#7 F2017

Residuals

Summary

- Inference for GLM
  - Wald test; Score test; LR test.
  - Goodness of Fit: Deviance, Pearson's χ^2
- Interval estimation for β

Reading

- McCullagh and Nelder (1989) Chapter 2 and Chapter 12.
- Dobson and Bartnett (2008) Chapter 7 and Chapter 9.

Residuals

- Revisit: linear model
  - $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\mu}_i$.
  - Residual plots (e.g., e_i vs. predicted .)
  - Normal $Q - Q$ plots for residuals
- In GLM, residuals can be used for
  - the adequacy of fit
  - check the choice of link function
  - detection of outliers

#7 — F2017

Types of residuals for GLM

Recall $\chi^2 = \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/m_i} = \sum_i (r_i^P)^2$

1. **Pearson residual**

$$r_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)/m_i}},$$

For normal $r_i^P = Y_i - \hat{\mu}_i = r_i^R$ (since $V(\hat{\mu}_i) = 1$, $m_i = 1$)

For $Y_i \sim Bin(p_i, m_i)$, consider Y_i/m_i,

$$r_i^P = \frac{Y_i/m_i - \hat{p}_i}{\sqrt{V(\hat{\mu}_i)/m_i}} = \frac{Y_i/m_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)/m_i}} = \frac{Y_i - m_i\hat{p}_i}{\sqrt{m_i\hat{p}_i(1-\hat{p}_i)}}$$

($\hat{\mu}_i$ is the variance function. Thus,)

In proc reg; output residual = r;

In Proc genmod; output resraw = response

$$\sum_{i=1}^n (r_i^p)^2 = X^2$$

- R: resid(glm.fit, "pearson")
- SAS: output **reschi**=pearson

For Poisson, $V(\mu_i) = \mu_i$, $m_i = 1$, $r_i^P = \frac{Y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$

2. **Deviance residual**

$$\sum_{i=1}^n (r_i^D)^2 \qquad r_i^D = \text{sign}(Y_i - \hat{\mu}_i)\sqrt{d_i},$$

where $D(\mathbf{Y}, \hat{\mu}) = \sum_{i=1}^n d_i$, d_i is the ith subject's contribution to the deviance. Thus,

$$\sum_{i=1}^n (r_i^D)^2 = \sum_{i=1}^n d_i = D(\mathbf{Y}, \hat{\mu}).$$

- R: resid(glm.fit, "deviance")
- SAS: output **resdev**=deviance

e.g., for Binomial data,

$$d_k = 2\left[Y_i \log\left(\frac{y_k}{n_k\hat{\pi}_k}\right) + (n_k - y_k)\log\left(\frac{n_k - y_k}{n_k - n_k\hat{\pi}_{i2}}\right)\right], \quad \hat{\pi}_k = \hat{\mu}_k,$$

See (7.9) on P. 1

Here, we use r_i^D for d_i.

See P.138 in D & B

3

3. **Working residual**

$$r_i^W = Z_i - \hat{\eta}_i = (Y_i - \hat{\mu}_i)\frac{\partial \hat{\eta}_i}{\partial \hat{\mu}_i}.$$

- R: resid(glm.fit, "working")

4. **Response residual**

$$r_i^R = Y_i - \hat{\mu}_i.$$

- R: resid(glm.fit, "response")
- SAS: output **resraw**=response

5. For the normal distribution, **Pearson = Deviance = Working = Response residual**

Standardized Residuals for GLM

- Revisit: linear model
  - Studentized residual $e_i = y_i - \hat{\mu}_i = y_i - \hat{y}_i$, since $Var(e_i) = \sigma^2(1-h_{ii})$

$$r_i = \frac{e_i}{\sqrt{\widehat{\text{Var}}(e_i)}} = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}},$$

where h_{ii} is the ith diagonal element of hat matrix $\mathbf{H}$ $= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)$.

- In GLM,
  - Define "Hat" matrix (similar idea as in linear model, see Section 7.6)

$\mathbf{H} =$ the same

Let h_{ii} be the ith diagonal element of $\mathbf{H}$.

  - Studentized standardized Pearson residual

$$r_i^{P'} = \frac{Y_i - \hat{\mu}_i}{\sqrt{a_i(\hat{\phi})V(\hat{\mu}_i)(1-h_{ii})}}$$

($a_i(\hat{\phi}) = \phi/m_i$)

Compared to $r_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$

    * SAS: output **stdreschi**=stdpearson
  - Studentized standardized deviance residual

$$r_i^{D'} = \frac{r_i^D}{\sqrt{\hat{\phi}(1-h_{ii})}}.$$

    * SAS: output **stdresdev**=stddeviance

Standardized Residuals for GLM (cont'd)

- In SAS Proc GENMOD,

  `reschi=` `resdev=` `resraw=` `stdreschi=` `stdresdev=`

- In R, Use package "boot" > library(boot) R help: > ?glm.diag

In SAS Proc genmod, Then in R

```
> glm.diag(glm.fit)
```

Std reschi = `> glm.diag(glm.fit)$rp` = standardized Pearson residual

Std resdev = `> glm.diag(glm.fit)$rd` = standardized deviance residual

`> glm.diag(glm.fit)$h` = leverage of the observation hii,

`> glm.diag(glm.fit)$res` = jackknife deviance residual

  - Always use *help*

    `>help('glm.diag')`

  - Guideline for detection of outliers and influence observations may be defined similarly as for linear models.

