STAT 371: LAB MANUAL: WINTER 2015

MACEWAN UNIVERSITY
KATHLEEN LAWRY-BATTY

Contingency Tables

A contingency table is a place where a joint distribution of two categorical variables, say X and Y, where X has I levels and Y has J levels, can be summarized. It is usual to consider an X variable to be an explanatory variable, and a Y variable a response variable. Both marginal distributions and conditional distributions can be of interest. Data that is used can either be offered in raw form (Format 1) or in summarized form (Format 2) in SPSS.

Example:
The file HockeyGenderRaw contains two variables. The response variable WatY (Yes-1, No-2) records whether or not an individual watches hockey and appears in column 1 and the explanatory variable GenX (Male -1, Female-2) records gender and appears in column 2.

(FORMAT 1)

WatY	GenX
1	1
1	1
….	
1	2
1	2
1	2
….	
2	1
2	1
2	1
….	
2	2
2	2
2	2
….	

The file WatchHockeyGenderSummarized summarizes the counts of observations for the (GenX, WatY) pairs (1,1), (1,2), (2,1) and (2,2).

(FORMAT 2)

GenX	WatY	Count
1	1	33
1	2	11
2	1	9
2	2	36

When working with raw data with columns of categorical variables, one can create a summary cross-tab table that counts the groupings. Here are the commands to do so for the hockey watching and gender raw data.

Commands

Analyze>Descriptive Statistics>Crosstabs
Row(s): GenX
Column(s): WatY
OK

GenX * WatY Crosstabulation

			WatY		Total
			1	2	
GenX	1	Count	33	11	44
		% within GenX	75.0%	25.0%	100.0%
	2	Count	9	36	45
		% within GenX	20.0%	80.0%	100.0%
Total		Count	42	47	89
		% within GenX	47.2%	52.8%	100.0%

3 ways of comparing two categorical variables (each with 2 levels)

It is usual to have X in the row, and Y in the column in a contingency table. Here n_{ij} represents the count of observations in the ijth cell, i=1,2 and j=1,2, while n_{i+} is the count in row I, n_{+J} is the count in column j, and n is the total number of observations.

		Y(Response)		
		1	2	Total
X(Explanatory)	1	$\mathbf{n_{11}}$	n_{12}	$\mathbf{n_{1+}}$
	2	$\mathbf{n_{21}}$	n_{22}	$\mathbf{n_{2+}}$
	Total	n_{+1}	n_{2+}	n

We write:

$P(Y=1|X=1) = p_1 = n_{11}/n_{1+}$

$P(Y=1|X=2) = p_2 = n_{21}/n_{2+}$

Compare proportions:

The **Wald (1-α)% CI** is: $(p_1-p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Relative Risk:

rr = π_1/π_2, $\widehat{rr} = p_1/p_2$

The distribution of $\widehat{rr}$ is very skewed, so we look at the distribution of **$\ln(\widehat{rr})$** instead.

The CI for ln(rr) is : $\ln(p_1/p_2) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1p_1} + \frac{1-p_2}{n_2p_2}}$

Convert (exponentiate) the endpoints to find a CI for rr.

Odds

$Odds_i$(of success in row i)= $p_i/(1-p_i)$, Hence: $p_i = odds_i/(odds_i + 1)$

Odds Ratio: $\theta = odds_1/odd_2$

$\hat{\boldsymbol{\theta}} = n_{11}n_{22}/n_{12}n_{21}$ (both **variables** response variables)

$\hat{\boldsymbol{\theta}}$ is MLE of θ

Distribution of $\hat{\boldsymbol{\theta}}$ very skewed, so look at distribution of $\ln(\hat{\boldsymbol{\theta}})$ instead

CI: for ln(θ): $\ln(\hat{\theta}) \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}}+\frac{1}{n_{22}}+\frac{1}{n_{12}}+\frac{1}{n_{21}}}$

Convert (exponentiate) the endpoints to find a CI for θ .

Class Example:

In SPSS

PCR	Relapse	Total
1	1	30
1	2	45
2	1	8
2	2	95

PCR is X (explanatory variable)

Relapse is Y (response variable)

SPSS Commands (to produce contingency table)

Data>Weight Cases *

Weight by Frequency Variable: Total

OK

Analysis>Descriptive Statistics>Crosstabs

Row(s): PCR

Column(s): Relapse

Statistics popup: ✓ Risk

Cells popup: Counts: ✓ Observed

Percentages: ✓ Row

*Note: If data is in two columns, one for PCR (X) and one for Relapse (Y), the weight cases command lines are unnecessary.

PCR * Relapse Crosstabulation

			Relapse		Total
			1	2	
PCR	1	Count	30	45	75
		% within PCR	40.0%	60.0%	100.0%
	2	Count	8	95	103
		% within PCR	7.8%	92.2%	100.0%
Total		Count	38	140	178
		% within PCR	21.3%	78.7%	100.0%

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for PCR (1 / 2)	**7.917**	**3.361**	**18.648**
For cohort Relapse = 1	**5.150**	**2.504**	**10.590**
For cohort Relapse = 2	.651	.536	.789
N of Valid Cases	178		

$P(Y=1|X=1) = p_1 = n_{11}/n_{1+} = 0.4$, $P(Y=1|X=2) = p2 = n_{21}/n_{2+} = 0.078$

(1-α)% Wald CI for (p_1-p_2) = $(p_1-p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$

95% CI is (.200, 0.445) (this must be calculated by hand)
Note that 0 is not in the 95% CI.
With 95% confidence, we have sufficient evidence that the proportion of PCR positive children who relapse is between 0.2 and 0.445 larger than the proportion of PCR- children who relapse.
Another way of saying this:
With 95% confidence, we have sufficient evidence that the proportion of children who relapse if they have a positive PCR is between 0.2 and 0.445 larger than the proportion of children who relapse if they have a negative PCR.

Estimated relative risk = $\widehat{rr} = p_1/p_2$ =0.4/0.78 = 5.15,
95% CI for rr is (2.504, 10.590) from above. (exponentiated endpoints of CI for $\ln(p_1/p_2)$)
Note that 1 is not in the 95% CI.
With 95% confidence, we have sufficient evidence that probability of a relapse is between 2.5 and 10.6 times higher for children with a positive PCR than a negative PCR.

Estimated odds ratio = $\widehat{\theta} = n_{11}n_{22}/n_{12}n_{21}$ = (30x95)/(8x45) = 7.917
95% CI for θ is (3.361, 18.648), from above. (exponentiated endpoints of CI for $\ln(\widehat{\theta})$)
Note that 1 is not in the 95% CI.
With 95% confidence, we have sufficient evidence that a relapse is associated with PCR. The odds of a relapse is 3.361 to 18.648 X higher for the PCR+ group than for the PCR- group.

Odds Ratio (another option for commands for a 2x2 contingency table with 2 levels per category)
SPSS Commands
Analysis>Descriptive Statistics>Crosstabs
Row(s): PCR
Column(s): Relapse
Statistics popup:
Cochran's and Mantel-Haenszel statistics
Test common odds ratio equals: 1
Cells popup: Counts: ✓ Observed

***These commands can only produce a 95% confidence interval.**
***These commands also produce a confidence interval for $\ln(\widehat{\theta})$**

Mantel-Haenszel Common Odds Ratio Estimate

Estimate			7.917
ln(Estimate)			2.069
Std. Error of ln(Estimate)			.437
Asymp. Sig. (2-sided)			.000
Asymp. 95% Confidence Interval	Common Odds Ratio	Lower Bound	3.361
		Upper Bound	18.648
	ln(Common Odds Ratio)	Lower Bound	1.212
		Upper Bound	2.926

The Mantel-Haenszel common odds ratio estimate is asymptotically normally distributed under the common odds ratio of 1.000 assumption. So is the natural log of the estimate.

Example: Gender/Hockey Data

Illustrate the gender and hockey watching variables in the three ways introduced above. ALWAYS BEST TO SET UP YOUR PROBLEM SO THAT $p_1 = P(Y=1|X=1)$ (in upper left corner of table) is the largest probability you expect to find.

X (explanatory) = Gender, Y(response) = Watch Hockey

$P(Y=1|X=1) = p_1 = n_{11}/n_{1+} = 0.75$, $P(Y=1|X=2) = p_2 = n_{21}/n_{2+} = 0.2$

Wald $(1-\alpha)\%$ CI for $(p_1-p_2) = (p_1-p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \rightarrow$ 95% CI becomes (.38, .72)

0 is not in 95% CI. With 95% confidence, we have sufficient evidence that the proportion of males who watch hockey differs from the proportion of females who watch hockey. The proportion of males who watch hockey is between 38% and 72% higher than the proportion of females who watch.

Estimated relative risk = $\widehat{rr} = p_1/p_2 = 0.75/0.2 = 3.75$

95% CI for rr is (2.040, 6.893) (exponentiated endpoints of CI for ln(p1/p2))

1 is not in 95% CI. With 95% confidence, we have sufficient evidence that probability of being a hockey watcher is 2.040 to 6.893 times higher for males than females.

Estimated odds ratio = $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21} = (33X36)/(9x11) = 12$

95% CI for θ is (4.416, 32.606)(exponentiated endpoints of CI for $\ln(\hat{\theta})$)

1is not in 95% CI. With 95% confidence, sufficient evidence that a gender is associated with regular hockey watching.

We can also say: The odds of being a hockey watcher are 4.416 to 32.606 times higher for a male than a female.

GenX * WatY Crosstabulation

			WatY		Total
			1	2	
GenX	1	Count	33	11	44
		% within GenX	75.0%	25.0%	100.0%
	2	Count	9	36	45
		% within GenX	20.0%	80.0%	100.0%
Total		Count	42	47	89
		% within GenX	47.2%	52.8%	100.0%

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for GenX (1 / 2)	12.000	4.416	32.606
For cohort WatY = 1	3.750	2.040	6.893
For cohort WatY = 2	.313	.184	.532
N of Valid Cases	89		

Example: University Student Data: Coding

Gen (F-Female, M-Male)
Hsleep (hours sleep weeknight)
Energy (scale of 1 to 10)
HrHwkCrs (weekly)
MTxH (minutes text per hour)
Nsiblings (number of siblings)
NChPlan (number of children have/planned)
Nlang (number of spoken languages)
Age (in years)
HH3 (homework: low 0.5-1.5→1, med 2-3→2, high 3,5-5→3)
EN3 (energy: low 2-4→1, med5-7→2 , high 8+→3)
HH2 (homework: low 0-1.5→1, high 2+→2)
HS3 (sleep, low 4-5→1, med 6-7→2, high 8-9 →3)
HS2 (sleep →, low 4-6 →1, high 7-9→2)
NL2 (languages: unilingual -1, multilingual -2)
University Student Data (how to write up findings if you do not have significance

Is gender related to number of languages spoken?

Gender (rows) (1 - Female, 2 - Male)

NL2: Languages (columns)(1 - Unilingual, 2 - Multilingual)

Gender * NL2 Crosstabulation

			NL2		
			1.00	2.00	Total
Gender	F	Count	38	15	53
		% within Gender	71.7%	28.3%	100.0%
	M	Count	35	18	53
		% within Gender	66.0%	34.0%	100.0%
Total		Count	73	33	106
		% within Gender	68.9%	31.1%	100.0%

Risk Estimate

		95% Confidence Interval	
	Value	Lower	Upper
Odds Ratio for Gender (F / M)	1.303	.571	2.972
For cohort NL2 = 1.00	1.086	.840	1.403
For cohort NL2 = 2.00	.833	.471	1.473
N of Valid Cases	106		

X (explanatory) = Gender (1 (Female),2 (Male))
Y(response) = NL2(1 (unilingual), 2 (multilingual))
$P(Y=1|X=1) = p_1 = n_{11}/n_{1+} = 0.72$,
$P(Y=1|X=2) = p_2 = n_{21}/n_{2+} = 0.66$
Wald (1-α)% CI for (p_1-p_2) =

$$(p_1-p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}} \rightarrow 95\%\ \text{CI becomes } (-.13,.25)$$

0 is in 95% CI. For α = 0.05, we do not have sufficient evidence that the proportion of females who are unilingual differs from the proportion of males who are unilingual. The proportion of females who are unilingual is between 13% less to 25% higher than the portion of males who are unilingual.

Estimated relative risk = $\widehat{rr} = p_1/p_2 = 0.72/0.66 = 1.086$
95% CI for rr is (0.840,1.403) (exponentiated endpoints of CI for ln(p1/p2))
1 is in 95% CI. For α = 0.05, we do not have significant evidence that the probability of being unilingual is a significant number of times higher for females than males. The probability of being unilingual is between 1.190 (1/.840) times higher for a male to 1.403 times higher for a female.

Estimated odds ratio = $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21} = (38X18)/(35x15) = 1.303$
95% CI for θ is (0.571, 2.972) (exponentiated endpoints of CI for $\ln(\hat{\theta})$)
1 is in 95% CI. For α = 0.05, we do not have sufficient evidence that gender is associated with language plurality. The odds of being unilingual is between 1.751 (1/.571) times higher for a male than a female to 2.972 times higher for a female than a male.

Possible Useful SPSS Commands for transforming/editing data

Use Transform > Recode into Different Variables to get new recoded explanatory variable
Use Data > Select Cases to get subset worksheets
Use copy/paste to create column for new explanatory variable

Data: Class Examples Flu/Vaccine

EXAMPLE 17 and 18 DATA	No Vaccine (1)	One Shot (2)	Two Shots (3)	Total
Flu (1)	24	9	13	46
No Flu (2)	289	100	565	954
Total	313	109	578	1000

Input Data to SPSS: Flu (Response), Vaccine (Explanatory):

Here data is presented counter to the usual way, with the response variable on the rows and the Explanatory variable on the columns. Data is in fluvaccine.sav

Flu	Vaccine	Count
1	1	24
1	2	9
1	3	13
2	1	289
2	2	100
2	3	565

We illustrate how to create a cross-tab table for flu and vaccine, and perform a test of independence. Standardized residuals are calculated to identify cells of possible influence.

SPSS Commands
Data>Weight Cases
Weight by Frequency Variable: Count
OK
Analysis>Descriptive Statistics>Crosstabs
Row(s): Flu
Column(s): Vaccine
Statistics popup: ✓ Chi-square
Cells popup: Counts: ✓ Observed
✓ Expected
✓ Residuals: Adjusted standardized*
*Adjusted Standardized Residuals in SPSS are Standardized Cell Residuals in Class Notes

Output: Class Example Flu/Vaccine

Flu * Vacine Crosstabulation

			Vacine			Total
			1	2	3	
Flu	1	Count	24	9	13	46
		Expected Count	14.4	5.0	26.6	46.0
		Adjusted Residual	3.1	1.9	-4.2	
	2	Count	289	100	565	954
		Expected Count	298.6	104.0	551.4	954.0
		Adjusted Residual	-3.1	-1.9	4.2	
Total		Count	313	109	578	1000
		Expected Count	313.0	109.0	578.0	1000.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	17.313[a]	2	.000
Likelihood Ratio	17.252	2	.000
Linear-by-Linear Association	14.915	1	.000
N of Valid Cases	1000		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 5.01.

Test of Independence

Question: Is catching flu and vaccination status associated?

Ho: Flu and Vaccine independent vs Ha: Ho not true

Ho: Flu and Vaccine not associated vs Ha:Ho not true

$\alpha = 0.05$

Assume: random sample, all $\mu_{ij} >= 5$

Test statistic: (HIGHLIGHTED ON OUTPUT)

Pearson: 17.313

Likelihood Ratio: 17.252

Decision: Reject Ho (p-values are both <=.001)

There is significant evidence that flu and vaccination status are associated.

Adjusted standardized residuals >= 1.96 in absolute magnitude identify cells contributing to the significance.

SEE CELLS 11, 13, 31, AND 33 IN THE OUTPUT. Here people without the vaccine are more likely to get flu than people with 2 shots of the vaccine.

Tests of Independence: return to University Data

Question: Is hours of homework associated with energy level?

Choose your coding and make new SPSS columns

Column HH2: Low(0-1.5 → 1), High(2+ → 2)

Column EN3: Low (2-4 →1), Medium (5-7 →2), High (8+ →3)

Perform a test of independence in SPSS and write up your hypothesis test.

SPSS Commands

Analysis>Descriptive Statistics>Crosstabs

Row(s): HH2

Column(s): EN3

Statistics popup: ✓ Chi-square

Cells popup: Counts: ✓ Observed

✓ Expected

✓ Residuals: Adjusted standardized

HH2 * EN3 Crosstabulation

			EN3			Total
			1.00	2.00	3.00	
HH2	1.00	Count	8	23	16	47
		Expected Count	14.2	21.3	11.5	47.0
		% within HH2	17.0%	48.9%	34.0%	100.0%
		Adjusted Residual	-2.6	.7	2.0	
	2.00	Count	24	25	10	59
		Expected Count	17.8	26.7	14.5	59.0
		% within HH2	40.7%	42.4%	16.9%	100.0%
		Adjusted Residual	2.6	-.7	-2.0	
Total		Count	32	48	26	106
		Expected Count	32.0	48.0	26.0	106.0
		% within HH2	30.2%	45.3%	24.5%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.215[a]	2	.016
Likelihood Ratio	8.491	2	.014
Linear-by-Linear Association	7.911	1	.005
N of Valid Cases	106		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.53.

University Data: Test of Independence for Hours Homework and Energy Level

Ho: Homework and Energy independent vs Ha: Ho not true

Ho: Homework and Energy not associated vs Ha:Ho not true

$\alpha = 0.05$

Assume: random sample, all μ_{ij} >=5

Test statistic:
Pearson χ^2: 8.215
Likelihood Ratio: 8.491
Decision: Reject Ho (p-values are both <=.05)
There is significant evidence that Homework and Energy are associated.
The adjusted standardized residuals >= 1.96 in absolute magnitude identify cells contributing to the significance. SEE CELLS 11, 13, 21, AND 23 IN THE OUTPUT. Here people doing less homework are more likely to have more energy and people doing more homework are more likely to be less energetic.

Test of Linear Trend: Ordinal Variables (looking for linear trend)
Data: Class Example on tartar and smoking levels

By hand	Tartar			
Smoking	None	Middle	Heavy	Total
No	284	236	48	568
Middle	606	983	209	1798
Heavy	1028	1871	425	3324
Total	1918	3090	682	5690

IN SPSS (Smoketartar.sav)

Smoke	Tartar	Count	Smoke2	Smoke3
1.00	1.00	284.00	.0	284.50
1.00	2.00	236.00	.0	284.50
1.00	3.00	48.00	.0	284.50
2.00	1.00	606.00	2.00	1467.50
2.00	2.00	983.00	2.00	1467.50
2.00	3.00	209.00	2.00	1467.50
3.00	1.00	1028.00	3.00	4028.50
3.00	2.00	1871.00	3.00	4028.50
3.00	3.00	425.00	3.00	4028.50

Here Smoking (X) on the rows explains Tartar (Y) on the columns
Tarter has 3 levels (none (1), middle (2), heavy (3)
Various numerical scores can reflect distances between smoking levels (No, Middle, Heavy)
Smoke: Scores are 1, 2, 3 for none, middle, heavy smoking
Smoke 2: Scores are 0, 2, 3 for none, middle, heavy smoking
Smoke 3: Scores are midranks for none, middle, heavy smoking

SPSS Commands
Data>Weight Cases
Weight by Frequency Variable: Count
OK
Analysis>Descriptive Statistics>Crosstabs

Row(s): Smoke1, Smoke2, Smoke3 (3 POSSIBLE SCORES)
Column(s): Tartar
Statistics popup: ✓ Chi-square
✓ Correlation
✓ Kendall's Tau-b
Cells popup: Counts: ✓ Observed
Pearson's r changes depending on scores, but generally "gives similar results" (Agresti)
Spearman's R takes the fact that the data is ordinal into account.
Kendal's τ-b examines the relationship between concordant and discordant sample observations.
Spearman's R and Kendall's Tau-b do not change with different scores.

Smoke1

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.717[a]	4	.000
Likelihood Ratio	76.226	4	.000
Linear-by-Linear Association	51.213	1	.000
N of Valid Cases	5690		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 68.08.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.080	.012	6.437	.000
	Spearman Correlation	.086	.013	6.545	.000[c]
Interval by Interval	Pearson's R	.095	.013	7.188	.000[c]
N of Valid Cases		5690			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

Smoke2

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.717[a]	4	.000
Likelihood Ratio	76.226	4	.000
Linear-by-Linear Association	60.830	1	.000
N of Valid Cases	5690		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 68.08.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.080	.012	6.437	.000
	Spearman Correlation	.086	.013	6.545	.000[c]
Interval by Interval	Pearson's R	.103	.013	7.841	.000[c]
N of Valid Cases		5690			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

Smoke3

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	79.717[a]	4	.000
Likelihood Ratio	76.226	4	.000
Linear-by-Linear Association	39.719	1	.000
N of Valid Cases	5690		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 68.08.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.080	.012	6.437	.000
	Spearman Correlation	.086	.013	6.545	.000[c]
Interval by Interval	Pearson's R	.084	.013	6.324	.000[c]
N of Valid Cases		5690			

a. Not assuming the null hypothesis.
b. Using the asymptotic standard error assuming the null hypothesis.
c. Based on normal approximation.

Ho: $\rho = 0$ versus Ha: $\rho \neq 0$
Random Samples

	Test statistic: $M_o^2 = (n-1)r^2$, df = 1	Pvalue = $P(\chi^2 > M_o^2)$
Pearson's R	Smoke1: 0.095, Smoke2: 0.103, Smoke3: 0.084	< 0.001
Spearman	0.086	< 0.001
Kendall's τ – b	0.080	< 0.001

There is significant evidence that there is a linear trend association. The more a person smokes the more tartar builds up.

Three Way Table:Class Example (alienwallethreewaydata.sav)

This example considers 3 categorical variables, X, Y, and Z, where we can control Z by looking at partial tables for X and Y for each level of Z.
Y – response variable (Rating(score as 1,2,3,4,5)
X – explanatory variable (Movie: Alien (1) or Wall-e(2))
Z – confounding variable (Gender: Female (1) and Male (2)

The data in the data file is:

Movie	Freq	Rating	Sex	Score	Sex_n	Movie_n
Alien	430.00	12	f	1.00	1.00	1.00
Wall-e	1104.00	12	f	1.00	1.00	2.00
Alien	356.00	34	f	2.00	1.00	1.00
Wall-e	441.00	34	f	2.00	1.00	2.00
Alien	1313.00	56	f	3.00	1.00	1.00
Wall-e	1274.00	56	f	3.00	1.00	2.00
Alien	5075.00	78	f	4.00	1.00	1.00
Wall-e	5998.00	78	f	4.00	1.00	2.00
Alien	6777.00	910	f	5.00	1.00	1.00
Wall-e	19106.00	910	f	5.00	1.00	2.00
Alien	1502.00	12	m	1.00	2.00	1.00
Wall-e	5654.00	12	m	1.00	2.00	2.00
Alien	1404.00	34	m	2.00	2.00	1.00
Wall-e	2261.00	34	m	2.00	2.00	2.00
Alien	7090.00	56	m	3.00	2.00	1.00
Wall-e	8199.00	56	m	3.00	2.00	2.00
Alien	49158.00	78	m	4.00	2.00	1.00
Wall-e	43116.00	78	m	4.00	2.00	2.00
Alien	77097.00	910	m	5.00	2.00	1.00
Wall-e	94079.00	910	m	5.00	2.00	2.00

The tables are below

Male	Rating					
Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	1502	1404	7090	49158	77097	136251
Wall-e	5654	2261	8199	43116	94079	153309
Total	7156	3665	15289	92274	171176	289560
Female	Rating					

Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	430	356	1313	5075	6777	13951
Wall-e	1104	441	1274	5998	19106	27923
Total	1534	797	2587	11073	25883	41874
Both	Rating					
Movie	1-2	3-4	5-6	7-8	9-10	Total
Alien	1932	1760	8403	54233	83874	150202
Wall-e	6758	2702	9473	49114	113185	181232
Total	8690	4462	17876	103347	197059	331434

A χ^2 test for conditional independence (female and male) and correlation , r, to test for **conditional trend** (female and male) using the chosen scores 1, 2, 3, 4, 5 considers the data for each gender separately. SPSS commands to obtain useful output follow.

