

MPH specialmodul Epidemiologi og Biostatistik

Lineær regression.

Analyse af tabeller.

Introduktion til logistisk regression.

23. marts og 20. april 2009

www.biostat.ku.dk/~pka/mpspec09

Per Kragh Andersen

1

Regression analysis.

The distribution of ONE (1):

- outcome (“udfalds”) variable
- response (“respons”) variable
- dependent (“afhængig”) variable
- Y variable

is related to ONE OR MORE

- explanatory (“forklarende”) variables
- independent (“uafhængig”) variables
- regression variables
- X variables
- covariates

2

Particularly, in epidemiological investigations, one distinguishes between explanatory variables which may be

- exposure variables (determinants)
- confounders

3

Types of outcome variables.

The type of OUTCOME variable determines which kind of regression model is relevant:

Y	Model
0-1 (“binary”)	logistic regression
Quantitative	linear regression
Survival time (“rate”)	Cox (Poisson) regression

4

The interpretation of the effect of an explanatory variable also depends on the type of outcome variable (i.e., the type of regression model):

Model	effect
logistic regression	$OR, \ln(OR)$
linear	difference between mean values
Cox (Poisson)	rate ratio, $\ln(\text{rate ratio})$

Types of explanatory variables.

In ALL types of regression models, TWO types of explanatory variables may be considered: categorical (esp. *binary*) “groups” and quantitative “lines”

For a *categorical* explanatory variable, the effect corresponds to difference between groups:

logistic regression	$\ln(OR)$
linear regression	mean values
Cox (Poisson) regression	$\ln(\text{rate ratio})$

For a *quantitative* explanatory variable, X , the effect corresponds to differences in

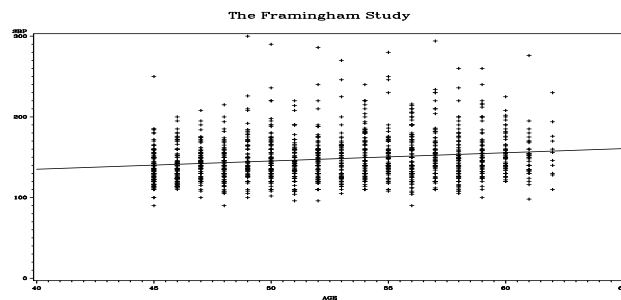
$\ln(OR)$ / mean value / $\ln(\text{rate ratio})$ *per unit of X*.

NOTE: This *linearity* is a model assumption to be checked!

Introduction to regression models: Linear regression.

Framingham data, no CHD at entry quantitative outcome:

$Y = SBP$, $X = \text{age in years}$ (quantitative explanatory variable)



Model: $Y_i = a + bX_i + \text{error}_i$

Line (a and b) fitted using “least-squares”, i.e., the estimates for a and b minimize the “sum of squares”:

$$\sum_i (Y_i - (a + bX_i))^2$$

b = effect of age (X) = slope of line, here measured in *mmHg* per year,

a = intercept of line = predicted value when $X = 0$.

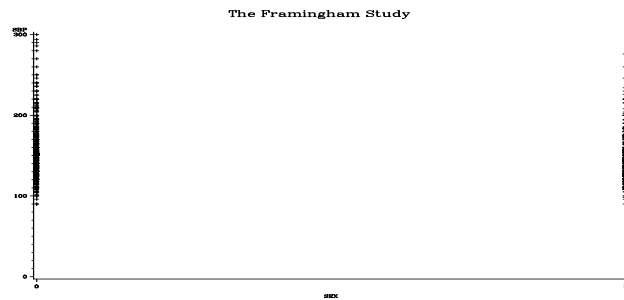
Estimated regression line: $SBP = 94.05 + 1.03 \cdot \text{age}$

Standard error (SE) of age effect: 0.16.

Test for no age effect $(1.03/0.16)^2 = 43.8, P < 0.0001$.

A binary explanatory variable: sex

$$Z = \begin{cases} 1 & \text{males} \\ 0 & \text{females} \end{cases}$$



9

$$\begin{aligned} \text{Model: } Y_i &= a + cZ_i + \text{error}_i \\ &= \begin{cases} (a + c) + \text{error}_i & \text{(males)} \\ a + \text{error}_i & \text{(females)} \end{cases} \end{aligned}$$

$c = \text{effect of sex } (Z) = \text{difference between mean for males and mean for females.}$

Note: this will depend on the chosen *reference category* (here females).

