# STAT 571A — Advanced Statistical Regression Analysis

## Chapter 14 NOTES

## Introduction to Logistic Regression (et al.)

# §14.1: Binary Response Data

- A common data format in regression analysis is where the response variable $Y_i$ is binary, i.e., $Y_i$=0 or $Y_i$=1, but nothing else!
- Typical examples: healthy vs. diseased, on vs. off, yes vs. no, alive vs. dead, etc.
- We still have a predictor variable $X_i$ that we feel can predict $E[Y_i]$.
- How to proceed?!?

# Binary Response

- Under the SLR model, we took

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

  with $E[\varepsilon_i] = 0$. Thus $E[Y_i] = \beta_0 + \beta_1 X_i$.

- But: notice that when $Y_i=0$ or $Y_i=1$ (only), $E[Y_i] = (0)P[Y_i = 0] + (1)P[Y_i = 1] = P[Y_i = 1]$.

- Call this $\pi_i = P[Y_i = 1] = E[Y_i]$, and recognize that $\pi_i = E[Y_i]$ is a probability: $0 \leq \pi_i \leq 1$.

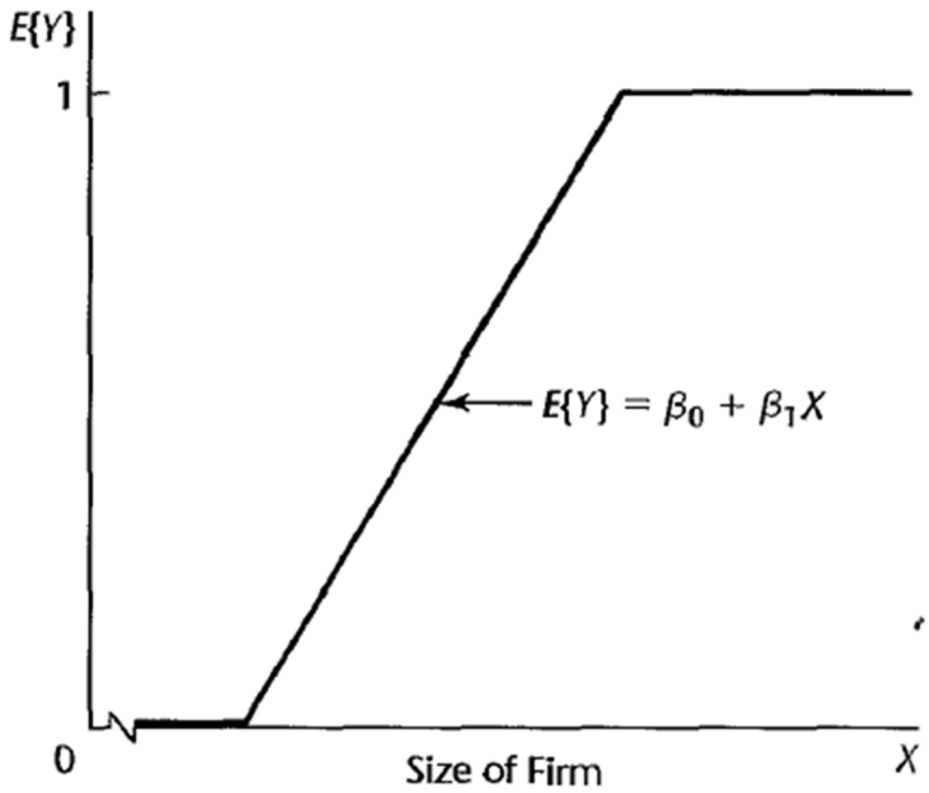
- Obviously the SLR model is inappropriate: the line $\beta_0 + \beta_1 X_i$ can't be constrained between 0 and 1!

# E[Y] for binary Y

- Fig. 14.1 illustrates the problem:

Over only a limited range will the straight line lie between 0 and 1; past this, we must truncate the mean response.

Probability That Firm Has Industrial Relations Department



# Binary Y Response

- For that matter, when $Y_i$=0 or $Y_i$=1 (only), $\varepsilon_i$ can<u>not</u> be $N(0,\sigma^2)$ in $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ as $Y_i$ is clearly a discrete random variable.

- And, it can be shown that when $Y_i$=0 or $Y_i$=1 (only) then $\sigma^2\{Y_i\} = \pi_i(1 - \pi_i)$, which is <u>non</u>-constant!

- Conclusion: binary data do not conform with our SLR model.

# §14.3: Simple Logistic Regression

- Instead of our previous model approach with $Y_i = E[Y_i] + \varepsilon_i$, for binary data we must move to a substantively *different formulation*.

