

BINOMIAL REGRESSION AND GLMS

Håkon K. Gjessing

Professor/Principal Investigator

- Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo
 - a Norwegian Research Council Centre of Excellence
- Department of Global Public Health and Primary Care, University of Bergen

Kathmandu

Friday 2 October 2026

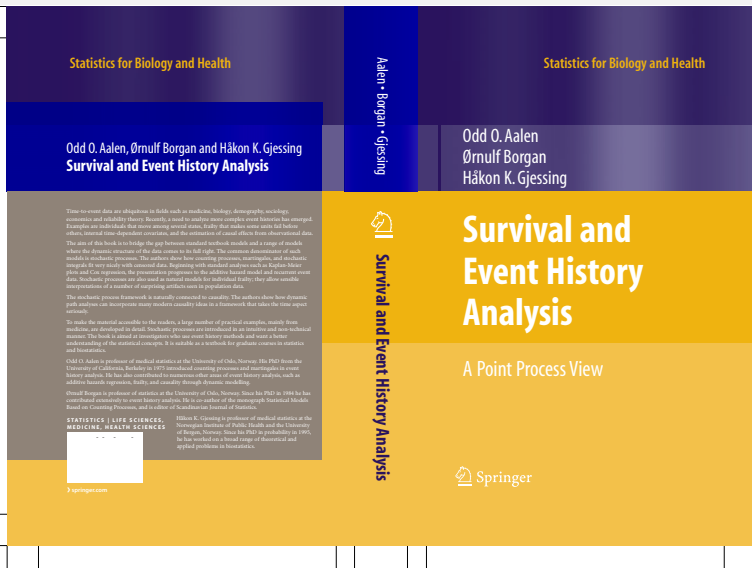
CAUSES OF DEATH IN THREE NORWEGIAN COUNTIES

- Data collection 1974–78
- All men and women aged 35–49 years
- Three rural Norwegian counties:
Oppland, Sogn og Fjordane, and Finmark
- Cardiovascular health screening examination
- More than 90% participation
- Self-report on past and current **smoking habits**
- Mortality in cohort followed-up to the end 2000
- Linked to **cause of death** registry at Statistics Norway
- A subset: 4000 (random) out of about 50000
(<https://www.mn.uio.no/math/english/people/aca/borgan/data/>)
- **Focus: Mortality 40-70 years**

→ Vollset SE, Tverdal A, Gjessing HK. Smoking and deaths between 40 and 70 years of age in women and men. *Annals of internal medicine*. 2006;144(6):381.



TOTALLY INDECENT SELF-PROMOTION



(Includes the example data analyzed using survival analysis)

RISK OF DEATH IN THREE NORWEGIAN COUNTIES

	dead	sex	bmi	smkgr
1	0	0	21.8	1
2	1	0	30.4	1
3	0	0	28.1	1
4	0	0	24.9	2
5	1	0	30.1	4
6	0	0	23.5	1

dead: indicator for death from any cause (0=censored, 1=dead)

sex: sex (0=female, 1=male)

bmi: body mass index at health screening exam

smkgr: smoking group (1=never smoked, 2=former smoker,
3=1-9 cigarettes per day, 4=10-19 cigarettes per day,
5=20+ cigarettes per day, 6=pipe or cigar)

SINGLE PROPORTION

- Risk of dying age 40–70
- $X = 586$ total number of deaths
- $n = 4000$ total number of individuals
- p is the proportion (risk) of dying

$$\hat{p} = \frac{X}{n} = \frac{586}{4000} = 0.1465 = 14.7\%$$

What is the **uncertainty** of this estimate? **Standard Error:**

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.1465(1-0.1465)}{4000}} = 0.00559$$

Standard 95% Confidence Interval (Normal approximation, Wald interval):

$$\hat{p} \pm 1.96 \cdot SE(\hat{p}) = 0.1465 \pm 1.96 \cdot 0.00559 = (0.136, 0.157) = (13.6\%, 15.7\%)$$

Wilson confidence interval:

$$(13.6\%, 15.8\%)$$

Better when $\hat{p}$ is very small (or very large), and when n small.

Suggest always use Wilson.

