Objectives:

- Motivation for the use of logistic regression for the analysis of binary response data.

- Review of simple linear regression and why it is inappropriate for binary response data.

- Curvilinear response model and the logit transformation.

- Beyond Traditional Statistical Methods Copyright 2000 D. Cook, P. Dixon, W. M. Duckworth, M. S. Kaiser, K. Koehler, W. Q. Meeker and W. R. Stephenson.

Objectives:

- The use of maximum likelihood methods to perform logistic regression.

- Assessing the fit of a logistic regression model.

- Determining the significance of explanatory variables.

Motivating Examples

- The Challenger disaster
- The sex of turtles
- BronchoPulmonary Dysplasia (BPD) in newborns
- College Mathematics placement/grades in a statistics class
- Credit card scoring/Market segmentation

The Challenger disaster

- On January 28, 1986 the space shuttle, *Challenger*, had a catastrophic failure due to burn through of an O-ring seal at a joint in one of the solid-fuel rocket boosters.

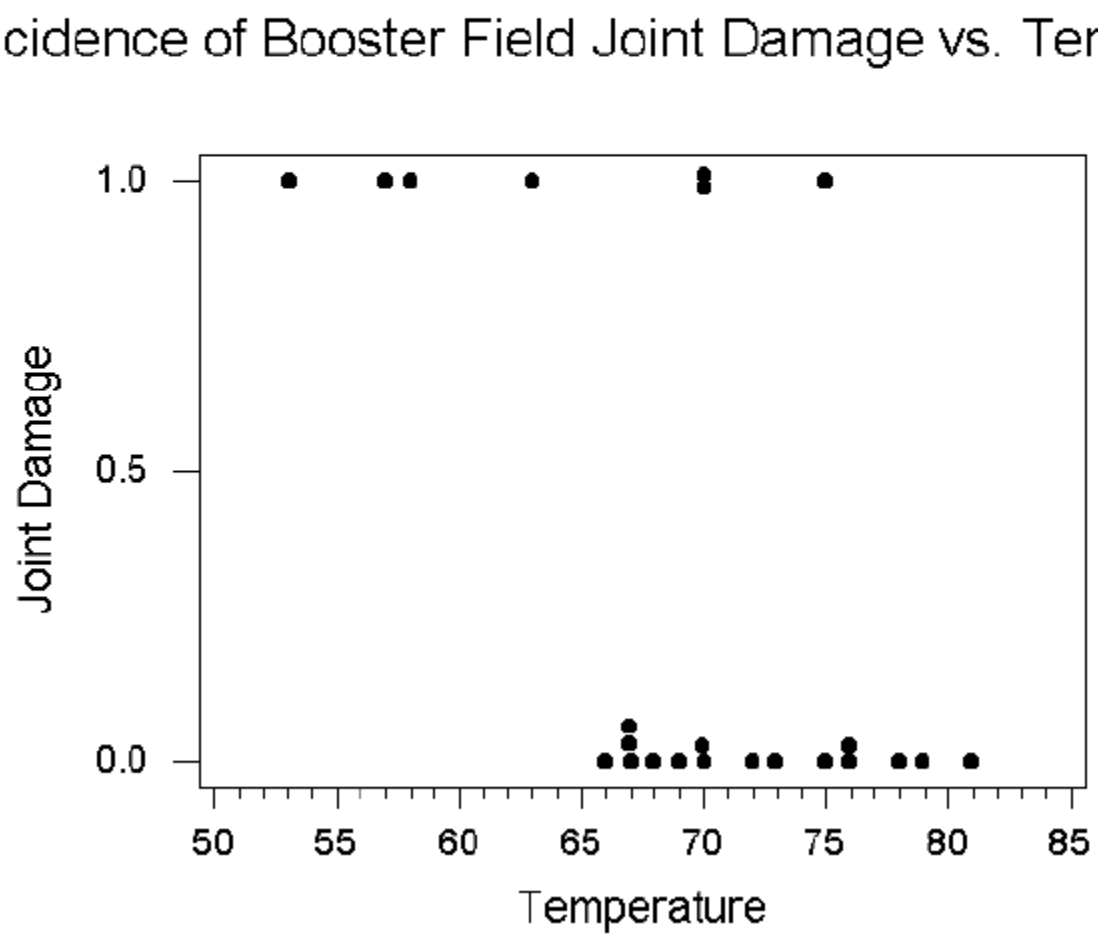
- Of 24 previous shuttle flights
  - 7 had incidents of damage to joints
  - 16 had no incidents of damage
  - 1 was unknown (booster not recovered after launch)

The Challenger disaster

- Could damage to solid-rocket booster field joints be related to cold weather at the time of launch?

- The following plot is derived from data from the Presidential Commission on the Space Shuttle *Challenger* Accident (1986). A 1 represents damage to field joints, and a 0 represents no damage.

The Challenger disaster



The Challenger disaster

- Incidence of joint damage
  - overall: $\frac{7}{23} = 30\%$
  - temperature $< 65^oF$: $\frac{4}{4} = 100\%$
  - temperature $\geq 65^oF$: $\frac{3}{19} = 16\%$
- Is there some way to predict the chance of booster field joint damage given the temperature at launch?

- library(faraway)
- orings
- oringsp=orings
- oringsp$b<-orings$Total
- oringsp$b[1]<-1
-

  glm.out<glm(b~temp,data=oringsp,family=
  binomial)

- glm.out<-glm(b~Temperature,data=oringsp,family=binomial)
- summary(glm.out)

- Call:
- glm(formula = b ~ Temperature, family = binomial, data = oringsp)

- Deviance Residuals:
- Min 1Q Median 3Q Max
- -1.0611 -0.7613 -0.3783 0.4524 2.2175

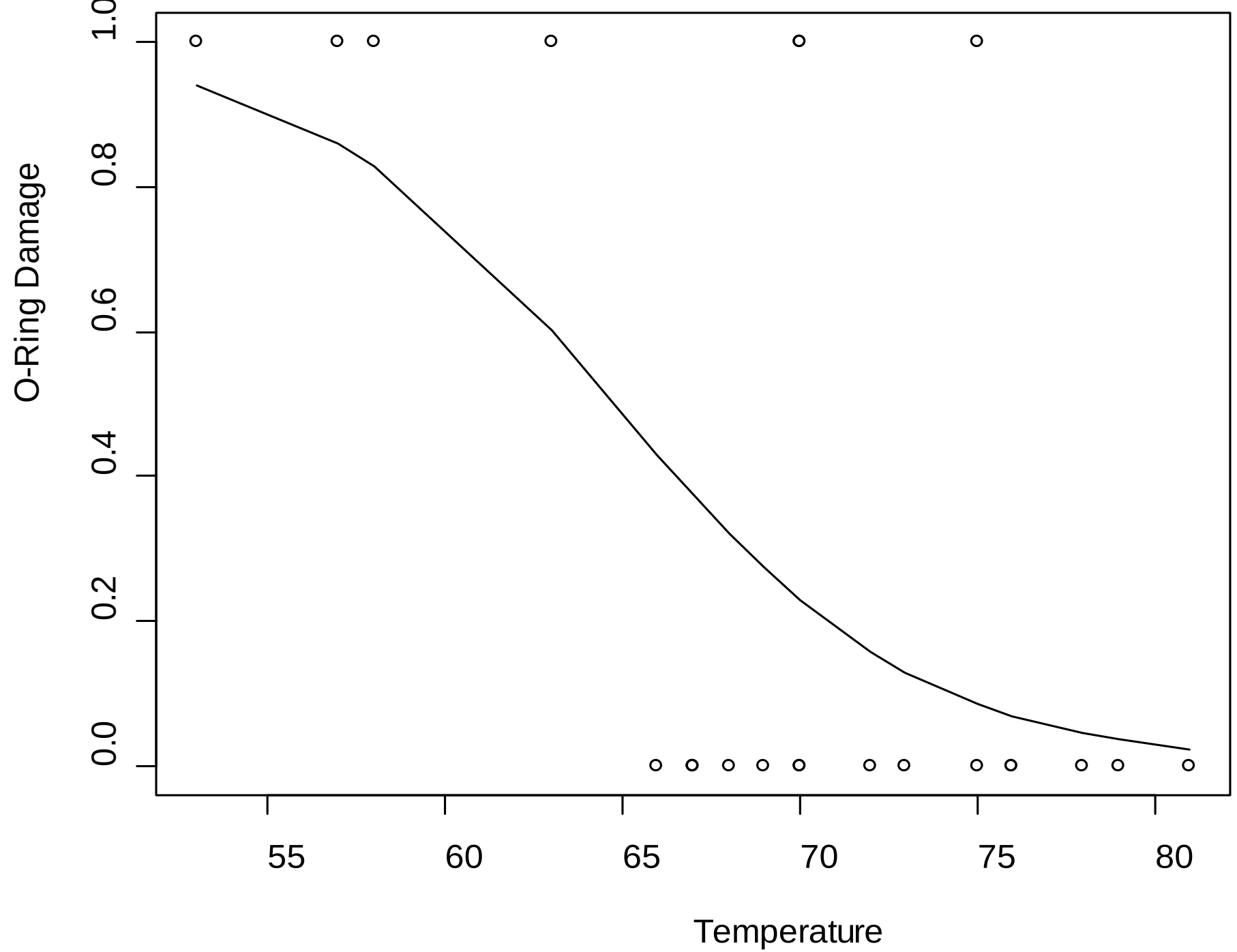
- Coefficients:
- Estimate Std. Error z value Pr(>|z|)
- (Intercept) 15.0429 7.3786 2.039 0.0415 *
- Temperature -0.2322 0.1082 -2.145 0.0320 *
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- (Dispersion parameter for binomial family taken to be 1)

- Null deviance: 28.267 on 22 degrees of freedom
- Residual deviance: 20.315 on 21 degrees of freedom
- AIC: 24.315

- Number of Fisher Scoring iterations: 5

- plot(b~Temperature,data=oringsp,ylab="O-Ring Damage")
- lines(orings$Temperature,predict(glm.out,type="response"))


O-Ring Damage
0.0
0.2
0.4
0.6
0.8
1.0
55
60
65
70
75
80
Temperature

The sex of turtles

- This example comes from a consulting project that Dr. Ken Koehler, ISU worked on.

