Biostatistics-Lecture 10
Regression

Ruibin Xi

Peking University

School of Mathematical Sciences

Analysis of Variance (ANOVA)

- Consider the Iris data again
- Want to see if the average sepal widths of the three species are the same
  - μ_1 , μ_2, μ_3 : the mean sepal width of Setosa, Versicolor, Virginica
  - Hypothesis:

    H0: $\mu_1 = \mu_2 = \mu_3$

    H1: at least one mean is different

Analysis of Variance (ANOVA)

- Used to compare ≥ 2 means
- Definitions
  - Response variable (dependent)—the outcome of interest, must be continuous
  - Factors (independent)—variables by which the groups are formed and whose effect on response is of interest, must be categorical
  - Factor levels—possible values the factors can take

Sources of Variation in One-Way ANOVA

- Partition the total variability of the outcome into components—source of variation
- $y_{i,j} \quad i = 1 \cdots k,\ j = 1 \cdots n_j$
  - the sepal width of the jth plant from the ith species (group)
  - $y_{ij} - \bar{y}_{..} = (y_{ij} - \bar{y}_{i\cdot}) + (\bar{y}_{i\cdot} - \bar{y}_{..})$

Grand mean

The ith group mean

Sources of Variation in One-Way ANOVA

- SST: sum of squares total

$$SST = SSB + SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_{\cdot\cdot} \right)^2$$

- SSB: sum of squares between

$$SSB = \sum_{i=1}^{k} n_i \left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot} \right)^2$$

- SSW (SSE): sum of squares within (error)

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_j} \left(y_{ij} - \bar{y}_{i\cdot} \right)^2$$

F-test in one-way ANOVA

- The test statistic is called F-statistic

$$F = \frac{MSB}{MSE} = \frac{SSB/(k-1)}{SSE/(n-k)}$$

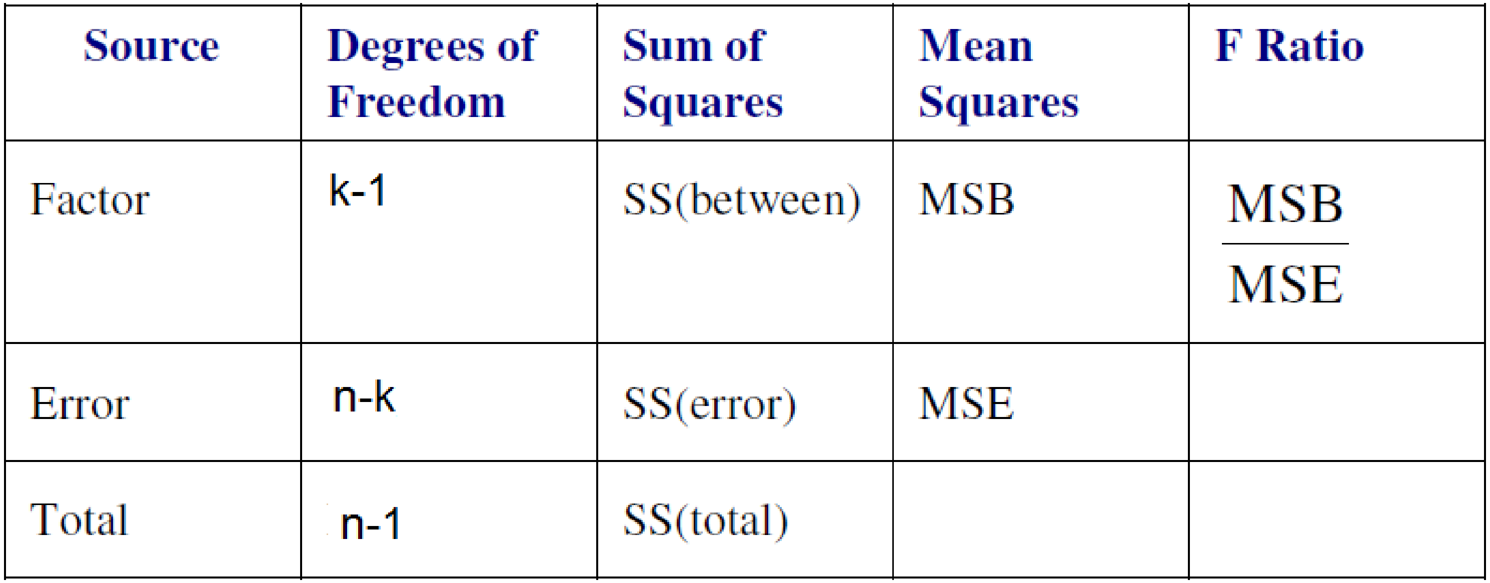
Under the null hypothesis, follows an F-distribution with $(df_1, df_2) = (k-1, n-k)$

- For the Iris data
  - SSB=11.34, MSB = 5.67, SSE=16.96, MSE=0.12
  - f = 49.16, $df_1=2, df_2=147$
  - Critical value 3.06 at α=0.05, reject the null
  - Pvalue = P(F>f)=4.49e-17

One-way ANOVA

- ANOVA table

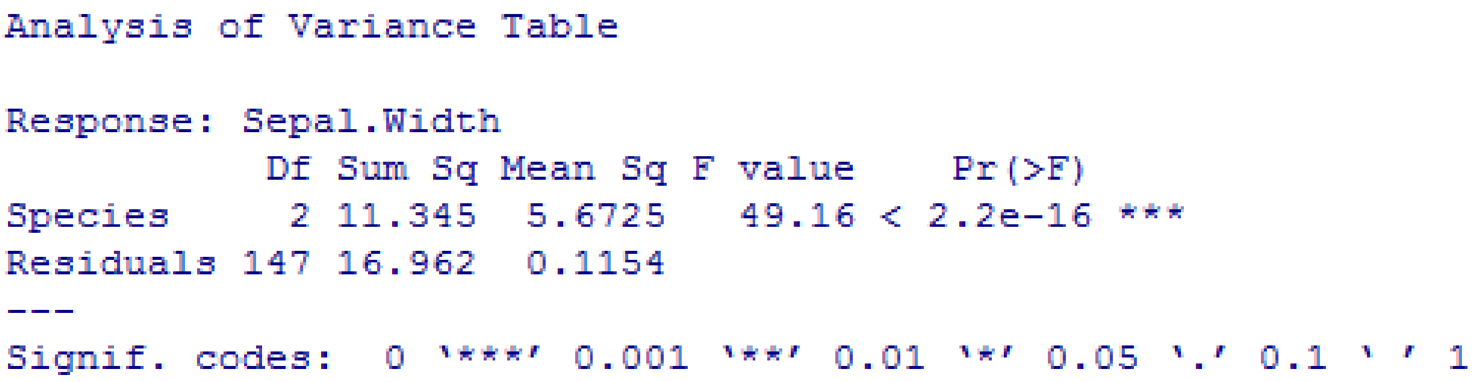
Table 15-2



Source	Degrees of Freedom	Sum of Squares	Mean Squares	F Ratio
Factor	k-1	SS(between)	MSB	$\frac{\text{MSB}}{\text{MSE}}$
Error	n-k	SS(error)	MSE	
Total	n-1	SS(total)		