VARIOUS CAUSES OF DEATH

Cause of death	Risk	95% CI
Overall	14.7%	(13.6%, 15.7%)
Cancer	5.4%	(4.7%, 6.2%)
Cardiovascular + Sudden	6.0%	(5.3%, 6.7%)
Other Medical	1.7%	(1.3%, 2.1%)
Alcohol + Accidents	1.5%	(1.1%, 1.9%)

COMPARE MALES AND FEMALES

- Overall death risk, by sex:

	Males	Females
X (deaths)	398	188
n (total)	2086	1914
$\hat{p}$	19.1%	9.8%
$SE(\hat{p})$	0.00860	0.00680
95% CI	(17.4%, 20.8%)	(8.6%, 11.2%)

COMPARE MALES AND FEMALES

- Non-overlapping confidence intervals means a significant difference (almost always, that is...)
- Warning: overlapping does NOT mean non-significant

BASIC CHI-SQUARED TEST:

	Males	Females	Total
X (deaths)	398	188	586
n - X (non-deaths)	1688	1726	3414
n (total)	2086	1914	4000

Chi-squared = 68.4, df = 1, p-value $< 2.2 \cdot 10^{-16}$
(simulated p-value $< 5.0 \cdot 10^{-4}$)

- BUT: Would also like to **measure** sex difference (not only whether significant/non-significant)

THREE MOST IMPORTANT MEASURES FOR GROUP COMPARISON

- ① Risk Difference (RD)
- ② Relative Risk (RR)
- ③ Odds Ratio (OR)

RISK DIFFERENCE (RD)

	Males	Females
X (deaths)	398	188
n (total)	2086	1914
$\hat{p}$	19.1%	9.8%
$SE(\hat{p})$	0.00860	0.00680

Risk Difference:

$$RD = \hat{p}_M - \hat{p}_F = 19.1\% - 9.8\% = 9.3\%$$

- Uncertainty of RD (i.e. standard error of RD):

$$SE(RD) = \sqrt{SE(\hat{p}_M)^2 + SE(\hat{p}_F)^2} = \sqrt{0.00860^2 + 0.00680^2} = 0.011$$

- 95% Confidence Interval of RD (Wald interval):

$$RD \pm 1.96 \cdot SE(RD) = 0.093 \pm 1.96 \cdot 0.011$$

$$(7.1\%, 11.4\%)$$

RELATIVE RISK (RISK RATIO, RR)

	Males	Females
X (deaths)	398	188
n (total)	2086	1914
$\hat{p}$	19.1%	9.8%
SE($\hat{p}$)	0.00860	0.00680

Relative Risk:

$$RR = \frac{\hat{p}_M}{\hat{p}_F} = \frac{19.1\%}{9.8\%} = 1.94$$

- Uncertainty of **log** RR (i.e. standard error of RR):

$$SE(\log RR) = \sqrt{\left(\frac{SE(\hat{p}_M)}{\hat{p}_M}\right)^2 + \left(\frac{SE(\hat{p}_F)}{\hat{p}_F}\right)^2} = \sqrt{\left(\frac{0.00860}{0.191}\right)^2 + \left(\frac{0.00680}{0.098}\right)^2} = 0.0826$$

- 95% Confidence Interval of log RR:

$$\log RR \pm 1.96 \cdot SE(\log RR) = 0.664 \pm 1.96 \cdot 0.0826 = (0.502, 0.826)$$

- 95% Confidence Interval of RR:

$$(0.502, 0.826) \longrightarrow e^{(0.502, 0.826)} \longrightarrow (1.65, 2.28)$$

ODDS RATIO (OR)

	Males	Females
X (deaths)	398	188
n - X (non-deaths)	1688	1726
$\hat{o}$ (odds)	$\frac{398}{1688}$	$\frac{188}{1726}$
	= 0.236	= 0.109

Odds Ratio:

$$OR = \frac{\hat{o}_M}{\hat{o}_F} = \frac{0.236}{0.109} = 2.16$$

$$OR = \frac{398 \cdot 1726}{1688 \cdot 188} = 2.16$$

- Uncertainty of **log** OR (i.e. standard error of OR, Woolf's formula):

$$SE(\log OR) = \sqrt{\left(\frac{SE(\hat{o}_M)}{\hat{o}_M}\right)^2 + \left(\frac{SE(\hat{o}_F)}{\hat{o}_F}\right)^2} = \sqrt{\frac{1}{398} + \frac{1}{1688} + \frac{1}{188} + \frac{1}{1726}} = 0.0949$$

