Topic2 - Logistic Regression --

1. Topics

- Review inference for logistic regression models -- estimates, standard errors, confidence intervals, tests of significance, nested models

- Classification using logistic regression: sensitivity, specificity, and ROC curves

- Checking the fit of logistic regression models: cross-validation, goodness-of-fit tests, AIC

- Keywords: logistic regression, inference, analysis of deviance, likelihood ratio tests, Wald test, kyphosis, prediction, classification, sensitivity, specificity, ROC curve, cross-validation, Hosmer-Lemeshow statistic, Akaike Information Criterion (AIC)

2. Learning objectives

- Use multiple logistic models to understand how risk of kyphosis (curvature of the spine) depends on several predictor variables

- Use logistic regression to classify subjects and assess the quality of a classification rule with its sensitivity, specificity and ROC curve

- Use cross-validation to make unbiased evaluations of classification rules

3. Inference for logistic regression (LR) models

3.1 LR Model

- Recall the LR model:

(1) Y_i are from a Binomial (n_i=1, μ_i) distribution

$\mu_i = Pr(Y_i=1 \mid X_i s)$, n observations

(2) Y_i are independent

(3) $log\ odds(Y=1) =$

$$log\left(\frac{\mu_i}{1-\mu_i}\right) =$$

$$\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

The LR model implies:

(a) $Odds(Y_i=1) = e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}$

(b) $Pr(Y_i=1) = \dfrac{Odds}{1 + Odds} =$

$$\mu_i = \frac{e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}}$$

(c) $Pr(Y_i=0) =$

$$1 - \mu_i = \frac{1}{1 + e^{\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}}}$$

(d) $Var(Y_i) = \mu_i \ (1 - \mu_i)$

3.2 Interpretation of coefficients

- $\beta_0 =$ *log odds* $(Y_i=1)$, given all Xs = 0

  $e^{\beta_0} =$ *odds* $(Y_i=1)$, given all Xs = 0

- $\beta_k =$ Difference in *log odds* $(Y_i=1)$ for $X_k + 1$ -vs- X_k, holding other Xs constant

  $e^{\beta_k} =$ odds ratio for $X_k + 1$ -vs- X_k, holding other Xs constant

3.3 Maximum likelihood estimates and standard errors

- The method of maximum likelihood estimation chooses values for parameter estimates which make the observed data “maximally likely.” Standard errors are obtained as a by-product of the maximization process

- Use **Stata** to get maximum likelihood estimates

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \quad (\text{ and } e^{\hat{\beta}_0}, e^{\hat{\beta}_1}, \ldots, e^{\hat{\beta}_p})$$

and standard errors

$$se_{\hat{\beta}_0}, se_{\hat{\beta}_1}, \ldots, se_{\hat{\beta}_p}$$

logit command gives $\acute{\beta}$s

logistic command gives the $e^{\hat{\beta}}$s

3.4 Comparing nested models

- Null model:

$$log\ odds\ (Y_i=1) =$$

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- Extended model:

$$log\ odds\ (Y_i=1) =$$

$$\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{p+1} X_{p+1} + \cdots + \beta_{p+s} X_{p+s}$$

s "new" *X*s

- Problem: Test hypothesis that multiple βs = *0*:

$$H_o: \quad \beta_{p+1} = \beta_{p+2} = \cdots = \beta_{p+s} = 0$$

- Solution: Use likelihood ratio test (LRT)

$$-2(\text{loglik}_{\text{NULL}} - \text{loglik}_{\text{EXT}}) \sim \chi^2_{S\ df} \text{ when } H_o \text{ is true}$$

3.5 p-value for β_j

- **p-value** for H_0 vs H_a (two-sided) for any given β_j can be obtained in two ways:

(1) Wald test: $z = \dfrac{\hat{\beta}_j}{se_{\hat{\beta}_j}}$

or,

(2) Likelihood ratio test (LRT) comparing null (X_j removed) and extended (X_j included) models:

$$X_1^2 = -2(LL_{X_j \, removed} - LL_{X_j \, included})$$

Likelihood ratio tests are valid under a wider range of conditions than Wald tests

- In **Stata**, the estimates table gives Wald tests; use **lrtest** as shown above in the example for nested models to get likelihood ratio tests

3.6 Confidence interval for β_j

- *100(1-α)%* CI for β_j

$$\hat{\beta}_j \pm z_{1-\alpha/2} \cdot \hat{se}_{\hat{\beta}_j}$$

- In **Stata**, the estimates table gives CIs

3.7 Standard error for linear combinations of coefficient estimates - FYI

- At times, need to calculate

  Variance $(w_1\hat{\beta}_1 + w_2\hat{\beta}_2)$

  for example,

  $\hat{\beta}_1 + \hat{\beta}_2 \quad (w_1 = w_2 = 1)$

or,

  $\hat{\beta}_1 - \hat{\beta}_2 \quad (w_1 = 1, w_2 = -1)$

- Recall formulas for variance calculations

  — *2* Independent samples

  — Want to estimate μ_1-μ_2, difference of means

  — $\bar{x}_1 - \bar{x}_2$ is best estimate

  — $Var(\bar{x}_1 - \bar{x}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$

3.7 Standard error for linear combinations of coefficient estimates - FYI (cont'd)

(variances add when samples are independent)

— $se_{\bar{x}_1 - \bar{x}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

- Example

SAMPLE:	# 1	# 2
$\bar{x}$	10	5
$se_{\bar{x}}$	3	4

$$se_{\bar{x}_1 - \bar{x}_2} = \sqrt{se^2_{\bar{x}_1} + se^2_{\bar{x}_2}}$$
$$= \sqrt{3^2 + 4^2} \quad = \quad 5$$

95% CI for $\mu_1 - \mu_2$:

$(10 - 5) \pm 2 \cdot 5 \;=\; 5 \pm 10$

3.7 Standard error for linear combinations of coefficient estimates - FYI (cont'd)

- More generally,

$$Var\,(w_1\hat{\beta}_1 + w_2\hat{\beta}_2) =$$

$$w_1^2\,Var(\hat{\beta}_1) + w_2^2\,Var(\hat{\beta}_2) +$$

$$2\,w_1 w_2 \sqrt{Var(\hat{\beta}_1)\,Var(\hat{\beta}_2)}\;corr(\hat{\beta}_1,\hat{\beta}_2)$$

So,

$$Var\,(\hat{\beta}_1 + \hat{\beta}_2) =$$

$$1^2\,Var(\hat{\beta}_1) + 1^2\,Var(\hat{\beta}_2) +$$

$$2\cdot 1\cdot 1\sqrt{Var(\hat{\beta}_1)\,Var(\hat{\beta}_2)}\;corr(\hat{\beta}_1,\hat{\beta}_2) =$$

$$Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2\sqrt{Var(\hat{\beta}_1)\,Var(\hat{\beta}_2)}\,corr(\hat{\beta}_1,\hat{\beta}_2)$$

- In **Stata**, use the **lincom** command after fitting a regression model to obtain the estimate, se, p-value, and *95% CI* for a linear combination of βs:

lincom x1+x2 for $\beta_1 + \beta_2$

lincom x2-x1 for $\beta_2 - \beta_1$

- END of FYI

4. Kyphosis example - Continuous predictor variables

- The data for this example are included with the S-PLUS system for statistical analysis and relate to 81 children who have undergone spinal surgery

- The response is the occurrence of a surgical complication – a post-operative spinal deformity known as kyphosis

4.1 Variables and questions

- Variables

  — Y_i indicates occurrence kyphosis following spinal surgery (1 = present, 0=absent)

  — *age* age of child in months

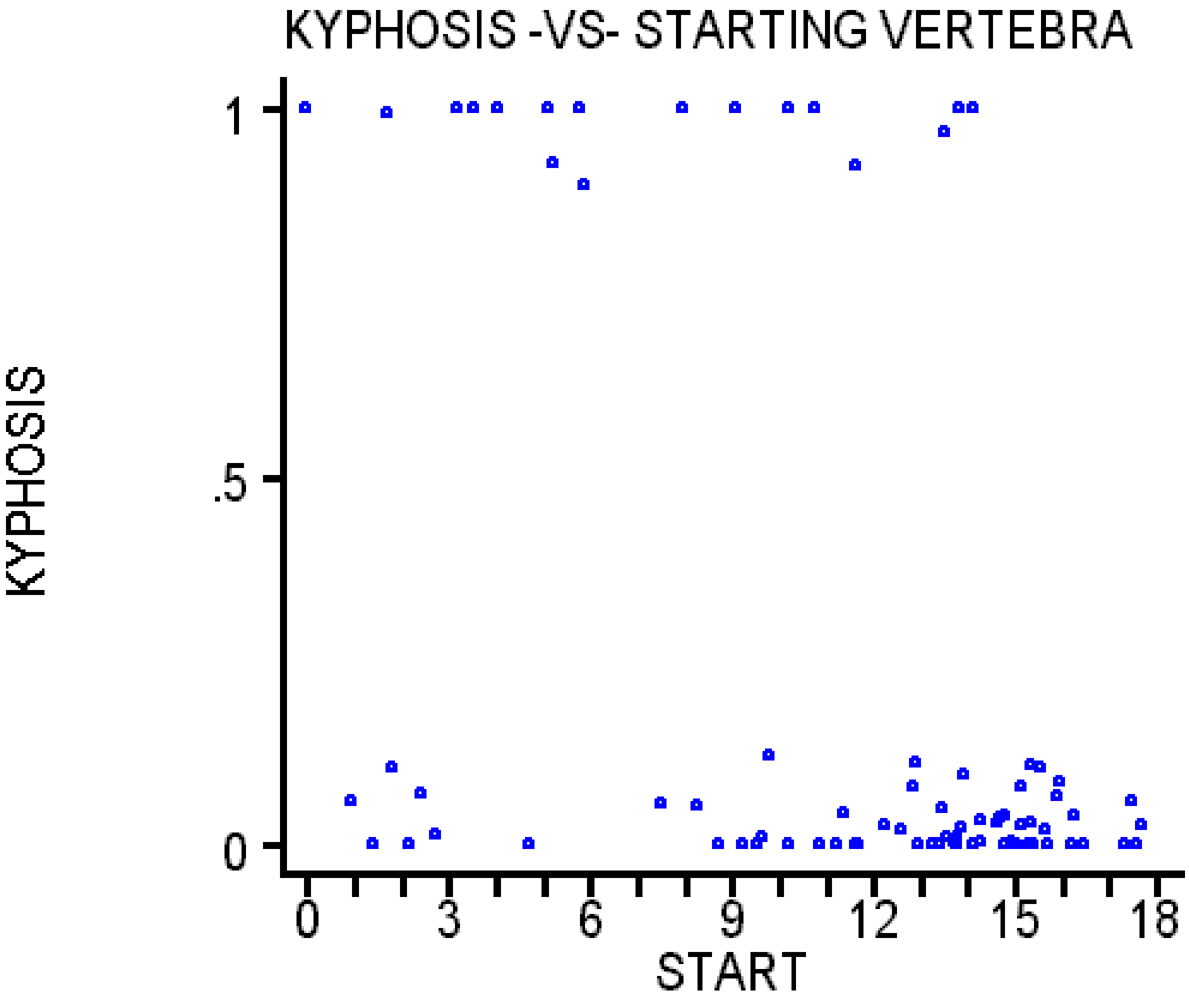
  — *number* number of vertebrae involved in spinal surgery

  — *start* first vertebra involved in spinal surgery

- Question: Which factors predict the occurrence of the post-operative spinal deformity, kyphosis?

