09/02/16 ALSM 2

DEVIANCE

- Output from GLM in R
- Extra criterion, for how good prediction is.
- Deviance is a log likelihood ratio statistic that compares the saturated model with the proposed GLM model

- Max likelihood estimate of the saturated model is computed by:
$(\hat{\theta}_1, \ldots, \hat{\theta}_N) = \arg\max \{ \log L(\theta_1, \ldots, \theta_N) \}$
and the value $\log L(\hat{\theta}_1, \ldots, \hat{\theta}_N)$ is therefore the maximum value of log likelihood function of the saturated models.

- When considering a GLM, a link function g is used to constrain the parameters such that $\theta_i = g^{-1}(x_i^T \beta)$ $\forall i = 1, \ldots, n$
- In this case likelihood is written $L(\beta)$ and log likelihood $\log[L(\beta)]$
- Parameter β often a lower dimension than θ's ($\dim \beta \leq N$) and the maximum
- likelihood is computed such that:
$\hat{\beta} = \arg\max \{ \log L(\beta) \}$
- Maximum likelihood value for a GLM is then $\log L(\hat{\beta})$

$$\log L(\hat{\beta}) = \sum_{i=1}^{N} a(y_i) b(g^{-1}(x_i^T \hat{\beta})) + c(g^{-1}(x_i^T \hat{\beta})) + d(y_i)$$

Deviance also called log likelihood ratio statistic compares the saturated model with the proposed GLM

$$D = 2[\log L(\hat{\theta}_1, \ldots, \hat{\theta}_N) - \log L(\hat{\beta})] = 2\log\left[\frac{L(\hat{\theta}_1, \ldots, \hat{\theta}_N)}{L(\hat{\beta})}\right] \quad \frac{\text{Saturated}}{\text{GLM}}$$

- $\log(\hat{\theta}_1, \ldots, \hat{\theta}_N)$ max value of log likelihood function for saturated model
- $\log(\hat{\beta})$ value of log likelihood function when fitting model $g(E[y]) = x^T \beta$

Approximation of log likelihood f^n when near its maximum
$\theta = (\theta_1, \ldots, \theta_N)^T$ $\quad \hat{\theta}_i = (\hat{\theta}_1, \ldots, \hat{\theta}_N)$
$\log L(\theta)$ near $(\hat{\theta})$

- Using Taylor expansion, Log likelihood can be approximated near the maximum likelihood estimate
- For Saturated model with notation $(\theta_1, \ldots, \theta_N)^T = 0$, when θ close to $\hat{\theta}$:

$$\log L(\theta) \simeq \log L(\hat{\theta}) + (\theta-\hat{\theta})^T \nabla_{\hat{\theta}} + \frac{1}{2}(\theta-\hat{\theta})^T H_{\hat{\theta}} (\theta-\hat{\theta})$$

$\nabla_{\hat{\theta}}$: Gradient of log L function computed at $\hat{\theta}$

$$\nabla_{\hat{\theta}} = \begin{bmatrix} \frac{d \log(L)}{d\theta_1}\Big|_{\hat{\theta}} \\ \vdots \\ \frac{d \log(L)}{d\theta_N}\Big|_{\hat{\theta}} \end{bmatrix}$$

Vector of $\dim(N)$

At max, derivative $= 0$ all equal 0 bc max $\hat{\theta}$ is max likelihood solution so term is going away

$H_{\hat{\theta}}$: Hessian Matrix of log L f^n computed at $\hat{\theta}$

Matrix will return a negative value (bc it is a second derivative)

Hessian Matrix is Symmetric and $\mathbb{R}$ $\Rightarrow$ can compute eigenvalues/vectors

Similarly for GLM model, when β is close to $\hat{\beta}$:

$$\log L(\beta) \simeq \log L(\hat{\beta}) + (\beta-\hat{\beta})^T \nabla_{\hat{\beta}} + \frac{1}{2}(\beta-\hat{\beta}) H_{\hat{\beta}} (\beta-\hat{\beta})$$