Note: different computer programs may choose the reference category differently.

10

a and c estimated using least squares

$$\begin{aligned} a &\sim \text{average } Y \text{ for females} && = 151.45 \\ c &\sim \text{average } Y \text{ for males minus} \\ &\quad \text{average } Y \text{ for females} && = 143.64 - 151.45 \\ &&& = -7.81 \end{aligned}$$

Standard error (SE) of sex effect: 1.49.

Test for no sex effect: $(-7.81/1.49)^2 = 27.4, P < 0.0001.$

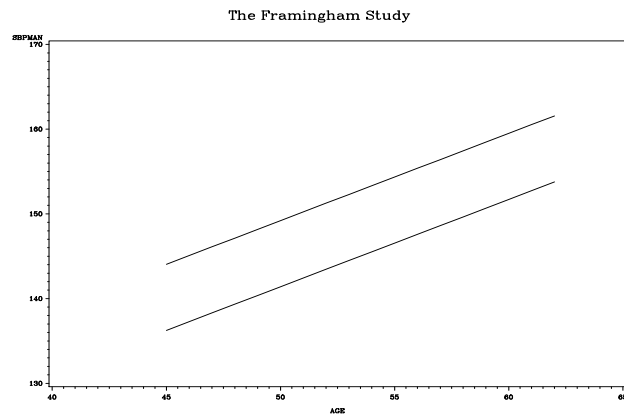
11

Two explanatory variables: age and sex

Possible model: NB just add terms for age and sex:

$$\begin{aligned} Y_i &= a + bX_i + cZ_i + \text{error}_i \\ &= \begin{cases} (a + c) + bX_i + \text{error}_i & \text{(males)} \\ a + bX_i + \text{error}_i & \text{(females)} \end{cases} \end{aligned}$$

12



NB: Same age effect (b) for both males and females
 Same sex effect (c) for all ages

13

$$\begin{aligned}
 Y_i &= a + bX_i + cZ_i + \text{error}_i \\
 &= \begin{cases} (a + c) + bX_i + \text{error}_i & \text{(males)} \\ a + bX_i + \text{error}_i & \text{(females)} \end{cases}
 \end{aligned}$$

NB: Same age effect (b) for both males and females
 Same sex effect (c) for all ages

$$\Rightarrow \text{No} \begin{cases} \text{interaction} \\ \text{effect-modification} \end{cases}$$

14

Here, parallel lines:

$$SBP = 97.68 + 1.03 \cdot \text{age} \text{ for females}$$

$$SBP = 97.68 - 7.83 + 1.03 \cdot \text{age} \text{ for males}$$

i.e.

$$SBP = 89.85 + 1.03 \cdot \text{age} \text{ for males}$$

NB: the effects for age and sex are *mutually adjusted*.

Standard error (SE) of age effect: 0.15.

Test for no age effect: $(1.03/0.15)^2 = 44.8, P < 0.0001$.

Standard error (SE) of sex effect: 1.47.

Test for no sex effect: $(-7.83/1.47)^2 = 28.4, P < 0.0001$.

15

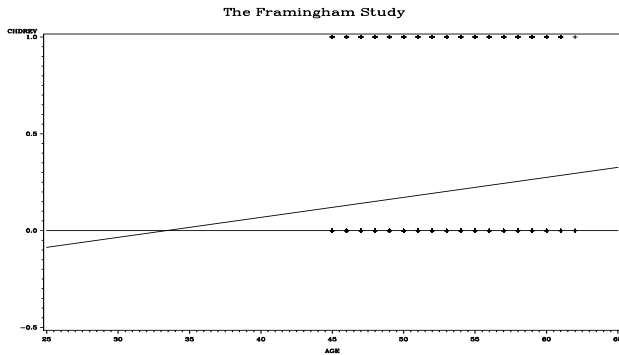
Logistic regression.

