Logistic regression in SAS version 8

Paul W. Dickman
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

paul.dickman@mep.ki.se

October 28, 2003

Introduction to logistic regression

- Assume that for each individual in a study we have collected data on a binary outcome variable, Y, and k explanatory variables, $X_1, X_2, \ldots, X_k$.
- The explanatory variables may be continuous or, to model a categorical variable we create a series of indicator variables (dummy variables).
- We may conveniently (but quite arbitrarily) code the two outcome categories as $Y = 0$ (individuals without the characteristic of interest) and $Y = 1$ (individuals with the characteristic of interest).
- In introductory statistics courses (such as BIOSTAT I in the epi program) one is introduced to linear regression models of the form

$$\mathsf{E}(Y|X_1, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \quad (1)$$

- The outcome variable, Y, is assumed, conditional on $X_1, \ldots, X_k$, to have a normal distribution with mean zero and constant variance.

- An observation of the outcome variable could be expressed as $y = \mathsf{E}(Y|X_1, \ldots, X_k) + \epsilon$ meaning the model could also be written as

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \epsilon \quad (2)$$

  where ϵ represents the random error (experimental error) and is generally assumed to be normally distributed with mean zero and constant variance.
- The right-hand side of Equation 1 can take any value between $-\infty$ and $+\infty$.
- If Y takes on the values 0 or 1 then the left hand side of Equation 1 represents a probability so must lie between 0 and 1.
- It is more reasonable to model $\mathsf{Pr}(Y = 1|X_1, \ldots, X_k)$ as the outcome, which is equal to $\mathsf{E}(Y|X_1, \ldots, X_k)$ when the response (Y) is coded as 0 or 1.
- To simplify notation, let the outcome of our model be $\pi(\boldsymbol{X}) = \mathsf{Pr}(Y = 1|X_1, \ldots, X_k)$, where $\boldsymbol{X}$ represents $X_1, \ldots, X_k$.

- $\pi(\boldsymbol{X})$ represents the probability of the outcome of interest and lies in the interval $[0, 1]$.
- $\pi(\boldsymbol{X})/(1 - \pi(\boldsymbol{X}))$ belongs to the interval $(0, \infty)$.
- $\log\{\pi(\boldsymbol{X})/(1 - \pi(\boldsymbol{X}))\}$ belongs to the interval $(-\infty, \infty)$; this is the same range of values to which the expression $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$ belongs.
- Thus, the basis for logistic regression is the equation (statistical model)

$$\log\left\{\frac{\pi(\boldsymbol{X})}{1 - \pi(\boldsymbol{X})}\right\} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k. \quad (3)$$

- Note that $\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}$ is the odds of the outcome of interest for an individual with covariates $\boldsymbol{X}$.
- $\log\left\{\frac{\pi(\boldsymbol{X})}{1-\pi(\boldsymbol{X})}\right\}$ is called the log odds or the logit.

- That is, instead of fitting a model with the probability of disease as the outcome, we fit a model where the logarithm of the odds of disease is the outcome.
- An equivalent way of specifying the model is via the equation

$$\pi(\boldsymbol{X}) = \mathsf{Pr}(Y = 1|\boldsymbol{X}) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots \beta_k X_k)} \quad (4)$$

- As in simple and multiple linear regression, if a particular regression coefficient, say β_j, is zero, then the corresponding explanatory variable, X_j, is not associated with the occurrence of the response, in which case we may wish to omit X_j from any final model for the observed data.

Interpreting the estimated regression coefficients in logistic regression

- The simplest case is when the logistic regression model involves only one explanatory variable, say X_1, and that X_1 takes only two values, 0 (unexposed) and 1 (exposed);
- A logistic regression model for these data would correspond to

$$\log\left\{\frac{\pi(X_1)}{1 - \pi(X_1)}\right\} = \beta_0 + \beta_1 X_1.$$

- More specifically, the model is

$$\log\left\{\frac{\pi(X_1 = 1)}{1 - \pi(X_1 = 1)}\right\} = \beta_0 + \beta_1$$

  for the exposed individuals ($X_1 = 1$) and

$$\log\left\{\frac{\pi(X_1 = 0)}{1 - \pi(X_1 = 0)}\right\} = \beta_0$$

  for the unexposed individuals ($X_1 = 0$).
- We see that β_0 represents the logarithm of the odds of response for unexposed individuals, whereas the logarithm of the odds of response for exposed individuals is given by $\beta_0 + \beta_1$.

