Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Nov 2, 2020 - Dec 16, 2020

Course material

Texts

1. Fox, J. (2015 3rd edition or 2008 2nd edition), *Applied Regression Analysis and Generalized Linear Models*, Sage
2. Faraway, J.J. (2002), *Practical Regression and Anova using R* - web text
3. Faraway, J.J. (2006 1st edition or 2016 2nd edition), *Extending the Linear Model with R*, Chapman & Hall /CRC
4. Optional, not used in course: McCulloch, C.E., Searle, S.R., and Neuhaus, J.M. (2008), *Generalized, Linear, and Mixed Models*, Wiley

ad 1 Text 1 is our main text book. We need extra appendices on Linear Algebra and Maximum Likelihood, see http://socserv.socsci.mcmaster.ca/jfox/Books/Applied-Regression-3E/Appendices.pdf).

ad 2 Text 2 is web based text (see http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf). Text is similar, but not identical to Faraway, J.J. (2006), *Linear Models with R*, Chapman & Hall /CRC or the second edition of this book, but does not contain the exercises. We will supply some exercises from the book, and use the R-package `faraway`.

ad 3 Text 3 is joint course material with course Mixed and Longitudinal Modeling in next semester. In the last 3 weeks of present course we will use the book.

ad 4 Text 4 is statistically more challenging, and contains joint material with course Mixed and Longitudinal Modeling. It weaves Linear, Generalized Linear and Mixed Models nicely together. Only for background reading, not needed for exam.

Organizational remarks

- When: classes on Monday, Tuesday and Wednesday, starting Monday Nov 2, 2020; 7 weeks; Mon+Tue: 9:15-13:00h, Wed: 14:15-18:00h
- Where: on-line classes
- What: lectures, practicals, case study
- Lectures:
  - on-line meeting using Kaltura Live room (9:15-9:45h on Mon+Tue, 14:15-14:45h on Wed); summary of course material of the day, ask questions
  - two prerecorded lectures on Brightspace (9:45-11:00h on Mon+Tue, 14:45-16:00h on Wed); prerecorded lectures can be watched at alternative times too
- Practical:
  - on-line using Kaltura Live rooms and Breakout rooms (11:15-13.00h on Mon+Tue, 16:15-18:00h on Wed), theory and practice using statistical software
- Examination
  - Two parts:
    1. Case study (group work): analyse practical dataset or study theoretical topic, starting in week four; hand in report on case study in week seven; give short oral presentation on case study in last meeting.
    2. Written exam: date in January to be decided
  - Final score is weighted average based on the case study (1/3) and the written exam (2/3) if the score of written exam is 5 or higher and score of case study report is 6 or higher; if score of written exam is lower than 5, then final score equals the score of written exam alone; score of case study needs to be at least 6.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 1, lecture 1

Content meeting 1: chapters 1-2 Fox book

- Chapter 1: Introduction
  - statistical models
  - observational vs experimental studies
  - populations and samples
  - word "regression"
- Chapter 2: What is regression analysis?
  - main idea of regression
  - utopian assumptions
  - nonparametric approaches

Statistical Models

- Model is abstraction of a real phenomenon or process isolating the aspects relevant to a particular question
- Statistical model is model with stochastic component, containing unknown parameters (constants), to be estimated from data. Describes distributional properties of response variables, thereby decomposing variability in known and unknown sources.
- Principle of parsimony: try to find a parsimonious model, i.e. a model that is as simple as possible, yet capturing the essence. This principle is also known as "Occam's razor".

- George Box: "All models are wrong but some are useful".

A few remarks

- Fox Book is written from Social Science point of view. We will touch upon examples from other fields of science, e.g. Life Sciences, too.
- Statistical data analysis describes outcomes of real social / biological processes and hardly ever the processes themselves. Therefore, we often attend to the descriptive nature of statistical models, and do not treat the models as if they were reality.
- Randomness plays big role in social and biological processes. Therefore, randomness should be important part of statistical models too. The stochastic component of statistical models handles this.

Observation and Experiment

- Make distinction between observational and experimental studies.
- Causal inference most certain in randomized experiments.
- Good experimental practice: avoid the confounding of experimentally manipulated explanatory variables with other factors that can influence the response using experimental design
- Sound analysis of observational data seeks to control statistically, i.e. through the statistical analysis, for potentially confounding factors.
- Observational data: distinguish between factor that is common prior cause of an explanatory and response variable, and factor that intervenes causally between the two.
  - Example observational data: occupational prestige, educational level of occupation, and income level of occupation.
  - Occupations with higher level of education tend to have higher prestige, occupations with higher levels of income also tend to have higher prestige. But income and educational level are also positively related.
  - Education is common prior cause of income and prestige, whereas income intervenes in the relation between education and prestige.
  - When education is controlled statistically, relationship between prestige and income decreases; likewise, when income is controlled, relationship between prestige and education decreases; but neither relationship disappears.
  - Helpful to use "causal model": education influences both income and prestige, while income potentially influences prestige.

Populations and Samples

- Statistical inference is built on framework of random sampling from population.
- Application is broader in practice, e.g. experiment in which subjects are assigned values of explanatory variables at random: inferences are properly made to hypothetical population of random rearrangements of subjects, even when subjects not sampled from larger population.
- Randomization and good sampling design are desirable in social research, but are not prerequisites for drawing statistical inference. Even when randomization or random sampling is employed, we typically want to generalize beyond strict bounds of statistical inference.

Regression

Word "regression" has historical background:

- Word coined by Francis Galton in 19^{th} century.
- Galton was English Victorian anthropologist, geneticist, explorer, inventor, meteorologist, psychometrician, and ...statistician; he was a cousin of Darwin
- Phenomenon described was "regression to the mean": heights of sons of tall fathers tend to be smaller than their fathers ("regress to the mean")
- For Galton this was biological phenomenon, but generally speaking it is the phenomenon that if a variable is extreme on its first measurement, it will tend to be closer to the average on its second measurement

What is Regression Analysis?

Regression Analysis is one of the many tools within statistical data analysis. Statistical data analysis is a craft, part art (skill, developed in practice) and part science; chapters 2-4 deal with elements of statistical data analysis.

- Regression analysis examines the relationship between quantitative response Y, and one or more explanatory variables $X_1, \ldots, X_k$.
- Regression analysis traces the conditional distribution of Y, or some aspect of it, like the mean, as a function of X's. In general: $p(y|x_1, \ldots, x_k) = f(x_1, \ldots, x_k)$.
- Example: dataset on Hourly Wages (Y) as function of Education in years (X).

```
> SLID <- read.table("SLID-Ontario.txt",header=T)
> head(SLID)
```

```
  age    sex wages education
1  40   Male  10.6        15
2  19   Male  11.0        13
3  46   Male  17.8        14
4  50 Female  14.0        16
5  31   Male   8.2        15
6  30 Female  17.0        13
```

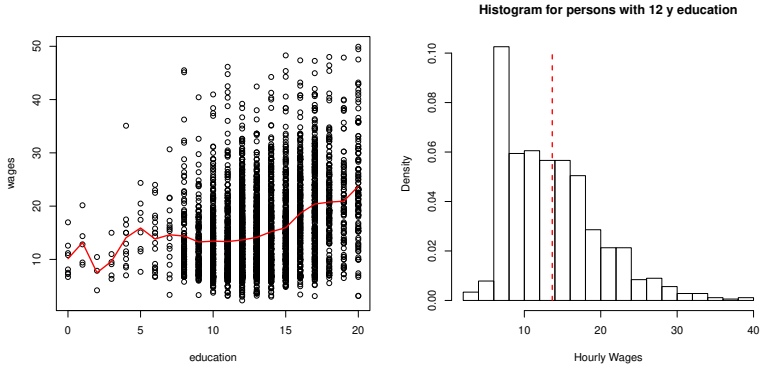
```
> summary(SLID)
```

```
      age         sex           wages          education
 Min.   :16   Female:2007   Min.   : 2.3   Min.   : 0.0
 1st Qu.:28   Male  :1990   1st Qu.: 9.2   1st Qu.:12.0
 Median :36                 Median :14.1   Median :13.0
 Mean   :37                 Mean   :15.5   Mean   :13.2
 3rd Qu.:46                 3rd Qu.:19.8   3rd Qu.:15.0
 Max.   :65                 Max.   :49.9   Max.   :20.0
```


The idea of regression

Relate the hourly wages to years of education. What is done below?

```
> attach(SLID)
> oldPar <- par(mfrow = c(1,2))
> plot(wages ~ education)
> meansperyear <- tapply(wages, education, mean)
> lines(0:20, meansperyear, col="red", lw=2)
> hist(wages[education==12], freq=FALSE, breaks=20,
+      main="Histogram for persons with 12 y education",
+      xlab="Hourly Wages")
> mean.wage.12y <- mean(wages[education==12])
> segments(mean.wage.12y, 0, mean.wage.12y, 0.1, lty=2, col="red", lw=2)
> par(oldPar)
```



Preliminaries

"Utopian" situation of linear regression analysis assumes that

- conditional distribution of response Y, $p(Y|x_1, \ldots, x_k)$, is normal distribution
- $var(Y)$ given Xs is constant
- linearity: $\mu \equiv E(Y|x_1, \ldots, x_k) = \alpha + \beta x_1 + \ldots + \beta_k x_k$

However (cf figure 2.3):

- distribution may be skewed, like in example of wages given education
- distribution may be multimodal (i.e. more than one top)
- distribution may have heavy tails
- spread may be unequal, e.g. $var(Y)$ may depend on X
- nonlinearity

This does not jeopardize usefulness of linear models. But be aware of limitations. Transform data to make assumptions of normality, equal variance and linearity more correct. Alternative approach is nonparametric regression.

Naive Nonparametric Regression

Binning and averaging

- discrete explanatory variables: examine conditional distribution of Y given the Xs directly, like in the wages versus education example, where we looked at $\bar{Y}|x$, the average wage given years of education.
- continuous explanatory variables: dissect X into small bins, and continue as in discrete case.
- wider bins have more bias, but less variance –> reduction in bias and variance cannot be achieved simultaneously.
- with more Xs binning becomes more and more impractible, because of "intrinsic sparseness of multivariate data" or "curse of dimensionality".

Local Averaging

In smaller samples, local averages of Y can be calculated in neighbourhood surrounding each x-value in the data: moving window. Defects of the procedure

- boundary bias: artifical flattening at edges of data.
- line connecting averages rather rough, because average jumps as observations enter and exit window.
- unusual data, outlier, unduly influences average.

Lowess (LOcally WEighted Scatterplot Smoothing) is better procedure. Looks like local-average smoother, but lowess smoother:

- computes fitted value based on least-squares regression.
- locally weighted regression with more weight to observations in neighbourhood close to focal x.
- makes provision for discounting outliers.

Span of lowess smoother (fraction of data used to compute each fitted value) must be specified. Larger span reduces variance, but increases bias (hence smoother regressions).

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 1, lecture 2

Content week 1, lecture 2: chapter 3 Fox book

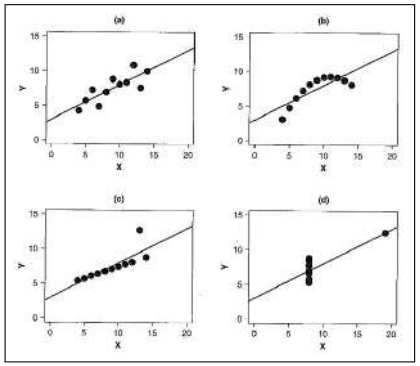
Chapter 3: Data examination

- data summaries: numerical and graphical
- univariate graphical: histogram, stem-and-leaf plot, QQ-plot to check distribution, boxplot
- bivariate graphical: scatterplot, side-by-side boxplot
- multivariate graphical: scatterplot matrix, coplot

Examining Data

- Careful data analysis should start with inspection of data.
- Four datasets with identical linear least-squares regression from Anscombe (1973).

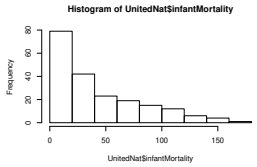
subplot (a) ok
subplot (b) curvilinear relationship
subplot (c) outlier
subplot (d) influential observation



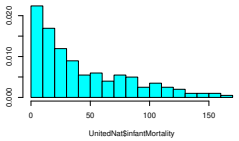
- Numerical summaries do not tell the whole story.
- Examine data graphically!

Univariate Displays: histograms and stem-and-leaf plots

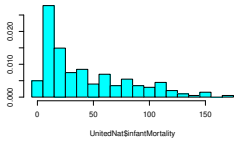
```
> UnitedNat <- read.table("UnitedNations.txt",header=T)
> hist(UnitedNat$infantMortality)
```




```
> require(MASS)
> truehist(UnitedNat$infantMortality, h=10, x0=0)
```




```
> truehist(UnitedNat$infantMortality, h=10, x0=-5)
```



Univariate Displays: stem-and-leaf plot

```
> stem(UnitedNat$infantMortality)
```

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

 0 | 2345555555666666667777777777788888999999
 1 | 000011222222333344444555566778888999
 2 | 00112223333444455556669
 3 | 00012334455777889999
 4 | 012344456889
 5 | 11246667888
 6 | 012455568
 7 | 122347788
 8 | 00222456669
 9 | 025678
10 | 234677
11 | 023445
12 | 2445
13 | 25
14 | 29
15 | 34
16 | 9
```

Disadvantages of histograms and stem-and-leaf plots:

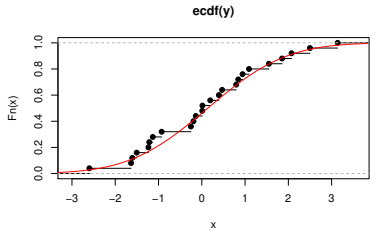
- binning (origin, bin width) influences form of histogram
- histogram is discontinuous

Nonparametric density estimators overcome these problems by smoothing histogram. Histogram itself is simple density estimator. Others are e.g. kernel density estimators. Kernel supplies weights for weighted averaging. R-function `density()` may be used.
Paul Eilers (Dutch statistician and former teacher in our MSc program) promotes density estimation based on P-splines.

QQ-plots (1)

Quantile-Quantile plots are useful for comparing an empirical distribution with a theoretical distribution.
Let $P(x) = P(X \leq x)$ be the *CDF=Cumulative Distribution Function* of x, to be compared with the data.
If the data are a sample from this distribution, a direct estimator of $P(x)$ would be *ECDF=Empirical Cumulative Distribution Function*: $\hat{P}(x) = \frac{\#_{i=1}^{n}(X_i \leq x)}{n}$.
Comparison with theoretical distribution is bit difficult, though, because ECDF should follow the curved theoretical CDF.

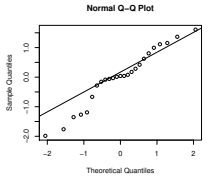
```
> y <- rnorm(25)
> EmpCDF <- ecdf(y)
> plot(EmpCDF)
> p <- seq(from=0.005, to=0.995, by=0.005)
> x <- mean(y) + sd(y)*qnorm(p)
> lines(x,p,col="red")
```



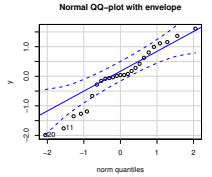
QQ-plots (2)

Quantile-Quantile plot shows quantiles, instead of cumulative probabilities:

1. Order data from smallest to largest: results are order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.
2. By convention, cumulative proportion below $X_{(i)}$ is $P_i = \frac{i-\frac{1}{2}}{n}$. (Other convention is $P_i = \left(i - \frac{1}{3}\right) / \left(n + \frac{1}{3}\right)$.)
3. Use inverse of CDF (quantile function) to find z_i corresponding to cumulative probability P_i: $z_i = P^{-1}\left(\frac{i-\frac{1}{2}}{n}\right)$, estimator of the expectation of order statistic $X_{(i)}$.
4. Plot z_i horizontally versus $X_{(i)}$ vertically. Plot should be approximately linear. If distributions are identical except for location and scale, $X_{(i)} \approx \mu + \sigma z_i$.
5. Comparison line often added to facilitate perception of departures from linearity. Line may be plotted through 1st and 3rd quartiles.
6. Deviations from linearity are expected due to sampling variation. Expected degree of sampling variation may be added as "envelope".



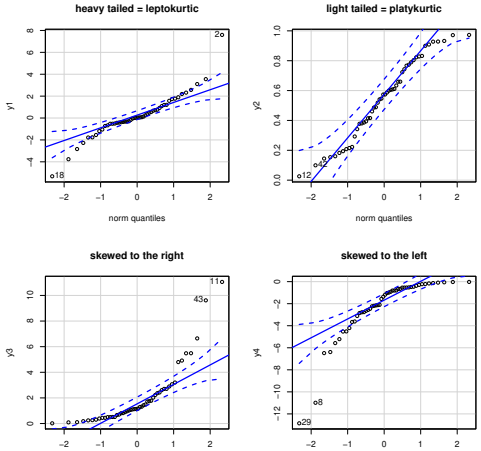

```
> y <- rnorm(25)
> qqnorm(y)    # check for normal distribution
> qqline(y)    # line through quartiles is added
> # Now load package car = Companion to Applied Regression by John Fox
> require(car)
> qqp <- qqPlot(y)   # normal QQ plot with 95% envelope
> title("Normal QQ-plot with envelope")
```



QQ-plots (3)

Some QQ-plots with typical deviations from normality.

```
> y1 <- rt(50,2)      # t-distribution with 2 df
> y2 <- runif(50)     # uniform(0,1) distribution
> y3 <- rchisq(50,2)  # chi-square distribution with 2 df
> y4 <- -rchisq(50,2)
> qq1 <- qqPlot(y1); title("heavy tailed = leptokurtic")
> qq2 <- qqPlot(y2); title("light tailed = platykurtic")
> qq3 <- qqPlot(y3); title("skewed to the right")
> qq4 <- qqPlot(y4); title("skewed to the left")
```



Boxplots

Boxplot is graphical representation of five-number summary of distribution: minimum, first quartile, median, third quartile, maximum, extended with outliers, if present.

- Line within box represents median.
- Box is drawn between hinges = Q1 and Q3; hence length of box is *IQR*=Inter Quartile Range.
- Points are plotted separately (are "outlying") if further away from hinge than $1.5 \times IQR$. This is default in `boxplot()` function. Other programs may use other values. Outliers may be defined in many different ways.

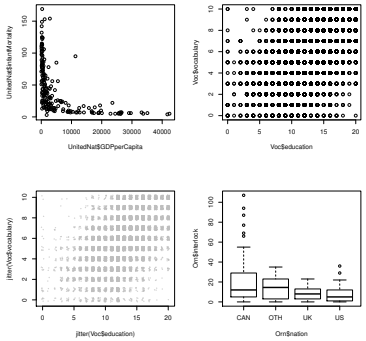
```
> boxplot(UnitedNat$infantMortality)
> title("Infant mortality",cex.main=2)
```

Plotting bivariate data

- Scatterplot is natural graphical display of relationship between two quantitative variables.
- Interpretation of scatterplot assisted by graphing nonparametric regression, summarizing relationship.
- Scatterplots of discrete data can be enhanced by randomly jittering of data.
- Parallel boxplots display relationship between quantitative response and discrete explanatory variable.

Note result of R-function `plot()` depending on type of argument.

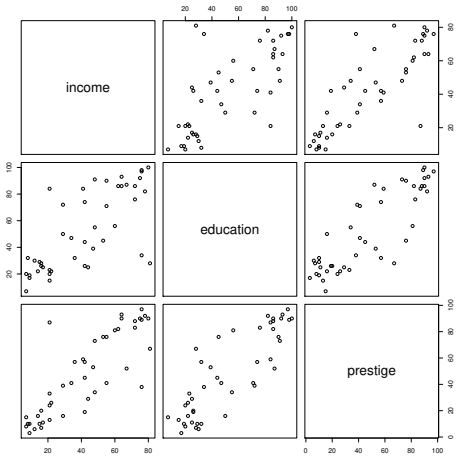
```
> Voc <- read.table("Vocabulary.txt",header=T)
> Orn <- read.table("Ornstein.txt",header=T)
> plot(UnitedNat$infantMortality ~ UnitedNat$GDPperCapita)
> plot(Voc$vocabulary ~ Voc$education)
> plot(jitter(Voc$vocabulary) ~ jitter(Voc$education),
+      pch=19, cex=0.1, col="grey")
> plot(Orn$interlock ~ Orn$nation)
```



Plotting multivariate data

- Scatterplot matrices: pairwise scatterplots

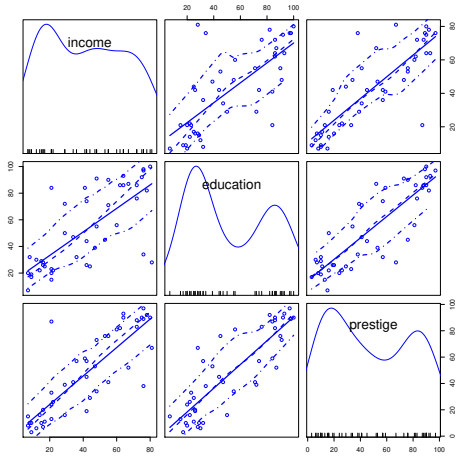
```
> pairs(Duncan[,2:4])
```



Plotting multivariate data

- More fancy scatterplot matrices:

```
> scatterplotMatrix(Duncan[,2:4])  # in package car
```



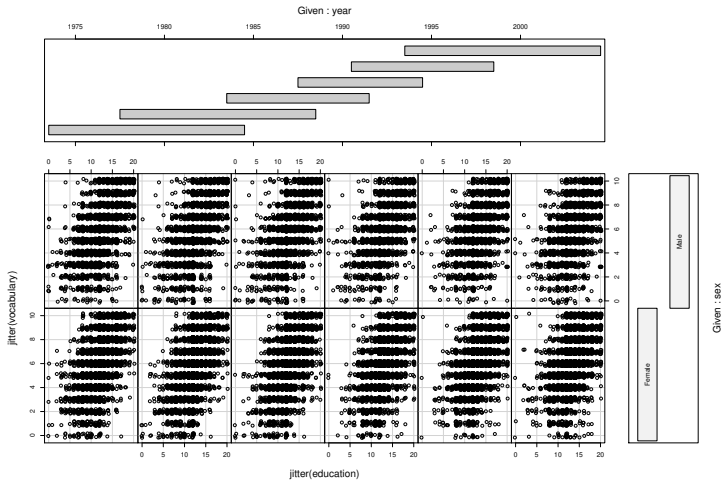
- Coded scatterplots: use value of third qualitative variable as plotting symbol or color.
- Three-dimensional scatterplots
- Conditioning plots

Coplots (1)

Conditioning plot = coplot

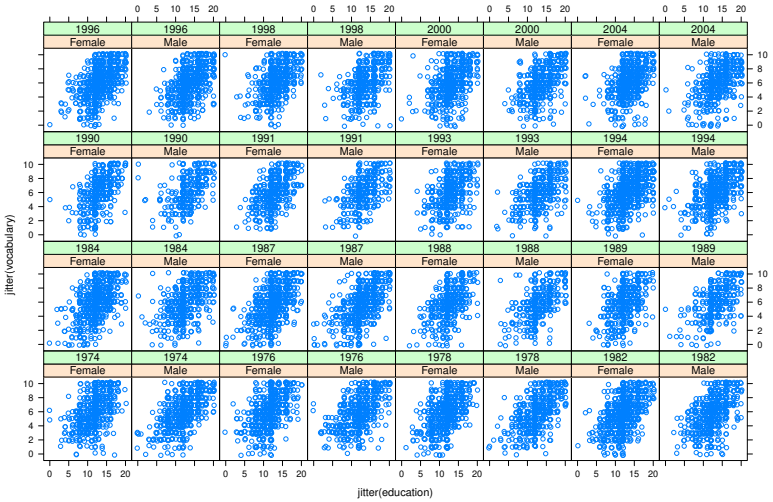
- Divide data into groups based on value(s) of other explanatory variable(s)=conditioning variables.
- If conditioning variable is continuous, use bins, or overlapping bins = shingles.

```
> coplot(jitter(vocabulary) ~ jitter(education) | year*sex, data=Voc)
```



Coplots (2)

```
> require(lattice)
> Voc$year <- as.factor(Voc$year)
> print(xyplot(jitter(vocabulary) ~ jitter(education) | sex*year, layout=c(8,4,1), data=Voc))
```



Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 1, lecture 3

Content week 1, lecture 3: chapter 4 Fox book

Data transformations

- enhance symmetry, linearize relationship, equalize variances
- power transformation / Box-Cox transformation for positive outcomes
- logit, probit, arcsine-square root for proportions
- bulging rule to assist in choice for linearizing transformation

Transforming Data

- Classical statistical methods, like linear least-squares regression, make strong assumptions about structure of data.
- But: assumptions may fail. Recall "utopian" assumptions.
- Different solutions, e.g.
  - abandon classical methods, use e.g. nonparametric regression
  - use more appropriate distributions and/or non-linear relationships: g.l.m. = generalized linear model (second part of course)
  - stick to linear model, but change response y and/or regressor(s) x by transformation so that data conform more closely to needed assumptions.
- Transformations may:
  - make distributions more symmetric
  - make relationships more linear
  - equalize variation across groups.

Family of Powers and Roots

- Family of power transformations: $X \to X^p$
  - e.g. $X^{-1} = 1/X$, $X^{1/2} = \sqrt{X}$
- Box-Cox family of transformations $X \to X^{(p)} \equiv \frac{X^p - 1}{p}$
  - Two families are in essence identical
  - Division by p preserves direction of X
  - For all p, at $X = 1$ we have $X^{(p)} = 0$ and slope=1
  - Descending "ladder" from $p = 1$ towards $p = -1$ compresses large values of X and spreads out small values
  - Ascending ladder from $p = 1$ towards $p = 2$ spreads out large values and compresses small values
  - By convention $X^{(0)} = \log_e(X) (= ln(X))$

Family of Powers and Roots (2)

- Often more convenient to take $\log_{10} X = \log_{10} e^{lnX} = lnX \log_{10} e = 0.4343 \times \ln X$. So, one is just a multiple of the other.
- Log transformation has nice multiplicative interpretation: $\log(XY) = \log(X) + \log(Y)$.
- Power transformation only sensible if $X > 0$.
- Only effective if ratio of largest to smallest data value is sufficiently large.
- Consider subtracting a "start" if this ratio < 5, like $X \to \log(X - 2000)$.
- Usually integer values of p are chosen, or "nice" fraction $\frac{1}{2}$ or $\frac{1}{3}$.

Transforming Skewness

- We like symmetrical distributions, because e.g.
  - classical statistical methods are based on assumptions of normality, which is symmetrical distribution
  - for symmetrical distribution means and standard deviations can be used as summaries for location and spread, which are less reasonable for skewed distribution.
- Select transformation to obtain (approximate) symmetry e.g. by examining median and hinges ($Q1$ and $Q3$), by trial and error, calculating

$$\frac{\text{Upper hinge} - \text{Median}}{\text{Median} - \text{Lower hinge}} \tag{1}$$

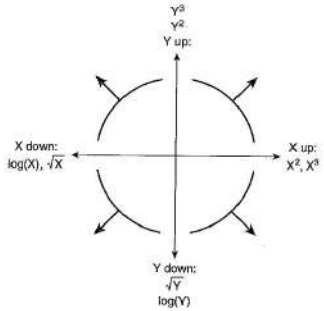
Value should be close to 1 for symmetrical distribution.

Transforming Skewness (2)

- Order statistics are preserved under power transformation, hence median of transformed variable is same as transformed median, but not for means and standard deviations. For skewed data, data summaries based on order statistics (median, IQR) generally make more sense.
- Positive skew (skewness to the right) is corrected by smaller powers (closer to 0); negative skew by larger powers (e.g. X^2).
- Often a range of transformations behaves approximately equal. Interpretation may aid in choice: multiplicative interpretation of $log(X)$, inverse of time required to travel a given distance is speed, square root of measure of area is linear measure of size, cube of linear measure of size is volume.

Transforming Nonlinearity

- Linear relationships are preferred because of
  - simple interpretation: in $\hat{Y} = a + bX$ unit increase in X, regardless of its level, corresponds with average change b of Y.
  - Simple and elegant statistical theory for linear models. Therefore we have the present course!
- Power transformations can make many nonlinear relationships more linear.
- Linearity may be achieved by power transformation of X, of Y, or of both.
- Tukey and Mosteller's bulging rule may be helpful in deciding what to do.



Transforming Nonconstant Spread

- Differences in spread often relates to differences in level: groups with higher levels tend to have higher spread as well. This is often observed in variables with a bound below, like counts or size.
- If there is positive association between level of variable in different groups and its spread (higher levels, higher spread), spreads can be made more equal by descending the ladder of powers, e.g. $\sqrt{}$ or $log()$ transformations.
- Negative association hardly ever occurs, but could be remedied by taking larger power.

Transforming Nonconstant Spread (2)

- From school of Tukey comes recipee to decide on power of transformation for variance stabilization: regress log(hinge spread) on log(median): $\log \text{spread} \approx a + b \log \text{level}$. Variance stabilizing transformation has power $p = 1 - b$.
- Unequal spread and skewness commonly occur together, because they have common origin, e.g. lower bound on possible values of variable of interest.

John Wilder Tukey
16/6/1915 - 26/7/2000

Transforming Proportions

- Power transformations can generally not be used for proportions, which have values between 0 and 1.
- Percentages and many rates (e.g. infant mortality per 1000 living births) are simple rescaled proportions, and are similarly affected.
- Sometimes "disguised" proportions, e.g. number of correct questions in an exam.
- Commonly employed transformations for proportions are:
  - *logit* transforms $P \rightarrow logit(P) = \log_e(\frac{P}{1-P})$
  - *probit* transforms $P \rightarrow probit(P) = \Phi^{-1}(P)$
  - *arcsine-square-root* transforms $P \rightarrow \sin^{-1}\sqrt{P}$

Logit transformation

- *logit* is $\log_e$ of the *odds*, with $odds = \frac{P}{(1-P)}$.
- *odds* transforms the interval $(0, 1)$ to $(0, \infty)$, and $\log_e$ transforms the interval $(0, \infty)$ to $(-\infty, \infty)$.

P	odds	logit
0.01	1/99	-4.60
0.05	1/19	-3.18
0.10	1/9	-2.20
0.30	3/7	-0.85
0.50	1	0.00
0.70	7/3	0.85
0.90	9/1	2.20
0.95	19/1	2.94
0.99	99/1	4.60

- Between 0.2 and 0.8 the *logit* transformation is nearly linear.
- If $P = 0$ or 1, $logit(P)$ cannot be calculated. If original counts k successes out of n are available, we may replace P by $P' = \frac{k+1/2}{n+1}$, and use $logit(P')$.
- $logit(P')$ is known as empirical logit.

Probit and arcsine-square-root transformations

- *probit*
  - $P \rightarrow probit(P) = \Phi^{-1}(P)$ with Φ^{-1} the inverse cumulative distribution function (quantile function) of the standard normal distribution.
  - *logit* and *probit* are nearly indistinguishable, once scales are equated: $logit \approx (\pi/\sqrt{3}) \times probit$
  - Use $probit(P')$ with $P' = \frac{k+1/2}{n+1}$ if $P = 0$ or 1.
- *arcsine-square-root*
  - $P \rightarrow \sin^{-1}\sqrt{P}$
  - Transformation stabilizes the variance of a binomial proportion: if $k \sim Binom(n, p)$, then $var(k/n) = p(1-p)/n$, and $var(sin^{-1}\sqrt{k/n}) \approx 1/(4n)$.
  - Minimum value of $\sin^{-1}\sqrt{P} = 0$ for $P = 0$, and maximum value is $\pi/2$ for $P = 1$.
  - Some authors dislike arcsine-square-root transformation: Warton, D.I., and F.K.C. Hui (2011). The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3-10.
- Functions like *logit* and *probit* will return later as link functions in generalized linear models.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 1, lecture 4

Content week 1, lecture 4: chapter 5 Fox book

In chapter 5 the regression model is described, but assumptions are not yet specified. Chapter 6 (lectures 6-7) introduces the formal regression model, including assumptions, leading to statistical inference.

- Utility of linear least-squares regression
- Simple regression:
  - Regression equation
  - Residuals, least-squares and normal equations by differentiation
  - Least-square estimators for slope and intercept
  - Interpretation slopes and intercept
  - Absolute goodness of fit (gof): residual standard error
  - Degrees of freedom for error
  - Model comparison: regression model vs null model
  - Sums of squares: residual SS and total SS
  - Analysis of variance: $TSS = RegSS + RSS$
  - Relative gof:$R^2 = RegSS/TSS$ = square of (multiple) correlation coefficient

Linear Least-Squares Regression

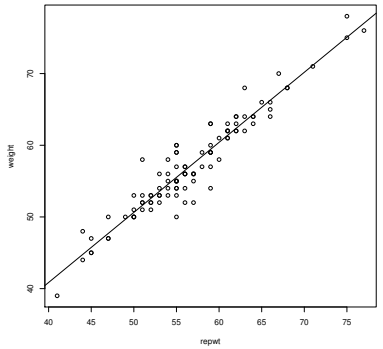
Linear Regression lies at heart of applied statistics:

- Some data may be adequately summarized by linear regression, but...
- By data transformation and diagnosis of problems the applicability of regression is considerably expanded.
- General linear models, a direct extension of regression models, can accommodate even more situations, both with qualitative explanatory variables and polynomial functions of quantitative regressors.
- Linear regression is starting point for wide variety of other generalizations, like weighted regression, robust regression, nonparametric regression, and generalized linear models.

Simple Regression: Least-Squares Fit (1)

Simple regression is regression with a single regressor.
Example: Davis data on measured and reported weight (kg) of 101 women. The relationship between measured weight (Y) and reported weight (X) appears linear.

```
> DavisF <- Davis[Davis$sex=="F" & Davis$weight<160,]  # select females and remove outlier
> plot(weight ~ repwt, data=DavisF)
> abline(lm(weight ~ repwt, data=DavisF))
> # function lm() fits linear models, function abline() plots regression line
```



Simple Regression: Least-Squares Fit (2)

- Line relating Y and X has equation $Y = A + BX$.
- Too simplistic, though, because no straight line passes perfectly through all points (X, Y), hence residual E is needed.
- Regression equation for observation $i = 1, \ldots, n$ is $Y_i = A + BX_i + E_i = \hat{Y}_i + E_i$, with $\hat{Y}_i$ the fitted value for observation i.
- Residual $E_i = Y_i - \hat{Y}_i =$ vertical distance between observed Y and predicted Y on regression line.
- What criterion could be used to find A and B? Good fitting line has predicted Y close to observed Y, hence small E_i.
- Making sum of residuals $\sum_{i=1}^{n} E_i$ as small as possible, does not work, because it is easy to make $\sum_{i=1}^{n} E_i = 0$: any line passing through $(\bar{X}, \bar{Y})$ has this property: line satisfies $\bar{Y} = A + B\bar{X}$, hence $Y_i - \bar{Y} = B(X_i - \bar{X}) + E_i$ (by subtraction), and $\sum_{i=1}^{n} E_i = \sum(Y_i - \bar{Y}) - B\sum(X_i - \bar{X}) = 0 - B \times 0 = 0$

Simple Regression: Least-Squares Fit (3)

- Option: find A and B to minimize sum of absolute residuals: $\sum |E_i|$ leads to *least-absolute-value regression* (*LAV*), but mathematically difficult.
- Option: find A and B to minimize sum of squared residuals, $\sum E_i^2$; leads to *least-squares regression*.
- Find intercept and slope A and B to minimize $S(A, B) = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - A - BX_i)^2$.
- Minimization can be done with high-school mathematics: equate derivatives w.r.t. A and B to 0, and solve.
- Leads to system of linear equations, called normal equations (check yourself):

$$\begin{aligned} An + B\sum X_i &= \sum Y_i \\ A\sum X_i + B\sum X_i^2 &= \sum X_i Y_i \end{aligned}$$

Simple Regression: Least-Squares Fit (4)

- Solution to normal equations:

$$\begin{aligned} \text{intercept} \quad A &= \bar{Y} - B\bar{X} \\ \text{slope} \quad B &= \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \end{aligned}$$

- Formula for A tells that line passes through $(\bar{X}, \bar{Y})$: regression line is $Y = A + BX = (\bar{Y} - B\bar{X}) + BX = \bar{Y} + B(X - \bar{X})$ and fill in $\bar{X}$ for X; any line that passes through $(\bar{X}, \bar{Y})$ has sum of residuals equal to 0, as seen earlier.
- Second normal equation implies that $\sum X_i E_i = 0$, because $\sum X_i E_i = \sum X_i (Y_i - A - BX_i) = \sum X_i Y_i - A\sum X_i - B\sum X_i^2$, which equals zero according to the second normal equation. This tells that residuals and explanatory variable X are uncorrelated.
- Residuals are uncorrelated with fitted values $\hat{Y}$: $\sum \hat{Y}_i E_i = 0$ (check by filling in $\hat{Y}_i = A + BX_i$).

Simple Regression: Least-Squares Fit (5)

- Example in book on regression of measured weight (Y) on reported weight (X) gives least-squares regression equation $\hat{Y} = 1.78 + 0.977 \times X$
- Interpretation of slope B=0.977: 1 kg increase in reported weight is associated with, on average, 0.977 kg increase in measured weight.
- Interpretation of intercept A: the fitted value of Y at $X = 0$; in this example without meaning, because no person has reported weight 0.
- Intercept usually of little interest.

Simple Regression: Residual Variation

- Residual variance $S_E^2 = \frac{\sum E_i^2}{n-2}$ measures how close the line fits to the scatter of points.
- S_E^2 is average squared deviation (of observed Y_i from fitted Y_i), but denominator is $n - 2$ and not n (as would be the case in ordinary average).
- $n - 2$ are called "degrees of freedom for error"; in simple regression two "degrees of freedom" are subtracted from n, because two parameters (intercept and slope) have been estimated first.
- Residual standard deviation $S_E = \sqrt{\frac{\sum E_i^2}{n-2}}$ is sometimes called standard error of the regression; it is better to reserve the word standard error for the estimated standard deviation of the sampling distribution of a statistic, like standard error of the mean.
- In weight example $S_E = \sqrt{\frac{418.87}{101-2}} = 2.06$ kg; predicting measured weight from reported weight leads to error of about 2.06 kg; if residuals are approximately normally distributed, 68% are in range $\pm 1 \times S_E \approx \pm 2$kg, and 95% are in range $\pm 2 \times S_E \approx 4$kg.
- S_E is absolute measure of goodness of fit of regression.

Simple Regression: R-square

- Another way to measure goodness of fit, is by comparing the current model, containing regressor X, with the model without X; this gives a relative measure of goodness of fit.
- Model without X is $Y_i = A' + E_i'$; hence model with only intercept, also called "null model" or "intercept-only model".
- Estimate A' by least squares:
  - minimize residual sum of squares $\sum E_i'^2 = \sum (Y_i - A')^2$
  - resulting in $A' = \bar{Y}$ (check!).
  - So, in intercept-only model the least-squares estimate of the intercept is the ordinary mean of Y!

Simple Regression: R-square (2)

- Now we have two sums of squares:
  - from regression of Y on X: $\sum(Y_i - \hat{Y}_i)^2$, the residual sum of squares of Y, or RSS.
  - from the null model: $\sum(Y_i - \bar{Y})^2$, the total (corrected) sum of squares of Y, or TSS.
- Necessarily $\sum(Y_i - \hat{Y}_i)^2 \leq \sum(Y_i - \bar{Y})^2$ (why?)
- Define regression sum of squares as $RegSS \equiv TSS - RSS$.
- R-square is proportional reduction in squared error

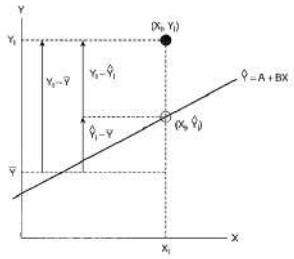
$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}$$

.

- Other names: R-square = fraction explained variation = coefficient of determination
- Correlation coefficient r of X and Y: r = square root of R^2, with sign according to slope of regression.

Simple Regression: Analysis of Variance

- Figure shows decomposition of deviations
  $Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$
- Squaring and summing gives
  $\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2 + 2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$
- Last term $2\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$, because $\sum(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum E_i(\hat{Y}_i - \bar{Y}) = \sum E_i\hat{Y}_i - \bar{Y}\sum E_i = 0$ as residuals and fitted values are uncorrelated and sum of residuals is zero.



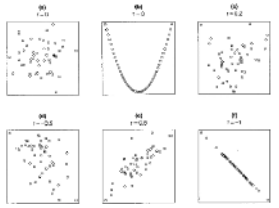
- So, total sum of squares around the mean can be written as sum of squares of fitted values around the mean and sum of squares of residuals
  $\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$

Simple Regression: Analysis of Variance (2)

- $\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$
- Regression sum of squares $RegSS$, earlier defined as $TSS - RSS$, must be $RegSS = \sum(\hat{Y}_i - \bar{Y})^2$.
- Analysis of variance for regression: $TSS = RegSS + RSS$
- Total sum of squares is split into two components. One component tells which part is "explained" by the model, due to regression. Second component tells which part remains "unexplained".

Simple Regression: correlation

- Correlation coefficient r is square root of R^2, with sign according to slope of regression.
- Check out examples in figure.



- Analogous to correlation between two random variables X and Y, $\rho = \sigma_{XY}/\sigma_X\sigma_Y$, sample correlation r can be defined as

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

- Note that for r it is irrelevant whether Y or X is response variable; not so for regression!
- Slope B measured in units of response per unit of explanatory variable. Correlation coefficient r is unitless.
- What happens to B if Y is divided by 1000 (like income Y expressed in thousands of euros instead of euros)? What happens to r?

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 1, lecture 5

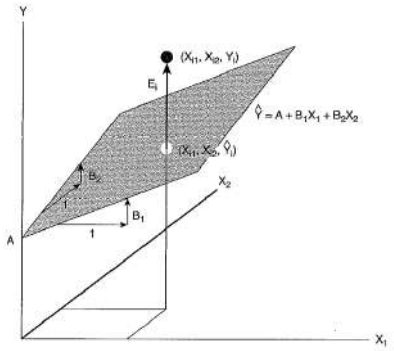
Content week 1, lecture 5: §5.2 Fox book

In chapter 5 the regression model is not yet fully described (assumptions not specified); §5.1 describes the simple regression case (one regressor), §5.2 the multiple regression case (more than one regressor). Chapter 6 introduces formal regression model, leading to statistical inference.

- Multiple regression (first two, later more than two regressors):
  - Regression equation and regression plane
  - Residuals, least-squares and normal equations by differentiation
  - Least-square estimators for slopes and intercept
  - Interpretation slopes and intercept
  - Residual standard error and degrees of freedom for error
  - Analysis of variance: $TSS = RegSS + RSS$
  - R^2 and R^2_{adj}
  - Multiple correlation coefficient $= \sqrt{R^2}$
  - Standardized regression coefficients

Multiple Regression: Two Explanatory Variables (1)

- Again, we start with quantitative dependent variable Y. Now there are two regressors: X_1 and X_2.
- With 2 regressors X_1 and X_2, regression equation is $\hat{Y} = A + B_1X_1 + B_2X_2$.
- Describes plane in 3-dimensional space, see figure.



- Residual $E_i = Y_i - \hat{Y}_i = Y_i - A - B_1X_1 - B_2X_2$
- As before, minimize $S(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1X_1 - B_2X_2)^2$, e.g. by differentation w.r.t. A, B_1, and B_2, and solve.

Multiple Regression: Two Explanatory Variables (2)

- Leads again to normal equations, now 3 linear equations with 3 parameters:

$$\begin{aligned} An + B_1 \sum X_{i1} + B_2 \sum X_{i2} &= \sum Y_i \\ A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1}X_{i2} &= \sum X_{i1}Y_i \\ A \sum X_{i2} + B_1 \sum X_{i1}X_{i2} + B_2 \sum X_{i2}^2 &= \sum X_{i2}Y_i \end{aligned}$$

- Explicit expressions for A, B_1, and B_2 can be written down, but we will not do that here.

Multiple Regression: Two Explanatory Variables (3)

- Least-squares coefficients are uniquely defined if $\sum X_1^{*2} X_2^{*2} \neq (\sum X_1^* X_2^*)^2$ with X^* the deviation from the mean $X_i - \bar{X}$. This condition is satisfied unless
  - X_1 and X_2 correlate perfectedly: they are collinear.
  - X_1 or X_2 do no vary.
- Interpretation of coefficients: B_1 and B_2 are partial coefficients, e.g. B_1 represents the average change in Y for a unit change in X_1, *holding value(s) of other explanatory variable(s) constant*.
- Check this interpretation algebraically, by comparing the $\hat{Y}$s for unit increase in X_1 at fixed value of $X_2 = x_2$:
  $[A + B_1(X_1 + 1) + B_2 x_2] - [A + B_1 X_1 + B_2 x_2] = B_1$

Example Multiple Regression: Two Explanantory Variables

- Example (Duncan's data: regression of prestige of occupation on education and income)

```
> Duncanreg <- lm(prestige ~ education + income, data=Duncan)
> coef(Duncanreg)
(Intercept)   education      income
     -6.065       0.546       0.599
```

- Unit increase in education, holding income constant, associated with increase of 0.55 units of prestige (which was percentage of respondents rating prestige of occupation as good or excellent). Unit increase in income, holding education constant, associated with increase of 0.60 of prestige. Intercept -6.1 does not have sensible interpretation, because there are no occupations with education value equal to zero and income value equal to zero (minimum value is 7 for both regressors).

Multiple Regression: Several Explanatory Variables

- Extension to more than 2 regressors is straightforward.
- Again quantitative dependent variable Y, but k regressors $X_1 \ldots X_k$.
- Minimize residual sum of squares
  $S(A, B_1, \ldots, B_k) = \sum E_i^2 = \sum (Y_i - (A + B_1 X_1 + \ldots + B_k X_k))^2$
- Result again normal equations, $k+1$ equations in $k+1$ unknown parameters.
- Unique solution exists unless one regressor is perfect linear combination of others, or regressor is invariant.

Example Multiple Regression: Several Explanatory Variables

- Example Canadian prestige data in dataframe `Prestige`
- Besides: regressors are defined differently compared to `Duncan` data: now `education` is average education years per occupation, `income` is average income per occupation.

```
> CanPrestige <- lm(prestige ~ education + income + women, data=Prestige)
> coef(CanPrestige)
(Intercept)   education      income       women
   -6.79433     4.18664    0.00131   -0.00891
```

- `prestige` is percentage, ranging 14.8 - 87.2, $IQR = 24.1$; `education` measured in years; `income` in dollars; `women` is percentage women in profession. Interpretation of coefficients: impact of education seems considerable: one year more education associated with increase in prestige of 4.2, holding others constant. Partial effect of income looks considerable too: 1 $ higher income associated with 0.0013 higher prestige, so 1.3 for each $1000; impact of gender is rather small. To judge the importance is rather difficult without information on range of values of `education` and `income`.

Multiple Regression: S_E and ANOVA

- Residual standard deviation $S_E = \sqrt{\frac{\sum E_i^2}{n-(k+1)}}$.
- Note denominator: degrees of freedom for error is $n-(k+1)$; $k+1$ degrees of freedom are "lost" because $k+1$ coefficients (intercept and k slopes) are estimated first.
- ANOVA decomposition of sums of squares as before:

$$TSS = RegSS + RSS$$

- $TSS = \sum(Y_i - \bar{Y})^2$ with $n-1$ df
- $RegSS = \sum(\hat{Y}_i - \bar{Y})^2$ with k df
- $RSS = \sum(Y_i - \hat{Y}_i)^2$ with $n-(k+1)$ df
- note that df add up: $n-1 = k+n-(k+1)$

Multiple Regression: R-squares and Multiple Correlation

- R-square defined as before:

$$R^2 \equiv \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  - "coefficient of determination"
  - R^2 is proportion of variation in Y that is captured by linear regression on X's.
- multiple correlation coefficient $= \sqrt{R^2} = r(Y, \hat{Y})$
- adjusted R-square

$$R^2_{adj} = 1 - \frac{RSS/(n-(k+1))}{TSS/(n-1)} = 1 - \frac{S_E^2}{TSS/(n-1)}$$

  - R^2_{adj} is alternative to R^2.
  - R^2 can never decrease if a regressor is added to model; R^2_{adj} can decrease if regressor is relatively unimportant. Therefore, R^2_{adj} may be used as criterion for model selection.

Example Multiple Regression

Duncan's regression of occupational prestige on education and income. Some "hand calculations".

```
> Duncanreg <- lm(prestige ~ education + income, data=Duncan)
> Y <- Duncan$prestige
> Yhat <- fitted(Duncanreg)
> Ybar <- mean(Y)
> Resid <- residuals(Duncanreg)
> (TSS <- sum((Y-Ybar)^2))
[1] 43688
> (RegSS <- sum((Yhat-Ybar)^2))
[1] 36181
> (RSS <- sum(Resid^2))
[1] 7507
> n <- nrow(Duncan); k <- 2
> (SE <- sqrt(RSS/(n-(k+1))))
[1] 13.4
> (Rsq <- RegSS/TSS); (adjRsq <- 1-(RSS/(n-(k+1))/(TSS/(n-1))))
[1] 0.828
[1] 0.82
```


Example Multiple Regression

```
> Duncanreg <- lm(prestige ~ education + income, data=Duncan)
> summary(Duncanreg)
Call:
lm(formula = prestige ~ education + income, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-29.54  -6.42   0.65   6.61  34.64

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0647     4.2719   -1.42     0.16
education     0.5458     0.0983    5.56  1.7e-06
income        0.5987     0.1197    5.00  1.1e-05

Residual standard error: 13.4 on 42 degrees of freedom
Multiple R-squared:  0.828,        Adjusted R-squared:  0.82
F-statistic:  101 on 2 and 42 DF,  p-value: <2e-16
> names(summary(Duncanreg))
 [1] "call"          "terms"         "residuals"     "coefficients"  "aliased"
 [6] "sigma"         "df"            "r.squared"     "adj.r.squared" "fstatistic"
[11] "cov.unscaled"
> summary(Duncanreg)$r.squared
[1] 0.828
> summary(Duncanreg)$adj.r.squared
[1] 0.82
> summary(Duncanreg)$sigma
[1] 13.4
```

Standardized Regression Coefficients (1)

- Suppose coefficients of different regressors are to be compared, but...
- regressors not measured in same units on same scale ("incommensurable"), making comparison difficult.
- Standardized coefficients are a way out: rescale regression coefficients according to measure of regressor spread.
- For example: Prestige data; which regressor has largest effect?
  $\widehat{\text{prestige}} = -6.794 + 4.187 \times \textit{education} + 0.00131 \times \textit{income} - 0.00891 \times \textit{gender}$
  $IQR(\textit{education}) = 4.2025$ years, $IQR(\textit{income}) = 4081.3$ dollars,
  $IQR(\textit{gender}) = 48.61\%$
  Changing each regressor over its IQR, holding remaining regressors constant, results in average changes in prestige:
  - Education: $4.2025 \times 4.187 = 17.59$
  - Income: $4081.1 \times 0.00131 = 5.361$
  - Gender: $48.61 \times -0.00891 = -0.4329$

  Thus education has larger effect than income; effect of gender small.
- Usually standardization is done using standard deviations of regressors, not IQR, and y is standardized as well.

Standardized Regression Coefficients (2)

- Start with multiple linear regression model $Y_i = A + B_1 X_{i1} + \ldots + B_k X_{ik} + E_i$.
- To get standardized coefficients:
  - Eliminate intercepts by subtracting $\bar{Y} = A + B_1 \bar{X}_1 + \ldots + B_k \bar{X}_k$.
  - This results in $Y_i - \bar{Y} = B_1(X_{i1} - \bar{X}_1) + \ldots + B_k(X_{ik} - \bar{X}_k) + E_i$;
    this is regression on centered variables.
  - Standardize: $\frac{Y_i - \bar{Y}}{S_Y} = (B_1 \frac{S_1}{S_Y}) \frac{(X_{i1} - \bar{X}_1)}{S_1} + \ldots + (B_k \frac{S_k}{S_Y}) \frac{(X_{ik} - \bar{X}_k)}{S_k} + \frac{E_i}{S_Y}$.
  - Rewrite: $Z_{iY} = B_1^* Z_{i1} + \ldots + B_k^* Z_{ik} + E_i^*$ with Z standardized variables.
- Standardized partial regression coefficient for j-th regressor is $B_j^* \equiv B_j(S_j/S_Y)$.

Example Standardized Regression Coefficients

- Example:

```
> OrdinaryCoef <- lm(prestige ~ education + income + women, data=Prestige)
> coef(OrdinaryCoef)
(Intercept)   education      income       women
   -6.79433     4.18664     0.00131    -0.00891
> options(digits=5)
> round(sd(Prestige$prestige),5)
[1] 17.204
> sd(Prestige$education)
[1] 2.7284
> sd(Prestige$income)
[1] 4245.9
> sd(Prestige$women)
[1] 31.725
```

  - Education: $4.187 \times 2.7284/17.204 = 0.664$
  - Income: $0.00131 \times 4245.9/17.204 = 0.3242$
  - Gender: $-0.00891 \times 31.725/17.204 = -0.01642$
- Note that standard deviations are difficult to justify in case of non-normal distributions, though.

```
> StCoef <- lm(scale(prestige) ~ -1 + scale(education) + scale(income) + scale(women),
+              data=Prestige) # What is scale doing?
> coef(StCoef)
scale(education)    scale(income)     scale(women)
        0.663955         0.324176        -0.016421
```


Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 1, lecture 6

Content week 1, lecture 6: Fox §6.1 + 6.2 partly

- Statistical inference for regression, first for simple, next for multiple regression
  - Formal linear regression model with assumptions
  - Errors: constant variance, Gaussian with expectation 0, independent; X without error
  - Properties least-squares estimators
  - Variance and standard error of slope and intercept
  - Hypothesis tests for individual slopes; t-distribution
  - Confidence intervals for individual slopes
  - Variance-inflation factor in multiple regression

Statistical Inference for Regression: simple linear regression

So far, linear regression has been described as descriptive technique. Now we discuss statistical inference, where we need a formal modeling approach:

- Statistical model for simple regression:

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

- Y_i is value of response variable Y on ith observation (out of n), x_i is value of explanatory variable X for ith observation; x_i is (assumed to be) known (hence small letter).
- α, β are population parameters, intercept and slope, to be estimated from the data.
- ϵ_i are the errors: aggregated omitted causes of Y, other missing explanatory variables, measurement error in Y, or other random components of Y.
- Assumptions w.r.t. errors ϵ_i are needed.
- Equivalently, assumptions w.r.t. distribution of Y conditional on X may be formulated.

Assumptions Simple Linear Regression

Assumptions for errors ϵ_i (or for Y conditional on X):

- Linearity: $E(\epsilon_i) \equiv E(\epsilon|x_i) = 0$, i.e. the average value of error is 0 (given the value of X);
  in terms of Y: $\mu_i \equiv E(Y_i) \equiv E(Y|x_i) = E(\alpha + \beta x_i + \epsilon_i) = \alpha + \beta x_i$
- Constant variance: $V(\epsilon|x_i) = \sigma^2_\epsilon$, i.e. variance of error is constant regardless value of X;
  in terms of Y: $V(Y|x_i) = \sigma^2_\epsilon$.
- Normality: errors are normally distributed $\epsilon_i \sim N(0, \sigma^2_\epsilon)$;
  in terms of Y: $Y_i \sim N(\alpha + \beta x_i, \sigma^2_\epsilon)$
- Independence: observations sampled independently: any pair of error ϵ_i and ϵ_j (or conditional response-variables Y_i and Y_j) are independent for $i \neq j$. Needs to be justified by procedures of data collection.
- Fixed X or X measured without error and independent of error. E.g. in experimental research X is usually fixed; in social research X is usually observed (sampled), and we assume that X is measured without error, and explanatory variable and error are independent.

Properties Least-Squares Estimator (1)

Under strong assumptions of simple linear regression, least-squares estimators have desirable properties as estimators of population coefficients α and β:

- Least-squares estimators are linear estimators in Y_i:
  e.g. $B = \sum_{i=1}^n m_i Y_i$, with $m_i = \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}$. Derivation of sampling distributions of A and B is simple now!
- Least-squares estimators are unbiased estimators of population regression coefficients: $E(A) = \alpha$ and $E(B) = \beta$.
- Variance of (the sampling distribution of) A and of B:

$$V(A) = \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} \sigma^2_\epsilon$$

.

$$V(B) = \frac{1}{\sum (x_i - \bar{x})^2} \sigma^2_\epsilon = \frac{1}{(n-1) S^2_X} \sigma^2_\epsilon$$

.

Note that $V(B)$ is small if 1) error variance σ^2_ϵ is small; 2) n is large; 3) X-values are widely spread (large S^2_X).

Properties Least-Squares Estimator (2)

- Least-squares estimators are most efficient of all linear unbiased estimators, i.e. have smallest variance (Gauss-Markov property).
- Under normality, least-squares estimators are most efficient of *all* unbiased estimators, not just the linear ones.
- Under all mentioned assumptions, least-squares coefficients A and B are the maximum likelihood estimators of α and β.
- Under normality assumptions, least-squares coefficients are themselves normally distributed: $A \sim N(\alpha, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma_\epsilon^2)$, $B \sim N(\beta, \frac{1}{\sum(x_i-\bar{x})^2}\sigma_\epsilon^2)$.

Sampling Distributions of Intercept and Slope

- Sampling distributions of A and of B:

$$A \sim N(\alpha, \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}\sigma_\epsilon^2)$$

$$B \sim N(\beta, \frac{1}{\sum(x_i-\bar{x})^2}\sigma_\epsilon^2)$$

- Therefore, hypothesis tests and confidence intervals for A and B could be based upon the (standard) normal distribution if σ_ϵ^2 would be known.
- However, σ_ϵ^2 is unknown, and needs to be estimated, using the residual variance $\hat{\sigma}_\epsilon^2 = S_E^2 = \frac{\sum E_i^2}{n-2}$.
- Estimated sampling variances are

$$\hat{V}(A) = \frac{\sum x_i^2}{n\sum(x_i-\bar{x})^2}S_E^2$$

$$\hat{V}(B) = \frac{1}{\sum(x_i-\bar{x})^2}S_E^2$$

t-distribution instead of normal distribution

- For inference, t-distributions are used instead of normal distribution, due to estimation of error variance.
- Recall (see e.g. Fox appendix, page 81) that if $Z \sim N(0,1)$, $X^2 \sim \chi_\nu^2$ and Z and X^2 independent, then $Z/\sqrt{X^2/\nu} \sim t_\nu$, i.e. has a t-distribution with ν degrees of freedom.
- Here, for slope β, $Z = (B-\beta)/\sigma_B \sim N(0,1)$ and $X^2 = (n-2)S_E^2/\sigma_\epsilon^2 \sim \chi_{n-2}^2$.
- The t_ν distribution has expected value 0 and variance $\nu/(\nu-2)$. The distribution is wider than the standard normal distribution. For $\nu \to \infty$ the t_ν distribution approaches the standard normal distribution.

Confidence Interval and Hypothesis Test for Slope

- $100(1-\alpha)\%$ confidence interval for slope is

$$B \pm t_{n-2;\alpha/2}SE(B)$$

with the standard error of the slope $SE(B) = \sqrt{\hat{V}(B)} = S_E/\sqrt{\sum(x_i-\bar{x})^2}$. $t_{n-2;\alpha/2}$ the $1-\alpha/2$ quantile from the t_{n-2} distribution.

- To test $H_0 : \beta = \beta_0$ use test statistic

$$t = \frac{B-\beta_0}{SE(B)}$$

which has under H_0 t-distribution with $n-2$ d.f.
This test statistic is the difference between estimate B of slope β and the hypothesized value (β_0) of the slope under the null hypothesis, expressed in standard errors of the slope. The typical test is testing whether the slope is zero, so $\beta_0 = 0$.

Example Confidence Intervals and Hypothesis Tests

```
> DavisF <- Davis[Davis$sex=="F" & Davis$weight<160 & !is.na(Davis$repwt),]  # clean up first
> DavisReg <- lm(weight ~ repwt, data=DavisF) # function lm fits linear model
> summary(DavisReg)[[4]] # 4th component of list; also try coef(summary(DavisReg))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.768     1.7532    1.01 3.16e-01
repwt          0.977     0.0307   31.86 1.66e-53
> summary(DavisReg)$sigma
[1] 2.07
> confint(DavisReg)  # confidence intervals for regression coefficients
             2.5 % 97.5 %
(Intercept) -1.711   5.25
repwt        0.916   1.04
```


Example Confidence Intervals and Hypothesis Tests

Now reproduce results by "hand" calculations for regression of observed weight on reported weight.

```
> Y <- DavisF$weight; X <- DavisF$repwt
> sumXY <- sum((X-mean(X))*(Y-mean(Y))); sumXX <- sum((X-mean(X))^2); sumYY <- sum((Y-mean(Y))^
> n<-nrow(DavisF)
> (B <- sumXY/sumXX)
[1] 0.977
> (A <- mean(Y) - B*mean(X))
[1] 1.77
> Yhat <- A+B*X; Resid <- Y-Yhat; RSS <- sum(Resid^2)
> (SE <- sqrt(RSS/(n-2)))
[1] 2.07
> (SEB <- SE/sqrt(sumXX))
[1] 0.0307
> (SEA <- SE*(sqrt(sum(X^2)/n))/(sqrt(sumXX)))
[1] 1.75
> t <- qt(0.975,n-2)
> (B.low <- B-t*SEB)
[1] 0.916
> (B.up  <- B+t*SEB)
[1] 1.04
```


Multiple Regression Model

Model $Y_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$.
k regressors, $n - (k+1)$ degrees of freedom for error.

- Assumptions identical to simple regression situation:
  - linearity: $E(\epsilon_i) = 0$
  - constant variance: $V(\epsilon_i) = \sigma^2_\epsilon$
  - normality: $\epsilon_i \sim N(0, \sigma^2_\epsilon)$
  - independence: ϵ_i and ϵ_j are independent for $i \neq j$
  - fixed Xs or Xs measured without error and independent of ϵ
  - further: Xs not invariant, and no X is perfect linear combination of others.
- Under these conditions least-squares estimators $B_1, \ldots, B_k$ are
  - linear functions of Y_i
  - unbiased
  - maximally efficient among unbiased estimators
  - maximum likelihood estimators
  - normally distributed.

Variance of Slope in Multiple Regression Model

- Slope B_j has variance:

$$V(B_j) = \frac{1}{1 - R^2_j} \times \frac{\sigma^2_\epsilon}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2} = \frac{\sigma^2_\epsilon}{\sum_{i=1}^n (x_{ij} - \hat{x}_{ij})^2}$$

  where R^2_j is R-square of the regression of X_j on all other Xs, and $\hat{x}_{ij}$ are fitted values of this auxiliary regression.

- Factor $1/(1 - R^2_j)$ is called the variance inflation factor=VIF.
- VIF is large if X_j is strongly correlated with other Xs.

Confidence Intervals and Hypothesis Tests: Individual Slopes

For individual coefficients procedures are identical to those in simple regression case, but...

- estimator of variance $\hat{\sigma}^2_\epsilon = S^2_E = \frac{\sum E_i^2}{n-(k+1)}$
- standard error of B_j is $SE(B_j) = \frac{1}{\sqrt{1-R^2_j}} \times \frac{S_E}{\sqrt{\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2}}$
- t-intervals and t-tests as before.

Example multiple regression of prestige on education and income.

