Chapter 12

Generalized Linear Models

12.1 Introduction

Generalized linear models are an important class of parametric 1D regression models that include multiple linear regression, logistic regression and loglinear Poisson regression. Assume that there is a response variable Y and a $k \times 1$ vector of nontrivial predictors $\boldsymbol{x}$. Before defining a generalized linear model, the definition of a one parameter exponential family is needed. Let $f(y)$ be a probability density function (pdf) if Y is a continuous random variable and let $f(y)$ be a probability mass function (pmf) if Y is a discrete random variable. Assume that the *support of the distribution* of Y is $\mathcal{Y}$ and that the *parameter space* of θ is Θ.

Definition 12.1. A *family* of pdfs or pmfs $\{f(y|\theta) : \theta \in \Theta\}$ is a **1-parameter exponential family** if

$$f(y|\theta) = k(\theta)h(y)\exp[w(\theta)t(y)] \tag{12.1}$$

where $k(\theta) \geq 0$ and $h(y) \geq 0$. The functions $h, k, t,$ and w are real valued functions.

In the definition, it is crucial that k and w do not depend on y and that h and t do not depend on θ. The parameterization is not unique since, for example, w could be multiplied by a nonzero constant m if t is divided by m. Many other parameterizations are possible. If $h(y) = g(y)I_{\mathcal{Y}}(y)$, then usually $k(\theta)$ and $g(y)$ are positive, so another parameterization is

$$f(y|\theta) = \exp[w(\theta)t(y) + d(\theta) + S(y)]I_{\mathcal{Y}}(y) \tag{12.2}$$

where $S(y) = \log(g(y))$, $d(\theta) = \log(k(\theta))$, and the support $\mathcal{Y}$ does not depend on θ. Here the indicator function $I_{\mathcal{Y}}(y) = 1$ if $y \in \mathcal{Y}$ and $I_{\mathcal{Y}}(y) = 0$, otherwise.

Definition 12.2. Assume that the data is $(Y_i, \boldsymbol{x}_i)$ for $i = 1, ..., n$. An important type of **generalized linear model (GLM)** for the data states that the $Y_1, ..., Y_n$ are independent random variables from a 1-parameter exponential family with pdf or pmf

$$f(y_i|\theta(\boldsymbol{x}_i)) = k(\theta(\boldsymbol{x}_i))h(y_i)\exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i\right]. \tag{12.3}$$

Here ϕ is a known constant (often a dispersion parameter), $a(\cdot)$ is a known function, and $\theta(\boldsymbol{x}_i) = \eta(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i)$. Let $E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i)$. The GLM also states that $g(\mu(\boldsymbol{x}_i)) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$ where the **link function** g is a differentiable monotone function. Then the **canonical link function** uses the function c given in (12.3), so $g(\mu(\boldsymbol{x}_i)) \equiv c(\mu(\boldsymbol{x}_i)) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i$, and the quantity $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ is called the **linear predictor** and the **sufficient predictor** (SP).

The GLM parameterization (12.3) can be written in several ways. By Equation (12.2),

$$\begin{aligned} f(y_i|\theta(\boldsymbol{x}_i)) &= \exp[w(\theta(\boldsymbol{x}_i))y_i + d(\theta(\boldsymbol{x}_i)) + S(y)]I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{c(\theta(\boldsymbol{x}_i))}{a(\phi)}y_i - \frac{b(c(\theta(\boldsymbol{x}_i))}{a(\phi)} + S(y)\right] I_{\mathcal{Y}}(y) \\ &= \exp\left[\frac{\nu_i}{a(\phi)}y_i - \frac{b(\nu_i)}{a(\phi)} + S(y)\right] I_{\mathcal{Y}}(y) \end{aligned}$$

where $\nu_i = c(\theta(\boldsymbol{x}_i))$ is called the natural parameter, and $b(\cdot)$ is some known function.

