Semester 2, 2012 (Last adjustments: August 16, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: August 16, 2012)

Monday, 30th July 2012

Lecture 1 - Content

- □ **Outline of MATH1905**
- □ **First definitions**
- □ **Types of variables**
- □ **A very short introduction to R**
- □ **Visualizing data**

See Phipps & Quine Chapter 1, Sections 1.1 and 1.2.

Outline of MATH1905

"Alea jacta est - The die has been cast."

Julius Caesar, 10 January 49 BC

- Mathematical problems on games of chance date back to 1494 (Pacioli from Italy): What is the distribution of revenue?

  Definition 1. Statistics is the science of collecting, organizing, interpreting and reporting data.

- Probability (theory) is the appropriate language of statistics.

Knowledge based on evidence

- The scientific method is about getting knowledge based on (hard) evidence.
- This involves the following steps:
  1. Formulate question
  2. Collect relevant data
  3. Do statistical analysis of data
  4. Draw conclusions

MATH1905 – Statistics (Advanced) will help to get you started.

Unit information sheet:

- Web sites: `www.maths.usyd.edu.au/...`
  - `/u/UG/JM/MATH1905/` (for School of Mathematics and Statistics material) or
  - `/u/jormerod/math1905/loc` (for John Ormerod's material).
- Lectures
- Tutorials
- Assessment
  - Exam
  - Quizzes
  - Assignments
- No textbook

Week-by-week lecture summary

1. Data analysis

- **Week 1:** Stem and leaf plots; relative frequencies and probability; histograms; 5-figure summaries; boxplots; R introduction.
- **Week 2:** Σ notation; sample mean; sample variance; bivariate data; correlation.
  $\Rightarrow$ tutorial classes start
- **Week 3:** Linear regression; residual plots; data analysis using R.

2. Probability

- ☐ **Week 4:** Axioms of probability; Venn diagrams; de Morgan's laws; inclusion-exclusion principle; counting principles; sampling; Bayes rule; independence.
- ☐ **Week 5:** Integer valued random variables; unordered selections; discrete random variables; mean and variance; probability generating functions.
- ☐ **Week 6:** Continuous rv's; mean and variance; standardized rv's; normal rv's.
- ☐ **Week 7:** Independent rv's; sums of independent normal rv's; sampling distributions; central limit theorem; normal approximation to binomial.
  $\Rightarrow$ QUIZ 1 in tutorial classes!

3. Statistical inference

- ☐ **Week 8:** Hypothesis testing; 1-sided and 2-sided test for a proportion; sign test.
- ☐ **Week 9:** Two sample binomial test; one sample Z-test; one sample t-test. $\Rightarrow$ ASSIGNMENT DUE!
- ☐ **Week 10:** Review of Z-test and t-test.
- ☐ **Week 11:** Two sample t-test; confidence intervals; confidence bounds.
- ☐ **Week 12:** χ^2 goodness of fit test. $\Rightarrow$ QUIZ 2 in tutorial classes!

4. Review

- ☐ **Week 13:** Review of data analysis, probability and statistical inference; past exam papers.

Introduction

Definition 2. A population is the set of all possible measurements of interest.

Definition 3. A sample is a subset of measurements from the population.

Definition 4. Data is the collection of measured variables.

Example (Length of words). The first three stanzas of Waltzing Matilda are:

> Once a jolly swagman camped by a billabong Under the shade of a coolibah tree, And he sang as he watched and waited till his billy boiled: "You'll come a-waltzing Matilda, with me."
>
> Waltzing Matilda, waltzing Matilda You'll come a-waltzing Matilda, with me And he sang as he watched and waited till his billy boiled: "You'll come a-waltzing Matilda, with me."
>
> Down came a jumbuck to drink at that billabong. Up jumped the swagman and grabbed him with glee. And he sang as he shoved that jumbuck in his tucker bag: "You'll come a-waltzing Matilda, with me."

(Source: http://en.wikipedia.org/wiki/Waltzing_Matilda)

Having collecting the above data we now consider organizing the above words in terms of the number of letters in each word.

Example (Length of words – continued). Variable character count of the $n = 15$ words in the first two linesof Waltzing Matilda, i.e.,

> Once a jolly swagman camped by a billabong Under the shade of a coolibah tree

is:

$$x_1 = 4,\ x_2 = 1,\ x_3 = 5,\ x_4 = 7,\ x_5 = 6,\ x_6 = 2,\ x_7 = 1,\ x_8 = 9,$$
$$x_9 = 5,\ x_{10} = 3,\ x_{11} = 5,\ x_{12} = 2,\ x_{13} = 1,\ x_{14} = 8,\ x_{15} = 4.$$

Population: all $N = 251$ possible word length counts in the entire poem; sample: the $n = 15$ words.

Types of variables

- **Nominal**: information given is a name, e.g. gender;
- **Ordinal**: the measurements can be naturally ordered, e.g. good, average, bad - or storm of category 1,2,3,4 or 5;
- **Quantitative**, which can be measured and is interpretable on either scale:
  - discrete, i.e. $\in \mathbb{N}$; from counting e.g. character count in previous example;
  - continuous ($\in \mathbb{R}$; e.g. length measurement)

Small and large data sets

- In general observations are denoted by $x_1, x_2, x_3, \ldots, x_n$.
- The sample size $= n$.
- $\boxed{\text{If } n < 30 \Rightarrow n \text{ is small;} \quad \text{if } n \geq 30 \Rightarrow n \text{ is large.}}$
- Other rules of thumb exist, e.g. $n = 25, 50$, or 100, to decide whether or not a data set is small or large.

Ordering observations

It's natural to order values. The ordered list of observations is

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}.$$

Definition 5. The ith smallest observation in a sample $x_1, x_2, x_3, \ldots, x_n$ is denoted by $x_{(i)}$ and is called the ith order statistic, $i = 1, \ldots, n$.

Example (Length of words).

The ordered values for our Waltzing Matilda data are

$$x_{(1)} = 1 \leq x_{(2)} = 1 \leq x_{(3)} = 1 \leq x_{(4)} = 2 \leq \ldots$$
$$\ldots \leq x_{(14)} = 8 \leq x_{(15)} = 9$$

This can be quickly obtained with the software R by executing

```
> x = c(4,1,5,7,6,2,1,9,5,3,5,2,1,8,4)
> sort(x)
 [1] 1 1 1 2 2 3 4 4 5 5 5 6 7 8 9
```

A very short introduction to R

What is R?

- R is a freeware 'clone' of the commercial package S-Plus based on the programming language S; (technically a 'function language'.)
- R can be downloaded (for free) from the R web site:
  http://cran.r-project.org/
- R has many 'inbuilt' mathematical & statistical commands.
- There are versions of R for all common operating systems.
- Reference PDF can be found on the course website.
- Many code examples in lecture/tutorial material.

A first dip into R

Elementary commands are either expressions or assignments.

- An expression is a command to simply display the result of a calculation, which is not retained in the computer's memory
- An assignment passes the result of a calculation to a variable name which is stored (but the result will not necessarily be printed out on the screen).

A simple R session

```
> 1*2 + sqrt(4) - 1/2
[1] 3.5
> x = 3.5
> (x+0.5)^2
[1] 16
```

Stored objects

- All assigned variables (or any other R objects) are stored by the computer until overwritten or explicitly deleted by the command `rm()` (for remove).
- To see what variables are stored, type `ls()` (for list) or `objects()`.

```
> x = 8; x;
[1] 8
> y = 3.1415
> ls()
[1] ".Last.value" "x"           "y"
> rm(x)
> objects()
[1] ".Last.value" "y"
```


Creating vectors in R

The command `c()` (for concatenate) creates R vectors.

```
> x = c(4,1,5,7,6,2,1,9,5,3,5,2,1,8,4)
> x
 [1] 4 1 5 6 2 1 9 5 3 5 2 1 8 4
```

The command `sort(x)` sorts the R vector `x`.

```
> sort(x)   # for sorting vectors
[1] 1 1 1 2 2 3 4 4 5 5 5 6 7 8 9
```

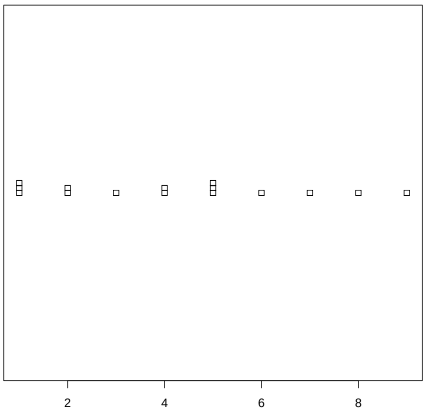
Exit R

```
> q()     # quits R
```

Visualizing data

Strip chart

- For small data sets;
- with `stripchart(x, method="stack")`.



Stem-and-leaf displays

- For small and not too large data sets;
- ordered or unordered; single, double or five stem version;
- with `stem(x, scale=1)`; if you change the `scale` parameter you get more/fewer stems - try `scale=`2^k, for $k = -2, -1, 0, 1, 2, 3$.

```
> stem(x, scale=2)

  The decimal point is at the |

  1 | 000
  2 | 00
  3 | 0
  4 | 00
  5 | 000
  6 | 0
  7 | 0
  8 | 0
  9 | 0
```

```
> stem(x, scale=1)

  The decimal point is at the |

  0 | 000
  2 | 000
  4 | 00000
  6 | 00
  8 | 00

> stem(x, scale=0.5)

  The decimal point is 1 digit(s) to the right of the |

  0 | 11122344
  0 | 5556789
```


Additional material for Lecture 1

General comment

At the end of each lecture I will provide some additional material and background information if appropriate.

More on stem-and-leaf displays

In a stem-and-leaf display all numbers are broken into two components: the stem = the leading digits; the leaf = the remaining digits. The R function `stem(x)` by default produces a stem-and-leaf display with default parameter `scale=1`. This does not mean that a single stem-and-leaf display is produced but rather what some underlying algorithm determines as the most appropriate. In practice you simply start with the default parameter. If you don't like the display either change the scale parameter to `scale=2` or `scale=0.5` or any other power of 2.

```
single stem version                double stem version         five stem version

stem | leafs                       stem | leafs                stem | leafs
   0 | 0 - 9                          0 | 0 - 4                   0 | 0 - 1
   1 | 1 1 2 3 9  (ordered)           0 | 5 - 9                   0 | 2 - 3
   2 | ...                            1 |                         0 | .
   3 | 4 2 3      (unordered)         1 |                         0 | .
   4 |                                .                           0 | 8 - 9
   .                                  .                           1 | 0 - 1
   .                                  .                           .
   .
```

Tuesday, 31st July 2012

Lecture 2 - Content

- □ **Absolute and relative frequencies**
- □ **Ordinate diagrams and histograms**
- □ **Cumulative frequencies and empirical distribution**
- □ **Five number summary**
- □ **Boxplot**

See Phipps & Quine Chapter 1, Sections 1.1, 1.2 and 2.

Absolute and relative frequencies

Example (Gold medals). A total of $n = 55$ countries had at least one olympic gold medal in Beijing 2008. In R absolute frequencies are obtained with `table(x)`, * $=$ AU:

```
x
 1   2   3   4   5   6   7   8   9  13 *14*  16  19  23  36  51
19   9   9   3   2   1   3   1   1   1 * 1*   1   1   1   1   1
```

Let $x_j \in \mathbb{R}$ denote possible measurements, here $j = 1, \ldots, 16$.

Definition 6. The (absolute) frequency with which the value x_j occurred is denoted by f_j.

Definition 7. The relative frequency $:= \widehat{p}_j = \frac{f_j}{n}$.

E.g. $x_{10} = 13 \in \mathbb{R}$ has $f_{10} = 1$ and $\widehat{p}_{10} = \frac{1}{55}$ or $x_1 = 1,\ f_1 = 19,\ \widehat{p}_1 = \frac{19}{55}$.

