Lecture 7 – Program

1. Data structure and basic questions

2. The logistic regression model

3. Odds and odds ratio

4. Maximum likelihood estimation

5. Deviance and sum of squares

6. Multiple logistic regression

Data structure and basic questions

As before the data have the form:

unit	response	covariates
1	y_1	$x_{11} \cdots x_{1p}$
2	y_2	$x_{21} \cdots x_{2p}$
.	.	$\cdots$
.	.	$\cdots$
.	.	$\cdots$
n	y_n	$x_{n1} \cdots x_{np}$

But the response is no longer measured on a quantitative scale.

The response is either

- taking the values 0 or 1, indicating a certain characteristic of a subject

or

- a proportion in a group where all subjects have the same values of *all* the covariates

Objective is as before: Explain the variation in the response y by variation in the covariates $x_1, \cdots, x_p$

Example: Coronary heart disease vs. age

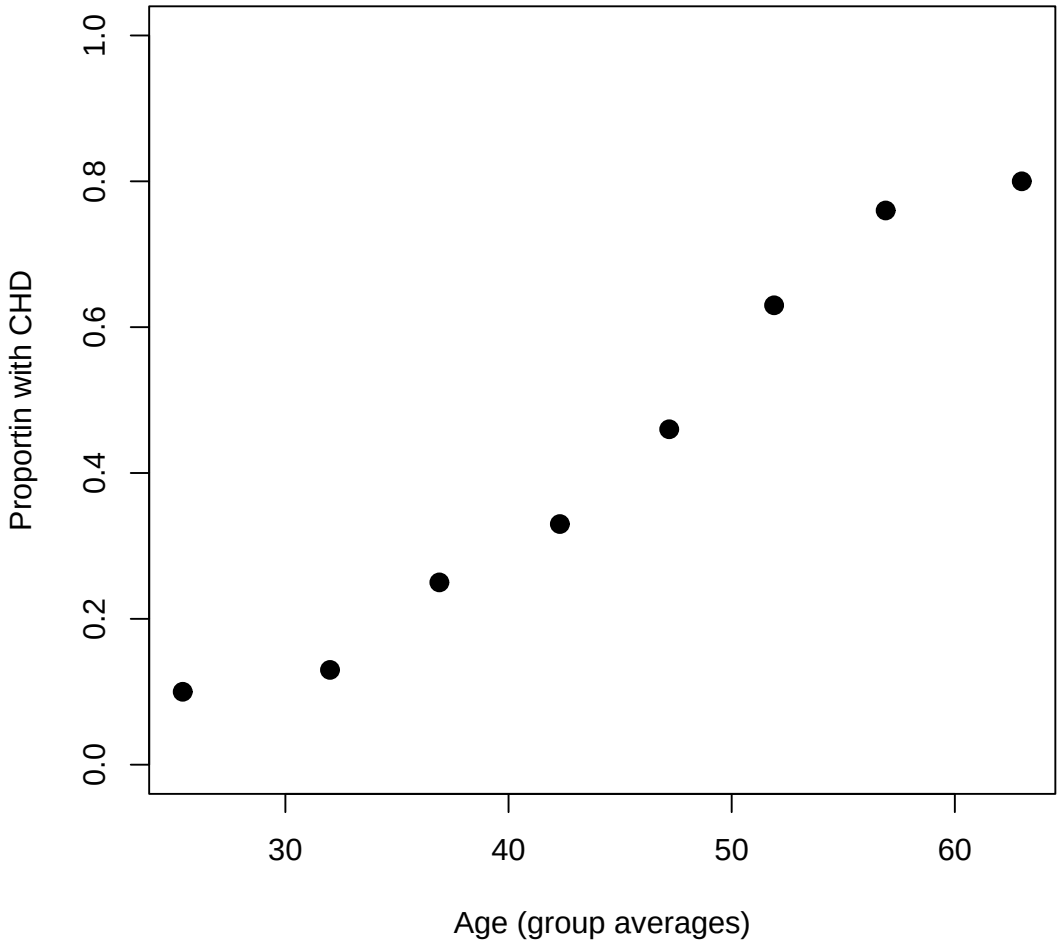
Either individual data

ind	chd	age
1	0	20
2	0	23
3	0	24
4	0	25
5	1	25
.	.	...
.	.	...
.	.	...
97	0	64
98	1	64
99	1	65
100	1	69

or grouped data

age	number	proportion with CHD
20-29	10	0.10
30-34	15	0.13
35-39	12	0.25
40-44	15	0.33
45-49	13	0.46
50-54	8	0.63
55-59	17	0.76
60-69	10	0.80

Proportion with coronary heart disease vs. age (based on the grouped data)



Why not linear regression?

