Introduction to Generalized Linear Models

I. Motivation

In this lecture we extend the ideas of linear regression to the more general idea of a generalized linear model (GLM). The essence of linear models is that the response variable is continuous and normally distributed: here we relax these assumptions and consider cases where the response variable is non-normal and in particular has a discrete distribution. Although these models are more general than linear models, nearly all of the techniques for testing hypotheses regarding the regression coefficients, and checking the assumptions of the model apply directly to a glm. In addition, a linear model is simply a special type of a generalised linear model and thus all of the discussion below applies equally to linear models.

In Lecture 1 we saw that a typical statistical model can be expressed as an equation that equates the mean(s) of the response variable(s) to some function of a linear combination of the explanatory variables:

$$E\left[Y \middle| X = x\right] = \eta(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p) = \eta[LC(X;\beta)], \qquad (1)$$

In equation (1), the form of the function $\eta(\ )$ is known, as are Y and X (the latter for a particular choice of explanatory variables). However, the ***parameters*** of the model, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$, are not known and must be estimated. The simple linear model is a special case of the above in which the function $\eta()$ is the identity function.

Generalized linear models extend the ideas underlying the simple linear model

$$E[Y \mid X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p, \quad Y \sim N\left(\mu, \sigma^2\right) \qquad (2)$$

where the Y_i are independent, to the following more general situations:

1. Response variables can follow distributions other than the Normal distribution. They can be discrete, or logical (one of two categories).

2. The relationship between the response and the explanatory variables does not have to take on the simple linear form above.

GLMs are based on the *exponential family of distributions* that shares many of the desirable statistical properties of the Normal distribution. Consider a single random variable Y whose probability distribution depends only on a single parameter θ. The distribution of Y belongs to the exponential family if it can be written in the form

$$f(y,\theta) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)] \quad (3)$$

where *a*, *b* and *c* are known functions. Furthermore, ϕ is a *scale parameter* that is treated as a nuisance parameter if it is unknown. If ϕ is known, this is an exponential-family model with *canonical parameter* θ. Some well-known distributions that belong to the exponential family of distributions include the Normal, exponential, Poisson and binomial distributions. For example, consider the discrete random variable Y that follows a Poisson distribution with parameter λ. The probability function for Y is

$$f(y,\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

where y takes the values $0, 1, 2, \ldots$. This can be rewritten as

$$f(y,\lambda) = \exp(y\log\lambda - \lambda - \log y!) = \exp(y\theta - \exp(\theta) - \log y!) = f(y,\theta).$$

This is exactly the form in (3) with $\theta = \log(\lambda)$, $\phi = 1$, $a(\phi) = 1$, $b(\theta) = \exp(\theta)$ and $c(y,\theta) = -\log y!$. Similarly, all the other distributions in the exponential family can be rewritten in form (2).

In the case of GLMs we require an extension of numerical estimation methods to estimate the parameters $\underline{\beta}$ from the linear model in (2), to a more general situation where there is some non-linear function, *g*, relating $E(Y_i) = \mu_i$ to the linear component $\underline{x}_i' \underline{\beta}$, that is

$$g(\mu_i) = \underline{x}_i' \underline{\beta}$$

where *g* is called the *link function*. The estimation of the parameters is typically based on maximum likelihood. Although explicit mathematical expressions can be found for the estimators in some special cases, numerical optimisation methods are usually needed. These methods are included in most modern statistical software packages. It is not the aim of this course to go into any detail on estimation, since we will focus on the application of these models rather than their estimation.

GLMs are extensively used in the analysis of binary data (e.g. logistic regression) and count data (e.g. Poisson and log-linear models). We will consider some of these applications in the following lecture. In this lecture we will introduce the basic GLM setup and assumptions.

