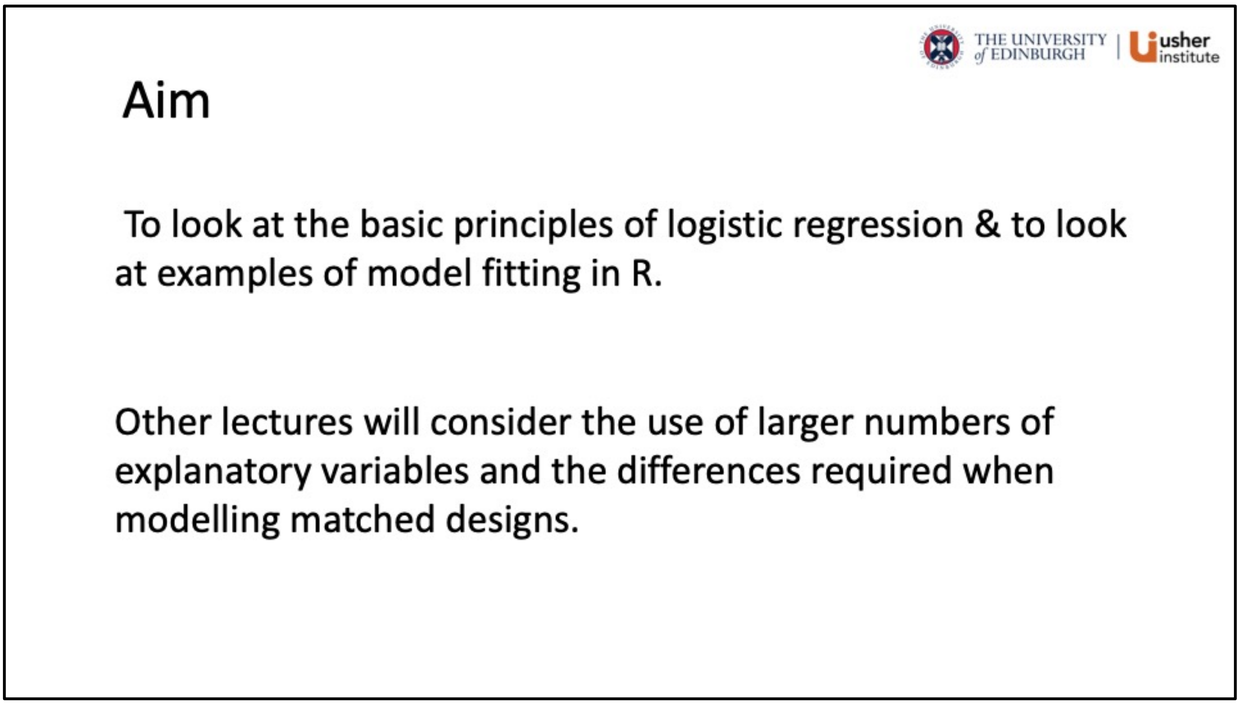
THE UNIVERSITY *of* EDINBURGH | usher institute

Introduction to Logistic Regression

Niall Anderson

Supported by

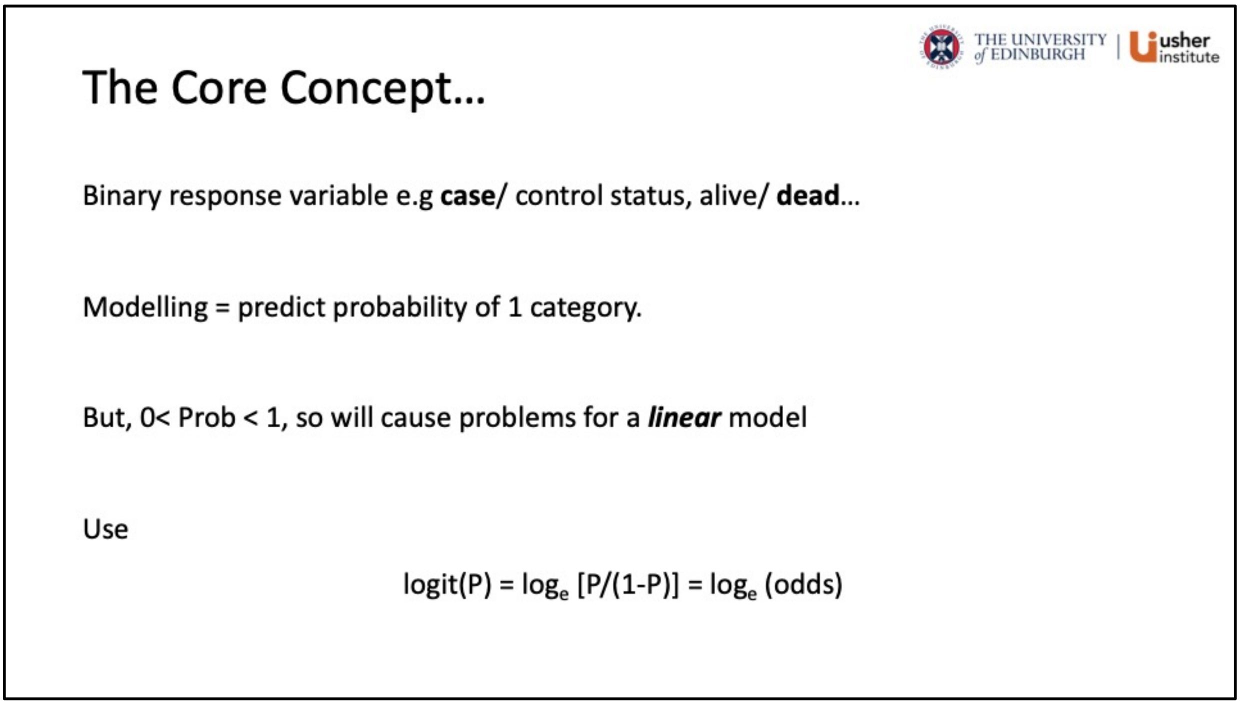
THE UNIVERSITY *of* EDINBURGH | Data-Driven Innovation