Notice that a GLM is a parametric model determined by the 1-parameter exponential family, the link function, and the linear predictor. Since the link function is monotone, the **inverse link function** $g^{-1}(\cdot)$ exists and satisfies

$$\mu(\boldsymbol{x}_i) = g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}_i). \tag{12.4}$$

Also notice that the Y_i follow a 1-parameter exponential family where

$$t(y_i) = y_i \text{ and } \mathrm{w}(\theta) = \frac{\mathrm{c}(\theta)}{\mathrm{a}(\phi)},$$

and notice that the value of the parameter $\theta(\boldsymbol{x}_i) = \eta(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i)$ depends on the value of $\boldsymbol{x}_i$. Since the model depends on $\boldsymbol{x}$ only through the linear predictor $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, a GLM is a 1D regression model. Thus the linear predictor is also a sufficient predictor.

The following three sections illustrate three of the most important generalized linear models. After selecting a GLM, the investigator will often want to check whether the model is useful and to perform inference. Several things to consider are listed below.

i) Show that the GLM provides a simple, useful approximation for the relationship between the response variable Y and the predictors $\boldsymbol{x}$.

ii) Estimate α and $\boldsymbol{\beta}$ using maximum likelihood estimators.

iii) Estimate $\mu(\boldsymbol{x}_i) = d_i \tau(\boldsymbol{x}_i)$ or estimate $\tau(\boldsymbol{x}_i)$ where the d_i are known constants.

iv) Check for goodness of fit of the GLM with an estimated sufficient summary plot = response plot.

v) Check for lack of fit of the GLM (eg with a residual plot).

vi) Check for overdispersion with an OD plot.

vii) Check whether Y is independent of $\boldsymbol{x}$; ie, check whether $\boldsymbol{\beta} = \boldsymbol{0}$.

viii) Check whether a reduced model can be used instead of the full model.

ix) Use variable selection to find a good submodel.

x) Predict Y_i given $\boldsymbol{x}_i$.

12.2 Multiple Linear Regression

Suppose that the response variable Y is quantitative. Then the multiple linear regression model is often a very useful model and is closely related to the GLM based on the normal distribution. To see this claim, let $f(y|\mu)$ be the $N(\mu, \sigma^2)$ family of pdfs where $-\infty < \mu < \infty$ and $\sigma > 0$ is known. Recall that μ is the mean and σ is the standard deviation of the distribution. Then the pdf of Y is

$$f(y|\mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right).$$

Since

$$f(y|\mu) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp(\frac{-1}{2\sigma^2}\mu^2)}_{k(\mu)\geq 0} \underbrace{\exp(\frac{-1}{2\sigma^2}y^2)}_{h(y)\geq 0} \exp(\underbrace{\frac{\mu}{\sigma^2}}_{c(\mu)/a(\sigma^2)} y),$$

this family is a 1-parameter exponential family. For this family, $\theta = \mu = E(Y)$, and the known dispersion parameter $\phi = \sigma^2$. Thus $a(\sigma^2) = \sigma^2$ and the canonical link is the **identity link** $c(\mu) = \mu$.

Hence the GLM corresponding to the $N(\mu, \sigma^2)$ distribution with canonical link states that $Y_1, ..., Y_n$ are independent random variables where

$$Y_i \sim N(\mu(\boldsymbol{x}_i), \sigma^2) \text{ and } E(Y_i) \equiv E(Y_i|\boldsymbol{x}_i) = \mu(\boldsymbol{x}_i) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i$$

for $i = 1, ..., n$. This model can be written as

$$Y_i \equiv Y_i|\boldsymbol{x}_i = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i + e_i$$

where $e_i \sim N(0, \sigma^2)$.

When the predictor variables are quantitative, the above model is called a multiple linear regression (MLR) model. When the predictors are categorical, the above model is called an analysis of variance (ANOVA) model, and when the predictors are both quantitative and categorical, the model is called an MLR or analysis of covariance model.

As a GLM, the MLR model states that $Y|SP \sim N(SP, \sigma^2)$, and the assumption that σ^2 is known is too strong. As a semiparametric model, the MLR model states that $Y = SP + e$ where the e_i are iid with zero mean and unknown constant variance σ^2. The semiparametric model is much more important than the GLM because the theory is similar for both models but the semiparametric model does not need the error distribution to be known. The semiparametric MLR model is discussed in detail in Chapters 2 and 3. Semiparametric ANOVA models also have theory similar to the normal GLM, and these models are discussed in Chapters 5 to 9.

