Logistic Regression

Part II

Logistic Regression: The Multivariate Problem

Q: What is the relationship between one (or more) exposure variables, ***E***, and a binary disease or illness outcome, ***Y***, while adjusting for potential confounding effects of C_1, C_2, …?

Example:

- ***Y*** is coronary heart disease (CHD). ***Y*** = 1 is "with CHD" and ***Y*** = 0 is "without CHD".
- Our exposure of interest is smoking, ***E*** = 1 for smokers (current, ever), and ***E*** = 0 for non-smokers.
- What is the extent of association between smoking and CHD?
- We want to "account for" or control for other variables (potential confounders) such as *age, race and gender*.

$$E, C_1, C_2, C_3 \Rightarrow Y$$

"independent" "dependent"

Logistic Regression: The Multivariate Problem

Independent variables: $\mathbf{X}=X_1, X_2,\ldots,X_p$

Dependent variable: $\boldsymbol{Y}$, binary

- We have a flexible choice for the type of independent variables. These may be continuous, categorical, binary.
- We can adopt a mathematical model to structure the *systematic variation* in the response variable $\boldsymbol{Y}$ as a function of $\boldsymbol{X}$.
- We can adopt a probability model to represent the *random variation* in the response.

Recall: Linear regression $Y = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p + e$

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

Instead, consider the equivalent representation …

$$Y \sim N(\mu(X), \sigma^2)$$

$$\mu(X) = \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$$

Binary Response Regression

Recall: For binary Y (= 0/1)

$$\mu = E(Y) = P(Y=1)=\pi$$

- The mean of a binary variable is a *probability*, i.e. $\pi \in (0,1)$.
- The mean may depend on covariates.
- This suggests considering: $\pi(X) = \beta_o + \beta_1 X_1 + \cdots + \beta_p X_p$
- Can we use linear regression for $\pi(X)$?
- Two issues:
  - If we model the mean for Y we'll need to impose the constraint $0 \leq \pi(\boldsymbol{X}) \leq 1$ for all $\boldsymbol{X}$.
    - o binary $\boldsymbol{X}$
    - o multi-categorical $\boldsymbol{X}$
    - o continuous $\boldsymbol{X}$
  - What is σ^2 for binary data?

The Logistic Function

The **logistic function** is given by: $f(z) = \dfrac{\exp(z)}{1+\exp(z)}$

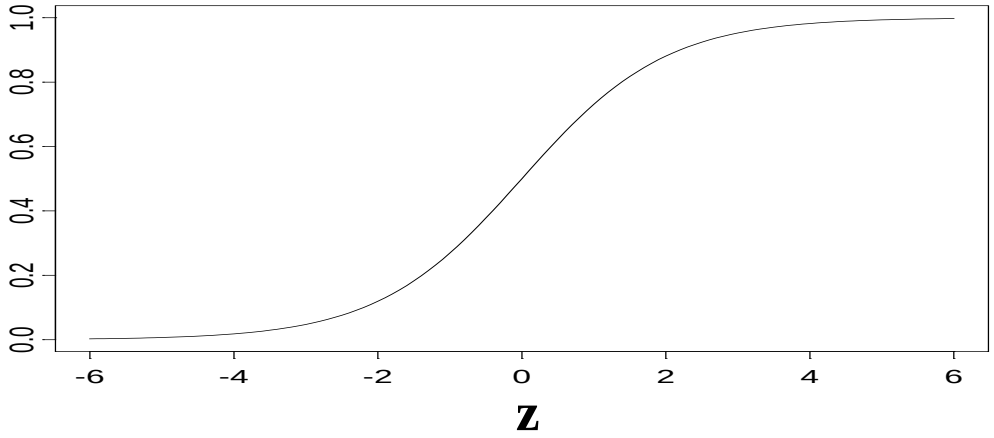
$$= \frac{1}{1+\exp(-z)}$$

Properties: an "S" shaped curve with:

$$\lim_{z \to +\infty} f(z) = 1/[1+0] = 1$$

$$\lim_{z \to -\infty} f(z) = 1/[1+\infty] = 0$$

$$f(0) = 1/2$$



The Logistic Regression Model

Define a "linear predictor" by

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + pX_p$$

Then the model for $\pi(X\beta)$ is:

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

this is called the "expit" transform

$$\Leftrightarrow$$

this is called the log odds or "logit" transform

$$\log\left(\frac{\pi(X\beta)}{1-\pi(X\beta)}\right) = X\beta$$

Generally, you see

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

1. This is a **non-linear** model for the mean (i.e. probability) , $\pi(X)$.
2. However, the model for the *log odds* is linear (i.e. $\mathbf{X\beta}$ is linear in $\mathbf{X}$).
3. The logit is the "link" function that relates the mean to the linear predictor.

The Logistic Regression Model

Q: Why is logistic regression so popular?

1. Dichotomous outcomes are common.
2. Logistic regression ensures that predicted probabilities lie between 0 and 1.
3. Regression parameters are log odds ratios – hence, estimable from case-control studies

Logistic Regression: Some special cases

Binary Exposure

Q: What is the logistic regression model for a simple binary exposure variable, ***E***, and a binary disease or illness outcome, ***D***?

Example: Pauling (1971)

	E=1 (Vit C)	E=0 (placebo)	
D (cold=Yes)	17	31	48
$\overline{D}$ (cold=No)	122	109	231
	139	140	279

Binary Exposure

X_1 : exposure:

X_1 = 1 if in group ***E*** (vitamin C)

X_1 = 0 if in group $\overline{E}$ (placebo)

Y: outcome

Y = 1 if in group ***D*** (cold)

Y = 0 if in group $\overline{D}$(no cold)

	X_1=1	X_1=0	
Y=1	17	31	48
Y=0	122	109	231
	139	140	279

$$\hat{P}[Y=1|X_1=1] = 17/139 = 0.122$$

$$\hat{P}[Y=1|X_1=0] = 31/140 = 0.221$$

$$\hat{RR} = 0.122/0.221 = 0.554$$

$$\hat{OR} = \frac{0.122/(1-0.122)}{0.221/(1-0.221)} = 0.490$$

Binary Exposure

Q: How would we approach these data using logistic regression?

Data:

Y	X_1	count
1	1	17
0	1	122
1	0	31
0	0	109

Model:

$$\pi(X) = \text{probability that Y=1 given X}$$

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 X_1$$

Binary Exposure – Parameter Interpretation

- Model $\text{logit}(\pi(X_1)) = \beta_0 + \beta_1 X_1$
- Probabilities:

$$P(Y=1|X_1=0) = \pi(X_1=0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

$$P(Y=1|X_1=1) = \pi(X_1=1) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$$

- Odds:

  Odds of disease when $(X_1 = 0) = \exp(\beta_0)$

  Odds of disease when $(X_1 = 1) = \exp(\beta_0 + \beta_1)$

- Odds Ratio:

$$OR = \frac{\text{Odds}(X_1=1)}{\text{Odds}(X_1=0)} = \exp(\beta_1)$$

- Log Odds ratio: β_1

Binary Exposure – Parameter Estimates

Q: How can we estimate the logistic regression model parameters?

In this simple case we could calculate by hand:

We know that:

$$\hat{\pi}(X_1 = 0) = 0.221,$$

$$\text{and } \log\left(\frac{\hat{\pi}(X_1=0)}{1-\hat{\pi}(X_1=0)}\right) = \hat{\beta}_0 .$$

$$\log\left(\frac{0.221}{1-0.221}\right) = -1.260$$

We also know that:

$$\hat{\pi}(X_1 = 1) = 0.122,$$

$$\text{and } \log\left(\frac{\hat{\pi}(X_1=1)}{1-\hat{\pi}(X_1=1)}\right) = \hat{\beta}_0 + \hat{\beta}_1.$$

$$\log\left(\frac{0.122}{1-0.122}\right) = -1.974$$

Hence:

$$\hat{\beta}_0 = -1.260$$

$$\text{and } \hat{\beta}_1 = -1.974 - (-1.260) = -0.713$$

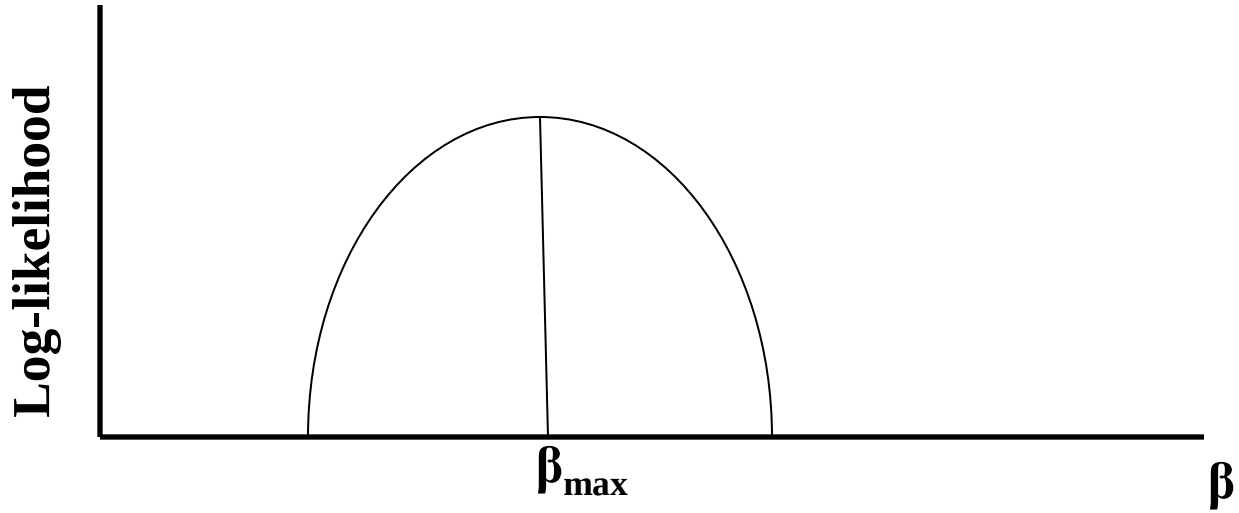
$$\hat{OR} = \exp(-0.713) = .490$$

Binary Exposure – Parameter Estimates

Q: How can we estimate the logistic regression model parameters?

A: More generally, for models with multiple covariates, the computer implements an estimation method known as "maximum likelihood estimation"

Generic idea – Find the value of the parameter (β) where the log-likelihood function, $\ell(\beta;\text{data})$, is maximum



- for multiple parameters β_1, β_2, … imagine a likelihood "mountain"

Maximum Likelihood Estimation

- In simple cases, corresponds with our common sense estimates; but applies to complex problems as well
- Maximum likelihood is the “best” method of estimation for any situation that you are willing to write down a probability model.
- We can use computers to find these estimates by maximizing a particular function, known as the likelihood function.
- We use comparisons in the value of the (log) likelihood function as a preferred method for testing whether certain variables (coefficients) are significant (i.e. to test H_0: $\beta=0$).

Vitamin C Study Example – STATA

```
. input count y x
        count         y         x
  1. 17 1 1
  2. 31 1 0
  3. 122 0 1
  4. 109 0 0
  5. end

. expand count
(275 observations created)


. cs y x, or
                |  x                    |
                |    Exposed   Unexposed |      Total
----------------+-----------------------+-----------
          Cases |         17          31 |        48
       Noncases |        122         109 |       231
----------------+-----------------------+-----------
          Total |        139         140 |       279
                |                       |
           Risk |   .1223022    .2214286 |    .172043
                |                       |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+----------------------
 Risk difference |        -.0991264       | -.1868592   -.0113937
      Risk ratio |         .5523323       |  .3209178    .9506203
 Prev. frac. ex. |         .4476677       |  .0493797    .6790822
 Prev. frac. pop |         .2230316       |
      Odds ratio |         .4899524       |  .2588072    .9282861  (Cornfield)
                +-------------------------------------------------
                           chi2(1) =     4.81  Pr>chi2 = 0.0283
```

Logistic Regression – Vitamin C Study Example

```
. logistic y x

Logistic regression                               Number of obs   =        279
                                                  LR chi2(1)      =       4.87
                                                  Prob > chi2     =     0.0273
Log likelihood =  -125.6561                       Pseudo R2       =     0.0190

------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4899524   .1613518    -2.17   0.030     .2569419    .9342709
------------------------------------------------------------------------------
```

```
. logit y x

Logit estimates                                   Number of obs   =        279
                                                  LR chi2(1)      =       4.87
                                                  Prob > chi2     =     0.0273
Log likelihood =  -125.6561                       Pseudo R2       =     0.0190

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.713447   .3293214    -2.17   0.030    -1.358905   -.0679889
       _cons |  -1.257361   .2035494    -6.18   0.000     -1.65631   -.8584111
------------------------------------------------------------------------------
```

Alternatives:

```
. glm y x, family(binomial) link(logit) eform
. binreg y x, or
```

Logistic Regression – Single Binary Predictor

Model: $\text{logit}[\pi(X_1)] = \beta_0 + \beta_1 X_1$

X_1	$\text{logit}(\pi(X_1))$
0	β_0
1	$\beta_0 + \beta_1$

i.e.,

- $\text{logit}(P(Y=1|X_1=0)) = \beta_0$
- $\text{logit}(P(Y=1|X_1=1)) = \beta_0 + \beta_1$
- $\text{logit}(P(Y=1|X_1=1)) - \text{logit}(P(Y=1|X_1=0)) = \beta_1$

β_1 is the log odds ratio of "success" (Y=1) comparing two groups with $X_1=1$ (first group) and $X_1=0$ (second group)

β_0 is the log odds (of Y) for $X_1=0$

Logistic Regression – Single Binary Predictor

Model: $$\text{logit}[\pi(X_1)] = \beta_0 + \beta_1 X_1$$

Probability

$X_1 = 1$	$X_1 = 0$
$\frac{\exp(\beta_0 + \beta_1)}{1+\exp(\beta_0 + \beta_1)}$	$\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$

Odds

$X_1 = 1$	$X_1 = 0$
$\exp(\beta_0 + \beta_1)$	$\exp(\beta_0)$

- $\exp(\beta_1)$ is the odds ratio that compares the odds of a success (Y=1) in the "exposed" (X_1=1) group to the "unexposed" (X_1=0) group.
- The logistic regression OR and the simple 2x2 OR are identical

$$\hat{OR} = \exp(\hat{\beta}_1) = ad/bc$$

- Also, expit($\hat{\beta}_0$) is P(Y = 1 | X = 0)

Logistic Regression – Case-Control Studies

Recall:

- In case-control studies we sample cases (**Y** = 1) and controls (**Y** = 0) and then ascertain covariates (**X**).
- From this study design we *cannot* estimate disease risk, P(**Y**=1|**X**), nor relative risk, but we *can* estimate exposure odds ratios.
- Exposure odds ratios are equal to disease odds ratios.
- The result is that we can use case-control data to estimate disease odds ratios, which for rare outcomes approximate relative risks.

Q: Can one do any regression modeling using data from a case-control study?

A: Yes, one can use standard logistic regression to estimate ORs but *not* disease risk probabilities

Logistic Regression – Case-Control Studies

The case-control study design is particularly effective when "disease" is rare

- If the "disease" affects only 1 person per 10,000 per year, we would need a very large prospective study
- But, if we consider a large urban area of 1,000,000, we would expect to see 100 cases a year
- We could sample all 100 cases and 100 random controls
- Sampling fractions, f, for cases and controls are then very different:

$$f_0 = \frac{100}{999,900} = .00001 \text{ for controls}$$

$$f_1 = \frac{100}{100} = 1 \text{ for cases}$$

Logistic Regression – Case-Control Studies

Key points:

1. We can "pretend that the case-control data was collected prospectively and use standard logistic regression (outcome=disease; covariate=exposure) to estimate regression coefficients and obtain standard errors

2. We need to be careful not to use the intercept, β_0, to estimate risk probabilities. In fact, $\hat{\beta}_0 = \beta_0 + \ln(f_1 / f_0)$ where β_0 is the true value you would get random sample of the population (e.g. if $f_1 = 1$ and $f_0 = .00001$, then $\hat{\beta}_0 = \beta_0 + 9.21$)

3. A key assumption is that the probability of being sampled for both cases and controls does not depend on the covariate of interest.

Logistic Regression – Case-Control Studies

Example 2: Keller (AJPH, 1965)

	Case	Control	Total
Smoker	484	385	869
Non-Smoker	27	90	117
Total	511	475	986

```
. logit cancer smoke

Logistic regression                               Number of obs   =        986
                                                  LR chi2(1)      =      45.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -659.89728                       Pseudo R2       =     0.0335

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   1.432814   .2298079     6.23   0.000     .9823992     1.88323
       _cons |  -1.203973   .2194269    -5.49   0.000    -1.634042   -.7739041
------------------------------------------------------------------------------
```

Interpret

OR = exp(1.43) = 4.19 (compare to hand calc)

β_0 is never directly interpretable in a case-control study!

