STK4900/9900 - Lecture 7

Program

1. Logistic regression with one predictor
2. Maximum likelihood estimation
3. Logistic regression with several predictors
4. Deviance and likelihood ratio tests
5. A comment on model fit

- Sections 5.1, 5.2 (except 5.2.6), and 5.6
- Supplementary material on likelihood and deviance

Logistic regression with one predictor

$y_i = \begin{cases} 0 & \text{if nothing happens} \\ 1 & \text{if a feature characteristic we are interested in is present} \end{cases}$

We have data $(x_1, y_1), \ldots, (x_n, y_n)$

Here y_i is a binary outcome (0 or 1) for subject i and x_i is a predictor for the subject

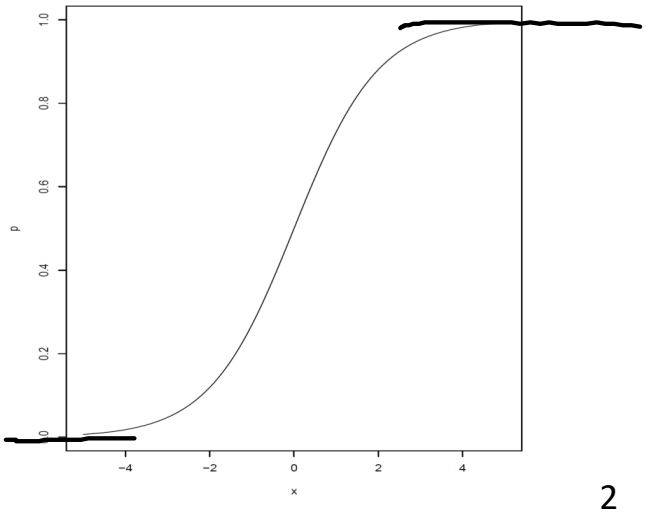
We let

$[0;1) \ni p(x) = E(y \mid x) = P(y = 1 \mid x)$

The logistic regression models take the form:

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}$$

This gives a "S-shaped" relation between $p(x)$ and x and ensures that $p(x)$ stays between 0 and 1



The logistic model may alternatively be given in terms of the odds:

$$\frac{p(x)}{1-p(x)} = \exp(\beta_0 + \beta_1 x) \qquad (*)$$

If we consider two subjects with covariate values $x+\Delta$ and x, respectively, their odds ratio becomes

$$\frac{p(x+\Delta)/[1-p(x+\Delta)]}{p(x)/[1-p(x)]} = \frac{\exp\left(\beta_0 + \beta_1 (x+\Delta)\right)}{\exp(\beta_0 + \beta_1 x)} = \exp(\beta_1 \Delta)$$

In particular e^{β_1} is the odds ratio corresponding to one unit's increase in the value of the covariate

By (*) the logistic regression model may also be given as:

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x \qquad (**)$$

Thus the logistic regression model is linear in the log-odds

Consider the WCGS study with CHD as outcome and age as predictor (individual age, not grouped age as we considered in Lecture 6)

R commands:

```
wcgs=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/data/wcgs.txt",
                sep="\t",header=T,na.strings=".")
fit=glm(chd69~age, data=wcgs,family=binomial)
summary(fit)
```

OR < 1 protective factor
OR = 1 does not change
OR > 1 risk factor

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9395	0.5493	-10.813	< 2e-16
age	0.0744	0.0113	6.585	4.56e-11

$e^{\beta_1 \Delta}$

The odds ratio for one year increase in age is $e^{0.0744} = 1.077$

while the odds ratio for a ten-year increase is $e^{0.0744 \cdot 10} = 2.10$

log OR < 0 protective
log OR = 0 no effect
log OR > 0 risk

(The numbers deviate slightly from those on slide 25 from Lecture 6, since there we used mean age for each age group while here we use the individual ages)

How is the estimation performed for the logistic regression model?

