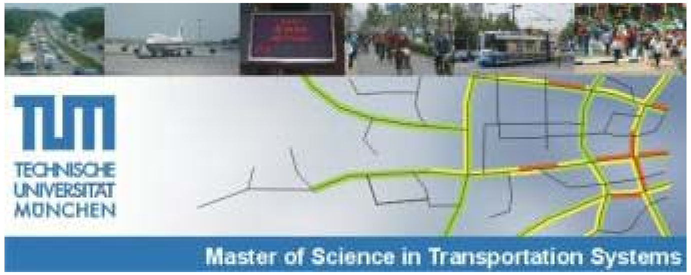
M.Sc. in ‚Transportation Systems‘



Applied Statistics in Transport
Correlation, Regression

Prof. Regine Gerike
Technische Universität München, mobil.TUM
regine.gerike@tum.de

Munich, 24/01/2012

Plan for Today's Lecture:

- Covariance and correlation
- Regression Analysis

Correlation – Regression - Introduction

- So far: tests for differences of random variables
- Regression and correlation analysis:
  describe and analyse the relationship between random variables
- Regression: describes the type of directional relationship between mainly ratio/interval scaled variables (the more … the more/less …)
- Correlation: describes the intensity of the non-directional relationship
- Example:
- Relation between the cubic capacity and the fuel consumption of a car
- Regression of the response variable PS as a function of income

Correlation – Regression - Introduction

The simplest regression analysis is linear regression: $y = b * x + a$

Simple linear regression:

Simple: one predictor variable (X) to describe the behaviour of dependent variable (Y).

Linear: It assumes a linear relationship between X and Y.

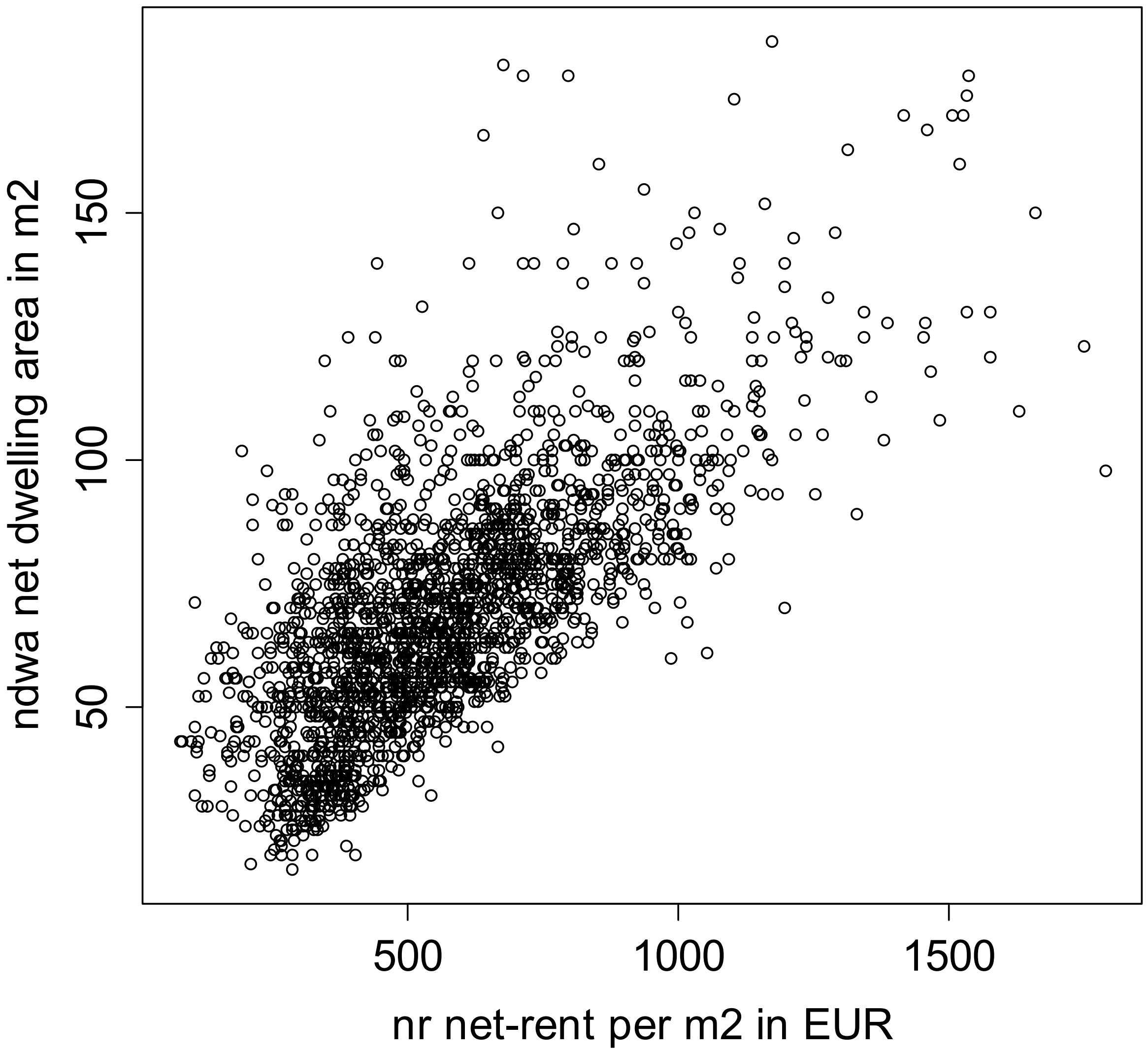
Steps Regression Analysis:

Step 1: Plot one variable against another (descriptive).

