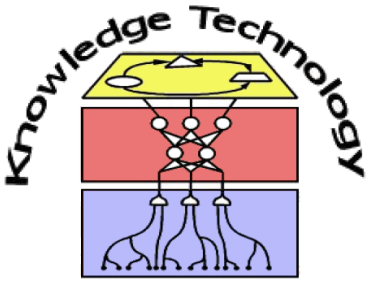
Research Methods

EDA with multiple variables

Dr. Sven Magg, Prof. Dr. Stefan Wermter



http://www.informatik.uni-hamburg.de/WTM/

Plan for today!

1. Multivariate EDA
2. Joint distribution of categorical data
   - Visualisation (contingency tables)
   - Statistics (chi-square)
3. Joint distribution of continuous data
   - Visualisation (scatterplots, line fitting)
   - Statistics (covariance, correlation coefficients)

Multivariate EDA

- Looking at one variable, we can
  - see that something influences this data
  - identify interesting areas

- Usually we are interested in how variables affect the outcome

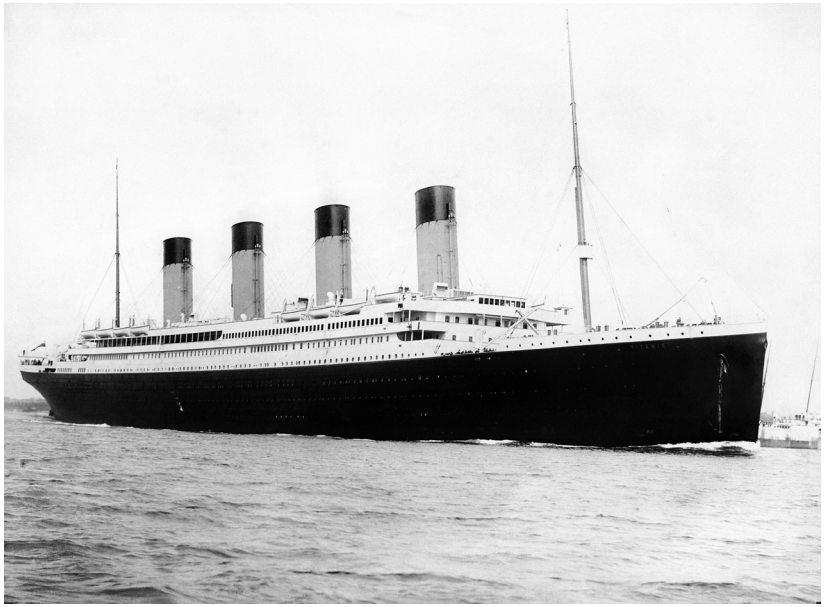
$$y = f(x, \varepsilon)$$

- OR: How do variables affect each other?

Joint distributions

- How can we see that variables influence each other?

- Data from the Titanic:
  - 2201 Passengers: 470 Female, 1731 Male
  - 711 Survivors, 1490 Dead



- Did your gender influence your survival rate?
  - Data is categorical

Contingency Tables

	Dead	Survived	
Female	126	344	470
Male	1364	367	1731
	1490	711	2201

- Females seem to have had a higher chance of survival!
- Make this more clear by dividing by marginal count

Division by row margins:

	Dead	Survived	
Female	26,81%	73,19%	470
Male	78,80%	21,20%	1731
	1490	711	2201

Division by column margins:

	Dead	Survived	
Female	8,46%	48,38%	470
Male	91,54%	51,62%	1731
	1490	711	2201

Contingency Tables

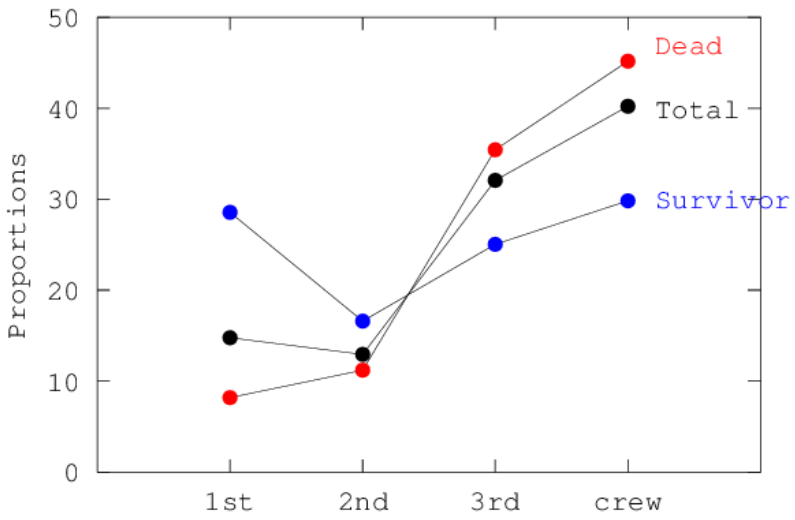
- Show how values of one variable are contingent on values of another variable
- Also called cross-classification tables
- Remember: They simply represent joint distributions, not causal relations!

