# Analysis of Categorical Data

SAS Institute

April 2008

# Binomial distribution

$$\begin{aligned} P(y|p) &= \begin{pmatrix} n \\ y \end{pmatrix} p^y (1-p)^{n-y} \\ &= \begin{pmatrix} n \\ y \end{pmatrix} (1-p)^n \left(\frac{p}{1-p}\right)^y \\ &= \begin{pmatrix} n \\ y \end{pmatrix} (1-p)^n \exp\left(y \log\left(\frac{p}{1-p}\right)\right). \end{aligned}$$

Canonical parameter $\log(\frac{p}{1-p})$

# Poisson distribution

$\lambda, 0 < \lambda < \infty$ er

$$
\begin{aligned}
P(y|\lambda) &= \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda} \\
&= \frac{1}{y!}\mathrm{e}^{-\lambda}\mathrm{e}^{y\log(\lambda)},
\end{aligned}
$$

Canonical parameter $\log(\lambda)$.

Generalized linear models (Nelder and Wedderburn (1972), *JRSS A):*

**Random component**:
$Y_1, \cdots, Y_n$ iid from an natural exponential family.
The distribution of each $Y_i$ depends on the parameter $\theta_i$ and has pdf of form:

$$f(y_i; \theta_i) = exp(y_i b(\theta_i) + c(\theta_i) + d(y_i))$$

**Systematic component**:
A set of parameters $\beta$ and explanatory variables $\mathbf{X_1}, \cdots, \mathbf{X_p}$, relating $\eta_i$ - (a linear predictor)

$$\eta_i = \sum_j \beta_j x_{ij}$$

**Link function**:
connects the random and the systematic components

$$g(\mu_i) = \sum_j \beta_j x_{ij},$$

where

$$\mu_i = E(Y_i)$$

-or simply, a GLM is a linear model for the transformed mean
of a response variable with distribution from a natural exponential family.

Contains:
Logistic regression
Logit models for multinomial responses
Poisson regression models - loglinear models
Contingency tables
Negative binomial regression
(and standard normality based linear models, gamma, inverse Gaussian ect.)

SAS: proc GENMOD

| Random Component | Link | Systematic Component | Model |
|---|---|---|---|
| Normal | Identity | Continous | Regression |
| Normal | Identity | Categorical | Analysis of variance |
| Normal | Identity | Mixed | Analysis of Covariance |
| Binomial | Logit | Mixed | Logistic regression |
| Poisson | Log | Mixed | Loglinear |
| Multinomial | Generalized logit | Mixed | Multinomial response |

- The GENMOD procedure fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector . There is, in general, no closed form solution for the maximum likelihood estimates of the parameters. The GENMOD procedure estimates the parameters of the model numerically through an iterative fitting process. Covariances, standard errors, and are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators.

# Logistic regression model

Binary response Y - 0/1

$\pi$=P(Y=1)

Z: explanatory variable

$$logit(\pi) = \alpha + \beta Z,$$

which corresponds to

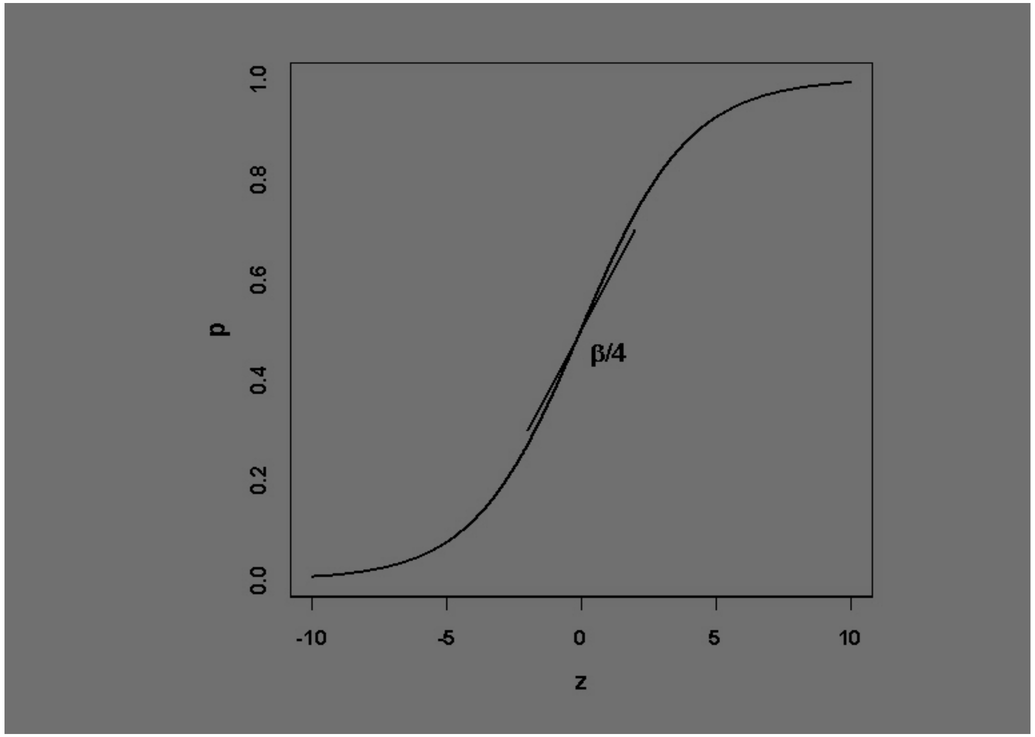
$$\pi = \frac{exp(\alpha + \beta Z)}{1 + exp(\alpha + \beta Z)}.$$