- If we subtract the latter model equation (where $X_1 = 0$) from the former (where $X_1 = 1$), we see that

$$\begin{aligned}
\beta_1 &= \log\left\{\frac{\pi(X_1 = 1)}{1 - \pi(X_1 = 1)}\right\} - \log\left\{\frac{\pi(X_1 = 0)}{1 - \pi(X_1 = 0)}\right\} \\
&= \log\left\{\frac{\pi(X_1 = 1)}{1 - \pi(X_1 = 1)} \Big/ \frac{\pi(X_1 = 0)}{1 - \pi(X_1 = 0)}\right\} \\
&= \log\left\{\frac{\text{odds of response when exposed}}{\text{odds of response when unexposed}}\right\}
\end{aligned}$$

- This equation reveals that β_1, the regression coefficient associated with X_1 represents the logarithm of the odds ratio.
- Stated another way, β_1 represents the change in the logarithm of the odds in favour of the response of interest when the corresponding explanatory variable, X_1, increases by one unit, i.e., from $X_1 = 0$ to $X_1 = 1$.

- Clearly, if $\beta_1 > 0$, the log-odds in favour of the response of interest increases as X_1 increases from 0 to 1; conversely, if $\beta_1 < 0$, the log-odds in favour of the response of interest decreases as X_1 increases from 0 to 1.
- It should also be evident that if $\beta_1 = 0$, then the log-odds in favour of the response of interest does not change as X_1 changes.
- We can show that the corresponding model for the probability of response,

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

  is an increasing function with respect to the regression coefficient, β_1, so that an increase in the log-odds in favour of response means that the probability of response increases.

Conditional vs. unconditional logistic regression

- Finely matched case-control studies (i.e. where the number of observations in each matched set (stratum) is small) should be analysed using conditional logistic regression.
- Frequency matched case-control studies may be analysed using unconditional logistic regression where the matching variables are included as explanatory variables.
- The likelihood function for conditional logistic regression is identical to the likelihood function for the Cox proportional hazards model meaning software designed for estimating the Cox model can be used to estimate the conditional logistic regression model.
- In SAS we use PROC PHREG to estimate the conditional logistic regression model.

An example of conditional logistic regression

- This code was used for Olof Stephansson's study of the association between maternal hemoglobin concentration during pregnancy and risk of stillbirth (JAMA (2000) 284:2611-7).
- The study was individually matched on delivery hospital and year of birth.

```
proc phreg data=olofs.main;
model time*case(0)= hb1 hp3 hb4 age1 age2 age3
        / ties=discrete risklimits;
strata hospital year;
hb: test hb1=hb3=hb4=0;
age: test age1=age2=age3=0;
run;
```

- The outcome of survival studies has two dimensions – the time at risk and whether or not the event of interest was observed. If the event of interest is not observed then the survival time is said to be censored.

- To estimate the Cox model using PHREG we specify a variable containing the survival time and a variable containing the vital status ('dead' or censored).
- To estimate the conditional logistic regression model we need to set up the data so that, within each stratum, the cases all 'die' at the same time and the controls are 'censored' at a later time.
- For example, we create a variable called `time` which takes the value 1 for cases and 2 for controls.
- We then tell SAS that all controls are censored – if the variable `case` takes the value 1 for cases and 0 for controls then we specify that anyone with case=0 is censored.
- The left-hand side of the model statement is `time*case(0)` where the values in parentheses indicate the values of the status variable that represent censoring.

- In order to use the correct likelihood function we need to specify the `ties=discrete` option on the model statement.
- To model categorical variables we must create indicator variables in a data step – I have created variables hb1–hb4 for the 4 categories of hemoglobin but only included 3 of these in the model (the one excluded is the reference).
- The TEST statement can be used to test the effect of a categorical variable.
- These are so-called Wald tests – likelihood ratio tests can be performed by fitting the full and reduced model and calculating the difference in the log likelihood.
- See example 14 from the book 'Logistic regression examples using the SAS system' for further details.
- In SAS version 9, PROC LOGISTIC can be used for conditional logistic regression using the new STRATA statement. Also new in version 9 is an experimental version of PROC PHREG that contains a CLASS statement.

Unconditional logistic regression in SAS

- Application of logistic regression in epidemiology primarily involves categorical explanatory variables.
- In SAS version 6, one was required to create dummy variables in a data step in order to model categorical variables using PROC LOGISTIC.
- PROC GENMOD, which contained a CLASS statement, was therefore preferable for logistic regression despite the disadvantage that it only provided estimates of log odds ratios (one was required to save the parameter estimates to a data set, exponentiate them in a DATA step, and print the resulting odds ratio estimates using PROC PRINT).
- PROC LOGISTIC in version 8 contains a CLASS statement, meaning that this is now the procedure of choice for logistic regression in SAS.
- An additional benefit of PROC LOGISTIC is that it contains options specific to logistic regression, such as goodness-of-fit tests and ROC curves.

