11

Simple Linear Regression and Correlation

LO

- Compute various sums of squares for a set of data pairs.
- Determine the equation of linear regression and the correlation coefficient; use linear regression to predict future values.
- Perform a hypothesis test in simple linear regression; test the significance of regression using t-test and F-test.

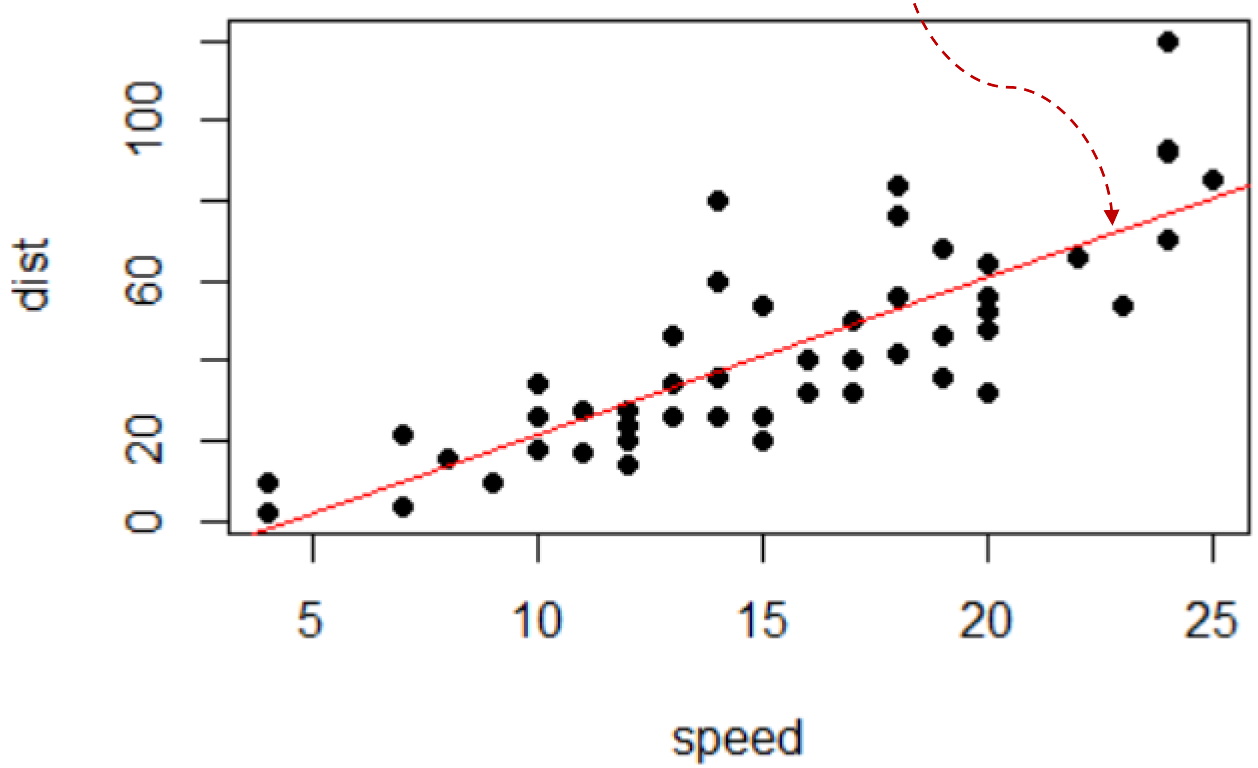
Introduction

- Speed and stopping distance of cars.