- 95% Confidence Interval of log OR:

$$\log OR \pm 1.96 \cdot SE(\log OR) = 0.772 \pm 1.96 \cdot 0.0949 = (0.586, 0.958)$$

- 95% Confidence Interval of OR:

$$(0.586, 0.958) \rightarrow e^{(0.586, 0.958)} \rightarrow (1.80, 2.61)$$

THREE MOST IMPORTANT MEASURES FOR GROUP COMPARISON

1 Risk Difference (RD):

- Intuitive (percentage points difference)
- High public health relevance

2 Relative Risk (RR)

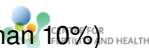
- Intuitive (percentage increase in risk)
- High relevance for risk assessment in research

3 Odds Ratio (OR)

- Percentage increase in odds
- Of high practical utility and relevance in research
- Applicable also in **case-control studies**

NOTE: OR always further away from 1 than RR (whether above or below 1)

NOTE: RR and OR are very close when prevalence is low (say, less than 10%)



GENERALIZED LINEAR MODELS (GLMs)

- A general type of regression model
- All GLMs can use both continuous and categorical covariates, just as in standard multiple regression.

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- The outcome can be *continuous*, *binomial*, or *count* data
- Regression model is determined by **Family** and **Link**

GENERALIZED LINEAR MODELS (GLMs)

GAUSSIAN REGRESSION

- **Family: Gaussian**
- **Link: Identity** (log link also useful)
- Gaussian continuous (like height) outcome
- Includes: Standard multiple linear regression, ANOVA, etc.

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

BINOMIAL REGRESSION

- **Family: Binomial**
- **Link: Identity, log, logit etc.**
- Binomial (dichotomous 0/1, like death) outcome
- Includes: additive risk regression, relative risk regression, logistic regression

$$\text{link}(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

POISSON REGRESSION

- **Family: Poisson**
- **Link: log** (Identity link also useful)
- Count (0, 1, 2, ..., like number of events) outcome
- Includes: additive risk regression, relative risk regression, logistic regression

$$\log(E(Y)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

REGRESSION METHODS (GLMs) FOR BINOMIAL OUTCOMES

Family: Binomial

1 Risk Difference (RD):

- Additive risk regression
- Link: **Identity**

$$p = \beta_0 + \beta_1 \text{sex}$$

2 Relative Risk (RR)

- Multiplicative risk regression
- Link: **log**

$$\log(p) = \beta_0 + \beta_1 \text{sex}$$

3 Odds Ratio (OR)

- Logistic regression (multiplicative odds regression)
- Link: **logit**

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \text{sex}$$

- $\frac{p}{1-p} = \text{odds}$

RELATIONSHIP BETWEEN ODDS AND RISK

ODDS FROM RISK

$$o = \frac{p}{1 - p}$$

RISK FROM ODDS

$$p = \frac{o}{o + 1}$$

Examples

MALES

$$p = \frac{o}{o + 1} = \frac{0.236}{0.236 + 1} = 0.191 = 19.1\%$$

FEMALES

$$p = \frac{o}{o + 1} = \frac{0.109}{0.109 + 1} = 0.098 = 9.8\%$$

LOGISTIC REGRESSION

The odds ratio compares odds for males with odds for females:

$$OR = \frac{o_M}{o_F}$$

so that

$$o_M = o_F \cdot OR$$

$$\log(o_M) = \log(o_F) + \log(OR)$$

- Regression model?

$$\log(\text{odds}) = b_0 + b_1 \cdot \text{sex}$$

sex is a “dummy” variable, 0 for women, 1 for men

$$b_0 = \log(o_F), \quad b_1 = \log(OR)$$

$$\log(\text{odds}) = b_0 + b_1 \text{sex}$$

Find OR associated with `sex`:

$$\text{OR} = \exp(b_1)$$

Advantages over standard analyses of tables:

- Adjust for other covariates
- Can use continuous covariates
- Employ general regression framework

LOGISTIC REGRESSION: EXAMPLE

Using R (www.r-project.org)

```
glm(dead ~ sex, family = binomial, data = d.dead)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.21712	0.07679	-28.871	< 2e-16
sex	0.77227	0.09488	8.139	3.97e-16

which means

$$b_0 = -2.21712, \quad b_1 = 0.77227$$

and

$$OR = \exp(b_1) = \exp(0.77227) = 2.16$$

CI for b_1 can be computed as:

$$\exp(0.77227 \pm 1.96 \cdot 0.09488) = (1.797, 2.607)$$

(compare with earlier result for OR)