4.1 Variables and questions (cont'd)

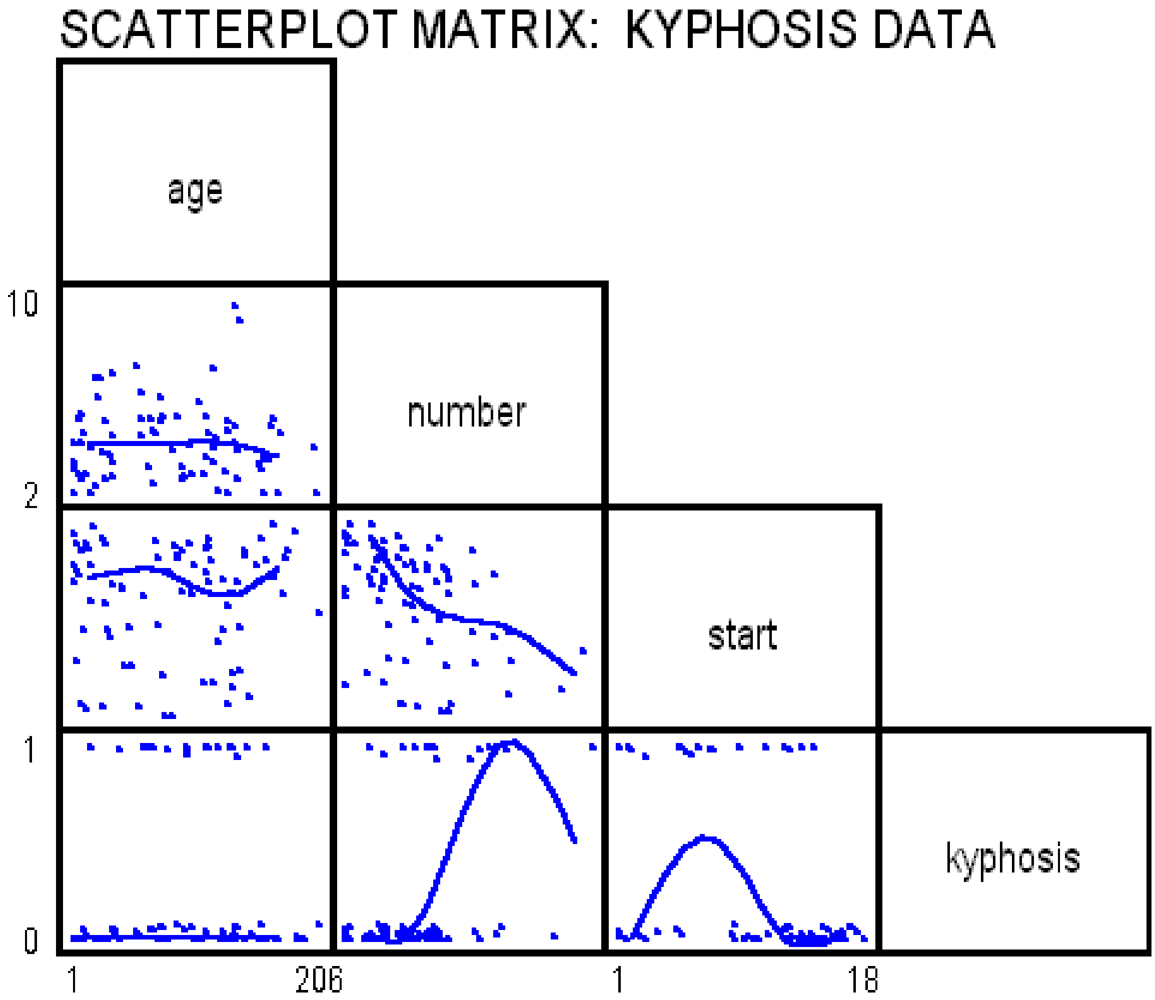
- For example, the following graph shows how the occurrence of kyphosis is related to the predictor *start*, the first vertebra involved in the surgery



4.2 Display data

- Scatterplot matrix

```
graph age number start kyphosis, matrix half
          jitter(5) symbol(.)connect(s) bands(4)
```



4.2 Display data (cont'd)

- Kernel smoothed plots of kyphosis proportions -vs- candidate predictors using **Stata**'s **ksm** non-parametric regression command

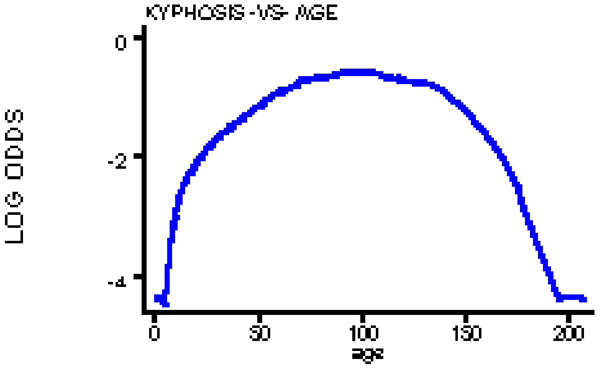
  This can be very useful for suggesting the shape of non-linear relationships without making any a priori assumptions about the mathematical form of the relationship between a predictor and the response -- see the **Stata** manual for more details and references

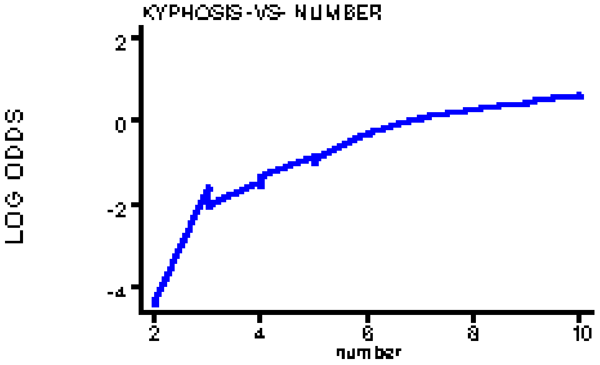
  You may need to vary the smoothing band width parameter (*bwidth* ranges from 0 to 1) to get the desired amount of smoothing (higher values give more smoothing); the *adjust* parameter is used with LR models; *logit* requests plots of log odds ratios

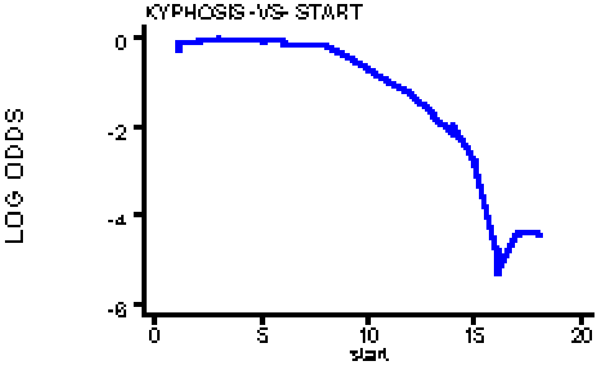
  ksm kyphosis age, lowess xlab ylab logit adjust symbol(.) jitter(5) bwidth(0.8)

