

### Logistic Regression (continued)

- Recall that logistic regression provides a way of relating predictor variables  $X$  to binary (0,1) outcome variable  $Y$ . This outcome variable is an indicator for "yes/no", "event/non-event", etc
- The model actually predicts  $\Pr(Y = 1|X)$ . This equals the expected value (mean) of  $Y$ , as in linear regression. How?
  - Note that for a binary variable, the mean of  $Y$  and covariate  $X$ ,  $\sum_{j=1}^{N_i} Y_i / N_i = p_i$ , or the proportion that equal 1.
  - For a binomial random variable (independent draws, probability  $p$  of success for each draw), this proportion equaling 1 is the same as  $\Pr(Y = 1)$

## Logistic Regression

- The model for  $\Pr(Y = 1|X)$  is nonlinear. We model linearly via the logit transform:

$$\log \left\{ \frac{p}{1-p} \right\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_q X_q$$

- $\beta_j$  in the model is the difference in log odds for a one unit increment in  $X_j$ . This is same as the **log odds ratio**.  
Exponentiating gives the **odds ratio**, reflecting the relative odds of event (being a '1') for a one-unit increment in  $X$ . Note that this is a multiplicative difference (rather than additive as in difference of means)
- Ex:  $\beta_{age} = 0.336$ . Then  $\exp(0.336) = 1.4$ . For a one-year increase in age, relative odds of event goes up by 1.4, or 40%

## Logistic Regression

- The model can produce a probability of event for each case in the dataset, based on the predictors  $X$ , via the equation

$$\Pr(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \dots + \beta_q X_q)}{1 + \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_q X_q)}$$

- With these values, one can assess how well individual cases may be classified with respect to outcome (will discuss shortly)

## Logistic Regression - another example model coefficient interpretation

- **Glioblastoma (GBM):** A randomized clinical trial was conducted comparing standard treatment (radiation + chemotherapy) to this treatment plus bevacizumab (Avastin) with the goal of improving two endpoints:
  - progression-free survival (time remaining alive and free of disease progression) and
  - survival time
- Here, we analyze PFS at 18 months from study entry
- While prognosis is generally unfavorable, important new tumor molecular markers are strongly associated w/outcomes and possibly treatment response. O(6)-methylguanine-DNA methyltransferase (MGMT) inactivation via methylation is also examined

## Logistic Regression model coefficient interpretation

Examining effects individually via tables:

```
. cc fail_18mo trt
```

	Exposed	Unexposed	Total	Proportion exposed
Cases	225	235	460	0.4891
Controls	87	74	161	0.5404
Total	312	309	621	0.5024

	Point estimate	[95% Conf. Interval]
Odds ratio	.81438	.5587526 1.185748 (exact)

Modest treatment effect - 19% lower odds of failure

```
. cc fail_18mo  mgmt_grp
```

			Proportion	
	Exposed	Unexposed	Total	exposed
Cases	354	106	460	0.7696
Controls	92	69	161	0.5714
Total	446	175	621	0.7182
	Point estimate		[95% Conf. Interval]	
Odds ratio	2.504717		1.678978 3.724206 (exact)	

MGMT unmethylated have 2.5 times greater odds of failure

## Logistic Regression model coefficient interpretation

### The model run:

```
. logit fail_18mo trt mgmt_grp
```

```
Iteration 0:   log likelihood = -355.38631
```

```
. . .
```

```
Iteration 3:   log likelihood = -343.79561
```

Logistic regression	Number of obs	=	621
	LR chi2(2)	=	23.18
	Prob > chi2	=	0.0000
Log likelihood = -343.79561	Pseudo R2	=	0.0326

fail_18mo	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
-----+-----					
trt	-.1996488	.1870863	-1.07	0.286	-.5663311 .1670335
mgmt_grp	.9167454	.1941661	4.72	0.000	.5361868 1.297304
_cons	.5330665	.1833776	2.91	0.004	.1736529 .8924801

**Adjusted ORs:** trt OR:  $e^{-.19964} = 0.819$ . MGMT OR:  $e^{.91674} = 2.50$

**Logistic Regression**  
**model coefficient interpretation**

We can generate 3 odds ratios relative to the reference group [no bev, MGMT methylate]d:

<b>MGMT Status</b>	<b>Treatment</b>	
	Placebo	Bevac
methylated	1.00	0.82
unmethylated	2.51	2.04

**Note that  $2.51 \times 0.82 = 2.04$ , i.e., the treatment effect is the same in both MGMT groups according to this model**



## Logistic Regression - another model

**We can add a treatment by MGMT interaction term to the model, permitting different trt effects in each MGMT group:**

```
. logit fail_18mo trt mgmt_grp trt_by_mgmt
Iteration 0:   log likelihood = -355.38631
I . .
Iteration 3:   log likelihood =  -342.3045
Logistic regression
```

	Number of obs	=	621
	LR chi2(3)	=	26.16
	Prob > chi2	=	0.0000
Log likelihood = -342.3045	Pseudo R2	=	0.0368

```
-----+-----
```

fail_18mo	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
trt	-.6310748	.3145175	-2.01	0.045	-1.247518	-.0146317
mgmt_grp	.561396	.2849644	1.97	0.049	.002876	1.119916
trt_by_mgmt	.6745599	.3920708	1.72	0.085	-.0938848	1.443005
_cons	.7646061	.2329773	3.28	0.001	.3079791	1.221233
-----+-----						

**There is a suggestion of interaction between treatment and MGMT status**

## Logistic Regression - another model

We again generate 3 odds ratios relative to [no bev, MGMT methylated]:

MGMT Status	Treatment	
	Placebo	Bevac
methylated	1.00	0.53
unmethylated	1.75	1.83

### Now note that

- Bev effect in methylated is larger, 47% reduction in odds of failure
- No Bevt effect in unmethylated, which we can see by computing  $1.83/1.75$  equals  $OR = 1.045$ . Another way to obtain is  $e^{-.6311+.6745} = 1.045$ . We can check these effects by estimating treatment effects separately by MGMT status.