In epidemiology, outcome variables are most frequently binary, e.g.,

$$Y_i = \begin{cases} 1 & \text{if } i \text{ is diseased} \\ 0 & \text{if } i \text{ is not diseased} \end{cases}$$

Then *linear regression is no good*

16



The regression line will go outside 0 and 1 and we need *another regression model* for binary outcomes.

Purpose of logistic regression:

Relate a **binary outcome variable**, e.g.,

$$Y_i = \begin{cases} 1 & \text{if } i \text{ gets CHD} \\ 0 & \text{if } i \text{ does not get CHD} \end{cases}$$

to *explanatory variables* for individual i . Let

$$p_i = \text{Prob}(\text{individual } i \text{ gets CHD}) = \text{Prob}(Y_i = 1).$$

To start simply, consider one binary explanatory variable, e.g., sex

$$Z_i = \begin{cases} 1 & \text{if } i \text{ is a man} \\ 0 & \text{if } i \text{ is a woman} \end{cases}$$

Since linear regression, in general, is no good in this case, we *need another idea*.

We look at $\ln(\text{odds}) : \ln\left(\frac{p_i}{1-p_i}\right)$

which is *unbounded*, i.e., $\ln\left(\frac{p_i}{1-p_i}\right)$

can take both very large negative and very large positive values.

Model:

$$\ln\left(\frac{p_i}{1-p_i}\right) = a + bZ_i = \begin{cases} a & \text{females} \\ a + b & \text{males} \end{cases}$$

That is,

$$\begin{aligned} b = (a + b) - a &= \ln(\text{odds for males}) \\ &- \ln(\text{odds for females}) \\ &= \ln(\text{OR for males vs. females}) \end{aligned}$$

Similarly,

$$-b = a - (a + b) = \ln(\text{OR for females vs. males})$$

Framingham example

	Z = 0 (females)	Z = 1 (males)
Y = 0 (no CHD)	616	479
Y = 1 (CHD)	104	164
	720	643

$$OR = \frac{164 \cdot 616}{104 \cdot 479} = 2.03, b = \ln(OR) = \ln\left(\frac{164 \cdot 616}{104 \cdot 479}\right) = 0.71$$

$$a = \ln(\text{odds for females}) = \ln\left(\frac{104}{616}\right) = -1.78$$

SE of sex effect: 0.14; test for no sex effect:

$$(0.71/0.15)^2 = 25.7, P < 0.0001.$$

Compare usual chi-square statistic: 26.31, 1 d.f.

SAS PROC GENMOD.

In SAS there are several procedures which can perform logistic regression. We will be using PROC GENMOD.

```
proc genmod data=framing descending;
  class sex;
  model chdny=sex/dist=bin type3;
  estimate "m vs. f" sex 1 -1 /exp;
run;
```

and $\ln(OR)$ is estimated with the *last* level of SEX as *reference category*. From the ESTIMATE statement, so is OR with 95% confidence limits.

21

We also get test statistics for the effect of SEX:

“Wald” $W = 25.74, df = 1, P < 0.0001$

and (from the TYPE3 option):

“Likelihood ratio” $LR = 26.38, df = 1, P < 0.0001$

These are almost (but not exactly) identical to the standard chi-square statistic based on the two by two table: 26.31

22

Next: *two explanatory variables*:

$$Z_i = \text{sex}_i \quad \text{and} \quad V_i = \begin{cases} 1 & \text{if } i \text{ smokes} \\ 0 & \text{otherwise} \end{cases}$$

Data can be *summarized as two 2 by 2 tables in two ways*

	<u>Males</u>			<u>Females</u>	
	V = 0	V = 1		V = 0	V = 1
Y = 0	191	288	Y = 0	423	192
Y = 1	57	107	Y = 1	77	27

	<u>Smokers (V = 1)</u>			<u>Non-smokers (V = 0)</u>	
	Males	Females		Males	Females
Y = 0	288	192	Y = 0	191	423
Y = 1	107	27	Y = 1	57	77

23

In this way, we can *either*

1. Study the *effect of smoking (V) adjusted for sex (Z)*
- or
2. Study the *effect of Sex (Z) adjusted for smoking (V)* using the Mantel-Haenszel method.

