Statistical Data Analysis, Lecture 10

dr. Dennis Dobler

Vrije Universiteit Amsterdam

22 April 2020

Topics in this course

1. Summarizing data
2. Exploring distributions
3. Density estimation
4. Bootstrap methods
5. Nonparametric tests
6. Analysis of categorical data
7. Multiple linear regression

Chapter 7: Analysis of categorical data

Contents of Chapter 7:

1. Fisher's exact test
2. Chisquare test
3. Extreme values
4. Bootstrap methods for contingency tables

Chapter 8: Linear regression analysis

Contents of Chapter 8:

1. The multiple linear regression model
   - parameter estimation
   - selection of explanatory variables
2. Diagnostics
   - plots
   - outliers
   - leverage points
   - influence points
3. Collinearity

SDA and spring vacation!

Scheme for the coming three weeks:

today, 22 April Lecture 10 and Open Office Hour (Zoom)
24 April computer class Assignment 6

27 April–1 May no lecture, no computer class!

5 May hand in Assignment 6
6 May feedback Assignment 6,
full lecture on Chapter 8,
Assignment 7 available

...
19 May hand in Assignment 7
...

models

General contingency table – Models II

	B_1	$\cdots\cdots\cdots$	B_j	$\cdots$	B_c	total
A_1	N_{11}	$\cdots\cdots\cdots$	N_{1j}	$\cdots$	N_{1c}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	N_{i1}	$\cdots\cdots\cdots$	N_{ij}	$\cdots$	N_{ic}	$N_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	N_{r1}	$\cdots\cdots\cdots$	N_{rj}	$\cdots$	N_{rc}	$N_{r\cdot}$
total	$N_{\cdot 1}$	$\cdots\cdots\cdots$	$N_{\cdot j}$	$\cdots$	$N_{\cdot c}$	$n = N_{\cdot\cdot}$

The general form of a contingency table, with
row variable A (r categories) and column variable B (c categories).

Model II A

Model II A One sample of size n. One rc-nomial distribution with probabilities p_{ij},

$$\sum_{i=1}^{r}\sum_{j=1}^{c} p_{ij} = 1$$

Null hypothesis No dependence between row and column variable, $p_{ij} = p_{i\cdot}p_{\cdot j}$ for $i = 1, \ldots, r, j = 1, \ldots, c$.

Test statistic

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(N_{ij} - n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}} \qquad \text{with } \hat{p}_{ij} = \frac{N_{i\cdot}N_{\cdot j}}{n^2}$$

Distribution $X^2 \sim \chi^2_{(r-1)(c-1)}$ under H_0, approximately.

Model II B

Model II B r independent samples of size $N_{i.}$ each. r c-nomial distributions with probabilities p_{ij},

$$\sum_{j=1}^{c} p_{ij} = 1 \qquad \text{for } i = 1, \ldots, r$$

Null hypothesis The r samples are homogeneous, i.e. $p_{1j} = p_{2j} = \cdots = p_{rj} \equiv p_j$ for $j = 1, \ldots, c$.

Test statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(N_{ij} - n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}} \qquad \text{with } \hat{p}_{ij} = \frac{N_{i.}N_{.j}}{n^2}$$

Distribution $X^2 \sim \chi^2_{(r-1)(c-1)}$ under H_0, approximately.

Model II C

Model II C c independent samples of size $N_{\cdot j}$ each. c r-nomial distributions with probabilities p_{ij},

$$\sum_{i=1}^{r} p_{ij} = 1 \qquad \text{for } j = 1, \ldots, c$$

Null hypothesis The c samples are homogeneous, i.e. $p_{i1} = p_{i2} = \cdots = p_{ic} \equiv p_i$ for $i = 1, \ldots, r$.