II. Basic GLM setup and assumptions

We will define the generalized linear model in terms of a set of independent random variables $Y_1, Y_2, \ldots, Y_N$ each with a distribution from the exponential family of distributions. The generalized linear model has three components:

1. The *random component*: the response variables, $Y_1, Y_2, \ldots, Y_N$, are assumed to share the same distribution from the exponential family of distributions introduced in the previous section, with $E(Y_i) = \mu_i$ and constant variance σ^2. This part describes how the response variable is distributed.

2. The *systematic component*: covariates (explanatory variables) $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_p$ produce a linear predictor $\underline{\eta}$ given by

$$\underline{\eta} = \sum_{i=1}^{p} \underline{x}_j \beta_j$$

3. The *link function,* g, between the random and systematic components. It describes how the covariates are related to the random component, i.e. $\eta_i = g(\mu_i)$, where $\mu_i = E(Y_i)$ and can be any monotone differentiable function.

An important aspect of generalized linear models that we need to keep in mind is that they assume independent (or at least uncorrelated) observations. A second important assumption is that there is a single error term in the model. This corresponds to *Assumption 1* for the linear model, namely, that the only error in the model has to do with the response variable: we will assume that the *X* variables are measured without error. In the case of generalized linear models we no longer assume constant variance of the residuals, although we still have to know how the variance depends on the mean. The *variance function* $V(\mu_i)$ relates the variance of Y_i is related to its mean μ_i. The form of this function is determined by the distribution that is assumed.

III. Goodness of fit and comparing models

Overview

A very important aspect of generalized models, and indeed all statistical models (Lectures 1-4), is to evaluate the relevance of our model for our data and how well it fits the data. In statistical terms this is referred to as goodness of fit. We are also interested in comparing different models and selecting a model that is reasonably simple, but that provides a good fit. This involves finding a balance between improving the fit on one side without unnecessary increasing the complexity of the model on the other. In a statistical modelling framework we perform hypothesis tests to compare how well two related models fit the data. In the generalized linear model framework, the two models we compare should have the same probability distribution and link function, but they

can differ with regards to the number of parameters in the systematic component of the model. The simpler model relating to the null hypothesis H_0 is therefore a special case of the other more general model. If the simpler model fits the data as well as the more general model it is retained on the grounds of parsimony and H_0 cannot be rejected. If the more general model provides a significantly better fit than the simple model, then H_0 is rejected in favour of the alternative hypothesis H_1, which corresponds to the more general model. In order to make these comparisons we need goodness of fit statistics to describe how well the models fit. These statistics can be based on any of a number of criteria such as the maximum value of the log-likelihood function, the minimum value of the sum of squares criterion or a composite statistic based on the residuals. If $f_Y(y,\theta)$ is the density function for a random variable Y given the parameter θ, then the *log-likelihood* based on a set of independent observations of Y, $y_1, y_2, \ldots, y_n$, is then defined as

$$l(\mu, y) = \sum_i \log f_i(y_i, \theta_i)$$

where $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$. It is important to note the subtle shift in emphasis from the density function. In the density function $f(y,\theta)$ is considered as a function in y for fixed θ, whereas the log-likelihood is primarily considered as a function of θ for the particular data observed ($y_1, y_2, \ldots, y_n$).

In order to test the hypotheses above, sampling distributions of the goodness of fit statistics are required. In the following subsection we consider one such goodness of fit criterion, the deviance, in a bit more detail.

Finally, we will say something about the choice of scale for the analysis. It is an important aspect of the model selection process, although scaling problems are considerably reduced in the generalized linear model setup. The normality and constant variance assumption of the linear regression model is for instance no longer a requirement. The choice of scale is largely dependent on the purpose for which the scale will be used. It is also important to keep in mind that no single scale will simultaneously produce all the desired properties.

