Chapter 13

Generalised Linear Models for Count Data

13.1 Introduction

We now specifically consider Generalise Linear Models for count data. We consider two models:

1. The Poisson Model
2. The Negative Binomial model.

13.1.1 Poisson

For the first, the model is: $Y_1, \ldots, Y_n$ are independent where $Y_i \sim P(\lambda_i)$ (Poisson with $\mathbb{E}[Y_i] = \lambda_i$). Recall that, for $Y \sim \text{Poiss}(\lambda)$,

$$\mathbb{P}(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda} = \exp\{y \log \lambda - \lambda - \log y!\}$$

so this is an exponential family, with sufficient statistic $T(y) = y$ and canonical parameter $\eta = \log \lambda$, with log partition function $A(\eta) = \lambda = e^{\eta}$.

We restrict attention to the *canonical* link:

$$\log \lambda_i = \sum_{j=1}^{p} x_{ij} \beta_j$$

where there are p known covariates; for observation Y_j, the values of the covariates are $x_{i1}, \ldots, x_{ip}$ and the parameters $\beta_1, \ldots, \beta_p$ are unknown.

If $Y \sim P(\lambda)$ then $\mathbb{E}[Y] = \text{Var}(Y) = \lambda$. The *variance function* for Poisson is therefore $V(\lambda) = \lambda$. If there is sufficient data, then the variance function may be estimated and it may turn out that $\frac{\widehat{V}(\widehat{\lambda})}{\widehat{\lambda}} > 1$. If the variance is *larger* than the mean, then the data is *overdispersed* and the Poisson model is not a good model.

13.1.2 Negative Binomial

The *negative binomial* distribution introduces an additional parameter, which enables modelling of data that is overdispersed. The accuracy of the parameter estimates decreases as more parameters are added; the negative binomial introduces only one extra parameter. The probability function is:

$$\mathbb{P}(Y=y|\lambda,\theta)=\frac{\Gamma(\theta+y)}{\Gamma(\theta)}\frac{1}{y!}\left(\frac{\lambda}{\theta+\lambda}\right)^y\left(\frac{\theta}{\theta+\lambda}\right)^\theta \qquad y=0,1,2,\ldots$$

This is a generalisation of the distribution

$$\mathbb{P}(Y=y|k,p)=\binom{y+k-1}{k-1}(1-p)^y p^k \qquad y=0,1,2,\ldots$$

which represents the number of failures *before* success k, with a sequence of independent Bernoulli trials, with success probability p.

Replacing $(y+k-1)!$ by $\Gamma(y+k)$ and $(k-1)!$ by $\Gamma(k)$, noting that the resulting expressions are well defined for non-integer k, reparametrising (by setting $k=\theta$ and $\lambda=\mathbb{E}[Y]\frac{k(1-p)}{p}=\lambda$), the probability function in terms of λ and θ is obtained. We write this as: $NB(\lambda,\theta)$. The mean and variance are:

$$\mathbb{E}[Y]=\lambda \qquad \mathrm{Var}(Y)=\lambda+\frac{\lambda^2}{\theta}$$

Note here that the overdispersion takes a particular form. When searching for a model, let $V(\lambda)$ denote the variance function. To estimate the θ parameter for a negative binomial, we estimate α in the expression $\widehat{V}(\lambda)=\lambda+\widehat{\alpha}\lambda^2$. If 0 lies within a suitable confidence interval for α, then the Negative Binomial model does not give advantage over the Poisson model.

Rewriting,

$$\mathbb{P}(Y=y|\lambda,\theta)=\frac{\Gamma(\theta+y)}{\Gamma(\theta)y!}\exp\left\{-y\log\left(1+\frac{\theta}{\lambda}\right)-\theta\log\left(1+\frac{\lambda}{\theta}\right)\right\}.$$

Note that this model is not an exponential family and not of the required form for a Generalised Linear Model if θ is taken as a dispersion parameter; it *only* falls into the framework of a GLM if θ is fixed and known. Nevertheless, it falls within the framework of a regular parametric family and, for i.i.d. sampling, satisfies conditions such that $\binom{\widehat{\lambda}}{\widehat{\theta}}$ is consistent and asymptotically normal.

