Logistic Regression

Logistic Regression: The Multivariate Problem

Q: What is the relationship between one or more exposure variables, $\boldsymbol{E}_j$, and a binary disease or illness outcome, $\boldsymbol{D}$?

Example:

- $\boldsymbol{D}$ is coronary heart disease (CHD). $\boldsymbol{D} = 1$ is "with CHD" and $\boldsymbol{D} = 0$ is "without CHD".
- Our exposure of interest is smoking, $\boldsymbol{E} = 1$ for smokers (current, ever), and $\boldsymbol{E} = 0$ 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`, `gender`.

$$\underbrace{\boldsymbol{E},\ \boldsymbol{C}_1,\ \boldsymbol{C}_2,\ \boldsymbol{C}_3}_{\text{independent}} \quad \Rightarrow \quad \underbrace{\boldsymbol{D}}_{\text{dependent}}$$

Logistic Regression: The Multivariate Problem

Independent variables: $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_p$

Dependent variable: $\boldsymbol{Y}$ binary

- We have a flexible choice for the 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

$$\begin{aligned} E(\boldsymbol{Y} \mid \boldsymbol{X}) &= \mu(\boldsymbol{X}) \\ &= \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots \beta_p \boldsymbol{X}_p \end{aligned}$$

$$\begin{aligned} \boldsymbol{Y} &= \mu(\boldsymbol{X}) + \epsilon \\ \epsilon &\sim N(0, \sigma^2) \end{aligned}$$

Binary Response Regression

Note: For a binary response variable we have

$$E(\boldsymbol{Y}) = \pi \in [0,1]$$

- The mean for a binary variable is a proportion or a probability.

- If we model the mean for $\boldsymbol{Y}$ we'll need to impose the constraint.

Q: Can we just use

$$\begin{aligned} E(\boldsymbol{Y} \mid \boldsymbol{X}) &= \pi(\boldsymbol{X}) \\ \pi(\boldsymbol{X}) &= \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p \end{aligned}$$

- for binary $\boldsymbol{X}$?

- for multi-categorical $\boldsymbol{X}$?

- for continuous $\boldsymbol{X}$?

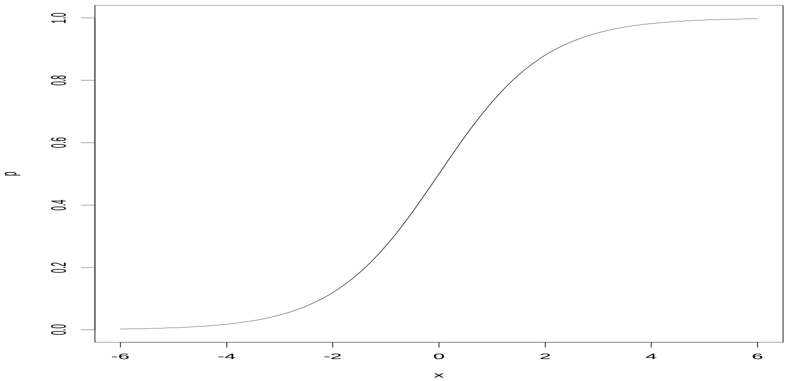
The Logistic Function

Q: Why is logistic regression so popular?

1. Dichotomous outcomes are common.
2. Logistic regression keeps means between 0 and 1.
3. Regression parameters are log odds ratios.

Q: How is the mean, $\pi(\boldsymbol{X})$, kept between 0 and 1?

The logistic function:



The Logistic Function

The **logistic function** is given by:

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

Properties:

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

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

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

The Logistic Model

$$
\begin{aligned}
\boldsymbol{Z} &= \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p \\
\boldsymbol{Z} &= \boldsymbol{X}\boldsymbol{\beta}
\end{aligned}
$$

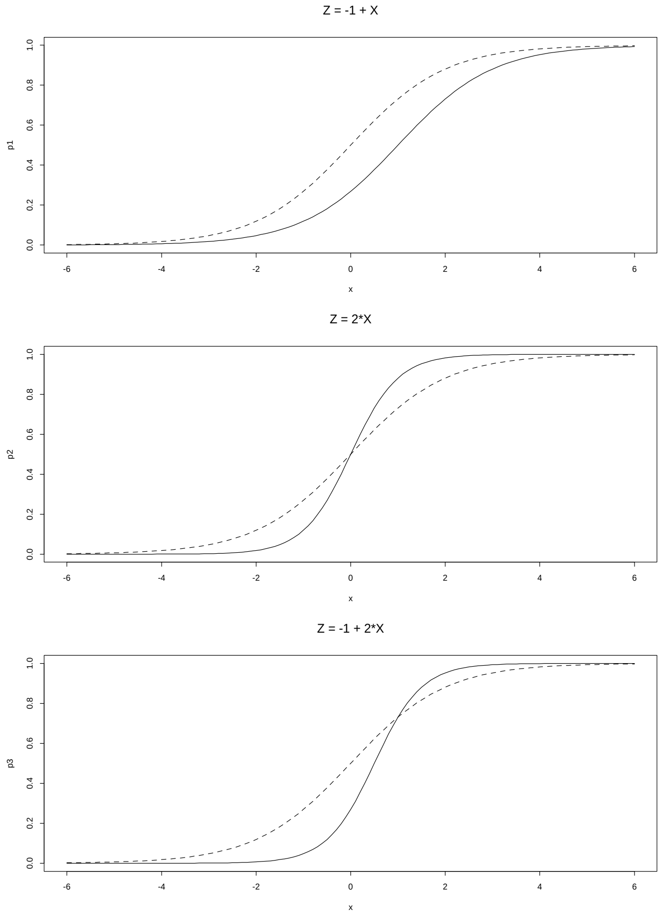
$$
\pi(\boldsymbol{Z}) = \frac{\exp(\boldsymbol{Z})}{1 + \exp(\boldsymbol{Z})}
$$

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

Note:

1. This is a **non-linear** model for the mean $\pi(\boldsymbol{X})$.
2. The combination of variables is still linear (ie. $\boldsymbol{Z}$ is linear in $\boldsymbol{X}$).
3. The quantity $\boldsymbol{Z}$ is called the **linear predictor**.

Some Examples



Applying the Logistic Model Formula

Example: from Kleinbaum (1994)

$$
\begin{aligned}
\boldsymbol{D} &= \text{CHD } (0,1) \\
\boldsymbol{X}_1 &= \text{CAT catecholamine level } 1 = \text{high}, \;\; 0 = \text{low} \\
\boldsymbol{X}_2 &= \text{AGE, in years} \\
\boldsymbol{X}_3 &= \text{ECG } 1 = \text{abnormal}, \;\; 0 = \text{normal} \\
n &= 609 \text{ males with 9-year follow-up}
\end{aligned}
$$

<u>Model:</u>

$$
\begin{aligned}
P(\boldsymbol{D} = 1 \mid \boldsymbol{X}) &= \pi(\boldsymbol{X}) \\
\pi(\boldsymbol{X}) &= \frac{\exp(\beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \beta_3 \boldsymbol{X}_3)}{1 + \exp(\beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \beta_3 \boldsymbol{X}_3)}
\end{aligned}
$$

<u>Estimates:</u> via maximum likelihood

$$
\begin{aligned}
\widehat{\beta}_0 &= -3.911 \\
\widehat{\beta}_1 &= 0.652 \\
\widehat{\beta}_2 &= 0.029 \\
\widehat{\beta}_3 &= 0.342
\end{aligned}
$$

Applying the Logistic Model Formula

Q: What is the probability for an individual with (CAT=$\boxed{1}$, AGE=40, ECG=0)?

$$
\begin{aligned}
\widehat{\pi}(\boldsymbol{X}^{(1)}) &= \frac{\exp(\,-3.911 + 0.652(1) + 0.029(40) + 0.342(0)\,)}{1 + \exp(\,-3.911 + 0.652(1) + 0.029(40) + 0.342(0)\,)} \\
&= \frac{\exp(-2.101)}{1 + \exp(-2.101)} \\
&= 0.109
\end{aligned}
$$

Q: What is the probability for an individual with (CAT=$\boxed{0}$, AGE=40, ECG=0)?

$$
\begin{aligned}
\widehat{\pi}(\boldsymbol{X}^{(0)}) &= \frac{\exp(\,-3.911 + 0.652(0) + 0.029(40) + 0.342(0)\,)}{1 + \exp(\,-3.911 + 0.652(0) + 0.029(40) + 0.342(0)\,)} \\
&= \frac{\exp(-2.751)}{1 + \exp(-2.751)} \\
&= 0.060
\end{aligned}
$$

Applying the Logistic Model Formula

Compare:

$$\widehat{RR} = \frac{\widehat{\pi}(\boldsymbol{X}^{(1)})}{\widehat{\pi}(\boldsymbol{X}^{(0)})}$$

$$= 0.109/0.060 = 1.817$$

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

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

Interpret:

Note: $\log(1.917) = 0.652$

Logit Transformation

Note:

$$\text{Odds}(\boldsymbol{X}) = \frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}$$

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

$$= \exp(\boldsymbol{X\beta})$$

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

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

$$= \beta_0 + \beta_1\boldsymbol{X}_1 + \ldots + \beta_p\boldsymbol{X}_p$$

Therefore, the logistic model is a linear model for the log odds.

Logit Contrasts

$$\begin{aligned}\text{logit}[\pi(\boldsymbol{X})] &= \boldsymbol{X\beta} \\ &= \text{ log odds for individual with } \boldsymbol{X} \\ \\ &= \beta_0 \;+\; \beta_1 \boldsymbol{X}_1 \;+ \ldots +\; \beta_p \boldsymbol{X}_p\end{aligned}$$

Q: What are the interpretations of the β_j's?

$\boxed{\beta_0}$

If all the $\boldsymbol{X}_j = 0$ then $\text{logit}[\pi(\boldsymbol{X})] = \beta_0$.

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.

Logit Contrasts

Back to the CHD example. We can represent the logistic model for CHD as:

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

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

$$\begin{aligned}\text{logit}[\pi(\boldsymbol{X}^{(1)})] &= \beta_0 + \beta_1(1) + \beta_2(40) \\ \\ \text{logit}[\pi(\boldsymbol{X}^{(1)})] &= \beta_0 + \beta_2(40) \\ \\ \text{logit}[\pi(\boldsymbol{X}^{(1)})] - \text{logit}[\pi(\boldsymbol{X}^{(0)})] &= \beta_1\end{aligned}$$

$\boxed{\beta_1}$

The difference in the log odds comparing an individual with CAT=1 to CAT=0 when AGE and ECG are fixed.

Q: Can we say: “The change in the log odds when CAT is changed from 0 to 1 and AGE and ECG are fixed.”

Odds Ratios

Any **odds ratio**, by definition, is a ratio of two odds, written here as **odds($\boldsymbol{X}^{(1)}$)** divided by **odds($\boldsymbol{X}^{(0)}$)** in which $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(0)}$ indicate two individuals, or two groups of individuals, that are being compared.
Groups/individuals are determined by their covariate values $(X_1, X_2, \ldots, X_p)$.

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

If we consider a logistic model and two covariate values, $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(0)}$, then we have:

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

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

$$\begin{aligned}
\text{OR}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)}) &= \frac{\text{odds}(\boldsymbol{X}^{(1)})}{\text{odds}(\boldsymbol{X}^{(0)})} \\
&= \frac{\exp(\boldsymbol{X}^{(1)}\boldsymbol{\beta})}{\exp(\boldsymbol{X}^{(0)}\boldsymbol{\beta})} \\
&= \exp\left(\sum_{j=1}^{p} \beta_j (\boldsymbol{X}_j^{(1)} - \boldsymbol{X}_j^{(0)}) \right)
\end{aligned}$$

Odds Ratios

In our CHD example we have:

$$\begin{aligned}
\boldsymbol{X}^{(1)} &= (1, 40, 0) \\
\boldsymbol{X}^{(0)} &= (0, 40, 0)
\end{aligned}$$

And we obtain the odds ratio comparison as:

$$\begin{aligned}
\text{OR}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)}) &= \exp\left(\sum_{j=1}^{3} \beta_j (\boldsymbol{X}_j^{(1)} - \boldsymbol{X}_j^{(0)}) \right) \\
&= \exp(\beta_1) \\
\widehat{\text{OR}}(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)}) &= \exp(0.652) \\
&= 1.919
\end{aligned}$$

Interpret:

Summary

1. The multivariate problem.
2. Define: the logistic function.
3. Properties of the logistic function.
4. Applying the logistic model to obtain probabilities.
5. Obtaining RR's and OR's.
6. The logit transformation.
7. Parameters in the logistic model.
8. $OR(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(0)})$.

Logistic Regression: Some simple special cases

Q: What does logistic regression provide for a simple binary exposure variable, $\boldsymbol{E}$, and a binary disease or illness outcome, $\boldsymbol{D}$?

Example: Pauling (1971)

	$\boldsymbol{D}$ (cold=Y)	$\overline{\boldsymbol{D}}$ (cold=N)	
$\boldsymbol{E}$ (vitamin C)	17	122	139
$\overline{\boldsymbol{E}}$ (placebo)	31	109	140
	48	213	279

Binary Exposure

- Let $\boldsymbol{X}_1$ be a dummy variable for exposure groups:
  $\boldsymbol{X}_1 = 1$ if in group $\boldsymbol{E}$, exposed, vitamin C
  $\boldsymbol{X}_1 = 0$ if in group $\overline{\boldsymbol{E}}$, unexposed, placebo

- Let $\boldsymbol{Y}$ be an indicator (binary) outcome variable:
  $\boldsymbol{Y} = 1$ if in group $\boldsymbol{D}$, disease, gets a cold
  $\boldsymbol{Y} = 0$ if in group $\overline{\boldsymbol{D}}$, no disease, no cold

	$\boldsymbol{Y} = 1$	$\boldsymbol{Y} = 0$	
$\boldsymbol{X}_1 = 1$	17	122	139
$\boldsymbol{X}_1 = 0$	31	109	140
	48	213	279

From these data we can estimate:

$$\begin{aligned}
\widehat{P}[\boldsymbol{Y} = 1 \mid \boldsymbol{X}_1 = 1] &= 17/139 = 0.122 \\
\widehat{P}[\boldsymbol{Y} = 1 \mid \boldsymbol{X}_1 = 0] &= 31/140 = 0.221 \\
\widehat{RR} &= 0.122/0.221 = 0.554 \\
\widehat{OR} &= \frac{0.122/(1-0.122)}{0.221/(1-0.221)} = 0.490
\end{aligned}$$

Binary Exposure

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

Data:

```
Y   X   count

1   1   17
0   1   122
1   0   31
0   0   109
```

Model:

$$\begin{aligned}
\pi(\boldsymbol{X}) &= \text{probability that } \boldsymbol{Y} = 1 \text{ given } \boldsymbol{X}. \\
\log\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right) &= \beta_0 + \beta_1 \boldsymbol{X}_1 \\
\text{logit}[\pi(\boldsymbol{X})] &= \beta_0 + \beta_1 \boldsymbol{X}_1
\end{aligned}$$

Binary Exposure - Parameter Interpretations

- Probabilities:

$$\pi(\boldsymbol{X}_1 = 0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

$$\pi(\boldsymbol{X}_1 = 1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$$

- Odds Ratio:

$$\frac{\text{Odds}(\boldsymbol{X}_1 = 1)}{\text{Odds}(\boldsymbol{X}_1 = 0)} = \exp(\beta_1)$$

$$\text{exposure log odds ratio} = \beta_1$$

Binary Exposure - Parameter Estimates?