## Logistic Regression - Stratified Treatment Effect

. cc fail\_18mo trt if mgmt\_grp==0

	Exposed	Unexposed	Total	Proportion exposed
Cases	48	58	106	0.4528
Controls	42	27	69	0.6087
Total	90	85	175	0.5143
	Point estimate		[95% Conf. Interval]	
Odds ratio	.5320197		.2736227 1.029957 (exact)	

. cc fail\_18mo trt if mgmt\_grp==1

	Exposed	Unexposed	Total	Proportion exposed
Cases	177	177	354	0.5000
Controls	45	47	92	0.4891
Total	222	224	446	0.4978
	Point estimate		[95% Conf. Interval]	
Odds ratio	1.044444		.6430068 1.697515 (exact)	

## Logistic Regression - Model Significance & Goodness of Fit

- As we discussed, a *likelihood ratio* test can be defined contrasting the model in question with a null model containing no predictors. This is analogous to the global F-test in SLR/MLR

- The test statistic:

$$D_{global} = -2(\log\text{-likelihood}_{null} - \log\text{-likelihood}_{full})$$

is  $\chi^2_{df}$  where degrees of freedom  $df = \text{number of predictors}$

- This same type of test is used for contrasting two nested models, say, dropping out 3 predictors out of 7 candidates

$$D_{nested} = -2(\log\text{-likelihood}_{smaller\ model} - \log\text{-likelihood}_{bigger\ model})$$

is  $\chi^2_3$  (3 df) and tests whether the three parameters considered for dropping simultaneously have  $\beta_j = 0$

## **Logistic Regression - Quantifying Fit**

- The global LR test evaluates whether there is a worthwhile model relative to no model at all. What about a test against 'the best model available', or a perfect fit?
- In the case of discrete predictors (where tables can be formed), such as test is available and can give a sense of what is the model that predicts best with the least number of predictors.

## - Model Goodness of Fit - Deviance: notation and definition

The quantity  $D$  above is known as the deviance and can be used to assess model fit: define

1.  $\hat{L}_c$ : the maximized likelihood (likelihood given the MLE) under the current model of interest
2.  $\hat{L}_f$ : the maximized likelihood under the model fits the data perfectly, which is termed the *full* or *saturated model*

**Then** the *Deviance*:  $D = -2\log(\hat{L}_c/\hat{L}_f) = -2\{\log\hat{L}_c - \log\hat{L}_f\}$

- $D$  measures the extent to which the current model deviates from the full model:
- Large  $D$  when  $\hat{L}_c$  is small relative to  $\hat{L}_f$ , indicating the current model fit is poor.
- Small  $D$  when  $\hat{L}_c$  is similar to  $\hat{L}_f$ , indicating 'good' current model fit.

## - Model Goodness of Fit - Deviance: notation and definition

### Deviance: formula

- Recall: the likelihood function based on observations  $y_i/n_i$ , for  $i = 1, 2, \dots, n$  groups defined by unique covariate combinations, with unknown  $p_i$  is

$$L = \prod_{i=1}^n \binom{n_i}{y_i} p_i^{y_i} (1 - p_i)^{n_i - y_i}$$

$$\log L = \sum_{i=1}^n \left\{ \log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i) \log(1 - p_i) \right\} \quad (1)$$

- Let  $\hat{p}_i$ ,  $i = 1, \dots, n$  be the fitted values under *current model*, then

$$\log \hat{L}_c = \sum_{i=1}^n \left\{ \log \binom{n_i}{y_i} + y_i \log \hat{p}_i + (n_i - y_i) \log(1 - \hat{p}_i) \right\} \quad (2)$$

- Define  $\tilde{p}_i = y_i/n_i$ ,  $i = 1, \dots, n$  which are the actual probabilities in

these table cells. We define this as the *full model*:

$$\log \hat{L}_f = \sum_{i=1}^n \left\{ \log \binom{n_i}{y_i} + y_i \log \tilde{p}_i + (n_i - y_i) \log (1 - \tilde{p}_i) \right\} \quad (3)$$

- The *Deviance* is then given by

$$\begin{aligned} D &= -2 \{ \log \hat{L}_c - \log \hat{L}_f \} \\ &= 2 \sum_{i=1}^n \left\{ y_i \log \left( \frac{\tilde{p}_i}{\hat{p}_i} \right) + (n_i - y_i) \log \left( \frac{1 - \tilde{p}_i}{1 - \hat{p}_i} \right) \right\} \end{aligned} \quad (4)$$



