Binary Data: Logistic Regression Models

Jan van den Broek

2023

1 Introduction

Imagine that one is interested in measuring something that is either present or absent. Think of measuring the occurrence of a disease, failure of a machine, the occurrence of a certain plant in an area, etc. These kinds of data are called binary data. In what follows I assume that a disease occurrence is measured. Usually one is interested in the relation between the disease and another variable.

Let's look at an example: In 1992 a study was done among 98 male HIV-patients. It was measured during approximately one year, whether or not they developed one or more episodes of urinary tract disease (UTD). (Hoepelman, A.I.M. et al, "Bacteriuria in men infected with HIV-1 is related to their immune status (CD4+ cell counts)", 1992, AIDS, volume 6, 179-194). 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. 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 of exposed (low) is coded as 1 and the unexposed (high) as 0.

1.1 The population: a model for the data generating process

The 98 HIV patients can be regarded as a sample from a larger population. Suppose we had measured the whole population. This population can be divided in two parts: the population group exposed (low immune status) and the population group unexposed (high immune status). 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 question now is: is immune status related to UTD? One can measure this with the odds ratio. 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 $1 : 2$. So the odds is the number of diseased per non-diseased (whereas a fraction is the number of diseased per individual in the population).

One can also calculate the odds in the population group unexposed: $\frac{\pi_0}{1-\pi_0}$ the proportion diseased vs non-diseased in the unexposed group. 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.

In practice of course we don't measure whole populations, so we don't know π_0, π_1 and ω. This definition of the population is useful though, as a model for the data that we are going to have in the sample.

1.2 The sample

From the population we have a sample of 98 HIV infected males. This sample can also be divided in two: the group with a low immune status (exposed) and the group with a high immune status (unexposed). Suppose the number of patients exposed is $c + d$ of which d have a UTD. The sample fraction diseased in the exposed group is then: $P_1 = \frac{d}{c+d}$. If the sample is good, the sample fraction P_1 will be a good estimator for the population fraction π_1. The same holds for the group unexposed: suppose there are $a + b$ patients

unexposed of which b patients developed a UTD, then the sample fraction in the unexposed group is $P_0 = \frac{b}{a+b}$.

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

So in the example $P_1 = \frac{14}{47} = 0.298$ and $P_0 = \frac{3}{51} = 0.059$. These are estimates for the population fractions, and the population odds ratio $\omega = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_0}{1-\pi_0}}$ can be estimated by the sample odds ratio :

$$OR = \frac{\frac{P_1}{1-P_1}}{\frac{P_0}{1-P_0}}$$

In the example the odds ratio is $OR = \frac{\frac{.298}{1-.298}}{\frac{.059}{1-.059}} = 6.77$. This means that the proportion diseased/non-diseased in the low immune group is 6.77 times higher as compared to the high immune group.

2 The logistic regression model

The logistic regression model relates the log [1] of the odds to the exposure in a linear manner. In the population this is $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot exposure$ in which *exposure* is a variable having values 0 for the unexposed and 1 for the exposed.

So in the group 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. In the exposed group 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

[1] Definition of a logarithm: ${}^b log(x) = y$ iff $b^y = x$ A log expresses a number x in a power of b. If $b = e = 2.718282\ldots$ then it is a natural logarithm denoted by ln. Some rules: $log(x \cdot y) = log(x) + log(y)$, $log(\frac{x}{y}) = log(x) - log(y)$, $log(x^y) = ylog(x)$, $ln(e^x) = x$, $e^{ln(x)} = x$ and $ln(\frac{\pi}{1-\pi}) = a$ then $\pi = \frac{e^a}{1+e^a}$

the exposed vs the unexposed.
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.

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: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot age$. 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 model for those at age x is: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha+\beta\cdot x$, for those with age it's $x+1$ $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha+\beta\cdot(x+1)$. Subtracting the model for those with age x from the model for 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 younger.

In the sample the estimates for α and β are again called a and b. Since the exposure variable is now continuous, it is not all that clear how to calculate a and b. We need a general method for this. The likelihood is this general method.

3 Likelihood

Let's forget about the exposure variable for a while. We have a sample of 98 HIV patients from a population of HIV patients. From every patient in the sample it is observed whether or not he has UTD. Let's call this measurement Y and the observation y. 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 randomly 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: $P(Y_i = y_i) = \pi^{y_i}(1-\pi)^{1-y_i}$. 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 0 or a 1 for patient 2 and ...etc. This is $P(Y_1 = y_1 \, and \, Y_2 = y_2 \, and \, \ldots \, and \, Y_{98} = y_{98}) = P(Y_1 = y_1) \cdot P(Y_2 = y_2) \cdots P(Y_{98} = y_{98})$. This equality holds because the observations are independent and thus we can multiply the separate probabilities. This probability of the observed data is called the likelihood, denoted by $L()$. It says how probable the observed data, the zeroes and the ones observed from the 98 patients, is. Although data is observed and the probabilities are gone, one might still wonder about what the probability was of observing the data.

We can now plug in the probability of each observation:

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

$\sum_{i=1}^{98}$ is the sum of the observations, the sum of the zeroes and ones we saw in the sample. It just counts the number of ones (number of UTD cases) in the observed data. So in the example it has a value of 17. Now, $L(\pi)$ gives the probability of the observed data, the probability of the 17 ones and 81 zeroes. The likelihood above is called the binomial or Bernoulli likelihood. This likelihood depends on π, the population fraction UTD patients, which means that the likelihood is seen as a function of π. Thus the likelihood is the probability of the observed data seen as a function the population fraction π.