```
> exp(confint(res))
```

Slightly more precise, built-in calculation:

	2.5 %	97.5 %
(Intercept)	0.09340824	0.1262432
sex	1.79996858	2.6114660

(Also) very close to earlier result

LOGISTIC REGRESSION: PREDICT INDIVIDUAL RISK

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.21712	0.07679	-28.871	< 2e-16
sex	0.77227	0.09488	8.139	3.97e-16

$$b_0 = -2.21712, \quad b_1 = 0.77227$$

Predicted log odds, for females:

$$Lo = -2.21712 + 0.77227 \cdot 0 = -2.21712$$

Predicted odds, for females:

$$o = \exp(Lo) = \exp(-2.21712) = 0.109$$

Predicted risk, for females:

$$p = \frac{o}{o + 1} = \frac{0.109}{0.109 + 1} = 0.098 = 9.8\%$$

- Similar calculations for males
- More variables can be added on in the same way (using b_0 , b_1 , b_2 etc.)
- The predicted risks will then not exactly match the original risk values

LOGISTIC REGRESSION: EXAMPLE, ADD SMOKING

```
glm(dead ~ sex + factor(smkgrr), family = binomial, data = d.dead)
```

NOTE: Here we have already calculated OR and CI from the $\exp(b)$ formula

	OR	2.5 %	97.5 %
(Intercept)	0.06770644	0.05432186	0.08347914
sex	1.78357836	1.46543998	2.17634462
smkgrr:2	1.46931064	1.11450772	1.94033190
smkgrr:3	2.71202457	1.98231062	3.70071875
smkgrr:4	2.57449499	1.97371130	3.36864305
smkgrr:5	3.21395668	2.29838255	4.48241506
smkgrr:6	4.32231409	2.48632792	7.36277695

smkgrr:

1=never smoked

2=former smoker

3=1-9 cigarettes per day

4=10-19 cigarettes per day

5=20+ cigarettes per day

6=pipe or cigar

LOGISTIC REGRESSION: EXAMPLE, ADD SMOKING

	OR	2.5 %	97.5 %
(Intercept)	0.06770644	0.05432186	0.08347914
sex	1.78357836	1.46543998	2.17634462
smkgr:2	1.46931064	1.11450772	1.94033190
smkgr:3	2.71202457	1.98231062	3.70071875
smkgr:4	2.57449499	1.97371130	3.36864305
smkgr:5	3.21395668	2.29838255	4.48241506
smkgr:6	4.32231409	2.48632792	7.36277695

smkgr:

1=never smoked

2=former smoker

3=1-9 cigarettes per day

4=10-19 cigarettes per day

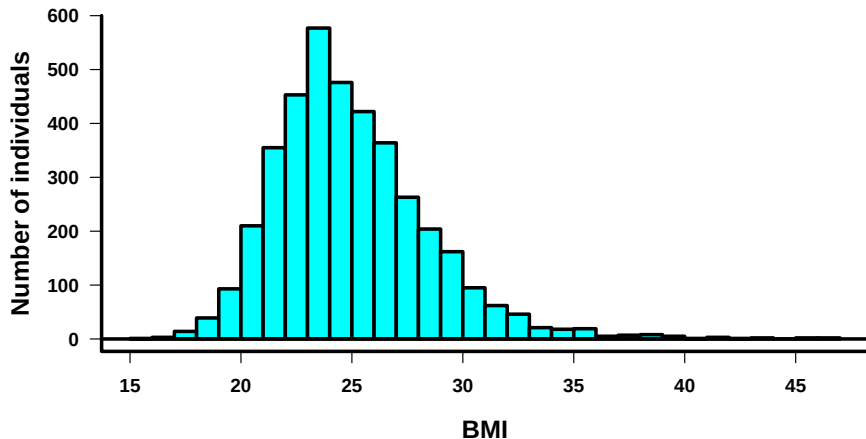
5=20+ cigarettes per day

6=pipe or cigar

Note:

- OR for sex (1.78) is slightly reduced compared to 2.16
... which is likely due to men smoking more than women
- OR for smoking increases for every step up in smoking dose

HOW ABOUT BMI?



How ABOUT BMI?

