Lecture 12: Review

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Modeling Approach Overview

- Model specification
  A model is specified in two parts: an equation linking the response and explanatory variables and the probability distribution of the response variable.

  Example: The equation linking each response variable Y and a set of explanatory variables $x_1, \ldots, x_m$ has the form

$$g(E(Y)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m.$$

- Estimation
  Estimate the parameters of the model.
- Adequacy
  Checking the adequacy of the model–how well it fits or summarizes the data.
- Inference
  Calculating confidence intervals, testing hypotheses about the parameters in the model and interpreting the results.
- Prediction
  Predict based on the model you developed, and assess the prediction accuracies based on different metrics.

Concept 1: Exponential family

Consider a single random variable Y whose probability distribution depends on a single parameter θ. The distribution belongs to the exponential family if it can be written in the form

$$f(y;\theta) = \exp(a(y)b(\theta) + c(\theta) + d(y)),$$

where $c(\theta) = log(t(\theta))$ and $d(y) = log(s(y))$.

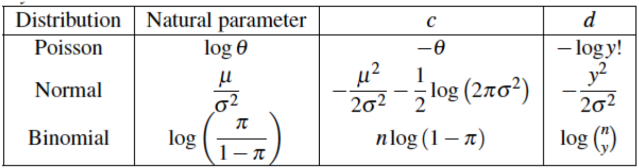
- Canonical (i.e. standard) form:

$$a(y) = y.$$

- Natural parameter of the distribution:

$$b(\theta).$$

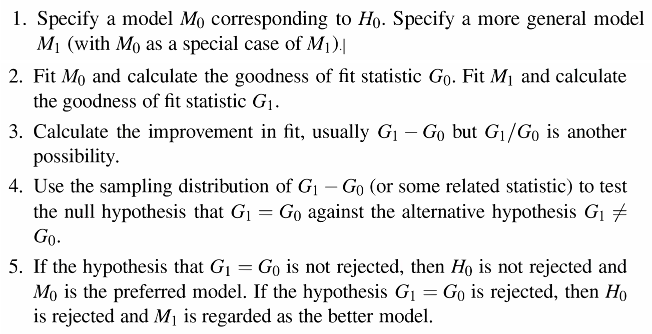
Many well-known distributions belong to the exponential family. For example, the Poisson, Normal and Binomial distributions can all be written in the canonical form.



Distribution	Natural parameter	c	d
Poisson	$\log\theta$	$-\theta$	$-\log y!$
Normal	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\left(2\pi\sigma^2\right)$	$-\frac{y^2}{2\sigma^2}$
Binomial	$\log\left(\frac{\pi}{1-\pi}\right)$	$n\log\left(1-\pi\right)$	$\log\binom{n}{y}$

Concept 2: Inference

The process and logic can be summarized as follows:



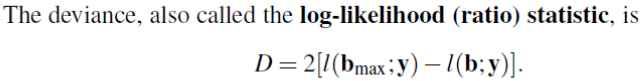
1. Specify a model M_0 corresponding to H_0. Specify a more general model M_1 (with M_0 as a special case of M_1).
2. Fit M_0 and calculate the goodness of fit statistic G_0. Fit M_1 and calculate the goodness of fit statistic G_1.
3. Calculate the improvement in fit, usually $G_1 - G_0$ but G_1/G_0 is another possibility.
4. Use the sampling distribution of $G_1 - G_0$ (or some related statistic) to test the null hypothesis that $G_1 = G_0$ against the alternative hypothesis $G_1 \neq G_0$.
5. If the hypothesis that $G_1 = G_0$ is not rejected, then H_0 is not rejected and M_0 is the preferred model. If the hypothesis $G_1 = G_0$ is rejected, then H_0 is rejected and M_1 is regarded as the better model.

