STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2018

Shivon Sue-Chee

UNIVERSITY OF TORONTO

February 13-15, 2018

STA 303/1002: Class 12- Logistic Regression

- What did we learn about Binary Logistic Regression?
  - Underlying probability distribution of response: Bernoulli
    - Outcome: Response variable, Y-binary
  - Model:
    $$\log\left(\frac{\pi}{1-\pi}\right) = f(\mathbf{X};\boldsymbol{\beta})$$
    where $f(\mathbf{X};\boldsymbol{\beta})$ is a linear function of the β's
    - Predictor variables, $\mathbf{X}$: categorical and/or continuous
  - Estimation: MLE via Fisher scoring algorithm
  - Interpretation of β's: Hold other X's constant, the odds of Y=1 change by factor of e^{β}.
  - Estimate Odds, Odds ratio, $e^{\beta(a-b)}$
  - Inference:
    - Wald tests and confidence intervals
    - Compare models: LRT: 1) $> 1\ \beta$, 2) $1\ \beta$, 3) Global

Binomial Logistic Regression

- What did we learn about Binomial Logistic Regression?
  - Underlying probability distribution of response: Binomial
    - Outcome: Response variable, Y-count variable
  - Model:
    $$\log\left(\frac{\pi}{1-\pi}\right) = f(\mathbf{X}; \boldsymbol{\beta})$$
    where $f(\mathbf{X}; \boldsymbol{\beta})$ is a linear function of the β's
  - Estimate Odds, Odds Ratio and π
  - Inference: Wald or LRT
  - We can do more tests for model adequacy than in Binary logistic regression.
- Deviance GOF test: Fitted vs Saturated
- Quote of the week: "All models are wrong but some are useful."- *Unknown*.

Which is an example of a Generalized Linear Model?

(a) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

(b) $\mu[Y|X_1] = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$

(c) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

(d) $\mu[Y|X_1, X_2] = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2)$

(e) $\mu[Y|X_1] = \beta_0 + \beta_1 10^{X_1}$

(f) $\mu[Y|X_1, X_2] = \dfrac{\beta_0 + \beta_1 X_1}{\beta_0 + \beta_2 X_2}$

(g) $\mu[Y|X_1] = \beta_0 + \exp(\beta_1 X_1)$

(h) $\mu[Y|X_1] = \beta_0 \exp(\beta_1 X_1)$

(i) $\mu[Y|X_1, X_2] = \beta_1 X_1 \exp(\beta_0 + \beta_2 X_2)$

Which is false?

(i) A Logistic regression model is a Generalized Linear Model.

(ii) Logistic regression assumes that there is a linear relationship between logits and explanatory variables.

(iii) Logistic regression describes population proportion or probability as a linear function of explanatory variables.

(iv) Logistic regression is a nonlinear regression model.

Model Assumptions for Binomial Logistic Regression

1. Underlying probability model for response is Binomial.
   - Variance is not constant; is a function of the mean.
2. Observations are independent.
3. The form of the model is correct
   - Linear relationship between logits and explanatory variables
   - All relevant variables are included; irrelevant ones excluded
4. Sample size is large enough for valid inference-tests and CIs. (Recall large-sample properties of MLEs.)
   - Check for outliers.

What is the SATURATED Model?

- Observed response proportion:

$$\bar{\pi}_i = \frac{y_i}{m_i}$$

- Observed or Empirical logits: (S-"saturated")

$$\log\left(\frac{\bar{\pi}_{S,i}}{1-\bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right)$$

- Fits the model exactly with the data
- Most general model possible for the data.

Which Models are often compared?

Consider one explanatory variable, X with n unique levels for the outcome, $Y \sim (Bin(m, \pi))$

- Saturated (FULL) Model: as many parameter coefficients as n

$$logit(\widehat{\pi}) = \widehat{\alpha}_0 + \widehat{\alpha}_1 \mathbb{1}_1 + \cdots + \widehat{\alpha}_{n-1} \mathbb{1}_{n-1}$$

- Fitted (REDUCED) Model: nested within a FULL model; has $(p+1)$ parameters

$$logit(\widehat{\pi}) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- NULL Model: Intercept only model

$$logit(\widehat{\pi}) = \widehat{\gamma}_0$$

Checking model adequacy: Form of the model

Deviance Goodness -Of -Fit (G-O-F) Test

- To check model adequacy in binomial logistic regression, we can use the Deviance Goodness -Of -Fit (G-O-F) Test.
- Analogous to GOF test for comparing 2 models in Linear Regression.