SPSS Commands

Data>Weight Cases
Weight by Frequency Variable: Freq
OK
Analysis>Descriptive Statistics>Crosstabs
Row(s): movie_n
Column(s): score
Layer 1 of 1:sex_n
Statistics popup: ✓ Chi-square ✓ Correlation
Cells popup: Counts: ✓ Observed

A χ^2 test for marginal independence and Correlation, r , to test **marginal trend** ((for chosen scores: 1,2,3,4,5) considers the data for both genders together. SPSS commands are as above, but no Layer(s) command is included.

Note the following:
1. Rows and Columns must be numerical in order for r to be calculated
2. Pearson's r will report the r corresponding to the scores chosen (here 1,2,3,4,5)
3. Spearman's r (based on mid-ranks scores)
4. Can calculate Chi-square statistic with Movie, Score, and Sex as string.

Output of interest is shown below. Necessary test statistic values and p-values are obtained from the output and summarized.

movie_n * score * sex Crosstabulation

sex				score					Total
				1.00	2.00	3.00	4.00	5.00	
f	movie_n	1.00	Count	430	356	1313	5075	6777	13951
			% within movie_n	3.1%	2.6%	9.4%	36.4%	48.6%	100.0%
			% within score	28.0%	44.7%	50.8%	45.8%	26.2%	33.3%
		2.00	Count	1104	441	1274	5998	19106	27923
			% within movie_n	4.0%	1.6%	4.6%	21.5%	68.4%	100.0%
			% within score	72.0%	55.3%	49.2%	54.2%	73.8%	66.7%
	Total		Count	1534	797	2587	11073	25883	41874
			% within movie_n	3.7%	1.9%	6.2%	26.4%	61.8%	100.0%
			% within score	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
m	movie_n	1.00	Count	1502	1404	7090	49158	77097	136251
			% within movie_n	1.1%	1.0%	5.2%	36.1%	56.6%	100.0%
			% within score	21.0%	38.3%	46.4%	53.3%	45.0%	47.1%
		2.00	Count	5654	2261	8199	43116	94079	153309
			% within movie_n	3.7%	1.5%	5.3%	28.1%	61.4%	100.0%
			% within score	79.0%	61.7%	53.6%	46.7%	55.0%	52.9%
	Total		Count	7156	3665	15289	92274	171176	289560
			% within movie_n	2.5%	1.3%	5.3%	31.9%	59.1%	100.0%
			% within score	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Chi-Square Tests

sex		Value	df	Asymp. Sig. (2-sided)
f	Pearson Chi-Square	1793.102[a]	4	.000
	Likelihood Ratio	1758.419	4	.000
	Linear-by-Linear Association	584.354	1	.000
	N of Valid Cases	41874		
m	Pearson Chi-Square	3778.475[b]	4	.000
	Likelihood Ratio	3927.323	4	.000
	Linear-by-Linear Association	160.043	1	.000
	N of Valid Cases	289560		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 265.53.

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1724.55.

Symmetric Measures

sex			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
f	Interval by Interval	Pearson's R	.118	.005	24.344	.000[c]
	Ordinal by Ordinal	Spearman Correlation	.181	.005	37.564	.000[c]
	N of Valid Cases		41874			
m	Interval by Interval	Pearson's R	-.024	.002	-12.654	.000[c]
	Ordinal by Ordinal	Spearman Correlation	.029	.002	15.343	.000[c]
	N of Valid Cases		289560			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

movie_n * score Crosstabulation

			score					Total
			1.00	2.00	3.00	4.00	5.00	
movie_n	1.00	Count	1932	1760	8403	54233	83874	150202
		% within movie_n	1.3%	1.2%	5.6%	36.1%	55.8%	100.0%
		% within score	22.2%	39.4%	47.0%	52.5%	42.6%	45.3%
	2.00	Count	6758	2702	9473	49114	113185	181232
		% within movie_n	3.7%	1.5%	5.2%	27.1%	62.5%	100.0%
		% within score	77.8%	60.6%	53.0%	47.5%	57.4%	54.7%
Total		Count	8690	4462	17876	103347	197059	331434
		% within movie_n	2.6%	1.3%	5.4%	31.2%	59.5%	100.0%
		% within score	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

movie_n * score Crosstabulation

			score					Total
			1.00	2.00	3.00	4.00	5.00	
movie_n	1.00	Count	1932	1760	8403	54233	83874	150202
		% within movie_n	1.3%	1.2%	5.6%	36.1%	55.8%	100.0%
		% within score	22.2%	39.4%	47.0%	52.5%	42.6%	45.3%
	2.00	Count	6758	2702	9473	49114	113185	181232
		% within movie_n	3.7%	1.5%	5.2%	27.1%	62.5%	100.0%
		% within score	77.8%	60.6%	53.0%	47.5%	57.4%	54.7%
Total		Count	8690	4462	17876	103347	197059	331434
		% within movie_n	2.6%	1.3%	5.4%	31.2%	59.5%	100.0%
		% within score	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Interval by Interval	Pearson's R	-.006	.002	-3.267	.001[c]
Ordinal by Ordinal	Spearman Correlation	.048	.002	27.571	.000[c]
N of Valid Cases		331434			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4692.375[a]	4	.000
Likelihood Ratio	4823.106	4	.000
Linear-by-Linear Association	10.671	1	.001
N of Valid Cases	331434		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 2022.13.

BY HAND SUMMARY	χ^2	Df	P
Conditional Independence (female)	1793.102	4	<.0001
Conditional Independence (male)	3778.475	4	<0.001
Marginal independence	4692.375	4	<0.001

BY HAND SUMMARY	r	M^2	df	P
Trend (conditional female)	0.118	583.03	1	<0.001
Trend (conditional male)	-0.024	166.79	1	<0.001
Marginal trend	-0.006	11.93	1	0.001

Test of Independence: Write-Up

There is significant evidence that rating and movie are dependent (associated) in both the partial table for males and the partial table for females …..so we say:

Rating and Movie are conditionally dependent, given gender. OR

There is a conditional association of movie and rating for males. OR

There is a conditional association of movie and rating for females.

There is significant evidence that rating and movie are dependent when we look at the marginal table (that ignores male and female and puts all the counts together).

We have marginal dependence (association)of rating and movie.

Test of Linear Association: Write-Up

There is significant evidence of a difference in median for the Alien and Wall-e groups when we condition on females. The positive sign on r indicates that Wall-e has a higher median rating than Alien for the female group. There is significant evidence of a difference in medians for the Alien and Wall-e groups when we condition on males. The negative sign on r indicates that Wall-e has a lower median rating than Alien for the male group. There is significant evidence of a difference in medians for the Alien and Wall-e groups when we look at the entire group of males and females together (the marginal result). The negative sign on r indicates that Wall-e has a lower median rating than Alien for the marginal situation.

Tests of Independence and Linear Trend University Data

Question: Is hours of homework associated with energy level? Confounding Variable is Gender!

Y: Response Variable: EN3:

Low (2-4 →1), Medium (5-7 →2), High (8+ →3)

X: Explanatory Variable: HH2:

Low(0-1.5 → 1), High(2+ → 2)

Z: Gender

(Female →1, Male →2)

Exercise: Perform a test of independence in SPSS. Write up your hypothesis test.

HH2 * EN3 * Gender Crosstabulation

Gender				EN3 1.00	EN3 2.00	EN3 3.00	Total
F	HH2	1.00	Count	6	7	5	18
			Expected Count	7.1	6.1	4.8	18.0
			Adjusted Residual	-.7	.5	.2	
		2.00	Count	15	11	9	35
			Expected Count	13.9	11.9	9.2	35.0
			Adjusted Residual	.7	-.5	-.2	
	Total		Count	21	18	14	53
			Expected Count	21.0	18.0	14.0	53.0
M	HH2	1.00	Count	2	16	11	29
			Expected Count	6.0	16.4	6.6	29.0
			Adjusted Residual	-2.7	-.2	2.9	
		2.00	Count	9	14	1	24
			Expected Count	5.0	13.6	5.4	24.0
			Adjusted Residual	2.7	.2	-2.9	
	Total		Count	11	30	12	53
			Expected Count	11.0	30.0	12.0	53.0

Chi-Square Tests

Gender		Value	df	Asymp. Sig. (2-sided)
F	Pearson Chi-Square	.486[a]	2	.784
	Likelihood Ratio	.490	2	.783
	Linear-by-Linear Association	.244	1	.622
	N of Valid Cases	53		
M	Pearson Chi-Square	12.561[b]	2	.002
	Likelihood Ratio	14.231	2	.001
	Linear-by-Linear Association	12.311	1	.000
	N of Valid Cases	53		

a. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 4.75.

b. 1 cells (16.7%) have expected count less than 5. The minimum expected count is 4.98.

Symmetric Measures

Gender			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
F	Ordinal by Ordinal	Kendall's tau-b	-.068	.128	-.533	.594
		Spearman Correlation	-.072	.135	-.517	.608[c]
	Interval by Interval	Pearson's R	-.068	.135	-.490	.626[c]
	N of Valid Cases		53			
M	Ordinal by Ordinal	Kendall's tau-b	-.464	.094	-4.526	.000
		Spearman Correlation	-.487	.099	-3.979	.000[c]
	Interval by Interval	Pearson's R	-.487	.099	-3.977	.000[c]
	N of Valid Cases		53			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

HH2 * EN3 Crosstabulation

			EN3 1.00	EN3 2.00	EN3 3.00	Total
HH2	1.00	Count	8	23	16	47
		Expected Count	14.2	21.3	11.5	47.0
		Adjusted Residual	-2.6	.7	2.0	
	2.00	Count	24	25	10	59
		Expected Count	17.8	26.7	14.5	59.0
		Adjusted Residual	2.6	-.7	-2.0	
Total		Count	32	48	26	106
		Expected Count	32.0	48.0	26.0	106.0

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	8.215[a]	2	.016
Likelihood Ratio	8.491	2	.014
Linear-by-Linear Association	7.911	1	.005
N of Valid Cases	106		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 11.53.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	-.261	.086	-3.022	.003
	Spearman Correlation	-.276	.091	-2.924	.004[c]
Interval by Interval	Pearson's R	-.274	.091	-2.911	.004[c]
N of Valid Cases		106			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

<u>Tests of Independence of Homework and Energy confounded by Gender</u>

There is significant evidence that homework level and energy level are dependent (associated) in both the partial table for males and the partial table for females …..<u>so we say:</u>

Homework and Energy and conditionally dependent, given gender.

There is a conditional association of homework and energy for males.

There is a conditional association of homework and energy for females.

There is significant evidence that homework and energy are dependent when we look at the marginal table (that ignores male and female and puts all the counts together).

We have marginal dependence (association)of homework and energy.

Test of Linear Association of Homework and Energy confounded by Gender

There is significant evidence of a difference in median for the low and high homework groups when we condition on females. The negative sign on r indicates that low homework performers have a higher median energy rating than high homework performers for the female group.

There is significant evidence of a difference in medians for the low and high homework groups when we condition on males. The negative sign on r indicates that low homework performers have a higher median energy rating than high homework performers for the male group.

There is significant evidence of a difference in medians for the low and high homework groups when we look at the entire group of males and females together (the marginal result). The negative sign on r indicates that low homework performers have a higher median energy rating than high homework performers for the marginal situation.

Generalized Linear Models (GLM)

1. Relate Explanatory and Response Variables
2. Models describe pattern of Association
3. Parameters describe nature/strength of Association
4. Test for Association: based on sample data while controlling for confounding variables
5. Can use models for Prediction

GLMs have both a random and a systemic component.

Random Component:This identifies that response variable Y and its distribution. The n observations Y1, Y2, …Yn are independent. The Ys have categorical outcomes.
Binary data: 2 possible outcomes (Success or Failure – coded 1 and 0 in raw data)
Example: Y = 1 or Y = 0 (Y is Bernouilli or Binomial with n =1)
Count data: Counts are outcomes
Example: Y = number of successes in a given time interval or a specified region of space (Y is Poisson)

Systematic Component:This specifies how explanatory variables are entered in linear way into model with a "linear predictor" $\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$

Link Functions:We will consider various link functions that link random and systematic components using a function g(μ), where μ is the mean of Y.

Model	E(Y) = μ is expressed in terms of observed x values and unknown parameters	Link Function
Linear ***EXAMPLE:*** ***Binary Y, X=x***	$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ $\boldsymbol{\mu = \pi(x) = \alpha + \beta x}$	Identity: g(μ) = μ
Logistic Regression ***EXAMPLE:*** ***Binary Y, X=x***	$\text{Log}(\mu/(1-\mu)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ $\boldsymbol{\mu = \pi(x) = \frac{exp(\alpha + \beta x)}{1+exp(\alpha + \beta x)}}$	Logit: g(μ) = log(μ/(1-μ))
Probit Regression **EXAMPLE:** ***Binary Y, X=x***	$\text{Probit}(\mu) = F^{-1}(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ where $\mu = P(Z <= \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) = F(\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$ for the standard normal cdf $\boldsymbol{\mu = \pi(x) = P(Z <= \alpha + \beta x)}$	Probit: g(μ) = inverse cumulative distribution of the standard normal distribution
Loglinear EXAMPLE: ***Poisson Y, X =x***	$\text{Log}(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ $\boldsymbol{\mu(x) = exp(\alpha + \beta x) = exp(\alpha)exp(\beta x)}$	Log: g(μ) = log(μ)

Exponential Family of Probability Distributions
These distributions can be written in a certain form, and include the normal, binomial, and poisson distribution.

Each of these distributions has a "natural" parameter upon which its probabilities depend
Normal: μ , μ = mean
Binomial: $\ln(\pi /1- \pi)$, π= success probability
Poisson: $\ln \mu$, μ = mean

The link function for "natural" parameter is called the "canonical" link.
We use Maximum Likelihood Estimators with GLMs.

Example: HSB (High School and Beyond): N = 600 US high school seniors
For the next example, we will use a sample of 600 high school seniors, taken from the
Large scale Longitudinal Study conducted by National Opinion Research Center (1980) under contract with National Center for Education Statistics. It includes the following variables.
ID Student Identification Number (3 digits)
SEX 1-Male 2-Female
RACE (Race or Ethnicity) 1-Hispanic 2-Asian 3 - Black 4- White
SES Socioeconomic Status 1-Low 2-Medium 3-High
SCTYP School Type 1-Public 2-Private
HSP High School Program 1-General 2-Academic Preparatory 3- Vocational/technical
LOCUS Locus of Control Standardized to mean 0 and st dev 1
CONCPT Self Concept Standardized to mean 0 and st dev 1
MOT Motivation Average of 3 motivation items
CAR Career Choice 1-Clerical 2-Craftsperson 3- Farmer 4-Homemaker 5-Laborer 6-Manager 7-Military 8-Operative 9-Professional[1], 10-Professional[2], 11-Proprietor 12-Protective 13-Sales 14-School 15-Service 16-Technical 17-Not working
RDG Reading T-score standardized to mean 50 and SD 10
WRTG Writing T-score standardized to mean 50 and SD 10
MATH Math T-score standardized to mean 50 and SD 10
SCI Science T-score standardized to mean 50 and SD 10
CIV Civics T-score standardized to mean 50 and SD 10
1-Accountatn, Nurse, Engineer, Librarian, Writer, Social Worker, Athlete, Politician, etc
2 – Clergy, Dentist, Physician, Lawyer, Scientist, College Teacher, etc

High School and Beyond (HSB Data)
Below we present SPSS commands to calculate the linear equation for our GLM of interest, and discuss and interpret results. In SPSS terminology, we "reference" the "non-academic" schools, as our level of most interest is "academic" schools, and we wish $\pi(x)$, the probability of success to be the probability of an academic school. We will look at two possible ways to input SPSS commands to obtain information of interest. Note that it is important to choose the correct reference category for our chosen dependent response variable.

Response: BINARY
Y: Non-Academic High School – 0 Academic High School – 1* OR
Program: Academic High School – 0 Non-Academic High School -1)

(*It is intuitive to think of higher number 1 being assigned to level of most interest where our $\pi(x)$ will be the probability of our "success". We will look at how SPSS handles this situation where the number 0 has been assigned to the level of most interest.)

Explanatory: (x = total of scores on 5 standardized achievement tests (Reading, Writing, Math, Science, Civics). This ranges from 164.7 to 350.

Random Component for Response Variable Y=program: Binomial
Systemic Component for Explanatory Variable X=x: Linear
Link Function: Identity $g(\mu) = \mu = (\alpha + \beta x)$

Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: Custom: **Distribution: Binomial** **Link Function: Identity** Response Tab: **Dependent Variable: y(change Binary reference category to first – lowest value)** (Academic High School – 1 or Non-Academic High School -0) Predictor Tab: x in the covariate box Model Tab: x as a main effect ✓ intercept in model All other tabs: left as defaults.	Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: Custom: **Distribution: Binomial** **Link Function: Identity** Response Tab: **Dependent Variable: program (default Binary reference category is last – highest value)** (Academic High School – 0 or Non-Academic High School -1) Predictor Tab: x in the covariate box Model Tab: x as a main effect ✓ intercept in model All other tabs: left as defaults.

Warnings

The maximum number of step-halvings was reached but the log-likelihood value cannot be further improved. Output for the last iteration is displayed.

The GENLIN procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

There is at least one invalid case in the last iteration. A case is invalid if there are errors in computing the inverse identity link function, the log-likelihood, the gradient, or the Hessian matrix in the iterative process. Only the iteration history is displayed.

Execution of this command stops.

Iteration History

Iteration	Update Type	Number of Step-halvings	Log Likelihood[a]	Parameter			Number of Invalid Cases[b]
				(Intercept)	x	(Scale)	
0	Initial	0	-327.270	-.968450	.005700	1	4
1	Scoring	0	-326.932	-.848214	.005253	1	
2	Newton	0	-326.789	-.897262	.005432	1	2
3	Newton	1	-326.788	-.893459	.005419	1	2
4	Newton	5	-326.788[c]	-.893459	.005419	1	2

Redundant parameters are not displayed. Their values are always zero in all iterations.
Dependent Variable: program
Model: (Intercept), x

a. The full log likelihood function is displayed.

b. These cases are invalid because there are errors in computing the inverse identity link function, the log-likelihood, the gradient, or the Hessian matrix in the iterative process. Invalid cases are excluded from iterations in which they cause computational errors.

c. The maximum number of step-halvings was reached but the log likelihood value cannot be further improved.

Intercept: Proportion of children in an academic high school with a total score of x=0 is -.893 (not sensible)
Slope: For every increase of 1 point in x, the probability of being in an academic high school increased by .0054

Discussion and Interpretation:

Problems with using the identity link function.

1. The model takes values over the entire real line, but probabilities are always between 0 and 1. But, with small or large enough x, $\alpha+\beta x$ will either be smaller than 0 or greater than 1.
2. ML estimators of α and β must be found by iterative methods as no formulas exist
3. The Fitting process fails when $\hat{\pi}(x) = \hat{\alpha} + \hat{\beta}x$ outside of 0 to 1 range

The fitted line we obtained was: $\hat{\pi}(x) = -0.893 + 0.0054x$

Note that these commands yield a column of $\hat{\pi}(x)$, and assign $\hat{\pi}(x) = 0$ for observation 133 (where x = 164.7 is smallest x) $\hat{\pi}(x) = 1$ for observation 84, (where x = 350 is largest x).

Logistic Regression Model

SPSS: Random Component for Response Variable Y=program: Binomial
Systemic Component for Explanatory Variable X=x: Linear
Link Function: Logit: $g(\mu) = ln(\frac{\mu}{1-\mu})$ = (α + βx)

Recall, for the Logit Link function,

$\text{Log}(\frac{\pi(x)}{1-\pi(x)}) = \alpha + \beta x$ OR $\pi(x) = \frac{\exp(a+\beta x)}{1+\exp(a+\beta x)}$

An advantage here is that $\pi(x) = \frac{\exp(a+\beta x)}{1+\exp(a+\beta x)}$ falls between 0 and 1 for all values of x.

SPSS COMMANDS FOR GLM APPROACH:

Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: Custom: Distribution: **Binomial** Link Function: **Logit** Response Tab: Dependent Variable: **y** **(change Binary reference category to first – lowest value)** (Academic High School – 1 or Non-Academic High School -0) Predictor Tab: x in the covariate box Model Tab: x as a main effect ✓ intercept in model All other tabs: left as defaults.	Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: Custom: Distribution: **Binomial** Link Function: **Logit** Response Tab: Dependent Variable: **program (default Binary reference category is last – highest value)** (Academic High School – 0 or Non-Academic High School - 1) Predictor Tab: x in the covariate box Model Tab: x as a main effect ✓ intercept in model All other tabs: left as defaults.

ONLY OUTPUT FOR THE DEPENDENT VARIABLE PROGRAM IS INCLUDED

Model Information

Dependent Variable	program[a]
Probability Distribution	Binomial
Link Function	Logit

a. The procedure models .00 as the response, treating 1.00 as the reference category.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	612.542	520	1.178
Scaled Deviance	612.542	520	
Pearson Chi-Square	530.969	520	1.021
Scaled Pearson Chi-Square	530.969	520	
Log Likelihood[a]	-325.509		
Akaike's Information Criterion (AIC)	655.019		
Finite Sample Corrected AIC (AICC)	655.039		
Bayesian Information Criterion (BIC)	663.813		
Consistent AIC (CAIC)	665.813		

Dependent Variable: program
Model: (Intercept), x

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Case Processing Summary

	N	Percent
Included	600	100.0%
Excluded	0	.0%
Total	600	100.0%

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
139.082	1	.000

Dependent Variable: program
Model: (Intercept), x

a. Compares the fitted model against the intercept-only model.

Categorical Variable Information

			N	Percent
Dependent Variable	program	.00	308	51.3%
		1.00	292	48.7%
		Total	600	100.0%

Continuous Variable Information

	N	Minimum	Maximum	Mean	Std. Deviation
Covariate x	600	134.70	360.00	259.3447	41.44849

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	103.104	1	.000
x	106.411	1	.000

Dependent Variable: program
Model: (Intercept), x

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-7.055	.6948	-8.417	-5.693	103.104	1	.000
x	.027	.0027	.022	.033	106.411	1	.000
(Scale)	1[a]						

Dependent Variable: program
Model: (Intercept), x

a. Fixed at the displayed value.

β determines the rate of increase or decrease of the S curve of the logistic regression model.

Logit $(\widehat{\pi(x)})$ = -7.055 + 0.027x is estimated equation

Interpretation: Since 0.027 is > 0, the estimated probability of attending an academic school increases as x (total score) increases.

