BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part V

Logistic Regression

Objectives

In this part, you will learn the following models

- Simple logistic regression with one continuous x (Ungrouped data)
- Logistic regression for contingency table (Grouped data)
- Multiple logistic regression (Mixed types of covariates)

Key

- Logistic model and its advantage on modeling binary response
- Estimation, inference, and interpretation for slope parameter
- Estimation and inference for the predicted probability
- Understand interaction in logistic regression
- Summarizing effects in logistic regression

Reading: Agresti (2002), Section 4.1 - 4.5, Section 5.1 and 5.2

Introduction

Logistic regression is the most popular regression techniques for modeling dichotomous (binary) dependent variables.

For each observation, the dependent variable is simply either equal to 1 (the event) or 0 (the non-event), e.g. in epi studies,

- Dead or alive
- Diseased or non-diseased
- Exposed or unexposed
- Incident case or control

Measures of occurrence of an event

Pr(Event) = Pr(Y=1) = π

Odds (Event) = Pr(Event)/(1 − Pr(Event)), briefly just use odds.

That is , odds $= \pi/(1-\pi) \Leftrightarrow \pi =$ odds/(odds + 1)

- If π is very small, the odds and π are similar.
- If $0 < \pi < 0.5 \Leftrightarrow 0 <$ odds < 1
- If $0.5 < \pi < 1 \Leftrightarrow 1 <$ odds $< \infty$

Questions: How to evaluate covariates' effect on π or odds?

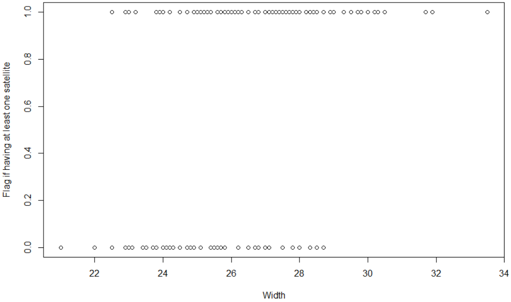
Crab Data

Previously, we have analyzed this data using Poisson model to study the association of number of satellites with the width of female crabs.

Now we dichotomize the count into a flag that takes two values:

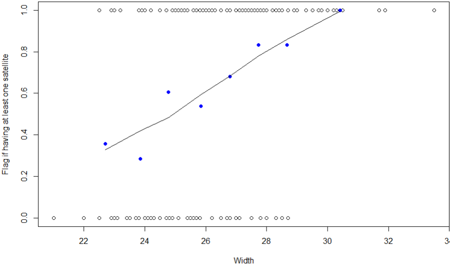
- 1, if has at least 1 satellite
- 0, if there is no satellite

The new goal is the study the effect of width on the probability of having at least 1 satellite.



Crab Data - Estimating Probability

Width	Number of Cases	Number Having Satellites	Sample Proportion
<23.25	14	5	0.36
23.25–24.25	14	4	0.29
24.25–25.25	28	17	0.61
25.25–26.25	39	21	0.54
26.25–27.25	22	15	0.68
27.25–28.25	24	20	0.83
28.25–29.25	18	15	0.83
>29.25	14	14	1.00



Simple Logistic Regression

Response variable $Y = 0$ or 1.

One explanatory variable X.

Assume Y has a binary distribution with parameter $\pi = P(Y = 1)$, which depends on the value of x. In this case, denote it as $\pi(x)$.

Logistic regression model:

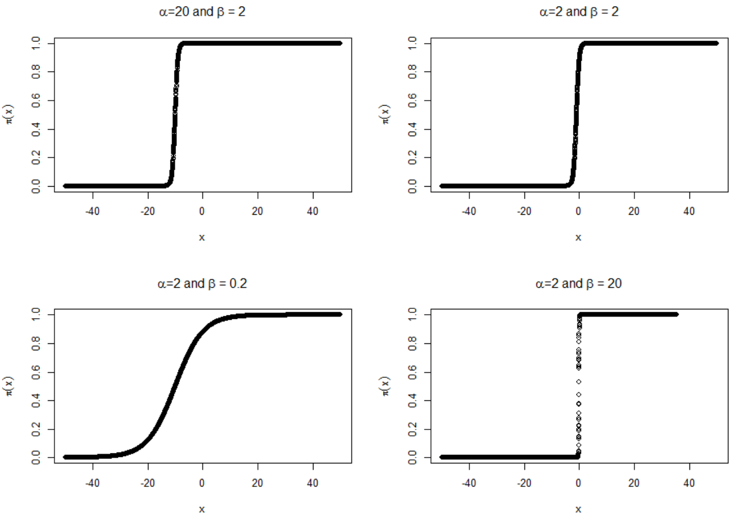
$$\text{logit}[\pi(x)] = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

That is, the logit of "success" probability has a linear form in x.

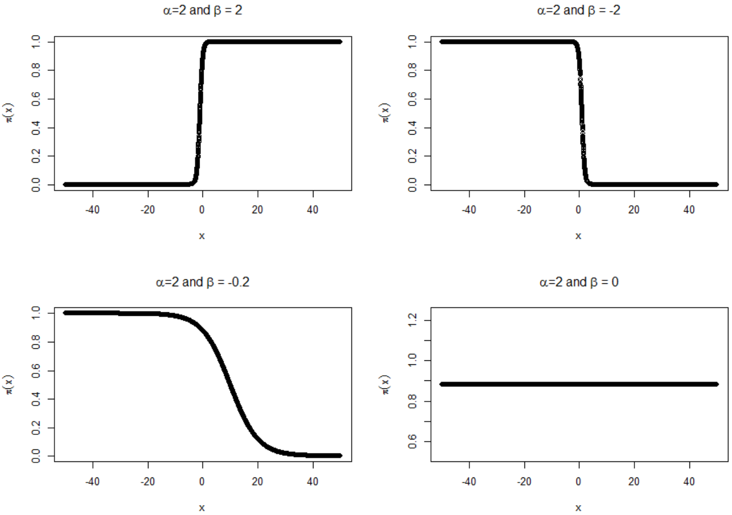
Equivalently,

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{-(\alpha+\beta x)}}$$

Plot of $\pi(x)$ Function



Plot of $\pi(x)$ Function



Properties of $\pi(x)$ function

The curve for $\pi(x)$ is S-shaped , nonlinear rather than linear in x.

$\pi(x) = \frac{1}{1+e^{-(\alpha+\beta x)}}$ is monotone in x.

If $\beta > 0$, then $\pi(x)$ increases as x increases.
If $\beta < 0$, then $\pi(x)$ decreases as x increases.

If $\beta = 0$, then $\pi(x) = 1/(1 + e^{-\alpha})$ is constant in x.

If $\pi(x) = 0.5$, then $x = -\alpha/\beta$ for simple logistic regression. This x is called median effect level and denoted as EL_{50}

Odds Ratio Interpretation

Recall that we have

$$\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$$

Therefore,

$$\text{odds} = \frac{\pi}{1-\pi} = \begin{cases} e^{\alpha+\beta x} & \text{at } x \\ e^{\alpha+\beta(x+1)} = e^{\beta} e^{\alpha+\beta x} & \text{at } (x+1) \end{cases}$$

$$\Rightarrow \frac{\text{odds at } (x+1)}{\text{odds at } x} = e^{\beta}$$

More generally,

$$\Rightarrow \frac{\text{odds at } (x+\Delta x)}{\text{odds at } x} = e^{\beta \Delta x}$$

- β is the change in the log odds when x increases in 1-unit.
- e^{β} is the multiplicative effect on odds when x increases by 1-unit, which is also the odds ratio at $x+1$ versus at x.