See figure . . .

$$logit(\pi) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_k Z_k,$$

# Interpretation

$$\pi = \frac{exp(\alpha + \beta Z)}{1 + exp(\alpha + \beta Z)}.$$



## Poisson regression

## Model $log(\mu_i) = log(n_i) + X\beta$

```
data insure;
n c car$ age;
ln = log(n);
datalines;
500 42 small 1
1200 37 medium 1
100 1 large 1
400 101 small 2
 500 73 medium 2
 300 14 large 2 ;
 run;
```

```
proc genmod data=insure;
    class car age;
    model c = car age / dist =
    poisson link = log offset = ln;
    run;
```

# Log-linear parameters

$X_{11}, \cdots X_{IJ}$, uafhængige, Poissonfordelte med parametre $\tau_{11}, \cdots, \tau_{IJ}$
Log-linear parameters

$$\log \tau_{ij} = \lambda_{ij}^{XY} + \lambda_i^X + \lambda_j^Y + \lambda_0,$$

or equivalent

$$\tau_{ij} = \exp(\lambda_{ij}^{XY} + \lambda_i^X + \lambda_j^Y + \lambda_0)$$

**Hypothesis:** $\lambda_{ij}^{XY} = 0$

<u>Poisson case</u>

$$\tau_{ij} = \exp(\lambda_i^X + \lambda_j^Y + \lambda_0) = \alpha_i \beta_j \rho$$

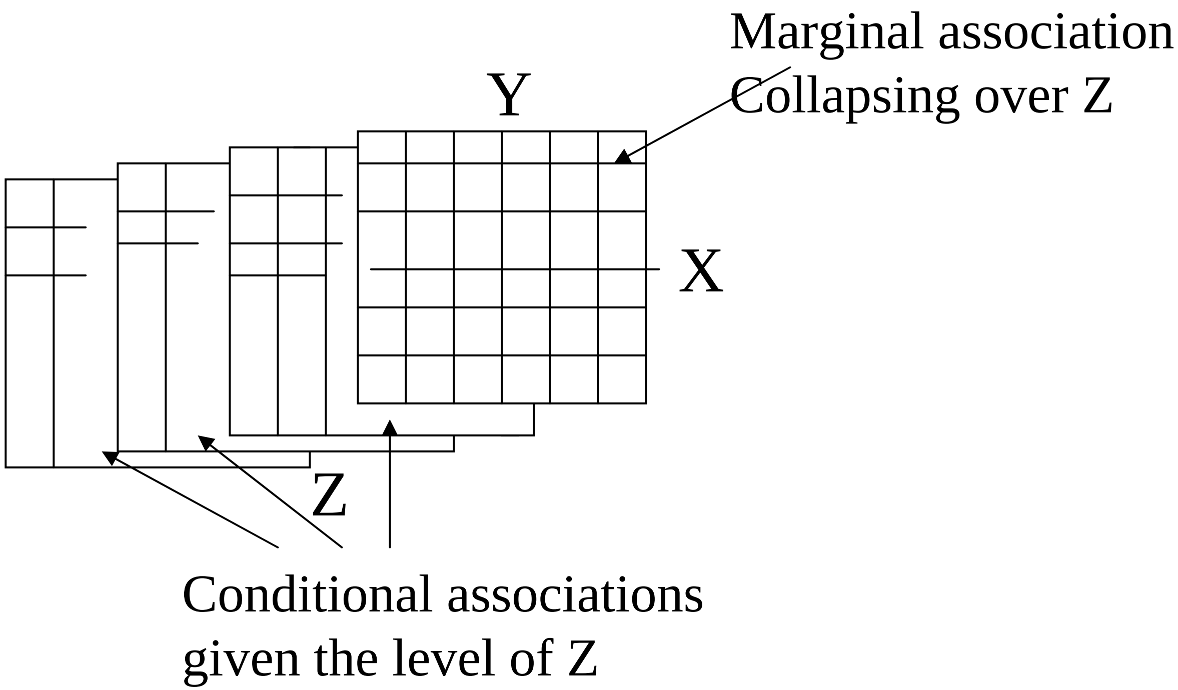
<u>Multinomial case</u>

$$p_{ij} = \frac{\tau_{ij}}{\tau_{..}} = \frac{\exp(\lambda_i^X + \lambda_j^Y)}{\sum_i \sum_j \exp(\lambda_i^X + \lambda_j^Y)} = p_i q_j$$

