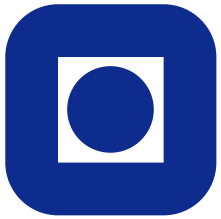


NTNU
Norwegian University of
Science and Technology

TMA4267 Linear Statistical Models V2014 (1)

Introduction to the course
Linear regression [1.1, 1.3]

Mette Langaas

To be lectured: January 6, 2014
wiki.math.ntnu.no/emner/tma4267/2014v/start/

TMA4267 Linear statistical methods

- Linear statistical methods?
- Learning outcome.
- Statistics.
- Background knowledge in probability and inference.
- TMA4267 topics.
- TMA4267 course information.
- Voting and questionnary.
- We start: correlation and simple linear regression.

Central obesity

Mostad et al. (2014): Cental obesity is associated with lower intake of wholegrain bread and less frequent breakfast and lunch. Results from the HUNT Study, an adult all-population survey.

- HUNT: Helseundersøkelsen i Nord-Trøndelag.
- Central obesity: waist circumference, waist-hip-ratio, BMI.
- Diet data: amounts or frequencies. Intake of vegetables, fruits, potato, alcohol, water, milk, soft drinks, sausages, white/wholegrain bread, ...
- Gender, age, physical activity, social status, smoking, ...

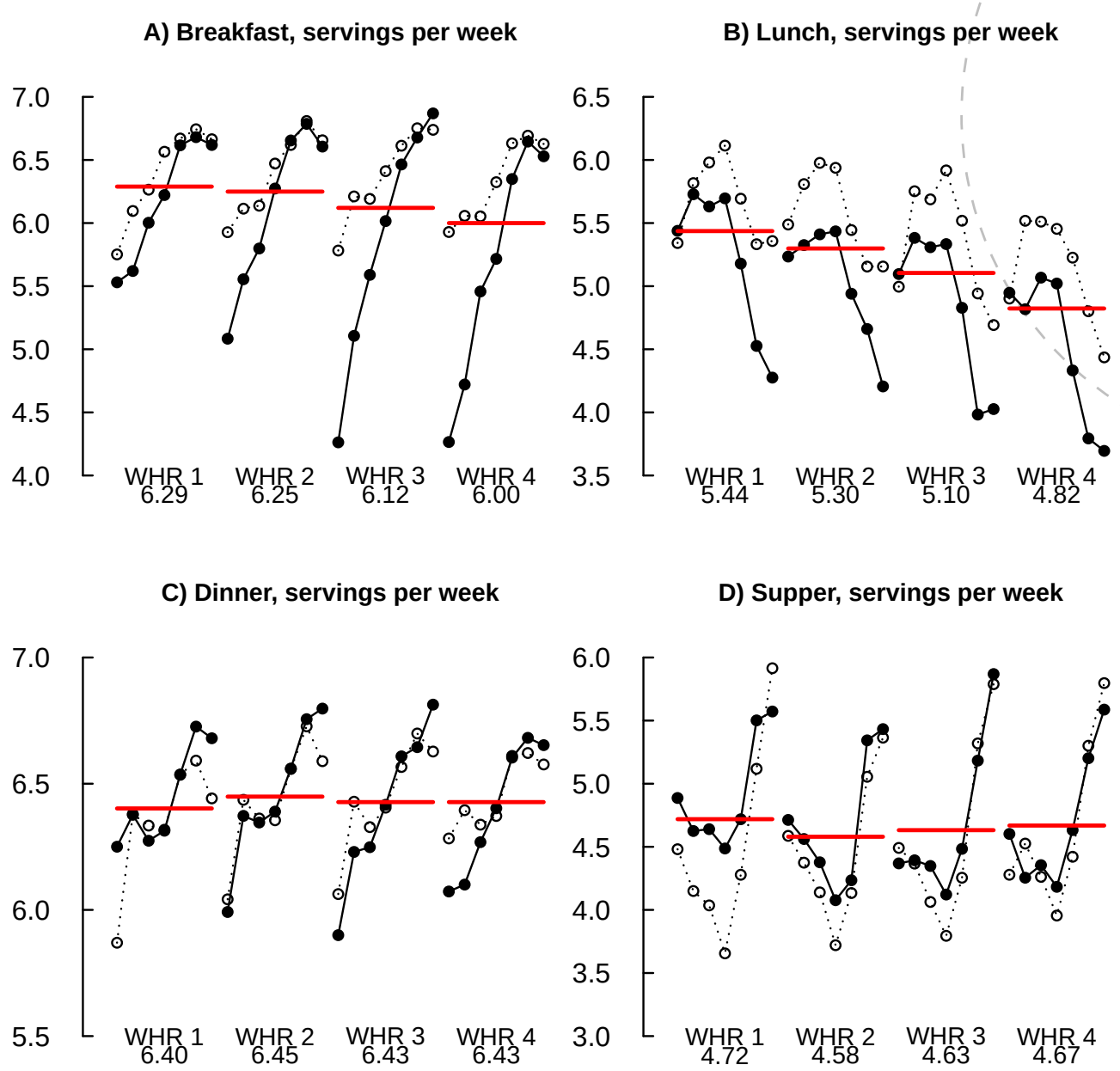

A) Breakfast, servings per week
7.0
6.5
6.0
5.5
5.0
4.5
4.0
WHR 1
6.29
WHR 2
6.25
WHR 3
6.12
WHR 4
6.00
B) Lunch, servings per week
6.5
6.0
5.5
5.0
4.5
4.0
3.5
WHR 1
5.44
WHR 2
5.30
WHR 3
5.10
WHR 4
4.82
C) Dinner, servings per week
7.0
6.5
6.0
5.5
WHR 1
6.40
WHR 2
6.45
WHR 3
6.43
WHR 4
6.43
D) Supper, servings per week
6.0
5.5
5.0
4.5
4.0
3.5
3.0
WHR 1
4.72
WHR 2
4.58
WHR 3
4.63
WHR 4
4.67

Figure caption

The figure depicts results using multiple linear regression with the dietary data as response and WHR group, age group and gender as explanatory variables.

- The vertical axis gives the recoded numerical value of the dietary question under study.
- To explain the effect of performing the current multiple linear regression,
  - divide the HUNT3 participants into 56 groups consisting of all the possible combinations of gender (males or females) and age groups (seven groups: $20 \leq \text{age} < 30$, $30 \leq \text{age} < 40$, $4 \leq \text{age} < 50$, $50 \leq \text{age} < 60$, $60 \leq \text{age} < 70$, $70 \leq \text{age} < 80$, $\text{age} \geq 80$) and WHR-group (four groups).
  - Then calculate the mean dietary response for each of these 56 groups, and depict them as circles (males filled circles females open circles). Lines have been added between circles (solid lines for males, broken for females).

Figure caption

