Binary data: Logistic regression

Jan van den Broek

January, 2024

Universiteit Utrecht

Odds

A research

In 1992 a study was done among 98 male HIV-patients.

It was measured during approximately one year, whether or not they developed one ore more episodes of urinary track disease (UTD).

This measure, UTD, is called the **disease variable**.

One of the research objectives was to determine the relation between UTD and the immune status of the patient.

The immune status of a patient is called the **exposure variable**. The immune status was measured as low (CD4+ cell count lower than 200×10^6) or as high.

A research

The data are in the following table:

	UTD	
Immune status	no	yes
high	48	3
low	33	14

The UTD-diseased are coded as 1, the non-diseased as 0.

The same coding is used for the exposure: the group exposed (low) is coded as 1 and the unexposed (high) as 0.

The population: a model for the data generating process

The fraction HIV patients who have UTD in the population group exposed is denoted as π_1, and the corresponding fraction in the population group unexposed is π_0.

So if $\pi_1 = 0.3$ this means that 30% of low immune status patients in the population had UTD.

The population: a model for the data generating process

In the group exposed the **odds** of the disease is defined as the fraction diseased divided by the fraction non-diseased: $\frac{\pi_1}{1-\pi_1}$.

Suppose, as an example, that the fraction diseased is $\frac{1}{3}$, the odds then is $\frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$.

This means that for every diseased individual there are two non-diseased individuals. So the proportion $\pi_1 : 1 - \pi_1$ is $1 : 2$.

Odds in the population group unexposed: $(\frac{\pi_0}{1-\pi_0})$ the proportion diseased vs non-diseased in the unexposed group.

The population: a model for the data generating process

The odds ratio now is

$$\omega = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$$

When for example the odds ratio is 3 this means that the proportion diseased/non-diseased in the group exposed is 3 times higher as compared to the group unexposed.

The sample

From the population we have a sample of 98 HIV infected males.

		UTD		
		No	Yes	
exposed	No	a	b	$P_0 = \frac{b}{a+b}$
	Yes	c	d	$P_1 = \frac{d}{c+d}$

So $P_1 = \frac{14}{70} = 0.208$ and $P_0 = \frac{3}{28} = 0.050$. These are estimates

The logistic regresion model

The logistic regression model in the population

The logistic regression model relates the log of the odds to the exposure in a **linear manner**.

In the population this is $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{exposure}$ in which (exposure) is a variable having values 0 for the unexposed and 1 for the exposed.

The logistic regression model in the population

- **unexposed**: The model is: $\ln\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha + \beta \cdot 0 = \alpha$ Thus α is the **log-odds in the group unexposed**, or, in other words it's the logarithm of the proportion diseased/non-diseased in group unexposed.
- **exposed**:The model is: $\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha + \beta \cdot 1$. Subtracting the model for the unexposed group from the model for the exposed group: $\ln\left(\frac{\pi_1}{1-\pi_1}\right) - \ln\left(\frac{\pi_0}{1-\pi_0}\right) = \ln\left(\frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}\right) = \ln(\omega) = \alpha + \beta - \alpha = \beta$. It follows that β is the **log-odds ratio for the exposed vs the unexposed** and e^{β} is the odds ratio for the exposed vs the unexposed.

The logistic regression model in the sample

In the sample we have to estimate the model, which means that we need estimates for α and β.

These estimates are called (a) and (b). We can find these estimates by replacing the population fractions in α and β by the sample fractions.

This gives $a = \ln\left(\frac{P_0}{1-P_0}\right)$ the log of the proportion diseased/non-diseased in the sample group unexposed, and $b = \ln(OR)$, the log of the sample odds ratio.

The logistic regression model in the sample

Above, the exposure variable can take only two values 0 and 1. It's a group indicator. It is also possible to have a continuous exposure variable like age or weight in the model.

The model looks just the same: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{age}$.

For age equal to 0 the model is $\ln\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha$. So α is the **log-odds for those who have exposure value 0**.

The logistic regression model in the sample

The model for those at age x is: $\ln\left(\frac{\pi_0}{1-\pi_0}\right) = \alpha + \beta \cdot x$, for those with age $x+1$ $\ln\left(\frac{\pi_1}{1-\pi_1}\right) = \alpha + \beta \cdot (x+1)$.

Subtracting the model for those with age x from the model from those with age $X+1$: $\ln\left(\frac{\pi_1}{1-\pi_1}\right) - \ln\left(\frac{\pi_0}{1-\pi_0}\right) = \ln\left(\frac{\frac{\pi_1}{1-\pi 1}}{\frac{\pi_0}{1-\pi_0}}\right) = \ln(\omega) = \alpha + \beta \cdot x + \beta - \alpha - \beta \cdot x = \beta$.

This means that β is the log-odds ratio and e^{β} the odds ratio for a one year increase in age. To put it differently: β is the **log-odds ratio if patients of a certain age are compared with patients who are one year older**. (It is the change in log-odds if the exposure changes with 1)

The likelihood

Likelihood

Let's forget about the exposure variable for a while. From every patient in the sample it is observed whether or not he has UTD. So for patient i we measure Y_i and observe outcome y_i, which can be 0 (no UTD) or 1 (UTD).

The fraction patients in the population who have UTD is π so the probability that a random selected patient has UTD is π.

This means that $Y_i = 1$ with probability π and $Y_i = 0$ with probability $1 - \pi$. This can be written more compactly as: $P(Y_i = y_i) = \pi^{y_i}(1 - \pi)^{1-y_i}$.

Likelihood

We have 98 patients.

The probability that patient 1 has UTD ($y_1 = 1$) or not ($y_1 = 0$) is: $P(Y_1 = y_1) = \pi^{y_1}(1-\pi)^{1-y_1}$.

For patient 2 this is: $P(Y_2 = y_2) = \pi^{y_2}(1-\pi)^{1-y_2}$ etc.

Now we want to know what the probability of the observations of all the 98 patients is. This is the probability that you observe a 0 or a 1 for patient 1 and observe a zero or a 1 for patient 2 and . . . etc.

This is $P(Y_1 = y_1 \textit{ and } Y_2 = y_2 \textit{ and } \ldots \textit{ and } Y_{98} = y_{98}) = P(Y_1 = y_1) \cdot P(Y_2 = y_2) \cdots P(Y_{98} = y_{98})$.

This **probability of the observed data** is called **the likelihood**, denoted by $L()$. It says how probable the observed data, the zero's and the one's observed from the 98 patients, is.

