# ST3241 Categorical Data Analysis I Logistic Regression

An Introduction and Some Examples

## Example Applications

- Business Applications
  - The probability that a subject pays a bill on time may use predictors such as the size of the bill, annual income, occupation, mortgage and debt obligations, percentage of bills paid on time in the past, and other aspects of an applicants credit history.
  - A company that relies on catalog sales may determine whether to send a catalog to a potential customer by modeling the probability of a sale as a function of indices of past buying behavior.

#### • Genetics

- Modeling the probability that an offspring inherits an allele of one type instead of another type a function of phenotypic values of various traits for that offspring.
- To model the probability that affected siblings pairs (ASP) have identity-by-descent allele sharing and tested its heterogeneity among the research centers.

## A Simple Logistic Regression Model

• For a binary response variable Y and an explanatory variable X, let  $\pi(x) = P(Y = 1 | X = x) = 1 - P(Y = 0 | X = x)$ .

$$\pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

• Equivalently, the log odds, called the *logit*, has the linear relationship

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

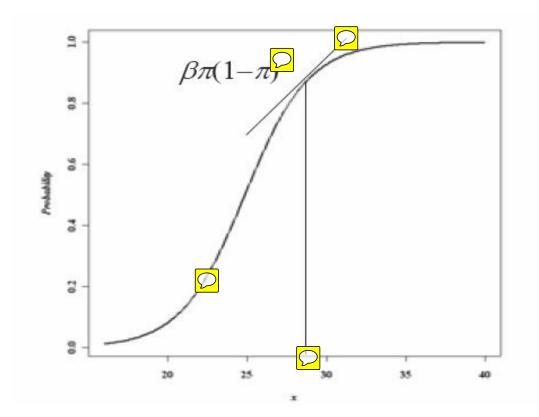
• This equates the logit link function the linear predictor

#### Interpretation of Parameters

- The parameter  $\beta$  determines the rate of increase or decrease of the S-shaped curve.
- The sign of  $\beta$  indicates whether the curve ascends or descends
- The rate of change increases as  $|\beta|$  increases.
- When the model holds with  $\beta = 0$ , then  $\pi(x)$  is identical at all x, so the curve becomes a horizontal straight line, and Y is then independent of X.  $\square$

# **Interpretation of Parameters**

• Liner Approximation Interpretation



#### **Linear Approximation Interpretations**

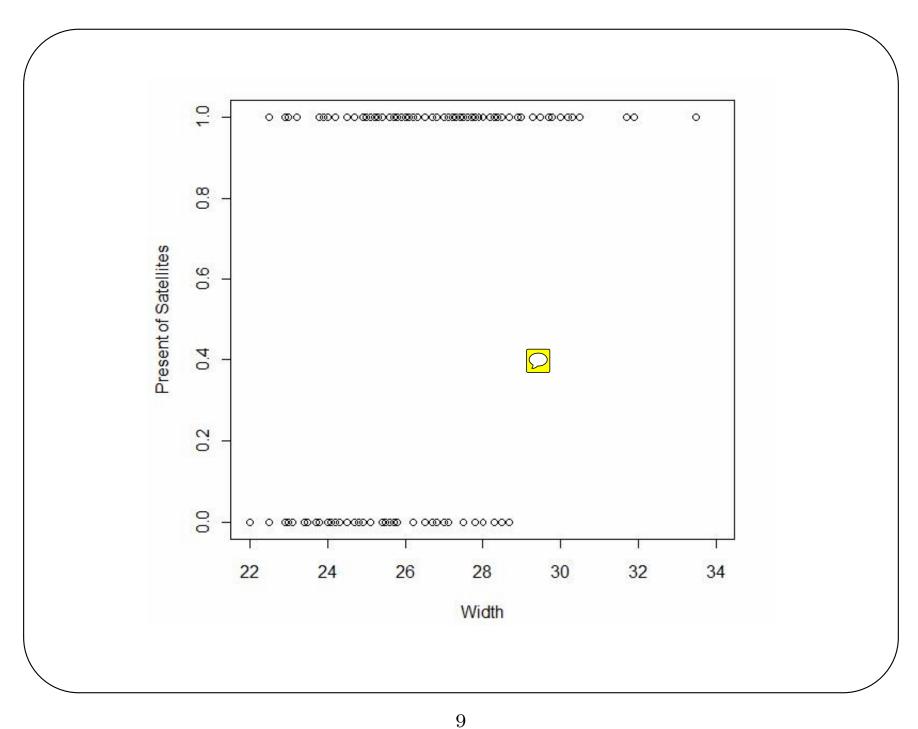
- A straight line drawn tangent to the curve at a particular x value, describes the rate of change at that point.
- For logistic regression parameter  $\beta$ , that line has slope equal to  $\beta \pi(x)[1-\pi(x)]$ .
- For example, the line tangent to the curve at x for which  $\pi(x) = 0.5$  as slope  $\beta(0.5)(0.5) = 0.25\beta$ .
- By contrast when  $\pi(x) = 0.9$  or 0.1, it has slope  $0.09\beta$

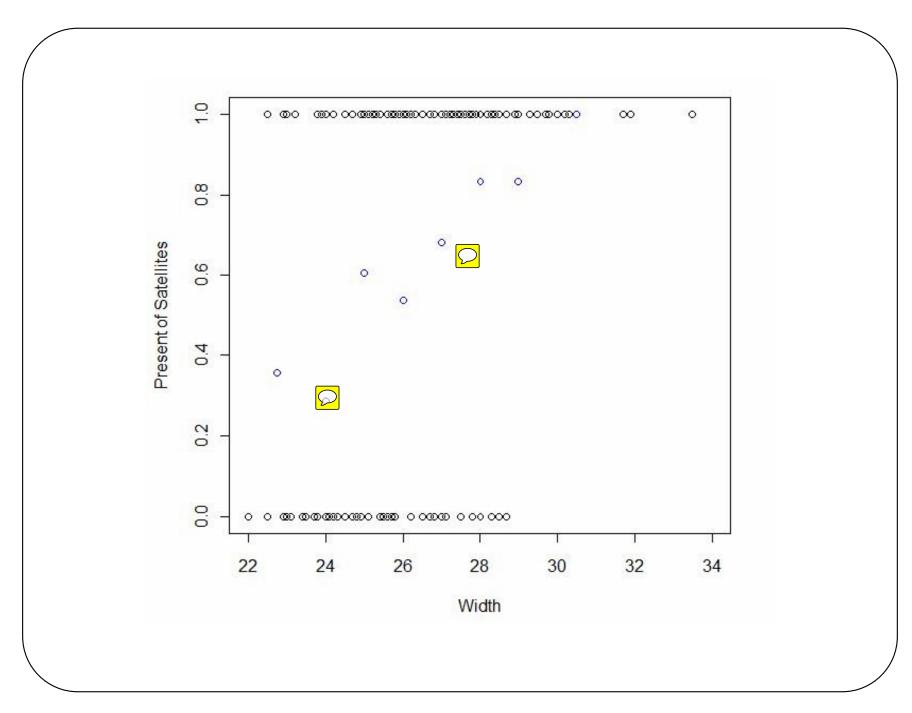
#### Linear Approximation Interpretations

- The slope approaches 0 as the probability approaches 1.0 or 0.
- The steepest slope of the curve occurs at x for which  $\pi(x) = 0.5$ ; that x value is  $x = -\alpha/\beta$ .
- This value of  $\mathfrak{S}$  sometimes called the *median effective level* and is denoted by  $EL_{50}$ .
- It represents the level at which each outcome has a 50% chance.