Linear Approximation Interpretation

Most of us do not think naturally on a logit (logarithm of the odds) scale, so we need to consider alternative interpretations.

Curve can be approximated by a straight line describing rate of change in $\pi(x)$ at a fixed value of x. Slope is $\beta\pi(x)(1-\pi(x))$.

- at x with $\pi(x) = 0.5$, slope is 0.25β
- at x with $\pi(x) = 0.1$ or 0.9, slope is 0.09β
- Steepest slope at the median effect level where $\pi(x) = 0.5$

Interpretation:

- The incremental rate of change in the fitted probability at x is $\hat{\beta}\hat{\pi}(x)(1-\hat{\pi}(x))$.
- (For positive $\hat{\beta}$) The estimated probability increases at the rate of $\hat{\beta}\hat{\pi}(x)(1-\hat{\pi}(x))$ per 1-unit increase in x.

Programs

SAS

```
Proc logistic data=crab descending;
     model flag = width;
run;

Proc genmod data=crab descending;
     model flag = width / dist=binomial;
run;
```

- If the data are coded 1 for disease and 0 for non-disease then the `descending` option is required to force SAS to estimate $\Pr[Y = 1|x]$ rather than the default of $\Pr[Y = 0|x]$
- Both methods are based on maximum likelihood ratio estimation.

R

```
fit = glm(flag ~ Width, family=binomial, data=crab)
summary(fit)
```

Why Ordinary Least Square (OLS) Estimation is not recommended?

Another possible method of fitting the model involves a transformation on the p value to its logit. From the model $\log(p/(1-p)) = \alpha + \beta x$, the logit is a linear function in the coefficients (α, β) and therefore it is possible to apply least square estimation. Although the transformation succeeds in linearizing the response function, two other problems remain.

- Since the true probability π is unknown, the values of the response logit are also unknown. Therefore, we must obtain estimate of the logit for each combination of the independent variables. This means we must have replicates for each of such combination, which is often rare in practical settings.
- Another issue is the concern of unequal variances. The variance of the response relates to $p(1-p)$ which is further depend on the x. That is, the regression errors are heterogeneous.

Crab Data - Model Fit

$$Y = \begin{cases} 1 & \text{if at least one satellite} \\ 0 & \text{if no satellite} \end{cases}$$

$X = \text{Width}$

Estimated model:

$$\text{logit}(\hat{\pi}(x)) = -12.35 + 0.497x$$

Or,

$$\hat{\pi}(x) = \frac{\exp(-12.35 + 0.497x)}{1 + \exp(-12.35 + 0.497x)}$$

- $\hat{\beta} > 0$, so $\hat{\pi} \uparrow$ as $x \uparrow$
- The median effect level $EL_{50} = -\frac{\hat{\alpha}}{\hat{\beta}} = 12.35/0.497 = 24.85$. This is where the function of $\pi(x)$ has the steepest slope.

Crab Data - Model Fit (cont.)

- At $x = 26$, the estimated probability is

$$\hat{\pi}(26) = \frac{\exp(-12.35 + 0.497 \times 26)}{1 + \exp(-12.35 + 0.497 \times 26)} = 0.64$$

- At $x = 26$, the estimated slope of the $\pi(x)$ is

$$\hat{\beta}\hat{\pi}(26)(1 - \hat{\pi}(26)) = 0.497 \times 0.64 \times (1 - 0.64) = 0.115$$

  For female with crabs with width 26 cm, the estimated probability of having at least one satellite increases at the rate of 0.11 per 1 cm increase in width.

- $e^{\hat{\beta}} = e^{0.497} = 1.64$
  If the width increases by 1 cm, the estimated odds increase by a factor 1.64. Or the multiplicative effect on odds associated with 1 cm increase in width is 1.64.

Inference for Simple Logistic Regression

MLE $\hat{\beta}$ is approximately normal for large samples.

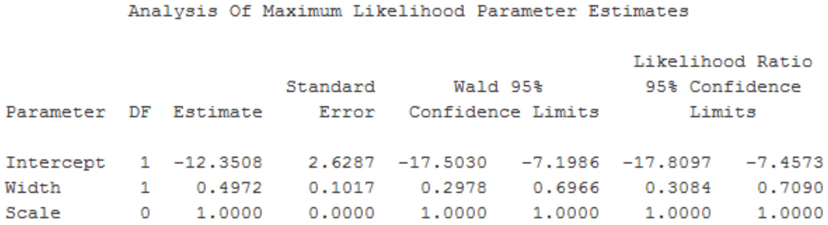
Confidence interval: Wald $(1-\alpha)$ CI for β is $\hat{\beta} \pm z_{\alpha/2}\mathrm{SE}(\hat{\beta})$.

[Crab data]:

- 95% CI for β: $0.4972 \pm 1.96 \times 0.1017 = (0.2979, 0.6965)$
- 95% CI for e^{β}: $(e^{0.2979}, e^{0.6965}) = (1.347, 2.007)$
  Odds estimate increases by at least 1.347 times, at most 2.007 times when the width increases 1 cm.
- 95% CI for $e^{3\beta}$: $(e^{3\times 0.2979}, e^{3\times 0.6965}) = (2.444, 8.082)$
  Odds estimate increases by at least 2.444 times, at most 8.082 times when the width increases 3 cm.

When the sample size is small, it is safer to use Likelihood-ratio (LR) CI.

- In SAS procedure PROC GENMOD, can use *lrci* and *waldci* to generate Likelihood-ratio CI and Wald CI (output shown below).
- In SAS procedure PROC LOGISTIC, can use $clparm = PL$ and $clparm = WALD$ to generate Likelihood-ratio CI and Wald CI.



```
                   Analysis Of Maximum Likelihood Parameter Estimates

                                                                     Likelihood Ratio
                                  Standard       Wald 95%             95% Confidence
Parameter   DF   Estimate            Error   Confidence Limits            Limits

Intercept    1   -12.3508           2.6287  -17.5030     -7.1986   -17.8097     -7.4573
Width        1     0.4972           0.1017    0.2978      0.6966     0.3084      0.7090
Scale        0     1.0000           0.0000    1.0000      1.0000     1.0000      1.0000
```

95% LR CI for e^{β}: $(e^{0.3084}, e^{0.7090}) = (1.361, 2.032)$.

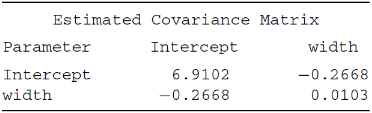
Confidence interval for π.

Recall that $\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$, we can first get the confidence interval for $\alpha + \beta x$ then obtain the confidence interval of $\pi(x)$ by the above relationship between $\alpha + \beta x$ and $\pi(x)$.

The estimated logit, i.e., estimate of $\alpha + \beta x = \hat{\alpha} + \hat{\beta} x$.

Variance of $\hat{\alpha} + \hat{\beta} x$ is $Var(\hat{\alpha} + \hat{\beta} x) = Var(\hat{\alpha}) + x^2 Var(\hat{\beta}) + 2x Cov(\hat{\alpha}, \hat{\beta})$

Estimates of $Var(\hat{\alpha})$, $Var(\hat{\beta})$, and $Cov(\hat{\alpha}, \hat{\beta})$ can be read from the covariance matrix, which can be requested in SAS by *covout* option. In R, you can use *summary.glm(model)$cov.unscaled* or *vcov(model)*.



