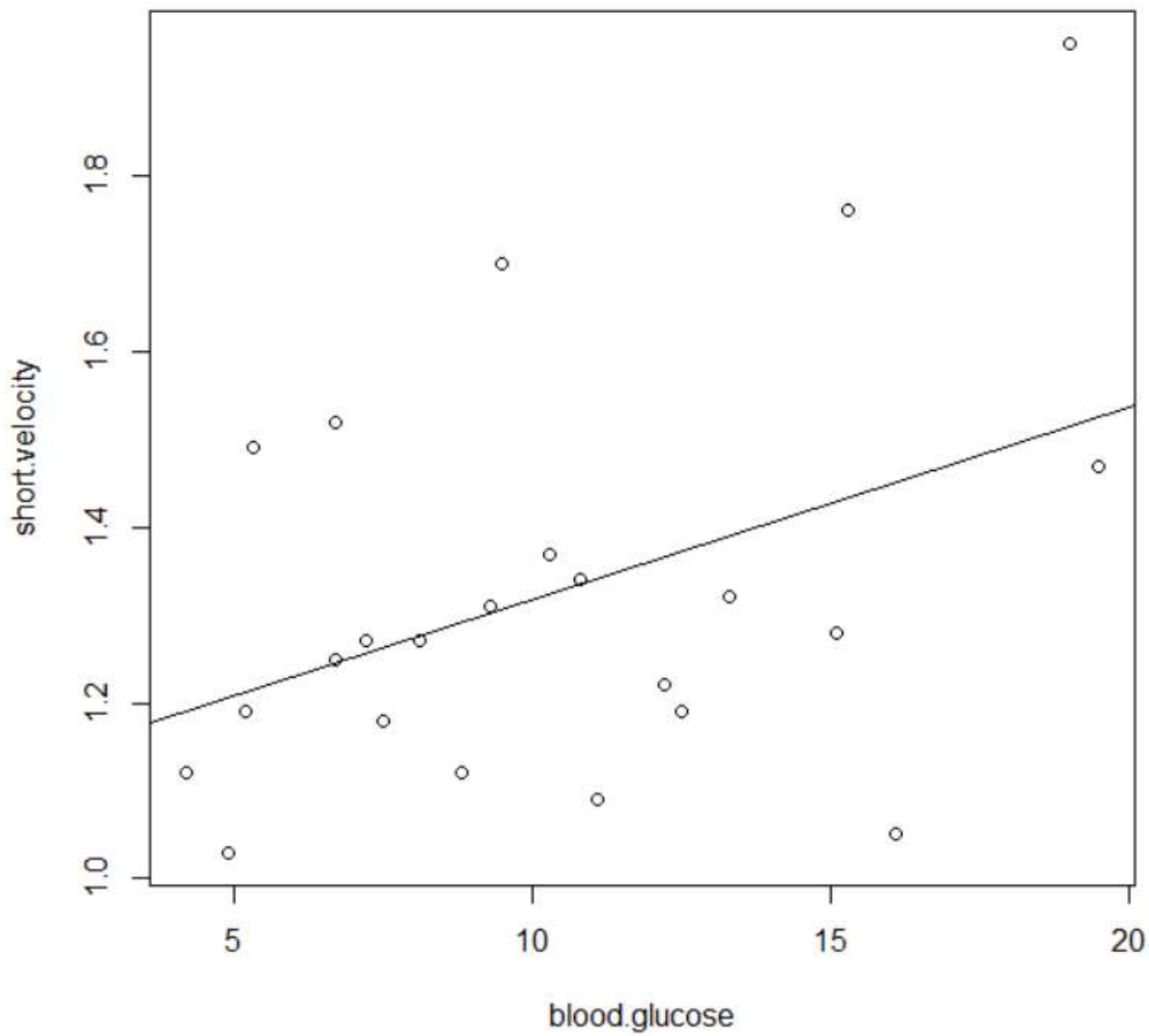
When the intercept term is present in the model, a measure of goodness of fit of the model is given by

$$R^2 = 1 - \frac{SS_{res}}{S_{yy}} = \frac{SS_{reg}}{S_{yy}}$$

This is known as the coefficient of determination. This measure is based on the concept that how much variation in y ’s stated by yy s is explainable by $\boldsymbol{SS_{reg}}$ and how much unexplainable part is contained in $\boldsymbol{SS_{res}}$.