```
> Duncanreg <- lm(prestige ~ education + income, data=Duncan)
> summary(Duncanreg)$sigma^2
[1] 179
> summary(Duncanreg)$coefficients
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.065     4.2719   -1.42 1.63e-01
education      0.546     0.0983    5.56 1.73e-06
income         0.599     0.1197    5.00 1.05e-05
> confint(Duncanreg)
              2.5 % 97.5 %
(Intercept) -14.686  2.556
education     0.348  0.744
income        0.357  0.840
```


Example Confidence Intervals and Hypothesis Tests: Individual Slopes

"Hand" calculations

```
> Y <- Duncan$prestige; X1 <- Duncan$education; X2 <- Duncan$income
> regY.X1X2 <- lm(Y~X1+X2); B1 <- coef(regY.X1X2)[2]; B2 <- coef(regY.X1X2)[3]
> Yhat <- regY.X1X2$fitted.values; resid <- Y-Yhat;
> SE2 <- sum(resid^2)/regY.X1X2$df.residual
> regX1.X2 <- lm(X1 ~ X2); regX2.X1 <- lm(X2 ~ X1)
> (R1.2 <- summary(regX1.X2)$r.squared)
[1] 0.525
> (R2.1 <- summary(regX2.X1)$r.squared)
[1] 0.525
> (VIF1 <- 1/(1-R1.2))
[1] 2.1
> (VIF2 <- 1/(1-R2.1))
[1] 2.1
> (SEB1 <- sqrt(VIF1* SE2 /sum((X1-mean(X1))^2)))
[1] 0.0983
> (SEB2 <- sqrt(VIF2* SE2 /sum((X2-mean(X2))^2)))
[1] 0.12
> t <- qt(0.975,regY.X1X2$df.residual)
> cat("95% ci for B1:", B1 - t*SEB1, B1 + t*SEB1,"\n")
95% ci for B1: 0.348 0.744
> cat("95% ci for B2:", B2 - t*SEB2, B2 + t*SEB2,"\n")
95% ci for B2: 0.357 0.84
```


Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 2, lecture 7

Content week 2, lecture 7: Multiple Regression: examples t-test for slope; §6.2 Fox book: F-test

- Example t-test for individual slope in multiple regression (H_a two-sided and one-sided)
- Statistical inference for multiple regression (continued)
  - F-test for all slopes: omnibus F-test; F-distribution
  - F-test for subset of slopes: F-test based upon incremental (extra) SS for nested models
  - Example F-test

Multiple Regression Model

Recall multiple regression model for response Y_i and k regressors $x_1 \dots x_k$:

$$Y_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

with assumptions for errors

- linearity: $E(\epsilon_i) = 0$;
- constant variance: $V(\epsilon_i) = \sigma_\epsilon^2$;
- normality: $\epsilon_i \sim N(0, \sigma_\epsilon^2)$;
- independence: ϵ_i and ϵ_j are independent for $i \neq j$;

Example multiple regression

Dataset `Duncan` on prestige of 45 occupations, to be explained by education and income.

```
> head(Duncan[,2:4])
           income education prestige
accountant     62        86       82
pilot          72        76       83
architect      75        92       90
author         55        90       76
chemist        64        86       90
minister       21        84       87
```

Model: $prestige_i = \alpha + \beta_1 education_i + \beta_2 income_i + \epsilon_i$
(i index for observations, $i = 1, \dots 45$)

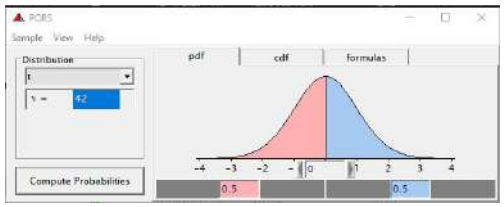
Least-squares estimates and t-tests for individual slopes:

```
> Duncanreg <- lm(prestige ~ education + income, data=Duncan)
> coef(summary(Duncanreg))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.065     4.2719   -1.42 1.63e-01
education      0.546     0.0983    5.56 1.73e-06
income         0.599     0.1197    5.00 1.05e-05
```


Example t-test for individual slope, two-sided H_a

Use significance level $\alpha = 0.05$ if not explicitly mentioned otherwise.
9 steps in hypothesis testing for slope of education:

1. $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 \neq 0$ In words: H_0: "education is not related to prestige (keeping income constant)", H_a: "education is related to prestige (keeping income constant)".
2. Test statistic: $t = \frac{B_1 - 0}{SE(B_1)}$ with B_1 the l.s. estimator of β_1.
3. If H_0 is true: $t \sim t_{42}$ ($42 = 45 - 3$, degrees of freedom for error)

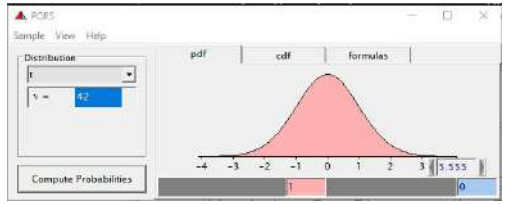


4. If H_a is true: t tends to smaller (if $\beta_1 < 0$) or larger (if $\beta_1 > 0$) values than prescribed by t_{42} distribution.
5. Two-tailed P-value is needed: $P = 2 \times P(t_{42} \geq |t|)$ (identical to $P = P(t_{42} \leq -|t|) + P(t_{42} \geq |t|)$)

Example t-test for individual slope, two-sided H_a (cont)

Steps 6-9 in hypothesis testing for slope of education:

6. Outcome of test statistic (R output previous slide): $t = \frac{0.546 - 0}{0.0983} = 5.555$
7. $P = 2 \times P(t_{42} \geq |5.555|)$ $= 2 \times 8.65 \times 10^{-7} =$ 1.73×10^{-6}



8. Conclusion: $P < 0.05$, so reject H_0
9. Conclusion in words: the data contains evidence that education is related to prestige (keeping income constant).

R reports two-tailed P-value for t-test of slope.

Example t-test for individual slope, one-sided H_a

Suppose we are very sure that a possible relationship, if existing, can only be positive. It makes sense then to test with a right-sided H_a.
Steps in hypothesis testing for slope of education:

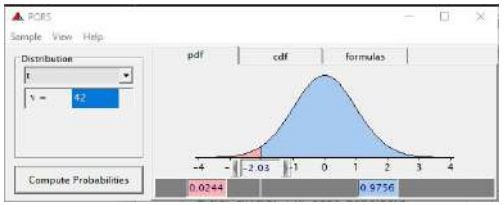
1. $H_0 : \beta_1 = 0$ versus $H_a : \beta_1 > 0$ In words: H_0: "education is not related to prestige (keeping income constant)", H_a: "education is positively related to prestige (keeping income constant)"
2. Test statistic: $t = \frac{B_1 - 0}{SE(B_1)}$
3. If H_0 is true: $t \sim t_{42}$
4. If H_a is true: t tends to larger values than prescribed by t_{42} distribution.
5. Right-tailed P-value is needed: $P = P(t_{42} \geq t)$
6. Outcome of test statistic (R output previous slide): $t = \frac{0.546 - 0}{0.0983} = 5.555$
7. $P = P(t_{42} \geq 5.555) = 8.65 \times 10^{-7}$
8. Conclusion: $P < 0.05$, so reject H_0
9. Conclusion in words: education is positively related to prestige (keeping income constant).

Here we can take half the P-value as reported by R.

Example t-test for individual slope, one-sided H_a (2)

Can two-tailed P-value, as reported by R, always be halved for one-sided H_a? No...

- In example outcome is in "correct" tail for right-sided H_a; therefore, P-value reported by R can be halved: $\frac{1}{2} \times 1.7 \times 10^{-6}$.
- What if outcome of test statistic is in "wrong" tail?
- Suppose slope estimate for education is negative, e.g. $B_1 = -0.2$, but we want to show $H_a : \beta_1 > 0$. How is testing affected?
- R would report two-tailed P-value: outcome test stat $t = \frac{-0.2 - 0}{0.0983} = -2.03$; $P = 2 \times P(t_{42} \geq |-2.03|) = 2 \times P(t_{42} \geq 2.03) = 2 \times 0.0244 = 0.048$.
- But, under H_a test statistic t expected to have larger values than t_{42} prescribes, so P-value still is $P(t_{42} \geq t) = P(t_{42} \geq -2.03) = 0.976$.



- Relation with P-value from R: $0.976 = 1 - 0.048/2$.

Example t-test for individual slope: $H_0 : \beta = \beta_0$

Suppose test for null value other than 0 is needed. R cannot be directly used (unless using some trick). Imagine that value 0.5 has some special meaning in education example, so we ask whether β_1 might be equal to 0.5 (given the data): $H_0 : \beta_1 = 0.5$.
Steps in hypothesis test for slope of education:

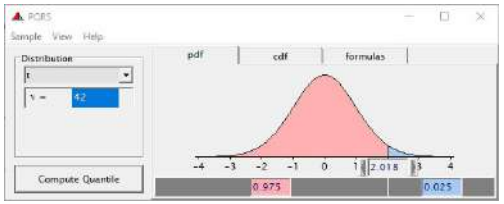
1. $H_0 : \beta_1 = 0.5$ versus $H_a : \beta_1 \neq 0.5$
2. Test statistic: $t = \frac{B_1 - 0.5}{SE(B_1)}$
3. If H_0 is true: $t \sim t_{42}$
4. If H_a is true: t tends to smaller (if $\beta_1 < 0$) or larger (if $\beta_1 > 0$) values than prescribed by t_{42} distribution.
5. Two-tailed P-value is needed: $P = 2 \times P(t_{42} \geq |t|)$
6. Outcome of test statistic (R output previous slide): $t = \frac{0.546 - 0.5}{0.0983} = 0.468$
7. $P = 2 \times P(t_{42} \geq |0.468|) = 2 \times 0.321 = 0.64$
8. Conclusion: $P > 0.05$, do not reject H_0
9. Conclusion in words: no evidence is found that the slope deviates from 0.5 (keeping income constant)

Example confidence interval for slope

- Point estimate for slope of education: $\hat{\beta}_1 = B_1 = 0.546$.
- $100(1-\alpha)\%$ confidence interval for slope is

$$B_1 \pm t_{45-3;\alpha/2} SE(B_1)$$

- 95% c.i. for β_1: $0.546 \pm t_{42;0.025} \times 0.0983 = 0.546 \pm 2.018 \times 0.0983 = (0.348, 0.744)$.



- Given the data, we are "95% confident" that the slope β_1 is in the interval $(0.348, 0.744)$.
- Connection confidence interval and hypothesis test: every value in 95% c.i. $(0.348, 0.744)$, if put to the test (using $\alpha = 0.05$), will not be rejected.

Hypothesis Test: All Slopes

Multiple regression model: $Y_i = \alpha + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$
- Global or "omnibus" test that all regressors are unimportant.
- F-test statistic: $F = \dfrac{RegSS/k}{RSS/(n-(k+1))}$
- Recall that $RegSS = TSS - RSS$, i.e. difference between residual sum of squares of the null model and current model. $RegSS$ tells how much better current model is compared to null model.
- Under H_0 F has F-distribution with k and $n-(k+1)$ d.f.
- Reject H_0 for large values of F, right-sided P-value and rejection region.
- Analysis of variance table or ANOVA table shows construction of F:

Source	*Sum of Squares*	*df*	*Mean Square*	*F*
Regression	$RegSS$	k	$\frac{RegSS}{k}$	$\frac{RegMS}{RMS}$
Residual	RSS	$n-(k+1)$	$\frac{RSS}{n-(k+1)}$	
Total	TSS	$n-1$		

Hypothesis Test: All Slopes

- F is ratio of two Mean Squares:
  1. denominator Residual Mean Square $RMS = S_E^2$ is estimator of error variance σ_ϵ^2.
  2. numerator Regression Mean Square is another estimator of σ_ϵ^2, but only if H_0 is true!

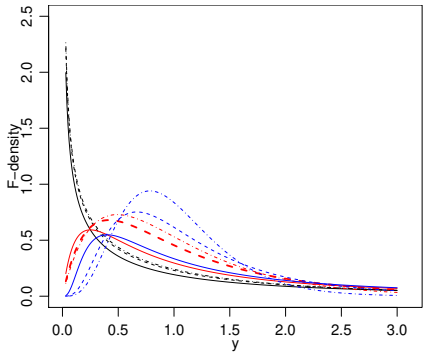
  Hence, under H_0, the ratio $RegMS/RMS$ is close to 1.
- Under H_a, $RegMS$ tends to be larger than σ_ϵ^2, so ratio $RegMS/RMS$ tends to be larger than 1.
- Reject H_0 if F is larger than critical value from F-distribution, or P-value smaller than chosen significance level.

Besides: F-distribution

- F-distribution obtained by taking ratio of two independent chi-square distributed variables, each divided by degrees of freedom: $F \equiv \frac{X_1^2/df_1}{X_2^2/df_2}$ (see Fox appendix p 82).
- F-distribution has degrees of freedom df_1 and df_2.
- F-distributed variable is positively valued.
- Expected value of F-distribution is $df_2/(df_2-2)$ (for $df_2 > 2$) ≈ 1 for large df_2.
- Square of t-distributed variable with df degrees of freedom has F-distribution with 1 and df degrees of freedom.
- In linear model situation: (informal) sums of squares / σ^2 have chi-square distributions, so ratio of mean squares (sum of squares / df), both estimating the same σ^2, has F-distribution (if sum of squares are independent).

Besides: examples F-distributions

```
> y <- seq(0.03, 3, by=0.01)
> F1.2  <- df(y,1,2);  F1.10  <- df(y,1,10);  F1.100  <- df(y,1,100)
> F4.2  <- df(y,4,2);  F4.10  <- df(y,4,10);  F4.100  <- df(y,4,100)
> F10.2 <- df(y,10,2); F10.10 <- df(y,10,10); F10.100 <- df(y,10,100)
> par(mar=c(3,3,0,1), mgp=c(1.5,0.75,0)); plot(c(0,2.5) ~ c(0,3.1), type="n",
>                                        ylab="F-density", xlab="y", lwd=2)
> lines(F1.2 ~ y, col="black", lty=1); lines(F1.10 ~ y, col="black", lty=2)
> lines(F1.100 ~ y, col="black", lty=4)
> lines(F4.2 ~ y, col="red",   lty=1); lines(F4.10 ~ y, col="red", lty=2, lwd=2)
> lines(F4.100 ~ y, col="red", lty=4)
> lines(F10.2 ~ y, col="blue", lty=1); lines(F10.10 ~ y, col="blue", lty=2)
> lines(F10.100 ~ y, col="blue", lty=4)
```



Example Hypothesis Test: All Slopes

```
> Duncanreg <- lm(prestige ~ education + income, data=Duncan)
> summary(Duncanreg)

Call:
lm(formula = prestige ~ education + income, data = Duncan)

Residuals:
    Min      1Q  Median      3Q     Max
-29.54  -6.42   0.65    6.61   34.64

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.0647     4.2719   -1.42     0.16
education     0.5458     0.0983    5.56  1.7e-06
income        0.5987     0.1197    5.00  1.1e-05

Residual standard error: 13.4 on 42 degrees of freedom
Multiple R-squared:  0.828,     Adjusted R-squared:  0.82
F-statistic:  101 on 2 and 42 DF,  p-value: <2e-16

> summary(Duncanreg)$fstatistic
value numdf dendf
  101     2    42
```

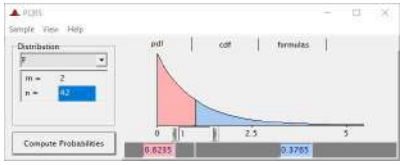

Example Hypothesis Test: All Slopes

```
> anova(Duncanreg)
Analysis of Variance Table

Response: prestige
          Df Sum Sq Mean Sq F value  Pr(>F)
education  1  31707   31707     177 < 2e-16
income     1   4474    4474      25 1.1e-05
Residuals 42   7507     179
> (RegSS <- sum((fitted.values(Duncanreg)-mean(fitted.values(Duncanreg)))^2))
[1] 36181
> (RSS <- deviance(Duncanreg)) # direct way to extract residual sum of squares
[1] 7507
> (F <- (RegSS/(Duncanreg$rank-1)) /(RSS/Duncanreg$df.residual))
[1] 101
```

R reports sums of squares of `education` and `income`.
Is sum of these two sums of squares equal to *RegSS*?
What would happen if order of `education` and `income` is reversed?

F-test in 9 steps

1. $H_0 : \beta_1 = \beta_2 = 0$ versus H_a: H_0 is not true
2. Test statistic: $F = \frac{RegSS/2}{RSS/(45-3)}$
3. If H_0 is true: $F \sim F_{2,42}$



4. If H_a is true: F tends to larger values than prescribed by $F_{2,42}$ distribution.
5. Right-tailed P-value needed: $P = P(F_{2,42} \geq F)$
6. Outcome test statistic: $F = \frac{36810/2}{7507/42} = 101.2$
7. $P = P(F_{2,42} \geq 101.2) = 1.1 \times 10^{-16}$
8. Conclusion: $P < 0.05$, so reject H_0
9. Conclusion in words: education and/or income are related to prestige.

Hypothesis Test: Subset of Slopes (1)

- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0$ with $1 \leq q \leq k$.
- For convenience, chose first q regressors, but any subset of βs may be tested.
- F-test is constructed by fitting two nested models:
  - Full model $FM : Y = \alpha + \beta_1 x_1 + \ldots + \beta_q x_q + \beta_{q+1} x_{q+1} + \ldots + \beta_k x_k + \epsilon$
  - Reduced model $RM : Y = \alpha + \beta_{q+1} x_{q+1} + \ldots + \beta_k x_k + \epsilon$
- FM and RM give residual SS's RSS_1 and RSS_0.
- Because the reduced model is nested within the full model (is special case of full model), RM fits worse than FM, and cannot have a smaller residual SS. Hence $RSS_0 \geq RSS_1$.
- F-statistic is $F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-(k+1))}$ with residual SS in numerator.
- By definition $F > 0$

Hypothesis Test: Subset of Slopes (2)

- F may be formulated using regression sums of squares, because $RSS_0 - RSS_1 = RegSS_1 - RegSS_0$, "Any increase in residual sum of squares, is decrease in regression sum of squares."
- So, $F = \frac{(RegSS_1 - RegSS_0)/q}{RSS_1/(n-(k+1))}$.

Example Hypothesis Test: Subset of Slopes

Test trivial $H_0 : \beta_1 = 0 \Leftrightarrow H_0$:"education has no association with prestige"

```
> FMreg <- lm(prestige ~ education + income, data=Duncan)
> RMreg <- lm(prestige ~ income, data=Duncan)
> (RSS1 <- anova(FMreg)[3,2]) # SS in second column of ANOVA table
[1] 7507
> (RSS0 <- anova(RMreg)[2,2])
[1] 13023
> (RSS1 <- deviance(FMreg)) # direct way to extract residual sum of squares
[1] 7507
> q <- 1
> (F <- ((RSS0-RSS1)/q)/(RSS1/FMreg$df.residual))
[1] 30.9
> (Pval <- 1-pf(F,q,FMreg$df.residual))
[1] 1.73e-06
```

F-test with $q = 1$ is equivalent to t-test:

```
> coef(summary(FMreg))[2,]
  Estimate Std. Error    t value   Pr(>|t|)
  5.46e-01   9.83e-02   5.56e+00   1.73e-06
```

Note that $t^2 = F$.

Empirical vs Structural Equations; Measurement Error in regressors (skip)

- Distinguish between two types of regression interpretations: descriptive or structural:
  - Descriptive: empirical association among variables.
  - Structural: causal relationships between variables.
- In structurally interpreted regression the regression model describes how response-variable is constructed: e.g. $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$, with all usual assumptions.
- In structural relations we can speak of bias produced by omitting a regressor that 1) is cause of Y and 2) is correlated with a regressor in the regression equation. Bias in least-squares estimation results from correlation between included regressor and the error, because the omitted regressor is incorporated in error.
- Measurement error in a regressor tends to attenuate (weaken) its regression coefficient and to make the variable an imperfect statistical control.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 2, lecture 8

Content week 2, lecture 8: Fox §7 until §7.3.4

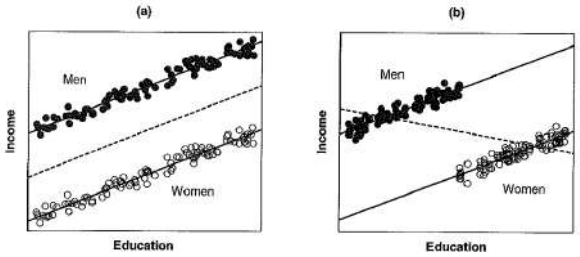
- Dummy variable regression (chapter 7)
  - dummies coding for factor levels
  - common-slope model
  - polytomous factors
  - non-parallel lines: interaction regressors
  - principle of marginality

Dummy-Variable Regression

- So far, in multiple linear regression a quantitative response was explained from quantitative explanatory variables.
- Often, qualitative explanatory variables, or factors, defining groups, need to be taken into account.
- Dummy variable regressors, or simply dummies, are used to do so.
- Dichotomous factor (binary factor, 2 levels) is represented by single dummy.
- Polytomous factor (>2 levels) is represented by set of dummies.

Dichotomous Factor

- One dichotomous factor and one quantitative explanatory variable.
- Assume additive relationship: partial effect of each explanatory variable does not depend on value of other explanatory variable(s).
- Example: relationship between education (x) and income (y) among women and men.
- Figure shows two situations, both with parallel within-gender regression lines, hence additive effects of education and gender.
  - Left figure: gender and education are unrelated; if gender was ignored, same slope would result, but size of error would be inflated.
  - Right figure: gender and education are related; ignoring gender, regression of income on education would result in biased assessment of education effect: slope would even become negative.



Common slope model (1)

- Possible solution: two separate regressions for women and women.
- Disadvantage:
  1. How to test for gender differences?
  2. If parallel regression is reasonable, more efficient estimation of common slope is achieved by pooling data from both groups.
- Better solution: common-slope model:

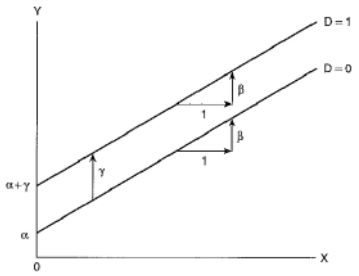
$$Y_i = \alpha + \beta X_i + \gamma D_i + \epsilon_i$$

with D = dummy = indicator variable, coded as $D_i = \begin{cases} 1 & \text{for men} \\ 0 & \text{for women} \end{cases}$

Common slope model (2)

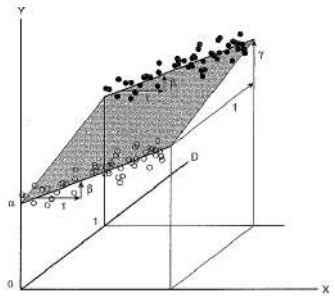
- Split model for women and men:
  - women: $Y_i = \alpha + \beta X_i + \gamma \times 0 + \epsilon_i = \alpha + \beta X_i + \epsilon_i$
  - men: $Y_i = \alpha + \beta X_i + \gamma \times 1 + \epsilon_i = (\alpha + \gamma) + \beta X_i + \epsilon_i$

  Women have regression line with intercept α, men have intercept $\alpha + \gamma$, and common slope β. Parameter γ is the difference in intercepts for men and women.



Common slope model (3)

- Common slope model is multiple regression model containing the rather special regressor dummy D.
- Plot below is 3-dimensional scatterplot for D, X, Y.
- Figure on the former slide is a projection onto the (X, Y) plane.



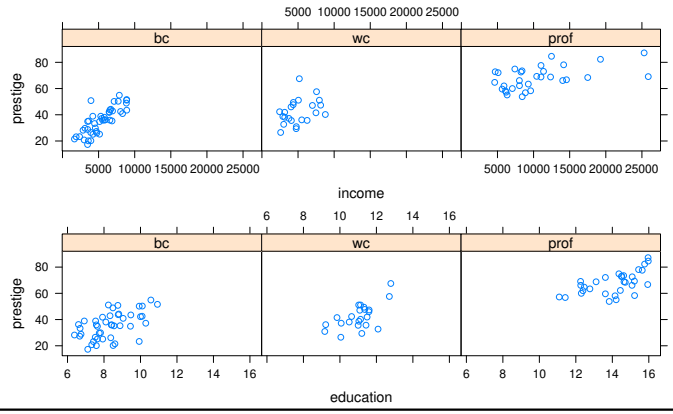
Common slope model (4)

- There is no special reason for D to be coded in the way we did. We could code $D = 0$ for men and $D = 1$ for women, making men the reference or baseline category. Sign of γ is reversed now, giving the difference in intercepts for women and men, whereas α represents intercept of reference group, men now.
- Testing gender effect, controlling for education, boils down to testing $H_0 : \gamma = 0$.
- So far, only dummy-variable regression with a single quantitative regressor is described, but this can easily extended to models with multiple quantitative regressors. For the common slope model the assumption is needed, that the slopes are the same in the two categories of the factor for each of the quantitative regressors.
- Fox makes distinction between explanatory variable and regressor: *gender* is explanatory variable, but D is regressor, used in the regression equation, and representing *gender*. *Education* is both explanatory variable and regressor.

Polytomous Factors (1)

- For polytomous factor with k levels, $k - 1$ dummy regressors are needed.
- Example `Prestige` dataset: occupational prestige as linear function of income and education years, for three types of occupation: professional, white collar and blue collar.

```
> type2 <- factor(Prestige$type, levels=c("bc","wc","prof")) # change level order to correspond
> require(lattice)
> print(xyplot(prestige ~ income    | type2, data=Prestige))  # in package lattice
> print(xyplot(prestige ~ education | type2, data=Prestige))
```



Polytomous Factors (2)

- Occupation type has 3 levels, 2 dummies are needed, e.g.:

Category	D_1	D_2
Professional and managerial	1	0
White collar	0	1
Blue collar	0	0

- Common slope model:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon$$

Polytomous Factors (3)

- Split for 3 groups:
  - Professional: $Y_i = (\alpha + \gamma_1) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon$
  - White collar: $Y_i = (\alpha + \gamma_2) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon$
  - Blue collar : $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon$
- Hence, α is intercept of regression model in reference group (blue collar), γ_1 is difference in intercepts between professional and reference, γ_2 is difference in intercepts between white collar and reference.
- Other group may be chosen as reference, redefine dummies accordingly.
- Sometimes "natural" reference group exists, like control group in experiment or clinical trial.
- Hypothesis test of interest: $H_0 : \gamma_1 = \gamma_2 = 0$. Use F-test based on incremental-sum-of-squares: fit nested Full and Restricted Models, and compare.

Polytomous Factors (4)

- Why not use 3 dummy variables for factor with 3 levels, like

Category	D_1	D_2	D_3
Professional and managerial	1	0	0
White collar	0	1	0
Blue collar	0	0	1

  In this case model for jth level is: $Y_i = (\alpha + \gamma_j) + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon$
- Problem is overparameterization: 4 parameters $(\alpha, \gamma_1, \gamma_2, \gamma_3)$ are used to describe only 3 intercepts. No unique values for 4 parameters exist: adding a constant to α and subtracting the same constant from γ_1, γ_2, γ_3 results in the same 3 intercepts. Another way of looking at the problem is: dummies are collinear: $D_3 = 1 - D_1 - D_2$.
- Problem of overparameterization is solved by exclusion of a parameter. Any of the γ_j may be removed, but also overall intercept α. What is the interpretation of γ_j in that case?

Polytomous Factors (5)

- Generalization to factors with m categories is straightforward: use $m - 1$ dummies. Choose f.i. last category as reference:

Category	D_1	D_2	...	D_{m-1}
1	1	0	$\cdots$	0
2	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$m-1$	0	0	$\cdots$	1
m	0	0	$\cdots$	0

Polytomous Factors (6)

- For multiple factors, and assuming *additive* effects, code set of dummies for each factor. For testing the effect of a factor, fit Full and Restricted Models, and use F-test.
- Example prestige continued. Fit parallel slope model.

```
> Pr2 <- Prestige[!is.na(Prestige$type),]  # remove some missing values
> #
> Y <- Pr2$prestige; X1 <- Pr2$income; X2 <- Pr2$education
> # Define two dummies
> D1 <- ifelse(Pr2$type=="prof",1,0)
> D2 <- ifelse(Pr2$type=="wc",1,0)
> #
> FM <- lm(prestige ~ income + education + D1 + D2, data=Pr2)
> coef(summary(FM))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.62293   5.227525  -0.119 9.05e-01
income       0.00101   0.000221   4.586 1.40e-05
education    3.67317   0.640502   5.735 1.21e-07
D1           6.03897   3.866855   1.562 1.22e-01
D2          -2.73723   2.513932  -1.089 2.79e-01
> summary(FM)$r.square
[1] 0.835
```

- Fitted model: $\hat{Y} = -0.6229 + 0.00101X_1 + 3.673X_2 + 6.039D_1 - 2.737D_2$

Polytomous Factors (7)

- Better let R do the work for you instead of defining dummies yourself.
- For factor like type, R by default makes dummies, and chooses first level as reference group. The levels of type are ordered as ("bc", "prof", "wc"), see below. So R chooses group blue collar as reference, as in Fox book.

```
> levels(Pr2$type)
[1] "bc"   "prof" "wc"
> FM2 <- lm(prestige ~ income + education + type, data=Pr2) # Automatic dummies for type
> coef(FM2)
(Intercept)      income   education    typeprof      typewc
   -0.62293     0.00101     3.67317     6.03897    -2.73723
```

- To change the reference group, you may redefine the factor, so that the group you want as reference becomes the first level of the factor. A more elegant approach is to use another "contrast" matrix.

```
> RefL <- contr.treatment(levels(Pr2$type), base=3) # wc becomes reference
> FM3 <- lm(prestige ~ income + education + type, contrasts=list(type = RefL), data=Pr2)
> coef(FM3)
(Intercept)      income   education      typebc    typeprof
   -3.36016     0.00101     3.67317     2.73723     8.77620
> model.matrix(FM3)[1:5,]  # First 5 lines of the model matrix
                    (Intercept) income education typebc typeprof
gov.administrators            1  12351      13.1      0        1
general.managers              1  25879      12.3      0        1
accountants                   1   9271      12.8      0        1
purchasing.officers           1   8865      11.4      0        1
chemists                      1   8403      14.6      0        1
```


Polytomous Factors (8)

Example continued

- Split into groups:
  - Professional: $\hat{Y} = 5.416 + 0.001013X_1 + 3.673X_2$
  - White collar: $\hat{Y} = -3.360 + 0.001013X_1 + 3.673X_2$
  - Blue collar: $\hat{Y} = -0.623 + 0.001013X_1 + 3.673X_2$
- Ordinary means are:

```
> tapply(Pr2$prestige, Pr2$type, mean)
  bc prof   wc
35.5 67.8 42.2
```

  Note that differences between means are quite different from differences between intercepts. Reason?

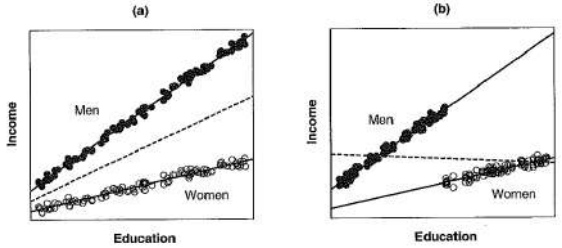
- To test $H_0 : \gamma_1 = \gamma_2 = 0$ fit FM and RM:

```
> RM <- lm(prestige ~ income + education, data=Pr2)
> RSS1 <- deviance(FM); RSS0 <- deviance(RM);
> (F <- ((RSS0-RSS1)/2)/(RSS1/FM$df.residual))
[1] 5.87
> (pval <- 1-pf(F,2,FM$df.residual))
[1] 0.00397
```

- Skip section 7.2.1 (Coefficient Quasi-Variances)

Modeling Interactions

- Two explanatory variables are said to interact in determining a response when the partial effect of one depends on the value of the other.
- Additive models are models without interaction.
- We now study interaction between a factor and a quantitative explanatory variable. This means that regression lines are *not* parallel.
- Example: effect of years of education on income for men and women.
  - Effect of education varies by gender (different slopes)
  - But also: effect of gender varies by education → interaction is symmetric concept.
  - Note that in figure a) gender and education are not related, but in b) they are: women have higher education levels than men.



- Don't mix up interaction and correlation of explanatory variables.

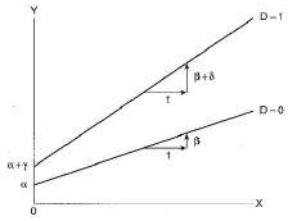
Constructing Interaction Regressors (1)

- Separate regressions for men and women could be fitted, but ...
- preferably a combined model is used, allowing to test the gender-by-education interaction.
- Separate regression and combined regression result in same estimates for intercepts and slopes.
- Model with different intercepts and slopes for women and men is fitted by inclusion of interaction regressor $XD = X \times D$, the product of the two regressors:

$$Y_i = \alpha + \beta X_i + \gamma D_i + \delta(X_i D_i) + \epsilon_i$$

Constructing Interaction Regressors (2)

- Splitting for women and men:
  - women: $Y_i = \alpha + \beta X_i + \gamma \times 0 + \delta(X_i \times 0) + \epsilon_i = \alpha + \beta X_i + \epsilon_i$
  - men: $Y_i = \alpha + \beta X_i + \gamma \times 1 + \delta(X_i \times 1) + \epsilon_i = (\alpha + \gamma) + (\beta + \delta) X_i + \epsilon_i$
- α and β are intercept and slope in reference group (women).
- γ is difference in intercept between men and women.
- δ is difference in slope between men and women.
- Testing for interaction: $H_0 : \delta = 0$.

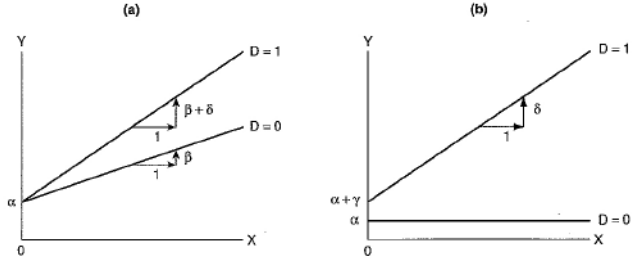


Principle of Marginality (1)

- Following John Nelder, effects of education and gender (the main effects), are marginal to the education-by-gender interaction.
- In general apply the following rules:
  - Do not test or interpret main effects of explanatory variables, if they interact.
  - If interaction can be ruled out on theoretical or empirical grounds, then proceed to test, estimate, and interpret main effects.
- Principle of marginality: model including a higher order term (such as an interaction), should normally also include the lower-order relatives of that term (main effects that "compose" the interaction).

Principle of Marginality (2)

- The Principle of Marginality is is not a strict principle. Models that violate the principle of marginality are not uninterpretable, but less broadly applicable.
- Below are two examples. Why do they violate the principle of marginality? Write down the models for the two examples.



Interactions With Polytomous Factors (1)

- Same method for modeling interactions by forming product regressors easily extends to situation of polytomous factors (more than two levels) and more quantitative regressors.
- Example `Prestige` dataset:
  - Response: Y=occupation prestige
  - One qualitative explanatory variable (factor) with 3 levels: occupational type; 2 dummies are used, coding for professional and white collar groups
  - Two quantitative explanatory variables X_1=income and X_2=education
  - Model:

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \gamma_1 D_{i1} + \gamma_2 D_{i2} \\ &\quad + \delta_{11} X_{i1} D_{i1} + \delta_{12} X_{i1} D_{i2} + \delta_{21} X_{i2} D_{i1} + \delta_{22} X_{i2} D_{i2} + \epsilon_i. \end{aligned}$$

Interactions With Polytomous Factors (2)

- Model can be split into:
  - Professional: $Y_i = (\alpha + \gamma_1) + (\beta_1 + \delta_{11}) X_{i1} + (\beta_2 + \delta_{21}) X_{i2} + \epsilon_i$.
  - White collar: $Y_i = (\alpha + \gamma_2) + (\beta_1 + \delta_{12}) X_{i1} + (\beta_2 + \delta_{22}) X_{i2} + \epsilon_i$.
  - Blue collar: $Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i$.

```
> lminteract <- lm(prestige ~ income + education + type + type:income + type:education, data=Pr
> coef(lminteract)
      (Intercept)            income         education          typeprof            typewc
          2.27575           0.00352           1.71327          15.35190         -33.53665
  income:typeprof      income:typewc education:typeprof   education:typewc
          -0.00290          -0.00207           1.38781           4.29087
> summary(lminteract)$r.squared
[1] 0.875
```


Interpreting Dummy-Regression Models With Interactions

- Model interpretation may be difficult by looking at individual coefficients.
- Helpful to write out the estimated regression equations for each group; sum some coefficients yourself to get the right intercepts and slopes.
- Alternative is to reparameterize, so that the intercepts and slopes per group are estimated directly. Below we first create dummies and products ourselves, and fit. Note that the "overall" intercept is removed from the model using -1.

```
> D1 <- ifelse(Pr2$type=="prof",1,0)
> D2 <- ifelse(Pr2$type=="bc",1,0)
> D3 <- ifelse(Pr2$type=="wc",1,0)
> X1D1 <- Pr2$income*D1; X1D2 <- Pr2$income*D2; X1D3 <- Pr2$income*D3;
> X2D1 <- Pr2$education*D1; X2D2 <- Pr2$education*D2; X2D3 <- Pr2$education*D3;
> lminteract2 <- lm(prestige ~  -1 + D1 + X1D1 + X2D1  +  D2 + X1D2 + X2D2
+                               + D3 + X1D3 + X2D3, data=Pr2)
> coef(lminteract2)
       D1      X1D1      X2D1        D2      X1D2      X2D2        D3      X1D3      X2D3
 17.62765   0.00062   3.10108   2.27575   0.00352   1.71327 -31.26090   0.00145   6.00415
> summary(lminteract2)$r.squared
[1] 0.986
```

- Be careful with R^2 in models without explicit intercept! R^2 is redefined in that case: as reference null model the model $Y = 0 + \epsilon$, is chosen, resulting in $R^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2}$.

Interpreting Dummy-Regression Models With Interact. (2)

- Better let R do the work for you. Again, remove the "overall" intercept, so that directly 3 intercepts are obtained, and ask for "interactions" like `type:income`. Don't use `type*income`, because R interprets that as `type + income + type:income`, and we would get difference parameters (for `income`) again.

```
> lminteract3 <- lm(prestige ~ -1 + type + type:income + type:education, data=Pr2)
> coef(lminteract3)
           typebc           typeprof             typewc      typebc:income    typeprof:in
          2.27575          17.62765         -31.26090            0.00352            0.0
    typewc:income  typebc:education typeprof:education   typewc:education
          0.00145            1.71327            3.10108            6.00415
```

- Fox suggests as aid in interpretation to focus on each high-order term (like type-income interaction), calculating predicted values over the range of values of one explanatory variable, while fixing the other(s) at its average value. A set of plots per explanatory variable is the result.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 2, lecture 9

Content week 2, lecture 9: Fox §7.3.5, §8.1

- Dummy variable regression
  - hypothesis testing for main effects and interactions (§7.3.5)
- One-way analysis of variance (§8.1)
  - one-way Analyis of variance model using dummies
  - effects model ($\mu + \alpha_j$) and means model (μ_j) notation
  - overparameterization
  - cornerstone restriction of parameters by use of dummy regressors
  - sum-to-zero restriction of parameters by use of deviation regressors
  - ANOVA table for one-way ANOVA

Recall Principle of Marginality

- Multiple regression model with at least one factor, represented by dummies, at least one quantitative regressor and their interaction(s).
- Principle of marginality: model including a higher order term (such as an interaction), should normally also include the lower-order relatives of that term (main effects that "compose" the interaction).
- In general apply the following rules:
  - Do not test or interpret main effects of explanatory variables, if they interact.
  - If interaction can be ruled out on theoretical or empirical grounds, then proceed to test, estimate, and interpret main effects.

Hypothesis Tests for Main Effects and Interactions (1)

- Use incremental F-tests to test null hypotheses, comparing Full and Reduced Models.
- Use the principle of marginality as guide to compare the right Full and Reduced Models:
  - First test interactions, or generally higher order terms.
  - Test main effects only if interactions are not important, or generally lower order terms if higher orders terms are not important.
  - Test each effect corrected for all others effects (except the higher-order terms it is marginal to) in the model.
- Example prestige: fit a series of models and construct the F-tests, using `anova`.

```
> Pr2 <- Prestige[!is.na(Prestige$type),]  # remove some missing values

> M1 <- lm(prestige ~ income + education + type + type:income + type:education, data=Pr2)
> M2 <- lm(prestige ~ income + education + type + type:income, data=Pr2)
> M3 <- lm(prestige ~ income + education + type + type:education, data=Pr2)
> M4 <- lm(prestige ~ income + education + type, data=Pr2)
> M5 <- lm(prestige ~ income + education, data=Pr2)
> M6 <- lm(prestige ~ income + type + type:income, data=Pr2)
> M7 <- lm(prestige ~ education + type + type:education, data=Pr2)
```

Hypothesis Tests for Main Effects and Interactions (2)

First test interactions:

```
> anova(M1,M2) # test interaction type:education
Analysis of Variance Table

Model 1: prestige ~ income + education + type + type:income + type:education
Model 2: prestige ~ income + education + type + type:income
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     89 3553
2     91 3791 -2      -238 2.99  0.056
> anova(M1,M3) # test interaction type:income
Analysis of Variance Table

Model 1: prestige ~ income + education + type + type:income + type:education
Model 2: prestige ~ income + education + type + type:education
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     89 3553
2     91 4505 -2      -952 11.9 2.6e-05
```


Hypothesis Tests for Main Effects and Interactions (3)

Next test main effects:

```
> anova(M2,M6) # test main effect of education, assuming absence of interaction type:education
Analysis of Variance Table

Model 1: prestige ~ income + education + type + type:income
Model 2: prestige ~ income + type + type:income
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     91 3791
2     92 4859 -1     -1068 25.6 2.1e-06
> anova(M4,M5) # test main effect of type, assuming absence of interactions with type
Analysis of Variance Table

Model 1: prestige ~ income + education + type
Model 2: prestige ~ income + education
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     93 4681
2     95 5272 -2      -591 5.87  0.004
> anova(M3,M7) # test main effect of income , assuming absence of interaction type:income
Analysis of Variance Table

Model 1: prestige ~ income + education + type + type:education
Model 2: prestige ~ education + type + type:education
  Res.Df  RSS Df Sum of Sq    F  Pr(>F)
1     91 4505
2     92 5637 -1     -1132 22.9 6.7e-06
```


Hypothesis Tests for Main Effects and Interactions (4)

- Results achieved on earlier slide are not exactly same as results in book:

Source	*Models Contrasted*	*Sum of Squares*	df	*F Fox*	*p Fox*	*F above*	*p above*
Income	3-7	1132	1	28.35	<0.0001	22.9	<0.0001
Education	2-6	1068	1	26.75	<0.0001	25.6	<0.0001
Type	4-5	592	2	7.41	0.0011	5.87	0.004
Income×Type	1-3	952	2	11.92	<0.0001	11.92	<0.0001
Education×Type	1-2	238	2	2.99	0.056	2.99	0.056
Residuals		3553	89				
Total		28347	97				

- We see slight discrepancy only for the main effects. The reason is the denominator of the F-statistics. In the Fox book the denominator is the residual mean square from model 1, i.e. the model with all main effects and interactions. In the testing we did by comparing two nested models, R used as denominator the residual mean square of the largest model of the two. The approach by Fox is to be preferred, because the full model ensures an unbiased estimate of the error variance.

Hypothesis Tests for Main Effects and Interactions (5)

- If the anova function is applied to the Full Model with all regressors in, then R constructs F-tests by sequentially building up the model, starting with a model containing only the intercept, and adding terms one at a time. Notice that completely different tests are performed now, which do not conform to testing principles formulated earlier.
- Resulting sums of squares are called type I SS or sequential SS or extra SS, corresponding to one specific order of terms in the model.

```
> M1 <- lm(prestige ~ income + education + type + type:income + type:education, data=Pr2)
> anova(M1)
Analysis of Variance Table

Response: prestige
               Df Sum Sq Mean Sq F value  Pr(>F)
income          1  14022   14022  351.24 < 2e-16
education       1   9053    9053  226.78 < 2e-16
type            2    591     296    7.40  0.0011
income:type     2    890     445   11.15 4.8e-05
education:type  2    238     119    2.99  0.0556
Residuals      89   3553      40
```

Hypothesis Tests for Main Effects and Interactions (6)

Alternative to `anova` function are functions `drop1` and `add1`.
Realize which two models (FM and RM) are being compared in each F-test.

- `drop1(M1,test="F")` drops higher order terms from the model, respecting marginality, and uses F-tests. Resulting SS are called type II SS or partial SS.

```
> drop1(M1, test="F")
Single term deletions

Model:
prestige ~ income + education + type + type:income + type:education
               Df Sum of Sq  RSS AIC F value  Pr(>F)
<none>                     3553 370
income:type     2       952 4505 389   11.92 2.6e-05
education:type  2       238 3791 372    2.99   0.056
```

- function add1 may be handy to add a single term to a model, given a scope, e.g. for adding a single interaction term to the additive model:

```
> add1(M4, ~ .^2, test="F")
Single term additions

Model:
prestige ~ income + education + type
                 Df Sum of Sq  RSS AIC F value  Pr(>F)
<none>                       4681 389
income:education  1       420 4261 382    9.07  0.0033
income:type       2       890 3791 372   10.68 6.8e-05
education:type    2       177 4505 389    1.78  0.1737
```


Caution Concerning Standardized Coefficients

- Don't use standardized regression coefficients for dummy variables or interaction regressors:
  - The straightforward interpretation of the unstandardized coefficient for a dummy, namely expected response difference between a group and the reference, is lost after standardization.
  - Furthermore, a dummy cannot be increased by one standard deviation: you can only go from 0 to 1.
  - Same story for interaction regressor.
  - A quantitative regressor may be standardized prior to taking the product with dummy.

Analysis of Variance

- So far, ANOVA (ANalysis Of VAriance) meant the partitioning of total sum of squares into "explained" and "unexplained" sums of squares. This generally applies to linear models.
- Historically, ANOVA refers to procedures for fitting and testing linear models in which all explanatory variables are categorical.
- ANOVA was developed by Ronald Fisher, who described the method in his famous 1925 book *Statistical Methods for Research Workers*.
- With single factor we have one-way ANOVA, with two factors two-way ANOVA, etc.
- Today we look at one-way ANOVA only, and postpone higher-way ANOVA.
- ANCOVA = ANalysis of COVAriance; an ANCOVA model is a model with at least one factor, and at least one quantitative explanatory variable, like we already saw in chapter 7.

One-Way Analysis of Variance: Regression on dummies

- In one-way analysis there is a single factor.
- Dummy regressors can be used to represent the effect of this factor.
- For example, response Y, factor X with 3 levels; define 2 dummies D_1 and D_2:

Group	D_1	D_2
1	1	0
2	0	1
3	0	0

  Model: $Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$
  This is the one-way ANOVA model written in regression notation: regressors (dummies) are explicitly mentioned.
- $E(Y_i) = \mu_j$ the expectation of Y in group j,

$$\mu_1 = \alpha + \gamma_1 \times 1 + \gamma_2 \times 0 = \alpha + \gamma_1$$
$$\mu_2 = \alpha + \gamma_1 \times 0 + \gamma_2 \times 1 = \alpha + \gamma_2$$
$$\mu_3 = \alpha + \gamma_1 \times 0 + \gamma_2 \times 0 = \alpha$$

  because $E(\epsilon_i) = 0$.
- So, we have a unique interpretation of the parameters in terms of group expectations:

$$\alpha = \mu_3$$
$$\gamma_1 = \mu_1 - \mu_3$$
$$\gamma_2 = \mu_2 - \mu_3$$

- As before, α is mean in reference group (group 3), γ_1 and γ_2 capture differences between group means and mean of reference group.

One-Way Analysis of Variance: ANOVA model (1)

- One-way ANOVA focuses on testing for differences among group means: omnibus F-tests: $H_0 : \gamma_1 = \gamma_2 = 0$, or phrased in group means: $H_0 : \mu_1 = \mu_2 = \mu_3$, or in words, H_0: "No differences among population group means".
- Traditionally, ANOVA model is written differently:
  - Y_{ij} uses two indices: j is index for group, i is index for observation within group.
  - We have m groups, so $j = 1, \cdots, m$; we have n_j replications per group, so $i = 1, \cdots, n_j$.
  - Note: Fox uses index j to indicate groups; other text books often use i; we follow Fox.
- One-way ANOVA model:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

where ϵ_{ij} has usual linear model assumptions: independent, normally distributed, equal variances, and zero expectation.

One-Way Analysis of Variance: ANOVA model (2)

The One-way ANOVA model can be written in two ways:

1. Regression notation:

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

   Regressors (here: dummies) are explicitly mentioned.

2. ANOVA notation:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

   Regressors are not explicitly mentioned.
   The idea of ANOVA notation is, of course, that μ represents the general level in "the population", and α_j represents "effect" of j-th level of the factor on the response.

- One-way ANOVA model in ANOVA notation can be written with or without intercept:
  - With intercept: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ is called the "effects-model", because parameters α_j are effects of factor level j
  - Without intercept: $Y_{ij} = \mu_j + \epsilon_{ij}$ is called the "means-model", because parameters μ_j are means (expected values) for factor level j.

One-Way Analysis of Variance: Overparameterization and restrictions

- Taking expectations: $\mu_j = \mu + \alpha_j$; hence overparameterization: we have $m + 1$ parameters (1 μ, and m α_j's), but only m μ_j's. E.g. with $m = 3$ groups:

$$\begin{aligned} \mu_1 &= \mu + \alpha_1 \\ \mu_2 &= \mu + \alpha_2 \\ \mu_3 &= \mu + \alpha_3 \end{aligned}$$

- Because of overparameterization, parameters cannot be uniquely estimated.
- Solution: place linear restriction on parameters: $w_0\mu + \sum_{j=1}^{m} w_j\alpha_j = 0$.
- Any restriction will do, e.g.
  - Corner stone parameterization: $\alpha_m = 0$ (so, $w_0 = w_1 = \cdots = w_{m-1} = 0, w_m = 1$). Dummy-coding scheme used earlier, where group m is the reference group, leads to this restriction: employ $m - 1$ dummies, coding for the first $m - 1$ levels of the factor. R uses $\alpha_1 = 0$.
  - Sum-to-zero parameterization or sigma constraint: $\sum_{j=1}^{m} \alpha_j = 0$ (so, $w_0 = 0, w_1 = \cdots = w_m = 1$). Using this restriction, we get $\mu_1 + \mu_2 + \mu_3 = 3\mu + \alpha_1 + \alpha_2 + \alpha_3 = 3\mu$, hence

$$\begin{aligned} \mu &= \frac{\sum_{j=1}^{m} \mu_j}{m} \equiv \mu_\cdot \\ \alpha_j &= \mu_j - \mu_\cdot \end{aligned}$$

    The dot (in $\mu_\cdot$) indicates averaging over index. Interpretation of parameters is now: μ is general mean, grand mean or overall mean; α_j is the difference of mean in group j and grand mean.
  - $H_0 : \mu_1 = \cdots = \mu_m \Leftrightarrow H_0 : \alpha_1 = \cdots = \alpha_m = 0$.

One-Way Analysis of Variance: Sum-to-zero parameterization

- Sum-to-zero parameterization is obtained by use of deviation regressors, instead of dummy regressors.
- Required are $m - 1$ deviation regressors $S_1, S_2, \cdots, S_{m-1}$ coded as

$$S_j = \begin{cases} 1 & : \text{ for observation in group } j \\ -1 & : \text{ for observation in group } m \\ 0 & : \text{ for observations in all other groups} \end{cases}$$

- For example, if $m = 3$

Group	S_1	S_2
1	1	0
2	0	1
3	-1	-1

- Model is $Y = \mu + \alpha_1 S_1 + \alpha_2 S_2 + \epsilon$; equations for group means are

$$\begin{aligned} \mu_1 &= \mu + \alpha_1 \times 1 + \alpha_2 \times 0 = \mu + \alpha_1 \\ \mu_2 &= \mu + \alpha_1 \times 0 + \alpha_2 \times 1 = \mu + \alpha_2 \\ \mu_3 &= \mu + \alpha_1 \times -1 + \alpha_2 \times -1 = \mu - \alpha_1 - \alpha_2 \end{aligned}$$

  From third group the restriction can be seen: $\alpha_3 = -\alpha_1 - \alpha_2$, so $\alpha_1 + \alpha_2 + \alpha_3 = 0$.

One-way Analysis of Variance: ANOVA table

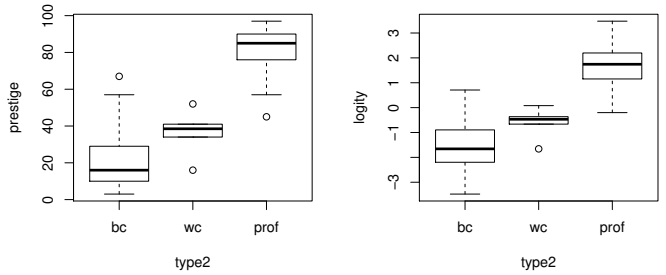
- Instead of fitting one-way ANOVA model by regression, estimates and sums of squares can also be calculated directly.
- Least-squares estimator of population mean μ_j in group j is sample mean : $\hat{\mu}_j = \overline{Y}_j$ in group j; then
  - $M \equiv \hat{\mu} = \frac{\sum_{j=1}^{m} \overline{Y}_j}{m} = \overline{Y}_.$
  - $A_j \equiv \hat{\alpha}_j = \overline{Y}_j - \overline{Y}_.$
  - Fitted values are $\widehat{Y}_{ij} = M + A_j = \overline{Y}_. + (\overline{Y}_j - \overline{Y}_.) = \overline{Y}_j$
  - $RegSS = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(\widehat{Y}_{ij} - \overline{Y}\right)^2 = \sum_{j=1}^{m} n_j \left(\overline{Y}_j - \overline{Y}\right)^2$ between-group sum of squares
  - $RSS = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(Y_{ij} - \widehat{Y}_{ij}\right)^2 = \sum\sum \left(Y_{ij} - \overline{Y}_j\right)^2$ within-group sum of squares
- ANOVA table:

Source	*Sum of Squares*	df	*Mean Square*	*F*	H_0
Between Groups	$\sum_{j=1}^{m} n_j \left(\overline{Y}_j - \overline{Y}\right)^2$	$m-1$	$\frac{RegSS}{m-1}$	$\frac{RegMS}{RMS}$	$\alpha_1 = \cdots = \alpha_m = 0$
Within Groups	$\sum\sum \left(Y_{ij} - \overline{Y}_j\right)^2$	$n-m$	$\frac{RSS}{n-m}$		
Total	$\sum\sum \left(Y_{ij} - \overline{Y}\right)^2$	$n-1$			

One-way Analysis of Variance: Example

- Duncan's data on prestige. Response `Prestige` is percentage. We analyze `logit(percentage/100)`, as it behaves slightly better: more symmetrical distributions within groups, less outliers.

```
> Duncan$type2 <-factor(Duncan$type,levels=c("bc","wc","prof")) # let order of levels correspond to book
> Duncan$logity <- log((Duncan$prestige)/(100-Duncan$prestige))
> boxplot(prestige ~ type2, data=Duncan)
> boxplot(logity ~ type2, data=Duncan)
```



One-way Analysis of Variance: Example continued

```
> (deviations <- contr.sum(levels(Duncan$type)))
     [,1] [,2]
bc      1    0
prof    0    1
wc     -1   -1
> anova.1 <- lm(logity ~ type, contrasts=list(type=deviations), data=Duncan)
> coef(anova.1)
(Intercept)       type1       type2
     -0.147      -1.335       1.779
> anova(anova.1)
Analysis of Variance Table

Response: logity
          Df Sum Sq Mean Sq F value  Pr(>F)
type       2   95.6    47.8      52 4.4e-12
Residuals 42   38.6     0.9
> model.matrix(anova.1)[1:9,]
           (Intercept) type1 type2
accountant           1     0     1
pilot                1     0     1
architect            1     0     1
author               1     0     1
chemist              1     0     1
minister             1     0     1
professor            1     0     1
dentist              1     0     1
reporter             1    -1    -1
```


One-way Analysis of Variance: Example continued

```
> anova.2 <- aov(logity ~ type, data=Duncan)  # aov is alternative for lm
> anova(anova.2)
Analysis of Variance Table

Response: logity
          Df Sum Sq Mean Sq F value  Pr(>F)
type       2   95.6    47.8      52 4.4e-12
Residuals 42   38.6     0.9
```

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 2, lecture 10

Content week 2, lecture 10

- One-way analysis of variance
  - After the F-test: pairwise comparisons
- Analysis of Covariance (§8.4)
  - wide sense: LM containing both qualitative (factor) and quantitative (covariate) explanatory variables
  - narrow sense: parallel lines model with main interest in factor, correcting for covariate
  - traditionally: centered covariate and sum-to-zero restriction of factor parameters
  - parallel lines model: adjusted means (exercise) and pairwise comparison of adjusted means

One-way Analysis of Variance: After F-test

- Quantitative response Y_{ij} in m groups.
- One-way ANOVA model (effects model notation):

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

  With F-test we compare means, testing $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_m = 0$, so "No difference at all".
- After a significant result from F-test, often next question is: which groups differ?
- Therefore: pairwise comparisons of means among groups, also called post-hoc tests.
- Each pairwise comparison can be done by t-test, using pooled error variance from ANOVA.
- We arrive here at problem of multiple comparisons or multiple testing: for a factor with m levels $m(m-1)/2$ pairs can be formed, and equally many pairwise t-tests can be done. The probability of at least one erroneous rejection of a null hypothesis is (much) larger than the nominal α used in an individual test! We postpone this topic for the moment.

One-way ANOVA: simple pairwise comparisons

- Simple function for pairwise comparisons: `pairwise.t.test()`
- Uses "pooled variance" estimate, i.e. error variance is estimated with `MSE` from one-way ANOVA. Information from different groups is "pooled" to get the error variance estimate.
- This only makes sense if we assume that variances are constant across different groups, as we do in one-way ANOVA.

```
> pairwise.t.test(Duncan$logity, Duncan$type,  p.adj="none") # ordinary t-tests, pooled variance
        Pairwise comparisons using t tests with pooled SD

data:  Duncan$logity and Duncan$type

     bc    prof
prof 8e-13 -
wc   0.05  1e-05

P value adjustment method: none
```

- Note option `p.adj="none"`, telling no adjustment for multiple comparisons is made.

Post-hoc t-tests after F-test: emmeans (1)

- R package emmeans contains tools to obtain model-based comparisons.
- Usable not only for one-way ANOVA, but also for more general linear models where groups are to be compared.
- It nicely processes the object obtained from the lm() function.