For $Y_1,\ldots,Y_n$ independent, we consider $Y_i\sim NB(\lambda_i,\theta)$ (i.e. each with the same θ parameter). We consider

$$g(\lambda_i):=\log\lambda_i=\sum_{j=0}^{p}x_{ij}\beta_j$$

and specialise to the setting $x_{i0}=1$ so that we may consider

$$\lambda_i = \lambda_0 t_i$$

where $\log \lambda_0 = \beta_0$ and $t_i = \exp\left\{\sum_{j=1}^{p} x_{ij}\beta_j\right\}$.

The log likelihood function is:

$$\log L(\beta, \theta; y_1, \ldots, y_n) = \left(\sum_{j=1}^{n} \log \Gamma(\theta + y_j) - n \log \Gamma(\theta) - \sum_{j=1}^{n} \log y_j!\right)$$
$$- \sum_{i=1}^{n} y_i \log\left(1 + \frac{\theta}{\lambda_0 \exp\left\{\sum_{j=1}^{p} x_{ij}\beta_j\right\}}\right) - \theta \sum_{i=1}^{n} \log\left(1 + \frac{\lambda_0}{\theta} \exp\left\{\sum_{j=1}^{p} x_{ij}\beta_j\right\}\right).$$

Negative binomial modelling is left to the computer lab exercises. The family falls into the GLM framework for a *fixed* θ, but does not fall into the GLM framework if θ is a *free* parameter. Nevertheless, the parameter estimators for (θ, β) can be shown to be asymptotically normal and consistent. The Fisher information and hence the asymptotic covariance can also be computed.

13.2 The Log-linear Model

Consider data which comes from a Poisson distribution with true rate λ. The rate λ may depend on a number of factors. We'll consider linear models, where all factors exert their influence *linearly.* For example, let Y denote the number of days absent from work during the year. This could be related to a number of factors, such as gender, type of work, etc ..... Then it could be reasonable to model $Y \sim P(\lambda)$ where $\lambda = \lambda(x_1, \ldots, x_p)$. For n independent observations $Y = (Y_1, \ldots, Y_n)^t$ where $\lambda_i = \mathbb{E}[Y_i]$, and $\lambda = (\lambda_1, \ldots, \lambda_n)^t$, where

$$\lambda_i = g(x_{i,1}, \ldots, x_{i,p}).$$

We consider only the *canonical link*;

$$\eta_i = \log \lambda_i = \sum_{j=1}^{p} x_{ij}\beta_j.$$

In this model, an additive increase in the independent variables will lead to a multiplicative increase in λ.

The likelihood is given by

$$L = \prod_{i=1}^{k} \frac{\lambda_i^{y_i}}{y_i!} \exp\{-\lambda_i\}$$

$$\log L = C + \sum_{i=1}^{k} (y_i \log \lambda_i - \lambda_i).$$

The saturated model is obtained by estimating each mean by what you observe, or setting $\tilde{\lambda}_i = y_i$ ($\tilde{\lambda}_i$ denotes the ML parameter estimate for the saturated model).

$$\log L_{\text{full}} = c + \sum_{i=1}^{k} (y_i \log y_i - y_i)$$

$$\log L_{\text{max}} = c + \sum_{i=1}^{k} (y_i \log \widehat{\lambda}_i - \widehat{\lambda}_i)$$

where, for example, $\widehat{\lambda}_i = \widehat{\alpha} + \widehat{\beta} x_i$ (this is the MLE for the current model).

$$\text{deviance} = -2 \log \frac{L_{\text{max}}}{L_{\text{full}}} = 2 \sum_{i=1}^{k} y_i \{\log \frac{y_i}{\widehat{\lambda}_i} + (y_i - \widehat{\lambda}_i)\}.$$

The following proposition is useful for *any* linear model connected with Poisson rates.

Proposition 13.1. *If the linear predictor $\eta_i = \sum_{i=1}^{p} x_{ij} \beta_j$ contains a constant term so that*

$$\eta = \beta_0 + \sum_j x_j \beta_j,$$

then the second term in the expression for deviance disappears.