Ordinate diagrams and histograms

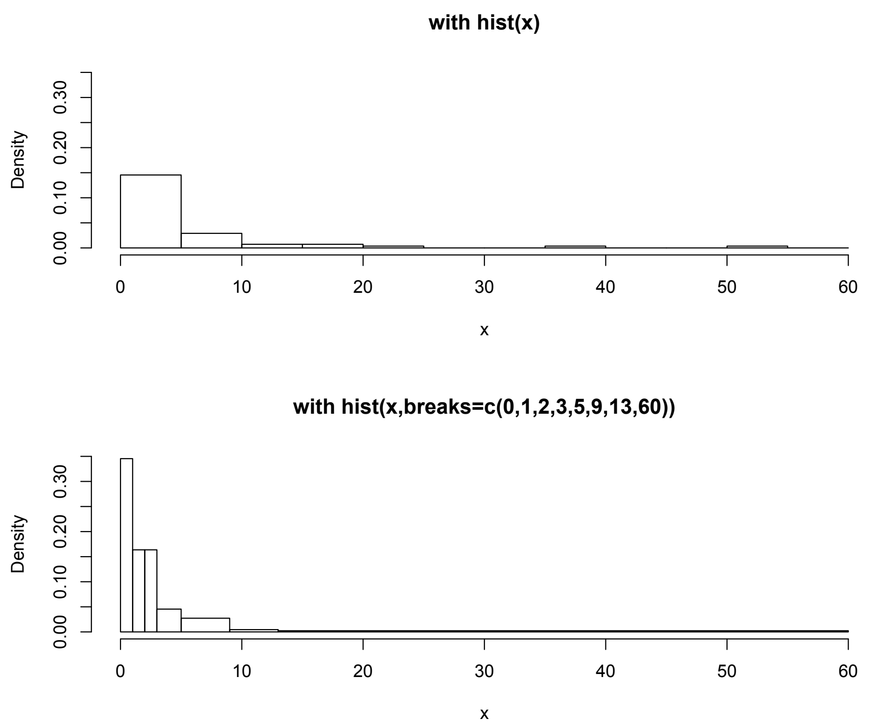
- If all values are discrete (i.e. $\in \mathbb{N}$) draw an

$$\text{ordinate diagram} = \text{plot of } j \text{ against } f_j;$$

  - with `barplot(table(x))` but this omits empty x-values,

  - `barplot(tabulate(x, nbins = max(1, x)))` has a scale preserving x-axis.

- Condensing information can be very useful $\Rightarrow$ slicing $\mathbb{R}$!
- Choose intervals (or midpoints) of the form $[\ l_j\ ,\ u_j\ )$.
  E.g. $\mathbb{R} = (-\infty, 1) \cup [1, 2) \cup [2, 3) \cup [3, 5) \cup [5, 9) \cup [9, 13) \cup [13, \infty)$.
- Produce a histogram with `hist(x)`. Remember: frequencies are represented by the area and not hight.



Optimal Binwidths for Histograms (Not examinable)

There are many ways to draw a boxplot. However, the choices in how a boxplot are represented does matter!

Scott (1992) proved that the asymptotically optimal binwidth (based on various assumptions such as differentiability of the underlying density) is

$$\left(\frac{24\sqrt{\pi}}{n}\right)^{1/3}.$$

This can be used as a reasonable rule of thumb for constructing histograms. This is automated by the R command `hist(x,breaks="Scott")`.

Cumulative frequencies

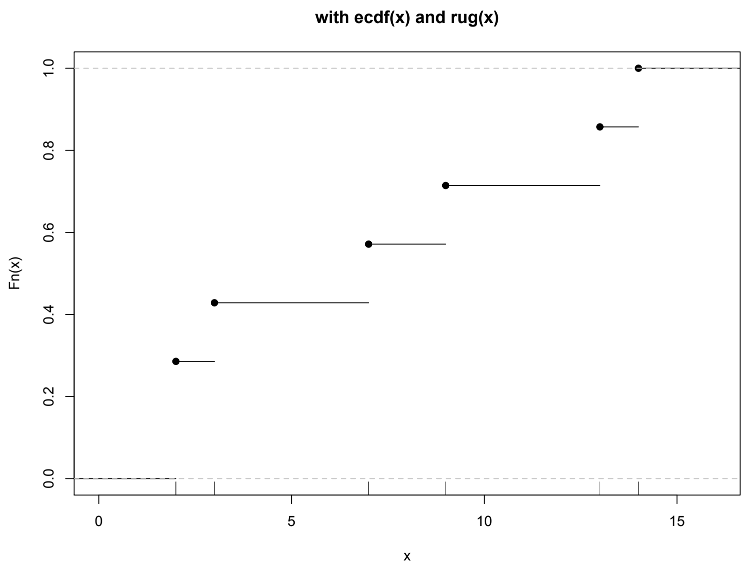
- Consider the number of gold medals for au, fr, jp, kr, nz, ch, and th:
  $x_1 = 14,\ x_2 = 7,\ x_3 = 9,\ x_4 = 13,\ x_5 = 3,\ x_6 = 2,\ x_7 = 2.$
- Ordering observations preserves the information!
  $x_{(1)} = 2 \leq x_{(2)} = 2 \leq \ldots \leq x_{(6)} = 13 \leq x_{(7)} = 14.$
- There are 6 different measurement values:

j	1	2	3	4	5	6
x_j	2	3	7	9	13	14
freq. f_j	2	1	1	1	1	1
cum. freq. F_j	2	3	4	5	6	7

  Hence, $F_j = f_1 + f_2 + \ldots + f_j$.
- Knowing frequencies or cumulative frequencies preserves the information!

Empirical distribution function (EDF)

EDFs are mathematically very useful, have many properties (monotone, continuous from the right) and can be drawn by plotting a step-function using x_j and F_j/n:



Five number summary

Definition 8. The minimum $= x_{(1)}$ and the maximum $= x_{(n)}$.

Definition 9. The range $= x_{(n)} - x_{(1)}$.

Definition 10. The median, $\widetilde{x}$, is a value such that at least half the observations (obs) are less than or equal to $\widetilde{x}$ and at least half the obs are greater or equal to $\widetilde{x}$.

Quartiles are medians of lower and upper half respectively:

Definition 11. The lower quartile, Q_1, is a value such that at least 25% of the obs are $\leq Q_1$ and at least 75% of the obs are $\geq Q_1$.

Definition 12. The upper quartile, Q_3, is a value such that at least 75% of the obs are $\leq Q_3$ and at least 25% of the obs are $\geq Q_3$.

Definition 13. The interquartile range, IQR $= Q_3 - Q_1$.

Five number summary (cont)

Definition 14. The five number summary is

$$(\text{min}, Q_1, \widetilde{x}, Q_3, \text{max}) = (Q_0, Q_1, Q_2, Q_3, Q_4)$$

and is visualized by the boxplot.

Example. Number of gold medals for au, fr, jp, kr, nz, ch, and th:

```
> x = c(14,7,9,13,3,2,2)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.500   7.000   7.143  11.000  14.000
```

or with

```
> quantile(x,c(0.00,0.25,0.50,0.75,1.00))
  0%  25%  50%  75% 100%
 2.0  2.5  7.0 11.0 14.0
```


Boxplot

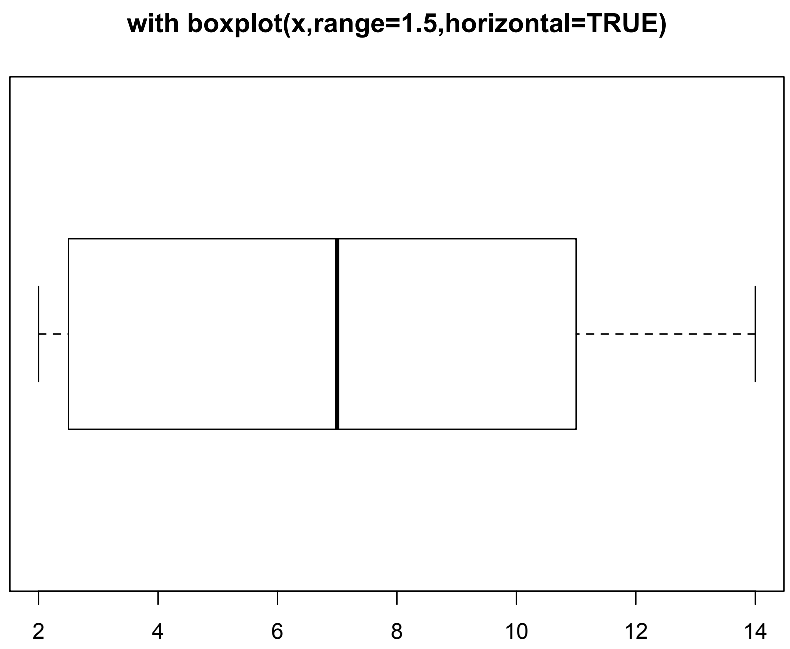
- Draw a box between Q_1 and Q_3;
- add midline at Q_2;
- draw whiskers to min and max if there are no outliers, otherwise to first point larger than LT and first point smaller than UT;
- draw all outlier candidates as points.

Definition 15. Potential outliers are points more than $r = c\times$ IQR beyond the ends of $[Q_1, Q_3]$, $c = 1.5$ is the default choice. Hence,

$$\begin{aligned}\text{Lower Threshold} &= \text{LT} = Q_1 - 1.5 \times \text{IQR},\\ \text{Upper Threshold} &= \text{UT} = Q_3 + 1.5 \times \text{IQR}.\end{aligned}$$

Other choices for c are $1, 1.5, 2, 2.5, 3, \ldots$. The larger c the fewer potential outliers are drawn as single points.

Boxplot (cont)



Boxplot (cont)

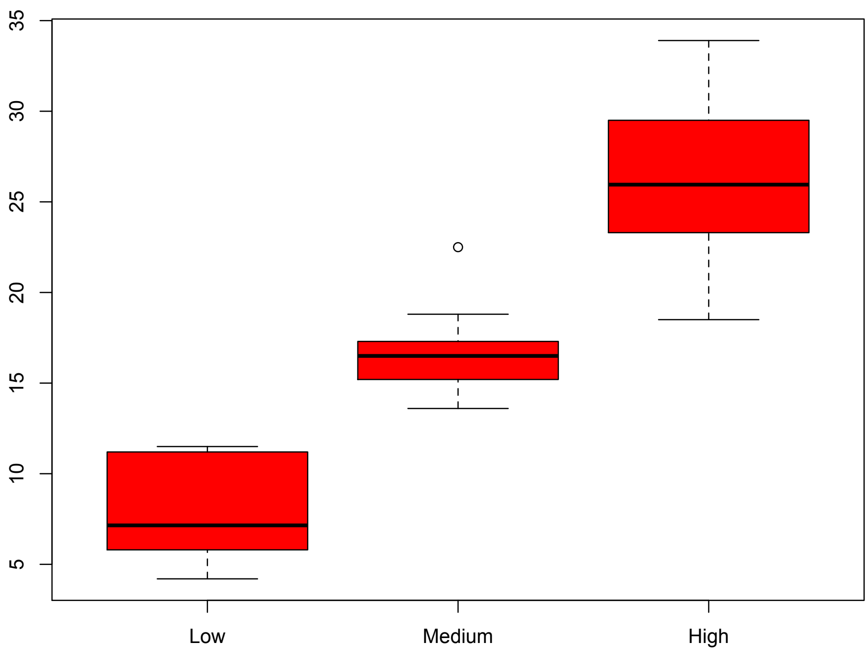
- A single boxplot is boring!
- Boxplots are powerful to compare a continuous variable (e.g. length, weight etc) with a nominal variable (e.g. treatment).
- Length of whisker in R is by default chosen to be $1.5 \times$ IQR, i.e. you don't need to specify `range = 1.5`.
- Boxplots give an easy impression of the shape of the data set:
  - Symmetrical: yes, no?
  - Skewed: left, right?
  - Right skewed = if boxplot is stretched to the right.

Example (Vitamin C and Tooth Growth).

