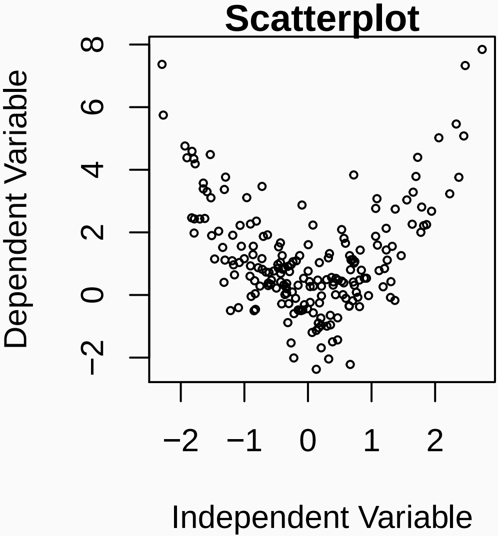
Lecture 3: Bivariate and Multivariate Data

Thomas Chadefaux

PO7001: Quantitative Methods I

Scatterplots

Dependent and Independent Variables



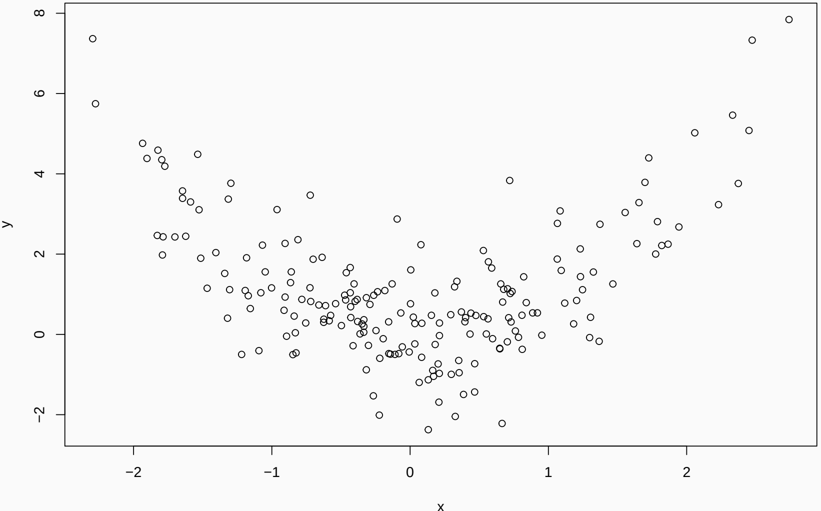
What to look for in a scatterplot

- Overall pattern: up, down, curvy, etc.?
- Strength of that relationship?
- Any major outliers?

Scatterplots in R

In its most basic form, a scatterplot is obtained as:

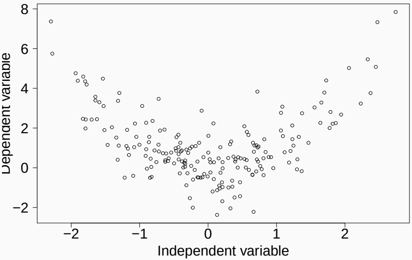
```r
plot(x, y)
```



Scatterplots in R

You can improve the plot by:

```
plot(x, y,
     # a. Labeling the axes
     xlab='Independent variable',
     ylab='Dependent variable',
     # making the axis horizontal
     las = 1,
     #increasing the label and axes sizes
     cex.axis = 2, cex.lab=2)
```

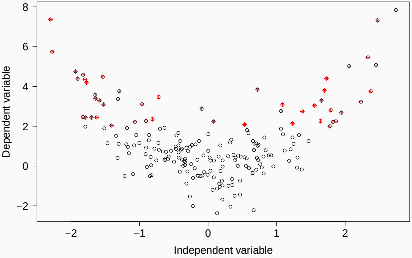


Scatterplots in R

You can color/depict points above a certain value differently

```
plot(x, y,
     xlab='Independent variable',
     ylab='Dependent variable',
     las = 1, cex.axis = 1.5, cex.lab=1.5)

points(x[y>2], y[y>2],
       col='red', #different color
       pch=3#different symbol
       )
```