- In both cases $\nabla_{\hat{\beta}}, \nabla_{\hat{\theta}}$ are zero vectors since $\hat{\beta}$ and $\hat{\theta}$ are maxima of log L
- Re-write Deviance: approximation of deviance near $\hat{\theta}$ and $\hat{\beta}$

$$D \simeq 2\{\log L(\hat{\theta}) - \log L(\hat{\beta})\}$$

$$\simeq 2\{\log L(\theta) - \tfrac{1}{2}(\theta-\hat{\theta})^T H_{\hat{\theta}} (\theta-\hat{\theta}) - \log L(\beta) - \tfrac{1}{2}(\beta-\hat{\beta})^T H_{\hat{\beta}} (\beta-\hat{\beta})\}$$

$$\simeq 2 \log\left[\frac{L_\theta}{L_\beta}\right] - (\theta-\hat{\theta})^T H_{\hat{\theta}} (\theta-\hat{\theta}) + (\beta-\hat{\beta}) H_{\hat{\beta}} (\beta-\hat{\beta})$$

- The term v is positive and will be near 0 if GLM model fits data almost as well as the Saturated model does.
- Hessian matrix computed at max $H_{\hat{\beta}}$ $H_{\hat{\theta}}$ will be negative

09/02/16 ALSM2

Sampling Distribution of Deviance

- The likelihood function for β can be approximated by a normal distribution near the estimate $\hat{\beta}$ such that

$$\mathcal{L}(\beta) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left[-\tfrac{1}{2}(\beta-\hat{\beta})^T \Sigma^{-1}(\beta-\hat{\beta})\right]$$

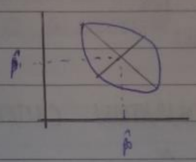
$$\log \mathcal{L}(\beta) = \text{constant} - \tfrac{1}{2}(\beta-\hat{\beta})^T \Sigma^{-1}(\beta-\hat{\beta})$$

$\Sigma^{-1} = -H_{\hat{\beta}}$ (Hessian Matrix)

Can find a set of eigen vectors

$\Sigma = L|\Lambda L|^T$

eigen values (diagonal matrix)



Σ controlling shape

Any Covariance Matrix Σ is symmetric and real

$\rightarrow \Sigma\, LL^T \Lambda\, LL$

Diagonal matrix which have eigenvalues on diagonal and zeros otherwise

eigen vectors (orthonormal basis of Σ) $\qquad LL^T LL = I$

$$(\beta-\hat{\beta})\overset{-H_\beta}{\Sigma^{-1}}(\beta-\hat{\beta}) = [LL(\beta-\hat{\beta})]^T \Lambda^{-1} [LL(\beta-\hat{\beta})]$$

$$\Lambda = \gamma^2 \qquad = [\gamma^{-1} LL(\beta-\hat{\beta})]^T I\, [\gamma^{-1} LL(\beta-\hat{\beta})]$$

Change of variable $z = \gamma^{-1} LL(\beta-\hat{\beta})$

$Z \sim N(0, I)$ (Identity Matrix)

$-(\beta-\hat{\beta})H_{\hat{\beta}}(\beta-\hat{\beta})$ is a sum of m variables with distribution $N(0,1)$

identity matrix in $\mathbb{R}^m$ with $m = \mathrm{Dim}(\beta)$

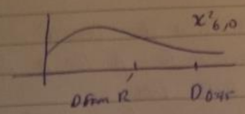
This term $\sum_{i=1}^{m} z_i^2$ has χ^2 dist with m d.f.

$(\theta-\hat{\theta})^T H_{\hat{\theta}}(\theta-\hat{\theta})$ follows $\chi^2(\dim(\theta) = N)$

$$D \simeq 2\,\mathrm{Log}\left[\frac{\mathcal{L}\hat{\theta}}{\mathcal{L}(\hat{\beta})}\right] \qquad \underbrace{-(\theta-\hat{\theta})^T H_{\hat{\theta}}(\theta-\hat{\theta}) + (\beta-\hat{\beta})^T H_{\hat{\beta}}(\beta-\hat{\beta})}_{\chi^2\ (\dim(\theta)-\dim(\beta))}$$

offset should be near zero if good model

$D \sim \chi^2(n-m, N=0)$ Both Dis



Is D value below $D_{0.95}$? if it is $\Rightarrow$ good model
(value from output below that of calculated)
- work backwards in stat table book
- Value from table 12.5
- Value from R: 11.2

