Lecture 2: The principle of model building

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

How to fit a model?

- 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.

Example 1: Chronic medical conditions

Data from the Australian Longitudinal Study on Women's Health show that women who live in country areas tend to have fewer consultations with general practitioners (family physicians) than women who live near a wider range of health services.

Table 2.1 *Number of chronic medical conditions of 26 town women and 23 country women with similar use of general practitioner services.*

Town
0 1 1 0 2 3 0 1 1 1 1 2 0 1 3 0 1 2 1 3 3 4 1 3 2 0
$n = 26$, mean = 1.423, standard deviation = 1.172, variance = 1.374
Country
2 0 3 0 0 1 1 1 1 0 0 2 2 0 1 2 0 0 1 1 1 0 2
$n = 23$, mean = 0.913, standard deviation = 0.900, variance = 0.810

The question of interest is: Do women who have similar levels of use of GP services in the two groups have the same need as indicated by their number of chronic medical conditions?

Model formulation

- The Poisson distribution provides a plausible way of modeling these data (why?).
- Y_{jk} be a random variable representing the number of conditions for the kth woman in the jth group, where $j = 1$ for the town group and $j = 2$ for the country group.
- The null hypothesis

$$H_0 : \theta_1 = \theta_2 = \theta$$

VS

$$H_2 : \theta_1 \neq \theta_2.$$

- That is, under H_0 (Model 1),

$$E(Y_{jk}) = \theta; \quad Y_{jk} \sim \text{Po}(\theta).$$

Under H_1 (Model 2),

$$E(Y_{jk}) = \theta_j; \quad Y_{jk} \sim \text{Po}(\theta_j).$$

If Model 2 is clearly better, then H_0 would be rejected in favor of H_1.

How to develop the test?

Model fitting: MLE

If H_0 is true, then the log-likelihood function is

$$l_0 = l(\theta; \mathbf{y}) = \sum_{j=1}^{J} \sum_{k=1}^{K_j} (y_{jk} \log(\theta) - \theta - \log(y_{jk}!)).$$

The maximum likelihood estimate is

$$\hat{\theta} = \sum_{j=1}^{J} \sum_{k=1}^{K_j} y_{jk} / N,$$

where $N = \sum_j K_j$. For these data the estimate is

$$\hat{\theta} = 1.184$$

and the maximum value of the log-likelihood function, obtained by substituting this value and data

$$\hat{l}_0 = -68.3868.$$

Model fitting: MLE

If H_1 is true, then the log-likelihood function is

$$l_1 = l(\theta_1, \theta_2; \mathbf{y}) = \sum_{k=1}^{K_1}(y_{1k}\log(\theta_1) - \theta_1 - \log(y_{1k}!)) + \sum_{k=1}^{K_2}(y_{2k}\log(\theta_2) - \theta_2 - \log(y_{2k}!)).$$

The maximum likelihood estimate is

$$\hat{\theta}_1 = \sum_{k=1}^{K_1} y_{1k}/K_1, \hat{\theta}_2 = \sum_{k=1}^{K_2} y_{2k}/K_2.$$

For these data the estimate is

$$\hat{\theta}_1 = 1.423, \quad \hat{\theta}_2 = 0.913.$$

The maximum value of the log-likelihood function, obtained by substituting this value and data

$$\hat{l}_1 = -67.0230.$$

What is your conclusion?

Residual analysis

Recall: If $Y \sim Po(\theta)$, then

$$E(Y) = Var(Y) = \theta.$$

The fitted value is $\hat{\theta}$. The residual is defined as $Y - \hat{\theta}$. A residual is usually standardized by dividing by its standard error. For the Poisson distribution an approximate standardized residual is

$$r = \frac{Y - \hat{\theta}}{\sqrt{\hat{\theta}}}.$$

Value of Y	Frequency	Standardized residuals from (2.1); $\widehat{\theta} = 1.184$	Standardized residuals from (2.2); $\widehat{\theta}_1 = 1.423$ and $\widehat{\theta}_2 = 0.913$
		Town	
0	6	−1.088	−1.193
1	10	−0.169	−0.355
2	4	0.750	0.484
3	5	1.669	1.322
4	1	2.589	2.160
		Country	
0	9	−1.088	−0.956
1	8	−0.169	0.091
2	5	0.750	1.138
3	1	1.669	2.184