12.3 Logistic Regression

Multiple linear regression is used when the response variable is quantitative, but for many data sets the response variable is categorical and takes on two values: 0 or 1. The occurrence of the category that is counted is labelled as a 1 or a "success," while the nonoccurrence of the category that is counted is labelled as a 0 or a "failure." For example, a "success" = "occurrence" could be a person who contracted lung cancer and died within 5 years of detection. For a binary response variable, a binary regression model is often appropriate.

Definition 12.3. The **binomial regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\mathrm{m_i}, \rho(\boldsymbol{x}_\mathrm{i})).$$

The **binary regression model** is the special case where $m_i \equiv 1$ for $i = 1, ..., n$ while the **logistic regression (LR) model** is the special case of binomial regression where

$$P(\text{success}|\boldsymbol{x}_\mathrm{i}) = \rho(\boldsymbol{x}_\mathrm{i}) = \frac{\exp(\alpha + \boldsymbol{\beta}^\mathrm{T}\boldsymbol{x}_\mathrm{i})}{1 + \exp(\alpha + \boldsymbol{\beta}^\mathrm{T}\boldsymbol{x}_\mathrm{i})}. \tag{12.5}$$

If the sufficient predictor $SP = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$, then the most used binomial regression models are such that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{binomial}(\mathrm{m_i}, \rho(\alpha + \boldsymbol{\beta}^\mathrm{T}\boldsymbol{x}_\mathrm{i})),$$

or

$$Y_i|SP_i \sim \text{binomial}(\mathrm{m_i}, \rho(\mathrm{SP_i})). \tag{12.6}$$

Note that the conditional mean function $E(Y_i|SP_i) = m_i\rho(SP_i)$ and the conditional variance function $V(Y_i|SP_i) = m_i\rho(SP_i)(1 - \rho(SP_i))$. Note that the LR model has

$$\rho(SP) = \frac{\exp(SP)}{1 + \exp(SP)}.$$

To see that the binary logistic regression model is a GLM, assume that Y is a binomial$(1, \rho)$ random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \binom{1}{y}\rho^y(1-\rho)^{1-y} = \underbrace{\binom{1}{y}}_{h(y)\geq 0}\underbrace{(1-\rho)}_{k(\rho)\geq 0}\exp[\underbrace{\log(\frac{\rho}{1-\rho})}_{c(\rho)}\,y].$$

Hence this family is a 1-parameter exponential family with $\theta = \rho = E(Y)$ and canonical link

$$c(\rho) = \log\left(\frac{\rho}{1-\rho}\right).$$

This link is known as the *logit link*, and if $g(\mu(\boldsymbol{x})) = g(\rho(\boldsymbol{x})) = c(\rho(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ then the inverse link satisfies

$$g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \frac{\exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})}{1 + \exp(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})} = \rho(\boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the binomial$(1, \rho)$ distribution with canonical link is the binary logistic regression model.

Although the logistic regression model is the most important model for binary regression, several other models are also used. Notice that $\rho(\boldsymbol{x}) = P(S|\boldsymbol{x})$ is the population probability of success S given $\boldsymbol{x}$, while $1 - \rho(\boldsymbol{x}) = P(F|\boldsymbol{x})$ is the probability of failure F given $\boldsymbol{x}$. In particular, for binary regression,

$$\rho(\boldsymbol{x}) = P(Y = 1|\boldsymbol{x}) = 1 - P(Y = 0|\boldsymbol{x}).$$

If this population proportion $\rho = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$, then the model is a 1D regression model. The model is a GLM if the link function g is differentiable and monotone so that $g(\rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$ and $g^{-1}(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) = \rho(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$. Usually the inverse link function corresponds to the cumulative distribution function of a location scale family. For example, for logistic regression, $g^{-1}(x) = \exp(x)/(1 + \exp(x))$ which is the cdf of the logistic $L(0, 1)$ distribution. For probit regression, $g^{-1}(x) = \Phi(x)$ which is the cdf of the Normal $N(0, 1)$ distribution. For the complementary log-log link, $g^{-1}(x) = 1 - \exp[-\exp(x)]$ which is the cdf for the smallest extreme value distribution. For this model, $g(\rho(\boldsymbol{x})) = \log[-\log(1 - \rho(\boldsymbol{x}))] = \alpha + \boldsymbol{\beta}^T\boldsymbol{x}$.

