BINOMIAL REGRESSION

Håkon K. Gjessing

Professor/Principal Investigator

Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo

Department of Global Public Health and Primary Care, University of Bergen

Makerere

Thursday, 8 June 2023

CENTRE FOR FERTILITY AND HEALTH

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
  (http://folk.uio.no/borgan/abg-2008/data/data.html)

- **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.

CENTRE FOR FERTILITY AND HEALTH

Causes of death in three Norwegian counties

```
  agestart agestop dead dead1 dead2 dead3 dead4 sex county sbp  bmi
1    40.00   60.80    0     0     0     0     0   0     14 110 2.18
2    44.43   57.65    1     0     0     1     0   0     14 120 3.04
3    40.00   60.38    0     0     0     0     0   0      5 156 2.81
4    41.11   66.29    0     0     0     0     0   0     14 130 2.49
5    47.06   65.97    1     0     1     0     0   0     14 148 3.01
6    48.51   70.00    0     0     0     0     0   0     14 154 2.35

  smkstart smkgr
1       NA     1
2       NA     1
3       NA     1
4       26     2
5       32     4
6       NA     1
```

agestart: age at health screening exam (or 40 years if screened before that age)
agestop: age in years at death or censoring
dead: indicator for death from any cause (0=censored, 1=dead)
sex: sex (0=female, 1=male)
county: county in Norway (5=Oppland, 14=Sogn og Fjordane, 20=Finmark)
sbp: systolic blood preasure at health screening exam
bmi: body mass index at health screening exam
smkstart: age started smoking
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)

CENTRE FOR FERTILITY AND HEALTH

SINGLE PROPORTION

- Risk of dying age 40–70, all causes (ignoring length of follow-up)
- X = 586 total number of deaths
- n = 4000 total number of individuals
- p is the proportion (risk) of dying

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

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

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

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

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

Wilson confidence interval:

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

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

Suggest always use Wilson.

CENTRE FOR FERTILITY AND HEALTH

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%)

- Overall death risk, by gender:

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

CENTRE FOR FERTILITY AND HEALTH

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** gender difference (not only whether significant/non-significant)

CENTRE FOR FERTILITY AND HEALTH

THREE MOST IMPORTANT MEASURES FOR GROUP COMPARISON

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

CENTRE FOR FERTILITY AND HEALTH

Risk Difference (RD)

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

Risk Difference:

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

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

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

- 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%)

CENTRE FOR FERTILITY AND HEALTH

Relative Risk (Risk Ratio, RR)

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

Relative Risk:

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

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

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

- 95% Confidence Interval of $\log \mathrm{RR}$:

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

- 95% Confidence Interval of RR:

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

CENTRE FOR FERTILITY AND HEALTH

Odds Ratio (OR)

	Males	Females
X (deaths)	398	188
n - X (non-deaths)	1688	1726
$\hat{o}$ (odds)	$\frac{398}{1688}$	$\frac{188}{1726}$
	= 0.195	= 0.099

Odds Ratio:

$$\mathrm{OR} = \frac{\hat{o}_M}{\hat{o}_F} = \frac{0.195}{0.099} = 2.16$$

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

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

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

- 95% Confidence Interval of $\log \mathrm{OR}$:

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

- 95% Confidence Interval of OR:

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

CENTRE FOR FERTILITY AND HEALTH

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**

CENTRE FOR FERTILITY AND HEALTH

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.

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

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

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

Poisson Regression

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

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

CENTRE FOR FERTILITY AND HEALTH

Family: Binomial

1. Risk Difference (RD):
   - Additive risk regression
   - Link: Identity

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

2. Relative Risk (RR)
   - Multiplicative risk regression
   - Link: log

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

3. Odds Ratio (OR)
   - Logistic regression (multiplicative odds regression)
   - Link: logit

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

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

CENTRE FOR FERTILITY AND HEALTH

Relationship between odds and risk:

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

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

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

so that

$$\begin{aligned} o_M &= o_F \cdot OR \\ \log(o_M) &= \log(o_F) + \log(OR) \end{aligned}$$