- What determines the sex (male or female) of turtles?
  - Genetics?
  - Environment?

The sex of turtles

- The following experiment was conducted with turtle eggs collected in the wild.
  - Turtle eggs (all of one species) are collected in Illinois.
  - Several eggs are put into boxes.
  - Boxes are incubated at different temperatures.
  - When turtles hatch, their sex is determined.

The sex of turtles

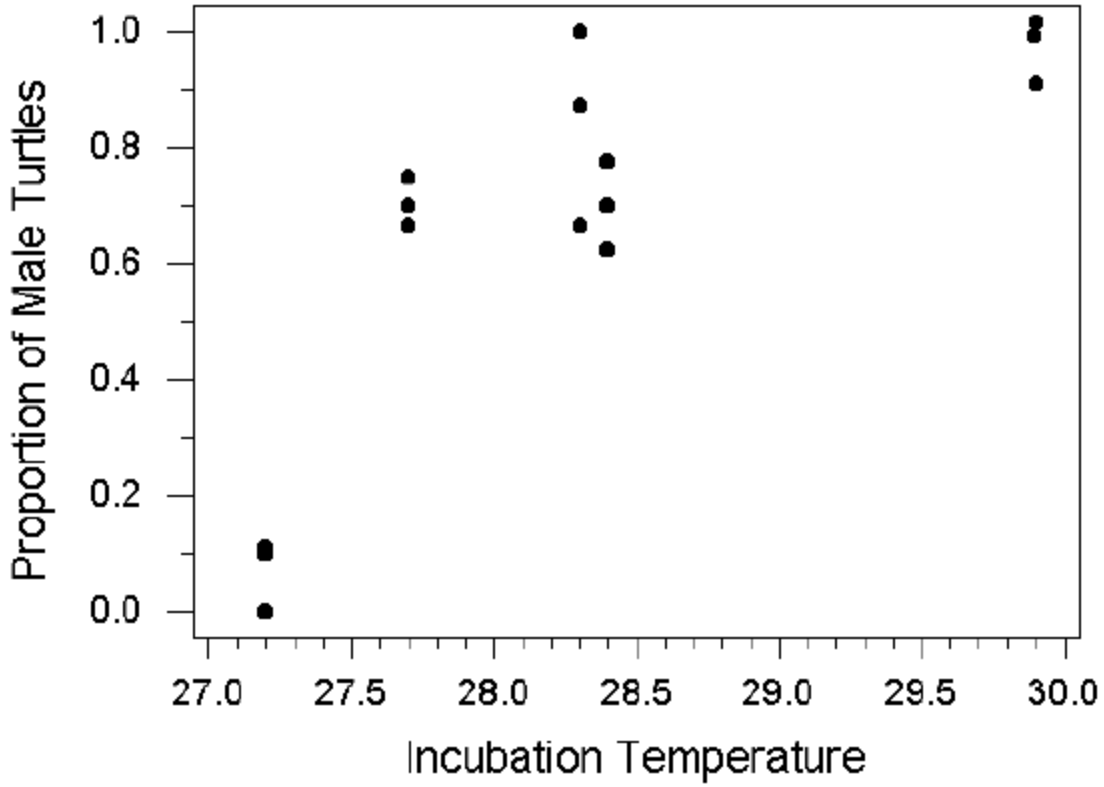
Temp	male	female	% male	Temp	male	female	% male
	1	9	10%		7	3	70%
27.2	0	8	0%	28.4	5	3	63%
	1	8	11%		7	2	78%
	7	3	70%		10	1	91
27.7	4	2	67%	29.9	8	0	100%
	6	2	75%		9	0	100%
	13	0	100%				
28.3	6	3	67%				
	7	1	88%				

The sex of turtles

- Proportion of male turtles for various temperature groups
  - overall: $\frac{91}{136} = 0.67$
  - temperature $< 27.5^oC$: $\frac{2}{25} = 0.08$
  - temperature $< 28.0^oC$: $\frac{19}{51} = 0.37$
  - temperature $< 28.5^oC$: $\frac{64}{108} = 0.59$
  - temperature $< 30.0^oC$: $\frac{91}{136} = 0.67$

The sex of turtles

Proportion of male turtles vs. incubation temperature



The sex of turtles

- Is there some way to predict the proportion of male turtles given the incubation temperature?

- At what temperature will you get a 50:50 split of males and females?

- data(turtle)
- temp male female
- 1 27.2 1 9
- 2 27.2 0 8
- 3 27.2 1 8
- 4 27.7 7 3
- 5 27.7 4 2
- 6 27.7 6 2
- 7 28.3 13 0
- 8 28.3 6 3
- 9 28.3 7 1
- 10 28.4 7 3
- 11 28.4 5 3
- 12 28.4 7 2
- 13 29.9 10 1
- 14 29.9 8 0
- 15 29.9 9 0

BPD in newborns

- This example comes from *Biostatistics Casebook,* by Rupert Miller, *et. al.*, (1980), John Wiley & Sons, New York.

- The data we will look at is a subset of a larger set of data presented in the casebook.

BPD in newborns

- Bronchopulmonary dysplasia (BPD) is a deterioration of the lung tissue.

- Evidence of BPD is given by scars on the lung as seen on a chest X-ray or from direct examination of lung tissue at death.

BPD in newborns

- Who gets BPD?
  - Those with respiratory distress syndrome (RDS) and oxygen therapy.
  - Those without RDS but who have gotten high levels of oxygen for some other reason.

BPD in newborns

- Response: BPD or no BPD (a binary response).

- Predictors: Hours of exposure to different levels of oxygen, O_2.
  - Low (21 to 39% O_2).
  - Medium (40 to 79% O_2).
  - High (80 to 100% O_2).

- The natural logarithm of the number of hours of exposure, lnL, lnM and lnH are used to model the response.

BPD in newborns

- Is there some way to predict the chance of developing BPD given the hours (or the natural logarithm of hours) of exposure to various levels of oxygen?

- Do the different levels of oxygen have differing effects on the chance of developing BPD?

Other examples

- College mathematics placement: Use ACT or SAT scores to predict whether individuals would receive a grade of C or better in an entry level mathematics course and so should be placed in a higher level mathematics course.

- Grades in a statistics course: Do things like interest in the course, feeling of pressure/stress and gender related to the grade (A, B, C, D, or F) one earns in a course?

Other examples

- Credit card scoring: Use various demographic and credit history variables to predict if individuals will be good or bad credit risks.

- Market segmentation: Use various demographic and purchasing information to predict if individuals will purchase from a catalog sent to their home.

Common Aspects

- All have a binary (or categorical) response:
  - damage/no damage.
  - male/female.
  - BPD/no BPD.
- All involve the idea of prediction of a chance, probability, proportion or percentage.
- Unlike other prediction situations, the response is bounded.

Logistic Regression

- Logistic regression is a statistical technique that can be used in binary response problems.

- It is different from ordinary least squares regression although there are similarities.

- It is important to recognize the similarities and differences between the two techniques (logistic regression and ordinary regression).

Simple Linear Regression:

- Review of ordinary least squares simple linear regression.
  - Data
  - Model
    - Structure on the means.
    - Error structure.

Simple Linear Regression:

- Data
  - Response, Y: numerical (continuous measurement).
  - Predictor, X: numerical (continuous measurement).
- Model
  - The mean response is a simple linear function of the predictor variable.
  - Individual responses vary around the mean.