Q: Can we estimate the model parameters?

$$\begin{aligned}
\widehat{\pi}(\boldsymbol{X}_1 = 0) &= 0.221 \\
\log\left(\frac{\pi(\boldsymbol{X}_1 = 0)}{1 - \pi(\boldsymbol{X}_1 = 0)}\right) &= \beta_0 \\
\log\left(\frac{0.221}{1 - 0.221}\right) &= -1.260
\end{aligned}$$

$$\begin{aligned}
\widehat{\pi}(\boldsymbol{X}_1 = 1) &= 0.122 \\
\log\left(\frac{\pi(\boldsymbol{X}_1 = 1)}{1 - \pi(\boldsymbol{X}_1 = 1)}\right) &= \beta_0 + \beta_1 \\
\log\left(\frac{0.122}{1 - 0.122}\right) &= -1.974
\end{aligned}$$

$$\begin{aligned}
\widehat{\beta}_0 &= -1.260 \\
\widehat{\beta}_1 &= -1.974 \; - \; -1.260 \; = \; -0.713
\end{aligned}$$

$$\log \widehat{OR} = \log(0.490) = -0.713$$

```
. expand count

. cs y x

                 | Treatment               |
                 |   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        |
                 +------------------------------------------------
                            chi2(1) =     4.81  Pr>chi2 = 0.0283

. cc y x

                 | Treatment               |             Proportion
                 |   Exposed   Unexposed   |      Total     Exposed
-----------------+-------------------------+----------------------
           Cases |        17          31   |         48      0.3542
        Controls |       122         109   |        231      0.5281
-----------------+-------------------------+----------------------
           Total |       139         140   |        279      0.4982
                 |                         |
                 |      Point estimate     |  [95% Conf. Interval]
                 |-------------------------+----------------------
      Odds ratio |         .4899524        |  .2588072    .9282861  (Cornfield)
 Prev. frac. ex. |         .5100476        |  .0717139    .7411928  (Cornfield)
 Prev. frac. pop |         .2693758        |
                 +------------------------------------------------
                            chi2(1) =     4.81  Pr>chi2 = 0.0283
```


Binary Exposure - Parameter Estimates

```
. logistic 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 | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |   .4899524   .1613518     -2.166   0.030       .2569419    .9342709
------------------------------------------------------------------------------

.
. logit

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     -2.166   0.030      -1.358905   -.0679889
   _cons |  -1.257361   .2035494     -6.177   0.000       -1.65631   -.8584111
------------------------------------------------------------------------------

.
```

Summary

1. For 2×2 tables logistic regression fits a pair of probabilities: $\widehat{\pi}(\boldsymbol{X} = 1)$ and $\widehat{\pi}(\boldsymbol{X} = 0)$.
2. Model: $\text{logit}\pi(\boldsymbol{X}) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{X}$.
3. $\boldsymbol{\beta}_0$ represents the reference "log odds".
4. $\boldsymbol{\beta}_1$ is the log odds ratio that compares the odds of response among the "exposed" ($\boldsymbol{X} = 1$) to the odds of response among the "unexposed" ($\boldsymbol{X} = 0$).
5. $\widehat{\boldsymbol{\beta}}_1 = \log[(ad)/(bc)]$: the logistic regression odds ratio and the simple 2×2 odds ratio are identical.
6. Note: the estimated standard errors (95% CI) from logistic regression are slightly different from the Woolf estimates in 2×2 table analysis.

Summary for single binary covariate

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

Probability

$\boldsymbol{X}_1 = 1$	$\boldsymbol{X}_1 = 0$
$\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$	$\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$

Odds

$\boldsymbol{X}_1 = 1$	$\boldsymbol{X}_1 = 0$
$\exp(\beta_0 + \beta_1)$	$\exp(\beta_0)$

Odds Ratios

$\boldsymbol{X}_1 = 1$	$\boldsymbol{X}_1 = 0$
$\exp(\beta_1)$	1.0

Logistic Regression: Some simple special cases

Q: What does logistic regression provide for a simple binary "exposure" variable, $\boldsymbol{E}$, a simple binary stratifying variable, $\boldsymbol{C}$, and a binary outcome, $\boldsymbol{Y}$?

Example: HIVNET Vaccine Trial Willingness

Previous Study

	Willing	Not willing	
Knowledge >= 8	22	112	134
Knowledge <= 7	32	153	185
	54	265	319

New Recruits

	Willing	Not willing	
Knowledge >= 8	39	146	185
Knowledge <= 7	67	179	246
	106	325	431

Cohort as a potenetial confounder?

```
.  table knowhigh, c(mean will6) f(%5.2f)

----------+------------
knowledge |
high      | mean(will6)
----------+------------
know6 <=7 |        0.24
know6 >=8 |        0.19
----------+------------

. table cohort, c(mean will6) f(%5.2f)

-----------+------------
recruitmen |
t cohort   | mean(will6)
-----------+------------
old cohort |        0.17
new cohort |        0.25
-----------+------------

. table cohort, c(mean knowhigh) f(%5.2f)

-----------+---------------
recruitmen |
t cohort   | mean(knowhigh)
-----------+---------------
old cohort |           0.42
new cohort |           0.43
-----------+---------------
```

Define Variables

- Let $\boldsymbol{X}_1$ be a dummy variable for "exposure" groups:
  $\boldsymbol{X}_1 = 1$ if subject has knowledge score ≥ 8.
  $\boldsymbol{X}_1 = 0$ if subject has knowledge score ≤ 7.

- Let $\boldsymbol{X}_2$ be a dummy variable for stratification groups:
  $\boldsymbol{X}_2 = 1$ if subject is a new recruit.
  $\boldsymbol{X}_2 = 0$ if subject is from a previous study (rollover).

- Let $\boldsymbol{Y}$ be an indicator (binary) outcome variable:
  $\boldsymbol{Y} = 1$ if subject is definitely willing at month 6.
  $\boldsymbol{Y} = 0$ if subject is not definitely willing at month 6.

$\boldsymbol{X}_2 = 0$

	$\boldsymbol{Y} = 1$	$\boldsymbol{Y} = 0$	
$\boldsymbol{X}_1 = 1$	22	112	134
$\boldsymbol{X}_1 = 0$	32	153	185
	54	265	319

$\boldsymbol{X}_2 = 1$

	$\boldsymbol{Y} = 1$	$\boldsymbol{Y} = 0$	
$\boldsymbol{X}_1 = 1$	39	146	185
$\boldsymbol{X}_1 = 0$	67	179	246
	106	325	431

From these data we can estimate:

Previous Study, $\boldsymbol{X}_2 = 0$:

$$
\begin{aligned}
\widehat{P}[\boldsymbol{Y}=1 \mid \boldsymbol{X}_1=1,\ \boldsymbol{X}_2=0] &= 22/134 = 0.164 \\
\widehat{P}[\boldsymbol{Y}=1 \mid \boldsymbol{X}_1=0,\ \boldsymbol{X}_2=0] &= 32/185 = 0.173 \\
\widehat{RR}_0 &= 0.164/0.173 = 0.948 \\
\widehat{OR}_0 &= \frac{0.164/(1-0.164)}{0.173/(1-0.173)} = 0.938
\end{aligned}
$$

New Recruits, $\boldsymbol{X}_2 = 1$:

$$
\begin{aligned}
\widehat{P}[\boldsymbol{Y}=1 \mid \boldsymbol{X}_1=1,\ \boldsymbol{X}_2=1] &= 39/185 = 0.211 \\
\widehat{P}[\boldsymbol{Y}=1 \mid \boldsymbol{X}_1=0,\ \boldsymbol{X}_2=1] &= 67/246 = 0.272 \\
\widehat{RR}_1 &= 0.211/0.272 = 0.776 \\
\widehat{OR}_1 &= \frac{0.211/(1-0.211)}{0.272/(1-0.272)} = 0.716
\end{aligned}
$$

Binary Exposure

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

Data:

Y	X1	X2	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

Possible Approaches / Models

Stratification / Additive Model

- Mantel-Haenszel estimates a Common Odds Ratio
- Regression Model:

$$\text{logit}\,(\pi(\boldsymbol{X})) = \beta_0 + \beta_1\,\text{KNOW}(1) + \beta_2\,\text{COHORT}(1)$$

Effect Modification / Interaction Model

- Separate Odds Ratio Estimates for Cohort Groups
- Regression Model:

$$\begin{aligned}\text{logit}\,(\pi(\boldsymbol{X})) = \ & \beta_0 \\ & +\beta_1\,\text{KNOW}(1) \\ & +\beta_2\,\text{COHORT}(1) \\ & +\beta_3\,\text{KNOW}(1)\cdot\text{COHORT}(1)\end{aligned}$$

Interpreting Adjusted Odds Ratios

1. What is the <u>outcome</u> of interest?
   (ie. Disease)
2. What are the <u>2 groups</u> being contrasted?
   (ie. Exposed vs Unexposed)
3. In the comparison what <u>other variables</u> are held fixed, or controlled?
   (ie. Age, Gender)

Logistic Regression Models

- Understand (graph) the logit structure.
  - additive model?
  - model with interaction?
  - exposure categories?

- logit $\Rightarrow$ probability or risk
  - $\uparrow$ logit $\Rightarrow$ $\uparrow$ odds $\Rightarrow$ $\uparrow$ probability

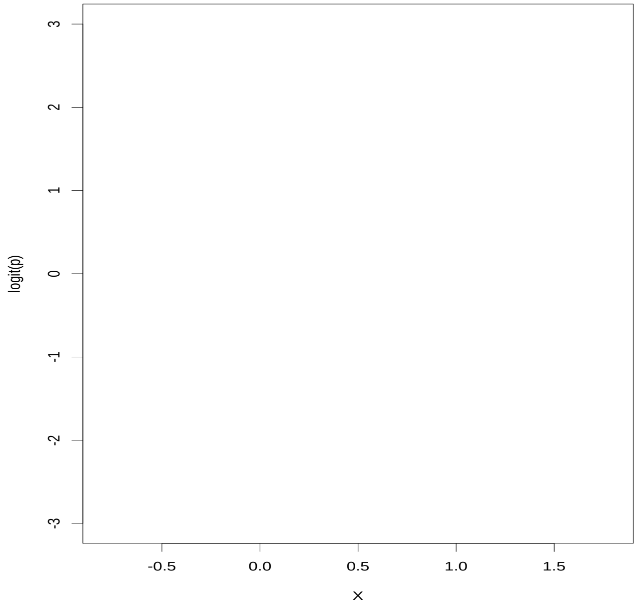
- Differences in logits are log Odds Ratios.

$$\text{logit}[\pi(\boldsymbol{X}^{(1)})] \; - \; \text{logit}[\pi(\boldsymbol{X}^{(0)})] \quad =$$

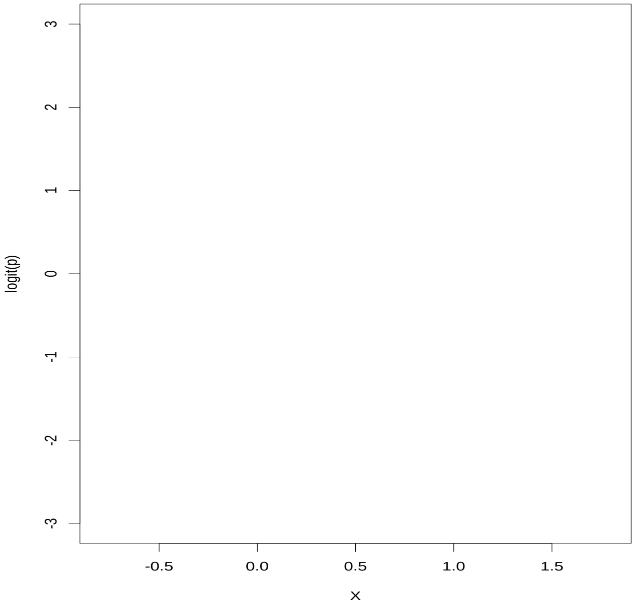
$$\log \; \text{Odds}(\boldsymbol{X}^{(1)}) \; - \; \log \; \text{Odds}(\boldsymbol{X}^{(0)}) \quad =$$

$$\log \left[\frac{\text{Odds}(\boldsymbol{X}^{(1)})}{\text{Odds}(\boldsymbol{X}^{(0)})} \right] \quad =$$

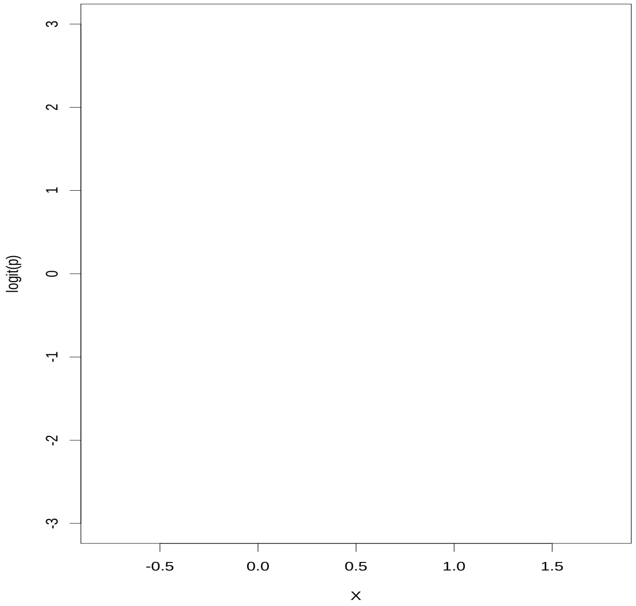
log odds ratio comaring $\boldsymbol{X}^{(1)}$ to $\boldsymbol{X}^{(0)}$



$$\text{logit}[\pi(\boldsymbol{X})] \; = \; \beta_0 + \beta_1 \text{KNOW}(1)$$



$$\text{logit}[\pi(\boldsymbol{X})] \;=\; \beta_0 + \beta_1\text{KNOW}(1) + \beta_2\text{COHORT}(1)$$



$$\begin{aligned}\text{logit}[\pi(\boldsymbol{X})] \;&=\; \beta_0 + \beta_1\text{KNOW}(1) + \beta_2\text{COHORT}(1)\\ &\quad + \beta_3\text{KNOW}(1)\cdot\text{COHORT}(1)\end{aligned}$$