Concept 2: Inference

- Wald statistic (one-parameter)

$$\frac{b-\beta}{se(b)} \sim N(0,1).$$

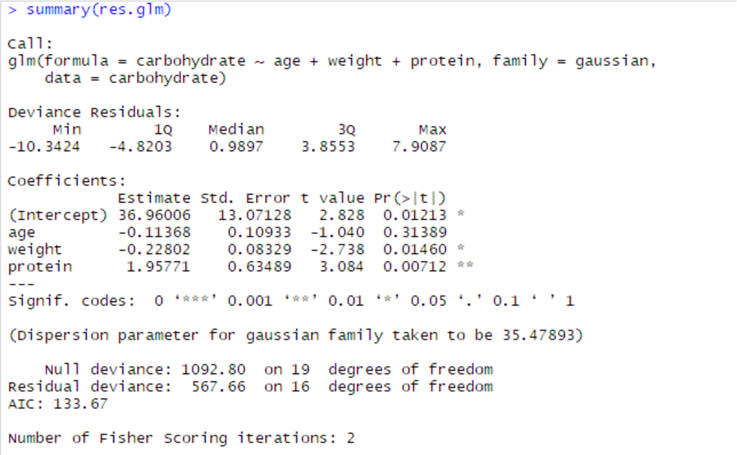
- Saturated model: One way of assessing the adequacy of a model is to compare it with a more general model with the **maximum number of parameters** that can be estimated.
- Deviance:



The deviance, also called the **log-likelihood (ratio) statistic**, is

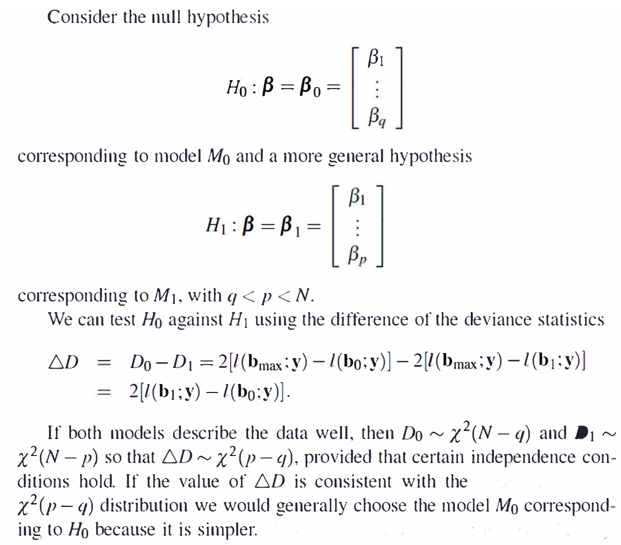
$$D = 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b};\mathbf{y})].$$

Example



```
> summary(res.glm)

Call:
glm(formula = carbohydrate ~ age + weight + protein, family = gaussian,
    data = carbohydrate)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-10.3424   -4.8203    0.9897    3.8553    7.9087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.96006   13.07128   2.828  0.01213 *
age         -0.11368    0.10933  -1.040  0.31389
weight      -0.22802    0.08329  -2.738  0.01460 *
protein      1.95771    0.63489   3.084  0.00712 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 35.47893)

    Null deviance: 1092.80  on 19  degrees of freedom
Residual deviance:  567.66  on 16  degrees of freedom
AIC: 133.67

Number of Fisher Scoring iterations: 2
```

Hypothesis testing



Consider the null hypothesis

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_q \end{bmatrix}$$

corresponding to model M_0 and a more general hypothesis

$$H_1 : \boldsymbol{\beta} = \boldsymbol{\beta}_1 = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

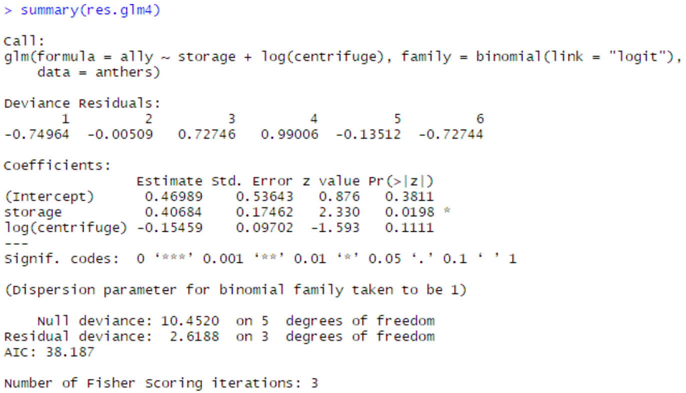
corresponding to M_1, with $q < p < N$.