Aim

To look at the basic principles of logistic regression & to look at examples of model fitting in R.

Other lectures will consider the use of larger numbers of explanatory variables and the differences required when modelling matched designs.



The Core Concept…

Binary response variable e.g **case**/ control status, alive/ **dead**…

Modelling = predict probability of 1 category.

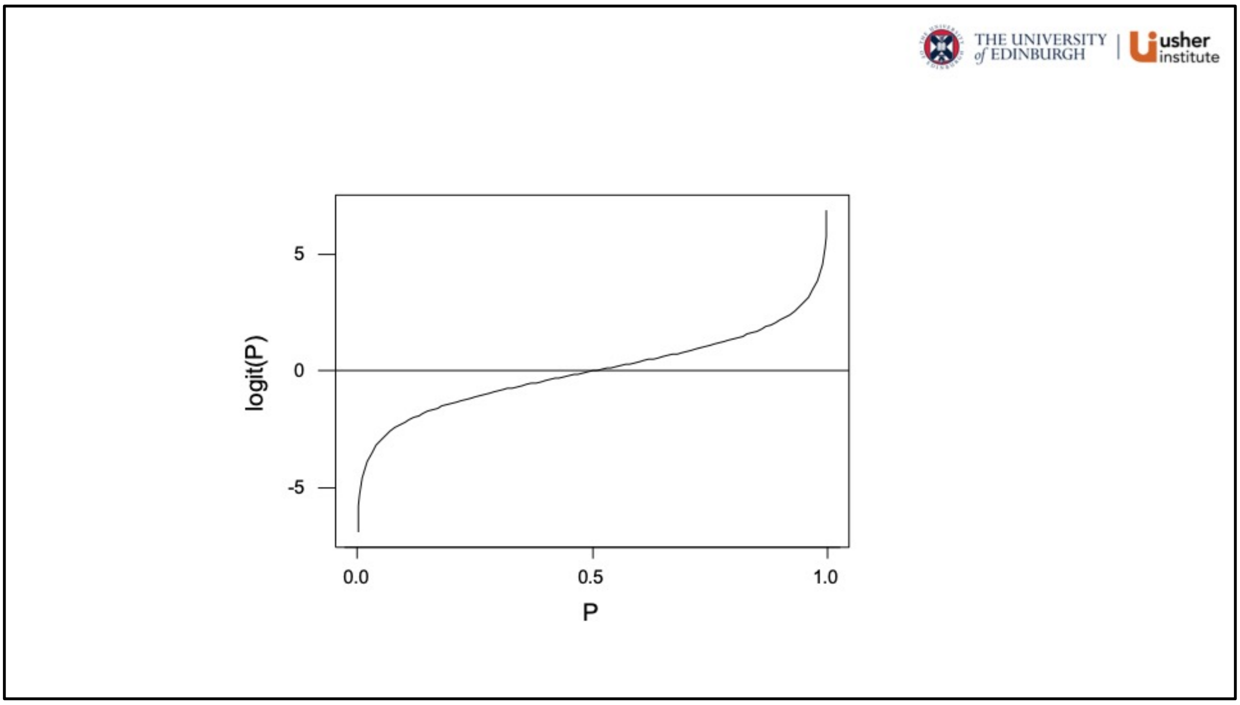
But, 0< Prob < 1, so will cause problems for a ***linear*** model