Goodness-of-fit statistics

The usual chi-squared goodness of fit statistic for count data which is often written as

$$X^2 = \sum \frac{(o_i - e_i)^2}{e_i} \sim \chi^2(m)$$

where o_i denotes the observed frequency and e_i denotes the corresponding expected frequency, and m is equal to the number of observations minus the number of parameters.

For Model 1,

$$\sum r_i^2 = \sum \frac{(Y_i - \hat{\theta}_i)^2}{\hat{\theta}_i} = 6 \times (-1.088)^2 + \ldots + 1 \times 1.669^2 = 46.759.$$

The degree of freedom $m = 23 + 26 - 1 = 48$.

For Model 2,

$$\sum r_i^2 = \sum \frac{(Y_i - \hat{\theta}_i)^2}{\hat{\theta}_i} = 6 \times (?1.193)^2 + \ldots + 1 \times 2.184^2 = 43.659.$$

The degree of freedom $m = 23 + 26 - 2 = 47$.

The difference between those two models is small: 46.759-43.659 = 3.10.

Example: Birthweight and gestational age

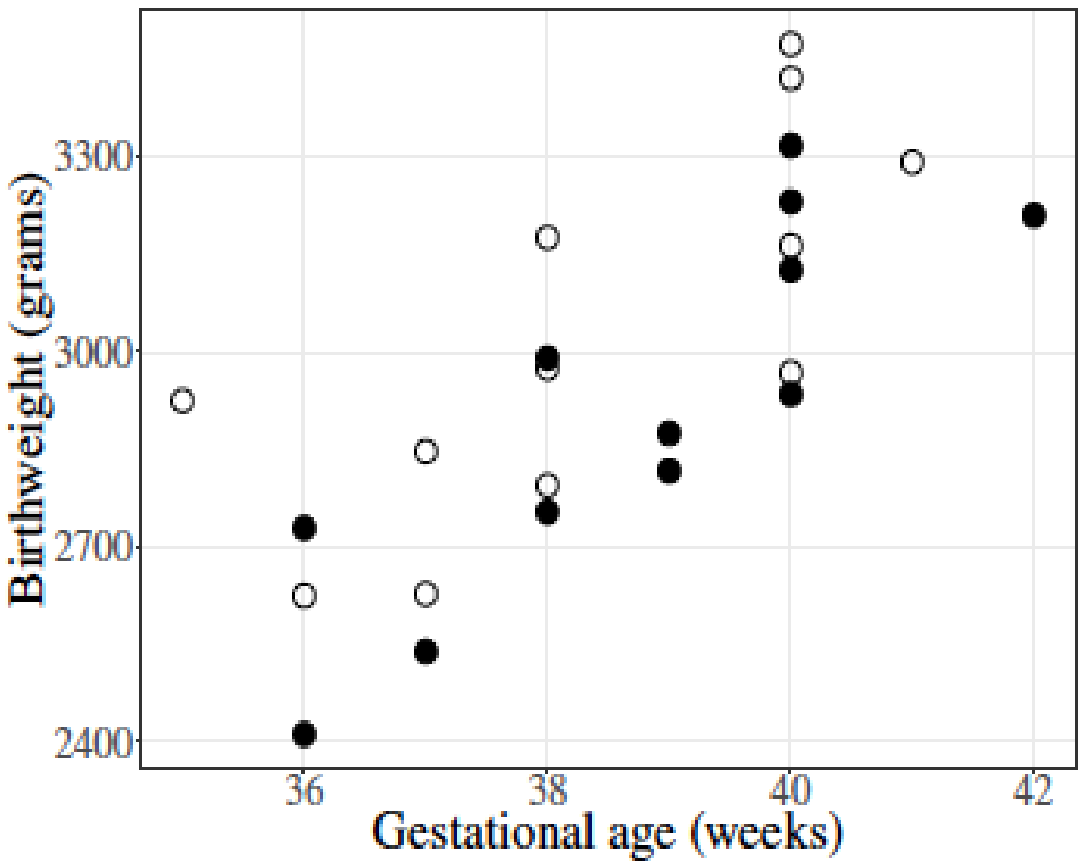
The data in the following Table are the birthweights (in grams) and estimated gestational ages (in weeks) of 12 male and female babies born in a certain hospital.