One-way ANOVA

- ANOVA table

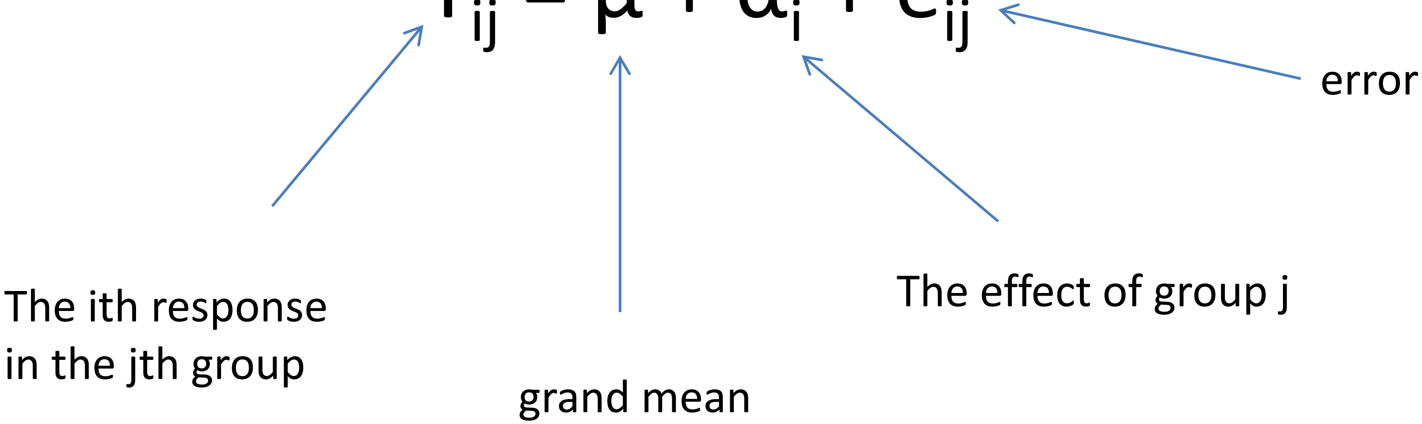


```
Analysis of Variance Table

Response: Sepal.Width
           Df Sum Sq Mean Sq F value    Pr(>F)    
Species     2 11.345  5.6725   49.16 < 2.2e-16 ***
Residuals 147 16.962  0.1154                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA model

- The statistical model

$$Y_{ij} = \mu + \alpha_i + e_{ij}$$



error

The effect of group j

The ith response in the jth group

grand mean

ANOVA assumptions

- Normality
- Homogeneity
- Independence

Regression—an example

- Cystic fibrosis (囊胞性纤维症) lung function data
  - PEmax (maximal static expiratory pressure) is the response variable
  - Potential explanatory variables
    - age, sex, height, weight,
    - BMP (body mass as a percentage of the age-specific median)
    - FEV1 (forced expiratory volume in 1 second)
    - RV (residual volume)
    - FRC(funcAonal residual capacity)
    - TLC (total lung capacity)

Regression—an example

- Let’s first concentrate on the age variable
- The model

$$y = \alpha + \beta x + e \ , \ \ e \sim N\left(0, \sigma^2\right)$$

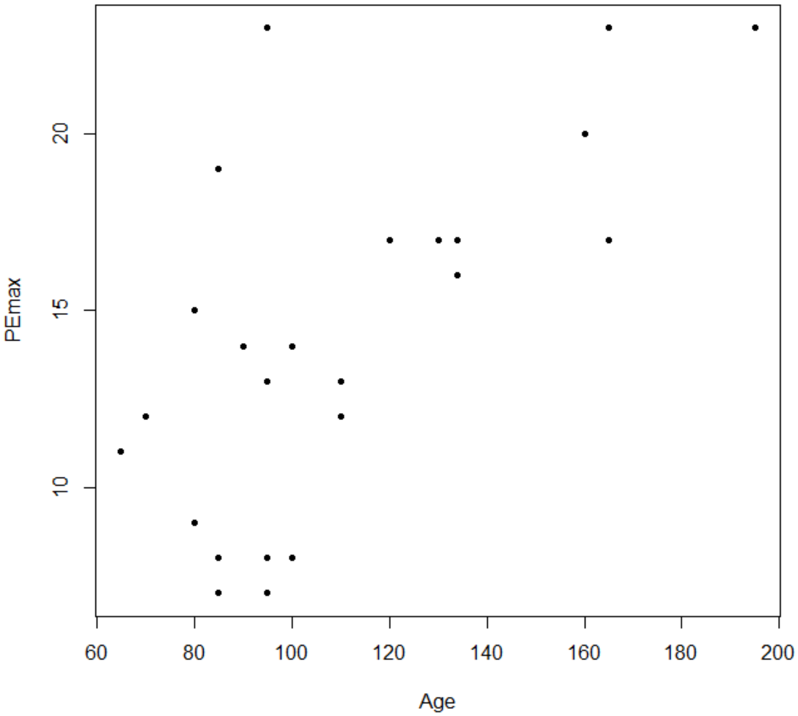
- Plot PEmax vs age

Regression—an example

- Let’s first concentrate on the age variable
- The model

y

- Plot PEma



Simple Linear regression

$$\boxed{y = \alpha + \beta x + e \ , \ \ e \sim N\left(0, \sigma^2\right)}$$

- *y*: dependent/response/outcome variable
- *x*: independent/explanatory/predictor variable
- *e*: error term
- α, β: coefficients/regression coefficients/model parameters
  - α: intercept
  - β: slope, describes the magnitude of association between X and Y
- For any give *x*, *y* = constant + normal random variable
- The values *x* are considered to be measured without error

Assumptions

- Normality
  - Given x, the distribution of y is normal with mean $\alpha+\beta x$ with standard deviation σ
- Homogeneity
  - σ does not depend on x
- Independence

Residuals

- Use the data from the sample to estimate α and β, the coefficients of the regression line

$$\boxed{y = \alpha + \beta x + e \ , \ \ e \sim N\left(0, \sigma^2\right)}$$

- Call the estimators *a* and *b*

$$\hat{y} = a + bx$$

