Chapter 4

Testing and Model Selection in Generalized Linear Models

4.1 Testing in Generalized Linear Models

4.1.1 Wald Test Statistic

– Let assume the random vector $\mathbf{y}$ is belonging to the exponential family of distributions, and let us consider the generalized linear model

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}. \tag{4.1}$$

– Consider general linear hypotheses

$$\begin{aligned} &H_0 : \mathbf{K}'\boldsymbol{\beta} = \mathbf{0}, \\ &H_1 : \mathbf{K}'\boldsymbol{\beta} \neq \mathbf{0}, \qquad \mathbf{K}' \in \mathbb{R}^{q,(p+1)}. \end{aligned}$$

– If the null hypothesis H_0 holds, then approximately

$$\mathbf{K}'\hat{\boldsymbol{\beta}} \sim N(\mathbf{0}, \mathbf{K}'(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{K}), \tag{4.2}$$

and hence furthermore approximately

$$Q = (\mathbf{K}'\hat{\boldsymbol{\beta}})'(\mathbf{K}'(\mathbf{X}'\widehat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{K})^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}} = (\mathbf{K}'\hat{\boldsymbol{\beta}})'\left(\widehat{\mathrm{Cov}}(\mathbf{K}'\hat{\boldsymbol{\beta}})\right)^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}} \sim \chi^2_q, \tag{4.3}$$

where χ^2_q denotes the central χ^2-distribution with $q = \mathrm{rank}(\mathbf{K})$ degrees freedom.

– The Wald test statistic is Q, and the p-value is obtained as the probability $p = P(\chi^2_q > Q_{obs})$, where Q_{obs} is the calculated observed value of the Wald test statistic.

4.1.2 Deviance and Likelihood Ratio Statistic

– The generalized linear model

$$\mathcal{S}: \ g(\boldsymbol{\mu}) = \mathbf{X}_s \boldsymbol{\beta}_s$$

is called the saturated model if $\mathbf{y} = \hat{\boldsymbol{\mu}}_s = g^{-1}(\mathbf{X}_s \hat{\boldsymbol{\beta}}_s))$.

– When $\mathrm{Var}(Y_i) = \phi \cdot v(\mu_i)$ in the exponential family of distributions, i.e., $a(\phi) = \phi$, the difference of the maximum values of the log-likelihood functions of $\mathcal{S}$ and $\mathcal{M}$ is

$$
\begin{aligned}
& 2\left(l(\hat{\boldsymbol{\beta}}_s|\mathbf{y}) - l(\hat{\boldsymbol{\beta}}|\mathbf{y})\right) \\
&= 2\left(\sum_{i=1}^{n} \frac{y_i \widehat{\Theta}_i(\hat{\boldsymbol{\beta}}_s) - b(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}_s))}{\phi} + c(y_i, \phi) - \sum_{i=1}^{n} \frac{y_i \widehat{\Theta}_i(\hat{\boldsymbol{\beta}}) - b(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}))}{\phi} + c(y_i, \phi)\right) \\
&= 2\left(\sum_{i=1}^{n} \frac{y_i(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}_s) - \widehat{\Theta}_i(\hat{\boldsymbol{\beta}})) - b(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}_s)) + b(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}))}{\phi}\right) \\
&= \frac{1}{\phi} \cdot 2\left(\sum_{i=1}^{n} y_i(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}_s) - \widehat{\Theta}_i(\hat{\boldsymbol{\beta}})) - b(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}_s)) + b(\widehat{\Theta}_i(\hat{\boldsymbol{\beta}}))\right) = \frac{1}{\phi} \cdot D(\mathcal{M}), \qquad (4.4)
\end{aligned}
$$

where $D(\mathcal{M})$ is called as the *deviance* of the model $\mathcal{M}$, and $\phi \cdot D(\mathcal{M})$ is called the *scaled deviance*.

