

# *Nested Model Comparison in General and Generalized Linear Models*

Mark Andrews  
Psychology Department, Nottingham Trent University

`mark.andrews@ntu.ac.uk`

## What are nested models

- ▶ Model  $M_0$  is nested in model  $M_1$  if the parameter space of  $M_0$  is a subset of the parameter space of  $M_1$ .
- ▶ For example, if  $M_0$  is the following linear model:

$$\text{for } i \in 1 \dots n, \quad y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

its parameter space is  $\beta_0, \beta_1, \sigma^2$ .

- ▶ If  $M_1$  is the following linear model:

$$\text{for } i \in 1 \dots n, \quad y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2),$$

its parameter space is  $\beta_0, \beta_1, \beta_2, \sigma^2$ .

- ▶ Any set of values of  $\beta_0, \beta_1, \sigma^2$  in  $M_0$  is a point in the parameter space of  $\beta_0, \beta_1, \beta_2, \sigma^2$  of  $M_1$  if we simply set  $\beta_2 = 0$ .
- ▶ In other words, we can make  $M_0$  with any given values of  $\beta_0, \beta_1, \sigma^2$  from  $M_1$  by setting  $\beta_0, \beta_1, \sigma^2$  in  $M_1$  to these same values and setting  $\beta_2 = 0$ .

## *Nested normal linear models*

- ▶ We can compare nested normal linear models using F tests.
- ▶ Assume  $M_0$  and  $M_1$  are normal linear models, with  $M_0$  nested in  $M_1$ .
- ▶ We calculate  $RSS_0$  and  $RSS_1$ , the residual sums of squares of  $M_0$  and  $M_1$ , respectively.
- ▶  $RSS_0$  will be greater than or equal to  $RSS_1$ .
- ▶ Then

$$\text{proportional increase in error} = \frac{\text{increase in error}}{\text{minimal error}},$$

$$= \frac{RSS_0 - RSS_1}{RSS_1},$$

## *Nested normal linear models*

```
M1 <- lm(Fertility ~ Agriculture + Education + Catholic, data =  
M0 <- lm(Fertility ~ Agriculture + Education, data = swiss)
```

```
RSS_0 <- sum(residuals(M0)^2)  
RSS_1 <- sum(residuals(M1)^2)
```

```
c(RSS_0, RSS_1)  
#> [1] 3953.270 2567.884
```

```
(RSS_0 - RSS_1)/RSS_1  
#> [1] 0.5395049
```

In other words,  $RSS_0$  is 1.54 greater than  $RSS_1$ .

## *Nested normal linear models*

- The F ratio is

$$F = \underbrace{\frac{RSS_0 - RSS_1}{RSS_1}}_{\text{effect size}} \times \underbrace{\frac{df_1}{df_0 - df_1}}_{\text{sample size}} = \frac{(RSS_0 - RSS_1)/(df_0 - df_1)}{RSS_1/df_1}.$$

where  $df_1$  is  $N - (K_1 + 1)$ , where  $K_1$  is number of (predictor; excluding intercept) coefficients in  $M_1$ .

```
df_0 <- M0$df.residual
df_1 <- M1$df.residual
c(df_0, df_1, df_0 - df_1, df_1/(df_0 - df_1))
#> [1] 44 43 1 43
```

## *Nested normal linear models*

```
RSS_0
#> [1] 3953.27
RSS_1
#> [1] 2567.884
RSS_0 - RSS_1
#> [1] 1385.386
df_0 - df_1
#> [1] 1
df_1
#> [1] 43
(RSS_0 - RSS_1)/(df_0 - df_1)
#> [1] 1385.386
RSS_1/df_1
#> [1] 59.71823
((RSS_0 - RSS_1)/(df_0 - df_1))/(RSS_1/df_1)
#> [1] 23.19871
```

## Nested normal linear models

```
anova(M0, M1)
#> Analysis of Variance Table
#>
#> Model 1: Fertility ~ Agriculture + Education
#> Model 2: Fertility ~ Agriculture + Education + Catholic
#>   Res.Df    RSS Df Sum of Sq    F    Pr(>F)
#> 1      44 3953.3
#> 2      43 2567.9  1    1385.4 23.199 1.842e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Nested normal linear models

```
drop1(M1, scope = ~ Catholic, test = 'F')
#> Single term deletions
#>
#> Model:
#> Fertility ~ Agriculture + Education + Catholic
#>           Df Sum of Sq    RSS    AIC F value    Pr(>F)
#> <none>                2567.9 196.03
#> Catholic  1      1385.4 3953.3 214.31  23.199 1.842e-05 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



## Nested normal linear models

```
drop1(M0, scope = ~ Education, test = 'F')
#> Single term deletions
#>
#> Model:
#> Fertility ~ Agriculture + Education
#>
#>           Df Sum of Sq    RSS    AIC F value    Pr(>F)
#> <none>                3953.3 214.31
#> Education  1      2329.8 6283.1 234.09  25.931 7.105e-06 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Nested normal linear models

```
drop1(M0, scope = ~ Education + Agriculture, test = 'F')
#> Single term deletions
#>
#> Model:
#> Fertility ~ Agriculture + Education
#>
#>           Df Sum of Sq    RSS    AIC F value    Pr(>F)
#> <none>                3953.3 214.31
#> Education      1    2329.85 6283.1 234.09 25.9312 7.105e-06 ***
#> Agriculture    1      61.97 4015.2 213.04  0.6897  0.4108
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