```
> lmo <- lm(logity ~ type, data=Duncan)
> emmeans(lmo, ~ type)  # model based means
 type emmean    SE df lower.CL upper.CL
 bc    -1.48 0.209 42    -1.90   -1.060
 prof   1.63 0.226 42     1.18    2.088
 wc    -0.59 0.391 42    -1.38    0.199

Confidence level used: 0.95
```

- Above model based means are calculated: estimates of $\mu + \alpha_i$ (using the effects model notation for the one-way ANOVA model).
- In this simple one-way ANOVA situation model based means are identical to ordinary means:

```
> summaryBy(logity ~ type, data=Duncan, FUN=mean) # ordinary group means
  type logity.mean
1   bc       -1.48
2 prof        1.63
3   wc       -0.59
```


Post-hoc t-tests after F-test: emmeans (2)

- Pairwise comparisons obtained in this way:

```
> emmeans(lmo, pairwise ~ type, adjust="none") # pairwise comparisons between levels of type
$emmeans
 type emmean    SE df lower.CL upper.CL
 bc    -1.48 0.209 42    -1.90   -1.060
 prof   1.63 0.226 42     1.18    2.088
 wc    -0.59 0.391 42    -1.38    0.199

Confidence level used: 0.95

$contrasts
 contrast  estimate    SE df t.ratio p.value
 bc - prof   -3.114 0.308 42 -10.110 <.0001
 bc - wc     -0.892 0.444 42  -2.010 0.0510
 prof - wc    2.222 0.452 42   4.920 <.0001
```

- Make sure you understand how standard errors of the mean (sem, e.g. for group bc=0.209), and standard errors of a difference (sed, e.g. for difference of means between bc and prof = 0.308) are obtained. To check you need the *MSE* from the ANOVA table and sample sizes in three groups (see below):

```
> anova(lmo)
Analysis of Variance Table

Response: logity
          Df Sum Sq Mean Sq F value   Pr(>F)
type       2  95.55   47.78   51.98 4.36e-12
Residuals 42  38.60    0.92
> table(Duncan$type)
  bc prof   wc
  21   18    6
```


Analysis of Covariance (§8.4)

- An ANCOVA model, as we use it here (ANCOVA in wider sense), is a linear model that contains both qualitative explanatory variables (factors) and quantitative explanatory variables (covariates).
- ANCOVA in narrow sense employs an additive model without interaction, i.e. a parallel lines model, with the aim to compare groups, but "corrected" for the covariates.
- In the last case the aim of inclusion of covariates in the model is to:
  1. obtain more precise comparisons among the groups, because part of the residual variation is explained by the covariates, leading to smaller $\hat{\sigma}^2_\epsilon$.
  2. obtain less biased results; the covariate may have slightly different means at the different levels of the treatment factor, which would lead to biased comparisons if not taken into account.
- Equivalent to dummy-variable regression we saw earlier, but usually parameterized differently:
  - ANOVA formulation for main effects and interactions of factors, i.e. $\mu + \alpha_i$
  - Covariates centered, i.e. expressed as deviations from their means.

ANCOVA models (wider sense) without and with interaction

- ANCOVA model with parallel lines (no interaction of covariate and factor):

$$Y_{ij} = \mu + \alpha_j + \beta\left(X_{ij} - \overline{X}\right) + \epsilon_{ij}$$

  with $j = 1, \ldots, m$ for m groups, i index for replicates within the group and single covariate X.
- ANCOVA model with interaction of covariate and factor:

$$Y_{ij} = \mu + \alpha_j + \beta\left(X_{ij} - \overline{X}\right) + \gamma_j\left(X_{ij} - \overline{X}\right) + \epsilon_{ij}$$

- In essence nothing new: earlier we wrote these models in regression notation (using dummies and products of dummies with quantative regressors) and not-centered. Assuming two groups ($m = 2$, so single dummy D) and single index $i = 1, \ldots, n$ with n total numbers of observations, we would write:
  - parallel lines (without interaction): $Y_i = \alpha + \beta X_i + \gamma D_i + \epsilon_i$
  - non-parallel lines (with interaction): $Y_i = \alpha + \beta X_i + \gamma D_i + \delta X_i D_i + \epsilon_i$

Analysis of Covariance (wider sense, including interaction) (1)

- Example Moore and Krupats study of conformity and authoritarianism. How is relationship between conformity and social status influenced by "authoritarianism"? Variable `conformity` counts how often a subject changes its judgment on 40 occasions (after having been confronted with judgement of partner). [So it actually is disguided proportion.] Variable `fscore` is "standard authoritarianism" score of the subject. Expectation is that low-authoritarian subjects (low `fscore`) are more responsive to social status of partner.

```
> head(Moore, n=2)
  partner.status conformity fcategory fscore
1            low          8       low     37
2            low          4      high     57
```

- First linear model with dummy for second level of `status`, `fscore` as covariate.

```
> ancova.1 <- lm(conformity ~ partner.status + fscore + partner.status:fscore, data=Moore)
> coef(ancova.1)
             (Intercept)        partner.statuslow                   fscore
                  20.793                  -15.534                   -0.151
partner.statuslow:fscore
                   0.261
> summary(ancova.1)$r.squared
[1] 0.294
> deviance(ancova.1)
[1] 853
```


Analysis of Covariance (wider sense, including interaction) (2)

- Traditional ANCOVA parameterization: centered covariate and sum-to-zero restriction on parameters.

```
> cf <- Moore$fscore-mean(Moore$fscore) # center the covariate
> ancova.2 <- lm(conformity ~ partner.status + cf + partner.status:cf,
+   contrasts=list(partner.status=contr.sum), data=Moore)
> coef(ancova.2)
      (Intercept)    partner.status1                 cf partner.status1:cf
          12.1406             2.1388            -0.0205            -0.1306
> summary(ancova.2)$r.squared
[1] 0.294
> deviance(ancova.2)
[1] 853
```

- Notice that R^2 and *RSS* are identical for two parameterizations, but parameters differ.
- Interpret parameters in two parameterizations!

Analysis of Covariance (narrow sense) - adjusted means

- Narrow sense ANCOVA: parallel lines model with main interest in factor, correcting for covariate; covariate typically centered.
- Model: $Y_{ij} = \mu + \alpha_j + \beta\left(X_{ij} - \overline{X}\right) + \epsilon_{ij}$
- Adjusted means are predicted means per group evaluated at the average value of the covariate.
- In case of centered covariate, average covariate value is zero, so adjusted mean estimates $\mu + \alpha_j$.
- In case of uncentered covariate, adjusted mean estimates $\mu + \alpha_j + \beta\overline{X}$.
- In both cases group differences are $\alpha_{j1} - \alpha_{j2}$, to be tested with t-tests.
- R function `emmeans()` gives adjusted means, and pairwise differences among adjusted means.

Analysis of Covariance (narrow sense) - example adjusted means (1)

- Example `Prestige` using single covariate `education` (X) and type of occupation (levels: `bc`, `prof`, `wc`; `bc` is reference group);
- Model, regression notation: $Y_i = \mu + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \beta X_i + \epsilon_i$
  i is index for observations $1, \ldots, 98$.
- Model, ANOVA notation: $Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$
  j is index for group ($j = 1, 2, 3$), i is index for observations within groups (counts within groups are: 44, 31, 23).

```
> Pr <- na.omit(Prestige)   #remove some missing values
> lmo <- lm(prestige ~ type + education, data=Pr)
> coef(summary(lmo))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.70      5.736   -0.47 6.39e-01
typeprof        6.14      4.259    1.44 1.53e-01
typewc         -5.46      2.691   -2.03 4.53e-02
education       4.57      0.672    6.81 9.16e-10
> emmeans(lmo, ~ type)
 type emmean   SE df lower.CL upper.CL
 bc     46.7 2.02 94     42.7     50.7
 prof   52.8 2.62 94     47.6     58.0
 wc     41.2 1.64 94     38.0     44.5

Confidence level used: 0.95
> mean(Pr$education) # ordinary mean of education
[1] 10.8
```

- Adjusted means for

bc $\hat{Y}_1 = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta}\bar{X} = -2.70 + 0 + 4.57 \times 10.8 = 46.7$

prof $\hat{Y}_2 = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta}\bar{X} = -2.70 + 6.14 + 4.57 \times 10.8 = 52.8$

wc $\hat{Y}_3 = \hat{\mu} + \hat{\alpha}_3 + \hat{\beta}\bar{X} = -2.70 - 5.46 + 4.57 \times 10.8 = 41.2$

Analysis of Covariance (narrow sense) - example adjusted means (2)

- Now with centered covariate X.
- Check how parameter estimates differ with centered covariate, and how adjusted means are just based on intercept parameters then.

```
> lmo2 <- lm(prestige ~ type + I(education-mean(education)), data=Pr)
> coef(summary(lmo2))
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                     46.67      2.016   23.15 1.32e-40
typeprof                         6.14      4.259    1.44 1.53e-01
typewc                          -5.46      2.691   -2.03 4.53e-02
I(education - mean(education))   4.57      0.672    6.81 9.16e-10
> emmeans(lmo2, ~ type)
 type emmean   SE df lower.CL upper.CL
 bc     46.7 2.02 94     42.7     50.7
 prof   52.8 2.62 94     47.6     58.0
 wc     41.2 1.64 94     38.0     44.5

Confidence level used: 0.95
```

- μ is predicted mean of bc at average value of X, so at $X - \bar{X} = 0$.
- Adjusted means for

$$\begin{aligned} \text{bc} \quad & \hat{Y}_1 = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta} \times 0 = 46.67 + 0 + 4.57 \times 0 = 46.7 \\ \text{prof} \quad & \hat{Y}_2 = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta} \times 0 = 46.67 + 6.14 + 4.57 \times 0 = 52.8 \\ \text{wc} \quad & \hat{Y}_3 = \hat{\mu} + \hat{\alpha}_3 + \hat{\beta} \times 0 = 46.67 - 5.46 + 4.57 \times 0 = 41.2 \end{aligned}$$

Analysis of Covariance (narrow sense) - example adjusted means (3)

- Finally with sum-to-zero restriction for α_j added.

```
> lmo3 <- lm(prestige ~ type + I(education-mean(education)), contrasts=list(type=contr.sum), data=Pr)
> coef(summary(lmo3))
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    46.894      0.852  55.016 2.67e-73
type1                          -0.228      2.158  -0.106 9.16e-01
type2                           5.914      2.279   2.595 1.10e-02
I(education - mean(education))  4.573      0.672   6.809 9.16e-10
> emmeans(lmo3, ~ type)
 type emmean   SE df lower.CL upper.CL
 bc     46.7 2.02 94     42.7     50.7
 prof   52.8 2.62 94     47.6     58.0
 wc     41.2 1.64 94     38.0     44.5

Confidence level used: 0.95
```

- α_3 is not shown in output; it is $\alpha_3 = -(\alpha_1 + \alpha_2)$ (so that $\sum \alpha_j = 0$).
- Adjusted means for

$$\begin{aligned} \text{bc} \quad & \hat{Y}_1 = \hat{\mu} + \hat{\alpha}_1 + \hat{\beta} \times 0 = 46.89 + -0.23 + 4.57 \times 0 = 46.7 \\ \text{prof} \quad & \hat{Y}_2 = \hat{\mu} + \hat{\alpha}_2 + \hat{\beta} \times 0 = 46.89 + 5.91 + 4.57 \times 0 = 52.8 \\ \text{wc} \quad & \hat{Y}_3 = \hat{\mu} + \hat{\alpha}_3 + \hat{\beta} \times 0 = 46.89 + (-(-0.23 + 5.91)) + 4.57 \times 0 = 41.2 \end{aligned}$$

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 2, lecture 11

Content week 2, lecture 11: Fox §9.1, Appendix (Fox) A,B.1.1,B.1.2

- Linear Model in matrix form: $y = X\beta + \epsilon$
- Linear Algebra: Notation
- Linear Algebra: Matrices
  - order, transpose, symmetry, triangular, diagonal, trace
  - simple matrix math: elementwise + - * / scalar multiplication
  - inner (dot) product, matrix multiplication, idempotency

Linear Models in Matrix Form

- So far, we have seen a number of models for quantitative response variables:
  - Linear regression models: quantitative explanatory variables.
  - Analysis of variance models: qualitative explanatory variables (factors).
  - Analysis of covariance models: both quantitative and qualitative explanatory variables.
- These models have a lot in common, like their assumptions:
  - Normality
  - Constant variance
  - Independence
  - Linearity in parameters, expected error zero.
- These models are all examples of Linear Models.
- It is very useful to write linear models in Matrix Form:

$$y = X\beta + \epsilon$$

- Before discussing this, first refresh knowledge of Matrix Algebra.

Matrix Algebra

- Use appendix B on Matrices, Linear Algebra, Vector Geometry, supplementary to Fox book
- For more thorough treatment of topic see e.g.
  - Searle, S.R. (1982) Matrix Algebra Useful for Statistics. Wiley
  - Harville, D.A. (2008) Matrix Algebra From a Statistician's Perspective. Springer
  - Lay, D.C. (2003) Linear Algebra and its Applications. Addison Wesley
  - Bretscher, O. (2008) Linear Algebra with Applications. Prentice Hall
  - Nicholson, W.K. (2006) Linear Algebra with Applications. McGraw-Hill Ryerson

Notation

Notation in appendix is (although we will deviate from it now and then):

- Known scalar constant: lowercase italic letter [a].
- Scalar random variable: uppercase italic letter [X], sometimes lowercase Greek letter [ϵ_i].
- Scalar parameters: lowercase Greek letter [α].
- Vectors and matrices: boldface, lowercase for vector [$\mathbf{x}_1$], uppercase for matrices [$\mathbf{A}$].
- Symbol $\equiv$ means "is equal to by definition".
- $E()$ expectation of scalar, vector, or matrix random variable.
- $V()$ variance of scalar random variable, or variance-covariance matrix of vector random variable.
- $C()$ covariance of two scalar random variables, or covariance matrix of two vector random variables.

Topics Matrix Algebra

Topics in appendix comprise:

Matrices

- Matrix is rectangular table of numbers with m rows and n columns
- For example (4 rows and 3 columns): $\mathbf{X} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix}$
- In general: $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$
- Matrix of order m by n has m rows and n columns, written $(m \times n)$.
- Column vector is matrix with one column: $\mathbf{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$
- Row vector is matrix with one row: $\mathbf{b}' = [b_1, b_2, \cdots, b_n]$.

Transpose of a Matrix

- Transpose of matrix $\mathbf{A}$, denoted $\mathbf{A}'$ or $\mathbf{A}^T$, is formed from $\mathbf{A}$ so that ith row of $\mathbf{A}'$ consists of elements of ith column of $\mathbf{A}$.
- For example $\mathbf{X} = \begin{bmatrix} 1 & -2 & 3 \\ 4 & -5 & -6 \\ 7 & 8 & 9 \\ 0 & 0 & 10 \end{bmatrix}$ and $\mathbf{X}' = \begin{bmatrix} 1 & 4 & 7 & 0 \\ -2 & -5 & 8 & 0 \\ 3 & -6 & 9 & 10 \end{bmatrix}$.
- $(\mathbf{A}')' = \mathbf{A}$.
- By convention, vector like $\mathbf{a}$ is column vector.
- Row vector is explicitly transposed, like $\mathbf{b}'$.

Square Matrix

- Square matrix of order n has n rows and n columns.
- Entries a_{ii} of square matrix form main diagonal of the matrix.
- Trace of the square matrix $\mathbf{A}$ is sum of diagonal elements: $\text{trace}(\mathbf{A}) \equiv \sum_{i=1}^{n} a_{ii}$.
- E.g. $\mathbf{B} = \begin{bmatrix} -5 & 1 & 3 \\ 2 & 2 & 6 \\ 7 & 3 & -4 \end{bmatrix}$ has diagonal elements -5, 2, -4, and $\text{trace}(\mathbf{B}) = \sum_{i=1}^{3} b_{ii} = -5 + 2 - 4 = -7$.
- Square matrix is symmetric if $\mathbf{A} = \mathbf{A}'$.
- E.g. matrix $\mathbf{B}$ is not symmetric, but $\mathbf{C} = \begin{bmatrix} -5 & 1 & 3 \\ 1 & 2 & 6 \\ 3 & 6 & -4 \end{bmatrix}$ is.
- Many matrices in statistical applications are symmetric: correlation matrices, variance-covariance matrices, matrices of sums of squares. Could matrix $\mathbf{C}$ above be a variance-covariance matrix?

Special matrices

- Upper triangular matrix is square matrix with zeroes below main diagonal:
  $\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ 0 & u_{22} & \cdots & u_{2n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & u_{nn} \end{bmatrix}$.
- Lower triangular matrix is square matrix with zeroes above main diagonal:
  $\mathbf{L} = \begin{bmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{bmatrix}$.
- Diagonal matrix is square matrix with all off-diagonal entries zero:
  $\mathbf{D} = \begin{bmatrix} d_{11} & 0 & \cdots & 0 \\ 0 & d_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & d_{nn} \end{bmatrix}$.
- Scalar matrix is diagonal matrix with all diagonal entries equal: $\mathbf{S} = \text{diag}(s, s, \cdots, s)$.
- Identity matrix is scalar matrix with all diagonal entries one: $\mathbf{I} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$.
- Zero matrix $\mathbf{0}$ with all entries 0.
- Unit vector $\mathbf{1}_n$ with all n elements 1.

Partitioned matrix

- Partitioned matrix is matrix whose elements are organized into submatrices
- Example $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ with submatrices $\mathbf{A}_{11} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}$, $\mathbf{A}_{12} = \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix}$, $\mathbf{A}_{21} = [a_{41}, a_{42}]$, and $\mathbf{A}_{22} = [a_{43}]$.

Simple matrix arithmetic: addition

- Two matrices may be added only if they are of same order.
- Matrix sum is formed by adding corresponding elements: if $\mathbf{A}$ and $\mathbf{B}$ have order $(m \times n)$ then $\mathbf{C} = \mathbf{A} + \mathbf{B}$ of order $(m \times n)$ and elements $c_{ij} = a_{ij} + b_{ij}$.
- Likewise, difference $\mathbf{D} = \mathbf{A} - \mathbf{B}$ with elements $d_{ij} = a_{ij} - b_{ij}$.
- Negative of matrix $\mathbf{E} = -\mathbf{A}$ with elements $e_{ij} = -a_{ij}$.
- Because matrix addition, subtraction, and negation involve element-wise operations only, same rules apply as in scalar operations:
  - $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$ (commutative).
  - $\mathbf{A} + (\mathbf{B} + \mathbf{C}) = (\mathbf{A} + \mathbf{B}) + \mathbf{C}$ (associative).
  - $\mathbf{A} - \mathbf{B} = \mathbf{A} + (-\mathbf{B}) = -(\mathbf{B} - \mathbf{A})$.
  - $\mathbf{A} - \mathbf{A} = \mathbf{0}$.
  - $\mathbf{A} + \mathbf{0} = \mathbf{A}$.
  - $-(-\mathbf{A}) = \mathbf{A}$.
  - $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$.

Simple matrix arithmetic: scalar product

- Product of scalar c and matrix $\mathbf{A}$ is $\mathbf{B} = c\mathbf{A}$ with elements $b_{ij} = ca_{ij}$
- Scalar-matrix product obeys rules:
  - $c\mathbf{A} = \mathbf{A}c$ (commutative).
  - $\mathbf{A}(b + c) = \mathbf{A}b + \mathbf{A}c$.
  - $c(\mathbf{A} + \mathbf{B}) = c\mathbf{A} + c\mathbf{B}$.
  - $0\mathbf{A} = \mathbf{0}$.
  - $1\mathbf{A} = \mathbf{A}$.
  - $(-1)\mathbf{A} = -\mathbf{A}$.

Matrix multiplication (1)

- Inner product or dot product of two vectors (each with n entries), say $\mathbf{a}'$ and $\mathbf{b}$, denoted as $\mathbf{a}' \cdot \mathbf{b}$ is scalar formed by multiplying corresponding entries of vectors and summing resulting products: $\mathbf{a}' \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$
- Example on whiteboard
- Two matrices $\mathbf{A}$ and $\mathbf{B}$ are comformable for multiplication in the given order $(\mathbf{AB})$, if number of columns of $\mathbf{A}$ is equal to number of rows of $\mathbf{B}$: $\mathbf{A}$ has order $m \times n$ and $\mathbf{B}$ has order $n \times p$.
- For example $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ are comformable for multiplication, but $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}$ are not.
- Let $\mathbf{C} = \mathbf{AB}$ be matrix product; with $\mathbf{a}_i'$ ith row of $\mathbf{A}$ and $\mathbf{b}_j$ jth column of $\mathbf{B}$. Then $\mathbf{C}$ is matrix of order $(m \times p)$ with $c_{ij} = \mathbf{a}_i' \cdot \mathbf{b}_j = \sum_{k=1}^{n} a_{ik} b_{kj}$.
- Example on whiteboard

Matrix multiplication (2)

- Rules for matrix multiplication:
  - $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$ (associative).
  - $(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$ (distributive).
  - $\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC}$ (distributive).
- Multiplication is in general not commutative:
  - if $\mathbf{A}$ is $m \times n$ and $\mathbf{B}$ is $n \times p$ then $\mathbf{AB}$ is defined, but $\mathbf{BA}$ only if $m = p$.
  - Even if $m = p$, $\mathbf{AB}$ and $\mathbf{BA}$ are of different order, unless matrices are square.
  - Even with square matrices, $\mathbf{AB}$ and $\mathbf{BA}$ are not necessarily equal.
  - If $\mathbf{AB} = \mathbf{BA}$ then $\mathbf{A}$ and $\mathbf{B}$ are said to commute with each other.
- Multiplication with identity matrices: $\mathbf{A}_{m\times n}\mathbf{I}_n = \mathbf{I}_m\mathbf{A}_{m\times n} = \mathbf{A}$.
- Transposing product: $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
- Powers of matrix are products of matrix with itself: $\mathbf{A}^2 = \mathbf{AA}$, $\mathbf{A}^3 = \mathbf{AAA}$, etc.
- Square root of $\mathbf{A}$ is matrix $\mathbf{B}$ with property $\mathbf{B}^2 = \mathbf{A}$, which is also written as $\mathbf{B} = \mathbf{A}^{1/2}$. The square root is not unique.
- If $\mathbf{A}^2 = \mathbf{A}$, $\mathbf{A}$ is said to be idempotent (e.g. projection matrix).

Matrix multiplication (3)

- For partioned matrices, addition, subtraction, and multiplication works as if submatrices were elements, e.g. with comformable submatrices: $\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \mathbf{B}_{21} & \mathbf{B}_{22} \end{bmatrix}$, $\mathbf{AB} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{B}_{11}+\mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12}+\mathbf{A}_{12}\mathbf{B}_{22} \\ \mathbf{A}_{21}\mathbf{B}_{11}+\mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12}+\mathbf{A}_{22}\mathbf{B}_{22} \end{bmatrix}$.
- Definition of matrix multiplication makes it simple to formulate systems of scalar equations as single matrix equation, e.g.

$$\begin{aligned} 2x_1 + 5x_2 &= 4 \\ x_1 + 3x_2 &= 5 \end{aligned}$$

becomes in matrix notation $\begin{bmatrix} 2 & 5 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$ or simply $\mathbf{Ax} = \mathbf{b}$.
That is what is needed in Linear Models!

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 2, lecture 12

Content week 2, lecture 12: Appendix (Fox) B.1.3,B.1.4,B.1.5,B2,B3

- Matrices (cont.)
  - matrix inverse, Gaussian elimination, elementary row operations, (non)singular matrix
  - determinant of square matrix
  - Kronecker (outer) product
- Basic vector geometry
  - vectors, coordinate space
  - vector addition, subtraction, length of vector, collinearity
- Vector space and subspace
  - n-dimensional vector space, subspace, vectors spanning subspace
  - linear independent set of vectors, linear dependency
  - dimension of subspace, basis for subspace, coordinates of y w.r.t. basis

Topics Matrix Algebra

Topics in appendix comprise:

B.1 Matrices (part)

B.2 Basic Vector Geometry

B.3 Vector Spaces and Subspaces (part)

B.4 Matrix Rank and Solving Set of Linear Equations

B.5 Eigenvalues and Eigenvectors

Matrix inverses

- In matrix algebra there is no direct analog of the scalar division, but a square matrix may have matrix inverse.
- Inverse of square matrix $\mathbf{A}$ is written as $\mathbf{A}^{-1}$, and has property $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.
- If square matrix has inverse, then it is called nonsingular.
- Square matrix without inverse is called singular.
- If inverse matrix exists, it is unique.
- If $\mathbf{AB} = \mathbf{I}$ then necessarily $\mathbf{BA} = \mathbf{I}$.
- Inverse of 2×2 matrix $\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \frac{1}{a_{11}a_{22} - a_{12}a_{21}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}$.
- Example of inverse of 2×2 matrix on whiteboard.

Matrix inverses (2)

- In scalar algebra only number 0 has no inverse. In contrast, many singular nonzero matrices exist.
- Example: hypothesize that $\mathbf{B}$ is inverse of matrix $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$. However, $\mathbf{AB} = \mathbf{A}\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} b_{11} & b_{12} \\ 0 & 0 \end{bmatrix} \neq \mathbf{I}_2$, contradicting the hypothesis. Hence $\mathbf{A}$ does not have an inverse.
- Generalized inverses exist for rectangular matrices and singular square matrices.

Finding the inverse of nonsingular square matrix

- Gaussian elimination.
- Steps for determining inverse of $\mathbf{A}$:
  - Augment matrix $\mathbf{A}$ at right hand side with identity matrix $\mathbf{I}$: $[\mathbf{A}, \mathbf{I}]$.
  - Reduce original matrix to identity matrix by applying elementary row operations of three sorts:
    E_I: multiply each entry in row by nonzero scalar constant.
    E_{II}: add scalar multiple of one row to another, replacing other row.
    E_{III}: exchange two rows of matrix.
  - Diagonal elements are called pivots. Sweep out other rows using pivot.
  - Example whiteboard

Why Gaussian elimination works

- Eliminiation method works, because each elementary row operation can be represented as multiplication on the left by appropriately formulated square matrix:
  - E_I e.g. multiply row 3 by e.g. 0.5 can be done by left-multiplication by diagonal matrix with diagonal $(1, 1, 0.5)$: $E_I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$.
  - E_{II} e.g. add row 1 to row 2, thereby replacing row 2 can be done by left-multiplication by $E_{II} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.
  - E_{III} e.g. swapping rows 2 and 3 can be done by left-multiplication by $E_{III} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$. *Try!*
- Elimination procedure applies sequence of (say p) elementary row operations to augmented matrix $\mathbf{A}, \mathbf{I}_n$: defining $\mathbf{E} \equiv \mathbf{E}_p \cdots \mathbf{E}_2 \mathbf{E}_1$ $\mathbf{E}[\mathbf{A}, \mathbf{I}_n] = [\mathbf{I}_n, \mathbf{B}]$, or $\mathbf{EA} = \mathbf{I}_n$, hence $\mathbf{E} = \mathbf{A}^{-1}$. Further $\mathbf{EI}_n = \mathbf{B}$, and consequently $\mathbf{B} = \mathbf{E} = \mathbf{A}^{-1}$.
- If $\mathbf{A}$ is singular it cannot be reduced to $\mathbf{I}$ by elementary row operations: at some point pivot is zero.

Rules for matrix inverses

- $\mathbf{I}^{-1} = \mathbf{I}$.
- $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.
- $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$.
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
- $(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$.
- For $\mathbf{D} = \text{diag}(d_1, \cdots, d_n)$ and all $d_i \neq 0$, then $\mathbf{D}$ is nonsingular, and $\mathbf{D}^{-1} = \text{diag}(1/d_1, \cdots, 1/d_n)$.
- Inverse of nonsingular symmetric matrix is itself symmetric.

Determinants (1)

- Each square matrix $\mathbf{A}$ associated with scalar, called determinant, written $\det\mathbf{A}$.
- For 2×2 matrix $\det\mathbf{A} = a_{11}a_{22} - a_{12}a_{21}$.
- General definition of determinant exists, but can also implicitly be defined by specifying properties:
  - D1 Multiplying row of square matrix by scalar constant multiplies determinant of matrix by same constant.
  - D2 Adding multiple of one row to another leaves determinant unchanged.
  - D3 Interchanging two rows changes the sign of determinant.
  - D4 $\det\mathbf{I} = 1$.

Determinants (2)

- Since Gaussian elimination reduces square matrix to identity matrix by above mentioned steps, value of determinant can be established. Determinant is product of pivot elements, possibly with change of sign of product.
- Example from Gaussian elimination:
  1. value 2 in pivot position 1, divide row 1 by 2, hence row of original matrix is $2\times$ corresponding row of new matrix $\mathbf{A}_1$, hence $\det\mathbf{A} = 2 \times \det\mathbf{A}_1$.
  2. subtract row 1 from row 2, determinant does not change
  3. subtract $4\times$ row 1 from row 3, determinant does not change
  4. interchange rows 2 and 3, determinant multiplied by -1
  5. value 8 in pivot position 2, so divide row 2 by 8, determinant old matrix is $8\times$ determinant of new matrix
  6. add row 2 to row 1: determinant does not change
  7. value 1 in pivot position 3, no action needed
  8. add $\frac{1}{2}\times$ row 3 to row 1: determinant does not change.
  9. add $\frac{1}{2}\times$ row 3 to row 2: determinant does not change.

  Going through all steps, results in $\det\mathbf{A} = (2) \times (-1) \times (8) \times (1) = -16$.
- Singular matrix has determinant 0 (because one or more of pivots are zero).
- Determinant can be viewed as the signed volume of the parallelepiped spanned by the column or row vectors of the matrix.

Kronecker product

- Let $\mathbf{A}$ be of order $m \times n$ and $\mathbf{B}$ $p \times q$. Then Kronecker product $\mathbf{A} \otimes \mathbf{B}$ is defined as $\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{2n}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{m1}\mathbf{B} & a_{m2}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$.

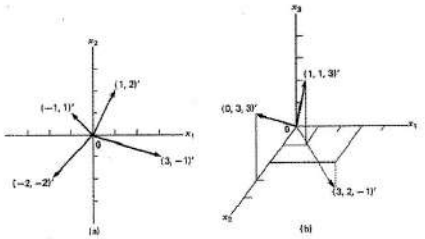
- Named after German mathematician Leopold Kronecker.
- Useful f.i. for compactly representing patterned matrices:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \otimes \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma_1^2 & \sigma_{12} & 0 & 0 \\ 0 & 0 & \sigma_{12} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma_1^2 & \sigma_{12} \\ 0 & 0 & 0 & 0 & \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

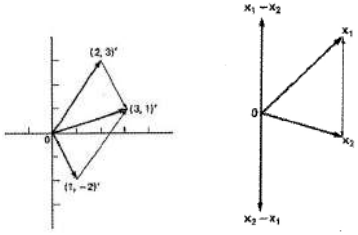
- Many properties similar to ordinary matrix multiplication:
  - $\mathbf{A} \otimes (\mathbf{B} + \mathbf{C}) = \mathbf{A} \otimes \mathbf{B} + \mathbf{A} \otimes \mathbf{C}$.
  - $(\mathbf{B} + \mathbf{C}) \otimes \mathbf{A} = \mathbf{B} \otimes \mathbf{A} + \mathbf{C} \otimes \mathbf{A}$.
  - $(\mathbf{A} \otimes \mathbf{B}) \otimes \mathbf{D} = \mathbf{A} \otimes (\mathbf{B} \otimes \mathbf{D})$.
  - $c(\mathbf{A} \otimes \mathbf{B}) = (c\mathbf{A}) \otimes \mathbf{B} = \mathbf{A} \otimes (c\mathbf{B})$.
  - $\otimes$ is generally not commutative: $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$.
  - $(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD}$.
  - $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$.
  - $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$.

Basic vector geometry

- Vector $\mathbf{x} = (x_1, \cdots, x_n)'$ is represented as directed line segment extending from origin of n-dimensional Cartesian coordinate space to the point defined by the entries (=coordinates) of vector. Figure shows vectors in (a) two-dimensional and (b) three-dimensional space.

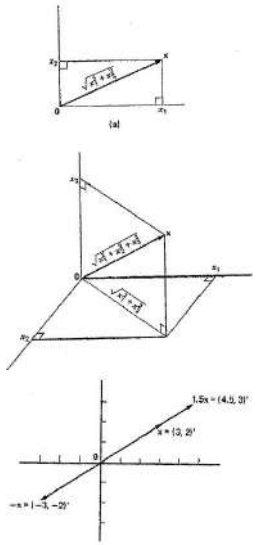


- Sum of two vectors $\mathbf{x}_1 + \mathbf{x}_2$: place "tail" of one at tip of other (left figure below)
- Difference of two vectors $\mathbf{x}_1 - \mathbf{x}_2$, or $\mathbf{x}_2 - \mathbf{x}_1$ (right figure below)



Length of a vector

- Length of vector $\mathbf{x}$: $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{n} x_i^2}$.
- Follows from theorem of Pythagoras in two dimensions.
- Distance between two vectors $\mathbf{x}_1$ and $\mathbf{x}_2$: $\|\mathbf{x}_1 - \mathbf{x}_2\| = \|\mathbf{x}_2 - \mathbf{x}_1\|$.
- Product of scalar and vector $a\mathbf{x}$ is vector of length $|a| \times \|\mathbf{x}\|$.
- If scalar a is positive, orientation of $a\mathbf{x}$ is same as of $\mathbf{x}$; if a is negative, $a\mathbf{x}$ points in opposite direction.

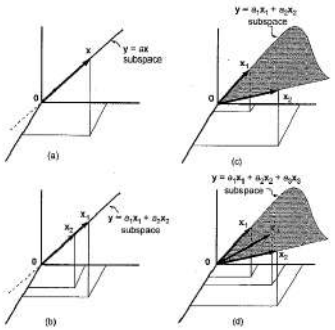


Vector spaces and subspaces (1)

- Vector space of dimension n is infinite set of all vectors $\mathbf{x} = (x_1, \ldots, x_n)$ with x_i any real number.
- Vector space of dimension 1 is real line; vector space of dimension 2 is plane.
- Subspace (of n-dimensional vector space), generated by set of k vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, consists of the subset of vectors $\mathbf{y}$ in the vector space that can be expressed as linear combinations of the generating set: $\mathbf{y} = a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + \cdots + a_k\mathbf{x}_k$.
- $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is said to span subspace that it generates.
- Set of vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ is linearly independent, if no vector in this set can be expressed as linear combination other vectors: it is not possible to write any $\mathbf{x}_j$ as $\mathbf{x}_j = a_1\mathbf{x}_1 + \cdots + a_{j-1}\mathbf{x}_{j-1} + a_{j+1}\mathbf{x}_{j+1} + \cdots + a_k\mathbf{x}_k$.
- Equivalently, set of vectors is linearly independent if there are no constants $b_1, b_2, \ldots, b_k$, not all 0, for which $b_1\mathbf{x}_1 + b_2\mathbf{x}_2 + \cdots + b_k\mathbf{x}_k = 0$.
- If such a relation does exist, the vectors comprise a linearly dependent set. The relation is called a linear dependency or collinearity.
- Dimension of subspace spanned by set of vectors is number of vectors in the largest linearly independent subset.
- Dimension of subspace cannot be larger than the smaller of k and n.

Vector spaces and subspaces (2)

- Example of 3-dimensional vector space:
  - (a) One-dimensional subspace (line) generated by single nonzero vector $\mathbf{x}$.
  - (b) One-dimensional subspace (line) generated by two collinear vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.
  - (c) Two-dimensional subspace (plane) generated by two linearly independent vectors $\mathbf{x}_1$ and $\mathbf{x}_2$.
  - (d) Two-dimensional subspace (plane) generated by three linearly dependent vectors.



- Linearly independent set of vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ provides basis for subspace it spans: any vector in this subspace can be written uniquely as linear combination of basis vectors $\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k$.
- Constants $c_1, c_2, \ldots, c_k$ are called coordinates of $\mathbf{y}$ with respect to basis $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$.
- Finding coordinates algebraically entails solving system of linear equations in which c_j's are unknowns: $\mathbf{y} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_k\mathbf{x}_k = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k] \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} = \mathbf{Xc}$.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

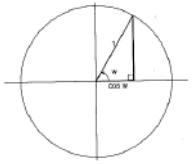
Week 3, lecture 13

Content week 3, lecture 13: Appendix (Fox) B.3.1,B.3.2,B4,B5,B6; calculation rules E / Var

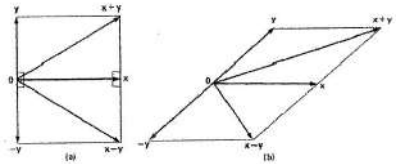
- Vector space and subspace (cont.)
  - cosine of angle
  - orthogonality of vectors x and y
  - orthogonal projection of vector y onto vector x, onto subspace
- Matrix rank and solution of linear simultaneous equations
  - row space of matrix; rank of matrix: dimension of row space
  - reduced row echelon form
  - linear simultaneous equations, unique solution, inconsistent (overdetermined) or consistent (underdetermined) system
- Eigenvalues and eigenvectors of square matrix A
- Quadratic form in x and positive-(semi)definite square matrix
- Calculation rules for expectations and variance-covariance matrices of linear combinations of random variables and random vectors

Orthogonality and orthogonal projections (1)

- Figure shows circle with radius 1 centered at origin.
- Angle w produces right triangle inscribed in circle; $cos(w)$ is signed length of side of triangle adjacent to angle.



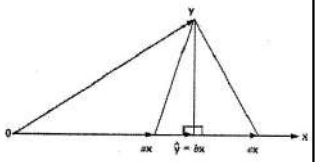
- Two vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal (perpendicular) if their inner product $\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i y_i = 0$.
- Figure shows geometry.
- In plot (a) $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. In that case $\|\mathbf{x}+\mathbf{y}\| = \|\mathbf{x}-\mathbf{y}\|$. As squared length of vector is inner product of vector with itself, we have $(\mathbf{x}+\mathbf{y}) \cdot (\mathbf{x}+\mathbf{y}) = (\mathbf{x}-\mathbf{y}) \cdot (\mathbf{x}-\mathbf{y})$, which simplifies into $\mathbf{x} \cdot \mathbf{y} = 0$.
- In plot (b) $\mathbf{x}$ and $\mathbf{y}$ are not orthogonal, and $\|\mathbf{x}+\mathbf{y}\| \neq \|\mathbf{x}-\mathbf{y}\|$, and $\mathbf{x} \cdot \mathbf{y} \neq 0$.



Orthogonality and orthogonal projections (2)

- Matrix $\mathbf{X}$ is orthogonal if each pair of its columns is orthogonal, i.e. $\mathbf{X}'\mathbf{X}$ is diagonal.
- Matrix $\mathbf{X}$ is orthonormal if $\mathbf{X}'\mathbf{X} = \mathbf{I}$.
- Orthogonal projection of vector $\mathbf{y}$ onto another vector $\mathbf{x}$, is scalar multiple $\widehat{\mathbf{y}} = b\mathbf{x}$ of $\mathbf{x}$, such that $(\mathbf{y} - \widehat{\mathbf{y}})$ is orthogonal to $\mathbf{x}$.
- To find b, note that $\mathbf{x} \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x} \cdot (\mathbf{y} - b\mathbf{x}) = 0$, thus $\mathbf{x} \cdot \mathbf{y} - b\mathbf{x} \cdot \mathbf{x} = 0$, and $b = (\mathbf{x} \cdot \mathbf{y})/(\mathbf{x} \cdot \mathbf{x})$.
- Orthogonal projection can be used to determine angle w between two vectors, by findings its cosine:

$$\cos(w) = \frac{\|\widehat{\mathbf{y}}\|}{\|\mathbf{y}\|} = \frac{b\|\mathbf{x}\|}{\|\mathbf{y}\|} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|^2} \times \frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \times \|\mathbf{y}\|}.$$



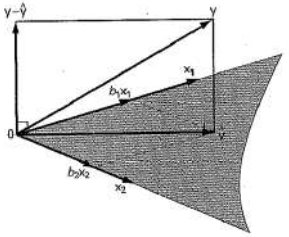
Orthogonality and orthogonal projections (3)

- Orthogonal projection of vector $\mathbf{y}$ onto subspace spanned by set of vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ is linear combination $\widehat{\mathbf{y}} = b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k$
- In that case $(\mathbf{y} - \widehat{\mathbf{y}})$ is orthogonal to every vector $\mathbf{x}_j$.
- Placing b_j's into vector $\mathbf{b}$ and vectors $\mathbf{x}_j$ into matrix $\mathbf{X} \equiv [\mathbf{x}_1, \ldots, \mathbf{x}_k]$, we have $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}$.
- Orthogonal projection: $\mathbf{x}_j \cdot (\mathbf{y} - \widehat{\mathbf{y}}) = \mathbf{x}_j \cdot (\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$ for all $j = 1, \ldots, k$.
- Equivalently, $\mathbf{X}'(\mathbf{y} - \mathbf{X}\mathbf{b}) = 0$, or $\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\mathbf{b}$.
- This matrix equation can be solved uniquely for $\mathbf{b}$ as long as $\mathbf{X}'\mathbf{X}$ is nonsingular:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

.

- Here is the least-squares solution for the linear model!



Matrix rank (1)

- Row space of $(m \times n)$ matrix $\mathbf{A}$ is subspace of n-dimensional vector space, spanned by the m rows of $\mathbf{A}$.
- Rank of $\mathbf{A}$ is dimension of its row space, that is, the maximum number of linearly independent rows in $\mathbf{A}$.
- $\text{rank}(\mathbf{A}) \leq \min(m, n)$.
- Matrix is in reduced row-echelon form (RREF) if it satisfies criteria:
  - R1 All of its nonzero rows (if any) precede all of its zero rows (if any).
  - R2 First nonzero entry (from left to right) in each nonzero row, called the leading entry in the row, is 1.
  - R3 Leading entry in each nonzero row (after the first row) is to the right of the leading entry in the previous row.
  - R4 All other entries are 0 in a column containing a leading entry.
- Rank of a matrix in RREF is equal to number of nonzero rows in the matrix.

Matrix rank (2)

- Matrix can be placed in RREF by sequence of elementary row operations, like in Gaussian elimination.
- Example white board.
- Rank of matrix $\mathbf{A}$ is equal to rank of its reduced row-echelon form $\mathbf{A}_R$.
- RREF of nonsingular square matrix is identity matrix, and rank equals order.
- Rank of $\mathbf{A}$ also equal to dimension of column space.

Linear simultaneous equations (1)

- System of m linear simultaneous equations in n unknowns $x_1, \ldots, x_n$ in matrix form:

$$\mathbf{Ax} = \mathbf{b}$$

  with $\mathbf{A}$ of order $m \times n$, $\mathbf{x}$ of order $n \times 1$, and $\mathbf{b}$ of order $m \times 1$; coefficient matrix $\mathbf{A}$ and rhs vector $\mathbf{b}$ are known constants, $\mathbf{x}$ is vector of unknowns.
- Suppose $m = n$: equal number of equations and unknowns.
  - If $\mathbf{A}$ is nonsingular, a unique solution exists: $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$.
  - If $\mathbf{A}$ is singular:
    - $\mathbf{A}$ can be transformed to RREF using elementary row operations: $\mathbf{A}_R = \mathbf{E}_p \cdots \mathbf{E}_2\mathbf{E}_1\mathbf{A} = \mathbf{EA}$.
    - Apply this to system of equations:

$$\mathbf{EAx} = \mathbf{Eb} \Leftrightarrow \mathbf{A}_R\mathbf{x} = \mathbf{b}_R$$

    - This system of equations is equivalent to original set $\mathbf{Ax} = \mathbf{b}$: any solution that satisfies one, satisfies other.
    - Let r be the rank of $\mathbf{A}$: because $\mathbf{A}$ is singular, $r < n$, and $\mathbf{A}_R$ contains r nonzero rows, and $n - r$ zero rows.
    - If any zero row of $\mathbf{A}_R$ associates with nonzero entry (say b) in $\mathbf{b}_R$, then system of equations is inconsistent or over-determined, for it contains the "equation" $0x_1 + \cdots + 0x_n = b \neq 0$.
    - If every zero row of $\mathbf{A}_R$ corresponds to zero entry in $\mathbf{b}_R$, then system of equations is consistent, and infinitely many solutions satisfy system: $n - r$ of unknowns may be given arbitrary values, which determine values of remaining r unknowns. System of equations is called under-determined.

Linear simultaneous equations (2)

- Suppose there are fewer equations than unknowns: $m < n$.
  - Then also $r < n$, and equations are either over-determined (if zero row of $\mathbf{A}_R$ corresponds to nonzero entry of $\mathbf{b}_R$) or under-determined (if system is consistent).
  - Example whiteboard.
- Suppose there are more equations than unknowns: $m > n$.
  - Suppose $\mathbf{A}$ has full-column rank ($r = n$):
    - $\mathbf{A}_R$ consists of order-n identity matrix, followed by $m - r$ zero rows.
    - If equations are consistent, they therefore have unique solution; otherwise they are over-determined.
  - Suppose $\mathbf{A}$ does not have full-column rank ($r < n$); then equations are either over-determined (if inconsistent) or under-determined (if consistent).
- Homogeneous systems of equations: $\mathbf{Ax} = \mathbf{0}$. Trivial solution: $\mathbf{x} = \mathbf{0}$. Non-trivial solutions exist if $\text{rank}(\mathbf{A}) < n$.

Eigenvalues and eigenvectors

- Let $\mathbf{A}$ be order-n square matrix; look at homogeneous system of linear equations:

$$(\mathbf{A} - \lambda\mathbf{I}_n)\mathbf{x} = \mathbf{0} \quad (1)$$

- Has nontrivial solution for certain values of scalar λ.
- Nontrivial solutions exist when matrix $(\mathbf{A} - \lambda\mathbf{I}_n)$ is singular, i.e.

$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0 \quad (2)$$

- This is called characteristic equation of matrix $\mathbf{A}$.
- Eigenvalues are values of λ for which equation holds. Other names: characteristic roots or latent roots.
- Eigenvector is vector $\mathbf{x}_1$ satisfying the equation for particular eigenvalue λ_1. Other names: characteristic vector or latent vector of $\mathbf{A}$ associated with λ_1.
- If the action of a matrix on a (nonzero) vector changes its magnitude but not its direction, then the vector is called an eigenvector of that matrix. The eigenvector is, in effect, multiplied by a scalar, called the eigenvalue corresponding to that eigenvector.

Eigenvalues and eigenvectors: example (2×2) case

- $\det\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} = 0 \Leftrightarrow (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$
  $\Leftrightarrow \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0$
- You may remember the abc-formula from high school, that solves the equation:
  $\lambda_1 = \frac{1}{2}\left[a_{11} + a_{22} + \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right]$
  $\lambda_2 = \frac{1}{2}\left[a_{11} + a_{22} - \sqrt{(a_{11} + a_{22})^2 - 4(a_{11}a_{22} - a_{12}a_{21})}\right]$
- Notice that $\lambda_1 + \lambda_2 = a_{11} + a_{22}$: sum of eigenvalues is trace.
- Notice that $\lambda_1\lambda_2 = a_{11}a_{22} - a_{12}a_{21}$: product of eigenvalues is determinant.
- Example $\mathbf{A} = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$. Then $\lambda_1 = 1.5$, $\lambda_2 = 0.5$ (fill in formulae above).
- Finding eigenvectors associated with $\lambda_1 = 1.5$:
  $\begin{bmatrix} 1 - 1.5 & 0.5 \\ 0.5 & 1 - 1.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Leftrightarrow \begin{bmatrix} -0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}\begin{bmatrix} x_{11} \\ x_{21} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ yielding
  $\mathbf{x}_1 = x_{21}^*\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Likewise for $\lambda_2 = 0.5$: $\mathbf{x}_2 = x_{22}^*\begin{bmatrix} -1 \\ 1 \end{bmatrix}$.
- Set of eigenvectors associated to each eigenvalue spans one-dimensional subspace.
- Eigenvectors are orthogonal $\mathbf{x}_1 \cdot \mathbf{x}_2 = -x_{21}^*x_{22}^* + x_{21}^*x_{22}^* = 0$.

Eigenvalues and eigenvectors: properties

- Characteristic equation of $(n \times n)$ matrix is nth order polynomial in λ; therefore, there are n eigenvalues, not necessarily distinct.
- Sum of eigenvalues of $\mathbf{A}$ is trace of $\mathbf{A}$.
- Product of eigenvalues of $\mathbf{A}$ is determinant of $\mathbf{A}$.
- Number of nonzero eigenvalues of $\mathbf{A}$ is rank of $\mathbf{A}$.
- Singular matrix has at least one zero eigenvalue.
- If $\mathbf{A}$ is symmetric, all eigenvalues are real numbers.
- If eigenvalues distinct, then set of eigenvectors associated with particular eigenvalue spans one-dimensional subspace. If k eigenvalues are equal, then common set of eigenvectors spans k dimensional subspace.
- Eigenvectors associated with different eigenvalues are orthogonal.

Quadratic forms and positive definite matrices

- Expression $\mathbf{x}'\mathbf{A}\mathbf{x}$ is called quadratic form in $\mathbf{x}$.
- From now $\mathbf{A}$ is symmetric matrix.
- $\mathbf{A}$ is positive definite if quadratic form is positive for all nonzero $\mathbf{x}$.
- $\mathbf{A}$ is positive semidefinite if quadratic form is non-negative for all nonzero $\mathbf{x}$.
- Eigenvalues of positive-definite matrix are all positive; eigenvalues of positive-semidefinite are all positive or zero.
- If $\mathbf{A}$ is $(n \times n)$ and positive definite, and $\mathbf{B}$ is $(n \times m)$ with full column rank $m < n$, then the $(m \times m)$ matrix $\mathbf{C}$, with $\mathbf{C} = \mathbf{B}'\mathbf{A}\mathbf{B}$, is also positive-definite.
- If $\text{rank}(\mathbf{B}) < m$ then $\mathbf{C}$ is positive-semidefinite.

Calculation rules Expectation and Variances for scalar expressions

Suppose we have random variables X_1, X_2, and X_3 with expected values μ_1, μ_2 and μ_3, variances σ_1^2, σ_2^2 and σ_3^2, and covariances $C(X_1, X_2) = \sigma_{12}$, $C(X_1, X_3) = \sigma_{13}$ and $C(X_2, X_3) = \sigma_{23}$.
Recall the calculation rules:

- $E(aX_1 + b) = aE(X_1) + b = a\mu_1 + b$
- $V(aX_1 + b) = a^2V(X_1) = a^2\sigma_1^2$
- $E(aX_1 + bX_2) = aE(X_1) + bE(X_2) = a\mu_1 + b\mu_2$
- $V(aX_1 + bX_2) = a^2V(X_1) + b^2V(X_2) + 2abC(X_1, X_2) = a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\sigma_{12}$
- $C(aX_1 + bX_2, cX_3) = ac\,C(X_1, X_3) + bc\,C(X_2, X_3) = ac\,\sigma_{13} + bc\,\sigma_{23}$

Calculation rules Expectation and Variances for matrix expressions

Suppose we collect variables X_1 and X_2 into vector $\mathbf{x} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with expected vector $E(\mathbf{x}) = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and variance-covariance matrix $V(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$.

Linear combinations $Y_1 = aX_1 + bX_2$ and $Y_2 = cX_1 + dX_2$ can be written as vector

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} aX_1 + bX_2 \\ cX_1 + dX_2 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \mathbf{A}\mathbf{x}, \text{ with } \mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

For matrices we have the calculation rules

- $E(A\mathbf{x}) = AE(\mathbf{x})$
- $V(A\mathbf{x}) = AV(\mathbf{x})A'$
- $C(A_1\mathbf{x}, A_2\mathbf{x}) = A_1V(\mathbf{x})A_2'$.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 3, lecture 14

Content lecture 14: Fox §9.1

- Statistical theory linear models
  - matrix form of linear model $\mathbf{y} = \mathbf{X}\beta + \epsilon$
  - model matrices for dummy-regression and anova models
  - contrasts; from intended $\beta = \mathbf{X}_B^{-1}\mu$ into $\mu = \mathbf{X}_B\beta$

Linear Models in Matrix Form

- So far, we have seen a number of models for quantitative response variables:
  - Linear regression models: quantitative explanatory variables.
  - Analysis of variance models: qualitative explanatory variables (factors).
  - Analysis of covariance models: both quantitative and qualitative explanatory variables.
- These models have a lot in common, like their assumptions:
  - Normality
  - Constant variance
  - Independence
  - Linearity in parameters, expected error zero.
- All are examples of Linear Models.
- It is very useful to write linear models in Matrix Form:

$$y = X\beta + \epsilon$$

Linear models in matrix form

- General linear model for observation $i(= 1, \ldots, n)$ given by $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$.
- Collecting regressor values of observation i into row vector, results in:

$$Y_i = [1, x_{i1}, x_{i2}, \ldots, x_{ik}] \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \epsilon_i = \mathbf{x}_i'\beta + \epsilon_i \ .$$

- Collecting all n observations into matrix equation:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad \text{or } \mathbf{y} = \mathbf{X}\beta + \epsilon$$

- $\mathbf{X}$ is called the model matrix or design matrix.

Assumptions linear models in matrix form

- Assumptions of linear model:
  - errors independent
  - errors normally distributed
  - errors common variance σ^2_ϵ
  - errors with zero expectation
- Written compactly in matrix form: $\epsilon \sim N_n(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_n)$: ϵ has multivariate normal distribution with
  - expectation $E(\epsilon) = \mathbf{0}$
  - covariance matrix $V(\epsilon) = E(\epsilon\epsilon') = \sigma^2_\epsilon \mathbf{I}_n$.
- For the response $\mathbf{y}$ we have : $\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2_\epsilon \mathbf{I}_n)$ because
  - $\mu \equiv E(\mathbf{y}) = E(\mathbf{X}\beta + \epsilon) = \mathbf{X}\beta + E(\epsilon) = \mathbf{X}\beta$
  - $V(\mathbf{y}) = E[(\mathbf{y} - \mu)(\mathbf{y} - \mu)'] = E[(\mathbf{y} - \mathbf{X}\beta)(\mathbf{y} - \mathbf{X}\beta)'] = E(\epsilon\epsilon') = \sigma^2_\epsilon \mathbf{I}_n$

Dummy regression and analysis of variance

- Example overparameterized one-way ANOVA model: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ for groups $j = 1, \ldots, m$.
  In matrix form

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{n_1,1} \\ \hline Y_{12} \\ \vdots \\ Y_{n_2,2} \\ \hline \vdots \\ Y_{1,m-1} \\ \vdots \\ Y_{n_{m-1},m-1} \\ \hline Y_{1m} \\ \vdots \\ Y_{n_m,m} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 1 & 0 & \cdots & 0 & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 1 & \cdots & 0 & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 1 & 0 \\ \hline 1 & 0 & 0 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \\ \alpha_m \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \vdots \\ \epsilon_{n_1,1} \\ \hline \epsilon_{12} \\ \vdots \\ \epsilon_{n_2,2} \\ \hline \vdots \\ \epsilon_{1,m-1} \\ \vdots \\ \epsilon_{n_{m-1},m-1} \\ \hline \epsilon_{1m} \\ \vdots \\ \epsilon_{n_m,m} \end{bmatrix}$$

- Model matrix $\mathbf{X}$ has rank m, one less than number of columns, because first column is sum of others.
- Solution: delete one column, implicitly setting corresponding parameter to 0. E.g. deleting second column sets $\alpha_1 = 0$, as R does.

Dummy regression and analysis of variance

- Matrices containing dummy regressors (e.g. in ANOVA and ANCOVA) are highly structured.
- Example 1: ANCOVA model with Y income, x years of education, dummy regressor d coding for men, we may form model $Y_i = \alpha + \beta x_i + \gamma d_i + \delta(x_i d_i) + \epsilon_i$. Suppose that first n_1 observations are women. Model in matrix form is:

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ \hline Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_1} & 0 & 0 \\ \hline 1 & x_{n_1+1} & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_{n_1} \\ \hline \epsilon_{n_1+1} \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Example sigma constraint

- Example overparameterized one-way ANOVA model with sigma constraint:

$$\mathbf{X}_F\beta = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 0 & \cdots & 0 \\ \hline 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 1 & \cdots & 0 \\ \hline \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ \hline 1 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & -1 & -1 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}$$

- Model matrix $\mathbf{X}_F$ is of full column rank.

Example sigma constraint continued

- Relationship between group ("cell") means $\mu = \mu_j$ and parameters of constrained model: $\begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_{m-1} \\ \mu_m \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 1 \\ 1 & -1 & -1 & \cdots & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}$ or $\mu = \mathbf{X}_B \beta_F$
- β_F contains parameters according to sum-to-zero parameterization.
- Rows of $\mathbf{X}_B$ form row basis of full-rank model matrix $\mathbf{X}_F$.
- $\mathbf{X}_B$ has full (column) rank, hence nonsingular, hence can be inverted, solving uniquely for the constrained parameters in terms of cell means: $\beta_F = \mathbf{X}_B^{-1}\mu$
- Solution is $\begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m-1} \end{bmatrix} = \begin{bmatrix} \mu_\cdot \\ \mu_1 - \mu_\cdot \\ \mu_2 - \mu_\cdot \\ \vdots \\ \mu_{m-1} - \mu_\cdot \end{bmatrix}$

Linear contrasts

- In the ANOVA model the relationship between group means and parameters is given by equation $\mu = \mathbf{X}_B \beta_F$.
- Thus, parameters are linear functions of group means: $\beta_F = \mathbf{X}_B^{-1}\mu$.
- Full rank parameterizations of one-way ANOVA (dummy coding, deviation coding) allow easy testing H_0 : no differences among group means, but in general without explicit interest in individual parameters.
- Sometimes, we want to formulate $\mathbf{X}_B$ so that individual parameters of β_F represent interesting contrasts among group means.

Linear contrasts (2)

- Example Friendly memory experiment: 3 experimental conditions: 1) SFR (words presented in random order) 2) B = "before" (words, remembered on previous trial, were presented prior to forgotten words) 3) M = "meshed" (remembered words interspersed with forgotten words, but in same order as originally recalled).
- Linear ("1 df") contrasts for two null hypotheses:
  1. $H_0 : \mu_1 = (\mu_2 + \mu_3)/2 \Leftrightarrow \mu_1 - \frac{1}{2}\mu_2 - \frac{1}{2}\mu_3 = 0$
  2. $H_0 : \mu_2 = \mu_3 \Leftrightarrow \mu_2 - \mu_3 = 0$.

  Code each hypothesis as parameter of model, employing matrix $\mathbf{X}_B^{-1}$:
  $\begin{bmatrix} \mu \\ \zeta_1 \\ \zeta_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$
- Parameter μ is the average of group means, ζ_1 represents the difference in means of group 1 and the average of groups 2 and 3, and ζ_2 represents the differences in group means of groups 2 and 3.
- Notice that the rows of $\mathbf{X}_B^{-1}$ are orthogonal: their dot-product is 0.

Linear contrasts (3)

- First note that $\mathbf{X}_B^{-1}(\mathbf{X}_B^{-1})' = \begin{bmatrix} \frac{3}{9} & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & 2 \end{bmatrix}$, because rows of $\mathbf{X}_B^{-1}$ are orthonogal.

  Therefore, $\mathbf{X}_B^{-1}(\mathbf{X}_B^{-1})' \begin{bmatrix} \frac{9}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{3}{9} & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} \frac{9}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

  It must be that the inverse of $\mathbf{X}_B^{-1}$ is

  $\mathbf{X}_B = (\mathbf{X}_B^{-1})' \begin{bmatrix} \frac{9}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{3} & 1 & 0 \\ \frac{1}{3} & -\frac{1}{2} & 1 \\ \frac{1}{3} & -\frac{1}{2} & -1 \end{bmatrix} \begin{bmatrix} \frac{9}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 1 & \frac{2}{3} & 0 \\ 1 & -\frac{1}{3} & \frac{1}{2} \\ 1 & -\frac{1}{3} & -\frac{1}{2} \end{bmatrix}$.
- So, columns of $\mathbf{X}_B$ are rows of $\mathbf{X}_B^{-1}$, divided by sum of squared entries in this row.
- Rescaling the columns does not have an effect on hypothesis testing, and makes coding more convenient: $\mathbf{X}_B = \begin{bmatrix} 1 & 2 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$

Linear contrasts (4)

- Orthogonal contrasts are convenient, but not necessary. Enough to have linear independent contrasts, and work back from $\mathbf{X}_B^{-1}$ towards the model $\mu = \mathbf{X}_B\beta_F$.
- With equal numbers of observations n per group, an orthogonal model matrix $\mathbf{X}_B$ implies an orthogonal full-rank matrix $\mathbf{X}_F$. Columns of an orthogonal model matrix represent independent sources of variation in response variable, and therefore set of orthogonal contrasts partitions regression sum of squares into one-degree-of-freedom components, each testing a hypothesis of interest.
- Linear comparisons are of interest even with unequal group frequencies, causing contrasts orthogonal in $\mathbf{X}_B$ to be correlated in $\mathbf{X}_F$.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 3, lecture 15

Content lecture 15: Fox §9.2-9.3

- Statistical theory linear models
  - least-squares in matrix notation
  - normal equations $\boxed{(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}}$
  - least-squares solution: $\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$
  - distribution least-squares estimator: $\boxed{\mathbf{b} \sim N_{k+1}\left[\beta, \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right]}$
  - Gauss-Markov theorem
  - maximum-likelihood estimation for linear models

Least-squares fit

Linear model: $\mathbf{y} = \mathbf{X}\beta + \epsilon$
with variance-covariance matrix for error vector: $V(\epsilon) = \sigma_\epsilon^2\mathbf{I}_n$

- Fitting model to data gives vectors of fitted values and residuals: $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$ with $\mathbf{b} = [B_0, B_1, \ldots, B_k]'$ vector of estimated coefficients, $\mathbf{e} = [E_1, E_2, \ldots, E_n]'$ vector of residuals. But how is $\mathbf{b}$ obtained?
- Find $\mathbf{b}$ that minimizes residual sum of squares: $S(\mathbf{b}) = \sum E_i^2 = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{y}'\mathbf{y} - (2\mathbf{y}'\mathbf{X})\mathbf{b} + \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}$ (why is $\mathbf{y}'\mathbf{X}\mathbf{b} = \mathbf{b}'\mathbf{X}'\mathbf{y}$?)
- To minimize $S(\mathbf{b})$ we need vector of partial derivatives w.r.t. $\mathbf{b}$: $\frac{\partial S(\mathbf{b})}{\partial \mathbf{b}} = \mathbf{0} - 2\mathbf{X}'\mathbf{y} + 2(\mathbf{X}'\mathbf{X})\mathbf{b}$.
- Setting derivative to $\mathbf{0}$, leads to normal equations:

$$\boxed{(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}}$$

- $k + 1$ equations for $k + 1$ coefficients.

Least-squares fit (2)

- If $\mathbf{X}'\mathbf{X}$ nonsingular (rank $k+1$), then unique least-squares solution is:

$$\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$$

- Rank of $\mathbf{X}'\mathbf{X}$ is equal to rank of $\mathbf{X}$.
- Matrix with second partial derivatives of sum of squared residuals is $\frac{\partial^2 S(\mathbf{b})}{\partial \mathbf{b} \partial \mathbf{b}'} = 2(\mathbf{X}'\mathbf{X})$. Because $\mathbf{X}'\mathbf{X}$ is positive definite when $\mathbf{X}$ is of full rank, solution $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ represents minimum of $S(\mathbf{b})$.

Least-squares fit: example

- Matrix $\mathbf{X}'\mathbf{X}$ contains sums of squares and products among regressors. $\mathbf{X}'\mathbf{y}$ contains sums of cross products of regressors and response.
- Example Duncan's `prestige` data.

