Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 18.3 Single Factor. Retrospective Analysis

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 Random 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 quantify the intensity of natural selection (Kettlewell data). This is called a prospective analysis. It is a longitudinal analysis because we followed individuals through time.

Today Retrospective analysis of binomial response across two levels of a single factor.

Wrap-up.

We used logistic regression to quantify the risk of cancer in smokers. The data were cross-sectional, in which we contrast groups with different histories at a single point in time. This is called retrospective analysis.

Binomial response variables –Retrospective Analysis.
Odds Ratio and Relative Risk

Does smoking increase the risk of cancer ?
What seems obvious today was once not obvious. Lung cancer emerged suddenly as a major health issue in the US in the middle of the 20th century. A rigorous study requires a longitudinal or prospective study, such as the moth mark-recapture study. In a prospective study, individuals who smoke would be carefully matched with non-smokers with similar characteristics (age, health status, *etc*.). This cohort is then followed over many years, to obtain the proportion of smokers that develop lung cancer, for comparison to the proportion of non-smokers that develop lung cancer. The result is rigorous, costly, and won't produce results for decades.

In a landmark publication, Jerome Cornfield showed that the relative risk in a population could be estimated from a case-control study. The data for such a study consist of patients with similar characteristics, sorted into those with and without the disease, to estimate the risk for those exposed and not exposed to a suspected cause. We then use the odds ratio to estimate the relative risk in the population. This is a retrospective or cross-sectional study. Unlike a prospective study, we do not begin with a known set of cases, then score them at a later time as having an attribute (disease) or not. Instead, we collect cases at a single point in time and assign them to categories in a 2 by 2 table: having or lacking the disease, and having or lacking exposure to the suspected cause. The result is a sample that is clearly far from representative of the population. Cornfield showed that the bias in the sample was large, and could be corrected. The results for lung cancer were clear, and set in motion research that established cigarette tars as the causal agent for lung cancer. Further, the publication established the mathematical basis for using case-control samples to estimate risk in a population.

To illustrate retrospective analysis in its modern form we will use the data presented by Cornfield (1951), even though the publication never mentions the odds ratio and was published well before modern methods for retrospective analysis (Breslow and Day 1980). Cornfield used percentage data and numbers (*N*) from Shrek *et al.* (1950) who reported data for US Army veterans with cancer, 35 with lung cancer and 171 with other cancers. The veterans were in the 40-49 age group, taken from records of over 5000 veterans presenting with tumors at Veteran's Administration hospitals in Chicago, from 1940 to 1942. The veterans in this age cohort were thus born between 1891 and 1902, and so were veterans of World War I. Cigarettes were issued as a ration to US troops in World War I (Goodman (2005). They were used as barter in the front lines, and were one of few reliefs from the psychological stress of trench warfare.

References

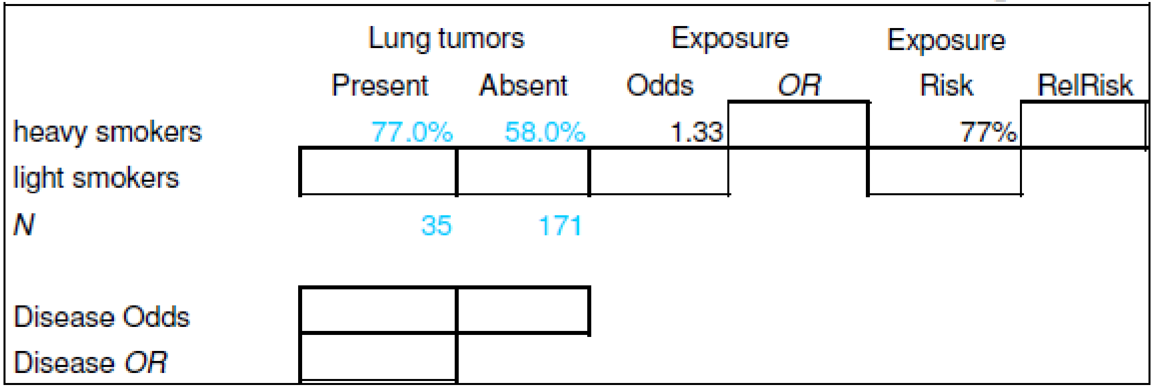
Breslow, N.E. and Day, N.E. (1980). *Statistical Methods in Cancer Research. Volume I – The Analysis of Case_Control Studies*. Lyon: International Agency for Research on Cancer.

Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Application to cancer of the lung, breast and cervix. *Journal of the National Cancer Institute* 11: 1269–1275.

Goodman, J. (Ed.). (2005). *Tobacco History and Culture: An Encyclopedia*. Detroit: Scribner's

Schrek, R. Baker, L.A., Ballard, G.P., Dolgoff, S. 1950. Tobacco smoking as an etiologic factor in disease. I. Cancer. *Cancer Research* 10:49–58.

Retrospective Analysis. Odds Ratios and Relative Risk.
Here are the percentages and cohort numbers (*N*) from Schrek *et al*, used by Cornfield (1951).



	Lung tumors		Exposure		Exposure	
	Present	Absent	Odds	*OR*	Risk	RelRisk
heavy smokers	77.0%	58.0%	1.33		77%	
light smokers						
N	35	171				
Disease Odds						
Disease *OR*						

The odds for heavy smokers are p/(1-p) = 77% / 58% = 1.33
Now calculate the percentage of light smokers with tumors and without.
From this, calculate the odds for light smokers.
Next calculate the exposure odds ratio for heavy relative to light smokers.
The exposure risk for heavy smoker is shown.
Fill in the exposure risk for light smokers and calculate the relative risk.

Next, calculate the disease odds for tumors present, defined as heavy smoker odds / light smoker odds. Then calculate the disease odds for tumors absent.

Did you discover that the disease *OR* is the same as the disease *OR*?
If not, go back and check your work.

Retrospective Analysis. Odds Ratios and Relative Risk.

The odds ratio is also known as the cross-product ratio.

The cross product ratio for counts in a 2 X 2 table is defined as follows.

$$CPR = \frac{a}{b} \cdot \left(\frac{c}{d}\right)^{-1}$$

Prospective Study

	Disease (Cancer)	
Exposure	Present	Absent
smokers	a	c
non-smokers	b	d

Here is the relative risk for smokers.

$$RR = \frac{a}{a+b} \cdot \left(\frac{c}{c+d}\right)^{-1}$$

Where the values in the table are percentages we have:

$$CPR = \frac{a}{1-a} \cdot \left(\frac{c}{1-c}\right)^{-1} \qquad \text{and} \qquad RR = \frac{a}{c}$$

$$CPR = \frac{a}{c} \cdot \left(\frac{1-a}{1-c}\right)^{-1}$$

Where the risk in the population is small, 1-*a* is large, the ratio (1-*a*)/(1-*c*) approaches a value of 1, and the CPR is an estimate of the relative risk (Cornfield 1951).

The CPR in the Schrek data for veterans age 40-49 was 2.4 times higher in smokers, compared to control. Can the observed increase in odds be dismissed as mere chance? We use the generic recipe for the Generalized Linear Model.

1. **Model and data equations**.

Verbal. Risk of tumors in male smokers age 40-49 increases relative to non-smokers.

Graphical Not useful. We have only 4 numbers.

Response variable: Odds of lung tumors

Explanatory variable: heavy smoking vs light smoking.

Note that Shrek *et al* reported percent smokers in two groups, those with and those without lung tumors.

1. **Model and data equations**.

Write formal model

Distribution $Ntumors{\sim}Binomial(N, \pi)$

Link Odds = e^{η}

$\eta = \beta_{ref} + \beta_{Smoke}Smoke$

$e^{\beta_{ref}} =$	Tumor odds, control group
$e^{\beta_{Smoke}} =$	Odds ratio, case (heavy) vs control (light smoker)
$e^{\beta_{ref}+\beta_{Smoke}\cdot Smoke}$	Tumor odds, case group (heavy smokers)

2. **Execute analysis.**

Place data in model format for generalized linear model routine. Data can be listed by patient. In this example 206 patients, each scored as LungTumor Y/N, and Smoke Light/Heavy. For simple analyses, with only a single explanatory variable, we can list data by variable (Lung tumor present or absent, Smoking light or heavy.