- The Simple Logistic Regression Model sets $E[Y_i] = \pi_i = 1/(1 + exp\{-\beta_0 - \beta_1 X_i\})$ and it discards the additive error assumption. (In effect, $\varepsilon_i$ no longer exists.)

- Formally, we simply assume

  $Y_i \sim \text{Binomial}(1, \pi_i)$   (for $i = 1,\ldots,n$).

# Logistic Function

The term "logistic regression" comes from use of a **logistic distribution model** for the mean response:

- The c.d.f. of the standard logistic dist'n is $F(\eta) = e^{\eta}/(1 + e^{\eta}) = 1/(1 + \mathit{exp}\{-\eta\})$
- As this is a c.d.f., it can be used to model any quantity that ranges between 0 and 1, such as our $E[Y_i] = \pi_i$.
- So, we take $\pi_i = 1/(1 + \mathit{exp}\{-\eta_i\})$ with $\eta_i = \beta_0+\beta_1X_i$. Recall that $\eta_i$ is called the linear predictor.
- The inverse function is the *logit function*

$$F^{-1}(\pi_i) = \mathit{logit}\{\pi_i\} = \mathit{log}\{\pi_i /(1 - \pi_i)\} \qquad (14.18a)$$

# Interpretation of $\beta_1$

- Under the logistic regr. model, the interpretation of $\beta_1$ differs from what we've seen previously.

- Notice that $logit\{\pi(X)\} = \beta_0 + \beta_1 X$, while $logit\{\pi(X+1)\} = \beta_0 + \beta_1(X+1)$. Then clearly
$$logit\{\pi(X+1)\} - logit\{\pi(X)\} = \cdots = \beta_1.$$

- But, we saw $logit\{\pi\} = log\{\pi/(1 - \pi)\}$, which is the logarithm of the odds
$$\pi/(1 - \pi).$$

# Interpretation of $\beta_1$ (cont'd)

- Thus we say that $\beta_1$ is the change in log-odds when we increase X by +1 unit.
- By the way: if $Odds(X) = \pi(X)/[1 - \pi(X)]$, then

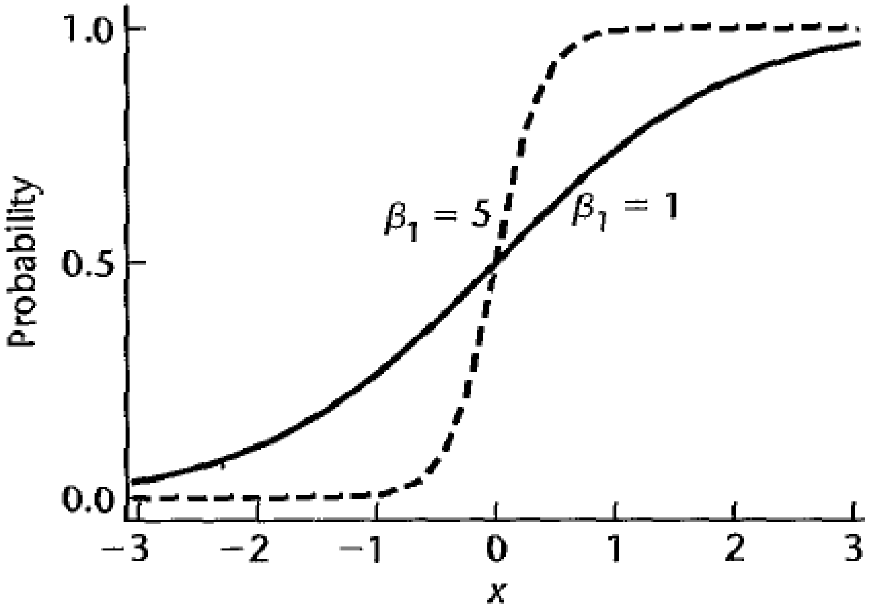
$$\begin{aligned} & log\{Odds(X+1)\} - log\{Odds(X)\} \\ = \quad & log\{Odds(X+1)/Odds(X)\} \end{aligned}$$

is called the log-odds ratio and it clearly equals $\beta_1$. The odds-ratio is then

$$OR = exp(\beta_1).$$

# Sigmoidal Response Function

- The logistic mean response is a sigmoidal ("S-shaped") function; see Fig. 14.2c:



- Other possibilities in the class of sigmoidal functions include the probit and complementary log-log ("CLL") functions. See §14.2.

# Maximum Likelihood

- We use weighted least squares (from §11.1) to fit the logistic regression model. This is equivalent to a **maximum likelihood** solution for the β parameters.
- Unfortunately, the equations do not produce a closed-form solution, so we must appeal to computer iteration.
- In R, we use the `glm()` function. (‘glm’ stands for generalized linear model, of which logistic regression is a special case; cf. §14.14.)