dist = -17.58 + 3.93speed

$$dist = \beta_0 + \beta_1 speed$$

	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
.	.	.
.	.	.
.	.	.
48	24	93
49	24	120
50	25	85



```
> summary(lm(cars$dist~cars$speed))

Call:
lm(formula = cars$dist ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max 
-29.069  -9.525  -2.272   9.215  43.201 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6438 
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

?

Regression (an empirical model)

We have two variables X, Y (numerical data)

We believe that Y depends in some way on X: $Y = f(X)$

Dependent variable
Response variable

Independent variable
Predictor
Explanatory variable
Regressor

Examples of (X, Y) pairs:

- X = study time and Y = score on a test.
- X = smoking frequency and Y = age of first heart attack.

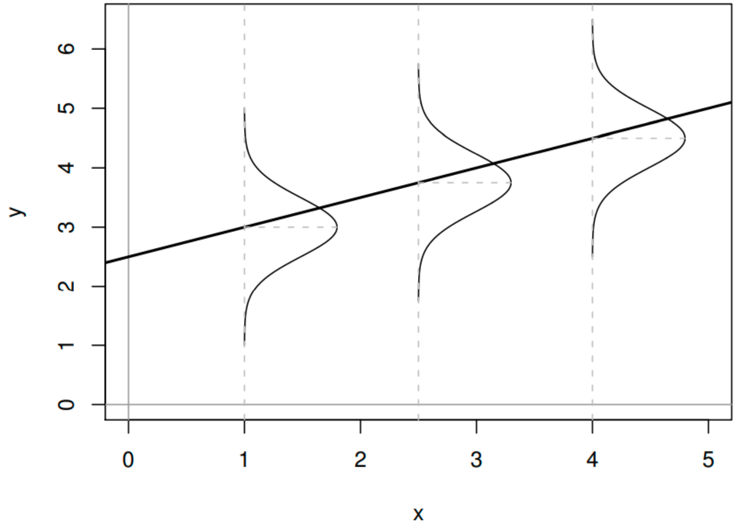
Given information about X and Y, we would like to predict future values of Y for particular values of X. ➔ Estimate $E[Y \mid X = x]$.

Simple linear regression

Intercept Slope

- $E[Y \mid X = x] = \beta_0 + \beta_1 x$,
- β_0 and β_1 are unknown regression coefficients ➔ to be estimated.
- We assume that each observation, Y, can be described by the model $Y = \beta_0 + \beta_1 x + \varepsilon$, // ε : random error
- $\varepsilon \sim N(0, \sigma^2)$

We never know exactly β_0, β_1, ε, σ^2