– Let us consider the following hypotheses in the partitioned generalized linear model $g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2$, when $\mathrm{Var}(Y_i) = \phi \cdot v(\mu_i)$:

$$\begin{aligned} H_0 &: \text{Model } g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 \text{ holds,} \\ H_1 &: \text{Model } g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 \text{ holds.} \end{aligned}$$

– Then the likelihood ratio statistic

$$\begin{aligned} LR &= -2 \cdot \log\left(\frac{\max_{\boldsymbol{\beta}_1} L_{H_0}(\boldsymbol{\beta}_1|\mathbf{y})}{\max_{\boldsymbol{\beta}_1,\boldsymbol{\beta}_2} L_{H_1}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2|\mathbf{y})}\right) \\ &= -2\left(l_{H_0}(\hat{\boldsymbol{\beta}}_1|\mathbf{y}) - l_{H_1}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2|\mathbf{y})\right) = 2 \cdot l_{H_1}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2|\mathbf{y}) - 2 \cdot l_{H_0}(\hat{\boldsymbol{\beta}}_1|\mathbf{y}) \\ &= 2\left(l_{H_0}(\hat{\boldsymbol{\beta}}_s|\mathbf{y}) - l_{H_0}(\hat{\boldsymbol{\beta}}_1|\mathbf{y})\right) - 2\left(l_{H_1}(\hat{\boldsymbol{\beta}}_s|\mathbf{y}) - l_{H_1}(\hat{\boldsymbol{\beta}}_1,\hat{\boldsymbol{\beta}}_2|\mathbf{y})\right) \\ &= \frac{1}{\phi} \cdot (D(H_0) - D(H_1)) = \frac{1}{\phi} \cdot \Delta D. \end{aligned} \tag{4.5}$$

– Since the likelihood ratio statistic generally follows asymptotically χ^2 distribution when H_0 holds, we have that $LR = \frac{1}{\phi} \cdot \Delta D \sim \chi^2_{(q)}$, when H_0 holds. Degrees of freedom are $q = \mathrm{rank}(\mathbf{X}_2)$.

– For Poisson and Bernoulli distributions, the parameter $\phi = 1$. Hence for those distributions the test statistic for the hypotheses H_0 and H_1 is the difference of the deviances ΔD, and the p-value is obtained as the probability $p = P(\chi^2_{(q)} > \Delta D_{obs})$, where ΔD_{obs} is the calculated observed value of the difference of the deviances.

– Since, when $\mathrm{Var}(Y_i) = \phi v(\mu_i)$, the X^2 statistic has asymptotically the property

$$\sum_{i=1}^{n} \left(\frac{y_i - \hat{\mu}_i}{\sqrt{\phi v(\hat{\mu}_i)}} \right)^2 = \frac{1}{\phi} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{v(\hat{\mu}_i)} = \frac{1}{\phi} X^2 \sim \chi^2_{(n - \mathrm{rank}(\mathbf{X}))}, \tag{4.6}$$

the statistic

$$F = \frac{\frac{1}{\phi} \Delta D / q}{\frac{1}{\phi} X^2 / \left(n - \mathrm{rank}(\mathbf{X})\right)} = \frac{\Delta D / q}{X^2 / \left(n - \mathrm{rank}(\mathbf{X})\right)} = \frac{\Delta D / q}{\tilde{\phi}}, \tag{4.7}$$

follows asymptotically F distribution with the $df_1 = q$ and $df_2 = n - \mathrm{rank}(\mathbf{X})$ degrees of freedoms when H_0 holds.

– Note that $\tilde{\phi}$ is calculated from the model $g(\boldsymbol{\mu}) = \mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2$.

– For those distributions when possible $\phi \neq 1$, the test statistic for the hypotheses H_0 and H_1 is the F statistic , and the p-value is obtained as the probability $p = P(F_{(q, n-(p+1))} > F_{obs})$, where F_{obs} is the calculated observed value of F statistic and $F_{(q,n-(p+1))}$ is the random variable following the F distribution with the $df_1 = q = \mathrm{rank}(\mathbf{X}_2)$ and $df_2 = n - (p+1) = n - \mathrm{rank}(\mathbf{X})$ degrees of freedoms.

Example 4.1.