#### Example: Horseshoe Crabs

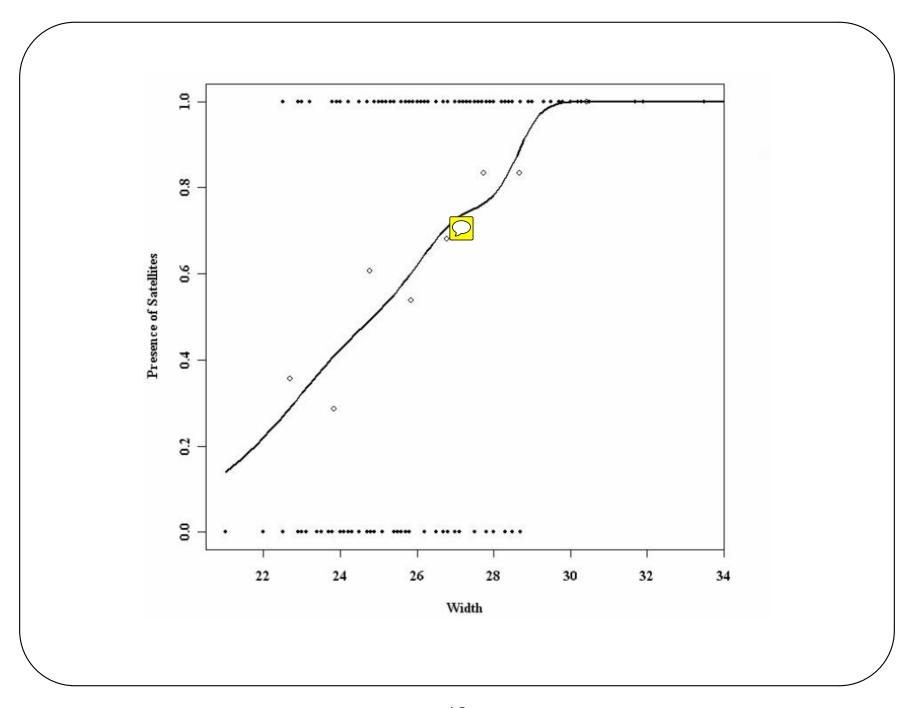
- Define Y = 1 if a female crab has at least one satellite, and Y = 0 if she has no satellite.
- We first use the female crab's width sole predictor
- It is difficult to determine whether a logistic regression model is reasonable by plotting Y against X.





## Example: Modeling

- Let  $\pi(x)$  denote the probability that a female horseshoe crab of width x has a satellite.
- Linear probability model:  $\pi(x) = \alpha + \beta x$
- Fitted model:  $\hat{\pi}(x) = -1.766 + 0.092x$
- This model provides a simple interpretation and realistic predictions over most of the width range.
- But it is inadequate for extreme values.
- e.g. at the maximum width in this sample of 33.5, its predicted probability equals  $-1.766 + 0.092(33.5) = 1.3\square$

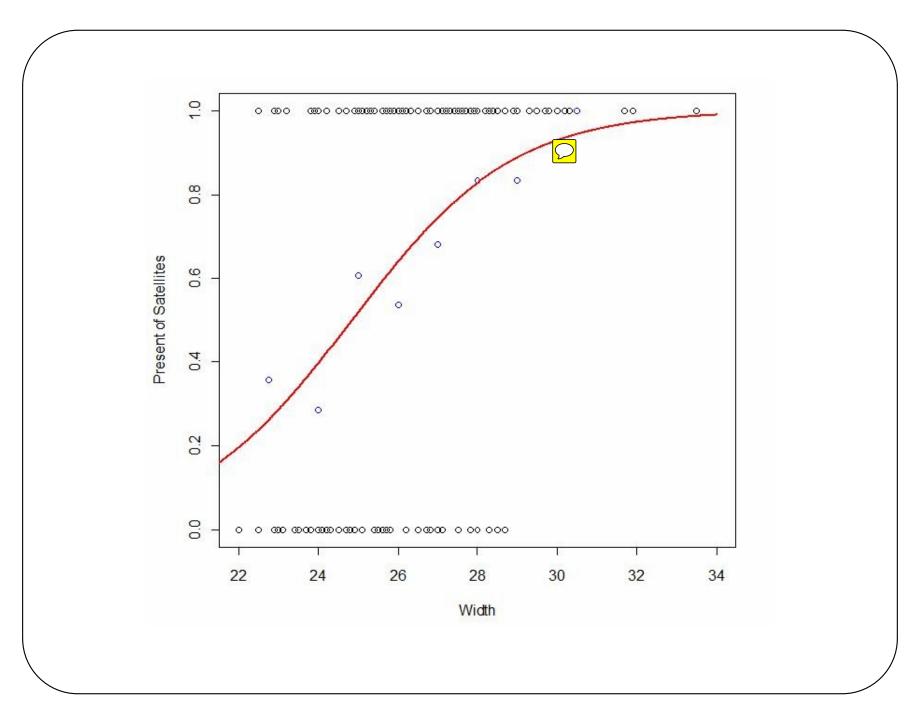


#### Example: Logistic Regression Model

• Using ML estimation, the predicted probability of having a satellite is

$$\hat{\pi}(x) = \frac{\exp(-12.351 + 0.497x)}{1 + \exp(-12.351 + 0.497x)}$$

- Since  $\hat{\beta} > 0$ , the predicted probability  $\hat{\pi}(x)$  is higher at larger width values.
- At the minimum width, x = 21, the predicted probability is  $\bigcirc$   $\exp(-12.351 + 0.0497(21))/[1 + \exp(-12.351 + 0.0497(21))] = 0.129$
- At the maximum width, x = 33.5, the predicted probability equals  $\exp(-12.351 + 0.0497(33.5))/[1 + \exp(-12.351 + 0.0497(33.5))] = 0.987$
- The median effective level:  $EL_{50} = -\hat{\alpha}/\hat{\beta} = 12.351/0.497 = 24.8$



#### **Notes**

- At the sample mean width of 26.3 cm, the predicted probability of a satellite equals 0.674.
- The incremental rate of change in the fitted probability at that point is

$$\hat{\beta}\hat{\pi}(26.3)(1 - \hat{\pi}(26.3)) = 0.497(0.674)(0.326) = 0.11$$

- For female crabs near the mean width, the estimated probability of a satellite increases at the rate of 0.11 per cm increase in width.
- The predicted rate of change is greatest at x = 24.8, the predicted probability increases at the rate of 0.25(0.497) = 0.12 per cm increase in width.