The Deviance

One way of assessing the adequacy of a model is to compare it with a more general model with the maximum number of parameters that can be estimated. It is referred to as the *saturated model*. In the saturated model there is basically one parameter per observation. The deviance assesses the goodness of fit for the model by looking at the difference between the log-likelihood functions of the saturated model and the model under investigation, i.e. $\ell(\underline{b}_{sat}, \underline{y}) - \ell(\underline{b}, \underline{y})$. Here $\underline{b}_{sat}$ denotes the maximum likelihood estimator of the parameter vector of the saturated model, $\underline{\beta}_{sat}$, and $\underline{b}$ is the maximum likelihood estimator of the parameters of the model under investigation, $\underline{\beta}$. The

maximum likelihood estimator is the estimator that maximises the likelihood function. The *deviance* is defined as

$$D = 2\{\ell(\underline{b}_{sat}, \underline{y}) - \ell(\underline{b}, \underline{y})\}.$$

A small deviance implies a good fit. The sampling distribution of the deviance is approximately $\chi^2(m-p, \upsilon)$, where ν is the non-centrality parameter. The deviance has an exact χ^2 distribution if the response variables Y_i are normally distributed. In this case, however, D depends on $\text{var}(Y_i) = \sigma^2$ which, in practice, is usually unknown. This prevents the direct use of the deviance as a goodness of fit statistic in this case. For other distributions of the Y_i, the deviance may only be approximately chi-square. It must be noted that this approximation can be very poor for limited amounts of data. In the case of the binomial and Poisson distributions, for example, D can be calculated and used directly as a goodness of fit statistic. If the scale parameter ϕ is unknown or known to have a value other than one, we us a scaled version of the deviance

$$\frac{D}{\phi}$$

and we call it the *scaled deviance*.

The deviance forms the basis for most hypothesis testing for generalized linear models. Suppose we are interested in comparing the fit of two models. These models need to have the same probability distribution and the same link function. The models also need to be hierarchical, which means that the systematic component of the simpler model M_0 is a special case of the linear component of the more general model M_1. Consider the null hypothesis

$$H_0 : \underline{\beta} = \underline{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

that corresponds to model M_0 and a more general hypothesis

$$H_1 : \underline{\beta} = \underline{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

that corresponds to model M_1, with $q < p < N$.

We can test H_0 against H_1 using the difference of the deviance statistics

$$\Delta D = D_0 - D_1 = 2\{\ell(\underline{b}_{sat}, \underline{y}) - \ell(\underline{b}_0, \underline{y})\} - 2\{\ell(\underline{b}_{sat}, \underline{y}) - \ell(\underline{b}_1, \underline{y})\} = 2\{\ell(\underline{b}_1, \underline{y}) - \ell(\underline{b}_0, \underline{y})\}$$

If both models describe the data well, then D_0 follows a $\chi^2(N-q)$ distribution and D_1 follows a $\chi^2(N-p)$distribution. It then follows that ΔD has a $\chi^2(p-q)$ distribution under certain independence assumptions. If the value of ΔD is consistent with the $\chi^2(p-q)$ distribution we would generally choose the model M_0 corresponding to H_0 because it is simpler. If the value of ΔD is in the critical region of the $\chi^2(p-q)$ distribution we would reject H_0 in favour of H_1 on the grounds that model M_1 provides a significantly better description of the data. It must be noted that this model too may not fit the data particularly well.

If the deviance can be calculated from the data, then ΔD provides a good method for hypothesis testing. The sampling distribution of ΔD is usually better approximated by the chi-squared distribution than is the sampling distribution of a single deviance.

Model checking

As was the case for the simple linear model, we should perform model checking after we have fitted a particular generalized linear model. As before we should look at whether the model is reasonable and we should investigate whether the various assumptions we make when we fit and draw inference using the model are satisfied. If the checks and investigations do reveal that there is a problem, then there are a number of different solutions available to us. These were discussed in Lecture Notes on model checking for linear models. In this section we discuss graphical techniques available to us for trying to detect systematic departures from our generalized linear model that may for example be the result of an incorrectly specified link function, variance function or a misspecification of the explanatory variables in the systematic component in the model. As was the case for the linear model, many of these graphical techniques entail using the residual values from our fitted model.