So, the likelihood is the probability of observing the 17 ones and 81 zeroes and that probability depends on π. What does this data tell us about π? What would be plausible values for π ? Would $\pi = 0.7$ be plausible? No! In that case one would expect approximately 70 ones and observing 17 ones would have low probability and thus a low likelihood. The likelihood of $\pi = 0.7$ is low since in that case the probability of observing 17 ones is very low. Does the data tells us that $\pi = 0.2$ might be plausible? That is a good possibility because in that case we would expect about 20 ones which is close. So observing 17 ones when $\pi = .2$ should have high probability and thus $\pi = .2$ has a high likelihood.

Apparently the process we are studying is such, that the observed data has a high probability of occurring, so the likelihood of the observed data should be high. One can calculate the likelihood for every value of π between 0 and 1. 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).

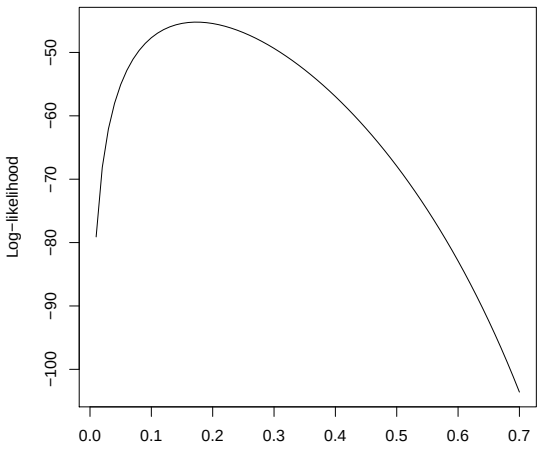

Figure 1: Log-likelihood for different values of π

Remember, the probability of observing a zero or one for person number i is $\pi^{y_i}(1-\pi)^{1-y_i}$, and the likelihood for independent data is the product of these probabilities. Instead of maximizing the likelihood one usually takes the log of the likelihood (see fig. 1). The log-likelihood is the sum of the log-probabilities:

$$\begin{aligned} l(\pi) = \ln[L(\pi)] &= \sum_{i=1}^{98} \ln\left[\pi^{y_i}(1-\pi)^{1-y_i}\right] \\ &= \sum_{i=1}^{98} \left[y_i \ln(\pi) + (1-y_i)\ln(1-\pi)\right] \end{aligned}$$

Maximizing the log-likelihood is the same as maximizing the 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. If one makes a plot of the log-likelihood for different values of π, like the one in figure 1, 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.

A measure for the peakedness of the log-likelihood is the second derivative at its 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 reverse 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.

Fortunately one does not have to do the calculations oneself, there are several computer programs available. One of the best is the open source program R which can be downloaded from http://www.r-project.org/.

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.

4 Putting it together.

4.1 Evidence in the data

The logistic model was given by: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot exposure$. This can also be written as : $\pi = \frac{e^{\alpha+\beta\cdot exposure}}{1+e^{\alpha+\beta\cdot exposure}}$. We want to estimate α and β. This can

be done by using the log-likelihood:

$$\begin{aligned} l(\pi) = \ln[L(\pi)] &= \sum_{i=1}^{98} \ln\left[\pi^{y_i}(1-\pi)^{1-y_i}\right] \\ &= \sum_{i=1}^{98} \left[y_i \ln(\pi) + (1-y_i)\ln(1-\pi)\right] \end{aligned}$$

For π we take $\frac{\exp^{\alpha+\beta\cdot exposure}}{1+\exp^{\alpha+\beta\cdot exposure}}$. 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. Below is a part of the R output for this model. In the column labeled Estimate you find a and b, the row starting with "intercept" gives a and the row starting with "immune" gives b, the log-odds ratio:

```
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
---

(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
```

Now, is there evidence in the data to state that immune status is related to UTD? If there is no relation between immune status and UTD then β in the logistic model $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot exposure$ is zero. Then the model states that the log-odds is constant and does not depend on immune status. This is model 0. 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:

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 for α, the second derivative gives the standard error. The log-likelihood has its maximum for a. The maximum of the

log-likelihood, l_0, is obtained by plugging in the value a for α. This is also denoted as $l(a)$ which shows that a is plugged in. The maximum for the likelihood then is $L_0 = e^{l_0}$. L_0(or $L(a)$) gives the probability of the data using model 0. More precise: L_0 is the likelihood of model zero meaning that if the estimated model zero was the generator of the data, L_0 gives the probability of the data.

Model 1 : The model is $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot exposure$ or $\pi = \frac{e^{\alpha+\beta\cdot exposure}}{1+e^{\alpha+\beta\cdot exposure}}$. Plug in this value for π in the log-likelihood. The first derivatives give the maximum likelihood estimators a for α and b for β, the second derivative gives the standard errors for these estimates. 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 model 1. More precise: L_1 is the likelihood of model one meaning that if the estimated model one was the generator of the data, L_1 gives the probability of the data.

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. Usually a model with a lot of exposure variables in it is better than a model with only a few. To take this number of exposure variables into account one can use 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. 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. This known as the principle of Occam's Razor (Wiliam Occam, 1300-1349) or as the principle of parsimony. This principle states roughly that one should keep things as

simple as possible. As a rough guide a difference in AIC's is considered small if it is smaller than 2.

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 the data as compared to the other models that are fitted.

