## Logistic regression

- When response variable is measured/counted, regression can work well.
- But what if response is yes/no, lived/died/ success/failure?
- Model probability of success.
- Probability must be between 0 and 1; need method that ensures this.
- Logistic regression does this; PROC LOGISTIC in SAS.
- Begin with simplest case.

### The rats, part 1

Rats given dose of some poison; either live or die:

- 0 lived
- 1 died
- 2 lived
- 3 lived
- 4 died
- 5 died

### Basic logistic regression analysis:

```
options linesize=80;
```

```
data rat;
infile "rat.dat";
input dose survival $;
```

```
proc logistic;
class survival;
model survival = dose;
output out=rat2 pred=pred;
```

```
proc print data=rat2;
```

# Output

#### The LOGISTIC Procedure

#### Model Information

| Data Set                   | WORK.RAT         |  |
|----------------------------|------------------|--|
| Response Variable survival |                  |  |
| Number of Response Levels  | 2                |  |
| Model                      | binary logit     |  |
| Optimization Technique     | Fisher's scoring |  |

| Number | of | Observations | Read | 6 |
|--------|----|--------------|------|---|
| Number | of | Observations | Used | 6 |

#### Response Profile

| rdered |          | Total     |
|--------|----------|-----------|
| Value  | survival | Frequency |
|        |          |           |
| 1      | died     | 3         |
| 2      | lived    | 3         |

Probability modeled is survival='died'.

# **Output part 2 (edited)**

... snip

Wald

Model Convergence Status

| Convergence cri  | lterion (GCONV=1E- | -8) satı | siled.     |
|------------------|--------------------|----------|------------|
| Testing Glob     | oal Null Hypothesi | IS: BETA | x=0        |
| Test             | Chi-Square         | DF       | Pr > ChiSq |
| Likelihood Ratio | 1.5449             | 1        | 0.2139     |
| Score            | 1.4286             | 1        | 0.2320     |
|                  |                    |          |            |

#### Analysis of Maximum Likelihood Estimates

1.2037

1

|           |    |          | Standard | Wald       |            |
|-----------|----|----------|----------|------------|------------|
| Parameter | DF | Estimate | Error    | Chi-Square | Pr > ChiSq |
|           |    |          |          |            |            |
| Intercept | 1  | -1.6841  | 1.7978   | 0.8774     | 0.3489     |
| dose      | 1  | 0.6736   | 0.6140   | 1.2037     | 0.2726     |

0.2726

### Interpreting the output

- Like (multiple) regression, get:
  - overall test of model ("global null hypothesis")
  - tests of significance of individual x's ("analysis of maximum likelihood estimates").
- Here none of them significant (only 6 observations).
- These tests all agree for regression, but don't for logistic regression. Look for consistent picture (Wald often different from others).
- Look at event "modeled", here "died".
- "Slope" for dose is positive, meaning that as dose increases, probability of event modelled (death) increases.
- Output data set contains predicted probabilities (next slide):

## **Predicted probabilities**

| Obs | dose | survival | _LEVEL_ | pred    |
|-----|------|----------|---------|---------|
| 1   | 0    | lived    | died    | 0.15656 |
| 2   | 1    | died     | died    | 0.26690 |
| 3   | 2    | lived    | died    | 0.41658 |
| 4   | 3    | lived    | died    | 0.58342 |
| 5   | 4    | died     | died    | 0.73310 |
| 6   | 5    | died     | died    | 0.84344 |
|     |      |          |         |         |

"Pred" is predicted probability of event named by \_LEVEL\_ (death). Goes up as dose increases.

