Logistic regression

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Generalized linear models

- Generalize regular regression to non-normal data $\{(Y_i, \mathbf{x}_i)\}_{i=1}^n$, most often Bernoulli or Poisson Y_i.
- The general theory of GLMs has been developed to outcomes in the exponential family (normal, gamma, Poisson, binomial, negative binomial, ordinal/nominal multinomial).
- The ith *mean* is $\mu_i = E(Y_i)$
- The ith *linear predictor is* $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \mathbf{x}_i' \boldsymbol{\beta}$.
- A GLM relates the mean to the linear predictor through a *link function* $g(\cdot)$:

$$g(\mu_i) = \eta_i = \mathbf{x}_i' \boldsymbol{\beta}.$$

14.1, 14.2 Binary response regression

Let $Y_i \sim \text{Bern}(\pi_i)$. Y_i might indicate the presence/absence of a disease.

We wish to relate the probability of "success" π_i to explanatory covariates $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})$.

$$Y_i \sim \text{Bern}(\pi_i),$$

implying $E(Y_i) = \pi_i$ and $\text{var}(Y_i) = \pi_i(1 - \pi_i)$.

Identity link $g(\mu) = \mu$

The **identity link** gives $\pi_i = \boldsymbol{\beta}' \mathbf{x}_i$. When $\mathbf{x}_i = (1, x_i)'$, this reduces to

$$Y_i \sim \text{Bern}(\beta_0 + \beta_1 x_i).$$

- When x_i is large or small, π_i can be less than zero or greater than one.
- The identity like is appropriate for a restricted range of x_i values.
- It can of course be extended to $\pi_i = \boldsymbol{\beta}' \mathbf{x}_i$ where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ik})$.
- This model can be fit in SAS `proc genmod`.

Individual Bernoulli vs. aggregated binomial

Data can be stored in one of two ways:

- If each subject has their own individual binary outcome Y_i, we can write `model y=x1 x2` in `proc genmod` or `proc logistic`.
- If data are grouped, so that there are $Y_{\cdot j}$ successes out of n_j with covariate $\mathbf{x}_j$, $j = 1, \ldots, c$, then write `model y/n=x1 x2`. This method is sometimes used to reduce a very large number of individuals n to a small number of distinct covariates c; it is essentially a product binomial model.

Association between snoring and heart disease

From Agresti (2013).

Let s be someone's snoring score, $s \in \{0, 2, 4, 5\}$.

		Heart disease		Proportion
Snoring	s	yes	no	yes
Never	0	24	1355	0.017
Occasionally	2	35	603	0.055
Nearly every night	4	21	192	0.099
Every night	5	30	224	0.118

This is fit in `proc genmod`:

```
data glm;
 input snoring disease total @@;
 datalines;
 0 24 1379 2 35 638 4 21 213 5 30 254
 ;
proc genmod data=glm; model disease/total = snoring / dist=bin link=identity;
run;
```

Extracting useful inferences

The fitted model is

$$\hat{\pi}(s) = 0.0172 + 0.0198s.$$

For every unit increase in snoring score s, the probability of heart disease increases by about 2%.

The p-values test $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. The latter is more interesting and we reject at the $\alpha = 0.001$ level. The probability of heart disease is strongly, *linearly* related to the snoring score.

We'll denote the maximum likelihood estimates by $\hat{\boldsymbol{\beta}}$ instead of $\mathbf{b}$. Both PROC LOGISTIC and PROC GENMOD give MLEs.

14.2, 14.3 Logistic regression

Often a fixed change in x has less impact when $\pi(x)$ is near zero or one.

Example: Let $\pi(x)$ be probability of getting an A in a statistics class and x is the number of hours a week you work on homework. When $x = 0$, increasing x by 1 will change your (very small) probability of an A very little. When $x = 4$, adding an hour will change your probability quite a bit. When $x = 20$, that additional hour probably won't improve your chances of getting an A much. You were at essentially $\pi(x) \approx 1$ at $x = 10$.

Of course, this is a *mean* model. Individuals will vary.