- With the *logit* option

4.2 Display data (cont'd)

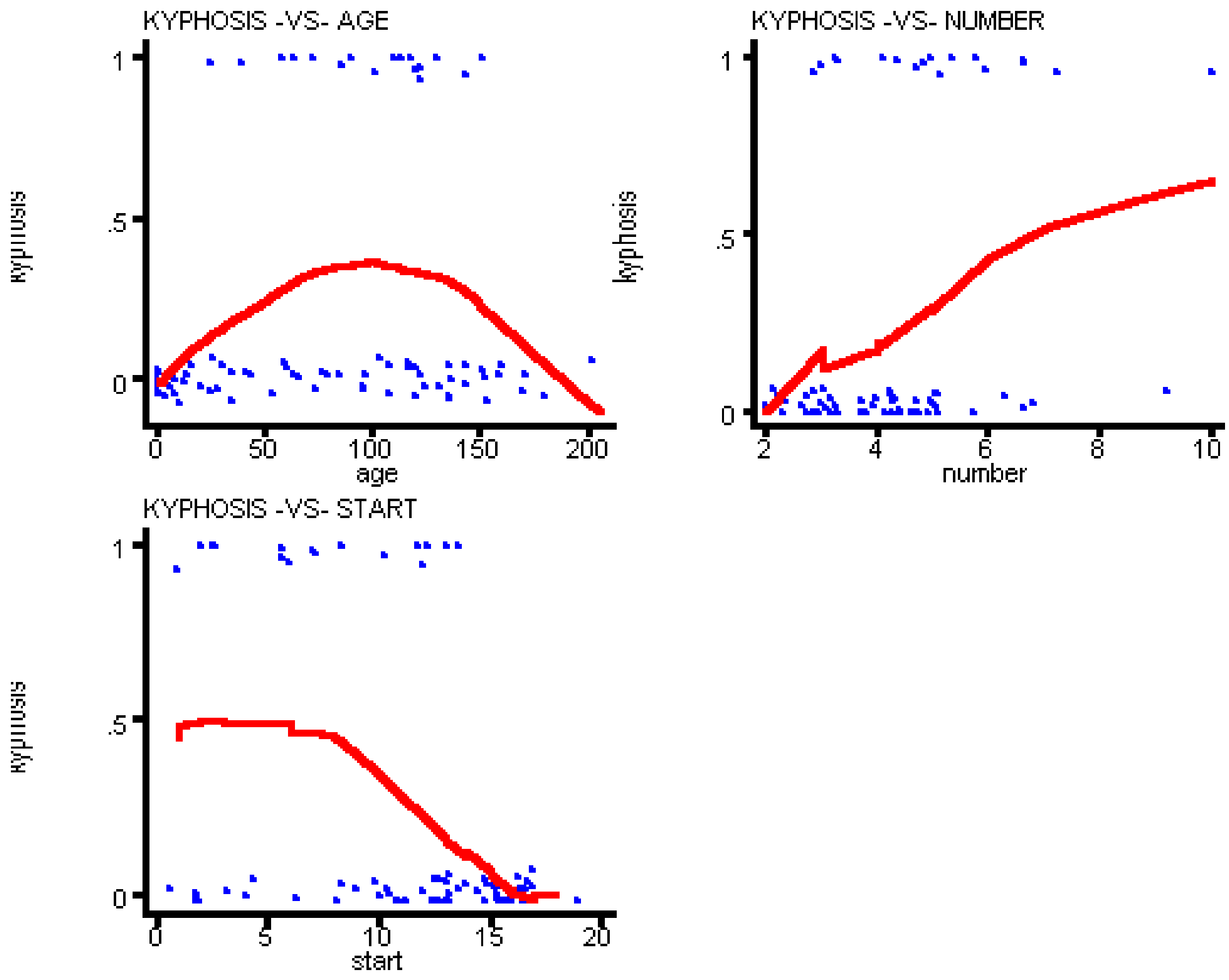






4.2 Display data (cont'd)

- Without the *logit* option



4.3 Model

log odds (kyphosis) =

log odds (Y=1) =

$$log\left(\frac{Pr\,(kyphosis)}{1-\ Pr\,(kyphosis)}\right) =$$

$$= \beta_0 + \beta_1(age\text{-}84) + \beta_2(age\text{-}84)^+$$

$$+ \beta_3(start\text{-}13) + \beta_4(number\text{-}4)$$

$$(age\text{-}84)^+ = \begin{cases} age-\ 84 & if\ age\ >\ 84 \\ 0 & if\ age\ \leq 84 \end{cases}$$

age, *start* and *number* are centered; *(age-84)*$^+$ is a linear spline term that allows for a differing rates of change in the log odds of kyphosis before and after age 84 months (7 years)

4.4 Parameter interpretation

- Model parameters (coefficients)

β_0 log odds of kyphosis for a 7 year old child whose surgery was for vertebrae 13,14,15,16

β_1 log ratio of the odds of kyphosis for children whose ages differ by one month, are ≤ 7 years of age and have same surgical variables (*start* and *number*)

$\beta_1+\beta_2$ log ratio of the odds of kyphosis for children whose ages differ by one month, are more than 7 years of age and who have the same surgical variables (*start* and *number*)

β_3 log ratio of the odds of kyphosis for children whose starting vertebrae involved in the surgery are one apart, are the same age and have same number of vertebrae involved

β_4 You do!

4.5 Results

- Logistic regression model estimates

Var	Estimate	se	z
Intercept	*-.843*	*.65*	*-1.29*
(age-84) mo	*.0486*	*.0192*	*2.53*
$(age-84)^+$ mo	*-.0663*	*.0293*	*-2.26*
(start-13)	*-.215*	*.0719*	*-2.99*
(number-4)	*.431*	*.236*	*1.83*

- **Stata** commands for logistic regression (*logit* coefficients that relate to log odds and *logistic* gives coefficients that relate to odds ratios):

```
logit kyphosis agec agep startc numberc

logistic  kyphosis agec agep startc numberc
```

4.5 Results (cont'd)

- **Stata** log:

```
------------------------------------------------------------------------------
kyphosis |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    agec |    .048641   .0192218      2.531   0.011       .0109669    .0863151
    agep |  -.0663331   .0293293     -2.262   0.024      -.1238174   -.0088488
  startc |  -.2151268   .0719186     -2.991   0.003      -.3560847   -.0741689
 numberc |   .4311881    .236198      1.826   0.068      -.0317514    .8941277
   _cons |  -.8425175   .6523282     -1.292   0.197      -2.121057    .4360223
------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------
kyphosis | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    agec |   1.049843   .0201799      2.531   0.011       1.011027     1.09015
    agep |   .9358191   .0274469     -2.262   0.024       .8835411    .9911903
  startc |   .8064392    .057998     -2.991   0.003       .7004133    .9285149
 numberc |   1.539085   .3635288      1.826   0.068       .9687473    2.445202
------------------------------------------------------------------------------
```

4.6 Likelihood ratio test

- Are any of the four predictors related to the occurrence of kyphosis? The hypothesis corresponding to this question is

  $H_{0:}\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

- Test this hypothesis, use either the analysis of deviance or the equivalent likelihood ratio test (LRT) for comparing nested models:

  Null: intercept only
  Extended: intercept and 4 predictors as
above

- Results: analysis of deviance (Dev) and log likelihoods

Model	*df*	*loglik*	*-2 loglik*
Intercept only	*1*	*-41.6*	*83.2*
4 predictors, as above	*5*	*-27.5*	*55.1*
Difference	*4*	*-14.1*	*28.1*

LRT:

4.6 Likelihood ratio test (cont'd)

$$\begin{aligned} -2(\text{Difference in log likelihoods}) &= -2(-41.6-(-27.5)) \\ &= -2(-14.1) \\ &= 28.2 \end{aligned}$$

$$\Pr(\chi^2_4 > 28.2) < .001$$

- **Stata** commands for the LRT

(*quietly* prefix suppresses estimation table)

For LRT:

```
quietly logit kyphosis age agep startc numberc
lrtest, saving(0)

quietly logit kyphosis
lrtest
```

4.7 Correlations among coefficients - FYI

- Coefficient estimates are usually not independent; they are correlated. In the kyphosis example, the correlation among all pairs of $\acute{\beta}$s are shown in the following "correlation matrix"

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
$\hat{\beta}_0$	.49	-.70	.26	.15
$\hat{\beta}_1$		$\underline{\underline{-.90}}$	-.28	.22
$\hat{\beta}_2$			.20	-.11
$\hat{\beta}_3$				.05

— To get the matrix of correlations among the $\acute{\beta}$s after a regression fit, use the **Stata** command:

vce , corr

- Example: Derive the variance of the estimate of

$\beta_1 + \beta_2$ -- the log odds ratio per year of age, after 7 years of age

Use the basic formula for the variance of a sum:

4.7 Correlations among coefficients - FYI (cont'd)

$$Var\,(\hat{\beta}_1 + \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2\,Cov\,(\hat{\beta}_1, \hat{\beta}_2)$$

$$= Var(\hat{\beta}_1) + Var(\hat{\beta}_2) +$$

$$2\sqrt{Var(\hat{\beta}_1) \times Var(\hat{\beta}_2)} \times corr(\hat{\beta}_1, \hat{\beta}_2)$$

$$= .0192^2 + .0293^2 + 2(.0192)(.0293)(-.90)$$

$$= .0146^2 = .00021316$$

- Then, derive

95% CI for $\beta_1 + \beta_2$:

$= (.0486+.0663) \pm 2(.0146)$

$= .1149 \pm .0292$

$= (-.0857, .1441)$

- Take antilogs, to get

95% CI for $e^{\beta_1 + \beta_2}$:

$= (e^{-.0467}, e^{.0115})$

$= (.954, 1.012)$

Compare with **Stata**'s **lincom** command -- gives the same results, apart from rounding:

```
. * Estimate beta1+beta2  - slope after age 7
.
. lincom agec+agep

 ( 1)  agec + agep = 0.0

------------------------------------------------------------------------------
kyphosis |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |  -.0176921   .0147775     -1.197   0.231      -.0466555    .0112713
------------------------------------------------------------------------------

------------------------------------------------------------------------------
kyphosis | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     (1) |   .9824635   .0145184     -1.197   0.231       .9544161    1.011335
-----------------------------------------------------------------
```

- END of FYI

4.8 Several alternative models

- Results of fitting several alternative logistic regression models to the kyphosis data

4.8 Several alternative models

					Residual	
Model	*Vars*	*Est*	*se*	*z*	*#parm*	*-2 loglik**
A	*intercept*	*-.127*	*.52*	*-.25*	*3*	*74.7*
	age-84	*.037*	*.014*	*2.5*		
	(age-84)$^+$	*-.057*	*.023*	*-2.5*		
B	*intercept*	*-1.94*	*.386*	*-5.0*	*2*	*68.1*
	start-13	*-.218*	*.060*	*-3.6*		
C	*intercept*	*-1.52*	*.31*	*-4.9*	*2*	*73.4*
	number-4	*.53*	*.19*	*2.9*		
D	*intercept*	*-.74*	*.64*	*-1.2*	*4*	*59.2*
	age-84	*.044*	*.017*	*2.5*		
	(age-84)$^+$	*-.063*	*.028*	*-2.3*		
	start-13	*-.244*	*.070*	*-3.5*		
E	*intercept*	*-.36*	*.56*	*-.64*	*4*	*65.4*
	(age-84)	*.040*	*.016*	*2.4*		
	(age-84)$^+$	*-.058*	*.025*	*-2.3*		
	number-4	*.55*	*.204*	*2.7*		
F	*intercept*	*1.6*	*.76*	*2.1*	*4*	*53.4*
	age-84	*.051*	*.019*	*2.73*		
	(age-84)$^+$	*-.076*	*.030*	*-2.52*		