Binomial logistic regression models are discussed in detail in Chapter 10.

12.4 Poisson Regression

If the response variable Y is a count, then the Poisson regression model is often useful. For example, counts often occur in wildlife studies where a region is divided into subregions and Y_i is the number of a specified type of animal found in the subregion.

Definition 12.4. The **Poisson regression model** states that $Y_1, ..., Y_n$ are independent random variables with

$$Y_i \sim \text{Poisson}(\mu(\boldsymbol{x}_{\text{i}})).$$

The **loglinear Poisson regression model** is the special case where

$$\mu(\boldsymbol{x}_i) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}_i). \tag{12.7}$$

To see that the loglinear regression model is a GLM, assume that Y is a Poisson(μ) random variable. For a one parameter family, take $a(\phi) \equiv 1$. Then the pmf of Y is

$$f(y) = P(Y = y) = \frac{e^{-\mu}\mu^y}{y!} = \underbrace{e^{-\mu}}_{k(\mu)\geq 0} \underbrace{\frac{1}{y!}}_{h(y)\geq 0} \exp[\underbrace{\log(\mu)}_{c(\mu)} y]$$

for $y = 0, 1, \ldots,$ where $\mu > 0$. Hence this family is a 1-parameter exponential family with $\theta = \mu = E(Y)$, and the canonical link is the log link

$$c(\mu) = \log(\mu).$$

Since $g(\mu(\boldsymbol{x})) = c(\mu(\boldsymbol{x})) = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, the inverse link satisfies

$$g^{-1}(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \exp(\alpha + \boldsymbol{\beta}^T \boldsymbol{x}) = \mu(\boldsymbol{x}).$$

Hence the GLM corresponding to the Poisson(μ) distribution with canonical link is the loglinear regression model.

Poisson regression models are discussed in detail in Chapter 11.

12.5 Inference and Variable Selection

This section gives a brief discussion of inference and variable selection for GLMs with emphasis on the logistic regression (LR) and loglinear regression (LLR) models. See Chapters 10 and 11 for more details. Inference for these two models is very similar to inference for the multiple linear regression (MLR) model and survival regression models. For all of these models, Y is independent of the $k \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_k)^T$ given the sufficient predictor $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$:

$$Y \perp\!\!\!\perp \boldsymbol{x} | (\alpha + \boldsymbol{\beta}^T \boldsymbol{x}).$$

To perform inference for LR and LLR, computer output is needed. Point estimators for the mean function are important. Given $\boldsymbol{x} = (x_1, ..., x_k)^T$, a major goal of binary logistic regression is to estimate the success probability $P(Y = 1|\boldsymbol{x}) = \rho(\boldsymbol{x})$ with the estimator

$$\hat{\rho}(\boldsymbol{x}) = \frac{\exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}{1 + \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x})}. \tag{12.8}$$

Similarly, a major goal of loglinear regression is to estimate the mean $E(Y|\boldsymbol{x}) = \mu(\boldsymbol{x})$ with the estimator

$$\hat{\mu}(\boldsymbol{x}) = \exp(\hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}). \tag{12.9}$$

Investigators also sometimes test whether a predictor X_j is needed in the model given that the other $k-1$ nontrivial predictors are in the model with the following **4 step Wald test of hypotheses**.
i) State the hypotheses Ho: $\beta_j = 0$ Ha: $\beta_j \neq 0$.
ii) Find the test statistic $z_{o,j} = \hat{\beta}_j / se(\hat{\beta}_j)$ or obtain it from output.
iii) The p–value $= 2P(Z < -|z_{oj}|) = 2P(Z > |z_{oj}|)$. Find the p–value from output or use the standard normal table.
iv) State whether you reject Ho or fail to reject Ho and give a nontechnical sentence restating your conclusion in terms of the story problem.

