Unit 5
Logistic Regression

"To all the ladies present and some of those absent"

- Jerzy Neyman

What behaviors influence the chances of developing a sexually transmitted disease? Comparing demographics, health education, access to health care, which of these variables are significantly associated with failure to obtain an HIV test? Among the several indicators of risk, including age, co-morbidities, severity of disease, which are significantly associated with surgical mortality among patients undergoing transplant surgery? In all of these examples, the outcome observed for each individual can take on only one of two possible values: positive or negative test, alive or dead, remission or non-remission, and so on. Collectively, the data to be analyzed are ***proportions***.

Proportions have some important features that distinguish them from data measured on a continuum. Proportions (1) are ***bounded*** from below by the value of zero (or zero percent) and bounded from above by one (or 100 percent); (2) as the proportion gets close to either boundary, the variance of the proportion gets smaller and smaller; thus, we ***cannot assume a constant variance***; and (3) proportions are ***not distributed normal***. **Normal theory regression models are <u>not</u> appropriate for the analysis of proportions.**

In unit 4, Categorical Data Analysis, emphasis was placed on contingency table approaches for the analysis of such data and it was highlighted that these methods should always be performed for at least two reasons: (1) they give a good feel for the data; and (2) they are free of the assumptions required for regression modeling.

Unit 5 is an introduction to <u>logistic regression</u> approaches for the analysis of proportions where it is of interest to explore the roles of possibly several influences on the observed proportions.

Table of Contents

Learning Objectives

When you have finished this unit, you should be able to:

- Explain why a normal theory regression model is ***not*** appropriate for a regression analysis of proportions.
- State the ***expected value*** (the mean) of a Bernoulli random variable.
- Define the ***logit*** of the mean of a Bernoulli random variable.
- State the ***logistic regression model*** and, specifically, the logit link that relates the logit of the mean of a Bernoulli random variable to a linear model in the predictors.
- Explain how to ***estimate odds ratio measures*** of association from a fitted logistic regression model.
- Explain how to ***estimate probabilities of event*** from a fitted logistic regression model.
- Perform and interpret ***likelihood ratio test*** comparisons of hierarchical models.
- Explain and compare ***crude versus adjusted*** estimates of odds ratio measures of association.
- Assess ***confounding*** in logistic regression model analyses.
- Assess ***effect modification*** in logistic regression model analyses.
- ***Draft an analysis plan*** for multiple predictor logistic regression analyses of proportions.

1. From Linear Regression To Logistic Regression
An Organizational Framework

In unit 2 (*Regression and Correlation*), we considered single and multiple predictor regression models for a single outcome random variable Y assumed continuous and distributed normal.

In unit 5 (*Logistic regression*), we consider single and multiple regression models for a single outcome random variable Y assumed discrete, binary, and distributed bernoulli.

	Unit 2 **Normal Theory Regression**	**Unit 5** **Logistic Regression**
Y	- univariate - continuous - Example: Y = cholesterol	- univariate - discrete, binary - Example: Y = dead/alive
$X_1, X_2, \ldots.., X_p$	- one or multiple - discrete or continuous - treated as fixed	- one or multiple - discrete or continuous - treated as fixed
$Y \mid X_1=x_1, .., X_p=x_p$	- Normal (Gaussian)	- Bernoulli (or binomial)
$E(Y\mid X_1=x_1, . X_p=x_p)$	$\mu_{Y\mid X_1 \ldots X_p} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$	$\mu_{Y\mid X_1 \ldots X_p} = \pi_{Y\mid X_1 \ldots X_p}$ $= \dfrac{1}{1+\exp\left[-\left(\beta_0+\beta_1 x_1+\ldots+\beta_p x_p\right)\right]}$
Right hand side of model	$\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$	$\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
Link	"natural" or "identity" $\mu_{Y\mid X_1 \ldots X_p}$ $= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$	"logit" $\text{logit}[\mu_{Y\mid X_1 \ldots X_p}]$ $= \text{logit}[\pi_{Y\mid X_1 \ldots X_p}]$ $= \ln\left[\pi_{Y\mid X_1 \ldots X_p} \big/ \left(1-\pi_{Y\mid X_1 \ldots X_p}\right)\right]$ $= \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
Estimation	Least squares (= maximum likelihood)	Maximum Likelihood
Tool	Residual sum of squares	Deviance statistic
Tool	Partial F Test	Likelihood Ratio Test

2. Use of Video Display Terminals and Spontaneous Abortion

Consider the following published example of logistic regression.

Source: Schnorr et al (1991) Video Display Terminals and the Risk of Spontaneous Abortion. *New England Journal of Medicine* 324: 727-33.

Background:

Adverse pregnancy outcomes were correlated with use of video display terminals (VDT's) beginning in 1980.

Subsequent studies were inconsistent in their findings.

Previous exposure assessments were self-report or derived from job title descriptions.

Electromagnetic fields were not previously measured.

Research Question:

What is the nature and significance of the association, as measured by the odds ratio, between exposure to electromagnetic fields emitted by VDTs and occurrence of spontaneous abortion, after controlling for

- History of prior spontaneous abortion
- Cigarette Smoking
- History of thyroid condition

Design: Retrospective cohort investigation of two groups of full-time female telephone operators.

		Spontaneous Abortion	
882 Pregnancies:	**N**	**n**	**%**
Exposed	**366**	**54**	**14.8%**
Unexposed	**516**	**82**	**15.9%**

The Data:

Variable	**Label**	**Range/Codes**
AVGVDT	average hours vdt in 1st trimester	continuous
NUMCIGS	# cigarettes/day	continuous
PRIORSAB	prior spontaneous abortion	1=yes, 0=no
SAB	spontaneous abortion	1=yes, 0=no
SMOKSTAT	smoker	1=yes, 0=no
PRTHYR	prior thyroid condition	1=yes, 0=no
VDTEXPOS	VDT exposure	1=yes, 0=no

AVGVDT	NUMCIGS	PRIORSAB	SAB	SMOKSTAT	PRTHYR	VDTEXPOS
0.000	15	0	0	1	0	0
0.000	10	0	0	1	0	0
0.000	20	0	0	1	0	0
	20	0	0	1	0	1
27.764	20	0	1	1	0	1
28.610	0	0	0	0	0	1
0.000	0	0	0	0	0	0
	0	0	0	0	0	1
19.717	0	0	0	0	0	1
0.000	0	0	0	0	0	0
25.022	0	0	0	0	0	1
…	…	…	…	…	…	…
0.000	0	1	0	0	0	0

3. Definition of the Logistic Regression Model

We suspect that multiple factors, especially use of video display terminals, contribute to an individual's odds of spontaneous abortion.

The outcome or dependent variable is Y=sab. Its value is y and

= 1 if spontaneous abortion occurred
0 otherwise

The predictors that might influence the odds of SAB are several:

X_1 = avgvdt
X_2 = numcigs
X_3 = priorsab
X_4 = smokstat
X_5 = prthyr, and
X_6 = vdtexpos

We are especially interested in

X_6 = vdtexpos (coded = 1 for exposed and = 0 for NON exposed) and
X_1 = avgvdt

Among the N=882 in our sample, we have potentially N=882 unique probabilities of spontaneous abortion.

$$\pi_1, \pi_2, \ldots, \pi_N.$$

For the i^{th} person

$$\pi_i = \text{Function}\ (X_{1i}, X_{2i}, X_{3i}, X_{4i}, X_{5i}, X_{6i})$$

$$\Pr [Y_i = 1] = \pi_i$$
$$\Pr [Y_i = 0] = (1 - \pi_i)$$

How do we model the N=882 individual probabilities π_i in relationship to the predictors?

Recall. Each profile of values, $\underline{X} = [\ X_1{=}x_1\ X_2{=}x_2, \ldots.\ X_6{=}x_6\ ]$, defines a sub-population with their own distribution of outcomes Y. For example the women with $X_3{=}1$ are the women with a history of prior spontaneous abortion, and are distinct from the women with $X_3{=}0$ (who have no such prior history). And so on; we can talk about distinct sub-populations based on the entire profile of values on $X_1, X_2, \ldots X_6$.

Review of normal theory linear regression analysis:

Y |[$X_1, X_2, X_3, X_4, X_5, X_6$] (read: "Y given [$X_1, X_2, X_3, X_4, X_5, X_6$]" is assumed to be distributed normal (Gaussian)

with mean = $\mu_{\ Y|\underline{x}}$ and variance=$\sigma^2_{Y|X}$.

The mean of Y at [$X_1, X_2, X_3, X_4, X_5, X_6$] is modeled linearly in $\underline{x}$ = [$X_1, X_2, X_3, X_4, X_5, X_6$]

Thus mean of Y | [$X_1, X_2, X_3, X_4, X_5, X_6$] = E [Y | ($X_1, X_2, X_3, X_4, X_5, X_6$)] = $\mu_{\ Y|\underline{x}}$

In normal theory linear regression:

$$\underbrace{E[Y|\ \underline{x}\ ] = \mu_{\underline{x}}}_{\text{“natural link”}} \quad = \underbrace{\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6}_{\text{“right hand side is linear in the predictors”}}$$

In a logistic model regression analysis, the framework is a little different:

Y is assumed to be distributed Bernoulli
with mean=$\pi_{\underline{x}}$ and variance= $\pi_{\underline{x}}$ (1-$\pi_{\underline{x}}$)

We do not model the mean of Y|$\underline{X}$=$\underline{x}$ = $\pi_{\underline{x}}$ linearly in $\underline{x}$ = [$X_1 \cdots X_6$].

*Instead, we model the **logit of the mean of** Y|$\underline{X}$=$\underline{x}$ = $\pi_{\underline{x}}$ linearly in $\underline{x}$ = [$X_1 \cdots X_6$].*

$$\underbrace{\text{Logit [E(Y|}\underline{X}\text{)] = logit[}\pi_{\underline{x}}\text{]}}_{\text{“logit link”}} = \ln\left[\frac{\pi_x}{1-\pi_x}\right] = \underbrace{\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \ldots + \beta_5X_5 + \beta_6X_6}_{\text{“right hand side is linear in the predictors”}}$$

Solution for Probability [Y=1| $X_1=x_1, X_2=x_2, \ldots, X_6=x_6$] = E[Y | $X_1=x_1, X_2=x_2, \ldots, X_6=x_6$] :

The formula for Pr [Y = 1| $X_1=x_1, X_2=x_2, \ldots, X_6=x_6$] can be written in either of two ways:

$$\pi_{\underline{x}} = \frac{\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_6x_6)}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_6x_6)}$$

$$= \frac{1}{1+\exp\left[-(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_6x_6)\right]}$$

Pr [Y = 0 | $X_1=x_1, X_2=x_2, \ldots, X_6=x_6$] is

$$(1-\pi_{\underline{x}}) = \frac{1}{1+\exp(\beta_0+\beta_1x_1+\beta_2x_2+\ldots+\beta_6x_6)}$$

Two other names for this model are "log-linear odds" and "exponential odds"

The logistic regression model focuses on the odds of event (in this case event of spontaneous abortion, SAB).

1) $\ln[\text{ odds } (\pi_{\underline{x}})] = \ln\left[\frac{\pi_{\underline{x}}}{1-\pi_{\underline{x}}}\right] = \beta_0 + \cdots + \beta_6X_6$ is a log-linear odds model.

2) $\left[\frac{\pi_{\underline{x}}}{1-\pi_{\underline{x}}}\right] = \exp\{\ \beta_0 + \cdots + \beta_6X_6\ \}$ is an exponential odds model.

We do not model E[Y | X] = $\pi_{\underline{x}} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6$?

1) $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6$
can range from $-\infty$ to $+\infty$ but $\pi_{\underline{x}}$ ranges from 0 to 1.

2) $\pi_{\underline{x}} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6$ is often not a good description of nature.

Assumptions:

1) Each Y_i follows a distribution that is Bernoulli with parameter E[Y | X] = $\pi_{\underline{x}_i}$.

2) The $Y_1, Y_2, \cdots , Y_N$ are independent.

3) The values of the predictors, $X_{i1}=x_{i1} \cdots X_{i6}=x_{i6}$, are treated as fixed.

4) The model is correct (this is also referred to as "linearity in the logit").

$$\text{logit}[\ E(Y)|\ X_1=x_1, X_2=x_2, \ldots, X_6=x_6\ ]$$

$$= \text{logit}\ [\ \pi_{\underline{x}}]$$

$$= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6$$

5) No multicollinearity

6) No outliers

7) Independence

3. Estimating Odds Ratios

For now, assume that we have a fitted model. We'll get to the details of estimation later.

Once a logistic regression model has been fit, the prediction equation can be used to estimate odds ratio (OR) measures of association.

Example 1: What is the estimated crude relative odds (OR) of spontaneous abortion (SAB) associated with any exposure (1 = exposed, 0 = not exposed) to a video display terminal (VDTEXPOS)?

Step 1:
To obtain crude odds ratios, either a 2x2 table can be used or a one predictor logistic regression model can be fit. Here, it is given by

$$\textbf{logit \{ probability [SAB=1] \}} = \beta_0 + \beta_1 \textbf{ VDTEXPOS}$$

Stata

```
. * The following assumes you have downloaded and opened vdt.dta.
.  logit sab vdtexpos
```

```
Logistic regression                                 Number of obs   =        882
                                                    LR chi2(1)      =       0.21
                                                       = Likelihood Ratio Statistic
                                                       for current model ("full") v
                                                    intercept only model ("reduced")
                                                             Analogous to Overall F
                                                    Prob > chi2     =     0.6443
Log likelihood = -379.08045                         Pseudo R2       =     0.0003
(-2) ln L = 758.1609
                                                    Wald Z    Wald Z p-value (2 sided) using Normal(0,1)

------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vtexpos  | -.0876939 = $\hat{\beta}_1$   .1903232    -0.46   0.645    -.4607204    .2853327
       _cons | -1.666325 = $\hat{\beta}_0$   .1204129   -13.84   0.000    -1.90233    -1.43032
------------------------------------------------------------------------------
```

z = Wald Z = [Coef] / [Std. Err.]
= [beta – 0] / [SE(beta)] ~ Normal(0,1) when $\beta_1 = 0$

Yielding the following prediction equation

$$\text{Fitted logit \{ pr[sab=1] \}} = -1.66633 - 0.08769{*}\text{vdtexpos}$$

Step 2:

Recognize a wonderful bit of algebra.