Estimated Covariance Matrix

Parameter	Intercept	width
Intercept	6.9102	−0.2668
width	−0.2668	0.0103

- $\widehat{Var}(\hat{\alpha}) = 6.9102$
- $\widehat{Var}(\hat{\beta}) = 0.0103$
- $\widehat{Cov}(\hat{\alpha}, \hat{\beta}) = -.2668$

[Crab data] The model fit is $\text{logit}(\hat{\pi}(x)) = -12.35 + 0.497x$. Obtain the confidence interval for $\pi(26.5)$.

The estimated logit is $-12.35 + 0.497 \times 26.5 = 0.825$.

The estimated variance of the estimated logit is
$6.9102 + (26.5)^2 \times 0.0103 + 2 \times 26.5 \times (-0.2668) = 0.038$

Therefore, the confidence interval for the estimated logit is
$0.825 \pm 1.96 \times \sqrt{0.038} = (0.44, 1.21)$.

Accordingly, the confidence interval for the $\pi(26.5)$ is

$$\left(\frac{e^{0.44}}{1 + e^{0.44}}, \frac{e^{1.21}}{1 + e^{1.21}} \right) = (0.61, 0.77)$$

Significance testing:

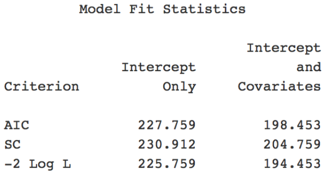
$H_0 : \beta = 0$ (Y is independent of X or $\pi(x)$ is constant in X)
$H_1 : \beta \neq 0$

Wald Test: $z = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$. Under H_0, $z \sim N(0,1)$ or $z^2 \sim \chi^2_{df=1}$.

Likelihood Ratio Test: Under H_0, $\beta = 0$ and denote the log-likelihood under H_0 as L_0. When $\beta = \hat{\beta}$, we have the alternative log-likelihood under H_1. The test statistic is $LR = -2(L_0 - L_1)$. Under H_0, $LR \sim \chi^2_{df=1}$.

[Crab data]:

- $z = \frac{0.4972}{0.1017} = 4.89$ and $z^2 = 23.91$.
  The p-value $= 2P(Z > 4.89) = P(\chi^2_{df=1} > 23.91) = 1.00836 \times 10^{-6}$
- $LR = 225.759 - 194.453 = 31.31$ and p-value $= P(\chi^2_{df=1} > 31.31)$.



```
                 Model Fit Statistics

                                     Intercept
                        Intercept          and
Criterion                    Only   Covariates

AIC                       227.759      198.453
SC                        230.912      204.759
-2 Log L                  225.759      194.453
```

Logistic Regression for Contingency Table

Logistic regression model can be developed for 2-way, 3-way, and muli-way tables to study the effect of the explanatory variables if one of the factors (dimension) can be treated as a dependent variable (response) and it has only **2 levels**. For example,

For a R×C two-way table, if the C (column) is a response variable with 2 levels and R (row) is an explanatory variable, a logistic regression model: $\text{logit}(\pi) = \alpha + \beta$ R can be built.

For a R×C×D table, if the C is a response variable with 2 levels, and R and D are explanatory variables, a logistic regression model: $\text{logit}(\pi) = \alpha + \beta_1 \text{R} + \beta_2 \text{D}$ can be built.

If R and D are categorical variables, dummy variables should be used for them. This can be specify in the *Class statement* in SAS and *factor* in R.

Logistic Regression for 2 × 2 Table

[Example] Kidney stone treatment

Stone	Treatment	Response Y	
Z	X	Success	Failure
Small	A	81	6
	B	234	36
Large	A	192	71
	B	55	25

Marginal Tables

Treatment	Response Y	
X	Success	Failure
A	273	77
B	289	61

Suppose here we would like to study the relationship between Treatment and Response. That is, we ignore the information of Stone size.

$$Y = \begin{cases} 1 & \text{if Success} \\ 0 & \text{if Failure} \end{cases} \qquad X = \text{Treatment} = \begin{cases} 1 & \text{Trt A} \\ 0 & \text{Trt B} \end{cases}$$

Model:

$$\text{logit}(\pi(x)) = \alpha + \beta X$$

.

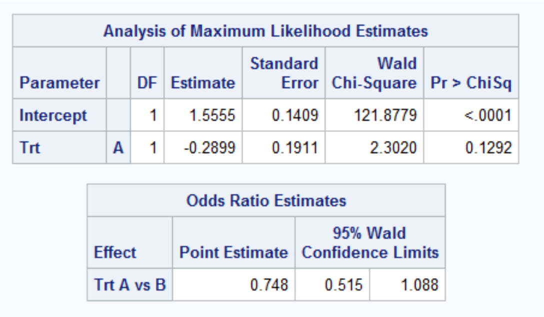
This means, Response has no effect on the Stone Size once Treatment information has been taken into account. That is, the Response and Stone Size are conditionally independent given Treatment.

Logistic model $\text{logit}(\pi(x)) = \alpha + \beta X$ implies:

$$\text{odds} = \frac{\pi}{1-\pi} = \begin{cases} e^{\alpha+\beta} & \text{if Trt is A} \\ e^{\alpha} & \text{if Trt is B} \end{cases}$$

Therefore,

- $\alpha =$ log odds of success for Trt B
- $\beta =$ increment in log odds for Trt A



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.5555	0.1409	121.8779	<.0001
Trt	A	1	-0.2899	0.1911	2.3020	0.1292

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Trt A vs B	0.748	0.515	1.088

Model fit: $\text{logit}(\hat{\pi}) = 1.56 - 0.29X$

$$\text{Estimated odds of heard attach} = \begin{cases} e^{1.56-0.29} = 3.56 & \text{if Trt A} \\ e^{1.56} = 4.76 & \text{if Trt B} \end{cases}$$

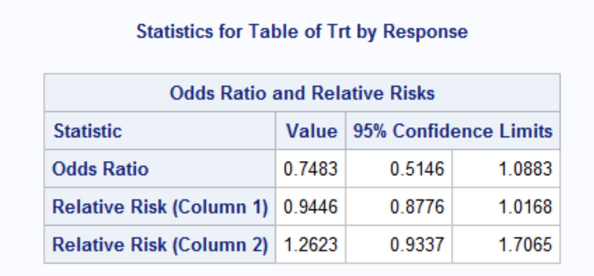
Estimated odds ratio $= e^{\hat{\beta}} = e^{-0.29} = 0.748$

Estimated odds of success in Trt A is 0.748 times of odds in Trt B.

95 % CI for β: $-0.29 \pm 1.96 \times 0.19 = (-0.6624, 0.0824)$

95 % CI for e^{β}: $(e^{-0.6624}, e^{0.0824}) = (0.52, 1.09)$.

Analyze this marginal table using PROC FREQ in SAS with *relrisk* option.



Statistics for Table of Trt by Response

Odds Ratio and Relative Risks			
Statistic	**Value**	**95% Confidence Limits**	
Odds Ratio	0.7483	0.5146	1.0883
Relative Risk (Column 1)	0.9446	0.8776	1.0168
Relative Risk (Column 2)	1.2623	0.9337	1.7065

Here, the estimate is the same as we obtain from logistic model. Thus, empirically, we find that analyzing a 2×2 table for relatedness is equivalent to logistic regression with a dummy variable.