For generalized linear models we require an extended definition of residuals, applicable to distributions other than the Normal distribution. Many of these these residuals are discussed in detail in McCullagh and Nelder (1989). These residuals can be used to assess various aspects of the adequacy of a fitted generalized linear model that we have mentioned above. The residuals should be unrelated to the explanatory variables and they can also be use to identify any unusual values that may require some further investigation. Various plots of the residuals can be used to assess these properties. Systematic pattern in the residual plots can for example be indicative of an unsuitable link function, wrong scale of one or more of the predictors, or omission of a quadratic term in a predictor. Examples of extended definitions of residuals that are widely used in model checking for GLMs include the Pearson and deviance residuals.

The *Pearson residuals* are just rescaled versions of the raw or response residuals and are defined as

$$r_P = \frac{(y_i - \hat{\mu}_i)}{\sqrt{V(\hat{\mu}_i)}}.$$

Here $V(\mu_i)$ is the *variance function*. The name is taken from the fact that for the Poisson distribution the Pearson residual is just the signed square root of the component of the Pearson X^2 goodness-of-fit statistic, so that

$$\sum r_p^2 = X^2.$$

If the deviance is used as a measure of discrepancy of a generalized linear model, then each unit contributes a quantity d_i to the deviance, so that $D = \sum_i d_i$. The *deviance residual* is defined as

$$r_D = sign(y_i - \mu_i)\sqrt{d_i}\ .$$

Most often standardised versions of the above residuals are used in model checking. The standardised versions of the Pearson and deviance residual are given by

$$r_P{}' = \frac{(y_i - \hat{\mu}_i)}{\sqrt{\hat{\phi}\, V(\hat{\mu}_i)(1-h)}} \text{ and } r_D{}' = \frac{r_D}{\sqrt{\hat{\phi}(1-h)}}$$

respectively. In the above h is the equivalent to the leverage that we have defined for the linear model. In general the deviance residual, either unstandardized or standardized, is preferred to the Pearson residual. Below we will discuss a few basic plots of the (standardized) residuals that can be used to check the validity of our model.

1. Informal checks using the residuals

It is almost standard procedure to consider scatterplots of the residuals against some function of the fitted values. *Scatterplots of the standardized deviance residuals against the estimated linear predictor* $\hat{\eta}$ or *against the fitted values transformed to a constant scale of the error distribution* are recommended for this purpose. For a few commonly used error distributions the following transformed values are recommended:

$\hat{\mu}$ for Normal errors,

$2\sqrt{\hat{\mu}}$ for Poisson errors,

$2\sin^{-1}\sqrt{\hat{\mu}}$ for binomial errors,

$2\log\hat{\mu}$ for gamma errors.

The plot should be centred at $\hat{\mu} = 0$ and a constant range. Typical deviations from this pattern include curvature in the residuals with the mean and a systematic change in

the range of the residuals with fitted value. Curvature may arise from several causes, including the wrong choice of link function, wrong choice of scale of one or more of the covariates, or omission of a quadratic term in a covariate.

A second useful scatterplot is that of the *residuals against an explanatory variable* in the linear predictor. This plot should exhibit the same form as the plot above. The presence of systematic trend usually arises for the same reasons as for the plot above. Additionally, the trend may also be the result of a faulty scale in another explanatory variable that is closely correlated with the one under investigation.

The third scatterplot that we will consider is termed an *added-variable plot*. This is equivalent to the partial regression plot that we considered for the linear model. This plot helps us to check if an omitted explanatory variable should be included in the linear predictor. The added-variable plot for a particular candidate explanatory variable is formed by (a.) finding the unstandardized residuals for the existing generalized linear model with response variable Y and any already included explanatory variables, (b.) finding the unstandardized residuals for another linear model in which the candidate explanatory variable is treated as the response, using the same linear predictor as for Y (here, the candidate explanatory variable is treated as the response variable), and (c.) plotting the first set of residuals against the second set. The presence of a trend in the points in this plot indicates that you might consider including the particular candidate variable in the model as an explanatory variable, and the shape of the trend can tell you what forms of the variable you might include.