					Residual	
Model	*Vars*	*Est*	*se*	*z*	*#parm*	*-2 loglik**
	(start-10)$^{+}$	*-.71*	*.200*	*-3.51*		
G	*intercept*	*1.66*	*.996*	*1.7*	*5*	*53.4*
	age-84	*.052*	*.019*	*2.7*		
	(age-84)$^{+}$	*-.077*	*.031*	*-2.5*		
	start-10	*-.017*	*.128*	*.135*		
	(start-10)$^{+}$	*-.746*	*.340*	*-2.2*		
H	*intercept*	*-.843*	*.652*	*-1.3*	*5*	*55.1*
	age-84	*.049*	*.019*	*2.5*		
	(age-84)$^{+}$	*-.066*	*.029*	*-2.3*		
	start-10	*-.215*	*.072*	*3.0*		
	(number-4)	*.431*	*.236*	*1.8*		

*relative to best possible

4.9 Interpretation

- By fitting multiple models, we learned:
  - — Starting vertebrae is best predictor of kyphosis, higher starting numbers have lower risk
    (Model B)
  - — Age and number of vertebrae in surgery are comparable predictors
    (Models A and C)

4.9 Interpretation (cont'd)

— Given *age*, *start* and *number* are both still important
(Models D and E)

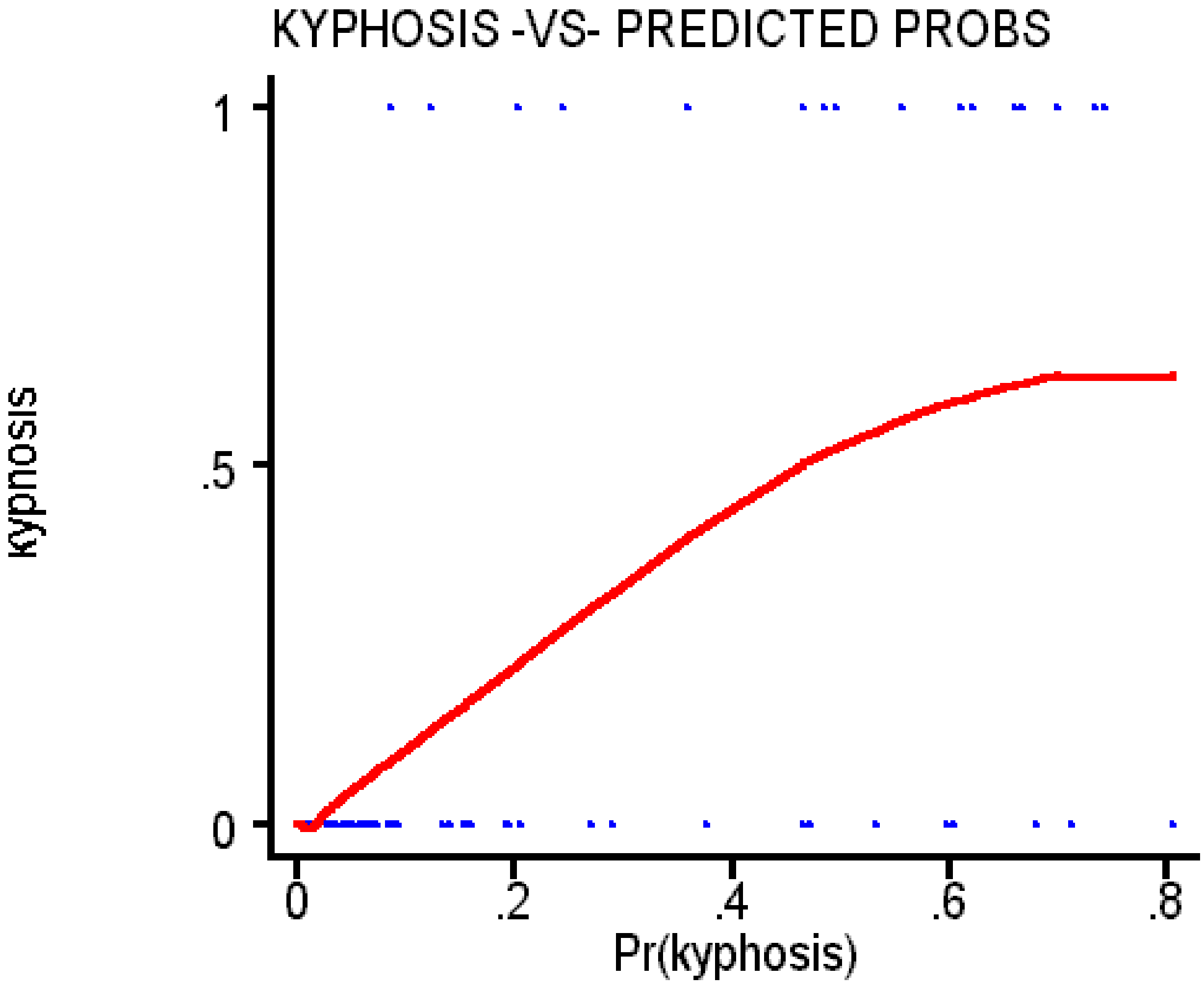
— Given *age* and *start*, *number* becomes marginal
(Model H)

— Risk is high but constant until starting vertebra 10 and then decreases linearly
(Model F -vs- G; Model F)

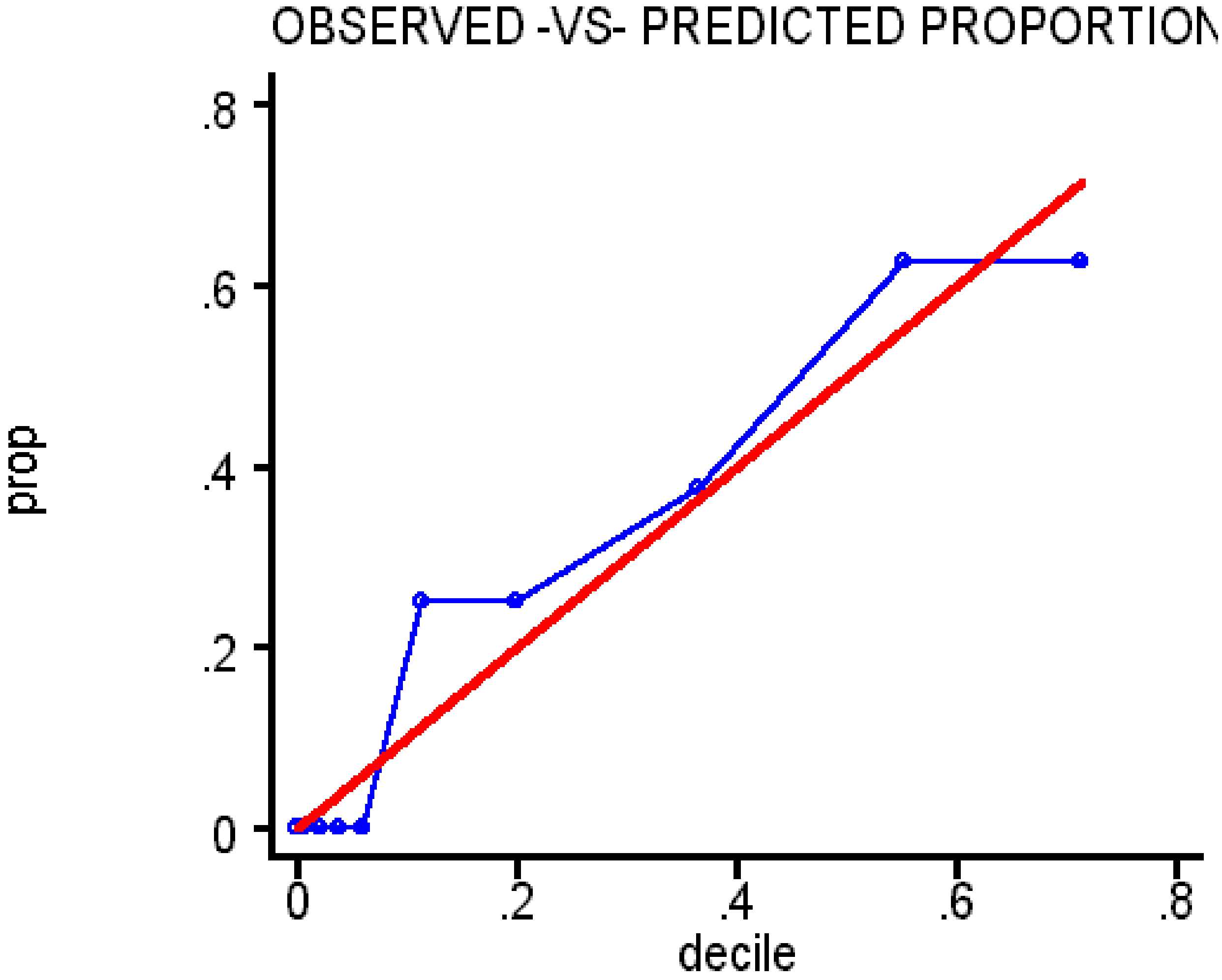
4.10 Checking LR model fit

- Plot binary response Y_i against predicted probability Y=1 ($\hat{\mu}_i$)

- Add smooth curve - ought to be roughly straight if the fit is good



- Divide predicted probabilities into deciles and plot observed proportions in each decile -vs- midpoints of decile bins – should give roughly a straight line indicating that the observed proportion of Y=1 ≈ $\hat{\mu}$



4.10 Checking LR model fit (cont'd)

- Hosmer-Lemeshow goodness-of-fit χ^2_{g-2}

Tests observed -vs- expected counts in cells defined by grouping the predicted probabilities into *g* groups, usually *g= 8 to 10*. Small p-values indicate poor fit, i.e., poor agreement between observed and expected counts.