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

The output for model 0 is:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5612     0.2668  -5.853 4.84e-09
---

(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
```

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 $\frac{L_1}{L_0} = 177.15$. So model 1 makes the observed data about 177 time as likely as model 0.

4.2 Hypothesis testing.

4.2.1 Likelihood ratio test.

In the literature usually p-values are calculated. In that case one can look at $2 \cdot \ln\left(\frac{L_1}{L_0}\right) = 2(l_1 - l_0)$ because one can show that, if the number of observations is large, this has a chi-squared distribution with *df* degrees of freedom where *df* is the difference of the number of parameters between the models. This p-value is used to test the hypothesis: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$. This is called the likelihood ratio test. This test compares one model to another, more general model.

4.2.2 Wald test.

Another way to test the hypothesis mentioned above is by the so-called Wald test. One can calculate this test statistic by dividing the estimate 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 preferred over the Wald test since in most applications the chi-square approximation to the likelihood ratio test statistic is better then the normal approximation for the Wald statistic.

4.2.3 Deviance.

Another way to compare models is with the so-called 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 formula for π the α and the β by their estimates a and b. Let's call this estimate of π, p. These values are called the fitted values for the model we are using here. Note that there are only two different fitted values in this case since the persons in the high immune status group all have exposure value zero and thus the same fitted values. The same is true for all the individuals in the low immune group: they all have the same exposure value one, and thus they all have the same fitted values. 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}.

Every observation has a fitted value, for every individual we look at the exposure value for this individual and calculate the individual's fitted value as: $\frac{e^{a+b\cdot exposure}}{1+e^{a+b\cdot exposure}}$. Instead of using the fitted values in the likelihood one can also use the observations themselves. So on every place in the likelihood were a π stands, plug in the observation itself. The argument for this is that the observation itself is the best fitted value in the sense that there are no fitted values closer to the observations than the observations themselves. Let's 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 than the likelihood of your model, indicating the model could use some improvement. 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.

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.

4.3 Testing with the urinary tract data

For model 1 and model 0 the deviances are in the output above. They 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. The outcome for the likelihood ratio test then is $(D_0 - D_1) = 10.354$. The degrees of freedom for this test is 1 since the difference in number of parameters between the models is 1. In order to calculate the p-value one has to use the chi-square distribution with 1 degree of freedom. This gives a p-value of 0.0013. When taking a significance level of 0.05 the null hypothesis is rejected.

5 Confidence Intervals

5.1 Profile likelihood

In section 7 the logistic regression model is extended to the case where there are more exposure variables in the model. When there is more than one parameter in the model the (log-) likelihood might not be that easy to understand. It is always easier to describe the (log-)likelihood with one parameter. It might also be that the interest is in just one parameter. The interest of the research might for example be in the odds ratio because the aim of the study was to find out whether the exposure variable is an risk factor. Also one might want to look at the effect of one variable at the time as with calculating confidence intervals.

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

To see how this method works consider the logistic regression model:

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

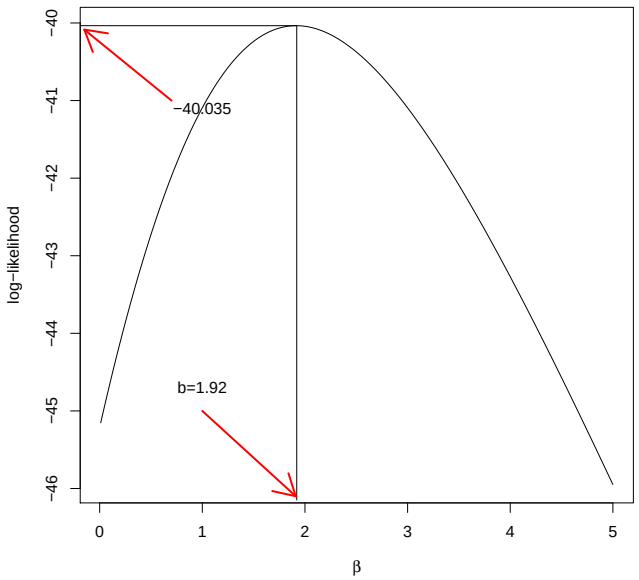

Figure 2: The profile likelihood which is maximized at $b = 1.92$.

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: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + 1.5 \cdot immune$. In this model there is only one parameter left: α. Rewrite this model as a model for a fraction: $\pi = \frac{e^{\alpha+1.5 \cdot exposure}}{1+e^{\alpha+1.5 \cdot exposure}}$. Plug this in the log-likelihood and maximize w.r.t α. 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: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + 1.7 \cdot immune$. Rewrite this model as a model for a fraction: $\pi = \frac{e^{\alpha+1.7 \cdot exposure}}{1+e^{\alpha+1.7 \cdot exposure}}$. 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

And we can go on like this: take $\beta = 2.1$.The logistic model then becomes: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + 2.1 \cdot immune$. Rewrite this model as a model for a fraction: $\pi = \frac{e^{\alpha+2.1 \cdot exposure}}{1+e^{\alpha+2.1 \cdot exposure}}$. Plug this in the log-likelihood and maximize w.r.t α. This give the maximum likelihood estimate $a = -2.92$ and a maximum for