Step 2: Calculate sample coefficient of correlation r.

Step 3: Compute the coefficients.

Scatterplot for rent.data



Correlation - Introduction

- Correlation describes the interdependence between two variables.
- We are interested in the direction and the intensity of interdependence.
- For to cardinally scaled variables: Bravais-Pearson-correlation coefficient
- For at least one ordinally scaled variable: Spearman's rank correlation coefficient (Spearman's rho, ρ)
- Kendall tau coefficient : number of concordant pairs/discordant pairs

Covariance, Bravais-Pearson-Correlation Coefficient

- The **Covariance** is a measure of association between two random variables obtained as the expected value of the product of the two random variables around their means:

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n} = \frac{\sum_{i=1}^{n} x_i * y_i - \frac{\sum_{i=1}^{n} x_i * \sum_{i=1}^{n} y_i}{n}}{n}$$

- The **Correlation Coefficient** is a dimensionless measure of the interdependence between two variables, lying in the interval from -1 to +1, with zero indicating the absence of correlation.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad r = \frac{cov(x, y)}{s_x * s_y}$$

Covariance, Correlation - Notation

$$r = \frac{cov(x,y)}{s_x * s_y} = \frac{s_{xy}}{s_x * s_y}$$

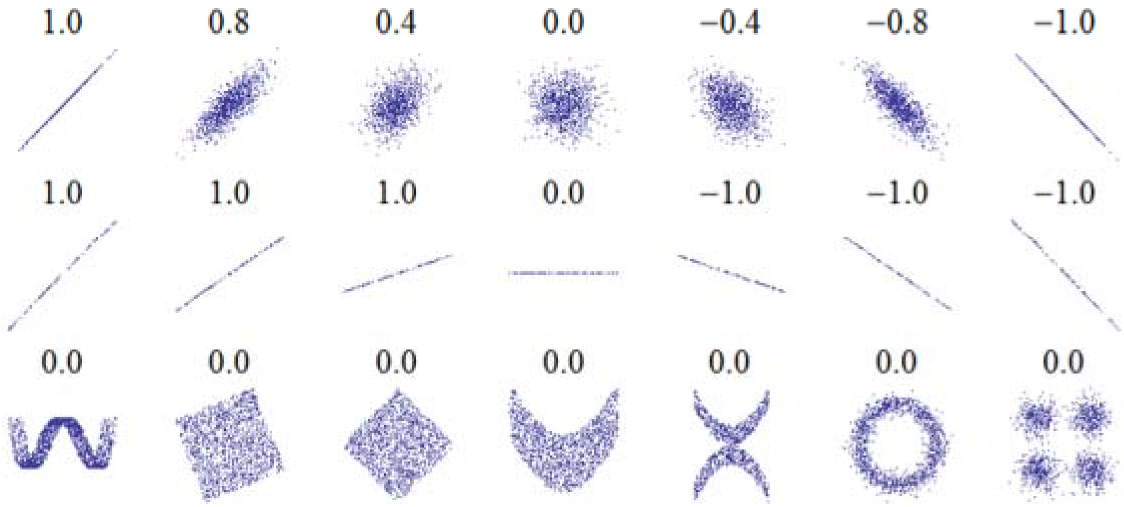
$$s_x{}^2 = \frac{1}{n-1} * S_{XX}\,,\; s_y{}^2 = \frac{1}{n-1} * S_{YY}\,,\; s_{XY} = \frac{1}{n-1} * S_{XY}$$

$$S_{XX} = \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n}\,,\; S_{YY} = \sum y_j^2 f_j - \frac{\left(\sum y_j f_j\right)^2}{n},\; S_{XY} = \sum x_i y_j f_{ij} - \frac{(\sum x_i f_i)(\sum y_j f_j)}{n}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}}$$

Properties of Bravais Correlation Coefficient:

- $r \in [-1,1]$
- Correlation is a measure of linear relationship.
- r=0 No linear relationship
- r=1 Perfect positive dependence
- r=-1 perfect negative dependence
- Symmetricity cor(x,y)=cor(y,x), ∀x,y
- Lacking causality



Spearman's Rank Correlation Coefficient

- The raw scores are converted to ranks, and the differences d_i between the ranks of each observation on the two variables are calculated:

$$r_{SP} = \frac{\sum(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{(rg(x_i) - \overline{rg}_X)^2 (rg(y_i) - \overline{rg}_Y)^2}}$$

$$\overline{rg}_X = \frac{1}{n}\sum_{i=1}^{n} rg(x_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2,\ \overline{rg}_Y = \frac{1}{n}\sum_{i=1}^{n} rg(y_i) = \frac{1}{n}\sum_{i=1}^{n} i = (n+1)/2$$

