SPSS TRAINING SESSION 2

STATISTICAL ANALYSIS

(SPSS 16.0)

Sun Li Centre for Academic Computing <u>lsun@smu.edu.sg</u>

OUTLINE

- Elementary Data Analysis
- Group Comparison & One-way ANOVA
 ANOVA
- Non-parametric Tests
- Correlations
- General Linear Regression
- Logistic Models
 - Binary Logistic Model
 - Ordinal Logistic Model
 - Multinomial Logistic Model

The Explore procedure

- Exploratory data analysis
 - Summary statistics
 - Distribution plots

Normality plots with tests

• The dependent variable must be a scale variable, while the grouping variables may be ordinal or nominal.

Analyze

Descriptive Statistics

Explore

🛃 Explore	
Age in years [age] Price of primary vehicle [car] Primary vehicle price category [carcat] Parts with current employer [employ] Parts with current employer [empcat] Vears with current employer [empcat] OK Paste Cancel Help	Statistics Plots Options

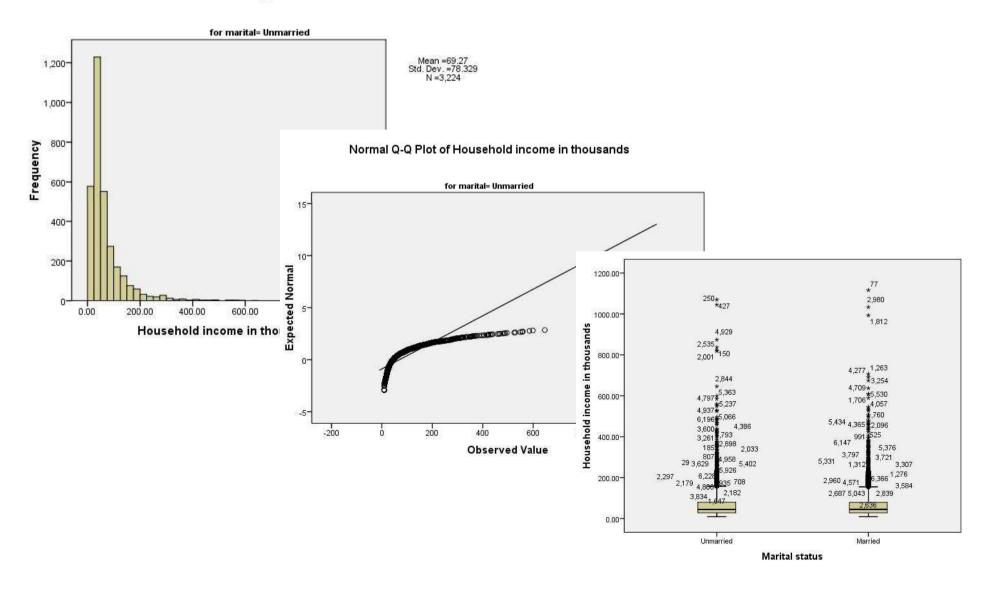
E.g.: *demo.sav*

Explore the household income for married and unmarried people.

	😤 Explore: Plots 🛛 🔀
 Explore: Statistics Descriptives Confidence Interval for Mean: 95 % M-estimators Outliers 	Boxplots Descriptive Factor levels together Dependents together Mone Stem-and-leaf Histogram Histogram Normality plots with tests
Continue Cancel Help	Spread vs Level with Levene Test None Power estimation Iransformed Power: Natural log
	Continue Cancel Help

Tests of Normality							
		Kolmogorov-Smirnov ^a Shapiro				napiro-Wilk	
	Marital status	Statistic	df	Siq.	Statistic	df	Siq.
Household income in	Unmarried	.221	3224	.000	.600	3224	.000
thousands	Married	.222	3176	.000	.602	3176	.000
a. Lilliefors Significance Correction							

Histogram



The Crosstabulation table

• Analysis of cross-classifications

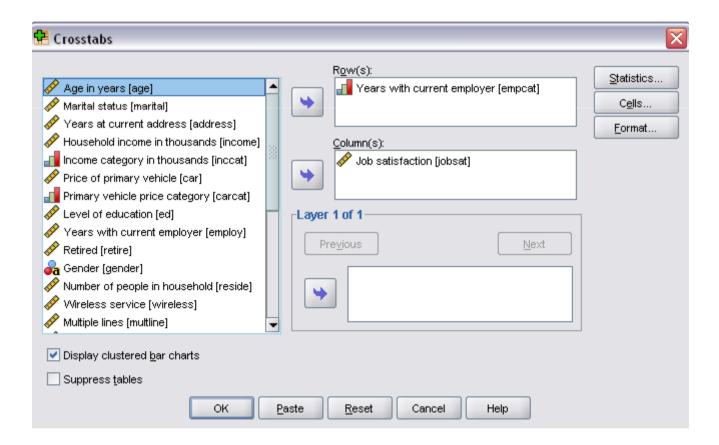
Analyze Descriptive Statistics Crosstabs

- To examine the relationship btw two categorical variables
 - Nominal-by-nominal relationships
 - Ordinal-by-ordinal relationships
 - Nominal-by-interval relationships
 - Relative risk measurement
 - Agreement measurement

🚰 Crosstabs: Statistics						
Chi-square	Correlations					
Nominal	Cordinal					
Contingency coefficient	<u>G</u> amma					
Phi and Cramer's V	Somers' d					
Lambda	Kendall's tau- <u>b</u>					
Uncertainty coefficient	Kendall's tau- <u>c</u>					
Nominal by Interval	<u>K</u> appa					
Eta	Risk					
	McNemar					
Cochr <u>a</u> n's and Mantel-Haenszel statistics Test common odds ratio equals:						
Continue Cancel Help						

E.g.:

Test and measure the relationship btw the job satisfactions and the number of year with current employer.



