Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 18.5 Single Explanatory Variable on an Ordinal Scale

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11), Part IV (Ch 13, 14)
18 Binomial Response Variables
18.1 Logistic Regression (Dose-Response)
18.2 Single Factor. Prospective Analysis
18.3 Single Factor. Retrospective Analysis
18.4 Single Fixed Factor.
18.5 Single Explanatory Variable. Ordinal Scale.
18.6 Two Categorical Explanatory Variables
18.7 Logistic ANCOVA

Ch18.xls

on chalk board

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning
ReCap Part II (Chapters 5,6,7) Hypothesis testing and estimation
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12,13,14,15) GLM with more than one explanatory variable
ReCap (Ch 16,17). Generalized Linear Model. Poisson response variables.
ReCap (Ch 18). We used logistic regression to compare the odds across 2 or more levels of a categorical variable.

Today: Binomial response variable with an explanatory variable on an ordinal (rank) scale.

Wrap-up.

Categorical variables sometimes occur on an ordinal scale. We use binary logistic regression to compare successive levels.

Binomial response variables with an ordinal scale explanatory variable.

We sometimes have an explanatory variable on an ordinal scale - we list categories from small to large, be we cannot state the exact numerical difference from one category to the next. An common example is survey data, with frequency of responses in categories ranging from strongly disagree, to disagree, to neutral, to agree, to strongly agree. Another common example is the construction of categories from interval or ratio scale data, where one of the categories is truncated at zero, or open ended (e.g. 10, 20, 30, 40 or more).

To illustrate the analysis of binary data aganst an ordinal scale explanatory variable, we will use another cross-sectional data (case-control study), taken from Zang and Wynder 1992, as reported in Sokal and Rohlf 2012 Ex 17.20, p815)

First, the odds and odds ratios for having cancer.

	Lung Cancer (males) present	absent	total	% present	cancer odds	Odds Ratio
smoke	522	866	1388	37.6%	0.600 : 1	33.3
non-smokers	15	822	837	1.79%	0.018 : 1	

Now, the relative risk (risk for smokers relative to non-smokers)

	Lung Cancer (males) present	absent	total	Risk(%)	Relative Risk
smoke	522	866	1388	37.6%	21
non-smokers	15	822	837	1.79%	

Example. Comparing several proportions in a retrospective study.
Here are the cross-sectional data (case-control study), taken from Zang and Wynder 1992.

	Lung Cancer (males) Present	absent	total	%	odds of cancer	odds ratio
non-smokers	15	822	837	1.79%	0.018 : 1	
1-10 cig	36	136	172	20.9%	0.265 : 1	14.51
11-20 cig	133	328	461	28.9%	0.405 : 1	22.22
21-40 cig	226	311	537	42.1%	0.727 : 1	39.82
≥41 cig	127	91	218	58.3%	1.396 : 1	76.48

The odds appear to increase substantially, depending on level of smoking. However, we cannot use regression because the last category is open ended.

1. Model and data equations.

Verbal. Risk of cancer in male smokers relative to non-smokers increases with number of cigarettes smoked.

Graphical Plot of Odds ratio for 4 groups (smoking) and 1 group (nonsmoking)

Response variable: odds of cancer

Explanatory variable: smoke cigarettes or not

Write formal model $Odds = e^{(\beta_o)} e^{(\beta_S S)}$

$e^{\beta_{ref}}$ = Cancer odds, reference or control group (non-smokers)

$e^{\beta_S S}$ = Odds ratios, one for each level of smoking

$e^{(\beta_o + \beta_S S)}$ = Cancer odds, smokers (not used)

2. Execute analysis.

Place data in model format for package with generalized linear model routine:

Binomial response variable in two columns, success and trials

Column = Ncancer, with response variable the number with cancer (presences)

Column = Ntot, with response variable the total number of people

Column labelled Smoke, with explanatory variable

Smoke = No, 1-10, 11-20, 21-40, 41+

```
MTB > print c1-c4
 Row   Ncancer   Ntot   Smoke

   1      15     837      0
   2      36     172      1-10
   3     133     461      11-20
   4     226     537      21-40
   5     127     218      41+
```

Minitab format

```
Data A;
  Input Ncancer Ntot Smoke $;
  Cards;
     15  837 No
     36  172 Cig1-10
    133  461 Cig11-20
    226  537 Cig21-40
    127  218 Cig41+
 ;
```

SAS command file

2. Execute analysis.

For categorical variables, some packages require the use of 'dummy variables' in a logistic regression routine. Here are the dummy variables for use in a regression routine.

```
MTB > print c1-c6
 Row   Ncancer   Ntot  1-10cig  11-20cig  21-40cig  41+cig

   1      15     837      0         0         0       0
   2      36     172      1         0         0       0
   3     133     461      0         1         0       0
   4     226     537      0         0         1       0
   5     127     218      0         0         0       1
```

Minitab format

Code the model statement in a statistical package according to the GzLM

```
MTB > BLogistic 'Ncancer' 'Ntot' = 'Cig1-10' 'Cig11-20' 'Cig21-40' 'Cig41+';
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.
```