- Form of hypotheses: H_0: REDUCED model, H_a: FULL model
- The DEVIANCE GOF test compares the fitted model (M) to the saturated model (S).

$$H_0 : (Fitted) logit(\widehat{\pi}) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

$$H_a : (Saturated) logit(\widehat{\pi}) = \widehat{\alpha}_0 + \widehat{\alpha}_1 \mathbb{1}_1 + \cdots + \widehat{\alpha}_{n-1} \mathbb{1}_{n-1}$$

Compared to Saturated model: Deviance G-O-F test

- Uses LRT
- Sometimes called "Drop-in-Deviance" test
- as extra-sum-of-squares tests; based on the deviance residual
- Hypotheses:

  H_0: $logit(\pi) = \alpha_0 + \alpha_1 X$

  (Fitted model fits data as well as Saturated model)

  H_a: $logit(\pi) = \beta_0 + \beta_1 \mathbb{1}_1 + \cdots + \beta_{n-1} \mathbb{1}_{n-1}$

  (Saturated model is better)
- Test Statistic:

$$Deviance = -2\log\left(\frac{\mathcal{L}_R}{\mathcal{L}_F}\right) = -2\log\left(\frac{\mathcal{L}_M}{\mathcal{L}_S}\right)$$

- Under H_0, $Deviance \sim$ Chi-square distribution with $n-(p+1)$ df.
- Warning: This is an asymptotic approximation, so it works better if each $m_i > 5$.)

Calculating the Deviance test statistic

Recall underlying model of Y: $Y_i \sim Binomial(m_i, \pi_i)$

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}, \quad y_i = 0, 1, \ldots, m_i$$

Hence the likelihood is:

$$\mathcal{L} = \Pi_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})}$$

Calculating the Deviance test statistic

$\hat{\pi}_i$

Then the log-likelihood is:

$$\log \mathcal{L} = \quad \sum_{i=1}^{n} [y_i \log(\pi_i) + (m_i - y_i)\log(1 - \pi_i) + \underline{\log \binom{m_i}{y_i}}]$$

The deviance test statistic is based on a ratio of likelihoods.

$$\begin{aligned} Deviance &= -2 \log \frac{\mathcal{L}_M}{\mathcal{L}_S} \\ &= -2(\log \mathcal{L}_M - \log \mathcal{L}_S) \\ &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \end{aligned}$$

- Q: A Saturated Model has *Deviance* =

Calculating the Deviance test statistic

$$
\begin{aligned}
Deviance &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \\
&= 2\sum_{i=1}^{n} \Big[y_i \log\left(\frac{y_i}{m_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i}\right) + \log\binom{m_i}{y_i} \\
&\qquad - y_i \log\left(\frac{\widehat{y}_i}{m_i}\right) - (m_i - y_i)\log\left(\frac{m_i - \widehat{y}_i}{m_i}\right) - \log\binom{m_i}{y_i} \Big] \\
&= 2\sum_{i=1}^{n} \Big[y_i \log(y_i) + (m_i - y_i)\log(m_i - y_i) \\
&\qquad - y_i \log(\widehat{y}_i) - (m_i - y_i)\log(m_i - \widehat{y}_i) \Big] \\
&= 2\sum_{i=1}^{n} \Big[y_i \log\left(\frac{y_i}{\widehat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \widehat{y}_i}\right) \Big]
\end{aligned}
$$

Case Study IV Exercise: Using Deviance

Using R output,

Q: Determine whether a saturated model is an improvement over the simpler model with linear function of log(*Area*).

(In R, we get deviance of a model by using `deviance('fittedmodel')`)

- Hypotheses:

- Test Statistic: Deviance=12.062

- Distribution of TS:

- P-value:

- Conclusion: The data are consistent with H_0; the simpler model with linear function of log(*Area*) is adequate (fits as well as the saturated model).

Binomial Logistic Regression: Interpreting Deviance

- Smaller deviance leads to larger *p*-value and vice versa.
- Large *p*-values means:
  - Fitted model is adequate, OR
  - Test is not powerful enough to detect inadequacies
- Small *p*-values means:
  - Fitted model is not adequate; consider a more complex model with more explanatory variables or higher order terms and so on, OR
  - Response distribution is not adequately modelled by the Binomial distribution, OR
  - There are severe outliers.