We can test H_0 against H_1 using the difference of the deviance statistics

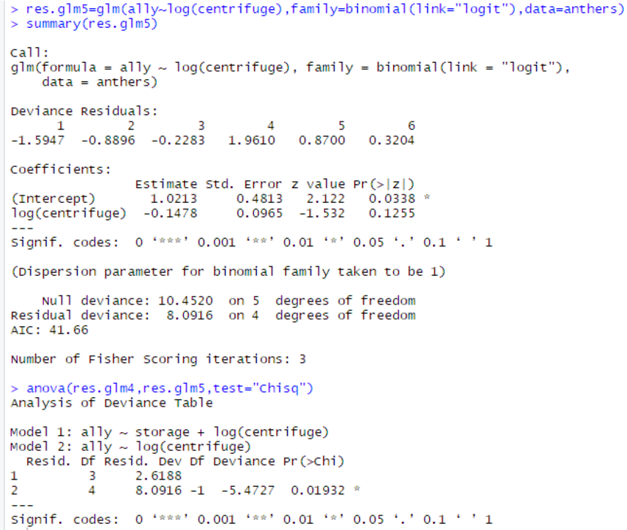
$$\begin{aligned} \triangle D &= D_0 - D_1 = 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b}_0;\mathbf{y})] - 2[l(\mathbf{b}_{\max};\mathbf{y}) - l(\mathbf{b}_1;\mathbf{y})] \\ &= 2[l(\mathbf{b}_1;\mathbf{y}) - l(\mathbf{b}_0;\mathbf{y})]. \end{aligned}$$

If both models describe the data well, then $D_0 \sim \chi^2(N-q)$ and $D_1 \sim \chi^2(N-p)$ so that $\triangle D \sim \chi^2(p-q)$, provided that certain independence conditions hold. If the value of $\triangle D$ is consistent with the $\chi^2(p-q)$ distribution we would generally choose the model M_0 corresponding to H_0 because it is simpler.