Use

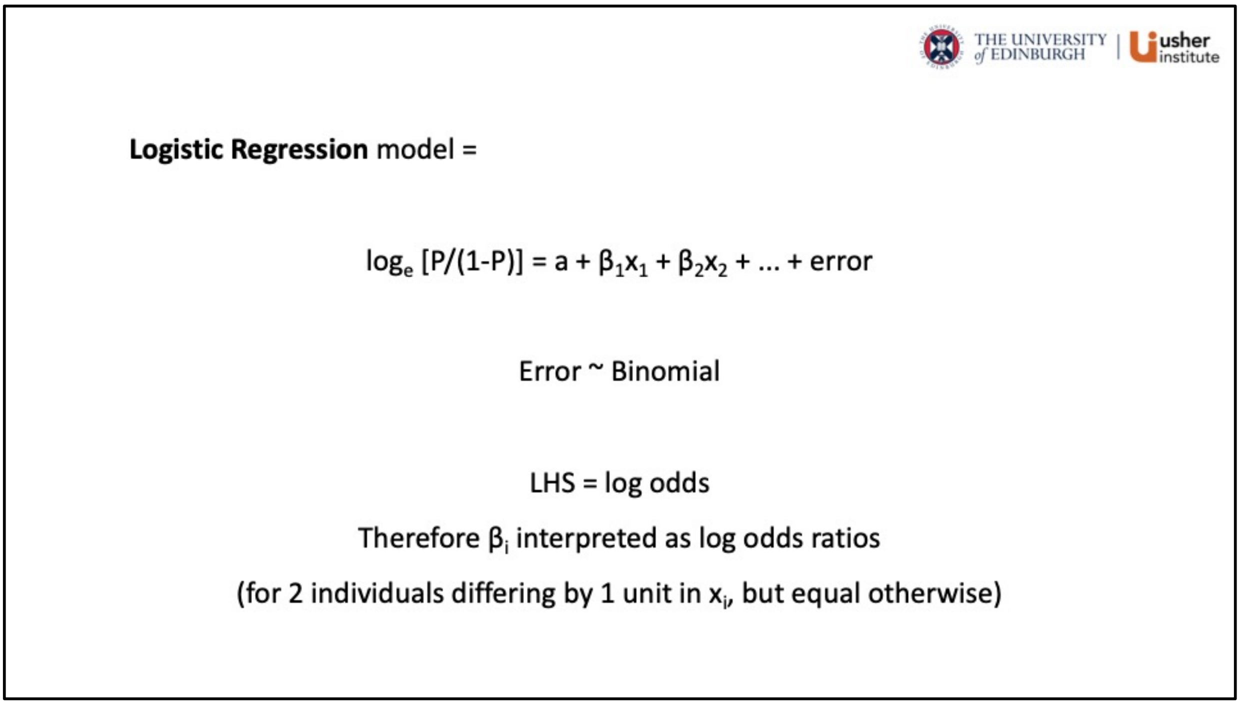
$$\text{logit}(P) = \log_e [P/(1-P)] = \log_e (\text{odds})$$

As mentioned before, we need a tool that is capable of modelling a **binary response variable**, as often we will be interested in factors affecting outcomes such as case/ control status, whether participants were alive or dead at the end of the study and so on. It is therefore quite convenient to think about the modelling task as predicting the **probability** of one of those categories (case/ dead/ etc).

However, probabilities are only found within a narrow range of numerical values (between 0 and 1), and this will be something that a simpler (linear model) framework will not be able to achieve – continuous outcome variables are not assumed to have limits of that sort. Therefore, we need to **transform** the probability of interest to a quantity that can vary over the entire range of Real numbers – we do this by using the **logit** transformation, shown here, and it turns out (perhaps quite conveniently) that P/(1-P) is actually the **odds** of the event of interest – something that helps with our interpretation of the model later.



This graph shows the effect of the logit transformation – the limited range of probabilities on the horizontal axis is modified to an infinite set of values on the vertical axis, with the curve heading down to minus infinity on the left and up to plus infinity on the right.