For a single exposure variable (1=exposed, 0=not exposed)

$$\mathbf{OR}_{\text{1 versus 0}} = \exp\{\,\beta\,\} \text{ where } \beta = \text{regression parameter for the exposure variable}$$

$$= \exp\{\,\text{logit}(\pi_1) - \text{logit}(\pi_0)\,\}$$

Proof (read if you are interested!):

$$\mathbf{OR} = \exp\{\ln[OR]\}$$

$$= \exp\left\{\ln\left[\frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}\right]\right\}$$

$$= \exp\left\{\ln\left[\frac{\pi_1}{1-\pi_1}\right] - \ln\left[\frac{\pi_0}{1-\pi_0}\right]\right\}$$

$$= \exp\{\text{logit}(\pi_1)\text{-logit}(\pi_0)\}$$

<u>"1" is the comparison and is vdtexpos=1:</u>

Estimated logit { prob[SAB=1|vdtexpos=1] } $= \hat{\beta}_0 + \hat{\beta}_1$

= -1.66633 - 0.08769

<u>"0" is the reference and is vdtexpos=0:</u>

Estimated logit { prob[SAB=1| vdtexpos=0] } $= \hat{\beta}_0$ = -1.66633

Step 3: Apply.

The odds ratio measure of association comparing the exposed telephone operator ("1") to the unexposed telephone operator ("0") is

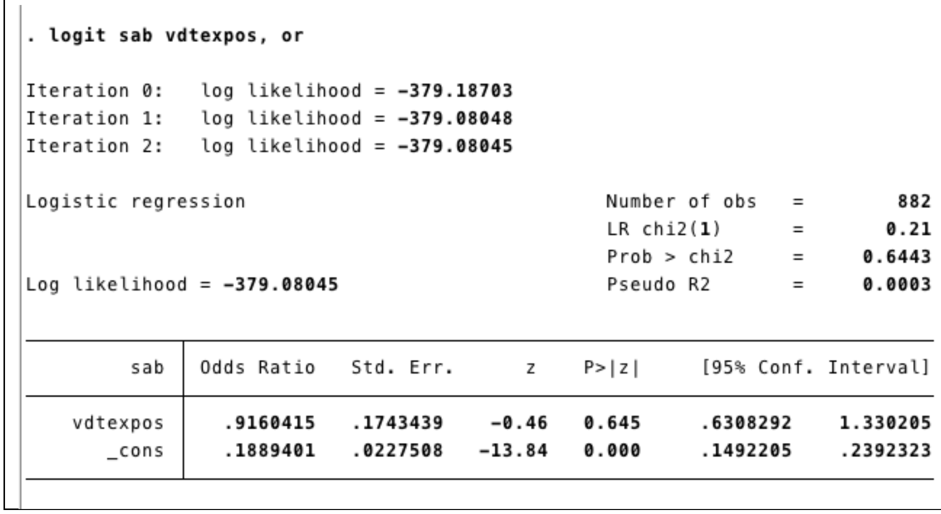
$$= \exp\{\,\text{logit}(\pi_1) - \text{logit}(\pi_0)\,\}$$
$$= \exp\{\,[\,\beta_0 + \beta_1\,] - [\,\beta_0]\,\}$$
$$= \exp\{\,\beta_1\,\}$$
$$= \exp\{-0.08769\}$$

= 0.9160 → *"Compared to the unexposed, the exposed have a relative odds of spontaneous abortion=.916"*

Stata Illustration – Obtaining estimated odds ratios after logistic regression

Method 1. Command logit with option or

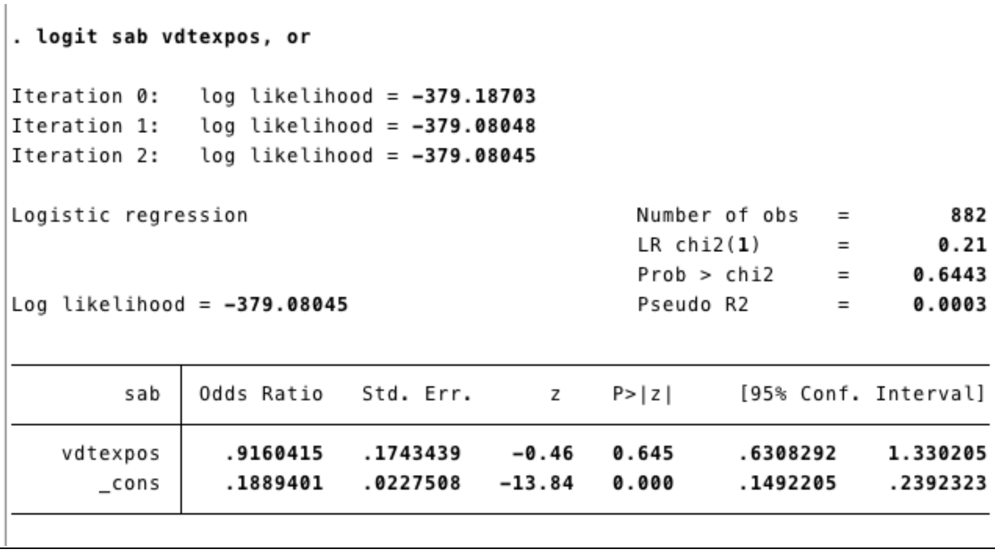
```
. logit sab vdtexpos, or
```



```
. logit sab vdtexpos, or

Iteration 0:   log likelihood = -379.18703
Iteration 1:   log likelihood = -379.08048
Iteration 2:   log likelihood = -379.08045

Logistic regression                             Number of obs     =        882
                                                LR chi2(1)        =       0.21
                                                Prob > chi2       =     0.6443
Log likelihood = -379.08045                     Pseudo R2         =     0.0003

------------------------------------------------------------------------------
         sab | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vdtexpos |   .9160415   .1743439    -0.46   0.645     .6308292    1.330205
       _cons |   .1889401   .0227508   -13.84   0.000     .1492205    .2392323
------------------------------------------------------------------------------
```

Method 2. Command logistic

```
. logistic sab vdtexpos
```



```
. logit sab vdtexpos, or

Iteration 0:   log likelihood = -379.18703
Iteration 1:   log likelihood = -379.08048
Iteration 2:   log likelihood = -379.08045

Logistic regression                             Number of obs     =        882
                                                LR chi2(1)        =       0.21
                                                Prob > chi2       =     0.6443
Log likelihood = -379.08045                     Pseudo R2         =     0.0003

------------------------------------------------------------------------------
         sab | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    vdtexpos |   .9160415   .1743439    -0.46   0.645     .6308292    1.330205
       _cons |   .1889401   .0227508   -13.84   0.000     .1492205    .2392323
------------------------------------------------------------------------------
```

The two profiles being compared can differ on several predictors! Let's try another one.

Here is another bit of wonderful algebra.

For two profiles of predictor variable values, "comparison" versus "reference"

$$\mathbf{OR}_{\text{comparison versus reference}} = \exp\{\text{logit}(\pi_{\text{comparison}}) - \text{logit}(\pi_{\text{reference}})\}$$

Example 2 - What is the estimated relative odds (OR) of spontaneous abortion (SAB) for a person who is not exposed to a VDT, smokes 10 cigarettes per day, has no history of prior SAB, and no thyroid condition relative to a person who has an average of 20 hours exposure to a VDT, is a nonsmoker, has a history of prior SAB and does have a thyroid condition?

Step 1:

Here the model fit is the 4 predictor model:

$$\text{logit}\{\text{probability}[\text{sab}=1]\} = \beta_0 + \beta_1\,\text{avgvdt} + \beta_2\,\text{numcigs} + \beta_3\,\text{priorsab} + \beta_4\,\text{prthyr}$$

Estimation now yields (output not shown).

$$\text{fitted logit}\{\text{prob}[\text{sab}=1]\} = -1.95958 + 0.00508(\text{avgvdt}) + 0.04267(\text{numcigs}) + 0.38500(\text{priorsab}) + 1.27420(\text{prthyr})$$

Step 2:

Calculate the two predicted logits and compute their difference.

	Value of Predictor for Person	
	"comparison"	**"reference"**
avgvdt	0	20
numcigs	10	0
priorsab	0	1
prthyr	0	1

"comparison"

$$\text{logit}[\pi_{\text{comparison}}] = -1.95958 + 0.00508(0) + 0.04267(10) + 0.38500(0) + 1.27420(0)$$
$$= -1.5329$$

"reference":

$$\text{logit}[\pi_{\text{reference}}] = -1.95958 + 0.00508(20) + 0.04267(0) + 0.38500(1) + 1.27420(1)$$
$$= -0.1988$$

$$\mathbf{logit}[\pi_{\text{comparison}}] - \mathbf{logit}[\pi_{\text{reference}}] = \mathbf{-1.5329 - [-0.1988]}$$
$$= \mathbf{-1.3341}$$

Step 3:

Exponentiate.

$$OR_{\text{comparison versus reference}} = \exp\{\ \text{logit}[\pi_{\text{comparison}}] - \text{logit}[\pi_{\text{reference}}]\ \}$$
$$= \exp\{-1.3341\}$$
$$= 0.2634$$

Interpetation - The estimated odds (OR) of spontaneous abortion (SAB) for a person who is not exposed to a VDT, smokes 10 cigarettes per day, has no history of prior SAB, and no thyroid condition *is 0.2634 times* that of the odds of spontaneous abortion (SAB) for a person who has an average of 20 hours exposure to a VDT, is a nonsmoker, has a history of prior SAB and does have a thyroid condition.

In General:

The Odds Ratio estimate ($\hat{OR}$) of association with outcome accompanying a unit change in the predictor X is a function of the estimated regression parameter $\hat{\beta}$

$$\hat{OR}_{\text{UNIT change in X}} = \exp\{\hat{\beta}\}$$

Tip – $OR_{\text{10 unit change in X}} = \exp[10*\beta]$

A hypothesis test of OR=1

Is equivalent to

A hypothesis test of $\beta = 0$

For a rare outcome (typically disease), the relative risk ($\hat{RR}$) estimate of association with outcome accompanying a unit change in the predictor X is reasonably estimated as a function of the estimated regression parameter β

$$\hat{RR}_{\text{UNIT change in X}} = \exp\{\hat{\beta}\}, \textit{approximately}$$

5. Estimating Probabilities

Again, let's assume that we have a fitted model. We'll get to the details of estimation later.

Once a logistic regression model has been fit, the prediction equation can also be used to estimate probabilities of event occurrence. The prediction equation can be used to estimate probabilities of event of disease if the study design is a cohort; it is used to estimate probabilities of history of exposure if the study design is case-control.

Reminder …– it is not possible to estimate probability of disease from analyses of case-control studies.

Recall that for Y distributed Bernoulli

$$E\,[\,Y\,] = \pi = \text{Probability of event occurrence}$$

Example 1- Under the assumption of a cohort study design, what is estimated probability of spontaneous abortion (sab) for a person with any exposure to a video display terminal? Consider the single predictor model containing the predictor vdtexpos)

Step 1:

Recall that we obtained the following equation for the fitted logit for the one predictor model containing VDTEXPOS:

$$\begin{aligned}\text{Predicted logit \{ prob[SAB=1| vdtexpos] \}} &= \hat{\beta}_0 + \hat{\beta}_1[\text{vdtexpos}] \\ &= -1.66633 \ - \ 0.08769*\text{VDTEXPOS}\end{aligned}$$

Step 2:

Utilizing the algebra on page 9, we have:

$$\text{Estimated pr[SAB=1]} \ = \ \hat{\pi}_{\text{VDTEXPOS=1}} = \frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1[\text{vdtexpos}]\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1[\text{vdtexpos}]\right)} = \frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1\right)}$$

Step 3:

Set VDTEXPOS=1, β_0 = -1.66633, β_1 =-0.08769 and solve

$$\text{Estimated pr[SAB=1]}=\frac{\exp(-1.66633\ \text{-}\ 0.08769[1])}{1+\exp(-1.66633\ \text{-}\ 0.08769[1])}$$

$$=\frac{0.1731}{1.1731}=0.148$$

6. The Deviance Statistic

”G Statistic”, “Log likelihood Statistic”, “Scaled Deviance”, Residual Deviance””

Where are we now? Recall the concept of “analysis of variance” introduced in Unit 2, Regression and Correlation. Analysis of variance is about the total variability of the observed outcome, and its partitioning into portions that are explained by the fitted model (due model/due regression) versus what’s left over as unexplained (due residual/due error). The deviance statistic in logistic regression is a measure of what remains left over as unexplained by the fitted model, analogous to the residual sum of squares in normal theory regression.

But first, a few words about likelihood, L.

$L_{saturated}$: We get the largest likelihood of the data when we fit a model that allows a separate predictor for every person. This is called the likelihood of the saturated model.

$L_{saturated}$ is a large number.

$L_{current}$: We get an estimated likelihood of the data when we fit the current model.

$L_{current}$ is a smaller number.

The **deviance statistic** in logistic regression is related to the two likelihoods, $L_{current}$ and $L_{saturated}$ in the following way.