### - Model Goodness of Fit - Deviance Statistic

- To use the deviance statistic, we have to have the following circumstances for the covariate set  $\mathbf{X}$ :
  - We assume there is a model that reproduces the probability of event at a given  $X$  combination perfectly (from a table of response by  $X$  category, the  $\tilde{p}$  from eqn. 4 above), and ...
  - there are model(s) with fewer parameters (for example, main effects only for  $\mathbf{X}$ , or a given  $X$  represented as a numeric index or on a continuous scale).
- Then, we can contrast models via estimated  $\hat{p}$  and  $\tilde{p}$

## Logistic Regression - D as Goodness of Fit Measure

### Deviance: distribution of the deviance statistic $D$

- We use  $D$  to evaluate the current model, and so we need to know its distribution. Under  $H_0$  that the *current model* requires no additional parameters to fit well, as the groups size  $n_i \rightarrow \infty$ ,  $D$  converges to  $\chi^2_{n-p}$ , where  $p = \#$  parameters fit,  $n = \#$  of groups.
- Thus, for reasonably large groups, the deviance provides a goodness of fit test for the model, and  $D \sim \chi^2_{n-p}$  approximately. In this case,
  - The value of deviance statistic,  $D$ , can be compared to tables of percentage points of the  $\chi^2_{n-p}$  distribution. Note - here we expect a large statistic (small p-value) under bad fit.
  - Since mean of a  $\chi^2_{n-p}$  r.v. is  $n - p$ , a useful rule of thumb is the  $D$  does not exceed  $n - p$ , the model may be satisfactory.

## Logistic Regression - D as Goodness of Fit Measure

**Deviance: distribution (continued): Case of 'continuous' X values:**

- For the case where X is (in practical terms) on a continuous scale, so that for each X value there is only one response (0/1) for Y (and thus no natural way to form a table),  $\log \hat{L}_f = 0$ , and  $D$  depends only on the fitted success  $\hat{p}_i$ :  
$$D = -2 \sum_{i=1}^n \{ \hat{p}_i \text{logit}(\hat{p}_i) + \log(1 - \hat{p}_i) \}.$$
 In this case, the deviance is uninformative about the goodness of fit of a model.
- In this case where  $n_i = 1$ , for all  $i$ ,  $D$  is not even approximately  $\chi^2$ .
- Even when the  $n_i$  all exceed unity (i.e, we have 'tables'), the  $\chi^2$  approximation may not be particularly good when the data are sparse, i.e., some of the binomial denominators  $n_i$  are very small.

We will look at an alternative test for this case a bit later . . .

## **Logistic Regression - D as Goodness of Fit Measure**

**Example: Donner Party:** In 1846 the Donner and Reed families left Springfield, Illinois, for California by covered wagon. In July, the Donner Party, as it became known, reached Fort Bridger, Wyoming. There, 87 people and 20 wagons crossed the Wasatch Range and the desert west of the Great Salt Lake and entering the eastern Sierra Nevada mountains in late October (later than planned). Heavy snows stranded the group in the mountains, and by the time of complete rescue in April, 1847, 40 of the 87 members had died from famine and exposure.

What factors were related to survival?

## Example - D as Goodness of Fit Measure

### Listing of part of the data for 45 individuals

```
list AGE SEX STATUS in 5/20, noobs clean
```

AGE	SEX	STATUS
15	MALE	DIED
25	MALE	SURVIVED
25	MALE	DIED
22	FEMALE	SURVIVED
20	MALE	SURVIVED
24	FEMALE	SURVIVED
25	MALE	DIED
25	FEMALE	DIED
15	FEMALE	SURVIVED
18	MALE	SURVIVED
25	FEMALE	SURVIVED
23	FEMALE	SURVIVED
28	MALE	SURVIVED
28	MALE	SURVIVED
21	FEMALE	SURVIVED
28	MALE	DIED

## Example - D as Goodness of Fit Measure

We form a discrete age categorical variable for illustrative purposes, classifying individuals as under 25 years or 25 and over.

The model:

```
. logistic died sex age25plus
```

```
Logistic regression               Number of obs   =           45
                                LR chi2(2)        =           8.09
                                Prob > chi2        =          0.0175
Log likelihood = -26.868228       Pseudo R2      =          0.1309
```

-----+-----							
died	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]		
-----+-----							
sex	3.597994	2.515823	1.83	0.067	.9138505	14.16595	
age25plus	3.84985	2.832213	1.83	0.067	.9104196	16.27968	
_cons	.2056279	.1604821	-2.03	0.043	.0445413	.9492959	
-----							

Note: \_cons estimates baseline odds.

We have two predictors of mortality: males and those over 25 have over 3-fold greater odds of death.

## Example - D as Goodness of Fit Measure

A full or saturated model here would include the sex by age group interaction, which will reproduce exactly the probabilities for the sex by age group by died/survived 3-way table:

```
. logit died sex age25plus agebysex
. . .
Iteration 10:  log likelihood = -25.096587
```

Logistic regression	Number of obs	=	45
	LR chi2(3)	=	11.63
	Prob > chi2	=	0.0088
Log likelihood = -25.096587	Pseudo R2	=	0.1882

died	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sex	17.63652	2388.794	0.01	0.994	-4664.313	4699.586
age25plus	17.57199	2388.794	0.01	0.994	-4664.378	4699.522
agebysex	-17.03307	2388.794	-0.01	0.994	-4698.983	4664.917
_cons	-17.34885	2388.794	-0.01	0.994	-4699.298	4664.6

### Example - D as Goodness of Fit Measure

This coefficient estimates 'go crazy' since the (female, under 25) group has zero deaths, but the model accurately reproduces the observed probabilities of death in the 4 unique groups. The log likelihood statistic (the part we need (-25.096587) ) is provided.