Consider the dataset butterfat.txt:

```
    Butterfat           Breed    Age
1        3.74        Ayrshire Mature
2        4.01        Ayrshire  2year
3        3.77        Ayrshire Mature
.
99       6.55          Jersey Mature
100      5.72          Jersey  2year

Average butterfat content (percentages) of milk for random samples
of twenty cows (ten two-year old and ten mature (greater than four years old))
from each of five breeds. The data are from Canadian records of
pure-bred dairy cattle.

Butterfat - butter fat content by percentage
Breed - a factor with levels Ayrshire Canadian Guernsey Holstein-Fresian Jersey
Age - a factor with levels 2year Mature
```

Denote the variables as following

$$Y = \text{Butterfat}, \quad X_1 = \text{Breed}, \quad X_2 = \text{Age}.$$

Let us consider modeling the expected value μ_i with the model

$$g(\mu_i) = \beta_0 + \beta_j + \alpha_h,$$

where index j is related to the categories of the variable $X_1 = \mathsf{Breed}$ and index h is related to the categories of the variable $X_2 = \mathsf{Age}$. Let us choose appropriate distribution and link function $g(\mu_i)$ for modeling the values of random variables Y_i. Let us then test at 5% significance level, is the explanatory variable $X_1 = \mathsf{Breed}$ statistically significant variable in the main effect model

$$g(\mu_i) = \beta_0 + \beta_j + \alpha_h.$$

Example 4.2.

The makiwara board can be made in different kinds of wood. In study, it was examined how much a makiwara board bends (in millimeters) of the force of the strike in different tree species. The makiwara boards used in study were made in two different ways. Dataset is given in file makiwaraboard.txt.

```
    WoodType BoardType Deflection
1     Cherry   Stacked      144.3
2     Cherry   Stacked      125.9
3     Cherry   Stacked      263.2
.
335      Oak   Tapered       73.3
336      Oak   Tapered       44.9
```

Denote variables as X_1=WoodType, X_2=BoardType, and Y=Deflection. Consider modeling the response variable $Y_i \sim Gamma(\mu_{jh}, \phi)$ by the following model

$$\mathcal{M}_{1|2}: \quad \log(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h,$$

where index j is related to the categories of the variable X_1=WoodType and index h is related to the categories of the variable X_2=BoardType. Test the hypotheses

$$H_0: \mu_{jh} - \mu_{j_* h_*} = 0,$$
$$H_1: \mu_{jh} - \mu_{j_* h_*} \neq 0,$$

for all possible differences.

4.1.3 Predictive Effect Size Testing

- Consider evaluating the effect on the response variable Y when the values of explanatory variables are changed from the set of values $\mathbf{x}_{1f}$ to the values $\mathbf{x}_{2f}$.
- Let the random variables Y_{1f} and Y_{2f} be unobserved values of the response variable Y given the explanatory values $\mathbf{x}_{1f}$ and $\mathbf{x}_{2f}$, respectively.
- Let us measure the effect size by predicting the difference $Y_{2f} - Y_{1f}$ and testing predictive hypothesis

$$
\begin{aligned}
H_0 &: y_{1f} = y_{2f}, \\
H_1 &: y_{1f} \neq y_{2f}.
\end{aligned}
$$

- Let us denote $\mathbf{y}_f = \binom{Y_{1f}}{Y_{2f}}$ and $\mathbf{X}_f = \begin{pmatrix} \mathbf{x}'_{1f} \\ \mathbf{x}'_{2f} \end{pmatrix}$.
- The maximum likelihood predictor for $\mathbf{y}_f$ is

$$\hat{\mathbf{y}}_f = g^{-1}(\mathbf{X}_f \hat{\boldsymbol{\beta}}). \tag{4.8}$$

- The prediction error $\mathbf{e}_f = \mathbf{y}_f - \hat{\mathbf{y}}_f$ has the covariance matrix

$$\operatorname{Cov}(\mathbf{e}_f) = \mathbf{V}_f + \mathbf{D}_f \mathbf{X}_f \operatorname{Cov}(\hat{\boldsymbol{\beta}}) \mathbf{X}'_f \mathbf{D}_f, \tag{4.9}$$

where $\mathbf{V}_f = \begin{pmatrix} \mathrm{Var}(Y_{1f}) & 0 \\ 0 & \mathrm{Var}(Y_{2f}) \end{pmatrix}$ and $\mathbf{D}_f = \begin{pmatrix} \frac{\partial \mu_{1f}}{\partial \eta_{1f}} & 0 \\ 0 & \frac{\partial \mu_{2f}}{\partial \eta_{2f}} \end{pmatrix}$.