Model:

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

$$\log\left(\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right) = \beta_0 + \beta_1\boldsymbol{X}_1 + \beta_2\boldsymbol{X}_2 + \beta_3\boldsymbol{X}_1\boldsymbol{X}_2$$

$$\text{logit}[\pi(\boldsymbol{X})] = \beta_0 + \beta_1\boldsymbol{X}_1 + \beta_2\boldsymbol{X}_2 + \beta_3\boldsymbol{X}_1\boldsymbol{X}_2$$

Parameter Interpretations

- Probabilities:

$$\pi(\boldsymbol{X}_1=0,\boldsymbol{X}_2=0) = \frac{\exp(\beta_0)}{1+\exp(\beta_0)}$$

$$\pi(\boldsymbol{X}_1=1,\boldsymbol{X}_2=0) = \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}$$

$$\pi(\boldsymbol{X}_1=0,\boldsymbol{X}_2=1) = \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)}$$

$$\pi(\boldsymbol{X}_1=1,\boldsymbol{X}_2=1) = \frac{\exp(\beta_0+\beta_1+\beta_2+\beta_3)}{1+\exp(\beta_0+\beta_1+\beta_2+\beta_3)}$$

Parameter Interpretations

- Odds Ratios:

$$\frac{\text{Odds}(\boldsymbol{X}_1=1,\boldsymbol{X}_2=0)}{\text{Odds}(\boldsymbol{X}_1=0,\boldsymbol{X}_2=0)} = \exp(\beta_1)$$

$$\text{knowledge log odds ratio for previous study} = \beta_1$$

$$\frac{\text{Odds}(\boldsymbol{X}_1=1,\boldsymbol{X}_2=1)}{\text{Odds}(\boldsymbol{X}_1=0,\boldsymbol{X}_2=1)} = \exp(\beta_1+\beta_3)$$

$$\text{knowledge log odds ratio for new recruits} = \beta_1+\beta_3$$

Two Binary Covariates - Parameter Estimates?

Q: Can we estimate the model parameters?

$$
\begin{aligned}
\widehat{\pi}(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 0) &= 0.173 \\
\log\left(\frac{\pi(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 0)}{1-\pi(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 0)}\right) &= \beta_0 \\
\log\left(\frac{0.173}{1-0.173}\right) &= -1.564
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\pi}(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 0) &= 0.164 \\
\log\left(\frac{\pi(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 0)}{1-\pi(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 0)}\right) &= \beta_0 + \beta_1 \\
\log\left(\frac{0.164}{1-0.164}\right) &= -1.629
\end{aligned}
$$

$$
\widehat{\beta}_0 = -1.564
$$

$$
\begin{aligned}
\widehat{\beta}_1 &= -1.629 \; - \; -1.564 \\
&= -0.064
\end{aligned}
$$

$$
\log \widehat{OR}_0 = \log(0.938) \; = \; -0.064
$$

$$
\begin{aligned}
\widehat{\pi}(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 1) &= 0.272 \\
\log\left(\frac{\pi(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 1)}{1-\pi(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 1)}\right) &= \beta_0 + \beta_2 \\
\log\left(\frac{0.272}{1-0.272}\right) &= -0.984
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\pi}(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 1) &= 0.211 \\
\log\left(\frac{\pi(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 1)}{1-\pi(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 1)}\right) &= \beta_0 + \beta_1 + \beta_2 + \beta_3 \\
\log\left(\frac{0.211}{1-0.211}\right) &= -1.319
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\beta}_2 &= -0.984 \; - \; -1.564 \\
&= 0.580 \\
\widehat{\beta}_3 &= -1.319 \; - \\
& \quad (-1.564 \; + \; -0.064 \; + \; 0.580) \\
&= -1.319 \; - \; -1.049 \\
&= -0.270
\end{aligned}
$$

$$
\begin{aligned}
\log \widehat{OR}_1 &= \log(0.716) \; = \; -0.334 \\
\log \widehat{OR}_1 - \log \widehat{OR}_0 &= \log(0.716) - \log(0.938) \\
&= -0.334 \; - \; -0.064 \\
&= -0.270
\end{aligned}
$$

Two Covariates - Parameter Estimates

```
. cc will6 knowhigh

                 | knowledge high         |             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       |  .5548161    1.133214  (Cornfield)
 Prev. frac. ex. |         .2071099       | -.1332137    .4451839  (Cornfield)
 Prev. frac. pop |         .0905667       |
                 +-----------------------------------------------
                               chi2(1) =     1.62  Pr>chi2 = 0.2035
.
. logistic will6 knowhigh

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

------------------------------------------------------------------------------
   will6 | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
knowhigh |   .7928901   .1448677     -1.270   0.204       .5542317    1.134318
------------------------------------------------------------------------------

. logit

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

------------------------------------------------------------------------------
   will6 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
knowhigh |  -.2320706   .1827084     -1.270   0.204      -.5901724    .1260312
   _cons |  -1.210015   .1145122    -10.567   0.000      -1.434455   -.9855753
------------------------------------------------------------------------------
```



```
. logistic will6 knowhigh cohort

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

------------------------------------------------------------------------------
   will6 | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
knowhigh |   .7879471   .1445962     -1.299   0.194       .5499117    1.129019
  cohort |   1.605702    .299915      2.535   0.011       1.113465    2.315546
------------------------------------------------------------------------------

. logit

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

------------------------------------------------------------------------------
   will6 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
knowhigh |  -.2383243   .1835101     -1.299   0.194      -.5979975    .1213488
  cohort |   .4735612   .1867812      2.535   0.011       .1074767    .8396457
   _cons |  -1.495195   .1650607     -9.058   0.000      -1.818708   -1.171682
------------------------------------------------------------------------------

.
. xi: logistic will6 i.knowhigh*i.cohort
i.knowhigh           Iknowh_0-1   (naturally coded; Iknowh_0 omitted)
i.cohort             Icohor_0-1   (naturally coded; Icohor_0 omitted)
i.knowhigh*i.cohort  IkXc_#-#     (coded as above)

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

------------------------------------------------------------------------------
   will6 | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Iknowh_1 |   .9391741   .2851272     -0.207   0.836       .5179965    1.702807
Icohor_1 |    1.78963   .4321054      2.410   0.016       1.114913    2.872668
IkXc_1_1 |    .759878   .2895239     -0.721   0.471       .3601011    1.603479
------------------------------------------------------------------------------
```

```
. logit

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

------------------------------------------------------------------------------
   will6 |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Iknowh_1 |  -.0627544   .3035936     -0.207   0.836      -.6577869    .5322781
Icohor_1 |   .5820088   .2414496      2.410   0.016       .1087763    1.055241
IkXc_1_1 |  -.2745974   .3810136     -0.721   0.471       -1.02137    .4721756
   _cons |  -1.564702   .1943861     -8.049   0.000      -1.945692   -1.183712
------------------------------------------------------------------------------
```


Summary

1. With two binary covariates we can model 4 probabilities:
$$\pi(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 0)$$
$$\pi(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 0)$$
$$\pi(\boldsymbol{X}_1 = 0, \boldsymbol{X}_2 = 1)$$
$$\pi(\boldsymbol{X}_1 = 1, \boldsymbol{X}_2 = 1)$$

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 ($\boldsymbol{\beta}_3$) is the difference between these log odds ratios.

4. **Q**: How to interpret the simpler model that doesn’t contain the interaction term, $\beta_3 \boldsymbol{X}_1 \boldsymbol{X}_2$?

Summary of the Odds

Saturated Model

	$\boldsymbol{X}_2 = 1$	$\boldsymbol{X}_2 = 0$
$\boldsymbol{X}_1 = 1$	$\exp(\beta_0 + \beta_1 + \beta_2 + \beta_3)$	$\exp(\beta_0 + \beta_1)$
$\boldsymbol{X}_1 = 0$	$\exp(\beta_0 + \beta_2)$	$\exp(\beta_0)$

Summary of the Odds Ratios

Saturated Model

	$\boldsymbol{X}_2 = 1$	$\boldsymbol{X}_2 = 0$
$\boldsymbol{X}_1 = 1$	$\exp(\beta_1 + \beta_2 + \beta_3)$	$\exp(\beta_1)$
$\boldsymbol{X}_1 = 0$	$\exp(\beta_2)$	1.0

Summary of the Odds

Additive log Odds Ratio Model
(Multiplicative Odds Ratio Model)

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

Summary of the Odds Ratios

Additive log Odds Ratio Model
(Multiplicative Odds Ratio Model)

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

Regression versus Mantel-Haenszel

```
.
. ***
. *** let's compare Mantel-Haenszel to logistic regression
. ***
.
. cc will6 knowhigh, by(cohort) bd

recruitment coho |       OR      [95% Conf. Interval]    M-H Weight
-----------------+-------------------------------------------------
      old cohort |   .9391741     .5208006   1.694425      11.23511 (Cornfield)
      new cohort |   .7136577     .4553849    1.11864      22.69606 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   .7928901     .5548161   1.133214               (Cornfield)
    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
```

Q: Similarities?

Q: Advantages / disadvantages ?

Logistic Regression: Estimation

Q: What is the RANDOM component of a logistic regression model?

(What's the analog of normal errors?)

Q: How are logistic regression parameters estimated?

(What's analog of Least Squares?)

Q: How can we perform inference regarding parameter estimates?

(What's analog of Wald and F statistics?)

Q: Why can we use logistic regression for either prospective data or case-control data ?

Q: What's the difference between <u>conditional</u> and <u>unconditional</u> methods?

What's Random?

A: Depends on the study design.

1. Follow-up Study: In this type of study we collect baseline information (covariates $\boldsymbol{X}$) and then follow subjects forward and measure the occurrence of an outcome, $\boldsymbol{Y}$.

   Fixed:

   Random:

2. Case Control Study: In this type of study we collect a certain number of cases ($\boldsymbol{Y} = 1$) and a certain number of controls ($\boldsymbol{Y} = 0$). We then ascertain exposure, or covariate information ($\boldsymbol{X}$).

   Fixed:

   Random:

Follow-up Study

We've seen that logistic regression can be used to model the probability of disease for population subgroups (specific values of $\boldsymbol{X}$). We model:

$$\begin{aligned} E(\boldsymbol{Y} \mid \boldsymbol{X}) &= \pi(\boldsymbol{X}) \\ P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}) &= \pi(\boldsymbol{X}) \end{aligned}$$

Characteristics:

- $\boldsymbol{Y}$ takes only the values 0 and 1.
- $\boldsymbol{Y}$ has a mean of $\pi(\boldsymbol{X})$.

Results:

- There's only one probability model for 0/1 data.

- Knowing $\pi(\boldsymbol{X})$ determines the probability for any outcome:

$P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}) =$

$P(\boldsymbol{Y} = 0 \mid \boldsymbol{X}) =$

- $V(\boldsymbol{Y} \mid \boldsymbol{X}) = \pi(\boldsymbol{X})[1 - \pi(\boldsymbol{X})]$.

The Model

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

SYSTEMATIC: $P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}) = \pi(\boldsymbol{X})$

$\text{logit}[\pi(\boldsymbol{X})] = \beta_0 + \beta_1 \boldsymbol{X}_1 + \ldots + \beta_p \boldsymbol{X}_p$

RANDOM: $\boldsymbol{Y} \sim b(1, \pi(\boldsymbol{X}))$

$\boldsymbol{Y}$'s are independent

The statement $\boldsymbol{Y} \sim b(1, \pi(\boldsymbol{X}))$ says:
"the variable $\boldsymbol{Y}$ is distributed as a binomial random variable with total $n = 1$ and success probability $\pi(\boldsymbol{X})$."

Using the logistic regression model

Q: When we use statistical models we want to criticize them (evaluate whether they are appropriate). What does that mean for logistic regression?

SYSTEMATIC: Are we using $\boldsymbol{X}$ appropriately?
(linear, quadratic)?
Which $\boldsymbol{X}$'s do I want in the model?

RANDOM: Are my observations independent?

Note: There is much less to fuss about in terms of the "errors" than with linear regression.

Note: We can use the logistic regression model for other types of binomial data where $n > 1$ but then we need to worry about the variance (random) assumptions.
Examples:

- Pups born to rats that have been exposed to a chemical (teratology).
- Condom use (Yes/No) for each episode in a month.

Violations of Independence

Examples:

- "Longitudinal Data" – Repeated outcomes (over time) from a sample of individuals.

- Observations from the same clinic if not stratifying on clinic.

- Community randomized studies (analyze at the unit of randomization).

- "Multilevel Data" – patients, nested within doctors, nested within hospitals.

Formally: After knowing the covariates, $\boldsymbol{X}$, I still have some information (a variable Z), that leads me to expect that outcomes Y_i and Y_j will be more similar if they share the same value of Z than if they do not.

This variable, Z, is a "cluster" identifier.
Examples:
- Longitudinal data – Subject ID
- Multilevel data – Doctor Name, Hospital Name.

Impact of Dependence

1. In general, $\widehat{\boldsymbol{\beta}}$, is consistent. (ie. this violation doesn't hurt modeling of the means, $\pi(\boldsymbol{X})$).
2. Estimates of precision (standard errors for $\widehat{\boldsymbol{\beta}}$) will be wrong. They may be too large or too small depending on the situation.
3. Since the standard errors are not valid we will have incorrect inference (hypothesis tests, confidence intervals).

Corrections for Dependence

1. Finesse it away! (take a summary for each person, cluster)
2. Find methods that can handle it appropriately:
   - GEE methods (Liang and Zeger, 1986)
     SAS proc genmod
     STATA `xtgee`
   - Mixed Models / Multilevel Models
     SAS proc mixed
     STATA `xtlogit`

   (covered in Biostat 536!)

Estimation

Least Squares: Linear regression.

- The mean estimates, $\widehat{\beta}_0 + \widehat{\beta}_1 X_i$ that are "closest" to the observed data, Y_i.

- If we assume normality of errors, then least squares is a special case of a more general statistical estimation method known as maximum likelihood.

Maximum Likelihood: Logistic regression.

