

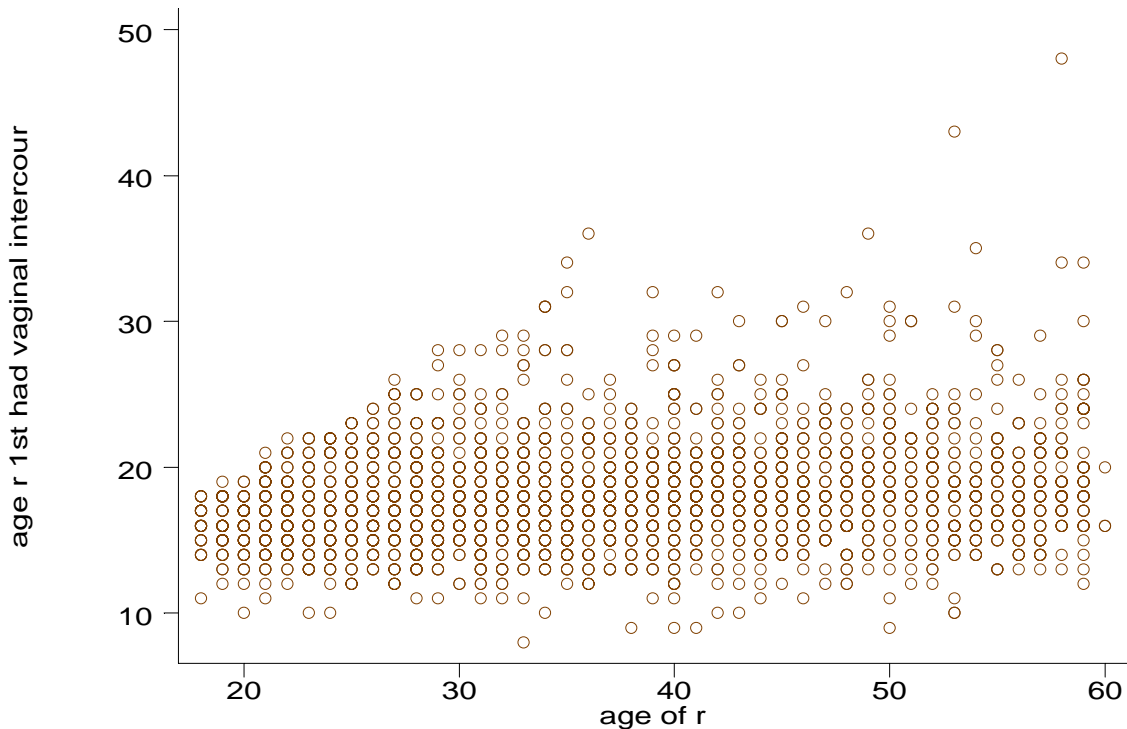
# Loglinear analysis

1. Why loglinear?
2. Basic Concepts.
3. Basic Models.
4. Reading SPSS outputs.
5. Model Comparison.

## 1. Why loglinear?

### a. Loglinear vs. Linear-regression

graph firstvi age



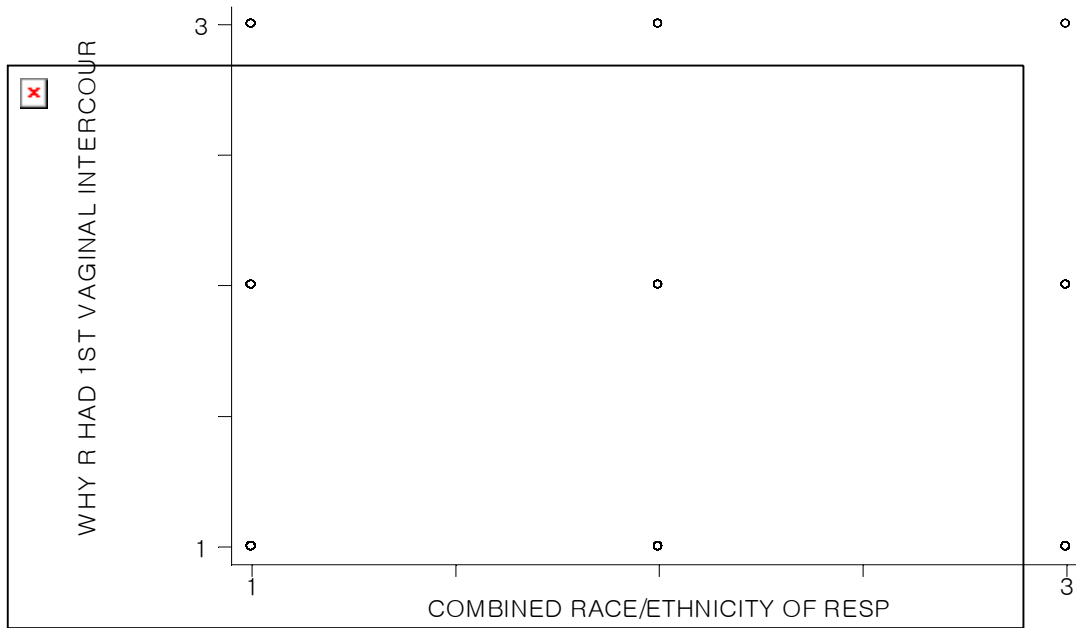
. regress firstvi age

Source	SS	df	MS			
Model	1679.64286	1	1679.64286	Number of obs =	3030	
Residual	30755.1786	3028	10.1569282	F( 1, 3028) =	165.37	
Total	32434.8215	3029	10.7080956	Prob > F =	0.0000	
				R-squared =	0.0518	
				Adj R-squared =	0.0515	
				Root MSE =	3.187	

firstvi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	.0684898	.005326	12.860	0.000	.0580469	.0789327
_cons	15.31821	.2041205	75.045	0.000	14.91798	15.71844

**No fundamental problem with this linear regression.**



COMBINED RACE/ETHNICITY OF RESP	WHY R HAD 1ST VAGINAL INTERCOUR			Total
	1	2	3	
1	1921 81.95	371 15.83	52 2.22	2344 100.00
2	373 70.38	136 25.66	21 3.96	530 100.00
3	243 79.15	57 18.57	7 2.28	307 100.00
Total	2537 79.75	564 17.73	80 2.51	3181 100.00

Pearson  $\chi^2(4) = 36.2854$  Pr = 0.000

**1: Whites, 2: Blacks, 3: Hispanics**

**1:something R wanted, 2: R went along with, 3: forced to do**

. regress fvwhy ethnic

Source	SS	df	MS	Number of obs = 3181		
Model	2.64004436	1	2.64004436	F( 1, 3179) =	11.71	
Residual	716.576554	3179	.225409423	Prob > F =	0.0006	
				R-squared =	0.0037	
				Adj R-squared =	0.0034	
Total	719.216599	3180	.226168742	Root MSE =	.47477	

fvwhy	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ethnic	.0442782	.0129381	3.422	0.001	.0189103	.0696461