Binomial and Poisson

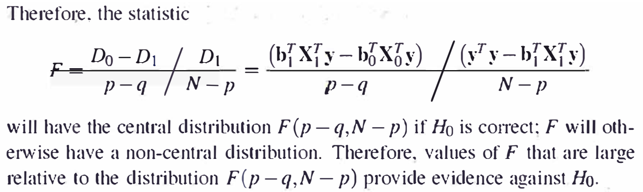


```
> summary(res.glm4)

Call:
glm(formula = ally ~ storage + log(centrifuge), family = binomial(link = "logit"),
    data = anthers)

Deviance Residuals:
       1         2         3         4         5         6
-0.74964  -0.00509   0.72746   0.99006  -0.13512  -0.72744

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.46989    0.53643   0.876   0.3811
storage          0.40684    0.17462   2.330   0.0198 *
log(centrifuge) -0.15459    0.09702  -1.593   0.1111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.4520  on 5  degrees of freedom
Residual deviance:  2.6188  on 3  degrees of freedom
AIC: 38.187

Number of Fisher Scoring iterations: 3
```



```
> res.glm5=glm(ally~log(centrifuge),family=binomial(link="logit"),data=anthers)
> summary(res.glm5)

Call:
glm(formula = ally ~ log(centrifuge), family = binomial(link = "logit"),
    data = anthers)

Deviance Residuals:
      1        2        3        4        5        6  
-1.5947  -0.8896  -0.2283   1.9610   0.8700   0.3204  

Coefficients:
                Estimate Std. Error z value Pr(>|z|)  
(Intercept)       1.0213     0.4813   2.122   0.0338 *
log(centrifuge)  -0.1478     0.0965  -1.532   0.1255  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 10.4520  on 5  degrees of freedom
Residual deviance:  8.0916  on 4  degrees of freedom
AIC: 41.66

Number of Fisher Scoring iterations: 3

> anova(res.glm4,res.glm5,test="Chisq")
Analysis of Deviance Table

Model 1: ally ~ storage + log(centrifuge)
Model 2: ally ~ log(centrifuge)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1         3     2.6188                       
2         4     8.0916 -1  -5.4727  0.01932 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

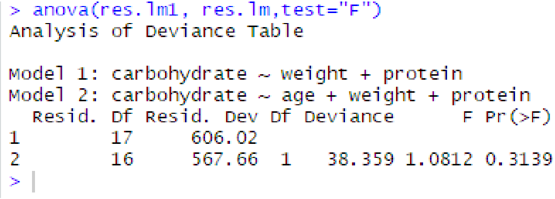
Normal distribution: F-test