```
> X <- cbind(1,Duncan$income,Duncan$education); y <- Duncan$prestige
> t(X) %*% X
     [,1]   [,2]   [,3]
[1,]   45   1884   2365
[2,] 1884 105148 122197
[3,] 2365 122197 163265
> t(X) %*% y
       [,1]
[1,]   2146
[2,] 118229
[3,] 147936
> solve(t(X) %*% X)
          [,1]      [,2]      [,3]
[1,]  0.102106 -8.50e-04 -8.43e-04
[2,] -0.000850  8.01e-05 -4.77e-05
[3,] -0.000843 -4.77e-05  5.40e-05
> (b <- solve(t(X) %*% X) %*% t(X) %*% y)
       [,1]
[1,] -6.065
[2,]  0.599
[3,]  0.546
```


Distribution of least-squares estimator

- $\mathbf{b}$ is linear estimator: $\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{M}\mathbf{y}}$ with $\mathbf{M} \equiv (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- $\mathbf{b}$ is unbiased estimator: $E(\mathbf{b}) = E(\mathbf{M}\mathbf{y}) = \mathbf{M}E(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \mathbf{X}\beta = \beta$.
- $\mathbf{b}$ has variance-covariance matrix:

$$\boxed{V(\mathbf{b}) = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}}$$

because
$V(\mathbf{b}) = \mathbf{M}V(\mathbf{y})\mathbf{M}' = [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\,\sigma_\epsilon^2\mathbf{I}_n\,[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\ \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.
Sampling variances and covariances only depend on model matrix and error variance.

- $\mathbf{b}$ has normal distribution, *if* $\mathbf{y}$ has normal distribution (because $\mathbf{b}$ is linear in $\mathbf{y}$).

Hence:

$$\boxed{\mathbf{b} \sim N_{k+1}\left[\beta, \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right]}$$

Gauss-Markov theorem

- If errors are independent with zero expectation and constant variance, then least-squares estimator $\mathbf{b}$ is the most efficient estimator within the class of linear unbiased estimators: it has smallest sampling variance.
- Least-squares estimator is BLUE = Best Linear Unbiased Estimator.
- Gauss-Markov theorem is justification of least-squares estimation!
- Under normality, least-squares estimator is most efficient of all unbiased estimators.

Gauss-Markov theorem: proof

Let's follow the proof of the Gauss-Markov theorem:

- Let $\tilde{\mathbf{b}}$ be best linear unbiased estimator of β.
- Least-squares estimator $\mathbf{b}$ is linear: $\mathbf{b} = \mathbf{My}$.
- Write $\tilde{\mathbf{b}} = (\mathbf{M} + \mathbf{A})\mathbf{y}$, with $\mathbf{A}$ the difference between the matrices for $\tilde{\mathbf{b}}$ and $\mathbf{b}$.
- We need to show that $\mathbf{A} = \mathbf{0}$:
  - $\tilde{\mathbf{b}}$ is unbiased and linear, so
    $\beta = E(\tilde{\mathbf{b}}) = E[(\mathbf{M} + \mathbf{A})\mathbf{y}] = E(\mathbf{My}) + E(\mathbf{Ay}) = E(\mathbf{b}) + \mathbf{A}E(\mathbf{y}) = \beta + \mathbf{AX}\beta$
    Therefore, $\mathbf{AX}\beta = \mathbf{0}$ for all β, so $\mathbf{AX} = \mathbf{0}$.
  - $V(\tilde{\mathbf{b}}) = (\mathbf{M}+\mathbf{A})\, V(\mathbf{y})\, (\mathbf{M}+\mathbf{A})' = (\mathbf{M}+\mathbf{A})\, \sigma^2_\epsilon \mathbf{I}_n\, (\mathbf{M}+\mathbf{A})' = \sigma^2_\epsilon(\mathbf{MM}'+\mathbf{MA}'+\mathbf{AM}'+\mathbf{AA}')$
  - Already shown that $\mathbf{AX} = \mathbf{0}$, hence $\mathbf{AM}' = \mathbf{AX}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$. So, $V(\tilde{\mathbf{b}}) = \sigma^2_\epsilon(\mathbf{MM}' + \mathbf{AA}')$
  - $\tilde{\mathbf{b}}$ has minimum variance; therefore, diagonal entries of $V(\tilde{\mathbf{b}})$ are as small as possible. Sampling variance of coefficient $\tilde{B}_j$ is jth diagonal entry of $V(\tilde{\mathbf{b}})$: $V(\tilde{B}_j) = \sigma^2_\epsilon(\sum_{i=1}^n m^2_{ji} + \sum_{i=1}^n a^2_{ji})$. Both sums are sums of squares, and cannot be negative. Because $V(\tilde{B}_j)$ is as small as possible, all a_{ji} must be 0. This argument applies to each coefficient in $\tilde{\mathbf{b}}$, so every row of $\mathbf{A}$ is $\mathbf{0}$, so $\mathbf{A} = \mathbf{0}$.
  - Finally: $\tilde{\mathbf{b}} = (\mathbf{M} + \mathbf{0})\mathbf{y} = \mathbf{My} = \mathbf{b}$. QED

Maximum-likelihood estimation

- Under assumptions of linear model, least-squares estimator $\mathbf{b}$ is also maximum-likelihood estimator of β, providing basis for generalizing the linear model.
- Linear model with assumptions for vector $\mathbf{y}$: $\mathbf{y} \sim N_n\left(\mathbf{X}\beta, \sigma^2_\epsilon \mathbf{I}_n\right)$, for ith observation: $Y_i \sim N(\mathbf{x}_i'\beta, \sigma^2_\epsilon)$.
- Probability density for observation i is $p(y_i) = \frac{1}{\sigma_\epsilon\sqrt{2\pi}} exp\left[-\frac{(y_i - \mathbf{x}_i'\beta)^2}{2\sigma^2_\epsilon}\right]$.
- For vector of n independent observations joint probability density is product of marginal densities:
  $p(\mathbf{y}) = \frac{1}{\left(\sigma_\epsilon\sqrt{2\pi}\right)^n} exp\left[-\frac{\sum(y_i - \mathbf{x}_i'\beta)^2}{2\sigma^2_\epsilon}\right] = \frac{1}{\left(\sigma_\epsilon\sqrt{2\pi}\right)^n} exp\left[-\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2_\epsilon}\right]$.
- Log likelihood is $\boxed{\log_e L(\beta, \sigma^2_\epsilon) = -\frac{n}{2}log_e(2\pi) - \frac{n}{2}log_e(\sigma^2_\epsilon) - \frac{1}{2\sigma^2_\epsilon}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}$.

Maximum-likelihood estimation (2)

- So, Log likelihood is $\boxed{\log_e L(\beta, \sigma^2_\epsilon) = -\frac{n}{2}log_e(2\pi) - \frac{n}{2}log_e(\sigma^2_\epsilon) - \frac{1}{2\sigma^2_\epsilon}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}$.
- To maximize, calculate partial derivatives w.r.t. β and σ^2_ϵ:
  $\frac{\partial \log_e L(\beta, \sigma^2_\epsilon)}{\partial \beta} = \frac{1}{2\sigma^2_\epsilon}(2\mathbf{X}'\mathbf{X}\beta - 2\mathbf{X}'\mathbf{y})$ and $\frac{\partial \log_e L(\beta, \sigma^2_\epsilon)}{\partial \sigma^2_\epsilon} = -\frac{n}{2}\frac{1}{\sigma^2_\epsilon} + \frac{1}{2\sigma^4_\epsilon}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$.
- Setting to 0 and solving brings m.l. estimators $\hat{\beta}$ and $\hat{\sigma}^2_\epsilon$:
  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ and $\hat{\sigma^2_\epsilon} = \frac{(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)}{n} = \frac{\mathbf{e}'\mathbf{e}}{n}$.
- Minimizing sum of squared residuals, maximizes the likelihood.
- $\hat{\sigma^2_\epsilon}$ is biased. Unbiased estimator $S^2_E = \mathbf{e}'\mathbf{e}/(n - (k+1))$ is preferred. Notice that bias shrinks to 0 if n increases: m.l. estimator is consistent.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 3, lecture 16

Content lecture 16: Fox §9.4

- Statistical theory linear models: statistical inference
  - inference for individual coefficients: t-tests and confidence intervals
  - inference for several coefficients: LR and F-tests
  - general linear hypotheses
  - joint confidence regions

General Linear Model Theory

Recall:

- Linear model: $\boxed{\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon}$
  with assumptions for error vector: $\boxed{\epsilon \sim N_n\left[\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n\right]}$
  and $\mathbf{X}$ model matrix of order $(n \times (k+1))$
- To obtain least-squares estimators $\mathbf{b}$ of $\boldsymbol{\beta}$ solve normal equations: $\boxed{(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}}$
- LS estimator (for $\mathbf{X}$ full column rank) is $\boxed{\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}}$
- $\mathbf{b}$ has variance-covariance matrix: $\boxed{V(\mathbf{b}) = \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}}$
- $\mathbf{b}$ is distributed as $\boxed{\mathbf{b} \sim N_{k+1}\left[\boldsymbol{\beta}, \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right]}$

Statistical inference for individual coefficients

- Vector of coefficients $\mathbf{b} \sim N_{k+1}\left(\boldsymbol{\beta}, \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\right)$
- Individual coefficient $B_j \sim N\left(\beta_j, \sigma_\epsilon^2 v_{jj}\right)$, with v_{jj} jth diagonal entry of $(\mathbf{X}'\mathbf{X})^{-1}$.
- So, $(B_j - \beta_j)/\sigma_\epsilon\sqrt{v_{jj}} \sim N(0,1)$.
- For testing $H_0 : \beta_j = \beta_j^{(0)}$, if σ_ϵ^2 would be known, test-statistic $Z = \frac{B_j - \beta_j^{(0)}}{\sigma_\epsilon\sqrt{v_{jj}}}$, which has $N(0,1)$-distribution if H_0 is true.
- But usually σ_ϵ^2 is not known...

Statistical inference for individual coefficients (2)

- Generally, σ_ϵ^2 is unknown, and is estimated by $S_E^2 = \mathbf{e}'\mathbf{e}/(n-(k+1))$, with $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b}$ is vector or residuals
- Estimator of covariance matrix $\boxed{\hat{V}(\mathbf{b}) = S_E^2(\mathbf{X}'\mathbf{X})^{-1}}$.
- Standard error $\boxed{SE(B_j) = S_E\sqrt{v_{jj}}}$.
- Recall:
  - $(n-(k+1))S_E^2/\sigma_\epsilon^2 \sim \chi^2$ with $n-(k+1)$ df.
  - Taking the ratio of a standard normally distributed quantity, and square root of chi-square distributed quantity $/df$, independent, gives the t-statistic $t = \dfrac{(B_j - \beta_j)/\sigma_\epsilon\sqrt{v_{jj}}}{\sqrt{S_E^2/\sigma_\epsilon^2}} = \dfrac{B_j - \beta_j}{S_E\sqrt{v_{jj}}}$, which has t-distribution with $n-(k+1)$ d.f.

Statistical inference for individual coefficients (3)

- To test $H_0 : \beta_j = \beta_j^{(0)}$, use test statistic $\boxed{t = \dfrac{B_j - \beta_j^0}{SE(B_j)}}$.
- If H_0 is true, $t \sim t_{n-(k+1)}$.
- $100(1-\alpha)\%$ confidence interval for β_j: $\boxed{B_j \pm t_{\alpha/2,n-(k+1)} SE(B_j)}$.

```
> lm1 <- lm(prestige ~ income + education, data=Duncan) # Use linear model facilities
> SE2 <- anova(lm1)[3,3]; print(SE2, digits=6)
[1] 178.731
> vcov(lm1)  # variance-covariance matrix of b
            (Intercept)   income education
(Intercept)      18.249 -0.15185  -0.15071
income           -0.152  0.01432  -0.00852
education        -0.151 -0.00852   0.00965
> coef(summary(lm1))  # ls estimates with se, t-test statistics and P-values
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -6.065     4.2719   -1.42 1.63e-01
income         0.599     0.1197    5.00 1.05e-05
education      0.546     0.0983    5.56 1.73e-06
> confint(lm1)  # 95% confidence intervals for coefficients
              2.5 % 97.5 %
(Intercept) -14.686  2.556
income        0.357  0.840
education     0.348  0.744
```


Statistical inference for individual coefficients (4)

- Reproduce some `lm()` results using matrix manipulations:

```
> e <- y-X %*% b  # vector of residuals
> (ete <- t(e) %*% e)  # residual sum of squares
     [,1]
[1,] 7507
> (SE2 <- ete/(nrow(X)-ncol(X)))  # residual variance
     [,1]
[1,]  179
> dim(SE2) <- NULL  # otherwise SE2 would be 1x1 matrix
> (Covb <- SE2 * solve(t(X) %*% X)) # variance-covariance matrix of b
       [,1]     [,2]     [,3]
[1,] 18.249 -0.15185 -0.15071
[2,] -0.152  0.01432 -0.00852
[3,] -0.151 -0.00852  0.00965
> sqrt(diag(Covb))  # standard errors of coefficients
[1] 4.2719 0.1197 0.0983
```


Statistical inference for several coefficients

- Inference on groups of coefficients may be needed because
  - least-squares estimators are often correlated (off-diagonal elements of $V(\mathbf{b})$ non-zero).
  - interest in related set of coefficients, like in ANOVA.
- FM: $Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \epsilon$
  giving ls estimate $\mathbf{b} = [B_0, B_1, \ldots, B_k]'$ and ml estimate $\hat{\sigma}^2_\epsilon = \mathbf{e}'\mathbf{e}/n$.
- Test $H_0 : \beta_1 = \cdots = \beta_q = 0$, corresponding to RM:
  $Y = \beta_0 + 0x_1 + \ldots + 0x_q + \beta_{q+1}x_{q+1} + \cdots + \beta_k x_k + \epsilon = \beta_0 + \beta_{q+1}x_{q+1} + \cdots + \beta_k x_k + \epsilon$
  giving ls estimate $\mathbf{b}_0 = [B'_0, 0, \cdots, 0, B'_{q+1}, \ldots, B'_k]'$ and ml estimate $\hat{\sigma}^2_{\epsilon_0} = \mathbf{e}'_0\mathbf{e}_0/n$.
- Note $\hat{\sigma}^2_\epsilon \leq \hat{\sigma}^2_{\epsilon_0}$, because $\mathbf{e}'\mathbf{e} < \mathbf{e}'_0\mathbf{e}_0$

Statistical inference for several coefficients: LR-test

- Maximized likelihood for FM $L_1 = \left(2\pi e \dfrac{\mathbf{e}'\mathbf{e}}{n}\right)^{-n/2}$ and for RM $L_0 = \left(2\pi e \dfrac{\mathbf{e}'_0\mathbf{e}_0}{n}\right)^{-n/2}$ (this is exercise 9.9).
- Likelihood ratio is $\dfrac{L_0}{L_1} = \left(\dfrac{\mathbf{e}'_0\mathbf{e}_0}{\mathbf{e}'\mathbf{e}}\right)^{-n/2} = \left(\dfrac{\mathbf{e}'\mathbf{e}}{\mathbf{e}'_0\mathbf{e}_0}\right)^{n/2}$.
- Likelihood ratio test statistic is $G^2_0 = -2log_e(L_0/L_1)$, asymptotically distributed as χ^2_q.
- Likelihood ratio test is hardly ever used in linear model situation, because exact result is available: F-test.

Statistical inference for several coefficients: F-test

- With $RSS = \mathbf{e}'\mathbf{e}$ residual sum of squares of FM and $RS_0S = \mathbf{e}_0'\mathbf{e}_0$ residuals sum of squares of RM, we have F-ratio

$$F_0 = \frac{(RSS_0 - RSS)/q}{RSS/(n-(k+1))} \sim F_{q,n-(k+1)}$$

  the well known incremental F-statistic.
- Recall that ratio of two independent χ^2 distributed random variables / d.f. has F-distribution:
  - $RSS/\sigma_\epsilon^2 = \mathbf{e}'\mathbf{e}/\sigma_\epsilon^2 \sim \chi^2_{n-(k+1)}$
  - $RSS_0/\sigma_\epsilon^2 = \mathbf{e}_0'\mathbf{e}_0/\sigma_\epsilon^2 \sim \chi^2_{n-(k+1-q)}$ if H_0 is true
  - Consequently, $(RSS_0 - RSS)/\sigma_\epsilon^2 \sim \chi^2_q$ if H_0 is true
  - $(RSS_0 - RSS)/\sigma_\epsilon^2$ and RSS/σ_ϵ^2 are independent.

Statistical inference for several coefficients: F-test (2)

- Incremental sum of squares $RSS_0 - RSS$ can be calculated directly from $\mathbf{b}$ and $(\mathbf{X}'\mathbf{X})^{-1}$ for FM (so without fitting RM):
  - Let $\mathbf{b}_1 = [B_1, \cdots, B_q]'$ be ls coefficients of interest from $\mathbf{b}$; let $\mathbf{V}_{11}$ be corresponding submatrix of $(\mathbf{X}'\mathbf{X})^{-1}$.
  - It can be shown that $RSS_0 - RSS = \mathbf{b}_1'\mathbf{V}_{11}^{-1}\mathbf{b}_1$, so incremental F-statistic $F_0 = \mathbf{b}_1'\mathbf{V}_{11}^{-1}\mathbf{b}_1/qS_E^2$.
  - Test general hypothesis $H_0 : \beta_1 = \beta_1^{(0)}$ (not necessarily $\mathbf{0}$):

$$F_0 = \frac{(\mathbf{b}_1 - \beta_1^{(0)})'\mathbf{V}_{11}^{-1}(\mathbf{b}_1 - \beta_1^{(0)})}{qS_E^2} \sim_{H_0} F_{q,n-(k+1)}$$

- Recall omnibus F-statistic for testing $H_0 : \beta_1 = \ldots = \beta_k = 0$:

$$F_0 = \frac{RegSS/k}{RSS/(n-(k+1))}$$

  - Denominator estimates σ_ϵ^2 whether or not H_0 is true.
  - Numerator has expectation $E(RegSS/k) = \beta_1'(\mathbf{X}^{*\prime}\mathbf{X}^*)\beta_1/k + \sigma_\epsilon^2$ (with $\mathbf{X}^*$ matrix of mean deviation regressors, omitting constant regressor). This is larger than σ_ϵ^2, unless H_0 is true.
  - Therefore, if H_0 is not true, the numerator of F tends to be larger than the denominator, and the F-statistic tends to values larger than 1.

General linear hypotheses

- Linear hypothesis: $H_0 : \mathbf{L}\beta = \mathbf{c}$.
- Hypothesis matrix $\mathbf{L}$ is of full row rank $q \leq k+1$.
- F-statistic:

$$F_0 = \frac{(\mathbf{Lb} - \mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{Lb} - \mathbf{c})}{qS_E^2} \sim_{H_0} F_{q,n-(k+1)}$$

- because
  - $\mathbf{b} \sim N_{k+1}[\beta, \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}]$
  - $\mathbf{Lb} \sim N_q[\mathbf{L}\beta, \sigma_\epsilon^2\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']$
  - Under H_0, $\mathbf{L}\beta = \mathbf{c}$, so $(\mathbf{Lb} - \mathbf{c})'[\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}']^{-1}(\mathbf{Lb} - \mathbf{c})/\sigma_\epsilon^2 \sim \chi^2_q$
- Example: Duncan's data
  - Test $H_0 : \beta_1 = \beta_2 = 0$, take $\mathbf{L} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ and $\mathbf{c} = [0, 0]'$.
  - Test $H_0 : \beta_1 - \beta_2 = 0$, take $\mathbf{L} = [0, 1, -1]$ and $\mathbf{c} = [0]$.

Joint confidence regions

- Joint confidence region for vector β_1 can be constructed from the F-test:

$$F_0 = \frac{(\mathbf{b}_1 - \beta_1^{(0)})'\mathbf{V}_{11}^{-1}(\mathbf{b}_1 - \beta_1^{(0)})}{qS_E^2} \sim_{H_0} F_{q,n-(k+1)}$$

- If $H_0 : \beta_1 = \beta_1^{(0)}$ is true, then

$$Pr\left[\frac{(\mathbf{b}_1 - \beta_1^{(0)})'\mathbf{V}_{11}^{-1}(\mathbf{b}_1 - \beta_1^{(0)})}{qS_E^2} \leq F_{\alpha,q,n-(k+1)}\right] = 1 - \alpha$$

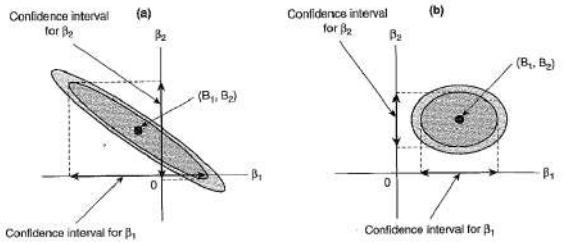
  with $F_{\alpha,q,n-(k+1)}$ the critical value of $F_{q,n-(k+1)}$ with right-tail probability α.
- Joint confidence region for β_1 consists of all β_1 for which $(\mathbf{b}_1 - \beta_1)'\mathbf{V}_{11}^{-1}(\mathbf{b}_1 - \beta_1) \leq qS_E^2F_{\alpha,q,n-(k+1)}$.
  - Boundary of region is ellipsoid centered at estimates $\mathbf{b}_1$ in q-dimensional space.
  - Contains all combinations of parameters $\beta_1, \ldots, \beta_q$ simultaneously acceptable at confidence level α.

Joint confidence regions (2)

- Example 2-dimensional case: $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$.
- Joint confidence interval for (β_1, β_2) are all (β_1, β_2) for which

  $$[B_1 - \beta_1, B_2 - \beta_2] \begin{bmatrix} \sum x_{i1}^{*2} & \sum x_{i1}^* x_{i2}^* \\ \sum x_{i1}^* x_{i2}^* & \sum x_{i2}^{*2} \end{bmatrix} \begin{bmatrix} B_1 - \beta_1 \\ B_2 - \beta_2 \end{bmatrix} \leq 2S_E^2 F_{\alpha,2,n-3},$$

  with $x_{ij}^* \equiv x_{ij} - \overline{x}_j$ deviations from means.
  - matrix $\mathbf{V}_{11}^{-1}$ holds mean-deviation sums of squares and products.
- Figure shows correlated regressors (lhs) and uncorrelated regressors (rhs).



Joint confidence regions (3)

- For single coefficient the confidence interval for β_1 is produced:
  $(B_1 - \beta_1)^2 \frac{\sum x_{i2}^{*2}}{\sum x_{i1}^{*2} \sum x_{i2}^{*2} - (\sum x_{i1}^* x_{i2}^*)^2} \leq 2S_E^2 F_{\alpha,1,n-3}$
  equivalent to the ordinary confidence interval based on t-distribution.
- Inner circles in figure on previous slide correspond to individual confidence intervals.
- Correlated regressors may lead to ambiguous inferences, see figure:
  - Individual intervals include 0, so separate hypotheses $H_0 : \beta_1 = 0$ or $H_0 : \beta_2 = 0$ are not rejected.
  - Joint region does not include (0,0), so $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ is rejected.
- Ellipsoid in q-dimensional space reflects correlational structure and dispersion of Xs.
- Skip 9.5-9.7

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 3, lecture 17

Content lecture 17: Fox §10.1

- Vector geometry of linear models
  - vector geometry of simple-regression model
  - vector geometry of least-squares fit: orthogonal projection
  - variables in mean-deviation form (centered variables)
  - degrees of freedom as dimensions of subspaces

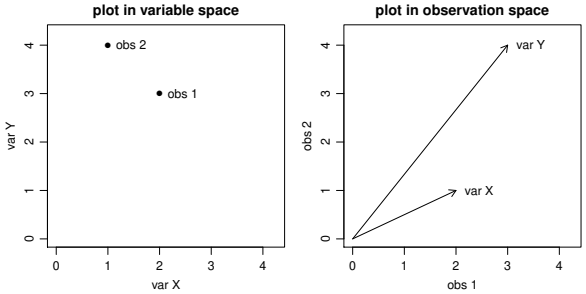
Simple regression

- Simple regression model in vector form: $\mathbf{y} = \alpha\mathbf{1}_n + \beta\mathbf{x} + \boldsymbol{\epsilon}$ assuming $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}_n, \sigma^2_\epsilon \mathbf{I}_n)$.
- Expected values are $E(\mathbf{y}) = \alpha\mathbf{1}_n + \beta\mathbf{x}$.
- Fitted regression equation in vector form: $\mathbf{y} = A\mathbf{1}_n + B\mathbf{x} + \mathbf{e}$.
- Fitted values are $\hat{\mathbf{y}} = A\mathbf{1}_n + B\mathbf{x}$.

Variable space versus observation space

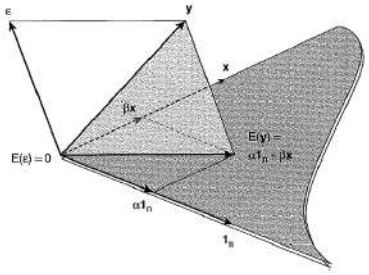
- Usual geometric representation of $\{X, Y\}$ data: scatterplot
  - variables X and Y define axes
  - n observations are represented as n points in two-dimensional space according to their $\{X, Y\}$ coordinates.
  - result is plot in variable space.
- Now switch roles of variables (columns) and observations (rows):
  - observations (rows) define axes
  - 2 variables are represented as arrows in higher (n)-dimensional space
  - result is plot in observation space
  - Usually more than 3 observations, making visualization of full vector space impossible; but often interest in two- and three-dimensional subspaces, which can be visualized.
- Example of 2×2 data matrix:

```
  X Y
1 2 3
2 1 4
```



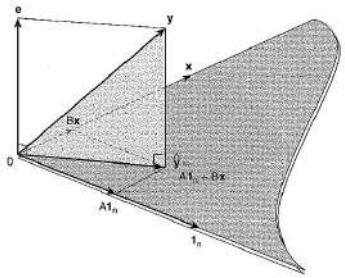
Simple regression plotted in observation space

- Plot represents n-dimensional coordinate space with observations as axes and variables as vectors.
- Simple linear regression is shown in figure.
- Subspace, spanned by vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{1}_n$, has dimension 3.
- $E(\mathbf{y}) = \alpha\mathbf{1}_n + \beta\mathbf{x}$ is linear combination of $\mathbf{x}$ and $\mathbf{1}_n$.
- Error vector $\boldsymbol{\epsilon} = \mathbf{y} - \alpha\mathbf{1}_n - \beta\mathbf{x}$.
- In this sample $\boldsymbol{\epsilon}$ is nonzero, but on average, over many samples, $E(\boldsymbol{\epsilon}) = \mathbf{0}_n$.



Simple regression: orthogonal projection

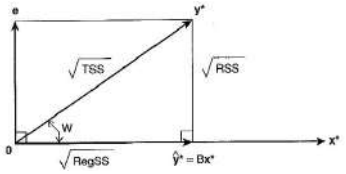
- Figure represents least-squares simple linear regression of Y on X, for same data as earlier figure.
- $\hat{\mathbf{y}}$ is linear combination of $\mathbf{1}_n$ and $\mathbf{x}$, so lies in $\{\mathbf{1}_n, \mathbf{x}\}$-plane.
- Residual $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ has length $\|\mathbf{e}\| = \sqrt{\sum E_i^2}$.
- Geometrical interpretation of least squares: length of $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$ is minimized, by taking as $\hat{\mathbf{y}}$ the orthogonal projection of $\mathbf{y}$ on $\{\mathbf{1}_n, \mathbf{x}\}$-plane.



Simple regression: variables in mean-deviation form

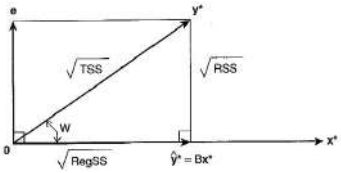
- Use variables in mean-deviation form for 2 reasons:
  - Dimension reduction allows simpler vector visualization.
  - ANOVA shows in vector diagram.
- Scalar form: $Y_i - \overline{Y} = B(x_i - \overline{x}) + E_i$, by subtraction of $\overline{Y} = A + B\overline{x}$ from $Y_i = A + Bx_i + E_i$
- Vector form: $\boxed{\mathbf{y}^* = B\mathbf{x}^* + \mathbf{e}}$, with $\mathbf{y}^* = \{Y_i - \overline{Y}\}$ and $\mathbf{x}^* = \{x_i - \overline{x}\}$.

- Find B by minimization of $\|\mathbf{e}\|$, i.e. orthogonal projection of $\mathbf{y}^*$ on $\mathbf{x}^*$:
  $B = \dfrac{\mathbf{x}^* \cdot \mathbf{y}^*}{\|\mathbf{x}^*\|^2} = \dfrac{\sum(x_i - \overline{x})(Y_i - \overline{Y})}{\sum(x_i - \overline{x})^2}$.



Simple regression: ANOVA and correlation

- See in figure the ANOVA = split of TSS into RSS and $RegSS$:
  - $RSS = \sum E_i^2 = \|\mathbf{e}\|^2$
  - $TSS = \sum(Y_i - \overline{Y})^2 = \|\mathbf{y}^*\|^2$
  - $RegSS = \sum(\hat{Y}_i - \overline{Y})^2 = \|\hat{\mathbf{y}}^*\|^2$.
- Pythagoras theorem: $TSS = RegSS + RSS$!



- Correlation coefficient $r = \sqrt{R^2} = \sqrt{\frac{RegSS}{TSS}} = \frac{\|\hat{\mathbf{y}}^*\|}{\|\mathbf{y}^*\|}$.
  - See that $r = \cos W$: correlation of 2 variables is cosine of angle separating their mean-deviation vectors.
  - Correlation also: $r = \frac{\|\hat{\mathbf{y}}^*\|}{\|\mathbf{y}^*\|} = \frac{\|B\mathbf{x}^*\|}{\|\mathbf{y}^*\|} = |B|\frac{\|\mathbf{x}^*\|}{\|\mathbf{y}^*\|} = \frac{\mathbf{x}^* \cdot \mathbf{y}^*}{\|\mathbf{x}^*\|^2}\frac{\|\mathbf{x}^*\|}{\|\mathbf{y}^*\|} = \frac{\mathbf{x}^* \cdot \mathbf{y}^*}{\|\mathbf{x}^*\|\|\mathbf{y}^*\|}$, the more familiar definition of correlation.

Simple regression: degrees of freedom

- Degrees of freedom associated with sum of squares represent dimension of subspace in which corresponding vector is confined.
- $\|\mathbf{y}\|^2$ (=Uncorrected sum of squares) has n df, because $\mathbf{y}$ can be anywhere in n-dimensional observation space.
- $\|\mathbf{y}^*\|^2 = TSS$ has $n-1$ df: $\mathbf{y}^*$ (mean-deviation form) is confined to $(n-1)$ dimensional subspace; by subtracting mean, one linear restriction is imposed: entries sum to zero, and only $n-1$ values are linearly independent.
- $\|\hat{\mathbf{y}}^*\|^2 = RegSS$ has 1 df, because $\hat{\mathbf{y}}^*$ is multiple of $\mathbf{x}^*$, which in turn is fixed and spans 1-dimensional subspace.
- $\|\mathbf{e}\|^2 = RSS$ has $n-2$ df, because $\mathbf{e}$ lies in orthogonal complement of subspace spanned by $\mathbf{x}$ and $\mathbf{1}_n$ within $\mathbb{R}^n$; least-squares residuals $\mathbf{e}$ satisfy two independent linear restrictions: $\mathbf{e} \cdot \mathbf{1}_n = 0$ and $\mathbf{e} \cdot \mathbf{x} = 0$.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 3, lecture 18

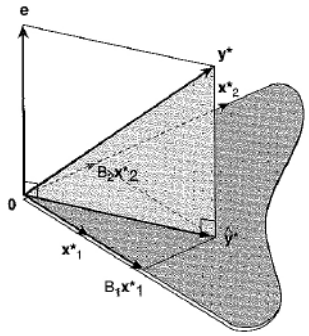
WAGENINGEN UNIVERSITY
WAGENINGEN UR

Content lecture 18: Fox §10.2 - 10.4

- Vector geometry of linear models (cont.)
  - vector geometry of least-squares fit in multiple regression
  - unbiased estimator of error variance
  - analysis of variance models

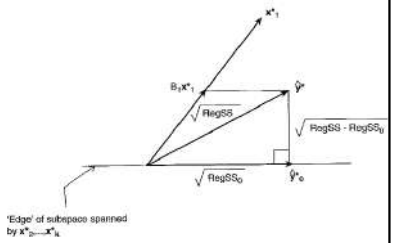
Multiple regression: vector geometry

- Limit to case of two explanatory variables.
- Model in vector form: $\mathbf{y} = A\mathbf{1}_n + B_1\mathbf{x}_1 + B_2\mathbf{x}_2 + \mathbf{e}$.
- Model in mean-deviation form $\boxed{\mathbf{y}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^* + \mathbf{e}}$ by subtraction of $\overline{Y} = A + B_1\overline{x}_1 + B_2\overline{x}_2$.
- Vector fitted values $\hat{\mathbf{y}}^* = B_1\mathbf{x}_1^* + B_2\mathbf{x}_2^*$ lies in $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane, by orthogonal projection of $\mathbf{y}^*$.
- Vector residuals $\mathbf{e}$ is orthogonal to $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane.
- Coefficients B_1 and B_2 uniquely defined as long as $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ are not collinear. (They would span a line in that case.)
- ANOVA for multiple regression appears in plane spanned by $\mathbf{y}^*$ and $\hat{\mathbf{y}}^*$: as before $TSS = RegSS + RSS$.
- Multiple correlation $R = \sqrt{RegSS/TSS} = \cos W$ equals simple correlation between observed and fitted values.



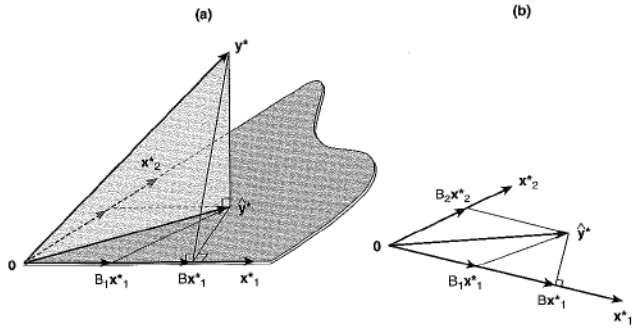
Multiple regression: F-test

- In model with k explanatory variables, test $H_0 : \beta_1 = 0$.
- Fitting full model gives $RegSS = \|\hat{\mathbf{y}}^*\|^2$.
- Fitting reduced model gives $RegSS_0 = \|\hat{\mathbf{y}}_0^*\|^2$.
- $RegSS$ is decomposed into two orthogonal components: $RegSS_0$ and $RegSS - RegSS_0$, giving (part of) numerator of F-statistic for testing $H_0 : \beta_1 = 0$.



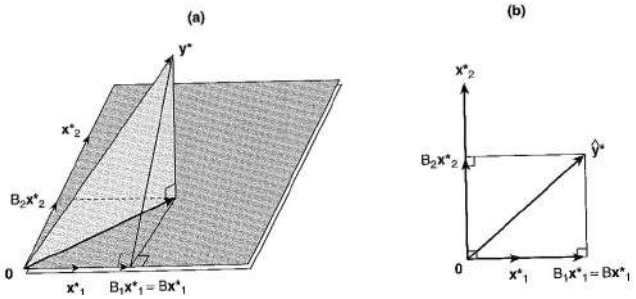
Multiple regression: multiple vs simple regression

- Suppose two positively correlated regressors.
- $\hat{\mathbf{y}}^*$ is orthogonal projection of $\mathbf{y}^*$ on $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane.
- Multiple regression B_1 found by projection of $\hat{\mathbf{y}}^*$, parallel to $\mathbf{x}_2^*$; likewise for B_2.
- Simple regression B of $\mathbf{x}_1^*$ is found by projection of $\mathbf{y}^*$ onto $\mathbf{x}_1^*$ alone.
- We might as well project $\hat{\mathbf{y}}^*$ onto $\mathbf{x}_1^*$, because $\mathbf{x}_1^* \cdot \mathbf{y}^* = \mathbf{x}_1^* \cdot \hat{\mathbf{y}}^*$.
- In this instance, the simple regression coefficient B exceeds the multiple regression coefficient B_1.



Multiple regression: uncorrelated regressors

- In case of uncorrelated regressors $B = B_1$.
- Regression SS is uniquely partitioned into components due to each of two regressors: $RegSS = \hat{\mathbf{y}}^* \cdot \hat{\mathbf{y}}^* = B_1^2 \mathbf{x}_1^* \cdot \mathbf{x}_1^* + B_2^2 \mathbf{x}_2^* \cdot \mathbf{x}_2^*$.
- With correlated regressors, this unique partitioning is not possible, because $RegSS = \hat{\mathbf{y}}^* \cdot \hat{\mathbf{y}}^* = B_1^2 \mathbf{x}_1^* \cdot \mathbf{x}_1^* + B_2^2 \mathbf{x}_2^* \cdot \mathbf{x}_2^* + 2B_1 B_2 \mathbf{x}_1^* \cdot \mathbf{x}_2^*$. The last term can be positive or negative.
- Degrees of freedom correspond to dimension of subspace of observation space, as before:
  - $\hat{\mathbf{y}}^*$ lies in $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane, and must have 2 df.
  - $\mathbf{e}$ is orthogonal to $\{\mathbf{x}_1^*, \mathbf{x}_2^*\}$ plane, and has $(n-1)-2 = n-3$ df.



Estimating error variance (1)

- We want to show that $S_E^2 = \sum E_i^2/(n-(k+1))$ is unbiased estimator of error variance σ_ϵ^2, using the vector geometry of regression.
- Least-squares residuals are not independent (though errors ϵ are: $\epsilon \sim N_n(\mathbf{0}_n, \sigma_\epsilon^2 \mathbf{I}_n)$): $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{Xb} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{Qy}$
  $\mathbf{Q}$ is of order $n \times n$.
  There are $k+1$ linear dependencies among the residuals.
- $V(\mathbf{e}) = V(\mathbf{Qy}) = \mathbf{Q}V(\mathbf{y})\mathbf{Q}' = \mathbf{Q}\sigma_\epsilon^2\mathbf{I}_n\mathbf{Q}' = \sigma_\epsilon^2\mathbf{QQ}' = \sigma_\epsilon^2\mathbf{QQ} = \sigma_\epsilon^2\mathbf{Q}$.
- $\mathbf{e} \sim N_n(\mathbf{0}, \sigma_\epsilon^2\mathbf{Q})$.
- $\mathbf{Q}$ has rank $n-(k+1)$, is singular and is non-diagonal, because:
  - $\mathbf{QX} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X} = \mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X} - \mathbf{X} = \mathbf{0}$; so, all rows of $\mathbf{Q}$ are orthogonal to all $k+1$ columns of $\mathbf{X}$; i.e. the rows of $\mathbf{Q}$ satisfy $k+1$ constraints and $\mathbf{Q}$ is singular; the rank is $n-(k+1)$.
  - $\mathbf{Q}$ cannot be diagonal, because a singular diagonal matrix would have some diagonal entries 0. $\mathbf{Q}$ has generally no diagonal entries zero, as the jth diagonal entry is 1- jth diagonal entry of $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Estimating error variance (2)

- Transform residuals into independent and identically distributed set by selection of orthonormal basis for error subspace:
  - $\mathbf{z} \equiv \mathbf{Ge}$ where $\mathbf{z}$ has $n-(k+1)$ elements and $\mathbf{G}$ has order $(n-(k+1) \times n)$.
  - Choose $\mathbf{G}$ so that
    - $\mathbf{G}$ is orthogonal to $\mathbf{X}$: $\mathbf{GX} = \mathbf{0}$: rows of $\mathbf{G}$ are orthogonal to columns of $\mathbf{X}$.
    - $\mathbf{G}$ is orthonormal: $\mathbf{GG}' = \mathbf{I}_{n-(k+1)}$: the $n-(k+1)$ rows of $\mathbf{G}$ are orthogonal to each other and have length 1.
  - Transformed residuals have properties: $\mathbf{z} = \mathbf{Gy}$, $E(\mathbf{z}) = \mathbf{0}$, $V(\mathbf{z}) = \sigma_\epsilon^2\mathbf{I}_{n-(k+1)}$.
    Check:
    - $\mathbf{z} = \mathbf{Ge} = \mathbf{G}(\mathbf{y} - \mathbf{Xb}) = \mathbf{Gy} - \mathbf{GXb} = \mathbf{Gy} - \mathbf{0b} = \mathbf{Gy}$
    - $E(\mathbf{z}) = E(\mathbf{Ge}) = \mathbf{G}E(\mathbf{e}) = \mathbf{G0} = \mathbf{0}$
    - $V(\mathbf{z}) = V(\mathbf{Ge}) = \mathbf{G}V(\mathbf{e})\mathbf{G}' = \mathbf{G}(\sigma_\epsilon^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'))\mathbf{G}' = \sigma_\epsilon^2(\mathbf{GG}' - \mathbf{GX}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{G}') = \sigma_\epsilon^2(\mathbf{I}_{n-(k+1)} - \mathbf{0}) = \sigma_\epsilon^2\mathbf{I}_{n-(k+1)}$.

Estimating error variance (3)

- Repeating: transformed residuals have properties: $\mathbf{z} = \mathbf{Gy}$, $E(\mathbf{z}) = \mathbf{0}$, $V(\mathbf{z}) = \sigma_\epsilon^2\mathbf{I}_{n-(k+1)}$.
- Entries of $\mathbf{z}$ have zero expectation and common variance σ_ϵ^2 so $E(\mathbf{z}'\mathbf{z}) = \sum_{i=1}^{n-(k+1)} E(Z_i^2) = (n-(k+1))\sigma_\epsilon^2$.
- So unbiased estimator of σ_ϵ^2 is $S_E^2 \equiv \frac{\mathbf{z}'\mathbf{z}}{n-(k+1)}$.
- Because Z_i are independent and normally distributed, $\frac{\mathbf{z}'\mathbf{z}}{\sigma_\epsilon^2} \sim \chi^2_{n-(k+1)}$.
- Explicit calculation of transformed residuals is not needed, because $\mathbf{e}'\mathbf{e} = \mathbf{z}'\mathbf{z}$: they have equal lengths. This result follows from observation that $\mathbf{e}$ and $\mathbf{z}$ are the same vector represented according to alternative bases: $\mathbf{e}$ gives coordinates of residuals relative to natural basis of n-dimensional observation space; $\mathbf{z}$ gives coordinates of the residuals relative to an arbitrary orthonormal basis for $n-(k+1)$ dimensional error subspace. A vector does not change length when basis changes.
- $$\boxed{S_E^2 = \frac{\mathbf{z}'\mathbf{z}}{n-(k+1)} = \frac{\mathbf{e}'\mathbf{e}}{n-(k+1)}}$$

Analysis of variance models

- Overparameterized one-way ANOVA model in effects model notation with m groups: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$.
- Model matrix $\mathbf{X}$ has $m + 1$ columns, which are collinear.
- Columns of $\mathbf{X}$ span subspace of dimension m.
- $\hat{\mathbf{y}}$ is orthogonal projection of $\mathbf{y}$ onto this subspace.
- Most simply done by selecting arbitrary basis for column space of model matrix.
- Dummy coding and deviation coding are two techniques for constructing basis for column space of $\mathbf{X}$.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 4, lecture 19

Content lecture 19: Faraway 16.1.5 Multiple comparisons

- Multiple comparisons
- Fisher's LSD
- Bonferroni method
- Tukey method
- Scheffé method

Multiple comparisons

- Starting point is one-way ANOVA: one factor, one quantitative response Y. Model: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$
- We know how to make ANOVA table, and test $H_0 : \alpha_1 = \cdots = \alpha_k = 0$ by F-test.
- Suppose from F-test we conclude that not all group means are equal. What next?
- One way to go is to make pairwise comparisons between groups, but other contrasts may be of interest as well. See part on General Linear Hypothesis.
- If many pairwise comparisons are made with individual t-tests, each at significance level α, probability to reject at least one H_0 falsely (family wise error rare) is (much) larger than α.
- Making arrangements so that the family wise error rate remains acceptable, is topic of study in multiple comparisons.

After the F-test

- Discriminate between comparisons were decided upon before or after examination of data.
- Suppose comparisons were decided upon beforehand. Faraway distinguishes 3 cases:
  - Single comparison: use standard t-test or t-based c.i. (confidence interval)
  - Few comparisons: use Bonferroni adjustment for t. With m comparisons, use α/m to get critical value from t-distribution.
  - Many comparisons: Bonferroni is increasingly conservative. Better to use Tukey, Scheffé or related method.
- If comparisons were decided upon after examination of the data, it may be best to adjust test or ci to allow for possibility that all comparisons were made.
- Two cases:
  - Only pairwise comparisons. Use Tukey method.
  - All contrasts, i.e. linear combinations. Use Scheffé method.

Pairwise comparison using t-distribution

- Suppose we want to estimate $\alpha_i - \alpha_j$ or test $H_0 : \alpha_i - \alpha_j = 0$.
- C.i. $\hat{\alpha}_i - \hat{\alpha}_j \pm t_{\alpha/2,n-I}\hat{\sigma}\sqrt{\frac{1}{J_i} + \frac{1}{J_j}}$
  (Notice that notation in Faraway deviates slightly from Fox.)
- LSD=Least Significant Difference $= t_{\alpha/2,n-I}\hat{\sigma}\sqrt{\frac{1}{J_i} + \frac{1}{J_j}}$:
  if $|\hat{\alpha}_i - \hat{\alpha}_j| > LSD$ ($\Leftrightarrow \{0 \notin \text{ci}\}$), then the difference is called significant.
- Suppose we test at $\alpha = 0.05$.
  - For 1 comparison, the probability P to reject H_0 if it is true, is 0.05.
  - For 2 independent comparisons, the probability P to reject at least one H_0, if all are true, is $1 - P(\text{none rejected}) = 1 - (1-\alpha)^2 = 1 - (1-0.05)^2 = 0.0975$.
  - For 3, $P = 1 - P(\text{none rejected}) = 1 - (1-\alpha)^3 = 1 - (1-0.05)^3 = 0.143$.
  - For 4, $P = 1 - P(\text{none rejected}) = 1 - (1-\alpha)^4 = 1 - (1-0.05)^4 = 0.185$.
  - For 5, $P = 1 - P(\text{none rejected}) = 1 - (1-\alpha)^5 = 1 - (1-0.05)^5 = 0.226$.
  - For 6, $P = 1 - P(\text{none rejected}) = 1 - (1-\alpha)^6 = 1 - (1-0.05)^6 = 0.265$.
- Hence, probability to falsely reject at least one H_0 can be much larger than α.
- Fisher's protected LSD: only make pairwise comparisons using LSD after a significant F-test.

Pairwise comparison using Studentized Range Distribution

- Idea: critical value must be made larger so that the familywise error rate, for all comparisons together, remains at 5% level.
- Leads to Tukey's Studentized Range Distribution.
- Let $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ and independent. Define range R as $R = max_i\{X_i\} - min_i\{X_i\}$. $R/\hat{\sigma}$ has the so-called studentized range distribution $q_{n,\nu}$ with ν number of degrees of freedom for estimating σ.
- Tukey's Honest Significant Difference(HSD) method gives c.i. $\hat{\alpha}_i - \hat{\alpha}_j \pm \frac{q_{I,n-I}}{\sqrt{2}}\hat{\sigma}\sqrt{\frac{1}{J_i} + \frac{1}{J_j}}$.
- Assumes equal group sizes; for unequal group sizes Tukey's HSD may be too conservative.

Example coagulation

- Example on coagulation times in 24 animals, randomly assigned to diets A, B, C and D; Faraway focuses on difference group B - C

```
> coagulation$diet <- relevel(coagulation$diet,ref="B")  # Make B reference group
> table(coagulation$diet) # number of animals per diet

B A C D
6 4 6 8

> tapply(coagulation$coag, coagulation$diet, mean) # mean coagulation per diet

 B  A  C  D
66 61 68 61
```

Different numbers of animals per group.

```
> g <- lm(coag ~ diet, data=coagulation)
> coef(summary(g))

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)       66     0.9661  68.316 3.532e-25
dietA             -5     1.5275  -3.273 3.803e-03
dietC              2     1.3663   1.464 1.588e-01
dietD             -5     1.2780  -3.912 8.636e-04

> summary(g)$sigma # residual standard deviation

[1] 2.366
```

- Do we have a single LSD for all pairwise comparisons?

Example coagulation continued

- Estimate of mean difference B-C is 2 with sed=1.366; residual $df = 24 - 4 = 20$.

```
> diffBC<-2; sedBC <- 1.366;
> (crit.t <- qt(0.975,20))
[1] 2.086
> cl.lo.t <- diffBC - crit.t * sedBC; cl.up.t <- diffBC + crit.t * sedBC  # CI based on t-distribution
> cat("CI (t-distr):", cl.lo.t, cl.up.t)
CI (t-distr): -0.8494 4.849
```

- Assume 2 pre-planned comparisons. Use Bonferroni:

```
> (crit.t.Bonf <- qt(1-0.025/2,20))
[1] 2.423
> cl.lo.B <- diffBC - crit.t.Bonf * sedBC; cl.up.B <- diffBC + crit.t.Bonf * sedBC  # CI based on t-distributio
> cat("CI (Bonferroni):", cl.lo.B, cl.up.B)
CI (Bonferroni): -1.31 5.31
```

- Use Tukey's method, applying Studentized Range Distribution

```
> (crit.SRD <- qtukey(0.95,4,20) / sqrt(2))
[1] 2.799
> cl.lo.T <- diffBC - crit.SRD * sedBC; cl.up.T <- diffBC + crit.SRD * sedBC
> cat("CI (Tukey):", cl.lo.T, cl.up.T)
CI (Tukey): -1.823 5.823
```


Example Tukey's Honest Significant Difference method

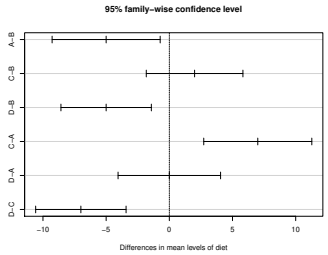
Using facilities for pairwise comparisons from the `emmeans` package, which processes objects of class `lm`:

```
> emmeans(g, pairwise ~ diet, adjust="tukey")  # this is also the default
$emmeans
 diet emmean    SE df lower.CL upper.CL
 B        66 0.966 20     63.4     68.6
 A        61 1.183 20     57.8     64.2
 C        68 0.966 20     65.4     70.6
 D        61 0.837 20     58.7     63.3

Confidence level used: 0.95
Conf-level adjustment: sidak method for 4 estimates

$contrasts
 contrast estimate   SE df t.ratio p.value
 B - A           5 1.53 20  3.273  0.0183
 B - C          -2 1.37 20 -1.464  0.4766
 B - D           5 1.28 20  3.912  0.0044
 A - C          -7 1.53 20 -4.583  0.0010
 A - D           0 1.45 20  0.000  1.0000
 C - D           7 1.28 20  5.477  0.0001

P value adjustment: tukey method for comparing a family of 4 estimates
```


Example Tukey's Honest Significant Difference method - 2

Using ANOVA procedure `aov()`, which is wrapper to `lm()`

```
> TukeyHSD(aov(coag ~ diet, data=coagulation))   # works on aov objects
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coagulation)

$diet
    diff     lwr     upr  p adj
A-B   -5 -9.275 -0.7246 0.0183
C-B    2 -1.824  5.8241 0.4766
D-B   -5 -8.577 -1.4229 0.0044
C-A    7  2.725 11.2754 0.0010
D-A    0 -4.056  4.0560 1.0000
D-C   -7 -10.577 -3.4229 0.0001
> plot(TukeyHSD(aov(coag ~ diet, data=coagulation)))
```



Scheffé's theorem for multiple comparisons

- Now look at all possible estimable linear combinations of parameters $\psi = c^T\beta$:
  - Linear combination of parameters $\psi = c^T\beta$ is estimable, if $a^T\mathbf{y}$ exists such that $E(a^T\mathbf{y}) = c^T\beta$.
  - Contrast is a special case of linear combination: $\sum c_i = 0$.
  - E.g. contrasts are estimable, but in overparameterized model α_i is not.
- Estimator $\hat{\psi} = a^T\mathbf{y}$ has $var(\hat{\psi}) = \sigma_\epsilon^2 a^T a$, estimated by $\hat{\sigma}_\epsilon^2 a^T a$.
- With q dimension of space of possible c (typically $I - 1$ with I the number groups) and rank($\mathbf{X}$)=r, we have Scheffé's theorem: $100(1-\alpha)\%$ simultaneous c.i. for all estimable ψ is

$$\hat{\psi} \pm \sqrt{qF_{\alpha;q,n-r}}\sqrt{\widehat{var}(\hat{\psi})}$$

Example Scheffé's method for multiple comparisons

- Example one-way ANOVA. Let $\psi = \sum_i c_i \alpha_i$, a contrast.
  - Verify that $\widehat{var}(\hat{\psi}) = \hat{\sigma}^2_\epsilon \sum_{i=1}^I \frac{c_i^2}{J_i}$
  - Simultaneous ci for ψ is $\sum_i c_i \hat{\alpha}_i \pm \sqrt{(I-1)F_{\alpha;I-1,n-I}}\hat{\sigma}_\epsilon \sqrt{\sum_{i=1}^I \frac{c_i^2}{J_i}}$
  - Suppose we look at the data and spot a "promising" contrast: $(B+C)/2 - (A+D)/2$, so $c_1 = -1/2$, $c_2 = c_3 = 1/2$, $c_4 = -1/2$. Use Scheffé's method allowing for all possible linear combinations.
  - In previous output we saw $\hat{\alpha}_A = -5$,$\hat{\alpha}_C = 2$, $\hat{\alpha}_D = -5$ and $\hat{\alpha}_B = 0$ (reference).

```
> (halfwidth <- sqrt(3*qf(0.95,3,20))*2.366*sqrt((1/4)*(1/4+1/6+1/6+1/8)))
[1] 3.036
> (psihat <- (2+0)/2 - (-5+-5)/2)
[1] 6
> c(psihat-halfwidth, psihat+halfwidth)
[1] 2.964 9.036
```


Pairwise comparisons with Tukey, Bonferroni, and Scheffé

- Pairwise comparisons with function `emmeans` with specific correction for multiple comparisons.

```
> emm.g <- emmeans(g, ~ diet)
> pairs(emm.g, adjust="none")
 contrast estimate   SE df t.ratio p.value
 B - A           5 1.53 20  3.273  0.0038
 B - C          -2 1.37 20 -1.464  0.1588
 B - D           5 1.28 20  3.912  0.0009
 A - C          -7 1.53 20 -4.583  0.0002
 A - D           0 1.45 20  0.000  1.0000
 C - D           7 1.28 20  5.477  <.0001
> pairs(emm.g, adjust="bonferroni")
 contrast estimate   SE df t.ratio p.value
 B - A           5 1.53 20  3.273  0.0228
 B - C          -2 1.37 20 -1.464  0.9527
 B - D           5 1.28 20  3.912  0.0052
 A - C          -7 1.53 20 -4.583  0.0011
 A - D           0 1.45 20  0.000  1.0000
 C - D           7 1.28 20  5.477  0.0001

P value adjustment: bonferroni method for 6 tests
> pairs(emm.g, adjust="scheffe")
 contrast estimate   SE df t.ratio p.value
 B - A           5 1.53 20  3.273  0.0323
 B - C          -2 1.37 20 -1.464  0.5549
 B - D           5 1.28 20  3.912  0.0088
 A - C          -7 1.53 20 -4.583  0.0021
 A - D           0 1.45 20  0.000  1.0000
 C - D           7 1.28 20  5.477  0.0003

P value adjustment: scheffe method with rank 3
```


Letter display of pairwise difference

- Hypothesis test results for pairwise differences of means can be shown using compact letter display: groups connected with same letter (number) are not significantly different.