- Data from an (old) experiment into the effects of vitamin C on tooth growth.
- 30 guinea pigs were divided (at random) into three groups of ten and treated with vitamin C (administered in orange juice).

- Group 1 dose was low, group 2 dose was medium and group 3 dose was high.
- Length of odontoblasts (teeth) measured as response variable.
- Reference: C. I. Bliss (1952) *The Statistics of Bioassay*. Academic Press.

Example (cont)



Example (cont)

```
> tapply(Length,Dose.fac,summary)
$Low
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20    5.95    7.15    7.98   10.90   11.50

$Medium
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.60   15.27   16.50   16.77   17.30   22.50

$High
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.50   23.38   25.95   26.14   28.80   33.90
```


Comments for the five number summary

- A median can be calculated for

$$n \text{ odd:} \quad \widetilde{x} = x_{(\frac{n+1}{2})}$$
$$n \text{ even:} \quad \widetilde{x} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}).$$

- If $n/4 \in \mathbb{N}$ then $k = \frac{n}{4}$ and

$$Q_1 = \frac{1}{2}(x_{(k)} + x_{(k+1)}), \quad Q_3 = \frac{1}{2}(x_{(n-k)} + x_{(n-k+1)});$$

  otherwise $k = \lceil \frac{n}{4} \rceil$ and $Q_1 = x_{(k)}, \quad Q_3 = x_{(n-k+1)}$.

- The range covers 100% of the obs

$$x_i \in [\, x_{(1)}\, ,\, x_{(n)}\,] \quad \text{for all } i = 1, \ldots, n,$$

  the IQR covers approximately 50% of the obs.

Density plots (Not examinable)

An alternative to histograms are (kernel) density plots. These are special smoothed positive functions which integrate to 1.

Example. Old Faithful is a cone geyser located in Wyoming, in Yellowstone National Park in the United States. It is also called the most predictable geographical feature on Earth erupting almost every 91 minutes. The data for length between consecutive eruptions can be obtained from the R code

```
> faithful$eruptions
  [1] 3.600 1.800 3.333 2.283 4.533 2.883 4.700 3.600 1.950 4.350 1.833 3.917 4.200 1.750 4.700 2.167 1.750 4.800
 [19] 1.600 4.250 1.800 1.750 3.450 3.067 4.533 3.600 1.967 4.083 3.850 4.433 4.300 4.467 3.367 4.033 3.833 2.017
 [37] 1.867 4.833 1.833 4.783 4.350 1.883 4.567 1.750 4.533 3.317 3.833 2.100 4.633 2.000 4.800 4.716 1.833 4.833
 [55] 1.733 4.883 3.717 1.667 4.567 4.317 2.233 4.500 1.750 4.800 1.817 4.400 4.167 4.700 2.067 4.700 4.033 1.967
 [73] 4.500 4.000 1.983 5.067 2.017 4.567 3.883 3.600 4.133 4.333 4.100 2.633 4.067 4.933 3.950 4.517 2.167 4.000
 [91] 2.200 4.333 1.867 4.817 1.833 4.300 4.667 3.750 1.867 4.900 2.483 4.367 2.100 4.500 4.050 1.867 4.700 1.783
[109] 4.850 3.683 4.733 2.300 4.900 4.417 1.700 4.633 2.317 4.600 1.817 4.417 2.617 4.067 4.250 1.967 4.600 3.767
[127] 1.917 4.500 2.267 4.650 1.867 4.167 2.800 4.333 1.833 4.383 1.883 4.933 2.033 3.733 4.233 2.233 4.533 4.817
[145] 4.333 1.983 4.633 2.017 5.100 1.800 5.033 4.000 2.400 4.600 3.567 4.000 4.500 4.083 1.800 3.967 2.200 4.150
[163] 2.000 3.833 3.500 4.583 2.367 5.000 1.933 4.617 1.917 2.083 4.583 3.333 4.167 4.333 4.500 2.417 4.000 4.167
[181] 1.883 4.583 4.250 3.767 2.033 4.433 4.083 1.833 4.417 2.183 4.800 1.833 4.800 4.100 3.966 4.233 3.500 4.366
[199] 2.250 4.667 2.100 4.350 4.133 1.867 4.600 1.783 4.367 3.850 1.933 4.500 2.383 4.700 1.867 3.833 3.417 4.233
[217] 2.400 4.800 2.000 4.150 1.867 4.267 1.750 4.483 4.000 4.117 4.083 4.267 3.917 4.550 4.083 2.417 4.183 2.217
```

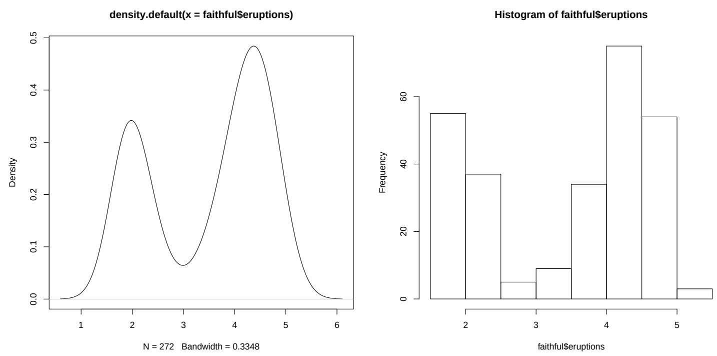


```
[235] 4.450 1.883 1.850 4.283 3.950 2.333 4.150 2.350 4.933 2.900 4.583 3.833 2.083 4.367 2.133 4.350 2.200 4.450
[253] 3.567 4.500 4.150 3.817 3.917 4.450 2.000 4.283 4.767 4.533 1.850 4.250 1.983 2.250 4.750 4.117 2.150 4.417
[271] 1.817 4.467
```

The density plot of this data can be obtained from the R code

```
> plot(density(faithful$eruptions))
```

Density plots are aesthetically pleasing to the eye when compared to histograms:



Some of the fundamental theory around density plots was developed by Australia Statisticians Matt Wand (my PhD supervisor) and Peter Hall (my grandsupervisor).

Additional material for Lecture 2

More on histograms

The best/nicest way to draw histograms is a matter of taste. The following rules serve as a guideline:

- Choose an appropriate number of intervals, e.g. $5 \leq k \leq 20$ or automated by $k = \lfloor \sqrt{n} \rfloor$, where $y = \lfloor x \rfloor \in \mathbb{N}$ is the function that returns the largest integer smaller or equal than x;
- choose appropriate interval boundaries of the form $[\ l_j\ ,\ u_j\ )$, $j = 1, \ldots, k$, e.g. equally spaced;
- determine the absolute/relative frequencies, i.e. the number of observations falling into each of the k intervals;
- draw the histogram such that the x-axis shows the *sliced* real numbers and draw rectangles on top of the histogram with **area proportional to the absolute/relative frequency**;
- don't forget to label both axes.

More on quartiles

Depending on the sample size and the sample itself it can occur that an entire interval satisfies the definition of the lower and upper quartile, respectively. To get a unique solution there exist multiple ways. The suggested unique solution on the previous slide is only one option and can be obtained in R by typing `quantile(x,type=2)`. Reading `help(quantile)` shows that this is the second unique solution out of nine implemented in R. The default option is `type=7`, i.e. if you just type `quantile(x)` this is what is done. What all definitions have in common is that a unique solution is produced by a particular weighted average of the two observations (order statistics) at either end of the interval $[x_{(k)}, x_{(k+1)}]$ and $[x_{(n-k-1)}, x_{(n-k)}]$, respectively, where $k = \lceil n/4 \rceil$.

Monday, 6th August 2012

Lecture 3 - Content

- **Σ notation**
- **Sample mean**
- **Sample variance**
- **Transformation of data to symmetry**

See Phipps & Quine Chapter 1, Sections 3 and 4.

Review Σ notation

For the values $x_1 = 2$, $x_2 = 1$, $x_3 = 5$, $x_4 = 3$, i.e $n = 4$ we have:

$$\begin{aligned}
\sum_{i=1}^{4} x_i &= x_1 + x_2 + x_3 + x_4 = 2 + 1 + 5 + 3 = 11; \\
\sum_{i=1}^{n} x_i^2 &= 2^2 + 1^2 + 5^2 + 3^2 = 39; \\
\sum_{i=2}^{n-1} (3x_i + 2) &= 3 \sum_{i=2}^{3} x_i + 2(\underbrace{n - 1 - 2 + 1}_{2}) = 18 + 4 = 22; \\
\sum_{i=1}^{n} cx_i &= c \sum_{i=1}^{n} x_i; \\
\sum_{i=1}^{n} c1 &= c \sum_{i=1}^{n} 1 = cn.
\end{aligned}$$

Review Σ notation

For the values $x_1 = 2$, $x_2 = 1$, $x_3 = 5$, $x_4 = 3$, $x_5 = 0$ i.e $n = 5$ we have:

$$\begin{aligned}
\sum_{i=1}^{5} \left(\frac{x_i - 3}{\sqrt{5}} \right)^2 &= \frac{1}{5} \sum_{i=1}^{5} (x_i - 3)^2 \\
&= \frac{1}{5} \sum_{i=1}^{5} x_i^2 - \frac{6}{5} \sum_{i=1}^{5} x_i + \frac{5}{5} 9 \\
&= \frac{1}{5} (39 - 66 + 45) \\
&= \frac{18}{5}.
\end{aligned}$$

Sample mean, standard deviation and variance

Definition 16. The sample mean is the simple average of the observations. For observations $x_1, x_2, \ldots, x_n$

$$\overline{x}_n = \overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Theorem 1. Given constants $a \in \mathbb{R}$ and $b \in \mathbb{R}$ and obs $x_1, x_2, \ldots, x_n$. Then the mean of the transformed observations $y_i = a \times x_i + b$, $i = 1, 2, \ldots, n$, is

$$\overline{y} = a \times \overline{x} + b.$$

Proof. Write down the left hand side of the equation and begin with the definition:

$$\begin{aligned}
\overline{y} &\overset{\text{def.}}{=} \frac{1}{n}\sum_{i=1}^{n} y_i \overset{\text{expand}}{=} \frac{1}{n}(y_1 + y_2 + \ldots + y_n) \\
&\overset{\text{replace}}{=} \frac{1}{n}\Big((ax_1 + b) + (ax_2 + b) + \ldots + (ax_n + b)\Big) \\
&\overset{\text{group}}{=} \frac{1}{n}\Big(a(x_1 + x_2 \ldots + x_n) + nb\Big) \\
&\overset{\text{simpl.}}{=} \frac{a}{n}\sum_{i=1}^{n} x_i + \frac{n}{n}b \\
&= a \times \overline{x} + b.
\end{aligned}$$

$\square$

Change of working origin and unit

The theorem helps to transform data to a new working origin, a, and a new working unit, h:

$$d_i = \frac{x_i - a}{h} = \frac{1}{h}x_i - \frac{a}{h}, \quad i = 1, \ldots, n.$$

Thus,

$$\overline{d} = \frac{1}{h}\overline{x} - \frac{a}{h}$$

and solving for $\overline{x}$ yields

$$\overline{x} = h\overline{d} + a.$$

Example. Find the mean of x_i: 9.80, 9.81, 9.82, 9.84.

$a = 9.80$ and $h = 0.01$

$\Rightarrow d_i : 0, 1, 2, 4$. Thus, $\overline{d} = \frac{7}{4} = 1.75$ and $\overline{x} = 0.01 \times 1.75 + 9.80 = 9.8175$

Mean vs median

- ☐ Mean, $\overline{x}$, easier to calculate and to handle than the median, $\widetilde{x}$.
- ☐ If the data are approximately symmetric then $\overline{x} \approx \widetilde{x}$.
- ☐ If the data are skewed then the mean is pulled toward the long tail.
- ☐ $\widetilde{x}$ is robust against outliers and incorrect readings whereas $\overline{x}$ is not.