### The rats, part 2

- More realistic: more rats at each dose (say 10).
- Listing each rat on one line makes a big data file.
- Use format below: dose, number of deaths, number of trials (rats):
  - 0 0 10
  - 1 3 10
  - 2 4 10
  - 3 6 10
  - 4 8 10
  - 5 9 10

Alter model line for PROC LOGISTIC to say: model deaths/trials = dose;

## SAS code for this logistic regression

```
options linesize=80;
```

```
data rat;
infile "rat2.dat";
input dose deaths trials;
```

```
proc logistic;
model deaths/trials = dose;
output out=rat2 pred=pred lower=lcl upper=ucl;
```

```
proc print data=rat2;
```

This time, have output data set also contain lower and upper limits of a 95% CI for each death probability.

# **Output part 1 (edited)**

| Number | of Observations Read | 6  |
|--------|----------------------|----|
| Number | of Observations Used | 6  |
| Sum of | Frequencies Read     | 60 |
| Sum of | Frequencies Used     | 60 |

#### Response Profile

| Ordered | Binary   | Total     |  |
|---------|----------|-----------|--|
| Value   | Outcome  | Frequency |  |
|         |          |           |  |
| 1       | Event    | 30        |  |
| 2       | Nonevent | 30        |  |
|         |          |           |  |

Model Convergence Status

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

The 6 lines of data correspond to 60 actual rats.

## **Output part 2 (edited)**

Testing Global Null Hypothesis: BETA=0

| Test             | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 25.0562    | 1  | <.0001     |
| Score            | 21.9657    | 1  | <.0001     |
| Wald             | 16.1449    | 1  | <.0001     |

Analysis of Maximum Likelihood Estimates

|           |    |          | Standard | Wald       |            |
|-----------|----|----------|----------|------------|------------|
| Parameter | DF | Estimate | Error    | Chi-Square | Pr > ChiSq |
|           |    |          |          |            |            |
| Intercept | 1  | -2.3619  | 0.6719   | 12.3585    | 0.0004     |
| dose      | 1  | 0.9448   | 0.2351   | 16.1449    | <.0001     |

- All 4 tests agree: significant effect of dose.
- Effect of larger dose is to increase death probability ("slope" positive).

### Just run PROC PRINT on output data set:

| 0bs | dose | deaths | trials | pred    | lcl     | ucl     |
|-----|------|--------|--------|---------|---------|---------|
| 1   | 0    | 0      | 10     | 0.08612 | 0.02463 | 0.26017 |
| 2   | 1    | 3      | 10     | 0.19511 | 0.08646 | 0.38304 |
| 3   | 2    | 4      | 10     | 0.38405 | 0.24041 | 0.55124 |
| 4   | 3    | 6      | 10     | 0.61595 | 0.44876 | 0.75959 |
| 5   | 4    | 8      | 10     | 0.80489 | 0.61696 | 0.91354 |
| 6   | 5    | 9      | 10     | 0.91388 | 0.73983 | 0.97537 |
|     |      |        |        |         |         |         |

- Predicted death probs increase with dose.
- Last 2 columns are 95% CI for prob of death at each dose (eg. dose 2, from 0.24 to 0.55).
- Intervals still quite wide even with n = 60 rats.
- Each rat doesn't contribute much information (just lived/died) so need n in hundreds to get precise intervals.

## **Multiple logistic regression**

- With more than one x, works much like multiple regression.
- Example: study of patients with blood poisoning severe enough to warrant surgery. Relate survival to other potential risk factors.
- Variables, 1=present, 0=absent:
  - survival (death from sepsis=1), response
  - shock
  - malnutrition
  - alcoholism
  - age (as numerical variable)
  - bowel infarction
- See what relates to death.

### Some SAS code

```
data x;
infile "sepsis.dat";
input death shock malnut alcohol age bowelinf;
proc logistic;
model death=shock malnut alcohol age bowelinf;
test malnut=0, bowelinf=0;
proc logistic;
model death=shock alcohol age bowelinf;
output out=z pred=p;
```

```
proc print data=z;
```

Use of PROC LOGISTIC resembles use of PROC REG, including "test".

### Output part 1

Number of Observations Used 106

#### Response Profile

| Total     |       | Ordered |
|-----------|-------|---------|
| Frequency | death | Value   |
| 85        | 0     | 1       |
| 21        | 1     | 2       |

Probability modeled is death=0.

