ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

Elements of Statistics and Probability

LECTURE 5 – Simple Regression

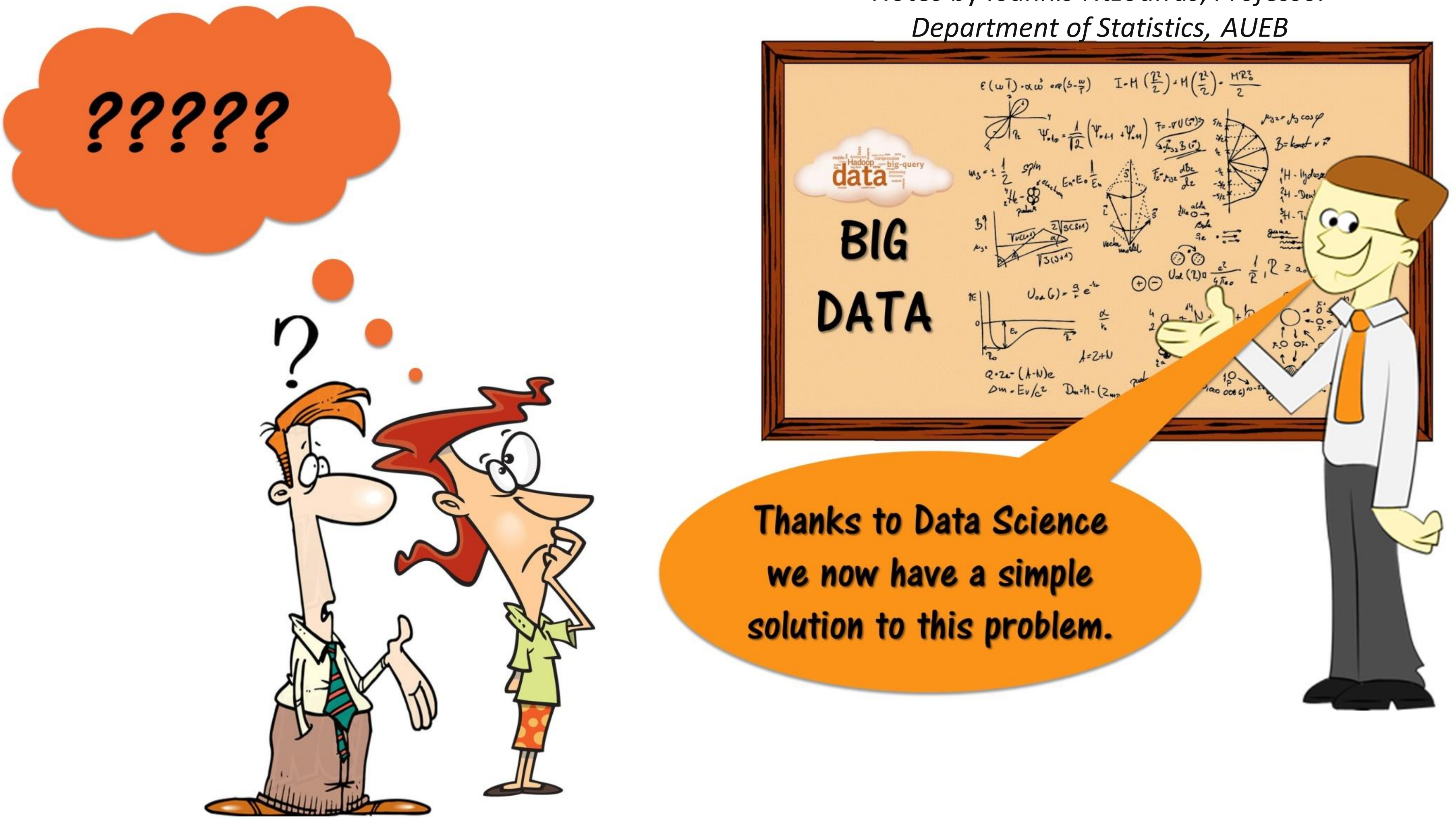
Xanthi Pedeli

Assistant Professor, xpedeli@aueb.gr

Department of Statistics, AUEB

Notes by Ioannis Ntzoufras, *Professor*

Department of Statistics, AUEB



5. *Correlation and Regression models*
Contents

- Introduction
  - ✓ Covariance between two variables
  - ✓ Pearson's correlation measure
  - ✓ Non-parametric correlation measures
  - ✓ The model of simple linear regression
- Multiple linear regression model
  - ✓ The simple linear regression model
  - ✓ Model assumptions
  - ✓ Parameter interpretation
  - ✓ Implementation in R (Example 5-3)
  - ✓ Testing for the model assumptions
  - ✓ Diagnostic residual plots
  - ✓ Transforming variable
- Comparison to the paired t-test

Pearson's correlation coefficient

- It is the normalized version of covariance $\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$
- It measures the degree of linear dependence/relationship
- Bounded and defined in the interval from -1 to 1
  - ✓ 1 = perfect (non-random) positive linear relationship
  - ✓ -1 = perfect (non-random) negative linear relationship
  - ✓ 0 = two variables are not correlated

    for normal data => variable are independent
- Free of units
- Quantifies the degree of linear relation
- Does not separates the response from the explanatory

Pearson's correlation coefficient

- Population correlation $\rho = \dfrac{Cov(X,Y)}{\sigma_x \sigma_y}$

- Sample estimator

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{S_{xy}}{S_x S_y}$$

```
> cor(salary$salbeg, salary$salnow)
[1] 0.8801175
```

5. *Correlation and Regression models*
5.1. Introduction – correlation

Pearson's correlation in R

- If X & Y independent $\Rightarrow$ Correlation = 0
- Correlation = 0 $\Rightarrow$ no linear dependence
  but not necessarily independence
- Correlation = 0 & X - Y normal $\Rightarrow$ independence

ΟΠΑ AUEB

Pearson's correlation & independence

- If X & Y independent ⇒ Correlation = 0

```
> z1<-rnorm(1000)
> z2<-rnorm(1000)
> cor(z1,z2)
[1] 0.01802764
```

```
> z1<-rgamma(1000,1,1)
> z2<-rgamma(1000,1,1)
> cor(z1,z2)
[1] 0.008469119
```

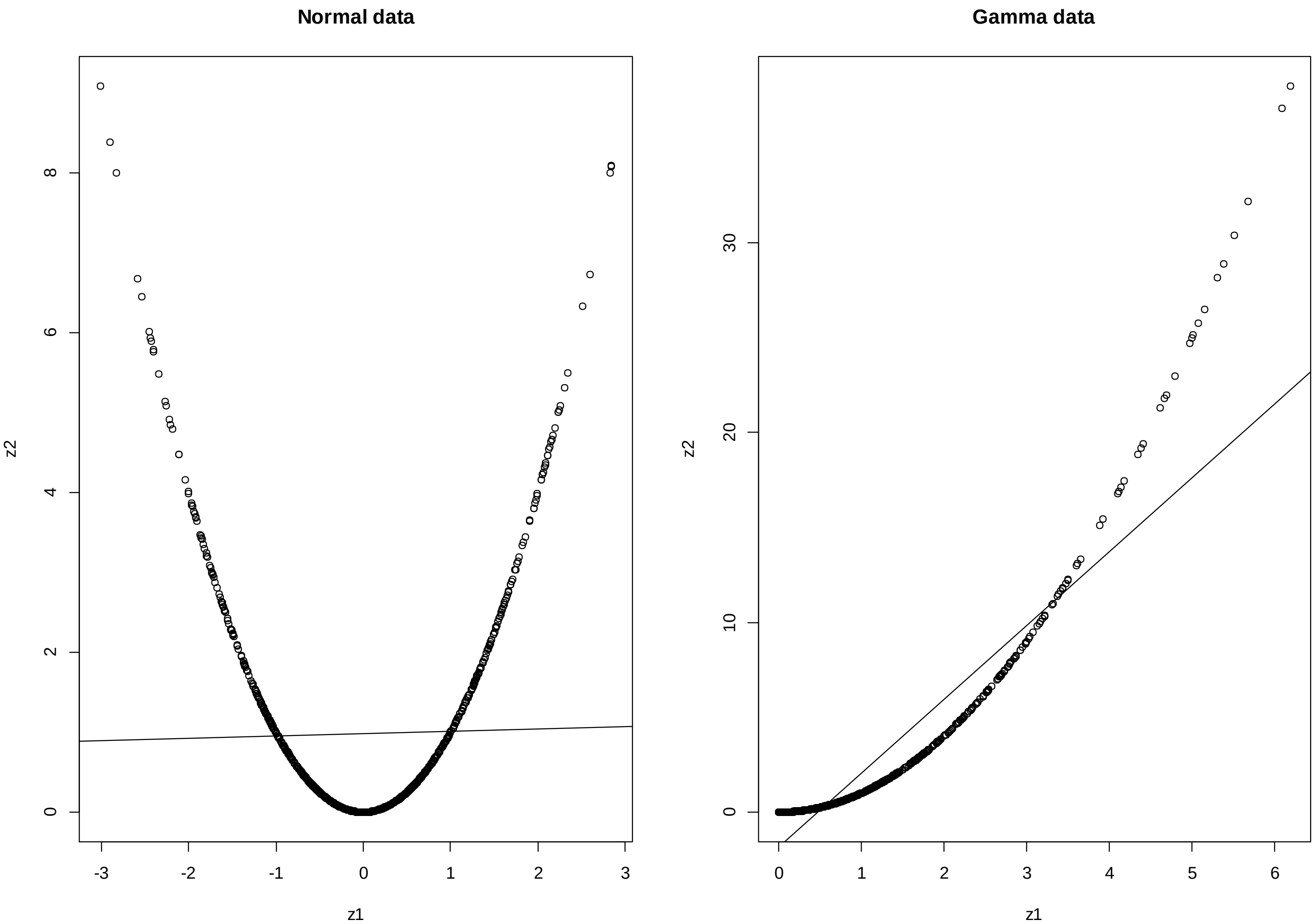
- Correlation = 0 ⇒ no linear dependence

  but not necessarily independence

```
> z1<-rnorm(1000)
> cor(z1,z1^2)
[1] 0.02178643
```

```
> z1<-rgamma(1000,1,1)
> cor(z1,z1^2)
[1] 0.9193777
```

5. Correlation and Regression models
5.1. Introduction – correlation



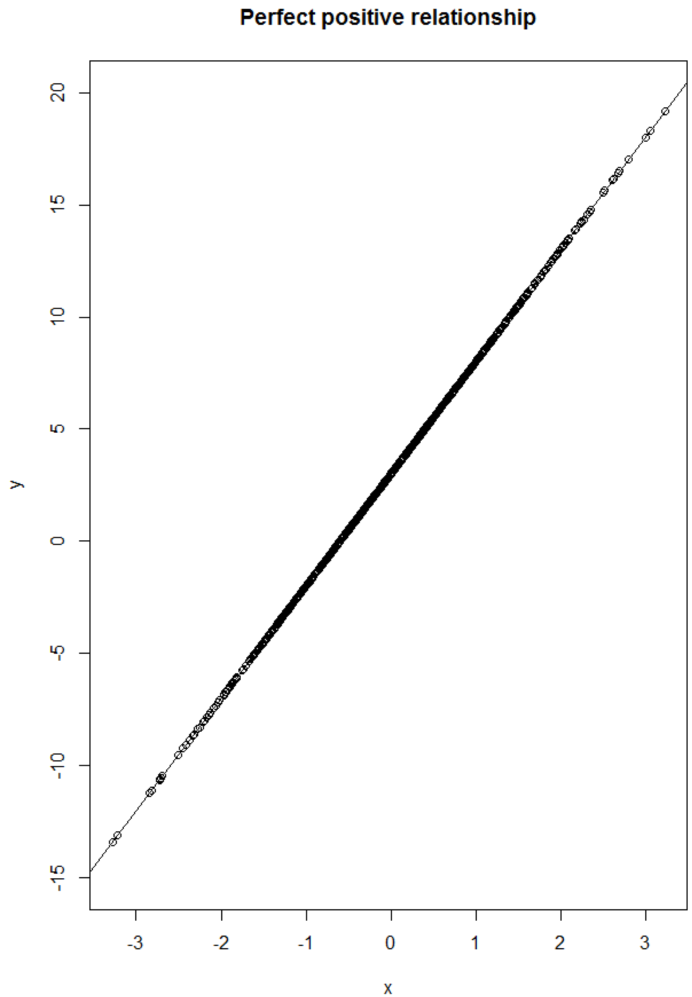
Pearson's correlation & linear functions

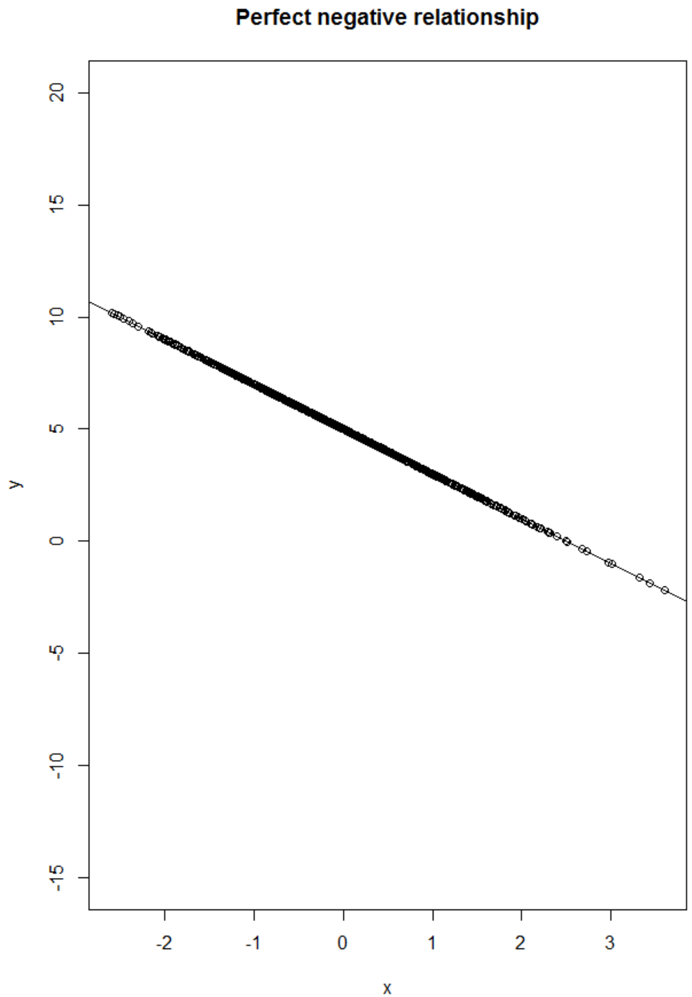
- If Y is a linear function of X $\Rightarrow$ Correlation = 1 or -1

```
> x<-rnorm(1000)
> y<- 5-2*x
> cor(x,y)
[1] -1
```

```
> x<-rnorm(1000)
> y<- 3+5*x
> cor(x,y)
[1] 1
```

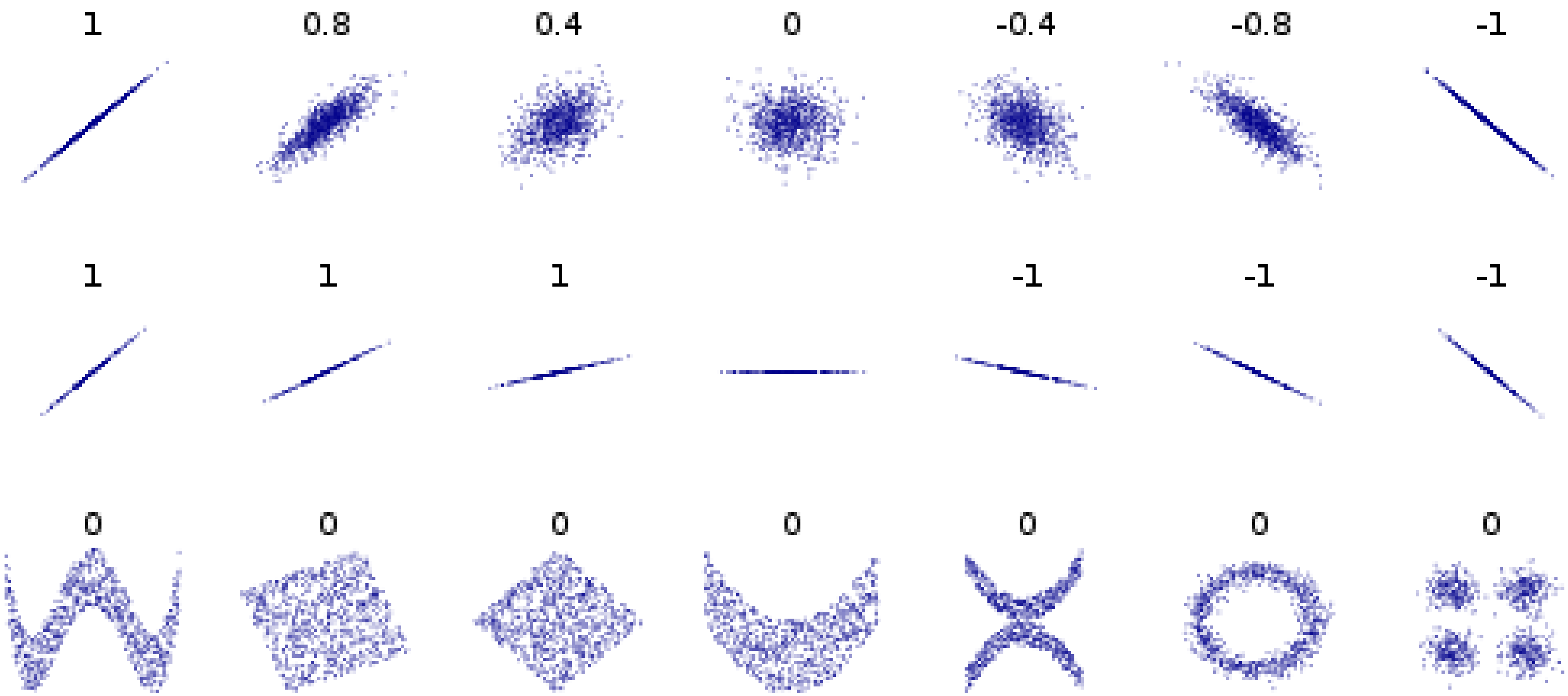
5. Correlation and Regression models
5.1. Introduction – correlation





5. *Correlation and Regression models*

5.1. Introduction – correlation



ΟΠΑ AUEB

Correlation matrix [using the observed data]

R is a pxp matrix with elements

- R_{jk} = Cor(X_j,X_k) – sample correlation between X_j and X_k
- R_{jj} = 1

  (the correlation of each variable with itself is one)

```
> cor(sal.num)
```

	id	salbeg	time	age	salnow	edlevel	work
id	1.00000000	-0.43118072	-0.012067260	0.10598470	-0.41863174	-0.33421423	0.018759273
salbeg	-0.43118072	1.00000000	-0.019753475	-0.01104036	0.88011747	0.63319565	0.045147858
time	-0.01206726	-0.01975347	1.000000000	0.05162975	0.08409227	0.04737878	0.002962074
age	0.10598470	-0.01104036	0.051629754	1.00000000	-0.14591032	-0.28084182	0.804397166
salnow	-0.41863174	0.88011747	0.084092267	-0.14591032	1.00000000	0.66055891	-0.097455333
edlevel	-0.33421423	0.63319565	0.047378777	-0.28084182	0.66055891	1.00000000	-0.252357836
work	0.01875927	0.04514786	0.002962074	0.80439717	-0.09745533	-0.25235784	1.000000000

The table is symmetric

Each element of the diagonal is 1 since each variable is fully correlated with itself (it is the identity function)

Example 5-1 [salary]

- Assess the possible linear relationships between starting and current salary

```
> x1<-salary$salbeg
> x2<-salary$salnow
> cor(x1,x2)
[1] 0.8801175
> cor.test(x1,x2)
```