Minitab command lines

Here is the model statement in a generalized linear model routine (SAS)

```
Proc Genmod; Classes Smoke;
  Model Ncancer/Ntot = Smoke/
  Link=logit dist=binomial type1 type3;
```

SAS command file

2. Execute analysis.

This is another saturated model. We have 5 fitted values, one for each observation. Fitted values from model output. The residuals are zero.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept		1	-4.0037	0.2605	-4.5143	-3.4930	236.13
Smoke	Cig1_10	1	2.6746	0.3210	2.0455	3.3036	69.44
Smoke	Cig11_20	1	3.1010	0.2801	2.5521	3.6500	122.58
Smoke	Cig21_40	1	3.6844	0.2748	3.1458	4.2231	179.75
Smoke	Cig41plu	1	4.3370	0.2945	3.7598	4.9143	216.84

SAS output file

2. Execute analysis.

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-4.0037	0.2605	-15.37	0.000			
1-10cig	2.6746	0.3209	8.33	0.000	14.51	7.73	27.21
11-20cig	3.1010	0.2801	11.07	0.000	22.22	12.83	38.47
21-40cig	3.6844	0.2748	13.41	0.000	39.82	23.24	68.24
41+cig	4.3370	0.2945	14.73	0.000	76.48	42.94	136.22

Minitab output

The first class listed (smoker) is the reference group (intercept).

$e^{\beta_{ref}} = e^{-4.00369} = 0.018$ Odds, reference group (non-smokers)

$e^{\beta_S S} = e^{2.6746 \cdot 1} = 14.51$ Odds ratio, smokers (1-10 cig/day) relative to nonsmokers. S=1 for this category

$e^{\beta_S S} = e^{3.1010 \cdot 1} = 22.22$ Odds ratio, smokers (11-20 cig/day) relative to nonsmokers. S=1 for this category.

$e^{\beta_S S} = e^{3.6844 \cdot 1} = 39.82$ Odds ratio, smokers (21-40 cig/day) relative to nonsmokers. S=1 for this category.

$e^{\beta_S S} = e^{4.3370 \cdot 1} = 76.48$ Odds ratio, smokers (41^+ cig/day) relative to nonsmokers. S=1 for this category

3. Use parameter estimates to calculate residuals, evaluate model.

Residuals all zero.

Binomial assumption appropriate assuming risk of developing cancer in one subject was not altered by another member developing cancer.

4. Population.

In this case, we plan to infer to a real population, not hypothetical population of "all possible measurements, given the protocol for measurement." We can estimate the odds ratio for the sample. Because the risk in the population is small, we can use the odds ratio from the case-control sample to estimate the relative risk in the population of people from which the patients with lung cancer came. Publications in the medical and health sciences routinely list characteristics of the sample (age, gender, etc) as guide to the relevant population.

5. Decide on mode of inference. Is hypothesis testing appropriate?
There is little doubt, from the parameter estimates, that risk increases with number of cigarettes smoked. It is of more interest to examine the change in risk at each level. We skip to step 10.

10. Analysis of parameters of biological interest.

The odds ratio (relative risk) of cancer increases with increasing cigarette use, relative to the reference group (nonsmokers) with similar characteristics. For case-control studies, the odds ratio is reported but the odds in each group is not reported because it is not a representative sample from that group.

Logistic Regression Table

					Odds	95% CI	
Predictor	Coef	SE Coef	Z	P	Ratio	Lower	Upper
Constant	-4.0037	0.2605	-15.37	0.000			
1-10cig	2.6746	0.3209	8.33	0.000	14.51	7.73	27.21
11-20cig	3.1010	0.2801	11.07	0.000	22.22	12.83	38.47
21-40cig	3.6844	0.2748	13.41	0.000	39.82	23.24	68.24
41+cig	4.3370	0.2945	14.73	0.000	76.48	42.94	136.22

Minitab output

The confidence limits are fairly wide at any one level of cigarette use. The odds ratios increase at each level, so we conclude that the odds (and hence relative risk) increase with increased smoking. We could estimate the rate of increase in risk with increase in use by defining cigarettes per day as an explanatory variable on a ratio type of scale, instead of the ordinal scale used in this example.

Confidence limits allow us to report the increase in risk relative to non-smokers at any level of cigarette use. This is more informative than an overall test (below) of whether risk differs among levels of use.

As a matter of information here is the overall test - the ANODEV table.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	551.2222			
Smoke	0.0000	4	551.22	<.0001

SAS output file

The goodness of fit of the null model to the data is $G^2 = 551.2$
The fit of the alternative model to the data is perfect $G^2 = 0.0$
The improvement is $\Delta G^2 = 551.2$

The classical goodness of fit test produces essentially the same result. The null hypothesis is the expected proportion: 537 with cancer / 2225 subjects = 0.24.

f	$= \hat{p} \cdot N_i$	+ residual	$2 \ln L = 2 f \ln(f / \hat{p} \cdot N_i)$
15	$= 0.24 \cdot 837$	- 187	-78
36	$= 0.24 \cdot 172$	- 6	-10
133	$= 0.24 \cdot 461$	+ 22	47
226	$= 0.24 \cdot 537$	+ 96	251
127	$= 0.24 \cdot 518$	+ 74	224