APPLICATIONS OF STATISTICS

Martina Vandebroek

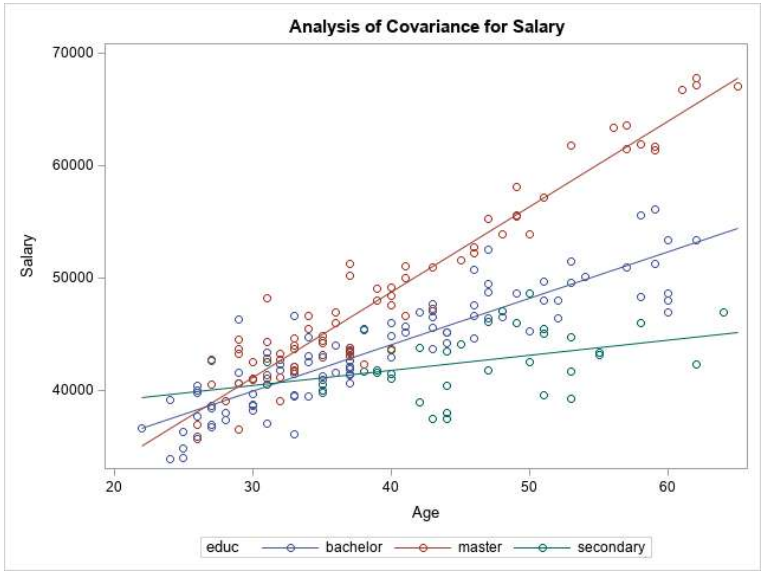
KU Leuven 2023-2024

TABLE OF CONTENTS:

EXAMPLES

regression: predict yearly salary based on age and education (the highest obtained degree: secondary, bachelor or master)
(with dummy coding)

$$\begin{aligned} \widehat{\text{salary}} &= 36362 + 136 * age \\ &- 8688 * bachelor \\ &- 17990 * master \\ &+ 274 * age * bachelor \\ &+ 623 * age * master \end{aligned}$$



two-way analysis of variance:

which combination of packaging and advertisement is best?
collect sales data from similar shops:

	Adv. 1	Adv. 2	Adv. 3	Adv. 4
Pack. 1	18000	10000	12000	24000
	17500	9000	13000	22000
Pack. 2	8000	5000	12000	14000
	9000	7000	10000	13000
Pack. 3	15000	11000	14000	12000
	16000	13000	14000	13000

are the differences between the means significant?

econometrics:

- what if nonlinear relationship?
- what if heteroscedasticity?
- what if autocorrelation?
- what if measurement error?
- what if regressors are stochastic?
- what if endogeneity?
- what if the time series are nonstationary?
- what if several units are measured at several time points?

logistic regression:

banks need to discriminate between good and bad clients such that only good clients are granted a loan based on salary, age, education, family size, ... (= *credit scoring*)

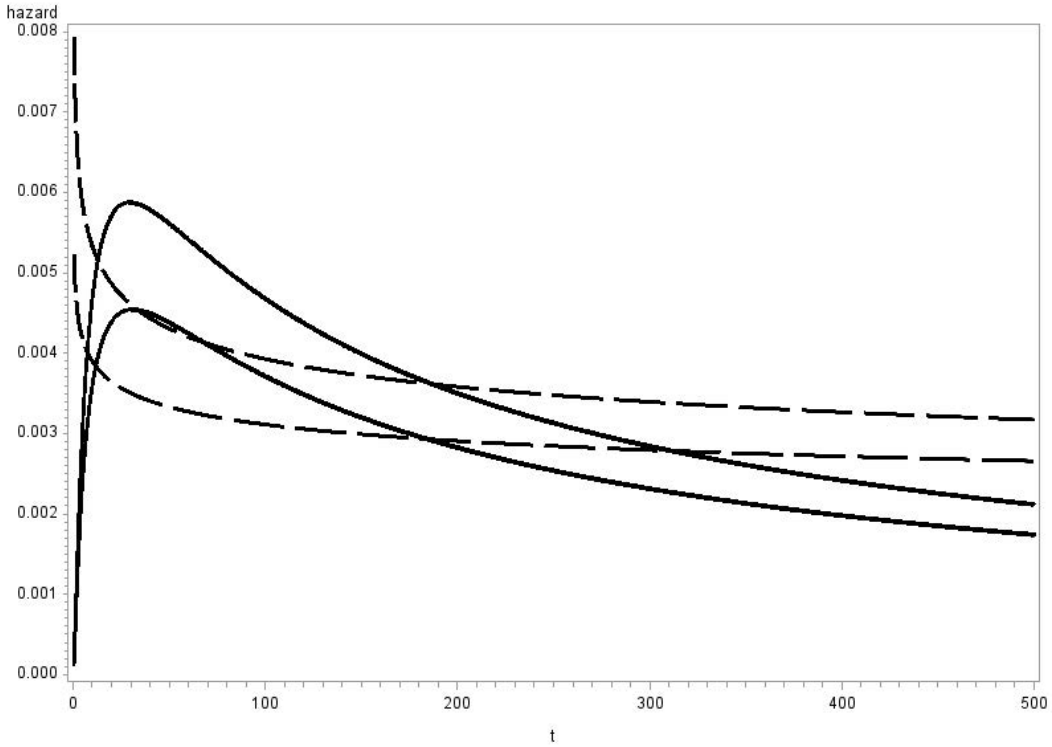
$\Rightarrow$ model the probability of being creditworthy
$\Rightarrow$ test the significance of the variables
$\Rightarrow$ use predicted probability to classify people

if more than 2 possible response values:

- ordered response values: cumulative logit or proportional odds
- no ordening in response values or categories:
  - multinomial logit model
  - conditional logit model

duration analysis (survival analysis):

model the time until an event
(for instance until finding a job, until the purchase of a new car, ...)



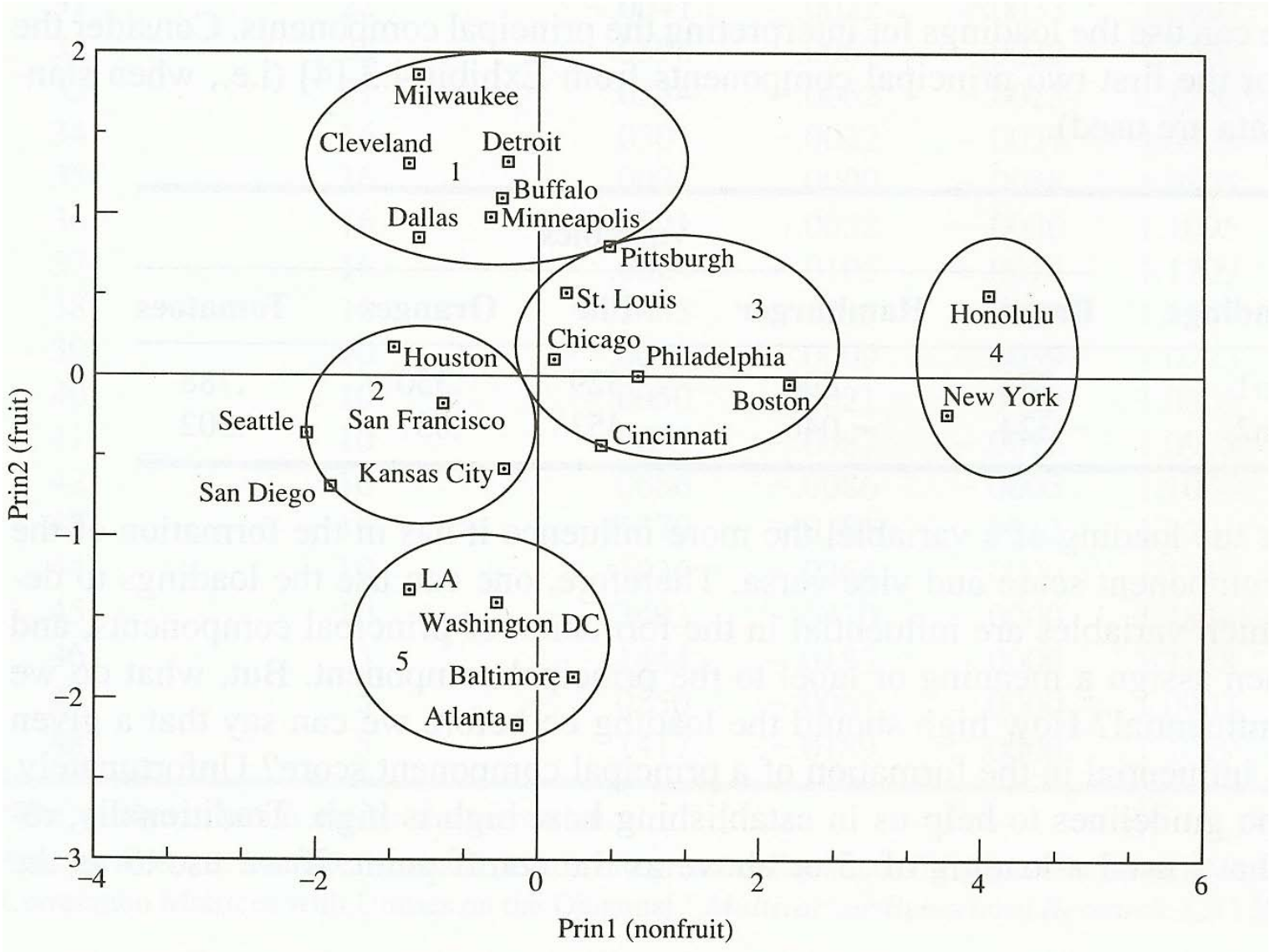
weibull: with reorientation course
weibull: without reorientation course
lognormal: with reorientation course
lognormal: without reorientation course

h(t) is the probability that someone that has not yet found a job at t, will find a job at t

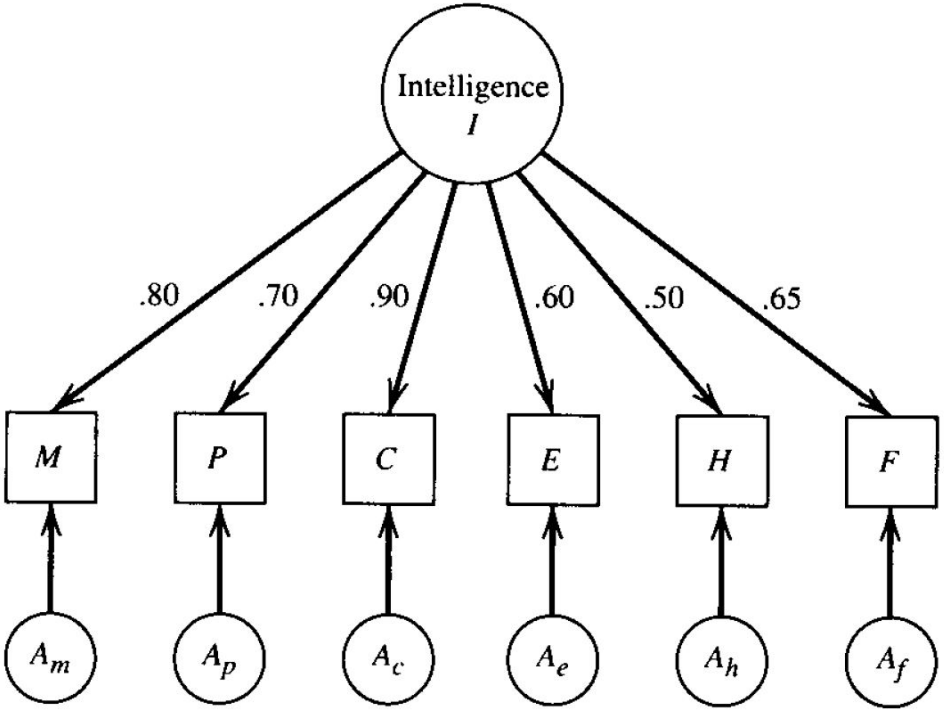
principal component analysis:

food price data : average price (in cents per pound)

city	Bread	Burger	Milk	Oranges	Tomatoes
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91.0	67.5	74.6	53.3
Boston	29.7	100.8	61.4	104.0	59.6
Buffalo	22.8	86.6	65.3	118.4	51.2
Chicago	26.7	86.7	62.7	105.9	51.2
Cincinnati	25.3	102.5	63.3	99.3	45.6
Cleveland	22.8	88.8	52.4	110.9	46.8
Dallas	23.3	85.5	62.5	117.9	41.8
Detroit	24.1	93.7	51.5	109.7	52.4
Honolulu	29.3	105.9	80.2	133.2	61.7
Houston	22.3	83.6	67.8	108.6	42.4
Kansas City	26.1	88.9	65.4	100.9	43.2
Los Angeles	26.9	89.3	56.2	82.7	38.4
Milwaukee	20.3	89.6	53.8	111.8	53.9
Minneapolis	24.6	92.2	51.9	106.0	50.7
New York	30.8	110.7	66.0	107.3	62.6
Philadelphia	24.5	92.3	66.7	98.0	61.7
Pittsburgh	26.2	95.4	60.2	117.1	49.3
St. Louis	26.5	92.4	60.8	115.1	46.2
San Diego	25.5	83.7	57.0	92.8	35.4
San Francisco	26.3	87.1	58.3	101.8	41.5
Seattle	22.5	77.7	62.0	91.1	44.9
Washington, DC	24.2	93.8	66.0	81.6	46.2


Milwaukee
Cleveland
Detroit
1
Buffalo
Dallas
Minneapolis
Pittsburgh
St. Louis
3
Chicago
Philadelphia
Houston
2
Boston
Seattle
San Francisco
Cincinnati
Kansas City
San Diego
Honolulu
4
New York
LA
Washington DC
5
Baltimore
Atlanta
Prin2 (fruit)
Prin1 (nonfruit)
2
1
0
−1
−2
−3
−4
−2
0
2
4
6

exploratory factor analysis: find the underlying dimension(s) that can explain the correlation among observed variables



discriminant analysis:

banks need to discriminate between good and bad clients such that only good clients are granted a loan (= *credit scoring*)

based on salary, age, education, family size, ...

$\Rightarrow$ describe the differences between good and bad clients

$\Rightarrow$ construct a classification rule

$\Rightarrow$ classify new clients as good or bad

(similar goal as logistic regression but different methodology)

(we only deal with 2 group discriminant analysis)

cluster analysis:

customer segmentation for personalized marketing based on

- family composition
- type of job
- leisure activities
- ...

$\Rightarrow$ divide customers in homogeneous groups

$\Rightarrow$ develop a different advertisement leaflet for each group

practical information:

- theory and examples using R (recordings)
- PC-sessions for problems with R
- (online) Q&A sessions on theory and R code
- extra applications to practice

- after each part an individual online task using TOLEDO (limited time to solve and submit!)(each graded on 1/20)

exam in January:

- a written open book exam in the exam period (16/20)
- result = max (exam + result on tasks ; exam result*20/16)

retake exam in August:

- a written open book exam in the exam period (16/20)
- a practical exam in PC-room analysing data with R (4/20)
- result = result on open book exam + result on practical exam

all the information can be found on TOLEDO/ATSTAT:

- recordings, datasets, ...
- discussion forum
- Q & A sessions
- the graded tasks
- the (continuously updated) list with errata
- (if necessary) changes in the schedule
- what material you can bring to the exam
- . . .

please inform me by mail (martina.vandebroek@kuleuven.be) about urgent practical problems (with the recordings, tasks, ...)

over the years I have consulted innumerable books and websites, a very brief selection:

part 1:

- Kleinbaum, D., Kupper, L., Nizam, A., Rosenberg, E.S., 2014, Applied regression analysis and other multivariable methods.
- Nachtsheim, C.J., Neter, J., Kutner M., Li W., 2006, Applied linear statistical models.

part 2:

- Hill, R.C., Griffiths, W.E., Lim, G.C., 2008, Principles of Econometrics.
- Stock J.H. and Watson, M.W., 2003, Introduction to econometrics.

part 3:

- Agresti Alan, 2013, Categorical Data Analysis.
- Allison, Paul, 2005, Survival analysis using SAS.

part 4:

- Johnson, R.A., Wichern, D.W., 2007, Applied multivariate statistical analysis.
- Tabachnick and Fidell, 2014, Using multivariate statistics.

PART 1

- REGRESSION
- ANOVA

1a. REGRESSION

REGRESSION

one distinguishes

- observational or empirical data
- experimental data (collected during a designed experiment)

and

- cross-sectional data
- time series data
- panel data (combines cross-sectional and time series data)

the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

y = dependent or response variable

β_i = parameters

x_i = independent, explanatory or regressor variables

u = stochastic residual or error term

assuming errors with mean zero, we can predict the mean response:

$$\mathbf{E}(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

example: the energy required for heating a house (kWh per year) depends on the size of the house (sq meter) and on the average outside temperature (degree C); assume the following relationship holds:

$$\text{energy} = 10000 + 200\ \text{size} - 400\ \text{temperature} + u$$

- the larger the house, the more energy is required: for a given outside temperature, an extra square meter increases the energy required on average with 200 kWh

- the higher the average outside temperature, the lower the need for energy; given the house size, the energy decreases on average with 400 kWh for each extra degree celcius

we consider only models that are linear in the *parameters*!

not OK:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \exp(\beta_2 x_2) + u \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2^{\beta_3} + u \end{aligned}$$

OK:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u \\ \ln(y) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u \\ y &= \beta_0 + \beta_1 x_1 + \beta_2 \exp(x_2) + u \end{aligned}$$

take a sample of n observation:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + u_i \qquad i = 1, \ldots, n$$

or in matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1k} \\ x_{20} & x_{21} & \cdots & x_{2k} \\ \vdots & & & \\ x_{n0} & x_{n1} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

with for all i: $\quad x_{i0} = 1$

let:

$\mathbf{Y}$: a column vector with response values

$\mathbf{X}$: the design matrix

$\boldsymbol{\beta}$: a column vector with parameters

$\mathbf{U}$: a column vector with residuals

then the model becomes:

$$\underset{(n\times 1)}{\mathbf{Y}} = \underset{(n\times(k+1))}{\mathbf{X}} \underset{((k+1)\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{U}}$$

Ordinary Least Squares (OLS) estimators $\mathbf{b}$ for $\boldsymbol{\beta}$

- minimize the sum of squared residuals

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - \ldots - b_k x_{ik})^2 = \min_{\mathbf{b}} \sum_{i=1}^{n} \hat{u}_i^2$$

$$\min_{\mathbf{b}} (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \hat{\mathbf{U}}'\hat{\mathbf{U}}$$

- taking derivates w.r.t. b_i gives the *normal equations*:
  $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}^*$

- in case $\mathbf{X}'\mathbf{X}$ is invertible, then $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$

*remark that $\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i'$ with $\mathbf{x}_i$ the ith row of $\mathbf{X}$ put in a column vector

- $\mathbf{b}$ is often denoted as $\hat{\boldsymbol{\beta}}$
- $\boldsymbol{\beta}$ denotes the unknown parameters that are estimated by $\mathbf{b}$
- $\mathbf{b}$ can denote a value (in case a sample has been taken) or can be considered a random variable which will take a different value for each sample (the following slides have the properties of this stochastic variable under specific assumptions)
- check the applet that illustrates the difference between the population and the sample line

properties of OLS estimators:

assumptions on the residuals*:

(A1) $E(u_i) = 0 \ (\forall i)$

(A2) $var(u_i) = \sigma^2 \ (\forall i)$
(homoscedasticity = equal variance for all residual terms)

(A3) $cov(u_i, u_j) = 0 \ \ (\forall i \neq j)$ (no serial or autocorrelation)

*these assumptions are called the Gauss-Markov assumptions

under assumptions **(A1), (A2), (A3)**:

- the OLS estimators are **unbiased**:

$$\begin{aligned}\mathbf{E}(\mathbf{b}) &= \mathbf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) &&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}(\mathbf{Y}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{E}(\mathbf{U})) &&= \boldsymbol{\beta}\end{aligned}$$

- the variance-covariancematrix of the estimators is:

$$\begin{aligned}\mathbf{cov}(\mathbf{b}) &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \;\; \mathbf{cov}(\mathbf{Y}) \;\; \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \;\; \left(\sigma^2\mathbf{I}\right) \;\; \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\,(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

as the OLS estimators are unbiased and their variances $\overset{n\to\infty}{\to} 0$,
the OLS estimators are **consistent***

*an estimator is consistent if its sample distribution gets more and more concentrated around the true value (so at least the asymptotic bias should be zero) and if its variance goes to zero too as n increases

if there is only 1 independent variable (*simple regression*):

$$
\begin{aligned}
var(b_0) &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}_1^2}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2} \right) \\
var(b_1) &= \frac{\sigma^2}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2} \\
cov(b_0, b_1) &= \frac{-\bar{x}_1 \sigma^2}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2} \text{ with } \bar{x}_1 = (\sum_{i=1}^{n} x_{i1})/n
\end{aligned}
$$

check the applet which shows the distribution of the intercept, the slope and their correlation

statistical inference

draw conclusions about β_i based on b_i by computing confidence intervals and/or testing hypotheses

we need an extra assumption:

(A4) $u_i \sim N\ (\forall i)$: the residuals are normally distributed*

the responses are then also normally distributed and as the estimators are linear combinations of the responses, also $b_0, b_1, ...b_k$ are normally distributed; consequently:

$$\mathbf{b} \sim N\left(\boldsymbol{\beta}, \sigma^2\ (\mathbf{X}'\mathbf{X})^{-1}\right)$$

*in case the residuals have another distribution, most of the results still hold asymptotically

remark that

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Total Sum of Squares: total variation in the response

$$SSTO = \sum_{i=1}^{n}(y_i - \bar{y})^2 \quad \text{with } \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Regression S.S.: variation explained by the model (as $\bar{\hat{y}} = \bar{y}$)

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \quad \text{with } \hat{y}_i = b_0 + b_1 x_{i1} + \ldots + b_k x_{ik}$$

Error S.S.: unexplained variation

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{u}_i^2 = \hat{\mathbf{U}}'\hat{\mathbf{U}}$$

hypothesis testing:

- consider the general hypothesis:

$$\mathbf{H_0} : \mathbf{L}'\boldsymbol{\beta} = \mathbf{h} \qquad \text{versus} \qquad \mathbf{H_a} : \mathbf{L}'\boldsymbol{\beta} \neq \mathbf{h}$$

  with $\mathbf{L}$: a $((k+1) \times l)$ matrix having rank $q \leq l$
  (= q linear independent restrictions on the parameters)

- general idea: compare the unexplained variation of the response in the full model and in the model under the H_0;
  accept* H_0 if the difference is small

*strictly speaking, one does not *accept the null hypothesis* but one *does not reject the null hypothesis*

let:

- $SSE(H_0)$ be the unexplained variation in the model that is reduced by H_0
- $SSE(F)$ the unexplained variation in the full model*

compare

SSE(H_0) - SSE(F)

if the difference is *relatively* small, we can accept H_0

to judge whether it is small, we need the distribution of this test statistic under the H_0

*in most cases the (F) is deleted in $SSE(F)$, so SSE stands for the SSE of the full model

under the null hypothesis:

$$\frac{\frac{SSE(H_0)-SSE(F)}{q}}{\frac{SSE(F)}{n-k-1}} = \frac{\frac{SSE(H_0)-SSE}{q}}{MSE} \overset{H_0}{\sim} F_{q,n-k-1}$$

this result tells us what values we can expect under H_0; if we get a value that is larger than we expect, we reject H_0

in case $(\mathbf{X}'\mathbf{X})$ is invertible, this teststatistic can be shown to be

$$\frac{(\mathbf{L}'\mathbf{b}-\mathbf{h})'(\mathbf{L}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L})^{-1}(\mathbf{L}'\mathbf{b}-\mathbf{h})}{q\ MSE}$$

it can be shown that $\mathbf{E}(MSE) = \sigma^2$, so $\hat{\sigma}^2 = MSE$ and

$$\widehat{\mathbf{cov}(\mathbf{b})} = MSE\ (\mathbf{X}'\mathbf{X})^{-1}$$

some frequently used tests:

- test the statistical significance of the model:
  is the model significantly better than the empty model?*

$\mathbf{H_0}: \beta_1 = \ldots = \beta_k = 0$ versus $\mathbf{H_a}:$ not all $\beta_i = 0$ $(i = 1 \ldots k)$

*remark that a significant model is not necessarily explaining a lot, the latter is assessed by $R^2 = SSR/SSTO$, the percentage of the variance in the response that is explained by the model

under H_0 we have that $\hat{y}_i = \bar{y}$ and therefore $SSR(H_0) = 0$ and $SSE(H_0) = SSTO$

the test statistic becomes:

$$\frac{\frac{SSTO - SSE(F)}{k}}{\frac{SSE(F)}{n-k-1}} = \frac{\frac{SSR(F)}{k}}{\frac{SSE(F)}{n-k-1}} = \frac{MSR}{MSE} \overset{H_0}{\sim} F_{k,n-k-1}$$

- MSR the Mean Squares due to Regression
- MSE the Mean Squares due to Error

check the applet which illustrates the distribution of the teststatistic under H_0 and H_a

- test the significance of the contribution of x_l:

  $\mathbf{H_0}$: $\beta_l = 0$ versus $\mathbf{H_a}$: $\beta_l \neq 0$

  the test statistic is

  $$\frac{\frac{SSE(H_0)-SSE}{1}}{\frac{SSE}{n-k-1}} \overset{H_0}{\sim} F_{1,n-k-1}$$

  with $SSE(H_0)$ the unexplained variation in the response when x_l is deleted from the model

– it tests whether the variable x_l can explain a significant part, *given that the other explanatory variables are already in the model*
– so it tests only the *marginal* explanatory value of the variable

it can be shown that this test is equivalent with the t-test*:

$$t = \frac{b_l - 0}{s(b_l)} \overset{H_0}{\sim} t_{n-k-1}$$

remark that with the F-test it is much more clear that only the marginal contribution of the variable is tested

*remember that the square of a t_n distribution is $F_{1,n}$ distributed

example: can building land prices be explained by the plot size and the distance to Brussels?

$$\begin{aligned} \text{price} &= \beta_0 + \beta_1 \text{ size} + \beta_2 \text{ distance} + u \\ \widehat{\text{price}} &= 963.04 + 0.32 \text{ size} - 19.17 \text{ distance} \end{aligned}$$

remember*

$$R^2 = \frac{SSR}{SSTO} = corr^2(y, \hat{y}) \quad \text{coefficient of determination}$$

$$R^2_{adj} = 1 - \frac{(n-1)(1-R^2)}{n-k-1} \quad \text{adjusted } R^2$$

$$\text{coeffient of variation} = \frac{\sigma}{\mu}, \text{ estimated by } \frac{\sqrt{MSE}}{\bar{y}}$$

*R^2 is often called *the proportion of the variance in the response that is explained by the model*

RESULTS

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2081885.632	1040942.816	12.74	0.0009
Error	13	1062157.554	81704.427		
Corrected Total	15	3144043.186			

R-Square	Coeff Var	Root MSE	price Mean
0.662168	40.3328	285.8399	708.7028

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	963.0393163	178.4908291	5.40	0.0001
size	0.3179532	0.1055010	3.01	0.0100
distance	-19.1682473	5.1574339	-3.72	0.0026

example: take into account the interaction between size and distance to Brussels

$$\text{price} = \beta_0 + \beta_1 \text{ size} + \beta_2 \text{ distance} + \beta_3 \text{ size * distance} + u$$

here:

$$\widehat{\text{price}} = 544.64 + 0.96 \text{ size} - 1.83 \text{ distance} - 0.03 \text{ size * distance}$$

so the price increases with size and this increase is less strong if the plot is further from Brussels: when the distance to Brussels is 10 km, the price increases by 0.96-0.30= 0.66 for each extra size unit; if the distance is 20 km, this increase is only 0.96-0.60 = 0.36 for each extra size unit

RESULTS

Dependent Variable: price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3005485.855	1001828.618	86.77	<.0001
Error	12	138557.331	11546.444		
Corrected Total	15	3144043.186			

R-Square	Coeff Var	Root MSE	price Mean
0.955930	15.16212	107.4544	708.7028

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	544.64273644	81.79712734	6.66	<.0001
size	0.96095641	0.08210823	11.70	<.0001
distance	-1.83034808	2.74171178	-0.67	0.5170
size*distance	-0.02771892	0.00309926	-8.94	<.0001

type I, II, III Sum of Squares

there are different ways to compute the contribution of each variable in explaining the variation of the response

type-I SS (sequential)

- the increase in SSR by sequentially adding variables to the model
- as SSTO $=$ SSR $+$ SSE and SSTO is independent of the model, it also gives the decrease in SSE by sequentially adding variables to the model
- remark that these SS depend on the order of the variables in the model

example* :

$$\begin{aligned} SS(x_1) &\equiv SSR(x_1) = SSTO - SSE(x_1) \\ SS(x_2|x_1) &\equiv SSR(x_1, x_2) - SSR(x_1) \\ &= SSE(x_1) - SSE(x_1, x_2) \\ SS(x_1 * x_2|x_1, x_2) &\equiv SSR(x_1, x_2, x_1 * x_2) - SSR(x_1, x_2) \\ &= SSE(x_1, x_2) - SSE(x_1, x_2, x_1 * x_2) \end{aligned}$$

this is the default sum of squares in anova() in R although it is rarely the best choice!

*$SSE(x_1)$ is the unexplained variation in the model with only x_1 next to the intercept

type II SS

- gives the decrease in SSE by adding a variable to the model which contains all the other variables except the interaction terms including that variable

- mainly used when there is no (significant) interaction

$$
\begin{aligned}
SS(x_1|x_2) &= SSR(x_1, x_2) - SSR(x_2) \\
SS(x_2|x_1) &= SSR(x_1, x_2) - SSR(x_1) \\
SS(x_1 * x_2|x_1, x_2) &= SSR(x_1, x_2, x_1 * x_2) - SSR(x_1, x_2)
\end{aligned}
$$

type III SS (partial)

- gives the increase in SSR (or decrease in SSE) by adding a variable to the model which contains all the other variables

- gives the marginal explanatory value of a variable:

$$
\begin{aligned}
SS(x_1|x_2, x_1 * x_2) &= SSR(x_1, x_2, x_1 * x_2) - SSR(x_2, x_1 * x_2) \\
SS(x_2|x_1, x_1 * x_2) &= SSR(x_1, x_2, x_1 * x_2) - SSR(x_1, x_1 * x_2) \\
SS(x_1 * x_2|x_1, x_2) &= SSR(x_1, x_2, x_1 * x_2) - SSR(x_1, x_2)
\end{aligned}
$$

- for models without interaction, type II = type III

- for models with interaction, we will use the type II sums of squares for the following reasons:
  - if the interaction is significant, that interaction has to be described and analysed further, the main effects should not be tested in that case
  - if the interaction is not significant, we can use the tests for the main effects in type II SS as we do not have to correct for the (not significant) interaction
  - type II SS do not depend on the coding scheme of the categorical variables (see later)
  - testing main effects with type II SS has been shown to be more powerful that testing them with type III SS

multicollinearity

in case the explanatory variables are (highly) correlated , the different types of SS will be (very) different; see for instance the following dataset with $corr(x_1, x_2) \approx 1$:

given x_1 is in the model, x_2 cannot contribute much and vice versa

y	x_1	x_2
42	4	2
39	4	2
48	4	2
51	4	2
49	6	3
53	6	3
61	6	3.01
60	6	3

$x_1 \approx 2x_2 \Rightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \approx \beta_0 + (2\beta_1 + \beta_2)x_2 + u$

and so only $(2\beta_1 + \beta_2)$ can precisely be estimated

Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	SSR
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$	27.000	-345.500	700.000	267.875
$y = \beta_0 + \beta_1 x_1 + u$	23.500	5.375	-	231.125
$y = \beta_0 + \beta_2 x_2 + u$	23.489	-	10.749	232.244

type I: $SS(x_1) = SSR(x_1) = 231.125$

type II: $SS(x_1|x_2) = SSR(x_1, x_2) - SSR(x_2) = 35.631$

so the contribution of x_1 is very different in the models with and without x_2 (and the same holds for x_2) and we can compare type I and type II sum of squares to detect strong multicollinearity

another check for multicollinearity: **Variance Inflation Factor**

$$(VIF)_i = \frac{1}{1 - {R_i}^2} \quad (i = 1 \ldots k)$$

with ${R_i}^2$ the coefficient of determination of

$$x_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_{i-1} x_{i-1} + \beta_{i+1} x_{i+1} + \ldots \beta_k x_k + u$$

when x_i is not correlated with the other explanatory variables at all, ${R_i}^2 = 0$ and $VIF_i = 1$

$VIF > 10$ indicates there is serious multicollinearity

for the land price example:

RESULTS

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	44.64274	81.79713	0.55	0.5952	0
size	1	0.96096	0.08211	11.70	<.0001	4.33224
distance	1	-1.83035	2.74171	-0.67	0.5170	2.02129
size*distance	1	-0.02772	0.00310	-8.94	<.0001	4.95365

for the multicollinear data: the VIF values for x_1 and x_2 are 26801!

detecting extreme and influential observations

- obs 1 is extreme in y-value, will only impact slightly the intercept
- obs 2 is extreme in x-value, will not influence the fitted line as the response value is not extreme, taking into account the x-value
- obs 3 is extreme in x and will have a large influence on the fitted line, so it is an influential observation

detection of outliers is important to detect errors in the data

what to do with outliers?

- if it is a typo: correct
- if it is not a typo but not influential: don't care
- if it is not a typo but influential, one can
  - run the analysis with and without the influential observations and report both
  - use a robust regression method (not dealt with)

- **extreme x-values:**

  the **hatmatrix*** is defined as:

  $$\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}' \quad \text{(remark that this depends not on } y\text{)}$$

  h_{ii} is related to the mahalanobisdistance* between $\mathbf{x}_i$ and $\bar{\mathbf{x}}$

  as $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$, the predicted values can be computed as

  $$\hat{y}_i = \sum_{j \neq i} h_{ij} y_j + h_{ii} y_i$$

*$\hat{y}$ is called y-hat and H is the matrix that puts a *hat* on y, hence the *hat* matrix

*this statistical distance takes into account variances and covariances; a formal definition is given on slide Principal Component Analysis - 17

so h_{ii}, the *hat-value* or *leverage value* of the ith observation measures how much y_i influences the prediction $\hat{y}_i$

if the ith observation has an extreme h_{ii} value, the regression line will be pulled to that observation

as it can be shown that $\sum_{i=1}^{n} h_{ii} = k + 1$, one uses the following rule of thumb:

if $h_{ii} > \frac{2(k+1)}{n}$, then the ith observation has an extreme x-value

observations 2 and 3 in the picture have extreme x-values but observation 2 is a *good leverage point* and 3 a *bad leverage point*

- **extreme y-values:**

  compute the *deleted residuals*:

  $$d_i = y_i - \hat{y}_{i(i)}$$

  with $\hat{y}_{i(i)}$ the predicted response for the ith observation based on the model that was fitted without the ith observation, this can be computed without refitting the model using the following relation:

  $$d_i = \frac{y_i - \hat{y}_i}{1 - h_{ii}}$$

it is common to used standardized deleted residuals, also called *studentized deleted residuals*:

$$d_i^* = \frac{y_i - \hat{y}_{i(i)}}{s(d_i)} \text{ with } s(d_i) = \sqrt{\frac{MSE_{(i)}}{1 - h_{ii}}}$$

these values should be $t_{(n-1)-(k+1)}$ distributed for non extreme observations

use the following rule of thumb:

if $|d_i^*| > t_{(n-1)-(k+1)}(95\%)$
then the ith observation has an extreme y-value

- **influential observations:**
  - to assess the influence of observation i on the predicted value $\hat{y}_i$, use:

$$dffits_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$ with dffit the *difference in fit*

use the following rule of thumb: the ith observation is influential

if $|dffits_i| > 1$ (for small or medium n, say 30 to 100)

if $|dffits_i| > 2\frac{\sqrt{k+1}}{\sqrt{n}}$ (for large n)

– Cook's distance is a more general measure to assess the influence of the ith observation on all the parameter estimates:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_{j(i)})^2}{k\ MSE} = \frac{(\mathbf{b} - \mathbf{b}_{(i)})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{b}_{(i)})}{k\ MSE}$$

with $\mathbf{b}_{(i)}$ the estimates obtained without the ith observation

use the following rule of thumb:

if $D_i > F_{k+1,n-k-1}(0.50)$

then the ith observation is influential

with

- k = 1
- n = 36
- cv-hat = 2*(k+1)/n
- cv-rstud = tinv(0.95,n-k-2)
- cv-dffits = 1
- cv-cookd = finv(0.50,k+1,n-k-1)

select observations for which

- hat > cv-hat
- abs(rstudent) > cv-rstud
- abs(dffits) > cv-dfits
- cookd > cv-cookd

RESULTS

	x	y	hat	cv-hat	rstudent	cv-rstud	dffits	cv-dffits	cookd	cv-cookd
28	49		0.02778	0.11111	4.26427	1.69236	0.72083	1	0.17258	0.70747
76	47		0.27548	0.11111	1.15373	1.69236	0.71142	1	0.25062	0.70747
76	11		0.27548	0.11111	-5.36759	1.69236	-3.30978	1	3.01288	0.70747

categorical information

example: explain y, the expenses for R&D, using:

- the global revenues (x) (quantitative or numerical information)
- being listed on the stock market or not ($type$) (qualitative/categorical)

use indicator variables: artificial variables representing the categorical information; there are 2 often used types of coding

1. dummy or treatment coding:

if the variable has c categories, we create c variables d_i such that

$d_i = 1$ for observations of category i and $d_i = 0$ otherwise

in the example, we get:

$$y = \beta_0 + \beta_1 x + \beta_L d_L + \beta_N d_N + u$$

d_L = 1 when the company is listed on the stock market

= 0 otherwise

d_N = 1 when the company is not listed

= 0 otherwise

as d_L and d_N are perfectly correlated, $(\mathbf{X}'\mathbf{X})$ is not invertible and the normal equations have infinitely many solutions

this nonuniqueness can be solved by adding an extra restriction: the default is to set the parameter of one dummy equal to zero (this category is called the reference category)

the estimates and interpretation of the parameters depend on the choice of this reference category (!!!); let for instance $listed$ be the reference category, then $\beta_L = 0$ and

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x \text{ for listed companies} \\ E(y) &= \beta_0 + \beta_N + \beta_1 x \text{ for nonlisted companies} \end{aligned}$$

so β_0 has the expected expenses for R&D for listed companies when revenues are zero (realistic?), β_1 has the effect on the response of an increase of x with 1, and β_N can be interpreted as the expected extra expenses for R&D of nonlisted companies compared to listed companies with similar revenues

2. sum or effects coding: can be defined in 2 equivalent ways:

- using dummies as before but with the extra constraint that the sum of the parameters over all categories is zero*
- defining c-1 indicators with (for example) the last category as the reference category as follows: $d_i = +1$ for observations of category i ($i \neq c$), $d_i = 0$ for observations not in category i or c, $d_i = -1$ for observations in category c

*and this should also be the case for the sum of the interaction parameters if any, see later

example: define a dummy d with is -1 for listed companies and $+1$ for nonlisted companies and use the following model

$$y = \beta_0 + \beta_1 x + \beta_d \, d + u$$

or

$$\begin{aligned} E(y) &= \beta_0 - \beta_d + \beta_1 x \text{ for listed companies} \\ E(y) &= \beta_0 + \beta_d + \beta_1 x \text{ for nonlisted companies} \end{aligned}$$

the same model is obtained with the previous dummy variables d_L and d_N and the constraint $\beta_L + \beta_N = 0$ instead of putting 1 parameter equal to zero (check that then $\beta_N = \beta_d$ with the choices made)

now β_0 has the expected expenses for R&D of companies with zero revenues, averaged over both types of companies, β_1 still has the effect on the response of an increase of x with 1 and $2\beta_d$ indicates how listed and nonlisted companies with similar revenues differ from each other with respect to expenses for R&D

in both cases, the significance test of β_1 tests the effect of the response and the test of β_d or β_N indicates whether there is a difference between listed and nonlisted firms w.r.t. expenses for R&D given the same revenues, only the test of the intercept has different information in both cases but this is typically less important

remark that the anova tables are independent of the arbitrary choice of the reference category!

for models with interactions, this choice of coding becomes even more important as it has more impact on the interpretation and on the different types of SS (because of the likely correlation of main effects with interaction terms)

test whether the effect of the revenues on R&D expenses is the same for listed and nonlisted companies (= whether there is interaction between revenues and type by adding the product of the indicators and x*)

*if the interaction is significant, *type* is called the *moderator, moderating the effect of revenue on R&D*

with dummy coding and $listed$ as the reference category, we get:

$$y = \beta_0 + \beta_1 x + \beta_N d_N + \beta_{1N} x \times d_N$$

$$\begin{aligned} E(y) &= \beta_0 + \beta_1 x \text{ for listed companies} \\ E(y) &= \beta_0 + \beta_N + (\beta_1 + \beta_{1N})x \text{ for nonlisted companies} \end{aligned}$$

whereas effects coding yields:

$$y = \beta_0 + \beta_1 x + \beta_d d + \beta_{1d} x \times d$$

$$\begin{aligned} E(y) &= \beta_0 - \beta_d + (\beta_1 - \beta_{1d})x \text{ for listed companies} \\ E(y) &= \beta_0 + \beta_d + (\beta_1 + \beta_{1d})x \text{ for nonlisted companies} \end{aligned}$$

in both cases the interaction (β_{1N} of $\beta 1d$) indicates whether the effect of x is the same for both types of companies and β_N and β_d whether there is a difference between the intercepts for both types of companies but β_0 and β_1 have different interpretations; in general the models with sum or effects coding are preferred because of the more *balanced* interpretation (although dummy coding parameters are often easier to interpret)

though the correlation between the indicator(s) and x remains the same with both types of codings, the correlation between the indicator and the interaction term is (in most cases) different, as well as the correlation between x and the interaction term, yielding different SS for both types of coding in a model with interaction
$\Rightarrow$ use sum coding in case there are interaction terms in the model

example: explain R&D expenses as a function of revenue and type:

a) the additive model with dummy or treatment coding gives:

RESULTS

Dependent Variable: expenses

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1504.413335	752.206667	72.50	<.0001
Error	17	176.386665	10.375686		
Corrected Total	19	1680.800000			

R-Square	Coeff Var	Root MSE	expenses Mean
0.895058	2.697765	3.221131	119.4000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
revenues	1	1188.167362	1188.167362	114.51	<.0001
type	1	316.245973	316.245973	30.48	<.0001

Source	DF	Type II SS	Mean Square	F Value	Pr > F
revenues	1	1358.613335	1358.613335	130.94	<.0001
type	1	316.245973	316.245973	30.48	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		40.18742309	B	7.23042902	5.56	<.0001
revenues		0.10174212		0.00889122	11.44	<.0001
type	nonlist	-8.05546921	B	1.45910570	-5.52	<.0001
type	listed	0.00000000	B	.	.	.

b) the additive model with sum or effects coding:

RESULTS

Dependent Variable: expenses

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1504.413335	752.206667	72.50	<.0001
Error	17	176.386665	10.375686		
Corrected Total	19	1680.800000			

R-Square	Coeff Var	Root MSE	expenses Mean
0.895058	2.697765	3.221131	119.4000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
revenues	1	1188.167362	1188.167362	114.51	<.0001
type	1	316.245973	316.245973	30.48	<.0001

Source	DF	Type II SS	Mean Square	F Value	Pr > F
revenues	1	1358.613335	1358.613335	130.94	<.0001
type	1	316.245973	316.245973	30.48	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		36.159688	B	7.309921	4.947	0.0001
revenues		0.10174212		0.00889122	11.44	<.0001
type	NONLISTED	4.027735	B	0.729553	-5.521	<.0001
type	LISTED	0.00000000	B	.	.	.

so for additive models:

- type II SS = type III SS as there is no interaction
- as type I SS $\neq$ type II SS, there is correlation between the indicator and revenues but this correlation is independent of the coding

c) the interaction model with dummy or treatment coding:

RESULTS

Dependent Variable: expenses

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1504.413335	752.206667	72.50	<.0001
Error	17	176.386665	10.375686		
Corrected Total	19	1680.800000			

R-Square	Coeff Var	Root MSE	expenses Mean
0.895058	2.697765	3.221131	119.4000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
revenues	1	1188.167362	1188.167362	114.51	<.0001
type	1	316.245973	316.245973	30.48	<.0001
revenues:type	1	0.005708	0.005708	0.00	0.9821

Source	DF	Type II SS	Mean Square	F Value	Pr > F
revenues	1	1358.613335	1358.613335	130.94	<.0001
type	1	316.245973	316.245973	30.48	<.0001
revenues:type	1	0.005708	0.005708	0.00	0.9821

Source	DF	Type III SS	Mean Square	F Value	Pr > F
revenues	1	691.60	691.60	62.7372	<.0001
type	1	2.89	2.89	0.2618	0.6159
revenues:type	1	0.005708	0.005708	0.00	0.9821

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		40.02185357	B	10.41556164	3.84	0.0014
revenues		0.10194777		0.01287108	7.92	<.0001
type	1nonlist	-7.71410896	B	15.07623731	-0.51	0.6159
type	2listed	0.00000000	B	.	.	.
revenues*type	1nonlist	-0.00041714	B	0.01833121	-0.02	0.9821
revenues*type	2listed	0.00000000	B	.	.	.

d) the interaction model with sum or effects coding:

RESULTS

Dependent Variable: expenses

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1504.419043	501.473014	45.49	<.0001
Error	16	176.380957	11.023810		
Corrected Total	19	1680.800000			

R-Square	Coeff Var	Root MSE	expenses Mean
0.895061	2.780747	3.320212	119.4000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
revenues	1	1188.167362	1188.167362	107.78	<.0001
type	1	316.245973	316.245973	28.69	<.0001
revenues*type	1	0.005708	0.005708	0.00	0.9821

Source	DF	Type II SS	Mean Square	F Value	Pr > F
revenues	1	1358.613335	1358.613335	123.24	<.0001
type	1	316.25	316.25	28.6875	<.0001
revenues*type	1	0.005708	0.005708	0.00	0.9821

Source	DF	Type III SS	Mean Square	F Value	Pr > F
revenues	1	1358.269138	1358.269138	123.21	<.0001
type	1	2.886140	2.886140	0.26	0.6159
revenues*type	1	0.005708	0.005708	0.00	0.9821

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		36.159688	B	7.309921	4.947	0.0001
revenues		0.101742		0.008891	11.44	<.0001
type	1nonlist	3.857000	B	7.538000	0.51	0.6159
type	2listed	0.000000	B	.	.	.
revenues*type	1nonlist	0.000209	B	0.009166	0.02	0.9821
revenues*type	2listed	0.000000	B	.	.	.

so for the model with interaction:

- type II $\neq$ type III because there is interaction *and* there is correlation between the main effects and the interaction
- type III SS are different with both types of coding as the correlation between the main effects and interaction depends on the coding!
- if there is interaction, we use sum coding and type II SS to draw conclusions

other applications of dummy coding:

- **Chow test:** test whether the same model is valid in 2 groups or in 2 periods; let:

  $E(y) = \delta_0 + \delta_1 x$ in group 1 $\qquad$ $E(y) = \gamma_0 + \gamma_1 x$ in group 2

  to test $H_0 : \delta_0 = \gamma_0$ and $\delta_1 = \gamma_1$ we introduce a dummy which is 0 for observations belonging to group 1 and 1 for observations of group 2; fit

  $$E(y) = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 x \times d$$

  then H_0 can be rewritten as $H_0 : \beta_2 = \beta_3 = 0$

- **piecewise linear regression:**

  assume we want to fit a different linear relationship between x and y for $x < x^*$ and for $x > x^*$; this can be achieved by estimating:

  $$y = \beta_0 + \beta_1 x + \beta_2 (x - x^*) \times d + u$$

  with $d = 0$ if $x < x^*$ and $d = 1$ for $x \geq x^*$

  if a jump in the response is allowed in $x = x^*$, use:

  $$y = \beta_0 + \beta_1 x + \beta_2 d + \beta_3 \; x \times d + u$$

  remark that we assume that x^* is known!
  if not, we have a nonlinear model (not covered here)

interaction vs multicollinearity

multicollinearity:

- highly correlated explanatory variable contain overlapping information which makes it hard to retrieve the effect of each variable on the response separately

- remember the example where $x_1 \approx 2 * x_2$ and only the total effect $(2\beta_1 + \beta_2)$ could be estimated

- that was an example with multicollinearity but no interaction: the effect of x_1 on the response is independent of the value of x_2 but the effect of x_1 (as measured by β_1) and of x_2 (as measured by β_2) could not be estimated accurately

interaction*:

- here the effect of a variable x_1 on the response depends on the value of the other variable(s)

- illustration: assume
  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}\ x_1 \times x_2 + u$
  then the effect of an increase of x_1 with 1 on y is on average equal to $\beta_1 + \beta_{12} x_2$

- remark that only the correlation between the explanatory variables is needed to check multicollinearity whereas the response variable is also required to check for interaction

*depending on the discipline this is also called **moderation** or, in case one variable is a dummy representing a treatment and control group, **difference in differences (DID)**

1b. ANOVA

ANALYSIS OF VARIANCE

- is used to test for significant differences between several means
- can be used for observational as well as experimental data
- only with carefully collected experimental data, one can show causal relationships

one-way analysis of variance

- one-way anova generalizes the t-test to more than 2 groups
- it can be regarded a special case of linear regression where the explanatory variables are dummies related to the factor levels
- one-way anova investigates the effect of one factor on the response

example: effect of package design of breakfast cereals on sales

- response: sales (in number of boxes sold) in the study period in a particular store
- factor: package design (with 4 levels)

the stores were chosen to be comparable in location and sales volume and other conditions that could effect sales (price, amount and location of shelf space, ...) were kept the same

P1	P2	P3	P4
12	14	19	24
18	12	17	30
	13	21	

$\Rightarrow$ is the expected sales the same for each package design?

we could use the following regression model with indicator variables:

$$y_i = \beta_0 + \beta_1 d_{i1} + \beta_2 d_{i2} + \beta_3 d_{i3} + \beta_4 d_{i4} + u_i$$

with ($j = 1 \ldots 4$ and $i = 1 \ldots n$)

with dummy coding and for instance reference P4, this would yield:

$$\begin{aligned}
\mathbf{E}(Y) &= \beta_0 + \beta_1 \text{ for P1} \\
\mathbf{E}(Y) &= \beta_0 + \beta_2 \text{ for P2} \\
\mathbf{E}(Y) &= \beta_0 + \beta_3 \text{ for P3} \\
\mathbf{E}(Y) &= \beta_0 \quad \text{for P4}
\end{aligned}$$

with sum coding this becomes:

$$\begin{aligned}
\mathbf{E}(Y) &= \beta_0 + \beta_1 \text{ for P1} \\
\mathbf{E}(Y) &= \beta_0 + \beta_2 \text{ for P2} \\
\mathbf{E}(Y) &= \beta_0 + \beta_3 \text{ for P3} \\
\mathbf{E}(Y) &= \beta_0 - \beta_1 - \beta_2 - \beta_3 \text{ for P4}
\end{aligned}$$

test the hypotheses $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs. $H_a : not\ H_0$

because these hypotheses depend on the chosen reference category, it is usually denoted in a more general way:

test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4$ vs. $H_a : not\ H_0$

however, in ANOVA one typically uses a different notation

one-way anova model with r factor levels:

$$y_{ij} = \mu_i + u_{ij} = \mu + \alpha_i + u_{ij} \quad \text{for } j = 1 \ldots n_i \quad \text{and } i = 1 \ldots r$$

y_{ij} : response value in the jth observation with the ith factor level

μ_i : mean response for the ith factor level

μ : overall mean

α_i : effect of ith factor level

u_{ij} : i.i.d. residuals, $u_{ij} \sim N(0, \sigma^2)$

n_i : number of observations for the ith factor level

we have for all i and j:

$$\begin{aligned} E(y_{ij}) &= \mu_i = \mu + \alpha_i \\ \sigma^2(y_{ij}) &= \sigma^2(u_{ij}) = \sigma^2 \\ y_{ij} &\sim N(\mu + \alpha_i, \sigma^2) \text{ and independent of each other} \end{aligned}$$

in the example:

$$sales_{ij} = \mu + \alpha_i + u_{ij}$$

although the notation is different, we use similar inference procedures as in regression (the same teststatistics can also be derived based on another but more restricted theory, that is where the typical ANOVA notation comes from)

the normal equations have a simple form in this case*:

$$\text{let } \bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} \quad ; \quad y_{..} = \sum_{i=1}^{r}\sum_{j=1}^{n_i} y_{ij} \quad ; \quad n = \sum_{i=1}^{r} n_i$$

$$\begin{aligned} n\widehat{\mu} + \sum_{i=1}^{r} n_i \hat{\alpha}_i &= y_{..} \\ \widehat{\mu + \alpha_1} &= \bar{y}_{1.} \\ &\dots \\ \widehat{\mu + \alpha_r} &= \bar{y}_{r.} \end{aligned}$$

*remark that a dot instead of a subscript stands for the summation over that subscript and a bar above the symbol stands for the division if that sum by the number of terms in the sum

which clearly shows the nonuniqueness of the individual parameter estimates for μ and α_i separately (there are only r linear independent equations for $r+1$ unknowns) but $\mu+\alpha_i$ has unique estimates (and only those are required for computing SSE and SSR):

$$\begin{aligned} SSE &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\widehat{y}_{ij})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\widehat{(\mu+\alpha_i)})^2 \\ &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2 \\ SSR &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\widehat{y}_{ij}-\bar{y}_{..})^2 = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2 \end{aligned}$$

the main hypothesis in one-way ANOVA is

$$\mathbf{H_0}: \mu_1 = \mu_2 = \ldots = \mu_r \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

or, in terms of the overparametrized model $\mu_i = \mu + \alpha_i$:

$$\mathbf{H_0}: \alpha_1 = \alpha_2 = \ldots = \alpha_r \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

we proceed as in regression by computing the SS under H_0:

$$SSE(H_0) = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = SSTO$$

$$
\begin{aligned}
F &= \frac{\frac{SSE(H_0)-SSE}{r-1}}{\frac{SSE}{n-(r-1)-1}} = \frac{\frac{SSTO-SSE}{r-1}}{\frac{SSE}{n-(r-1)-1}} = \frac{\frac{SSR}{r-1}}{\frac{SSE}{n-(r-1)-1}} \\
&= \frac{\left[\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{..})^2 - \sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2\right]/(r-1)}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2/(n-r)} \\
&= \frac{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(\bar{y}_{i.}-\bar{y}_{..})^2/(r-1)}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})^2/(n-r)} = \frac{MSR}{MSE} \overset{H_0}{\sim} F_{r-1,n-r}
\end{aligned}
$$

MSR: *between or treatment sum of squares*
measures the differences *between* the group means

MSE: *within sum of squares*
measures the variation *within* the groups

ANOVA: *analysis of variance* because 2 *variances* are compared to decide on the difference between the *means*

remark that to compute type I, II, III sums of squares, all indicator variables corresponding to the same factor (or to the same interaction in later sections) are included or removed simultaneously

example: analyse the sales data

RESULTS

```
                          Class Level Information
                    Class           Levels    Values
                    DESIGN               4    P1 P2 P3 P4

                   Number of Observations          10

Dependent Variable: SALES
                               Sum of
Source                 DF     Squares   Mean Square     F Value  Pr > F
Model                   3  258.0000000  86.0000000       11.22    0.0071
Error                   6   46.0000000   7.6666667
Corrected Total         9  304.0000000

          R-Square     Coeff Var      Root MSE    SALES   Mean
          0.848684      15.38264      2.768875        18.00000

Source                 DF     Type I SS     Mean Square    F Value  Pr > F
DESIGN                  3     258.000000    86.0000000     11.22    0.0071

Source                 DF     Type II SS    Mean Square    F Value  Pr > F
DESIGN                  3     258.000000    86.0000000     11.22    0.0071

Source                 DF     Type III SS   Mean Square    F Value  Pr > F
DESIGN                  3     258.000000    86.0000000     11.22    0.0071
```

Level of DESIGN	N	SALES Mean	SALES Std Dev
P1	2	15.0000000	4.24264069
P2	3	13.0000000	1.00000000
P3	3	19.0000000	2.00000000
P4	2	27.0000000	4.24264069

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	27.00000000 B	1.95789002	13.79	<.0001
DESIGN P1	-12.00000000 B	2.76887462	-4.33	0.0049
DESIGN P2	-14.00000000 B	2.52762515	-5.54	0.0015
DESIGN P3	-8.00000000 B	2.52762515	-3.17	0.0194
DESIGN P4	0.00000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular,and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

---

remark that the general F-test and the average sales per factor level do not depend on the nonunique estimates but the parameter estimates depend on the coding scheme used (here dummy coding)

power

- power = the probability that H_0 is rejected while H_a is true

$$= P\left(\frac{MSR}{MSE} > F_{r-1,n-r}(1-\alpha)|H_a\ true\right)$$

- the power is a function of the true values of μ_i (denoted by μ_i^T), of σ and of the sample sizes
- if H_a is true, $\frac{MSR}{MSE}$ is no longer $F_{r-1,n-r}$ distributed but has a non-central $F_{r-1,n-r,\lambda}$ distribution with non-centrality parameter:

$$\lambda = \frac{\sum_i^r n_i \left(\mu_i^T - \bar{\mu}^T\right)^2}{\sigma^2} \text{ with } \bar{\mu}^T = \sum_{i=1}^{r} \frac{\mu_i^T}{r}$$

- as with all statistical tests, the power can be increased by increasing the significance level α but then the probability of a type I error increases too

compute the power for an overall F-test with $r = 4$, $\alpha = 5\%$, $\bar{\mu}^T = (25, 25, 26, 30)$ and $\sigma = 3$ if the number of observations per group is 2 or 3 or ...