In general, analyzing a $I \times 2$ table for relatedness is equivalent to logistic regression with $(I - 1)$ dummy variables.

Case-control Studies

$X =$ smoked at least one cigarette per day for at least a year

$Y =$ lung cancer indicator

	Lung Cancer	
Smoked	Yes	No
Yes	688	650
No	21	59
Total	709	709

Case-control studies are retrospective sampling designs, we can estimate $P(X = i|Y = j)$ but not $P(Y = j|X = i)$. Odds ratio can be used to measure the strength of association.

For a case-control study, we can still fit a logistic regression. The β have the same meaning as that in a prospective study, but the α is not meaningful anymore.

Logistic Regression for $2 \times 2 \times 2$ Table (Main effect model)

[Example] Kidney stone treatment

Stone	Treatment	Response Y	
Z	X	Success	Failure
Small	A	81	6
	B	234	36
Large	A	192	71
	B	55	25

Suppose

$$Y = \begin{cases} 1 & \text{if Success} \\ 0 & \text{if Failure} \end{cases}$$

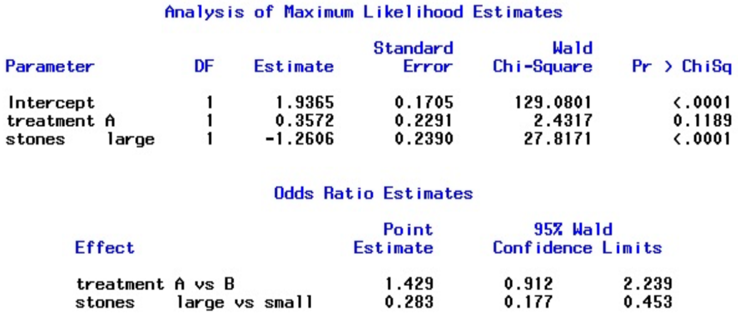
$$X_1 = \text{Treatment} = \begin{cases} 1 & \text{Trt A} \\ 0 & \text{Trt B} \end{cases} \qquad X_2 = \text{Stone size} = \begin{cases} 1 & \text{Large} \\ 0 & \text{Small} \end{cases}$$

Logistic model $\text{logit}(\pi(x)) = \alpha + \beta_1 X_1 + \beta_2 X_2$ implies:

$$\text{log odds} = \log(\frac{\pi}{1-\pi}) = \begin{cases} \alpha & \text{treat small stone with Trt B} \\ \alpha + \beta_1 & \text{treat small stone with Trt A} \\ \alpha + \beta_2 & \text{treat large stone with Trt B} \\ \alpha + \beta_1 + \beta_2 & \text{treat large stone with Trt A} \end{cases}$$

Hence,

- α: log odds of success for treating small stone with Trt B.
- β_1: increment to log odds for using Trt A.
- β_2: increment to log odds for treating large stone.



```
                    Analysis of Maximum Likelihood Estimates

                                               Standard          Wald
Parameter            DF      Estimate             Error    Chi-Square    Pr > ChiSq

Intercept             1        1.9365            0.1705      129.0801        <.0001
treatment A           1        0.3572            0.2291        2.4317        0.1189
stones      large     1       -1.2606            0.2390       27.8171        <.0001

                            Odds Ratio Estimates

                                          Point          95% Wald
       Effect                          Estimate      Confidence Limits

       treatment A vs B                   1.429       0.912       2.239
       stones     large vs small          0.283       0.177       0.453
```

Model fit: $\text{logit}(\hat{\pi}) = 1.94 + 0.36X_1 - 1.26X_2$

Controlling for stone size, estimated odds of success for using Trt A is $e^{0.36} = 1.43$ times of the estimated odds for Trt B.

95 % CI for β_1: $0.36 \pm 1.96(0.23) = (-0.09, 0.81)$

95 % CI for e^{β_1}: $(e^{-0.09}, e^{0.81}) = (0.91, 2.24)$.

Closer look at the model $\text{logit}(\pi(x)) = \alpha + \beta_1 X_1 + \beta_2 X_2$.

- No interaction means that
  - The relationship of Y and X_1 is the same at each level of X_2 ($e^{0.36} = 1.43$)
  - The relationship of Y and X_2 is the same at each level of X_1 ($e^{-1.26} = 0.28$)

  That is, the homogeneous association that the same odds ratio at each level of other variable.
- If $\beta_1 = 0$, then Y is conditionally independent of X_1 given X_2. To test for $H_0 : \beta = 0$ vs $H_1 : \beta_1 \neq 0$,

  $$z = \frac{\hat{\beta}_1}{\text{SE}(\hat{\beta}_1)} = \frac{0.36}{0.23} = 1.56, p\text{-}value = 2 \times P(Z > 1.56) = 0.12$$

  The p-value > 0.05. That is, controlling for stone size, the odds of success are similar for Trts A and B.

- Do we need stone size in the model?

  $H_0 : \beta_2 = 0$ (given trt, Y is independent of stone size)
  $H_1 : \beta_2 \neq 0$

$$z = \frac{\hat{\beta}_2}{\mathrm{SE}(\hat{\beta}_2)} = \frac{-1.26}{0.24} = -5.27, \textit{p-value} = 2 \times P(Z > 5.27)$$

Or $z^2 = (-5.27)^2 = 27.8, \textit{p-value} = P(\chi^2_{df=1} > 27.8)$
Both p-value $= 1.34 \times 10^{-7} < 0.05$. That is, there is evidence that controlling for treatment, success less likely for large stones than small stones.

Logistic Regression for $2 \times 2 \times 2$ Table (Interaction model)

Previous model with only main effects assumes homogenous association. Let's check this assumption by including the interaction term.

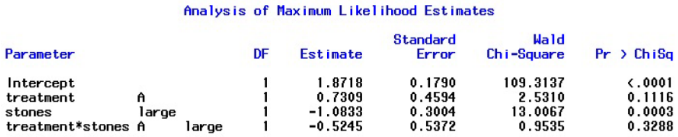
Model:

$$\text{logit}(\pi(x)) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

This model is a saturated model (that is, the most complex model, which provides perfect fit to the data).

$$X_1 = \text{Treatment} = \begin{cases} 1 & \text{Trt A} \\ 0 & \text{Trt B} \end{cases} \qquad X_2 = \text{Stone size} = \begin{cases} 1 & \text{Large} \\ 0 & \text{Small} \end{cases}$$

- α: log odds of success for treating small stone with Trt B.
- $\alpha + \beta_1$: log odds of success for treating small stone with Trt A.
- $\alpha + \beta_2$: log odds of success for treating large stone with Trt B.
- $\alpha + \beta_1 + \beta_2 + \beta_3$: log odds of success for treating large stone with Trt A



```
                   Analysis of Maximum Likelihood Estimates

                                             Standard        Wald
Parameter                  DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept                   1      1.8718      0.1790      109.3137        <.0001
treatment         A         1      0.7309      0.4594        2.5310        0.1116
stones            large     1     -1.0833      0.3004       13.0067        0.0003
treatment*stones  A   large 1     -0.5245      0.5372        0.9535        0.3288
```

Model fit: $\text{logit}(\hat{\pi}) = 1.87 + 0.73X_1 - 1.08X_2 - 0.52X_1X_2$

For small stones, estimated odds of success for treatment A is $e^{0.7309} = 2.0769$ times of estimated odds for treatment B.