Example. Assume in the previous example 9.80 is misread as 3.80.

```
> x = c(3.80,9.81,9.82,9.84)
> mean(x)
[1] 8.3175
> median(x)
[1] 9.815
```

The mean is a Least Squares (LS) estimate!

Definition of Least Squares:

$$S(a) := \sum_{i=1}^{n}(x_i - a)^2; \quad \text{minimise} \;\; S(a).$$

Hence,

$$\begin{aligned} S(a) &= \sum(x_i^2 - 2ax_i + a^2) = \sum x_i^2 - 2a\left(\sum x_i\right) + n \times a^2 \\ &= \sum x_i^2 - 2an\overline{x} + na^2 \\ &\Rightarrow \frac{\partial S(a)}{\partial a} = S'(a) = -2n\overline{x} + 2na \end{aligned}$$

$S'(a)$ equals 0 if $a = \overline{x}$.

The median is a Least Absolute Deviation (LAD) estimate!

Definition of Least Absolute Deviation:

$$D(a) := \sum_{i=1}^{n}|x_i - a|; \quad \text{minimise} \;\; D(a).$$

For simplicity assume that $x_i \neq a$ for all x_i. Then

$$\frac{\partial D(a)}{\partial a} = -\sum_{i=1}^{n} \text{sign}(x_i - a)$$

where

$$\text{sign}(z) = \begin{cases} 1 & z > 0 \\ 0 & z = 0 \\ -1 & z < 0 \end{cases}$$

Thus a solution to $D'(a) = 0$ is the value a such that

$\{$ The number of x_is greater than a $\} = \{$ The number of x_is less than a$\}$.

In other words the sample median!

Sample variance and standard deviation

Definition 17. For data $x_1, x_2, \ldots, x_n$ the sample standard deviation s_x is

$$s_x = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2}$$

and the sample variance is

$$s_x^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{1}{n-1} S_{xx}.$$

Alternative formula for s_x^2

($\sum$ index omitted)

$$\begin{aligned} S_{xx} &= \sum (x_i - \overline{x})^2 = \sum (x_i^2 - 2\overline{x} x_i + \overline{x}^2) \\ &= \sum x_i^2 - 2\overline{x} \sum x_i + n \times \overline{x}^2; \quad \text{with } \sum x_i = n \times \overline{x}, \\ &= \sum x_i^2 - n\overline{x}^2 \\ &= \sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2. \end{aligned}$$

Hence,

$$s_x^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} x_i \right)^2 \right) = \frac{1}{n-1} \left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2 \right).$$

Change of working origin and unit (cont)

Exercise. Show that for working origin a and working unit h the variance of data $x_1, x_2, \ldots, x_n$ equals $h^2\times$ the variance of the transformed data

$$d_i = \frac{x_i - a}{h} \quad \text{i.e.} \quad x_i = h \cdot d_i + a \Rightarrow s_x^2 = h^2 \times s_d^2,$$

and therefore $s_x = h \times s_d$.

Example. Data: 340, 350, 360, 370, 380.

For $a = 360$, $h = 10$ we get

$d_i : -2, -1, 0, 1, 2$ and $\sum d_i = 0$, $\sum d_i^2 = 10$:

and

$$s_d^2 = \frac{1}{5-1}\left(10 - \frac{1}{5} \times 0^2\right) = 2.5.$$

So $s_x^2 = h^2 \times s_d^2 = 100 \times 2.5 = 250$.

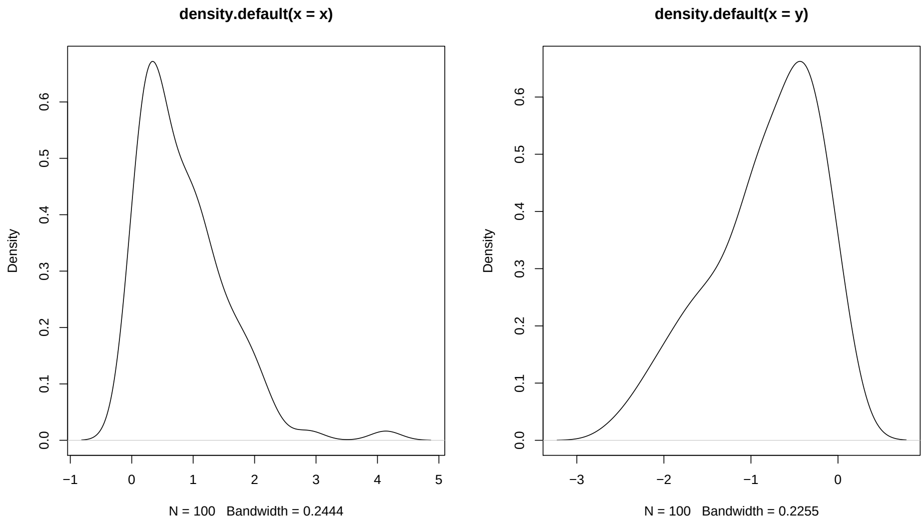
```
> x = c(340,350,360,370,380)
> var(x)
[1] 250
> sd(x)
[1] 15.81139
```

Skewed data

Definition 18. Data are said to be left skewed if the left tail of the density is longer than the right.

Definition 19. Data are said to be right skewed if the right tail of the density is longer than the left.

The data in the left hand side plot are right skewed whereas the data on the right hand side are left skewed.



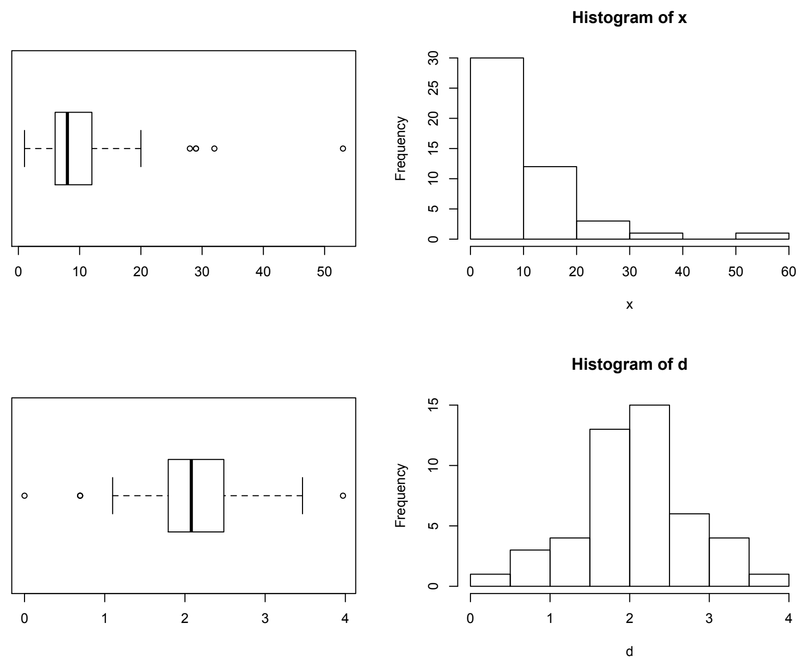
Transformations of data

- To have symmetric data can be a desirable property for some statistical methods.
- Data obtained as differences (e.g. from before/after studies) are often approximately symmetric.
- For right skewed data $\{x_i\}$
  - $d_i = x_i^a$, for various values of $a \in (0, 1)$.
  - $d_i = \log x_i$
  - $d_i = -x_i^{-a}$, for $a > 0$.
- Left skewed data $\{x_i\}$: transform into right skewed data by $d_i = -x_i$.
- For data $\{x_i\}$ recorded as proportions, i.e. $x_i \in (0, 1)$, the logit transform can be used: $d_i = \log \frac{x_i}{1-x_i}$.

Example (Swiss fertility data). Execute the following code in R:

```
> data(swiss)
> help(swiss)
> names(swiss)
> x = swiss$Education
> d = log(x)
> par(mfrow=c(2,2))
> boxplot(x,horizontal=TRUE)
> hist(x)
> boxplot(d,horizontal=TRUE)
> hist(d)
> summary(x)
> summary(d)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    6.00    8.00   10.98   12.00   53.00
> summary(d)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.792   2.079   2.099   2.485   3.970
```

Example (cont)



Additional material for Lecture 3

More on the sample variance and sample standard deviation

The sample variance is almost the average of squared distances to the sample mean. But instead of using n in the denominator a $(n-1)$ term is used. This will make perfect sense after STAT2011/2911 but probably not too much sense at this stage. Note that it is almost the average, particularly when n is large since $\lim_{n\to\infty} \frac{n}{n-1} = 1$ and

$$s_x^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{n}{(n-1)}\Big(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\Big).$$

To fully appreciate to correct by $\frac{n}{(n-1)}$ the following concepts have to be understood first: random variables, function of random variables, expected value, covariance of random variables, unbiasedness.

Both the range of the sample and the standard deviation of the sample measure aspects of spread or scale. They are to a certain extent depend of each other. An equality is given in the theorem below.

Theorem (Thomson, 1955) Let $w = x_{(n)} - x_{(1)}$ be the range of the observations $x_1, \dots, x_n$ then the sample standard deviation satisfies

$$\sqrt{\frac{1}{2(n-1)}} \le \frac{s_x}{w} \le \begin{cases} \frac{1}{2}\sqrt{\frac{n}{n-1}}; & n \text{ even,} \\ \frac{1}{2}\sqrt{\frac{n+1}{n}}; & n \text{ odd.} \end{cases}$$

Tuesday, 7th August 2012

Lecture 4 - Content

- **Bivariate data**
- **Scatterplot**
- **Correlation coefficient**

See Phipps & Quine Chapter 1, Section 5.

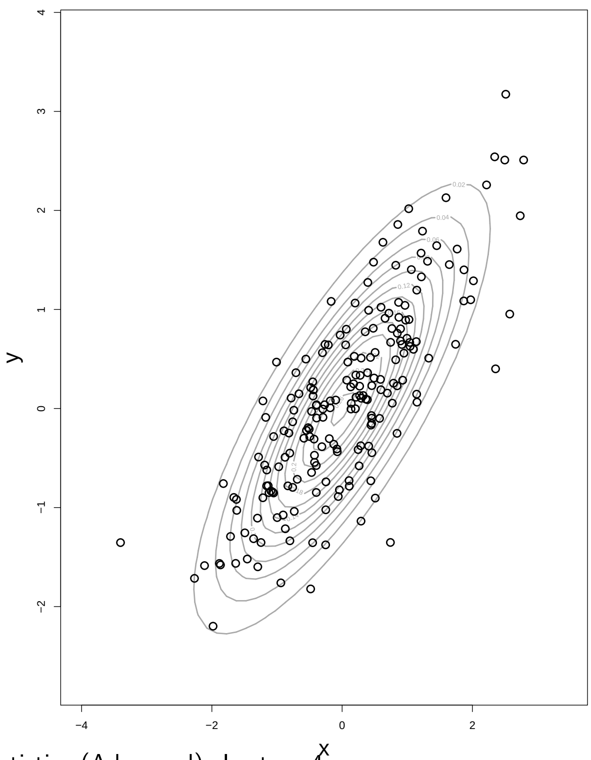
Bivariate data

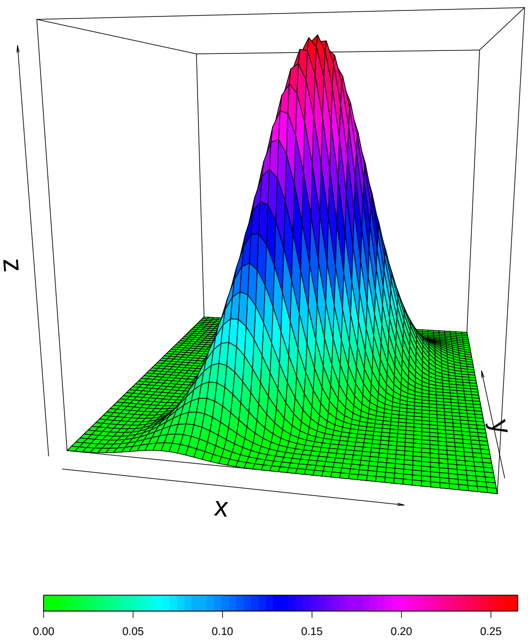
- So far univariate data only, i.e. observations on a single feature.
- In general multivariate data, e.g. bivariate data

$$\begin{aligned} x &= \text{patient's age} \\ y &= \text{patient's reaction time} \end{aligned}$$

- The first step in the analysis of multivariate data is visualisation!