SPSS COMMANDS FOR REGRESSION: BINOMIAL APPROACH

Random Component for Response Variable Y=program: Binomial

Systemic Component for Explanatory Variable X=x: Linear

Link Function: Logit: $g(\mu) = ln(\frac{\mu}{1-\mu})$ = (α + βx)

Analysis>Regression>Binary Logistic

Dependent: y

Covariate(s):x

Leave other defaults

Ok

<u>This automatically References</u> : (your lowest coded category) For y: "non-academic schools"

OUTPUT

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000
	Model	139.082	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1a	x	.027	.003	106.411	1	.000	1.028
	Constant	-7.055	.695	103.104	1	.000	.001

a. Variable(s) entered on step 1: x.

We can readily find the coefficients for the equation Logit ($\widehat{\pi(x)}$) = -7.055 + 0.027x above.

Probit Regression Model

Random Component for Response Variable Y=program: Binomial
Systemic Component for Explanatory Variable X=x: Linear
Link Function: Probit (F(α + βx)) = Probit (P(Z <= (α + βx))
Probit(π(x)) = α + βx where π(x) = left tail probability for standard normal distribution with z score α + βx OR
π(x)= P(Z <= α + βx)=F(α + βx) on -∞<x<∞ for standard normal distribution
Transforms π(x) so that the regression curve for π(x) has the appearance of the normal cdf
Advantage: π(x) falls between 0 and 1 for all x.

Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: Custom: Distribution: Binomial Link Function: Probit Response Tab: Dependent Variable: y (change Binary reference category to first – lowest value) (Academic High School – 1 or Non-Academic High School -0) Predictor Tab: x in the covariate box Model Tab: x as a main effect ✓ intercept in model All other tabs: left as defaults.	Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: Custom: Distribution: Binomial Link Function: Probit Response Tab: Dependent Variable: program (default Binary reference category is last – highest value) (Academic High School – 0 or Non-Academic High School -1) Predictor Tab: x in the covariate box Model Tab: x as a main effect ✓ intercept in model All other tabs: left as defaults.

<u>OUTPUT is for dependent variable program only</u>

Model Information

Dependent Variable	program[a]
Probability Distribution	Binomial
Link Function	Probit

a. The procedure models .00 as the response, treating 1.00 as the reference category.

Case Processing Summary

	N	Percent
Included	600	100.0%
Excluded	0	.0%
Total	600	100.0%

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
138.947	1	.000

Dependent Variable: program
Model: (Intercept), x

a. Compares the fitted model against the intercept-only model.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	612.677	520	1.178
Scaled Deviance	612.677	520	
Pearson Chi-Square	532.430	520	1.024
Scaled Pearson Chi-Square	532.430	520	
Log Likelihood[a]	-325.577		
Akaike's Information Criterion (AIC)	655.153		
Finite Sample Corrected AIC (AICC)	655.174		
Bayesian Information Criterion (BIC)	663.947		
Consistent AIC (CAIC)	665.947		

Dependent Variable: program
Model: (Intercept), x

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Continuous Variable Information

	N	Minimum	Maximum	Mean	Std. Deviation
Covariate x	600	164.70	350.00	259.9447	40.44849

Categorical Variable Information

			N	Percent
Dependent Variable	program	.00	308	51.3%
		1.00	292	48.7%
		Total	600	100.0%

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-4.252	.3932	-5.023	-3.482	116.965	1	.000
x	.017	.0015	.014	.019	121.093	1	.000
(Scale)	1[a]						

Dependent Variable: program
Model: (Intercept), x

a. Fixed at the displayed value.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	116.965	1	.000
x	121.093	1	.000

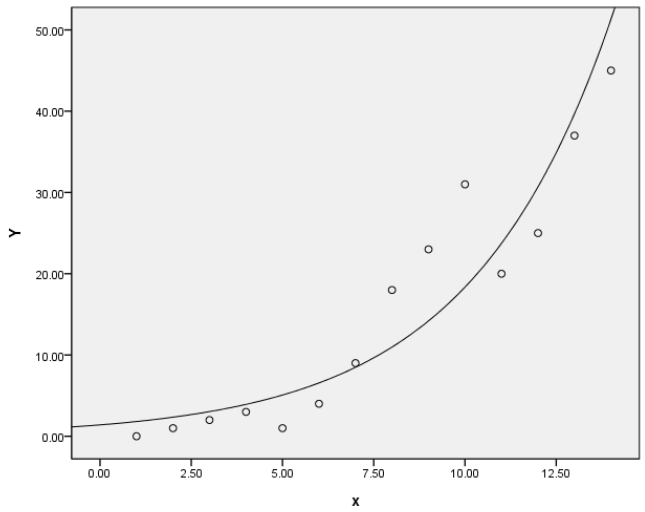
Dependent Variable: program
Model: (Intercept), x

β determines the rate of increase or decrease of the S curve of the probit regression model.
Probit($\widehat{\pi(x)}$) = -4.252 + 0.017x is estimated equation (see highlighted parameter estimate output on following output slide)
Interpretation: Since 0.017 is > 0, the higher the total score, the higher a probability is that a person attends an academic high school program.

Example 29: Aids/Australia
Model: Poisson Loglinear Model

- Data is number of deaths due to AIDS per 3 month period from Jan 1983 -Jun 1986
- x_i (Period #) y_i (# Deaths)
- 1.00 0.00
- 2.00 1.00
- 3.00 2.00
- 4.00 3.00
- 5.00 1.00
- 6.00 4.00
- 7.00 9.00
- 8.00 18.00
- 9.00 23.00
- 10.00 31.00
- 11.00 20.00
- 12.00 25.00
- 13.00 37.00
- 14.00 45.00

We make a scatterplot of the points.
COMMANDS
Graphs>LegacyDialogs>Scatterdot
Simple Scatter, Define
Y Axis Y ,X Axis x, OK



The scatterplot suggests a quadratic model.

Random Component for Response Variable Y= # deaths: Poisson
Systemic Component for Explanatory Variable X= x =time: Linear
Link Function: Ln(μ) = α + βx

COMMANDS

Analysis>Generalized Linear Models>Generalized Linear Models Type of Model Tab: – Custom: – Distribution: Poisson – Link Function: Log Response Tab: – Dependent Variable: Y Predictor Tab: – x in the covariate box Model Tab: – x as a main effect – include intercept in model All other tabs: left as defaults.	To put the equation on the scatterplot, Double click graph so pops up Right click on graph Select: Add Reference Line from Equation Custom Equation: Type exp(0.340 + 0.257*x) Apply $\hat{\mu}(x) = \exp(0.340 + 0.257x)$ is estimated equation and can be read from the output below

OUTPUT

Model Information

Dependent Variable	Y
Probability Distribution	Poisson
Link Function	Log

Case Processing Summary

	N	Percent
Included	14	100.0%
Excluded	0	.0%
Total	14	100.0%

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
177.619	1	.000

Dependent Variable: Y
Model: (Intercept), x

a. Compares the fitted model against the intercept-only model.

Continuous Variable Information

		N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variable	Y	14	.00	45.00	15.6429	14.98516
Covariate	x	14	1.00	14.00	7.5000	4.18330

Goodness of Fit[b]

	Value	df	Value/df
Deviance	29.654	12	2.471
Scaled Deviance	29.654	12	
Pearson Chi-Square	28.847	12	2.404
Scaled Pearson Chi-Square	28.847	12	
Log Likelihood[a]	-41.290		
Akaike's Information Criterion (AIC)	86.581		
Finite Sample Corrected AIC (AICC)	87.672		
Bayesian Information Criterion (BIC)	87.859		
Consistent AIC (CAIC)	89.859		

Dependent Variable: Y
Model: (Intercept), x

a. The full log likelihood function is displayed and used in computing information criteria.

b. Information criteria are in small-is-better form.

Tests of Model Effects

Source	Type III		
	Wald Chi-Square	df	Sig.
(Intercept)	1.828	1	.176
x	135.477	1	.000

Dependent Variable: Y
Model: (Intercept), x

Parameter Estimates

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test			Exp(B)	95% Wald Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	.340	.2512	-.153	.832	1.828	1	.176	1.404	.858	2.298
x	.257	.0220	.213	.300	135.477	1	.000	1.292	1.238	1.349
(Scale)	1[a]									

Dependent Variable: Y
Model: (Intercept), x

a. Fixed at the displayed value.

LOGLINEAR MODEL

$\text{Log}(\mu) = \alpha + \beta x$ OR $\mu = \exp(\alpha + \beta x) = e^{\alpha}e^{\beta x}$

A one unit increase in x has a multiplicative impact of e^{β} on μ. If $\beta > 0$, then $e^{\beta} > 1$, and μ increases as x increases. If $\beta < 0$, then μ decreases as x increases.

$\mu(x) = e^{0.340 + 0.257x}$ is estimated equation (from output)

Interpretation: The number of AIDS deaths in OZ over a 3 month period was in average $e^{0.257} = 1.29$ times higher than in the previous 3 month period.

Inference re Model Parameter β

Ho: $\beta = 0$ versus Ha: $\beta \neq 0$

Assumptions: Random samples large enough that ML estimators approximately normal

	Possible Test Statistics	P-value
Wald Z	$z_o = \hat{\beta}/SE(\hat{\beta})$ (is Z under Ho)	$2P(Z > \lvert z_o \rvert)$
Wald χ^2	$\chi_o^2 = z_o^2 = (\hat{\beta}/SE(\hat{\beta}))^2$, df=1 (is χ^2 under Ho)	$P(\chi^2 > \chi^2_o)$
Likelihood ratio	$-2\ln(l_0/l_1) = -2(L_0-L_1)$, df=1 (is χ^2 under Ho)	$P(\chi^2 > \chi^2_o)$

$\hat{\beta}$ = maximum likelihood estimate of β

l_0 = maximized value of likelihood function under Ho

l_1 = maximized value of likelihood function when β need not equal 0

$L_o = \ln(l_1)$ and $L_1 = \ln(l_0)$ are the maximized log-likelihood functions

From output:

Wald χ^2	$\chi_o^2 = 135.477$, df=1	$P(\chi^2 > \chi^2_o) < 0.001$

Reject Ho. Have sufficient evidence that time has an effect on the number of AIDS deaths in Oz.

Generalization: Compare Model of Interest M to Saturated Model

Ho:proposed model M fits as well as saturated model S vs Ha:Ho not correct

Ho:all parameters in model S that are not in model M = 0 (said another way)

Assumptions: Random samples, large samples

Rejection of Ho suggests that we need a full saturated model (model of "perfect fit" with one parameter per "measurement"). Being able to not reject Ho suggests that model M might suffice. Rejection of Ho means model M has significantly worse fit than model S. Therefore we hope test is **not** significant!

We test this with a:

GOODNESS OF FIT TEST BASED ON **DEVIANCE** = $-2(L_M-L_S)$

Likelihood ratio test (Deviance test)	Deviance$_0$ = -2(L_M-L_S)	For some GLMs*, test statistic is χ^2 with df=difference in number of parameters in models S and M when Ho is true *holds for Poisson loglinear, for example	P($\chi^2 > \chi^2_o$) when Ho is true
l_M = maximized value of likelihood function under proposed model M l_S = maximized value of likelihood function under saturated model S L_M = maximized log likelihood of model of interest and L_S = maximized log likelihood of saturated model			

Measures of Deviance/df "close" to 1 indicate good model fit.

Aids in Australia: Ho: model of interest fits as well as (full) saturated model (Poisson loglinear)

Goodness of Fit test on output	-2(L_M-L_S) = 29.654,df=14-2 = 12	Deviance/df = 29.654/12 = 2.471

We would reject Ho. The saturated model fits significantly better than the proposed model.

Generalization: Compare two models, M_0 is nested in M_1

Ho:model M_0 fits as well as model M_1 vs Ha:Ho not correct
Ho:all parameters in model M_1 but not in model M_0=0 (said another way)

Assumptions: Random samples, large samples

Likelihood ratio	-2(L_0-L_1) = Deviance$_0$ – Deviance$_1$ is χ^2 with df=difference in number of parameters in models M_0 and M_1	P($\chi^2 > \chi^2_o$) when Ho is true
L_o = maximized log likelihood of model Mo and L_1 = maximized log likelihood of Model M_1 Deviance$_0$ – Deviance$_1$ = -2(L_0-L_S) – 2(L_1-L_S) = -2(L_0-L_1)		

Rejection of Ho with a "relatively" large test statistic means and small p-value means that model M_o fits poorly in comparison to model M_1, and suggests the larger model is significantly more appropriate.

Aids in Australia: Ho:Intercept only model fits as well as model including time

Likelihood ratio	-2ln(l_0/l_1) = -2(L_0-L_1) = 177.619,df=2-1=1 (See Omnibus test on Output)	P($\chi^2 > \chi^2_o$) <0.001

There is sufficient evidence that the model including time fits better than the intercept only model.

Residuals

These are "standardized" differences between the observed and estimated values, and should be investigated when they exceed 2 in absolute value.

Pearson's Residuals $ei = \frac{yi - \hat{\mu}i}{\sqrt{\widehat{\mathrm{Var}}(yi)}}$

If you group the x variable (so it is categorical), these are approximately normal.
If you don't group the x variable, these are not approximately normal.

The Standardized Pearson's Residuals $ei = \frac{yi - \hat{\mu}i}{SE}$ are approximately normal, where
SE is estimated standard error of residual $yi - \hat{\mu}_i$.

Deviance Residuals measure how much individual measurements add to the deviance of a model.

Random Component for Response Variable Y=: Poisson
Systemic Component for Explanatory Variable X= Time: Linear
Link Function: Log

The following commands can be used to obtain the necessary residuals.

COMMANDS
Analysis>Generalized Linear Models>Generalized Linear Models
Type of Model Tab:
Counts: Poisson loglinear
Response Tab:
Dependent Variable: Y
Predictor Tab:
x in the covariate box
Model Tab:
x as a main effect
include intercept in model
Save Tab:
✓**Pearson residual**
✓ **Standardized Pearson residual**
✓ **Standardized deviance residual**
All other tabs:
left as defaults.

After save: the datasheet columns are as follows

X	Y	MeanPredicted	PearsonResidual	DevianceResidual	StdPearsonResidual
1.00	.0	1.82	-1.347	-1.905	-1.417
2.00	1.00	2.35	-.879	-.993	-.928
3.00	2.00	3.03	-.593	-.632	-.627
4.00	3.00	3.92	-.464	-.484	-.492
5.00	1.00	5.06	-1.806	-2.210	-1.916
6.00	4.00	6.55	-.995	-1.073	-1.054
7.00	9.00	8.46	.186	.184	.196
8.00	18.00	10.93	2.137	1.953	2.249
9.00	23.00	14.13	2.359	2.161	2.475
10.00	31.00	18.26	2.980	2.708	3.128
11.00	20.00	23.60	-.742	-.762	-.785
12.00	25.00	30.51	-.997	-1.029	-1.084
13.00	37.00	39.43	-.386	-.391	-.449

14.00	45.00	50.96	-.834	-.851	-1.134

NOTE LARGER RESIDUALS AT x = 8, 9, AND 10. At that time, it appears that other factors were involved.

The HSB data will be used to create several GLMs for investigation. Commands, output, and interpretations will be offered.

The following commands allow for the creation of a logit expression when Y is the response variable and the explanatory variable is X, the academic score.

Example 35: **(Using Regression path)**
NOTE: DEFAULT SUCCESS PROBABILITY IS THE ONE WITH THE HIGHEST LEVEL IN THE RESPONSE VARIABLE (ACADEMIC)
ANALYZE>REGRESSION>BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move x to Covariate box
Covariate box will read
x
-Method: Enter
(Use Save and Options buttons as needed)
-Ok

Response: Y – (0-non-academic, 1-academic)
Explanatory:
X – academic score (total of reading, writing, math, science, civics)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x	.027	.003	106.411	1	.000	1.028
	Constant	-7.055	.695	103.104	1	.000	.001

a. Variable(s) entered on step 1: x.

$\pi(x)$, success prob is prob of attending academic school
Equation: $\text{logit}(\widehat{\pi(x)}) = -7.055 + 0.027x$
Slope: a 1 point increase in the total test score x multiplies the odds of a student being in an academic program by a factor of $e^{0.027} = 1.02$. The odds of being in an academic program increase by 2% for each unit increase in x.

The following commands allow information necessary for the calculation of the Wald Z, Wald χ^2, and the Likelihood Ratio test statisics to be obtained, in addition to the logit expression, and the predicted $\pi(x)$ values.

Example 35: **Commands for Inference Results**:
Response: Y – (0-non-academic, 1-academic)
Explanatory: X – academic score (total of reading, writing, math, science, civics)
NOTE: DEFAULT SUCCESS PROBABILITY IS THE ONE WITH THE HIGHEST LEVEL IN THE RESPONSE VARIABLE (ACADEMIC)
ANALYZE>REGRESSION>BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move x to Covariate box
Covariate box will read x
-Method: Enter
(Use Save and Options buttons as shown below)
-Ok
Option box: Check "CI for exp(B)" – will provide 95% CI for e^{β} in the Variables in the Equation box in the output file
Save Box: Check "Probabilities" – will provide a column of predicted values for $\pi(x)$ in the data file

To obtain predicted values for $\pi(x)$ when your x of interest is not one of the x values in the data, add the x value of interest to the bottom of the x column, in the row below the last row of data.

Example: For x = 260, SPSS provides 0.5162 as predicted value for $\pi(260)$ (in the data file).

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000
	Model	139.082	1	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	x	.027	.003	106.411	1	.000	1.028	1.022	1.033
	Constant	-7.055	.695	103.104	1	.000	.001		

a. Variable(s) entered on step 1: x.

Example 35 (commands for inference results) (continued)
We can use the above information to perform inference:
Tests of significance for the model: logit($\pi(x) = \alpha + \beta x$):
Ho: $\beta = 0$ versus Ha: $\beta \neq 0$, choose $\alpha = 5\%$
Assumptions: Random samples large enough that ML estimators approximately normal
$\hat{\beta}$ = maximum likelihood estimate of β

coloured information is from output above

	Test Statistic	P-value
Wald Z	$z_o = \hat{\beta}/SE(\hat{\beta})$ (is Z under Ho) $= \sqrt{106.411} = 10.32$	$2P(Z > \lvert z_o \rvert)$ <0.001 (from tables)
Wald χ^2	$\chi_o^2 = z_o^2 = (\hat{\beta}/SE(\hat{\beta}))2$, df=1 (is χ^2 under Ho) =106.411, df =1	$P(\chi^2 > \chi^2_o)$ <0.001
Likelihood ratio	$-2\ln(l_0/l_1) = -2(L_0-L_1)$, df=1 (is χ^2 under Ho) = 139.082, df =1	$P(\chi^2 > \chi^2_o)$ <0.001

Reject Ho
At a significance level of 5%, we have sufficient evidence that the odds of being in an academic program are related to the test results. (That is; we have sufficient evidence to support the model M_1 .)

Wald CI for β: 0.027 +/- 1.96(0.003) = (.022,.033) (by hand using SPSS output information)

Wald CI for e^{β} = (1.022, 1.033)

Example 35: *Additional inference output needed for **confidence intervals for predicted values of $\pi(x)$** can be found when one uses the **Generalized Linear Model** path to perform your binary logistic regression.*
Response: Y – (0-non-academic, 1-academic)
Explanatory:
X – academic score (total of reading, writing, math, science, civics)

Analysis>Generalized Linear Models>Generalized Linear Models
Type of Model Tab: Choose "Binary Logistic" button
Response Tab: Select y for the Dependent Variable *Binary button popup box: Change reference category to first (lowest value). This will ensure that your $\pi(x)$ probability of success will be the probability of an academic school.*
Predictors Tab: Select x for the Covariates box
Model Tab: Choose Type "Main Effects" and move the x variable into the Model box
Estimation Tab: Leave as default
Statistics Tab: Leave all defaults AND also check "Include exponential parameter estimates" and "Covariance matrix for parameter estimate". (Also note that you can change the level of confidence of your CIs here, and that the defaults for the inference tests and the confidence interval types are Wald.)
EM Means Tab: Leave as default
Save Tab: Check "Predicted value of mean of response, "Lower bound of confidence interval for mean of response", "Upper bound of confidence interval for mean of response", "Predicted Value of linear predictor", "Estimated standard error of predicted value of linear predictor".

Click OK

Model Information

Dependent Variable	y[a]
Probability Distribution	Binomial
Link Function	Logit

a. The procedure models 1.00 as the response, treating .00 as the reference category.

Goodness of Fit[b]

	Value	df	Value/df
Deviance	612.542	520	1.178
Scaled Deviance	612.542	520	
Pearson Chi-Square	530.969	520	1.021
Scaled Pearson Chi-Square	530.969	520	
Log Likelihood[a]	-325.509		
Akaike's Information Criterion (AIC)	655.019		
Finite Sample Corrected AIC (AICC)	655.039		
Bayesian Information Criterion (BIC)	663.813		
Consistent AIC (CAIC)	665.813		

Dependent Variable: y
Model: (Intercept), x
a. The full log likelihood function is displayed and used in computing information criteria.
b. Information criteria are in small-is-better form.

Omnibus Test[a]

Likelihood Ratio Chi-Square	df	Sig.
139.082	1	.000

Dependent Variable: y
Model: (Intercept), x
a. Compares the fitted model against the intercept-only model.

Tests of Model Effects

	Type III		
Source	Wald Chi-Square	df	Sig.
(Intercept)	103.104	1	.000
x	106.411	1	.000

Dependent Variable: y
Model: (Intercept), x

Parameter Estimates

			95% Wald Confidence Interval		Hypothesis Test				95% Wald Confidence Interval for Exp(B)	
Parameter	B	Std. Error	Lower	Upper	Wald Chi-Square	df	Sig.	Exp(B)	Lower	Upper
(Intercept)	-7.055	.6948	-8.417	-5.693	103.104	1	.000	.001	.000	.003
x	.027	.0027	.022	.033	106.411	1	.000	1.028	1.022	1.033
(Scale)	1[a]									

Dependent Variable: y
Model: (Intercept), x
a. Fixed at the displayed value.

Profile Likelihood Confidence Intervals

It appears that SPSS can calculate Profile Likelihood Confidence Intervals. You tell it on the Statistics tab to use a "Likelihood ratio" Chi-square statistics and a "Profile likelihood" Confidence Interval Type. It produces the output below. I'm not convinced it is really calculating the "Profile Likelihood" confidence intervals for β, given that they appear close to the "Wald" confidence intervals. However, the confidence intervals for the intercept are different, and if you look at further decimal places for the CI for β, they do differ. And when I ran some other models with more explanatory variables I did get some larger differences in CIs for various parameters for the two types of confidence intervals.

Output

Tests of Model Effects

Source	Type III		
	Likelihood Ratio Chi-Square	df	Sig.
(Intercept)	133.838	1	.000
x	139.082	1	.000

Dependent Variable: y
Model: (Intercept), x

Parameter Estimates

Parameter	B	Std. Error	95% Profile Likelihood Confidence Interval		Hypothesis Test			Exp(B)	95% Profile Likelihood Confidence Interval for Exp(B)	
			Lower	Upper	Wald Chi-Square	df	Sig.		Lower	Upper
(Intercept)	-7.055	.6948	-8.457	-5.730	103.104	1	.000	.001	.000	.003
x	.027	.0027	.022	.033	106.411	1	.000	1.028	1.023	1.033
(Scale)	1[a]									

Dependent Variable: y
Model: (Intercept), x
a. Fixed at the displayed value.