Likelihood

We can now plug in the probability of each observation:

$$
\begin{aligned}
L(\pi) &= P(Y_1 = y_1 \text{ and } Y_2 = y_2 \text{ and } \ldots \text{ and } Y_{98} \\
&= P(Y_1 = y_1) \cdot P(Y_2 = y_2) \cdots P(Y_{98} = y_{98}) \\
&= \pi^{y_1} \cdot (1-\pi)^{1-y_1} \cdot \pi^{y_2} \cdot (1-\pi)^{1-y_2} \cdots \pi^{y_{98}} \cdot (1-\pi)^{1-y_{98}} \\
&= \pi^{\sum_{i=1}^{98} y_i} \cdot (1-\pi)^{\sum_{i=1}^{98}(1-y_i)}
\end{aligned}
$$

Likelihood

The idea is now to estimate π in such a manner that the probability of the observed data is maximal or, differently, that the likelihood is maximal (**maximum likelihood estimates**).

Instead of maximizing the likelihood one usually takes the **log of the likelihood** :

$$l(\pi) = \ln[L(\pi)] = \sum_{i=1}^{98} [y_i \ln(\pi) + (1 - y_i) \ln(1 - \pi)]$$

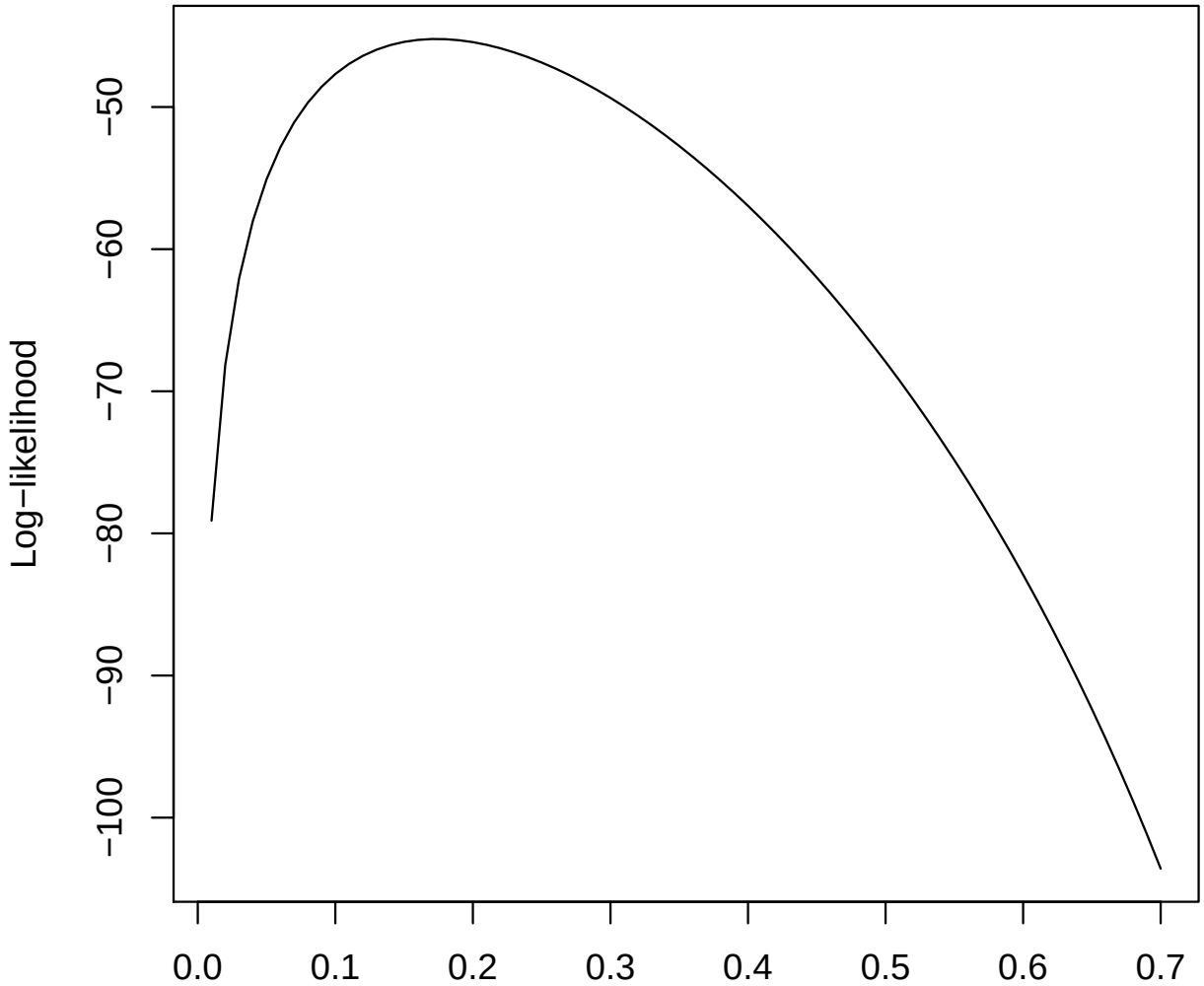
Likelihood

To see for which π the log-likelihood is maximal, one calculates the derivative with respect to π: $l'(\pi)$.

Putting this equal to zero and solving, gives the value of π for which the log-likelihood, and thus the likelihood, is maximal.

Let's call this value p. Maximizing the log-likelihood is the same as maximizing the likelihood.

Likelihood



Likelihood

If one makes a plot of the log-likelihood for different values of π, one can observe two cases in the neighborhood of the maximum:

Log-likelihood is flat If this is the case, the position of the maximum is not well determined. There is then not much information about the value for π for which the log-likelihood is maximal. In this case the standard error of the estimate of π is large.

Log-likelihood is peaked In this case, the position of the maximum is very well determined. There is a lot of information about the value for π for which the log-likelihood is maximal. In this case the standard error of the estimate of π is small.

Likelihood

A measure for the peakedness of the log-likelihood is the second derivative at it's maximum : $l''(p)$.

For a flat function the second derivative is small, for a peaked function it is large. The amount of information is found by calculating the negative of the second derivative:

$$Information = -l''(p)$$

From this the standard error can be calculated:

$$standard\ error = \sqrt{\frac{1}{Information}}$$

This relation shows that one should be careful with large standard errors since this means that the information is small. In that case there is not enough information to get a good estimate.

Likelihood

Summary: The likelihood is the probability of the observed data, given the model, seen as a function of the parameter π.
The first derivative gives the maximum likelihood estimator, that value of the parameter for which the log-likelihood, and thus the likelihood is maximal.
The second derivative gives the standard error.