- Or simplified: $r_{SP} = 1 - \frac{6\sum d_i^2}{(n^2-1)n}$ (condition: no ties)

- $d_i=x_i-y_i$ → difference between the ranks of corresponding values X_i and Y_i
- n → number of values in each data set (same for both sets)

Range: $-1 \leq r_{SP} \leq 1$

$r_{SP} > 0$ same direction of relationship: x high → y high and x low → y low

$r_{SP} < 0$ opposite direction of relationship: x high → y low and x low → y high

$r_{SP} \approx 0$ no monotonous relationship

Spearman's Rank Correlation Coefficient - Example

Two business associates intend to establish a new branch for their company. They can choose among four locations which they rank as follows:

Location	I	II	III	IV
Ranking A	2	4	1	3
Ranking B	1	3	2	4

Do they have similar or controversial opinions about the location choice?

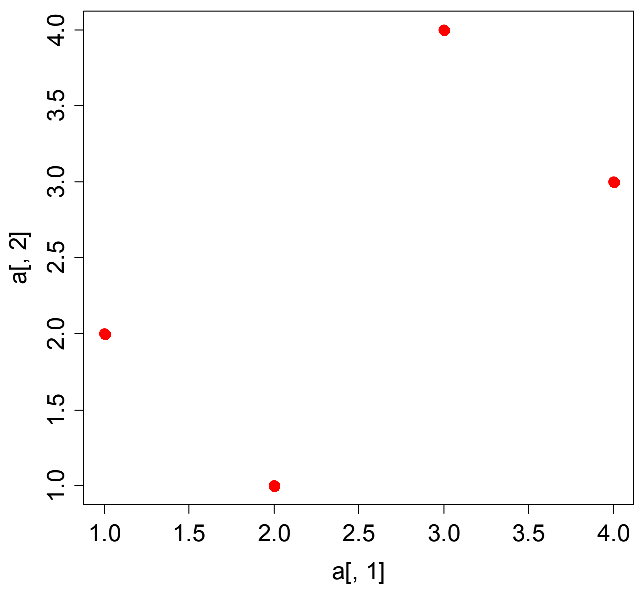
Location	Ranking A	Ranking B	D_i	D_i^2
I	2	1	1	1
II	4	3	1	1
III	1	2	-1	1
IV	3	4	-1	1

$$r_{SP} = 1 - \frac{6\sum d_i^2}{(n^2-1)n} = 1 - \frac{6*4}{(16-1)4} = 1 - \frac{6}{15} = 0.6$$

There is a good consensus among the business associates.

Spearman's Rank Correlation Coefficient – Example – R-Code

```
> (a=matrix(c(2,4,1,3,1,3,2,4),nrow=4))
     [,1] [,2]
[1,]    2    1
[2,]    4    3
[3,]    1    2
[4,]    3    4
> cor(a[,1],a[,2],method="spearman") #gives 0.6
[1] 0.6
> plot(a[,1],a[,2],pch=19,col="red",lwd=1.5,cex=1.5,cex.axis=1.5,cex.lab=1.5)
```



Spearman measures not the intensity of a linear but of a monotonic relationship.

Kendall Tau Coefficient

- Less parametric / less strict than Spearman's Rank Correlation Coefficient

$$\tau = \frac{concordant\ pairs - disconcordant\ pairs}{\sqrt{concordant + disconcordant - extra_y}\sqrt{concordant + disconcordant - extra_x}}$$

- x and y equal: concordant, x and y not equal: disconcordant
- Only x equal: $extra_x$, only y equal: $extra_y$

- Range: $-1 \leq \tau \leq 1$
- 1: perfect concordance, -1: perfect disconcordance
- Reflects also monotonous non-linear relationships

- More robust than Pearson
- Kendall τ is recommended for small sample sizes, non normally distributed data, unequal scales
- Kendall τ and Spearman are highly correlated, show in most cases the same direction and intensity of relationship

Please see

HensherSmith1984.pdf

for more correlation formulae for the different scales of variables:

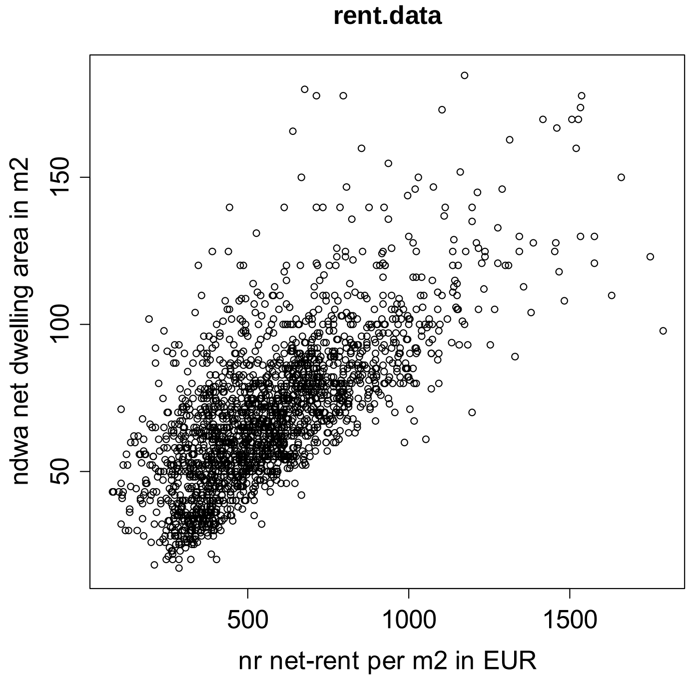
ratio, interval, ordinal, dichotomous, nominal.