The Deviance statistic  $D = -2\{\log\hat{L}_{current} - \log\hat{L}_{full}\}$  is then

$$D = -2\{-26.868228 - (-25.096587)\} = 3.5432$$

The associated p-value for a  $\chi_1^2$  distribution is 0.06. So, the model with two predictors passes as a possibly suitable model relative to the 'perfect' model



### Example - D as Goodness of Fit Measure

To do this in Stata, need to install a module called *ldev* and run after the model of interest (here, the main effects model)

```
. ldev
```

```
Logistic model deviance goodness-of-fit test
```

number of observations =	45
number of covariate patterns =	4
deviance goodness-of-fit =	3.54
degrees of freedom =	1
Prob > chi2 =	0.0598

## Goodness of Fit Measures - continuous covariates

- What if we have continuous predictors (no tables or very sparse tables)?
- the Hosmer-Lemeshow test goodness of fit measure in logistic regression groups the  $n$  observations into groups (according to their estimated probability of event) and calculates the corresponding generalized Pearson  $\chi^2$  statistic. Usually deciles (10 groups) are used.

$$H = \sum_{g=1}^G \frac{(O_g - E_g)^2}{n_g(\hat{p}_g(1 - \hat{p}_g))}$$

Where  $O_g$  is the number of events in group  $g$  and  $E_g = n_g * \hat{p}_g$  is the expected number of events

- In this type of test, a *large* p-value ( $> 0.05$ ) indicates good correspondence between observed and predicted outcomes.

## Goodness of Fit Measures - HL Test

The test can also be run on discrete data - Donner party data

```
. estat gof
```

Logistic model for died, goodness-of-fit test

number of observations =	45
number of covariate patterns =	28
Pearson chi2(25) =	24.35
Prob > chi2 =	0.4991

## Logistic Regression - Model Significance and Goodness of Fit Measures

```
. logit sta typ age
```

```
Iteration 0:   log likelihood = -100.08048
Iteration 1:   log likelihood = -87.895217
. . .
Iteration 5:   log likelihood = -86.537821
```

```
Logistic regression               Number of obs   =           200
                                LR chi2(2)        =           27.09
                                Prob > chi2        =           0.0000
Log likelihood = -86.537821       Pseudo R2       =           0.1353
```

-----						
sta	Coef.	Std. Err.	z	P> z	[95\% Conf. Interval]	
-----+-----						
typ	2.453535	.75257	3.26	0.001	.978525	3.928545
age	.0340162	.0106944	3.18	0.001	.0130556	.0549767
_cons	-5.508762	1.033511	-5.33	0.000	-7.534407	-3.483118
-----						

```
. estat gof, group(10) table
```

Logistic model for sta, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

+-----+						
Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
+-----+						
1	0.0362	0	0.5	20	19.5	20
2	0.0478	1	0.9	20	20.1	21
3	0.0812	1	1.2	18	17.8	19
4	0.1209	2	1.8	18	18.2	20
5	0.1916	2	3.1	18	16.9	20
+-----+						
6	0.2531	7	4.6	14	16.4	21
7	0.2936	6	6.1	16	15.9	22
8	0.3376	5	5.8	13	12.2	18
9	0.3846	8	7.3	12	12.7	20
10	0.5186	8	8.7	11	10.3	19
+-----+						

number of observations = 200  
 number of groups = 10  
 Hosmer-Lemeshow chi2(8) = 2.95  
 Prob > chi2 = 0.9372

## Logistic Regression - Goodness of Fit Measures

This result looks very positive - good fit. However,  $\text{Pseudo-}R^2 = 13\%$ , so not that much variation explained.

$\text{Pseudo-}R^2$  measures proportion reduction in log-likelihood over null model. This is a useful measure, but more like another F-test than a measure of model explanatory power.

The Hosmer-Lemeshow g.o.f. test is more valuable as a means to identify major systematic variation that is not explained. Large p-value does not assure that prediction will be highly accurate, etc.

## Logistic Regression - Goodness of Fit for the Donner Data

With continuous age predictor and sex:

```
. logistic died sex AGE
```

```
Logistic regression               Number of obs   =           45
                                LR chi2(2)        =           10.57
                                Prob > chi2        =           0.0051
Log likelihood = -25.628142       Pseudo R2      =           0.1710
```

-----						
died	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
sex	4.939645	3.731909	2.11	0.034	1.1236	21.71599
AGE	1.081343	.0403206	2.10	0.036	1.005135	1.16333
_cons	.0395411	.0548425	-2.33	0.020	.0026089	.5992974
-----						

Note: \_cons estimates baseline odds.

```
. estat gof, group(10) table
```

Logistic model for died, goodness-of-fit test

(Table collapsed on quantiles of estimated probabilities)

+-----+						
Group	Prob	Obs_1	Exp_1	Obs_0	Exp_0	Total
+-----+						
1	0.1928	0	0.8	5	4.2	5
2	0.3257	1	1.3	4	3.7	5
3	0.4827	1	1.8	3	2.2	4
4	0.5662	3	2.2	1	1.8	4
5	0.5798	7	4.6	1	3.4	8
+-----+						
6	0.6226	1	0.6	0	0.4	1
7	0.6637	3	3.2	2	1.8	5
8	0.6878	3	2.7	1	1.3	4
9	0.8770	2	4.0	3	1.0	5
10	0.9692	4	3.8	0	0.2	4
+-----+						

```

number of observations =      45
      number of groups =      10
Hosmer-Lemeshow chi2(8) =     10.94
          Prob > chi2 =      0.2049