Visualisation!





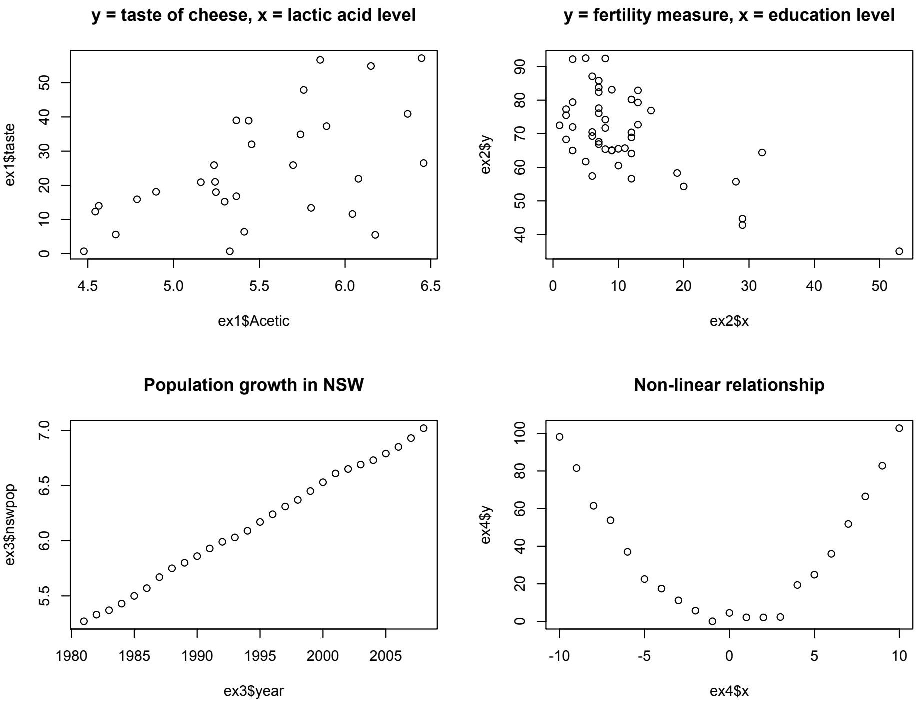
Scatterplot

For bivariate data $(x_1, y_1), \dots, (x_n, y_n)$ simply plot the points.

Four examples:

- Taste of matured cheese and lactic acid level. $(r = 0.55)$.
- Education level (x_i) and fertility level (y_i) of Swiss provinces (French speaking part) in $n = 47$. $(r = -0.66)$.
- Population growth in NSW between 1981 - 2008. $(r = 0.9992)$.
- Noisy non-linear functional relationship. $(r = 0.015)$.

Four examples (cont)



Correlation coefficient

Definition 20. The correlation coefficient is a numerical index that measures the degree of linear association between x and y,

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2) \times (\sum_{i=1}^{n}(y_i - \overline{y})^2)}}.$$

Note that,

$$\begin{aligned}
S_{xx} &= \sum_{i=1}^{n}(x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 \\
S_{yy} &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \left(\sum_{i=1}^{n} y_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\
S_{xy} &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \left(\sum_{i=1}^{n} x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)
\end{aligned}$$

In R: calculate r with `cor(x,y)`.

Example

Dose (in grams)	x	30	40	50	60	70	80	90	100
Breathing rate	y	16	14	13	13	11	12	9	9

To calculate r we need $n = 8$, $\sum_{i=1}^n x_i y_i = 5910$, $\sum_{i=1}^n x_i = 520$, $\sum_{i=1}^n y_i = 97$, $\sum_{i=1}^n x_i^2 = 38000$ and $\sum_{i=1}^n y_i^2 = 1217$. So

- $S_{xy} = \left(\sum_{i=1}^n x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right) = 5910 - \frac{1}{8}520 \times 97 = -395$
- $S_{xx} = 38000 - \frac{1}{8}520^2 = 4200$
- $S_{yy} = 1217 - \frac{1}{8}97^2 = 40.875$
- $r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-395}{\sqrt{4200 \times 40.875}} = -0.9533$ (to 4 d.p)

Properties of the correlation coefficient

i) The correlation coefficient is always between -1 and 1: $r \in [-1, 1]$.

ii) If $r = 1$ then all obs. (x_i, y_i) lie on a straight line with positive slope.

iii) If $r = -1$ then all obs. (x_i, y_i) lie on a straight line with negative slope.

iv) If $r = 0$ it does not follow that there is no relationship between x and y!

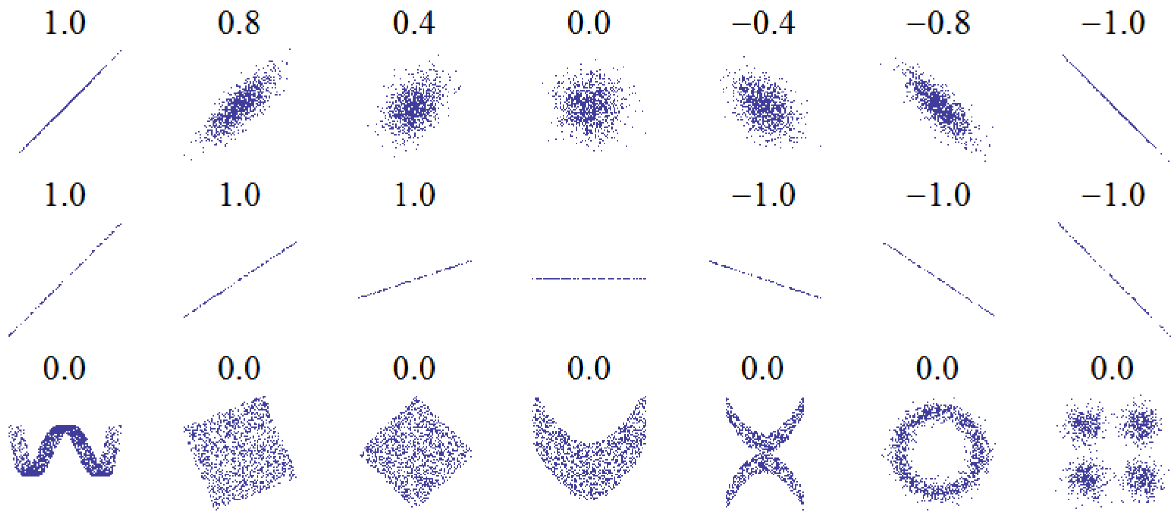
v) For high r (close to 1 or -1) it does not follow that there must be a relationship between x and y!

Proof. i) Is true because,

$$0 \le \sum_{i=1}^n \left(\frac{x_i - \overline{x}}{\sqrt{S_{xx}}} - \frac{y_i - \overline{y}}{\sqrt{S_{yy}}}\right)^2 = \underbrace{\frac{\sum (x_i - \overline{x})^2}{S_{xx}}}_{=1} + \underbrace{\frac{\sum (y_i - \overline{y})^2}{S_{yy}}}_{=1} - 2\underbrace{\frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{S_{xx}S_{yy}}}}_{=r} = 2 - 2r.$$

Hence it follows that $r \le 1$. Similarly for $r \ge -1$ but with $\sum_{i=1}^n \left(\frac{x_i - \overline{x}}{\sqrt{S_{xx}}} + \frac{y_i - \overline{y}}{\sqrt{S_{yy}}}\right)^2$. □

Correlation Examples



Alternative formula for S_{xy}

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

Expanding and simplifying yields

$$\begin{aligned} S_{xy} &= \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum (x_i y_i - \overline{x} y_i - \overline{y} x_i + \overline{x}\,\overline{y}) \\ &= \sum x_i y_i - \underbrace{\overline{x}}_{\frac{1}{n}\sum x_i} \left(\sum y_i\right) - \overline{y} \underbrace{\sum x_i}_{n\overline{x}} + n\overline{x}\,\overline{y} \\ &= \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) \underbrace{-n\overline{x}\,\overline{y} + n\overline{x}\,\overline{y}}_{=0} \end{aligned}$$

Theorem 2. Linear rescaling and translating of x or y values does not change the correlation coefficient r.

Proof: Let $u_i = a + bx_i$ and $v_i = c + dy_i$ where $b > 0$ and $d > 0$ then

$$S_{uv} = \sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v}).$$

Then

$$\begin{aligned} u_i - \overline{u} &= a + bx_i - \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(a + bx_i) \\ &= a + bx_i - \tfrac{1}{n}\textstyle\sum_{i=1}^{n} a - \tfrac{b}{n}\textstyle\sum_{i=1}^{n} x_i \\ &= b(x_i - \overline{x}) \end{aligned}$$

Hence, $S_{uv} = bd\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = bdS_{xy}$. Similarly, $S_{uu} = b^2 S_{xx}$ and $S_{vv} = d^2 S_{yy}$. Then

$$r_{uv} = \frac{S_{uv}}{\sqrt{S_{uu}S_{vv}}} = \frac{bdS_{xy}}{\sqrt{b^2d^2S_{xx}S_{yy}}} = r_{xy}.$$

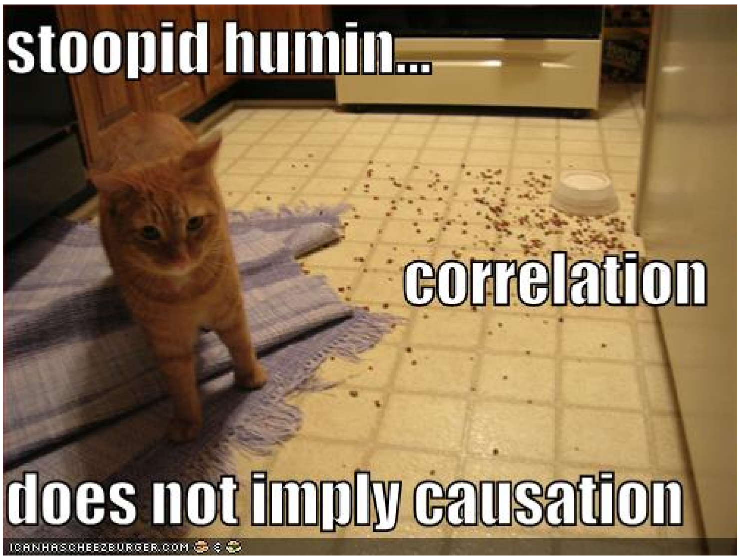
Example (Swiss fertility data). With R:

```
> x = swiss$Education
> y = swiss$Fertility
> length(x)
[1] 47
> c(sum(x),sum(x**2))
[1]  516 9918
> c(sum(y),sum(y**2))
[1]   3296.7 238416.9
> sum(x*y)
[1] 32526
> cor(x,y)
[1] -0.6637889
```

Thus, $S_{xx} = 9918 - \frac{(516)^2}{47} = 4252.979$, $S_{yy} = 7177.955$, $S_{xy} = 32526 - \frac{516\times 3296.7}{47}$. Hence,

$$r = \frac{-3667.557}{\sqrt{4252.979 \times 7177.955}} = -0.6637888.$$

Common misconception – Correlation is not cause!



Common misconception – Correlation is not cause!

Causation between two events implies a dependence between the two events.