➔ Estimation, hypothesis testing on these parameters ➔ prediction



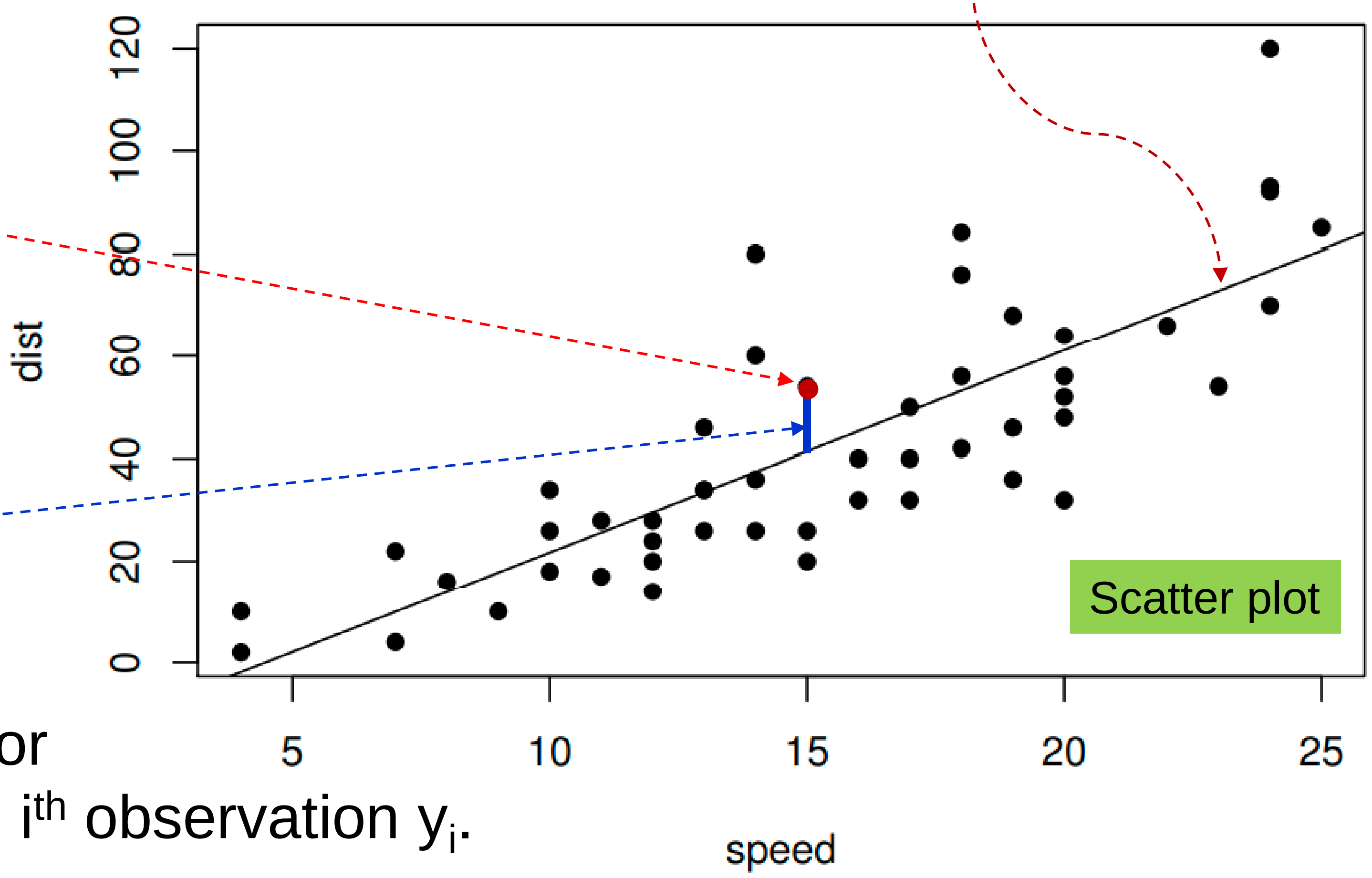
Example - cars

- Speed and stopping distance of cars.

Estimated regression line
$\hat{y}$ = -17.58 + 3.93x

Data point (x_{26}, y_{26}) = (speed = 15, dist = 54)

$e_{26} = y_{26} - (-17.58 + 3.93x_{26})$
$= 12.63$



Error term $e_i = y_i - \hat{y}_i$, is called i^{th} residual, the error in the fit of the model to the i^{th} observation y_i.

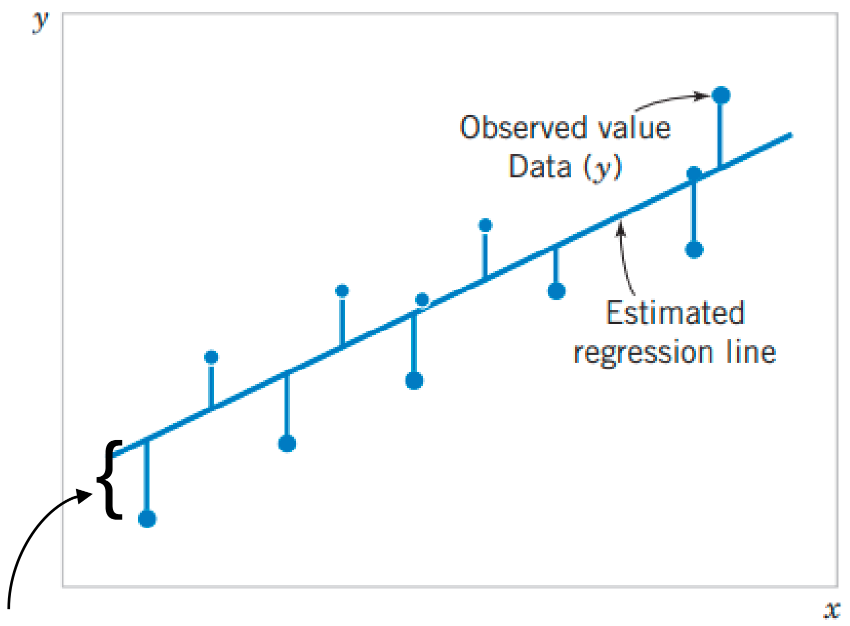
Method of least squares

- The **method of least squares** is used to estimate the parameters, β_0 and β_1 by minimizing L, the sum of the squares of the vertical deviations

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$



Error term $e_i = y_i - \hat{y}_i$, or residual

Least squares estimates of β_0, β_1

Least squares normal equations

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i \end{cases}$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{xy} = \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

$$\Rightarrow \begin{cases} \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n} = S_{xy}/S_{xx} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \end{cases}$$

Cars - Estimates of β_0, β_1

True regression line

dist $= \beta_0 + \beta_1$speed

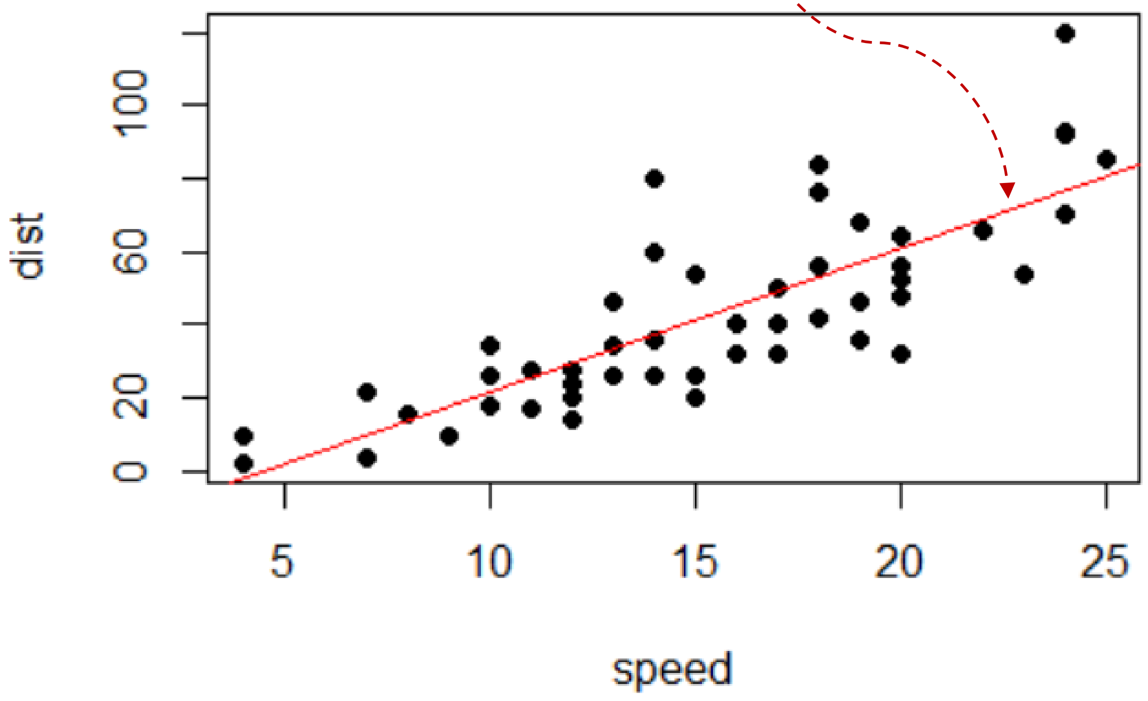
Estimated regression line

dist = -17.58 + 3.93speed