#### Odds Ratio Interpretation

• The odds of a success. e. Y = 1 at X = x is:

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

• The odds of a success at X = x + 1 is:

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

• Therefore, the odds ratio:

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

• Therefore,  $\beta$  can be considered as a  $\log$  odds ratio for one unit width increase.

#### Example: Horseshoe Crabs

- The estimated odds of a satellite multiply by  $\exp(\hat{\beta}) = \exp(0.497) = 1.64$  for each centimeter increase in width.
- The predicted probability at the mean value of width, x = 26.3, is 0.674, and odds = 0.674/0.326 = 2.07.
- At x = 27.3 the predicted probability equals 0.773 and odds = 0.773/0.227 = 3.40.
- Now note that, 3.40 = 2.07(1.64).

#### Confidence Interval For Effects

• For a simple logistic regression model:

$$logit[\pi(x)] = \alpha + \beta x$$

a large sample 95% confidence interval is

$$\hat{\beta} \pm z_{\alpha/2}(ASE)$$

• Exponentiating the endpoints of this interval yields one for  $e^{\beta}$  the odds ratio for a 1-unit increase in X.

#### Example: Horseshoe Crab Data

- The estimated effect of width in the fitted equation for the probability of a satellite  $\hat{\beta} = 0.497$  with ASE = 0.102.
- A 95% confidence interval for  $\beta$  is:  $0.497 \pm 1.96(0.102) = (0.298, 0.697)$
- The confidence interval for the effect on the odds per centimeter increase in width equals  $(e^{0.298}, e^{0.697}) = (1.35, 2.01)$

#### Example: Continued · · ·

- At  $\pi = 0.5$ , the estimated rate of change is  $.25\beta$
- A 95% confidence interval for  $0.25\beta$  equals 0.25 times the endpoints of the interval for  $\beta$ , or [0.25(0.298), 0.25(0.697)] = (0.074, 0.174).
- If the logistic regression model holds, for values of x near the width at which  $\pi = 0.5$ , the rate of increase in the probability of a satellite per centimeter increase in width falls between 0.07 and 0.17.

## Tests of Significance

- Wald's Test Statistic  $z = \hat{\beta}/ASE$
- Under  $H_0: \beta = 0, \mathbf{z}^2$  has approximately a chi-squared distribution with 1 d.f.
- The likelihood ratio test uses the log likelihoods at  $\hat{\beta}$  under null and alternative hypotheses.
- It uses twice the difference between the maximized log likelihood at  $\hat{\beta}$  and at  $\beta = 0$ .
- It also has approximately a chi-squared distribution with  $1 \ d.f.$

#### Tests of Significance

- The score test uses the log likelihood at  $\beta = 0$  through the derivative of the log likelihood at that point
- Under  $H_0$ , the score test statistic also has a large sample chi-squared distribution with  $1 d.f. \square$
- For large samples, these three tests usually give similar results.
- The likelihood ratio test is preferred over the others as it uses more information
- When  $|\beta|$  is relatively large, the Wald test is not as powerful as the likelihood ratio test.

#### **Estimates of Probability**

• The estimated probability that Y = 1 at X = x is

$$\hat{\pi}(x) = \frac{\exp(\hat{\alpha} + \hat{\beta}x)}{1 + \exp(\hat{\alpha} + \hat{\beta}x)}$$

• The large sample standard error of the estimated logit ::

$$Var(\hat{\alpha} + \hat{\beta}x) = Var(\hat{\alpha}) + x^{2}Var(\hat{\beta}) + 2xCov(\hat{\alpha}, \hat{\beta})$$

• A 95% confidence interval for the true logit is then

$$(\hat{\alpha} + \hat{\beta}x) \pm 1.96 \times \sqrt{Var(\hat{\alpha} + \hat{\beta}x)}$$

• Substituting each endpoint into the inverse transformation  $\pi(x) = \exp(\log it)/[1 + \exp(\log it)]$  gives a corresponding interval for  $\pi(x)$ 

#### Example: Horseshoe Crab Data Using SAS

```
%include 'd:\class\st3241\crab.sas';
proc logistic data=crab desc;
  model y = width;
  output out=predict p=pi_hat lower=LCL
  upper=UCL;
run;
proc print data=predict;
  where width=21 or width=26.3 or width=33.5;
  var width pi_hat LCL UCL;
run;
```

#### Output

The LOGISTIC Procedure

Model Information

Data Set WORK.CRAB

Response Variable y

Number of Response Levels 2

Model binary logit

Optimization Technique Fisher's scoring

Number of Observations Read 173

Number of Observations Used 173

```
Response Profile
```

Ordered Total

Value y Frequency

1 1 111

2 0 62

Probability modeled is y=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied

#### 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

```
Testing Global Null Hypothesis: BETA=0

Test Chi-Square DF Pr > ChiSq

Likelihood Ratio 31.3059 1 <.0001

Score 27.8752 1 <.0001

Wald 23.8872 1 <.0001
```

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

			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

Obs	width	pi_hat	LCL	UCL
14	21.0	0.12910	0.05109	0.28981
107	26.3	0.67400	0.59147	0.74700
141	33.5	0.98670	0.93878	0.99722

#### To Print Estimated Covariances

```
proc logistic data=crab desc covout outest=temp;
  model y = width;
  output out=predict p=pi_hat upper=ucl lower=lcl;
run;

proc print data=temp noobs;
  where _type_='COV';
  var _name_intercept width;
run;
```

# Output

\_NAME\_ Intercept width

Intercept 6.91023 -0.26685

width -0.26685 0.01035

#### Horseshoe Crab Data Example with R

```
>crabs<-
  read.table('d:/class/st3241/crab.txt', header=T)
>names(crabs)
[1] "color" "spine" "width" "satell" "weight"
>crabs$psat<-crabs$satell>0
>crabs.lg<-glm(psat~width, family=binomial(), data=crabs)
>summary(crabs.lg)
```

#### Output

#### Output

```
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: 194.45 on 171 degrees of freedom
AIC: 198.45
Number of Fisher Scoring iterations: 4
```

#### Prediction in R

# Sample Proportions vs Model-based Fits

- In some situations, one may ignore the model fit and simply use sample proportions (i.e. the saturated model) to estimate such probabilities.
- For example, 6 female crabs in the sample had x = 26.5, and 4 of them had satellites.
- The sample proportion estimate at x = 26.5 is

$$\hat{\pi} = 4/6 = 0.67$$

• The 95% confidence interval based on these six observations alone equals (0.30, 1.04)

# Sample Proportions vs Model-based Fits

• The model based estimate of probability at x = 26.5 is

$$\hat{\pi}(x) = 0.695$$

- The corresponding 95% confidence interval is (0.61, 0.77)
- When the logistic regression model truly holds, the model based estimator of a probability is considerably better than the sample proportions.
- The standard error for the model based estimate is 0.04, whereas the standard error for the sample proportions is  $\sqrt{(0.67)(1-0.67)/6} = 0.19$

# Sample Proportions vs Model-based Fits

- In practice, the model is not exactly the true relationship  $\bigcirc$  between  $\pi(x)$  and x.
- However, if it approximates the true probabilities decently, its estimator still tends to be closer than the sample proportions.
- The model smoothes the sample data, somewhat dampening the observed variability.