	Dead	Survived	
1.Class	122	203	325
2.Class	167	118	285
3.Class	528	178	706
Crew	673	212	885
	1490	711	2201

Contingency Tables

Survival by class and gender (column proportions):

	Dead	Survived		
1.Class	8,19%	28,55%	325	14,8%
2.Class	11,21%	16,60%	285	12,9%
3.Class	35,44%	25,04%	706	32,1%
Crew	45,17%	29,82%	885	40,2%
	1490	711	2201	100%

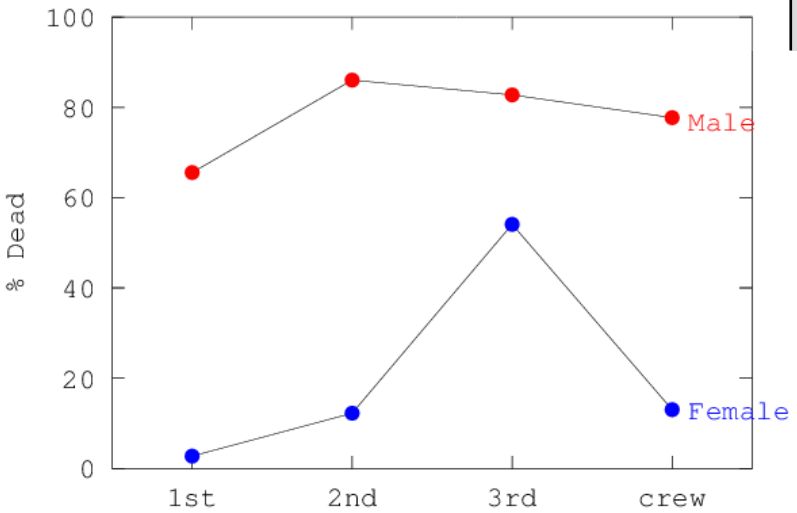


- Plot column/row proportions to visualise dependencies
- Clear dependency between status and death rate
- Combination of gender and status?

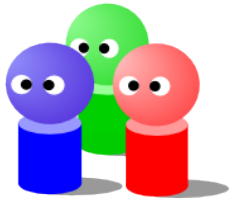
Contingency Tables

Survival by class and gender (row proportions):

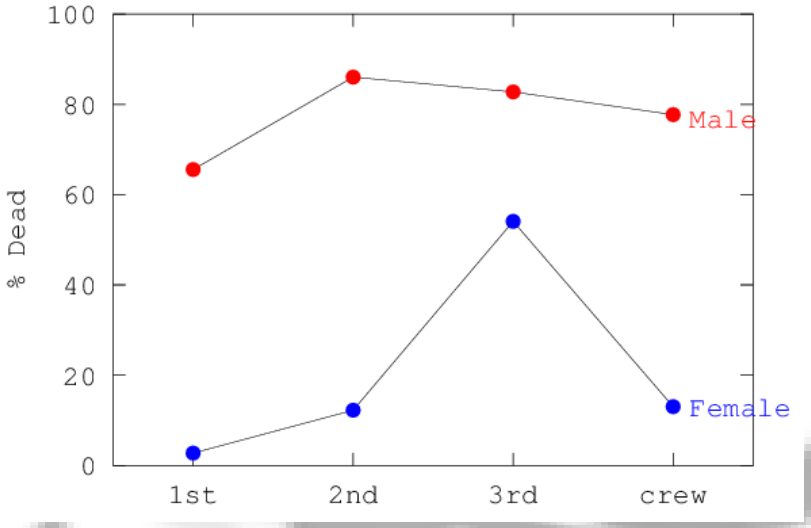
		Dead	Survived	
Male	**1.Class**	118 (65.6%)	62 (34.4%)	180
	2.Class	154 (86,0%)	25 (14.0%)	179
	3.Class	422 (82.7%)	88 (17.3%)	510
	Crew	670 (77.7%)	192 (22.3%)	862
Female	**1.Class**	4 (2.8%)	141 (97.2%)	145
	2.Class	13 (12.3%)	93 (87.7%)	106
	3.Class	106 (54.1%)	90 (45.9%)	196
	Crew	3 (13.0%)	20 (87.0%)	23
		1490	711	2201



Group Task!