Can we do a Deviance GOF test in Binary case?

In Binary logistic regression case, $m_i = 1$ for all i, and $y_i = \begin{cases} 0 \\ 1 \end{cases}$

Then deviance becomes:

$$\begin{aligned} \mathit{Deviance} &= 2\sum_{i=1}^{n} \big[y_i \log(y_i) + (1 - y_i)\log(1 - y_i) \\ &\qquad - y_i \log(\widehat{y}_i) - (1 - y_i)\log(1 - \widehat{y}_i) \big] \\ &= 2\sum_{i=1}^{n} \big[- y_i \log(\widehat{y}_i) - (1 - y_i)\log(1 - \widehat{y}_i) \big]. \end{aligned}$$

Notice that the terms that came from the saturated model, $\log \mathcal{L}_S$ are gone, so deviance is no longer useful to compare $\mathcal{L}_M$ with $\mathcal{L}_S$.

Model assessment in Binomial Logistic Regression

- Is linear relationship appropriate?
  - Plot observed logit versus quantitative explanatory variable
- Is the form of the model correct?
  - Use Wald or LRT tests
- Is saturated model better than fitted model?
  - Deviance GOF test
- Are there outliers?
  - Examine standardized residuals: Pearson and Deviance Residuals
- Consider other model fit statistics: AIC, BIC
- Other issues/concerns in model fitting

Residuals: Pearson and Deviance

- Response (raw) residuals: (*observed* − *fitted*) proportion

$$\widehat{\pi}_{S,i} - \widehat{\pi}_{M,i} = \frac{y_i}{m_i} - \widehat{\pi}_{M,i}$$

$\pi \in (0,1)$
$\pi_i - \hat{\pi}_i \in (-1,1)$
$(-\infty, \infty)$

- Standardized residuals:
  (1) Pearson Residuals: uses estimate of s.d. of Y (in denominator)

$$P_{res,i} = \frac{y_i - m_i\widehat{\pi}_{M,i}}{\sqrt{m_i\widehat{\pi}_{M,i}(1-\widehat{\pi}_{M,i})}}$$

$(-2,2)$

  (2) Deviance Residuals: defined so that the sum of the squares of the residuals is the deviance

$\sum_i d_i^2 = G^2$ = Deviance of fitted model.

$$D_{res,i} = \text{sign}(y_i - m_i\widehat{\pi}_{M,i}) \times \sqrt{2\left\{y_i \log\left(\frac{y_i}{m_i\widehat{\pi}_{M,i}}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - m_i\widehat{\pi}_{M,i}}\right)\right\}}$$

Response, Pearson and Deviance Residuals in R

- Response residuals

```
residuals(fitbl, type="response")
```

$(-1,1)$.

- Pearson residuals

```
residuals(fitbl, type="pearson")
```

$(-\infty,\infty)$

- Deviance residuals

```
residuals(fitbl, type="deviance")
```

$(-\infty,\infty)$

Case Study IV Example: Were there outliers in the data?

	Pearson, $P_{res,i}$	Deviance, $D_{res,i}$
Asymptotic Dist.	$N(0,1)$	$N(0,1)$
R code	`pearson`	`deviance`
Possible outlier if	$\lvert P_{res,i}\rvert > 2$	$\lvert D_{res,i}\rvert > 2$
Outlier if	$\lvert P_{res,i}\rvert > 3$	$\lvert D_{res,i}\rvert > 3$
Under small n	D_{res} closer to $N(0,1)$ than P_{res}	
$\hat{\pi}$ close to 0 or 1	P_{res} are unstable; related to instability of Wald	

- Results: Both are $< |2|$, so no outliers

Case IV Residuals

```
rres<-residuals(fitbl, type=c("response"))
pres<-residuals(fitbl, type=c("pearson"))
dres<-residuals(fitbl, type=c("deviance"))
rbind(pis,phats,rres, pres,dres)
```