– For the difference $Y_{2f} - Y_{1f}$, the variance of the prediction error $e_f = Y_{2f} - Y_{1f} - (\hat{Y}_{2f} - \hat{Y}_{1f})$ hence is

$$\mathrm{Var}(e_f) = \mathrm{Var}(Y_{1f}) + \mathrm{Var}(Y_{2f}) + \mathbf{k}'\mathbf{D}_f\mathbf{X}_f\,\mathrm{Cov}(\hat{\boldsymbol{\beta}})\mathbf{X}'_f\mathbf{D}_f\mathbf{k}, \qquad \mathbf{k} = \left(\begin{smallmatrix} -1 \\ 1 \end{smallmatrix} \right). \tag{4.10}$$

– After replacing unknown parameters by their estimates, the $100(1-\alpha)\%$ prediction interval for the difference $Y_{2f} - Y_{1f}$ is

$$\left[\hat{Y}_{2f} - \hat{Y}_{1f} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(Y_{1f}) + \widehat{\mathrm{Var}}(Y_{2f}) + \mathbf{k}'\widehat{\mathbf{D}}_f\mathbf{X}_f\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{X}'_f\widehat{\mathbf{D}}_f\mathbf{k}} \right]. \tag{4.11}$$

– Equivalently, the test statistic for the predictive hypothesis is

$$Q = \frac{\hat{Y}_{2f} - \hat{Y}_{1f}}{\sqrt{\widehat{\mathrm{Var}}(Y_{1f}) + \widehat{\mathrm{Var}}(Y_{2f}) + \mathbf{k}'\widehat{\mathbf{D}}_f\mathbf{X}_f\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}})\mathbf{X}'_f\widehat{\mathbf{D}}_f\mathbf{k}}} \tag{4.12}$$

– Similarly as p-value is calculated in testing of parameters, the so-called d-value is obtained for the predictive hypothesis as

$$d = 2 \times P(Z > |Q_{obs}|), \tag{4.13}$$

where $Z \sim N(0,1)$

Example 4.3.

The dream of every Karate Kid is to find such a punching board (makiwara board) that will withstand the blows but which would not be so rigid or hard that training would then harm hands. The makiwara board can be made in different kinds of wood. In study, it was examined how much a makiwara board bends (in millimeters) of the force of the strike in different tree species. The makiwara boards used in study were made in two different ways. Dataset is given in file makiwaraboard.txt.

```
    WoodType BoardType Deflection
1    Cherry   Stacked      144.3
2    Cherry   Stacked      125.9
3    Cherry   Stacked      263.2
.
335     Oak   Tapered       73.3
336     Oak   Tapered       44.9
```

Denote explanatory variables as X_1=WoodType and X_2=BoardType. Consider modeling the response variable Y=Deflection by the following model

$$\mathcal{M}_{1|2}: \quad \log(\mu_{jh}) = \beta_0 + \beta_j + \alpha_h,$$

where index j is related to the categories of the variable X_1=WoodType and index h is related to the categories of the variable X_2=BoardType.

Assume $Y_i \sim Gamma(\mu_{jh}, \phi)$. Let us consider the (predictive) effect size $Y_{2f} - Y_{1f}$ in situation where the explanatory variables are changed from the values

$$X_1 = \textsf{Cherry},$$
$$X_2 = \textsf{Stacked}.$$

to the values

$$X_1 = \textsf{Oak},$$
$$X_2 = \textsf{Tapered}.$$

4.2 Model Selection in Generalized Linear Models

4.2.1 Model Selection within Distribution

– Let us consider the model selection between two competing models when the assume distribution for the random variables Y_i is the same in both models.