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:

- $y_i \sim \mathrm{N}(\mu_i, \sigma^2)$ 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 (cf slide 12 from Lecture 1):

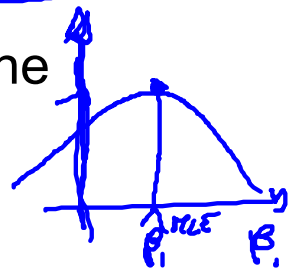
$$f(y_i, \mu_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(y_i - \mu_i)^2\}$$

The likelihood is the simultaneous density

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

$\beta_0 + \beta_1 x_i$

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

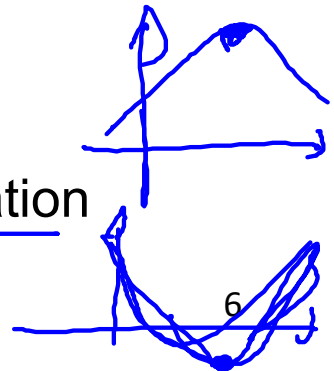
Maximizing the likelihood L is the same as maximizing

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

$$\max_{\mu_i} \left(-\sum_{i=1}^{n}(y_i - \mu_i)^2\right)$$

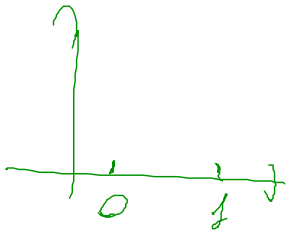
which is the same as minimizing $\sum_{i=1}^{n}(y_i - \mu_i)^2$

$$\min \left(\sum_{i=1}^{n}(y_i - \mu_i)\right)^2$$



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

We then consider the situation for logistic regression



We have data $(x_1, y_1), \ldots, (x_n, y_n)$, where y_i is a binary outcome (0 or 1) for subject i and x_i is a predictor

Here we have

$$P(y_i = 1 \mid x_i) = p_i$$

$$P(y_i = 0 \mid x_i) = 1 - p_i$$

where

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

Thus the distribution of y_i may be written as

$$P(y_i \mid x_i) = p_i^{y_i} (1 - p_i)^{1 - y_i}$$

$P(y_i = 1 \mid x) = p_i$

$y_i = 1 \rightarrow p_i^1 (1 - p_i)^{1-1} = p_i$

$y_i = 0 \rightarrow p_i^0 (1 - p_i)^{1-0} = 1 - p_i$

The likelihood becomes

$$L = \prod_{i=1}^{n} P(y_i | x_i) = \prod_{i=1}^{n} p_i^{\,y_i} (1 - p_i)^{1-y_i}$$

Since

$$p_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

the likelihood is, for given observations, a function of the unknown parameters β_0 and β_1

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

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

Confidence interval for β_1 and odds ratio

95% confidence interval for β_1 (based on the normal approximation):

$$\hat{\beta}_1 \pm 1.96 \cdot se(\hat{\beta}_1)$$

$OR = \exp(\beta_1)$ is the odds ratio for 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

In the CHD example we have $\hat{\beta}_1 = 0.0744$ and $se(\hat{\beta}_1) = 0.0113$

95% confidence interval for β_1 :

$0.0744 \pm 1.96 \cdot 0.0113$ i.e. from 0.052 to 0.096

Estimate of odds ratio $OR = \exp(0.0744) = 1.077$

95% confidence interval for *OR* :

from $\exp(0.052) = 1.053$ to $\exp(0.096) = 1.101$

R function for computing odds ratio with 95% confidence limits

```
expcoef=function(glmobj)
{
regtab=summary(glmobj)$coef
expcoef=exp(regtab[,1])
lower=expcoef*exp(-1.96*regtab[,2])
upper=expcoef*exp(1.96*regtab[,2])
cbind(expcoef,lower,upper)
}

expcoef(fit)
```

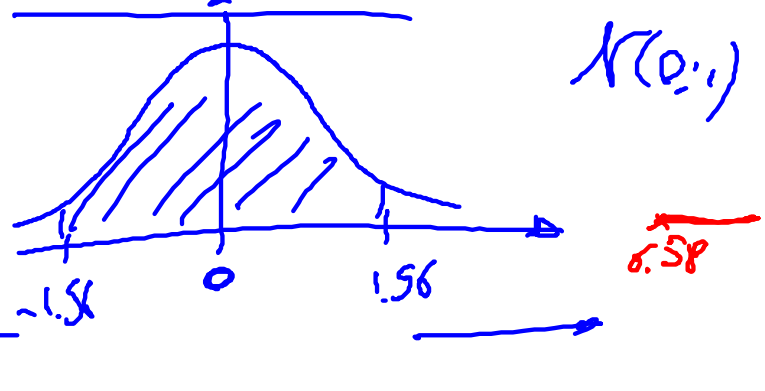
R output (edited):