Evidence

Evidence

The logistic model: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{exposure}$ or $\pi = \frac{e^{\alpha+\beta\cdot \textit{exposure}}}{1+e^{\alpha+\beta\cdot \textit{exposure}}}$.

We want to estimate α and β. This can be done by using the log-likelihood:

$$l(\pi) = \ln[L(\pi)] = \sum_{i=1}^{98} [y_i \ln(\pi) + (1 - y_i) \ln(1 - \pi)]$$

For π we take $\frac{e^{\alpha+\beta\cdot \textit{exposure}}}{1+e^{\alpha+\beta\cdot \textit{exposure}}}$.

Evidence

The log-likelihood now depends on α and β. Maximizing with respect to α and β gives the values *a* and *b* for which the log-likelihood is maximal. The second derivatives give the standard errors of *a* and *b*.

```
##
## Call:
## glm(formula = episode ~ immune, family = binomial, data = lr1)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7726     0.5951  -4.659 3.18e-06 ***
## immune        1.9151     0.6752   2.836  0.00456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.424  on 97  degrees of freedom
## Residual deviance: 80.070  on 96  degrees of freedom
## AIC: 84.07
##
## Number of Fisher Scoring iterations: 5
```

Evidence

Is there evidence in the data to state that immune status is related to UTD?

The logistic model $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{exposure}$

no relation with exposure: Then the model states that the log-odds is constant and does not depend on immune status so $\beta = 0$.This is model 0.

relation with exposure: The logistic model in which β is not equal to zero is model 1.

So actually two models are fitted and then these models are compared to see which one is **best supported by the data**.

Evidence

Model 0: The model is $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha$ or $\pi = \frac{e^{\alpha}}{1+e^{\alpha}}$. This model for π is used in the log-likelihood.

The derivative w.r.t. α gives the maximum likelihood estimator a, the second derivative gives the standard error.

The log-likelihood has its maximum for a. The maximum of the log-likelihood, l_0 or $l(a)$, is obtained by plugging in the value a for α.

The maximum for the likelihood then is $L_0 = e^{l_0}$. L_0 or $L(a)$ gives the probability of the data using model0.

Evidence

Model 1: The model is $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \mathit{exposure}$ or $\pi = \frac{e^{\alpha+\beta\cdot \mathit{exposure}}}{1+e^{\alpha+\beta\cdot \mathit{exposure}}}$.

Plug in this value for π in the log-likelihood. The first derivatives gives the maximum likelihood estimator a and b , the second derivatives gives the standard errors.

The log-likelihood has its maximum for a and b. The maximum of the log-likelihood, l_1 or $l(a, b)$, is obtained by plugging in the value a and b for α and β.

The maximum for the likelihood then is $L_1 = e^{l_1}$. L_1 or $L(a, b)$ gives the probability of the data using model1.

Evidence

If L_1 is larger than L_0 then the observed data is more likely using model 1, this model makes the data more probable.

In that case under model 1, one better understands why the data is observed as it is.

To see if model 1 makes the data more likely than model 0 one calculates **the likelihood ratio:** $\frac{L_1}{L_0}$.

If this ratio has for instance an outcome of three, it means that the probability of observing the data is 3 times higher using model 1 as compared to model 0.

More precise: the likelihood of model 1 is 3 times larger as the likelihood of model 0 meaning, that if the estimated model 1 was the generator of the data, the probability of the data would be 3 times larger as when the estimated model 0 was the generator of the data.

Evidence

Usually a model with a lot of exposure variables in it is better than a model with only a few.

Akaike's information criterion (AIC):
$AIC = -2 \cdot (\text{ log-likelihood }) + 2 \cdot (\text{ number of parameters in the model }) = -2 \cdot l + 2 \cdot p$

were p is the number of parameters (α's and β's) in the model.

One can fit different models, calculate for every model the AIC and see which model is best as compared to the other models.

Evidence

That model is best which has the largest likelihood and thus the largest log-likelihood. So that model is best that has the lowest AIC as compared to the others.

If the difference in AIC between models is small then the model with the smallest number of parameters is chosen. One then can say that there is not much evidence in the data to keep these variables in the model. **Occam's Razor**(Wiliam Occam, 1300-1349).

As a rough guide a difference in AIC's is considered small if it is smaller then 2.

Evidence

As should be clear the AIC is not a measure of how good a model fits the data, **it is a measure of how good the model fits to the data as compared to the other models** that are fitted.

Evidence

From the output for model 1 above it can seen that the AIC equals 84.07 so the log-likelihood for this model is $l_1 = -40.035$.

```
##
## Call:
## glm(formula = episode ~ 1, family = binomial, data = lr1)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5612     0.2668  -5.853 4.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.424  on 97  degrees of freedom
## Residual deviance: 90.424  on 97  degrees of freedom
## AIC: 92.424
##
## Number of Fisher Scoring iterations: 3
```

Evidence

So the AIC for this model is 92.424 which is much larger than that for model 1.

The log-likelihood is $l_0 = -45.212$. From this $l_1 - l_0 = 5.177$ and $\frac{L_1}{L_0} = 175.9$.

So model 1 makes the oserved data about 176 time as likely as model 0.

Evidence

hypothesis: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$

In that case one can look at $2 \cdot \ln\left(\frac{L_1}{L_0}\right) = 2(l_1 - l_0)$

this has approximately a chi-squared distribution with *df* degrees of freedom where *df* is the difference of the number of parameters between the models

This approximate is good if the number of observations is large.

This is called the **likelihood ratio test**.

Evidence

An other way to test the hypothesis mentioned above is by the so called Wald test.

One can calculate this test statistic by deviding the estmate of β by its standard error: $\frac{b}{se(b)}$.

The p-value of the outcome of this test statistic is calculated by using the standard normal distribution.

The likelihood ratio test is prefered over the wald test since in most applications the chi-square approximation to the likelihood ratio test statistics is better then the normal approximation for the wald statistic.