Summary comparison of PROC GENMOD and PROC LOGISTIC for unconditional logistic regression

		LOGISTIC	
Characteristic	GENMOD	v6	v8
CLASS statement	yes	no	yes
Odds ratio estimates directly	no	yes	yes
Options specific to logistic regression	no	yes	yes
Advanced capabilities			
GEE (for correlated data)	yes	no	no
Easy parameterisation of interactions	yes	no	no?

- The REPEATED statement in PROC GENMOD facilitates the estimation of marginal models (using generalised estimating equations) to correlated data (e.g. twin data or repeated measures data). This capability is not available in PROC LOGISTIC.
- PROC GENMOD had a nice syntax for parameterising models containing interactions. For example, one could easily obtain estimates of the exposure odds ratio with confidence intervals for each level of an effect modifier. So far I haven't been able to do this easily using PROC LOGISTIC (see slide 44).

An example of PROC LOGISTIC in SAS version 8

- I'll use the CAHRES breast cancer data as an example and will reproduce some of the results published in Cecilia Magnusson's doctoral thesis.

  Magnusson C *et al.*, Breast-cancer risk following long-term oestrogen- and oestrogen-progestin-replacement therapy. *Int J Cancer* 1999;81:339-44.

- We are interested in the effect of ever exclusive use of unopposed estrogen (`eox`) and wish to adjust for parity (`parity`), height (`f2`), BMI (`bmi`), age at first birth (`agefb`), age at menopause (`mpage`), menopause type (surgical/natural) (`mpty`), and age (`f1`).

- All confounders are modelled as categorical variables except for parity.

- Categories are created using PROC FORMAT.


```
proc format library=emma;

value mpage
low-<45='<45'
45-<50='[45,50)'
50-<52='[50,52)'
52-<55='[52,55)'
55-high='55+'
;

value bmi
low-<22.16='BMI Q1'
22.16-<24.09='BMI Q2'
24.09-<25.85='BMI Q3'
25.85-<28.31='BMI Q4'
28.31-high='BMI Q5'
;
run;
```


Estimating the model without any options

```
proc logistic data=emma.analysis;
class mpage f1 bmi agefb f2;
model case=eox parity f2 bmi agefb mpage mpty f1;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```

- Any variable that appears in a CLASS statement is modelled as a categorical variable.

- Any variable that is in the MODEL statement but not the CLASS statement is modelled as a continuous variable. That is, the estimated odds ratio applies to a one unit increase in the variable.

- A variable can appear in the CLASS statement and not the MODEL statement, although SAS will exclude all observations with missing values for this variable despite it not being in the model. This behaviour is useful for comparing models estimated using the same observations.

First we should check the log

```
67  proc logistic data=emma.analysis;
68  class mpage f1 bmi agefb f2;
69  model case=eox parity f2 bmi agefb mpage mpty f1;
70  format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
71  run;

NOTE: PROC LOGISTIC is modeling the probability that CASE='CASE'.
      One way to change this to model the probability that
      CASE='CTRL' is to specify the response variable option
      EVENT='CTRL'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 5354 observations read from the data set
      EMMA.ANALYSIS.
```

- We confirm that we are modelling the correct outcome (the probability of being a case).

Now let's look at the output

```
Data Set                    EMMA.ANALYSIS
Response Variable           CASE
Number of Response Levels   2
Number of Observations      4195
Model                       binary logit

        Response Profile
 Ordered                  Total
   Value     CASE     Frequency

       1     CASE          1888
       2     CTRL          2307

Probability modeled is CASE='CASE'.

NOTE: 1159 observations were deleted due to missing values for the
      response or explanatory variables.
```



```
                Class Level Information

                                   Design Variables

Class     Value          1     2     3     4     5

BMI       BMI Q1         1     0     0     0
          BMI Q2         0     1     0     0
          BMI Q3         0     0     1     0
          BMI Q4         0     0     0     1
          BMI Q5        -1    -1    -1    -1

AGEFB     35+            1     0     0     0     0
          <20            0     1     0     0     0
          [20,25)        0     0     1     0     0
          [25,30)        0     0     0     1     0
          [30,35)        0     0     0     0     1
          nuliparous    -1    -1    -1    -1    -1
```



```
                Model Convergence Status
        Convergence criterion (GCONV=1E-8) satisfied.

