STAT 226 Lecture 18
Goodness of Fit and the Deviance
Section 5.2.1-5.2.3

Yibi Huang

Binomial response data in grouped data, wide format:

	condition of the trials (explanatory variables)				number of successes	number of failures
Condition 1	x_{11}	x_{12}	$\ldots$	x_{1k}	y_1	$n_1 - y_1$
Condition 2	x_{21}	x_{22}	$\ldots$	x_{2k}	y_2	$n_2 - y_2$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
Condition N	x_{N1}	x_{N2}	$\ldots$	x_{Nk}	y_N	$n_N - y_N$

where $y_1, y_2, \ldots, y_N$ are independent and

$$y_i \sim \text{Binomial}(n_i, \pi(\mathbf{x}_i)).$$

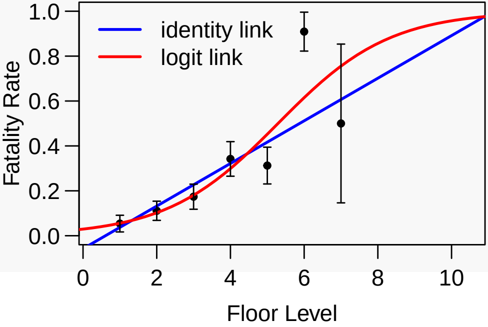
where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$.

E.g., the fatal falls data in Slides L09.pdf are of this form.

Fatal Falls Data

```r
ff = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/falls.txt",
  h=T)
ff
```

```
  floor fatal live
1     1     2   35
2     2     6   48
3     3     8   38
4     4    13   25
5     5    10   22
6     6    10    1
7     7     1    1
```



Which model fits data better, identity link or logit link?

Likelihood Revisit

A way to choose models is to compare their max. (log-)likelihoods.

$$\text{likelihood : } \prod_i [\pi(\mathbf{x}_i)]^{y_i}[1-\pi(\mathbf{x}_i)]^{n_i-y_i}$$

$$\text{log-likelihood : } \sum_i \{y_i \log \pi(\mathbf{x}_i) + (n_i - y_i)\log[1-\pi(\mathbf{x}_i)]\}$$

where

$$\pi(x) = \begin{cases} \alpha + \beta x & \text{for linear prob model (identity link)} \\ \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} & \text{for logistic model (logit link)} \end{cases}$$

For the fatal falls data:

Model	Max. Log-Likelihood
linear (identity link)	-102.4135
logistic (logit link)	-101.1594

The logistic model has a higher max. log-likelihood. Is it better?

Upper Bound of Maximized (Log-)Likelihood

Regardless of the functional form of $\pi(\mathbf{x}_i)$, the likelihood and log-likelihood must be of the form

$$\text{likelihood : } \prod_i [\pi(\mathbf{x}_i)]^{y_i}[1-\pi(\mathbf{x}_i)]^{n_i-y_i}$$

$$\text{log-likelihood : } \sum_i \{y_i \log \pi(\mathbf{x}_i) + (n_i - y_i)\log[1-\pi(\mathbf{x}_i)]\}$$

Since $y_i \log \pi(\mathbf{x}_i) + (n_i - y_i)\log[1-\pi(\mathbf{x}_i)]$ is the log-likelihood for a single observation $y_i \sim$ binomial$(n_i, \pi(\mathbf{x}_i))$, which reaches its max when $\pi(\mathbf{x}_i)$ equals its MLE y_i/n_i, we know

$$y_i \log \widehat{\pi}(\mathbf{x}_i) + (n_i - y_i)\log[1-\widehat{\pi}(\mathbf{x}_i)] \le y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right).$$

So

$$\begin{aligned} &\text{the max. possible log-likelihood of } \textbf{any} \text{ model} \\ &= \sum_i \{y_i \log \widehat{\pi}(\mathbf{x}_i) + (n_i - y_i)\log[1-\widehat{\pi}(\mathbf{x}_i)] \\ &\le \sum_i \left\{y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right)\right\} \end{aligned}$$

floor level x	total falls n_x	fatal falls y_x
1	37	2
2	54	6
3	46	8
4	38	13
5	32	10
6	11	10
7	2	1