Covariances of Parameter Estimates

	(Intercept)	x
(Intercept)	.48271	-.00183
x	-.00183	7.04654E-6

Profile Likelihood CI for β: (.022,.033)
Profile Likelihood CI for e^{β} = (1.022, 1.033)

Confidence Intervals for the Probabilities, π(x)

From Data File (because of Save tab):

New Columns Appear (this is row 601, which has predicted value information for π(260)

XBPredicted	SBStandardError	MeanPredicted	CIMeanPredictedLower	CIMeanPredictedUpper
0.065	0.092	0.516	0.471	0.561

95% CI for logit(π(260)) is 0.065 +/- 1.96(0.092) = (-0.115,0.244) (by hand, using computer information)
Exponentiating, by hand, on the value of 0.065 gives .516 (for $\widehat{\pi(260)}$)
Exponentiating, by hand gives (0.471,0.561) as a 95% CI for π(260))

$\widehat{\pi(260)}$ = 0.516 (from output)
95% CI for π(260) is (0.471, 0.561) (from output)

SO WE DO HAVE A MATCH:

ROUNDING PROBLEMS WHEN USING COMPUTER OUTPUT: BEWARE:

A suitable formula for calculating a CI for logit(π(x)) is:

logit($\widehat{\pi(x)}$) ± 1.96$\sqrt{SE^2}$, where SE^2 = $\widehat{Var(\alpha)}$ + $x^2\widehat{Var(\beta)}$ + 2x $\widehat{Cov(\hat{\alpha},\hat{\beta})}$

If you calculate logit($\widehat{\pi(260)}$) using the ROUNDED formula -7.055 + 0.027 provided in the SPSS output, you will obtain -0.035 as your logit($\widehat{\pi(260)}$) If you calculate logit($\widehat{\pi(260)}$) using 12 decimal places (which you can obtain by popping up the boxes from the SPSS output), you will obtain 0.065 as your logit(π(260)).

YOU WOULD NOT WANT TO USE THIS INCORRECT VALUE OF -0.35 IN CALCULATING YOUR CI BY HAND.

FURTHERMORE, IF YOU USE THE COVARIANCE OF PARAMETER ESTIMATES TABLE PROVIDED BY THE COMPUTER OUTPUT (SEE ABOVE) AND SUBSTITUTE THESE VALUES INTO THE EQUATION ABOVE, FURTHER ROUNDING ERROR WILL OCCUR AND YOUR INTERVAL WIDTH WILL BE INCORRECT.

HOWEVER, IF YOU CARRY 12 DECIMAL PLACES THROUGHOUT ALL YOUR BY HAND WORK, YOUR BY HAND ANSWERS WILL MATCH THE COMPUTER OUTPUT.

IT'S NICE WHEN COMPUTERS DO THINGS FOR US.

THIS PROBLEM OCCURS DUE TO THE SMALL MAGNITUDE OF THE logit($\widehat{\pi(260)}$)

Example 37: NOTE: DEFAULT INDICATOR REFERENCE CATEGORY FOR EXPLANATORY VARIABLES IS HIGHER NUMBERED LEVEL ANALYZE>REGRESSION> BINARY LOGISTIC Logistic Regression Screen -Move y to Dependent box	Response: Y – (0-non-academic, 1-academic) Explanatory: SES (1-low, 2-middle, 3-high) (express SES with 2 dummy variables) d_1 = 0 – not low, 1 – low d_2 = 0 – not med, 1 - medium **Model Summary**

-Move SES to Covariate box
Covariate box will read
SES
-Method: Enter
-Click Categorical Button
Covariate Screen
-Move SES from Covariate box to categorical Covariates box
Categorical Covariate box will read
SES(indicator)
-Continue
Logistic Regression Screen
Covariate box will read
SES(cat)
(Use Save and Options buttons as needed)
-Ok

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	779.390[a]	.083	.111

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	ses			47.358	2	.000	
	ses(1)	-1.725	.253	46.488	1	.000	.178
	ses(2)	-.989	.210	22.152	1	.000	.372
	Constant	.956	.175	29.673	1	.000	2.600

a. Variable(s) entered on step 1: ses.

π(x), success prob is prob of attending academic school
Equation: logit($\widehat{\pi(x)}$)= .956 – 1.725 d_1 -.989d_2
SES Odds Ratio: oddslow/oddshigh = $e^{-1.725}$ = .1781
(indicator reference category is high)
(.1781 times higher for low income than high income)
(1/.1781 = 5.61X higher for high income than low income)
SES Odds Ratio: oddsmed/oddshigh = $e^{-.989}$ = .3719
(indicator reference category is high)
((.3719 times higher for medium income than high income)
(1/.3719 = 2.6889X higher for high income than medium income)

Example 38:
ANALYZE>REGRESSION>
BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move Gender and School to Covariate box
Covariate box will read
Gender
School
-Method: Enter
-Click Categorical Button
Covariate Screen
-Move Gender and School from Covariate box to categorical

Response: Y – (0-non-academic, 1-academic)
Explanatory:
Gender (0-female, 1-male)
School (0-public, 1 – private)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	792.754[a]	.062	.083

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender(1)	-.070	.169	.173	1	.678	.932

Covariates box
Categorical Covariate box will read
Gender(indicator)
School(indicator)
-Continue
Logistic Regression Screen
Covariate box will read
Gender(cat)
School (cat)
(Use Save and Options buttons as needed)
-Ok

school(1)	-1.533	.272	31.791	1	.000	.216
Constant	1.413	.275	26.476	1	.000	4.109

a. Variable(s) entered on step 1: gender, school.

π(x), success prob is prob of attending academic school

Equation: logit($\widehat{\boldsymbol{\pi}(\boldsymbol{x})}$)= = 1.413 – 0.70gender – 1.533school

Gender Odds Ratio: oddsfemale/oddsmale = $e^{-0.70}$ = 1.0725

(indicator reference category male)

(odds 1.0275 times higher to attend academic school for females than males, other variables held constant)

School Odds Ratio: oddspublic/oddsprivate= $e^{-1.533}$ = 0.2159

(indicator reference category private)

(odds .2159 times higher to attend an academic school for public schools than private schools, other variables held constant)

(odds 4.6321 times higher to attend an academic school for private schools than public schools, other variables held constant)

Example 38: (2nd suggested approach)
(Note: set indicator reference on explanatory variable appropriately)
ANALYZE>REGRESSION>BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move Gender and School to Covariate box
Covariate box will read
Gender
School
-Method: Enter
-Click Categorical Button
Covariate Screen
-Move Gender and School from Covariate box to categorical Covariates box
-Highlight School
-Change to Indicator(first)
Categorical Covariate box will read
Gender(indicator)
School(indicator(first))
-Continue
Logistic Regression Screen
Covariate box will read
Gender(cat)
School (cat)
(Use Save and Options buttons as needed)
-Ok

Response: Y – (0-non-academic, 1-academic)
Explanatory:
Gender (0-female, 1-male)
School (0-public, 1 – private)

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	792.754[a]	.062	.083

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	gender(1)	-.070	.169	.173	1	.678	.932
	school(1)	1.533	.272	31.791	1	.000	4.634
	Constant	-.120	.128	.889	1	.346	.887

a. Variable(s) entered on step 1: gender, school.

π(x), success prob is prob of attending academic school

Equation: logit($\widehat{\boldsymbol{\pi}(\boldsymbol{x})}$)= -.120 – 0.70gender + 1.533school

Gender Odds Ratio: oddsfemale/oddsmale = $e^{-0.70}$ = 1.0725

(indicator reference category male)

(odds 1.0275 times higher to attend academic school for females than males)

School Odds Ratio: oddsprivate/oddspublic

(indicator reference category public)

School Odds Ratio: oddsprivate/oddspublic = $e^{1.533}$ = 4.6321

(odds 4.6321 times higher to attend an academic school for private schools than public schools, other variables held constant)

Example 5: From notes:
Model Mo: $\beta_0 + \beta_1 x + \beta_2$ school $+\beta_3 d_1 + \beta_4 d_2$
where we use two dummy variables to characterize the SES variable.

1. Ho: being in an academic program is independent from SES when controlling for academic score and school type (i.e. Ho: $\beta_3 = \beta_4 = 0$) versus Ha: Ho is not true, $\alpha = 0.05$

2. Assumptions are met

3. Test Statistic $\chi_o^2 = -2(L_o - L_1) = -2(-332.058 - -325.980) = -2(-332.058 + 325.980) = 12.156$, df = 5 -3 = 2 (from following output)

4. P-valule = $P(\chi^2 > 12.156) < 0.005$ (by table) (from following output)

5. Reject Ho since the p-value is $< \alpha$.

6. At a significance level of 5%, the data suggests that SES is associated with the chance of a student being in an academic program even when controlling for academic score and school type.

L_0 = log likelihood under the Ho model
L_1 = log likelihood under the H_1 model

df = difference in the number of parameters of the two models

There are many ways to find L_o and L_1 and the Test Statistic χ_o^2 in SPSS. We will examine several.

Characteristics of the Binary Regression approach:
●It only allows a lowest value reference category on the response variable.
●It allows for a switch in reference category on any explanatory variable.
●It allows for a stepwise entry of parameters.
●It allows for the saving of predicted probabilities and residuals.

Characteristics of the GLM approach:
●It allows for the binary reference category to be switched on the response variable.
●It only allows for all high value reference category (or all low value) on all explanatory variables.
●It allows for the saving of predicted probabilities, confidence intervals, and residuals.
And much more!

Note that the 12.147 in the examples below is the number you get when you use L_o and L_1 to more decimal places. So the 12.147 corresponds to the 12.156 from the classroom notes.

Example 5 REGRESSION PATH: MODEL L_0 (NO SES). METHOD ENTER

ANALYZE>REGRESSION>BINARY LOGISTIC <u>Logistic Regression Screen</u> -Move y to Dependent box -Move x, School to Covariate box	*Response: Y – (0 – nonacademic, 1 – academic) Explanatory: X – academic score (total of reading, writing, math, science, civics) School (0 – public, 1 – private)

Covariate box will read
x
School
-Method: Enter
-Click Categorical Button
Covariate Screen
-Move School from Covariate box to categorical Covariates box
Categorical Covariate box will read
School(indicator)
-Continue
Logistic Regression Screen
Covariate box will read
x
School (cat)
(Use Save and Options buttons as needed)
-Ok

$-2L_0 = 664.107$
$L_0 = -332.058$

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	664.107[a]	.243	.324

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x	.027	.003	99.251	1	.000	1.027
	school(1)	-1.421	.288	24.337	1	.000	.241
	Constant	-5.738	.747	58.952	1	.000	.003

a. Variable(s) entered on step 1: x, school.

$\pi(x)$, success prob is prob of attending academic school

Equation: $\widehat{\text{logit}(\pi(x))} = -5.738 + 0.027x - 1.421\text{school}$

Total Test Score: A 1 point increase in the total test score x multiplies the odds of a student being in an academic program by a factor of $e^{0.027} = 1.027$, other variables held constant. The odds of being in an academic program increase by 2.7% for each unit increase in x, other variables held constant.

School Odds Ratio: oddspublic/oddsprivate = $e^{-1.421} = .2415$
(indicator reference category private)
(odds .2415 times higher to attend academic school for public schools than private schools, other variables held constant)
(odds 1/.2415 = 4.1413 times higher to attend academic school for private schools than public schools, other variables held constant)

Example 5: REGRESSION PATH: MODEL L_1 (INCLUDES SES), METHOD ENTER

ANALYZE>REGRESSION>
BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move x, School, SES to Covariate box
Covariate box will read
x
School
SES
-Method: Enter
-Click Categorical Button
Covariate Screen
-Move School, SES from Covariate box to categorical Covariates box
Categorical Covariate box will read
School(indicator)
SES (indicator)

*Response: Y – (0 – nonacademic, 1 – academic)
Explanatory:X – academic score (total of reading, writing, math, science, civics)
School (0 – public, 1 – private)
SES (1-low, 2-middle, 3-high)
(categorize SES as 2 dummies)
$d_1 = 0$ – not low, 1 – low
$d_2 = 0$ – not med, 1 – medium

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	651.960[a]	.258	.345

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)

-Continue
Logistic Regression Screen
Covariate box will read
x
School (cat)
SES (cat)
(Use Save and Options buttons as needed)
-Ok
$-2L_1 = 651.960$, $L_1 = -325.980$
SES Odds Ratio: oddslow/oddshigh = $e^{-0.891} = 0.4102$
(indicator reference category is high)
(odds 0.4102 times higher to attend academic school for low income than high income, other variables held constant)
(odds 1/0.4102 = 2.4378 times higher to attend academic school for high income than low income, other variables held constant)
SES Odds Ratio: oddsmed/oddshigh = $e^{-.706} = .4936$
(indicator reference category is high)
(odds 0.4936 times higher to attend academic school for medium income than high income, other variables constant)
(odds 1/0.4102 =2.0259 times higher to attend academic school for high income than low income, other variables constant)

Step 1[a]	x	.025	.003	80.320	1	.000	1.025
	school(1)	-1.346	.291	21.369	1	.000	.260
	ses			11.890	2	.003	
	ses(1)	-.891	.285	9.756	1	.002	.410
	ses(2)	-.706	.235	8.978	1	.003	.494
	Constant	-4.722	.801	34.798	1	.000	.009

a. Variable(s) entered on step 1: x, school, ses.

$\pi(x)$, success prob is prob of attending academic school
Equation: $\text{logit}(\widehat{\pi(x)})$=-4.722+0.025x-1.346school-0.891d1-0.706d2
Total Test Score: A 1 point increase in the total test score x multiplies the odds of a student being in an academic program by a factor of $e^{0.025} = 1.025$, other variables held constant. The odds of being in an academic program increase by 2.5% for each unit increase in x, other variables held constant.
School Odds Ratio: oddspublic/oddsprivate = $e^{-1.346} = .2603$
(indicator reference category private)
(odds .2603 times higher to attend academic school for public schools than private schools, other variables held constant)
(odds 1/.2603 = 3.8417 times higher to attend academic school for private schools than public schools, other variables held constant)

There is another approach for obtaining the necessary L_o and L_1 for this problem.
Example 5: REGRESSION PATH: ALL PARAMETERS, METHOD FORWARD (LR)

ANALYZE>REGRESSION>BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move x, School, and SES to Covariate box
Covariate box will read
x
School
SES
-Method: Forward (LR)
-Click Categorical Button
Covariate Screen
-Move School, SES from Covariate box to categorical Covariates box
Categorical Covariate box will read
School(indicator)
SES (indicator)
-Continue
Logistic Regression Screen

*Response: Y – (0 – nonacademic, 1 – academic)
Explanatory: X – academic score (total of reading, writing, math, science, civics)
School (0 – public, 1 – private)
SES (1-low, 2-middle, 3-high)
(SES characterized with 2 dummy variables)
$d_1 = 0$ – not low, 1 – low
$d_2 = 0$ – not med, 1 – medium

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276
2	664.107[a]	.243	.324
3	651.960[b]	.258	.345

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000

Covariate box will read
x
School (cat)
SES (cat)
(Use Save and Options buttons as needed)
-Ok

$-2L_0 = 664.107$
$L_0 = -332.058$

$-2L_1 = 651.960$
$L_1 = -325.980$

Step 3: Chi-square 12.147

	Model	139.082	1	.000
Step 2	Step	28.161	1	.000
	Block	167.243	2	.000
	Model	167.243	2	.000
Step 3	Step	12.147	2	.002
	Block	179.390	4	.000
	Model	179.390	4	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	x	.027	.003	106.411	1	.000	1.028
	Constant	-7.055	.695	103.104	1	.000	.001
Step 2[b]	x	.027	.003	99.251	1	.000	1.027
	school(1)	-1.421	.288	24.337	1	.000	.241
	Constant	-5.738	.747	58.952	1	.000	.003
Step 3[c]	x	.025	.003	80.320	1	.000	1.025
	school(1)	-1.346	.291	21.369	1	.000	.260
	ses			11.890	2	.003	
	ses(1)	-.891	.285	9.756	1	.002	.410
	ses(2)	-.706	.235	8.978	1	.003	.494
	Constant	-4.722	.801	34.798	1	.000	.009

a. Variable(s) entered on step 1: x.

b. Variable(s) entered on step 2: school.

c. Variable(s) entered on step 3: ses.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	x	-415.675	139.082	1	.000
Step 2	x	-396.464	128.820	1	.000
	school	-346.134	28.161	1	.000
Step 3	x	-375.324	98.688	1	.000
	school	-338.157	24.354	1	.000
	ses	-332.053	12.147	2	.002

$\pi(x)$, success prob is prob of attending academic school

Equation: $\text{logit}(\widehat{\pi(x)}) = -4.722 + 0.025x - 1.346\text{school} - 0.891d1 - 0.706d2$

Example 5 GLM PATH: MODEL L_O (NO SES) LOG LIKELIHOOD KERNAL (to obtain L_0)

Analysis>Generalized Linear Models>Generalized Linear Models
Type of Model Tab:
Binary Logistic
Response Tab:
Dependent Variable: y
(change Binary reference category to first – lowest value)
Predictor Tab:
x in the Covariate box
school in Factors box
Model Tab:
x as a main effect
school as main effect
✓ intercept in model
Statistics tab:

*Response: Y – (0 – nonacademic, 1 – academic)
Explanatory:
X – academic score (total of reading, writing, math, science, civics)
School (0 – public, 1 – private)
SES (1-low, 2-middle, 3-high) (characterize with 2 dummies)
$d_1 = 0$ – not low, 1 – low
$d_2 = 0$ – not med, 1 – medium

Goodness of Fit[b]

	Value	df	Value/df
Deviance	598.611	535	1.119
Scaled Deviance	598.611	535	
Pearson Chi-Square	533.771	535	.998
Scaled Pearson Chi-Square	533.771	535	
Log Likelihood[a]	-332.053		
Akaike's Information Criterion (AIC)	670.107		
Finite Sample Corrected AIC (AICC)	670.147		

Log-Likelihood Function:
Kernal
All other tabs:
left as defaults.

$-2L_0 = 664.107$
$L_0 = -332.058$

Bayesian Information Criterion (BIC)	683.298		
Consistent AIC (CAIC)	686.298		

Dependent Variable: y
Model: (Intercept), x, school
a. The kernel of the log likelihood function is displayed and used in computing information criteria.

Parameter Estimates

Parameter	B	Std. Error	95% Profile Likelihood Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-5.738	.7474	-7.236	-4.301	58.952	1	.000
x	.027	.0027	.022	.032	99.251	1	.000
[school=.00]	-1.421	.2881	-2.009	-.875	24.337	1	.000
[school=1.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: y
Model: (Intercept), x, school
a. Set to zero because this parameter is redundant.
b. Information criteria are in small-is-better form.

$\pi(x)$, success prob is prob of attending academic school

Equation: $\widehat{\text{logit}(\pi(x))} = -5.738 + 0.027x - 1.421\text{school}$

Example 5 MODEL L_1: GLM APPROACH: LOG LIKELIHOOD KERNAL (to obtain L_1)

Analysis>Generalized Linear Models>Generalized Linear Models
Type of Model Tab:
Binary Logistic
Response Tab:
Dependent Variable: y
(change Binary reference category to first – lowest value)
Predictor Tab:
x in the Covariate box
school in Factors box
ses in Factors box
Model Tab:
x as a main effect
school as main effect
ses in Factors box
✓ intercept in model
Statistics tab:
Log-Likelihood Function:
Kernal
All other tabs:
left as defaults.

$-2L_1 = 651.960$
$L_1 = -325.980$

ses Likelihood Ratio Chi-Square
12.147

*Response: Y – (0 – nonacademic, 1 – academic)
Explanatory: X – academic score (total of reading, writing, math, science, civics)
School (0 – public, 1 – private)
SES (1-low, 2-middle, 3-high)
(View SES as an "x" with 3 levels)
$d_1 = 0$ – not low, 1 – low
$d_2 = 0$ – not med, 1 – medium

Goodness of Fit[b]

	Value	df	Value/df
Deviance	632.552	571	1.108
Scaled Deviance	632.552	571	
Pearson Chi-Square	570.021	571	.998
Scaled Pearson Chi-Square	570.021	571	
Log Likelihood[a]	-325.980		
Akaike's Information Criterion (AIC)	661.960		
Finite Sample Corrected AIC (AICC)	662.061		
Bayesian Information Criterion (BIC)	683.944		
Consistent AIC (CAIC)	688.944		

Dependent Variable: y
Model: (Intercept), x, school, ses
a. The kernel of the log likelihood function is displayed and used in computing information criteria.
b. Information criteria are in small-is-better form.

Tests of Model Effects

Source	Type III		
	Likelihood Ratio Chi-Square	df	Sig.
(Intercept)	76.689	1	.000
x	98.688	1	.000
school	24.354	1	.000
ses	12.147	2	.002

Dependent Variable: y
Model: (Intercept), x, school, ses

Parameter Estimates

Parameter	B	Std. Error	95% Profile Likelihood Confidence Interval	Hypothesis Test

			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-4.722	.8005	-6.319	-3.176	34.798	1	.000
x	.025	.0028	.020	.031	80.320	1	.000
[school=.00]	-1.346	.2912	-1.939	-.793	21.369	1	.000
[school=1.00]	0[a]	.	.	.	.	.	.
[ses=1.00]	-.891	.2853	-1.455	-.335	9.756	1	.002
[ses=2.00]	-.706	.2355	-1.173	-.248	8.978	1	.003
[ses=3.00]	0[a]	.	.	.	.	.	.
(Scale)	1[b]						

Dependent Variable: y
Model: (Intercept), x, school, ses
a. Set to zero because this parameter is redundant.
b. Fixed at the displayed value.

$\pi(x)$, success prob is prob of attending academic school

Equation: $\text{logit}(\widehat{\pi(x)})=-4.722+0.025x-1.346\text{school}-0.891d1-0.706d2$

Model Building and Application in Logistic Regression

Choosing Predictors:

-At least 10 of each outcome for every predictor
-Avoid too many predictors
-Avoid multi-collinearity (caused by linear dependence of predictors)
-If have interaction terms in model, all lower order terms should be included
-all of none of a dummy variable should be included
-Include all variables important to design of study, even if effect non-significant.

HSB model (could have up to 29 predictors (because 292/10 = 29)

Analyze > Descriptive Statistics > Frequencies
Variable(s): y
Ok

y

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.00	292	48.7	48.7	48.7
	1.00	308	51.3	51.3	100.0
	Total	600	100.0	100.0	

Finding Models:

Stepwise selection:

Backward: Drop existing variable with the "least significant' effect at each step. Stop when another step shows no further improvement in the model

Forward: Add variable that shows the "most significant" effect at each step. Stop when another step shows no further improvement in the model.