Simple Linear Regression:

- Model:

$$Y_i = \beta_o + \beta_1 X_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Simple Linear Regression:

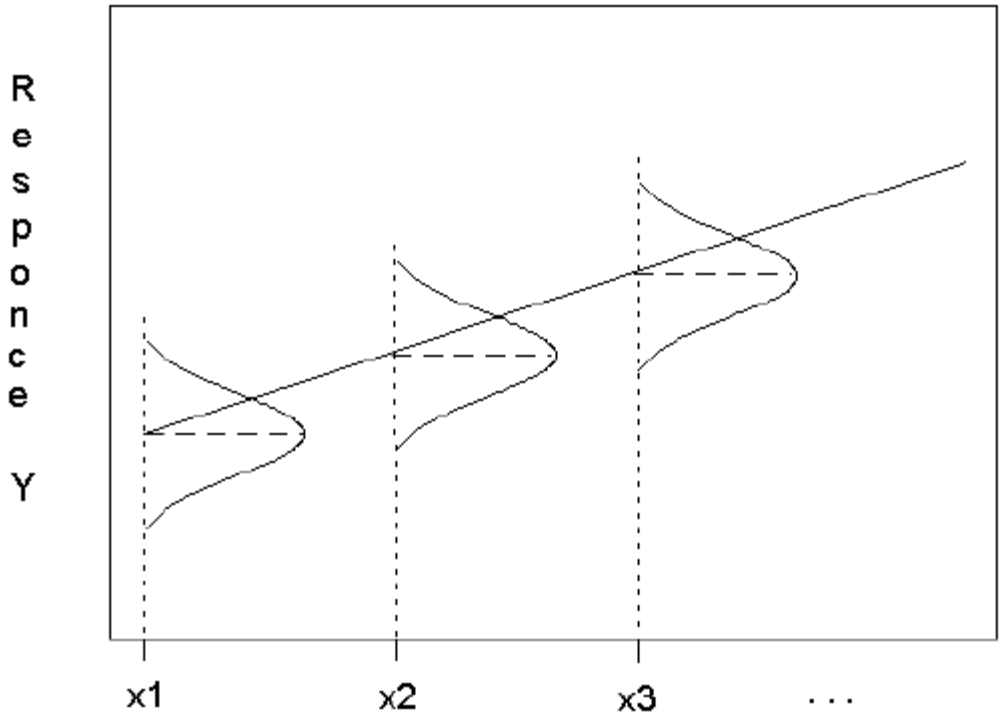
- Model:
  - Structure on the means

$$E(Y_i|X_i) = \beta_o + \beta_1 X_i$$

  - Error structure

$$\epsilon_i \sim N(0, \sigma^2)$$

Simple Linear Regression



Challenger disaster

- Binary response:
  - $Y_i = 1$ damage to field joint
  - $Y_i = 0$ no damage to field joint
- Probability:
  - Prob$(Y_i = 1) = \pi_i$
  - Prob$(Y_i = 0) = 1 - \pi_i$

Binary response

- In general

$$E(Y_i) = 0 * (1 - \pi_i) + 1 * \pi_i = \pi_i$$

- With predictor variable, X_i

$$E(Y_i|X_i) = \beta_o + \beta_1 X_i = \pi_i \quad (1)$$

Simple Linear Regression?

- Constraint on the response

$$0 \leq E(Y_i|X_i) = \pi_i \leq 1$$

- Non-constant variance

$$Var(\epsilon_i) = Var(Y_i) = \pi_i(1 - \pi_i)$$

  - the variance depends on the value of X_i.

Simple Linear Regression?