Logistic Regression – Case-Control Studies

Disease OR = Exposure OR $\Rightarrow$

```
. logit cancer smoke

Logistic regression                               Number of obs   =        986
                                                  LR chi2(1)      =      45.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -659.89728                       Pseudo R2       =     0.0335

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   1.432814   .2298079     6.23   0.000     .9823992     1.88323
       _cons |  -1.203973   .2194269    -5.49   0.000    -1.634042   -.7739041
------------------------------------------------------------------------------

. logit smoke cancer

Logistic regression                               Number of obs   =        986
                                                  LR chi2(1)      =      45.78
                                                  Prob > chi2     =     0.0000
Log likelihood = -336.26115                       Pseudo R2       =     0.0637

------------------------------------------------------------------------------
       smoke |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cancer |   1.432814   .2298079     6.23   0.000     .9823992     1.88323
       _cons |   1.453434   .1170834    12.41   0.000     1.223954    1.682913
------------------------------------------------------------------------------
```

Binary Regression – Other "links"

1) "logit" link: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1$
   - most common form of binary regression
   - guarantees that $\pi(X)$ is between 0 and 1
   - coefficients are log(OR) (for 0/1 covariate)
   - computationally stable

2) "identity" link" $\pi(X) = \beta_0 + \beta_1 X_1$
   - predicted $\pi(X)$ can be $-\infty$ to ∞
   - coefficients are RD (for 0/1 covariate)
   - computationally stable

3) "log" link" $\log[\pi(X)] = \beta_0 + \beta_1 X_1$
   - predicted $\pi(X)$ can be 0 to ∞
   - coefficients are log(RR) for (0/1 covariate)
   - less computationally stable

Binary Regression – Other "links"

Example 1 – Vitamin C Study

```
.* logistic link regression
. binreg y x, or
. glm y x, family(binomial) link(logit) eform
```

```
             |                 OIM
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4899524   .1613518    -2.17   0.030     .2569419    .9342709
```

```
.* identity link regression
. binreg y x, rd
. glm y x, family(binomial) link(identity)
```

```
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.0991264   .0447624    -2.21   0.027    -.1868592   -.0113937
       _cons |   .2214286   .0350915     6.31   0.000     .1526505    .2902067

Coefficients are the risk differences
```

```
.* log link regression
. binreg y x, rr
. glm y x, family(binomial) link(log) eform
```

```
             |                 OIM
           y | Risk Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .5523323   .1530115    -2.14   0.032     .3209178    .9506203
```

Summary

1. For 2 x 2 tables logistic regression fits a pair of probabilities:

$$\hat{\pi}(X=1) \text{ and } \hat{\pi}(X=0)$$

2. Model: $\mathbf{logit}(\pi(X)) = \beta_0 + \beta_1 X.$

3. β_0 represents the reference "log odds" when X=0.

4. β_1 is the log odds ratio that compares the log odds of response among the "exposed" (***X*** = 1) to the log odds of response among the "unexposed" (***X*** = 0).

5. The logistic regression odds ratio and the simple 2x2 odds ratio are identical.

6. Note: the estimated standard errors (95% CI) from logistic regression may be slightly different from those for the 2 x 2 table analysis.

7. Other links are possible

Applying the *Multiple* Logistic Model

Example: from Kleinbaum (1994)

$\boldsymbol{Y} = \boldsymbol{D}$ = CHD (0, 1)

$\boldsymbol{X}_1$ = CAT catecholamine level 1 = high, 0 = low

$\boldsymbol{X}_2$ = AGE, in years

$\boldsymbol{X}_3$ = ECG 1 = abnormal, 0 = normal

n = 609 males with 9-year follow-up

Model: $\boldsymbol{P(D=1|X) = \pi(X) = \frac{\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)}{1+\exp(\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3)}}$

Estimates: via maximum likelihood

$$\hat{\beta}_0 = -3.911$$
$$\hat{\beta}_1 = 0.652$$
$$\hat{\beta}_2 = 0.029$$
$$\hat{\beta}_3 = 0.342$$

Applying the *Multiple* Logistic Model

Q: What is the estimated probability of CHD for an individual with (**CAT = 1**, AGE = 40, ECG = 0)?

$$\hat{\pi}\left(\mathbf{X}^{(1)}\right) = \frac{\exp\left(-3.911+0.652(1)+0.029(40)+0.342(0)\right)}{1+\exp\left(-3.911+0.652(1)+0.029(40)+0.342(0)\right)}$$

$$= \frac{\exp(-2.101)}{1+\exp(-2.101)}$$

$$= 0.109$$

Q: What is the estimated probability of CHD for an individual with (**CAT = 0**, AGE = 40, ECG = 0)?

$$\hat{\pi}\left(\mathbf{X}^{(0)}\right) = \frac{\exp\left(-3.911+0.652(0)+0.029(40)+0.342(0)\right)}{1+\exp\left(-3.911+0.652(0)+0.029(40)+0.342(0)\right)}$$

$$= \frac{\exp(-2.751)}{1+\exp(-2.751)}$$

$$= 0.060$$

Applying the *Multiple* Logistic Model

Compare CHD risk by CAT for persons with AGE=40 and ECG=0:

$$\hat{RR} = \frac{\hat{\pi}(\mathbf{X}^{(1)})}{\hat{\pi}(\mathbf{X}^{(0)})}$$

$$= 0.109/0.060 = 1.817$$

$$\hat{OR} = \frac{\hat{\pi}(\mathbf{X}^{(1)})/[1-\hat{\pi}(\mathbf{X}^{(1)})]}{\hat{\pi}(\mathbf{X}^{(0)})/[1-\hat{\pi}(\mathbf{X}^{(0)})]}$$

$$= \frac{0.109/(1-0.109)}{0.060/(1-0.060)} = 1.917$$

Note:

1) $\log(1.917) = 0.652 = \hat{\beta}_1$

2) Does the estimated OR associated with CAT change if AGE or ECG changes?

3) Does the estimated RR associated with CAT change if AGE or ECG changes?

Logit Coefficients

We can represent the logistic model for CHD as:

$$\mathbf{logit}[\pi(\mathbf{X})] = \beta_0 + \beta_1\mathbf{CAT} + \beta_2\mathbf{AGE} + \beta_3\mathbf{ECG}$$

Then we can evaluate (1,40,0) and (0,40,0) as before:

$$\mathbf{logit}[\pi(\mathbf{X}^{(1)})] = \beta_0 + \beta_1 x 1 + \beta_2 x 40$$

$$\mathbf{logit}[\pi(\mathbf{X}^{(0)})] = \beta_0 + \beta_2 x 40$$

$$\mathbf{logit}[\pi(\mathbf{X}^{(1)})] - \mathbf{logit}[\pi(\mathbf{X}^{(0)})] = \beta_1$$

Q: What is the interpretation of β_1?

We can say: "β_1 represents the change in the log odds when CAT changed from 0 to 1 and AGE and ECG are fixed"

Q: What is the interpretation of β_2?

β_2 represents the change in the log odds when AGE changes by one year and CAT and ECG are fixed.

Logit Coefficients

Q: What is the interpretation of β_0 ?

- If all the $X_j = 0$ then $\text{logit}(\pi(X)) = \beta_0$
- Therefore, β_0 is the *log odds for an individual with all covariates equal to zero*.

Q: Does this make sense?

Sometimes – but not in our CHD example. CAT=0 is meaningful, ECG=0 is meaningful, but AGE=0 is not.

Recall: β_0 is never directly interpretable in a case-control study

Odds Ratios

We can also consider an odds ratio between two individuals (or populations) characterized by two covariate values, $X^{(1)}$ and $X^{(0)}$:

$$\text{odds}(X^{(1)}) = \exp(X^{(1)}\beta)$$

$$\text{odds}(X^{(0)}) = \exp(X^{(0)}\beta)$$

$$\text{OR}(X^{(1)}, X^{(0)}) = \frac{\text{odds}(X^{(1)})}{\text{odds}(X^{(0)})}$$

$$= \frac{\exp(X^{(1)}\beta)}{\exp(X^{(0)}\beta)}$$

$$= \exp\left(\sum_{j=1}^{p} \beta_j (X_j^{(1)} - X_j^{(0)})\right)$$

Odds Ratios

In the CHD example we have:

$$\mathbf{X^{(1)}} = (1, 40, 1)$$

$$\mathbf{X^{(2)}} = (0, 40, 0)$$

And we can obtain the odds ratio comparison as:

$$\mathbf{OR}(\mathbf{X}^{(1)},\mathbf{X}^{(0)}) = \exp\left(\sum_{j=1}^{3} \beta_j (X_j^{(1)} - X_j^{(0)})\right)$$

$$= \exp(\beta_1 + \beta_3)$$

$$\widehat{OR}\left(X^{(1)}, X^{(0)}\right) = \exp(0.652 + 0.342)$$

$$= 2.702$$

Interpret:

Summary
Multiple Logistic Regression

1. Define: logistic function.
2. Properties of the logistic function.
3. Multiple covariates.
4. Applying the logistic model to obtain probabilities.
5. Interpreting the parameters in the logistic model.
6. Obtaining RRs and ORs.

Logistic Regression: 2 Binary Covariates

Q: What is the logistic regression model for a simple binary "exposure" variable, ***E***, a simple binary stratifying variable, ***C***, and a binary outcome, ***Y***?

Example: HIVNET Vaccine Trial - willingness to participate

Previous Study	High Knowledge (≥8)	Low Knowledge(<=7)	
Willing	22	32	54
Not Willing	112	153	265
	134	185	319

New Recruits	High Knowledge (≥8)	Low Knowledge(<=7)	
Willing	39	67	106
Not Willing	146	179	325
	185	246	431

Logistic Regression: 2 Binary Covariates

- Role of cohort (Previous study/ New recruits)?

  Y=Willingness to participate

  E=Knowledge

  C=Cohort (recruitment time)

- Crude table – willingness vs knowledge OR = .79
  - For previous study cohort OR = .94
  - For new recruits cohort OR = .71
- Cohort as confounder ?
- Cohort as effect modifier (on OR scale)?

Logistic Regression: Defining Variables

- Let X_1 be a dummy variable for "exposure" groups:

  $X_1 = 1$ if subject has knowledge score ≥ 8.

  $X_1 = 0$ if subject has knowledge score ≤ 7.

- Let X_2 be a dummy variable for "stratification" groups:

  $X_2 = 1$ if subject is a new recruit

  $X_2 = 0$ if subject is from a previous study (rollover).

- Let Y be an indicator (binary) outcome variable:

  $Y = 1$ if subject is definitely willing to participate.

  $Y = 0$ if subject is not definitely willing to participate.

HIVNET Vaccine Trial Data

Y will	X1 know	X2 cohort	count count
1	1	0	22
0	1	0	112
1	0	0	32
0	0	0	153
1	1	1	39
0	1	1	146
1	0	1	67
0	0	1	179

(In what follows, assume that the data have been expanded on “count”.)

Crude Analysis Using STATA

```
. cc will know   /* Combined Data */
                                                      Proportion
                |   Exposed   Unexposed  |     Total     Exposed
----------------+------------------------+----------------------
          Cases |        61          99  |       160      0.3812
       Controls |       258         332  |       590      0.4373
----------------+------------------------+----------------------
          Total |       319         431  |       750      0.4253
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+----------------------
     Odds ratio |         .7928901       |  .5440409    1.150404  (exact)
 Prev. frac. ex. |         .2071099       | -.1504037    .4559591  (exact)
 Prev. frac. pop |         .0905667       |
                +-------------------------------------------------
                           chi2(1) =     1.62  Pr>chi2 = 0.2035
```

```
. logistic will know

Logistic regression                               Number of obs   =        750
                                                  LR chi2(1)      =       1.63
                                                  Prob > chi2     =     0.2018
Log likelihood =    -387.94                       Pseudo R2       =     0.0021

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |   .7928901   .1448677    -1.27   0.204     .5542317    1.134318
------------------------------------------------------------------------------
```

Possible Approaches/Models

Adjustment for confounding by stratification:

- Mantel-Haenszel estimate of common OR
- Logistic regression using an additive model on the logit scale

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 \text{KNOW} + \beta_2 \text{COHORT}$$

Effect modification or interaction

- Separate odds ratio estimates for each stratum
- Logistic regression using *interaction model*

$$\begin{aligned}\text{logit}(\pi(X)) = \beta_0 &+ \beta_1 \text{KNOW} \\ &+ \beta_2 \text{COHORT} \\ &+ \beta_3 \text{KNOW} * \text{COHORT}\end{aligned}$$

Adjusted Analyses Using STATA

```
./* cohort as confounder */
. cc will know, by(cohort) bd

         cohort |      OR      [95% Conf. Interval]    M-H Weight
----------------+-------------------------------------------------
              0 |   .9391741     .4916112   1.769653      11.23511 (exact)
              1 |   .7136577     .4413983   1.145675      22.69606 (exact)
----------------+-------------------------------------------------
          Crude |   .7928901     .5440409   1.150404               (exact)
    M-H combined |   .7883295     .5504114   1.129089
---------------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.52  Pr>chi2 = 0.4711
Test of homogeneity (B-D)      chi2(1) =     0.52  Pr>chi2 = 0.4708

                 Test that combined OR = 1:
                             Mantel-Haenszel chi2(1) =      1.69
                                                Pr>chi2 =    0.1941
```

```
. logistic will know cohort

Logistic regression                               Number of obs   =        750
                                                  LR chi2(2)      =       8.24
                                                  Prob > chi2     =     0.0163
Log likelihood = -384.63529                       Pseudo R2       =     0.0106

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |   .7879471   .1445962    -1.30   0.194     .5499117    1.129019
      cohort |   1.605702    .299915     2.54   0.011     1.113465    2.315546
------------------------------------------------------------------------------
```

Adjusted Analyses Using STATA

```
. logit

Logit estimates                                   Number of obs   =        750
                                                  LR chi2(2)      =       8.24
                                                  Prob > chi2     =     0.0163
Log likelihood = -384.63529                       Pseudo R2       =     0.0106

------------------------------------------------------------------------------
        will |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |  -.2383243   .1835101    -1.30   0.194    -.5979975    .1213488
      cohort |   .4735612   .1867812     2.54   0.011     .1074767    .8396457
       _cons |  -1.495195   .1650607    -9.06   0.000    -1.818708   -1.171682
------------------------------------------------------------------------------
```

Parameter Interpretation
"Additive" Model (on logit scale)

Model: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

X_1	X_2	$\text{logit}(\pi(X))$
0	0	β_0
1	0	$\beta_0+\beta_1$
0	1	$\beta_0+\beta_2$
1	1	$\beta_0+\beta_1+\beta_2$

- Odds Ratios for willingness to participate:

$$\frac{\text{Odds}(X_1=1, X_2=0)}{\text{Odds}(X_1=0, X_2=0)} = \exp(\beta_1)$$

log odds ratio for KNOW, previous study $= \beta_1$

$$\frac{\text{Odds}(X_1=1, X_2=1)}{\text{Odds}(X_1=0, X_2=1)} = \exp(\beta_1)$$

log odds ratio for KNOW, new recruits $= \beta_1$

Adjusted Analyses Using STATA: Additive Model

```
. predict logodds, xb
. gen estprob=invlogit(logodds)
. predict stderr, stdp
. gen lodds_lb=logodds-1.96*stderr
. gen lodds_ub=logodds+1.96*stderr
. gen prob_lb=invlogit(lodds_lb)
. gen prob_ub=invlogit(lodds_ub)
. collapse  (mean) estprob lodds_lb lodds_ub prob_lb prob_ub, by(will know cohort)
. list will know cohort  estprob prob_lb prob_ub, noobs
```

will	know	cohort	estprob	prob_lb	prob_ub
0	0	0	.1831433	.1395883	.2365522
0	1	0	.150138	.1089749	.2033019
1	0	0	.1831433	.1395883	.2365522
1	1	0	.150138	.1089749	.2033019
0	0	1	.2647093	.2168735	.3188011
0	1	1	.2209811	.1725128	.2784841
1	0	1	.2647093	.2168735	.3188011
1	1	1	.2209811	.1725128	.2784841

Q: Suppose I gave you the logit output on slide 173 and asked you to estimate $\pi(1,0)$

A: $\text{logit}(\hat{\pi}(1,0) = -1.495 - .238 = -1.733 \implies \hat{\pi}(1,0) = .150$

Parameter Interpretation: Effect Modification

Model: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$

X_1	X_2	$\text{logit}(\pi(X))$
0	0	β_0
1	0	$\beta_0+\beta_1$
0	1	$\beta_0+\beta_2$
1	1	$\beta_0+\beta_1+\beta_2+\beta_3$

- Odds Ratios for willingness to participate:

$$\frac{\text{Odds}(X_1=1, X_2=0)}{\text{Odds}(X_1=0, X_2=0)} = \exp(\beta_1)$$

log odds ratio for knowledge, previous study $= \beta_1$

$$\frac{\text{Odds}(X_1=1, X_2=1)}{\text{Odds}(X_1=0, X_2=1)} = \exp(\beta_1+\beta_3)$$

log odds ratio for knowledge, new recruits $= \beta_1+\beta_3$

Effect Modification Model

```
. xi: logistic will i.know*i.cohort

Logistic regression                               Number of obs   =        750
                                                  LR chi2(3)      =       8.76
                                                  Prob > chi2     =     0.0327
Log likelihood = -384.37645                       Pseudo R2       =     0.0113

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iknow_1 |   .9391741   .2851272    -0.21   0.836     .5179965    1.702807
  _Icohort_1 |    1.78963   .4321054     2.41   0.016     1.114913    2.872668
_IknoXcoh_~1 |    .759878   .2895239    -0.72   0.471     .3601011    1.603479
------------------------------------------------------------------------------

. logit

------------------------------------------------------------------------------
        will |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iknow_1 |  -.0627544   .3035936    -0.21   0.836    -.6577869    .5322781
  _Icohort_1 |   .5820088   .2414496     2.41   0.016     .1087763    1.055241
_IknoXcoh_~1 |  -.2745974   .3810136    -0.72   0.471     -1.02137    .4721756
       _cons |  -1.564702   .1943861    -8.05   0.000    -1.945692   -1.183712
------------------------------------------------------------------------------
```

Notes:

1) The interaction (β_3) is not significant p = 0.47

2) Care must be taken in interpreting the main effects (β_1 and β_2) in a model with interactions

Effect Modification Model

Calculate estimated probabilities and CIs as before:

```
. list will know cohort  estprob prob_lb prob_ub, noobs
```

will	know	cohort	estprob	prob_lb	prob_ub
0	0	0	.172973	.1250231	.2343866
0	1	0	.1641791	.1106093	.2367862
1	0	0	.172973	.1250231	.2343866
1	1	0	.1641791	.1106093	.2367862
0	0	1	.2723577	.2203884	.3313729
0	1	1	.2108108	.1579798	.2755281
1	0	1	.2723577	.2203884	.3313729
1	1	1	.2108108	.1579798	.2755281

Q: Suppose I gave you the logit output on slide 177 and asked you to estimate $\pi(1,1)$

A: $\text{logit}(\hat{\pi}(1,1) = -1.565 - .063 + .582 - .274 = -1.32 \implies \hat{\pi}(1,1) = .211$

Logistic Regression vs Mantel-Haenszel: EM

```
. cc will know, by(cohort) bd

          cohort |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
      old cohort |   .9391741     .4916112   1.769653      11.23511 (exact)
      new cohort |   .7136577     .4413983   1.145675      22.69606 (exact)
-----------------+-------------------------------------------------
           Crude |   .7928901     .5440409   1.150404               (exact)
    M-H combined |   .7883295     .5504114   1.129089
-----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.52  Pr>chi2 = 0.4711
Test of homogeneity (B-D)      chi2(1) =     0.52  Pr>chi2 = 0.4708

                 Test that combined OR = 1:
                              Mantel-Haenszel chi2(1) =      1.69
                                                Pr>chi2 =    0.1941
```

```
. xi: logistic will i.know*i.cohort

Logistic regression                               Number of obs   =        750
                                                  LR chi2(3)      =       8.76
                                                  Prob > chi2     =     0.0327
Log likelihood = -384.37645                       Pseudo R2       =     0.0113

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iknow_1 |   .9391741   .2851272    -0.21   0.836     .5179965    1.702807
  _Icohort_1 |    1.78963   .4321054     2.41   0.016     1.114913    2.872668
_IknoXcoh_~1 |    .759878   .2895239    -0.72   0.471     .3601011    1.603479
------------------------------------------------------------------------------
```

Q: What is the analogue of BD test for homogeneity?