- Have a S-shaped plot, so only approximately linear in the middle

- When the responses are recorded as binary data (0 or 1), they are far from being normally distributed. (We will have approximate normality for grouped data.)

- For a binary y we have

$$P(y = 1) = 1 - P(y = 0) = p$$

Then $\mathrm{E}(y) = p$ and $\mathrm{Var}(y) = p(1 - p)$. Thus the variance is not constant, but a function of the mean.

Linking response and covariate

We start out with the situation where there is only one covariate

Let y be the observed proportion for a unit, and let m be the number of individuals in the unit. (For individual data we have $m = 1$, and y can only take the values 0 or 1.)

Let the unit have covariate value x.

Considered as a random variable we know that

$$\mathrm{E}(y) = p = p(x)$$

is varying with x.

We want to model how $p(x)$ depends on x.

Linking response and covariate, contd.

One option is to consider a linear model where the dependency is described by

$$p(x) = \beta_0 + \beta_1 \, x$$

This will *not* capture:

- the S-shaped curve typically observed
- the fact that the expression $\beta_0 + \beta_1 x$ can take on all values in $(-\infty \, , \, \infty)$, while the probability $p(x)$ should stay within the interval $(0, 1)$

Linking response and covariate contd.

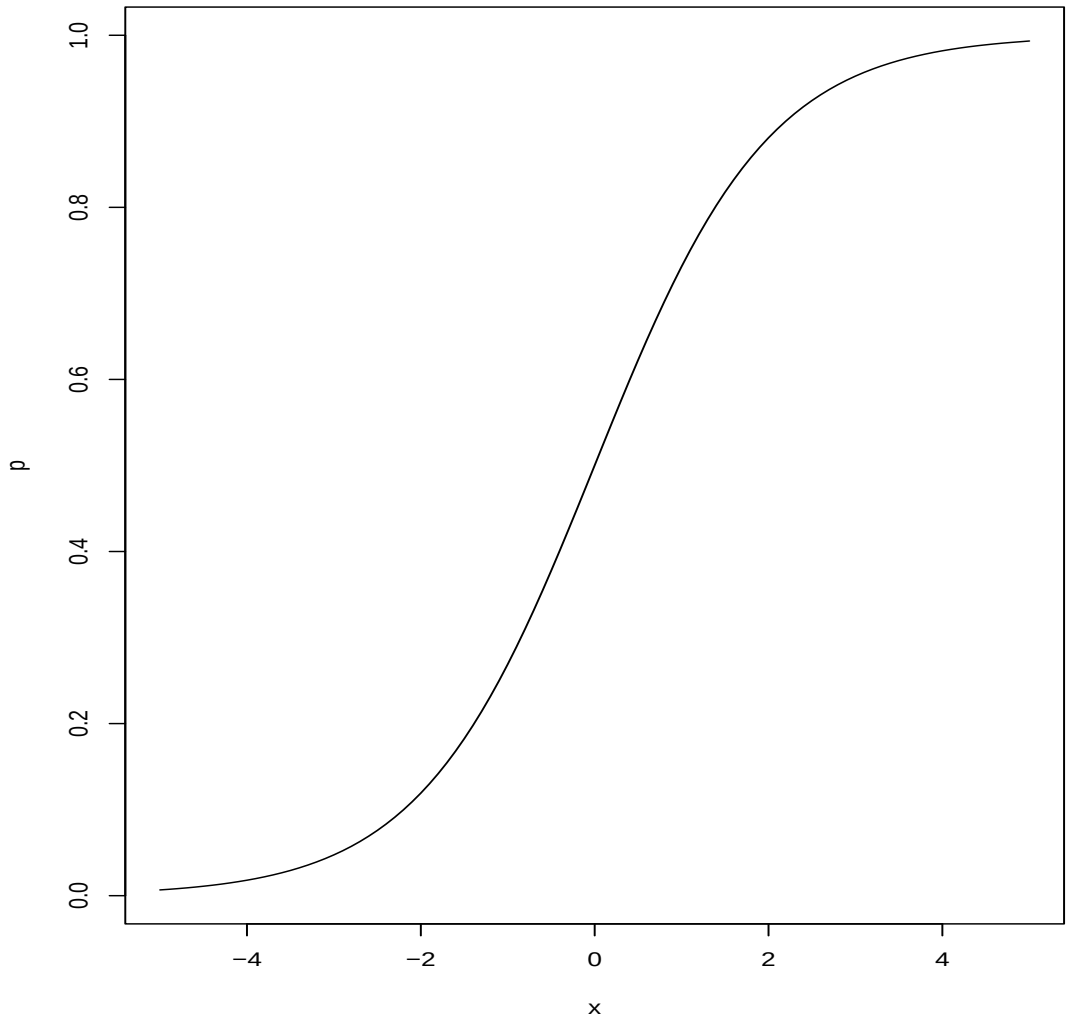
The most common solution is to link the expectation of the response $p(x)$ and the linear predictor $\eta(x) = \beta_0 + \beta_1 x$ by the relation