Deviance

$$L(\pi) = \pi^{y_1} \cdot (1-\pi)^{1-y_1} \cdot \pi^{y_2} \cdot (1-\pi)^{1-y_2} \cdots \pi^{y_{98}} \cdot (1-\pi)^{1-y_{98}}$$

$$l(\pi) = \log(L) = \sum_{i=1}^{98} \log\left(\pi^{y_i}(1-\pi)^{(1-y_i)}\right)$$

The logistic model: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{exposure}$ or $\pi = \frac{e^{\alpha+\beta\cdot \textit{exposure}}}{1+e^{\alpha+\beta\cdot \textit{exposure}}}$. estimates: $a = -2.776$ and $b = 1.9151$

Fitted values $p = \frac{e^{a+b\cdot \textit{exposure}}}{1+e^{a+B\cdot \textit{exposure}}}$.

un-exposed: $p_0 = 0.058$

expose: $p_1 = 0.2976$

Deviance

Consider the logistic regression model we are fitting here: the model with an intercept and immune status in it.

We can estimate the population fraction UTD , π, with $\frac{e^{a+b\cdot exposure}}{1+e^{a+b\cdot exposure}}$, so replace in the formulea for π the α and the β by their estmates a and b. Let's call this estimate of π, p.

These values are called **the fitted values** for the model

(Note that there are only two different fitted values in this case)

As stated above we can plug in the fitted values in the likelihood and obtain the maximum value for the likelihood for this model. Let's call this L_{model}.

Deviance

The observation self is the best fitted value

There are no fitted values closer to the observations than the observations themself.

Instead of using the fitted values in the likelihood one can also use the observations themselfs. So on every place in the likelihood were a π stands, plug in the observation itself.

Lets call this likelihood value $L_{observation}$.

Now the ratio $\frac{L_{observation}}{L_{model}}$ shows how far away the model is from the observations.

If this is for instance 10 then the likelihood with the observations as fitted values is ten times larger then the likelihood of your model, indicating the model could use some improvement.

Deviance

The deviance is twice the logarithm of this likelihood ratio: $D = 2 \cdot \ln\left(\frac{L_{observation}}{L_{model}}\right) = 2 \cdot (l_{observation} - l_{model})$.

So the deviance shows how far away the model is from the data in terms of the difference of log-likelihoods.

The deviance can be calculated for every model of interest, in our case model 0 and model 1.

Then that model is chosen which has the smallest deviance since that model is closest to the data and thus describes the data better.

Deviance

One can use the difference in deviance between model 1 and model 0 to test the hypothesis above.

The deviance for model 0 is:

$$D_0 = 2 \cdot (l_{observation} - l_0)$$

and for model 1 the deviance is :

$$D_1 = 2 \cdot (l_{observation} - l_1)$$

So the difference is
$D_0 - D_1 = 2 \cdot (l_{observation} - l_0) - 2 \cdot (l_{observation} - l_1) = 2(l_1 - l_0)$
which is exactly the likelihood ratio statistic.

Deviance

Difference in deviance gives the likelihood ratio test statistic, but the $l_{observation}$ drops out.

For this reason, the deviance is also defined as minus two times the log-likelihood : $-2 \cdot log(L)$.

This is a short version of the definition of the deviance and this one is used by R.

Deviance

for model 1 the R summary gave:

```
##
## Call:
## glm(formula = episode ~ immune, family = binomial, data = lr1)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.7726     0.5951  -4.659 3.18e-06 ***
## immune        1.9151     0.6752   2.836  0.00456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 90.424  on 97  degrees of freedom
## Residual deviance: 80.070  on 96  degrees of freedom
## AIC: 84.07
##
## Number of Fisher Scoring iterations: 5
```

The deviances are called residual deviances there.

The Null deviance is the deviance for the model with only an intercept. So for model 0 the Null deviance is the same as the residual deviance which is 90.424. The deviance for model 1 is 80.07.

Profile Likelihood

Profile likelihood

A way to look at the (log-)likelihood as a function of one parameter, e.g. a log odds ratio β, is to focus on this parameter by holding it constant.

Then for each such a constant value of this parameter the log likelihood is maximized over all other parameters.

To plot such a (log-)likelihood just put the value of the parameter (β) on the horizontal axis and the value of the maximized (log-)likelihood on the vertical axis.

This (log-)likelihood is called the **profile (log-)likelihood**

Profile likelihood

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{immune}$$

This model has two parameter: α the log-odds if immune is zero and β the log-odds ratio.

Let's focus attention on the log-odds ratio. Consider a specific value for β e.g 1.5. The logistic model then becomes: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + 1.5 \cdot \textit{immune}$.

In this model there is only one parameter left: α. Rewrite this model as a model for a fraction: $\pi = \frac{e^{\alpha+1.5\cdot \textit{exposure}}}{1+e^{\alpha+1.5\cdot \textit{exposure}}}$.

Plug this in the log-likelihood and maximize w.r.t α.

Profile likelihood

This give the maximum likelihood estimate $a = -2.46$ and a maximum value for the log-likelihood of -40.237.

So the maximum value of the log-likelihood for $\beta = 1.5$ is -40.237.

Now, take another value for β, 1.7.

The logistic model then becomes: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + 1.7 \cdot \mathit{immune}$.

Rewrite this model as a model for a fraction: $\pi = \frac{e^{\alpha+1.7\cdot \mathit{exposure}}}{1+e^{\alpha+1.7\cdot \mathit{exposure}}}$.

Plug this in the log-likelihood and maximize w.r.t α. This give the maximum likelihood estimate $a = -2.81$ and a maximum for the log-likelihood of -40.088.

So the maximum value of the log-likelihood for $\beta = 1.7$ is -40.088

Profile likelihood

And we can go on like this:

the maximum value of the log-likelihood for $\beta = 2.1$ is -40.071

Note that the value for a is changing if other values for β are taken.

Look for which value of β, the log-likelihood is at its maximum. This maximized log-likelihood was calculated for 500 values of β from 0.01 to 5.

The log-likelihood is at its maximum if β is equal to 1.92. This maximum value of the log-likelihood is -40.035. So the maximum likelihood estimate for β is $b = 1.92$. The estimated α is $a = -2.776$. These are the same as in section 4.1.