Additive vs Effect Modification Model

Estimated probabilities

X_1	X_2	Additive	Interaction
0	0	.183	.173
1	0	.150	.164
0	1	.265	.272
1	1	.221	.211

Estimated odds ratios

Contrast	Additive	Interaction
X_1=1 vs X_1=0; X_2=0	.788	.939
X_1=1 vs X_1=0; X_2=1	.788	.714

Additive vs Effect Modification Model

Notes:

1) In the additive model, the estimated probabilities do not exactly match the observed probabilities for the 4 combinations of X_1 and X_2

2) Here, the effect modification model is an example of a "saturated" model, so estimated probabilities exactly match the observed probabilities

Regression or Stratification?

"Recently there has been some dispute between 'modellers', who support the use of regression models, and 'stratifiers' who argue for a return to the methods described in Part I of this book. Logically this dispute is based on a false distinction – there is no real difference between the methods. In practice the difference lies in the inflexibility of the older methods which thereby imposes a certain discipline on the analyst. Firstly, since stratification methods treat exposures and confounders differently, any change in the role of a variable requires a new set computations. This forces us to keep in touch with the underlying scientific questions. Secondly, since strata must be defined by cross classification, relatively few confounders can be dealt with and we are forced to control only for confounders of a priori importance. These restraints can be helpful in keeping a data analyst on the right track but once the need for such a discipline is recognized, there are significant advantages to the regression modelling approach."

Clayton and Hills (1993), Statistical Methods in Epidemiology, page 273

SUMMARY

1. With two binary covariates we can model 4 probabilities:

$$\pi(X_1=0,\ X_2=0)$$
$$\pi(X_1=1,\ X_2=0)$$
$$\pi(X_1=0,\ X_2=1)$$
$$\pi(X_1=1,\ X_2=1)$$

2. We model two odds ratios:

$$\text{Odds}(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 0) \,/\, \text{Odds}(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 0)$$

$$\text{Odds}(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 1) \,/\, \text{Odds}(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 1)$$

3. The "interaction" coefficient (β_3) is the difference between these log odds ratios.

4. How to estimate and interpret coefficients in the simpler model that doesn't contain the interaction term, X_1X_2 ?

SUMMARY: Additive Model

Table of odds

	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	$\exp(\beta_0 + \beta_1 + \beta_2)$	$\exp(\beta_0 + \beta_1)$
$X_1 = 0$	$\exp(\beta_0 + \beta_2)$	$\exp(\beta_0)$

Table of odds ratios (relative to $X_1=0$, $X_2=0$)

	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	$\exp(\beta_1)\exp(\beta_2)$	$\exp(\beta_1)$
$X_1 = 0$	$\exp(\beta_2)$	1.0

SUMMARY: Interaction Model

	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)$	$\exp(\beta_0 + \beta_1)$
$X_1 = 0$	$\exp(\beta_0 + \beta_2)$	$\exp(\beta_0)$

Table of odds ratios (relative to X_1=0, X_2=0)

	$X_2 = 1$	$X_2 = 0$
$X_1 = 1$	$\exp(\beta_1 + \beta_2 + \beta_3)$	$\exp(\beta_1)$
$X_1 = 0$	$\exp(\beta_2)$	1.0

Logistic Regression – by the way…

```
. logistic will know cohort

Logistic regression                               Number of obs   =        750
                                                  LR chi2(2)      =       8.24
                                                  Prob > chi2     =     0.0163
Log likelihood = -384.63529                       Pseudo R2       =     0.0106

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |   .7879471   .1445962    -1.30   0.194     .5499117    1.129019
      cohort |   1.605702    .299915     2.54   0.011     1.113465    2.315546
------------------------------------------------------------------------------

. logistic know will cohort

Logistic regression                               Number of obs   =        750
                                                  LR chi2(2)      =       1.77
                                                  Prob > chi2     =     0.4133
Log likelihood = -510.58282                       Pseudo R2       =     0.0017

------------------------------------------------------------------------------
        know | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        will |    .787947   .1445962    -1.30   0.194     .5499117    1.129019
      cohort |    1.05728   .1588732     0.37   0.711     .7875594    1.419373
------------------------------------------------------------------------------
```

- Implication for confounder adjustment in case-control study?

The "dreaded" zero

$X_2 = 0$

	$Y = 1$	$Y = 0$	
$X_1 = 1$	**0**	112	134
$X_1 = 0$	32	153	185
	54	265	319

$X_2 = 1$

	$Y = 1$	$Y = 0$	
$X_1 = 1$	39	146	185
$X_1 = 0$	67	179	246
	106	325	431

$$P(Y=1 \mid X_1=1, X_2=0) = \pi(1,0) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} = 0$$

Q: What value of $\beta_0+\beta_1$ will give $\pi(1,0) = 0$?

The "dreaded" zero

```
. logit will know cohort [freq=count]
Logistic regression                               Number of obs   =        728
                                                  LR chi2(2)      =      36.75
                                                  Prob > chi2     =     0.0000
Log likelihood = -335.13143                       Pseudo R2       =     0.0520

------------------------------------------------------------------------------
        will |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        know |    -.75103   .2100932    -3.57   0.000    -1.162805    -.339255
      cohort |   1.047106   .2202911     4.75   0.000     .6153435    1.478869
       _cons |  -1.880522   .1952178    -9.63   0.000    -2.263142   -1.497902
------------------------------------------------------------------------------

. xi:logit will i.know*i.cohort [freq=count]
Logistic regression                               Number of obs   =        728
                                                  LR chi2(3)      =      57.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -324.54073                       Pseudo R2       =     0.0819

------------------------------------------------------------------------------
        will |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Iknow_1 |  -17.22683   .2302223   -74.83   0.000    -17.67805    -16.7756
  _Icohort_1 |   .5820088   .2414496     2.41   0.016     .1087763    1.055241
_IknoXcoh_~1 |   16.88948          .        .       .            .           .
       _cons |  -1.564702   .1943861    -8.05   0.000    -1.945692   -1.183712
------------------------------------------------------------------------------
Note: 112 failures and 0 successes completely determined.
```

- Point estimates "okay"; SE and CI not okay
- Use exact logistic regression - Stata `exlogistic`

The Logistic Regression Model Formally Defined

The binomial model that we use for binary outcome data can be specified by the following statements:

Systematic component:
$$P(Y=1|X) = \pi(X)$$
$$\text{logit}[\pi(X)] = \beta_0+\beta_1X_1+\ldots+\beta_pX_p$$

Random component:
$$Y \sim Bin(1,\pi(X))$$

Assume that ***Y***'s are independent

The statement $Y \sim Bin\big(1,\pi(X)\big)$ means:

"the variable ***Y*** is distributed as a *Binomial* random variable with $n = 1$ "trials" and success probability π(X)" This special case of the Binomial is called a Bernoulli random variable.

Using the Logistic Regression Model

Evaluation of the appropriateness of statistical models is an important feature of data analysis. For logistic regression this includes checking two components.

SYSTEMATIC:

• Which ***X***'s should be included in the model?

Choice of variables depends on goals of the analysis

- Prediction
- Hypothesis generating (EDA)
- Hypothesis confirmation (analytic study)

• Appropriate model forms (linear, quadratic, smooth) for continuous variables.

RANDOM:

• Are the observations independent?

Note: There is less concern about "error" terms than with linear regression.

Violations of Independence

Examples:

- "Longitudinal Data" – Repeated outcomes (over time) from a sample of individuals.
- Clustered sampling. (e.g. select female rats and measure outcome on each pup in a litter).
- Community randomized studies (analyze at the unit of randomization).
- "Multilevel Data" – e.g. patients, nested within doctors, nested within hospitals.

Formally: After knowing the covariates, $\boldsymbol{X}$, there is still some information (a variable $\boldsymbol{W}$), that leads one to expect that outcomes $\boldsymbol{Y}_i$ and $\boldsymbol{Y}_j$ will be more similar if they share the same value of W than if they do not. This variable, W, is a "cluster" identifier.

Examples:

- Longitudinal data – Subject ID
- Multilevel data – Doctor Name, Hospital Name

Impacts of Dependence

1. In general, $\hat{\beta}$ is consistent (dependence doesn't affect modeling of the means, $\pi(X)$).
2. Estimates of precision (standard errors of $\hat{\beta}$) may be wrong. They may be too large or too small depending on the situation.
3. Since the standard errors are not valid (random part of the model is mis-specified) we will have incorrect inference (hypothesis tests, confidence intervals).

Corrections for Dependence

1. Finesse it away! Take a summary for each person, cluster (if reasonable)
2. Find methods that can handle it appropriately:

- GEE methods (Liang and Zeger, 1986)
  - SAS proc genmod
  - STATA xtgee, xtlogit
  - R/S gee functions
- Mixed Models / Multilevel Models
  - SAS proc mixed, nlmixed, glimmix
  - STATA xtlogit, gllamm, cluster option
  - R/S lmer, nlmer

Many of these topics are covered in Biostat 536 (Autumn quarter) and 540 (Spring Quarter).

SUMMARY

1. The RANDOM part of the logistic model is simple: binary outcomes
2. The assumption of independence is important
3. Deciding which X's to include in the model depends on the scientific question you are asking
4. The logistic regression coefficient estimate $\hat{\beta}$ is obtained via maximum likelihood.
5. We will discuss how the maximized log likelihood, log L, is used for statistical inference.

Wald Tests

Most statistical packages produce tables of the form:

estimate	s.e.	Z
$\hat{\beta}_0$	s_0	$\hat{\beta}_0 / s_0$
$\hat{\beta}_1$	s_1	$\hat{\beta}_1 / s_1$
$\hat{\beta}_2$	s_2	$\hat{\beta}_2 / s_2$
$\vdots$		
$\hat{\beta}_p$	s_p	$\hat{\beta}_p / s_p$

from which we can obtain the following CI and "**Wald tests**":

- $\hat{\beta}_j \pm 1.96 s_j$ as a 95% confidence interval for β_j.
- $2 \times P[Z > |\hat{\beta}_j / s_j|] =$ p-value for testing $H_0: \beta_j = 0$

Q: What about hypotheses on multiple parameters e.g. H_0: $\beta_1 = \beta_2 = \ldots = 0$?

A: 1) Extend Wald test to multivariate situation

2) **Likelihood ratio tests**

"Nested" Models

When a scientific hypothesis can be formulated in terms of restrictions on a set of parameters (e.g. some β's equal to 0), we can formulate a pair of models: one that imposes the restriction (null model); and one that does not impose the restriction (alternative model). For example:

Model 1: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1$

Model 2: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

- Model 1 ("reduced") is a special case of Model 2 ("full").
- Model 1 is said to be *nested* within Model 2.
- Model 1 has a *subset* of the variables contained in Model 2.
- Model 1 is formed from Model 2 by the constraint : $H_0: \beta_2 = \beta_3 = 0$

By looking at the relative goodness-of-fit (as measured by log-likelihood) of these two models we can judge whether the additional flexibility in Model 2 is important.

Likelihood Ratio Tests

We can use the maximum likelihood fits from nested models to test if the "difference" between these models is significant. For example:

Model 1: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1$

Model 2: $\text{logit}[\pi(X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Model 1 is formed from Model 2 by the hypothesis: $H_0 : \beta_2 = \beta_3 = 0$

From the fitting of these models we obtain maximizing log likelihoods:

Model 1: $\log L_1$

Model 2: $\log L_2$

We can then use the *Likelihood Ratio Statistic*:

$$LR = 2 * (\log L_2 - \log L_1)$$

Which **under the null hypothesis** has a $\chi^2(d)$ distribution where d is the difference in the number of parameters for the two models.

Likelihood Ratio Tests: STATA Example

HIV vaccine trial data: Willingness to participate in an HIV Vaccine trial

```
. logistic will cohort
Logistic regression                               Number of obs   =        750
                                                  LR chi2(1)      =       6.54
                                                  Prob > chi2     =     0.0106
Log likelihood = -385.48728                       Pseudo R2       =     0.0084

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cohort |    1.60057   .2985952     2.52   0.012     1.110397    2.307124
------------------------------------------------------------------------------
. estimates store model1

. logistic will cohort know
Logistic regression                               Number of obs   =        750
                                                  LR chi2(2)      =       8.24
                                                  Prob > chi2     =     0.0163
Log likelihood = -384.63529                       Pseudo R2       =     0.0106

------------------------------------------------------------------------------
        will | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cohort |   1.605702    .299915     2.54   0.011     1.113465    2.315546
        know |   .7879471   .1445962    -1.30   0.194     .5499117    1.129019
------------------------------------------------------------------------------
. estimates store model2

. lrtest model2 model1
Likelihood-ratio test                             LR chi2(1)  =      1.70
(Assumption: model1 nested in model2)             Prob > chi2 =    0.1918
```

Logistic Regression: More Complex Covariates

Types of covariates :

- Nominal/qualitative variable (unordered categories)
  - o stratified adjustment with dummy variables

	X_1	X_2	logit(π(X))
Stratum 1	0	0	β_0
Stratum 2	1	0	$\beta_0+\beta_1$
Stratum 3	0	1	$\beta_0+\beta_2$

- Ordinal variable (ordered categories)
  - o stratified adjustment via dummy variables
  - o linear adjustment with an assigned score – implies linear increase in log odds ratio across categories (multiplicative increase in odds ratio)

$$X_j = \begin{cases} 1 & \text{stratum } 1 \\ 2 & \text{stratum } 2 \\ 3 & \text{stratum } 3 \end{cases}$$

  - o linear + quadratic or more complicated adjustment using score

Logistic Regression: More Complex Covariates

- Quantitative (continuous) variable.
  - o group and then treat as an ordinal variable, e.g.

$$\text{AGE} = \begin{cases} 1 & <40 \text{ years} \\ 2 & 40\text{-}49 \text{ years} \\ 3 & 50\text{-}59 \text{ years} \\ 4 & 60+ \text{ years} \end{cases}$$

  - o Group and treat as nominal (more flexibility)
  - o linear fit to measured values – implies log odds increases by β for each one unit change in covariate
  - o quadratic or more complicated adjustment ?

The Framingham Study

- 5209 subjects identified in 1948 in a small Massachusetts town
- Biennial exams for BP, serum cholesterol, relative weight
- Major endpoints include occurrence of coronary heart disease (CHD) and deaths from
  - CHD including sudden death (MI)
  - Cerebrovascular accident (CVA)
  - Cancer (CA)
  - Other causes
- We will look at follow up from first twenty years.

The Framingham Study

"*It is the function of longitudinal studies, like that of coronary heart disease in Framingham, to investigate the effects of a large variety of variables, both singly and jointly, on the effects of disease. The traditional approach of the epidemiologist, multiple cross-classification, quickly becomes impracticable as the number of variables to be investigated increases. Thus if 10 variables are under consideration, and each is to be studied at only 3 levels … there would be 59,049 cells in the multiple cross-classification.*" Truett, Cornfield & Kannel, 1967, p 511.

