

MPH specialmodul Epidemiologi og Biostatistik

Logistisk regression:

Kvantitative forklarende variable

Interaktion

27. april 2009

www.biostat.ku.dk/~pka/mphspec09

Per Kragh Andersen

1

SAS PROC GENMOD

In PROC GENMOD, the variable **AGENCY** is included in the model as an explanatory variable but NOT as a 'CLASS' variable:

```
proc genmod data=framing descending;  
  model chdny=agency/dist=bin type3;  
  estimate 'age odds ratio' agency 1/exp;  
run;
```

NB: To use **AGENCY** as a quantitative explanatory variable its levels must be *ordered*.

3

Quantitative explanatory variables

When analysing **AGE** we saw that the risk (odds) increased with age. Perhaps we can use *only 1 parameter* to describe the increase in $\ln(\text{odds})$ between neighbouring categories:

$$\ln\left(\frac{p_i}{1-p_i}\right) = a + bX_i = \begin{cases} a & X_i = 0 \\ a + b & X_i = 1 \\ a + 2b & X_i = 2 \\ a + 3b & X_i = 3 \end{cases}$$

The resulting estimate $b = 0.29$ is a sort of *average* between: 0.21, 0.42-0.21, 0.86-0.42

Does this model fit? "Straight line?"

2

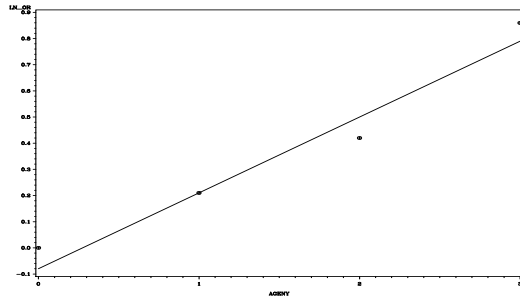
Output includes:

- An estimate of $b = \ln(OR)$, increase/decrease per unit of the quantitative explanatory variable
- $b = 0.29$, $SE = 0.062$
- The corresponding *OR*-estimate, $\exp(b) = 1.33$ with 95% confidence limits: from 1.18 to 1.50
- A "Wald" chi-square test for no effect of **AGENCY**:
Test = 21.29, 1 *df*, $P < 0.0001$ (this is found in the table which also gives $b = \ln(OR)$.)
- An "LR" chi-square test for no effect of **AGENCY**:
Test = 21.74, 1 *df*, $P < 0.0001$ (this is given because of the **TYPE3** option and is found in the bottom of the output)

4

Tests for trend and tests for linearity.

The Wald and LR tests for a quantitative explanatory variable always have **1 d.f.** It is called a TEST FOR TREND, takes the linear effect for granted and tests whether the SLOPE IS ZERO.



Tests for trend and tests for linearity.

To examine *whether the linear model is acceptable* we need to compare the two models

- (1) AGENY as a CLASS variable
- (2) AGENY as a quantitative variable

This can be done using SAS PROC GENMOD by fitting both models (1) and (2) and subtracting the two values of 'Deviance'. These are:

Model (2):	1329.51
Model (1):	1328.65
Difference	0.86

This difference is a chi-squared distributed test statistic with $3-1=2$ df. where: 3=parameters for AGENY in model (1) 1=parameters for AGENY in model (2). The value 0.86 in χ^2_2 gives $P = 0.65$.

Tests for trend and tests for linearity.

CONCLUSION:

the test is NOT significant

⇒

the linear model IS acceptable.

This test statistic is called the (“likelihood ratio”)

TEST FOR DEPARTURES FROM TREND

or

TEST FOR LINEARITY.

NOTE: the test for trend always has 1 df. (it corresponds to 1 “slope”).

The test for linearity has a number of df which depends on the number of levels for the CLASS variable (df=number of levels - 2).

If the model fits then we may include the variable
AGE (in years)

in the model (i.e., *NO grouping*)

$$\ln\left(\frac{p_i}{1-p_i}\right) = a + b \cdot AGE_i$$

The analysis gives

$$b = 0.066$$

Interpretation: For each year

$$\exp(b) = 1.07$$

is the *factor by which odds for CHD increases.*

The model with AGE as a truly quantitative variable can not be tested against the model with the grouped AGENY.

To test the new model we may define a new variable by including a command like:

AGESQ = AGE*AGE;

in a DATA step (a non-linear (“quadratic”) term).

This new variable is included in the model including AGE and we may test whether AGESQ may be left out:

	Estimate (= ln(OR))	Standard Error	Wald Chi-Square	P
AGE	0.0577	0.3424	0.03	0.87
AGESQ	0.0001	0.0032	0.00	0.98

CONCLUSION: AGESQ may be left out \Rightarrow the linear model (including only AGE) IS acceptable.

Including quantitative explanatory variables: Pros and cons.

Pros.

- The effect of arbitrarily chosen cut points is avoided.
- The number of parameters in the model is reduced
- and, thereby, in principle, the model is simpler.

Cons.

- However, the slope parameter may be more difficult to interpret than OR's relative to a reference category.