- Non-Normal error terms

$$\epsilon_i = Y_i - (\beta_o + \beta_1 X_i)$$

  - when $Y_i = 1$

$$\epsilon_i = 1 - (\beta_o + \beta_1 X_i)$$

  - when $Y_i = 0$

$$\epsilon_i = 0 - (\beta_o + \beta_1 X_i)$$

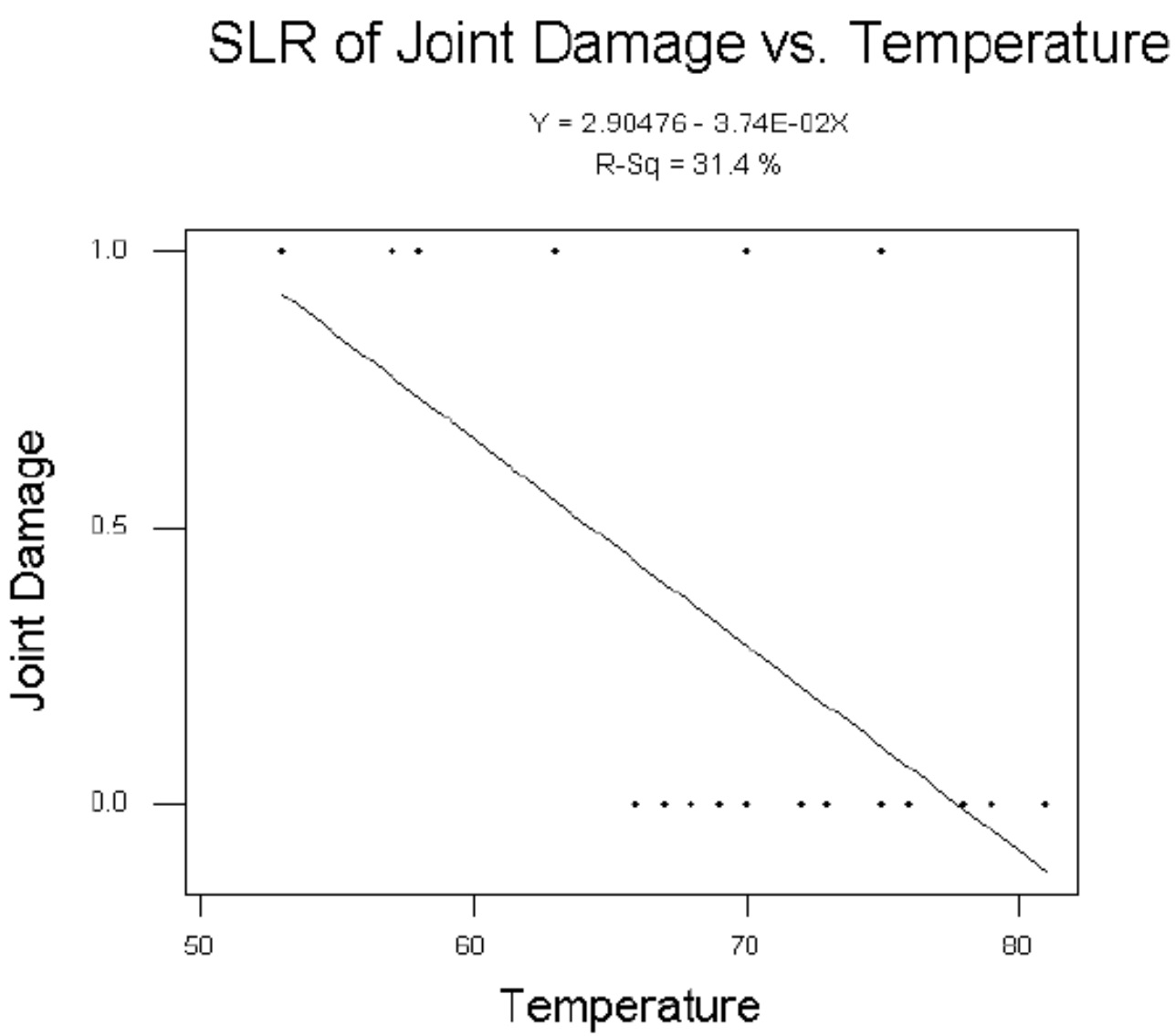
Use SLR anyway

- Challenger disaster
  - SLR prediction equation

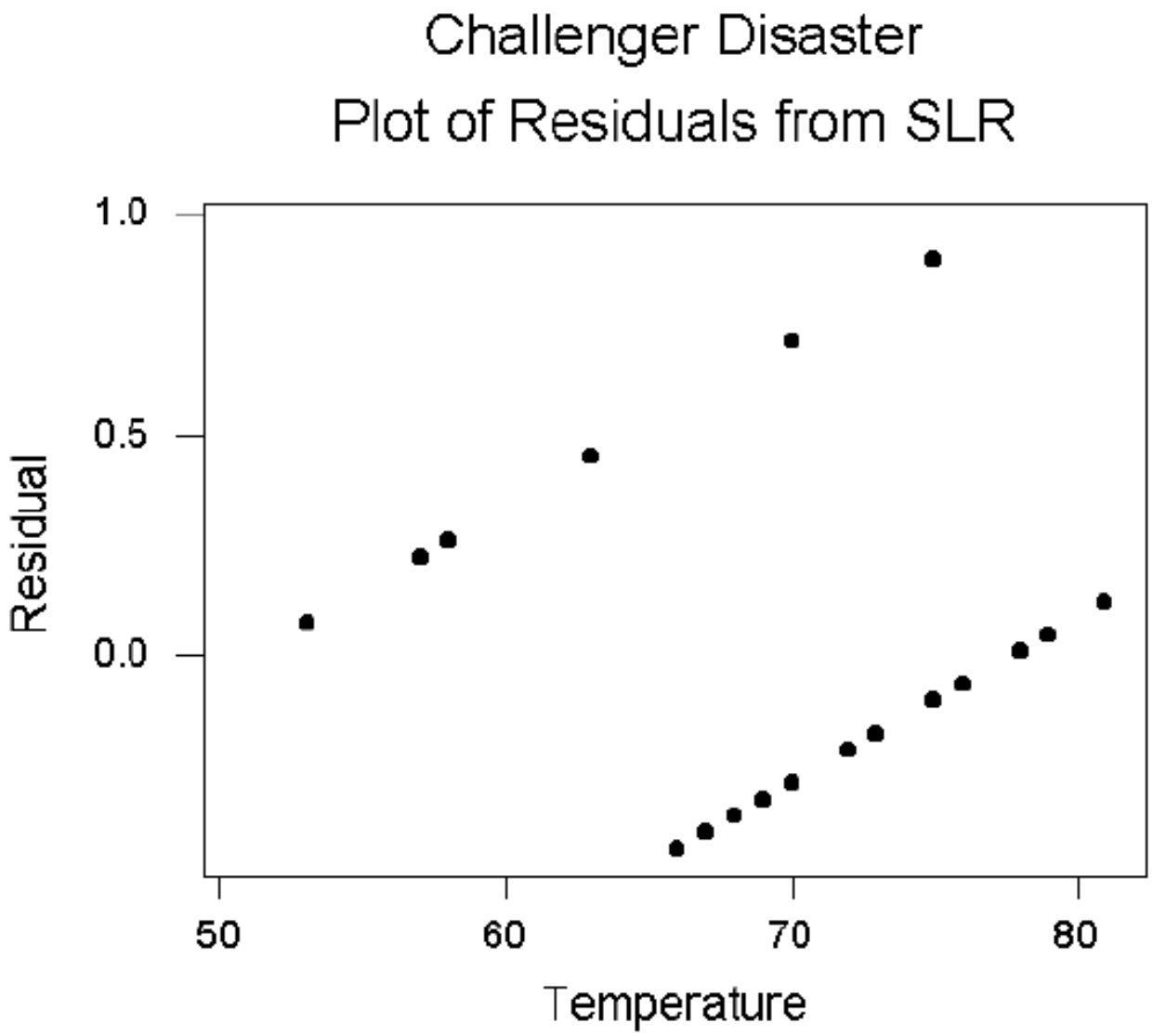
    $\widehat{Y} = 2.905 - 0.0374X$

    * for $X = 77$ $\quad \widehat{Y} = 0.025$
    * for $X = 65$ $\quad \widehat{Y} = 0.474$
    * for $X = 51$ $\quad \widehat{Y} = 0.998$

Fitted line plot



Plot of residuals



SLR anyway

- When we use ordinary least squares to fit a simple linear regression model we will get unbiased estimates of the regression parameters (intercept and slope).

- Since we are violating the equal variance assumption, the standard errors for these estimates will be larger.

- The unusual pattern in the residuals indicates that the simple linear model is not capturing the pattern in the response.

Weighted Least Squares

- To overcome the unequal variances, we can weight the observations by the inverse of the variance.

$$w_i = \frac{1}{\pi_i(1-\pi_i)}$$

- Since π_i is not known, we need to use estimated weights

$$\widehat{w}_i = \frac{1}{\widehat{Y}_i(1-\widehat{Y}_i)}$$

Weighted Least Squares

- Fit Ordinary Least Squares (simple linear model)
  - Obtain estimates, $\widehat{Y}_i$
  - If an estimate is less than 0 or more than 1, set it to 0.001 or 0.999, respectively.
  - Compute weights, $\widehat{w}_i$
- Fit Weighted Least Squares (simple linear model)

Challenger disaster

- Ordinary Least Squares (simple linear model)

Coefficient	Estimate	Std Error
Intercept	2.905	0.8421
Slope	−0.0374	0.0120

- Weighted Least Squares (simple linear model)

Coefficient	Estimate	Std Error
Intercept	2.344	0.5324
Slope	−0.0295	0.0067

Comments

- Weighted least squares reduces the standard errors of the estimated parameters (intercept and slope).

- Weighted least squares does not affect the lack of normality for the errors.

- Weighted least squares does not address the possibility that fitted responses can fall below zero or above one.

- A curvilinear response function is a more appropriate model than a simple linear relationship between predictor and response.

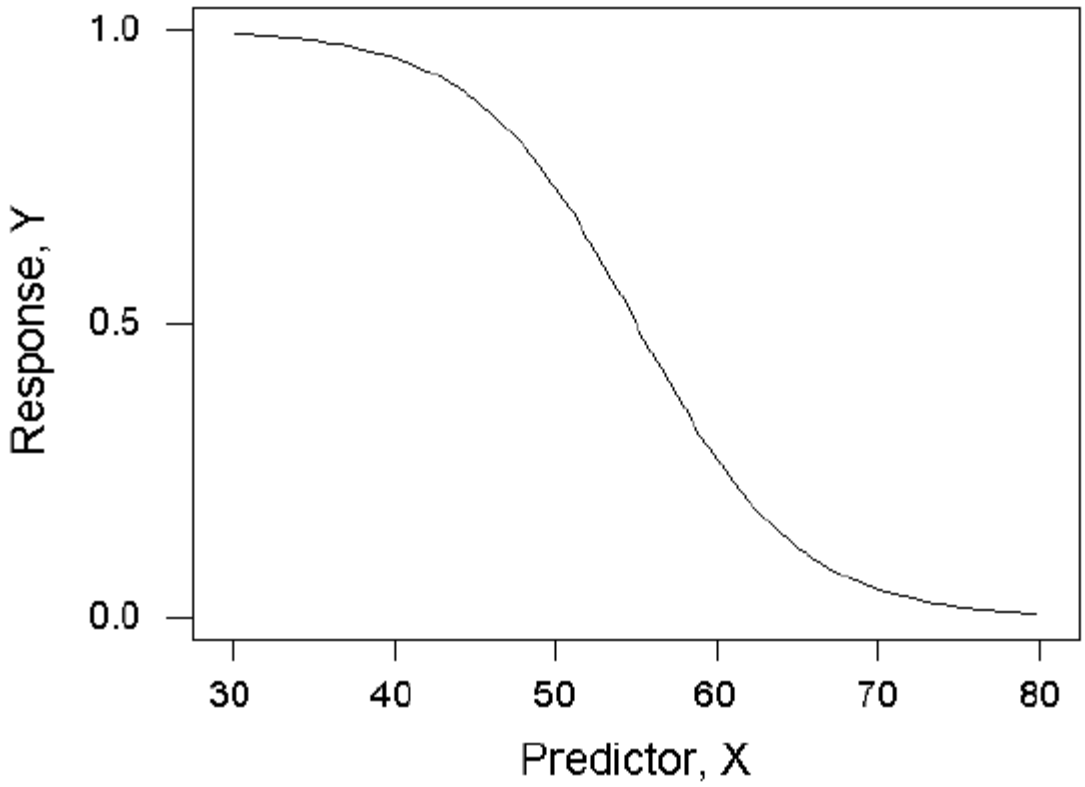
Curvilinear model

- When the response variable is binary, or a binomial proportion, the expected response is more appropriately modeled by some curved relationship with the predictor variable.

- One such curved relationship is given by the logistic model

$$E(Y_i|X_i) = \pi_i = \frac{e^{(\beta_o+\beta_1 X_i)}}{1 + e^{(\beta_o+\beta_1 X_i)}} \quad (2)$$

Logistic model



Logistic model

- The logit transformation.

$$\pi_i' = ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_o + \beta_1 X_i \quad (3)$$

- Estimate π_i by p_i, the observed proportion, and apply the logit transformation.

$$ln\left(\frac{p_i}{1-p_i}\right)$$

The sex of turtles

- Combined data

Temp	male	female	total	pmale, p_i
27.2	2	25	27	0.0741
27.7	17	7	24	0.7083
28.3	26	4	30	0.8667
28.4	19	8	27	0.7037
29.9	27	1	28	0.9643

The sex of turtles

- Ordinary least squares linear model on the raw proportions

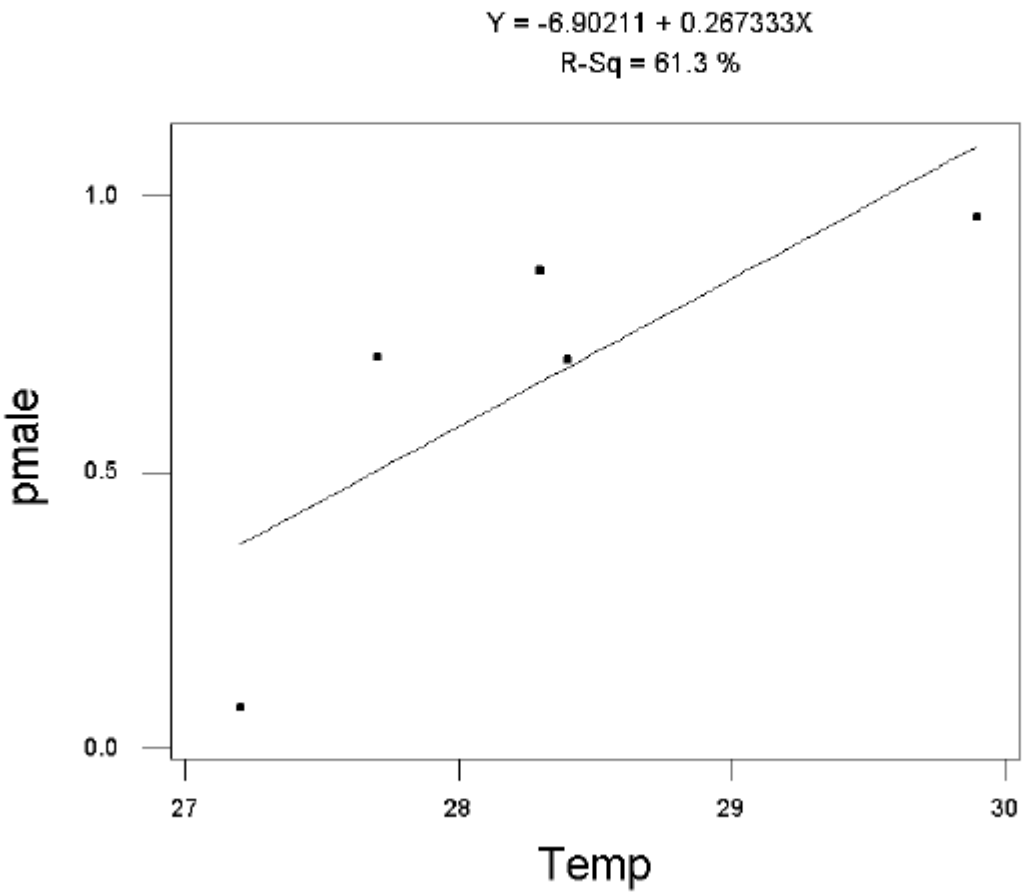
$$\widehat{\pi} = -6.902 + 0.2673 Temp \quad (4)$$

Temp	27.2	27.7	28.3	28.4	29.9
Fit prop., $\widehat{\pi}$	0.369	0.503	0.663	0.690	1.091

- Note that the ordinary least squares linear model on the raw proportions gives a fitted values that is above 1.