The Framingham Study

```
. infile lexam surv cause cexam chd cva ca oth sex age ht wt sc1
    sc2 dbp sbp mrw smoke using "C:\fram.dat"

. mvdecode ht  wt  sc1  sc2  mrw  smoke, mv(-1)
        ht: 6 missing values generated
        wt: 6 missing values generated
       sc1: 2037 missing values generated
       sc2: 626 missing values generated
       mrw: 6 missing values generated
     smoke: 36 missing values generated

. summarize
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       lexam |      5209    14.03532    3.449479          2         16
        surv |      5209    .3822231    .4859773          0          1
       cause |      5209    1.563064    2.368246          0          9
       cexam |      5209    2.561912    4.739291          0         16
         chd |      5209    .1161451    .3204296          0          1
-------------+--------------------------------------------------------
         cva |      5209    .0725667    .2594489          0          1
          ca |      5209    .1034748    .3046072          0          1
         oth |      5209    .2660779    .4419479          0          1
         sex |      5209    1.551545    .4973837          1          2
         age |      5209    44.06873    8.574954         28         62
-------------+--------------------------------------------------------
          ht |      5203    64.81318    3.582707       51.5       76.5
          wt |      5203    153.0867    28.91543         67        300
         sc1 |      3172    221.2393    45.01786         96        503
         sc2 |      4583    228.1778    44.81669        115        568
         dbp |      5209    85.35861    12.97309         50        160
-------------+--------------------------------------------------------
         sbp |      5209    136.9096     23.7396         82        300
         mrw |      5203    119.9575     19.9834         67        268
       smoke |      5173    9.366518    12.03145          0         60
```

High Blood Pressure and CHD

Outcome: CHD in first 20 years of Framingham study

Exposure:

$$\text{BPHI} = \begin{cases} 0 & \text{SBP} < 167 \text{ mm Hg} \\ 1 & \text{SBP} \geq 167 \text{mm Hg} \end{cases}$$

Potential Confounder:

$$\text{AGE} = \begin{cases} 1 & \text{40-44 years} \\ 2 & \text{45-49 years} \\ 3 & \text{50-54 years} \\ 4 & \text{55-59 years} \\ 5 & \text{60+ years} \end{cases}$$

The Framingham Study

Restrict analysis to males, 40+, with known values of baseline serum cholesterol, smoking and relative weight, who had no evidence of CHD at first exam:

```
. drop if sex>1 | age<40 | sc1==.| mrw==.| cexam==1|smoke==.
```

Create categorical variables for BP and age:

```
. gen bpg=sbp
. recode bpg min/126=1 127/146=2 147/166=3 167/max=4
. gen agp=age
. recode agp 40/44=1 45/49=2 50/54=3 55/59=4 60/62=5
. tab chd agp
```

chd	agp 1	2	3	4	5	Total
0	226	174	153	138	41	732
1	41	35	48	38	16	178
Total	267	209	201	176	57	910

```
. tab chd bpg
```

chd	bpg 1	2	3	4	Total
0	213	302	140	77	732
1	32	63	42	41	178
Total	245	365	182	118	910

Framingham: Logistic Regression Model

- Model 0 : First fit "null" model (constant term only),

$$\text{logit}[\pi(X)] = \beta_0$$

```
. logit chd

Iteration 0:   log likelihood = -449.76578
Iteration 1:   log likelihood = -449.76578

Logistic regression                               Number of obs   =        910
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -449.76578                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |  -1.413997   .0835709   -16.92   0.000    -1.577793   -1.250201
------------------------------------------------------------------------------

. estimates store model0
```

- Model 1: Next fit a model with binary BP only

```
. gen bphi=bpg
. recode bphi 1/3=0 4=1
```

$$\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{BPHI}$$

Framingham: Logistic Regression Model

```
. logit chd bphi
Logistic regression                               Number of obs   =        910
                                                  LR chi2(1)      =      17.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -440.99047                       Pseudo R2       =     0.0195

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bphi |    .934421   .2149493     4.35   0.000      .513128    1.355714
       _cons |  -1.564654   .0939467   -16.65   0.000    -1.748787   -1.380522
------------------------------------------------------------------------------
```

```
. estimates store model1
```

- Model 2: Next fit a model with age categories only

$$\text{logit}[\pi(\text{X})] = \beta_0 + \sum_{j=1}^{4} \beta_j AGP(j+1)$$

```
. xi: logit chd i.agp
Logistic regression                               Number of obs   =        910
                                                  LR chi2(4)      =       9.38
                                                  Prob > chi2     =     0.0522
Log likelihood = -445.07467                       Pseudo R2       =     0.0104

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iagp_2 |   .1032557    .251264     0.41   0.681    -.3892126     .595724
     _Iagp_3 |    .547726   .2370323     2.31   0.021     .0831513    1.012301
     _Iagp_4 |   .4172954   .2497543     1.67   0.095    -.0722139    .9068048
     _Iagp_5 |   .7659796   .3401548     2.25   0.024     .0992885    1.432671
       _cons |  -1.706963   .1697499   -10.06   0.000    -2.039667   -1.374259
------------------------------------------------------------------------------
```

```
. est store model2
```

Framingham: Logistic Regression Model

- LR for the (unadjusted) effect of high blood pressure on CHD

```
. lrtest model0 model1

Likelihood-ratio test                                 LR chi2(1)  =     17.55
(Assumption: model0 nested in model1)                 Prob > chi2 =    0.0000
```

- LR test for age effect H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$

```
. lrtest  model0  model2
Likelihood-ratio test                                 LR chi2(4)  =      9.38
(Assumption: model0 nested in model2)                 Prob > chi2 =    0.0522
```

- Also possible to test H_0: $\beta_1=\beta_2=\beta_3=\beta_4=0$ using a Wald type test (command must immediately follow model fitting)

```
. test  _Iagp_2  _Iagp_3  _Iagp_4  _Iagp_5

 ( 1)  _Iagp_2 = 0
 ( 2)  _Iagp_3 = 0
 ( 3)  _Iagp_4 = 0
 ( 4)  _Iagp_5 = 0
           chi2(  4) =    9.38
         Prob > chi2 =    0.0523
```

Framingham: Logistic Regression Model

- Model 3: BP and age groups

```
. xi: logit chd bphi i.agp, or
Logistic regression                               Number of obs   =        910
                                                  LR chi2(5)      =      24.93
                                                  Prob > chi2     =     0.0001
Log likelihood = -437.29993                       Pseudo R2       =     0.0277
------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bphi |   2.429429   .5289034     4.08   0.000     1.585593    3.722345
     _Iagp_2 |   1.061081   .2689046     0.23   0.815     .6457025     1.74367
     _Iagp_3 |   1.626755   .3897112     2.03   0.042     1.017198    2.601591
     _Iagp_4 |   1.353106   .3438848     1.19   0.234     .8222487    2.226692
     _Iagp_5 |   2.006542   .6909787     2.02   0.043     1.021706    3.940674
------------------------------------------------------------------------------

. estimates store model3

. lrtest model2 model3
Likelihood-ratio test                             LR chi2(1)  =     15.55
(Assumption: model2 nested in model3)             Prob > chi2 =    0.0001
```

Framingham: Logistic Regression Model

- Model 3: BP and age groups

```
. xi: logit chd bphi i.agp
```

```
Logistic regression                               Number of obs   =        910
                                                  LR chi2(5)      =      24.93
                                                  Prob > chi2     =     0.0001
Log likelihood = -437.29993                       Pseudo R2       =     0.0277

------------------------------------------------------------------------------
         chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        bphi |   .8876561   .2177069     4.08   0.000     .4609584    1.314354
     _Iagp_2 |   .0592878   .2534252     0.23   0.815    -.4374165    .5559921
     _Iagp_3 |   .4865875   .2395635     2.03   0.042     .0170517    .9561233
     _Iagp_4 |   .3024023   .2541448     1.19   0.234    -.1957124    .8005171
     _Iagp_5 |   .6964127    .344363     2.02   0.043     .0214736    1.371352
       _cons |  -1.797441   .1729173   -10.39   0.000    -2.136352   -1.458529
------------------------------------------------------------------------------
```

```
. predict logit3, xb
```

Framingham: Logistic Regression Model

Model 3:

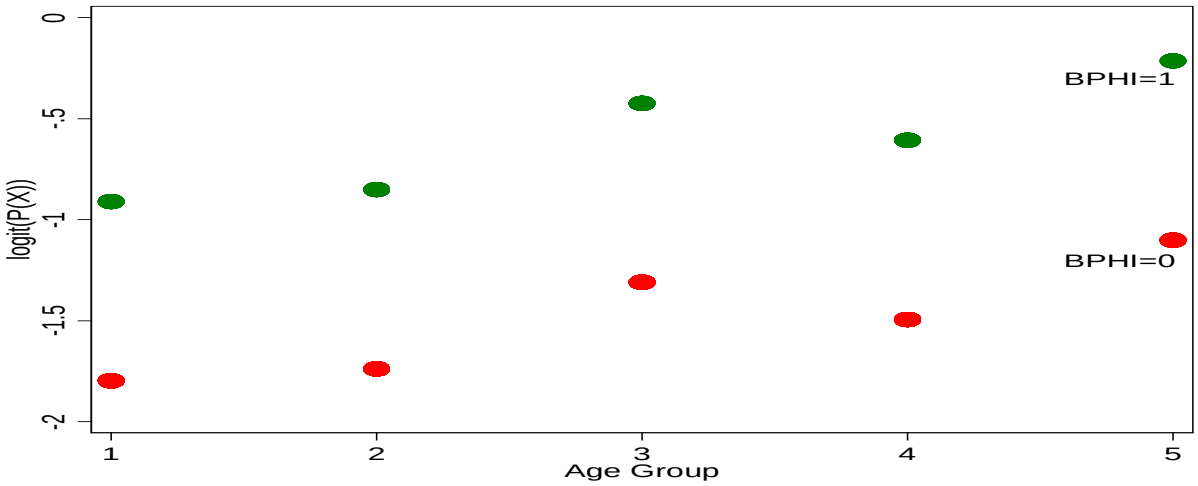
$$\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{BPHI} + \sum_{j=2}^{5} \beta_j \text{AGP}(j)$$

logit structure:

AGE	SBP<167	SBP≥167
40-44 years	β_0	$\beta_0 + \beta_1$
45-49 years	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2$
50-54 years	$\beta_0 + \beta_3$	$\beta_0 + \beta_1 + \beta_3$
55-59 years	$\beta_0 + \beta_4$	$\beta_0 + \beta_1 + \beta_4$
60+ years	$\beta_0 + \beta_5$	$\beta_0 + \beta_1 + \beta_5$

Framingham: Logistic Regression Model

```
. twoway (scatter logit3 agp if bphi==0, mcol(red) msize(large) msymbol(circle)
mlwidth(medthick) )  (scatter logit3 agp if bphi==1, msize(large) msymbol(circle)
mlwidth(medthick) mcol(green) ), ytitle(logit(P(X))) xtitle(Age Group) legend(off)
scheme(s1mono) text(-1.2 4.8 "BPHI=0") text(-.3 4.8 "BPHI=1")
```



Framingham: Logistic Regression Model

Model 4:

$$\text{logit}[\pi(X)] = \beta_0 + \beta_1 \text{BPHI} + \sum_{j=2}^{5} \beta_j \text{AGP(j)} + \sum_{k=2}^{5} \beta_{k+4} BPHI * AGP(k)$$

logit structure:

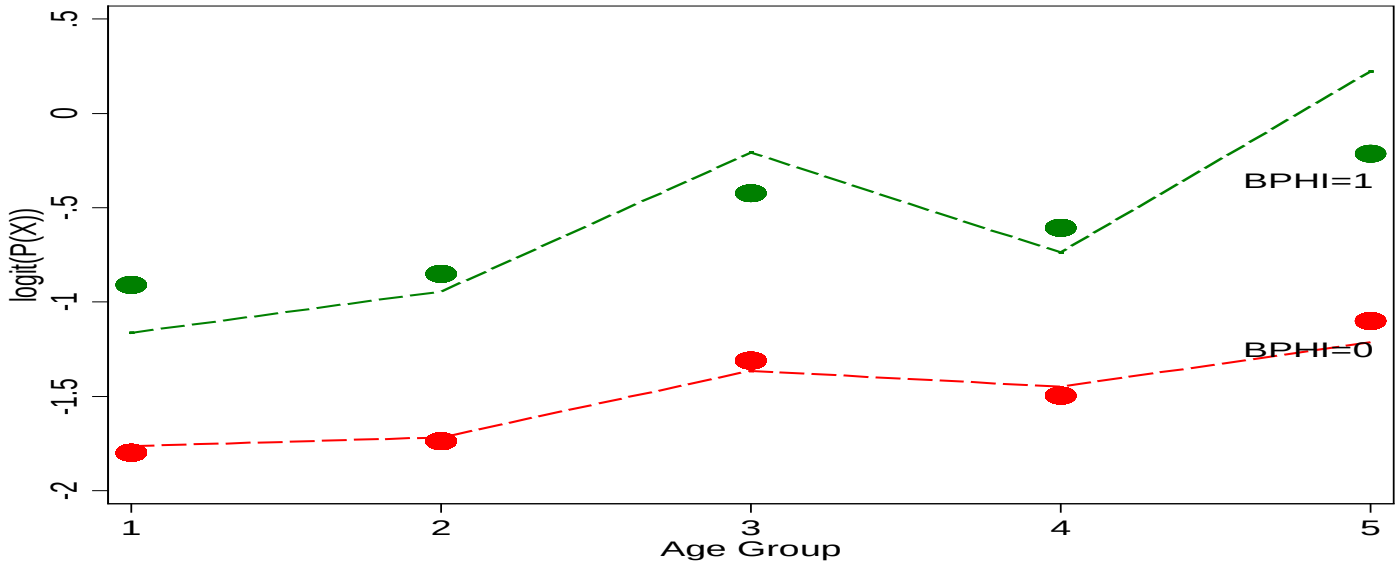
	SBP<167	SBP≥167
40-44 years	β_0	$\beta_0 + \beta_1$
45-49 years	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_6$
50-54 years	$\beta_0 + \beta_3$	$\beta_0 + \beta_1 + \beta_3 + \beta_7$
55-59 years	$\beta_0 + \beta_4$	$\beta_0 + \beta_1 + \beta_4 + \beta_8$
60+ years	$\beta_0 + \beta_5$	$\beta_0 + \beta_1 + \beta_5 + \beta_9$

Framingham: Logistic Regression Model

```
. xi: logit chd i.bphi*i.agp

Logistic regression                               Number of obs   =        910
                                                  LR chi2(9)      =      26.40
                                                  Prob > chi2     =     0.0018
Log likelihood = -436.56689                       Pseudo R2       =     0.0293

------------------------------------------------------------------------------
         chd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    _Ibphi_1 |   1.822917   .9901638     1.11   0.269     .6286587    5.285897
     _Iagp_2 |   1.047009   .2860925     0.17   0.866     .6128603    1.788706
     _Iagp_3 |   1.490268   .3897851     1.53   0.127     .8925433     2.48828
     _Iagp_4 |   1.369565    .383166     1.12   0.261     .7914808    2.369873
     _Iagp_5 |   1.734234   .6727305     1.42   0.156     .8108049    3.709361
_IbphXagp_~2 |   1.188571   .8698337     0.24   0.813     .2831975    4.988399
_IbphXagp_~3 |   1.744653   1.196498     0.81   0.417     .4549333    6.690683
_IbphXagp_~4 |    1.11746   .7702859     0.16   0.872     .2893899    4.315001
_IbphXagp_~5 |   2.306494   2.142652     0.90   0.368     .3734357    14.24586
------------------------------------------------------------------------------

. estimates store model4
. lrtest model4 model3, stats

Likelihood-ratio test                             LR chi2(4)  =      1.47
(Assumption: model3 nested in model4)             Prob > chi2 =    0.8326

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
      model3 |    910   -449.7658   -437.2999      6     886.5999    915.4805
      model4 |    910   -449.7658   -436.5669     10     893.1338    941.2682
-----------------------------------------------------------------------------

. predict logit4, xb
```

Framingham: Logistic Regression Model

```
. twoway (scatter logit3 agp if bphi==0, mcol(red) msize(large) msymbol(circle)
mlwidth(medthick) )  (scatter logit3 agp if bphi==1, msize(large) msymbol(circle)
mlwidth(medthick) mcol(green) ) (line logit4 agp if bphi==0, sort  lcol(red)
lpat(dash))  (line logit4 agp if bphi==1, sort lcol(green) lpat(dash) ),
ytitle(logit(P(X))) xtitle(Age Group) legend(off) scheme(s1mono) text(-1.25 4.8
"BPHI=0") text(-.35 4.8 "BPHI=1")
```



Framingham: Model Summary

Summary of models

Model	Description	#pars	BP OR	logL	AIC	LR Test	Test statistic (df, p)
0	Intercept only	1	-	-449.8	901.5	-	-
1	BP only	2	2.55	-441.0	886.0	1 vs 0	17.55 (1,<0.00005)
2	Factor Age	5	-	-445.1	900.1	2 vs 0	9.38 (4, .052)
3	BP+Factor Age	6	2.43	-437.3	886.6	3 vs 2	15.55 (1, .0001)
4	BP +Factor Age+ Interaction	10	1.82 2.17 3.18 2.04 4.20	-436.6	893.1	4 vs 3	1.47 (4, .83)

Conclusion?

Ille-et-Vilaine Case-control Study

Cases: 200 males diagnosed in one of regional hospitals in French department of Ille-et-Vilaine (Brittany) between Jan 1972 and Apr 1974

Controls: Random sample of 778 adult males from electoral lists in each commune (775 with usable data)

Exposures: Detailed dietary interview on consumption of various foods, tobacco and alcoholic beverages

Background: Brittany was a known "hot spot" of esophageal cancer in France and also had high levels of alcohol consumption, particularly of the local (often homemade) apple brandy known as Calvados

Reference: Tuyns AJ, Pequinot G, Jensen OM. (1977) Le cancer de l'oesophage en Ille-et-Vilaine en fonction des niveaux de consommation d'alcohol et de tabac. *Bull Canc* **64**: 45-60.