```
> CLD(emm.g, adjust="tukey")
 diet emmean    SE df lower.CL upper.CL .group
 A        61 1.183 20     57.8     64.2  1
 D        61 0.837 20     58.7     63.3  1
 B        66 0.966 20     63.4     68.6   2
 C        68 0.966 20     65.4     70.6   2

Confidence level used: 0.95
Conf-level adjustment: sidak method for 4 estimates
P value adjustment: tukey method for comparing a family of 4 estimates
significance level used: alpha = 0.05
```


Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 4, lecture 20

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Content lecture 20: Fox §8.2

- Two-way ANOVA
  - patterns of means and profile plot
  - interaction and main effects
  - means model and effects model; overparameterization
  - parameter interpretation with corner stone restriction and sum-to-zero restriction

Recall One-Way ANOVA model (lecture 9)

- Quantitative response Y to be explained from a factor (group variable).
- Typically one-way ANOVA model is written (using ANOVA notation) as:

$$Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$$

  ϵ_{ij} has usual linear model assumptions: independent, normally distributed, equal variances, and zero expectation.
  - Y_{ij} uses two indices: j is index for group, i is index for observation within group.
  - m groups, so $j = 1, \cdots, m$; we have n_j replications per group, so $i = 1, \cdots, n_j$.
  - One-way ANOVA model in ANOVA notation can be written with or without intercept:
    - With intercept: $Y_{ij} = \mu + \alpha_j + \epsilon_{ij}$ is called the "effects-model", because parameters α_j are effects of factor level j
    - Without intercept: $Y_{ij} = \mu_j + \epsilon_{ij}$ is called the "means-model", because parameters μ_j are means (expected values) for factor level j.
- Regression notation is another way to write down the model. E.g. for factor with 3 levels:

$$Y_i = \alpha + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \epsilon_i$$

  Regressors (here: dummies D_{i1} and D_{i2} for levels 1 and 2 of the factor) are explicitly mentioned.

Two-way analysis of variance: patterns of means

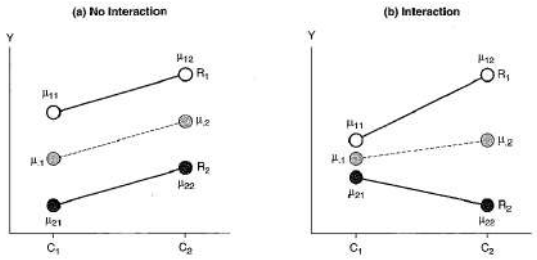
- Fox §8.2: from one-way ANOVA to two-way ANOVA by including second factor
- Inclusion of second factor allows study of interaction.
- Focus on Two-way ANOVA with interaction
- Two factors: R (with r levels) and C (with c levels).
- Suppose we have access to population means μ_{jk} of responses Y_{ijk}. Notation:

	C_1	C_2	$\cdots$	C_c	
R_1	μ_{11}	μ_{12}	$\cdots$	μ_{1c}	$\mu_{1.}$
R_2	μ_{21}	μ_{22}	$\cdots$	μ_{2c}	$\mu_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
R_r	μ_{r1}	μ_{r2}	$\cdots$	μ_{rc}	$\mu_{r.}$
	$\mu_{.1}$	$\mu_{.2}$	$\cdots$	$\mu_{.c}$	$\mu_{..}$

- Within cell (j, k) of the design, response variable Y has population mean μ_{jk}.
- Margins contain marginal means, in rows and in columns.
- Grand mean $\mu_{..} \equiv \frac{\sum_j \sum_k \mu_{jk}}{r \times c} = \frac{\sum_j \mu_{j.}}{r} = \frac{\sum_k \mu_{.k}}{c}$.

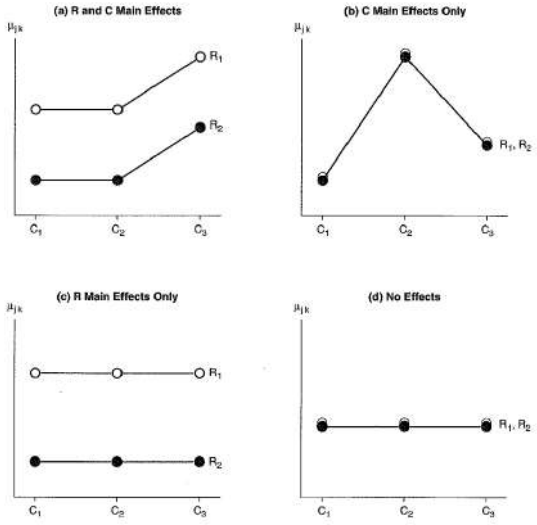
Two-way analysis of variance: patterns of means

- If R and C do not interact, then partial relationship between each factor and Y does not depend on category at which other factor is "held constant":
  $\mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{j'k'} = \mu_{j.} - \mu_{j'.} \Leftrightarrow$ "Profiles are parallel"
  $\Leftrightarrow$ row differences constant across columns
  $\Leftrightarrow$ column differences constant across rows
  $\Leftrightarrow$ Factors R and C have additive effects.
- Interaction between two factors means that row difference changes across columns and vice versa $\Leftrightarrow$ "Profiles are not parallel".
- When interactions are absent, partial effect of one factor (main effect) is given by difference in population marginal means.
- Below "profile plots" are shown:



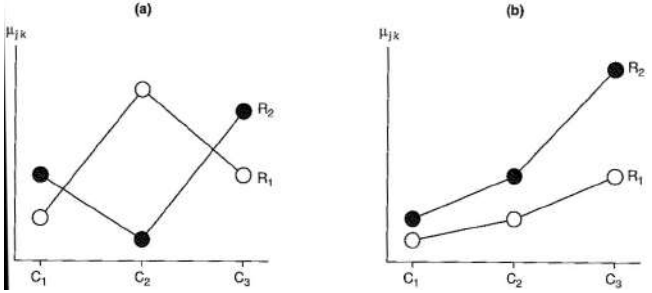
Profile plots without interaction

- Examples of patterns of relationship in two-way classification, all showing no interaction.



Profile plots with interaction

- Lhs plot shows example with dramatic interaction: lines cross, population means for 3 categories of C are ordered differently within R_1 and R_2.
- Rhs plot shows less dramatic interaction. This type of interaction can sometimes be transformed away, e.g. by taking logs.



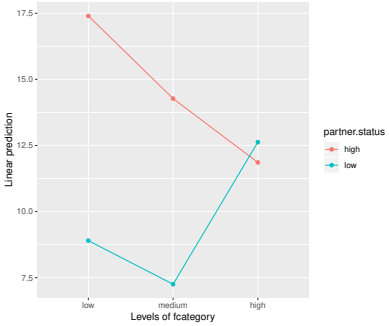
Checking interactions and main effects

- Sample means will always show (some) interaction, even if interactions are absent in population, due to sampling error.
- We have to determine whether departures from parallelism observed in sample are sufficiently large to be statistically significant, or whether they could be a product of chance alone.
- In large samples, we want to determine whether "statistically significant" interactions are of sufficient magnitude to be of substantive interest. We may well decide to ignore interactions that are statistically significant, but trivially small.
- If interactions are present and non-negligible, then typically do not interpret main effects of factors: to conclude that two variables interact is to deny that they have separate effects. This refers to principle of marginality, earlier discussed: main effects of R and C are marginal to RC interaction.

Example interaction (=profile) plot

Moore dataset: 45 subjects in social-psychological experiment, faced with manipulated disagreement from a partner of either low or high status; Y=`conformity` is number of conforming responses in 40 trials; two factors: `partner.status` with levels low and high, `fcategory` categorized F-scale score on authoritarianism with levels low, medium and high. Design is unbalanced. Function `emmip` makes interaction plot based upon predicted means from a `lm` object:

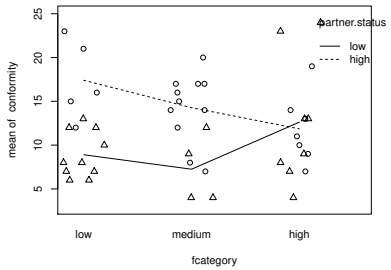
```
> Moore$fcategory <- factor(Moore$fcategory, levels=c("low","medium","high"))  # reorder levels
> table(Moore$partner.status, Moore$fcategory) # notice unequal numbers of observations over 6 groups
       low medium high
  high   5     11    7
  low   10      4    8
> lmo <- lm(conformity ~ partner.status + fcategory + partner.status:fcategory, data=Moore)
> emmip(lmo, partner.status ~ fcategory, pch=c(15,16), cex=1.5)
```



Example interaction plot

Simpler way of plotting means (not based upon `lm` object):

```
> with(Moore, interaction.plot(fcategory,partner.status,conformity,ylim=c(3,25)))
> with(Moore, points(jitter(as.numeric(fcategory)),conformity,pch=c(1,2)[partner.status]))
```



Two-way ANOVA model with interaction: means and effects model

- Two ways of writing two-way ANOVA model (in ANOVA notation) with interaction:
  - Means model: $\boxed{Y_{ijk} = \mu_{jk} + \epsilon_{ijk}}$
  - Effects model: $\boxed{Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}}$
  - both have errors ϵ_{ijk} with all usual linear model assumptions.
  - cell means: $\mu_{jk} \equiv EY_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$.
- Means model:
  - cell means μ_{jk} are expected values EY_{ijk}
  - as many parameters as groups ($r \times c$), no overparameterization
  - interaction is implicit: no explicit interaction parameters: interaction contrast $(\mu_{jk} - \mu_{j'k}) - (\mu_{jk'} - \mu_{j'k'}) \neq 0$.
- Effects model:
  - μ general mean, α_j main-effect parameters for row, β_k main-effect parameters for column, γ_{jk} interaction parameters
  - more parameters $(1 + r + c + r \times c)$ than groups $(r \times c)$, overparameterization
  - interaction is explicit: parameters γ_{jk}
  - solve overparameterization by $1 + r + c$ restrictions on parameters.

Two-way ANOVA model with interaction - testing scheme

- Interpretation of two-way ANOVA depends on presence/absence of interaction.
- Start testing for interaction:
  - Means model:
    $H_0 : \mu_{jk} - \mu_{j'k} = \mu_{jk'} - \mu_{jk'} \Leftrightarrow$ row effects same within all levels of column factor.
  - Effects model: $H_0 : \gamma_{jk} = 0$ (all j, k) $\Leftrightarrow$ all interaction parameters zero
- After interaction, look at main effects.
  - Convenient to express main effect hypotheses in marginal means, so using derived parameters from means model:
    for row classification: $H_0 : \mu_{1.} = \mu_{2.} = \cdots = \mu_{r.}$
    for column classification: $H_0 : \mu_{.1} = \mu_{.2} = \cdots = \mu_{.c}$
  - Formulating main effect hypotheses with effect model parameters more complex:
    - row classification: $H_0 : \alpha_1 + \gamma_{1.} = \alpha_2 + \gamma_{2.} = \cdots = \alpha_r + \gamma_{r.}$ (see next slide)
    - column classification: $H_0 : \beta_1 + \gamma_{.1} = \beta_2 + \gamma_{.2} = \cdots = \beta_c + \gamma_{.c}$ (see next slide)
    - Notice that these "main effect null hypotheses" not only contain main effect parameters, but also (averages of) interaction parameters!
- Such formulated main-effect null hypotheses are testable, with or without interactions, though usually only sensible when interactions are absent.
- Main effect H_0's simplify if specific restrictions are placed on parameters, e.g. with sum-to-zero restriction on interaction parameters we have $\gamma_{j.} = \gamma_{.k} = 0$!

Overparameterized two-way ANOVA model: corner stone restriction

- Default parameterization in R's `lm()` function: first level of factor is reference group
  - $\alpha_1 = 0$
  - $\beta_1 = 0$
  - $\gamma_{1k} = 0$ for all $k = 1, \ldots, c$
    $\gamma_{j1} = 0$ for all $j = 1, \ldots, r$;
    these 2 sets contain one redundant restriction.
- Corner stone restriction, first level reference, interpretation of parameters in terms of cell means:
  - $\mu = \mu_{11}$
  - $\alpha_j = \mu_{j1} - \mu_{11}$
  - $\beta_k = \mu_{1k} - \mu_{11}$
  - $\gamma_{jk} = (\mu_{jk} - \mu_{1k}) - (\mu_{j1} - \mu_{11})$
- Main-effects null hypothesis, e.g. for row main effects:
  $H_0 : \mu_{1.} = \mu_{2.} = \cdots = \mu_{r.} (= \mu_{..})$
  $\Leftrightarrow H_0 : \mu + \alpha_1 + \beta_. + \gamma_{1.} = \mu + \alpha_2 + \beta_. + \gamma_{2.} = \cdots = \mu + \alpha_r + \beta_. + \gamma_{r.}$
  $\Leftrightarrow H_0 : 0 = \alpha_2 + \gamma_{2.} = \cdots = \alpha_r + \gamma_{r.}$
  because μ and $\beta_.$ vanish, and $\alpha_1 = 0$ and $\gamma_{1.} = 0$.
  This "main effect" hypothesis still contains interaction parameters! Not nice...

Overparameterized two-way ANOVA model: sigma constraint

- Sigma constraints:
  - $\sum_{j=1}^{r} \alpha_j = 0$
  - $\sum_{k=1}^{c} \beta_k = 0$
  - $\sum_{j=1}^{r} \gamma_{jk} = 0$ for all $k = 1, \ldots, c$
    $\sum_{k=1}^{c} \gamma_{jk} = 0$ for all $j = 1, \ldots, r$;
    these 2 sets contain one redundant restriction.
- Sigma constraint, interpretation of parameters in terms of population means:
  - $\mu = \mu_{..}$
  - $\alpha_j = \mu_{j.} - \mu_{..}$
  - $\beta_k = \mu_{.k} - \mu_{..}$
  - $\gamma_{jk} = \mu_{jk} - (\mu + \alpha_j + \beta_k) = \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..}$
- Main-effects null hypothesis, e.g. for row main effects:
  $H_0 : \mu_{1.} = \mu_{2.} = \cdots = \mu_{r.} (= \mu_{..})$
  $\Leftrightarrow H_0 : \mu_{1.} - \mu_{..} = \mu_{2.} - \mu_{..} = \cdots = \mu_{r.} - \mu_{..} (= 0)$
  $\Leftrightarrow H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$
  Now "main effect" hypothesis does not contain interaction parameters anymore! Nice...

Fitting two-way ANOVA model with interaction to data

- Least-squares estimator of μ_{jk}: $\overline{Y}_{jk} = \frac{\sum_{i=1}^{n_{jk}} Y_{ijk}}{n_{jk}}$.
- Least-squares estimators of sum-to-zero constrained model parameters follow easily:
  - $M \equiv \hat{\mu} = \overline{Y}_{..} = \sum\sum \overline{Y}_{jk}/(r \times c)$
  - $A_j \equiv \hat{\alpha}_j = \overline{Y}_{j.} - \overline{Y}_{..}$
  - $B_k \equiv \hat{\beta}_k = \overline{Y}_{.k} - \overline{Y}_{..}$
  - $C_{jk} \equiv \hat{\gamma}_{jk} = \overline{Y}_{jk} - \overline{Y}_{j.} - \overline{Y}_{.k} + \overline{Y}_{..}$
- Residuals are $E_{ijk} = Y_{ijk} - (M + A_j + B_k + C_{jk}) = Y_{ijk} - \overline{Y}_{jk}$
- For testing hypotheses incremental sums of squares are needed, and in general no easy formulae exist, unless all cell frequencies are equal.
- Deviation-coded regressors may be used, and incremental sums of squares calculated.
- Example 2×3 classification:
  - Row parameters (1 remains): replace α_2 by $-\alpha_1$, because of model restriction $\alpha_1 + \alpha_2 = 0$
  - Column parameters (2 remain): replace β_3 by $-\beta_1 - \beta_2$, because of model restriction $\beta_1 + \beta_2 + \beta_3 = 0$
  - Interaction parameters (2 remain):
    replace γ_{13} by $-\gamma_{11} - \gamma_{12}$, because of model restriction $\gamma_{11} + \gamma_{12} + \gamma_{13} = 0$
    replace γ_{21} by $-\gamma_{11}$, because of model restriction $\gamma_{11} + \gamma_{21} = 0$
    replace γ_{22} by $-\gamma_{12}$, because of model restriction $\gamma_{12} + \gamma_{22} = 0$
    replace γ_{23} by $-\gamma_{13} = \gamma_{11} + \gamma_{12}$, because of model restriction $\gamma_{13} + \gamma_{23} = 0$

Fitting two-way ANOVA model with interaction to data (2)

Cell		(α_1)	(β_1)	(β_2)	(γ_{11})	(γ_{12})
Row	Column	R_1	C_1	C_2	R_1C_1	R_1C_2
1	1	1	1	0	1	0
1	2	1	0	1	0	1
1	3	1	-1	-1	-1	-1
2	1	-1	1	0	-1	0
2	2	-1	0	1	0	-1
2	3	-1	-1	-1	1	1

- For example: $\mu_{13} = \mu + \alpha_1 - \beta_1 - \beta_2 - \gamma_{11} - \gamma_{12}$
- "Mechanical" rules:
  - $r - 1$ regressors for row main effects; jth regressor is 1 if observation belongs to row j, -1 if to (last) row r, 0 otherwise.
  - $c - 1$ regressors for column main effects; kth regressor is 1 if observation belongs to column k, -1 if to (last) column c, 0 otherwise.
  - $(r - 1)(c - 1)$ regressors for interactions; form product of corresponding main effect regressors for rows and column.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 4, lecture 21

Content lecture 21: Fox 8.2-8.3; Far 16.3-6 (Exp designs)

- Two-way ANOVA (cont)
  - F-tests for interaction and main effects, and principle of marginality
  - type I, II, III sums of squares
  - equal cell frequencies, balanced design, orthogonality
  - analysis of unbalanced design
  - two-way ANOVA without interaction: additive model
- Higher-way ANOVA
- Special experimental designs
  - complete randomized design
  - randomized complete block design
  - Latin square
  - balanced incomplete block design
  - fractional factorial design

Testing hypotheses in two-way ANOVA with interaction

- Tests for interactions and main effects constructed by incremental-sum-of-squares approach.
- As always, incremental sums of squares are given by differences between residual sums of squares of a smaller nested model versus a larger model, or, equivalently, between regression sums of squares of a larger and a smaller model.
- Notation: $SS(\alpha, \beta, \gamma)$ is regression sum of squares for full model with main effects and interaction.
- Notation: $SS(\gamma|\alpha, \beta) = SS(\alpha, \beta, \gamma) - SS(\alpha, \beta)$
  "Sum of squares for interaction after main effects". Others likewise.
- $RSS = \sum\sum\sum E_{ijk}^2 = \sum\sum\sum(Y_{ijk} - \overline{Y}_{jk})^2 = TSS - SS(\alpha, \beta, \gamma)$
- Testing interactions: $H_0 : \gamma_{jk} = 0$ by $SS(\gamma|\alpha, \beta)$
- Testing row main-effects: $H_0 : \alpha_j = 0$ can be done in two ways:
  1. by $SS(\alpha|\beta)$: assumes interactions absent, conforms to principle of marginality; SAS Type II SS.
  2. by $SS(\alpha|\beta, \gamma)$: interactions present, does not conform to marginality principle; not very interesting if interaction is present; SAS Type III SS.
- Likewise for column main-effects.
- Famous statisticians have different opinions about this, although the majority (certainly within the R world) seems to prefer $SS(\alpha|\beta)$.

Example two-way ANOVA with interaction

```
> twoway <- lm(conformity ~ partner.status + fcategory + partner.status:fcategory,
+ contrasts=list(fcategory=contr.sum, partner.status=contr.sum), data=Moore)
> anova(twoway)
Analysis of Variance Table

Response: conformity
                     Df Sum Sq Mean Sq F value   Pr(>F)
partner.status        1 204.33 204.332  9.7448 0.003381
fcategory             2  11.61   5.807  0.2770 0.759564
partner.status:fcategory  2 175.49  87.744  4.1846 0.022572
Residuals            39 817.76  20.968
> Anova(twoway, type="II")   # Anova function is in car package
Anova Table (Type II tests)

Response: conformity
                     Sum Sq Df F value   Pr(>F)
partner.status       212.21  1 10.1207 0.002874
fcategory             11.61  2  0.2770 0.759564
partner.status:fcategory 175.49  2  4.1846 0.022572
Residuals            817.76 39
> Anova(twoway, type="III")  # for type III SS sum-to-zero restriction (contr.sum) is essential!!
Anova Table (Type III tests)

Response: conformity
                     Sum Sq Df  F value    Pr(>F)
(Intercept)          5752.8  1 274.3592 < 2.2e-16
partner.status        239.6  1  11.4250  0.001657
fcategory              36.0  2   0.8589  0.431492
partner.status:fcategory  175.5  2   4.1846  0.022572
Residuals            817.8 39
```


Equal cell frequencies

- With equal cell frequencies, deviation regressors for different sets of effects are uncorrelated.
- Equal-cell-frequency ANOVA design is both balanced and orthogonal.
- Balanced design is design in which all treatment pairs can be estimated with the same precision.
- Term orthogonal refers to properties of subspaces relating to sets of regressors. There are ANOVA designs with unequal cell frequencies, which are orthogonal, but not balanced.
- Life is easier with equal cell frequencies:
  $SS(\alpha|\beta, \gamma) = SS(\alpha|\beta) = SS(\alpha)$
  $SS(\beta|\alpha, \gamma) = SS(\beta|\alpha) = SS(\beta)$
  $SS(\gamma|\alpha, \beta) = SS(\gamma)$
  In this case simple formulae for sums of squares exist.

Cautionary remarks (1)

- Quite some confusion about analysis of unbalanced data.
- Dates back to Frank Yates, who proposed methods both for $SS(\alpha|\beta)$ and $SS(\alpha|\beta,\gamma)$.
- Part of confusion comes from type of restrictions (or other methods) used to solve overparameterization.

Cautionary remarks (2)

- Compare dummy coding and deviation coding:
  - Let $SS^*(.)$ represent regression sum of squares for dummy-coded model and $SS(.)$ for deviation-coded model.
  - Regression sums of squares do not depend on parameterization for:
    interaction model: $SS(\alpha,\beta,\gamma) = SS^*(\alpha,\beta,\gamma)$
    additive model: $SS(\alpha,\beta) = SS^*(\alpha,\beta)$
    single main effect models: $SS(\alpha) = SS^*(\alpha)$ and $SS(\beta) = SS^*(\beta)$.
  - Hence, incremental sums of squares depending on them are identical, like we do by respecting marginality: test for interaction: $SS(\gamma|\alpha,\beta) = SS^*(\gamma|\alpha,\beta)$
    tests for main effect ("type II"): $SS(\alpha|\beta) = SS^*(\alpha|\beta)$.
  - In general, however, $SS(\alpha,\gamma)$ and $SS^*(\alpha,\gamma)$ are not identical, and consequently main effect "type III" SS are not identical: $SS(\beta|\alpha,\gamma) = SS(\alpha,\beta,\gamma) - SS(\alpha,\gamma)$ not equal to $SS^*(\beta|\alpha,\gamma) = SS^*(\alpha,\beta,\gamma) - SS^*(\alpha,\gamma)$.
- General lesson: tests that violate principle of marginality do depend on restriction used.
- Deviation coding gives reasonable results for $SS(\alpha|\beta,\gamma)$, but dummy coding does not.

Three-way ANOVA (1)

- Label factors as A, B, and C, with a, b, and c levels.
- Model $Y_{ijkm} = \mu_{jkm} + \epsilon_{ijkm} =$
  $\mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(m)} + \alpha_{AB(jk)} + \alpha_{AC(jm)} + \alpha_{BC(km)} + \alpha_{ABC(jkm)} + \epsilon_{ijkm}$.
- Use e.g. sigma constraints on all parameters to solve overparameterization.
- Three-way ANOVA model includes parameters for main effects, two-way interactions, and three-way interactions.
- Keep principle of marginality in mind: do not interpret a lower-order effect (e.g. interaction AB) if a non-null higher-order relative (ABC) is in the model: if joint effects of A and B are different in different categories of C, generally not sensible to speak of unconditional AB effects without reference to specific category of C.

Three-way ANOVA (2)

- Deviation regressors defined in usual way, with deviation regressors for interactions by taking products of main-effect regressors.
- In that case, constrained parameters in terms of population means are:

$$\mu = \mu_{...}$$
$$\alpha_{A(j)} = \mu_{j..} - \mu_{...}$$
$$\alpha_{AB(jk)} = \mu_{jk.} - \mu_{j..} - \mu_{.k.} + \mu_{...}$$
$$\alpha_{ABC(jkm)} = \mu_{jkm} - \mu_{jk.} - \mu_{j.m} - \mu_{.km} + \mu_{j..} + \mu_{.k.} + \mu_{..m} - \mu_{...}$$

- Null hypotheses can be formulated using parameters, or using (marginal) cell means.
- F-tests based on incremental sums of squares can be formulated.
- Type II sums of squares (respecting marginality) and Type III sums of squares may be used.

Higher-order classifications (1)

- Extension to more than 3 factors, say p factors, is straightforward.
- Principle of marginality guides in interpreting effects.
- Three-way interactions are reasonably complex. Higher-order interactions even more difficult to interpret.
- Higher-order interactions may be observed, e.g. if specific combination of characteristics predisposes individuals to act in certain manner.
- But higher-order interactions may be not-significant, or negligibly small relative to other effects.
- Not all interactions need to be taken into model. Sometimes models with main effects (additive models) only, or main effects and two-way interactions only, are fitted, limiting consideration to effects of theoretical interest.

Higher-order classifications (2)

- Fox shows example 4-way ANOVA. A few remarks:
  - ANOVA table shows sums of squares respecting principle of marginality ("Type II").
  - In large datasets even trivial effects can prove to be "statistically significant"; we may wish to ignore such effects.
  - Criteria, other than the P-value, may assist in model selection (see chapter 22).
  - In example, a model with main effects and one two-way interaction (15 parameters) had $R^2 = 0.267$, whereas the full model (100 parameters) had $R^2 = 0.275$.
  - "Effect display" guides interpretation of results: it shows each effect, keeping other explanatory variables at average values.

Empty cells in ANOVA

- With increasing number of factors, number of cells grows much faster, and some cells in p-way classification may be empty.
- Complete factorial model is not possible then.
- Empty cell does not give problem, if marginal frequency tables, corresponding to effects that are included, contain no empty cells.
- Example, in 2×2 classification, with one empty cell, we may fit the additive model $Y_{ijk} = \mu + \alpha_j + \beta_k + \epsilon_{ijk}$
  but not the complete factorial model $Y_{ijk} = \mu + \alpha_j + \beta_k + \gamma_{jk} + \epsilon_{ijk}$.

Some special experimental designs

So far, ANOVA model is described in factorial situations, with one or more treatment factors, and each level of a factor combined with each level of other factor(s).
Faraway looks at number of special experimental designs, that can be analyzed with ANOVA:

- Randomized Complete Block Design
- Latin squares
- Balanced Incomplete Block Design
- Fractional Factorial Design

Randomized complete block design

- In Completely Randomized Design or *CRD*, treatments are assigned to experimental units at random.
- This is appropriate if units are homogeneous.
- With heterogeneous units, it may be better to arrange experimental units into blocks where the intrablock variation is small, but interblock variation large.
- If number of experimental units within block equals number of treatments, so that all treatments can be assigned (randomized) to units within the block, design is called *RCBD* = Randomized Complete Block Design.
- Example: 3 crop varieties on 5 fields. Divide each field in 3 plots, and randomly assign 3 crop varieties to 3 plots within field.
- Model: $Y_{ij} = \mu + \tau_i + \rho_j + \epsilon_{ij}$: additive model with main effects of block and treatment factor.
- Assessment of interaction is difficult, because there is single observation per block-treatment combination. Tukey-1 df test for nonadditivity is possibility.
- Relative advantage of *RCBD* over *CRD* by relative efficiency $= \frac{\hat{\sigma}^2_{CRD}}{\hat{\sigma}^2_{RCBD}}$. Estimates of residual variance are obtained by fitting models with and without the block effect.
- Penicilline example: relative efficiency is 1.62, meaning that 62% more observations are needed for *CRD* to obtain same precision as *RCBD*.

Balanced Incomplete Block design

- In complete block, block size equals number of treatments.
- If block size is less than number of treatments, incomplete block design can be used.
- In incomplete block design, treatments and block are not orthogonal. Some treatment contrasts are not identifiable from certain block contrasts: confounding.
- In Balanced Incomplete Block Design = *BIBD*, all pairwise differences are identifiable and have equal standard error. Pairwise differences may be more interesting than other contrasts.
- Example: 4 treatments (A,B,C,D), block size $k = 3$, number of blocks $b = 4$.

1	A	B	C
2	A	B	D
3	A	C	D
4	B	C	D

- Each pair of treatments occurs in same block $\lambda = 2$ times.
- We cannot have *BIB* for any number of treatments, block size, and number of blocks, only for special cases.
- Model: $Y_{jik} = \mu + \tau_i + \rho_j + \epsilon_{ijk}$. Again, main effects only.

Latin squares

- You might call this a "Sudoku" design...
- Useful if there are two blocking variables.
- Example 1: agricultural experiment: in field moisture level may vary in one direction, and fertility in another direction.
- Example 2: industrial experiment: suppose we wish to compare 4 production methods (4 treatments A,B,C,D). Available are 4 machines (1,2,3,4), and 4 operators (I,II,III,IV).
- Latin square design could be:

	1	2	3	4
I	A	B	C	D
II	B	D	A	C
III	C	A	D	B
IV	D	C	B	A

- Each treatment assigned to each row and column one and once only.
- Block sizes must be equal to number of treatments.
- Model: $Y_{jik} = \mu + \tau_i + \beta_j + \gamma_k + \epsilon_{ijk}$. Main effects only!
- Latin square can also be used for comparing 3 treatment factors; only t^2 runs are required instead of t^3 for full factorial. This is an example of a fractional factorial design.

Fractional factorial design

- Suppose we have number of factors, each with number of levels.
- Full factorial design has at least one run for each combination of the levels.
- Total number of combinations can quickly grow very large with increasing number of factors.
- In full factorial design, all interactions (up to highest order) are estimable.
- Fractional factorial design uses only fraction of number of runs of full factorial design, e.g. to reduce cost, or because only limited experimental material is available.
- Suppose 7 factors, each with 2 levels: 1 mean, 7 main effects, 21 two-way interactions, 35 3-way interactions, 35 4-way interactions, 21 5-way interactions, 7 6-way interactions, 1 7-way interaction. Is it reasonable to assume that higher order interactions are negligible? If so, not all $2^7 = 128$ runs are needed.
- Maybe a quarter may be run, and still all main effects and 2-way interactions are estimable.
- Idea of fractional factorial experiment: try to estimate many parameters with as little data as possible.
- Fractional factorial design popular in engineering applications, like product design.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 4, lecture 22

Content lecture 22: Faraway: 3.9 + 4-6

- Overview problems in linear models, diagnostics
- Errors in predictors
- Generalized and weighted least squares
- Testing for lack of fit: regression vs ANOVA

Problems in Linear models

- We study what might go wrong in fitting linear models to data, how to diagnose the problems, and how to repair (some of) them.
- We will start with some sections from the Faraway text, and in next lectures continue with Fox chapters 11-13.
- Fox treats the following topics:
  - Chapter 11: Unusual and influential data: about leverage, outliers, and influence diagnostics;
  - Chapter 12: Diverse topics: nonlinearity, nonconstant variance, and non-normality;
  - Chapter 13: Collinearity
- We will skip some details from Fox; chapters 12-13 next week

What can go wrong?

Faraway categorizes things that can go wrong:

- Sources and quality of data: how data was collected directly influences type of conclusions.
  1. We may have biased sample from population of interest. Extrapolation towards population is difficult then.
  2. Important predictors may have been missed. Prediction may be poor, or relationship between predictors and response is misinterpreted; predictors may have been measured with error.
  3. Observational data make causal conclusion problematic. Confounding between predictors makes disentangling into separate effects difficult.
  4. Range and quantitative nature of data may limit predictions. Don't extrapolate.
- Error component. Remember model assumptions: $\epsilon \sim N_n(0, \sigma^2 \mathbf{I}_n)$.
  1. Errors may be heterogeneous (unequal variance).
  2. Errors may be correlated.
  3. Errors may not be normally distributed. This point is less serious, because even without normal errors, $\hat{\beta}$'s are approximately normal due to central limit theorem. Especially in larger datasets non-normality is not big issue.
- Systematic part of model $E(\mathbf{y}) = \mathbf{X}\beta$ (Faraway: structural part of model) may be incorrect.
  - Actual model may come from physical theory, in which case a simple linear model can only be regarded as an approximation to a complex reality.
  - Best one can hope for, is that model is fair representation of reality.
  - "Model can be no more than a good portrait".
  - Remember George Box: "All models are wrong, but some are useful".

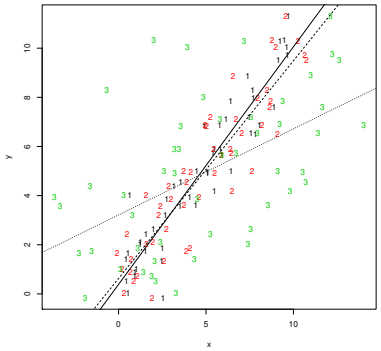
Diagnostics

- We will study most topics mentioned on earlier slide (but not all).
- Estimation and inference from regression models depend on several assumptions, which need checking.
- Assumptions are checked using regression diagnostics.
- Potential problems fall in 3 categories:
  - Random part: do we have constant variance, uncorrelatedness, normal distribution?
  - Systematic part: fixed part of model may not be correct;
  - Unusual observations: sometimes a few observations might change choice and fit of model.
- Diagnostic techniques can be graphical or numerical.
- Regression diagnostics may suggest improvements.
- Model building is iterative and interactive.

Errors in predictors

- This topic was touched upon in Fox section 6.4.
- We are not going into any details, but the main idea is that errors in predictors cause bias in regression coefficients, typically towards zero: attenuation.
- Books are written full about this topic, e.g. Fuller, W. A. (1987). Measurement error models.
- Below, Faraway shows ideas by simulation.

```
>  # Data generating model: y = 0 + 1*x + e:
> x<- 10*runif(50); y <- x + rnorm(50)
> gx <- lm(y ~x); coef(gx)
(Intercept)           x
     0.3870      0.9671
> z <- x + rnorm(50) # Add some noise to regressor
> gz <- lm(y ~ z); coef(gz)
(Intercept)           z
     0.6278      0.8862
> z2 <- x + 5*rnorm(50) # Add more noise
> gz2 <- lm(y ~ z2); coef(gz2)
(Intercept)          z2
     3.2139      0.3504
> matplot(cbind(x,z,z2),y,xlab="x",ylab="y")
> abline(gx,lty=1); abline(gz, lty=2); abline(gz2, lty=3)
```



Generalized least squares (1)

- Model assumptions: $\epsilon \sim N_n(0, \sigma^2 \mathbf{I}_n)$. But it could be that errors have non-constant variance or are correlated.
- Suppose covariance matrix $V(\epsilon) = \sigma^2 \Sigma$, where σ^2 is unknown, but Σ known.
- So, we know the correlation and relative variance between errors, but not absolute scale.
- Generalized least squares (GLS) may be used, minimizing

$$(\mathbf{y} - \mathbf{X}\beta)^T \Sigma^{-1} (\mathbf{y} - \mathbf{X}\beta)$$

solved by $\hat{\beta} = \left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{y}$.

Generalized least squares (2)

- Instead of OLS (Ordinary Least Squares) estimators now GLS (Generalized Least Squares) estimators are obtained:

$$\hat{\beta}_{GLS} = \left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \Sigma^{-1} \mathbf{y}$$

- Some background:
  - Cholesky decomposition: any real symmetric positive-definite matrix, like Σ, can be written as $\Sigma = \mathbf{S}\mathbf{S}^T$, where $\mathbf{S}$ is lower triangular matrix.
  - Hence, premultiplying $\mathbf{y} = \mathbf{X}\beta + \epsilon$ with $\mathbf{S}^{-1}$ gives $\mathbf{S}^{-1}\mathbf{y} = \mathbf{S}^{-1}\mathbf{X}\beta + \mathbf{S}^{-1}\epsilon$, or, renaming, $\mathbf{y}' = \mathbf{X}'\beta + \epsilon'$ (note that $'$ is not transposing now.)
  - The new error vector ϵ' has covariance matrix $V(\epsilon') = V(\mathbf{S}^{-1}\epsilon) = \mathbf{S}^{-1} V(\epsilon) (\mathbf{S}^{-1})^T = \mathbf{S}^{-1}\sigma^2\Sigma(\mathbf{S}^{-1})^T = \sigma^2 \mathbf{S}^{-1}\mathbf{S}\mathbf{S}^T(\mathbf{S}^{-1})^T = \sigma^2 \mathbf{I}_n$.
  - So, new variable $\mathbf{y}'$ and model matrix $\mathbf{X}'$ are related by regression equation with uncorrelated errors with equal variance: back home!
- $V(\hat{\beta}) = \left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right)^{-1} \sigma^2$.
- GLS is described in Fox in chapter 16, which we largely skip.

Weighted least squares

- Sometimes errors are uncorrelated, but have unequal variance, where form of inequality is known.
- Weighted least squares (WLS) may be used, which is just special case of GLS.
- $V(\epsilon) = \sigma^2 \text{diag}(1/w_1, \ldots, 1/w_n)$ with known weights w_i.
- Finding Cholesky decomposition is very easy: $S = \text{diag}(\sqrt{1/w_1}, \ldots, \sqrt{1/w_n})$.
- In essence, WLS is regressing $\sqrt{w_i}y_i$ on $\sqrt{w_i}x_i$ (and column of 1's in $\mathbf{X}$ replaced by $\sqrt{w_i}$.
- Observations with low variability get high weight, and vice versa.
- Examples:
  - Errors proportional to predictor: $var(\epsilon_i) \propto x_i$, hence $w_i = x_i^{-1}$.
  - Y_i is average over n_i observations; $var(Y_i) = var(\epsilon_i) = \sigma^2/n_i$, hence $w_i = n_i$.
- WLS is described in Fox section 12.2 under Nonconstant error variance.

Testing for Lack of fit

- How to tell if a model fits the data? One aspect is looking at $\hat{\sigma}^2_\epsilon$, based on the chosen regression model: should be unbiased estimate of σ^2_ϵ .
- If model is not complex enough or takes wrong form, $\hat{\sigma}^2_\epsilon$ will overestimate σ^2_ϵ .
- If model is too complex, $\hat{\sigma}^2_\epsilon$ may underestimate σ^2_ϵ .
- So, for testing, we could compare $\hat{\sigma}^2_\epsilon$ with σ^2_ϵ .
- If σ^2_ϵ is known, $\dfrac{(n-p)\hat{\sigma}^2_\epsilon}{\sigma^2_\epsilon} \sim \chi^2_{n-p}$, and test of lack of fit could be based on this.
- More realistically, σ^2_ϵ is unknown; $\hat{\sigma}^2_\epsilon$ from chosen regression model should be compared with some model-free estimate of σ^2_ϵ.
- This can be done if there are repeated values of y for one or more fixed x. Repeated measurements should be independent replicates, not just repeated measurements on same object or unit.
- Independent repeated measurements give pure error estimate of σ^2_ϵ: SS_{PE}/df_{PE}, with $SS_{PE} = \sum_j \sum_i \left(y_{ij} - \overline{y}_j\right)^2$ with $df_{PE} = \sum_j (\text{number replicates} - 1) = n - \text{nr groups}$.
- You may recognize SS_{PE}: it is the within groups sum of squares from one-way ANOVA in which regressor x is treated as factor!

Testing for Lack of fit: regression versus ANOVA

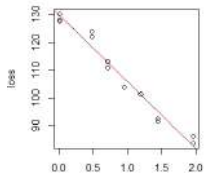
- Lack of fit test is comparison of regression model with ANOVA model.
- Partition the residual sum of squares RSS into 2 components:

	df	SS	MS	F
Lack of fit	$n - p - df_{PE}$	$RSS - SS_{PE}$	$\frac{RSS - SS_{PE}}{n-p-df_{PE}}$	Ratio of MS's
Pure Error	df_{PE}	SS_{PE}	SS_{PE}/df_{PE}	
Residual	$n - p$	RSS		

- Note that, as always, not rejecting H_0: "model fits adequately" does not necessarily mean that H_0 is true.

```
> g <- lm(loss ~ Fe, data=corrosion)  # Simple linear regression
> deviance(g) # Residual sum of squares from simple linear regression
[1] 102.9
> ga <- lm(loss ~ factor(Fe), data=corrosion)  # Fe is factor: ANOVA
> deviance(ga) # Residual sum of squares from one-way ANOVA
[1] 11.78
> deviance(ga)/ga$df.residual # Pure error variance estimate
[1] 1.964
> anova(g,ga)
Analysis of Variance Table

Model 1: loss ~ Fe
Model 2: loss ~ factor(Fe)
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     11 102.9
2      6  11.8  5      91.1 9.28 0.0086
```

```
> corrosion
    Fe  loss
1 0.01 127.6
2 0.01 130.1
3 0.01 128.0
4 0.48 124.0
5 0.48 122.0
6 0.71 110.8
7 0.71 113.1
8 0.95 103.9
9 1.19 101.5
10 1.44  92.3
11 1.44  91.4
12 1.96  83.7
13 1.96  86.2
```



Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 4, lecture 23

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Content lecture 23: Fox §11.1-11.3; Faraway §7.1-7.3

- Leverages and hat matrix
- Outliers: residuals, standardized and studentized residuals

Outliers, Leverage, and Influence

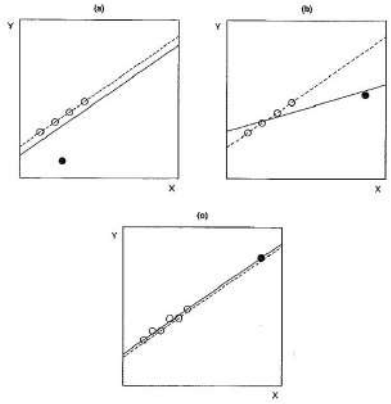
- Why bother about outliers? Unusual data are problematic in linear models fit by least squares, because they may unduly influence results of analysis.
- Regression outlier is observation whose response-variable value is conditionally unusual given value of explanatory variable(s).
- Observation has high leverage if its regressor values are extreme so that it potentially has strong leverage (influence) on regresion coefficients.
- Observation has high influence if it both is discrepant (a regressino outlier) and has high leverage:

$$\text{Influence on coefficients} = \text{Leverage} \times \text{Discrepancy}$$

Example Outliers, Leverage, and Influence

Figures show basic distinctions for simple linear regression.

- Figure a) Low leverage, but regression outlier, so influence is weak. Observation has X at center of X-distribution, leverage is low. Y is extreme, given the value of X, so it is regression outlier: highly discrepant. Deletion of observation hardly has impact on slope, slightly affects intercept.
- Figure b) High leverage, and regression outlier, so influence is strong. X is unusually large, so high leverage. Y given X value is extreme, so it is regression outlier. Deletion of observation will markedly affect slope and intercept.
- Figure c) High leverage, but not regression outlier, so influence is weak. X is unusually large (high leverage), but Y given X is not extreme. Deletion will not change slope and intercept substantially.



Assessing Leverage: Hat-values (1)

- Hat-value h_i (or h_{ii}) is measure of leverage in regression.
- Hat-values are diagonal elements of hat-matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, which "puts the hat on Y":
  - Remember least squares estimator: $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; then fitted values are $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{H}\mathbf{y}$; so, $\mathbf{H}$ transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$;
  - $\mathbf{H}$ is projection matrix, it orthogonally projects $\mathbf{y}$ on the plane spanned by the columns of $\mathbf{X}$;
  - $\mathbf{H}$ is idempotent ($\mathbf{H}^2 = \mathbf{H}$);
  - $\mathbf{H}$ is symmetric.

Assessing Leverage: Hat-values (2)

- Property of hat-values h_i: $0 \leq h_i \leq 1$. This follows from $h_i = \sum_{j=1}^{n} h_{ij}^2 = h_i^2 + \sum_{j \neq i} h_{ij}^2$ (because of symmetry and idempotency of $\mathbf{H}$). Hence $h_i > 0$ (as h_i is sum of squares of h_{ij}) and $h_i \leq 1$ (as $h_i^2 = h_i - \sum_{j \neq i} h_{ij}^2$, so $h_i^2 \leq h_i$, which is the case for $h_i \leq 1$). If intercept is in model $1/n \leq h_i \leq 1$.
- Rank of $\mathbf{H}$ is $k+1$ (for regression model with k regressors and intercept); as the trace of an idempotent matrix equals its rank (not proven), $\text{trace}(\mathbf{H}) = \sum h_i = k+1$. Hence average h_i value is $(k+1)/n$.
- Leverage h_i measures distance from the centroid of the Xs, taking into account correlational and variational structure of Xs (Mahalanobis distance).
- Fitted values $\hat{Y}_j = h_{1j}Y_1 + h_{2j}Y_2 + \ldots + h_{jj}Y_j + \ldots + h_{nj}Y_n$. So, weight h_{ij} captures contribution of observation Y_i to fitted value $\hat{Y}_j$.
- $\mathbf{H}$ depends only on regressors, not on $\mathbf{y}$!

Leverages with R

- Example Davis data: one large outlier of measured weight.
- What would be average value of leverage? We have $n = 183$ observations, and $k = 3$ regressors (i.e. $k+1 = 4$ parameters, intercept included).

```
> g1 <- lm(repwt ~ weight + factor(sex) + weight:factor(sex), data=Davis)
> lev <- lm.influence(g1)$hat
> sort(lev,decreasing=T)[1:10]  # 10 largest leverages
     12      21      97      54      30     156      65      82     118     169
0.71419 0.16684 0.07321 0.06878 0.06451 0.05254 0.04912 0.04895 0.04569 0.04569
> # Alternative way of getting leverages
> X <- model.matrix(g1)
> lev2 <- hat(X)
> sort(lev2,decreasing=T)[1:10]
 [1] 0.71419 0.16684 0.07321 0.06878 0.06451 0.05254 0.04912 0.04895 0.04569 0.04569
```


Detecting Outliers: Residuals

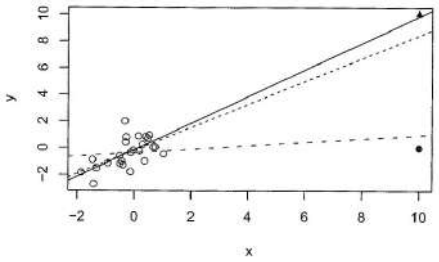
- Least squares residuals $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{X}\beta + \epsilon) - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \epsilon) = \epsilon - \mathbf{H}\epsilon = (\mathbf{I}_n - \mathbf{H})\epsilon$.
- Hence, $E(\mathbf{e}) = E((\mathbf{I}_n - \mathbf{H})\epsilon) = (\mathbf{I}_n - \mathbf{H})E(\epsilon) = (\mathbf{I}_n - \mathbf{H})\mathbf{0}_n = \mathbf{0}_n$ Residuals have expectation zero.
- Further, $V(\mathbf{e}) = V((\mathbf{I}_n - \mathbf{H})\epsilon) = (\mathbf{I}_n - \mathbf{H})\sigma_\epsilon^2\mathbf{I}_n(\mathbf{I}_n - \mathbf{H})' = \sigma_\epsilon^2(\mathbf{I}_n - \mathbf{H})$, because, like $\mathbf{H}$ also $\mathbf{I}_n - \mathbf{H}$ is symmetric and idempotent. Residuals generally do not have equal variance, and are not uncorrelated!
- Single residual: $V(E_i) = \sigma_\epsilon^2(1 - h_i)$; so, a large leverage will make variance of residual small, in other words: fit will be forced to be close to Y_i.
- Notice that $V(\hat{\mathbf{y}}) = \sigma_\epsilon^2\mathbf{H}$, so for single fitted value: $V(\hat{Y}_i) = \sigma_\epsilon^2 h_i$.

Detecting Outliers: Standardized Residuals

- Residuals can be standardized to have variance 1 as $E_i' \equiv \dfrac{E_i}{S_E\sqrt{1 - h_i}}$.
- These residuals are named differently by different authors:
  - Fox calls them standardized residuals;
  - Faraway calls them (internally) studentized residuals.
- Value of standardized residual gives idea about outlyingness of observation. Values larger than 3 or smaller than -3 are unlikely to occur.

Problems with Standardized Residuals

- Outliers can conceal themselves.
- Example: 2 high leverage observations: ▲ and •.
  solid line: including ▲, excluding •;
  dashed line: including •, excluding ▲;
  dotted line: both excluded.
- ▲ is not influential observation, because regression line hardly changes after removal; • is influential, because regression line changes dramatically.
- How can you detect the influential observation •? Not by checking E or E_i'...



- Other problem with standardized residuals is that numerator and denominator are not independent, preventing E_i' to follow t-distribution: if $|E_i|$ is large, estimate of standard deviation $S_E = \sqrt{\sum E_i^2/(n-(k+1))}$ tends to be large as well.

Detecting Outliers: Studentized Residuals

- Do leave-one-out cross validation for residuals: exclude point i, recompute estimates to get $\mathbf{b}_{(-i)}$ and $\hat{\sigma}^2_{(-i)}$.
  Next make prediction for observation i: $\hat{Y}_{(-i)} = \mathbf{x}_i'\mathbf{b}_{(-i)}$.
- If $Y_i - \hat{Y}_{(-i)}$ is large, then case i is outlier.
- Ordinary residuals may miss nasty points, because the regression line is pulled so close to them that true status is concealed.
- How large is large? $V\left(Y_i - \hat{Y}_{(-i)}\right) = \sigma^2\left(1 + \mathbf{x}_i'\left(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}\right)^{-1}\mathbf{x}_i\right)$
  estimated as $\hat{V}\left(Y_i - \hat{Y}_{(-i)}\right) = \hat{\sigma}^2_{(-i)}\left(1 + \mathbf{x}_i'\left(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}\right)^{-1}\mathbf{x}_i\right)$
- Externally studentized residual is defined as:

$$E_i^* \equiv \frac{\left(Y_i - \hat{Y}_{(-i)}\right)}{\hat{\sigma}_{(-i)}\left(1 + \mathbf{x}_i'\left(\mathbf{X}_{(-i)}'\mathbf{X}_{(-i)}\right)^{-1}\mathbf{x}_i\right)^{1/2}}$$

- If model is correct, $E_i^* \sim t_{n-1-(k+1)}$.
- Other names, simply studentized residuals, jackknife residuals, cross-validated residuals, deleted residuals.
- Naming in R is as in Fox: standardized and studentized residuals (functions `rstandard` and `rstudent`).

Detecting Outliers: Studentized Residuals (2)

- Instead of doing n regressions, externally studentized residuals may be calculated directly as

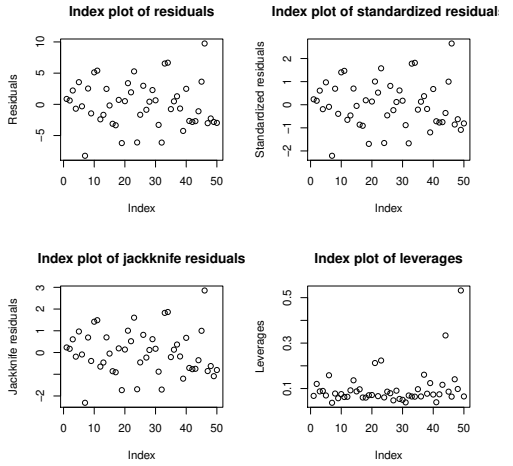
$$E_i^* = \frac{E_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_i}} = E_i'\sqrt{\frac{n-k-2}{n-k-1-E_i'^2}}.$$

- If n is large, factor under square root is close to 1, and distinction between internally and externally studentized residuals essentially disappears.
- Outlier tests may be done based on E_i^*, as $E_i^* \sim t_{n-k-2}$.
  Bonferroni adjustment of P-value is suggested to correct for multiple comparisons, with Bonferroni P-value = $2 \times n \times$ one-sided P-value from t-distribution.
  `car` package contains function `outlierTest`.
- Alternative approach is QQ-plot for externally studentized residuals against quantiles of proper t-distribution.
- We skip Fox section 11.3.2.

Savings data Faraway: R example

Some index plots (x-axis shows observation numbers, $i = 1, \ldots, 50$) of residuals and leverages:

```
> g <- lm(sr ~ pop15 + pop75 +dpi + ddpi, , data=savings)
> plot(g$res, ylab="Residuals", main="Index plot of residuals")
> plot(rstandard(g), ylab="Standardized residuals", main="Index plot of standardized residuals")
> plot(rstudent(g), ylab="Jackknife residuals", main="Index plot of jackknife residuals")
> plot(lm.influence(g)$hat, ylab="Leverages", main="Index plot of leverages")
```



Some further remarks about outliers

- General remarks:
  - Two or more outliers next to each other can hide each other.
  - Outlier in one model may not be outlier in another when variables have been changed or transformed.
  - Error distribution may be non-normal, so that larger residuals may be expected.
  - Individual outliers much less of a problem in larger datasets: single point will not have leverage to affect fit considerably. However, clusters of outliers may.
- What to do about outliers?
  - Check data-entry errors first.
  - Examine physical context: what did happen? Discovery of outlier may be of great interest.
  - Exclude point from analysis, try reinclude later, compare results. Report honestly about existence of outliers, even if not included in model.
  - Robust regression may be preferred if outliers exist, which cannot be identified as mistakes or aberrations.
  - Don't exclude outliers in automated way.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 4, lecture 24

Content lecture 24: Fox §11.4-11.6; Faraway §7.4-7.5, 7.7

- Further diagnostics for influence: DFBETAS, Cook's distance, COVRATIO
- Residual plots
  - residuals vs fitted values to check constant variance
  - residuals vs regressor to check linearity
  - added variable plot to check joint influence on individual slope

Influential observations: influence on $\hat{\beta}$ and $\hat{\mathbf{y}}$

- Influential point is one whose removal from dataset would cause large change in the fit.
- Influential point may or may not be an outlier, and may or may not have large leverage, but will tend to have at least one of these properties.
- Measures of influence include
  - Change in coefficients $D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}$ (for $i = 1, \ldots, n$, $j = 0, 1, \ldots, k$), often called DFBETA_{ij}.
    Standardized version, dividing by (deleted) standard error, called DFBETAS_{ij}:
    $$D^*_{ij} = \frac{D_{ij}}{SE_{-i}(B_j)}$$
  - Change in fit $\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)} = \mathbf{X}(\hat{\beta} - \hat{\beta}_{(-i)})$
- Disadvantage is that there are so many to look at: $n \times (k+1)$ DFBETAS, and n vectors of length n.

Influential observations: Cook's distance

- Alternative: Cook's distance $D_i = \dfrac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(-i)})}{(k+1)\hat{\sigma}^2} = \dfrac{E_i'^2}{k+1} \times \dfrac{h_i}{1-h_i}$.
  - Based upon F-statistics for "hypothesis" that $\beta_j = B_{j(-i)}$, measuring "distance" between B_j and $B_{j(-i)}$.
  - First term in D_i is measure of discrepancy; second term is measure of leverage.
  - D_i measures overall influence of observation i on regression coefficients; alternative interpretation is that it measures aggregate influence of observation i on fitted values $\hat{\mathbf{y}}$.
  - Look for value of D_i that stand out from rest.

```
> cook <- cooks.distance(g)
> range(cook)
[1] 4.737e-05 2.681e-01
```

- Alternative to Cook's distance is $DFFITS_i = E_i^* \sqrt{\dfrac{h_i}{1-h_i}}$.
- All deletion statistics depend on hat-values and residuals. Therefore plot of E_i^* versus h_i is useful. Look for observations with both big.

Influential observations: COVRATIO

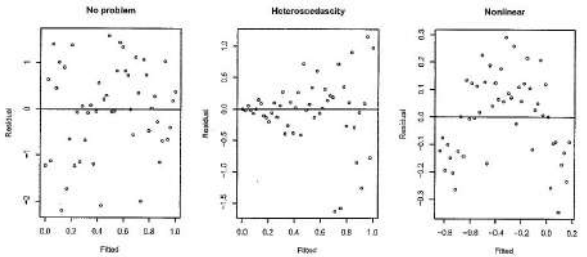
- Not only beta's and fitted values are subject to influence.
- Residual standard deviation, and consequently, confidence intervals and confidence regions change if individual observations are deleted.
- Influence measure COVRATIO approximates squared ratio of volumes of deleted and full-data confidence regions for regression coefficients:
  $\text{COVRATIO}_i = \dfrac{1}{(1-h_i)\left(\dfrac{n-k-2+E_i^{*2}}{n-k-1}\right)^{k+1}}$.
  - Observations that increase precision of estimation have values of COVRATIO larger than 1.
  - Observations that decrease precision of estimation have values of COVRATIO smaller than 1.
  - Look for values that differ considerably from 1.
  - Large hat-value produces large COVRATIO.
  - Low hat-value may not change coefficients much, but it decreases precision of estimation by increasing estimated error variance, leading to small COVRATIO.

Numerical cutoffs for diagnostic statistics

- Hat-values: use $h_i > 2 \times \overline{h} = 2 \times (k+1)/n$ (though this tends to select too many points, alternative is $3 \times \overline{h}$)
- Externally studentized residuals: draw lines at ± 2 to draw attention to observations outside this range; under ideal conditions about 5% of studentized residuals are outside this range.
- |DFBETAS| > 2 may be used as cutoff, as it is a standardized coefficient.
- Cook's distance $D_i > \dfrac{4}{n-k-1}$.
- $|\text{COVRATIO}_i - 1| > \dfrac{3(k+1)}{n}$.

Residual plots

- Looking at all these influential diagnostics, we almost forget to do the basic check for model assumptions: look at residual plots!
- Most important plot: residuals **e** (on y-axis) versus fitted values $\hat{\mathbf{y}}$ (on x-axis).
- If all is well, you should see constant variance in vertical direction, and scatter should be symmetric vertically about 0.
- Look for: heteroscedasticity (nonconstant variance), and nonlinearity.

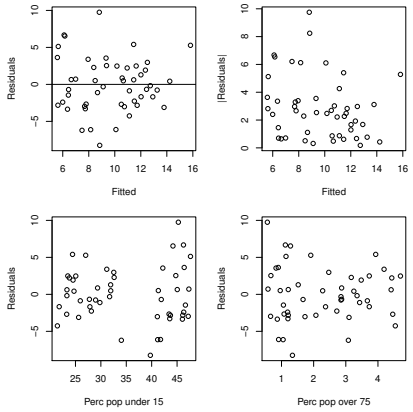


- Also plot residuals **e** versus regressors **x** for regressors that are included in the model and that are excluded from the model.
- Look for the same type of distortions; for regressors, excluded from the model, look for any relationship that may indicate that the predictor should be included.

Residual plots: example

Following Faraway with `savings` data

```
> g <- lm(sr ~ pop15 + pop75 +dpi + ddpi, , data=savings)
> plot(g$res ~ g$fit, xlab="Fitted", ylab="Residuals"); abline(h=0)
> plot(abs(g$res) ~ g$fit, xlab="Fitted", ylab="|Residuals|")
> # Plotting absolute values simplifies diagnosis of non-constant variance
> plot(g$res ~ savings$pop15, xlab="Perc pop under 15", ylab="Residuals")
> plot(g$res ~ savings$pop75, xlab="Perc pop over 75", ylab="Residuals")
```



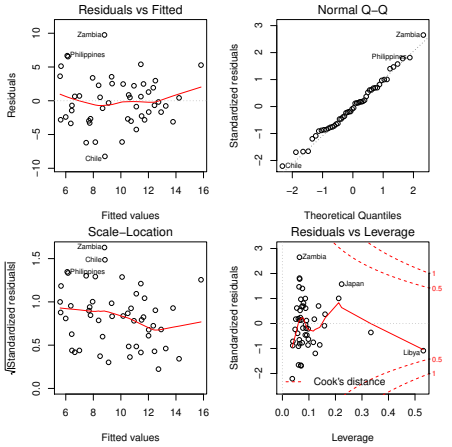
Default residual plots in R: plot() function

R's function `plot()` processes objects of class `lm`, producing four residual plots:

```
> plot(g)
```

1. Residuals versus fitted values
2. Normal QQ-plot for standardized residuals
3. $\sqrt{|\text{Standardized residuals}|}$ versus fitted values
4. Standardized residuals versus leverages with added contour lines of equal Cook's distance (by default 0.5 and 1)

Two other plots concerning Cook's distance are optional.

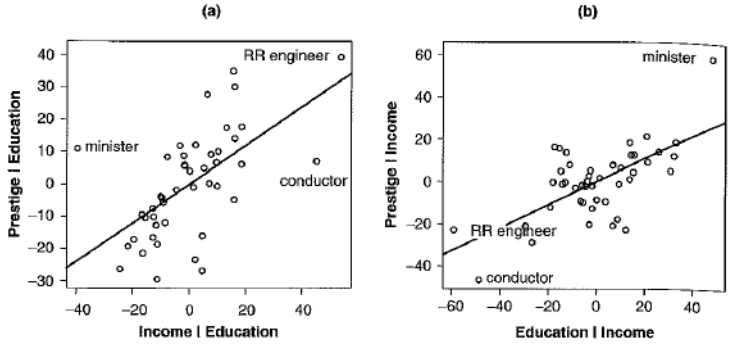


Added-variable plots

- Subsets of observations can be jointly influential or can offset each others influence.
- Deletion statistics could be generalized to subsets of several points, but this approach is rather impractical given the large number of subsets.
- Alternative: use graphical methods, especially added-variable plot, also called partial-regression plot.
- In added-variable plot the effect of a regressor, say X_1, is studied, after correction for all other regressors:
  - $Y_i^{(1)}$ are residuals from regression of Y on all Xs with exception of X_1.
  - $X_i^{(1)}$ are residuals from regression of X_1 on all Xs (but X_1).
- These residuals have following properties:
  - Residuals $Y_i^{(1)}$ and $X_i^{(1)}$ are parts of Y and X_1 when effects of $X_2, \ldots, X_k$ are "removed".
  - Slope from simple regression of $Y_i^{(1)}$ on $X_i^{(1)}$ is slope B_1 from full multiple regression.
  - Residuals from simple regression of $Y_i^{(1)}$ on $X_i^{(1)}$ are same as those from full regression.
  - Variation of $X_i^{(1)}$ is conditional variation of X_1 holding other Xs constant; standard error of slope B_1 from simple regression of $Y_i^{(1)}$ on $X_i^{(1)}$ is same as multiple regression standard error of slope.

Added-variable plots (2)

- Now plot $X^{(j)}$ versus $Y^{(j)}$ for each $j = 1, \ldots, k$.
  - Plot provides visual impression of precision of slope B_1.
  - Plot permits examination of leverage and influence of observations on individual slope B_1.
- Example Added-variable plots Duncan's regression of prestige on education and income.



Partial residual plot

- Partial residual plot is competitor to added-variable plot.
  - Faraway discusses this type of plots next to added-variable plots.
  - Fox postpones the topic to chapter 12, and calls them component-plus-residual plots.
  - Following Fox, we will discuss these in a next lecture.
- We skip sections 11.6.2 (Forward Search), some matrix algebra results 11.8.3, and 11.8.4.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 5, lecture 25

Content lecture 25: Fox §12.1-12.2

- Graphical methods to check assumptions in LM
  - non-normality; QQ-plots with/without envelope
  - heteroscedasticity; plot (studentized) residuals or sqrt(abs(residuals)) vs fitted values; weighted least squares

Diagnosing Non-normality and Non-constant Error Variance

- Chapter 11 focused on problems with specific observations.
- Chapter 12-13 deal with more general problems of model specification.
- First we focus on graphical methods to check
  - non-normally distributed errors;
  - nonconstant error variance;
- Basic example: SLID data. Regression of hourly wage rate ($/hour) on dummy variable for sex (1=males), education (years), age (years). Model may be too simplistic.