Logit link and logistic regression

The most widely used nonlinear function to model probabilities is the logit link:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i.$$

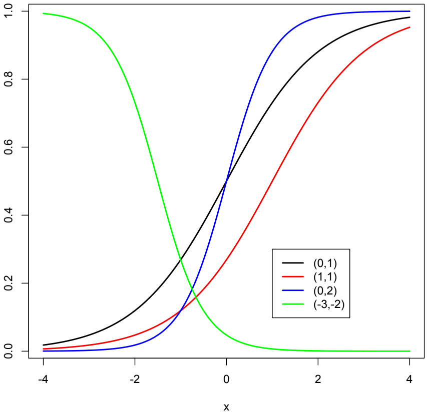
Solving for π_i and then dropping the subscripts we get the probability of success $(Y = 1)$ as a function of x:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} = [1 + \exp(-\beta_0 - \beta_1 x)]^{-1}.$$

When $\beta_1 > 0$ the function increases from 0 to 1; when $\beta_1 < 0$ it decreases. When $\beta = 0$ the function is constant for all values of x and Y is unrelated to x.

The logistic function is $\text{logit}^{-1}(x) = e^x/(1 + e^x)$.

Logistic functions for various β_0 and β_1



Logistic curves $\pi(x) = e^{\beta_0+\beta_1 x}/(1+e^{\beta_0+\beta_1 x})$ with $(\beta_0, \beta_1) = (0,1)$, $(1,1)$, $(0,2)$, $(-3,-2)$. What about $(\beta_0, \beta_1) = (\log 2, 0)$?

Logistic regression on snoring data

To fit the snoring data to the logistic regression model we use the same SAS code as before (`proc genmod`), except we specify LINK=LOGIT (or drop the LINK option, since LOGIT is the default) and obtain $b_0 = -3.87$ and $b_1 = 0.40$ as maximum likelihood estimates.

```
proc genmod data=glm;
*We dropped DIST=BIN too, though it’s better practice to include it;
model disease/total = snoring;
run;
```

Logistic output

You can also use `proc logistic` to fit binary regression models.

```
proc logistic data=glm;
model disease/total = snoring;
run;
```

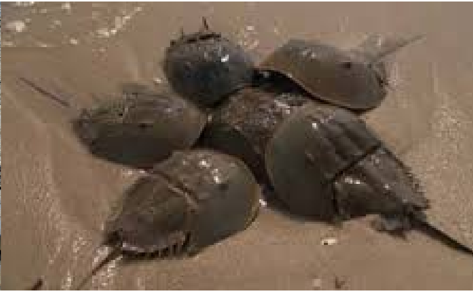
The fitted model is $\hat{\pi}(x) = \frac{\exp(-3.87+0.40x)}{1+\exp(-3.87+0.40x)}$. As before, we reject $H_0 : \beta_1 = 0$; there is a strong, positive association between snoring score and developing heart disease.

Horseshoe Crab facts

- Horseshoe crabs aren't that closely related to crabs.
- Their mass spawning events (e.g., at Delaware Bay in DE and NJ) attract thousands of shorebirds, including the threatened Red Knot
- These events also attract(ed) commercial fishermen (eel and conch fisheries), fertilizer companies (no longer), and the biomedical industry (unique antibacterial properties of their blue blood)
- Exploitation of horseshoe crabs has greatly affected migrating shorebirds as well (see Red Knot above).

Horseshoe Crab spawning

(a) Horseshoe Crab spawning event



(b) Female Horseshoe Crab with mate and satellite males

Horseshoe crabs

Horseshoe crabs arrive on the beach in pairs and spawn in the high intertidal during the springtime, new and full moon high tides. Unattached males also come to the beach, crowd around the nesting couples and compete with attached males for fertilizations. Satellite males form large groups around some couples while ignoring others, resulting in a nonrandom distribution that cannot be explained by local environmental conditions or habitat selection.

Based on the evidence from observations and experiments, the most likely explanation for the nonrandom distribution of satellite males among nesting pairs is that unattached males are preferentially attracted to some females over others. [1]

[1] Brockmann H.J. (1996) Satellite Male Groups in Horseshoe Crabs, Limulus polyphemus. Ethology.

Crab mating (Agresti, 2013)

Data on $n = 173$ female horseshoe crabs.