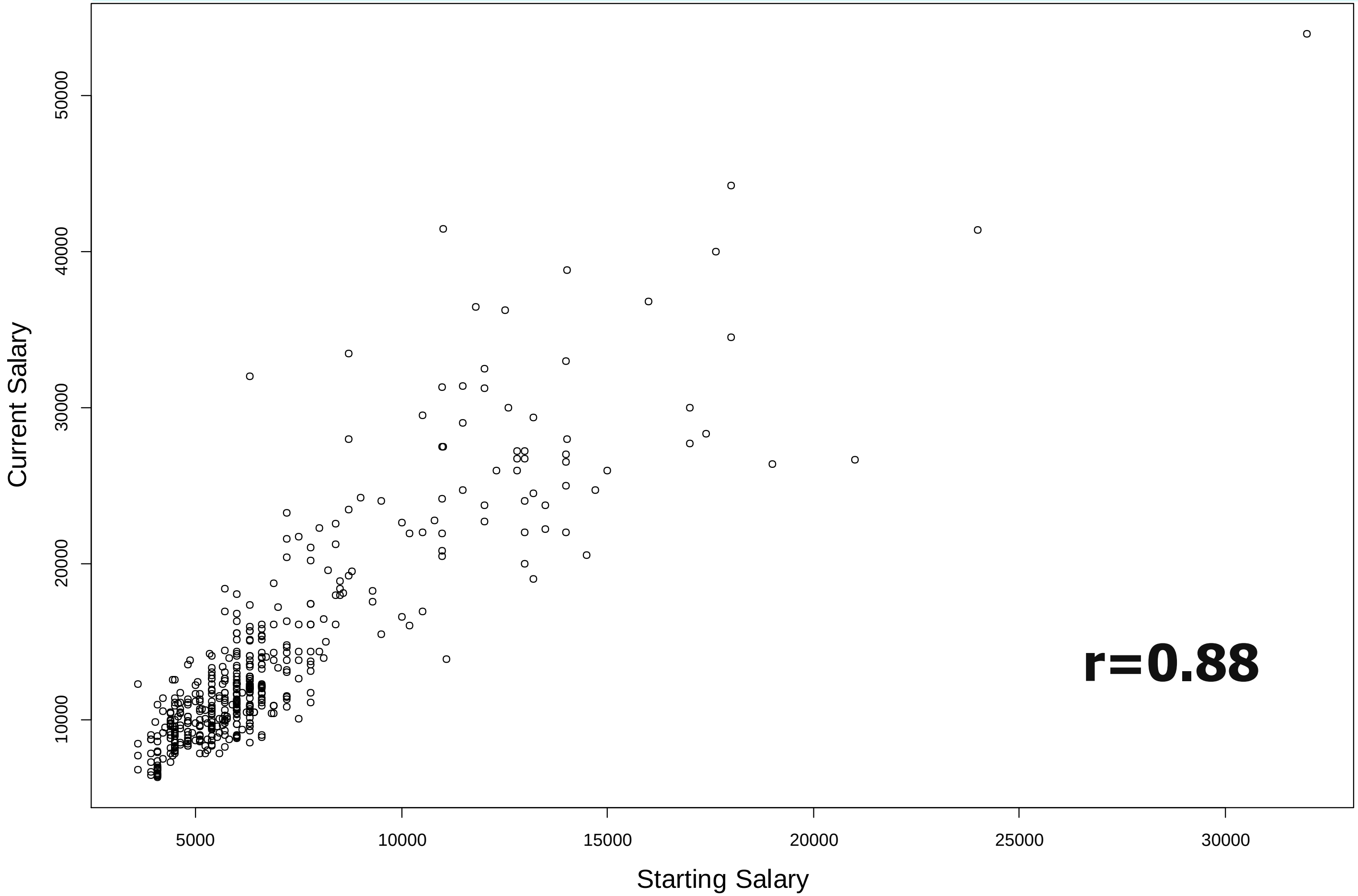
H_0: $\rho=0$

i.e. there is no linear relationship between the current and the starting salary

```
        Pearson's product-moment correlation

data:  x1 and x2
t = 40.2755, df = 472, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8580696 0.8989267
sample estimates:
      cor
0.8801175
```

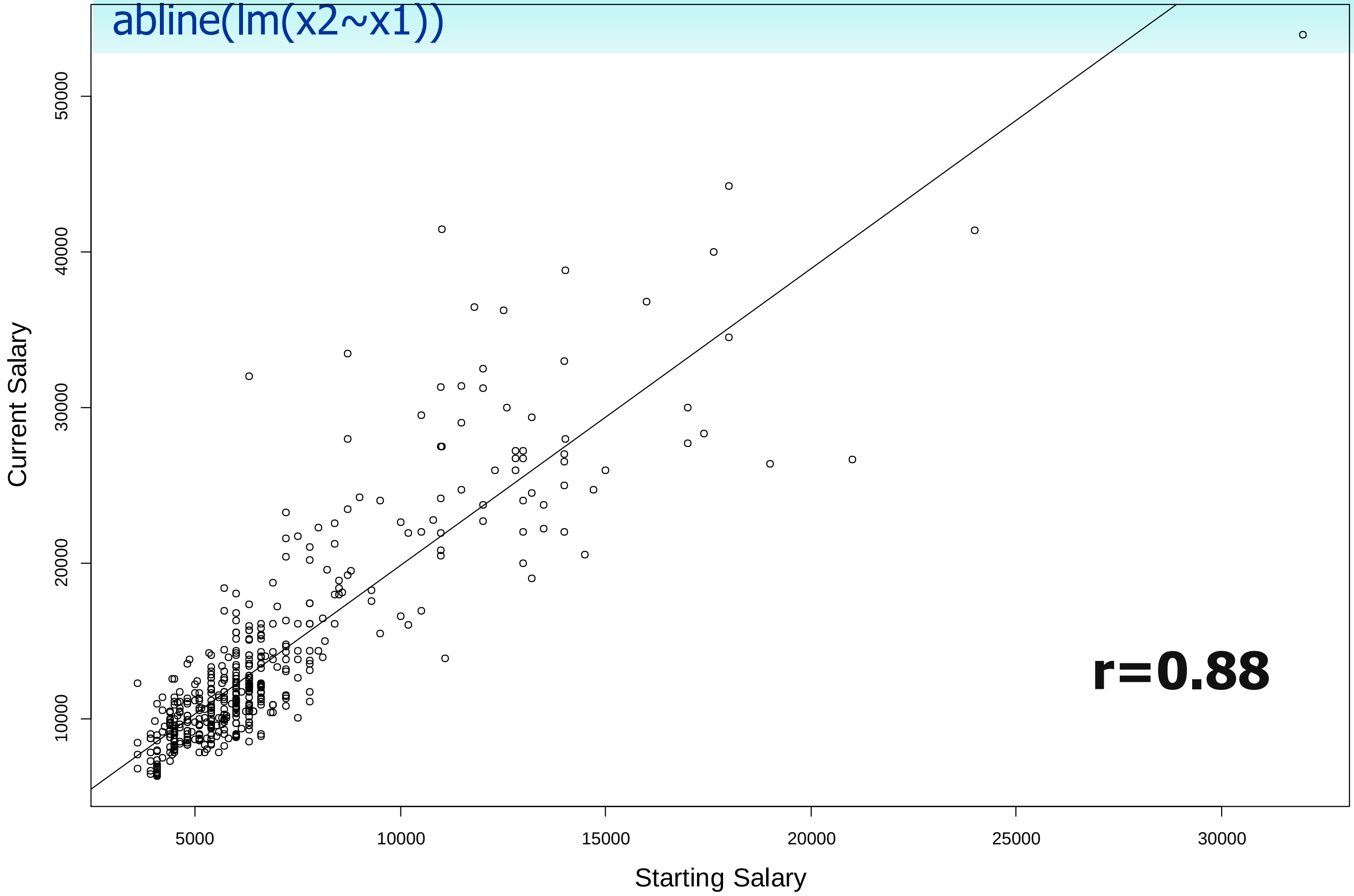
5. Correlation and Regression models
5.1. Introduction – correlation

```
plot(x1,x2, xlab='Starting Salary', ylab='Current Salary', cex.axis=1.5)
```



5. *Correlation and Regression models*
5.1. *Introduction – correlation*

```
plot(x1,x2, xlab='Starting Salary', ylab='Current Salary', cex.axis=1.5)
abline(lm(x2~x1))
```



Further comments (1)

- The coefficient assumes that both X and Y are random variables
- It can be used as a measure of linearity
- The hypothesis test assumes normality or large sample
- Alternatively, non-parametric correlation measures can be used
- If the relationship is strong but non-linear then the Pearson correlation coefficient will show how well this is approximated by a linear function

5. *Correlation and Regression models*
5.1. *Introduction – correlation*

Further comments (2)

According with Chatfield & Collins (1980, p. 40-41)

- The test is conservatory i.e. small values of r will give significant relationship (of some kind) especially for large samples
- Empirical rule:
  - strong linear dependence for $|r|>0.70$
  - Medium linear dependence for $0.4<|r|<0.70$
  - Weak linear dependence for $|r|< 0.4$
- The coefficient is not estimated reliably for small samples ($n<12$)

Example 5-1 [salary]

- Assess the possible linear relationships between age and the id?

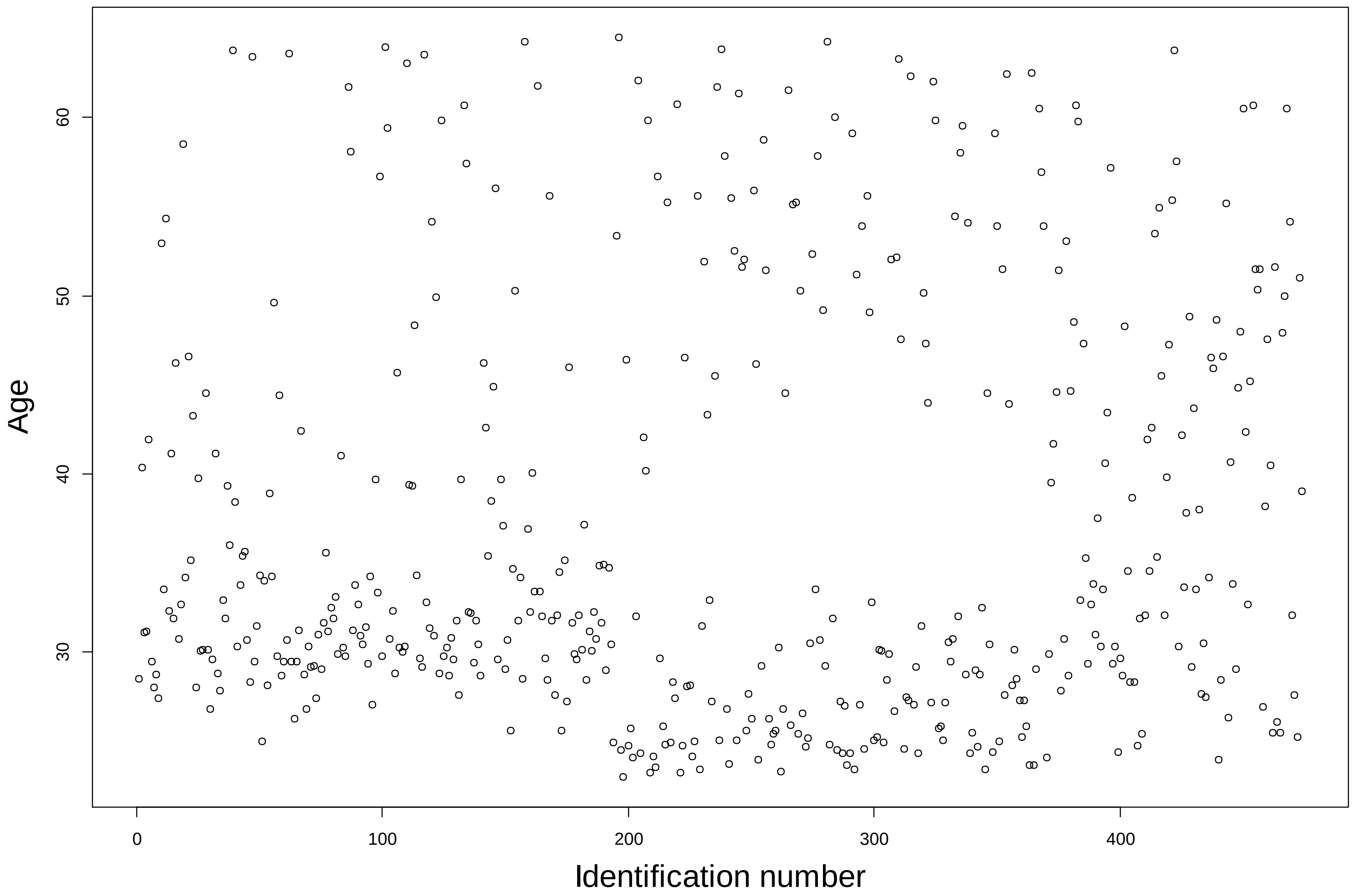
```
> x1<-salary$id
> x2<-salary$age
> cor(x1,x2)
[1] 0.1059847
> cor.test(x1,x2)

        Pearson's product-moment correlation

data:  x1 and x2
t = 2.3156, df = 472, p-value = 0.02101
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.01607248 0.19419663
sample estimates:
      cor
0.1059847
```

It seams that there is significant negative linear dependence between the the age and the id!!!

Does this makes sense?

Is the value of the coefficient large?

5. *Correlation and Regression models*
5.1. *Introduction – correlation*



Example 5-1 [salary]

- To assess the possible linear relationships between starting and current salary

```
library(sjPlot)
sjt.corr(x, corMethod = "pearson", showPValues = TRUE,
         pvaluesAsNumbers = FALSE,  fadeNS = TRUE,  digits = 3)
```

	salbeg	*salnow*
salbeg		0.880***
salnow	0.880***	

Computed correlation used pearson-method with pairwise-deletion.

Pearson's correlation between starting and current salary

* 0.01<p-value<0.05

** 0.001<p-value<0.01

*** p-value <0.001

ΟΠΑ AUEB

Example 5-1 [salary]

- To assess the possible linear relationships between starting and current salary

```
library(sjPlot)
sjt.corr(x, corMethod = "pearson", showPValues = TRUE,
        pvaluesAsNumbers = FALSE,  fadeNS = TRUE,  digits = 3)
```

	salbeg	*salnow*
salbeg		0.880 (0.000)
salnow	0.880 (0.000)	
Computed correlation used pearson-method with pairwise-deletion.		

Pearson's correlation between starting and current salary

P-value is given in brackets.

The current salary is highly correlated to the starting salary

5. *Correlation and Regression models*
5.1. *Introduction – correlation*

Example 5-1 [salary]

- To assess the possible linear relationships between starting and current salary

```
sjt.corr(sal.num,
         corMethod = "pearson",
         showPValues = TRUE,
         pvaluesAsNumbers = TRUE,
         fadeNS = TRUE,  digits = 3,
         triangle = "both")
```

	id	*salbeg*	*time*	*age*	*salnow*	*edlevel*	*work*
id		-0.431 (0.000)	-0.012 (0.793)	0.106 (0.021)	-0.419 (0.000)	-0.334 (0.000)	0.019 (0.684)
salbeg	-0.431 (0.000)		-0.020 (0.668)	-0.011 (0.811)	0.880 (0.000)	0.633 (0.000)	0.045 (0.327)
time	-0.012 (0.793)	-0.020 (0.668)		0.052 (0.262)	0.084 (0.067)	0.047 (0.303)	0.003 (0.949)
age	0.106 (0.021)	-0.011 (0.811)	0.052 (0.262)		-0.146 (0.001)	-0.281 (0.000)	0.804 (0.000)
salnow	-0.419 (0.000)	0.880 (0.000)	0.084 (0.067)	-0.146 (0.001)		0.661 (0.000)	-0.097 (0.034)
edlevel	-0.334 (0.000)	0.633 (0.000)	0.047 (0.303)	-0.281 (0.000)	0.661 (0.000)		-0.252 (0.000)
work	0.019 (0.684)	0.045 (0.327)	0.003 (0.949)	0.804 (0.000)	-0.097 (0.034)	-0.252 (0.000)	

Computed correlation used pearson-method with pairwise-deletion.

Non significant correlations are faded with grey color

Example 5-1 [salary]

- To assess the possible linear relationships between starting and current salary

```
sjt.corr(sal.num,
         corMethod = "pearson",
         showPValues = TRUE,
         pvaluesAsNumbers = TRUE,
         fadeNS = TRUE,  digits = 3,
         triangle = "lower")
```

	id	*salbeg*	*time*	*age*	*salnow*	*edlevel*	*work*
id							
salbeg	-0.431 *(0.000)*						
time	-0.012 *(0.793)*	-0.020 *(0.668)*					
age	0.106 *(0.021)*	-0.011 *(0.811)*	0.052 *(0.262)*				
salnow	-0.419 *(0.000)*	0.880 *(0.000)*	0.084 *(0.067)*	-0.146 *(0.001)*			
edlevel	-0.334 *(0.000)*	0.633 *(0.000)*	0.047 *(0.303)*	-0.281 *(0.000)*	0.661 *(0.000)*		
work	0.019 *(0.684)*	0.045 *(0.327)*	0.003 *(0.949)*	0.804 *(0.000)*	-0.097 *(0.034)*	-0.252 *(0.000)*	

Computed correlation used pearson-method with pairwise-deletion.

Non significant correlations are faded with grey color

Back to correlation matrices

```
> cor(sal.num)
                 id      salbeg         time         age      salnow     edlevel         work
id       1.00000000 -0.43118072 -0.012067260  0.10598470 -0.41863174 -0.33421423  0.018759273
salbeg  -0.43118072  1.00000000 -0.019753475 -0.01104036  0.88011747  0.63319565  0.045147858
time    -0.01206726 -0.01975347  1.000000000  0.05162975  0.08409227  0.04737878  0.002962074
age      0.10598470 -0.01104036  0.051629754  1.00000000 -0.14591032 -0.28084182  0.804397166
salnow  -0.41863174  0.88011747  0.084092267 -0.14591032  1.00000000  0.66055891 -0.097455333
edlevel -0.33421423  0.63319565  0.047378777 -0.28084182  0.66055891  1.00000000 -0.252357836
work     0.01875927  0.04514786  0.002962074  0.80439717 -0.09745533 -0.25235784  1.000000000
```

5. *Correlation and Regression models*
5.1. Introduction – correlation

How to tide up and make correlation matrices readable

- Keep only correlation measures (no p-values)
- Keep only one or two decimals
- Eliminate irrelevant variables (e.g. id)
- Group correlated variables
- Uses symbols or colors for high or significant correlations
- If even these changes, it does not makes any sense
  - Eliminate numbers and keep only colors or symbols
  - Use path diagrams

Correlation matrices

- Eliminate decimal numbers & other values

```
> round(cor(sal.num),1)
           id salbeg time  age salnow edlevel work
id        1.0   -0.4  0.0  0.1   -0.4    -0.3  0.0
salbeg   -0.4    1.0  0.0  0.0    0.9     0.6  0.0
time      0.0    0.0  1.0  0.1    0.1     0.0  0.0
age       0.1    0.0  0.1  1.0   -0.1    -0.3  0.8
salnow   -0.4    0.9  0.1 -0.1    1.0     0.7 -0.1
edlevel  -0.3    0.6  0.0 -0.3    0.7     1.0 -0.3
work      0.0    0.0  0.0  0.8   -0.1    -0.3  1.0
```

Correlation matrices

- Eliminate irrelevant values

```
> round(cor(sal.num),1)[-1,-1]
        salbeg time  age salnow edlevel work
salbeg     1.0  0.0  0.0    0.9     0.6  0.0
time       0.0  1.0  0.1    0.1     0.0  0.0
age        0.0  0.1  1.0   -0.1    -0.3  0.8
salnow     0.9  0.1 -0.1    1.0     0.7 -0.1
edlevel    0.6  0.0 -0.3    0.7     1.0 -0.3
work       0.0  0.0  0.8   -0.1    -0.3  1.0
```