If Ho is rejected, then conclude that X_j is needed in the GLM model for Y given that the other $k-1$ predictors are in the model. If you fail to reject Ho, then conclude that X_j is not needed in the GLM model for Y given that the other $k-1$ predictors are in the model. Note that X_j could be a very useful GLM predictor, but may not be needed if other predictors are added to the model.

The Wald confidence interval (CI) for β_j can also be obtained from the output: the large sample 100 $(1-\delta)$ % CI for β_j is $\hat{\beta}_j \pm z_{1-\delta/2}\ se(\hat{\beta}_j)$.

For a GLM, often 3 models are of interest: the **full model** that uses all k of the predictors $\boldsymbol{x}^T = (\boldsymbol{x}_R^T, \boldsymbol{x}_O^T)$, the **reduced model** that uses the r predictors $\boldsymbol{x}_R$, and the **saturated model** that uses n parameters $\theta_1, ..., \theta_n$ where n is the sample size. For the full model the $k+1$ parameters $\alpha, \beta_1, ..., \beta_k$ are estimated while the reduced model has $r+1$ parameters. Let $l_{SAT}(\theta_1, ..., \theta_n)$ be the likelihood function for the saturated model and let $l_{FULL}(\alpha, \boldsymbol{\beta})$ be the likelihood function for the full model. Let

$$L_{SAT} = \log\ l_{SAT}(\hat{\theta}_1, ..., \hat{\theta}_n)$$

be the log likelihood function for the saturated model evaluated at the maximum likelihood estimator (MLE) $(\hat{\theta}_1, ..., \hat{\theta}_n)$ and let

$$L_{FULL} = \log\ l_{FULL}(\hat{\alpha}, \hat{\boldsymbol{\beta}})$$

be the log likelihood function for the full model evaluated at the MLE $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$. Then the **deviance**

$$D = G^2 = -2(L_{FULL} - L_{SAT}).$$

The degrees of freedom for the deviance $= df_{FULL} = n - k - 1$ where n is the number of parameters for the saturated model and $k + 1$ is the number of parameters for the full model.

The saturated model for logistic regression states that $Y_1, ..., Y_n$ are independent binomial(m_i, ρ_i) random variables where $\hat{\rho}_i = Y_i/m_i$. The saturated model for loglinear regression states that $Y_1, ..., Y_n$ are independent Poisson(μ_i) random variables where $\hat{\mu}_i = Y_i$.

Assume that the response plot has been made and that the logistic or loglinear regression model fits the data well in that the nonparametric step or lowess estimated mean function follows the estimated model mean function closely and there is no evidence of overdispersion. The deviance test is used to test whether $\boldsymbol{\beta} = \mathbf{0}$. If this is the case, then the predictors are not needed in the GLM model. If $H_o : \boldsymbol{\beta} = \mathbf{0}$ is not rejected, then for loglinear regression the estimator $\hat{\mu} = \overline{Y}$ should be used while for logistic regression

$$\hat{\rho} = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} m_i$$

should be used. Note that $\hat{\rho} = \overline{Y}$ for binary logistic regression.

The 4 step **deviance test** follows.

i) $H_o : \boldsymbol{\beta} = \mathbf{0} \quad H_A : \boldsymbol{\beta} \neq \mathbf{0}$

ii) test statistic $G^2(o|F) = G^2_o - G^2_{FULL}$

iii) The p–value $= P(\chi^2 > G^2(o|F))$ where $\chi^2 \sim \chi^2_k$ has a chi–square distribution with k degrees of freedom. Note that $k = k + 1 - 1 = df_o - df_{FULL} = n - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that there is a GLM relationship between Y and the predictors $X_1, ..., X_k$. If p–value $\geq \delta$, then fail to reject H_o and conclude that there is not a GLM relationship between Y and the predictors $X_1, ..., X_k$.

If the reduced model leaves out a single variable X_i, then the change in deviance test becomes $H_o : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. This change in deviance test is usually better than the Wald test if the sample size n is not large, but for large n the test statistics from the two tests tend to be very similar (asymptotically equivalent tests).

If the reduced model is good, then the **EE plot** of $ESP(R) = \hat{\alpha}_R + \hat{\boldsymbol{\beta}}_R^T \boldsymbol{x}_{Ri}$ versus $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ should be highly correlated with the identity line with unit slope and zero intercept.