Profile likelihood

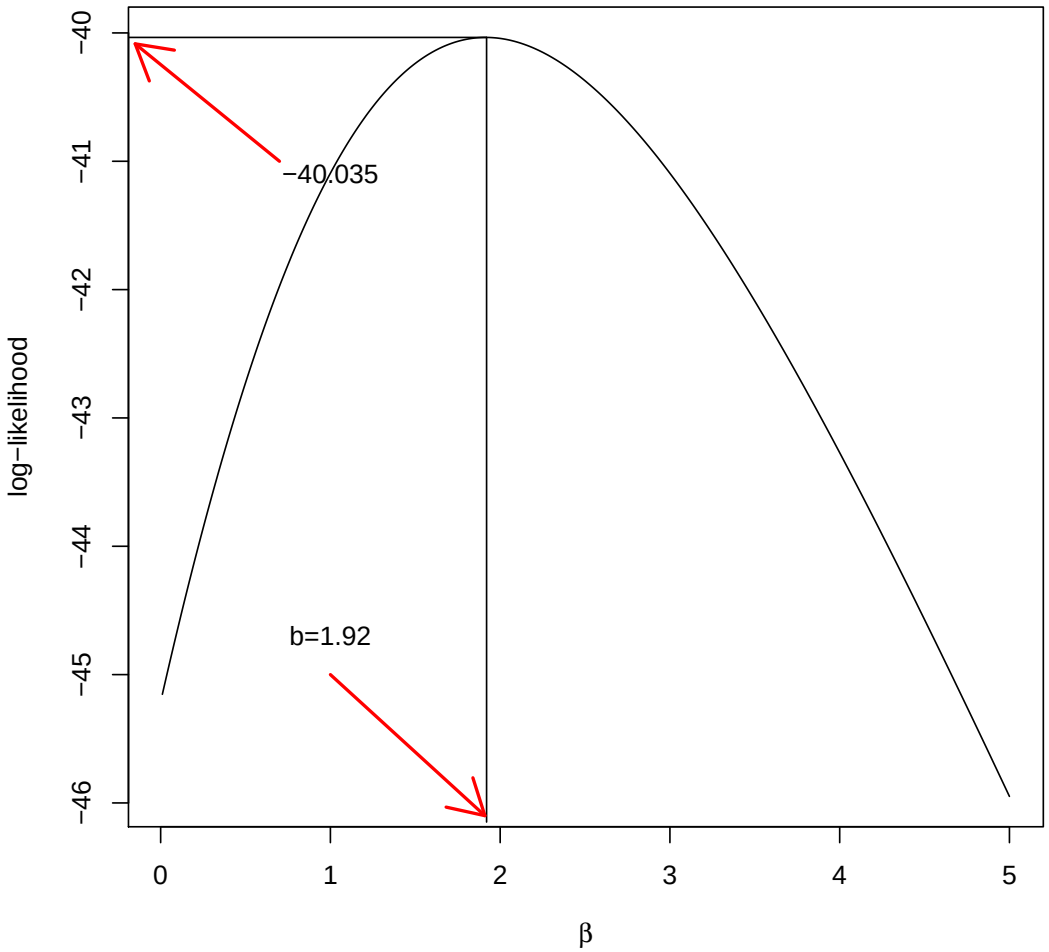


Figure 2: Profile log-Likelihood

Confidence intervals

Profile log-likelihood confidence intervals

The likelihood ratio test statistic was $2 \cdot \ln\left(\frac{L_1}{L_0}\right) = 2(l_1 - l_0)$

l_1 is the maximum value of the log-likelihood if the alternative hypothesis is true.

It is the maximum value when the log likelihood is maximized over α and β giving the maximum likelihood estimates a and b.

This maximum value can be denoted as $l(a, b)$ and is equal to -40.035.

Profile log-likelihood confidence intervals

l_0 is the maximum value of the log-likelihood if the null-hypothesis is true, that is it is the maximum value for the log-likelihood for a specific value for β.

The null-hypothesis might be $\beta = 1.5$. The log likelihood with this specific value of β is maximized over α. This maximum value of the log-likelihood can in this case be denoted as $l(a, \beta)$ and can thus be seen as the maximum value of the log-likelihood under the null-hypothesis.

So the likelihood ratio test statistic can be written as $2(l(a, b) - l(a, \beta))$.

For $l(a, b)$ two parameters needed to be estimated, for $l(a, \beta)$ only one, so the difference in number of parameters to estimate is 1.

Profile log-likelihood confidence intervals

The likelihood ratio test statistic has approximately a chi-squared distribution with one degree of freedom.

So the likelihood ratio test with a significance level of 0.05 rejects the null-hypothesis if the likelihood ratio test statistic $(2(l(a,b) - l(a,\beta))\ )$ is larger then $\chi^2_{0.95}$ which is in the case of 1 degree of freedom equal to 3.84

Profile log-likelihood confidence intervals

For all those values for β for which the likelihood ratio test statistic is smaller than 3.84, the conclusion is that the null-hypothesis is not rejected.

This is an interpretation of a 95% confidence interval: **all those values for the parameter of interest that, when put in the null-hypothesis would lead to not rejecting the null-hypothesis**. So the null-hypothesis would not be rejected for values of β for which

```
\begin{align*}
2(l(a,b)-l(a,\beta))<3.84
\end{align*} or \begin{align*}
-l(a,\beta)< \frac{3.84}{2}-l(a,b)
\end{align*} or \begin{align*}
l(a,\beta)> l(a,b)-\frac{3.84}{2}=-40.035-\frac{3.84}
\end{align*}
```