For the data of fatal falls, this upper bound for the max. log-likelihood is

$$
\begin{aligned}
& 2\log\left(\frac{2}{37}\right) + (37-2)\log\left(\frac{37-2}{37}\right) \\
& + 6\log\left(\frac{6}{54}\right) + (54-6)\log\left(\frac{54-6}{54}\right) \\
& + \cdots \\
& + 1\log\left(\frac{1}{2}\right) + (2-1)\log\left(\frac{2-1}{2}\right) \\
& = -96.89521
\end{aligned}
$$

Model	Max. Log-Likelihood
linear (identity link)	-102.4135
logistic (logit link)	-101.1594
upper bound	-96.8952

Deviance

The deviance of a model is twice the diff. of its maximized log-likelihood and the upper bound.

$$\begin{aligned}
\text{Deviance} &= -2(\text{max. log-likelihood} - \text{upper bound}) \\
&= -2\Big(\sum\nolimits_i \{y_i \log \widehat{\pi}(\mathbf{x}_i) + (n_i - y_i) \log[1 - \widehat{\pi}(\mathbf{x}_i)]\} \\
&\qquad\qquad - \sum\nolimits_i \left\{ y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i}\right) \right\} \Big) \\
&= 2 \sum\nolimits_i \left\{ y_i \log\left(\frac{y_i}{n_i \widehat{\pi}(\mathbf{x}_i)}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i(1 - \widehat{\pi}(\mathbf{x}_i))}\right) \right\} \\
&= 2 \sum_i (\text{observed}) \log\left(\frac{\text{observed}}{\text{fitted}}\right) \\
&= G^2
\end{aligned}$$

For the logistic model of the fatal falls data,

floor level	observed fatal count	fitted fatal count	observed live count	fitted live count
1	2	2.06	35	34.94
2	6	5.52	48	48.48
3	8	8.31	38	37.69
4	13	11.36	25	26.64
5	10	14.47	22	17.53
6	10	6.76	1	4.24
7	1	1.51	1	0.49

$$
\begin{aligned}
\text{Deviance} = 2\Big[& 2\log\left(\frac{2}{2.06}\right) + 35\log\left(\frac{35}{34.94}\right) \\
&+ 6\log\left(\frac{6}{5.52}\right) + 48\log\left(\frac{48}{48.48}\right) \\
&+ \dots \\
&+ 1\log\left(\frac{1}{1.51}\right) + 1\log\left(\frac{1}{0.49}\right)\Big] \approx 8.5283
\end{aligned}
$$

```r
ff = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/falls.txt",
  h=T)
ff.logit = glm(cbind(fatal,live) ~ floor,
                  family = binomial(link="logit"),data=ff)
```

See next page.

```
summary(ff.logit)

Call:
glm(formula = cbind(fatal, live) ~ floor, family = binomial(link = "log
    data = ff)

Deviance Residuals:
      1        2        3        4        5        6        7
-0.0417   0.2112  -0.1194   0.5726  -1.6135   2.2206  -0.7780

Coefficients:
            Estimate Std. Error z value        Pr(>|z|)
(Intercept)   -3.492      0.501   -6.97 0.0000000000031
floor          0.660      0.125    5.27 0.0000001384974

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 42.0319  on 6  degrees of freedom
Residual deviance:  8.5283  on 5  degrees of freedom
AIC: 33.45
```

The Saturated Model

The upper bound for maximized log-likelihoods itself is also the maximized likelihood for a model — the **saturated model**.

The *saturated model* is the most complex model possible for the data, which has a separate parameter $\pi_i = \pi(\mathbf{x}_i)$ for each (n_i, y_i) and fits the data perfectly that

$$\widehat{\pi_i} = \frac{y_i}{n_i}.$$

Example (Fatal Falls). The saturate model has a separate parameter π_i for each floor level $i = 1, 2, 3 \ldots, 7$.

The Saturated Model

- In the *saturated* model

$$\text{number of parameters} = \text{number of } \textit{rows} \text{ in the data}$$

- If the number of parameters in a model is identical to the number of rows in the data, the model is usually the saturated model.

Example (Mouse Muscle Tension). The saturate model is the 3-way interaction model, for it has 8 parameters, and the data have 8 rows.