After obtaining an acceptable full model where

$$SP = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_R^T \boldsymbol{x}_R + \boldsymbol{\beta}_O^T \boldsymbol{x}_O$$

try to obtain a **reduced model**

$$SP = \alpha + \beta_{R1} x_{R1} + \cdots + \beta_{Rr} x_{Rr} = \alpha_R + \boldsymbol{\beta}_R^T \boldsymbol{x}_R$$

where the reduced model uses r of the predictors used by the full model and $\boldsymbol{x}_O$ denotes the vector of $k - r$ predictors that are in the full model but not the reduced model. For logistic regression, the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Binomial$(m_i, \rho(\boldsymbol{x}_{Ri}))$ while for loglinear regression the reduced model is $Y_i | \boldsymbol{x}_{Ri} \sim$ independent Poisson$(\mu(\boldsymbol{x}_{Ri}))$ for $i = 1, ..., n$.

Assume that the response plot looks good and that there is no evidence of overdispersion. Then we want to test H_o: the reduced model is good (can be used instead of the full model) versus H_A: use the full model (the full model is significantly better than the reduced model). Fit the full model and the reduced model to get the deviances G^2_{FULL} and G^2_{RED}.

The 4 step **change in deviance test** is
i) H_o: the reduced model is good H_A: use the full model
ii) test statistic $G^2(R|F) = G^2_{RED} - G^2_{FULL}$
iii) The p–value $= P(\chi^2 > G^2(R|F))$ where $\chi^2 \sim \chi^2_{k-r}$ has a chi–square distribution with k degrees of freedom. Note that k is the number of nontrivial predictors in the full model while r is the number of nontrivial predictors in the reduced model. Also notice that $k - r = (k+1) - (r+1) = df_{RED} - df_{FULL} = n - r - 1 - (n - k - 1)$.

iv) Reject H_o if the p–value $< \delta$ and conclude that the full model should be used. If p–value $\geq \delta$, then fail to reject H_o and conclude that the reduced model is good.

Next some rules of thumb are given for GLM variable selection. Before performing variable selection, a useful full model needs to be found. The process of finding a useful full model is an iterative process.

The full model will often contain factors and interactions. If w is a nominal variable with J levels, make w into a factor by using use $J-1$ (indicator or) dummy variables $x_{1,w}, ..., x_{J-1,w}$ in the full model. For example, let $x_{i,w} = 1$ if w is at its ith level, and let $x_{i,w} = 0$, otherwise. An interaction is a product of two or more predictor variables. Interactions are difficult to interpret. Often interactions are included in the full model, and then the reduced model without any interactions is tested.

A **scatterplot** of x versus Y is used to visualize the conditional distribution of $Y|x$. A **scatterplot matrix** is an array of scatterplots and is used to examine the marginal relationships of the predictors and response. Place Y on the top or bottom of the scatterplot matrix. Variables with outliers, missing values or strong nonlinearities may be so bad that they should not be included in the full model. Suppose that all values of the variable x are positive. The **log rule** says add $\log(x)$ to the full model if $\max(x_i)/\min(x_i) > 10$. For the binary logistic regression model, it is often useful to mark the plotted points by a 0 if $Y = 0$ and by a + if $Y = 1$.

To make a full model, use the above discussion and then make a response plot to check that the full model is good. The number of predictors in the full model should be much smaller than the number of data cases n. Suppose that the Y_i are binary for $i = 1, ..., n$. Let $N_1 = \sum Y_i =$ the number of 1's and $N_0 = n - N_1 =$ the number of 0's. A rough rule of thumb is that the full model should use no more than $\min(N_0, N_1)/5$ predictors and the final submodel should have r predictor variables where r is small with $r \leq \min(N_0, N_1)/10$. For loglinear regression, a rough rule of thumb is that the full model should use no more than $n/5$ predictors and the final submodel should use no more than $n/10$ predictors.

Variable selection, also called subset or model selection, is the search for a subset of predictor variables that can be deleted without important loss of information. A *model for variable selection* for a GLM can be described by

$$SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x} = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_E^T \boldsymbol{x}_E = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S \tag{12.10}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$ is a $k \times 1$ vector of nontrivial predictors, $\boldsymbol{x}_S$ is a $r_S \times 1$ vector and $\boldsymbol{x}_E$ is a $(k - r_S) \times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model.

