# Additional Lecture – Working with Epidemiologic Data

## Collect with your research question in mind!

- Collect more, not less
- Design questionnaire with database in mind – proper design would help to avoid entry errors
- Select useful categories
- Be specific, but not over-specific
- Keep coding consistent across variables (for example, low risk group=0, Yes/No 1/0, don't know, missing)

## Analysis protocol

- Specify variables of interest, both dependent and independent
- Specify modeling approaches: continuous, categorical, etc.
- Analytical techniques
- Covariates and their modeling
- Any subset or secondary analyses
- The protocol is dynamic and not written in stone!

## Getting started

- Keeping record of decision making process is extremely important!
- Explore data dictionary- in database management systems, a file that defines the basic organization of a database. A data dictionary contains a list of all files in the database, the number of records in each file, and the names and types of each field
- Keep clear documentation of programs and outputs- smaller files are easier to work with in the future
- Save SAS programs with different date stamps
- Clearly label analysis sections in the SAS program-you can say something like "Table 1", etc. Have a preface to describe the program objectives, date created or modified, important variable definitions
- Document program specifics either separately or in SAS (as /*….*/)
- Run the program step-by-step to catch mistakes

## Data clean-up

- Familiarize yourself with the data
  - Continuous: run proc univariate, proc means, how many missing data points, any unusual values?
  - Categorical: any unusual coding? Any unusual distribution? How many missing?

## Dealing with missing data

- Create a separate category for unknown
- Impute median values of a given variable in controls for any missing – if large, could lead to biased estimation → conduct sensitivity analysis later
- Search the literature for possible imputation approaches

## Try different modeling approaches

- 1 unit change
- below vs. above median
- Quartiles based on distribution
- 10, 50, 100 unit change

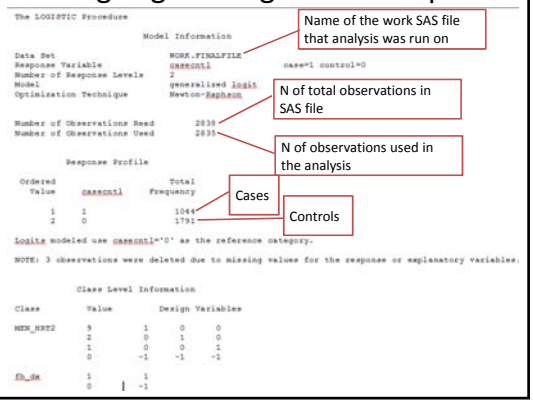
## Binary logistic regression in SAS

Need to specify "descending" if cases=1 and controls=0

```
proc logistic data=alcohol1 descending;
class MEN_HRT2/descending;
class FH_DX/descending;
class PAR_LR/descending;
class smoking/descending;
class MENAR_LR/descending;
class cavgpct2/descending;
class AGEMEN_LR/descending;
class BBDC_DX/descending;
model CASEcntl=Cavgpct2 BBDC_DX AGE_DX CBMI_DX MEN_HRT2
PAR_LR FH_DX smoking MENAR_LR AGEMEN_LR/link=glogit
scale=none aggregate waldcl waldrl lackfit rsquare;
title 'Percent density as usual categories-stratified model, alcohol=1';
run;
```

AGGREGATE- determines subpopulations for Pearson chi-square and deviance
LACKFIT - requests Hosmer and Lemeshow goodness-of-fit test
SCALE - specifies method to correct overdispersion
RSQUARE - requests a generalized measure for the fitted model
waldcl waldrl –requests to calculate Parameter Estimates and Wald Confidence Intervals

## Reading Logistic Regression Output



```
The LOGISTIC Procedure

Model Information

Data Set                          WORK.FINALFILE
Response Variable                 casecntl          case=1 control=0
Number of Response Levels         2
Model                             generalized logit
Optimization Technique            Newton-Raphson

Number of Observations Read       2835
Number of Observations Used       2835

Response Profile

Ordered                           Total
  Value     casecntl   Frequency

      1            1        1044
      2            0        1791

Logits modeled use casecntl='0' as the reference category.

NOTE: 3 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information

Class         Value        Design Variables

MEN_HRT2      9            1     0     0
              2            0     1     0
              1            0     0     1
              0           -1    -1    -1

fh_dx         1            1
              0           -1
```