- Regression model?

$$\boxed{\log(\text{odds}) = \beta_0 + \beta_1 \cdot \text{sex}}$$

`sex` is a "dummy" variable, 0 for women, 1 for men

$$\beta_0 = \log(o_F), \quad \beta_1 = \log(OR)$$

CENTRE FOR FERTILITY AND HEALTH

LOGISTIC REGRESSION

$$\boxed{\log(\text{odds}) = \beta_0 + \beta_1 \texttt{sex}}$$

Find OR associated with `sex`:

$$\text{OR} = \exp(\beta_1)$$

Advantages over standard analyses of tables:

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

CENTRE FOR FERTILITY AND HEALTH

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
```

As before, CI can be computed (manually) as:

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

(compare earlier result for OR)

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

CENTRE FOR FERTILITY AND HEALTH

LOGISTIC REGRESSION: EXAMPLE, ADD SMOKING

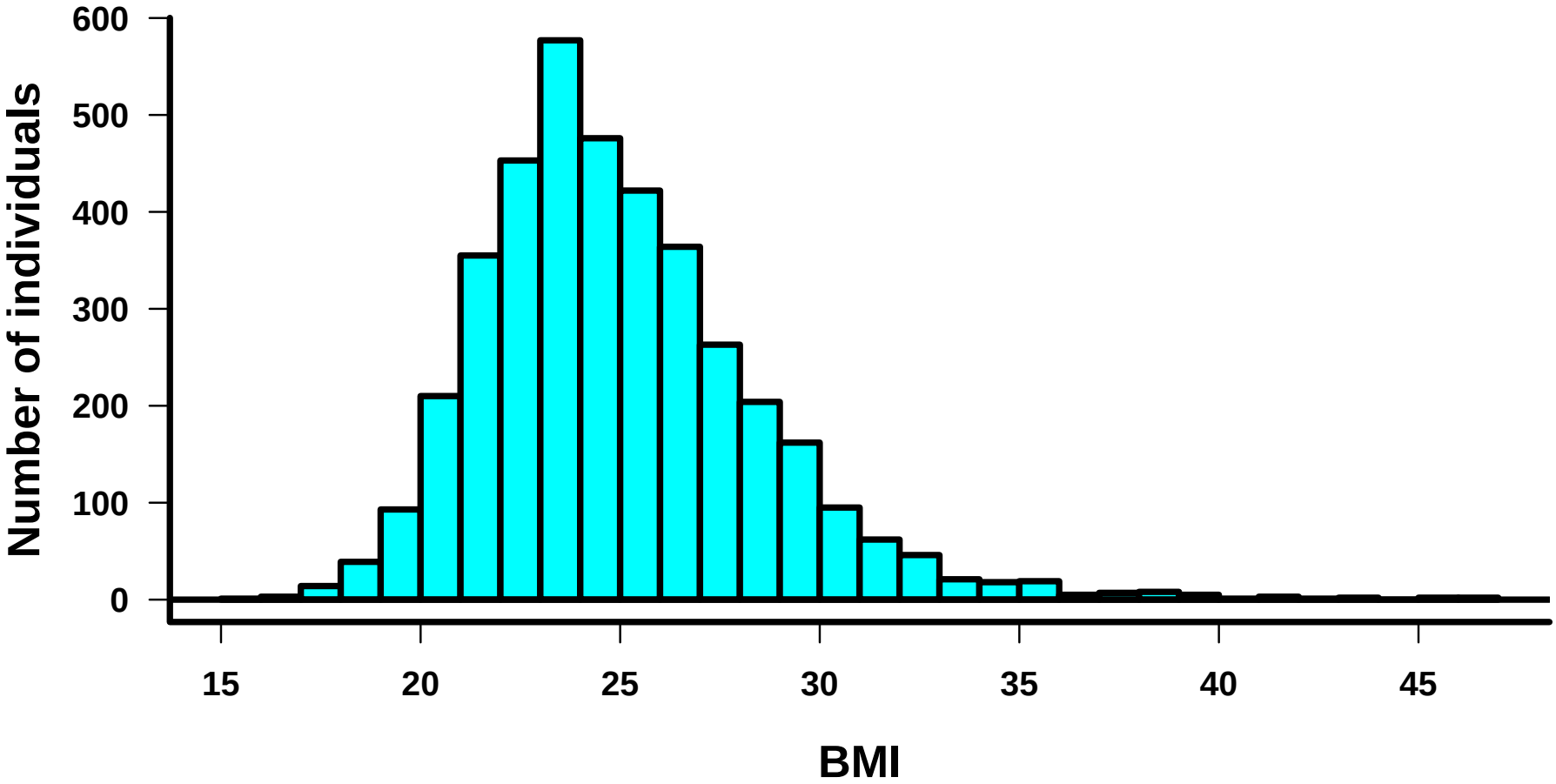
```
glm(dead ~ sex + factor(smkgr), family = binomial, data = d.dead)
```

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

CENTRE FOR FERTILITY AND HEALTH

How about BMI?



CENTRE FOR FERTILITY AND HEALTH

How about BMI?