For large stones, estimated odds of success for treatment A is $e^{0.7309-0.5245} = 1.2292$ times estimated odds for treatment B.

Under the interaction model, the odds ratios for treatment A versus Treatment B are different under different stone sizes.

Do we need interaction in the model?

$H_0 : \beta_3 = 0$ v.s. $H_1 : \beta_3 \neq 0$

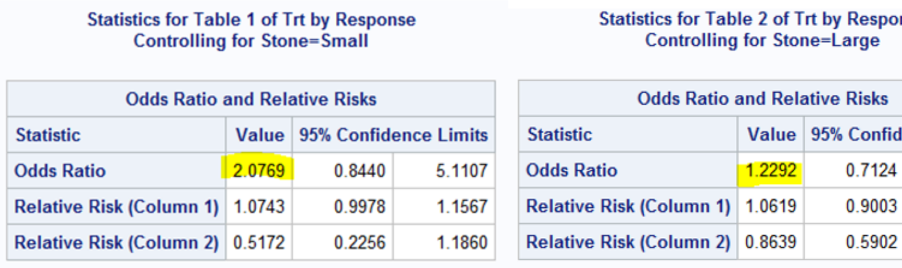
$$z = \frac{\hat{\beta}_3}{\mathrm{SE}(\hat{\beta}_3)} = \frac{-0.52}{0.54} = -0.98, p\text{-}value = 2 \times P(Z > 0.98)$$

Or $z^2 = (-0.98)^2 = 0.96, p\text{-}value = P(\chi^2_{df=1} > 0.96)$

Both p-value $= 0.33 > 0.05$. That is, no strong evidence to reject the assumption of homogenous association.

Recall, Breslow-Day Test for testing homogenous association for $2 \times 2 \times K$ table.

Analyze this partial table using PROC FREQ in SAS with *relrisk* option.



Statistics for Table 1 of Trt by Response
Controlling for Stone=Small

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	2.0769	0.8440	5.1107
Relative Risk (Column 1)	1.0743	0.9978	1.1567
Relative Risk (Column 2)	0.5172	0.2256	1.1860

Statistics for Table 2 of Trt by Respo
Controlling for Stone=Large

Odds Ratio and Relative Risks		
Statistic	Value	95% Confid
Odds Ratio	1.2292	0.7124
Relative Risk (Column 1)	1.0619	0.9003
Relative Risk (Column 2)	0.8639	0.5902

Here, the estimate from controlling stone size, is the same as we obtain from logistic model. So, there is an equivalence in fitting a contingency $2 \times 2 \times K$ table($X \times Y \times Z$) with a saturated logistic model of Y with 1 dummy variable for X, $K-1$ dummy variables for Z, and their interaction terms.

Further Comments on Logistic Regression for $2 \times 2 \times K$ Table

A typical example is multi-center clinical trials.

$\pi = P(Y = \text{success})$.

$$X = \text{Treatment} = \begin{cases} 1 & \text{Trt A} \\ 0 & \text{Trt B} \end{cases}$$

Center $= 1, 2, ..., K$.

Model:

$$\text{logit} = \alpha + \beta X + \beta_1 C_1 + \beta_2 C_2 + \ldots + \beta_{K-1} C_{K-1},$$

where $C_1, C_2, \ldots, C_{K-1}$ are dummy/indicator variables for center. Here the K-th center is the baseline. For all other centers, $C_i = 1$ if the study is conducted in center i.

- β_i is the effect for center i relative to the K-th center.
- No center effect means $\beta_1 = \beta_2 = \ldots = \beta_{K-1} = 0$.

- $\beta = 0$ means no treatment effect.
    - Note: this model assumes homogeneous association. Because the odds ratio e^{β} is the same for each center. That is, no treatment by center interaction.
    - e^{β} can be understood as the common XY odds ratio for each of the K partial tables.
    - In previous lecture, we use Cochran-Mantel-Hazenszel Test for the similar purpose.
    - Under the logistic model, we can use the Wald test or the likelihood-ratio test.

Coding Scheme for Categorical Predictors

A categorical variable of K levels can be represented by $K-1$ variables. Two commonly used coding schemes are dummy coding and effect coding.

- Different programs use different default schemes. PROC GENMOD uses dummy coding and PROC LOGISTIC uses effect coding. This can be changed by *PARAM=XXX* option.
- Because the same number of variables are used to represent the categorical variables, both coding schemes result in the same overall fit. However, the interpretation of the coefficients change.
- Both schemes need to specify a reference level. In SAS, this level can be changed by the *Ref=XX* option. For dummy coding, this reference level is considered as the baseline, that all the other levels need to be compared with this baseline. For effect coding, all levels are compared with the grand mean. The difference of this reference level and the other levels is, the effect of this reference level is not directly available and need to be computed additionally.

Illustrate using $2 \times 2 \times K$ table main effect model with $K = 5$. That is, logit $= \alpha + \beta X + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4$.

Dummy coding

Dummy coding uses dummy variables which only takes 0 or 1 value. If the 5th center is the baseline, then

		Dummy coding			
		C_1	C_2	C_3	C_4
Center	Center 1	1	0	0	0
	Center 2	0	1	0	0
	Center 3	0	0	1	0
	Center 4	0	0	0	1
	Center 5	0	0	0	0

- $\alpha =$ log odds for the 5-th center using trt B
- $\beta_i =$ difference of log odds for i-th center relative to the 5-th center using trt B

To see this,

$$\begin{array}{l}\text{Log odds of success using trt B}\\ = \alpha + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4\end{array} = \begin{cases} \alpha + \beta_1 & \text{Center 1} \\ \alpha + \beta_2 & \text{Center 2} \\ \alpha + \beta_3 & \text{Center 3} \\ \alpha + \beta_4 & \text{Center 4} \\ \alpha & \text{Center 5} \end{cases}$$

Effect coding

Here, the reference level always equals -1

		Dummy coding			
		C_1	C_2	C_3	C_4
Center	Center 1	1	0	0	0
	Center 2	0	1	0	0
	Center 3	0	0	1	0
	Center 4	0	0	0	1
	Center 5	-1	-1	-1	-1

- $\alpha =$ grand average of the log odds of success using trt B across all centers
- $\beta_i =$ difference of the log odds for i-th center relative to the grand average log odds using trt B
- For the reference center, the effect is $-\beta_1 - \beta_2 - \beta_3 - \beta_4$

To see this,

$$\begin{aligned}&\text{Log odds of success using trt B}\\&= \alpha + \beta_1 C_1 + \beta_2 C_2 + \beta_3 C_3 + \beta_4 C_4\end{aligned} = \begin{cases} \alpha + \beta_1 & \text{Center 1} \\ \alpha + \beta_2 & \text{Center 2} \\ \alpha + \beta_3 & \text{Center 3} \\ \alpha + \beta_4 & \text{Center 4} \\ \alpha - \beta_1 - \beta_2 - \beta_3 - \beta_4 & \text{Center 5} \end{cases}$$

Average these 5 terms,

$$\frac{(\alpha + \beta_1) + (\alpha + \beta_2) + (\alpha + \beta_3) + (\alpha + \beta_4) + (\alpha - \beta_1 - \beta_2 - \beta_3 - \beta_4)}{5} = \alpha$$

Building Meaningful Models

The model building process becomes more challenging as the number of explanatory variables increases because of the rapid increase in possible effects and interactions.