<u>Stratified sampling</u>

$$p_{ij} = \frac{\tau_{ij}}{\tau_{i.}} = \frac{\exp(\lambda_i^X + \lambda_j^Y)}{\sum_j \exp(\lambda_i^X + \lambda_j^Y)} = \frac{\exp(\lambda_j^Y)}{\sum_j \exp(\lambda_j^Y)} = p_j$$

# Three-way tables




$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$$

# Associations

X , Y ,Z **mutually independent** :

$$\log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

X and Y **conditional independent** given Z if X and Y are independent in each partial table:

$$\log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

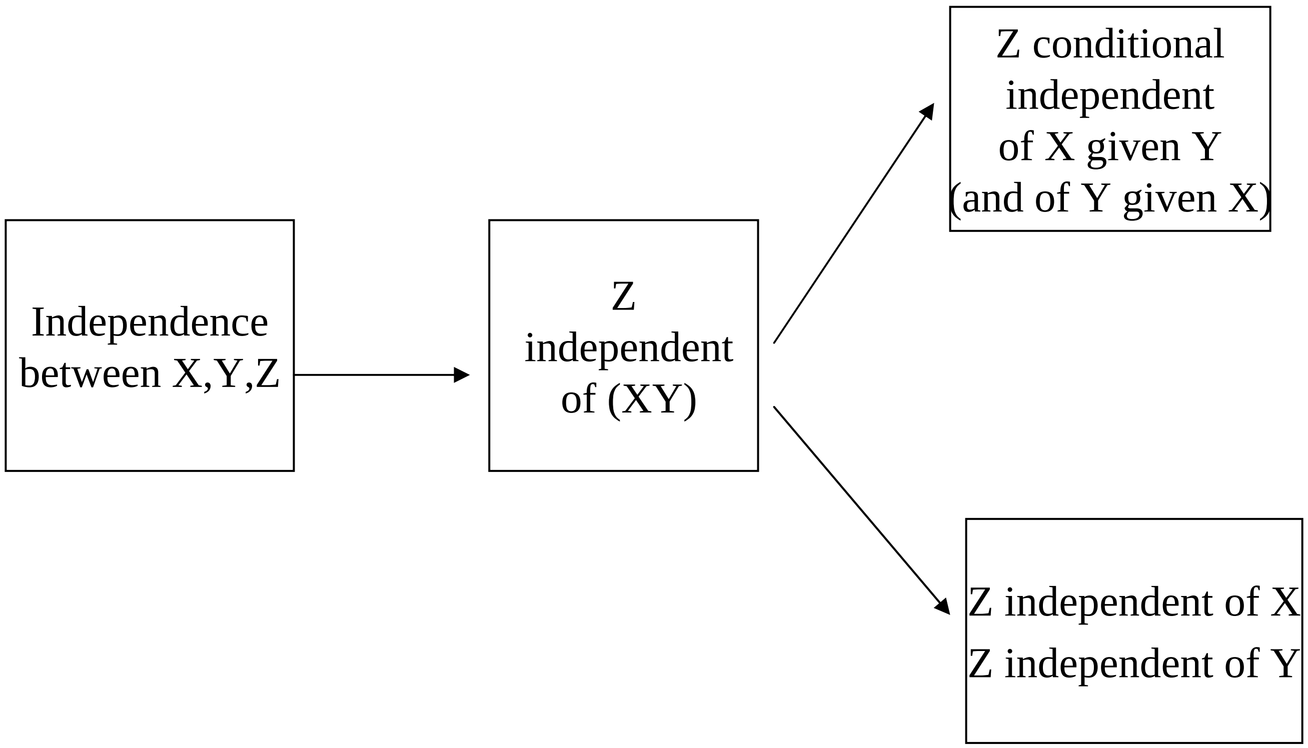
**Homogeneous X-Y association:**

**No interaction** between X and Y in their effect on Z:

$$\log\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$$

- Conditional independence *does not* imply marginal independence.
- Conditional independence is a special case of homogeneous association.
- Homogeneous XY association implies homogeneity for the other associations too (a symmetric property).
- If homogeneous association there is said to be *no interaction* between two variables in their effect on the other variable.

# Associations in three way tables



# Equivalent log-linear and logit models for three-way table with binary response Y

| Loglinear model | Logit model |
|---|---|
| (Y,XZ) | $\beta_0$ |
| (XY,XZ) | $\beta_0 + \beta_i^X$ |
| (YZ,XZ) | $\beta_0 + \beta_k^Z$ |
| (XY,YZ,XZ) | $\beta_0 + \beta_i^X + \beta_k^Z$ |
| (XYZ) | $\beta_0 + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$ |

# Multicategory Logit Models

$Y$ nominal variable with J categories
$p_1, \ldots, p_J$ response probabilities
such that
$P(Y = i) = p_i$ and $\sum_i p_i = 1$

Multicategory logit models simultaneously refer to all pairs of categories.