	Chi-Square rests						
	Value	df	Asymp. Sig. (2-sided)				
Pearson Chi-Square	1.690E3	8	.000				
Likelihood Ratio	1747.380	8	.000				
Linear-by-Linear Association	1525.767	1	.000				
N of Valid Cases	6400						

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 315.37.

Chi Squaro Tosts

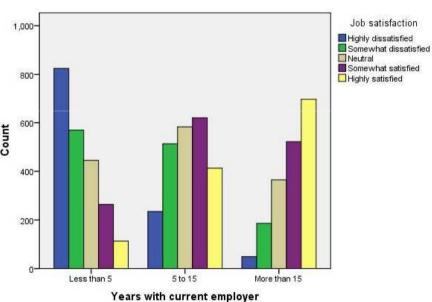
	Directional Measures						80
			Value	Asymp. Std. Errorª	Approx.		
Ordinal by Ordinal	Somers' d	Symmetric	.418	.009	47.6	322	60
		Years with current employer Dependent	.382	.008	47.6	ount	
		Job satisfaction Dependent	.461	.010	47.6	0	40

a. Not assuming the null hypothesis.

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

Symmetric Measures

		Value	Asymp. Std. Error ^a	Approx. T ^b	Approx. Sig.
Ordinal by Ordinal	Kendall's tau-b	.420	.009	47.655	.000
	Kendall's tau-c	.458	.010	47.655	.000
	Gamma	.560	.011	47.655	.000
	N of Valid Cases	6400			



Bar Chart

a. Not assuming the null hypothesis.

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

<u>T Tests</u>

- The one-sample T test
- The paired-samples T test (*skip*)
- The independent-samples T test

Analyze Compare Means

Ī	<u>A</u> nalyze	<u>G</u> raphs	<u>U</u> tilities	Add	- <u>o</u> ns	<u>W</u> in	idow	<u>H</u> el	р		
	Repor	ts		►			÷	-			7
1	D <u>e</u> scr	riptive Statis	tics	►	Г						
	Ta <u>b</u> le:	s		►	ati	stics					
1	Comp	are Means		→	М	<u>M</u> eans					Ì
	<u>G</u> enei	ral Linear M	odel	→	t	One- <u>S</u>	ample	T Tes	t		-
1	Gener	rali <u>z</u> ed Line:	ar Models	►	A-B	Indepe	ndent	-Samp	les T	Test	
-	Mi <u>×</u> ed	Models		→	A1-A2	<u>P</u> aired-	-Samp	les T	Test		-
	<u>C</u> orre	late		→	A	<u>O</u> ne-W	/ay Al	NOVA			
	<u>R</u> egre	ession		•							
	L <u>og</u> lin	ear		►	Ŀ		Leve	ene's (Test /ariar	for Ec	quality o
	Class	i <u>f</u> y		►	Ŀ				anai	1000	
	<u>D</u> ata f	Reduction		►	Ŀ						
	Sc <u>a</u> le			►	Ŀ			F			Siq.
9	<u>N</u> onpa	arametric Te	ests	►	Г			.0	37		.8
	Time S	Series		►	Ŀ						
-	<u>S</u> urviv	val		►	⊢						
	🔀 Missir	ng Value Ar	al <u>⊻</u> sis		Ŀ						
1	M <u>u</u> ltip	le Respons	е	•	L						
	<u>Q</u> ualit	y Control		•	L						
	🚺 ROC (Cur <u>v</u> e									

E.g.:

- 1. Test if the average household income equals to 75k.
- 2. Test if the average household income for married and unmarried people has no significant difference.

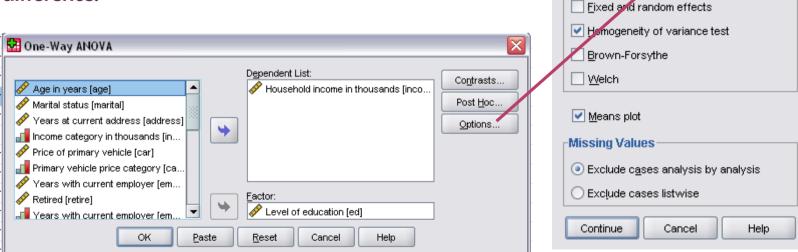
🔛 One-Sample T Test	
✓ Age in years [age] ✓ ✓ Marital status [marital] ✓ ✓ Years at current add Income category in t ✓ Price of primary vehic ✓ ✓ Primary vehicle price ✓ ✓ Level of education [ed] ✓ ✓ Years with current e ✓ ✓ Retired Iretire1 ✓ ✓ OK Paste Reset Cancel Help	Independent-Samples T Test Age in years [age] Years at current add Income category in t Primary vehicle price Primary vehice<

One-way analysis of variance

• to test the hypothesis that the means of two or more groups are not significantly different.

E.g.:

Test if the average household income for the five education groups are different. If there is significant difference, identify which groups have the major difference.