- Fisher (1922) invented this general method.

Problem: Unknown model parameters, θ.

Set-up: Write the probability of the data, $\boldsymbol{Y}$, in terms of the model parameter and the data, $P(\boldsymbol{Y}, \theta)$.

Solution: Choose as your estimate the value of the unknown parameter that makes your data look as likely as possible. Pick $\widehat{\theta}$ that puts the largest possible probability on your data.

(*) Examples

(*) = this material is for your information (not required).

You have a simple sample of $Y_i = 0/1$ for $i = 1, 2, \ldots, n$ and you want to estimate $P(Y = 1) = \pi$.

Probability: $P(Y_1, \ldots, Y_n;\ \pi\ ) = \prod_{i=1}^{n} \pi^{Y_i}(1-\pi)^{(1-Y_i)}$

Let's fix the number in the sample at $n = 20$. Then all we need to know is the number of Y_i's that were 1, call that $Z = \sum Y_i$.

$$P(Y_1, Y_2, \ldots Y_{20};\ \pi\ ) = \pi^{\sum Y_i}(1-\pi)^{(20-\sum Y_i)}$$

$$P(Z;\ \pi\ ) = \binom{20}{Z}\pi^{Z}(1-\pi)^{(20-Z)}$$

Two ways to look at this:

- Fix π and look at the probability of different values of Z:

$\pi = 0.1$

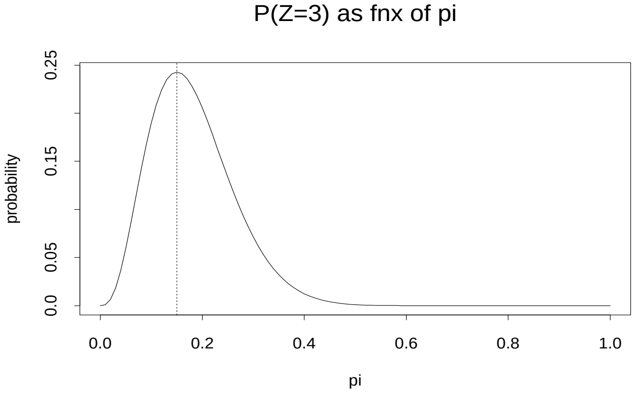
Z	$P(Z, \pi)$
0	0.122
1	0.270
2	0.285
3	0.190
4	0.090
5	0.032

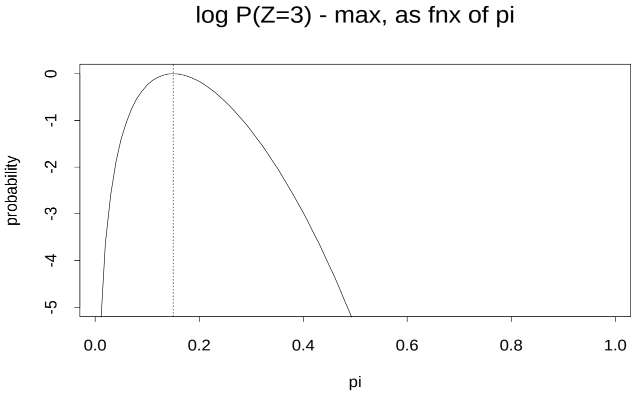
- Fix Z and look at the probability under different values of π:

$Z = 3$

π	$P(Z, \pi)$
0.01	0.001
0.05	0.060
0.10	0.190
0.20	0.205
0.30	0.072
0.40	0.012

If you observe the data $Z = 3$ then the likelihood function is being shown in the second table, or the plots below:





Comments:

- The maximum likelihood estimate in this example is simply the observed proportion $\widehat{\pi} = Z/20 = 3/20 = 0.15$.

- The likelihood function lets us compare different values of the parameters – showing which values are supported by the data and which values we can essentially eliminate.

- Maximum likelihood is the "best" method of estimation for any situation that you are willing to write down a probability model.

- Maximum likelihood can be used even when the unknown parameter has several components (ie. $\beta_0, \beta_1, \ldots, \beta_p$).

- Computers can find these maximum values!

Logistic Regression(*)

Q: What is the likelihood function when using logistic regression?

A: Similar to before – we write a $\pi(\boldsymbol{X}_i\boldsymbol{\beta})$ for every $Y_i = 1$ and a $1 - \pi(\boldsymbol{X}_i\boldsymbol{\beta})$ for every $Y_i = 0$:

$$P(Y_1, Y_2, \ldots, Y_n) = \prod_{i=1}^{n} \pi(\boldsymbol{X}_i\boldsymbol{\beta})^{Y_i} \left[1 - \pi(\boldsymbol{X}_i\boldsymbol{\beta})\right]^{(1-Y_i)}$$

where we have the logistic model for the π's:

$$\pi(\boldsymbol{X}_i\boldsymbol{\beta}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}$$

Maximum Likelihood estimate of $\boldsymbol{\beta}$ is denoted $\widehat{\boldsymbol{\beta}}$. This value makes the probability of the observed data as large as possible – "fits" the data best.

Logistic Regression and Likelihood

Q: If I'm not a theoretician, but simply want to analyze my data, then why should I care about likelihoods?

A: We use comparisons in the value of the likelihood function as the preferred method for testing whether certain variables (coefficients) are significant (ie. to test $H_o : \beta_j = 0$).

In Linear Regression we used the change in the residual sum of squares (partial F test) as a method for seeing if variables were significant.

In Logistic Regression we will use the change in the log likelihood as a method for seeing if variables are significant.

Summary

1. The RANDOM part of the logistic model is simple.
2. The assumption of independence is important.
3. The probability of the observed data is called the "likelihood".
4. The logistic regression coefficient estimate $\widehat{\boldsymbol{\beta}}$ is obtained via maximum likelihood.
5. We will discuss how the maximized log likelihood, $-2 \log L$, is used for statistical inference.

Logistic Regression: Inference

- "Nested" models
- Maximized log likelihood, $\log L$, & Likelihood Ratio Tests
- $\widehat{\boldsymbol{\beta}}$ and standard errors – Wald Tests
- Inference for linear combinations of $\widehat{\boldsymbol{\beta}}$

“Nested” Models

When a scientific hypothesis can be formulated in terms of restrictions on a set of parameters (ie. 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).

Example:

$$\begin{aligned} \text{Model 1}: \quad \text{logit}[\pi(\boldsymbol{X})] &= \beta_0 + \beta_1 \boldsymbol{X}_1 \\ \text{Model 2}: \quad \text{logit}[\pi(\boldsymbol{X})] &= \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \beta_3 \boldsymbol{X}_3 \end{aligned}$$

- Model 1 is a special case of Model 2.
- Model 1 is said to be <u>nested</u> within Model 2.
- Model 1 has a <u>subset</u> of the variables contained in Model 2.

◦ By looking at the relative goodness-of-fit of these two models we can judge whether the additional flexibility in Model 2 was important.

Likelihood Ratio Statistics

We can use the maximum likelihood fits from nested models to test if the “difference” between these models is significant.

Example:

$$\begin{aligned} \text{Model 1}: \quad \text{logit}[\pi(\boldsymbol{X})] &= \beta_0 + \beta_1 \boldsymbol{X}_1 \\ \text{Model 2}: \quad \text{logit}[\pi(\boldsymbol{X})] &= \beta_0 + \beta_1 \boldsymbol{X}_1 + \beta_2 \boldsymbol{X}_2 + \beta_3 \boldsymbol{X}_3 \end{aligned}$$

Model 1 is formed from Model 2 by the hypothesis:

$$H_o: \quad \beta_2 = \beta_3 = 0$$

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

$$\begin{aligned} \text{Model 1}: &\quad \log L_1 \\ \text{Model 2}: &\quad \log L_2 \end{aligned}$$

We can then use the Likelihood Ratio Statistic:

$$LR \;=\; 2 \times (\log L_2 - \log L_1)$$

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.

Example

```
infile age alc tob y count using /biostat/.../NewTuyns.dat

label variable age "Age Group"
label variable alc "Alcohol"
label variable tob "Tobacco"
label variable y "Case/Control Status"

label define agegps 1 "25-34" 2 "35-44" 3 "45-54" 4 "55-64" 5 "65-74" 6 "75+"
label define alcgps 1 "<40g/day" 2 "40-79g/day" 3 "80-119g/day" 4 "120+g/day"
label define tobgps 1 "0-9g/day" 2 "10-19g/day" 3 "20-29g/day" 4 "30+g/day"
label define status 1 "Case" 0 "Control"

label values age agegps
label values alc alcgps
label values tob tobgps
label values y status

tabodds y age [freq=count], or
tabodds y tob [freq=count], or


drop if count==0
expand count


xi:  logistic y i.age
logit
lrtest, saving(1)

xi: logistic y i.age i.tob
logit
lrtest, saving(2)


lrtest, using(2) model(1)
```



```
. do NewTclass

. infile age alc tob y count using c:\YOUR-PATH\NewTuyns.dat
(176 observations read)
.
(label definitions)
.
.
. tabodds y age [freq=count], or

------------+------------------------------------------------------------
        age |  Odds ratio       chi2       P>chi2     [95% Conf. Interval]
------------+------------------------------------------------------------
      25-34 |    1.000000          .            .             .          .
      35-44 |    5.534759       3.26       0.0711      0.682304  44.897259
      45-54 |   31.676647      26.29       0.0000      3.943092 254.472873
      55-64 |   52.650602      43.21       0.0000      6.304213 439.719592
      65-74 |   59.669811      46.18       0.0000      6.674741 533.426917
        75+ |   48.225806      32.67       0.0000      4.682406 496.695189
------------+------------------------------------------------------------
Test of homogeneity (equal odds): chi2(5)  =    95.98
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    82.57
                                  Pr>chi2  =   0.0000

. tabodds y tob [freq=count], or

------------+------------------------------------------------------------
        tob |  Odds ratio       chi2       P>chi2     [95% Conf. Interval]
------------+------------------------------------------------------------
   0-9g/day |    1.000000          .            .             .          .
  10-19g/da |    1.899341      11.02       0.0009      1.292147   2.791862
  20-29g/da |    1.910256       7.72       0.0055      1.200295   3.040153
   30+g/day |    3.483409      25.31       0.0000      2.074288   5.849783
------------+------------------------------------------------------------
Test of homogeneity (equal odds): chi2(3)  =    29.61
                                  Pr>chi2  =   0.0000

Score test for trend of odds:     chi2(1)  =    26.99
                                  Pr>chi2  =   0.0000
.
. drop if count==0
(41 observations deleted)

. expand count
(837 observations created)
```

```
.
.
. xi:  logistic y i.age
i.age              Iage_1-6     (naturally coded; Iage_1 omitted)

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(5)      =     119.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -434.08202                       Pseudo R2       =     0.1214

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Iage_2 |   5.534759    5.87086     1.613   0.107       .6921617    44.25781
  Iage_3 |   31.67665   32.24812     3.394   0.001       4.307064    232.9684
  Iage_4 |    52.6506   53.37903     3.910   0.000       7.218139    384.0444
  Iage_5 |   59.66981   60.74304     4.017   0.000       8.114156    438.7993
  Iage_6 |   48.22581   50.98864     3.666   0.000       6.071739    383.0416
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(5)      =     119.94
                                                  Prob > chi2     =     0.0000
Log likelihood = -434.08202                       Pseudo R2       =     0.1214

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Iage_2 |   1.711048   1.060725     1.613   0.107      -.3679356    3.790032
  Iage_3 |    3.45558   1.018041     3.394   0.001       1.460256    5.450903
  Iage_4 |   3.963678   1.013835     3.910   0.000       1.976597    5.950758
  Iage_5 |   4.088826   1.017986     4.017   0.000        2.09361    6.084042
  Iage_6 |   3.875894   1.057289     3.666   0.000       1.803645    5.948144
   _cons |  -4.744932   1.004331    -4.724   0.000      -6.713384    -2.77648
------------------------------------------------------------------------------

. lrtest, saving(1)

.
. xi: logistic y i.age i.tob
i.age              Iage_1-6     (naturally coded; Iage_1 omitted)
i.tob              Itob_1-4     (naturally coded; Itob_1 omitted)
```



```
Logit estimates                                   Number of obs   =        972
                                                  LR chi2(8)      =     156.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -415.74964                       Pseudo R2       =     0.1585

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Iage_2 |   6.140108   6.544626     1.703   0.089       .7601446    49.59704
  Iage_3 |   36.17285   37.07026     3.501   0.000       4.853599    269.5886
  Iage_4 |   61.72942   63.03597     4.037   0.000        8.34208    456.7831
  Iage_5 |   83.48177   85.76944     4.307   0.000        11.1446    625.3438
  Iage_6 |   60.39319   64.45659     3.842   0.000       7.456163    489.1707
  Itob_2 |   1.842308   .3797414     2.964   0.003       1.230014    2.759397
  Itob_3 |   1.944706   .4874833     2.653   0.008       1.189821     3.17853
  Itob_4 |   5.696028   1.721364     5.757   0.000       3.150181    10.29933
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(8)      =     156.61
                                                  Prob > chi2     =     0.0000
Log likelihood = -415.74964                       Pseudo R2       =     0.1585

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  Iage_2 |   1.814842   1.065881     1.703   0.089      -.2742466    3.903931
  Iage_3 |   3.588309   1.024809     3.501   0.000       1.579721    5.596897
  Iage_4 |   4.122761   1.021166     4.037   0.000       2.121313    6.124209
  Iage_5 |   4.424628   1.027403     4.307   0.000       2.410955    6.438302
  Iage_6 |   4.100876   1.067283     3.842   0.000       2.009041    6.192712
  Itob_2 |    .611019   .2061227     2.964   0.003       .2070259    1.015012
  Itob_3 |   .6651108    .250672     2.653   0.008       .1738028    1.156419
  Itob_4 |   1.739769   .3022042     5.757   0.000        1.14746    2.332078
   _cons |  -5.367645   1.017863    -5.273   0.000       -7.36262    -3.37267
------------------------------------------------------------------------------

. lrtest, saving(2)
.
.
. lrtest, using(2) model(1)
Logistic:  likelihood-ratio test                      chi2(3)     =      36.66
                                                      Prob > chi2 =     0.0000
```

Wald Statistics

Most statistical packages produce tables:

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

From this table we can obtain the following:

- $\widehat{\beta}_j \pm 1.96 s_j$ is a 95% confidence interval for β_j.

- $2 \times P[Z > |\widehat{\beta}_j/s_j|]$ = p-value for testing $H_o : \beta_j = 0$.