Choose number of groups so that expected counts are all >5, if possible

Stata results for H-L test of Model F with 8 groups:

```
. quietly logit kyphosis age agep startp

. lfit , group(8)

Logistic model for kyphosis, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

       number of observations =        81
             number of groups =         8
      Hosmer-Lemeshow chi2(6) =         1.87
                  Prob > chi2 =         0.9310
```

4.11 Summary of kyphosis analysis

- A **prospective** study of **81** children, ages 1 to 206 months, who underwent spinal surgery, was conducted to identify risk factors for kyphosis, a post-operative spinal deformity. Three predictors were considered: child's age, first vertebra, and number of vertebrae involved in the surgery

- Considering the three predictors, the odds of kyphosis is **estimated** to **increase** by **5.0% per month** of age **(95% CI = 1.0%, 9.0%)** for the first 7 years and then to **decrease** by -1.8% per month afterwards (95% CI = -4.6%, 1.1%).

  ... (Model H)

- At a given age, the starting vertebra is a more important predictor than the number of vertebrae. The odds of kyphosis decreases by approximately -22% per vertebrae (95%CI = -32%, -11.2%) although the data provide some evidence that the risk decreases slowly at first and then faster after vertebrae 10.

  ... (Models D&E; F&G)

4.11 Summary of kyphosis analysis (cont'd)

- At a given age and starting vertebrae, the odds of kyphosis increase with a greater number of vertebrae involved (OR 1.5 per vertebrae, 95% CI = .97, 2.4) although this effect is not statistically significant.

  ... (Model H)

- We used the logistic regression model with age, first vertebrae and number of vertebrae as a screening instrument. In the group of patients observed, we estimate that this model can screen for likely kyphosis cases with a sensitivity of 90% at a specificity of 75%, defining positive screen as a predicted probability of kyphosis $\geq$ 0.3

  ... (Model H)

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI

- How well can we screen for kyphosis based on the predictors age, starting vertebra, and number of vertebrae?

- Logistic regression gives $\hat{\mu}_i = Pr\ (Y_i = 1 \mid X_i)$

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

- Define a screening “test” based on whether the predicted probabilities from the LR model fall above (= "positive") or below (= "negative") some cut point *c*

Screening “test” = “positive” if $\hat{\mu}_i > c$

e.g., take $c = .5$ -- *c* can be set to any value between 0 and 1; usually must try several choices and compare results

Note: The LR model H was used for illustration

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

- Results comparing the logistic regression classification (screening “test” above the cut point *c=0.5*) compared with the true classification into kyphosis or not

Screen	Kyphosis	
c =.5	+	-
+	*8*	*3*
-	*9*	*61*

- $Sensitivity_c = Pr\,(screen = "+ " \mid kyphosis = +)$

- $Specificity_c = Pr\,(screen = "- " \mid kyphosis = -)$

- Can estimate the sensitivity and specificity from data above for a cut point of .5

$$\widehat{Sens}_{c=.5} = 8 / (8 + 9) = .47$$

$$\widehat{Spec}_{c=.5} = 61 / (61+3) = .95$$

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

- Thus, a cut point of *c* = *.5* gives a classification with poor sensitivity.

  What if we reduce the cut point to *c*=*.3*?

Screen	Kyphosis	
c =.3	+	-
+	*13*	*8*
-	*4*	*56*

$$\hat{Sens}_{c=.3} = 13 / (13 + 4) = .76$$

$$\hat{Spec}_{c=.3} = 56 / (56 + 8) = .88$$

- For a particular cutoff point, *c*, **Stata** can calculate sensitivity, specificity, and other characteristics (an epidemiologist’s feast!) for the screening test defined by any given cut point

  For cut point of *c=0.3*:

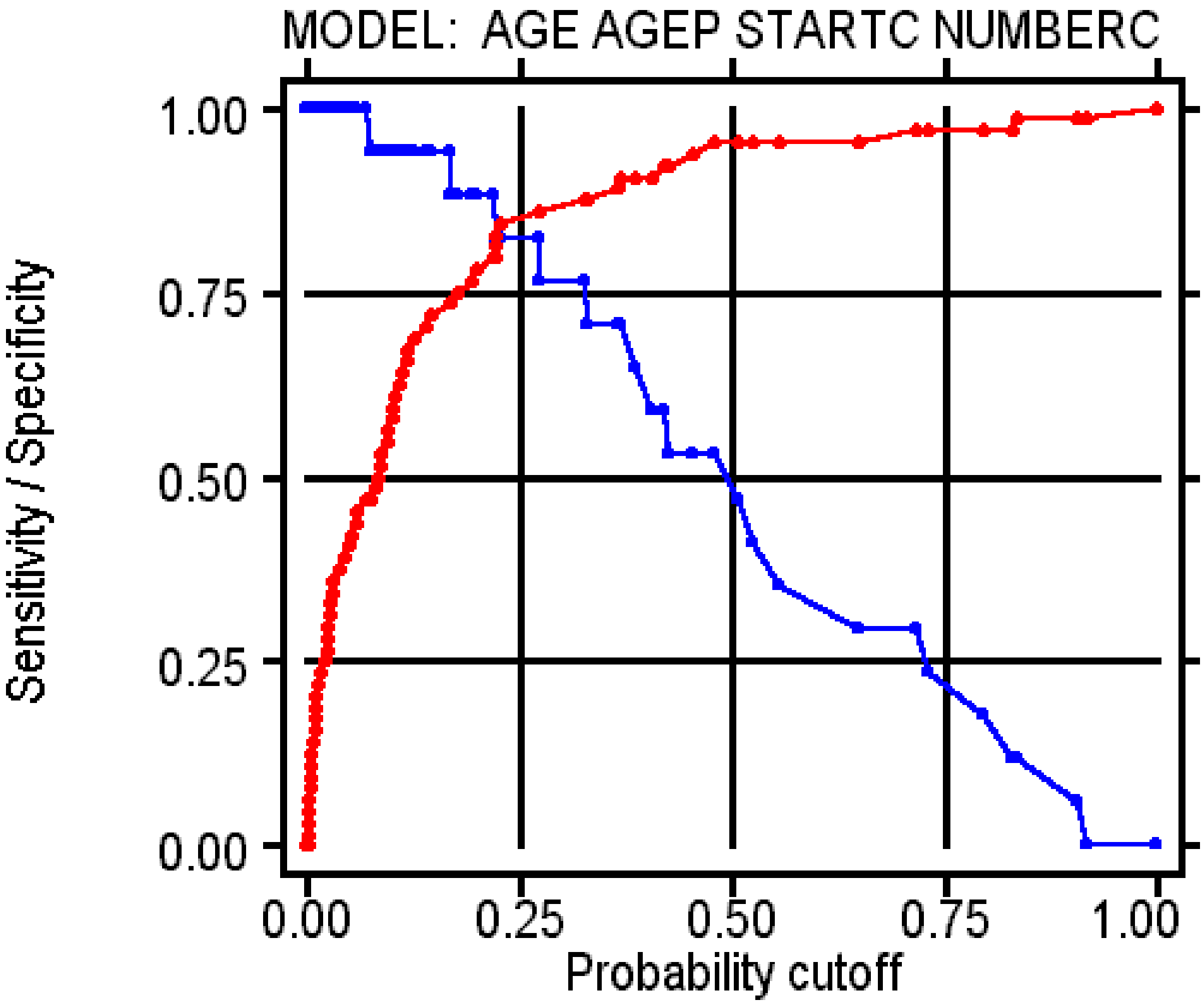
4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

```
. lstat , cutoff(0.3)
Logistic model for kyphosis
              -------- True --------
Classified |         D            ~D        Total
-----------+--------------------------+-----------
     +     |        13             8  |        21
     -     |         4            56  |        60
-----------+--------------------------+-----------
   Total   |        17            64  |        81

Classified + if predicted Pr(D) >= .3
True D defined as kyphosis ~= 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   76.47%
Specificity                     Pr( -|~D)   87.50%
Positive predictive value       Pr( D| +)   61.90%
Negative predictive value       Pr(~D| -)   93.33%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   12.50%
False - rate for true D         Pr( -| D)   23.53%
False + rate for classified +   Pr(~D| +)   38.10%
False - rate for classified -   Pr( D| -)    6.67%
--------------------------------------------------
Correctly classified                        85.19%
--------------------------------------------------
```

- Now, repeat for all possible cut point values *c* in *(0,1)* and display the results in a plot of sensitivity and specificity -vs- *c* with the **Stata** command:

  lsens

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)



4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

- Another good way to evaluate the usefulness of the classification is to consider all cut point values *c* in the interval *(0,1)* and combine the resulting sensitivity and specificity into an “ROC” Curve – look ahead for an example

- ROC stands for “Receiver-Operating Characteristic” – first used in signal detection theory

  — Plot *Sensitivity* -vs- (*1 minus Specificity*), for all cut points *c* ranging from 0 to 1

  — In other words, plot “True Positive Rate” -vs- “False Positive Rate”

  — The curve starts at *(0,0)* corresponding to *c=1* and stops at *(1,1)* corresponding to *c=0*