the log-likelihood of -40.071. So 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.

So for all possible values of β we can calculate the maximum value of the log-likelihood. We then can 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 and was plotted in figure 2. 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.

5.2 Profile confidence intervals

5.2.1 95 % two-sided profile-likelihood confidence interval

This profile likelihood is used to determine confidence intervals that are based on the likelihood ratio test. 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.

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. 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.

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

$$2(l(a,b) - l(a,\beta)) < 3.84$$

or

$$-l(a,\beta) < \frac{3.84}{2} - l(a,b)$$

or

$$l(a,\beta) > l(a,b) - \frac{3.84}{2} = -40.035 - \frac{3.84}{2} = -41.955$$

This line is drawn in figure 3. Now one can look up the values for β for which this holds (see figure 3) by using the calculations for the profile likelihood above. So a 95% two sided profile confidence interval for β is $0.7 \leq \beta \leq 3.44$.

An other interpretation for a 95% confidence interval is: The probability that this interval contains the population value of the parameter is 0.95.

5.2.2 Direct likelihood interpretation

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.

$$2(l(a,b) - l(a,\beta)) < 3.84$$

or

$$2\log\left(\frac{L(a,b)}{L(a,\beta)}\right) < 3.84$$

dividing by 2 and taking anti-logs:

$$\frac{L(a,b)}{L(a,\beta)} < \exp(\frac{3.84}{2}) = 6.83 \approx 7$$

So the interval also has the important interpretation:
All those values for β for which the likelihood ratio w.r.t. the maximum likelihood estimate, is smaller than 7.
This is another way of saying what values of β are consistent with the data without referring to a significance level. Those values for β are consistent with the data, for which the likelihood of the maximum likelihood value is no more than 7 times better.

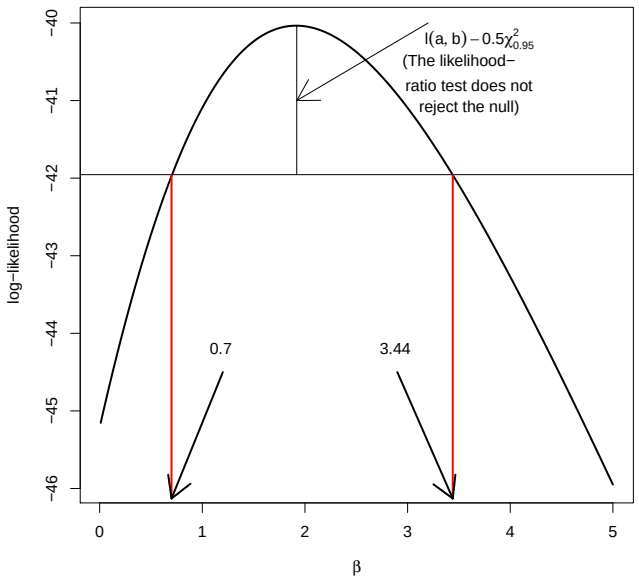

Figure 3: The 95 % profile confidence interval: those values not rejected by the likelihood ratio test are the values of β between 0.7 and 3.44

5.3 Wald confidence intervals

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 (se section 4.1) $1.92 - 1.96 \times 0.6752 \leq \beta \leq 1.92 + 1.96 \times 0.6752$ or $0.60 \leq \beta \leq 3.24$.

For the same reason the likelihood ratio test is preferred over the Wald test, is the above profile interval (based on the likelihood ratio test) preferred over the confidence interval based on the Wald test.

6 Model checking

As said, 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: $y_1^{y_1} \cdot (1-y_1)^{1-y_1}$. 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 amount is large then this individual contributes a large amount 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 amount to the deviance, implying that the model fits this individual's observation. The same can be calculated for individuals 2, 3 and so on. For individual i the contribution to the deviance is:

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

which is twice the difference in log-likelihood contributions. 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. The deviance residuals are defined as the square root of the individual deviance contributions multiplied with a plus or a minus sign, depending on whether or not the observation are larger or smaller than 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-likelihood contributions is large for this individual.

One can now make a plot of the deviance residuals and the fitted values. One can then check whether or not there are 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 more extreme negative residuals. For the observations which are one: smaller fitted values will give larger positive residuals.

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 reasonably large deviance residuals. This is because the fitted value 0.059 is far away from the observation 1.

7 More exposure variables

One can also 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: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 \cdot immune + \beta_2 \cdot age$ or $\pi = \frac{e^{\alpha+\beta_1 \cdot immune+\beta_2 \cdot age}}{1+e^{\alpha+\beta_1 \cdot immune+\beta_2 \cdot age}}$. 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. One can now fit four models:

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

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

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

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

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.

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).

To perform the likelihood ratio test for comparing model 3 with model 0 for instance, calculate $2 \cdot \ln \frac{L_3}{L_0} = 2(l_3 - l_0)$ and then calculate the p-value using a chi-squared distribution with 2 degrees of freedom.
The R output for these models is:

Model 0 : Model with no exposures: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha$.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.5612     0.2668  -5.853 4.84e-09
---

(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
```