_cons	1.167399	.0195015	59.862	0.000	1.129162	1.205636

**Problems with this linear regression**

1. **Linearity assumption**
2. **Normal Distribution assumption**
3. **Equidistance assumption for dependent and independent variables.**

**b. Loglinear vs. Other regressions (linear regression with categorical variables, ordinal logit model, multinomial logit model, logistic model)**

**\*Linear Regression with categorical independent variable**

```
. xi: regress fvwhy i.ethnic

i.ethnic          Iethni_1-3  (naturally coded; Iethni_1 omitted)

-----+-----
Source |           SS          df           MS                Number of obs =   3181
-----+-----+-----+-----
Model |   7.67432539         2   3.83716269                F( 2, 3178) =   17.14
Residual |  711.542273       3178   .223896247                Prob > F      =   0.0000
-----+-----+-----+-----
Total |  719.216599       3180   .226168742                R-squared     =   0.0107
                                           Adj R-squared =   0.0100
                                           Root MSE     =   .47318

-----+-----
fvwhy |           Coef.      Std. Err.      t    P>|t|     [95% Conf. Interval]
-----+-----+-----+-----+-----
Iethni_2 |   .133204      .0227588     5.853  0.000     .0885805   .1778275
Iethni_3 |   .0286253     .0287197     0.997  0.319    -.0276858   .0849364
   _cons |   1.202645     .0097734    123.053 0.000     1.183482   1.221808
-----+-----
```

**Problems?**

- 1. Linearity assumption**
- 2. Normality assumption**
- 3. Equidistance for dependent variable**