  — ROC curve for a model without any predictive power is a 45° line

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

— The steeper the ROC curve, the greater the predictive power

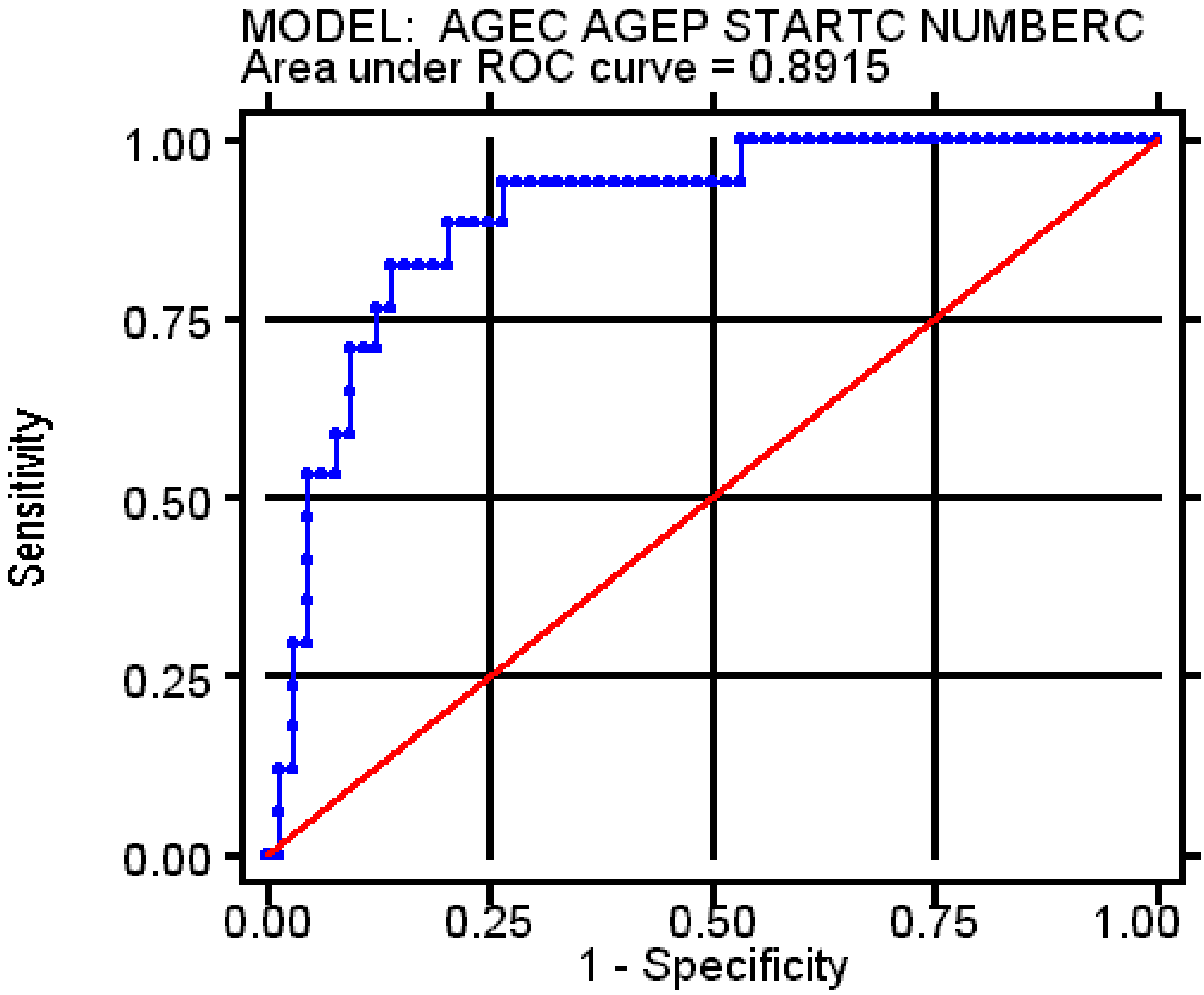
— More area beneath the ROC curve indicates greater predictive power

area= 0.5 for no predictive power
1.0 for perfect predictive power

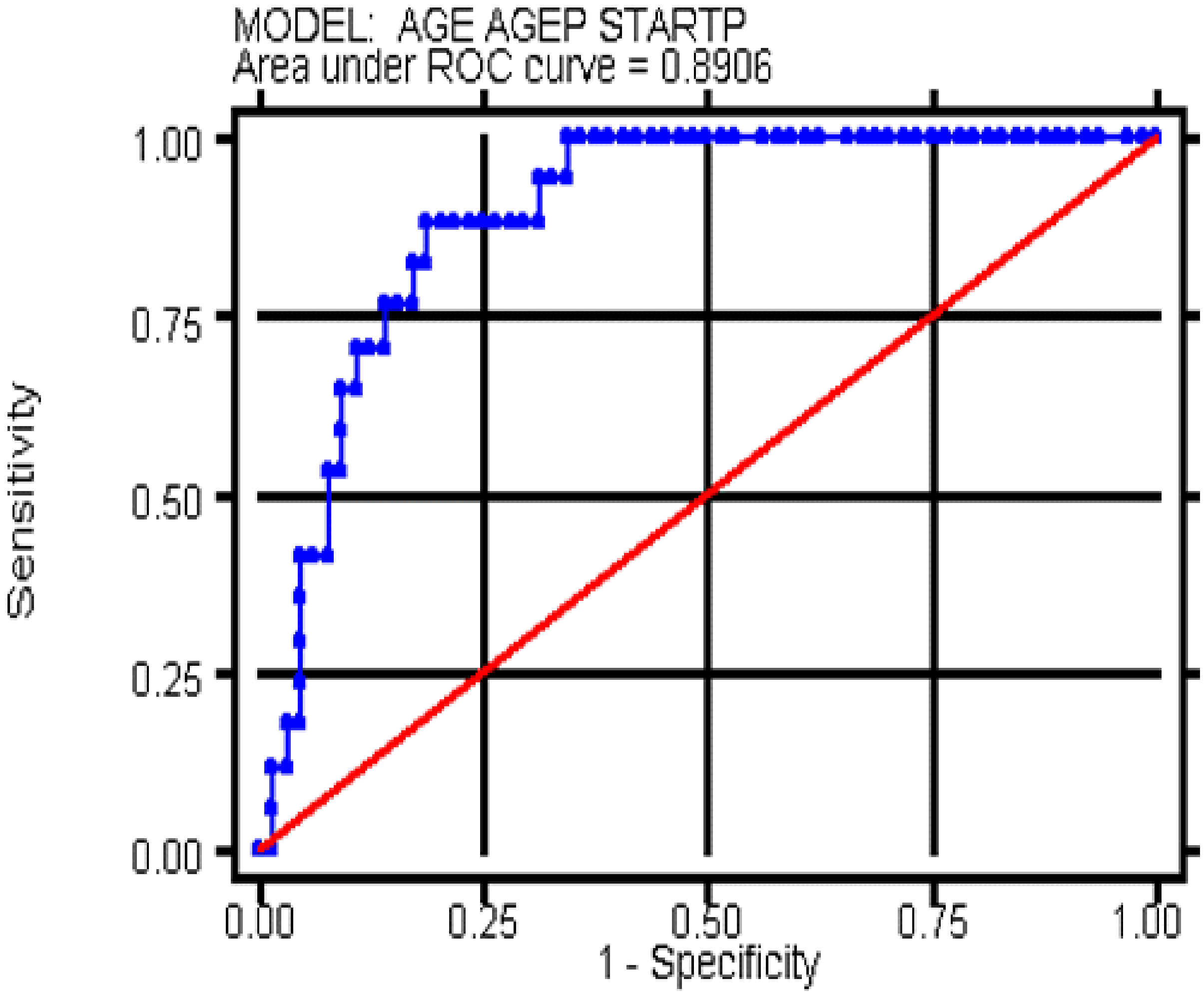
— ROC curves are most useful for comparing two or more competing classifications

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

— The ROC curves below compare Models F and H



4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)



- Additional material on ROC analysis covering software for statistical comparison of ROC curves, including tests and 95% confidence limits for areas under the curve:

4.12 Classification using LR: sensitivity, specificity, ROC curves - FYI (cont'd)

Stata command: *roctab*, written by Mario Cleves at the Stata Corporation, is included in the **Stata** Technical Bulletin (STB-52, sg120) and can be downloaded from the **Stata** website: *www.stata.com*. Documentation is in STB-52, which can be purchased from the website

— *Rockit*, is a public domain, stand-alone program for PCs and other platforms. This software for ROC analysis was written by Charles Metz from the Radiology Department at the University of Chicago. The software and documentation, including an extensive set of references, is available from

http://www-radiology.uchicago.edu/sections/roc

Note: the URL must be entered exactly as above, including the http:// and, yes, that really is a dash after www

4.13 Split sample cross-validation - FYI

- The sensitivities and specificities calculated from the observed data are overly optimistic

- Why?

  — Because the same data used to fit the "optimal" model (max likelihood or least squares) was used to test the classification power of the model as indicated by sensitivity and specificity

  — "Optimization capitalizes on chance" (Tukey speaking "Tukish")

- Solution - Use the method of Cross-Validation

  — Leave out a small fraction of the data (as small as 1/n -- leave out a fraction 1/n)

  — Use the rest of the data to fit the logistic model

  — Predict the left out values and classify them using $\hat{\mu}_i > c$

4.13 Split sample cross-validation - FYI (cont'd)

— Repeat for all possible “small fractions of the data” -- most practical when the fraction is 1/n

— Calculate sensitivity and specificity with predictions from the models and data that does not include the observations used to fit the model

— Thus, the strength of the method of cross validation for comparing models is that it uses different data for fitting models than it uses for evaluating models

- Cross-validation of the ROC analysis for the kyphosis data, deleting 1 observation at a time -- fraction deleted =1/81 (n= 81 observations)

4.13 Split sample cross-validation - FYI (cont'd)

— Use a modification of the *crossval.ado* **Stata** macro written by Rick Thompson of the Johns Hopkins Biostatistics Consulting Center. The macro file can be downloaded from the “Classes” page of the course website. For use, it must be placed in the current folder for the **Stata** session or in the personal “ADO” folder

— The macro also calculates cross-validated goodness-of-fit tests -- see below for results