- The model should be complex enough to fit the data well,
- but simpler models are easier to interpret.

Confirmatory Analyses: Model building is guided by a theory, for instance, certain terms are included.

Exploratory Analyses: Typically there is a lack of underlying theory. A search among many models is conducted in a hope to find clues about which predictors are associated with the response and suggest questions for future research.

How Many Predictors Can You Use?

One guideline is that there should be at least 10 observations with $y = 0$ or $y = 1$ for every predictor included. For instance, if $y = 1$ only 30 times out of 1000 observations (that is, $y = 0$ 970 times), the model should nave no more than about 3 predictors even though the overall sample size is large.

When this guideline is violated,

- software still fits the model
- ML estimates may be quite biased and estimates of standard errors may be poor

Models with too many predictors often suffer from multicollinearity. That is, correlations among predictors cause none of the predictors is significant even though collectively they are significant.

General Consideration

For both grouped data and ungrouped data,

- Examine significance of parameters using Wald test and LR test
- Compare with saturated model with some measure for goodness of fit
- Compare two nested models using LR test
  (As null model is nested within any other models, so a special case here is to compare with the null model.)
- Compare two non-nested models using prediction performance and information criterion
- Influential diagnosis

For grouped data,

- Compare with saturated model using LR test
- Examine residuals
- Check overdispersion

For ungrouped data,

- Compare with saturated model using Hosmer and Lemeshow test which forms groups for continuous variables

Grouped Data, Ungrouped Data, and Continuous Predictors

- The grouped data are the totals of successes and failures at each combination of the predictor values
- The ungrouped data are the raw observations
- Although the ML estimates of parameters are the same for either form of data, the X^2 and G^2 statistics are not
- X^2 and G^2 only make sense for the grouped data. The large-sample theory applies when the fitted counts mostly exceed 5
- For logistic regression models with continuous or nearly continuous predictors, the X^2 and G^2 statistics do not have approximate chi-squared distributions. One way of getting around of this problem is to create categories so that there are adequate number of observations in each categories

Goodness of Fit and the Deviance

The most complex model possible is the saturated model, denoted by M_s: There is one parameter for each observation.

Goodness-of-fit examines adequacy of the current model to see if the current model is close enough to the saturated model.

Deviance evaluates the "distance" between the saturated model and current/working model (denoted by M_c) in terms of model fit. Use L to denote the log-likelihood.

$$Deviance = G^2(\mathrm{M}_c) = -2[L(\mathrm{M}_c) - L(\mathrm{M}_s)]$$

Goodness of Fit Test

In terms of tests, we can conduct goodness of fit test by testing whether all parameters in M_s but not in M_c are equal to 0.

If predictors are all categorical, we can use deviance as a test statistic. This test is exactly the LR test for testing nested models. Or we can use Pearson's χ^2 test. Both tests have the null distribution as χ^2_{df} where the $df =$ (n-number of parameters in M_c). Note, because we are under the grouped data case, therefore, the n is exactly the same as the number of groups.

For the ungrouped data, one may create groups on continuous variables to apply goodness of fit. One common approach is to form groups of approximately equal number of observations ro of equal class width. This is the basic principle of the Hosmer-Lemeshow test. The null distribution is χ^2_{df} with $df =$ the number of groups -2.

If there are K combinations across the levels of categorical variables, then the group number is K. Assume there are n_k observations belong to the k-th group, among which n_{1k} are the "success". Under the working model, the fitted count of "success" is $n_k\hat{\pi}_k$ with $\hat{\pi}_k$ computed from model M_c.

The deviance test statistics is

$$\begin{aligned} G^2(\mathrm{M}_c) &= 2\sum \text{observed} \cdot \log\left[\frac{\text{observed}}{\text{fitted}}\right] \\ &= 2\sum_k \left[n_{1k}\log\left(\frac{n_{1k}}{n_k\hat{\pi}_k}\right) + (n_k - n_{1k})\log\left(\frac{n_k - n_{1k}}{n_k - n_k\hat{\pi}_k}\right)\right] \end{aligned}$$

The corresponding Pearson statistic is

$$\begin{aligned} X^2(\mathrm{M}_c) &= \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} \\ &= \sum_k \left[\frac{(n_{1k} - n_k\hat{\pi}_k)^2}{n_k\hat{\pi}_k} + \frac{(n_{1k} - n_k\hat{\pi}_k)^2}{n_k - n_k\hat{\pi}_k}\right] = \sum_k \frac{(n_{1k} - n_k\hat{\pi}_k)^2}{n_k\hat{\pi}_k(1 - \hat{\pi}_k)} \end{aligned}$$

Hosmer-Lemeshow test takes a similar form as the Pearson's χ^2 test with all the counts obtained from the created groups, instead of natural groups. Little guidance for choose the group number here, but generally it should be larger than the number of parameters in the model.

Goodness of fit measure: pseduo R^2

In ordinary least square, R^2 is often used as a goodness of fit measure which accounts for the percentage of variation in the data that can be explained by the model. Therefore, it represents an improvement from null model to fitted model. Another name for R^2 is the multiple correlation, given the fact that it evaluates the relevance of all the predictors as a whole to the response.

For logistic regression, a pseduo R^2 measure can be obtained in similar way. Use M_{null} to represent the model with intercept only. Here the L is the log-likelihood.

- McFadden's ρ: $1 - \frac{L(\mathrm{M}_c)}{L(\mathrm{M}_{null})}$
- Adjusted McFadden's ρ: $1 - \frac{L(\mathrm{M}_c) - \text{no. of parameters in } \mathrm{M}_c}{L(\mathrm{M}_{null})}$

There are other measures developed for the same purpose. McFadden's are the most parallel to the R^2 in form.

Nested Models

When there are multiple choices of models available, we need to perform model selection among those candidate models. For any two models we compare, they may be nested within one another, or non-nested.

Let's consider our crab data as an example. Previously, we have fit a logistic model with a single continuous variable Width (denoted as x). Now, we further consider the variable Color. For Color, the values are

- 1, light medium
- 2, medium
- 3, dark medium
- 4, dark

Because there is a natural order of those color values, we can either model Color as a categorical variable, or a continuous variable.

Therefore, we may come up with the following models.

- Model (a): $\text{logit}[\Pr(Y = 1)] = \alpha + \beta_1 x$
- Model (b): $\text{logit}[\Pr(Y = 1)] = \alpha + \beta_2 c$, where c treats Color as a continuous variable.
- Model (c): $\text{logit}[\Pr(Y = 1)] = \alpha + \beta_1 x + \beta_2 c$
- Model (d): $\text{logit}[\Pr(Y = 1)] = \alpha + \beta_1 x + \beta_3 c_1 + \beta_4 c_2 + \beta_5 c_3$, where c_1, c_2, and c_3 are dummy variables for Color that takes value 1, 2, and 3, respectively. The level 4 is the baseline level.

Questions:

- Are Model (a) and Model (b) nested models?
- Are Model (a) and Model (c) nested models?
- Are Model (c) and Model (d) nested models?

Answers:

- Are Model (a) and Model (b) nested models?

  No. Model (a) is based on variable Weight, and Model (b) is based on variable Color. Both models contain a variable that the other one does not use.