Profile log-likelihood confidence intervals

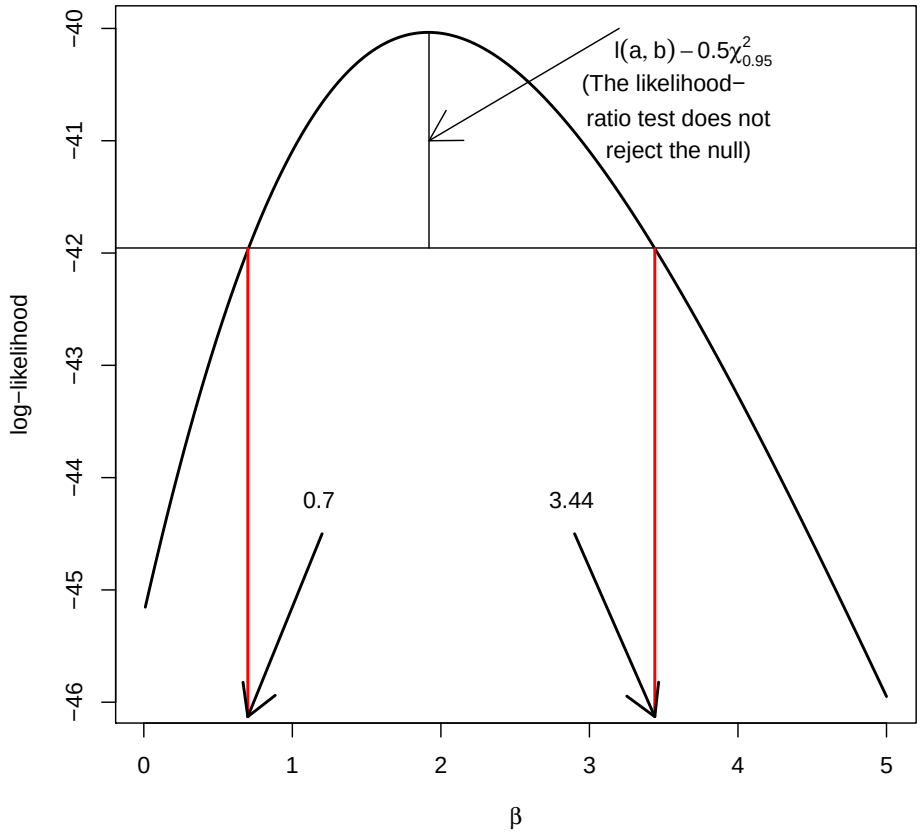


Figure 3: The 95 % profile confidence interval: those values not rejected by the likelihood ratio test are the values of β between 0.7 and 3.44

Profile log-likelihood confidence intervals

Do not reject the null: for all those values for β for which the likelihood ratio test statistic is smaller than 3.84.

```
\begin{align*}
2(l(a,b)-l(a,\beta))<3.84
\end{align*}
```

or

```
\begin{align*}
2\log\left(\frac{L(a,b)}{L(a,\beta)} \right)<3.84
\end{align*}
```

dividing by 2 ad taking anti-logs:

```
\begin{align*}
\frac{L(a,b)}{L(a,\beta)} < \exp(\frac{3.84}{2})=6.83
\end{align*}
```

Confidence intervals: Wald interval

One can also calculate a confidence interval based on the Wald test.

This is done by using $\frac{b-\beta}{se(b)}$ which has approximately a standard normal distribution, so that: $-196 \leq \frac{b-\beta}{se(b)} \leq 1.96$ or $b - 1.96se(b) \leq \beta \leq b + 1.96se(b)$.

So for the urinary tract example this becomes (see section 4.1) $1.92 - 1.96 \times 0.6752 \leq \beta \leq 1.92 + 1.96 \times 0.6752$ or $0.60 \leq \beta \leq 3.24$.

Model Checking

Model Checking

Every observation **has its own fitted value**, although for the data used here many fitted values are the same.

For instance for the first observation the fitted value can be calculated from: $p_1 = \frac{e^{a+b\cdot exposure}}{1+e^{a+b\cdot exposure}}$ with for exposure the value of the exposure variable for that individual.

The contribution for this observation to the maximum value of the likelihood is: $p_1^{y_1} \cdot (1-p_1)^{1-y_1}$.

In order to see how far the fitted value for this individual is from the observation of this individual one could calculate this likelihood contribution when the fitted value is replaced by the observation: $y_1^{y_1} \cdot (1-y_1)^{1-y_1}$.

Model Checking

The contribution for this individual to the deviance is

$$D(y_1) = 2 \cdot \ln\left(y_1^{y_1} \cdot (1-y_1)^{1-y_1}\right) - 2 \cdot \ln\left(p_1^{y_1} \cdot (1-p_1)^{1-y_1}\right).$$

If this number is large then this individual contributes a large number to the deviance and, if there are more of these individuals, to a bad fit.

If this number is small then this individual contributes a small value to the deviance implying the model fits this individuals observation.

Model Checking

The same can be calculated for individual 2 and 3 and so one.

For individual number i the contribution to the deviance is:

$$D(y_i) = 2 \cdot \ln\left(y_i^{y_i} \cdot (1-y_i)^{1-y_i}\right) - 2 \cdot \ln\left(p_i^{y_i} \cdot (1-p_i)^{1-y_i}\right)$$

After some manipulation with logarithms this becomes:

$$D(y_i) = 2\left[\ln\left(\frac{y_i}{p_i}\right)^{y_i} + \ln\left(\frac{1-y_i}{1-p_i}\right)^{1-y_i}\right]$$

Note that the sum of all these individual contributions is the deviance for the model used here.

Model Checking

The deviance residuals are defined as the square root from the individual deviance contributions multiplied with a plus or a minus sign depending on whether or not the observation are larger or smaller then the fitted values:

$$res_{dev} = sign(y_i - p_i) \cdot \sqrt{D(y_i)}$$

A large residual means that this individual contributes a large individual deviance, thus contributing to a bad fit.

Or, to put it differently: a large deviance residual means that the difference between the fitted value and the observation in terms of log-likelihoods is large for this individual

Model Checking

One can now make a plot of the deviance residuals and the fitted values. One can then check whether or not there large residuals and for these try to find out what is going on there.

Very often one sees two separate residuals patterns. This is due to the 0-1 character of the dependent variable.

The pattern in the residuals is usually one that goes down. For the zero observations: larger fitted values will give smaller negative residuals. For the observations which are one: smaller fitted values will give larger positive residuals.

Model Checking

For the data used here there are only two different fitted values: one for the low immune group which is 0.298 and one for the high immune group which is 0.059. So there are only 4 different residuals depending on whether UTD is one or zero.

The deviance residuals are:

	UTD	
Immune status	no	yes
high	-0.348	2.38
low	-0.841	1.556

So for instance all 33 individuals with low immune status and no UTD have a deviance residual of -0.841.

The 3 observation with high immune status and UTD seem to have resonable large deviance residuals. This is because the fitted value 0.059 is far away from the observation 1

More exposure variables

More Exposures

One can olso have a model with two or more exposures in it.

For instance a model with immune status and the age of the patient in it, looks like:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 \cdot \mathit{immune} + \beta_2 \cdot \mathit{age} \text{ or}$$

$$\pi = \frac{e^{\alpha+\beta_1\cdot \mathit{immune}+\beta_2\cdot \mathit{age}}}{1+e^{\alpha+\beta_1\cdot \mathit{immune}+\beta_2\cdot \mathit{age}}}.$$

In this model α is the log-odds for those patients for which immune status and age is zero.

β_1 is the log of the odds ratio for immune status holding age constant.

β_2 is the log of the odds ratio for age holding immune status constant.

More exposures

Model 0 : Model with **no exposures**: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha$. The maximal value of the likelihood is L_0. AIC=92.424 and maximum of log-likelihood is -45.212.

Model 1 : Model with **only immune status** in it: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{immune}$. The maximal value of the likelihood is L_1. AIC=84.07 and maximum of log-likelihood is -40.035.

