# Chapter 5: Logistic Regression-I

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

[Acknowledgements to Tim Hanson and Haitao Chu]

# 5.1.1 Model Interpretation

The logistic regression model is

$$Y_i \sim \text{bin}(n_i, \pi_i), \quad \pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1})}.$$

- $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i,p-1})$ is a $p$-dimensional vector of explanatory variables including a place holder for the intercept.
- $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ is the $p$-dimensional vector of regression coefficients. These are the unknown population parameters.
- $\eta_i = \mathbf{x}_i' \boldsymbol{\beta}$ is called the linear predictor.
- Many, many uses including credit scoring, genetics, disease monitoring, etc, etc...
- Many generalizations: ordinal data, complex random effects models, discrete choice models, etc.

Lets start with simple logistic regression:

$$Y_i \sim \text{bin}\left(n_i, \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}\right).$$

*An odds ratio*: let's look at how the odds of success changes when we increase $x$ by one unit:

$$\begin{aligned}\frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} &= \frac{\left[\frac{e^{\alpha+\beta x+\beta}}{1+e^{\alpha+\beta x+\beta}}\right] / \left[\frac{1}{1+e^{\alpha+\beta x+\beta}}\right]}{\left[\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}\right] / \left[\frac{1}{1+e^{\alpha+\beta x}}\right]} \\ &= \frac{e^{\alpha+\beta x+\beta}}{e^{\alpha+\beta x}} = e^{\beta}.\end{aligned}$$

When we increase $x$ by one unit, the odds of an event occurring increases by a factor of $e^{\beta}$, *regardless of the value of* $x$.

So $e^{\beta}$ is an odds ratio. We also have

$$\frac{\partial \pi(x)}{\partial x} = \beta \pi(x)[1 - \pi(x)].$$

Note that $\pi(x)$ changes more when $\pi(x)$ is away from zero or one than when $\pi(x)$ is near 0.5.
This gives us *approximately* how $\pi(x)$ changes when $x$ increases by a unit.
This increase depends on $x$, unlike the odds ratio.

# 5.1.3 Horseshoe Crab Data

Let's look at $Y_i = 1$ if a female crab has one or more satellites, and $Y_i = 0$ if not. So

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}},$$

is the probability of a female having more than her nest-mate around as a function of her width $x$.

| Parameter | DF | Estimate | Standard Error | Wald Chi−Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | −12.3508 | 2.6287 | 22.0749 | <.0001 |
| width | 1 | 0.4972 | 0.1017 | 23.8872 | <.0001 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| width | 1.644 | 1.347 | 2.007 |

- We estimate the probability of a satellite as

$$\hat{\pi}(x) = \frac{e^{-12.35+0.50x}}{1 + e^{-12.35+0.50x}}.$$

- The odds of having a satellite increases by a factor between 1.3 and 2.0 times for every *cm* increase in carapace width.
- The coefficient table houses estimates $\hat{\beta}_j$, se$(\hat{\beta}_j)$, and the Wald statistic $z_j^2 = \{\hat{\beta}_j/\text{se}(\hat{\beta}_j)\}^2$ and $p$-value for testing $H_0 : \beta_j = 0$.
- What do we conclude here?

# 5.1.2 Looking at data

- With a single predictor $x$, can plot $p_i = y_i/n_i$ versus $x_i$. This approach works well when $n_i \neq 1$. The plot should look like a "lazy s."
- Alternatively, the sample logits $\log p_i/(1-p_i) = \log y_i/(n_i - y_i)$ versus $x_i$ should be approximately straight.
- If some categories have all successes or failures, an ad hoc adjustment is $\log\{(y_i + 0.5)/(n_i - y_i + 0.5)\}$.
- When many $n_i$ are small, you can group the data yourself into, say, 10-20 like categories and plot them.