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \eta(x) = \beta_0 + \beta_1 x$$

or equivalently by the relation

$$p(x) = \frac{e^{\eta(x)}}{1+e^{\eta(x)}} = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

The logistic function $\frac{e^{\eta(x)}}{1+e^{\eta(x)}}$ when $\eta(x) = x$



Alternatively we could have used another strictly increasing continuous function defined on $(-\infty\,,\,\infty)$ with values in $(0, 1)$.

Odds and odds ratio

The quantity $\frac{p(x)}{1-p(x)}$ is called the **odds**.

For the logistic model the odds becomes:

$$\frac{p(x)}{1-p(x)} = e^{\beta_0+\beta_1 x}$$

The **odds ratio** between two individuals with covariates x' and x is

$$\text{OR} = \frac{p(x')/(1-p(x'))}{p(x)/(1-p(x))} = e^{\beta_1(x'-x)}$$

which shows that as in linear regression case, the coefficient β_1 measure the influence of covariates changes. In logistic regression the change is in the odds ratio, not in the mean or expectation of the response.

Maximum likelihood estimation

Estimation in the logistic model is performed using maximum likelihood estimation

We first describe maximum likelihood estimation for the linear regression model:

- $y_i \sim$ N(μ_i, σ^2) with $\mu_i = \beta_0 + \beta_1 x_i$
- the y_i are independent

For ease of presentation, we assume that σ^2 is known

The density of y_i takes the form:

$$f(y_i, \mu_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_- \mu_i)^2\}$$

Maximum likelihood estimation contd.

The likelihood is the simultaneous density

$$L = \prod_{i=1}^{n} f(y_i, \mu_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2\}$$

considered as a *function of the parameters* β_0 and β_1 for the observed values of the y_i

We estimate the parameters by *maximizing the likelihood*. This corresponds to finding the parameters that make the observed y_i as likely as possible

Maximizing L is the same as maximizing

$$\log L = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2,$$

which is the same as minimizing $\sum_{i=1}^{n}(y_i - \mu_i)^2$.

For the linear regression model, maximum likelihood estimation coincides with least squares estimation

Maximum likelihood for logistic regression

Data: (y_i, m_i, x_i); for $i = 1, 2, \ldots, n$

We assume that the frequencies $f_i = m_i y_i$ are binomially distributed, i.e.

$$P(f_i = f) = \binom{m_i}{f} p(x_i)^f \{1 - p(x_i)\}^{m_i - f}$$

In other words, for each covariate value x_i we have m_i subjects, and the number of subjects having the attribute or characteristic in question are $f_i \sim$ bin$(m_i, p(x_i))$

We further assume that the frequencies $f_1, f_2, \ldots, f_n$ are independent

Max. likelihood for logistic regression, contd.