# Example: Program’g Task Data (CH14TA01)

- Y = Programming task result (0 = failure, 1 = success)
  X = Months of experience
- Logistic regression analysis in R:

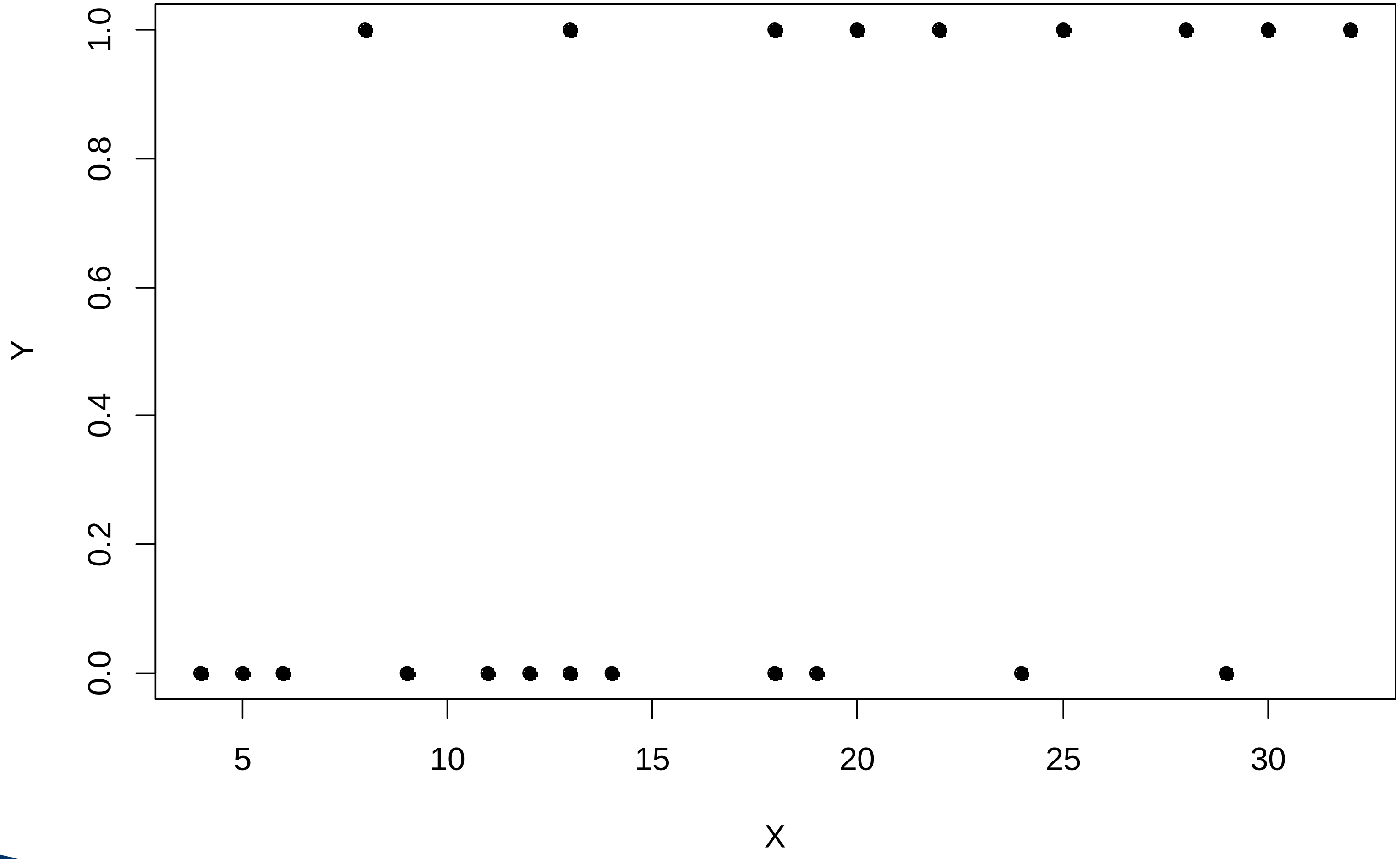
```
> plot( Y ~ X )     #not very informative
> CH14TA01.glm = glm( Y~X, family=binomial(logit) )
> summary( CH14TA01.glm )
```

```
Coefficients:
            Estimate  Std. Error  z value   Pr(>|z|)
(Intercept) -3.05970     1.25935   -2.430   0.0151
 X           0.16149     0.06498    2.485   0.0129
(Dispersion parameter for binomial family taken to be 1)
 Number of Fisher Scoring iterations: 4
```