Since S is unknown, candidate subsets will be examined. Let $\boldsymbol{x}_I$ be the vector of r terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining terms (out of the candidate submodel). Then

$$SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I + \boldsymbol{\beta}_O^T \boldsymbol{x}_O. \tag{12.11}$$

Definition 12.5. The model with $SP = \alpha + \boldsymbol{\beta}^T \boldsymbol{x}$ that uses all of the predictors is called the *full model.* A model with $SP = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I$ that only uses the constant and a subset $\boldsymbol{x}_I$ of the nontrivial predictors is called a *submodel.* The full model is always a submodel.

Suppose that S is a subset of I and that model (12.10) holds. Then

$$SP = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S = \alpha + \boldsymbol{\beta}_S^T \boldsymbol{x}_S + \boldsymbol{\beta}_{(I/S)}^T \boldsymbol{x}_{I/S} + \boldsymbol{0}^T \boldsymbol{x}_O = \alpha + \boldsymbol{\beta}_I^T \boldsymbol{x}_I \tag{12.12}$$

where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if the set of predictors S is a subset of I. Let $(\hat{\alpha}, \hat{\boldsymbol{\beta}})$ and $(\hat{\alpha}_I, \hat{\boldsymbol{\beta}}_I)$ be the estimates of $(\alpha, \boldsymbol{\beta})$ and $(\alpha, \boldsymbol{\beta}_I)$ obtained from fitting the full model and the submodel, respectively. Denote the ESP from the *full model* by $ESP = \hat{\alpha} + \hat{\boldsymbol{\beta}}^T \boldsymbol{x}_i$ and denote the ESP from the *submodel* by $ESP(I) = \hat{\alpha}_I + \hat{\boldsymbol{\beta}}_I \boldsymbol{x}_{Ii}$.

Definition 12.6. An **EE plot** is a plot of $ESP(I)$ versus ESP.

Variable selection is closely related to the change in deviance test for a reduced model. You are seeking a subset I of the variables to keep in the model. The $AIC(I)$ statistic is used as an aid in backward elimination and forward selection. The full model and the model I_{min} found with the smallest AIC are always of interest. Burnham and Anderson (2004) suggest that if $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline, and models with $\Delta(I) > 10$ should not be used as the final submodel. Create a full model. The full model has a deviance at least as small as that of any submodel. The final submodel should have an EE plot that clusters tightly about the identity line. As a rough rule of thumb, a good submodel I has corr($ESP(I), ESP$) $\geq$ 0.95. Look at the submodel I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Submodel I_I is the initial submodel to examine.

Backward elimination starts with the full model with k nontrivial variables, and the predictor that optimizes some criterion is deleted. Then there are $k - 1$ variables left, and the predictor that optimizes some criterion is deleted. This process continues for models with $k - 2, k - 3, ..., 2$ and 1 predictors.

Forward selection starts with the model with 0 variables, and the predictor that optimizes some criterion is added. Then there is 1 variable in the model, and the predictor that optimizes some criterion is added. This process continues for models with $2, 3, ..., k - 1$ and k predictors. Both forward selection and backward elimination result in a sequence, often different, of k models $\{x_1^*\}, \{x_1^*, x_2^*\}, ..., \{x_1^*, x_2^*, ..., x_{k-1}^*\}, \{x_1^*, x_2^*, ..., x_k^*\}$ = full model.

All subsets variable selection can be performed with the following procedure. Compute the ESP of the GLM and compute the OLS ESP found by the OLS regression of Y on $\boldsymbol{x}$. Check that |corr(ESP, OLS ESP)| $\geq$ 0.95. This high correlation will exist for many data sets. Then perform multiple linear regression and the corresponding all subsets OLS variable selection with the $C_p(I)$ criterion. If the sample size n is large and $C_p(I) \leq 2(r + 1)$ where the subset I has $r + 1$ variables including a constant, then corr(OLS ESP, OLS ESP(I)) will be high by the proof of Proposition 3.2, and hence corr(ESP, ESP(I)) will be high. In other words, if the OLS ESP and GLM ESP are highly correlated, then performing multiple linear regression and the corresponding MLR variable selection (eg forward selection, backward elimination or all subsets selection) based on the $C_p(I)$ criterion may provide many interesting submodels.

