

10.4.1 Testing for symmetry in a square $I \times I$ table

Consider an $I \times I$ table which cross-classifies (X, Y) on the same outcomes.

	$Y = 1$	$Y = 2$	\dots	$Y = I$
$X = 1$	π_{11}	π_{12}	\dots	π_{1I}
$X = 2$	π_{21}	π_{22}	\dots	π_{2I}
\vdots	\vdots	\vdots	\ddots	\vdots
$X = I$	π_{I1}	π_{I2}	\dots	π_{II}

Marginal homogeneity happens when $P(X = i) = P(Y = i)$ ($\pi_{+i} = \pi_{+i}$) for $i = 1, \dots, I$. This is important, for example, when determining if classifiers (like X-ray readers) tend to classify in roughly the same proportions. If not, perhaps one reader tends to diagnose a disease more often than another reader.

Symmetry, a stronger assumption, implies marginal homogeneity.

An $I \times I$ table is *symmetric* if $P(X = i, Y = j) = P(X = j, Y = i)$ ($\pi_{ij} = \pi_{ji}$).

This simply reduces the number of parameters from I^2 (subject to summing to one) to $I(I + 1)/2$ (subject to summing to one). For example, in a 3×3 table this forces

	$Y = 1$	$Y = 2$	$Y = 3$
$X = 1$	π_1	π_2	π_3
$X = 2$	π_2	π_4	π_5
$X = 3$	π_3	π_5	π_6

subject to $\pi_1 + \pi_4 + \pi_6 + 2\pi_2 + 2\pi_3 + 2\pi_5 = 1$.

The symmetric model is easily fit by specifying the cell probabilities by hand in GENMOD. A test of the symmetric model versus the saturated model is a test of $H_0 : \pi_{ij} = \pi_{ji}$ and can be carried out by looking at the Deviance statistic (yielding a LRT).

The following table is from Yule (1900)

Husband	Wife		
	Tall	Medium	Short
Tall	18	28	14
Medium	20	51	28
Short	12	25	9

Let (X, Y) be the heights of the (Husband, Wife). The table is symmetric if $P(X = i, Y = j) = P(X = j, Y = i)$. For example, symmetry forces the same proportion of pairings of (Husband, Wife)=(Tall, Short) and (Husband, Wife)=(Short, Tall). This assumes the following structure

Husband	Wife		
	Tall	Medium	Short
Tall	π_1	π_2	π_3
Medium	π_2	π_4	π_5
Short	π_3	π_5	π_6

subject to $\pi_1 + 2\pi_2 + 2\pi_3 + \pi_4 + 2\pi_5 + \pi_6 = 1$.

```

data hw;
  input h w symm count @@;
  datalines;
1 1 1 18  1 2 2 28  1 3 3 14
2 1 2 20  2 2 4 51  2 3 5 28
3 1 3 12  3 2 5 25  3 3 6  9
;
proc genmod; class symm;
  model count=symm / link=log dist=poi;

```

The GENMOD output gives us

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	3	1.6635	0.5545

A test of symmetry versus the saturated model gives a p -value of $P(\chi_3^2 > 1.66) = 0.65$. We accept that the symmetric model fits.

Symmetry implies marginal homogeneity, $P(X = i) = P(Y = i)$.

Husbands and wives are tall, medium, or short in the same proportions.

Furthermore, for example, short wives and tall husbands occur with the same probability as tall wives with short husbands.

10.5.4 Kappa measure of agreement

Consider an $I \times I$ table where X and Y are cross-classified on the same scale. Below are $n = 118$ slides classified for carcinoma of the uterine cervix by two pathologists as (1) negative, (2) atypical squamous hyperplasia, (3) carcinoma *in situ*, or (4) squamous or invasive carcinoma.