Testing Global Null Hypothesis: BETA=0

| Test             | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
|                  |            |    |            |
| Likelihood Ratio | 52.4060    | 5  | <.0001     |
| Score            | 43.8921    | 5  | <.0001     |
| Wald             | 16.2433    | 5  | 0.0062     |

Model as a whole is significant: at least one of the x's helps predict death (actually modelling P(survival)).

# Finding significant *x*'s

Analysis of Maximum Likelihood Estimates

|           |    |          | Standard | Wald       |            |
|-----------|----|----------|----------|------------|------------|
| Parameter | DF | Estimate | Error    | Chi-Square | Pr > ChiSq |
|           |    |          |          |            |            |
| Intercept | 1  | 9.7539   | 2.5417   | 14.7267    | 0.0001     |
| shock     | 1  | -3.6739  | 1.1648   | 9.9479     | 0.0016     |
| malnut    | 1  | -1.2166  | 0.7282   | 2.7909     | 0.0948     |
| alcohol   | 1  | -3.3549  | 0.9821   | 11.6691    | 0.0006     |
| age       | 1  | -0.0922  | 0.0303   | 9.2353     | 0.0024     |
| bowelinf  | 1  | -2.7976  | 1.1640   | 5.7767     | 0.0162     |

- Only marginal one is malnut.
- Test that both malnut and bowelinf can be removed (suspect not):

|            |    | Wald       |        |
|------------|----|------------|--------|
| Pr > ChiSq | DF | Chi-Square | Label  |
| 0.0329     | 2  | 6.8302     | Test 1 |

■ Indeed, not.

### **Predictions from model without "malnut"**

### So fit model without malnut and obtain predictions.

### A few chosen at random:

| Obs | death | shock | malnut | alcohol | age | bowelinf | _LEVEL_ | p       |
|-----|-------|-------|--------|---------|-----|----------|---------|---------|
| 4   | 0     | 0     | 0      | 0       | 26  | 0        | 0       | 0.99858 |
| 1   | 0     | 0     | 0      | 0       | 56  | 0        | 0       | 0.97945 |
| 2   | 0     | 0     | 0      | 0       | 80  | 0        | 0       | 0.84658 |
|     |       |       |        |         |     |          |         |         |
| 11  | 1     | 0     | 0      | 1       | 66  | 1        | 0       | 0.06871 |
| 32  | 1     | 0     | 0      | 1       | 49  | 0        | 0       | 0.78700 |

- Survival chances pretty good if no risk factors, though decreasing with age.
- Having more than one risk factor reduces survival chances dramatically.
- Usually model does a good job of predicting survival, but occasionally someone dies who was predicted to survive.

## **Changing the response category**

- In first rats example, got prob of death but maybe wanted prob of living.
- Change model line to this:
  - model survival(event='lived') = dose;
- Output now includes:

|           |     |          | Standa   | ırd      | Wa] | _d        |   |
|-----------|-----|----------|----------|----------|-----|-----------|---|
| arameter  | DF  | Estimate | Error    | Chi-Squa | are | Pr > ChiS | q |
| intercept | 1   | 1.6841   | 1.7978   | 0.87     | 74  | 0.348     | 9 |
| lose      | 1   | -0.6736  | 0.6140   | 1.20     | )37 | 0.272     | б |
|           |     |          |          |          |     |           |   |
|           | Obs | dose     | survival | _LEVEL_  | I   | pred      |   |
|           | 1   | 0        | lived    | lived    | 0.8 | 34344     |   |
|           | 2   | 1        | died     | lived    | 0.7 | 73310     |   |
|           | 3   | 2        | lived    | lived    | 0.5 | 58342     |   |
|           | 4   | 3        | lived    | lived    | 0.4 | l1658     |   |
|           | 5   | 4        | died     | lived    | 0.2 | 26690     |   |
|           | 6   | 5        | died     | lived    | 0.1 | 5656      |   |