Correlation matrices

- Add colors

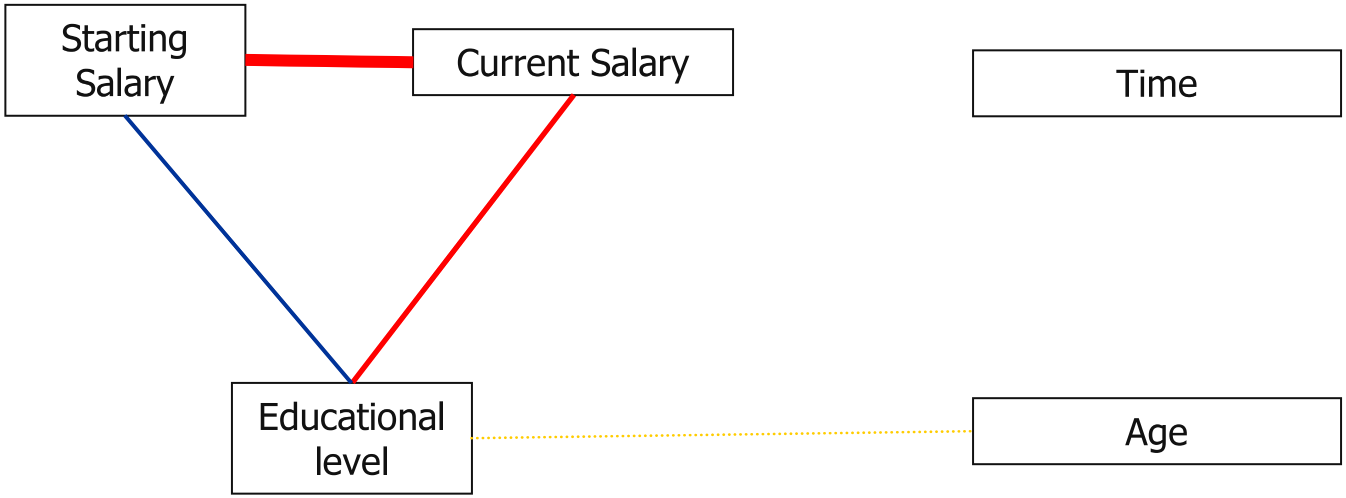
```
> temp<-round(cor(sal.num),1)[-1,-1]
> index<-c(1,4,5,3,2)
> temp[index,index]
         salbeg salnow edlevel  age time
salbeg      1.0    0.9     0.6  0.0  0.0
salnow      0.9    1.0     0.7 -0.1  0.1
edlevel     0.6    0.7     1.0 -0.3  0.0
age         0.0   -0.1    -0.3  1.0  0.1
time        0.0    0.1     0.0  0.1  1.0
```

Correlation matrices

- Re-arrange the matrix according to the correlations

```
> temp<-round(cor(sal.num),1)[-1,-1]
> index<-c(1,4,5,3,2)
> temp[index,index]
         salbeg salnow edlevel  age time
salbeg      1.0    0.9     0.6  0.0  0.0
salnow      0.9    1.0     0.7 -0.1  0.1
edlevel     0.6    0.7     1.0 -0.3  0.0
age         0.0   -0.1    -0.3  1.0  0.1
time        0.0    0.1     0.0  0.1  1.0
```

Path diagram

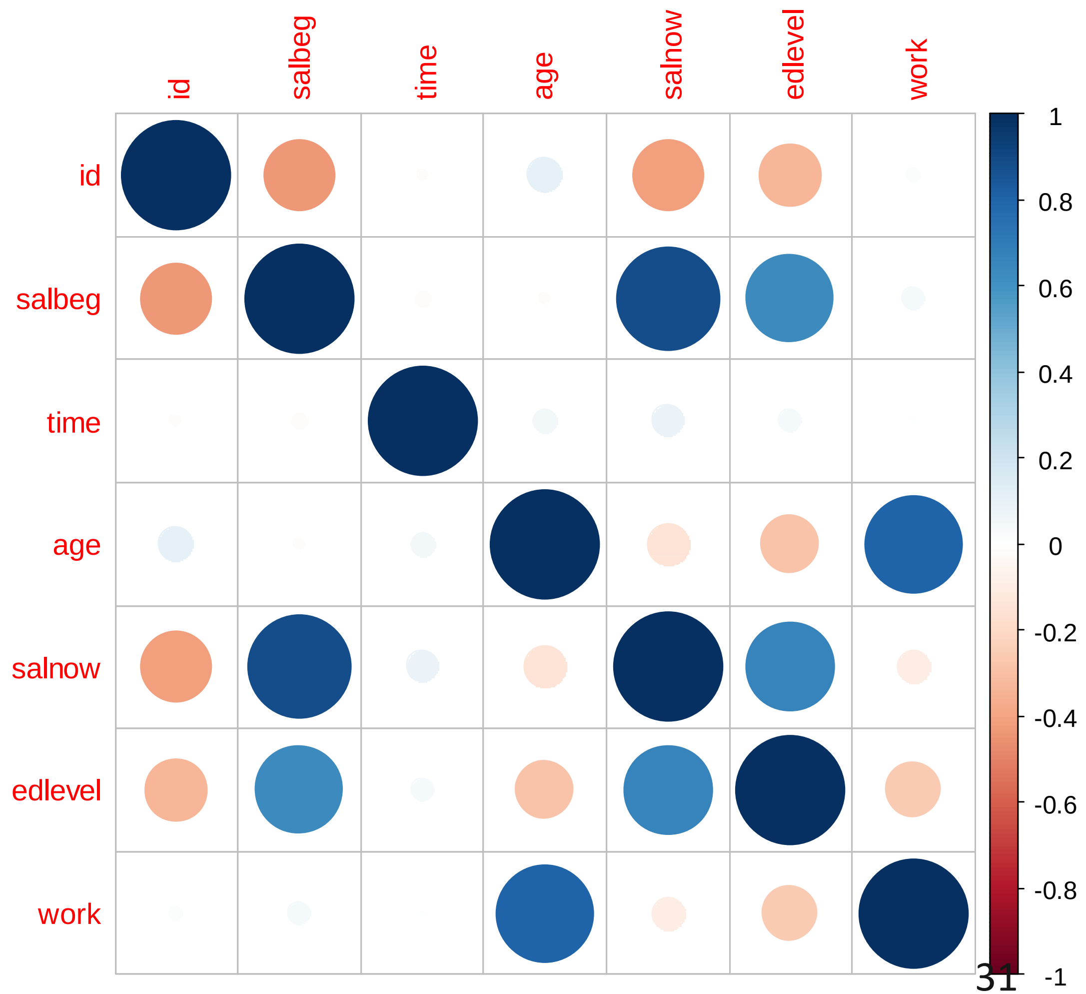


Fancy plots using sjPlot

```
x<-sal.num
libray(sjPlot); sjp.corr(x, corMethod = "pearson")
```

Fancy plots using corrplot

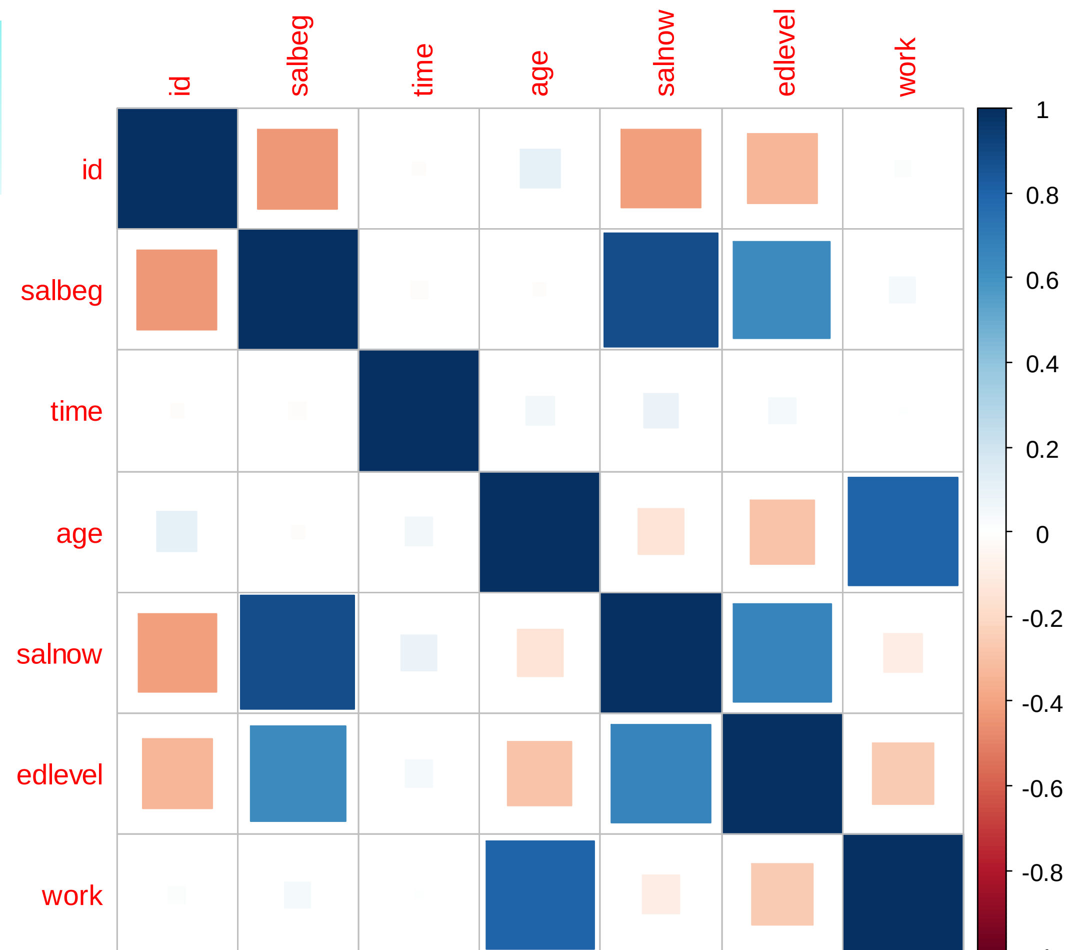
```
library(corrplot)
corrplot(cor(sal.num))
```



5. *Correlation and Regression models*
5.1. *Introduction – correlation*

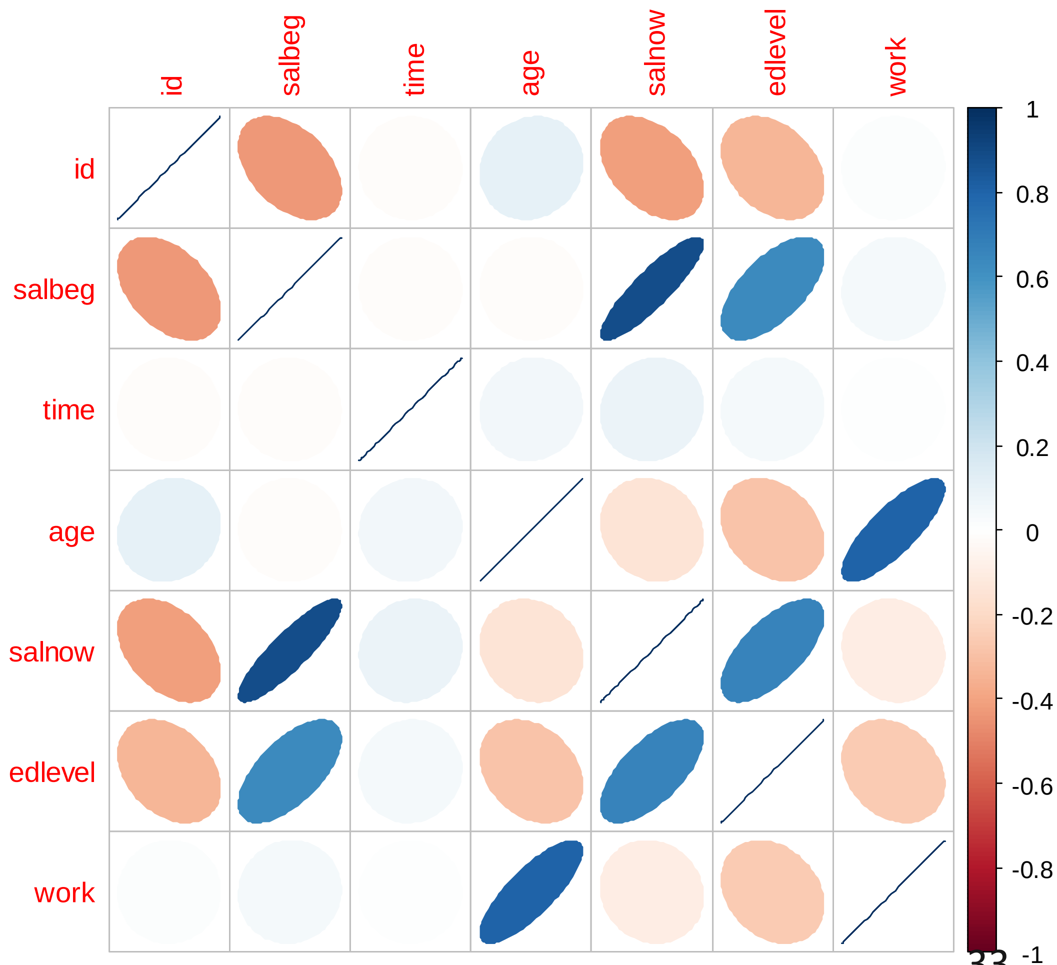
Fancy plots using corrplot

```
library(corrplot)
corrplot(cor(sal.num),
              method= "square")
```



Fancy plots using corrplot

```
library(corrplot)
corrplot(cor(sal.num),
                method= " ellipse ")
```

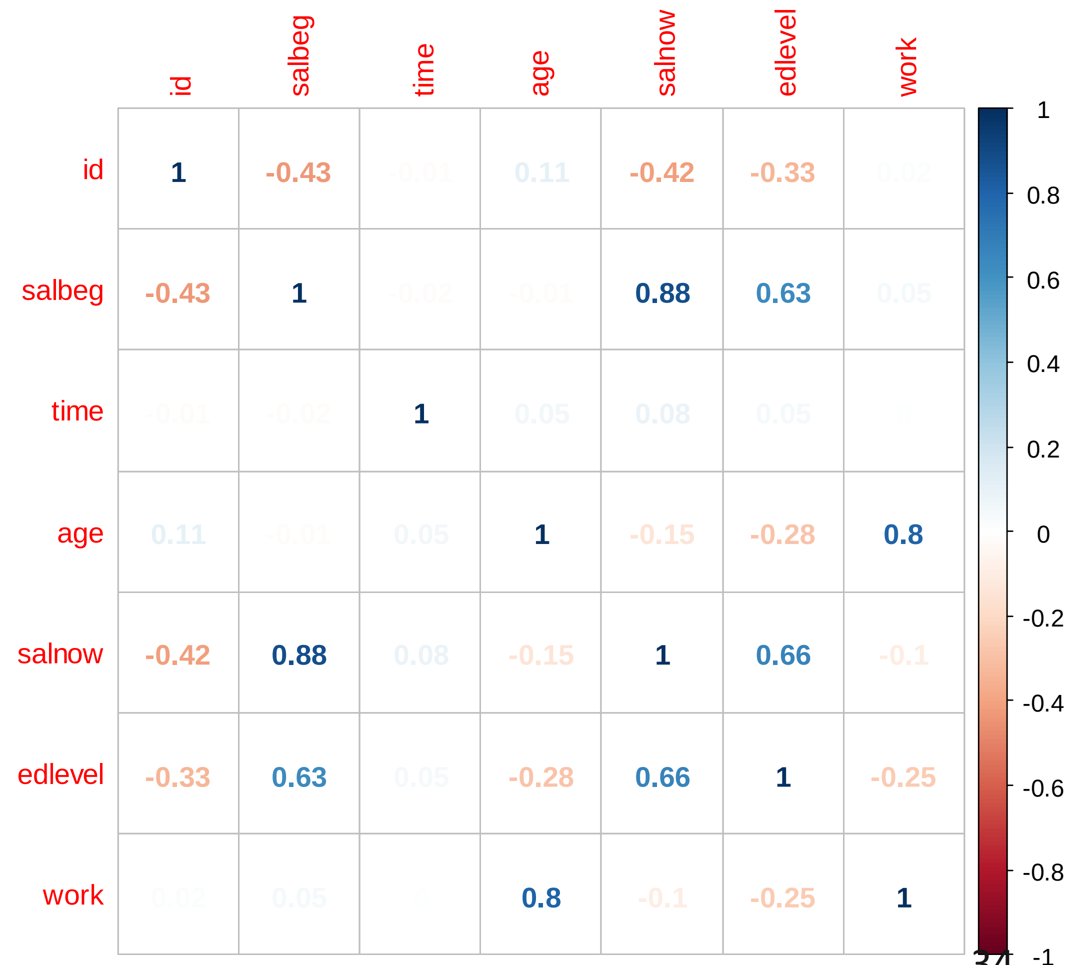


5. *Correlation and Regression models*

5.1. *Introduction – correlation*

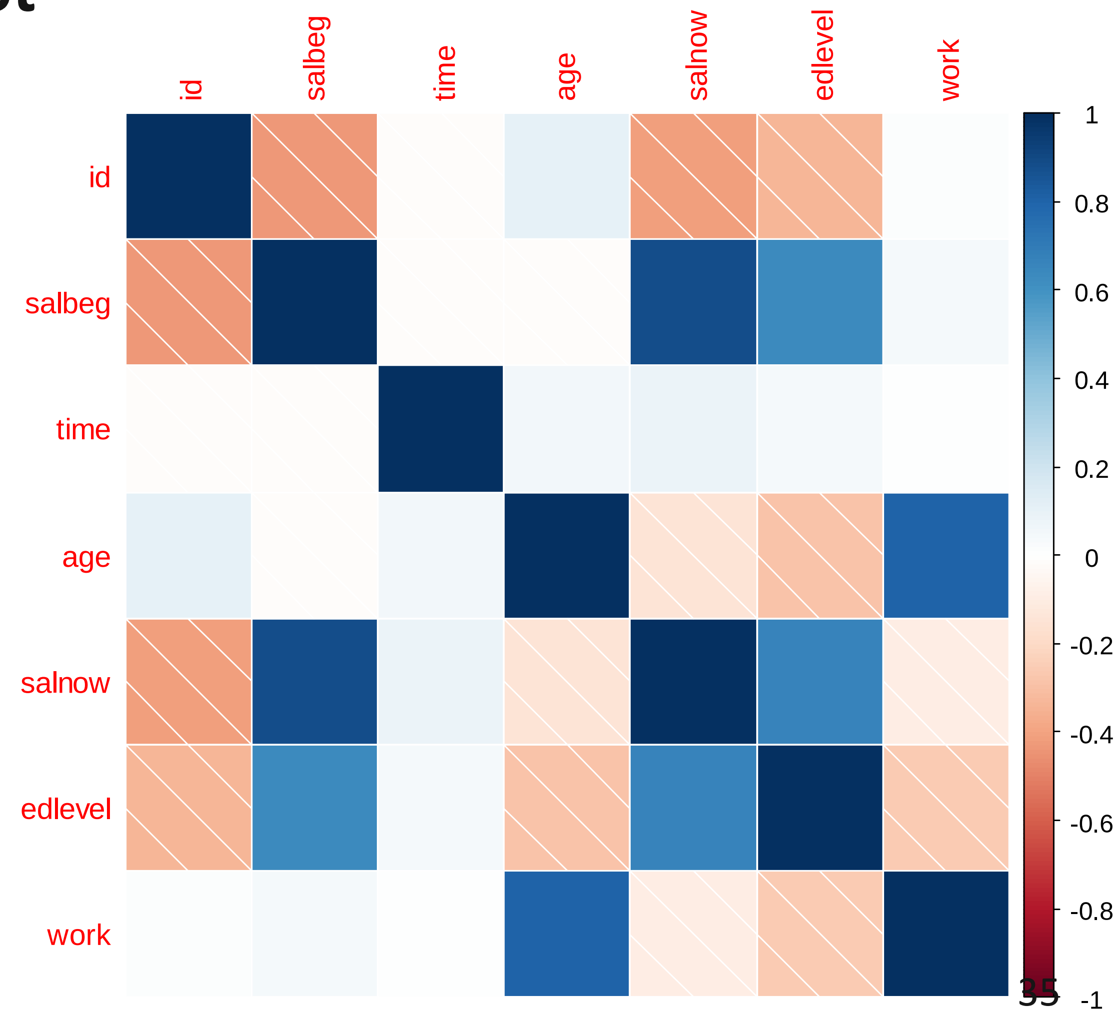
Fancy plots using corrplot

```
library(corrplot)
corrplot(cor(sal.num),
            method= " number")
```