# Nominal Responses
# Baseline-Category Logits

$Y$ nominal variable with J categories

$p_1, \ldots, p_J$ response probabilities

$$\log(\frac{p_i}{p_J}) = \alpha_i + \beta_i z, i = 1, \ldots, J-1$$

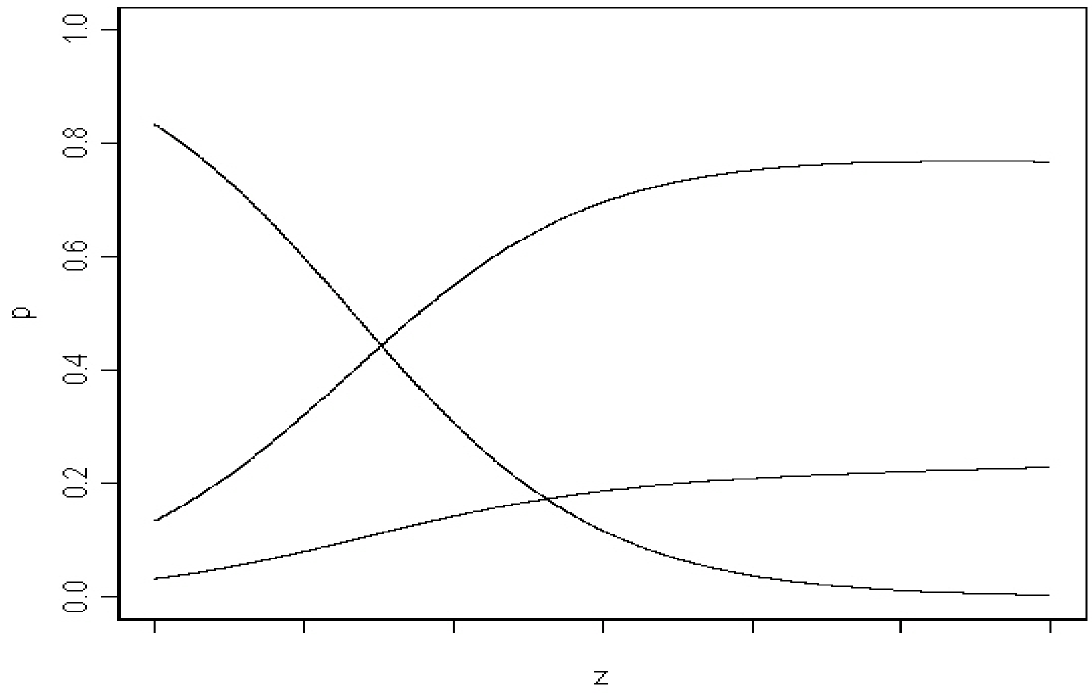
J-1 logit equations, with separate parameters.

If J=2 : ordinary logistic regression model.

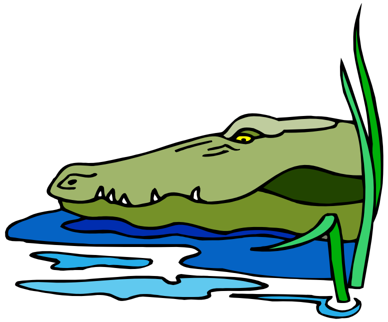
Here the last category (J) is the baseline, but any category can be used

$$\widehat{p}_i = \frac{exp(\widehat{\alpha}_i + \widehat{\beta}_i z)}{1 + \sum_1^{J-1} exp(\widehat{\alpha}_i + \widehat{\beta}_i z)},$$
$$\widehat{p}_J = \frac{1}{1 + \sum_1^{J-1} exp(\widehat{\alpha}_i + \widehat{\beta}_i z)},$$

# Baseline-category logit model



# Baseline Category
# Alligator Food Choice

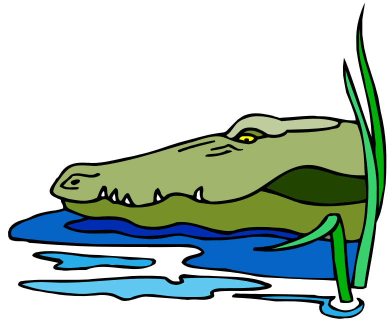


59 alligators, sampled in Florida
Alligator length (L) in meters

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| I 1.24 | I 1.30 | I 1.30 | F 1.32 | F 1.32 | F 1.40 | I 1.42 | F 1.42 |
| I 1.45 | O 1.45 | I 1.47 | F 1.47 | I 1.50 | I 1.52 | I 1.55 | I 1.60 |
| I 1.63 | O 1.65 | I 1.65 | F 1.65 | F 1.65 | F 1.68 | I 1.70 | O 1.73 |
| I 1.78 | I 1.78 | O 1.78 | I 1.80 | F 1.80 | I 1.85 | I 1.88 | I 1.93 |
| I 1.98 | F 2.03 | F 2.03 | F 2.16 | F 2.26 | F 2.31 | F 2.31 | F 2.36 |
| . | . | . | . | . | | | |