### **Testing fit: seroconversion example**

- Seroconversion: body develops specific antibodies to microorganisms in blood (as when person gets certain disease).
- Seropositive: still have antibodies in blood after recovery from the disease.
- Malaria survey: ages plus seropositiveness recorded. Data, with variables: age group number, middle of age group, #individuals, #seropositive:

11.5123824.0132637.518218412.514014517.513820625.016139735.013319847.09225960.07444

### **Does seropositiveness depend on age?**

Calculate observed pct of seropositives for each age group in DATA step:

```
data sero;
infile "sero.dat";
input group age n r;
obspos=r/n;
```

proc print;
with this result:

| Obs | group | age  | n   | r  | obspos  |
|-----|-------|------|-----|----|---------|
|     |       |      |     |    |         |
| 1   | 1     | 1.5  | 123 | 8  | 0.06504 |
| 2   | 2     | 4.0  | 132 | 6  | 0.04545 |
| 3   | 3     | 7.5  | 182 | 18 | 0.09890 |
| 4   | 4     | 12.5 | 140 | 14 | 0.10000 |
| 5   | 5     | 17.5 | 138 | 20 | 0.14493 |
| 6   | 6     | 25.0 | 161 | 39 | 0.24224 |
| 7   | 7     | 35.0 | 133 | 19 | 0.14286 |
| 8   | 8     | 47.0 | 92  | 25 | 0.27174 |
| 9   | 9     | 60.0 | 74  | 44 | 0.59459 |

### **Does a logistic regression fit?**

- Prob of being seropositive generally increases with age, but age group 6 has too many seropositives and age group 7 too few.
- Fit logistic model anyway, and test for fit.
- Hosmer-Lemeshow test:
  - null: logistic regression is appropriate
  - alternative: it is not.
- Code (note "events/trials" syntax and "lackfit"): proc logistic;

```
model r/n = age / lackfit;
```

# **Hosmer-Lemeshow test output**

Partition for the Hosmer and Lemeshow Test

|       |       |          | rent     | Nonevent |          |  |
|-------|-------|----------|----------|----------|----------|--|
| Group | Total | Observed | Expected | Observed | Expected |  |
| 1     | 123   | 8        | 8 14     | 115      | 114 86   |  |
| 2     | 132   | 6        | 9.69     | 126      | 122.31   |  |
| 3     | 182   | 18       | 15.43    | 164      | 166.57   |  |
| 4     | 140   | 14       | 14.53    | 126      | 125.47   |  |
| 5     | 138   | 20       | 17.46    | 118      | 120.54   |  |
| 6     | 161   | 39       | 27.11    | 122      | 133.89   |  |
| 7     | 133   | 19       | 31.97    | 114      | 101.03   |  |
| 8     | 92    | 25       | 32.30    | 67       | 59.70    |  |
| 9     | 74    | 44       | 36.38    | 30       | 37.62    |  |

Hosmer and Lemeshow Goodness-of-Fit Test

| Chi-Square | DF | Pr > ChiSq |
|------------|----|------------|
| 21.3185    | 7  | 0.0033     |

### Interpretation

- Actually a chi-squared test based on division of x (age) into groups (here, 9 age groups).
- P-value 0.0033 small, so logistic regression not appropriate.
- Maybe age groups 6 and 7 are wrong way around. Assume this (in practice wouldn't, of course)
- Fit same model again and re-do Hosmer-Lemeshow.