	expcoef	lower	upper
(Intercept)	0.0026	0.0009	0.0077
age	1.077	1.054	1.101

Wald test for $H_0 : \beta_1 = 0$

To test the null hypothesis $H_0 : \beta_1 = 0$ versus the two-sided alternative $H_A : \beta_1 \neq 0$ we often use the Wald test statistic:

$$z = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)}$$



We reject H_0 for large values of $|z|$

Under H_0 the test statistic is approximately standard normal

P-value (two-sided): $P = 2\,P(Z > |z|)$ where Z is standard normal

In the CHD example we have $\hat{\beta}_1 = 0.0744$ and $se(\hat{\beta}_1) = 0.0113$

Wald test statistic

$$z = 0.0744 / 0.0113 = 6.58$$

which is highly significant (cf. slide 4)

Multiple logistic regression

multiple logistic regression
=
multivariable logistic regression
≠
multivariate logistic regression

Assume now that we for each subject have

- a binary outcome y
- predictors $x_1, x_2, ..., x_p$

We let

$$p(x_1, x_2, ..., x_p) = E(y \mid x_1, x_2, ..., x_p) = P(y = 1 \mid x_1, x_2, ..., x_p)$$

Logistic regression model:

$$p(x_1, x_2, ..., x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}$$

Alternatively the model may be written:

$$\log\left(\frac{p(x_1, x_2, ..., x_p)}{1 - p(x_1, x_2, ..., x_p)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

The logistic model may also be given in terms of the odds:

$$\frac{p(x_1, x_2, ..., x_p)}{1 - p(x_1, x_2, ..., x_p)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)$$

If we consider two subjects with values $x_1 + \Delta$ and x_1, for the first covariate and the same values for all the others, their odds ratio becomes

$$\frac{p(x_1 + \Delta, x_2, ..., x_p) \Big/ \left[1 - p(x_1 + \Delta, x_2, ..., x_p)\right]}{p(x_1, x_2, ..., x_p) \Big/ \left[1 - p(x_1, x_2, ..., x_p)\right]}$$

$$= \frac{\exp(\beta_0 + \beta_1 (x_1 + \Delta) + \beta_2 x_2 + ... + \beta_p x_p)}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)} = \exp(\beta_1 \Delta)$$

In particular e^{β_1} is the odds ratio corresponding to one unit's increase in the value of the first covariate *holding all other covariates constant*

A similar interpretation holds for the other regression coefficients

Wald tests and confidence intervals

- $\hat{\beta}_j$ = MLE for β_j
- $se(\hat{\beta}_j)$ = standard error for $\hat{\beta}_j$

To test the null hypothesis $H_{0j} : \beta_j = 0$ we use the Wald test statistic:

$$z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

which is approximately N(0,1)-distributed under H_{0j}

95% confidence interval for β_j : $\hat{\beta}_j \pm 1.96 \cdot se(\hat{\beta}_j)$

$OR_j = \exp(\beta_j)$ is the odds ratio for one unit's increase in the value of the *j*-th covariate *holding all other covariates constant*

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

Consider the WCGS study with CHD as outcome and age, cholesterol (mg/dL), systolic blood pressure (mmHg), body mass index (kg/m^2), and smoking (yes, no) as predictors (as on page 152 in the text book we omit an individual with an unusually high cholesterol value)

R commands:

```
wcgs.mult=glm(chd69~age+chol+sbp+bmi+smoke, data=wcgs, family=binomial,
                    subset=(chol<600))
summary(wcgs.mult)
```

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-12.3110	0.9773	-12.598	< 2e-16
age	0.0644	0.0119	5.412	6.22e-08
chol	0.0107	0.0015	7.079	1.45e-12
sbp	0.0193	0.0041	4.716	2.40e-06
bmi	0.0574	0.0264	2.179	0.0293
smoke	0.6345	0.1401	4.526	6.01e-06

Odds ratios with confidence intervals

R command (using the function from slide 10):

```
expcoef(wcgs.mult)
```

R output (edited):

	expcoef	lower	upper
(Intercept)	4.50e-06	6.63e-07	3.06e-05
age	1.067	1.042	1.092
chol	1.011	1.008	1.014
sbp	1.019	1.011	1.028
bmi	1.059	1.006	1.115
smoke	1.886	1.433	2.482