```



## Logistic Regression - Predictive Ability

Since the model generates  $\Pr(Y = 1)$  for all cases, we can assess how well we might predict failure or success with the model. Specifically:

- We can simply assess how many correct and incorrect for a given  $\Pr(Y = 1)$  criterion, such as 0.5. Ordinary sensitivity/specificity type calculations can be used to assess the result
- We can try to select a probability value that optimizes the above parameters. Note that as in sensitivity/specificity problems, there is a trade-off between predicting cases and non-cases correctly

## Logistic Regression - Example of Predictive Ability

Glioblastoma (GBM) randomized clinical trial comparing standard treatment (radiation + chemotherapy) to same plus bevacizumab (Avastin), Analysis of survival

We will additionally include MGMT and composite prognostic indicator known as *RPA class*, as predictors

(Gilbert et al *NEJM* 2014)

## Logistic Regression - Predictive Ability

- GBM data:

```
.* TRT effect = exposed here is trt = 1 (bev), use cohort with MGMT
. cc survival trt if cohort==1
```

	Exposed	Unexposed	Total	Proportion Exposed
Cases	210	193	403	0.5211
Controls	97	108	205	0.4732
Total	307	301	608	0.5049
	Point estimate		[95% Conf. Interval]	
Odds ratio	1.211474		.8530073	1.721077 (exact)
Attr. frac. ex.	.1745591		-.1723229	.4189684 (exact)
Attr. frac. pop	.0909613			
chi2(1) = 1.25 Pr>chi2 = 0.2639				

```

.* MGMT effect - exposed here is bad marker value (=1)
. cc survival mgmt_grp if cohort==1

```

	Exposed	Unexposed	Total	Proportion Exposed
Cases	321	82	403	0.7965
Controls	115	90	205	0.5610
Total	436	172	608	0.7171
	Point estimate		[95% Conf. Interval]	
Odds ratio	3.063627		2.085235	4.496513 (exact)
Attr. frac. ex.	.6735895		.5204378	.7776055 (exact)

chi2(1) = 37.16 Pr>chi2 = 0.0000

- No treatment benefit for survival (trt/control OR = 1.21, indicating 21% *greater odds* of failure on Avastin, although not significantly different from 1.0). MGMT- unmethylated patients have 3-fold greater risk of death. In GBM, this marker represents significant heterogeneity in prognosis

## Logistic Regression - Predictive Ability

Examine treatment, MGMT, and RPA class (a baseline prognostic class indicator) together via logistic regression

```
. logit survival trt mgmt_grp rpa_class, or
```

```
Iteration 0:    log likelihood = -388.59785
```

```
Iteration 1:    log likelihood = -361.36406
```

```
. . .
```

```
Iteration 4:    log likelihood = -361.10958
```

Logistic regression

Number of obs = 608

LR chi2(3) = 54.98

Prob > chi2 = 0.0000

Log likelihood = -361.10958

Pseudo R2 = 0.0707

-----						
survival	Odds Ratio	Std. Err.	z	P> z	[95% Conf. Interval]	
-----+-----						
trt	1.253899	.2259569	1.26	0.209	.8807916	1.785057
mgmt_grp	3.354802	.6471911	6.27	0.000	2.298569	4.896394
rpa_class	1.967377	.3278095	4.06	0.000	1.419248	2.727201
_cons	.0507661	.0358217	-4.22	0.000	.0127335	.2023957
-----						

- Predict probabilities and assess

```
. predict phat
. list trt mgmt_grp rpa_class survival phat in 20/40, clean
```

	trt	mgmt_grp	rpa_cl~s	survival	phat
20.	1	1	3	1	.6192176
21.	1	1	5	1	.8629051
22.	1	1	4	0	.7618647
23.	1	1	4	0	.7618647
24.	0	1	5	1	.8338792
25.	1	1	3	0	.6192176
. . .					
34.	1	1	4	0	.7618647
35.	1	0	4	0	.4881366
36.	1	1	3	1	.6192176
37.	1	1	4	1	.7618647
38.	0	1	3	1	.5646291
39.	1	0	3	0	.3264767
40.	1	1	4	1	.7618647

One can 'assign' all those with, say,  $phat > .5$ , to failure, and compare to actual failure outcome

- **A summary function to do this is provided in Stata**

```
. estat classification
```

```
Logistic model for survival
```

		----- True -----		
Classified		D	~D	Total
-----+-----+-----				
+		340	131	471
-		63	74	137
-----+-----+-----				
Total		403	205	608

```
Classified + if predicted Pr(D) >= .5
```

```
True D defined as survival != 0
```

-----		
Sensitivity	Pr( +  D)	84.37%
Specificity	Pr( - ~D)	36.10%
Positive predictive value	Pr( D  +)	72.19%
Negative predictive value	Pr(~D  -)	54.01%
-----		
Correctly classified		68.09%
-----		

- **The above rule uses 0.5 as a classification probability.** What about other values? List the predicted probabilities.

```
. tab phat
```