- Deviance for the saturated model $= 0$

```
mouse = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/mousemuscle_wide.txt",
  header=T)
mouse
  drug weight muscle tension.High tension.Low
1    1   High      1            3           3
2    1   High      2           23          41
3    1    Low      1           22          45
4    1    Low      2            4           6
5    2   High      1           21          10
6    2   High      2           11          21
7    2    Low      1           32          23
8    2    Low      2           12          22
```

```
mouse$W= mouse$weight
mouse$M= mouse$muscle
mouse$D= as.factor(mouse$drug)
glm3 = glm(cbind(tension.High,tension.Low) ~ W*M*D,
           family=binomial, data=mouse)
```

```
summary(glm3)
```

See next page.

```
glm(formula = cbind(tension.high, tension.low) ~ W * M * D, family = bi
    data = mouse.muscle)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9743     3.4157  -0.285    0.775
WLow         -2.3438     3.8528  -0.608    0.543
M             0.2324     1.7956   0.129    0.897
D             1.5524     1.8611   0.834    0.404
WLow:M        1.3243     2.3163   0.572    0.568
WLow:D        0.7400     2.1398   0.346    0.729
M:D          -0.8105     1.0103  -0.802    0.422
WLow:M:D     -0.4360     1.3071  -0.334    0.739

    Null deviance: 1.9019e+01  on 7  degrees of freedom
Residual deviance: 1.1324e-14  on 0  degrees of freedom
AIC: 46.117
```

Goodness of Fit (GOF) Test and the Deviance

Let L_M be the max. log-likelihood of some Model M of interest. As the upper bound for max. log-likelihood itself is the max. log-likelihood for the saturated model L_S, the *deviance* of Model M

$$\text{Deviance} = -2[L_M - \text{(upper bound)}] = -2(L_M - L_S),$$

is just the likelihood ratio test statistic comparing

$$H_0 : \text{Model } M \quad \text{v.s.} \quad H_a : \text{saturated model.}$$

Deviance has an approx. **chi-squared** distribution w/

$$\begin{aligned} \text{df} &= (\text{\# of parameters in saturated model}) \\ &\qquad - (\text{\# of parameters in Model } M) \\ &= (\text{\# of rows in the data}) - (\text{\# of parameters in Model } M) \end{aligned}$$

However, this approx. is good only when most $y_i \geq 5$ and $n_i - y_i \geq 5$.

Goodness of Fit and the Deviance

- Large deviance indicates lack of fit
- Small deviance means the model fits nearly as good as the best possible model

Goodness of Fit test for the four models of fatal falls data:

Model	Deviance	d.f.	P-value
linear (identity link)	11.0365	5	0.0507
logistic (logit link)	8.5283	5	0.1294

```r
# pchisq(deviance, df, lower.tail=FALSE)  # GOF test P-value
pchisq(11.0365, df=5, lower.tail=FALSE)
[1] 0.05066
pchisq(8.5283, df=5, lower.tail=FALSE)
[1] 0.1294
```

Goodness-of-fit tests show the logistic model fit a bit better than the model w/ identity link.

For the mouse muscle tension data, the saturated model is the 3-way interaction model, the Goodness of fit test of a model is simply comparing the model with the 3-way interaction model.

```
glm3 = glm(cbind(tension.High,tension.Low) ~ W*M*D, family=binomial, da
glm2 = glm(cbind(tension.High,tension.Low) ~ M*D, family=binomial, data
glm2$deviance
[1] 1.529
anova(glm2, glm3,test="Chisq")
Analysis of Deviance Table

Model 1: cbind(tension.High, tension.Low) ~ M * D
Model 2: cbind(tension.High, tension.Low) ~ W * M * D
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4       1.53
2         0       0.00  4     1.53     0.82
```

Observe the deviance of the $M * D$ model 1.53 is exactly the LR statistic comparing $M * D$ with $W * M * D$.