Example: HSB with predictors x (= reading + writing + math + science + civics), science, SES and school (Model M2 for our purposes here)

Y – (0-non-academic, 1-academic)
X – academic score (total of reading, writing, math, science, civics)
School (0-public, 1 – private)
Science (score on Science test)
SES (1-low, 2-middle, 3-high)
(categorize SES with 2 dummies)
d_1 = 0 – not low, 1 – low
d_2 = 0 – not med, 1 - medium

ANALYZE>REGRESSION>BINARY LOGISTIC
Logistic Regression Screen
-Move y to Dependent box
-Move SES, school, x, science to Covariate box
Covariate box will read
x
School
science
SES
-Method: Forward LR
-Click Categorical Button
Covariate Screen
-Move School and SES from Covariate box to categorical Covariates box
-Highlight School
-Change to Indicator(first)
Categorical Covariate box will read
School(indicator(first))
SES(indicator)
-Continue
Logistic Regression Screen
Covariate box will read
x
school (cat)
science
ses(Cat)
Ok

Categorical Variables Codings

		Frequency	Parameter coding (1)	Parameter coding (2)
ses	1.00	139	1.000	.000
	2.00	299	.000	1.000
	3.00	162	.000	.000
school	.00	506	.000	
	1.00	94	1.000	

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	139.082	1	.000
	Block	139.082	1	.000
	Model	139.082	1	.000
Step 2	Step	28.161	1	.000
	Block	167.243	2	.000
	Model	167.243	2	.000
Step 3	Step	13.506	1	.000
	Block	180.749	3	.000
	Model	180.749	3	.000
Step 4	Step	12.806	2	.002
	Block	193.555	5	.000
	Model	193.555	5	.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	95% C.I.for EXP(B) Upper
Step 1[a]	x	.027	.003	106.411	1	.000	1.028	1.022	1.033
	Constant	-7.055	.695	103.104	1	.000	.001		
Step 2[b]	school(1)	1.421	.288	24.337	1	.000	4.143	2.355	7.287
	x	.027	.003	99.251	1	.000	1.027	1.022	1.033
	Constant	-7.160	.714	100.606	1	.000	.001		
Step 3[c]	school(1)	1.449	.291	24.699	1	.000	4.257	2.404	7.537
	x	.040	.005	72.679	1	.000	1.041	1.031	1.051
	science	-.063	.018	12.879	1	.000	.939	.907	.972
	Constant	-7.308	.725	101.481	1	.000	.001		
Step 4[d]	ses			12.506	2	.002			
	ses(1)	-.926	.289	10.256	1	.001	.396	.225	.698
	ses(2)	-.731	.238	9.413	1	.002	.481	.302	.768
	school(1)	1.374	.295	21.738	1	.000	3.952	2.218	7.043
	x	.039	.005	65.371	1	.000	1.039	1.030	1.049
	science	-.065	.018	13.502	1	.000	.937	.905	.970
	Constant	-6.193	.786	62.014	1	.000	.002		

Options:
Check CI for exp(B) 95%
Continue
Ok

AIC = -2(log likelihood - # parameters) = -2log likelihood + 2(#parameters)

M1: (predictors x, school, and science)
AIC = 650.601 + 8 = 658.601
(Parameters for intercept, x, school, and science)

M2: (predictors x, school, science, and SES)
AIC = 637.795 + 12 = 649.795
(Parameters for intercept, x, school, science, and 2 dummy variables for SES)
AIC SMALLER – BETTER MODEL

a. Variable(s) entered on step 1: x.
b. Variable(s) entered on step 2: school.
c. Variable(s) entered on step 3: science.
d. Variable(s) entered on step 4: ses.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	692.268[a]	.207	.276
2	664.107[b]	.243	.324
3	650.601[b]	.260	.347
4	637.795[b]	.276	.368

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.
b. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

$\pi(x)$ (success prob is prob of attending academic school)

Eqn: $\text{logit}(\widehat{\pi(x)}) = = -6.193 + 0.039x + 1.374\text{school} - 0.065\text{science} - 0.926d_1 - 0.731d_2$

Classification Tables:

If $\widehat{\pi(x)}$ is >= 0.5, predict success for this individual more likely than failure (Classify Individual as Success)
If $\widehat{\pi(x)}$ is < 0.5, predict failure for this individual less likely than success (Classify Individual as Failure)

Create Classification table to compare actual measurements with predicted categories.

Model M2 above provides the following classification table. We are interested in the one at Step 4, when all predictors are in the model.

Our model has percentage corrects of over 70% for both Y=0 and Y=1.

Classification Table[a]

			Predicted		
			y		Percentage
	Observed		.00	1.00	Correct
Step 1	y	.00	198	94	67.8
		1.00	81	227	73.7
	Overall Percentage				70.8
Step 2	y	.00	208	84	71.2
		1.00	73	235	76.3
	Overall Percentage				73.8
Step 3	y	.00	202	90	69.2
		1.00	75	233	75.6

	Overall Percentage				72.5
Step 4	y	.00	208	84	71.2
		1.00	74	234	76.0
	Overall Percentage				73.7

a. The cut value is .500

If you run M2 with Enter, and put all the predictors in at once, you obtain:

Classification Table[a]

			Predicted		
			y		
	Observed		.00	1.00	Percentage Correct
Step 1	y	.00	208	84	71.2
		1.00	74	234	76.0
	Overall Percentage				73.7

a. The cut value is .500

Correlation:

Run model M2 as above, entering all variables at once, but pull up the Save pop-up, and choose Probabilities under predicted values when you run. This will create a column labeled PRE_1 in your data file.

Then follow the path Analyze > Correlate > Bivariate and put the variables y and Predicted Probability [PRE_1] in the Variables box. Say ok. You will receive the output found below.

(I suspect that Karen ran a GLM approach, which will also provide predicted values, given that her correlation is slightly different from mine, viz. 0.533 versus 0.536.)

Correlations

		y	Predicted probability
y	Pearson Correlation	1	.536**
	Sig. (2-tailed)		.000
	N	600	600
Predicted probability	Pearson Correlation	.536**	1
	Sig. (2-tailed)	.000	
	N	600	600

**. Correlation is significant at the 0.01 level (2-tailed).

$R^2 = .536^2 = 0.287$

When you run model M2 as above, and enter all the variables at once, the Model Summary returned looks like this. Note that it reminds you that it did one step in order to obtain the output.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	637.795[a]	.276	.368

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

"Influence Diagnostics"

How influential single observations are on the model fit is of interest. Some measures that look at this include.

1. DFBETA – the change in the parameter estimate divided by its standard error, when excluding the individual
2. c– confidence displacement – this measures the change in the joint confidence intervals for all parameters when excluding the individual
3. The change in the Pearson χ^2 or G^2 (Deviance) when the observation is omitted

The larger the measure, the greater the influence of the individual.

We can run model M2 as above, bring up the save box and select Leverage values and DFBeta(s) under the Influence group. Say Continue and Ok to run.

The leverage values in SPSS consider observations to be influential when they have values larger than 2(k+1)/n where n is the sample size, and k is the number of predictors.

To create a descriptive statistics box to look at a summary of your results, follow the path Analyze>Descriptive Statistics > Descriptives and bring the Leverage value variable, and the DFBETA variables to the Variable(s) box. Drag and drop them to get them to appear in the same order as in Karen's notes.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Leverage value	600	.00255	.03216	.0100000	.00502988
DFBETA for x	600	-.00116	.00050	.0000000	.00019665
DFBETA for science	600	-.00250	.00346	.0000000	.00072190
DFBETA for school(1)	600	-.07069	.05189	-.0000021	.01207167

DFBETA for ses(1)	600	-.04132	.03338	.0000014	.01171271
DFBETA for ses(2)	600	-.03367	.03515	-.0000003	.00967199
Valid N (listwise)	600				

We can see that there are no standout individual values by looking at these summary statistics.

Multicategory Logit Models – Nominal Response

Variable Y has J categories, i= 1,…, J
$\pi_1, \ldots.\pi_J$ are the probabilities that observations fall into the categories

Question: What effect do certain predictors have on the $\pi_1, \ldots.\pi_J$?

Binary Logistic: Model odds for success (probability of success in relationship to probability of failure)
Multicategory Logit: Simultaneously model all relationships between probabilities for pairs of categories(model odds of falling within one category instead of another)

Baseline Logits approach:
Pair a baseline category with all remaining categories
Usually last category (J)

Define:
Baseline Logit
$\text{Log}(\pi_i/\pi_J)$, i = 1,… J-1

Model with one predictor x
$\text{Log}(\pi_i/\pi_J) = \alpha_i + \beta_i x$, i = 1, …. , J-1

For each category i compared, a new set of parameters introduced

SPSS models these J-1 model equations simultaneously.

Baseline category logit model allows comparison of categories a and b

$\text{Log}(\pi_a/\pi_b) = \log ((\pi_a/\pi_J)/(\pi_b/\pi_J)) = \log (\pi_a/\pi_J) – \log(\pi_b/\pi_J) = (\alpha_a – \alpha_b) – (\beta_a – \beta_b)x$

Example: Contraceptive Use in Dependency on Age
File is Contraceptive.sav
3165 currently married women
Variables
contra (s(sterilization), o(other), n(none))

age (midpoints of 5 year intervals, viz: 17.5, 22.5…47.5)
a2 – age squared

Other columns:
freq -frequency in each age midpoint/contraception level cell
s – frequencies in s category for each age midpoint
o – frequencies in o category for each age midpoint
n - frequencies in n category for each age midpoint
lsn – log (π_s/π_n)
lon – log (π_o/π_n)

It is of interest to create a crosstab table, for visual acuity. We will also find a Chi-square statistic, in order to test for independence. (Below, we see that we reject our null hypothesis of no relationship between age and method of contraception, and conclude that the two variables are related (associated))

Commands:
Data>Weight Cases
Weight cases by Frequency Variable: freq
OK

Analyze >Descriptive Statistics > Crosstabs
Row(s): age
Column(s): contra
Statistics button: check chi square
Continue
OK

age * contra Crosstabulation

Count

		contra			Total
		n	o	s	
age	17.50	232	61	3	296
	22.50	400	137	80	617
	27.50	301	131	216	648
	32.50	203	76	268	547
	37.50	188	50	197	435
	42.50	164	24	150	338
	47.50	183	10	91	284
Total		1671	489	1005	3165

The o and n numbers were entered into their own columns in the contraceptive.sav file.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	430.028[a]	12	.000
Likelihood Ratio	521.103	12	.000
N of Valid Cases	3165		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 43.88.

We will be looking at using a multicategory logit model.

It is therefore of interest to **graph age versus log(π_s/π_n) and age versus log(π_o/π_n).** These log values can be obtained for our age midpoints by using the models below, but they can also be readily obtained if we note that for age midpoints of interest: Log(π_s/π_n) = Log(s/n) and Log(π_o/π_n) = Log(o/n)

Commands:
Transform>Compute Variable
Target Variable lsn
Numeric Expression: LN(s/n)

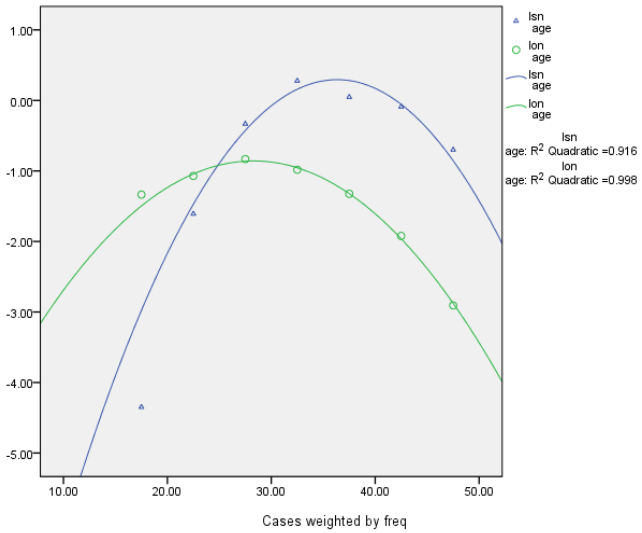
Transform>Compute Variable
Target Variable:lon
Numeric Expression: LN(o/n)

Graph>Legacy Dialogs>Scatter/Dot> Overlay Scatter
Define
X-Y pairs

Pair	Y Variable	X Variable
1	lsn	age
2	lon	age

Ok

Double click to activate graph
Click on lsn on legend, and change type to a triangle in the popup, apply, close
Right click on an lsn triangle (or a lon circle), select add fit line at total from the dropdown menu, select quadratic in the popup, and apply.
Close the activated graph window to return it to the output file.



SES Odds Ratio: oddslow/oddshigh = $e^{-0.891}$ = 0.4102

(reference category is high)

(odds 0.4102 times higher to attend academic school for low income than high income, other variables held constant)

(odds 1/0.4102 = 2.4378 times higher to attend academic school for high income than low income, other variables held constant)

SES Odds Ratio: oddsmed/oddshigh = $e^{-.706}$ = .4936

(reference category is high)

(odds 0.4936 times higher to attend academic school for medium income than high income, other variables held constant)

(odds 1/0.4102 =2.0259 times higher to attend academic school for high income than low income, other variables held constant)We see that a quadratic model appears to fit our data well.

We are interested in comparing two models, a linear model and a quadratic model. We will obtain logit equations, and perform tests of significance. Our results are summarized here, and the output and commands needed to obtain the output are provided below.

Linear Model (baseline n)	Quadratic Model (baseline n)
$\text{Log}(\pi_s/\pi_n) = \alpha_s + \beta_{1s}\text{age}$	$\text{Log}(\pi_s/\pi_n) = \alpha_s + \beta_{1s}\text{age} + \beta_{2s}\text{age}^2$
$\text{Log}(\pi_o/\pi_n) = \alpha_o + \beta_{1o}\text{age}$	$\text{Log}(\pi_o/\pi_n) = \alpha_o + \beta_{1o}\text{age} + \beta_{2o}\text{age}^2$
From output BELOW:	From output BELOW:
$\text{Log}(\widehat{\pi s}/\widehat{\pi n}) = -2.236 + 0.053\text{age}$	$\text{Log}(\widehat{\pi s}/\widehat{\pi n}) = -12.618 + .710\text{age} - .010\text{age}^2$
$\text{Log}(\widehat{\pi o}/\widehat{\pi n}) = -.152 - 0.037\text{ age}$	$\text{Log}(\widehat{\pi o}/\widehat{\pi n}) = -4.550 + .264\text{age} - .005\text{age}^2$

Linear Model (baseline o)	Quadratic Model (baseline o)
From output BELOW:	From output BELOW:
$\text{Log}(\widehat{\pi s}/\widehat{\pi o}) = -2.084 + 0.091\text{age}$	$\text{Log}(\widehat{\pi s}/\widehat{\pi o}) = -8.068 + .446\text{age} - 0.005\text{age}^2$

EXAMPLE OF INTERPRETATION: Given that the chosen contraceptive is either sterilization or none, find how the odds that a woman used sterilization at 26 changed from the odds that a woman used sterilization at 25.

$\text{Log}(\widehat{\pi s}/\widehat{\pi n}) = -12.6 + .710\text{age} - .010\text{age}^2$

$(\widehat{\pi s}/\widehat{\pi n}) = e^{(-12.6 + .710\text{age} - .010\text{age}2)}$

For age = 25, $(\widehat{\pi s}/\widehat{\pi n})$ = odds (25) = $e^{(-12.6 + .710(25) - .010(25)2)}$

For age = 26, $(\widehat{\pi s}/\widehat{\pi n})$ = odds (26) = $e^{(-12.6 + .710(26) - .010(26)2)}$

Odds(26)/Odds(25) = $e^{(.710(26-25) - .010(262-252))} = e^{(0.710(1) - .010(51))} = 1.22$

There was a 22% increase in the odds from the age of 25 to 26.

Linear Model (Test of Fit):	Quadratic Model (Test of Fit):
From output BELOW	From output BELOW
Pearson χ^2 = 268.169 with df = 10, p-value =0.000	Pearson χ^2 = 18.869 with df = 8, p-value =0.016
Deviance = 293.979 with df =10, p-value = 0.000	Deviance = 20.475 with df =9, p-value = 0.009

The quadratic model shows a marginally "acceptable" model.

Linear Model (Wald χ^2 signif test on odds ratios)
OUTPUT BELOW

Odds-ratio	Variable	χ^2	Df	P-value
Ster-None	age	130.295	1	<.001
Other-None	age	33.008	1	<.001
Ster-Other	age	168.869	1	<.001

Quadratic Model(Wald χ^2 signif test on odds ratios)
OUTPUT BELOW

Odds-ratio	Variable	χ^2	Df	P-value
Ster-None	age	232.22	1	<.001
	age^2	218.25	1	<.001
Other-None	age	31.47	1	<.001
	age^2	39.23	1	<.001
Ster-Other	age	54.79	1	<.001
	age^2	28.24	1	<.001

Here we see that both age and age^2 have a significant effect on the log odds for sterilization versus none,other versus none, and sterilization versus other.

QUADRATIC MODEL COMMANDS AND OUTPUT:

To obtain parameter estimate output for 1st two equations:

Levels (categories) for the variable contra will be n, o, s (and the first will be the one lowest in the alphabet)

Run with baseline logit of n (none)

Commands:

Analyze>Regression>Multinomial Logistic

Dependent: contra

Reference Category: first

Covariate(s):

age

a2

Model button:

Multinomial Logistic Regression: Model popup window
Choose Custom
Build Terms: Main effects
Select age and a2 from Factors & Covariates window and bring them over to Forced Entry Terms window.
Continue
Statistics Window: leave defaults (Under model: Case processing summary, Pseudo R-square, Step summary, Model fitting information, and under Parameters: Estimates and Likelihood ratio tests, and under Define Subpopulations: Covariate patterns defined by factors and covariates) Also check "Goodness of Fit" under model.
Criteria Window: leave defaults
Options Window: leave defaults
Save Window: select estimated response probabilities
To obtain parameter output estimate for 3rd equation, rerun above commands using baseline logit other. When you click on reference category, it offers a custom option. Set it to o (other).

Parameter Estimates

contra[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
o	Intercept	-4.550	.694	42.998	1	.000			
	age	.264	.047	31.472	1	.000	1.302	1.187	1.428
	a2	-.005	.001	39.233	1	.000	.995	.994	.997
s	Intercept	-12.618	.757	277.545	1	.000			
	age	.710	.046	243.225	1	.000	2.033	1.860	2.223
	a2	-.010	.001	218.250	1	.000	.990	.989	.992

a. The reference category is: n.

Parameter Estimates

contra[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
n	Intercept	4.550	.694	42.998	1	.000			
	age	-.264	.047	31.472	1	.000	.768	.700	.842
	a2	.005	.001	39.233	1	.000	1.005	1.003	1.006
s	Intercept	-8.068	.949	72.285	1	.000			
	age	.446	.060	54.789	1	.000	1.561	1.388	1.757
	a2	-.005	.001	28.843	1	.000	.995	.993	.997

a. The reference category is: o.

The Goodness-of-Fit output for this quadratic model can be found on both runs.

Goodness-of-Fit

	Chi-Square	df	Sig.

Pearson	18.869	8	.016
Deviance	20.475	8	.009

LINEAR MODEL COMMANDS AND OUTPUT

The Linear model commands are similar to the QUADRATIC MODEL COMMANDS. The only difference is that there is only one explanatory variable, age, rather than two explanatory variables, age and a2.

Parameter Estimates

contra[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
o	Intercept	-.152	.190	.638	1	.424			
	age	-.037	.006	33.008	1	.000	.964	.951	.976
s	Intercept	-2.236	.159	198.226	1	.000			
	age	.053	.005	130.295	1	.000	1.055	1.045	1.065

a. The reference category is: n.

Parameter Estimates

contra[a]		B	Std. Error	Wald	df	Sig.	Exp(B)	95% Confidence Interval for Exp(B)	
								Lower Bound	Upper Bound
n	Intercept	.152	.190	.638	1	.424			
	age	.037	.006	33.008	1	.000	1.038	1.025	1.051
s	Intercept	-2.084	.215	93.818	1	.000			
	age	.091	.007	168.869	1	.000	1.095	1.080	1.110

a. The reference category is: o.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	268.169	10	.000
Deviance	293.978	10	.000

TO TEST IF AGE HAS AN EFFECT ON CONTRACEPTIVES:

WE WILL NEED **FURTHER QUADRATIC MODEL OUTPUT** HERE, VIZ: A TABLE WITH MODEL FITTING INFORMATION, AS BELOW. IT IS OBTAINED IN THE OUTPUT FROM THE COMMANDS ABOVE.

Ho: all the age and age^2 parameters are 0 ($\beta_{1s} = \beta_{2s} = \beta_{1o} = \beta_{2o} = 0$), and the intercept model suffices) versus
Ha: at least one of the parameters is non-zero, choose $\alpha = 0.05$
Assumptions hold
$\chi^2 = -2(L_o - L_1) = 500.628$ with df =4 (see output below)
Where L_o is intercept only model and L_1 is model with age and age^2

P-value <0.001
Reject Ho at the 5% significance level. The data provides sufficient evidence that age has an effect on the choice of contraceptive.

Model Fitting Information

Model	Model Fitting Criteria	Likelihood Ratio Tests		
	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	601.377			
Final	100.749	500.628	4	.000

<u>COMPARING ESTIMATED PROBABILITIES FROM CONTINGENCY TABLES WITH PROBABILITIES ESTIMATED FROM THE QUADRATIC MODEL</u>

Karen's notes have the formulas for calculating the $\hat{\pi}\ s$.

Commands:
SPSS calculates the $\hat{\pi}\ s$ for us when we run the quadratic model as we did above, and, in addition, use the Save button, and in the Saved variables box, check "estimated response probabilities".

This will yield three new columns in the data file contraceptive.sav . They will be labeled, EST1-1, EST2_1, AND EST3_1. For age= 22.5, we have

EST1_1 EST2_2 EST3_3
.64 .23 .13
And these values match the values found in Karen's notes for $\widehat{\pi n}$, $\widehat{\pi o}$ and $\widehat{\pi s}$ when age = 22.5

Ordinal Response

Model
Y is an ordinal variable with categories 1, 2,…J
Cumulative probability for category j = $P(Y<=J) = \pi_1 + \ldots. + \pi_j$, j = 1,2,….J
P(Y<=1) <= P(Y<=2) <= …. P(Y<=J) = 1

Model an ordinal response:
Model the cumulative response probabilities (or cumulative odds)

Cumulative Logits:
Logit (P(Y<=j)) = $\log\left(\frac{P(Y\leq j)}{1-P(Y\leq j)}\right) = \log\left(\frac{\pi_1+\pi_2+\cdots+\pi_j}{\pi_{j+1}+\cdots+\pi_J}\right)$

Categories are then: I (1 2 … j) and II (j+1 … J)

For one predictor variable x, proportional odds model is , for j = 1, 2, … J-1,

Logit (p(Y<=j)) = $\alpha_j + \beta x$

NOTE: slope β the same for all cumulative logits
β describes the effect of x on the odds of falling into categories 1 to j.
The model assumes that this effect is the same for all cumulative odds.

NOTE: Property of this model (see textbook page 181 for diagram) is that, for β >0, as x increases, the probability of falling in a lower category increases.
NOTE: SPSS models logit(Y<=j) = $\alpha_j - \beta x$
And reports the negative of the slope!

Example:
HSB Data:
Y, the ordinal variable here, is SES (low – 1, middle – 2, high – 3)
Math
Writing
Sex (1 – Male, 2 - Female)
Race (1-Hispanic 2-Asian 3 - Black 4- White)

We will use the model equation below for j = 1, 2.
Logit(P(Y<=j) = $\alpha j+\beta_1$math +$\beta_2$writing+$\beta_3$sex + β_{41}race1 to β_{42}race2 + β_{43} race3

We first check to be sure we have no ses/sex cells or ses/race cells with very few observations, as this would cause problems in analyzing our model.

Commands:
Analyze>Descriptive Statistics > Crosstabs
Row(s): ses

Column(s): sex
Ok

Analyze>Descriptive Statistics > Crosstabs
Row(s): ses
Column(s): race
Ok

ses * sex Crosstabulation

Count

		sex		
		1.00	2.00	Total
ses	1.00	50	89	139
	2.00	144	155	299
	3.00	79	83	162
Total		273	327	600

ses * race Crosstabulation

Count

		race				
		1.00	2.00	3.00	4.00	Total
ses	1.00	23	8	24	84	139
	2.00	34	14	24	227	299
	3.00	14	12	10	126	162
Total		71	34	58	437	600

There is no indication of empty or “small” cells to worry about.