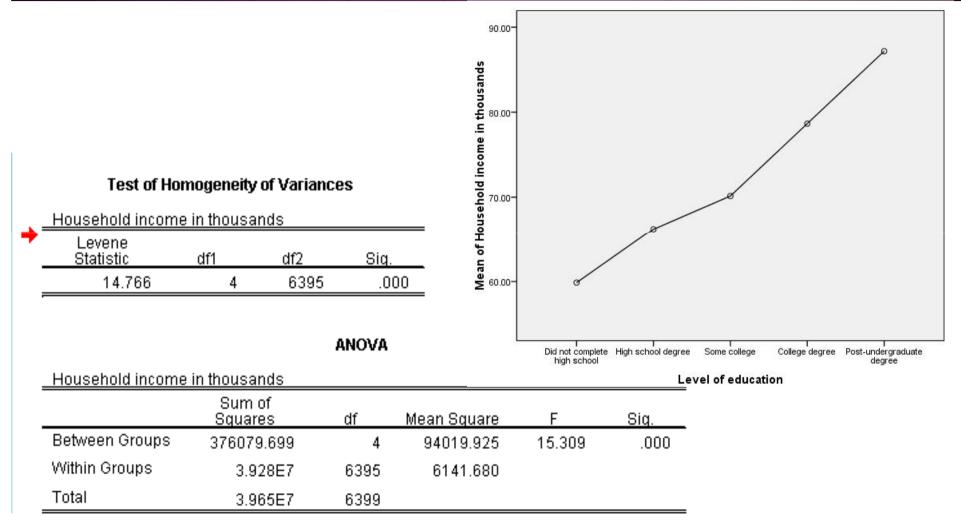


X

🚰 One-Way ANOVA: Options

-Statistics

Descriptive



Multiple Comparisons

뜸 One-Way ANOVA: Post Hoc Multiple Compariso	Household income in tho Tamhane	usands					
┌Equal Variances Assumed						95% Confid	ence Interval
SD S-N-K Waller-Dun	(1) Level of education	(J) Level of education	Mean Difference (I- J)	Std. Error	Sig.	Lower Bound	Upper Bound
	Did not complete high school	High school degree	-6.34094	2.33139	.064	-12.8724	.1905
Sidak 🗌 Tukey's-b 🗌 Dunnett	301001	Some college	-10.26837*	2.74450	.002	-17.9587	-2.5781
Scheffe Duncan Control Categ		College degree	-18.78400*	3.01158	.000	-27.2233	-10.3447
R-E-G-WF Hochberg's GT2 Test		Post-undergraduate degree	-27.30373*	5.26659	.000	-42.1229	-12.4846
R-E-G-W Q Gabriel 2-sided	High school degree	Did not complete high school	6.34094	2.33139	.064	1905	12.8724
r Equal Variances Not Assumed		Some college	-3.92743	2.74970	.811	-11.6318	3.7769
		College degree	-12.44306*	3.01632	.000	-20.8952	-3.9909
Tamhane's T2 Dunnett's T3 Games-How		Post-undergraduate degree	-20.96279*	5.26930	.001	-35.7894	-6.1362
Significance level: 0.05	Some college	Did not complete high school	10.26837*	2.74450	.002	2.5781	17.9587
Continue Cancel		High school degree	3.92743	2.74970	.811	-3.7769	11.6318
		College degree	-8.51563	3.34591	.105	-17.8907	.8594
		Post-undergraduate degree	-17.03536	5.46465	.019	-32.4016	-1.6691
	College degree	Did not complete high school	18.78400 [*]	3.01158	.000	10.3447	27.2233
		High school degree	12.44306 [*]	3.01632	.000	3.9909	20.8952
		Some college	8.51563	3.34591	.105	8594	17.8907
		Post-undergraduate degree	-8.51973	5.60355	.749	-24.2704	7.2309
	Post-undergraduate degree	Did not complete high school	27.30373*	5.26659	.000	12.4846	42.1229
		High school degree	20.96279 [*]	5.26930	.001	6.1362	35.7894
		Some college	17.03536*	5.46465	.019	1.6691	32.4016
		College degree	8.51973	5.60355	.749	-7.2309	24.2704

*. The mean difference is significant at the 0.05 level.

Non-parametric tests

- Two-independent samples tests
- Tests for several independent samples

Analyze

Nonparametric Tests 2 Independent Samples K Independent Samples

<u>A</u> nalyze	<u>G</u> raphs	<u>U</u> tilities	Add	l- <u>o</u> ns	<u>W</u> indo	W	<u>H</u> elp
Report	S		•		è 🌑		
D <u>e</u> scrip	otive Statist	ics	►			La	bel
Ta <u>b</u> les			►	je in	years		
Compa	re Means		►	arita	l status	З	
<u>G</u> enera	al Linear Mo	del	►	ears	at curr	ent a	address
Genera	ali <u>z</u> ed Linea	r Models	►	puse	hold in	com	e in thousands
Mi <u>×</u> ed №	Models		►	com	e categ	aory	in thousands
<u>C</u> orrela	ite		►	ice (of prima	ary v	ehicle
<u>R</u> egres	sion		►	imar	ry vehic	cle p	rice category
Logline	ar		►	vel d	of educ	atior	ı
Classi <u>f</u>	у		►	ears	with cu	urren	t employer
<u>D</u> ata Re	eduction		►	etire	d		
Sc <u>a</u> le			►	ears with current employ		t employer	
<u>N</u> onpar	ametric Te	sts	►	X2 <u>C</u>	<u>hi-Squa</u>	re	
Time Se	eries		►	0/1 E	<u>)</u> inomial.		
<u>S</u> urvival		►	AAAB E	<u>l</u> uns			
🔀 Missing Value Anal <u>y</u> sis			🔼 <u>1</u>	-Sample	K-S.		
Multiple Response		►	<u> </u>	Indeper	ndent	Samples	
Quality Control		►	🔛 н	(Indepei	ndent	Samples	
🖊 ROC CI	ur <u>v</u> e			a 2	Related	Sam	ples
4		0		M K	Related	l Sam	ples

Two-independent samples tests:

Test if the household income varies btw married and unmarried people.