AKAIKE INFORMATION CRITERION

- $\{(y_i, x_i)\}$ $i = 1, \ldots, N$

  likelihood: $L(\theta_1, \ldots, \theta_N) = \prod_{i=1}^{N} P_{y|\theta}(y_i \mid \theta_i)$
- with $P_{y|\theta}$ from an exponential family probability distribution

  $L(\theta_1, \ldots, \theta_N) = \prod_{i=1}^{N} \operatorname{Exp}[a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)]$
- Log transform usually computed instead:

  $\log L(\theta_1, \ldots, \theta_N) = \sum_{i=1}^{N} [a(y_i)b(\theta_i) + c(\theta_i) + d(y_i)]$
- MLE for saturated model then

  $(\hat{\theta}_1, \ldots, \theta_N) = \operatorname{argmax} \log L(\theta_1, \ldots, \theta_N)$

- When a GLM is used, a link function g is used to constrain the parameters such that $\theta_i \propto g^{-1}(x_i^T; \beta)$ $\forall i = 1, \ldots, N$
- The parameter β has often a lower dimension than θs i.e. $\dim(\beta) \leq N$ and the maximum likelihood estimator computes: $\hat{\beta} = \operatorname{argmax} \log L(\beta)$

AIC is a measure of goodness of fit defined by: $AIC = -2\log L(\hat{\beta}) + 2p$
where: • $p = \dim(\beta)$ # of parameters to be estimated by the model
• $\hat{\beta}$ are the estimated parameters that maximise the likelihood or log likelihood
• $\log L(\hat{\beta})$ is the maximum value of the log likelihood.
- Best model is a trade off between one that maximises the likelihood with also having the minimum number of parameters
- Select model with lowest AIC

CHI-Square Distribution

Assuming that $Z \sim N(0,1)$ $i=1,\dots,N$ show that $X = Z_i^2$ has a χ^2 distribution

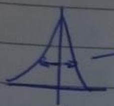
$P_x(x) = \int P_{xz}(x,z)\,dz$

$= \int P_{x|z}(x|z)\,P_z(z)\,dz$ conditional distribution $Z \sim N(0,1)$

No uncertainty in X when given Z (x is z^2 remember)

Dirac Distribution has variance of 0: no uncertainty

$P_{x|z}(x|z) = \delta(x - z^2)$



width of 0 $\sigma^2 = 0$

z^2

$\int_{-\infty}^{\infty} \delta(x-z)^2\,dx = 1$ (Property of dirac function)

$= \int \delta(x-z)^2 \frac{1}{2\pi} \exp\left(\frac{-z^2}{2}\right) dz$

Use Dirac property to rewrite dirac part

$\int \frac{1}{2|\sqrt{x}|} \left[\delta(\sqrt{x}+z) + \delta(\sqrt{x}-z)\right]$

$= \int \frac{1}{2|\sqrt{x}|} \left[\delta(\sqrt{x}+z) + \delta(\sqrt{x}-z)\right] \frac{1}{2\pi} \exp\left[\frac{-z^2}{2}\right] dz$ two

$= \frac{1}{2\sqrt{x}} \left| \int \delta(\sqrt{x}-z) \exp\left(\frac{-z^2}{2}\right) + \int \delta(\sqrt{x}+z) \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right] dz \right.$

Next Property: $\int_{-\infty}^{\infty}$ ~~f(t) d(t-T)~~ $f(t)\,d(t-T)\,dt = f(T)$

$= \frac{1}{2\sqrt{x}} \left| \left[\frac{1}{\sqrt{2\pi}} \exp\left[\frac{-x}{2}\right] + \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-x}{2}\right]\right] \right.$

Chi-Square $f(x,k) = \dfrac{x^{\left(\frac{k}{2}-1\right)} e^{-\frac{x}{2}}}{2^{k/2}\,\Gamma\left(\frac{k}{2}\right)}$ $x \geq 0$

gamma

$E[x] = k$

- look for deviance near value of k as an indicator

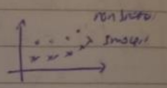
16/02/16

ALSM2

Smoking Dataset

- Split into rows according to yes/no outcome
  $\hat{\theta}$ for saturated system $= y_i/n_i$ for each row