- The discrepancies between the observed and fitted values are called residuals

$$\begin{aligned} d &= y - \hat{y} \\ &= y - a - bx \end{aligned}$$

Fitting the model

- One mathematical technique for fitting a straight line to a set of points is known as the method of least squares
- To apply this method, note that each data point (x_i, y_i) lies some vertical distance d_i from an arbitrary line (d_i is measured parallel to the vertical axis)
- Ideally, all residuals would be equal to 0
- Since this is impossible, we choose another criterion: we minimize the sum of squared

$$\boxed{S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2}$$

Fitting the model

- The resulting line is the **least squares line**
- Using calculus, it can be shown that

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

- Once *a* and *b* are known, we can substitute various values of *x* into the regression and compute *y*.

Goodness of Fit

- After estimating the model parameters, we need to evaluate how well the model fits the data
- Three criteria:
  - Inference about beta
  - R^2
  - Residual plots

- These concepts will hold for more complex cases, such as multiple regression, logistic regression, and Cox regression

Inference about β

- Because the parameter β describes the relationship between *X* and *Y*, inference about β tells us about the strength of the linear relationship.
- After estimating the model parameters, we can do hypothesis testing and build confidence intervals for β.
- The standard error of *b* in a simple linear regression is estimated as

$$\boxed{\widehat{s.e.}(b) = \sqrt{\frac{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}}$$

Inference about β

- To test the hypotheses H_0: β=0, we calculate the test statistic

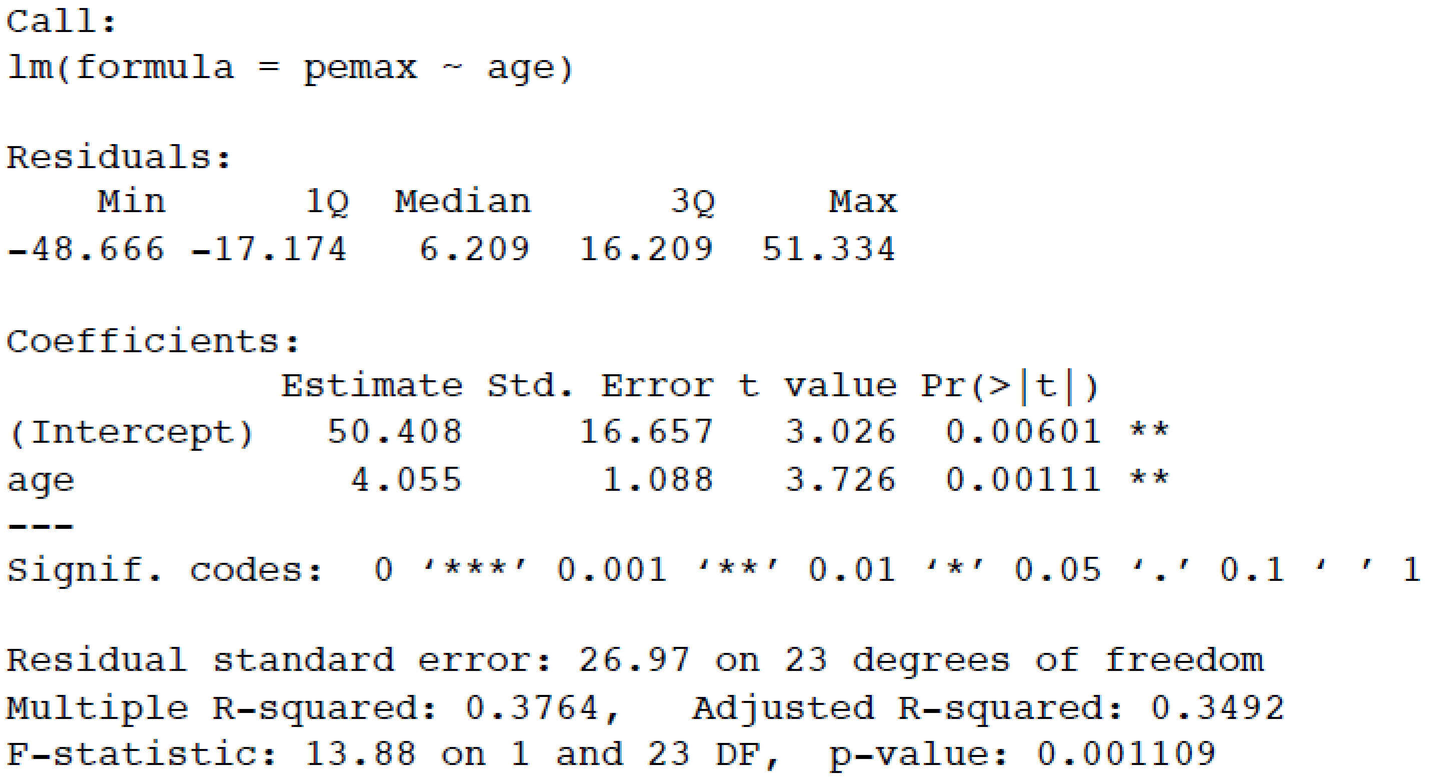
$$\boxed{t = \frac{b}{\widehat{s.e.}(b)}}$$

- Under H_0, this has a t distribution with n-2 df
- If the true population slope is equal to 0, there is no linear relationship between x and y; x is of no value in predicting y
- 100(1-α) CI for β:

$$\boxed{b \pm t_{n-2,\,1-\frac{\alpha}{2}}\,\widehat{s.e.}(b)}$$

- We can also carry out a similar procedure for α

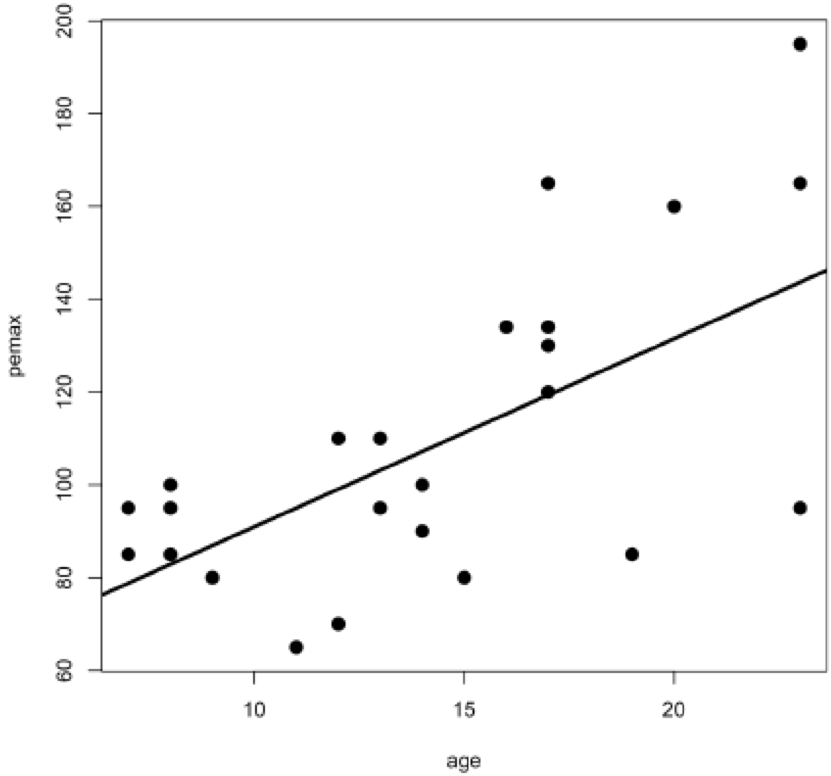
Inference about β: the CF data