# Model Checking

- There is no guarantee that a certain logistic regression model will fit the data well.
- To detect lack of fit, one can use likelihood ratio tests to compare the model to more complex ones.
- If more complex models do not fit better, this provides some assurance that the model chosen is reasonable.

# Model Checking

- Fitted logistic regression models provide predicted probabilities that Y = 1.
- At each setting of X, one can multiply the predicted probability by the total number of observations to obtain fitted count.
- The test of the null hypothesis that the model holds compares the fitted and observed counts using a Pearson's  $\chi^2$  or likelihood ratio  $G^2$  statistic.

# Asymptotic Distributions and Degrees of Freedom

- For a fixed number of settings, when most fitted counts equal at least 5,  $\chi^2$  or  $G^2$  have approximate chi-squared distributions.
- The degrees of freedom, called the residual d.f. for the model, equal the number of sample logits minus the number of model parameters.

#### Goodness of Fit for Models w. Continuous Predictors

- For models with continuous predictors, we may have too many settings of explanatory variable resulting in very small fitted counts.
- e.g. in the Horseshoe crab data, width takes 66 distinct values for the 173 crabs.
- In such cases, we compare the observed and fitted values in the grouped form.
- When explanatory variables are continuous, it is difficult to analyze lack of fit without some type of grouping.

# Example: Horseshoe Crab Data

	Obs	erved	Fitted		
Wiph	Yes	No	Yes	No	
<23.25	5	9	3.64	10.36	
23.25-24.25	4	10	5.31	8.69	
24.25-25.25	17	11	13.78	14.22	
25.25-26.25	21	18	24.23	14.77	
26.25-27.25	15	7	15.94	6.06	
27.25-28.25	20	4	19.38	4.62	
28.25-29.25	15	3	15.65	2.35	
>29.25	14	0	13.08	0.92	

# Example: Continued

- In each category, the fitted value for response yes is the average of predicted probabilities  $\hat{\pi}(x)$  for all crabs having width in that category.
- The fitted value for response no is the **average** of  $1 \hat{\pi}(x)$  for those crabs.
- The test statistics are:  $\chi^2 = 2.18$  and  $G^2 = 2.13$  with d.f. = 8  $2 = 6 \square$
- Since P-value> 0.3, none of these tests show a lack of fit for the model.

#### SAS Codes: Read The Data And Create The Groups

```
%include 'd:\class\st3241\crab.sas';
data crab1;
set crab;
wcat=0;
if width<=23.25 then wcat=1;
if 23.25< width<=24.25 then wcat=2;
if 24.25< width<=25.25 then wcat=3;
if 25.25< width<=26.25 then wcat=4;
if 26.25< width<=27.25 then wcat=5;
if 27.25< width<=28.25 then wcat=6;
if 28.25< width<=29.25 then wcat=6;
if 29.25< width<=29.25 then wcat=7;
if 29.25< width then wcat=8;
run</pre>
```

#### Fit The Model And Create The Table of Probabilities

```
proc logistic data=crab1 desc noprint;
  model y = width ;
  output out=predict p=pi_hat;
run;
proc sql;
  create table pred2 as
  select pi_hat, wcat, sum(y) as Num_yes, sum(1-y) as
  Num_no, sum(pi_hat) as Fitted_yes,
    sum(1-pi_hat) as Fitted_no
  from predict
  group by wcat;
quit;
```

# Sort and Print The Table

```
proc sort data=pred2;
  by wcat;
run;
data pred3;
set pred2;
by wcat;
if first.wcat;
run;
proc print data = pred2;
  format wcat wcat.;
var wcat Num_yes Num_no Fitted_yes Fitted_no;
run;
```

# Output

				Fitted_	Fitted_
Obs	wcat	Num_yes	Num_no	yes	no
1	<=23.25	5	9	3.6354	10.3646
2	23.25-24.25	4	10	5.3060	8.6940
3	24.25-25.25	17	11	13.7776	14.2224
4	25.25-26.25	21	18	24.2277	14.7723
5	26.25-27.25	15	7	15.9378	6.0622
6	27.25-28.25	20	4	19.3833	4.6167
7	28.25-29.25	15	3	15.6502	2.3498
8	>29.25	14	0	13.0820	0.9180

# Likelihood Ratio Tests for Goodness of Fit

- Let  $M_0$  and  $M_1$  be two competing models
- Let  $L_0$  and  $L_1$  be the maximized log-likelihoods under the models  $M_0$  and  $M_1$  respectively.
- Similarly, let  $L_S$  denote the maximized log likelihood of the saturated model.
- Then the deviances for the models  $M_0$  and  $M_1$  are  $G^2(M_0) = -2(L_0 L_S)$  and  $G^2(M_1) = -2(L_1 L_S)$ .

#### <u>Likelihood Ratio Tests for Goodness of Fit</u>

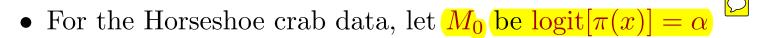
- Denote the likelihood ratio statistic for testing  $M_0$ , given that  $M_1$  holds, by  $G^2(M_0|M_1)$ .
- Then

$$G^{2}(M_{0}|M_{1}) = -2(L_{0} - L_{1}) = -2(L_{0} - L_{S}) - [-2(L_{1} - L_{S})]$$

$$= G^{2}(M_{0}) - G^{2}(M_{1})$$

- This statistic is large when  $M_0$  fits poorly compared to  $M_1$ .
- It has a large sample chi-squared distribution with d.f. equal to the difference between the residual d.f. values for the two models.  $\square$

#### Example



- Let  $M_1$  be the model  $\operatorname{logit}[\pi(x)] = \alpha + \beta x$
- The simpler model posits independence of width and having a satellite.
- Then  $G^2(M_0) = 225.76$  with d.f. = 173 1 = 172 and  $G^2(M_1) = 194.45$  with d.f. = 173 2 = 171.
- Therefore  $G^2(M_0|M_1) = G^2(M_0) G^2(M_1) = 225.76 194.45 = 31.31$  with d.f. = 172 171 = 1.
- We conclude that the model  $M_0$  is not a good fit compared to model  $M_1$  since the associated P-value is 2.199438e 08.