Birthweight (grams) and gestational age (weeks) for boys and girls.

	Boys		Girls	
	Age	Birthweight	Age	Birthweight
	40	2968	40	3317
	38	2795	36	2729
	40	3163	40	2935
	35	2925	38	2754
	36	2625	42	3210
	37	2847	39	2817
	41	3292	40	3126
	40	3473	37	2539
	37	2628	36	2412
	38	3176	38	2991
	40	3421	39	2875
	38	2975	40	3231
Mean	38.33	3024.00	38.75	2911.33

Example: Birthweight and gestational age

The question of interest is whether the rate of increase of birthweight with gestational age is the same for boys and girls.



Example: Birthweight and gestational age

Let Y_{jk} be a random variable representing the birthweight of the kth baby in group j where $j = 1$ for boys and $j = 2$ for girls and $k = 1, ..., 12$.

A fairly general model relating birthweight to gestational age is

$$E(Y_{jk}) = \mu_{jk} = \alpha_j + \beta_j x_{jk},$$

where x_{jk} is the gestational age of the kth baby in group j. The intercept parameters α_1 and α_2 are likely to differ because, on average, the boys were heavier than the girls.

We want to test

$$H_0 : \beta_1 = \beta_2 = \beta$$

VS

$$H_1 : \beta_1 \neq \beta_2.$$

That is, we can test H_0 against H_1 by fitting two models

$$E(Y_{jk}) = \alpha_j + \beta x_{jk}; \quad Y_{jk} \sim N(\mu_{jk}, \sigma^2)$$

VS

$$E(Y_{jk}) = \alpha_j + \beta_j x_{jk}; \quad Y_{jk} \sim N(\mu_{jk}, \sigma^2).$$

Model fitting

$$l_1(\alpha_1,\alpha_2,\beta_1,\beta_2;\mathbf{y}) = \sum_{j=1}^{J}\sum_{k=1}^{K}[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_{jk}-\mu_{jk})^2]$$

$$= -\frac{1}{2}JK\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{jk}-\alpha_j-\beta_j x_{jk})^2,$$

where $J=2$ and $K=12$ in this case. When obtaining maximum likelihood estimates of $\alpha_1, \alpha_2, \beta_1$ and β_2, the parameter σ^2 is treated as a known constant, or **nuisance parameter**, and is not estimated.

This is equivalent to the least squares estimation, i.e., minimizing the expression

$$S_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ij}-\mu_{jk})^2 = \sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ij}-\alpha_j+\beta_j x_{jk})^2.$$

The least squares estimates are the solutions of the equations

$$\frac{\partial S_1}{\partial \alpha_j} = -2\sum_{k=1}^{K}(y_{jk}-\alpha_j-\beta_j x_{jk}) = 0,$$

$$\frac{\partial S_1}{\partial \beta_j} = -2\sum_{k=1}^{K}x_{jk}(y_{jk}-\alpha_j-\beta_j x_{jk}) = 0. \qquad (2.10)$$