- Are Model (a) and Model (c) nested models?

  Yes. Model (c) is based on variables Weight and Color, therefore, it is a bigger model than Model (a). If we exclude variable Color, then the Model (c) can be reduced to exactly the Model (a).

- Are Model (c) and Model (d) nested models?

  Surprisingly, yes. Sometimes, the nested models are less obvious to detect. In fact, we can reduce model (d) to model (c) by assuming

$$\beta_3 = 3\beta_5, \beta_4 = 2\beta_5,$$

$$\begin{aligned}
\rightarrow \text{logit}[\Pr(Y=1)] &= \alpha + \beta_1 x + 3\beta_5 c_1 + 2\beta_5 c_2 + \beta_5 c_3 \\
&= \alpha + \beta_1 x + \beta_3(3c_1 + 2c_2 + c_3) \\
&= \alpha + \beta_1 x + \beta_3(4 - c) \\
&= (\alpha + 4\beta_3) + \beta_1 x + (-\beta_3)c
\end{aligned}$$

Testing Nested Models

For nested models, we can use likelihood ratio test to determine whether the model with more parameters has similar model fit as the model with reduced number of parameters.

If both models lead to similar model fit, then we may go with the model with reduced parameters. In general,

$$\begin{aligned} \mathrm{H}_0 &: \text{the simpler model } \mathrm{M}_1 \text{ is true} \\ \mathrm{H}_1 &: \text{the more complext model } \mathrm{M}_2 \text{ is true} \end{aligned}$$

We can also write the hypothesis in terms of tests for the parameters, e.g.,

- Model (a) vs. Model (c) $\mathrm{H}_0 : \beta_2 = 0$
- Model (a) vs. Model (d) $\mathrm{H}_0 : \beta_3 = \beta_4 = \beta_5 = 0$
- Model (c) vs. Model (d) $\mathrm{H}_0 : \beta_3 = 3\beta_5, \beta_4 = 2\beta_5$

The likelihood ratio test statistic is

$$2L(\mathrm{M}_2) - 2L(\mathrm{M}_1),$$

where L is the log-likelihood.

We can express this test statistics in terms of deviance $G^2(\mathrm{M}_1) - G^2(\mathrm{M}_2)$. Because

$$\begin{aligned} G^2(\mathrm{M}_1) &= -2[L(\mathrm{M}_1) - L(\mathrm{M}_s)] \\ G^2(\mathrm{M}_2) &= -2[L(\mathrm{M}_2) - L(\mathrm{M}_s)] \\ \rightarrow \quad 2L(M_2) - 2L(M_1) &= G^2(\mathrm{M}_1) - G^2(\mathrm{M}_2) \end{aligned}$$

The null distribution is χ^2_{df} given large sample size, where $df =$ no. of parameters in M_2 - no. of parameters in M_1.

An asymptotically equivalent test is based on the Pearson statistic

$$X^2(\mathrm{M}_1) - X^2(\mathrm{M}_2).$$

For example,

- Model (a) vs. Model (c) H$_0$: $\beta_2 = 0$ null distn $= \chi^2_1$
- Model (a) vs. Model (d) H$_0$: $\beta_3 = \beta_4 = \beta_5 = 0$ null distn $= \chi^2_3$
- Model (c) vs. Model (d) H$_0$: $\beta_3 = 3\beta_5, \beta_4 = 2\beta_5$ null distn $= \chi^2_2$

Comments:

Both G^2(M$_1$) and G^2(M$_2$) are likelihood ratio test statistic comparing the fit of each model to the saturated model. Recall that, for the individual G^2 statistics to be well approximated by χ^2, we need the group data. Or to say, at least 80% of the $\hat{\mu}_i$ and $(n_i - \hat{\mu}_i)$ are greater than 5.

However, even if the individual G^2s are far away from χ^2, their difference may be reasonably approximated by χ^2. For a good approximation, we need to have a large sample size and small *df*. The same applies to the Pearson statistic.

Crab Data Example: Backward Elimination

Backward Elimination starts with a complex model and then determines whether a certain term can be taken out.

Use Crab data as example. For those 173 samples, we consider covariates

- C = color (light medium, medium, dark medium, and dark)
- S = spine condition (both good, one broken, both broken)
- W = width

Model	Predictors	Deviance	df	AIC	Models Compared	Deviance Difference
1	C:S + C:W + S:W	173.7	155	209.7	–	
2	C + S + W	186.6	166	200.6	(2)–(1)	12.9 (df = 11)
3a	C + S	208.8	167	220.8	(3a)–(2)	22.2 (df = 1)
3b	S + W	194.4	169	202.4	(3b)–(2)	7.8 (df = 3)
3c	C + W	187.5	168	197.5	(3c)–(2)	0.9 (df = 2)
4a	C	212.1	169	220.1	(4a)–(3c)	24.6 (df = 1)
4b	W	194.5	171	198.5	(4b)–(3c)	7.0 (df = 3)
5	(C = dark) + W	188.0	170	194.0	(5)–(3c)	0.5 (df = 2)
6	None	225.8	172	227.8	(6)–(5)	37.8 (df = 2)

Information Criterion

We have learned that LR test can be usd to compare two nested models, where the difference of 2× log-likelihood between the two models is used as the test statistic. Here we know that the large model has better model fit, that is higher log-likelihood value; the question is, whether it is a significant higher value.

Another way to evaluate the model fit is directly using a criterion that considers both the model fit and model complexity. A common way is using information criterion and the most commonly seen are AIC and BIC (known as SC, Schwarz Criterion).

$$\begin{aligned} AIC &= -2 \text{ log-likelihood} + 2 \times \text{number of } \beta\text{s} \\ BIC &= -2 \text{ log-likelihood} + \log(n) \times \text{number of parameters in model} \end{aligned}$$

A small AIC and BIC are preferred. Both can be applied to non-nested models. BIC tends to select a model with less parameters than AIC does.

Prediction Performance

We may assume that a better model has superior predictive ability.

Classification table is useful tool to measure the agreement between the observed dichotomous response variable and the **predicted binary label** created by a certain rule. For example, if the $\hat{\pi}_i > 0.5$, the predicted label is 1, otherwise 0.

Another way is to measure the correlation between **predicted probabilities** and observed dichotomous response variable. The higher the correlation, the better the performance.

- Somer's D
- Goodman-Kruskal Gamma
- Kendall's Tau-α
- Area under the ROC curve (c)

Classification Table

The classification table is a cross-tabulation between

- observed binary outcome y. These are true/actual values
- predicted binary label $\hat{y}$. These are model-based and can be different by models and cutoff rules.