        Model Fit Statistics
                                Intercept
                Intercept         and
Criterion         Only        Covariates

AIC             5775.585       5644.795
SC              5781.927       5803.336
-2 Log L        5773.585       5594.795


       Testing Global Null Hypothesis: BETA=0
Test                 Chi-Square       DF     Pr > ChiSq

Likelihood Ratio       178.7899       24         <.0001
Score                  174.7536       24         <.0001
Wald                   167.0518       24         <.0001
```



```
      Type III Analysis of Effects

                              Wald
Effect       DF    Chi-Square    Pr > ChiSq

EOX           1       22.4354        <.0001
PARITY        1       25.9749        <.0001
F2            4       15.2362        0.0042
BMI           4       41.2119        <.0001
AGEFB         5       10.6389        0.0590
MPAGE         4       25.4645        <.0001
MPTY          1        0.0583        0.8092
F1            4        8.5054        0.0747
```

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.1234	0.0830	2.2123	0.1369
EOX		1	0.6605	0.1394	22.4354	<.0001
PARITY		1	-0.1625	0.0319	25.9749	<.0001
F2	175+	1	-0.0504	0.1612	0.0976	0.7548
F2	<160	1	-0.1543	0.0726	4.5203	0.0335
F2	[160,165)	1	-0.0478	0.0643	0.5524	0.4573
F2	[165,170)	1	-0.0236	0.0663	0.1264	0.7222
BMI	BMI Q1	1	-0.2791	0.0680	16.8285	<.0001
BMI	BMI Q2	1	-0.0764	0.0649	1.3854	0.2392
BMI	BMI Q3	1	-0.0354	0.0644	0.3024	0.5824
BMI	BMI Q4	1	0.0483	0.0628	0.5914	0.4419
AGEFB	35+	1	0.2178	0.1381	2.4871	0.1148
AGEFB	<20	1	-0.1008	0.1024	0.9692	0.3249
AGEFB	[20,25)	1	-0.1787	0.0644	7.7027	0.0055

AGEFB	[25,30)	1	-0.0655	0.0654	1.0010	0.3171
AGEFB	[30,35)	1	0.1246	0.0878	2.0136	0.1559
MPAGE	55+	1	0.1696	0.0809	4.3896	0.0362
MPAGE	<45	1	-0.4497	0.1090	17.0204	<.0001
MPAGE	[45,50)	1	-0.0460	0.0591	0.6062	0.4362
MPAGE	[50,52)	1	0.2033	0.0617	10.8529	0.0010
MPTY		1	0.0392	0.1622	0.0583	0.8092
F1	70+	1	-0.1422	0.0596	5.6884	0.0171
F1	[50,55)	1	0.1564	0.0992	2.4868	0.1148
F1	[55,60)	1	0.0759	0.0704	1.1606	0.2813
F1	[60,65)	1	-0.0881	0.0648	1.8512	0.1736

Odds Ratio Estimates

Effect				Point Estimate	95% Wald Confidence Limits	
EOX				1.936	1.473	2.544
PARITY				0.850	0.798	0.905
F2	175+	vs	[170,175)	0.722	0.472	1.104
F2	<160	vs	[170,175)	0.650	0.522	0.811
F2	[160,165)	vs	[170,175)	0.723	0.590	0.887
F2	[165,170)	vs	[170,175)	0.741	0.602	0.912
BMI	BMI Q1	vs	BMI Q5	0.537	0.440	0.656
BMI	BMI Q2	vs	BMI Q5	0.658	0.543	0.797
BMI	BMI Q3	vs	BMI Q5	0.685	0.566	0.830
BMI	BMI Q4	vs	BMI Q5	0.745	0.618	0.899
AGEFB	35+	vs	nuliparous	1.240	0.859	1.790
AGEFB	<20	vs	nuliparous	0.902	0.652	1.246
AGEFB	[20,25)	vs	nuliparous	0.834	0.647	1.076
AGEFB	[25,30)	vs	nuliparous	0.934	0.729	1.198

AGEFB	[30,35)	vs	nuliparous	1.130	0.857	1.490
MPAGE	55+	vs	[52,55)	1.048	0.843	1.302
MPAGE	<45	vs	[52,55)	0.564	0.423	0.752
MPAGE	[45,50)	vs	[52,55)	0.845	0.713	1.001
MPAGE	[50,52)	vs	[52,55)	1.084	0.911	1.290
MPTY				1.040	0.757	1.429
F1	70+	vs	[65,70)	0.869	0.734	1.029
F1	[50,55)	vs	[65,70)	1.172	0.901	1.524
F1	[55,60)	vs	[65,70)	1.081	0.890	1.313
F1	[60,65)	vs	[65,70)	0.917	0.766	1.098

- We see that ever exclusive users of unopposed estrogen have an estimated 94% higher risk of breast cancer compared to never users of any form of HRT.
- The variable eox is coded as 1 for ever exclusive users, 0 for never users of any form of HRT, and missing (.) for women who used more than one type of HRT (who are excluded from the analysis).

- eox is not listed in the CLASS statement so the estimates refer to a 1 unit increase. Because of the coding, this corresponds to a comparison of ever to never users. The same odds ratio estimates would be obtained if eox was in the CLASS statement.
- The parameter estimate for eox is 0.6605 and we see that $\exp(0.6605) = 1.936$. That is, the parameter estimate has an interpretation as a log odds ratio.
- BMI is listed in the CLASS statement so is modelled as a categorical variable. You can think of this as having SAS create dummy variables in the background.
- SAS has chosen to use quintile 5 as the reference (I'll show you how to change this shortly) and we see that the odds ratio for Q1 vs Q5 is 0.537. The corresponding parameter estimate is -0.2791 and we see that $\exp(-0.2791) = 0.756$.
- The exponentiated parameter estimate is not the same as the odds ratio!

- This is because, by default, SAS uses what is known as effect coding for the parameter estimates whereas we are more familiar with reference cell coding.
- However, SAS always uses reference cell coding when reporting odds ratio estimates.
- With reference cell coding each parameter represents the difference between the given level and the 'reference level' whereas with effect coding each parameter represents the difference between the given level and the 'average response' (see slide 57).

- You can tell SAS to use reference cell coding by specifying the `param=ref` option on the CLASS statement.