Model Estimation

The estimating Equations (2.10) can be simplified to

$$\sum_{k=1}^{K} y_{jk} - K\alpha_j - \beta_j \sum_{k=1}^{K} x_{jk} = 0,$$

$$\sum_{k=1}^{K} x_{jk} y_{jk} - K\alpha_j \sum_{k=1}^{K} x_{jk} - \beta_j \sum_{k=1}^{K} x_{jk}^2 = 0,$$

for $j = 1$ or 2. These are called the **normal equations**. The solution is

$$b_j = \frac{K\sum_k x_{jk} y_{jk} - (\sum_k x_{jk})(\sum_k y_{jk})}{K\sum_k x_{jk}^2 - (\sum_k x_{jk})^2},$$

$$a_j = \bar{y}_j - b_j \bar{x}_j,$$

where a_j is the estimate of α_j and b_j is the estimate of β_j, for $j = 1$ or 2.

Model fitting

To test $H_0 : \beta_1 = \beta_2 = \beta$, we have

$$S_0 = \sum_{j=1}^{J} \sum_{k=1}^{K} (y_{jk} - \alpha_j - \beta x_{jk})^2.$$

Similarly,

$$\frac{\partial S_0}{\partial \alpha_j} = -2 \sum_{k=1}^{K} (y_{jk} - \alpha_j - \beta x_{jk}) = 0,$$

$$\frac{\partial S_0}{\partial \beta} = -2 \sum_{j=1}^{J} \sum_{k=1}^{K} x_{jk} (y_{jk} - \alpha_j - \beta x_{jk}) = 0,$$

for $j = 1$ and 2. The solution is

$$b = \frac{K \sum_j \sum_k x_{jk} y_{jk} - \sum_j (\sum_k x_{jk} \sum_k y_{jk})}{K \sum_j \sum_k x_{jk}^2 - \sum_j (\sum_k x_{jk})^2},$$

$$a_j = \bar{y}_j - b \bar{x}_j.$$

Residual anlaysis

For model 1, the fitted values are

$$\hat{y}_{jk} = a_j + bx_{jk},$$

and for model 2, the fitted values are

$$\hat{y}_{jk} = a_j + b_j x_{jk}.$$

Model	Slopes	Intercepts	Minimum sum of squares
(1)	$b = 120.894$	$a_1 = -1610.283$	$\widehat{S}_0 = 658770.8$
		$a_2 = -1773.322$	
(2)	$b_1 = 111.983$	$a_1 = -1268.672$	$\widehat{S}_1 = 652424.5$
	$b_2 = 130.400$	$a_2 = -2141.667$	

The residual is defined as $y_{jk} - \hat{y}_{jk}$, and the standardized residual is

$$\frac{y_{jk} - \hat{y}_{jk}}{s},$$

where s is the sample standard deviation of the residuals.

Residual checking

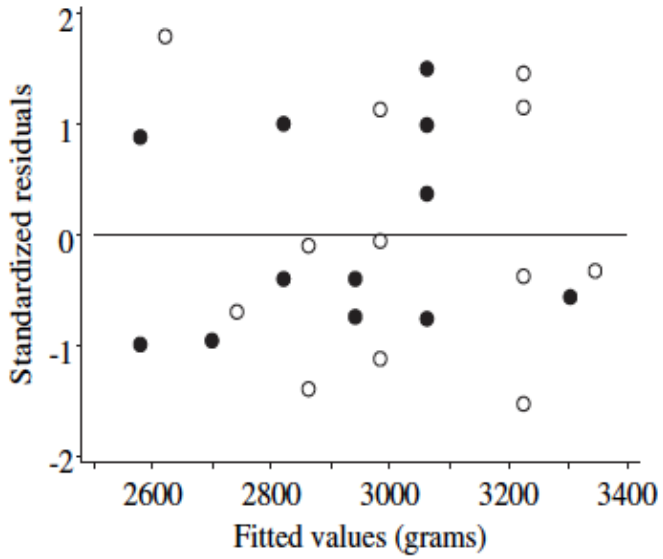
Checking:

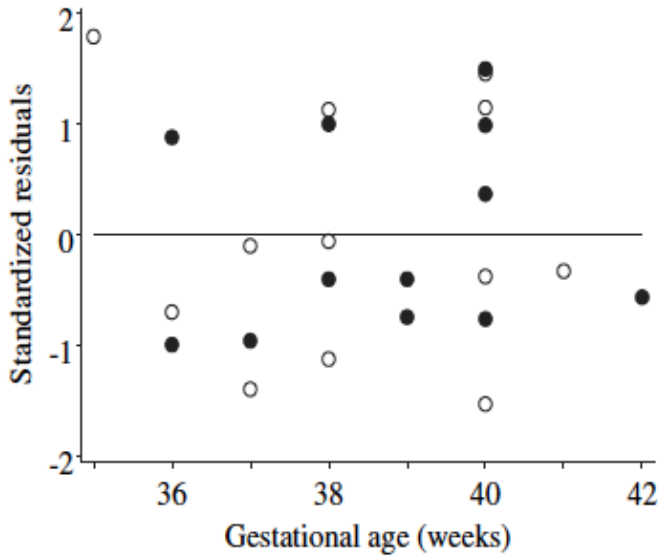
- Independence
- Approximately Normal with a mean of zero and constant variance
- Unrelated to the explanatory variables

How to do this:

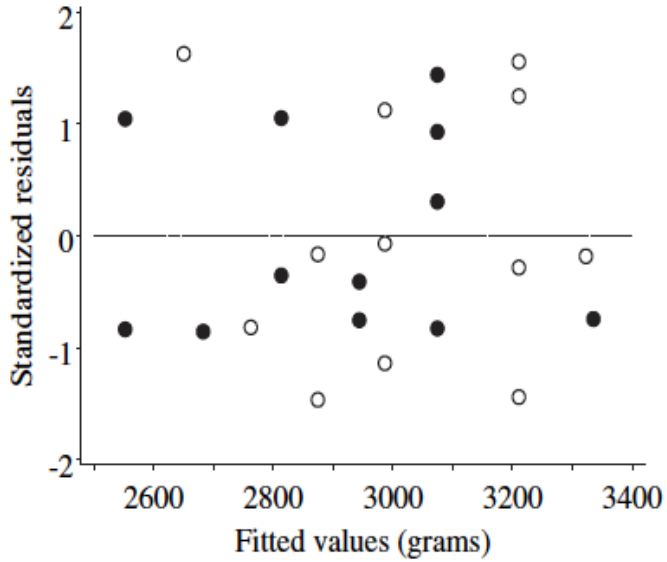
- standardized residuals vs explanatory variables
- standardized residuals should be plotted against the fitted values
- sequence plot of the residuals
- Normality: PP plot and QQ plot

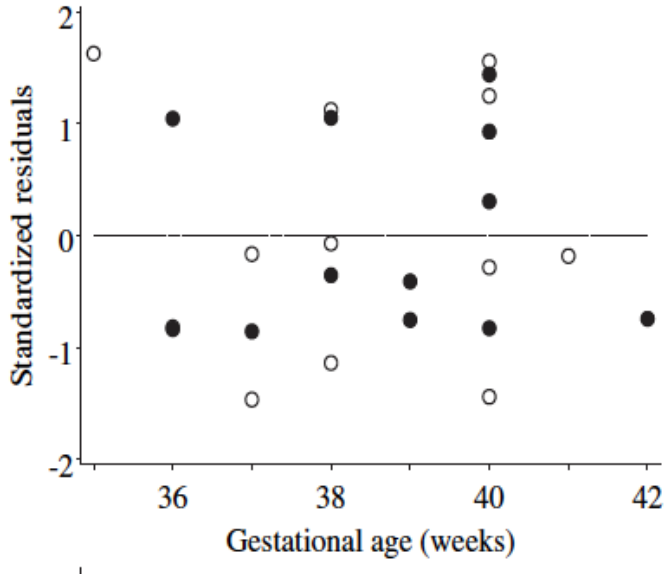
Residuals from Model 1





Residuals from Model 2





Which model to use?