- Nature of explanatory variables
  - nominal red, green, black
  - ordinal natural ordering - week days
  - continuous e.g. weight, age, temperature



For the $\hat{\theta}$'s $\left(\frac{y_i}{n_i}\right)$ draw a graph of all values to see difference
- GLM tries to fit a curve between these points
- Need to determine $\beta^T x$
  → could use age^2
  → Smoking is a binary variable, $smoke^2$ has no effect
  → Multiply smoke by age → 1,2,3,4,5,6,0,0,0,0 ← called "mixed effect" when these variables are used
  → Multiply age^2 by smoke → 1,4,9,16,25,0,0,0,0,0
  → i.e. can make many explanatory variables out of 2 variables

$$\beta_0 + \beta_1 x^{age} + \beta_2 x^{smoke} + \beta_3 x^{age} x^{smoke} + \beta_4 (x^{age})^2 + \beta_5 (x^{age})^2 x^{smoke}$$

→ largest model for what we have defined in our table

- When smokes $x^{smoke} = 1$: $\beta_0 + \beta_2 + (\beta_1 + \beta_3) x^{age} + (\beta_4 + \beta_5)(x^{age})^2$ "mixed model"
  $x^{smoke} = 0$: $\beta_0 + \beta_1 x^{age} + \beta_4 (x^{age})^2$
- 6 parameters here
- Binomial or poisson is choice here

- Try all possible models and choose one with lowest AIC
- Fitted model: $\beta_0 + \beta_1 x^A + \beta_2 x^S + \beta_3 x^A x^S + \beta_4 (x^A)^2$ has lowest AIC.
- AIC: Poisson + log link : 66.7
  Binomial + logit : 66.63
  Binomial + probit : 66.33
- Tried with age^2 removed - made AIC worse

- When mixed effect was removed, AIC increased
- Write the GLM $g(\mu) = \beta^T x$
- Poisson + log: $Log(\mu) = \hat{\beta}_0 + \hat{\beta}_1 x^a + \hat{\beta}_1 x^s + \hat{\beta}_1 x^a x^s + \hat{\beta}_4 x^{a^2}$
- Binomial logit: $\log\left[\frac{\theta}{1-\theta}\right] = \hat{\beta}_0 \ldots$ β's will be different because of different likelihood calculation

Poisson: $P(y|\lambda) = \frac{\lambda^y}{y!} \exp(-\lambda)$ 10 groups, $N=10$

$L(\lambda_1, \ldots, \lambda_{10}) = \prod_{i=1}^{10} \frac{\lambda_i^{y_i}}{y_i!} \exp(-\lambda_i)$

$Log(L) = \sum_{i=1}^{10} y_i \, Log(\lambda_i) - \lambda_i - \log(y_i!)$

$\hat{\lambda}_i = y_i$ Saturated model - Max likelihood situation

Deviance/AIC: $Log\, L(\hat{\lambda}_1, \ldots, \hat{\lambda}_{10})$

$-n\lambda_i + \sum y_i \log(\lambda) - \log(\prod y_i)$

$\frac{d\,Log(L)}{d\lambda} = -n + \sum y_i \frac{1}{\lambda}$

$\hat{\lambda} = \sum y / n$

Binomial: $P(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}$

$L(\theta_1, \ldots, \theta_{10}) = \prod_{i=1}^{10} \binom{n_i}{y_i} \theta_i^{y_i} (1-\theta_i)^{n-y_i}$ (Applied to each group)

$L(\theta_1, \ldots, \theta_{10}) = \left(\prod_{i=1}^{10} \left(\frac{n!}{x_i!(n-x_i)!}\right)\right) p^{\sum x_i} (1-p)^{n-\sum x_i}$

$Log\, L = \sum y_i \log(p) + (n - \sum y_i) \log(1-p)$

$\frac{d\,Log\, L}{dp} = \frac{1}{p_i} \sum y_i + \frac{1}{1-p}(n - \sum y_i) = 0$

$\hat{\theta}_i = \sum y_i / n_i$

Saturated model replace θ_i with $\hat{\theta}_i = y_i / n_i$

Difference between age? look at coefficients β's to determine.
- look at standard errors associated with the β's
- "couldn't remove an explanatory variable from model → variable is important"