We will create our estimated model. Some commands below are included to help us check our fit.

COMMANDS:
Analyze > Regression >Ordinal
Dependent: ses
Factor(s):
sex
race
Covariate(s):
writing
math
Output button:
Display area:

(Leave defaults: Goodness of fit statistics, Summary statistics, parameter estimates)
Check "Test of parallel lines"
Saved Variables area: Check "Predicted category"
Continue
Ok

Output of use from these commands is posted below with explanations.

Model Fitting Information

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Intercept Only	1221.075			
Final	1148.593	72.482	6	.000

Link function: Logit.

Here we see that the model that includes all the predictors fits significantly better than the model that omits all the predictors (the intercept model): Reject Ho (intercept only model), with $\chi^2 = 74.482$, df =6, and p-value <0.001.

Goodness-of-Fit

	Chi-Square	df	Sig.
Pearson	1148.222	1138	.410
Deviance	1125.438	1138	.599

Link function: Logit.

Here both the Pearson χ^2 (1148.22, df = 1138) and the Deviance (1125.438, df = 1138) suggest good fit.

Pseudo R-Square

Cox and Snell	.114
Nagelkerke	.130
McFadden	.058

Link function: Logit.

The Pseudo R^2 values give the impression that the model fits somewhat well.

Here are the parameter estimates.

Parameter Estimates

		Estimate	Std. Error	Wald	df	Sig.	95% Confidence Interval	
							Lower Bound	Upper Bound
Threshold	[ses = 1.00]	2.509	.555	20.422	1	.000	1.421	3.597

	[ses = 2.00]	4.927	.587	70.497	1	.000	3.777	6.077
Location	writing	.027	.011	5.833	1	.016	.005	.049
	math	.043	.012	14.121	1	.000	.021	.066
	[sex=1.00]	.456	.170	7.210	1	.007	.123	.788
	[sex=2.00]	0[a]	.	.	0	.	.	.
	[race=1.00]	-.143	.255	.314	1	.575	-.644	.357
	[race=2.00]	-.151	.345	.193	1	.660	-.827	.524
	[race=3.00]	-.476	.279	2.910	1	.088	-1.024	.071
	[race=4.00]	0[a]	.	.	0	.	.	.

Link function: Logit.

a. This parameter is set to zero because it is redundant.

Threshold values. The estimates tell us the values used to differentiate socioeconomic classes. 2.509 differentiates low ses from middle ses and 4.927 differentiates middle ses from high ses.

Math estimate:

Math is a significant predictor of SES (χ^2 =14.121, df = 1, p-value <0.001)

A one point increase in math results in an increase in the ordered log odds of being in a higher ses category of 0.043 (all other variables in the model held constant). Since $e^{.043}$ = 1.044, a one point increase in math increases your chance of being in a higher SES category by 4.4%.

Writing estimate:

Writing is a significant predictor of SES (χ^2 =5.833, df = 1, p-value <0.016)

A one point increase in writing results in an increase in the ordered log odds of being in a higher ses category of 0.027 (all other variables in the model held constant). Since $e^{.027}$ = 1.027, a one point increase in math increases your chance of being in a higher SES category by 2.7%.

Sex estimate:

Sex is a significant predictor of SES (χ^2 =7.210, df = 1, p-value <0.007)

If a person is male, this results in increase in the ordered log odds of being in a higher socioeconomic category of 0.456 (all other variables held constant). Since $c^{.456}$ = 1.578, males are 57.8% more likely to be in a higher SES category than females (all other variables held constant). ALTERNATIVE INTERPRETATION. Is it better to say that the ordered log odds of being in a higher socioeconomic category are 1.578X times higher for males than females (all other variables held constant)?

Race estimate:

Race is not a significant predictor of SES. (all p-values for all race dummy variables are large)

We need to check if it is reasonable to assume that the slopes for the cumulative odds are all the same (that is: that the S-curves for all SES categories are truly parallel. We are checking if the chance of falling into a higher SES is equally affected by the predictor variables for all SES categories.)

Test of Parallel Lines[a]

Model	-2 Log Likelihood	Chi-Square	df	Sig.
Null Hypothesis	1148.593			
General	1142.142	6.450	6	.375

The null hypothesis states that the location parameters (slope coefficients) are the same across response categories.

a. Link function: Logit.

Here Ho is that the lines are parallel. Our result is non-significant, indicating that we cannot reject the possibility that the effect is the same for all categories. This supports our model, but we should, as always be cautious when "accepting" our Ho because we don't know the error probability for that decision.

Commands above have created a Predicted Response Category Variable in the HSB data file. It will be to the far right of all the columns in your HSB data file should be labeled PRE_# (# will depend on whether this is your first predictor variable for this set of data, or if you have been doing several examinations are keeping the new columns that are appearing in the HSB data file when you do), and when you hover over the label, you will see that it is a column of Predicted Response Category.

We create a crosstab table of SES and the predicted response categories for SES with our model in order to see how well the model does.

COMMANDS

Analyze>Descriptive Statistics>Crosstabs

Row(s): ses

Column(s): Predicted Response Category

OK

ses * Predicted Response Category Crosstabulation

Count

		Predicted Response Category			
		1.00	2.00	3.00	Total
ses	1.00	18	117	4	139
	2.00	10	259	30	299
	3.00	7	131	24	162
Total		35	507	58	600

Because all observations do not fall on the diagonal, we know that our model is not very powerful for prediction of the SES of an individual.

Loglinear Models for Contingency Tables

Consider the GLM:

Link function g(mu) = log(mu)
Poisson -random component
Linear predictor - systematic component

Contingency table:
Response Variable: Count
Predictors: Categorical Variables that define the contingency table

Suppose have two categorical variables, X (with I levels) and Y(with J levels)

π_{ij} = joint probability = $\pi_{i+}\pi_{+j}$, i = 1,…, I and j = 1,… J

Expected cell count $\mu_{ij} = n\pi_{ij}$, i = 1,…, I and j = 1,… J

If the two variables are independent,
Expected cell count $\mu_{ij} = n\pi_{i+}\pi_{+j}$ i = 1,…, I and j = 1,… J

$Log(\mu ij) = log(n) + log(\pi_{i+}) + log(\pi_{+j})$ i = 1,…, I and j = 1,… J

This is the **loglinear model of independence.**
It only fits if the two variables are independent.

Writing convention:

$Log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$ i = 1,…, I and j = 1,… J

- λ is the overall effect (included to ensure the $\Sigma\pi_{ij} = n$)

- λ_i^X is the (main or marginal) effect of category i of variable X on the log of the expected cell count

- λ_j^Y is the (main or marginal) effect of category j of variable Y on the log of the expected cell count

Model fit statistics for the model (Pearson's χ^2 and the Likelihood Ratio Statistic) are used to test if the two variables are independent.

Example: Two way table: (alien_walle.sav)

Model is:
$Log(\mu ij) = \lambda + \lambda_i^{Movie} + \lambda_j^{Rating}$ i = 1,2 and j = 1,2,3,4,5

Movie_n (1 – Alien, 2-Wall-e)
Rating (12 – 1&2, 34 – 3&4, 56 - 5&6, 78 – 7&8, 910 – 9&10)

We will examine the fit of the model mentioned above with this data.

Commands:
Analyze>Loglinear>General
Distribution of Cell Counts:
Poisson
Factor(s):
Movie_n
Rating
Model tab:
Custom
Build Term(s): Choose Main effects
Highlight movie_n and rating in the Factors and Covariates box
Move them (with the arrow) to the Terms in Model box
Continue
Save tab (just note there are no default options and cancel)
Options tab (just note the defaults and cancel)
Display defaults: Frequencies, Residuals
Plot defaults: Adjusted residuals, Normal probability for adjusted
Continue
OK

NOTE: At no point did we tell it a dependent count variable here. It looked at the data file and made its own cell counts and residuals table. It matches the movie-n values and ratings values with the counts from the freq column. Pretty cool, eh. HOWEVER,

It is probably better to do a Data>Weight Cases command, though, as it might not always be the case that it is obvious in what column SPSS should look for the frequencies.

(See how it counted)

Cell Counts and Residuals[a,b]

		Observed		Expected			Standardized	Adjusted	
movie_n	rating	Count	%	Count	%	Residual	Residual	Residual	Deviance
1.00	12	1932	.6%	3938.206	1.2%	-2006.206	-31.969	-43.810	-35.505
	34	1760	.5%	2022.127	.6%	-262.127	-5.829	-7.933	-5.962
	56	8403	2.5%	8101.193	2.4%	301.807	3.353	4.662	3.333
	78	54233	16.4%	46835.648	14.1%	7397.352	34.181	55.721	33.336
	910	83874	25.3%	89304.827	26.9%	-5430.827	-18.173	-38.596	-18.362
2.00	12	6758	2.0%	4751.794	1.4%	2006.206	29.104	43.810	27.349
	34	2702	.8%	2439.876	.7%	262.124	5.307	7.931	5.216
	56	9473	2.9%	9774.807	2.9%	-301.807	-3.053	-4.662	-3.069

	78	49114	14.8%	56511.352	17.1%	-7397.352	-31.118	-55.721	-31.837
	910	113185	34.2%	107754.173	32.5%	5430.827	16.544	38.596	16.408

a. Model: Poisson

b. Design: Constant + movie_n + rating

You will get a lot more output, but we are only interested in the Goodness of Fit tests table.

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	4823.100	4	.000
Pearson Chi-Square	4692.375	4	.000

a. Model: Poisson

b. Design: Constant + movie_n + rating

Ho: independence between movie_n and rating versus Ha: is a relationship (dependence) between movie_n and rating

Reject Ho (Pearson χ^2=4692.4, df = 4, p-value <0.001). We do not have independence.

WE SHOULD NOT USE THIS MODEL. WE NEED A MODEL THAT TAKES INTERACTION INTO EFFECT. WE DO NOT HAVE INDEPENDENCE OF RATING AND MOVIE…..

However, for interest ONLY, and to help us get ready to read the output for more complex models, I include the parameter estimates here.

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	11.588	.003	4210.148	.000	11.582	11.593
[movie_n = 1.00]	-.188	.003	-53.821	.000	-.195	-.181
[movie_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12]	-3.121	.011	-284.761	.000	-3.143	-3.100
[rating = 34]	-3.788	.015	-250.354	.000	-3.818	-3.758
[rating = 56]	-2.400	.008	-307.255	.000	-2.415	-2.385
[rating = 78]	-.645	.004	-168.046	.000	-.653	-.638
[rating = 910]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + movie_n + rating

FOR INTEREST ONLY, WE LOOK AT READING THIS OUTPUT, AND INTERPRETING IT IF IT WERE REASONABLE TO USE IT.

X-rating i = 1, 2, 3, 4, 5

Y-movie j = 1, 2

$\text{Log}(\mu_{11}) = \lambda + \lambda_1^{\text{Rating}} + \lambda_1^{\text{Movie}} = 11.588 - .3.121 - .188$

$\text{Log}(\mu_{21}) = \lambda + \lambda_2^{\text{Rating}} + \lambda_1^{\text{Movie}} = 11.588 - 3.788 - .188$

$\text{Log}(\mu_{31}) = \lambda + \lambda_3^{\text{Rating}} + \lambda_1^{\text{Movie}} = 11.588 - 2.400 - .188$

$\text{Log}(\mu_{41}) = \lambda + \lambda_4^{\text{Rating}} + \lambda_1^{\text{Movie}} = 11.588 - .645 - .188$

$\text{Log}(\mu_{51}) = \lambda + \lambda_5^{\text{Rating}} + \lambda_1^{\text{Movie}} = 11.588 - .188$

$\text{Log}(\mu_{12}) = \lambda + \lambda_1^{\text{Rating}} + \lambda_2^{\text{Movie}} = 11.588 - 3.121$

$\text{Log}(\mu_{22}) = \lambda + \lambda_2^{\text{Rating}} + \lambda_2^{\text{Movie}} = 11.588 - 3.788$

$\text{Log}(\mu_{32}) = \lambda + \lambda_3^{\text{Rating}} + \lambda_2^{\text{Movie}} = 11.588 - 2.400$

$\text{Log}(\mu_{42}) = \lambda + \lambda_4^{\text{Rating}} + \lambda_2^{\text{Movie}} = 11.588 - .645$

$\text{Log}(\mu_{52}) = \lambda + \lambda_5^{\text{Rating}} + \lambda_2^{\text{Movie}} = 11.588$

Examples of interpretation: (IF OUR MODEL WERE TRUE AND WE HAD INDEPENDENCE)

$\text{Log}(\mu_{i1}/\mu_{i2}) = \lambda_1^{\text{Movie}} - \lambda_2^{\text{Movie}} = \lambda_1^{Y} - \lambda_2^{Y} = -0.188$

does not depend on the level i of X(rating). That is, it is the same for all levels I of X(rating).

$e^{-.188} = 0.8286$ = oddsAlien/oddsWalle is the same for each rating

The odds of a movie being Alien are 0.8286 as high as the odds of a movie being Walle, for each rating.
The odds of a movie being Walle are 1/.8286 = 1.2068 times higher than the odds of a movie being Alien, for each rating.

$\text{Log}(\mu_{1j}/\mu_{5j}) = \lambda_1^{\text{Rating}} - \lambda_5^{\text{Rating}} = \lambda_1^{X} - \lambda_5^{X} = -3.121$

does not depend on the level j of Y(movie). It is the same for all levels, j, of Y. That is, it is the same for both movies.

$e^{-3.121} = .0441$ = odds Rating 1/odds Rating 5 is the same for both movies.

The odds of a rating of 1 are .0441 times as high as the odds of a rating of 5, for each movie.
The odds of a rating of 5 are 1/0.441 = 22.6757 times higher than a rating of 1, for each movie.

TWO WAY SATURATED MODEL (using movie_n, rating, and movie_n*rating)

Model is:

$\text{Log}(\mu ij) = \lambda + \lambda_i^{\text{Movie}} + \lambda_j^{\text{Rating}} + \lambda_{ij}^{\text{MovieRating}}$ j = 1,2 and i = 1,2,3,4,5

Movie_n (1 – Alien, 2-Wall-e)

Rating (12 – 1&2, 34 – 3&4, 56 - 5&6, 78 – 7&8, 910 – 9&10)

$\text{Log}(\mu_{11}) = \lambda + \lambda_1^{\text{Rating}} + \lambda_1^{\text{Movie}} + \lambda_{11}^{\text{Movie*Rating}}$

$$\text{Log}(\mu_{21}) = \lambda + \lambda_2^{\text{Rating}} + \lambda_1^{\text{Movie}} + + \lambda_{12}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{31}) = \lambda + \lambda_3^{\text{Rating}} + \lambda_1^{\text{Movie}} + + \lambda_{13}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{41}) = \lambda + \lambda_4^{\text{Rating}} + \lambda_1^{\text{Movie}} + + \lambda_{14}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{51}) = \lambda + \lambda_5^{\text{Rating}} + \lambda_1^{\text{Movie}} + \lambda_{15}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{21}) = \lambda + \lambda_1^{\text{Rating}} + \lambda_2^{\text{Movie}} + \lambda_{21}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{22}) = \lambda + \lambda_2^{\text{Rating}} + \lambda_2^{\text{Movie}} + \lambda_{22}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{32}) = \lambda + \lambda_3^{\text{Rating}} + \lambda_2^{\text{Movie}} + \lambda_{23}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{42}) = \lambda + \lambda_4^{\text{Rating}} + \lambda_2^{\text{Movie}} + \lambda_{24}^{\text{Movie*Rating}}$$
$$\text{Log}(\mu_{52}) = \lambda + \lambda_5^{\text{Rating}} + \lambda_2^{\text{Movie}} + \lambda_{25}^{\text{Movie*Rating}}$$

Note: a saturated model gives a perfect fit for any sample. Since we seek a smaller set of parameters in order to have a more parsimonious model to describe the population, a saturated model is not our goal. Commands below are included in order that students can run this model, and see that the residuals are all zero when the saturated model is fit.

COMMANDS:

Data> Weight Cases
Weight cases by:
Frequency Variable: freq
Ok
Analyze>Loglinear>General
Distribution of Cell Counts:
Poisson
Factor(s):
Rating
Movie_n
Model tab:
Choose Saturated, Continue
Save tab (just note there are no default options and cancel)
Options tab (just note the defaults and cancel)
Display defaults: Frequencies, Residuals
Plot defaults: Adjusted residuals, Normal probability for adjusted
OK

SPSS adds .5 to each observed count for the saturated model. Later, I will show you how to change this.

Cell Counts and Residuals[a,b]

		Observed		Expected			Standardized	Adjusted	
movie_n	rating	Count	%	Count	%	Residual	Residual	Residual	Deviance
1.00	12	1932.500	.6%	1932.500	.6%	.000	.000	.000	.000
	34	1760.500	.5%	1760.500	.5%	.000	.000	.000	.000
	56	8403.500	2.5%	8403.500	2.5%	.000	.000	.000	.000
	78	54233.500	16.4%	54233.500	16.4%	.000	.000	.000	.000
	910	83874.500	25.3%	83874.500	25.3%	.000	.000	.000	.000
2.00	12	6758.500	2.0%	6758.500	2.0%	.000	.000	.000	.000
	34	2702.500	.8%	2702.500	.8%	.000	.000	.000	.000
	56	9473.500	2.9%	9473.500	2.9%	.000	.000	.	.000
	78	49114.500	14.8%	49114.500	14.8%	.000	.000	.	.000
	910	113185.500	34.1%	113185.500	34.1%	.000	.000	.000	.000

a. Model: Poisson

b. Design: Constant + movie_n + rating + movie_n * rating

Note that for the saturated model, observed and expected counts are exactly the same.

THREE WAY TABLE MODEL

THREE WAY EXAMPLE: ALIEN WALL-E

For each gender, we create a cross tab tables for movie * rating.

Commands
Analysis>Descriptive Statistics >Crosstabs
Row(s): movie
Column(s): rating
Layer: sex

movie * rating * sex Crosstabulation

Count

sex			rating					Total
			12	34	56	78	910	
f	movie	Alien	430	356	1313	5075	6777	13951
		Wall-e	1104	441	1274	5998	19106	27923
	Total		1534	797	2587	11073	25883	41874
m	movie	Alien	1502	1404	7090	49158	77097	136251
		Wall-e	5654	2261	8199	43116	94079	153309
	Total		7156	3665	15289	92274	171176	289560

We will examine various models for looking at this data when we take the categorical predictors movie, gender and rating all into account.

MODEL 1: SATURATED (3 WAY INTERACTION, ALL 2 WAY INTERACTION, ALL MAIN EFFECTS)

$Log(\mu_{ijk}) = \lambda + \lambda_i^{movie} + \lambda_j^{rate} + \lambda_k^{sex} + \lambda_{jk}^{movierate} + \lambda_{ik}^{moviesex} + \lambda_{ij}^{sexrate} + \lambda_{ijk}^{movieratesex}$

$I = 1,2,\ j = 1,\ldots 5,\ k = 1,2$

Commands:
Analyze>Loglinear>General
Distribution of Cell Counts:
Poisson
Factor(s):
Movie_n
Rating

Sex_n
Model tab:
Saturated
Continue
Save tab (just note there are no default options and cancel)
Options tab (just note the defaults and cancel)
Display defaults: Frequencies, Residuals
Plot defaults: Adjusted residuals, Normal probability for adjusted
OK

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	.000	0	.
Pearson Chi-Square	.000	0	.

a. Model: Poisson
b. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n + rating * sex_n + movie_n * rating * sex_n

Cell Counts and Residuals[a,b]

movie_n	rating	sex_n	Observed Count	Observed %	Expected Count	Expected %	Residual	Standardized Residual	Adjusted Residual	Deviance
1.00	12	1.00	430.500	.1%	430.500	.1%	.000	.000	.000	.000
		2.00	1502.500	.5%	1502.500	.5%	.000	.000	.000	.000
	34	1.00	356.500	.1%	356.500	.1%	.000	.000	.000	.000
		2.00	1404.500	.4%	1404.500	.4%	.000	.000	.000	.000
	56	1.00	1313.500	.4%	1313.500	.4%	.000	.000	.000	.000
		2.00	7090.500	2.1%	7090.500	2.1%	.000	.000	.	.000
	78	1.00	5075.500	1.5%	5075.500	1.5%	.000	.000	.000	.000
		2.00	49158.500	14.8%	49158.500	14.8%	.000	.000	.	.000
	910	1.00	6777.500	2.0%	6777.500	2.0%	.000	.000	.000	.000
		2.00	77097.500	23.3%	77097.500	23.3%	.000	.000	.000	.000
2.00	12	1.00	1104.500	.3%	1104.500	.3%	.000	.000	.000	.000
		2.00	5654.500	1.7%	5654.500	1.7%	.000	.000	.000	.000
	34	1.00	441.500	.1%	441.500	.1%	.000	.000	.000	.000
		2.00	2261.500	.7%	2261.500	.7%	.000	.000	.	.000
	56	1.00	1274.500	.4%	1274.500	.4%	.000	.000	.000	.000
		2.00	8199.500	2.5%	8199.500	2.5%	.000	.000	.	.000
	78	1.00	5998.500	1.8%	5998.500	1.8%	.000	.000	.000	.000
		2.00	43116.500	13.0%	43116.500	13.0%	.000	.000	.000	.000
	910	1.00	19106.500	5.8%	19106.500	5.8%	.000	.000	.	.000
		2.00	94079.500	28.4%	94079.500	28.4%	.000	.000	.000	.000

a. Model: Poisson
b. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n + rating * sex_n + movie_n * rating * sex_n

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	11.452	.003	3512.569	.000	11.446	11.458
[movie_n = 1.00]	-.199	.005	-40.978	.000	-.209	-.190
[movie_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12]	-2.812	.014	-205.348	.000	-2.839	-2.785
[rating = 34]	-3.728	.021	-175.198	.000	-3.770	-3.686
[rating = 56]	-2.440	.012	-211.909	.000	-2.463	-2.417
[rating = 78]	-.780	.006	-134.160	.000	-.792	-.769
[rating = 910]	0[a]	.	.	.	.	.
[sex_n = 1.00]	-1.594	.008	-200.891	.000	-1.610	-1.579
[sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 12]	-1.126	.029	-38.271	.000	-1.184	-1.069
[movie_n = 1.00] * [rating = 34]	-.277	.034	-8.079	.000	-.345	-.210
[movie_n = 1.00] * [rating = 56]	.054	.017	3.175	.001	.021	.087
[movie_n = 1.00] * [rating = 78]	.330	.008	40.301	.000	.314	.346
[movie_n = 1.00] * [rating = 910]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 12]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 34]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 56]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 78]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 910]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [sex_n = 1.00]	-.837	.015	-56.012	.000	-.867	-.808
[movie_n = 1.00] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12] * [sex_n = 1.00]	-.039	.034	-1.151	.250	-.105	.027
[rating = 12] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 34] * [sex_n = 1.00]	-.039	.053	-.750	.453	-.143	.064
[rating = 34] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 56] * [sex_n = 1.00]	-.267	.031	-8.588	.000	-.328	-.206
[rating = 56] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 78] * [sex_n = 1.00]	-.378	.016	-23.789	.000	-.409	-.347
[rating = 78] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 910] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[rating = 910] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 12] * [sex_n = 1.00]	1.220	.066	18.625	.000	1.092	1.349
[movie_n = 1.00] * [rating = 12] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 34] * [sex_n = 1.00]	1.100	.080	13.698	.000	.942	1.257
[movie_n = 1.00] * [rating = 34] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 56] * [sex_n = 1.00]	1.013	.045	22.466	.000	.924	1.101
[movie_n = 1.00] * [rating = 56] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 78] * [sex_n = 1.00]	.539	.025	21.466	.000	.490	.588
[movie_n = 1.00] * [rating = 78] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 910] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 910] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 12] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 12] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 34] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 34] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 56] * [sex_n = 1.00]	0[a]	.	.	.	.	.