EXERCISE:

```
> library(ISwR)
> data(thuesen)
> attach(thuesen)
> thuesen
   blood.glucose short.velocity
1           15.3            1.76
2           10.8            1.34
3            8.1            1.27
4           19.5            1.47
5            7.2            1.27
6            5.3            1.49
7            9.3            1.31
8           11.1            1.09
9            7.5            1.18
10          12.2            1.22
11           6.7            1.25
12           5.2            1.19
13          19.0            1.95
14          15.1            1.28
15           6.7            1.52
16           8.6              NA
17           4.2            1.12
18          10.3            1.37
19          12.5            1.19
20          16.1            1.05
21          13.3            1.32
22           4.9            1.03
23           8.8            1.12
24           9.5            1.70
```


short.velocity
1.0
1.2
1.4
1.6
1.8
5
10
15
20
blood.glucose

```
> lm(short.velocity~blood.glucose)

Call:
lm(formula = short.velocity ~ blood.glucose)

Coefficients:
  (Intercept)  blood.glucose
      1.09781        0.02196


> summary(lm(short.velocity~blood.glucose))

Call:
lm(formula = short.velocity ~ blood.glucose)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40141 -0.14760 -0.02202  0.03001  0.43490

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.09781    0.11748   9.345 6.26e-09 ***
blood.glucose  0.02196    0.01045   2.101   0.0479 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2167 on 21 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.1737,    Adjusted R-squared:  0.1343
F-statistic: 4.414 on 1 and 21 DF,  p-value: 0.0479
```

```
> lm.velo <- lm(short.velocity~blood.glucose)
> lm.velo

Call:
lm(formula = short.velocity ~ blood.glucose)

Coefficients:
  (Intercept)  blood.glucose
      1.09781        0.02196
```

```
> fitted(lm.velo)
       1        2        3        4        5        6        7        8
1.433841 1.335010 1.275711 1.526084 1.255945 1.214216 1.302066 1.341599
       9       10       11       12       13       14       15       17
1.262534 1.365758 1.244964 1.212020 1.515103 1.429449 1.244964 1.190057
      18       19       20       21       22       23       24
1.324029 1.372346 1.451411 1.389916 1.205431 1.291085 1.306459
```

```
> predict(lm.velo,int="c")
        fit      lwr      upr
1  1.433841 1.291371 1.576312
2  1.335010 1.240589 1.429431
3  1.275711 1.169536 1.381887
4  1.526084 1.306561 1.745607
5  1.255945 1.139367 1.372523
6  1.214216 1.069315 1.359118
7  1.302066 1.205244 1.398889
8  1.341599 1.246317 1.436881
9  1.262534 1.149694 1.375374
10 1.365758 1.263750 1.467765
11 1.244964 1.121641 1.368287
12 1.212020 1.065457 1.358583
13 1.515103 1.305352 1.724854
14 1.429449 1.290217 1.568681
15 1.244964 1.121641 1.368287
17 1.190057 1.026217 1.353898
18 1.324029 1.230050 1.418008
19 1.372346 1.267629 1.477064
20 1.451411 1.295446 1.607377
21 1.389916 1.276444 1.503389
22 1.205431 1.053805 1.357057
23 1.291085 1.191084 1.391086
24 1.306459 1.210592 1.402326
```

```
> resid(lm.velo)
           1            2            3            4            5            6
 0.326158532  0.004989882 -0.005711308 -0.056084062  0.014054962  0.275783754
           7            8            9           10           11           12
 0.007933665 -0.251598875 -0.082533795 -0.145757649  0.005036223 -0.022019994
          13           14           15           17           18           19
 0.434897199 -0.149448964  0.275036223 -0.070057471  0.045971143 -0.182346406
          20           21           22           23           24
-0.401411486 -0.069916424 -0.175431237 -0.171085074  0.393541161
```

```
> res<-resid(lm.velo)
> shapiro.test(res)

        Shapiro-Wilk normality test

data:  res
W = 0.92413, p-value = 0.08173
```