2 5



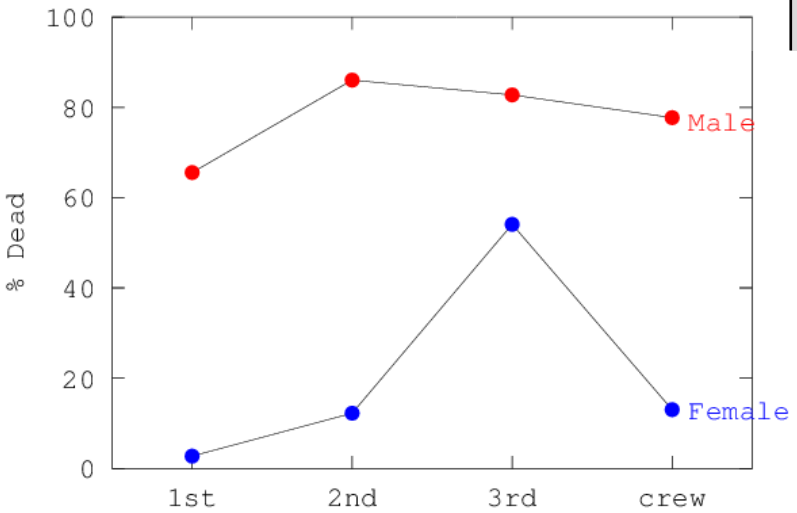
What would it mean if

1. **the lines would be parallel?**
2. **the lines would be identical**

Contingency Tables

Survival by class and gender (row proportions):

		Dead	Survived	
Male	**1.Class**	118 (65.6%)	62 (34.4%)	180
	2.Class	154 (86,0%)	25 (14.0%)	179
	3.Class	422 (82.7%)	88 (17.3%)	510
	Crew	670 (77.7%)	192 (22.3%)	862
Female	**1.Class**	4 (2.8%)	141 (97.2%)	145
	2.Class	13 (12.3%)	93 (87.7%)	106
	3.Class	106 (54.1%)	90 (45.9%)	196
	Crew	3 (13.0%)	20 (87.0%)	23
		1490	711	2201

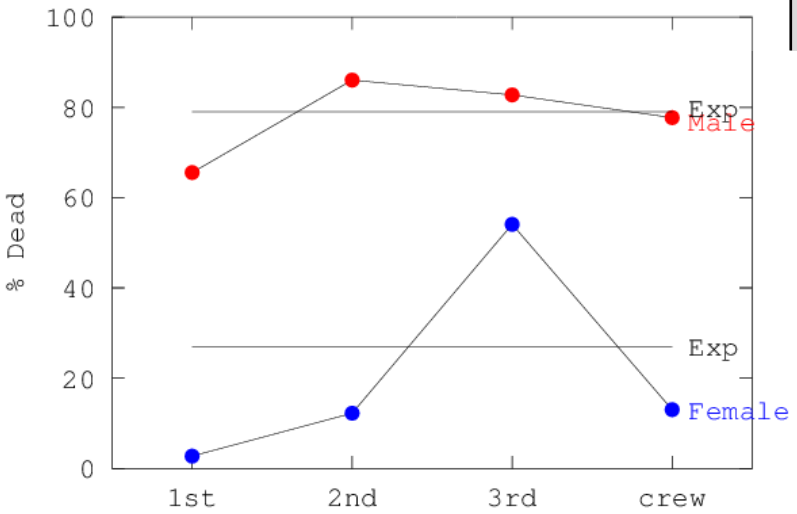


- The effect of status on survival depends on gender
- Quantifying this by differences? Male – Female: (63%, 74%, 28%, 65%)

Contingency Tables

Survival by class and gender (row proportions):

		Dead	Survived	
Male	**1.Class**	118 (65.6%)	62 (34.4%)	180
	2.Class	154 (86,0%)	25 (14.0%)	179
	3.Class	422 (82.7%)	88 (17.3%)	510
	Crew	670 (77.7%)	192 (22.3%)	862
Female	**1.Class**	4 (2.8%)	141 (97.2%)	145
	2.Class	13 (12.3%)	93 (87.7%)	106
	3.Class	106 (54.1%)	90 (45.9%)	196
	Crew	3 (13.0%)	20 (87.0%)	23
		1490	711	2201



- It seems that “female” has a higher impact on class than “male”
- Can we quantify this effect?