For the horseshoe crab data let's use the categories defined in Chapter 4. A new variable $w$ is created that is the midpoint of the width categories:

```
data crab1; input color spine width satell weight;
  weight=weight/1000; color=color-1;
  y=0; n=1; if satell >0 then y=1; w=22.75;
   if width>23.25 then w=23.75;
   if width>24.25 then w=24.75;
   if width>25.25 then w=25.75;
   if width>26.25 then w=26.75;
   if width>27.25 then w=27.75;
   if width>28.25 then w=28.75;
   if width>29.25 then w=29.75;
run;
proc sort data=crab1; by w;
proc means data=crab1 noprint; by w; var y n;
    output out=crabs2 sum=sumy sumn;
data crabs3; set crabs2; p=sumy/sumn;
     logit =log((sumy+0.5)/(sumn-sumy+0.5));
proc gplot; plot p*w; plot logit *w; run;
```

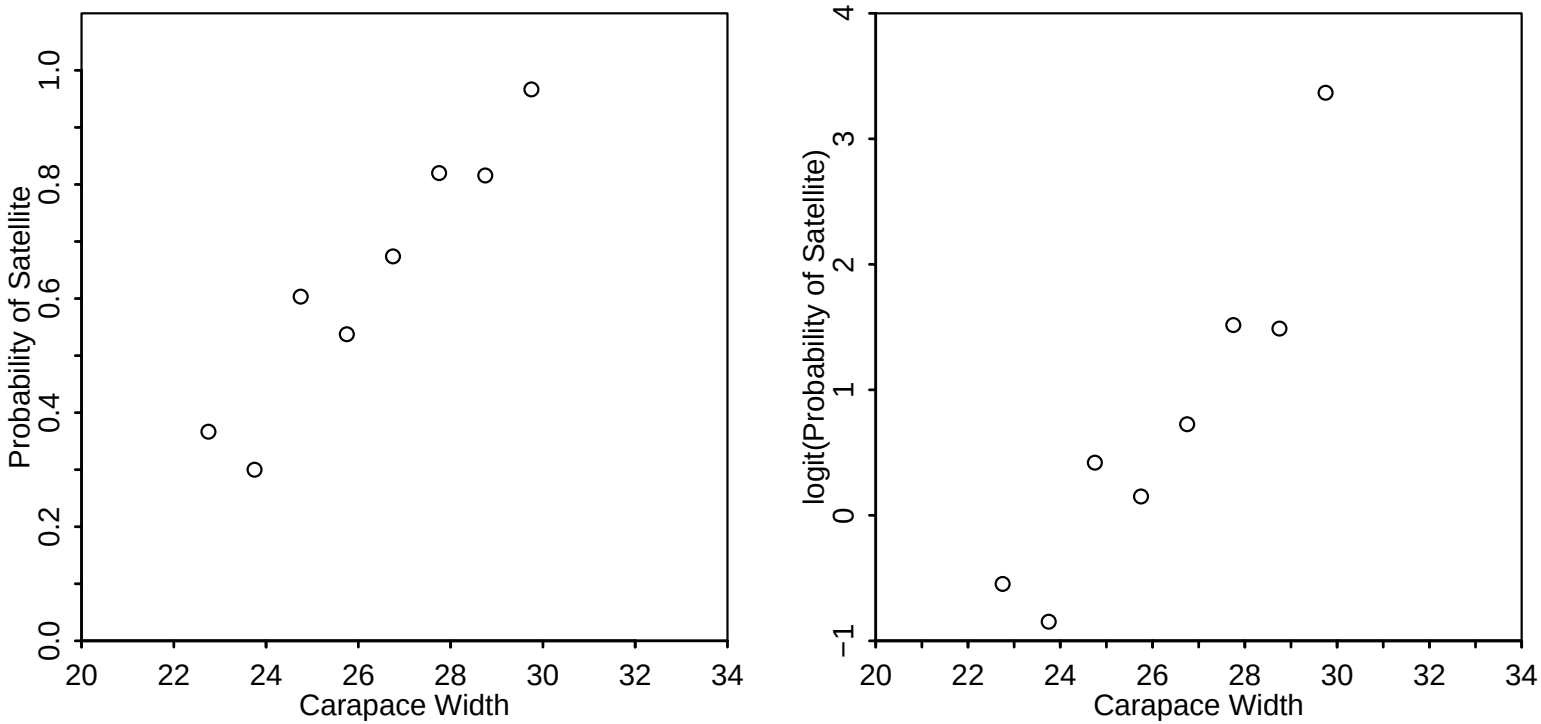


Figure : Sample P & logit(P) versus width; Is it "lazy s" or "straight?"