The likelihood becomes

$$L = \prod_{i=1}^{n} \binom{m_i}{f_i} p(x_i)^{f_i} \{1 - p(x_i)\}^{m_i - f_i}.$$

Since

$$p(x_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

the likelihood is, for given observations, a function of the unknown parameters β_0 and β_1, i.e. $L = L(\beta_0, \beta_1)$.

We estimate β_0 and β_1 by the values of these parameters that maximizes the likelihood.

These estimates are called the **maximum likelihood estimates** (MLE) and are denoted $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Example, grouped CHD data

```
avage<-c(25.4, 32, 36.9, 42.3, 47.2, 51.9, 56.9, 63)
totno<-c(10,15, 12, 15, 13, 8, 17, 10)
nochd<-c(1, 2, 3, 5, 6, 5, 13, 8)

nochd/totno
0.10000 0.13333 0.25000 0.33333
0.46153 0.62500 0.76471 0.80000

mod<-glm(cbind(nochd,totno-nochd)~avage, family=binomial)
summary(mod)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.19938    1.11650  -4.657 3.21e-06 ***
avage        0.10857    0.02371   4.580 4.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.7015  on 7  degrees of freedom
Residual deviance:  0.3903  on 6  degrees of freedom
AIC: 25.527

Number of Fisher Scoring iterations: 4
```

Wald test for $H_0 : \beta_1 = 0$

$\widehat{\beta}_1$ is approximately $N(\beta_1, se_1^2)$-distributed, where $\widehat{se}_1$ is computed by a statistical package.

Hence, $H_0 : \beta_1 = 0$ is rejected at the 5% level by the so-called **Wald test** if

$$\left| \frac{\widehat{\beta}_1}{\widehat{se}_1} \right| > 1.96.$$

The P-value for the test is given in the computer output (based on the normal approximation)

Example, grouped CHD data

Here $\widehat{\beta}_1 = 0.1086$ and $\widehat{se}_1 = 0.0237$, and the Wald test statistic becomes $0.1086/0.0237 = 4.58$, which is highly significant

Confidence intervals for β_1 and OR

An 95% confidence interval for β_1 (based on the normal approximation) is given by

$$\widehat{\beta}_1 \pm 1.96 \times \widehat{se}_1$$

OR= exp(β_1) is the odds ratio of one unit's increase in x

We obtain a 95% confidence interval for OR by transforming the lower and upper limits of the confidence interval for β_1

Example, grouped CHD data

Here $\widehat{\beta}_1 = 0.1086$ and $\widehat{se}_1 = 0.0237$.
A 95% confidence interval for β_1 has limits $0.1086 \pm 1.96 \times 0.0237$, i.e. from 0.0621 to 0.1550.

An estimate of the odds ratio OR= e^{β_1} is $\widehat{\text{OR}} = e^{0.1086} = 1.115$ with 95% confidence limits from $e^{0.0621} = 1.064$ to $e^{0.0621} = 1.168$

Deviance and sum of squares

For linear regression the sum of squares was a key quantity in connection with testing and for assessing the fit of a model.

We want to define a quantity for logistic regression that corresponds to the sum of squares.

To this end we start out by considering the relation between the sum of squares and the log-likelihood for the linear regression model.

Deviance and sum of squares, contd.

For the linear regression model the log-likelihood $l = \log L$ takes the form:

$$l = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$$

It obtains its *largest* value for the *saturated model*, i.e. the model where there are no restrictions on the μ_i. Then the μ_i are estimated by $\tilde{\mu}_i = y_i$, and the log-likelihood becomes:

$$\tilde{l} = -\frac{n}{2}\log(2\pi\sigma^2)$$

It follows that the *deviance*

$$D = 2\left(\tilde{l} - l\right) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2$$

is the sum of squares divided by σ^2.

Deviance for binomial data

We then consider data of the form

$$(y_i, m_i, x_i) \qquad i = 1, \ldots n$$

where the frequencies $f_i = m_i y_i$ are independent and binomially distributed $f_i \sim$ bin(m_i, p_i).

The log-likelihood $l = l(p_1, \ldots, p_n)$ is a function of $p_1, \ldots, p_n$

For the *saturated model*, i.e. the model where there are no restrictions on the p_i, the p_i are estimated by the observed proportions

$$\tilde{p}_i = y_i$$

and the log-likelihood takes the value

$$\tilde{l} = l(\tilde{p}_1, \ldots, \tilde{p}_n)$$

Deviance for binomial data, contd.