- C = color (1,2,3,4=light medium, medium, dark medium, dark).
- S = posterior spine condition (1,2,3=both good, one worn or broken, both worn or broken). Males attach to posterior spines when mating.
- W = carapace width (cm).
- Wt = weight (kg).
- Sa = number of satellites (additional male crabs besides her nest-mate husband) nearby. Satellite males fertilize some of the female's eggs.

We are initially interested in the probability that a female horseshoe crab has one or more satellites ($Y_i = 1$) as a function of carapace width.

Shorebirds Feeding

(a) Shore birds feeding on horseshoe crab eggs

(b) Red Knot with tag B95–the so-called Moon Bird–has migrated over a quarter-million miles since first tagged in 1995

Crab data in SAS

```
data crabs;
weight=weight/1000; color=color-1;
*Convert satellite to a binary variable rather than a count;
y=0; if satell>0 then y=1; id=_n_; run;
proc logistic data=crabs;
model y(event=’1’)=width / link=logit; run;
```

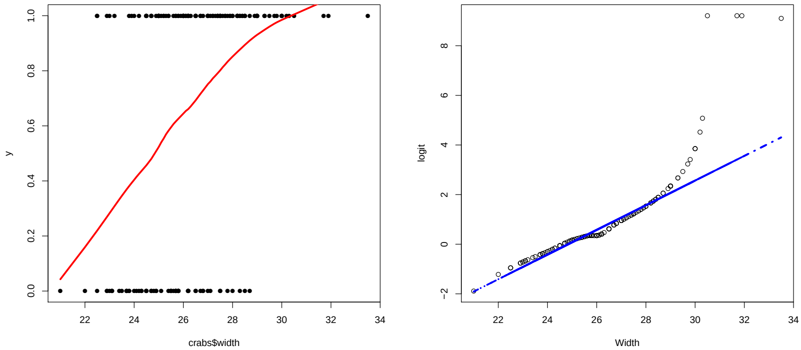
event=’1’ tells SAS to model $\pi_i = P(Y_i = 1)$ rather than $\pi_i = P(Y_i = 0)$. The default link is logit (giving logistic regression) – Specify it here anyway for transparency.

Crab data in R

```
url<-"http://people.stat.sc.edu/hoyen/Stat705/Data/crabs.txt"
crabs<-read.table(file=url, stringsAsFactors=F, header=T)
str(crabs)
table(crabs$color)
y<-ifelse(crabs$satell > 0, 1, 0)
mylogit<-glm( y ˜ width, data=crabs, family="binomial")
summary(mylogit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45

Number of Fisher Scoring iterations: 4
```


y
0.0
0.2
0.4
0.6
0.8
1.0
22
24
26
28
30
32
34
crabs$width
logit
−2
0
2
4
6
8
22
24
26
28
30
32
34
Width

14.3 Model interpretation

For simple logistic regression

$$Y_i \sim \text{Bern}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right).$$

An odds ratio: let's look at how the odds of success changes when we increase x by one unit:

$$\begin{aligned}\frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} &= \frac{\left[\frac{e^{\beta_0+\beta_1 x+\beta_1}}{1+e^{\beta_0+\beta_1 x+\beta_1}}\right] \Big/ \left[\frac{1}{1+e^{\beta_0+\beta_1 x+\beta_1}}\right]}{\left[\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}\right] \Big/ \left[\frac{1}{1+e^{\beta_0+\beta_1 x}}\right]} \\ &= \frac{e^{\beta_0+\beta_1 x+\beta_1}}{e^{\beta_0+\beta_1 x}} = e^{\beta_1}.\end{aligned}$$

When we increase x by one unit, the odds of an event occurring increases by a factor of e^{β_1}, *regardless of the value of x*.

Horseshoe crab data

Let's look at $Y_i = 1$ if a female crab has one or more satellites, and $Y_i = 0$ if not. So

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}},$$

is the probability of a female having more than her nest-mate around as a function of her width x.

From regression output we obtain a table with estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ as well as standard errors, χ^2 test stattistics, and p-values that $H_0 : \beta_0 = 0$ and $H_0 : \beta_1 = 0$. We also obtain an estimate of the odds ratio e^{b_1} and a 95% CI for e^{β_1}.