$\Delta = 1$

$\Delta = 10$

For a numerical covariate it may be more meaningful to present an odds ratio corresponding to a larger increase than one unit (cf. slide 13)

This is easily achieved by refitting the model with a rescaled covariate

If you (e.g) want to study the effect of a ten-years increase in age, you fit the model with the covariate **age_10=age/10**

R commands:

```
wcgs.resc=glm(chd69~age_10+chol_50+sbp_50+bmi_10+smoke, data=wcgs,
                    family=binomial, subset=(chol<600))
summary(wcgs.resc)
```

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.006	0.116	-12.598	< 2e-16
age_10	0.644	0.119	5.412	6.22e-08
chol_50	0.537	0.076	7.079	1.45e-12
sbp_50	0.965	0.205	4.716	2.40e-06
bmi_10	0.574	0.264	2.179	0.0293
smoke	0.634	0.140	4.526	6.01e-06

Note that values of the Wald test statistic are not changed (cf. slide 15)

Odds ratios with confidence intervals:

R command (using the function from slide 10):

```
expcoef(wcgs.resc)
```

R output (edited):

	expcoef	lower	upper
(Intercept)	0.0494	0.0394	0.0621
age_10	1.9050	1.5085	2.4057
chol_50	1.7110	1.4746	1.9853
sbp_50	2.6240	1.7573	3.9180
bmi_10	1.7760	1.0595	2.9770
smoke	1.8860	1.4329	2.4824

An aim of the WCGS study was to study the effect on CHD of certain behavioral patterns, denoted A1, A2, B3 and B4

Behavioral pattern is a categorical covariate with four levels, and must be fitted as a factor in R

R commands:

```
wcgs$behcat=factor(wcgs$behpat)
wcgs.beh=glm(chd69~age_10+chol_50+sbp_50+bmi_10+smoke+behcat,
                    data=wcgs, family=binomial, subset=(chol<600))
summary(wcgs.beh)
```

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7527	0.2259	-12.19	< 2e-16
age_10	0.6064	0.1199	5.057	4.25e-07
chol_50	0.5330	0.0764	6.980	2.96e-12
sbp_50	0.9016	0.2065	4.367	1.26e-05
bmi_10	0.5536	0.2656	2.084	0.0372
smoke	0.6047	0.1411	4.285	1.82e-05
behcat2	0.0660	0.2212	0.298	0.7654
behcat3	-0.6652	0.2423	-2.746	0.0060
behcat4	-0.5585	0.3192	-1.750	0.0802

Here we may be interested in :

- Testing if behavioral patterns have an effect on CHD risk
- Testing if it is sufficient to use two categories for behavioral pattern (A and B)

In general we consider a logistic regression model:

$$p(x_1, x_2, ..., x_p) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p)}$$

Here we want to test the null hypothesis that q *of* the β_j's are equal to zero, or equivalently that there are q linear restrictions among the β_j's

Examples:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \qquad (q = 4)$$

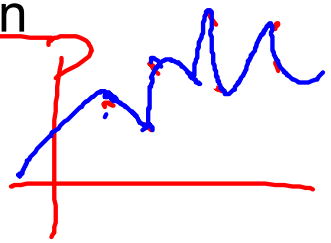
$$H_0 : \beta_1 = \beta_2 \text{ and } \beta_3 = \beta_4 \qquad (q = 2)$$

Deviance and sum of squares

For the linear regression model 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 log-likelihood and the sum of squares for the linear regression model



For the linear regression model $l = \log L$ takes the form (cf. slide 6):

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

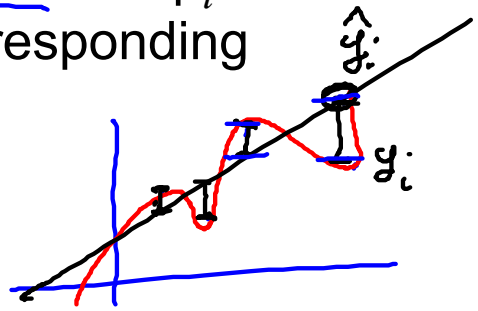
The log-likelihood obtains its *largest* value for the *saturated model*, i.e. the model where there are no restrictions on the μ_i

For the saturated model the μ_i are estimated by $\tilde{\mu}_i = y_i$, and the log-likelihood becomes