For a fitted logistic regression model we obtain the estimated probabilities

$$\widehat{p}_i = \widehat{p}(x_i) = \frac{e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}{1 + e^{\widehat{\beta}_0 + \widehat{\beta}_1 x_i}}$$

and the corresponding value

$$\widehat{l} = l(\widehat{p}_1, \ldots, \widehat{p}_n)$$

of the log-likelihood.

The **deviance** for the model is defined as

$$\hat{D} = 2\left(\tilde{l} - \widehat{l}\right)$$

The deviance measures the fit of the model.

If *all* m_i are large, $\hat{D}$ is approximately χ^2_{n-2} distributed when the logistic model holds true.

This can be used as a **goodness-of-fit** test.

Example, grouped CHD data

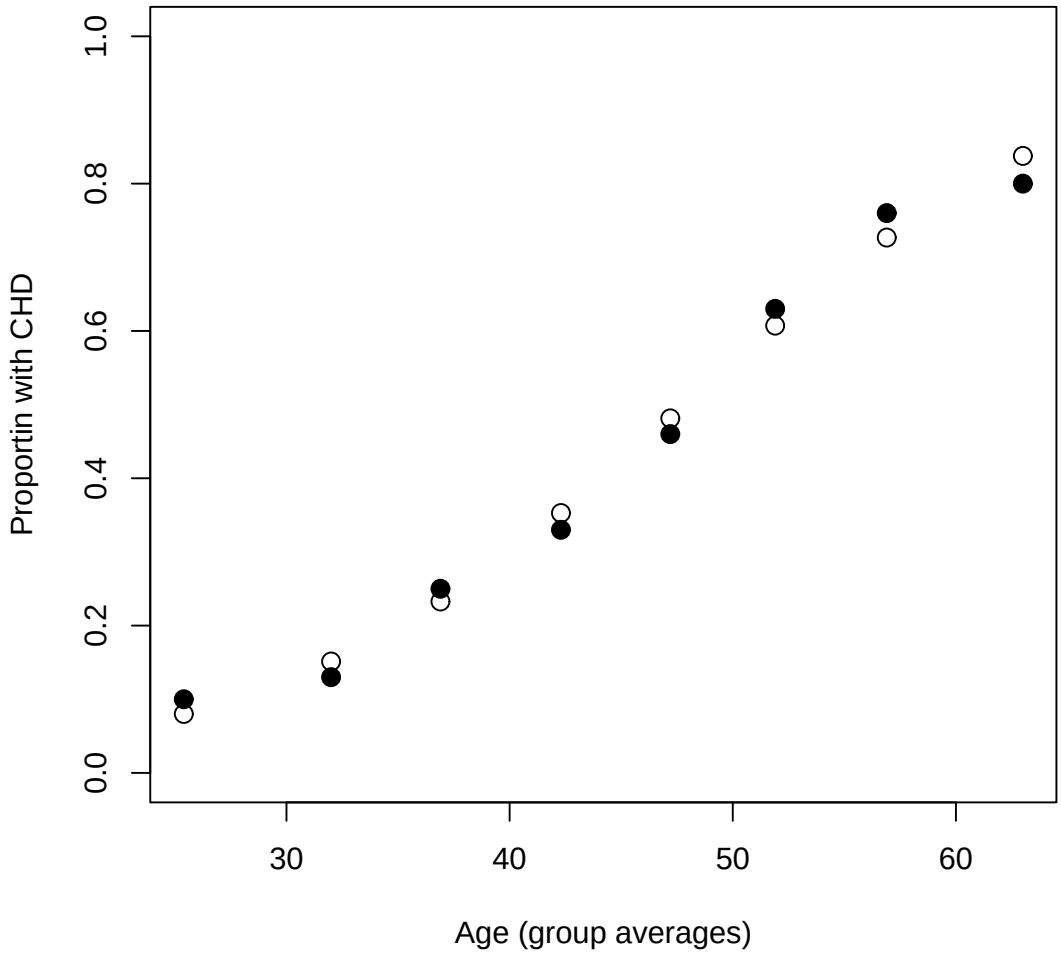
```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.19938    1.11650  -4.657 3.21e-06 ***
avage        0.10857    0.02371   4.580 4.65e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 28.7015  on 7  degrees of freedom
Residual deviance:  0.3903  on 6  degrees of freedom
AIC: 25.527
```

The deviance for the fitted logistic model ("residual deviance") is 0.39.