SAS output

```
                                   Standard          Wald
Parameter     DF     Estimate         Error    Chi-Square    Pr > ChiSq
Intercept      1     -12.3508        2.6287       22.0749        <.0001
width          1       0.4972        0.1017       23.8872        <.0001

                       Odds Ratio Estimates

                        Point           95% Wald
          Effect      Estimate      Confidence Limits
          width          1.644       1.347        2.007
```

We estimate the probability of a satellite as

$$\hat{\pi}(x) = \frac{e^{-12.35+0.50x}}{1+e^{-12.35+0.50x}}.$$

The odds of having a satellite increases by a factor between 1.3 and 2.0 times for every *cm* increase in carapace width.

The coefficient table houses estimates $\hat{\beta}_j$, $\text{se}(\hat{\beta}_j)$, and the Wald statistic $z_j^2 = \{\hat{\beta}_j/\text{se}(\hat{\beta}_j)\}^2$ and p-value for testing $H_0 : \beta_j = 0$. What do we conclude here?

14.4 Multiple predictors

Now we have $k = p - 1$ predictors $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})$ and fit

$$Y_i \sim \text{bin}\left(n_i, \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})} \right).$$

- Many of these predictors may be sets of dummy variables associated with categorical predictors.
- e^{β_j} is now termed the *adjusted* odds ratio. This is how the odds of the event occurring changes when x_j increases by one unit *keeping the remaining predictors constant*.
- This interpretation may not make sense if two predictors are highly related.

Crab data

There are two categorical predictors, C and S, and two continuous predictors W and Wt. Let $Y = 1$ if a randomly drawn crab has one or more satellites and $\mathbf{x} = (C, S, W, Wt)$ be her covariates. An *additive* model including all four covariates would look like

$$\begin{aligned} \text{logit } \pi(\mathbf{x}) &= \beta_0 + \beta_1 I\{C = 1\} + \beta_2 I\{C = 2\} + \beta_3 I\{C = 3\} \\ &\quad + \beta_4 I\{S = 1\} + \beta_5 I\{S = 2\} + \beta_6 W + \beta_7 Wt \end{aligned}$$

This model is fit via

```
proc logistic data=crabs descending;
 class color spine / param=ref;
 model y = color spine width weight / lackfit;
```

R Code

```
> fit2<-glm( y ~ color + spine + width + weight, data=crabs, family="binomial")
> summary(fit2)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.0650134  3.9285518  -2.053   0.0401 *
color2      -0.1029024  0.7825912  -0.131   0.8954
color3      -0.4888642  0.8531183  -0.573   0.5666
color4      -1.6086658  0.9355326  -1.720   0.0855 .
spine1      -0.0959809  0.7033698  -0.136   0.8915
spine2       0.4002868  0.5027043   0.796   0.4259
width        0.2631279  0.1952986   1.347   0.1779
weight       0.0008258  0.0007038   1.173   0.2407
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 185.20  on 165  degrees of freedom
AIC: 201.2
```

R:Analysis of Deviance Table

```
> anova(fit2, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                     172     225.76
color   3  13.6977       169     212.06  0.003347 **
spine   2   3.2270       167     208.83  0.199185
width   1  22.2219       166     186.61 2.429e-06 ***
weight  1   1.4099       165     185.20  0.235073
```

Likelihood Ratio Test

```
> fit3<-glm( y ~ color + width, data=crabs, family="binomial")
> summary(fit3)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.38519    2.87346  -3.962 7.43e-05 ***
color2        0.07242    0.73989   0.098    0.922
color3       -0.22380    0.77708  -0.288    0.773
color4       -1.32992    0.85252  -1.560    0.119
width         0.46796    0.10554   4.434 9.26e-06 ***
> anova(fit2, fit3, test="Chisq")
Analysis of Deviance Table

Model 1: y ~ color + spine + width + weight
Model 2: y ~ color + width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       165     185.20
2       168     187.46 -3   -2.255   0.5212
> AIC(fit2, fit3)
     df     AIC
fit2  8 201.202
fit3  5 197.457
```

Deviance

The **deviance** of a fitted model compared the log-likelihood of the fitted model to the log-likelihood of a model with n parameters that fits the n observation perfectly. It can be shown that the likelihood of this **saturated model** is equal to 1 (log-likelihood 0). Therefore, the deviance for the logistic regression model is

$$\text{Deviance} = -2\sum_{i=1}^{n}[Y_i \log(\widehat{\pi}_i) + (1 - Y_i)\log(1 - \widehat{\pi}_i)]$$

The **larger** the deviance, the **poorer** the fit.