The sex of turtles

SLR of proportion male on incubation temperature



The sex of turtles

- Combined data

Temp	pmale, p_i	$ln\left(\frac{p_i}{1-p_i}\right)$
27.2	0.0741	−2.5257
27.7	0.7083	0.8873
28.3	0.8667	1.8718
28.4	0.7037	0.8650
29.9	0.9643	3.2958

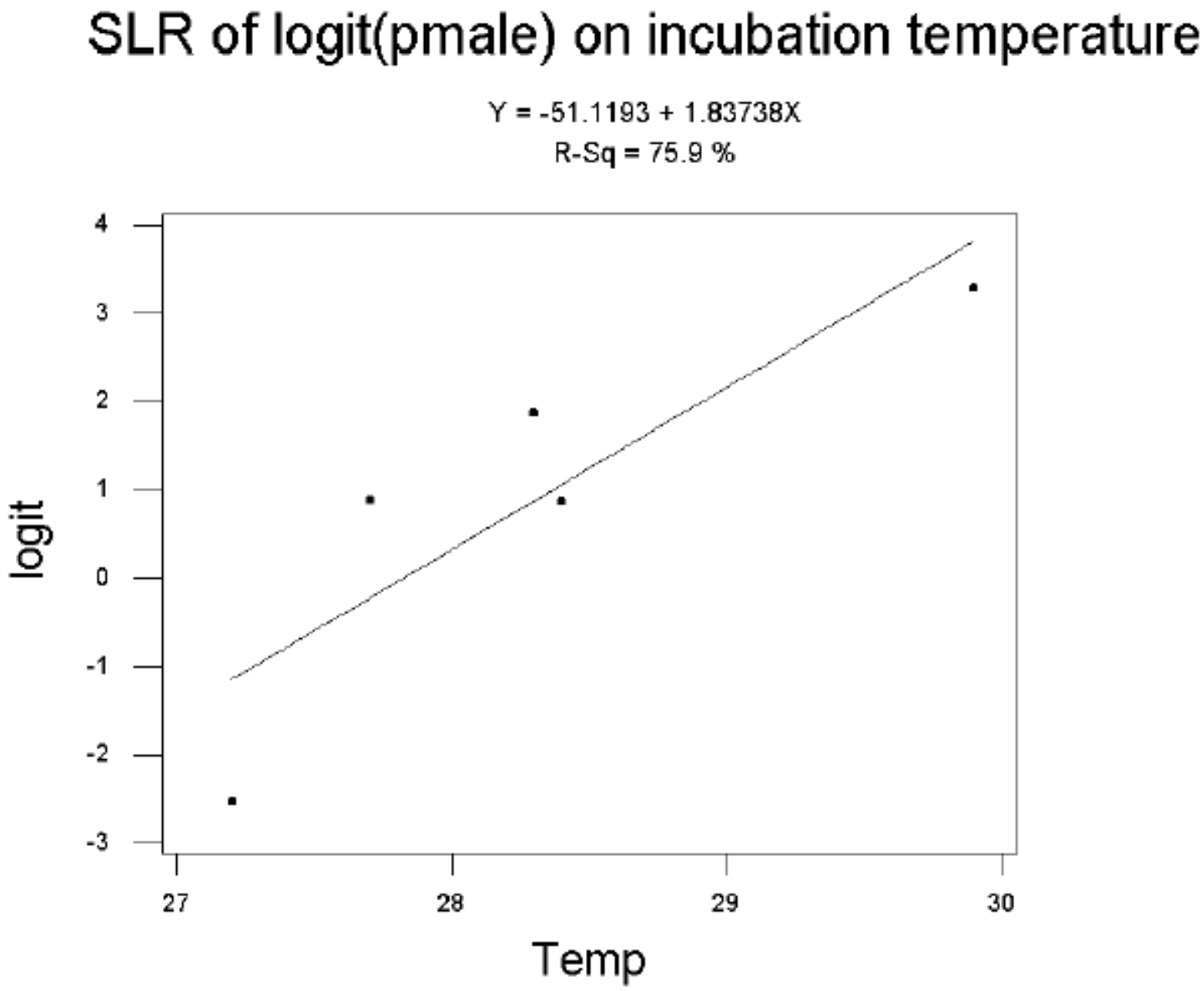
The sex of turtles

- Ordinary least squares on the logit transformed proportions

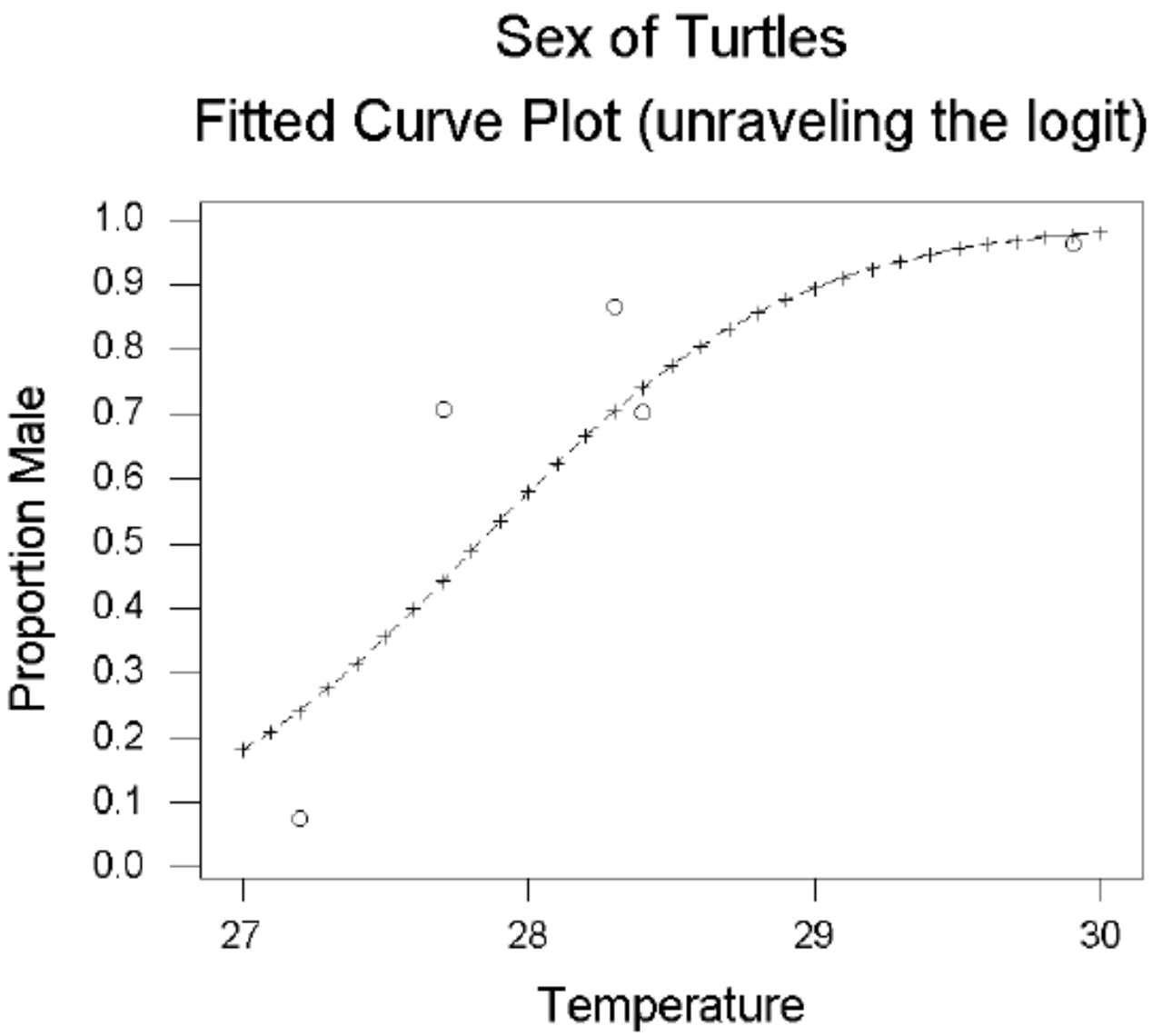
$$\widehat{\pi}' = -51.1116 + 1.8371 Temp \quad (5)$$

Temp	27.2	27.7	28.3	28.4	29.9
Fit logit, $\widehat{\pi}'$	−1.1420	−0.2334	0.8788	1.0626	3.8182
Fit prop., $\widehat{\pi}$	0.242	0.444	0.707	0.743	0.979

The sex of turtles



Linear model on logit



The sex of turtles

- Temperature to give a 50:50 split
  - Ordinary least squares linear model on proportions.
    - A temperature of 27.69^oC will give a predicted proportion male of 0.50.
  - Ordinary least squares linear model on the logit transformed proportions.
    - A temperature of 27.82^oC will give a predicted proportion male of 0.50.