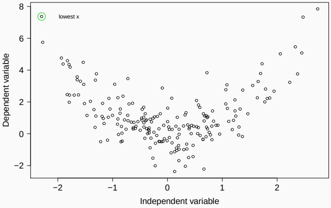


Scatterplots in R

Add some labels

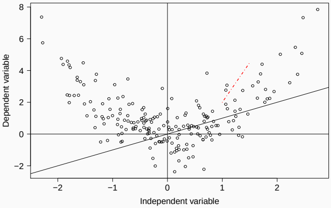
```
plot(x, y,
     xlab='Independent variable',
     ylab='Dependent variable',
     las = 1, cex.axis = 1.5, cex.lab=1.5)

#find upper left point
toPlot <- which(x==min(x))
# add text next to it
text(x[toPlot] + 0.5, y[toPlot], 'lowest x')
#draw a big circle around it
points(x[toPlot], y[toPlot], cex=3, col=3)
```



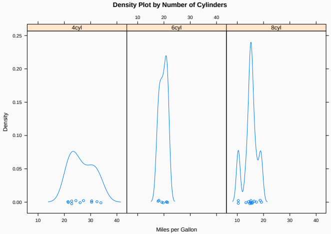
Draw lines

```
plot(x, y,
     xlab='Independent variable',
     ylab='Dependent variable',
     las = 1, cex.axis = 1.5, cex.lab=1.5)
library(fields)
xline(0) # Vertical line
yline(0) # Horizontal line
abline(0,1) # "slope"
segments(1,2,1.5,4.5, col=2, lwd=2, lty=4)
```



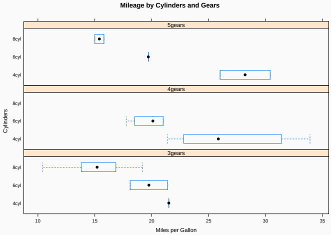
Lattice Plots

```
library(lattice)
attach(mtcars)
# create factors with value labels
gear.f<-factor(gear,levels=c(3,4,5), labels=c("3gears","4gears","5gears"))
cyl.f <-factor(cyl,levels=c(4,6,8),labels=c("4cyl","6cyl","8cyl"))
# kernel density plots by factor level
densityplot(~mpg|cyl.f,
       main="Density Plot by Number of Cylinders",
   xlab="Miles per Gallon")
```



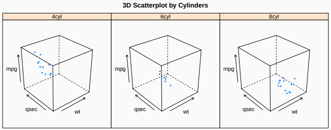
Scatterplots in R

```
# boxplots for each combination of two factors
bwplot(cyl.f~mpg|gear.f,
    ylab="Cylinders", xlab="Miles per Gallon",
   main="Mileage by Cylinders and Gears",
   layout=(c(1,3)))
```



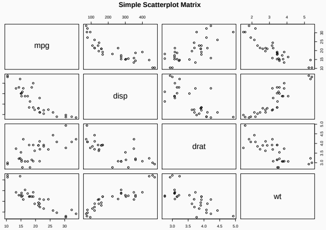
Scatterplots in R

```r
# 3d scatterplot by factor level
cloud(mpg~wt*qsec|cyl.f,
      main="3D Scatterplot by Cylinders")
```



Scatterplots in R

```r
# Basic Scatterplot Matrix
pairs(~mpg+disp+drat+wt,data=mtcars,
   main="Simple Scatterplot Matrix")
```



Covariance and Correlation

Defining the covariance

Remember that the variance is defined as

$$Var(x) = \sigma^2 = \frac{\sum_i (x_i - \bar{x})^2}{N-1},$$

which can easily be rewritten as

$$Var(x) = \sigma^2 = \frac{\sum_i (x_i - \bar{x})(x_i - \bar{x})}{N-1}.$$

The variance measures how much a variable deviates "from itself".

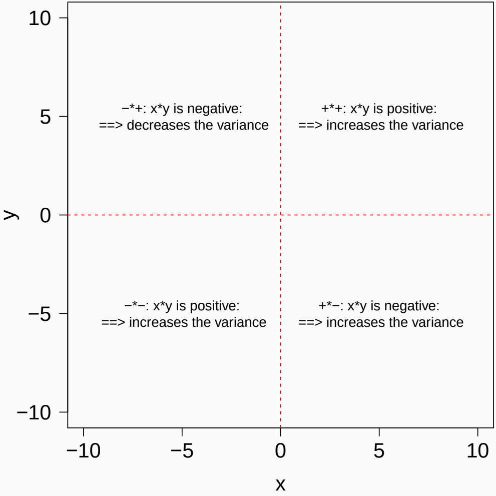
The covariance measures how two variables co-vary. I.e., how they move together (or not). It is defined in very much the same way as the variance:

$$Cov(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{N-1}.$$

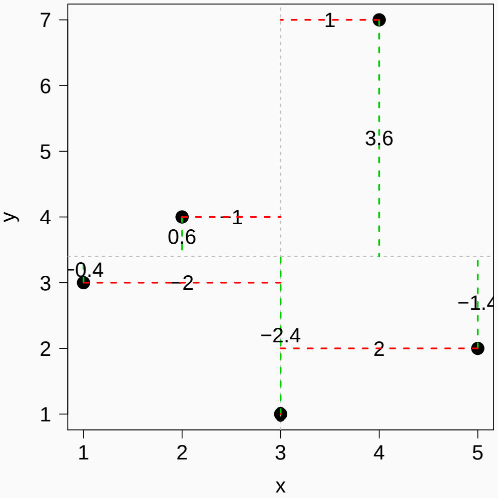
Notice how the *variance* of x was really the covariance of x with itself. I.e.,

$$Var(x) = Cov(x, x)$$

Intuition for the covariance



Intuition for the covariance



Covariance: An example

x	y	$x-\bar{x}$	$y-\bar{x}$	$(x-\bar{x})(y-\bar{x})$
1	4	1-3=-2	4-4=0	0
2	3	2-3=-1	3-4=-1	1
3	2	3-3=0	2-4 = -2	0
4	6	4-3=1	6-4=2	2
5	5	5-3=2	5-4=1	2
				sum = 5

So the covariance is $5/5 = 1$. But actually we were supposed to divide by N-1, so $5/4 = 1.25$. Let's check

```
x <- c(1:5)
y <- c(4, 3, 2, 6, 5)
cov(x, y)
```

```
## [1] 1.25
```

Covariance

The only problem with the covariance is that it has no 'natural' scale. If I double the size of every value x and y, the covariance increases, even though the linear relationship is the same. If I add points, the covariance changes. As a result, you cannot easily compare the covariance of two different samples

```
x <- c(1,2,3,4)
y <- c(4,3,2,6)
cov(x,y)
```

```
## [1] 0.8333333
```

```
cov(2*x, 2*y)
```

```
## [1] 3.333333
```

Correlation: Pearson's r

- Because the covariance has no boundaries or natural unit, the correlation r, or Pearson's correlation coefficient, is a normalized variance. It is defined as

$$r = \frac{cov(x, y)}{\sigma_x \sigma_y}$$

- Note that this is just the covariance devided by the standard deviations of x and y. Doing this normalizes the covariance to a number between -1 and 1. A negative number means that x and y are negatively correlated (a negative slope).

Correlation: Pearson's r

In R, all you need is 'cor(x,y)' Note that the correlation coefficient is not affected by the units.

```r
x <- c(1,2,3,6)
y <- c(3,4,2,5)
cor(x,y)
```

```
## [1] 0.5976143
```

```r
cor(2*x, 2*y)
```

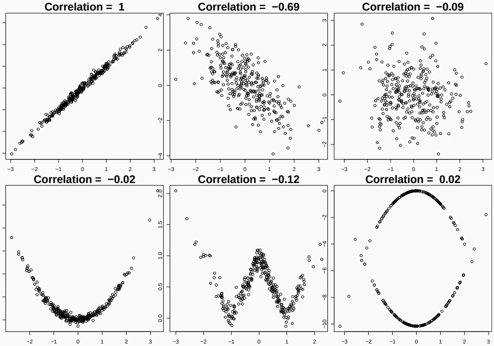
```
## [1] 0.5976143
```

```r
# BUT a non-linear change does affect the correlation
cor(x^2, y^2)
```

```
## [1] 0.7632795
```

Correlation

Some examples:



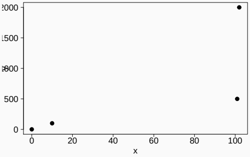
Correlation: Spearman's rank correlation coefficient

Sometimes we have non-linear data, or ordinal data, for which Pearson's r is not well suited. An alternative is Spearman's rank correlation coefficient, which is the same as Pearson' r but uses the correlation of the *ranks* rather than the correlation of the values. I.e.,

$$r_s = r_{rank_x, rank_y}$$

Correlation: Pearson vs Spearman: an example

```r
x <- c(0, 10, 101, 102)
y <- c(1, 100, 500, 2000)
plot(x, y, cex=2, pch=19, cex.axis=2, cex.lab=2, las=1)
```




```r
cor(x, y, method = 'pearson')
```

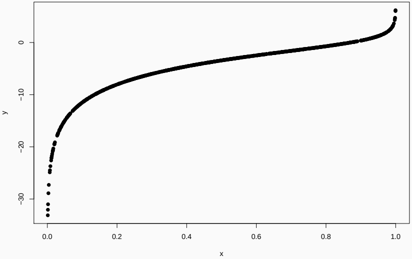
```
## [1] 0.7544237
```

```r
cor(x, y, method = 'spearman')
```

```
## [1] 1
```