This indicates a good fit of the model.

Example, grouped CHD data

Observed (●) and fitted proportions (○)



Deviance for binomial data, contd.

We have seen how the deviance can be used to check if the logistic model

$$p(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}$$

gives a good fit to the data (when compared with the saturated model).

We can also use the deviance to test the null hypothesis

$$\mathsf{H}_0 : \beta_1 = 0$$

i.e. if the logistic model can be simplified.
(This gives an alternative to the Wald test.)

We then compare the deviance of the logistic model with the deviance of the model under H_0.

Deviance for binomial data, contd.

Under H_0 all the p_i are assumed to be equal. The common value of the p_i is estimated by

$$p_i^\star = \frac{f_1 + \cdots + f_n}{m_1 + \cdots + m_n}$$

The corresponding value of the log-likelihood is $l^* = l(p_1^*, \ldots, p_n^*)$ and the deviance becomes

$$D^* = 2(\tilde{l} - l^*)$$

If H_0 is true, the difference in deviance between the two models

$$G = D^* - \hat{D} = 2(\hat{l} - l^*)$$

is approximately χ_1^2 distributed.

We reject H_0 if G is large compared to the percentiles in the χ_1^2-distribution.

Example, grouped CHD data

```
mod0<-glm(cbind(nochd,totno-nochd)~1, family=binomial)

mod1<-glm(cbind(nochd,totno-nochd)~avage, family=binomial)

mod2<-glm(cbind(nochd,totno-nochd)~factor(avage),
                                        family=binomial)

anova(mod0,mod1,mod3,test="Chisq")

Analysis of Deviance Table

Model 1: cbind(nochd, totno - nochd) ~ 1
Model 2: cbind(nochd, totno- nochd) ~ avage
Model 3: cbind(nochd, totno - nochd) ~factor(avage)

  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         7    28.7015
2         6     0.3903  1  28.3112 1.033e-07
3         0  1.776e-15  6   0.3903    0.9989
```

The approach based on comparing deviances may be used in more complex situations than the one considered here, very much in the same way as when comparing sums of squares in multippel linear regression and ANOVA models

Multiple logistic regression

Several covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)$

Model: $p(\mathbf{x}) = \dfrac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$

or

$$\log\left(\frac{p(\mathbf{x})}{1 - p(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

With $\mathbf{x}' = (x_1 + 1, x_2, \ldots, x_p)$ one has

$$\text{OR}_1 = \exp(\beta_1) = \frac{p(\mathbf{x}')/(1 - p(\mathbf{x}'))}{p(\mathbf{x})/(1 - p(\mathbf{x}))} = \frac{\text{Odds}(\mathbf{x}')}{\text{Odds}(\mathbf{x})}$$

Thus the $\exp(\beta_1)$ can be interpreted as odds ratio when x_1 is increased with 1 unit while the other covariates remain the same. (Similarly for the other regression coefficients.)

Tests and confidence intervals

$\widehat{\beta}_j$ = MLE for β_j

$\widehat{se}_j$ = estimated standard error for $\widehat{\beta}_j$

Wald test statistic for $\mathsf{H}_{0j} : \beta_j = 0$

$$Z_j = \frac{\widehat{\beta}_j}{\widehat{se}_j} \sim \mathsf{N}(0,1) \text{ under } \mathsf{H}_{0j}$$

95% confidence interval for β_j

$$\widehat{\beta}_j \pm 1.96 \times \widehat{se}_j$$

We obtain a 95% confidence interval for $\mathsf{OR}_j = \exp(\beta_j)$ by transforming the lower and upper limits of the confidence interval for β_j

Tests based on the deviance

Consider the logistic regression model

$$\text{Model: } p(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}$$

We will test the null hypothesis H_0 that q of the β_js are equal to zero. (Equivalently that there are q linear restrictions among the β_js.)