Plot follows →

# Programming Task Example (cont'd)

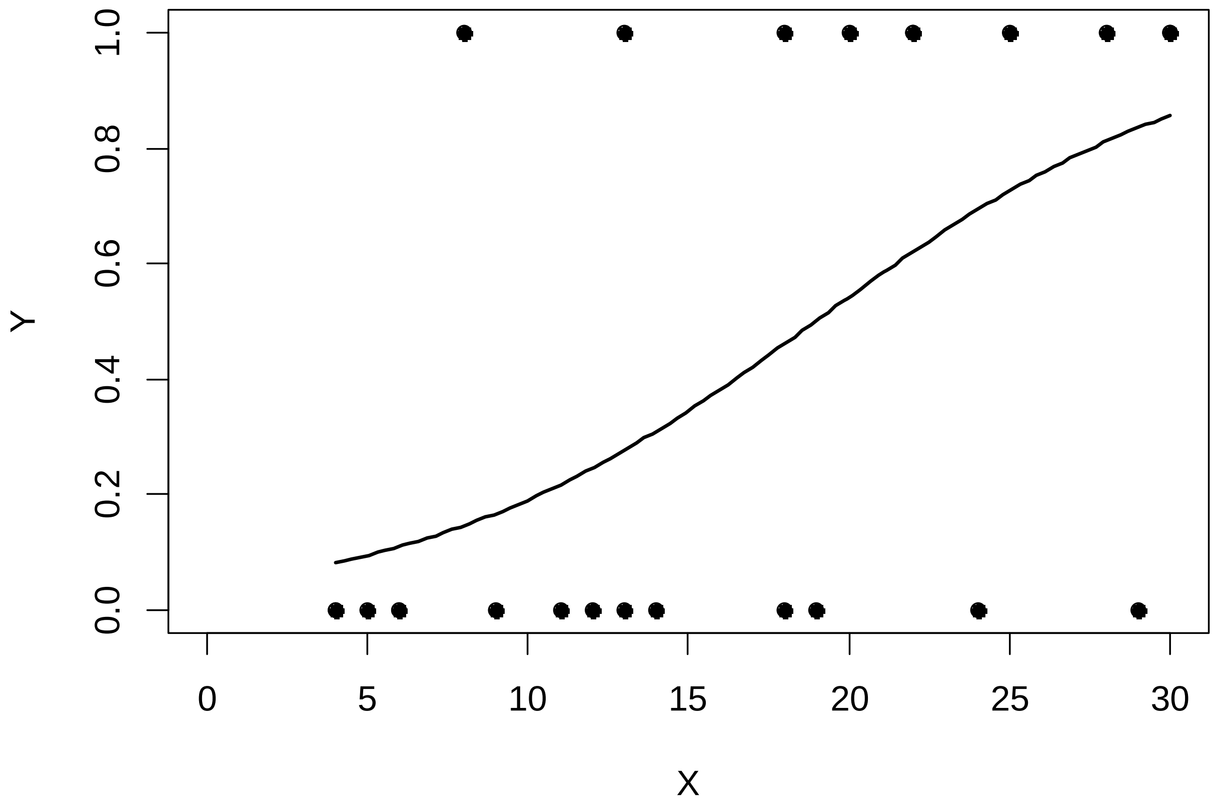
**Because the Y-response data are binary (0 or 1), the scatterplot isn't very illustrative:**



# Programming Task Example (cont'd)

**Overlay plot with fitted logistic regression curve (cf. Fig. 14.5):**

```
> plot( Y ~ X )
> b0=coef(CH14TA01.glm)[1]; b1=coef(CH14TA01.glm)[2]
> curve( 1/(1 + exp(-b0-b1*x)), xlim=c(4,30), add=T )
```



# Replication → Binomial Proportion Data

When multiple $Y_{ij}$s are observed at the same $X_j$, we have replication:

- The binary observations are $Y_{ij}$ for $i=1,\ldots,n_j$ and $j=1,\ldots,c$.
- Sum over *i* to produce bounded counts:
  $$Y_{\bullet j} = \sum_i Y_{ij} \sim \text{Binomial}(n_j,\pi_j), \text{ at each } X_j.$$
- This results in proportions, $Y_{\bullet j}/n_j$, at each $X_j$. In effect, these are nonparametric estimates of $\pi_j$.
- Can continue to model $\pi_j$ as logistic:
  $$\pi_j = 1/(1 + exp\{-\beta_0 - \beta_1 X_j\})$$
- Can still use `glm()` to fit the logistic regression model to such proportion data.