Q: What about combinations of parameters? (ie. $\beta_2 - \beta_1$)

Linear Combinations

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

$\boxed{\text{For example:}}$ An odds ratio that compares TOB(4) to TOB(3) adjusting for AGE.

We use the following result:

$$\widehat{V}(a\widehat{\beta}_j + b\widehat{\beta}_k) \;=\; a^2 \times s_j^2 \;+\; 2ab \times c_{jk} \;+\; b^2 \times s_k^2$$

For example:

- $\widehat{V}(\widehat{\beta}_j + \widehat{\beta}_k) \;=$

- $\widehat{V}(\widehat{\beta}_j - \widehat{\beta}_k) \;=$

- $\widehat{V}(10 \times \widehat{\beta}_j) \;=$

Or the more general result:

$$\widehat{V}(\sum_j a_j \widehat{\beta}_j) \;=\; \sum_j a_j^2 \times s_j^2 \;+\; 2\sum_{j<k} a_j a_k \times c_{jk}$$

Variance / Covariance Matrix

Similar to linear regression we obtain an estimate of the regression coefficient and standard errors for these regression coefficients.

More generally we obtain a <u>covariance</u> matrix:

$$\widehat{V}(\widehat{\boldsymbol{\beta}}) = \begin{bmatrix} s_1^2 & c_{12} & c_{13} & \dots & c_{1p} \\ c_{21} & s_2^2 & c_{23} & \dots & c_{2p} \\ c_{31} & c_{32} & s_3^2 & \dots & c_{3p} \\ \vdots & & & & \\ c_{p1} & c_{p2}^2 & c_{p3} & \dots & s_p^2 \end{bmatrix}$$

where

$$\begin{aligned} s_j^2 &= \text{the estimated variance for } \widehat{\beta}_j \\ c_{jk} &= \text{the estimated covariance between } \widehat{\beta}_j \text{ and } \widehat{\beta}_k \end{aligned}$$

Logistic Regression

- Discussion / practice with general logistic model.
- Example relating logistic regression to 2×2 table analysis (binary exposure).
- Example relating logistic regression to stratified 2×2 table analysis (binary exposure, categorical confounder).
- Example discussing effect modification (interaction) with a binary exposure and a binary effect modifier.

Q: What about a more general treatment of exposures, confounders, and/or effect modifiers?

Confounders:

$\circ$ Several levels of a qualitative variable (unordered categories)

- stratified adjustment with dummy variables

$$X_j(2) = \begin{cases} 0 & \text{stratum 1} \\ 1 & \text{stratum 2} \\ 0 & \text{stratum 3} \end{cases} \qquad X_j(3) = \begin{cases} 0 & \text{stratum 1} \\ 0 & \text{stratum 2} \\ 1 & \text{stratum 3} \end{cases}$$

$\circ$ Several levels of an ordinal variable (ordered categories)

- stratified adjustment with dummy variables
- linear adjustment with an assigned score

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

- linear + quadratic or more complicated adjustment using score

Confounders: (continued)

$\circ$ Quantitative (continuous) variable.

- group and then treat as an ordinal variable

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

- linear adjustment in in measured values
- quadratic or more complicated adjustment (nuisance?)

Example

Outcome: oesophageal cancer

Exposure:

$$\text{NEWTOB} = \begin{cases} 0 & \text{tobacco use is} <10 \text{ g/day} \\ 1 & \text{tobacco use is} \geq 10 \text{ g/day} \end{cases}$$

Confounder:

$$\text{AGE} = \begin{cases} 1 & \text{25-44 years} \\ 2 & \text{45-54 years} \\ 3 & \text{55-64 years} \\ 4 & \text{65-74 years} \\ 5 & \text{75+ years} \end{cases}$$


```
. cc y newtob, by(newage)

          newage |       OR       [95% Conf. Interval]   M-H Weight
-----------------+-------------------------------------------------
           25-44 |      5.664     1.332866          .      .8012821 (Cornfield)
           45-54 |   2.671614     1.339354   5.321494      5.061033 (Cornfield)
           55-64 |   2.536216     1.441307   4.460941      7.644628 (Cornfield)
           65-74 |   1.385399     .7159378   2.682346       7.31677 (Cornfield)
             75+ |   2.121212     .5892296   7.650982           1.5 (Cornfield)
-----------------+-------------------------------------------------
           Crude |   2.151243     1.566262   2.954605               (Cornfield)
    M-H combined |   2.274106     1.617971   3.196322
-----------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(4) =     3.80  Pr>chi2 = 0.4341

                   Test that combined OR = 1:
                                Mantel-Haenszel chi2(1) =     22.86
                                                Pr>chi2 =    0.0000
```

Example

Model:

$$\text{logit}[\pi(\boldsymbol{X})] = \alpha + \beta_1 \text{TOB} + \sum_{j=2}^{5} \beta_j \text{AGE}(j)$$

logit structure:

	TOB=0	TOB=1
25-44 years	α	$\alpha + \beta_1$
45-54 years	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2$
55-64 years	$\alpha + \beta_3$	$\alpha + \beta_1 + \beta_3$
65-74 years	$\alpha + \beta_4$	$\alpha + \beta_1 + \beta_4$
75+ years	$\alpha + \beta_5$	$\alpha + \beta_1 + \beta_5$

Example

Model:

$$\text{logit}[\pi(\boldsymbol{X})] = \alpha + \beta_1 \text{TOB} + \beta_2 \text{AGE}$$

logit structure:

	TOB=0	TOB=1
25-44 years	$\alpha + (1)\beta_2$	$\alpha + \beta_1 + (1)\beta_2$
45-54 years	$\alpha + (2)\beta_2$	$\alpha + \beta_1 + (2)\beta_2$
55-64 years	$\alpha + (3)\beta_2$	$\alpha + \beta_1 + (3)\beta_2$
65-74 years	$\alpha + (4)\beta_2$	$\alpha + \beta_1 + (4)\beta_2$
75+ years	$\alpha + (5)\beta_2$	$\alpha + \beta_1 + (5)\beta_2$

```
.
. logistic y newtob

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(1)      =      22.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -482.62705                       Pseudo R2       =     0.0231

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   2.151243   .3490804      4.721   0.000       1.565185     2.95674
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(1)      =      22.85
                                                  Prob > chi2     =     0.0000
Log likelihood = -482.62705                       Pseudo R2       =     0.0231

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   .7660456   .1622692      4.721   0.000       .4480039    1.084087
   _cons |   -1.74585   .1227096    -14.227   0.000      -1.986356   -1.505343
------------------------------------------------------------------------------

. lrtest, saving(0)
```

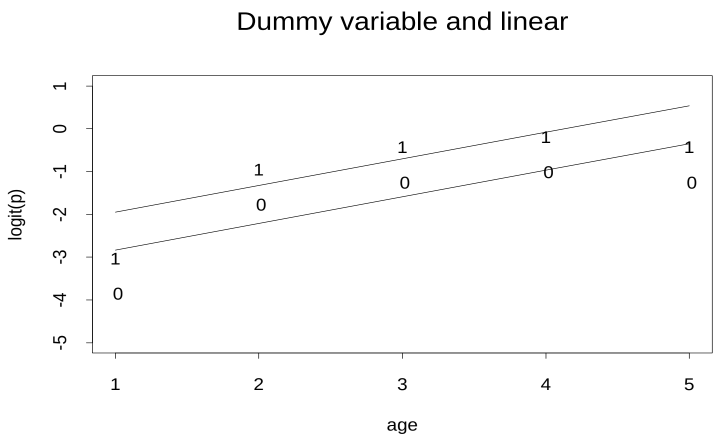


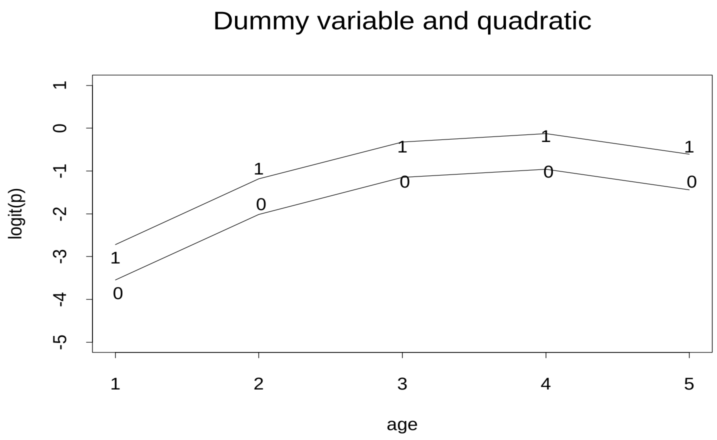
```
.
. xi: logistic y newtob i.newage
i.newage             Inewag_1-5   (naturally coded; Inewag_1 omitted)

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(5)      =     139.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -424.50072                       Pseudo R2       =     0.1408

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   2.269541    .392561      4.738   0.000       1.616986    3.185443
Inewag_2 |   7.993391   2.907852      5.714   0.000       3.918134    16.30733
Inewag_3 |   13.44772   4.730177      7.388   0.000       6.749043    26.79508
Inewag_4 |   17.11149   6.250276      7.774   0.000       8.363174    35.01098
Inewag_5 |   13.44159   6.267469      5.573   0.000        5.38961     33.5231
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(5)      =     139.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -424.50072                       Pseudo R2       =     0.1408

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   .8195776   .1729693      4.738   0.000        .480564    1.158591
Inewag_2 |   2.078615    .363782      5.714   0.000       1.365616    2.791615
Inewag_3 |    2.59881   .3517457      7.388   0.000       1.909401    3.288218
Inewag_4 |    2.83975   .3652679      7.774   0.000       2.123838    3.555662
Inewag_5 |   2.598354   .4662743      5.573   0.000       1.684473    3.512235
   _cons |  -3.835052   .3393721    -11.300   0.000      -4.500209   -3.169895
------------------------------------------------------------------------------

. lrtest, saving(1)
```

```
.
. logistic y newtob newage

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(2)      =     105.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -441.20282                       Pseudo R2       =     0.1070

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   2.428903   .4174832      5.163   0.000       1.734221    3.401856
  newage |   1.862119    .133999      8.640   0.000       1.617166    2.144176
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(2)      =     105.70
                                                  Prob > chi2     =     0.0000
Log likelihood = -441.20282                       Pseudo R2       =     0.1070

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   .8874398   .1718814      5.163   0.000        .5505585    1.224321
  newage |   .6217153   .0719605      8.640   0.000        .4806754    .7627552
   _cons |  -3.455834    .255333    -13.535   0.000       -3.956277    -2.95539
------------------------------------------------------------------------------

. lrtest, saving(2)
```



```
.
. generate newage2 = newage*newage

.
. logistic y newtob newage newage2

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(3)      =     135.27
                                                  Prob > chi2     =     0.0000
Log likelihood = -426.41899                       Pseudo R2       =     0.1369

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |    2.29559   .3965052      4.811   0.000       1.636329     3.22046
  newage |   12.72331   4.961669      6.522   0.000       5.924606    27.32376
 newage2 |    .714783   .0468169     -5.127   0.000       .6286691    .8126926
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(3)      =     135.27
                                                  Prob > chi2     =     0.0000
Log likelihood = -426.41899                       Pseudo R2       =     0.1369

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   .8309898   .1727248      4.811   0.000        .4924555    1.169524
  newage |   2.543435    .389967      6.522   0.000        1.779114    3.307757
 newage2 |  -.3357763   .0654981     -5.127   0.000       -.4641502   -.2074024
   _cons |  -5.755692   .5608276    -10.263   0.000       -6.854894    -4.65649
------------------------------------------------------------------------------

. lrtest, saving(3)

.
. lrtest, using(1) model(2)
Logistic:  likelihood-ratio test                      chi2(3)     =      33.40
                                                      Prob > chi2 =     0.0000

. lrtest, using(1) model(3)
Logistic:  likelihood-ratio test                      chi2(2)     =       3.84
                                                      Prob > chi2 =     0.1469
```





Logistic Regression: More Complicated Models

Exposures: (primary predictors)

○ Several levels of an ordinal variable.

- dummy variables for the different levels comparing each level to the lowest (reference) level.
- linear term in exposure score – implies linear increase in log odds ratio comparing category j to the lowest – implies multiplicative increases in the odds ratio comparing category j to the lowest.
- quadratic or more complicated relationship

○ Quantitative (continuous) variable.

- group and treat as ordinal exposure

$$\text{ALC} = \begin{cases} 1 & <40 \text{ g/day} \\ 2 & 40\text{-}79 \text{ g/day} \\ 3 & 80+ \text{ g/day} \end{cases}$$

- linear term in exposure
- quadratic or more general form of dose-response

Logistic Regression: More Complicated Models

Effect Modifiers:

$\boxed{\circ}$ Several levels of a qualitative variable

- dummy variable adjustments and interaction terms

$\boxed{\circ}$ Several levels of a quantitative variable

- dummy variable adjustments and interaction terms
- linear or more complicated interaction terms.
  NOTE: adjustment always at least as thorough as interaction model.