Correlation Coefficients in R

```
#from help documentation:
var(x, y = NULL, na.rm = FALSE, use)
cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))
cor(x, y = NULL, use = "everything",
     method = c("pearson", "kendall", "spearman"))
cov2cor(V)
```

Correlation Coefficients in R , Example rent.data

```
> rent.data<-read.table("rent.asc", header=T, sep="\t")
> attach(rent.data)
> plot(nr,ndwa,main="rent.data",xlab="nr net-rent per m2 in EUR",
+ ylab="ndwa net dwelling area in m2",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
> str(rent.data)
'data.frame':   2053 obs. of  13 variables:
 $ nr        : num  741 716 528 554 698 ...
 $ nrsqm     : num  10.9 11.01 8.38 8.52 6.98 ...
 $ ndwa      : int  68 65 63 65 100 81 55 79 52 77 ...
 $ rooms     : int  2 2 3 3 4 4 2 3 1 3 ...
 $ yc        : num  1918 1995 1918 1983 1995 ...
 $ n         : int  2 2 2 16 16 16 6 6 6 6 ...
 $ agood     : int  1 1 1 0 1 0 0 0 0 0 ...
 $ abest     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hw        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ch        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ tb        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ bathextra: int  0 0 0 1 1 0 1 0 0 0 ...
 $ kextra    : int  0 0 0 0 1 0 0 0 0 0 ...
```



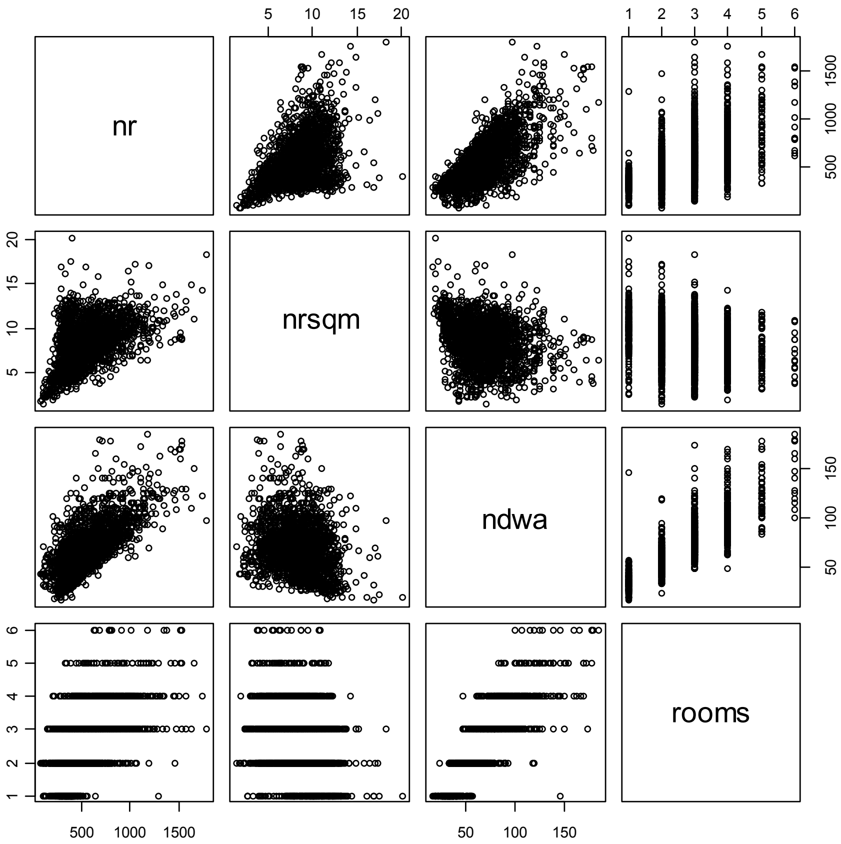
Correlation Coefficients in R , Example rent.data

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \qquad r = \frac{cov(x,y)}{s_x * s_y}$$

```
> var(nr)
[1] 60238.1
> var(ndwa)
[1] 633.1543
> var(nr,ndwa)
[1] 4369.12
> var(nr,ndwa)/sqrt(var(nr)*var(ndwa))
[1] 0.7074627
> cor(nr,ndwa)
[1] 0.7074627
```