R Code

```
> crabs$ncolor<-ifelse(crabs$color==4, 1, 0)
> fit4<-glm( y ~ ncolor + width, data=crabs, family="binomial")
> summary(fit4)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -11.6790     2.6925  -4.338 1.44e-05 ***
ncolor       -1.3005     0.5259  -2.473   0.0134 *
width         0.4782     0.1041   4.592 4.39e-06 ***
```

Hosmer-Lemeshow Goodness of Fit

```
> hl<-hoslem.test(fit4$y, fitted(fit4), g=10)
> cbind(hl$observed,hl$expected)
               y0 y1      yhat0     yhat1
[0.0504,0.327] 13  5 13.8636464  4.136354
(0.327,0.45]   11  7 11.0369104  6.963090
(0.45,0.539]    8  8  7.9984462  8.001554
(0.539,0.615]   7 12  8.0622291 10.937771
(0.615,0.68]   10 10  6.8118836 13.188116
(0.68,0.73]     5 11  4.6247054 11.375295
(0.73,0.801]    3 11  3.2266434 10.773357
(0.801,0.854]   2 15  2.9623188 14.037681
(0.854,0.899]   3 19  2.6660602 19.333940
(0.899,0.987]   0 13  0.7471565 12.252843

> h1

Hosmer and Lemeshow goodness of fit (GOF)
test

data:  fit4$y, fitted(fit4)
X-squared = 4.0228, df = 8, p-value = 0.8551
```

The H-L GOF test with p=0.85 suggests there's no evidence of gross lack of fit.

Comments

- The odds of having satellite(s) significantly increases by $e^{1.3005} \approx 3.67$ for medium vs. dark crabs.
- The odds of having satellite(s) significantly increases by a factor of $e^{0.4782} \approx 1.6$ for every *cm* increase in carapace width when fixing color.
- Lighter, wider crabs tend to have satellite(s) more often.
- The H-L GOF test shows no gross LOF.
- We didn't check for interactions. If an interaction between color and width existed, then the odds ratio of satellite(s) for different colored crabs would change with how wide she is.

AIC & model selection

"All models are wrong; some models are useful." – George Box*.

It is often of interest to examine several competing models. In light of underlying biology or science, one or more models may have relevant interpretations within the context of why data were collected in the first place.

In the absence of scientific input, a widely-used model selection tool is the Akaike information criterion (AIC),

$$\text{AIC} = -2[L(\hat{\beta}; \mathbf{y}) - p].$$

The term $L(\hat{\beta}; \mathbf{y})$ represents model fit. If you add a parameter to a model, $L(\hat{\beta}; \mathbf{y})$ has to increase. If we only used $L(\hat{\beta}; \mathbf{y})$ as a criterion, we'd keep adding predictors until we ran out. The p penalizes for the number of the predictors in the model.

Crab data

The AIC has very nice properties in large samples in terms of prediction. The smaller the AIC is, the better the model fit (asymptotically).

Model	AIC
W	198.8
$C + W$	197.5
$C + W + Wt + W * C + W * Wt$	196.8

The best model is the most complicated one, according to AIC. One might choose the slightly "worse" model $C + W$ for its enhanced interpretability.

Aggregating over distinct covariates

Binomial data is often recorded as individual (Bernoulli) records:

i	y_i	n_i	x_i
1	0	1	9
2	0	1	14
3	1	1	14
4	0	1	17
5	1	1	17
6	1	1	17
7	1	1	20

Grouping the data yields an identical model:

i	y_i	n_i	x_i
1	0	1	9
2	1	2	14
3	2	3	17
4	1	1	20

Comments

- There are $c = 66$ distinct widths $\{\mathbf{x}_i\}$ out of $n = 173$ crabs.
- The Hosmer and Lemeshow test gives a p-value of 0.73 based on $g = 10$ groups.
- GOF tests are meant to detect *gross* deviations from model assumptions.