The current model explains a lot	**The current model does NOT explain a lot**
$L_{current} \approx L_{saturated}$	$L_{current} << L_{saturated}$
$\dfrac{L_{current}}{L_{saturated}} \approx 1$	$\dfrac{L_{current}}{L_{saturated}} << 1$
$\ln\left[\dfrac{L_{current}}{L_{saturated}}\right] \approx 0$	$\ln\left[\dfrac{L_{current}}{L_{saturated}}\right] << 0$
Deviance $= (-2)\ \ln\left[\dfrac{L_{current}}{L_{saturated}}\right]$ ≈ 0	**Deviance** $= (-2)\ \ln\left[\dfrac{L_{current}}{L_{saturated}}\right]$ $>> 0$
A number close to 0	A large positive number

Evidence that the current model explains a lot of the variability in outcome

Deviance ≈ small
p-value ≈ large

$$\textbf{Deviance Statistic, D} = -2 \ln\left[\frac{L_{current}}{L_{saturated}}\right]$$

$$= (-2)\ln(L_{current}) - (-2)\ln(L_{saturated})$$

Deviance df = [Sample size] – [# fitted parameters]

where

$L_{current}$ **= likelihood of data using current model**
$L_{saturated}$ **= likelihood of data using the saturated model**

Notes -
(1) By itself, the deviance statistic does not have a well defined distribution
(2) However, differences of deviance statistics that compare hierarchical models do have well defined distributions, namely chi square distributions.

A Feel for the Deviance Statistic

(1) Roughly, the **deviance statistic D** is a measure of what remains unexplained.
Hint – The analogue in normal theory regression is the residual sum of squares (SSQ error)

(2) A deviance statistic value **close to zero** says that a lot is explained and, importantly, that little remains unexplained. → The current model with its few predictors performs similarly to the saturated model that permits a separate predictor for each person.

(3) **WARNING!** The deviance statistic D is **NOT** a measure of goodness-of-fit. Recall that we said the same thing about the overall F-statistic in normal theory regression.

(4) The **deviance statistic D** is the basis of the **likelihood ratio test** .

(5) The **likelihood ratio test** is used for the comparison of **hierarchical models**.
Recall – In normal theory regression, hierarchical models are compared using the Partial F-test.

a. The Likelihood Ratio (LR) Test

Likelihood Ratio (LR) Test

Under the assumptions of a logistic regression model and the comparison of the hierarchical models:

$$\text{Reduced: } \operatorname{logit}[\pi \mid X_1, X_2 \ldots, X_p] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

$$\text{Full: } \operatorname{logit}[\pi \mid X_1, X_2 \ldots, X_p, X_{p+1}, X_{p+2}, \ldots, X_{p+k}] = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \beta_{p+1} X_{p+1} + \ldots + \beta_{p+k} X_{p+k}$$

For testing:

$$H_O: \ \beta_{p+1} = \beta_{p+2} = \ldots = \beta_{p+k} = 0$$

$$H_A: \ \text{not}$$

A Likelihood Ratio Test Statistic LR, defined

$$\mathbf{LR = Deviance_{REDUCED} - Deviance_{FULL}}$$

$$= [\,(-2)\ln(L)_{REDUCED} - (-2)\cancel{\ln(L)}_{SATURATED}\,] - [\,(-2)\ln(L)_{FULL} - (-2)\cancel{\ln(L)}_{SATURATED}\,]$$

$$= [\,(-2)\ln(L)_{REDUCED}\,] - [\,(-2)\ln(L)_{FULL}\,]$$

has null hypothesis distribution that is Chi Square$_{DF=k}$

Thus, rejection of the null hypothesis occurs for

Test statistic values, LR = large
and accompanying p-value= small

Tip – In practice, we obtain LR using the 2nd formula; it says: LR = [(-2) ln (L) $_{REDUCED}$] - [(-2) ln (L) $_{FULL}$]

Example: Controlling for prior spontaneous abortion (PRIORSAB), is 0/1 exposure to VDT associated with spontaneous abortion?

The idea here is similar to the idea of the partial F test in normal theory linear regression. Two models that are **hierarchical** are compared: a "reduced/reference" versus a "full/comparison".

Step 1: Fit the **"reduced/reference"** model, defined as containing the control variable(s) only. (Note – The available sample size here is 881)

It estimates that logit {pr [sab=1]} = $\beta_0 + \beta_1$ PRIORSAB

$$\begin{aligned} \textbf{(-2) ln } \mathbf{L}_{\textbf{reduced}} &= \textbf{754.56} \\ \textbf{Deviance DF}_{\textbf{reduced}} &= \textbf{881 – 1 = 880} \end{aligned}$$

Step 2: Fit the **"full/comparison"** model, defined as containing the control variable(s) + predictor(s) of interest. It estimates that logit {pr [sab=1]} = $\beta_0 + \beta_1$ PRIORSAB + β_2 VDTEXPOS

$$\begin{aligned} \textbf{(-2) ln } \mathbf{L}_{\textbf{full}} &= \textbf{753.81} \\ \textbf{Deviance DF}_{\textbf{full}} &= \textbf{881 – 2 = 879} \end{aligned}$$

Step 3: Compute the change in deviance and the change in deviance df, remembering that in logistic regression the subtraction is of the form "reduced" - "full".

$$\begin{aligned} \textbf{Likelihood Ratio Test LR} &= \textbf{(-2) ln } \mathbf{L}_{\textbf{reduced}} \textbf{ - (-2) ln } \mathbf{L}_{\textbf{full}} \\ &= \textbf{754.56 - 753.81} \\ &= \textbf{0.75} \end{aligned}$$

$$\begin{aligned} \mathbf{\Delta} \textbf{ Deviance Df} &= \textbf{Deviance DF}_{\textbf{reduced}} \textbf{ - Deviance DF}_{\textbf{full}} \\ &= \textbf{880 - 879} \\ &= \textbf{1} \end{aligned}$$

Example – continued.

H_0: VDTEXPOS, controlling for PRIORSAB, is **not** associated with SAB
$\boldsymbol{\beta}_{\textbf{VDTEXPOS}} = \mathbf{0}$ in the model that also contains PRIORSAB

H_A: VDTEXPOS, controlling for PRIORSAB, is associated with SAB
$\boldsymbol{\beta}_{\textbf{VDTEXPOS}} \neq \mathbf{0}$ in the model that also contains PRIORSAB

Suppose we obtain:
Likelihood Ratio Statistic $\chi^2(\text{df}=1) = 0.75$
p-value = .39

Interpretation. Assumption of the null hypothesis $\boldsymbol{\beta}_{\textbf{VDTEXPOS}} = \mathbf{0}$ and its application to the observed data yields a result that is reasonably plausible (p-value=.39). The null hypothesis is **NOT** rejected. Conclude that there is **not statistically significant evidence** that exposure to VDT, after controlling for prior spontaneous abortion, is associated with spontaneous abortion.

Note - A little algebra (not shown) reveals that there are two, equivalent, formulae for the LR test:

Solution #1
LR Test = Δ Deviance Statistic
[Deviance (reduced model)] - [Deviance (full model)]

Solution #2: this works because ln likelihood (saturated) drops out… see page 20
LR Test = Δ Deviance
= Δ { (-2) ln (likelihood)]
= [(-2) ln likelihood (reduced model)] - [(-2) ln likelihood (full model)]

b. Model Development

Recall from Unit 2, Regression and Correlation …. with apologies, the following is a duplication

There are ***no*** rules ***nor a single best strategy***. Different study designs and research questions call for different approaches. ***Tip –*** Before you begin model development, make a list of your study design, research aims, outcome variable, primary predictor variables, and covariates.

As a general suggestion, the following approach has the advantages of providing a reasonably thorough **exploration of the data and relatively little risk of missing something important.**

Preliminary – Be sure you have: (1) checked, cleaned and described your data, (2) screened the data for multivariate associations, and (3) thoroughly explored the bivariate relationships.

Step 1 – Fit the "maximal" model.
The maximal model is the large model that contains all the explanatory variables of interest as predictors. This model also contains all the covariates that might be of interest. It also contains all the interactions that might be of interest. Note the amount of variation explained.

Step 2 – Begin simplifying the model.
Inspect each of the terms in the "maximal" model with the goal of removing the predictor that is the least significant. Drop from the model the predictors that are the least significant, beginning with the higher order interactions (***Tip*** -interactions are complicated and we are aiming for a simple model). Fit the reduced model. Compare the amount of variation explained by the reduced model with the amount of variation explained by the "maximal" model.

If the deletion of a predictor has little effect on the variation explained ….
Then leave that predictor out of the model.|

And inspect each of the terms in the model again.

If the deletion of a predictor has a significant effect on the variation explained …
Then put that predictor back into the model.

Step 3 – Keep simplifying the model.
Repeat step 2, over and over, until the model remaining contains nothing but significant predictor variables.

Beware of some important caveats

- Sometimes, you will want to keep a predictor in the model regardless of its statistical significance (an example is randomization assignment in a clinical trial)
- The order in which you delete terms from the model matters!
- You still need to be flexible to considerations of biology and what makes sense.

So what's new here?

In logistic regression, this is done using the likelihood ratio test.

If the likelihood ratio statistic is statistically significant (small p-value), we say that the added variables are statistically significant after adjustment for the control variables.

Example – Depression Among Free-Living Adults.

Among free-living adults of Los Angeles County, what is the prevalence of depression and what are its correlates? In particular, in a given data set containing information on several candidate predictors, which predictors are the significant ones?

A reasonable analysis approach *for this particular example* is the following:

Step 1. Fit single predictor models. Retain for further consideration:

- Predictors with crude significance levels of association p<.25
- Predictors of *a priori* interest

Step 2. Evaluate candidate predictors for evidence of multicollinearity:

Step 3. Fit a multivariable model containing the "candidates" from step 1. Retain for further consideration

- Predictors with adjusted significance levels p < .10

Step 4. Fit the multivariable model containing the reduced set of "candidates" from step 3.

- Compare the step 3 and step 4 models using the likelihood ratio (LR) test.

Step 5. Investigate confounding. For each confounder

- Begin with the step 4 model. --- **reduced model ---**
- Fit an enhanced model that includes the suspected confounder.
  Note the estimated β's and deviance statistic values. -- **full model --**
- Assess the adjusted statistical significance of the suspected confounder using a likelihood ratio (LR) test.

- Compute relative change in the estimated β's:

$$\Delta\hat{\beta}=\left(\frac{|\ \hat{\beta}_{\text{without confounder}}-\hat{\beta}_{\text{with confounder}}\ |}{\hat{\beta}_{\text{with confounder}}}\right)\text{x}100$$

Criteria for Retention of Suspected Confounder

1. **Likelihood ratio (LR) test of its adjusted association is significant; and**
2. $\Delta\beta$ $\geq$ **15% or so.**

Step 6. Investigate effect modification

- Begin with the "near final" model identified in step 5
- Fit, one at a time, enhanced models that contain each pairwise interaction
- Assess statistical significance of each interaction using the LR test

7. Illustration
Depression Among Free-Living Adults

Source: Frerich RR, Aneshensel CS and Clark VA (1981) Prevalence of depression in Los Angeles County. *American Journal of Epidemiology* 113: 691-99.

Before you begin: Download from the course website: depress_small.dta

Background

The data for this illustration is a **subset of n=294** observations from the original study of 1000 adult residents of Los Angeles County. The purpose of the original study was to estimate the prevalence of depression and to identify the predictors of, and outcomes associated with, depression. The study design was a longitudinal one that included four interviews

In this illustration, only data from the first interview are used. Thus, this example is a cross-sectional analysis to identify the correlates of prevalent depression. Among these n=294, there are **50 events** of prevalent depression.

Codebook:

Variable	Label	Range/Codes
depressed	Case of depression	1=yes, 0 =no
age	Age, years	continuous
income	Income, thousands of dollars	continuous
female	Female gender	1=female, 0=male
unemployed	Unemployed	1=unemployed, 0=other
chronic	Chronic illness in past year	1=yes, 0=no
alcohol	Current alcohol use	1=yes, 0=no

Goal

Perform a multiple logistic regression analysis of these data to identify the correlates of prevalent depression.

Illustration for Stata Users.

Before you begin: Download from the course website: **depress_small.dta**
Launch Stata. From the toolbar: FILE > OPEN to read in the data set depress_small.dta

Preliminary. Describe the analysis sample.

```
(Depression Data Small Version)

. codebook, compact

Variable     Obs Unique      Mean  Min   Max  Label
-----------------------------------------------------------------------------------------
age          294     66  44.41497   18    89  age in years at last birthday
alcohol      294      2  .7959184    0     1  
chronic      294      2  .5068027    0     1  
depressed    294      2   .170068    0     1  
female       294      2   .622449    0     1  
income       294     30  20.57483    2    65  thousands of dollars per year
unemployed   294      2   .047619    0     1  
-----------------------------------------------------------------------------------------
            Looks reasonable.  There are no missing data.
            All of the binary variables are coded 0/1.
            The 2 continuous variables have reasonable ranges.

. * Continuous variable distributions:  by depression status
. sort depressed


. tabstat age, by(depressed) col(stat) stat(n mean sd min q max) format(%8.2f) longstub

depressed      variable |         N      mean        sd       min       p25       p50       p75       max
------------------------+--------------------------------------------------------------------------------
normal              age |    244.00     45.24     18.15     18.00     29.00     43.50     59.00     89.00
depressed           age |     50.00     40.38     17.40     18.00     26.00     34.50     51.00     79.00
------------------------+--------------------------------------------------------------------------------
Total               age |    294.00     44.41     18.09     18.00     28.00     42.50     59.00     89.00
---------------------------------------------------------------------------------------------------------
            Depressed persons tend to be younger.  Variability is comparable.


. tabstat income, by(depressed) col(stat) stat(n mean sd min q max) format(%8.2f)
longstub

depressed      variable |         N      mean        sd       min       p25       p50       p75       max
------------------------+--------------------------------------------------------------------------------
normal           income |    244.00     21.68     15.98      2.00      9.00     17.00     28.00     65.00
depressed        income |     50.00     15.20      9.84      2.00      7.00     13.00     23.00     45.00
------------------------+--------------------------------------------------------------------------------
Total            income |    294.00     20.57     15.29      2.00      9.00     15.00     28.00     65.00
---------------------------------------------------------------------------------------------------------
            Depressed persons tend to be lower income.  Also, the variability in
            income is less (sd=9.84 for depressed, sd=15.98 for non-depressed).
```

```
. *
. * Discrete variable distributions:  by depression status

. tab2 alcohol depressed, row exact
           |       depressed
   alcohol |    normal  depressed |     Total
-----------+----------------------+----------
non-drinker |        51          9 |        60
           |     85.00      15.00 |    100.00
-----------+----------------------+----------
   drinker |       193         41 |       234
           |     82.48      17.52 |    100.00
-----------+----------------------+----------
     Total |       244         50 |       294
           |     82.99      17.01 |    100.00

           Fisher's exact =                 0.705
   1-sided Fisher's exact =                 0.402
```