```
> cars.lm$coefficients
(Intercept)  cars$speed
 -17.579095    3.932409
```

Estimated regression line

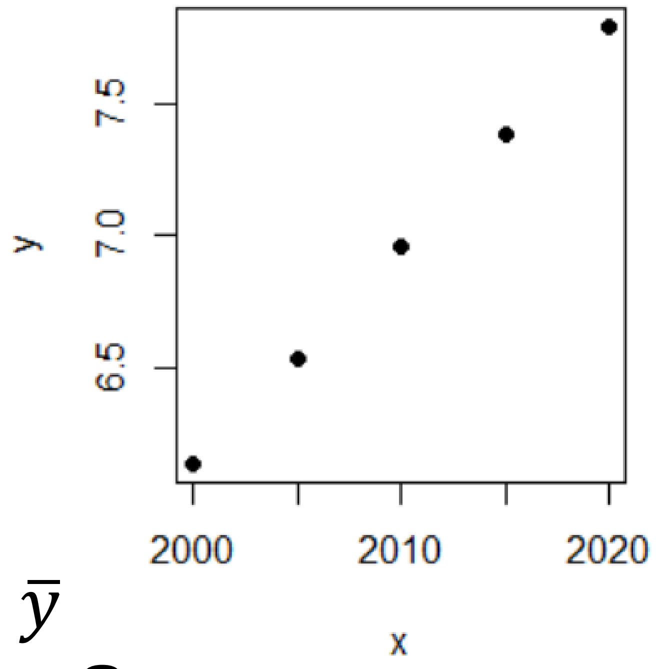
y = -17.58 + 3.93x



Exercise - World Population

Year	**World Population (B)**
2000	6.14
2005	6.54
2010	6.96
2015	7.38
2020	7.79

Population of the world, 2000 – 2020
(Source: worldometers.info)



a/ Compute $\bar{x}$, $\bar{y}$

b/ Compute S_{xx}, S_{xy}

c/ Find the estimated regression line

d/ Predict the world population in 2025

Exercise - World Population

x	2000	2005	2010	2015	2020
y	6.14	6.54	6.96	7.38	7.79

a/ $\bar{x}$ = 2010, $\bar{y}$ = 6.962

b/ $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ = 250, $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ = 20.7

c/ Find the estimated regression line

$\hat{\beta}_1 = S_{xy}/S_{xx} = 0.0828$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -159.466 \Rightarrow y = 0.0828x - 159.466$

d/ Predict the world population in 2025

y(2025) = 0.0828*(2025) - 159.466 = 8.204 (B)

Estimating σ^2
Error sum of squares SS_E

- Error Sum of Squares SS_E

$$\boxed{SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}\left(y_i - y_i\right)^2 = SS_T - \beta_1 S_{xy}}$$

- $E(SS_E) = (n - 2)\sigma^2$
- Estimate σ^2 using SS_E

$$\boxed{\hat{\sigma}^2 = \frac{SS_E}{n-2}}$$

$$SS_T = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

Total Sum of Squares