Therefore, the statistic

$$F = \frac{D_0 - D_1}{p-q} \Big/ \frac{D_1}{N-p} = \frac{(\mathbf{b}_1^T \mathbf{X}_1^T \mathbf{y} - \mathbf{b}_0^T \mathbf{X}_0^T \mathbf{y})}{p-q} \Big/ \frac{(\mathbf{y}^T \mathbf{y} - \mathbf{b}_1^T \mathbf{X}_1^T \mathbf{y})}{N-p}$$

will have the central distribution $F(p-q, N-p)$ if H_0 is correct; F will otherwise have a non-central distribution. Therefore, values of F that are large relative to the distribution $F(p-q, N-p)$ provide evidence against H_0.



```
> anova(res.lm1, res.lm,test="F")
Analysis of Deviance Table

Model 1: carbohydrate ~ weight + protein
Model 2: carbohydrate ~ age + weight + protein
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1        17     606.02
2        16     567.66  1   38.359 1.0812 0.3139
> 
```

Concept 3: Residuals

- Pearson, or chi-squared, residual

$$X_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}.$$

- Standardized Pearson residuals

$$r_{Pk} = \frac{X_k}{1-h_k},$$

  where h_k is the leverage.

- Deviance residuals

$$d_i = \mathrm{sign}(y_i - n_i\hat{\pi}_i)\left(2\left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right]\right)^{1/2}.$$

- Deviance residual

$$r_{D_i} = \frac{d_i}{1-h_i}.$$

Concept 4: Goodness of fit statistics

- Pearson chi-squared statistic

$$X^2 = \sum \frac{(O - e)^2}{e}.$$

- For the binomial regression

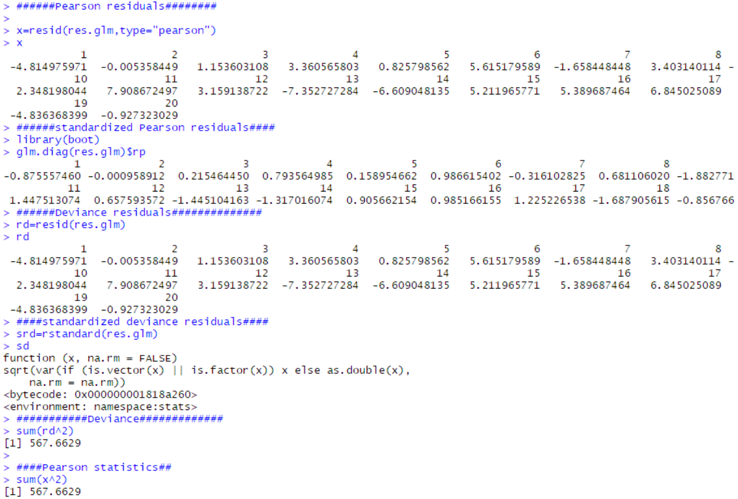
$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\pi_i)^2}{n_i\pi_i(1 - \pi_i)}.$$

- Estimated expected frequencies

$$X^2 = \sum_{i=1}^{N} \frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1 - \hat{\pi}_i)}.$$

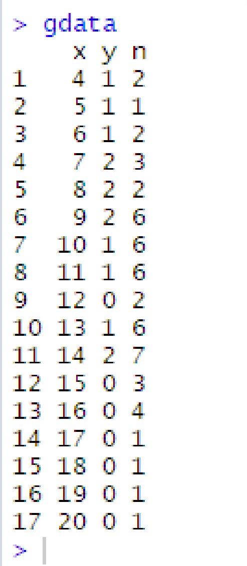
- Asymptotically equivalent to the deviances

$$D = 2\sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right].$$

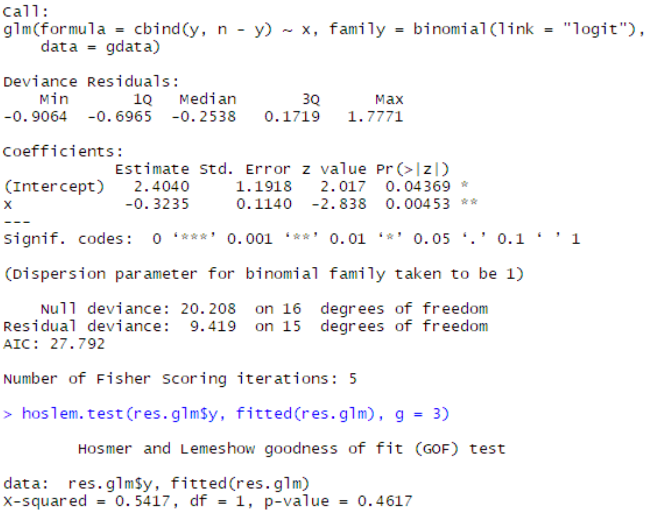