### **Output from this analysis**

#### The LOGISTIC Procedure

#### Analysis of Maximum Likelihood Estimates

|           |    |          | Standard | Wald       |            |
|-----------|----|----------|----------|------------|------------|
| Parameter | DF | Estimate | Error    | Chi-Square | Pr > ChiSq |
|           |    |          |          |            |            |
| Intercept | 1  | -2.8107  | 0.1565   | 322.5387   | <.0001     |
| age       | 1  | 0.0476   | 0.00457  | 108.4657   | <.0001     |

Hosmer and Lemeshow Goodness-of-Fit Test

| Chi-Square | DF | Pr | > | ChiSq |
|------------|----|----|---|-------|
|------------|----|----|---|-------|

8.4427 7 0.2952

- No problems with logistic model now.
- Probability of being seropositive definitely increases with age.

# **Predicted probabilities**

| 0bs | age  | n   | r  | pobs    | pred    | lcl     | ucl     |
|-----|------|-----|----|---------|---------|---------|---------|
|     |      |     |    |         |         |         |         |
| 1   | 1.5  | 123 | 8  | 0.06504 | 0.06069 | 0.04588 | 0.07989 |
| 2   | 4.0  | 132 | 6  | 0.04545 | 0.06783 | 0.05227 | 0.08759 |
| 3   | 7.5  | 182 | 18 | 0.09890 | 0.07914 | 0.06258 | 0.09961 |
| 4   | 12.5 | 140 | 14 | 0.10000 | 0.09830 | 0.08042 | 0.11963 |
| 5   | 17.5 | 138 | 20 | 0.14493 | 0.12147 | 0.10230 | 0.14366 |
| б   | 25.0 | 133 | 19 | 0.14286 | 0.16494 | 0.14313 | 0.18934 |
| 7   | 35.0 | 161 | 39 | 0.24224 | 0.24115 | 0.21102 | 0.27409 |
| 8   | 47.0 | 92  | 25 | 0.27174 | 0.35991 | 0.30883 | 0.41437 |
| 9   | 60.0 | 74  | 44 | 0.59459 | 0.51061 | 0.43049 | 0.59018 |

Plenty of data, so CIs are mostly short. Note clear upward trend in probabilities.

### More than 2 response categories

- With 2 response categories, model the probability of one, and prob of other is one minus that. So doesn't matter which category you model.
- With more than 2 categories, have to think more carefully about the categories: are they
  - ordered: you can put them in a natural order (like low, medium, high)
  - nominal: ordering the categories doesn't make sense (like red, green, blue).
- SAS handles both kinds of response; learn how.

### **Ordinal response: the miners**

- Model probability of being in given category or lower.
- Example: coal-miners often suffer disease pneumoconiosis. Likelihood of disease believed to be greater among miners who have worked longer.
- Severity of disease measured on categorical scale: 1 = none, 2 = moderate, 3 = severe.
- Data are frequencies:

Exposure None Moderate Severe

| 5.8  | 98 | 0  | 0  |
|------|----|----|----|
| 15.0 | 51 | 2  | 1  |
| 21.5 | 34 | б  | 3  |
| 27.5 | 35 | 5  | 8  |
| 33.5 | 32 | 10 | 9  |
| 39.5 | 23 | 7  | 8  |
| 46.0 | 12 | б  | 10 |
| 51.5 | 4  | 2  | 5  |

### Data setup

- Set up data file with one frequency on each line, like this: exposure, response category, frequency.
  - 5.8
     98
     15
     1
     51
     51
     2
     2
     15
     3
     1
     21.5
     1
     34
- Don't need to enter zero frequencies.
- Multiple response categories treated as ordered by default.
- Make sure ordering in data is the right one! (I use numbers to keep ordering straight.)

### Code

```
data miners;
infile "miners.dat";
input exposure severity frequency;
```

```
proc logistic;
class severity;
freq frequency;
model severity = exposure;
output out=miners2 pred=pred;
```

```
proc print data=miners2;
Note:
```

class statement turns numbers into ordered response

freq statement ensures frequencies are read as such.