F=Fish, I=Invertebrates, O=Other

# Alligator Food Choice



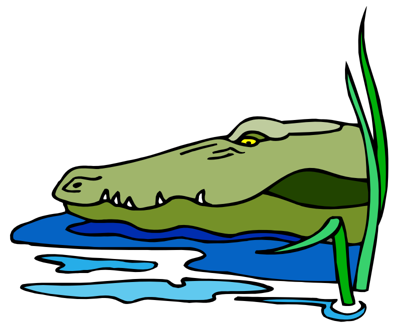
Model:

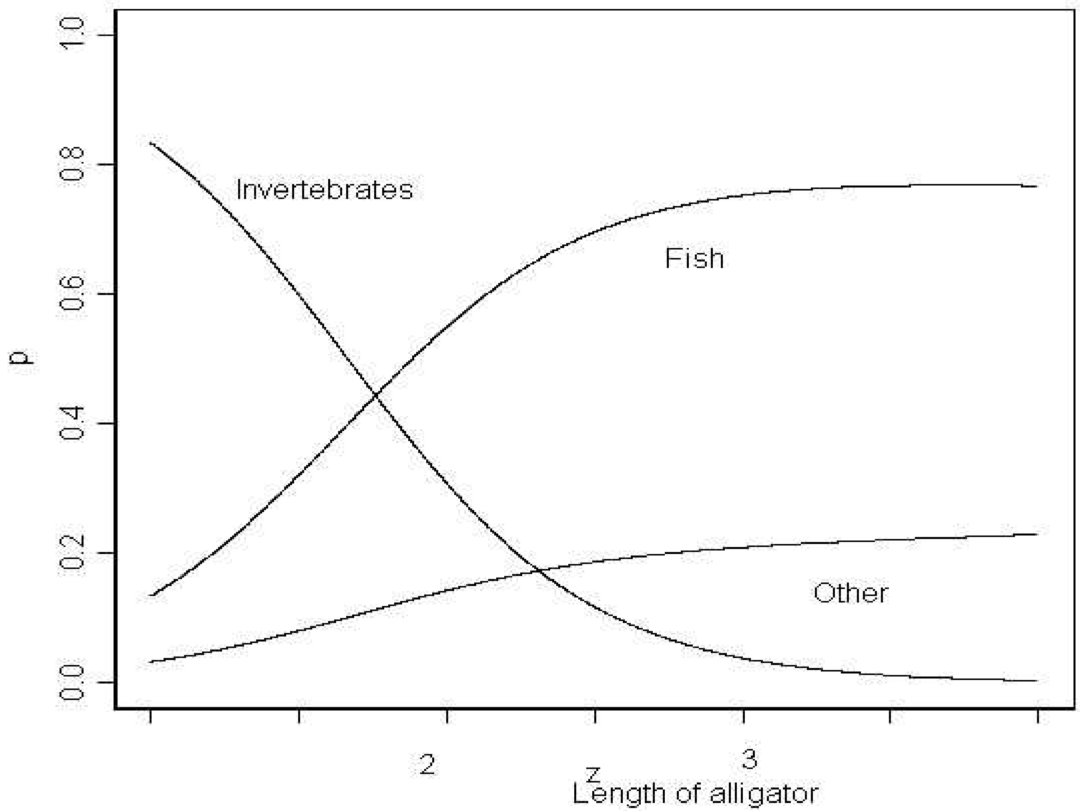
$$log(\frac{p_i}{p_3}) = \alpha_i + \beta_i L$$ , **i=1, 2**

Parameter estimates and Standard Errors (in parentheses):

| Parameter | Fish/Other | Invertebrate/Other |
|---|---|---|
| Intercept | 1.490 | 5.716 |
| Length | -0.070 (.521) | -2.473 (.901) |