# Example: Coupon Data (CH14TA02)

- Y = # households redeeming coupons out of n=200 households
- X = Price reduction per coupon ($)
- Logistic regr. analysis in R (note need for cbind(Y,n-Y) syntax in formula's response variable):

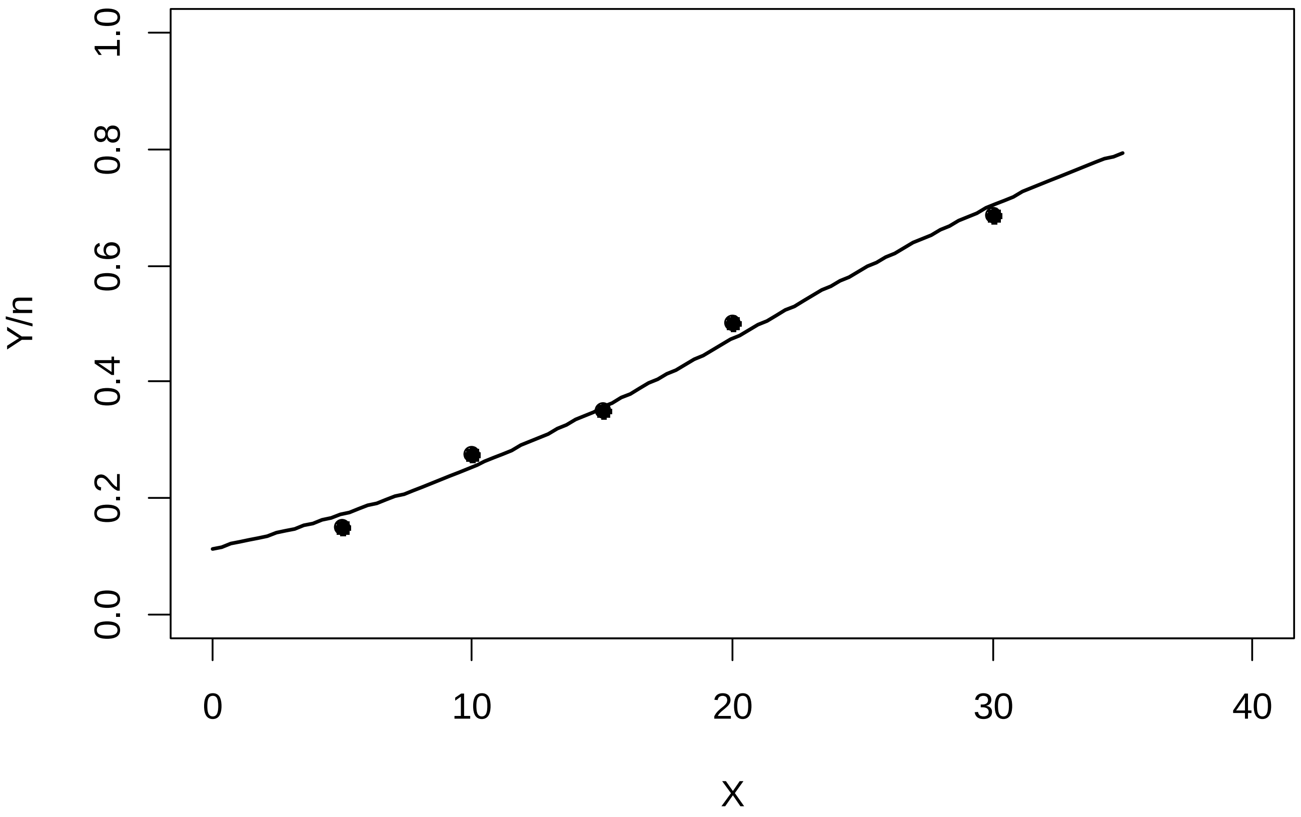
```
> CH14TA02.glm = glm( cbind(Y,n-Y) ~ X,
                      family = binomial(logit) )
> summary( CH14TA02.glm )
```

```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept) -2.044348   0.160977  -12.70   <2e-16
 X           0.096834   0.008549   11.33   <2e-16
(Dispersion parameter for binomial family taken to be 1)
 Number of Fisher Scoring iterations: 3
```

# Example: Coupon Data (cont'd)

**Overlay plot with fitted logistic regression curve (cf. Fig. 14.7):**

```
> plot( Y/n ~ X, pch=19 )
> b0=coef(CH14TA02.glm)[1]; b1=coef(CH14TA02.glm)[2]
> curve( 1/(1 + exp(-b0-b1*x)), xlim=c(0,35), add=T )
```



# §14.4: Multiple Logistic Regression

- The extension to multiple X variables $X_1, X_2, \ldots, X_{p-1}$ is straightforward. Take $\pi_i = 1/(1 + exp\{-\beta_0 - \beta_1 X_{i1} - \ldots - \beta_{p-1} X_{i,p-1}\})$
- Continue to use weighted least squares/maximum likelihood to estimate the β parameters.
- This still requires the computer: in R, modify the `formula` input in the `glm()` function in an obvious fashion.