Test statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(N_{ij} - n\hat{p}_{ij}\right)^2}{n\hat{p}_{ij}} \qquad \text{with } \hat{p}_{ij} = \frac{N_{i\cdot} N_{\cdot j}}{n^2}$$

Distribution $X^2 \sim \chi^2_{(r-1)(c-1)}$ under H_0, approximately.

chisquare test

The theorems needed

Theorem

Let for $m = 2, 3, \ldots$, the ℓ-vector $N^m = (N_1, \ldots, N_\ell)$ with $\sum_{j=1}^{\ell} N_j = m$ be multinomially distributed with parameters $m, p_1, \ldots, p_\ell$ which satisfy $p_j > 0$ for all j and $\sum_{j=1}^{\ell} p_j = 1$. Then it holds that

$$\sum_{j=1}^{\ell} \frac{(N_j - mp_j)^2}{mp_j} \xrightarrow{\mathcal{D}} \chi^2_{\ell-1}, \qquad m \to \infty,$$

where χ^2_ν denotes a chi-square distribution with ν degrees of freedom.

Theorem

The sum of s independent χ^2_ν distributed random variables has a $\chi^2_{s\nu}$ distribution.

Application to models

We apply these theorems to models (II) A, B and C.

Because we don't know p_{ij} we need estimates under the three H_0's respectively

$$\begin{aligned}
\text{model A: } \hat{p}_{ij} &= \hat{p}_{i.}\hat{p}_{.j} = \frac{N_{i.}}{n}\frac{N_{.j}}{n} \\
\text{model B: } \hat{p}_{j} &= \frac{N_{.j}}{n} \\
\text{model C: } \hat{p}_{i} &= \frac{N_{i.}}{n}
\end{aligned}$$

For each maximum likelihood estimated parameter the number of degrees of freedom of the test statistic distribution is reduced by 1.

The chisquare test

Theorem

Under the null hypotheses of models II A, II B, and II C, and for n, the row totals, and the column totals, respectively, sufficiently large, the test statistic

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(N_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}},$$

with

$$\hat{p}_{ij} = \frac{N_{i.} N_{.j}}{n^2}$$

approximately has a χ^2-distribution with $(r-1)(c-1)$ degrees of freedom.

Remark The approximation by $\chi^2_{(r-1)(c-1)}$ is reasonable if $\mathrm{E}_{H_0} N_{ij} > 1$ for all (i,j) and at least 80% of the $\mathrm{E}_{H_0} N_{ij} > 5$.

Example 1 (0)

Question Are kind of study and gender independent?
Consider the following data (numbers given are counts):

	exact	arts
men	23	17
women	7	13

Example 1 (1)

If we apply the chisquare test to the `study` data

```
> study
     [,1] [,2]
[1,]   23    7
[2,]   17   13
```

we first need to check the rule of thumb:

```
> chisq.test(study)$expected
     [,1] [,2]
[1,]   20   10
[2,]   20   10
```

The rule of thumb is fulfilled.

(The attribute `expected` of the function `chisq.test` just computes $\frac{N_{i\cdot}N_{\cdot j}}{n}$.)

Example 1 (2)

Then we apply the chisquare test to the study data (model A, B or C):

```
> chisq.test(study)

Pearson's Chi-squared test with Yates' continuity correction

data:  study
X-squared = 1.875, df = 1, p-value = 0.1709
```

R applies a continuity correction for 2 x 2 tables. Without that correction:

```
> chisq.test(study,correct=F)

Pearson's Chi-squared test

data:  study
X-squared = 2.7, df = 1, p-value = 0.1003
```

This correction makes sense for small tables, like 2×2.

Example 1 (3)

The function `chisq.test` has an argument `simulate.p.value`.

If it is set to `TRUE` *R* performs a bootstrap test and does not use the $\chi^2_{(k-1)(r-1)}$ distribution.

```
> chisq.test(study,simulate.p.value=T)

Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)

data:  study
X-squared = 2.7, df = NA, p-value = 0.1839
> chisq.test(study,simulate.p.value=T)

Pearson's Chi-squared test with simulated p-value (based on 2000
replicates)

data:  study
X-squared = 2.7, df = NA, p-value = 0.1679
```

This setting is useful if the rule of thumb is not fulfilled.