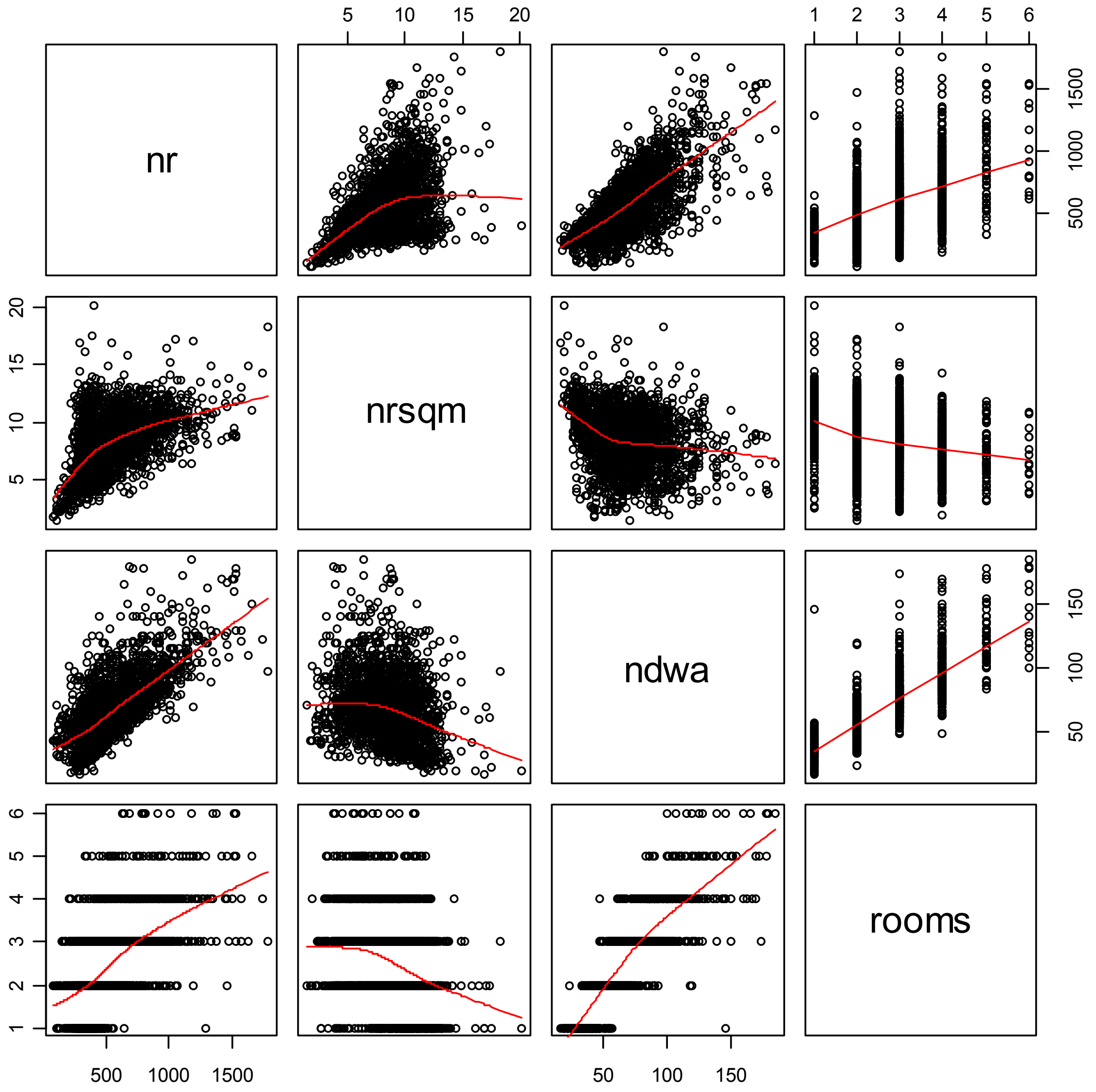
Correlation Coefficients in R , Example rent.data

```
> cor(rent.data[1:4])
             nr      nrsqm       ndwa      rooms
nr    1.0000000  0.4747967  0.7074627  0.5442473
nrsqm 0.4747967  1.0000000 -0.2268304 -0.2729057
ndwa  0.7074627 -0.2268304  1.0000000  0.8406454
rooms 0.5442473 -0.2729057  0.8406454  1.0000000
> pairs(rent.data[1:4])
```



Correlation Coefficients in R , Example rent.data

```
> pairs(rent.data[1:4],panel=panel.smooth)
```



Correlation Coefficients in R , Example rent.data

```
> cor.test(nr,ndwa)

        Pearson's product-moment correlation

data:  nr and ndwa
t = 45.3336, df = 2051, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6851715 0.7284298
sample estimates:
      cor
0.7074627
```