### Output part 1

#### Model Information

| Number | of Observations Read | 22  |
|--------|----------------------|-----|
| Number | of Observations Used | 22  |
| Sum of | Frequencies Read     | 371 |
| Sum of | Frequencies Used     | 371 |

#### Response Profile

| Total     |          | Ordered |
|-----------|----------|---------|
| Frequency | severity | Value   |
| 289       | 1        | 1       |
| 38        | 2        | 2       |
| 44        | 3        | 3       |

Probabilities modeled are cumulated over the lower Ordered Values.

22 lines in data file; frequencies indicate 371 miners total.

Response profile shows number in each severity category in total.

# Output part 2

#### Testing Global Null Hypothesis: BETA=0

| Test             | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 88.2432    | 1  | <.0001     |
| Score            | 80.7246    | 1  | <.0001     |
| Wald             | 64.5206    | 1  | <.0001     |

#### Analysis of Maximum Likelihood Estimates

|           |   |    |          | Standard | Wald       |            |
|-----------|---|----|----------|----------|------------|------------|
| Parameter |   | DF | Estimate | Error    | Chi-Square | Pr > ChiSq |
| Intercept | 1 | 1  | 3.9559   | 0.4096   | 93.2527    | <.0001     |
| Intercept | 2 | 1  | 4.8691   | 0.4437   | 120.4349   | <.0001     |
| exposure  |   | 1  | -0.0959  | 0.0119   | 64.5206    | <.0001     |

Severity of disease definitely depends on exposure. To see how:

# **Predicted severity probs (edited)**

### as they depend on exposure:

| Obs | exposure | severity | frequency | _LEVEL_ | pred    |
|-----|----------|----------|-----------|---------|---------|
| 1   | 5.8      | 1        | 98        | 1       | 0.96769 |
| 2   | 5.8      | 1        | 98        | 2       | 0.98678 |
| 3   | 15.0     | 1        | 51        | 1       | 0.92535 |
| 4   | 15.0     | 1        | 51        | 2       | 0.96865 |
| 9   | 21.5     | 1        | 34        | 1       | 0.86920 |
| 10  | 21.5     | 1        | 34        | 2       | 0.94306 |
| 15  | 27.5     | 1        | 35        | 1       | 0.78893 |
| 16  | 27.5     | 1        | 35        | 2       | 0.90306 |
| 21  | 33.5     | 1        | 32        | 1       | 0.67766 |
| 22  | 33.5     | 1        | 32        | 2       | 0.83974 |
| 27  | 39.5     | 1        | 23        | 1       | 0.54181 |
| 28  | 39.5     | 1        | 23        | 2       | 0.74666 |
| 33  | 46.0     | 1        | 12        | 1       | 0.38799 |
| 34  | 46.0     | 1        | 12        | 2       | 0.61241 |
| 39  | 51.5     | 1        | 4         | 1       | 0.27225 |
| 40  | 51.5     | 1        | 4         | 2       | 0.48251 |

### **Understanding the predicted probs**

- Miner with 5.8 years exposure has prob 0.968 of no disease, and prob 0.987 of moderate disease or lower (and prob 1 of severe disease or lower).
- Subtracting: prob of no disease 0.968, moderate disease 0.987 0.968 = 0.019, severe disease 1 0.987 = 0.013.
- Compare with miner with 51.5 years exposure: prob 0.272 of no disease, prob 0.483 - 0.272 = 0.211 of moderate disease, prob 1 - 0.483 = 0.517 of severe disease.
- Summary:

| Exposure | P(none) | P(moderate) | P(severe) |
|----------|---------|-------------|-----------|
| 5.8      | 0.968   | 0.019       | 0.013     |
| 27.5     | 0.789   | 0.115       | 0.097     |
| 51.5     | 0.272   | 0.211       | 0.517     |

Miner with more exposure has higher prob of having worse disease.

### **Unordered responses**

- With unordered (nominal) responses, can use generalized logit.
- Example: 735 people, record age and sex (male 0, female 1), which of 3 brands of some product preferred.
- Data in mlogit.dat separated by commas.
- Tell SAS that sex and brand numbers only distinguish categories.
- For predictions, get output data set and inspect.