#### Residuals

- Let  $y_i$  denote the number of successes for  $n_i$  trials at the *i*-th setting of the explanatory variables.
- Let  $\hat{\pi}_i$  denote the predicted probability of success for the model fit.
- Then the Pearson residual for the setting i is:

$$e_i = \frac{y_i - n_i \hat{\pi}_i \mathcal{O}}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

• The Pearson statistic for testing the model fit satisfies

$$\chi^2 = \sum e_i^2$$

#### Some SAS Codes: Reading Grouped Data

```
data grouped;
input width cases satell;
cards;
 22.69 14
 23.84 14
           4
 24.77 28
           17
 25.84 39
            21
 26.79 22
           15
 27.74 24
           20
 28.67
        18
           15
 30.41
        14
           14
run;
proc format;
 value width 22.69='<=23.25' 23.84='23.25-24.25'
 24.77='24.25-25.25' 25.84='25.25-26.25'
 26.79='26.25-27.25' 27.74='27.25-28.25' 28.67='28.25-
 29.25' 30.41='>29.25';
run;
```

#### Fitting The Null Model

```
data grouped;
  set grouped;
  id = _n_;
run;
proc logistic data = grouped desc noprint;
  model satell/cases = ;
  output out=temp1 reschi=pearsona p=pi_hata;
run;
data temp1;
  set temp1;
  keep id Fitted_yesa pearsona;
  Fitted_yesa= pi_hata*cases;
run;
```

# Compute Residuals

```
proc logistic data = grouped desc;
  model satell/cases= width /influence;
  output out=temp2 reschi=pearson p=pi_hat h=h;
run;
data temp2;
  set temp2;
  Fitted_yes=pi_hat*cases;
  adjres = pearson/sqrt(1-h);
  keep pearson Fitted_yes adjres cases satell
width id pi_hat;
run;
```

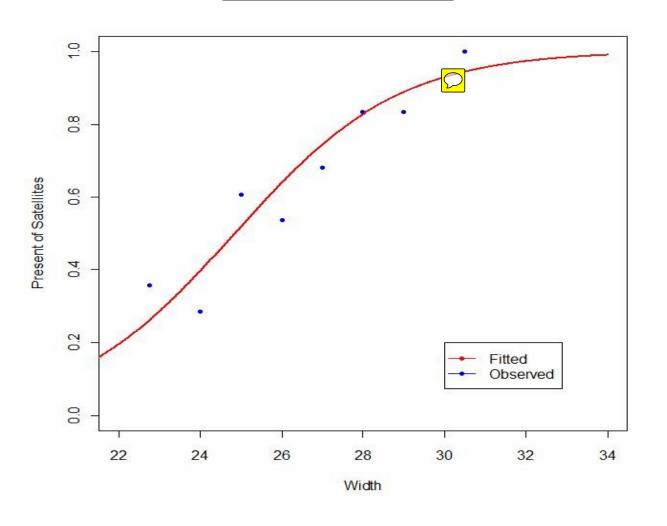
# Print A Nice Table

```
data combo;
  merge temp1 temp2;
  by id;
run;
proc print data = combo;
  format width width.;
  var width cases satell
  fitted_yesa pearsona Fitted_yes
  pearson adjres;
run;
```

# **Partial Output**

				${ t Fitted}_{ t }$		${ t Fitted}_{ t }$		
Obs	width	cases	satell	yesa	pearsona	yes	pearson	adjres
1	<=23.25	14	5	8.9827	-2.21972	3.8473	0.69012	0.85323
2	23.25-24.25	14	4	8.9827	-2.77706	5.4975	-0.81957	-0.93058
3	24.25-25.25	28	17	17.9653	-0.38043	13.9724	1.14434	1.34923
4	25.25-26.25	39	21	25.0231	-1.34344	24.2136	-1.06063	-1.24356
5	26.25-27.25	22	15	14.1156	0.39321	15.7962	-0.37724	-0.41511
6	27.25-28.25	24	20	15.3988	1.95862	19.1604	0.42716	0.49492
7	28.25-29.25	18	15	11.5491	1.69621	15.4644	-0.31464	-0.36050
8	>29.25	14	14	8.9827	2.79639	13.0469	1.01131	1.13725





#### Measures of Influence: Leverage

- Some observations may have much *influence* in determining parameter estimates.
- The fit could be quite different if they were deleted.
- Several measures describe various aspects of influence.
- Leverage the diagonal elements of the so-called  $hat\ matrix$ .
- Roughly, the *hat matrix* is a matrix that, when applied to the *standardized sample logits*, yields the *standardized predicted logit* values for the models.

# Influence Measures: Leverage

- The greater an observation's leverage, the greater its potential influence.
- Let **X** be the design matrix, whose *i*-th row contains the values of the explanatory variables for the *i*-th observation  $(1 \ x_{i1} \cdots x_{ip})$ .
- Let **W** be the diagonal matrix with its *i*-th diagonal entry equals  $w_i = n_i \hat{\pi}_i (1 \hat{\pi}_i)$
- Then the hat matrix,  $\mathbf{H}$  is defined as  $\mathbf{H} = \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2} \ \ \bigcirc$

# Other Influence Measures

- Coefficient Change For each parameter in the model, the change in the parameter estimate when the observation is deleted. This change, divided by its standard error, is called *Dfbeta*.
- Confidence Interval Change A measure of the change in a joint confidence interval for the parameters produced by deleting the observation. This confidence interval displacement diagnostic is denoted by C.
- Test Statistic Change The change in  $\chi^2$  or  $G^2$  goodness-of-fit statistics, when the observation is deleted.

# Influence Measures: Adjusted Residuals

- For each measure, the larger the value, the greater the observation's influence.
- One can also use leverage values to construct an adjustment to the Pearson residual  $e_i$  that is slightly larger in absolute value and does have an approximate standard normal distribution when the model holds.

