MA50260 Statistical Modelling

Lecture 12: GLM - Model Selection and Modelling Aspects

Ilaria Bussoli

March 15, 2024

Model Comparison

In the last lecture, we compared nested GLMs using the deviance.

- If ϕ is known, we use

$$D(\mathcal{M}_2, \mathcal{M}_1) = 2\left[\ell\left(\hat{\underline{\beta}}^{(2)}\right) - \ell\left(\hat{\underline{\beta}}^{(1)}\right)\right] = \frac{D_1 - D_2}{\phi} \sim \chi^2_{p_2 - p_1}.$$

- If ϕ is unknown, we consider

$$\frac{(D_1 - D_2)/(p_2 - p_1)}{D_2/(n - p_2)} \sim F_{p_2 - p_1, n - p_2}.$$

How do we compare models that are not nested?

AIC and BIC

The AIC and BIC incorporate the complexity to assess model fit.

Akaike's Information Criterion (AIC) is

$$\text{AIC} = -2\,\ell\left(\hat{\underline{\beta}}\right) + 2p,$$

and **Schwarz Information Criterion (BIC)** is

$$\text{BIC} = -2\,\ell\left(\hat{\underline{\beta}}\right) + p\log n,$$

where p is the number of explanatory variables.

For a better fitting model, we want a **lower** AIC or BIC.

Example - Contraceptive Use (I)

Data for $n = 1607$ women in Fiji across multiple age groups

```
##      age education wantsMore notUsing using
## 1    <25       low       yes       53     6
## 2    <25       low        no       10     4
## 3    <25      high       yes      212    52
## 4    <25      high        no       50    10
## 5  25-29       low       yes       60    14
## 6  25-29       low        no       19    10
```

We fit a binomial GLM with the factor **age**:

```
fit1 <- glm(cbind(using, notUsing) ~ age,
            family = binomial, data = cuse)
fit1$deviance
```

```
## [1] 86.58064
```

Example - Contraceptive Use (II)

Let's check if we should another explanatory variable

```
add1(fit1, ~. + education + wantsMore, test = "Chisq")
```

```
## Single term additions
##
## Model:
## cbind(using, notUsing) ~ age
##             Df Deviance    AIC    LRT  Pr(>Chi)
## <none>           86.581 166.09
## education  1   80.418 161.93  6.162   0.01305 *
## wantsMore  1   36.888 118.40 49.693 1.798e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Example - Contraceptive Use (III)

Results indicate that both could be added

```
fit2 <- update(fit1, ~. + education + wantsMore)
```

Do we improve the model fit?

```
fit2$deviance
```

```
## [1] 29.91722
```

What are the estimates?

```
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.8082     0.1590 -5.0832   0.0000
## age25-29         0.3894     0.1759  2.2143   0.0268
## age30-39         0.9086     0.1646  5.5194   0.0000
## age40-49         1.1892     0.2144  5.5460   0.0000
## educationlow    -0.3250     0.1240 -2.6202   0.0088
## wantsMoreyes    -0.8330     0.1175 -7.0908   0.0000
```

Forward and Backward selection

How can we select the best fitting model?

Forward and Backward selection

How can we select the best fitting model?

There are generally two strategies:

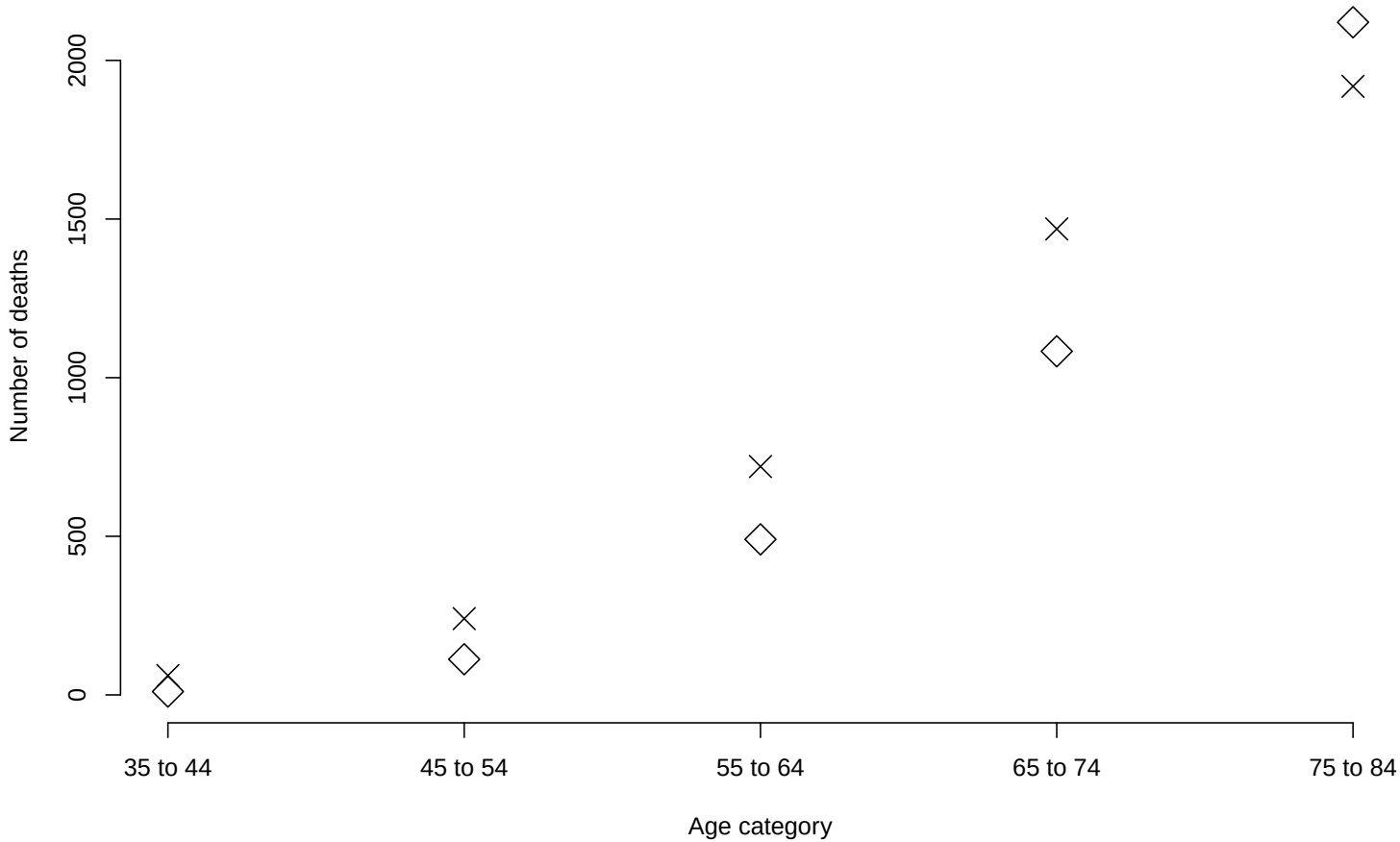
Forward Selection:
Start with the simplest model and add single explanatory variables sequentially to see if they improve the model.

Backward Selection:
Start with the full model and remove single explanatory variables sequentially, and check whether the model fit changes substantially.

Doctor Deaths - Nonlinear Predictors (I)