# 5.1.4 Retrospective sampling & logistic regression

- In case-control studies the number of cases and the number of controls are set ahead of time. It is not possible to estimate the probability of being a case *from the general population* for these types of data, but just as with a $2 \times 2$ table, we *can still estimate an odds ratio* $e^{\beta}$.
- Let $Z$ indicate whether a subject is sampled (1=yes,0=no). Let $P_1 = P(Z = 1|y = 1)$ be the probability that a case is sampled and let $P_0 = P(Z = 1|y = 0)$ be the probability that a control is sampled.
- In a simple random sample, $P_1 = P(Y = 1)$ and $P_0 = P(Y = 0) = 1 - P_1$.
- Assume the logistic regression model

$$\pi(x) = P(Y = 1|x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

- Assume that the probability of choosing a case is independent of $x$, $P(Z = 1|y = 1, x) = P(Z = 1|y = 1)$ and the same for a control $P(Z = 1|y = 0, x) = P(Z = 1|y = 0)$. This is the case, for instance, when a fixed number of cases and controls are sampled retrospectively, regardless of their $x$ values.
- Bayes' rule gives us

$$\begin{aligned} P(Y = 1|z = 1, x) &= \frac{P_1\pi(x)}{P_1\pi(x) + P_0(1 - \pi(x))} \\ &= \frac{e^{\alpha^* + \beta x}}{1 + e^{\alpha^* + \beta x}}, \end{aligned}$$

  where $\alpha^* = \alpha + \log(P_1/P_0)$.
- The parameter $\beta$ *has the same interpretation* in terms of odds ratios as with simple random sampling.

Comments:

- This is very powerful & another reason why logistic regression is widely used.
- Other links (e.g. identity, probit) do not have this property.
- *Matched* case/controls studies require more thought; Chapter 11.2.5.

# 5.2.1 Inference about Model Parameters and Probabilities

Consider the full model

$$\text{logit}\{\pi(\mathbf{x})\} = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} = \mathbf{x}'\boldsymbol{\beta}.$$

Most types of inferences are functions of $\boldsymbol{\beta}$, say $g(\boldsymbol{\beta})$. Some examples:

- $g(\boldsymbol{\beta}) = \beta_j$, $j^{th}$ regression coefficient.
- $g(\boldsymbol{\beta}) = e^{\beta_j}$, $j^{th}$ odds ratio.
- $g(\boldsymbol{\beta}) = e^{\mathbf{x}'\boldsymbol{\beta}}/(1 + e^{\mathbf{x}'\boldsymbol{\beta}})$, probability $\pi(\mathbf{x})$.

If $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$, then $g(\hat{\boldsymbol{\beta}})$ is the MLE of $g(\boldsymbol{\beta})$. This provides an estimate. The *delta method* is an all-purpose method for obtaining a standard error for $g(\hat{\boldsymbol{\beta}})$.

We know

$$\hat{\beta} \overset{\bullet}{\sim} N_p(\beta, \widehat{\text{cov}}(\hat{\beta})).$$

Let $g(\beta)$ be a function from $\mathbb{R}^p$ to $\mathbb{R}$. Taylor's theorem implies, as long as the MLE $\hat{\beta}$ is somewhat close to the true value $\beta$, that

$$g(\beta) \approx g(\hat{\beta}) + [Dg(\hat{\beta})](\beta - \hat{\beta}),$$

where $[Dg(\beta)]$ is the vector of first partial derivatives

$$Dg(\beta) = \begin{bmatrix} \frac{\partial g(\beta)}{\partial \beta_1} \\ \frac{\partial g(\beta)}{\partial \beta_2} \\ \vdots \\ \frac{\partial g(\beta)}{\partial \beta_p} \end{bmatrix}.$$

Then

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\bullet}{\sim} N_p(\mathbf{0}, \widehat{\text{cov}}(\hat{\boldsymbol{\beta}})),$$

implies

$$[Dg(\boldsymbol{\beta})]'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\bullet}{\sim} N(0, [Dg(\boldsymbol{\beta})]'\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})[Dg(\boldsymbol{\beta})]),$$

and finally

$$g(\hat{\boldsymbol{\beta}}) \overset{\bullet}{\sim} N(g(\boldsymbol{\beta}), [Dg(\hat{\boldsymbol{\beta}})]'\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})[Dg(\hat{\boldsymbol{\beta}})]).$$