Example 2 (0)

Question: Does frequency of nucleotides in DNA depend on its position in the DNA sequence?
Consider the following data of 100 DNA sequences of length 5:

position	1	2	3	4	5	total
A	33	34	19	20	21	127
G	22	27	23	24	21	117
C	31	18	34	30	25	138
T	14	21	24	26	33	118
total	100	100	100	100	100	500

Example 2 (1)

The DNA data:

```
> dna
   1  2  3  4  5
A 33 34 19 20 21
G 22 27 23 24 21
C 31 18 34 30 25
T 14 21 24 26 33
```

Which model is appropriate here?

Model C: 5 samples of 100 nucleotides.
H_0 : the probabilities of having A, G, C or T (p_A, p_G, p_C and p_T) is the same for each position.

```
> chisq.test(dna)$expected
     [,1] [,2] [,3] [,4] [,5]
[1,] 25.4 25.4 25.4 25.4 25.4
[2,] 23.4 23.4 23.4 23.4 23.4
[3,] 27.6 27.6 27.6 27.6 27.6
[4,] 23.6 23.6 23.6 23.6 23.6
```

Example 2 (2)

The chisquare test yields:

```
> chisq.test(dna)

Pearson's Chi-squared test

data:  dna
X-squared = 23.4967, df = 12, p-value = 0.02379
```

Conclusion H_0 is rejected (at level $\alpha = 5\%$): the probability of the nucleotides is not the same for all positions.

Question Where are the differences from H_0 in these data?

extremes

General contingency table

	B_1	$\cdots\cdots\cdots$	B_j	$\cdots$	B_c	total
A_1	N_{11}	$\cdots\cdots\cdots$	N_{1j}	$\cdots$	N_{1c}	$N_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_i	N_{i1}	$\cdots\cdots\cdots$	N_{ij}	$\cdots$	N_{ic}	$N_{i\cdot}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
A_r	N_{r1}	$\cdots\cdots\cdots$	N_{rj}	$\cdots$	N_{rc}	$N_{r\cdot}$
total	$N_{\cdot 1}$	$\cdots\cdots\cdots$	$N_{\cdot j}$	$\cdots$	$N_{\cdot c}$	$n = N_{\cdot\cdot}$

Again, the general form of a contingency table, with row variable A (r categories) and column variable B (c categories).

Residuals, contributions and normalized contributions

Residuals (`data-expected`)

$$N_{ij} - n\hat{p}_{ij}$$

Contributions (`residuals`)

$$\frac{N_{ij} - n\hat{p}_{ij}}{\sqrt{n\hat{p}_{ij}}}$$

Normalized contributions (`stdres`)

$$V_{ij} = \frac{N_{ij} - n\hat{p}_{ij}}{\sqrt{\frac{N_{i\cdot}(n-N_{i\cdot})N_{\cdot j}(n-N_{\cdot j})}{n^2(n-1)}}} \overset{H_0}{\sim} N(0,1) \text{ approx.}$$

Looking at residuals (1)

Compute residuals and contributions to X^2

```
> dna-chisq.test(dna)$expected    ## "residuals" in syllabus
     1    2    3    4    5
A  7.6  8.6 -6.4 -5.4 -4.4
G -1.4  3.6 -0.4  0.6 -2.4
C  3.4 -9.6  6.4  2.4 -2.6
T -9.6 -2.6  0.4  2.4  9.4
> chisq.test(dna)$residuals       ## "contributions" in syllabus
      1     2     3     4     5
A  1.51  1.71 -1.27 -1.07 -0.87
G -0.29  0.74 -0.08  0.12 -0.50
C  0.65 -1.83  1.22  0.46 -0.49
T -1.98 -0.54  0.08  0.49  1.93
```

Largest contributions:

- (A,1), (A,2) and (T,5) much (?) more often than expected under H_0
- (C,2) and (T,1) much (?) less often than expected under H_0