$\hat{\pi}_i = \frac{e^{\hat{\mu}_i}}{1+e^{\hat{\mu}_i}}$

5/75

π_i — $\hat{\pi}_i$ —

```
##              1        2       3        4        5       6       7        8
## pis   0.066667  0.04478 0.15152  0.11765  0.10714 0.20000 0.18605  0.09677
## phats 0.060173  0.07036 0.09854  0.13800  0.15946 0.16205 0.16389  0.17125
## rres  0.006493 -0.02558 0.05298 -0.02035 -0.05232 0.03795 0.02216 -0.07448
## pres  0.236464 -0.81883 1.44400 -0.42139 -0.75619 0.46058 0.39247 -1.10075
## dres  0.232656 -0.87369 1.34958 -0.43071 -0.79584 0.44746 0.38577 -1.18097
##               9       10      11      12      13     14     15       16
## pis    0.178571  0.18750 0.26667  0.1000 0.29032 0.3125 0.4667  0.24242
## phats  0.185415  0.20524 0.22264  0.2516 0.25158 0.2603 0.2842  0.30185
## rres  -0.006844 -0.01774 0.04403 -0.1516 0.03875 0.0522 0.1825 -0.05943
## pres  -0.093181 -0.24850 0.57969 -1.5622 0.49717 0.4759 1.5673 -0.74367
## dres  -0.093632 -0.25127 0.56727 -1.7173 0.48934 0.4666 1.4954 -0.75939
##              17     18
## pis    0.325000 0.5000
## phats  0.327828 0.3998
## rres  -0.002828 0.1002
## pres  -0.038101 0.5008
## dres  -0.038129 0.4957
```