Proof

$$\log L = c + \sum_{i=1}^{k} (y_i \log \lambda_i - \lambda_i)$$

so that

$$\begin{aligned} \frac{\partial}{\partial \beta_j} \log L &= \sum_{i=1}^{k} \left(\frac{y_i}{\lambda_i} - 1 \right) \frac{\partial \lambda_i}{\partial \beta_j} \\ &= \sum_{i=1}^{k} \left(\frac{y_i}{\lambda_i} - 1 \right) \frac{\partial \lambda_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} \\ &= \sum_{i=1}^{k} (y_i - \lambda_i) x_{ij} \end{aligned}$$

so that $\sum_{i=1}^{k} (y_i - \widehat{\lambda}_i) x_{ij} = 0$ for all j. Considering the constant term β_0 is equivalent to the case where $x_{0j} = 1$ and this gives the result. □

If the model contains a constant term, then

$$\text{deviance} = 2 \sum_{i=1}^{k} y_i \log \frac{y_i}{\lambda_i} = 2 \sum_{i=1}^{k} O_i \log \frac{O_i}{E_i}.$$

13.3 Loglinear Model: Variance of Parameter Estimator

Loglinear models, like logistic regression models, are generalised linear models using the canonical link $\eta_i = \log \lambda_i$ where $\lambda_i = \mathbb{E}[Y_i]$ and $\eta_i = \sum_j x_{ij}\beta_j$, $\beta = (\beta_1, \ldots, \beta_p)^t$ is a vector of unknown parameters.

For independent Poisson sampling, the log likelihood involves the parameters of the loglinear model $\log \lambda_i = \sum_j x_{ij}\beta_j$ through

$$\log L(\lambda) = \sum_i y_i \log \lambda_i - \sum_i \lambda_i = \sum_i y_i (\sum_j x_{ij}\beta_j) - \sum_i \exp\left\{\sum_j x_{ij}\beta_j\right\}$$

using $\lambda_i = \exp\{\sum_j x_{ij}\beta_j\}$. Hence:

$$\frac{\partial}{\partial \beta_j} \log L(\lambda) = \sum_i y_i x_{ij} - \sum_i \lambda_i x_{ij}$$

so that, when we equate these derivatives equal to zero, using X to denote the matrix with entries x_{ij},

$$X^t y = X^t \widehat{\lambda}.$$

The matrix of second partial derivatives of the loglikelihood has elements

$$\frac{\partial^2}{\partial \beta_j \partial \beta_k} \log L(\lambda) = -\sum_i x_{ij} \frac{\partial \lambda_i}{\partial \beta_k} = -\sum_i x_{ij} \left\{ \frac{\partial}{\partial \beta_k} \exp\left\{\sum_h x_{ih}\beta_h\right\}\right\} = -\sum_i x_{ij} x_{ik} \lambda_i$$

so that (since this is a *canonical* exponential family) the Fisher information is:

$$I(\beta) = -\nabla_\beta \nabla_\beta \log L(\beta) = X^t \mathrm{Diag}(\lambda) X.$$

This is estimated by:

$$X^t \mathrm{Diag}(\widehat{\lambda}) X,$$

where $\mathrm{Diag}(\widehat{\lambda})$ has the elements of $\widehat{\lambda}$ on the diagonal.

Since this is an exponential family with canonical parametrisation, therefore, asymptotically:

$$I^{1/2}(\beta) \left(\widehat{\beta}_{ML} - \beta\right) \overset{n \to +\infty}{\longrightarrow}_{(d)} N(0, J)$$

where J is the $p \times p$ identity matrix. This is shown later in the lecture.

The estimated covariance matrix of $\widehat{\beta}$ is

$$\mathrm{Cov}(\widehat{\beta}_{ML}) = (X^t \mathrm{Diag}(\widehat{\lambda}) X)^{-1}.$$

13.4 Huber Sandwich Estimators

This section gives an informal account of the so-called 'Huber Sandwich Estimator'. We discuss the algorithm, and mention some of the ways in which it is applied. In brief, under rather stringent conditions, the algorithm can be used to estimate the variance of the MLE when the underlying model is incorrect. However, the algorithm ignores *bias*, which may be appreciable. Thus, results are liable to be misleading.

Let i index observations whose values are y_i . Let $\theta \in \mathbb{R}^p$ be a p-vector of parameters. Let $y \mapsto p_i(y|\theta)$ be a positive density (or mass) function. For example, suppose that y takes only values 0 or 1 (the case dealt with most fully here) and $p_i(0|\theta) > 0$, $p_i(1|\theta) > 0$ and $p_i(0|\theta) + p_i(1|\theta) = 1$. Assume $\theta \mapsto p_i(y|\theta)$ is smooth. Let Y_i be independent with density $p_i(.|\theta)$ (so they are not identically distributed). The data are modelled as observed values of $Y_i : i = 1, \ldots, n$. The likelihood function is therefore