# Alligator Food Choice





# Ordinal responses
# Cumulative Logit Models

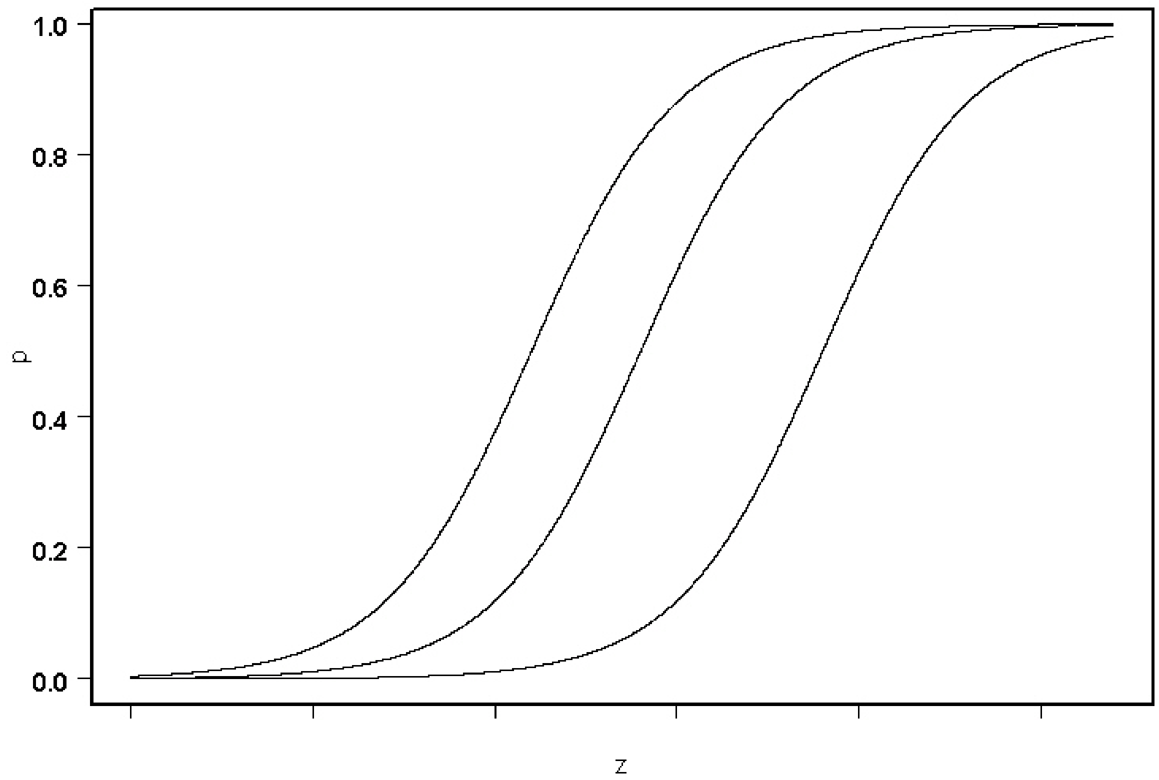
Cumulative logits model:

$$\text{logit}(P(Y \leq i)) = \log(\tfrac{P(Y \leq i)}{1-P(Y \leq i)}) = \alpha_i + \beta z \quad \text{i=1, \ldots,J-1}$$

NOTE: no subscript on $\beta$, so identical effect of $x$ for all $J-1$ collapsings of the response into binary outcomes.

Sometimes called a *proportional odds model*:
McCullagh (1980)
$\text{logit}(P(Y \leq i|x_1)) - \text{logit}(P(Y \leq i|x_2)) = \beta(x_i - x_2)$

# Ordinal response categories



**Collapsibility** property:

Effect parameters are invariant to the choice of categories for Y - $\alpha_i$ will be affected.

**Appropriateness** can be tested:

Score test is provided by SAS –compares the model with a more complex model with varying parameters for the effects

If poor fit try baseline-category logit model or a nonsymmetric link (eg log-log)

**Complementary log-log link:**

$\log(\text{-}\log[1 - P(Y \leq i|x)]) = \alpha_i + \beta x$
$P(Y > i|x_1) = (P(Y > i|x_2))^{exp(\beta(x_1 - x_2))}$
Proportional hazards model for survival data for grouped survival times.

| Life length | Males | | Females | |
|---|---|---|---|---|
| | White | Black | White | Black |
| 0-20 | 2.4 | 3.6 | 1.6 | 2.7 |
| 20-40 | 3.4 | 7.5 | 1.4 | 2.9 |
| 40-50 | 3.8 | 8.3 | 2.2 | 4.4 |
| 50-60 | 17.5 | 25.0 | 69.9 | 16.3 |
| Over60 | 72.9 | 55.6 | 84.9 | 73.7 |

$P(Y > i|G = male, R = r) = (P(Y > i|G = female, R = r))^{1.93}$

**Diagnostics.**

Deviance $D(y;\hat{\mu}) = -2(l(\hat{\mu};y) - l(y;y))$
or $\chi^2$ goodness-of-fit not too large compared to N-p

**Overdispersion** D/(N-p) is larger than expected (=1) indicates inadequate model - eg wrong link or missing explanatory variables

More complex model: include an extra parameter

## Dispersion models

$$f(y) = f(y) = \exp\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\}$$

$\phi$ is called the dispersion parameter. ($\phi$ known just exp.family)
$\theta$ the natural parameter.
**Poisson:**

$$f(y) = \frac{\mu^y e^{-\mu}}{y!} \; y = 0, 1, 2, ... \quad \phi = 1 \;\; , Var(Y) = \mu$$

**Binomial:**

$$f(y) = \binom{n}{r}\mu^r(1-\mu)^{n-r} \; y = \frac{r}{n}, r = 0, 1, 2, ..., n \; \phi = 1 \; , Var(Y) = \frac{\mu(1-\mu)}{n}$$

$\mu = E(Y)$
$Var(Y) = \frac{V(\mu)\phi}{w}$

Overdispersion $\phi > 1$
Poisson: $Var(\mu) = \phi\mu$
Binomial: $Var(\mu) = \phi\mu(1 - \mu)$

Estimates for parameters $\beta$'s are the same as ML estimates in the poisson resp. binomial model - but inflates their standard error.

The estimated $cov(\hat{\beta})$ is $\phi$ times that for the standard model.