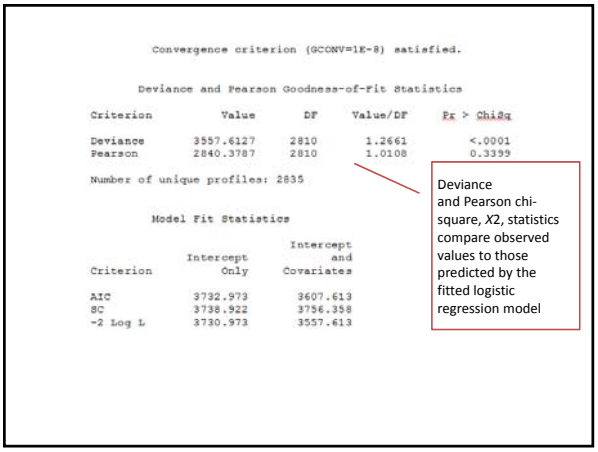
Name of the work SAS file that analysis was run on

N of total observations in SAS file

N of observations used in the analysis

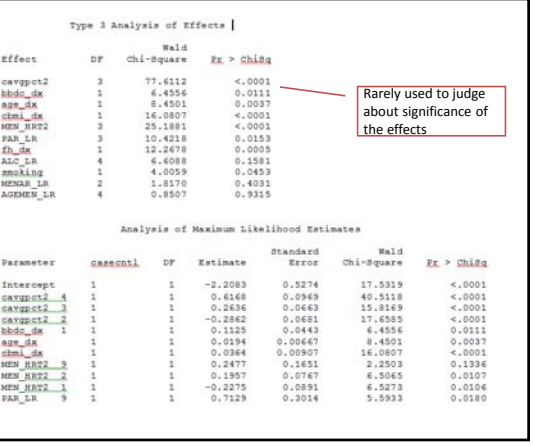
Cases

Controls



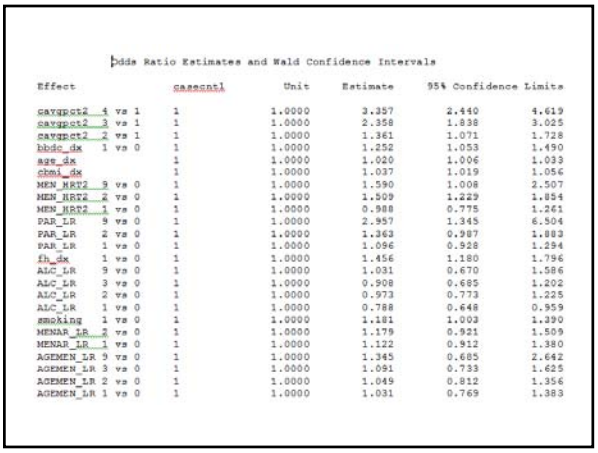
```
Convergence criterion (GCONV=1E-8) satisfied.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion          Value      DF     Value/DF    Pr > ChiSq

Deviance        3557.6127    2810       1.2661        <.0001
Pearson         2840.3787    2810       1.0108        0.3399

Number of unique profiles: 2835

Model Fit Statistics

                              Intercept
               Intercept          and
Criterion           Only   Covariates

AIC             3732.973     3607.613
SC              3738.922     3756.358
-2 Log L        3730.973     3557.613
```

Deviance and Pearson chi-square, *X2*, statistics compare observed values to those predicted by the fitted logistic regression model



```
Type 3 Analysis of Effects

                        Wald
Effect          DF    Chi-Square    Pr > ChiSq

cavgpct2         3       77.6112        <.0001
bbdc_dx          1        6.4556        0.0111
age_dx           1        8.4501        0.0037
cbmi_dx          1       16.0807        <.0001
MEN_HRT2         3       25.1881        <.0001
PAR_LR           3       10.4218        0.0153
fh_dx            1       12.2678        0.0005
ALC_LR           4        6.6088        0.1581
smoking          1        4.0059        0.0453
MENAR_LR         2        1.8170        0.4031
AGEMEN_LR        4        0.8507        0.9315

Analysis of Maximum Likelihood Estimates

                                          Standard          Wald
Parameter       casecntl   DF   Estimate      Error    Chi-Square    Pr > ChiSq

Intercept              1    1    -2.2083     0.5274       17.5319        <.0001
cavgpct2     4         1    1     0.6168     0.0969       40.5118        <.0001
cavgpct2     3         1    1     0.2636     0.0663       15.8169        <.0001
cavgpct2     2         1    1    -0.2862     0.0681       17.6585        <.0001
bbdc_dx      1         1    1     0.1125     0.0443        6.4556        0.0111
age_dx                 1    1     0.0194    0.00667        8.4501        0.0037
cbmi_dx                1    1     0.0364    0.00907       16.0807        <.0001
MEN_HRT2     9         1    1     0.2477     0.1651        2.2503        0.1336
MEN_HRT2     2         1    1     0.1957     0.0767        6.5065        0.0107
MEN_HRT2     1         1    1    -0.2275     0.0891        6.5273        0.0106
PAR_LR       9         1    1     0.7129     0.3014        5.5933        0.0180
```