Goodness-of Fit Based on Pearson's Chi-Squared

One can also use Pearson's Chi-Squared statistic

$$X^2 = \sum_i \left\{ \frac{(y_i - n_i\pi(\mathbf{x}_i))^2}{n_i\widehat{\pi}(\mathbf{x}_i)} + \frac{[n_i - y_i - n_i(1 - \widehat{\pi}(\mathbf{x}_i))]^2}{n_i(1 - \widehat{\pi}(\mathbf{x}_i))} \right\}$$

$$= \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}}$$

to do goodness-of-fit test comparing

$$H_0 : \text{Model } M \quad \text{v.s.} \quad H_a : \text{saturated model.}$$

X^2 is different from Deviance but it has an approx. **chi-squared** distribution w/ same d.f. as Deviance.

Like deviance, the approx. for X^2 is good only when all observations (n_i, y_i) have large n_i.

Grouped Data v.s. Ungrouped Data

Although the ML estimates of parameters are the same for grouped or ungrouped data, the deviances are different.

For ungrouped data, $n_i = 1$ for all i and $y_i = 0$ or 1, so

$$\begin{aligned} L_S &= \sum\nolimits_i \left\{ y_i \log\left(\frac{y_i}{n_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i}\right)\right\} \\ &= \sum\nolimits_i \{y_i \log(y_i) + (1 - y_i)\log(1 - y_i)\} = 0 \end{aligned}$$

and hence

$$\text{Deviance} = -2(L_M - L_S) = -2L_M.$$

Grouped Data v.s. Ungrouped Data

```
ff = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/falls.txt",
  h=T)  # Grouped data
ff.ug = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/fallsUG.txt",
  h=T)   # Ungrouped data
```

```
ff.ug
    floor outcome
1       2    live
2       5    live
3       5    live
4       2    live
5       1    live
(... omitted...)
219     1    live
220     4    live
```

Grouped Data v.s. Ungrouped Data

```
ff.logit = glm(cbind(fatal,live) ~ floor, family=binomial, data=ff)
ffug.logit =glm((outcome == "fatal")~floor,family=binomial, data=ff.ug)

ff.logit$coef
(Intercept)       floor
     -3.492       0.660
ffug.logit$coef             # same coefficient estimates
(Intercept)       floor
     -3.492       0.660

ff.logit$deviance
[1] 8.528
ffug.logit$deviance         # different deviances
[1] 202.3

ff.logit$df.residual        # different df for deviances
[1] 5
ffug.logit$df.residual
[1] 218
```

Grouped Data, Ungrouped Data, Continuous Predictors

- *GOF test only apply on Grouped Data*.
  Deviances computed from ungrouped data don’t not have an approx. chi-squared distribution.

Grouped Data, Ungrouped Data, Continuous Predictors

- *GOF test only apply on Grouped Data*.
  Deviances computed from ungrouped data don't not have an approx. chi-squared distribution.
- **Continuous predictors** usually have too many levels (e.g., Width in horseshoe crabs data) that the deviance of model w/ such predictors do not have approx. chi-squared dist. if the number of observations at each level is too small.

Grouped Data, Ungrouped Data, Continuous Predictors

- *GOF test only apply on Grouped Data*.
  Deviances computed from ungrouped data don't not have an approx. chi-squared distribution.
- **Continuous predictors** usually have too many levels (e.g., Width in horseshoe crabs data) that the deviance of model w/ such predictors do not have approx. chi-squared dist. if the number of observations at each level is too small.
- Even though deviances may not have approx. chi-squared dist., the difference of deviances of two models is often approx. chi-squared.
  One can *safely* use the **diff. of deviances** to do **likelihood ratio test** for model comparison no matter the data are grouped or not grouped.

Summary for Deviance

For a Model M of interest

$$
\begin{aligned}
\text{Deviance} &= -2(L_M - L_S) \\
&= 2 \textstyle\sum_i \left\{ y_i \log\left(\frac{y_i}{n_i \widehat{\pi}(\mathbf{x}_i)}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i(1 - \widehat{\pi}(\mathbf{x}_i))}\right)\right\} \\
&= 2 \sum_i (\text{observed}) \log\left(\frac{\text{observed}}{\text{fitted}}\right) \\
&= G^2
\end{aligned}
$$

where

$$
\begin{aligned}
L_M &= \text{max. log-likelihood for Model } M \\
L_S &= \text{max. log-likelihood for the saturated model} \\
&= \text{the upper bound for max. log-likelihood of ANY model}
\end{aligned}
$$

Deviance can be used to do goodness-of-fit test.