Know how to find good models from output. Neither the full model nor the final submodel should show evidence of overdispersion. The following rules of thumb (roughly in order of decreasing importance) may be useful. It is often not possible to have all 10 rules of thumb to hold simultaneously. Let submodel I have $r_I + 1$ predictors, including a constant. Do not use more

predictors than submodel I_I, which has no more predictors than the minimum AIC model. It is possible that $I_I = I_{min} = I_{full}$. Then the submodel I is good if i) the response plot for the submodel looks like the response plot for the full model.
ii) Want corr(ESP,ESP(I)) ≥ 0.95.
iii) The plotted points in the EE plot cluster tightly about the identity line.
iv) Want the p-value ≥ 0.01 for the change in deviance test that uses I as the reduced model.
v) Want $r_I + 1 \leq n/10$, but for binary LR want $r_I + 1 \leq \min(N_1, N_0)/10$ where N_0 is the number of 0s and N_1 is the number of 1s.
vi) Want the deviance $G^2(I)$ close to $G^2(full)$ (see iv): $G^2(I) \geq G^2(full)$ since adding predictors to I does not increase the deviance).
vii) Want AIC(I) $\leq AIC(I_{min}) + 7$ where I_{min} is the minimum AIC model found by the variable selection procedure.
viii) Want hardly any predictors with p-values > 0.05.
ix) Want few predictors with p-values between 0.01 and 0.05.
x) Want $G^2(I) \leq n - r_I - 1 + 3\sqrt{n - r_I - 1}$.

Suppose that the full model is good and is stored in M1. Let M2, M3, M4 and M5 be candidate submodels found after forward selection, backward elimination, et cetera. Make a scatterplot matrix of the ESPs for M2, M3, M4, M5 and M1. Good candidates should have estimated sufficient predictors that are highly correlated with the full model estimated sufficient predictor (the correlation should be at least 0.9 and preferably greater than 0.95). For binary logistic regression, mark the symbols (0 and +) using the response variable Y.

The final submodel should have few predictors, few variables with large Wald p–values (0.01 to 0.05 is borderline), a good response plot, no evidence of overdispersion and an EE plot that clusters tightly about the identity line. If a factor has $J-1$ dummy variables, either keep all $J-1$ dummy variables or delete all $J-1$ dummy variables, do not delete some of the dummy variables.

12.6 Complements

GLMs were introduced by Nelder and Wedderburn (1972). Most of the models in the first 12 chapters of this text are GLMs. Other books on generalized linear models (in roughly decreasing order of difficulty) include McCullagh and Nelder (1989), Fahrmeir and Tutz (2001), Myers, Montgomery and Vining (2002), Dobson and Barnett (2008). Also see Fox (2008), Hardin and Hilbe (2007), Hoffman (2003), Hutcheson and Sofroniou (1999) and Lindsey (2000). Cook and Weisberg (1999a, ch. 21-23) also has an excellent discussion. Texts on categorical data analysis that have useful discussions of GLMs include Agresti (2002), Le (1998), Lindsey (2004), Simonoff (2003) and Powers and Xie (2000) who give econometric applications.

Barndorff-Nielsen (1982) is a very readable discussion of exponential families. Also see Olive (2008, 2009b).

The response plot of the ESP versus Y is crucial for visualizing the GLM. The estimated mean function and a scatterplot smoother (a nonparametric estimator of the mean function) can be added as visual aids. Model and nonparametric estimators estimated SD function can also be computed. Then the estimated mean function ± the estimated SD function can be plotted.

Olive and Hawkins (2005) give a simple all subsets variable selection procedure that can be applied to generalized linear models, such as logistic regression and Poisson regression, using readily available OLS software.

12.7 Problems

PROBLEMS WITH AN ASTERISK * ARE USEFUL.

12.1. Draw a typical response plot for the following models.

a) multiple linear regression

b) logistic regression for a binary response variable

c) loglinear Poisson regression