RESULTS

Computed Power

Index	N Per Group	Power
1	2	0.163
2	3	0.322
3	4	0.479
4	10	0.947
5	20	>.999

compute how many observations per group are necessary to obtain a power of 0.90 in case of $r = 4$, $\alpha = 5\%$, $\bar{\mu}^T = (25, 25, 26, 30)$ and $\sigma = 3$

RESULTS

```
Computed N Per Group
Actual     N Per
Power     Group
0.917         9
```

post hoc analysis

if the global F-test indicates that not all means are equal, one typically wants to know where the differences are located by testing contrasts or looking at multiple comparisons

a **contrast** is a linear combination of the group means such that the coefficients sum to 0 (it has an estimator that is independent of the coding!):

$$C \equiv \sum_{i=1}^{r} c_i \mu_i = \sum_{i=1}^{r} c_i \alpha_i \quad \text{with} \quad \sum_{i=1}^{r} c_i = 0$$

often used: pairwise comparisons: $\mu_i - \mu_j = 0$

this hypothesis can be tested with our general approach:

$$\frac{\frac{SSE(H_0)-SSE}{q}}{\frac{SSE}{n-r}} = \frac{\frac{SSE(H_0)-SSE}{1}}{MSE} = \frac{SS_{contrast}}{MSE} \overset{H_0}{\sim} F_{1,n-r}$$

example: package design P1 and P2 used 3-color printing and P3 and P4 5-color printing; is there a significant difference between the expected sales with 3-color and 5-color package designs?

$$H_0 : \mu_1 + \mu_2 = \mu_3 + \mu_4 \qquad \text{versus} \qquad \mathbf{H_a} : not\ H_0$$

RESULTS

Dependent Variable: SALES

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
3 vs 5 colors	1	194.4000000	194.4000000	25.36	0.0024

problem 1: significance level is valid for one particular contrast

$\Rightarrow$ if many contrasts have to be tested, the probability of rejecting at least one H_0 incorrectly can get large

$\Rightarrow$ this is called α-*inflation* or the *multiple testing* or *multiple comparison* problem

assume you want to test 6 contrasts:

$$\begin{aligned}
&\mathcal{P}(\text{ at least 1 hypothesis is rejected incorrectly}) \\
&\quad = \quad 1 - \mathcal{P}(\text{ none of the null hypothesis is rejected incorrectly}) \\
&\quad = \quad 1 - (1-\alpha)^6 \quad \Rightarrow \text{ with } \alpha = 5\% \text{ , } \mathcal{P} = 26\%
\end{aligned}$$

remember that testing the hypothesis $H_0 : C = 0$, using α as the probability of a type I error, is equivalent with checking whether the $(1 - \alpha)100\%$ confidence interval for C contains zero, so we have similarly that

$\mathcal{P}$(at least 1 hypothesis is rejected incorrectly)

$= \mathcal{P}$(at least 1 confidence interval does not contain 0 when it should)

$= 1 - \mathcal{P}$(none of the confidence intervals does not contain 0 when it should)

$= 1 - (1 - \alpha)^6$ voor $\alpha = 5\%$ is $\mathcal{P} = 26\%$

$\Rightarrow$ how to control the *overall error rate* or *overall confidence level*?

problem 2: *data snooping*:

- the significance level is only valid if the contrast was not suggested by the data!

- if one tests systematically the pair of means corresponding to the largest and smallest sample mean, this invalidates the classical t-test as one will make more type 1 errors (rejecting H_0 while it is true) than specified by α

- check the applet that illustrates the concept of data snooping

multiple comparisons procedures

the studentized range distribution $\mathcal{Q}_{r,n-r}$ is defined as follows:

- let $z_1, z_2, ...z_r$ be a sample from a $N(\mu, \sigma^2)$ distribution
- define $w = max_i(z_i) - min_i(z_i)$
- let s^2 be an estimator for σ^2 with $n - r$ degrees of freedom
- then $\frac{w}{s}$ has a *studentized range* distribution $\mathcal{Q}_{r,n-r}$

the **Tukey-method** uses this result with $z_i = \bar{y}_{i.} - \mu_i$ and $s^2 = \frac{MSE}{k}$ to test all pairwise comparisons (assuming $n_i = k$):

$$\mathcal{P}\left(\frac{\max_i(\bar{y}_{i.}-\mu_i)-\min_i(\bar{y}_{i.}-\mu_i)}{\sqrt{MSE/k}}\leq \mathcal{Q}_{r,n-r}(1-\alpha)\right)=1-\alpha$$

$$\mathcal{P}\Big(\bar{y}_{i.}-\bar{y}_{j.}-\mathcal{Q}_{r,n-r}(1-\alpha)\sqrt{MSE/k}\leq \mu_i-\mu_j$$
$$\leq \quad \bar{y}_{i.}-\bar{y}_{j.}+\mathcal{Q}_{r,n-r}(1-\alpha)\sqrt{MSE/k} \text{ for all i and j}\Big)=1-\alpha$$

- simultaneous $100(1-\alpha)\%$ confidence intervals for all $\mu_i-\mu_j$
- if 0 does not lie in one of the intervals, the associated means are said to be significantly different
- $(1-\alpha)$ is also called the *family confidence coefficient* or the *experimentwise error rate* (this procedure guarantees that in repeated sampling, in $100(1-\alpha)\%$ of the samples all the intervals contain correctly $\mu_i-\mu_j$ for all i and j)

example:

RESULTS

Tukey's Studentized Range (HSD) Test for sales

NOTE: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	6
Error Mean Square	7.666667
Critical Value of Studentized Range	4.89559

Comparisons significant at the 0.05 level are indicated by ***.

design Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
P4 - P3	8.000	-0.750	16.750	
P4 - P1	12.000	2.415	21.585	***
P4 - P2	14.000	5.250	22.750	***
P3 - P4	-8.000	-16.750	0.750	
P3 - P1	4.000	-4.750	12.750	
P3 - P2	6.000	-1.826	13.826	
P1 - P4	-12.000	-21.585	-2.415	***
P1 - P3	-4.000	-12.750	4.750	
P1 - P2	2.000	-6.750	10.750	
P2 - P4	-14.000	-22.750	-5.250	***
P2 - P3	-6.000	-13.826	1.826	
P2 - P1	-2.000	-10.750	6.750	

if the sample sizes are unequal, one has to use the harmonic mean $\bar{n}_{ij}$ of n_i and n_j:

$$\frac{2}{\frac{1}{n_i} + \frac{1}{n_j}}$$

in our packaging example $\bar{n}_{34} = 2.4$ and therefore the confidence interval for $\mu_4 - \mu_3$ is

$$\left[8 \pm 4.896\sqrt{\frac{7.6667}{2.4}}\right] = [-0.750, 16.750]$$

the **Bonferroni method** can be used to derive simultaneous confidence intervals based on the following idea:

let $\mathbf{I_1}$ and $\mathbf{I_2}$ be confidence intervals for C_1 and C_2

$\Rightarrow \mathcal{P}(C_1 \in \mathbf{I_1}) = 1 - \alpha_1$ and $\mathcal{P}(C_2 \in \mathbf{I_2}) = 1 - \alpha_2$

then:

$$\begin{aligned} \mathcal{P}(C_1 \in \mathbf{I_1} \textbf{ and } C_2 \in \mathbf{I_2}) &= 1 - \mathcal{P}(C_1 \notin \mathbf{I_1} \textbf{ or } C_2 \notin \mathbf{I_2}) \\ &= 1 - \alpha_1 - \alpha_2 \\ &\quad + \mathcal{P}(C_1 \notin \mathbf{I_1} \textbf{ and } C_2 \notin \mathbf{I_2}) \\ \Rightarrow \mathcal{P}(C_1 \in \mathbf{I_1} \textbf{ and } C_2 \in \mathbf{I_2}) &\geq 1 - \alpha_1 - \alpha_2 \end{aligned}$$

to compute simultaneous $100(1-\alpha)\%$ confidence intervals for m contrasts, compute confidence intervals with $100(1-\alpha/m)\%$ confidence level

remark that if all pairwise comparisons are required, Tukey will give smaller confidence intervals

RESULTS

```
            General Linear Models Procedure
     Bonferroni (Dunn) T tests for variable: SALES

NOTE: This test controls the type I experimentwise error rate
      but generally has a higher type II error rate than
      Tukey's for all pairwise comparisons.

      Alpha= 0.05  Confidence= 0.95  df= 6  MSE= 7.666667
                   Critical Value of T= 3.86299
```

Comparisons significant at the 0.05 level are indicated by '***'.

VERPAKK Comparison		Simultaneous Lower Confidence Limit	Difference Between Means	Simultaneous Upper Confidence Limit	
D	- C	-1.764	8.000	17.764	
D	- A	1.304	12.000	22.696	***
D	- B	4.236	14.000	23.764	***
C	- D	-17.764	-8.000	1.764	
C	- A	-5.764	4.000	13.764	
C	- B	-2.733	6.000	14.733	
A	- D	-22.696	-12.000	-1.304	***
A	- C	-13.764	-4.000	5.764	
A	- B	-7.764	2.000	11.764	
B	- D	-23.764	-14.000	-4.236	***
B	- C	-14.733	-6.000	2.733	
B	- A	-11.764	-2.000	7.764	

---

6 contrasts with overall confidence of 95%, so the pairwise confidence should be $1 - \frac{0.05}{6}$; remark that $t_6(1 - \frac{0.05}{12}) = 3.863$

ANOVA assumptions

here too, we need assumptions (A1) until (A4)

the effect of deviations in the context of anova has been studied:

- normality:
  the F-test is quite robust to nonnormality (the p-value is still quite realiable) IF the distribution is not extremely skewed

- homoscedasticity: unequal variances will
  - not have much impact on the validity of the F-test IF the sample sizes are approximately equal
  - yield invalid confidence intervals for pairwise comparisons (try to transform the data)

blocked experiments

- goal:
  - reduce the residual variance
  - increase the validity of the conclusions

- group the subjects* in homogeneous groups and assign the treatments at random within the block

*also called **experimental units** in this context

example 1: prior to introducing widespread training in the firm, the firm tested three training methods: (1) study at home with programmed training materials, (2) training sessions at local offices conducted by local staff, and (3) training at the headquarter

- one can assign n employees randomly to the 3 training programs and analyse the test results by one-way anova
- it is likely that the difference in the effects is relative small compared to the variation among the employees
- reduce the unexplained variation by forming blocks of 3 similar employees with respect to age, education, ...

example 2: compare the effect of 4 different diets

- assign 20 women randomly to the diets and compare the mean loss by one way anova
- the age and/or initial weight might also have a large effect
- form 5 blocks of 4 *similar* women and assign the diets randomly within each block

the model with r factor levels and s blocks

$$y_{ij} = \mu_{ij} + u_{ij} = \mu + \alpha_i + \beta_j + u_{ij} \ (i = 1, \ldots r \text{ and } j = 1, \ldots s)$$

with

μ : the intercept

α_i : meausing the effect of treatment i

β_j : measuring the effect of block j

u_{ij} : $\sim N(0, \sigma^2)$

s : the number of blocks

this model can also be written as a regression model by creating indicator variables x_i for the different treatments and indicator variables z_j for the different blocks

$$y = \mu + \alpha_1 x_1 + \ldots \alpha_r x_r + \beta_1 z_1 + \ldots + \beta_s z_s + u$$

all previous remarks about different types of coding (or extra equations to get unique parameters) and choice of reference categories remain valid

is there a treatment effect?

$$\mathbf{H_0}: \bar{\mu}_{1.} = \bar{\mu}_{2.} = \ldots = \bar{\mu}_{r.} \qquad \text{versus} \qquad \mathbf{H_a}: \text{not } H_0$$

with $\overline{\mu}_{i.} = \sum_{j=1}^{s} \mu_{ij}/s$, or equivalently:

$$\mathbf{H_0}: \alpha_1 = \ldots = \alpha_r \qquad \text{versus} \qquad \mathbf{H_a}: \text{not } H_0$$

- example 1: do the training programs yield the same average test results or are the effects significantly different?
- example 2: is there a significant difference in the average weight loss with the 4 diets?

as before we compare the SSE of the full model

$$y_{ij} = \mu_{ij} + u_{ij} = \mu + \alpha_i + \beta_j + u_{ij}$$

with that of the following reduced model

$$y_{ij} = \mu + \beta_j + u_{ij}$$

(i= 1,. . . r and j = 1,. . . s)

this is type II SS = type III SS: $SS(x_1, x_2, \ldots, x_r | z_1, z_2, \ldots, z_s)$, the marginal value of the x-variables

was blocking useful? test

$$\mathbf{H_0}: \bar{\mu}_{.1} = \bar{\mu}_{.2} = \ldots = \bar{\mu}_{.s} \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

with $\overline{\mu}_{.j} = \sum_{i=1}^{r} \mu_{ij}/r$, or equivalently

$$\mathbf{H_0}: \beta_1 = \ldots = \beta_s \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

by comparing the SSE of the full model with that of the following reduced model

$$y_{ij} = \mu + \alpha_i + u_{ij}$$

example: difference between 3 training programs?

analyzing the scores with one-way anova yields an F-test with p-value of 0.4369

RESULTS

```
     Class Level Information
     Class          Levels    Values
     training            3    training1 training2 training3
     block               8    1 2 3 4 5 6 7 8

Number of Observations  24

Dependent Variable: score

Source                DF   Sum of Squares     Mean Square    F Value    Pr > F
Model                  9     8616.916667      957.435185      42.84    <.0001
Error                 14      312.916667       22.351190
Corrected Total       23     8929.833333

        R-Square     Coeff Var      Root MSE    score Mean
        0.964958      7.454987      4.727705      63.41667

Source              DF      Type II SS     Mean Square    F Value    Pr > F
training             2      677.083333      338.541667      15.15    0.0003
block                7     7939.833333     1134.261905      50.75    <.0001
```

Tukey's Studentized Range (HSD) Test for treatment

NOTE: This test controls the Type I experimentwise error rate

Alpha	0.05
Error Degrees of Freedom	14
Error Mean Square	22.35119
Critical Value of Studentized Range	3.70139
Minimum Significant Difference	6.1869

Means with the same letter are not significantly different.

Tukey Grouping	Mean	N	training
A	68.625	8	training2
A			
A	65.500	8	training3
B	56.125	8	training1

Tukey's Studentized Range (HSD) Test for score

NOTE: This test controls the Type I experimentwise error rate

Alpha	0.05
Error Degrees of Freedom	14
Error Mean Square	22.35119
Critical Value of Studentized Range	4.99029
Minimum Significant Difference	13.621

Means with the same letter are not significantly different.

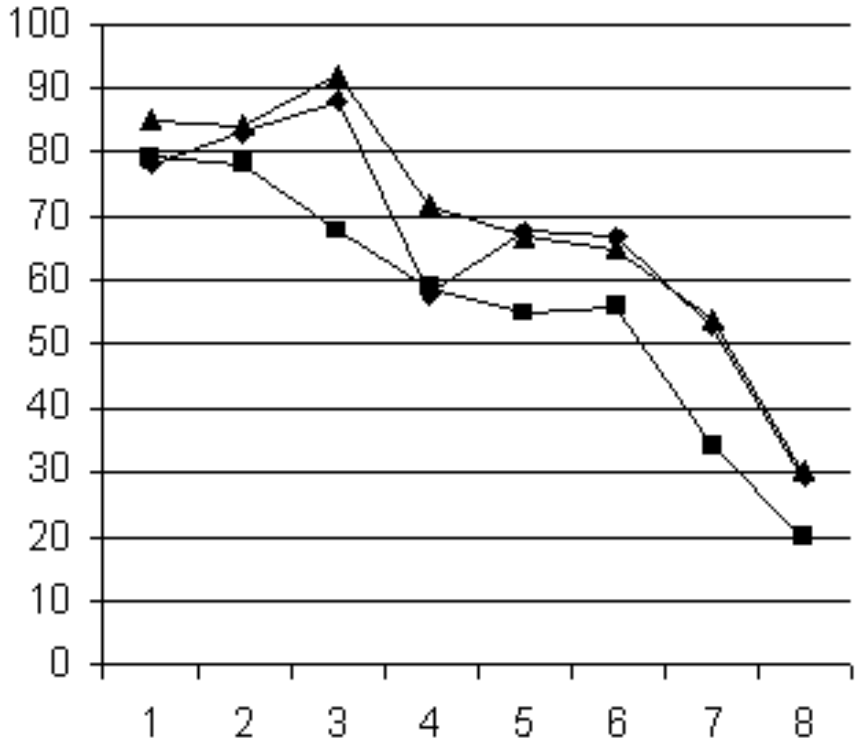
Tukey Grouping	Mean	N	block
A	82.667	3	3
A			
A	81.667	3	2
A			
A	80.667	3	1
B	63.333	3	5
B			
B	63.000	3	4
B			
B	62.667	3	6
C	47.000	3	7
D	26.333	3	8

so there is a significant difference between the mean scores of the 3 programs and it was useful to block the employees as the mean scores are significantly different among blocks

check that *type I* $=$ *type II* $=$ *type III*, independent of the coding:

- type II $=$ type III because there is no interaction
- type I $=$ type II because it is a balanced design; the number of blocks is the same for each treatment and the number of treatments is the same for all blocks: knowing that an observation is in block j contains no information on the treatment and vice versa, so there is no overlap in the variation explained by the x dummies and the variance explained by the z dummies, and this is independent of the coding

this model assumes that the effects of treatment and block are additive (no interaction); one can check this additivity visually by checking whether the lines are parallel (if there is more than 1 observation in the cells, we can test this assumption)



analysis of covariance

- we augment the analysis of variance model with quantitative variables *(covariates)* that are related to the response in order to reduce the unexplained variance
- so the goal is the same as with blocked experiments: improve the power of the anova test
- we use the extra information while analysing the results if it is impossible to form homogeneous blocks
- the covariate has to be related with the response (otherwise it would not reduce the SSE) but should not be influenced by the treatments in any way (so preferably measured before the study)

example: investigate what the best way is to teach statistics: flipped classroom, lectures or pc-labs

- a one-way anova would assign the students randomly to the 3 treatments and compare the average exam results of the 3 groups
- it is likely that the treatment effects will be small compared to the variation within the groups
- blocking would assume that it is possible to make blocks of students with the same aptitude for statistics (measured for instance by the score on the previous exam of statistics)
- if blocking is impossible due to practical reasons, add the covariate information to the model when analysing the data

remark that *the time devoted to the course* would not be a valid covariate as that might well be influenced by the treatment

example: is there a significant difference between starting salaries of business engineers that get a job in the accounting, marketing and finance department of firms?

take into account the *capacity* of the person (for instance measured by the obtained result when graduating) and include it as a covariate

example: we will compare the effect of 3 advertising campaigns for breakfast cereals:

- A: give coupons at the shop entrance to be used the same day
- B: use a larger display at the shelf
- C: use a special booth where people can taste the product

15 comparable shops were selected and each of the 3 campaigns was randomly assigned to 5 shops on a particular Saturday

the results of a one-way anova analysis are:

RESULTS

Dependent Variable: sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	112.9333333	56.4666667	2.03	0.1742
Error	12	334.0000000	27.8333333		
Corrected Total	14	446.9333333			

R-Square	Coeff Var	Root MSE	sales Mean
0.252685	14.95954	5.275731	35.26667

Source	DF	Type I SS	Mean Square	F Value	Pr > F
type	2	112.9333333	56.4666667	2.03	0.1742

$\Rightarrow$ how to reduce the unexplained variance?

$\Rightarrow$ collect how many boxes were sold the previous Saturday in each of the shops and use this as a covariate

model with r factor levels and 1 covariate:

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + u_{ij} \quad (i = 1 \ldots r, j = 1 \ldots n_i)$$

with

- μ: intercept
- α_i: effect of the ith treatment
- γ: effect of the covariate
- x_{ij}: the value of the covariate for the jth observation under the ith treatment
- u: i.i.d. $N(0, \sigma^2)$ residuals

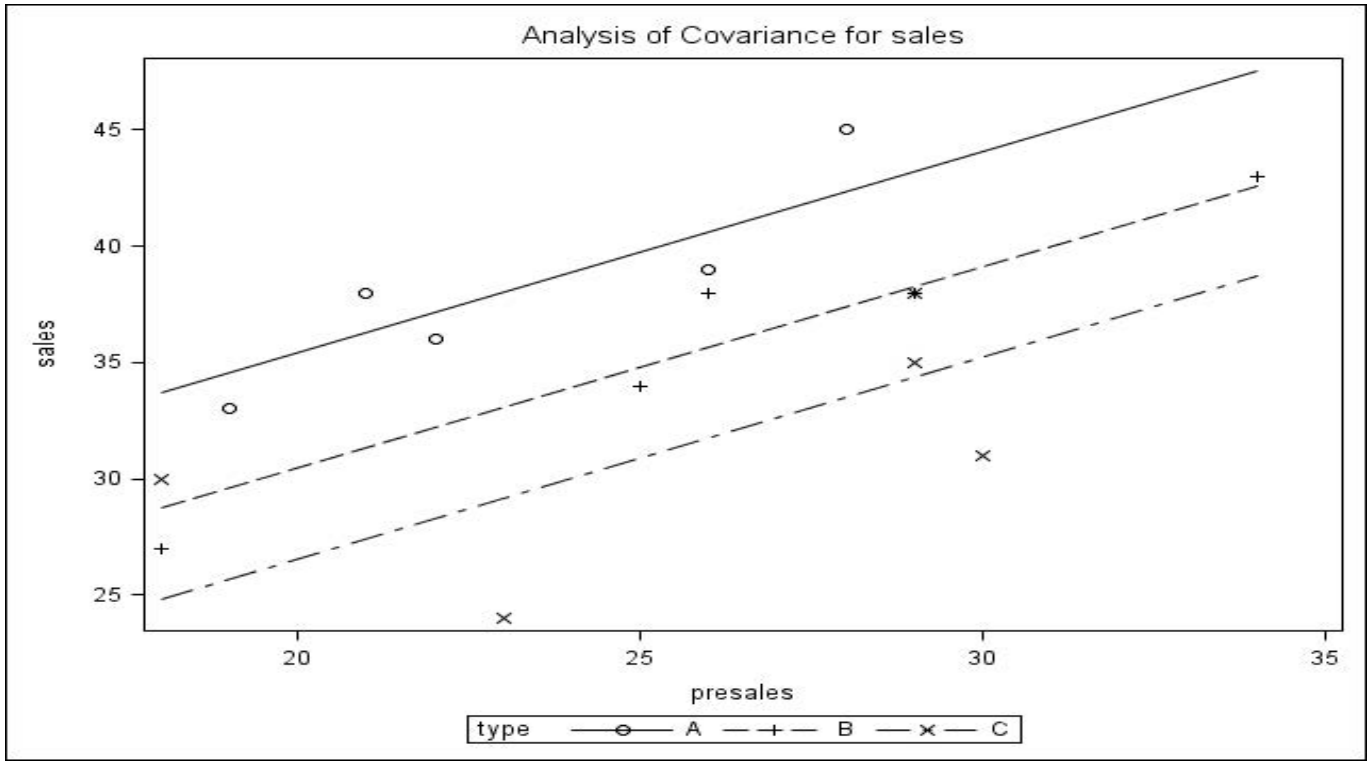
which yields:

$$\begin{aligned} E(y_{ij}) &= \mu + \alpha_i + \gamma x_{ij} \\ \sigma^2(y_{ij}) &= \sigma^2 \\ y_{ij} &\sim N(\mu + \alpha_i + \gamma x_{ij}, \sigma^2) \end{aligned}$$

for example:

$$\begin{aligned} E(y_{1j}) &= \mu + \alpha_1 + \gamma x_{1j} \quad \text{for A} \\ E(y_{2j}) &= \mu + \alpha_2 + \gamma x_{2j} \quad \text{for B} \\ E(y_{3j}) &= \mu + \alpha_3 + \gamma x_{3j} \quad \text{for C} \end{aligned}$$

in general r parallel lines are fitted



$\Rightarrow \alpha_i - \alpha_j$ are still measuring the treatment effects

- first we test whether the effect of the covariate is the same for each factor level, fit the model including interaction terms*:

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + (\alpha\gamma)_i \; x_{ij} + u_{ij}$$

for the example:

$$\begin{aligned} y_{1j} &= \mu + \alpha_1 + (\gamma + (\alpha\gamma)_1) \; x_{1j} + u_{1j} \quad \text{for A} \\ y_{2j} &= \mu + \alpha_2 + (\gamma + (\alpha\gamma)_2) \; x_{2j} + u_{2j} \quad \text{for B} \\ y_{3j} &= \mu + \alpha_3 + (\gamma + (\alpha\gamma)_3) \; x_{3j} + u_{3j} \quad \text{for C} \end{aligned}$$

*remark that this is a similar model as the model explaining expenses for R&D but here in ANOVA notation

$$\mathbf{H_0}: (\alpha\gamma)_1 = \ldots = (\alpha\gamma)_r \qquad \text{versus} \qquad \mathbf{H_a}: \text{ niet } H_0$$

is tested by comparing the SS of the full model

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + (\alpha\gamma)_i \; x_{ij} + u_{ij}$$

with the SS of the reduced model

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + u_{ij}$$

if the interaction is significant: report the effect of the covariate for each treatment; if no interaction is required, test the main effect of the treatment

- **if there is no interaction, test the treatment main effect**

$$\mathbf{H_0}: \alpha_1 = \alpha_2 = \ldots = \alpha_r \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

by comparing the SS of the full model*

$$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + u_{ij}$$

with the SS of the reduced model

$$y_{ij} = \mu + \gamma x_{ij} + u_{ij}$$

*one can refit the model without interaction terms or assume that the interaction is so small that it will not make a difference - there is still discussion among statisticians what is the best approach - though it seems that using type II SS is more powerful for testing the main effects than type III SS if there is no interaction

- **if there is no interaction, test whether there is a significant covariate effect**

  $\mathbf{H_0}: \gamma = 0$ versus $\mathbf{H_a}: \gamma \neq 0$

  by comparing the SS of the full model

  $$y_{ij} = \mu + \alpha_i + \gamma x_{ij} + u_{ij}$$

  with the SS of the reduced model

  $$y_{ij} = \mu + \alpha_i + u_{ij}$$

example: we use sum coding and type II SS:

RESULTS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	349.6925427	69.9385085	6.47	0.0081
Error	9	97.2407906	10.8045323		
Corrected Total	14	446.9333333			

R-Square	Coeff Var	Root MSE	sales Mean
0.782427	9.320486	3.287025	35.26667

Source	DF	Type II SS	Mean Square	F Value	Pr > F
type	2	185.384	92.692	8.579	0.0082
presales	1	223.4623829	223.4623829	20.68	0.0014
presales*type	2	13.2968265	6.6484133	0.62	0.5617

so the lines can be considered to be parallel, there is a significant different effect between the types and the presales have a significant impact too

treatment means and adjusted treatment means:

assume you need to predict the expected sales with each type of promotion, then you have the following 2 options:

- *treatment means*: gives the treatment means: $\bar{y}_{i.} = \hat{\mu} + \hat{\alpha}_i + \hat{\gamma}\bar{x}_{i.}$
  - is the average response value of the ith treatment
  - is misleading if the covariate values are very different for different treatments

- *adjusted treatment means:* $\bar{y}^a_{i.} = \hat{\mu} + \hat{\alpha}_i + \hat{\gamma}\bar{x}_{..}$
  - is the predicted treatment mean at $\bar{x}_{..}$
  - this is the better (more reliable) approach

RESULTS

Level of type	N	sales Mean	sales Std Dev	presales Mean	presales Std Dev
A	5	38.2000000	4.43846820	23.2000000	3.70135110
B	5	36.0000000	5.95818764	26.4000000	5.85662019
C	5	31.6000000	5.31977443	25.8000000	5.16720427

Least Squares Means (based on the additive model)

type	sales LSMEAN
A	39.8719322
B	34.9045962
C	31.0234717

- remark that in regression analysis the term *covariate* is used for any numerical regressor and *factor* for any categorical regressor whereas in anova, the term *covariate* is used to denote that it is a nuisance factor that is taken into account to improve the analysis of variance where the focus is on the effect of the factor levels
- although anova has to do with the analysis of factor level effects, the term *factor analysis* is used for a completely different statistical methodology (see part 4)

analysis of covariance or blocked design?

advantages of blocking:

- no assumption is required about the relationship of the blocking variable and the response (in ancova we assume a linear relationship)

- the treatment effect and the blocking effect can be estimated independently of each other (type I = type III SS)

two-way anova (balanced case)

example: assume the sales corresponding to different types of campaign (factor 1) and advertising media (factor 2) are as follows:

	newspaper	radio
A	9 7 9 7	10 15 13 9
B	3 2 5 8	14 15 12 9
C	9 6 8 9	8 11 7 10

the model including interaction effects is:

$$y_{ijk} = \mu_{ij} + u_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_{ijk}$$

for $i = 1, \ldots, r, j = 1, \ldots, s, k = 1, \ldots, m$

start by testing the interaction terms as the rest of the analysis depends on the significance of these terms!

- **test the interaction effect:** compare the SS of the full model

$$y_{ijk} = \mu_{ij} + u_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + u_{ijk}$$

with that of the following reduced model:

$$y_{ijk} = \mu_{ij} + u_{ijk} = \mu + \alpha_i + \beta_j + u_{ijk}$$

$(i = 1, \ldots, r, j = 1, \ldots, s, k = 1, \ldots, m)$

example:

RESULTS

Class Level Information

Class	Levels	Values
type	3	A B C
medium	2	newspaper radio

Number of observations 24

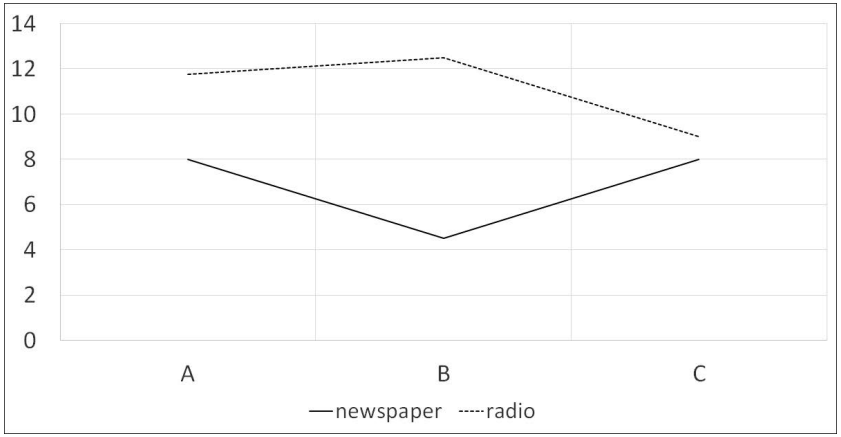
Dependent Variable: sales

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	168.2083333	33.6416667	7.15	0.0008
Error	18	84.7500000	4.7083333		
Corrected Total	23	252.9583333			

R-Square	Coeff Var	Root MSE	sales Mean
0.664965	24.22180	2.169869	8.958333

Source	DF	Type II SS	Mean Square	F Value	Pr > F
type	2	10.0833333	5.0416667	1.07	0.3636
medium	1	108.3750000	108.3750000	23.02	0.0001
type*medium	2	49.7500000	24.8750000	5.28	0.0157

- with sum coding and balanced designs, type I SS = type II SS = type III SS
- as the interaction term is significant, the difference in cell means cannot be explained by additive effects only
- the main effects of the factors are no longer relevant as their effect depends on the level of the other factor
- test *simple effects* instead: the factor effect for a specific level of the other factor



example: simple effects

RESULTS

Least Squares Means

type	medium	sales LSMEAN
A	newspaper	8.0000000
A	radio	11.7500000
B	newspaper	4.5000000
B	radio	12.5000000
C	newspaper	8.0000000
C	radio	9.0000000

type*medium Effect Sliced by type for sales

type	DF	Sum of Squares	Mean Square	F Value	Pr > F
A	1	28.125000	28.125000	5.97	0.0250
B	1	128.000000	128.000000	27.19	<.0001
C	1	2.000000	2.000000	0.42	0.5228

– for types A and B radiospots are significantly better than newspaper ads
– for type C: radiospots and newspaper ads give rise to the same sales on average

- **if there is no significant interaction, test the main effects**
  - only makes sense when there are no significant interaction effects*
  - test the treatment effect of the first factor†:

$$\mathbf{H_0}: \bar{\mu}_{1.} = \bar{\mu}_{2.} \ldots = \bar{\mu}_{r.} \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

comparing the SS of the following 2 models

$$y_{ijk} = \mu + \alpha_i + \beta_j + u_{ijk}$$

$$y_{ijk} = \mu + \beta_j + u_{ijk}$$

*one can refit the model without interaction or assume that it is so small that it will not make a difference - there is still discussion among statisticians what is the best approach; type III corresponds to not refitting the model, type II corresponds approximately to refitting the model but with a better estimate for MSE

†if there are significant interaction terms, this is equivalent to: $\alpha_i + (\alpha\beta)_{i.}$ is equal $\forall i$, which does not make much sense but without interaction, this boils down to equal α_i's as before

– similarly one can test the effect of the second factor:

$$\mathbf{H_0}: \bar{\mu}_{.1} = \bar{\mu}_{.2} \ldots = \bar{\mu}_{.s} \qquad \text{versus} \qquad \mathbf{H_a}: \text{ not } H_0$$

by comparing the SS of the models

$$y_{ijk} = \mu + \alpha_i + \beta_j + u_{ijk}$$

and

$$y_{ijk} = \mu + \alpha_i + u_{ijk}$$

- **factor level effects**
  - one can use tukey intervals to analyse the pairwise comparisons of main effects if there is no interaction
  - contrasts can be tested as before but it gets more complicated (we do not deal with this)

what if unequal sample sizes?

- type I SS are no longer equal to type II or type III SS
- if there are no empty cells: the common approach is to use type II SS with sum coding but as there is overlapping information in the sets of indicators, the effects cannot be disentangled clearly
- if there are empty cells: not all interaction terms are estimable:
  - assume that there is no interaction and fit only main effects
  - test the contrasts that seem most sensible

three-way anova

example: is the quality of beans depending on the type of soil, the plant season and the time in storage?

	may		september	
	grey	red	grey	red
7 months	4 3 3 3 4 5 3 1 3 2	4 3 3 4 3 2 3 6 4 4	3 5 2 3 4 3 6 2 5 4	3 3 2 1 3 4 3 4 3 3
1 month	3 2 2 2 3 2 4 3 2 2	4 6 5 4 7 4 5 5 6 6	4 1 4 4 6 5 4 3 4 4	4 5 6 4 5 5 7 4 4 7

quality is given on a scale from 1 (= very bad) till 7 (= very good)

model $(i = 1 \ldots r, j = 1 \ldots s, k = 1 \ldots t, l = 1 \ldots m)$:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + u_{ijkl}$$

where $(\alpha\beta\gamma)_{ijk}$ denotes the three-way interaction and measures whether the two-way interactions between 2 factors depend on the level of the third factor

this is the direct generalization of a two-way interaction which measures whether the first-order or main effect of one factor depends on the level of the other factor

as the design is balanced, type I SS = type II SS = type III SS (with sum coding!)

RESULTS

Class Level Information

Class	Levels	Values
SOIL	2	grey red
SEASON	2	may sept
STORAGE	2	one seven

Number of observations in data set = 80

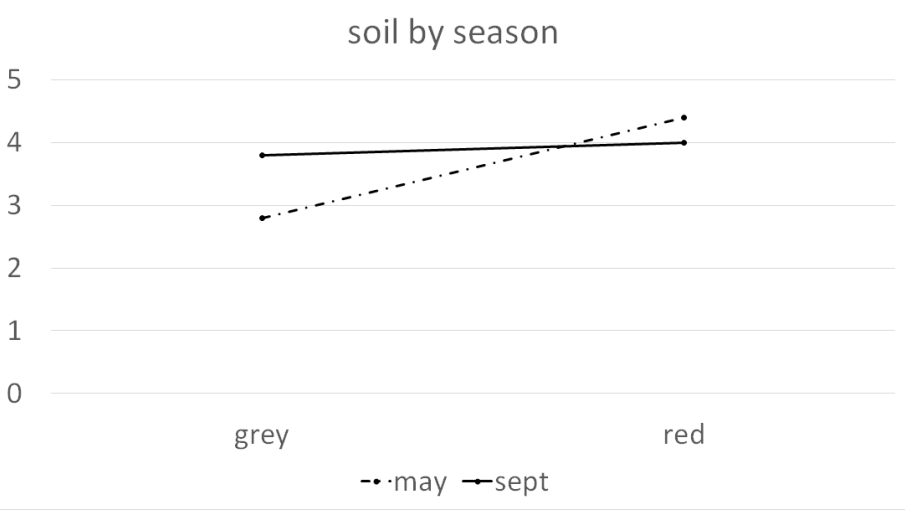
Dependent Variable: QUALITY

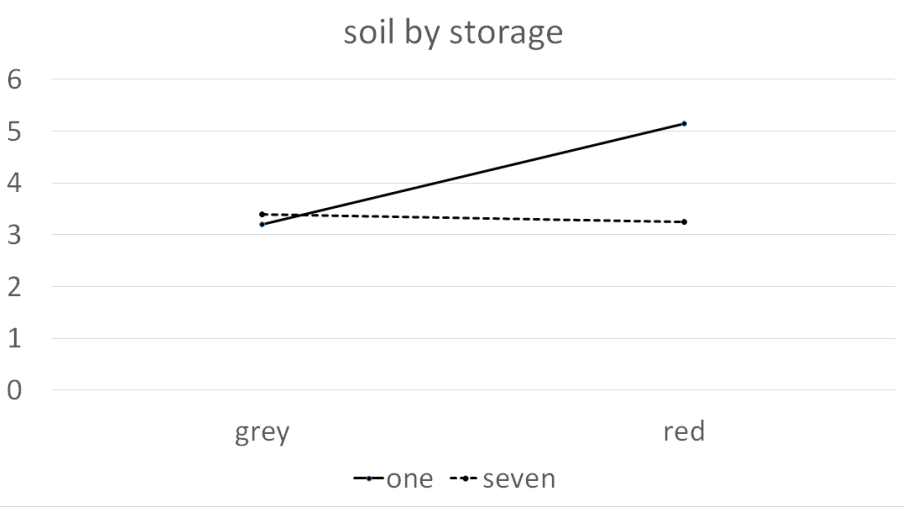
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	66.80000000	9.54285714	7.97	0.0001
Error	72	86.20000000	1.19722222		
Corrected Total	79	153.00000000			

Source	DF	Type II SS	Mean Square	F Value	Pr > F
soil	1	16.20000000	16.20000000	13.53	0.0004
season	1	1.80000000	1.80000000	1.50	0.2241
storage	1	14.45000000	14.45000000	12.07	0.0009
soil*season	1	9.80000000	9.80000000	8.19	0.0055
soil*storage	1	22.05000000	22.05000000	18.42	<.0001
season*storage	1	2.45000000	2.45000000	2.05	0.1569
soil*season*storage	1	0.05000000	0.05000000	0.04	0.8386

the *3-way interaction* is negligible as is *season*storage*

but *soil*season* and *soil*storage* are significant:





testing simple effects:

RESULTS

Least Squares Means

soil	season	quality LSMEAN
grey	may	2.80000000
grey	sept	3.80000000
red	may	4.40000000
red	sept	4.00000000

soil *season Effect Sliced by soil for quality

soil	DF	Sum of Squares	Mean Square	F Value	Pr > F
grey	1	10.000000	10.000000	8.35	0.0051
red	1	1.600000	1.600000	1.34	0.2515

RESULTS

Least Squares Means

soil	storage	quality LSMEAN
grey	one	3.20000000
grey	seven	3.40000000
red	one	5.15000000
red	seven	3.25000000

soil *storage Effect Sliced by soil for quality

soil	DF	Sum of Squares	Mean Square	F Value	Pr > F
grey	1	0.400000	0.400000	0.33	0.5651
red	1	36.100000	36.100000	30.15	<.0001

conclusion:

- on red soil the plant season is unimportant but on grey soil it is better to plant in september

- if one plans to keep the beans 7 months in storage, the type of soil is not important (the quality will not be good anyway) but if one wants to utilize them after one month, planting them on red soil is most appropriate

random or stochastic factor levels

- until now models with *fixed effects* or *fixed factor levels*:
  - all relevant factor levels were included in the design
  - f.i. gender with 2 levels, 7 days a week, the 4 package designs that were developed

- in models with *random or stochastic factor levels*:
  - the factor levels in the design are a random sample from the set of all possible factor levels
  - the individual factor level effects are not important, the goal is to estimate the variance of the factor effects
  - observations are no longer independent from each other

random block effects

similar idea and goal as the previous blocked design but the blocks are now considered a random sample of all possible blocks:

$$y_{ij} = \mu_{ij} + u_{ij} = \mu + \alpha_i + \beta_j + u_{ij} \ \ (i = 1, \ldots r \text{ and } j = 1, \ldots s)$$

but now

μ : the overall mean

α_i : the effect of factor level i

β_j : the random block effect of block j, i.i.d. $\sim N(0, \sigma^2_\rho)$

u_{ij} : $\sim N(0, \sigma^2)$ and all u_{ij} and β_j are assumed independent

example: a police department investigates whether the wear characteristics of 2 paints for road markings are significantly different; they select random locations where they use the 2 paints; after a suitable period of exposure to weather and traffic a measure of wear was obtained; selecting a location = taking a draw from $N(0, \sigma^2_\rho)$

this introduces correlation between the observations which expresses that 2 observations from the same block will be more similar than observations from different blocks:

$$\begin{aligned} var(y_{ij}) &= var(\beta_j + u_{ij}) = \sigma^2_\rho + \sigma^2 \\ cov(y_{ij}, y_{i'j}) &= cov(\mu + \alpha_i + \beta_j + u_{ij}, \mu + \alpha'_i + \beta_j + u_{i'j}) \\ &= var(\beta_j) = \sigma^2_\rho \\ cov(y_{ij}, y_{i'j'}) &= 0 \end{aligned}$$

so the covariance matrix of the observations is block diagonal and any two observations of the same block have the same correlation $\sigma^2_\rho \Rightarrow$ this pattern is called **compound symmetry**

there exist adaptations of the F-test we dealt with earlier to take into account this correlation structure

more recent approaches use maximum likelihood estimators (see chapter on logistic regression for more information on such an approach), taking into account mean and covariance structure of the responses; we do not go into details but look at some examples

the significance test for σ^2_ρ is complex and not always valid (and therefore not given by default) but the test for the fixed effects is valid

example: compare the efficacy of different paints

RESULTS

Class Level Information

Class	Levels	Values
location	15	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
paint	2	1 2

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
location	0.04012
Residual	0.09593

Fit Statistics

-2 Res Log Likelihood	27.8
AIC (smaller is better)	31.8
AICC (smaller is better)	32.2
BIC (smaller is better)	33.2

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
paint	1	14	87.40	<.0001

Estimates for Fixed Effects with sum coding

Effect	paint	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		3.67667	0.07663	14	47.979	<.0001
paint	1	-0.52867	0.05655	14	-9.349	<.0001
paint	2	0	.	.	.	.

---

so there is a significant paint effect and the first paint is on average $2 * 0.52867 = 1.06$ worse than the second paint

$\hat{\sigma}_\rho^2 = 0.04$, so the size of the effect of a randomly chosen location is expected to be between $[-2\hat{\sigma}_\rho; +2\hat{\sigma}_\rho] = [-0.4; 0.4]$, so much smaller than the paint effect

repeated measures

a frequently used **random block design** is one in which a random block consists of a randomly chosen test person*

suppose test persons have to rate 4 different wines; use the model:

$$y_{ij} = \mu + \alpha_i + \pi_j + u_{ij}$$

- α_i: effect of wine i
- π_j: effect of test person j, $\sim N(0, \sigma_\pi^2)$
- u_{ij}: error terms $\sim N(0, \sigma^2)$, independent of π_j

*remark that this is a generalization of the paired samples t-test

as before

$$\begin{aligned} cov(y_{ij}, y_{kj}) &= \sigma^2_\pi,\ i \neq k \\ cov(y_{ij}, y_{kl}) &= 0,\ j \neq l \end{aligned}$$

this is an efficient experiment in that it does not need many respondents; there are however some pitfalls: try to avoid

- *order effects*: first wine is rated different than the last one
  $\Rightarrow$ randomize the order in which the persons rate the wines

- *carry over effects*: rating might depend on previous wine (which might have been very good or bad)
  $\Rightarrow$ allow sufficient time between wines

example: wine tasting

the factor wine is now called a *within subjects factor*

RESULTS

Class Level Information

Class	Levels	Values
judge	6	1 2 3 4 5 6
wine	4	1 2 3 4

Convergence criteria met.

Covariance Parameter Estimates

Cov Parm	Estimate
judge	8.4000
Residual	1.0667

Fit Statistics

-2 Res Log Likelihood	82.6
AIC (smaller is better)	86.6
AICC (smaller is better)	87.3
BIC (smaller is better)	86.2

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
wine	3	15	57.50	<.0001

Differences of Least Squares Means

Effect	wine	_wine	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P
wine	1	2	-2.0000	0.5963	15	-3.35	0.0043	Tukey-Kramer	0.0202
wine	1	3	-6.6667	0.5963	15	-11.18	<.0001	Tukey-Kramer	<.0001
wine	1	4	-6.0000	0.5963	15	-10.06	<.0001	Tukey-Kramer	<.0001
wine	2	3	-4.6667	0.5963	15	-7.83	<.0001	Tukey-Kramer	<.0001
wine	2	4	-4.0000	0.5963	15	-6.71	<.0001	Tukey-Kramer	<.0001
wine	3	4	0.6667	0.5963	15	1.12	0.2811	Tukey-Kramer	0.6844

Estimates for Fixed Effects with sum coding

Effect	wine	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		23.6667	1.2019	5	19.692	<.0001
wine	1	2.3333	0.3651	15	6.390	<.0001
wine	2	-3.6667	0.3651	15	-10.042	<.0001
wine	3	-1.6667	0.3651	15	-4.564	0.0004
wine	4	.	.	.	.	.