```
> x<-cars$speed
> y<-cars$dist
> n<-length(x)
> Sxy <- sum((x-mean(x))*(y-mean(y)))
> Sxx<-sum((x-mean(x))^2)
> betha1hat<- Sxy/Sxx
> SST<- sum((y-mean(y))^2)
> SSE<-SST-betha1hat*Sxy
> Sigmahat <- sqrt(SSE/(n-2))
> Sigmahat
[1] 15.37959
```

Properties of the Least Squares Estimators

- $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_0) = \beta_0$
- $V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}}$
- $V(\hat{\beta}_0) = \sigma^2\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right]$
- Estimated Standard Errors: $se\left(\beta_1\right) = \sqrt{\dfrac{\sigma^2}{S_{xx}}}$ and $se\left(\beta_0\right) = \sqrt{\sigma^2\left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{S_{xx}}\right)}$

```
> Se1<- sqrt(Sigmahat^2/Sxx)
> Se0 <- sqrt(Sigmahat^2*(1/n+mean(x)^2/Sxx))
> c(Se1,Se0)
[1] 0.4155128 6.7584402
```

Exercises

x:	1	2	3	4	5
y:	3	4	4	5	6

a/ Compute SS_E, SS_T

b/ Estimate σ^2

c/ Estimate standard error

of the slope and the intercept

Exercise – World population

x	2000	2005	2010	2015	2020
y	6.14	6.54	6.96	7.38	7.79

a/ Compute SS_E, SS_T

b/ Estimate σ^2

c/ Estimate standard error of the slope and the intercept

```
> x<-c(2000,2005,2010,2015,2020)
> y<-c(6.14,6.54,6.96,7.38,7.79)
> n<-length(x)
> Sxy <- sum((x-mean(x))*(y-mean(y)))
> Sxx<-sum((x-mean(x))^2)
> betha1hat<- Sxy/Sxx
> SST<- sum((y-mean(y))^2)
> SSE<-SST-betha1hat*Sxy
> Sigmahat <- sqrt(SSE/(n-2))
> Se1<- sqrt(Sigmahat^2/Sxx)
> Se0 <- sqrt(Sigmahat^2*(1/n+mean(x)^2/Sxx))
> c(SSE,SST,Sigmahat^2,Se1,Se0)
[1] 0.000120 1.714080 0.000040 0.000400 0.804005
```

Exercise

- Show that

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

$$SS_T = SS_R + SS_E$$

ANOVA

ANOVA Identity

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

$$SS_T = SS_R + SS_E$$

Total variation = Explained variation + Unexplained variation.

SS_R: Regression Sum of Squares
➔ (variation explained by linear model)

SS_E: Error Sum of Squares
➔ (unexplained variation)

F-test

F-test: $$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E}$$

$F_{1,n-2}$ distribution

Reject H_0: $\beta_1 = 0$ if $f_0 > f_{\alpha,1,n-2}$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

We reject H_0: $\beta_1 = 0$ when F is large – that is, when the explained variation is large relative to the unexplained variation.

ANOVA

```
> anova(cars.lm)

Analysis of Variance Table

Response: dist
          Df Sum Sq Mean Sq F value    Pr(>F)
speed      1  21186 21185.5  89.567 1.490e-12 ***
Residuals 48  11354   236.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Verify $f_0 > f_{\alpha,1,n-2}$

Here we see that the F statistic is 89.57 with a p-value very close to zero. Conclusion: there is very strong evidence that $H_0 : \beta_1 = 0$ is false, that is, there is strong evidence that $\beta_1 \neq 0$. Moreover, we conclude that the regression relationship between dist and speed is significant.