Chi-Square χ^2 Statistics

Survival by class and gender (row proportions):

		Dead	Survived	
Male	**1.Class**	118 (65.6%)	62 (34.4%)	180
	2.Class	154 (86,0%)	25 (14.0%)	179
	3.Class	422 (82.7%)	88 (17.3%)	510
	Crew	670 (77.7%)	192 (22.3%)	862
		1364	367	1731

- Cells contain frequencies of conjunct events
- Probability of conjunct event if the events are independent: $\Pr(A\&B) = \Pr(A)\Pr(B)$
- $\Rightarrow \Pr(1.Class\ \&Dead) = \Pr(1.Class)\Pr(Dead)$ if Status and Gender are independent
- Problem: We don't know $\Pr(1.Class)$ or $\Pr(Dead)$

Chi-Square χ^2 Statistics

Survival by class and gender (row proportions):

		Dead	Survived	
Male	**1.Class**	118 (65.6%)	62 (34.4%)	180
	2.Class	154 (86,0%)	25 (14.0%)	179
	3.Class	422 (82.7%)	88 (17.3%)	510
	Crew	670 (77.7%)	192 (22.3%)	862
		1364	367	1731

- Estimate probabilities from margins:
- $Pr(Status = 1.Class) = 180/1731 = 0.104$
- $\Pr(Outcome = Dead) = 1364/1731 = 0.788$
- $\Pr(Status = 1.Class \ \& Outcome = Dead) = \frac{180*1364}{1731} = 0.082$
- Now we can calculate the expected frequency of the conjunct event: $f_e = 0.082 * 1731 = 141.84$

Chi-Square χ^2 Statistics

Observed frequencies f_o and expected frequencies f_e:

		Dead	Survived	
Male	**1.Class**	118 (142)	62 (38)	180
	2.Class	154 (141)	25 (38)	179
	3.Class	422 (402)	88 (108)	510
	Crew	670 (679)	192 (183)	862
		1364	367	1731

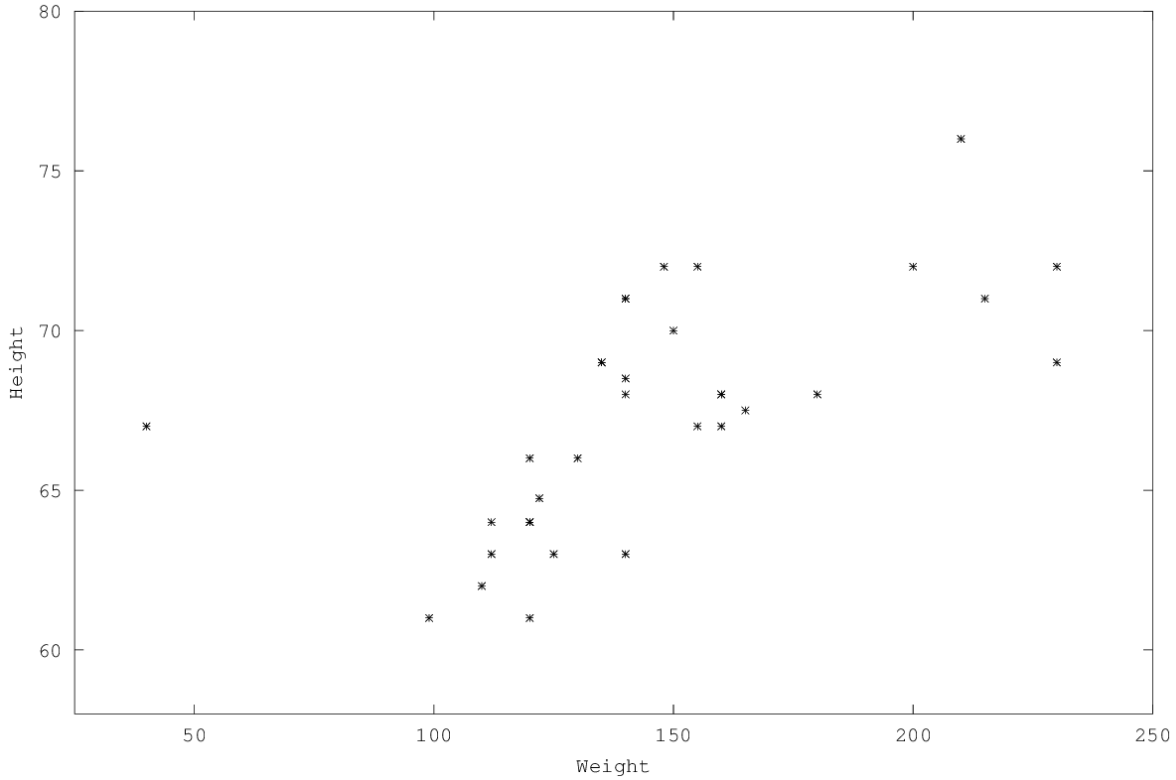
Chi-Square χ^2

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

- $\chi^2{}_{male} = 29.852$
- $\chi^2{}_{female} = 130.69$
- χ^2 assumes that row and column variables are independent and is a measure of difference between observed and expected frequencies
- What do large values of χ^2 mean?