$$L(\theta; y_1, \ldots, y_n) = \prod_{i=1}^{n} p_i(Y_i|\theta).$$

The log likelihood function is therefore

$$\log L(\theta) = \sum_{i=1}^{n} \log p_i(Y_i|\theta).$$

We can write the vector of first derivatives and matrix of second derivatives as:

$$\nabla_\theta \log L(\theta) = \sum_{i=1}^{n} g_i(Y_i|\theta) \qquad \nabla_\theta \nabla_\theta \log L(\theta) = \sum_{i=1}^{n} h_i(Y_i|\theta)$$

where $g_i(y|\theta)$ is the vector $\nabla_\theta \log p_i(y|\theta)$ and $h_i(y|\theta)$ is the matrix with components $\nabla_\theta \nabla_\theta \log p_i(y|\theta)$. Expanding a Taylor series around θ_0,

$$\log L(\theta) = \log L(\theta_0) + (\theta - \theta_0)^t \nabla_\theta \log L(\theta_0) + \frac{1}{2}(\theta - \theta_0)^t \nabla_\theta \nabla_\theta \log L(\theta_0)(\theta - \theta_0) + \ldots$$

We consider asymptotic results, where we ignore higher order terms. The maximum likelihood estimator satisfies $\nabla_\theta \log L(\theta) = 0$, which is (if θ_0 is close to $\widehat{\theta}_{ML}$ and taking Taylor expansion of $\nabla \log L(\theta)$)

$$\nabla_\theta \log L(\theta_0) + \nabla_\theta \nabla_\theta \log L(\theta_0)(\widehat{\theta}_{ML} - \theta_0) = 0$$

giving:

$$\widehat{\theta}_{ML} - \theta_0 = (-\nabla_\theta \nabla_\theta \log L(\theta_0))^{-1} \nabla_\theta \log L(\theta_0).$$

Assuming the quantity $-\frac{1}{n} \nabla_\theta \nabla_\theta \log L(\theta_0)$ can be approximated by:

$$-\frac{1}{n}\nabla_\theta\nabla_\theta \log L(\theta_0) = \frac{1}{n}\sum_{i=1}^n h_i(Y_i|\theta_0) \simeq \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\theta_0}[h_i(Y_i|\theta_0)]$$

(which, under mild assumptions, is simply a law of large numbers)

and using

$$\frac{1}{n}\mathrm{Cov}_{\theta_0}(\nabla_\theta \log L(\theta_0)) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[g_i(Y_i|\theta_0)g_i^t(Y_i|\theta_0)]$$

then:

$$n\mathrm{Cov}_{\theta_0}(\widehat{\theta}_{ML}) \simeq \left(-\frac{1}{n}\nabla_\theta\nabla_\theta \log L(\theta_0)\right)^{-1}\left(\frac{1}{n}\mathrm{Cov}_{\theta_0}(\nabla_\theta \log L(\theta_0))\right)\left(-\frac{1}{n}\nabla_\theta\nabla_\theta \log L(\theta_0)\right)^{-1}.$$

The covariance $\mathrm{Cov}_{\theta_0}(\widehat{\theta})$ is then estimated as

$$\widehat{V} = (-A)^{-1}B(-A)^{-1}$$

where

$$A = \nabla_\theta\nabla_\theta \log L(\widehat{\theta}) \qquad B = \sum_{i=1}^n g_i(Y_i|\widehat{\theta})g_i^t(Y_i|\widehat{\theta}).$$

The quantity $\widehat{V}$ is the *Huber Sandwich Estimator* and the square roots of the diagonal elements of $\widehat{V}$ are the *robust standard errors* or *Huber-White standard errors.*

13.5 Generalised Linear Models: Residuals and Diagnostics

In linear regression, diagnostics are built around residuals and residual sums of squares. For generalised linear models, we look for quantities that can provide similar information. For generalised linear models, there are several different kinds of residuals and hence different qwuantities that are analagous to the residual sum of squares in linear regression analysis.

Pearson Residuals The *Pearson residual* is based on the idea of subtracting the mean and dividing through by the standard deviation;

$$r_i = \frac{y_i - \widehat{\mu}_i}{\sqrt{V(\widehat{\mu}_i)}}.$$

For Bernoulli data, $\widehat{\mu}_i = \widehat{\pi}_i$, the estimated 'success' probability for y_i and $V(\widehat{\pi}_i) = \widehat{\pi}_i(1-\widehat{\pi}_i)$.