Model Diagnostics

- Using residuals (just like in linear regression),
  - Residual vs. each covariate X_j: check linearity
  - Normal $Q-Q$ plot of standardized residuals: approximately normal
  - Residual vs. index of measurements: check correlation
- See the example of seizure data.

Example: Seizure Data

- These data arise from a clinical trial of $N = 59$ epileptics. Patients suffering from simple or complex partial seizures were randomized to receive either the antiepileptic drug progabide or a placebo, as an adjuvant to standard chemotherapy. At each of four successive post-randomization visits, the number of seizures occurring over the previous two weeks was collected. Baseline data at entry included the number of epileptic seizures recorded in the preceding 8-week period and patient age in years. [See Thall and Vail (1990), *Biometrics*, **46**, 657-671 for details].

  The R and SAS code for this example is attached.

Example: Seizure Data (cont'd)

- Variables:
  - $y1$. seizure count for weeks 0-2 post treatment assignment
  - $y2$: seizure count for weeks 2-4 post treatment assignment
  - $y3$: seizure count for weeks 4-6 post treatment assignment
  - $y4$: seizure count for weeks 6-8 post treatment assignment
  - trt: treatment assignment (0=placebo, 1=progabide)
  - base: seizure count for weeks 8 weeks prior to treatment assignment
  - age: subject age (in years)
- Seizure data:

baseline data

y1	y2	y3	y4	trt	base	age
5	3	3	3	0	11	31

↓ # of epileptic seizures

Model Diagnostics: GOF

- Recall: If model is correct:

Recall : $\frac{D(y,\hat{\mu})}{\phi} = 2\phi[\ell(y,\phi|y)-\ell(\hat{\mu},\phi|y)] \overset{\cdot}{\sim} \frac{D}{\phi}\left(\text{ or } \frac{X^2}{\phi}\right) \sim \chi^2_{n-q}$.

D does not depend on ϕ

Thus, we expect

$$\frac{D}{n-q}\left(\text{ or } \frac{X^2}{n-q}\right) \approx \phi.$$

- Example:
  - Normal

$$\frac{D}{n-q} = \frac{X^2}{n-q} = \frac{\sum(Y_i-\hat{\mu}_i)^2}{n-q} \approx \sigma^2$$

  - Binomial

$Y_i^* \sim Bin(p_i, m_i)$, $\frac{X^2}{n-q} = \sum_i \frac{m_i(Y_i-\hat{\mu}_i)^2}{(n-q)\hat{\mu}_i(1-\hat{\mu}_i)} \approx 1.$

$Y_i = Y_i^*/m_i$.

  - Poisson

$$\chi^2 = \sum \frac{(Y_i-\hat{\mu}_i)^2}{(n-q)\hat{\mu}_i} \approx 1$$

Model Diagnostics: GOF (cont'd)

- For Binomial and Poisson models, if $X^2/(n-q)$ (or $D(\mathbf{Y},\hat{\mu})/(n-q)$) is much larger than 1, then the model is suspect. The data are said to be *over-dispersed*. Then, all the inferences based on the current model are invalid.
- Example. Seizure data use y_4 as response variable, see seizure.sas,
  - See the SAS or R outputs.

$df = n - q = 59 - 4 = 55$, $\frac{D(\mathbf{Y},\hat{\mu})}{n-q} = \frac{144.5692}{59-4} = 2.6285 > 1$

$\frac{X^2}{n-q} = \frac{133.585}{59-4} = 2.4288 > 1$

Thus, these data are over-dispersed (i.e. inflated variance).

Implying $\phi > 1$

Over-dispersion

- Reading:
  - Dobson and Bartnett Chapter 7 (Binomial), Chapter 9 (Poisson)
  - MN Section 4.5 (Binomial), Section 6.2.3 (Poisson)
- Binomial and Poisson data often have greater variance than expected:
  - Binomial: $Y^* = Y/m$, $E[Y^*] = \mu$, $V[Y^*] > \frac{\mu(1-\mu)}{m}$
  - Poisson: $E[Y] = \mu$, $V[Y] > \mu$

  Such data are *over-dispersed*.