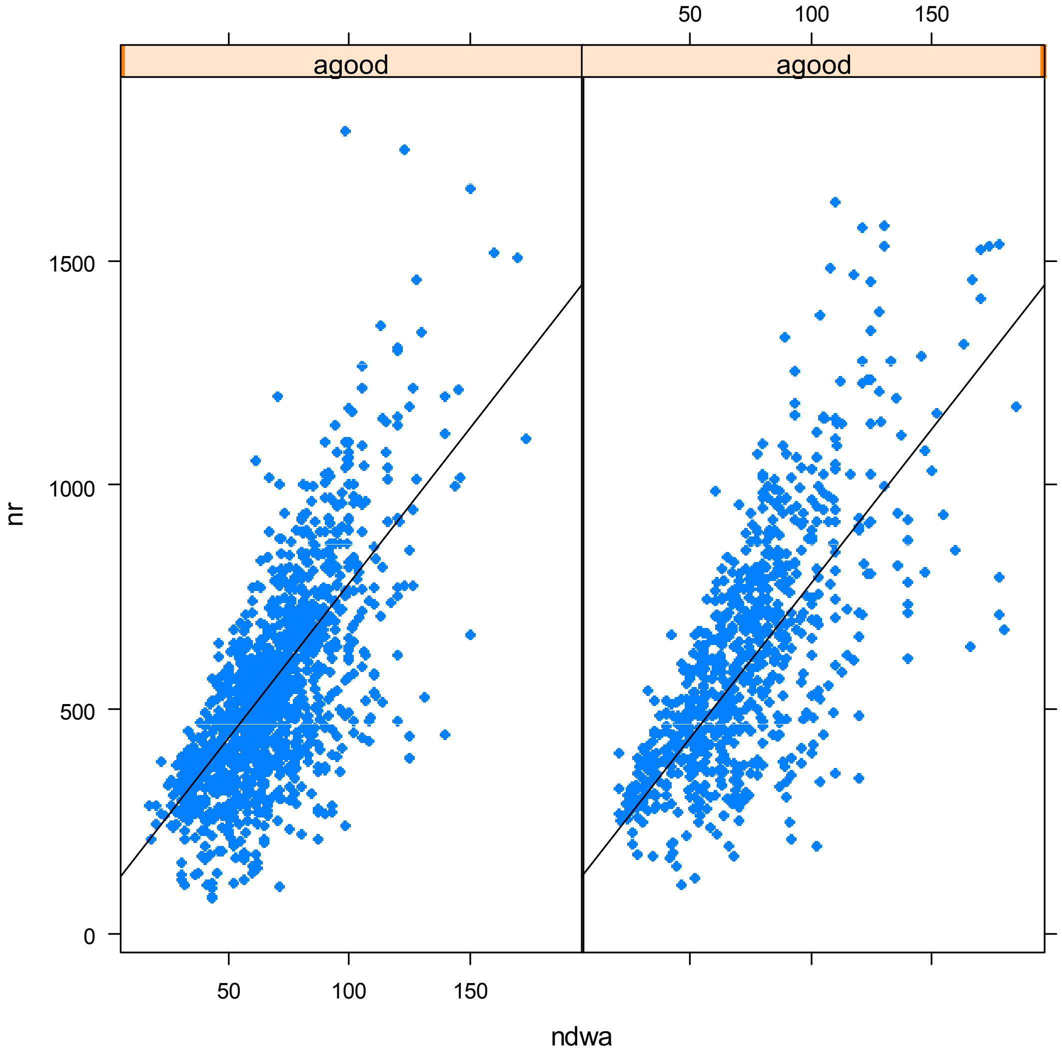
$$t_{d.f.} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

$$t_{d.f.} = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.7074627 * \sqrt{2053-2}}{\sqrt{1-0.7074627^2}} = 45.33359$$