# Interpretation of β parameters

- For the multiple linear-logistic model, the interpretation of the β parameters extends naturally from the simple linear-logistic case.
- $\beta_k$ is the log-odds ratio associated with a unit (+1) increase in $X_k$ <u>when</u> <u>all</u> <u>other</u> <u>X's</u> <u>are</u> <u>held</u> <u>fixed</u>.
- Special cases include polynomial logistic regression with $X_{ik} = X_i^k$ (set $k$ no larger than about 2 or 3 in practice), and logistic ANCOVA models with mixed quant./qual. predictors.

## Example: Disease Outbreak Data (CH14TA03)

**Multiple logistic regression data:**

- **Y = Disease status (Y=1 if present, Y=0 otherwise)**
- **$X_1$ = Age (yrs.)**
- **$X_2$ = Socioeconomic status 'M' (1 = middle class, 0 otherwise; see p. 573)**
- **$X_3$ = Socioeconomic status 'L' (1 = lower class, 0 otherwise; see p. 573)**
- **$X_4$ = City location ("sector") indicator**

# Disease Outbreak Data: R Code

- Multiple logistic regression analysis in R:

```
> CH14TA03.glm = glm( Y ~ X1 + X2 + X3 + X4,
                    family = binomial(logit) )
```

- Output estimated regr. coefficients with std. errors, etc.:

```
> summary( CH14TA03.glm )
```

- Print Var.-Cov. matrix of $\mathrm{b}$ vector, $\mathrm{s}^2\{\mathrm{b}\}$ (load *MASS* package 1st):

```
> library( MASS )
> vcov( CH14TA03.glm )
```

Output follows →

# Disease Outbreak Data: R Output

**Begin with summary() results:**

```
Call:  glm(formula = Y ~ X1 + X2 + X3 + X4,
                        family = binomial(logit))
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) -2.31293    0.64259  -3.599  0.000319
 X1          0.02975    0.01350   2.203  0.027577
 X2          0.40879    0.59900   0.682  0.494954
 X3         -0.30525    0.60413  -0.505  0.613362
 X4          1.57475    0.50162   3.139  0.001693
(Dispersion parameter for binomial family taken
to be 1)
Number of Fisher Scoring iterations: 4
```

# Disease Outbreak Data: R Output (cont’d)

**Next print Var.-Cov. matrix $s^2\{b\}$ from vcov( ):**

```
      (Intercept)      X1      X2      X3      X4
(Intercept) 0.4129 -0.0057 -0.1836 -0.2010 -0.1632
 X1        -0.0057  0.0002  0.0011  0.0007  0.0003
 X2        -0.1836  0.0011  0.3588  0.1482  0.0129
 X3        -0.2010  0.0007  0.1482  0.3650  0.0623
 X4        -0.1632  0.0003  0.0129  0.0623  0.2516
```

(cf. Table 14.4)

# §14.5: Inference in Logistic Regression

- To test if a particular $X_k$-variable is important in a logistic regression, we use a variant of the partial t-test, called a Wald Test.
- Test $H_o:\beta_k = 0$ vs. $H_a:\beta_k \neq 0$ (two-sided is default) using the Wald statistic $z^* = b_k/s\{b_k\}$, where $b_k$ is the MLE of $\beta_k$ and $s\{b_k\}$ is its std. error.
- Refer to $z^* \sim N(0,1)$ (not the t-dist'n) for the rejection region or p-value; e.g., $P = 2P[N(0,1) > |z^*|]$.
- As usual, this is a pointwise inference. Must apply a Bonferroni adjustment for multiple inferences on $g > 1$ different $\beta_k$s.

# Notes on Logistic Wald Test

1. The Wald test here is only an approximation that improves as $n \to \infty$. For small samples, it may not control the false positive error rate.

2. In R, Wald test results are provided in output from the `summary()` function.

3. IMPORTANT: Do NOT use the Wald test when $p = 2$, i.e., when there is only one X-variable. It is *known to be unstable* (Hauck & Donner, 1977, *JASA* vol. 72, pp. 851-853). Instead, use the likelihood ratio (LR) test, described next →

# LR test in Logistic Regression

- For testing multiple $\beta_k$s in a single $H_o$, say
  $$H_o: \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0,$$
  use the likelihood ratio (LR) statistic:
  $$G^2 = -2 \log\{L(RM)/L\ (FM)\}$$
  where $L$(FM) is the "likelihood" under the full model and $L$(RM) is the "likelihood" under the reduced model when $H_o$ is true. Note that q = p−1 is possible (1 d.f. alternative to Wald test).
- Reject $H_o$ when $G^2 > \chi^2(1-\alpha;p-q)$. Two-sided p-value is $P[\chi^2(p-q) > G^2]$.
- The details are nuanced & extend beyond our scope. See advanced texts on logistic regression.