Deviance Residuals The *deviance residual* d_i is based on the contribution to the log likelihood. For example, for regression on binary data,

$$\log L_{\text{model}} = \sum_{i=1}^{n} y_i \log \widehat{\pi}_i + (1-y_i)\log(1-\widehat{\pi}_i).$$

By analogy with linear regression, each term in the sum should be equal to $-\frac{1}{2}d_i^2$. Therefore, the deviance residual is defined as:

$$d_i = s_i\sqrt{-2(y_i \log \widehat{\pi}_i + (1-y_i)\log(1-\widehat{\pi}_i))}$$

where

$$s_i = \begin{cases} 1 & y_i = 1 \\ -1 & y_i = 0. \end{cases}$$

Adding the squares of the Pearson residuals gives the *Pearson statistic*

$$X^2 = \sum_{i=1}^{n} r_i^2$$

while

$$D = \sum_{i=1}^{n} d_i^2 = -2\log L_{\text{model}}.$$

For a model with p fitted parameters, these could in principle be compared with χ^2_{n-p}, but this test does not work well in practise.

13.6 Model Checking

Model checking mimics linear regression; recall the algorithm for parameter estimation. For Gaussian linear regression with $\epsilon \sim N(0, \sigma^2 I_n)$, we used the residuals $Y - \widehat{\mu}$ Instead of the vector Y and $\widehat{\mu}$ considered for linear regression, we consider the *adjusted dependent variate* z and the *linear predictor* $\widehat{\eta}$. The residual variance is replaced by an estimate of the dispersion parameter ϕ and the crucial hat matrix is the hat matrix from the weighted linear regression;

$$H = W^{1/2}X(X^tWX)^{-1}X^tW^{1/2}$$

where $W^{1/2} = \text{diag}(W_1^{1/2}, \dots, W_n^{1/2})$. This is simply equivalent to replacing X by $W^{1/2}X$ in linear regression. To close approximation, the vector of Pearson residuals satisfies:

$$V^{-1/2}(\widehat{\mu} - \mu) \simeq HV^{-1/2}(Y - \mu),$$

where $V = \text{diag}(V(\mu_i))$.

Residual Checks Standardised deviance residuals are recommended, plotted against some function of the fitted values. The transform should be to a constant-information scale of the error distribution. This is the *variance stabilising transform* discussed in Statistics (there was an exercise in the tutorials). For example, $2\sqrt{\widehat{\mu}}$ for Poisson errors and $2\sin^{-1}(\sqrt{\widehat{\mu}})$ for Bernoulli errors. On this scale, the errors should appear i.i.d..

Variance Function A plot of the absolute value of the residuals against the fitted values gives an informal check on the adequacy of the assumed variance function. To make a formal test, the variance function can be embedded in a suitable family. For example for Poisson, $V(\mu) = \mu$, so consider the family μ^α and test $H_0 : \alpha = 1$ versus $H_1 : \alpha \neq 1$.

Link Function An informal check involves examining the plot of the adjusted dependent variable z against $\widehat{\eta}$, the estimated linear predictor. If the link function is correct, then these points should lie, at least approximately, on a straight line.

13.7 Asymptotic Normality of Parameter Estimators

In many situations (all that we deal with in this course), the parameter estimators are *asymptotically normal*, under some minor conditions on the Fisher information. We show this for *canonical* link functions. When the link is *canonical*,

$$-\nabla\nabla \log L(\beta) = I(\beta).$$

For the non-canonical case, $-\nabla\nabla \log L(\beta) = I(\beta) + H(\beta)$, where $\mathbb{E}[H(\beta)] = 0$. Additional conditions are needed to establish asymptotic Gaussianity of the parameter estimators; such a condition is that there is an $\alpha > 0$ such that $I(\beta) + H(\beta) - \alpha I(\beta)$ is positive definite for all β. This is (of course) difficult to verify in practise.

For canonical links, the conditions we require are:

1. Let $I_n(\beta)$ denote the Fisher information of the parameter vector β based on n observations and let $\lambda_{\text{min},n}$ denote the smallest eigenvalue of $I_n(\beta_0)$ where β_0 is the true parameter value, then $\lambda_{\text{min},n} \overset{n\to+\infty}{\longrightarrow} +\infty$.