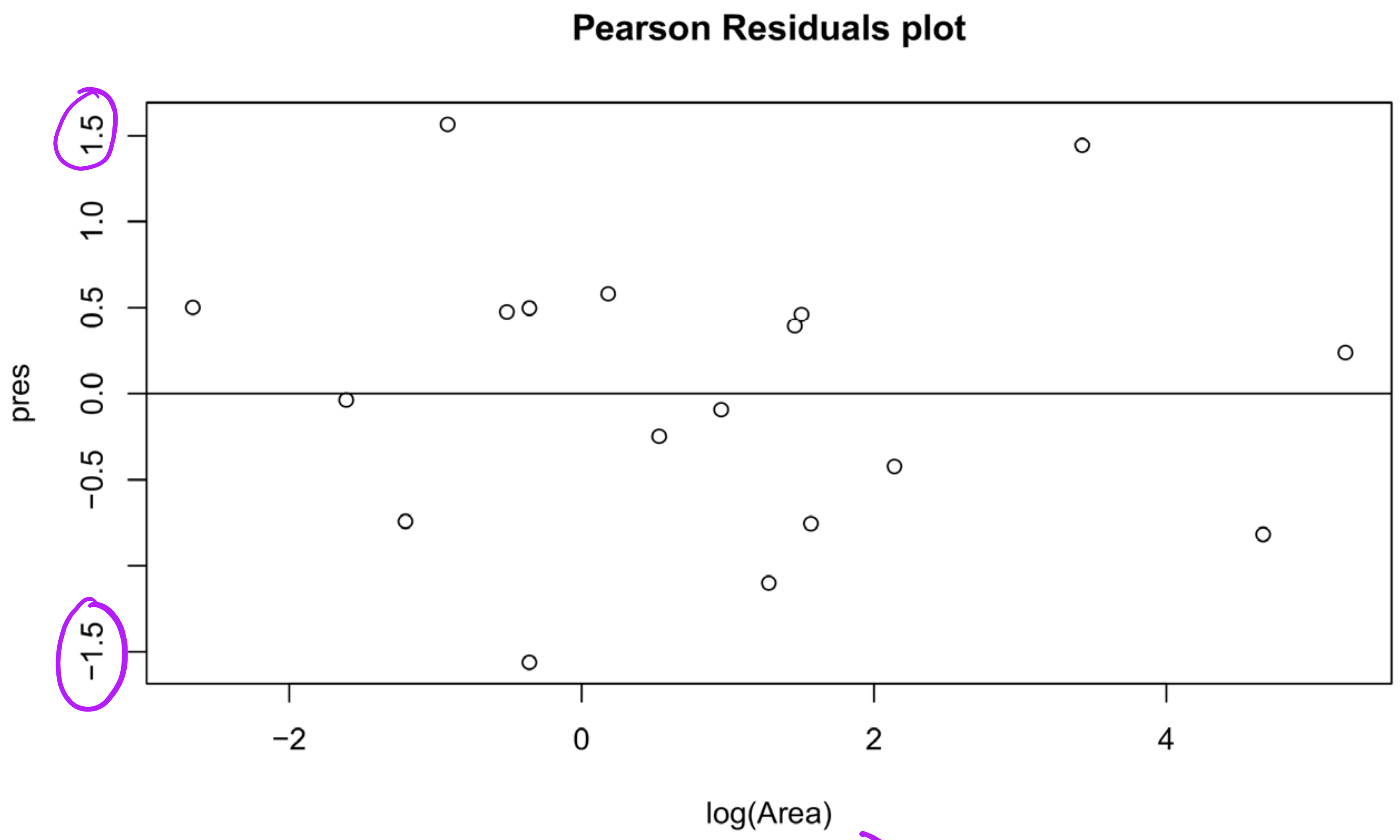
$\pi_i - \hat{\pi}_i$

$P_{res,i}$

$D_{res,i}$

Case IV Residuals Plot

```
plot(log(Area), pres, main="Pearson Residuals plot")
abline(h=0)
```

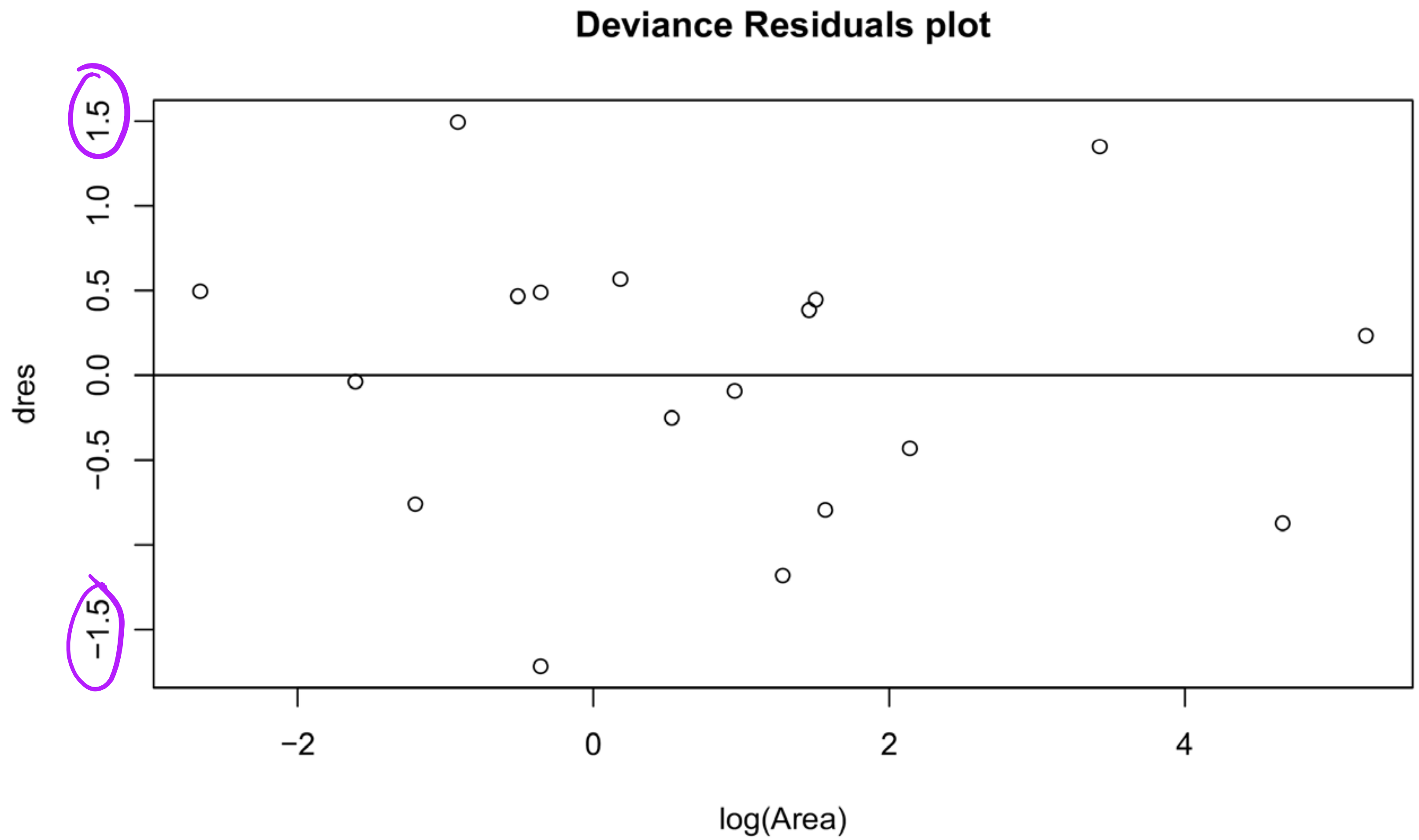


x_i

$\hat{\pi}_i$

Case IV Residuals Plot

```r
plot(log(Area), dres, main="Deviance Residuals plot")
abline(h=0)
```