Continuous Variables

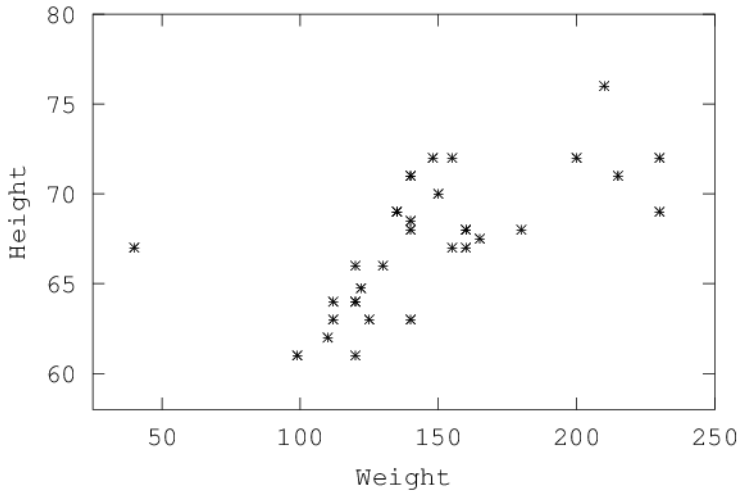
- We could use Bins and contingency tables…
- Better to use human pattern recognition abilities!

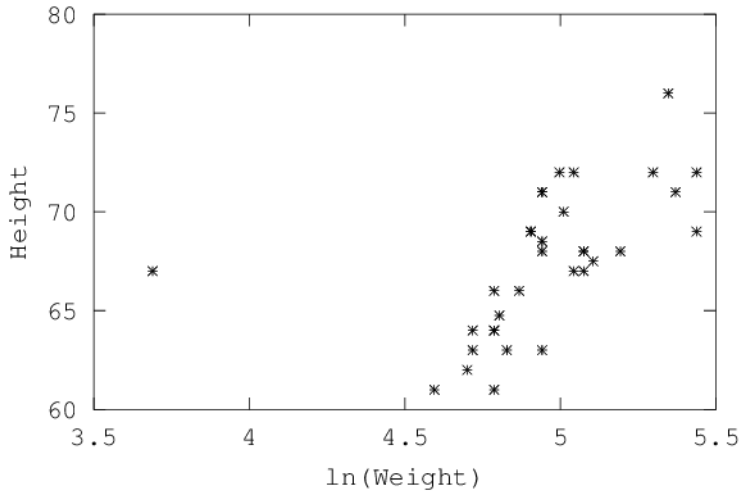


Heights and weights of 33 students

Scatterplots

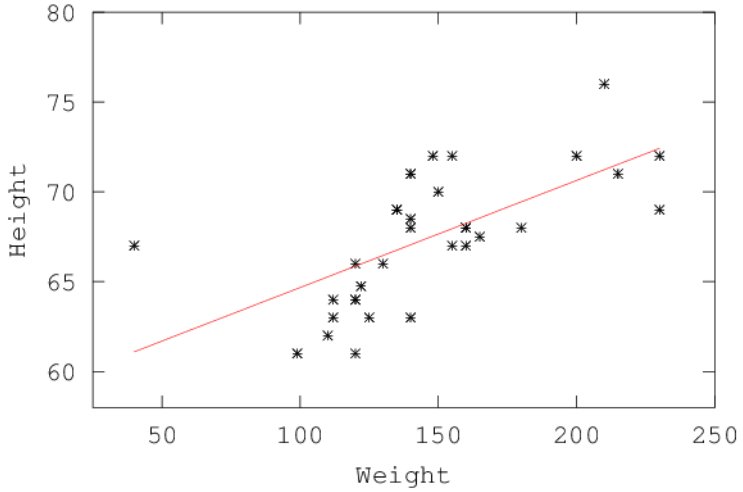
- How can I see independency?
  - Horizontal/Vertical lines
  - Uniformly/Randomly distributed
- We can change the axis (data transformation)