Model 1 : Model with only immune status in it: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot immune$.

```
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
---

(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
```

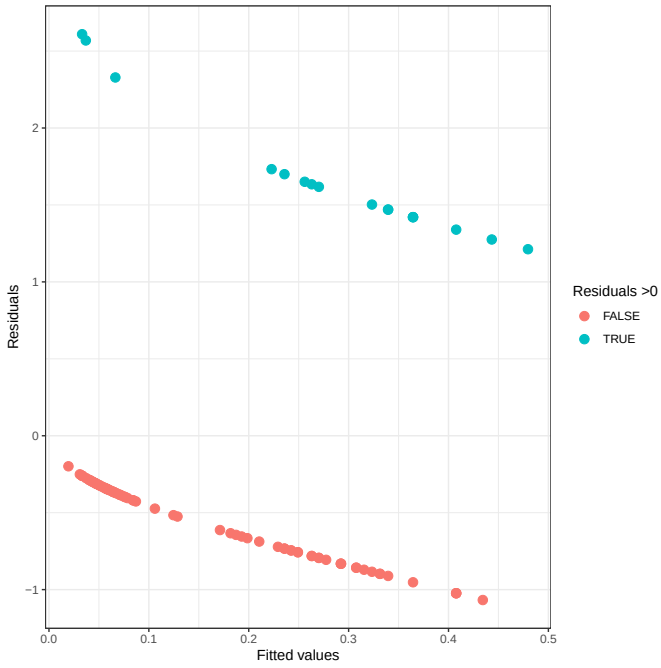

Figure 4: Deviance residuals

Model 2 : Model with only age in it: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta \cdot age$.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.55807    1.18849  -2.994  0.00276
age          0.04821    0.02694   1.790  0.07346
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 90.424  on 97  degrees of freedom
Residual deviance: 87.063  on 96  degrees of freedom
AIC: 91.063

Number of Fisher Scoring iterations: 4
```

Model 3 : Model with both immune status and age in it: $\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 \cdot immune + \beta_2 \cdot age$.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.21067    1.32569  -3.176  0.00149
immune       1.79460    0.68254   2.629  0.00856
age          0.03647    0.02876   1.268  0.20474
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 90.424  on 97  degrees of freedom
Residual deviance: 78.427  on 95  degrees of freedom
AIC: 84.427

Number of Fisher Scoring iterations: 5
```

Model 1 has the lowest AIC so this model is best supported by the data, although the difference with model 3 is not large. The maximum log-likelihood values are -45.212 for model 0, -40.035 for model 1, -43.532 for model 2 and -39.2135 for model 3. From this output one can calculate for instance the likelihood ratio for model 1 compared to model 0: $\frac{L_1}{L_0} = \frac{\exp^{-40.035}}{\exp^{-45.212}} = 177.15$. So model 1 makes the data approximately 177 times as probable as model 0! The deviance residuals vs the fitted values plot for model 3, the model with the most exposure variables in it, is in figure 4.

8 Predictive accuracy

Suppose we have a data set like the ones we discussed so far, and suppose we 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 a, b_1, b_2, b_3 and b_4. 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.

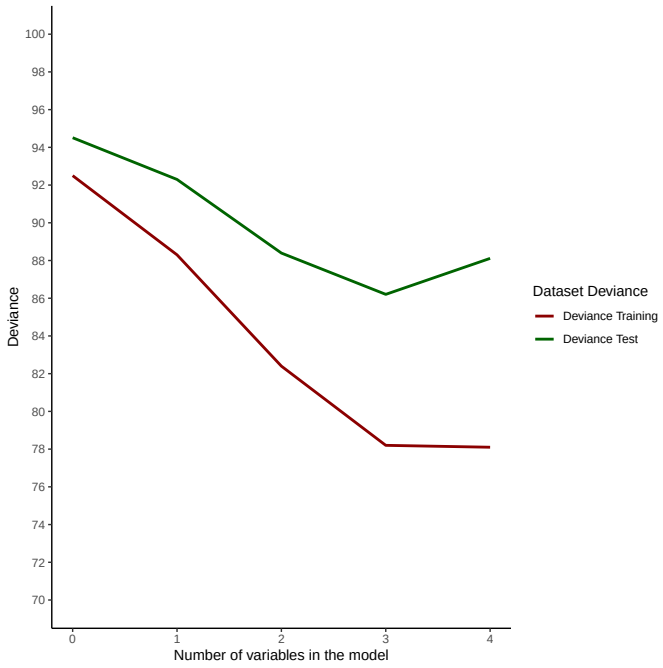

Figure 5: Deviance from the training data and the test data

As an example 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:

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

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

In most cases an important question for the researcher is 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 in (1) 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. 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. Consider 5 models:

1. a model with only a constant
2. a model with a constant and 1 variable
3. a model with a constant and 2 variables
4. a model with a constant and 3 variables
5. a model with a constant and 4 variables

For now, we are interested in the number of variables in the model, not which variables. (In this case, for a model with 1 variable there are 4 possibilities.) 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.

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 (see figure 5).

The first thing to notice from figure 5 is that the out-of-sample deviance is always larger than the in-sample deviance. This is because, 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. 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. In general:

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

And that is magic:
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)!

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.

9 An application of logistic regression: Ordinal data[2]

9.1 Introduction

In the years 1995 until 1998 a research was done among 1243 Dalmatian pups. It was determined whether or not they were deaf in at least one ear. The research question was if deafness was related to pigmentation. In order to answer this question it was measured whether or not there were many spots on the skin, whether or not the pup had a spot on the head and whether or not the pup had blue eyes.

Here, as an example, we look at the spot measurement and will treat it as the variable of interest, that is to say as the dependent variable. The

[2]This section is optional

number of spots was measured in three classes: light, moderate and heavy. A research question might be: are the number of spots different between males and females. First the data:

	Spots			
Gender	light	moderate	severe	total
male	86	458	75	619
female	105	468	51	624
total	191	926	126	1243

This table is called a frequency table. The six numbers in the table are the frequencies. These frequencies count the number of males with light spots (86), the number of males with moderate spots etc.

The dependent variable in this example is the number of spots divided in three classes. It's a variable with three possible outcomes. Such a variable is called a categorical variable. Binary variables are also categorical variables, but with only two possible outcomes. If the categories of a categorical variable cannot be ordered then it is called a nominal variable. If these categories can be ordered then the variable is called ordinal. So the variable spots in this research is an ordinal variable: light is less than moderate which in turn is less than heavy.