```
> glm(dead ~ sex + factor(smkgr) + 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
factor(smkgr)2	1.51762626	1.146066451	2.01377857
factor(smkgr)3	2.89946443	2.108662464	3.97822311
factor(smkgr)4	2.70912217	2.064086299	3.56822858
factor(smkgr)5	3.43382238	2.438414134	4.82413364
factor(smkgr)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 20$\mathrm{kg/m^2}$ to 21$\mathrm{kg/m^2}$ increases the odds with about 5%
- Or, an increase from 20$\mathrm{kg/m^2}$ to 30$\mathrm{kg/m^2}$ gives

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

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

CENTRE FOR FERTILITY AND HEALTH

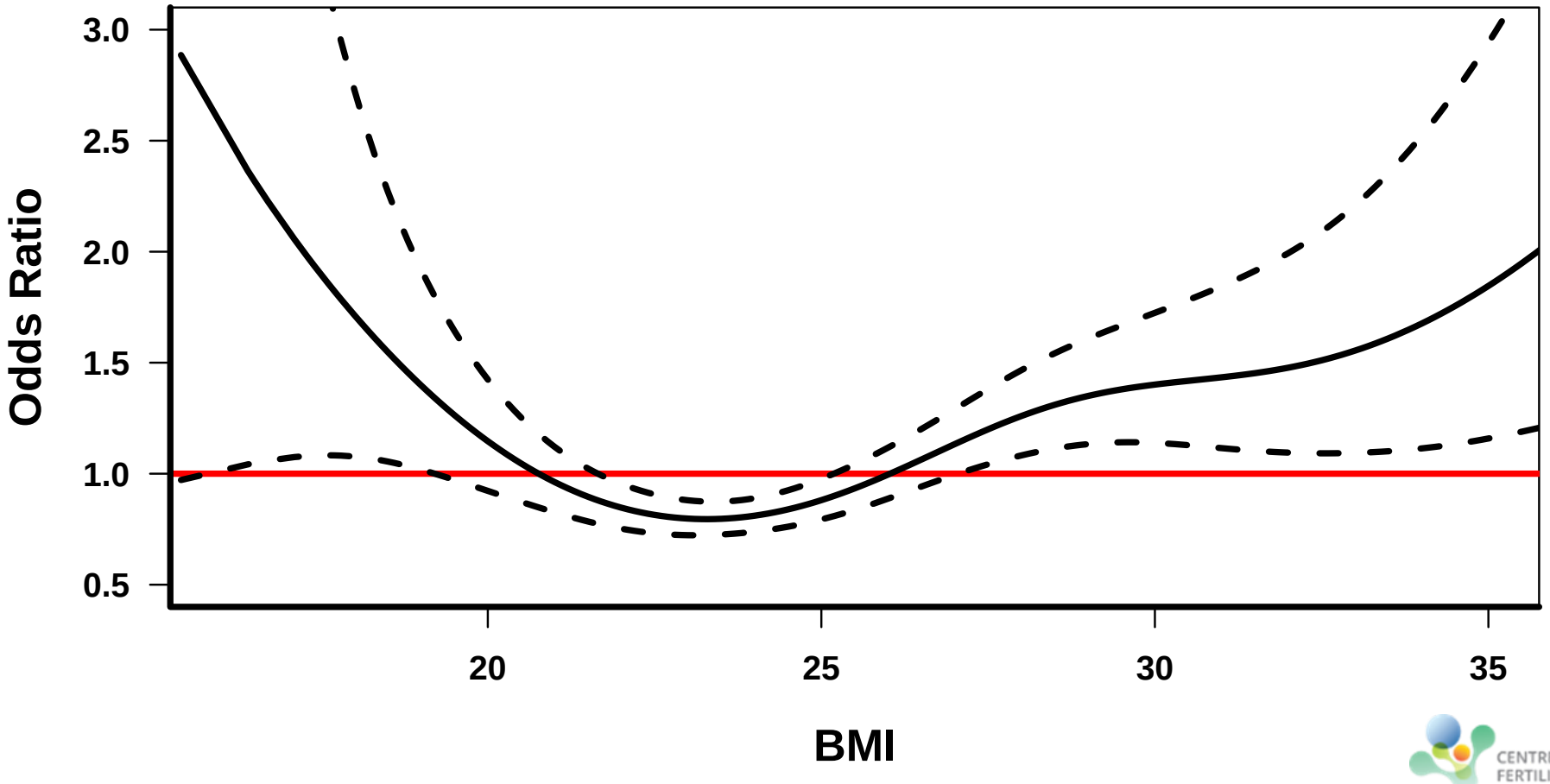
How about BMI?