Pr(survival)	Freq.	Percent	Cumul.
-----+-----			
.2788	9	1.48	1.48
.3264767	11	1.81	3.29
.4319939	56	9.21	12.50
.4881366	61	10.03	22.53
.5646291	38	6.25	28.78
.5994036	17	2.80	31.58
.6192176	25	4.11	35.69
.6523169	18	2.96	38.65
.718427	141	23.19	61.84
.7618647	161	26.48	88.32
.8338792	40	6.58	94.90
.8629051	31	5.10	100.00
-----+-----			
Total	608	100.00	

**Re-run the classification with all those with predicted probability greater than 0.66 declared failures**



```
. estat classification, cutoff(.667)
```

Logistic model for survival

		----- True -----		
Classified		D	~D	Total
-----+-----+-----				
+		284	89	373
-		119	116	235
-----+-----+-----				
Total		403	205	608

Classified + if predicted  $\Pr(D) \geq .667$

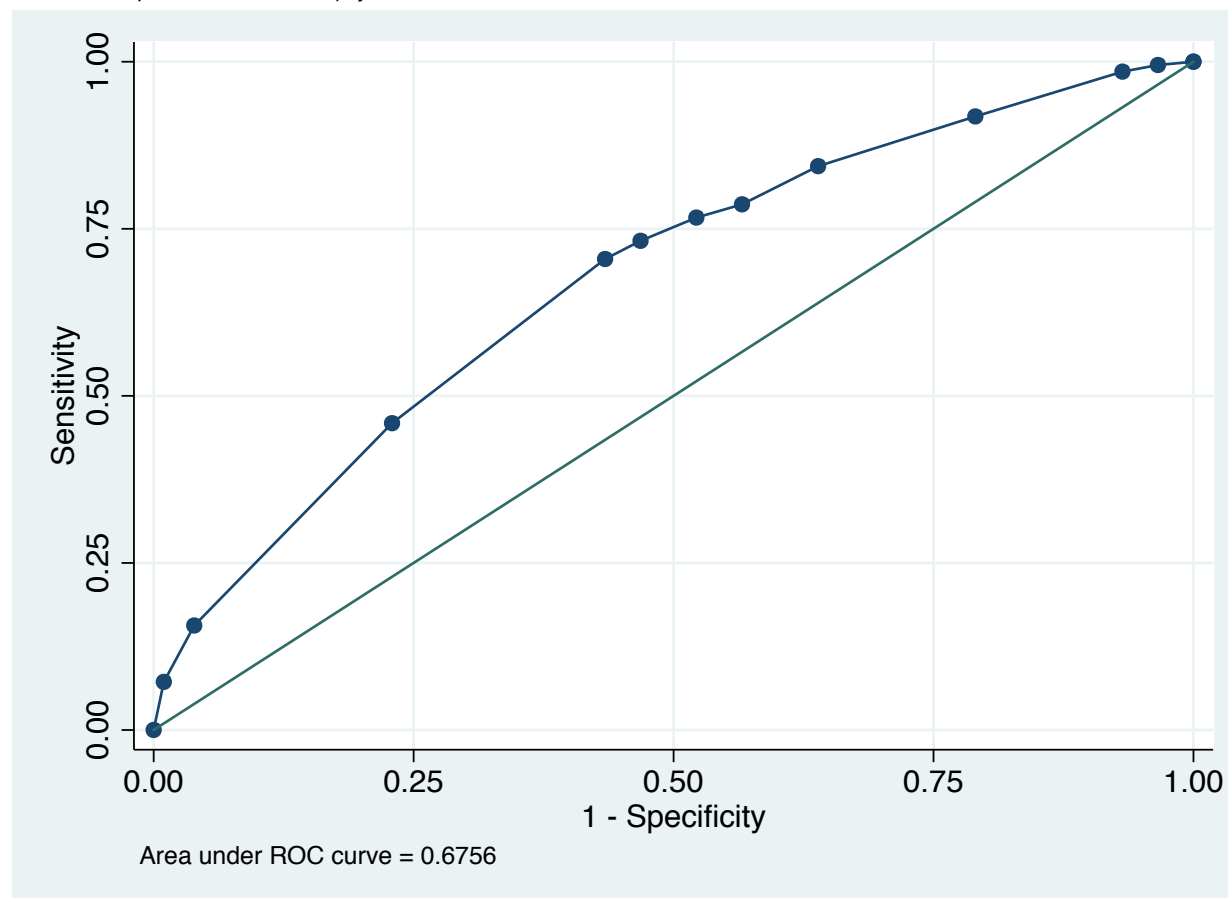
True D defined as survival  $\neq 0$

-----		
Sensitivity	Pr( +   D)	70.47%
Specificity	Pr( -   ~D)	56.59%
Positive predictive value	Pr( D   +)	76.14%
Negative predictive value	Pr( ~D   -)	49.36%
-----		
Correctly classified		65.79%
-----		

- This rule is better on some measures, worse on others. To look at all probability cutoff values at once, use ROC curve (command *lroc*)

## Logistic Regression - ROC Curve for Prediction

- Receiver Operating Characteristic (ROC) curve: plot of Sensitivity ( $\Pr(\text{predict } + | \text{case is } +)$ ) vs. False Positive Rate ( $\Pr(\text{predict } + | \text{case is } -)$ ) for all possible probability cut points.