Depression is more prevalent among drinkers.
but this is not statistically significant.

```
. tab2 chronic depressed, row exact

                |       depressed
        chronic |    normal  depressed |     Total
----------------+----------------------+----------
          other |       126         19 |       145
                |     86.90      13.10 |    100.00
----------------+----------------------+----------
chronic illness |       118         31 |       149
                |     79.19      20.81 |    100.00
----------------+----------------------+----------
          Total |       244         50 |       294
                |     82.99      17.01 |    100.00

           Fisher's exact =                 0.089
   1-sided Fisher's exact =                 0.054
```

Depression is slightly more prevalent among the ill.

```
. tab2 female depressed, row exact

           |       depressed
    female |    normal  depressed |     Total
-----------+----------------------+----------
      male |       101         10 |       111
           |     90.99       9.01 |    100.00
-----------+----------------------+----------
    female |       143         40 |       183
           |     78.14      21.86 |    100.00
-----------+----------------------+----------
     Total |       244         50 |       294
           |     82.99      17.01 |    100.00

           Fisher's exact =                 0.004
   1-sided Fisher's exact =                 0.003
```

Depression is more prevalent among females.

```
. tab2 unemployed depressed, row exact

           |       depressed
unemployed |    normal  depressed |     Total
-----------+----------------------+----------
     other |       236         44 |       280
           |     84.29      15.71 |    100.00
-----------+----------------------+----------
unemployed |         8          6 |        14
           |     57.14      42.86 |    100.00
-----------+----------------------+----------
     Total |       244         50 |       294
           |     82.99      17.01 |    100.00

           Fisher's exact =                 0.018
   1-sided Fisher's exact =                 0.018
```

Depression is more prevalent among the unemployed.

Step 1. Fit single predictor models - Using Wald Z-score, retain predictors with significance levels < .25 or that are of a priori interest.

```
. logit depressed age
Logistic regression                               Number of obs     =        294
                                                  LR chi2(1)        =       3.10
                                                     = Likelihood Ratio Statistic
                                                     for current model ("full") v
                                                 intercept only model ("reduced")
                                                           Analogous to Overall F
                                                  Prob > chi2       =     0.0785
Log likelihood = -132.51436                       Pseudo R2         =     0.0115
(-2) ln L = 265.50287
```

Wald Z   Wald Z p-value (2 sided)

```
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0156211   .0090668    -1.72   0.085    -.0333917    .0021495
       _cons |  -.9171994   .4043128    -2.27   0.023    -1.709638   -.1247608
------------------------------------------------------------------------------
```

```
. logit depressed alcohol
Logistic regression                               Number of obs     =        294
                                                  LR chi2(1)        =       0.22
                                                  Prob > chi2       =     0.6387
Log likelihood = -133.95203                       Pseudo R2         =     0.0008
(-2) ln L = 267.90406
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     alcohol |   .1854829    .400363     0.46   0.643    -.5992142      .97018
       _cons |  -1.734601   .3615508    -4.80   0.000    -2.443228   -1.025975
------------------------------------------------------------------------------
```

```
. logit depressed chronic
Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       3.12
                                                  Prob > chi2     =     0.0775
Log likelihood = -132.50414                       Pseudo R2       =     0.0116
(-2) ln L = 265.00828
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     chronic |   .5551455   .3182777     1.74   0.081    -.0686675    1.178958
       _cons |  -1.891843   .2461058    -7.69   0.000    -2.374201   -1.409484
------------------------------------------------------------------------------

. logit depressed female
Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       8.73
                                                  Prob > chi2     =     0.0031
Log likelihood = -129.69883                       Pseudo R2       =     0.0325
(-2) ln L = 259.39766
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |    1.03857   .3766882     2.76   0.006     .3002749    1.776866
       _cons |  -2.312535   .3315132    -6.98   0.000    -2.962289   -1.662782
------------------------------------------------------------------------------

. logit depressed income
Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       8.72
                                                  Prob > chi2     =     0.0031
Log likelihood = -129.70102                       Pseudo R2       =     0.0325
(-2) ln L = 259.40204
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      income |  -.0358267   .0134794    -2.66   0.008    -.0622458   -.0094076
       _cons |  -.9375673   .2658415    -3.53   0.000    -1.458607   -.4165276
------------------------------------------------------------------------------

. logit depressed unemployed
Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       5.46
                                                  Prob > chi2     =     0.0195
Log likelihood = -131.33315                       Pseudo R2       =     0.0204
(-2) ln L = 262.6663
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  unemployed |    1.39196   .5644743     2.47   0.014     .2856108    2.498309
       _cons |  -1.679642   .1642089   -10.23   0.000    -2.001486   -1.357799
------------------------------------------------------------------------------
```

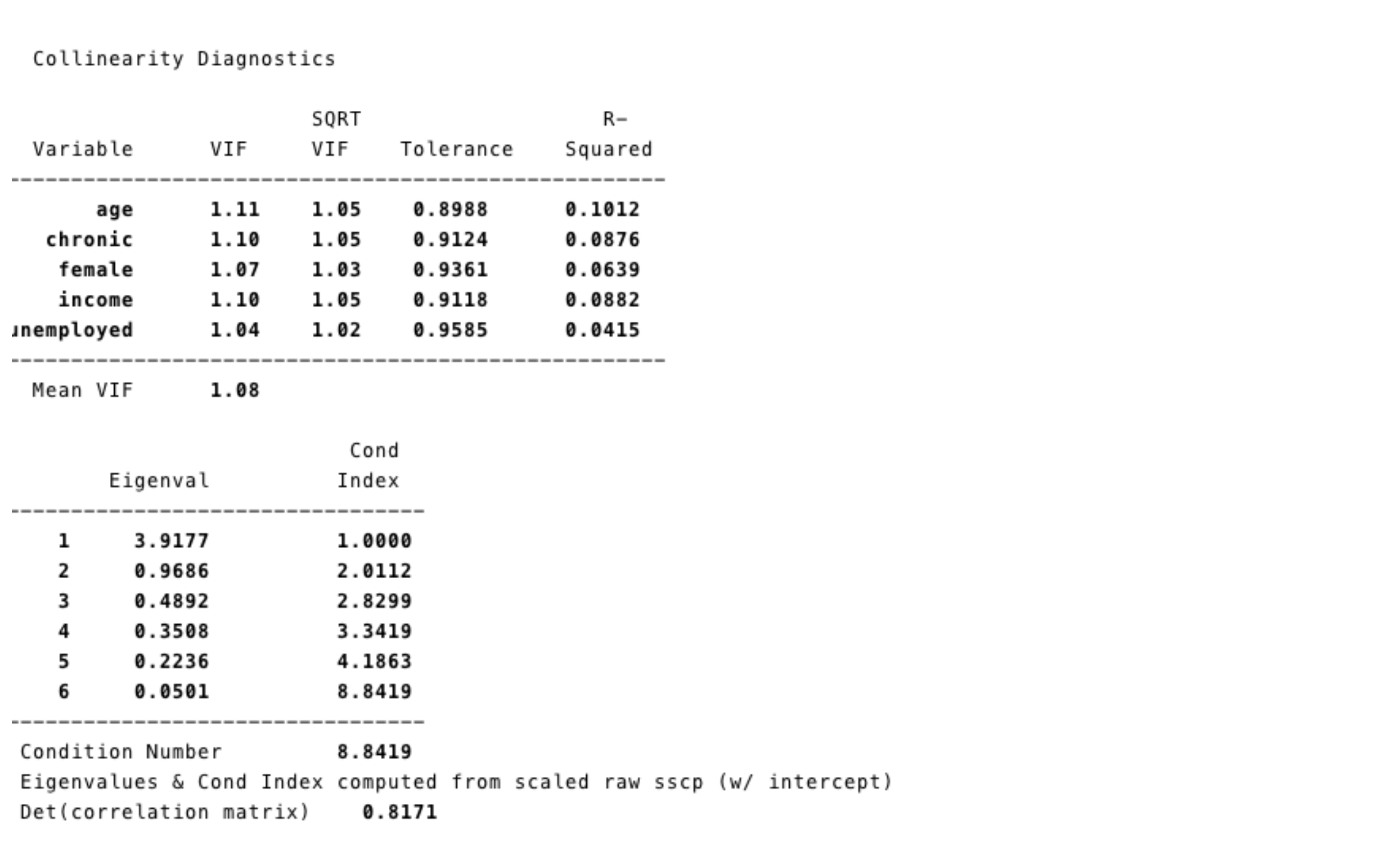
Step 1 – Summary

Predictor	Significance of Wald Z	Remark
age	.085	Consider further – pvalue is < .25
alcohol	.643	Drop
chronic	.081	Consider further. pvalue is < .25
female	.006	Consider further. pvalue is < .25
income	.008	Consider further. pvalue is < .25
unemployed	.014	Consider further. pvalue is < .25.

Step 2 – Assess candidate predictors for evidence of multicollinearity

Note – This assumes you have downloaded and installed collin.ado

```
. collin age alcohol chronic female income unemployed
```



```
 Collinearity Diagnostics

                        SQRT                   R-
 Variable      VIF      VIF    Tolerance    Squared
----------------------------------------------------
      age      1.11    1.05    0.8988      0.1012
  chronic      1.10    1.05    0.9124      0.0876
   female      1.07    1.03    0.9361      0.0639
   income      1.10    1.05    0.9118      0.0882
unemployed     1.04    1.02    0.9585      0.0415
----------------------------------------------------
 Mean VIF      1.08

                           Cond
       Eigenval           Index
---------------------------------
    1     3.9177          1.0000
    2     0.9686          2.0112
    3     0.4892          2.8299
    4     0.3508          3.3419
    5     0.2236          4.1863
    6     0.0501          8.8419
---------------------------------
 Condition Number         8.8419
 Eigenvalues & Cond Index computed from scaled raw sscp (w/ intercept)
 Det(correlation matrix)    0.8171
```

Collinearity occurs when the predictors are themselves inter-related
If extreme, this is a problem for at least 2 reasons:
1. Model is unstable   2. Model is uninterpretable
Multicollinearity problem is suggested if VIF > 10 or Tolerance < .10
Here, things look reasonable.

Step 3. Fit multiple predictor model using step 1 predictors having crude significance < .25

```
. logit depressed age chronic female income unemployed


Logistic regression                               Number of obs   =        294
                                                  LR chi2(5)      =      26.04
                                                  Prob > chi2     =     0.0001
Log likelihood = -121.04134                       Pseudo R2       =     0.0971
(-2) ln L = 242.08268  Deviance df = 294-(5) = 289
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0219383    .009494    -2.31   0.021    -.0405462   -.0033305
     chronic |    .594859   .3508664     1.70   0.090    -.0928265    1.282545
      female |   .8121316   .3968805     2.05   0.041     .0342602    1.590003
      income |  -.0320672   .0141399    -2.27   0.023    -.0597809   -.0043534
  unemployed |   1.069739   .5989254     1.79   0.074    -.1041334    2.243611
       _cons |  -1.031844   .6121359    -1.69   0.092    -2.231608    .1679207
------------------------------------------------------------------------------
```

Step 3 – Summary

Predictor	Adjusted Significance (Wald)	Remark
age	.021	Retain – pvalue is < .10
chronic	.090	For illustration purposes, let's consider dropping this variable, despite pvalue < .10 (it's close!)
female	.041	Retain – pvalue is < .10
income	.023	Retain – pvalue is < .10
unemployed	.074	Retain – pvalue is < .10.

Step 4. Fit the multivariable model containing predictors with adjusted significance levels < .10 from step 3.

We will then compare the step 3 model with the step 4 model using a likelihood ratio test.

```
. logit depressed age female income unemployed

Logistic regression                               Number of obs   =        294
                                                  LR chi2(4)      =      23.09
                                                  Prob > chi2     =     0.0001
Log likelihood = -122.51896                       Pseudo R2       =     0.0861
(-2) ln L = 245.03792  Deviance df = 294-(4) = 290

------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.018802   .0091785    -2.05   0.041    -.0367917   -.0008124
      female |    .938952   .3887469     2.42   0.016      .177022    1.700882
      income |  -.0334314   .0141518    -2.36   0.018    -.0611684   -.0056944
  unemployed |   .9634566   .5921991     1.63   0.104    -.1972324    2.124146
       _cons |  -.8968284   .5978889    -1.50   0.134    -2.068669    .2750123
------------------------------------------------------------------------------
```

By Hand: Likelihood ratio test comparing step 3 and step 4 models

$$
\begin{aligned}
\text{LR Test} &= [\,(-2)\ln(L)_{\text{REDUCED}}\,] - [\,(-2)\ln(L)_{\text{FULL}}\,] \\
&= [\,245.04\,] - [\,242.08\,] \\
&= 2.96
\end{aligned}
$$

$$\text{LR Test df} = \Delta \text{ Deviance df} = \Delta \text{ \# predictors in model} = 290\text{-}289 = 1$$

$$\text{p-value} = \Pr\{\text{Chi square with 1 degree of freedom} \geq 2.96\} = .0853$$

This is not significant. Possibly, we can drop chronic

Stata: Likelihood ratio test comparing step 2 and step 3 models.

```
. * REDUCED model using command quietly: to suppress output. Don't forget the colon.
. quietly: logit depressed age female income unemployed
. * Save results using stata command estimates store NAME
. estimates store reduced

. * FULL model using command quietly: to suppress output. Don't forget the colon.
. quietly: logit depressed age chronic female income unemployed
. * Save results using stata command estimates store NAME
. estimates store full

. * Obtain LR test using stata command lrtest
. lrtest reduced full

Likelihood-ratio test                                 LR chi2(1)  =       2.96
(Assumption: reduced nested in full)                  Prob > chi2 =     0.0856   match!
```

Step 5. Investigate confounding.