4.13 Split sample cross-validation - FYI (cont'd)

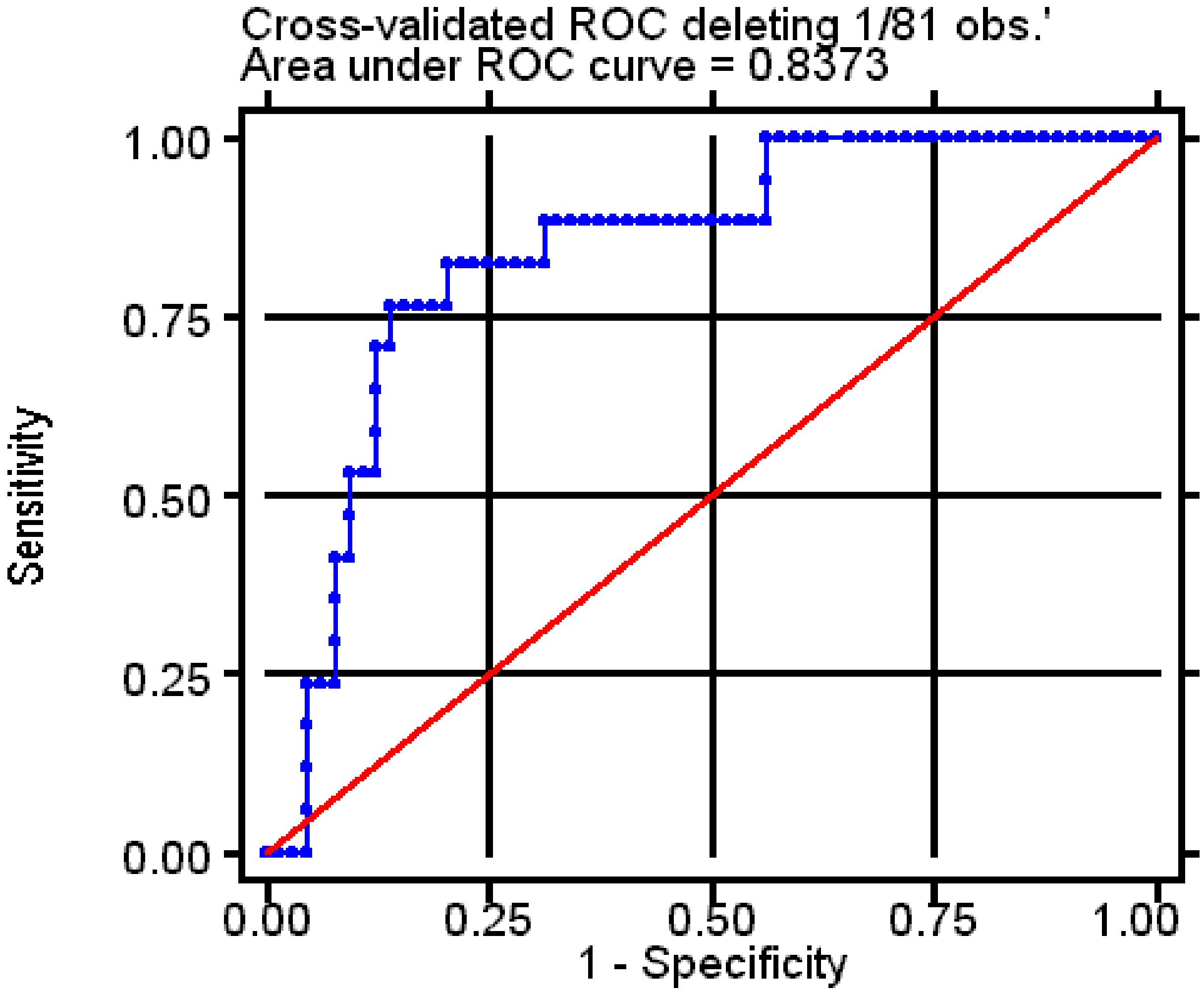
— RESULTS:

Cross-validated ROC curve for the model:

kyphosis = age agep startp

```
.
. crossval kyphosis age agep startp, numgrp(81)


CROSS-VALIDATED ROC CURVE DELETING 1/81 OBS.:

Logistic model for kyphosis

number of observations =       81
area under ROC curve   =   0.8373


CROSS-VALIDATED GOODNESS-OF-FIT DELETING 1/81 OBS.:

Logistic model for kyphosis, goodness-of-fit test

       number of observations =        81
 number of covariate patterns =        80
             Pearson chi2(79) =        91.30
                  Prob > chi2 =         0.1625



CROSS-VALIDATED HOSMER-LEMESHOW GOODNESS-OF-FIT DELETING 1/81 OBS.:

Logistic model for kyphosis, goodness-of-fit test
(Table collapsed on quantiles of estimated probabilities)

       number of observations =        81
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        12.19
                  Prob > chi2 =         0.1431
```

4.13 Split sample cross-validation - FYI (cont'd)



— The area under the cross-validated ROC curve is .84 compared with the somewhat more optimistic .89 from the earlier ROC analysis, which fit and tested the model with the full set of observed data

5. Stata do-file script: cl7ex1.do

- A complete **Stata** do-file script for this example is included below and is posted on the course website:

  cl7ex1.do

- The data (in raw data format) are also posted on the website:

  kyphosis.dat

- The example requires two **Stata** macros (*ado* files), which must be downloaded form the course website and placed either in the working folder (where the two above files are) or in your personal ADO folder (wherever that may be)