```
Call:
lm(formula = pemax ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-48.666 -17.174   6.209  16.209  51.334 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)   50.408     16.657   3.026  0.00601 **
age            4.055      1.088   3.726  0.00111 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 26.97 on 23 degrees of freedom
Multiple R-squared: 0.3764,	Adjusted R-squared: 0.3492 
F-statistic: 13.88 on 1 and 23 DF,  p-value: 0.001109
```

Plotting the regression line

```
plot(age,pemax,cex=2,pch=20)
names(my.model)
abline(my.model$coeff[1],my.model$coeff[2],lw=3)
```



R^2

- Another measure is R^2, sometimes called the coefficient of determination:

$$\boxed{R^2 = \frac{\text{Reg SS}}{\text{Total SS}} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$

- **This is the proportion of variation explained by the model**
- It is also the square of Pearson’s correlation coefficient

```
> cor(pemax,age)^2
[1] 0.3763505
```

Residual plot

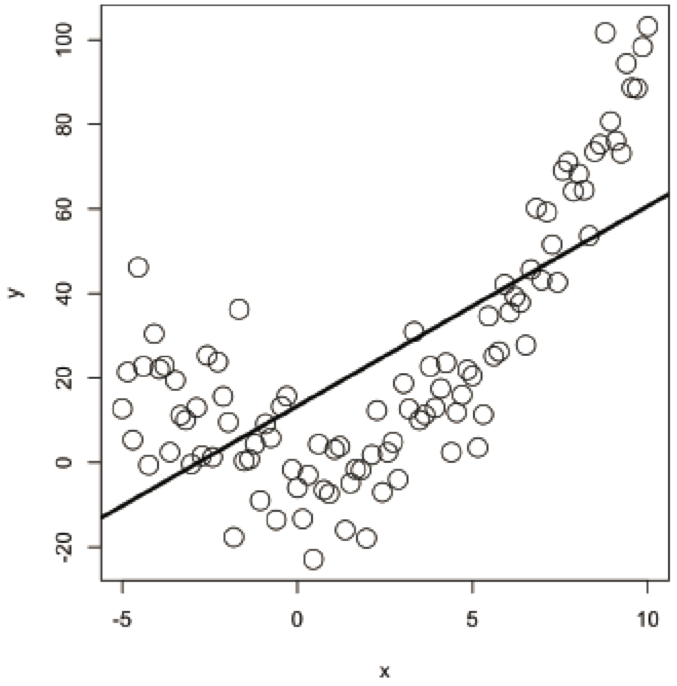
- We've been assuming that the association between *X* and *Y* in the population is truly linear.
- Even if the association is nonlinear, these methods may still fit a line without detecting a problem. In this case, inferences from the model will not be correct.
- Previously we defined a point's **residual:**

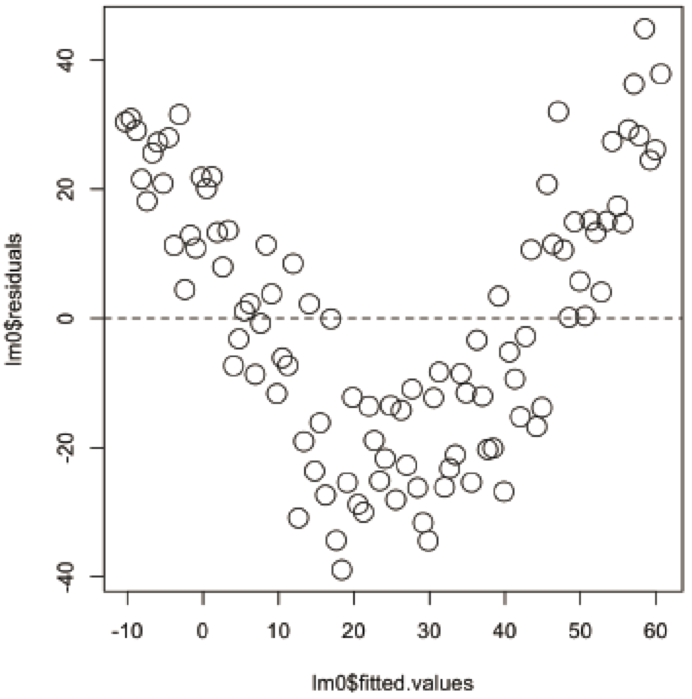
$$d_i = y_i - \hat{y}_i = y_i - a - bx_i$$

- Because of the assumptions of linear regression, we expect all the residuals to be normally distributed with the same mean (0) and the same variance.
- Violations of the linear regression assumptions can often be detected on a residual plot.

Residual plot

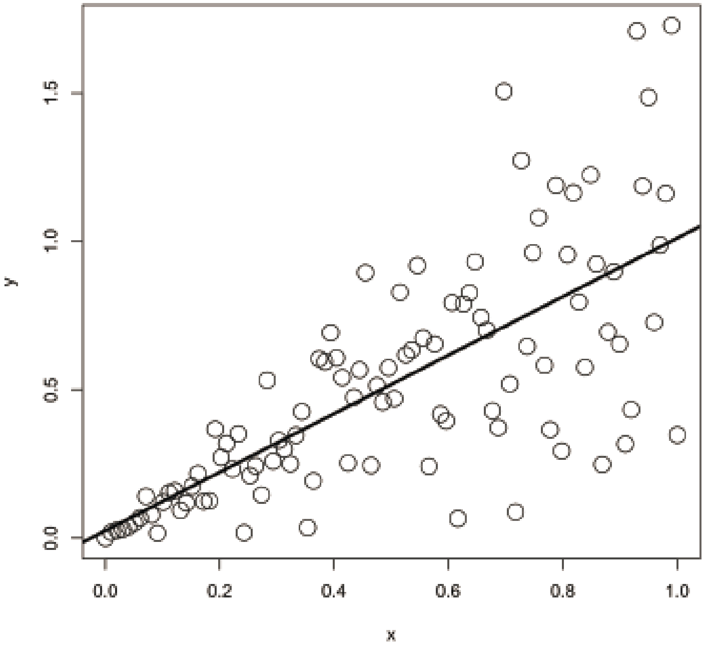
- Plot the predicted *y*-values on the *x*-axis and the residuals on the *y*-axis
- Are the residuals normally distributed with constant variance?

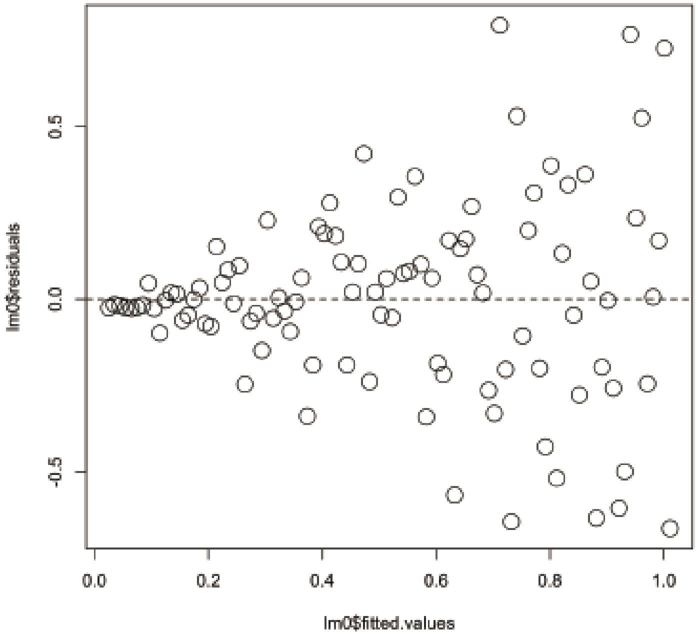




Residual plot

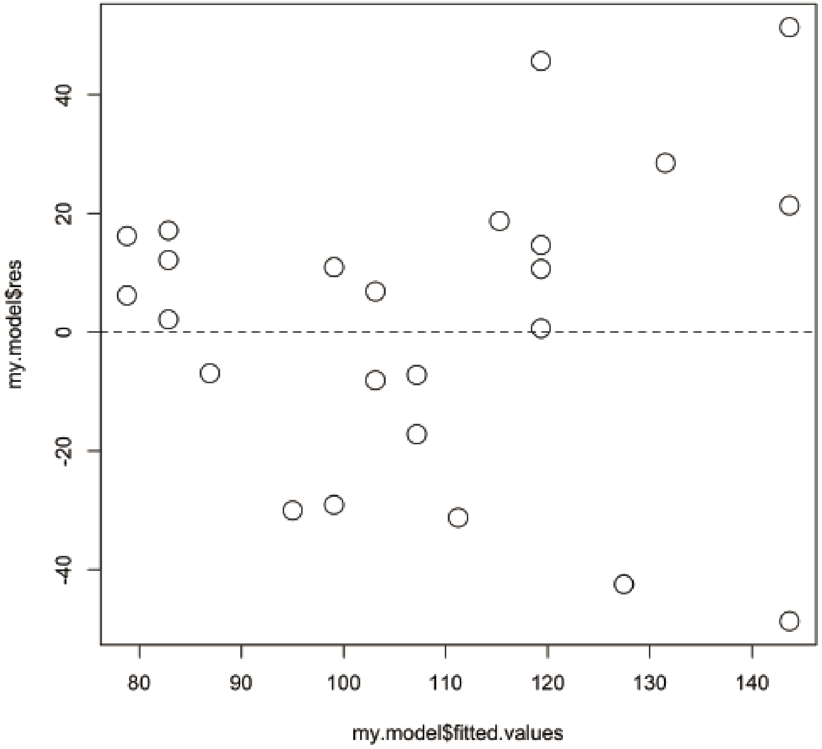
- Another example:





Residual plot

- The CF patients data



Does this model violate the assumption for constant variance?

Linear Regression

- Which models are ‘linear’?
  - $y = a + bx$
  - $y = bx$
  - $y = a + b_1x_1 + b_2x_2$
  - $y = a + b \log(x)$
  - $y = a + b\, x_1^2$
  - $\log(y) = a + bx$

- In fact, linear regression is not so restrictive

Summary: simple linear regression

- Linear model

$$\boxed{y = \alpha + \beta x + e\ ,\ \ e \sim N\left(0, \sigma^2\right)}$$

- Method of Least Squares

$$\boxed{S = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} \left(y_i - a - bx_i\right)^2}$$

- Testing for significance of coefficients