The horizontal axis is used in a hierarchical manner, with WHR-group as the first level and age group as the second level.

- To arrive at the estimated regression coefficients for each WHR group average the 14 mean values (for each gender and for seven age groups) for the WHR group in question.
- In this way both genders and all age groups carry equal weight in the calculation of the regression coefficients for WHR group.
- This averaging procedure stands for the expression “correcting for gender and age group” in the multiple linear regression. The estimated regression coefficients are depicted by horizontal solid red lines.

TMA4267 Linear statistical methods Learning outcome, Knowledge

— The student has strong theoretical knowledge about the most popular statistical models and methods that are used in science and technology, with emphasis on regression-type statistical models.

— The statistical properties of the multivariate normal distribution are well known to the student, and the student is familiar with the role of the multivariate normal distribution within linear statistical models.

TMA4267 Linear statistical methods Learning outcome, Skills

- The student knows how to design an experiment and
- how to collect informative data of high quality to study a phenomenon of interest.
- Subsequently, the student is able to choose a suitable statistical model,
- apply sound statistical methods, and
- perform the analyses using statistical software.
- The student knows how to present the results from the statistical analyses, and how to draw conclusions about the phenomenon under study.

Statistics

Originally: collection and presentation of data
Today: much more!

- Design and collection of data from statistical investigations.
- Build a model for the random (stochastic) mechanisms behind the data.
- Drawing conclusions about this mechanisms based on using statistical methods to analyse the data.
- Evaluating the strength of the conclusions (effect sizes, *p*-values, confidence intervals, power).

New York Times, August 2009

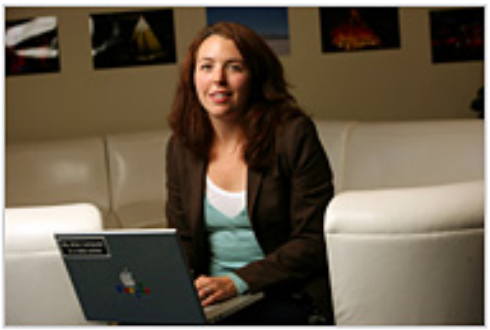
For Today's Graduate, Just One Word: Statistics

By STEVE LOHR
Published: August 5, 2009

MOUNTAIN VIEW, Calif. — At Harvard, Carrie Grimes majored in anthropology and archaeology and ventured to places like Honduras, where she studied Mayan settlement patterns by mapping where artifacts were found. But she was drawn to what she calls "all the computer and math stuff" that was part of the job.