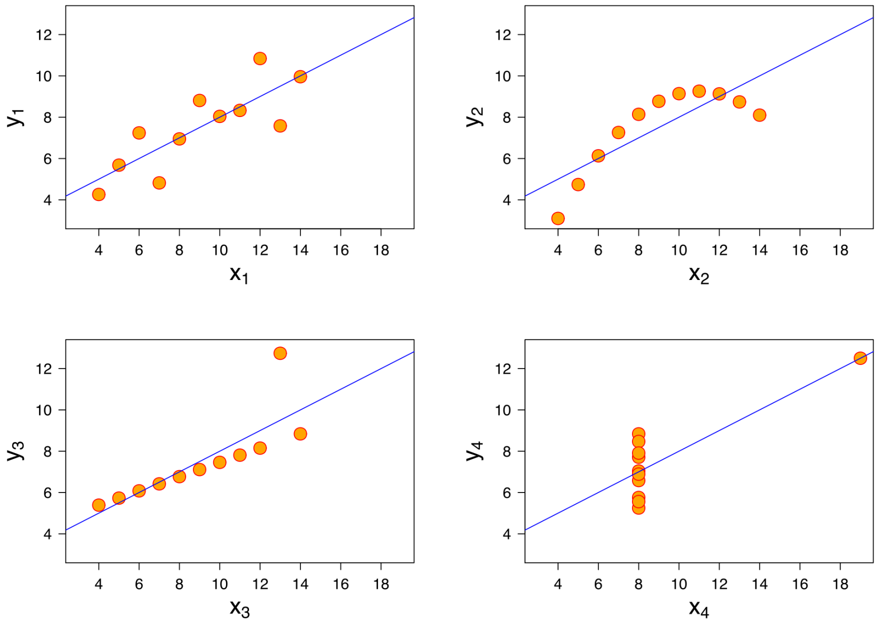
However, correlation cannot be used to infer a causal relationship between the variables because the cause of the underlying the correlation may be indirect and unknown, and high correlations can occur where no causal process exists.

For example, one may observe a correlation between the lecturer John Ormerod waking up and daybreak, though there is no direct causal relationship between these events, i.e. John Ormerod does not cause the sun to rise.

A correlation between age and height in children is fairly causally transparent, but a correlation between mood and health in people is less so. Does improved mood lead to improved health, or does good health lead to good mood, or both?

Common misconception – Correlation does not mean linearity!

All of the examples below have a correlation of 0.816.

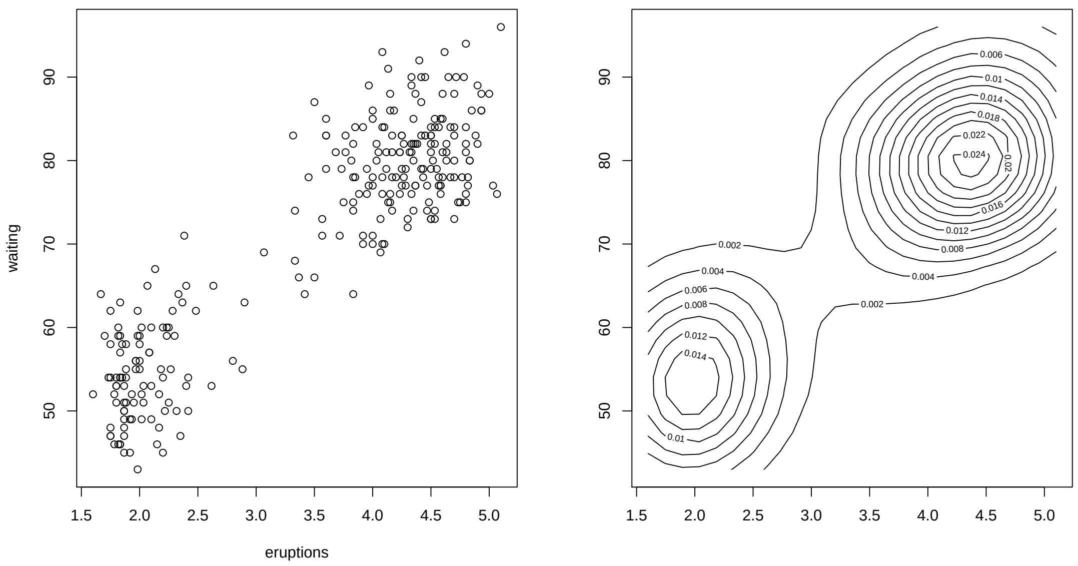


Contour plots (Not examinable)

Contour plots (based on density methods) are another useful way of looking at data.

On the next page the left hand side plot below is a scatterplot of the "Old faithful" dataset we saw earlier whereas the right hand side plot corresponds to the R command:

```
> library(MASS) # load an R library into memory
> contour(kde2d(faithful$eruptions,faithful$waiting))
```



More on measuring correlation (Not examinable)

The correlation coefficient r, often called Pearson correlation, is just one quantity that measures if two set of observations are correlated, i.e. are related. There are many more. Probably the second most famous is the Spearman rank correlation. Instead of using the original observations $(x_1, y_1), \ldots, (x_n, y_n)$ the corresponding ranks are analysed:

$$x_i \mapsto u_i = \mathsf{rank}(x_i)$$

$$y_i \mapsto v_i = \mathsf{rank}(y_i)$$

The Spearman rank correlation coefficient is simply the Pearson correlation coefficient for the u's and v's,

$$\rho = \frac{\sum \left(u_i - \overline{u}\right)\left(v_i - \frac{n+1}{2}\right)}{\sqrt{\sum \left(u_i - \frac{n+1}{2}\right)^2 \sum \left(v_i - \overline{v}\right)^2}}.$$

A toy example in R:

```
> x = c(5,4,2,1.5,3)
> y = c(5.2,4.7,2.8,1.9,4.1)
> u = rank(x)
> v = rank(y)
> cor(x,y)
[1] 0.9696742
> cor(u,v)
[1] 1
> plot(x,y)
```

The Spearman correlation coefficient measures to what extent the relationship of x and y is monotone. In the toy example the scatterplot of x and y shows a perfectly monotone relationship. Therefore, $\rho = 1$ whereas $r = 0.97$, i.e. is not exactly one because there could well be some quadratic relationship.

Monday, 13th August 2012

Lecture 5 - Content

- **Simple linear regression**

See Phipps & Quine Chapter 1, Section 5.

Quotes about regression

Yale Law Professor Ian Ayres on regression:

> William Grove, completed a meta-analysis of 136 human versus machine studies. In only 8 out of 136 studies was expert opinion found to be appreciably more accurate than statistical prediction... Indeed, regression equations are so much better than humans... that even very crude regressions with just a few variables have been found to outpredict humans.

Cognitive psychologists Richard Nisbett and Lee Ross on regression:

> Human judges are not merely worse than optimal regression equations; they are worse than almost any regression equation.

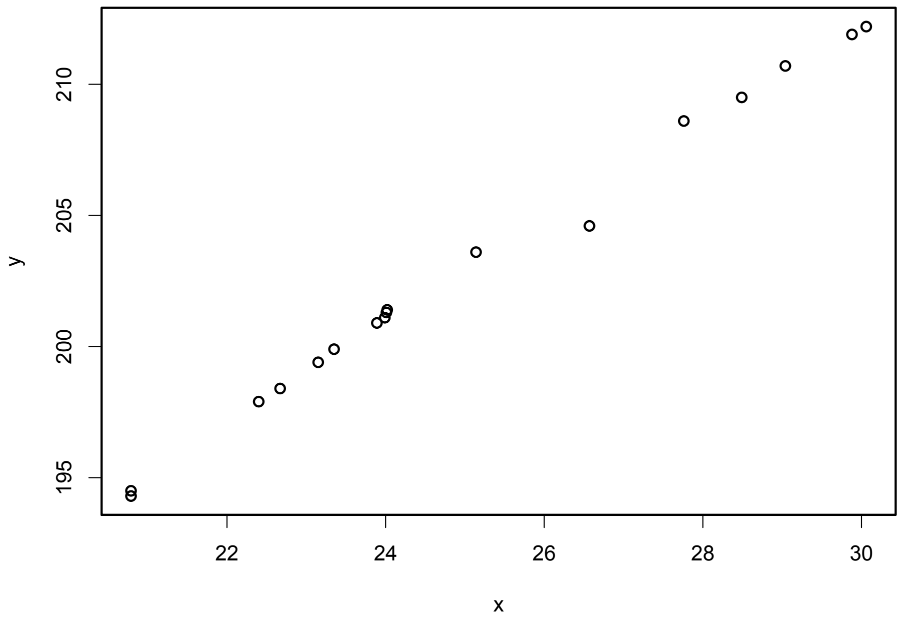
Simple linear regression

Linear regression seeks to model the relationship between the mean of a response variable, y, and a single explanatory variable x.

Example (Boiling point data).

- Data on boiling point in degrees Fahrenheit (y) and pressure in inches of mercury (x), collected during an expedition in the Alps.
- Reference: Hand et al. (1994). *A Handbook of Small Data Sets*, London: C. & Hall.

The scatterplot shows that there is a clear relationship between y (temperature) and x (pressure).



Regression lines

For data $(x_1, y_1), \ldots, (x_n, y_n)$ we want to find a regression line that "fits" the data points. A simple linear regression model is

$$y_i = \underbrace{a + bx_i}_{\widehat{y}_i} + e_i = \widehat{y}_i + e_i, \quad i = 1, \ldots, n, \tag{1}$$

where

- a is the intercept of the regression line,
- b is the slope of the regression line,
- $e_i = y_i - \widehat{y}_i$ is called the residual (error) of observation i.

The "best" regression line

Suppose we want to fit the "best" line $y = a + bx$ to the data.

There are a number of ways to define "best". We could choose a and b such that the sum of squared residuals is minimised:

$$M(a, b) = \sum_{i=1}^{n} (y_i - a - bx_i)^2 = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

or where the sum of absolute residuals is minimised:

$$D(a, b) = \sum_{i=1}^{n} |y_i - a - bx_i| = \sum_{i=1}^{n} |y_i - \widehat{y}_i| = \sum_{i=1}^{n} |e_i|$$

or where the maximum absolute residual is minimised:

$$H(a, b) = \mathsf{max}_i |y_i - a - bx_i| = \mathsf{max}_i |y_i - \widehat{y}_i| = \mathsf{max}_i |e_i|$$

The first problem corresponds to the "least squares" (LS) method, which chooses values of a and b which minimise the sum of the squares of these residuals. The other criteria are **much** harder to minimise.

The least squares regression line

Theorem 3. The least squares regression line, i.e. with a and b such that $M(a, b)$ is minimal, has intercept

$$a = a_{\mathsf{LS}} = \overline{y} - b_{\mathsf{LS}} \overline{x}$$

and slope

$$b = b_{\mathsf{LS}} = \frac{S_{xy}}{S_{xx}}.$$

Recall

- $S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{1}{n} (\sum x_i)(\sum y_i)$;
- $S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum x_i^2 - \frac{1}{n} (\sum x_i)^2$.

Proof. Let $M = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$.
First minimise over a:

$$\frac{\partial M}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i)(-1) = 0$$

Hence,

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \Leftrightarrow n\overline{y} - na - nb\overline{x} = 0 \Rightarrow a = \overline{y} - b\overline{x}.$$

Then, substitute for a in the expression for M to get

$$M = \sum_{i=1}^{n}(y_i - \overline{y} - b(x_i - \overline{x}))^2 = S_{yy} - 2bS_{xy} + b^2 S_{xx}.$$

Minimise over b:

$$\frac{\partial M}{\partial b} = -2S_{xy} + 2bS_{xx} = 0 \Leftrightarrow b = \frac{S_{xy}}{S_{xx}}.$$

□

Example

In a study on the absorption of a drug, the dose x (in grams) and concentration in the urine y (in mg/g) were recorded as:

Dose (in grams)	x	46	53	37	42	34	29	60	44	41	48	33	40
Urine Concentration	y	12	14	11	13	10	8	17	12	10	15	9	13

$$\sum_{i=1}^{n} x_i = 507, \quad \sum_{i=1}^{n} y_i = 144, \quad n = 12$$
$$\sum_{i=1}^{n} x_i^2 = 22265, \quad \sum_{i=1}^{n} y_i^2 = 1802, \quad \sum_{i=1}^{n} x_i y_i = 6314$$

- □ $S_{xy} = \left(\sum_{i=1}^{n} x_i y_i\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 6314 - \frac{1}{12}(507)(144) = 230$
- □ $S_{xx} = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 22265 - \frac{1}{12}(507)^2 = 844.25$
- □ $b = \frac{S_{xy}}{S_{xx}} = \frac{230}{844.25} = 0.272431$ (to 6 d.p)
- □ $a = \overline{y} - b\overline{x} = \frac{144}{12} - 0.272431 \times \frac{507}{12} = 0.489782$ (to 6 d.p)

Fitted regression line

Because $a = \overline{y} - b\overline{x}$ we can write

$$y = a + bx = \overline{y} - b\overline{x} + bx = \overline{y} + b(x - \overline{x}),$$

so the regression line passes through the component-wise mean $(\overline{x}, \overline{y})$.