crossval.ado	Modification of Rick Thompson's macro for cross-validating logistic models via ROCs and goodness-of-fit tests
mlfit.ado	Calculates *AIC*s after a regression fit

```
version 7.0

*  Cl7EX1.DO  Logistic Regression
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
*  Kyphosis Data from S-PLUS system
*
*
*  Raw data: kyphosis.dat




* Assumes files are in folder  [path]\bio623

*    If files are in another folder, change cd  command below to
*      point Stata to the correct folder

* To run this program,  use the following Stata commands:

*          cd [path]\bio623    ... change directory to folder bio623

*          do  cl7ex1





* OUTLINE:


* Part a.  Input and display data, calculate new variables
* Part b.  Scatterplot Matrix + ksm lowess smoothed curves in a single image
* Part c.  Boxplots by Kyphosis  -- put into one image

* Part d.  Non-linearity display -- logs odds vs quintiles of X
*                   put plots in one image

* Part e.  Fit Logistic Regression Model
* Part f.  Fit Alternative models, get Deviances using glm
* Part h.  Sensitivity, Specificity, ROC curves
* Part i.  Cross Validation with split samples
* Part j.  Check Fit using Hosmer-Lemeshow chi-square -- 10 groups
* Part k.  Calculations of Deviance and AIC



* Housekeeping

* Clear workspace
clear


* Turn off -more- pause
set more off
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
* Save log file on disk, use .txt so Notepad will open it

capture log close
log using cl7ex1.log, replace


* Make subfolder for graphs
shell md cl7ex1


* Extend linesize for log

set linesize 100




* Part a.  Input and display data, calculate new variables

infile obs str7 kyph age number start using kyphosis.dat

gen     kyphosis=1  if kyph=="present"

replace kyphosis=0  if kyph=="absent"


* Get means, medians, etc

codebook kyphosis age number start , tabulate(2)


* Center continuous Xs -- so beta0 has interpretation -- at medians, rounded

gen agec = age-84

gen startc = start-13

gen numberc = number -4

gen startp=(start-10)*(start>10)  if start~=.


* Make "plus" function for broken-arrow spline with a break at 84 months
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
gen agep = (age-84)*(age>84)  if age~=.




* Part b.  Scatterplot Matrix + ksm lowess smoothed curves in a single image


*  Allow for long lines split over several lines separated by semi-colons

#delimit ;

set textsize 150;


graph age number start kyphosis, matrix half jitter(5)
     symbol(.)connect(s) bands(4) t1("SCATTERPLOT MATRIX:  KYPHOSIS DATA");
#delimit cr;

gphprint , saving(cl7ex1\figb1.wmf,replace)

#delimit ;

set textsize 150;


ksm kyphosis age, lowess xlab ylab adjust sy(.i) jitter(5) bwidth(0.8)
        saving(cl7ex1\one.gph,replace) t1("KYPHOSIS -VS- AGE");
gphprint , saving(cl7ex1\figb2.wmf,replace) ;



ksm kyphosis number, lowess xlab ylab adjust sy(.i) jitter(5) bwidth(0.8)
        saving(cl7ex1\two.gph,replace) t1("KYPHOSIS -VS- NUMBER");
gphprint , saving(cl7ex1\figb3.wmf,replace) ;



ksm kyphosis start, lowess xlab ylab adjust sy(.i) jitter(5) bwidth(0.8)
        saving(cl7ex1\three.gph,replace)  t1("KYPHOSIS -VS- START");
gphprint , saving(cl7ex1\figb4.wmf,replace) ;



set textsize  170;

graph using cl7ex1\one.gph cl7ex1\two.gph cl7ex1\three.gph,
        saving(cl7ex1\figb.wmf,replace);
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
set textsize 150;

* Make a better graph of kyphosis by start -- use more jitter and label;
*         the axes more completely, including tick marks;

graph kyphosis start, jitter(10)
     symbol(o) xtick(0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18)
     xlabel (0 3 6 9 12 15 18)  ylabel(0 .5  1.0)
     l1(" ") l2("KYPHOSIS") b2("START")   t1("KYPHOSIS -VS- STARTING VERTEBRA");

gphprint , saving(figb5.wmf,replace);


#delimit cr;

set textsize 100





* Part c.  Boxplots by Kyphosis  -- put into one image


sort kyphosis

set textsize 150

graph age , by (kyphosis) box t2("AGE") ylab saving(cl7ex1\one.gph,replace)

gphprint , saving(cl7ex1\figc1.wmf,replace)


graph number, by (kyphosis) box t2("NUMBER") ylab saving(cl7ex1\two.gph,replace)

gphprint , saving(cl7ex1\figc2.wmf,replace)


graph start, by (kyphosis) box t2("START")   ylab saving(cl7ex1\three.gph,replace)
b2("KYPHOSIS")

gphprint , saving(cl7ex1\figc3.wmf,replace)


set textsize 170

graph using cl7ex1\one.gph cl7ex1\two.gph cl7ex1\three.gph,
saving(cl7ex1\figc.wmf,replace)

set textsize 100
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
* Part d.  Non-linearity display -- logs odds vs quintiles of X
*                   put plots in one image

set textsize 150

* Age



egen q1 = pctile(age), p(20)
egen q2 = pctile(age), p(40)
egen q3 = pctile(age), p(60)
egen q4 = pctile(age), p(80)
egen qmin = min(age)
egen qmax = max(age)

gen     quintile = (q4+qmax)/2  if age >  q4
replace quintile = (q3+q4  )/2  if age <= q4
replace quintile = (q2+q3  )/2  if age <= q3
replace quintile = (q1+q2  )/2  if age <= q2
replace quintile = (qmin+q1)/2  if age <= q1
replace quintile =.  if age==.


egen qcat = group(quintile)

tab quintile qcat

egen prop = mean(kyphosis), by (qcat)

gen lodds = log(prop/(1-prop))

graph lodds quintile , connect(l) symbol(O)  xlab ylab t1("AGE")  l1(" ") l2("LOG
ODDS") b2("QUINTILES") saving(cl7ex1\one.gph,replace)


gphprint , saving(cl7ex1\figd1.wmf,replace)

drop q1-q4 quintile qmin qmax qcat prop lodds


* Start

egen q1 = pctile(start), p(20)
egen q2 = pctile(start), p(40)
egen q3 = pctile(start), p(60)
egen q4 = pctile(start), p(80)
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
egen qmin = min(start)
egen qmax = max(start)

gen     quintile = (q4+qmax)/2  if start >  q4
replace quintile = (q3+q4  )/2  if start <= q4
replace quintile = (q2+q3  )/2  if start <= q3
replace quintile = (q1+q2  )/2  if start <= q2
replace quintile = (qmin+q1)/2  if start <= q1
replace quintile =.  if start==.


egen qcat = group(quintile)

tab quintile qcat

egen prop = mean(kyphosis), by (qcat)

gen lodds = log(prop/(1-prop))

graph lodds quintile , connect(l) symbol(O)  xlab ylab t1("START") l1(" ") l2("LOG
ODDS") b2("QUINTILES")  saving(cl7ex1\two.gph,replace)


gphprint , saving(cl7ex1\figd1.wmf,replace)

drop q1-q4 quintile qmin qmax qcat prop lodds



* Number

egen q1 = pctile(number), p(20)
egen q2 = pctile(number), p(40)
egen q3 = pctile(number), p(60)
egen q4 = pctile(number), p(80)
egen qmin = min(number)
egen qmax = max(number)

gen     quintile = (q4+qmax)/2  if number >  q4
replace quintile = (q3+q4  )/2  if number <= q4
replace quintile = (q2+q3  )/2  if number <= q3
replace quintile = (q1+q2  )/2  if number <= q2
replace quintile = (qmin+q1)/2  if number <= q1
replace quintile =.  if start==.


egen qcat = group(quintile)

tab quintile qcat

egen prop = mean(kyphosis), by (qcat)
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
gen lodds = log(prop/(1-prop))

graph lodds quintile , connect(l) symbol(O)  xlab ylab t1("NUMBER") l1(" ") l2("LOG
ODDS") b2("QUINTILES") saving(cl7ex1\three.gph,replace)


gphprint , saving(cl7ex1\figd1.wmf,replace)


graph using cl7ex1\one.gph cl7ex1\two.gph cl7ex1\three.gph,
saving(cl7ex1\figd.wmf,replace)

set textsize 100




* Part e.  Fit Logistic Regression Model

logit kyphosis agec agep startc numberc


* Estimate beta1+beta2  - slope after age 7

lincom agec+agep

*  Get Odds Ratios

logistic kyphosis agec agep startc numberc
lincom agec+agep



* Verify global test for Xs=0 fitting null vs extended models

quietly logit kyphosis age agep startc numberc
lrtest, saving(0)

quietly logit kyphosis
lrtest



* Display correlation matrix for estimates -- rarely done/needed
*     -- show correlations among betas for verifying beta1+beta2 estimate

quietly logit kyphosis age agep startc numberc
vce , corr
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
* Part f.  Fit Alternative models, get Deviances using glm


* Model with intercept only -- to confirm global test

logit kyphosis

quietly glm  kyphosis , family(binomial) link(logit)
display "Deviance:  " e(deviance_s) "   d.f.:  " e(df) "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model with all Xs:  age agep startc numberc

logit kyphosis agec agep startc numberc

quietly glm  kyphosis agec agep startc numberc, family(binomial) link(logit)
display "Deviance:  " e(deviance_s) "   d.f.:  " e(df) "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model A:  agec agep

logistic kyphosis agec agep
logit

quietly glm  kyphosis agec agep, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model B:  start
logistic kyphosis startc
logit

quietly glm  kyphosis startc, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model C:  numberc
logistic kyphosis numberc
logit

quietly glm  kyphosis numberc, family(binomial) link(logit)
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model D:  age agep startc
logistic kyphosis agec agep startc
logit

quietly glm  kyphosis agec agep startc, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model E:  age agep numberc
logistic kyphosis agec agep numberc
logit

quietly glm  kyphosis agec agep numberc, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model F:  age agep startp


logistic kyphosis agec agep startp
logit

quietly glm  kyphosis agec agep startp, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model G: age agep startc startp

replace startc=(start-10)  if start~=.

logistic kyphosis agec agep startc startp
logit


quietly glm  kyphosis agec agep startc startp, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))



* Model H: age agep startc numberc
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
logistic kyphosis agec agep startc numberc
logit

quietly glm  kyphosis agec agep startc numberc, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)  "   P-Value:  "
chiprob(e(df),e(deviance_s))




* Part g.  Graph fitted -vs- observed

quietly logistic kyphosis age agep startp

lpredict muhat


* Lowess smoothed plot

set textsize 150

ksm kyphosis muhat, lowess ylab xlab s(.i) bwidth(0.8) t1("KYPHOSIS -VS- PREDICTED
PROBS")

gphprint , saving(cl7ex1\figg1.wmf,replace)



* Plot deciles fitted -vs- observed

drop q1-q4 quintile qmin qmax qcat prop lodds


* muhat

egen q1 = pctile(muhat), p(10)
egen q2 = pctile(muhat), p(20)
egen q3 = pctile(muhat), p(30)
egen q4 = pctile(muhat), p(40)
egen q5 = pctile(muhat), p(50)
egen q6 = pctile(muhat), p(60)
egen q7 = pctile(muhat), p(70)
egen q8 = pctile(muhat), p(80)
egen q9 = pctile(muhat), p(90)
egen qmin = min(muhat)
egen qmax = max(muhat)

gen     decile = (q9+qmax)/2  if muhat >  q9
replace decile = (q8+q9  )/2  if muhat <= q9
replace decile = (q7+q8  )/2  if muhat <= q8
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
replace decile = (q6+q7  )/2  if muhat <= q7
replace decile = (q5+q6  )/2  if muhat <= q6
replace decile = (q4+q5  )/2  if muhat <= q5
replace decile = (q3+q4  )/2  if muhat <= q4
replace decile = (q2+q3  )/2  if muhat <= q3
replace decile = (q1+q2  )/2  if muhat <= q2
replace decile = (qmin+q1)/2  if muhat <= q1
replace decile =.  if muhat==.


egen deccat = group(decile)

tab decile deccat

egen prop = mean(kyphosis), by (deccat)


* Plot obs -vs- pred  and y=x for comparison with perfect fit

graph prop decile decile , connect(ll) symbol(Oi)  xlab ylab t1("OBSERVED -VS-
PREDICTED PROPORTIONS --  DECILES")


gphprint , saving(cl7ex1\figg2.wmf,replace)


set textsize 100





* Part h.  Sensitivity, Specificity, ROC curves





set textsize 125

* Model with age agep startc numberc

quietly logit kyphosis age agep startc numberc

* Sensitivity, Specificity, etc for a give cut point

lstat , cutoff(0.5)

lstat , cutoff(0.3)
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
* Plot Sensitivity -vs- Specificity -- all possible cut points

lsens , t1("MODEL:  AGE AGEP STARTC NUMBERC") l1(".") l2("Sensitivity / Specificity")

gphprint , saving(cl7ex1\figh1.wmf,replace)


* ROC CURVES

lroc , t1("MODEL:  AGEC AGEP STARTC NUMBERC") l1(" ") l2("Sensitivity")

gphprint , saving(cl7ex1\figh2.wmf,replace)


* Better model: double broken-arrow model - age agep startp

quietly logit kyphosis age agep startp

lroc , t1("MODEL:  AGE AGEP STARTP") l1(" ") l2("Sensitivity")

gphprint , saving(cl7ex1\figh3.wmf,replace)




* Part i.  Cross Validation with split samples

*  Use modification of Rick Thompson's macro:  crossval.ado
*     -- can be downloaded from course website

*  Put crossval.ado in the current folder or in the personal ADO folder

*  Note: Parameter numgrps = k , where 1/k = random fraction of data deleted
*          for cross-validation. k=2 splits data in random halves.
*        Recommend setting k=n, which deletes one observation at a time

crossval kyphosis age agep startp, numgrp(81)




* Part j.  Check Fit using Hosmer-Lemeshow chi-square -- 10 groups

quietly logit kyphosis age agep startc numberc
lfit , group(10)
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
quietly logit kyphosis age agep startp
lfit , group(10)


set textsize 100





* Part k.  Calculations of Deviance and AIC


* Generate 4 new "predictors", all containing random numbers

set seed 568123457

gen x5 = uniform()
gen x6 = uniform()
gen x7 = uniform()
gen x8 = uniform()



* Get Deviances and AICs (uses mlfit.ado macro -- downloadable from website)

quietly glm  kyphosis agec agep startc numberc , family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)

quietly logit kyphosis agec agep startc numberc
mlfit


quietly glm  kyphosis agec agep startc numberc x5, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)

quietly logit kyphosis agec agep startc numberc x5
mlfit


quietly glm  kyphosis agec agep startc numberc x5 x6, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)

quietly logit kyphosis agec agep startc numberc x5 x6
mlfit



quietly glm  kyphosis agec agep startc numberc x5 x6 x7, family(binomial) link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)
```

5. Stata do-file script: cl7ex1.do (cont'd)

```
quietly logit kyphosis agec agep startc numberc x5 x6 x7
mlfit


quietly glm  kyphosis agec agep startc numberc x5 x6 x7 x8, family(binomial)
link(logit)
display "Deviance:  " e(deviance_s)  "   d.f.:  " e(df)

quietly logit kyphosis agec agep startc numberc x5 x6 x7 x8
mlfit



* Close log file -- Only when all errors have been fixed

*log close
```