Example

Model:

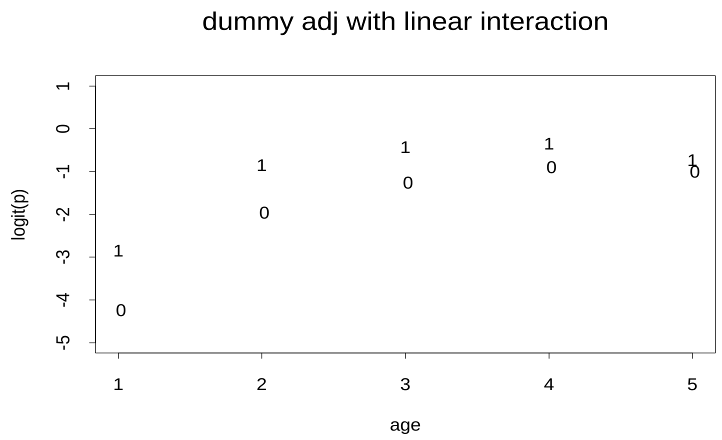
$$\text{logit}[\pi(\boldsymbol{X})] = \alpha + \beta_1 \text{TOB} + \sum_{j=2}^{5} \beta_j \text{AGE}(j) + \beta_6 \text{ TOB} \cdot \text{AGE}$$

- logit structure:

	TOB=0	TOB=1
25-44 years	α	$\alpha + \beta_1 + (1)\beta_6$
45-54 years	$\alpha + \beta_2$	$\alpha + \beta_1 + \beta_2 + (2)\beta_6$
55-64 years	$\alpha + \beta_3$	$\alpha + \beta_1 + \beta_3 + (3)\beta_6$
65-74 years	$\alpha + \beta_4$	$\alpha + \beta_1 + \beta_4 + (4)\beta_6$
75+ years	$\alpha + \beta_5$	$\alpha + \beta_1 + \beta_5 + (5)\beta_6$

Q: What is the disease odds ratio comparing TOB=1 to TOB=0 for AGE category j?

```
.
. generate newtage = newtob*newage

.
. xi: logistic y newtob i.newage newtage
i.newage            Inewag_1-5   (naturally coded; Inewag_1 omitted)

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(6)      =     141.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -423.1915                       Pseudo R2       =     0.1434

------------------------------------------------------------------------------
       y | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |    5.29514   2.978646      2.963   0.003       1.758149    15.94774
Inewag_2 |   9.729572   3.811189      5.808   0.000       4.515107    20.96619
Inewag_3 |   19.62995   8.588636      6.804   0.000       8.327128    46.27463
Inewag_4 |   28.22598   14.02035      6.725   0.000       10.66209     74.7233
Inewag_5 |   25.42508    15.7996      5.207   0.000       7.521601    85.94373
 newtage |   .7561485   .1322673     -1.598   0.110       .5366769    1.065372
------------------------------------------------------------------------------

. logit

Logit estimates                                   Number of obs   =        972
                                                  LR chi2(6)      =     141.73
                                                  Prob > chi2     =     0.0000
Log likelihood =  -423.1915                       Pseudo R2       =     0.1434

------------------------------------------------------------------------------
       y |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  newtob |   1.666789   .5625245      2.963   0.003       .5642616    2.769317
Inewag_2 |    2.27517   .3917118      5.808   0.000       1.507429    3.042911
Inewag_3 |   2.977056   .4375272      6.804   0.000       2.119519    3.834594
Inewag_4 |   3.340243   .4967178      6.725   0.000       2.366694    4.313792
Inewag_5 |   3.235736    .621418      5.207   0.000       2.017779    4.453693
 newtage |  -.2795174   .1749224     -1.598   0.110       -.622359    .0633241
   _cons |  -4.218451   .4364217     -9.666   0.000      -5.073822   -3.363081
------------------------------------------------------------------------------
```

Recall:

- In case-control studies with takes cases ($\boldsymbol{Y} = 1$) and controls ($\boldsymbol{Y} = 0$) and then ascertain covariates ($\boldsymbol{X}$).

- From this study design we can't estimate disease risk, $P(\boldsymbol{Y} = 1 \mid \boldsymbol{X})$ nor relative risk, but can estimate exposure odds ratios.

- Exposure odds ratios were shown to equal 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 we do any regression modeling using data from this study design?

A: Yes. We can use standard logistic regression to estimate odds ratios but not disease risk probabilities.

- Prentice and Pike (1979)
- Breslow and Day (1981)

Justification:

Let $\boldsymbol{X}$ denote covariates, $\boldsymbol{Y}$ case/control status, and Z be an indicator of whether a subject is sampled ($Z = 1$ if selected, and 0 otherwise).

In case-control studies we typically sample cases and controls with different probabilities:

$$
\begin{aligned}
P(Z = 1 \mid \boldsymbol{Y} = 0) &= p_0 \\
P(Z = 1 \mid \boldsymbol{Y} = 1) &= p_1
\end{aligned}
$$

(*)Logistic Regression: Case-Control Analysis

There are two important (mathematical) results that allow logistic regression to work in this design:

$\boxed{1.}$ If we make no assumptions regarding the distribution of $\boldsymbol{X}$ then we can focus only on the contribution $P(\boldsymbol{Y} \mid \boldsymbol{X})$.

$$
\begin{aligned}
\text{case-control likelihood} &: P(\boldsymbol{X} \mid \boldsymbol{Y}) \\
P(\boldsymbol{X} \mid \boldsymbol{Y}) &= P(\boldsymbol{Y} \mid \boldsymbol{X}) \frac{P(\boldsymbol{X})}{P(\boldsymbol{Y})}
\end{aligned}
$$

$\boxed{2.}$ If we assume that sampling of cases and controls is independent of the covariates X,

$$P(Z = 1 \mid \boldsymbol{Y}, \boldsymbol{X}) = P(Z = 1 \mid \boldsymbol{Y})$$

then using the likelihood function for the sampled data, $P(\boldsymbol{Y} \mid \boldsymbol{X}, Z = 1)$ allows estimation of logistic regression coefficients although the intercept term is impacted.

(*)Logistic Regression: Case-Control Analysis

Q: How does the likelihood function for the sampled data compare to a prospective data likelihood function?

$$P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}, Z = 1) \quad ? \quad P(\boldsymbol{Y} = 1 \mid \boldsymbol{X})$$

$$
\begin{aligned}
P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}, Z = 1) &= \frac{P(\boldsymbol{Y} = 1, Z = 1 \mid \boldsymbol{X})}{P(Z = 1 \mid \boldsymbol{X})} \\
\text{and} & \\
P(\boldsymbol{Y} = 1, Z = 1 \mid \boldsymbol{X}) &= P(Z = 1 \mid \boldsymbol{Y} = 1, \boldsymbol{X}) P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}) \\
P(Z = 1 \mid \boldsymbol{X}) &= P(Z = 1 \mid \boldsymbol{Y} = 0, \boldsymbol{X}) P(\boldsymbol{Y} = 0 \mid \boldsymbol{X}) + \\
& \quad P(Z = 1 \mid \boldsymbol{Y} = 1, \boldsymbol{X}) P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}) \\
\text{so that} & \\
P(\boldsymbol{Y} = 1 \mid \boldsymbol{X}, Z = 1) &= \frac{p_1 P(\boldsymbol{Y} = 1 \mid \boldsymbol{X})}{p_0 P(\boldsymbol{Y} = 0 \mid \boldsymbol{X}) + p_1 P(\boldsymbol{Y} = 1 \mid \boldsymbol{X})} \\
&= \frac{p_1 \exp(\beta_0 + \beta_1 \boldsymbol{X})}{p_0 + p_1 \exp(\beta_0 + \beta_1 \boldsymbol{X})} \\
&= \frac{\exp(\alpha + \beta_1 \boldsymbol{X})}{1 + \exp(\alpha + \beta_1 \boldsymbol{X})}
\end{aligned}
$$

where the intercept $\alpha = \beta_0 + \log(p_1/p_0)$.

Logistic Regression: Case-Control Analysis

Summary:

1. We can "pretend" that the case-control data was collected prospectively and use standard logistic regression to estimate regression coefficients (β_j for $j \neq 0$) and obtain standard errors.

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

3. We need to be careful <u>not</u> to use the fitted logistic regression model to estimate risk probabilities since the intercept that we estimate is incorrect: $\alpha = \beta_0 + \log(p_1/p_0)$.

Logistic Regression - Case Study

- These data come from Appendix 1 of Hosmer and Lemeshow (1989), and were collected at Baystate Medical Center, Springfield MA, during 1986.

- Low birth weight is an outcome that has been of concern to physicians for years. This is due to the fact that infant mortality rates and birth defect rates are very high for low birth weight babies. A woman's behavior during pregnancy (including diet, smoking habits, and receiving prenatal care) can greatly alter the chances of carrying the baby to term and, consequently, of delivering a baby of normal birth weight.

Q: What is the contribution of maternal smoking to the likelihood of having a child born with low birth weight?

Data

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, 1 = 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


```
. infile id lbw age lwt race smoke ptl hyper urirr pvft weight using lowbwt.dat
(189 observations read)

*** NAMES and LABELS not shown ***
.
. 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
.
. generate newpvft=pvft
. recode newpvft 2/max=2
. label define pgps 0 "0" 1 "1" 2 "2+"
. label values newpvft pgps
.
. tabulate agecat lbw, chi2 row

           |  Low birth weight
    agecat |  >=2500g    < 2500g |     Total
-----------+----------------------+----------
       <20 |        36         15 |        51
           |     70.59      29.41 |    100.00
-----------+----------------------+----------
     20-24 |        44         25 |        69
           |     63.77      36.23 |    100.00
-----------+----------------------+----------
     25-29 |        27         15 |        42
           |     64.29      35.71 |    100.00
-----------+----------------------+----------
       30+ |        23          4 |        27
           |     85.19      14.81 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(3) =   4.6641   Pr = 0.198
```

```
. tabulate wcat lbw, chi2 row

           |   Low birth weight
      wcat |  >=2500g    < 2500g |     Total
-----------+----------------------+----------
      <105 |        17         20 |        37
           |     45.95      54.05 |    100.00
-----------+----------------------+----------
   106-120 |        41         14 |        55
           |     74.55      25.45 |    100.00
-----------+----------------------+----------
   121-130 |        20         11 |        31
           |     64.52      35.48 |    100.00
-----------+----------------------+----------
   131-150 |        23          7 |        30
           |     76.67      23.33 |    100.00
-----------+----------------------+----------
      151+ |        29          7 |        36
           |     80.56      19.44 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(4) =  13.2923   Pr = 0.010

. tabulate race lbw, chi2 row

           |   Low birth weight
      Race |  >=2500g    < 2500g |     Total
-----------+----------------------+----------
     white |        73         23 |        96
           |     76.04      23.96 |    100.00
-----------+----------------------+----------
     black |        15         11 |        26
           |     57.69      42.31 |    100.00
-----------+----------------------+----------
     other |        42         25 |        67
           |     62.69      37.31 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

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



```
. tabulate smoke lbw, chi2 row
           |   Low birth weight
     Smoke |  >=2500g    < 2500g |     Total
-----------+----------------------+----------
        No |        86         29 |       115
           |     74.78      25.22 |    100.00
-----------+----------------------+----------
       Yes |        44         30 |        74
           |     59.46      40.54 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(1) =   4.9237   Pr = 0.026

. tabulate anyptl lbw, chi2 row
           |   Low birth weight
    anyptl |  >=2500g    < 2500g |     Total
-----------+----------------------+----------
         0 |       118         41 |       159
           |     74.21      25.79 |    100.00
-----------+----------------------+----------
        1+ |        12         18 |        30
           |     40.00      60.00 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(1) =  13.7590   Pr = 0.000

. tabulate hyper lbw, chi2 row
History of |
hypertensi |   Low birth weight
        on |  >=2500g    < 2500g |     Total
-----------+----------------------+----------
        No |       125         52 |       177
           |     70.62      29.38 |    100.00
-----------+----------------------+----------
       Yes |         5          7 |        12
           |     41.67      58.33 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(1) =   4.3880   Pr = 0.036
```

```
. tabulate urirr lbw, chi2 row

  Presence |
of uterine |
irritabili |   Low birth weight
        ty |   >=2500g    < 2500g |     Total
-----------+----------------------+----------
        No |       116         45 |       161
           |     72.05      27.95 |    100.00
-----------+----------------------+----------
       Yes |        14         14 |        28
           |     50.00      50.00 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(1) =   5.4008   Pr = 0.020

.
. tabulate newpvft lbw, chi2 row

           |   Low birth weight
   newpvft |   >=2500g    < 2500g |     Total
-----------+----------------------+----------
         0 |        64         36 |       100
           |     64.00      36.00 |    100.00
-----------+----------------------+----------
         1 |        36         11 |        47
           |     76.60      23.40 |    100.00
-----------+----------------------+----------
        2+ |        30         12 |        42
           |     71.43      28.57 |    100.00
-----------+----------------------+----------
     Total |       130         59 |       189
           |     68.78      31.22 |    100.00

          Pearson chi2(2) =   2.5384   Pr = 0.281
```



```
.
. tabulate agecat smoke, chi2 row

           |         Smoke
    agecat |        No        Yes |     Total
-----------+----------------------+----------
       <20 |        28         23 |        51
           |     54.90      45.10 |    100.00
-----------+----------------------+----------
     20-24 |        44         25 |        69
           |     63.77      36.23 |    100.00
-----------+----------------------+----------
     25-29 |        26         16 |        42
           |     61.90      38.10 |    100.00
-----------+----------------------+----------
       30+ |        17         10 |        27
           |     62.96      37.04 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

          Pearson chi2(3) =   1.0742   Pr = 0.783

. tabulate wcat smoke, chi2 row

           |         Smoke
      wcat |        No        Yes |     Total
-----------+----------------------+----------
      <105 |        19         18 |        37
           |     51.35      48.65 |    100.00
-----------+----------------------+----------
   106-120 |        35         20 |        55
           |     63.64      36.36 |    100.00
-----------+----------------------+----------
   121-130 |        18         13 |        31
           |     58.06      41.94 |    100.00
-----------+----------------------+----------
   131-150 |        19         11 |        30
           |     63.33      36.67 |    100.00
-----------+----------------------+----------
      151+ |        24         12 |        36
           |     66.67      33.33 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

          Pearson chi2(4) =   2.2704   Pr = 0.686
```

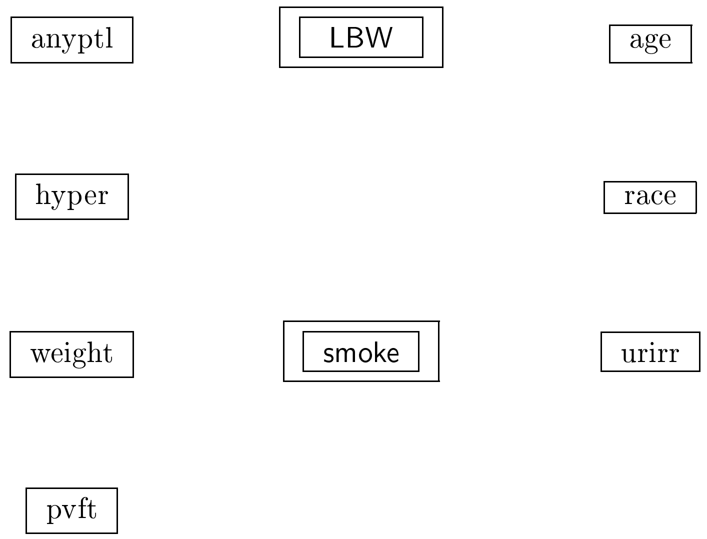
```
. tabulate race smoke, chi2 row

           |        Smoke
      Race |        No        Yes |     Total
-----------+----------------------+----------
     white |        44         52 |        96
           |     45.83      54.17 |    100.00
-----------+----------------------+----------
     black |        16         10 |        26
           |     61.54      38.46 |    100.00
-----------+----------------------+----------
     other |        55         12 |        67
           |     82.09      17.91 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

          Pearson chi2(2) =  21.7790   Pr = 0.000

. tabulate anyptl smoke, chi2 row
           |        Smoke
    anyptl |        No        Yes |     Total
-----------+----------------------+----------
         0 |       103         56 |       159
           |     64.78      35.22 |    100.00
-----------+----------------------+----------
        1+ |        12         18 |        30
           |     40.00      60.00 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