$$\widehat{S}_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{jk} - a_j - b_j x_{jk})^2$$

and

$$\widehat{S}_0 = \sum_{j=1}^{J}\sum_{k=1}^{K}(Y_{jk} - a_j - b x_{jk})^2.$$

and further,

$$\widehat{S}_1 = \sum_{j=1}^{J}\sum_{k=1}^{K}[Y_{jk} - (\alpha_j + \beta_j x_{jk})]^2 - K\sum_{j=1}^{J}(\overline{Y}_j - \alpha_j - \beta_j \overline{x}_j)^2$$
$$- \sum_{j=1}^{J}(b_j - \beta_j)^2\Big(\sum_{k=1}^{K} x_{jk}^2 - K\overline{x}_j^2\Big)$$

Then,

$$\frac{\hat{S}_1}{\sigma^2} \sim \chi^2(JK - J - J), \quad \frac{\hat{S}_0}{\sigma^2} \sim \chi^2(JK - J - 1).$$

Which model to use?

If H_0 is correct, then the minimum values $\hat{S}_1$ and $\hat{S}_0$ should be nearly equal. Otherwise, $\hat{S}_0$ should be much larger than $\hat{S}_1$. If H_0 is correct, we have

$$\frac{1}{\sigma^2}(\hat{S}_0 - \hat{S}_1) \sim \chi^2(J-1).$$

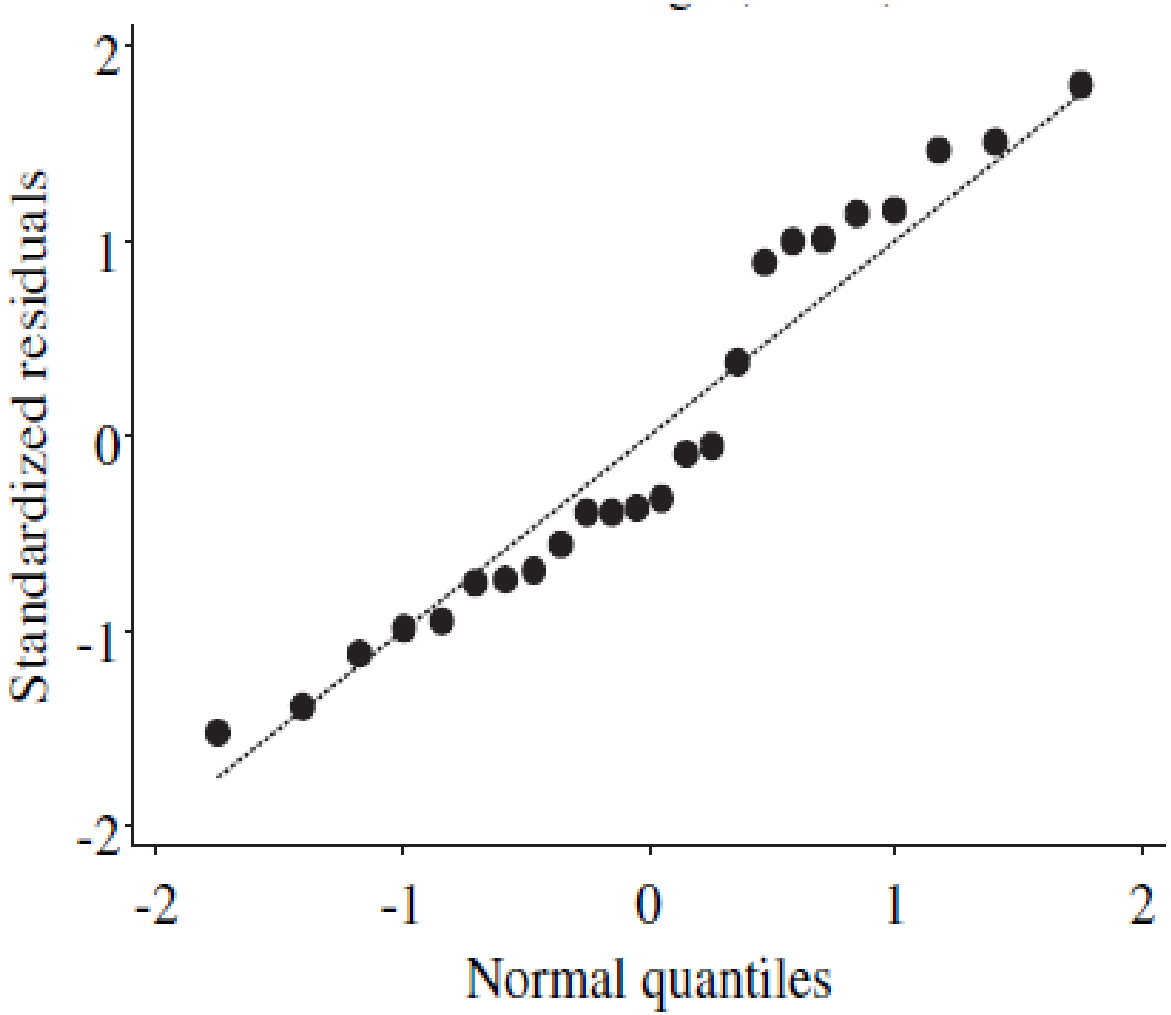
Since σ^2 is unknown, we use S_1^2. That is,