PART 2

- ECONOMETRICS

2. ECONOMETRICS

ECONOMETRICS

- is the application of statistical methods to economic data
- the basic tool for econometrics is the multiple linear regression model
- as the results are often used for policy recommendations, econometricians stress the importance of unbiasedness, efficiency, and consistency of the estimators and therefore investigate the consequences of violations of the assumptions

model misspecification

- **what if we include irrelevant variables?**

  assume that the correct model is:

  $$y = \beta_0^* + \beta_1^* x_1 + u$$

  whereas the model that is estimated is:

  $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \tilde{u} \qquad \text{(with } \beta_2 = 0\text{)}$$

  – if the assumptions hold for u, they also hold for $\tilde{u}$ and therefore the estimators for β_1, β_2 and σ^2 are still unbiased

– however $var(b_1) > var(b_1^*)$ in most cases as:

$$\begin{aligned} var(b_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} \times \frac{1}{1 - corr^2(x_1, x_2)} \\ &= var(b_1^*) \times \frac{1}{1 - corr^2(x_1, x_2)} \end{aligned}$$

more general: the estimators of the model with irrelevant estimators are still unbiased and consistent but they are estimated with less precision than in the model without these variables except when the irrelevant variables in the model are uncorrelated with the variables that should be in the model

- **what if we omit relevant variables?**

  assume the true model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

  but the estimated model is:

$$y = \beta_0^* + \beta_1^* x_1 + \tilde{u}$$

  then

$$\begin{aligned}
\mathbf{E}(b_0^*) &= \beta_0 + \frac{1}{n\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} \left(\sum_i x_{i1}^2 \sum_i x_{i2} - \sum_i x_{i1} \sum_i x_{i1} x_{i2} \right) \beta_2 \\
\mathbf{E}(b_1^*) &= \beta_1 + \frac{cov(x_1, x_2)}{var(x_1)} \beta_2
\end{aligned}$$

  where the last terms are called **omitted variable bias**

as we assume that $\beta_2 \neq 0$ (otherwise x_2 would not be a relevant variable) we can derive from these expressions that

- the estimator b_0^* only remains unbiased in the unlikely case that $\sum_i x_{i1}^2 \sum_i x_{i2} - \sum_i x_{i1} \sum_i x_{i1} x_{i2} = 0$
- the estimator b_1^* remains unbiased if $corr(x_1, x_2) = 0$

more general: the intercept will almost always be biased but the other parameters will be unbiased if the omitted variables are uncorrelated with the variables in the model

so in case $cov(x_1, x_2) > 0$ (which is equivalent with $corr(x_1, x_2) > 0$) and $\beta_2 > 0$, there will be upward bias and similarly we get downward bias if $cov(x_1, x_2) < 0$ and $\beta_2 > 0$

example: if the true model in a study of the effect of class attendance on grades is:

$$\text{grade} = \beta_0 + \beta_1 \text{ attendance} + \beta_2 \text{ studyhours} + u$$

but one estimates

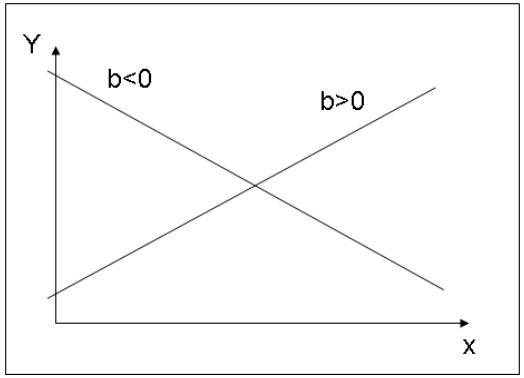
$$\text{grade} = \beta_0 + \beta_1 \text{ attendance} + u$$

one can expect upward bias (so overestimation of the effect of attendance) as $corr(\text{attendance, studyhours}) > 0$ is likely to be

- **what if the relationship is not linear?**

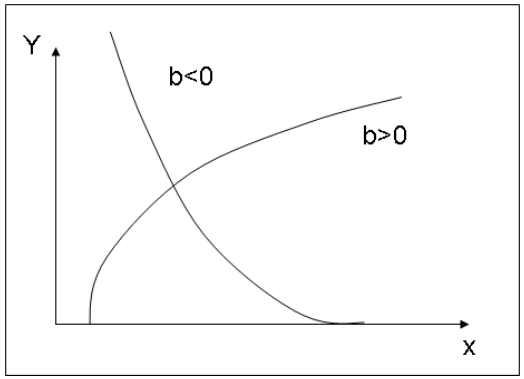
  sometimes the model can be transformed to a linear model by using the natural logarithm; we list 4 often used simple regressions.

  - **linear relationship:** $\hat{y} = a + b\ x$



  - **logarithmic relationship:** $\hat{y} = a + b\ \ln x$

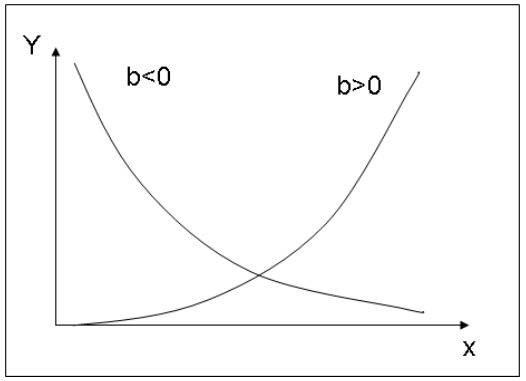
    if x increases with 1%, $\hat{y}$ will on average increase with (approximately) $\frac{b}{100}$



– **exponential relationship or loglinear relationship:**

$\widehat{\ln y} = a + b\ x$ or $\hat{y} = c\ d^x$ $(c = \exp(a),\ d = \exp(b))$

if x increases with 1 unit, $\hat{y}$ will increase on average (approximately) with 100b%



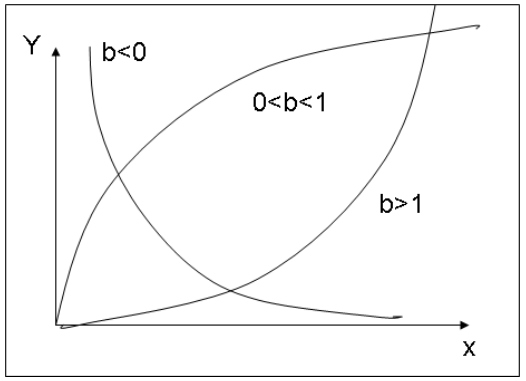
– **power/multiplicative:**

$\widehat{\ln y} = a + b\ \ln x$ of $\hat{y} = c\ x^b$ $(c = \exp(a))$

if x increases with 1%,

$\hat{y}$ will on average increase approximately by b%

$(=$ constant elasticities as $\frac{d\hat{y}/\hat{y}}{dx/x} = b)$



comparing models with transformed response:

to compare the goodness of fit of the following models

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u \qquad \text{with} \quad R_1^2 = corr^2(y_i, \hat{y}_i) \\ \ln y &= \alpha_0 + \alpha_1 x + u \qquad \text{with} \quad R_2^2 = corr^2\left(ln(y_i), \widehat{ln(y_i)}\right) \end{aligned}$$

one cannot compare the corresponding R^2 values as these represent the explained variance of y and of $\ln(y)$ respectively and those are not comparable; one has to proceed as follows:

- fit the loglineair model and compute $\widehat{\ln y_i}$
- compute the correlation r between the y_i and $\exp(\widehat{\ln y_i})$
- this r^2 is then comparable with R_1^2

violations of the standard assumptions

remark that the OLS estimates can always be computed to find the best fitting hyperplane (minimizing the sum of squared residuals)

to show that they are unbiased and consistent and that $\mathbf{cov}(\mathbf{b})$ is equal to $\sigma^2\ (\mathbf{X}'\mathbf{X})^{-1}$, one needs to assume that*

(A1) the residuals have expected value zero: $\mathbf{E}(u_i) = 0$

(A2) the residuals have constant variance: $var(u_i) = \sigma^2$

(A3) the residuals are uncorrelated: $corr(u_i, u_j) = 0, \forall i, j$

*remark that assumption (A4 - normality) is only necessary for inference (hypothesis testing and confidence intervals) in small samples as the estimators are asymptotically normal no matter the distribution of the residuals

under assumptions (A1), (A2) and (A3), the OLS estimators are **BLUE** (best linear unbiased estimators): there are no other estimators that are linear combinations of the responses, that are unbiased and have smaller variance

when assumption (A2) and/or (A3) are not satisfied, the OLS estimators are still unbiased and consistent, so $\mathbf{E}(\mathbf{b}^{OLS}) = \boldsymbol{\beta}$ and $\mathbf{var}(\mathbf{b}^{OLS}) \overset{n\to\infty}{\longrightarrow} 0$ but they are no longer BLUE and

$$\mathbf{cov}(\mathbf{b}^{OLS}) \neq \sigma^2 \, (\mathbf{X}'\mathbf{X})^{-1}$$

which invalidates the default OLS inference

(A2) is violated: HETEROSCEDASTICITY

OLS estimators are still unbiased and consistent but $\mathbf{cov}(\mathbf{b}) \neq \sigma^2 \, (\mathbf{X}'\mathbf{X})^{-1}$, the correct formula is

$$\mathbf{cov}(\mathbf{b^{OLS}}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{V}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

with $\mathbf{V}$ the covariance matrix of the responses:

$$\mathbf{V} = \mathbf{cov}\left(\mathbf{Y}\right) = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

remark that $\mathbf{cov}(\mathbf{b})^{\mathbf{OLS}}$ is estimated by a **sandwich** estimator, where the estimator of the middle term $\mathbf{X}'\mathbf{V}\mathbf{X}$ is called the **meat**

one has several options*:

a) use GLS (generalized least squares estimation) which is BLUE

b) use the OLS estimators with appropriate standard errors (White)

c) try to transform the model to get homoscedasticity,
the $\ln$ function often helps
(not dealt with here as solution b) is easy to use)

*remark that there exist several tests for heteroscedasticity, each testing for a specific type of heteroscedasticity (not dealt with here)

a) GLS

assume that the variances σ_i^2 are known, then the model

$$\begin{aligned} \frac{y_i}{\sigma_i} &= \frac{\beta_0}{\sigma_i} + \beta_1 \frac{x_{i1}}{\sigma_i} + \beta_2 \frac{x_{i2}}{\sigma_i} + \ldots + \beta_k \frac{x_{ik}}{\sigma_i} + \frac{u_i}{\sigma_i} \\ y_i^* &= \beta_0^* + \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \ldots + \beta_k x_{ik}^* + u_i^* \end{aligned}$$

is again a homoscedastic model for which the OLS estimators are BLUE, so the GLS estimators minimize $\sum u_i^{*2} = \sum \left(\frac{u_i}{\sigma_i}\right)^2$, the weighted residuals (the method is also called WLS)

problem: one needs a good estimator for the variances (we will not use this approach as working with incorrect weights does also not yield estimators that are BLUE)

b) OLS: using the OLS estimators with **heteroscedastic consistent estimators** (HC or HCC) for the covariance matrix (this is always feasible but is in theory less efficient than GLS)

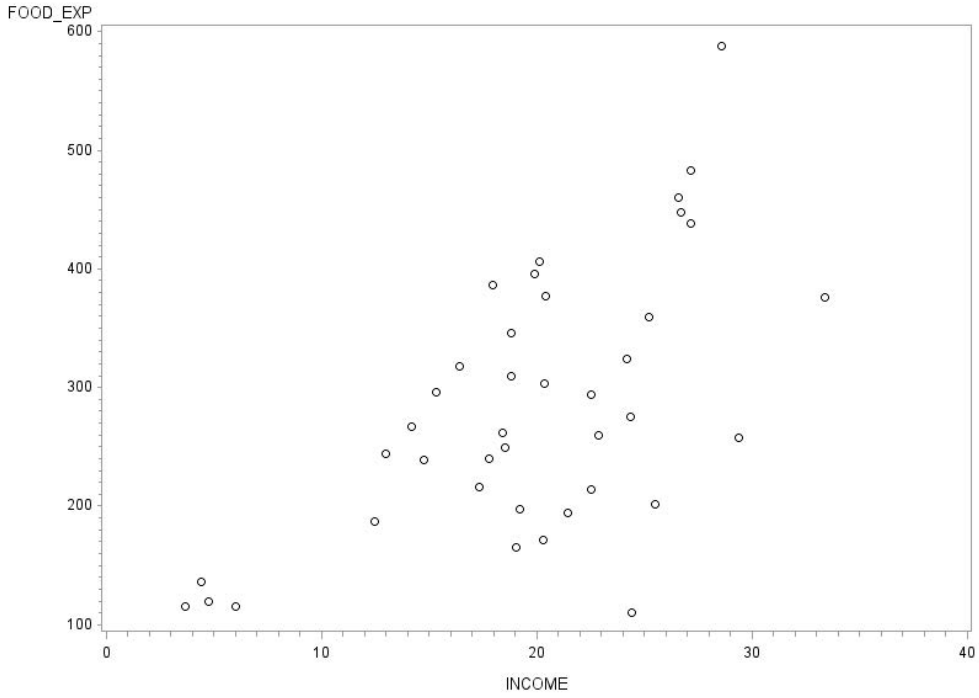
White showed that estimating the $\mathbf{V}$ in $\mathbf{cov}(\mathbf{b})^{\mathbf{OLS}}$ by

$$\hat{\mathbf{V}} = \begin{pmatrix} \hat{u}_1^2 & 0 & \cdots & 0 \\ 0 & \hat{u}_2^2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{u}_n^2 \end{pmatrix}$$

with $\hat{u}_i$ the OLS residuals, yields an asymptotically consistent estimator of the covariance matrix $\mathbf{V}$

example: model expenditures for food as a function of income

in case of simple regression a scatterplot helps to identify the problem; in case of multiple regression, heteroscedasticity is much harder to detect without statistical tests



RESULTS

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	--Heteroscedasticity Consistent-- Standard Error	t Value	Pr > |t|
Intercept	1	83.41600	43.41016	1.92	0.0622	26.76835	3.12	0.0035
INCOME	1	10.20964	2.09326	4.88	<.0001	1.76327	5.79	<.0001

as this OLS approach with robust standard errors does not require any extra assumptions, it is good practice to always ask for the robust standard errors: in case of homoscedasticity they are very similar to the classical standard errors, in case of heteroscedasticity they are more reliable

(A3) is violated: AUTOCORRELATION

time series models such as*

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \ldots \ \beta_k x_{tk} + u_t, \quad t = 1 \ldots T$$

often suffer from autocorrelated or serially correlated residuals:

$$corr(u_t, u_{t-1}) \neq 0$$

then the OLS estimators are still unbiased and consistent but not BLUE and $\mathbf{cov}(\mathbf{b^{OLS}}) \neq \sigma^2(\mathbf{X'X})^{-1}$

*it is more common to use T instead of n to denote the number of observations in time series

the solutions are similar as in the heteroscedasticity case:

a) use GLS which is BLUE (in practice FGLS (feasible GLS))
   (in case of AR(1)* residuals: Cochrane-Orcutt)

b) use OLS estimators with appropriate standard errors
   (Newey-West)

c) get rid of the autocorrelation by including lags of the dependent variable (autoregressive models)

*a first order autoregressive model

we only deal with errors that are AR(1):

$$u_t = \rho u_{t-1} + \nu_t \text{ (assume } |\rho| < 1)$$

with $E(\nu_t) = 0$, $var(\nu_t) = \sigma_\nu^2$, $cov(\nu_t, \nu_{t-s}) = 0$

then $E(u_t) = 0$, $var(u_t) = \frac{\sigma_\nu^2}{1-\rho^2}$,

$cov(u_t, u_{t-s}) = \rho^s var(u_t)$ and $corr(u_t, u_{t-s}) = \frac{\rho^s var(u_t)}{var(u_t)} = \rho^s$

so correlation decreases as time passes; let $\mathbf{V} = \mathbf{cov}(\mathbf{Y})$ then

$$\mathbf{V} = \begin{pmatrix} \frac{\sigma_\nu^2}{1-\rho^2} & \rho\frac{\sigma_\nu^2}{1-\rho^2} & \cdots & \rho^{n-1}\frac{\sigma_\nu^2}{1-\rho^2} \\ \rho\frac{\sigma_\nu^2}{1-\rho^2} & \frac{\sigma_\nu^2}{1-\rho^2} & \cdots & \rho^{n-2}\frac{\sigma_\nu^2}{1-\rho^2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1}\frac{\sigma_\nu^2}{1-\rho^2} & \rho^{n-2}\frac{\sigma_\nu^2}{1-\rho^2} & \cdots & \frac{\sigma_\nu^2}{1-\rho^2} \end{pmatrix} = \frac{\sigma_\nu^2}{1-\rho^2} \begin{pmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \cdots & 1 \end{pmatrix}$$

test for autocorrelation:

AR(1) in residuals can be detected with the **Durbin-Watson test**

$$dw = \frac{\sum_{t=2}^{n}(u_t - u_{t-1})^2}{\sum_{t=1}^{n} u_t^2} \approx 2 - 2\hat{\rho} = 2 - 2corr(u_t, u_{t-1})$$

Durbin Watson tests the following hypotheses:

$H_0 : \rho = 0$ vs $H_a : \rho > 0$ and $H_0 : \rho = 0$ vs $H_a : \rho < 0$

(remark that you hope here that you do not have to reject H_0)

example: model the inflation rate as a function of time and test whether the residuals are AR(1)

$$infln_t = \beta_0 + \beta_1 t + u_t$$

RESULTS

```
              Durbin-Watson Statistics
            Order             DW    Pr < DW    Pr > DW
              1          1.4098     <.0001     1.0000
NOTE: Pr<DW is the p-value for testing positive autocorrelation,
      and Pr>DW is the p-value for testing negative autocorrelation.
```

so the residuals suffer from significant positive autocorrelation

a) GLS: assume ρ is known, then we can transform the model to

$$\begin{aligned} y_t - \rho y_{t-1} &= \beta_0 - \rho\beta_0 + \beta_1(x_{t1} - \rho x_{t-1,1}) + \beta_2(x_{t2} - \rho x_{t-1,2}) + \\ &\quad \ldots + \beta_k(x_{tk} - \rho x_{t-1,k}) + u_t - \rho u_{t-1} \end{aligned}$$

for which the OLS estimators are BLUE

Cochrane-Orcutt use this idea to get estimators that are more efficient than OLS though they are biased (but still consistent):

- compute OLS estimates for $y_t = \beta_0 + \beta_1 x_{t1} + \ldots + \beta_k x_{tk} + u_t$
- regress u_t on u_{t-1} in $u_t = \rho u_{t-1} + \nu_t$ (no intercept!)
- use $\hat{\rho}$ to compute $y_t - \hat{\rho} y_{t-1}, x_{t1} - \hat{\rho} x_{t-1,1}, \ldots$
- use OLS to estimate this model

example: we use Cochrane-Orcutt estimation for

$$infln_t = \beta_0 + \beta_1 t + u_t$$

(remark that there are several slightly different implementations depending on how they treat the first observation)

we first show the OLS estimation results which are unbiased and consistent but the standard errors and therefore also the p-values are invalid!

OLS estimation:

Dependent Variable: INFLN

Number of Observations Read	270
Number of Observations Used	269

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.29802	0.29802	6.83	0.0095
Error	267	11.65286	0.04364		
Corrected Total	268	11.95088			

Root MSE	0.20891	R-Square	0.0249
Dependent Mean	0.25602	Adj R-Sq	0.0213
Coeff Var	81.59870		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.31432	0.02569	12.24	<.0001
t	1	-0.00042863	0.00016403	-2.61	0.0095

Estimate of Autoregressive Parameter

Lag	Coefficient	Standard Error	t Value
1	0.286426	0.058745	-4.88

one-step Cochrane-Orcutt estimation:

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	0.3158	0.0344	9.18	<.0001
t	1	-0.000433	0.000219	-1.97	0.0497

---

remark that here the Cochrane-Orcutt idea is used only once whereas the approach in R uses this idea iteratively until convergence

remark: in case not all residuals are correlated with each other but there is only a temporary relationship, one can use **moving average** residuals instead of autoregressive residuals but these models are harder to estimate and we will not deal with them here; an **MA(1) process** is defined as

$$u_t = \nu_t + \alpha \nu_{t-1} \text{ with } |\alpha| < 1$$

which implies

$$\begin{aligned} cov(u_t, u_{t-1}) &= \alpha \sigma_\nu^2 \\ cov(u_t, u_{t-s}) &= 0 \quad \text{for} \quad s \geq 2 \end{aligned}$$

b) OLS: using OLS with the correct standard errors

$$\mathbf{cov}(\mathbf{b^{OLS}}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

with $\mathbf{V}$ the cov matrix of the responses (no longer a diagonal matrix)

Newey and West have developed a general approach to estimate $\mathbf{X}'\mathbf{V}\mathbf{X}$ which assumes that the residuals are autocorrelated and heteroscedastic but the parameters of the autocorrelation function are NOT required: only the maximum lag L and the decay function by which the correlation is assumed to diminish over time

there are many ways to select a decay function and a maximum lag, we use the default HAC (heteroscedasticity and autocorrelation consistent) version in the R package sandwich

RESULTS

```
ols estimation:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.3143155  0.0256885  12.236  < 2e-16 ***
time        -0.0004286  0.0001640  -2.613  0.00948 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2089 on 267 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.02494,Adjusted R-squared:  0.02128
F-statistic: 6.828 on 1 and 267 DF,  p-value: 0.009481



with the default Newey West standard errors:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept)  0.31431548  0.03543437  8.8704  < 2e-16 ***
time        -0.00042863  0.00022204 -1.9304  0.05461 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c) get rid of the autocorrelated errors: add lagged terms of the response variable (autoregressive = regressing on itself)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \gamma_1 x_{1t} + \gamma_2 x_{2t} \ldots + u_t$$

remark that the DW test is no longer valid in autoregressive models but the Breusch-Godfrey (also called Lagrange multiplier) test can always be used; it also tests whether the residuals are AR(1)* (we do not look at the details):

$H_0 : \rho = 0$ vs $H_a : \rho \neq 0$ (hope you do not have to reject H_0!)

*it can also be used for AR(m) residuals, we use it only for m=1

example: use the Breusch-Godfrey test on the residuals of

$$infln_t = \beta_0 + \beta_1 t + u_t$$

RESULTS

Godfrey's Serial Correlation Test		
Alternative	LM	Pr > LM
AR(1)	22.1878	<.0001

which shows again (see DW test) significant autocorrelation; check whether adding 1 lag removes the autocorrelation in the residuals:

$$infln_t = \beta_0 + \beta_1 t + \beta_2\ infln_{t-1} + u_t$$

RESULTS

Godfrey's Serial Correlation Test		
Alternative	LM	Pr > LM
AR(1)	8.4646	0.0036

as there is still autocorrelation in the residuals, add another lag:

$$infln_t = \beta_0 + \beta_1 t + \beta_2\ infln_{t-1} + \beta_3\ infln_{t-2} + u_t$$

RESULTS

Godfrey's Serial Correlation Test

Alternative	LM	Pr > LM
AR(1)	1.7968	0.1801

now the autocorrelation is gone and the OLS estimates and inference can be used

stochastic regressors

until now we have considered the regressors or explanatory variables as fixed values (except when including lagged dependent variables in the previous slides or with stochastic factor levels in anova)

in econometric models (some of) the regressors should often be considered as stochastic variables as their values will change too if one takes another sample

the model looks similar as before but, depending on the application, the x can have fixed values or be stochastic variables

$$y = \beta_0 + \beta_1 x_1 + \ldots \beta_k x_k + u$$

it is very important to distinguish the following situations:

- only **exogenous** regressors: $corr(x_j, u) = 0, \forall j \in (1 \ldots k)^*$
  - if there is no endogeneity everything remains valid
  - so if (A1), (A2) and (A3) are satisfied as well (note that these should now hold **conditional on given values of** $x_1, x_2, \ldots x_k$), the OLS estimators are still unbiased and consistent and $\mathbf{cov}(\mathbf{b^{OLS}}) = \sigma^2 \ (\mathbf{X'X})^{-1}$

*strictly speaking we do not need zero correlation but independence; 2 stochastic variables are independent if there is no relation between them whereas correlation only looks at linear relationships - independence implies zero correlation but not vice versa

- **endogeneity:** $corr(x_j, u) \neq 0$ for at least one regressor x_j
  - a regressor x_j which is correlated with the residuals is called an endogenous regressor
  - for an endogenous regression the b_j^{OLS} estimator is **biased** with

$$E\left(b_j^{OLS}\right) = \beta_j + \frac{cov(x_j, u)}{var(x_j)}$$

    and the estimator is even **inconsistent**!
  - intuitively it means that the effect of a change in x_j on y is hard to measure because a change in x_j also affects the residual u

endogeneity typically occurs in the following 4 situations:

- **dynamic models**: autoregressive model with autocorrelated residuals (add lags of the response variables to get rid of the autocorrelated residuals)

- **simultaneity or reverse causality**:
  in $demand = \beta_0 + \beta_1\ price + u$ one can expect the price to be influenced by the demand too, so $price = \delta_0 + \delta_1 demand + \nu$
  solving for price yields $price = \frac{\delta_0+\delta_1\beta_1}{1-\delta_1\beta_1} + \frac{\delta_1}{1-\delta_1\beta_1}u + \frac{\nu}{1-\delta_1\beta_1}$
  assuming u and ν are independent, we get:
  $cov(price, u) = cov(\frac{\delta_1}{1-\delta_1\beta_1}u, u) = \frac{\delta_1}{1-\delta_1\beta_1}var(u)$

– **omitted variable bias** (similar as with fixed regressors): assume that one estimates $salary = \beta_0 + \beta_1\, education + u$
it is likely that the years of education and the salary are both correlated with ability which is omitted and therefore appears in the residual

– **measurement error** or **error in variables**:
assume the correct relation is

$$y = \beta_0 + \beta_1 x + u \text{ with } u \sim N(0, \sigma^2)$$

but x cannot directly be observed and the proxy
$x^* = x + \nu$ is used with ν i.i.d. and $N(0, \sigma_\nu^2)$

let $y = \beta_0 + \beta_1 x^* + \epsilon$ be the model that is estimated, then

$$\begin{aligned} y &= \beta_0 + \beta_1 x^* + \epsilon = \beta_0 + \beta_1 (x + \nu) + \epsilon \\ &= \beta_0 + \beta_1 x + \beta_1 \nu + \epsilon = \beta_0 + \beta_1 x + u \end{aligned}$$

which leads to

$$
\begin{aligned}
cov(x^*, \epsilon) &= cov(x + \nu, u - \beta_1 \nu) = -\beta_1 \sigma_\nu^2 \\
\text{and as } var(x^*) &= \sigma_x^2 + \sigma_\nu^2 \\
E(b_1^{OLS}) &= \beta_1 - \frac{\beta_1 \sigma_\nu^2}{\sigma_x^2 + \sigma_\nu^2} \\
&= \beta_1 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_\nu^2}
\end{aligned}
$$

which is called **softening or attenuation bias**

Instrumental Variables Estimation:

if $corr(x_j, u) \neq 0$ one can get consistent estimators by using IV:
a variable z is a good instrumental variable for x_j if

- it is uncorrelated with the residuals: $corr(z, u) = 0$
- it is highly correlated with x_j: $corr(z, x_j) > 0$
- it has no direct effect on y

in case there are several endogenous regressors, use at least one instrumental variable for each of these regressors

examples: it is often very difficult to find justifiable instruments!

- salary (y) and education (x) are both depending on ability which is not directly observable: we can use the mother's education (z) to control for endogeneity

- consumption (y) and income (x) are both determined by macroeconomic factors: we can use investments (z) to control for endogeneity

IV estimation is a 2 step procedure (2 stage LS $=$ 2SLS):

assume we want to estimate

$y = \beta_0 + \beta_1 x + u$ where $cov(x, u) \neq 0$

assume further that z is an instrumental variable, then we fit

$x = \gamma_0 + \gamma_1 z + \epsilon$

and compute the predicted values $\hat{x}$ which are used in

$y = \beta_0 + \beta_1 \hat{x} + u^*$

then the OLS estimator of this model is the **IV estimator** b_1^{IV}

(remark that the standard deviation has to be adapted to take into account the extra uncertainty in $\hat{x}$)

this estimator b_1^{IV} is a consistent estimator for β_1 but the variance is larger than in standard OLS

$$var(b_1^{IV}) \approx \frac{var(b_1^{OLS})}{corr^2(z,x)}$$

so the smaller the correlation between x and z, the less precise the estimator for β_1

similar results hold for more endogenous variables and more instrumental variables where one replaces all endogenous variables by their predicted values based on all instrumental variables and all exogenous variables

example: $ln(wage) = \beta_0 + \beta_1 educ + u$

regress $\ln(wage)$ on education based on a sample of 428 women

as we can anticipate that education is positively correlated with ability which is omitted, the OLS estimate is likely to overestimate the impact of education

we use mother-education as instrumental variable

OLS estimation neglecting endogeneity:

RESULTS

Dependent Variable: lnwage
Number of Observations Read 428

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	26.32642	26.32642	56.93	<.0001
Error	426	197.00102	0.46244		
Corrected Total	427	223.32744			

Root MSE	0.68003	R-Square	0.1179
Dependent Mean	1.19017	Adj R-Sq	0.1158
Coeff Var	57.13724		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.18520	0.18523	-1.00	0.3180
educ	1	0.10865	0.01440	7.55	<.0001

2SLS with *mother's education* as instrumental variable:

Two-Stage Least Squares Estimation
Dependent Variable lnwage

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.496428	0.496428	1.02	0.3138
Error	426	207.9598	0.488169		
Corrected Total	427	223.3274			

Root MSE	0.69869	R-Square	0.00238
Dependent Mean	1.19017	Adj R-Sq	0.00004
Coeff Var	58.70495		

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.702174	0.485099	1.45	0.1485
educ	1	0.038550	0.038228	1.01	0.3138

using also *father's education* as extra instrumental variable:

RESULTS

Two-Stage Least Squares Estimation
Dependent Variable lnwage

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.182931	1.182931	2.46	0.1172
Error	426	204.5444	0.480151		
Corrected Total	427	223.3274			

Root MSE	0.69293	R-Square	0.00575
Dependent Mean	1.19017	Adj R-Sq	0.00342
Coeff Var	58.22088		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.551020	0.408581	1.35	0.1782
educ	1	0.050490	0.032168	1.57	0.1172

IV and OLS are both consistent estimators but as IV is less efficient than OLS, it should only be used if necessary:

test for endogeneity by the **Wu-Hausman test**:

$H_0: \ cov(x_j, u) = 0, \forall j \quad$ vs. $\quad H_a: \ \exists j : cov(x_j, u) \neq 0$

(so you hope you do not have to reject H_0 to be able to use OLS)

as the test statistic compares the OLS and IV estimators, it depends on the instrumental variables that are used! (we do not deal with the details of this test)

example: Wu-Hausman test for $ln(wage) = \beta_0 + \beta_1 educ + u$

- with only mother-education as IV:

RESULTS

```
Diagnostic tests:
                 df1 df2 statistic p-value
Wu-Hausman         1 425     4.206  0.0409 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6987 on 426 degrees of freedom
Multiple R-Squared: 0.06881,Adjusted R-squared: 0.06663
Wald test: 1.017 on 1 and 426 DF,  p-value: 0.3138
```

- with mother and father education as IV:

RESULTS

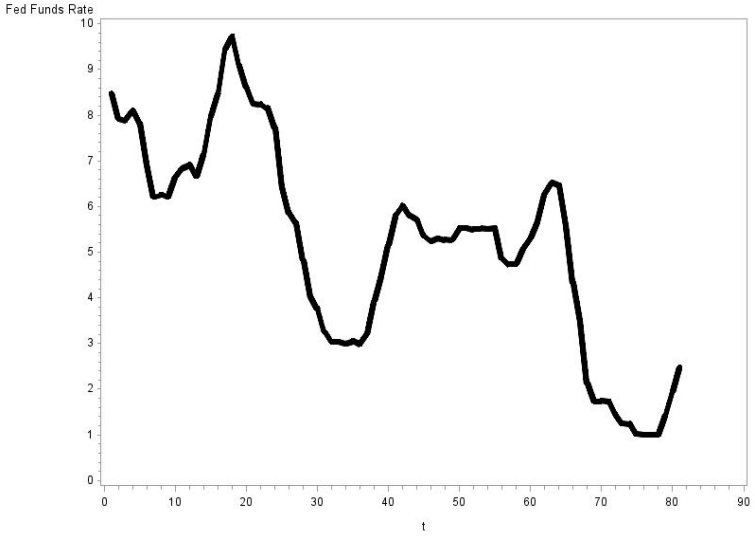
```
Diagnostic tests:
                 df1 df2 statistic p-value
Wu-Hausman         1 425     4.319  0.0383 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6929 on 426 degrees of freedom
Multiple R-Squared: 0.08411,Adjusted R-squared: 0.08196
Wald test: 2.464 on 1 and 426 DF,  p-value: 0.1172
```

so we reject the null hypothesis in both cases: there is endogeneity and the IV estimators will have to be used

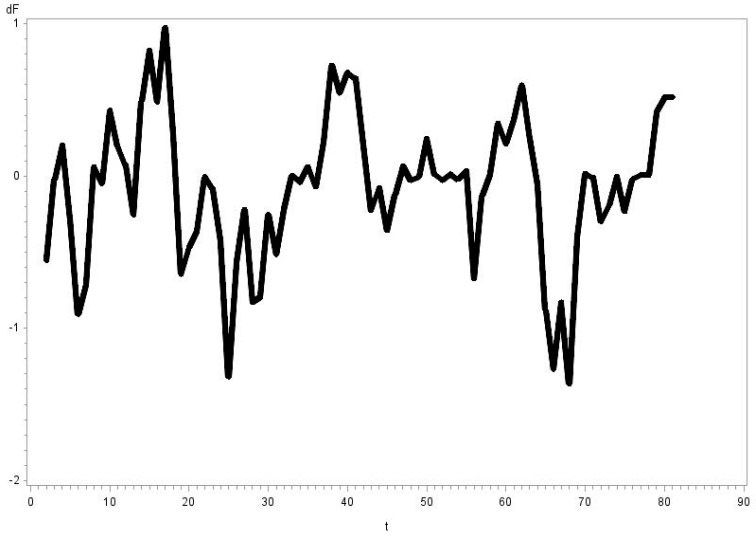
time series analysis

it is always best to start by plotting the time series to check for problems and/or get ideas about how to model the series

F_t



$\Delta F_t = F_t - F_{t-1}$



stationarity

a time series y_t $(t = 1 \ldots T)$ is stationary if

$$
\begin{aligned}
E(y_t) &= \mu && \text{(is the same for all t)} \\
& && \text{(is called mean stationary)} \\
var(y_t) &= \sigma^2 && \text{(is finite and the same for all t)} \\
& && \text{(is called variance stationary)} \\
cov(y_t, y_{t-s}) &= \gamma_s && \text{(depends only on the lag s, not on t } (s = 1 \ldots T) \\
& && \text{(is called covariance stationary)}
\end{aligned}
$$

so a time series is stationary if you can start the series at any time and the properties will remain the same

some frequently used time series*: **white noise**

$y_t = u_t$ with u_t i.i.d. with mean 0 and variance σ^2

$$\begin{aligned} E(y_t) &= 0 \\ var(y_t) &= \sigma^2 \\ cov(y_t, y_{t-s}) &= 0 \end{aligned}$$

so white noise is stationary

remark that the residuals in a classic regression model are assumed to be white noise which is normally distributed; we will assume that all white noise is normally distributed

*check the applet on time series

some frequently used time series: **random walk**

$$y_t = y_{t-1} + u_t, \quad \text{with} \quad u_t \ \text{white noise} \quad (\text{assume } y_0 = 0)$$

the residual in time series analysis is often called the **shock** or **innovation** as it is unpredictable; remark that

$$\begin{aligned} E(y_t) &= 0 \\ var(y_t) &= t\sigma^2 \\ cov(y_t, y_{t-s}) &= var(y_{t-s}) = (t-s)\sigma^2 \end{aligned}$$

so a random walk is nonstationary but $\Delta y_t = y_t - y_{t-1}$ is stationary; a series that becomes stationary after taking first differences is called **difference stationary** or **integrated of order** $1 = \mathrm{I}(1)^*$

*if the time series can be made stationary by differencing it m times, it is called I(m)

some frequently used time series: **AR(1)**

$$y_t = \alpha + \phi y_{t-1} + u_t, \quad \text{with} \quad u_t \ \text{white noise}$$

(so a random walk is an AR(1) model for which $\alpha = 0$ and $\phi = 1$)

$$\begin{aligned} E(y_t) &= \alpha(1 + \phi + \phi^2 + \phi^3 + \ldots + \phi^{t-1}) \\ var(y_t) &= \sigma^2(1 + \phi^2 + \phi^4 + \ldots + \phi^{2(t-1)}) \\ cov(y_t, y_{t-s}) &= \phi^s cov(y_{t-s}, y_{t-s}) = \phi^s var(y_{t-s}) \end{aligned}$$

the behavior depends on the value of ϕ:

- if $|\phi| \geq 1$ the mean and variance will go to infinity and the time series is nonstationary

- if $|\phi| < 1$, we can use that $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \ldots$ to obtain

$$\begin{aligned} E(y_t) &= \frac{\alpha}{1-\phi} \\ var(y_t) &= \frac{\sigma^2}{1-\phi^2} \\ cov(y_t, y_{t-s}) &= \frac{\phi^s \sigma^2}{1-\phi^2} \end{aligned}$$

  so in this case the AR(1) time series is indeed stationary

  remark that in this case α is related to the nonzero mean of the series whereas it reflects the deterministic trend if $\phi = 1$

different types of nonstationarity:

- **deterministic trend:** if the trend can be described by a function of t as in $y_t = \alpha + \beta t + u_t$ the series is nonstationary but it is called **stationary around a linear trend** or **trend stationary**: except for small variations, the series is predictable

  the model $y_t = \alpha + \beta t + \phi y_{t-1} + u_t$ is also called trend stationary if $|\phi| < 1$; otherwise it is called nonstationary as the autoregressive term produces a stochastic trend

- **unit-root nonstationarity:**

  consider the AR(1) model $y_t = \alpha + \phi y_{t-1} + u_t$ (assume $\phi > 0$)

  this series is unit-root nonstationary if $\phi = 1$ (or larger)

  the model is often rewritten by subtracting y_{t-1} from both sides:

$$\begin{aligned} y_t - y_{t-1} &= \alpha + \phi y_{t-1} - y_{t-1} + u_t \\ \Delta y_t &= \alpha + \Phi y_{t-1} + u_t \quad \text{with } \Phi = \phi - 1 \end{aligned}$$

  so testing whether $\phi = 1$ is equivalent with testing $\Phi = 0$

so we want to test: $H_0 : \Phi = 0$ vs $H_a : \Phi < 0$

(we assume a unit root until strong evidence against it!)

the usual teststatistic for the significance of the lagged term parameter will be used but:

- the test can only be used if the residuals are uncorrelated
- the teststatistic is no longer t-distributed under H_0
  (so the usual p-values are not valid!)

to get rid of autocorrelated residuals, one has to add autoregressive terms; select the appropriate number of lags by the Breusch-Godfrey test (remark that we can test the residuals of the model in y_t or of the model in Δy_t as those residuals are the same)

suppose we end up with an AR(m) model with uncorrelated residuals:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_m y_{t-m} + u_t$$

it can be shown that such an AR(m) model is unit-root nonstationary if $\phi_1 + \phi_2 + \ldots + \phi_m = 1$

the model can again be rewritten (but it is more complicated) as

$$\Delta y_t = \alpha + \Phi y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{m-1} \Delta y_{t-m+1} + u_t$$

with $\Phi = \phi_1 + \phi_2 + \ldots \phi_m - 1$ and the γ's functions of the ϕ's
and $\Delta y_{t-s} = y_{t-s} - y_{t-s-1}$

so also after adding lagged terms to get rid of autocorrelated residuals, the significance of the parameter Φ of the y_{t-1} term still indicates whether the series y_t is unit-root nonstationary or not

to test the significance of Φ, the regular teststatistic is used (see regression, hypothesis testing) but because y_t is not stationary under H_0, it is no longer t distributed under H_0 (and it is therefore called the τ-teststatistic)

the distribution of the τ-teststatistic depends on whether an intercept and a deterministic trend are required in the model or not, so the p-values for the significance of Φ will be different for*†:

$$\begin{aligned}
\Delta y_t &= \Phi y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{m-1} \Delta y_{t-m+1} + u_t \\
\Delta y_t &= \alpha + \Phi y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{m-1} \Delta y_{t-m+1} + u_t \\
\Delta y_t &= \alpha + \beta t + \Phi y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_{m-1} \Delta y_{t-m+1} + u_t
\end{aligned}$$

*several people have investigated these distributions (and have come up with slightly different p-values!)

†the models are called respectively *the model with no drift and no trend*, *the model with drift and no trend*, *the model with drift and trend*

example: is the Federal Funds Rate F a stationary series?

remark that it is easy to get results for several lags and all 3 models with the adf.test in R; if all these tests have small p-values (or if all have large p-values), there is no need to check for the correct model and number of lags as all p-values then lead to the same conclusion

we check which test we need: start by using Breusch-Godfrey to check for autocorrelated residuals in

$$F_t = \alpha + \beta t + \phi F_{t-1} + u_t$$

(based on the plot of F_t, we add an intercept and trend)

RESULTS

Godfrey's Serial Correlation Test		
Alternative	LM	Pr > LM
AR(1)	38.0243	<.0001

as there is significant autocorrelation, we check whether the autocorrelation can be removed by including another lag; so we test the residuals of $F_t = \alpha + \beta t + \phi F_{t-1} + \phi_2 F_{t-2} + u_t$

RESULTS

Godfrey's Serial Correlation Test		
Alternative	LM	Pr > LM
AR(1)	0.0767	0.7819

as the autocorrelation is gone, we fit the model

$$\Delta F_t = \alpha + \beta t + \Phi F_{t-1} + \gamma_1 \Delta F_{t-1} + u_t$$

RESULTS

OLS Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	0.6529	0.2293	2.85	0.0057
t	1	-0.005842	0.002550	-2.29	0.0248
lagf	1	-0.0812	0.0259	-3.14	0.0024
lagdf	1	0.6946	0.0836	8.31	<.0001

so we need indeed an intercept and deterministic trend (cfr *with drift and trend*) and the appropriate teststatistic is -3.14 but the corresponding p-value is not valid

correct p-values obtained with the Augmented Dickey Fuller test:

RESULTS

```
Augmented Dickey-Fuller Unit Root Tests

Type                        Tau       Pr < Tau
no drift no trend       -1.1631        0.2214
with drift no trend     -2.0903        0.2489
with drift and trend    -3.1399        0.1044
```

which shows, with p-value=0.1044, that we have to accept H_0 and that F_t is unit-root nonstationary

is the F series I(1)? or, equivalently, is $\Delta F_t = F_t - F_{t-1}$ stationary?

you can check that there is no need to add lags here to get rid of autocorrelated residuals

RESULTS

Augmented Dickey-Fuller Unit Root Tests

Type	Tau	Pr < Tau
no drift no trend	-4.0074	0.0001
with drift no trend	-3.9875	0.0024
with drift and trend	-3.9617	0.0139

based on the time series plot of ΔF, the second model looks most appropriate*, so $\tau = -3.9875$ and the p-value is 0.0024

we can now reject H_0 and conclude that the series ΔF is stationary or that F is I(1)

*one can also check that the intercept is not significantly different from 0

regression with time series: consider the regression model*

$$y_t = \alpha + \beta x_t + u_t$$

then y_t and x_t should have the same properties for this to work!

- if y_t and x_t are both **stationary** then the classical OLS estimates and inference can be used
- if y_t and x_t are both **trend stationary**, detrend them and use the residuals after detrending; the classical OLS estimates and inference can be used

*we look here at the simplest case but the same holds for *distributed lag models* where lags of x_t are included as regressors as well as for *autoregressive distributed lag (ADL) models* where lags of y_t are added as regressors to a distributed lag model

- y_t and x_t have **unit roots***:
  - OLS will often give very misleading results: even if the series are completely unrelated, the R^2 can be quite high (**spurious regression**) $\Rightarrow$ if the series are I(1), use first differences
  - OLS still works if the residuals in $y_t = \alpha + \beta x_t + u_t$ are stationary; then the series x_t and y_t are said to be **cointegrated**
    (so the nonstationarity of x_t and y_t cancel each other out; it is said that they have a common stochastic trend or that there is an equilibrium equation between them)

*check the applet on time series

examples of cointegrated series:

- let x_t be the price of regular oranges and y_t the price of organic oranges; then these prices are related and we expect y_t to be somewhat larger than x_t

  x_t and y_t can be expected to be cointegrated as the difference between the 2 prices will not become large: if the gap becomes large, people no longer pay the extra premium for organic oranges and buy the regular oranges

- short and long term interest rates

Engle-Granger test for cointegration: H_0 : no cointegration

no extra tests are required, we test the residuals of the OLS regression $y_t = \alpha + \beta x_t + u_t$ for a unit root*
(if there is no unit root, then the series are cointegrated)

example: the bond rate (B) is also I(1) and therefore unit root nonstationary as is the Federal Funds rate (F); are these series cointegrated?

we use an ADF test on the residuals of $F_t = \beta_0 + \beta_1 B_t + u_t$

it can be checked that one needs to add 1 lag to remove autocorrelation in the residuals of $\hat{u}_t = \phi_1 \hat{u}_{t-1} + \nu_t$, the appropriate ADF tests are therefore

*there exist corrected p-values that take into account that the residuals are estimated (neglected here)

RESULTS

Augmented Dickey-Fuller Unit Root Tests

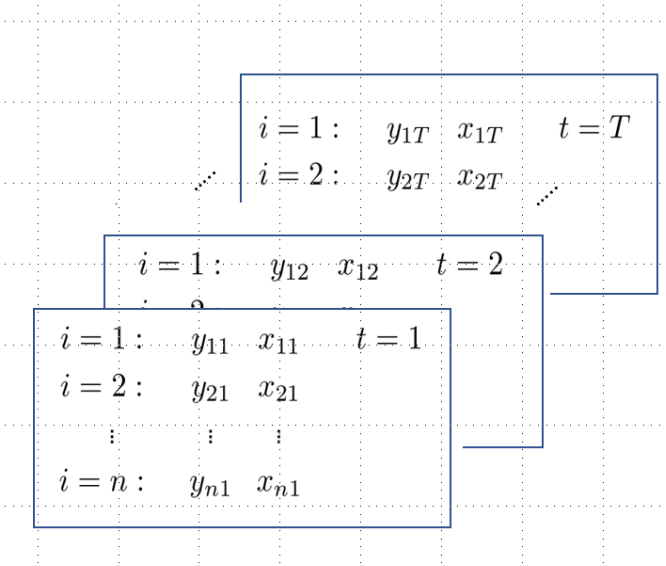
Type	Tau	Pr < Tau
no drift no trend	-4.2336	<.0001
with drift no trend	-4.2129	0.0012
with drift and trend	-4.1549	0.0080

no matter which test is most appropriate (although for residuals no mean or trend is expected), the p-value is small and we can reject the null hypothesis that $\Phi = 0$; we conclude that B and F are cointegrated

panel data

n units (individuals, countries, households, firms, ...) are measured on T occasions; a model with 1 regressor might look like this:

$$y_{it} = \alpha + \beta x_{it} + u_{it}, \text{ for } i = 1 \ldots n \text{ and } t = 1 \ldots T$$



i	t	y	x
1	1	y_{11}	x_{11}
⋮	⋮	⋮	⋮
n	1	y_{n1}	x_{n1}
1	2	y_{12}	x_{12}
⋮	⋮	⋮	⋮
n	2	y_{n2}	x_{n2}
⋮	⋮	⋮	⋮
1	T	y_{1T}	x_{1T}
⋮	⋮	⋮	⋮
n	T	y_{nT}	x_{nT}

examples:

- relate the evolution of the house prices in different cities to the unemployment rate ($i = 1 \ldots n, t = 1 \ldots T$):

$$\text{HousePrice}_{it} = \alpha + \beta \ \text{UnemploymentRate}_{it} + u_{it}$$

- consider the effect of the stock market value and capital on Investments

$$\text{Investments}_{it} = \alpha + \beta_1 \text{Capital}_{it} + \beta_2 \text{Value}_{it} + u_{it}$$

(Grunfeld dataset: data for n=10 firms and T=20 years)

the error term u_{it} can be rewritten as the sum of a term ζ_t that depends only on t, a term ν_i that depends only on i and a term η_{it} that depends on both i and t:

$$u_{it} = \zeta_t + \nu_i + \eta_{it} \qquad = \textbf{error component model}$$

- ζ_t denotes the time trend that was not captured in the model; we can easily get rid of this error by including a function of time or by including $T-1$ dummies:

$$y_{it} = \alpha + \beta x_{it} + \delta_2 D_2 + \delta_3 D_3 + \ldots \delta_T D_T + \nu_i + \eta_{it}$$

from now on we assume that there is no unobserved time trend or that an appropriate function of time has been included

- ν_i is called the **unobserved heterogeneity**, we assume that these are independent of each other, $E(\nu_i) = 0$ and $var(\nu_i) = \sigma_\nu^2$

- η_{it} is called the **idiosyncratic error**, we assume that these are independent of each other, $E(\eta_{it}) = 0$ and $var(\eta_{it}) = \sigma_\eta^2$

- we assume that $cov(\nu_i, \eta_{it}) = 0, \forall i, t$

so the model is $y_{it} = \alpha + \beta x_{it} + u_{it}$ with $u_{it} = \nu_i + \eta_{it}$

we have to check:

- whether the residuals suffer from heteroscedasticity or autocorrelation (then the OLS estimators are still unbiased and consistent but not BLUE and the standard cov matrix is not valid*)
- whether there is endogeneity (if the stochastic regressors are correlated with the residuals, the OLS estimators are biased and inconsistent!); we will assume throughout that $corr(x_{it}, \eta_{it}) = 0, \forall i, t$ but we worry about $corr(x_{it}, \nu_i)$
- whether the time series are stationary or cointegrated
  (we assume throughout that this is the case, otherwise use detrended variables and/or differences of variables)

*$\mathbf{cov}(\mathbf{b^{OLS}}) = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$ with $\mathbf{V} = \mathbf{cov}(\mathbf{u})$ ($\mathbf{cov}(\mathbf{u}) \neq \mathbf{cov}(\mathbf{y})$ as x is stochastic)

exogeneous case: assuming $corr(x_{it}, \nu_i) = 0$

residuals related to the same unit are likely to be correlated*:

$$cov(u_{is}, u_{it}) = cov(\nu_i + \eta_{is}, \nu_i + \eta_{it}) = \sigma_\nu^2 \text{ for } s \neq t$$

as before we have several options:

a) we can use **OLS** which is unbiased and consistent but no longer BLUE and the formula for the standard errors is incorrect; use **cluster corrected standard errors** (use an estimate for $\mathbf{V} = \mathbf{cov}(\mathbf{u})$ to compute $\mathbf{cov}(\mathbf{b^{OLS}})$ similarly as White and NeweyWest) (different approaches with all slightly different results)

*see the compound symmetric structure in Part 1, here often in combination with heteroscedasticity

b) use **FGLS = Random Effects estimation**

GLS: transform the variables by

$$\begin{aligned} y_{it}^* &= y_{it} - \psi \bar{y}_{i.} \text{ with } \psi = 1 - \sqrt{\frac{\sigma_\eta^2}{T\sigma_\nu^2 + \sigma_\eta^2}} \\ x_{it}^* &= x_{it} - \psi \bar{x}_{i.} \\ u_{it}^* &= u_{it} - \psi \bar{u}_{i.} \qquad \text{(cfr: if } \sigma_\nu^2 = 0, \psi = 0) \end{aligned}$$

these residuals u_{it}^* have constant variance and are uncorrelated, so the OLS estimators of this model are BLUE

use OLS estimates of the original model to get estimates for σ_η^2 and σ_ν^2 (which is complicated here)(so we end up with Feasible GLS which is not BLUE but consistent and more efficient than OLS with corrected cov matrix)

example: the Grunfeld dataset

to compare with, we first compute the OLS estimators with default standard errors:

RESULTS

Dependent Variable: investments

Model Description

Number of Cross Sections	10
Time Series Length	20

Fit Statistics

SSE	1749127.640	DFE	197
MSE	8878.8205	Root MSE	94.2275
R-Square	0.8131		

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-43.0245	9.4979	-4.53	<.0001
value	1	0.115374	0.00583	19.79	<.0001
capital	1	0.231931	0.0255	9.11	<.0001

with cluster corrected standard errors:

RESULTS

Dependent Variable: investments

Model Description

Estimation Method	Pooled
Number of Cross Sections	10
Time Series Length	20
Hetero. Corr. Cov. Matrix Estimator	4

Fit Statistics

SSE	1749127.640	DFE	197
MSE	8878.8205	Root MSE	94.2275
R-Square	0.8131		

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-43.0245	19.1950	-2.24	0.0261
value	1	0.115374	0.0149	7.76	<.0001
capital	1	0.231931	0.0798	2.90	0.0041

the random effects estimates:

RESULTS

Fit Statistics

SSE	546061.5552	DFE	197
MSE	2771.8861	Root MSE	52.6487
R-Square	0.7696		

Variance Component Estimates

Variance Component for Cross Sections	7724.105
Variance Component for Error	2781.144

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-57.9342	29.9464	-1.93	0.0545
value	1	0.109493	0.0106	10.36	<.0001
capital	1	0.308856	0.0172	17.98	<.0001

remark that $\hat{\sigma}_\nu^2 = 7724.105$, $\hat{\sigma}_\eta^2 = 2781.144$

endogeneous case: $corr(\nu_i, x_{it}) \neq 0$

OLS and FGLS=RE are both biased and inconsistent!

there are several solutions, among which

- **fixed effect estimation = individual effects estimation**
- **within group estimation**
- IV estimation (not dealt with here)

the first 2 approaches lead to the same estimates and significance for the parameters of the *time varying* regressors

we have $y_{it} = \alpha + \beta x_{it} + u_{it}$ with $u_{it} = \nu_i + \eta_{it}$ and $corr(x_{it}, \nu_i) \neq 0$

get rid of the unobserved heterogeneity by transforming the model:

- **individual effects estimation or fixed effects estimation:** capture the individual differences in individual intercepts:
  $y_{it} = \alpha_i + \beta x_{it} + u_{it}$ and $u_{it} \approx \eta_{it}$

- **the within group estimator**: remove the individual differences by fitting a model with *time demeaned* variables:
  $y^*_{it} = y_{it} - \bar{y}_{i.}$, $x^*_{it} = x_{it} - \bar{x}_{i.}$, $u^*_{it} = u_{it} - \bar{u}_{i.} = \eta_{it} - \bar{\eta}_{i.}$

as the endogeneity has been removed, we can use

- standard OLS if there is no heteroscedascticity or autocorrelation
- OLS + cluster corrected covariance matrix if there is heteroscedascticity or autocorrelation (if is safe to always use these)

remark that

- one cannot include regressors that only depend on i and not on t (in the fixed effects model they cannot be distinguished from the individual intercepts and in the demeaned model they drop out)
- in the model with demeaned variables there is no need to estimate the (often large number of) individual intercepts of the fixed effects model (which is less of an issue with computers nowadays)
- the individual effects model is almost always to be preferred over the regression model with only 1 intercept (called the pooled model): if the α_i are significantly different from each other (almost always the case!), the estimator for β will suffer from omitted variable bias if only 1 intercept is included and this bias can become very large

remark that the terms *fixed effects* and *random effects* are used here differently than in ANOVA (which is quite confusing):

- in ANOVA *random effects* means that factor levels are chosen at random whereas *fixed effects* means that we are interested in the individual factor levels (so the terminology is determined by the set up of the experiment)

- in panel data the fixed effects model and fixed effects estimators are used to solve an endogeneity problem and the random effects model and random effects estimators are used to get efficient estimators in case of exogeneity (so the terminology is determined by the estimation method)

example: on the Grunfeld dataset

the fixed effect estimates (no corrected standard errors):

RESULTS

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	8837969.792	803451.799	288.89	<.0001
Error	188	522855.136	2781.144		
Corrected Total	199	9360824.928			

R-Square	Coeff Var	Root MSE	investments Mean
0.944144	36.14401	52.73656	145.9068

Source	DF	Type III SS	Mean Square	F Value	Pr > F
I	9	1226272.503	136252.500	48.99	<.0001
value	1	238453.570	238453.570	85.74	<.0001
capital	1	889468.669	889468.669	319.82	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		-6.5462682	B	11.81986618	-0.55	0.5803
I	1	-62.5971812	B	50.30696924	-1.24	0.2149
I	2	107.4086770	B	26.92852116	3.99	<.0001
I	3	-228.5724612	B	26.49588483	-8.63	<.0001
I	4	-21.0887013	B	18.03866020	-1.17	0.2439
I	5	-108.7706332	B	18.42532863	-5.90	<.0001
I	6	-16.5272964	B	17.11168236	-0.97	0.3354
I	7	-60.1366577	B	17.43587876	-3.45	0.0007
I	8	-50.8123227	B	17.97745010	-2.83	0.0052
I	9	-80.7307422	B	17.36690092	-4.65	<.0001
I	10	0.0000000	B	.	.	.
value		0.1097711		0.01185490	9.26	<.0001
capital		0.3106441		0.01737039	17.88	<.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

the within group estimates:

first without and then with cluster corrected standard errors:

RESULTS

Model Description

Number of Cross Sections	10
Time Series Length	20

Fit Statistics

SSE	522855.1364	DFE	188
MSE	2781.1443	Root MSE	52.7366
R-Square	0.9441		

F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
9	188	48.99	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.54627	11.8199	-0.55	0.5803
value	1	0.109771	0.0119	9.26	<.0001
capital	1	0.310644	0.0174	17.88	<.0001

Hetero. Corr. Cov. Matrix Estimator

Fit Statistics

SSE	522855.1364	DFE	188
MSE	2781.1443	Root MSE	52.7366
R-Square	0.9441		

F Test for No Fixed Effects

Num DF	Den DF	F Value	Pr > F
9	188	48.99	<.0001

Parameter Estimates

Variable	DF	Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-6.54627	1.2125	-5.40	<.0001
value	1	0.109771	0.0143	7.69	<.0001
capital	1	0.310644	0.0498	6.24	<.0001

in case $corr(\nu_i, x_{it}) = 0$ both the fixed effects estimator and the random effects estimator can be used but the random effects estimator is much more efficient than the fixed effect estimator (and can estimate the effect of time-constant regressors), so the random effects estimator is to be used whenever possible

use the **Hausman test*** to test $H_0 : corr(\nu_i, x_{it}) = 0$

- if H_0 is accepted, use the random effects estimator
- if H_0 is rejected, use the fixed effects estimator

the test statistic compares $\hat{\beta}^{RE}$ and $\hat{\beta}^{FE}$, these are both consistent under H_0 but only β^{FE} is consistent if H_0 is not true

*there are several versions which give slightly different p-values, we do not look at the details of the test

example: on the Grunfeld dataset:

RESULTS