THE UNIVERSITY of EDINBURGH | usher institute

Logistic Regression model =

$$\log_e [P/(1-P)] = a + \beta_1 x_1 + \beta_2 x_2 + \ldots + \text{error}$$

Error ~ Binomial

LHS = log odds

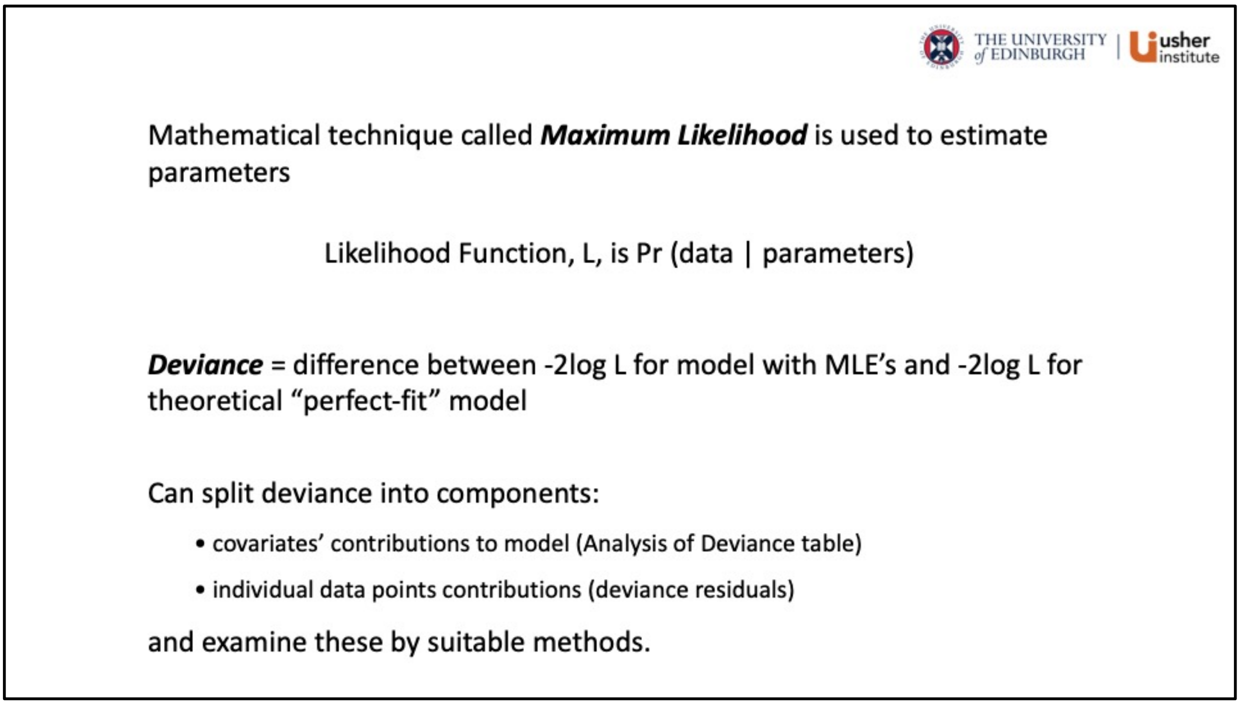
Therefore β_i interpreted as log odds ratios

(for 2 individuals differing by 1 unit in x_i, but equal otherwise)

With this adaptation, we can then use a model structure like this – we express the logit event probability in terms of a **linear predictor** involving the explanatory variables we need to include (the x terms), each modified by a parameter (the beta terms). Because the event is part of a binary outcome variable, we use a Binomial probability distribution as the model for the error term in the model: this error term is a theoretical construct to make the model behave in the way we want – it is not estimated or calculated in its own right.

As the left hand side of this formula is the log of the odds of outcome, the right hand side must be a sum of log odds also. Thus, we can interpret the beta terms as the change in the log odds of the event for a 1 unit difference in an x variable. Where x represents a categorical variable (e.g. Sex), then this gives a log odds ratio for Male relative to Female (or vice versa).

In either case, by back transforming (anti-log or e^{β}), we obtain actual **odds ratios** for the event of interest, a measure which we already know and use…



Mathematical technique called ***Maximum Likelihood*** is used to estimate parameters

Likelihood Function, L, is Pr (data | parameters)

Deviance = difference between -2log L for model with MLE's and -2log L for theoretical "perfect-fit" model

Can split deviance into components:

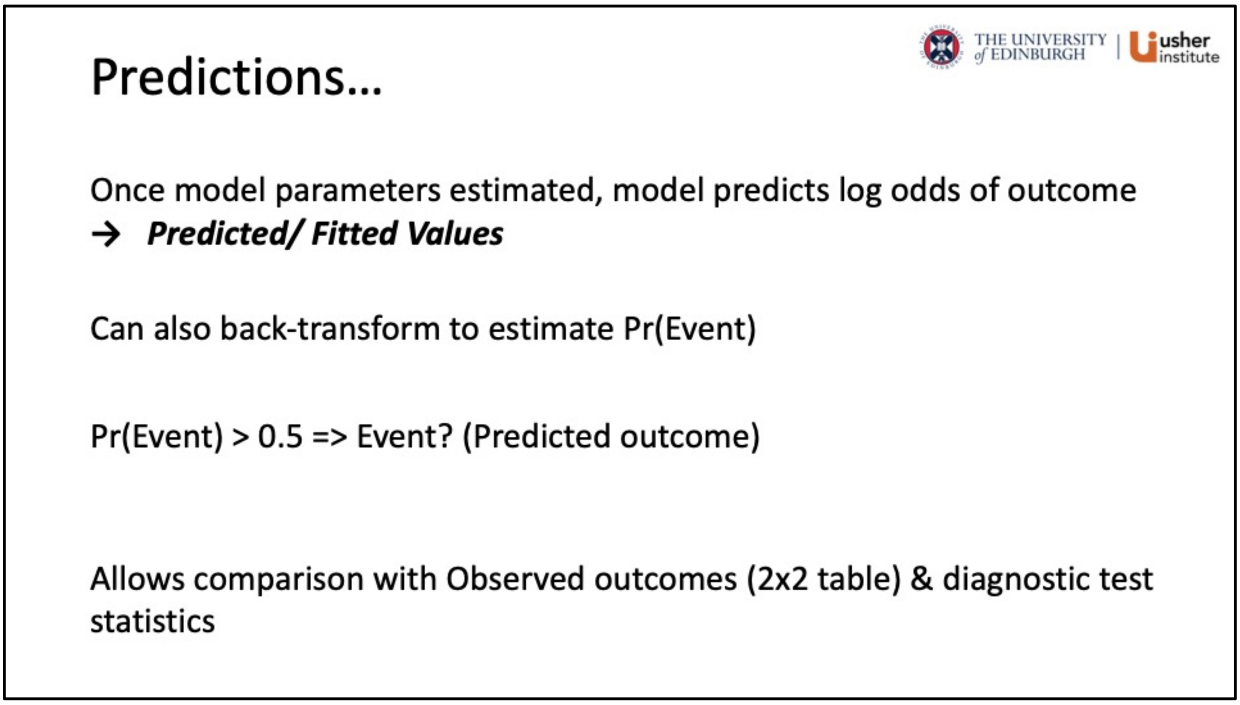
- covariates' contributions to model (Analysis of Deviance table)
- individual data points contributions (deviance residuals)