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \boxed{\hat{s.e.}(b) = \sqrt{\frac{\left(\frac{1}{n-2}\right)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}} \quad \boxed{t = \frac{b}{\hat{s.e.}(b)}}$$

Multiple regression

- See blackboard

Regression

GENERALIZED LINEAR MODELS

Generalized liner models

- GLM allow for response distributions other than normal
  - Basic structure $g(\mu_i) = \mathbf{X}_i\beta$

$$\mu_i \equiv \mathbb{E}(Y_i)$$

g is a smooth monotonic ‘link function’

- The distribution of Y is usually assumed to be independent and

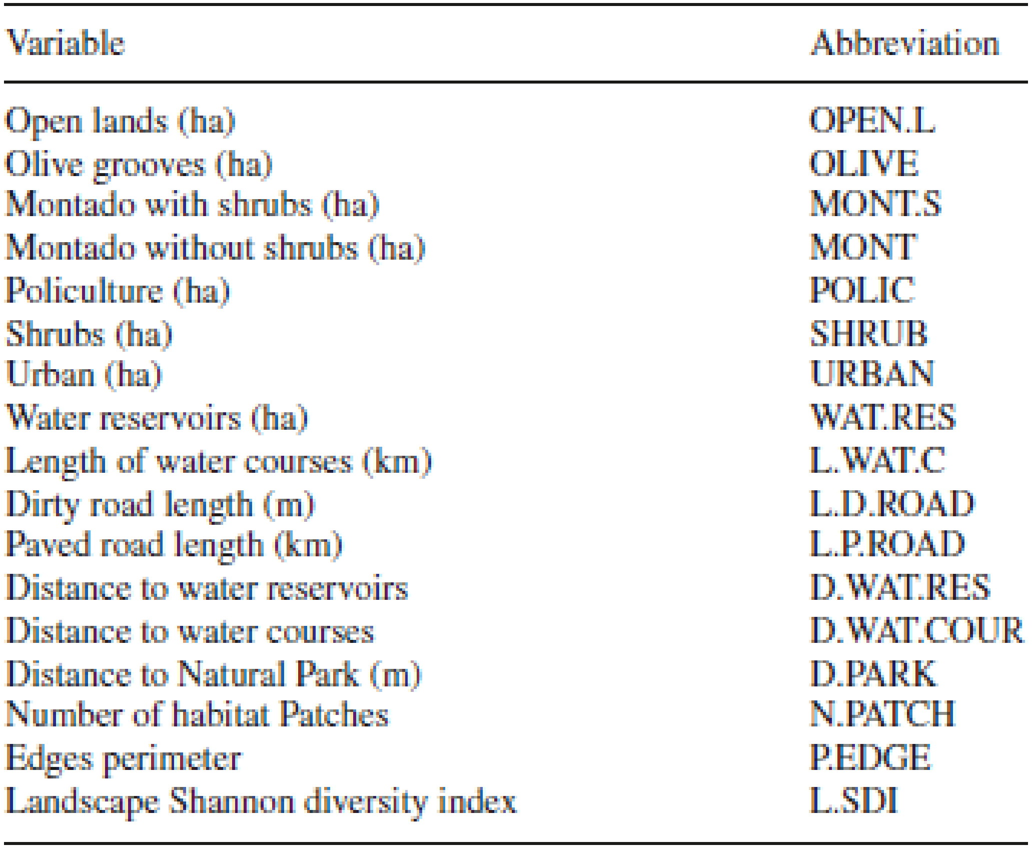
$Y_i \sim$ some exponential family distribution.

Generalized Linear model-an example

- An example
  - A study investigated the roadkills of amphibian
    - Response variable: the total number of amphibian fatalities per segment (500m)
    - Explanatory variables

Generalized Linear model-an example

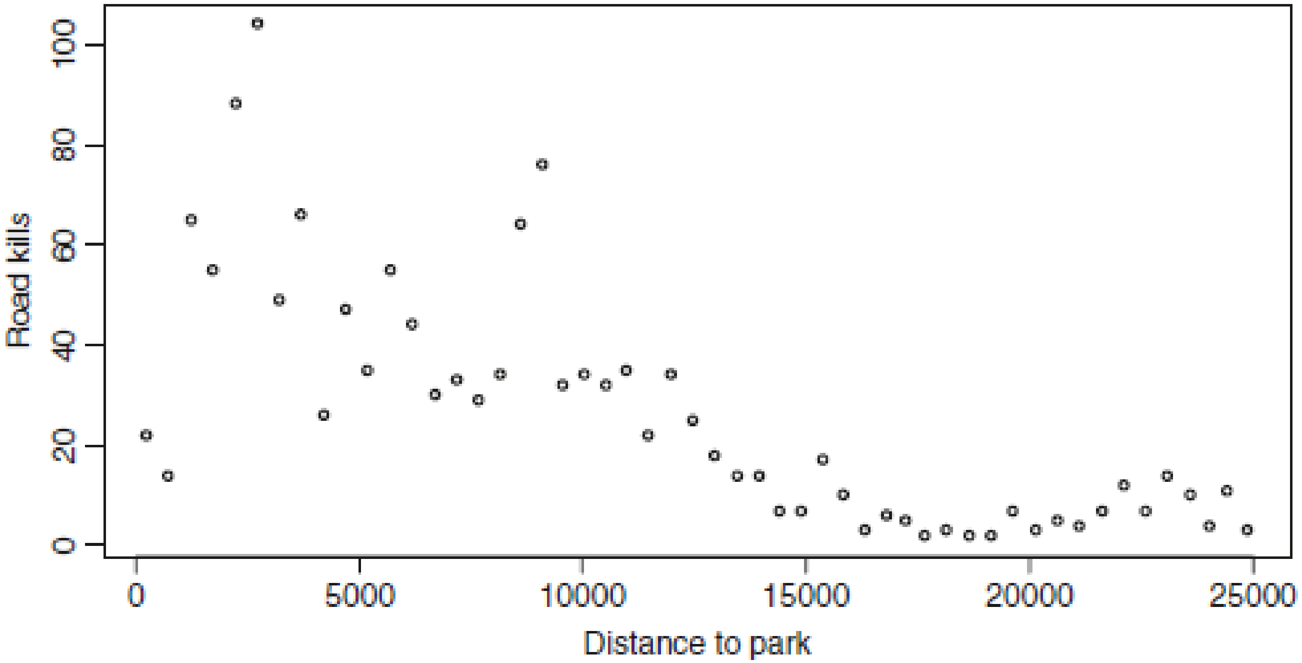
- An example
  - A stuc ibian
    - Res ibian fata
    - Exp



Variable	Abbreviation
Open lands (ha)	OPEN.L
Olive grooves (ha)	OLIVE
Montado with shrubs (ha)	MONT.S
Montado without shrubs (ha)	MONT
Policulture (ha)	POLIC
Shrubs (ha)	SHRUB
Urban (ha)	URBAN
Water reservoirs (ha)	WAT.RES
Length of water courses (km)	L.WAT.C
Dirty road length (m)	L.D.ROAD
Paved road length (km)	L.P.ROAD
Distance to water reservoirs	D.WAT.RES
Distance to water courses	D.WAT.COUR
Distance to Natural Park (m)	D.PARK
Number of habitat Patches	N.PATCH
Edges perimeter	P.EDGE
Landscape Shannon diversity index	L.SDI

Generalized Linear model-an example

- An example
  - – n



ural

Generalized Linear model-an example

- An example
  - A study investigated the roadkills of amphibian
    - Response variable: the total number of amphibian fatalities per segment (500m)
    - Explanatory variables
    - For now, we are only interested in *Distance to Natural Park*

Generalized Linear model-an example

- An over-simplified model

$$Y_i \sim p(\mu_i)$$

$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i$$

$$\log(\mu_i) = \alpha + \beta \times D.PARK_i \quad \text{or} \quad \mu_i = e^{\alpha + \beta \times D.PARK_i}$$

- For Poisson we have

$$f(y_i; \mu_i) = \frac{\mu_i^{y_i} \times e^{-\mu_i}}{y_i!} \qquad y_i \geq 0,\ y_i \text{ integer}$$

GLM-exponential families

- Exponential families
  - The density

$$f_\theta(y) = \exp\left[\{y\theta - b(\theta)\}/a(\phi) + c(y,\phi)\right]$$

b, a and c are arbitrary functions.

ϕ an arbitrary 'scale' parameter.

θ the 'canonical parameter' of the distribution

  - Normal distributions is an exponential family

$$\begin{aligned}