```
        Hausman Test for Random Effects
Coefficients          DF    m Value    Pr > m
           2           2       1.31    0.5193
```

the Hausman test shows that there is no problem of correlation between ν_i and x_{it} and so the random effects estimator can be used

check that the random effects estimators for the β's have indeed smaller variance than the fixed effects estimators

PART 3

- LOGISTIC REGRESSION
- DURATION ANALYSIS

3a. LOGISTIC REGRESSION

LOGISTIC REGRESSION

consider a binary response variable (coded as 0 or 1):

- predict whether a firm goes bankrupt within the coming 12 months
- predict whether a person will buy the product or not
- predict whether a student will pass or fail

based on some variables (called covariates) which can be numerical or categorical

the linear probability model

$y_i = \beta_0 + \beta_1 x_{i1} + u_i$ with $y_i = 0$ or 1 and $i = 1 \ldots n$

assume $E(u_i) = 0 \Rightarrow E(y_i) = \beta_0 + \beta_1 x_{i1}$

as y_i is Bernoulli distributed; let π_i be the probability that the event will occur for the ith observation, then:

$Prob(y_i = 1) = \pi_i$ $\equiv$ probabilty of an *event*

$Prob(y_i = 0) = 1 - \pi_i$ $\equiv$ probabilty of a *nonevent*

$\Rightarrow E(y_i) = 1 \times \pi_i + 0 \times (1 - \pi_i) = \pi_i = \beta_0 + \beta_1 x_{i1}$

$\Rightarrow$ linear relationship between the covariate and prob(event)

drawbacks of this model:

- the predicted probability lies between $-\infty$ and ∞ instead of between 0 and 1

- the residuals are not normally distributed: if $y_i = 1$ then $u_i = 1 - \beta_0 - \beta_1 x_{i1}$ and if $y_i = 0$ then $u_i = -\beta_0 - \beta_1 x_{i1}$
  $\Rightarrow$ the ordinary least squares estimators are still unbiased but have only approximately a normal distribution if n is large

- heteroscedasticity (the variance depends on x_i): use weighted least squares estimators to get efficient estimators, with weights: $\frac{1}{var(y_i)} = \frac{1}{\pi_i(1-\pi_i)} \approx \frac{1}{\hat{\pi}_i(1-\hat{\pi}_i)} = \frac{1}{\hat{y}_i^{OLS}(1-\hat{y}_i^{OLS})}$

- the model is not realistic: the marginal effect of x on y is the same for all values of x
  - a S-shaped curve is more realistic
  - every cumulative distribution function has a S-shape: we will use the logistic distribution function

if we would use the normal distribution, we would end up with the *probit*-model (not covered)

simple logistic regression

$$\pi_i = E(y_i = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1})}} = \frac{e^{(\beta_0 + \beta_1 x_{i1})}}{1 + e^{(\beta_0 + \beta_1 x_{i1})}}$$

$\Rightarrow$ the probability is always between 0 and 1

$$\begin{aligned} 1 - \pi_i &= \frac{1}{1 + e^{(\beta_0 + \beta_1 x_{i1})}} \\ \frac{\pi_i}{1 - \pi_i} &= e^{\beta_0 + \beta_1 x_{i1}} \\ \ln(\frac{\pi_i}{1 - \pi_i}) &= \beta_0 + \beta_1 x_{i1} \end{aligned}$$

$\frac{\pi}{1-\pi}$ are the **odds** in favor of the event

$\ln(\frac{\pi}{1-\pi})$ is the **logit**

there exists a unique relationship between probabilities and odds:

- given $\pi \Rightarrow$ odds $= \frac{\pi}{1-\pi}$
- given odds $\Rightarrow \pi = \frac{odds}{1+odds}$

odds ratio

interpretation of the parameters of a logit model is more difficult than in linear models: the effect of an increase of x on the probability depends on the value of x

use the odds ratio instead: if x_1 increases with 1 unit, then

$$\begin{aligned} logit|_{x_1+1} &= logit|_{x_1} + \beta_1 \\ odds|_{x_1+1} &= odds|_{x_1} \times e^{\beta_1} \\ \frac{odds|_{x_1+1}}{odds|_{x_1}} &= odds\ ratio = e^{\beta_1} \end{aligned}$$

so the effect of an increase of x on the *odds* is easier to express as it is independent of the value of x

- e^{β_1} indicates how much the odds in favor of the event change if the covariate is increased with 1 unit

- for β_1 larger than 0, the odds in favor of the event will increase with $(e^{\beta_1} - 1)100\%$ if the covariate is increased with 1 unit

- for β_1 smaller than 0, the odds in favor of the event will decrease with $(1 - e^{\beta_1})100\%$ if the covariate is increased with 1 unit

estimation and inference

the ordinary least squares estimators of the linear logit model only exist if there are grouped observations (= several observations for the same x values) as for nongrouped observations we have either $\ln(\frac{1}{0}) = \infty$ or $\ln(\frac{0}{1}) = -\infty$

use the *maximum likelihood estimators*:
maximize the likelihood function w.r.t. β_0 and β_1

$$
\begin{aligned}
L &= \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i} \\
&= \prod_{i=1}^{n} \left(\frac{e^{(\beta_0+\beta_1 x_{i1})}}{1+e^{(\beta_0+\beta_1 x_{i1})}} \right)^{y_i} \left(\frac{1}{1+e^{(\beta_0+\beta_1 x_{i1})}} \right)^{1-y_i}
\end{aligned}
$$

ML looks for the parameters that maximize the probability of observing the sample at hand

notice that the model parameters cannot be estimated uniquely if there is *perfect separation*: if for all small x values the corresponding response is zero ($y = 0$) and for all large x values $y = 1$, there are infinitely many parameters that can fit the data perfectly (*the separation problem*)

- **significance of the model**
  - the **likelihood ratio test**: let L_0 be the maximum value of L under the nulhypothesis $H_0 : \beta_1 = 0$ and L_1 the maximum value of L for the complete model; under H_0*:

$$-2 \ln \frac{L_0}{L_1} = [-2 \ln(L_0)] - [-2 \ln(L_1)] \overset{n \to \infty}{\sim} \chi^2_1$$

  - the **score test** or the **lagrange multiplier test**: this test is based on $\frac{\partial \ln L}{\partial \beta_1}$ evaluated in $\beta_1 = 0$: the teststatistic is also asymptotically χ^2_1 distributed under $H_0 : \beta_1 = 0$

*$-2 \ln(L_0)$ is also called the **null deviance** and $-2 \ln(L_1)$ is called the **model deviance** or **residual deviance**; it is the *deviance* from the perfect model for which $-2 \ln(L) \equiv -2 \ln(1) = 0$

– **Wald test**: under $H_0 : \beta_1 = 0$

$$\frac{\hat{\boldsymbol{\beta}}_1^2}{var(\hat{\boldsymbol{\beta}}_1)} \overset{H_0}{\sim} \chi_1^2$$

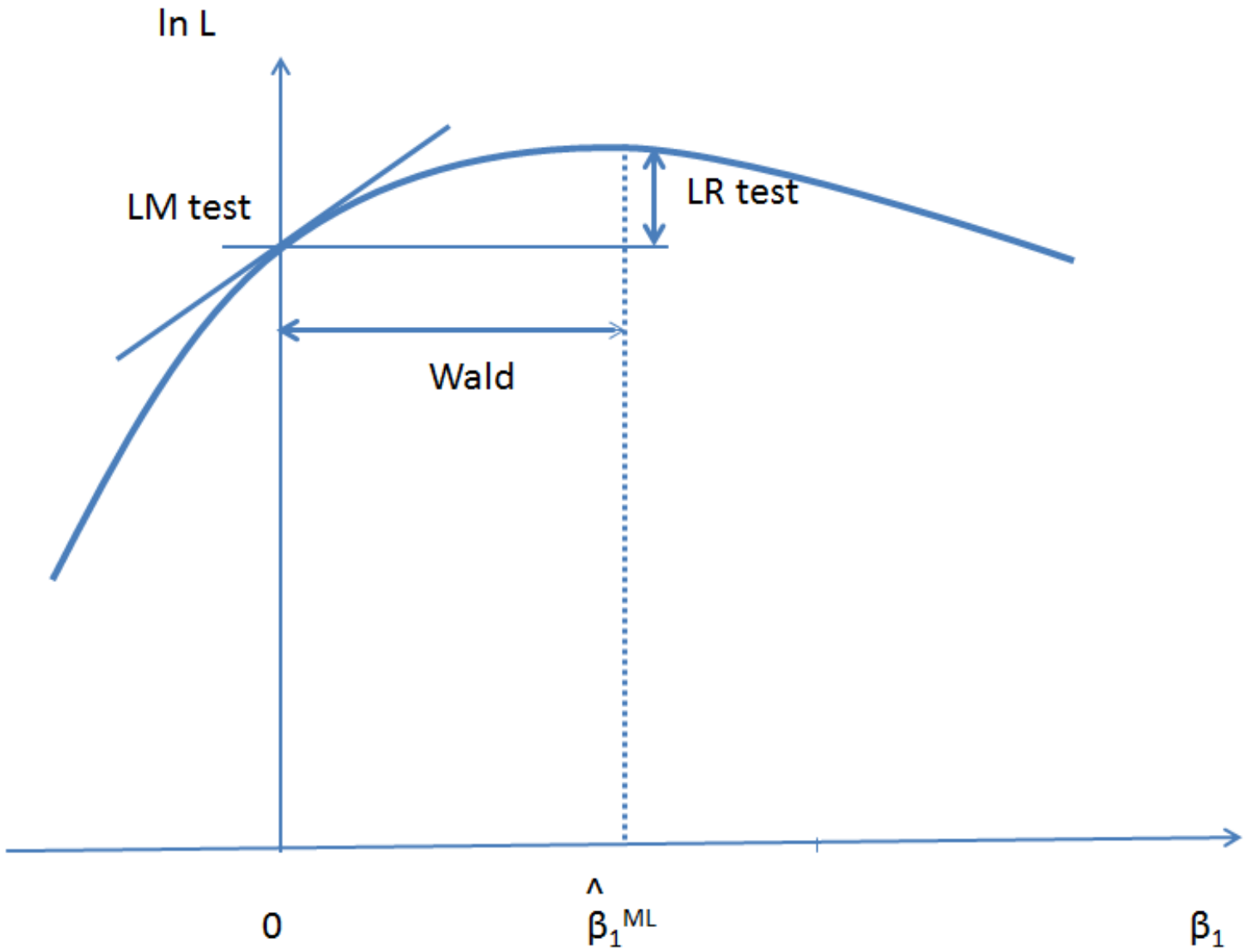
with

$$var(\hat{\boldsymbol{\beta}}_1) = -\mathbf{E}\left(\frac{\partial^2 \ln L}{\partial \beta_1^2}\right)$$

this teststatistic is related to the Mahalanobisdistance between the estimates $\hat{\boldsymbol{\beta}}$ and the null vector $\mathbf{0}$ and checks whether they differ significantly

– these 3 tests are asymptotically equivalent

these tests can be visualized as follows:



– the Akaike Information Criterion (AIC) and the BIC or Schwartz Criterion (BIC or SC) are also based on $-2\ln L$ but with a correction for the degrees of freedom (as does R^2_{adj}):

$$\begin{aligned} AIC &= -2\ln L + 2(\text{nr of parameters}) \\ BIC &= -2\ln L + (\text{nr of parameters})\ln n \end{aligned}$$

with n the number of observations

these criteria can only be used to compare different models fitted on the same data: the smaller the criterion, the better

- **significance of individual parameters**
  - test $H_0 : \beta_1 = 0$ based on $\hat{\beta}_1 \overset{n\to\infty}{\sim} N(\beta_1, \sigma_1^2)$:

$$\frac{\hat{\beta}_1 - \beta_1}{s_1} \sim N(0,1) \text{ and therefore } \frac{\hat{\beta}_1}{s_1} \overset{H_0}{\sim} N(0,1)$$

  - the Wald-test is the squared version of this test:

$$\text{under } H_0 : \beta_1 = 0 \quad \left(\frac{\hat{\beta}_1}{s_1}\right)^2 \sim \chi_1^2$$

example: predict whether a person who took a car insurance will be involved in an accident the coming year or not as a function of the horse power of the car (hp)

RESULTS

```
Number of Observations  21
        Response Profile
Ordered                      Total
  Value     accident     Frequency
      1     acc                 10
      2     noacc               11
Probability modeled is accident='acc'.
```

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	31.065	14.839
BIC	32.109	16.928
-2 Log L	29.065	10.839

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.2259	1	<.0001
Score	13.0380	1	0.0003
Wald	5.3869	1	0.0203

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-20.1158	8.7955	5.2307	0.0222
hp	1	0.1220	0.0526	5.3869	0.0203

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
hp	1.130	1.019	1.252

- the likelihood ratio test, Wald test and score test agree that the model is significantly better than the empty model

- fitted model: $\ln\left(\frac{\widehat{P(accident)}}{P(no\ accident)}\right) = -20,1158 + 0,1220\ hp$

- if hp increases with 1 unit, the odds in favor of an accident will increase with $(1.13 - 1) = 0.13 = 13\%$

- assume $p = 0.25$ and therefore the odds are $1/3$
  $\Rightarrow$ an increase of hp with 1 unit, will increase the odds to $1.13 * 1/3 = 0.377$ which corresponds to a probability of 0.274 which is an increase of 9.4% compared to 0.25

- assume $p = 0.50 \Rightarrow$ an increase of hp with 1 unit, will increase the odds to 1.13 and therefore the probability to 0.53 which is an increase of 6.1% compared to 0.50

RESULTS

Obs	hp	accident	_LEVEL_	predprob
1	153	noacc	acc	0.18955
2	142	noacc	acc	0.05761
3	184	noacc	acc	0.91121
4	105	noacc	acc	0.00067
5	160	noacc	acc	0.35456
6	142	noacc	acc	0.05761
.....				
19	193	acc	acc	0.96852
20	174	acc	acc	0.75188
21	181	acc	acc	0.87681

classification

use the predicted probability of an event to classify an observation as an event or a nonevent based on a *cutoff*-value, for instance: predict an event if $\hat{p} > 0.5$ and a nonevent otherwise

classification table or confusion table:

	predicted	
observed	event	nonevent
event	(tp) true positives	(fn) false negatives
nonevent	(fp) false positives	(tn) true negatives

- **% correct:** % of all observations that is classified correctly:

$$\frac{number\ correct\ events\ and\ correct\ nonevents}{number\ observations} = \frac{tp+tn}{tp+fp+fn+tn}$$

- **true positive rate (TPR):** % of events correctly classified:

$$\frac{number\ correct\ events}{number\ observed\ events} = \frac{tp}{tp+fn} \equiv \textbf{sensitivity, recall, hit rate}$$

- **true negative rate (TNR):** % of nonevents correctly classified:

$$\frac{number\ correct\ nonevents}{number\ observed\ nonevents} = \frac{tn}{fp+tn} \equiv \textbf{specificity}$$

- **false positive rate (FPR):** % of nonevents predicted as events:

$$\frac{number\ incorrect\ events}{number\ observed\ nonevents} = \frac{fp}{fp+tn} \equiv \textbf{1 - specificity = fall out}$$

- **false negative rate (FNR):** % of events predicted as nonevents:

$$\frac{number\ incorrect\ nonevents}{number\ observed\ events} = \frac{fn}{tp+fn}$$

- **concordant pairs** (n_c): all pairs of an event and a nonevent for which the predicted probability that an event will occur is larger for the event than for the nonevent

- **disconcordant pairs** (n_d): all pairs of an event and a nonevent for which the predicted probability that an event will occur is smaller for the event than for the nonevent

- several measures (Somer's D, Gamma, Tau-a, c) are based on these numbers; for instance: $\gamma = \frac{n_c - n_d}{n_c + n_d}$

(remark that some people use slightly different definitions for FPR and FNR which is very confusing)

Association of Predicted Probabilities and Observed Responses

Percent Concordant	94.5	Somers' D	0.900
Percent Discordant	5.5	Gamma	0.908
Percent Tied	0.0	Tau-a	0.471
Pairs	110	c	0.950

Classification Table

Prob Level	Correct Event (tp)	Correct Non-Event (tn)	Incorrect Event (fp)	Incorrect Non-Event (fn)
0.000	10	0	11	0
0.020	10	5	6	0
0.040	10	5	6	0
0.060	10	6	5	0
.....				
0.480	9	10	1	1
0.500	9	10	1	1
0.520	9	10	1	1
.....				
0.980	2	11	0	8
1.000	0	11	0	10

remarks

- prob level: cut-off values
  - $problevel = 0$: all observations classified as events
    $\Rightarrow$ sensitivity $= 100$
  - $problevel = 1$: no observation classified as an event
    $\Rightarrow$ specificity $= 100$

- obs 3 has a high predicted probability but had no accident

Receiver Operating Characteristic (ROC) curve

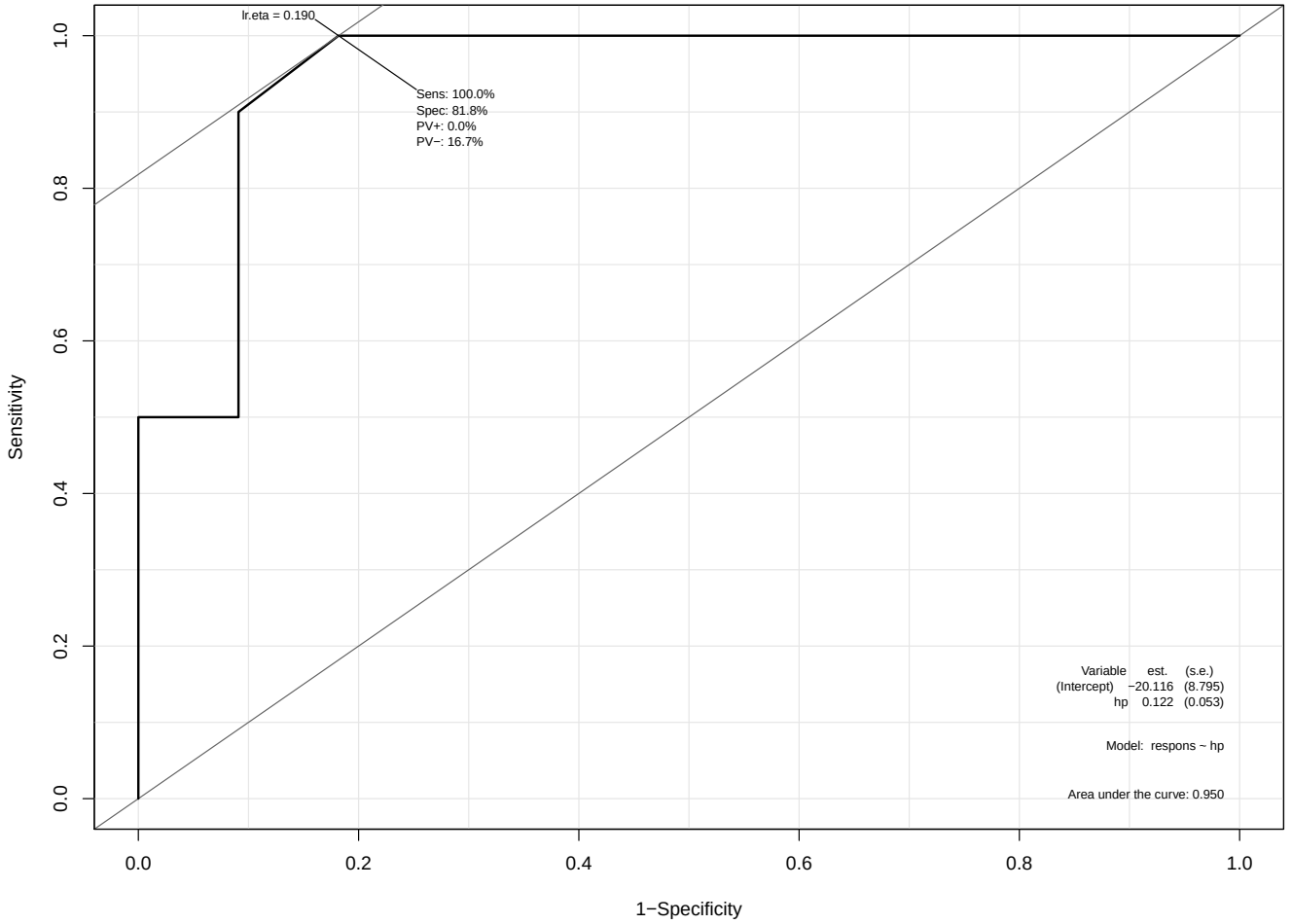
- depicts the true positive rate (or sensitivity) against the false positive rate (or 1 - specificity) for different cutoff values
- perfect classification in (0,1): in this case all events are classified as events and no nonevents are incorrectly classified as events
- an increase in sensitivity (for instance by reducing the cutoff value from 0.5 to 0.1) will almost always induce a decrease in specificity
- use the ROC curve to decide on a proper cutoff value
- the Area Under the Curve (AUC) is equal to the percentage *concordant pairs* plus half of the percentage *tied pairs*

random allocation of observations would result in the diagonal ROC curve:

- if the cutoff value is for instance 25%, random allocation would classify 75% randomly chosen observations as events and the rest as nonevents

- this would result in a sensitiviy of 75% (as 75% of the events would be classified as events) and a specificity of 25% (as 25% of the nonevents are classified as nonevents) and therefore 1-specificity would be equal to 75% too

$\frac{tp}{tp+fn}$



$\frac{fp}{fp+tn}$

the point at the bottom left corner corresponds to the cutoff value of 1, the point at the top right corner with a cutoff value of 0

multiple logistic regression

we still consider a binary response variable but multiple explanatory variables:

$$\pi_i = \frac{e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik})}}{1+e^{(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik})}} = \frac{e^{logit_i}}{1+e^{logit_i}}$$

or $$\ln\frac{\pi_i}{1-\pi_i} = logit_i = \beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}$$

now the likelihoodfunction has to be maximized w.r.t. $\beta_0, \beta_1, \ldots, \beta_k$:

$$L = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ can again be tested with the likelihood ratio test:

$$[-2\ln(L_0)] - [-2\ln(L_1)] \overset{H_0}{\sim} \chi^2_k$$

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ can also be tested with a generalized version of the score test and of the Wald test

- individual parameters can be tested with the Wald test as before

- interpretation of the parameters with the odds ratio: the odds in favor of the event are multiplied by e^{β_i} if x_i increases with 1 unit and the other variables remain constant

example: take into account the experience of the driver (in years)

RESULTS

Number of Observations 21

Response Profile

Ordered Value	accident	Total Frequency
1	acc	10
2	noacc	11

Probability modeled is accident='acc'.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	31.065	16.547
BIC	32.109	19.680
-2 Log L	29.065	10.547

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	18.5180	2	<.0001
Score	13.1632	2	0.0014
Wald	6.0899	2	0.0476

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-17.7259	8.6241	4.2246	0.0398
HP	1	0.1139	0.0492	5.3533	0.0207
EXPERIENCE	1	-0.0570	0.1058	0.2896	0.5904

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
HP	1.121	1.018	1.234
EXPERIENCE	0.945	0.768	1.162

---

based on the AIC and BIC values as well as based on the p-values related to experience, we can conclude that experience does not improve the model significantly

example: consider the following information on 100 women:

- work =1 in case the woman has paid employment, =0 otherwise
- child1 =1 in case the woman has kids aged under 2, =0 otherwise
- child2 =1 in case the woman has kids aged between 2 and 6, =0 otherwise
- husbinc = income of the partner
- educ = years of education
- age (in years)

and predict who has paid employment based on the other variables

Number of Observations 100

Response Profile

Ordered Value	work	Total Frequency
1	1	72
2	0	28

Probability modeled is work=1.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	120.591	103.369
BIC	123.196	119.000
-2 Log L	118.591	91.369

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	27.2219	5	<.0001
Score	26.4798	5	<.0001
Wald	17.2036	5	0.0041

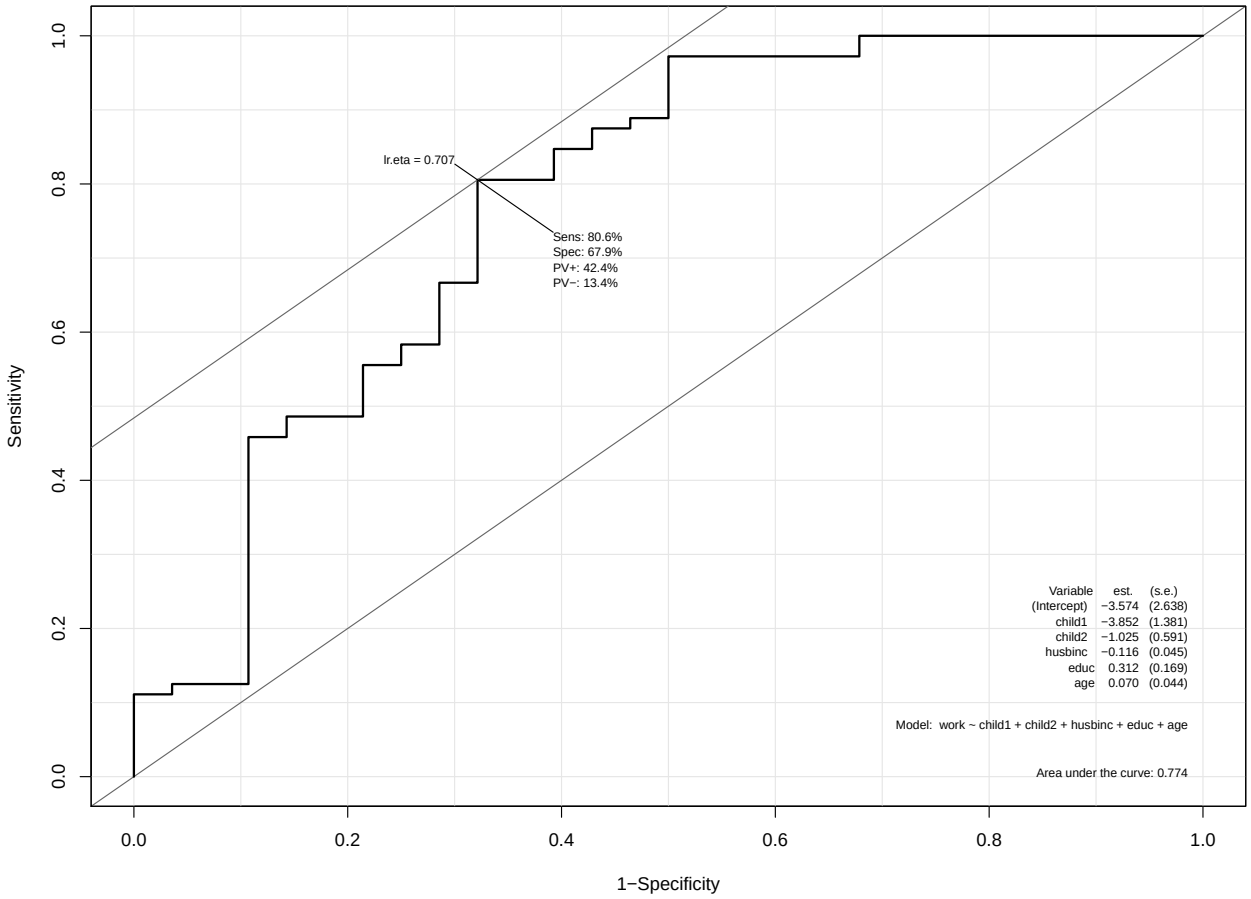
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5736	2.6378	1.8353	0.1755
child1	1	-3.8517	1.3809	7.7804	0.0053
child2	1	-1.0253	0.5907	3.0129	0.0826
husbinc	1	-0.1160	0.0448	6.7084	0.0096
educ	1	0.3117	0.1695	3.3811	0.0659
age	1	0.0699	0.0441	2.5147	0.1128

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
child1	0.021	0.001	0.318
child2	0.359	0.113	1.142
husbinc	0.891	0.816	0.972
educ	1.366	0.980	1.904
age	1.072	0.984	1.169

the corresponding ROC-curve is:



models for more than 2 response values

we consider 3 frequently used models*:

- the cumulative logit model: for ordered response values
- the multinomial logit model: no ordered response values and the classification depends on characteristics of the item to be classified or the person that has to make the choice
- the discrete choice model or conditional logit model: no ordered response values and the classification depends on characteristics of the alternatives from which to choose

*all these models simplify to the previous binary logit if there are only 2 responses

cumulative logit or proportional odds model

examples: predict based on some relevant variables

- whether a family has no, one or more than one car
- whether a person will strongly disagree, disagree, be neutral, agree or strongly agree with some statement
- whether a student will pass in June, will pass in September or will fail in September

assume there are r response values and y_i is the response of the ith observation: the trick is to transform it to $r-1$ binary logit models:

$$
\begin{aligned}
logit_{1i} = \ln \frac{P(y_i \leq 1)}{P(y_i > 1)} = \ln \frac{\pi_{1i}}{\pi_{2i} + \ldots + \pi_{ri}} &= \beta_{10} - (\beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \\
logit_{2i} = \ln \frac{P(y_i \leq 2)}{P(y_i > 2)} = \ln \frac{\pi_{1i} + \pi_{2i}}{\pi_{3i} + \ldots + \pi_{ri}} &= \beta_{20} - (\beta_1 x_{i1} + \ldots + \beta_k x_{ik}), \\
&\vdots \\
logit_{(r-1)i} = \ln \frac{P(y_i < r)}{P(y_i = r)} = \ln \frac{\pi_{1i} + \ldots \pi_{(r-1)i}}{\pi_{ri}} &= \beta_{(r-1)0} - (\beta_1 x_{i1} + \ldots + \beta_k x_{ik})
\end{aligned}
$$

the negative sign in the logits was introduced to get a direct correspondence between the slope and the ranking: a positive coefficient indicates that as the value of the explanatory variable increases, the odds (and therefore also the probability) in favor of a *higher* response value increase

remark the strong assumption that is made: the effect of an increase of a variable is the same for each logit:

$$\frac{\text{the odds in favor of the event } (y \leq j) \text{ at } x_k + 1}{\text{the odds in favor of the event } (y \leq j) \text{ at } x_k} = e^{-\beta_k} \text{ for all } j$$

this assumption is tested with the **score test for the proportional odds assumption** which tests whether the more general model is significantly better or not:

$$\begin{aligned}
\ln \frac{\pi_{1i}}{\pi_{2i} + \ldots + \pi_{ri}} &= \beta_{10} - (\beta_{11} x_{i1} + \ldots + \beta_{1k} x_{ik}), \\
&\vdots \\
\ln \frac{\pi_{1i} + \ldots + \pi_{(r-1)i}}{\pi_{ri}} &= \beta_{(r-1)0} - (\beta_{(r-1)1} x_{i1} + \ldots + \beta_{(r-1)k} x_{ik})
\end{aligned}$$

test: $H_0 : \beta_{11} = \beta_{21} = \ldots = \beta_{(r-1)1}; \ldots; \beta_{1k} = \beta_{2k} = \ldots = \beta_{(r-1)k}$

the logits lead to the following probabilities:

$$\begin{aligned}
\pi_{1i} &= \frac{e^{logit_{1i}}}{1+e^{logit_{1i}}}, \\
\pi_{1i}+\pi_{2i} &= \frac{e^{logit_{2i}}}{1+e^{logit_{2i}}}, \\
&\vdots \\
\pi_{1i}+\pi_{2i}+\ldots\pi_{(r-1)i} &= \frac{e^{logit_{(r-1)i}}}{1+e^{logit_{(r-1)i}}}, \\
\pi_{ri} &= 1-(\pi_{1i}+\pi_{2i}+\ldots\pi_{(r-1)i})
\end{aligned}$$

the parameters are estimated by maximizing the likelihood and the likelihood ratio test can be used to check whether the model is significantly better than the empty model

example: car insurance data with 3 response values: no accident (accident = 1), limited damage (accident =2), large damage (accident = 3)

RESULTS

Number of Observations 21

Response Profile

Ordered Value	accident	Total Frequency
1	1noacc	11
2	2small	5
3	3large	5

Probabilities modeled are cumulated over the lower Ordered Values.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
0.8310	2	0.6600

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	46.927	26.513
BIC	49.017	30.691
-2 Log L	42.927	18.513

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.4149	2	<.0001
Score	14.2326	2	0.0008
Wald	8.1842	2	0.0167

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 1noacc	1	19.2871	7.6801	6.3066	0.0120
Intercept 2small	1	22.5376	8.5270	6.9859	0.0082
hp	1	0.1285	0.0466	7.6083	0.0058
experience	1	-0.1007	0.0931	1.1696	0.2795

- fitted model:

$$\ln \frac{P(\widehat{no\ accident})}{P(accident)} = 19.2871 - (0,1285\ hp - 0,1007\ experience),$$

$$\ln \frac{P(no\ or\ \widehat{small}\ accident)}{P(large\ accident)} = 22.5376 - (0,1285\ hp - 0,1007\ experience).$$

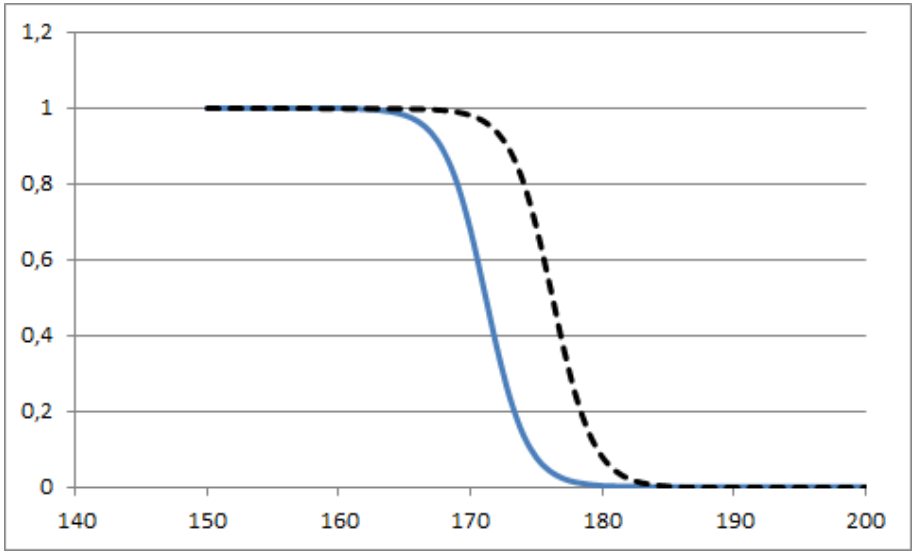
- p-value of the *score test for the proportional odds assumption* is large $(0.66) \Rightarrow$ the more general model is not significantly better

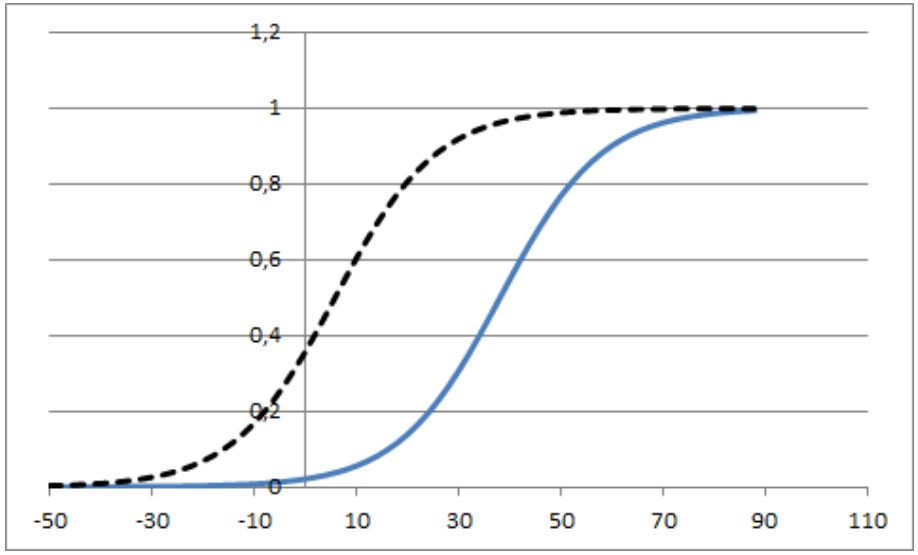
- p-values of the score test and the likelihood ratio test are small $\Rightarrow$ the model is significantly better than the empty model

RESULTS

Obs	experience	hp	accident	_LEVEL_	p
1	23	153	1noacc	1noacc	0.87534
2	23	153	1noacc	2small	0.99451
3	18	142	1noacc	1noacc	0.94577
4	18	142	1noacc	2small	0.99778
5	16	184	1noacc	1noacc	0.06079
6	16	184	1noacc	2small	0.62549
7	25	105	1noacc	1noacc	0.99976
8	25	105	1noacc	2small	0.99999
9	32	160	1noacc	1noacc	0.87609
10	32	160	1noacc	2small	0.99455
....					
39	14	204	3large	1noacc	0.00404
40	14	204	3large	2small	0.09471
41	18	193	3large	1noacc	0.02431
42	18	193	3large	2small	0.39130

the model is visualized as a function of hp in case experience $= 20$ (left) and as a function of experience in case hp $= 180$ (right):





the solid line corresponds to $P(no\ accident)$, the dashed line to $P(at\ most\ a\ small\ accident)$

multiple outcomes without inherent ordering

examples:

- predict whether a person commutes to work by bus, car or train
- predict whether a person will vote NVA, CD&V, Open Vld, sp.a, Groen, VB...

let there be r possible outcomes, then r logits are defined as:

$$
\begin{aligned}
logit_{1i} &= \beta_{10} + \beta_{11}x_{i11} + \ldots + \beta_{1k}x_{i1k} \\
&\vdots \\
logit_{ji} &= \beta_{j0} + \beta_{j1}x_{ij1} + \ldots + \beta_{jk}x_{ijk} \\
&\vdots \\
logit_{ri} &= \beta_{r0} + \beta_{r1}x_{ir1} + \ldots + \beta_{rk}x_{irk}
\end{aligned}
$$

not all parameters of all logits will be estimable in all cases!

if

- individual* i gets utility U_{ji} from choosing j

$$U_{ji} = logit_{ji} + \epsilon_{ji}$$

- the error terms ϵ_{ji} have independent *extreme value* distributions with mean 0 and the same variance (see Duration Analysis)

- and if we assume that individual i maximizes his utility

then it can be shown that $\pi_{ji} = P(y_i = j)$ is given by

$$P(y_i = j) = \frac{e^{logit_{ji}}}{e^{logit_{1i}} + \ldots + e^{logit_{ji}} + \ldots + e^{logit_{ri}}}$$

*this *individual* is often a person but the model can also be used if it is a product or concept or something else

similarly as for the binary logit models, the parameters β_{jl} are estimated by maximizing the likelihood:

$$L = \prod_{i=1}^{n} \pi_{1i}^{y_{1i}} \, \pi_{2i}^{y_{2i}} \, \ldots \, \pi_{ri}^{y_{ri}} \quad \text{with } y_{ji} = 1 \text{ if } y_i = j \text{ and 0 otherwise}$$

we start by looking at 2 special cases:

- the explanatory variables only depend on i, not on j
  (so x represents characteristics of the individual i)

- the explanatory variables only depend on j, not on i
  (so x represents characteristics of the outcome j)

the multinomial logit model

examples: (here x_{ijl} is independent of outcome j)

- predict whether a person commutes to work by bus, car or train based on the characteristics of the person (income, age, gender...)
- predict a person will vote NVA, CD&V, OVld, SP.a, ... based on the characteristics of the person (income, age, education,...)

the r logits become:

$$\begin{aligned} logit_{1i} &= \beta_{10} + \beta_{11}x_{i1} + \ldots + \beta_{1k}x_{ik} \\ &\ldots \\ logit_{ri} &= \beta_{r0} + \beta_{r1}x_{i1} + \ldots + \beta_{rk}x_{ik} \end{aligned}$$

and

$$\begin{aligned} P(y_i = j) &= \frac{e^{logit_{ji}}}{e^{logit_{1i}} + \ldots + e^{logit_{ji}} + \ldots + e^{logit_{ri}}} \\ &= \frac{e^{\beta_{j0}+\beta_{j1}x_{i1}+\ldots+\beta_{jk}x_{ik}}}{e^{\beta_{10}+\beta_{11}x_{i1}+\ldots+\beta_{1k}x_{ik}} + \ldots + e^{\beta_{r0}+\beta_{r1}x_{i1}+\ldots+\beta_{rk}x_{ik}}} \\ &= \frac{e^{(\beta_{j0}-\beta_{r0})+(\beta_{j1}-\beta_{r1})x_{i1}+\ldots+(\beta_{jk}-\beta_{rk})x_{ik}}}{e^{(\beta_{10}-\beta_{r0})+(\beta_{11}-\beta_{r1})x_{i1}+\ldots+(\beta_{1k}-\beta_{rk})x_{ik}} + \ldots + 1} \end{aligned}$$

so only the difference between the parameters is estimable

to solve the nonuniqueness, we put $\beta_{r0} = \beta_{r1} = \ldots = \beta_{rk} = 0$ such that $logit_{ri} = 0$ (which category is chosen as the reference category is not important: the tests and resulting probabilities do not depend on this choice) which yields the **multinomial model** or **generalized logit** or **baseline category logit model**: for $j = 1, \ldots, r-1$:

$$P(y_i = j) = \frac{e^{\beta_{j0}+\ldots+\beta_{jk}x_{ik}}}{e^{\beta_{10}+\ldots+\beta_{1k}x_{ik}} + \ldots + e^{\beta_{j0}+\ldots+\beta_{jk}x_{ik}} + \ldots + 1}$$

remark the **alternative specific parameters** β_{jk} and **alternative specific constants** (ASC) β_{j0} (the latter representing the expected utility of each outcome if the variables are zero)

remark that we get

$$
\begin{aligned}
logit_{1i} &= \ln \frac{P(y_i = 1)}{P(y_i = r)} = \beta_{10} + \beta_{11} x_{i1} + \ldots + \beta_{1k} x_{ik}, \\
logit_{2i} &= \ln \frac{P(y_i = 2)}{P(y_i = r)} = \beta_{20} + \beta_{21} x_{i1} + \ldots + \beta_{2k} x_{ik}, \\
logit_{(r-1)i} &= \ln \frac{P(y_i = (r-1))}{P(y_i = r)} = \beta_{(r-1)0} + \ldots + \beta_{(r-1)k} x_{ik} \\
&\vdots \\
logit_{ri} &= \ln \frac{P(y_i = r)}{P(y_i = r)} = 0
\end{aligned}
$$

so these logits do not have the interpretation of *log(odds)* but of *log(***relative risk**); be careful in interpreting the parameters!

example: predict travel mode (bus, car, train) based on age and gender (gender $= 1$ for males, gender $= 0$ for females)

you need one observation for each choice made, for instance if everyone had the choice between bus, car and train, the data might look like

age	gender	choice
25	0	car
28	0	car
41	1	train
45	0	bus
...		

Model Information

Data Set	WORK.TRAVEL
Response Variable	chosen
Number of Response Levels	3
Model	generalized logit
Optimization Technique	Newton-Raphson

Number of Observations Read	107
Number of Observations Used	107

Response Profile

Ordered Value	chosen	Total Frequency
1	bus	39
2	car	28
3	train	40

Logits modeled use chosen='train' as the reference category.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	236.514	196.234
SC	241.859	212.271
-2 Log L	232.514	184.234

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	48.2794	4	<.0001
Score	44.9772	4	<.0001
Wald	37.9691	4	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age	2	13.7976	0.0010
genderF0	2	24.5318	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	chosen	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	bus	1	-1.5585	1.0618	2.1544	0.1422
Intercept	car	1	1.8818	1.0509	3.2064	0.0734
age	bus	1	0.0771	0.0272	8.0383	0.0046
age	car	1	-0.0231	0.0296	0.6073	0.4358
genderF0	bus	1	-2.5810	0.5996	18.5286	<.0001
genderF0	car	1	-2.7151	0.6391	18.0505	<.0001

Relative Risk Ratios

Effect	chosen	Point Estimate	95% Wald Confidence Limits	
age	bus	1.080	1.024	1.139
age	car	0.977	0.922	1.036
genderF0	bus	0.076	0.023	0.245
genderF0	car	0.066	0.019	0.232

Obs	age	gender F0	chosen	_LEVEL_	phat
1	25	0	car	bus	0.23591
2	25	0	car	car	0.60102
3	25	0	car	train	0.16306
4	28	0	car	bus	0.29115
5	28	0	car	car	0.54917
6	28	0	car	train	0.15968
7	41	1	train	bus	0.24348
8	41	1	train	car	0.10916
9	41	1	train	train	0.64736
.	.	.	.	.	.

$$\frac{\text{RR(bus vs train)} \mid_{age+1}}{\text{RR(bus vs train)} \mid_{age}} = \frac{[\text{P(bus)/P(train)}] \mid_{age+1}}{[\text{P(bus)/P(train)}] \mid_{age}} = 1.08$$

$$\begin{aligned} Logit_{bus} &= -1.5585 + 0.0771 * age - 2.5810 * genderF0 \\ Logit_{car} &= 1.8818 - 0.0231 * age - 2.7151 * genderF0 \end{aligned}$$

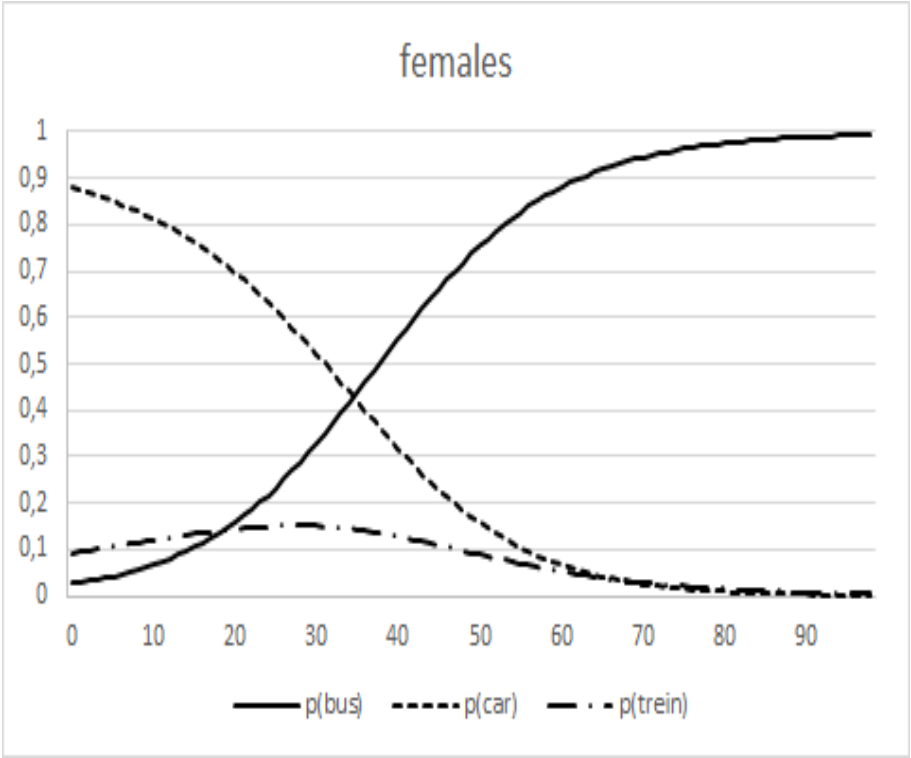
for the first observation age $= 25$ and gender $= 0$ and therefore:

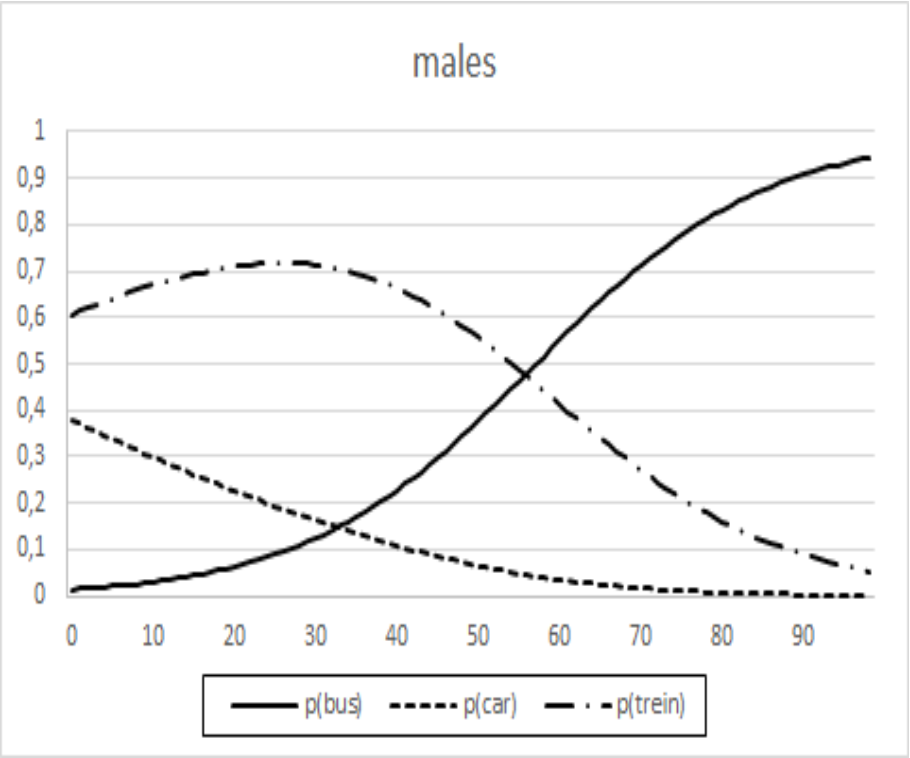
$$logit_{bus} = 0.3690 \quad \& \quad logit_{car} = 1.3043$$

$$\begin{aligned} P(bus) &= \frac{e^{logit_{bus}}}{e^{logit_{bus}} + e^{logit_{car}} + 1} = 0.23599 \\ P(car) &= \frac{e^{logit_{car}}}{e^{logit_{bus}} + e^{logit_{car}} + 1} = 0.60103 \\ P(train) &= \frac{1}{e^{logit_{bus}} + e^{logit_{car}} + 1} = 1 - P(bus) - P(car) = 0.16310 \end{aligned}$$

for the third observation age $= 41$ and gender $= 1$ we get similarly:

$$P(bus) = 0.24339 \quad P(car) = 0.10914 \quad P(train) = 0.64747$$


females
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0 10 20 30 40 50 60 70 80 90
p(bus)
p(car)
p(trein)


males
1
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
0 10 20 30 40 50 60 70 80 90
p(bus)
p(car)
p(trein)

discrete choice or conditional logit model

examples: (here x_{ijl} is independent of individual i)

- predict whether a person travels to work by bus, car or train based on the characteristics of the travel mode (travel time, costs, comfort, ...)

- predict which product a person will buy based on the characteristics of the product (price, size, color, ...)

now the logits become:

$$\begin{aligned} logit_1 &= \beta_{10} + \beta_{11}x_{11} + \ldots + \beta_{1k}x_{1k} \\ &\vdots \\ logit_r &= \beta_{r0} + \beta_{r1}x_{r1} + \ldots + \beta_{rk}x_{rk} \end{aligned}$$

McFadden suggested to use the following simplified logits:

$$\begin{aligned} logit_1 &= \beta_1 x_{11} + \ldots + \beta_k x_{1k} \\ &\vdots \\ logit_r &= \beta_1 x_{r1} + \ldots + \beta_k x_{rk} \end{aligned}$$

as the parameters β_l now do not depend on j they are called **generic parameters** and interpreted as part-worths, the values that people attach to the different characteristics

the probabilities (that are now independent of i!) become:

$$P(y=j) = \frac{e^{\beta_1 x_{j1} + \ldots + \beta_k x_{jk}}}{e^{\beta_1 x_{11} + \ldots + \beta_k x_{1k}} + \ldots + e^{\beta_1 x_{r1} + \ldots + \beta_k x_{rk}}}$$

if there are different choice sets to choose from, the characteristics x_{ij}, the logits $logit_j$ and the probabilities π_j need an extra index s denoting the choice set: x_{ijs}, $logit_{js}$ and π_{js}

remark that a generic intercept cannot be estimated as it drops out of the probabilities

example: predicting travel mode (bus, car, train) based on the corresponding cost and time

subject	mode	travtime	travcost	chosen
1	bus	1.0	8.9	0
1	car	2.4	4.3	1
1	train	0.4	8.3	0
2	bus	23.0	3.0	0
2	car	5.5	4.0	1
2	train	21.5	2.0	0
3	bus	7.1	8.6	0
3	car	8.5	5.9	0
3	train	8.0	4.8	1
...				

RESULTS

Dependent Variable	chosen
Number of Observations	107
Log Likelihood	-89.29792
Number of Iterations	6
AIC	182.59585
BIC	187.94151

Discrete Response Profile

Index	CHOICE	Frequency	Percent
0	1	39	36.45
1	2	28	26.17
2	3	40	37.38

Conditional Logit Estimates
Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
travtime	1	-0.3599	0.0973	-3.70	0.0002
travcost	1	-0.3172	0.0640	-4.96	<.0001

the estimated model is therefore

$$logit = -0.3599\ travtime - 0.3172\ travcost$$

and for the first choice set we have:

$$\begin{aligned} Logit(bus) &= -0.3599 * 1.0 - 0.3172 * 8.9 = -3.18298 \\ Logit(car) &= -0.3599 * 2.4 - 0.3172 * 4.3 = -2.22772 \\ Logit(train) &= -0.3599 * 0.4 - 0.3172 * 8.3 = -2.77672 \end{aligned}$$

$$P(bus) = \frac{e^{-3.18298}}{e^{-2.22772} + e^{-3.18298} + e^{-2.77672}} = 0.19606$$

$$P(car) = \frac{e^{-2.22772}}{e^{-2.22772} + e^{-3.18298} + e^{-2.776725}} = 0.50962$$

$$P(train) = \frac{e^{-2.77672}}{e^{-2.22772} + e^{-3.18298} + e^{-2.77672}} = 0.29432$$

RESULTS

Obs	phat	subject	mode	travtime	travcost	chosen
1	0.19605	1	bus	1.0	8.9	0
2	0.50963	1	car	2.4	4.3	1
3	0.29432	1	train	0.4	8.3	0
4	0.00250	2	bus	23.0	3.0	0
5	0.99159	2	car	5.5	4.0	1
6	0.00590	2	train	21.5	2.0	0
7	0.20673	3	bus	7.1	8.6	0
8	0.29413	3	car	8.5	5.9	0
9	0.49915	3	train	8.0	4.8	1
.	.	.	.	.	.	.

we only consider models where the partworths are assumed to be the same for all persons; in a **mixed logit model** or **random parameter logit model** one estimates the distribution of the β_i in the population

what if **respondent specific** + **alternative specific** information?

including respondent specific information with generic parameters is impossible at these terms drop out of the probabilities

get alternative specific coefficients from a model with generic parameters by creating dummy variables D_j for each alternative j (except for the reference level) and using interactions of these dummies and the individual specific information x_i

for instance (with train the reference level):

subject	mode	travtime	travcost	Dcar	Dbus	agecar	agebus	chosen
1	bus	1.0	8.9	0	1	0	25	0
1	car	2.4	4.3	1	0	25	0	1
1	train	0.4	8.3	0	0	0	0	0
2	bus	23.0	3.0	0	1	0	28	0
2	car	5.5	4.0	1	0	28	0	1
2	train	21.5	2.0	0	0	0	0	0
3	bus	7.1	8.6	0	1	0	41	0
3	car	8.5	5.9	1	0	41	0	0
3	train	8.0	4.8	0	0	0	0	1

. . .

RESULTS

Dependent Variable	chosen
Number of Observations	107
Log Likelihood	-79.67595
Number of Iterations	6
AIC	171.35190
BIC	187.38888

Discrete Response Profile

Index	CHOICE	Frequency	Percent
0	1	39	36.45
1	2	28	26.17
2	3	40	37.38

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Dbus	1	-2.3243	1.1208	-2.07	0.0381
Dcar	1	1.3644	1.0518	1.30	0.1945
travtime	1	-0.3031	0.0999	-3.04	0.0024
travcost	1	-0.3332	0.0636	-5.24	<.0001
agebus	1	0.0591	0.0272	2.17	0.0298
agecar	1	-0.0546	0.0292	-1.87	0.0617

$$
\begin{aligned}
Logit &= -2.3243 * Dbus + 1.3644 * Dcar \\
&\quad -0.3031 * travtime - 0.3332 * travcost \\
&\quad +0.0591 * agebus - 0.0546 * agecar
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
Logit(bus) &= -2.3243 - 0.3031 * travtime - 0.3332 * travcost + 0.0591 * age \\
Logit(car) &= 1.3644 - 0.3031 * travtime - 0.3332 * travcost - 0.0546 * age \\
Logit(train) &= -0.3031 * travtime - 0.3332 * travcost
\end{aligned}
$$

remark that the intercepts in the 3 logits are often interpreted as a measure of *comfort*: the results show that the bus is considered as least comfortable, the train (with intercept 0) as having medium comfort and the car as having most comfort

for the first choice set, we get

$$\begin{aligned} Logit(bus) &= -2.3243 - 0.3031 * 1.0 - 0.3332 * 8.9 + 0.0591 * 25 = -4.1154 \\ Logit(car) &= 1.3644 - 0.3031 * 2.4 - 0.3332 * 4.3 - 0.0546 * 25 = -2.1608 \\ Logit(train) &= -0.3031 * 0.4 - 0.3332 * 8.3 = -2.8868 \end{aligned}$$

from which the probabilities can be computed:

$$
\begin{aligned}
P(bus) &= e^{-4.1154}/[e^{-2.1608} + e^{-4.1154} + e^{-2.8868}] = 0.087128 \\
P(car) &= e^{-2.1608}/[e^{-2.1608} + e^{-4.1154} + e^{-2.8868}] = 0.615209 \\
P(train) &= e^{-2.8868}/[e^{-2.1608} + e^{-4.1154} + e^{-2.8868}] = 0.297663
\end{aligned}
$$

RESULTS

Obs	phat	subject	mode	travtime	travcost	Dcar	Dbus	agecar	agebus	chosen
1	0.08725	1	bus	1.0	8.9	0	1	0	25	0
2	0.61496	1	car	2.4	4.3	1	0	25	0	1
3	0.29779	1	train	0.4	8.3	0	0	0	0	0
4	0.00410	2	bus	23.0	3.0	0	1	0	28	0
5	0.97831	2	car	5.5	4.0	1	0	28	0	1
6	0.01759	2	train	21.5	2.0	0	0	0	0	0
7	0.24703	3	bus	7.1	8.6	0	1	0	41	0
8	0.14972	3	car	8.5	5.9	1	0	41	0	0
9	0.60325	3	train	8.0	4.8	0	0	0	0	1

drawback: the **independence of irrelevant alternatives** (IIA):

as

$$P(y_i = j) = \frac{e^{logit_{ji}}}{e^{logit_{1i}} + \ldots + e^{logit_{ji}} + \ldots + e^{logit_{ri}}}$$

the ratio $\frac{P(y_i=j)}{P(y_i=k)}$ for any 2 responses j and k is independent of all other alternatives in the choice set which is often unrealistic

consider the *blue and red bus example*:

- assume a choice set with a blue bus and a train with $P(\text{train}) = P(\text{blue bus}) = 50\%$
- introduce a red bus as an extra travel mode
- as $P(\text{blue bus})/P(\text{train}) = 1$ and it can be expected that $P(\text{blue bus}) = P(\text{red bus})$ and the total probability has to be 1, we get $P(\text{train}) = P(red\ bus) = P(\text{blue bus}) = 1/3$
- this is not very realistic (the model is wrong because the error terms of the utilities are not independent; *use a nested logit model* $>$ not dealt with)

3b. DURATION ANALYSIS

DURATION ANALYSIS

duration analysis models the time T until a specified event occurs; this event can be death (then it is called survival analysis), machine failure, bankruptcy, employment, divorce, graduation, job change, change of provider, length of stay ... (most of the terminology is however related to *death* events)

duration analysis can be based on only the times of events but it is common to estimate causal or predictive models in which the risk of an event depends on covariates (= regressors or explanatory variables)

censoring

typical about survival data are the **censored data**; the information one has about subjects at the time of analysis is one of the following:

- died at time t (more general: the event has occurred at time t)
- still alive at the end of the study ($=$ *right* censored), so the only thing known is that t is larger than some value
- withdrawn at time t ($=$ *right* censored) (for instance died while in a study on divorce, or moved to an unknown address)
- (less frequent) *left* censored, the only thing known is that t is smaller than some value (for instance because the event happened before you observed it but you don't know the exact time)

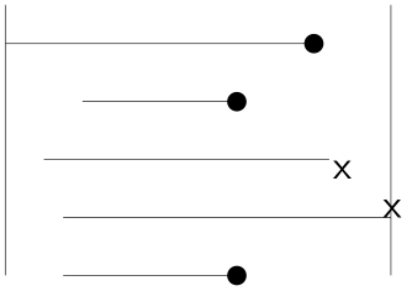
- (less frequent) *interval* censored, the only thing known is that t is in some interval (for instance because observations are collected at discrete time points)

- we only look at random type I censoring: the study is designed to end after C years, but censored subjects do not all have the same censoring time as they may have entered after the study began

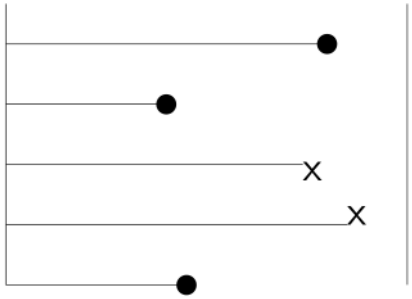
- in type II censoring, the study ends when a specific number of events have occurred

- this type of analysis is also called **failure time analysis, reliabiliby analysis, event history analysis, survival analysis, time-to-event analysis...**

there is a difference between *calendar time* and *time in the study*: it is common for subjects to enter the study continuously throughout the length of the study, as is the case in the left picture (a circle means an event ($\delta = 1$), a cross means censoring ($\delta = 0$))



calendar time



study time

T_i	δ_i
80	1
40	1
74	0
85	0
45	1

study time deals with the time that subjects were part of the study: they all start at study time zero and ending points corresponding to an event or to censoring

regardless of the type of censoring, we must assume that it is non-informative about the event; that is, the censoring is caused by something else than the impending event (counter examples: in a study on divorce, people with marital problems may drop out relatively more or, in a study about unemployment duration, people move abroad because that cannot find a job in their home country)

the usual regression models cannot be used for modelling T as

- time till death is always positive
- time till death has a skewed distribution and therefore a normal distribution is not appropriate

describing survival functions

- $F(t) = P(T \leq t)$ is the cumulative distribution function of T
- $S(t) = P(T > t) = 1 - F(t)$ is the **survival function**, gives the probability that a subject will survive past time t
  - it is non-increasing
  - at time $t = 0$, $S(t) = 1$, so surviving past time 0 is 1
  - $S(\infty) = 0$ in most cases; if $S(\infty) > 0$: *cure* model
- $f(t) = F'(t) = -S'(t)$ the probability that the event occurs between t and $t + \Delta t$ for a very small Δt*

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$$

*the proportion of all people at risk at t=0 that have an event between t and $t + \Delta t$

- $h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \ln S(t)$
  - is the **hazard function or rate** or **risk function** or **failure rate**
  - is the probability the event occurs between t and $t + \Delta t$ for a very small Δt, given that the event has not yet occurred at time t^*
  - it is also called **the instantaneous failure rate**:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T \leq t + \Delta t \mid T \geq t)}{\Delta t}$$

  - remark that

$$S(t) = \exp\left\{ - \int_0^t h(s) ds \right\}$$

*the proportion of all people still at risk at t that have an event between t and $t + \Delta t$

- $mrl(t)$ is the **mean residual lifetime** at t or **the mean time until the event**, given that the event has not yet occurred at t:

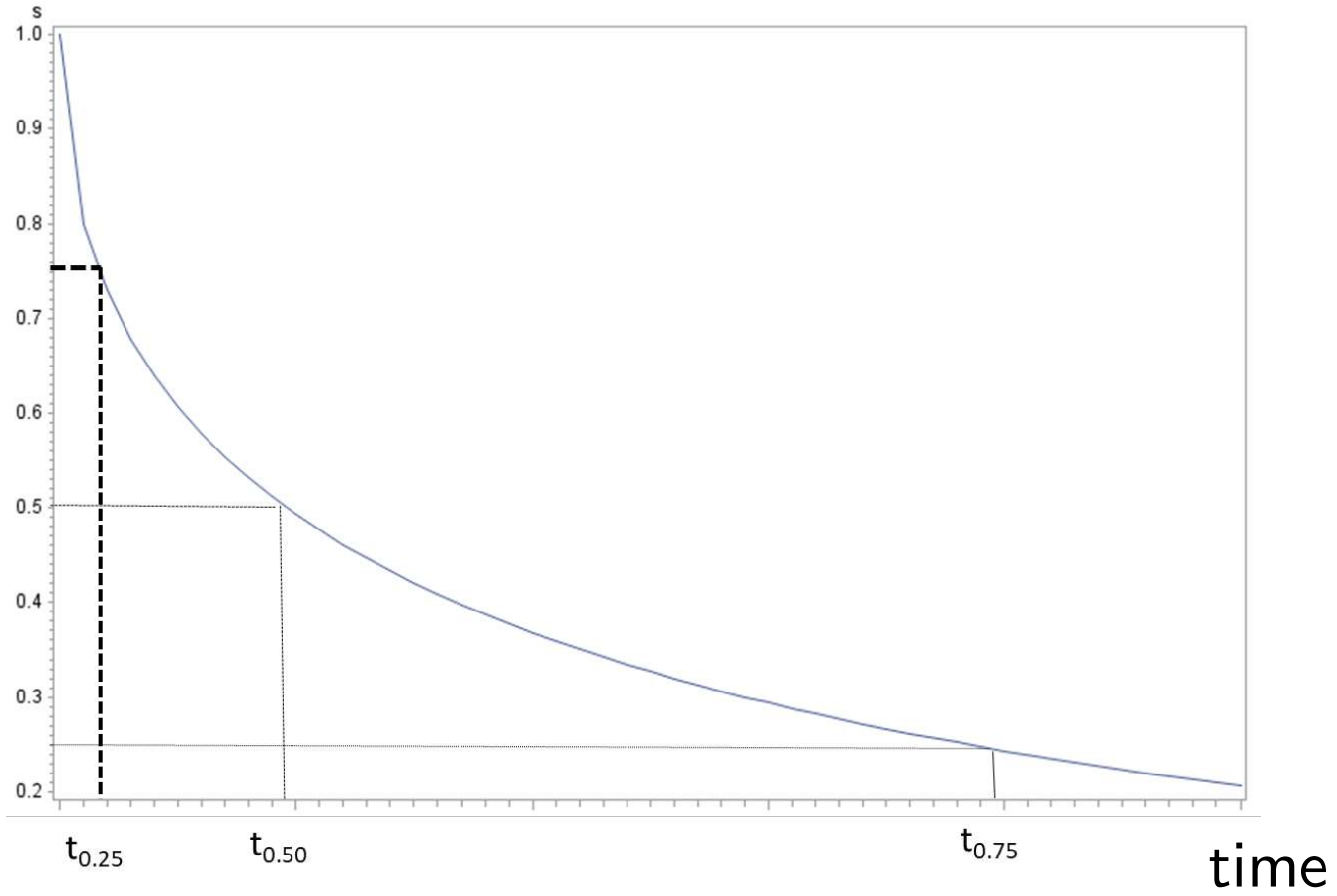
$$mrl(t) = E(T - t \mid T > t) = \frac{\int_t^\infty S(x)dx}{S(t)}$$

$mrl(0)$ is the mean time until the event:

$$mrl(0) = E(T) = \int_0^\infty S(x)dx$$

- the p^{th} percentile t_p is the smallest t for which $S(t_p) \leq 1 - p$ or $F(t_p) \geq p$ (if p is 25, 50 or 75, it is called a quartile)

- the 50^{th} percentile is the median time until the event occurs


S
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
$t_{0.25}$
$t_{0.50}$
$t_{0.75}$
time

nonparametric description of survival functions

if no theoretical distribution is assumed; the Kaplan-Meier Product-Limit estimator is most often used to visualize the survival function: let $t(1) < t(2) < \ldots t(m)$ the ordered time points at which something happens (an event or censoring) then

$$\hat{S}(t) = \prod_{i:t(i)<t} \left(1 - \frac{d_i}{n_i}\right)$$

with

- n_i the number at risk just before time $t(i)$
- d_i the number of events at $t(i)$

suppose we have the following data:

t	nr at risk	death	δ	$S(t)$	mrl(0)
0	5			$5/5 = 1.0$	
40	5	yes	1	$1.0 \times (1 - 1/5) = 0.8$	(40-0) * 1 = 40
45	4	yes	1	$0.8 \times (1 - 1/4) = 0.6$	+(45-40)*0.80 = 4
74	3	no	0		
80	2	yes	1	$0.6 \times (1 - 1/2) = 0.3$	+(80-45)*0.6 =21
85	1	no	0		
					65

when the last time point corresponds to a censored observation, $S(t)$ is undetermined after that time (several approximations exist)

the table at the right shows the computation of the mean residual lifetime at 0

on toy example:

RESULTS

Product-Limit Survival Estimates

t	Survival	Failure	Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	5
40.0000	0.8000	0.2000	0.1789	1	4
45.0000	0.6000	0.4000	0.2191	2	3
74.0000*	.	.	.	2	2
80.0000	0.3000	0.7000	0.2387	3	1
85.0000*	.	.	.	3	0

NOTE: The marked survival times are censored observations.

Summary Statistics for Time Variable t

Quartile Estimates

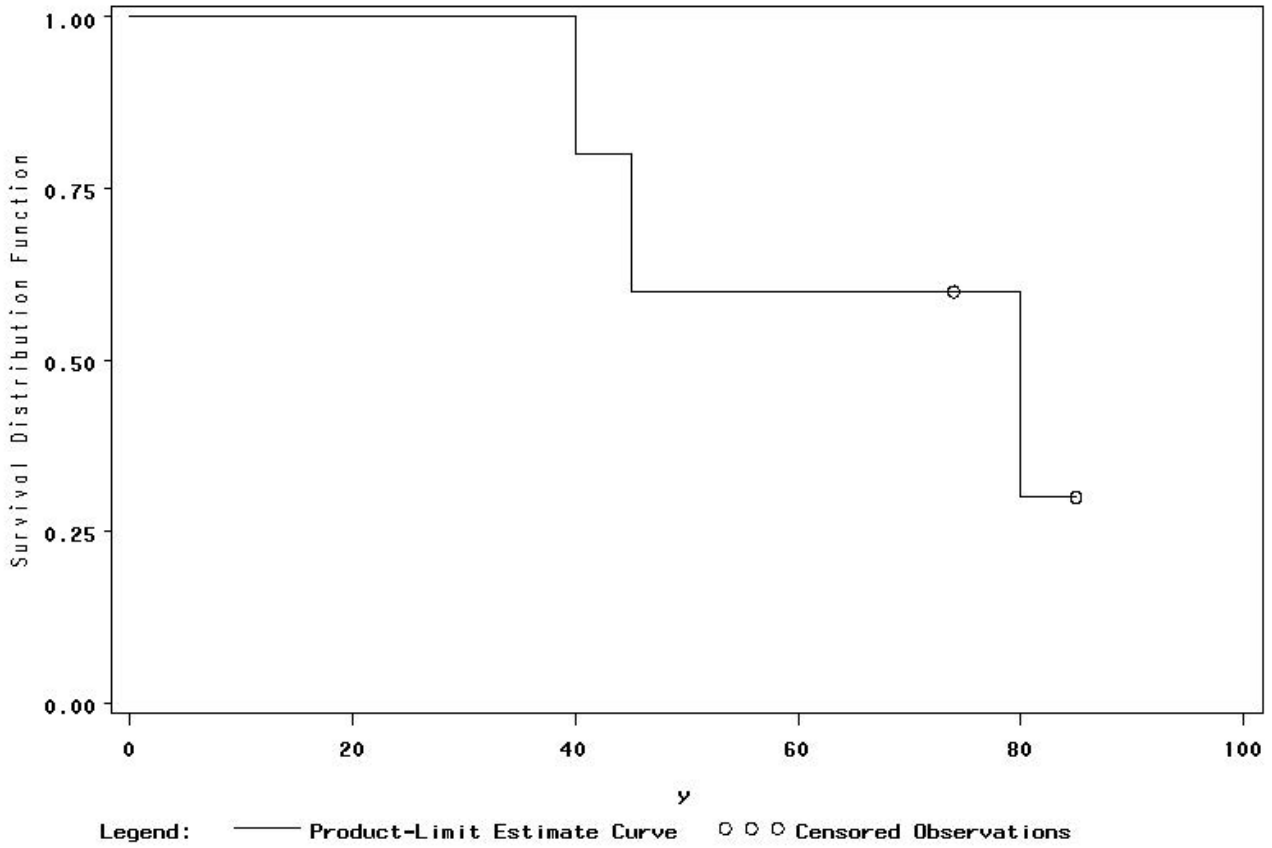
Percent	Point Estimate	95% Confidence Interval [Lower	Upper)
75	.	45.0000	.
50	80.0000	40.0000	.
25	45.0000	40.0000	.

Mean	Standard Error
65.0000	10.0995

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
5	3	2	40.00



example: T is the time (in days) until employment

- people can attend a re-orientation course (course $= 1$ if they did)
- also the number of previous jobs (prevj) is given (used later)
- the first 10 observations are

Obs	ID	age	time	censor	prevj	course
1	1	39	188	1	3	0
2	2	33	26	1	3	0
3	3	33	207	1	2	0
4	4	32	144	1	3	1
5	5	24	551	0	2	0
6	6	30	32	1	1	0
7	7	39	459	1	3	0
8	8	27	22	1	3	0
9	9	40	210	1	2	0
10	10	36	184	1	2	0

the Kaplan-Meier estimation is done separately for those that did not take the course and for those that did take it (only part of the output is shown)

RESULTS

Stratum 1: course = 0

Product-Limit Survival Estimates

time	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00	1.0000	0	0	0	308
2.00	0.9968	0.00325	0.00324	1	307
3.00	0.9935	0.00649	0.00458	2	306
4.00	.	.	.	3	305
4.00	0.9870	0.0130	0.00645	4	304
5.00	.	.	.	5	303
5.00	.	.	.	6	302
5.00	0.9773	0.0227	0.00849	7	301
6.00	0.9740	0.0260	0.00906	8	300

.....