# SAS Codes: Fit The Null Model

```
proc logistic data = grouped desc
  noprint;
  model satell/cases= ;
  output out=temp1
  difchisq=Pearson_diffa
  difdev=Likelihood_ratio_diffa;
run;
data temp1;
  set temp1;
  keep id Pearson_diffa
  Likelihood_ratio_diffa;
run;
```

#### SAS Codes: Fit With Width

```
proc logistic data = grouped desc noprint;
  model satell/cases = width ;
  output out=temp2 difchisq=Pearson_diff
difdev=Likelihood_ratio_diff c=c
  dfbetas=dfbeta_int dfbeta_width;
run;
data temp2;
  set temp2;
keep id width Pearson_diff
  Likelihood_ratio_diff c dfbeta_width;
run;
```

# SAS Codes: Print Diagnostics

```
data combo;
  merge temp2 temp1;
  by id;
run;
proc print data = combo;
  format width width.;
  var width dfbeta_width c Pearson_diff
  Likelihood_ratio_diff Pearson_diffa
  Likelihood_ratio_diffa;
run;
```

# Output

	$dfbeta_{-}$		Pearson_	$Likelihood_{-}$	Pearson_	$Likelihood_{-}$
width	width	C	diff	ratio_diff	diffa	ratio_diffa
<=23.25	-0.54098	0.38481	0.72800	0.70315	5.36099	5.0931
23.25-24.25	0.37397	0.25049	0.86599	0.89483	8.39114	8.0007
24.25-25.25	-0.42952	0.71024	1.82042	1.83044	0.17268	0.1708
25.25-26.25	-0.01498	0.57945	1.54644	1.52094	2.33013	2.2704
26.25-27.25	-0.09352	0.03633	0.17232	0.16890	0.17714	0.1799
27.25-28.25	0.21486	0.08387	0.24494	0.25366	4.45410	4.9507
28.25-29.25	-0.17208	0.04064	0.12996	0.12534	3.21126	3.5837
>29.25	0.54688	0.34220	1.29335	2.24483	8.50836	13.1139

# **Qualitative Predictors**

- Suppose the binary response Y has two binary predictors X and Z.
- For the  $2 \times 2 \times 2$  contingency table, the model:  $\operatorname{logit}(\pi) = \alpha + \beta_1 x + \beta_2 z$  has separate main effects for the two predictors and no interaction effect.
- The variables X and Z in this model are  $dummy \ variables$  that indicates categories for the predictors.

# Coefficient Interpretations

• At a fixed level z of Z, the effect on the logit of changing from x = 0 to x = 1 is

$$[\alpha + \beta_1 \times 1 + \beta_2 z] - [\alpha + \beta_1 \times 0 + \beta_2 z] = \beta_1$$

- It equals the log odds ratio between X and Y at Z = z.
- Thus  $exp(\beta_1)$  describes the conditional odds ratio between X and Y.
- The lack of interaction term in this model implies that the model satisfies the *homogeneous association*.

# Conditional Independence

- Conditional independence between X and Y, controlling for Z implies  $\beta_1 = 0$ .
- The simpler model  $logit(\pi) = \alpha + \beta_2 z$  then applies to the three way model.
- One can test whether  $\beta_1 = 0$  using a Wald statistic or a likelihood ratio statistic comparing the two models.

# Example: Effects of AZT

		Response		
Race	AZT Use	Yes	No	
White	Yes	14	93	
	No	32	81	
Black	Yes	11	52	
	No	12	43	

- It is believed that use of AZT slows down the development of AIDS symptoms.
- X = AZT Treatment
- Z = Race
- Y = Development of AIDS symptoms

#### Example: Continued ···

- The MLE of  $\beta_1$  (the effect of AZT) is = -0.720 (ASE = 0.279).
- The estimated *odds ratio* between immediate AZT use and development of AIDS symptoms equals  $\exp(-0.720) = 0.49$ .
- To test for conditional independence,  $H_0: \beta_1 = 0$ .
  - The LR statistic = 6.9 with d.f. = 1, p-value = 0.01.
  - The Wald Test statistic =  $(-0.720/0.279)^2 = 6.6$ .
- Goodness of fit statistics,  $\chi^2 = 1.4$  and  $G^2 = 1.4$  with residual d.f. = 4 3 = 1 suggests the model fits adequately.

#### SAS Codes

```
data aids;
input race $ azt $ yes no @@;
cases = yes + no;
datalines;
white y 14 93 white n 32 81
black y 11 52 black n 12 43
;
run;
proc genmod data=aids order=data;
class race azt;
model yes/cases = race azt / dist=bin link=logit
  obstats type3;
run;
```

Class Levels Values

race 2 white black

azt 2 y n

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	1.3835	1.3835
Scaled Deviance	1	1.3835	1.3835
Pearson Chi-Square	1	1.3910	1.3910
Scaled Pearson X2	1	1.3910	1.3910
Log Likelihood		-167.5756	

#### Analysis Of Parameter Estimates

			Standard	Wald 95%	Confidence	Chi-	
Parameter	DF	Estimate	Error	Lin	nits	Square	Pr>ChiSq
Intercept	1	-1.0736	0.2629	-1.5889	-0.5582	16.6	<.0001
race white	1	0.0555	0.2886	-0.5102	0.6212	0.04	0.8476
race black	0	0.0000	0.0000	0.0000	0.0000		
azt y	1	-0.7195	0.2790	-1.2662	-0.1727	6.65	0.0099
azt n	0	0.0000	0.0000	0.0000	0.0000		
Scale	0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis
Chi-

Source	DF	Square	$\mathtt{Pr} > \mathtt{ChiSq}$
race	1	0.04	0.8473
azt	1	6.87	0.0088

#### **ANOVA Type Representations**

- A factor having two levels requires only a single dummy variable.
- A factor having I levels requires I-1 dummy variables.
- The model formula  $logit(\pi_{ik}) = \alpha + \beta_i^X + \beta_k^Z$  represents the effects of X through parameters  $\{\beta_i^X\}$  and the effects of Z through parameters  $\{\beta_k^Z\}$ .
- This model for applies to any number of levels of X and Z.

#### **Notes**

- Each factor has as many parameters as it has levels, but one is redundant.
- For instance, if X has I levels, it has I-1 non-redundant parameters.
- $\beta_i^X$  denotes the effects on the logit of being classified in level i of X.
- Conditional independence between X and Y, given Z, corresponds to  $\beta_1^X = \beta_2^X = \cdots = \beta_I^X$

## Redundancy In Parameters

- To account for the redundancy in parameters, one can set the parameter for the last category to be zero.
- An analogous approach is to set the parameter for the first category to be zero.
- Alternatively, one can impose the restriction  $\beta_1^X + \beta_2^X + \dots + \beta_I^X = 0$

#### Some Codes

```
data aids1;
input race1 azt1 symptoms freq race2 azt2 race3 azt3;
datalines;
               1 1 14 0 0
            1
                  0 93 0 0
            1 1
            1
               0
                  1
                     32
                        0 1
                              1 -1
                        0 1 1 -1
            1 0
                    81
                  0
            0
                     11
                        1 0 -1
               1
                  0
                     52 1 0
            0
                             -1
                        1 1 -1 -1
                     12
            0
            0
               0 0 43
                        1 1
                              -1 -1
run;
```

#### Notes

- For any coding schemes, the differences  $\beta_1^X \beta_2^X$  and  $\beta_1^Z \beta_2^Z$  are identical and represent the conditional log odds ratios of X and Z with the response given the other variable.
- The estimate of a parameter for a single category of a factor is irrelevant; different ways of handling redundancy may result in different parameter estimates.
- An estimate makes sense only by comparison with one for another category.
- Similarly, different coding schemes yield the same estimated probability.