Pathologist A	Pathologist B				Total
	1	2	3	4	
1	22	2	2	0	26
2	5	7	14	0	26
3	0	2	36	0	38
4	0	1	17	10	28
Total	27	12	69	10	118

If A and B were the *same person* then $\pi_{ij} = 0$ when $i \neq j$, i.e. there'd only be nonzero diagonal elements. Nonzero off-diagonal elements reflect disagreement and the further off the diagonal they are, the more severe the disagreement.

For example there are two slides classified by B as carcinoma *in situ* (not metastasized beyond the original site) that A classified as negative.

Perfect agreement occurs when $\pi_{11} + \pi_{22} + \pi_{33} + \pi_{44} = 1$. The strength of agreement has to do with how close this is to one.

Marginal homogeneity occurs when the two classifiers agree on the proportion of each classification in the population, but not necessarily the classifications themselves. If marginal homogeneity is not satisfied, then one classifier tends to classify a fixed category more often than the other.

Classifiers are independent if $P(X = i, Y = j) = P(X = i)P(Y = j)$, and in this case agreement for category i happens with probability $P(X = i, Y = i) = P(X = i)P(Y = i) = \pi_{i+}\pi_{+i}$. The kappa statistic looks at the difference between the probability of agreement $\sum_{i=1}^I \pi_{ii}$ and agreement due to “chance” $\sum_{i=1}^I \pi_{i+}\pi_{+i}$, normalized by the largest this can be when $\sum_{i=1}^I \pi_{ii} = 1$:

$$\kappa = \frac{\sum_{i=1}^I \pi_{ii} - \pi_{i+}\pi_{+i}}{1 - \sum_{i=1}^I \pi_{i+}\pi_{+i}},$$

and is estimated by simply replacing π_{ij} by $\hat{\pi}_{ij} = n_{ij}/n_{++}$.

```

data table;
  input A B count @@;
  datalines;
1 1 22 1 2 2 1 3 2 1 4 0
2 1 5 2 2 7 2 3 14 2 4 0
3 1 0 3 2 2 3 3 36 3 4 0
4 1 0 4 2 1 4 3 17 4 4 10
;
proc freq order=data; weight count; tables A*B / plcorr agree;

```

The FREQ Procedure

Statistic	Value	ASE
Gamma	0.9332	0.0340
Polychoric Correlation	0.9029	0.0307

Test of Symmetry

Statistic (S)	30.2857
DF	6
Pr > S	<.0001

Kappa Statistics

Statistic	Value	ASE	95% Confidence Limits	
Simple Kappa	0.4930	0.0567	0.3818	0.6042
Weighted Kappa	0.6488	0.0477	0.5554	0.7422

Sample Size = 118

- There's a test for symmetry! The statistic is the same as the Pearson GOF test for the symmetric log-linear model, i.e. a score test for testing $H_0 : \pi_{ij} = \pi_{ji}$. What do we conclude?
- How about $\hat{\gamma} = 0.93$ and $\hat{\rho} = 0.90$, both highly significant? What does that tell us?
- Finally, $\hat{\kappa} = 0.49$ with 95% CI about (0.4, 0.6). The difference between observed agreement and that expected purely by chance is between 0.4 and 0.6, moderately strong agreement.
- The weighted kappa statistic is valid for an ordinal response and weights differences in classifications according to how "severe" the discrepancy. See p. 435.
- κ is *one number* summarizing agreement. It may be much more interesting to quantify *where* or *why* disagreement occurs via models.

Test of marginal homogeneity

Recall that McNemar's test tests $H_0 : P(X = 1) = P(Y = 1)$ for a 2×2 table. This is output from PROC FREQ in SAS using AGREE.