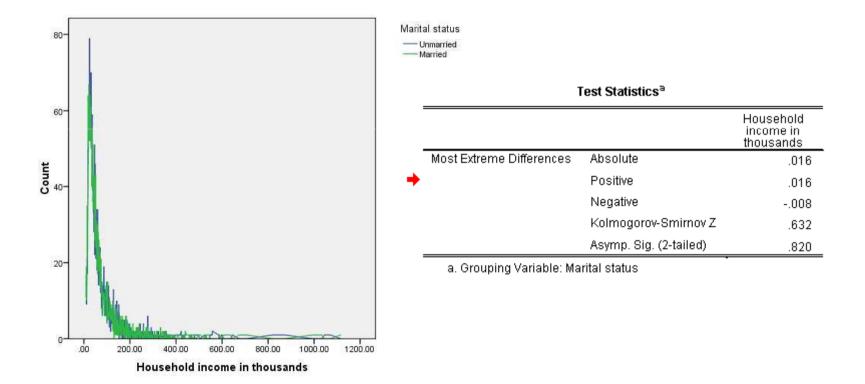
Two-Independent-Samples Tests Test Variable List: Option Option	
Age in years [age] Age in years [address] Income category in thousands [incode] Price of primary vehicle [car] Primary vehicle price category [carcat]	Mann-Whitne Wilcoxon test
 Comparison [ed] Years with current employer [employ] Retired [retire] Years with current employer [empcat] 	
Test Type	Test Statistics ^a
Mann-Whitney U Kolmogorov-Smirnov Z Moses extreme reactions Wald-Wolfowitz runs	Hi ir th
OK Paste Reset Cancel Help	Mann-Whitney U 51
n	 Wilcoxon W

Whitney and on tests--

	Household income in thousands
Mann-Whitney U	5108032.500
Wilcoxon W	1.031E7
Z	158
Asymp. Sig. (2-tailed)	.874

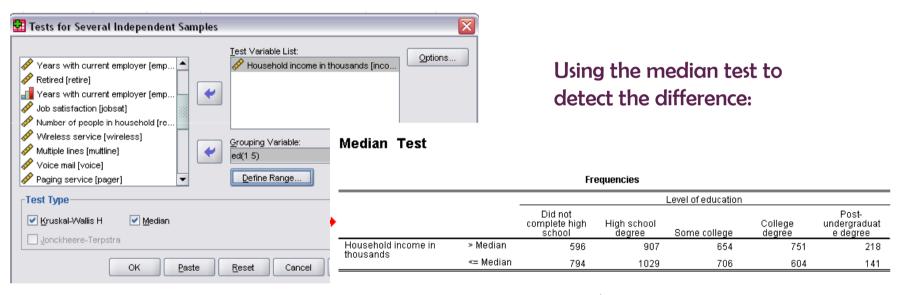
a. Grouping Variable: Marital status

The two-sample Kolmogorov-Smirnov test



K-independent samples tests:

Test if the household income varies among people with different educations.



	Household income in thousands
N	6400
Median	45.0000
Chi-Square	66.957ª
df	4
Asymp. Sig.	.000

Test Statistics^b

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 175.3.

b. Grouping Variable: Level of education

Using Kruskal-Wallis to Test Ordinal Outcomes

Kruskal-Wallis Test

Ranks				
	Level of education	N	Mean Rank	
Household income in thousands	Did not complete high school	1390	2923.64	
	High school degree	1936	3117.33	
	Some college	1360	3195.49	
	College degree	1355	3460.83	
	Post-undergraduate degree	359	3757.41	
	Total	6400		

	Test Statistics ^{a,b}
	Household income in thousands
Chi-Square	94.672
df	4
Asymp. Sig.	.000

a. Kruskal Wallis Test

b. Grouping Variable: Level of education

CORRELATION

Correlation

Analyze Correlation

- Bivariate correlation:
 - Describe relationship btw two variables.
- Partial correlation:
 - Describe relationship btw two variables while controlling for the effects of one or more additional variables.
- Distances: (skip)
 - Similarities/dissimilarities btw pairs of variables/cases.

CORRELATION

E.g.:

Compute the correlation coefficient btw the household income and the price of primary vehicle.

Bivariate Correlations Image: Status Veriables: Image: Age in years Image: Status Image: Marital status Image: Status Image: Primary vehicle price Status Image: Primary vehicle price Status Image: Primary vehicle price Status Image: Price of education For the status Image: Price of education Price Image: Price of education <th></th> <th>etric Correlations C:\Program Files\SPSS</th> <th>Inc\SPSS16\Samples\</th> <th>demo.sav</th> <th></th>		etric Correlations C:\Program Files\SPSS	Inc\SPSS16\Samples\	demo.sav	
OK Paste Reset Cancel Help			Correlations		
				Household income in thousands	Price of primary vehicle
	Spearman's rh		Correlation Coefficient	1.000	.998**
	_	thousands	Sig. (2-tailed)		.000
	-		N	6400	6400
		Price of primary vehicle	Correlation Coefficient	.998**	1.000
			Sig. (2-tailed)	.000	
			N	6400	6400
	**. Correlatio	on is significant at the 0.01 leve	el (2-tailed).		

CORRELATION

E.g.:

Compute the correlation coefficient btw the household income and the price of primary vehicle after controlling factor age, marriage status and education.