and examine these by suitable methods.

A mathematical technique called maximum likelihood estimation is used to estimate the parameters of logistic regression models. This depends on the Binomial probability distribution for our data given the set of parameters, but we instead regard our data as a fixed (non-varying) quantity, so that we can view this probability as a function of the possible values of the parameters given the data (essentially turning the function on its head!) – when given this interpretation, we call the function the **Likelihood**, and then choose parameters to achieve the largest possible value of this function.

This mathematical framework allows us to calculate a very useful quantity – the **Deviance**. This is a log scale difference in the likelihood for our "best fit" model compared to the theoretical model that would give a perfect fit to the data. The latter would be a model that contained one parameter for every data point, perfectly describing the data but actually just re-casting it into a different form, so this is not a model that would be useful in practice. However, it does allow us to assess how effective our "best fit" model is, and to compare the effectiveness of different competing models for the same data.

Furthermore, we can split the deviance for our model into components that represent the individual covariates’ (explanatory variables’) contributions to the model, allowing production of an **Analysis of Deviance** table. We can also break the Deviance down further, into components matched to individual data points, and this allows us to carry out some model checking – see the Statistical Modelling for Epidemiology course for more details.



THE UNIVERSITY *of* EDINBURGH | usher institute

Predictions…

Once model parameters estimated, model predicts log odds of outcome
→ ***Predicted/ Fitted Values***

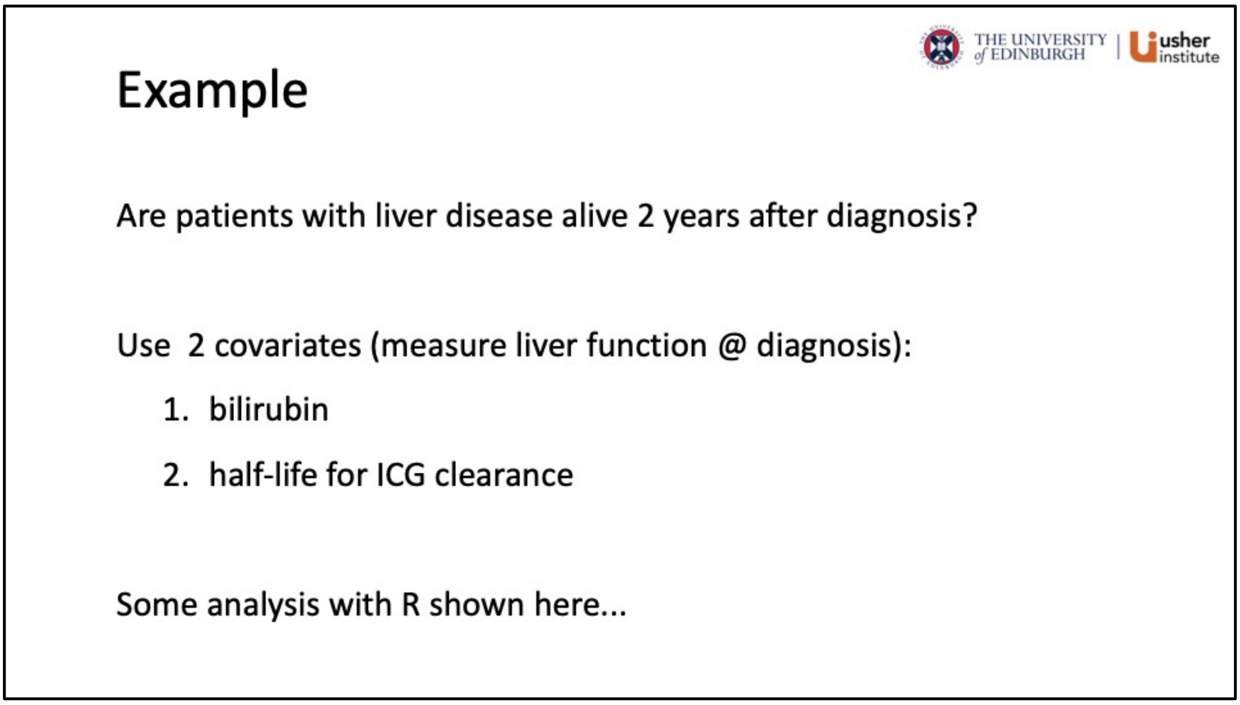
Can also back-transform to estimate Pr(Event)

Pr(Event) > 0.5 => Event? (Predicted outcome)

Allows comparison with Observed outcomes (2x2 table) & diagnostic test statistics

As well as using the odds ratios (from the parameters) of each explanatory variable, it is possible to think about the predictions made by the model, which are (directly) log odds of outcome events, but can be back-transformed to estimates of the probability of an event for each study participant.