Comments

- The logit transformation has adjusted for the curved nature of the response. With the linear model on the logit transformed proportions we will not get predicted values less than zero or greater than one.

- There is, however, still the problem of unequal variances and non-normal errors.

Comments

- By doing ordinary least squares we are trying to force binary response data into a familiar method of analysis.

- What we really need is a new way of looking at this problem and a new means of analyzing the data.

Likelihood

- The likelihood function is a function of the data and the parameters of the model. We maximize the likelihood by finding estimates of the model parameters that are most likely to give us the data.

- With binary response the form of the likelihood function is relatively simple.

Likelihood

- For the binary response

  - $Y_i = 1$ with probability π_i

  - $Y_i = 0$ with probability $1 - \pi_i$

- The general form of the likelihood function is

$$L((\beta_o, \beta_1); Data) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i} \quad (6)$$

Likelihood

- With a logistic model the probability, π_i, is a curvilinear function of the explanatory variable, X_i given by

$$\pi_i = \frac{e^{(\beta_o + \beta_1 X_i)}}{1 + e^{(\beta_o + \beta_1 X_i)}}$$

and

$$(1 - \pi_i) = \frac{1}{1 + e^{(\beta_o + \beta_1 X_i)}}$$

Likelihood

$$
\begin{aligned}
L((\beta_o, \beta_1); Data) &= \\
&= \prod_{i=1}^{n} \left(\frac{e^{(\beta_o + \beta_1 X_i)}}{1 + e^{(\beta_o + \beta_1 X_i)}} \right)^{Y_i} \left(\frac{1}{1 + e^{(\beta_o + \beta_1 X_i)}} \right)^{1 - Y_i} \\
&= \prod_{i=1}^{n} \frac{\left(e^{(\beta_o + \beta_1 X_i)} \right)^{Y_i}}{\left(1 + e^{(\beta_o + \beta_1 X_i)} \right)} \qquad (7)
\end{aligned}
$$

Log Likelihood

- It is often easier to work with the natural logarithm of the likelihood function, the log likelihood.

$$log\left[L((\beta_o, \beta_1); Data)\right] = \sum_{i=1}^{n} Y_i(\beta_o + \beta_1 X_i)$$

$$- \sum_{i=1}^{n} log\left[1 + e^{(\beta_o + \beta_1 X_i)}\right] \qquad (8)$$

Maximum Likelihood

- Choose β_o and β_1 so as to maximize the log likelihood. These choices will also maximize the likelihood.

- Similar to ordinary least squares, we will obtain two equations with two unknowns (β_o and β_1).

- Unlike ordinary least squares, the equations are not linear and so must be solved by iteration (start with initial values for β_o and β_1, evaluate the log likelihood, choose a new value for β_o or β_1 that reduces the log likelihood, repeat until the log likelihood does not change).

Splus

- General Linear Model
  - family=binomial
    - uses the binomial (binary response) likelihood function as given on slide 32.
  - link=logit
    - uses the logistic model as given on slide 33.

The sex of turtles

- Splus logistic regression on combined turtle data

$$\widehat{\pi}' = -61.3183 + 2.2110 Temp \quad (9)$$

Temp	27.2	27.7	28.3	28.4	29.9
Fit logit, $\widehat{\pi}'$	−1.1791	−0.0736	1.2530	1.4741	4.7906
Fit prop., $\widehat{\pi}$	0.235	0.482	0.778	0.814	0.992

- data(turtle)
- turtle$n<-turtle$male+turtle$female
- glm.turtle<-glm((male/n)~temp,weights=n,family=binomial,data=turtle)
- summary(glm.turtle)

- Call:
- glm(formula = (male/n) ~ temp, family = binomial, data = turtle,
- weights = n)

- Deviance Residuals:
- Min 1Q Median 3Q Max
- -2.0721 -1.0291 -0.2714 0.8087 2.5550

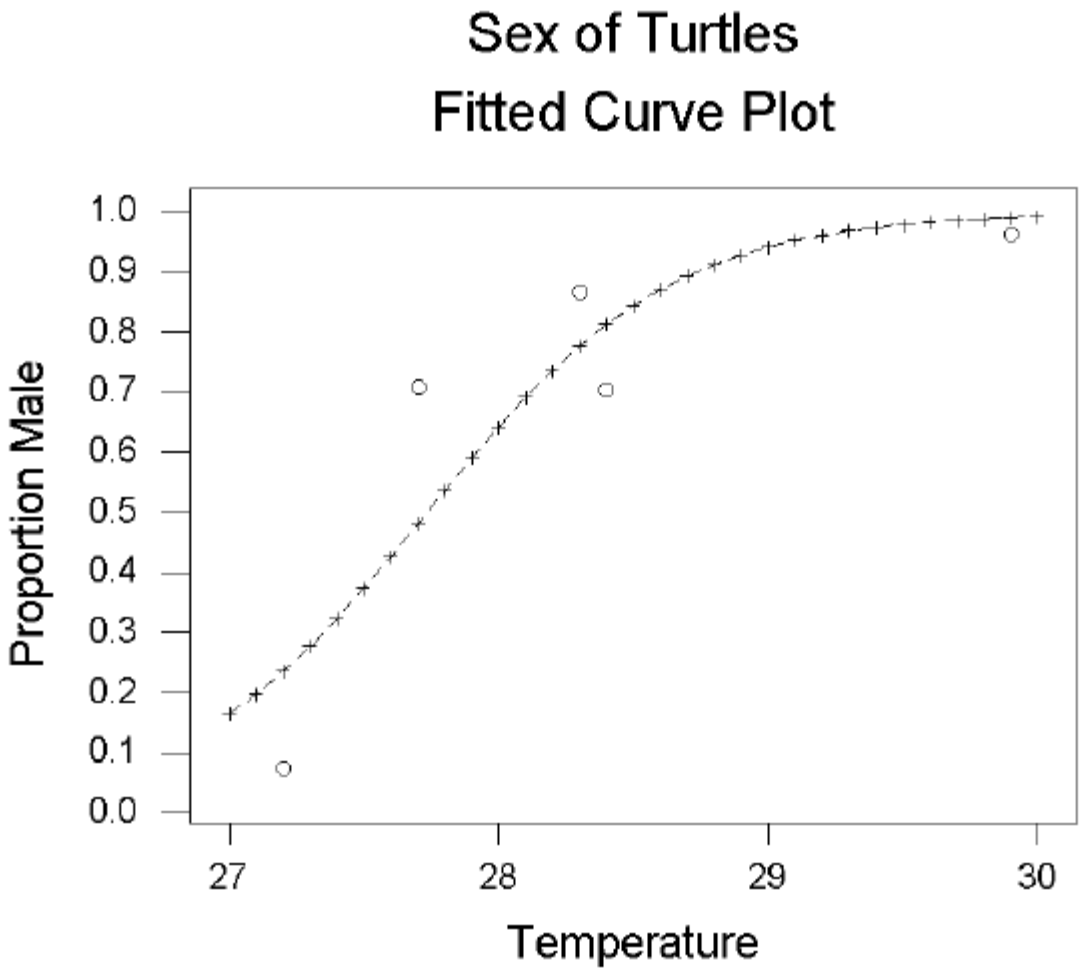
- Coefficients:
- Estimate Std. Error z value Pr(>|z|)
- (Intercept) -61.3183 12.0224 -5.100 3.39e-07 ***
- temp 2.2110 0.4309 5.132 2.87e-07 ***
- ---
- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- (Dispersion parameter for binomial family taken to be 1)

- Null deviance: 74.508 on 14 degrees of freedom
- Residual deviance: 24.942 on 13 degrees of freedom
- AIC: 53.836

- Number of Fisher Scoring iterations: 5

Splus: glm(binomial,logit)



Comments

- When p_i equals 0 or 1, use

$$p_i = \frac{1}{2n_i} \text{ or } p_i = 1 - \frac{1}{2n_i}$$

  respectively.

- Some computer programs may adjust all p_i by some small amount, *e.g.*

$$p_i + \frac{(0.5 - p_i)}{n_i}$$

Comments

- Different computer programs may use different adjustments, starting values, round off, and algorithms than Splus. Even within Splus there is more than one way to run glm. Below are the fits for various programs for the combined turtle data.