## Logistic Regression - ROC Curve for Prediction

- **Area under curve is the metric of interest.** Perfect prediction has area = 1.0. Area under 'Guessing' line = 0.50. Area can be thought of as equalling, if given two cases (event and non-event), the probability of correctly classifying one as event and other as non-event
- Point closest to upper left corner is the best classifier (best trade-off between sensitivity and specificity)
- This model has area = 0.68, best probability cutpoint near specificity = 75%, sensitivity = 50%. Can examine ordered list of predicted probabilities to find (1roc in Stata will not list plot values (?))

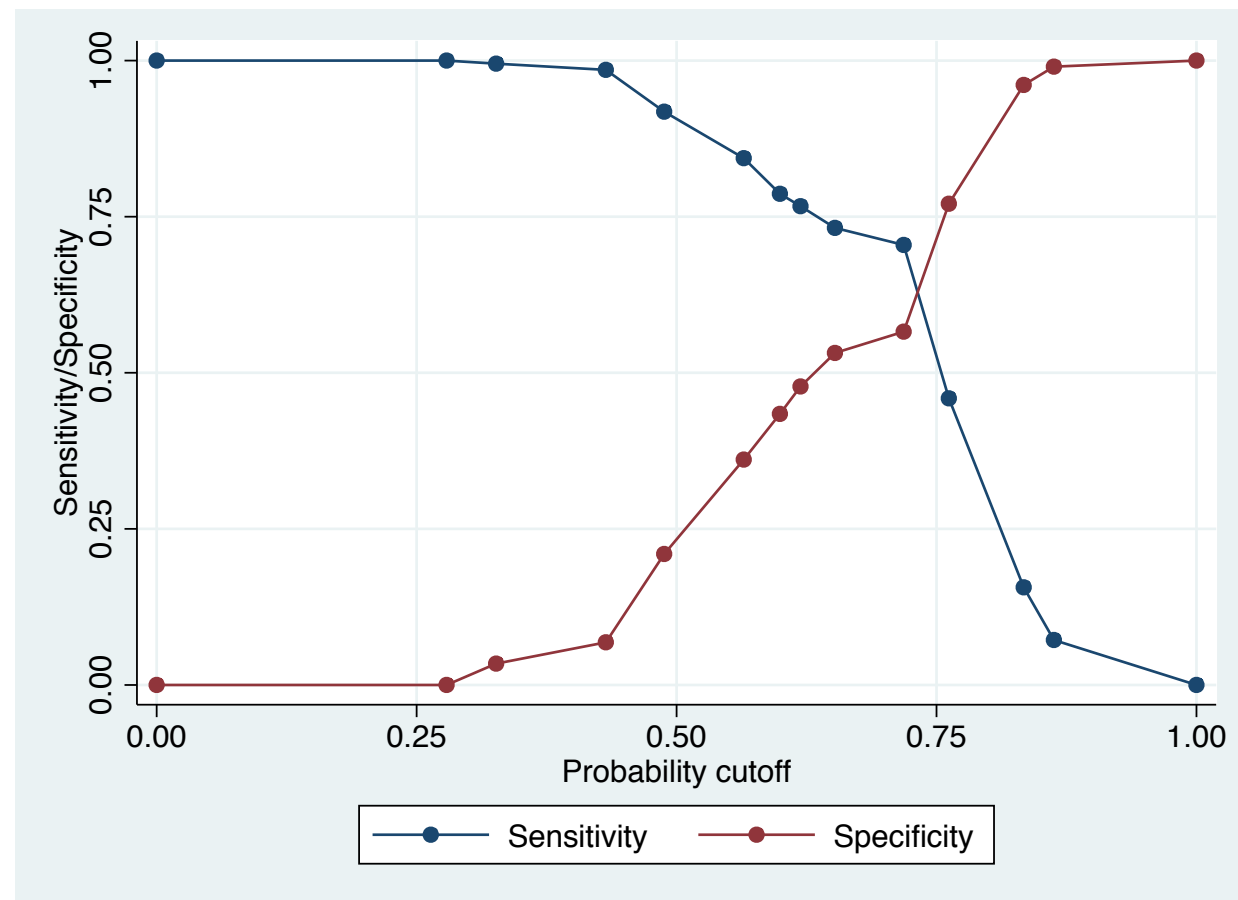
A related option 1sens in Stata will provide the sensitivity/specificity values at each model predicted probability.

## Logistic Regression - ROC Curve for Prediction

```
. lsens, gense(sensval) gensp(specval)  
. list p sensval specval in 1/12, clean
```

	p	sensval	specval
1.	.8629051	0.000000	1.000000
2.	.8338792	0.071960	0.990244
3.	.7618647	0.156328	0.960976
4.	.718427	0.459057	0.770732
5.	.6523169	0.704715	0.565854
6.	.6192176	0.732010	0.531707
7.	.5994036	0.766749	0.478049
8.	.5646291	0.786600	0.434146
9.	.4881366	0.843672	0.360976
10.	.4319939	0.918114	0.209756
11.	.3264767	0.985112	0.068293
12.	.2788	0.995037	0.034146

- These are the points on the ROC curve. A plot is also produced



- IMPORTANT NOTE: Optimizing cut point on a given dataset does not establish that classifier will work on novel data - requires validation in an independent sample (not used to build the predictor model)

## Classification - Donner Party Data

```
. estat classification
Logistic model for died
```

----- True -----			
Classified	D	~D	Total
-----+-----+-----			
+	23	8	31
-	2	12	14
-----+-----+-----			
Total	25	20	45