Often, when comparing raters, we have more than 2 categories. A general test of marginal homogeneity tests $H_0 : P(X = i) = P(Y = i)$ for $i = 1, \dots, I$. mh is a small program written by John Uebersax to perform overall tests of marginal homogeneity, among other things.

```
MH Program: Marginal Homogeneity Tests for N x N Tables
Version 1.2 - John Uebersax
2008-04-24  2:19 PM
```

```
***INPUT***
```

```
Diagnoses of Carcinoma (Agresi Table 10.8)
```

```
4 categories
```

```
Path A is row variable
```

```
Path B is column variable
```

```
ordered categories
```

22	2	2	0
5	7	14	0
0	2	36	0
0	1	17	10

```
Total number of cases:      118
```

BASIC TESTS

Four-fold tables tested

22	4	5	87
7	19	5	87
36	2	33	47
10	18	0	90

McNemar Tests for Each Category

Level (k)	Frequency		Proportion (Base Rate)		Chi- squared(a)	p
	Path A	Path B	Path A	Path B		
1	26	27	0.220	0.229	exact test	1.0000
2	26	12	0.220	0.102	8.167	0.0043*
3	38	69	0.322	0.585	27.457	0.0000*
4	28	10	0.237	0.085	18.000	0.0000*

(a) or exact test

* p < Bonferroni-adjusted significance criterion of 0.017.

Tests of Overall Marginal Homogeneity

Bhapkar chi-squared = 38.528 df = 3 p = 0.0000
 Stuart-Maxwell chi-squared = 29.045 df = 3 p = 0.0000

Bowker Symmetry Test

Chi-squared = 30.286 df = 6 p = 0.0000

TESTS FOR ORDERED-CATEGORY DATA

McNemar Test of Overall Bias
or Direction of Change

Cases where Path A level is higher: 25
Cases where Path B level is higher: 18

Chi-squared = 1.140 df = 1 p = 0.2858

Four-fold tables tested (for thresholds tests)

22	4	5	87
36	16	3	63
90	0	18	10

Tests of Equal Category Thresholds

Level (k)	Proportion of cases below level k		Threshold(a)		Chi- squared(b)	p
	Path A	Path B	Path A	Path B		
2	0.220	0.229	-0.771	-0.743	exact test	1.0000
3	0.441	0.331	-0.149	-0.439	8.895	0.0029*
4	0.763	0.915	0.715	1.374	18.000	0.0000*

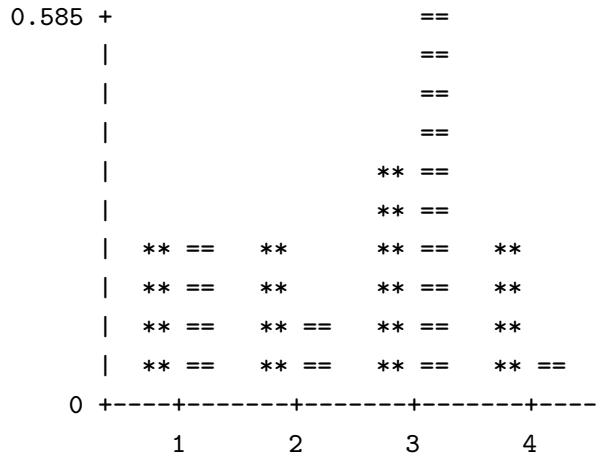
(a) for probit model

(b) or exact test

* p < Bonferroni-adjusted significance criterion of 0.017.

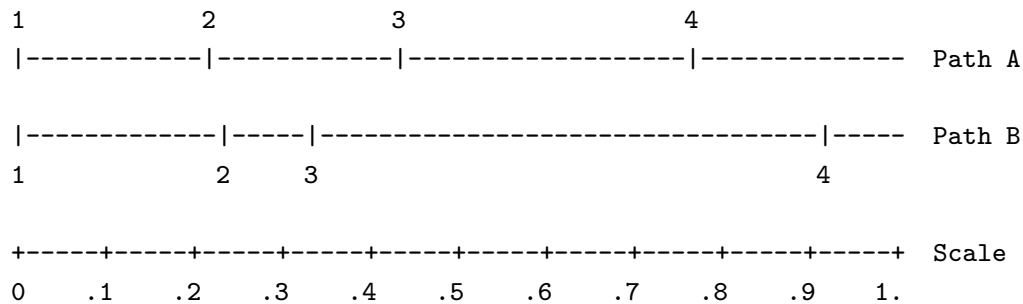
GRAPHIC OUTPUT

Marginal Distributions of Categories
for Path A (**) and Path B (==)



Notes: x-axis is category number or level.
y-axis is proportion of cases.