Years at current add		ns	Corre	lations				
Income category in t Primary vehicle price Years with current e	of p			Household income in thousands	Price of primary vehicle	Age in years	Marital status	Level of education
Retired [retire] Controlling	for: -none-a	Household income in thousands	Correlation	1.000	.792	.335	.003	.09
📲 Years with current e 🛛 🔗 Age in	ı ye	linuusanus	Significance (2-tailed)		.000	.000	.836	.00
🔗 Job satisfaction (jobs 🛛 💊 Marita			df	0	6398	6398	6398	639
🞸 Number of people in 🔄 🛄 💞 Level	ofe	Price of primary vehicle	Correlation	.792	1.000	.376	002	.10
🖋 Wireless service Iwi 💌	_		Significance (2-tailed)	.000		.000	.846	.00
Test of Significance			df	6398	0	6398	6398	639
Two-tailed One-tailed		Age in years	Correlation	.335	.376	1.000	.003	10
			Significance (2-tailed)	.000	.000		.812	.00
Display actual significance level			df	6398	6398	0	6398	639
OK Paste Reset		Marital status	Correlation	.003	002	.003	1.000	03
			Significance (2-tailed)	.836	.846	.812		.01
			df	6398	6398	6398	0	639
		Level of education	Correlation	.096	.102	126	030	1.00
			Significance (2-tailed)	.000	.000	.000	.016	
			df	6398	6398	6398	6398	
	Age in years & Marital status & Level of	Household income in thousands	Correlation	1.000	.757			
	education	lilousailus	Significance (2-tailed)		.000			
			df	0	6395			
		Price of primary vehicle	Correlation	.757	1.000			
			Significance (2-tailed)	.000				
			df	6395	0			

The GLM Univariate procedure

Analyze

General Linear Model Univariate Analyze Regression Linear

In SPSS, there are two menus about linear regressions under pull-down menu "Analyze". The difference btw the two menus is that "Linear" function under "Regression" treats every predictor to be continuous variables, while in "General Linear Model", there are options to define categorical predictors and continuous predictors.

- Factors: Categorical predictors should be selected as factors in the model.
 - Fixed-effects factors are generally thought of as variables whose values of interest are all represented in the data file.
 - Random-effects factors are variables whose values in the data file can be considered a random sample from a larger population of values. They are useful for explaining excess variability in the dependent variable.
- Covariates: Scale predictors should be selected as covariates in the model.

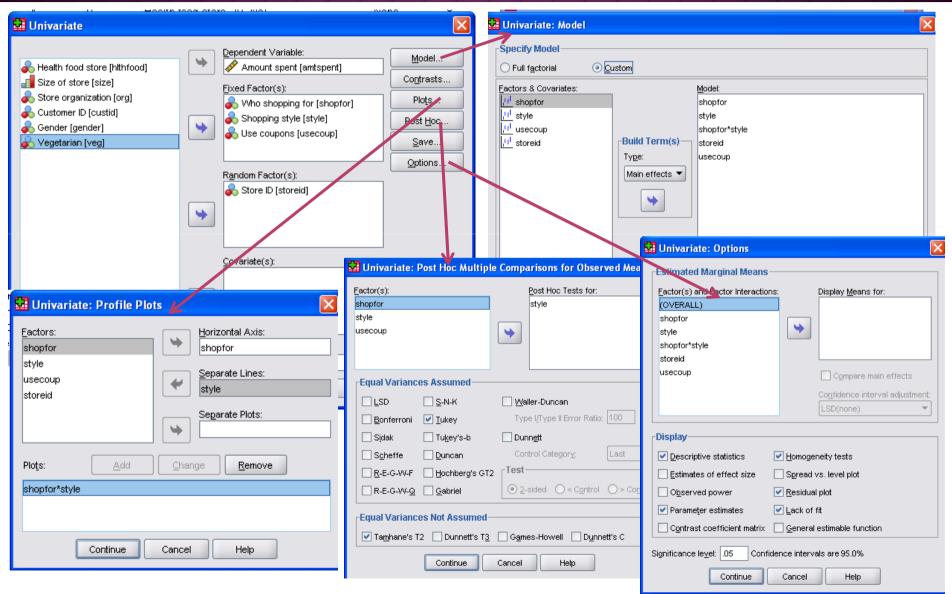
E.g.: grocery_1month.sav

Use the GLM Univariate procedure to fit a model with fixed and random effects to the amounts customers spent in grocery stores.

E.g.:

A grocery store chain surveyed a set of customers concerning their purchasing habits. Given the survey results and how much each customer spent in the previous month, the store wants to see the factors.

Variable	Variable information
storeid	Store id
shop for	Who shop for: 1 = self 2 = self and spouse 3 = self and family
style	Shopping style: 1 = biweekly, in bulk 2 = weekly, similar items 3 = often, what's on sale
usecoup	Use coupons: 1 = no 2 = from newspaper 3 = from mailings 4 = from both
amtspent	Amount spent last month



Levene's Test of Equality of Error Variances^a Dependent Variable:Amount spent F df1 df2 Sig. .650 320 30 .961 Tests the null hypothesis that the error variance of the dependent variable is equal across groups. a. Design: Intercept + shopfor + style + shopfor * style + storeid + usecoup

Tests of Between-Subjects Effects

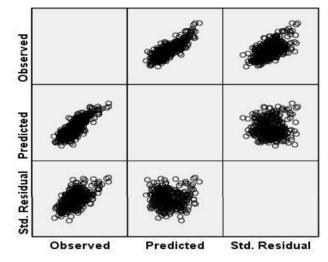
Dependent Variable:Amount spent

Source		Type III Sum of Squares	df	Mean Square	F	Siq.
Intercept	Hypothesis	3.362E7	1	3.362E7	5219.104	.000
	Error	672713.222	104.418	6442.505ª		
shopfor	Hypothesis	761142.660	2	380571.330	94.249	.000
	Error	1130619.470	280	4037.927 ^b		
style	Hypothesis	64332.117	2	32166.059	7.966	.000
	Error	1130619.470	280	4037.927 ^b		
shopfor * style	Hypothesis	12766.092	4	3191.523	.790	.532
	Error	1130619.470	280	4037.927 ^b		
storeid	Hypothesis	479152.932	59	8121.236	2.011	.000
	Error	1130619.470	280	4037.927 ^b		
usecoup	Hypothesis	191472.817	3	63824.272	15.806	.000
	Error	1130619.470	280	4037.927 ^b		

a. .589 MS(storeid) + .411 MS(Error)