1. $OR_{MH} = 0.97 \quad (= \exp(-0.034))$
 $X_{MH}^2 = 0.052$
2. $OR_{MH} = 2.03 \quad (= \exp(0.709))$
 $X_{MH}^2 = 22.96$

Conclusion: there is **no effect of smoking** adjusted for sex *but* there is **an effect of sex** adjusted for smoking.

24

Corresponding logistic regression model:

$$\ln\left(\frac{p_i}{1-p_i}\right) = a + b_1 Z_i + b_2 V_i$$

$$= \begin{cases} a & F \text{ no-smoke} \\ a + b_1 & M \text{ no-smoke} \\ a + b_2 & F \text{ smoke} \\ a + b_1 + b_2 & M \text{ smoke} \end{cases}$$

Note: $b_1 = (a + b_1) - a$
 $= (a + b_1 + b_2) - (a + b_2)$

$= \ln OR$ (males vs. females for given V_i)

and $b_2 = (a + b_2) - a$
 $= (a + b_1 + b_2) - (a + b_1)$

$= \ln OR$ (smoke vs. non-smoke for given Z_i)

SAS PROC GENMOD.

In PROC GENMOD both SEX and SMOKE are included in the logistic regression model as CLASS explanatory variables.

```
proc genmod data=framing descending;
    class sex smoke;
    model chdny=sex smoke/dist=bin type3;
    estimate 'm vs. f' sex 1 -1/exp;
    estimate 'smoke vs. not' smoke -1 1/exp;
run;
```

Thereby, $\ln(OR)$'s, *mutually adjusted*, are estimated for both variables with the last (highest) level as reference category. From the ESTIMATE statement, also OR 's with last level of SEX and first level of SMOKE as reference are obtained.

	<i>OR</i>	95% c.i.	Wald-test
m vs. f	2.00 (= exp(0.695))	1.50 to 2.67	22.57
sm. vs. non-sm.	1.03 (= exp(0.033))	0.78 to 1.37	0.05

Relation to/discrepancies from stratified (Mantel-Haenszel) analysis:

In logistic regression:

- *1 analysis only!* In the same model, we estimate the effect of sex adjusted for smoking and the effect of smoking adjusted for sex!
- In the calculations, the exposure variable and the confounder are treated identically

advantage/drawback?

The results from logistic regression and Mantel-Haenszel analysis are *not numerically exact identical but they are close:*

Mantel-Haenszel: $lnOR$ (males vs. females) = 0.709
 $lnOR$ (smoke vs. non-smoke) = -0.034

Logistic regression: $lnOR$ (males vs. females) = 0.695
 $lnOR$ (smoke vs. non-smoke) = 0.033

Also: chi-square tests are close

Explanatory variable with several levels:

age= X

$$X_i = \begin{cases} 0 & \text{if } i \text{ has age } 45 - 48 \\ 1 & \text{if } i \text{ has age } 49 - 52 \\ 2 & \text{if } i \text{ has age } 53 - 56 \\ 3 & \text{if } i \text{ has age } 57 - 62 \end{cases}$$

Summarize in 2 by 4 table

	$X = 0$	$X = 1$	$X = 2$	$X = 3$	
	45-48	49-52	53-56	57-62	
$Y = 0$	308	298	254	235	1095
$Y = 1$	51	61	64	92	268
	359	359	318	327	1363

(NB: Both males and females)

Again, we may test whether the explanatory variable, X , affects the outcome, Y , using a chi-square test statistic.

In the Mantel-Haenszel chi-square test statistic for a 2 by 2 table, we have calculated the *EXPECTED NUMBER* in one cell.

In the general chi-square test, we calculate the corresponding expected number in *ALL CELLS* and add, for all cells,

$$(OBS - EXP)^2 / EXP$$

E.g., $Y = 0, X = 2$: $OBS = 254$

$$EXP = \frac{1095}{1363} \cdot 318 = 255.5$$

Contribution to test statistic

$$\frac{(254 - 255.5)^2}{255.5} = 0.008$$

Similarly, for the other 7 cells in the table

$$\sum \frac{(OBS - EXP)^2}{EXP} = 23.29 \sim \chi_3^2, P < 0.001$$

$$\begin{aligned} 3. \text{ d.f.} &= (\text{columns} - 1) \times (\text{rows} - 1) \\ &= (4 - 1) \times (2 - 1) \end{aligned}$$