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

For a given specification of the linear regression model the μ_i are estimated by the fitted values, i.e. $\hat{\mu}_i = \hat{y}_i$, with corresponding log-likelihood

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



The deviance for the model is defined as $D = 2(\tilde{l} - \hat{l})$ and it becomes

$$D = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2$$

For the linear regression model the deviance is just the sum of squares for the fitted model divided by σ^2

Deviance for binary data

We then consider logistic regression with data

$$(y_i, x_{1i}, x_{2i}, ..., x_{pi}) \qquad i = 1, 2, ..., n$$

where y_i is binary response and the x_{ji} are predictors

We introduce $p_i = P(y_i = 1 \mid x_{1i}, x_{2i}, ..., x_{pi})$ and note that the log-likelihood $l = l(p_1, ..., p_n)$ is a function of $p_1, ..., p_n$ (cf. slide 8)

For the *saturated model*, i.e. the model where there are no restrictions on the p_i, the p_i are estimated by $\tilde{p}_i = y_i$ and the log-likelihood takes the value $\tilde{l} = l(\tilde{p}_1, ..., \tilde{p}_n)$

For a fitted logistic regression model we obtain the estimated probabilities

$$\hat{p}_i = \hat{p}(x_{1i}, x_{2i}, ..., x_{pi}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_p x_{pi})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_p x_{pi})}$$

and the corresponding value $\hat{l} = l(\hat{p}_1, ..., \hat{p}_n)$ of the log-likelihood

The deviance for the model is defined as

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

The deviance itself is not of much use for binary data

But by comparing the deviances of two models, we may check if one gives a better fit than the other.

Consider the WCGS study with age, cholesterol, systolic blood pressure, body mass index, smoking and behavioral pattern as predictors (cf slide 19)

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7527	0.2259	-12.19	< 2e-16
age_10	0.6064	0.1199	5.057	4.25e-07
chol_50	0.5330	0.0764	6.980	2.96e-12
sbp_50	0.9016	0.2065	4.367	1.26e-05
bmi_10	0.5536	0.2656	2.084	0.0372
smoke	0.6047	0.1411	4.285	1.82e-05
behcat2	0.0660	0.2212	0.298	0.7654
behcat3	-0.6652	0.2423	-2.746	0.0060
behcat4	-0.5585	0.3192	-1.750	0.0802

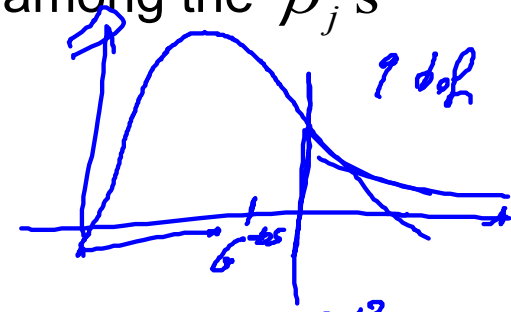
Null deviance: 1774.2 on 3140 degrees of freedom
Residual deviance: 1589.6 on 3132 degrees of freedom

The deviance of the fitted model is denoted "residual deviance" in the output

The "null deviance" is the deviance for the model with no covariates, i.e. for the model where all the p_i are assumed to be equal

Deviance and likelihood ratio tests

We want to test the null hypothesis H_0 that q *of* the β_j's are equal to zero, or equivalently that there are q linear restrictions among the β_j's



To test the null hypothesis, we use the test statistic

$$G = D_0 - D$$

where D_0 is the deviance under the null hypothesis and D is the deviance for the fitted model (not assuming H_0)

We reject H_0 for large values of G

To compute P-values, we use that the test statistic G is chi-square distributed (χ^2) with q degrees of freedom under H_0

We will show how we may rewrite G in terms of the likelihood ratio

We have

$$D = 2(\tilde{l} - \hat{l}) \quad \text{and} \quad D_0 = 2(\tilde{l} - \hat{l}_0)$$

Here

$$\hat{l} = \log \hat{L} \quad \text{and} \quad \hat{l}_0 = \log \hat{L}_0$$

where

$$\hat{L} = \max_{model} L \quad \text{and} \quad \hat{L}_0 = \max_{H_0} L$$

Thus

$$G = D_0 - D = 2(\tilde{l} - \hat{l}_0) - 2(\tilde{l} - \hat{l}) = -2(\hat{l}_0 - \hat{l}) = -2\log\left(\hat{L}_0 / \hat{L}\right)$$