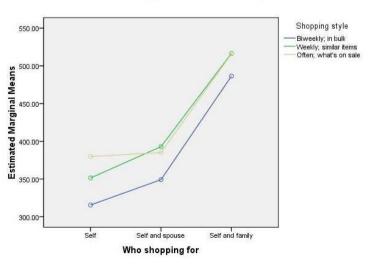
b. MS(Error)

Dependent Variable: Amount spent



Model: Intercept + shopfor + style + shopfor * style + storeid + usecoup

Estimated Marginal Means of Amount spent



- Hays, W. L. 1981. Statistics, 3rd ed. New York: Holt, Rinehart, and Winston.
- Brown, M. B., and A. B. Forsythe. 1974b. Robust tests for the equality of variances. Journal of the American Statistical Association, 69:, 364-367.
- Milliken, G., and D. Johnson. 1992. Analysis of Messy Data: Volume 1. Designed Experiments. New York: Chapman & Hall.
- Neter, J., W. Wasserman, and M. H. Kutner. 1990. Applied Linear Statistical Models, 3rd ed. Homewood, III.: Irwin.
- Siegel, S., and N. J. Castellan. 1988. Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill, Inc..
- Conover, W. J. 1980. Practical Nonparametric Statistics, 2nd ed. New York: John Wiley and Sons.
- Horton, R. L. 1978. The General Linear Model. New York: McGraw-Hill, Inc..

LOGISTIC MODELS

Logistic Models

Binary logistic model: dichotomous response outcomes
 e.g.: presence or absence of an event

 $\pi_i = E(y_i \mid x_i) \qquad \text{logit}(\pi) = \log(\pi/(1-\pi)) = g(\pi) = \alpha + \beta' X$

• Ordinal logistic model: ordinal response variable with more than two ordered categories

e.g.: a 5-point Likert scale

 $g(\Pr(Y \le i \mid X)) = \alpha_i + \beta' X, \quad i = 1, \dots, k$

• Multinomial logistic model: nominal response variables with more than two categories

e.g.: different types of programs in school

$$\log\left(\frac{\Pr(Y=i \mid X)}{\Pr(Y=k+1 \mid X)}\right) = \alpha_i + \beta'_i X, \quad i = 1, \dots, k$$

Using binary logistic regression to assess credit risk

Analyze Regression Binary Logistic

E.g.: bankloan.sav

If you are a loan officer at a bank, then you want to be able to identify characteristics that are indicative of people who are likely to default on loans, and use those characteristics to identify good and bad credit risks.

We have information on 850 past and prospective customers. The first 700 cases are customers who were previously given loans. Use these 700 customers to create a logistic regression model. Then use the model to classify the 150 prospective customers as good or bad credit risks.

Variable name	Variable information		
age	Age in years		
ed	Level of education 1= didn't complete high school 2= high school degree 3= college degree 4= undergraduate 5= postgraduate		
employ	Years with current employer		
address	Years in current address		
income	Household income in thousands		
debtinc	Debt to income ratio (*100)		
creddebt	Credit card debt in thousands		
othdebt	Other debts in thousands		
default	Previously defaulted 1= Yes O = No		

Step 1: select the first 700 obs for logistic model

Data Select Cases

Age in years [age] Level of education [ed] Years with current employer [employ] Years at current address [address] Household income in thousands [income] Debt to income ratio (×100) [debtinc] Credit card debt in thousands [creddebt] Other debt in thousands [othdebt] Previously defaulted [default]	Select All cases If gondition is satisfied [f] Rangom sample of cases Sample Based on time or case range Range Use filter variable: First Case Last Case Observation: 1 700 Output Continue Cancel Help Filter out unselected cases Copy selected cases to a new dataset Dataset name:
	O Delete unselected cases

Step 2: construct logistic model

				Johnada	
Logistic Regression	X		Probabilities	<u>U</u> nstandardized	
		1	Group membership	Logit	
🔗 Age in years [age]	Dependent:			Studentized	
Age in years [age] Level of education [ed]	Previousiy detautted [detautt]			-	
Years with current employer [employ]	Block 1 of 1		Cook's	Sta <u>n</u> dardized	
Years at current address [address]	Previous Devices		Leverage values	De <u>v</u> iance	
Ausehold income in thousands [income]	Covariates:		DfBeta(s)		
P Debt to income ratio (x100) [debtinc]	age				
Credit card debt in thousands [creddebt] Other debt in thousands [othdebt]	ed ed		Export model information t	o XML file	
	employ address			Browse	
	>a*b> income		Include the covariance mat	trix	
	Method: Forward: LR		Continue Cancel	Help	
	Selection Varia <u>b</u> le:				
	Rule		- · · · · · · · · · · · · · · · · · · ·		
			Logistic Regression: Options		
ок	Paste Reset Cancel Help		Statistics and Plots		
		- 1	Classification plots	Correlations of estimate	
Logistic Regression: De	efine Categorical Variables 🖌 🔀				~~
			Mosmer-Lemeshow goodness-o	of-fit 📃 Iteration history	
Covariates:	Categorical Covariates: ed(Indicator)		Casewise listing of residuals	CI for exp(B): 95	%
Years with current empl	eu(muicator)		Outliers outside 2 std.	dou	
Age in years [age]				uev.	
Credit card debt in thou	★		O <u>A</u> ll cases		
Other debt in thousands			-Display		
Debt to income ratio (x1 Years at current addres			In the each step ○ At last step		
Years at current addres					
	Change Contrast		Probability for Stepwise	Classification cutoff:	0.5
	Contrast: Indicator 💌 Change		Entry: 0.05 Removal: 0.10	Maximum Iterations:	20
	Reference Category: Last Eirst				20
			Include constant in model		
Co	ntinue Cancel Help		Continue	ancel Help	
			Continue		