Conclusion: age affects CHD-risk significantly; the *risk seems to increase with age:*

Measures for the strength of the effect:

$$\begin{aligned} OR_1(X = 1 \text{ vs. } X = 0) &= \frac{308 \cdot 61}{298 \cdot 51} = 1.24 \\ &= \exp(0.21) \\ OR_2(X = 2 \text{ vs. } X = 0) &= 1.52 \\ &= \exp(0.42) \\ OR_3(X = 3 \text{ vs. } X = 0) &= 2.36 \\ &= \exp(0.86) \end{aligned}$$

The chi-square statistic tests whether

$$OR_1 = OR_2 = OR_3 = 1 \quad (3 \text{ d.f.})$$

**One categorical variable
with several (4) categories:**

$$X_i = \begin{cases} 0 & \text{if } i \text{ has age } 45 - 48 \\ 1 & \text{if } i \text{ has age } 49 - 52 \\ 2 & \text{if } i \text{ has age } 53 - 56 \\ 3 & \text{if } i \text{ has age } 57 - 62 \end{cases}$$

In *logistic regression*, $\ln(OR)$'s for the categories relative to a **reference category** are estimated:

$$\ln\left(\frac{p_i}{1 - p_i}\right) = \begin{cases} a & \text{if } i \text{ has age } 45 - 48 \\ a + b_1 & \text{if } i \text{ has age } 49 - 52 \\ a + b_2 & \text{if } i \text{ has age } 53 - 56 \\ a + b_3 & \text{if } i \text{ has age } 57 - 62 \end{cases}$$

$$\begin{aligned} b_1 &= \ln(OR)(1 \text{ vs. } 0) = 0.21 \\ b_2 &= \ln(OR)(2 \text{ vs. } 0) = 0.42 \\ b_3 &= \ln(OR)(3 \text{ vs. } 0) = 0.86 \end{aligned}$$

NB: Different computer programs may choose the reference category differently.

SAS PROC GENMOD.

In SAS PROC GENMOD we get, with the present coding of AGE

	$\ln(OR)$	95% c.i.	<i>LR</i> -test
0 vs. 3	-0.86	-1.24 to -0.48	22.60
1 vs. 3	-0.65	-1.01 to -0.28	
2 vs. 3	-0.44	-0.81 to -0.08	

with 3 d.f.

If we want the same results as above we need to let the youngest become the reference category (use ESTIMATE):

	<i>OR</i>	95% c.i.
1 vs. 0	1.24	0.83 to 1.85
2 vs. 0	1.52	1.02 to 2.28
3 vs. 0	2.36	1.61 to 3.46

37

SAS-øvelser.

De følgende spørgsmål drejer sig alle om Framingham studiet, og der henvises til variabellisten i forelæsningsnoterne. De første øvelser har til formål at rekonstruere resultater fra forelæsningsnoterne

- **1.** Brug programmet `framing.sas` til at læse filen `framing.txt` og konstruere en ny udfaldsvariabel, `CHDNY`, som er
 - 1 hvis `CHD=2, 3, ..., 10`
 - 0 hvis `CHD=0`
 - “missing” (‘.’) hvis `CHD=1`.

Tilføj kommandoer, som laver en anden ny variabel, `SMOKE`, som er

- 1 hvis `cig>0`
- 0 hvis `cig=0`

38

- “missing” (‘.’) hvis `cig=.`

og udfør Mantel-Haenszel analyserne af køn stratificeret for rygning og rygning stratificeret for køn.

- **2.** Konstruer variabelen `AGENCY` ved at inddele i 4 grupper efter `AGE` og sammenlign 18-års risikoen for `CHD` i de 4 aldersgrupper (ved et χ^2 -test).
- **3.** Udfør logistiske regressionsanalyser af 18-års risikoen for `CHD` med forklarende variable:
 - køn
 - køn og rygning
 - alder (dvs. `AGENCY`)
 - alle 3

Fortolk resultaterne.

39

- **4.** Konstruer en variabel svarende til en passende inddeling efter `SBP` og estimer i en logistisk regressionsanalyse effekten af denne variabel med og uden justering for køn og alder.

40