9.2 The continuation ratio model

One way to take the ordinal character into account, is to split the table into sub-tables. Firstly, attention is focused on the two lowest classes, the light and moderate classes. Then one can describe how many animals in the light and moderate class are in the higher of these two classes (moderate). This is done for both males and females. So, for the males, if one looks at the light and moderate number of spots (the two lowest outcome classes of the spot variable), 458 of those males are in the higher class (moderate) and 86 are in the lower class.

Then the light and moderate class are added up and compared to the severe class, also separately for males and females. How many animals are now in the higher class (severe) and how many are not. For the males there are 75 in the higher class (severe) and $86 + 458 = 544$ are not (light + moderate class).

One can also argue from the highest class to the lower classes: first it is determined how many animals there are in the highest class (severe) as compared to the rest (light+moderate). Then conditionally on all animals in the rest of the table (light+moderate), how many are there in the higher

class (moderate) as compared to the other class (light).

So one can make the following table.

Part of the table	gender	number in higher class	number in lower class
lower	male	458	86
lower	female	468	105
upper	male	75	544
upper	female	51	573

As can be seen from the table, for every dog it is determined whether or not the dog is in the higher class. This is binary data. The table contains the number of dogs in the higher class as compared to the other class. So it is grouped binary data and one can use the logistic regression model. The splitting of the original table and then using a logistic regression is known as the continuation ratio model. In this way the probability of getting a higher spot-class is modeled, first in the lower part of the table and then in the higher part. So there is a clear direction in the analysis reflecting the ordinal character of the spot variable.

One can fit the logistic regression model:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 PART\_TABLE + \beta_2 GENDER + \beta_3 PART\_TABLE \cdot GENDER$$

To see if there is an interaction a drop1() can be used on the fit of the model:

```
Df  Deviance     AIC
<none>                                      0.0000      32.142
factor(part.table):factor(gender)       1   2.8763      33.018
```

So one can not show that there is an interaction with these data. The gender effect then is taken to be the same for moderate versus light (lower part) as compared to severe versus not severe (upper part):

```
Deviance Residuals:
1        2        3        4
-0.8235   0.7279   0.8675  -0.9570

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)           1.7704     0.1059  16.720  < 2e-16 ***
factor(part.table)1  -3.8602     0.1236 -31.223  < 2e-16 ***
factor(gender)1      -0.3538     0.1222  -2.896  0.00378 **
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1511.8333  on 3  degrees of freedom
Residual deviance:    2.8763  on 1  degrees of freedom
AIC: 33.018

Number of Fisher Scoring iterations: 3
```

There seems to be a gender effect: the log oddsratio for females versus males is -0.3538 indicating that females have fewer spots.

10 The data sets

Lowbirth.dat

Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight. The variables identified in the code sheet given in the table have been shown to be associated with low birth weight in the obstetrical literature. The goal of the current study was to ascertain if these variables were important in the population being served by the medical center where the data were collected. So one wants to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of whom had low birth weight babies and 130 of whom had normal birth weight babies. Four variables which were thought to be of importance were age, weight of the subject at her last menstrual period, race, and the number of physician visits during the first trimester of pregnancy.

```
LIST OF VARIABLES:

Variable                                              Abbreviation
-------------------------------------------------------------
```

```
Identification Code                                          id
Low Birth Weight (0 = Birth Weight >= 2500g,                 lowb
             1 = Birth Weight < 2500g)
Age of the Mother in Years                                   age
Weight in Pounds at the Last Menstrual Period                lwt
Race (1 = White, 2 = Black, 3 = Other)                       race
Smoking Status During Pregnancy (1 = Yes, 0 = No)            smoke
History of Premature Labor (0 = None,1 = One,etc.)           ptl
History of Hypertension (1 = Yes, 0 = No)                    ht
Presence of Uterine Irritability (1 = Yes, 0 = No)           ui
Number of Physician Visits During the First                  ftv
  Trimester ftv (0 = None, 1 = One, 2 = Two, etc.)
Birth Weight in Grams                                        bwt
-------------------------------------------------------------

References:

1. Hosmer and Lemeshow, \\
Applied Logistic Regression, Wiley, (1989).
```

pdd.csv

The Proventricular Dilatation Disease is causing the death of captive and free-ranging birds on a global scale. The disease, well-known since the end of the 70's, appeared for the first time in macaws and this led to the so-called Macaw Wasting Disease. Its progressive appearance in other Psitacines and in other groups of birds has generated a variety of other synonyms. PDD can present subacute, acute and chronic stages. Most diseased birds die within several months to a year after developing clinical signs. The most common of these include depression, weight loss, regurgitation and the occurrence of undigested food in the faeces. It can also affect the nervous system provoking, in this case, a lack of co-ordination and abnormal head movements. Only in the Psitaciform order, Proventricular Dilatation Disease has been reported in more than 50 species, including such disparate groups as cockatoos, Agapornis spp., amazons and macaws. It poses a serious threat, not only for aviculture, but also for the management of threatened species in the wild.

During the last six years, the investigations have established in an incontrovertible way, that Proventricular Dilatation Disease is caused by a virus, although the polemic among the investigators about the concrete virus in question has not finished yet. Another important

issue is the way in which this virus, which according to some authors is quite similar to the paramyxovirus that affects chickens (Gallus gallus), started to affect the macaws population.

In a large Dutch study it was determined from 308 parrots whether or not they had PDD. Besides this it was measured whether or not the parrots were coming from the NOP, a kind of center for diseased parrots. Also the gender was measured, as well as whether or not the parrot was suffering from arteriosclerosis. The research question was which of these were risk factors for PDD.