So

$$\text{se}\{g(\hat{\boldsymbol{\beta}})\} = \sqrt{[Dg(\hat{\boldsymbol{\beta}})]'\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})[Dg(\hat{\boldsymbol{\beta}})]}.$$

This can be used to get confidence intervals for probabilities, etc.

```
proc logistic data=crabs1 descending;
    model y = width; output out=crabs2 pred=p lower=l upper=u;
proc sort data=crabs2; by width;
proc gplot data=crabs2;
     title "Estimated probabilities with pointwise 95% CI's";
    symbol1 i=join color=black; symbol2 i=join color=red line=3;
    symbol3 i=join color=black; axis1 label =('');
     plot (l p u)*width / overlay vaxis=axis1;
run;
```

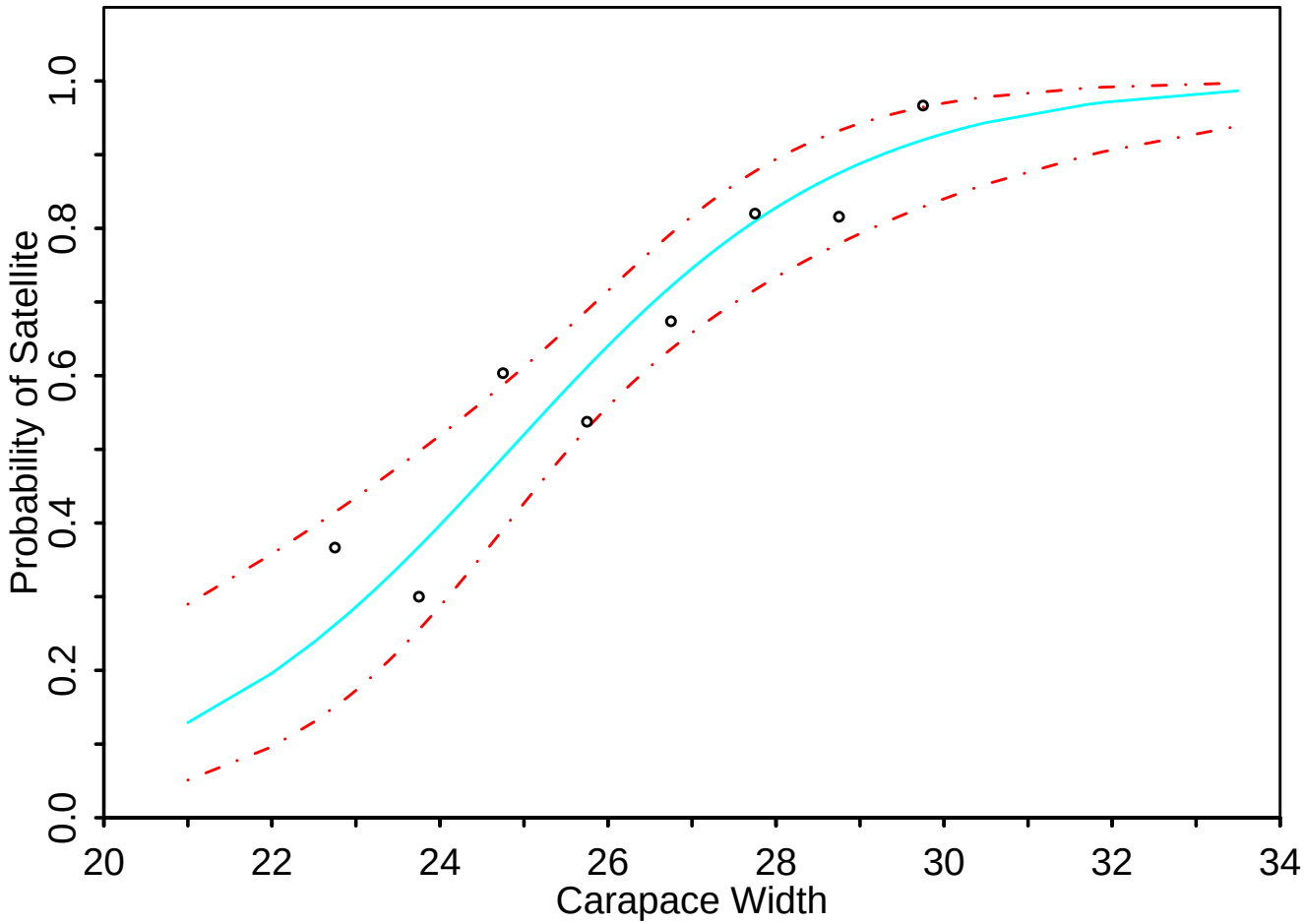


Figure : Fitted probability of satellite as a function of width & 95% CIs.

# 5.2.3 & 5.2.4 Goodness of fit

The deviance GOF statistic is defined to be

$$D = 2\sum_{i=1}^{s}\left\{y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right\},$$

where $\hat{\pi}_i = \frac{e^{\mathbf{x}_i'\hat{\beta}}}{1+e^{\mathbf{x}_i'\hat{\beta}}}$ are fitted values.
Pearson's GOF statistic is

$$X^2 = \sum_{i=1}^{s}\frac{(y_i - n_i\hat{\pi}_i)^2}{n_i\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Both statistics are approximately $\chi^2_{s-p}$ in large samples assuming that the number of *trials* $n = \sum_{i=1}^{s} n_i$ increases in such a way that each $n_i$ increases.

# 5.2.5 Group your data

Binomial data is often recorded as individual (Bernoulli) records:

| $i$ | $y_i$ | $n_i$ | $x_i$ |
|---|---|---|---|
| 1 | 0 | 1 | 9 |
| 2 | 0 | 1 | 14 |
| 3 | 1 | 1 | 14 |
| 4 | 0 | 1 | 17 |
| 5 | 1 | 1 | 17 |
| 6 | 1 | 1 | 17 |
| 7 | 1 | 1 | 20 |

Grouping the data yields an identical model:

| $i$ | $y_i$ | $n_i$ | $x_i$ |
|---|---|---|---|
| 1 | 0 | 1 | 9 |
| 2 | 1 | 2 | 14 |
| 3 | 2 | 3 | 17 |
| 4 | 1 | 1 | 20 |

- $\hat{\boldsymbol{\beta}}$, se$(\hat{\beta}_j)$, and $L(\hat{\boldsymbol{\beta}})$ don't care if data are grouped.
- The quality of residuals and GOF statistics *depend on how data are grouped*. $D$ and Pearson's $X^2$ will change!
- In PROC LOGISTIC type AGGREGATE and SCALE=NONE after the MODEL statement to get $D$ and $X^2$ based on grouped data. This option *does not* compute residuals based on the grouped data. You can aggregate over all variables or a subset, e.g. AGGREGATE=(width).

- The Hosmer and Lemeshow test statistic orders observations $(\mathbf{x}_i, Y_i)$ by fitted probabilities $\hat{\pi}(\mathbf{x}_i)$ from smallest to largest and divides them into (typically) $g = 10$ groups of roughly the same size. A Pearson test statistic is computed from these $g$ groups.
- The statistic would have a $\chi^2_{g-p}$ distribution if each group had *exactly the same predictor* $\mathbf{x}$ for all observations (but the observations in a group do not have the same predictor $\mathbf{x}$ and they do not share a common success probability). In general, the null distribution is *approximately* $\chi^2_{g-2}$ when the number of distinct patterns of covariate values equals the sample size (see text). Termed a "near-replicate GOF test" (Hosmer and Lemeshow 1980). The LACKFIT option in PROC LOGISTIC gives this statistic.
- Can also test $\text{logit}\{\pi(x)\} = \beta_0 + \beta_1 x$ versus more general model $\text{logit}\{\pi(x)\} = \beta_0 + \beta_1 x + \beta_2 x^2$ via $H_0 : \beta_2 = 0$.