Exercise - World population

x	2000	2005	2010	2015	2020
y	6.14	6.54	6.96	7.38	7.79

Complete the ANOVA table

Source	Sum of squares	Degrees of freedom	Mean squares	F_0
Regression	SS_R = ?	1	$MS_R = SS_R/1$ = ?	MS_R/MS_E = ?
Error	SS_E = ?	n - 2 = ?	$MS_E = SS_E/(n-2)$ = ?	
Total	SS_T = ?	n -1 = ?		

t-test on β_1

- Suppose we wish to test:
- H_0: $\beta_1 = \beta_{1,0}$
- H_1: $\beta_1 \neq \beta_{1,0}$
- Test statistic $T_0 = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \dfrac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$

We would reject H_0:

$|t_0| > t_{\alpha/2,\, n-2}$

the t distribution with n - 2 degrees of freedom

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

t-test on β_0

$$H_0\colon \beta_0 = \beta_{0,0}$$

$$H_1\colon \beta_0 \neq \beta_{0,0}$$

We would reject H_0:

$|t_0| > t_{\alpha/2,\, n-2}$

use the statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)}$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -17.5791     6.7584  -2.601   0.0123 *  
cars$speed    3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H_0: $\beta_0 = 0$

H_1: $\beta_0 \neq 0$

Regression and correlation

- *Covariance* *Cov(X,Y)* is a measure of *linear relationship* between the random variables X and Y.
- *Correlation coefficient*

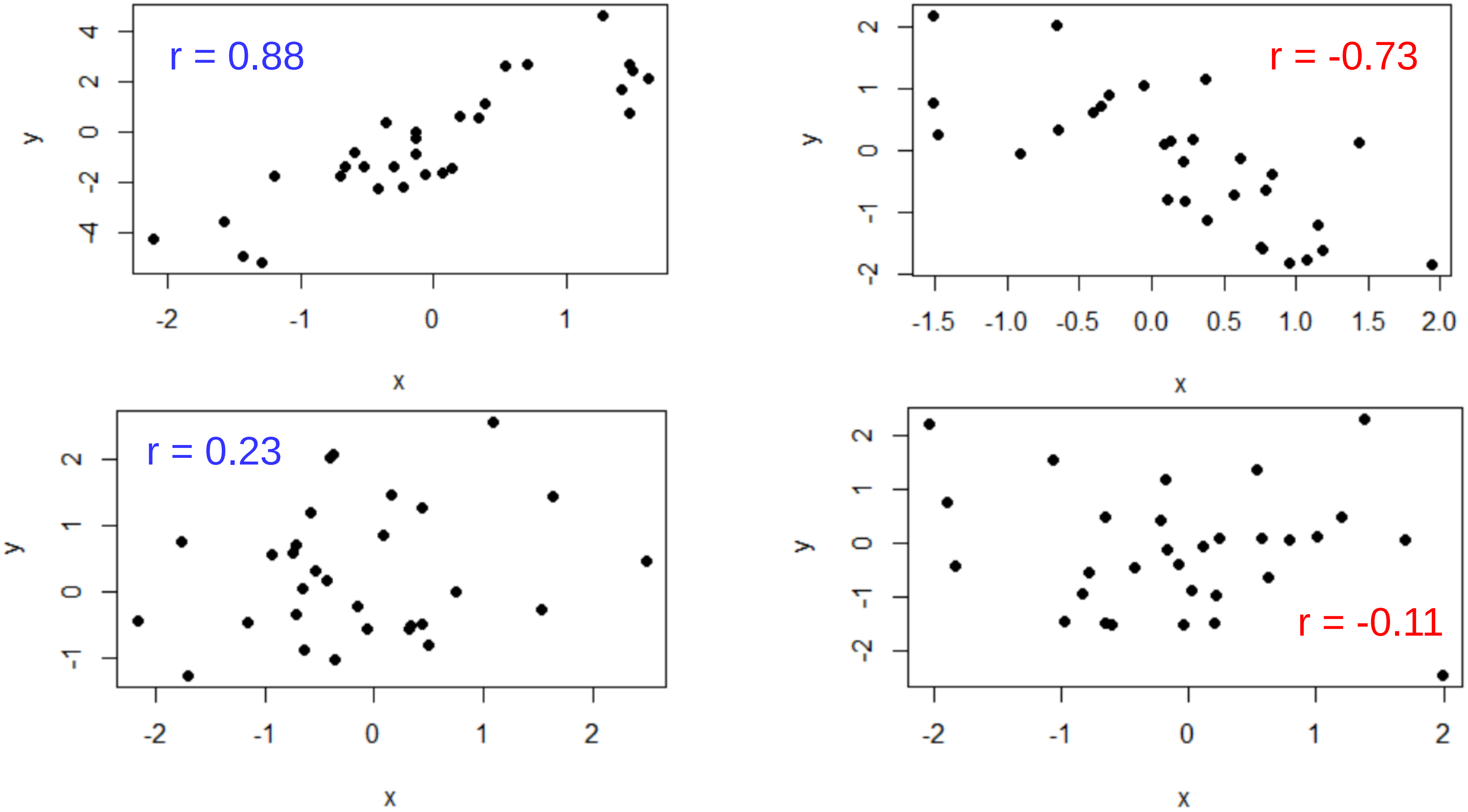
$$\rho = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$