Correlation coefficient and regression slope

Recall that the correlation coefficient between vectors x and y is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \in [-1, 1].$$

Because,

$$b = \frac{S_{xy}}{S_{xx}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{xx}}}\frac{\sqrt{S_{yy}}}{\sqrt{S_{yy}}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}.$$

Therefore, b and r have the same sign, both positive or both negative.

Example (Boiling point, cont). The $n = 17$ observations are:

```
> x
 [1] 20.79 20.79 22.40 22.67 23.15 23.35 23.89 23.99 24.02
[10] 24.01 25.14 26.57 28.49 27.76 29.04 29.88 30.06
> y
 [1] 194.5 194.3 197.9 198.4 199.4 199.9 200.9 201.1 201.4
[10] 201.3 203.6 204.6 209.5 208.6 210.7 211.9 212.2
```

To obtain a and b we first calculate the following auxiliary numbers:

$\sum x_i = 426$ $\qquad \sum y_i = 3,450.2$

$\sum x_i^2 = 10,821$ $\qquad \sum y_i^2 = 700,759$

$\sum x_i y_i = 86,735.5$

Thus,

$S_{xx} = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2 = 10,821 - \frac{426^2}{17} = 145.9412$

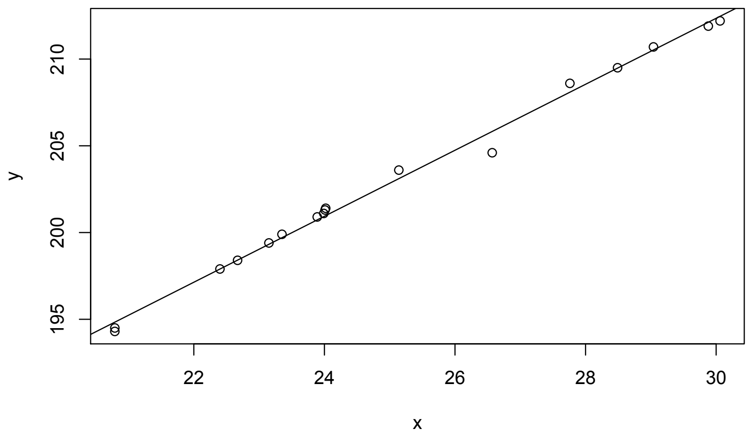
$S_{yy} = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2 = 530.7824$

$S_{xy} = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i) = 277.5421 \Rightarrow r = \frac{S_{xy}}{\sqrt{S_{xx}S_{xy}}} = \frac{277.5421}{278.3185} = 0.9972$

and we have $b = \frac{S_{xy}}{S_{xx}} = \frac{277.5421}{145.9412} = 1.90$ and $a = \frac{3450.2}{17} - 1.90 \times \frac{426}{17} = 155.3$.

In R: Execute,

```
> plot(x,y)       # as before, produces scatterplot of x against y
> abline(lm(y~x)) # lm(y~x) : lm() = linear model function;
                  # y~x means model y by x
> lm(y~x)
(Intercept)            x
    155.296        1.902
```



Additional material for Lecture 5

More on simple linear regression

A function $f(\beta) : \mathbb{R}^2 \mapsto \mathbb{R}$ is linear (a linear map) if and only if it preserves addition and scalar multiplication, i.e.

1. for all $\beta, \gamma \in \mathbb{R}^2$ we have $f(\beta + \gamma) = f(\beta) + f(\gamma)$,
2. for all $c \in \mathbb{R}$ we have $f(c\beta) = cf(\beta)$.

A simple linear regression is considered to be a function of the intercept a and slope b, given the information of the data, i.e. $(x_1, y_1), \ldots, (x_n, y_n)$. Therefore one can do much more with simple linear regression than just fitting a straight line. For example consider a simple transformation of the explanatory variable x such as $z = \log(x)$. Then,

$$y_i = a + b\log(x_i) + e_i = f(a, b|x_i) + e_i = a + bz_i + e_i = f(a, b|z_i) + e_i, \quad i = 1, \ldots, n,$$

is clearly a simple linear regression model since

$$y = f(a, b|x) = a + bz \Rightarrow f(ca, cb|z) = ca + cbz = cf(a, b|z) \quad \text{and} \quad f(a_0 + a, b_0 + b|z) = a_0 + b_0 z + a + bz = f(a_0, b_0|z) + f(a, b|z).$$

Linear regression models will turn out to be a very powerful instrument in the analysis of higher dimensional data and are in statistics as powerful as are Taylor or Fourier series in calculus. You can learn more on linear models in STAT2912 and much more in STAT3912.

Tuesday, 14th August 2012

Lecture 6 - Content

- □ **Semi-log transformation**
- □ **Residual plots**
- □ **Explaining variability**

See Phipps & Quine Chapter 1, Section 5.

Semi-log Transformations

Suppose an exponential trend of the type $y = A \times B^x$ is expected.
Take (natural) logs of both sides to obtain

$$\begin{aligned} \log(y) &= \log(A \times B^x) \\ &= \log(A) + \log(B) \times x \end{aligned}$$

and so if we put $Y = \log(y)$, $X = x$, $a = \log(A)$ and $b = \log(B)$ the line we now want to estimate is $Y = a + bX$.

Procedure: Perform a semi-log transform, i.e. $X_i = x_i$ and $Y_i = \log(y_i)$, then find a LSR line for the points (X_i, Y_i) for the line $Y = a + bX$ in the usual way. Lastly, transform back to obtain the fitted curve $y = A \times B^x$ (using $A = e^a$ and $B = e^b$).

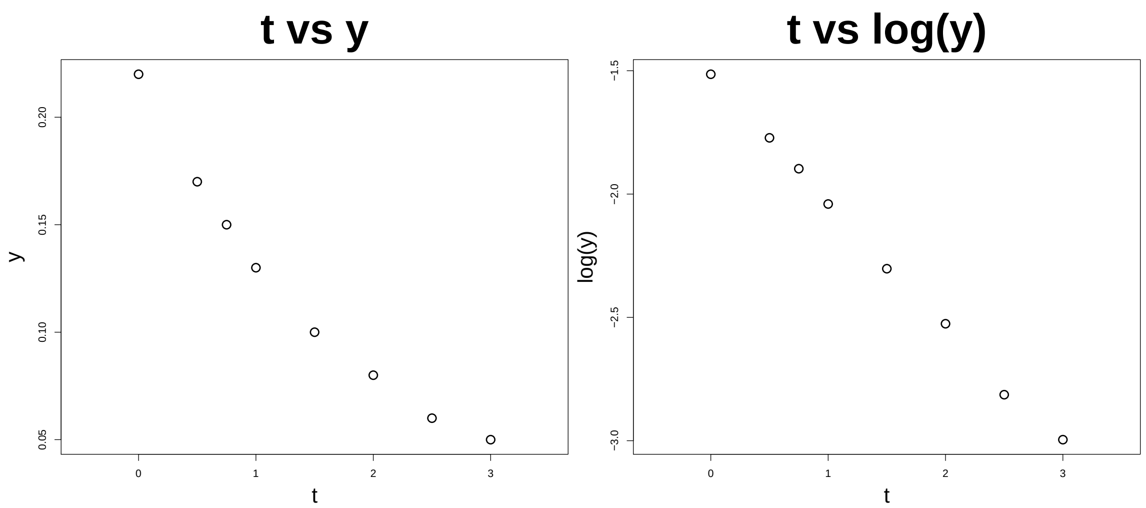
The Semi-log Transformation – Example

The alcoholic content, y (mg/ml) of a person's blood, t hours after drinking whisky, is displayed in the table below:

Time (h)	t	0.00	0.50	0.75	1.00	1.50	2.00	2.50	3.00
Alcohol (mg/ml)	y	0.22	0.17	0.15	0.13	0.10	0.08	0.06	0.05

Why do you think that an exponential relationship, $y = A \times B^t$, might be an appropriate relationship?

The Semi-log Transformation – Plots



The Semi-log Transformation – Working Out

The original data is

t	0.00	0.50	0.75	1.00	1.50	2.00	2.50	3.00
y	0.22	0.17	0.15	0.13	0.10	0.08	0.06	0.05

Using a semi-log transformation we have

$X = t$	0.00	0.50	0.75	1.00	1.50	2.00	2.50	3.00
$Y = \log(y)$	-1.51	-1.77	-1.90	-2.04	-2.30	-2.53	-2.81	-3.00

Using these values:

$$\sum_{i=1}^{n} X_i = 11.25, \quad \sum_{i=1}^{n} Y_i = -17.86, \quad n = 8$$
$$\sum_{i=1}^{n} X_i^2 = 23.31, \quad \sum_{i=1}^{n} Y_i^2 = 41.77, \quad \sum_{i=1}^{n} X_i Y_i = -28.89$$

The Semi-log Transformation – Working Out

Using the values:

$$\sum_{i=1}^{n} X_i = 11.25, \quad \sum_{i=1}^{n} Y_i = -17.86, \quad n = 8$$
$$\sum_{i=1}^{n} X_i^2 = 23.31, \quad \sum_{i=1}^{n} Y_i^2 = 41.77, \quad \sum_{i=1}^{n} X_i Y_i = -28.89$$

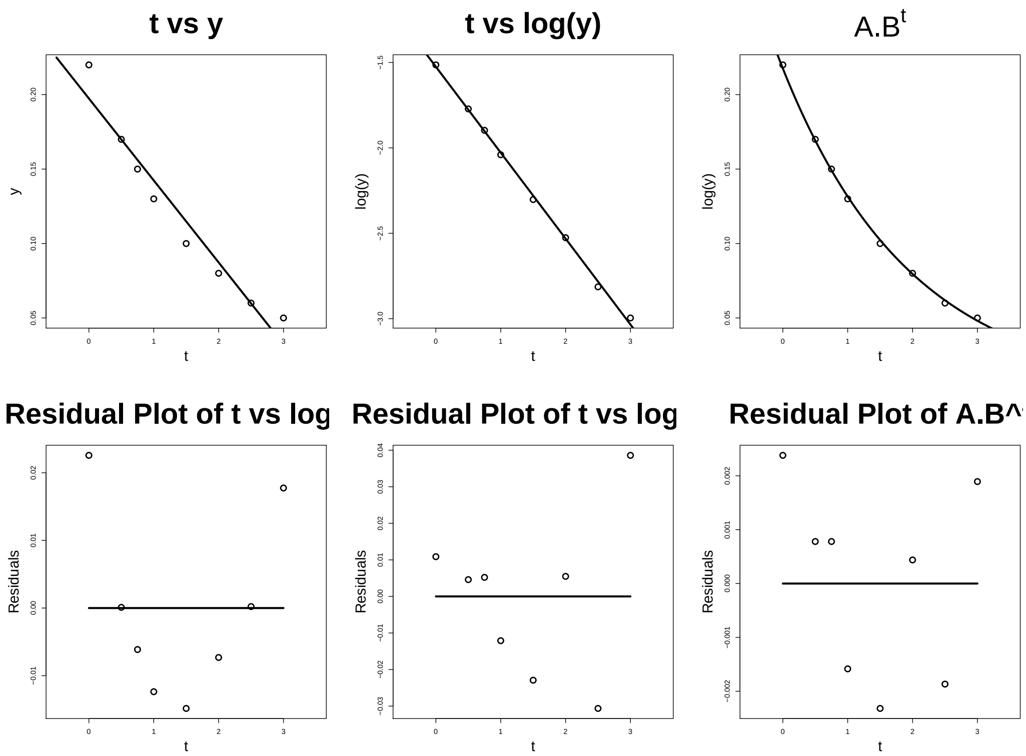
we have

- $S_{xy} = -3.76$
- $S_{xx} = 7.49$

Hence,

- $b = S_{xy}/S_{xx} = -0.5031074$
- $a = \overline{y} - b\overline{x} = -1.525005$
- $A = e^a = 0.2176199$
- $B = e^b = 0.6046488$

The Semi-log Transformation – Plots



Residual plots

- The scatterplot of y and x already indicates whether or not a straight line is a good model.
- A scatterplot of the residuals e against the explanatory variable x gives further insight and is called residual plot:
  - Is there any curvature left?
  - Are there any non horizontal patterns left?