#### Logit Model for $2 \times 2 \times K$ Tables

- Consider X to be binary and Z is a control variable with K levels.
- In the model  $logit(\pi_{ik}) = \alpha + \beta_i^X + \beta_k^Z$  conditional independence exists between X and Y controlling for Z, if  $\beta_1^X = \beta_2^X$ .
- In such a case, common odds ratio  $\exp(\beta_1^X \beta_2^X)$  for the K partial tables equal 1.
- The CMH statistic used earlier is the efficient score statistic for testing X-Y conditional independence in this model.
- The ML estimate of the common odds ratio  $\exp(\beta_1^X \beta_2^X)$  is an alternative to the Mantel-Haenszel estimator.

## Multiple Logistic Regression

- Denote a set of k predictors for a binary response Y by  $X_1, X_2, \dots, Xk$ .
- Model for the logit of the probability  $\pi$  that Y = 1 generalizes to  $\operatorname{logit}(\pi(x)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$
- The parameters  $\beta_i$  refers to the effect of  $X_i$  on the log odds that Y = 1, controlling for other  $X_i$ s.
- e.g.  $\exp(\beta_i)$  is the multiplicative effect on the odds of a 1-unit increase in  $X_i$ , at fixed levels of other  $X_i$ s.

## Horseshoe Crab Example

- $\bullet$  Consider the female crab's colrwidth and color as predictors.
- Color has five categories: light, medium light, medium, medium dark, dark.
- The sample contained no light crabs, so our model uses only the other four categories.

# Example: The Model

- Let x denotes the width and we introduce three dummy variables for color:
  - $-c_1 = 1$  for medium light color, and 0 otherwise
  - $-c_2 = 1$  for medium color, and 0 otherwise
  - $-c_3=1$  for medium dark color, and 0 otherwise
- The crab color is dark, when  $c_1 = c_2 = c_3 = 0$ .
- The model is:

$$logit(\pi) + \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x$$

#### **SAS Codes**

```
%inc 'd:\class\st3241\crab.sas';
data dummy;
set crab;
c1 = 0;
if color=1 then c1=1;
c2 = 0;
if color=2 then c2=1;
c3 = 0;
if color=3 then c3=1;
output;
run;
```

## **SAS Codes**

```
proc logistic data = dummy
  desc;
  model y = c1 c2 c3 width;
  output out=temp p=pi_hat;
run;
```

#### Model Fit Statistics

In	t	e	r	C	6	b	t
	_	_	_	_	•	_	_

	Intercept	and
Criterion	Only	Covariates
AIC	227.759	197.457
SC	230.912	213.223
-2 Log L	225.759	187.457

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	$\mathtt{Pr}>\mathtt{ChiSq}$
Likelihood Ratio	38.3015	4	<.0001
Score	34.3384	4	<.0001
Wald	27.6788	4	<.0001

## Analysis of Maximum Likelihood Estimates

			Standard	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr> ChiSq
Intercept	1	-12.7151	2.7618	21.1965	<.0001
c1	1	1.3299	0.8525	2.4335	0.1188
c2	1	1.4023	0.5484	6.5380	0.0106
<b>c</b> 3	1	1.1061	0.5921	3.4901	0.0617
width	1	0.4680	0.1055	19.6573	<.0001

#### Odds Ratio Estimates

	Point	95% <b>Wa</b>	ld
Effect	Estimate	Confidence	Limits
c1	3.781	0.711	20.102
<b>c</b> 2	4.065	1.387	11.909
<b>c</b> 3	3.023	0.947	9.646
width	1.597	1.298	1.964

## **Example: Discussions**

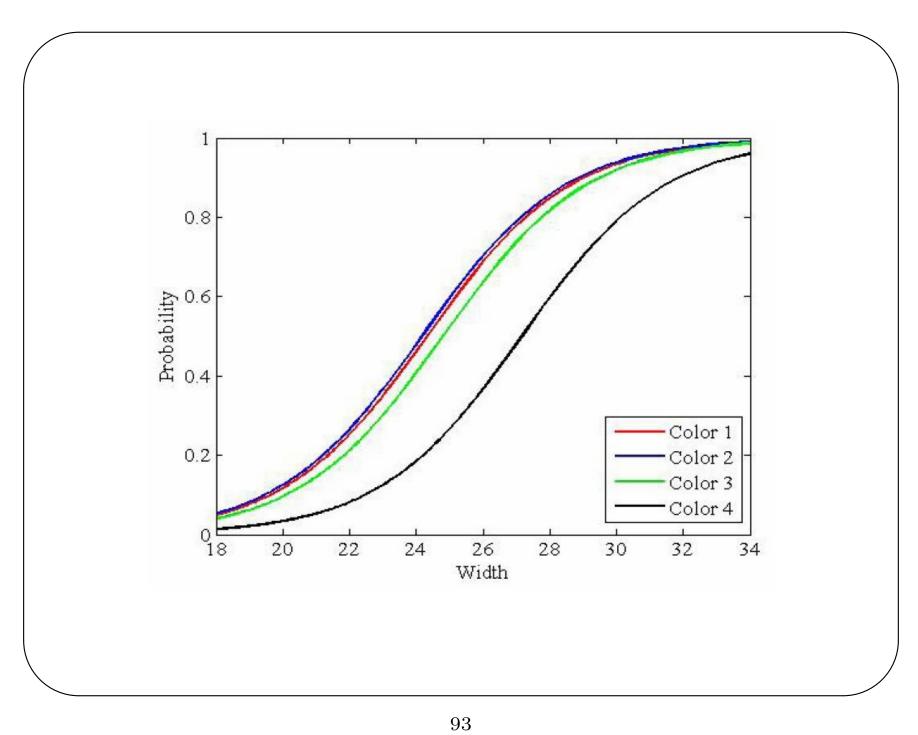
• For dark crabs,  $c_1 = c_2 = c_3 = 0$ , and the prediction equation is:

$$logit(\hat{\pi}) = -12.715 + 0.468x$$

• For medium light crabs,  $c_1 = 1$  and the prediction equation is:

$$logit(\hat{\pi}) = (-12.715 + 1.330) + 0.468x = -11.385 + 0.468x$$

• The model assumes a lack of interaction between color and width in their effects on the response.