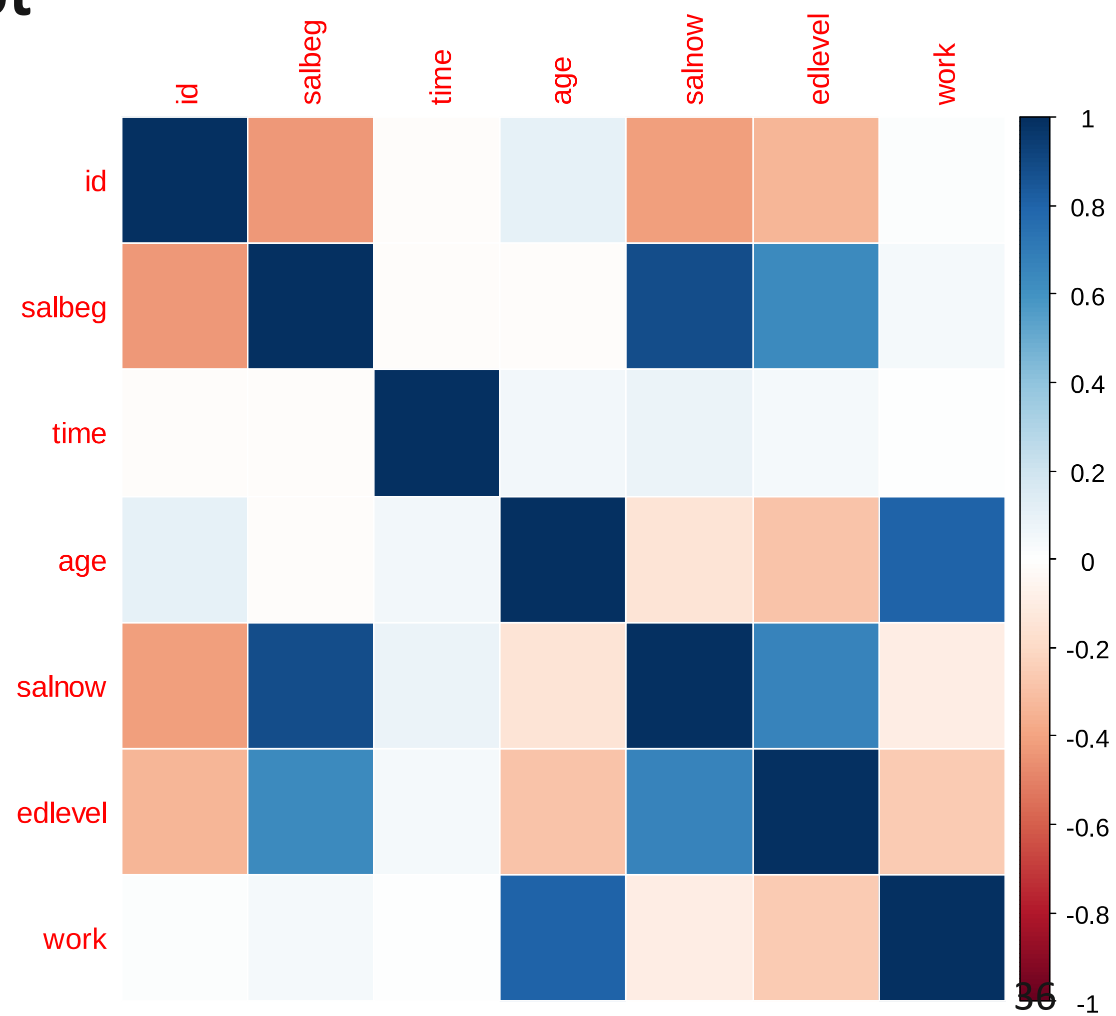
Fancy plots using corrplot

```
library(corrplot)
corrplot(cor(sal.num),
         method= " shade")
```



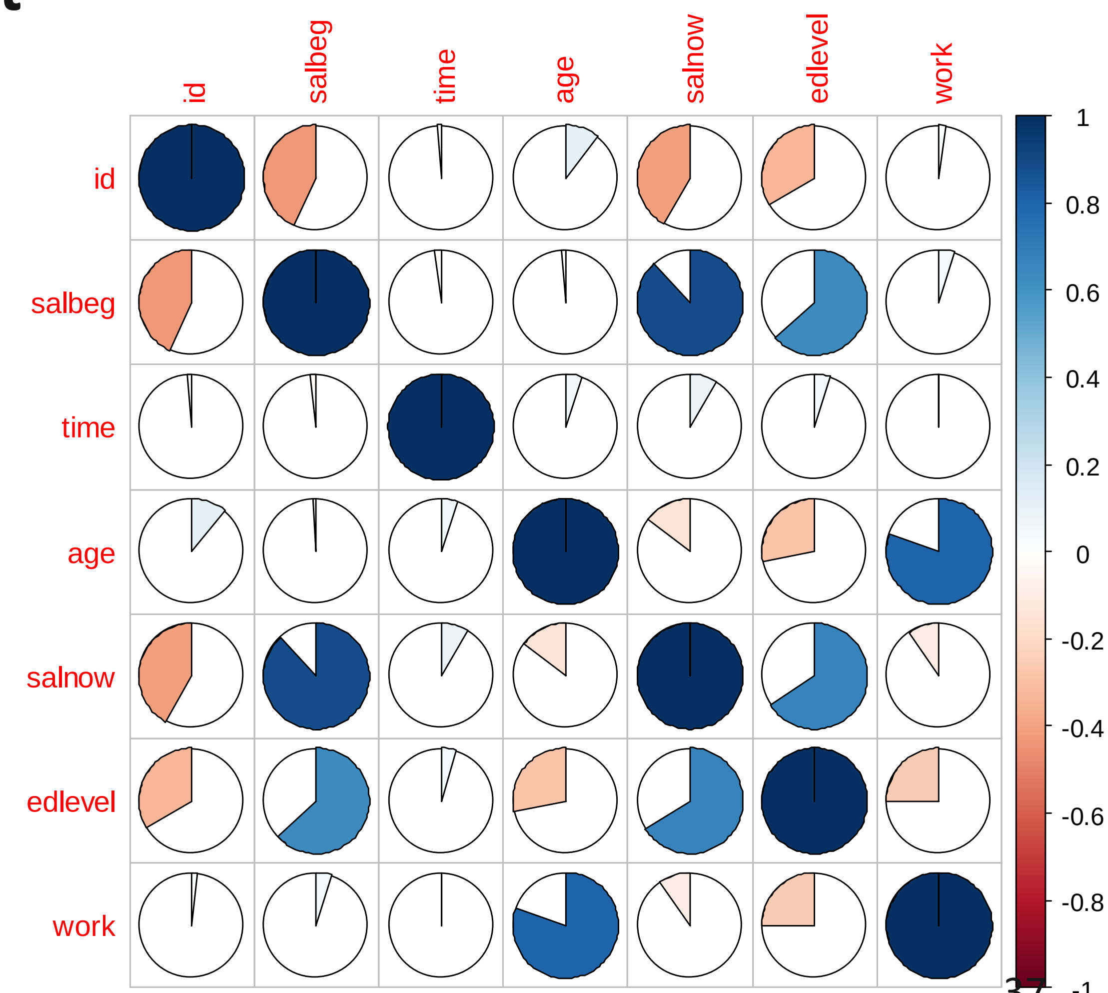
Fancy plots using corrplot

```
library(corrplot)
corrplot(cor(sal.num),
              method= " color")
```



Fancy plots using corrplot

```
library(corrplot)
corrplot(cor(sal.num),
                        method= "pie")
```



Example 5-2 [world95]

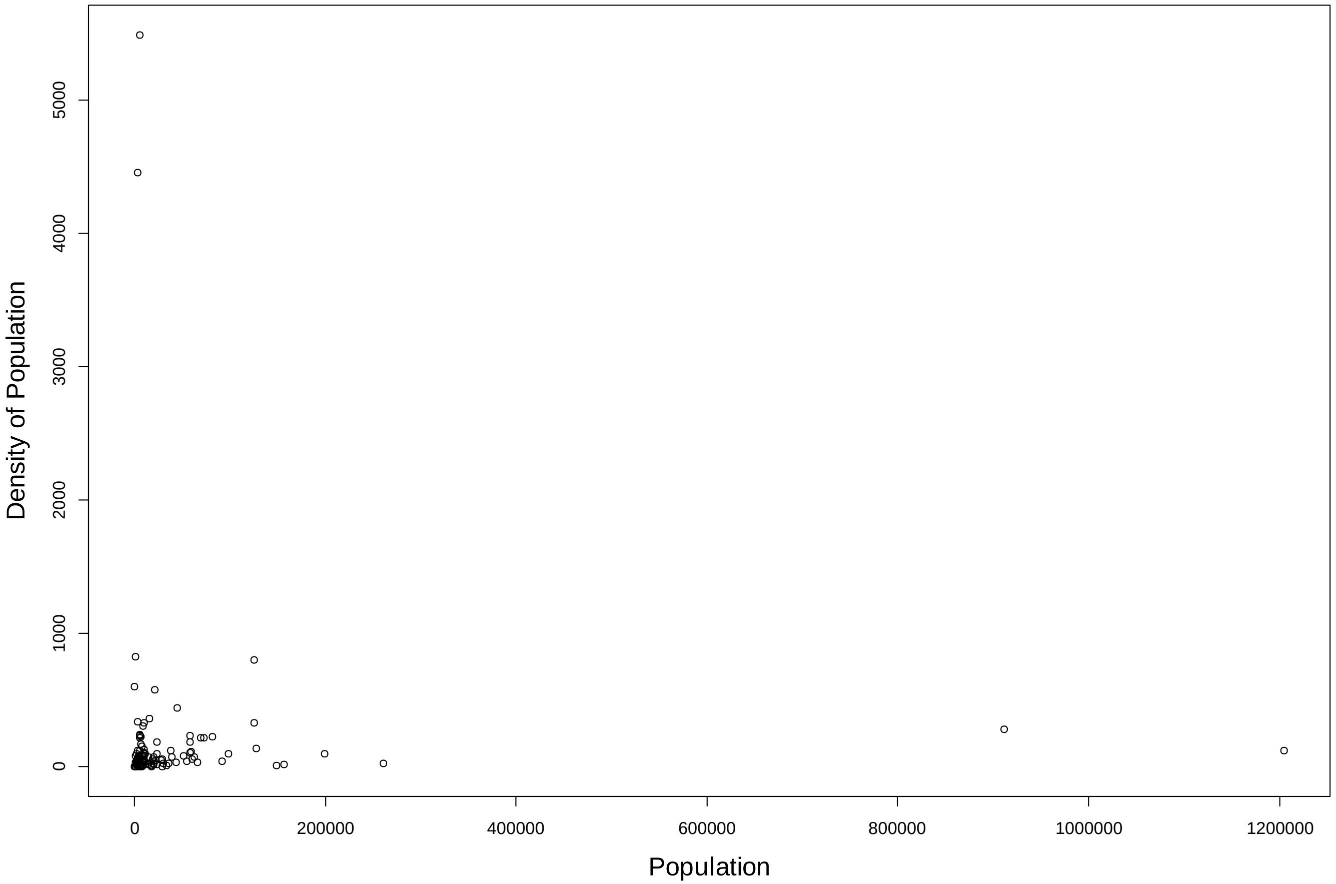
We would like to assess the correlation between the population and the density

```
> cor.test(world95$popul,world95$density)

        Pearson's product-moment correlation

data:  world95$popul and world95$density
t = -0.1894, df = 107, p-value = 0.8501
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2057032  0.1703786
sample estimates:
        cor
-0.01830997
```

Non-significant linear relationship between the population and the density.

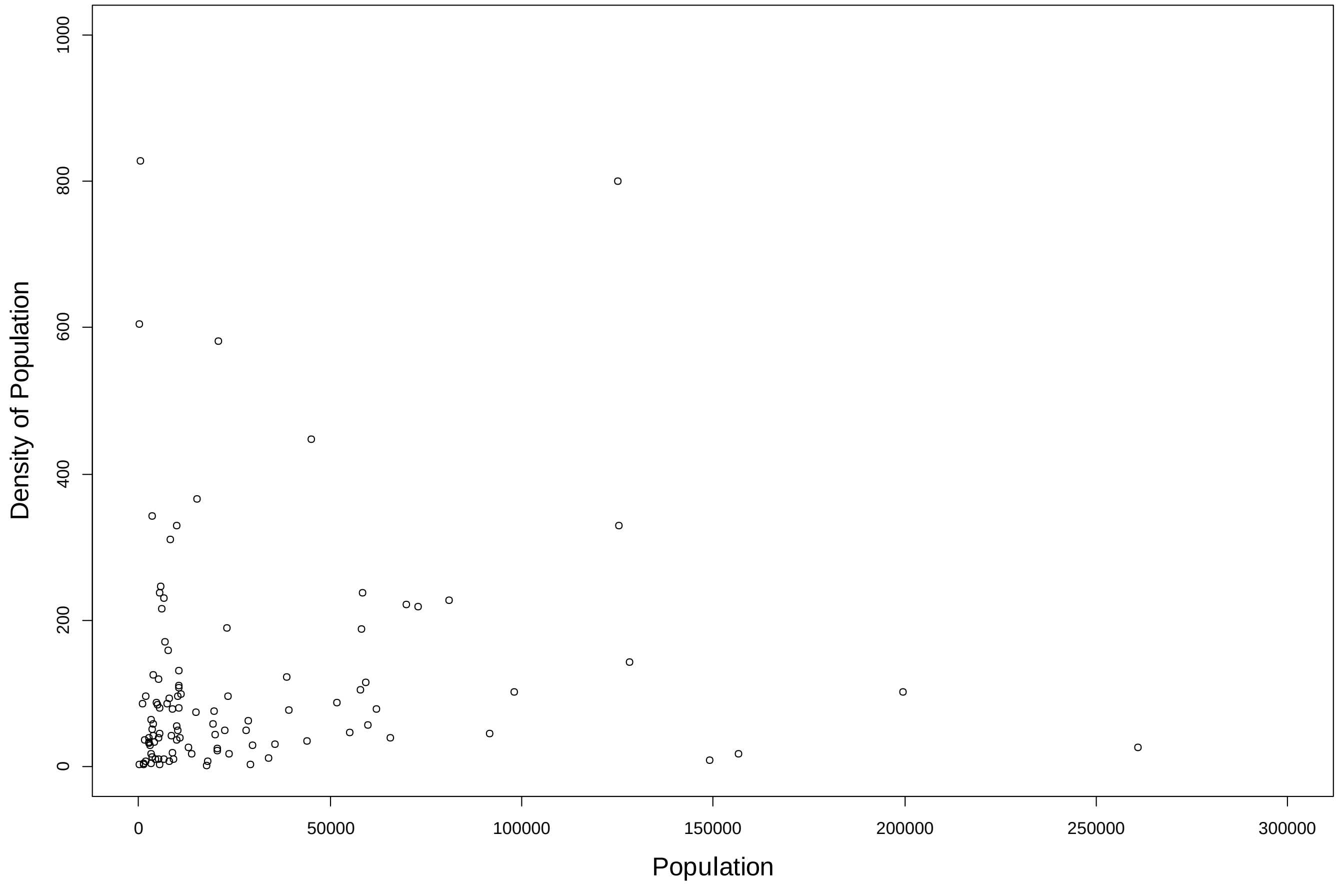
Also the coefficient is very small indicating minor or no linear relationship

5. Correlation and Regression models
5.1. Introduction – correlation



5. *Correlation and Regression models*

5.1. *Introduction – correlation*



Example 5-2 [world95]

We would like to assess the correlation between the population and the density

But by definition

DENSITY = POPULATION/AREA (in sq meters)

= a + b * POPULATION

with a=0 and b=1/AREA !!!!

So why r≈0 instead of r=1????

5. *Correlation and Regression models*
5.1. *Introduction – the simple linear model*

Let us assume that we have two quantitative variables

- X: explanatory or independent variable
- Y: response or dependent variable

If we believe that X influences (or affects) in a some way the response Y then it is sensible to assume that a function h(x) exists such that:

$$y = h(x)$$

[perfect/deterministic relationship]

Since we mainly study random phenomena/experiments then it is sensible to add a random (unpredicted) component (i.e. error term)

$$\begin{aligned} y &= h(x) + \varepsilon \\ \varepsilon &\sim \text{Distribution}(\theta) \end{aligned}$$

5. *Correlation and Regression models*
5.1. Introduction – the simple linear model

Two quantitative variables

- X: explanatory or independent variable
- Y: response or dependent variable

Regression model assumes

- linear relationship (function) between X and Y

$$h(x) \; = \; \beta_0 + \beta_1 x$$

- Normal errors

$$\varepsilon \sim N(0, \sigma^2)$$

So the regression model is now given by

$$\begin{aligned} y &= \beta_0 + \beta_1 x + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2) \end{aligned}$$

ΟΠΑ AUEB

Two quantitative variables

- X: explanatory or independent variable
- Y: response or dependent variable

Regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

- WHY LINEAR?
- WHY NORMAL?
- WHY ZERO MEAN OF ERRORS?
- WHAT σ^2 means?

5. *Correlation and Regression models*
5.1. *Introduction – the simple linear model*

More general approach [GLM]

(and more appropriate in terms of modeling)

- X: explanatory or independent variable
- Y: response or dependent variable

$$
\begin{aligned}
Y &\sim \text{Distribution}(\theta) \\
g(\theta) &= h(x)
\end{aligned}
$$

- ✓ Distribution(**θ**): stochastic (random) component
- ✓ h(x): deterministic (non random) component
- ✓ g(θ): link function between stochastic and deterministic component
- ✓ Usually h(x) ⇔ linear function of X ⇔ also called linear predictor

More general approach [GLM]

(and more appropriate in terms of modeling)

- X: explanatory or independent variable
- Y: response or dependent variable

- $Y \sim \text{Normal}(\mu, \sigma^2)$ $[\boldsymbol{\theta}^T = (\mu, \sigma^2)]$
- $\mu = \beta_0 + \beta_1 x$ $[g(\theta) = \mu]$

Two ways to write a regression model:

- Using the error term representation

$$y = \beta_0 + \beta_1 x + \varepsilon$$
$$\varepsilon \sim N(0, \sigma^2)$$

or equivalently

- Using the stochastic response (GLM type) representation

$$Y \sim N(\mu, \sigma^2)$$
$$\mu = \beta_0 + \beta_1 x$$

5. *Correlation and Regression models*
5.2. *The simple linear regression model*

The two ways to write a regression model when data are introduced.

We need to introduce an indicator for the study unit/observation :

Representing by Y_i, X_i (for i=1,2, … , n) the pairs of the response & explanatory values for each study unit <*i*>

- Using the error term representation

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i &\sim N(0, \sigma^2) \end{aligned}$$

or equivalently

- Using the stochastic response (GLM type) representation

$$\begin{aligned} Y_i &\sim N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 x_i \end{aligned}$$

ΟΠΑ AUEB

Terminology and estimators

- $\hat{\beta}_0, \hat{\beta}_1$: Sample estimators/estimates of β_0 and β_1
- $\hat{y}_i$: Expected value according to the model or fitted value for <*i*> study unit/ observation/subject
- e_i : Regression residual (estimate of ε_i)
- $\hat{\sigma}^2$: Estimator/estimate of the error variance

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

ΟΠΑ AUEB

Terminology and estimators

- R^2 : Coefficient of determination
  - ✓ This is a goodness of fit measure
  - ✓ Takes values from 0 to 1
  - ✓ **Interpretation**: % of variability explained by the model
  - ✓ In simple regression $R^2=r^2$
- R_{adj}^2 : Adjusted coefficient of determination
  - ✓ Takes values from 0 to 1
  - ✓ **Interpretation**: % of variance explained by the model
  - ✓ More useful in multiple regression

$$R^2 = 1 - \frac{(n-2)\hat{\sigma}^2}{(n-1)s_Y^2}$$

$$R_{adj}^2 = 1 - \frac{\hat{\sigma}^2}{s_Y^2}$$

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

Terminology and estimators

Sample estimators of model coefficients β_0 & β_1

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

$$= \frac{\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}} \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2}} = \frac{s_y}{s_x} r$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

5.2.1. Model assumptions (summary)

ASSUMPTIONS (to be checked):

- Independence of errors (and of Y_i)
- Normality of errors (and of Y_i)
- Homoscedasticity of errors (and Y_i)
- Linearity between X & Y

- We work with the residuals e_i

We will discuss in more detail about regression diagnostics and residual analysis later on in this presentation

ΟΠΑ AUEB

We use a regression model to

- Describe and understand the association between the two variables
- To predict future values of Y
- Both

When we are interested in the relationship between X & Y:

- **Primary test**: H_0: $\beta_1=0$ vs. H_1: $\beta_1 \neq 0$
- **Test of secondary importance**: H_0: $\beta_0=0$ vs. H_1: $\beta_0 \neq 0$

In case that we are interested in prediction:

- we need to know if we can use the fitted model for prediction

Testing for the relationship between X & Y

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

- ✓ Equivalent to testing for the correlation between X & Y
- ✓ It provides the slope of the fitted line
- ✓ We are interested in the interpretation of CAUSAL relationships between variables (i.e. characteristics or phenomena).