Tentatively, a “good” final model is the four predictor model with predictors: **age, female, income,** and **unemployed**. Here, we explore possible confounding of the four predictor model by the omitted variable **chronic**. Specifically, we assess **chronic** as a potential confounder using 2 criteria:

__1. Likelihood Ratio test < .10 (or .05 or threshold of choice).
__2. Relative Change in estimated betas > 15% (or threshold of choice) using the following formula:

$$\Delta\hat{\beta}=\left(\frac{|\ \hat{\beta}_{\text{without confounder}}-\hat{\beta}_{\text{with confounder}}\ |}{\hat{\beta}_{\text{with confounder}}}\right)x100$$

Fit of tentative “good” final model (shown again…)

```
. logit depressed age female income unemployed
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.018802   .0091785    -2.05   0.041    -.0367917   -.0008124
      female |    .938952   .3887469     2.42   0.016      .177022    1.700882
      income |  -.0334314   .0141518    -2.36   0.018    -.0611684   -.0056944
  unemployed |   .9634566   .5921991     1.63   0.104    -.1972324    2.124146
       _cons |  -.8968284   .5978889    -1.50   0.134    -2.068669    .2750123
------------------------------------------------------------------------------
```

Fit of enhanced model with chronic

```
. logit depressed age chronic female income unemployed
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0219383    .009494    -2.31   0.021    -.0405462   -.0033305
     chronic |    .594859   .3508664     1.70   0.090    -.0928265    1.282545
      female |   .8121316   .3968805     2.05   0.041     .0342602    1.590003
      income |  -.0320672   .0141399    -2.27   0.023    -.0597809   -.0043534
  unemployed |   1.069739   .5989254     1.79   0.074    -.1041334    2.243611
       _cons |  -1.031844   .6121359    -1.69   0.092    -2.231608    .1679207
------------------------------------------------------------------------------
```

Looking for ≥ 15% Change in Betas for Predictors in Model

Potential confounding of age, female, income, unemployed
By: chronic

$\hat{\beta}_{\text{age}}$(w/o chronic) = -.018802; $\hat{\beta}_{\text{age}}$(w chronic) = -.0219383; Change = 14.30%

$\hat{\beta}_{\text{female}}$(w/o chronic) = .938952; $\hat{\beta}_{\text{female}}$(w chronic) = .8121316; Change = 15.62%

$\hat{\beta}_{\text{income}}$(w/o chronic) = -.0334314; $\hat{\beta}_{\text{income}}$(w chronic) = -.0320672; Change = 2.32%

$\hat{\beta}_{\text{unemployed}}$(w/o chronic) = .9634566; $\hat{\beta}_{\text{age}}$(w chronic) = 1.069739; Change = 9.94%

The relative change in the beta for female is borderline at 15.6%. For parsimony, let's drop chronic.

Step 6. Investigate effect modification.

Are individuals who are both unemployed and with low income more likely to be depressed? For this illustration, we will create a new variable called **low** to capture individuals whose income is less than $10,000. Then we will create an interaction of **low** and **unemployed**. **Tip –** When assessing interaction, it is necessary to include the main effects of both of the variables contributing to the interaction. Thus, this model includes the main effects **low** and **unemployed** in addition to the interaction **low_unemployed.**

```
. *  Create new variable low
. generate low=income
. recode low (min/10=1) (10/max=0)
. label define lowf 0 "other" 1 "low (<$10K)"
. label values low lowf
. fre low
low
-----------------------------------------------------------------
                      |      Freq.    Percent      Valid       Cum.
----------------------+------------------------------------------
Valid   0 other       |        203      69.05      69.05      69.05
        1 low (<$10K) |         91      30.95      30.95     100.00
        Total         |        294     100.00     100.00
-----------------------------------------------------------------
```

```
. *  Create interaction of the two variables:  low and unemployed
. generate low_unemployed=low*unemployed
. label define lowunemployedf 0 "other" 1 "unemployed and low"
. label values low_unemployed lowunemployedf
. fre low_unemployed
low_unemployed
------------------------------------------------------------------------
                              |      Freq.    Percent      Valid       Cum.
------------------------------+-----------------------------------------
Valid   0 other               |        287      97.62      97.62      97.62
        1 unemployed and low  |          7       2.38       2.38     100.00
        Total                 |        294     100.00     100.00
------------------------------------------------------------------------
```

Hmmmm …. We have only 7 individuals who are both UNEMPLOYED and with income < $10,000

```
. * fit of near final model + low  + interaction
. logit depressed age female income unemployed low low_unemployed

Logistic regression                               Number of obs   =        294
                                                  LR chi2(6)      =      27.74
                                                  Prob > chi2     =     0.0001
Log likelihood = -120.19036                       Pseudo R2       =     0.1035

--------------------------------------------------------------------------------
     depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
           age |  -.0147588    .009597    -1.54   0.124    -.0335685    .0040509
        female |   1.036787   .3984331     2.60   0.009     .2558726    1.817702
        income |  -.0543487   .0201008    -2.70   0.007    -.0937456   -.0149517
    unemployed |   .2545214   .8759089     0.29   0.771    -1.462229    1.971271
           low |  -.9450088   .4722731    -2.00   0.045    -1.870647   -.0193705
low_unemployed |   1.544647   1.247604     1.24   0.216    -.9006125    3.989906
         _cons |  -.4746871   .6837299    -0.69   0.488    -1.814773    .8653989
--------------------------------------------------------------------------------

. *  LR test of interaction
. * reduced model
. quietly: logit depressed age female income unemployed low
. estimates store reduced

. * full model
. quietly: logit depressed age female income unemployed low low_unemployed
. estimates store full

. lrtest reduced full

Likelihood-ratio test                                 LR chi2(1)  =      1.60
(Assumption: reduced nested in full)                  Prob > chi2 =    0.2055
```

Note – The lack of statistical significance is not surprising given the small number, 7, who are both UNEMPLOYED and with income < $10,000. Again for parsimony, let's drop low.

Conclusion:

A reasonable multiple predictor model of depression in this sample contains the following predictors: **age, female, income,** and **unemployed.** Let's fit the final model one more time, in two ways: (1) using the command **logit** to obtain the prediction equation and (2) using the command **logistic** to obtain odds ratios instead of betas.

```
. logit depressed age female income unemployed
------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.018802   .0091785    -2.05   0.041    -.0367917   -.0008124
      female |    .938952   .3887469     2.42   0.016      .177022    1.700882
      income |  -.0334314   .0141518    -2.36   0.018    -.0611684   -.0056944
  unemployed |   .9634566   .5921991     1.63   0.104    -.1972324    2.124146
       _cons |  -.8968284   .5978889    -1.50   0.134    -2.068669    .2750123
------------------------------------------------------------------------------
```

→

Logit { pr[depressed=1] } = -0.90 - 0.02***age** + 0.94***female** – 0.03***income** +0.97***unemployed**

```
. logistic depressed age female income unemployed

------------------------------------------------------------------------------
   depressed | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9813736   .0090076    -2.05   0.041     .9638769    .9991879
      female |     2.5573   .9941424     2.42   0.016     1.193657    5.478777
      income |   .9671213   .0136865    -2.36   0.018     .9406648    .9943218
  unemployed |    2.62074      1.552     1.63   0.104     .8209998    8.365746
       _cons |   .4078612   .2438557    -1.50   0.134     .1263538    1.316547
------------------------------------------------------------------------------
```

Examination of this model fit suggests that, in adjusted analysis:

(1) Older age is marginally associated with lower prevalence of depression.
Relative odds (OR) of depression associated with 1 year increase = .98 (p=.04)

(2) Females, compared to males are more likely to be depressed.
Relative Odds (Odds ratio), OR = 2.6 (p=.016)

(3) Higher income is associated with lower prevalence of depression.
Relative odds (OR) of depression associated with $1K increase = .97 (p=.018)

(4) Unemployed persons, are marginally significantly more likely to be depressed.
Relative Odds, OR = 2.6 (p=.010)

8. Regression Diagnostics

With a fitted model come two applications, prediction and hypothesis tests.

- We've seen that a **prediction** is a guess of the expected outcome for a person with a particular profile of values of the explanatory variables (eg – value of vdtexpos) using the values of the estimated betas is obtained using the estimated betas:

$$\text{Predicted probability}_{\text{vdtexpos}} = \hat{\pi} = \frac{\exp\left(\hat{\beta}_0+\hat{\beta}_1[\text{vdtexpos}]\right)}{1+\exp\left(\hat{\beta}_0+\hat{\beta}_1[\text{vdtexpos}]\right)}$$

- An example of an **hypothesis test** is the hypothesis test of the significance of VDTEXPOS. The likelihood ratio test that the β for VDTEXPOS is equal to zero compares

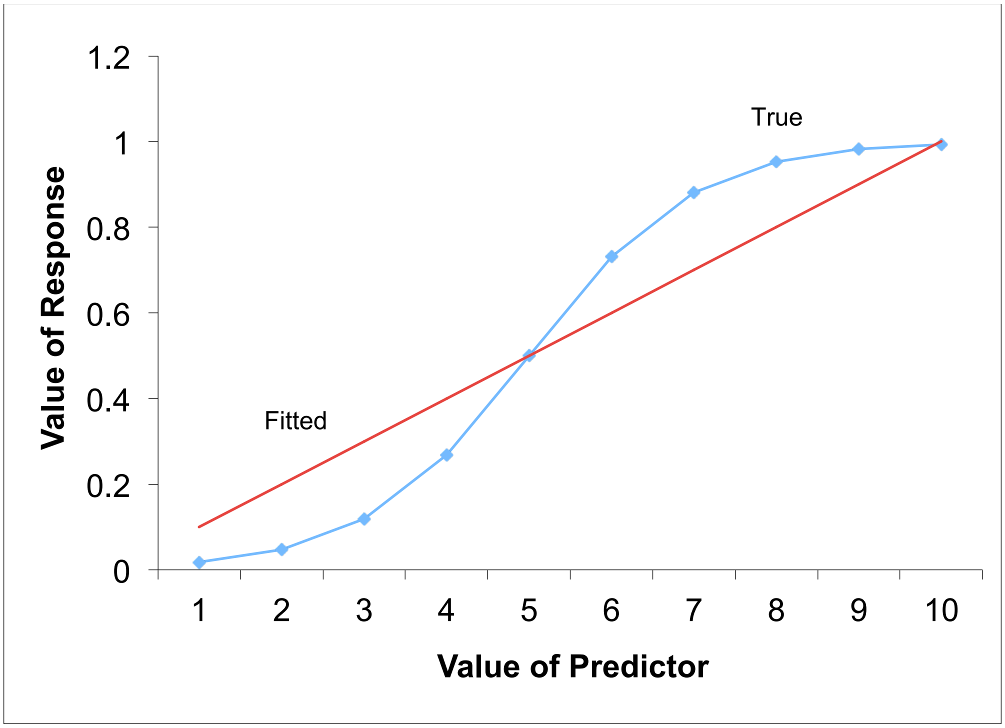
  1) the odds of SAB for exposed persons ("comparison"), versus
  2) the odds of SAB for Unexposed ("reference") persons.

Neither prediction nor hypothesis tests have meaning when the model is a poor fit to the data.

Reasons for a poor fit include the following:

(1) The wrong relationship was fit.
(2) The data include extreme values which influence too greatly the fitted line.
(3) Important explanatory variables have not been included.

We need regression diagnostics for the detection of a poor fit:



Example - The fit is poor here because the true relationship is quadratic, not linear.

We notice that the discrepancies between the observed and the fitted values are not of consistent size.

Some are large and some are small.

Goodness-of-fit assessments are formal techniques for identifying such inconsistencies.

These techniques become especially important when a picture is not possible, as when the number of predictors is greater than one.

Assessing regression model adequacy was introduced previously (Unit 2, Regression and Correlation). Regression diagnostics are of two types:

- **Systematic component**
    - Is the assumption of linearity on the ln(odds) scale correct?
    - Is the logistic model formulation a reasonably good fit?
    - Should we have fit a different model?
    - Does the fitted model predict well?

- **Case analysis**
    - Is the fitted model excessively influenced by one or a small number of individuals?

There exist methods to address each of these regression diagnostic questions.

Question	**Method of Assessment**
Is the assumption of linearity on the ln(odds) scale correct?	a. Assessment of linearity
Is the logistic model formulation a reasonably good fit?	b. Hosmer-Lemeshow test for overall goodness of fit.
Should we have fit a different model?	c. Linktest
Does the fitted model predict well?	d. Classification table e. The ROC Curve
Is the fitted model excessively influenced by one or a small number of individuals or *covariate patterns*? *Note – Here we might look at covariate patterns instead of individuals.*	f. Pregibon Delta beta statistic

a. Assessment of Linearity

A logistic regression model assumes that the **logit of the probability (π) of event occurrence** (eg – spontaneous abortion) is **linear** in the predictors X_1, X_2, … etc.