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



```
. tabulate hyper smoke, chi2 row

History of |
hypertensi |        Smoke
        on |        No        Yes |     Total
-----------+----------------------+----------
        No |       108         69 |       177
           |     61.02      38.98 |    100.00
-----------+----------------------+----------
       Yes |         7          5 |        12
           |     58.33      41.67 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

          Pearson chi2(1) =   0.0340   Pr = 0.854

. tabulate urirr smoke, chi2 row

  Presence |
of uterine |
irritabili |        Smoke
        ty |        No        Yes |     Total
-----------+----------------------+----------
        No |       100         61 |       161
           |     62.11      37.89 |    100.00
-----------+----------------------+----------
       Yes |        15         13 |        28
           |     53.57      46.43 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

          Pearson chi2(1) =   0.7302   Pr = 0.393
```

```
. tabulate newpvft smoke, chi2 row

           |        Smoke
   newpvft |        No        Yes |     Total
-----------+----------------------+----------
         0 |        55         45 |       100
           |     55.00      45.00 |    100.00
-----------+----------------------+----------
         1 |        35         12 |        47
           |     74.47      25.53 |    100.00
-----------+----------------------+----------
        2+ |        25         17 |        42
           |     59.52      40.48 |    100.00
-----------+----------------------+----------
     Total |       115         74 |       189
           |     60.85      39.15 |    100.00

          Pearson chi2(2) =   5.1262   Pr = 0.077
```


Summarize the EDA



```
. infile id lbw age lwt race smoke ptl hyper urirr pvft weight using lowbwt.dat
.

*** VARIABLE NAMES, LABELS, and RECODING ***


. logistic lbw smoke

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

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .7040592   .3196386      2.203   0.028        .0775791   1.330539
   _cons |  -1.087051   .2147299     -5.062   0.000       -1.507914  -.6661886
------------------------------------------------------------------------------

. lrtest, saving(1)

.
. logistic lbw smoke anyptl

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(2)      =      15.34
                                                  Prob > chi2     =     0.0005
Log likelihood = -109.66445                       Pseudo R2       =     0.0654

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .5364762   .3335101      1.609   0.108       -.1171916   1.190144
  anyptl |   1.351661   .4214984      3.207   0.001         .525539   2.177782
   _cons |  -1.261694   .2280439     -5.533   0.000       -1.708652  -.8147365
------------------------------------------------------------------------------

. lrtest, saving(2)
```



```
.
. xi:  logistic lbw smoke anyptl i.race
i.race            Irace_1-3    (naturally coded; Irace_1 omitted)

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(4)      =      23.77
                                                  Prob > chi2     =     0.0001
Log likelihood = -105.45115                       Pseudo R2       =     0.1013

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .9326674   .3830806      2.435   0.015        .1818433   1.683492
  anyptl |   1.284558   .4305061      2.984   0.003        .4407818   2.128335
 Irace_2 |   1.092107   .5003836      2.183   0.029         .111373   2.072841
 Irace_3 |   1.031102   .4123075      2.501   0.012        .2229943    1.83921
   _cons |  -1.970956   .3670825     -5.369   0.000       -2.690425  -1.251488
------------------------------------------------------------------------------

. lrtest, saving(3)

.
. xi:  logistic lbw smoke anyptl i.race i.newpvft
i.race            Irace_1-3    (naturally coded; Irace_1 omitted)
i.newpvft         Inewpv_0-2   (naturally coded; Inewpv_0 omitted)

                                                  LR chi2(6)      =      25.20
                                                  Prob > chi2     =     0.0003
Log likelihood = -104.73508                       Pseudo R2       =     0.1074

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .8234415   .3953754      2.083   0.037          .04852   1.598363
  anyptl |   1.395546   .4483789      3.112   0.002        .5167394   2.274352
 Irace_2 |   1.049959   .5019423      2.092   0.036        .0661704   2.033748
 Irace_3 |   .9315888    .423713      2.199   0.028        .1011265   1.762051
Inewpv_1 |  -.5371209   .4567105     -1.176   0.240       -1.432257   .3580153
Inewpv_2 |  -.1390592   .4256898     -0.327   0.744       -.9733959   .6952776
   _cons |  -1.750285   .4229536     -4.138   0.000       -2.579259   -.921311
------------------------------------------------------------------------------

. lrtest, saving(4)
```

```
.
. xi:  logistic lbw smoke anyptl i.race i.newpvft hyper
i.race              Irace_1-3     (naturally coded; Irace_1 omitted)
i.newpvft           Inewpv_0-2    (naturally coded; Inewpv_0 omitted)

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(7)      =      28.86
                                                  Prob > chi2     =     0.0002
Log likelihood = -102.90493                       Pseudo R2       =     0.1230

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .8324501   .3999549      2.081   0.037       .048553    1.616347
  anyptl |   1.417874    .452653      3.132   0.002      .5306901    2.305057
 Irace_2 |   .9856591    .509341      1.935   0.053      -.012631    1.983949
 Irace_3 |   .9482569   .4296207      2.207   0.027      .1062159    1.790298
Inewpv_1 |  -.5214505    .462202     -1.128   0.259      -1.42735    .3844489
Inewpv_2 |  -.0592196   .4303705     -0.138   0.891     -.9027304    .7842911
   hyper |   1.230846   .6480027      1.899   0.058     -.0392163    2.500908
   _cons |  -1.868184    .436218     -4.283   0.000     -2.723155   -1.013212
------------------------------------------------------------------------------
. lrtest, saving(5)

. xi:  logistic lbw smoke anyptl i.race i.newpvft hyper urirr
i.race              Irace_1-3     (naturally coded; Irace_1 omitted)
i.newpvft           Inewpv_0-2    (naturally coded; Inewpv_0 omitted)

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(8)      =      31.91
                                                  Prob > chi2     =     0.0001
Log likelihood =  -101.3815                       Pseudo R2       =     0.1360

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .8361867   .4045585      2.067   0.039      .0432666    1.629107
  anyptl |   1.298265   .4592108      2.827   0.005      .3982284    2.198302
 Irace_2 |   1.017674      .5096      1.997   0.046      .0188761    2.016471
 Irace_3 |   .9426682    .435952      2.162   0.031       .088218    1.797118
Inewpv_1 |  -.4606427   .4677612     -0.985   0.325     -1.377438    .4561525
Inewpv_2 |  -.0271993   .4337993     -0.063   0.950     -.8774303    .8230318
   hyper |   1.370196   .6512502      2.104   0.035      .0937687    2.646623
   urirr |   .8041039   .4562035      1.763   0.078     -.0900385    1.698246
   _cons |  -2.018308   .4551316     -4.435   0.000      -2.91035   -1.126267
------------------------------------------------------------------------------
. lrtest, saving(6)
```



```
.
. xi:  logistic lbw smoke anyptl i.race i.newpvft hyper urirr i.agecat
i.race              Irace_1-3     (naturally coded; Irace_1 omitted)
i.newpvft           Inewpv_0-2    (naturally coded; Inewpv_0 omitted)
i.agecat            Iageca_1-4    (naturally coded; Iageca_1 omitted)

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(11)     =      36.16
                                                  Prob > chi2     =     0.0002
Log likelihood = -99.255578                       Pseudo R2       =     0.1541

------------------------------------------------------------------------------
     lbw |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   .8426154   .4094339      2.058   0.040      .0401398    1.645091
  anyptl |   1.398253   .4823508      2.899   0.004      .4528629    2.343643
 Irace_2 |   .9631186   .5106114      1.886   0.059     -.0376613    1.963899
 Irace_3 |    .864313   .4455825      1.940   0.052     -.0090126    1.737639
Inewpv_1 |  -.4412121   .4806857     -0.918   0.359     -1.383339    .5009146
Inewpv_2 |     .10775   .4530354      0.238   0.812      -.780183    .9956831
   hyper |   1.352725   .6502511      2.080   0.037      .0782565    2.627194
   urirr |   .7519809   .4646972      1.618   0.106     -.1588088    1.662771
Iageca_2 |   .2761735     .44373      0.622   0.534     -.5935212    1.145868
Iageca_3 |   .2839491   .5012037      0.567   0.571     -.6983921     1.26629
Iageca_4 |  -.9558746    .720959     -1.326   0.185     -2.368928    .4571791
   _cons |  -2.093934   .5209528     -4.019   0.000     -3.114983   -1.072885
------------------------------------------------------------------------------

. lrtest, saving(7)
```

```
.
.
. lrtest, using(7) model(6)
Logistic:  likelihood-ratio test                        chi2(3)     =       4.25
                                                        Prob > chi2 =     0.2355

. lrtest, using(6) model(5)
Logistic:  likelihood-ratio test                        chi2(1)     =       3.05
                                                        Prob > chi2 =     0.0809

. lrtest, using(5) model(4)
Logistic:  likelihood-ratio test                        chi2(1)     =       3.66
                                                        Prob > chi2 =     0.0557

. lrtest, using(4) model(3)
Logistic:  likelihood-ratio test                        chi2(2)     =       1.43
                                                        Prob > chi2 =     0.4887


*** ODDS RATIOS  ***

Logit estimates                                   Number of obs   =        189
                                                  LR chi2(11)     =      36.16
                                                  Prob > chi2     =     0.0002
Log likelihood = -99.255578                       Pseudo R2       =     0.1541

------------------------------------------------------------------------------
     lbw | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   smoke |   2.322433   .9508828      2.058   0.040       1.040956    5.181482
  anyptl |   4.048122   1.952615      2.899   0.004       1.572809    10.41913
 Irace_2 |   2.619854   1.337727      1.886   0.059       .9630391    7.127058
 Irace_3 |   2.373375   1.057534      1.940   0.052       .9910279    5.683905
Inewpv_1 |   .6432563   .3092041     -0.918   0.359         .25074     1.65023
Inewpv_2 |   1.113769   .5045769      0.238   0.812       .4583221    2.706573
   hyper |   3.867952    2.51514      2.080   0.037         1.0814     13.8349
   urirr |   2.121198   .9857146      1.618   0.106       .8531595    5.273903
Iageca_2 |   1.318077   .5848701      0.622   0.534       .5523788    3.145171
Iageca_3 |   1.328365   .6657816      0.567   0.571       .4973844    3.547667
Iageca_4 |   .3844757   .2771913     -1.326   0.185        .093581    1.579612
------------------------------------------------------------------------------
```


Regression Analysis & Model Building

Q: What are the goals of regression analysis?

A: Estimation, Testing, & 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.
  - ⋆ Remove bias.

- Testing whether variables are associated with the response.

- Prediction of a response variable given a collection of covariates.

Regression Analysis

- First step is to identify the scientific question. Apriori consideration of the specific goals of analysis is crucial!

Classification of Variables:

∘ Response variable.

⋆ Dependent variable.
⋆ Outcome variable.

∘ Predictor of interest.

⋆ Exposure variable.
⋆ Treatment assignment.

∘ Confounding variables.

⋆ Associated with response and POI.
⋆ Not intermediate.

∘ Precision variables.

⋆ Associated with response and not POI.
⋆ Reduces response incertainty.

Regression Analysis

Example: Impact of maternal smoking on low birth weight?

- Measured variables:

```
birth weight (g)
maternal smoking (yes/no)
maternal age (yrs)
maternal weight at last menses (kg)
race (1=white, 2=black, 3=other)
history of premature labor (0=no, 1=yes)
history of hypertension (0=no, 1=yes)
```

Response:

Predictor of Interest:

Confounder:

Precision:

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 <u>not</u> provided:
1. how variables chosen / selected
2. how effect modifiers assessed
3. how confounders assessed

Model Goals

Main Model Goals:

- Valid "effect" estimation (exposure $\rightarrow$ disease)
- Good prediction of the outcome
- Parsimonious description of correlations / associations

Classification of Analysis:

- Confimatory Data Analysis (CDA)
  - Formal hypothesis testing
  - Protocol contains analysis plans
- Exploratory Data Analysis (EDA)
  - Hypothesis generating
  - Write / create analysis plans
  - Confirmatory studies to follow

Alternative Modeling Goals

1. Valid "effect" estimation (exposure $\rightarrow$ disease)
   - Logistic regression for **inference**
   - Confirmatory data analysis (CDA)
     - Formal hypothesis testing
     - Protocol contains analysis plans
2. Good prediction of the outcome (screening)
   - Logistic regression for **prediction**
   - Little/no attempt to interpret parameters (black box)
   - Tradeoff between *bias* (from too small a model) and *variance* (from too large a model)
   - Role for "neural networks"?
3. Data mining
   - Exploratory Data Analysis (EDA)
   - Hypothesis generating
   - Write / create study design and protocol
   - Confirmatory studies to follow

Confirmatory Data Analysis

- Carefully decide the *goals*: scientific question to be answered

  Outcome: D

  Exposure of primary interest: E

  Control or confounding variables: restrict attention to clinically or biologically meaningful variables known *a priori* to be causally related to disease outcome
  - Literature review
  - Theoretical basis
- Decide how to use the variables
  - Carefully (and parsimoniously) model exposure effect
  - Use "rich" model for confounder control: no attempt to interpret parameters
  - What interactions will be considered?
    Structure interaction analyses *a priori* to detect most plausible form(s) of "effect modification"
  - Respect "marginality": don't include interactions without corresponding main effects

- Plan a short series of different analyses
  - Main effect of exposure adjusted for
    1. No confounders
    2. Primary confounders
    3. Primary and secondary confounders
  - Interaction of exposure effect (separately) with short list of potential effect modifiers
- Write the plan down and stick to it

CDA (continued)

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

Controversy remains regarding next step

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

"The object of stratification or covariance adjustment ... is ... to alleviate the distortion in the estimated effect of exposure caused by confounding or selection bias. If another factor is causally related to disease, and if there is a chance it could be correlated with the exposure of interest, statistical adjustment is needed to produce a valid estimate. Since one rarely has good prior information about the degree of association between various risk factors in the population, a reasonable and prudent policy is to take account in the analysis of all known causal factors regardless of whether they may appear to be related to the exposure of interest in the data at hand. Significance testing of their relationship to either disease or exposure is irrelevant to the issue of whether adjustment for such effects modifies the association of interest."

(Breslow, Ann. Rev. Publ. Health, 1982, p. 38-39)

Guidelines for CDA

⋆ Carefully decide the scientific question that you want answered.