We let:

- $\widehat{D}$ be the deviance and $\widehat{l}$ the log-likelihood under the full logistic regression model
- D^* be the deviance and l^* be the log-likelihood under the null hypothesis H_0

Then

$$G = D^* - \widehat{D} = 2(\widehat{l} - l^*)$$

is approximately χ^2_q distributed if H_0 is true. We reject H_0 if G is large compared to the percentiles in the χ^2_q-distribution.

Example, insulin injections

Fam. hist		Yes		No	
Dep. on injec.		Yes	No	Yes	No
Age at onset	< 45	6	1	16	2
	≥ 45	6	36	8	48

Read the data, and fit a sequence of models

```
dep<-c(6,6,16,8)
no<-c(7,42,18,56)
age<-c("lt45","ge45","lt45","ge45")
fam<-c("Yes","Yes","No","No")

mod0<-glm(cbind(dep,no-dep)~1,family=binomial)
mod1<-glm(cbind(dep,no-dep)~factor(age), ....)
mod2<-glm(cbind(dep,no-dep)~factor(age)+factor(fam), ...)
mods<-glm(cbind(dep,no-dep)~factor(age)*factor(fam), ...)
```

Example, insulin injections, contd.

Compare the models:

```
anova(mod0,mod1,mod2,mod3,test="Chisq")
```

```
Analysis of Deviance Table

Model 1: cbind(dep,no-dep)~1
Model 2: cbind(dep,no-dep)~factor(age)
Model 3: cbind(dep,no-dep)~factor(age)+factor(fam)
Model 4: cbind(dep,no-dep)~factor(age)*factor(fam)


  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         3     50.034
2         2      0.047  1   49.987 1.548e-12
3         1      0.039  1    0.007     0.932
4         0  2.665e-15  1    0.039     0.843
```

Only age seems to have an effect.

Example, insulin injections, contd.

Summarize the results for the model with age

```
summary(mod1)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.7918     0.2887  -6.207 5.41e-10 ***
factor(age)lt45   3.7842     0.6798   5.567 2.60e-08 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 50.033593  on 3  degrees of freedom
Residual deviance:  0.046666  on 2  degrees of freedom
AIC: 15.682

Number of Fisher Scoring iterations: 4
```

Interpretation

Let p_{ij} be the probability of injection dependency when factor "age" has level i (2 : $<$ 45 1 : $\geq$ 45) and factor "fam" has level j (2: Yes; 1: No). The full (and saturated) model is

$$p_{ij} = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)}$$

where

$$x_1 = \begin{cases} 0 & \text{if } i = 1 \\ 1 & \text{if } i = 2 \end{cases}, \; x_2 = \begin{cases} 0 & \text{if } j = 1 \\ 1 & \text{if } j = 2 \end{cases}$$

and

$$x_3 = \begin{cases} 0 & \text{if } (i,j) \neq (2,2) \\ 1 & \text{if } (i,j) = (2,2) \end{cases}$$

Interpretation, contd.

In the model without interaction and main effect for "fam", the odds for being dependent on insulin injections is therefore

$$\frac{p_{ij}}{1-p_{ij}} = \exp(\beta_0 + \beta_1 x_1) \qquad \text{for} \quad j = 1, 2.$$

which does not depend on j (i.e. on "fam").

Hence the odds-ratios for age ≥ 45 vs. age < 45 are (irrespectively of "fam"):

$$\text{OR}_j = \frac{p_{2j}}{1-p_{2j}} / \frac{p_{1j}}{1-p_{1j}} = e^{\beta_1}.$$

Interpretation, contd.

Since $\widehat{\beta}_1 = 3.78$ with $\widehat{se}_1 = 0.68$,
a 95% confidence interval for β_1 is
$3.78 \pm 1.96 \times 0.68$.

We have $\widehat{\text{OR}}_j = \exp(3.78) = 43.8$.

The 95% confidence interval for OR is (by exponentiating the limits above) from 11.4 to 160.0.

This quantifies how much bigger the risk of being dependent on injections is for persons who get the disease when they are less than 45 years. Remark that this is regardless of the family history.