Proportion of cases below each level



Comments...

- The Bhapkar test (p. 422, 10.3.3; more powerful than Stuart-Maxwell) for marginal homogeneity is highly significant with $p = 0.0000$. We reject marginal homogeneity. The graphical output indicates that both pathologists tend to classify ‘negative’ in roughly the same proportion, but that B classifies ‘carcinoma *in situ*’ more often than A , whereas A classifies ‘atypical squamous hyperplasia’ and ‘squamous or invasive carcinoma’ more often than B .
- There is also an individual test for each category.
 $H_0 : P(X = i) = P(Y = i)$ is rejected for $i = 2, 3, 4$ but not $i = 1$.

- We are interested in whether one rater tends to classify slides ‘higher’ or ‘lower’ than the other. Off-diagonal elements above the diagonal are when B classifies higher than A ; elements below the diagonal are when B classifies lower than A . The McNemar test of overall bias is not significant, indicating that one rater does not tend to rate higher or lower than the other.
- The test for symmetry has the same test statistic and p -value as from SAS.
- The program is easy to run on a Windows-based PC and free. There is a users guide and sample input and output files. Web location: <http://ourworld.compuserve.com/homepages/jsuebersax/mh.htm>.

4.8 Generalized additive models

Consider a linear regression problem:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i,$$

where $e_1, \dots, e_n \stackrel{iid}{\sim} N(0, \sigma^2)$.

Diagnostics (residual plots, added variable plots) might indicate poor fit of the basic model above. Remedial measures might include transforming the response, transforming one or both predictors, or both. One also might consider adding quadratic terms and/or an interaction term.

Note: we only consider transforming *continuous* predictors!

When considering a transformation of one predictor, an added variable plot can suggest a transformation (e.g. $\log(x)$, $1/x$) that might work *if the other predictor is “correctly” specified*.

In general, a transformation is given by a function $g(x)$. Say we decide that x_{i1} should be log-transformed and the reciprocal of x_{i2} should be used. Then the resulting model is

$$Y_i = \beta_0 + \beta_1 \log(x_{i1}) + \beta_2/x_{i2} + e_i = \beta_0 + g_{\beta_1}(x_{i1}) + g_{\beta_2}(x_{i2}) + e_i,$$

where $g_{\beta_1}(x)$ and $g_{\beta_2}(x)$ are two functions specified by β_1 and β_2 .

Here we are specifying forms for $g_{\beta_1}(x)$ and $g_{\beta_2}(x)$ based on exploratory data analysis, but we could from the outset specify *models* for $g_{\theta_1}(x)$ and $g_{\theta_2}(x)$ that are rich enough to capture interesting and predictively useful aspects of how the predictors affect the response and *estimate these functions from the data*.

This is an example of “nonparametric regression,” which ironically connotes the inclusion of *lots* of parameters rather than fewer.

This approach has gained more favor from Bayesians, but is not the approach taken in SAS PROC GAM. PROC GAM makes use of *cubic smoothing splines*.

For simple regression data $\{(x_i, y_i)\}_{i=1}^n$, a cubic spline smoother $g(x)$ minimizes

$$\sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int_{-\infty}^{\infty} g''(x)^2 dx.$$

Good fit is achieved by minimizing the sum of squares $\sum_{i=1}^n (y_i - g(x_i))^2$. The $\int_{-\infty}^{\infty} g''(x)^2 dx$ term measures how wiggly $g(x)$ is and $\lambda \geq 0$ is how much we will penalize $g(x)$ for being wiggly.

So the spline trades off between goodness of fit and wiggleness.

Although not obvious, the solution to this minimization is a cubic spline: a piecewise cubic polynomial with the pieces joined at the unique x_i values.