```
##          age     smoking deaths person-years
## 1   35 to 44      smoker     32         52407
## 2   45 to 54      smoker    104         43248
## 3   55 to 64      smoker    206         28612
## 4   65 to 74      smoker    186         12663
## 5   75 to 84      smoker    102          5317
## 6   35 to 44 non-smoker      2         18790
## 7   45 to 54 non-smoker     12         10673
## 8   55 to 64 non-smoker     28          5710
## 9   65 to 74 non-smoker     28          2585
## 10 75 to 84 non-smoker     31          1462
```

Doctor Deaths - Nonlinear Predictors (II)



Doctor Deaths - Nonlinear Predictors (III)

For the Poisson regression model, we have

```
docglma <- glm(deaths ~ agecat + smoking,
               family = poisson(), data = doctors)
deviance(docglma)
```

```
## [1] 157.5874
```

We get good improvements by adding a quadratic term:

```
docglmb <- glm(deaths ~ agecat + I(agecat^2) + smoking,
               family = poisson(), data = doctors)
deviance(docglmb)
```

```
## [1] 14.65234
```

Doctor Deaths - Varying Exposure

So far, the Poisson models ignore the varying number of doctors across groups.

Doctor Deaths - Varying Exposure

So far, the Poisson models ignore the varying number of doctors across groups.

We account for this **varying exposure** by defining

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i) \\ \log(\mu_i) &= \log(u_i) + \mathbf{x}_i^{\mathrm{T}} \underline{\beta}. \end{aligned}$$

The term $\log(u_i)$ is called the **offset**.

This new model achieves the lowest deviance

```
## [1] 12.17555
```

Further improvements are achieved by including an interaction effect between age and smoking status.

Overdispersion (I)

Observations are **overdispersed** if their variance is much higher than their mean.

Overdispersion (I)

Observations are **overdispersed** if their variance is much higher than their mean.

A high deviance despite a good model fit can signify overdispersion.

Overdispersion (I)

Observations are **overdispersed** if their variance is much higher than their mean.

A high deviance despite a good model fit can signify overdispersion.

We can look for this feature via the standardized residuals

$$r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

and check if they are greater than 1, or via

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} (r_i^P)^2,$$

which we can compare to a χ^2_{n-p} distribution.

Overdispersion (II)

How can we address the issue of overdispersion in a Poisson or binomial model?

- Fit a quasi-Poisson or quasi-binomial model with

$$\mathrm{Var}(Y_i) = \phi V(\mu_i) = \phi \mu_i.$$

- Use a negative-binomial distribution with

$$f(y \mid \mu, \vartheta) = \frac{\Gamma(y + \vartheta)}{\Gamma(\vartheta) y!} \left(\frac{\mu}{mu + \vartheta} \right)^y \left(\frac{\vartheta}{mu + \vartheta} \right)^\vartheta,$$

The first option prevents us from using the likelihood summaries (e.g., AIC), but we can still use the deviance for model comparison.

Example - Disease Incidents

Let's fit a Poisson model to the **citydisease** data set

```
fitp <- glm(Incidents ~ Month, family = "poisson",
            data = citydisease)
sum( (citydisease$Incidents - fitted(fitp) )^2 /
       fitted(fitp) ) / 18
```

```
## [1] 2.609231
```

```
fitp$deviance
```

```
## [1] 45.70303
```

The estimate $\hat{\phi}_P > 1$. So let's try a negative-binomial model

```
fitnb2 <- glm.nb(Incidents ~ Month, data = citydisease)
fitnb2$deviance
```

```
## [1] 26.99114
```

This model reduces the deviance from 45.7 to 27.0.