```
proc logistic data=emma.analysis;
class mpage f1 bmi agefb f2 \ param=ref;
model case=eox parity f2 bmi agefb mpage mpty f1;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```

- We see that SAS has now constructed the design variables using the more familiar reference cell coding.

Class Level Information

Class	Value	Design Variables 1	2	3	4	5
BMI	BMI Q1	1	0	0	0	
	BMI Q2	0	1	0	0	
	BMI Q3	0	0	1	0	
	BMI Q4	0	0	0	1	
	BMI Q5	0	0	0	0	
AGEFB	35+	1	0	0	0	0
	<20	0	1	0	0	0
	[20,25)	0	0	1	0	0
	[25,30)	0	0	0	1	0
	[30,35)	0	0	0	0	1
	nuliparous	0	0	0	0	0

- These design matrices are not particularly interesting and we can suppress their display by specifying the `nodummyprint` option on the model statement.
- By default, SAS chooses the category with the highest value as the reference level. This choice is made using the formatted value, not the underlying data value. Consider, for example, the coding of age at first birth

```
value agefb
0='nuliparous'
1-<20='<20'
20-<25='[20,25)'
25-<30='[25,30)'
30-<35='[30,35)'
35-high='35+'
;
```

- The highest category based on the formatted value is 'nuliparous' whereas the highest category based on the data value is 35+. I'll describe shortly how this behaviour can be modified.

- We can specify a reference category for any variable listed in the CLASS statement.

```
proc logistic data=emma.analysis;
class mpage f1 bmi(ref='BMI Q3') agefb(ref='[25,30)') f2
                                / param=ref;
model case=eox parity f2 bmi agefb mpage mpty f1 / nodummyprint;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```


Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
EOX			1.936	1.473	2.544
PARITY			0.850	0.798	0.905
F2	175+	vs [170,175)	0.722	0.472	1.104
F2	<160	vs [170,175)	0.650	0.522	0.811
F2	[160,165)	vs [170,175)	0.723	0.590	0.887
F2	[165,170)	vs [170,175)	0.741	0.602	0.912
BMI	BMI Q1	vs BMI Q3	0.784	0.637	0.964
BMI	BMI Q2	vs BMI Q3	0.960	0.785	1.174
BMI	BMI Q4	vs BMI Q3	1.087	0.893	1.323
BMI	BMI Q5	vs BMI Q3	1.459	1.206	1.767
AGEFB	35+	vs [25,30)	1.328	0.944	1.866
AGEFB	<20	vs [25,30)	0.965	0.756	1.232
AGEFB	[20,25)	vs [25,30)	0.893	0.761	1.048
AGEFB	[30,35)	vs [25,30)	1.209	0.968	1.510
AGEFB	nuliparous	vs [25,30)	1.070	0.835	1.372

- We can also specify that the lowest, rather than the highest, category should be the default reference category.

```
proc logistic data=emma.analysis;
class mpage f1 bmi agefb f2 / param=ref ref=first;
model case=eox parity f2 bmi agefb mpage mpty f1 / nodummyprint;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```

- This may not be exactly what we want, however, since the ranking is based on the formatted values.

Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
EOX			1.936	1.473	2.544
PARITY			0.850	0.798	0.905
F2	<160	vs 175+	0.901	0.596	1.362
F2	[160,165)	vs 175+	1.003	0.670	1.501
F2	[165,170)	vs 175+	1.027	0.685	1.541
F2	[170,175)	vs 175+	1.386	0.906	2.121
BMI	BMI Q2	vs BMI Q1	1.225	0.995	1.508
BMI	BMI Q3	vs BMI Q1	1.276	1.037	1.570
BMI	BMI Q4	vs BMI Q1	1.387	1.131	1.701
BMI	BMI Q5	vs BMI Q1	1.862	1.525	2.274
AGEFB	<20	vs 35+	0.727	0.492	1.075
AGEFB	[20,25)	vs 35+	0.673	0.479	0.945
AGEFB	[25,30)	vs 35+	0.753	0.536	1.059
AGEFB	[30,35)	vs 35+	0.911	0.631	1.315
AGEFB	nuliparous	vs 35+	0.806	0.559	1.164

- We can tell SAS to instead use the 'internal' order. That is, the order according to the underlying data values.

```
proc logistic data=emma.analysis;
class mpage f1 bmi agefb f2 / param=ref ref=first order=internal;
model case=eox parity f2 bmi agefb mpage mpty f1 / nodummyprint;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```

- This means that the lowest value of age at first birth will be 0 (nuliparous) whereas when ordering was based on the formatted vales it was '35+' (see slide 47 for details of the sort order).

```
0='nuliparous'
1-<20='<20'
20-<25='[20,25)'
25-<30='[25,30)'
30-<35='[30,35)'
35-high='35+'
```


Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
EOX			1.936	1.473	2.544
PARITY			0.850	0.798	0.905
F2	[160,165)	vs <160	1.112	0.936	1.323
F2	[165,170)	vs <160	1.140	0.953	1.362
F2	[170,175)	vs <160	1.538	1.234	1.917
F2	175+	vs <160	1.110	0.734	1.677
BMI	BMI Q2	vs BMI Q1	1.225	0.995	1.508
BMI	BMI Q3	vs BMI Q1	1.276	1.037	1.570
BMI	BMI Q4	vs BMI Q1	1.387	1.131	1.701
BMI	BMI Q5	vs BMI Q1	1.862	1.525	2.274
AGEFB	<20	vs nuliparous	0.902	0.652	1.246
AGEFB	[20,25)	vs nuliparous	0.834	0.647	1.076
AGEFB	[25,30)	vs nuliparous	0.934	0.729	1.198
AGEFB	[30,35)	vs nuliparous	1.130	0.857	1.490
AGEFB	35+	vs nuliparous	1.240	0.859	1.790

Class Level Information

Class	Value	Design Variables				
BMI	BMI Q1	0	0	0	0	
	BMI Q2	1	0	0	0	
	BMI Q3	0	1	0	0	
	BMI Q4	0	0	1	0	
	BMI Q5	0	0	0	1	
AGEFB	nuliparous	0	0	0	0	0
	<20	1	0	0	0	0
	[20,25)	0	1	0	0	0
	[25,30)	0	0	1	0	0
	[30,35)	0	0	0	1	0
	35+	0	0	0	0	1

- We can even set the default reference category to be the lowest category (based on the unformatted values) while specifying specific reference categories for one or more variables.

```
proc logistic data=temp.analysis;
class mpage f1 bmi(ref='BMI Q3') agefb f2
     / param=ref ref=first order=internal;
model case=eox parity f2 bmi agefb mpage mpty f1 / nodummyprint;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```


Odds Ratio Estimates

Effect			Point Estimate	95% Wald Confidence Limits	
EOX			1.936	1.473	2.544
PARITY			0.850	0.798	0.905
F2	[160,165)	vs <160	1.112	0.936	1.323
F2	[165,170)	vs <160	1.140	0.953	1.362
F2	[170,175)	vs <160	1.538	1.234	1.917
F2	175+	vs <160	1.110	0.734	1.677
BMI	BMI Q1	vs BMI Q3	0.784	0.637	0.964
BMI	BMI Q2	vs BMI Q3	0.960	0.785	1.174
BMI	BMI Q4	vs BMI Q3	1.087	0.893	1.323
BMI	BMI Q5	vs BMI Q3	1.459	1.206	1.767
AGEFB	<20	vs nuliparous	0.902	0.652	1.246
AGEFB	[20,25)	vs nuliparous	0.834	0.647	1.076
AGEFB	[25,30)	vs nuliparous	0.934	0.729	1.198
AGEFB	[30,35)	vs nuliparous	1.130	0.857	1.490
AGEFB	35+	vs nuliparous	1.240	0.859	1.790