Outcome: $\boldsymbol{D}$
Exposure of Interest: $\boldsymbol{E}$

Variable Specification

- Restrict attention to clinically or biologically meaningful variables.
  - study goals
  - literature review
  - theoretical basis
  - Define these variables as $C_1, C_2, \ldots, C_p$

Guidelines for CDA

- Decide how you will use the variables
  - Will you include interactions among these covariates?
  - Call the combinations V_j (ie. this may be just be the C's or may include some $C_j C_k$ terms)
- Decide what interactions between $\boldsymbol{E}$ and the V_j are to be considered.
  - In most cases the number of *apriori* interactions is limited.

Guidelines for CDA

- Make sure that your model is

  "Hierarchically Well-Formulated"

ie. all interactions have the lower order terms (main effects)

NOTE: Having completed these steps you will have identified the statistical model that is of interest. There is no data-dependent variable selection. The next task will be to decide if this complete model can be simplified.

Backward Elimination - Interaction

- Do you really need the $\boldsymbol{E}\, V_i\, V_j$ terms?
  - Use likelihood ratio tests
  - Use a "chunk test" – are all coefficients zero?
- Do you really need the $\boldsymbol{E}V_j$ terms (other than any required by 3-way interactions)?
  - Use likelihood ratio tests
  - Use a "chunk test" – are all coefficients zero?

Guidelines for CDA

Backward Elimination - Confounders

- Do you really need the V_j terms (other than any required by 2-way and/or 3-way interactions)?

  ◦ Primary goal of confounder adjustment is the removal of bias in the exposure estimate.

  ◦ Why consider any removal?
    A: Precision.

  ◦ Monitor changes in the effect measure (odds ratio).

  ◦ Which subsets can we use?

Guidelines for CDA

Kleinbaum (1994):
"In general, regardless of the number of V's in one's model, 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

Multiple Comparisons

- If we conduct enough significance tests then we are bound to find "a significant assocaiation".

Q: After trying a few (dozen?) regression models I finally settled on this last model – and here the p-value of `exposure` is 0.038 – what can I say?

Example:

Hilsenbeck, Clark, and McGuire
"Why do so many prognostic factors fail to pan out?"
Breast Cancer Research and Treatment
(1992) **22**: 197-206

Regression Models

"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 (1993)
Statistical Methods in Epidemiology
page 273

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 of 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 tracks but once the need for such discipline is recognized, there are significant advantages to the regression modelling approach."

Clayton and Hills (1993)
<u>Statistical Methods in Epidemiology</u>
page 273

Empirical Model Building

A 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.
   - Document pairwise associations:
     - correlations
     - mean differences with s.e.'s
     - log odds ratios with s.e.'s
   - Additional stratified analyses for key variables.

4. Regression models developed:
   - Main effects regression.
   - Forward selection for strong interactive effects.
   - Backward selection for stable effects.
   - Summary of standardized regression coefficient(s) for each omitted variable if added individually in order to reassure that no important further effects have been overlooked.
5. Presentation of model(s):
   - Coefficients and s.e.'s
   - Graphical display

Empirical Model Building

Model Development Cox and Wermuth (1996) section 7.3

1. Specify required explanatory variables:
   - Predictor of interest.
   - Known important variables (a priori).
2. Regress response on all explanatory variables considered. Include nonlinear terms (ie. `age`^{2}) that are *a priori* considered important.
3. Eliminate (individually) variables that have small standardized regression coefficients. Identify one or more simple, well-fitting models.
4. Examine, one term at a time, the impact of adding squared terms and cross-product terms into the model.

   Any significant terms uncovered in this way need detailed interpretation before we decide how to incorporate them into the "final" model.
5. List the contribution of omitted primary explanatory variables if added back to the model one at a time (individually) to check that no important effect has been overlooked.

Statistical Thinking

Breslow (1999)

"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 is 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."

Sources of Variation

Breslow (1999)

"When queried by students seeking a general recipe on how to apply statistical methods to any particular area of application, I usually punt. This, I explain, is part of the 'art' of statistics whose practice is perfected through immersion in the subject matter area, careful study of statistical applications in that area and, if one is lucky, through apprenticeship under a master practicioner."

"Once the study goals are defined and the response variables are identified, the next step is to consider sources of variation in those variables. This goes well beyond the identification of potential explanatory variables for inclusion in a regression equation. It is important to recognize that there may be multiple sources and multiple levels of unexplained (random) variation that could, or should, be taken into account."

Model Selection / Building Summary

$\boxed{\star}$ Determine the scientific question.

$\boxed{\star}$ Classify the measured variables.

$\boxed{\star}$ Thorough univariate and bivariate summaries.

$\boxed{\star}$ Choose an analysis plan.

- Confirmatory regression.
  formal analysis plan *a priori*
  limited number of models (three!)

- Exploratory / Empirical regression.
  more flexible
  have a strategy! (ie. reproducible analysis)

Regression and Predicted Values

$\boxed{\bullet}$ Logistic regression for inference with binary outcomes.

$\boxed{\bullet}$ Logistic regression for prediction of binary outcomes.

1. We can use logistic regression to obtain probabilities $\pi(\boldsymbol{X})$.
2. Assessment of Accuracy
   - Data to fit and check?
   - Sensitivity
   - Specificity
   - ROC curve
3. Examples
   - LBW Data
   - Duh et al. (1998) AJE
   - Gypsy moths

LBW Data

- Recall that we used a logistic regression model to assess the contribution of smoking toward the probability of a LBW outcome.

- Using this fitted logistic regression we obtain fitted probabilities, $\widehat{\pi}(\boldsymbol{X})$, that range from 3% to 89%.

```
   p.hat id y age mom.weight    race smoke ptl hyper urirr npvft   bw pvft
1 0.4064 85 0  19        182 : black     0   0     0     1     0 2523    0
2 0.1113 86 0  33        155 : other     0   0     0     0     3 2551   2+
3 0.1952 87 0  20        105 : white     1   0     0     0     1 2557    1
4 0.4712 88 0  21        108 : white     1   0     0     1     2 2594   2+
5 0.3777 89 0  18        107 : white     1   0     0     1     0 2600    0
   agecat any.ptl
1  : <=19       0
2   : 30+       0
3 : 20-24       0
4 : 20-24       0
5  : <=19       0
```


Predicted Values

These predicted values are fitted probabilities:

$$\widehat{P}[Y = 1 \mid \boldsymbol{X}] = \widehat{\pi}(\boldsymbol{X})$$

One approach to assessing the accuracy of these predicted values is to summarize the the distribution of the fitted values for the $Y = 0$ and the $Y = 1$ outcomes:

$$\text{sensitivity} \quad : \quad P[\text{Test} > c \mid Y = 1]$$

$$\text{specificity} \quad : \quad P[\text{Test} \leq c \mid Y = 0]$$

- We could consider the "test" to be a single predictor, X_j.

- We could consider the "test" to be the combination of predictors given by the fitted probability, $\widehat{\pi}(\boldsymbol{X})$.

ROC Curve

We can actually define a collection of "positive" and "negative" tests by considering a variety of values for the cut-off, c.

Define:

$$\text{sensitivity}(c) \quad : \quad P[\text{Test} > c \mid Y = 1]$$

$$\text{specificity}(c) \quad : \quad P[\text{Test} \leq c \mid Y = 0]$$

Then the ROC curve plots the values

$$\begin{array}{ccc} [\ \text{1-specificity}(c) & , & \text{sensitivity}(c)\ ] \\ (\text{False Positive using } c & , & \text{True Positive using } c) \end{array}$$

For all possible values of c.

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

Predicted values

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

- Cross-validation is one option to obtain a more honest assessment of the predictive capacity of a model. This involves using separate data to fit the model and assess the model.

Examples:

◦ LBW data

◦ Prediction of Hepatic disorders (Duh et al. AJE, 1998)

◦ Prediction of Gypsy moths

STATA Example with LBW Data

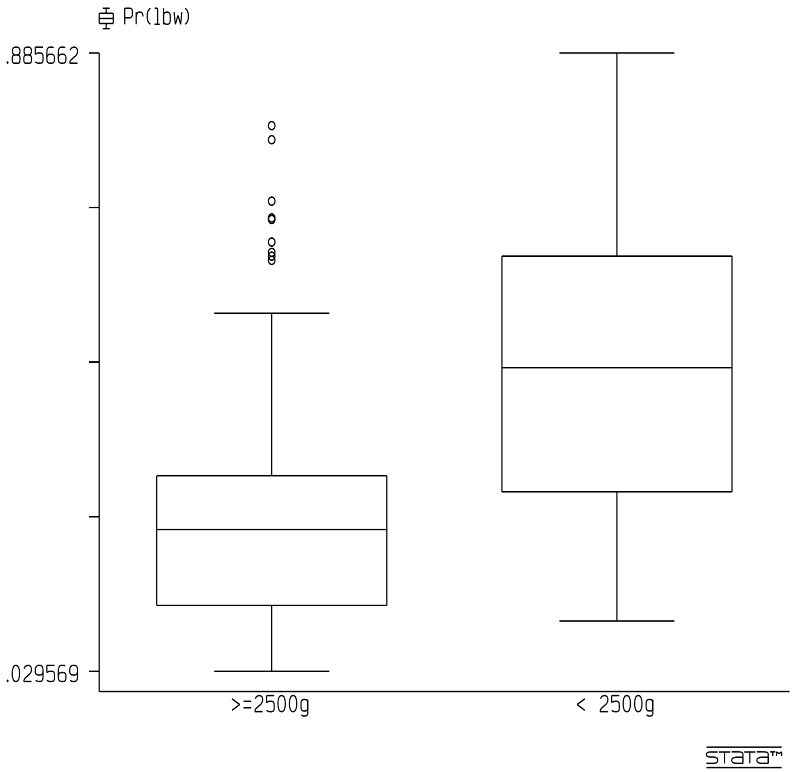
```
infile id lbw age lwt race smoke ptl hyper urirr pvft weight using lowbwt.dat

*** VARIABLE LABELS HERE ***

***
*** fit the logistic regression model
***

xi:  logistic lbw smoke anyptl i.race i.newpvft hyper urirr i.agecat

***
*** obtain the fitted probabilities
***

predict fitp

summarize fitp
summarize lbw

***
*** display data for 1:5
***

list fitp lbw smoke anyptl race newpvft hyper urirr agecat in 1/5

***
*** look at fitted value by outcome
***

sort lbw
graph fitp, box by(lbw) saving("lbw-FitBox.plot")
gphpen -deps "lbw-FitBox.plot"
```

```
***
*** obtain sensitivity and specificity based on Test+ = ( fitp > 0.5 )
***

lstat

***
*** look at full range of sensitivity and specificity
***

lsens, saving("lbw-SensSpec.plot")
gphpen -deps "lbw-SensSpec.plot"

***
*** look at the roc curve
***

lroc, saving("lbw-ROC.plot")
gphpen -deps "lbw-ROC.plot"
```


Fitted Values and Data

```
. list fitp lbw smoke anyptl race newpvft hyper urirr agecat in 1/20

Observation 1

        fitp    .1583131       lbw    >=2500g      smoke          No
      anyptl           0      race      other    newpvft           1
       hyper          No     urirr         No     agecat         <20

Observation 2

        fitp    .7107266       lbw    < 2500g      smoke         Yes
      anyptl          1+      race      white    newpvft           0
       hyper          No     urirr        Yes     agecat         <20

Observation 3

        fitp    .1096878       lbw    >=2500g      smoke          No
      anyptl           0      race      white    newpvft           0
       hyper          No     urirr         No     agecat         <20

Observation 4

        fitp     .222472       lbw    >=2500g      smoke         Yes
      anyptl           0      race      white    newpvft           0
       hyper          No     urirr         No     agecat         <20

Observation 5

        fitp    .2262478       lbw    >=2500g      smoke          No
      anyptl           0      race      other    newpvft           0
       hyper          No     urirr         No     agecat         <20
```

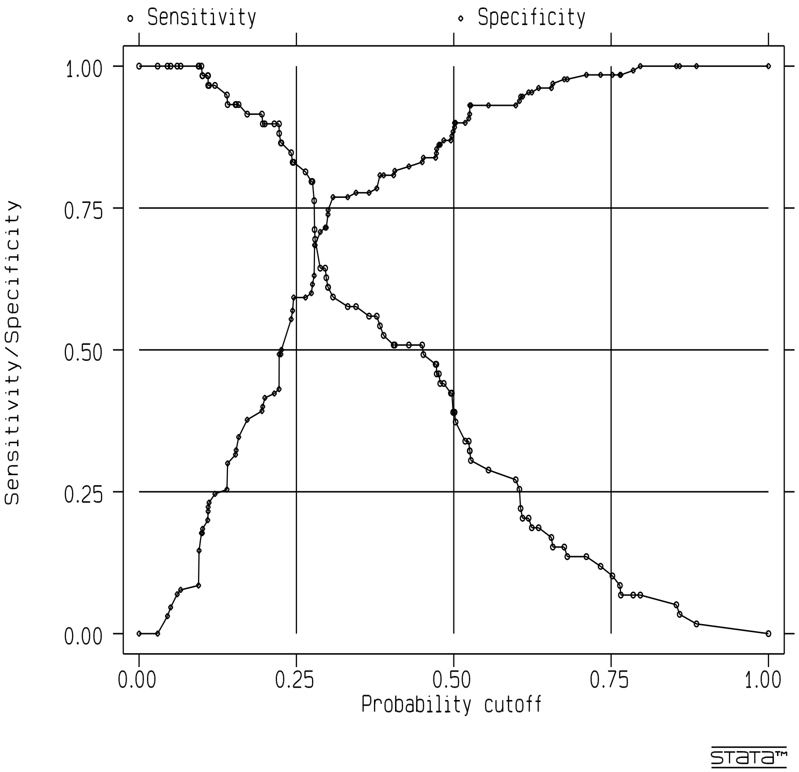
Fitted Values by LBW



Default Sensitivity / Specificity

```
. lstat

Logistic model for lbw

              -------- True --------
Classified |         D            ~D         Total
-----------+--------------------------+-----------
     +     |        23            14  |          37
     -     |        36           116  |         152
-----------+--------------------------+-----------
   Total   |        59           130  |         189

Classified + if predicted Pr(D) >= .5
True D defined as lbw ~= 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   38.98%
Specificity                     Pr( -|~D)   89.23%
Positive predictive value       Pr( D| +)   62.16%
Negative predictive value       Pr(~D| -)   76.32%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   10.77%
False - rate for true D         Pr( -| D)   61.02%
False + rate for classified +   Pr(~D| +)   37.84%
False - rate for classified -   Pr( D| -)   23.68%
--------------------------------------------------
Correctly classified                        73.54%
--------------------------------------------------
```

Sensitivity and Specificity



ROC curve