- 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

CENTRE FOR FERTILITY AND HEALTH

How about BMI?

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



CENTRE FOR FERTILITY AND HEALTH

Relative Risk from regression models

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

```
                    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?

CENTRE FOR FERTILITY AND HEALTH

The Logit Link Function for Logistic Regression

Recall:

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

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

Link function:

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

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

$$\begin{aligned} \log(\text{odds}_M) &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \\ \log(\text{odds}_F) &= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \end{aligned}$$

Subtract:

$$\begin{aligned} \log(\text{odds}_M) - \log(\text{odds}_F) &= \beta_1 \\ \text{OR} = \frac{\text{odds}_M}{\text{odds}_F} &= \exp(\beta_1) \end{aligned}$$

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

CENTRE FOR FERTILITY AND HEALTH

FORMULAS REALLY DO MATTER (??)

The Log Link Function for Multiplicative Risk Regression

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

Link function:

$$\log(p)$$

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

$$\begin{aligned} \log(p_M) &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \\ \log(p_F) &= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \end{aligned}$$

Subtract:

$$\begin{aligned} \log(p_M) - \log(p_F) &= \beta_1 \\ RR = \frac{p_M}{p_F} &= \exp(\beta_1) \end{aligned}$$

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

CENTRE FOR FERTILITY AND HEALTH

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

Link function:

$$p$$

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

$$\begin{aligned} p_M &= \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \\ p_F &= \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k \end{aligned}$$

Subtract:

$$\begin{aligned} p_M - p_F &= \beta_1 \\ RD = p_M - p_F &= \beta_1 \end{aligned}$$

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

CENTRE FOR FERTILITY AND HEALTH

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(smkgr) + 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(smkgr)2  0.031162994  0.006464203 0.057440655
factor(smkgr)3  0.110709769  0.072054540 0.152704055
factor(smkgr)4  0.100713913  0.069421898 0.133465102
factor(smkgr)5  0.155151870  0.104278229 0.209814228
factor(smkgr)6  0.213873564  0.109425349 0.330340778
bmi             0.004244154  0.001947703 0.006499868
```