If this probability is greater than 0.5, we would tend to predict that the participant will have an event, whereas if less than 0.5 we would predict that they would not. We could therefore construct a 2 x 2 table of observed and predicted events as a way of investigating the usefulness of the model. Additionally, one could then use measures such as sensitivity, specificity and ROC curves to summarise the predictive ability of the current model – this is quite frequently encountered in the literature.



Example

Are patients with liver disease alive 2 years after diagnosis?

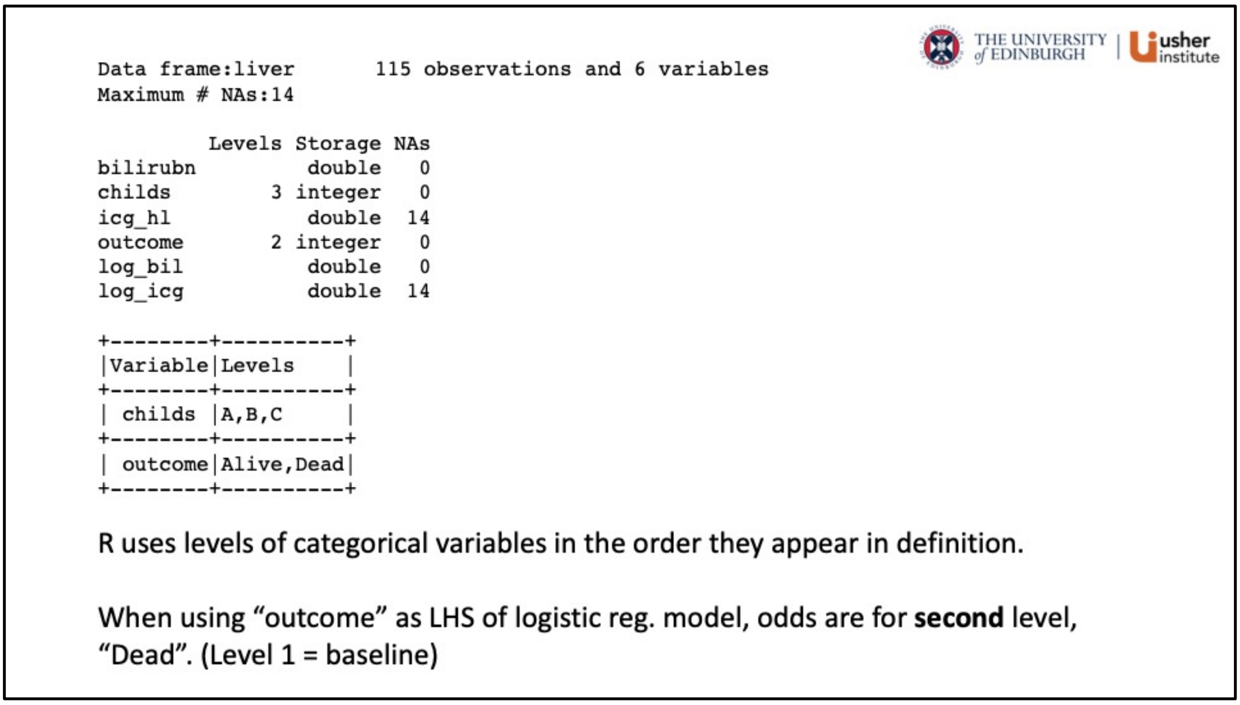
Use 2 covariates (measure liver function @ diagnosis):

1. bilirubin
2. half-life for ICG clearance

Some analysis with R shown here...

Let’s look at an example of a study designed to investigate 2 key measures of liver function (bilirubin and half-life for ICG clearance), in terms of predicting whether patients with liver disease would be alive 2 years after diagnosis.

We’ll look at the output generated by an analysis of the data in R.