# Disease Outbreak Data (CH14TA03, cont’d)

- Recall that we had p−1=4 predictor variables, so consider the “full” LR test of

$$H_o: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

- In R, find the `CH14TA03.glm` object for the FM, also fit the RM, and then apply the `anova()` function with the `test=‘Chisq’` option:

```
> CH14TA03rm.glm = glm( Y ~ 1,
                 family = binomial(logit) )
> anova( CH14TA03rm.glm, CH14TA03.glm,
                           test=‘ChiSq’)
```

Output follows →

# Disease Outbreak Data (cont’d)

**“Full” LR test of $H_o$: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$:**

**R output from the anova( ) function (notice the title “Analysis of Deviance Table” to distinguish from the ANOVA table in normal-data MLRs):**

```
Analysis of Deviance Table
Model 1: Y ~ 1
Model 2: Y ~ X1 + X2 + X3 + X4
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        97     122.32
2        93     101.05  4   21.264 0.0002808
```

**P-value is 0.0003 so “full” model is clearly significant.**

# Disease Outbreak Data (cont'd)

Now consider LR test of $H_o$: $\beta_1 = 0$:

```
> CH14TA03rm1.glm = glm( Y ~ X2+X3+X4,
                    family = binomial(logit) )
> anova( CH14TA03rm1.glm, CH14TA03.glm,
                         test='ChiSq')
 Model 1: Y ~ X2 + X3 + X4
 Model 2: Y ~ X1 + X2 + X3 + X4
   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
 1        94     106.20
 2        93     101.05  1   5.1495  0.02325
```

P-value is 0.0233 so retain $X_1$ in model; see p.581.

# Pointwise Confidence Intervals

For a 1−α pointwise conf. interval on a single $\beta_k$, there are 2 options:

- The **Wald interval** is the familiar form

  $$b_k \pm z(1-\{\alpha/2\})s\{b_k\}$$

  where z(1−{α/2}) is the upper-{α/2} critical point from Z ~ N(0,1).

→ *Avoid this if p=2*, due to Wald test's instability.

- Preferred: "Invert" a level-α LR test of $H_o$: $\beta_k = 0$ into a 1−α **LR conf. interval**, a.k.a. "**profile likelihood interval**." This has no closed form, but it can be computed in R.

# Programming Task Data (CH14TA01, cont'd)

- Recall:
  Y = Programming task result (0 = failure, 1 = success)
  X = Months of experience
- 95% logistic profile likelihood conf. interval for $\beta_1$:

```
> library( MASS )        #load MASS package
> confint( CH14TA01.glm, parm=2 )

   Waiting for profiling to be done...
        2.5 %      97.5 %
   0.05002505 0.31403972
```

⇒ can report $0.050 < \beta_1 < 0.314$.

- If desired (not recommended), compute Wald interval by hand from output of `summary(CH14TA01.glm)`; see p. 579.

# §14.8: Logistic Regression Diagnostics

- For a Residual Analysis, the usual, “raw” residual isn’t that useful with binary data. Instead, in logistic regression we find the **Pearson Residual**

$$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad (14.79)$$

where $\hat{\pi}_i$ is the *i*th predicted response.

- A *studentized* form, $r_{SPi}$, also exists; see Equation (14.81).

# Deviance Residuals

- With logistic regression models, a slightly more stable form is the **Deviance Residual**
  $dev_i$ =
  $$sign(Y_i - \hat{\pi}_i)\sqrt{-2[Y_i \log(\hat{\pi}_i) + (1-Y_i)\log(1-\hat{\pi}_i)]}$$
  as in Equation (14.83).

- **Residual plots**: One can plot $r_{Pi}$ or $dev_i$ against $\hat{\pi}_i$, but this will always produce a two-curve pattern ⇒ not that useful with binary data. See Fig. 14.12.