Statistical test for effect modification

- To test whether the effect of eox is modified by BMI we fit the interaction term between these two variables.

```
proc logistic data=temp.analysis;
class mpage f1 bmi agefb f2
     / param=ref ref=first order=internal;
model case=eox parity f2 bmi agefb mpage mpty f1 eox*bmi
     /nodummyprint expb;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```


Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
EOX	1	2.3313	0.1268
PARITY	1	25.9509	<.0001
F2	4	15.4054	0.0039
BMI	4	40.1326	<.0001
AGEFB	5	10.3026	0.0671
MPAGE	4	25.8703	<.0001
MPTY	1	0.0207	0.8856
F1	4	8.4851	0.0753
EOX*BMI	4	2.5542	0.6350

- There is no evidence of a statistically significant interaction.

Estimating the effect of eox for each category of BMI

- The previous example showed how to formally test for effect modification although the parameter estimates of the resulting model do not have a useful interpretation.
- To estimate the effect of eox for each category of BMI we use the following.

```
proc logistic data=temp.analysis;
class mpage f1 bmi agefb f2
     / param=ref ref=first order=internal;
model case=parity f2 bmi agefb mpage mpty f1 eox(bmi)
     /nodummyprint expb;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```

- This estimates the same model as the previous slide but some parameters have different interpretations.

- The term eox(bmi) provides estimates of the effect of eox nested within BMI.
- SAS does not seem to report odds ratios for any variables that figure in interaction terms in the 'Table of odds ratio estimates'.
- The expb option makes SAS report the exponentiated parameter estimates in the table of estimates, but unfortunately there are no confidence intervals.

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Wald ChiSq	Pr(ChiSq)	Exp(Est)
Intercept		−0.5947	0.2130	7.7927	0.0052	0.552
PARITY		−0.1625	0.0319	25.9509	<.0001	0.850
F2	[160,165)	0.1110	0.0884	1.5796	0.2088	1.117
F2	[165,170)	0.1361	0.0912	2.2274	0.1356	1.146
F2	[170,175)	0.4341	0.1126	14.8702	0.0001	1.544
F2	175+	0.1120	0.2109	0.2821	0.5953	1.119

BMI	BMI Q2	0.1684	0.1099	2.3466	0.1256	1.183
BMI	BMI Q3	0.2393	0.1091	4.8114	0.0283	1.270
BMI	BMI Q4	0.3062	0.1073	8.1429	0.0043	1.358
BMI	BMI Q5	0.6217	0.1049	35.1421	<.0001	1.862
AGEFB	<20	−0.1044	0.1654	0.3986	0.5278	0.901
AGEFB	[20,25)	−0.1794	0.1300	1.9026	0.1678	0.836
AGEFB	[25,30)	−0.0649	0.1269	0.2617	0.6089	0.937
AGEFB	[30,35)	0.1209	0.1411	0.7332	0.3919	1.128
AGEFB	35+	0.2078	0.1874	1.2290	0.2676	1.231
MPTY		0.0234	0.1629	0.0207	0.8856	1.024
EOX(BMI)	BMI Q1	0.4777	0.3129	2.3313	0.1268	1.612
EOX(BMI)	BMI Q2	0.9736	0.2880	11.4276	0.0007	2.648
EOX(BMI)	BMI Q3	0.5398	0.3064	3.1034	0.0781	1.716
EOX(BMI)	BMI Q4	0.8430	0.3216	6.8714	0.0088	2.323
EOX(BMI)	BMI Q5	0.4552	0.2888	2.4850	0.1149	1.576

- The estimates of the effect of eox are similar for each category of BMI (as we might expect since there was no evidence of a statistically significant interaction).

Sort order for character variables

- From the smallest to largest displayable character, the English-language ASCII sequence is

```
blank ! " # $ % & ' ( ) * + , - . /0 1 2 3 4 5 6 7 8 9 :
; < = > ? @
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z[ \] ^_
a b c d e f g h i j k l m n o p q r s t u v w x y z { } ~
```

- The main features of the ASCII sequence are that digits are sorted before uppercase letters, and uppercase letters are sorted before lowercase letters. The blank is the smallest displayable character.
- Missing (blank) values of character variables are smaller than any printable character value.