🔜 Logistic Regression: Save

-Predicted Values

1

×

-Residuals

				Model Summary
	Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
	1	701.429ª	.137	.200
	2	631.083 ^b	.219	.321
•	3	575.636 ^b	.279	.408
	4	556.732°	.298	.436

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.160	8	.924
2	4.158	8	.843
3	6.418	8	.600
4	8.556	8	.381

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.

c. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

														95.0% C.I.	for EXP(B)
		Classification Tal	ble ^a					В	S.E.	Wald	df	Siq.	Exp(B)	Lower	Upper
				Predicte	d	Step 1ª	debtinc	.132	.014	85.377	1	.000	1.141	1.109	1.173
			P	reviously det	faulted		Constant	-2.531	.195	168.524	1	.000	.080		
	Observed		No	Yes	Percentage Correct	Step 2 ^b	employ	141	.019	53.755	1	.000	.868	.836	.902
Step 1	Previously defaulted	No	490	27	94.8		debtinc	.145	.016	87.231	1	.000	1.156	1.122	1.192
		Yes	137	46	25.1		Constant	-1.693	.219	59.771	1	.000	.184		
		Overall Percentage			76.6	Step 3°	employ	244	.027	80.262	1	.000	.783	.743	.826
Step 2	Previously defaulted	No	481	36	93.0		debtinc	.088	.018	23.328	1	.000	1.092	1.053	1.131
		Yes	110	73	39.9		creddebt	.503	.081	38.652	1	.000	1.653	1.411	1.937
		Overall Percentage			79.1		Constant	-1.227	.231	28.144	1	.000	.293		
Step 3	Previously defaulted	No	477	40	92.3	Step 4 ^d	employ	243	.028	74,761	1	.000	.785	.743	.829
		Yes	99	84	45.9		address	081	.020	17.183	1	.000	.100	.887	.958
		Overall Percentage			80.1						1				
Step 4	Previously defaulted	No	478	39	92.5		debtinc	.088	.019	22.659	1	.000	1.092	1.053	1.133
		Yes	91	92	50.3		creddebt	.573	.087	43.109	1	.000	1.774	1.495	2.104
		Overall Percentage			81.4		Constant	791	.252	9.890	1	.002	.453		

Variables in the Equation

a. The cut value is .500

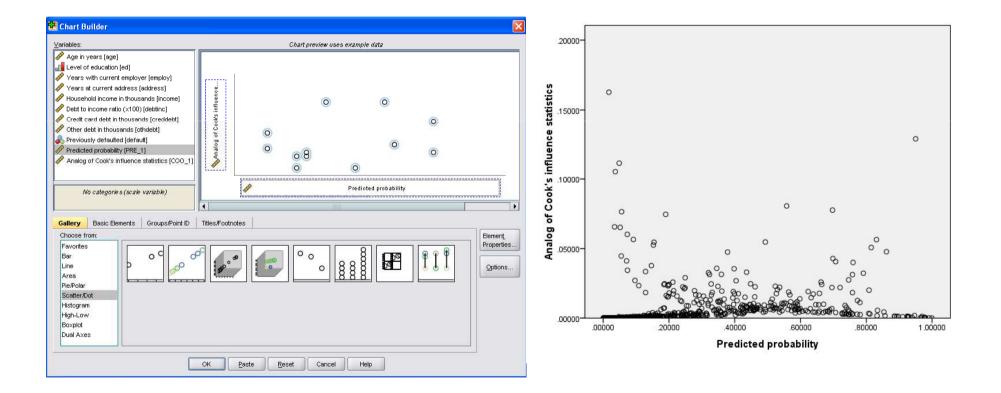
a. Variable(s) entered on step 1: debtinc.

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

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

d. Variable(s) entered on step 4: address.

Step 3: identify possible outlying obs Scatter plot: predicted probability vs Cook's D statistics



Step 4: draw ROC curve and find the optimal cut-off point

a. Under the nonparametric assumption b. Null hypothesis: true area = 0.5

Analyze **ROC Curve**

RO RO	OC Curve															
∎∎ L ∳⁄ Y ∳∕ Y	Age in years [age] _evel of education [Years with current Years at current ad	employer [employ] dress [address]	>	: I probability [PRE_1]					R	oc c	urve					
		n thousands [income]	State Variable			1.00	1		TT	27	-				/	
of D A∕C	Debt to income ratio	(x100) [debtinc] housands [creddebt]	Previous	ly defaulted [default]				_	+ +			F	-	1		
	Other debt in thousa		<u>V</u> alue of State Variable	e: 1						T				1		
		luence statistics [C	Display						1							
			Dispidy			0.80			1				1			
			ROC C <u>u</u> rve						1							
			Vith Diagona	al reference line				1	°							
			Standard <u>e</u> rror a	nd confidence interval		N 60		1								
			Coordinate points	s of the ROC Curve		€		1			/					
						i i i		1						_	_	
		ок	<u>P</u> aste <u>R</u> eset	Cancel Help		Sensitivity		1	+ +	-	4			_	-	
						ў _{0.40}		_	+	X	-	+		_		
1									+		-	+				
						+						+ +			-	
			Area Under the	Curve			f	-			-	+ +		_		
	Toot Booul	lt) (ariable/c)·P	redicted probabil	itu		0.20			X			 			-	
	TestResu	L Vallaple(S).F		Asymptotic 95	% Confidence rval			X		1						
+	Area	Std. Error ^a	Asymptotic Sig. ^b	Lower Bound	Upper Bound	0.00	0.00	().20	0.40	0	L .60	0.80		1,00	
	.856	.016	.000	.825	.886					1.4	Specific	itv				