Ille-et-Vilaine Case-control Study Using STATA

```
. use "ille-et-vilaine.dta", clear

. tabodds case tob [freq=count], or

------------------------------------------------------------------------------
        tob |  Odds Ratio       chi2       P>chi2     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        0-9 |    1.000000          .           .               .           .
      10-19 |    1.867329      10.46       0.0012        1.271188    2.743040
      20-29 |    1.910256       7.72       0.0055        1.200295    3.040153
        30+ |    3.483409      25.31       0.0000        2.074288    5.849783
------------------------------------------------------------------------------
Test of homogeneity (equal odds): chi2(3)  =    29.33
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    26.93
                                  Pr>chi2  =   0.0000

. logistic case tob  [freq=count]

Logistic regression                               Number of obs   =        975
                                                  LR chi2(1)      =      25.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -482.05896                       Pseudo R2       =     0.0256

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         tob |   1.474687   .1123596     5.10   0.000     1.270121    1.712201
------------------------------------------------------------------------------
```

The estimated esophageal cancer odds ratio comparing (any) adjacent groups of tobacco consumption is 1.47. How do we interpret this odds ratio?

Ille-et-Vilaine Case-control Study

For this logistic model, how would you estimate the esophageal cancer OR comparing tobacco consumption 10-19 with 0-9 g/day?

1.47

For this logistic model, how would you estimate the esophageal cancer OR comparing tobacco consumption 20-29 with 0-9 g/day?

1.47^2=2.17

For this logistic model, how would you estimate the esophageal cancer OR comparing tobacco consumption 30+ with 0-9 g/day?

1.47^3=3.21

For this logistic model, how would you estimate the esophageal cancer OR comparing tobacco consumption 30+ with 10-19g/day?

1.47^2=2.17

Ille-et-Vilaine Case-control Study

```
. logit case
…
. estimates store model0
. xi: logistic case i.age
Logistic regression                               Number of obs   =        975
                                                  LR chi2(5)      =     121.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -434.22195                       Pseudo R2       =     0.1223

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |   5.447368   5.777946     1.60   0.110     .6812858    43.55562
     _Iage_3 |   31.67665   32.24812     3.39   0.001     4.307063    232.9685
     _Iage_4 |    52.6506   53.37904     3.91   0.000     7.218137    384.0445
     _Iage_5 |   59.66981   60.74305     4.02   0.000     8.114154    438.7995
     _Iage_6 |   48.22581   50.98864     3.67   0.000     6.071737    383.0417
------------------------------------------------------------------------------

. logit
Logit estimates                                   Number of obs   =        975
                                                  LR chi2(5)      =     121.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -434.22195                       Pseudo R2       =     0.1223

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |   1.695133   1.060686     1.60   0.110    -.3837734    3.774039
     _Iage_3 |    3.45558   1.018041     3.39   0.001     1.460256    5.450903
     _Iage_4 |   3.963678   1.013835     3.91   0.000     1.976597    5.950758
     _Iage_5 |   4.088826   1.017986     4.02   0.000      2.09361    6.084043
     _Iage_6 |   3.875894    1.05729     3.67   0.000     1.803645    5.948144
       _cons |  -4.744932   1.004331    -4.72   0.000    -6.713384    -2.77648
------------------------------------------------------------------------------

. estimates store model1
. lrtest model0 model1
Likelihood-ratio test                             LR chi2(5)  =    121.04
(Assumption: model0 nested in model1)             Prob > chi2 =    0.0000
```

Ille-et-Vilaine Case-control Study

```
. xi: logistic case i.age i.tob
Logistic regression                               Number of obs   =        975
                                                  LR chi2(8)      =     157.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -415.90235                       Pseudo R2       =     0.1594

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |   6.035932   6.433579     1.69   0.092     .7472494    48.75544
     _Iage_3 |   36.20831   37.10768     3.50   0.000     4.858071    269.8688
     _Iage_4 |   61.79318   63.10318     4.04   0.000     8.350141    457.2853
     _Iage_5 |   83.56952   85.86284     4.31   0.000     11.15546    626.0487
     _Iage_6 |   60.45383   64.52342     3.84   0.000     7.463141    489.6954
     _Itob_2 |   1.835482   .3781838     2.95   0.003     1.225655    2.748731
     _Itob_3 |   1.945172   .4877329     2.65   0.008     1.189947    3.179717
     _Itob_4 |   5.706139   1.725687     5.76   0.000     3.154398     10.3221
------------------------------------------------------------------------------

. logit
Logit estimates                                   Number of obs   =        975
                                                  LR chi2(8)      =     157.68
                                                  Prob > chi2     =     0.0000
Log likelihood = -415.90235                       Pseudo R2       =     0.1594

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |    1.79773    1.06588     1.69   0.092    -.2913563    3.886817
     _Iage_3 |   3.589289   1.024839     3.50   0.000     1.580642    5.597936
     _Iage_4 |   4.123793     1.0212     4.04   0.000     2.122278    6.125308
     _Iage_5 |   4.425679   1.027442     4.31   0.000     2.411929    6.439428
     _Iage_6 |    4.10188   1.067317     3.84   0.000     2.009976    6.193784
     _Itob_2 |   .6073073   .2060406     2.95   0.003     .2034753    1.011139
     _Itob_3 |   .6653504   .2507403     2.65   0.008     .1739085    1.156792
     _Itob_4 |   1.741543   .3024264     5.76   0.000     1.148798    2.334287
       _cons |  -5.367792   1.017907    -5.27   0.000    -7.362852   -3.372731
------------------------------------------------------------------------------

. estimates store model2
```

Ille-et-Vilaine Case-control Study

```
. lrtest model1 model2
likelihood-ratio test                                 LR chi2(3)  =     36.64
(Assumption: model1 nested in model2)                 Prob > chi2 =    0.0000

. xi: logistic case i.age i.tob i.alc
Logistic regression                                   Number of obs   =       975
                                                      LR chi2(11)     =    285.62
                                                      Prob > chi2     =    0.0000
Log likelihood = -351.93592                           Pseudo R2       =    0.2887

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     _Iage_2 |   7.249153   8.003167     1.79   0.073      .8328154    63.09948
     _Iage_3 |   43.65363   46.62157     3.54   0.000       5.38204    354.0738
     _Iage_4 |   76.33883   81.30049     4.07   0.000      9.467161    615.5611
     _Iage_5 |    133.808   144.0209     4.55   0.000      16.22978    1103.193
     _Iage_6 |   124.7787   139.9078     4.30   0.000      13.85905    1123.434
     _Itob_2 |   1.549686   .3538287     1.92   0.055      .9905925    2.424334
     _Itob_3 |   1.669657   .4557781     1.88   0.060      .9778419    2.850925
     _Itob_4 |   5.160313   1.775732     4.77   0.000      2.628853    10.12945
     _Ialc_2 |   4.198086   1.049782     5.74   0.000      2.571568    6.853377
     _Ialc_3 |    7.24794   2.063937     6.96   0.000      4.147867    12.66497
     _Ialc_4 |   36.70338   14.13218     9.36   0.000      17.25685    78.06397
------------------------------------------------------------------------------

. est store model3
. lrtest model2 model3, stats
Likelihood-ratio test                                 LR chi2(3)  =    127.93
(Assumption: model2 nested in model3)                 Prob > chi2 =    0.0000

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
      model2 |    975   -494.7442   -415.9023      9     849.8047    893.7466
      model3 |    975   -494.7442   -351.9359     12     727.8718    786.4611
-----------------------------------------------------------------------------
```

Ille-et-Vilaine Case-control Study

- **Do we need a more complex model?**

```
. xi: logistic case i.age i.tob*i.alc

. estimates store model4

. lrtest model3 model4

Likelihood-ratio test                                 LR chi2(9)  =      5.45
(Assumption: model3 nested in model4)                 Prob > chi2 =    0.7934
```

- **Can we simplify the model?**

```
. xi: logistic case i.age tob alc

. estimates store model5

. lrtest model3 model5

Likelihood-ratio test                                 LR chi2(4)  =      8.78
(Assumption: model5 nested in model3)                 Prob > chi2 =    0.0667
```

Note: Not obvious, but model 5, that treats the categories of tob and alc as linear, is nested in model 3

Ille-et-Vilaine Case-control Study
Goodness-of-Fit

To check goodness-of-fit of Model 3, compare expected number of cases with observed:

```
. xi: logistic case i.age i.tob i.alc
. predict fitted, pr
. preserve
. collapse (sum) fitted case, by(alc tob)
. list, noobs

  |    alc     tob     fitted  case|
  |--------------------------------|
  |   0-39     0-9   13.68551    9 |
  |   0-39   10-19   7.091438   10 |
  |   0-39   20-29    3.44484    5 |
  |   0-39     30+   4.778213    5 |
  |  40-79     0-9   30.96526   34 |
  |--------------------------------|
  |  40-79   10-19   18.84762   17 |
  |  40-79   20-29   16.02676   15 |
  |  40-79     30+   9.160364    9 |
  | 80-119     0-9   17.83326   19 |
  | 80-119   10-19   19.91053   19 |
  |--------------------------------|
  | 80-119   20-29   6.668531    6 |
  | 80-119     30+   6.587675    7 |
  |   120+     0-9   15.51597   16 |
  |   120+   10-19   12.15041   12 |
  |   120+   20-29   6.859873    7 |
  |--------------------------------|
  |   120+     30+   10.47375   10 |
  +--------------------------------+
```

Ille-et-Vilaine Case-control Study
Goodness-of-Fit

For grouped data the Pearson Chi-squared statistic has a Chi-squared distribution with J-p degrees of freedom, where J=number of covariate patterns and p is the number of fitted coefficients in the model (including the intercept).

```
. xi: logistic case i.age i.tob i.alc
. lfit

Logistic model for case, goodness-of-fit test

      number of observations =       975
 number of covariate patterns =        88
            Pearson chi2(76) =        86.56
                 Prob > chi2 =         0.1913
```

Ille-et-Vilaine Case-control Study
Goodness-of-Fit

When J increases with N (e.g. quantitative covariate): the Hosmer-Lemeshow test is based on collapsing the data into G groups based on ordering of the fitted probabilities and then calculating the Pearson Chi-square statistic:

```
. lfit, group(10) table

Logistic model for case, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0070 |     0 |   0.3 |    99 |  98.7 |    99 |
  |     2 | 0.0299 |     1 |   1.9 |   125 | 124.1 |   126 |
  |     3 | 0.0456 |     4 |   3.4 |    76 |  76.6 |    80 |
  |     4 | 0.0717 |     4 |   6.7 |   100 |  97.3 |   104 |
  |     5 | 0.1193 |    12 |  12.1 |    96 |  95.9 |   108 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.1790 |    12 |  10.8 |    56 |  57.2 |    68 |
  |     7 | 0.2450 |    25 |  24.8 |    82 |  82.2 |   107 |
  |     8 | 0.3590 |    34 |  31.4 |    59 |  61.6 |    93 |
  |     9 | 0.4891 |    44 |  39.9 |    49 |  53.1 |    93 |
  |    10 | 0.9625 |    64 |  68.7 |    33 |  28.3 |    97 |

      number of observations =       975
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =      4.31
                Prob > chi2 =      0.8284
```

Ille-et-Vilaine Case-control Study

Using model 3 we can construct a table of estimated esophageal cancer ORs by Alcohol and Tobacco, adjusted for Age:

Alcohol (g/day)	**Tobacco (g/day)**			
	0-9	**10-19**	**20-29**	**30+**
0-39	1.0	1.55	1.67	5.16
40-79	4.20	6.51	7.01	21.66
80-119	7.25	11.23	12.10	37.40
120+	36.70	56.88	61.28	189.40

Note: each odds ratio in the interior of the table is obtained by multiplying together the respective marginal odds ratios

Linear Combinations

In some situations we want to compute confidence intervals and/or tests for combinations of the β_j's

Linear combinations are expressions of the form

$$a_1\beta_1 + a_2\beta_2 + a_3\beta_3 + \ldots$$

For example: Suppose, in model 1, we wish to consider pooling data from 65-74 year olds with those 75+. This is equivalent to testing H_o: $\beta_{Iage\_6} - \beta_{Iage\_5} = 0$.

Interpretation of $\beta_{Iage\_6} - \beta_{Iage\_5}$?

Linear Combinations – Example

H_0: $\beta_{Iage\_6} - \beta_{Iage\_5} = 0$.

Note that $\hat{\beta}_{Iage6} - \hat{\beta}_{Iage5} = 3.876 - 4.088 = -.212$

```
. test _Iage_6 - _Iage_5 = 0
 ( 1) - _Iage_5 + _Iage_6 = 0

           chi2(  1) =    0.33
         Prob > chi2 =    0.5648
```

```
. lincom _Iage_6 - _Iage_5, or
```

```
 ( 1) - _Iage_5 + _Iage_6 = 0

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .8082111   .2989255    -0.58   0.565     .3914703    1.668595
------------------------------------------------------------------------------
```

Conclusion?

Regression Analysis & Model Building

- The first step is to identify the scientific question.
- *A priori* specification of the goals of analysis is crucial, particularly for interpretation of p-values

Q: What are the statistical goals of regression analysis?

A: Estimation, Testing and/or Prediction

- **Estimation** of the "effect" of one variable (exposure), called the *predictor of interest* (POI), after "adjusting", or controlling for other measured variables.
  - Remove confounding effects.
  - Identify effect modification.
- **Testing** of a limited number of hypotheses, specified *a priori* in the study protocol, regarding the association of response and POI
- **Prediction** of outcomes for "new" subjects from the same population, given covariates.

Modeling Goals

1. CDA: "Effect" estimation (exposure → disease)
   - Logistic regression for **inference**
   - Parsimonious description of associations
   - Parameter interpretation desired
   - Avoid automated methods
2. Good prediction of the outcome
   - Logistic regression for **prediction**
   - Less focus on interpreting parameters (black box)
   - Tradeoff between *bias* (from underfitting) and *variance* (from overfitting)
   - Automated methods useful
   - Now widely used with microarrays: identify gene expression "signatures" for diagnosis or prognosis
3. "Data mining"
   - Exploratory Data Analysis (EDA)
   - Hypothesis generating
   - Write / create study design and protocol
   - Confirmatory studies to follow

Guidelines for CDA – Effect Estimation

- Carefully decide the scientific question that you want answered.

  Outcome: ***D***
  Exposure of Interest: ***E***

- Carefully (and parsimoniously) model the exposure effect
- Restrict attention to clinically or biologically meaningful variables.

  study goals
  literature review
  theoretical basis
  Define these variables as $C_1, C_2, \ldots, C_p$

- Use a "rich" model for confounder control
- Structure interaction analyses *a priori*
- Make sure that your model is "hierarchically well-formulated" i.e. don't include interactions without corresponding main effects

Model Building Strategies

Kleinbaum, *Logistic Regression* Chapter 6:

"*Most epidemiologic research studies in the literature … provide a minimum of information about modeling methods used in the data analysis*."

"*Without meaningful information about the modeling strategy used, it is difficult to assess the validity of the results provided. Thus, there is need for guidelines regarding modeling strategy to help researchers know what information to provide*."

"*In practice, most modeling strategies are ad hoc; in other words, researchers often make up a strategy as they go along in their analysis. The general guidelines that we recommend here encourage more consistency in the strategy used by different researchers*."

Information often ***not*** provided:

1. how variables chosen / selected
2. how confounders assessed
3. how effect modifiers assessed

Model Building Strategies

Classification of Variables:

- Response variable.
  - Dependent or outcome variable.
- Predictor of interest (POI)
  - Exposure variable.
  - Treatment assignment.
- Confounding variables.
  - Associated with response and POI.
  - Not intermediate.
- Design variables
  - Used for (frequency) matching or stratification
  - Must be considered in analysis
- Effect modifiers.
  - Identifies subgroups

Guidelines for Effect Estimation

- Plan a short (2-3) series of analyses
  - Main effect of exposure adjusted for
    - No confounders
    - Primary confounders
    - Primary and secondary confounders
  - (Possibly) Interaction of exposure effect with short list of potential effect modifiers

- Write down the plan and adhere to it

- Having completed these steps you will have identified the statistical model that is of interest. There is no data-dependent variable selection.

Guidelines for Effect Estimation
Model Simplification

Controversy remains regarding next step:

1. Use "backwards elimination" to discard
   - o interactions that are not statistically significant (LR test)
   - o confounders that do not change effect estimates by specified amount

   or

2. Stay put!

Guidelines for Effect Estimation
Model Simplification

- Plan how you will simplify the model e.g.,
  1) Check for interaction between POI and other covariates
  2) Check for confounding of POI
     - do not assess confounding for EM
     - generally difficult to assess confounding in presence of any EM
  3) Discard covariates which are not EM or confounders!
- Decide on strategy for dropping variables a priori (i.e., "I will drop interactions if …", "I will include a confounder if …")
  - Effect modification decision often based on p-value
  - Confounding decision based on *change in coefficient*
  - Dropping variables *may* increase precision
- Write the plan down and stick to it

Guidelines for Effect Estimation
Model Simplification

Kleinbaum (1994):

"In general, …, the method for assessing confounding when there is no interaction is to monitor changes in the effect measure corresponding to different subsets of potential confounders in the model."

"To evaluate how much of a change is a **meaningful** change when considering the collection of coefficients … is quite **subjective**."

Recommendations:

- 10% change in the measure of interest
  - Mickey & Greenland (1989) AJE
  - Maldonado & Greenland (1993) AJE

Statistical Thinking

"*As a medical statistician, I am appalled by the large number of irreproducible results published in the medical literature. There is a general, and likely correct, perception that this problem is associated more with statistical, as opposed to laboratory, research. I am convinced, however, that results of clinical and epidemiological investigations could become more reproducible if only the investigators would apply more rigorous statistical thinking and adhere more closely to well established principles of the scientific method. While I agree that the investigative cycle is an iterative process, I believe that it works best when it is hypothesis driven*."