$$-1 \leqslant \rho \leqslant +1$$

- Sample correlation coefficient R

$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX}SS_T)^{1/2}}$$

Sample correlation and scatter plot



Test Statistic for Zero Correlation

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

Reject H_0: $\rho = 0$ if

$t_0 > t_{\alpha/2, n-2}$

t distribution with n - 2 degrees of freedom if H_0: $\rho = 0$ is true

$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX}SS_T)^{1/2}}$$

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{SS_T} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$$

Note.

$R^2 = 1 - SS_E/SS_T$

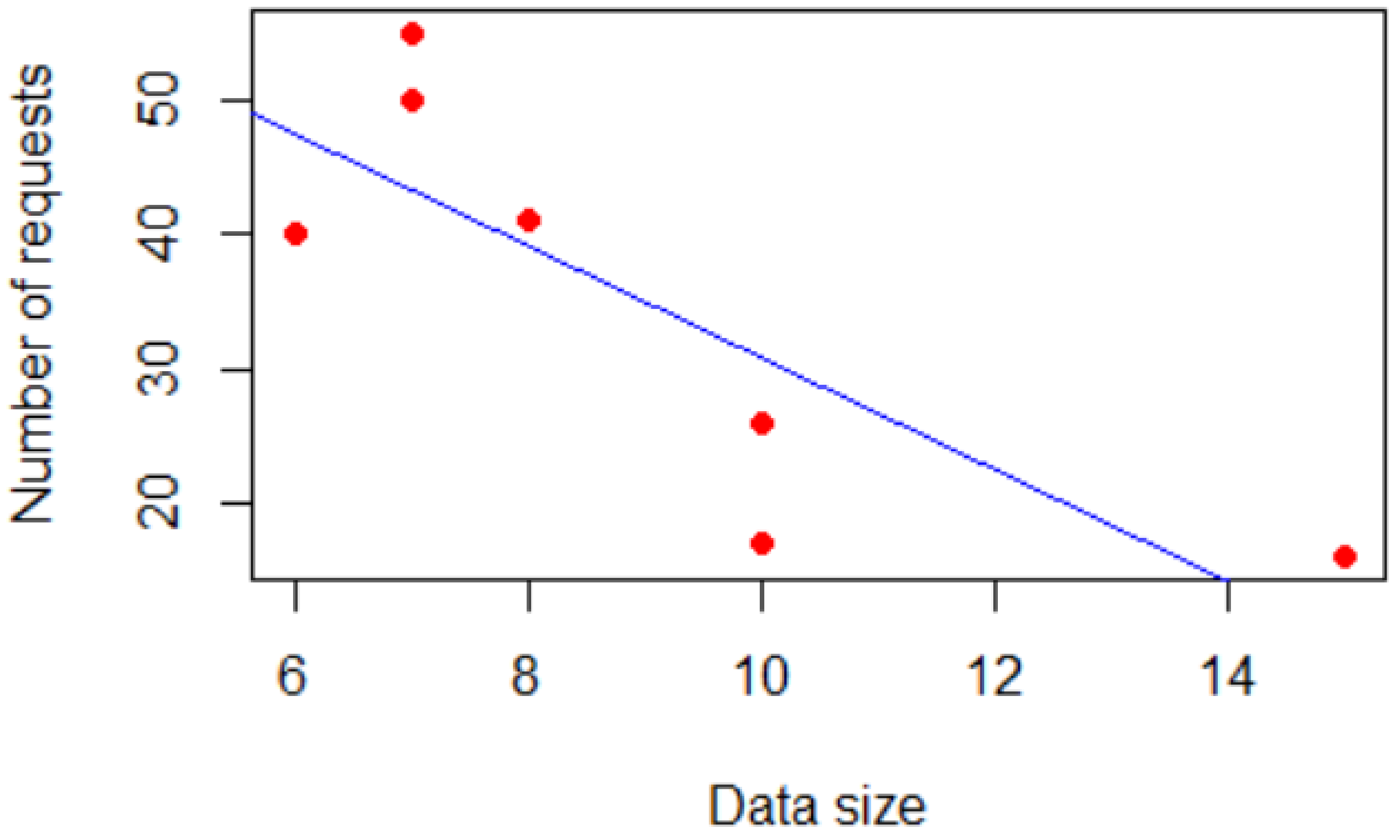
$0 \leqslant R^2 \leqslant 1$: coefficient of determination