Interpretation: It tests how much we expect that Y will increase if X increases by one unit

- ✓ The value of β_1 is affected by the scale and the units of measurement of both X & Y.
- ✓ The correlation measures (ρ & r) and the corresponding tests (for ρ or β1) are not affected by linear changes.

Testing for the relationship between X & Y

Secondary hypothesis test: H_0: $\beta_0=0$ vs. H_1: $\beta_0 \neq 0$

- ✓ Intercept of the fitted line
- ✓ It provides the point where the fitted line intersects with the vertical axis YY' i.e. the value of Y when X=0

Interpretation: Is the expected value of Y when X=0.

- ✓ Many times this value does not have direct interpretation (since this value is not possible or outside the observed range
- ✓ Sometimes we constraint $\beta_0=0$ due to logic or an assumed theory
- ✓ Other times it is convenient to consider instead of X, the centered version $X'=X - \bar{X}$. Then
  - ✓ β_1 remains the same
  - ✓ β_0 gives the expected value of Y when X is equal to the sample mean

5. *Correlation and Regression models*
5.2.2. Parameter interpretation

Deciding whether we can use the fitted model for prediction

- We can predict the expected value of Y for each X
- The error variance σ^2 & R^2 quantify the precision of the prediction
  - ✓ $R^2>0.7$ ⇔ good predictions
  - ✓ $R^2>0.9$ ⇔ very good predictions

Predicting outside the observed values

[Extrapolation – a trip to the unknown?]

BECAREFUL: predictions are reliable and acceptable only for values of X that we have observed (and hence we have some information about it)

- ✓ We cannot predict something that we have not any information about it and therefore we have not studied it
- ✓ Sometimes we are forced to make predictions outside the observed range of X (extrapolation)
  - ➢ This predictions should be used only as a rough yardstick
  - ➢ We assume the same (linear) relationship is valid also for these unobserved values of X

Example 5-3 [data frame cargo]

- The head of the logistics department of a large company is interested to estimate the delivery time and therefore the corresponding cost of each cargo depending on the distance
- For this reason, we randomly selected 10 cargo deliveries and recorded the distance in miles and the days until the delivery
- Construct a model that can assist the manager in his aim

Cargo delivery	1	2	3	4	5	6	7	8	9	10
Distance in Miles	825	215	1070	550	480	920	1350	325	670	1215
Delivery time in days	3.5	1.0	4.0	2.0	1.0	3.0	4.5	1.5	3.0	5.0

Example 5-3

- Study Unit: cargo
- Sample size: n=10 cargos
- Characteristics: p=3
  - ✓ Cargo id
  - ✓ Distance
  - ✓ Delivery time
- Which is X & which is Y?

	id	distance	delivery
1	1	825	3.5
2	2	215	1
3	3	1070	4
4	4	550	2
5	5	480	1
6	6	920	3
7	7	1350	4.5
8	8	325	1.5
9	9	670	3
10	10	1215	5
11			

ΟΠΑ AUEB

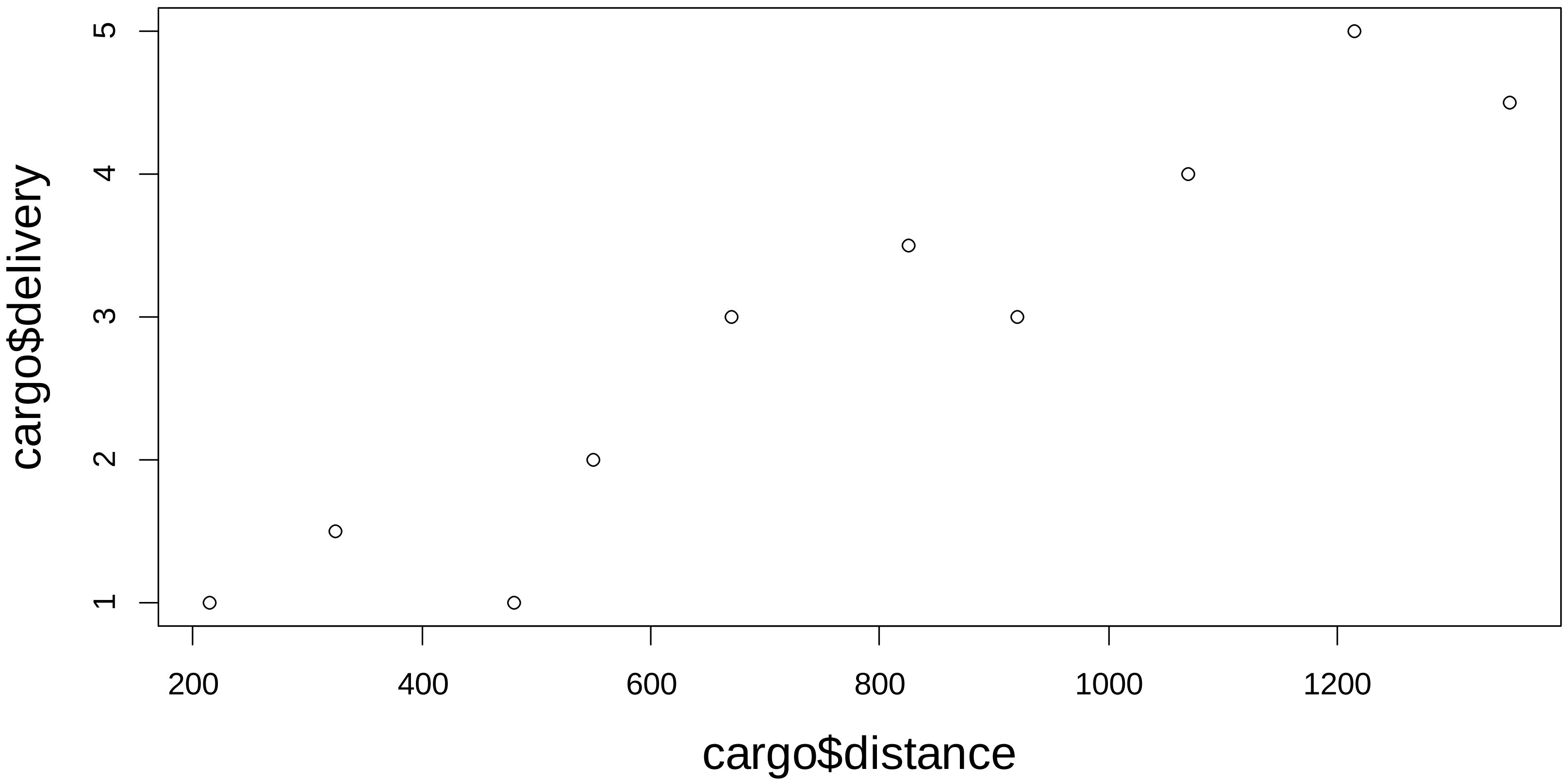
Example 5-3

Analysis in steps

- Analysis of each variable separately
- Visualization using a scatter-plot
- Correlation measures
- Regression model
- Testing for the assumptions (residual analysis)
- Revise model if necessary

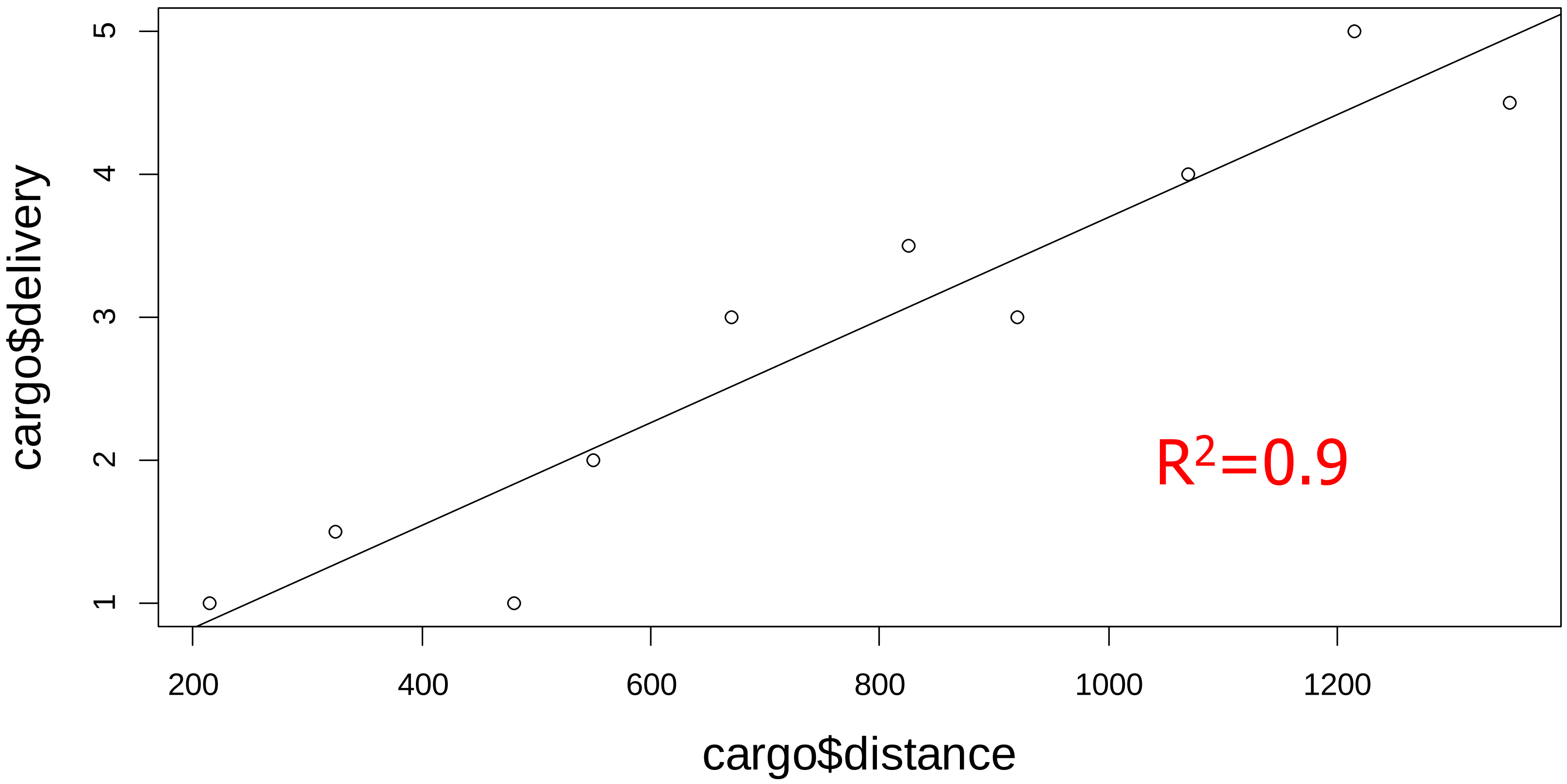
Example 5-3: Visualization

SCATTERPLOT



Example 5-3: Visualization

SCATTERPLOT

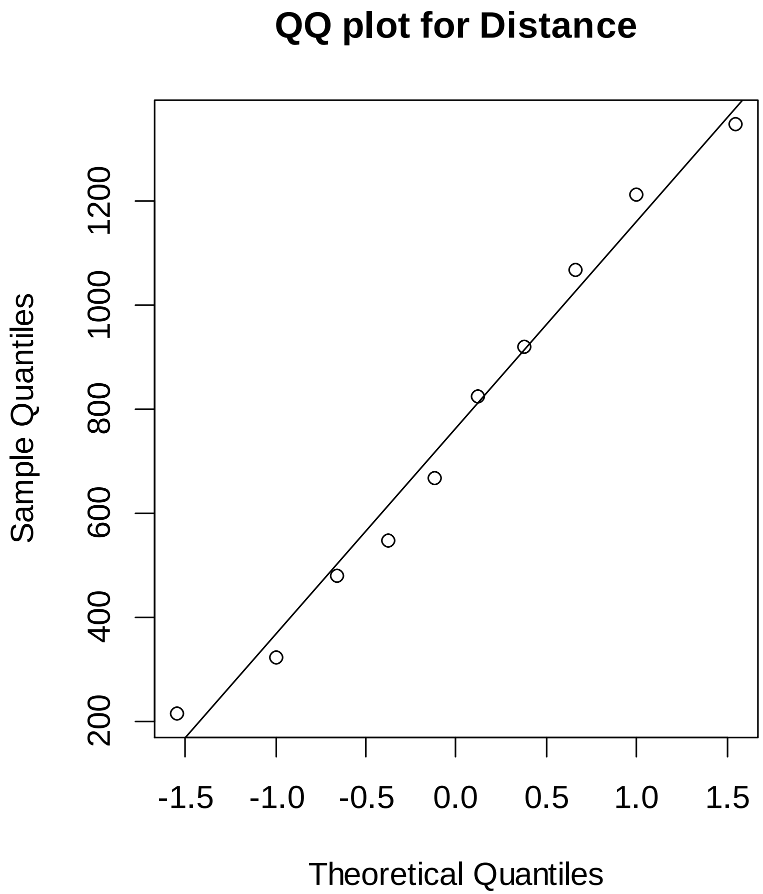


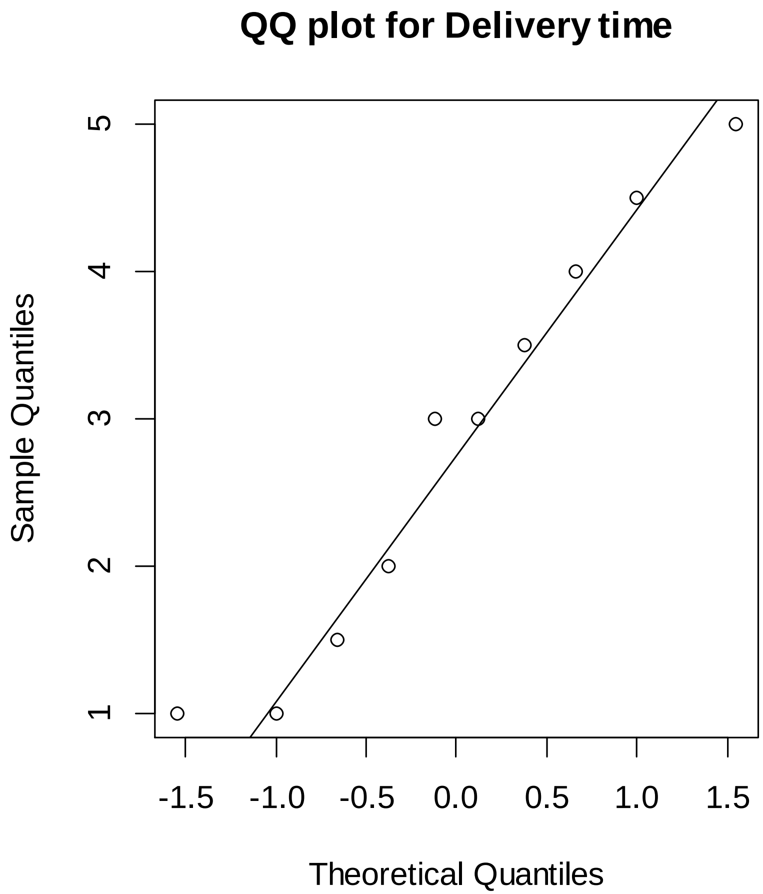
Example 5-3: Testing for normality of the original variables

```
> library(nortest)
> lillie.test(cargo$distance)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  cargo$distance
D = 0.1117, p-value = 0.9769

> shapiro.test(cargo$distance)

        Shapiro-Wilk normality test

data:  cargo$distance
W = 0.9701, p-value = 0.8915
```

```
> lillie.test(cargo$delivery)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  cargo$delivery
D = 0.1416, p-value = 0.8243

> shapiro.test(cargo$delivery)

        Shapiro-Wilk normality test

data:  cargo$delivery
W = 0.937, p-value = 0.5203
```

Example 5-3: Testing for normality of the original variables





Example 5-3: Monitoring correlation

```
> cor.test(cargo$distance, cargo$delivery )

        Pearson's product-moment correlation

data:  cargo$distance and cargo$delivery
t = 8.5086, df = 8, p-value = 2.795e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7932921 0.9881624
sample estimates:
      cor
0.9489428
```

Example 5-3: Fitting the regression model

Response — delivery
Explanatory — distance
Linear model — lm

```
> lm( delivery~distance, data=cargo )

Call:
lm(formula = delivery ~ distance, data = cargo)

Coefficients:
(Intercept)     distance
   0.118129     0.003585
```

$$Y=0.12+0.0036\,X+\varepsilon$$

```
> res_ex53 <-lm( delivery~distance, data=cargo )
> names(res_ex53)
 [1] "coefficients"  "residuals"     "effects"       "rank"
 [5] "fitted.values" "assign"        "qr"            "df.residual"
 [9] "xlevels"       "call"          "terms"         "model"
>
```

Example 5-3: Summarizing the regression model

```
> summary(res_ex53)

Call:
lm(formula = delivery ~ distance, data = cargo)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83899 -0.33483  0.07842  0.37228  0.52594 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.1181291  0.3551477   0.333    0.748    
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
Multiple R-squared:  0.9005,    Adjusted R-squared:  0.8881 
F-statistic:  72.4 on 1 and 8 DF,  p-value: 2.795e-05
```

Example 5-3: Summarizing the regression model

```
> summary(res_ex53)

Call:
lm(formula = delivery ~ distance, data = cargo)

Residuals:
     Min       1Q   Median       3Q      Max
-0.83899 -0.33483  0.07842  0.37228  0.52594

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
Multiple R-squared:  0.9005,	Adjusted R-squared:  0.8881
F-statistic:  72.4 on 1 and 8 DF,  p-value: 2.795e-05
```

Summary statistics for residuals

ΟΠΑ AUEB

Example 5-3: Summarizing the regression model

```
> summary(res_ex53)

Call:
lm(formula = delivery ~ distance, data = cargo)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83899 -0.33483  0.07842  0.37228  0.52594 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.1181291  0.3551477   0.333    0.748    
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
Multiple R-squared:  0.9005,    Adjusted R-squared:  0.8881 
F-statistic:  72.4 on 1 and 8 DF,  p-value: 2.795e-05
```

Summary table for regression coefficients

Model:

$Y=0.12+0.0036\,X + \varepsilon$

P-value for testing whether parameters are zero
Intercept = Not significant
Slope = Significant effect of distance on delivery

Example 5-3: Summarizing the regression model

Parameter estimates of the model

Days of Delivery = 0.118 + 0.00359 Miles+ ε,

ε~NORMAL(0, 0.48^2)

```
Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
```

Example 5-3: Summarizing the regression model

Standard errors of the estimates

$$\hat{\sigma}_{\hat{\beta}_0} = \sqrt{Var(\hat{\beta}_0)} = 0.355, \hat{\sigma}_{\hat{\beta}_1} = \sqrt{Var(\hat{\beta}_1)} = 0.000421$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
```

Example 5-3: Summarizing the regression model

$$t = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}}$$

Test functions

$$t_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}} = \frac{0.118}{0.355} = 0.333, \quad t_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{0.00359}{0.000421} = 8.527$$

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
```

Example 5-3: Summarizing the regression model

P-values for testing the hypothesis that each coefficient is zero

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
```

Example 5-3: Summarizing the regression model

Standardized coefficients (or beta coefficients)

- ✓The are the regression coefficients when we standardize all variables
- ✓We can use the command scale within the formula in lm in R
- ✓The beta coefficient of β_0 is always zero (0)
- ✓**Interpretation of b_1:** How many standard deviations of Y we expect Y to change when X increases by one standard deviation (of X)