```
  630.00*          .           .           .      242          8
  633.00*          .           .           .      242          7
  634.00*          .           .           .      242          6
  655.00*          .           .           .      242          5
  658.00*          .           .           .      242          4
  659.00      0.1582      0.8418      0.0490      243          3
  708.00*          .           .           .      243          2
  720.00*          .           .           .      243          1
 1172.00*      0.1582      0.8418           .      243          0
    NOTE: The marked survival times are censored observations.

          Summary Statistics for Time Variable time
                      Quartile Estimates
                 Point          95% Confidence Interval
 Percent     Estimate     Transform       [Lower      Upper)
      75       439.00     LOGLOG          355.00          .
      50       192.00     LOGLOG          175.00      224.00
      25        85.50     LOGLOG           68.00      113.00

                        Mean    Standard Error
                      275.21             13.12
NOTE: The mean survival time and its standard error were underestimated because
      the largest observation was censored and the estimation was restricted to
      the largest event time.
```

so 85.5 if the smallest time such that the probability of finding a job earlier is larger than 25% or, assuming a lot of data, approximately 25% of the unemployed people will find a job within 85.5 days

The LIFETEST Procedure
Stratum 2: course = 1

Product-Limit Survival Estimates

time	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00	1.0000	0	0	0	320
3.00	.	.	.	1	319
3.00	0.9938	0.00625	0.00441	2	318
4.00	.	.	.	3	317
4.00	0.9875	0.0125	0.00621	4	316
5.00	0.9844	0.0156	0.00693	5	315
6.00	.	.	.	6	314
.....					
624.00*	.	.	.	265	8
625.00*	.	.	.	265	7
654.00*	.	.	.	265	6
655.00*	.	.	.	265	5
659.00*	.	.	.	265	4
734.00*	.	.	.	265	3
762.00*	.	.	.	265	2
763.00*	.	.	.	265	1
805.00*	0.1641	0.8359	.	265	0

NOTE: The marked survival times are censored observations.

The LIFETEST Procedure

Stratum 2: course = 1

Summary Statistics for Time Variable time

Quartile Estimates

Percent	Point Estimate	Transform	95% Confidence Interval [Lower	Upper)
75	290.00	LOGLOG	243.00	384.00
50	131.50	LOGLOG	113.00	154.00
25	74.00	LOGLOG	55.00	83.00

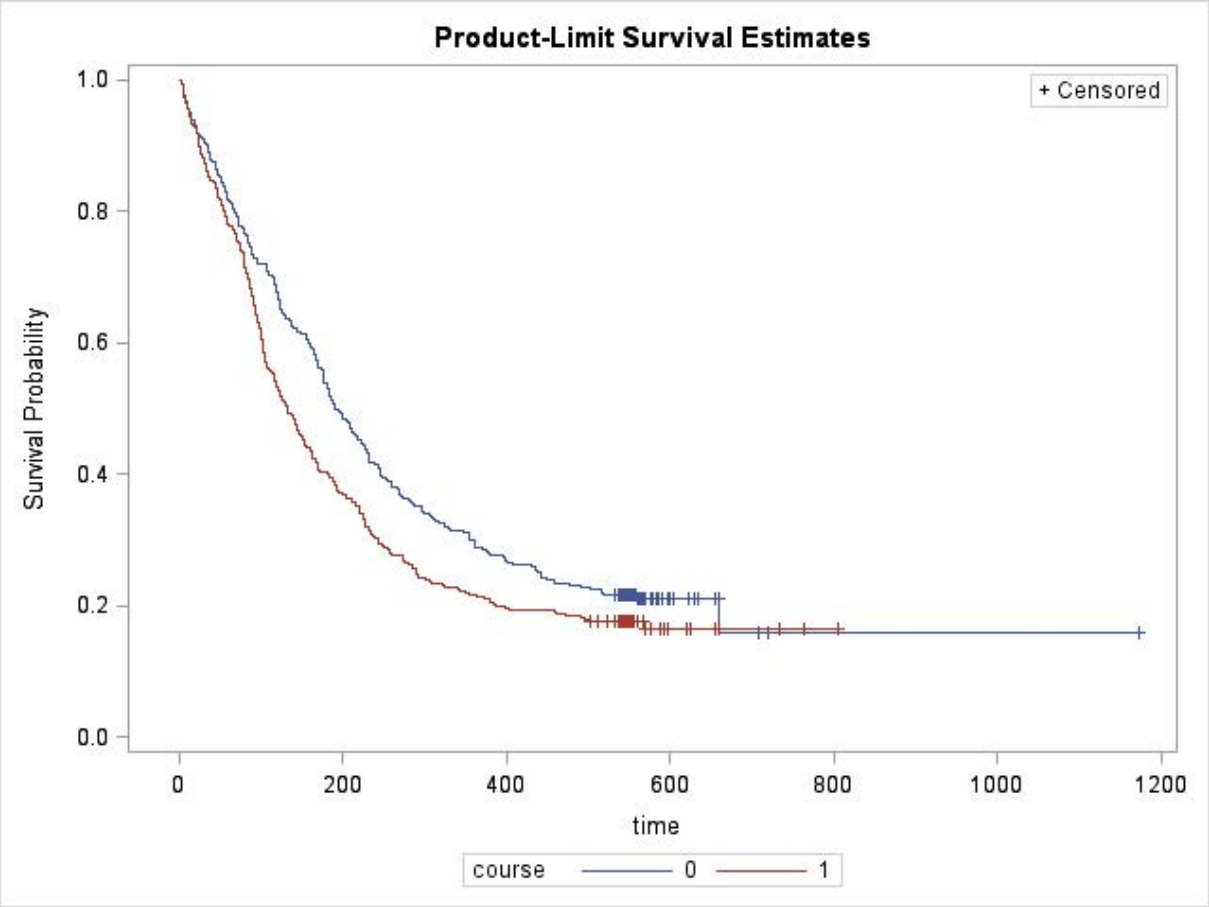
Mean	Standard Error
210.93	10.70

NOTE: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Summary of the Number of Censored and Uncensored Values

Stratum	course	Total	Failed	Censored	Percent Censored
1	0	308	243	65	21.10
2	1	320	265	55	17.19
Total		628	508	120	19.11

here are the estimated survival functions; the upper curve for people that did not take the course ($course = 0$), the lower curve for those that did ($course = 1$)



Homogeneity of Survival Curves: $H_0 : S_1(t) = S_0(t)$ vs $H_a : not\ H_0$

one can use the LogRank test and the WIlcoxon test where the latter is a weighted version of the first test which puts more weight to earlier times which is typically more reliable because there are more observations at earlier times

RESULTS

Testing Homogeneity of Survival Curves for time over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	6.7979	1	0.0091
Wilcoxon	9.4608	1	0.0021

we can reject the null hypothesis, the difference between both survival functions is not due to random fluctation, they are significantly different from each other

parametric description of survival functions

assume a theoretical distribution and estimate its parameters

some frequently used simple theoretical distributions:

- **exponential** $(\alpha > 0)$

$$f(t) = \frac{1}{\alpha}\exp(-\frac{t}{\alpha}) \qquad mrl(0) = \alpha$$

$$S(t) = \exp(-\frac{t}{\alpha}) \qquad mrl(t) = \alpha$$

$$h(t) = -\frac{d\ln S(t)}{dt} = \frac{1}{\alpha}$$

no memory: having survived until t has no effect on the risk of dying in the next instant nor on the mean residual life time

- **lognormal**: T has a lognormal distribution if $X = \ln(T)$ has a normal distribution (μ is called the intercept and σ the scale)

$$\begin{aligned} f(t) &= \frac{1}{\sigma t}\,\phi\left(\frac{\ln(t)-\mu}{\sigma}\right) \\ S(t) &= 1-\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right) \\ mrl(0) &= \exp\left(\mu+0.5\sigma^2\right) \\ t_p &= \exp(\mu+\sigma z_p) \end{aligned}$$

with ϕ and Φ the density and the cumulative standard normal distribution and z_p the p^{th} percentile of a N(0,1)

- **Weibull** (there exist a lot of different parameterizations!*)

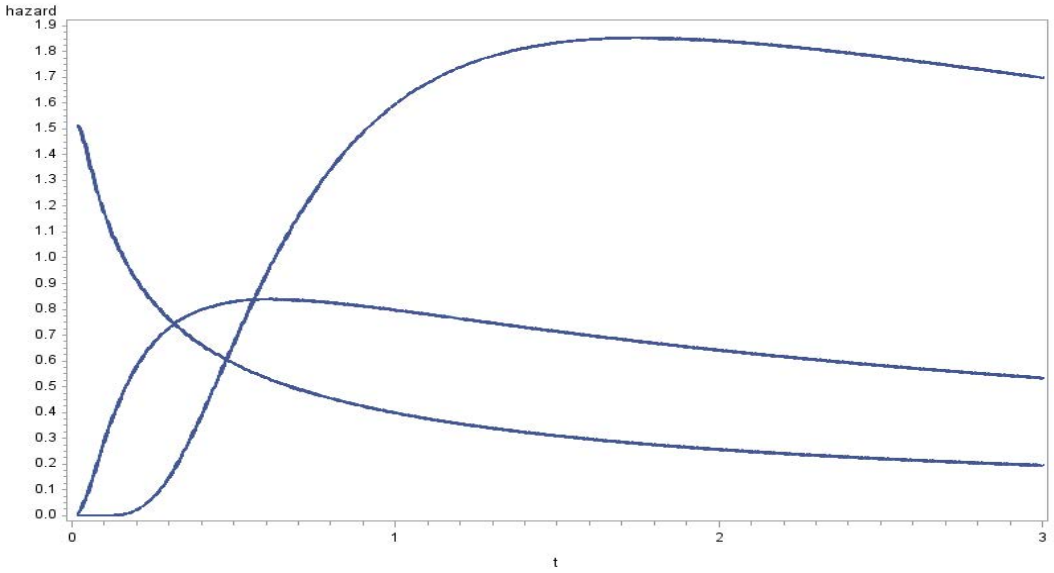
$$
\begin{aligned}
f(t) &= \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1}\exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right) \\
S(t) &= exp\left(-\left(\frac{t}{\alpha}\right)^{\gamma}\right) \qquad\qquad mrl(0) = \alpha\,\Gamma\left(1+\frac{1}{\gamma}\right) \\
h(t) &= \frac{\gamma}{\alpha}\left(\frac{t}{\alpha}\right)^{\gamma-1} \qquad\qquad t_p = \alpha\left[-\ln(1-p)\right]^{1/\gamma}
\end{aligned}
$$

with γ the **shape** parameter which defines whether the hazard function decreases ($\gamma < 1$), increases ($\gamma > 1$) or is constant ($\gamma = 1$) in which case the Weibull is equal to the exponential

the parameter α is called the **scale** parameter and denotes how much the distribution is stretched.

*Γ stands for the gamma function; for positive integers m: $\Gamma(m) = (m-1)!$

lognormal hazard functions

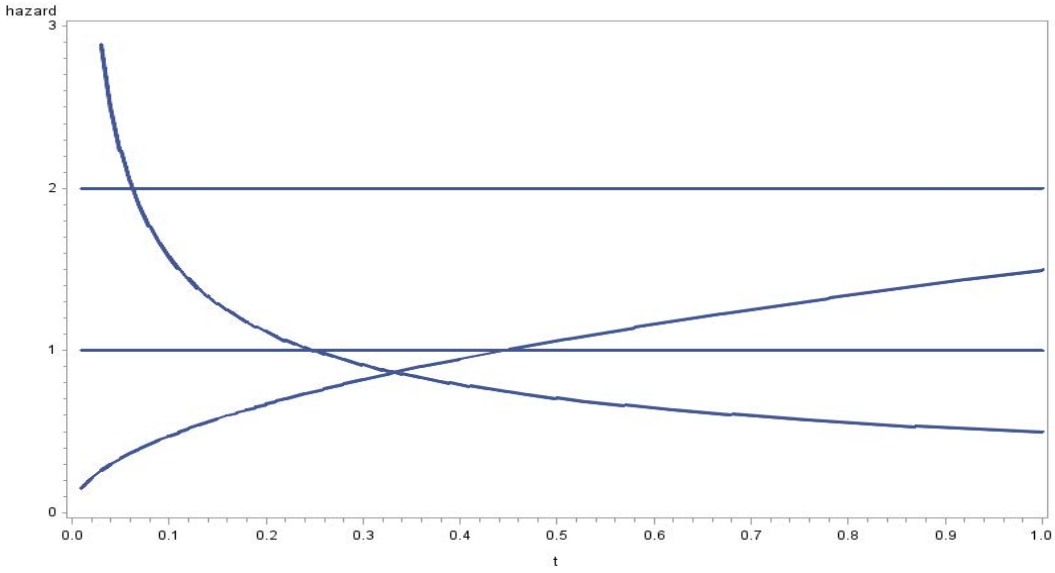


$\mu = 0,\ \sigma = 0.5$

$\mu = 0,\ \sigma = 1.0$

$\mu = 0,\ \sigma = 2.0$

Weibull hazard functions



$\alpha = 0.5,\ \gamma = 1.0$

$\alpha = 1.0,\ \gamma = 1.5$

$\alpha = 1.0,\ \gamma = 1.0$

$\alpha = 1.0,\ \gamma = 0.5$

if T has a Weibull distribution with parameters α and γ, then $X = \ln(T)$ has a **Gumbel or extreme value distribution** with parameters μ and σ:

$$f(x) = \frac{1}{\sigma} \exp\left(\frac{x-\mu}{\sigma} - \exp\left(\frac{x-\mu}{\sigma}\right)\right)$$

where σ is called the **scale** parameter and μ the **intercept**; the parameters of the two models are related as follows:

$$\sigma = \gamma^{-1} \qquad \mu = \ln(\alpha)$$

(remark that the *scale* parameter of this extreme value distribution is related to the *shape* parameter of the corresponding Weibull distribution!)

- if T has a lognormal (μ, σ^2) distribution,
  then $X = \ln T$ has a normal (μ, σ^2) distribution and
  $X = \ln T$ can be written as $X = \ln T = \mu + \sigma\epsilon$ with ϵ a standard normal distribution with $\mu = 0$ and $\sigma^2 = 1$

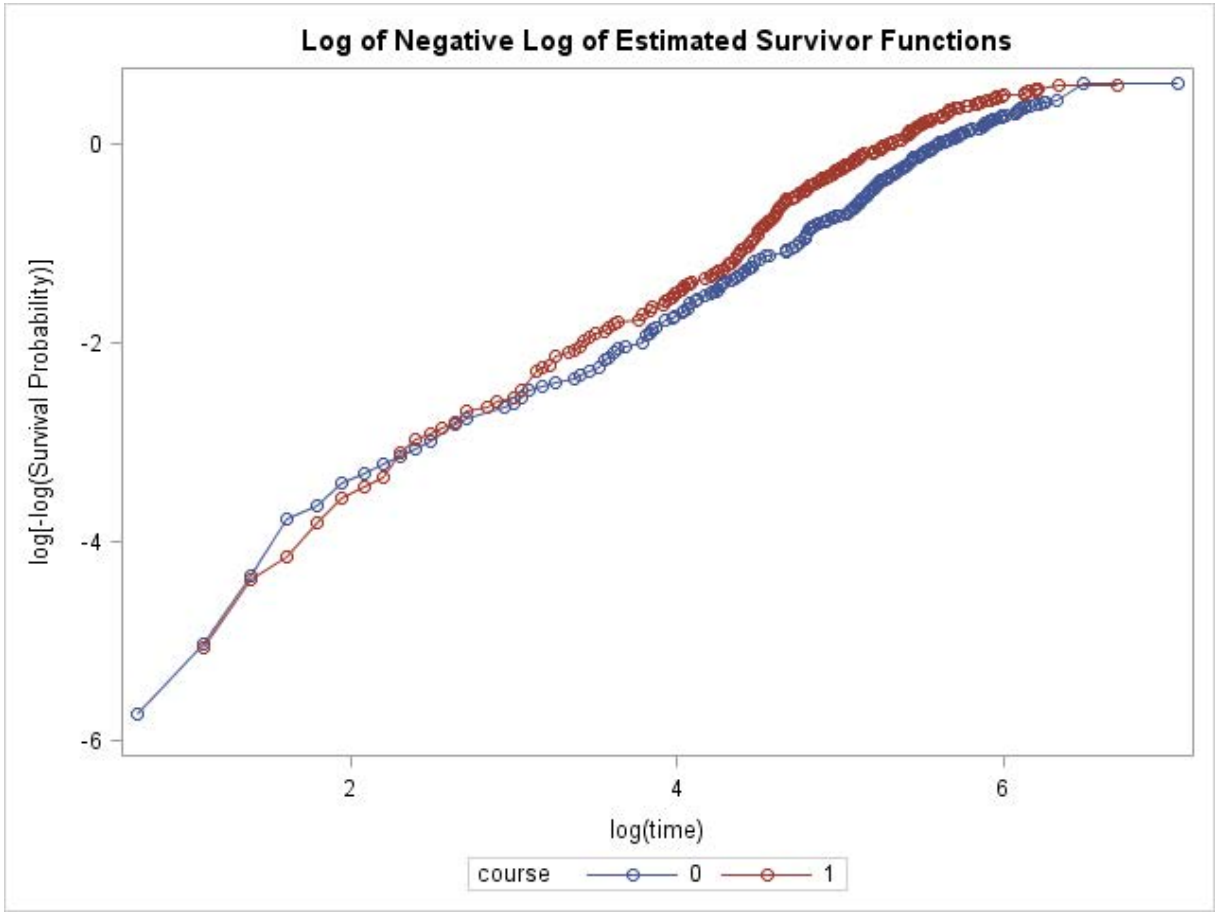
- if T has a Weibull (α, γ) distribution,
  then $X = \ln T$ has a Gumbel (μ, σ^2) distribution and
  $X = \ln T$ can be written as $X = \ln T = \mu + \sigma\epsilon$ with ϵ a standard Gumbel distribution with $\mu = 0$ and $\sigma^2 = 1$

visual check: remark that for a Weibull distribution:

$$\ln(-\ln S(t)) = -\gamma \ln(\alpha) + \gamma \ln(t)$$

to check the goodness-of-fit of the Weibull:

- plot $\ln(-\ln S(t))$ against $\ln(t)$
- check whether you get a straight line
- if so, you can derive estimates for α and γ based on this plot


Log of Negative Log of Estimated Survivor Functions
log[-log(Survival Probability)]
0
-2
-4
-6
2
4
6
log(time)
course
0
1

based on the intercepts and slopes in this plot, we get approximate values for α and γ (lines very similar, we fit only one line):

based on the line through (5.8,0) and (2,-3.5) (the part which is most reliable) we get:

$$\ln(-\ln S(t)) = -5.34 + 0.92 \ln(t)$$

from which we get the following estimated values:

$$\begin{aligned} \hat{\gamma} &= 0.92 \\ \hat{\gamma}\ \widehat{\ln(\alpha)} &= 5.34 \Rightarrow \hat{\mu} = \widehat{\ln(\alpha)} = 5.34/0.92 = 5.80 \\ \hat{\alpha} &= \exp(\hat{\mu}) = 330.30 \end{aligned}$$

$\gamma < 1 \Rightarrow$ hazard decreases: increasingly more difficult to find a job

more precise estimates can be obtained by maximizing the likelihood which is computed as

$$L = \prod_{i\ events} f(t_i) \prod_{i\ censored} S(t_i).$$

where the appropriate functions for $f(t)$ and $S(t)$ are inserted depending on the assumptions made

we fit a Weibull distribution for course = 0 we get:

RESULTS

Model Information

Data Set	WORK.EMPLOYMENT
Dependent Variable	Log(time)
Censoring Variable	censor
Censoring Value(s)	0
Number of Observations	308
Noncensored Values	243
Name of Distribution	Weibull
Log Likelihood	-499.9966709

Fit Statistics	
-2 Log Likelihood	999.993
AIC (smaller is better)	1003.993
BIC (smaller is better)	1011.454

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	3297.246
Weibull AIC (smaller is better)	3301.246
Weibull BIC (smaller is better)	3308.706

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.7850	0.0712	5.6455	5.9245	6604.01	<.0001
Scale	1	1.1089	0.0603	0.9968	1.2335		
Weibull Scale	1	325.3836	23.1630	283.0097	374.1020		
Weibull Shape	1	0.9018	0.0490	0.8107	1.0032		

from which we get

$$\begin{aligned}\hat{\mu} &= 5.785 \text{ or } \hat{\alpha} = \exp(\hat{\mu}) = 325.3836 \\ \hat{\sigma} &= 1.1089 \text{ or } \hat{\gamma} = \frac{1}{\hat{\sigma}} = 0.9018\end{aligned}$$

so pretty close to the values we estimated based on the plot of $\ln(-\ln S(t))$

for course $= 1$: $\hat{\alpha} = 249.2667$ and $\hat{\gamma} = 0.8671$ (output not shown)

$$\begin{aligned}mrl(0)|course = 0 &= \alpha\Gamma\left(1 + \frac{1}{\gamma}\right) = 341.9912 \\ mrl(0)|course = 1 &= \alpha\Gamma\left(1 + \frac{1}{\gamma}\right) = 267.9143\end{aligned}$$

AIC and BIC can be used to compare the fit of different models:

$$\begin{aligned} AIC &= -2\ln L + 2(p+k) \\ BIC &= -2\ln L + (p+k)\ln(n) \end{aligned}$$

with p the number of covariates (is zero until now) and k the number of parameters in the distribution function (1 for exponential, 2 for Weibull and lognormal...) (the smaller, the better)

if we fit a lognormal survival time we get the following estimates (remark that according to AIC, the lognormal fit is slightly better than the Weibull fit for both groups):

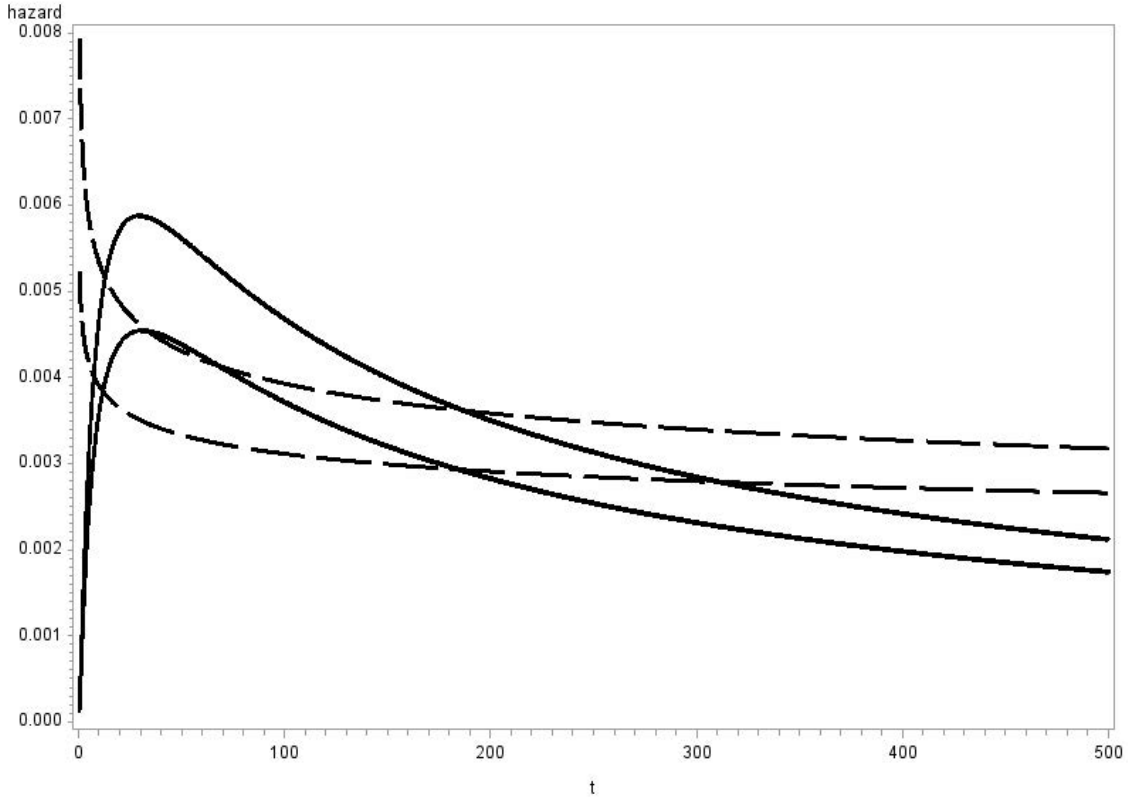
$$course = 0: \quad \hat{\mu} = 5.2465; \hat{\sigma} = 1.4504$$

$$course = 1: \quad \hat{\mu} = 4.9529; \hat{\sigma} = 1.3754$$

and therefore

$$\begin{aligned} mrl(0)|course = 0 &= \exp(5.2465 + 0.50 \times 1.4504^2) = 543.663 \\ mrl(0)|course = 1 &= \exp(4.9529 + 0.50 \times 1.3754^2) = 364.586 \end{aligned}$$

a plot of the 4 hazard functions looks like this:



weibull: course = 1
weibull: course = 0
lognormal: course = 1
lognormal: course = 0

the accelerated failure time (AFT) model or the loglinear model (can deal with explanatory variables)

assume now that $\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$:

$\ln(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \sigma\epsilon$ and ϵ a s.v.

or $T = e^{\beta_0 + \sigma\epsilon} \; e^{\beta_1 x_1 + \ldots + \beta_k x_k}$

$$\begin{aligned} \text{let} \quad S_0(t) &= P(T > t | \forall i : x_i = 0) = P(e^{\beta_0 + \sigma\epsilon} > t) \\ \text{then} \quad S_{\mathbf{x}}(t) &= P(T > t | x_1, \ldots x_p) = P\left(e^{\beta_0 + \sigma\epsilon} > t \; e^{-(\beta_1 x_1 + \ldots + \beta_k x_k)}\right) \\ &= S_0\left(t \; e^{-(\beta_1 x_1 + \ldots + \beta_k x_k)}\right) \end{aligned}$$

so the effect of the covariates is to change the timescale: if $e^{-(\beta_1 x_1 + \ldots + \beta_k x_k)} > 1$, the time is accelerated, otherwise extended

- $T_{x_i+1} = e^{\beta_i} T_{x_i}$ so if $\beta_i < 0$ an increase in the corresponding covariate accelerates the time as it decreases the survival time; if $\beta_i > 0$ an increase in the corresponding variable extends the time as the survival time increases

- $S_{x_i+1}(t) = S_{x_i}(te^{-\beta_i})$ or $S_{x_i}(t) = S_{x_i+1}(te^{\beta_i})$

- the parameters are again estimated by maximum likelihood

- the residuals are often assumed to be normally distributed (then the survival time is lognormally distributed) or Gumbel distributed (then the survival time is Weibull distributed)

example: we will study the effect of age, attending the re-orientation course and the number of previous jobs on the time until employment (and compute the 0.1, 0.5 and 0.9 percentiles) assuming a Weibull survival distribution

RESULTS

Model Information

Data Set	WORK.EMPLOYMENT
Dependent Variable	Log(time)
Censoring Variable	censor
Censoring Value(s)	0
Number of Observations	623
Noncensored Values	504
Name of Distribution	Weibull
Log Likelihood	-1016.831971