Looking at residuals (2)

Compare normalized residuals to $\Phi^{-1}(\alpha/2)$ and $\Phi^{-1}(1-\alpha/2)$ (e.g. ± 1.96):

```
> round(chisq.test(dna)$stdres,2)
      1     2     3     4     5
A  1.95  2.21 -1.64 -1.39 -1.13
G -0.37  0.95 -0.11  0.16 -0.63
C  0.85 -2.40  1.60  0.60 -0.65
T -2.53 -0.68  0.11  0.63  2.48
```

Conclusion H_0 is rejected because

- (A,2) and (T,5) more often than expected under H_0
- (C,2) and (T,1) less often than expected under H_0

bootstrap methods

The bootstrap test for contingency tables

Apply a bootstrap test, conditionally on the row and column marginals, using some sensible test statistic T.

- Generate B new $r \times c$-contingency tables with the same marginals but following the null hypothesis H_0!
- Compute the B bootstrap values $T_1^*, \ldots, T_B^*$.
- The empirical distribution of the B bootstrap values $T_1^*, \ldots, T_B^*$ estimates the conditional distribution of T under H_0 given the marginals of the original data set.

In case the χ^2 approximation does not hold, one can do a bootstrap test using $T = X^2$. Other interesting statistics: smallest entry, largest entry, maximum absolute contribution.

Example (1)

Apply a bootstrap test to the study data using the maximum entry, the minimum entry or the maximum absolute contribution as statistic T (one-sided hypothesis tests).

```
> t=max(study)
> bootval=bootstrapcat(study,1000,max)
> mean(bootval>=t)
[1] 0.061
> mean(bootval>=t)
[1] 0.047

> t=min(study)
> bootval=bootstrapcat(study,1000,min)
> mean(bootval<=t)
[1] 0.059
> bootval=bootstrapcat(study,1000,min)
> mean(bootval<=t)
[1] 0.045
```

Example(2)

```
> t=maxcontributionscat(study)
> bootval=bootstrapcat(study,1000,maxcontributionscat)
> mean(bootval>=t)
[1] 0.054
```

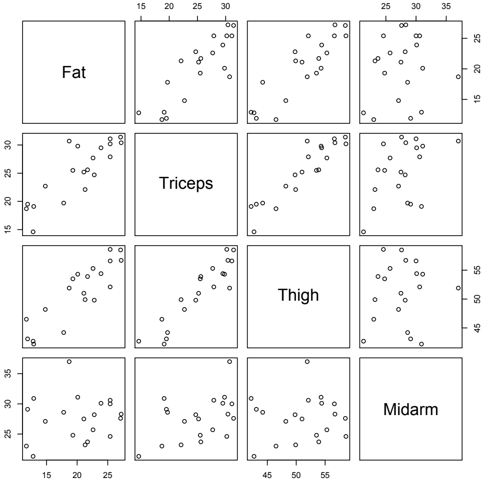
Conclusion The results for these test statistics (based on the study data) are on the verge of rejecting/not rejecting H_0 $(\alpha = 5\%)$.

Functions `bootstrapcat` and `maxcontributionscat` are available together with Assignment 6.

multiple linear regression

Idea (1)

Example Data on bodyfat & different girths of 20 females.



Fat: difficult to measure.

Question:
Predict Fat
from other variables?

Idea (2)

Regression:

response variable $=$ f(explanatory variables) $+$ measurement error
(dependent) (independent)

Linear regression: f linear function.