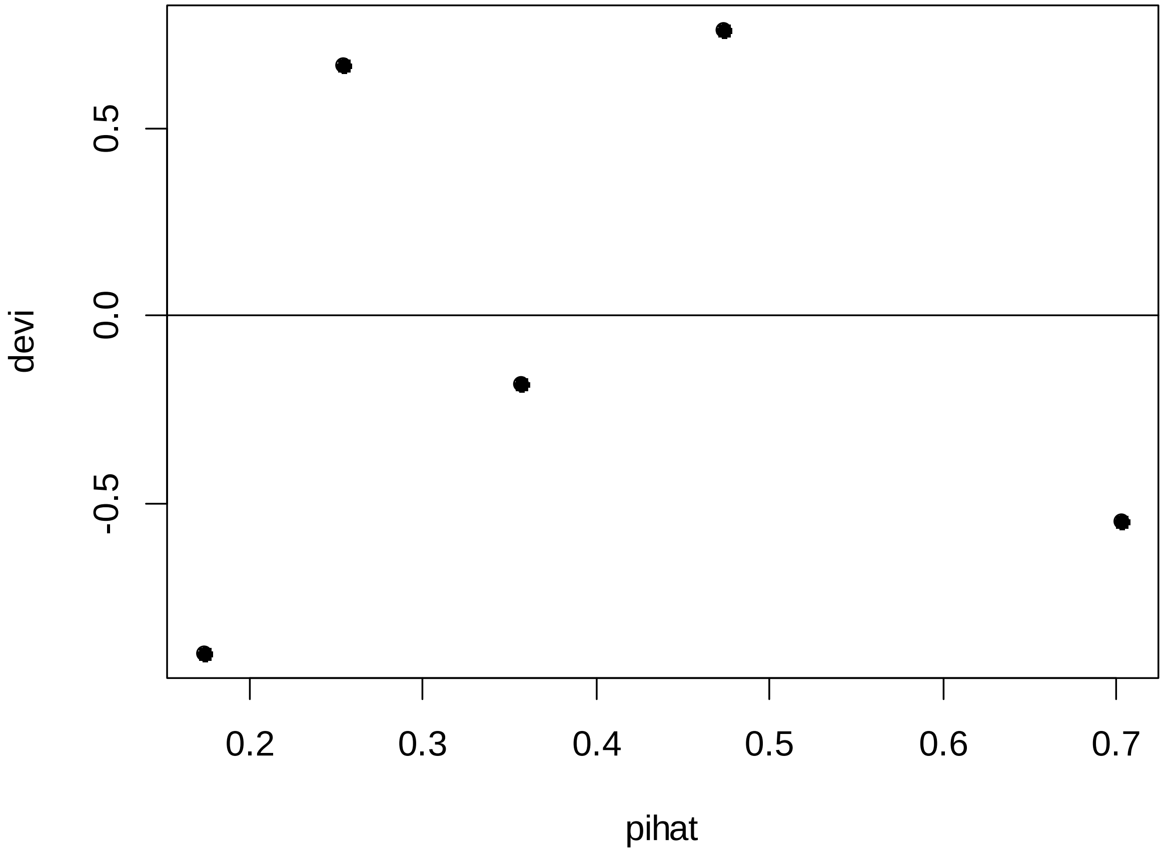
# Residual Plots with Proportion Data

- If replication in the binary response at multiple values of X produces proportion data, residual plots are more informative.
- Example: for the Coupon Data (CH14TA02), find the deviance residuals and plot in R:

```
> devi = residuals( CH14TA02.glm,
                    type=‘deviance’ )
> pihat = predict( CH14TA02.glm,
                    type=‘response’ )
> plot( devi ~ pihat ); abline ( h=0 )
```

# Deviance Residual Plot

**Here, the deviance residual plot is not very illustrative (due to limited number of distinct X values) but as given it shows no substantial problems:**



# Other Logistic Regr. Diagnostics

- More generally, the Deviance of a logistic regr. fit measures the adequacy of the model fit, using the likelihood function. The expression is complicated; see Eqn. (14.75).
  NOTATION: *DEV*($\mathrm{X}$) where $\mathrm{X}$ is the design matrix of the posited model.
- A rule-of-thumb diagnostic indicates serious model inadequacy if

$$\frac{DEV(\mathrm{X})}{n-p} > 1 + \frac{2.8}{\sqrt{n-p}}$$

- Other diagnostics for logistic regression include a form of Cook's distance; see pp. 599-601.

# Example: Disease Outbreak Data (CH14TA03, cont’d)

- Recall that we had p−1 = 4 predictor variables.
- In R, using the `CH14TA03.glm` object, calculate the terms for the adequacy measure rule-of-thumb:

```
> residDF = CH14TA03.glm$df.residual
> CH14TA03.glm$deviance/residDF     #adequacy measure
   [1] 1.086604
> 1 + ( 2.8/sqrt(residDF) )         #threshold
   [1] 1.290346
```

- We see *DEV*(X)/(n−p) = 1.0866 does not exceed the rule-of-thumb threshold of 1.2903, so we conclude that the model fits the data here in an adequate fashion.

# §14.14: Generalized Linear Models

- The logistic regr. model is a special case of a much larger family of regression models, called Generalized Linear Models (GLiMs).

- GLiMs also include:
  - MLR Normal (Gaussian) models from Chs. 1-11.
  - Poisson log-linear regression: $Y_i \sim \text{Poisson}(\lambda_i)$ with $\log\{\lambda_i\} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}$.
  - Gamma regression: $Y_i \sim \text{Gamma}(a_i, b_i)$ with $\log\{a_i b_i\} = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1}$.

- Continue to use `glm()` but now modify the `family=` option; see `help(glm)`.