```
glm(dead ~ sex + factor(smkggr) + bmi, family = binomial, data = d.dead)
```

	OR	2.5 %	97.5 %
(Intercept)	0.01916708	0.009510923	0.03863632
sex	1.81279208	1.484486977	2.21969468
smkggr:2	1.51762626	1.146066451	2.01377857
smkggr:3	2.89946443	2.108662464	3.97822311
smkggr:4	2.70912217	2.064086299	3.56822858
smkggr:5	3.43382238	2.438414134	4.82413364
smkggr:6	4.35540058	2.477296366	7.49219025
bmi	1.04923150	1.022158182	1.07653223

Interpretation bmi OR:

- *One unit* increase in BMI leads to a 1.049 times increase in odds
- E.g., an increase from 20kg/m² to 21kg/m² increases the odds with about 5%
- Or, an increase from 20kg/m² to 30kg/m² gives

$$OR = 1.049^{10} = 1.62$$

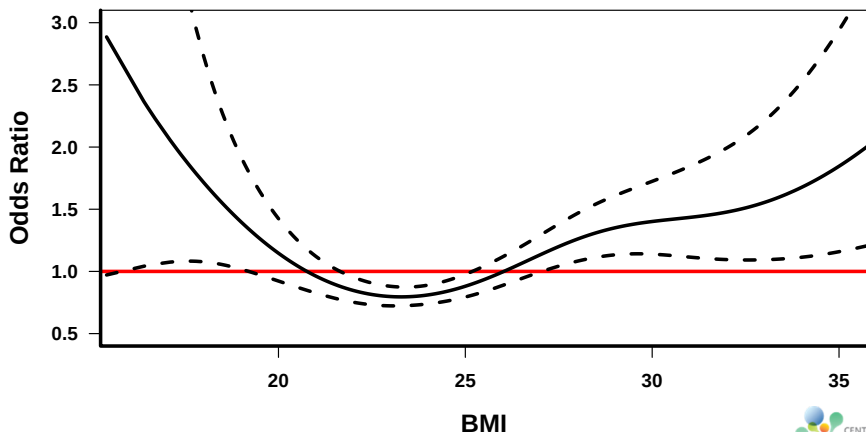
- I.e. an increase of 10 units increases the odds with about 62%

HOW ABOUT BMI?

- Weakness: Assumes linear effect (on log scale) of BMI!
- Possible solutions:
 - Categorize BMI, for instance 0–20, 20–25, 25–30, 30+
 - Or better(?): Extend logistic regression to a `gam` model

HOW ABOUT BMI? GAM MODEL

```
> gam(dead ~ sex + factor(smkg) + s(bmi), family = binomial, data =
```



RELATIVE RISK FROM REGRESSION MODELS

```
> glm(dead ~ sex, family = binomial(link = "log"), data = d)
```

	RR	2.5 %	97.5 %
(Intercept)	0.09822362	0.08542848	0.1120922
sex	1.94246343	1.65517562	2.2890282

Very good match to previous RR calculation

Why does it work?

THE LOGIT LINK FUNCTION FOR LOGISTIC REGRESSION

Recall:

$$\log(\text{odds}) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Link function:

$$\log \frac{p}{1-p}$$

Links p to the covariates $\beta_0 + \beta_1 \cdot x_1 + \cdots$

$$\log(\text{odds}_M) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

$$\log(\text{odds}_F) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Subtract:

$$\log(\text{odds}_M) - \log(\text{odds}_F) = \beta_1$$

$$\text{OR} = \frac{\text{odds}_M}{\text{odds}_F} = \exp(\beta_1)$$

NOTE: Effect of x_1 is assumed independent of other covariates!

FORMULAS REALLY DO MATTER (??)



THE LOG LINK FUNCTION FOR MULTIPLICATIVE RISK REGRESSION

$$\log(p) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Link function:

$$\log(p)$$

Links p to the covariates $\beta_0 + \beta_1 \cdot x_1 + \cdots$

$$\log(p_M) = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

$$\log(p_F) = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Subtract:

$$\log(p_M) - \log(p_F) = \beta_1$$

$$RR = \frac{p_M}{p_F} = \exp(\beta_1)$$

NOTE: Effect of x_1 is assumed independent of other covariates!

THE IDENTITY LINK FUNCTION FOR ADDITIVE RISK REGRESSION

$$p = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Link function:

p

Links p to the covariates $\beta_0 + \beta_1 \cdot x_1 + \cdots$

$$p_M = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

$$p_F = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Subtract:

$$p_M - p_F = \beta_1$$

$$RD = p_M - p_F = \beta_1$$

NOTE: Effect of x_1 is assumed independent of other covariates!

EXAMPLE, ADDITIVE RISK REGRESSION

```
glm(dead ~ sex, family = binomial(link = "identity"), data = d.dead)
```

	RD	2.5 %	97.5 %
(Intercept)	0.09822362	0.08542912	0.1120930
sex	0.09257217	0.07108642	0.1141103

Once more, fits old results very well!