**\*Ordinal logit**

```
. xi: ologit fvwhy i.ethnic

i.ethnic          Iethni_1-3  (naturally coded; Iethni_1 omitted)

Iteration 0:  Log Likelihood =-1844.1989
Iteration 1:  Log Likelihood =-1827.7506
Iteration 2:  Log Likelihood =-1827.4046
Iteration 3:  Log Likelihood =-1827.4045

Ordered Logit Estimates                Number of obs =   3181
                                         chi2(2)       =   33.59
                                         Prob > chi2   =   0.0000
Log Likelihood = -1827.4045            Pseudo R2     =   0.0091

-----+-----
fvwhy |           Coef.      Std. Err.      z    P>|z|     [95% Conf. Interval]
-----+-----+-----+-----+-----
Iethni_2 |   .6459637     .1086218     5.947  0.000     .433069   .8588585
Iethni_3 |   .1746721     .1499426     1.165  0.244    -.11921   .4685543
-----+-----
   _cut1 |   1.512525     .0536716                                (Ancillary parameters)
   _cut2 |   3.812296     .117999
-----+-----

. predict wanted along forced
(option p assumed; predicted probabilities)

. sort ethnic
```

. by ethnic: summ wanted along forced

```
-> ethnic= 1
Variable | Obs      Mean      Std. Dev.      Min      Max
-----+-----
wanted  | 2344    .8194351      0      .8194351    .8194351
along   | 2344    .1589453      0      .1589453    .1589453
forced  | 2344    .0216196      0      .0216196    .0216196
```

```
-> ethnic= 2
Variable | Obs      Mean      Std. Dev.      Min      Max
-----+-----
wanted  | 530     .7040296      0      .7040296    .7040296
along   | 530     .2555179      0      .2555179    .2555179
forced  | 530     .0404525      0      .0404525    .0404525
```

```
-> ethnic= 3
Variable | Obs      Mean      Std. Dev.      Min      Max
-----+-----
wanted  | 307     .7921365      0      .7921365    .7921365
along   | 307     .1822234      0      .1822234    .1822234
forced  | 307     .0256401      0      .0256401    .0256401
```

**Odds ratio of 'along' vs. 'wanted'**  
**Black vs. White:  $(0.26/0.7)/(0.16/0.82) = 1.9$**   
**Black vs. Hispanic:  $(0.26*0.7)/(0.18*0.79)=1.6$**   
**Hispanic vs. White:  $(0.18*0.79)/(0.16*0.82)=1.2.$**

**From loglinear (Row effect model):**

Goodness-of-Fit test statistics

```
Likelihood Ratio Chi Square = 2.80412   DF = 2   P = .246
Pearson Chi Square = 2.78004   DF = 2   P = .249
```

Estimates for Parameters

```
ETHNIC
Parameter   Coeff.   Std. Err.   Z-Value   Lower 95 CI   Upper 95 CI
-----+-----
1  1.445288212   .07569   19.09600   1.29694   1.59363
2  -.695399684   .09343   -7.44284   -.87853   -.51227
FVWHY
Parameter   Coeff.   Std. Err.   Z-Value   Lower 95 CI   Upper 95 CI
-----+-----
3  1.560035109   .05285   29.51670   1.45644   1.66363
4  .1647811098   .04747   3.47151   .07175   .25782
ETHNIC BY F
Parameter   Coeff.   Std. Err.   Z-Value   Lower 95 CI   Upper 95 CI
-----+-----
5  -.216937411   .05659   -3.83316   -.32786   -.10601
6  .3003860149   .06693   4.48807   .16920   .43157
```

**Odds ratio of 'along' vs. 'wanted'**  
**Black vs. White:  $\exp(.3-(-.22))=1.7$**   
**Black vs. Hispanic:  $\exp(0.3-(-0.08))=1.5$**   
**Hispanic vs. White:  $\exp(-0.08-(-0.22))=1.2.$**

**Problem of ordinal logit: model fitting**

## From another loglinear (RC model)

### GOODNESS-OF-FIT INFORMATION

```

PEARSON CHI-SQUARE          =          .14366
LIKELIHOOD-RATIO CHI-SQUARE =          .14788

DEGREES OF FREEDOM         =              1

INDEX OF DISSIMILARITY     =          .00068

FINAL ITERATION            =              59
MAXIMUM DEVIATION          =          .00098114
  