```
> ######Pearson residuals########
>
> x=resid(res.glm,type="pearson")
> x
           1            2            3            4            5            6            7            8
-4.814975971 -0.005358449  1.153603108  3.360565803  0.825798562  5.615179589 -1.658448448  3.403140114 -
          10           11           12           13           14           15           16           17
 2.348198044  7.908672497  3.159138722 -7.352727284 -6.609048135  5.211965771  5.389687464  6.845025089
          19           20
-4.836368399 -0.927323029
> ######standardized Pearson residuals####
> library(boot)
> glm.diag(res.glm)$rp
           1            2            3            4            5            6            7            8
-0.875557460 -0.000958912  0.215464450  0.793564985  0.158954662  0.986615402 -0.316102825  0.681106020 -1.882771
          11           12           13           14           15           16           17           18
 1.447513074  0.657593572 -1.445104163 -1.317016074  0.905662154  0.985166155  1.225226538 -1.687905615 -0.856766
> ######Deviance residuals##############
> rd=resid(res.glm)
> rd
           1            2            3            4            5            6            7            8
-4.814975971 -0.005358449  1.153603108  3.360565803  0.825798562  5.615179589 -1.658448448  3.403140114 -
          10           11           12           13           14           15           16           17
 2.348198044  7.908672497  3.159138722 -7.352727284 -6.609048135  5.211965771  5.389687464  6.845025089
          19           20
-4.836368399 -0.927323029
> ####standardized deviance residuals####
> srd=rstandard(res.glm)
> sd
function (x, na.rm = FALSE)
sqrt(var(if (is.vector(x) || is.factor(x)) x else as.double(x),
    na.rm = na.rm))
<bytecode: 0x000000001818a260>
<environment: namespace:stats>
> ##########Deviance#############
> sum(rd^2)
[1] 567.6629
>
> ####Pearson statistics##
> sum(x^2)
[1] 567.6629
```

- Asymptotic distribution
$$X^2 \sim \chi^2(N-p).$$
  There is some evidence to suggest that X^2 is often better than D because D is unduly influenced by very small frequencies.
- If each observation has a different covariate pattern, so y_i is zero or one, then neither D nor X^2 provides a useful measure of fit. This can happen if the explanatory variables are continuous.
- Hosmer and Lemeshow (HL) Statistic: Group observations into g categories on the basis of their predicted probabilities. Typically about 10 groups are used with approximately equal numbers of observations in each group.
$$X_{HL}^2 \sim \chi^2(g-2).$$



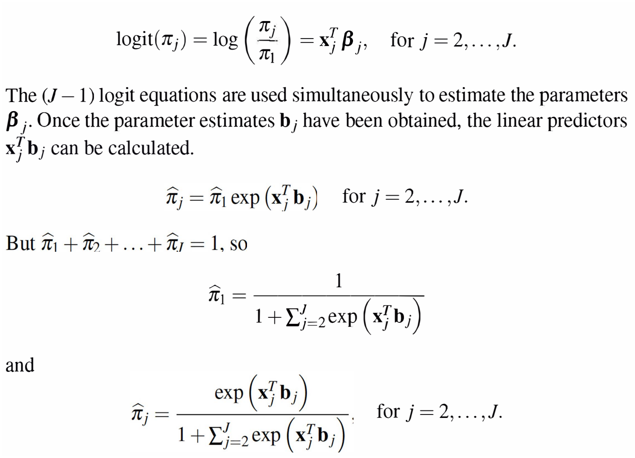
```
> gdata
    x y n
1   4 1 2
2   5 1 1
3   6 1 2
4   7 2 3
5   8 2 2
6   9 2 6
7  10 1 6
8  11 1 6
9  12 0 2
10 13 1 6
11 14 2 7
12 15 0 3
13 16 0 4
14 17 0 1
15 18 0 1
16 19 0 1
17 20 0 1
> 
```