```
glm(dead ~ sex + factor(smkg) + bmi,  
     family = binomial(link = "identity"), data = d.dead)
```

	RD	2.5 %	97.5 %
(Intercept)	-0.046274867	-0.098122776	0.011151717
sex	0.059191784	0.036496825	0.082263035
factor(smkg)2	0.031162994	0.006464203	0.057440655
factor(smkg)3	0.110709769	0.072054540	0.152704055
factor(smkg)4	0.100713913	0.069421898	0.133465102
factor(smkg)5	0.155151870	0.104278229	0.209814228
factor(smkg)6	0.213873564	0.109425349	0.330340778
bmi	0.004244154	0.001947703	0.006499868

- **Additive Risk Regression:**

- Estimates risk differences (RD)
- High public health relevance
- Model not “well defined”:
 - May predict negative risk, or risk above 100%
- This may cause software to crash
- Problematic for very low or very high prevalences

- **Multiplicative Risk Regression:**
 - Estimates relative risk (RR)
 - High relevance for risk assessment in research
 - Model not “well defined”:
 - May predict risk above 100%
 - This may cause software to crash
 - Problematic for high prevalences

ISSUES?? WHAT MODEL TO CHOOSE

● Logistic Regression:

- Estimates odds ratio (OR)
- Of high practical utility and relevance in research
- Applicable also in case-control studies
- Model “well defined”, always a result
- Sometimes (unfairly?) avoided because OR is harder to communicate than RR and RD

NOTE: Sometimes claimed that RR “must” be used instead of OR with high prevalences. This is wrong, and RR does not always work with high prevalences

NOTE: For high prevalences (say, $> 90\%$), a logistic regression is equivalent to a multiplicative regression for the **opposite** outcome

NOTE: If covariate effects are not very strong, all models tend to produce similar conclusions

PREDICT INDIVIDUAL RISK

- First, estimate a model from data:

```
> result <- gam(dead ~ sex + factor(smkg) + s(bmi),  
                 family = binomial, data = d.dead)
```

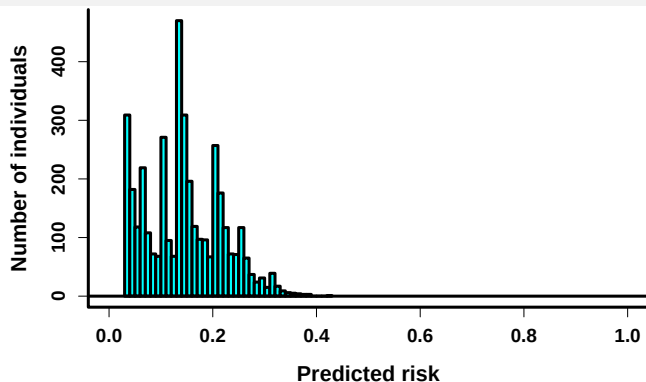
- From model, predict individual risk:

```
> pred <- predict.gam(result, type = "response")
```

Predicted risk values:

dead	sex	bmi	smkg	pred
0	0	21.8	1	0.04052002
1	0	30.4	1	0.10041696
0	0	28.1	1	0.07743467
0	0	24.9	2	0.07115030
1	0	30.1	4	0.19815446
0	0	23.5	1	0.03419502
...	...	...	...	...

DISTRIBUTION OF PREDICTED INDIVIDUAL RISK



In groups, when BMI = 25:

		smkgr					
sex		1	2	3	4	5	6
	0	4.2%	7.2%	15.3%	14.2%	19.2%	25.5%
	1	10.8%	13.8%	21.9%	20.9%	25.8%	32.2%

ROC CURVE

Sensitivity: *Among the dead*, Probability of being positive

Specificity: *Among the survivors*, Probability of being negative