```
> SLID <- read.table("SLID-Ontario.txt",header=T)
> slidreg <- lm(wages ~ sex + age + education, data=SLID)
> coef(summary(slidreg))
             Estimate Std. Error t value   Pr(>|t|)
(Intercept)  -8.1242   0.598977  -13.56  5.268e-41
sexMale       3.4737   0.207009   16.78  4.038e-61
age           0.2613   0.008664   30.16 3.424e-180
education     0.9296   0.034257   27.14 5.473e-149
> summary(slidreg)$r.squared
[1] 0.3074
```

Non-Normally Distributed Errors (1)

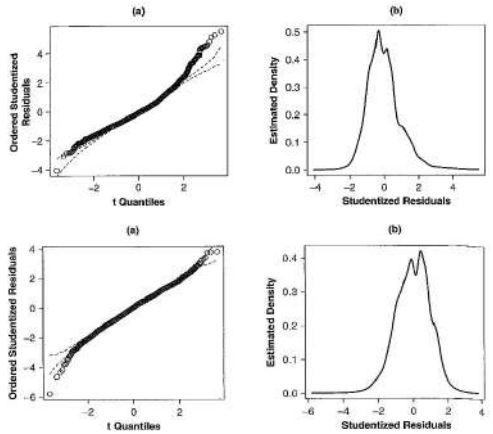
- Central limit theorem states that, under broad conditions, inference (hypothesis tests, confidence intervals) based on least-squares estimator is approximately valid in all but small samples.
- Why bother then?
  - Validity of least-squares estimation is robust, but efficiency is not: under normality least-squares estimator is the most efficient unbiased estimator, but with other error distributions this is not the case. Robust estimators may be more efficient. Heavy-tailed error distributions are problematic because they give rise to outliers.
  - Highly skewed error distributions generate outliers in direction of skew, but also compromise interpretation of least-squares fit. Fit is conditional mean (of Y given Xs). But mean is not good measure of center of highly skewed distribution. Preferable to transform data to produce more symmetric distribution.
  - Multimodal error distribution suggests omission of discrete explanatory variable. Examination of distribution of residuals may motivate respecification of model.

Non-Normally Distributed Errors (2)

- Graphical check of normality: QQ-plot of residuals
  - Plot preferably the externally studentized residual E_i^* versus normal or t_{n-k-2} distribution.
  - Only for small samples is difference between normal and t distribution important.
  - In larger samples internally studentized residuals or raw residuals will give same impression.
  - Note that studentized residuals are not independent random sample from t_{n-k-2}. Dependencies are generally negligible.
  - QQ-plot effective in displaying tail behaviour.

Non-Normally Distributed Errors (3)

- Other univariate graphical displays: in large samples histograms with many bars may reveal multimodal distribtutions better. In smaller samples, smoothed histogram.
- Positive skew corrected by moving response down the ladder, e.g. square root- or log-transforms.
- Below QQ-plot and smoothed histogram of cross-validated residuals; further down after log-transformation.



Confidence Envelopes by Simulated Sampling

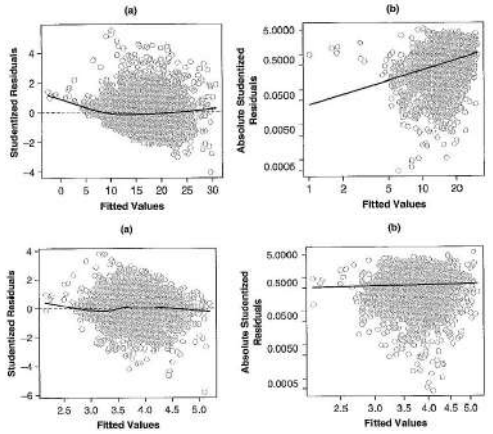
- Atkinson (1985) suggested procedure for construction of approximate confidence "envelope" in residual quantile-comparison plot.
  - Repeatedly (m times) simulate vectors $\mathbf{y}$, based on fitted regression model.
  - Regress simulated responses on Xs.
  - Calculate studentized residuals from each regression.
  - Determine the appropriate quantiles of simulated residuals.
- Procedure is example of parametric bootstrap.

Nonconstant Error Variance (1)

- One of assumptions of linear model is that variation of response around regression line/surface - error variance - is constant: $V(\epsilon) = V(Y|x_1 \ldots, x_k) = \sigma^2_\epsilon$.
- Heteroscedasticity = nonconstant error variance.
  Homoscedasticity = constant error variance.
- Least-squares estimator remains unbiased and consistent even with nonconstant variance, but efficiency is impaired (we can do better) and usual formulas for standard errors are inaccurate.
- Graphical method for detection of nonconstant error variances: Residual plots
  - Plot residuals E_i against fitted values $\hat{Y}_i$: especially useful if error variance increases with expectation of Y.
  - Plot residuals against each X: especially useful if error variance changes with X.
  - R's default residual diagnostic `plot()` function plots $\sqrt{|E_i|}$ against fitted values $\hat{Y}_i$.
- Do not plot E_i against response itself Y_i, because plot is "tilted": they are not uncorrelated, $\text{corr}(E_i, Y_i) = \sqrt{1 - R^2}$.
- Correlation between residuals and fitted values is 0.

Nonconstant Error Variance (2)

- Best plot externally studentized residuals E^*_i versus fitted values, because ordinary residuals have unequal variances, even with constant error variance.
- Pattern of changing spread more easily seen by plotting $|E^*_i|$, or squared studentized residuals E^{*2}_i against $\hat{Y}$.
- Plot log-spread ($log(|E^*_i|)$) against log-level ($log(\hat{Y})$). Line through cloud of points with slope b suggests variance stabilizing transformation $1 - b$.
- Below residual plot and log(spread-level) plot; further down after log-transformation.



Nonconstant Error Variance (3)

- Log-transformation of wages made distribution of studentized residuals more symmetrical.
- Now same transformation stabilized residual variance.
- Often goes together, because heavy right tail of residual distribution and nonconstant spread are both consequences of lower bound 0 of response variable.
- Transforming Y changes shape of error distribution, but also alters shape of regression of Y on X.
- Sometimes same transformation that stabilizes variance and normalizes distribution, also makes relationship more linear; but this is not necessarily case. Check nonlinearity also after transformation!
- Nonconstant residual spread can be caused by omission of important variable from model.

Weighted Least Squares (1)

- WLS is alternative approach to estimation in presence of nonconstant error variance.
- Earlier WLS was presented as a special case of GLS (General Least Squares).
- Errors have different variances: $\epsilon_i \sim N(0, \sigma^2_i)$, but variances are known up to constant of proportionality σ^2_ϵ: $\sigma^2_i = \sigma^2_\epsilon / w^2_i$.
- Covariance matrix of errors $\Sigma = \sigma^2_\epsilon \times \text{diag}(1/w^2_1, \ldots, 1/w^2_n) \equiv \sigma^2_\epsilon \times \mathbf{W}^{-1}$.
- M.I. estimators are $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}$ and $\hat{\sigma}^2_\epsilon = \frac{\sum(w_i E_i)^2}{n}$.
- $V(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2_\epsilon(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}$.
- E.g. if known that standard deviation is proportional to X_1, use $1/X_{i1}$ as weights w_i.

Weighted Least Squares (2)

- Instead of WLS, ordinary least squares (OLS) can be applied by multiplying left-hand and right-hand size of regression equation by w_i.
- E.g. if standard deviation is proportional tot X_1, weights $w_i = 1/X_{i1}$:

$$\frac{Y_i}{X_{i1}} = \alpha\frac{1}{X_{i1}} + \beta_1 + \beta_2\frac{X_{i2}}{X_{i1}} + \ldots + \beta_k\frac{X_{ik}}{X_{i1}} + \frac{\epsilon_i}{X_{i1}}$$

  Slope β_1 of X_1 in original model, becomes the intercept in this regression model.

Correcting OLS Standard errors for Nonconstant Variance

- Covariance matrix of OLS estimator is $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V(\mathbf{y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, which simplifies to $V(\mathbf{b}) = \sigma^2_\epsilon(\mathbf{X}'\mathbf{X})^{-1}$.
- In case of heteroscedastic, but independent errors, (so diagonal covariance matrix $V(\mathbf{y}) = \Sigma$), then $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Sigma\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.
- Because $E(\epsilon_i) = 0$, variance of ϵ_i is $\sigma^2_i = E(\epsilon^2_i)$. This suggests as estimator of σ^2_i the squared residual E^2_i, and Σ estimated by $\hat{\mathbf{\Sigma}} = \text{diag}(E^2_1, \ldots, E^2_n)$.
- This brings the "sandwich estimator of variance" or White-adjusted standard errors:

$$\tilde{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

- Topic for case study project.

How Nonconstant Error Variance Affects OLS Estimator

- Harm produced by heteroscedasticity is relatively mild.
- Only when standard deviation of the errors varies by more than a factor 3 (largest variance 10 times smallest variance). A more conservative rule of thumb: worry if largest variance is 4 times smallest variance.
- Topic for case study project: study effects of ignored heteroscedasticity.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 5, lecture 26

Content lecture 25: Fox §12.3-12.4

- Graphical methods to check assumptions in LM (cont.)
  - non-normality; QQ-plots with/without envelope
  - heteroscedasticity; plot (studentized) residuals or sqrt(abs(residuals)) vs fitted values; weighted least squares
  - non-linearity; residuals vs regressors; component-plus-residual plot
- Testing methods to check assumptions in LM:
  - test for lack of fit with discrete regressor values (anova vs regression)
  - Levene's test for constant variance
  - m.l. method for optimal Box-Cox transformation

Nonlinearity

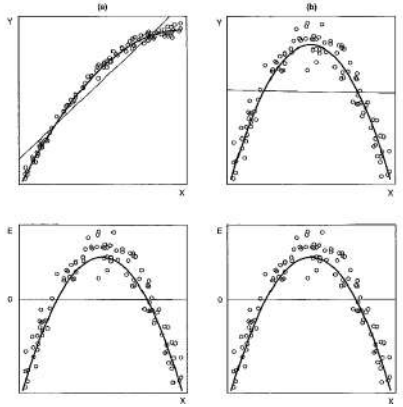
- Assumption that average error $E(\epsilon) = 0$ implies that regression surface accurately reflects dependency of conditional average value of Y on Xs.
- Conversely, violating assumption of linearity implies model fails to capture systematic pattern of relationship between response and explanatory variables.
- Term nonlinearity used in broad sense; it does not necessarily, but may mean that assumed linear relationship is nonlinear. E.g. if additive model is specified, but two regressors interact, then average error is not 0 for all combinations of X values, giving nonlinearity in broader sense.
- Regression surface is generally high dimensional.
- Therefore, focus on particular patterns of departure of linearity.
- Graphical diagnostics: two- (and three-) dimensional projections of (k+1) dimensional point cloud of observations $\{Y_i, X_{i1}, \ldots, X_{ij}\}$.

Component-Plus-Residual Plots (1)

- To check systematic part of model, plotting Y against each X is useful.
- However, these plots do not tell whole story: they look at marginal relationship of Y and X ("ignoring" other Xs), and not at partial relationship ("correcting" for other Xs), that we are mainly interested in.
- Therefore, residual based plots are more promising.
- Plotting (studentized) residuals against Xs may help (and we did this already to check for heteroscedasticity), but drawback is that they do not distinguish between monotone and non-monotone nonlinearity, see figure next slide.
- This is not unimportant, because monotone nonlinearity can often be "corrected" by simple transformation.

Component-Plus-Residual Plots (2)

- Bottom two residual plots are identical, but they come from two different Y-X relationships:
  - left plot: "monotone" nonlinear relationship between Y and X; simple transformation $\sqrt{\cdot}$ does job: replacing βX by $\beta\sqrt{X}$ transforms away this nonlinear relationship.
  - right plot: non-monotone non-linear relationship between Y and X. Now βX needs to be replaced by by $\beta_1 X + \beta_2 X^2$) to capture the systematic trend.

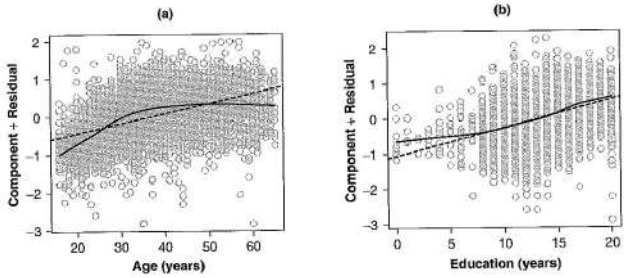


Component-Plus-Residual Plots (3)

- Added-variable plots, introduced for detection of influential data, can reveal nonlinearity and suggest monotonicity of relationship.
- However, these plots are less useful for establishing a transformation, because it is the unadjusted X_j itself that is transformed, not the X_j adjusted for the other Xs.
- Component-plus-residual plots (other name: partial residual plots) are alternative (though not useful for revealing leverage and influence):
  - Partial residual for jth regressor is $E_i^{(j)} = E_i + B_j X_{ij}$: add back linear component of partial relationship between Y and X_j to least-squares residual.
  - Next, plot $E^{(j)}$ versus X_j.
  - By construction, multiple regression coefficient B_j is slope of simple regression of $E^{(j)}$ on X_j.
  - Nonlinearity may be apparent in plot.

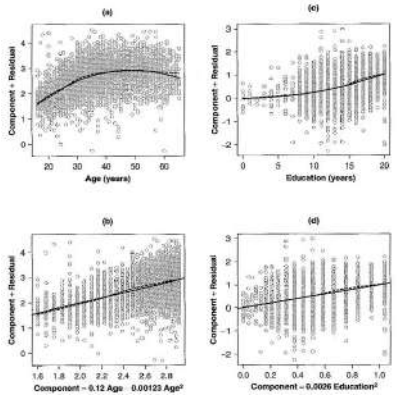
Component-Plus-Residual Plots (4)

- Example SLID (solid lines lowess smooths, broken lines least-squares fits):



Component-Plus-Residual Plots (5)

- Trial and error suggested linear and quadratic specification for age, and quadratic of education.
- Next, check component-plus-residual plots for new fit:
  - Partial residuals are:
    $E_i^{(Age)} = 0.1198 \times Age_i - 0.00123 \times Age^2 + E_i$
    $E_i^{(Education)} = 0.00265 \times Education_i^2 + E_i$.
    Plot these against original Xs
  - Plot these also against partial fits $0.1198 \times Age_i - 0.00123 \times Age^2$ and $0.00265 \times Education_i^2$.



- Fox extends analysis by allowing interaction of age and education with gender. Interpretation of results is supported by effect displays (plots of fitted values for one regressor, keeping others at average values). Component-plus-residual plots are made for men and women separately, plotting partial residual against partial fit.

When do component-plus-residual plots work?

- Lower dimensional displays cannot always uncover what happens in higher dimensions.
- With component-plus-residual plots we would like to estimate the partial relationship between Y and (say) X_1, let's call it $f(X_1)$.
- This will work if one of the following two conditions hold:
  - function $f(X_1)$ is linear after all;
  - Other explanatory variables $X_2, \ldots, X_k$ are each linearly related to X_1. Nonlinear relationships between other Xs and X_1 can cause component-plus-residual plot for X_1 not to reflect the true partial regression $f(X_1)$.
- We skip further details.
- Summary: Simple forms of nonlinearity can often be detected in component-plus-residual plots. Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an explanatory variable, for example).
  Component-plus-residual plots reliably reflect nonlinearity when there are not strong nonlinear relationships among the explanatory variables in a regression.

Discrete Data

- Plots made for discrete explanatory and response variables may be difficult to interpret, because multiple observations with same values are hidden. Jittering may help.
- Discrete response variable violates assumption that errors in linear model are normally distributed. But this problem is only serious in extreme cases, e.g. if there are only a few response categories. In that case statistical models for categorical response data are preferred.
- Discrete explanatory variables pose no problem in linear models. They partition data into groups, facilitating tests of nonlinearity and nonconstant error variance.
- Testing for Lack of Fit. This topic was studied earlier, based on Faraway. Remember comparison of simple linear regression with one-way ANOVA by incremental F-test. Fox extends the procedure for multiple linear regression situation, by inclusion of dummies for all levels (but one) of each discrete regressor separately, and jointly. Duncan's data are studied with lack of fit tests for age and for education.

Discrete Data: Testing for Nonconstant Error Variance

- Discrete X partitions data into (say) m groups; Y_{ij} is ith observation in group j.
- If error variance is constant across groups, then within group sample variances $S_j^2 = \frac{\sum_{i=1}^{n_j}(Y_{ij} - \overline{Y}_j)^2}{n_j - 1}$ should be similar.
- Traditional test to compare variances is Bartlett's test, but it is notoriously sensitive to nonnormality.
- More robust procedure is Levene's test:
  - Calculate $Z_{ij} \equiv |Y_{ij} - \tilde{Y}_j|$ with $\tilde{Y}_j$ median of Y_{ij} in group j.
  - Next, perform one-way ANOVA of Z_{ij} over m groups. If error variance is not constant across groups, then group means $\overline{Z}_j$ will tend to differ, producing large values of F-statistic.
- Package `car` has function `leveneTest`.

```
> with(Moore, leveneTest(conformity, fcategory))
Levene Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2    0.05   0.96
      42
```


Maximum Likelihood Methods for Transformation (1)

- Sophisticated approach to find transformation of Y or X is to embed usual linear model in more general nonlinear model that contains a parameter for the transformation.
- Suppose transformation is indexed by single parameter λ (e.g. power transformation $Y \to Y^\lambda$, and likelihood for model can be written down as function of λ and usual regression parameters: $L(\lambda, \alpha, \beta_1, \ldots, \beta_k, \sigma_\epsilon^2)$.
- By maximizing likelihood m.l. estimates of λ and other parameters are obtained.
- Likelihood ratio tests can be used, comparing maximized likelihood with null-likelihood based on $H_0 : \lambda = \lambda_0$, e.g. $\lambda_0 = 1$, i.e. no transformation.
- Alternatively, Wald tests or score tests may be done.

Maximum Likelihood Methods for Transformation (2)

- These ideas can be used to find the "optimal" Box-Cox transformation of Y, based on model $Y_i^{(\lambda)} = \alpha + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \epsilon_i$
  with $Y_i^{(\lambda)} = \begin{cases} \frac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$
- We are not going into details, during the practical we follow the example described in the Faraway text (pp 96-98).
- Box-Tidwell Transformation of Xs is another possibility for case study.

Skipped topics

- We skip section 12.5.3 about score tests proposed by Breusch-Pagan, Cook and Weisberg, and White to test for heteroscedasticity.
  Candidate for case study.
- We skip section 12.6 on Structural Dimension.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 5, lecture 27

Content lecture 27: Fox §13; Faraway §9.2+9.4

- Collinearity
- Principal components

Collinearity and Remedies

- If perfect linear relationship among regressors exist, least-squares coefficient are no longer uniquely defined.
- Strong, but less-than-perfect linear relationship among Xs causes least-squares coefficients to be unstable: large standard errors of coefficients, broad confidence intervals, hypothesis tests with low power.
  Small changes in data (in extreme cases even due to rounding errors) can greatly change coefficients. Large changes in coefficients coincide with only very small changes in residual sum of squares.
- Fox makes 3 starting remarks:
  1. Collinearity is relatively rare problem in social-science applications of linear models. More frequent problems are insufficient variation in regressors, small samples, large error variance.
  2. Methods employed as cures for collinearity (biased estimation, variable selection) may be worse than the disease.
  3. Not obvious that detection of collinearity in data has practical implications. Usually impossible to redesign study to decrease correlations between Xs.

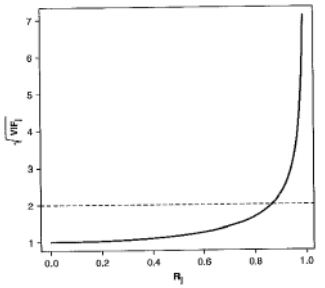
Detecting Collinearity

- Suppose perfect linear relationship exists between Xs: $c_1X_{i1} + c_2X_{i2} + \ldots + c_kX_{ik} = c_0$. Then:
  - least squares normal equations do not have unique solution;
  - sampling variances of regression coefficients are infinite,

  because $\mathbf{X}'\mathbf{X}$ matrix is singular.
  Perfect collinearity is often product of some error in formulating linear model, like too many dummies.
- Suppose linear relationship is less than perfect. Then:
  - Sampling variance of slope B_j : $V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\epsilon^2}{(n-1)S_j^2}$ with R_j^2 squared multiple correlation for regression of X_j on other Xs, and $S_j^2 = \sum(X_{ij} - \overline{X}_j)^2/(n-1)$ the variance of X_j.
  - Variance Inflation Factor $VIF = \frac{1}{1-R_j^2}$.
  - VIF indicates directly impact of collinearity on precision of B_j.
  - Fox works with $\sqrt{VIF}$, because it relates to standard error of B_j and hence confidence interval.

Detecting Collinearity (2)

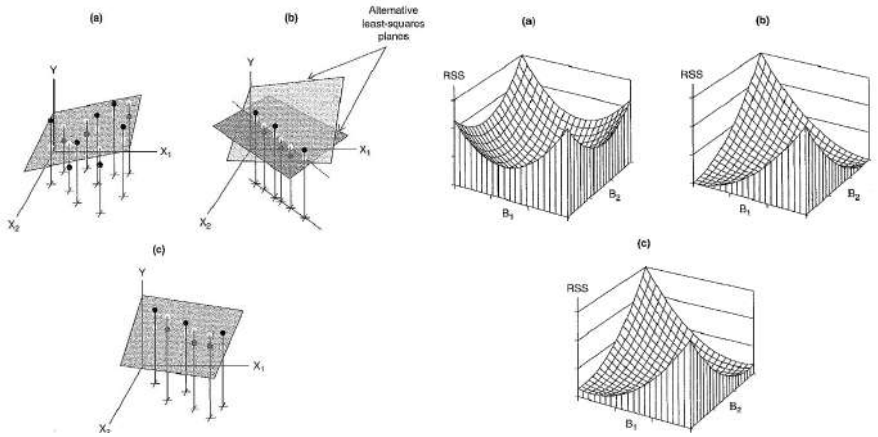
- Figure shows that linear relationship between variables must be quite strong before collinearity seriously impairs precision of estimation: precision of estimation is halved if $R_j \approx 0.9$.



- VIF is basic diagnostic for collinearity.
- Not applicable to sets of related regressors, like dummies.
- Collinearity also called multicollinearity.
- If X_1 strongly collinear with other Xs, residuals $X^{(1)}$ from regression on other Xs show little variation. Added-variable plot has $X^{(1)}$ on x-axis, which is nearly invariant. Consequently, slope B_1 is subject to substantial sampling variation.
- Confidence intervals for individual coefficients are projections of ellipses. These confidence intervals will be wide. But considerable information about sum of regression coefficients exists, if Xs are positively correlated.

Detecting Collinearity (3)

- Suppose 2 explanatory variables in regression.
- Examples of no collinearity, complete collinearity and strong collinearity on stability of least-squares plane, and on residual sum of squares as function of slopes B_1 and B_2.

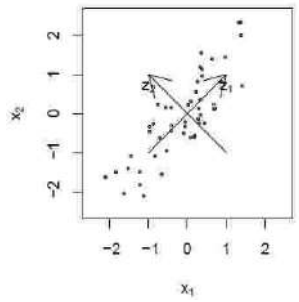


Principal Components (1)

- Fox describes in quite some detail Principal Components, which I find too much at this moment.
- Instead we follow Faraway text pages 107-113. We follow the red line, and you can try the R-code yourself during the practical.
- Principal Components is a data-analytic technique, with which we try to summarize the (maybe many, and maybe collinear) regressors into (only) a few principal components.
- It has no underlying (linear) model (in contrast to the related Factor Analysis). It is a data-summarizing technique, that can be used in any situation.
- The principal components themselves are linear combinations of the variables (regressors in our case), and they are orthogonal. Orthogonality is a nice property, simplifying testing and interpretation. Interpretation of the linear combinations of variables may be problematic, though.

Principal Components (2)

- Figure below (figure 9.1 in Faraway) gives the idea. Suppose we have 2 predictors x_1 and x_2, which are highly correlated. The X matrix is not orthogonal. Now we try to rotate the coordinate axes so that in the new system the predictors are orthogonal. Furthermore, the rotation is such, that first axis lies in direction of greatest variation in data, second in second greatest direction of variation, and so on. Rotated directions z_1 and z_2 (in example) are linear combinations of original predictors. We replace two regressors by two pc's. No information is lost.



Principal Components (3)

Construction of principal components in matrix algebra:

- Find $p \times p$ rotation matrix $\mathbf{U}$ such that $\mathbf{Z} = \mathbf{XU}$ and $\mathbf{Z}'\mathbf{Z} = \mathbf{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. [A rotation matrix is an orthonormal matrix, so $\mathbf{U}'\mathbf{U} = \mathbf{I}_p$ and $\mathbf{U}' = \mathbf{U}^{-1}$.]
- Hence, the columns of $\mathbf{Z}$ are orthogonal, and have squared length λ_i.
- $\mathbf{Z}'\mathbf{Z} = (\mathbf{XU})'\mathbf{XU} = \mathbf{U}'\mathbf{X}'\mathbf{XU}$. As $\mathbf{Z}'\mathbf{Z}$ is the mentioned diagonal matrix $\mathbf{\Lambda}$, we have $\mathbf{U}'\mathbf{X}'\mathbf{XU} = \mathbf{\Lambda}$, so $\mathbf{UU}'\mathbf{X}'\mathbf{XU} = \mathbf{U\Lambda}$, leading to $\mathbf{X}'\mathbf{XU} = \mathbf{U\Lambda}$ (because $\mathbf{U}' = \mathbf{U}^{-1}$ for a rotation matrix, hence $\mathbf{UU}' = \mathbf{I}_p$). Here you see the definition of eigenvalues and eigenvectors: the diagonal elements of $\mathbf{\Lambda}$: $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ must be the eigenvalues of $\mathbf{X}'\mathbf{X}$, and the columns of $\mathbf{U}$ are the eigenvectors of $\mathbf{X}'\mathbf{X}$.
- Columns of $\mathbf{Z}$ are the principal components, i.e. the linear combinations of the original variables; principal components are orthogonal to each other.
- λ_i is the variance of $\mathbf{Z}_i$; λ_1, corresponding to the first column of $\mathbf{Z}$, is largest, telling that the first principal component has largest variance (direction of largest variance); λ_2 second largest, etc.
- Zero eigenvalues indicate non-identifiability, which would occur if we have perfect collinearity.

Principal Components (4)

- Another approach is to find linear combinations of columns of regressors which have maximum variation. Find $\mathbf{u}_1$ such that $var(\mathbf{u}_1 X)$ is maximized, subject to $\mathbf{u}_1'\mathbf{u}_1 = 1$. Next find $\mathbf{u}_2$ which maximizes $var(\mathbf{u}_2 X)$ subject to $\mathbf{u}_2'\mathbf{u}_2 = 1$ and $\mathbf{u}_1'\mathbf{u}_2 = 0$. Keep finding directions of greatest variation orthogonal to those already found.

Concluding remarks:

- Principal components are linear combinations of predictors. Little is gained if these are not interpretable. Generally predictors have to be measurements of comparable quantities for interpretation to be possible.
- Principal components do not use the response y. Therefore, it is possible that not the first, but may be the second or third principle component (or none whatsoever) is important in predicting response y.
- Instead of using $\mathbf{X}'\mathbf{X}$, correlation matrix of predictors may be used.

Collinearity revisited, Faraway

- If $\mathbf{X}'\mathbf{X}$ is singular, i.e. some regressors are linear combinations of others, we have exact collinearity. Then no unique least-squares estimator of β exists.
- Detect collinearity by e.g.
  - Examination of correlation matrix of predictors: large pairwise collinearities may present themselves.
  - Regress X_i on all other Xs, giving R_i^2, and repeat for all predictors. R_i^2 close to one indicates problem.
  - Examine eigenvalues of $\mathbf{X}'\mathbf{X}$: small eigenvalues indicate problem.

    Condition number is $\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}}$. Condition number $\kappa > 30$ is considered large.

    Other condition numbers, $\sqrt{\lambda_1/\lambda_i}$ may also be considered, because they may indicate whether there is more than one linear combination to blame.
  - Variance Inflation Factor $VIF = \frac{1}{1-R_j^2}$.

Collinearity revisited, Faraway (2)

- Variance Inflation Factor $VIF = \frac{1}{1-R_j^2}$.

  We earlier saw the variance of coefficient $V(B_j) = \frac{1}{1-R_j^2} \times \frac{\sigma_\epsilon^2}{(n-1)S_j^2}$.

  Notice that variance of coefficient will be large if
  - if regressor X_j hardly varies, because then S_j will be small. It would be good to maximize S_j, by spreading observations as much as possible. This has consequences for (experimental) design.
  - if collinearity exists, because then $VIF = 1/(1-R_j^2)$ will be large.
  - if error variance is large.
- Collinearity leads to
  - imprecise estimates of β; even signs of coefficients may be misleading.
  - t-tests fail to reveal significant factors.
  - missing importance of predictors.

Coping With Collinearity: No Quick Fix (Fox)

- Model Respecification: although collinearity is data problem, not necessarily deficiency of model, sometimes respecifying model may help: e.g. combine highly correlated regressors into single predictor.
- Variable Selection: to be discussed next time.
- Biased Estimation: e.g. Ridge Regression. Many modern regression techniques exist, which combine ideas from variable selection and biased estimation.
- Prior Info About Regression Coefficients. Think e.g. of Bayesian approach.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 5, lecture 28

Content week 5, lecture 28: Faraway text §8.2-8.4

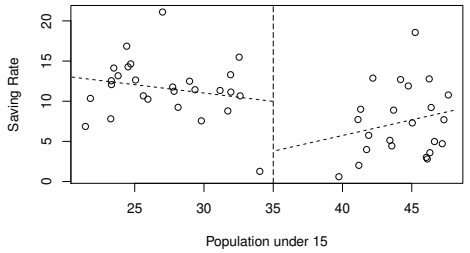
More flexible functional relationship between y and x:

- Broken stick regression
- Polynomial regression, orthogonal polynomials
- Splines (introduction)

Broken Stick Regression

- We saw earlier that predictors may need transformation to get the systematic part of the model right. Partial regression plots or, more specifically, component-plus-residual plots may be helpful.
- In some cases different linear regression models apply in different regions of predictor.
- Example: `savings` data, focusing on `pop15` predictor. We fit 2 simple linear regressions, depending on whether `pop15` greater or smaller than 35%.

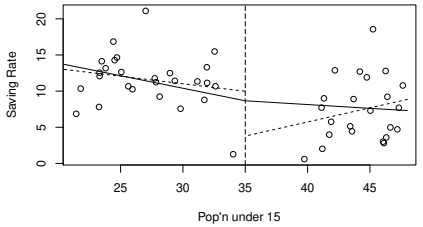
```
> g1 <- lm(sr ~ pop15, savings, subset=(pop15<35))
> g2 <- lm(sr ~ pop15, savings, subset=(pop15>35))
> plot(savings$pop15, savings$sr, xlab="Population under 15", ylab="Saving Rate")
> abline(v=35, lty=5)
> segments(20, g1$coef[1]+g1$coef[2]*20, 35, g1$coef[1]+g1$coef[2]*35, lty=2)
> segments(48, g2$coef[1]+g2$coef[2]*48, 35, g2$coef[1]+g2$coef[2]*35, lty=2)
```



Broken Stick Regression (2)

- Notice that two parts of fit do not meet at join, which is undesirable.
- Solution: broken stick regression, or more generally called segmented regression or piecewise regression.
- Define two basis functions:
  $B_l(x) = \begin{cases} c - x & \text{if } x < c \\ 0 & \text{otherwise} \end{cases}$ and $B_r(x) = \begin{cases} x - c & \text{if } x > c \\ 0 & \text{otherwise} \end{cases}$
  with c the predictor value where division occurs.
- Next, fit linear regression model $Y = \beta_0 + \beta_1 B_l(X) + \beta_2 B_r(X) + \epsilon$.
- Two regression lines are guaranteed to meet at c (why?). Regression function is continuous now.
- Model contains two slopes, and single intercept, so 3 instead of original 4 parameters.

```
> lhs <- function(x) {ifelse(x<35,35-x,0)}
> rhs <- function(x) {ifelse(x<35,0,x-35)}
> gb <- lm(sr ~ lhs(pop15) + rhs(pop15), savings)
> x <- seq(20,48,by=1)
> predy <- gb$coef[1] +
+   gb$coef[2]*lhs(x) + gb$coef[3]*rhs(x)
```



Polynomial regression

- Another way of generalizing $\mathbf{X}\beta$ part of model is to add polynomial terms.
- One-predictor case: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_d X^d + \epsilon$ allowing for more flexible relationship, though generally not believed to be exactly true.

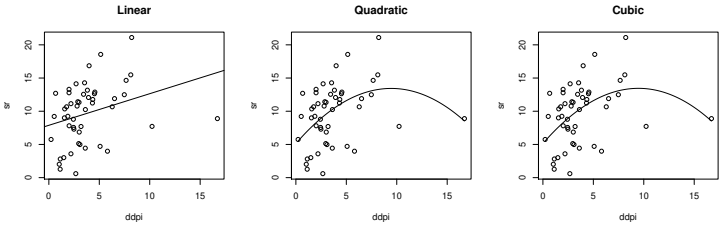
```
> g1<-lm(sr ~ ddpi,savings);
> #plot(sr ~ ddpi, savings); abline(g1); title("Linear"); summary(g1)$coef
>
> g2<-lm(sr ~ ddpi + I(ddpi^2),savings)
> summary(g2)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.13038	1.43472	3.576	0.0008211
ddpi	1.75752	0.53772	3.268	0.0020259
I(ddpi^2)	-0.09299	0.03612	-2.574	0.0132617

```
> g3<-lm(sr ~ ddpi + I(ddpi^2) + I(ddpi^3),savings)
> summary(g3)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.145e+00	2.198606	2.340282	0.02366
ddpi	1.746e+00	1.380455	1.264812	0.21231
I(ddpi^2)	-9.097e-02	0.225598	-0.403227	0.68865
I(ddpi^3)	-8.497e-05	0.009374	-0.009064	0.99281

```
> x <- seq(min(savings$ddpi), max(savings$ddpi), length.out=100)
> predy2 <- coef(g2)[1]+coef(g2)[2]*x + coef(g2)[3]*x^2
> #plot(sr ~ ddpi, savings); lines(x, predy2); title("Quadratic")
```



Polynomial regression: respect marginality

- Principle of marginality: do not remove lower order terms from model, even if not statistically significant.
- Additive change in scale would change estimates and t-statistics for coefficients of all but highest order term. Hence, don't pay too much attention to lower order terms in presence of higher order terms: just leave them in model.
- Check this for `savings` data with quadratic regression as before, using `ddpi-10` as regressor. How do new coefficients relate to old coefficients?

```
> g2.2 <- lm(sr ~ I(ddpi-10) + I((ddpi-10)^2),savings)
> summary(g2.2)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	13.40705	1.42401	9.4150	2.160e-12
I(ddpi - 10)	-0.10219	0.30274	-0.3375	7.372e-01
I((ddpi - 10)^2)	-0.09299	0.03612	-2.5741	1.326e-02

Orthogonal polynomials

- Each time a term is removed from or added to model, coefficients change and model needs to be refitted.
- High order polynomial models may be numerically unstable.
- Orthogonal polynomials help: replace old set of predictors $X, X^2, X^3, \cdots$ by new, orthogonal, set of predictors $Z_1, Z_2, Z_3, \cdots$:
  $Z_1 = a_1 + b_1 X$
  $Z_2 = a_2 + b_2 X + c_2 X^2$
  $Z_3 = a_3 + b_3 X + c_3 X^2 + d_3 X^3$, etc.
  such that $Z_i^T Z_j = 0$ and $Z_i^T Z_i = 1$ (length 1).

```
> g2 <- lm(sr ~ poly(ddpi,2),savings)
> summary(g2)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.671	0.5769	16.765	1.841e-21
poly(ddpi, 2)1	9.559	4.0790	2.343	2.339e-02
poly(ddpi, 2)2	-10.500	4.0790	-2.574	1.326e-02

```
> g4<- lm(sr ~ poly(ddpi,4),savings)
> summary(g4)$coef
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.67100	0.5846	16.542880	9.477e-21
poly(ddpi, 4)1	9.55899	4.1338	2.312421	2.539e-02
poly(ddpi, 4)2	-10.49988	4.1338	-2.540030	1.461e-02
poly(ddpi, 4)3	-0.03737	4.1338	-0.009041	9.928e-01
poly(ddpi, 4)4	3.61197	4.1338	0.873773	3.869e-01

Check orthogonality

- Check that orthogonal polynomials are orthogonal to each other and have length 1:

```
> x <- model.matrix(g4)
> dimnames(x) <- list(NULL,c("Int","power1","power2","power3","power4"))
> round(t(x) %*% x,3)
```

	Int	power1	power2	power3	power4
Int	50	0	0	0	0
power1	0	1	0	0	0
power2	0	0	1	0	0
power3	0	0	0	1	0
power4	0	0	0	0	1

- You can have more than one (not necessarily orthogonal) polynomial predictor as can be seen in this quadratic response surface model:

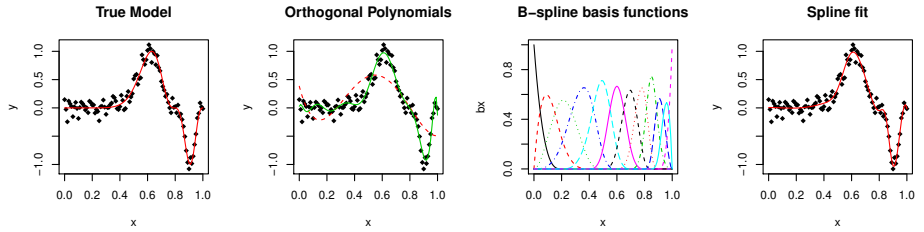
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

Splines

- Polynomials have advantage of smoothness, but disadvantage that each datapoint affects fit globally. Reason is that predictors have non-zero values across whole range of predictor.
- Segmented regression localizes influence of each datapoint to its particular segment, but lacks smoothness.
- Combine both (smoothness and local influence) by using B-spline basis functions.
- Define cubic B-spline basis on interval $[a, b]$ using knots at $t_1, \ldots, t_k$:
  1. Given basis function is non-zero on interval defined by four successive knots and zero elsewhere. This ensures local influence.
  2. Basis function is cubic polynomial for each sub-interval between successive knots.
  3. Basis function is continuous, with continuous first and second derivative at each knot. This ensures smoothness of fit.
  4. Basis function integrates to one over its support.
- Different types of splines exist, e.g. B-splines, natural splines, P-splines; details are beyond scope of the course.

Example

```
> funky <- function(x) sin(2 * pi * x^3)^3
> x <- seq(0, 1, by = 0.01)
> y <- funky(x) + 0.1 * rnorm(101)
> matplot(x, cbind(y, funky(x)), type = "pl", ylab = "y", pch = 18, lty = 1,
+     main = "True Model")
> g4 <- lm(y ~ poly(x, 4))
> g12 <- lm(y ~ poly(x, 12))
> matplot(x, cbind(y, g4$fit, g12$fit), type = "pll", ylab = "y", pch = 18,
+     lty = c(1, 2), main = "Orthogonal Polynomials")
> library(splines)
> knots <- c(0, 0, 0, 0, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 1, 1, 1,
+     1)
> bx <- splineDesign(knots, x)
> gs <- lm(y ~ bx)
> matplot(x, bx, type = "l", main = "B-spline basis functions")
> matplot(x, cbind(y, gs$fit), type = "pl", ylab = "y", pch = 18, lty = 1,
+     main = "Spline fit")
```



Modern regression methods

- Finding good transformations on several predictors simultaneously may be problematic.
- Nonparametric regression techniques, like the B-splines approach, way out.
- E.g. additive models, regression trees, MARS (=Multivariate Adaptive Regression Splines), neural networks.
- Fox chapter 18 describes some.
- Skip now.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 5, lecture 29

Content week 5, lecture 29: Fox: §22-skip some details; Far: §10.1-10.2

- Model selection
  - Problems with model selection using hypothesis testing
  - Stepwise versus criterion based procedures
  - Stepwise procedures: backward, forward, stepwise

Model Selection: Caution!

- Suppose we have many predictor variables, not all necessarily related to response. We want to select the "best" subset of predictors.
- Fox starts with cautionary remarks about model selection based on hypothesis tests:
  - Problem of simultaneous inference.
  - Failing to reject a null hypothesis is not same as demonstrating that null hypothesis is supported by data. Distinguish small, but precisely estimated coefficient (narrow c.i., but containing 0), from large, imprecisely estimated coefficient. Eliminating imprecisely estimated term can seriously bias other estimates if true population coefficient is large and variable is strongly related to others (see e.g. section 9.7).
  - Impact of large samples on hypothesis tests: trivially small effects become statistically significant if dataset is large.
  - Exaggerated precision: after model selection, coefficient standard errors tend to become too small when variables correlated to those staying in model are eliminated, leading to artifically narrow c.i.'s and small P-values.

General strategies

To address these concerns Fox suggests:

- Use alternative model-selection criteria instead of statistical significance.
- Compensate for simultaneous inference, e.g. by Bonferroni adjustments, or by holding back some data to validate a statistical model selected by another approach
- Avoid model selection. Specify maximally complex and flexible model without trying to simplify it. But: (1) often not possible to specify fully adequate model in advance, (2) retaining unimportant terms in model violates modelling principle of parsimony.
- Model averaging: don't select single model discarding all others, integrate information from different models using weights according to degree of model support from data.

Model Selection: motivating remarks from Faraway

- Reasons for selecting "best" subset of regressors:
  - Explain data in simplest way, removing redundant predictors. Principle of Occam's Razor (parsimony) states that among several plausible explanations for phenomenon, simplest is best.
  - Unnecessary predictors will add noise to estimation of other quantities that are of interest, degrees of freedom are wasted. More precise estimates and predictions may be achieved with smaller model.
  - Collinearity is caused by having too many variables doing same job. Remove excess predictors.
  - If the model is to be used for prediction, save time and money by not measuring redundant predictors.
- Variable selection is part of process of model building, like identification of outliers and influential points, and variable transformation.

Types of variable selection

- Two main types of variable selection:
  1. Stepwise approach, comparing successive models
     - Stepwise approach may use hypothesis testing to select the next step, but other criteria may be used too (e.g. R's function `stepAIC` uses AIC criterion)
  2. Criterion approach, finding model that optimizes some measure of goodness of fit.
- Don't forget marginality principle: for polynomial models, keep lower order terms in model, if higher order term is important.
- Model selection is conceptually simplest if prediction is goal: develop regression model that will predict new data as accurately as possible.
- Computer power and availability of huge datasets (e.g. in genomics) facilitates model selection as point of focus in research.
- Datamining is related topic.

Stepwise procedures

If hypothesis testing is used:

- Backward Elimination
  1. Start with all predictors in model.
  2. Remove predictor with highest p-value greater than threshold p-value-to-stay α_{crit}.
  3. Refit model and goto 2)
  4. Stop if all p-values of terms remaining in the model are smaller than α_{crit}.
- Forward Selection
  1. Start with no predictor in model.
  2. Enter predictor with smallest p-value, smaller than threshold p-value-to-enter α_{crit}.
  3. Refit model and goto 2)
  4. Stop if all p-values of terms not in the model are higher than α_{crit}.
- Stepwise Regression
  This is combination of backward elimination and forward selection. After entering variable, all variables in model are candidate for removal. Thresholds p-value-to-enter and p-value-to-stay need to be specified.

Stepwise procedures may also be used in combination with other criteria, e.g. AIC.

Drawback stepwise procedures

- Drawback related to earlier mentioned caveats:
  - "Optimal" model may be missed due to adding / dropping of single variables.
  - If using p-values: don't treat p-values literally! Recall multiple testing problem, but also overstating importance of remaining predictors after removal of less significant ones.
  - Procedures are not linked to final objective of prediction or explanation. Model selection should not be divorced from final purpose of research.
  - Stepwise selection tends to pick models smaller than desirable for prediction purposes.

Example

Backward elimination "by hand", p-value-to-stay is 0.05.

```
> g <- lm(Life.Exp ~ ., data=statedata)
> coef(summary(g))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.09e+01   1.75e+00 40.5859 2.51e-35
Population   5.18e-05   2.92e-05  1.7748 8.32e-02
Income      -2.18e-05   2.44e-04 -0.0892 9.29e-01
Illiteracy   3.38e-02   3.66e-01  0.0923 9.27e-01
Murder      -3.01e-01   4.66e-02 -6.4590 8.68e-08
HS.Grad      4.89e-02   2.33e-02  2.0979 4.20e-02
Frost       -5.74e-03   3.14e-03 -1.8246 7.52e-02
Area        -7.38e-08   1.67e-06 -0.0443 9.65e-01
> summary(g)$r.squared
[1] 0.736
> g <- update(g, . ~ . -Area)
> coef(summary(g))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   1.39e+00 51.1652 3.69e-40
Population   5.19e-05   2.88e-05  1.8023 7.85e-02
Income      -2.44e-05   2.34e-04 -0.1043 9.17e-01
Illiteracy   2.85e-02   3.42e-01  0.0833 9.34e-01
Murder      -3.02e-01   4.33e-02 -6.9634 1.45e-08
HS.Grad      4.85e-02   2.07e-02  2.3454 2.37e-02
Frost       -5.78e-03   2.97e-03 -1.9446 5.84e-02
>
```

```
> g <- update(g, . ~ . -Illiteracy)
> coef(summary(g))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.11e+01   1.03e+00  69.067 1.66e-46
Population   5.11e-05   2.71e-05   1.888 6.57e-02
Income      -2.48e-05   2.32e-04  -0.107 9.15e-01
Murder      -3.00e-01   3.70e-02  -8.099 2.91e-10
HS.Grad      4.78e-02   1.86e-02   2.569 1.37e-02
Frost       -5.91e-03   2.47e-03  -2.395 2.10e-02
> g <- update(g, . ~ . -Income)
> coef(summary(g))
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.10e+01   9.53e-01   74.54 8.61e-49
Population   5.01e-05   2.51e-05    2.00 5.20e-02
Murder      -3.00e-01   3.66e-02   -8.20 1.77e-10
HS.Grad      4.66e-02   1.48e-02    3.14 2.97e-03
Frost       -5.94e-03   2.42e-03   -2.46 1.80e-02
> g <- update(g, . ~ . -Population)
> coef(summary(g))
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 71.03638    0.98326   72.25 5.25e-49
Murder      -0.28307    0.03673   -7.71 8.04e-10
HS.Grad      0.04995    0.01520    3.29 1.95e-03
Frost       -0.00691    0.00245   -2.82 6.99e-03
> summary(g)$r.squared
[1] 0.713
```


Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 5, lecture 30

Content week 5, lecture 30: §22.1.1 until p 675; Far text: §10.3-10.4)

- Criteria for model selection:
  - R^2_{adj}
  - Mallow's C_p
  - *PRESS*
  - *AIC* and *BIC*
- Model validation; cross validation

Criterion-Based Procedures

- Criterion-based procedures typically compare all possible models ("All possible subsets regression"), or circumvent checking all models by clever algorithms.
- Checking all possible subsets can be extremely laborious. Model with k regressors has 2^k possible submodels!
- Different criteria may be used, e.g.:
  - R^2_{adj}
  - Mallow's C_p
  - *PRESS*
  - *AIC* and *BIC*

Criterion-Based Procedure: adjusted R^2

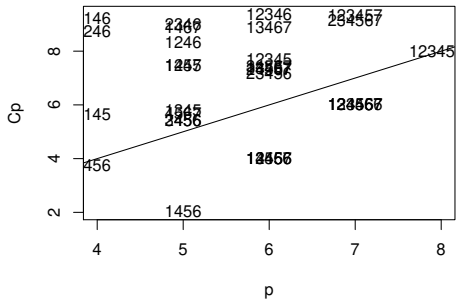
- R^2 cannot be used as criterion for model selection, because it would always choose the largest possible model (why?).
  Recall $R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}$.
- $R^2_{adj} = 1 - \dfrac{RSS/(n-p)}{TSS/(n-1)} = 1 - \dfrac{\hat{\sigma}^2_{model}}{\hat{\sigma}^2_{null}}$.
- Notice that $\hat{\sigma}^2_{null}$ does not depend on selected model: it is the estimate of error variance based on "empty" model, i.e. model containing intercept only.
- R^2_{adj} will only increase by changing a model, if the estimate of error variance based on new model $\hat{\sigma}^2_{model}$ decreases.
- Estimate of error variance will only decrease, if "change" in residual sum of squares is compensated by change in residual d.f.

Criterion-Based Procedure: Mallow's C_p

- Good model should predict well, so total *MSE* (Mean-Squared Error) of prediction should be small. We predict true $E(Y_i)$ with $\hat{Y}_i$.
- *MSE* of prediction is population parameter. Normed version is:
  $$\frac{1}{\sigma^2_\epsilon}\sum_{i=1}^{n} E(\hat{Y}_i - E(Y_i))^2 = \frac{1}{\sigma^2_\epsilon}\sum_{i=1}^{n}\left\{V\left(\hat{Y}_i\right) + \left[E\left(\hat{Y}_i\right) - E(Y_i)\right]^2\right\}$$
- Fitted values $\hat{Y}_i$ come from current model, which contains p coefficients.
- Two components: variance $V\left(\hat{Y}_i\right)$ and squared bias $\left[E\left(\hat{Y}_i\right) - E(Y_i)\right]^2$.
- So, prediction *MSE* contains sum of variance and squared bias.
- Question is, whether by removing a variable, the decrease in variance offsets any increase in bias.
- Prediction MSE is estimated by Mallow's C_p: $C_p = \dfrac{RSS_p}{\hat{\sigma}^2_\epsilon} + 2p - n$, with $\hat{\sigma}^2_\epsilon$ from full model, and RSS_p from current model.
- Good model should have C_p close to or below p. Model with bad fit has C_p much bigger than p.
- For full model, $C_p = p$ (check).
- Usually C_p is plotted against p. Look for models with small p and with C_p around or less than p.

Example Mallow's C_p

```
> library(leaps)
> g <- lm(Life.Exp ~., data=statedata)
> x <- model.matrix(g)[,-1]
> y <- statedata$Life
> g <- leaps(x,y)
> Cpplot(g)
```



- Seven predictors, giving $2^7 = 128$ different models.
- Only best are plotted.
- Competition between models "456" and "1456": both are on or below $C_p = p$ line, indicating good fits. Smaller model is more parsimonious, but larger models fits slightly better. Even some larger models are on or below $C_p = p$ line, but are not chosen in presence of smaller models that fit.

Example R^2_{adj}

```
> adjr <- leaps(x,y,method="adjr2")
> maxadjr(adjr,8)
    1,4,5,6   1,2,4,5,6   1,3,4,5,6   1,4,5,6,7 1,2,3,4,5,6 1,3,4,5,6,7 1,2,4,5,6,7
      0.713       0.706       0.706       0.706       0.699       0.699       0.699
    4,5,6
    0.694
```

- Model with largest R^2_{adj} is "1456".
- Best 3-predictor model is"456", on 8th place.
- Variable selection methods sensitive to outliers, influential points, and transformation.

Criterion-Based Procedure: *PRESS* and cross-validation criterion *CV*

- *PRESS* = Predicted REsidual Sum of Squares $= \sum_{i=1}^{n} \left(\hat{Y}_{-i} - Y_i\right)^2$, i.e. sum of squared deleted residuals, or cross-validated residuals.
- Example of Leave-one-out cross-validation: fit model n times, omitting ith observation at step i, and use fitted model to predict omitted observation: $\hat{Y}_{-i}$.
- Cross-validation criterion estimates mean-squared error of prediction as $CV \equiv \dfrac{\sum_{i=1}^{n} \left(\hat{Y}_{-i} - Y_i\right)^2}{n} = PRESS/n$.
- Prefer model with smallest value of *CV* or *PRESS*.
- In linear models, leave-one-out fitted values $\hat{Y}_{-i}$ can be computed without refitting model n times.
- Other cases, refitting is needed, and becomes computationally expensive.
- Alternative is p-fold cross validation: divide data into small number of subsets (e.g. 10) of roughly equal size, fit model omitting each subset in turn, obtaining fitted values for all observations in omitted subset.
- Generalized cross-validation criterion is approximation of *CV*: $GCV \equiv \dfrac{n \times RSS}{(n-p)^2}$

Criterion-Based Procedure: *AIC* and *BIC* (1)

- *AIC*=Akaike Information Criterion
- *BIC*=Bayesian Information Criterion = Schwarz's Bayesian Criterion
- Both belong to family of penalized model-fit statistics: $-2logL(\hat{\boldsymbol{\theta}}) +$ penalty:
  - $L(\hat{\boldsymbol{\theta}})$ is maximized likelihood under current model, and $\boldsymbol{\theta}$ vector of parameters of model, including regression coefficients and error variance, and $\hat{\boldsymbol{\theta}}$ is m.l.e.
  - penalty $= cp$ with p number of parameters in model.
  - Magnitude of criterion is not interpretable, but differences are.
  - Model with smallest value of information criterion is preferred.
- $AIC \equiv -2logL(\hat{\boldsymbol{\theta}}) + 2p$
  $BIC \equiv -2logL(\hat{\boldsymbol{\theta}}) + log(n)\ p$

Criterion-Based Procedure: *AIC* and *BIC* (2)

- Linear model with normal errors (check): $-2logL(\hat{\boldsymbol{\theta}}) = nlog(\hat{\sigma}^2_\epsilon)$ with $\hat{\sigma}^2_\epsilon = \mathbf{e}'\mathbf{e}/n$ the MLE of error variance. So
  - $AIC = nlog(\hat{\sigma}^2_\epsilon) + 2p$
  - $BIC = nlog(\hat{\sigma}^2_\epsilon) + log(n)p$.
- Penalty of *BIC* grows with sample size, not so for *AIC*.
- Penalty of *BIC* larger than for *AIC* if $n \geq 8$, so *BIC* tends to select smaller models.
- *AIC* is based on Kullback-Leibler information, measuring distance between true distribution of data and distribution of data under particular model. Skip details.
- *BIC* has basis in Bayesian hypothesis testing, comparing degree of support for two models. Skip details.

Example Model selection by AIC

- Function `step` doesn't perform all-possible-subset regression, but employs sequential stepwise method (by repeated `add1` and `drop1` statements) without giving dubious p-values. Partial output below.

```
> g <- lm(Life.Exp ~., data=statedata)
> stepres <- step(g, steps=2)
Start:  AIC=-22.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost + Area

             Df Sum of Sq  RSS   AIC
- Area        1      0.00 23.3 -24.2
- Income      1      0.00 23.3 -24.2
- Illiteracy  1      0.00 23.3 -24.2
<none>                    23.3 -22.2
- Population  1      1.75 25.0 -20.6
- Frost       1      1.85 25.1 -20.4
- HS.Grad     1      2.44 25.7 -19.2
- Murder      1     23.14 46.4  10.3

Step:  AIC=-24.18
Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
    Frost

             Df Sum of Sq  RSS   AIC
- Illiteracy  1      0.00 23.3 -26.2
- Income      1      0.01 23.3 -26.2
<none>                    23.3 -24.2
- Population  1      1.76 25.1 -22.5
- Frost       1      2.05 25.3 -22.0
- HS.Grad     1      2.98 26.3 -20.2
- Murder      1     26.27 49.6  11.6

Step:  AIC=-26.17
```


Summary model selection

- Stepwise procedures use restricted search through space of potential models.
- Testing-based procedures use dubious hypothesis testing.
- Criterion-based procedures typically search a wider space of models ("All possible subsets regression"), and compare models using a particular criterion.
- Criterion-based procedures are preferred.
- Variable selection is not end in itself, but means to an end. Aim: construct model that predicts well or explains relationships in data well. Automatic selections are not guaranteed to be consistent. Use methods as guide only.
- Accept possibility that several models are suggested which fit equally well. Then consider:
  - Do models have similar qualitative consequences?
  - Do they make similar predictions?
  - What is cost of measuring predictors?
  - Which has best diagnostics?

Model validation

- Model validation: split dataset into training subsample and validation subsample:
  - Training subsample is used to specify statistical model.
  - Validation subsample is used to evaluate the fitted model.
- Cross-validation is application of this idea, where roles of training and validation subsamples are interchanged or rotated.
- Statistical modeling: iterative sequence of data exploration, model fitting, model criticism, model-respecification.
  Variables may be dropped, interactions may be incorporated or deleted, variables may be transformed, unusal data may be corrected, removed, or otherwise accommodated.
  Resulting model should accurately reflect principal characteristics of data.
- Danger is capitalization on chance: overfitting and overstating strength of results.
- Ideal solution: collect new data with which to validate model; often not possible.
- Model validation simulates the collection of new data by randomly dividing data into two parts: one for exploration and model formulation, second for checking adequacy of model, formal estimation, and testing.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 6, lecture 31

Content week 6, lecture 31: Fox app D6.1-D6.4

- Maximum likelihood
  - likelihood function, log-likelihood
  - maximum likelihood estimator (MLE) and properties
  - asymptotic variance of mle; Fisher information
  - likelihood ratio test (LRT)
  - inference for single parameter: Wald-test, likelihood ratio test, score test
  - inference for several parameters, Fisher information matrix, asymptotic var-covar matrix

Maximum-Likelihood Estimation

- Most general estimation principle in statistics.
- Provides estimators with reasonable intuitive basis and many desirable statistical properties.
- Broadly applicable, relatively simple to apply.
- Theory of maximum-likelihood estimation provides standard errors, statistical tests, and other results useful for inference.
- Disadvantage: frequently requires strong assumptions about structure of data.

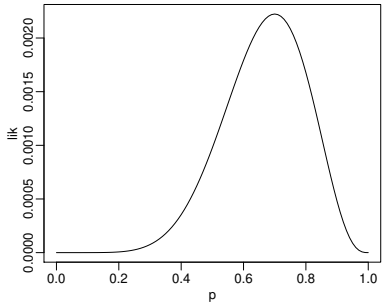
M.L.: Preliminary Example

- Example of flipping a coin. What is probability π of getting head?
- Suppose coin is flipped 10 times ($n = 10$) with results: $HHTHHHTTHH$.
- Probability of this sequence is function of unknown parameter π:
  $Pr(data|parameter) = Pr(HHTHHHTTHH|\pi) = \pi\pi(1-\pi)\pi\pi\pi(1-\pi)(1-\pi)\pi\pi = \pi^7(1-\pi)^3$
- Given the collected data, we can think of the probability as function of π: let it vary over its range $0-1$.
  Likelihood function is this probability as function of π:
  $L(\text{parameter}|\text{data}) = L(\pi|HHTHHHTTHH) = \pi^7(1-\pi)^3$.
- So, probability function and likelihood function are same equation, first is function of data, second is function of parameter.

M.L.: Preliminary Example (2)

Plot of likelihood function $L(\text{parameter}|\text{data}) = L(\pi|HHTHHHTTHH) = \pi^7(1-\pi)^3$:

```
> p <- seq(0,1,by=0.01)
> lik <- p^7*(1-p)^3
> plot(p,lik, type="l")
```



M.L.: Preliminary Example (3)

- Although each value of $L(\pi|data)$ is notional probability, function $L(\pi|data)$ is not probability distribution or density function: it does not integrate to 1, for example.
- In example: probability of obtaining the sample of data as found is small, regardless of true value of π. This is usually case: usually any specific sample result (including the realized one) will have low probability.
- Nevertheless, likelihood contains useful information about unknown parameter π:
  - E.g. π cannot be 0 or 1, because in that case the observed data could not have been observed at all.
  - Value of π that is most supported by the data is the one for which likelihood is largest: it is the maximum-likelihood estimate (MLE), denoted as $\hat{\pi}$.
- Here, $\hat{\pi} = 0.7$, which is just the sample proportion of heads: $7/10$.

Generalization of example

- Suppose n independent flips of coin, producing particular sequence that include x heads and $n - x$ tails.
- Then: $L(\pi|data) = Pr(data|\pi) = \pi^x(1-\pi)^{n-x}$.
- Find value of π that maximizes $L(\pi|\text{data})$, abbreviated as $L(\pi)$.
- Often simpler to maximize log of the likelihood:
  $\log L(\pi) = x \log \pi + (n-x)\log(1-\pi)$.
- Differentiate w.r.t. π: $\dfrac{d \log L(\pi)}{d\pi} = \dfrac{x}{\pi} + (n-x)\dfrac{1}{1-\pi}(-1)$, setting to 0, and solving for π produces the maximum-likelihood estimator (MLE).
- Maximum-likelihood estimator of π is $\hat{\pi} = X/n$.

Properties of Maximum-Likelihood Estimators

- Under very broad conditions, maximum-likelihood estimators have following properties:
  - Consistent
  - Asymptotically unbiased, although may be biased in finite samples
  - Asymptotically efficient: no asymptotically unbiased estimator has smaller asymptotic variance
  - Asymptotically normally distributed
  - If sufficient statistic for parameter exists, then MLE of parameter is function of sufficient statistic.
- Asymptotic sampling variance of MLE $\hat{\alpha}$ of single parameter α can be obtained from second derivative of log likelihood:

$$\mathcal{V}(\hat{\alpha}) = \frac{1}{-E\left[\dfrac{d^2 \log L(\alpha)}{d\alpha^2}\right]}$$

Denominator of $\mathcal{V}(\hat{\alpha})$ is called Fisher information: $\mathcal{I}(\alpha) \equiv -E\left[\dfrac{d^2 \log L(\alpha)}{d\alpha^2}\right]$

In practice, MLE $\hat{\alpha}$ is substituted, to get an estimate of asymptotic sampling variance $\hat{\mathcal{V}}(\hat{\alpha})$

Properties of Maximum-Likelihood Estimators (2)

Intuition:

- If likelihood has sharp peak, then MLE is clearly differentiated from nearby values. Then, second derivative of log likelihood is large negative number and the Fisher information will be larger positive number: there is lot of "information" in data concerning value of parameter, and sampling variance of MLE is small.
- If log likelihood is relatively flat at its maximum, alternative estimates quite different from MLE are nearly as good as MLE: second derivatice will be small, there is little information in data concerning value of parameter, and sampling variance of MLE is large.

Log Likelihood Ratio statistic

- $L(\hat{\alpha})$ is value of likelihood function at MLE $\hat{\alpha}$, while $L(\alpha)$ is the likelihood at the true (but generally unknown) parameter α.
- Log Likelihood Ratio Statistic $G^2 \equiv -2 \log \frac{L(\alpha)}{L(\hat{\alpha})} = 2\left(\log L(\hat{\alpha}) - \log L(\alpha)\right)$
- Often simply called Likelihood Ratio Statistic.
- G^2 has asymptotically a chi-square distribution with 1 d.f.
- As by definition MLE maximizes likelihood for the particular sample under consideration, the value of likelihood at true parameter value α is generally smaller than at MLE $\hat{\alpha}$.
- If $\hat{\alpha}$ is MLE of α, and if $\beta = f(\alpha)$ is function of α, then $\hat{\beta} = f(\hat{\alpha})$ is MLE of β.

Statistical Inference: Wald, Likelihood-Ratio, and Score Tests

Properties of MLE lead to 3 general procedures for testing a single parameter $H_0 : \alpha = \alpha_0$:

1. Wald test
2. Likelihood-ratio test
3. Score test

Statistical Inference: Wald, Likelihood-Ratio, and Score Tests

- Wald test: relying on asymptotic normality of MLE $\hat{\alpha}$, calculate test statistic as:

$$Z_0 \equiv \frac{\hat{\alpha} - \alpha_0}{\sqrt{\hat{\mathcal{V}}(\hat{\alpha})}}$$

  which is asymptotically distributed as $N(0,1)$ under H_0.
- Likelihood-ratio test: test statistic is:

$$G_0^2 \equiv -2 \log \frac{L(\alpha_0)}{L(\hat{\alpha})} = 2\left(\log L(\hat{\alpha}) - \log L(\alpha_0)\right)$$

  which is asymptotically distributed as χ_1^2 under H_0.

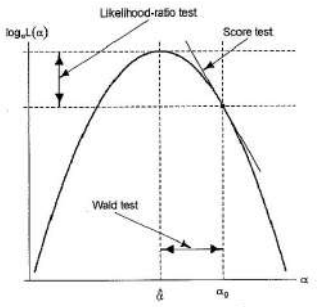
  Wald test and LR test can be "turned around" to produce confidence intervals.
- Score test: "score" $S(\alpha)$ is the derivative (slope) of log-likelihood at particular value of α. At the MLE the slope must necessarily be 0: $S(\hat{\alpha}) = 0$. It can be shown that score statistic

$$S_0 \equiv \frac{S(\alpha_0)}{\sqrt{\mathcal{I}(\alpha_0)}}$$

  is asymptotically distributed as $N(0,1)$ under H_0.

Statistical Inference: Wald, Likelihood-Ratio, and Score Tests (2)

- Unless log likelihood is quadratic, 3 test statistics produce somewhat different results, although they are asymptotically equivalent.
- In some contexts, score test has practical advantage of not requiring the computation of MLE $\hat{\alpha}$ (because S_0 depends only on null value α_0, specified under H_0.
- In most cases, LR test is more reliable than Wald and score tests.
- Relationships between 3 test statistics shown in figure:
  - Wald test measures distance between $\hat{\alpha}$ and α_0 using standard error to calibrate this distance. E.g., if $\hat{\alpha}$ is far from α_0, doubt is cast on H_0.
  - Likelihood ratio test measures distance between $\log L(\hat{\alpha})$ and $\log L(\alpha_0)$: if $\log L(\hat{\alpha})$ is much larger than $\log L(\alpha_0)$, then H_0 is probably wrong.
  - Score test measures slope of log likelihood at α_0: if slope is very steep, then we are far from the peak of likelihood function, casting doubt on H_0.



- Example on blackboard.

Several Parameters

- Maximum-likelihood can be generalized to simultaneous estimation of several parameters.
- Suppose we have multivariate observations collected in matrix $\mathbf{X}$ with n observations on m variables, depending on k parameters collected in vector $\boldsymbol{\alpha}$.
- Let $p(\mathbf{X}|\boldsymbol{\alpha})$ represents the joint probability or probability density for m-dimensional observations.
- Likelihood $L(\boldsymbol{\alpha}) \equiv L(\boldsymbol{\alpha}|\mathbf{X})$ is function of parameter vector $\boldsymbol{\alpha}$; find values $\hat{\boldsymbol{\alpha}}$ that maximize this function.
- More convenient with log's: $\log L(\boldsymbol{\alpha})$.
- To maximize likelihood, first find vector of partial derivatives $\partial \log L(\alpha)/\partial\alpha$, set this derivative vector to $\mathbf{0}$, and solve the matrix equation for $\hat{\boldsymbol{\alpha}}$. If there is more than one root, choose solution that produces largest likelihood.

Several Parameters: properties

- MLE is consistent, asymptotically unbiased, asymptotically efficient, asymptotically (multivariate) normal
- Asymptotic $(k \times k)$ variance-covariance matrix of MLE is

$$\mathcal{V}(\hat{\boldsymbol{\alpha}}) = \left\{ -E\left[\frac{\partial^2 \log L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right] \right\}^{-1}$$

- Fisher information matrix or expected information matrix is

$$\mathcal{I}(\boldsymbol{\alpha}) \equiv -E\left[\frac{\partial^2 \log L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} \right]$$

- Notice how formulae for several parameters closely parallel those for single parameter.