Other Model Fit Statistics

- Useful for comparing models with same response and same data
- Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty
  1. Akaike's Information Criterion (AIC)

  $$AIC = -2\log\mathcal{L} + 2(p+1)$$

  2. Schwarz's (Bayesian Information) Criterion (BIC)

  $$BIC = -2\log\mathcal{L} + (p+1)\log N$$

  where

  - p-number of explanatory variables, and
  - $N = \sum_{i=1}^{n} m_i$.
- Example: see AIC, BIC for Case IV model

Case IV Fit Statistics

```r
AIC(fitbl)
```

```
## [1] 75.39
```

```r
BIC(fitbl)
```

```
## [1] 77.17
```

Problems and Solutions in Logistic Regression

Problems and Complications common to Linear and Logistic Regression

- *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
  - Unstable fitted equation
  - Coefficient that should be statistically significant is not
  - Coefficient may have the wrong sign
  - Sometimes, large s.e. of $\widehat{\beta}$
  - Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
  - Unstable fitted equation
  - Coefficient that should be statistically significant is not
  - Coefficient may have the wrong sign
  - Sometimes, large s.e. of $\widehat{\beta}$
  - Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\widehat{\beta}$'s, deviance)
- *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Problems and Complications common to Linear and Logistic Regression

- *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\widehat{\beta}$'s, deviance)
- *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Two problems specific to Logistic Regression

$Y_i \sim Bin(m_i, \pi_i)$.

1. Extra-binomial variation

- variance of Y_i greater than $m_i\pi_i(1-\pi_i)$
- also called "over dispersion"
- does not bias $\widehat{\beta}$'s but s.e. of $\widehat{\beta}$'s will be too small (too small p-values, too narrow CIs)

$X_1, X_2, \dots, X_{m_i}$

$X_i \overset{iid}{\sim} Ber(\pi_i)$.

SOLUTION: add one more parameter to the model, ψ- dispersion parameter. Then $\text{Var}(Y_i) = \psi m_i \pi_i (1-\pi_i)$.

$Y_i = \sum_{i=1}^{m_i} X_i$

$\hat{\psi}$

$\boxed{\hat{\psi}\, m_i \pi_i (1-\pi_i).}$ Quasi-binomial

Example of Extra-binomial variation

Hints of Over dispersion:

1. Non iid Ber per grp.
2. Deviance GoF. (Small p-value).
3. Outliers

- Suppose $X_1, \ldots, X_{m_1}$ are not independent but identical Bernoulli(π)
- Suppose all pairs (X_i, X_j) have a common correlation ρ
- Let $Y_1 = \sum_i^{m_1} X_i$

$$
\begin{aligned}
Var(Y_1) &= Var(\sum_i^{m_1} X_i) \\
&= \sum_i^{m_1} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j), \\
&= m_1\pi(1-\pi) + \sum_{i \neq j} \rho \sqrt{Var(X_i)Var(X_j)} \\
&= m_1\pi(1-\pi) + n(n-1)\rho\pi(1-\pi), \quad \texttt{assume} \quad \rho > 0 \\
&> m_1\pi(1-\pi)
\end{aligned}
$$

Estimating Extra-binomial variation

- Model for variance: $Var(Y_i) = \psi m_i \pi_i (1 - \pi_i)$
- Estimate of $\hat{\psi}$: scaled Pearson chi-square statistic,

$$\hat{\psi} = \frac{\sum_i^n P_{res,i}^2}{n-(p+1)} = \frac{\text{sum of squared Pearson residuals}}{d.f.}$$

$$\frac{\chi_{n-(p+1)}}{n-(p+1)}.$$

  - $\hat{\psi} >> 1$ indicates evidence of overdispersion
  - ψ does not affect $E(Y_i)$, hence using overdisperion does not change $\hat{\beta}$
  - $SE(\hat{\beta})$ is multiplied by $\sqrt{\hat{\psi}}$,

$$SE_{\psi}(\hat{\beta}) = \sqrt{\hat{\psi}} SE_{\psi=1}(\hat{\beta})$$

- Overdispersion does not apply to Bernoulli data. If y_i only takes on 0 or 1, then it must be Bernoulli(π_i) and its variance must be $\pi_i(1-\pi_i)$ (McCullagh and Nelder(1989)).