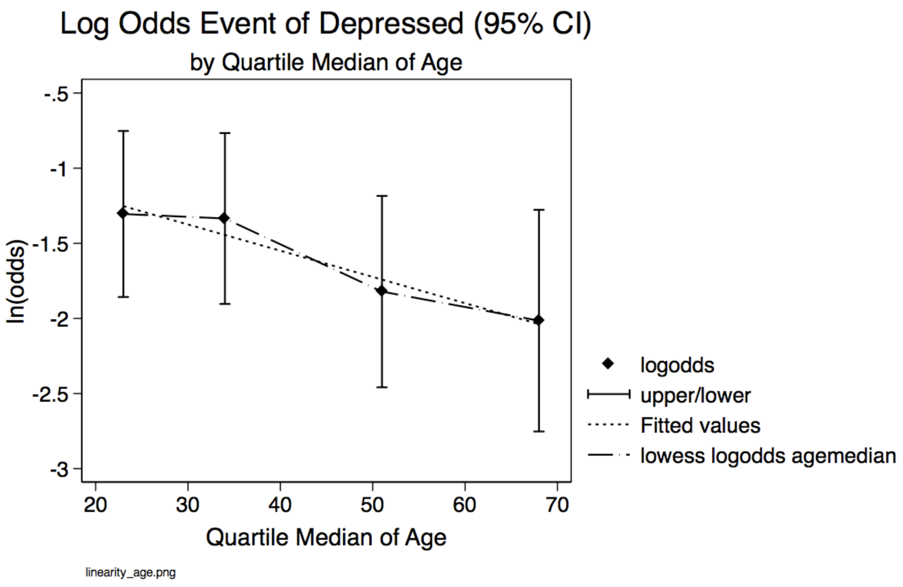
$$\text{logit}[\pi_x] = \text{Logit}[E(Y)] = \ln\left[\frac{\pi_x}{1-\pi_x}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_5 X_5 + \beta_6 X_6$$

Violation of the assumption of linearity of the logit in a continuous predictor can lead to incorrect estimates and incorrect conclusions. A variety of approaches are available for assessing the assumption of linearity in logistic regression but are *beyond the scope of these notes.*

A graphical assessment of linearity of Y = logit with changes in X=predictor can be performed in Stata (we'll do this in lab). It involves five steps

1. Collapse the predictor values of X into groups (eg; quartiles)
2. In each group, obtain the median value of the predictor variable X.
3. In each group, obtain the observed proportion experiencing the event Y.
4. In each group, obtain the observed logit [proportion experiencing event]
   Tip – Obtain 95% CI limits as well.
5. Produce a two-way plot of X=midpoint versus Y=logit, perhaps with some overlays.

Example – continued.



Not bad! The plot looks reasonable enough that it is okay to model the logit linearly in age.

b. The Hosmer-Lemeshow Test of Goodness-of-Fit

The **Hosmer-Lemeshow Goodness of Fit Test** compares observed versus predicted counts of outcome events in each of several "meaningful" subgroups of the data, in a manner similar to the Chi Square Goodness of Fit Test introduced in Unit 4, Categorical Data. If the fit is good (null hypothesis is true), the observed and (model based) expected counts will be close and their differences will be small. The actual test statistic is a sum of (observed – expected)/expected^2 and is distributed chi square under the null hypothesis.

Null Hypothesis: "Good fit" is indicated by similar counts of observed and predicted counts in all the subgroups.

The difference between the two counts is then close to zero.

The sum, taken over the subgroups, is also small.

The Groups Used in a Hosmer-Lemeshow Test are defined by the predicted probabilities

Within each group, members have similar predicted probabilities of outcome event.

The most commonly used groups are 10 subgroups defined by deciles of predicted.

1st subgroup: This is the 1/10th of sample of persons who have the **lowest** predicted probabilities of outcome event.

2nd subgroup: This is the next 1/10 of sample of persons. These persons have the **next lowest** predicted probabilities of outcome event.

And so on ….

10th subgroup: This is the last 1/10 of sample of persons. These persons have the **highest** predicted probabilities of outcome event.

Hosmer-Lemeshow Goodness of Fit Test

H_O: The current model is a "good" fit to the data.
H_A: not.

$$\chi^2_{\text{Hosmer-Lemeshow; DF=\# groups-2}} = \sum_{\text{decile of risk}} \left\{ \frac{\left[\text{Observed count - Predicted count}\right]^2}{\text{Predicted count}} \right\}$$

Rejection occurs for large values of the chi square statistic with associated small p-values

Calculation of observed and (model fit) predicted counts:

$$\text{Observed count} = \text{Actual number of events in decile}$$

$$\text{Predicted count} = (\text{\# in group})\ (\text{Average predicted probability})$$

When the null hypothesis of a "good" fit is true,

$\chi^2_{Hosmer-Lemeshow}$ is distributed Chi Square, approximately. With df= (# groups) – (2)

For example, with 8 groups, the degrees of freedom = 6

Large values of this statistic suggest a poor fit.

Statistically significant values of the Hosmer-Lemeshow statistic evidence ONLY that the fit is poor. We do not learn why. Further assessments are necessary to understand their nature.

Stata Illustration

Example: Depression Among Free-Living Adults – *continued.*

```
. *-- must have fit the "final" model before doing test --*
. logit depressed age female income unemployed

          -- some output omitted –

. *-- Use command estat gof to obtain Hosmer Lemeshow Test --*
. estat gof, group(8) table

Logistic model for depressed, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.0598 |     2 |   1.5 |    35 |  35.5 |    37 |
  |     2 | 0.0804 |     2 |   2.6 |    35 |  34.4 |    37 |
  |     3 | 0.1180 |     4 |   3.7 |    33 |  33.3 |    37 |
  |     4 | 0.1575 |     5 |   5.1 |    31 |  30.9 |    36 |
  |     5 | 0.1800 |     5 |   6.3 |    32 |  30.7 |    37 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.2232 |     8 |   7.5 |    29 |  29.5 |    37 |
  |     7 | 0.3034 |    11 |   9.7 |    26 |  27.3 |    37 |
  |     8 | 0.6457 |    13 |  13.6 |    23 |  22.4 |    36 |
  +--------------------------------------------------------+

       number of observations =       294
             number of groups =         8
      Hosmer-Lemeshow chi2(6) =         0.97
                  Prob > chi2 =         0.9867
```

KEY -

- **Column "TOTAL"** – These are the stratum specific sample sizes.

- **Column "PROB"** – The groups are defined by the predicted probabilities. Individuals in group 1 have the "lowest" predicted probabilities and range from a 0% probability to a 5.98% probability. Individuals in group 2 have the "next lowest" predicted probabilities. These range from 5.98% to 8.04%. And so on.

- **Columns "OBS_1 and EXP_1"** – These are the observed and expected counts of **depressed= yes** in each group. For example, in group 4, there were 5 observed events of depressed=yes compared to a logistic model expected number of events of depressed=yes equal to 5.1.

- **Columns "OBS_0 and EXP_0"** –. These are the observed and expected counts of **depressed= no** in each group. For example, in group 4, there were 31 observed events of depressed=no compared to a logistic model expected number of events of depressed=no equal to 30.9

- The Hosmer_Lemeshow test (p=.9867) suggests no statistically significant departure from a good fit. The null hypothesis of "good fit" is NOT rejected. **Good news!**

c. The Linktest

The **Link Test** is an example of a **specification test**.

Like the Hosmer-Lemeshow statistic, the **Link Test** is a simple check of the fitted model. It assesses whether or not the fitted model is adequate fit (null hypothesis) to the data or, if not, if there is still some additional modeling that needs to be done (alternative hypothesis). The crudeness of the Link Test is that what we learn is limited. If the null hypothesis is rejected, we know only that some alternative modeling is needed, but we don't know what alternative modeling is needed.

Link Test

H_O: The current model is an adequate fit to the data.
H_A: Alternative modeling is needed.

A Likelihood Ratio (LR) Test is performed and compares a "null hypothesis" adequate model (reduced) with an "alternative hypothesis enhanced (full) model:

$$\text{Reduced: } \operatorname{logit}[\pi] = \beta_0 + \beta_1[\hat{\pi}_{\text{model}}]$$

$$\text{Full: } \operatorname{logit}[\pi] = \beta_0 + \beta_1[\hat{\pi}_{\text{model}}] + \beta_2[\hat{\pi}^2_{\text{model}}]$$

Thus,

H_O: $\beta_2 = 0$

H_A: not

Key -

$\hat{\pi}_{\text{model}}$: This is the predicted probability from our model; we hope this is significant.

$\hat{\pi}^2_{\text{model}}$: If the null is true (the model is adequate),this should be non-significant.

Rejection of the null occurs for large values of the LR Test and associated small p-values.

Stata Illustration

Example: Depression Among Free-Living Adults – *continued.*

```
. *-- Here, too - must have fit the "final" model before doing test --*
. logit depressed age female income unemployed

          -- some output omitted -

. * --  Linktest --*
. linktest

          -- some output omitted -


------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   1.075812   .6569617     1.64   0.102    -.2118091    2.363434
      _hatsq |   .0251889   .2041306     0.12   0.902    -.3748998    .4252775
       _cons |   .0438939   .5070363     0.09   0.931     -.949879    1.037667
------------------------------------------------------------------------------
```

$\_hat = \hat{\pi}_{\text{model}}$: This is marginally significant (p=.10); perhaps we'd hoped for better. But okay.

$\_hatsq = \hat{\pi}^2_{\text{model}}$: This is non-significant (p=.90). Good news.

The Link Test (p=.902) suggests no statistically significant departure from model adequacy. The null hypothesis of "model adequacy" is NOT rejected. **Good news!**

d. The Classification Table

Rationale

- Just because the fitted model is a good fit overall doesn't mean that individual predictions are correct most of the time.

- The classification table, and associated plots, are useful in a selected analysis setting:

  The investigator wishes to use the fitted equation to make predictions as to which group (event or non-event) a person belongs, based on his/her covariate profile.

Method

- For each individual, there are two quantities to work with
  - Actual outcome: Yes/No indicator of event occurrence
  - Estimated probability of event: Between 0 and 1

- Choose a threshold probability for event declaration by model.
  - Default is usually 0.5
  - This can be reset.
  - Consideration of several permits construction of ROC curve.

A separate classification table is produced for each cut-off you select

		Observed (True)		
		Event	**Non-Event**	
Predicted	**Event**			
	Non-Event			

Example:
Suppose that for subject id=103 observed event = YES predicted probability = .68

When cut-off=.60 observed event is still = YES Now, predicted event = YES Because .68 > .60
When cut-off=.70 observed event is still = YES But, predicted event = NO Because .68 < .70

Stata Illustration

Example: Depression Among Free-Living Adults – *continued.*

```
. *-- Check. Must have fit the “final” model first --*
. logit depressed age female income unemployed

. *--- default cutoff = .5 So no need to specify the cutoff value --
. estat classification

Logistic model for depressed

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         2             1  |          3
     -     |        48           243  |        291
-----------+--------------------------+-----------
   Total   |        50           244  |        294

Classified + if predicted Pr(D) >= .5
True D defined as depressed != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    4.00%
Specificity                     Pr( -|~D)   99.59%
Positive predictive value       Pr( D| +)   66.67%
Negative predictive value       Pr(~D| -)   83.51%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.41%
False - rate for true D         Pr( -| D)   96.00%
False + rate for classified +   Pr(~D| +)   33.33%
False - rate for classified -   Pr( D| -)   16.49%
--------------------------------------------------
Correctly classified                        83.33%   = (2+243)/294 = .8333
--------------------------------------------------
```

Key and some checks:

- **Concordance** is (2+243)/294 = .8333, or 83.33% This matches the “correctly classified – 84.33%”
- Different software packages produce different amounts of detail. STATA happens to provide lots of detail.
- Check: Sensitivity = % of true event that is predicted to be event = 2/50 = 0.50, or 4%
- Check: Predictive value positive = % of predicted positive that are actual events = 2/3 = .667, or 66.67%
- Check: Predictive value negative = % of predicted negative that are actual NON events = 243/291, 83.51%

Example - Stata allows different cut-offs

```
. *--- cutoff=0.6 --
. estat classification, cutoff(.6)
Logistic model for depressed

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |         1             1  |          2
     -     |        49           243  |        292
-----------+--------------------------+-----------
   Total   |        50           244  |        294

Classified + if predicted Pr(D) >= .6
True D defined as depressed != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)    2.00%
Specificity                     Pr( -|~D)   99.59%
Positive predictive value       Pr( D| +)   50.00%
Negative predictive value       Pr(~D| -)   83.22%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    0.41%
False - rate for true D         Pr( -| D)   98.00%
False + rate for classified +   Pr(~D| +)   50.00%
False - rate for classified -   Pr( D| -)   16.78%
--------------------------------------------------
Correctly classified                        82.99%
--------------------------------------------------
```

```
. *--- cutoff=0.1 --
. estat classification, cutoff(.1)
Logistic model for depressed
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        43           160  |        203
     -     |         7            84  |         91
-----------+--------------------------+-----------
   Total   |        50           244  |        294

Classified + if predicted Pr(D) >= .1
True D defined as depressed != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   86.00%
Specificity                     Pr( -|~D)   34.43%
Positive predictive value       Pr( D| +)   21.18%
Negative predictive value       Pr(~D| -)   92.31%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   65.57%
False - rate for true D         Pr( -| D)   14.00%
False + rate for classified +   Pr(~D| +)   78.82%
False - rate for classified -   Pr( D| -)    7.69%
--------------------------------------------------
Correctly classified                        43.20%
--------------------------------------------------
```

e. The ROC Curve

One of the uses of a fitted logistic model is to make predictions for new individuals; eg – **is this new person predicted to experience the event or not?**

An ROC curve ("Receiver-Operating Characteristic) is a visual display of the overall performance of a fitted logistic model and its associated equation for predicted probabilities. It takes into consideration that there are **two kinds of errors of prediction**: (1) a true event is predicted to be a non-event (false negative) and (2) a true non-event is predicted to be an event (false positive, which is the same as 1 - specificity).