```

### CROSS-RATIOS

```

ROWS          COLS          OBSERVED          EXPECTED
=====
1, 2          1, 2          1.8879          1.8882
1, 2          2, 3          1.1017          1.1028
2, 3          1, 2          .6433          .6294
2, 3          2, 3          .7953          .9312
  
```

### MODEL PARAMETERS

```
=====
```

```

THETA (PHI FOR MODEL II) =          .41187

ROW          ALPHA          GAMMA(MU FOR MODEL II)

1           869.90354        -.57957
2           266.57839        .78785
3           124.63738        -.20828

COLUMN          BETA          DELTA(NU FOR MODEL II)

1           1.81982         -.81031
2           .46111          .31825
3           .06635          .49205
  
```

## Odds ratio of 'along' vs. 'wanted'

**Black vs. White: = 1.9**

**Black vs. Hispanic: = 1/0.6 = 1.7**

**Hispanic vs. White: 1.9/1.7 = 1.1**

There is no problem with ordinal logit model above. Actually, their estimates are almost identical to ones from loglinear models. However each model has its own limitations.

### Ordinal logit model

1. What if we want to test other functional forms (or models)? Or if we have symmetric table requiring quasi-models? Ordinal model has fixed function form<sup>1</sup>. This (and other regressions too) lacks 'model flexibility'.

---

<sup>1</sup>  $\Pr(\text{outcome}=i, \text{ when dependent is } j) = \frac{1}{1 + \exp(-k_i + \sum \beta_j \chi_j)} - \frac{1}{1 + \exp(-k_{i-1} + \sum \beta_j \chi_j)}$

2. Where are 'goodness-of-fit' tests? ( $R^2$  has no statistical distribution allowing significance test).

Loglinear model

1. It is not easy to handle many variables at the same time: three or four ways are practical maximum.

**c. Loglinear vs. Simple  $\chi^2$  test.**

If  $\chi^2$  test coming with cross tabulation says that there is no relationship between two variables (say, p-value=0.3), does that really mean there is no relationship? We cannot be sure because it is just possible other model (say, 'row effect' model) turns out to be better model. Even if independence model fits to the data, there can be room for model improvement.

## 2. Basic Concepts

### a. Odds

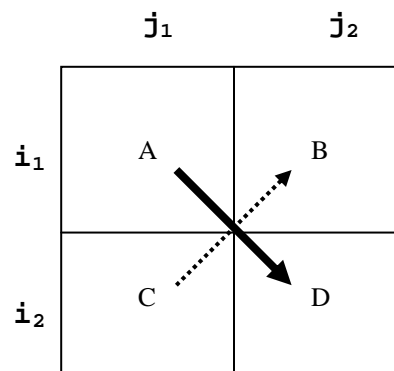
Odds means ratio of two probabilities.  
 Odds of A versus B =  $P(A)/P(B)=f_A/f_B$

### b. Odds ratio: the ratio of two odds: engine of the analysis

Let's examine the following table.

NO. OF PARTNERS LAST YEAR CONST	esatis		Total
	1	2	
1	1945	444	2389
	81.41	18.59	100.00
	85.05	66.07	80.74
more than 1	342	228	570
	60.00	40.00	100.00
	14.95	33.93	19.26
Total	2287	672	2959
	77.29	22.71	100.00
	100.00	100.00	100.00

Pearson  $\chi^2(1) = 120.2346$  Pr = 0.000



How can we see the relationship between the number of partners and emotional satisfaction with a primary partner?

First, somebody can argue that among the monogamous people, 81% are extremely satisfied while only 60% of polygamous people are extremely satisfied. This means exactly that (A vs. B) is greater than (C vs. D). In other words, ( $j_1$  vs.  $j_2$  in the row  $i_1$ ) is greater than ( $j_1$  vs.  $j_2$  in the row  $i_2$ ).

We can express this by using odds. (odds of  $j_1$  vs.  $j_2$  in the row  $i_1$ ) is greater than (odds of  $j_1$  vs.  $j_2$  in the row  $i_2$ ). If we think 'odds ratio' as ratio of two odds ( $j_1$  vs.  $j_2$ ) between each rows ( $i_1$  vs.  $i_2$ ), we can express it as 'odds ratio is greater than 1'. So now, we can express the relationship in two by two table using just one value! - odds-ratio.