"*The epidemiology literature is replete with irreproducible results stemming from the failure to clearly distinguish between analyses that were specified in the protocol and that test the a priori hypotheses whose specification was needed to secure funding, and those that were performed post-hoc as part of a serendipitous process of data exploration*."

Breslow (1999)

Statistical Thinking

"It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient a priori importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as data dredging or blind fishing and carry considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it. There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it – findings will inevitably be biased. Confounders should be chosen a priori and not on the basis of statistical significance."

Clayton and Hills, Statistical Methods in Epidemiology, 1993, p. 273

Statistical Thinking

- "When you go looking for something specific, your chances of finding it are very bad, because of all the things in the world, you're only looking for one of them.

- "When you go looking for anything at all, your chances of finding it are very good, because of all the things in the world, you're sure to find some of them."

Daryl Zero in "The Zero Effect"

Multiple Comparisons Problem

- "When you go looking for something specific, your chances of finding ~~it~~ **[a spurious association by chance]** are very bad, because of all the things in the world, you're only looking for one of them.

- "When you go looking for anything at all, your chances of finding ~~it~~ **[a spurious association by chance]** are very good, because of all the things in the world, you're sure to find some of them."

Logistic Regression Model Building
Low Birthweight Case Study

Is maternal smoking a risk factor for having a child born with low birth weight?

The variables identified in the table below have been shown to be associated with low birth weight in the obstetrical literature.

Table: Code Sheet for the Variables in the Low Birth Weight Data Set.

Columns	Variable	Abbreviation
2-4	Identification Code	ID
10	Low Birth Weight (0 = Birth Weight ge 2500g, l = Birth Weight < 2500g)	LBW
17-18	Age of the Mother in Years	AGE
23-25	Weight in Pounds at the Last Menstrual Period	LWT
32	Race (1 = White, 2 = Black, 3 = Other)	RACE
40	Smoking Status During Pregnancy (1 = Yes, 0 = No)	SMOKE
48	History of Premature Labor (0 = None, 1 = One, etc.)	PTL
55	History of Hypertension (1 = Yes, 0 = No)	HYPER
61	Presence of Uterine Irritability (1 = Yes, 0 = No)	URIRR
67	Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)	PVFT
73-76	Birth Weight in Grams	BWT

Logistic Regression Model Building
Low Birthweight Case Study

Scientific question: *Is maternal smoking a risk factor for having a child born with low birth weight?*

Outcome: LBW

Exposure: smoking during pregnancy

Potential confounders: mother's age, weight, race, history of premature labor, history of hypertension

Potential effect modifier: history of hypertension

Logistic Regression Model Building
Low Birthweight Case Study

```
. infile id lbw age lwt race smoke ptl hyper urirr pvft weight using "lowbwt.dat"
. generate agecat=age
. recode agecat min/19=1 20/24=2 25/29=3 30/max=4
. label define newgps 1 "<20" 2 "20-24" 3 "25-29" 4 "30+"
. label values agecat newgps
. generate wcat=lwt
. recode wcat min/105=1 106/120=2 121/130=3 131/150=4 151/max=5
. label define wgps 1 "<=105" 2 "106-120" 3 "121-130" 4 "131-150" 5 "151+"
. label values wcat wgps
. generate anyptl=ptl
. recode anyptl 1/max=1
. label define agps 0 "0" 1 "1+"
. label values anyptl agps
. tabulate lbw smoke, chi2 row col
```

lbw	smoke 0	1	Total
0	86	44	130
	66.15	33.85	100.00
	74.78	59.46	68.78
1	29	30	59
	49.15	50.85	100.00
	25.22	40.54	31.22
Total	115	74	189
	60.85	39.15	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 4.9237 Pr = 0.026

Logistic Regression Model Building
Low Birthweight Case Study

```
. tabulate lbw agecat, chi2 row
```

lbw	agecat <20	20-24	25-29	30+	Total
0	36	44	27	23	130
	27.69	33.85	20.77	17.69	100.00
	70.59	63.77	64.29	85.19	68.78
1	15	25	15	4	59
	25.42	42.37	25.42	6.78	100.00
	29.41	36.23	35.71	14.81	31.22
Total	51	69	42	27	189
	26.98	36.51	22.22	14.29	100.00
	100.00	100.00	100.00	100.00	100.00

Pearson chi2(3) = 4.6641 Pr = 0.198

```
. tabulate lbw wcat, chi2 row
```

lbw	wcat <=105	106-120	121-130	131-150	151+	Total
0	17	41	20	23	29	130
	13.08	31.54	15.38	17.69	22.31	100.00
	45.95	74.55	64.52	76.67	80.56	68.78
1	20	14	11	7	7	59
	33.90	23.73	18.64	11.86	11.86	100.00
	54.05	25.45	35.48	23.33	19.44	31.22
Total	37	55	31	30	36	189
	19.58	29.10	16.40	15.87	19.05	100.00
	100.00	100.00	100.00	100.00	100.00	100.00

Pearson chi2(4) = 13.2923 Pr = 0.010

Logistic Regression Model Building
Low Birthweight Case Study

```
. tabulate lbw race, chi2 row

           |                race
       lbw |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |        73         15         42 |       130
           |     56.15      11.54      32.31 |    100.00
           |     76.04      57.69      62.69 |     68.78
-----------+---------------------------------+----------
         1 |        23         11         25 |        59
           |     38.98      18.64      42.37 |    100.00
           |     23.96      42.31      37.31 |     31.22
-----------+---------------------------------+----------
     Total |        96         26         67 |       189
           |     50.79      13.76      35.45 |    100.00
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   5.0048   Pr = 0.082
```

```
. tabulate lbw anyptl, chi2 row

           |        anyptl
       lbw |         0          1+ |     Total
-----------+----------------------+----------
         0 |       118          12 |       130
           |     90.77        9.23 |    100.00
           |     74.21       40.00 |     68.78
-----------+----------------------+----------
         1 |        41          18 |        59
           |     69.49       30.51 |    100.00
           |     25.79       60.00 |     31.22
-----------+----------------------+----------
     Total |       159          30 |       189
           |     84.13       15.87 |    100.00
           |    100.00      100.00 |    100.00

          Pearson chi2(1) =  13.7590   Pr = 0.000
```

Logistic Regression Model Building
Low Birthweight Case Study

```
. tabulate lbw hyper, chi2 row col
```

lbw	hyper 0	1	Total
0	125	5	130
	96.15	3.85	100.00
	70.62	41.67	68.78
1	52	7	59
	88.14	11.86	100.00
	29.38	58.33	31.22
Total	177	12	189
	93.65	6.35	100.00
	100.00	100.00	100.00

Pearson chi2(1) = 4.3880 Pr = 0.036

```
. tabulate agecat smoke, chi2 row col
```

smoke	agecat <20	20-24	25-29	30+	Total
0	28	44	26	17	115
	24.35	38.26	22.61	14.78	100.00
	54.90	63.77	61.90	62.96	60.85
1	23	25	16	10	74
	31.08	33.78	21.62	13.51	100.00
	45.10	36.23	38.10	37.04	39.15
Total	51	69	42	27	189
	26.98	36.51	22.22	14.29	100.00
	100.00	100.00	100.00	100.00	100.00

Pearson chi2(3) = 1.0742 Pr = 0.783

Logistic Regression Model Building
Low Birthweight Case Study

```
. tabulate smoke wcat, chi2 row
```

smoke	wcat <105	106-120	121-130	131-150	151+	Total
0	19	35	18	19	24	115
	16.52	30.43	15.65	16.52	20.87	100.00
1	18	20	13	11	12	74
	24.32	27.03	17.57	14.86	16.22	100.00
Total	37	55	31	30	36	189
	19.58	29.10	16.40	15.87	19.05	100.00

Pearson chi2(4) = 2.2704 Pr = 0.686

```
. tabulate smoke race, chi2 row
```

smoke	race 1	2	3	Total
0	44	16	55	115
	38.26	13.91	47.83	100.00
1	52	10	12	74
	70.27	13.51	16.22	100.00
Total	96	26	67	189
	50.79	13.76	35.45	100.00

Pearson chi2(2) = 21.7790 Pr = 0.000

Logistic Regression Model Building
Low Birthweight Case Study

```
. tabulate smoke anyptl, chi2 row

           |       anyptl
     smoke |         0         1+ |     Total
-----------+----------------------+----------
         0 |       103         12 |       115
           |     89.57      10.43 |    100.00
-----------+----------------------+----------
         1 |        56         18 |        74
           |     75.68      24.32 |    100.00
-----------+----------------------+----------
     Total |       159         30 |       189
           |     84.13      15.87 |    100.00

          Pearson chi2(1) =   6.5050   Pr = 0.011
```

```
. tabulate smoke hyper, chi2 row

           |        hyper
     smoke |         0          1 |     Total
-----------+----------------------+----------
         0 |       108          7 |       115
           |     93.91       6.09 |    100.00
-----------+----------------------+----------
         1 |        69          5 |        74
           |     93.24       6.76 |    100.00
-----------+----------------------+----------
     Total |       177         12 |       189
           |     93.65       6.35 |    100.00

          Pearson chi2(1) =   0.0340   Pr = 0.854
```

Logistic Regression Model Building
Low Birthweight Case Study

Crude analysis:

```
. logit lbw smoke, or

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(1)      =       4.87
                                                  Prob > chi2     =     0.0274
Log likelihood =  -114.9023                       Pseudo R2       =     0.0207

------------------------------------------------------------------------------
         lbw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   2.021944   .6462912     2.20   0.028     1.080668    3.783083
------------------------------------------------------------------------------

. estimates store model1
```

Logistic Regression Model Building
Low Birthweight Case Study

Effect of smoking, adjusted for confounding:

```
. xi: logit lbw smoke i.agecat i.wcat i.race anyptl hyper, or

Logistic regression                               Number of obs   =        189
                                                  LR chi2(12)     =      42.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -96.147927                       Pseudo R2       =     0.1806

------------------------------------------------------------------------------
         lbw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   2.262309   .9348211     1.98   0.048      1.00652     5.08489
  _Iagecat_2 |   1.392663   .6378679     0.72   0.470     .5675163     3.41754
  _Iagecat_3 |   1.471076   .7585693     0.75   0.454     .5354361    4.041684
  _Iagecat_4 |   .4195708   .3026273    -1.20   0.229     .1020592    1.724878
    _Iwcat_2 |   .3322387   .1645385    -2.22   0.026     .1258635    .8770022
    _Iwcat_3 |   .4294435   .2488233    -1.46   0.145     .1379471    1.336902
    _Iwcat_4 |   .3403724   .2053118    -1.79   0.074     .1043546     1.11019
    _Iwcat_5 |   .1483812   .0991536    -2.86   0.004     .0400476     .549771
    _Irace_2 |   3.349107   1.834262     2.21   0.027      1.14482    9.797629
    _Irace_3 |   2.041873   .9271941     1.57   0.116     .8385072    4.972226
      anyptl |   3.908954   1.864624     2.86   0.004     1.534709    9.956236
       hyper |   5.126167   3.631541     2.31   0.021     1.278717    20.54997

. est store model3
```

Logistic Regression Model Building
Low Birthweight Case Study

Assess potential effect modification:

```
. xi: logit lbw i.smoke*hyper i.agecat i.wcat i.race anyptl, or

Logistic regression                               Number of obs   =        189
                                                  LR chi2(13)     =      43.47
                                                  Prob > chi2     =     0.0000
Log likelihood = -95.602101                       Pseudo R2       =     0.1852

------------------------------------------------------------------------------
         lbw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   _Ismoke_1 |   2.033485   .8714053     1.66   0.098     .8779652    4.709824
       hyper |   2.823694   2.529321     1.16   0.247     .4879269    16.34107
_IsmoXhype~1 |   4.358472   6.180988     1.04   0.299     .2705195    70.22147
  _Iagecat_2 |   1.460021   .6726449     0.82   0.411     .5918377    3.601765
  _Iagecat_3 |   1.614789   .8454232     0.92   0.360     .5787204    4.505703
  _Iagecat_4 |   .4043117   .2968872    -1.23   0.217     .0958687    1.705124
    _Iwcat_2 |    .325607   .1617741    -2.26   0.024     .1229657    .8621908
    _Iwcat_3 |   .3948697   .2325164    -1.58   0.115     .1245172    1.252213
    _Iwcat_4 |   .3531531   .2120446    -1.73   0.083     .1088614     1.14565
    _Iwcat_5 |   .1226656   .0865586    -2.97   0.003     .0307663    .4890693
    _Irace_2 |   3.678809   2.054487     2.33   0.020     1.231235    10.99192
    _Irace_3 |   2.063504   .9400548     1.59   0.112     .8449476    5.039422
      anyptl |   4.214317   2.045116     2.96   0.003     1.628012     10.9093

. est store model4
```

Logistic Regression Model Building
Low Birthweight Case Study

Assess potential effect modification:

```
. lrtest model3 model4

Likelihood-ratio test                                 LR chi2(1)  =      1.09
(Assumption: model3 nested in model4)                 Prob > chi2 =    0.2961

. lincom  _Ismoke_1+ _IsmoXhyper_1


         lbw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   2.181873   1.381393     1.58   0.114    -.5256074    4.889353
------------------------------------------------------------------------------

. lincom  _Ismoke_1+ _IsmoXhyper_1, or


         lbw | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   8.862888   12.24313     1.58   0.114     .5911962    132.8675
------------------------------------------------------------------------------
```

Conclusions?

Guidelines for Prediction

- Confounding not an issue (won't interpret coefficients)
- Automated methods (forward, backward selection; all possible regression) are appropriate
  - Hosmer & Lemeshow (1989) "*the analyst, not the computer, is ultimately responsible for the review and evaluation of the model*."
  - Draper & Smith (1981) "*The screening of variables should never be left to the soul discretion of any statistical procedure*".
- Estimate error rates
  - Internal (same dataset as used for fitting)
  - External (reserved dataset; split sample; cross-validation)
- Still important to have a plan!

Guidelines for Prediction

There are a number of different approaches to variable selection, once the form of the "full" model has been determined.

- **Purposeful selection**
- **Stepwise selection algorithms**
- **Best subsets algorithms**

All possible models may be compared based on a formal, explicit criterion. For instance:

- Akaike's information criterion (AIC) = -2logL +2p
  where p is the number of parameters estimated.

Low Birthweight Study: Prediction Accuracy

The variables identified in the table below have been shown to be associated with low birth weight in the obstetrical literature. The goal of the study was to ascertain if these variables were important in the population being served by the medical center where the data were collected.

Table: Code Sheet for the Variables in the Low Birth Weight Data Set.

Columns	Variable	Abbreviation
2-4	Identification Code	ID
10	Low Birth Weight (0 = Birth Weight ge 2500g, l = Birth Weight < 2500g)	LBW
17-18	Age of the Mother in Years	AGE
23-25	Weight in Pounds at the Last Menstrual Period	LWT
32	Race (1 = White, 2 = Black, 3 = Other)	RACE
40	Smoking Status During Pregnancy (1 = Yes, 0 = No)	SMOKE
48	History of Premature Labor (0 = None, 1 = One, etc.)	PTL
55	History of Hypertension (1 = Yes, 0 = No)	HYPER
61	Presence of Uterine Irritability (1 = Yes, 0 = No)	URIRR
67	Number of Physician Visits During the First Trimester (0 = None, 1 = One, 2 = Two, etc.)	PVFT
73-76	Birth Weight in Grams	BWT

Stepwise Algorithms

Variables are included or excluded from the model, based on their statistical significance. Suppose we have k variables for consideration in the model. A forward selection algorithm would proceed as follows:

Step 0: Fit a model with the intercept only and evaluate the likelihood, L_0. Fit each of the k possible univariate models, evaluate their likelihood, L_{j0}, j=1,2,..,k and carry out the LRT comparing L_0 and L_{j0}. The variable (say the 1st) with smallest p-value is included in the model, provided this p-value is less than some pre-specified p_E.

Step 1: All k-1 models containing the intercept, the 1st variable and one of the remaining variables are fitted. The log-likelihoods are compared with those from the model containing just the intercept and the 1st variable. Say the 2nd variable has the smallest p-value, p_2. It is then included in the model, provided $p_2 < p_E$.

Step 2: Carry out a LRT to assess whether the 1st variable, given the presence of the 2nd variable, can be dropped from the model. Compare the p-value from the LRT with a pre-specified p-value, p_R. etc

Step S: All k variables have been included in the model or all variables in the model have p-values less than p_R and all variables not in the model have p-values greater than p_E.

The same principles can be applied, working backward from the "full" model with all k variables.