Testing for interaction (effect-modification) using logistic regression

Two explanatory variables *interact* (= the effect of one variable is modified by the value of another variable) *if*

the effect of one variable on the outcome depends on the value of another variable, e.g.,

the age-effect on the CHD-risk *may be different* for men and women.

Interaction in SAS PROC GENMOD.

In SAS PROC GENMOD we may add INTERACTION TERMS to the logistic regression model and thereby TEST FOR NO INTERACTION. For the two variables AGENY and SEX the interaction term added to the model is: AGENY*SEX

If both are CLASS variables then the LR chi-square statistic for no interaction (obtained from the TYPE3 option) has:

$$3 = (4 - 1) \cdot (2 - 1) \quad d.f.$$

when AGENY has 4 levels and SEX has 2 levels.

The statistic 5.46 then gives $P = 0.14 \Rightarrow$ NO significant interaction (NO significant effect-modification).

If AGENY is quantitative then

AGENY*SEX has only 1 d.f.

and with the test statistic 4.59 we get $P = 0.03$

⇒ INTERACTION.

With the quantitative variable AGE (in years, i.e. no grouping) the LR chi-square statistic for

AGE*SEX is 3.17 with 1 d.f.,

$P = 0.08$.

What to believe??

We need some estimates to see what is going on.

If we want to see estimates for AGENY within categories of SEX then we may write the model *without* the term AGENY, i.e.

```
model chdny=sex ageny*sex/dist=bin type3;
```

instead of

```
model chdny=sex ageny ageny*sex/dist=bin type3;
```

With the oldest as reference we get $\ln(OR)$'s:

AGENY	0	1	2	3
SEX=1: M	-0.53	-0.48	-0.27	0
SEX=2: K	-1.50	-0.79	-0.61	0

Using the ESTIMATE command we may (with some efforts)

1. choose an appropriate reference category (e.g., the youngest)
2. obtain OR 's rather than just $\ln(OR)$'s

yielding the OR 's:

AGENY	0	1	2	3
SEX=1: M	1	1.05	1.30	1.69
SEX=2: K	1	2.04	2.44	4.49

```
proc genmod data=framing descending;
class sex ageny;
model chdny=sex ageny*sex/dist=bin type3;
estimate '2-1 m' ageny*sex -1 1 0 0 0 0 0 0/exp;
estimate '3-1 m' ageny*sex -1 0 1 0 0 0 0 0/exp;
estimate '4-1 m' ageny*sex -1 0 0 1 0 0 0 0/exp;
estimate '2-1 k' ageny*sex 0 0 0 0 -1 1 0 0/exp;
estimate '3-1 k' ageny*sex 0 0 0 0 -1 0 1 0/exp;
estimate '4-1 k' ageny*sex 0 0 0 0 -1 0 0 1/exp;
run;
```

With AGENY as a quantitative explanatory variable we get:

	$\ln(OR)$	OR
Males	0.18	1.20
Females	0.46	1.58

Conclusion: the age effect is stronger for females than for males. The age effects may be considered linear and in the linear model there is significant difference between the age effects for males and females.

Interaction: summary.

Testing for interaction in SAS PROC GENMOD is quite easy!

When testing for interaction, there are two possibilities:

- (I) The interaction is unimportant/insignificant
- (II) The interaction is important/significant

If (I): The good old additive model is used.

If (II): E.g., the age effect is different for men and women and we then want to see the two effects.

These effects may be computed using SAS PROC GENMOD with some efforts.

Concluding remarks concerning logistic regression:

Getting out estimates from logistic regression in, e.g., SAS is **easy** (technically).

The **“art”** is *which estimates to ask for*.

1. *Which variables to include?*
(The “exposure” variable(s) should, of course, be there but which confounders?)
2. *How should variables be included?*
(Categorical (“CLASS”) variables, quantitative variables with few categories, or “real” quantitative variables)
3. *Which interactions to study?*

Some recommendations:

- Model strong confounders in a “robust” way, i.e., do not make too strict assumptions about their effect.
- Limit the number of interactions to study to those which are pre-specified.

There are **many-many** possibilities!

In published results using logistic regression, all these choices have been made (perhaps without thinking too much about it) before presenting tables and graphs and other results.

⇒ Remember your scepticism when reading articles!

SAS-øvelser.

De følgende spørgsmål drejer sig alle om Framingham studiet, og der henvises til variabellisten i forelæsningsnoterne.

- **1.** Konstruer en variabel svarende til en passende inddeling efter SBP (f.eks. delt i 4 grupper ved 120, 140 og 180) og estimer i en logistisk regressionsanalyse effekten af denne variabel efter justering for køn og alder.
- **2.** Undersøg om denne variabel kan indgå *lineært* i modellen (dvs. test for “departures from trend”).
- **3.** Undersøg om der er interaktion mellem kn og blodtryk - både når blodtryk indgår lineært (som i spm. **2.**) og som CLASS variabel. Estimer i begge tilfælde effekten af blodtryk for både mænd og kvinder.

- **4.** Erstat den grupperede variabel i spm. **1.** med den oprindelige variabel SBP målt i *mmHg*. Fortolk resultaterne.