f_\mu(y) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] \\
&= \exp\left[\frac{-y^2 + 2y\mu - \mu^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right] \\
&= \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log(\sigma\sqrt{2\pi})\right],
\end{aligned}$$

GLM-exponential families

- Consider the log-likelihood of a general exponential families

$$l(\theta) = [y\theta - b(\theta)]/a(\phi) + c(y, \phi)$$

$$\frac{\partial l}{\partial \theta} = [y - b'(\theta)]/a(\phi)$$

$$\mathbb{E}\left(\frac{\partial l}{\partial \theta}\right) = [\mathbb{E}(Y) - b'(\theta)]/a(\phi).$$

Since $\mathbb{E}(\partial l/\partial \theta) = 0$

$$\mathbb{E}(Y) = b'(\theta).$$

GLM-exponential families

- Differentiating the likelihood one more time

$$\frac{\partial^2 l}{\partial \theta^2} = -b''(\theta)/a(\phi),$$

using the equation $\mathbb{E}(\partial^2 l/\partial\theta^2) = -\mathbb{E}[(\partial l/\partial\theta)^2]$

$$b''(\theta)/a(\phi) = \mathbb{E}\left[(Y - b'(\theta))^2\right]/a(\phi)^2.$$

$$\operatorname{var}(Y) = b''(\theta)a(\phi).$$

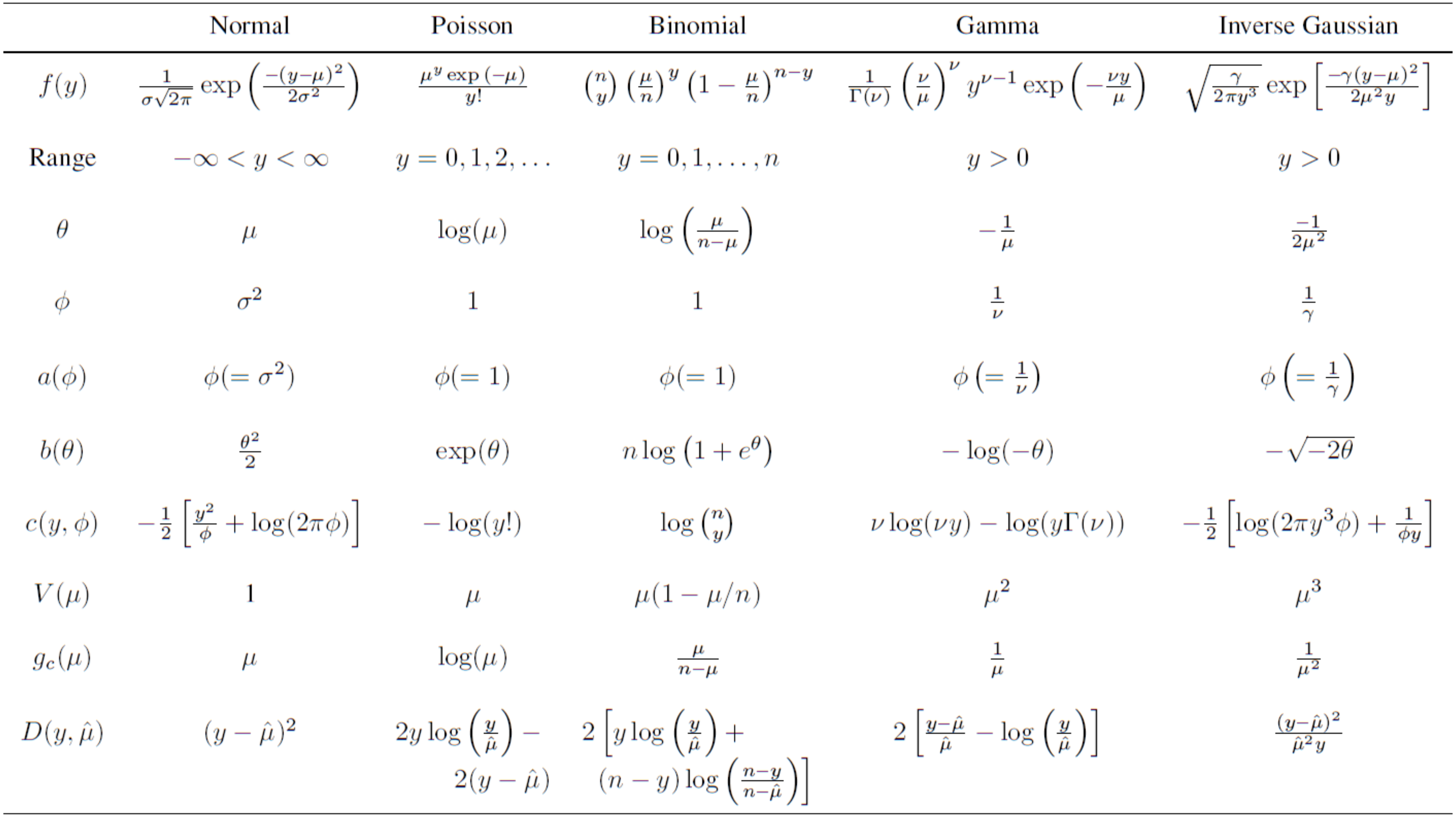
We often assume $a(\phi) = \phi/\omega.$

$$\operatorname{var}(Y) = b''(\theta)\phi/\omega.$$

Define $V(\mu) = b''(\theta)/\omega.$

$$\operatorname{var}(Y) = V(\mu)\phi.$$

GLM-exponential families



	Normal	Poisson	Binomial	Gamma	Inverse Gaussian
$f(y)$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$	$\frac{\mu^y\exp(-\mu)}{y!}$	$\binom{n}{y}\left(\frac{\mu}{n}\right)^y\left(1-\frac{\mu}{n}\right)^{n-y}$	$\frac{1}{\Gamma(\nu)}\left(\frac{\nu}{\mu}\right)^{\nu}y^{\nu-1}\exp\left(-\frac{\nu y}{\mu}\right)$	$\sqrt{\frac{\gamma}{2\pi y^3}}\exp\left[\frac{-\gamma(y-\mu)^2}{2\mu^2 y}\right]$
Range	$-\infty < y < \infty$	$y = 0, 1, 2, \ldots$	$y = 0, 1, \ldots, n$	$y > 0$	$y > 0$
θ	μ	$\log(\mu)$	$\log\left(\frac{\mu}{n-\mu}\right)$	$-\frac{1}{\mu}$	$\frac{-1}{2\mu^2}$
ϕ	σ^2	1	1	$\frac{1}{\nu}$	$\frac{1}{\gamma}$
$a(\phi)$	$\phi(=\sigma^2)$	$\phi(=1)$	$\phi(=1)$	$\phi\left(=\frac{1}{\nu}\right)$	$\phi\left(=\frac{1}{\gamma}\right)$
$b(\theta)$	$\frac{\theta^2}{2}$	$\exp(\theta)$	$n\log\left(1+e^{\theta}\right)$	$-\log(-\theta)$	$-\sqrt{-2\theta}$
$c(y,\phi)$	$-\frac{1}{2}\left[\frac{y^2}{\phi}+\log(2\pi\phi)\right]$	$-\log(y!)$	$\log\binom{n}{y}$	$\nu\log(\nu y)-\log(y\Gamma(\nu))$	$-\frac{1}{2}\left[\log(2\pi y^3\phi)+\frac{1}{\phi y}\right]$
$V(\mu)$	1	μ	$\mu(1-\mu/n)$	μ^2	μ^3
$g_c(\mu)$	μ	$\log(\mu)$	$\frac{\mu}{n-\mu}$	$\frac{1}{\mu}$	$\frac{1}{\mu^2}$