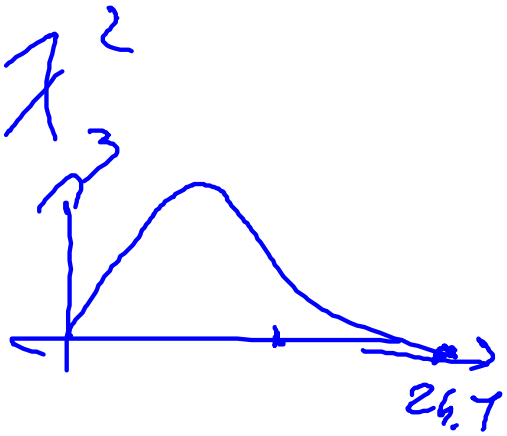
Thus large values of G corresponds to small values of the likelihood ratio $\hat{L}_0 / \hat{L}$ and the test based on G is equivalent to the likelihood ratio test

For the model with age, cholesterol, systolic blood pressure, body mass index, smoking, and behavioral pattern as predictors (cf slide 25) the deviance becomes $D = 1589.6$

For the model without behavioral pattern (cf slide 17) the deviance takes the value $D_0 = 1614.4$

The test statistic takes the value:

$$G = D_0 - D = 1614.4 - 1589.6 = 24.8$$



R commands:

```
anova(wcgs.resc,wcgs.beh,test="Chisq")
```

R output (edited):

Analysis of Deviance Table

Model 1: chd69 ~ age_10 + chol_50 + sbp_50 + bmi_10 + smoke
Model 2: chd69 ~ age_10 + chol_50 + sbp_50 + bmi_10 + smoke + behcat

	Resid.Df	Resid.Dev	Df	Deviance	P(>|Chi|)
1	3135	1614.4			
2	3132	1589.6	3	24.765	1.729e-05

Model fit for linear regression (review)

1. Linearity
2. Constant variance
3. Independent responses
4. Normally distributed error terms and no outliers

Model fit for logistic regression

1. Linearity: Still relevant, see following slides
2. Heteroscedastic model, $Var(y_i \mid x_i) = p_i(1-p_i)$, i.e. depends on $E(y_i \mid x_i) = p_i$.
   However this non-constant variance is taken care of by the maximum likelihood estimation.
3. Independent responses: See Lecture 10 on Friday.
4. Not relevant, data are binary, no outliers in responses (but there could well be extreme covariates, influential observations).

Checking linearity for logistic regression

We want to check if the probabilities can be adequately described by the linear expression

$$\log\left(\frac{p(x_1, x_2, ..., x_p)}{1 - p(x_1, x_2, ..., x_p)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$$

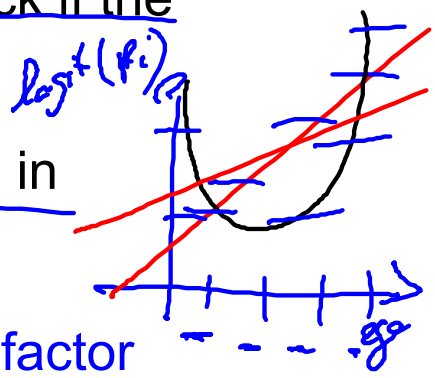
We will discuss 3 approaches:

1. Grouping the covariates
2. Adding square terms or logarithmic terms to the model
3. Extending the model to generalized additive models (GAM)

$$\log\left(\frac{p(x_1, x_2, ..., x_p)}{1 - p(x_1, x_2, ..., x_p)}\right) = \beta_0 + f_1(x_1) + f_2(x_2) + ... + f_p(x_p)$$

1. Grouping the variables

For a simple illustration we consider the situation where age is the only covariate in the model for CHD, and we want to check if the effect of age is linear (on the log-odds scale)



The procedure will be similar if there are other covariates in addition to age

We may here fit a model considering the age group as a factor (age groups: 35-40, 41-45, 46-50, 51-55, 56-60)

Or we may fit a model where the mean age in each age group is used as numerical covariate (means: 39.5, 42.9, 47.9, 52.8, 57.3)

We may then use a deviance test to check if flexible a categorical model gives a better fit than the linear numerical.