The binomial response variable is listed in two columns, success and trials

Column *N*	number of patients in each group
Column *Ntmr*	number with lung tumors in each group
Column *Smoke*	factor with two levels, heavy or light smoker

	Lung Tumors		
	Present	Absent	Total
heavy smokers	27	99	126
light smokers	8	72	80

Code the model statement in a statistical package. Here is the logistic regression code in Minitab.

```
MTB > BLogistic 'Ntmr' 'Ntot' = 'Smoker';
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.                              Minitab commands
```

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

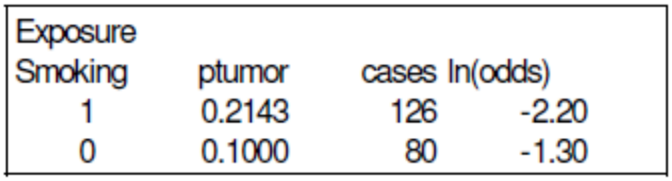
```
Proc Genmod; Classes Smoker;
  Model Ntmr/Ntot = Smoker/
  Link=logit dist=binomial type1 type3;       SAS command file
```

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

```
glm(Ntmr/Ntot ~ Smoker,
      family = binomial(link = logit), data = Cornfield,
      weights = cases)                                  R code
```

2. Execute analysis.

For R/S+ we can recode the response variable *Ntumor* to a proportion *ptumor*. Note the coding of exposure as 1 (heavy smoking).= and zero (light smoking).



Exposure Smoking	ptumor	cases	ln(odds)
1	0.2143	126	-2.20
0	0.1000	80	-1.30

```
Call: glm(formula = ptumor ~ Smoking,
      family = binomial(link = logit), data = Cornfield,
      weights = cases)
```

Code generated by S+

The binomial approach based on 206 cases (0 or 1) with 205 degrees of freedom appears to be far better than the approach based on only 1 degree of freedom for the aggregated approach. However, we will be testing the improvement in fit due to only 1 parameter, which drops the degrees of freedom either from 205 to 204 df, or from 1 to 0 df. The improvement in fit will be the same for both approaches.

Residuals will be zero for the aggregated approach because this is a saturated model (two parameters and two observations).

Generalized linear model routines require a statement of:
-the error structure (binomial in this case)
-the link function (logit in this case).
-the structural model (explanatory variables).

2. Execute analysis.

Here are the parameter estimates.

	Value	Std. Error	t value	exp(Value)
(Intercept)	-2.20	0.3727	-5.90	0.111
Smoking	0.90	0.4313	2.08	2.455

The intercept is the logarithm of the odds for one of the cases. In this example the intercept is the log odds for light smokers because we listed light smoking as a lower value (zero) than heavy smoking (value of 1). The parameter estimate is Odds = exp(-2.20) = 0.111. The smoking coefficient is the logarithm of the odds ratio for heavy smokers relative to light smokers. The parameter estimate is *OR* = exp(0.9) = 2.455

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

Residuals are all zero, so we can't use them for evaluation. We assume that the 35+171 = 206 trials were independent events. That is, developing a lung tumor does not depend on the chance of another participant in the study developing a lung tumor.

4. What is the evidence?

$LR = e^{4.814\,/\,2} = 11.1$

There is some evidence (LR>10) for greater odds of tumors in heavy than light smokers.

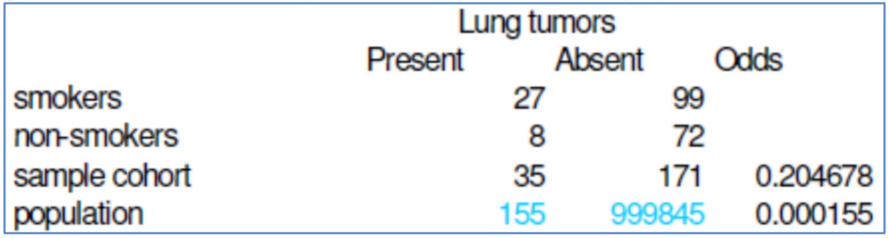
	Df	Deviance	Resid df	Residual Deviance
NULL			1	4.814
Exposure	1	4.814	0	0

5. Decide on mode of inference. Is hypothesis testing appropriate?

At the time of the Schrek study there was no knowledge of how cigarette smoking causes cancer. There was considerable debate, driven by economics and public health concerns. The debates about cigarette smoking, in the 1950s, were similar in many ways to subsequent public debates about the effect on community health of adding flouride to public water supplies, or the debates in the century about the safety of oil and gas extraction by hydraulic fracturing. Evidence (an odds ratio) and rational treatment of uncertainty (reporting Type I error, or effect size and confidence limits) is important in any rational debate. Evidence and a measure of uncertainty are also guides to allocation of public funds to research. We will use confidence limits to evaluate uncertainty on the evidence reported by Cornfield (1951) on the Schrek *et al* (1950) result.