Case Study IV: Logisitc Model with logged explanatory variable

```
fitbl<-glm(cbind(Extinct,NExtinct)~log(Area), family=binomial, data=krunnit)
summary(fitbl)
```

$Var(y_i) = m_i \pi_i (1-\pi_i)$

```
##
## Call:
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = binomial,
##     data = krunnit)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.71726  -0.67722   0.09726   0.48365   1.49545
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
## log(Area)   -0.29710    0.05485  -5.416 6.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 45.338  on 17  degrees of freedom
## Residual deviance: 12.062  on 16  degrees of freedom
## AIC: 75.394
##
## Number of Fisher Scoring iterations: 4
```

Case IV Estimating ψ

$\hat{\psi}$

```
(psihat=sum(residuals(fitbl, type="pearson")^2/fitbl$df.residual))
```

```
## [1] 0.7326
```

```
summary(fitbl, dispersion=psihat)
```

```
##
## Call:
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = binomial,
##     data = krunnit)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7173  -0.6772   0.0973   0.4837   1.4954
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.1962     0.1014  -11.80  < 2e-16 ***
## log(Area)    -0.2971     0.0469   -6.33  2.5e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 0.7326)
##
##     Null deviance: 45.338  on 17  degrees of freedom
## Residual deviance: 12.062  on 16  degrees of freedom
```

$\sqrt{\hat{\psi}}\, SE(\hat{\beta}) = SE_{\psi}(\hat{\beta})$

Case IV: As a Quasi-Binomial

```
fitbl2=glm(cbind(Extinct,NExtinct)~log(Area), family=quasibinomial)
summary(fitbl2)
```

$Var(y_i) = \psi m_i \pi_i (1-\pi_i)$.

```
##
## Call:
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = quasibinomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7173  -0.6772   0.0973   0.4837   1.4954
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.1962     0.1014  -11.80  2.6e-09 ***
## log(Area)    -0.2971     0.0469   -6.33  1.0e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.7326)
##
##     Null deviance: 45.338  on 17  degrees of freedom
## Residual deviance: 12.062  on 16  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Same as p.35

Two problems specific to logistic regression

2. Complete and Quasi-complete separation

- *Complete separation*:
  - one or a linear combination of explanatory variables perfectly predict whether $Y = 1$ or $Y = 0$
  - In Binary response, when $y_i = 1$, $\hat{y}_i = 1$, then $\sum_{i=1}^{n}\{y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\} = 0$.
  - MLE's cannot be computed
- *Quasi-complete separation*:
  - explanatory variables predict $Y = 1$ or $Y = 0$ almost perfectly (just a few points wrong)
  - MLE's are numerically unstable

SOLUTION: simplify the model. Other options- penalized maximum likelihood, exact logistic regression, bayesian methods

Using Logistic Regression for Classification

Using Logistic Regression for Classification

- Want: predict outcome as

$$y^*|(x_1^*, x_2^*, \dots, x_p^*) = \begin{cases} 1 \\ 0 \end{cases}$$

- Do: calculate $\widehat{\pi}_M^*$- the estimated probability that $y^* = 1$ based on the fitted model given $X_1 = x_1^*, X_2 = x_2^*, \dots, X_p = x_p^*$. From this we want to predict that

$$y^* = \begin{cases} 1 & \text{if } \widehat{\pi}_M^* \text{ is large} \\ 0 & \text{if } \widehat{\pi}_M^* \text{ is small} \end{cases}$$

- Need: choose a cut-off probability to distinguish between large and small.

Classification: Approaches to choosing a threshold

Approach 1 - Set cut-off probability as 0.5

- If $\widehat{\pi}_M^* > 0.5$, classify y^* as 1
- Useful if there are equal numbers of 1's and 0's
- Useful if false negatives and false positives are equally bad.

Classification: Approaches to choosing a threshold

Approach 2- Find "best" cut-off probability from data.

- Try different cut-offs and see which gives fewest incorrect classifications
- Useful if proportions of 1's and 0's in data reflect their relative proportions in the population
- Likely to overestimate the proportions of correct predictions that model makes. Then, one should assess model correct classification rates on different data than was used to fit the model.