The predicated label $\hat{y}$ is obtained by applying a cutoff π_0 to $\hat{\pi}$.

$$\begin{cases} \hat{y} = 1, & \text{if } \hat{\pi} > \pi_0 \\ \hat{y} = 0, & \text{if } \hat{\pi} <= \pi_0 \end{cases}$$

Classification Table

		Prediction	
		$\hat{y} = 1$	$\hat{y} = 0$
True/	$y = 1$	a	b
Actual	$y = 0$	c	d

Three useful summaries of predictive power can be calculated:

sensitivity $= P(\hat{y} = 1 | y = 1) = \frac{a}{a+b}$

specificity $= P(\hat{y} = 0 | y = 0) = \frac{d}{c+d}$

P(correct classification) $= \frac{a+d}{a+b+c+d}$

Classification Table for Crab Data under simple logistic regression with variable Width and two cutoffs for prediction

		Prediction, $\pi_0 = 0.64$		Prediction, $\pi_0 = 0.50$		
		$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
True/	$y = 1$	74	37	94	17	111
Actual	$y = 0$	20	42	37	25	62

For $\pi_0 = 0.64$

sensitivity = $P(\hat{y} = 1|y = 1) = \frac{74}{111}$

specificity = $P(\hat{y} = 0|y = 0) = \frac{42}{62}$

P(correct classification) = $\frac{74+42}{111+62}$

For $\pi_0 = 0.5$

sensitivity = $P(\hat{y} = 1|y = 1) = \frac{94}{111}$

specificity = $P(\hat{y} = 0|y = 0) = \frac{25}{62}$

P(correct classification) = $\frac{94+25}{111+62}$

Classification table has limitations:

It collapses continuous predictive value $\hat{\pi}$ into binary ones. The choice of π_0 is arbitrary and results may be sensitive.

Correlation Measurements

Let n be the total number of subjects. Then there are $n(n-1)/2$ distinct pairs of subjects. Let t be the total evaluated pairs, which is formed by each subject with event to every subject without event. The pair is

- tied: if the two predicted probabilities are within 0.002 of one-another
- concordant: if the subject with the higher predicted probability has the higher value for the response variable
- discordant: if the subject with the higher predicted probability has the lower value for the response variable
- Let n_t, n_c, and n_d denote the number of tied, concordant, and discordant pairs, respectively. $t = n_t + n_c + n_d$.

$$\begin{aligned}
\text{Somer's D} &= (n_c - n_d)/t \\
\text{Gamma} &= (n_c - n_d)/(n_c + n_d) \\
\text{Tau-}\alpha &= (n_c - n_d)/[n(n-1)/2] \\
c &= 0.5 \times (1 + \text{Somer's D}) = (n_c + 0.5 \times n_t)/t
\end{aligned}$$

Kendall's Tau-α is the most conservative of the three and closest in spirit to the R^2 statistic in linear regression

[SAS Output]

```
Association of Predicted Probabilities and Observed Responses

Percent Concordant     73.5    Somers' D    0.485
Percent Discordant     25.0    Gamma        0.492
Percent Tied            1.5    Tau-a        0.224
Pairs                  6882    c            0.742
```

- 73.5% of the 6882 total evaluated pairs are concordant
- Kendall's Tau-α statistics is 0.224 indicating a moderate, positive association between the predicted probabilities and the response variable

Residuals for Logit Models

What is the influence of individual observations? With categorical predictors, we can use residuals to compare observed and fitted counts.

Note that, under ungrouped data where explanatory variables are continuous, each $n_i = 1$. Then the residuals are usually uninformative.

Let y_i denote the number of "successes" for n_i trials at setting i of the explanatory variables. Let $\hat{\pi}_i$ denote the estimated probability of success for the model fit. The fitted number of successes is $n_i\hat{\pi}_i$

$$\text{Pearson residual} = e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

It is obvious that the Pearson statistic $X^2(M) = \sum_i e_i^2$.

When n_i is large, e_i has an approximate normal distribution. When the model holds, $\{e_i\}$ has an approximate expected value of zero but a smaller variance than a standard normal variable.

Therefore, we introduce the standardized residuals,

$$\text{Standardized residual} = \frac{y_i - n_i\hat{\pi}_i}{SE(y_i - n_i\hat{\pi}_i)} = \frac{e_i}{\sqrt{1-h_i}} \sim N(0,1)$$

where $0 < h_i < 1$ is the "leverage" of the ith observation. It is the ith diagonal of the "hat matrix" $\mathbf{H}$ (more details later). It takes into account the variation introduced in using $\hat{\pi}_i$ instead of π_i.

Similar to the Pearson residuals, we can also define deviance residuals d_i.

$$\begin{aligned}\text{Devianc residual} &= |d_i| \\ &= \left\{2\left[n_{1i}\log\left(\frac{n_{1i}}{n_i\hat{\pi}_i}\right) + (n_i - n_{1i})\log\left(\frac{n_i - n_{1i}}{n_i - n_i\hat{\pi}_i}\right)\right]\right\}^{1/2}.\end{aligned}$$

The sign of d_i is set to be the same as that of $n_{1i} - n_i\hat{\pi}_i$. It is obvious that Deviance statistic $G^2(M) = \sum_i d_i^2$.

Usually, if residuals fall outside $(-3, 3)$, there is evidence of lack of fit for those observations. Some sources use $(-2, 2)$.

Influence Diagnostics for Logistic Regression

Influence pertains to how much parameter estimates would change if one observation was omitted. The diagnostic tools are very similar to those used in linear regression $Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon$.

Leverage

- In linear regression, $\hat{\beta} = (X^TX)^{-1}X^TY \to \hat{Y} = X\hat{\beta} = X(X^TX)^{-1}X^TY$. We define hat matrix $H_{n\times n} = X(X^TX)^{-1}X$, which transforms Y to $\hat{Y}$. The i-th diagonal element of H, $\hat{h}_i$, is leverage for the i-th observation. It evaluates how far x_i lies from the centroid of $x_1, \ldots, x_n$.
- Leverage values greater than $2p/n$ or $3p/n$ are often flagged as potentially influential. It may be helpful to identify those observations, remove them one at a time, and see how the parameters change.
- Unlike linear regression, the leverage $\hat{h}_i$ depends on both the model fit $\hat{\beta}$ as well as the covariates X. Points have extreme predictor values may not have high leverage $\hat{h}_i$ if $\hat{\pi}_i$ is close to 0 to 1. Still may be useful for detecting extreme predictor values x_i.

Leverage

- The hat matrix for logistic regression is

$$H^*_{n\times n} = X^*({X^*}^T X^*)^{-1}{X^*}^T,$$

  where the $X^* = W^{1/2}X$ and W is a diagonal matrix with the i-th element $n_i\hat{\pi}_i(1-\hat{\pi}_i)$. Because $\hat{\pi}$ contains $\hat{\beta}$, the hat matrix H^* relates to the model.

- Recall that, this h_i has used to adjust Pearson residuals to obtain standardized residual, which is standard normal. $r_i = e_i/\sqrt{1-h_i}$.

- Comment on X^*: In GLM we mentioned the Fisher scoring algorithm, which can be written as an iterative re-weighted least squares. For logistic model, the weight is W. There are other analogues between linear model and logistic model. The estimated covariance matrix is $(X^TX)^{-1}$ for linear regression, and $(X^TWX)^{-1}$ for logistic model.

Model Diagnostics: Leave-one-out measures

- Let $\hat{\beta}_i$ be an estimated effect. Let $\hat{\beta}_{i(j)}$ be its updated estimate after removing the j observation.

  $$Dfbeta_{ij} = \frac{\hat{\beta}_i - \hat{\beta}_{i(j)}}{SE(\hat{\beta}_i)}$$

  $Dfbeta_{ij}$ measures the standardized difference in $\hat{\beta}_i$ when one observation is removed.

- Confidence interval displacement diagnostic $c_i = e_i^2 \hat{h}_i/(1 - \hat{h}_i)^2$. It measures the change in the joint confidence region for β when one observation is removed.

- Similarly, we can measure the change in X^2 and G^2 when one observation is removed.