– If there is two competing models with different link function,

$$\mathcal{M}_1 : \quad g_1(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$
$$\mathcal{M}_2 : \quad g_2(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta},$$

then the model having smaller Akaike information criterion value, $AIC = 2(p+1) - 2l(\hat{\boldsymbol{\beta}}|\mathbf{y})$, is usually preferred one.

– Hypothesis testing can be used for selection of appropriate explanatory variables in the model. For example, if there is two competing hierarchical models

$$\mathcal{M}_1 : \quad g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1,$$
$$\mathcal{M}_{1|2} : \quad g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2,$$

then the choice of the model can be based on testing the hypotheses

$$H_0 : \text{Model } \mathcal{M}_1 \text{ is the true model,}$$
$$H_1 : \text{Model } \mathcal{M}_{1|2} \text{ is the true model.}$$

– Model selection of two competing hierarchical models

$$\begin{aligned} \mathcal{M}_1: &\quad g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1, \\ \mathcal{M}_{1|2}: &\quad g(\boldsymbol{\mu}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2, \end{aligned}$$

can also be based on AIC values of the models or the coefficient of determination value

$$R^2 = 1 - \frac{D(\mathcal{M})}{D(\mathcal{M}_0)}, \tag{4.14}$$

where $\mathcal{M}_0: \ g(\boldsymbol{\mu}) = \mathbf{1}\beta_0$.

– Stepwise Procedures - Forward Selection and Backward Elimination methods can be used to obtain the best set of explanatory variables.

– The fitted values $\hat{\boldsymbol{\mu}}$ and the raw residuals $\mathbf{e} = \mathbf{y} - \hat{\boldsymbol{\mu}}$ should not have any modellable patterns.

– If the assumption on distribution, the choice of link function, and the selection of explanatory variables have been made successfully, the Pearson residuals

$$o_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathrm{Var}}(Y_i)}} \tag{4.15}$$

follows approximately standard normal distribution.

4.2.2 Model Selection between Distributions

– Let us consider the distribution selection for the model $g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$.

– If there is two competing distributions $f_1(y_i|\boldsymbol{\beta})$ and $f_2(y_i|\boldsymbol{\beta})$, then the mean squared error

$$\mathrm{MSE}(\mathcal{M}) = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{n}$$

can be used to compare the effect of the competing distributions.

– If the Pearson residuals

$$o_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathrm{Var}}(Y_i)}} \tag{4.16}$$

can be modeled by the linear model

$$o_i^2 = \alpha_0 + \alpha_1 \hat{\mu}_i + \epsilon_i \tag{4.17}$$

with $H_0 :\ \alpha_1 = 0$ rejected (Breusch–Pagan test), then the assumed distribution is most likely not correct.

– With correct distributional assumption, the observed prediction interval coverage corresponds with the constructed prediction interval.

Example 4.4.

In biodiesel study, methyl ester was produced from waste canola oil. In experiments, it was measured what kind of effect the factors $X_1 = \mathsf{Time}$ (15,30,45min), $X_2 = \mathsf{Temperature}$ (240,255,270C), and level of Methanol/Oil weight ratio (1,1.5,2), $X_3 = \mathsf{Methanol}$, have on yield of methyl ester, $Y = \mathsf{Yield}$. Data obtained from experiments is available in a file canoladiesel.txt.

```
   Time Temp Methanol Yield
1    15  240      1.0   1.5
2    15  240      1.5   3.2
.
19   45  270      2.0 102.0
```

Let us consider the model

$$\mathcal{M}1_{\mathsf{inverse}}: \quad \frac{1}{\mu_i} = \beta_0 + \beta_1 x_{i1}, \quad \mathcal{M}1_{\log}: \quad \log(\mu_i) = \beta_0 + \beta_1 x_{i1}.$$

Consider the following linear models for Pearson residuals o_i

$$o_i^2 = \alpha_0 + \alpha_1 \hat{\mu}_i + \varepsilon_i$$

and estimate the prediction coverage under different distributional assumptions.