2. Checking the variance function

A plot of the absolute residuals against the fitted values gives an informal check on the adequacy of the assumed variance function. An ill-chosen variance function will result in a trend in the mean. A positive trend indicates that the current variance function is increasing to slowly with the mean, and vice versa.

3. Checking the link function

An informal check involves examining the plot of the scale-adjusted dependent variable against $\hat{\eta}$, the estimated linear predictor. This should approximately be a straight line. For link functions of the power family an upward curvature in the plot points to a link with higher power than that used. A downward curvature points to a lower power. For binary data this plot is uninformative. McCullagh and Nelder (1989) discuss more formal methods for this situation. Note that checks for the link function are affected by failure to establish the correct scales for one or more of the explanatory variables in the linear predictor. This can be validated using partial residual plots. They are described below.

4. Checking the scales of explanatory variables

The partial residual plot is a useful tool for checking whether the correct scales have been used for the explanatory variables. In its generalized form the partial residual is

defined by

$$u = z - \hat{\eta} + \hat{\gamma}x$$

where z is the adjusted dependent variable, $\hat{\eta}$ the fitted linear predictor and $\hat{\gamma}$ the parameter estimate for the explanatory variable x. The plot of u against x provides an informal check if the scale of x is satisfactory. A correctly specified scale should result in an approximately linear plot. The form of the plot may suggest a suitable alternative if the scale is not appropriate. Note that distortions in this plot may also occur when the scales of other explanatory variables are wrong, which may require that we look at partial residual plots for several explanatory variables.

5. Checks for outlying or influential points

We also need to check for individual points that may differ from the general pattern set by the remainder of the points. We defined the *leverage* h for individual points in the context of the linear model by which to judge their influence on the fit. We can consider these measures for GLMs as well, but we must note that a point in the extreme of the explanatory variable's range will not necessarily have a high leverage if its weight is small.

The Cook's distance was introduced as a measure of influence for the linear model in a previous lecture. Adapted versions of the Cook's distance can be used for generalized linear models. The Studentized residuals

$$e_i^* = r_i^* = \frac{Y_i - \hat{Y}_{(i)}}{s\hat{e}(Y_i - \hat{Y}_{(i)})}$$

that were introduced in the context of the linear model can also be used to assess the consistency of individual points. One-step approximations of these residuals exist that are appropriate for GLMs. To interpret the n values that we get for the leverage, Cook's distance and Studentized residuals respectively, we need some measure to assess how large extreme values would be in a sample of a given size even if no unusual points were present. Normal plots can be used for this purpose. There are two forms of Normal plots: the half-Normal plot and the Normal plot. We will not go into the theoretical detail underlying these plots. We will only mention that the half-Normal plot is appropriate for non-negative quantities like the leverage and Cook's distance, while for the Studentized residuals there are two options, either a half-Normal plot of $|r^*|$ or a full Normal plot of r^* itself. For either plot the ordered values of the statistic are plotted against the expected order statistics of a Normal sample. Extreme points will appear at the extremes of the plot, and may possibly deviate from the trend indicated by the remainder of the points. Note that is trend is not necessarily linear in the case of the leverage or Cook's distances.

6. Checks for correlations in the errors

An index plot of the residuals should can be used to assess correlation in the residuals. If the residuals are independent this should fluctuate randomly without systematic pattern. If the residuals are correlated special modelling methods are needed.

K. Javaras and W. Vos (2002)

References

Davidson, A. C. and Snell, E. J. (1991). *Residuals and diagnostics*. Chapter 4 of Hinkley et al. (1991).

Dobson, A. (1990). *An Introduction to Generalized Linear Models (2nd ed.)*. Boca Raton, FL: Chapman and Hall/CRC.

Hinkley, D. V., Reid, N. and Snell, E. J. eds (1991). *Statistical Theory and Modelling. In Honour of Sir David Cox*. London: Chapman & Hall.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models (2nd ed.)*. London: Chapman and Hall.