Sort order for numeric variables

Sort order	Symbol	Description
smallest	._	underscore
	.	period
	.A-.Z	special missing values
	$-n$	negative numbers
	0	zero
largest	$+n$	positive numbers

The ORDER= option on the CLASS statement in PROC LOGISTIC

ORDER=DATA order of appearance in the input data set
ORDER=FORMATTED external formatted value, except for numeric variables with no explicit format, which are sorted by their unformatted (internal) value
ORDER=FREQ descending frequency count; levels with the most observations come first in the order
ORDER=INTERNAL unformatted value

Options for confidence intervals for odds ratios — the CLODDS option to the MODEL statement

- By default, confidence intervals are based on individual Wald tests (`CLODDS=WALD`).
- Confidence intervals based on the profile likelihood can be obtained by specifying `CLODDS=PL`.
- By specifying `CLPARM=BOTH`, the procedure computes two sets of confidence intervals for the odds ratios, one based on the profile likelihood and the other based on the Wald tests.
- The confidence coefficient can be specified with the `ALPHA=` option.

The Hosmer-Lemeshow Goodness-of-Fit Test

- The Pearson and deviance goodness-of-fit tests are not valid for sparse data. Hosmer and Lemeshow (1989) proposed an alternative test.
- First, the observations are sorted in increasing order of their estimated event probability.
- The observations are then divided into approximately ten groups.
- The observed and expected number of events are then tabulated for each group.
- The test statistic takes the form of the standard comparison of observed to expected events.
- SAS calculates the Hosmer-Lemeshow goodness-of-fit test when the `LACKFIT` option is specified on the MODEL statement.


```
                    CASE = CASE          CASE = CTRL
Group  Total    Observed   Expected   Observed   Expected

    1    420         113     117.66        307     302.34
    2    422         151     145.64        271     276.36
    3    418         168     158.88        250     259.12
    4    420         172     170.65        248     249.35
    5    420         192     181.67        228     238.33
    6    420         182     192.75        238     227.25
    7    421         197     204.66        224     216.34
    8    420         209     217.78        211     202.22
    9    420         233     235.00        187     185.00
   10    414         271     263.23        143     150.77

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square       DF     Pr > ChiSq
    5.5258        8          0.7002
```


Specifying the units for odds ratio estimates

- By default, odds ratios for continuous explanatory variables are estimated for each one unit change in the corresponding explanatory variable.
- In the CAHRES study, duration of HRT use is recorded in days.
- The UNITS statement enables you to specify units of change so that customized odds ratios can be estimated. For example, we may wish to estimate the odds ratio for each year of use.

```
proc logistic data=temp.analysis;
class mpage f1 bmi agefb f2 / param=ref ref=first order=internal;
model case=eodu parity f2 bmi agefb mpage mpty f1;
units eodu=365.25;
format mpage mpage. f1 age. bmi bmi. agefb agefb. f2 height.;
run;
```

- We could, alternatively, create a new variable containing duration in years.

Interpreting the estimated regression coefficients when using effect coding

- Consider again the case when the logistic regression model involves only one explanatory variable, but we instead code $X_1 = 1$ for the exposed and $X_1 = -1$ for the unexposed.
- The underlying logistic regression model is still the same,

$$\log\left\{\frac{\pi(X_1)}{1-\pi(X_1)}\right\} = \beta_0 + \beta_1 X_1,$$

although the parameters now have different interpretations.

- The log odds for the exposed and unexposed, expressed as functions of the parameters are,

$$\log\left\{\frac{\pi(X_1=1)}{1-\pi(X_1=1)}\right\} = \beta_0 + \beta_1$$

for the exposed individuals ($X_1 = 1$) and

$$\log\left\{\frac{\pi(X_1=-1)}{1-\pi(X_1=-1)}\right\} = \beta_0 - \beta_1$$

for the unexposed individuals ($X_1 = -1$).

- The estimated log OR is given by

$$\begin{aligned}\log\left\{\frac{\text{odds} \mid \text{exposed}}{\text{odds} \mid \text{unexposed}}\right\} &= \log(\text{odds} \mid \text{exposed}) - \log(\text{odds} \mid \text{unexposed}) \\ &= (\beta_0 + \beta_1) - (\beta_0 - \beta_1) \\ &= 2\beta_1\end{aligned}$$

- The estimated odds ratio under this parameterisation is therefore $\exp(2\beta_1)$.

Interpretation of β_1 when using effect coding

- Consider again the case where the model involves only one explanatory variable coded $X_1 = 1$ for the exposed and $X_1 = -1$ for the unexposed.

$$\begin{aligned}\beta_1 &= \frac{1}{2}\{\text{logodds(exposed)} - \text{logodds(unexposed)}\} \\ &= \text{logodds(exposed)} - \frac{\{\text{logodds(exposed)} + \text{logodds(unexposed)}\}}{2}\end{aligned}$$

- β_1 represents the log OR comparing the exposed to the (unweighted) average of the exposed and unexposed.
- The interpretation is the same for more than two categories.