Here we find no improvement, p=0.269

R commands:

```
fit.catage=glm(chd69~factor(agec), data=wcgs,family=binomial)
summary(fit.catage)
wcgs$agem=39.5*(wcgs$agec==0)+42.9*(wcgs$agec==1)+47.9*(wcgs$agec==2)+
                52.8*(wcgs$agec==3)+57.3*(wcgs$agec==4)
fit.linage=glm(chd69~agem, data=wcgs,family=binomial)
summary(fit.linage)
anova(fit.linage, fit.catage,test="Chisq")
```

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8043	0.1850	-15.162	< 2e-16
factor(agec)1	-0.1315	0.2310	-0.569	0.569
factor(agec)2	0.5307	0.2235	2.374	0.018
factor(agec)3	0.8410	0.2275	3.697	0.0002
factor(agec)4	1.0600	0.2585	4.100	4.13e-05

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9466	0.5616	-10.588	< 2e-16
agem	0.0747	0.0116	6.445	1.15e-10

Model 1: chd69 ~ agem
Model 2: chd69 ~ factor(agec)

	Resid. Df	Resid. Dev	Df	Deviance	P(>|Chi|)
1	3152	1740.2			
2	3149	1736.3	3	3.928	0.269

2. Adding square terms or log-terms

The simple model $\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x$

can be extended to more flexible models such as

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x + \beta_2 x^2$$

or

$$\log\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x + \beta_2 \log(x)$$

We may then use a deviance test to check if the flexible models give a better fit than the original.

Here we neither find any improvement, p=0.79 and p=0.89

R commands:

```
fit=glm(chd69~age, data=wcgs,family=binomial)
fita2=glm(chd69~age+I(age^2), data=wcgs,family=binomial)
anova(fit,fita2,test="Chisq")

fitlog=glm(chd69~age+log(age), data=wcgs,family=binomial)
anova(fit,fitlog,test="Chisq")
```

R output (edited):

```
> anova(fit,fita2,test="Chisq")
Model 1: chd69 ~ age
Model 2: chd69 ~ age + I(age^2)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3152     1738.4
2      3151     1738.3  1 0.069473   0.7921

 > anova(fit,fitlog,test="Chisq")
 Model 1: chd69 ~ age
 Model 2: chd69 ~ age + log(age)
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
 1      3152     1738.4
 2      3151     1738.3  1 0.019058   0.8902
```

3. Generalized additive model

In this example just with one covariate: $\log\left(\frac{p(x_1)}{1-p(x_1)}\right) = \beta_0 + f_1(x_1)$

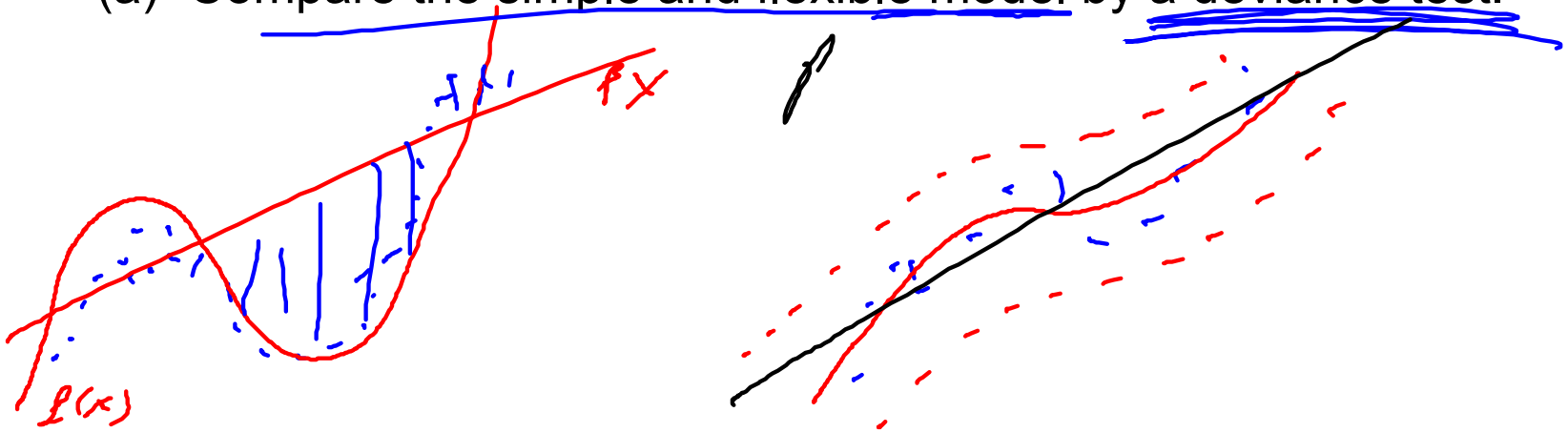
where $f_1(x_1)$ is a smooth function estimated by the program.