```
Call:
glm(formula = cbind(y, n - y) ~ x, family = binomial(link = "logit"),
    data = gdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9064  -0.6965  -0.2538   0.1719   1.7771

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4040     1.1918   2.017  0.04369 *
x            -0.3235     0.1140  -2.838  0.00453 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 20.208  on 16  degrees of freedom
Residual deviance:  9.419  on 15  degrees of freedom
AIC: 27.792

Number of Fisher Scoring iterations: 5

> hoslem.test(res.glm$y, fitted(res.glm), g = 3)

        Hosmer and Lemeshow goodness of fit (GOF) test

data:  res.glm$y, fitted(res.glm)
X-squared = 0.5417, df = 1, p-value = 0.4617
```

Concept 5: Nominal logistic regression

Nominal logistic regression models are used when there is no natural order among the response categories. One category is arbitrarily chosen as the reference category.



$$\operatorname{logit}(\pi_j) = \log\left(\frac{\pi_j}{\pi_1}\right) = \mathbf{x}_j^T \boldsymbol{\beta}_j, \quad \text{for } j = 2, \ldots, J.$$

The $(J-1)$ logit equations are used simultaneously to estimate the parameters $\boldsymbol{\beta}_j$. Once the parameter estimates $\mathbf{b}_j$ have been obtained, the linear predictors $\mathbf{x}_j^T \mathbf{b}_j$ can be calculated.

$$\widehat{\pi}_j = \widehat{\pi}_1 \exp\left(\mathbf{x}_j^T \mathbf{b}_j\right) \quad \text{for } j = 2, \ldots, J.$$

But $\widehat{\pi}_1 + \widehat{\pi}_2 + \ldots + \widehat{\pi}_J = 1$, so

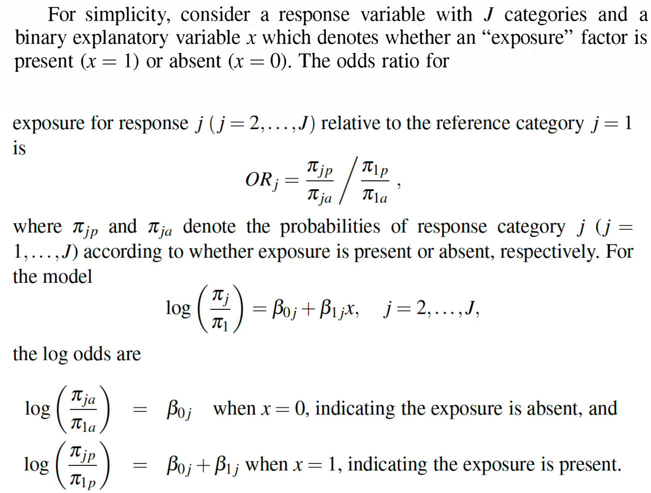
$$\widehat{\pi}_1 = \frac{1}{1 + \sum_{j=2}^{J} \exp\left(\mathbf{x}_j^T \mathbf{b}_j\right)}$$

and

$$\widehat{\pi}_j = \frac{\exp\left(\mathbf{x}_j^T \mathbf{b}_j\right)}{1 + \sum_{j=2}^{J} \exp\left(\mathbf{x}_j^T \mathbf{b}_j\right)}, \quad \text{for } j = 2, \ldots, J.$$

Interpretation

Often it is easier to interpret the effects of explanatory factors in terms of odds ratios than the parameters β.



For simplicity, consider a response variable with J categories and a binary explanatory variable x which denotes whether an "exposure" factor is present ($x = 1$) or absent ($x = 0$). The odds ratio for

exposure for response j ($j = 2, \ldots, J$) relative to the reference category $j = 1$ is

$$OR_j = \frac{\pi_{jp}}{\pi_{ja}} \bigg/ \frac{\pi_{1p}}{\pi_{1a}},$$

where π_{jp} and π_{ja} denote the probabilities of response category j ($j = 1, \ldots, J$) according to whether exposure is present or absent, respectively. For the model

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \beta_{0j} + \beta_{1j}x, \quad j = 2, \ldots, J,$$

the log odds are

$$\begin{aligned} \log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) &= \beta_{0j} \quad \text{when } x = 0\text{, indicating the exposure is absent, and} \\ \log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) &= \beta_{0j} + \beta_{1j} \text{ when } x = 1\text{, indicating the exposure is present.} \end{aligned}$$

Interpretation

Therefore, the logarithm of the odds ratio can be written as

$$\begin{aligned}\log OR_j &= \log\left(\frac{\pi_{jp}}{\pi_{1p}}\right) - \log\left(\frac{\pi_{ja}}{\pi_{1a}}\right) \\ &= \beta_{1j}.\end{aligned}$$