For various choices of "cut-off" **(recall - this is the value above which a predicted probability is classified as a predicted event)** an ROC curve is plot of X=false positive against Y = true positive values for various choices of "cutoff":

"Cutoff"	**.10**	**.20**	**etc**	**.80**	**.90**
X = false positive = 1 - specificity					
Y = correct positive = sensitivity					

Key

- In a real world application, **the choice of "cutoff" has real world implications** as when a predicted event=yes prompts the initiation of treatment.
- **A diagonal line with slope=1 is a reference line**. It represents the ROC curve for test that performs no better than the **flip of a coin**.
- The area under the ROC curve is often denoted c-statistic. It has a defined meaning:

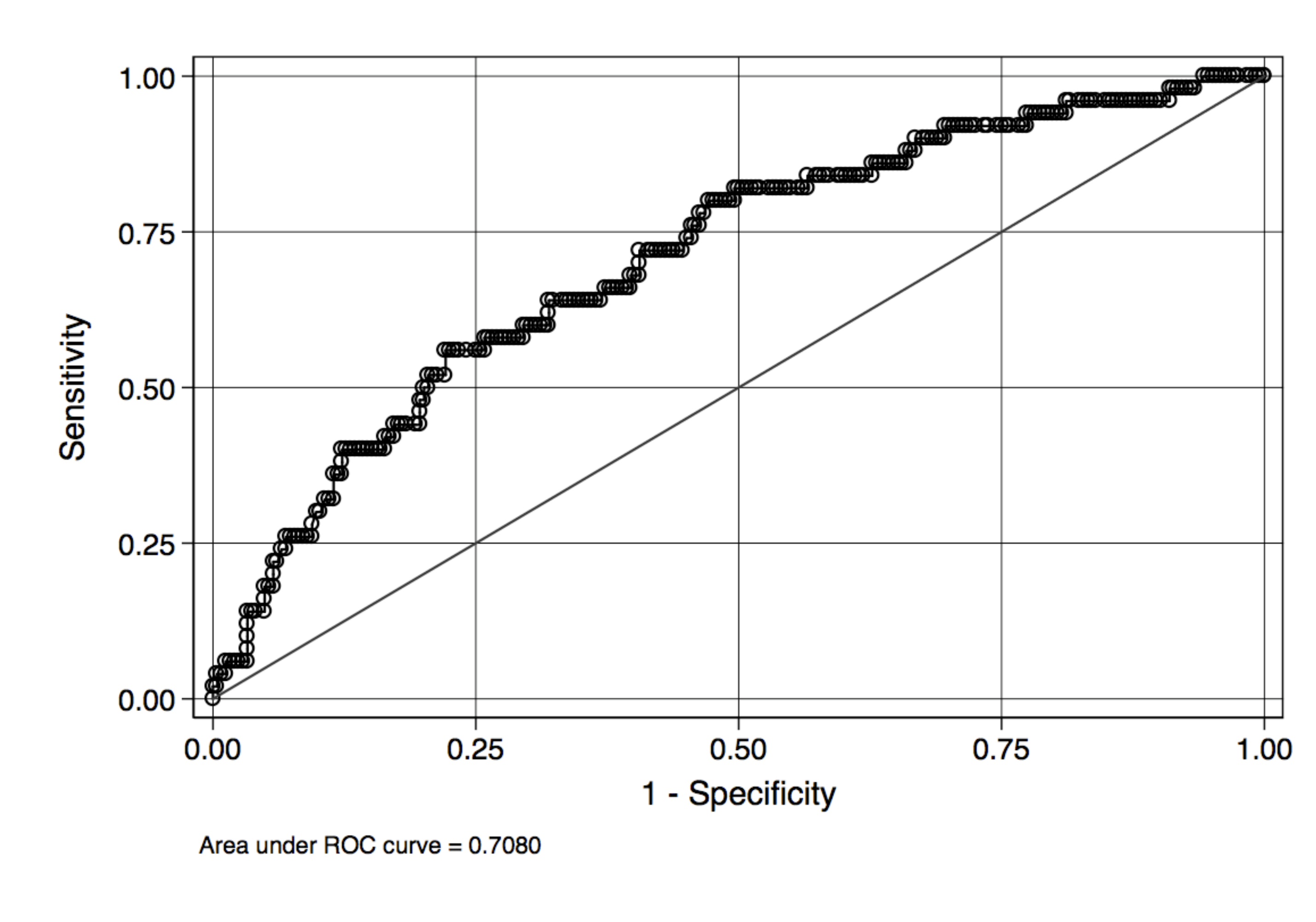
ROC Curve

c-statistic = Overall % correctly classified

Stata Illustration

Example: Depression Among Free-Living Adults – *continued.*

```
. *-- Again, be sure to have fit the "final" model first --*
. logit depressed age female income unemployed

. *-- obtain predicted logits
. predict xb, xb

. *-- obtain ROC Plot
. lroc
```



Area under ROC curve = 0.7080

Key -

- Recall - The straight line with slope =1 is a reference line; it corresponds to the ROC curve where chance alone is operating (coin toss with probability heads = .50)

- **ROC c-statistic = .7080** says that the overall % who are correctly classified is 70.8%. This is not very impressive, actually. We typically hope to do better.

f. The Pregibon Delta Beta Statistic

Recall the Cook's Distance Statistic introduced in unit 2, Regression and Correlation. This statistic provides a measure of the extent to which inclusion or non-inclusion of an individual changes the estimated betas.

The plot is of X=Subject ID versus Y=Cook's Distance
Spikes in the plot identify individuals whose inclusion are influential on the fit.

The analogue in logistic regression is the Pregibon Delta Beta Statistic, dbeta. The formula is beyond the scope of this course. However, a feel for it is the following:

dbeta = function of { standardized difference in betas w deletion of individual or deletion of covariate pattern }

The Pregibon Delta Beta Statistic can be computed for study individuals or for covariate patterns instead of study id.

- A **covariate pattern** is a unique profile (or combination) of values on the variables.

- The **maximum number** of covariate patterns in a data set occurs when every individual is unique in his/her pattern of values of the predictors. In this extreme case, the number of covariate patterns = sample size = n.

- Often, however, the same covariate pattern is shared by more than one individual (eg – 4 subjects have age=50, sex=male, exposure=yes). Thus, often, the **number of covariate patterns < n**.

The plot is of X=predicted probability versus Y=dbeta

- **Small values of dbeta: individual or covariate pattern is not influential**
  Small: dbeta values less than 1 or so, approx

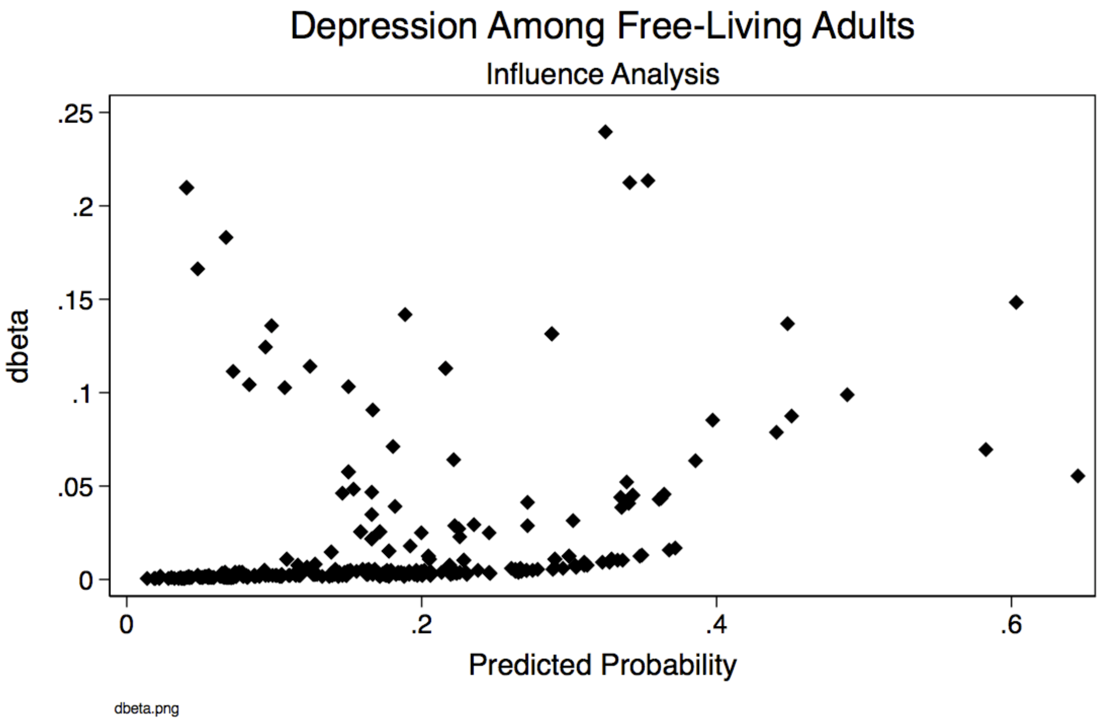
- **Large values of dbeta: individual or covariate pattern is influential**
  Large: dbeta values > 1

Tip – **Regardless of the magnitudes of the dbeta, be on the look out for spikes**
Spikes are suggestive of comparative influence

Stata Illustration

Example: Depression Among Free-Living Adults – *continued.*

```
. *-- Again, be sure to have fit the “final” model first --*
. logit depressed age female income unemployed

. *-- Pregibon Delta Beta Plot

. * -- Xaxis = predicted probabilities using variable named phat
. *-- use command predict NAME, p
. predict phat, p
. label variable phat "Predicted Probability"


. * -- Yaxis = Pregibon delta beta values using variable named dbeta
. *-- use command predict NAME, dbeta
. predict dbeta, dbeta
. label variable dbeta "Pregibon Delta Beta"

. *-- Plot --*
. graph twoway (scatter dbeta phat, msymbol(d)), title("Depression Among Free-Living
Adults") subtitle("Influence Analysis") ytitle("dbeta") caption("dbeta.png",
size(vsmall))
```



- The dbeta values are all less than .25, suggesting the absence of influential points. **Good news!**

9. Example - Disabling Knee Injuries in the US Army

Source: **Sulsky SI, et al . Risk Factors for Disability Discharge from the US Army Related to Occupational Knee Injury (2000).**

Background:

The strongest correlate of lost time from work, lost productivity, and lost working years of life is occupational injuries.

Occupational activities have been found to be associated with knee disorders.

Poorly understood, however, are the differences in risk of knee disorders associated with socio-demographic versus occupational task characteristics.

Better understanding of the socio-demographic variations in risk of occupational knee injury is important to future studies of occupational risks.

Therefore, Sulsky et al conducted a case-control study to investigate selected socio-demographic risk factors for occupational knee injury in the US Army.

Research Question:

What are the separate and joint effects of gender, age, and race/ethnicity in the odds of disabling knee injury among enlisted Army personnel on active duty between 1980 and 1994?

Design: **Nested case-control** investigation of knee related disability within the occupational cohort of enlisted US Army personnel on active duty between 1980 and 1994.

Total Army Injury and Health Outcomes Data Base (TAIHOD)

2.1 million men
283,000 women
≈ 2.4 million

↓

Data Library	
Cases	**Controls**
First record of any of 11 eligible codes 7868 men 860 women 8728 total	Density sampling[*] of TAIHOD by year, separately for each gender 11,758 men (control:case = 1.5:1) 5,109 women (control:case = 6:1) 16,867 Total (control:case = 2:1)

↓

Analysis Sample

	Cases	Controls	Control:Case
Women	860: all cases	2580: density sampling by year	3:1
Men	1005: equal random sampling by year over 15 years (67/year)	3009: equal random sampling by year over 15 years (201/year)	3:1
Total	1865	5589	7454

[*] ***For the unfamiliar* - *Density Sampling by Year*: For each year, controls were drawn in proportion to the number of cases for that year. (E.g. – A year with 2 cases and 3:1 sampling of controls yields 6 controls for that year.)**

Estimated Distribution of Risk Factors:
Age and Race/Ethnicity, by Gender

Our estimates will have to take into account the method of sampling employed. How does this work?

Let's look at a simple illustration. Suppose ….

Men	Women
Source Population, N=2000 Size of random sample, n=100 Probability[inclusion] = 100/2000 = .05 Weight per person included = 1/.05 = 20 Each man in the sample represents 20 men in the source population.	Source Population, N=1000 Size of random sample, n=100 Probability[inclusion] = 100/1000 = .10 Weight per person included = 1/.10 = 10 Each woman in the sample represents 10 women in the source population.
The number of men <21 years of age in the sample is # = 50. Therefore, estimated number of men <21 years of age in the source population is 50 x (weight=20) = 1000	The number of women <21 years of age in the sample is # = 25 Therefore, estimated number of women <21 years of age in the source population is 25 x (weight=10) = 250

What is the *overall* relative frequency of age < 21 years?

Unweighted estimate describes the sample: (50+25)/200 = 37.5%.
Weighted estimate describes the population: = (1000+250)/3000 = 41.7%

REMINDER
When a study calls for stratified sampling with disproportionate sampling of selected groups, estimates of population characteristics must take sample weights and stratified sampling into account.

**Estimated Distribution of Risk Factors:
Age and Race/Ethnicity, by Gender**

			Relative Frequency* Among	
			Cases	Controls
MEN	**Age**	**<21**	15	20
		21-23	19	19
		23-26	26	20
		26-30.36	20	18
		30.36-54	19	23
	Race/Ethnicity	**Unknown**	0	0
		White	71	62
		Black	22	29
		Other	7	9
WOMEN	**Age**	**<21**	19	19
		21-23	18	20
		23-26	19	22
		26-30.36	24	23
		30.36-54	20	16
	Race/Ethnicity	**Unknown**	0.2	0
		White	68	47
		Black	26	45
		Other	6	8

- **Estimated relative frequencies take sample weights and stratified sampling into account.**

We'll use quintiles of age.

Race/Ethnicity will be categorized as White/Non-White.

A multivariable logistic regression model analysis will explore the separate and joint associations with disabling knee injury of age, gender, and race/ethnicity.

Recall the Research Question:

What are the separate and joint effects of gender, age, and race/ethnicity in the odds of disabling knee injury among enlisted Army personnel on active duty between 1980 and 1994?

We are especially interested in identifying possible interactions.

- This analysis is to guide future analyses of occupational risk factors.
- A "traditional" analysis of occupational risk factors might simply control for age, gender, and race/ethnicity.
- If interactions exist among age, gender, and race/ethnicity, inclusion of only main effects might lead to incorrect inferences.