  - Splus: $\pi_i' = -61.31828 + 2.21103 X_i$
    or $\pi_i' = -61.318299 + 2.211031 X_i$

  - Minitab: $\pi_i' = -61.32 + 2.2110 X_i$

  - SAS: $\pi_i' = -61.3183 + 2.2110 X_i$

Interpretation of results

- The coefficients in a logistic regression are often difficult to interpret because the effect of increasing X by one unit varies depending on where X is. This is the essence of a nonlinear model.

- Consider first the interpretation of the odds:

$$\frac{\pi_i}{(1-\pi_i)}$$

If $\pi_i = 0.75$, then the odds of getting a male turtle are 3 to 1. That is, a male turtle is 3 times as likely as a female turtle.

Interpretation of results

- In logistic regression we model the log-odds. The predicted log-odds, $\widehat{\pi}_i'$ are given by the linear equation given in slide 38.

- The predicted odds are

$$e^{\widehat{\pi}_i'} = \frac{\widehat{\pi}_i}{1-\widehat{\pi}_i}$$

- If we increase X_i by 1 unit, we multiply the predicted odds by $e^{\widehat{\beta}_1}$.

Interpretation of results

- At 27^o C the predicted odds for a male turtle are 0.20, about 1 in 5. That is, it is 5 times more likely to get a female than a male at this temperature.

- At 28^o C the predicted odds for a male are $e^{2.2110} = 9.125$ times the odds at 27^o C, or 1.825. At 28^o C getting a male is almost twice as likely as getting a female.

- At 29^o C the predicted odds for a male are $e^{2.2110} = 9.125$ times the odds at 28^o C, or 16.65. At 29^o C getting a male is over 16 times more likely than getting a female.

Interpretation of results

- The intercept can be thought of as the predicted log-odds when X_i is zero. The anti-log of the intercept may have some meaning as a baseline odds, especially if zero is within the range of the data for the predictor variable, X.

- In the turtle example, all data comes from temperatures, values or X, between 27^o C and 30^o C. The values $X = 0$ is well outside the range of the data. In the turtle example, the intercept, or the anti-log of the intercept, has no practical interpretation.

Inference

- So far we have only looked at the fitting of a model by estimating parameters using maximum likelihood techniques.

- Estimates of model parameters are subject to variation.

- We must be able to quantify this variation in order to make inferences; tests of hypotheses and confidence intervals for model parameters.

Inference

- Testing hypotheses
  - Is there a statistically significant relationship between the predictor variable and the binary response?
  - Is the logistic model a good fit for the binary response data or is there a statistically significant lack of fit?
- Confidence intervals
  - Confidence intervals for the model parameters.
  - Confidence intervals for the predicted proportions.

A word of caution

- Inference techniques for logistic regression appear to be similar to, or at least analogous to, inference techniques for ordinary least squares regression with a linear model.

- Inference for logistic regression is based on asymptotic theory. That is, the inference techniques are approximate with the approximations getting better when you have larger amounts of data (larger sample sizes).

Inference for logistic regression

- Just as with ordinary least squares regression we need some means of determining the significance of the estimates of the model parameters. We also need a means of assessing the fit, or lack of fit, of the logistic model.

- Inference for logistic regression is often based on the deviance (also known as the residual deviance).

- The deviance is twice the log-likelihood ratio statistic.

Inference for logistic regression

- The deviance for a logistic model can be likened to the residual sum of squares in ordinary least squares regression for the linear model.

- The smaller the deviance the better the fit of the logistic model.

- A large value for the deviance is an indication that there is a significant lack of fit for the logistic model and some other model may be more appropriate.

Inference for logistic regression

- The deviance can be compared to a chi-square distribution, which approximates the distribution of the deviance.

- The degrees of freedom is determined by the number of observations and the number of parameters in the model that are estimated.

df = n - # parameters estimated

The sex of turtles

- The residual deviance for the logistic model fit to the combined turtle data is 14.863 on 3 degrees of freedom.

- The P-value, the chance that a χ^2 with 3 degrees of freedom exceeds 14.863, is 0.0019.

- There is a significant lack of fit with the logistic model.

- There is room for improvement in the model.

The sex of turtles

- What went wrong?
  - It could be that the dip in the number of males at 28.4^o C compared to the number of males at 28.3^o C is causing the logistic model not to fit well.
  - It could be that the curvilinear relationship between temperature and the proportion of male turtles is not symmetric. The logistic model is symmetric in its rise and then leveling off.

Significance of temperature?

- Although there is some lack of fit with the logistic model, does temperature and the logistic model give us statistically significant information about proportion of male turtles?

- We need to be able to measure if we are doing better predicting the proportion of male turtles using temperature and the logistic model than if were to make predictions ignoring the temperature.

Null Deviance

- The null deviance summarizes the fit of a logistic model that just includes an intercept.

$$\pi_i = \frac{e^{\beta_o}}{1+e^{\beta_o}}$$

- Such a model would predict a constant value for the response proportion regardless of the value of the predictor variable.

- By looking at the change in the deviance when a predictor variable is added, we can determine whether or not that predictor variable is adding significantly to the predictive ability of the model.

Change in Deviance

- The difference between the null deviance and the residual deviance represents the effect of adding a predictor variable to the logistic model.

- The change in deviance can be compared to a χ^2 distribution to determine statistical significance.

- The degrees of freedom for the χ^2 is equal to the number of predictor variables added to the model, in this case, 1.

The sex of turtles

- Null deviance (model with a constant proportion of males)
  - 64.4285 on 4 degrees of freedom
- Residual deviance (logistic model relating proportion of males to temperature)
  - 14.8630 on 3 degrees of freedom
- Change in deviance ("Importance" of temperature in the logistic model)
  - 49.5655 on 1 degrees of freedom

The sex of turtles

- Comparing the change in deviance, 49.5655, to a χ^2 with 1 degree of freedom, the P-value is virtually zero.

- This change in deviance is not attributable to chance alone, rather including temperature in the logistic model is adding significantly to your ability to predict the proportion of male turtles.

Summary

- Temperature is statistically significant in the logistic regression model for the sex of turtles. Using temperature and the fitted logistic model will give you better predictions for the proportion of males than using a constant proportion as your prediction.

- Although the logistic model using temperature is better than a constant proportion model, it may not give the best predictions. There is a significant lack of fit for the logistic model. This indicates that other curvilinear models may provide a better fit.

Comment

- The analysis we have done is on the combined data. Combining the three separate observations at each temperature into one is analogous to averaging observations in ordinary least squares linear regression.

- We can do logistic regression on the 15, three for each temperature, separate observations. The equation of the fitted logistic regression will be the same, however the deviances and degrees of freedom will change.

- The conclusions for this analysis are similar to those for the analysis of the combined data.

Alternative inference

- An alternative to the change in deviance for determining statistical significance of predictor variables in logistic regression is given by an approximate z-test statistic.

$$z = \frac{estimated\ \ parameter}{standard\ \ error}$$

- This z-test statistic has an approximate standard normal distribution for large samples.

Connection

- For very large samples (an asymptotic result) the change in deviance and the square of the z-test statistic should give approximately the same value.

- In small to moderate size samples, the two statistics can give different results.

- When in doubt, use the change in deviance.

The sex of turtles

- z-test statistic
  - $z = \frac{2.211}{0.4306} = 5.13$
- Square of the z-test statistic
  - $z^2 = (5.13)^2 = 26.32$
- Change in deviance
  - 49.57

The sex of turtles

- Both the z-test statistic and the change in deviance indicate that temperature is highly significant.

- Sample sizes for the combined turtle data are moderate, between 25 and 30 for each temperature.

- The P-values derived from either test will be approximate, at best.

Multiple Logistic Regression

- Often we have several predictor variables with a binary response.

- The ideas of logistic regression can be extended to the case with multiple predictor variables.

- As with multiple linear regression, we must be aware of problems introduced by multicollinearity (correlated predictor variables).

Multiple Logistic Regression

- The general multiple logistic regression equation

$$\pi_i = \frac{e^{(\beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki})}}{1 + e^{(\beta_o + \beta_1 X_i + \beta_2 X_{2i} + ... + \beta_k X_{ki})}} \quad (10)$$

Bronco Pulmonary Dysplasia

- Response: 1 if bronco pulmonary dysplasia (BPD) is present, 0 if BPD is absent.

- Predictors: the number of hours of exposure to Low, Medium and High levels of O_2. Since these numbers are quite spread out, a log transformation is used. Since some values are zero, the log transformation is applied to the number of hours plus 1.

BPD: One predictor at a time

- Single predictor: lnL=ln(Low + 1)

$$\widehat{\pi}_i' = -1.9193 + 0.3822\text{lnL}$$

- Null deviance: 50.4465, df = 39
- Residual deviance: 42.4022, df = 38
- Change in deviance: 8.0443, df = 1, P-value = 0.005

BPD: One predictor at a time

- Single predictor: lnM=ln(Medium + 1)

$$\widehat{\pi}_i' = -4.8411 + 0.9103\text{lnM}$$

- Null deviance: 50.4465, df = 39
- Residual deviance: 34.1814, df = 38
- Change in deviance: 16.2651, df = 1, P-value = 0.000

BPD: One predictor at a time

- Single predictor: lnH=ln(High + 1)

$$\widehat{\pi}_i' = -55.6682 + 11.0679\text{lnH}$$

- Null deviance: 50.4465, df = 39
- Residual deviance: 10.0584, df = 38
- Change in deviance: 40.3881, df = 1, P-value = 0.000

BPD: One predictor at a time

- The single best predictor is lnH=ln(High + 1). This results in the largest change in deviance, leaving the smallest residual deviance.