$D(y,\hat{\mu})$	$(y-\hat{\mu})^2$	$2y\log\left(\frac{y}{\hat{\mu}}\right)-$ $2(y-\hat{\mu})$	$2\left[y\log\left(\frac{y}{\hat{\mu}}\right)+\right.$ $\left.(n-y)\log\left(\frac{n-y}{n-\hat{\mu}}\right)\right]$	$2\left[\frac{y-\hat{\mu}}{\hat{\mu}}-\log\left(\frac{y}{\hat{\mu}}\right)\right]$	$\frac{(y-\hat{\mu})^2}{\hat{\mu}^2 y}$

Fitting the GLM

- In a GLM $g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$

$$Y_i \sim f_{\theta_i}(y_i)$$

- The joint likelihood is

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_{\theta_i}(y_i),$$

- The log likelihood

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \log[f_{\theta_i}(y_i)] \\ &= \sum_{i=1}^{n} [y_i\theta_i - b_i(\theta_i)]/a_i(\phi) + c_i(\phi, y_i), \end{aligned}$$

Fitting the GLM

- Assuming $a_i(\phi) = \phi/\omega_i$ (ω_i is known)

$$l(\beta) = \sum_{i=1}^{n} \omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)$$

- Differentiating the log likelihood and setting it to zero

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b_i'(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} \right)$$

Fitting the GLM

- By the chain rule

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

- Since $\mu_i = b'(\theta_i)$

$$\frac{\partial \mu_i}{\partial \theta_i} = b_i''(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b_i''(\theta_i)},$$

- We have

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{[y_i - b_i'(\theta_i)]}{b_i''(\theta_i)/\omega_i} \frac{\partial \mu_i}{\partial \beta_j}.$$

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0$$

Canonical Link Function

- The Canonical Link Function g_c is such that

$$g_c(\mu_i) = \theta_i$$

- Remember that $g(\mu_i) = \mathbf{X}_i\boldsymbol{\beta}$

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi}\sum_{i=1}^{n}\omega_i\left(y_i\frac{\partial\theta_i}{\partial\beta_j} - b_i'(\theta_i)\frac{\partial\theta_i}{\partial\beta_j}\right)$$

$$\mu_i = b'(\theta_i)$$

- So

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\omega_i\left(y_i\frac{\partial\theta_i}{\partial\beta_j} - \mu_i\frac{\partial\theta_i}{\partial\beta_j}\right) = 0$$

$$\partial\theta_i/\partial\beta_j = X_{ij}$$

Iteratively Reweighted Least Squares (IRLS)

- We first consider fitting the nonlinear model

$$\mathbb{E}(\mathbf{y}) \equiv \boldsymbol{\mu} = \mathbf{f}(\beta)$$

  by minimizing

$$\mathcal{S} = \sum_{i=1}^{n} \{y_i - f_i(\beta)\}^2 = \|\mathbf{y} - \mathbf{f}(\beta)\|^2$$

  where f is a nonlinear function

Iteratively Reweighted Least Squares (IRLS)

- Given a good guess $\hat{\beta}^{[k]}$

  by using the Taylor expansion

  $$\mathcal{S} \approx \mathcal{S}^{[k]} = \|\mathbf{y} - \mathbf{f}(\hat{\beta}^{[k]}) + \mathbf{J}^{[k]}\hat{\beta}^{[k]} - \mathbf{J}^{[k]}\beta\|^2$$