[movie_n = 2.00] * [rating = 56] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 78] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 78] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 910] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 910] * [sex_n = 2.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.
b. Model: Poisson
c. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n + rating * sex_n + movie_n * rating * sex_n

MODEL 2: HOMOGENOUS MODEL: Drop the 3 way interaction term, but include all 2 way interactions and main effects

$Log(\mu_{ijk}) = \lambda + \lambda_i^{movie} + \lambda_j^{rate} + \lambda_k^{sex} + \lambda_{jk}^{movierate} + \lambda_{ik}^{moviesex} + \lambda_{ij}^{sexrate}$
I = 1,2, j = 1,…5, k = 1,2

Commands:
Data> Weight Cases
Weight cases by:
Frequency Variable: freq
Ok

Analyze > Loglinear >General
General Loglinear Analysis Box:
Distribution of Cell Counts: Poisson
Factors:
movie_n
rating
sex_n

Model Button
Custom
Main effects dropdown:
Highlight factors one at a time, and click arrow to move them over to terms in model box
Movie_n
Rating
Sex_n
Interaction:
Highlight movie_n and rating, click arrow to bring over to terms in model box
Highlight movie_n and sex_n, click arrow to bring over to terms in model box

Highlight rating and sex_n, click arrow to bring over to terms in the model box
Continue

Options
Display:
Frequencies (default)
Residuals (default)
Estimates (check)
Plot:
Adjusted Residuals (default)
Normal Probability for adjusted (default)
Save: (leave defaults)
OK

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	1121.297	4	.000
Pearson Chi-Square	1162.251	4	.000

a. Model: Poisson
b. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n + rating * sex_n

Cell Counts and Residuals[a,b]

movie_n	rating	sex_n	Observed Count	Observed %	Expected Count	Expected %	Residual	Standardized Residual	Adjusted Residual	Deviance
1.00	12	1.00	430	.1%	232.873	.1%	197.127	12.918	15.219	11.540
		2.00	1502	.5%	1699.127	.5%	-197.127	-4.782	-15.219	-4.880
	34	1.00	356	.1%	232.127	.1%	123.873	8.130	10.602	7.532
		2.00	1404	.4%	1527.873	.5%	-123.873	-3.169	-10.602	-3.213
	56	1.00	1313	.4%	919.188	.3%	393.812	12.989	17.963	12.196
		2.00	7090	2.1%	7483.812	2.3%	-393.812	-4.552	-17.963	-4.593
	78	1.00	5075	1.5%	4455.475	1.3%	619.525	9.281	15.090	9.078
		2.00	49158	14.8%	49777.525	15.0%	-619.525	-2.777	-15.090	-2.783
	910	1.00	6777	2.0%	8111.336	2.4%	-1334.336	-14.816	-30.277	-15.252
		2.00	77097	23.3%	75762.664	22.9%	1334.336	4.848	30.277	4.834
2.00	12	1.00	1104	.3%	1301.127	.4%	-197.127	-5.465	-15.219	-5.612
		2.00	5654	1.7%	5456.873	1.6%	197.127	2.669	15.219	2.653
	34	1.00	441	.1%	564.873	.2%	-123.873	-5.212	-10.601	-5.422
		2.00	2261	.7%	2137.127	.6%	123.873	2.680	10.602	2.654

	56	1.00	1274	.4%	1667.812	.5%	-393.812	-9.643	-17.963	-10.066
		2.00	8199	2.5%	7805.188	2.4%	393.812	4.458	17.963	4.421
	78	1.00	5998	1.8%	6617.525	2.0%	-619.525	-7.616	-15.090	-7.739
		2.00	43116	13.0%	42496.475	12.8%	619.525	3.005	15.090	2.998
	910	1.00	19106	5.8%	17771.664	5.4%	1334.336	10.009	30.277	9.888
		2.00	94079	28.4%	95413.336	28.8%	-1334.336	-4.320	-30.277	-4.330

a. Model: Poisson

b. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n + rating * sex_n

Parameter Estimates[b,c]

					95% Confidence Interval	
Parameter	Estimate	Std. Error	Z	Sig.	Lower Bound	Upper Bound
Constant	11.466	.003	3578.338	.000	11.460	11.472
[movie_n = 1.00]	-.231	.005	-48.520	.000	-.240	-.221
[movie_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12]	-2.861	.014	-208.969	.000	-2.888	-2.835
[rating = 34]	-3.799	.021	-179.580	.000	-3.840	-3.757
[rating = 56]	-2.503	.011	-219.850	.000	-2.526	-2.481
[rating = 78]	-.809	.006	-142.636	.000	-.820	-.798
[rating = 910]	0[a]	.	.	.	.	.
[sex_n = 1.00]	-1.681	.008	-220.493	.000	-1.696	-1.666
[sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 12]	-.936	.026	-35.622	.000	-.988	-.885
[movie_n = 1.00] * [rating = 34]	-.105	.031	-3.373	.001	-.166	-.044
[movie_n = 1.00] * [rating = 56]	.189	.016	11.985	.000	.158	.219
[movie_n = 1.00] * [rating = 78]	.389	.008	50.179	.000	.374	.404
[movie_n = 1.00] * [rating = 910]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 12]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 34]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 56]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 78]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 910]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [sex_n = 1.00]	-.554	.011	-49.926	.000	-.575	-.532
[movie_n = 1.00] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12] * [sex_n = 1.00]	.247	.029	8.497	.000	.190	.304
[rating = 12] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 34] * [sex_n = 1.00]	.350	.040	8.785	.000	.272	.428
[rating = 34] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 56] * [sex_n = 1.00]	.137	.022	6.135	.000	.093	.181
[rating = 56] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 78] * [sex_n = 1.00]	-.179	.012	-14.738	.000	-.203	-.155
[rating = 78] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[rating = 910] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[rating = 910] * [sex_n = 2.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n + rating * sex_n

WE WISH TO TEST IF THE RATING, SEX, AND MOVIE ARE HOMOGENOUS.

Ho: $\lambda_{ijk}^{ratesexmovie} = 0$ (homogenous) versus Ha: saturated model, $\alpha = 0.05$

Reject Ho. (Pearson $\chi^2 = 1121.297$, df =4, pvalue <.001, Likelihood Ration = 1162.251, df =4, p-value<0.001).

We need the three way interaction term. The relationship between sex, rating and movie is not homogenous.

Model 3: ALL MAIN EFFECTS, INTERACTION TERMS FOR RATE*MOVIE AND SEX*MOVIE (RATING*SEX IS EXCLUDED)

Model is:
Log $\mu ijk = \lambda + \lambda_i^{sex} + \lambda_j^{rate} + \lambda_k^{movie} + \lambda_{ij}^{ratemovie} + \lambda_{ik}^{sexmovie}$
I = 1,2, j = 1,…5, k = 1,2

Commands:
Data> Weight Cases
Weight cases by:
Frequency Variable: freq
Ok

Analyze > Loglinear >General
General Loglinear Analysis Box:
Distribution of Cell Counts: Poisson
Factors:
movie_n
rating
sex_n

Model Button
Custom
Main effects dropdown:
Highlight factors one at a time, and click arrow to move them over to terms in model box
Movie_n
Rating
Sex_n
Interaction:
Highlight movie_n and rating, click arrow to bring over to terms in model box
Highlight movie_n and sex_n, click arrow to bring over to terms in model box
Continue

Options
Display:
Frequencies (default)
Residuals (default)
Estimates (check)
Plot:
Adjusted Residuals (default)
Normal Probability for adjusted (default)
Save: (leave defaults)
OK

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.

Likelihood Ratio	1599.934	8	.000
Pearson Chi-Square	1788.937	8	.000

a. Model: Poisson

b. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n

Cell Counts and Residuals[a,b]

			Observed		Expected			Standardized	Adjusted	
movie_n	rating	sex_n	Count	%	Count	%	Residual	Residual	Residual	Deviance
1.00	12	1.00	430	.1%	179.447	.1%	250.553	18.704	19.766	15.826
		2.00	1502	.5%	1752.553	.5%	-250.553	-5.985	-19.762	-6.137
	34	1.00	356	.1%	163.472	.0%	192.528	15.058	15.904	13.003
		2.00	1404	.4%	1596.528	.5%	-192.528	-4.818	-15.898	-4.921
	56	1.00	1313	.4%	780.484	.2%	532.516	19.061	20.598	17.346
		2.00	7090	2.1%	7622.516	2.3%	-532.516	-6.099	-20.598	-6.173
	78	1.00	5075	1.5%	5037.247	1.5%	37.753	.532	.699	.531
		2.00	49158	14.8%	49195.753	14.8%	-37.753	-.170	-.699	-.170
	910	1.00	6777	2.0%	7790.350	2.4%	-1013.350	-11.481	-18.140	-11.744
		2.00	77097	23.3%	76083.650	23.0%	1013.350	3.674	18.140	3.666
2.00	12	1.00	1104	.3%	1041.227	.3%	62.773	1.945	2.156	1.926
		2.00	5654	1.7%	5716.773	1.7%	-62.773	-.830	-2.156	-.832
	34	1.00	441	.1%	416.306	.1%	24.694	1.210	1.326	1.199
		2.00	2261	.7%	2285.694	.7%	-24.694	-.517	-1.326	-.517
	56	1.00	1274	.4%	1459.536	.4%	-185.536	-4.856	-5.424	-4.965
		2.00	8199	2.5%	8013.464	2.4%	185.536	2.073	5.424	2.065
	78	1.00	5998	1.8%	7567.153	2.3%	-1569.153	-18.038	-22.970	-18.723
		2.00	43116	13.0%	41546.847	12.5%	1569.153	7.698	22.970	7.651
	910	1.00	19106	5.8%	17438.779	5.3%	1667.221	12.625	22.402	12.432
		2.00	94079	28.4%	95746.221	28.9%	-1667.221	-5.388	-22.402	-5.404

a. Model: Poisson

b. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n

Parameter Estimates[b,c]

					95% Confidence Interval	
Parameter	Estimate	Std. Error	Z	Sig.	Lower Bound	Upper Bound
Constant	11.469	.003	3656.315	.000	11.463	11.476
[movie_n = 1.00]	-.230	.005	-48.520	.000	-.239	-.221
[movie_n = 2.00]	0[a]	.	.	.	.	.
[rating = 12]	-2.818	.013	-225.062	.000	-2.843	-2.794
[rating = 34]	-3.735	.019	-191.873	.000	-3.773	-3.697
[rating = 56]	-2.481	.011	-231.923	.000	-2.502	-2.460
[rating = 78]	-.835	.005	-154.512	.000	-.845	-.824
[rating = 910]	0[a]	.	.	.	.	.

[sex_n = 1.00]	-1.703	.007	-261.736	.000	-1.716	-1.690
[sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [rating = 12]	-.952	.026	-36.357	.000	-1.004	-.901
[movie_n = 1.00] * [rating = 34]	-.129	.031	-4.165	.000	-.190	-.068
[movie_n = 1.00] * [rating = 56]	.180	.016	11.483	.000	.149	.211
[movie_n = 1.00] * [rating = 78]	.399	.008	51.683	.000	.384	.414
[movie_n = 1.00] * [rating = 910]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 12]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 34]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 56]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 78]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [rating = 910]	0[a]	.	.	.	.	.
[movie_n = 1.00] * [sex_n = 1.00]	-.576	.011	-52.282	.000	-.598	-.554
[movie_n = 1.00] * [sex_n = 2.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [sex_n = 1.00]	0[a]	.	.	.	.	.
[movie_n = 2.00] * [sex_n = 2.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.
b. Model: Poisson
c. Design: Constant + movie_n + rating + sex_n + movie_n * rating + movie_n * sex_n

WE WISH TO TEST IF THE RATING IS INDEPENDENT FROM THE RATERS SEX, GIVEN DIFFERENT MOVIES. (THAT IS, WE WISH TO TEST IF IS IT TRUE THAT THE RELATIONSHIP BETWEEN RATING AND SEX VANISHES WHEN WE HOLD MOVIE CONSTANT.)

Ho: $\lambda_{jk}^{ratesex}=\lambda_{ijk}^{ratesexmovie} = 0$ Ha: Ho not true, $\alpha = 0.05$
Reject Ho. (Pearson $\chi^2 = 1788.9$, df =8, pvalue <.001, Likelihood Ration = 1599.9, df =8, p-value<0.001). Sex and rating are not independent, given the movie, and thus sex has a significant effect on the rating of a movie. (This follows from our previous result that the relationship between sex, rating, and movie is not homogenous)

3 WAY TABLE: SUPERVISOR, WORKER SATISFACTION

Coding: manage_n (0 - bad, 1 - good)
Superv_n(0 - low, 1-high)
Worker_n(0 - low, 1 - high)

File name: Satisfactionrecode.sav

WILL DO FOUR MODELS
SATURATED MODEL (ALL MAIN, 2 WAY, AND 3 WAY EFFECTS)
HOMOGENOUS MODEL (ALL MAIN AND 2 WAY EFFECTS)
CONDITIONAL INDEPENDENCE MODEL (ALL MAIN EFFECT, MW INTERACTION AND MS INTERACTION)
INDEPENDENT MODEL (MAIN EFFECTS ONLY)

SATURATED MODEL

Data>Weight Cases
Weight Cases by: freq
Ok

Analysis>Loglinear>General
Factor(s)
Manage_n
Superv_n
Worker_n
Model:
Custom
Put manage_n, superv_n, and worker_n in as main effects terms
Put manage_n*superv_n, manage_n*worker_n, and superv_n*worker_n in as 2way interaction terms
Put manage_n*surperv_n*worker_n in as a 3 way interaction term
Continue
Options:
Display: Include Frequencies, Residuals, Design matrix and estimates
Plot: Include Adjusted Residuals, Normal probability for adjusted
Continue
Save: Predicted Values
Continue
Ok

Below you will find the Goodness of Fit Test, Cell Counts and Residuals and Parameter Estimates tables obtained from running the commands above. Because this is the saturated model, SPSS adds .5 to all cell counts. This default is specified as Delta = .5 in the Criteria area of the options window.

Goodness-of-Fit Testsa,b

	Value	df	Sig.
Likelihood Ratio	.000	0	.
Pearson Chi-Square	.000	0	.

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n + manage_n * superv_n * worker_n

Cell Counts and Residualsa,b

			Observed		Expected			Standardized	Adjusted	
manage_n	superv_n	worker_n	Count	%	Count	%	Residual	Residual	Residual	Deviance
.00	.00	.00	103.500	14.4%	103.500	14.4%	.000	.000	.	.000
		1.00	87.500	12.2%	87.500	12.2%	.000	.000	.	.000
	1.00	.00	32.500	4.5%	32.500	4.5%	.000	.000	.000	.000
		1.00	42.500	5.9%	42.500	5.9%	.000	.000	.000	.000
1.00	.00	.00	59.500	8.3%	59.500	8.3%	.000	.000	.000	.000
		1.00	109.500	15.2%	109.500	15.2%	.000	.000	.000	.000
	1.00	.00	78.500	10.9%	78.500	10.9%	.000	.000	.	.000
		1.00	205.500	28.6%	205.500	28.6%	.000	.000	.000	.000

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n + manage_n * superv_n * worker_n

Parameter Estimatesb,c

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	5.325	.070	76.342	.000	5.189	5.462
[manage_n = .00]	-1.576	.169	-9.352	.000	-1.906	-1.246
[manage_n = 1.00]	0a	.	.	.	.	.
[superv_n = .00]	-.630	.118	-5.321	.000	-.861	-.398
[superv_n = 1.00]	0a	.	.	.	.	.
[worker_n = .00]	-.962	.133	-7.253	.000	-1.222	-.702
[worker_n = 1.00]	0a	.	.	.	.	.
[manage_n = .00] * [superv_n = .00]	1.352	.221	6.109	.000	.918	1.785
[manage_n = .00] * [superv_n = 1.00]	0a	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00]	0a	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00]	0a	.	.	.	.	.
[manage_n = .00] * [worker_n = .00]	.694	.268	2.588	.010	.169	1.220
[manage_n = .00] * [worker_n = 1.00]	0a	.	.	.	.	.
[manage_n = 1.00] * [worker_n = .00]	0a	.	.	.	.	.
[manage_n = 1.00] * [worker_n = 1.00]	0a	.	.	.	.	.
[superv_n = .00] * [worker_n = .00]	.352	.209	1.689	.091	-.057	.761
[superv_n = .00] * [worker_n = 1.00]	0a	.	.	.	.	.
[superv_n = 1.00] * [worker_n = .00]	0a	.	.	.	.	.
[superv_n = 1.00] * [worker_n = 1.00]	0a	.	.	.	.	.
[manage_n = .00] * [superv_n = .00] * [worker_n = .00]	.084	.345	.243	.808	-.592	.760
[manage_n = .00] * [superv_n = .00] * [worker_n = 1.00]	0a	.	.	.	.	.
[manage_n = .00] * [superv_n = 1.00] * [worker_n = .00]	0a	.	.	.	.	.
[manage_n = .00] * [superv_n = 1.00] * [worker_n = 1.00]	0a	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00] * [worker_n = .00]	0a	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00] * [worker_n = 1.00]	0a	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00] * [worker_n = .00]	0a	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00] * [worker_n = 1.00]	0a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n + manage_n * superv_n * worker_n

If you reset Delta = 0, SPSS will return a Cell counts and Residuals table without the extra .5 added. The Goodness of Fit test results are not affected, but the Parameter Estimates will change slightly.

goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	.000	0	.
Pearson Chi-Square	.000	0	.

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n + manage_n * superv_n * worker_n

Cell Counts and Residuals[a,b]

manage_n	superv_n	worker_n	Observed		Expected		Residual	Standardized Residual	Adjusted Residual	Deviance
			Count	%	Count	%				
.00	.00	.00	103	14.4%	103.000	14.4%	.000	.000	.	.000
		1.00	87	12.2%	87.000	12.2%	.000	.000	.000	.000
	1.00	.00	32	4.5%	32.000	4.5%	.000	.000	.000	.000
		1.00	42	5.9%	42.000	5.9%	.000	.000	.000	.000
1.00	.00	.00	59	8.3%	59.000	8.3%	.000	.000	.000	.000
		1.00	109	15.2%	109.000	15.2%	.000	.000	.000	.000
	1.00	.00	78	10.9%	78.000	10.9%	.000	.000	.000	.000
		1.00	205	28.7%	205.000	28.7%	.000	.000	.000	.000

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n + manage_n * superv_n * worker_n

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	5.323	.070	76.214	.000	5.186	5.460
[manage_n = .00]	-1.585	.169	-9.360	.000	-1.917	-1.253
[manage_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = .00]	-.632	.119	-5.329	.000	-.864	-.399
[superv_n = 1.00]	0[a]	.	.	.	.	.
[worker_n = .00]	-.966	.133	-7.263	.000	-1.227	-.706
[worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [superv_n = .00]	1.360	.222	6.121	.000	.924	1.795
[manage_n = .00] * [superv_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [worker_n = .00]	.694	.270	2.574	.010	.166	1.223
[manage_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = .00] * [worker_n = .00]	.352	.209	1.684	.092	-.058	.763
[superv_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[superv_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [superv_n = .00] * [worker_n = .00]	.088	.347	.255	.799	-.591	.767
[manage_n = .00] * [superv_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [superv_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[manage_n = .00] * [superv_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00] * [worker_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n + manage_n * superv_n * worker_n

HOMOGENOUS MODEL
(INCLUDES MAIN EFFECTS AND ALL 2 WAY EFFECTS)

Data>Weight Cases
Weight cases by Frequency Variable: freq

OK

Analysis>Loglinear>General
Factor(s)
Manage_n
Superv_n
Worker_n
Model:
Custom
Put manage_n, superv_n, and worker_n in as main effects terms
Put manage_n*superv_n, manage_n*worker_n, and superv_n*worker_n in as 2way interaction terms
Continue
Options:
Display: Include Frequencies, Residuals, Design matrix and estimates
Plot: Include Adjusted Residuals, Normal probability for adjusted
Continue
Save: Predicted Values
Continue
OK

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	.065	1	.799
Pearson Chi-Square	.065	1	.799

a. Model: Poisson
b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n

Cell Counts and Residuals[a,b]

			Observed		Expected			Standardized	Adjusted	
manage_n	superv_n	worker_n	Count	%	Count	%	Residual	Residual	Residual	Deviance
.00	.00	.00	103	14.4%	102.264	14.3%	.736	.073	.255	.073
		1.00	87	12.2%	87.736	12.3%	-.736	-.079	-.255	-.079
	1.00	.00	32	4.5%	32.736	4.6%	-.736	-.129	-.255	-.129
		1.00	42	5.9%	41.264	5.8%	.736	.115	.255	.114
1.00	.00	.00	59	8.3%	59.736	8.4%	-.736	-.095	-.255	-.095
		1.00	109	15.2%	108.264	15.1%	.736	.071	.255	.071
	1.00	.00	78	10.9%	77.264	10.8%	.736	.084	.255	.084
		1.00	205	28.7%	205.736	28.8%	-.736	-.051	-.255	-.051

a. Model: Poisson

Cell Counts and Residuals[a,b]

manage_n	superv_n	worker_n	Observed Count	Observed %	Expected Count	Expected %	Residual	Standardized Residual	Adjusted Residual	Deviance
.00	.00	.00	103	14.4%	102.264	14.3%	.736	.073	.255	.073
		1.00	87	12.2%	87.736	12.3%	-.736	-.079	-.255	-.079
	1.00	.00	32	4.5%	32.736	4.6%	-.736	-.129	-.255	-.129
		1.00	42	5.9%	41.264	5.8%	.736	.115	.255	.114
1.00	.00	.00	59	8.3%	59.736	8.4%	-.736	-.095	-.255	-.095
		1.00	109	15.2%	108.264	15.1%	.736	.071	.255	.071
	1.00	.00	78	10.9%	77.264	10.8%	.736	.084	.255	.084
		1.00	205	28.7%	205.736	28.8%	-.736	-.051	-.255	-.051

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
Constant	5.327	.068	78.002	.000	5.193	5.460
[manage_n = .00]	-1.607	.148	-10.826	.000	-1.897	-1.316
[manage_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = .00]	-.642	.112	-5.757	.000	-.861	-.423
[superv_n = 1.00]	0[a]	.	.	.	.	.
[worker_n = .00]	-.979	.123	-7.955	.000	-1.221	-.738
[worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [superv_n = .00]	1.396	.171	8.189	.000	1.062	1.731
[manage_n = .00] * [superv_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [worker_n = .00]	.748	.169	4.423	.000	.416	1.079
[manage_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = .00] * [worker_n = .00]	.385	.167	2.309	.021	.058	.711
[superv_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[superv_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n + superv_n * worker_n

CONDITIONAL INDEPENDENCE MODEL (CONDITIONED ON SUPERV-WORKER) (INCLUDES MAIN EFFECTS, MW AND MS TWO WAY INTERACTION TERMS)

Data>Weight Cases

Weight cases by Frequency Variable: freq
OK

Analysis>Loglinear>General
Factor(s)
Manage_n
Superv_n
Worker_n
Model:
Custom
Put manage_n, superv_n, and worker_n in as main effects terms
Put manage_n*superv_n, and manage_n*worker_n, in as 2way interaction terms
Continue
Options:
Display: Include Frequencies, Residuals, Design matrix and estimates
Plot: Include Adjusted Residuals, Normal probability for adjusted
Continue
Save: Predicted Values
Continue
Ok