5. Population.

Cornfield’s 1951 publication showed how to infer from the sample to a larger population, all males age 40-49 in the American midwest in 1940-42. At the time, the risk of developing lung cancer for this population was 155 in a million.
Here are the odds of lung cancer for the sample, compared to the population.



	Lung tumors		
	Present	Absent	Odds
smokers	27	99	
non-smokers	8	72	
sample cohort	35	171	0.204678
population	155	999845	0.000155

The sample is hugely biased (veterans with cancer of all types). To correct the bias, Cornfield used the population disease risk (155 per million) to recompute the relative risk for the population.
Subsequent publications in the medical and health sciences routinely list characteristics of the sample (age, gender, *etc*) as a guide to the relevant population. They rarely report the disease risk in the sample versus the population (as Cornfield did) or the exposure risk. Cornfield corrected for bias in the disease risk, but did not correct for exposure risk in the sample, compared to the population. A higher exposure risk would be expected for the sample (WW I veterans) given free access to a highly addictive substance (nicotine in cigarettes) in a war zone.

10. Analysis of parameters of biological interest.

Odds for light smokers $e^{\beta_o} = e^{-2.2} = 0.111$
Odds ratio for heavy relative to light smokers $e^{\beta_{Smoke}} = e^{0.9} = 2.455$
Confidence limits for β_{Smoke} are $\hat{\beta}_{Smoke} \pm 1.96 \cdot sterr$
The standard error on the estimate of $\hat{\beta}_{Smoke}$ was 0.4313.
The confidence interval is from 0.0526 to 1.74
The confidence interval for the *OR* is exp(0.0526) to exp(1.74),
i.e. from 1.05 to 5.1
We can exclude the null hypthesis (*OR* = 1) and we can exclude odds ratios greater than 5 times higher for heavy smokers.

In a decision-theoretic context we can reject the null hypothesis (OR = 1) at a 5% limit on Type I error.
G = 4.814
p = 0.029 from a $\chi 2$ distribution with a single degree of freedom.

10. Analysis of parameters of biological interest.

In his 1951 publication Cornfield argued that the relative risk in the sample could be used to estimate the relative risk in the population, if the disease risk in the population was small. Here are Cornfield's recalculated proportions for lung tumor presence/absence in light and heavy smokers in the population, after applying the disease risk of 155 per million for the same age cohort in the population. Odds and odds ratios, which Cornfield did not use, have been added.

	Lung tumors		Exposure			
	Present	Absent	Odds	OR	Risk	RelRisk
heavy smokers	0.00011935	0.579910	2.06E-04	2.42	1.E-04	3.35
light smokers	0.00003565	0.419935	8.49E-05		4.E-05	
population	0.00015500	0.999845	1.55E-04			
Disease Odds	3.35	1.38				
Disease OR	2.42					
Disease Risk	0.00011935	0.579910				
RelRisk	0.00020581					

Veterans smoke more than non-veterans and thus the sample is not representative of the population with respect to exposure. A correction similar to that for disease risk in the population could also be applied, to improve the inference from the case-control sample to the population.

Binomial response variables –Retrospective Analysis. Multiple categories.

Cornfield considered only two categories, light and heavy smoking.
What is the risk relative to non-smokers?
Does number of cigarettes/day increase risk ?
Here are data from a cross-sectional (case-control) study by Zang and Wynder (1992 *Cancer* 70: 69-76) who report frequency of tumors in 2225 subjects at 5 levels of cigarette smoking. *cf* Sokal and Rohlf 1995, Exercise 17.20

```
                    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.

Here is the result for a classical goodness of fit test. The null hypothesis is the expected proportion: 537 with cancer / 2225 subjects = 0.24.

f	$= \hat{p} \cdot N_i$	$+$	residual	$2\ln L = 2f\ln(f / \hat{p} \cdot N_i)$
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
$G^2 = -2\sum f\ln(f / \hat{p}N_i) =$				434

Here is the ANODEV table for the Zang and Wynder data.

```
                LR Statistics For Type 1 Analysis
                                                   Chi-
        Source          Deviance         DF      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 $\Delta G^2 = 551.2$

Do you get the same result?
Compare the odds for smokers and non-smokers, as an odds ratio.
Calculate the confidence limits on this odds ratio, and interpret.