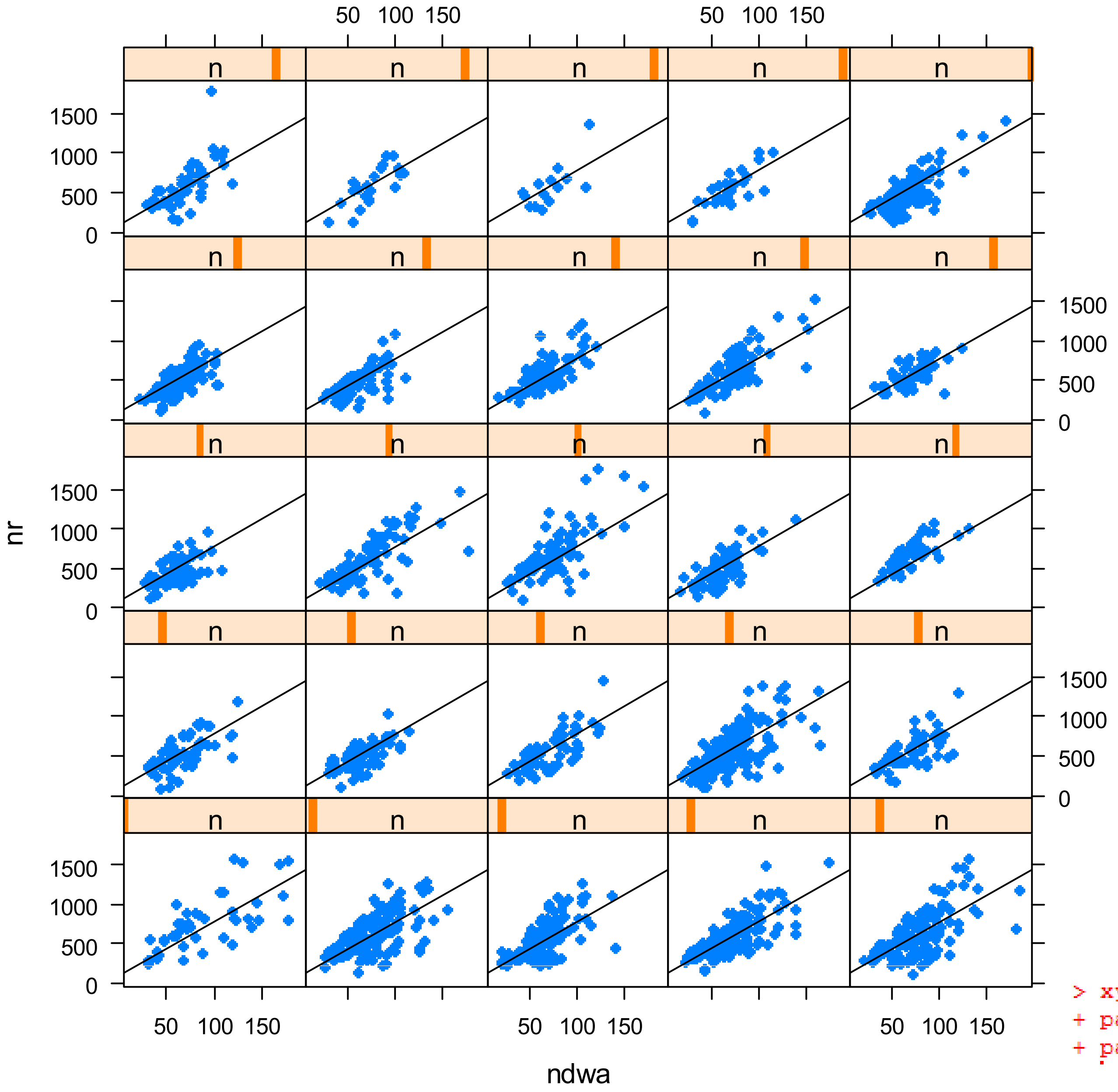
```
> #Spearman
> cor(nr,ndwa,method="spearman") #0.6970837
[1] 0.6970837
> cor.test(nr,ndwa,method="spearman")

        Spearman's rank correlation rho

data:  nr and ndwa
S = 436855767, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.6970837
```

Correlation Coefficients in R , Example rent.data

```
> library(lattice)
> xyplot(nr~ndwa|agood,panel=function(x,y){
+ panel.xyplot(x,y,pch=16)
+ panel.abline(lm(nr~ndwa))})
```



Correlation Coefficients in R , Example rent.data




```
> xyplot(nr~ndwa|n,panel=function(x,y){
+ panel.xyplot(x,y,pch=16)
+ panel.abline(lm(nr~ndwa))})
```

Regression Analysis

- Regression analysis is a statistical technique for analysing the relationships between two or more variables.

Simple linear regression:

- Simple: It uses only one predictor variable (X) to describe the behaviour of dependent variable (Y).
- Linear: It assumes a linear relationship between X and Y (in the coefficients).

$$Y = \beta_0 + \beta_1 * X + \varepsilon$$

Not in this course but easily to be implemented in R:

- polynomial regression (often used to test for non-linearity in a relationship)
- piecewise regression (two or more adjacent straight lines)
- robust regression (models that are less sensitive to outliers)
- multiple regression (numerous explanatory variables)

Simple Linear Regression

- Regression line (or curve): A graphical display of a regression model, usually with the response/dependent variable y on the ordinate and the regressor x on the abscissa.
- Regressor variable: The independent or predictor variable in a regression model.
- Regression coefficient(s): The parameter(s) in a regression model.
- Notation:

$$S_{XX} = \sum x_i^2 f_i - \frac{(\sum x_i f_i)^2}{n} \quad S_{YY} = \sum y_j^2 f_j - \frac{(\sum y_j f_j)^2}{n} \quad S_{XY} = \sum x_i y_j f_{ij} - \frac{(\sum x_i f_i)(\sum y_j f_j)}{n}$$

$$s_X{}^2 = \frac{1}{n-1} * S_{XX} \quad s_Y{}^2 = \frac{1}{n-1} * S_{YY} \quad s_{XY} = \frac{1}{n-1} * S_{XY}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}} = \frac{s_{xy}}{s_x * s_y} = \frac{cov(x,y)}{s_x * s_y}$$

$$\boxed{\hat{y} = a_y + b_y * x}$$

$$\boxed{b_y = \frac{S_{XY}}{S_{XX}} = \frac{cov(x,y)}{s_x^2}} \quad \boxed{a_y = \overline{y} - b_y * \overline{x}}$$

$$\boxed{\hat{x} = a_x + b_x * y}$$

$$\boxed{b_x = \frac{S_{XY}}{S_{YY}} = \frac{cov(x,y)}{s_Y^2} \quad a_x = \overline{x} - b_x * \overline{y}}$$

Coefficient of Determination

- In statistics, the coefficient of determination, R^2, is the proportion of variability in a data set that is accounted for by a statistical model.
- For linear regression: R^2 is the square of a correlation coefficient (Pearson).

$$R^2 = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2}$$

- Numerator: deviation of y_i from the mean that is “explained” by regression line
- Hence, the coefficient of determination can be defined as the quotient of the (squared) “declared” (computed/estimated) deviations and the observed deviations from the mean.