```
    Number of Observations Read          628
    Number of Observations Used          623
    Missing Values                         5

                  Fit Statistics
-2 Log Likelihood                    2033.664
AIC (smaller is better)              2043.664
BIC (smaller is better)              2065.837


       Fit Statistics (Unlogged Response)
-2 Log Likelihood                    6707.083
Weibull AIC (smaller is better)      6717.083
Weibull BIC (smaller is better)      6739.256

Algorithm converged.
```

(remark that these fit statistics cannot be compared with those of the previous outputs because those were based on a subset of the observations for which $course = 0$)

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	5.3765	0.2923	4.8036	5.9493	338.40	<.0001
age	1	0.0149	0.0081	-0.0009	0.0306	3.41	0.0649
course	1	-0.2527	0.0999	-0.4485	-0.0569	6.40	0.0114
prevj	1	-0.0389	0.0547	-0.1462	0.0684	0.50	0.4777
Scale	1	1.1169	0.0413	1.0388	1.2009		
Weibull Shape	1	0.8953	0.0331	0.8327	0.9627		

Obs	ID	age	time	censor	prevj	course	_PROB_	pred
1	1	39	188	1	3	0	0.1	27.830
2	1	39	188	1	3	0	0.5	228.190
3	1	39	188	1	3	0	0.9	872.242
4	2	33	26	1	3	0	0.1	25.455
5	2	33	26	1	3	0	0.5	208.722
6	2	33	26	1	3	0	0.9	797.824
.....								

the time until employment is not significantly affected by age or the number of previous employments but is does decrease significantly by attending the course: $\frac{T_{course=1}}{T_{course=0}} = \exp(-0.2527) = 0.78$ (all else equal), so the time until employment decreases with 22% by attending the course

or, in terms of the survival function:

$$S_{course=1}(t) = S_{course=0}(te^{-(-0.2527)}) = S_{course=0}(1.29\ t)$$

for 39 year old persons with prevj = 3 and course = 0, we predict that 10% finds a job within 27.830 days, 50% within 228.190 days and 90% within 872.242 days.

remark that these values can be computed by

$$t_p = \alpha(-\ln(1-p))^{1/\gamma} = \exp(\mu)(-\ln(1-p))^{\sigma}$$

with $age = 39, prevj = 3$ and $course = 0$ we get:

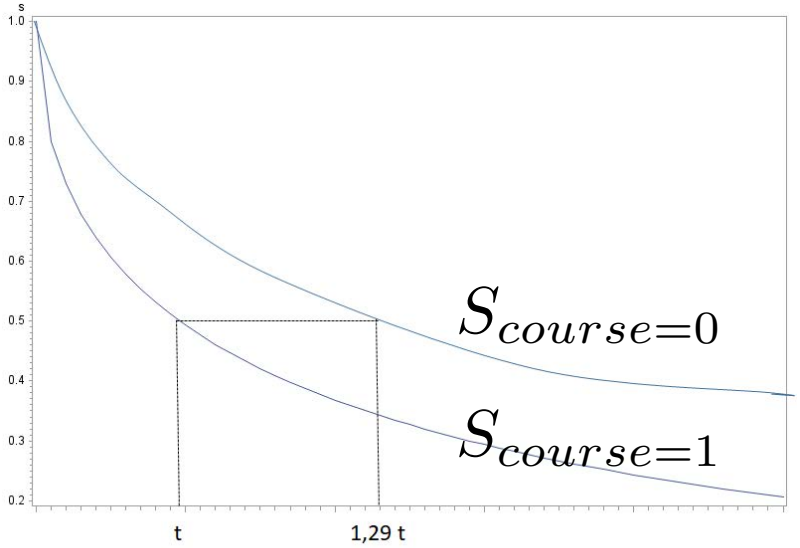
$$\hat{\alpha} = \exp(5.3765 + 0.0149 \times 39 - 0.0389 \times 3) = \exp(5.84) = 344.0889$$

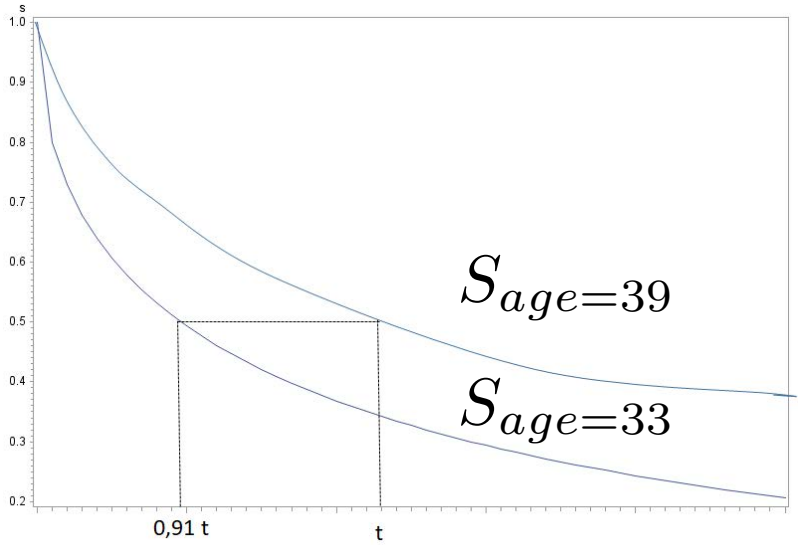
and therefore

$$\begin{aligned} t_{0.10} &= 344.0889\ (-\ln(1-0.10))^{1/0.8953} = 27.83 \\ t_{0.50} &= 344.0889\ (-\ln(1-0.50))^{1/0.8953} = 288.19 \\ t_{0.90} &= 344.0889\ (-\ln(1-0.90))^{1/0.8953} = 873.24 \end{aligned}$$

the left plot illustrates that $S_{course=1}(t) = S_{course=0}\ (1.29\ t)^*$

the right plot illustrates that $S_{age=39}(t) = S_{age=33}\ (0.91\ t)$
with $0.91 = e^{-6*0.0149}$





*remark that the scales are far from correct

with a normally distributed error term, we get a slightly better fit:

RESULTS

Model Information

Data Set	WORK.EMPLOYMENT
Dependent Variable	Log(time)
Censoring Variable	censor
Censoring Value(s)	0
Number of Observations	623
Noncensored Values	504
Right Censored Values	119
Name of Distribution	Lognormal
Log Likelihood	-1000.805474

Number of Observations Read	628
Number of Observations Used	623
Missing Values	5

Fit Statistics	
-2 Log Likelihood	2001.611
AIC (smaller is better)	2011.611
BIC (smaller is better)	2033.784

Fit Statistics (Unlogged Response)	
-2 Log Likelihood	6675.030
Lognormal AIC (smaller is better)	6685.030
Lognormal BIC (smaller is better)	6707.203

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.9108	0.3272	4.2694	5.5522	225.19	<.0001
age	1	0.0175	0.0093	-0.0008	0.0357	3.50	0.0614
course	1	-0.2855	0.1136	-0.5082	-0.0629	6.32	0.0120
prevj	1	-0.1163	0.0632	-0.2401	0.0075	3.39	0.0656
Scale	1	1.3813	0.0456	1.2948	1.4736		

some predicted percentiles:

Obs	ID	age	time	censor	course	prevj	_PROB_	pred
1885	1001	39	.	.	1	3	0.1	24.21
1886	1001	39	.	.	1	3	0.5	142.18
1887	1001	39	.	.	1	3	0.9	834.90
1888	1002	39	.	.	0	3	0.1	32.21
1889	1002	39	.	.	0	3	0.5	189.16
1890	1002	39	.	.	0	3	0.9	1110.81
1891	1003	33	.	.	1	3	0.1	21.80
1892	1003	33	.	.	1	3	0.5	128.04
1893	1003	33	.	.	1	3	0.9	751.90
1894	1004	33	.	.	0	3	0.1	29.01
1895	1004	33	.	.	0	3	0.5	170.35
1896	1004	33	.	.	0	3	0.9	1000.37

---

attending a course now decreases the time with 25% (as $\exp(-0.2855) = 0.75$)

remark that percentiles can here be computed by

$$t_p = \exp(\mu + \sigma z_p)$$

with μ estimated as

$$4.9108 + 0.0175 \times 39 - 0.2855 \times 0 - 0.1163 \times 3 = 5.2444$$

and with $z_{0.10} = -1.2855$, $z_{0.50} = 0$ and $z_{0.90} = 1.2855$ this yields

$$\begin{aligned} t_{0.1} &= 32.096 \\ t_{0.5} &= 189.502 \\ t_{0.9} &= 1118.855 \end{aligned}$$

alternatively one can use the **COX proportional hazard model** in which not the survival time but the hazard function is modeled:

$$h(t) = h_0(t)\ e^{\theta_1 x_1 + \theta_2 x_2 + ... + \theta_k x_k}$$

with $h_0(t)$ the baseline hazard function which does not have to be specified; it is called the *proportional* hazard model because the *hazard ratio* does not depend on t:

$$\exp(\theta_j) = \frac{h(t)\ at\ (x_1, \ldots, x_j + 1, \ldots, x_k)}{h(t)\ at\ (x_1, \ldots, x_j, \ldots, x_k)}$$

(we will not deal with this model...)

extensions:

- models with time-dependent covariates
- models with random parameters are called **frailty models**
- multivariate survival models: model for correlated lifetimes (of husband & wives, of siblings)
- competing risks: duration until one of several types of events (time till bankruptcy or acquisition)

PART 4: MULTIVARIATE STATISTICS

- PRINCIPAL COMPONENT ANALYSIS
- EXPLORATORY FACTOR ANALYSIS
- DISCRIMINANT ANALYSIS
- CLUSTER ANALYSIS

4a. PRINCIPAL COMPONENT ANALYSIS

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA): construct new variables which are linear combinations of the original variables such that the first PC captures most information, the second PC captures most of the remaining information, etcetera; the maximum number of new variables that can be constructed is equal to the number of original variables, and the new variables are uncorrelated among themselves.

example: consider the weight and height of a group of people; these are often heavily correlated and can therefore be summarized

by calculating a person's 'figure', it is possible to capture a reasonable part of the information of two variables in one variable

however, when the dataset would consist of the weight and income of a group of people, the variables will not be correlated much: performing a PCA on these data is possible, but representing the data by just one new variable will lead to a substantial loss of information (and the new variable would be meaningless)

this shows that the use of PCA depends highly on the structure of the covariance matrix: highly correlated variables can well be summarized in less dimensions without losing much information

data preprocessing

we start from the matrix $\mathbf{X}$ with n observations on k variables:

	$variables$					
obs	x_1	x_2	$\ldots$	x_j	$\ldots$	x_k
1	x_{11}	x_{12}	$\ldots$	x_{1j}	$\ldots$	x_{1k}
2	x_{21}	x_{22}	$\ldots$	x_{2j}	$\ldots$	x_{2k}
$\vdots$						
i	x_{i1}	x_{i2}	$\ldots$	x_{ij}	$\ldots$	x_{ik}
$\vdots$						
n	x_{n1}	x_{n2}	$\ldots$	x_{nj}	$\ldots$	x_{nk}

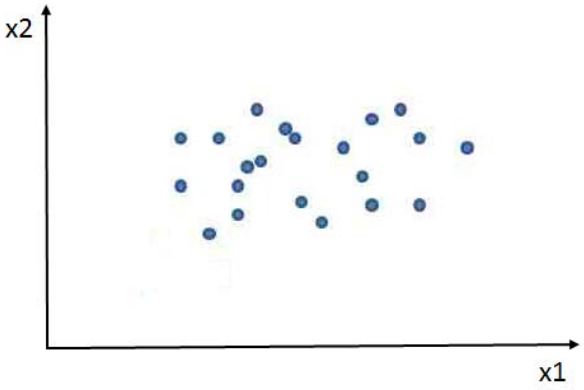
we will assume that all variables are numerical!

PCA needs data that are either mean-corrected or standardized

- **mean-corrected or centered data:** subtract the mean $\bar{x}_j$ from each x_{ij} ($=$ moving the origin to the *centroid* of the dataset)
  (the sign contains some information about the observation)

- **standardized data:** divide the mean-corrected x_j variable by the standard deviation s_j of that variable
  (the value contains information about the relative position of the observation)

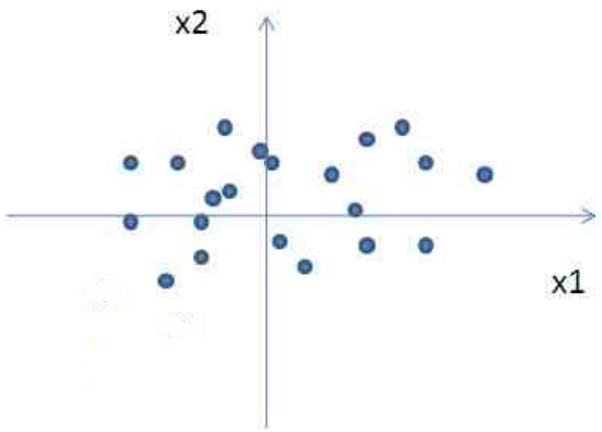
raw data



$$mean = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}$$

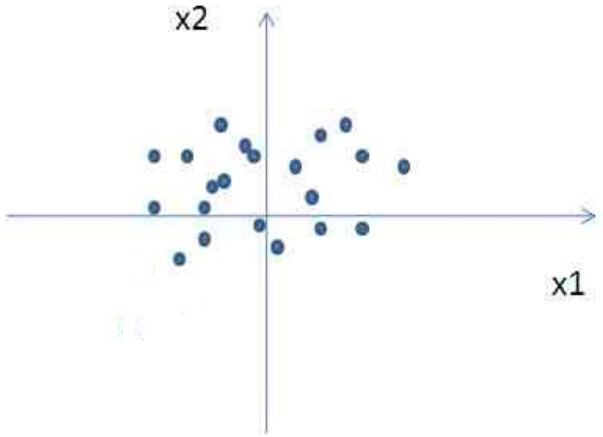
mean-corrected data



$$mean = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} s_1^2 & s_{12} \\ s_{12} & s_2^2 \end{pmatrix}$$

standardized data



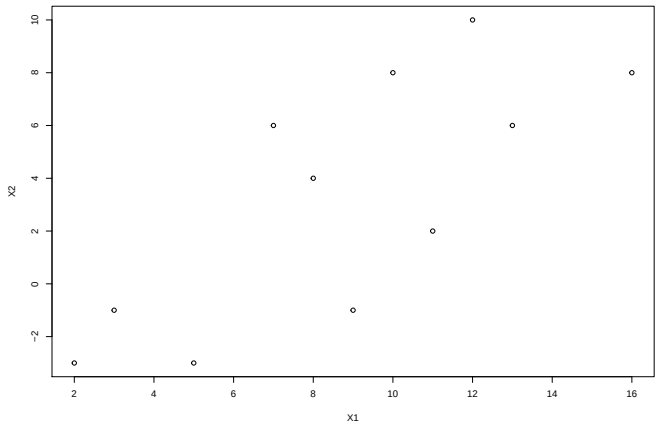
$$mean = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\mathbf{S} = R = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

$\mathbf{S}$ stands for the sample covariance matrix (which estimates $\mathbf{\Sigma}$, the population covariance matrix) and $\mathbf{R}$ for the correlation matrix (population and sample version!)

the method is explained using these data (k=2):

obs.	original x_1	original x_2	mean-centered centered x_1	mean-centered centered x_2
1	16	8	8	5
2	12	10	4	7
3	13	6	5	3
4	11	2	3	-1
5	10	8	2	5
6	9	-1	1	-4
7	8	4	0	1
8	7	6	-1	3
9	5	-3	-3	-6
10	3	-1	-5	-4
11	2	-3	-6	-6
12	0	0	-8	-3
mean	8	3	0	0
variance	23.091	21.091	23.091	21.091



$$\mathbf{S} = \begin{pmatrix} 23.091 & 16.455 \\ 16.455 & 21.091 \end{pmatrix}$$

$$\mathbf{R} = \begin{pmatrix} 1.000 & 0.746 \\ 0.746 & 1.000 \end{pmatrix}$$

theorem

- let $x_1, x_2, \ldots, x_k$ be stochastic variables and $\mathbf{\Sigma} = cov(\mathbf{x})$ the corresponding covariance matrix; let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$ be the k eigenvalues of $\mathbf{\Sigma}$ and $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$ be the corresponding eigenvectors

- the principal components PC_j, $j = 1 \ldots k$, corresponding with the variables $x_1, x_2, \ldots, x_k$, are given by

$$PC_j = \mathbf{e}_j' \mathbf{x} = e_{j1} x_1 + e_{j2} x_2 + \ldots + e_{jk} x_k = \mathbf{x}' \mathbf{e}_j$$

where $var(PC_j) = \lambda_j$ and $cov(PC_i, PC_j) = 0$ for $i \neq j$.

proof:

- first, look for the linear combination $\mathbf{c}$ that maximizes

$$\begin{aligned} var(\mathbf{c}'\mathbf{x}) &= var(c_1x_1 + c_2x_2 + \ldots + c_kx_k) \\ &= c_1^2\ var(x_1) + c_2^2\ var(x_2) + \ldots + c_k^2\ var(x_k) \\ &\quad +2\sum_i \sum_{j>i} c_ic_jcov(x_i, x_j) = \mathbf{c}'\mathbf{\Sigma}\mathbf{c} \end{aligned}$$

- normalize $\mathbf{c}$ to prevent the variance from going to infinity: suppose $\sum_{i=1}^{p} c_i^2 = \mathbf{c}'\mathbf{c} = 1$ (this comes down to fixing the scale of the new variable)

- the corresponding Lagrange function and first derivative are:

$$\begin{aligned} L &= \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} + \lambda(1 - \mathbf{c}'\mathbf{c}) \\ \frac{\partial L}{\partial \mathbf{c}} &= 2\boldsymbol{\Sigma}\mathbf{c} - 2\lambda\mathbf{c} = 2(\boldsymbol{\Sigma} - \lambda I)\mathbf{c} = \mathbf{0} \end{aligned}$$

- so, λ has to be an eigenvalue of $\boldsymbol{\Sigma}$ and $\mathbf{c}$ the corresponding eigenvector

- multiplication with $\mathbf{c}'$ gives $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} - \lambda\mathbf{c}'\mathbf{c} = 0$ or $\lambda = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$

- since $\mathbf{c}'\boldsymbol{\Sigma}\mathbf{c}$ has to be maximized take the largest eigenvalue λ_1 and the eigenvector $\mathbf{e}_1$

similarly one can prove that the remaining principal components PC_j, $j = 2 \ldots k$ which have maximal variance under the constraint that they are uncorrelated to the previous PC_i's, are given by $\mathbf{e}_j'\mathbf{x}$

for the example we need the eigenvalues and eigenvectors of

$$\mathbf{S} = \begin{pmatrix} 23.09 & 16.45 \\ 16.45 & 21.09 \end{pmatrix}$$

the eigenvalues are defined by

$$\begin{vmatrix} 23.09 - \lambda & 16.45 \\ 16.45 & 21.09 - \lambda \end{vmatrix} = 0$$

$$\Rightarrow (23.09 - \lambda) * (21.09 - \lambda) - 16.45^2 = 0$$

or $216.36 - 44.18\lambda + \lambda^2 = 0$

from which $\lambda_1 = 38.57$ and $\lambda_2 = 5.61$

the corresponding eigenvectors are defined by

$$\begin{pmatrix} 23.09-\lambda_1 & 16.45 \\ 16.45 & 21.09-\lambda_1 \end{pmatrix} \begin{pmatrix} e_{11} \\ e_{12} \end{pmatrix} = 0$$

$$\begin{pmatrix} 23.09-\lambda_2 & 16.45 \\ 16.45 & 21.09-\lambda_2 \end{pmatrix} \begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix} = 0$$

the normalized and orthogonal eigenvectors are* :

$$\mathbf{e}_1 = (0.728 \;\; 0.685)' \text{ and } \mathbf{e}_2 = (-0.685 \;\; 0.728)'$$

*remark that the original matrix can be reconstructed based on the eigenvalues and eigenvectors; check that

$$\begin{pmatrix} 23.09 & 16.45 \\ 16.45 & 21.09 \end{pmatrix} = 38.57 \begin{pmatrix} 0.728 \\ 0.685 \end{pmatrix} (0.728 \;\; 0.685) + 5.606 \begin{pmatrix} -0.685 \\ 0.728 \end{pmatrix} (-0.685 \;\; 0.728) = \lambda_1 \mathbf{e}_1 \mathbf{e}_1' + \lambda_2 \mathbf{e}_2 \mathbf{e}_2'$$

the total information does not change:

$$
\begin{aligned}
& var(PC_1) + var(PC_2) + \ldots + var(PC_k) \\
\overset{by\ construction}{=} \quad & \lambda_1 + \lambda_2 + \ldots \lambda_k \\
\overset{property\ of\ eigenvalues}{=} \quad & trace(\mathbf{\Sigma}) \\
\overset{definition\ of\ trace}{=} \quad & var(x_1) + var(x_2) \ldots + var(x_k)
\end{aligned}
$$

Covariance Matrix

	X1	X2
X1	23.09090909	16.45454545
X2	16.45454545	21.09090909

Total Variance 44.181818182

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	38.5758133	32.9698084	0.8731	0.8731
2	5.6060049		0.1269	1.0000

Eigenvectors

	PC1	PC2
X1	0.728238	-.685324
X2	0.685324	0.728238

Simple Statistics

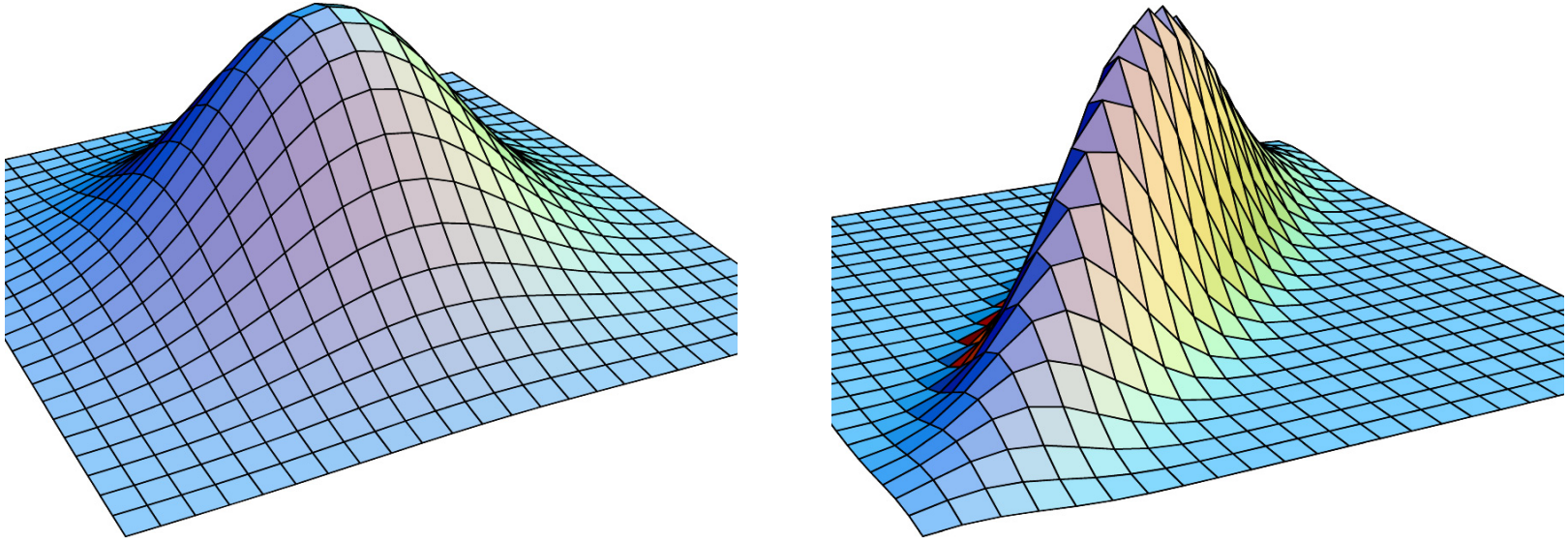
Variable	N	Mean	Std Dev	Sum
X1	12	8.00000	4.80530	96.00000
X2	12	3.00000	4.59248	36.00000
PC1	12	0	6.21094	0
PC2	12	0	2.36770	0

geometrical interpretation

consider a **multivariate normal distribution**:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

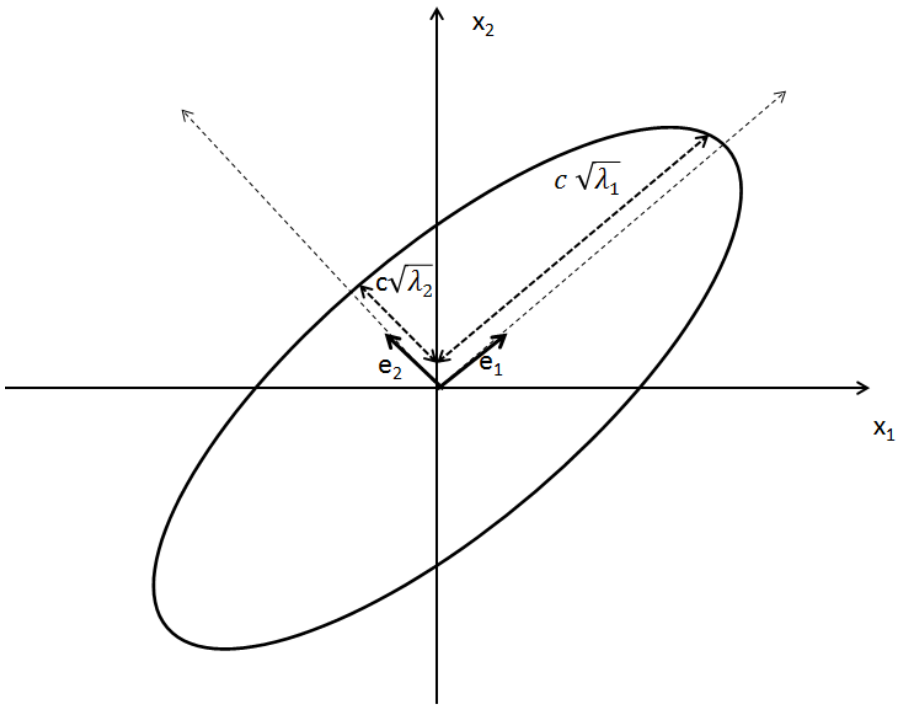
for $k = 2$: (left: without correlation, right: with correlation)



all points with the same density satisfy (for some constant c):

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2$$

these points lie on an ellipsoid around $\boldsymbol{\mu}$ with axes along the eigenvectors $\mathbf{e}_1, \mathbf{e}_2, \ldots$ of $\boldsymbol{\Sigma}$ with length $c\sqrt{\lambda_1}, c\sqrt{\lambda_2}, \ldots$ where λ_i are the eigenvalues of $\boldsymbol{\Sigma}$



$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ is the squared **Mahalanobis or statistical distance** between $\boldsymbol{\mu}$ and $\mathbf{x}$

more generally, the distance between points $\mathbf{x}_i$ and $\mathbf{x}_j$ is:

$$d_s(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\left[(\mathbf{x}_i - \mathbf{x}_j)' \ \ \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right]}$$

so all the observations with the same density have the same Mahalanobis distance to $\boldsymbol{\mu}$

PC scores

- are the values of the new variables for each observation
- use the eigenvectors as weights to calculate the PC scores using the mean-corrected data:

$$\begin{aligned} PC_1 &= 0.728x_1 + 0.685x_2 \\ PC_2 &= -0.685x_1 + 0.728x_2 \end{aligned}$$

- for the first observation: $0.728*(16-8)+0.685*(8-3) = 9.2525$

  (if one starts from standardized variables, one has to use the standardized values to compute the scores too)

RESULTS

Obs	X1	X2	PC1	PC2
1	16	8	9.25253	-1.84140
2	12	10	7.71022	2.35637
3	13	6	5.69716	-1.24191
4	11	2	1.49939	-2.78421
5	10	8	4.88310	2.27054
6	9	-1	-2.01306	-3.59828
7	8	4	0.68532	0.72824
8	7	6	1.32773	2.87004
9	5	-3	-6.29666	-2.31346
10	3	-1	-6.38249	0.51367
11	2	-3	-8.48137	-0.25748
12	0	0	-7.88188	3.29788

the loadings that define the principal components (here the eigenvectors) are called the **pattern loadings**

structural or structure loadings

- to what extent are the original variables important for the $PC's$?
- the structural loadings are $corr(PC_j, x_i)$
- e.g.: $corr(PC_1, x_1) = 0.941$ and $corr(PC_1, x_2) = 0.927$
  $\Rightarrow$ both x_1 and x_2 are important in forming PC_1
- loadings are useful for interpreting and labeling the $PC's$; traditionally one uses as cut-off point for the (absolute values of) structural loadings: 0.5

RESULTS

Pearson Correlation Coefficients, N = 12
Prob > |r| under H0: Rho=0

	X1	X2	PC1	PC2
X1	1.00000	0.74562	0.94126	-0.33768
		0.0054	<.0001	0.2831
X2	0.74562	1.00000	0.92684	0.37545
	0.0054		<.0001	0.2291
PC1	0.94126	0.92684	1.00000	0.00000
	<.0001	<.0001		1.0000
PC2	-0.33768	0.37545	0.00000	1.00000
	0.2831	0.2291	1.0000	

effect of standardization of data

- one can perform PC-analysis on mean-corrected data or standardized data (the latter is the default)

- if based on mean-corrected data, the $PC's$ are scale dependent!

  e.g.: let $var(x_1) \approx var(x_2)$ and rescale x_2: $x_2^* = 1000 \times x_2$

  $\Rightarrow var(x_2^*) = 1000000 \times var(x_2)$ and $PC_1 \approx x_2^*$

- when all variables are equally important, but are differently scaled $\Rightarrow$ use standardized data

- when the variances of the variables contain important information (for instance if all variables are measured on the same Likert scale) $\Rightarrow$ use mean-corrected data

number of PC's to extract

- extract $PC's$ to explain a given percentage of the variance
- scree plot: plot the eigenvalues in decreasing order and find the elbow that distinguishes the *mountain* from the *debris*
- retain only $PC's$ with eigenvalue larger than one (only for standardized data)
- **Horn's** Parallel procedure: compute eigenvalues associated with many simulated uncorrelated normal variables - retain the ith PC if the corresponding eigenvalue is larger than the 95th (or 5th or 50th) percentile of the distribution of the ith largest eigenvalue of the random data (same idea as the previous rule but taking random variation into account)

example: foodprice data

look for a weighted sum of the various food prices that summarizes how expensive or cheap a city's food items are

RESULTS

Observations	23
Variables	5

Correlation Matrix

	bread	burger	milk	oranges	tomatoes
bread	1.0000	0.6817	0.3282	0.0367	0.3822
burger	0.6817	1.0000	0.3334	0.2109	0.6319
milk	0.3282	0.3334	1.0000	-.0028	0.2544
oranges	0.0367	0.2109	-.0028	1.0000	0.3581
tomatoes	0.3822	0.6319	0.2544	0.3581	1.0000

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.42246795	1.31779306	0.4845	0.4845
2	1.10467489	0.36619436	0.2209	0.7054
3	0.73848053	0.24486921	0.1477	0.8531
4	0.49361132	0.25284601	0.0987	0.9518
5	0.24076530		0.0482	1.0000

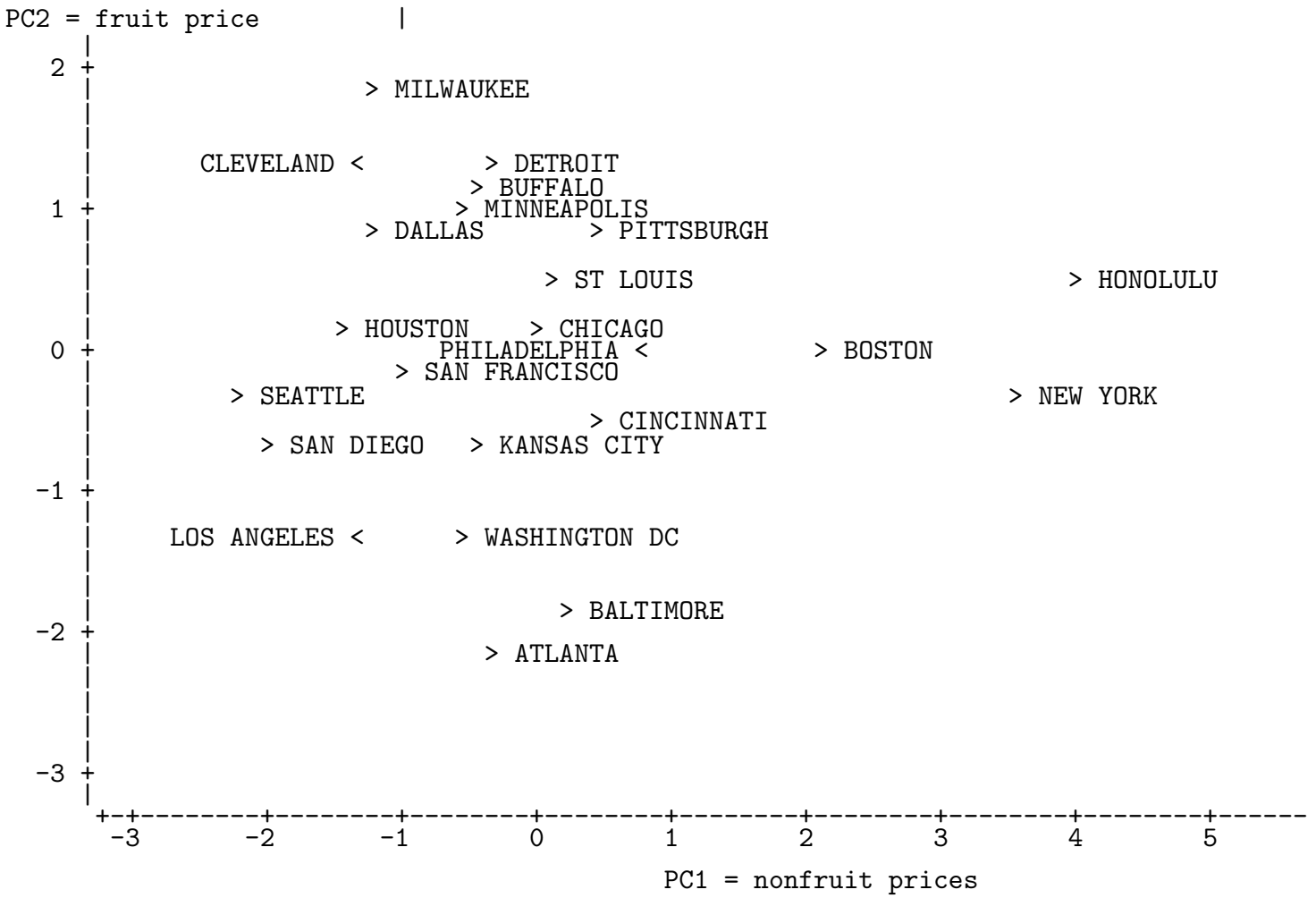
Eigenvectors

	PC1	PC2	PC3	PC4	PC5
bread	0.496149	-.308620	-.386394	0.509305	0.499899
burger	0.575702	-.043802	-.262472	-.028137	-.772635
milk	0.339570	-.430809	0.834640	0.049100	-.007882
oranges	0.224990	0.796777	0.291607	0.479016	0.005967
tomatoes	0.506434	0.287028	-.012266	-.712706	0.391201

Obs	city	bread	burger	milk	orange	tomato	PC1	PC2	PC3	PC4	PC5
1	ATLANTA	24.5	94.5	73.9	80.1	41.6	-0.227	-2.188	0.966	-0.187	-0.819
2	BALTIMORE	26.5	91.0	67.5	74.6	53.3	0.281	-1.883	-0.120	-1.094	0.544
3	BOSTON	29.7	100.8	61.4	104.0	59.6	2.247	-0.073	-1.094	-0.125	0.523
4	BUFFALO	22.8	86.6	65.3	118.4	51.2	-0.341	1.105	1.239	-0.175	0.169
5	CHICAGO	26.7	86.7	62.7	105.9	51.2	0.113	0.086	0.066	0.177	0.934
6	CINCINNATI	25.3	102.5	63.3	99.3	45.6	0.592	-0.451	-0.320	0.141	-1.252
7	CLEVELAND	22.8	88.8	52.4	110.9	46.8	-1.215	1.306	-0.533	-0.114	-0.270
8	DALLAS	23.3	85.5	62.5	117.9	41.8	-1.096	0.840	0.868	0.775	-0.099
9	DETROIT	24.1	93.7	51.5	109.7	52.4	-0.274	1.317	-1.045	-0.440	-0.224
10	HONOLULU	29.3	105.9	80.2	133.2	61.7	4.077	0.493	1.642	0.692	0.021
11	HOUSTON	22.3	83.6	67.8	108.6	42.4	-1.287	0.148	1.534	0.247	-0.083
12	KANSAS CITY	26.1	88.9	65.4	100.9	43.2	-0.317	-0.601	0.317	0.648	0.172
...											
22	SEATTLE	22.5	77.7	62.0	91.1	44.9	-2.091	-0.367	0.649	-0.554	0.687
23	WASHINGTON DC	24.2	93.8	66.0	81.6	46.2	-0.395	-1.400	0.111	-0.681	-0.561

structural loadings	bread	hamburger	milk	oranges	tomatoes
PC1	.772	.896	.529	.350	.788
PC2	-.324	-.046	-.453	.837	.302

plot of PC-scores:

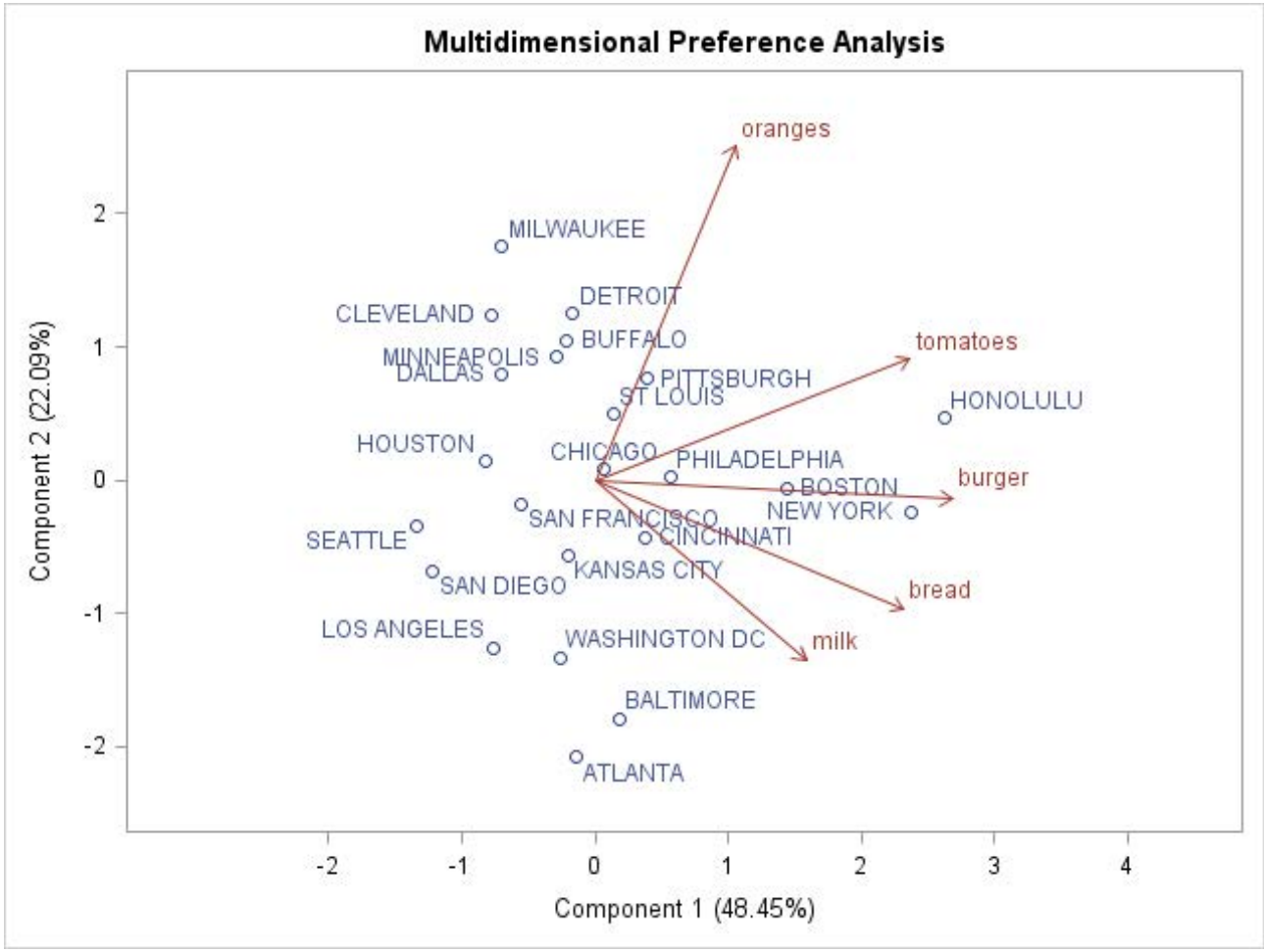


biplots

- in a biplot* the PC scores are plotted as points representing the observations and also the variables are plotted on the same dimensions and represented by vectors
- based on the theory of Singular Value Decomponsitions it can be shown that one can retrieve the original x_{ij} values by projecting the PC scores on the vectors (if the number of principal components is smaller than k, this is only an approximation)

*remark that there exist also biplots that are defined somewhat differently, this here is the default biplot

example:


Multidimensional Preference Analysis
oranges
tomatoes
burger
bread
milk
MILWAUKEE
DETROIT
CLEVELAND
BUFFALO
MINNEAPOLIS
DALLAS
PITTSBURGH
ST LOUIS
HONOLULU
HOUSTON
CHICAGO
PHILADELPHIA
BOSTON
NEW YORK
SAN FRANCISCO
SEATTLE
CINCINNATI
KANSAS CITY
SAN DIEGO
LOS ANGELES
WASHINGTON DC
BALTIMORE
ATLANTA
Component 2 (22.09%)
Component 1 (48.45%)
2
1
0
-1
-2
-2
-1
0
1
2
3
4

remark:

- the projection of the PC-score of Milwaukee on the projected variables indicates that in Milwaukee oranges are quite expensive while the other food items are relatively cheap

- the representation of the 5 dimensions in 2 dimensions is not perfect: compare the price of oranges in Atlanta and Baltimore with the representation in the biplot

4b. EXPLORATORY FACTOR ANALYSIS

EXPLORATORY FACTOR ANALYSIS

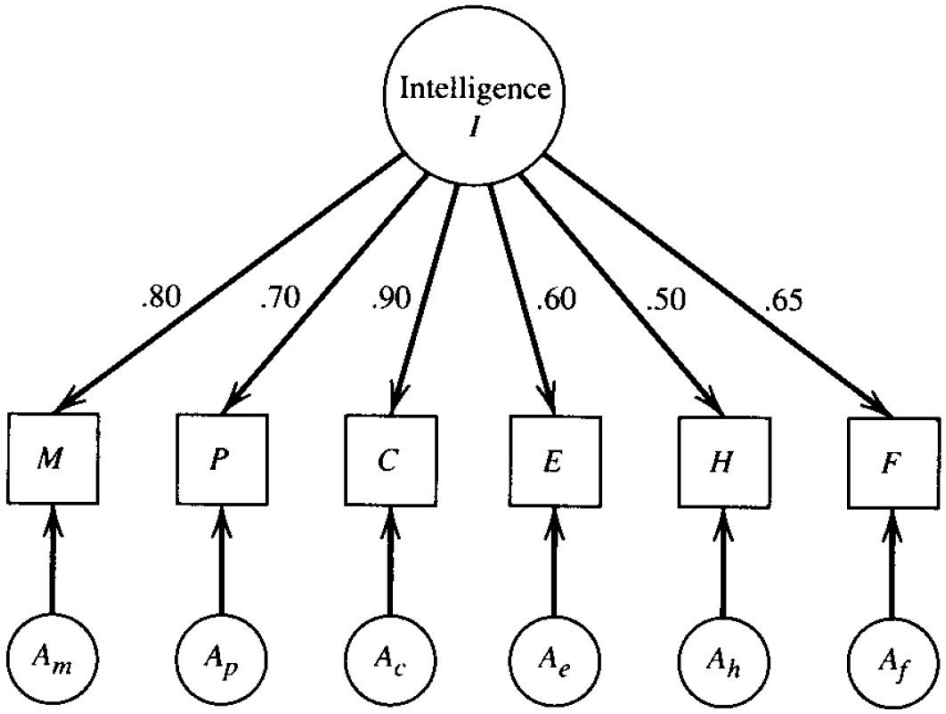
introduction

example of a one factor model

assume the exam scores are available for mathematics (M), physics (P), chemistry (C), english (E), history (H) and french (F)

$\Rightarrow$ try to explain the correlation structure observed

assume for instance that the correlations can be explained by one underlying latent factor (which could then be called Intelligence)



standard notation:

- circles contain latent unobservable variables or factors
- squares contain the observed variables or indicators

$$
\begin{aligned}
M &= 0.80I + A_m & \qquad H &= 0.50I + A_h \\
P &= 0.70I + A_k & \qquad E &= 0.60I + A_e \\
C &= 0.90I + A_c & \qquad F &= 0.65I + A_f
\end{aligned}
$$

- since I cannot be directly observed, it is referred to as a *common/latent factor* or *unobservable construct* ξ
- (0.80, 0.70, . . . 0.65) are the *pattern loadings* or *factor loadings*
- the observed variables M, P, C, ... are *indicators* or *measures* x_i of the construct I
- $A_M, \ldots, A_F$ are the specific factors (denoted as ϵ_i)

then for measure or variable x_i we have in general*

$$x_i = \lambda_i \xi + \epsilon_i \qquad i = 1 \ldots k \qquad \text{with } k \text{ the number of indicators}$$

*remark that λ is not an eigenvalue here

then:

$$\begin{aligned} E(x_i) &= \lambda_i \; E(\xi) + E(\epsilon_i) \\ var(x_i) &= var(\lambda_i \xi + \epsilon_i) = \lambda_i^2 \; var(\xi) + var(\epsilon_i) + 2 \; \lambda_i \; cov(\xi, \epsilon_i) \end{aligned}$$

assumptions:

- we will use standardized variables $\Rightarrow E(x_i) = 0$ and $var(x_i) = 1$
- $E(\xi) = 0$ and therefore also $E(\epsilon_i) = 0$
- $var(\xi) = 1$
- $corr(\epsilon_i, \xi) = 0$
- $corr(\epsilon_i, \epsilon_j) \overset{i \neq j}{=} 0$

given these assumptions, the total variance of any indicator can be decomposed

$$\begin{aligned} \Rightarrow 1 \equiv var(x_i) &= \lambda_i^2 + var(\epsilon_i) \\ &= \text{communality} + \text{specificity} \end{aligned}$$

- the *communality* is the variance in common with I (the explained variance)
- the *specificity or uniqueness* is the variance specific to that indicator (the unexplained or unique variance)

the correlation between any indicators is given by the product of their respective pattern loadings

$$corr(x_i, x_j) \;=\; corr(\lambda_i \xi + \epsilon_i, \lambda_j \xi + \epsilon_j) = \lambda_i \lambda_j$$

this is an important result: it states that the factor should explain the correlation structure completely

the correlation between an indicator x_i and the latent factor is the *structural loading*

$$corr(x_i, \xi) \;=\; corr(\lambda_i \xi + \epsilon_i, \xi) = \lambda_i$$

remark that for a one-factor model, the structure and pattern loadings are always the same

Communalities

Variable	Communality	Error or Unique Variance	Pattern Loading	Structural Loading	Shared Variance
M	.640	.360	.800	.800	.640
P	.490	.510	.700	.700	.490
C	.810	.190	.900	.900	.810
E	.360	.640	.600	.600	.360
H	.250	.750	.500	.500	.250
F	.423	.577	.650	.650	.423
Total	2.973	3.027			2.973

Correlation Matrix for One-Factor Model

	M	P	C	E	H	F
M	1.000					
P	.56	1.000				
C	.72	.63	1.000			
E	.48	.42	.54	1.000		
H	.40	.35	.45	.30	1.000	
F	.52	.46	.59	.39	.33	1.000

remarks:

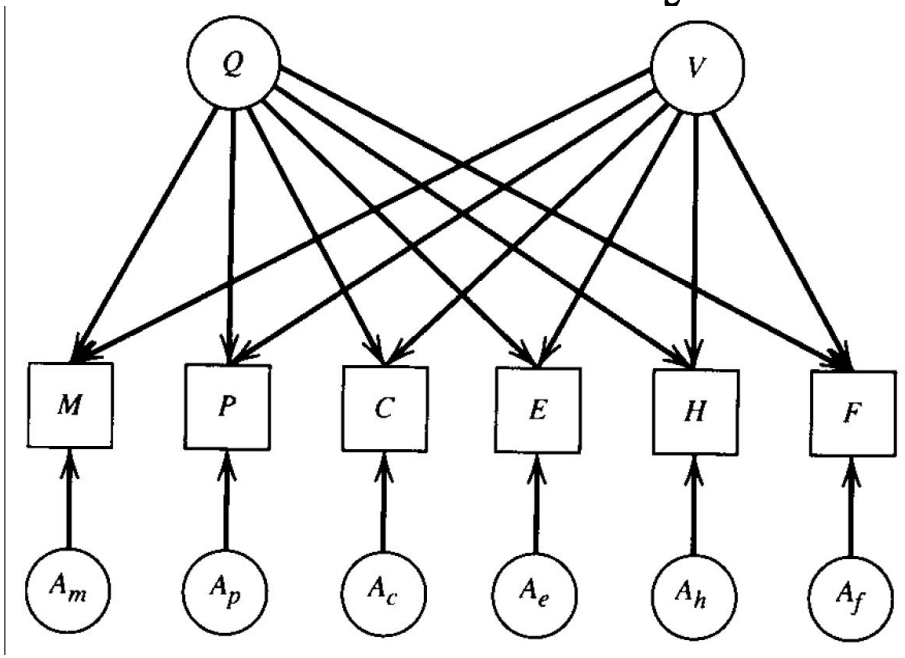
- the last matrix is the *induced or implied* correlation matrix $R^i =$ the correlation matrix as predicted by the model; which should be clearly distinguished from the initial correlation matrix R^{obs} of the raw data; these matrices will differ, unless the model is perfect

- e.g. $corr^i(M, P) = \lambda_M \lambda_P = 0.8 \times 0.7 = 0.56$

- e.g. $var(M) = (0.80)^2 + var(A_m)$

example of a two factor model

students' grades as functions of factors Q and V



$$
\begin{aligned}
M &= 0.800Q + 0.200V + A_m \qquad & H &= 0.150Q + 0.820V + A_h \\
P &= 0.700Q + 0.300V + A_k \qquad & E &= 0.200Q + 0.800V + A_e \\
C &= 0.600Q + 0.300V + A_c \qquad & F &= 0.250Q + 0.850V + A_f
\end{aligned}
$$

in general: $x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + \epsilon_i, \quad i = 1 \ldots k$:

$$
\begin{aligned}
1 = var(x_i) &= \lambda_{i1}^2 + \lambda_{i2}^2 + var(\epsilon_i) + 2\lambda_{i1}\lambda_{i2}\ cov(\xi_1, \xi_2) \\
corr(x_i, x_j) &= \lambda_{i1}\lambda_{j1} + \lambda_{i2}\lambda_{j2} + (\lambda_{i1}\lambda_{j2} + \lambda_{i2}\lambda_{j1}) corr(\xi_1, \xi_2) \\
corr(x_i, \xi_1) &= \lambda_{i1} + \lambda_{i2}\ corr(\xi_1, \xi_2) \\
corr(x_i, \xi_2) &= \lambda_{i2} + \lambda_{i1}\ corr(\xi_1, \xi_2)
\end{aligned}
$$

- with $corr(x_i, \xi_j)$ the *structural loadings* and $\lambda_{i1}, \lambda_{i2}$ the *pattern loadings*[*]
- for an orthogonal model: $cov(\xi_1, \xi_2) = 0 = corr(\xi_1, \xi_2)$

[*] as can be seen from the formulas, the structure and pattern loadings will only be the same if the factors are uncorrelated, i.e. if the factor model is orthogonal

Communalities

	Communalities			
Variable	**Q**	**V**	**Total**	**Unique Variance**
M	.640	.040	.680	.320
P	.490	.090	.580	.420
C	.360	.090	.450	.550
E	.040	.640	.680	.320
H	.023	.672	.695	.305
F	.063	.723	.786	.214
Total	1.616	2.255	3.871	2.129

Pattern and Structure Loadings and Shared Variance

	Pattern Loading		Structure Loading		Shared Variance	
Variable	**Q**	**V**	**Q**	**V**	**Q**	**V**
M	.800	.200	.800	.200	.640	.040
P	.700	.300	.700	.300	.490	.090
C	.600	.300	.600	.300	.360	.090
E	.200	.800	.200	.800	.040	.640
H	.150	.820	.150	.820	.023	.672
F	.250	.850	.250	.850	.063	.723
Total					1.616	2.255

Correlation Matrix

	M	P	C	E	H	F
M	1.000					
P	.620	1.000				
C	.540	.510	1.000			
E	.320	.380	.360	1.000		
H	.284	.351	.336	.686	1.000	
F	.370	.430	.405	.730	.735	1.000

remark:

- the R^i for the two-factor model is different from that for the one factor model; by adding factors, one tries to explain a larger part of the correlations among the indicators
- by examining the communalities of factor loadings, it is possible to interpret the factors
  - communalities of E, H and F with factor V are much greater than those with factor Q
    $\Rightarrow V$ measures the verbal abilities
  - communalities of M, P and C with factor Q are much greater than those with factor V
    $\Rightarrow Q$ measures the quantitative abilities

factor indeterminacy

the factor analysis solution is *not unique*; there are several factor models that are equally good in explaining the correlation structure

consider an alternative two-factor model

$$
\begin{aligned}
M &= 0.667Q - 0.484V + A_m \\
P &= 0.680Q - 0.343V + A_k \\
C &= 0.615Q - 0.267V + A_c
\end{aligned}
\qquad\qquad
\begin{aligned}
H &= 0.725Q + 0.412V + A_h \\
E &= 0.741Q + 0.361V + A_e \\
F &= 0.812Q + 0.355V + A_f
\end{aligned}
$$

Communalities

Variable	Communalities			Unique Variance
	Q	V	Total	
M	.445	.234	.679	.321
P	.462	.118	.580	.420
C	.378	.071	.449	.551
E	.549	.130	.679	.321
H	.526	.170	.696	.304
F	.659	.126	.785	.215
Total	3.019	.849	3.868	2.132

Pattern and Structure Loadings and Shared Variance

Variable	Pattern Loading		Structure Loading		Shared Variance	
	Q	V	Q	V	Q	V
M	.667	−.484	.667	−.484	.445	.234
P	.680	−.343	.680	−.343	.462	.118
C	.615	−.267	.615	−.267	.378	.071
E	.741	.361	.741	.361	.549	.130
H	.725	.412	.725	.412	.526	.170
F	.812	.355	.812	.355	.659	.126
Total					3.019	.849

Correlation Matrix

	M	P	C	E	H	F
M	1.000					
P	.620	1.000				
C	.540	.510	1.000			
E	.320	.380	.360	1.000		
H	.284	.351	.336	.686	1.000	
F	.370	.430	.405	.730	.735	1.000

compare the 2 two-factor models: although the models produce the same R^i, the contribution of each factor in the explained variance of each variable is different

- the factor solution is not unique (∞ many equally good solutions)!
- different interpretation of the factors for each solution:
  here Q stands for a general intelligence factor and V for a latent factor that differentiates between quantitative and verbal ability (so the letters Q and V are not really appropriate here, better to replace them by something like I and Q-V)
- the factor model has been rotated
- look for a factor model that is easy to interpret

exploratory versus confirmatory factor analysis

exploratory factor analysis:

- explain the correlation structure among observed variables
- try to find underlying dimensions that can explain the observed correlations
- example: the correlation between scores on mathematics, statistics and physics exams can be explained because they all measure somehow quantitative intelligence

$\Rightarrow$ no assumptions are made on the number of factors and the relations between factors and observed variables

confirmatory factor analysis:

- how to measure intelligence?
- collect data on some related variables; these should be highly correlated if they really measure the same underlying concept

$\Rightarrow$ a factor model is hypothesized; assumptions are made on the number of factors and the relations between factors and observed variables.

$\Rightarrow$ here the focus is on goodness-of-fit of the hypothesized model

methodology

a general factor model:

$$\begin{cases} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \ldots + \lambda_{1m}\xi_m + \epsilon_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \ldots + \lambda_{2m}\xi_m + \epsilon_2 \\ & \vdots \\ x_k &= \lambda_{k1}\xi_1 + \lambda_{k2}\xi_2 + \ldots + \lambda_{km}\xi_m + \epsilon_k \end{cases}$$

(assume $m << k$)

remark that we look for the λ_{ij}, not for the ξ_j!

$$\underset{(k\times 1)}{\mathbf{x}} = \underset{(k\times m)}{\boldsymbol{\Lambda}} \underset{(m\times 1)}{\boldsymbol{\xi}} + \underset{(k\times 1)}{\boldsymbol{\epsilon}}$$

assuming standardized data and factors + uncorrelated factors:

$var(\xi_i) = 1$, $corr(\xi_i, \xi_j) = 0, corr(\xi_i, \epsilon_j) = 0$, $corr(\epsilon_i, \epsilon_j) \overset{i\neq j}{=} 0$ $(\forall\ i, j)$

$$\Rightarrow 1 = var(x_i) = \sum_l \lambda_{il}^2 + var(\epsilon_i)$$

$$corr(x_i, x_j) \overset{i\neq j}{=} \sum_l \lambda_{il}\lambda_{jl}$$

with

$$\Psi = \begin{pmatrix} var(\epsilon_1) & 0 & \dots & 0 \\ 0 & var(\epsilon_2) & \dots & 0 \\ \dots & & & \\ 0 & 0 & \dots & var(\epsilon_k) \end{pmatrix}$$

we get $R^i = \Lambda\Lambda' + \Psi$ with $\Lambda\Lambda'$ containing the explained variances and induced correlations and Ψ containing the unexplained variances

so, given $R^{obs} \Rightarrow$ estimate Λ by one of the methods such that the implied correlations are close to the observed correlations; once Λ is obtained, Ψ is computed as a diagonal matrix with the unexplained variances on the diagonal and $R^i = \Lambda\Lambda' + \Psi$

example: the methods to obtain a factor model are illustrated with the correlation matrix R^{obs}

$$R^{obs} = \begin{array}{c} \begin{array}{cccccc} M & P & C & E & H & F \end{array} \\ \left(\begin{array}{cccccc} 1.000 & & & & & \\ 0.620 & 1.000 & & & & \\ 0.540 & 0.510 & 1.000 & & & \\ 0.320 & 0.380 & 0.360 & 1.000 & & \\ 0.284 & 0.351 & 0.336 & 0.686 & 1.000 & \\ 0.370 & 0.430 & 0.405 & 0.730 & 0.735 & 1.000 \end{array} \right) \end{array}$$

principal component factoring

compute eigenvalues λ_i* and eigenvectors $\mathbf{e}_i$ of R^{obs}:

$$\begin{aligned}
\lambda_i &= 3.367, 1.194, 0.507, 0.372, 0.313, 0.247 \\
\\
\mathbf{e}_1 &= (0.368, 0.391, 0.372, 0.432, 0.422, 0.456)' \\
\mathbf{e}_2 &= (0.510, 0.409, 0.383, -.375, -.421, -.329)' \\
&\vdots \\
\mathbf{e}_6 &= (0.042, 0.039, 0.024, 0.343, 0.447, -.824)'
\end{aligned}$$

*remark that λ refers here again to eigenvalues!

$$PC_1 = 0.368\ M + 0.391\ P + \ldots + 0.456\ F$$
$$PC_2 = 0.510\ M + 0.409\ P + \ldots - 0.329\ F$$
$$\vdots$$
$$PC_6 = 0.042\ M + 0.039\ P + \ldots - 0.824\ F$$

$\Downarrow$ $\mathbf{V}$(eigenvector-matrix): $\mathbf{V}\mathbf{V}' = \mathbf{I}$ or $\mathbf{V}^{-1} = \mathbf{V}'$

$$M = 0.368\ PC_1 + 0.510\ PC_2 + \ldots + 0.042\ PC_6$$
$$P = 0.391\ PC_1 + 0.409\ PC_2 + \ldots + 0.039\ PC_6$$
$$\vdots$$
$$F = 0.456\ PC_1 - 0.329\ PC_2 + \ldots - 0.824\ PC_6$$

but we assumed $var(\xi_i) = 1$, so divide PC_i by $\sqrt{\lambda_i}$ because:

$$var\left(\frac{PC_i}{\sqrt{\lambda_i}}\right) = \left(\frac{1}{\sqrt{\lambda_i}}\right)^2 \times var(PC_i) = 1$$

$$\begin{aligned}
M &= 0.368\sqrt{\lambda_1}\,\frac{PC_1}{\sqrt{\lambda_1}} + 0.510\sqrt{\lambda_2}\,\frac{PC_2}{\sqrt{\lambda_2}} + \ldots + 0.042\sqrt{\lambda_6}\,\frac{PC_6}{\sqrt{\lambda_6}} \\
P &= 0.391\sqrt{\lambda_1}\,\frac{PC_1}{\sqrt{\lambda_1}} + 0.409\sqrt{\lambda_2}\,\frac{PC_2}{\sqrt{\lambda_2}} + \ldots - 0.039\sqrt{\lambda_6}\,\frac{PC_6}{\sqrt{\lambda_6}} \\
&\vdots \\
F &= 0.456\sqrt{\lambda_1}\,\frac{PC_1}{\sqrt{\lambda_1}} - 0.329\sqrt{\lambda_2}\,\frac{PC_2}{\sqrt{\lambda_2}} + \ldots - 0.824\sqrt{\lambda_6}\,\frac{PC_6}{\sqrt{\lambda_6}}
\end{aligned}$$

let $\xi_i = \frac{PC_i}{\sqrt{\lambda_i}}$

$$
\begin{aligned}
M &= 0.675\ \xi_1 + 0.557\ \xi_2 + \ldots + 0.021\ \xi_6 \\
P &= 0.717\ \xi_1 + 0.447\ \xi_2 + \ldots + 0.019\ \xi_6 \\
&\vdots \\
F &= 0.837\ \xi_1 - 0.359\ \xi_2 + \ldots - 0.409\ \xi_6
\end{aligned}
$$

$\Rightarrow$ with 6 factors one can explain the correlation structure perfectly!

cfr: here $R^i = R^{obs}$, so in this model the communalities for all the variables are equal to one or, equivalently, all the specificities are zero and all correlations are perfectly explained*

*this is no surprise as R^{obs} can completely be reconstructed based on all its eigenvalues and eigenvectors

determine number of $PC's$ to retain*:

- the minimum eigenvalue criterion: consider only the PC's with eigenvalues ≥ 1
- the scree plot
- Horn's Parallel procedure

*remark that it is weird that these criteria do not look at the explained correlation

we retain 2 factors based on the minimum eigenvalue criterion

$\Rightarrow$ our model becomes:

$$\begin{aligned} M &= 0.675\ \xi_1 + 0.557\ \xi_2 + \epsilon_m \\ P &= 0.717\ \xi_1 + 0.447\ \xi_2 + \epsilon_k \\ &\vdots \\ F &= 0.837\ \xi_1 - 0.359\ \xi_2 + \epsilon_f \end{aligned}$$

this factor model explains the correlation between the variables by modeling the variables as the sum of two parts:

- a linear combination of two rescaled $PC's$, the *common factors*
- a sum of remaining error components ϵ, the *unique factors* (remark that $corr(\epsilon_i, \epsilon_j)$ will not be zero but hopefully small)

in matrix notation:

$$\begin{pmatrix} M \\ \vdots \\ F \end{pmatrix} \approx \begin{pmatrix} 0.675 & 0.557 \\ \vdots & \vdots \\ 0.837 & -0.359 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \epsilon_m \\ \vdots \\ \epsilon_f \end{pmatrix}$$

$$= \Lambda^{(1)} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \epsilon_m \\ \vdots \\ \epsilon_f \end{pmatrix}$$

and $R^i = \Lambda^{(1)}\Lambda^{(1)'} + \Psi^{(1)}$

summary of the results:

	factor loadings			specific variance
variable	ξ_1	ξ_2	communalities	ϵ
M	.675	.557	.766	.234
P	.717	.447	.714	.286
C	.683	.418	.641	.359
E	.793	-.410	.797	.203
H	.774	-.461	.812	.188
F	.837	-.359	.829	.171

- the variance explained by ξ_1 is 3.365 ($=.675^2 + .717^2 + \ldots + .837^2 = \lambda_1$)
- the variance explained by ξ_2 is 1.194 ($=.557^2 + .447^2 + \ldots + .359^2 = \lambda_2$)
- total variance is 6 of which 4.459 ($=3.365+1.194$) is explained

(remark that symbol λ_i here is an eigenvalue, not a loading!)

$$\Lambda^{(1)}\Lambda^{(1)'} = \begin{pmatrix} .766 & .733 & .694 & .307 & .266 & .365 \\ .733 & .716 & .677 & .385 & .349 & .440 \\ .694 & .677 & .641 & .370 & .336 & .422 \\ .307 & .385 & .370 & .797 & .803 & .811 \\ .266 & .349 & .336 & .803 & .812 & .813 \\ .365 & .440 & .422 & .811 & .813 & .830 \end{pmatrix}$$

$$R^{obs} - \Lambda^{(1)}\Lambda^{(1)'} = \begin{pmatrix} .234 & -.114 & -.154 & .013 & .018 & .005 \\ -.114 & .284 & -.167 & -.006 & .010 & -.011 \\ -.154 & -.167 & .359 & -.010 & .000 & -.017 \\ .013 & -.006 & -.010 & .203 & -.117 & -.082 \\ .018 & .001 & .000 & -.117 & .188 & -.079 \\ .005 & -.011 & -.017 & -.082 & -.079 & .170 \end{pmatrix}$$

$RMSR = \sqrt{\frac{\sum_{i=1}^{k}\sum_{j>i}^{k} res_{ij}^2}{k(k-1)/2}}$, here: 0.078

iterative improvement

- as Λ only determines the explained variances and the induced correlations, a diagonal matrix Ψ with specificities has to be added to approximate R^{obs}
- replace therefore the diagonal in R^{obs} by estimates of the communalities, yielding a *reduced correlation matrix in which the diagonals are adjusted for the unique factors*
- compute eigenvalues and eigenvectors of this reduced matrix
- use the previous solution to computed estimated communalities in an iterative procedure
- the different methods start with different initial priors, the rest of the iterations proceed in the same way

- **principal axis factoring**
  - use the previous method to obtain: $R^{obs} \approx \Lambda^{(1)}\Lambda^{(1)'} + \Psi^{(1)}$
  - replace the diagonal elements of R^{obs} by the explained variances of this approximation and call the resulting matrix $R^{(1)}$
  - use the previous method to obtain $\Lambda^{(2)}$ based on the eigenvalues and eigenvectors of $R^{(1)}$
  - replace the diagonal elements of R^{obs} by the explained variances of this approximation and call the resulting matrix $R^{(2)}$
  - ... until little change in the communalities

- **image analysis**
  is similar to principal axis factoring but in the first step the ith diagonal element of R^{obs} is replaced by the following estimate of the communality: the Squared Multiple Correlation (SMC) coefficient (R^2) from the regression

$$x_i = \alpha_0 + \alpha_1 x_1 + \ldots \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \ldots \alpha_k x_k + \epsilon$$

  and this for all $i = 1 \ldots k$

  notice that R^2 is the proportion of the variance in x_i predicted by the other variables; this is called the image of the variable x_i, reflected by the other variables

iterative factor improvement, with prior communatilities one

RESULTS

Prior Communality Estimates: ONE

Preliminary Eigenvalues: Total = 6 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	3.36708606	2.17290691	0.5612	0.5612
2	1.19417915	0.68717306	0.1990	0.7602
3	0.50700609	0.13515892	0.0845	0.8447
4	0.37184717	0.05869130	0.0620	0.9067
5	0.31315586	0.06643018	0.0522	0.9589
6	0.24672568		0.0411	1.0000

2 factors will be retained by the MINEIGEN criterion.

The FACTOR Procedure
Initial Factor Method: Iterated Principal Factor Analysis

Iteration	Change	Communalities					
1	0.3594	0.76582	0.71564	0.64062	0.79670	0.81162	0.83086
2	0.1277	0.69838	0.62622	0.51292	0.72426	0.74467	0.78387
3	0.0422	0.67946	0.59762	0.47074	0.69784	0.71918	0.77400
4	0.0135	0.67487	0.58806	0.45722	0.68774	0.70846	0.77438
5	0.0051	0.67443	0.58455	0.45287	0.68358	0.70333	0.77691
6	0.0028	0.67509	0.58304	0.45140	0.68171	0.70053	0.77935
7	0.0019	0.67593	0.58224	0.45084	0.68078	0.69884	0.78122
8	0.0013	0.67670	0.58173	0.45059	0.68028	0.69775	0.78257
9	0.0009	0.67734	0.58136	0.45044	0.67999	0.69702	0.78350

Convergence criterion satisfied.

Eigenvalues of the Reduced Correlation Matrix: Total = 3.86960282 Average = 0.6449338

	Eigenvalue	Difference	Proportion	Cumulative
1	3.02841573	2.18717781	0.7826	0.7826
2	0.84123793	0.83966796	0.2174	1.0000
3	0.00156997	0.00045288	0.0004	1.0004
4	0.00111709	0.00236669	0.0003	1.0007
5	-.00124960	0.00023870	-0.0003	1.0004
6	-.00148830		-0.0004	1.0000

Factor Pattern

	Factor1	Factor2
m	0.63577	0.52263
p	0.65778	0.38559
c	0.59807	0.30457
e	0.76220	-0.31471
h	0.74932	-0.36815
f	0.83151	-0.30345

Variance Explained by Each Factor

Factor1	Factor2
3.0284157	0.8412379

Final Communality Estimates: Total = 3.869654

m	p	c	e	h	f
0.67734414	0.58135862	0.45044494	0.67999290	0.69701543	0.78349763

Residual Correlations With Uniqueness on the Diagonal

	m	p	c	e	h	f
m	0.32266	0.00028	0.00059	-0.00011	0.00001	-0.00006
p	0.00028	0.41864	-0.00084	-0.00001	0.00006	0.00005
c	0.00059	-0.00084	0.54956	0.00001	-0.00002	0.00012
e	-0.00011	-0.00001	0.00001	0.32001	-0.00099	0.00072
h	0.00001	0.00006	-0.00002	-0.00099	0.30298	0.00021
f	-0.00006	0.00005	0.00012	0.00072	0.00021	0.21650

Root Mean Square Off-Diagonal Residuals: Overall = 0.00042558

remark:

- as all indicators have high pattern loadings on factor 1, this factor can be labelled as "general intelligence level"

- as the three quantitative subjects have positive pattern loadings on factor 2 and the three verbal subjects have negative pattern loadings, the second factor discriminates between verbal and quantitative ability

maximum likelihood method

for a multivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{R} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi$:

$$f(\mathbf{x}) = (2\pi)^{-k/2}|\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi|^{-1/2} e^{-\frac{1}{2}\mathbf{x}'(\boldsymbol{\Lambda}\boldsymbol{\Lambda}'+\Psi)^{-1}\mathbf{x}}$$

so the likelihoodfunction is

$$\mathbf{L} = \prod_{i=1}^{n} f(\mathbf{x}_i) = (2\pi)^{-nk/2}|\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi|^{-n/2} e^{-\frac{1}{2}\ \sum_{i=1}^{n} \mathbf{x}_i'\ (\boldsymbol{\Lambda}\boldsymbol{\Lambda}'+\Psi)^{-1}\ \mathbf{x}_i}$$

maximize this likelihood w.r.t. $\boldsymbol{\Lambda}$ (the Ψ matrix again ensures that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \Psi$ has 1's on the diagonal)

this approach has the advantage that one can test the appropriateness of a factor model; for instance the hypotheses

- $H0 : \mathbf{R} = \mathbf{I}$ ($m = 0$ factors) versus $Ha : m > 0$
- $H0 : \mathbf{R} = \mathbf{\Lambda}\mathbf{\Lambda}' + \Psi$ (m factors) versus $Ha : > m$ *necessary*

can be tested with a likelihoodratio test:

$$-2 \ln \frac{\text{likelihood of hypothesized model}}{\text{likelihood with } \mathbf{R} = \mathbf{R}^{obs}} \overset{H_0}{\sim} \chi^2_{\frac{1}{2}[(k-m)^2-k-m]}.$$

Significance Tests

Test	DF	Chi-Square	Pr > ChiSq
H0: No common factors	15	28364.4258	<.0001
HA: At least one common factor			
H0: 2 Factors are sufficient	4	0.0004	1.0000
HA: More factors are needed			

Akaike's Information Criterion	-7.999633
Schwarz's Bayesian Criterion	-36.840995

Factor Pattern

	Factor1	Factor2
m	0.57154	-0.59441
p	0.60852	-0.45794
c	0.55875	-0.37120
e	0.79324	0.22432
h	0.78615	0.27815
f	0.86193	0.20633

Variance Explained by Each Factor

Factor	
Factor1	2.99934825
Factor2	0.87108561

Total Communality: 3.870434

MSA (Measure of Sampling Adequacy):

how small are partial correlations relative to ordinary correlations?

$$MSA = \frac{\sum_{i,j} \text{corr}^2(x_i, x_j)}{\sum \text{corr}^2(x_i, x_j) + \sum \text{partial corr}^2(x_i, x_j)}$$

if Kaiser's MSA > 0.8, the covariance matrix can well be factored

RESULTS

Partial Correlations Controlling all other Variables

	m	p	c	e	h	f
m	1.00000	0.44624	0.30877	0.01370	-0.03204	0.06098
p	0.44624	1.00000	0.20254	0.05115	0.02581	0.09909
c	0.30877	0.20254	1.00000	0.04790	0.03147	0.08633
e	0.01370	0.05115	0.04790	1.00000	0.31717	0.41595
h	-0.03204	0.02581	0.03147	0.31717	1.00000	0.45155
f	0.06098	0.09909	0.08633	0.41595	0.45155	1.00000

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.81296561

m	p	c	e	h	f
0.76886810	0.81244223	0.86692066	0.83196352	0.81209703	0.79666443

partial correlation between x_i and x_j controlling for the other variables or negative anti-image correlations

$=$ correlation between u_i and u_j from:

$$\begin{aligned} x_i &= \alpha_0 + \alpha_1 x_1 + \ldots \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \ldots \alpha_{j-1} x_{j-1} \\ &+ \alpha_{j+1} x_{j+1} + \ldots + \alpha_k x_k + u_i \\ x_j &= \alpha_0 + \alpha_1 x_1 + \ldots \alpha_{i-1} x_{i-1} + \alpha_{i+1} x_{i+1} + \ldots \alpha_{j-1} x_{j-1} \\ &+ \alpha_{j+1} x_{j+1} + \ldots + \alpha_k x_k + u_j \end{aligned}$$

this represents the correlation between x_i and x_j that is not common with the other variables and therefore should be small
(these partial correlations can be computed using only R^{obs})

factor rotation

assume one has a factor model, then R^i is given by:

$$R^i = \Lambda\Lambda' + \Psi$$

for an orthogonal matrix $\mathbf{T}$, $\mathbf{T}\ \mathbf{T}' = \mathbf{I}$, we can rewrite this:

$$\begin{aligned} R^i &= \Lambda\Lambda' + \Psi = \Lambda\ \mathbf{T}\ \mathbf{T}'\ \Lambda' + \Psi \\ &= \Lambda^*\Lambda^{*'} + \Psi \quad \text{with} \quad \Lambda^* = \Lambda\ \mathbf{T} \end{aligned}$$

$$\begin{aligned}\mathbf{x} &= \Lambda\boldsymbol{\xi} + \boldsymbol{\epsilon} = \Lambda\ \mathbf{T}\ \mathbf{T}'\ \boldsymbol{\xi} + \boldsymbol{\epsilon} \\ &= \Lambda^*\boldsymbol{\xi}^* + \boldsymbol{\epsilon} \ \text{ with } \ \boldsymbol{\xi}^* = \mathbf{T}'\ \boldsymbol{\xi}\end{aligned}$$

$\boldsymbol{\xi}^*$ has the same properties as $\boldsymbol{\xi}$; the pattern loadings Λ and Λ^* are equally good in approximating R^{obs}.

$\Rightarrow$ choose $\mathbf{T}$ such that loadings are easy to interpret; we discuss 2 techniques to obtain rotated loadings: quartimax and varimax (we only look at orthogonal rotations, not at oblique rotations).

quartimax rotation (look at rows)

identify factor structure such that all variables have fairly high loadings (in absolute value) on a few factors and have near zero loadings on the other factors

$$Q_i = \frac{\sum_{j=1}^{m} (\lambda_{ij}^{*2} - \bar{\lambda}_{i.}^{*2})^2}{m} \qquad \Rightarrow \max Q = \sum_{i=1}^{k} Q_i$$

(maximize variance of (squared) loadings across factors)

varimax rotation (look at columns)

identify factor structure such that every factor has high loadings on a few variables and low loadings for the other variables

$$V_j = \frac{\sum_{i=1}^{k} (\lambda_{ij}^{*2} - \bar{\lambda}_{.j}^{*2})^2}{k} \quad \Rightarrow \max V = \sum_{j=1}^{m} V_j$$

(maximize variance of (squared) loadings for each factor)

Rotation Method: Varimax
Orthogonal Transformation Matrix

	1	2
1	0.76675	0.64194
2	-0.64194	0.76675

Rotated Factor Pattern

	Factor1	Factor2
m	0.15198	0.80885
p	0.25683	0.71791
c	0.26305	0.61745
e	0.78645	0.24798
h	0.81087	0.19875
f	0.83236	0.30112

Variance Explained by Each Factor

Factor1	Factor2
2.1270972	1.7425564

Final Communality Estimates: Total = 3.869654

m	p	c	e	h	f
0.67734414	0.58135862	0.45044494	0.67999290	0.69701543	0.78349763

remark: now the quantitative courses have high pattern loadings on factor 2 and low loadings on factor 1, whereas the "verbal courses" have high loadings on the first factor and low loadings on factor 2

we conclude that factor 1 can be labelled the verbal ability and factor 2 the quantitative ability (same interpretation for the quartimax rotation in this case)

factor scores

- have the same interpretation as principal component scores: a factor score is the value of the factor for each observation

- if one uses the principal component method or principal axis factoring method, one can use the corresponding rescaled PC-scores as factor scores

- if one uses another method such as maximum likelihood, the factors scores cannot be computed easily and have to be estimated by regression type methods

- one needs the raw data to be able to compute scores

example: detergent data

given is the correlation matrix between the following variables:

```
v1 = "gentle to fabrics",
v2 = "won't harm colors",
v3 = "wont't harm synthetics",
v4 = "safe for lingerie",
v5 = "strong, powerful",
v6 = "gets dirt out",
v7 = "makes colors bright",
v8 = "remove grease stains",
v9 = "good for greasy oil",
v10= "pleasant fragrance",
v11= "removes collar soil",
v12= "removes stubborn stains";
```

part of the output (priors = SMC and quartimax rotation):

RESULTS

```
 Initial Factor Method: Iterated Principal Factor Analysis

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.90231483

    v1       v2       v3       v4       v5       v6
0.8085   0.8493   0.8544   0.8669   0.9086   0.9422

    v7       v8       v9      v10      v11      v12
0.8924   0.9144   0.9401   0.9054   0.8932   0.9432


      Prior Communality Estimates: SMC

   v1       v2       v3       v4       v5       v6
0.4198   0.3996   0.5649   0.5655   0.6050   0.5792

   v7       v8       v9      v10      v11      v12
0.6976   0.7454   0.6658   0.5932   0.7131   0.6444
```

```
Preliminary Eigenvalues: Total = 7.19345032  Average = 0.59945419

             Eigenvalue    Difference    Proportion    Cumulative

       1    5.93553363    4.60259208        0.8251        0.8251
       2    1.33294155    1.07730400        0.1853        1.0104
   ......
      11    -.15038006    0.03442370       -0.0209        1.0257
      12    -.18480376                     -0.0257        1.0000

           2 factors will be retained by the PROPORTION criterion.


Iteration  Change                              Communalities

  1      0.0673  0.44449  0.43474  0.63225  0.61637  0.56941  0.57728  0.66152
                 0.76524  0.68336  0.54781  0.67609  0.65991

...
  7      0.0009  0.43845  0.43928  0.67955  0.63563  0.56072  0.57814  0.65061
                 0.77688  0.69197  0.53824  0.66514  0.66717

             Convergence criterion satisfied.
```

Rotated Factor Pattern

		Factor1	Factor2
v1	gentle to fabrics,	0.21243	0.62715
v2	won't harm colors,	0.22797	0.62234
v3	wont't harm synthetics,	0.35681	0.74313
v4	safe for lingerie,	0.39430	0.69293
v5	strong, powerful,	0.74705	0.05136
v6	gets dirt out,	0.75380	0.09967
v7	makes colors bright,	0.79845	0.11435
v8	remove grease stains,	0.87735	0.08449
v9	good for greasy oil,	0.82787	0.08119
v10	pleasent fragrance,	0.72012	0.14025
v11	removes collar soil,	0.80360	0.13916
v12	removes stubborn stains	0.81228	0.08587

---

the factors can be labelled as follows:

- factor1 = efficacy or ability to clean clothes
- factor2 = mildness quality

4c. DISCRIMINANT ANALYSIS

2 GROUP DISCRIMINANT ANALYSIS

goal of discriminant analysis = goal of logistic regression

is called **supervised learning** in datamining

some examples: use some explanatory variables to predict whether

- client is creditworthy / not creditworthy
- student passes first year at university / does not pass
- car is involved in an accident or not

test for discriminating power

test the multivariate means for equality

$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ versus $H_a : \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$

(assume $\Sigma_1 = \Sigma_2$ and multivariate normal distributions)

$$\frac{n_1 + n_2 - k - 1}{k(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} \, (\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)' S_{pooled}^{-1} (\bar{\mathbf{x}}^1 - \bar{\mathbf{x}}^2)$$

$$\overset{H_0}{\sim} F_{k, n_1+n_2-k-1}$$

this is called **Hotelling's test**

if H_0 can be rejected, there is discriminating power in the variables

example: are the mean experience and mean hp significantly different in the acc and noacc groups?

$$\bar{\mathbf{x}}_{acc} = \begin{pmatrix} 13.60 \\ 184.80 \end{pmatrix} \quad \bar{\mathbf{x}}_{noacc} = \begin{pmatrix} 19.73 \\ 137.18 \end{pmatrix} \quad \mathbf{S}_{pooled} = \begin{pmatrix} 66.978 & -23.750 \\ -23.750 & 381.749 \end{pmatrix}$$

$$\begin{aligned} MD^2 &= (-6.127\ \ 47.618)\mathbf{S}_{pooled}{}^{-1}(-6.127\ \ 47.618)' = 6.093 \\ F &= \frac{(10+11-2-1)}{2(10+11-2)}\frac{10*11}{10+11}6.093 = 15.12 \\ p-value &= P(F_{2;18} \geq 15.12) = 0.0001 \end{aligned}$$

RESULTS

Multivariate Analysis of Variance

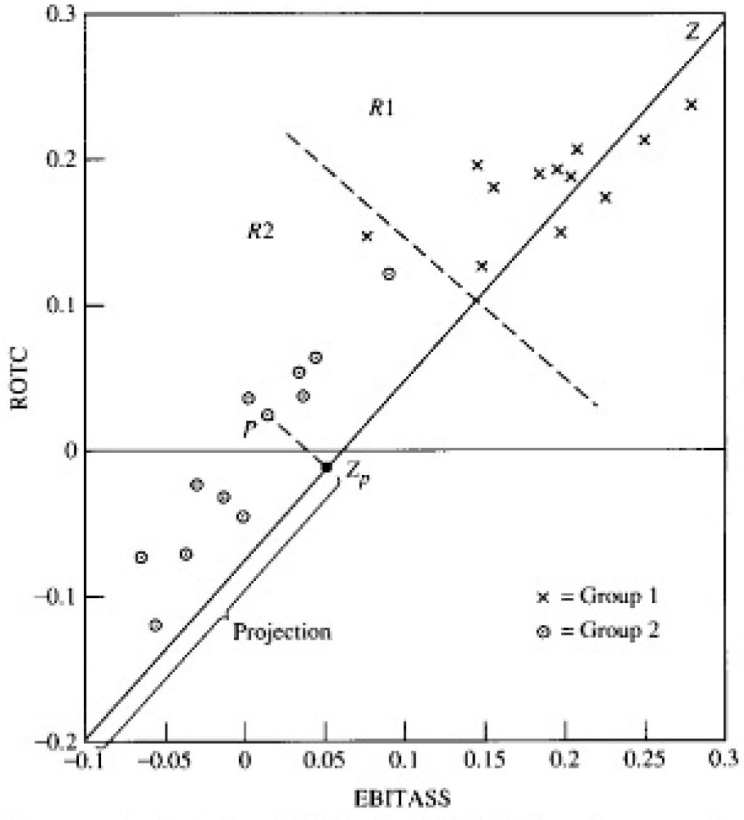
F Value	Num DF	Den DF	Pr > F
15.12	2	18	0.0001

Fisher's approach (descriptive or canonical discriminant analysis)

- compose an index (= a linear combination of the variables, also called the **discriminant function**) that distinguishes the groups
- this index can be used to find out what variables are discriminating
- use the index to predict membership

try to find a linear combination of the k variables, $z = \boldsymbol{\gamma}'\mathbf{x} = \sum_{i=1}^{k} \gamma_i x_i$, such that

- scores of subjects in the same group are very similar
- scores of subjects from different groups are quite different



notations and assumptions:

- assume that the (within) covariancematrices of the groups are equal to $\boldsymbol{\Sigma}$
- the mean score of the subjects of the first group is given by $\boldsymbol{\gamma}$'$\boldsymbol{\mu_1}$ and for the second group $\boldsymbol{\gamma}$'$\boldsymbol{\mu_2}$
- the variance of the scores of one group is given by $\sigma_z^2 = var(\boldsymbol{\gamma}'\mathbf{X}) = \boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}$

so we're looking for a $\boldsymbol{\gamma}$

- that maximizes the difference between $\boldsymbol{\gamma}$'$\boldsymbol{\mu_1}$ and $\boldsymbol{\gamma}$'$\boldsymbol{\mu_2}$
- while minimizing the intragroup variance $\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}$

theorem:

$$\max_{\boldsymbol{\gamma}} \frac{(\boldsymbol{\gamma}'\boldsymbol{\mu_1} - \boldsymbol{\gamma}'\boldsymbol{\mu_2})^2}{\boldsymbol{\gamma}'\boldsymbol{\Sigma}\boldsymbol{\gamma}} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$$

with $\boldsymbol{\gamma} = c\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ and c any constant $\neq 0$

remark that $(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$ is the squared Mahalanobis-distance between $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$

classifying new objects with the Fisher method: let

- $\mathbf{x}_0$ be the values of the k variables for the object to be classified
- $z_0 = \boldsymbol{\gamma}'\mathbf{x}_0 = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0$ the corresponding score

classify the object in the first group if:

- $\boldsymbol{\gamma}'\mathbf{x}_0$ is closer to $\boldsymbol{\gamma}'\boldsymbol{\mu_1}$ than to $\boldsymbol{\gamma}'\boldsymbol{\mu_2}$ in euclidean distance:

$$|\boldsymbol{\gamma}'\mathbf{x}_0 - \boldsymbol{\gamma}'\boldsymbol{\mu_1}| \leq |\boldsymbol{\gamma}'\mathbf{x}_0 - \boldsymbol{\gamma}'\boldsymbol{\mu_2}|$$

- which is equivalent to:

$$\boldsymbol{\gamma}'\mathbf{x}_0 \geq \frac{\boldsymbol{\gamma}'\boldsymbol{\mu_1} + \boldsymbol{\gamma}'\boldsymbol{\mu_2}}{2} \quad = \text{the cutoff value}$$

  remark that as $\boldsymbol{\Sigma}^{-1}$ is positive semidefinite*
  $0 \leq (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2}) = \boldsymbol{\gamma}'\boldsymbol{\mu_1} - \boldsymbol{\gamma}'\boldsymbol{\mu_2}$

- and this is equivalent to:
  classify an object $\mathbf{x}_0$ in the first group if the Mahalanobisdistance between $\mathbf{x}_0$ and $\boldsymbol{\mu_1}$ is smaller than between $\mathbf{x}_0$ and $\boldsymbol{\mu_2}$:

$$(\mathbf{x}_0 - \boldsymbol{\mu_1})'\Sigma^{-1}(\mathbf{x}_0 - \boldsymbol{\mu_1}) \leq (\mathbf{x}_0 - \boldsymbol{\mu_2})'\Sigma^{-1}(\mathbf{x}_0 - \boldsymbol{\mu_2})$$

*remark that a square $(k \times k)$ matrix **A** is positive semidefinite if for all $(k \times 1)$ vectors $\mathbf{x}$: $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$; use this with $\mathbf{A} = \boldsymbol{\Sigma}^{-1}$ and $\mathbf{x} = (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})$

- in practice the $\boldsymbol{\mu_i}$ will be estimated by $\bar{\mathbf{x}}_\mathbf{i}$ and $\boldsymbol{\Sigma}$ by S_{pooled}
- different functions and softwares rescale these scores differently; you can for instance rescale the scores such that the mean score is zero:

$$\hat{\gamma}'\bar{\mathbf{x}} = \hat{\gamma}' \left(\frac{n_1 \bar{\mathbf{x}}_{\cdot}^{1} + n_2 \bar{\mathbf{x}}_{\cdot}^{2}}{n_1 + n_2} \right) = 0$$

to this end the following value is subtracted from the raw scores:

$$\frac{n_1(\hat{\gamma}\bar{\mathbf{x}}_{\cdot}^{1}) + n_2(\hat{\gamma}\bar{\mathbf{x}}_{\cdot}^{2})}{n_1 + n_2} \equiv \frac{n_1 \bar{z}_1 + n_2 \bar{z}_2}{n_1 + n_2}$$

example

- predict whether a person who took a car insurance will be involved in an accident the coming year or not
- explanatory variables: the power of the car (hp = horsepower) and the years of driving experience

RESULTS

```
      Class Level Information
            Variable
accident    Name        Frequency
acc         acc                10
noacc       noacc              11

Raw Canonical Coefficients
hp              0.0493170974
experience      -.0195745442

Structural loadings
hp                   0.995231
experience          -0.462100
```