14.8 Diagnostics

GOF tests are *global* checks for model adequacy. Residuals and influential measures can refine a model inadequacy diagnosis.

The data are $(\mathbf{x}_j, Y_{\cdot j})$ for $j = 1, \ldots, c$. The j^{th} *fitted value* is an estimate of $\mu_j = E(Y_{\cdot j})$, namely $\widehat{E(Y_{\cdot j})} = \hat{\mu}_j = n_j \hat{\pi}_j$ where $\pi_j = \frac{e^{\boldsymbol{\beta}' \mathbf{x}_j}}{1+e^{\boldsymbol{\beta}' \mathbf{x}_j}}$ and $\hat{\pi}_j = \frac{e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_j}}{1+e^{\hat{\boldsymbol{\beta}}' \mathbf{x}_j}}$. The raw residual e_j is what we see $(Y_{\cdot j})$ minus what we predict $(n_j \hat{\pi}_j)$. The Pearson residual divides this by an estimate of $\sqrt{\text{var}(Y_{\cdot j})}$:

$$r_{P_j} = \frac{y_{\cdot j} - n_j \hat{\pi}_j}{\sqrt{n_j \hat{\pi}_j (1 - \hat{\pi}_j)}}.$$

The Pearson GOF statistic is

$$X^2 = \sum_{j=1}^{c} r_{P_j}^2.$$

Diagnostics

The standardized Pearson residual is given by

$$r_{SP_j} = \frac{y_{.j} - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j(1-\hat{\pi}_j)(1-\hat{h}_j)}},$$

where $\hat{h}_j$ is the j^{th} diagonal element of the *hat* matrix $\hat{\mathbf{H}} = \hat{\mathbf{W}}^{1/2}\mathbf{X}(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{W}}^{1/2}$ where $\mathbf{X}$ is the design matrix

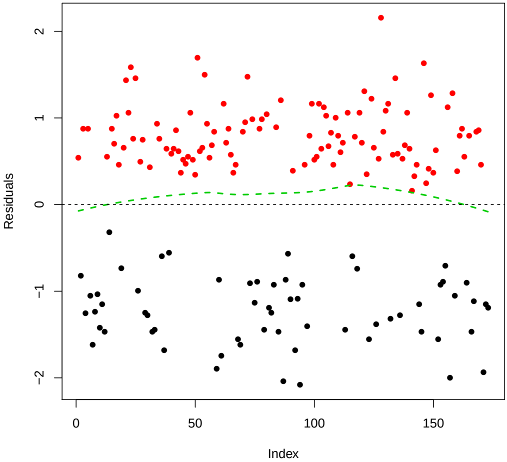
$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{c1} & \cdots & x_{c,p-1} \end{bmatrix},$$

and

$$\hat{\mathbf{W}} = \begin{bmatrix} n_1\hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & n_2\hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_c\hat{\pi}_c(1-\hat{\pi}_c) \end{bmatrix}.$$

Diagnostics

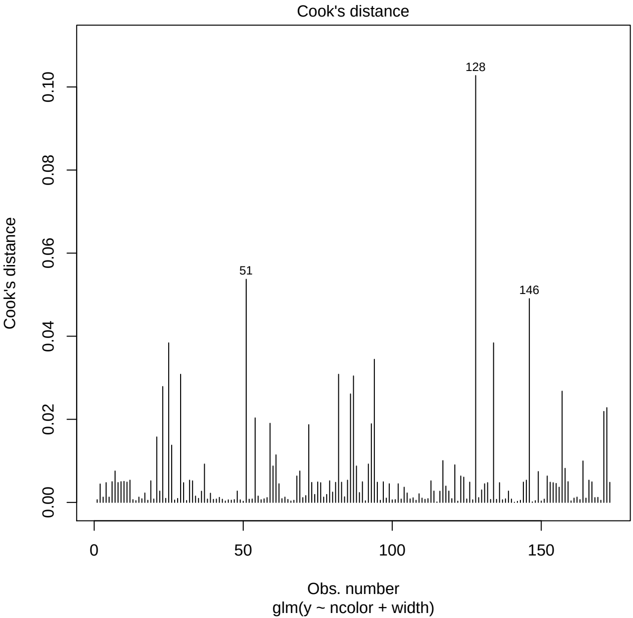
- Your book suggests lowess smooths of residual plots (pp. 594–595), based on the identity $E(Y_i - \hat{\pi}_i) = E(e_i) = 0$ for Bernoulli data. You are looking for a line that is *approximately* zero, not perfectly zero. The line will have a natural increase/decrease at either end if there are lots of zeros or ones – e.g. last two plots on p. 595.
- Residual plots for individual predictors might show curvature; adding quadratic terms or interactions can improve fit.
- An overall plot is a smoothed r_{SP_j} versus the linear predictor $\hat{\eta}_j = \hat{\boldsymbol{\beta}}'\mathbf{x}_j$. This plot will tell you if the model tends to over or underpredict the observed data for ranges of the linear predictor.
- You can look at individual r_{SP_j} to determine model fit. For the crab data, this might flag some individual crabs as ill-fit or unusual relative to the model. I usually flag $|r_{SP_j}| > 3$ as being ill-fit by the model.