Low Birthweight Study
Example of Variable Selection – Forward Stepwise

```
. gen newpvft= pvft
. recode newpvft 2/max=2
. xi:sw,forw pe(.2) pr(.25) lr: logit lbw lwt smoke anyptl i.race newpvft
   hyper urirr age

LR test               begin with empty model

p = 0.0004 <  0.2000  adding   anyptl
p = 0.0320 <  0.2000  adding   age
p = 0.0410 <  0.2000  adding   hyper
p = 0.0171 <  0.2000  adding   lwt
p = 0.1048 <  0.2000  adding   _Irace_2
p = 0.1071 <  0.2000  adding   urirr
p = 0.1492 <  0.2000  adding   smoke
p = 0.0696 <  0.2000  adding   _Irace_3
p = 0.3130 >= 0.2500  removing age

Logistic regression                               Number of obs   =        189
                                                  LR chi2(7)      =      36.82
                                                  Prob > chi2     =     0.0000
Log likelihood = -98.925785                       Pseudo R2       =     0.1569

------------------------------------------------------------------------------
         lbw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      anyptl |   1.128857   .4503896     2.51   0.012     .2461093    2.011604
    _Irace_3 |   .8544142   .4409119     1.94   0.053    -.0097573    1.718586
       hyper |   1.866895   .7073782     2.64   0.008     .4804594    3.253331
         lwt |  -.0159185   .0069539    -2.29   0.022    -.0295478   -.0022891
    _Irace_2 |   1.300856    .528489     2.46   0.014     .2650362    2.336675
       urirr |   .7506488   .4588171     1.64   0.102    -.1486163    1.649914
       smoke |   .8665818   .4044737     2.14   0.032      .073828    1.659336
       _cons |   -.125326   .9675725    -0.13   0.897    -2.021733    1.771081
------------------------------------------------------------------------------
```

Low Birthweight Study
Example of Variable Selection – Backward Stepwise

```
. xi:sw, pe(.2) pr(.25) lr: logit lbw lwt smoke anyptl i.race
  newpvft hyper urirr age

LR test              begin with full model

p = 0.8616 >= 0.2500  removing newpvft
p = 0.3130 >= 0.2500  removing age

Logistic regression                                     Number of obs   =        189
                                                        LR chi2(7)      =      36.82
                                                        Prob > chi2     =     0.0000
Log likelihood = -98.925785                             Pseudo R2       =     0.1569

------------------------------------------------------------------------------
         lbw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0159185   .0069539    -2.29   0.022    -.0295478   -.0022891
       smoke |   .8665818   .4044737     2.14   0.032      .073828    1.659336
      anyptl |   1.128857   .4503896     2.51   0.012     .2461093    2.011604
    _Irace_2 |   1.300856    .528489     2.46   0.014     .2650362    2.336675
    _Irace_3 |   .8544142   .4409119     1.94   0.053    -.0097573    1.718586
       urirr |   .7506488   .4588171     1.64   0.102    -.1486163    1.649914
       hyper |   1.866895   .7073782     2.64   0.008     .4804594    3.253331
       _cons |   -.125326   .9675725    -0.13   0.897    -2.021733    1.771081
------------------------------------------------------------------------------
```

Logistic Regression & Prediction of Binary Outcomes

1. We can use logistic regression to obtain estimates of probabilities, $\pi(X)$
2. Assessment of Accuracy
   - Comparison of fitted and observed counts within subgroups to assess goodness-of-fit of the model
   - Prediction of individual outcomes for future subjects (screening):
     - Sensitivity
     - Specificity
     - ROC curve

Using Modeling Results for Prediction

The predicted probabilities, $\hat{P}[Y = 1 \,|X] = \hat{\pi}(X)$, may be used to predict the outcome of a subject with covariates X using a decision rule such as

Predict Y=1 whenever $\hat{\pi}(X) > \frac{1}{2}$

or, more generally,

Predict Y=1 whenever $X\hat{\beta} > c$

where c is a constant and $X\hat{\beta}$ is the linear predictor (LP).

Define for any LP:

Sensitivity: $P\left[\mathbf{LP} > c \,|\, Y = 1\right]$

Specificity: $P\left[\mathbf{LP} \leq c \,|\, Y = 0\right]$

The ROC Curve

For each value of the "cutoff" or criterion, *c*, we have an associated sensitivity and specificity. Which threshold, *c*, to choose can depend on such factors as the "costs" assigned to two different types of error: falsely predicting Y=0 when, in fact, Y=1, and *vice versa* and the "benefits" of correct assignments.

Define:

Sensitivity(*c*): $P\left[\text{Test} > c \mid Y = 1\right]$

Specificity(*c*): $P\left[\text{Test} \leq c \mid Y = 0\right]$

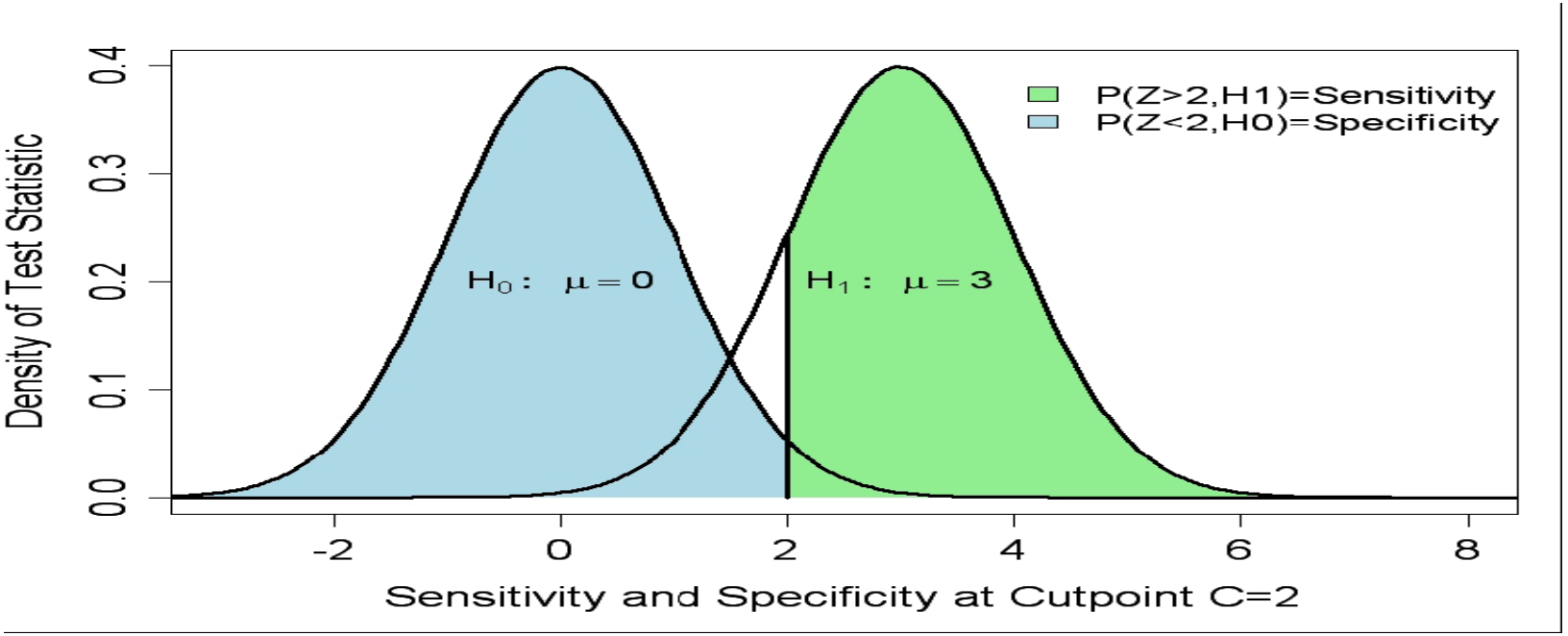
Then the "ROC" (Receiver Operating Characteristic) curve plots the values

[1-Specificity(*c*) , Sensitivity(*c*)]
(False Positive using *c* , True Positive using *c*)

for all possible values of *c*.

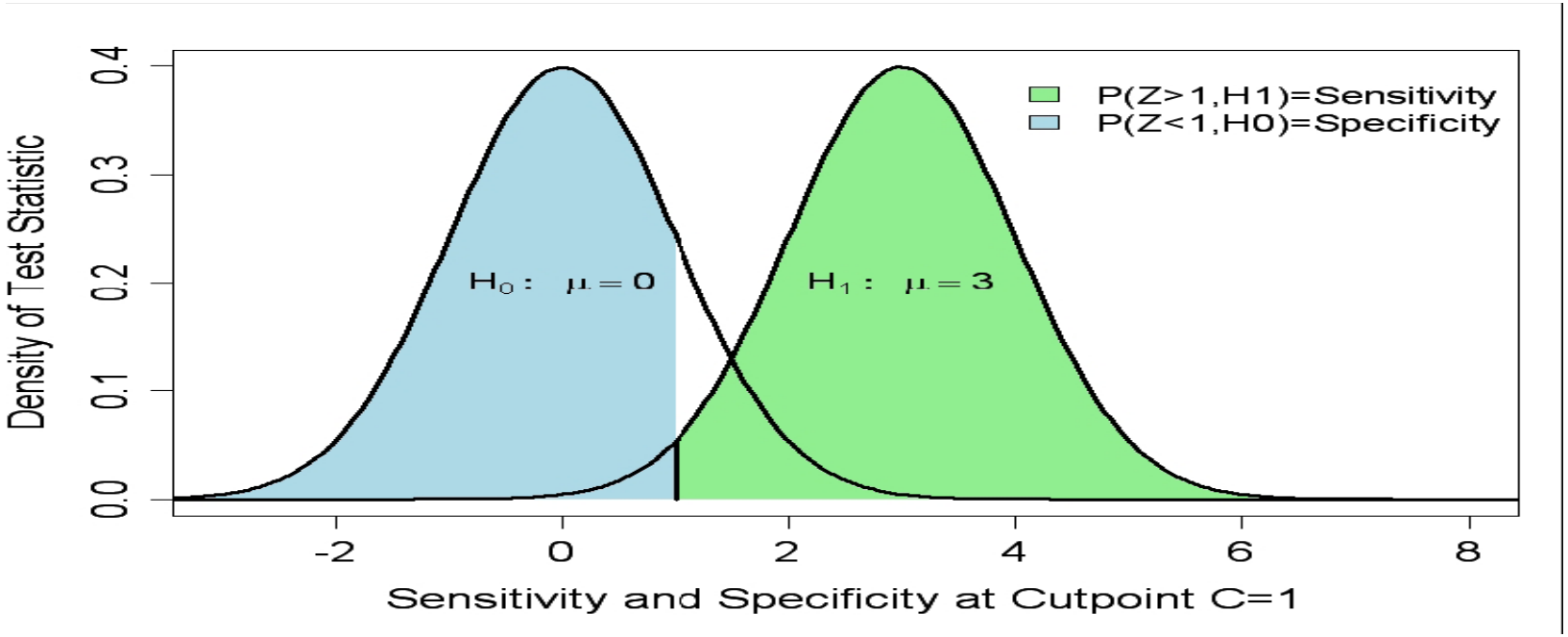
Sensitivity and Specificity

E.g. assume normal distributions (equal variance) for LP, $X\hat{\beta}$



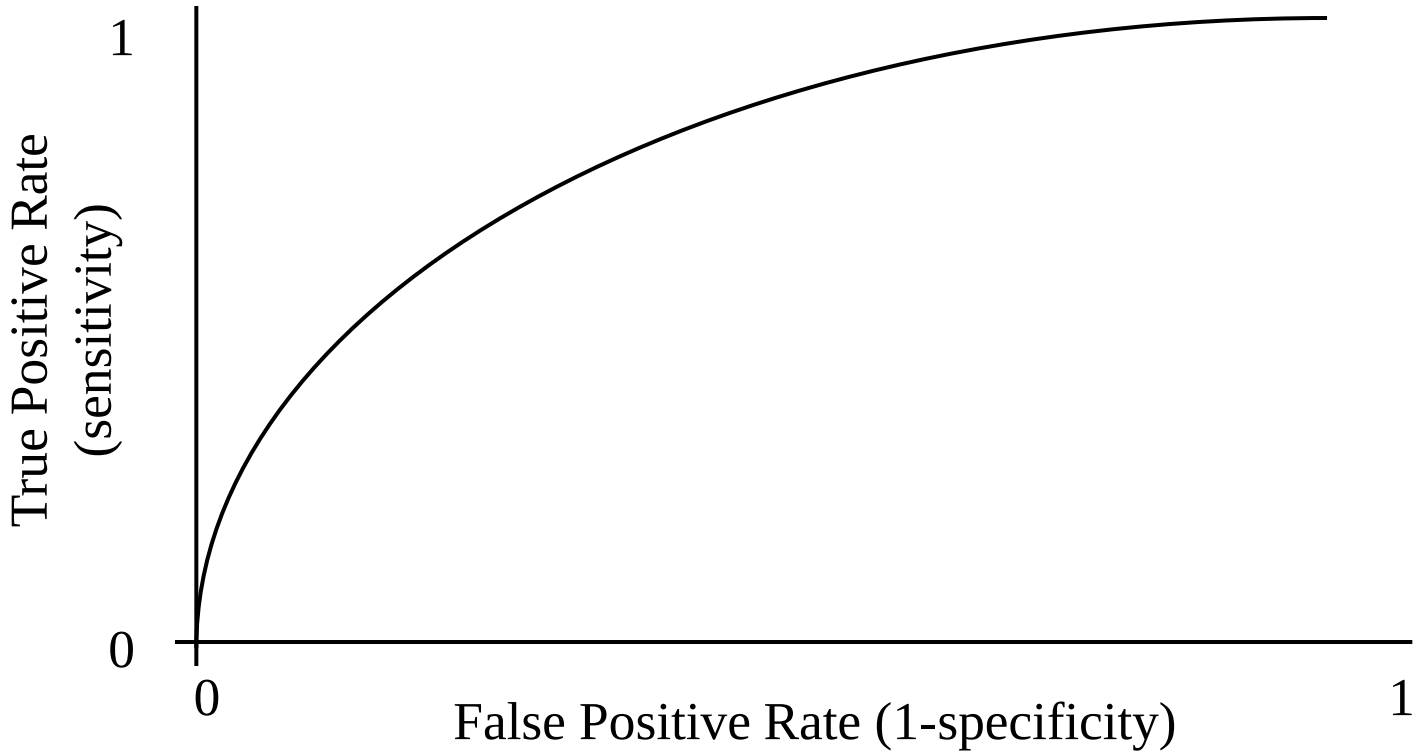
Sensitivity and Specificity

Different choices of *c* give different error rates



ROC (receiver operating characteristic) Curve

By varying *c* over its entire range ($-\infty$, $+\infty$) and plotting Sensitivity vs 1-Specificity we obtain



ROC Curve

An overall summary of an ROC curve is the *area* under the curve. This is attractive since:

1. A perfect test would have area equal to 1.0.

2. A (basically) worthless test would have an area of 0.5.

3. Interpretation:

   The area under the ROC curve corresponds to the probability that a randomly chosen case ($Y = 1$) would have a higher "test" value (higher LP value) than a randomly chosen control ($Y = 0$).

Accurate Estimates of Predition Error

- If we use the same data to *fit* a model and *assess* a model we generally obtain biased, overly optimistic prediction summaries (unless we explicitly correct the summaries).

- Use *training* data to build the model and *test* data to evaluate

- Akaike's Information Criterion (AIC) approximates the MSE of prediction and is useful for comparing models fitted to the same data. In particular, AIC tries to pick the model that would have the lowest prediction error when applied to *new* data.

Low Birthweight Study: Prediction Accuracy

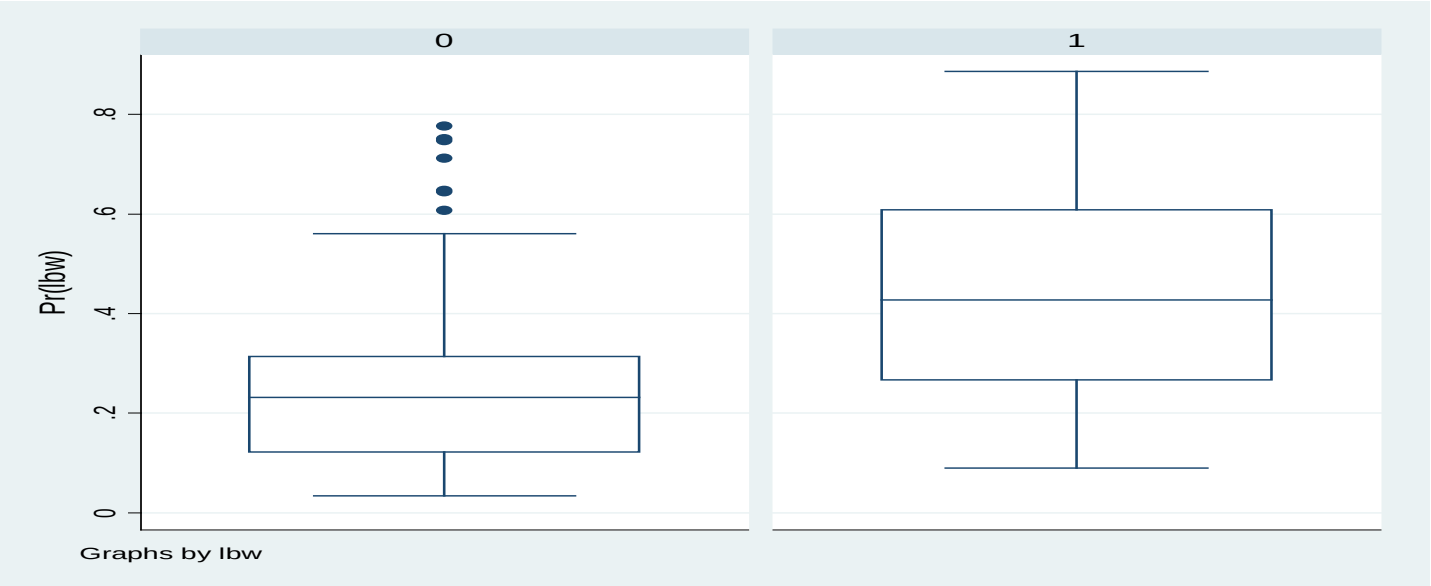
```
. xi: logit lbw lwt smoke anyptl i.race hyper urirr
Logistic regression                               Number of obs   =        189
                                                  LR chi2(7)      =      36.82
                                                  Prob > chi2     =     0.0000
Log likelihood = -98.925785                       Pseudo R2       =     0.1569

------------------------------------------------------------------------------
         lbw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0159185   .0069539    -2.29   0.022    -.0295478   -.0022891
       smoke |   .8665818   .4044737     2.14   0.032      .073828    1.659336
      anyptl |   1.128857   .4503896     2.51   0.012     .2461093    2.011604
    _Irace_2 |   1.300856    .528489     2.46   0.014     .2650362    2.336675
    _Irace_3 |   .8544142   .4409119     1.94   0.053    -.0097573    1.718586
       hyper |   1.866895   .7073782     2.64   0.008     .4804594    3.253331
       urirr |   .7506488   .4588171     1.64   0.102    -.1486163    1.649914
       _cons |   -.125326   .9675725    -0.13   0.897    -2.021733    1.771081
------------------------------------------------------------------------------

. predict fitp
(option p assumed; Pr(lbw))

. summarize fitp
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fitp |       189    .3121693    .2025102   .0341958   .8862907

. summarize lbw
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         lbw |       189    .3121693    .4646093          0          1
```