In a nutshell, if odds-ratio is greater than 1, then  $i_2$  (instead of  $i_1$ ) is more likely to be  $j_2$  (instead of  $j_1$ ): solid line arrow.

If odds-ratio is 1, then odds in  $i_1$  is the same as odds in  $i_2$  and there is no relationship between two variables.

If odds-ratio is less than 1, then odds in  $i_1$  is less than odds in  $i_2$ . In other words, ( $j_1$  vs.  $j_2$  in the row  $i_2$ ) is greater than ( $j_1$  vs.  $j_2$  in the row  $i_1$ ): dotted line arrow.

### Variety of Interpretation

$$\theta \text{ (odds-ratio)} = \frac{AD}{BC}$$

$$= \frac{\frac{A}{B}}{\frac{C}{D}} : (j_1 \text{ vs. } j_2 \text{ in } i_1) \text{ vs. } (j_1 \text{ vs. } j_2 \text{ in } i_2) = \frac{\frac{A}{C}}{\frac{B}{D}} : (i_1 \text{ vs. } i_2 \text{ in } j_1) \text{ vs. } (i_1 \text{ vs. } i_2 \text{ in } j_2)$$

$$= \frac{\frac{D}{C}}{\frac{B}{A}} : (j_2 \text{ vs. } j_1 \text{ in } i_2) \text{ vs. } (j_2 \text{ vs. } j_1 \text{ in } i_1) = \frac{\frac{D}{B}}{\frac{C}{A}} : (i_2 \text{ vs. } i_1 \text{ in } j_2) \text{ vs. } (i_2 \text{ vs. } i_1 \text{ in } j_1)$$

If this  $\theta$  is less than 1, it is more natural to express the association using  $1/\theta$ .

$$1/\theta = \frac{\frac{B}{A}}{\frac{D}{C}} : (j_2 \text{ vs. } j_1 \text{ in } i_1) \text{ vs. } (j_2 \text{ vs. } j_1 \text{ in } i_2) : \text{dotted line.}$$

Any form of associations (and thus, models) can be expressed by using this odds-ratio.

### c. $\chi^2$ (Pearson chi-square) and $G^2$ (Likelihood-ratio chi-square)

$$\chi^2 = \sum_i \sum_j \frac{(F_{ij} - \hat{F}_{ij})^2}{\hat{F}_{ij}}, \quad G^2 = 2 \sum_i \sum_j F_{ij} \ln \frac{F_{ij}}{\hat{F}_{ij}}, \text{ where } F_{ij} \text{ is observed frequencies and } \hat{F}_{ij} \text{ is}$$

expected frequencies from the model.

1. What are the meanings of these statistics? Difference between observed and expected frequencies. In other words, it measures how much the model, which produces expected frequencies, does not fit to the data (i.e., observed frequencies). So less statistics means more fit. So  $\chi^2$  must be small enough to make p-value greater than  $\alpha$  level (say, 0.05) in order to make the model fit to the data.
2. Try to calculate them for a simple model (say, independence model). Think about the above two by two table regarding # of partners and 'emotional satisfaction'.

$$\hat{F}_{11} \cong 2959 \times 0.8074 \times 0.7729 = 1846.53$$

$$\hat{F}_{12} \cong 2959 \times 0.8074 \times 0.2271 = 542.564$$

$$\hat{F}_{21} \cong 2959 \times 0.1926 \times 0.7729 = 440.478$$

$$\hat{F}_{22} \cong 2959 \times 0.1926 \times 0.2271 = 129.425$$

$$\chi^2 \cong [(1945-1846.53)^2/(1846.53) + (444-542.564)^2/(542.564) + (342-440.478)^2/(440.478) + (228-129.425)^2/(129.425)] = 120.252$$

3. For large sample sizes, these statistics are equivalent. So that we can use the difference between these two as a sign for stability of estimation. Also if zero frequency (or small frequency) occupies many cells, these two diverge in different directions. Which statistic goes larger?
4. The advantage of  $G^2$  statistic is that it, like the total sums of squares in the analysis of variance, can be subdivided into interpretable parts that add up to the total. So when we compare nested models, we have to use  $G^2$ . So for the goodness of test itself, we use  $\chi^2$  (you might want to compare this to  $G^2$  if you have many small-frequency cells) and we use  $G^2$  when we do comparisons of models<sup>2</sup>.
5. Suppose we make all the cell frequencies  $n$  times larger than ones as in the original table. This only change makes  $\chi^2$   $n$  times larger with the same degree of freedom. So if we have large frequencies of cells, the fact itself makes the independence model more likely to be rejected and conclude that there is some association. As many people argue, is this a problem of  $\chi^2$  statistic?