Example

(Efficiency of computer programs). A computer manager needs to know how efficiency of her new computer program depends on the size of incoming data. Efficiency will be measured by the number of processed requests per hour. Applying the program to data sets of different sizes, she gets the following results,

x (data size, Gigabytes)	6	7	7	8	10	10	15
y (processed requests)	40	55	50	41	17	26	16

In general, larger data sets require more computer time, and therefore, fewer requests are processed within 1 hour. The response variable here is the number of processed requests (y), and we attempt to predict it from the size of a data set (x).



Number of requests
20
30
40
50
6
8
10
12
14
Data size

Example (cont.)

1/ The regression line

$n = 7$, $\bar{x} = 9$, $\bar{y} = 35$, $S_{xx} = 56$, $S_{xy} = -232$, $S_{yy} = 1452$.

Estimates of the slope and the intercept

$\hat{\boldsymbol{\beta}}_1 = S_{xy}/S_{xx} = -4.14$, $\hat{\beta}_0 = \bar{y} - \hat{\boldsymbol{\beta}}_1 * \bar{x} = 35 - (-4.14)9 = 72.29$

➔The estimated regression line: $y = 72.29 - 4.14x$

➔ Increasing incoming data sets by 1 gigabyte, we expect to process 4.14 fewer requests per hour.

Example (cont.)

2/ ANOVA table

We have $SS_T = S_{yy} = 1452$ partitioned into $SS_R = \hat{\beta}_1^2 S_{xx} = 961$ and $SS_E = SS_T - SS_R = 491$.

Source	**Sum of squares**	**Degrees of freedom**	**Mean squares**	F_0
Regression	$SS_R = 961$	1	$MS_R = SS_R/1 = 961$	$MS_R/MS_E = 9.79$
Error	$SS_E = 491$	$n - 2 = 5$	$MS_E = SS_E/(n-2) = 98.2$	
Total	$SS_T = 1452$	$n - 1 = 6$		

Example (cont.)

3/ (t-test on the slope β_1) Does the number of processed requests really depend on the size of data sets?

We wish to test

$H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$

T-statistic:

$$t_0 = \frac{\beta_1}{\sqrt{\sigma^2 / S_{xx}}} = \frac{-4.14}{\sqrt{98.2 / 56}} = -3.13$$

Use $\alpha = 0.05$, $3.13 = |t_0| > t_{\alpha/2,n-2} = t_{0.025,5} = 2.571$ ➔ Reject H_0 at the 0.05 level of significance

Example (cont.)

4/ ANOVA F-test

A similar result is suggested by the F-test.

$f_0 = MS_R/MS_E = 9.79$

$9.79 = f_0 > f_{\alpha,1,n-2} = f_{0.05,1,5} = 6.61$ ➔ Reject H_0: $\beta_1 = 0$ at the 0.05 level of significance

5/ R^2 (R-square)

$R^2 = SS_R/SS_T = 961/1452 = 0.6619$

➔ 66.19% of the total variation of the number of processed requests is explained by sizes of data sets only.

Example (cont.)

6/ t-test on correlation

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \bar{X})}{\left[\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX}SS_T)^{1/2}}$$

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{SS_T} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$$

$t_0 = -3.13$

➔ Reject H_0: $\rho = 0$ at the 0.05 level

ρ: *correlation coefficient*

```
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
      1       2       3       4       5       6       7 
 -7.429  11.714   6.714   1.857 -13.857  -4.857   5.857 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   72.286     12.491   5.787  0.00217 **
x             -4.143      1.324  -3.129  0.02599 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.908 on 5 degrees of freedom
Multiple R-squared:  0.6619,	Adjusted R-squared:  0.5943 
F-statistic:  9.79 on 1 and 5 DF,  p-value: 0.02599
```

x	6	7	7	8	10	10	15
y	40	55	50	41	17	26	16

THANKS