Model 2 : Model with **only age** in it: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot \textit{age}$. The maximal value of the likelihood is L_2. AIC=91.063 and maximum of log-likelihood is -43.532.

Model 3 : Model with **both immune status and age** in it: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 \cdot \textit{immune} + \beta_2 \cdot \textit{age}$. The maximal value of the likelihood is L_3. AIC=84.427

More Exposures

Now several comparisons can be made:

Model 3 and 1 : $\frac{L_3}{L_1}$, do we need age given immune status is already in the model.

Model 3 and 2 : $\frac{L_3}{L_2}$, do we need immune status given age is already in the model.

Model 3 and 0 : $\frac{L_3}{L_0}$, do we need immune status, age or both.

Model 2 and 0 : $\frac{L_2}{L_0}$, do we need age alone.

Model 1 and 0 : $\frac{L_1}{L_0}$, do we need immune status alone.

More Exposures

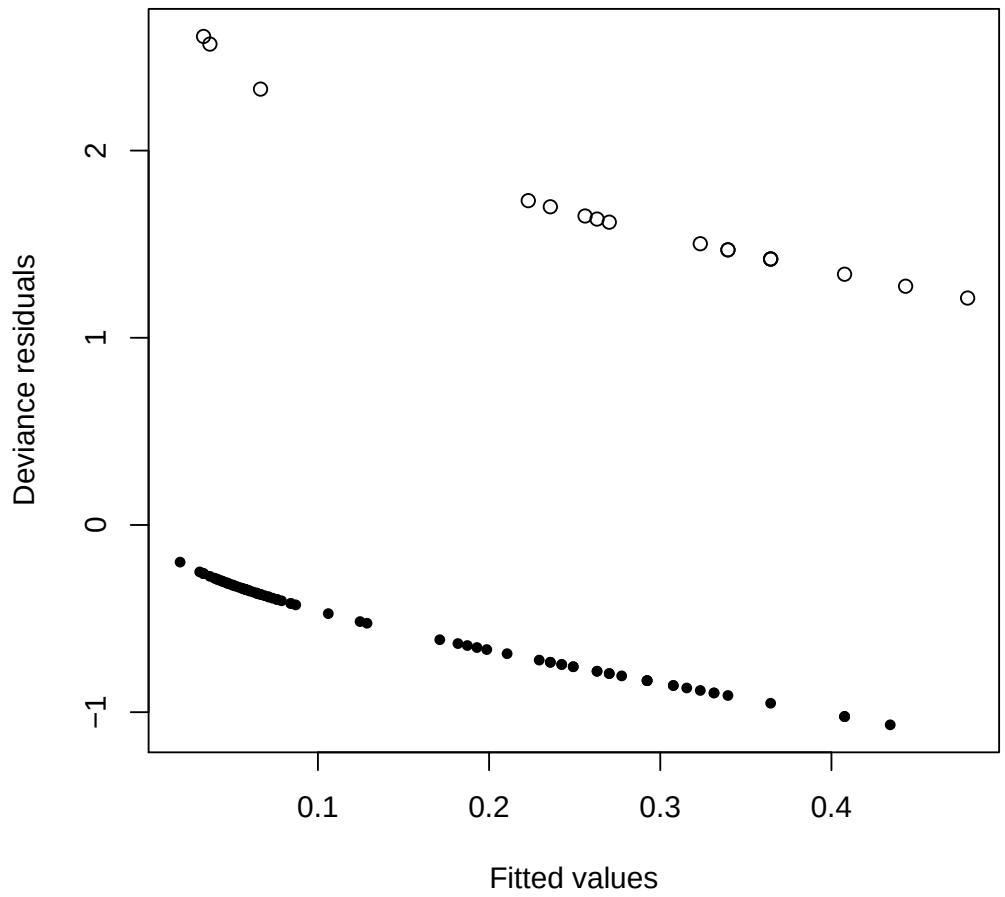
One can calculate the AIC for every model and see which model has the smallest AIC. That model is then best supported by the data. If two models have approximately the same AIC then pick the simplest one (Occam's razor).

Model 1 has the lowest AIC so this model is best supported by the data, although the difference with model 3 is not large. One can calculate for instance the likelihood ratio for model 1 compared to model 0: $\frac{L_1}{L_0} = \frac{\exp^{-40.035}}{\exp^{-45.212}} = 175.9$.

So model 1 makes the data approximately 176 times as probable as model 0!

Residual plot

The deviance residuals vs the fitted values plot for model 3, the model with the most exposure variables in it.



Predictive accuracy

Predictive accuracy

Suppose fit a model to that data with 4 independent variables. The data set on which the model is fitted, is also called the **training data set** since the model is trained with this data set to give the estimates of the parameters. We learn about the parameters.

The deviance can be used to see how well this model fits the data. In this case the short version, $-2\log(\text{Likelihood})$, is used. It measures how well the fitted model describes the dependent variable in this data set. This deviance is called the **in-sample deviance**.

It reflects the fact that it is the deviance on the data set used to estimate the parameters. It is the deviance for the fitted model in the training data.

Predictive accuracy

Consider the UTD data with immune status, age and 2 disease history variables DH1 and DH2.

The data set is considered as the training data set. The logistic model is fitted to give the estimates a, b_1, b_2, b_3 and b_4. With these estimates one can calculate for each individual the predicted or fitted values:

$$p = \frac{e^{a+b_1 imune+b_2 age+b_3 DH1+b_4 DH2}}{1+e^{a+b_1 imune+b_2 age+b_3 DH1+b_4 DH2}} \quad (1)$$

and from this one can calculate the deviance which shows how far away the fitted values are from the observations.

Predictive accuracy

An important question: **how well this model works on new data**.

Suppose, one gathers a new data set, get the parameter estimates from the training data (a, b_1, b_2, b_3 and b_4) and see how well the model predicts the dependent variable from this new data set. So for each individual in this new data set the values of the 4 independent variables used to obtain the predicted values.

The model is not fitted to the new data set but instead the parameter estimates from the training data set are used. This new data set is called the **test data set**.

Predictive accuracy

One can use the calculated predicted values to determine the deviance for this case, called **the out-of-sample deviance**.

It reflects the fact that it is the deviance on the data set not used to estimate the parameters.

It is the deviance for the predicted model in the test data set. It measures how accurate the model is in predicting new data.

Predictive accuracy

Consider 5 models:

1. a model with only a constant
2. a model with a constant and 1 variable (immune)
3. a model with a constant and 2 variables (immune+age)
4. a model with a constant and 3 variables (immune+age+DH1)
5. a model with a constant and 4 variables (immune+age+DH1+DH2)

So, the second model contains an intercept and immune, the third model contains an intercept, immune and age , the fourth model contains an intercept, immune, age and DH1 and finally the fifth model contains an intercept, immune, age, DH1 and DH2.