---

<sup>2</sup> Let me summarize some confusing statistics here.  $G^2=2[\log\text{-likelihood of the saturated model} - \log\text{-likelihood of tested model}]$ .  $L^2=2[\log\text{-likelihood of tested model} - \log\text{-likelihood of constant rate model}]$ . Pseudo  $R^2=[1 - \ln(\log\text{-likelihood of tested model})/\ln(\log\text{-likelihood of constant rate model})]$ .  $G^2$  is used for loglinear model comparisons and  $L^2$  is used for diverse 'regression' model comparisons.

### 3. Basic Models

How can we model the pattern of associations? Let's review some basic models with mathematical equations.

#### a. Independence model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C$ , where  $\sum_i \lambda_i^R = 0$  and  $\sum_j \lambda_j^C = 0$ , R and C is superscript to indicate corresponding variables not powers.

How does this mathematical equation mean 'independence'?

Always, think about it in terms of odds-ratio.

$$\ln(\theta) = \left( \frac{F_{ij} F_{i'j'}}{F_{i'j} F_{ij'}} \right) = \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{i'j}) - \ln(\hat{F}_{ij'}) = 0.$$

Then,  $\theta = 1$ . So there is no association.

Number of parameters used: one for  $\lambda$ , (I-1) for  $\lambda_i^R$ , and (J-1) for  $\lambda_j^C$ .

So degree of freedom =  $I*J - (1 + I - 1 + J - 1) = (I-1)*(J-1)$ .

#### b. Uniform association model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \beta * i * j$ , ( $i$  and  $j$  is not symbols but they are just values, i.e., scores that are not estimated)  $\rightarrow i$  and  $j$  are treated as interval variables just as linear-regression. So, at least they must be pre-ordered.

$$\begin{aligned} \ln(\theta) = \left( \frac{F_{ij} F_{i'j'}}{F_{i'j} F_{ij'}} \right) &= \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{i'j}) - \ln(\hat{F}_{ij'}) = \beta * i * j + \beta * i' * j' \\ &\quad - \beta * i' * j - \beta * i * j' = \beta (i - i')(j - j') \end{aligned}$$

So in uniform association model, log-odds are the same for any adjacent sub-tables. How many more parameters are introduced than independence model?

#### c. row-effect model

$\ln(\hat{F}_{ij}) = \lambda + \lambda_i^R + \lambda_j^C + \mu_i * j$ ,  $j$  is a score (value of the column variable) not a parameter to be estimated.

$$\ln(\theta) = \left( \frac{F_{ij} F_{i'j'}}{F_{i'j} F_{ij'}} \right) = \ln(\hat{F}_{ij}) + \ln(\hat{F}_{i'j'}) - \ln(\hat{F}_{i'j}) - \ln(\hat{F}_{ij'}) = \mu_i * j + \mu_{i'} * j' - \mu_{i'} * j - \mu_i * j'$$

$$= (\mu_i - \mu_{i'})(j - j')$$

So in row-effect model, the adjacent sub-table odds-ratio is determined by difference of 'row' parameters. Here, we assume column variable as interval variable. So, at least they must be ordinal.

How many more parameters are used than in uniform association model?

How is uniform association model nested within row-effect model?

Try other models, say, column-effect model, RC model, R+C model, RC<sub>H</sub>, and (R+C)<sub>H</sub> models.