Rarely used to judge about significance of the effects



```
Odds Ratio Estimates and Wald Confidence Intervals

Effect                        casecntl    Unit      Estimate    95% Confidence Limits

cavgpct2   4 vs 1                    1  1.0000        3.357        2.440        4.619
cavgpct2   3 vs 1                    1  1.0000        2.358        1.838        3.025
cavgpct2   2 vs 1                    1  1.0000        1.361        1.071        1.728
bbdc_dx    1 vs 0                    1  1.0000        1.252        1.053        1.490
age_dx                               1  1.0000        1.020        1.006        1.033
cbmi_dx                              1  1.0000        1.037        1.019        1.056
MEN_HRT2   9 vs 0                    1  1.0000        1.590        1.008        2.507
MEN_HRT2   2 vs 0                    1  1.0000        1.509        1.229        1.854
MEN_HRT2   1 vs 0                    1  1.0000        0.988        0.775        1.261
PAR_LR     9 vs 0                    1  1.0000        2.957        1.345        6.504
PAR_LR     2 vs 0                    1  1.0000        1.363        0.987        1.883
PAR_LR     1 vs 0                    1  1.0000        1.096        0.928        1.294
fh_dx      1 vs 0                    1  1.0000        1.456        1.180        1.796
ALC_LR     9 vs 0                    1  1.0000        1.031        0.670        1.586
ALC_LR     3 vs 0                    1  1.0000        0.908        0.685        1.202
ALC_LR     2 vs 0                    1  1.0000        0.973        0.773        1.225
ALC_LR     1 vs 0                    1  1.0000        0.788        0.648        0.959
smoking    1 vs 0                    1  1.0000        1.181        1.002        1.390
MENAR_LR   2 vs 0                    1  1.0000        1.179        0.921        1.509
MENAR_LR   1 vs 0                    1  1.0000        1.122        0.912        1.380
AGEMEN_LR  9 vs 0                    1  1.0000        1.345        0.685        2.642
AGEMEN_LR  3 vs 0                    1  1.0000        1.091        0.733        1.625
AGEMEN_LR  2 vs 0                    1  1.0000        1.049        0.812        1.356
AGEMEN_LR  1 vs 0                    1  1.0000        1.031        0.769        1.383
```

Partition for the Hosmer and Lemeshow Test

| Group | Total | chdcntl = 1 Observed | chdcntl = 1 Expected | chdcntl = 0 Observed | chdcntl = 0 Expected |
|---|---|---|---|---|---|
| 1 | 284 | 60 | 54.30 | 224 | 229.70 |
| 2 | 284 | 64 | 60.83 | 220 | 215.17 |
| 3 | 284 | 70 | 78.86 | 214 | 205.14 |
| 4 | 284 | 91 | 87.71 | 193 | 196.29 |
| 5 | 284 | 99 | 96.31 | 185 | 187.69 |
| 6 | 284 | 104 | 105.33 | 180 | 178.67 |
| 7 | 284 | 107 | 116.62 | 177 | 167.38 |
| 8 | 284 | 144 | 128.51 | 140 | 155.49 |
| 9 | 284 | 137 | 142.18 | 147 | 141.82 |
| 10 | 279 | 168 | 165.34 | 111 | 113.66 |

Hosmer and Lemeshow Goodness-of-Fit Test

| Chi-Square | DF | Pr > ChiSq |
|---|---|---|
| 8.1259 | 8 | 0.4213 |

Overall model fit, the larger p value, the better

By specifying lackfit option for model statement

## Modeling interactions

- Continuous independent variables
  - Model BMI=X1 X2 X1*X2
  - Interpretation is tricky and is rarely used, other than saying that interaction exists hard to make sense of that interaction

## Two categorical variables

- Consider a study of the analgesic effects of anti-pain treatments on elderly patients with neuralgia
  - Dependent variable: PAIN - Yes/No
  - Independent: TREATMENT - Two treatments + placebo
  - Covariates: Age (continuous) and Gender (F, M)

proc logistic descending;
class Treatment Sex;
model Pain= Treatment Sex Age Duration/link=glogit scale=none aggregate waldcl waldrl lackfit rsquare;
run;

proc logistic data=Neuralgia;
class Treatment Sex; model Pain= Treatment Sex Treatment*Sex Age Duration/expb; run;