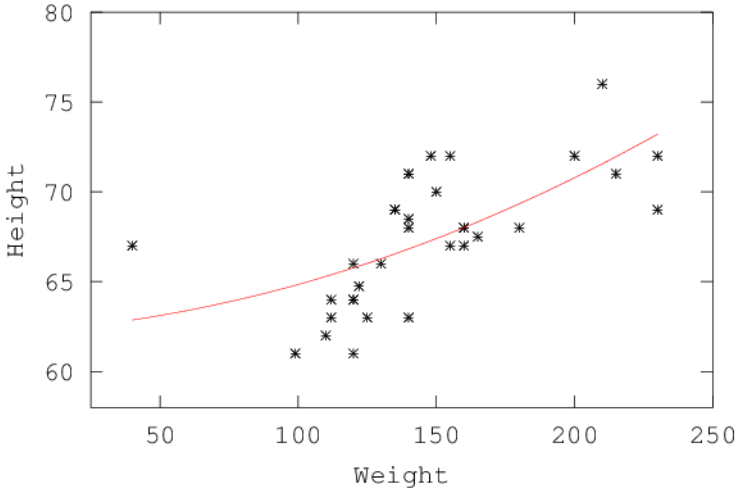




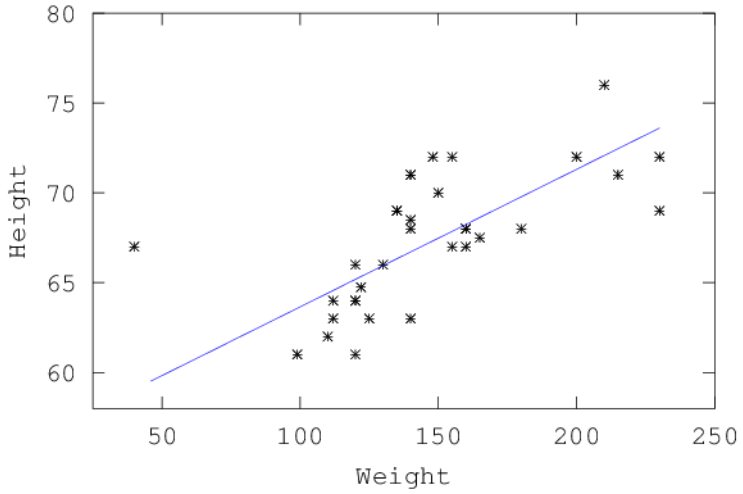
- Plot seems to suggest a nonlinear dependency

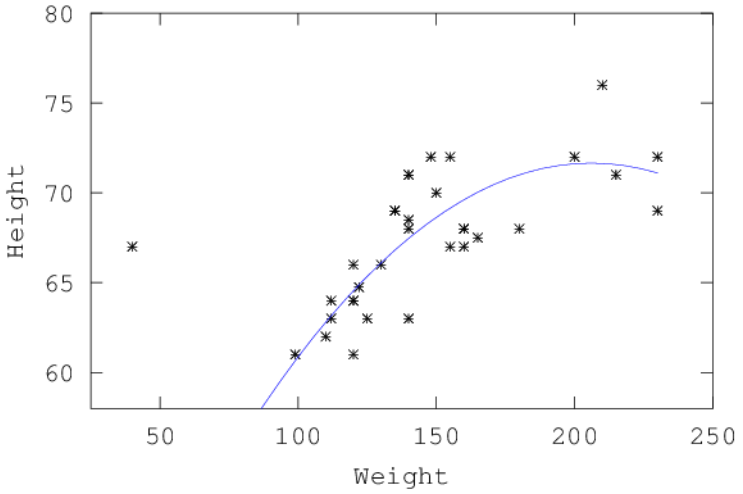
Fitting lines to scatters





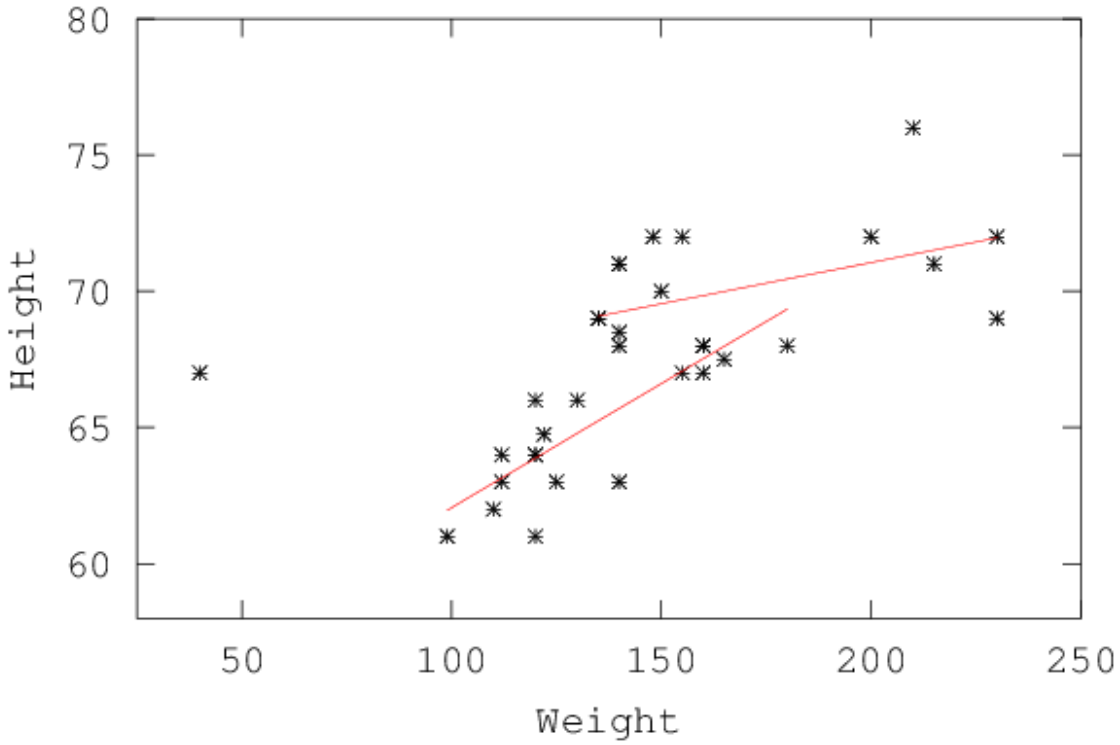
Or ignoring the lower outlier:





Fitting lines to scatters

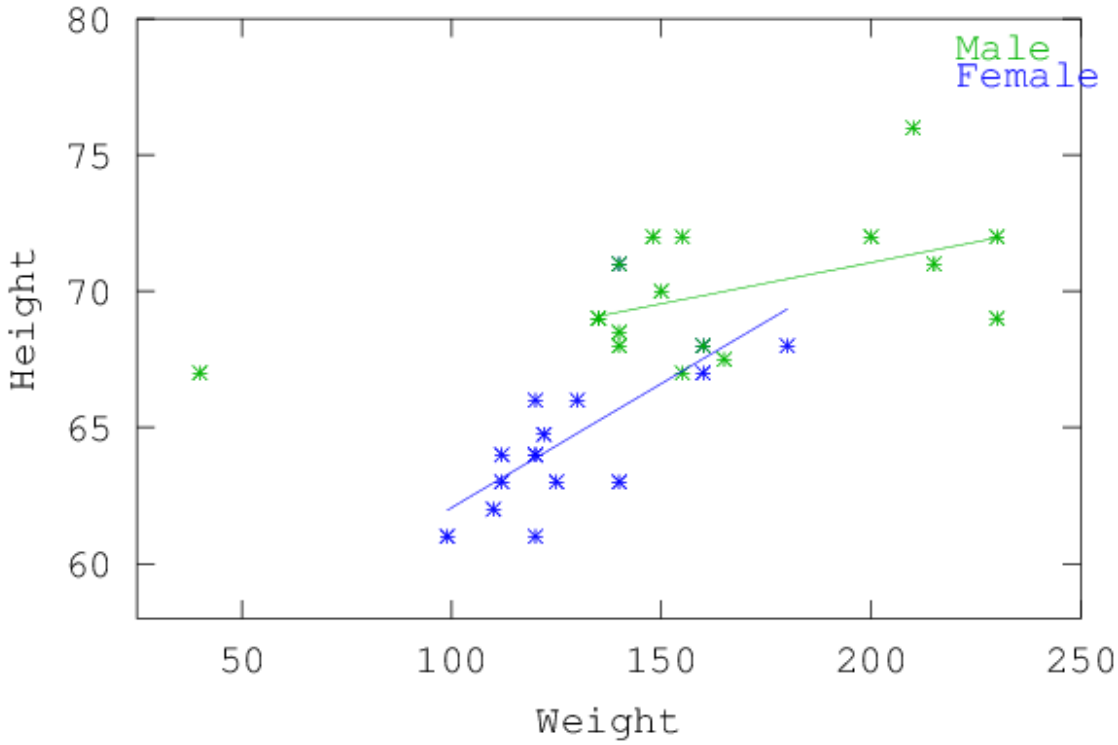
- Also possible piece-wise linear fits:



- Piece-wise fits hint at subgroups within the data

Fitting lines to scatters

- Colour scatterplots to reveal subgroup memberships



Sample Covariance

- Sample variance captures the average amount a value deviates from the mean

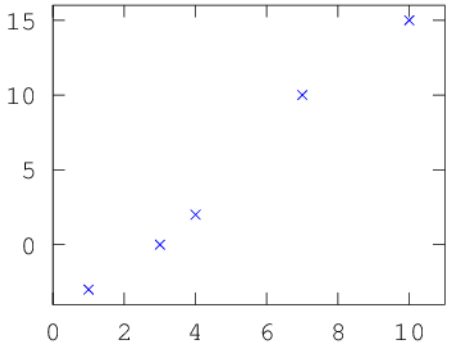
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