	Added parameters to independence model	ln(θ)
C	$v_j * i$	$(v_j - v_{j'})(i - i')$
RC	$\phi * \mu_i * v_j$	$\phi(\mu_i - \mu_{i'})(v_j - v_{j'})$
R+C	$\mu_i * j + v_j * i$	$(\mu_i - \mu_{i'})(j - j') + (v_j - v_{j'})(i - i')$
RC <sub>H</sub>	$\phi * \mu_i * \mu_i$	$\phi(\mu_i - \mu_{i'})(u_j - u_{j'})$
(R+C) <sub>H</sub>	$\mu_i * j + \mu_i * i$	$(\mu_i - \mu_{i'})(j - j') + (\mu_j - \mu_{j'})(i - i')$

## 4. Reading SPSS outputs

```

1 0 set length=none/width=100
2 0 data list list/ ethnic fvwhy freq
3 0 weight by freq
4 0 compute f=fvwhy
5 0 loglinear ethnic fvwhy (1,3) with f
6 0 /print=estim, cor
7 0 /design=ethnic, fvwhy, ethnic by f → Model 1
8 0 /design=ethnic, fvwhy, ethnic by fvwhy → Model 2
9 0 /contrast(ethnic)=deviation(2)
10 0 /design=ethnic, fvwhy, ethnic by f → Model 3

```

### - Model 1 -

Goodness-of-Fit test statistics

```

Likelihood Ratio Chi Square = 2.80412 DF = 2 P = .246
Pearson Chi Square = 2.78004 DF = 2 P = .249

```

### **Does this model fit to the data? (think about the meaning of Chi Square)**

ETHNIC BY F (White if ethnic=1, Black if ethnic=2, and Hispanic if ethnic=3)

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
5	<b>-.216937411</b>	.05659	-3.83316	-.32786	-.10601
6	<b>.3003860149</b>	.06693	4.48807	.16920	.43157

Covariance(below) and Correlation(above) Matrices of Parameter Estimates

Parameter	Parameter					
	1	2	3	4	5	6
1	.00573	-.11432	-.54038	.07386	-.93050	.07827
2	-.00081	.00873	-.12805	-.08631	.07506	-.92498
3	-.00216	-.00063	.00279	.33679	.56524	.17707
4	.00027	-.00038	.00084	.00225	-.07858	.09464
5	-.00399	.00040	.00169	-.00021	<b>.00320</b>	-.05840
6	.00040	-.00578	.00063	.00030	<b>-.00022</b>	<b>.00448</b>

### **What is the last parameter not shown?**

### **Could you argue that there is a significant difference between White and Black regarding the reason for the first virginal intercourse?**

$$\frac{(0.3 - (-0.22))}{\sqrt{(0.0032) + (0.00448) - 2 * (-0.00022)}} = 5.77, \text{ significant at } 0.1\% \text{ level.}$$

### **Could you conclude that there is a significant difference between White and Hispanic? The following is an algebraic method.**

Suppose a = parameter for Whites and b = parameter for Blacks. Then the parameter for Hispanic = -(a + b).

Difference = a - (-(a+b)) = 2\*a + b.

Var(2a + b) = Var(2a) + Var(b) + 2\*Cov(2a, b) = 4\*Var(a) + Var(b) + 2\*2\*Cov(a, b).

(∵ Var(aX) = a<sup>2</sup>\*VAR(X), Cov(X,Y) = E(XY) - E(X)E(Y)).

However there is a more convenient alternative. Please take a look at Model3.

**Another little trick for the significance test of (a+b+c)**

$$z = (a+b+c)/(\text{sqrt}(\text{var}(a+b+c))).$$

$$\text{var}(a+b+c) = \text{var}[(a+b)+c] = \text{var}(a+b) + \text{var}(c) + 2*\text{cov}[(a+b),c]$$

$$\begin{aligned} \text{cov}[(a+b), c] &= E[(a+b)(c)] - E[a+b]*E(c) = E[ac + bc] - [E(a) + E(b)]*E(c) \\ &= E[ac] + E[bc] - E(a)E(c) - E(b)E(c) = \text{cov}(a,c) + \text{cov}(b,c). \end{aligned}$$

$$\text{Thus, } z = \frac{a + b + c}{\sqrt{\text{var}(a) + \text{var}(b) + \text{var}(c) + 2\text{cov}(a,b) + 2\text{cov}(b,c) + 2\text{cov}(a,c)}}$$

**- Model 2 -**

**This is a saturated model.**

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square = .00000 DF = 0 P = 1.000  
 Pearson Chi Square = .00000 DF = 0 P = 1.000