```
    Squared
  Canonical
Correlation
   0.626821

 Eigenvalue
     1.6797
```

- the structural loading $= corr(x_i, \hat{\gamma}'\bar{\mathbf{x}}) = corr(x_i, z)$

- the squared canonical correlation gives the discriminating power*:
  $CR^2 = \frac{SS_b(z)}{SS_t(z)} = \frac{\sum_{g=1}^{2} n_g(\bar{z}^g - \bar{z})^2}{\sum_{g=1}^{2}\sum_{i=1}^{n_g}(z_i^g - \bar{z})^2} = 0.626821$
  (the closer to 1, the better)

- as does the *eigenvalue* $= \frac{SS_b(z)}{SS_w(z)} = \frac{SS_b(z)}{SS_t(z) - SS_b(z)}$ =1.6797
  (the larger, the better)

*see the chapter on cluster analysis for the definition of *total, between and within SS*

Class Means on Canonical Variables

accident	Can1
acc	1.292934281
noacc	-1.175394801

Obs	experience	hp	accident	Can1
1	25	105	noacc	-2.86572
2	23	108	noacc	-2.67862
3	27	120	noacc	-2.16511
4	15	129	noacc	-1.48636
5	28	138	noacc	-1.29698
6	2	128	noacc	-1.28121
7	18	142	noacc	-0.90397
8	8	142	noacc	-0.70822
9	23	153	noacc	-0.45935
10	32	160	noacc	-0.29030
11	23	160	acc	-0.11413
12	18	170	acc	0.47691
13	8	174	acc	0.86993
14	15	181	acc	1.07812
15	12	181	acc	1.13685
16	16	184	noacc	1.20650
17	4	186	acc	1.54003
18	22	194	acc	1.58223
19	18	193	acc	1.61121
20	14	204	acc	2.23199
21	2	205	acc	2.51620

these class means and discriminant scores can be retrieved as follows:

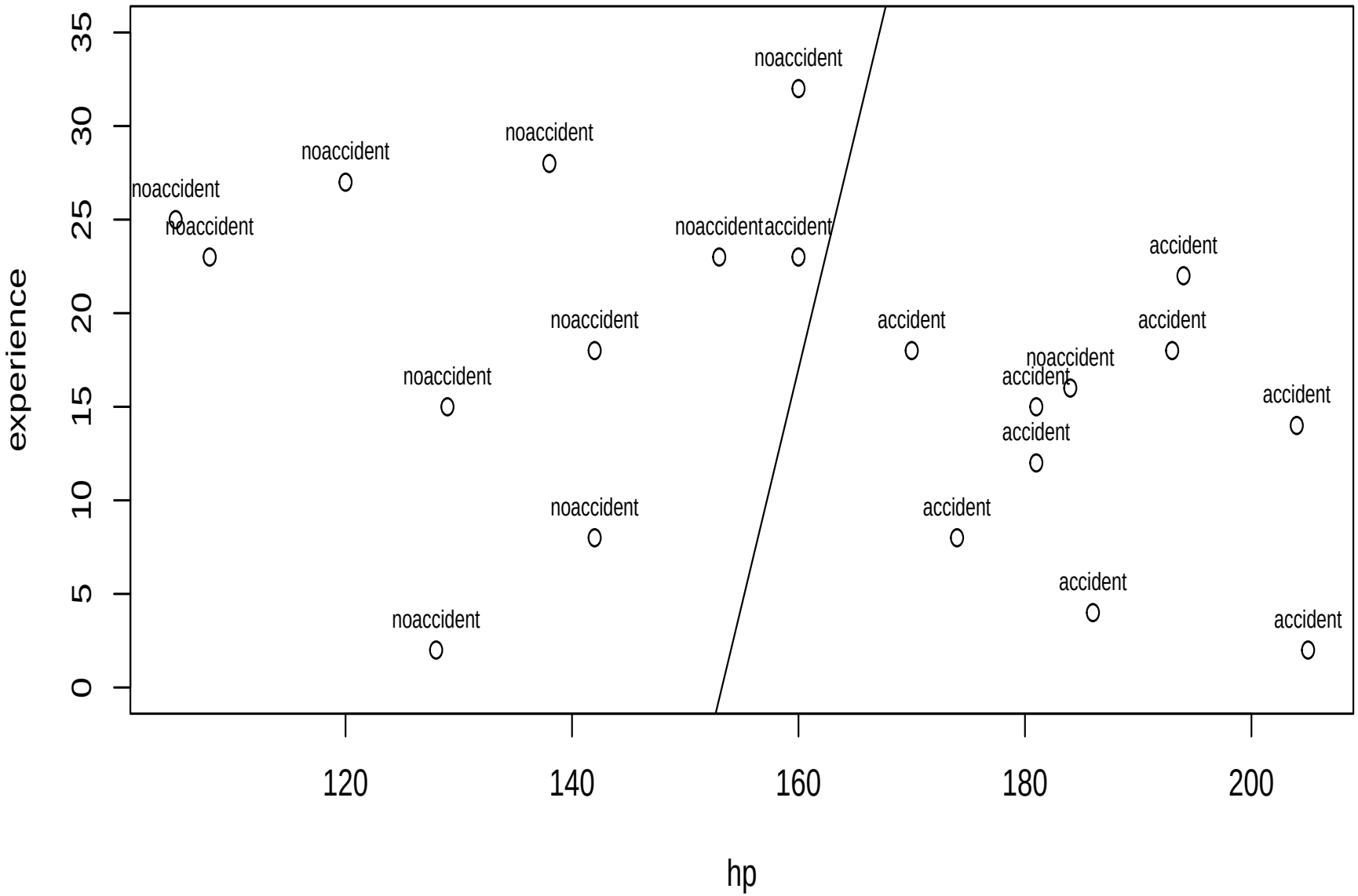
$$\bar{z}_{acc} = (0,0493 \quad -0,0196) \begin{pmatrix} 184,80 \\ 13,600 \end{pmatrix} \qquad \bar{z}_{noacc} = (0,0493 \quad -0,0196) \begin{pmatrix} 137,18 \\ 19,727 \end{pmatrix}$$
$$= 8,8476 \qquad = 6,3793$$

the scores are rescaled by the overall mean $\frac{n_1\bar{z}_1+n_2\bar{z}_2}{n_1+n_2} = 7,5547$

so the rescaled class means are

$$\begin{aligned} \bar{z}_{acc} &= 8,8476 - 7,5547 = 1,2929 \\ \bar{z}_{noacc} &= 6,3793 - 7,5547 = -1,1754 \end{aligned}$$

the following plot shows the points labelled with the variable accident and the points for with $0.0493 * hp - 0.0196 * experience = 7.55457$


experience
hp
0
5
10
15
20
25
30
35
120
140
160
180
200
noaccident
accident

the misclassification approach

2 possible errors and associated costs:

belonging to	classified into group 1	classified into group 2
group 1	0	$c(2\|1)$
group 2	$c(1\|2)$	0

look for some rule that minimizes *the expected misclassification cost* taking into account the prior probabilities of belonging to a group

- $f_i(\mathbf{x})$ is the density function of $\mathbf{X}$ in group i
- p_1 and p_2: prior probabilities (relative size of each population)
- $c(i|j)$: misclassification cost if object of group j is classified in group i
- $\mathbf{x}_0$: the object to be classified

the expected misclassification cost when classifying the object $\mathbf{x}_0$ in group 1 is*:

$$
\begin{aligned}
&0\ \mathcal{P}(object \in group\ 1\ |\mathbf{x}_0) + c(1|2)\mathcal{P}(object \in group\ 2\ |\mathbf{x}_0) \\
&\quad = c(1|2)\frac{\mathcal{P}(\mathbf{x}_0|object\ \in group\ 2\ )\ p_2}{\mathcal{P}(\mathbf{x}_0|object\ \in group\ 1\ )\ p_1 + \mathcal{P}(\mathbf{x}_0|object\ \in group\ 2\ )\ p_2} \\
&\quad = c(1|2)\frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}
\end{aligned}
$$

the expected misclassification cost classifying the object in group 2

$$
c(2|1)\frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}
$$

*Bayes' rule: $\mathcal{P}(H_k|A) = \frac{\mathcal{P}(A|H_k)\mathcal{P}(H_k)}{\mathcal{P}(A|H_1)\mathcal{P}(H_1)+\mathcal{P}(A|H_2)\mathcal{P}(H_2)}$ with H_1 and H_2 a partition

$\Rightarrow$ classify the object $\mathbf{x}_0$ in group 1 if

$$c(1|2)p_2 f_2(\mathbf{x}_0) \leq c(2|1)p_1 f_1(\mathbf{x}_0) \quad \text{or} \quad \boxed{\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} \geq \frac{c(1|2)p_2}{c(2|1)p_1}}$$

- so depending on the priors and misclassification costs, one obtains different classification rules
- only the value $\frac{c(1|2)p_2}{c(2|1)p_1}$ is important not the individual values

deal with classification costs in practice: as most software cannot deal with costs, we can use the following trick if there are only 2 groups:

let $c(2|1) = 50$, $c(1|2) = 100$, $p_1 = 40\%$ and $p_2 = 60\%$,

so $c(2|1)p_1 = 50 * 0.40 = 20$ and $c(1|2)p_2 = 100 * 0.60 = 60$

we rescale them to sum to 1:

$p_1^* = \frac{20}{20+60} = \frac{20}{80} = 25\%$ and $p_2^* = \frac{60}{20+60} = \frac{60}{80} = 75\%$

if we use these priors we get the same classification as if we used the original priors and costs; so from now on we can neglect the misclassification costs, assuming they are equal or have been put in the prior values

- the **posterior probability** of belonging to group 1, given the value $\mathbf{x}_0$ is $\mathcal{P}(object\ belongs\ to\ first\ group|\mathbf{x}_0)$ which is

$$\frac{p_1 f_1(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}$$

the **posterior probability** of belonging to group 2, given the value $\mathbf{x}_0$ is

$$\frac{p_2 f_2(\mathbf{x}_0)}{p_1 f_1(\mathbf{x}_0) + p_2 f_2(\mathbf{x}_0)}$$

so classify where the posterior probability is largest

special case: $f_i(\mathbf{x})$ **multivariate normal distribution** $N(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i})$

- $\boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2} = \boldsymbol{\Sigma} \Rightarrow$ classify object $\mathbf{x}_0$ in group 1 if:

$$(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}\mathbf{x}_0 \geq \frac{1}{2}(\boldsymbol{\mu_1} - \boldsymbol{\mu_2})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu_1} + \boldsymbol{\mu_2}) + ln\left(\frac{p_2}{p_1}\right)$$

so Fisher minimizes the expected misclassification cost if:

1. populations are multivariate normally distributed
2. populations have the same covariancematrix
3. prior probabilities are equal

- $\boldsymbol{\Sigma_1 \neq \Sigma_2}$: $\Rightarrow$ classify object $\mathbf{x}_0$ in group 1 if:

$$-\frac{1}{2}\mathbf{x}_0'(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\mathbf{x}_0 + (\boldsymbol{\mu_1}'\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu_2}\boldsymbol{\Sigma}_2^{-1})\mathbf{x}_0 - \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right)$$
$$-\frac{1}{2}\left(\boldsymbol{\mu_1}'\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu_1} - \boldsymbol{\mu_2}'\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu_2}\right) \geq \ln\left(\frac{p_2}{p_1}\right)$$

$=$ quadratic function in $\mathbf{x}_0$

evaluation of classification rules

compute the % of misclassified observations by

- **resubstitution**

  compute the confusion- or classificationmatrix:

		classified into	
		group 1	group 2
belonging	group 1	n_{1C}	n_{1F}
to	group 2	n_{2F}	n_{2C}

  % of misclassifications $= \frac{n_{1F}+n_{2F}}{n_1+n_2}$

- advantage: simple measure
- disadvantage: underestimates the true misclassification rate because the same sample is used to determine the classification rule and to evaluate the rule

- **holdout method**: better to split the sample in two:
  - one part is used to derive the classification rule (= training sample)
  - other part is used to evaluate rule (= validation sample)

  and compute the % of misclassifications in the validation set

- **leave-one-out crossvalidation**

  1. derive classificationrule with $n_1 + n_2 - 1$ observations
  2. classify the deleted observation by this rule
  3. repeat step 1 and 2 for all $n_1 + n_2$ objects
  4. count the number of misclassifications n_{1F}^U and n_{2F}^U

  compute the % of misclassifications

example: accident data using equal priors, equal covariance matrices and multivariate normal distributions:

RESULTS

Total Sample Size 21 Variables 2

Class Level Information

accident	Variable Name	Frequency	Prior Probability
acc	acc	10	0.500000
noacc	noacc	11	0.500000

Classification Results

Resubstitution Results using Linear Discriminant Function

From accident	acc	noacc	Total
acc	9	1	10
noacc	1	10	11

Cross-validation Results using Linear Discriminant Function

From accident	acc	noacc	Total
acc	9	1	10
noacc	2	9	11

using $p_1 = 0.10$ and $p_2 = 0.90$ we get:

RESULTS

Total Sample Size 21 Variables 2

Class Level Information

accident	Variable Name	Frequency	Prior Probability
acc	acc	10	0.100000
noacc	noacc	11	0.900000

Classification Results

Resubstitution Results using Linear Discriminant Function

From accident	acc	noacc	Total
acc	7	3	10
noacc	1	10	11

test the equality of the covariance matrices:

test $H_0 : \boldsymbol{\Sigma_1} = \boldsymbol{\Sigma_2}$ by one of the many tests that are available

the Box M test for instance yields:

RESULTS

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
3.069399	3	0.3810

Since the Chi-Square value is not significant at the 0.1 level, a pooled covariance matrix will be used in the discriminant function.

as the covariance matrices are not significantly different, the quadratic classfication rule is not expected to yield better results than the linear discriminant function

4d. CLUSTER ANALYSIS

CLUSTER ANALYSIS

combine observations in groups (clusters) such that

- observations in a group are similar (w.r.t. some criterion)
- observations in different groups are different from each other

example: a marketing manager is interested in identifying similar clients to send more personalized leaflets

is called **unsupervised learning** in datamining

hierarchical versus nonhierarchical methods

an **agglomerative hierarchical algorithm** starts with n clusters (each observation is one cluster), combines the two most similar clusters, combines the next two most similar clusters and so on; (different methods because of different distance measures between clusters); a **divisive hierarchical algorithm** goes from 1 to n clusters

a **nonhierarchical algorithm** or *partitioning algorithm* chooses K initial clusters and reassigns observations until no improvement can be obtained; the results depend highly on the choice of initial centroids

distance and similarity measures

- **Euclidean distance** between observation i and j is

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ik} - x_{jk})^2}$$

the Euclidean distance is not scale invariant, cfr. if one of the variables has a much larger variance than the others it will dominate the distance measure; if this is not appropriate, use:

$$d_s(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\frac{(x_{i1} - x_{j1})^2}{s_1^2} + \ldots + \frac{(x_{ik} - x_{jk})^2}{s_k^2}}$$

$=$ euclidean distance for standardized data

- **association measures** for binary data

		W	1	0
V	1		a	b
	0		c	d

with $a + b + c + d = p$

where a: the number of attributes that V and W have in common
b: the number of attributes of V that are not in W
c: the number of attributes of W that are not in V
d: the number of attributes they are both lacking

the euclidean distance is not very useful here!

possible measures (extra useful with asymmetric information: having a characteristic is more informative than not having it):

$$
\begin{aligned}
s_1(V,W) &= \frac{a}{a+b+c} \quad \text{(Jaccard similarity coefficient)} \\
s_2(V,W) &= \frac{2(a+d)}{2(a+d)+b+c} \\
s_3(V,W) &= \frac{2a}{2a+b+c} \\
s_4(V,W) &= \frac{a}{k}
\end{aligned}
$$

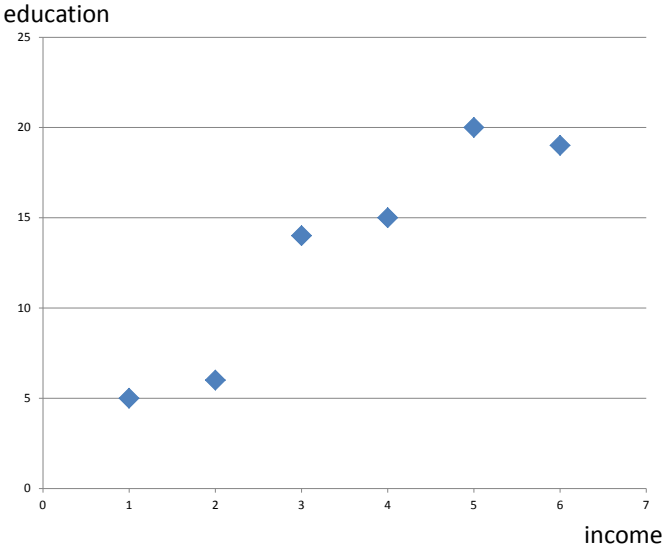
transform similarity measures into dissimilarities, f.i. by

$$d(v,w) = 1 - s(V,W) \text{ or } d(v,w) = \frac{1}{s(V,W)}$$

hierarchical clustering

we illustrate all methods with these data*:

subject	income	education
s1	5	5
s2	6	6
s3	15	14
s4	16	15
s5	25	20
s6	30	19



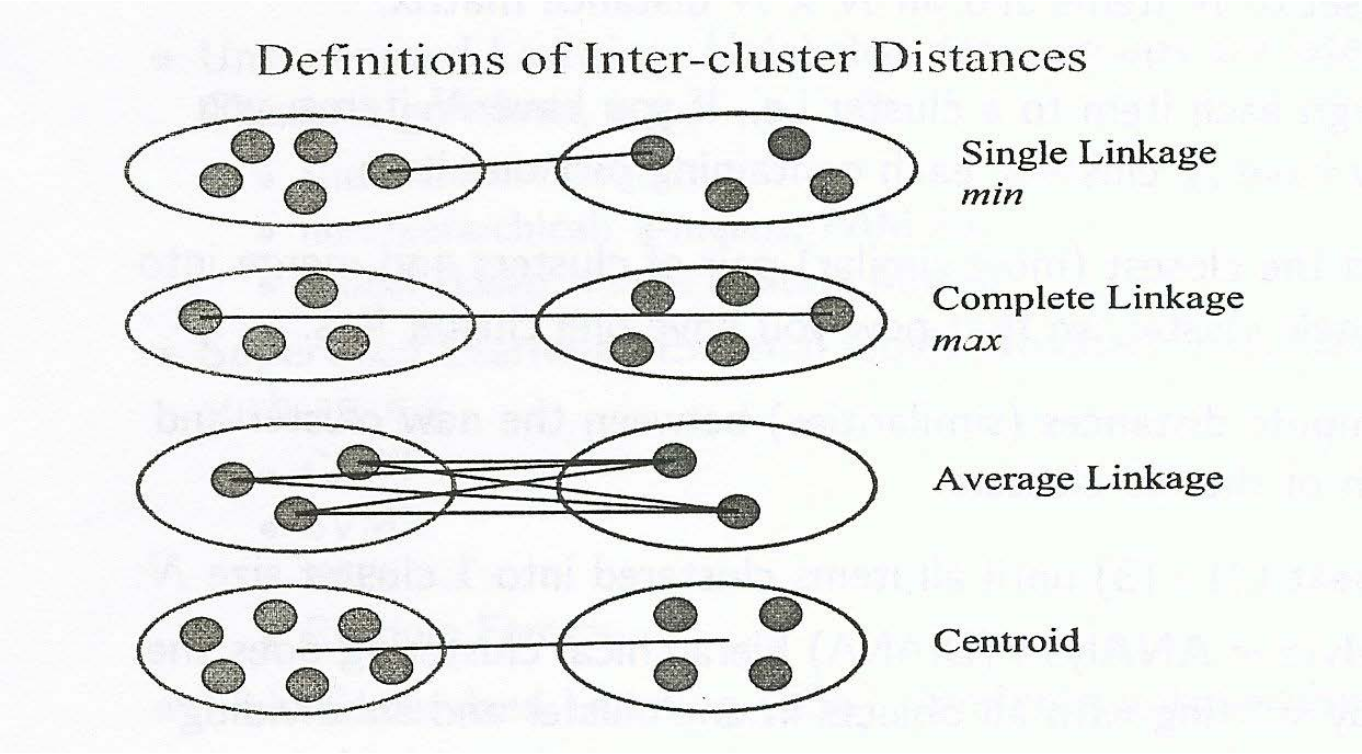
*as mentioned earlier, one should almost always standardize the variables, it is not done here because it is easier to manually recompute the results if the variables are not standardized

$$d^2(s1, s2) = (5-6)^2 + (5-6)^2 = 2$$
$$d^2(s2, s4) = (6-16)^2 + (6-15)^2 = 181$$

the **squared** Euclidean distances:

	S1	S2	S3	S4	S5	S6
S1	0.00	2.00	181.00	221.00	625.00	821.00
S2	2.00	0.00	145.00	181.00	557.00	745.00
S3	181.00	145.00	0.00	2.00	136.00	250.00
S4	221.00	181.00	2.00	0.00	106.00	212.00
S5	625.00	557.00	136.00	106.00	0.00	26.00
S6	821.00	745.00	250.00	212.00	26.00	0.00

all methods start by combining the two observations that are closest, in this case s1 and s2 or s3 and s4 (one can choose which one); the following steps depend on the way the distance between clusters is defined



the centroid method

- each cluster is replaced by the *average observation* which is called the **centroid** of that group

- the distance between 2 clusters is then defined as the distance between the 2 centroids

the subsequent steps of the algorithm are:

Data for Five Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3	15.0	14.0
3	S4	16.0	15.0
4	S5	25.0	20.0
5	S6	30.0	19.0

the squared Euclidean distances are now:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	162.50	200.50	590.50	782.50
S3	162.50	0.00	2.00	135.96	250.00
S4	200.50	2.00	0.00	106.00	212.00
S5	590.50	135.96	106.00	0.00	26.00
S6	782.50	250.00	212.00	26.00	0.00

$$\begin{aligned} d^2(C1, s5) &= (5.5 - 25)^2 \\ &+ (5.5 - 20)^2 \\ &= 590.50 \end{aligned}$$

Data for Four Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3&S4	15.5	14.5
3	S5	25.0	20.0
4	S6	30.0	19.0

the squared Euclidean distances are now:

	S1&S2	S3&S4	S5	S6
S1&S2	0.00	181.00	590.50	782.50
S3&S4	181.00	0.00	120.50	230.50
S5	590.50	120.50	0.00	26.00
S6	782.50	230.50	26.00	0.00

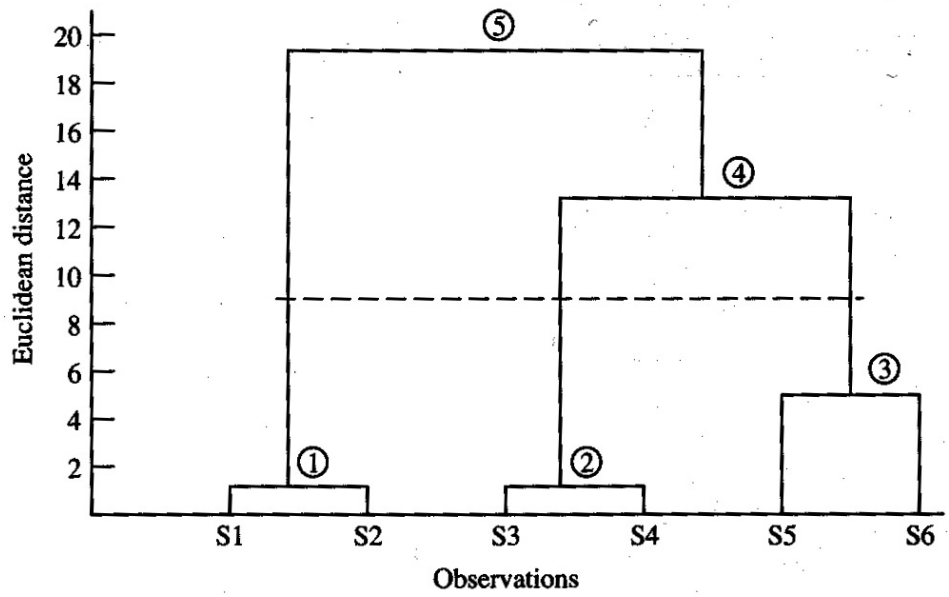
$$\begin{aligned} d^2(C1, C2) &= (5.5 - 15.5)^2 \\ &+ (5.5 - 14.5)^2 \\ &= 181 \end{aligned}$$

Data for Three Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3&S4	15.5	14.5
3	S5&S6	27.5	19.5

the squared Euclidean distances:

	S1&S2	S3&S4	S5&S6
S1&S2	0.00	181.00	680.00
S3&S4	181.00	0.00	169.00
S5&S6	680.00	169.00	0.00



the different steps are visualized in a dendrogram or tree (the cluster distances are represented on the vertical axis)

the single linkage or nearest neighbor method

distance between the clusters $C1$ and $C2$ is defined as:

$$d^2(C1, C2) = \min_{s_i \in C1, s_j \in C_2} d^2(s_i, s_j)$$

and can be computed easily by the following relation: suppose the cluster $C2$ was created by combining the clusters U and V:

$$d^2(C1, C2) = \min(d^2(C1, U), d^2(C1, V))$$

this method can suffer from *chaining* leading to elongated clusters

after the first step of the algorithm:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	145.00	181.00	557.00	745.00
S3	145.00	0.00	2.00	136.00	250.00
S4	181.00	2.00	0.00	106.00	212.00
S5	557.00	136.00	106.00	0.00	26.00
S6	745.00	250.00	212.00	26.00	0.00

$$\begin{aligned} d^2(C1, s5) &= \min(d^2(s1, s5); d^2(s2, s5)) \\ &= min(625; 557) = 557 \end{aligned}$$

second step: combine S3&S4 ...

the complete linkage or furthest neighbor method

distance between the clusters $C1$ and $C2$ is defined as:

$$d^2(C1, C2) = \max_{s_i \in C1, s_j \in C_2} d^2(s_i, s_j)$$

and can be computed easily by the following relation: suppose the cluster $C2$ was created by combining the clusters U and V:

$$d^2(C1, C2) = \max(d^2(C1, U), d^2(C1, V))$$

after the first step of the algorithm:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	181.00	221.00	625.00	821.00
S3	181.00	0.00	2.00	136.00	250.00
S4	221.00	2.00	0.00	106.00	212.00
S5	625.00	136.00	106.00	0.00	26.00
S6	821.00	250.00	212.00	26.00	0.00

$$\begin{aligned} d^2(C1, s6) &= \max(d^2(s1, s6); d^2(s2, s6)) \\ &= \max(821; 745) = 821 \end{aligned}$$

the average linkage method

the distance between the clusters $C1$ and $C2$:

$$d^2(C1, C2) = \frac{\sum_{s_i \in C1} \sum_{s_j \in C2} d^2(s_i, s_j)}{\text{number of obs in C1} \times \text{number of obs in C2}}$$

first step of algorithm:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	163.00	201.00	591.00	783.00
S3	163.00	0.00	2.00	136.00	250.00
S4	201.00	2.00	0.00	106.00	212.00
S5	591.00	136.00	106.00	0.00	26.00
S6	783.00	250.00	212.00	26.00	0.00

$$d^2(C1, s3) = \frac{d^2(s1, s3) + d^2(s2, s3)}{2} = \frac{181 + 145}{2} = 163$$

for the Ward method we need to introduce some notation:

	X_1	X_2	$\ldots$	X_k
$cluster\ 1:$	x_{11}^1	x_{12}^1	$\ldots$	x_{1k}^1
	x_{21}^1	x_{22}^1	$\ldots$	x_{2k}^1
		$\vdots$		
	$x_{n_1 1}^1$	$x_{n_1 2}^1$	$\ldots$	$x_{n_1 k}^1$
average	$\bar{x}_1^1$	$\bar{x}_2^1$	$\ldots$	$\bar{x}_k^1$
$cluster\ 2:$	x_{11}^2	x_{12}^2	$\ldots$	x_{1k}^2
		$\vdots$		
	$x_{n_2 1}^2$	$x_{n_2 2}^2$	$\ldots$	$x_{n_2 k}^2$
average	$\bar{x}_1^2$	$\bar{x}_2^2$	$\ldots$	$\bar{x}_k^2$
		$\vdots$		
$cluster\ G:$	x_{11}^G	x_{12}^G	$\ldots$	x_{1k}^G
		$\vdots$		
	$x_{n_G 1}^G$	$x_{n_G 2}^G$	$\ldots$	$x_{n_G k}^G$
average	$\bar{x}_1^G$	$\bar{x}_2^G$	$\ldots$	$\bar{x}_k^G$
total average	$\bar{x}_1$	$\bar{x}_2$	$\ldots$	$\bar{x}_k$

$$SS_t = \sum_{j=1}^{k}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ij}^g - \bar{x}_j)^2$$

$$(= \text{independent of cluster solution!})$$

$$SS_b = \sum_{j=1}^{k}\sum_{g=1}^{G} n_g(\bar{x}_j^g - \bar{x}_j)^2$$

$$SS_w = SS_t - SS_b = \sum_{j=1}^{k}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ij}^g - \bar{x}_j^g)^2$$

with G the number of clusters, k the number of variables and n_g number of observation in cluster g

note that $SS_t = SS_w + SS_b$ and that in the first step $SS_w = 0$ and $SS_b = SS_t$ and in the last step $SSw = SS_t$ and $SS_b = 0$

Ward's error or trace method

based on the **loss in information** (or the increase in SS_w or the increase in *merging cost*) that is incurred by replacing observations by their cluster centroid:

$$SS_w = ESS = \sum_{g=1}^{G} \sum_{j=1}^{k} \sum_{i=1}^{n_g} (x_{ij}^g - \bar{x}_j^g)^2$$

for G clusters with n_g observations in the g-th cluster and x_{ij}^g the value of the j-th variable of the i-th observation in the g-th cluster

combine observations or clusters that minimize the increase in ESS:

s1,s2:

$$(5-5.5)^2+(5-5.5)^2$$

$$+(6-5.5)^2+(6-5.5)^2=1$$

s2,s5:

$$(6-15.5)^2+(6-13)^2$$

$$+(25-15.5)^2+(20-13)^2=278.5$$

s1,s2,s6:

$$(5-13.66)^2+(5-10)^2$$

$$+(6-13.66)^2+(6-10)^2$$

$$+(30-13.66)^2+(19-10)^2=522.66$$

	Members in Cluster					
Cluster Solution	**1**	**2**	**3**	**4**	**5**	**ESS**
(a) All Possible Five-Cluster Solutions						
1	S1,S2	S3	S4	S5	S6	1.0
2	S1,S3	S2	S4	S5	S6	90.5
3	S1,S4	S2	S3	S5	S6	110.5
4	S1,S5	S2	S3	S4	S6	312.5
5	S1,S6	S2	S3	S4	S5	410.5
6	S2,S3	S1	S4	S5	S6	72.5
7	S2,S4	S1	S3	S5	S6	90.5
8	S2,S5	S1	S3	S4	S6	278.5
9	S2,S6	S1	S3	S4	S5	372.5
10	S3,S4	S1	S2	S5	S6	1.0
11	S3,S5	S1	S2	S4	S6	68.0
12	S3,S6	S1	S2	S4	S5	125.0
13	S4,S5	S1	S2	S3	S6	53.0
14	S4,S6	S1	S2	S3	S5	106.0
15	S5,S6	S1	S2	S3	S4	13.0
(b) All Possible Four-Cluster Solutions						
1	S1,S2,S3	S4	S5	S6		109.333
2	S1,S2,S4	S3	S5	S6		134.667
3	S1,S2,S5	S3	S4	S6		394.667
4	S1,S2,S6	S3	S4	S5		522.667
5	S1,S2	S3,S4	S5	S6		2.000
6	S1,S2	S3,S5	S4	S6		69.000
7	S1,S2	S3,S6	S4	S5		126.000
8	S1,S2	S4,S5	S3	S6		54.000
9	S1,S2	S4,S6	S3	S5		107.000
10	S1,S2	S5,S6	S3	S4		14.000

RESULTS

Centroid Hierarchical Cluster Analysis
Root-Mean-Square Total-Sample Standard Deviation = 8.376555

Cluster History

NCL	--Clusters Joined--		FREQ	RMS STD	SPRSQ	RSQ	Cent Dist	Tie
5	s1	s2	2	0.7071	0.0014	.999	1.4142	T
4	s3	s4	2	0.7071	0.0014	.997	1.4142	
3	s5	s6	2	2.5495	0.0185	.979	5.099	
2	CL4	CL3	4	5.5227	0.2409	.738	13	
1	CL5	CL2	6	8.3766	0.7378	.000	19.704	

3-cluster solution

---------- CLUSTER=1 ----------

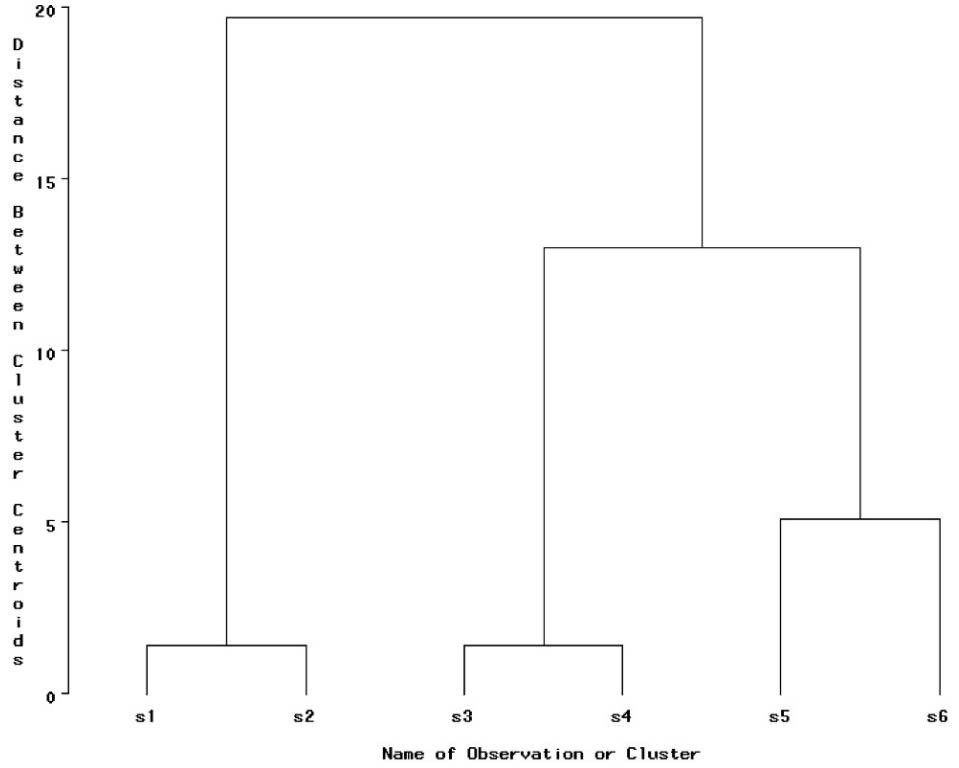
Obs	sid	income	educ
1	s1	5	5
2	s2	6	6

---------- CLUSTER=2 ----------

Obs	sid	income	educ
3	s3	15	14
4	s4	16	15

---------- CLUSTER=3 ----------

Obs	sid	income	educ
5	s5	25	20
6	s6	30	19



evaluation of cluster solution

- RMSSTD: a measure of the standard deviation of all variables over all observations (= scale dependent)

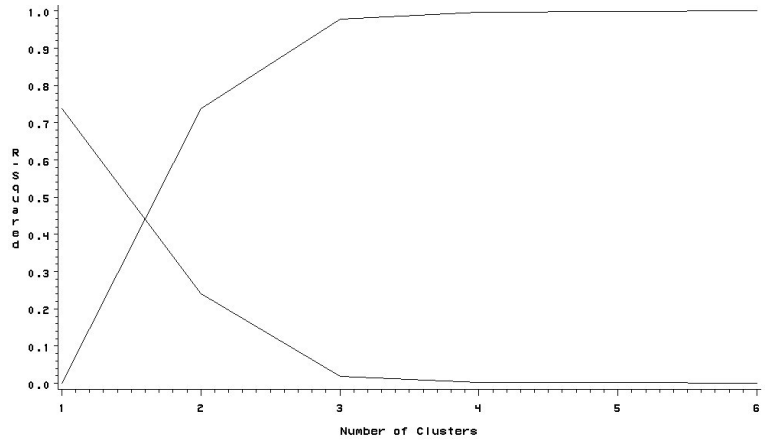
$$\sqrt{\frac{\sum_{j=1}^{k} s_j^2}{k}} \quad (here : 8.37) \quad \text{(similar to } \sqrt{SS_t})$$

  the smaller this value, the more homogeneous the data

- RMSSTD of new cluster: (should be small)
  (similar to $\sqrt{SS_w}$ of the new cluster)
  a measure of the standard deviation of all variables over all observations in the new cluster

- semipartial R-squared: increase in SS_w/SS_t or decrease in R^2-value (should be small):

- R-squared:
  (should be large)
  $R^2 = \frac{SS_b}{SS_t} = 1 - \frac{SS_w}{SS_t}$



- distance between the 2 clusters that were last combined
  (should be small)

example: consider different fooditems

fooditem	calories	protein	fat	calcium	iron
Braised beef	340	20	28	9	2.6
Hamburger	245	21	17	9	2.7
Roast beef	420	15	39	7	2.0
Beef steak	375	19	32	9	2.6
Canned beef	180	22	10	17	3.7
Broiled chicken	115	20	3	8	1.4
Canned chicken	170	25	7	12	1.5
Beef heart	160	26	5	14	5.9
Roast lamb leg	265	20	20	9	2.6
Roast lamb shoulder	300	18	25	9	2.3
Smoked ham	340	20	28	9	2.5
Roast pork	340	19	29	9	2.5
Simmered pork	355	19	30	9	2.4
Beef tongue	205	18	14	7	2.5
Veal cutlet	185	23	9	9	2.7
Baked bluefish	135	22	4	25	0.6
Raw clams	70	11	1	82	6.0
Canned clams	45	7	1	74	5.4
Canned crabmeat	90	14	2	38	0.8
Fried haddock	135	16	5	15	0.5
Broiled mackerel	200	19	13	5	1.0
Canned mackerel	155	16	9	157	1.8
Fried perch	195	16	11	14	1.3
Canned salmon	120	17	5	159	0.7
Canned sardines	180	22	9	367	2.5
Canned tuna	170	25	7	7	1.2
Canned shrimp	110	23	1	98	2.6

because of the different scales, the variables have to be standardized first!

Single Linkage Cluster Analysis
Root-Mean-Square Total-Sample Standard Deviation

Cluster History

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Min Dist	Tie
...							
10	CL19	CL12	6	0.0225	.899	0.975	
9	CL14	CL11	13	0.1347	.764	0.9922	
8	CL9	CL10	19	0.1079	.656	1.0382	
7	Raw clams	Canned clams	2	0.0043	.652	1.0608	
6	CL8	Roast beef	20	0.0674	.584	1.2801	
5	CL6	Canned shrimp	21	0.0410	.543	1.5361	
4	CL5	CL13	23	0.0793	.464	1.7509	
3	CL4	Beef heart	24	0.0810	.383	1.841	
2	CL3	Canned sardines	25	0.1455	.238	3.0861	
1	CL2	CL7	27	0.2376	.000	3.4184	

Complete Linkage Cluster Analysis
Root-Mean-Square Total-Sample Standard Deviation

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Maximum Distance	Tie
...							
10	CL12	Canned shrimp	5	0.0149	.936	1.7705	
9	CL14	Roast beef	7	0.0166	.919	1.869	
8	CL17	CL19	4	0.0137	.906	1.9279	
7	CL8	CL15	6	0.0340	.872	2.2996	
6	CL11	CL10	10	0.0484	.823	2.6444	
5	CL6	CL7	16	0.0792	.744	3.1163	
4	CL5	Beef heart	17	0.0758	.668	4.5596	
3	CL4	Canned sardines	18	0.1336	.535	5.1892	
2	CL3	CL13	20	0.2098	.325	5.7243	
1	CL9	CL2	27	0.3250	.000	6.0693	

Centroid Hierarchical Cluster Analysis

Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	False Centroid Distance	Tie
...							
10	CL11	CL13	11	0.0415	.892	0.8734	
9	Raw clams	Canned clams	2	0.0043	.887	1.0608	
8	CL10	CL16	13	0.0485	.839	1.2709	
7	CL8	Canned shrimp	14	0.0213	.818	1.2966	
6	CL7	CL12	16	0.0568	.761	1.3796	
5	CL20	Roast beef	7	0.0166	.744	1.4122	
4	CL5	CL6	23	0.2801	.464	1.7562	
3	CL4	Beef heart	24	0.0810	.383	2.5645	
2	CL3	CL9	26	0.2403	.143	3.0711	
1	CL2	Canned sardines	27	0.1428	.000	3.5113	

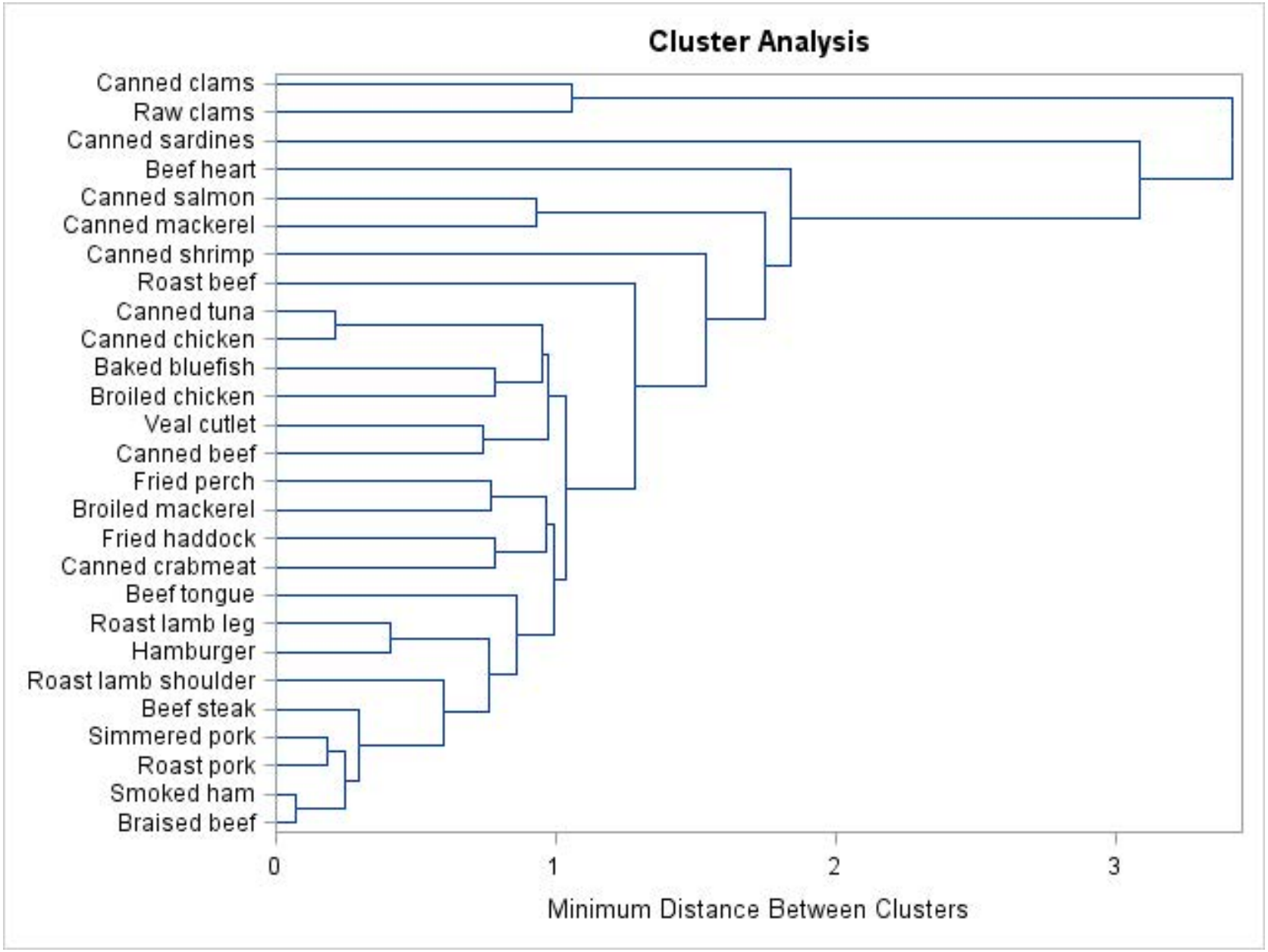
Ward's Minimum Variance Cluster Analysis

Root-Mean-Square Total-Sample Standard Deviation

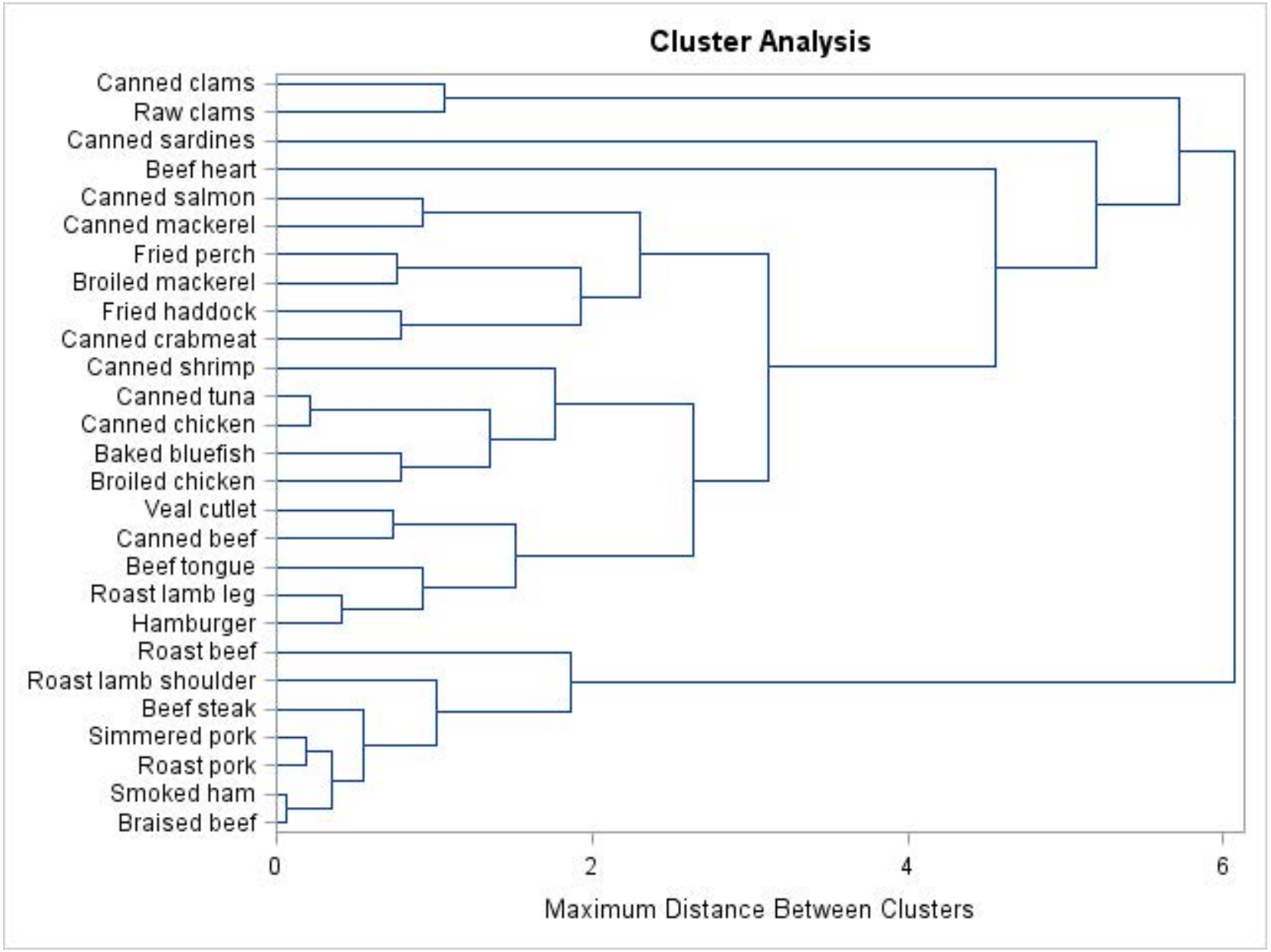
Cluster History

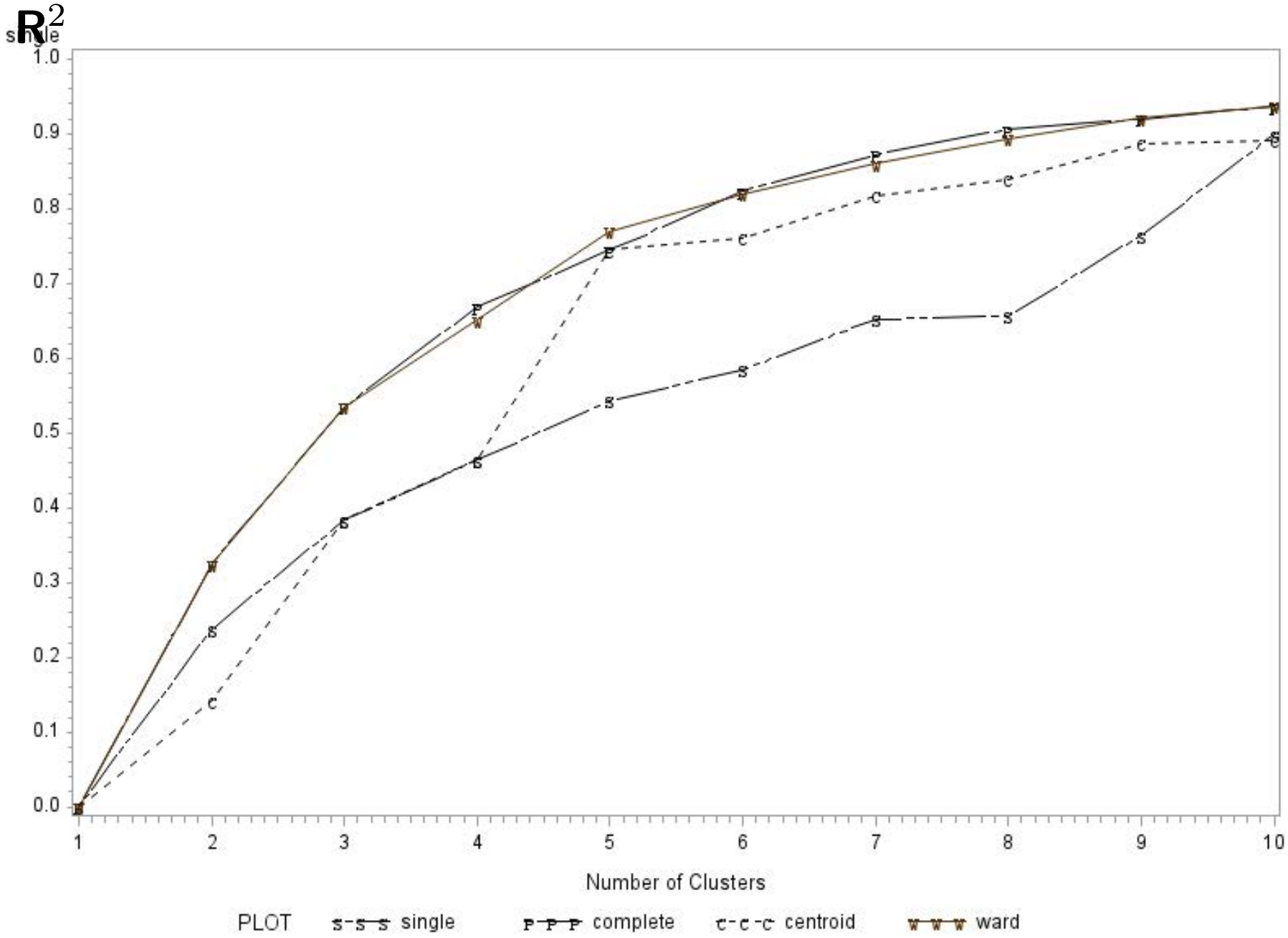
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	False Between Cluster SSq	Tie
...							
10	CL17	CL19	4	0.0137	.937	1.0131	
9	CL15	Roast beef	7	0.0166	.921	1.2104	
8	CL11	Beef heart	4	0.0271	.894	1.3838	
7	CL10	CL16	6	0.0340	.860	1.792	
6	CL14	CL8	7	0.0402	.819	1.859	
5	CL6	CL12	11	0.0498	.770	2.3509	
4	CL7	Canned sardines	7	0.1182	.651	3.1069	
3	CL5	CL4	18	0.1166	.535	3.6896	
2	CL3	CL13	20	0.2098	.325	5.2287	
1	CL9	CL2	27	0.3250	.000	9.0979	

method=single (note the typical *chaining* effect):



method=complete:




R^2
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10
Number of Clusters
PLOT s-s-s single p-p-p complete c-c-c centroid w-w-w ward

NAME	complete +centroid	NAME	ward	NAME	single
Beef steak	1	Beef steak	1	Baked bluefish	1
Braised beef	1	Braised beef	1	Beef steak	1
Roast beef	1	Roast beef	1	Beef tongue	1
Roast lamb shoulder	1	Roast lamb shoulder	1	Braised beef	1
Roast pork	1	Roast pork	1	Broiled chicken	1
Simmered pork	1	Simmered pork	1	Broiled mackerel	1
Smoked ham	1	Smoked ham	1	Canned beef	1
Baked bluefish	2	Baked bluefish	2	Canned chicken	1
Beef tongue	2	Beef heart	2	Canned crabmeat	1
Broiled chicken	2	Beef tongue	2	Canned shrimp	1
Broiled mackerel	2	Broiled chicken	2	Canned tuna	1
Canned beef	2	Canned beef	2	Fried haddock	1
Canned chicken	2	Canned chicken	2	Fried perch	1
Canned crabmeat	2	Canned shrimp	2	Hamburger	1
Canned mackerel	2	Canned tuna	2	Roast beef	1
Canned salmon	2	Hamburger	2	Roast lamb leg	1
Canned shrimp	2	Roast lamb leg	2	Roast lamb shoulder	1
Canned tuna	2	Veal cutlet	2	Roast pork	1
Fried haddock	2	Broiled mackerel	3	Simmered pork	1
Fried perch	2	Canned crabmeat	3	Smoked ham	1
Hamburger	2	Canned mackerel	3	Veal cutlet	1
Roast lamb leg	2	Canned salmon	3	Canned mackerel	2
Veal cutlet	2	Fried haddock	3	Canned salmon	2
Raw clams	3	Fried perch	3	Canned clams	3
Canned clams	3	Canned clams	4	Raw clams	3
Beef heart	4	Raw clams	4	Beef heart	4
Canned Sardines	5	Canned Sardines	5	Canned Sardines	5

nonhierarchical methods: K-means clustering

the nr of clusters (K) has to be chosen a priori!

different steps:

- select K initial seeds ($\approx$ cluster centroids)
- assign each observation to the cluster to which it is "closest"
- recompute the centroids
- reassign observations to the cluster of which the centroid is closest
- stop if no reallocation

methods for obtaining initial seeds:

- select the first K observations as initial seeds
- take the first nonmissing observation as the first seed;
  the next observation that is at least a certain distance separated from the first seed is the second seed, the third seed should be at least a certain distance separated from the previous seeds...
- take K observations randomly
- refine selected seeds by some rule
- use a heuristic to identify the cluster centers (which means that everything that you can think of is OK as long as it leads to K separated centroids)
- supplied by the expert
- use results from hierarchical analysis

Initial Seeds

Cluster	income	educ
1	5.000	5.000
2	16.000	15.000
3	30.000	19.000

Minimum Distance Between Initial Seeds = 14.56022

Iteration History

		Relative Change in Cluster Seeds		
Iteration	Criterion	1	2	3
1	1.5811	0.0486	0.1751	0.0486
2	1.1180	0	0	0

Convergence criterion is satisfied.

Obs	Cluster	Distance from Seed
1	1	0.7071
2	1	0.7071
3	3	0.7071
4	3	0.7071
5	2	2.5495
6	2	2.5495

we look for 3 clusters in the food data (eliminating the 2 outlying observations) and use the cluster means of the remaining clusters of the centroid approach as initial centroids

RESULTS

Obs	CLUSTER	_TYPE_	_FREQ_	calories	protein	fat	calcium	iron
1	1	0	7	1.43714	-0.10080	1.48009	-0.45171	0.02245
2	2	0	16	-0.39740	0.19110	-0.43697	-0.09163	-0.44924
3	3	0	2	-1.48118	-2.35200	-1.10877	0.43618	2.27093
4	4	0	1	-0.46842	1.64640	-0.75344	-0.38397	2.40779
5	5	0	1	-0.27080	0.70560	-0.39811	4.13968	0.08110

RESULTS

Initial Seeds

Cluster	calories	protein	fat	calcium	iron
1	1.437139501	-0.100800099	1.480085852	-0.451707752	0.022448583
2	-0.397399263	0.191100188	-0.436969614	-0.091632362	-0.449243216
3	-1.481184246	-2.352002316	-1.108771840	0.436180722	2.270927626

Minimum Distance Between Initial Seeds = 2.734591

Iteration History

		Relative Change in Cluster Seeds		
Iteration	Criterion	1	2	3
1	0.6293	0.1207	0.1040	0.2411
2	0.6093	0.0610	0.0404	0
3	0.6065	0	0	0

Convergence criterion is satisfied.

Criterion Based on Final Seeds = 0.6065

Cluster Summary

Cluster	Frequency	RMS Std Deviation	Maximum Distance from Seed to Observation	Radius Exceeded	Nearest Cluster	Distance Between Cluster Centroids
1	9	0.3956	1.6642		2	2.6316
2	14	0.7757	2.7899		1	2.6316
3	2	0.3510	0.5550		2	4.2509

Cluster Means

Cluster	calories	protein	fat	calcium	iron
1	1.152910901	0.097014250	1.174679952	-0.515909977	0.173140192
2	-0.532536046	0.270253982	-0.594217493	0.187911410	-0.482748025
3	-1.460346734	-2.328342000	-1.126537331	1.006215027	2.600105313

Cluster Standard Deviations

Cluster	calories	protein	fat	calcium	iron
1	0.516426239	0.420084025	0.560112034	0.014659310	0.159227953
2	0.355131911	0.884397545	0.349708059	1.211104178	0.714800958
3	0.168839287	0.685994341	0.000000000	0.124388369	0.318455906

Obs	fooditem	CLUSTER	DISTANCE	calories	protein	fat	calcium	iron
Cluster=1								
1	Braised beef	1	0.27849	1.23781	0.33955	1.21319	-0.51102	0.27322
2	Hamburger	1	1.33383	0.33046	0.58209	0.25997	-0.51102	0.34828
3	Roast beef	1	1.66424	2.00189	-0.87313	2.16642	-0.55500	-0.17714
4	Beef steak	1	0.57800	1.57209	0.09701	1.55982	-0.51102	0.27322
5	Roast lamb leg	1	0.94670	0.52148	0.33955	0.51994	-0.51102	0.27322
6	Roast lamb shoulder	1	0.46025	0.85577	-0.14552	0.95322	-0.51102	0.04804
7	Smoked ham	1	0.26108	1.23781	0.33955	1.21319	-0.51102	0.19816
8	Roast pork	1	0.15338	1.23781	0.09701	1.29985	-0.51102	0.19816
9	Simmered pork	1	0.31537	1.38107	0.09701	1.38651	-0.51102	0.12310
Cluster=2								
10	Canned beef	2	1.78953	-0.29035	0.82462	-0.34663	-0.33511	1.09889
11	Broiled chicken	2	0.90429	-0.91116	0.33955	-0.95322	-0.53301	-0.62751
12	Canned chicken	2	1.43897	-0.38586	1.55223	-0.60660	-0.44506	-0.55245
13	Beef tongue	2	1.33148	-0.05158	-0.14552	0.00000	-0.55500	0.19816
14	Veal cutlet	2	1.38713	-0.24260	1.06716	-0.43328	-0.51102	0.34828
15	Baked bluefish	2	1.04526	-0.72014	0.82462	-0.86657	-0.15920	-1.22799
16	Canned crabmeat	2	1.69071	-1.14994	-1.11566	-1.03988	0.12666	-1.07787
17	Fried haddock	2	1.36952	-0.72014	-0.63059	-0.77991	-0.37909	-1.30306
18	Broiled mackerel	2	1.13689	-0.09933	0.09701	-0.08666	-0.59898	-0.92775
19	Canned mackerel	2	2.71879	-0.52912	-0.63059	-0.43328	2.74334	-0.32726
20	Fried perch	2	1.21122	-0.14709	-0.63059	-0.25997	-0.40108	-0.70257
21	Canned salmon	2	2.78988	-0.86341	-0.38806	-0.77991	2.78732	-1.15293
22	Canned tuna	2	1.51789	-0.38586	1.55223	-0.60660	-0.55500	-0.77763
23	Canned shrimp	2	1.80402	-0.95892	1.06716	-1.12654	1.44599	0.27322
Cluster=3								
24	Raw clams	3	0.55497	-1.34096	-1.84327	-1.12654	1.09417	2.82529
25	Canned clams	3	0.55497	-1.57973	-2.81341	-1.12654	0.91826	2.37492

PAM - using medoids instead of centroids

sometimes, the representative observation of a cluster has to be one of the observations and not the average of all observations in the cluster (because averaging does not make sense, or to increase the interpretability of the clusters)

this representative observation is called the **medoid** of the cluster instead of the centroid (so it is the data point in the cluster that is *most similar to* or *in the center of* all other points in the cluster)

the corresponding K-means type algorithm is called **PAM**, **Partitioning Around Medoids**, and often uses a distance measure that is more robust to outliers, such as the Manhattan distance (this is much more demanding in terms of memory and computation time)

THE END