#### Example: Predicted Probabilities

• For a medium light crab of width 26.3 cm, the predicted probability of having a satellite is

$$\frac{\exp(-11.385 + 0.468 \times 26.3)}{1 + \exp(-11.385 + 0.468 \times 26.3)} = 0.715$$

• A dark crab of the same width has predicted probability

$$\frac{\exp(-12.715 + 0.468 \times 26.3)}{1 + \exp(-12.715 + 0.468 \times 26.3)} = 0.399$$

#### Model Comparison

- One can use the likelihood-ratio method to test hypotheses about parameters in multiple logistic regression models.
- To test whether color makes a significant contribution, we test  $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ .
- We compare the maximized log-likelihood  $L_1$  of the full model to the maximized log-likelihood  $L_0$  for the simpler model in which those parameters equal 0, using test statistic  $-2(L_0 L_1) = 7.0$ .
- The chi-squared d.f. = 3, the difference between the number of parameters in the two models.
- The p value = 0.07 provides slight evidence of a color effect.

```
proc logistic data = dummy desc;
  model y = c1 c2 c3 width;
run;
proc logistic data = dummy desc;
  model y=width;
run;
```

#### Model Fit Statistics

Intercept
-----------

	Intercept	and
Criterion	Only	Covariates
AIC	227.759	197.457
SC	230.912	213.223
-2 Log L	225.759	187.457

Intercept

#### Model Fit Statistics

Criterion

AIC

and	ntercept
Covariates	Only
198.453	227.759

230.912 SC 204.759 -2 Log L 225.759 194.453

Intercept

#### **Notes**

- More generally, one can compare maximized log-likelihoods for any pair of models such that one is a special case of the other.
- One such comparison checks whether the model requires interaction terms.
- The test analyzes whether a better fitting model results from adding the interaction of color and width to the model.

## **SAS Codes**

```
data dummy2;
  set dummy;
  cw1 = c1*width;
  cw2 = c2*width;
  cw3 = c3*width;

run;
proc logistic data = dummy2 desc;
  model y = c1 c2 c3 width cw1 cw2 cw3;
run;
```

#### Model Fit Statistics

T	nt	ar	CE	pt
_		CT	CE	:PL

	Intercept	and
Criterion	Only	Covariates
AIC	227.759	199.081
SC	230.912	224.307
-2 Log L	225.759	183.081

## Quantitative Treatment of Ordinal Predictors

- Color has a natural ordering of categories, from lightest to darkest.
- Color may have a linear effect, for a set of monotone scores assigned to its categories.
- Let us assign scores c = 1, 2, 3, 4 to the color categories and fit the model

$$logit(\pi) = \alpha + \beta_1 c + \beta_2 x$$

• The fitted model is:

$$logit(\hat{\pi}) = -10.071 - 0.509c + 0.458x$$

## Example: Continued

- At a given width, for every one-category increase in color darkness, the estimated odds of a satellite multiply by exp(-0.509) = 0.60.
- A LR test compares the fit of this model to the model with separate parameters for each color.
- The test statistic equals 1.7 based on d.f. = 2 with p value = 0.44.
- It suggests that the simpler model is permissible.

#### Model Selection: An Example

- The horseshoe crab data set has four predictors: color (4 categories), spine condition (3 categories), weight and width of the carapace shell.
- Let us fit a model containing main effects for the four predictors, treating color and spine condition as qualitative (factors).

## Example: SAS Codes

```
data main;
  set crab;
  c1 = 0;
  if color=1 then c1=1;
  c2 = 0;
  if color=2 then c2=1;
  c3 = 0;
  if color=3 then c3=1;
  spine1=0;
```

## Example: SAS Codes

```
if spine=1 then spine1=1;
spine2=0;
if spine=2 then spine2=1;
run;
proc logistic data = main desc;
model y = c1 c2 c3 spine1 spine2 width
weight;
test c1=c2=c3=0;
test spine1=spine2=0;
run;
```

.1	resting Global Nu.	II Hypotnesis:	BEI	rA=U
	Test	Chi-Square	DF	${ t Pr} > { t ChiSq}$
	Likelihood Ratio	40.5565	7	<.0001
	Score	36.3068	7	<.0001
	Wald	29.4763	7	0.0001

#### Analysis of Maximum Likelihood Estimates

			Standard	Wald	
Parameter	DF	Estimate	Error	Chi-Square	Pr> ChiSq
Intercept	1	-9.2734	3.8378	5.8386	0.0157
c1	1	1.6087	0.9355	2.9567	0.0855
<b>c</b> 2	1	1.5058	0.5667	7.0607	0.0079
<b>c</b> 3	1	1.1198	0.5933	3.5624	0.0591
spine1	1	-0.4003	0.5027	0.6340	0.4259
spine2	1	-0.4963	0.6292	0.6222	0.4302
width	1	0.2631	0.1953	1.8152	0.1779
weight	1	0.8258	0.7038	1.3765	0.2407

Linear Hypotheses Testing Results
Wald

Label Chi-Square DF Pr> ChiSq
Test 1 7.1610 3 0.0669
Test 2 1.0105 2 0.6034

#### Model Selection

- ullet A model with several predictors has the potential for multicollinearity
  - Strong correlations among predictors making it seem that no one variable is important when all others are in the model.
- The very small p-value for the overall test, yet the lack of significance of individual predictors is a warning signal for potential multicollinearity.
- There is a strong linear relationship between width and weight with a correlation of 0.887.
- It does not make much sense to analyze an effect of width while controlling for weight, since weight naturally increases as width does.

				Models	
Model	Predictors	Deviance	DF	Compared	Difference
(1)	C*S*W	170.44	152	-	-
(2)	C*S+C*W+S*W	173.68	155	(2)-(1)	$3.2(\mathrm{df}=3)$
(3a)	C*S+S*W	177.34	158	(3a)-(2)	$3.7(\mathrm{df} = 3)$
(3b)	C*W+S*W	181.56	161	(3b)-(2)	$7.9(\mathrm{df} = 6)$
(3c)	C*S+C*W	173.69	157	(3c)-(2)	$0.0(\mathrm{df} = 2)$
(4a)	C*W	181.64	163	(4a)-(3c)	$8.0(\mathrm{df} = 6)$
(4b)	C*S	177.61	160	(4b)-(3c)	$3.9(\mathrm{df}=3)$
(5)	C+S+W	186.61	166	(5)-(4b)	$9.0(\mathrm{df} = 6)$
(6a)	C+S	208.83	167	(6a)-(5)	$22.2(\mathrm{df} = 1)$
(6b)	S+W	194.42	169	(6b)-(5)	$7.8(\mathrm{df} = 3)$
(6c)	C+W	187.46	168	(6c)-(5)	$0.8(\mathrm{df}=2)$
(7a)	С	212.06	169	(7a)-(6c)	$24.5(\mathrm{df} = 1)$
(7b)	W	194.45	171	(7b)-(6c)	$7.0(\mathrm{df} = 3)$
(8)	C = dark + W	187.96	170	(8)-(6c)	$0.5(\mathrm{df}=2)$
(9)	None	225.76	172	(9)-(8)	$37.8(\mathrm{df}=2)$

#### **Backward Elimination**

- To select a model, we can use a backward elimination procedure, starting with a complex model and successively taking out the terms.
- At each stage, we eliminate the term in the model that has the largest p-value when we test that its parameters equal to zero.
- We test only the highest order terms for each variable.
- It is inappropriate to remove a main effect term if the model contains higher-order interactions involving that term.

# Some Other Topics of Interest • Sample size and power for logistic regression • Exact inference for logistic regression