# Raw (Bernoulli) data with `aggregate scale=none lackfit;`

Deviance and Pearson Goodness–of–Fit Statistics

| Criterion | Value | DF | Value/DF | Pr > ChiSq |
|---|---|---|---|---|
| Deviance | 69.7260 | 64 | 1.0895 | 0.2911 |
| Pearson | 55.1779 | 64 | 0.8622 | 0.7761 |

Number of unique profiles: 66

Partition for the Hosmer and Lemeshow Test

| | | y = 1 | | y = 0 | |
|---|---|---|---|---|---|
| Group | Total | Observed | Expected | Observed | Expected |
| 1 | 19 | 5 | 5.39 | 14 | 13.61 |
| 2 | 18 | 8 | 7.62 | 10 | 10.38 |
| 3 | 17 | 11 | 8.62 | 6 | 8.38 |
| 4 | 17 | 8 | 9.92 | 9 | 7.08 |
| 5 | 16 | 11 | 10.10 | 5 | 5.90 |
| 6 | 18 | 11 | 12.30 | 7 | 5.70 |
| 7 | 16 | 12 | 12.06 | 4 | 3.94 |
| 8 | 16 | 12 | 12.90 | 4 | 3.10 |
| 9 | 16 | 13 | 13.69 | 3 | 2.31 |
| 10 | 20 | 20 | 18.41 | 0 | 1.59 |

Hosmer and Lemeshow Goodness–of–Fit Test

| Chi–Square | DF | Pr > ChiSq |
|---|---|---|
| 5.2465 | 8 | 0.7309 |

Comments:

- There are 66 distinct widths $\{\mathbf{x}_i\}$ out of $N = 173$ crabs. For $\chi^2_{66-2}$ to hold, we must keep sampling crabs that only have one of the 66 *fixed number of widths*! Does that make sense here?
- The Hosmer and Lemeshow test gives a $p$-value of 0.73 based on $g = 10$ groups. Are assumptions going into this $p$-value met?
- None of the GOF tests have assumptions that are met in practice for continuous predictors. Are they still useful?
- The raw statistics do not tell you *where* lack of fit occurs. Deviance and Pearson residuals do tell you this (later). Also, the table provided by the H-L tells you which groups are ill-fit should you reject $H_0$ : logistic model holds.
- GOF tests are meant to detect *gross* deviations from model assumptions. No model ever truly fits data except hypothetically.

## 5.3.1 Categorical predictors

Let's say we wish to include variable $X$, a categorical variable that takes on values $x \in \{1, 2, \ldots, I\}$. We need to allow each level of $X = x$ to affect $\pi(x)$ differently. This is accomplished by the use of dummy variables. This is typically done one of two ways.

Define $z_1, z_2, \ldots, z_{I-1}$ as follows:

$$z_j = \begin{cases} 1 & X = j \\ -1 & X \neq j \end{cases}$$

This is the default in PROC LOGISTIC with a CLASS X statement. Say $I = 3$, then the model is

$$\text{logit } \pi(x) = \beta_0 + \beta_1 z_1 + \beta_2 z_2.$$

which gives

$$\begin{aligned} \text{logit } \pi(x) &= \beta_0 + \beta_1 - \beta_2 \quad \text{when} \quad X = 1 \\ \text{logit } \pi(x) &= \beta_0 - \beta_1 + \beta_2 \quad \text{when} \quad X = 2 \\ \text{logit } \pi(x) &= \beta_0 - \beta_1 - \beta_2 \quad \text{when} \quad X = 3 \end{aligned}$$

At alternative method uses "zero/one" dummies instead:

$$z_j = \begin{cases} 1 & X = j \\ 0 & X \neq j \end{cases}$$

This is the default if PROC GENMOD with a CLASS X statement. This can also be obtained in PROC LOGISTIC with the PARAM=REF option. This sets class $X = I$ as baseline. Say $I = 3$, then the model is

$$\text{logit } \pi(x) = \beta_0 + \beta_1 z_1 + \beta_2 z_2.$$

which gives

$$\begin{aligned} \text{logit } \pi(x) &= \beta_0 + \beta_1 \quad \text{when} \quad X = 1 \\ \text{logit } \pi(x) &= \beta_0 + \beta_2 \quad \text{when} \quad X = 2 \\ \text{logit } \pi(x) &= \beta_0 \quad \text{when} \quad X = 3 \end{aligned}$$