- PROC LOGISTIC displays a table of the Type 3 analysis of effects
- Since the model contains the Treatment*Sex interaction term, the odds ratios for Treatment and Sex are not computed

Type 3 Analysis of Effects

| Effect | DF | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|
| Treatment | 2 | 11.9886 | 0.0025 |
| Sex | 1 | 5.3104 | 0.0212 |
| Treatment*Sex | 2 | 0.1412 | 0.9318 |
| Age | 1 | 7.2744 | 0.0070 |
| Duration | 1 | 0.0247 | 0.8752 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| Age | 0.764 | 0.629 | 0.929 |
| Duration | 1.005 | 0.942 | 1.073 |

### Test for multiplicative interaction with a cross-product

```
proc logistic data = practice descending;
model disease = famhx smoking famhx* smoking;
oddsratio famhx / at(smoking = 0 1);
oddsratio smoking / at(famhx = 0 1);
run;
```

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -4.5591 | 0.7108 | 41.1409 | <.0001 |
| **Famhx** | 1 | 2.3869 | 0.7931 | 9.0576 | 0.0026 |
| **smoking** | 1 | 1.1579 | 1.0109 | 1.3120 | 0.2520 |
| **famhx* smoking** | **1** | **0.2411** | **1.0984** | **0.0482** | **0.8262** |

Wald Confidence Interval for Odds Ratios

| Label | Estimate | 95% Confidence |
|---|---|---|
| famhx at smoking =0 | 10.880 | 2.299 51.486 |
| famhx at smoking =1 | 13.846 | 3.122 61.408 |
| smoking at famhx =0 | 3.183 | 0.439 23.086 |
| smoking at famhx =1 | 4.051 | 1.745 9.405 |

multiplicative interaction is not significant

## Best strategy

- Use p-value for Type 3 effects to determine whether the interaction is significant
- Perform stratified analysis to report strata-specific ORs

## Testing for additive interaction with a cross-product

```
proc genmod data practice descending;
model disease= famhx   smoking   famhx* smoking /link = identity dist = binomial lrci;
estimate 'RD of famhx when smoking = 0' famhx 1;
estimate 'RD of famhx when smoking = 1' famhx 1 famhx*smoking 1;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates  Likelihood Ratio

| Parameter | DF | Estimate | Standard Error | 95% Confidence Limits | | Wald Chi-Square | Pr>ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 0.0104 | 0.0073 | 0.0017 | 0.0317 | 2.02 | 0.1551 |
| famhx | 1 | 0.0919 | 0.0331 | 0.0368 | 0.1675 | 7.70 | 0.0055 |
| smoking | 1 | 0.0219 | 0.0236 | -0.0139 | 0.0870 | 0.86 | 0.3534 |
| famhx *smoking | 1 | **0.1916** | **0.0667** | **0.0588** | **0.3219** | **8.26** | **0.0040** |

**Interaction is statistically significant "additive interaction".**

## Variable Manipulation

- If continuous variables by nature, you can create a new variable that represents median values in each category and use that variable in interaction term
- Example: BMI in kg/m$^2$
  - Categorical BMI <25, 25-<30, 30-<35
  - Median for BMI: 23, 27, 32
  - New variable: 23, 27, 32

## Another option: create a combined variable (joint effects)

- Family history of cancer and smoking
- Variable combined:
  - 0-no family history, no smoking
  - 1-family history, no smoking
  - 2-no family history, smoking
  - 3-family history and smoking
- And then model it without class statement to get p-value for significance
- But this is usually used to demonstrate joint effects rather than to test interaction significance

## More on interaction…

- In other words, just because a significant effect is found in one group and not in the other, does NOT mean the effects are necessarily different in the two groups
- Remember, statistical significance is not only a function of the effect but also the sample size and the baseline risk. Both of these can differ across groups.

## Types of logistic regression

- Binary (ordinary) Yes/No
- Ordinal or ordered – For ordinal outcome variables assumes that the coefficients that describe the relationship between the lowest versus all higher categories of the response variable are the same as those that describe the relationship between the next lowest category and all higher categories, etc. (proportional odds assumption or the parallel regression assumption) - for example BMI
- Polychotomous logistic regression – for nominal outcome variables, for example, lobular, ductal breast cancer, controls