LIST OF VARIABLES:

Variable	Abbreviation
Dutch medical center for parrots	nop (0=n0,1=yes)
Gender	gender (0=male,1=female)
Wether or not arteriosclerosis	arterio (0=no,1=yes)
Number of pdd parrots	pdd
Total number of parrots	n

osteochon.csv

From each of 30 stallions a number of daughters were sampled. It was measured for these mares, all approximately 3 years old, whether or not they had osteochondroses in one of their joints. The purpose of the study was to find risk factors for osteochondrosis. The data set contains the following variables:

LIST OF VARIABLES:

Variable	Abbreviation
The father of the mare	father
The amount of extra food given to the mare	food (1=01-kg,2=2-3kg,3=>3kg)
The kind of ground on which the mare was raised	ground (2=sand,1= other)
Height at withers	height
Osteochondrosis	oc

dalmatian.csv

In the years 1995 until 1998 a research was done among 1243 Dalmatian pups. It was determined whether or not they were deaf in at least one ear. The research question was if deafness was related to pigmentation. In order to answer this it was measured whether or not there were many spots on the skin, whether or not the pup had a spot on the head and whether or not the pup had blue eyes. Besides this, one wants to determine if there was heritability involved. In order to look at this the family history score was determined. This is a method to cope with litter-effects (heritability): of every pup it was determined how many

brothers and sisters were deaf. Call this number m. Then from the whole data set the fraction of dogs who are deaf is determined. This fraction is multiplied by $littersize-1$. This the expected number of deaf brothers and sisters when there are no differences between litters. The family history score is now defined as $fhs = m - fraction * (litersize - 1)$

```
LIST OF VARIABLES:
Variable                                    Abbreviation
-----------------------------------------------------------------
Wether or not the pup is deaf               deaf(0=no,1=yes)
Amount of spots on the skin                 spot(1=light,2=moderate,
                                                3=heavy)
Whether or not the pup has blue eyes        blueeye (0=no,1=yes)
Whether or not the pup has a spot           headspot(0=n0,1=yes)
            on the head
Gender                                      gender(0=male,1=female)
Family history score                        fhs
-----------------------------------------------------------------
```

Exercises

R commands First the data file needs to be read in. The data is in episode.txt. It is a text file. The first lines are:

```
episode followup  cd4  age
    0   24  125   35
    0   12   50   34
    1    6   30   37
    0    6   80   36
    0    3  170   35
    0    6   95   26
    0    4   35   44
    0    3   50   42
    2    6   25   64
```

The first line contains the column names. This can be read in with the command read.table() or with the rstudio menu. This results in a data frame object. A data frame contains several columns of data. These columns can be of different type: they can be a grouping variable, a continuous variable or a variable containing characters. We will call the data frame ep:

```
lr1  <- read.table(file="episode.txt", header=TRUE)
```

The header=TRUE states that the first line contains the column names.
In this file the columns are separated by spaces. Often a different separator is used, for instance a comma, called a csv (comma separated value) file. Then one can use :

```
read.table(file="episode.txt", header=TRUE, sep=",")
```

To see what the names of the columns are: names(ep). To look at a specific column, e.g. cd4: lr1$cd4. So if you want to use a variable from a data frame, use the name of that data frame, then a dollar sign' followed by the column name.
In the cd4 column the cd4 values of the patients are stored. From this we need to make a new column which has a 1 if the cd4 value is smaller than 200 and a zero otherwise. To do this use $lr1\$cd4 < 200$. If you do this you see that you get a column with TRUE and FALSE in it. To make this a column with 0 and 1, multiply the statement with 1 and put the result in a column called immune: lr1$immune <- $1 * (em\$cd4 < 200)$.
Now we are ready to fit the models. If there is an exposure in the data file that is a group variable, coded other than 0-1, then you should tell this to

R by using the function factor(). So factor(group) tells R that group is not a numeric variable but that its values should be used as group labels. To fit the logistic regression model with immune as an exposure variable use

```
fit <- glm(episode <- immune,family=binomial,data=lr1)
```

For every patient also the follow-up time is recorded. It might sometimes be a good idea to model the odds per month follow-up, thus to use the model

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

Rewriting gives :

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

, that is followup is in the model without a coefficient attached to it. To achieve this in R the term **offset** is used:

```
fit <- glm(episode~immune+offset(log(followup)),
       family=binomial,data=lr1)
```

To fit the logistic regression model 3 use:

```
epifit.3 <- glm(episode~immune+age,family=binomial,data=lr1)
```

epifit.3 will contain the result. It will be an object of glm-type because you used glm to create it. To see what is in it use names(epifit.3) and if you want to see something specific use e.g. epifit.3$coefficients. To get the tables from the text : summary(epifit.3). Profile confidence intervals can be obtained by: confint(epifit.3) and the wald intervals by confint.default(epifit.3). The deviance residuals can be obtained by residuals(model3) and the fitted values by fitted.values(epifit.33).
Now you can leave out 1 variable from the model and look at the differences in AIC's. You can do this by fitting all the different models and then comparing them. You can also use the function drop1(fit). This function looks at the terms in fit, then leaves the terms out one by one and calculates for every term left out the AIC of the model. The command drop1(fit, test="Chisq") calculates the likelihood ratio test for every term left out. Then you can fit a model by leaving out the variable with the least influence and then start the procedure all over again using this last model as a starting point, etc. For the AIC this can also be done automatically:

```
# backward:
step(epifit.3,direction = "backward")
#forward
step(epifit.0,direction = "forward",scope = ~immune+age)
```

where epifit.3 is the fit for model3 and epifit.0 for model0.
To make the residual plot:

```
lr.res <- data.frame(lr1,res=residuals(epifit.3),fit=fitted(epifit.3))
ggplot2::ggplot(lr.res,aes(x=fit,y=res))+
          geom_point()+theme_bw()
```

Exercises 1

1. episode.txt

   (a) Reproduce the output for the models for episode from the text. (First read in the data from episode.txt)

   (b) This datafile also contains the variable followup. This is the time a patient is in the study. Fit a logistic regression model with the log of the follow up time as an exposure variable and compare this model with the one that only contains an intercept using the AIC.