```
> res_ex53beta

Call:
lm(formula = scale(delivery) ~ scale(distance), data = cargo)

Coefficients:
    (Intercept)  scale(distance)
     -7.022e-17        9.489e-01

> round(res_ex53beta$coef, 3)
    (Intercept) scale(distance)
          0.000           0.949
```

Example 5-3: Summarizing the regression model

Standardized coefficients (or beta coefficients)

✓In simple linear regression the beta coefficient is equal to the Pearson's correlation coefficient

```
> res_ex53beta

Call:
lm(formula = scale(delivery) ~ scale(distance), data = cargo)

Coefficients:
    (Intercept)  scale(distance)
     -7.022e-17        9.489e-01

> round(res_ex53beta$coef, 3)
    (Intercept) scale(distance)
          0.000           0.949
```

```
> round(cor(cargo[,-1]),3)
         distance delivery
distance    1.000    0.949
delivery    0.949    1.000
```

ΟΠΑ AUEB

Why the standardized coefficient is equal to the correlation

$$\hat{\beta}_0^{(st)} = \bar{Z}_y - \hat{\beta}_1^{(st)} \bar{Z}_x = 0 \qquad \hat{\beta}_1^{(st)} = \frac{s_{Z_y}}{s_{Z_x}} r_{Z_x Z_y} = r_{Z_x Z_y}$$

$$r_{Z_x Z_y} = \frac{\sum_{i=1}^{n} (Z_{x,i} - \bar{Z}_x)(Z_{y,i} - \bar{Z}_y)}{\sqrt{\sum_{i=1}^{n} (Z_{x,i} - \bar{Z}_x)^2 \sum_{i=1}^{n} (Z_{y,i} - \bar{Z}_y)^2}} = \frac{\sum_{i=1}^{n} Z_{x,i} Z_{y,i}}{\sqrt{\sum_{i=1}^{n} Z_{x,i}^2 \sum_{i=1}^{n} Z_{y,i}^2}} =$$

$$= \frac{\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_x} \frac{Y_i - \bar{Y}}{s_y} \right)}{\sqrt{\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_x} \right)^2 \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{s_y} \right)^2}} = r_{XY}$$

Example 5-3: Summarizing the regression model

```
> summary(res_ex53)

Call:
lm(formula = delivery ~ distance, data = cargo)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.83899 -0.33483  0.07842  0.37228  0.52594 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.1181291  0.3551477   0.333    0.748    
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
                                                   0.8881
                                                   -05
```

Residual standard deviation

$\sigma=0.48$

✓It measures the precision of the model predictions

It means that the accuracy of the prediction is 0,5 day

Fitted value ± 0,5 day will include 66% of the cases

Fitted value ± 1 day will include 95% of the cases

Full Model:

$$Y=0.12+0.0036\,X+\varepsilon$$

$$\varepsilon \sim N(0,\ 0{,}48^2)$$

Example 5-3: Summarizing the regression model

R^2= % of variability explained by the model

✓It uses the biased estimates of variance

✓It is used as a measure of goodness of fit

✓Increases with **every** covariate we add (even if it is rubbish)

✓Therefore **it should not be used** as a variable or model selection criterion

✓We can only compare models with the same number of covariate and same response

✓In simple linear regression $R^2=r^2$

```
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
Multiple R-squared:  0.9005,    Adjusted R-squared:  0.8881
F-statistic:  72.4 on 1 and 8 DF,  p-value: 2.795e-05
```

Coefficients of determination

90% of the variability is explained only using the distance as covariate

Example 5-3: Summarizing the regression model

R_{adj}^2= % of variance explained by the model adjusted for the number of covariates

- ✓It considers the number of covariates
- ✓It uses the unbiased variance estimators
- ✓It is used as a measure of goodness of fit
- ✓It does not increases always (adding very bad covariates will decrease R_{adj}^2)
- ✓It can be used as a variable or model selection criterion
- ✓In simple linear regression it does not differ a lot from R^2.

```
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
Multiple R-squared:  0.9005,     Adjusted R-squared:  0.8881
F-statistic:  72.4 on 1 and 8 DF,  p-value: 2.795e-05
```

Coefficients of determination

88% of the variability is explained only using the distance as covariate

Example 5-3: Summarizing the regression model

ANOVA table details for regression models

- ✓**In simple regression it tests for: H_0: $\beta_1=0$ vs. H_1: $\beta_1 \neq 0$**
- ✓**Be careful: in multiple regression the assumption involves all covariate effects!**
- ✓**Generally tests how much the current model differs from the constant (or null) model (that is, $y=\beta_0+\varepsilon$)**

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1181291  0.3551477   0.333    0.748
distance    0.0035851  0.0004214   8.509 2.79e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.48 on 8 degrees of freedom
Multiple R-squared:  0.9005,	Adjusted R-squared:  0.8881
F-statistic:  72.4 on 1 and 8 DF,  p-value: 2.795e-05
```

Anova table details

We reject the null hypothesis, so the model is different from the constant the delivery is significant for the model

ΟΠΑ AUEB

Example 5-3: ANOVA table for the regression model

ANOVA table details for regression models

✓**In simple regression it tests for: H_0: $\beta_1=0$ vs. H_1: $\beta_1 \neq 0$**

✓**Be careful: in multiple regression the assumption involves all covariate effects!**

✓**Generally tests how much the current model differs from the constant (or null) model (that is, $y=\beta_0+\varepsilon$)**

```
> anova(res_ex53)
Analysis of Variance Table

Response: delivery
          Df  Sum Sq Mean Sq F value    Pr(>F)
distance   1 16.6816 16.6816  72.396 2.795e-05 ***
Residuals  8  1.8434  0.2304
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis, so the model is different from the constant the delivery is significant for the model

Example 5-3: Interpretation of the results

Parameter β_1=0.00359 (the slope)

- Is there a linear effect? YES

  P=0.000<0.05 i.e. we reject the null (H_0)=> Therefore the distance influences the delivery time

- Of what direction is the relationship? POSITIVE

  β_1>0 which implies positive relationship => the longer the distance, the more delayed is the delivery

- How much the distance influences the delivery?
  - Each extra mile of distance increases the expected time by 0.00359 days (approximately 5 minutes)
  - With every extra 100 miles, the expected delivery increases by 0.359 days (approximately 8.6 hours)

ΟΠΑ AUEB

Example 5-3: Interpretation of the results

Why this interpretation?

Parameter β_1

- Let us assume two different explanatory values $X_1=X$ & $X_2=X+1$ then
- $\mu_1= \beta_0+\beta_1 X_1= \beta_0+\beta_1 X$
- $\mu_2= \beta_0+\beta_1 X_2 = \beta_0+\beta_1 (X+1)$
- $\Delta\mu=\mu_2-\mu_1=\beta_0+\beta_1 (X+1)-\beta_0-\beta_1 X=\beta_1$

ΟΠΑ AUEB

Example 5-3: Interpretation of the results

Parameter β_0=0.118 (the intercept)

- ✓ Can be removed from the equation without changing much the fit/predictions? YES

  P=0.748>0.05 i.e. we do not reject the null (H_0)=> Therefore the constant/intercept can be assumed to be equal to zero and be removed from the model

Example 5-3: Interpretation of the results

Parameter β_0=0.118 (the intercept)

- ✓ INTERPRETATION:
  - When the distance is zero then the delivery time is 0.118 days (2.8 ώρες)
  - It shows the delivery time when the cargo destination is very close
  - BE CAREFUL this value is outside the range of X since the smallest destination is 215 miles away

```
> range(cargo$distance)
[1]  215 1350
```

- ✓ Shall we remove it? Possibly YES.

  The logic here says that we should remove this term from the model

Example 5-3: Interpretation of the results

Predictive performance and goodness of fit

- ✓ R=r=0.95 & R^2=0.89;
  - High correlation between the two variables
  - Well fitted model and accurate predictions
  - 89% of the variance is explained by the model

    which means that if we know the distance we can accurately predict the delivery time

Example 5-3: Interpretation of the results

Standardized coefficient b_1=0.949

- ✓ If the distance increases by a standard deviation (i.e. 380 miles) then the delivery time is expected to increase by 0.95 standard deviations of Y (that is, by 0.949*1.435=1.36 days).

```
> sapply( cargo[,-1], sd)
  distance   delivery
379.745529   1.434689
```

ASSUMPTIONS (to be checked):

- Normality of errors (and of Y_i)
- Homoscedasticity of errors (and Y_i)
- Independence of errors (and of Y_i)
- Linearity between X & Y

- We work with the residuals e_i

ΟΠΑ AUEB

Types of residuals:

- (Unstandardised) Residuals $e_i = y_i - \widehat{y}_i$
- Standardized residuals

SPSS → $$e_i^* = \frac{y_i - \widehat{y}_i}{\widehat{\sigma}^2}$$

R – Wikipedia → SPSS Wikipedia →
$$e_i^* = \frac{y_i - \hat{y}_i}{s.e.(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

h_{ii} is the diagonal elements of the hat matrix **H**

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

- Studentized residuals
  (internally studentized)

Types of residuals:

- Standardized residuals

R – Wikipedia (internally studentized)

SPSS
Wikipedia

$$e_i^* = \frac{y_i - \hat{y}_i}{s.e.(y_i - \hat{y}_i)} = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

h_{ii} is the diagonal elements of the hat matrix **H**

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$$

- Studentized residuals
- (Deleted) Studentized residuals

  (or jack-knife residuals)

(externally studentized)
When using estimating the standard error from the regression model without using the i-th observation

Types of residuals in R:

- (Unstandardized) Residuals

```
res_ex53$residuals
residuals(res_ex53)
resid(res_ex53)
```

- Standardized residuals

```
rstandard(res_ex53)
library(MASS)
round(stdres(res_ex53),3)
```

- Studentized residuals (Jack-knife residuals)

```
rstudent(res_ex53)
library(MASS)
studres(res_ex53)
```

- **NOTE: That all "standardized" residuals will be similar for reasonably large n**

ASSUMPTIONS (to be checked):

Theoretical errors

$$
\begin{aligned}
E(\varepsilon_i) &= 0 \\
Var(\varepsilon_i) &= \sigma^2 \\
Cov(\varepsilon_i, \varepsilon_j) &= 0
\end{aligned}
$$

Estimated sample residuals

$$
\begin{aligned}
E(e_i) &= 0 \\
Var(e_i) &= \sigma^2(1 - h_{ii}) \\
Cov(e_i, e_j) &= -\sigma^2 h_{ij}
\end{aligned}
$$

ASSUMPTIONS (to be checked):

- Normality of errors (and of Y_i)

  Use studentized residuals

- Homoscedasticity of errors (and Y_i)

  Use standardized or studentized residuals (with expected variance eq. to 1)

- Independence of errors (and of Y_i)

  Use studentized/Jack-knife residuals

  (expected correlation eq. to 0)

- Linearity between X & Y

(for reasonably large n you can use any of them since they will be similar)

ASSUMPTIONS: The Normality assumption

Consequences of departures from Normality:

- The performance of hypothesis tests and confidence intervals can be compromised.
- Though, these procedures are generally robust to small departures from Normality.

How to cure the problem:

- **Use transformations (log or Box-Cox)**
- **Use non-normal errors**
- **Use GLM models for non-normal responses**
- **Use non-parametric regression models**

ASSUMPTIONS: The normality assumption

Use un-standardized residuals

- **Normality QQ-plots for unstandardized residuals**
- **Student QQ-plots for studentized residuals**
- **Lilliefors KS & Shapiro test**
- **Other normality tests**

ΟΠΑ AUEB

ASSUMPTIONS : Checking for independence

Error independence cannot be checked easily.

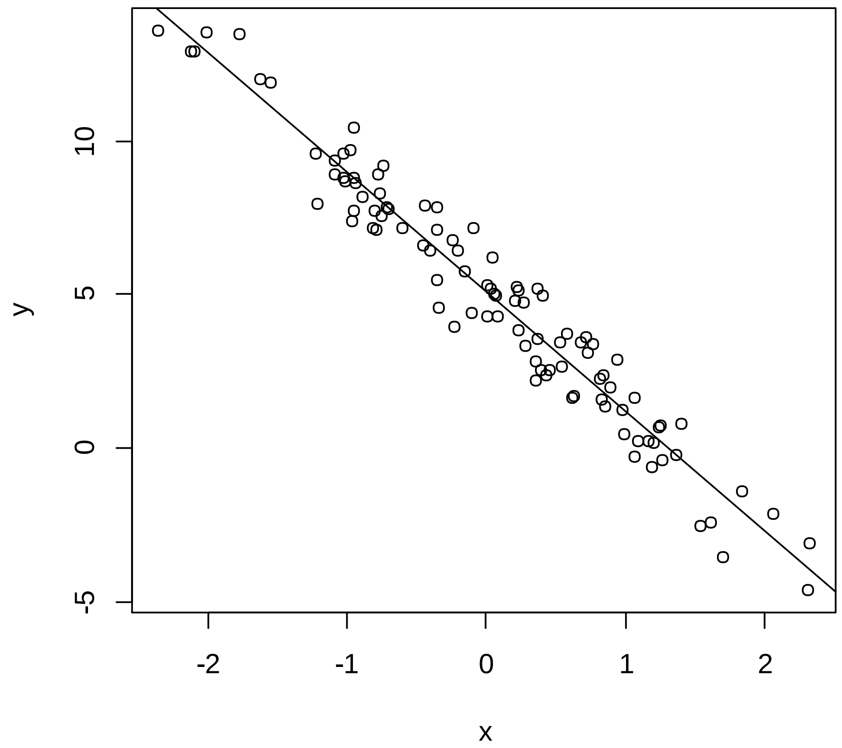
Some diagnostics are the following:

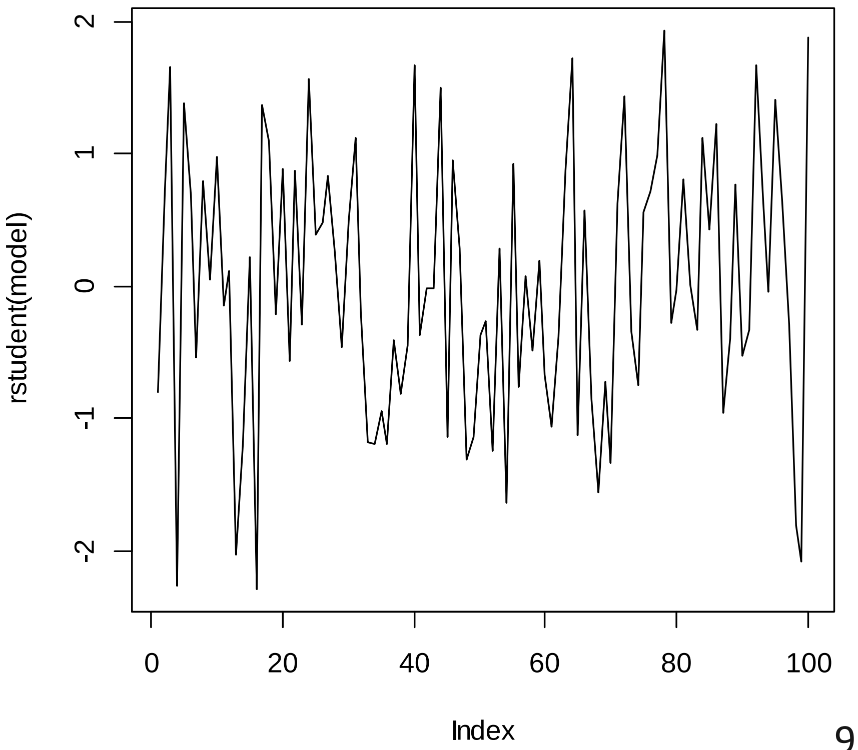
- If the data have meaning in terms of time sequence then this analysis should be skipped since it is not possible to check for indepdendence
- Time sequence plot (against id or any variable with chronological meaning)
- Test for non randomness using the runs test
- Tests for auto-correlations
  - ✓ Durbin – Watson test (testing for serial correlation of order one)
  - ✓ ACF Plots & Tests for autocorrelations
  - ✓ AR models

For details see Ryan 1997 p. 46-47

ASSUMPTIONS : Checking for independence

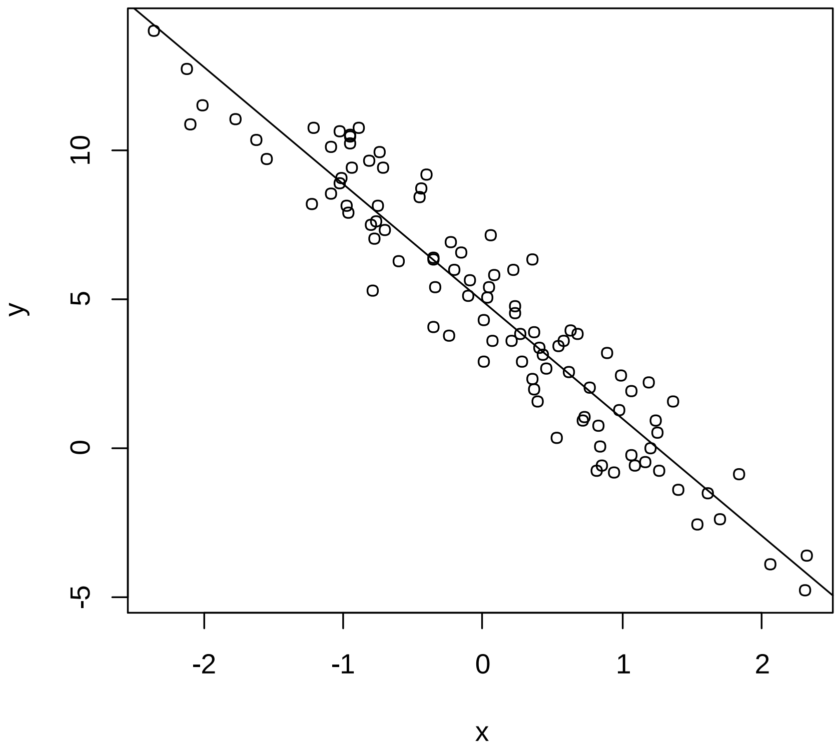
Simple time-sequence plot - Example of independence

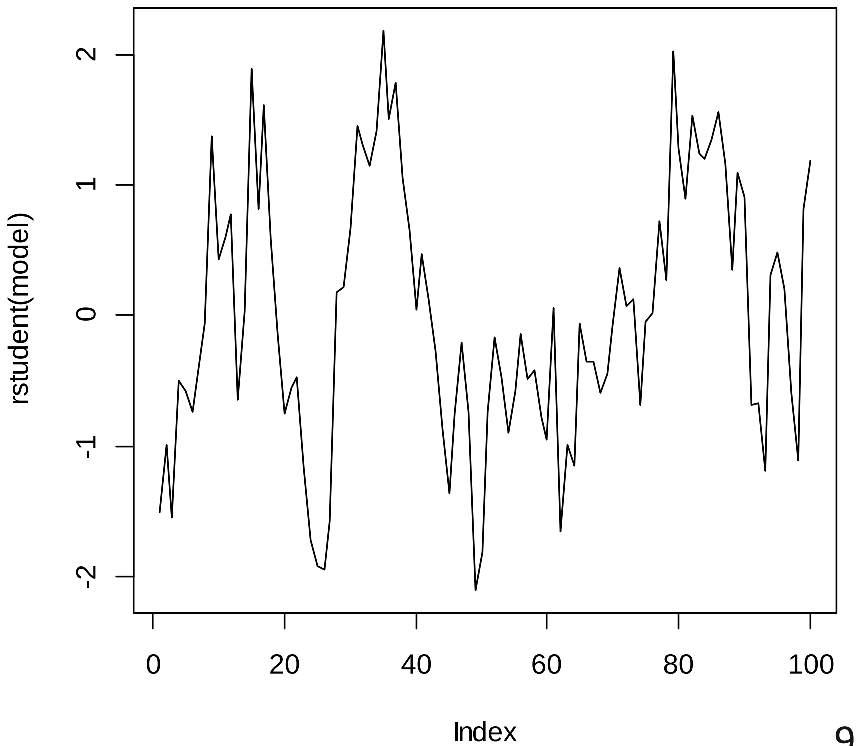




ASSUMPTIONS : Checking for independence

Simple time-sequence plot - Examples of dependence

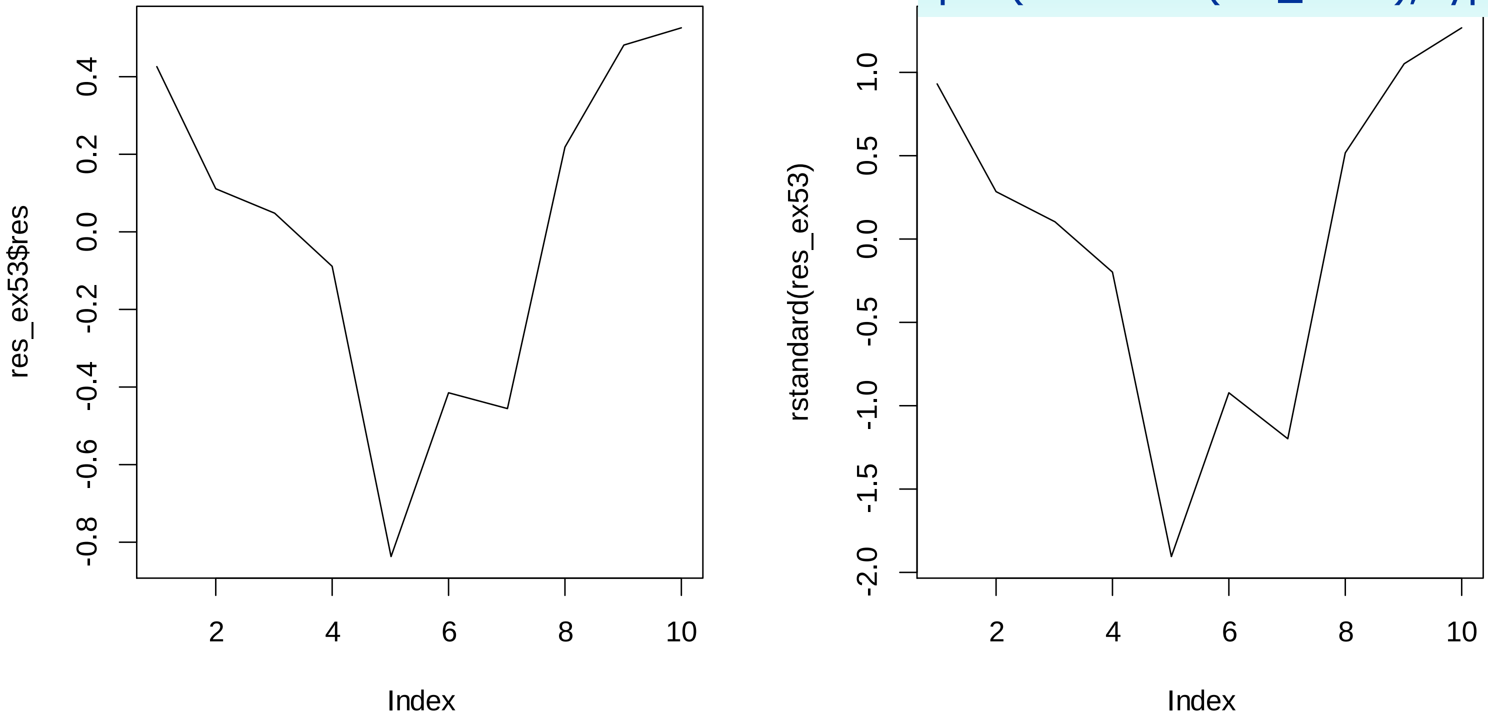




ASSUMPTIONS : Checking for independence

Simple time-sequence plot