Other: nonlinear/generalized linear regression (Statistical Models)

Multiple linear regression model

$$
\begin{aligned}
Y_i &= \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + e_i \\
\mathrm{E}e_i &= 0 \\
\mathrm{E}e_i e_j &= \begin{cases} \sigma^2, & i = j, \\ 0, & i \neq j, \end{cases}
\end{aligned}
$$

where

- Y_i: i^{th} response observation
- x_{ij}: (known) value of j^{th} explanatory variable for i^{th} obs.
- $\beta_0, \beta_1, \ldots, \beta_p$, and σ^2: unknown parameters
- e_i: unknown stochastic measurement error in i^{th} observation

Model in matrix notation

$$
\begin{aligned}
Y &= X\beta + e \\
\mathrm{E}e &= 0 \\
\mathrm{Cov}(e) &= \sigma^2 I_{n\times n}
\end{aligned}
$$

with

- $Y = (Y_1, \ldots, Y_n)^T$ stochastic vector of observations
- $X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$ design matrix,

  (known) values of explanatory var. (assume $\mathrm{rank}(X) = p + 1$)
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ vector of unknown parameters
- σ^2 unknown variance
- $e = (e_1, \ldots, e_n)^T$ stochastic vector of measurement errors

Further common assumptions

$$e_i \sim N(0, \sigma^2) \quad \text{i.i.d.} \qquad i = 1, \ldots, n$$

$$\Rightarrow \quad Y_i \sim N(\beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p, \sigma^2)$$

Note: Y_i not identically distributed!

parameter estimation

Least squares approach

$\hat\beta$ minimizes $S(\beta) = \|Y - X\beta\|^2$.
Result: parameter estimator $\qquad \hat\beta = (X^TX)^{-1}X^TY$.
We have

$$\begin{aligned}
\mathrm{E}\hat\beta &= (X^TX)^{-1}X^T\mathrm{E}Y = (X^TX)^{-1}X^TX\beta = \beta \\
\mathrm{Cov}(\hat\beta) &= \sigma^2(X^TX)^{-1}
\end{aligned}$$

Residuals: $\; R_i = Y_i - \hat Y_i$, where $\hat Y_i = \hat\beta_0 + x_{i1}\hat\beta_1 + \cdots + x_{ip}\hat\beta_p$.

Residual sum of squares:

$$RSS = S(\hat\beta) = \sum_{i=1}^n (Y_i - \hat Y_i)^2 = \|Y - X\hat\beta\|^2$$

Finally, $\quad \hat\sigma^2 = \frac{RSS}{n-p-1} \quad$ and $\quad \widehat{\mathrm{Cov}}(\hat\beta) = \hat\sigma^2(X^TX)^{-1}$.

Example (1): bodyfat data

```
> bodyfat=read.table("bodyfat.txt",header=TRUE)

> bodyfat
    Fat Triceps Thigh Midarm
1  11.9    19.5  43.1   29.1
2  22.8    24.7  49.8   28.2
....
20 21.1    25.2  51.0   27.5

> is.data.frame(bodyfat)
[1] TRUE

> is.matrix(bodyfat)
[1] FALSE
```

`bodyfat`: `dataframe` in *R* (default of `read.table`).

Needed for `lm`.

Example (2): bodyfat data

```
> fatlm=lm(Fat~Triceps+Thigh+Midarm,data=bodyfat)
> fatlm

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Coefficients:
(Intercept)      Triceps        Thigh       Midarm
    117.085        4.334       -2.857       -2.186
```

`lm` takes `model formula`: `response` $\sim$ `var1+...+varp`.

Default: include intercept. Drop? `response` $\sim$ `var1+...+varp-1`.

`lm` output: `linear model`. Apply summary; see `help(lm)`.

Example (3): bodyfat data

```
> summary(fatlm);

Call:
lm(formula = Fat ~ Triceps + Thigh + Midarm)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7263 -1.6111  0.3923  1.4656  4.1277

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
Triceps        4.334      3.016   1.437    0.170
Thigh         -2.857      2.582  -1.106    0.285
Midarm        -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06
```

to finish

To wrap up

Today we discussed

1. Contingency tables
   - Chisquare test
   - Extreme values
   - Bootstrap methods
2. Multiple linear regression
   - Idea
   - Parameter estimation

in two weeks Multiple linear regression

- variable selection
- plots
- outliers
- leverage points
- influence points