   (c) Fit a model with log(followup) as an offset:

   ```
   glm(episode <- offset(log(followup)),
       family=binomial,data=ep)
   ```

   Fit this model and compare the AIC with the former one.

   (d) Write down the logistic regression model for the model with the offset. Give an interpretation of this model.

2. lowbirth.dat

   (a) Read in the dataset lowbirth.dat

   (b) Fit three models, one with exposure age, one with exposure smoke and one with exposure ht.

   (c) For all 3 models give an interpretation of the estimate of β for the specific exposure at hand.

   (d) Compare all 3 models to the model with only an intercept in it using AIC. Calculate for each the likelihood ratio and give this an interpretation.

   (e) Also compare the 3 models with each other by comparing AIC. Also calculate likelihood ratios.

   (f) Which model fits the data best? Give your argumentation.

3. pdd.csv

   (a) Read in the data file pdd.csv
   Note that this file is different from the other files. It doesn't have a 0-1 variable in it. Instead the data is grouped. The column pdd contains the number of parrots with PDD and the column n contains the total number of parrots. So $n - pdd$ is the number of parrots without PDD. As an example, line number 6 states: there

were 16 male parrots, from the NOP, having no arteriosclerosis, and 5 of these had PDD. To fit a logistic regression model the dependent variable is not just one column. It is a matrix containing the number of PDD-cases and the number without PDD. So the dependent variable is : cbind(pdd,n-pdd). cbind stands for column bind: it binds together two columns into a matrix. The model can now be fitted with:

```
glm(cbind(pdd,n-pdd) ~ exposure, family=binomial,
    data=ppd)
```

(b) How many PDD cases are there from the NOP center and how many parrots in total ? And from other then the NOP center? use:

```
with(lr3,tapply(pdd,list(nop),sum))
```

and

```
with(lr3,tapply(n,list(nop),sum))
```

with means apply a function (here tapply) with a dataset (here: lr3). tapply applies a function, sum, to a data collumn, pdd, for several groups, nop.

(c) Use the previous exercise to make a table of pdd by nop.

(d) Calculate the odds ratio and give it an interpretation. Why does the outcome seem logical ?

(e) Fit the logistic regression model with nop as exposure and compare the results with those from the previous question.

4. Dalmatian.cvs

(a) Read in the data file dalmatian.csv.

(b) Explain how the variable fhs deals with the heritability.

(c) Fit the logistic regression model for deaf, with fhs as an exposure variable.

(d) Compare the AIC of the previous model with the model that only contains a constant. Also calculate the likelihood ratio and interpret the results.

Exercises 2

1. osteochon.csv

   (a) Read in the data file osteochon.csv

   (b) Fit a logistic regression model with the exposure variables food, ground and height.

   (c) Give for each exposure variable the likelihood ratio test, when it is left out of the model. Decide which exposure should be left out.

   (d) Fit the model without that exposure and do the same with this model as above. Continue until you decide nothing can be left out anymore.

   (e) Describe the final model you are left with and interpret the result.

   (f) Give profile confidence intervals for the estimates in the final model and

   (g) Write a short account of the analysis you just did. It should contain what the analysis was and its results.

2. episode.txt

   (a) Read in the file episode.txt

   (b) Fit the logistic regression model with exposures immune and age and with log(followup) as an offset.

   (c) Interpret the parameters and discuss the difference with the model without the offset.

   (d) Can immune or age or both be left out? Use AIC to check this.

3. osteochon.csv

   (a) Read in the data file osteochon.csv

   (b) Fit the logistic regression model with exposures father, food, ground and heigt. (Remember to use factor() for food, ground and father)

   (c) Use likelihood ratio tests to see which exposure can be left out, then fit that model and again see which exposure can be left out. Continue until no more exposures can be left out.

   (d) Discuss the final model. Give a possible interpretation of the terms in the model and why they are likely to be related to osteochondrosis.

(e) Start off with the full model again and try to reduce it by using AIC.

(f) Is the final model the same as the one you got with the likelihood ratio tests? Can you explain this?

(g) Give the 95% profile confidence interval for height and give the interpretation of this.

4. lowbirth .dat

(a) Fit the logistic regression model with exposures age, lwt, race, smoke, ptl, ht, ui and ftv.

(b) Find out which exposures can be left out using AIC.

(c) Discuss the final model.

(d) Write a short report about your findings. Include the statistical model you used, the method you used to reduce this model and the final results (estimates and standard errors).

Exercises 3

To fit the continuation ratio model:

```
part.table <- c(0,0,1,1)
gender <- c(0,1,0,1)
high <- c(458,468,75,51)
low <- c(86,105,544,673)
fit.cont <- glm(cbind(high,low)~factor(part.table)+ factor(gender) +
factor(part.table):factor(gender),family=binomial)
summary(fit.cont)
```

The lung cancer data can be read in with

```
read.table("cancer", header=TRUE)
```

Exercise: repeat the analysis from the text.