```
par( mfrow=c(1,2) )
plot(res_ex53$res, type='l')
plot(rstandard(res_ex53), type='l')
```



ASSUMPTIONS : Checking for independence

The Durbin-Watson test for serial correlation

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

- ✓ 0<D<4
- ✓ 0<D<2 positive autocorrelation
- ✓ 2<D<4 negative autocorrelation
- ✓ D=2 ⇔ no autocorrelation

```
library(lmtest)
dwtest(res_ex53)
```

```
> dwtest(res_ex53)

        Durbin-Watson test

data:  res_ex53
DW = 0.7533, p-value = 0.01374
alternative hypothesis: true autocorrelation is greater than 0
```

Uses asymptotic test

ASSUMPTIONS : Checking for independence

The Durbin-Watson test for serial correlation

$$d = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

```
library(car)
durbinWatsonTest(res_ex53)
dwt(res_ex53)
dwt(res_ex53$resid)
```

```
> library(car)
> durbinWatsonTest(res_ex53)
 lag Autocorrelation D-W Statistic p-value
   1       0.4995069     0.7533433   0.038
 Alternative hypothesis: rho != 0
> dwt(res_ex53)
 lag Autocorrelation D-W Statistic p-value
   1       0.4995069     0.7533433   0.024
 Alternative hypothesis: rho != 0
```

Uses bootstrap

ASSUMPTIONS: Homoscedasticity of errors (and Y_i)

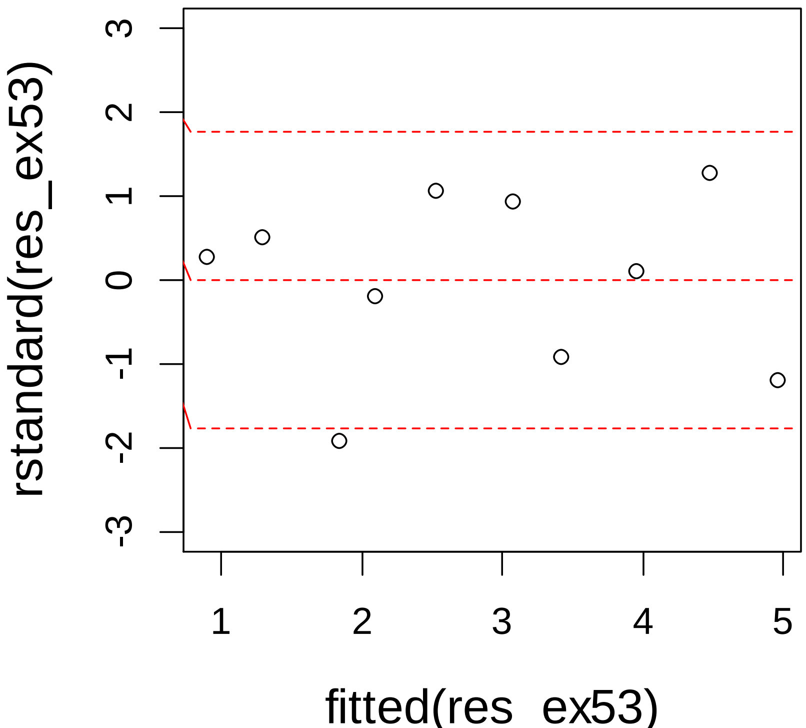
- Plot of covariates vs. residuals
- Plot fitted values vs. residuals
- Plot fitted values vs. **squared** residuals
- Plot of fitted values vs. **squared root** residuals
- Checking for equality of variance in quartiles of fitted values
- Score tests for nonconstant error variance (Breusch & Pagan, 1979 – Cook & Weisberg, 1983)

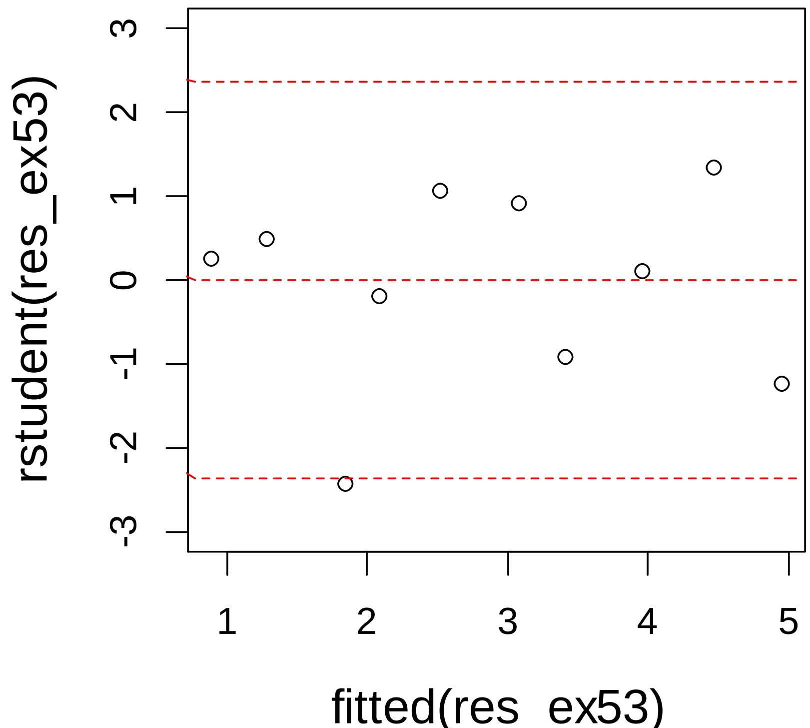
For more details see

- Fox (2002. 1st edition p. 206-209)
- Draper & Smith (1998, 3rd edition, p. 56-59, 62-67)
- Gunst & Mason (1980, p 237)

ASSUMPTIONS: Homoscedasticity of errors

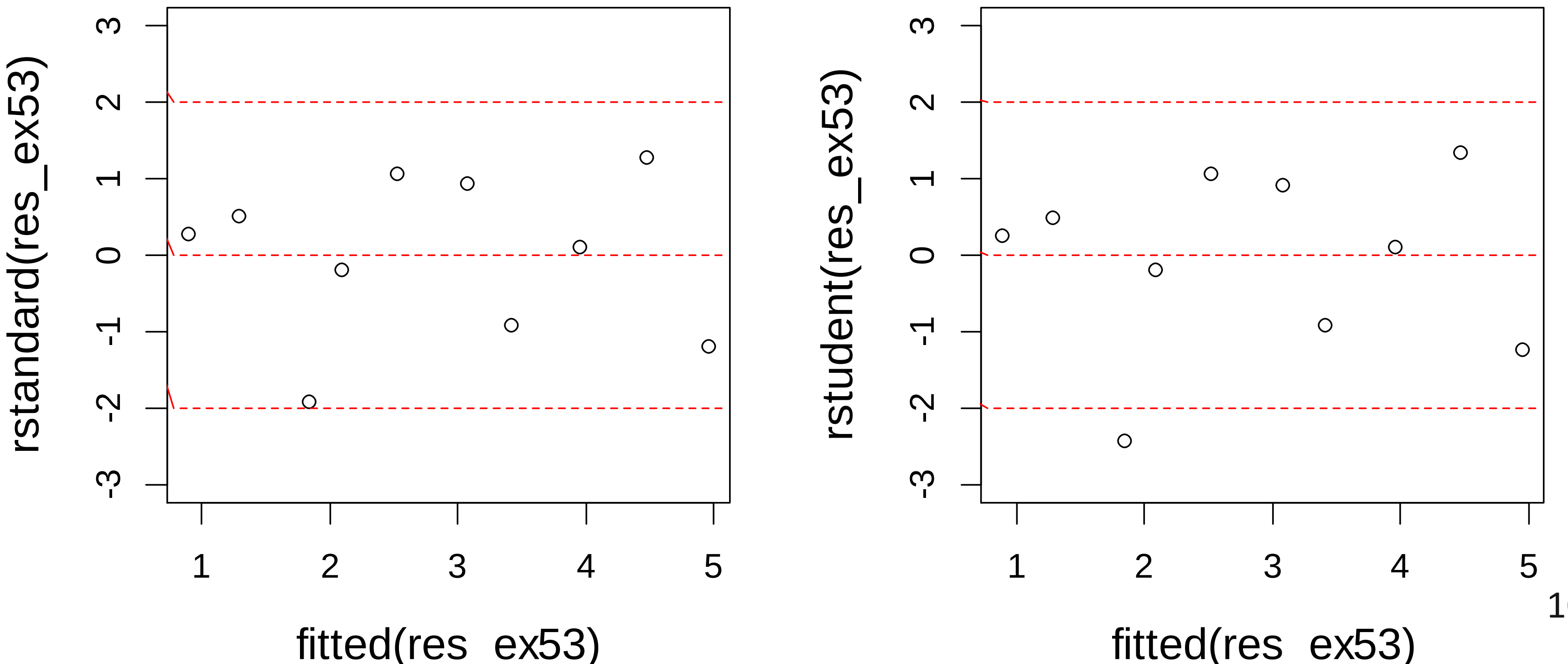
- Fitted values vs. standardized or studentized residuals using 95% quantiles from the correct distributions





ASSUMPTIONS: Homoscedasticity of errors

– Fitted values vs. standardized or studentized residuals using ± 2 (i.e. 95% quantiles assuming approximate normality)



ASSUMPTIONS: Homoscedasticity of errors

- Fitted values vs. standardized or studentized residuals using 95% quantiles from the correct distributions

```
par( mfrow=c(1,2), cex=1.3, cex.lab=1.3)
 n<-nrow(cargo)
 p<-2
 plot( fitted(res_ex53), rstandard(res_ex53), ylim= range( c(-3,3,
rstandard(res_ex53)) ) )
 ub <- sqrt(qbeta( 0.95, 0.5, 0.5*(n-p-1) )*(n-p-1))
 abline( h=c(-ub,0,ub), col=2,lty=2 )

 plot( fitted(res_ex53), rstudent(res_ex53), ylim= range( c(-3,3,
rstandard(res_ex53)) ) )
 ub <- qt( 0.975, (n-p-1) )
 abline( h=c(-ub,0,ub), col=2,lty=2 )
```

ASSUMPTIONS: Non-linearity

Consequences of departures from linearity: if linearity fails

- The error variance will appear as non-constant even if it is constant due to the model misspecification
- the model is inadequate, especially for prediction.

How to cure the problem:

- Transform the response
- Transform the covariates
- Use polynomial regression or non-parametric regression models
- Use non-linear models

ASSUMPTIONS: Non-linearity

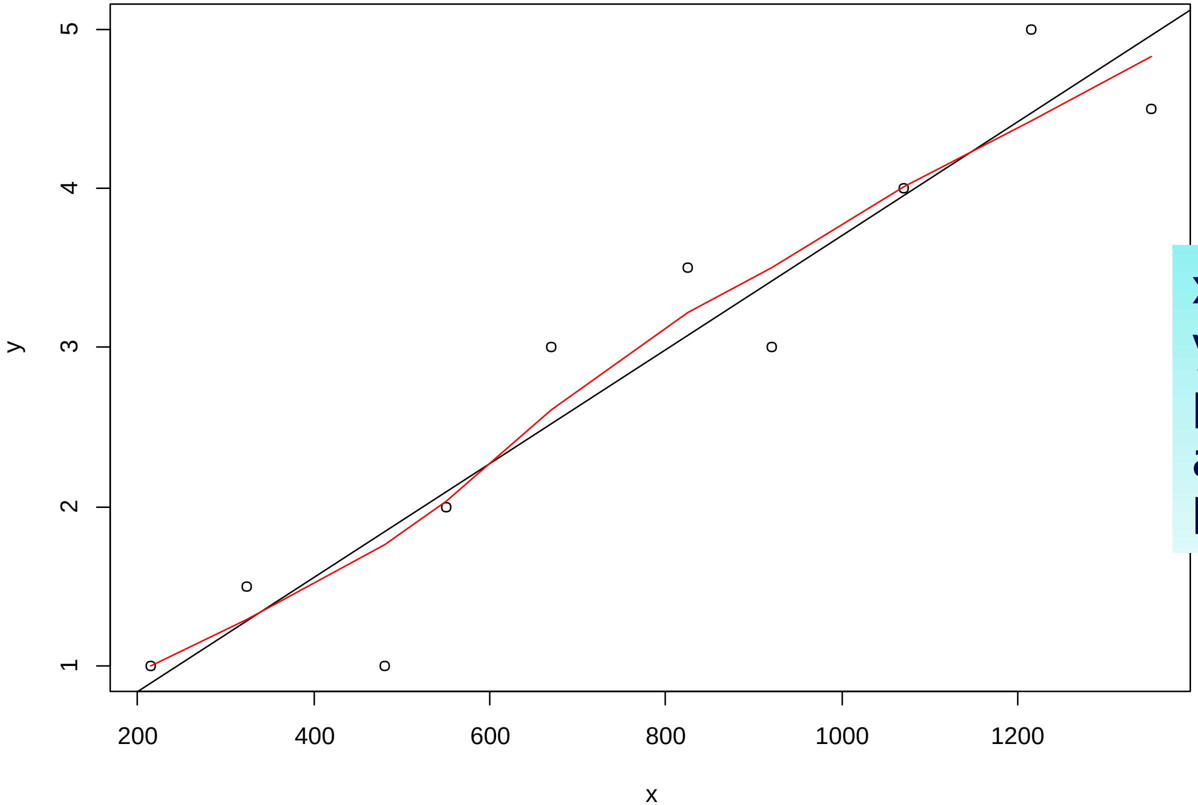
- Plot of X vs. Y
- Plot of residuals vs. covariates
- Tukey's test and residualPlot
- Fit polynomial models
- Partial residual plots (cr.plot)

5. *Correlation and Regression models*
5.2.4. *Checking for model assumptions*

ASSUMPTIONS: Non-linearity

- Plot of X vs. Y

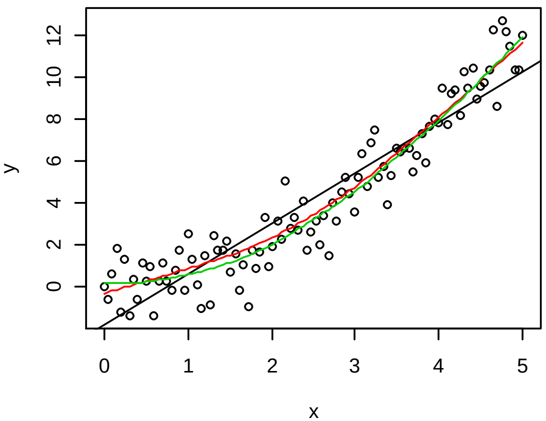
- There are several types of nonparametric regression. The most commonly used is the ***lowess (or loess)*** procedure first developed by Cleveland (1979)
  - Lowess (or loess) is an acronym for ***locally weighted scatterplot smoothing***
  - These models fit local polynomial regressions and join them together



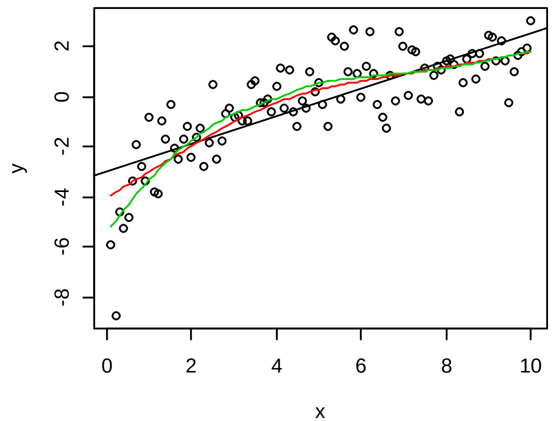