Hastie and Tibshirani (1986, 1990) point out that the meaning of λ depends on the units x_i is measured in, but that λ can be picked to yield an “effective degrees of freedom” df or an “effective number of parameters” being used in $g(x)$. Then the complexity of $g(x)$ is equivalent to $(df - 1)$ -degree polynomial, but with the coefficients “spread out” more yielding a more flexible function that fits data better.

Alternatively, λ can be picked through cross validation, by minimizing

$$CV(\lambda) = \sum_{i=1}^n (y_i - g_{\lambda}^{-i}(x_i))^2.$$

Both options are available in SAS.

We don't have $\{(x_i, y_i)\}_{i=1}^n$ where y_1, \dots, y_n are continuous, but rather $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where y_i is categorical (e.g. Bernoulli) or Poisson. The generalized additive model (GAM) is given by

$$h\{E(Y_i)\} = \beta_0 + g_1(x_{i1}) + \dots + g_{ip}(x_{ip}),$$

for p predictor variables. Y_i is a member of an exponential family such as binomial, Poisson, normal, etc. h is a link function.

Each of $g_1(x), \dots, g_p(x)$ are modeled via cubic smoothing splines, each with their own smoothness parameters $\lambda_1, \dots, \lambda_p$ either specified as df_1, \dots, df_p or estimated through cross-validation. The model is fit through "backfitting." See Hastie and Tibshirani (1990) or the SAS documentation for details.

Let's fit a GAM to the O-ring space shuttle data:

```
data shut1;
  input temp td @@;
  datalines;
66 0 70 1 69 0 68 0 67 0 72 0 73 0 70 0 57 1 63 1 70 1 78 0 67 0
53 1 67 0 75 0 70 0 81 0 76 0 79 0 75 1 76 0 58 1
;
ods html; ods graphics on;
proc gam plots(clm) data=shut1;
  model td = spline(temp) / dist=binomial;
run; quit; ods graphics off; ods html close;
```

Output:

```
                The GAM Procedure
                Dependent Variable: td
Smoothing Model Component(s): spline(temp)
```

Summary of Input Data Set

Number of Observations	23
Number of Missing Observations	0
Distribution	Binomial
Link Function	Logit

Iteration Summary and Fit Statistics

Number of local score iterations	15
Local score convergence criterion	5.925073E-10
Final Number of Backfitting Iterations	1
Final Backfitting Criterion	8.5164609E-9
The Deviance of the Final Estimate	12.445020758

The local score algorithm converged.

Regression Model Analysis

Parameter Estimates

Parameter	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	5.18721	14.01486	0.37	0.7156
Linear(temp)	-0.08921	0.19693	-0.45	0.6560

Smoothing Model Analysis

Fit Summary for Smoothing Components

Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline(temp)	0.999976	3.000000	136344	16