Classified + if predicted  $\Pr(D) \geq .5$ . True D defined as died != 0

-----		
Sensitivity	$\Pr(+ D)$	92.00%
Specificity	$\Pr(- \sim D)$	60.00%
Positive predictive value	$\Pr(D +)$	74.19%
Negative predictive value	$\Pr(\sim D -)$	85.71%
-----		
False + rate for true ~D	$\Pr(+ \sim D)$	40.00%
False - rate for true D	$\Pr(- D)$	8.00%
False + rate for classified +	$\Pr(\sim D +)$	25.81%
False - rate for classified -	$\Pr(D -)$	14.29%
-----		
Correctly classified		77.78%

## Classification - Donner Party Data

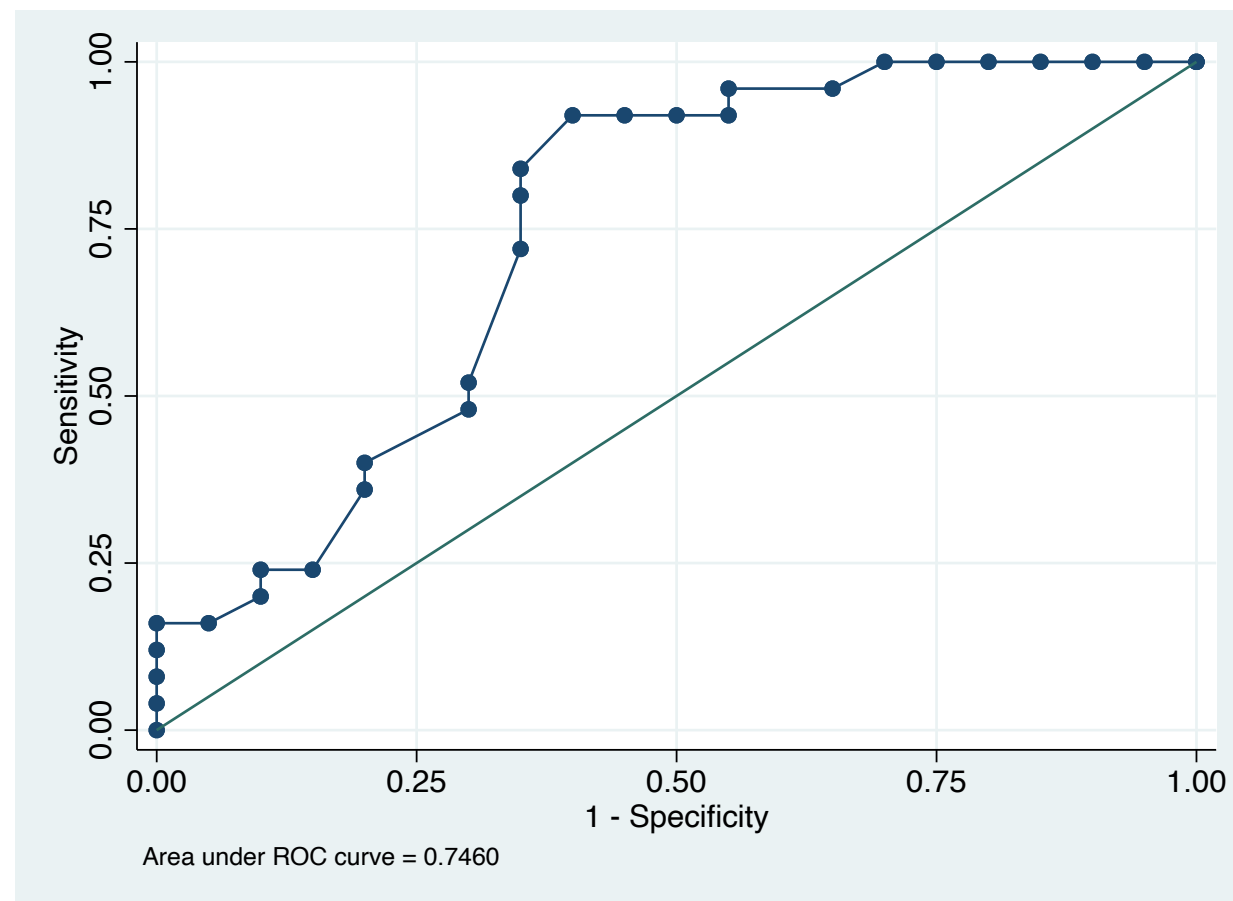
All possible cutpoints (except the extremes):

```
. predict phat
. list phat sensval specval, clean
. . .
10.    .6636918    0.360000    0.800000
11.    .635664    0.400000    0.800000
12.    .6094921    0.480000    0.700000
13.    .5798071    0.520000    0.700000
14.    .5716949    0.720000    0.650000
15.    .5606443    0.800000    0.650000
16.    .541299    0.840000    0.650000
17.    .4827471    0.920000    0.600000
18.    .4744595    0.920000    0.550000
19.    .4438758    0.920000    0.500000
20.    .3869718    0.920000    0.450000
21.    .3256594    0.960000    0.450000
. . .
```

probability cutpoint around 0.54 is best, as also shown in the ROC curve

## Classification - Donner Party Data

```
.* get ROC curve  
. lroc
```





## **Summary: Logistic Regression Models**

- Several diagnostic quantities, aiming to detect outliers and influential points, are defined (C&H 12.5, not covered in SPRM). These borrow concepts from linear regression, discussed next . . .