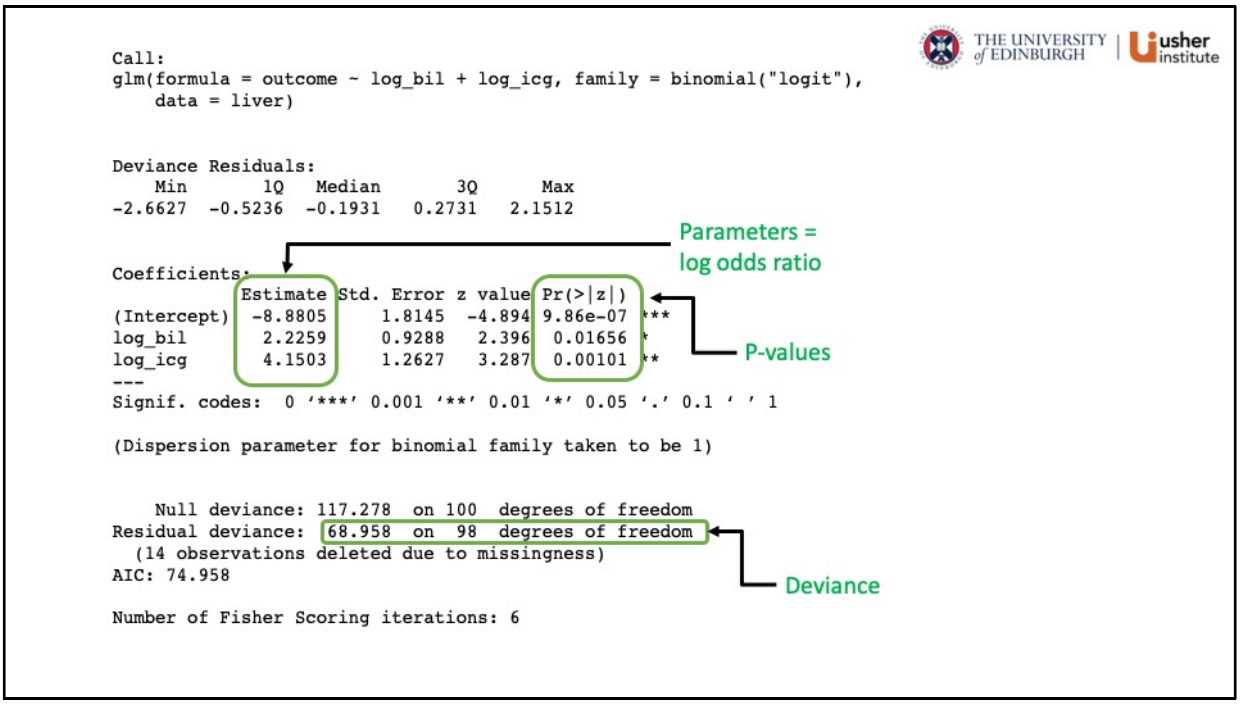
```
Data frame:liver      115 observations and 6 variables
Maximum # NAs:14

         Levels Storage NAs
bilirubn         double   0
childs        3 integer   0
icg_hl           double  14
outcome       2 integer   0
log_bil          double   0
log_icg          double  14

+--------+----------+
|Variable|Levels    |
+--------+----------+
| childs |A,B,C     |
+--------+----------+
| outcome|Alive,Dead|
+--------+----------+
```

R uses levels of categorical variables in the order they appear in definition.

When using “outcome” as LHS of logistic reg. model, odds are for **second** level, “Dead”. (Level 1 = baseline)

The data set contains a small number of variables, including a measurement (Childs’ scores) that we won’t make use of in this example. We will concentrate on the last two variables as independent variables in our model – these are $\log_{10}$ transformed versions of the bilirubin and ICG measurements (as these variables in their original form are highly positively skewed, and therefore are unlikely to contribute in a **linear** way to the linear predictor of our logistic regression model).

It is important to note that R will always (by default) use the levels of a categorical variable in the order in which they appear in this listing – that is, alphabetical or strictly numerical order. When we use the *outcome* variable as the dependent variable, R will fit the model in terms of the probability of the second level as the event of interest, so our models will estimate the probability of **being dead at 2 years**. The first level will typically be regarded as the baseline level for the problem.