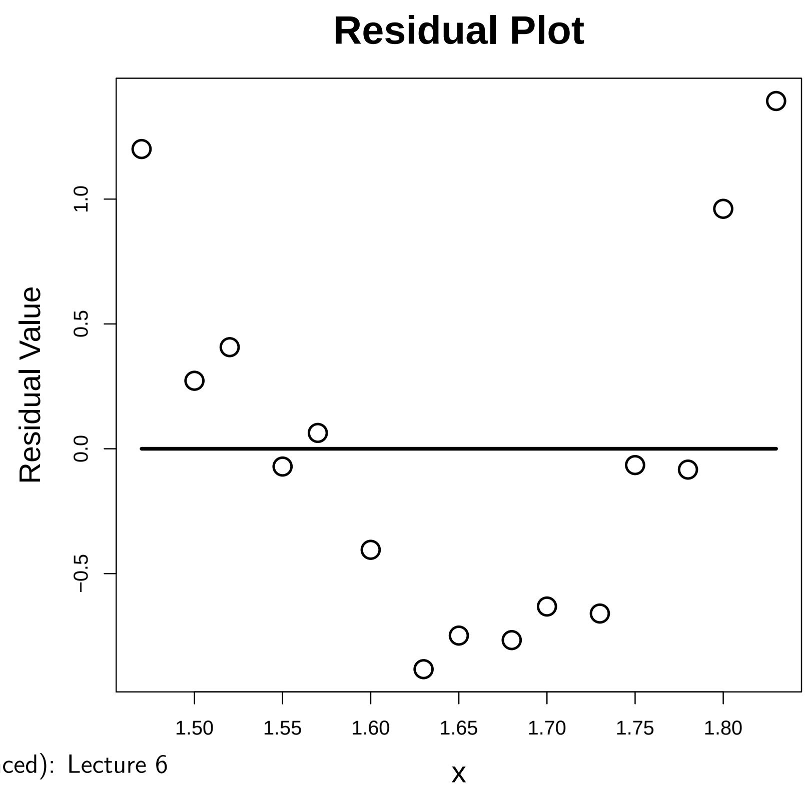
Remarks

- It is a property of the least squares method that

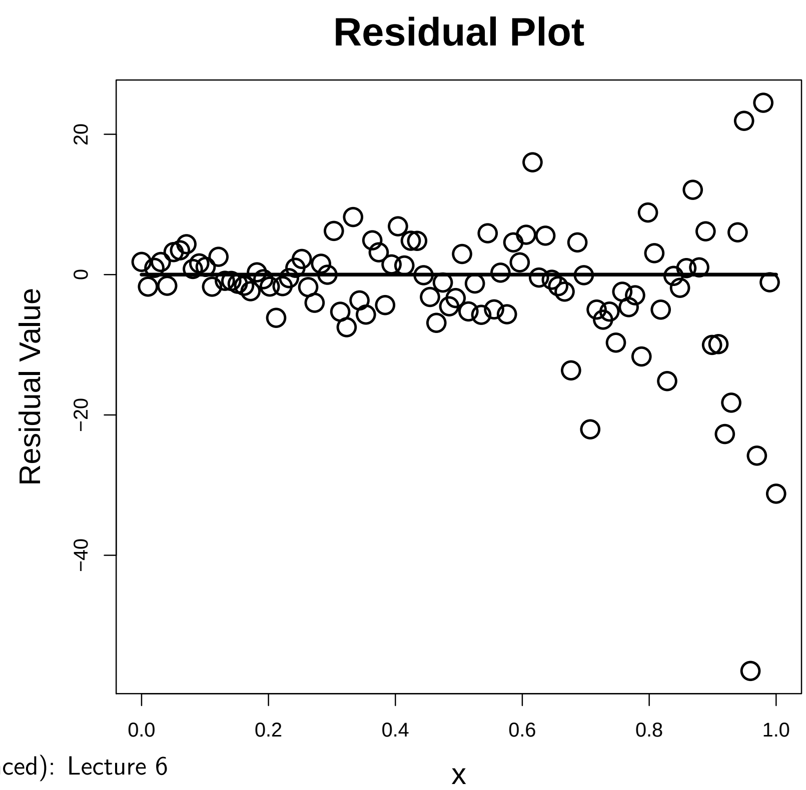
$$\sum_{i=1}^{n} e_i = 0 \;\Rightarrow\; \overline{e}_i = 0$$

  'local' failures, i.e. regions where there is curvature indicate that 'locally' a straight line is not an appropriate model.
- A boxplot and histogram of the residuals e can be drawn to assess symmetry and other aspects of the residuals.
- Overall, residuals should appear randomly scattered about zero.
- Long sequences of positive residuals followed by sequences of negative residuals in e_i vs x_i plot suggests that the error terms are not independent.
- Outliers can severely effect the quality of the fit.

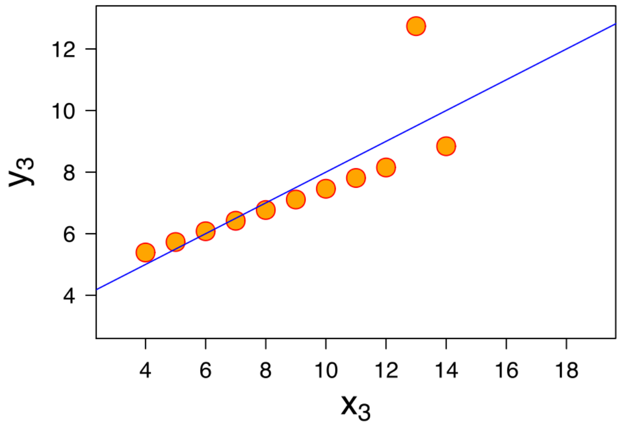
Residual plots - Nonlinear Example

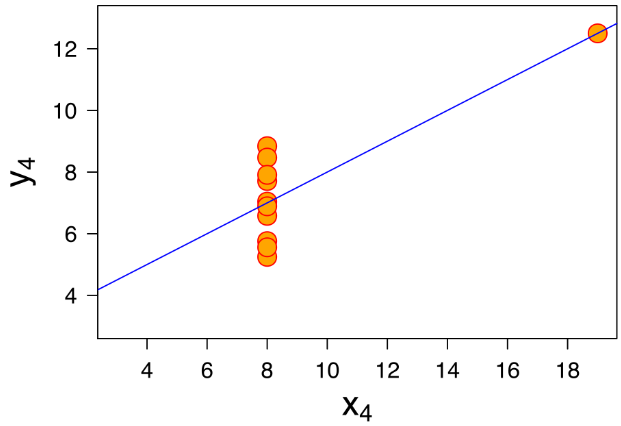


Residual plots - Heteroscedasticity



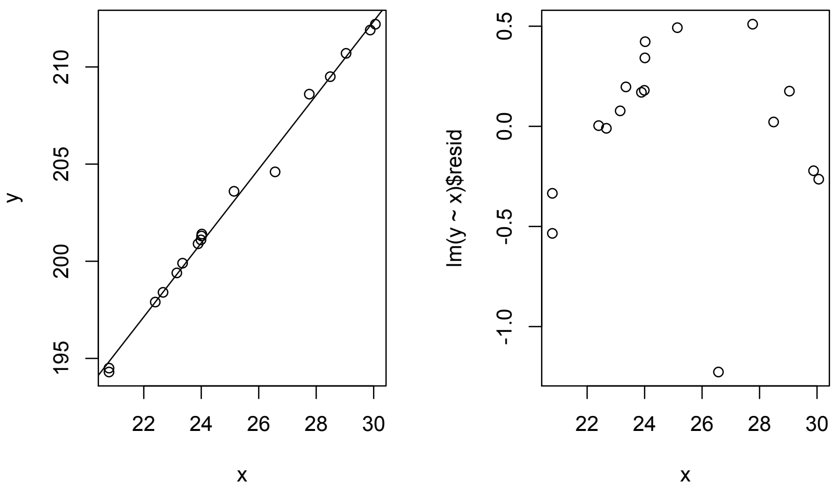
Residual plots - Outlier Example



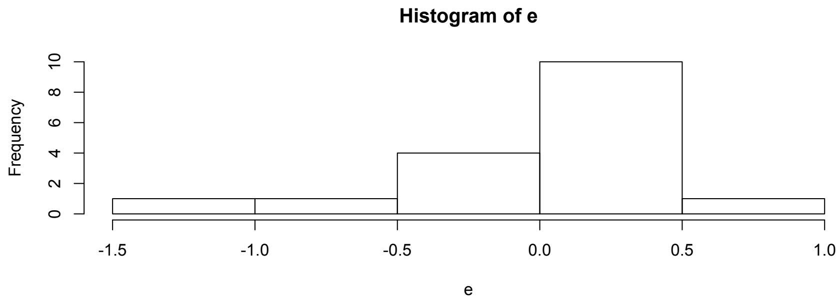


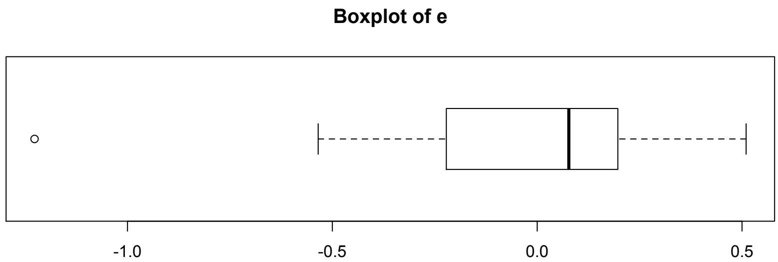
Example (Boiling point, cont). After executing the following lines in R:

```
> par(mfrow=c(1,2));
> plot(x,y)
> abline(coef(lm(y~x)))
> plot(x, lm(y~x)$resid)
```



```
> e = lm(y~x)$resid
> hist(e)
> boxplot(e, main="Boxplot of e")
```





Explaining variability

- The proportion of variability of y's explained by the regression on x is r^2, i.e.

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}}.$$

- The variance of the y's is $s_y^2 = S_{yy}/(n-1)$.
- Recall, $\widehat{y}_i = a + bx_i = \overline{y} + \frac{S_{xy}}{S_{xx}}(x_i - \overline{x})$:

$$\begin{aligned}
\text{RSS} = \sum e_i^2 &= \sum (y_i - \widehat{y}_i)^2 = \sum \left(y_i - \overline{y} - \frac{S_{xy}}{S_{xx}}(x_i - \overline{x}) \right)^2 \\
&= S_{yy} + \frac{S_{xy}^2}{S_{xx}^2} \sum (x_i - \overline{x})^2 - 2\frac{S_{xy}}{S_{xx}} \sum (y_i - \overline{y})(x_i - \overline{x}) \\
&= S_{yy} + \frac{S_{xy}^2}{S_{xx}} - 2\frac{S_{xy}^2}{S_{xx}} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}.
\end{aligned}$$

- Hence, $\frac{S_{yy} - \text{RSS}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = r^2$.