CENTRE FOR FERTILITY AND HEALTH

Issues?? What model to choose

- **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

CENTRE FOR FERTILITY AND HEALTH

Issues?? What model to choose

- **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

CENTRE FOR FERTILITY AND HEALTH

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

CENTRE FOR FERTILITY AND HEALTH

PREDICT INDIVIDUAL RISK

- First, estimate a model from data:

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

- From model, predict individual risk:

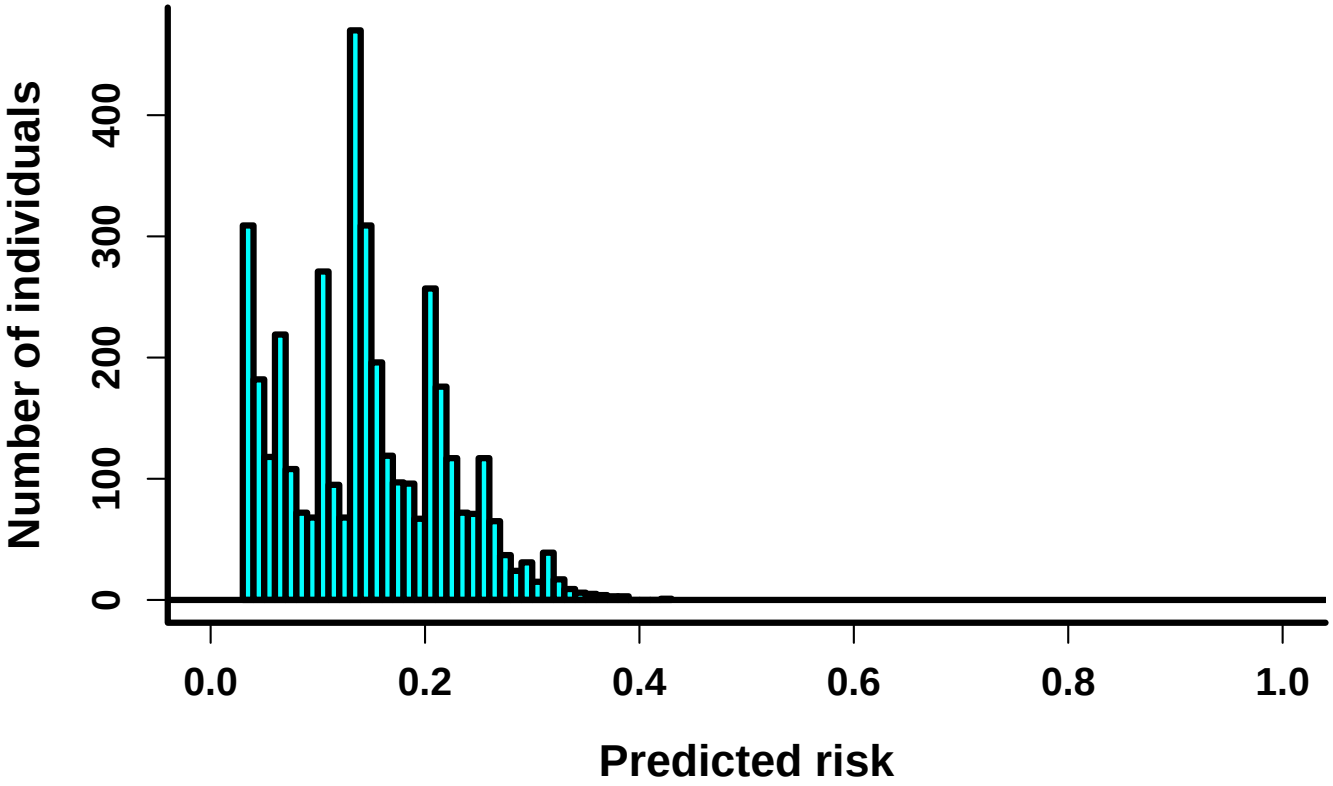
```
> pred <- predict.gam(result, type = "response")
```

Predicted risk values:

dead	sex	bmi	smkgr	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
...	...	...	...	...