```
x<-cargo$distance
y<-cargo$delivery
plot(x,y)
abline(res_ex53)
lines(lowess(x,y), col=2)
```

5. *Correlation and Regression models*
5.2.4. *Checking for model assumptions*

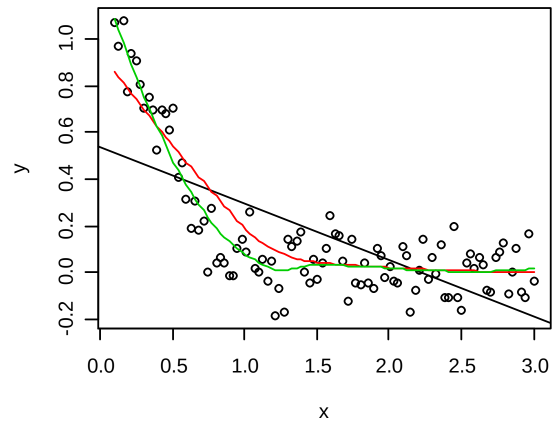
$$y = 0.5x^2 + N(0,1)$$



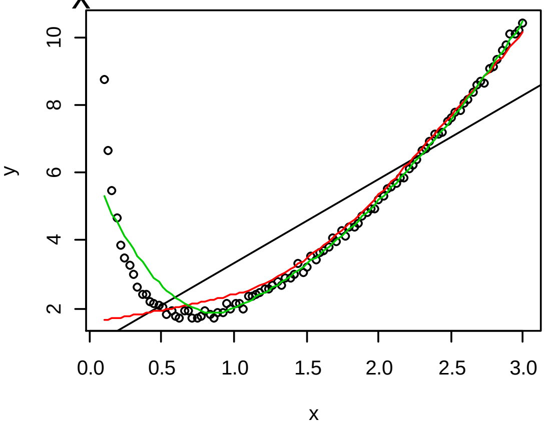
$$y = -3 + 2\log(x) + N(0,1)$$



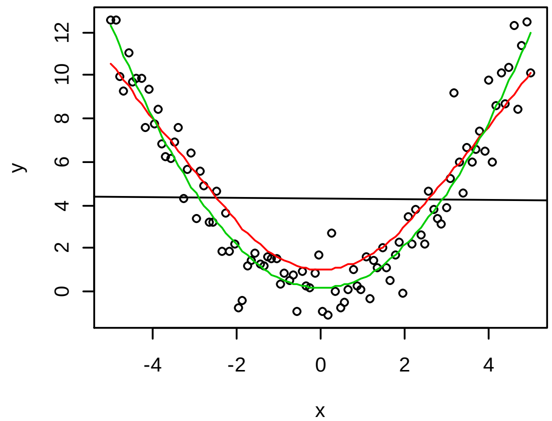
$$y = e^{-3x^2} + N(0, 0.01)$$



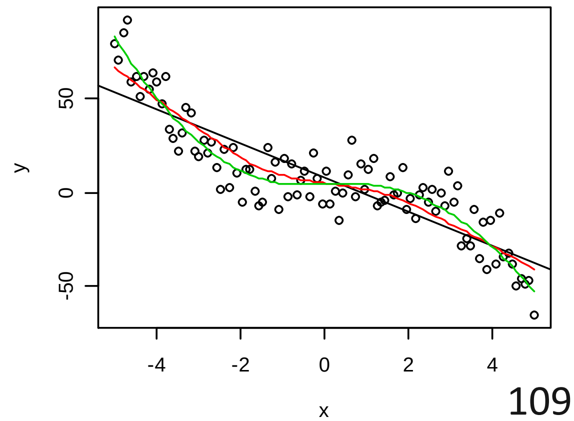
$$y = \frac{1}{x} + e^{-3x^2} + x^2 + \log(x) + N(0, 0.01$$



$$y = 0.5x^2 + N(0,1)$$

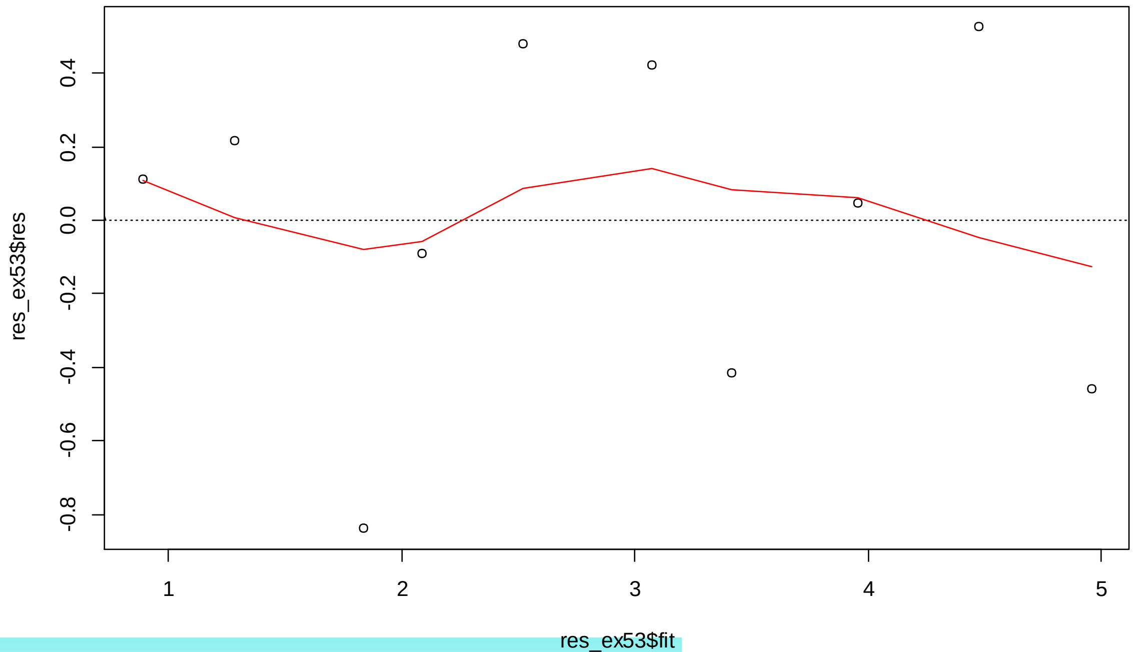


$$y = 5 + 0.5x^2 - 0.6x^3 + N(0,100)$$



ASSUMPTIONS: Non-linearity

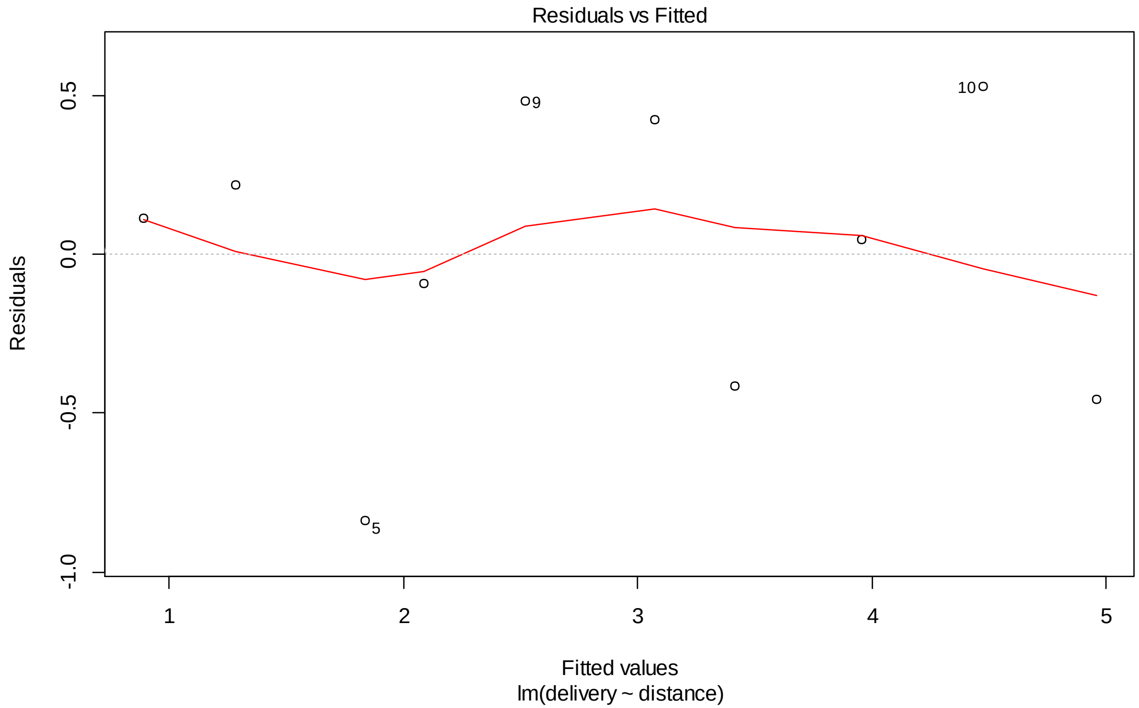
- Plot of residuals vs. covariates




```
plot(res_ex53$fit, res_ex53$res)
abline(h=0, lty=3)
lines(lowess(res_ex53$fit,res_ex53$res), col=2)
```

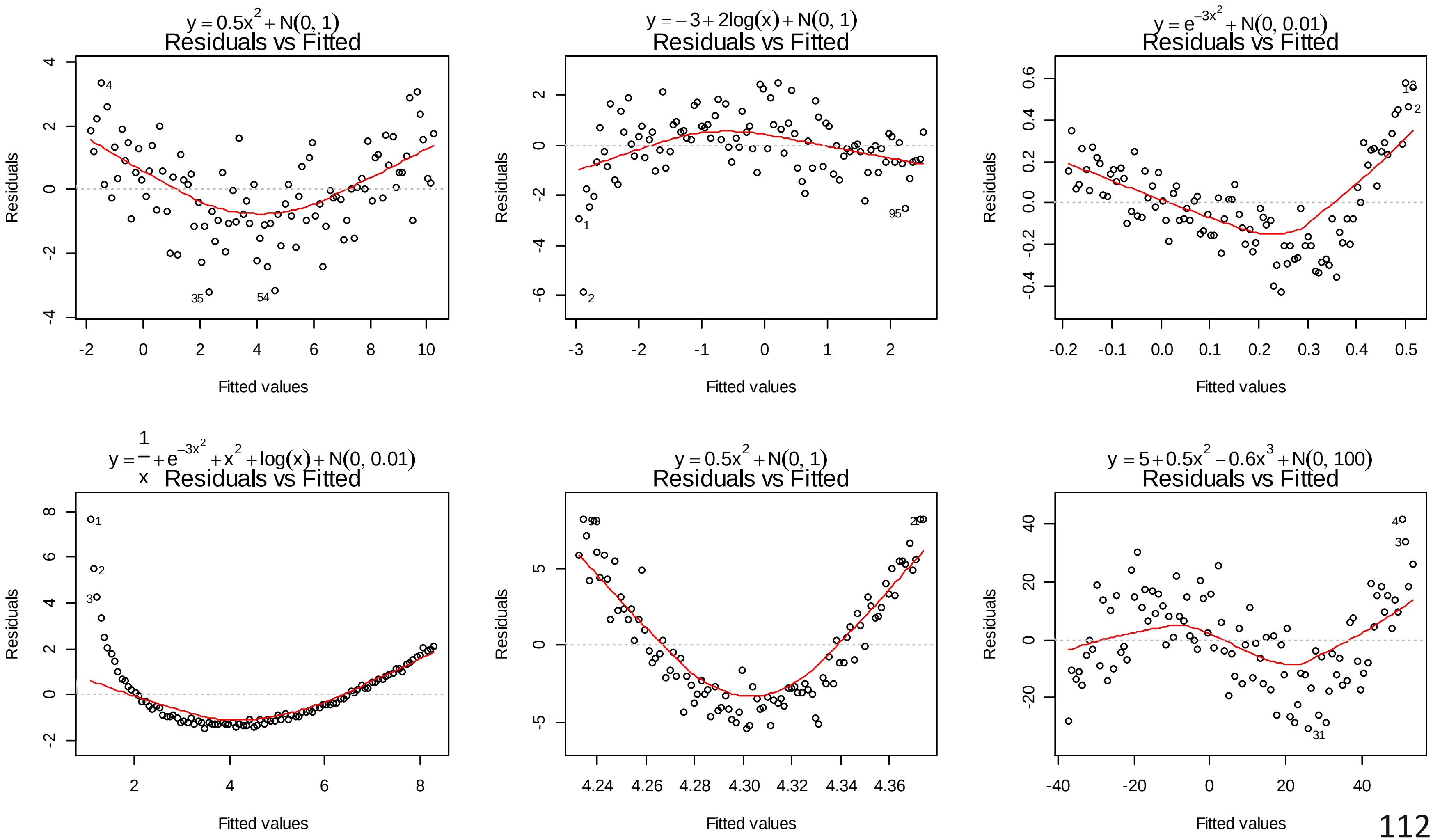
ASSUMPTIONS: Non-linearity

- Plot of residuals vs. covariates




```
plot(res_ex53, which=1)
```

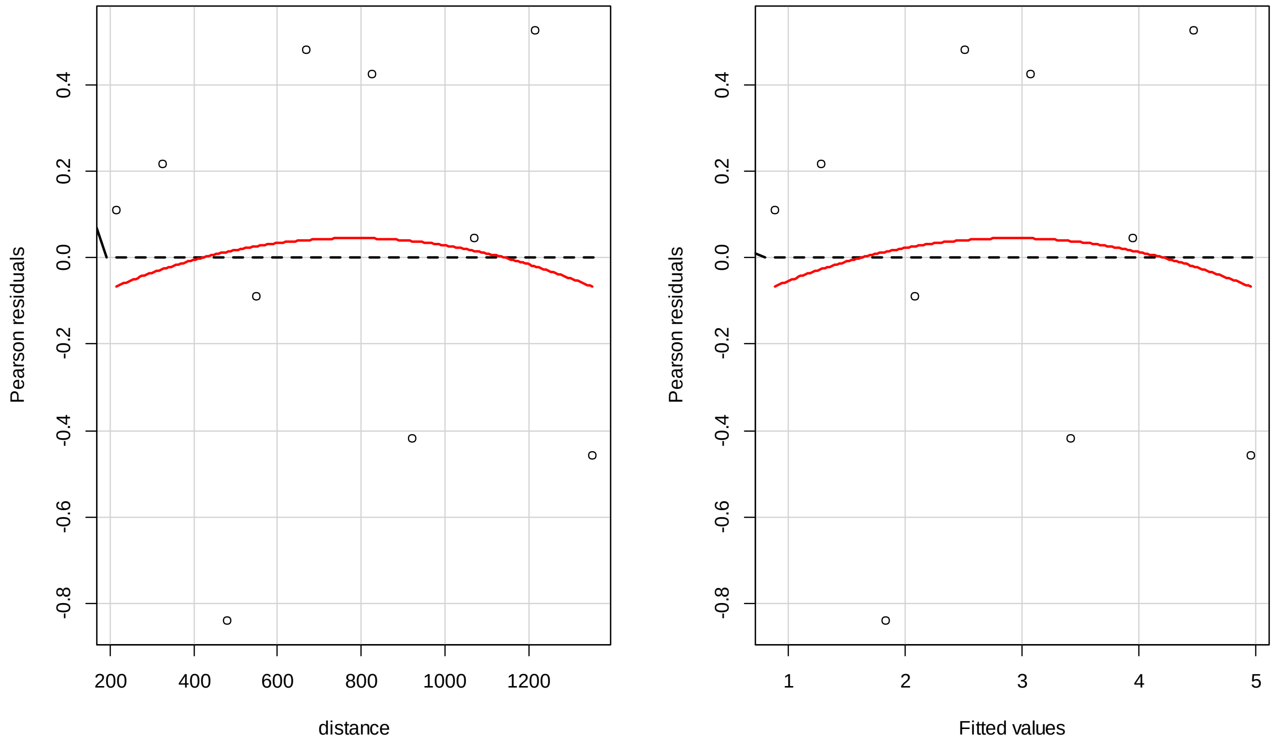
5. *Correlation and Regression models*
5.2.4. Checking for model assumptions



ASSUMPTIONS: Non-linearity

- Tukey's test and residualPlot

```
> residualPlots(res_ex53)
             Test stat Pr(>|t|)
distance         -0.25    0.810
Tukey test       -0.25    0.803
```



ASSUMPTIONS: Non-linearity

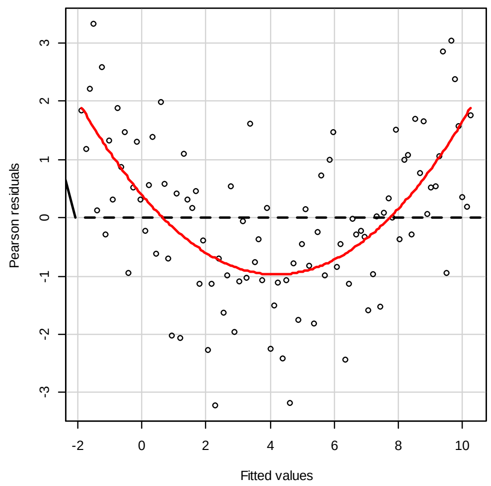
- Tukey's test and residualPlot

```
> residualPlots(res_ex53)
           Test stat Pr(>|t|)
distance       -0.25    0.810
Tukey test     -0.25    0.803
```

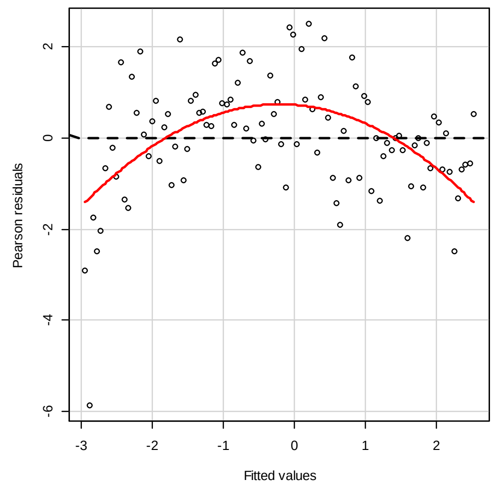
```
> summary(lm( delivery~distance+I(distance^2), data=cargo ))

Call:
lm(formula = delivery ~ distance + I(distance^2), data = cargo)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.8527 -0.3224  0.1033  0.3457  0.5461 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.824e-02  7.664e-01  -0.063    0.952
distance       4.127e-03  2.216e-03   1.863    0.105
I(distance^2) -3.465e-07  1.389e-06  -0.250    0.810

Residual standard error: 0.5109 on 7 degrees of freedom
Multiple R-squared:  0.9014,    Adjusted R-squared:  0.8732 
F-statistic: 31.99 on 2 and 7 DF,  p-value: 0.0003013
```

5. Correlation and Regression models
5.2.4. Checking for model assumptions

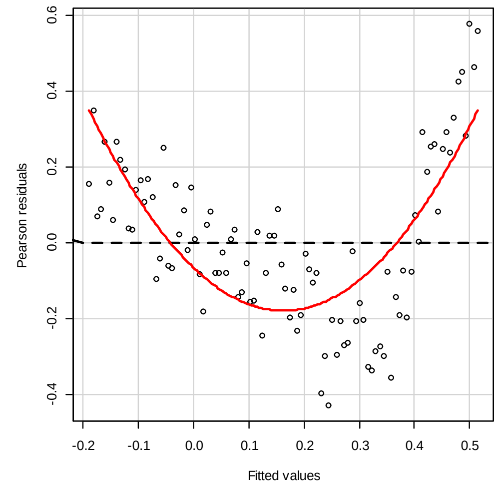
$y = 0.5x^2 + N(0,1)$



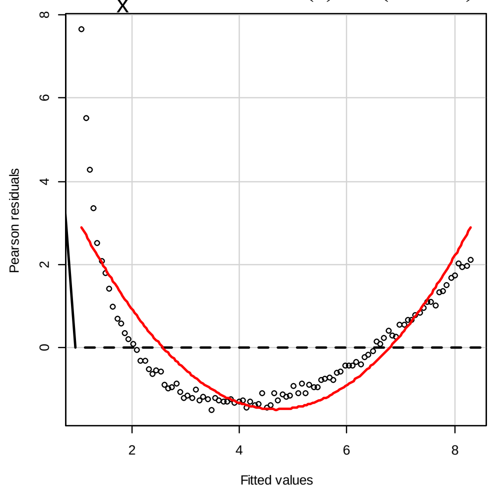
$y = -3 + 2\log(x) + N(0,1)$



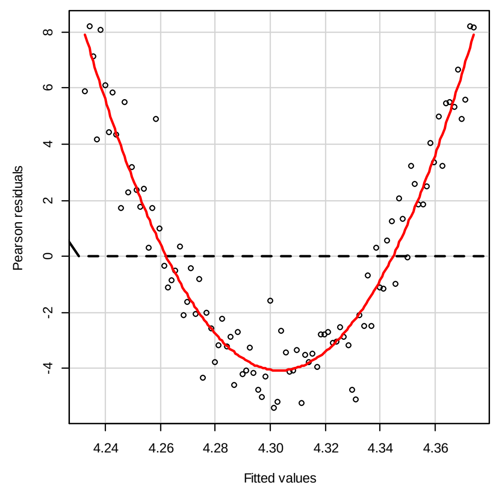
$y = e^{-3x^2} + N(0, 0.01)$



$y = \frac{1}{x} + e^{-3x^2} + x^2 + \log(x) + N(0, 0.01)$



$y = 0.5x^2 + N(0,1)$



$y = 5 + 0.5x^2 - 0.6x^3 + N(0,100)$