- Does adding a variable to the single predictor models improve the overall fit of the model; reduce the residual deviance significantly?

- We can talk about various selection procedures, such as forward selection and backward selection.

Model Selection Procedures

- Forward selection: add variables one by one
- Backward selection: delete variables one by one
- Stepwise selection: add/delete variables
- All possible models

Forward selection

- Start with a null (intercept only) model and add terms one at a time looking at the change in deviance to see if adding the term has caused a significant change.

- The final model can depend on the order of the variables entered.
  - Forward selection entering lnL first.
  - Forward selection entering lnH first.

Forward selection, lnL entered first

Model	Deviance	Change	P-value
Null	50.4465		
lnL	42.4022	8.0443	0.005
lnL + lnM	34.0134	8.3888	0.004
lnL + lnM + lnH	1.3409	32.6725	0.000

The final model contains all three variables. Each variable, when added, reduced the residual deviance significantly.

Splus

- The anova() command in Splus will summarize the sequential addition of terms.

  $->$ attach(bpd)

  $->$ bpd.logistic.full $<-$
  glm(BPD $\sim$ lnL+lnM+lnH,family=binomial)

  $->$ anova(bpd.logistic.full,test="Chisq")

  Note that in this command, lnL will be added first, lnM second and lnH last.

> stepAIC(bpd2)

- Start: AIC= 8
- BPD ~ lnH + lnL + lnM

- Df Deviance AIC
- <none> 2.490e-07 8.000
- - lnL 1 6.478 12.478
- - lnM 1 8.272 14.272
- - lnH 1 34.013 40.013

- Call: glm(formula = BPD ~ lnH + lnL + lnM, family = binomial, data = bpd)

- Coefficients:
- (Intercept) lnH lnL lnM
- -6844.33 1239.44 48.53 92.38

- Degrees of Freedom: 39 Total (i.e. Null); 36 Residual
- Null Deviance: 50.45
- Residual Deviance: 2.49e-07 AIC: 8

Forward selection, lnH entered first

Model	Deviance	Change	P-value
Null	50.4465		
lnH	10.0584	40.3881	0.000
lnH + lnL	8.2717	1.7867	0.181
lnH + lnM	6.4784	3.5800	0.058

If lnH is entered first then adding either lnL or lnM will not significantly change the residual deviance. This forward selection would stop at this point. One might suggest using an $\alpha = 0.10$. If one does, lnM would be entered at this step and lnL would be entered at the next step. The final model would then contain all three variables.

Other selection procedures

- Backward elimination
  - Begin with a full model with all variables included.
  - Eliminate a variable if by doing so you do not significantly change the residual deviance.
  - Continue eliminating variables until doing so would significantly increase the residual deviance.

Backward elimination

Model	Deviance	Change	P-value
lnL + lnM + lnH	1.3409		
lnM + lnH	6.4784	5.1375	0.023
lnL + lnH	8.2717	6.9308	0.008
lnL + lnM	34.0134	32.6725	0.000

Dropping any variable from the full model will significantly change the residual deviance giving a worse fit. This indicates that the model with all three variables is the best fit.

Other selection procedures

- Stepwise
  - Start with the null (intercept only) model.
  - Add a term that significantly reduces the residual deviance.
  - Continue to add terms that significantly reduces the residual deviance. Check other terms in the model to see if any can be eliminated. Stop when no more terms can be added or eliminated.

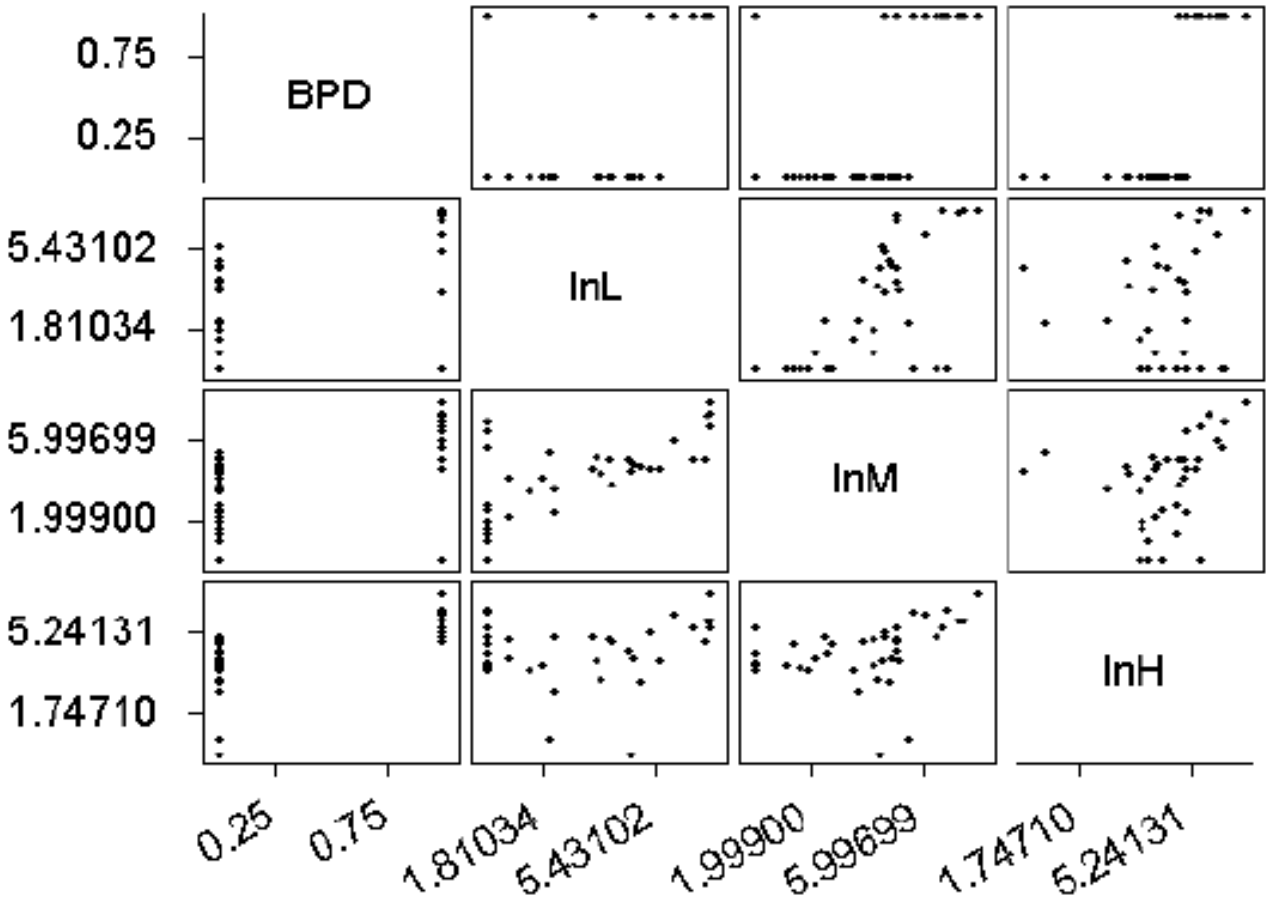
Stepwise selection

- A stepwise selection procedure will run into the same problems a forward selection procedure will in terms of stopping too soon.

- One can also do a stepwise selection procedure by starting with a backward elimination and checking to see if any of the previously eliminated variables might be added at a later time.

BPD in newborns

Incidence of BPD vs. ln(hours) at various levels of O2



- Download data from www.owlnet.rice.edu/~stat553/

- bpd$lnL <- log(bpd$Low+1) bpd$lnM <- log(bpd$Medium+1) bpd$lnH <- log(bpd$High+1)
- # Use the glm function to fit a # logistic regression model bpd2 <- glm(BPD ~ lnL + lnM + lnH, family=binomial, data=bpd)

- summary(bpd2)
- anova(bpd2, test="Chisq")

- bpd2 <- glm(BPD ~ lnH + lnL + lnM, family=binomial, data=bpd) summary(bpd2) anova(bpd2, test="Chisq")

All Possible Models

- As the name implies, every possible combination of variables is used to fit the data.

- The model that has the smallest residual deviance with all variables statistically significant would be chosen as the "best" model.

- This exhaustive search can become quite burdensome when there are many explanatory variables to consider.

Summary

- Binary response data abounds in many different application areas.

- Binary response data presents special problems because
  - the nature of the relationship is often curved.
  - the response is bounded below by zero and above by one.
  - equal variance and normal distribution assumptions are unreasonable.

Summary

- The logit transformation can account for the curved nature of the response as well as the bounds.

- Maximum likelihood estimation techniques are easy to apply to binary data provided one has access to a statistical computing package such as S+.

- The computing package accounts for the binomial nature of the response and uses the logit transformation.

Summary

- Inference for logistic regression is asymptotic in nature and so requires large amounts of data.

- Multiple logistic regression is a straightforward extension of one variable logistic regression.

- Various selection procedures can be employed to search for the "best" model.