Several Parameters: hypothesis tests

- Wald test for $H_0 : \boldsymbol{\alpha} = \boldsymbol{\alpha}_0$ uses test statistic

$$Z_0^2 \equiv (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)' \hat{\mathcal{V}}(\hat{\boldsymbol{\alpha}})^{-1} (\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0)$$

  asymptotically χ^2_k-distributed under H_0.
- Likelihood ratio test generalizes straightforwardly:

$$G_0^2 \equiv -2 \log \frac{L(\boldsymbol{\alpha}_0)}{L(\hat{\boldsymbol{\alpha}})}$$

  asymptotically distributed as χ^2_k under H_0.
- Score test uses score vector $S(\boldsymbol{\alpha}) \equiv \partial \log L(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$, resulting in score statistic:

$$S_0^2 \equiv S(\boldsymbol{\alpha}_0)' \mathcal{I}(\boldsymbol{\alpha}_0)^{-1} S(\boldsymbol{\alpha}_0)$$

  asymptotically χ^2_k-distributed under H_0.

Several Parameters: hypothesis tests (2)

- Each test can be adapted to more complex hypotheses, e.g. test H_0 that p out of k elements of $\boldsymbol{\alpha}$ are equal to particular values.
- Let $L(\hat{\boldsymbol{\alpha}}_0)$ represents maximized likelihood under constraint specified by hypothesis, i.e. setting p parameters to values specified under H_0, but estimating the remaining $k - p$; $L(\hat{\boldsymbol{\alpha}})$ represents globally maximized likelihood. Then, under H_0:

$$G_0^2 \equiv -2 \log \frac{L(\hat{\boldsymbol{\alpha}}_0)}{L(\hat{\boldsymbol{\alpha}})}$$

  has asymptotically χ^2_p distribution.
- Example on blackboard.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 6, lecture 32

Content week 6, lecture 32: Fox §14.1, 14.3.1; Faraway ELM §2)

- Generalized linear model (GLM) for binomial data
  - binomial and binary data
  - problems with linear regression
  - generalized linear model:
    - binomial distribution; expectation and variance
    - linear predictor
    - link functions: logit, probit, complementary log-log
  - maximum likelihood for binomial glm
  - logistic regression; odds, OR, interpretation regression coefficient: $e^{\beta} = OR$

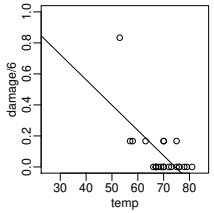
Generalized Linear Models: Binomial data

- We now study extension of linear models: Generalized Linear Models.
- Response no longer needs to be normally distributed, but distribution can be chosen from a class of distributions.
- Look at trials with binary outcome: success or failure. The number of successes out of total number of trials n has (under some conditions) binomial distribution. Outcomes are: $0, 1, \ldots, n$.
- With single (binary) trial ($n = 1$), only two outcomes possible: 0 or 1. Binomial distribution with $n = 1$ is called the Bernoulli distribution.

Example: Challenger disaster (Faraway)

- In 1986 space shuttle Challenger exploded after launch due to failure of rubber O-ring seals in rocket boosters. At lower temperatures rubber is more brittle. At time of launch temperature was $31^{\circ}\text{F} = \frac{5}{9}(31 - 32)^{\circ}\text{C} = -0.5^{\circ}\text{C}$.
- Could the failure of O-rings have been predicted?
- Data are available from 23 previous shuttle missions. Each shuttle has 2 boosters, each with 3 O-rings.
- Plot proportion of damaged O-rings, and fit simple linear regression model.
- Anything peculiar w.r.t. shown predicted values?

```
> plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim=c(0,1))
> linmod <- lm(damage/6 ~ temp, orings)
> abline(linmod)
> pred <- predict(linmod)
> pred[pred<0]
       17        18        19        20        21        22        23
-0.008929 -0.008929 -0.025238 -0.025238 -0.057857 -0.074167 -0.106786
```

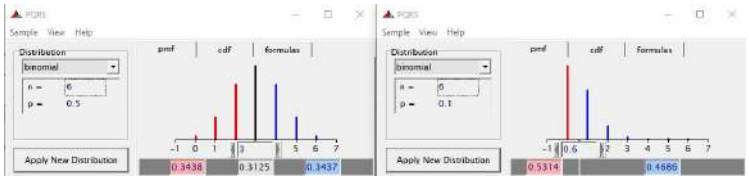


Problems with linear regression

- With simple regression, predicted probabilities can be greater than 1 or less than 0!! Possible solution: truncate predicted probabilities to 0 or 1, but the model would not be credible: for temperatures low enough the probability of failure would suddenly be exactly 1, and temperatures high enough suddenly exactly 0. Smooth transitions would be preferred.
- Sensible distribution for number of damaged O-rings is binomial distribution: $Y_i \sim B(6, \pi_i)$. In simple linear regression we assume normal distribution of fraction, though.
  This flaw may not be very serious, due to central limit theorem.
- Variance of the binomial count Y_i is $6\pi_i(1-\pi_i)$. Variance of binomial fraction $Y_i/6$ is $\pi_i(1-\pi_i)/6$. Both are not constant.
  However, the linear model assumes constant variance.
- Hence, assumptions of linear model are violated in multiple ways.
- Way out: use transformation of response and use weighted least squares.
- Better: develop model directly suited for binomial data.

Binomial distribution

- Suppose response Y_i has binomial distribution $B(n_i, \pi_i)$, with $i = 1, \ldots, n$, and Y_i independent: $P(Y_i = y_i) = \binom{n_i}{y_i} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$.
  Binomial distribution assumes that n_i individual trials with outcomes 0 (failure) or 1 (success), composing the total count of successes Y_i, are independent, and have same probability of success π_i.
- Below two binomial distributions are shown with $n = 6$ (as in spaceshuttle example) and with $\pi = 0.5$ and $\pi = 0.1$.



- What is the expected value of Y_i for the two cases shown above? What is the general expression for $E(Y_i)$? And for the binomial fraction $E(Y_i/n_i)$?
- Notice that for $\pi = 0.1$ the variance is smaller than for $\pi = 0.5$. What is variance for the two cases shown above? What is the general expression of $var(Y_i)$? And of the variance of the binomial fraction $var(Y_i/n_i)$?

Towards the binomial glm: linear predictor and link function

- Individual trials composing response Y_i are assumed to be subject to one set of q predictor values $(x_{i1}, \ldots, x_{iq})$. Such group of trials is called covariate class.
- We need model that describes relationship between the set of predictors $x_1, \ldots, x_q$ and the probability of success π:
  construct linear predictor $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq}$.
- In g.l.m. world the symbol η often used for linear predictor.
- Linear predictor η_i can accommodate quantitative and qualitative predictors (by use of dummies), and allows for transformation of predictors. It is as flexible as in linear model case, and retains interpretability.
- Setting directly $\eta_i = \pi_i$ does not work, because η_i is in principle unrestricted, whereas $0 \leq \pi_i \leq 1$.
- We need a link function g; this function should transform a value in the range $0-1$ (of π) into a value in range $-\infty, \infty$ (of linear predictor η).

Link functions in binomial case

- So, link function g is needed. It should have the following properties:
  - g must be monotone.
  - g must be such, that $g(\pi_i) = \eta_i$, or that the inverse of g transforms η_i back to π_i: $0 \leq g^{-1}(\eta_i) \leq 1$.
- Many functions g could be chosen, e.g. any inverse cumulative distribution function (cdf), because a cumulative distribution function transforms a value in range $(-\infty, \infty)$ to range $(0, 1)$, and is monotone.
- Three common choices:
  - Logit: $\eta = log(\pi/(1-\pi))$ (this is inverse of logistic cdf $\frac{1}{1+e^{-\eta}}$).
  - Probit: $\eta = \Phi^{-1}(\pi)$ (this is inverse of normal cdf).
  - Complementary log-log: $\eta = log(-log(1-\pi))$ (this is inverse of extreme value cdf).
- For the moment we choose the logit-link function. We saw this function in week 1 as a transformation for probabilities: $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \log(\text{odds}(\pi))$!

Fitting the binomial glm

- Maximum likelihood is used to estimate the parameters of the model.
- We need the log of the binomial likelihood. This turns out to be (try to derive this during practical): $\log L(\beta) = \sum_{i=1}^{n} \left[y_i \eta_i - n_i \log(1 + e_i^{\eta_i}) + \log \begin{pmatrix} n_i \\ y_i \end{pmatrix} \right]$
- Realize that in the expression above $\eta_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$.
- Maximizing the log likelihood (w.r.t. β's) gives MLE $\hat{\beta}$'s.
- Standard m.l. theory gives approximate standard errors of $\hat{\beta}$'s and hypothesis tests for parameters (Wald test, likelihood ratio tests).

Fitting the binomial glm: spaceshuttle example

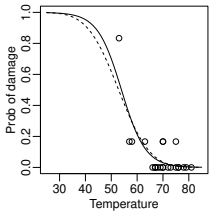
- Binomial responses consist of two parts: total number of trials n, and number of successes y. In R pass this information through two-column matrix containing number of successes y and number of failures $n - y$.
- Linear predictor: $\eta_i = \beta_0 + \beta_1 temp_i$ $(i = 1, \ldots 23)$
- For binomial distribution default link function is logit: logistic regression; $logit(\pi_i) = \eta_i$
- M.l. estimates of β_0 and β_1 are 11.6 and -0.216 with approximate standard errors.

```
> logitm <- glm(cbind(damage, 6-damage) ~ temp, family=binomial, orings)
> summary(logitm)

Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.953  -0.735  -0.439  -0.208   1.956

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.6630     3.2963    3.54    4e-04
temp         -0.2162     0.0532   -4.07  4.8e-05

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
AIC: 33.67

Number of Fisher Scoring iterations: 6
```


Fitting the binomial glm: example continued

- Fitted logistic regression model is below (black line).
- Also try the probit link function, and plot results into same scatterplot (dashed line).

```
> plot(damage/6 ~temp, orings, xlim=c(25,85), ylim=c(0,1),
+ xlab="Temperature", ylab="Prob of damage")
> x <- seq(25,85,1)
> lines(x,ilogit(11.6630-0.2162*x))
> # below probit regression
> probitm <- glm(cbind(damage, 6-damage) ~ temp,
+ family=binomial(link=probit), orings)
> coef(summary(probitm))
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   5.5915    1.71055   3.269 1.080e-03
temp         -0.1058    0.02656  -3.984 6.791e-05
> lines(x,pnorm(5.5915-0.1058*x), lty=2) # probit results
```



- Although coefficients from logit and probit links are quite different, fits are similar.

Predicting Challenger probabilities

- Predict response at 31°F (but notice that we are extrapolating...)

```
> ilogit(11.6630-0.2162*31) # according to logistic regression
[1] 0.993
> pnorm(5.5915-0.1058*31)   # according to probit regression
[1] 0.9896
> # nicer way:
> predict.at <- data.frame(temp=31)
> predict(logitm, predict.at, type="response")
    1
0.993
> predict(probitm, predict.at, type="response")
     1
0.9896
> # on scale of linear predictor:
> predict(logitm, predict.at, type="link")
   1
4.96
> predict(probitm, predict.at, type="link")
    1
2.312
>
```

Interpreting β's in logistic regression

- $\log(\text{odds}) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
  What interpretation does β_1 have? Unit increase in x_1 (holding other covariates constant), increases odds of success by factor of e^{β_1}.
  e^{β_1} is OR=Odds Ratio: the ratio of the odds at $x_1 + 1$ and the odds at x_1.
- If two groups (e.g. females versus males) are compared in logistic regression and β_1 is the regression coefficient for the corresponding dummy variable, e^{β_1} is the odds ratio of the group coded with value 1 versus the group coded with 0.
  To see this, take logistic regression model $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ in which x_1 is dummy variable coding for the two levels of the group factor:
  $x_1 = 1$ (e.g. females): $\log\left(\frac{\pi_F}{1-\pi_F}\right) = \beta_0 + \beta_1 1 + \beta_2 x_2 = \beta_0 + \beta_1 + \beta_2 x_2$
  $x_1 = 0$ (e.g. males): $\log\left(\frac{\pi_M}{1-\pi_M}\right) = \beta_0 + \beta_1 0 + \beta_2 x_2 = \beta_0 + \beta_2 x_2$
  Subtraction: $\log\left(\frac{\pi_F}{1-\pi_F}\right) - \log\left(\frac{\pi_M}{1-\pi_M}\right) = \beta_1$
  Rewriting lhs: $\log(\text{odds}_F) - \log(\text{odds}_M) = \log(\text{OR}_{FvsM}) = \beta_1$ or $\text{OR}_{FvsM} = e^{\beta_1}$
- Relation of odds-ratio to relative risk RR.
  $RR = \pi_1/\pi_2$ with π_1 the probability of success in group 1, and π_2 in group 2. For rare outcomes (small π_1 and π_2) $RR \approx OR = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 6, lecture 33

Content week 6, lecture 33: Fox §14.1; ELM §2)

- Statistical inference in GLM for binomial data
  - deviance
    - comparing deviances: likelihood ratio test
    - deviance test for goodness of fit with binomial (n not too small) data
  - Wald test for individual coefficient (or single linear combination of coefficients)
  - goodness of fit: deviance and Pearson's X^2
- Miscellanea
  - link function motivated through tolerance distribution
  - prospective vs retrospective sampling
  - predictions: link scale and response scale; effective dose

Statistical Inference in (binomial) GLM

- Recall the three methods for hypothesis testing / confidence intervals in maximum likelihood estimation:
  1. Wald test
  2. likelihood ratio test
  3. score test
- Score tests are hardly found in standard GLM applications, but Wald and likelihood ratio tests are.
- For likelihood ratio tests we work with deviances. Deviances in GLMs replace sums of squares in LM.

Deviance

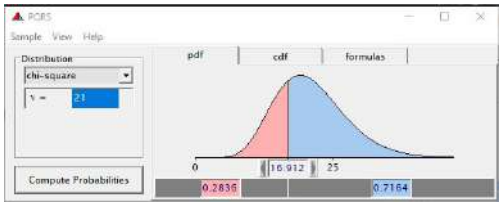
- Nested models (Full Model FM and Reduced Model RM) can be compared using likelihood ratio tests, as we know:
  $LRT = 2\log\frac{L(FM)}{L(RM)} = 2log(L(FM)) - 2log(L(RM))$.
- Saturated model (SM) is largest possible model for dataset with n observations: model with a parameter for every observation, so a model with n parameters.
- As model SM is useless, because it does not bring any simplification: it is as complex as the dataset we start with. But it bringS an upperbound for likelihood!
- Maximized likelihood for saturated model is highest possible likelihood for any model for this specific dataset.
- Now compare current model (call it CM) with the saturated model. CM is nested within SM, so a LRT statistic can be formed.
- In the binomial case the LRT statistic is $D = 2logL(SM) - 2logL(CM) = 2\sum_{i=1}^{n}\{y_i\log(y_i/\hat{y}_i) + (n_i - y_i)\log((n_i - y_i)/(n_i - \hat{y}_i))\}$
  in which $\hat{y}_i$ are fitted values from current model; y_i are "fitted values" in saturated model: fitted values are simply observed y_i.
- D is called the deviance. It measures how close the current model comes to perfection.
- D is measure of goodness of fit.

Example deviance

- Continue with spaceshuttle example: fit logistic regression model, explaining fraction failing o-rings by temperature.
- R-output gives a `Residual deviance`, i.e. deviance of current model, and `Null deviance`, i.e. deviance for null model, containing intercept only.

```
> logitm <- glm(cbind(damage, 6-damage) ~ temp, family=binomial, data=orings)
> summary(logitm)
    ......
    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

- What is residual deviance telling us?
- Check chi-square distribution with 21 df!



Deviance for goodness of fit

- If Y is binomially distributed with n_i relatively large and an adequately fitting model, deviance has approximately χ^2_{n-p} distribution.
  What value do you expect for this distribution? What is the variance and standard deviation?
- Deviance test can be used to check whether model has adequate fit. Conclusion?

```
> pchisq(deviance(logitm), df.residual(logitm), lower=FALSE)
[1] 0.7164
```

- If deviance is much larger than its d.f., null hypothesis of "adequately fitting model" can be rejected.
- χ^2 distribution is only approximation, that becomes more accurate as n_i increase.
- For binary response, so $n_i = 1$, and $y_i = 0$ or $y_i = 1$, deviance reduces to $D = -2\sum_{i=1}^{n}\{\hat{\pi}_i\text{logit}(\hat{\pi}_i) + \log(1 - \hat{\pi}_i)\}$.
  Deviance D is function of $\hat{\pi}_i$ only: it does not compare fitted values $\hat{\pi}_i$ to the data y_i, hence it does not assess goodness of fit, and it is not χ^2 distributed.
  In binary case other methods for judging goodness of fit must be employed.
- Approximation is very poor for small n_i. Values $n_i \geq 5$ has been suggested.

Analysis of deviance

- Better use of deviances is to construct the LRT statistic to compare two nested models:
  - Difference in deviances $D_{RM} - D_{FM}$ is just the LRT statistic, that we already know: $D_{RM} - D_{FM} = (2log(L(SM) - 2log(L(RM))) - (2log(L(SM) - 2log(L(FM))) = 2log(L(FM)) - 2log(L(RM))$
  - So, deviance difference of RM and FM is same as twice log likelihood diffence of FM and RM, which is the LRT.
  - Differences of residual sums of squares in LM, generalize to differences of deviances in GLM! Instead of Analysis of Variance (in LM) we apply Analysis of Deviance in GLM.
  - Test statistic has approximately χ^2_{l-s} distribution with l residual df of RM and s residual df of FM.
- Example: test effect of temperature, i.e. $H_0 : \beta = 0$. Compare models with temp (resulting in `Residual deviance`), and without temp (then we arrive at null model, resulting in `Null deviance`)

Analysis of deviance example

- LRT for temperature "by hand":

```
> pchisq(logitm$null.deviance - logitm$deviance, logitm$df.null - logitm$df.residual, lower=FALSE)
[1] 2.747e-06
```

- LRT for temperature using `anova` (this is type I test, order dependent):

```
> anova(logitm, test="LRT")
Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(damage, 6 - damage)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    22       38.9
temp  1       22        21       16.9  2.7e-06
```


Wald test for single coefficient

- Alternative is Wald test-statistic $\hat{\beta}/\text{se}(\hat{\beta})$, which under H_0 has approximately standard normal distribution.
- Notice that p-values for LRT and Wald test are not identical (but they would with normal distributions).
- In some cases, especially with sparse data, standard errors can be overestimated and z-values too small, so Wald test may miss the effect –> Hauck-Donner effect.
- In general: LRT is preferred.

```
> # Wald test for temp:
> coef(summary(logitm))
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)  11.6630    3.29626   3.538 4.028e-04
temp         -0.2162    0.05318  -4.066 4.777e-05
```

- In other statistical programs Wald test statistic is presented as square of the above: $\hat{\beta}^2/\text{var}(\hat{\beta})$, which under H_0 has approximately χ^2 with 1 df.

Confidence intervals

- Confidence intervals based on normal approximations, related to Wald test: $100(1-\alpha)\%$ c.i. for β_i is $\hat{\beta}_i \pm z_{a/2}\text{se}(\hat{\beta}_i)$.
- Better to construct (profile) likelihood-based confidence interval, using library `MASS`

```
> # c.i. based on normal approximation
> b <- coef(summary(logitm))[2,1]
> se.b <- coef(summary(logitm))[2,2]
> lb <- b - qnorm(1-0.05/2) * se.b ; ub <- b + qnorm(1-0.05/2) * se.b
> c(lb,ub)
[1] -0.3205 -0.1120
> #profile likelihood c.i.
> library(MASS)
> confint(logitm)
              2.5 %  97.5 %
(Intercept)  5.5752 18.7376
temp        -0.3327 -0.1202
```


Goodness of Fit

- Besides deviance, Pearson's X^2 statistic can be used as measure of goodness of fit. We know the X^2 statistic as: $X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$ with O_i observed counts and E_i expected counts.
- In binomial case this becomes

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}$$

- Pearson residuals are $r_i^P = (y_i - n_i\hat{\pi}_i)/\sqrt{var\hat{y}_i}$, which are sort of standardized residuals.
- Pearson's X^2 is sum of squared Pearson residuals.
- Like the residual deviance Pearson's X^2 has approximately a χ^2_{n-p} distribution under the null hypothesis (of a "well fitting" model).

```
> modl <- glm(cbind(dead,alive) ~conc, family=binomial, data=bliss)
> sum(residuals(modl,type="pearson")^2)
[1] 0.3673
> deviance(modl)
[1] 0.3787
```

Tolerance Distribution

- Link functions can be motivated by modeling a (latent) variable on an underlying scale, with a specific tolerance distribution.
- E.g. suppose students answer questions on test; specific student has (unobserved) aptitude T, and particular question has difficulty d. Student will answer question correct only if $T > d$. Suppose that randomly selected student has aptitude $T \sim N(\mu, \sigma^2)$. Probability that randomly selected student answers question wrongly is:
  $p = P(T \leq d) = P((T-\mu)/\sigma \leq (d-\mu)/\sigma) = P(z \leq (d-\mu)/\sigma) = \Phi((d-\mu)/\sigma)$
  Hence $\phi^{-1}(p) = (d-\mu)/\sigma = -\mu/\sigma + (1/\sigma)d$.
- We see here probit regression model with intercept $-\mu/\sigma$ and slope $1/\sigma$.
- By modeling aptitude with logistic distribution, we arrive at logistic regression.
- By modeling aptitude with extreme value distribution, we arrive at complementary log-log regression.

Prospective and Retrospective Sampling (1)

- Example: study into the relation between type of feeding (breast fed or bottle fed) for babies and occurrence of bronchitis. Are breast fed babies less diseased?
- Prospective sampling: predictors are fixed and outcome is observed.
  E.g. select sample of new born whose parents had chosen particular feeding method, and observe number children developing bronchitis.
  Called cohort study.
- Retrospective sampling: outcome is fixed and predictors are observed.
  E.g. infants come to doctor with respiratory disease, and we record the method of feeding.
  Called case control study.
- It seems that prospective sampling is needed to answer the question of interest. However, retrospective sampling (case control study) is equally effective!

Example Prospective and Retrospective Sampling (2)

- Data:

	diseased	healthy	total
bottle fed	77	381	458
breast fed	47	447	494
total	124	828	952

- Cohort study (prospective sampling):
  - 458 babies are bottle fed and 494 are breast fed (494). After some time baby's disease status is determined.
  - Given infant is bottle fed, log-odds of having disease is $\log\left(\frac{77}{458}/(1-\frac{77}{458})\right) = \log\left(\frac{77}{458}/\frac{381}{458}\right) = \log(77/381) = -1.60$
  - Given infant is breast fed, log-odds of having disease is $\log(47/447) = -2.25$
  - Hence difference in log-odds is $\Delta = -1.60 - -2.25 = 0.65$, representing increased risk of respiratory disease incurred by bottle feeding relative to breast feeding. This is regression coefficient for dummy for bottle fed, analyzing the fraction diseased babies.
- Case control study (retrospective sampling):
  - 124 diseased and 828 healthy babies enter the study. For each baby the method of feeding is determined.
  - Given infant is diseased, log-odds of being bottle fed are $\log(77/47)$
  - Given infant is healthy, log-odds of being bottle fed are $\log(381/447)$
  - Difference in log-odds is $\Delta = \log(77/47) - \log(381/447) = 0.65$. Same! This is regression coefficient for dummy for diseased, analyzing fraction bottle fed babies.
- Hence, retrospective design is as effective as prospective design!

Logistic regression for Prospective and Retrospective Sampling (3)

```
> # prospective study
> n<- c(458,494)
> k <- c(77,47)
> d.bottle <- c(1,0) #dummy for bottle fed
> prospective <- glm(cbind(k,n-k) ~ d.bottle, family=binomial(link=logit))
> coef(summary(prospective))
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  -2.2524     0.1533 -14.689 7.603e-49
d.bottle      0.6534     0.1978   3.303 9.552e-04
> # retrospective study
> n<- c(124,828)
> k <- c(77,381)
> d.disease <- c(1,0) #dummy for bronchitis
> retrospective <- glm(cbind(k,n-k) ~ d.disease, family=binomial(link=logit))
> coef(summary(retrospective))
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)  -0.1598     0.06973  -2.291 0.0219506
d.disease     0.6534     0.19780   3.303 0.0009552
```

- Odds ratio of being diseased for bottle versus breast fed babies is $e^{0.65342} = 1.92$.
- Odds ratio of being bottle fed for diseased versus healthy babies is $e^{0.65342} = 1.92$.
- Identical regression coefficients in the two logistic regressions!

Prediction and Effective Doses

- Suppose we want to predict outcome for given values of covariates. For binary data this means estimating probability of success.
- For given covariates x_0, $\hat{\eta} = x_0\hat{\beta}$ with approximate variance given by $x_0'\hat{\mathcal{V}}(\hat{\beta})x_0$.
- Approximate c.i. using normal approximation.
- For answer on probability scale, we need to transform back using inverse link.

```
> lmod <- glm(cbind(dead,alive) ~conc, family=binomial, data=bliss)
> lmodsum <- summary(lmod)
> x0 <- c(1,2.5)
> eta0 <- sum(x0 * coef(lmod))
> ilogit(eta0)
[1] 0.6413
```

- So, 64% predicted chance of death at this dose.
- 95% c.i. for probability.

```
> (cm <- lmodsum$cov.unscaled)   #extract variance-covariance matrix
            (Intercept)     conc
(Intercept)     0.17463 -0.06582
conc           -0.06582  0.03291
> se <- sqrt( t(x0) %*% cm %*% x0)
> ilogit(c(eta0 - 1.96*se, eta0+1.96*se))
[1] 0.5343 0.7358
> # or use predict function
```


Prediction and Effective Doses (2)

- Sometimes we wish to estimate value of x corresponding to chosen probability π.
- For example, determine which dose $ED50$ will lead to probability of $\pi = 0.5$ of success.
- ED stands for Effective Dose, LD for Lethal Dose.
- $\widehat{\text{ED50}} = -\hat{\beta}_0/\hat{\beta}_1$.

```
> (ld50 <- -lmod$coef[1]/lmod$coef[2])
(Intercept)
          2
```

- To determine standard error delta method can be used, giving approximate variance: $\text{var}\left(g(\hat{\theta})\right) \approx g'(\hat{\theta})^T var(\hat{\theta}) g'(\hat{\theta})$
- Theoretical background can be found in Fox Appendix, D.6.5. We return to the topic next week.
- Library MASS contains function `dose.p`

```
> library(MASS)
> dose.p(lmod,p=c(0.5,0.9))
          Dose     SE
p = 0.5: 2.000 0.1784
p = 0.9: 3.891 0.3450
```


Estimation Problems

- Sometimes the (Fisher scoring) algorithm fails to converge.
- This may happen if groups are linearly separable: in that case a perfect fit is possible, but it results in unstable estimates of parameters.
- Example below: 10 observations with $x \leq 5$ all with failures, 10 observations with $x > 5$ all with successes.
- R issues warnings:

```
1:  glm.fit:  algorithm did not converge
2:  glm.fit:  fitted probabilities numerically 0 or 1 occurred
```

```
> k <- c(rep(0,10), rep(1,10))
> n <- 1
> x <- c(1,4,3,5,4,3,5,4,1,2, 7,8,6,10,11,9,7,8,9,10)
> glmo <- glm(cbind(k,n-k) ~ x, family=binomial (link=logit))
> coef(summary(glmo))
            Estimate Std. Error   z value Pr(>|z|)
(Intercept)  -241.22     226827 -0.001063   0.9992
x              43.79      40991  0.001068   0.9991
```


Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 6, lecture 34

Content lecture 34: ELM 2.11 Overdispersion

- Overdispersion (binomial case)
  - reasons for overdispersion
  - extra scale parameter
  - inference: F-statistic based on deviances and extra scale parameter
  - other methods for handling overdispersion

Overdispersion binomial case

- Binomial response variable
- Suppose binomial GLM specification is correct. In that case (and provided that binomial totals are not too small) residual deviance has approximately χ^2_{n-p} distribution.
- Sometimes (often?) we observe residual deviance (and/or Pearson's X^2) larger than expected according to χ^2_{n-p}, i.e. larger than $\approx (n-p) + 2 \times \sqrt{2(n-p)}$ (why this value?).
- It is essential that you recognize this situation and take action!

Causes for binomial overdispersion

- Wrong structural form of model: miss important predictors, did not rightly transform. Try to check this.
- Occurrence of outliers. Try to locate these.
- Sparse data: remember that in most extreme case of binary response, residual deviance does not even approximately follows χ^2_{n-p} distribution, and does not not contain any information about model fit. In less extreme cases approximation is poor. With small binomial totals we are not able to draw conclusions about fit of model based on the residual deviance.
- Deficiency in random part of model: binomial distribution is not appropriate for response. Remember that binomial distribution appears if we have constant probability of success and independent binary trials. In that case, for binomial total m, $var(Y) = mp(1-p)$. If assumptions of constant probability and independence of individual trials are not fulfilled, variance of binomial count may by greater than $mp(1-p)$ (overdispersion), or, less frequently, less than $mp(1-p)$ (underdispersion).

Overdispersion due to deficiency in random part

- Overdispersion can arise due to violation of independence of individual binary trials or of the assumption of identical probabilities of individual trials.
- There may be unexplained heterogeneity within a group that leads to variation in p.
- Example spaceshuttle: position of O-ring on one of the two booster rockets, or the booster itself, may affect failure probability, but this variable is not recorded and cannot be included in model.
- Heterogeneity can also result from clustering. Instead of sampling individuals from population, clusters are sampled. Spaceshuttle example may be looked upon in this way: we do not sample individual O-rings, but they come in groups of 6 (actually 2 groups of 3) from a single spaceshuttle launch.

Overdispersion due to clustering of binary trials

Suppose we have c clusters, each with cluster size k, so that total binomial sample size is $m = ck$. Spaceshuttle example: 2 clusters (boosters), each with 3 o-rings. Let Z_i be number of successes in cluster i, with $Z_i \sim B(k, p_i)$ and Z_i independent. Now suppose that the binomial probability itself p_i is a random variable with $E(p_i) = p$ and $var(p_i) = \tau^2 p(1-p)$. Total number of successes is $Y = Z_1 + \cdots + Z_c$. Then

- $E(Y) = E(\sum_{i=1}^{c} Z_i) = \sum E(Z_i) = \sum E_{p_i}(E_{Z_i}(Z_i|p_i)) = \sum E_{p_i}(kp_i) = \sum kE(p_i) = \sum_{i=1}^{c} kp = mp$.
  So, expected count is identical as in standard case.
- $var(Y) = var(\sum_{i=1}^{c} Z_i) = \sum var(Z_i) = \sum\{E(var(Z_i|p_i)) + var(E(Z_i|p_i))\} = \sum\{E(kp_i(1-p_i)) + var(kp_i)\} = \sum\{kE(p_i(1-p_i)) + k^2 var(p_i)\} = \sum\{k(E(p_i) - E(p_i^2)) + k^2 var(p_i)\} = \sum\{k(E(p_i) - (var(p_i) + (E(p_i))^2)) + k^2 var(p_i)\} = \sum\{k(p - (\tau^2 p(1-p) + p^2)) + k^2\tau^2 p(1-p)\} = \sum\{(1 + (k-1)\tau^2)kp(1-p)\} = (1 + (k-1)\tau^2)mp(1-p)$
- $var(Y) = (1 + (k-1)\tau^2)mp(1-p)$, or, $var(Y) = \phi mp(1-p)$ with $\phi = (1 + (k-1)\tau^2)$. Overdispersion, because $\phi = (1 + (k-1)\tau^2) \geq 1$.
- In sparse case with $m = 1$ (or $k = 1$), the problem cannot occur.

Overdispersion due to dependence

- Overdispersion or underdispersion can result from dependence between binary trials.
- E.g. subjects in animal trials may be influenced in their response by other subjects. If food supply is limited, probability of survival of one animal may be increased if another animal dies, resulting in underdispersion.

Modeling overdispersion: extra scale parameter

- Simplest approach is introduction of extra dispersion (scale) parameter ϕ: $var(Y) = \phi mp(1-p)$ with ϕ constant but unknown. Note that in earlier formula overdispersion parameter ϕ depended on cluster size k and τ.
- In standard binomial case $\phi = 1$: no overdispersion
- Estimate ϕ from Pearson's X^2 or residual deviance:

$$\hat{\phi} = \frac{X^2}{n-p}$$

  Faraway advises not to use residual deviance.
- Estimation of β is unaffected, but variances and standard errors are, because $\widehat{var}(\hat{\beta}) = \hat{\phi}(X'\hat{W}X)^{-1}$. Standard errors need to be "blown up" with factor $\sqrt{\hat{\phi}}$.
- Deviance differences cannot be used for comparisons of nested models anymore. Deviance differences have approximately $\phi\chi^2$ distributions. Instead, use F-statistics, with approximately F-distributions:

$$F = \frac{(D_{RM} - D_{FM})/(dfe_{RM} - dfe_{FM})}{\hat{\phi}}$$

Example overdispersion

- Data from Manly: boxes of trout eggs buried on 5 stream locations and retrieved at four time points. Number of surviving eggs was recorded.
- In one case all eggs survive, in other none.

```
> head(troutegg, n=4)
  survive total location period
1      89    94        1      4
2     106   108        2      4
3     119   123        3      4
4     104   104        4      4
```

- Fit logistic regression model for fraction surviving eggs with additive effects of location and period.

```
> bmod <- glm(cbind(survive, total-survive) ~ location + period, family=binomial, data=troutegg)
> coef(summary(bmod))
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   4.6358     0.2813  16.479 5.191e-61
location2    -0.4168     0.2461  -1.694 9.035e-02
location3    -1.2421     0.2194  -5.660 1.511e-08
location4    -0.9509     0.2288  -4.157 3.230e-05
location5    -4.6138     0.2502 -18.439 6.339e-76
period7      -2.1702     0.2384  -9.103 8.775e-20
period8      -2.3256     0.2429  -9.573 1.043e-21
period11     -2.4500     0.2341 -10.466 1.243e-25
> deviance(bmod)
[1] 64.5
> df.residual(bmod)
[1] 12
```

Example overdispersion (2)

- Residual deviance is 64.5 on 12 d.f. Overdispersion! E.g. because $64.5 >> 12 + 2\sqrt{2 \cdot 12} = 21.8$ (or look in χ^2_{12} distribution)
- Reason? Check for outliers, check whether interaction is needed (e.g. by plotting empirical logits).
- Lacking other reasons for overdispersion, put blame on random part of model.
- Estimate scale parameter, here based upon Pearson's X^2 as:

```
> (phi <- sum( residuals(bmod, type="pearson")^2 )/12)
[1] 5.33
```


Example overdispersion (3)

- Make F-tests for predictors instead of ordinary χ^2-tests for deviance differences!
- Here we use `scale` to specify the scale parameter estimated from Pearson's X^2 (which is the default), but you could change this into a scale parameter based upon the residual deviance.
- Below column `Deviance` shows the residual deviance of the model after dropping that specific term.

```
> drop1(bmod, scale=phi, test="F")
Single term deletions

Model:
cbind(survive, total - survive) ~ location + period

scale:  5.33

         Df Deviance AIC F value  Pr(>F)
<none>            64 157
location  4      914 308    39.5 8.1e-07
period    3      229 182    10.2  0.0013
```

- Ignoring overdispersion leads to wrong tests and P-values!

```
> drop1(bmod, test="LRT")
Single term deletions

Model:
cbind(survive, total - survive) ~ location + period
         Df Deviance AIC LRT Pr(>Chi)
<none>            64 157
location  4      914 998 849   <2e-16
period    3      229 315 164   <2e-16
```


Example overdispersion (4)

- Compare standard errors and tests for individual coefficients with and without handling overdispersion.

```
> coef(summary(bmod, dispersion=phi))  # extra scale parameter for variance used
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   4.6358     0.6495  7.1376 9.494e-13
location2    -0.4168     0.5682 -0.7335 4.632e-01
location3    -1.2421     0.5066 -2.4517 1.422e-02
location4    -0.9509     0.5281 -1.8004 7.180e-02
location5    -4.6138     0.5777 -7.9868 1.385e-15
period7      -2.1702     0.5504 -3.9429 8.051e-05
period8      -2.3256     0.5609 -4.1462 3.380e-05
period11     -2.4500     0.5405 -4.5330 5.815e-06
> coef(summary(bmod))                # without extra scale parameter; wrong!
            Estimate Std. Error z value  Pr(>|z|)
(Intercept)   4.6358     0.2813  16.479 5.191e-61
location2    -0.4168     0.2461  -1.694 9.035e-02
location3    -1.2421     0.2194  -5.660 1.511e-08
location4    -0.9509     0.2288  -4.157 3.230e-05
location5    -4.6138     0.2502 -18.439 6.339e-76
period7      -2.1702     0.2384  -9.103 8.775e-20
period8      -2.3256     0.2429  -9.573 1.043e-21
period11     -2.4500     0.2341 -10.466 1.243e-25
```

- Realize that by using extra scale parameter model is not any longer a binomial GLM, because no binomial distribution (nor other distribution within the GLM setting) exists with variance equal to $\phi mp(1-p)$.
  This type of GLM is called a quasi-binomial GLM.

Example overdispersion (5)

- Using `family=quasibinomial()` dispersion parameter is automatically estimated from Pearson's X^2, and tests are modified accordingly:

```
> bmod2 <- glm(cbind(survive, total-survive) ~ location + period, family=quasibinomial(), data=troutegg)
> coef(summary(bmod2))
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)   4.6358     0.6495  7.1376 1.184e-05
location2    -0.4168     0.5682 -0.7335 4.773e-01
location3    -1.2421     0.5066 -2.4517 3.050e-02
location4    -0.9509     0.5281 -1.8004 9.697e-02
location5    -4.6138     0.5777 -7.9867 3.824e-06
period7      -2.1702     0.5504 -3.9429 1.953e-03
period8      -2.3256     0.5609 -4.1462 1.356e-03
period11     -2.4500     0.5405 -4.5330 6.861e-04
> deviance(bmod2)
[1] 64.5
> drop1(bmod2,test="F")
Single term deletions

Model:
cbind(survive, total - survive) ~ location + period
         Df Deviance F value  Pr(>F)
<none>            64
location  4      914    39.5 8.1e-07
period    3      229    10.2  0.0013
```

Other approaches for handling overdispersion

- Use approach with extra scale parameter only if binomial totals are roughly equal for different covariate classes.
- Otherwise consider the following:
  - Use one of the overdispersion methods by Williams (available in R package `dispmod`)
  - Assume a distribution for the probability p of the binomial distribution, often the beta distribution, and "mix" this beta distribution and binomial distribution into compound distribution of y: beta-binomial distribution e.g. available in R package `glmmTMB` using `family=betabinomial`.
- Details are beyond scope of course.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 6, lecture 35

Content lecture 35: Poisson GLM (ELM 3, Fox 15.2, 15.2.1 partly)

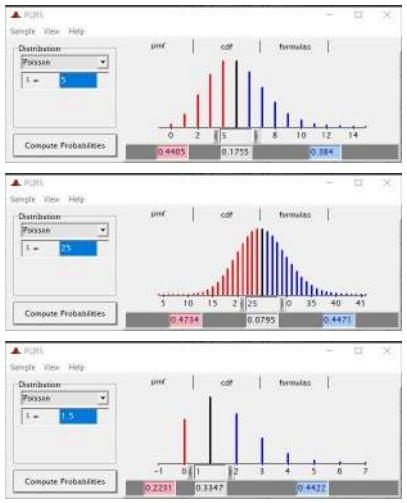
- Generalized linear model (GLM) for Poisson (count) data
  - count data
  - generalized linear model:
    - Poisson distribution; expectation and variance; some properties
    - linear predictor
    - link functions: log
  - maximum likelihood for Poisson glm
  - Poisson deviance
  - goodness of fit: deviance and Pearson's X^2
  - overdispersion
  - rate models, offset

Count regression

- Response is a count (positive integer).
- If total count is bounded, binomial distribution comes into picture.
- If counts are unbounded and sufficiently large, normal approximation (may be after square root transformation of count) may be justified.
- In other cases we may use Poisson distribution or negative binomial distribution, allowing for overdispersion.

Poisson regression

- Suppose $Y \sim Pois(\mu)$. Recall that $P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$.
- Recall further that $E(Y) = \mu$ and $var(Y) = \mu$. Variance is equal to mean!
- Below a few examples: $Pois(5)$, $Pois(25)$, $Pois(1.5)$.



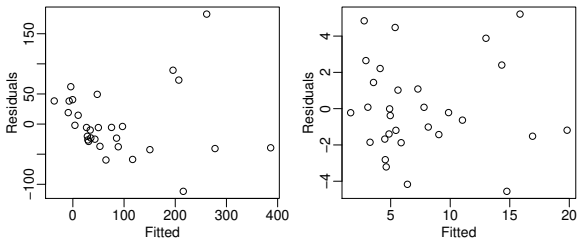
Poisson regression

- Poisson distribution arises naturally in several ways:
  - As approximation of binomial distribution if binomial n is large and p is small.
    E.g. modeling incidence of rare forms of cancer, number of affected people is small proportion of population.
  - Suppose probability of occurrence of event in given time interval is proportional to length of time interval, independent of occurrence of other events. Then number of events in any specified time interval has Poisson distribution.
    E.g. number of incoming telephone calls per day, number of earthquakes per month.
    Note, though, that rate of calls may not be constant over 24 h.
  - Suppose time between events is independent and identically exponentially distributed. Count the number of events in given time period. Count has Poisson distribution (equivalent to former case).
- Important property of independent Poisson variables: suppose $Y_i \sim Pois(\mu_i)$ for $i = 1, \ldots, n$ and Y_i independent; then $\sum_i Y_i \sim Pois(\sum_i \mu_i)$.
  If we only have access to aggregated data, but know that underlying counts have Poisson distributions, then sum has Poisson distribution.

Example count data

- Count: number of tortoise species per Galapagos island; 5 geographic variables.
- First use ordinary linear regression, make residual plots. Notice nonconstant variance. Notice negative $\hat{y}$.
- Next, transform response using variance stabilizing transformation $\sqrt{\cdot}$.

```
> modl <- lm(Species ~ ., gala)
> plot(predict(modl), residuals(modl), xlab="Fitted", ylab="Residuals")
> modt <- lm(sqrt(Species) ~ ., gala)
> plot(predict(modt), residuals(modt), xlab="Fitted", ylab="Residuals")
```



- Fit for transformed response doesn't look so bad: all predicted (transformed) responses are positive, and variance is approximately constant.
- But some counts are rather small. Can we rely on normal approximation? And interpretation of transformed variables is difficult.

Poisson regression

- Components of Poisson regression model:
  - Random part of model: $Y_i \sim Pois(\mu_i)$, independent
  - Fixed part of model: linear predictor $\eta_i = x_i^T \beta$ for vector of regressor values x_i for observation i.
  - Link function to link expected count μ_i to linear predictor η_i. As expected count $\mu > 0$, desirable link function transform interval $(0, \infty)$ into $(-\infty, \infty)$. Log is just doing that:
    $$log(\mu_i) = \eta_i = x_i^t \beta$$
    Log is canonical link function. A multiplicative model for μ becomes a linear model $log(\mu)$.
- Log-likelihood is
  $logL(\beta) = \sum_{i=1}^n [y_i \eta_i - e^{\eta_i} - log(y_i!)] = \sum_{i=1}^n [y_i x_i^T \beta - e^{x_i^T \beta} - log(y_i!)]$
- Find m.l.e. by differentiating w.r.t. β and solving $\sum_{i=1}^n (y_i - e^{x_i^T \hat{\beta}}) x_{ij} = 0$ for all j, or more compactly: $X^T y = X^T \hat{\mu}$.
- These equations resemble the normal equations from linear regression (where $\hat{\mu} = X\hat{\beta}$). We have this simple form only if the link function is canonical link function.
- Still, solving for β is not easy, and iterative numerical methods are needed (Newton-Raphson, IRWLS).

Example Poisson regression, continued

```
> modp <- glm(Species ~ ., family=poisson, data=gala)
> summary(modp)

Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.275  -4.497  -0.944   1.917  10.185

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.15e+00   5.17e-02   60.96  < 2e-16
Area        -5.80e-04   2.63e-05  -22.07  < 2e-16
Elevation    3.54e-03   8.74e-05   40.51  < 2e-16
Nearest      8.83e-03   1.82e-03    4.85  1.3e-06
Scruz       -5.71e-03   6.26e-04   -9.13  < 2e-16
Adjacent    -6.63e-04   2.93e-05  -22.61  < 2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.7

Number of Fisher Scoring iterations: 5
```


Poisson deviance

- Residual deviance is twice the difference in log-likelihood between the saturated and current model.
- Saturated model will produce as fitted values exactly $\hat{\mu}_i^{sat} = y_i$, whereas current model produces $\hat{\mu}_i$.
- Hence, deviance D is
  $2\sum_{i=1}^{n}[(y_i log(\hat{\mu}_i^{sat}) - \hat{\mu}_i^{sat} - log(y_i!)) - (y_i log(\hat{\mu}_i) - \hat{\mu}_i - log(y_i!))]$
- So, Poisson deviance is

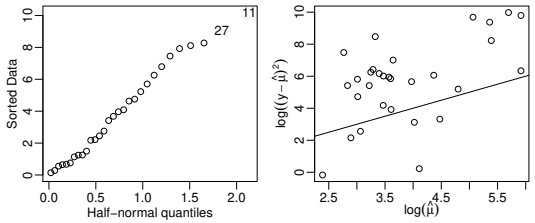
$$D = 2\sum_{i=1}^{n}[y_i log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)]$$

- Same inference as binomial model:
  - Check goodness of fit through residual deviance or (preferably) Pearson's $X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$, against χ^2_{n-p}.
  - Compare nested models by taking differences of deviances, and compare with $\chi^2_{p_2-p_1}$.
  - Test significance of individual predictors and construct c.i. using standard errors (Wald procedures), but likelihood methods are preferred.

Example Poisson regression continued

- Example residual deviance is 717 on 24 df; indicates that Poisson regression model does not fit well.
- Check residuals by half-normal plot. Outliers? [Faraway likes half-normal plots of residuals; in half-normal plot the absolute values of residuals are plotted against the quantiles from the half-normal distribution, i.e. the distribution of the absolute value of a standard-normally distributed variable].
- Study relationship between variance and mean (Poisson: variance=mean): plot (on log-scales) $(y_i - \hat{\mu}_i)^2$ (which served as crude approximation of variance of y_i given value of μ) against $\hat{\mu}_i$

```
> halfnorm(residuals(modp))
> plot(log(fitted(modp)), log((gala$Species-fitted(modp))^2), xlab=expression(log(hat(mu))),
+  ylab=expression(log((y-hat(mu))^2)))
> abline(0,1)
```



- Variance seems to be larger than mean. It could be that variance is proportional to mean. In that case the line would be shifted upwards on log-log plot.

Overdispersion Poisson regression

- Poisson distribution has single parameter, and very strict relationship between variance and mean, allowing no flexibility in variance structure.
- Like in binomial case, it is sometimes (often?) needed to allow for over- or underdispersion.
- Overdispersion can occur if Poisson response Y has rate λ, which itself has distribution, e.g. gamma distribution with $E\lambda = \mu$ and $var(\lambda) = \mu/\phi$. Then Y has negative binomial distribution with mean $EY = \mu$, and $var(Y) = \mu(1+\phi)/\phi$ which is larger than μ, hence overdispersion relative to Poisson.
- Easy, but simplistic, way out is, like in binomial case, to introduce extra scale parameter ϕ:
  - $var(Y) = \phi E(Y) = \phi\mu$
  - $\phi > 1$: overdispersion, $\phi < 1$: underdispersion
  - Estimate ϕ from Pearson's X^2: $\hat{\phi} = \frac{X^2}{n-p} = \frac{\sum_i (y_i - \hat{\mu}_i)^2/\hat{\mu}_i}{n-p}$
  - Adjust standard errors, use F-tests based upon deviance differences instead of χ^2 tests.

Example Poisson regression continued

```
> (phi <- sum(residuals(modp, type="pearson")^2)/modp$df.res)
[1] 31.75
> summary(modp, dispersion=phi)
Call:
glm(formula = Species ~ ., family = poisson, data = gala)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.275  -4.497  -0.944   1.917  10.185

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.154808   0.291590   10.82  < 2e-16
Area        -0.000580   0.000148   -3.92  8.9e-05
Elevation    0.003541   0.000493    7.19  6.5e-13
Nearest      0.008826   0.010262    0.86     0.39
Scruz       -0.005709   0.003525   -1.62     0.11
Adjacent    -0.000663   0.000165   -4.01  6.0e-05

(Dispersion parameter for poisson family taken to be 31.75)

    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  716.85  on 24  degrees of freedom
AIC: 889.7

Number of Fisher Scoring iterations: 5
```


Example Poisson regression continued

```
> (modp2 <- glm(Species ~ ., family=quasipoisson(), data=gala))
Call:  glm(formula = Species ~ ., family = quasipoisson(), data = gala)

Coefficients:
(Intercept)         Area    Elevation      Nearest        Scruz     Adjacent
   3.154808    -0.000580     0.003541     0.008826    -0.005709    -0.000663

Degrees of Freedom: 29 Total (i.e. Null);  24 Residual
Null Deviance:          3510
Residual Deviance: 717          AIC: NA
> drop1(modp2, test="F")
Single term deletions

Model:
Species ~ Area + Elevation + Nearest + Scruz + Adjacent
          Df Deviance F value  Pr(>F)
<none>           717
Area       1     1204   16.32 0.00048
Elevation  1     2390   56.00   1e-07
Nearest    1      739    0.76 0.39336
Scruz      1      814    3.24 0.08444
Adjacent   1     1341   20.91 0.00012
```


Rate models

- Number of events may depend on some size variable. E.g. number of burglaries reported in different cities, will depend on number of households in cities. Size variable could also be time.
- Sometimes binomial distribution could be used if totals are known (like e.g. in example of burglaries with known city sizes)
- Example `dicentric`: effect of gamma radiation on counts of chromosomal abnormalities ca. Numbers of exposed cells (in hundreds) `cells` differed. Predictors are dose amount and rate.
  - One way to analyze would be binomial glm for `ca / cells`.
  - Alternative would be Poisson glm, assuming that $\mu = E(ca)$ is proportional to `cells`: starting with twice the number of cells, would result on average in twice as many abnormal cells. So, with μ_0 the expected number of abnormal cells per starting cell: $\mu = cells \times \mu_0$, so that $log(\mu) = log(cells) + log(\mu_0)$.
- Rate model for counts: $y \sim Pois(\mu)$ with $log(\mu) = log(cells) + log(\mu_0) = log(cells) + \beta x$.
- Note that $log(cells)$ is a regressor with regression coefficient identical to 1. This is called an offset. Other regression coefficient(s) β have to be estimated as before.

Example Rate model

```
> rmod <- glm(ca ~ offset(log(cells)) + log(doserate)*factor(doseamt), family=poisson, data=dicentric)
> summary(rmod)
Call:
glm(formula = ca ~ offset(log(cells)) + log(doserate) * factor(doseamt),
    family = poisson, data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4910  -0.6247  -0.0508   0.7679   1.5912

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -2.7467     0.0343  -80.16  < 2e-16
log(doserate)                     0.0718     0.0352    2.04  0.04130
factor(doseamt)2.5                1.6254     0.0495   32.86  < 2e-16
factor(doseamt)5                  2.7611     0.0435   63.49  < 2e-16
log(doserate):factor(doseamt)2.5  0.1612     0.0483    3.34  0.00084
log(doserate):factor(doseamt)5    0.1935     0.0424    4.56  5.1e-06

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4753.00  on 26  degrees of freedom
Residual deviance:   21.75  on 21  degrees of freedom
AIC: 209.2

Number of Fisher Scoring iterations: 4
```

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

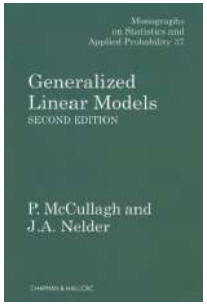
Biometris,
Wageningen University, Wageningen, The Netherlands

Week 6, lecture 36

Content lecture 36: GLM theory (ELM §6, Fox §15.3-15.4)

- exponential family, canonical parameter, dispersion parameter
- mean, variance function, variance
- canonical link function
- iterated reweighted least squares
- inference: LRT using deviance
- (scaled) deviance
- diagnostics: residuals (response, Pearson, deviance), leverage and influence

GLM theory



- Bible of GLM: McCullagh and Nelder (1989)
- GLM specified by 3 components:
  1. Random part of model: specification of distribution of response variable y; distribution comes from exponential family.
  2. Systematic part of model: linear predictor $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = x^T\beta$
  3. Link function g, which links mean $\mu = E(y)$ to linear predictor: $g(\mu) = \eta$.
- Exponential family of distributions:

$$f(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

- Canonical parameter: θ, representing the location.
- Dispersion parameter: ϕ, representing the scale.

Example normal distribution

- Distribution in exponential family:

$$f(y|\theta,\phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right]$$

- E.g. normal distribution:

$$f(y|\theta,\phi) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] = exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}(\frac{y^2}{\sigma^2} + log(2\pi\sigma^2))\right]$$

  - canonical parameter $\theta = \mu$
  - dispersion parameter $\phi = \sigma^2$
  - $a(\phi) = \phi$
  - $b(\theta) = \theta^2/2$
  - $c(y,\phi) = -(y^2/\phi + log(2\pi\phi))/2$.

GLM theory

- Example Poisson distribution (try yourself)
  $f(y|\theta,\phi) = \ldots$
  - canonical parameter $\theta = \ldots$
  - dispersion parameter $\phi = \ldots$
  - $a(\phi) = \ldots$
  - $b(\theta) = \ldots$
  - $c(y,\phi) = \ldots$
- Example binomial distribution (try yourself)
  $f(y|\theta,\phi) = \ldots$
  - canonical parameter $\theta = \ldots$
  - dispersion parameter $\phi = \ldots$
  - $a(\phi) = \ldots$
  - $b(\theta) = \ldots$
  - $c(y,\phi) = \ldots$
- Less common members of exponential family: gamma and inverse Gaussian distribution.
- Exponential and χ^2-distributions are special cases of the gamma distribution.
- Other distributions, like negative binomial and Weibull distribution, are not members of exponential family. Still GLMs can be fitted with some modifications.

Mean and variance in exponential family

Let Y has distribution in exponential family. Then

- $E(Y)(=\mu) = b'(\theta)$
- $var(Y) = b''(\theta)a(\phi)$.

Hence:

- Mean is function of θ only
- Variance is product of functions of location and scale
- $b''(\theta)$ is called variance function $V(\mu)$, and describes how variance depends on mean.

These properties follow from likelihood results:

- $E\left(\frac{\partial logL}{\partial \theta}\right) = 0$. Note: $\frac{\partial logL}{\partial \theta}$ is the score. So: expected score is zero.
  This gives the result for EY. (Why? Hint: fill formula for exponential family distribution into the score and work out.)
- $E\left(\frac{\partial^2 logL}{\partial \theta^2}\right) + E\left(\left(\frac{\partial logL}{\partial \theta}\right)^2\right) = 0.$
  Taking second order derivatives and expectations gives result for $var(Y)$ (why?).
- Gaussian case is rather special, because $b''(\theta) = 1$: variance does not depend on mean.
- Weights may be introduced by setting $a(\phi) = \phi/w$, with w known weights, possibly varying between observations.

Linear predictor and Link function

- Effect of regressor enters model through linear predictor: $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
- Link function g describes how mean response $E(Y) = \mu$ is linked to regressors through linear predictor $g(\mu) = \eta$.
- In principle, any monotone continuous and differentiable function will do, but convenient and common choice for standard GLMs exist:
  - Gaussian GLM: identity link $\eta = g(\mu) = \mu$.
    Another choice would give $\mu = g^{-1}(\eta)$, so that $y = g^{-1}(\eta) + \epsilon$.
    Notice that this not same as transforming the response $g(y) = \eta + \epsilon$.
  - Poisson GLM: because an expected count must be positive, standard choice for inverse link is $\mu = e^{\eta}$, giving the canonical log-link function: $\eta = log(\mu)$. With this link function additive effects of x lead to multiplicative effects on μ.
  - Binomial GLM: because expected fraction is in the interval $(0,1)$, a preferred link function transforms the interval $(0,1)$ to $(-\infty,\infty)$.
    Common choices: logistic link, probit-link, and complementary log-log link.

Canonical link functions

- Canonical link function is link function that follows naturally from exponentially family notation: it is the function g that transform μ to the scale of the canonical parameter θ. Because $\mu = b'(\theta)$, the canonical link function must be $g = (b')^{-1}$.
  Examples:

Family	$b(\theta)$	$b'(\theta) = \mu$	Can.link $\eta = (b')^{-1}(\mu)$	Var.function $(b''(\theta))$
Normal	$\theta^2/2$	θ	$\eta = \mu$	1
Poisson	e^{θ}	e^{θ}	$\eta = log(\mu)$	μ
Binomial	$log(1+e^{\theta})$	$\frac{e^{\theta}}{1+e^{\theta}}$	$\eta = log(\mu/(1-\mu))$	$\mu(1-\mu)$
Gamma			$\eta = \mu^{-1}$	μ^2
Inv. Gaussian			$\eta = \mu^{-2}$	μ^3

- If canonical link is used, then $X^T Y$ is sufficient for β.
- Canonical link is mathematically and computationally efficient, and often natural choice, but not required.

Fitting GLM: IRWLS

- Parameters are estimated using maximum likelihood.
- Log likelihood for single observation, with $a_i(\phi) = \phi/w_i$ is

$$logL(\theta_i, \phi; y_i) = w_i \left[\frac{y_i\theta_i - b(\theta_i)}{\phi} \right] + c(y_i, \phi)$$

  and for independent observations log likelihood is $\sum_i logL(\theta_i, \phi; y_i)$.
- Only in Gaussian GLM can we maximize analytically and find exact solution.
- In other cases use numerical optimization: Newton-Raphson method (for finding roots of functions, here: root of derivative of log likelihood) with Fisher scoring (using expected Fisher information matrix), is equivalent to IRWLS = iteratively reweighted least squares.
- In GLM we have a linear model, but it sits on the scale of the linear predictor: $g(\mu) = X\beta$, whereas response is on different scale. Another complication is that variance is not constant, but we already know how to handle that in linear models: use weighted least squares.

Fitting GLM: IRWLS continued

- To arrive on the scale of linear predictor, we may calculate $g(y)$, and regress it on X. However, $g(y)$ may not exist, like in binomial case if $y = 0$.
- Instead, linearize $g(y)$ using Taylor expansion around current value of μ:

$$g(y) \approx g(\mu) + (y - \mu)g'(\mu) = \eta + (y - \mu)\frac{d\eta}{d\mu} \equiv z$$

- Further, $\widehat{var}(z) = \left(\frac{d\eta}{d\mu}\right)^2 V(\hat{\mu}) = \frac{1}{w}$.
- Here is the IRWLS algorithm:
  1. Set initial estimates $\hat{\eta}_0$ (from which $\hat{\mu}_0$ follow).
  2. Form "adjusted dependent variable" $z_0 = \hat{\eta}_0 + (y - \hat{\mu}_0)\frac{d\eta}{d\mu}|_{\hat{\mu}_0}$.
  3. Form weights w_0 with $w_0^{-1} = \left(\frac{d\eta}{d\mu}\right)^2 |_{\hat{\mu}_0} V(\hat{\mu}_0)$.
  4. Regress z_0 on X using weights w_0 to get (new) estimates of β.
  5. Get new $\hat{\eta}_1$ and $\hat{\mu}_1$
  6. Iterate steps 2-5 until convergence.
- Only $\eta = g(\mu)$ and $V(\mu)$ are needed, no further distributional information!
- Estimator of variance-covariance matrix of regression coefficients:

$$\hat{V}(\hat{\beta}) = (X^T W X)^{-1}\hat{\phi}$$

- Follow example on pp 118-119 during practical.

Null - Current - Saturated models

- Null model is smallest model possible (that we will entertain), and full or saturated model is most complex.
- Null model represents situation where there is no relationship between regressors and response:
  - Usually common mean μ is fitted for all y (intercept-only-model).
  - Sometimes (e.g. contingency tables) additional parameters enter the null model to get the right marginal totals.
- In saturated model data is reproduced exactly by model:
  - Typically, n parameters needed for n observations, e.g. by treating numerical values of quantitative regressors as factor levels.
  - Saturated model lacks important characteristic of models: it is not a simplification, and in that respect is uninformative.
- Statistical model describes how we partition data into systematic structure and random variation:
  - One extreme is null model: data is represented entirely by random variation.
  - Other extreme is saturated model: data is represented entirely by systematic structure.

Goodness of fit

- Saturated model gives measure how well any model could possibly fit. Therefore consider difference between log-likelihood of saturated model ($\ell(y, \phi|y)$) and current model ($\ell(\hat{\mu}, \phi|y)$), expressed as likelihood ratio statistic:

$$2(\ell(y, \phi|y) - \ell(\hat{\mu}, \phi|y))$$

- With independent observations and exponential family distributions when $a_i(\phi) = \phi/w_i$ this simplifies to:

$$\sum_i 2w_i(y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i))/\phi$$

  with $\tilde{\theta}_i$ estimator under saturated model, and $\hat{\theta}_i$ under current model.
- Can be written as $D(y, \hat{\mu})/\phi$.
- $D(y, \hat{\mu})$ is called deviance, whereas $D(y, \hat{\mu})/\phi$ is called scaled deviance.
- Another measure of discrepancy is Pearson's X^2:

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Hypothesis tests (1)

- Two main types of hypothesis tests:
  - Goodness of fit test, checking whether current model fits data
  - Comparison of two nested models, where smaller model (ω) is obtained by putting linear restriction(s) on parameters of larger model (Ω).
- Goodness of fit tests:
  - Under certain conditions scaled deviance and Pearson's X^2 are asymptotically χ^2_{n-p} distributed.
  - For GLMs like Gaussian, value of dispersion parameter ϕ is not known, and this test cannot be used.
  - For binomial and Poisson distribution $\phi = 1$, so goodness of fit test can be used. It is dubious for smaller datasets, and useless in binary case.

Hypothesis tests (2)

- Comparison of larger model (Ω) and nested smaller model (ω):
  - Difference of two scaled deviances $D_\omega - D_\Omega$ is asymptotically $\chi^2_{\Delta df}$ with Δdf the difference in numbers of residual degrees of freedom ω and Ω (or equivalenty, difference in numbers of model parameters of Ω and ω).
  - For Gaussian and others where dispersion parameters has to be estimated by $\hat{\phi}$, instead an F-statistic can be computed:

$$F = \frac{(D_\omega - D_\Omega)/\Delta df}{\hat{\phi}}$$

    with $\hat{\phi} = X^2/(n-p)$ estimator of dispersion.
  - For Gaussian situation $\hat{\phi} = RSS_\Omega/df_\Omega$, and resulting F-statistic has exact F-distribution under null hypothesis.

Example Hypothesis tests

```
> modl <- glm(cbind(dead,alive) ~ conc, family=binomial, data=bliss)
> summary(modl)

Call:
glm(formula = cbind(dead, alive) ~ conc, family = binomial, data = bliss)

Deviance Residuals:
      1       2       3       4       5
-0.4510  0.3597  0.0000  0.0643 -0.2045

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -2.324      0.418   -5.56  2.7e-08
conc           1.162      0.181    6.40  1.5e-10

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 64.76327  on 4  degrees of freedom
Residual deviance:  0.37875  on 3  degrees of freedom
AIC: 20.85

Number of Fisher Scoring iterations: 4

> 1-pchisq(deviance(modl), df.residual(modl))

[1] 0.9446
```


Example Hypothesis tests (2)

```
> anova(modl, test="Chi")

Analysis of Deviance Table

Model: binomial, link: logit

Response: cbind(dead, alive)

Terms added sequentially (first to last)


     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                     4       64.8
conc  1     64.4         3        0.4    1e-15

> coef(summary(modl))[2,] # Wald test

  Estimate Std. Error    z value   Pr(>|z|)
 1.162e+00  1.814e-01  6.405e+00  1.508e-10

> modl2 <- glm(cbind(dead,alive) ~ conc + I(conc^2), family=binomial, data=bliss)
> anova(modl,modl2,test="Chi")

Analysis of Deviance Table

Model 1: cbind(dead, alive) ~ conc
Model 2: cbind(dead, alive) ~ conc + I(conc^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3      0.379
2         2      0.195  1    0.183     0.67
```

GLM diagnostics: Residuals

- Check adequacy of assumptions underlying GLM as is done in linear models.
- Residuals represent differences between responses and fitted values.
- $\hat{\epsilon} = y - \hat{\mu}$ are called response residuals. Their variance is not constant.
- Pearson residuals are comparable to standardized residuals used for linear models:

$$r_P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}}$$

  - Rescaling of response residuals
  - Notice that $\sum r_P^2 = X^2$
  - Can be skewed for non-normal responses.
- Deviance residuals r_D are set up such that $\sum r_D^2 = Deviance = \sum d_i$: $r_D = sign(y - \hat{\mu})\sqrt{d_i}$

```
> r.d <- residuals(modl)             # deviance residuals
> r.p <- residuals(modl, "pearson")
> r.r <- residuals(modl, "response")
> r.w <- residuals(modl, "working") # by-product of IRWLS from weighted least squares
> r.w2 <- modl$residuals            # gives working residuals; beware!
```


GLM diagnostics: Leverage and influence

- Linear model: $\hat{y} = Hy$ with H the hat matrix and leverages h_I on diagonal, representing potential of observation i to influence fit.
- In GLM slightly different: IRWLS uses weights w_i, which do affect leverage. Hat matrix $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$ with $W = diag(w_i)$ containing leverages on diagonal.
- Large leverage typically means that regressor values are unusual in some way.
- Difference with linear model: h_i also depends on response through weights.