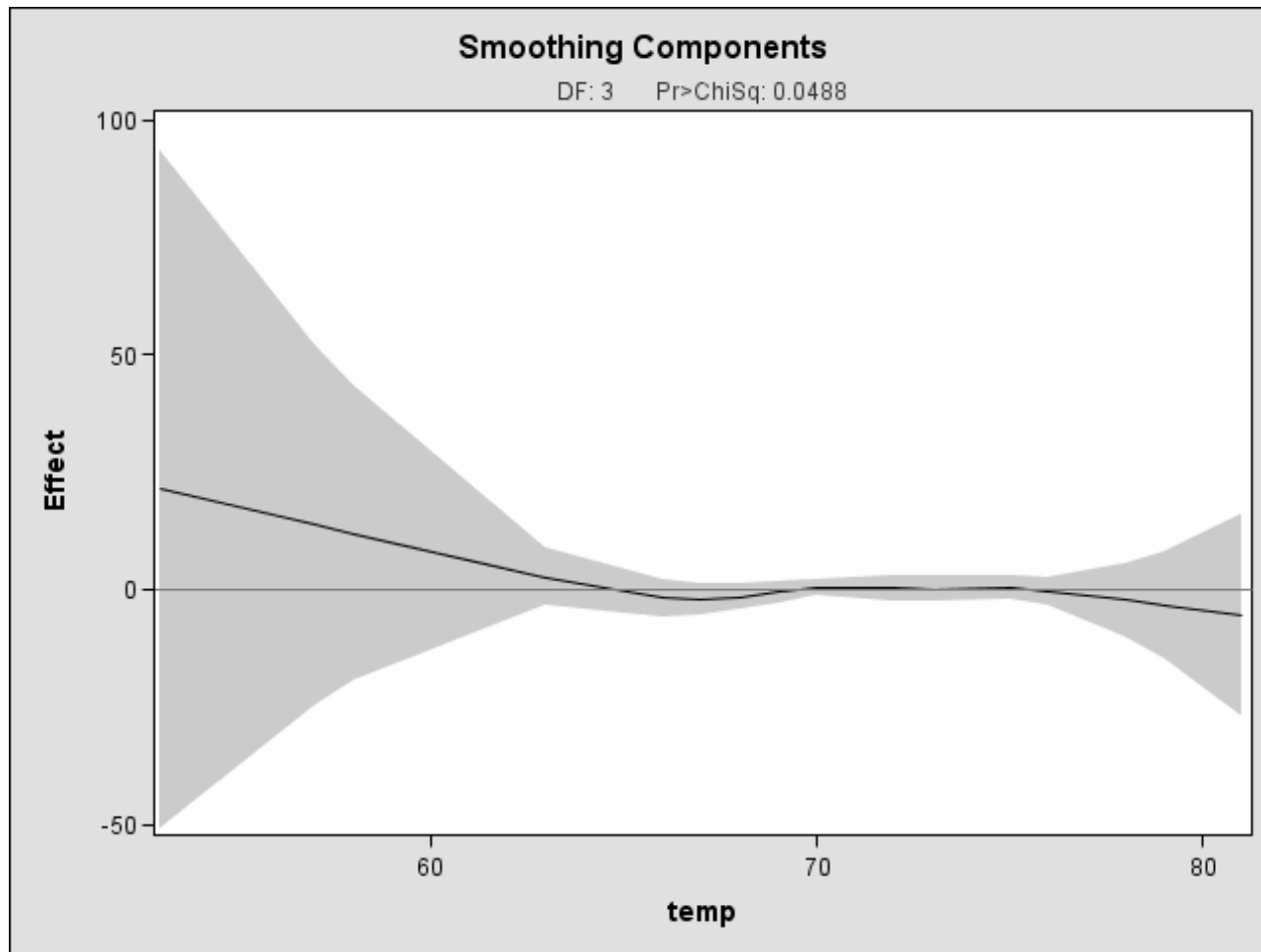
Smoothing Model Analysis

Analysis of Deviance

Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline(temp)	3.00000	7.870171	7.8702	0.0488

The Analysis of Deviance table gives a χ^2 -test from comparing the deviance between the full model and the model with this variable dropped – here the intercept model *plus a linear effect in temperature*. We see that temperature effect is significantly nonlinear at the 5% level. The default $df = 3$ corresponds to a smoothing spline with the complexity of a cubic polynomial.

The following plot was obtained from the `plots(c1m)` statement. The plot has the estimated smoothing spline function with the linear effect subtracted out. The plot includes a 95% curvewise Bayesian confidence band. We visually inspect where this band does not include zero to get an idea of where significant nonlinearity occurs. This plot can suggest simpler transformations of predictor variables than use of the full-blown smoothing spline.



The band basically includes zero for most temperature values; at a few points it comes close to not including zero.

The plot spans the range of temperature values in the data set and becomes highly variable at the ends. Do you think extrapolation is a good idea using GAMs?

We want to predict the probability of a failure at 39 degrees. I couldn't get GAM to predict beyond the range of predictor values.