ETHNIC BY FVWHY

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
5	.1890796885	.06354	2.97594	.06455	.31361
6	-.089820214	.07087	-1.26738	-.22873	.04909
7	-.271554420	.07447	-3.64668	-.41751	-.12560
8	.0862648331	.08250	1.04563	-.07544	.24797

**Just for an example of reading parameters.**

	Wanted	Get along	Forced
<b>White</b>	<b>0.19</b>	<b>-0.09</b>	<b>-(0.19-0.09)</b>
<b>Black</b>	<b>-0.27</b>	<b>0.086</b>	<b>-(-0.27+0.086)</b>
<b>Hispanic</b>	<b>-(0.19-0.27)</b>	<b>-(-0.09+0.086)</b>	<b>(0.19-0.27-0.09+0.086)</b>

**Why is not this saturated model our goal even though it always fits the data perfectly?**

**- Model 3 -**

ETHNIC BY F

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
5	-.216937411	.05659	-3.83316	-.32786	-.10601
6	-.083448604	.08509	-.98072	-.25022	.08333

Covariance(below) and Correlation(above) Matrices of Parameter Estimates

Parameter	Parameter					
	1	2	3	4	5	6
1	.00573	-.57364	-.54038	.07386	-.93050	.55734
2	-.00492	.01284	.46650	.02183	.55959	-.93097
3	-.00216	.00279	.00279	.33679	.56524	-.51523
4	.00027	.00012	.00084	.00225	-.07858	-.02217
5	-.00399	.00359	.00169	-.00021	.00320	-.61919
6	.00359	-.00898	-.00232	-.00009	-.00298	.00724

***This makes us to argue about the significance of difference between White and Hispanic without using algebraic method.***

***What happened if you type 'fvwhy by ethnic' instead of 'ethnic by fvwhy'?***

***Why is it?***

```

1 0 set length=none/width=100
2 0 data list list/ ethnic fvwhy freq
3 0 weight by freq
4 0 compute f=fvwhy
5 0 loglinear ethnic fvwhy (1,3) with f
6 0 /print=estim, cor
7 0 /design=ethnic, fvwhy, ethnic by f
8 0 /design=ethnic, fvwhy, fvwhy by ethnic
9 0 /contrast(ethnic)=deviation(2)
10 0 /design=ethnic, fvwhy, ethnic by f

```

FVWHY BY ETHNIC

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
5	.1890796885	.06354	2.97594	.06455	.31361
6	-.089820214	.07087	-1.26738	-.22873	.04909
7	-.271554420	.07447	-3.64668	-.41751	-.12560
8	.0862648331	.08250	1.04563	-.07544	.24797

## 5. Model Comparison

- Write down the model name, number of degrees of freedom,  $\chi^2$ ,  $G^2$ , and p-value.
- Take a look at the difference between  $\chi^2$  and  $G^2$ . Small difference means reliable estimation.
- From the p-value of  $\chi^2$ , determine which model fits to the data.
- Make pair-wise comparisons of a model fitting to the data with all the other nested models.

### An example.

From the 'ethnic' and 'fvwhy' table.

	D.F.	$G^2$	$\chi^2$	p-value
Independence	4	33.8	36.3	<0.001
Uniform	3	22.7	23.2	<0.001
Row effect	2	2.8	2.8	0.3879
RC	1	0.15	0.14	0.7

So 'Row effect' and 'RC' models fit to the data. So comparisons are as follows.

R vs. I: 2, 31, p-value <0.001

R vs. U: 1, 19.9, p-value <0.001

R vs. RC: 1, 2.65, p-value >0.1

RC vs. I: 3, 33.65, p-value <0.001

RC vs. U: 2, 22.55, p-value <0.001

**1. If p-value is significant, that means added parameters are significantly greater than 0, so the model with more parameters is better: added parameter improves the model significantly.**

**2. If p-value is not significant, that means added parameters are not significantly greater than 0, so the model with less parameter is better: added parameters do not improve the model significantly.**

**3. Why do we have to include non-fitting models in the comparisons?**

**4. After the first three comparisons, do we have to do the last two comparisons too?**

*Think about the following example.*

*Model 1: 14, 11.84*

*Model 2: 16, 17.47*

*Model 3: 17, 21.12*

**What can be our conclusion about the above situation?**

**5. For comparison, two models must be nested. What can be exceptional case?**