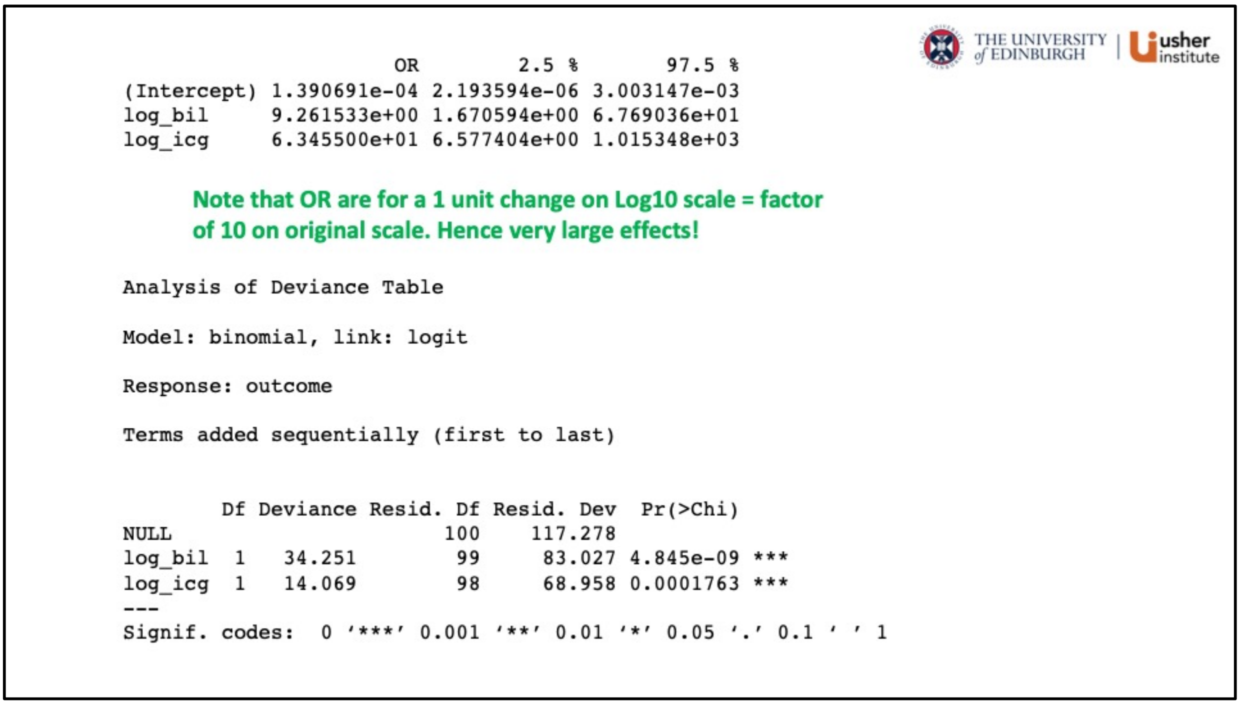


```
Call:
glm(formula = outcome ~ log_bil + log_icg, family = binomial("logit"),
    data = liver)


Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6627  -0.5236  -0.1931   0.2731   2.1512

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.8805     1.8145  -4.894 9.86e-07 ***
log_bil       2.2259     0.9288   2.396  0.01656 *
log_icg       4.1503     1.2627   3.287  0.00101 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 117.278  on 100  degrees of freedom
Residual deviance:  68.958  on  98  degrees of freedom
  (14 observations deleted due to missingness)
AIC: 74.958

Number of Fisher Scoring iterations: 6
```

We fit a logistic regression model for the log odds of death by 2 years in terms of log bilirubin and log ICG. In R, this uses a function called **glm()**, which stands for **generalized linear models**. This is a large taxonomy of related statistical model types, that share some underlying characteristics, but can be separated by their **family** type (i.e. the type of error structure – here, Binomial), and the nature of any transformation of the outcome variable needed (called the **link** function – here, the logit transformation). You will notice the definition of family type and link function in the glm() call…

The parameters of the model are displayed in the Coefficients column, and these are log odds ratios overall (intercept) and for the two log-transformed variables (these are continuous, so have only 1 parameter each). The contribution of each variable to the model is evaluated via the p-value in the same row in the last column of that table.

Lower down we see the deviance (R calls this the Residual Deviance), representing the variability that has not been explained by the model as it stands.



```
                     OR        2.5 %       97.5 %
(Intercept) 1.390691e-04 2.193594e-06 3.003147e-03
log_bil     9.261533e+00 1.670594e+00 6.769036e+01
log_icg     6.345500e+01 6.577404e+00 1.015348e+03
```

Note that OR are for a 1 unit change on Log10 scale = factor of 10 on original scale. Hence very large effects!

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: outcome

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      100    117.278
log_bil  1   34.251        99     83.027 4.845e-09 ***
log_icg  1   14.069        98     68.958 0.0001763 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The first table on this slide shows what happens when we back transform the parameters to obtain odds ratios. These are quite large values and (as with the p-values in the previous slide) are output in scientific notation – the log bilirubin OR is 9.26 and the log ICG OR is 63.45 ($6.345\text{x}10^1$). Because of the log-transform, these correspond to changes of a factor of 10 on the original scale of bilirubin or ICG, which are very large changes in their own right, and perhaps much larger than we would expect to see between any 2 study participants.

We then see the Analysis of Deviance table, which indicates (in a strict sequence) the effect of fitting first bilirubin in the model, then ICG in addition to bilirubin. Both additions are well supported by the data – the chi-squared p-value is testing H0: parameter = 0, and both H0 are rejected strongly. You should note that these tables work somewhat differently others you may encounter (e.g. ANOVA tables - these usually do not depend on the order of fitting, but the Deviance tables do). This can lead to different p-values being observed for the same variables in the same model, depending on the precise ordering adopted – this can’t really be avoided, unfortunately.

In summary, the two liver function measurements seem to give separate but important information on estimating the probability of death by 2 years in this population of liver disease patients. The precise magnitude of the effect is slightly concealed by the necessary $\log_{10}$ transformations, but in both cases, larger values are associated with greater probabilities of death.