Residuals
2
1
0
−1
−2
0
50
100
150
Index

Influential observations

Unlike linear regression, the leverage $\hat{h}_j$ in logistic regression depends on the model fit $\hat{\beta}$ as well as the covariates $\mathbf{x}_j$. Points that have extreme predictor values $\mathbf{x}_j$ may not have high leverage $\hat{h}_j$ if $\hat{\pi}_j$ is close to 0 or 1. Here are the influence diagnostics available in PROC LOGISTIC:

- Leverage $\hat{h}_j$. Still may be useful for detecting "extreme" predictor values $\mathbf{x}_j$.
- $c_j = r_{SP_j}^2 \hat{h}_j / [p(1 - \hat{h}_j)^2]$ measures the change in the joint confidence region for β when j is left out (Cook's distance).
- DFBETA$_{js}$ is the standardized change in $\hat{\beta}_s$ when observation j is left out.
- The change in the X^2 GOF statistic when obs. j is left out is DIFCHISQ$_j = r_{SP_j}^2/(1 - \hat{h}_j)$. ($\Delta X_j^2$ in your book)


Cook's distance
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
128
51
146
0
50
100
150
Obs. number
glm(y ~ ncolor + width)

Multicollinearity

```
>library(car)
> vif(fit4)
  ncolor    width
1.000031 1.000031
```

As a rule of thumb, a VIF value that exceeds 10 indicates a problematic amount of collinearity.

Connection to 2 × 2 Tables

CHD	Smoke	Coffee	n
yes	no	no	15
no	no	no	42
yes	yes	no	11
no	yes	no	8
yes	no	yes	15
no	no	yes	21
yes	yes	yes	25
no	yes	yes	14

n=151

CHD, Smoking and Coffee

- Case-control (disease=CHD)
- Is smoking and/or coffee related to an increased odds of CHD?
- Is the association of coffee with CHD higher among smoker? That is, is smoking an effect modifier of the coffee-CHD associations?

Pooled data (ignore smoking)

Coffee	CHD	Control
No	26	50
Yes	40	35
% Coffee	0.39	0.59

OR$= \frac{\frac{0.39}{1-0.39}}{\frac{0.59}{1-0.59}} = 2.2$

95%CI$=$ (1.15, 4.27)

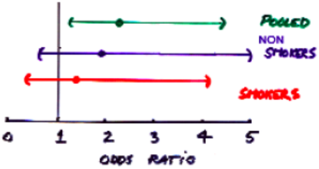
Stratify

Coffee	CHD	Control
No	15	42
Yes	15	21

Table: Non-Smoker, OR=2.06 (0.82, 4.9)

Coffee	CHD	Control
No	11	8
Yes	25	14

Table: Smoker, OR=1.29 (0.42, 4.0)



Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

- e^{β_1}: Odds ratio of having CHD for coffee drinkers versus non-drinkers among **non-smokers**

Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

- e^{β_1}: Odds ratio of having CHD for coffee drinkers versus non-drinkers among **non-smokers**
- $e^{\beta_1+\beta_3}$: Odds ratio of having CHD for coffee drinkers versus non-drinkers among **smokers**

Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

- e^{β_2}: Odds ratio of having CHD for smokers versus non-smoker among **non-coffee drinkers**

Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

- e^{β_2}: Odds ratio of having CHD for smokers versus non-smoker among **non-coffee drinkers**
- $e^{\beta_2+\beta_3}$: Odds ratio of having CHD for smokers versus non-smoker among **coffee drinkers**

Logistic Regression Model

- $Y_i = 1$ if CHD, 0 if control
- $p_i = Pr(Y_i = 1)$

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

- e^{β_2}: Odds ratio of having CHD for smokers versus non-smoker among **non-coffee drinkers**
- $e^{\beta_2+\beta_3}$: Odds ratio of having CHD for smokers versus non-smoker among **coffee drinkers**
- e^{β_3}: factor by which odds ratio of being a CHD case for coffee drinkers -vs- nondrinkers is multiplied for smokers as compared to non-smokers

Common Idea: Additional multiplicative change in the odds ratio beyond the smoking or coffee drinking effect alone when you have both of these risk factors present.

R Code

```
> n<-c(15,42, 11, 8, 15, 21, 25, 14)
> smoke<-rep(c("no", "yes"), 2)
> coffee<-rep(c("no", "yes"), each=2)
> chd<-matrix(n, ncol=2, byrow=T)
> dat<-data.frame(smoke, coffee, chd)
>dat
>dat
  smoke coffee chd1 chd0
1    no     no 15 42
2   yes     no 11  8
3    no    yes 15 21
4   yes    yes 25 14
> chd1<-as.numeric(dat[,3])
> chd0<-as.numeric(dat[,4])
```

Logistic Regression and 2 × 2 Table

```
> fit1<-glm(cbind(chd1, chd0) ~ coffee, data=dat, family="binomial")
> summary(fit1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.6539     0.2418  -2.705  0.00684 **
coffeeyes     0.7875     0.3347   2.353  0.01864 *
> exp(confint(fit1))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.3191732 0.8272657
coffeeyes   1.1469951 4.2732289
```

$e^{0.7875} = 2.2$ (95%CI: 1.15, 4.27)

Coffee	CHD	Control
No	26	50
Yes	40	35
% Coffee	0.39	0.59

OR$= \frac{\frac{0.39}{1-0.39}}{\frac{0.59}{1-0.59}} = 2.2$

95%CI$=$ (1.15, 4.27)

Main effects of coffee and smoke

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i$$

```
> fit2<-glm(cbind(chd1, chd0) ˜ coffee + smoke, data=dat, family="binomial")
> summary(fit2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9572     0.2703  -3.541 0.000398 ***
coffeeyes     0.5270     0.3542   1.488 0.136798
smokeyes      1.1020     0.3610   3.053 0.002269 **
---
```

Is smoking a confounding variable?

	OR	2.5% CI	97.5% CI
Model 1			
Intercept	0.52	0.32	0.83
coffeeyes	2.20	1.15	4.27
Model 2			
Intercept	0.38	0.22	0.64
coffeeyes	1.69	0.84	3.40
smokeyes	3.01	1.49	6.18

Smoking does not confound the relationship between coffee drinking and CHD.

- since 1.7 is in the 95% CI from the model without smoking.

Is smoking an effect modifier of CHD-coffee association?

```
> fit3<-glm(cbind(chd1, chd0) ~ coffee + smoke + coffee*smoke, data=dat, family
> summary(fit3)
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.0296     0.3008  -3.423 0.000619 ***
coffeeyes            0.6931     0.4525   1.532 0.125573
smokeyes             1.3481     0.5535   2.435 0.014873 *
coffeeyes:smokeyes  -0.4318     0.7295  -0.592 0.553899
---
> anova(fit2, fit3)
Analysis of Deviance Table

Model 1: cbind(chd1, chd0) ~ coffee + smoke
Model 2: cbind(chd1, chd0) ~ coffee + smoke + coffee * smoke
  Resid. Df Resid. Dev Df Deviance
1         1     0.3511
2         0     0.0000  1   0.3511
```

And we conclude there is little evidence that smoking is an effect modifier!

Multiple predictors

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i$$

Smoking	Coffee:No	Yes
No	β_0	$\beta_0 + \beta_1$
Yes	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$

Table: Log-odds for CHD