2. Let β_0 denote the true parameter value. For a symmetric non-negative matrix A with decomposition $A = PDP^t$ where D is diagonal, let $A^{1/2} = PD^{1/2}P^t$. where the entries of $D^{1/2}$ are the non-negative square roots of the entries of D. Let

$$N_n(\delta) = \{\beta : \|I_n^{1/2}(\beta_0)(\beta - \beta_0)\| \leq \delta\}$$

Then we require that, for all $\delta > 0$:

$$\max_{\beta \in N_n(\delta)} \|V_n(\beta) - I\| \longrightarrow 0$$

where $V_n(\beta) = I_n^{-1/2}(\beta_0) I_n(\beta) I_n^{-1/2}(\beta_0)$.

Note: Comparison with i.i.d. sampling In 'Statistics', we showed asymptotic normality for the ML estimator of the parameter vector when the sampling was i.i.d. under general conditions, which included canonical exponential families. In this setting, the two conditions are clearly satisfied for a family of full rank; if $I_1(\eta)$ denotes the Fisher information for a single observation, then $I_n(\eta) = nI_1(\eta)$ where $I_n(\eta)$ denotes the Fisher information for n i.i.d. observations. Hence, $\lambda_{\text{min},n} = n\lambda_{\text{min},1}$ and, since the family is of full rank (so that $\lambda_{\text{min},1} > 0$, the first condition is clearly satisfied.

The second follows in the same way; if $I_1(\beta)$ is continuous in β, then $V_n(\beta) = I_1^{-1/2}(\beta_0) I_1(\beta) I_1^{-1/2}(\beta_0)$, which does not depend on n and the continuity condition is satisfied. These two conditions are therefore satisfied for i.i.d. sampling from a canonical exponential family of full rank.

Recall the definition of the *score function*, the gradient of the log-likelihood, which is zero at $\widehat{\beta}_{ML}$. Let

$$U_i(\beta) := \frac{1}{a(\phi)V(\mu_i)g'(\mu_i)}$$

then the score funtion may be written as:

$$s_n(\beta) = \nabla_\beta \log L_n(\beta) = \sum_{i=1}^n U_i(\beta)(y_i - \mu_i(\beta))x_i.$$

Theorem 13.2. *Under the conditions above:*

1.

$$I_n^{-1/2}(\beta_0)s_n(\beta_0) \longrightarrow N(0, I).$$

2.

$$I_n^{1/2}(\beta_0)(\widehat{\beta}_{ML,n} - \beta_0) \longrightarrow N(0, I).$$

Proof For statement 1., we consider the moment generating function and show that it converges to the $N(0, I)$ moment generating function. Let v denote any unit vector; $v'v = 1$. Consider

$$M_n(\delta) = \mathbb{E}\left[\exp\left\{\delta v' I_n^{-1/2}(\beta_0)s_n(\beta_0)\right\}\right]$$

Now use a Taylor expansion and let

$$\beta_n := \beta_0 + \delta I_n^{-1/2} v.$$

Clearly, under the condition that $\lambda_{n,\min} \to +\infty$, $\beta_n \to \beta_0$. The Taylor expansion gives:

$$\log L_n(\beta_n) = \log L_n(\beta_0) + (\beta_n - \beta_0)' s(\beta_0) + \frac{1}{2}(\beta_n - \beta_0)' I_n(\beta_n^*)(\beta_n - \beta_0)$$

for suitable β_n^*. Taking exponentials gives:

$$L_n(\beta_n) = L_n(\beta_0) \exp\{\delta v' I_{nr}^{-1/2\prime} s(\beta_0)\} \exp\{\frac{1}{2}\delta^2 v' V_n(\beta_n^*) v\}$$

The conditions now imply that $\beta_n^* \to \beta_0$ and $V_n(\beta_n^*) \to I$, so that:

$$\mathbb{E}_{\beta_0}\left[\exp\left\{\delta v' I_n^{-1/2} S_n(\beta_0)\right\}\right] \longrightarrow \exp\left\{\frac{1}{2}\delta^2\right\}$$

This is the m.g.f. of a $N(0,1)$, hence 1. follows.

The second part (asymptotic normality of $\widehat{\beta}$) follows simply from the Taylor expansion:

$$0 = s(\widehat{\beta}_{n,ML}) = s(\beta_0) + I_n(\beta^*)(\widehat{\beta}_{n,ML} - \beta_0)$$

for suitable β^*. Therefore

$$I_n^{-1/2}(\beta^*) s(\beta_0) = I_n^{1/2}(\beta^*)(\widehat{\beta}_{n,ML} - \beta_0)$$

and, after some details, the result follows.

□