The approach can easily be extended to several covariates.

We can then

(a) Plot the estimated function with confidence intervals. Will a straight line fit within the confidence limits?

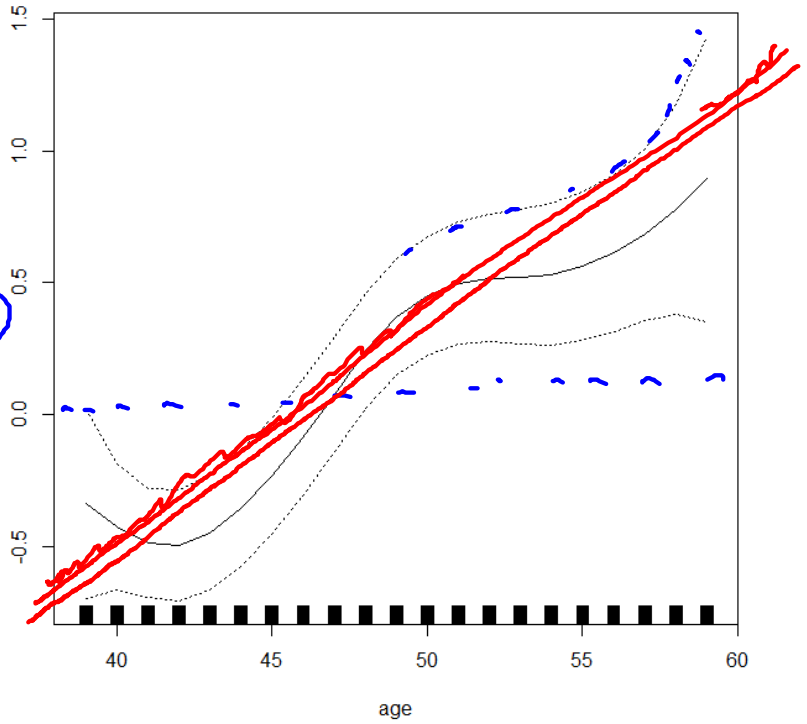
(a) Compare the simple and flexible model by a deviance test.



R output (edited):

```
> library(gam)
>fitgam=gam(chd69~s(age), data=wcgs,
                          family=binomial)
> plot(fitgam,se=T)
> anova(fit,fitgam,test="Chisq")
Analysis of Deviance Table

Model 1: chd69 ~ age
Model 2: chd69 ~ s(age)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      3152     1738.4
2      3149     1729.6  3   8.7622  0.03263 *
```



For these data

(a) The informal graphical check just allows a straight line within confidence limits.

(a) However, the deviance test gives a weakly significant deviation from linearity (p=0.032)

There may thus be some unimportant deviation from linearity.

Deviance and grouped data

On slides 26-29 in Lecture 6 we saw that we got the same estimates and standard errors when we fitted the model with mean age in each age group as numerical covariate using binary data and grouped data

R commands:

```
summary(fit.linage)

chd.grouped=read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v16/chd_grouped.txt ",
header=T)
fit.grouped=glm(cbind(chd,no-chd)~agem, data=chd.grouped, family=binomial)
summary(fit.grouped)
```

R output (edited):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9466	0.5616	-10.588	< 2e-16
agem	0.0747	0.0116	6.445	1.15e-10

Null deviance: 1781.2 on 3153 degrees of freedom
Residual deviance: 1740.2 on 3152 degrees of freedom

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.9466	0.5616	-10.588	< 2e-16
agem	0.0747	0.0116	6.445	1.15e-10

Null deviance: 44.95 on 4 degrees of freedom
Residual deviance: 3.928 on 3 degrees of freedom

We see that the "residual deviance" and the "null deviance" are not the same when we use binary data and when we use grouped data

However, the difference between the two is the same in both cases

As long as we look at differences between deviances, it does not matter whether we used binary or grouped data