$$F = \frac{(\hat{S}_0 - \hat{S}_1)/\sigma^2}{J-1} \Big/ \frac{\hat{S}_1/\sigma^2}{JK-2J} = \frac{(\hat{S}_0 - \hat{S}_1)/(J-1)}{\hat{S}_1/(JK-2J)} \sim F(J-1, JK-2J).$$

If H_0 is not correct, F has a non-central F-distribution and the calculated value of F will be larger than expected from the central.
In our example, the value of F is

$$\frac{(658770.8 - 652424.5)/1}{652424.5/20} = 0.19.$$

Normal probability plot



Some notations

For the notations used in the generalized linear regression:

$$g[\mathrm{E}(Y)] = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m.$$

For responses $Y_1, \ldots, Y_N$, this can be written in matrix notation as

$$g[\mathrm{E}(\mathbf{y})] = \mathbf{X}\boldsymbol{\beta},$$

where

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_N \end{bmatrix} \quad \text{is a vector of responses,}$$

$$g[E(\mathbf{y})] = \begin{bmatrix} g[\mathrm{E}(Y_1)] \\ . \\ . \\ . \\ g[\mathrm{E}(Y_N)] \end{bmatrix}$$

denotes a vector of functions of the terms $\mathrm{E}(Y_i)$ (with the same g for every element),

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ . \\ . \\ . \\ \beta_p \end{bmatrix} \quad \text{is a vector of parameters,}$$

Example 1

For the data on chronic medical conditions, the equation in the model

$$\mathrm{E}(Y_{jk}) = \theta_j; \quad Y_{jk} \sim \mathrm{Po}(\theta_j),\ j = 1,2$$

can be written in the matrix form with g as the identity function (i.e., $g(\theta_j) = \theta_j$),

$$\mathbf{y} = \begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,26} \\ Y_{2,1} \\ \vdots \\ Y_{2,23} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix},$$

Example 2

The more general model for the data on birthweight and gestational age is

$$\mathrm{E}(Y_{jk}) = \mu_{jk} = \alpha_j + \beta_j x_{jk}; \quad Y_{jk} \sim \mathrm{N}(\mu_{jk}, \sigma^2).$$

This can be written in the matrix form if g is the identity function,

$$\mathbf{y} = \begin{bmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1K} \\ Y_{21} \\ \vdots \\ Y_{2K} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & 0 & x_{11} & 0 \\ 1 & 0 & x_{12} & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{1K} & 0 \\ 0 & 1 & 0 & x_{21} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{2K} \end{bmatrix}.$$