Therefore, the analysis plan seeks to estimate

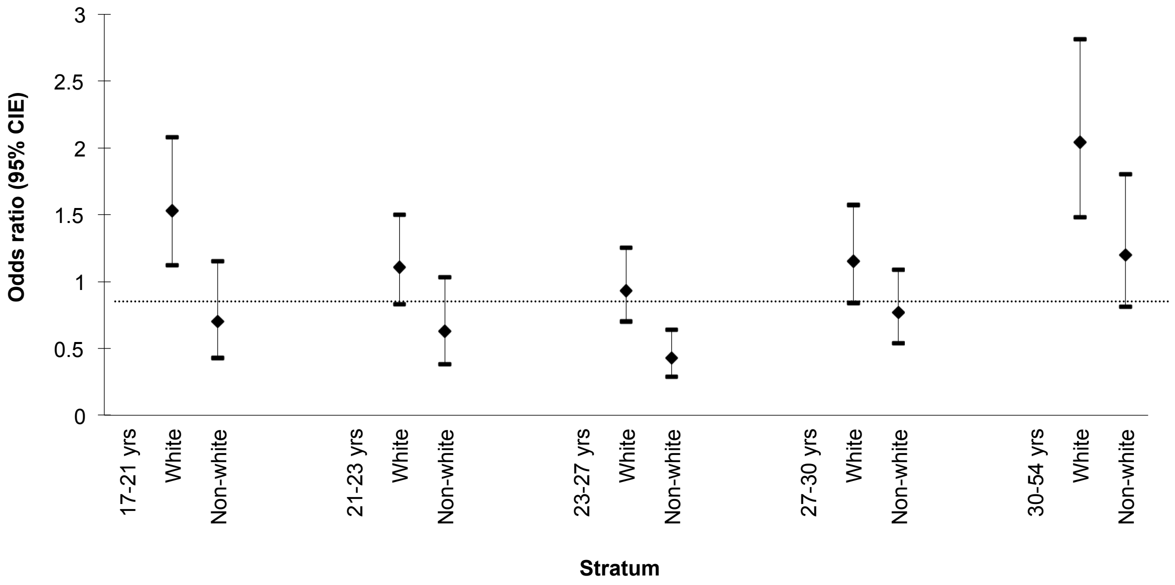
- The separate effects of gender on risk of disabling knee injury among groups defined by age | and race/ethnicity.

  e.g. – Is the effect of gender different among young workers compared to the effect of gender among older workers?

- The separate effects of increasing age on risk of disabling knee injury among groups defined by gender and race/ethnicity.

  e.g. – Is the effect of increasing age different among men and women?

Figure 1: Relative odds of discharge for disabling knee injury among enlisted women compared to men, stratified by age (quintiles) and race.



- **Among Whites:**

Women are at *higher* risk of disabling knee injury than men at all ages *except* among persons aged 23-27.

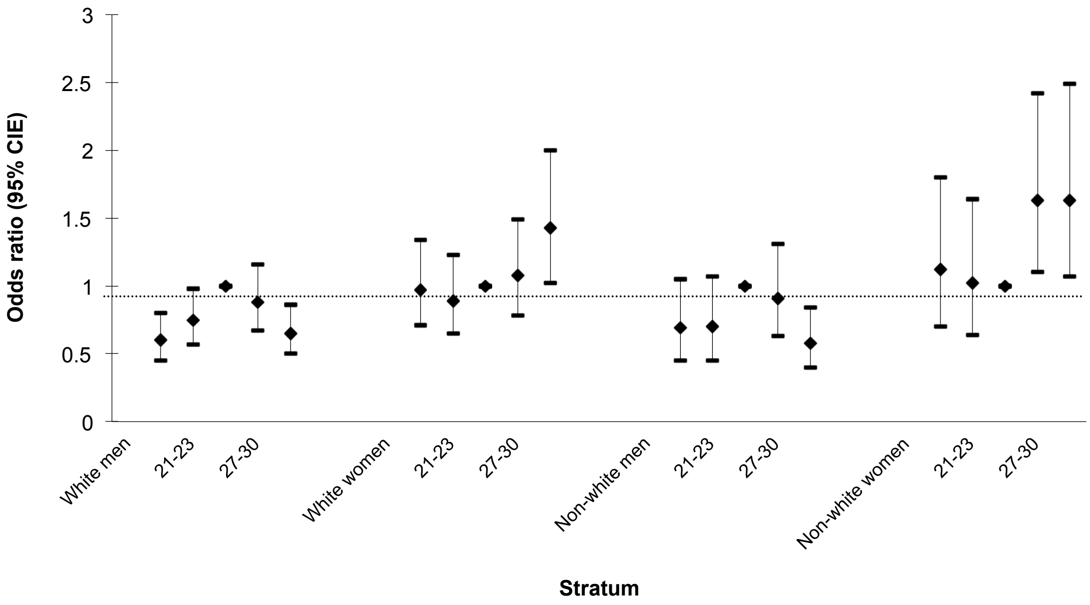
The gender effect is greatest among the youngest (17-21 years) and oldest (30-54) persons. ("U" shape)

- **Among non-Whites:**

Women are at *lower* risk of disabling knee injury than men at all ages *except* among persons aged 30-54.

The gender effect is greatest among persons in the middle age group (23-27 years). ("U" shape)

Figure 2: Relative odds of discharge for disabling knee injury with increasing age, stratified by sex and race.



note: The reference age group is age 23-27 years.

- **Among Men:**

With *increasing age*, the change in risk of disabling knee injury exhibits a "∩" pattern.

The "∩" pattern among Whites is *stronger* than the "∩" pattern among non-Whites.

- **Among Women:**

With *increasing age*, the change in risk of disabling knee injury exhibits a "J" pattern.

The "J" pattern among Whites is *more precise* than the "J" pattern among non-Whites.

This example is a nice illustration of the distinction between *confounding* and *effect modification*

CAUTION!!

Confounding and effect modification are not simply about sampling and variations in nature. Their identification in statistical analysis is also a function of the choice of scale of measurement.

In the analysis of the relative odds of disabling knee injury, we are actually speaking of

Odds ratio confounding
Odds ratio modification

A (odds ratio) relationship between "E" and "D" that is confounded by X means:

1) X is related to both "E" and "D"

2) The unadjusted association between "E" and "D" is spuriously large or small because of the confounding effects of X

3) However, at each level of X, the association between "E" and "D" is the same.

4) A logistic regression analysis of the "E"-"D" relationship should include the predictor variable X.

A (odds ratio) relationship between "E" and "D" that is modified by X means:

1) X is related to both "E" and "D"

2) With changes in the level of X, the association between "E" and "D" changes also.

3) A logistic regression analysis of the "E"-"D" relationship should reveal these changes with X through the inclusion of "E"-"X" interactions.

Appendix
Overview of Maximum Likelihood Estimation

The method of maximum likelihood estimation is used to obtain “good” guesses of the values of the regression coefficients, $\beta_0 \cdots \beta_6$.

What do we mean by “good”?

1) Recall that, in linear model regression, “good” was conceptualized as obtaining guesses of $\beta_0 \cdots \beta_6$ that make as small as possible the total of the vertical distances between the observed data Y and the fitted values $\hat{Y}$. We use the method of least squares and choose guesses, represented as $\hat{\beta}_0 \ldots \hat{\beta}_6$, which minimize the residual sum of squares:

$$\text{Residual sum of squares} = \sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{N}\left(Y_i - \left[\hat{\beta}_0 + \ldots + \hat{\beta}_6 x_6\right]\right)^2$$

When the distribution of the errors is normal, we have a very nice result:

Method of least squares = Method of maximum likelihood; where

“maximum likelihood estimation” is described below.

2) In logistic model regression, “good” is conceptualized as obtaining guesses of $\beta_0 \cdots \beta_6$ which make as large as possible the likelihood of obtaining the observed data. This is the method of maximum likelihood.

A Feel for Maximum Likelihood Estimation

A box contains two coins, A and B. One is selected.

"A" is fair and lands "heads" with probability π =.50.

"B" is not fair. It lands "heads" with probability π =.67.

Game: Toss the coin n=20 times. Note how many times the coin lands "heads". Call this X. Suppose X=15.

Question: Which choice of π , .50 or .67, maximizes the chances that the coin lands "heads" 15 times?

$\binom{20}{15} \pi^{15} (1-\pi)^{20-15}$	**π=.50**	**π=.67**
Likelihood, L **L = Prob [X=15]**	**=.10**	**=.45**

Review: The expression $\binom{20}{15}$ is a binomial coefficient and represents the number of ways to choose 15 items from 20. It is equal to 20!/[15! 5!].

There is a 10% chance of 15 "heads" when π=.50. There is a 45% chance of 15 "heads" when π=.67.

Even though scenarios of low probability do occur, the maximum likelihood estimate of the unknown probability of heads is chosen to be the one that makes as large as possible, the likelihood of the actual data.

⇒ The maximum likelihood guess of π=.67.

Overview of Maximum Likelihood Estimation in Logistic Regression

Preliminaries

(1) It is assumed that the n outcomes $Y_1, \ldots., Y_n$ are independent

(2) It is also assumed that each Y_i is the outcome of a Bernoulli (π_i) trial

(3) We'll use the notation L_i to represent each individual "likelihood", also called the probability density:

$$\begin{aligned} L_i &= \text{Probability}[Y_i = y_i] \\ &= \pi_i^{y_i} \left(1-\pi_i\right)^{1-y_i} \\ &= \left[\frac{\pi_i}{1-\pi_i}\right]^{y_i} \left(1-\pi_i\right)^1 \end{aligned}$$

(4) We'll use the notation **L** to represent the likelihood of all n observations in the data. This is also called the "probability density of the data"

L = likelihood of the data

$$\begin{aligned} L &= \text{Probability}[Y_1=y_1,\ Y_2=y_2,\ \ldots,\ Y_p=y_p] \\ &= \text{Probability}[Y_1=y_1]\ \text{Probability}[\ Y_2=y_2]\ \ldots\ \text{Probability}[\ Y_p=y_p] \text{ by independence} \\ &= \prod_{i=1}^{n} \text{Probability}[Y_i=y_i] \\ &= \prod_{i=1}^{n} L_i \end{aligned}$$

(4) The logistic model with predictors $\beta_0, \beta_1, \ldots., \beta_p$ is defined

$$\ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

$x_{1i} =$ value of the variable x_1 for the "ith" person, etc.

(5) The logistic model with predictors β_0, β_1, …., β_p also means that

$$\ln\left(1-\pi_{\underline{x}}\right) = \ln\left[\frac{1}{1+\exp\left(\beta_0+\beta_1 x_1+...+\beta_p x_p\right)}\right]$$

$$= \ln[1] \ - \ \ln\left[1+\exp\left(\beta_0+\beta_1 x_1+...+\beta_p x_p\right)\right] \text{because } \ln(a/b) = \ln(a) - \ln(b)$$

$$= 0 - \ln\left[1+\exp\left(\beta_0+\beta_1 x_1+...+\beta_p x_p\right)\right] \text{because } \ln[1]=0$$

$$= \ - \ln\left[1+\exp\left(\beta_0+\beta_1 x_1+...+\beta_p x_p\right)\right]$$

Overview

- Maximum likelihood estimation of β_0, β_1, …., β_p is accomplished by maximizing the natural logarithm of the likelihood L of the data.

- We'll let L (β) = ln { L } represent the natural logarithm of the data under the assumption of the logistic regression model.

Solution for L (β).
This is the function of the data that we seek to maximize with respect to $\beta_0, \beta_1, \ldots, \beta_p$

L (β) = ln { L }

$$= \ln\left[\prod_{i=1}^{n} L_i\right]$$

$$= \sum_{i=1}^{n}\{\ln[L_i]\} \quad \text{because } \ln[(a)(b)] = \ln(a) + \ln(b)$$

$$= \sum_{i=1}^{n} \ln\left\{\left[\frac{\pi_i}{1-\pi_i}\right]^{y_i}(1-\pi_i)^1\right\} \quad \text{by preliminary \#3}$$

$$= \sum_{i=1}^{n}\left\{\ln\left[\frac{\pi_i}{1-\pi_i}\right]^{y_i} + \ln(1-\pi_i)^1\right\} \quad \text{again because } \ln[(a)(b)] = \ln(a) + \ln(b)$$

$$= \sum_{i=1}^{n}\left\{y_i \ln\left[\frac{\pi_i}{1-\pi_i}\right] + \ln(1-\pi_i)\right\} \quad \text{because } \ln(a^b) = (b)\ln[a]$$

$$= \sum_{i=1}^{n}\left\{y_i \ln\left[\frac{\pi_i}{1-\pi_i}\right]\right\} + \sum_{i=1}^{n}\{\ln(1-\pi_i)\}$$

$$= \sum_{i=1}^{n}\{y_i\left[\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi}\right]\} + \sum_{i=1}^{n}\{\ln(1-\pi_i)\} \quad \text{by preliminary \#4}$$

$$= \sum_{i=1}^{n}\{y_i\left[\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi}\right]\} - \sum_{i=1}^{n}\{\ln(1+\exp[\beta_0+\beta_1 x_{1i}+\ldots+\beta_p x_{pi}])\} \quad \text{by preliminary \#5}$$

Maximization of the Log-Likelihood L (β) = ln { L }

Maximizing L (β) = ln { L } with respect to each of $\beta_0, \beta_1, \ldots, \beta_p$ is not the straightforward solution that was seen for estimating β_0 and β_1 in simple linear regression. It is beyond the scope of this course to develop the solution required here.

In brief, the solution for the maximum likelihood estimates is obtained by a method called Newton Raphson iteration. In brief, this iterative procedure for maximizing L (β) = ln { L } works with a linear approximation of the derivative of L (β) = ln { L } with respect to $\beta_0, \beta_1, \ldots, \beta_p$ and an initial estimate of $\beta_0, \beta_1, \ldots, \beta_p$. From there an updated estimate of $\beta_0, \beta_1, \ldots, \beta_p$ is obtained. Iteration continues until a convergence criterion is reached.