  $$J_{ij}^{[k]} = \partial f_i / \partial \beta_j$$

  Define the pseudodata

  $$\mathbf{z}^{[k]} = \mathbf{y} - \mathbf{f}(\hat{\beta}^{[k]}) + \mathbf{J}^{[k]}\hat{\beta}^{[k]}$$

  $$\mathcal{S}^{[k]} = \|\mathbf{z}^{[k]} - \mathbf{J}^{[k]}\beta\|^2$$

Iteratively Reweighted Least Squares (IRLS)

- Note that in GLM, we are trying to solve

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0$$

- If $V(\mu_i)$ are known, this is equivalent to minimizing

$$\mathcal{S} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{V(\mu_i)}$$

Iteratively Reweighted Least Squares (IRLS)

- We are inspired to use the following algorithm
  - At the kth iteration, define

    $$\eta_i^{[k]} = \mathbf{X}_i \hat{\beta}^{[k]} \qquad \mu_i^{[k]} = g^{-1}(\eta_i^{[k]})$$

    - Calculate the $V(\mu_i^{[k]})$ terms implied by the current $\hat{\beta}^{[k]}$
    - update $\hat{\beta}^{[k+1]}$ as in the nonlinear model case
    - set k to be k+1
  - But the second step also involves iteration, we may perform one step iteration here to obtain $\hat{\beta}^{[k+1]}$

Deviance

- The deviance is defined as

$$\begin{aligned} D &= 2[l(\hat{\beta}_{\max}) - l(\hat{\beta})]\phi \\ &= \sum_{i=1}^{n} 2\omega_i \left[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i) \right] \end{aligned}$$

where $l(\hat{\beta}_{\max})$ is the log likelihood with the saturated model: the model with one parameter per data point

Also note that deviance is defined to be independent of ϕ

Deviance

- GLM does not have R^2
- The closest one is the explained deviance

$$100 \times \frac{\text{null deviance} - \text{residual deviance}}{\text{null deviance}}$$

- The over-dispersion parameter may be estimated by

$$\hat{\phi} = \frac{D}{n-p}$$

or by the Pearson statistic

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

Model comparison

- Scaled deviance

$$D^* = D/\phi$$

- For the hypothesis testing problem

$$\mathrm{H}_0 : \mathrm{g}(\boldsymbol{\mu}) = \mathbf{X}_0\boldsymbol{\beta}_0$$

under H_0

$$D_0^* - D_1^* \sim \chi^2_{p_1 - p_0}$$

$$F = \frac{(D_0 - D_1)/(p_1 - p_0)}{D_1/(n - p_1)} \dot{\sim} F_{p_1 - p_0, n - p_1}$$

Residuals

- Pearson Residuals $\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$
  - Approximately zero mean and variance ϕ
- Deviance Residuals

$$\hat{\epsilon}_i^d = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$$

$$D = \sum_{i=1}^{n} d_i$$

Negative binomial

- Density

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k) \times \Gamma(y+1)} \times \left(\frac{k}{\mu+k}\right)^k \times \left(1 - \frac{k}{\mu+k}\right)^y$$

$$E(Y) = \mu \qquad \text{var}(Y) = \mu + \frac{\mu^2}{k}$$