I prefer the latter method because it's easier to think about for me. You can choose a different baseline category with REF=FIRST next to the variable name in the CLASS statement. Table 3.7 (p. 89):

| | Drinks per day | | | | |
|---|---|---|---|---|---|
| Malformation | 0 | $< 1$ | $1-2$ | $3-5$ | $\geq 6$ |
| Absent | 17,066 | 14,464 | 788 | 126 | 37 |
| Present | 48 | 38 | 5 | 1 | 1 |

```
data mal;
    input cons present absent score @@;
    total = present+absent;
    datalines ;
   1 48 17066 0 2 38 14464 0.5 3 5 788 1.5 4 1 126 4.0 5 1 37 7.0;
run;
proc logistic data=mal;
    class cons / param=ref ref=last;
    model present/total = cons;
run;
```

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 6.2020 | 4 | 0.1846 |
| Score | 12.0821 | 4 | 0.0168 |
| Wald | 9.2811 | 4 | 0.0544 |

Type 3 Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| cons | 4 | 9.2811 | 0.0544 |

Analysis of Maximum Likelihood Estimates

| Parameter | | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|
| Intercept | | 1 | −3.6109 | 1.0134 | 12.6956 | 0.0004 |
| cons | 1 | 1 | −2.2627 | 1.0237 | 4.8858 | 0.0271 |
| cons | 2 | 1 | −2.3309 | 1.0264 | 5.1577 | 0.0231 |
| cons | 3 | 1 | −1.4491 | 1.1083 | 1.7097 | 0.1910 |
| cons | 4 | 1 | −1.2251 | 1.4264 | 0.7377 | 0.3904 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| cons 1 vs 5 | 0.104 | 0.014 | 0.774 |
| cons 2 vs 5 | 0.097 | 0.013 | 0.727 |
| cons 3 vs 5 | 0.235 | 0.027 | 2.061 |
| cons 4 vs 5 | 0.294 | 0.018 | 4.810 |

The model is

$$\text{logit } \pi(X) = \beta_0 + \beta_1 I\{X = 1\} + \beta_2 I\{X = 2\} + \beta_3 I\{X = 3\} + \beta_4 I\{X = 4\}$$

where $X$ denotes alcohol consumption $X = 1, 2, 3, 4, 5$.

- Type 3 analyses test whether all dummy variables associated with a categorical predictor are simultaneously zero, here $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$. If we accept this then the categorical predictor is not needed in the model.
- PROC LOGISTIC gives estimates and CIs for $e^{\beta_j}$ for $j = 1, 2, 3, 4$. Here, these are interpreted as the odds of developing malformation when $X = 1, 2, 3$, or 4 versus the odds when $X = 5$.
- We are not as interested in the *individual* Wald tests $H_0 : \beta_j = 0$ for a categorical predictor. Why is that? Because they only compare a level $X = 1, 2, 3, 4$ to baseline $X = 5$, not to each other.

- The `Testing Global Null Hypothesis: BETA=0` are three tests that *no predictor* is needed; $H_0 : \text{logit}\{\pi(x)\} = \beta_0$ versus $H_1 : \text{logit}\{\pi(x)\} = \mathbf{x}'\boldsymbol{\beta}$.
- Anything wrong here? 1) p-values = 0.18, 0.02, 0.05 from LR, Score and Wald tests respectively; 2) the p-vlues using the exact conditional distribution of $X^2$ and $G^2$ are 0.03 and 0.13, providing mixed signals. The table 3.7 has a mixture of very small, moderate, and extremely large counts, even though n=32,574, the null distributions of $X^2$ and $G^2$ may not be close to chi-squared. In any case, these statistics ignore the ordinality of alcohol consumption.

- Note that the Wald test for $H_0 : \beta = 0$ is the same as the Type III test that consumption is not important. Why is that?
- Let $Y = 1$ denote malformation for a randomly sampled individual. To get an odds ratio for malformation from increasing from, say, $X = 2$ to $X = 4$, note that

$$\frac{P(Y = 1|X = 2)/P(Y = 0|X = 2)}{P(Y = 1|X = 4)/P(Y = 0|X = 4)} = e^{\beta_2 - \beta_4}.$$

This is estimated with the CONTRAST command.