Over-dispersion (cont'd)

- **Over-dispersion**: if for any GLM,

$$V[Y] > a(\phi)V(\mu).$$

  - Consequences
    - Usual GLM inferences invalid
    - SE's too small since variance is underestimated
  - Correction
    - Dispersion (scale) parameter
    - Mixture model

Example: Over-dispersion

- Example [Seed data. Crowder M.J. (1978), *JRSS C, Appl. Stat.*, **27** 34-37]. The R and SAS code for this example is attached.
  - Study of germination of 2 types of seeds treated with 2 root extracts.
  - Variables
    - seed=types 1, 2; extract=extracts 1,2;
    - r = number of germinated seeds; m = number of seeds on plate.
  - Binomial GLM with logit link:
    - $Y = r/m$ = proportion of germinated seeds
    - $E[Y] = \mu$ = probability of germination
    - $V[Y] = \mu(1-\mu)/m$,

$$\log\left(\frac{\mu}{1-\mu}\right) = \beta_0 + \beta_1 S + \beta_2 E + \beta_3 S \times E.$$

  - Seed data:

seed	extract	r	m		Y
1	1	10	39	0	2564103
1	1	23	62	0	3709677
1	1	23	81	0	2839506
1	1	26	51	0	5098039
1	1	17	39	0	4358974
1	2	5	6	0	8333333
1	2	53	74	0	7162162

Example: Over-dispersion (cont'd)

- Interpretation of over-dispersion for seed data.
- Is the binomial distribution reasonable?
  - Yes, for a **Single plate**
    - seeds are independent with same probability of germination.
  - No, for a **Across plates** within test conditions
    - Probability of germination may vary across plates due to

      amount of extract

      extract batch

      seed batch

      incubation time, temperature etc.
  - This is an example of a mixture model: a mixture of plates (clusters) with possibly different germination probabilities.

Over-dispersion (cont'd)

- General approach to over-dispersed GLMs:
- Is the binomial distribution reasonable?
  - Assume

$$V[\mu]^* = \sigma^2 V(\mu).$$

Since $w_i = \frac{1}{Var(Y_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2} = \frac{1}{V(\mu_i)a(\phi)[g'(\mu_i)]}$

then $(w_i/\sigma^2)^{-1} = \sigma^2 w_i^{-1}$

  - Properties: (later in quasi-likelihood theory)
    - Distribution of $\hat{\beta}$ Recall $\hat{\beta} \sim MVN(\beta, (X^T W X)^{-1})$ replace W by W/σ^2, then

$$\hat{\beta} \sim MVN(\beta, \sigma^2(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}).$$

    - Analysis of deviance:

$$D(\text{Reduced model}) - D(\text{Full model}) \sim \sigma^2 \chi^2_{p-q}.$$

If σ^2 is larger than 1 the p-value becomes larger

Recall: in our definition, $D = \phi \log(\text{likelihood ratio})$

$= \sigma^2 \log(\text{likelihood ratio})$

Scaled deviance $D^* = D/\phi \sim \log(\text{likelihood ratio}) \sim \chi^2(p-q)$

STAT 635-GLM-Lecture Notes 7, Diagnostics for Generalized Linear Models, Fall 2017

    - Estimation of dispersion parameter:

$$\tilde{\sigma}^2 = \frac{X^2}{n-q} = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{(n-q)V(\hat{\mu})},$$

R uses this based on some quasi distribution, e.g., quasi binomial or quasiPoisson

R uses this,

or

$$\tilde{\sigma}^2 = \frac{D(\mathbf{Y}, \hat{\mu})}{n-q},$$

SAS uses this or $\chi^2/(n-q)$ depending on scale = deviance or scale = pearson.

SAS uses this or $\frac{X^2}{n-q}$ depends on the option in Proc GENMOD, scale= deviance or pearson.

Recall Note#6, $D/\phi \sim \chi^2(n-q)$

Over-dispersion (cont'd)

- Example 1. Seed data
- Example 2. Seizure data
- R and SAS commands for correcting standard errors automatically by using the scale parameter
  - R
    - Logistic regression:
      `glm( , family=quasibinomial, )` R uses $\frac{X^2}{n-g}$
    - Poisson regression:
      `glm( , family=quasipoisson, )`
  - SAS
    - Logistic regression:
      `proc genmod ; model / dist=bin scale=pearson;`

SAS uses $\frac{D}{n-g}$ (scale = deviance) or $\frac{X^2}{n-g}$ (scale = pearson)

Over-dispersion (cont'd)

- Where over-dispersion does not exist:
  - Bernoulli 0, 1 variable (i.e. upgrouped binary data):
    - e.g., logistic regression with continuous covariates
  - One batch per condition
    - e.g., in Seed data example, if for each seed-extract combination only one batch of seed had been observed. But in the current example, there are five batches for each combination.
- Underdispersion
  - The opposite of overdispersion with $\sigma^2 < 1$.
  - It occurs less frequently in practice (usually not our concern)
  - e.g.,
    - in competition situations: seeds in a batch compete for fertilizer the winner grows, the looser does not.

Pearson's χ^2

- Revisit: Pearson's χ^2

$$X^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)/m_i}$$ see Note #6

  - Large X^2 implies
    - systematic deficiencies of the model (poor fit) possibly due to
      - wrong link
      - missing covariate
      - necessity to transform some covariates
      - outlying observations
    - unexplained random variation (over-dispersion)
      - random variation in response probabilities
      - correlation between binary responses
- Remove systematic deficiencies before looking into over-dispersion.

- For examples of over-dispersion, see SAS codes `budworm.sas`, `seed sas` and `seizure.sas` and the attached R and SAS programs.