Correlation: Pearson vs Spearman: another example

```
x <- runif(1000)
y <- log(x^5/(1-x^5))
plot(x, y, pch=19)
```




```
cor(x, y, method = 'pearson')
```

```
## [1] 0.9027593
```

```
cor(x, y, method = 'spearman')
```

```
## [1] 1
```

Least square regression

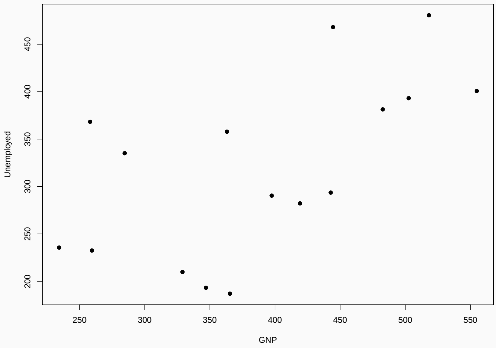
Least square regression

NB: this is just a brief introduction to the concept—not a full treatment, which will be discussed later in the module.

Least square regression

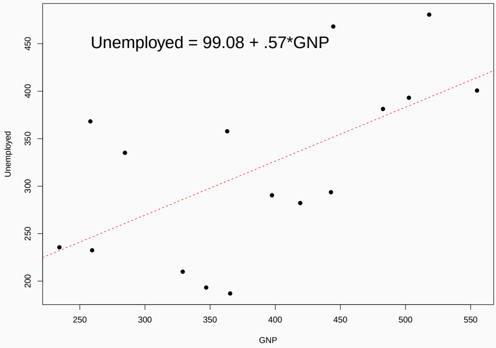
An example:

```
par(mfrow=c(1,1), mar=c(4,4,1,1))
attach(longley)
plot(GNP, Unemployed, pch=19)
```



Least square regression

An example:



Least square regression: interpretation

- An increase in GNP increases the expected number of the unemployed
- For every one unit increase in GNP, the expected number of unemployed increases by .57 units.
- For a GNP of 400, the *expected* number of unemployed would be: $99.08 + 0.57 \times 400 = 327.08$

Least square regression: In R

```r
x <- c(1,2,3,4)
y <- c(15,12,16,25)
lm(y ~ x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##         8.5          3.4
```

Least square regression: The model

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

- y_i is your dependent variable, and x_i is your independent variable. They are fixed and are given to you (or you collected them, but either way they are not to be estimated).
- α and β are *parameters*. They are what you don't know and want to estimate.
  - α is the intercept—the value of y when x is 0
  - β is the slope, or how much does y increase when x increases by 1 unit.
- Because the relationship is never a perfect straight line, there are deviations from the line. We call these deviations the *error term* or *residual*.

Least square regression: Estimating the coefficients

We want to find the values of α and β that fit the data best. By fit, we mean that we want to minimize the deviations from the line, and more specifically we are going to minimize the sum of squared errors. The least square model does just that: it returns the line such that the sum of squared deviations from the line is minimized.

- One solution is to just try out all possible values until you find the best one.
- A better solution is to use the formulae:

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

  and

$$a = \bar{y} - b\bar{x}$$

  You'll learn *a lot more* about this and where the equation comes from throughout this course.

Least square regression: Estimating the coefficients—an example

x <- c(1,2,3,4) y <- c(15,12,16,25)

$$b = \frac{(1-2.5)(15-17)+(2-2.5)(12-17)+(3-2.5)(16-17)+(4-}{(1-2.5)^2+(2-2.5)^2+(3-2.5)^2+(4-2.5)^2)}$$

$$= \frac{(3+2.5-0.5+12)}{2.25+0.25+0.25+2.25}$$

$$= \frac{17}{5} = 3.4$$

and $a = \bar{y} - b\bar{x} = 17 - 3.4 \times 2.5 = 8.5$

Let's check with R:

```r
lm(y~x)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##         8.5          3.4
```

Model fit: R^2

- Also called coefficient of determination
- r^2 is the proportion of variation in Y determined by variation in X.
- $0 \leq r^2 \leq 1$
- $r^2 = 1 - \frac{SS_{res}}{SS_{tot}}$

$SS_{res} = \sum_i e_i^2$ is the sum of squared residuals (the blue squares in the plot below). $SS_{tot} = \sum_i (y_i - \bar{y})^2$ is the total sum of squares (red in plot below)