# The code

```
data prefs;
infile "mlogit.dat" delimiter=",";
input brand sex age;
```

```
proc logistic;
class brand;
class sex;
model brand=sex age / link=glogit;
output out=mlogit2 pred=pred;
```

```
proc print data=mlogit2;
```

# Output part 1

#### Model Information

| Response Variable           | brand             |
|-----------------------------|-------------------|
| Number of Response Levels   | 3                 |
| Model                       | generalized logit |
| Number of Observations Used | 735               |

#### Response Profile

| Total     |       | Ordered |
|-----------|-------|---------|
| Frequency | brand | Value   |
| 207       | 1     | 1       |
| 307       | 2     | 2       |
| 221       | 3     | 3       |

Logits modeled use brand=3 as the reference category.

## Output part 2

#### Testing Global Null Hypothesis: BETA=0

| Test             | Chi-Square | DF | Pr > ChiSq |
|------------------|------------|----|------------|
| Likelihood Ratio | 185.8502   | 4  | <.0001     |
| Score            | 163.9538   | 4  | <.0001     |
| Wald             | 129.7966   | 4  | <.0001     |

#### Type 3 Analysis of Effects

|        |    | Wald       |            |
|--------|----|------------|------------|
| Effect | DF | Chi-Square | Pr > ChiSq |
| sex    | 2  | 7.6704     | 0.0216     |
| age    | 2  | 123.3880   | <.0001     |

At least one of sex and age makes a difference to the predicted probs; the bottom table says they both do.

# **Predicted probabilities (a few)**

| Obs  | brand | sex | age | _LEVEL_ | pred    |
|------|-------|-----|-----|---------|---------|
| 4    | 1     | 0   | 26  | 1       | 0.89429 |
| 5    | 1     | 0   | 26  | 2       | 0.09896 |
| 6    | 1     | 0   | 26  | 3       | 0.00674 |
| 10   | 1     | 1   | 27  | 1       | 0.77288 |
| 11   | 1     | 1   | 27  | 2       | 0.20869 |
| 12   | 1     | 1   | 27  | 3       | 0.01843 |
| 2149 | 3     | 0   | 38  | 1       | 0.02598 |
| 2150 | 3     | 0   | 38  | 2       | 0.23855 |
| 2151 | 3     | 0   | 38  | 3       | 0.73547 |
| 2152 | 2     | 1   | 38  | 1       | 0.01623 |
| 2153 | 2     | 1   | 38  | 2       | 0.25162 |
| 2154 | 2     | 1   | 38  | 3       | 0.73215 |

### **Understanding them**

- Many combinations of age, sex and brand-preferred.
- Obs 4, 5 and 6 are for males (sex=0) age 26; prob of preferring brand 1 is 0.894, brand 2 is 0.099, brand 3 is 0.007.
- Summarize whole table from previous page:

| Sex    | Age | P(prefer 1) | P(prefer 2) | P(prefer 3) |
|--------|-----|-------------|-------------|-------------|
| Male   | 26  | 0.894       | 0.099       | 0.007       |
| Female | 27  | 0.773       | 0.209       | 0.018       |
| Male   | 38  | 0.026       | 0.239       | 0.735       |
| Female | 38  | 0.016       | 0.252       | 0.732       |

Younger people prefer brand 1, older prefer brand 3.

Females (a little) less likely to prefer brand 1 and more likely to prefer brand 2. (Sex difference is significant.)

### **Alternative data format**

Summarize all people of same brand preference, same sex, same age on one line of data file with frequency on end:

1024110262102741028410297103003

•••

Whole data set in 65 lines not 735!

### **Code for alternative data format**

```
data prefs;
infile "mlogit2.dat";
input brand sex age frequency;
```

```
proc logistic;
class brand;
class sex;
freq frequency;
model brand=sex age / link=glogit;
output out=mlogit2 pred=pred;
```

Add freq line in analysis. Output same as before.