Hence, $OR_j = \exp(\beta_{1j})$ which is estimated by $\exp(b_{1j})$. If $\beta_{1j} = 0$, then $OR_j = 1$ which corresponds to the exposure factor having no effect. Also, for example, 95% confidence limits for OR_j are given by $\exp[b_{1j} \pm 1.96 \times \text{s.e.}(b_{1j})]$, where $\text{s.e.}(b_{1j})$ denotes the standard error of b_{1j}. Confidence intervals which do not include unity correspond to β values significantly different from zero.

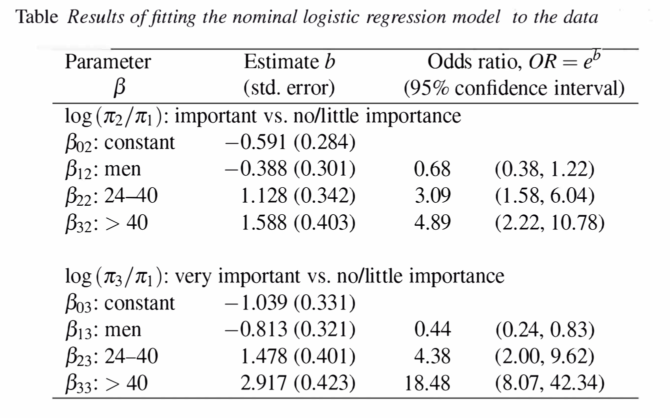


Table *Results of fitting the nominal logistic regression model to the data*

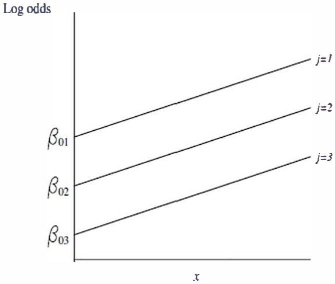
Parameter β	Estimate b (std. error)	Odds ratio, $OR = e^b$ (95% confidence interval)	
$\log(\pi_2/\pi_1)$: important vs. no/little importance			
β_{02}: constant	-0.591 (0.284)		
β_{12}: men	-0.388 (0.301)	0.68	(0.38, 1.22)
β_{22}: 24–40	1.128 (0.342)	3.09	(1.58, 6.04)
β_{32}: > 40	1.588 (0.403)	4.89	(2.22, 10.78)
$\log(\pi_3/\pi_1)$: very important vs. no/little importance			
β_{03}: constant	-1.039 (0.331)		
β_{13}: men	-0.813 (0.321)	0.44	(0.24, 0.83)
β_{23}: 24–40	1.478 (0.401)	4.38	(2.00, 9.62)
β_{33}: > 40	2.917 (0.423)	18.48	(8.07, 42.34)

Concept 6: Ordinal logistic regression

If there is an obvious natural order among the response categories, then this can be taken into account in the model specification. The **proportional odds model** is based on the assumption that the effects of the covariates $X_1, \ldots, X_{p-1}$ are the same for all categories on the logarithmic scale.

$$\log \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} = \beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}.$$

As for the nominal logistic regression model, the odds ratio associated with an increase of one unit in an explanatory variable x_k is $\exp(\beta_k)$, where $k = 1, \ldots, p-1$.



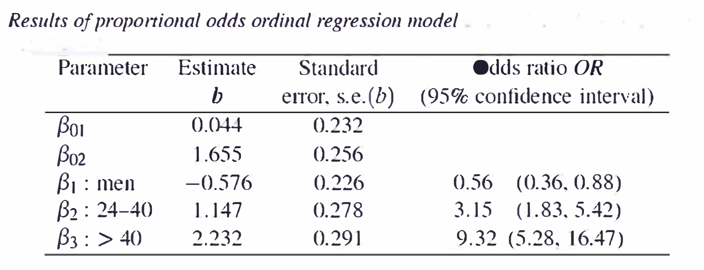
Interpretation

- In this model, intercept β_{0j} is the log-odds of falling into or below category j when $x_1 = x_2 = \ldots = x_{p-1} = 0$.
- A single parameter β_k describes the effect of x_k on Y such that β_k is the increase in log-odds of falling into or below any category associated with a one-unit increase in x_k, holding all the other X-variables constant.
- The proportional-odds condition forces the lines corresponding to each cumulative logit to be parallel. It is because the intercepts can differ, but that slope for each variable stays the same across different equations.
- Note that the description of the model given on is perhaps a bit counterintuitive, in that high values of $\beta_{0j} + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$ are associated with low values of Z.
- For this reason, many people prefer to specify the model as

$$\log\left(\frac{P(z \leq j)}{P(z > j)}\right) = \beta_{0j} - \beta_1 x_1 - \ldots - \beta_{p-1} x_{p-1}.$$

so that the sign of β's has the usual meaning (i.e., if positive,an increase in x is associated with an increase in z).

Example: car



Results of proportional odds ordinal regression model

Parameter	Estimate b	Standard error, s.e.(b)	Odds ratio OR (95% confidence interval)
β_{01}	0.044	0.232	
β_{02}	1.655	0.256	
β_1 : men	-0.576	0.226	0.56 (0.36, 0.88)
β_2 : 24–40	1.147	0.278	3.15 (1.83, 5.42)
β_3 : > 40	2.232	0.291	9.32 (5.28, 16.47)