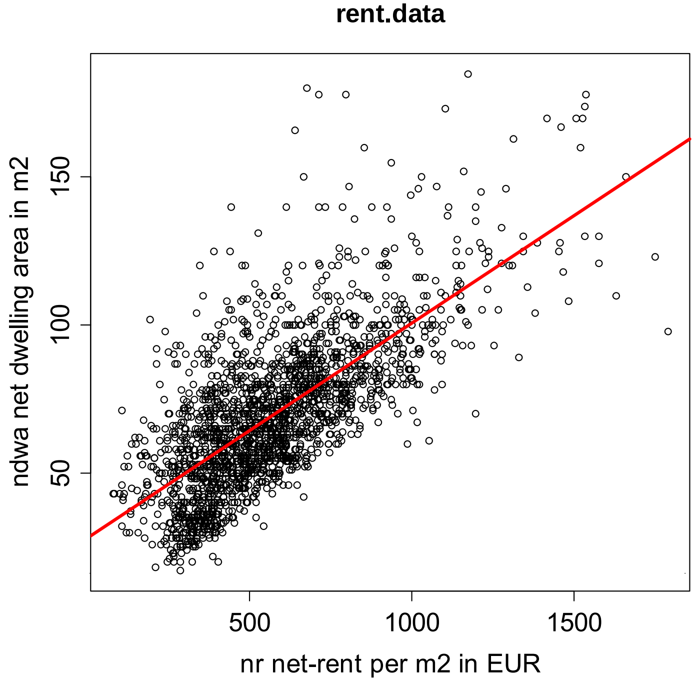
$0 < R^2 < 1$, $R^2 = 1$ if all observed values y_i lie exactly on the regression line. The more the observed values differ from the regression line, the lower R^2 is.

Linear Regression - Conditions

- Response variable: interval scale, normally distributed
- Independent variables: interval scale & normally distributed or dichotomous (coded as dummy variables)
- Independence of observations
- Linear relationship can be assumed.
- The variances of the y-values for one x-value should be the same over the whole range of x-values (homoscedasticity).

Linear Regression in R – Example rent.data

```
> rent.data<-read.table("rent.asc", header=T, sep="\t")
> attach(rent.data)
> plot(nr,ndwa,main="rent.data",xlab="nr net-rent per m2 in EUR",
+ ylab="ndwa net dwelling area in m2",cex.axis=1.5,cex.main=1.5,cex.lab=1.5)
> str(rent.data)
'data.frame':   2053 obs. of  13 variables:
 $ nr        : num  741 716 528 554 698 ...
 $ nrsqm     : num  10.9 11.01 8.38 8.52 6.98 ...
 $ ndwa      : int  68 65 63 65 100 81 55 79 52 77 ...
 $ rooms     : int  2 2 3 3 4 4 2 3 1 3 ...
 $ yc        : num  1918 1995 1918 1983 1995 ...
 $ n         : int  2 2 2 16 16 16 6 6 6 6 ...
 $ agood     : int  1 1 1 0 1 0 0 0 0 0 ...
 $ abest     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ hw        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ ch        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ tb        : int  0 0 0 0 0 0 0 0 0 0 ...
 $ bathextra: int  0 0 0 1 1 0 1 0 0 0 ...
 $ kextra   : int  0 0 0 0 1 0 0 0 0 0 ...
```




```
> abline(lm(ndwa~nr),col="red",lwd=3)
```

Linear Regression in R – Example rent.data

```
> model<-lm(nr~ndwa)
> summary(model)

Call:
lm(formula = nr ~ ndwa)

Residuals:
     Min       1Q   Median       3Q      Max
-655.178  -97.409    7.404   98.729 1023.448

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  89.8469    11.2644   7.976 2.49e-15 ***
ndwa          6.9006     0.1522  45.334  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 173.5 on 2051 degrees of freedom
Multiple R-squared: 0.5005,     Adjusted R-squared: 0.5003
F-statistic:  2055 on 1 and 2051 DF,  p-value: < 2.2e-16
```

```
var(nr)  #60238.1
var(ndwa)  #633.1543
var(nr,ndwa)  #4369.12
var(nr,ndwa)/sqrt(var(nr)*var(ndwa))  #0.7074627
cor(nr,ndwa)  #0.7074627
(cor(ndwa,nr)^2)   #coefficient of determination, R-squared, 0.5005034
#50 percent of the (squared) deviations of the observed values
#from the mean can be explained by the regression line.

(b_xy<-var(nr,ndwa)/var(ndwa))  #6.90056
(a_xy<-mean(nr)-b_xy*mean(ndwa))  #89.84691
```