Low Birthweight Study: Prediction Accuracy

```
. list fitp lbw smoke anyptl lwt race hyper urirr if age==15, table
compress noobs
```

fitp	lbw	smoke	any~l	lwt	race	hyper	urirr
.4050466	0	0	0	98	2	0	0
.1328077	1	0	0	110	1	0	0
.4131672	1	0	0	115	3	0	1

```
. sort lbw
. graph box fitp, medtype(line) by(lbw) intensity(0)
```



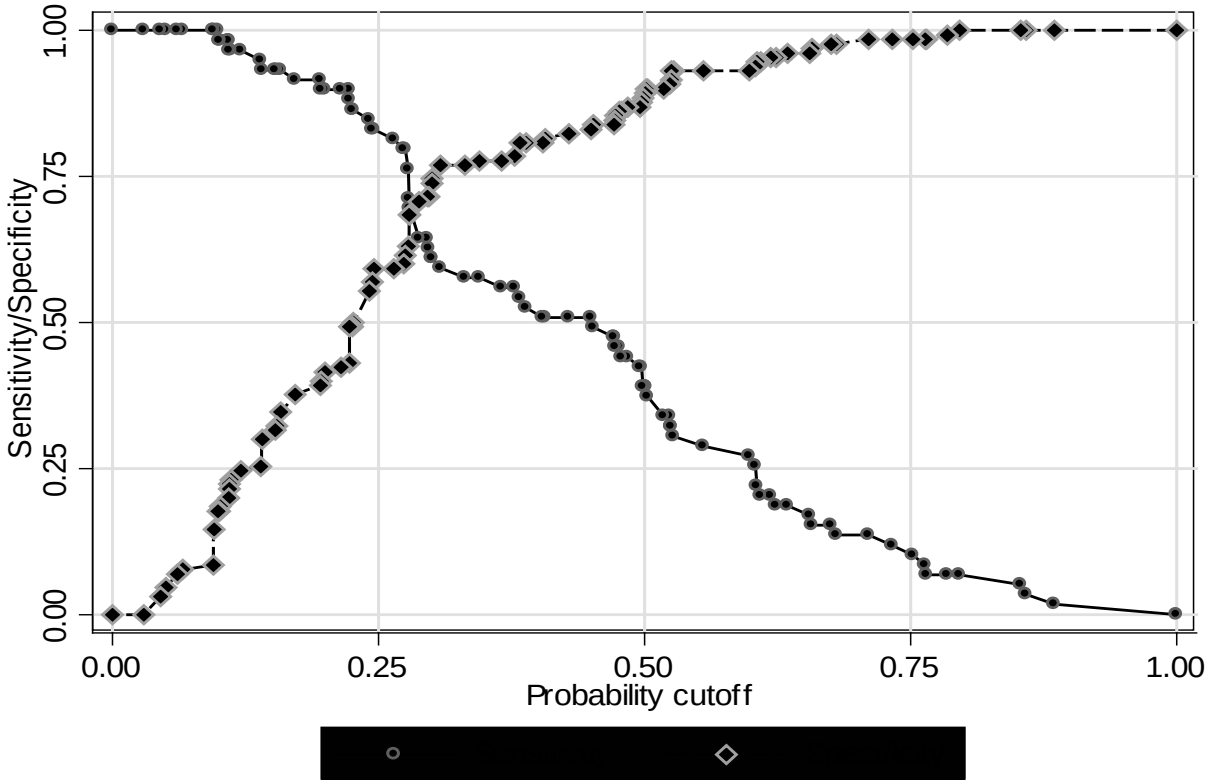
Low Birthweight Study: Prediction Accuracy

Results of prediction when cutoff=0.5:

```
. estat classification

Logistic model for lbw

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        26            13  |         39
     -     |        33           117  |        150
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as lbw != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   44.07%
Specificity                     Pr( -|~D)   90.00%
Positive predictive value       Pr( D| +)   66.67%
Negative predictive value       Pr(~D| -)   78.00%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   10.00%
False - rate for true D         Pr( -| D)   55.93%
False + rate for classified +   Pr(~D| +)   33.33%
False - rate for classified -   Pr( D| -)   22.00%
--------------------------------------------------
Correctly classified                        75.66%
--------------------------------------------------
```

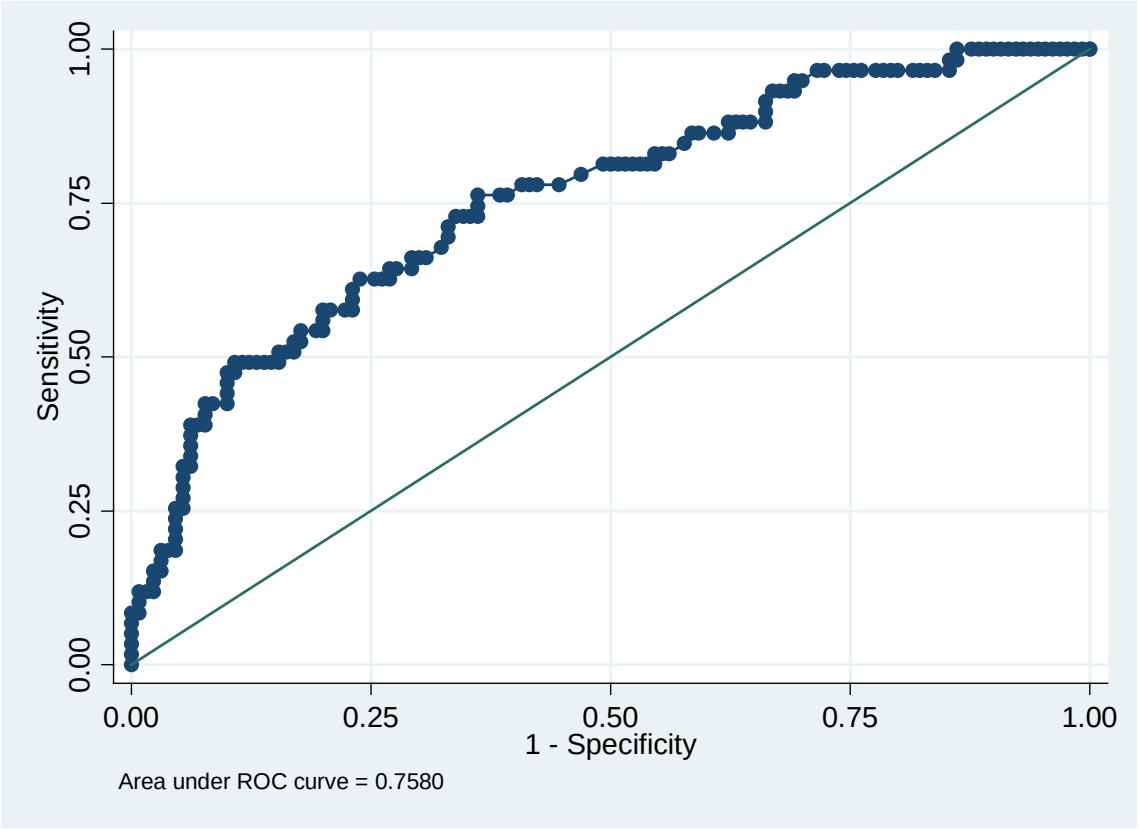
Low Birthweight Study: Prediction Accuracy

```
. lsens
```



Low Birthweight Study: Prediction Accuracy

```
. lroc
Logistic model for lbw

number of observations =      189
area under ROC curve   =   0.7580
```



Low Birthweight Study: Prediction Accuracy

Using a training sample:

```
. set seed 4514
. gen uni=uniform()
. sort uni
. gen sample=0
. replace sample=1 if _n>100
. tab sample
```

sample	Freq.	Percent	Cum.
0	100	52.91	52.91
1	89	47.09	100.00
Total	189	100.00	

```
. xi: logit lbw lwt smoke anyptl i.race hyper urirr if sample==0
Logistic regression                               Number of obs   =        100
                                                  LR chi2(7)      =      22.31
                                                  Prob > chi2     =     0.0022
Log likelihood =  -51.53012                       Pseudo R2       =     0.1780

------------------------------------------------------------------------------
         lbw |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0009498    .008417    -0.11   0.910    -.0174468    .0155472
       smoke |   1.033608   .5691081     1.82   0.069    -.0818234    2.149039
      anyptl |   .5845931   .6443695     0.91   0.364    -.6783479    1.847534
    _Irace_2 |     1.6412   .7254575     2.26   0.024     .2193291     3.06307
    _Irace_3 |   1.176396   .6257745     1.88   0.060    -.0500994    2.402891
       hyper |   1.744725   1.031183     1.69   0.091    -.2763558    3.765806
       urirr |   1.800833   .6206835     2.90   0.004     .5843157     3.01735
       _cons |  -2.346305   1.234023    -1.90   0.057    -4.764946    .0723361
------------------------------------------------------------------------------
```

Low Birthweight Study: Prediction Accuracy

```
. estat classification

Logistic model for lbw

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        16             9  |         25
     -     |        16            59  |         75
-----------+--------------------------+-----------
   Total   |        32            68  |        100

Classified + if predicted Pr(D) >= .5
True D defined as lbw != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   50.00%
Specificity                     Pr( -|~D)   86.76%
Positive predictive value       Pr( D| +)   64.00%
Negative predictive value       Pr(~D| -)   78.67%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   13.24%
False - rate for true D         Pr( -| D)   50.00%
False + rate for classified +   Pr(~D| +)   36.00%
False - rate for classified -   Pr( D| -)   21.33%
--------------------------------------------------
Correctly classified                        75.00%
--------------------------------------------------
```

```
. lroc

Logistic model for lbw

number of observations =      100
area under ROC curve   =   0.7576
```

Low Birthweight Study: Prediction Accuracy

```
. estat classification if sample==1

Logistic model for lbw

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        10            11  |         21
     -     |        17            51  |         68
-----------+--------------------------+-----------
   Total   |        27            62  |         89

Classified + if predicted Pr(D) >= .5
True D defined as lbw != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   37.04%
Specificity                     Pr( -|~D)   82.26%
Positive predictive value       Pr( D| +)   47.62%
Negative predictive value       Pr(~D| -)   75.00%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   17.74%
False - rate for true D         Pr( -| D)   62.96%
False + rate for classified +   Pr(~D| +)   52.38%
False - rate for classified -   Pr( D| -)   25.00%
--------------------------------------------------
Correctly classified                        68.54%
--------------------------------------------------
```

```
. lroc if sample==1

Logistic model for lbw

number of observations =       89
area under ROC curve   =   0.7004
```

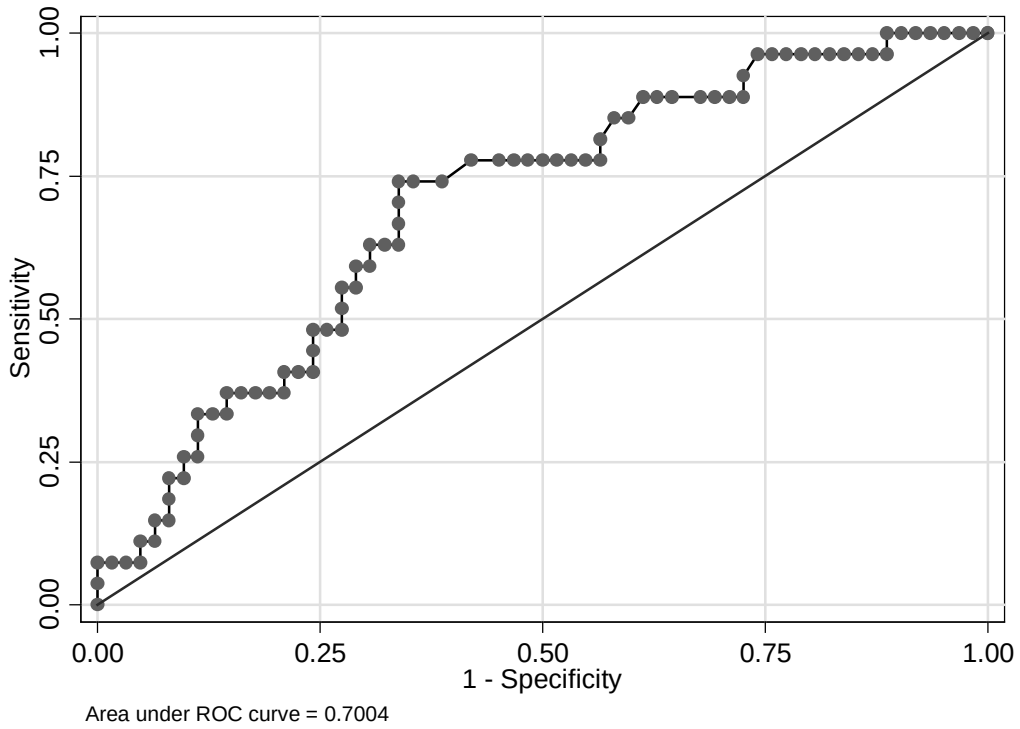
Low Birthweight Study: Prediction Accuracy

Assessing prediction via the validation sample with a different threshold:

```
. estat classification if sample==1, cutoff(.2)

Logistic model for lbw

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        21            35  |         56
     -     |         6            27  |         33
-----------+--------------------------+-----------
   Total   |        27            62  |         89

Classified + if predicted Pr(D) >= .2
True D defined as lbw != 0
-------------------------------------------------
Sensitivity                     Pr( +| D)   77.78%
Specificity                     Pr( -|~D)   43.55%
Positive predictive value       Pr( D| +)   37.50%
Negative predictive value       Pr(~D| -)   81.82%
-------------------------------------------------
False + rate for true ~D        Pr( +|~D)   56.45%
False - rate for true D         Pr( -| D)   22.22%
False + rate for classified +   Pr(~D| +)   62.50%
False - rate for classified -   Pr( D| -)   18.18%
-------------------------------------------------
Correctly classified                        53.93%
-------------------------------------------------
```

Low Birthweight Study: Prediction Accuracy

```
. lroc if sample==1, scheme(s1mono)

Logistic model for lbw

number of observations =       89
area under ROC curve   =   0.7004
```



Note: Highly statistically significant risk factors (predictors) do not guarantee successful prediction!

Guidelines for Statistical Analysis

Strategy: Cox and Wermuth (1996) section 7.2

1. Establish the main scientific research question.
2. Check Data Quality
   - Look for:
     - Possible errors in coding.
     - Outliers.
     - Missing values.
   - Produce univariate summaries.
3. Classification of variables based on substantive grounds: Outcome, known important variables, others
4. Document pairwise associations:
   - Correlations
   - mean differences with SE's
   - log odds ratios with SE's
   - Additional stratified analyses for key variables.

Guidelines for Statistical Analysis

Strategy: Cox and Wermuth (1996) section 7.2

5. Develop regression models.
6. Presentation of model(s): Coefficients and SE's, Graphical display
7. There is no best model

George Box: *All models are wrong, but some are useful!*

Statistical Thinking

"It will rarely be necessary to include a large number of variables in the analysis, because only a few exposures are of genuine scientific interest in any one study, and there are usually very few variables of sufficient a priori importance for their potential confounding effect to be controlled for. Most scientists are aware of the dangers of analyses which search a long list of potentially relevant exposures. These are known as data dredging or blind fishing and carry considerable danger of false positive findings. Such analyses are as likely to impede scientific progress as to advance it. There are similar dangers if a long list of potential confounders is searched, either with a view to explaining the observed relationship between disease and exposure or to enhancing it – findings will inevitably be biased. Confounders should be chosen a priori and not on the basis of statistical significance."

(Clayton and Hills, Statistical Methods in Epidemiology, 1993, p. 273)

Statistical Thinking

"*As a medical statistician, I am appalled by the large number of irreproducible results published in the medical literature. There is a general, and likely correct, perception that this problem is associated more with statistical, as opposed to laboratory, research. I am convinced, however, that results of clinical and epidemiological investigations could become more reproducible if only the investigators would apply more rigorous statistical thinking and adhere more closely to well established principles of the scientific method. While I agree that the investigative cycle is an iterative process, I believe that it works best when it is hypothesis driven.*"

"*The epidemiology literature is replete with irreproducible results stemming from the failure to clearly distinguish between analyses that were specified in the protocol and that test the a priori hypotheses whose specification was needed to secure funding, and those that were performed post-hoc as part of a serendipitous process of data exploration.*"

Breslow (1999)

Other Issues in Model Building

- Multiple Comparisons
  - If we conduct enough significance tests then we are bound to find a significant association
  - Having a systematic plan at least allows a reviewer to understand the risk
  - Hilsenbeck, Clark and McGuire (1992). "Why do so many prognostic factors fail to pan out?" *Breast Cancer Research and Treatment* **22**: 197-206
- Multicollinearity
  - One or more covariates are highly correlated
  - Yields unstable coefficients
- Influential observations
  - Check delta-beta statistics

Additional Topics

- Conditional logistic regression
  - many strata
  - matched data
- Ordinal and polytomous logistic regression
- Analysis with correlated data
  - GEE
  - hierarchical models (random effects)