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	5.387	2	.068
Pearson Chi-Square	5.410	2	.067

a. Model: Poisson
b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n

Cell Counts and Residuals[a,b]

			Observed		Expected			Standardized	Adjusted	
manage_n	superv_n	worker_n	Count	%	Count	%	Residual	Residual	Residual	Deviance
.00	.00	.00	103	14.4%	97.159	13.6%	5.841	.593	1.601	.587
		1.00	87	12.2%	92.841	13.0%	-5.841	-.606	-1.601	-.613
	1.00	.00	32	4.5%	37.841	5.3%	-5.841	-.950	-1.600	-.976
		1.00	42	5.9%	36.159	5.1%	5.841	.971	1.600	.947
1.00	.00	.00	59	8.3%	51.033	7.1%	7.967	1.115	1.687	1.088
		1.00	109	15.2%	116.967	16.4%	-7.967	-.737	-1.687	-.745

	1.00	.00	78	10.9%	85.967	12.0%	-7.967	-.859	-1.687	-.873
		1.00	205	28.7%	197.033	27.6%	7.967	.568	1.684	.564

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	5.283	.067	78.772	.000	5.152	5.415
[manage_n = .00]	-1.695	.148	-11.440	.000	-1.986	-1.405
[manage_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = .00]	-.521	.097	-5.355	.000	-.712	-.331
[superv_n = 1.00]	0[a]	.	.	.	.	.
[worker_n = .00]	-.829	.102	-8.101	.000	-1.030	-.629
[worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [superv_n = .00]	1.464	.168	8.713	.000	1.135	1.794
[manage_n = .00] * [superv_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [superv_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = .00] * [worker_n = .00]	.875	.160	5.464	.000	.561	1.189
[manage_n = .00] * [worker_n = 1.00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [worker_n = .00]	0[a]	.	.	.	.	.
[manage_n = 1.00] * [worker_n = 1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n

INDEPENDENCE MODEL (ALL MAIN EFFECTS)

Data>Weight Cases
Weight cases by Frequency Variable: freq
OK

Analysis>Loglinear>General
Factor(s)
Manage_n
Superv_n
Worker_n
Model:
Custom
Put manage_n, superv_n, and worker_n in as main effects terms
Continue

Options:
Display: Include Frequencies, Residuals, Design matrix and estimates
Plot: Include Adjusted Residuals, Normal probability for adjusted
Continue
Save: Predicted Values
Continue
Ok

Goodness-of-Fit Tests[a,b]

	Value	df	Sig.
Likelihood Ratio	117.997	4	.000
Pearson Chi-Square	128.086	4	.000

a. Model: Poisson
b. Design: Constant + manage_n + superv_n + worker_n

Cell Counts and Residuals[a,b]

manage_n	superv_n	worker_n	Observed		Expected		Residual	Standardized Residual	Adjusted Residual	Deviance
			Count	%	Count	%				
.00	.00	.00	103	14.4%	50.286	7.0%	52.714	7.434	9.404	6.502
		1.00	87	12.2%	81.899	11.5%	5.101	.564	.793	.558
	1.00	.00	32	4.5%	50.145	7.0%	-18.145	-2.562	-3.240	-2.746
		1.00	42	5.9%	81.670	11.4%	-39.670	-4.390	-6.171	-4.845
1.00	.00	.00	59	8.3%	85.905	12.0%	-26.905	-2.903	-4.130	-3.078
		1.00	109	15.2%	139.911	19.6%	-30.911	-2.613	-4.270	-2.720
	1.00	.00	78	10.9%	85.665	12.0%	-7.665	-.828	-1.177	-.841
		1.00	205	28.7%	139.520	19.5%	65.480	5.544	9.051	5.178

a. Model: Poisson

Cell Counts and Residuals[a,b]

manage_n	superv_n	worker_n	Observed		Expected		Residual	Standardized Residual	Adjusted Residual	Deviance
			Count	%	Count	%				
.00	.00	.00	103	14.4%	50.286	7.0%	52.714	7.434	9.404	6.502
		1.00	87	12.2%	81.899	11.5%	5.101	.564	.793	.558
	1.00	.00	32	4.5%	50.145	7.0%	-18.145	-2.562	-3.240	-2.746
		1.00	42	5.9%	81.670	11.4%	-39.670	-4.390	-6.171	-4.845
1.00	.00	.00	59	8.3%	85.905	12.0%	-26.905	-2.903	-4.130	-3.078
		1.00	109	15.2%	139.911	19.6%	-30.911	-2.613	-4.270	-2.720
	1.00	.00	78	10.9%	85.665	12.0%	-7.665	-.828	-1.177	-.841
		1.00	205	28.7%	139.520	19.5%	65.480	5.544	9.051	5.178

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n

Parameter Estimates[b,c]

Parameter	Estimate	Std. Error	Z	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Constant	4.938	.067	73.791	.000	4.807	5.069
[manage_n = .00]	-.536	.077	-6.911	.000	-.687	-.384
[manage_n = 1.00]	0[a]	.	.	.	.	.
[superv_n = .00]	.003	.075	.037	.970	-.144	.149
[superv_n = 1.00]	0[a]	.	.	.	.	.
[worker_n = .00]	-.488	.077	-6.332	.000	-.639	-.337
[worker_n = 1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

b. Model: Poisson

c. Design: Constant + manage_n + superv_n + worker_n

We can use the data above to create a table of actual and expected cell counts for our models of interest. These expected cell counts are useful in obtaining estimated odds ratios, as will be illustrated below.

				Expected

Manage	Superv	Worker	Observed	Saturated	Homogenous	Condl Indep Superv-Wrkr	Independ
Bad(0)	Low(0)	Low(0)	103.00	103.00	102.26	97.16	50.29
Bad(0)	Low(0)	High(1)	87.00	87.00	87.74	92.84	81.90
Bad(0)	High(1)	Low(0)	32.00	32.00	32.74	37.84	50.15
Bad(0)	High(1)	High(1)	42.00	42.00	41.26	36.16	81.67
High(1)	Low(0)	High(1)	59.00	59.00	597.74	51.03	85.91
High(1)	Low(0)	Low(0)	109.00	109.00	108.26	116.9	139.91
High(1)	High(1)	High(1)	78.00	78.00	77.26	85.97	85.67
High(1)	High(1)	Low(0)	205.00	205.00	205.74	197.03	139.52

NOTE; THESE ARE CORRECT, AS THEY WORK IN CALCULATING ODDS.

We can check for the fitness of these models by creating a table that takes information from the Goodness of Fit tables above.

Model	Like Rat χ^2	Df	P-value	Pearson χ^2	Df	P-value
Saturated	0	0	1	0	0	1
Homogenous	0.065	1	.799	0.065	1	.799
Cond Ind (Superv-Wrkr)	5.387	2	0.068	5.410	2	0.067
Independence	117.997	4	<0.001	128.086	4	.001

We see that the conditional independent model is acceptable (with a p-value of 0.068 it is not rejected at the 5% significance level). This model has less parameters than the homogenous model, which makes it desirable.

We bring down the cell count and residual table from above and take another look at it for this conditional independent model. It does not indicate that there are any high standardized residuals to concern us.

Cell Counts and Residuals[a,b]

			Observed		Expected			Standardized	Adjusted	
manage_n	superv_n	worker_n	Count	%	Count	%	Residual	Residual	Residual	Deviance
.00	.00	.00	59	8.3%	51.033	7.1%	7.967	1.115	1.687	1.088
		1.00	109	15.2%	116.967	16.4%	-7.967	-.737	-1.687	-.745
	1.00	.00	78	10.9%	85.967	12.0%	-7.967	-.859	-1.687	-.873
		1.00	205	28.7%	197.033	27.6%	7.967	.568	1.684	.564
1.00	.00	.00	103	14.4%	97.159	13.6%	5.841	.593	1.601	.587
		1.00	87	12.2%	92.841	13.0%	-5.841	-.606	-1.601	-.613
	1.00	.00	32	4.5%	37.841	5.3%	-5.841	-.950	-1.600	-.976
		1.00	42	5.9%	36.159	5.1%	5.841	.971	1.600	.947

a. Model: Poisson

b. Design: Constant + manage_n + superv_n + worker_n + manage_n * superv_n + manage_n * worker_n

We can also test
Ho: $\lambda_{ij}^{SW} = 0$ versus Ha: $\lambda_{ij}^{SW} \neq 0$,α= 0.05
Test statistic = $-2(L_o - L_1)$ = 5.387 – 0.065 (both from output) = 5.321, df = 2-1 = 1
From Table 7 in the textbook, 0.01 <p-value <0.025
We reject Ho. The data provides evidence at the 5% significance level that this interaction term, $\lambda_{ij}^{SW} \neq 0$. It appears from here that λ_{ij}^{SW} should remain in the model.

We have contradictory results. We will have to decide whether we might prefer the parsimonious model given that the residuals do not indicate trouble.

We can also check AIC = -2(Log likelihood -#parameters) for both models, again using -2Log Likelihoods from the output. Below we can see that it is smaller for the homogenous model, thus favouring the homogenous model.

Model	AIC
Homogenous	.065 – 2(7) = -13.94
Cond Indep (Super Worker)	5.387 - 2(6) = -6.61

Fitted Odds Ratios:

Before we use the data above to look at how to obtain fitted odd ratios, we will take a bit of a background theoretical refresher to help us out.

Loglinear models use $\mu_{ij} = n\pi_{ij}$ rather than π_{ij}.

With **2 variables X and Y**, and two levels to each variable, (X=0,1 and Y=0,1), we can write

Log $\mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$, i =0,1 and j = 0,1
For our work, we have assumed that the count is a Poisson variable.
In this case, when we look at the odds we note:

$\Theta_{XY} = \frac{n_{00}n_{11}}{n_{01}n_{10}} = \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}} = \frac{\mu_{00}\mu_{11}}{\mu_{01}\mu_{10}}$ and our table of interest might be set up this way

	Y(0)	Y(1)
X(0)	n_{00}	n_{01}
X(1)	n_{10}	n_{11}

We can't go wrong with the odds if we say:
The odds of observing Y level 0 are #times higher for X level 0 than X level 1.

Now let us look at a case with **3 variables, X, Y, and Z**, each with two levels (0 and 1).

Log $\mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}$ i=0,1, j = 0,1, and k = 0,1

If we look at levels 0 and 1 separately for Z, we can have two tables of interest, viz:
Z (0)

	Y(0)	Y(1)
X(0)	n_{00}	n_{01}
X(1)	n_{10}	n_{11}

Here we look at $\text{Log}(\mu_{ij0}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_0^Z + \lambda_{ij}^{XY} + \lambda_{i0}^{XZ} + \lambda_{j0}^{YZ} + \lambda_{ij0}^{XYZ}$ i= 0,1 and j=0,1

$\Theta_{XY(0)} = \frac{n_{00(0)}n_{11(0)}}{n_{01(0)}n_{10(0)}} = \frac{\pi_{00(0)}\pi_{11(0)}}{\pi_{01(0)}\pi_{10(0)}} = \frac{\mu_{00(0)}\mu_{11(0)}}{\mu_{01(0)}\mu_{10(0)}}$

We would say that the odds of observing Y level 0 are #times higher for X level 0 than X level 1 (conditioned on Z=0).

Z (1)

	Y(0)	Y(1)
X(0)	n_{00}	n_{01}
X(1)	n_{10}	n_{11}

Here we look at $\text{Log}(\mu_{ij1}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_1^Z + \lambda_{ij}^{XY} + \lambda_{i1}^{XZ} + \lambda_{j1}^{YZ} + \lambda_{ij1}^{XYZ}$ i= 0,1 and j=0,1
With corresponding odds ratios of interest expressed as:

$\Theta_{XY(1)} = \frac{n_{00(1)}n_{11(1)}}{n_{01(1)}n_{10(1)}} = \frac{\pi_{00(1)}\pi_{11(1)}}{\pi_{01(1)}\pi_{10(1)}} = \frac{\mu_{00(1)}\mu_{11(1)}}{\mu_{01(1)}\mu_{10(1)}}$

We would say that the odds of observing Y level 0 are #times higher for X level 0 than X level 1 (conditioned on Z=1).

We can create two 2x2 tables in Y and Z if we look at levels 0 and 1 separately for X.
We can create two 2x2 tables in X and Z if we look at levels 0 and 1 separately for Y.

For our example, we are considering various loglinear Poisson models, as follows.
Saturated
Homogenous
Conditional Independent (Supervisor-Worker)
Independent
Equations for these are as follows:

Saturated : $\text{Log}\,\mu_{ijk} = \lambda + \lambda_i^M + \lambda_j^S + \lambda_k^W + \lambda_{ij}^{MS} + \lambda_{ik}^{MW} + \lambda_{jk}^{SW} + \lambda_{ijk}^{MSW}$ i=0,1, j = 0,1, and k = 0,1

Homogen : $\text{Log}\,\mu_{ijk} = \lambda + \lambda_i^M + \lambda_j^S + \lambda_k^W + \lambda_{ij}^{MS} + \lambda_{ik}^{MW} + \lambda_{jk}^{SW}$ i=0,1, j = 0,1, and k = 0,1

CondI(SW): $\text{Log}\,\mu_{ijk} = \lambda + \lambda_i^M + \lambda_j^S + \lambda_k^W + \lambda_{ij}^{MS} + \lambda_{ik}^{MW}$ i=0,1, j = 0,1, and k = 0,1

Independ : $\text{Log}\,\mu_{ijk} = \lambda + \lambda_i^M + \lambda_j^S + \lambda_k^W$ i=0,1, j = 0,1, and k = 0,1

Note that each of these equations has eight possible values, corresponding to

Manage	Superv	Worker
Bad(0)	Low(0)	Low(0)
Bad(0)	Low(0)	High(1)
Bad(0)	High(1)	Low(0)
Bad(0)	High(1)	High(1)
High(1)	Low(0)	Low(0)

High(1)	Low(0)	High(1)
High(1)	High(1)	Low(0)
High(1)	High(1)	High(1)

Now let us consider some conditional odds that might be of interest for our models.

Model	M and S (condition on W)	M and W (condition on S)	S and W (condition on M)
Saturated – level 0	$\Theta_{MS(0)}$	$\Theta_{M(0)W}$	$\Theta_{(0)SW}$
Saturated – level 1	$\Theta_{MS(1)}$	$\Theta_{M(1)W}$	$\Theta_{(1)SW}$
Homogenous	$\Theta_{MS(0)} = \Theta_{MS(1)}$	$\Theta_{M(0)W} = \Theta_{M(1)W}$	$\Theta_{(0)SW} = \Theta_{(1)SW}$
Cond. Indep (Super-Worker)	$\Theta_{MS(0)} = \Theta_{MS(1)}$	$\Theta_{M(0)W} = \Theta_{M(1)W}$	$\Theta_{(0)SW} = \Theta_{(1)SW}$
Independence	$\Theta_{MS(0)} = \Theta_{MS(1)}$	$\Theta_{M(0)W} = \Theta_{M(1)W}$	$\Theta_{(0)SW} = \Theta_{(1)SW}$

For the Homogenous and the Conditional Independence Model, the conditional odds are the same for both levels of the variable on which we condition. We prove this for the Conditional Independence Model (Super-Worker) situation above for the M and S (condition on W) situation. Other proofs are analogous.

$\text{Log } \Theta_{MS(k)} = \text{Log } (\frac{\mu_{00(k)}\mu_{11(k)}}{\mu_{01(k)}\mu_{10(k)}}) = \log\mu_{00k} + \log\mu_{11k} - \log\mu_{01k} - \log\mu_{10k} =$

$\lambda + \lambda_0^M + \lambda_0^S + \lambda_k^W + \lambda_{00}^{MS} + \lambda_{0k}^{MW}$
$+ \lambda + \lambda_1^M + \lambda_1^S + \lambda_k^W + \lambda_{11}^{MS} + \lambda_{1k}^{MW}$
$-(\lambda + \lambda_0^M + \lambda_1^S + \lambda_k^W + \lambda_{01}^{MS} + \lambda_{0k}^{MW})$
$-(\lambda + \lambda_1^M + \lambda_0^S + \lambda_k^W + \lambda_{10}^{MS} + \lambda_{1k}^{MW})$
$= \lambda_{00}^{MS} + \lambda_{11}^{MS} - \lambda_{01}^{MS} - \lambda_{10}^{MS}$

which does not depend on k.

So $\text{Log } \Theta_{MS(0)} = \text{Log } \Theta_{MS(1)}$, and exponentializing gives us $\Theta_{MS(0)} = \Theta_{MS(1)}$.

Furthermore, for us, this is equal to λ_{00}^{MS} because SPSS sets up the equation so that the only non-zero parameter estimates returned are when levels of the all categories involved in the estimated parameter are 0.

Further algebra will provide us with the following equations.

Model	M and S (condition on W)	M and W (condition on S)	S and W (condition on M)
Saturated – level 0	$\Theta_{MS(0)} = e^{\lambda_{00}^{MS} + \lambda_{000}^{MSW}}$	$\Theta_{M(0)W} = e^{\lambda_{00}^{MW} + \lambda_{000}^{MSW}}$	$\Theta_{(0)SW} = e^{\lambda_{00}^{SW} + \lambda_{000}^{MSW}}$
Saturated – level 1	$\Theta_{MS(1)} = e^{\lambda_{00}^{MS}}$	$\Theta_{M(1)W} = e^{\lambda_{00}^{MW}}$	$\Theta_{(1)SW} = e^{\lambda_{00}^{SW}}$
Homogenous	$\Theta_{MS(0)} = \Theta_{MS(1)} = e^{\lambda_{00}^{MS}}$	$\Theta_{M(0)W} = \Theta_{M(1)W} = e^{\lambda_{00}^{MW}}$	$\Theta_{(0)SW} = \Theta_{(1)SW} = e^{\lambda_{00}^{SW}}$
Cond. Indep (Super-Worker)	$\Theta_{MS(0)} = \Theta_{MS(1)} = e^{\lambda_{00}^{MS}}$	$\Theta_{M(0)W} = \Theta_{M(1)W} = e^{\lambda_{00}^{MW}}$	1
Independence	1	1	1

For our models, our output provides the estimated parameters for the 0 levels as below.

	λ	λ_0^M	λ_0^S	λ_0^W	λ_{00}^{MS}	λ_{00}^{MW}	λ_{00}^{SW}	λ_{000}^{MSW}
Saturated	5.323	-1.584	-.632	-.966	1.360	.694	.352	.088
	λ	λ_0^M	λ_0^S	λ_0^W	λ_{00}^{MS}	λ_{00}^{MW}	λ_{00}^{SW}	
Homogen	5.327	-1.607	-.642	-.979	1.396	.748	.385	

	λ	λ_0^M	λ_0^S	λ_0^W	λ_{00}^{MS}	λ_{00}^{MW}		
CondI(SW)	5.283	-1.695	-.521	-.829	1.464	.875		
	λ	λ_0^M	λ_0^S	λ_0^W				
Indep	4.938	-.536	.003	.488				

Substituting from above, we obtain, for our estimated odds, the following.

Model	M and S (condition on W)	M and W (condition on S)	S and W (condition on M)
Saturated – level 0	$\widehat{\Theta_{MS(0)}} = e^{1.36+0.088} = 4.26$	$\widehat{\Theta_{M(0)W}} = e^{.694+.088} = 2.19$	$\widehat{\Theta_{(0)SW}} = e^{.352+.088} = 1.55$
Saturated – level 1	$\widehat{\Theta_{MS(1)}} = e^{1.36} = 3.90$	$\widehat{\Theta_{M(1)W}} = e^{.694} = 2.00$	$\widehat{\Theta_{(1)SW}} = e^{.352} = 1.42$
Homogenous	$\widehat{\Theta_{MS(0)}} = \widehat{\Theta_{MS(1)}} = e^{1.396} = 4.04$	$\widehat{\Theta_{M(0)W}} = \widehat{\Theta_{M(1)W}} = e^{.748} = 2.11$	$\widehat{\Theta_{(0)SW}} = \widehat{\Theta_{(1)SW}} = e^{.385} = 1.47$
Cond. Indep (SW)	$\widehat{\Theta_{MS(0)}} = \widehat{\Theta_{MS(1)}} = e^{1.464} = 4.32$	$\widehat{\Theta_{M(0)W}} = \widehat{\Theta_{M(1)W}} = e^{.875} = 2.40$	1
Independence	1	1	1

We note that we can also check these numbers by calculating the estimated log(odds) using the appropriate 2x2 table of expected frequencies. We include the following summary table of observed and expected frequencies for our models (these numbers were obtained from the cell counts and residuals tables in our output).

				Expected			
Manage	Superv	Worker	Observed	Saturated	Homogenous	Condl Indep Superv-Wrkr	Independ
Bad(0)	Low(0)	Low(0)	103.00	103.00	102.26	97.16	50.29
Bad(0).	Low(0)	High(1)	87.00	87.00	87.74	92.84	81.90
Bad(0)	High(1)	Low(0)	32.00	32.00	32.74	37.84	50.15
Bad(0)	High(1)	High(1)	42.00	42.00	41.26	36.16	81.67
High(1)	Low(0)	Low(0)	59.00	59.00	59.74	51.03	85.91
High(1)	Low(0)	High(1)	109.00	109.00	108.26	116.90	139.91
High(1)	High(1)	Low(0)	78.00	78.00	77.26	85.97	85.67
High(1)	High(1)	High(1)	205.00	205.00	205.74	197.03	139.52

Mod	M and S (condition on W)	M and W (condition on S)	S and W (condition on M)
Sat– 0	(103x78)/(32x59)=4.26	(103x109)/(87x59) = 2.19	(103x42)/(32x87) = 1.554
Sat – 1	(87x205)/(42x109)=3.90	(32x205)/(42x78)=2.00	(59x205)/(78x109) = 1.423
Homo	(102.26x77.26)/(32.74x59.74)=4.04 (87.74x205.74)/(41.26x108.26)=4.04	(102.26x108.26)/(87.74x59.74)=2.11 (32.74x205.74)/(41.26x77.26)=2.11	(102.26x41.26)/(87.74x32.74)=1.47 (59.74x205.75)/(108.26x77.26)=1.47
Cond In(SW)	(97.16x85.97)/(37.84x51.03)=4.32 (92.84x197.03)/(36.16x116.90)=4.32	(97.16x116.9)/(92.84x51.03) = 2.40 (37.84x197.03)/(36.16x85.97)=2.40	(97.16x36.16)/(92.84x37.84)=1.00 (51.03x197.03)/116.90x85.97)= 1.00
Indep	(50.29x85.67)/(50.15x85.91) = 1.00 (81.90x139.52)/(81.67x139.91)=1.00	(50.29x139.91)/(81.90x85.91)=1.00 (50.15x139.52)/(81.67x85.67)=1.00	(50.29x81.67)/(81.90x50.15)=1.00 (85.91x139.52)/(139.91x85.67)=1.00

We illustrate how the entry is obtained for one of the cells above. Other entries are obtained in an analogous manner.

Consider $\widehat{\Theta_{M(0)W}}$ for our Homogenous Model.

For S =0, our 2x2 table of expected frequencies follows.

	W(0)	W(1)
M(0)	102.26	87.74
M(1)	59.74	108.26

$\widehat{\Theta_{MS(0)}}$= (103)(78)/(59)(32)