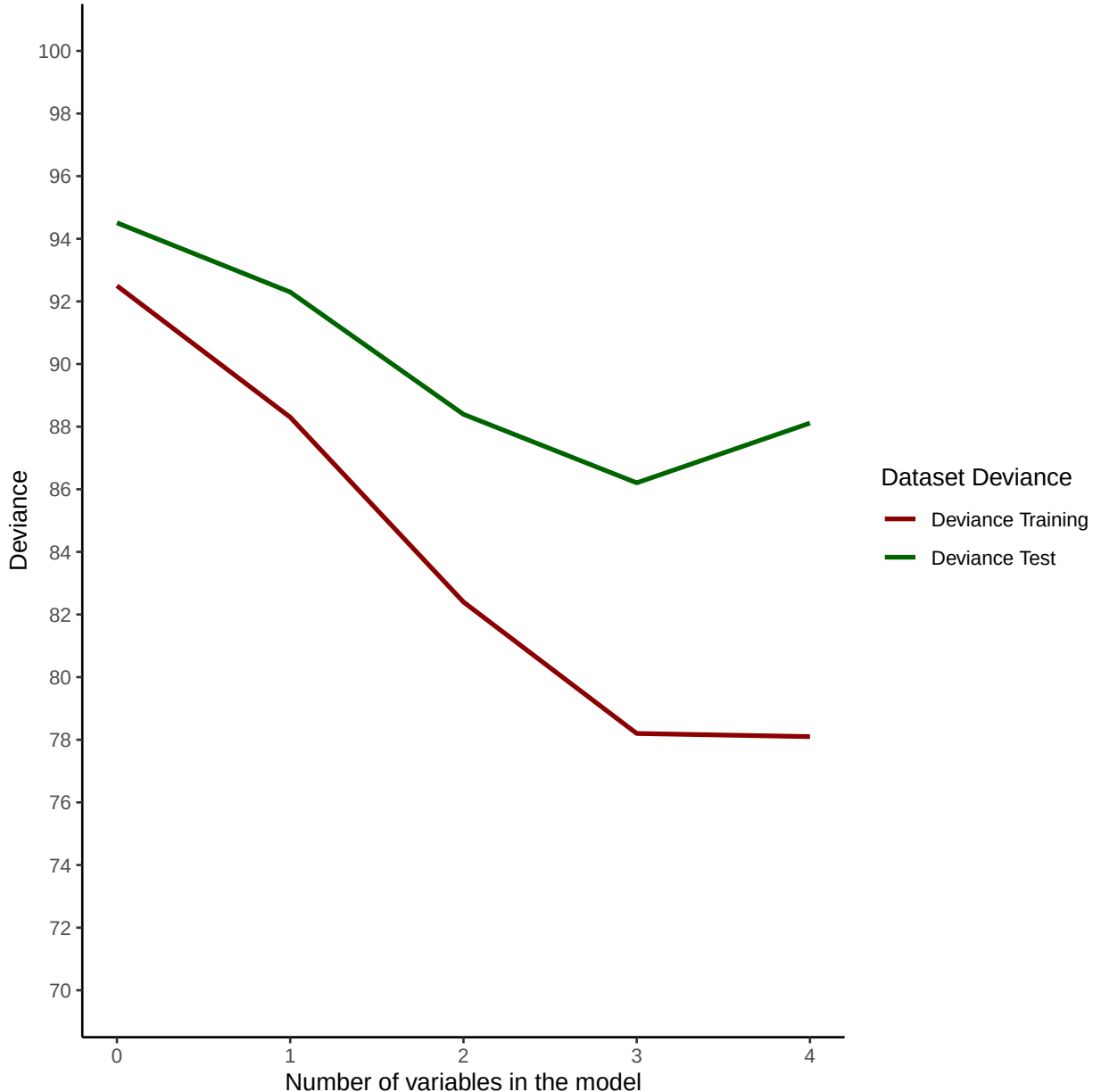
Predictive accuracy

For every one of the 5 models the training set is used to estimate the parameters and to calculate the in-sample deviance.

For every model the test data set is used with the estimated parameters to see how well the model fits this data and to calculate the out-of-sample deviance.

Then we can make a plot of these to deviance: the horizontal axis represents the number of variables and the vertical the deviances

Predictive accuracy



Predictive accuracy

The first thing to notice from figure 5 is that the out-of-sample deviance is always larger than the in-sample deviance.

When the model is fitted to the training set data the fitting procedure - the likelihood and the maximum likelihood estimating- listens carefully to the data.

So the data from the training set with its possible peculiarities has a substantial influence on the parameter estimates: **the model is adapted to the training data**.

This is not the case for the test data. These data have no influence at all on the parameter estimates. As a consequence the training data will always give a better fit as compared to the test data.

Predictive accuracy

A second point can be made by looking at the difference between the deviance for the predicted model in the test data set and the deviance for the fitted model in the training data set.

If there are no variables in the model, there will be just be one parameter in that model. The difference between out-of-sample deviance and the in-sample deviance is approximately 2.

If there is 1 variable in the model, so there are 2 parameters to estimate, the difference between out-of-sample deviance and the in-sample deviance is approximately 4.

With 2 variables and 3 parameters in the model, this difference will be approximately 6.

Predictive accuracy

In general:

$$\begin{aligned}\text{out-of-sample deviance} \approx \text{in-sample deviance} + \\ 2 \cdot \text{number of parameters} \\ = \text{AIC}\end{aligned}$$

Predictive accuracy

Predictive accuracy

In most researches there is no test data set, there is just one sample so we can not calculate the out-of-sample deviance. We then don't know the predictive accuracy of the model.

However AIC is a approximation of this out of sample deviance. If we look for models with the smallest AIC we are looking for models with the lowest out-of-sample deviance, although we do not have a test data set.

Looking for models with the low AIC values (wrt other models) is looking for models with high predictive accuracy (wrt other models)!

Predictive accuracy

Akaike's information criterion is a measure of predictive accuracy (like other information criteria).

Predictive accuracy is important for two reasons:

1. It measures the performance of a model (approximate out-of-sample deviance).
2. It can be used to compare models.

So Akaike's information criterion is an important tool to compare models if one tries to come up with relatively "simple" models that make good sense and that have good predictive performance.