## Nested normal linear models

```
anova(M0)
#> Analysis of Variance Table
#>
#> Response: Fertility
#>
#>           Df Sum Sq Mean Sq F value    Pr(>F)
#> Agriculture  1   894.8   894.84   9.9596 0.002886 **
#> Education    1 2329.8 2329.85 25.9312 7.105e-06 ***
#> Residuals   44 3953.3    89.85
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- If we have two models,  $M_0$  and  $M_1$ , with  $M_0$  nested in  $M_1$ , and with residual sums of squares  $RSS_0$  and  $RSS_1$ , respectively, we can calculate:

$$\begin{aligned}\text{proportional decrease in error} &= \frac{\text{decrease in error (from } M_0 \text{ to } M_1)}{\text{error in } M_0}, \\ &= \frac{RSS_0 - RSS_1}{RSS_0}, \\ &= R^2\end{aligned}$$

```
(RSS_0 - RSS_1) / RSS_0
```

```
#> [1] 0.3504405
```

- In other words, the reduction in error from  $M_0$  to  $M_1$  is 0.35 of the error of  $M_0$ .

## $R^2$ : The coefficient of determination

- It can be shown that

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{RSS}},$$

where TSS is *total* sum of squares, ESS is *explained* sum of squares, and RSS is *residual* sum of squares.

- The coefficient of determination  $R^2$  is defined as

$$\begin{aligned} R^2 &= \frac{\text{ESS}}{\text{TSS}} = \text{Proportion of variation that is explained,} \\ &= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \end{aligned}$$

## $R^2$

- ▶ If  $M_0$  is a *null* model, i.e. no predictors, then  $TSS = RSS_0$ .
- ▶ It can be shown that

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{RSS_0} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{RSS_0 - RSS_1} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{RSS_1}.$$

- ▶ As such,  $R^2$  is defined as

$$R^2 = \frac{RSS_0 - RSS_1}{RSS_0} = 1 - \frac{RSS_1}{RSS_0},$$

or 1 minus the error of  $M_1$  relative to  $M_0$ .

```
M_null <- lm(Fertility ~ 1, data = swiss)
RSS_null <- sum(residuals(M_null)^2)
RSS_0 / RSS_null
#> [1] 0.5507516
1 - RSS_0 / RSS_null
#> [1] 0.4492484
(RSS_null - RSS_0) / RSS_null
#> [1] 0.4492484
summary(M0)$r.squared
#> [1] 0.4492484
```

## Adjusted $R^2$

- By explaining proportion of variance explained,  $R^2$  is used a *goodness of fit* measure.
- However,  $R^2$  will always grow with  $K$ , the number of predictors.
- $R^2$  can be *adjusted* to counteract the artificial effect of increasing numbers of predictors as follows:

$$R^2_{\text{Adj}} = 1 - \underbrace{\frac{\text{RSS}}{\text{TSS}}}_{R^2} \underbrace{\frac{n-1}{n-K-1}}_{\text{penalty}},$$

where  $n$  is sample size.

- $R^2_{\text{Adj}}$  is not identical to the proportion of variance explained in the *sample*, but is an unbiased measured of the population  $R^2$ .



## Adjusted $R^2$

```
n <- nrow(M0$model)
K <- length(coef(M0)) - 1 # no. of predictor coefs
penalty <- (n - 1)/(n - K - 1)
1 - (RSS_0 / RSS_null) * penalty
#> [1] 0.4242143
summary(M0)$adj.r.squared
#> [1] 0.4242143
```

# Deviance

- ▶ The *deviance* of a model is defined

$$-2 \log L(\hat{\beta}|\mathcal{D}),$$

where  $\hat{\beta}$  are the mle estimates.

- ▶ The better the model fit, the *lower* the deviance.
- ▶ This can be seen as equivalent to RSS for generalized linear models.

## Deviance

```
swiss_df <- mutate(swiss, y = Fertility > median(Fertility))
M1 <- glm(y ~ Agriculture + Education + Catholic,
          data = swiss_df, family = binomial())
M0 <- glm(y ~ Agriculture + Education, data = swiss_df,
          family = binomial())

D_0 <- deviance(M0)
D_1 <- deviance(M1)

c(D_0, D_1)
#> [1] 54.37716 40.70328
(D_0 - D_1) / D_1 # prop. incr. error
#> [1] 0.3359405
(D_0 - D_1) / D_0 # equiv to  $R^2$ ?
#> [1] 0.2514637
```

## *Model comparison with deviance*

- ▶ Let us assume we have two models:  $M_1$  and  $M_0$  where  $M_0$  is nested in  $M_1$ .
- ▶ The deviance of  $M_0$  minus the deviance of the  $M_1$  is

$$\Delta_D = D_0 - D_1.$$

- ▶ Under the null hypothesis,  $\Delta_D$  is distributed as  $\chi^2$  with  $K_1 - K_0$  df, where  $K_1$  is the number of parameters in  $M_1$  and  $K_0$  is the number of parameters in  $M_0$ .

## Model comparison with deviance

```
K_0 <- length(coef(M0))
K_1 <- length(coef(M1))

pchisq(D_0 - D_1,
      df = K_1 - K_0,
      lower.tail = F)
#> [1] 0.0002174578
anova(M0, M1, test = 'Chisq')
#> Analysis of Deviance Table
#>
#> Model 1: y ~ Agriculture + Education
#> Model 2: y ~ Agriculture + Education + Catholic
#>   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
#> 1          44      54.377
#> 2          43      40.703  1   13.674 0.0002175 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```