```
> h <- influence(modl)$hat
```

- Residuals may be studentized:

$$r_{SD} = \frac{r_D}{\sqrt{\hat{\phi}(1 - h_i)}}$$

- Jackknife residuals by exclusion of case i are rather expensive. Instead approximation due to Williams could be used, in R available under:

```
> r.s <- rstudent(modl)
```

- Leverage measures potential influence. Measures of influence directly assess effect on fit, e.g. by

```
> i.c. <- influence(modl)$coef
> c.d <- cooks.distance(modl)
```


GLM diagnostics: Diagnostic methods

- Two types:
  1) Detect single case or small groups of cases that do not fit the pattern of rest: e.g. outlier detection.
  2) Check assumptions of model, which can be split into:
  2a) Check structural form of model, like choice and transformation of regressors
  2b) Check stochastic part of model, like nature of variance about mean response.
- Methods for checking assumptions of model:
  - Linear model: plot residuals versus fitted values.
  - GLM: plot residuals versus linear predictor $\hat{\eta}$ or fitted values $\hat{\mu}$. Generally better to plot linear predictor $\hat{\eta}$.
    - Check plot for nonlinear relationship. If present, change choice of regressors or transformation of regressors.
    - Check variance of residuals with respect to fitted values. Possibly change the variance function (and the distribution along with, maybe quasi-likelihood), or use weights. In deviance and Pearson residuals variance function is scaled out, so provided a correct variance function, constant variance is expected in plot.
  - With binary response, residuals plots are not very helpful, as only two values for given predicted response. Less extreme cases (binomial with small n and low Poisson counts) show milder artefacts.
  - Plot response against regressors, even before fitting model.
  - Partial residual plot ("Components plus residual" plots) are possible.

GLM diagnostics: half-normal plots

- Normal model: use QQ-plot of residuals to check normality. GLM: we do not expect residuals to be normally distributed, but detection of outliers is needed. Use half-normal plot to this end: compares sorted absolute residuals and quantiles of half-normal distribution. Don't look for straight line, but look for outliers to be identified as points off the trend. Best to use jack-knife residuals.
- Half-normal plots can also be used for necessarily positive quantities, like leverages and Cook's distances.

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 7, lecture 37

Content week 7, lecture 37: gamma GLM Far ELM: §7.1

- Analyzing positive continuous responses: gamma GLM
  - gamma distribution: density, parameters, canonical link, variance function
  - gamma GLM vs LM with log-transformation of y

Gamma GLM: gamma distribution

- Let y be a positive continuous response variable ($y > 0$).
- Gamma distribution may describe random behaviour of such variable y. Density of gamma distribution:

$$f(y) = \frac{1}{\Gamma(\nu)} \lambda^{\nu} y^{\nu-1} e^{-\lambda y}$$

  with ν shape parameter, and λ rate parameter. R calls $1/\lambda$ the scale parameter.
- $\Gamma(x)$ is the so-called gamma function, taking care that area under density equals 1; for integer n, $\Gamma(n) = (n-1)!$.
- $EY = \mu = \nu/\lambda$ and $var(Y) = \nu/\lambda^2 = (1/\nu)(\nu/\lambda)^2 = (1/\nu)\mu^2$
- For GLM purposes reparametrization is useful, putting $\lambda = \nu/\mu$ to get

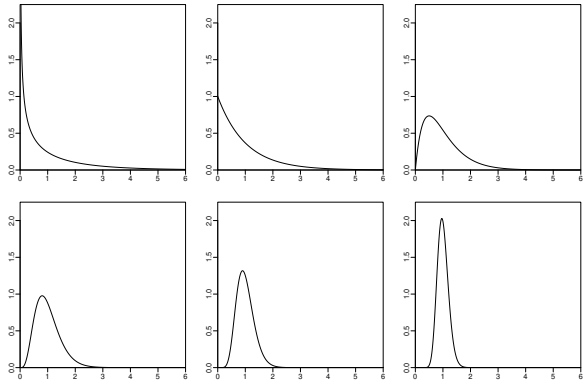
$$f(y) = \frac{1}{\Gamma(\nu)} \left(\frac{\nu}{\mu}\right)^{\nu} y^{\nu-1} e^{-\frac{\nu}{\mu} y}$$

  with $E(Y) = \mu$ and $\text{var}(Y) = \nu^{-1}\mu^2$. Dispersion parameter is $\phi = \nu^{-1}$.

Shapes of gamma densities

- Plot gamma densities with shape parameters ν=0.5, 1, 2, 5, 10, 25. By choosing $\lambda = \nu$, $E(Y) = \mu = \nu/\lambda = 1$. As $\text{var}(Y) = \nu^{-1}\mu^2$, variances are 2, 1, 0.5, 0.2, 0.1, 0.04.

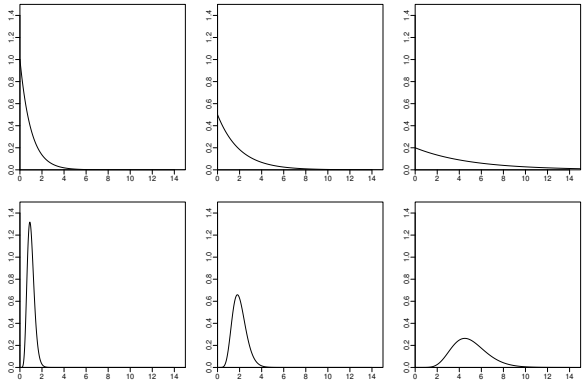
```
> x <- seq(0,6,by=0.01)
> plot(x, dgamma(x,shape=0.5,rate=0.5), type="l",ylab="", xlab="",ylim=c(0,2.25),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=1,  rate=1),   type="l",ylab="", xlab="",ylim=c(0,2.25),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=2,  rate=2),   type="l",ylab="", xlab="",ylim=c(0,2.25),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=5,  rate=5),   type="l",ylab="", xlab="",ylim=c(0,2.25),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=10, rate=10),  type="l",ylab="", xlab="",ylim=c(0,2.25),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=25, rate=25),  type="l",ylab="", xlab="",ylim=c(0,2.25),xaxs="i",yaxs="i")
```



Variance-mean relationship

- Keeping shape ν constant, but increasing μ (so, decreasing λ), gamma distributions shift to the right and are wider ($\text{var}(Y) = \nu^{-1}\mu^2$). Top 3 plots: $\nu = 1$; bottom 3 plots: $\nu = 10$; both: $\mu = 1, 2, 5$. Gamma GLM functions: assuming constant shape ν, model mean μ as function of explanatory variables.

```
> x <- seq(0,15,by=0.01)
> plot(x, dgamma(x,shape=1,  rate=1),   type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=1,  rate=0.5), type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=1,  rate=0.2), type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=10, rate=10),  type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=10, rate=5),   type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dgamma(x,shape=10, rate=2),   type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
```



Gamma GLM: distribution (2)

- Some further properties
  - Exponential distribution with rate λ is gamma distribution with shape parameter $\nu = 1$ and rate parameter λ.
  - Sum of ν independent and identically distributed exponential random variables with rate λ has gamma distribution with shape parameter ν and rate parameter λ.
  - χ^2 distribution is special case of gamma where $\lambda = 1/2$ and $\nu = df/2$.
- From exponential family notation: $b(\theta) = -\log(-\theta)$
- First derivative gives the mean: $\mu = b'(\theta) = -1/\theta$. Inverting this relationship gives canonical link function: $g(\mu) = \eta = -1/\mu$. Minus is removed, inverse link is used.
- Second derivative gives variance function: $V(\mu) = b''(\theta) = \mu^2$; so, the variance function is the square.
- (Unscaled) deviance is $D(y, \hat{\mu}) = -2\sum\{\log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)/\hat{\mu}_i\}$.

Gamma GLM or transform?

- Gamma GLM may be used for continuous, positive, right-skewed response y, but also if relationship between mean and variance of response (namely: variance proportional to square of mean) points towards it.
- Suppose you prefer a Gaussian linear model, but variance is not constant. Options:
  - If variances (up to a constant factor) are known, weighted least squares could be used.
  - Transformation of Y may help, e.g. $\log Y$ or $\sqrt{Y}$.
    - Suppose variance-mean relationship is $\text{var}Y \propto EY$, then use $\sqrt{Y}$ as variance stabilizing transformation.
    - GLM version of this variance-mean relationship found with Poisson distribution. Fact that Poisson is discrete distribution is not problematic, because fitting GLM only depends on mean and variance of distribution. Other characteristics of distribution are not used.
    - Suppose variance-mean relationship is $\text{var}Y \propto (EY)^2$, then use $\log Y$ as variance stabilizing transformation. $\text{SD}(Y) \propto E(Y)$ is equivalent to constant coefficient of variation $\text{CV} = \text{SD}(Y)/E(Y)$. Fitting Gaussian linear model on log-transformed data, implies a lognormal distribution for original response.
    - GLM version of this variance-mean relationship found with gamma distribution.

Gamma GLM: link functions

- Canonical link function is reciprocal $\eta = \mu^{-1}$.
  - Note that with this link function a negative linear predictor value is not transformed back to allowed range for μ: $(0, \infty)$. This could cause problems and might require restrictions on β or on range of possible predictor values.
  - Sometimes the reciprocal link is advantageous, like in the Michaelis-Menten model:
    $$E(y) = \mu = \frac{\alpha_0 x}{1 + \alpha_1 x}$$
    that can be reexpressed as $\mu^{-1} = \eta = \frac{\alpha_1}{\alpha_0} + \frac{1}{\alpha_0}\frac{1}{x}$. If x increases, $\eta \to \frac{\alpha_1}{\alpha_0}$, so mean response is bounded.
- log link: $\eta = \log \mu$. Use this link when effect of predictors is suspected to be multiplicative on mean. If variance is small, the approach is similar to Gaussian model with logged response.
- identity link: $\eta = \mu$. Useful for modelling sums of squares or variance components which have χ^2 distributions.
- Estimation of the dispersion parameter $\phi = \nu^{-1}$ preferably from Pearson's X^2 (according to McCullagh & Nelder): $\hat{\phi} = \frac{1}{\hat{\nu}} = \frac{X^2}{n-p}$.
  Maximum likelihood estimator and estimator from deviance $D/(n-p)$ sensitive to unusually small values of response.

Example: semiconductors

- Study of waver resistivity (of semiconductors): 4 factors at two levels, full factorial design; from previous experience a right-skewed response was expected. Box-Cox method or past experience suggest log transformation of response.

```
> llmdl <- lm(log(resist) ~ .^2, wafer)
> rlmdl <- step(llmdl)
> summary(rlmdl)

Call:
lm(formula = log(resist) ~ x1 + x2 + x3 + x4 + x1:x3 + x2:x3 +
    x3:x4, data = wafer)

Residuals:
     Min       1Q   Median       3Q      Max
-0.0812 -0.0365 -0.0004  0.0386  0.0836

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3111     0.0476  111.53  4.7e-14
x1+           0.2009     0.0476    4.22  0.00292
x2+          -0.2107     0.0476   -4.43  0.00221
x3+           0.4372     0.0673    6.49  0.00019
x4+           0.0354     0.0476    0.74  0.47892
x1+:x3+      -0.1562     0.0673   -2.32  0.04896
x2+:x3+      -0.1782     0.0673   -2.65  0.02941
x3+:x4+      -0.1830     0.0673   -2.72  0.02635

Residual standard error: 0.0673 on 8 degrees of freedom
Multiple R-squared:  0.947,        Adjusted R-squared:  0.901
F-statistic: 20.5 on 7 and 8 DF,  p-value: 0.000165
```


Example: semiconductors (2)

- Now fit gamma GLM; use family name `Gamma`

```
> gmdl <- glm(resist ~ .^2, family=Gamma(link=log), wafer)
> rgmdl <- step(gmdl)
> summary(rgmdl)

Call:
glm(formula = resist ~ x1 + x2 + x3 + x4 + x1:x3 + x2:x3 + x3:x4,
    family = Gamma(link = log), data = wafer)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.08319  -0.03679  -0.00065   0.03820   0.08139

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.3120     0.0476  111.68  4.6e-14
x1+           0.2003     0.0476    4.21  0.00295
x2+          -0.2110     0.0476   -4.44  0.00218
x3+           0.4367     0.0673    6.49  0.00019
x4+           0.0354     0.0476    0.74  0.47836
x1+:x3+      -0.1555     0.0673   -2.31  0.04957
x2+:x3+      -0.1763     0.0673   -2.62  0.03064
x3+:x4+      -0.1819     0.0673   -2.70  0.02687

(Dispersion parameter for Gamma family taken to be 0.004525)

    Null deviance: 0.697837  on 15  degrees of freedom
Residual deviance: 0.036266  on  8  degrees of freedom
AIC: 139.2

Number of Fisher Scoring iterations: 4
```


Example: some remarks

- Coefficients and standard errors in two models are remarkably similar.
- Also root of dispersion corresponds to residual standard deviation of linear model:

```
> sqrt(summary(rgmdl)$dispersion)  # square root of estimated dispersion parameter
[1] 0.06727
> summary(rlmdl)$sigma             # residual standard deviation from linear model
[1] 0.06735
```

- Maximum likelihood estimate of ϕ much smaller:

```
> library(MASS)
> gamma.dispersion(rgmdl)
[1] 0.002266
> sqrt(gamma.dispersion(rgmdl))
[1] 0.0476
```

- In this example shape $\nu = 1/\phi = 1/0.004525 = 221$ is large, and gamma distribution is well approximated by normal.
  Likewise lognormal distribution with small variance is well approximated by normal.
  Not much difference between the two then.
- Gamma GLM has advantage of modelling response directly, lognormal model has advantage of working with standard linear model (but for log transformed response).

Motor ensurance example: gamma GLM

- Swedish data on payment of ensurance claims, predicted from mileage, bonus (yes/no) and type of car. Total amount of claims supposed to be proportional to number insured; therefore treat log(number insured) as offset.

```
> g.glm <- glm(Payment ~ offset(log(Insured)) + Kilometres + Make + Bonus, family=Gamma(link=log), motori)
> summary(g.glm)
Call:
glm(formula = Payment ~ offset(log(Insured)) + Kilometres + Make +
    Bonus, family = Gamma(link = log), data = motori)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.037  -0.635  -0.096   0.265   2.703

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.5273     0.1777   36.72  < 2e-16
Kilometres    0.1201     0.0311    3.85  0.00014
Make2         0.4070     0.1782    2.28  0.02313
Make3         0.1553     0.1796    0.87  0.38767
Make4        -0.3439     0.1915   -1.80  0.07355
Make5         0.1447     0.1810    0.80  0.42473
Make6        -0.3456     0.1782   -1.94  0.05352
Make7         0.0614     0.1824    0.34  0.73689
Make8         0.7504     0.1873    4.01  7.9e-05
Make9         0.0320     0.1782    0.18  0.85778
Bonus        -0.2007     0.0215   -9.33  < 2e-16

(Dispersion parameter for Gamma family taken to be 0.556)

    Null deviance: 238.97  on 294  degrees of freedom
Residual deviance: 155.06  on 284  degrees of freedom
AIC: 7168
```

Motor ensurance example: lognormal model

- We use the `glm` function with `gaussian` family.

```
> lgn.glm <- glm(log(Payment) ~ offset(log(Insured)) + Kilometres + Make + Bonus, family=gaussian, motori)
> summary(lgn.glm)
Call:
glm(formula = log(Payment) ~ offset(log(Insured)) + Kilometres +
    Make + Bonus, family = gaussian, data = motori)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2852  -0.4176   0.0273   0.4614   2.4738

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.51403    0.18634   34.96  < 2e-16
Kilometres   0.05713    0.03265    1.75   0.0813
Make2        0.36387    0.18686    1.95   0.0525
Make3        0.00692    0.18824    0.04   0.9707
Make4       -0.54786    0.20076   -2.73   0.0067
Make5       -0.02179    0.18972   -0.11   0.9087
Make6       -0.45881    0.18686   -2.46   0.0147
Make7       -0.32118    0.19126   -1.68   0.0942
Make8        0.20958    0.19631    1.07   0.2866
Make9        0.12545    0.18686    0.67   0.5025
Bonus       -0.17806    0.02254   -7.90  6.2e-14

(Dispersion parameter for gaussian family taken to be 0.611)

    Null deviance: 238.56  on 294  degrees of freedom
Residual deviance: 173.53  on 284  degrees of freedom
AIC: 704.6

Number of Fisher Scoring iterations: 2
```


Motor ensurance example: compare gamma GLM with log transformation

- Differences quite large now. E.g. mileage significant in gamma GLM, but not in lognormal model.
- Two models are not nested and have different distributions for response, making direct comparison difficult.
- *AIC* ($= -2$maximized log likelihood $+ 2p$) (smaller is better) cannot be used, because of common practice to discards parts of likelihood that are not functions of parameters. This is ok for comparing models with same distribution. For comparison of different distributions it is essential that all parts of likelihood are retained. Here *AIC* values for two models are quiet different, hence precaution not taken.
- Taking differences between residual deviance and corresponding null deviance, results in larger "jump" for gamma GLM, indicating preference for gamma GLM here:

```
> (summary(g.glm)$null.deviance - summary(g.glm)$deviance)
[1] 83.92
> (summary(lgn.glm)$null.deviance - summary(lgn.glm)$deviance)
[1] 65.04
```

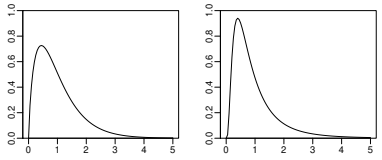

Motor ensurance example: comparing shapes of distributions

- Dispersion parameters from the two models:

```
> (disp.g.glm <- summary(g.glm)$dispersion)
[1] 0.556
> (disp.lgn.glm <- summary(lgn.glm)$dispersion)
[1] 0.611
```

- Gamma distribution has shape $\hat{\nu} = 1/0.556 = 1.80$, so quite skewed distribution. Below, the rate λ is given the same value, so that expected value $\mu = 1$.
- For the linear model after log-transformation, residual variance is $\hat{\sigma}^2_\epsilon = 0.611$. To plot the log-normal distribution, the mean after log-transformation μ is specified so that the expected value for the log-normal distributed response equals 1: $\exp(\mu + \frac{1}{2}\hat{\sigma}^2_\epsilon) = 1 \Leftrightarrow \mu + \frac{1}{2}\hat{\sigma}^2_\epsilon = 0 \Leftrightarrow \mu = -\frac{1}{2}\hat{\sigma}^2_\epsilon = -\frac{1}{2}0.611$.
- Comparing shapes of distributions: lognormal more peaked.

```
> x <- seq(0,5,by=0.05)
> plot(x, dgamma(x,shape=1/disp.g.glm, rate=1/disp.g.glm), type="l",ylab="", xlab="",ylim=c(0,1))
> plot(x, dlnorm(x,meanlog=-0.5*disp.lgn.glm, sdlog=sqrt(disp.lgn.glm)), type="l",ylab="", xlab="",ylim=c(
```



Motor ensurance example: predict mean response

- Making a prediction at plausible values of explanatory variables for the gamma GLM:

```
> x0 <- data.frame(Make="1", Kilometres=1, Bonus=1, Insured=100)
> predg <- predict(gl,new=x0, se=T, type="response")   # for gamma glm
> predg$fit; predg$se.fit
    1
63061
    1
9711.4
```

- Making the prediction for the lognormal model:

```
> predln <-predict(llg, new=x0,se=T,type="response")   # for lognormal
> predln$fit; predln$se.fit
    1
10.998
[1] 0.16145
```

  - Results are on the log-scale.
  - By applying inverse link function we get prediction on original scale. An approximate standard error $se(exp(\hat{y}))$ is obtained by applying the delta-method (see lecture 39):

```
> c(exp(predln$fit), exp(predln$fit)*predln$se.fit)
       1        1
59771.0  9649.8
```

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 7, lecture 38

Content lecture 38: Miscellanea (Far ELM: 7.2-7.4)

- Inverse Gaussian distribution
- Joint modeling mean and dispersion
- Quasi-likelihood

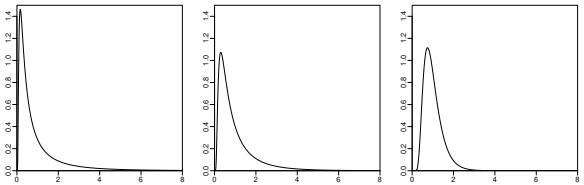
Inverse Gaussian GLM

- Density of inverse Gaussian random variable $Y \sim IG(\mu, \lambda)$ (with $y, \mu, \lambda > 0$) is

$$f(y|\mu,\lambda) = (\lambda/2\pi y^3)^{1/2} exp\left[-\lambda(y-\mu)^2/2\mu^2 y\right]$$

  mean is μ, variance is μ^3/λ. Canonical link is $\eta = 1/\mu^2$, variance function $V(\mu) = \mu^3$. Deviance is $D = \sum(y_i - \hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$.

```
> library(SuppDists)
> x <- seq(0,8,by=0.01)
> plot(x, dinvGauss(x,1,0.5), type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dinvGauss(x,1,1),   type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
> plot(x, dinvGauss(x,1,5),   type="l",ylab="", xlab="",ylim=c(0,1.5),xaxs="i",yaxs="i")
```



- Used in modeling lifetime distributions.
- Variance function increases more rapidly with mean than gamma GLM.
- Example is given in Faraway; I have never used inverse Gaussian distribution in practice.

Joint modeling of mean and dispersion

- In GLM's so far, we modeled the mean response $\mu = EY$, where variance has form $\text{var}Y_i = \phi V(\mu_i)$: dispersion parameter $\times$ variance function (of the mean)
- Dispersion parameter ϕ is e.g.
  - ϕ is variance in Gaussian model
  - $\phi = 1$ in binomial and Poisson model
  - ϕ is squared coefficient of variation in gamma model
- We may bring in unequal dispersions through known weights w_i: $\phi \equiv \phi_i = \phi/w_i$.
- Now, we want to allow the dispersion ϕ_i to vary with covariates X. This occurs f.i. in industrial experiments, where an item should be manufactured with target mean, but also with small variance. Effect of different predictors on mean and on dispersion is studied.

Joint modeling of mean and dispersion (2)

- Joint model specification:
  1. For mean use standard GLM approach:
     $E(Y_i) = \mu_i$; $\eta_i = g(\mu_i) = \sum_j x_{ij}\beta_j$; $\text{var}(Y_i) = \phi_i V(\mu_i)$; use as weights $w_i = 1/\phi_i$
  2. For dispersion ϕ_i, not longer considered fixed: suppose estimate d_i of dispersion is obtained from the GLM for the mean; model d_i with gamma GLM:
     $E(d_i) = \phi_i$; linear predictor is $\zeta_i = \log(\phi_i) = \sum_j z_{ij}\gamma_j$; $\text{var}(d_i) = \tau\phi_i^2$: quadratic variance function of the gamma GLM.
- Notice connection between two models: model for mean produces response d_i for the dispersion model, which in turn produces weights w_i for mean model.
- Model for dispersion not necessarily gamma, but we wish to model strictly positive, continuous, and typically skewed dispersion.
- Dispersion predictors usually are subset of mean model predictors.
- Candidates for d_i (estimates of ϕ_i) are r_P^2 and r_D^2: squared Pearson or deviance residuals.
- Example about welding-strength can be followed during practical.

Quasi-likelihood

- Suppose we are able to specify link and variance functions of model for some new type of data, but have no clear idea about distribution.
- E.g. we specify identity link and constant variance; this is typical of standard regression setting, use least-squares to estimate regression coefficients; formally, for inference the Gaussian distribution for errors is needed, but we know that inference is fairly robust to nonnormality if sample size gets larger.
  Important part of model specification is link and variance.
- Same holds for other GLMs. Provided a not-too-small sample, results are not sensitive to small deviations from distributional assumptions. Link, variance, and independence assumptions are far more important.
- Suppose combination of link and variance is specified, which does not correspond to any of standard GLMs. This is not really a problem, because for fitting procedure for GLMs (IRWLS) only link and variance functions are used without any distributional assumptions.
- However, estimation of β and their standard errors is often not enough: some form of inference is required. For the deviance, a likelihood is needed; for a likelihood a distribution is needed, which is lacking. This seems to be a problem.
- What we need is a suitable substitute for a likelihood that can be computed without assuming a distribution. We can find such a substitute: quasi-likelihood.

Quasi-likelihood (2)

- Let Y_i have mean μ_i and variance $\phi V(\mu_i)$. Assume that Y_i are independent.
  Define score U_i: $U_i = \dfrac{Y_i - \mu_i}{\phi V(\mu_i)}$.
- U_i has following properties:
  - $E(U_i) = 0$.
  - $\text{var}(U_i) = \dfrac{1}{\phi V(\mu_i)}$.
  - $-E\dfrac{\partial U_i}{\partial \mu_i} = -E\dfrac{-\phi V(\mu_i) - (Y_i - \mu_i)\phi V'(\mu_i)}{[\phi V(\mu_i)]^2} = \dfrac{1}{\phi V(\mu_i)}$

  These are the properties of the derivative of a log-likelihood ℓ' (the score)!
- This suggests that we can use U in place of ℓ'. The (elements of the) log-likelihood itself may be replaced by:
$$Q_i = \int_{y_i}^{\mu_i} \frac{y_i - t}{\phi V(t)} dt$$
- The quasi-likelihood $Q = \sum_{i=1}^n Q_i$ should behave like the log-likelihood.
- Asymptotic properties of maximum likelihood estimators also hold for quasi-likelihood based estimators.
- Choice of variance function is associated with random structure of model, while link function determines relationship with systematic part of model.

Quasi-likelihood (3)

- For variance functions associated with members of exponential family, quasi-likelihood corresponds exactly to log-likelihood.
- For variance functions corresponding to binomial and Poisson distribution, quasi-likelihood approach has advantage: regular GLM assumes $\phi = 1$, whereas corresponding quasi-binomial and quasi-poisson allow the dispersion ϕ to be free parameter, useful for modeling overdispersion, as we already saw.
- Some choices of $V(\mu)$ do not even correspond to a known (or existing) distribution.
- $\hat{\beta}$ is obtained by maximizing Q. Everything proceeds as in standard GLMs, except for estimation of ϕ. The likelihood approach is not reliable. Instead use $\hat{\phi} = \dfrac{X^2}{n-p}$.
- Quasi-likelihood estimators are generally less efficient than corresponding regular likelihood based estimator. So use distribution, if possible.

Quasi-likelihood (4)

- Inference as in standard GLMs. Regular deviance is $D(y, \hat{\mu}) = -2\phi \sum_i (\ell(\hat{\mu}_i|y_i) - \ell(y_i|y_i))$.
- By analogy, quasi-deviance is $-2\phi Q$, because contribution of saturated model is zero. The ϕ cancels, so quasi-deviance is just

$$-2 \sum_i \int_{y_i}^{\mu_i} \frac{y_i - t}{V(t)} dt$$

- An example of a quasibinomial GLM would be the model for a continuous fraction, e.g. the proportion of sleeping time spent dreaming. Such proportion does not have a binomial total n, so a binomial distribution cannot be used. But instead the variance function $\phi\mu(1-\mu)$ may be reasonable in combination with e.g. a logit link function (Faraway ELM §7.4 pp 149-150).

Linear & Generalized Linear Models and Linear Algebra

Master Statistical Science

Gerrit Gort
gerrit.gort@wur.nl

Biometris,
Wageningen University, Wageningen, The Netherlands

Week 7, lecture 39

Content lecture 39: delta method (App Fox)

- Delta method for approximate standard errors and confidence intervals of non-linear function of estimators
- Optional: Contingency tables, Simpson's paradox (Far ELM 4)
- Optional: Multinomial data (Far ELM 5)

Delta Method for approximate standard errors

- Suppose $\hat{\alpha}$ is m.l. estimator of parameter α. We are interested in parameter $\beta = f(\alpha)$, where f is a nonlinear function.
- We know that $\hat{\beta} = f(\hat{\alpha})$ is m.l. estimator of β (because $\hat{\alpha}$ is m.l. estimator of α).
- But for inference we need a standard error!
- Suppose we have an (approximate) variance $V(\hat{\alpha})$.
- Delta method gives approximate standard error of $\hat{\beta} = f(\hat{\alpha})$:
  - Based on first order Taylor-series approximation: develop $\hat{\beta} = f(\hat{\alpha})$ around the true value of the parameter α (which we don't know!):

$$f(\hat{\alpha}) \approx f(\alpha) + f'(\alpha)(\hat{\alpha} - \alpha)$$

  - Realize that, though we don't know α, and hence don't know neither $f(\alpha)$, nor $f'(\alpha)$, these are just constants.
  - Therefore, $V(f(\hat{\alpha})) = [f'(\alpha)]^2 V(\hat{\alpha})$.
  - This still brings no relief, because α is not known. Therefore, plug the MLE $\hat{\alpha}$ for α, to get the estimated asymptotic variance of $\hat{\beta}$:

$$\hat{V}(\hat{\beta}) = \left[f'(\hat{\alpha})\right]^2 V(\hat{\alpha})$$

Delta Method: example

- Suppose we have a sample of binary data, giving sample proportion $\hat{\pi}$ as m.l. estimator of population proportion π.
- We already know that $V(\hat{\pi}) = \pi(1-\pi)/n$.
- Suppose we are interested in $f(\pi) = \text{logit}(\pi) = \log \frac{\pi}{1-\pi}$. Call this parameter Λ.
- MLE of Λ is: $\hat{\Lambda} = \log \frac{\hat{\pi}}{1-\hat{\pi}}$.
- Asymptotic variance of this sample logit is $V(\hat{\Lambda}) = [f'(\pi)]^2 V(\hat{\pi}) = \left[\frac{1-\pi}{\pi}\frac{(1-\pi)-\pi(-1)}{(1-\pi)^2}\right]^2 \frac{\pi(1-\pi)}{n} = \left[\frac{1}{\pi(1-\pi)}\right]^2 \frac{\pi(1-\pi)}{n} = \frac{1}{n\pi(1-\pi)}$
- So, estimated asymptotic variance of logit is

$$\hat{V}(\hat{\Lambda}) = \frac{1}{n\hat{\pi}(1-\hat{\pi})}$$

Delta Method for function of several parameters

- Suppose $\beta \equiv f(\alpha_1, \alpha_2, \ldots, \alpha_k) = f(\boldsymbol{\alpha})$ (so $\boldsymbol{\alpha}$ is vector now).
  Let vector $\hat{\boldsymbol{\alpha}}$ be MLE of $\boldsymbol{\alpha}$, and let $V(\hat{\boldsymbol{\alpha}})$ be asymptotic variance covariance matrix of $\hat{\boldsymbol{\alpha}}$.
- Then asymptotic variance of $\hat{\beta} = f(\hat{\boldsymbol{\alpha}})$ is

$$V(\hat{\beta}) = [\mathbf{g}(\boldsymbol{\alpha})]' V(\hat{\boldsymbol{\alpha}}) [\mathbf{g}(\boldsymbol{\alpha})] = \sum_{i=1}^{k}\sum_{j=1}^{k} v_{ij} \times \frac{\partial\hat{\beta}}{\partial\alpha_i} \times \frac{\partial\hat{\beta}}{\partial\alpha_j}$$

  where $\mathbf{g}(\boldsymbol{\alpha}) \equiv \partial\hat{\beta}/\partial\boldsymbol{\alpha}$ and $v_{ij}$is i,jth entry of $V(\hat{\boldsymbol{\alpha}})$.
  Notice that $\mathbf{g}(\boldsymbol{\alpha})'$ is the vector of partial derivatives as row vector.
- Estimated asymptotic variance of $\hat{\beta}$ is obtained by filling in the maximum likelihood estimates $\hat{\alpha}$:

$$V(\hat{\beta}) = [\mathbf{g}(\hat{\boldsymbol{\alpha}})]' V(\hat{\boldsymbol{\alpha}}) [\mathbf{g}(\hat{\boldsymbol{\alpha}})]$$

Delta Method for function of several parameters: examples

- Fox pp 451-452 shows example for function of two parameters: estimate in a quadratic regression model $E(Y) = \ldots + \beta_1 X + \beta_2 X^2 \ldots$ the value of X where the function reaches a minimum or maximum. For this value of X the derivative w.r.t X needs to be 0, so $\beta_1 + 2\beta_2 X = 0$, leading to $X_{opt} = \frac{-\beta_1}{2\beta_2}$, a non-linear function of two parameters. An approximate variance of X_{opt} is obtained using the delta method.
- Faraway ELM pp 42-43 shows comparable example for LD50 from logistic regression: look for the value of X so that the probability equals 0.5: $\text{logit}(0.5) = \beta_0 + \beta_1 X$, so $\log(0.5/(1-0.5) = 0 = \beta_0 + \beta_1 X$, leading to LD50: $X = \frac{-\beta_0}{\beta_1}$, again a non-linear function of two parameters.
- Approximate confidence interval for ratio of regression coefficients can also be constructed using Fieller's method (not discussed).

Predictions in GLM

- Predictions (with standard errors) for glm's can be made on linear-predictor scale, or on fitted-value (response) scale, using function `predict()` with options `type="link"` or `type="response"`
- Predictions on the response scale are obtained by applying the inverse-link function to the predictions on the linear scale. The inverse-link function is a non-linear function, so to obtain an approximate variance or standard error for the prediction on the response scale the delta method may be used.
- Example on next slide (Faraway ELM pp 41-42) is about logistic regression for death rates of insects, which were treated with different concentrations of insecticide. We want to make a prediction for the fraction dead insects at dose 2.5, together with standard errors and confidence intervals.
  Prediction on linear predictor scale: $\hat{\eta} = \hat{\beta}_0 + 2.5\hat{\beta}_1$
  Prediction on response scale:
  $\hat{\mu} = \text{ilogit}(\hat{\beta}_0 + 2.5\hat{\beta}_1) = \frac{e^{\hat{\beta}_0+2.5\hat{\beta}_1}}{1+e^{\hat{\beta}_0+2.5\hat{\beta}_1}} = \frac{1}{1+e^{-(\hat{\beta}_0+2.5\hat{\beta}_1)}}$.
- The prediction on response scale involves a non-linear function of parameters.

Predictions for the linear predictor and for the mean response

- Fit logistic regression model and get predictions at link and response scales, using function `predict()` and by basic calculations. After that, get se's and ci's.

```
> data(bliss); modl <- glm(cbind(dead,alive) ~ conc, family=binomial, bliss); lmodsum <- summary(modl)
> #prediction at link scale:
> predict(modl, newdata=data.frame(conc=2.5))
     1
0.5809
> cf <- coef(modl)
> (lp <- c(1,2.5) %*% cf)
       [,1]
[1,] 0.5809
> # prediction at response scale:
> predict(modl, newdata=data.frame(conc=2.5), type="response")
     1
0.6413
> (fv <- 1/(1+exp(-lp)))
       [,1]
[1,] 0.6413
```

- The estimated probability to die with given concentration 2.5 is 0.64,

Predictions in GLM: standard errors of prediction

```
> # se of prediction at link scale:
> predict(modl, newdata=data.frame(conc=2.5),se=T, type="link" )$se.fit
[1] 0.2263
> b0 <- coef(modl)[1]; b1 <- coef(modl)[2]
> V.b <- lmodsum$cov.unscaled  # var-covar matrix of coefs
> (se.lp <- sqrt(c(1,2.5) %*% V.b %*% c(1,2.5)))
       [,1]
[1,] 0.2263
> # se of prediction at response scale: use delta method!
> predict(modl, newdata=data.frame(conc=2.5),se=T, type="response" )$se.fit
      1
0.05206
> dfdb0 <- exp(-(b0+b1*2.5))/(1+exp(-(b0+b1*2.5)))^2
> dfdb1 <- 2.5*exp(-(b0+b1*2.5))/(1+exp(-(b0+b1*2.5)))^2
> (se.delta <- sqrt(c(dfdb0,dfdb1) %*% V.b %*% c(dfdb0,dfdb1)))
        [,1]
[1,] 0.05206
```

- At the bottom part of R-script the delta method is applied: calculate the partial derivatives w.r.t. b0 and b1, and form the standard error of the prediction for mean response.
- Notice that the result is identical to what is obtained with the `predict` function.

Predictions in GLM: confidence intervals

- First form confidence intervals for prediction on linear predictor scale, so for linear combination $\beta_0 + 2.5\beta_1$.
- Next backtransform this interval into an interval for mean response applying the inverse-link function.
- This interval has better properties than the interval obtained by taking the prediction for the mean response +/- 1.96 standard error (mean response). The latter interval might e.g. contain negative values or values larger than 1.

```
> # 95% confidence interval for prediction on link scale
> lb <- lp - qnorm(0.975) * se.lp; ub <- lp + qnorm(0.975) * se.lp
> c(lb, ub)
[1] 0.1374 1.0245
> # 95% confidence interval for prediction on response scale: backtransform c.i. on link scale
> c(1/(1+exp(-lb)), 1/(1+exp(-ub)))
[1] 0.5343 0.7358
> # this is not the same as using the standard error as calculated with delta method!
> # backtransformed interval is better.
> lb.delta <- fv - qnorm(0.975) * se.delta; ub.delta <- fv + qnorm(0.975) * se.delta
> c(lb.delta, ub.delta)
[1] 0.5393 0.7433
```


Skip next part

The remaining part is optional and can be skipped.

Contingency tables

- Contingency table is frequency table of two or more cross-classified categorical variables.
- Variables can be nominal or ordinal.
- Example ordinal variable: variable on five-point Likert scale with values "strongly disagree", "disagree" , "neutral" , "agree", "strongly agree" .
- Definition of interval scale (discretized continuous data) in Faraway is strange.

Two-by-Two Tables

- Below are data from semiconductor experiment. Sample of wafers was drawn and cross-classified according to whether particle was found on die producing wafer and quality of wafer itself.

Quality	No Particles	Particles	Total
Good	320	14	334
Bad	80	36	116
Total	400	50	450

- Data may have arisen under several sampling schemes:
  - Observe manufacturing scheme certain time period; 450 wafer observed, which were next cross-classified. Use Poisson model.
  - Sample 450 wafers, which were next cross-classified. Use multinomial model.
  - Select 400 wafer without particles, and 50 with particles; next score good or bad outcome. Use binomial model.
  - Select 400 wafer without particles, and 50 with particles, that also included by design 334 good wafers and 116 bad ones. Use hypergeometric model.
- Hypergeometric scheme not very likely here.
- Question of interest: does presence of particles on die affect quality outcome?
- All four sampling schemes lead to same conclusion.

Two-by-Two Tables (2)

```
> y <- c(320,14,80,36)
> particle <- gl(2,1,4,labels=c("no","yes"))   # generate levels
> quality  <- gl(2,2,labels=c("good","bad"))
> (wafer <- data.frame(y,particle,quality))
    y particle quality
1 320       no    good
2  14      yes    good
3  80       no     bad
4  36      yes     bad
> (ov <- xtabs(y ~ quality + particle))
        particle
quality  no yes
   good 320  14
   bad   80  36
```

- Poisson Model: suppose process is observed for some time; count number of occurrences of possible outcomes. It would be natural to view these outcomes occurring at different rates, and form Poisson model for rates. Suppose we fit additive model:

```
> modl <- glm(y ~ particle + quality, poisson)
> coef(summary(modl))
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)     5.693     0.0572  99.535 0.000e+00
particleyes    -2.079     0.1500 -13.863 1.061e-43
qualitybad     -1.058     0.1078  -9.813 9.908e-23
> summary(modl)$null.deviance
[1] 474.1
> summary(modl)$deviance
[1] 54.03
```


Two-by-Two Tables: Poisson approach

- Null model tells that all 4 outcomes occur at same rate; model does not fit, because deviance is 474,1 on 3 d.f.
- Additive model is improvement, because deviance changed into 54.03 on 1 d.f. LR test:

```
> drop1(modl, test="Chi")
Single term deletions

Model:
y ~ particle + quality
          Df Deviance AIC LRT Pr(>Chi)
<none>           54  84
particle  1      364 392 310   <2e-16
quality   1      164 192 110   <2e-16
```

- Looking at coefficients we see that wafers without particles occur at significantly higher rate, and similarly with good-quality wafers.
- Fitted values of this additive model are functions of marginal totals.
- Log-likelihood (ignoring terms not depending on μ_i): $\log L = \sum_i y_i \log(\mu_i)$

Two-by-Two Tables: Poisson approach

- Additive model: $\log(\mu) = \gamma + \alpha_i + \beta_j$ with α_i is particle effect, and β_j quality effect.
- Due to log link, predicted rate for response is formed from product of rates for corresponding levels.
- No interaction term though, so variables particle and quality are supposed be independent.
- Deviance is 54.03 on 1 d.f. Hence, model does not fit.
- Test for independence is the test based on residual deviance from additive model:
  - Fit model with interaction; this is saturated model, so deviance is 0.
  - Fit additive model, deviance is 54.03
  - LR test statistic for independence is deviance difference: 54.03 - 0 = 54.03 on 1 d.f. Conclusion: null hypothesis of no interaction is rejected.

Two-by-Two Tables: Multinomial Model

- Suppose total sample size is fixed at 450; frequency of 4 outcomes was recorded.
- Multinomial model makes sense now.
- With n total sample size, y_{ij} count in cell (i,j) and p_{ij} probability that observation falls in that cell, probability of observed vector of counts is

$$\frac{n!}{\prod_i \prod_j y_{ij}!} \prod_i \prod_j p_{ij}^{y_{ij}}$$

- Maximize likelihood: $\log L = \sum_i \sum_j y_{ij} \log(p_{ij})$: same form as in Poisson case!
- Main question: are quality and particle-presence independent?
- With p_{ij} probability of joint outcome in cell (i,j), $p_{i.}$ probability of outcome of quality class i, and $p_{.j}$ probability of outcome of particle class j. Under independence $p_{ij} = p_{i.}p_{.j}$.
- M.l. estimators are $\hat{p}_{i.} = \sum_j y_{ij}/n$ and $\hat{p}_{.j} = \sum_i y_{ij}/n$

```
> (pi <- prop.table(xtabs(y ~ quality)))
quality
  good    bad
0.7422 0.2578
> (pj <- prop.table(xtabs(y ~ particle)))
particle
    no    yes
0.8889 0.1111
> (fv <- outer(pi,pj)*450)  # fitted values under independence
         particle
quality    no   yes
   good 296.9 37.11
   bad  103.1 12.89
```


Two-by-Two Tables: Multinomial Model (2)

- To test fit, compare this model against saturated model, for fitted values $\hat{\mu}_{ij} = y_{ij}$. So deviance is $2\sum_i \sum_j y_{ij} \log(y_{ij}/\mu_{ij})$:

```
> (2*sum(ov*log(ov/fv)))
[1] 54.03
```

- Same deviance as in Poisson case: test for independence in multinomial model coincides with test for no interaction in Poisson model.
- Connection between Poisson and multinomial model also clear from following: Let $Y_1, \ldots, Y_k$ be independent Poisson random variables with rates $\lambda_1, ldots, \lambda_k$. Then joint distribution of $Y_1, \ldots, Y_k | \sum_i Y_i = n$ is multinomial distribution with probabilities $p_j = \lambda_j / \sum_i \lambda_i$.
- Alternative to deviance is Pearsons X^2 statistic: $X^2 = \sum_{i,j} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$

```
> (sum((ov-fv)^2/fv))
[1] 62.81
> prop.test(ov) # X^2 with with Yates correction
        2-sample test for equality of proportions with continuity correction

data:  ov
X-squared = 60, df = 1, p-value = 9e-15
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1757 0.3611
sample estimates:
prop 1 prop 2
0.9581 0.6897
```


Two-by-Two Tables: Binomial Model

- It makes sense to view presence of particle as affecting quality of wafer. Then quality is response, and particle status is predictor.
- Fix number of wafers without particles at 400, and with particles at 50, and observe outcome.
- Use binomial model for response in two groups.

```
> (m <- matrix(y,nrow=2))
     [,1] [,2]
[1,]  320   80
[2,]   14   36
> modb <- glm(m ~ 1, family=binomial)
> deviance(modb)
[1] 54.03
```

- We fit null model saying that probability of good quality is equal for two groups of particle status.
- Hypothesis of homogeneity corresponds exactly to test of independence and deviance coincides.

Two-by-Two Tables: Hypergeometric Model

- Suppose that both margins are fixed.
- Not very common.
- Example: suppose you are given 10 true or false statements and told that 5 are true and 5 are false. Sort statements into true and false.
- Here is 2×2 table with given marginal totals

	drawn	not drawn	total
successes	k	$m-k$	m
failures	$n-k$	$N-n-m+k$	$N-m$
total	n	$N-n$	N

- Probability of table with k successes is given by hypergeometric distribution with probabilities of form $\dfrac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$

Two-by-Two Tables: Hypergeometric Model (2)

- Fishers exact test: total hypergeometric probability of all outcomes more extreme than one observed.

```
> fisher.test(ov)

        Fishers Exact Test for Count Data

data:  ov
p-value = 3e-13
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  5.091 21.544
sample estimates:
odds ratio
     10.21
```

- Odds ratio is $(y_{11}y_{22})/(y_{12}y_{21})$ measures association; exact confidence interval may be calculated (see above),

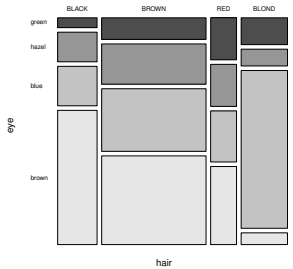
```
> (OR <- (320*36)/(14*80))
[1] 10.29
```

- Fishers exact test is attractive, because it can be used in small samples, where LR-tests needed in GLMs may fail.

Larger Two-Way Tables

```
> data(haireye)
> haireye[1:2,]
   y   eye  hair
1  5 green BLACK
2 29 green BROWN
> (ct <- xtabs( y ~ hair + eye, haireye))
       eye
hair    green hazel blue brown
  BLACK     5    15   20    68
  BROWN    29    54   84   119
  RED      14    14   17    26
  BLOND    16    10   94     7
> chisq.test(ct)

        Pearsons Chi-squared test

data:  ct
X-squared = 138, df = 9, p-value <2e-16
```

```
> mosaicplot(ct,color=TRUE,main=NULL,las=1)
```



Larger Two-Way Tables

- Fit Poisson GLM

```
> modc <- glm(y ~ hair + eye, family = poisson, haireye)
> summodc <-summary(modc)
> summodc$coef
             Estimate Std. Error z value  Pr(>|z|)
(Intercept)   2.4575     0.1523  16.136 1.424e-58
hairBROWN     0.9739     0.1129   8.623 6.539e-18
hairRED      -0.4195     0.1528  -2.745 6.045e-03
hairBLOND     0.1621     0.1309   1.238 2.157e-01
eyehazel      0.3737     0.1624   2.301 2.139e-02
eyeblue       1.2118     0.1424   8.510 1.741e-17
eyebrown      1.2347     0.1420   8.694 3.499e-18
> summodc$null.deviance
[1] 453.3
> summodc$deviance
[1] 146.4
```

- Deviance of 146.4 on 9 d.f. tells us that hair and eye color are not independent.
- In practice further study should reveal where the differences are, e.g. using correspondence analysis. We skip this now.
- We also skip 4.3. on matched pairs.

Three-Way Contingency Tables

```
> data(femsmoke)
> femsmoke[1:3,]
  y smoker dead   age
1 2    yes  yes 18-24
2 1     no  yes 18-24
3 3    yes  yes 25-34
> (ct <- xtabs( y ~ smoker + dead, femsmoke))
      dead
smoker yes  no
   yes 139 443
   no  230 502
> prop.table(ct,1)
      dead
smoker    yes     no
   yes 0.2388 0.7612
   no  0.3142 0.6858
> summary(ct)
Call: xtabs(formula = y ~ smoker + dead, data = femsmoke)
Number of cases in table: 1314
Number of factors: 2
Test for independence of all factors:
        Chisq = 9, df = 1, p-value = 0.003
```

- Smoking seems to be beneficial?! 76% of smokers survive, but only 69% of nonsmokers.
- Now look at what happens within age groups

```
> (cta <- xtabs( y ~ smoker + dead, femsmoke, subset=(age=="55-64")))
      dead
smoker yes no
   yes  51 64
   no   40 81
> prop.table(cta,1)
      dead
smoker    yes     no
   yes 0.4435 0.5565
   no  0.3306 0.6694
```

- Now 56% of smokers survive, and 67% of nonsmokers.
- This holds for all age groups!

Simpson's paradox

- Simpson's paradox: marginal association adding over age groups may be different from conditional association observed within age groups.

```
> prop.table(xtabs(y ~ smoker + age, femsmoke),2)
      age
smoker  18-24  25-34  35-44  45-54  55-64  65-74    75+
   yes 0.4701 0.4413 0.4739 0.6250 0.4873 0.2182 0.1688
   no  0.5299 0.5587 0.5261 0.3750 0.5127 0.7818 0.8312
```

- Smokers are more concentrated in younger age groups and younger peoples are more likely to live for another 20 years.
- Skip part on Mantel-Haenszel statistic (p 83).

Interactions with 3 factors

- Let p_{ijk} probability that observation falls in cell (i, j, k); $p_{i..}$ is marginal probability that observation falls in cell i of first variable; $p_{.j.}$ is marginal probability that observation falls in cell j of second variable; $p_{..k}$ is marginal probability that observation falls in cell k of third variable;
- Types of interaction
  - Mutual independence
  - Joint independence
  - Conditional independence
  - Uniform association

Mutual independence

- If all 3 variables are independent, then $p_{ijk} = p_{i..}p_{.j.}p_{..k}$.
- $EY_{ijk} = np_{ijk}$, so

$$\log(EY_{ijk}) = \log n + \log p_{i..} + \log p_{.j.} + \log p_{..k}$$

- Hence, main effects model corresponds to mutual independence.
- This is null model, i.e. simplest model.
- Don't use the model $\log(EY_{ijk}) = \mu$ as the null model, as this model is not reproducing the observed marginal totals.

```
> modi <- glm(y ~ smoker + dead + age, femsmoke, family=poisson)
> c(deviance(modi), df.residual(modi))
[1] 735  19
```

- Very large deviance for d.f., so this model does not fit the data.
- Coefficients of model correspond to marginal proportions, e.g. for smoke

```
> (coefsmoke <- exp(c(0,coef(modi)[2])))
        smokerno
   1.000    1.258
> coefsmoke/sum(coefsmoke)   # these are just marginal proportions for smokers
        smokerno
  0.4429   0.5571
```

- Hence, main effects are needed just to convey information that is already known, and are not main interest of study.

Joint independence

- Let $p_{ij.}$ be marginal probability that observation falls into cell $(i, j, .)$.
- Suppose first and second variable are dependent, but jointly independent with third.
- Then $p_{ijk} = p_{ij.}p_{..k}$, which can be represented as:
  $\log(EY_{ijk}) = \log(n) + \log(p_{ij.}) + \log(p_{..k})$
- Now include main effects and interaction term $\log(p_{ij.})$.
- Model with just one interaction is model with joint independence.
- E.g.

```
> modj <- glm(y ~ smoker*dead + age, femsmoke, family=poisson)
> c(deviance(modj), df.residual(modj))
[1] 725.8  18.0
```

- Small improvement over mutual independence model, but deviance is still very high. Model does not fit well.
- Other two models with joint independence appear to fit badly as well.

Conditional independence

- Let $p_{ij|k}$ be probability that observation falls in cell $(i, j, .)$ given that third variable has value k.
- Suppose we assert that first and second variables are independent given value of third variable, then $p_{ij.|k} = p_{i..|k}p_{.j.|k}$, which leads to $p_{ijk} = p_{i.k}p_{.jk}/p_{..k}$ (because e.g. $p_{i..|k} = p_{i.k}/p_{..k}$), resulting in model
  $\log(EY_{ijk}) = \log(n) + \log(p_{i.k}) + \log(p_{.jk}) - \log(p_{..k})$
- Using hierarchy principle, also include main effects corresponding to interaction terms; here we have model with 3 main effects and two interaction terms. The minus for $\log(p_{..k})$ is irrelevant.

```
> modc <- glm(y ~ smoker*age + age*dead, femsmoke, family=poisson)
> c(deviance(modc), df.residual(modc))
[1] 8.327 7.000
```

- Now deviance is only slightly larger than d.f., indicating good fit.
- So, smoking is independent of life status, given age.
- Some doubts about accuracy of χ^2 approximation due to small counts.

Uniform association

- Still one step to go before we reach the saturated model: model with all pairwise interactions.
  $\log(EY_{ijk}) = \log(n) + \log(p_{i..}) + \log(p_{.j.}) + \log(p_{..k}) + \log(p_{ij.}) + \log(p_{i.k}) - \log(p_{.jk})$
- No simple interpretation in terms of independence.

```
> modu <- glm(y ~ (smoker + age + dead)^2, femsmoke, family=poisson)
> ctf <- xtabs(fitted(modu) ~  smoker + dead + age, femsmoke)
> apply(ctf, 3, function(x) (x[1,1]*x[2,2]/(x[1,2]*x[2,1])))
18-24 25-34 35-44 45-54 55-64 65-74   75+
1.533 1.533 1.533 1.533 1.533 1.533 1.533
```

- Odds ratio for smoke / dead is same at each level of age: uniform association model asserts that for every level of one variable, we have the same association for other two variables.

Model selection

- Studied log-linear models are hierarchical.
- Start with most complex model and see how far it can be reduced by analysis of deviance.
- Start with saturated model.

```
> modsat <- glm(y ~ smoker*age*dead, femsmoke, family=poisson)
> drop1(modsat, test="Chi")
Single term deletions

Model:
y ~ smoker * age * dead
              Df Deviance AIC  LRT Pr(>Chi)
<none>             0.00 190
smoker:age:dead  6     2.38 181 2.38     0.88
```

- Three-way interaction term may be dropped.
- Can two-way terms be dropped?

```
> drop1(modu, test="Chi")
Single term deletions

Model:
y ~ (smoker + age + dead)^2
           Df Deviance AIC LRT Pr(>Chi)
<none>              2 181
smoker:age   6       93 259  90   <2e-16
smoker:dead  1        8 185   6    0.015
age:dead     6      632 798 630   <2e-16
```

- Two interaction terms are highly significant; `smoker:dead` only just significant. This term corresponds to test for conditional independence of smoking and life status given age group. So, conditional independence does not hold.

Binomial model

- Some 3-way tables one variable may be considered response, other two predictors.
- Example: life status may be considered response; has 2 levels, so binomial GLM; for > 2 levels multinomial model required.

```
>   ybin <- matrix(femsmoke$y,ncol=2)
> modbin <- glm(ybin ~ smoker*age, femsmoke[1:14,], family=binomial)
> drop1(modbin, test="Chi")
Single term deletions

Model:
ybin ~ smoker * age
           Df Deviance  AIC  LRT Pr(>Chi)
<none>           0.00 75.0
smoker:age  6     2.38 65.4 2.38     0.88
```

- Interaction term may be dropped. Further simplification possible? ■binmod2, keep.source=T■= modbinr <- glm(ybin smoker + age, femsmoke[1:14,], family=binomial) drop1(modbinr, test="Chi")
- This model equivalent to uniform association model.

Binomial model: compare

```
>   deviance(modbinr)
[1] 2.381
> deviance(modu)
[1] 2.381
> exp(-coef(modbinr)[2])
smokerno
   1.533
```

- So binomial GLM can be identified with Poisson GLM.
- Binomial analysis preferred.
- Difference: null models.

```
>   modbinull <- glm(ybin ~ 1, femsmoke[1:14,], family=binomial)
> deviance(modbinull)
[1] 641.5
> modj <- glm(y ~ smoker*age + dead, femsmoke, family=poisson)
> deviance(modj)
[1] 641.5
```

- Binomial model implicitly asumes an association between smoker and age, which indeed is present in this example.
- Poisson model would allow to drop an extra if it is not important; would be impossible for binomial model.
- Skip 4.5 on ordinal variables

Multinomial data

- Multinomial distribution is extension of binomial distribution to situation where response can take more than two values.
- Let Y_i be random variable, taking one of values $1, 2, \ldots, J$; let $p_{ij} = P(Y_i = j)$, so $\sum_{j=1}^{J} p_{ij} = 1$.
- Binomial situation if $J = 2$.
- Data may be grouped or ungrouped.
- Let Y_{ij} be number of observations falling into category j for group or individual i, and let $n_i = \sum_j Y_{ij}$.
- For ungrouped data, $n_i = 1$ and one and only one of $Y_{i1}, \ldots, Y_{iJ}$ is equal to one and rest to zero.
- Y_{ij} conditional on total n_i has multinomial distribution:

$$P(Y_{i1} = y_{i1}, \ldots, Y_{iJ} = y_{iJ}) = \frac{n_i!}{y_{i1}! \cdots y_{iJ}!} p_{i1}^{y_{i1}} \cdots p_{iJ}^{y_{iJ}}$$

- Distinguish between nominal and ordinal multinomial data.

Multinomial logit model

- Probabilities p_{ij} must be linked to predictor x_i, while ensuring that probabilities are restriced between 0 and 1.
- Similar idea as in binomial case: $\eta_{ij} = x_i^T \beta_j = \log \frac{p_{ij}}{p_{i1}}$ $(j = 2, \ldots, J)$; note that $\eta_{i1} = 0$.
- Constraint is $\sum_{j=1}^{J} p_{ij} = 1$, so convenient to declare one (first) category baseline, and set $p_{i1} = 1 - \sum_{j=2}^{J} p_{ij}$.
- We have $p_{ij} = \dfrac{exp(\eta_{ij})}{1 + \sum_{j=2}^{J} exp(\eta_{ij})}$.
- Estimate parameter by m.l., and use standard methods of inference.
- Example: response is party pref (Democrat, Independent, Republican), predictors are age, education level and income group.

```
> summary(sPID)
   Democrat Independent  Republican
        380         239         325
> summary(nincome)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.5    23.5    37.5    46.6    67.5   115.0
> table(nes96$educ)
   MS HSdrop     HS   Coll  CCdeg  BAdeg  MAdeg
   13     52    248    187     90    227    127
```

Multinomial logit model

- Multinomial logit model can be fitted with special `multinom` function.
- Poisson glm can also be used: independent Poisson variates conditional on their total are multinomially distributed.
- To this end, declare factor that has level for each multinomial observation in data: response factor; next treat individual components of multinomial response as Poisson responses. For ungrouped data (like present example), one response will be one, and rest zero.
- Below graphical output for multinomial logit model for different example from Fox (figure 14.6).