COMMENTS (58)
SIGN IN TO E-MAIL
PRINT
REPRINTS
SHARE

Enlarge This Image



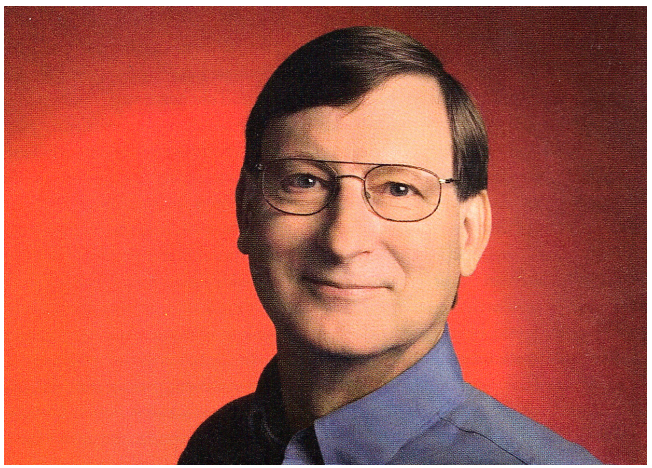
Thor Swift for The New York Times

"People think of field archaeology as Indiana Jones, but much of what you really do is data analysis," she said.

Now Ms. Grimes does a different kind of digging. She works at Google, where she uses statistical analysis of mounds of data to come up with ways to improve its search engine.

Google: Hal Varian

"I keep saying that the sexy job the next 10 years is statistician".



Significance, March 2011.

The data revolution is upon us. The data we have and the way we treat it has changed beyond measure – as witness a quote from Hal Varian of Google: "Back in the early days of the Web, every document had at the bottom, 'Copyright 1997. Do not redistribute.' Now every document has at the bottom, 'Copyright 2008. Click here to send to your friends.'" Hal Varian has made another famous quote about statistics in the new data age. **Julian Champkin** interviewed him.

Hal Varian is not actually a statistician himself. He is an economist. He is in fact the chief economist at Google. As such he is the spokesman for the organisation which is presumably the biggest transmitter, organiser, analyser and general handler of data that the world has ever seen. He is also of course the man who said: "The sexy profession of the next decade will be statistician." He made the quote in 2008, and it will not let him go.

Do you know this?

- Probability: random variables, probability distribution function (pdf), cumulative distribution function (cdf), mean E, variance Var, covariance Cov, correlation Corr, momentgenerating function (MFG), Normal, (chisq), t and (F-distribution).
- Inference: population and sample philosophy, parameter estimation, confidence interval, hypothesis test, *p*-value, power.
- Linear methods: vector and matrix algebra, real vector spaces, orthogonality.

TMA4267 topics

Some probability + mostly applied statistical inference.

- Simple linear regression and the bivariate Normal.
- Working with random vectors and matrices.
- Multivariate normal distribution.
- Multiple linear regression (MLR).
- Analysis of variance (quadratic forms and connections to DFs).
- Extensions and practical issues with MLR.
- Experimental design (special case of MLR).
- Model selection and regularization (how to select or handle many covariates).
- Contingency tables (goodness of fit, and from continuous to categorical data).
- If time permits: introduction to generalized linear and mixed models.

TMA4267 Linear Statistical Models
Course information

https://wiki.math.ntnu.no/tma4267/2014v/start

- Course literature and reading list.
- Lectures and handouts.
- Statistical software.
- Exercises.
- Compulsory project.
- Exam

Electronic voting

– more than an anonymous show of hands?

For student: check that topics are understood, compare to class, focus on the question asked, while preserving anonymity.

For lecturer: collect data to design sessions that are more contingent.

Future studies?

What is your current plan of topic for future studies?

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don’t know

Electronic voting

Use your smart phone, or other devise with internet access and go to **http://clicker.math.ntnu.no/**, and then select TMA4267 as classroom.
Answers

- A: Statistics
- B: Mathematics
- C: Numerics
- D: Other
- E: Don’t know

Start voting now!

Start-up questionnary in TMA4267

https://innsida.ntnu.no/sso/?target=EvalProd&returnargs=16731
Link posted at the course www-page (under messages) and at It's learning.

Linear regression

Bingham and Fry (2010): 1.1-1.4

- Covariance, correlation.
- Galton height data.
- Least squares.
- Simple linear regression.
- Correlation vs. regression slope?

RVs X and Y

X and Y are RVs (random variables) with joint pdf $f(x, y)$ and marginal means μ_X and μ_Y.

The covariance of X and Y is

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\ &= E(X \cdot Y) - \mu_X \mu_Y \end{aligned}$$

The correlation coefficient of X and Y is

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}$$

From population to sample

A random sample from the (X,Y) population is given as $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.
Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
Define the following sums of squares:

- $S_{XX} = \sum_{i=1}^{n}(X_i - \bar{X})^2$
- $S_{YY} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$
- $S_{XY} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$

The sample covariance and correlations are given as (note 1/n, not 1/(n-1))

$$\widehat{\mathrm{Cov}(X, Y)} = \frac{1}{n} S_{XY}$$

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$