- With two variables that have a linear association, we expect:
  - x deviates from the mean $\Rightarrow$ y deviates from the mean in the same direction (positive linear association)
  - x deviates from the mean $\Rightarrow$ y deviates from the mean in the opposite direction (negative linear association)
- We can use the cross-product deviations

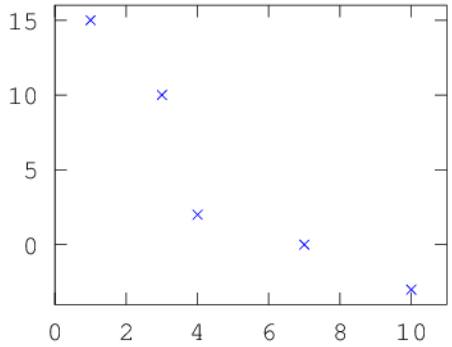
Sample Covariance

Covariance

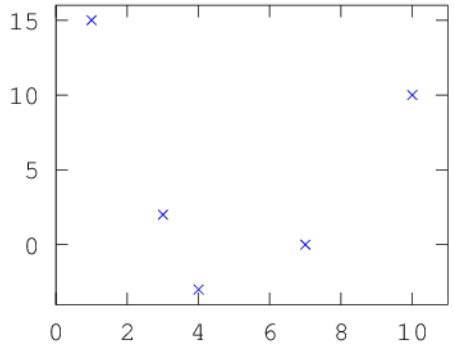
$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y})}{n-1}$$



26.25



-24.25



-2.75

Correlation Coefficients

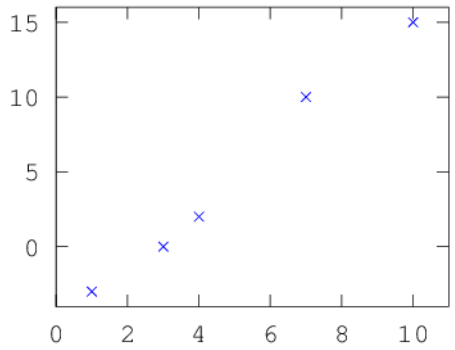
- Covariance can be an arbitrarily large number depending on the data
- Standardised covariance: Correlation coefficient

Pearson's Correlation Coefficient

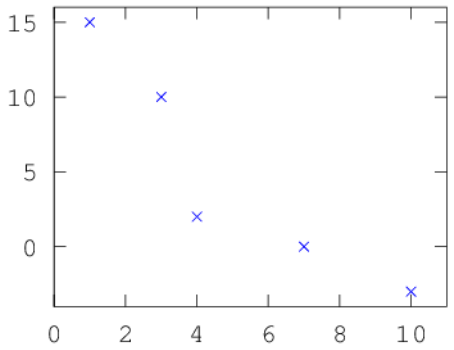
$$r_{XY} = \frac{cov(x,y)}{s_X s_Y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_X s_Y}$$

- Dividing by $s_X s_Y$ leads to values between -1 and 1

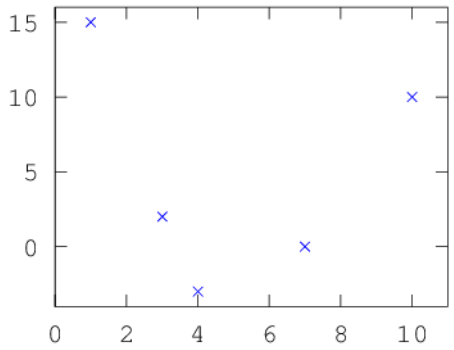
Pearson‘s Correlation Coefficient



0.995



-0.919



-0.104

- 1.0: Perfect positive correlation
- -1.0: Perfect negative correlation
- Values near 0.0 do not mean x and y are not associated, just not linearly associated!
- Only needs data on an interval scale

What have we learned?

1. Visualisations of joint distributions to utilise the human pattern recognition mechanism
2. Contingency tables for categorical data
3. Chi-Square measure to test independence
4. Scatterplots helpful to fit lines and find associations between variables
5. Colour scatterplots to find subgroups and influences
6. Many different fits possible with different interpretations!
7. Covariance and correlation coefficients quantify linear associations

Homework!

1. **In the RM CommSy, you can find data on a survey amongst students of a statistics lecture**
2. **Perform Exploratory Data Analysis to find unexpected results and associations between variables**
3. Get used to Octave/MatLab (or other package)
4. Example scripts can be found together with the lecture slides

Deadlines: **- November 12, 12:00 noon, Report with Graphs**

- November 13, Discussion after lecture