Estimate $\phi$ by Pearson/N-p or D/N-p.
That is multiply the std errors by $\sqrt{\phi}$

```
•	data a;
•	input yes n temp moisture co2 freshair dust ventil;
•	cards;
•	18 19 22.0 30 0.09 8.5 0.20 230
•	16 20 21.5 25 0.11 6.1 0.08 230
•	 4 19 21.5 25 0.11 4.8 0.06 230
•	13 18 18.5 25 0.09 9.2 0.07 236
•	12 14 20.0 25 0.05 8.7 0.08 236
•	 4 18 20.0 25 0.11 5.2 0.12 236
•	14 17 20.5 30 0.08 13.1 0.09 249
•	18 19 21.0 30 0.08 12.5 0.06 249
•	 9 16 21.5 30 0.09 8.7 0.07 215
•	 8 18 21.0 30 0.07 9.3 0.09 215
•	;
•	data b;
•	set a;
•
•	proc genmod;
•	make 'obstats' out=pred;
•	model yes/n=freshair temp moisture dust ventil co2/
•	        dist=bin link=logit obstats type1;
•
•	title1 "Eksempel fra E.B. Andersen (1991):";
•	title2 " The Statistical Analysis of Categorical Data. Springer-Verlag.";
```

- Criteria For Assessing Goodness Of Fit

| Criterion | DF | Value | Value/DF |
|---|---|---|---|
| Deviance | 3 | 6.4741 | 2.1580 |
| Scaled Deviance | 3 | 6.4741 | 2.1580 |
| Pearson Chi-Square | 3 | 7.6520 | 2.5507 |
| Scaled Pearson X2 | 3 | 7.6520 | 2.5507 |
| Log Likelihood | | -88.0226 | |

- Algorithm converged.

- Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 4.6563 | 9.0648 | -13.1103 | 22.4230 | 0.26 | 0.6075 |
| freshair | 1 | 1.4486 | 0.3269 | 0.8080 | 2.0893 | 19.64 | <.0001 |
| temp | 1 | 1.3204 | 0.3420 | 0.6501 | 1.9908 | 14.90 | 0.0001 |
| moisture | 1 | -1.1412 | 0.2941 | -1.7175 | -0.5648 | 15.06 | 0.0001 |
| dust | 1 | 25.3048 | 7.5654 | 10.4769 | 40.1326 | 11.19 | 0.0008 |
| ventil | 1 | -0.0705 | 0.0363 | -0.1417 | 0.0006 | 3.78 | 0.0519 |
| co2 | 1 | 20.2955 | 17.7058 | -14.4073 | 54.9983 | 1.31 | 0.2517 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

- NOTE: The scale parameter was held fixed.

| | | | |
|---|---|---|---|
| Deviance | 3 | 6.4741 | 2.1580 |
| Scaled Deviance | 3 | 3.0000 | 1.0000 |
| Pearson Chi-Square | 3 | 7.6520 | 2.5507 |
| Scaled Pearson X2 | 3 | 3.5458 | 1.1819 |
| Log Likelihood | | -40.788 | |

Analysis Of Parameter Estimates

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 4.6563 | 13.3164 | -21.4433 | 30.7560 | 0.12 | 0.7266 |
| freshair | 1 | 1.4486 | 0.4802 | 0.5075 | 2.3898 | 9.10 | 0.0026 |
| temp | 1 | 1.3204 | 0.5024 | 0.3357 | 2.3052 | 6.91 | 0.0086 |
| moisture | 1 | -1.1412 | 0.4320 | -1.9878 | -0.2945 | 6.98 | 0.0082 |
| dust | 1 | 25.3048 | 11.1138 | 3.5222 | 47.0873 | 5.18 | 0.0228 |
| ventil | 1 | -0.0705 | 0.0533 | -0.1750 | 0.0340 | 1.75 | 0.1858 |
| co2 | 1 | 20.2955 | 26.0104 | -30.6839 | 71.2749 | 0.61 | 0.4352 |
| Scale | 0 | 1.4690 | 0.0000 | 1.4690 | 1.4690 | | |

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

- The dispersion parameter is also estimated by maximum likelihood or, optionally, by the residual deviance or by Pearson's chi-square divided by the degrees of freedom

- Models for matched pairs
- Repeated Categorical response data
- Random effects GLMM
- Graphical models (directed)

Reference:

Agresti A, 2002, **Categorical Data Analysis**, Wiley (homepage with SAS programs)

# Estimation

$$
\begin{aligned}
\mathrm{l}(\underline{\tau}|\underline{x}) = \quad & \textstyle\sum_i \sum_j x_{ij}\tau_{ij}^{AB} + \sum_i x_{i.}\tau_i^A + \sum_j x_{.j}\tau_j^B + x_{..}\tau_0 \\
& - \textstyle\sum_i \sum_j \log x_{ij}! - \sum_i \sum_j \exp(\tau_{ij}^{AB} + \tau_i^A + \tau_j^B + \tau_0),
\end{aligned}
$$