CENTRE FOR FERTILITY AND HEALTH

DISTRIBUTION OF PREDICTED INDIVIDUAL RISK



In groups, when BMI = 25:

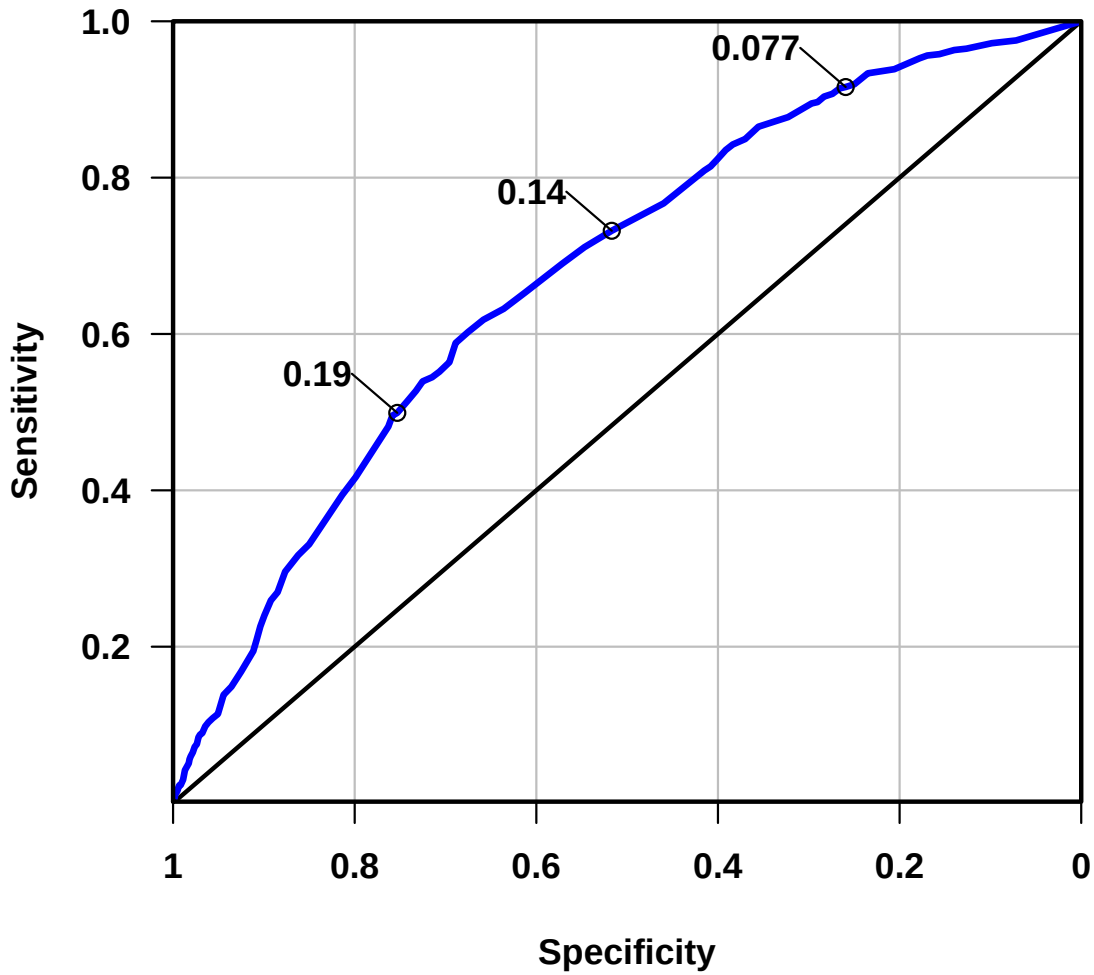
	smkgr					
sex	1	2	3	4	5	6
0	4.2%	7.2%	15.3%	14.2%	19.2%	25.5%
1	10.8%	13.8%	21.9%	20.9%	25.8%	32.2%

CENTRE FOR FERTILITY AND HEALTH

ROC CURVE

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

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



CENTRE FOR FERTILITY AND HEALTH