$x_{ij}$ er sufficient for $\tau_{ij}^{AB}$, $x_{i.}$ er sufficient for $\tau_i^A$, $x_{.j}$ er sufficient for $\tau_{jB}$ og $x_{..}$ er sufficient for $\tau_0$

$$
\begin{aligned}
x_{ij} &= \mathrm{E}(X_{ij}), \quad i = 1, \cdots, I-1, j = 1, \cdots, J-1, \\
x_{i.} &= \mathrm{E}(X_{i.}), \quad i = 1, \cdots, I-1, \\
x_{.j} &= \mathrm{E}(X_{.j}), \quad j = 1, \cdots, J-1, \\
x_{..} &= \mathrm{E}(X_{..}).
\end{aligned}
$$

# Multiple Logistic regression With dummies at two levels

**Two binary predictors, X and Z.**

**For the 2×2×2 contingency table the model for $\pi$=P(Y=1) is**

$$logit(\pi) = \alpha + \beta_1 x + \beta_2 z,$$

**Denote the levels for each predictor variable by (0,1).**
**For each level of Z the conditional OR is**

$$OR(x = 1/x = 0) = exp(\beta_1)$$

**The same as homogeneous association.**

**Conditional independence between X and Y given Z if**

$$logit(\pi) = \alpha + \beta_2 z,$$

# Multiple Logistic regression
# With dummies at two levels

**Two binary predictors, X and Z.**

**For the 2×2×2 contingency table the model for $\pi$=P(Y=1) is**

$$logit(\pi) = \alpha + \beta_1 x + \beta_2 z,$$

**Denote the levels for each predictor variable by (0,1).**
**At fixed levels for Z the effect on logit changing x from 0 to 1 is**

$$= (\alpha + \beta_1 \cdot 1 + \beta_2 z) - (\alpha + \beta_1 \cdot 0 + \beta_2 z) = \beta_1$$

**so,** $log(odds_{x=1}) - log(odds_{x=0}) = \beta_1$
**and therefore**

$$OR(x = 1/x = 0) = exp(\beta_1)$$

**If levels of x are (-1,1)**

$$OR(x = 1/x = -1) = exp(2\beta_1)$$

# Estimation

Disse ligninger har samme løsning som

$$x_{ij} = \mathrm{E}(X_{ij}), i = 1, \cdots, I, j = 1, \cdots, J$$

Da $\mathrm{E}(X_{ij}) = \lambda_{ij}$, er $\hat{\lambda}_{ij} = x_{ij}$

$x_{ij}$, $x_{i.}$, $x_{.j}$ og $x_{..}$ kaldes for de **sufficiente marginaler**

# Estimation under $\mathrm{H}_0$: $\tau_{ij}^{AB} = 0$

$$\mathrm{l}(\underline{\tau}|\underline{x}) = \sum_i x_{i.}\tau_i^A + \sum_j x_{.j}\tau_j^B + x_{..}\tau_0 - \sum_i \sum_j \log x_{ij}! \\ - \sum_i \sum_j \exp(\tau_i^A + \tau_j^B + \tau_0),$$

De sufficiente marginaler er $x_{i.}$, $x_{.j}$ og $x_{..}$, og likelihoodligningerne er

$$\begin{aligned} x_{i.} &= \mathrm{E}(X_{i.}), \quad && i = 1, \cdots, I-1, \\ x_{.j} &= \mathrm{E}(X_{.j}), \quad && j = 1, \cdots, J-1, \\ x_{..} &= \mathrm{E}(X_{..}). \end{aligned}$$

```
data a;
input year collisions miles @@;
lmiles=log(miles);
cards;
1970    3     281      1977    4     264
1971    6     276      1978    1     267
1972    4     268      1979    7     265
1973    7     269      1980    3     267
1974    6     281      1981    5     260
1975    2     271      1982    6     231
1976    2     265      1983    1     249
;
run;
proc genmod;
model collisions =  /dist=poisson offset = lmiles;
run;
proc genmod;
model collisions = year /dist=poisson offset = lmiles;
run;
proc print;
```

| | | | |
|---|---|---|---|
| **Deviance** | 12 | 15.8992 | 1.3249 |

**Analysis Of Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 22.0562 | 65.4260 | -106.176 | 150.2887 | 0.11 | 0.7360 |
| year | 1 | -0.0133 | 0.0331 | -0.0782 | 0.0516 | 0.16 | 0.6885 |
| Scale | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

| | | | |
|---|---|---|---|
| **Deviance** | 13 | 16.0602 | 1.2354 |

**Analysis Of Parameter Estimates**

| Parameter | DF | Estimate | Standard Error | Wald 95% Confidence Limits | | Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|---|---|
| **Intercept** | 1 | -4.1768 | 0.1325 | -4.4364 | -3.9172 | 994.41 | <.0001 |
| **Scale** | 0 | 1.0000 | 0.0000 | 1.0000 | 1.0000 | | |

$$exp(-4.1768) = 0.01534$$

$$exp(-4.1768 \pm 1.96 * 0.1325) = (0.0118, 0.0199)$$