1 - Specificity

Step 5: predict credit risk

Select the rest 150 obs and compute predicted probability with model coefficients

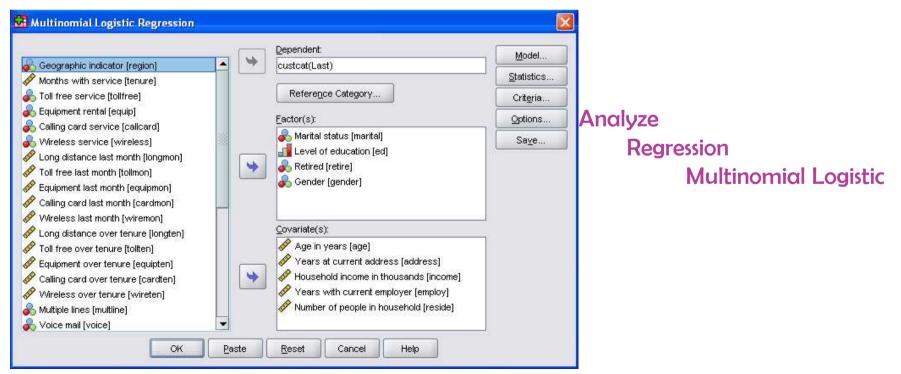
😤 Compute Variable		
Compute Variable Iarget Variable: pred Type & Label Image: Age in years [age] Level of education [ed] Years with current employer [employ] Years at current address [address] Household income in thousands [income] Debt to income ratio (x100) [debtinc] Credit card debt in thousands [orteddebt] Other debt in thousands [othdebt] Previously defaulted [default] Predicted probability [PRE_2] Analog of Cook's influence statistics [-	Numgric Expression: EXP(-0.791-0.243*employ-0.081*address+0.088*debtinc+0.573*creddebt)/(1+EXP(-0.791-0.243*employ-0.081*address+0.088*debtinc+0.573*creddebt)) + > 7 8 9 - <=
[f] (optional case selection condition)		DK Paste Reset Cancel Help

MULTINOMIAL LOGISTIC MODELS

Using Multinomial Logistic Regression to Classify Telecommunications Customers

E.g.: telco.sav

A telecommunications provider has segmented its customer base by service usage patterns, categorizing the customers into four groups. If demographic data can be used to predict group membership, you can customize offers for individual prospective customers.



MULTINOMIAL LOGISTIC MODELS

🚾 Multinomial Logistic Reg	gression: Model		×	Multinomial Logistic Regression: Statistics
Multinomial Logistic Reg	Build Terms	Forced Entry Terms: Forced Entry Terms: Stepwise Terms: reside income age retire marital ed gender		Multinomial Logistic Regression: Statistics Case processing summary Model Pseudo R-square Cell probabilities Step summary Model Model Step summary Model Model Step summary Model Model Model Model Step summary Other Step summary Model Model Model Model Step summary Cell probabilities Model Step summary Cell probabilities Model
✓ Include intercept in model		address emplov Step <u>w</u> ise Method: Forward entry	•	Image: Age in years [age] Image: Age in years [age]
	Continue	Cancel Help		Years with current e Continue Cancel Help

MULTINOMIAL LOGISTIC MODELS

				Paramet	ter Estimate	s			
								95% Confidence (E	e Interval for Exp 3)
Customer cate	edorv ^a	В	Std. Error	Wald	df	Siq.	Exp(B)	Lower Bound	Upper Bound
Basic service	Intercept	181	.431	.176	1	.675			
	reside	258	.068	14.418	1	.000	.773	.677	.883
	[ed=1]	3.762	.532	50.070	1	.000	43.047	15.183	122.044
	[ed=2]	1.959	.427	21.042	1	.000	7.095	3.072	16.390
	[ed=3]	1.453	.435	11.171	1	.001	4.276	1.824	10.025
	[ed=4]	.584	.425	1.893	1	.169	1.794	.780	4.123
	[ed=5]	0p			0				
	address	022	.012	3.498	1	.061	.978	.956	1.001
	employ	042	.012	12.437	1	.000	.958	.936	.981
E-service	Intercept	132	.351	.141	1	.707			
	reside	110	.066	2.761	1	.097	.896	.787	1.020
	[ed=1]	1.592	.481	10.938	1	.001	4.913	1.913	12.619
	[ed=2]	.452	.345	1.717	1	.190	1.571	.799	3.087
	[ed=3]	.482	.345	1.948	1	.163	1.620	.823	3.188
	[ed=4]	092	.326	.080	1	.778	.912	.481	1.728
	[ed=5]	0 ⁶			0				
	address	.015	.011	1.860	1	.173	1.015	.993	1.037
	employ	016	.011	1.969	1	.161	.984	.962	1.006
Plus service	Intercept	-1.732	.572	9.173	1	.002			
	reside	173	.067	6.583	1	.010	.841	.737	.960
	[ed=1]	4.318	.642	45.173	1	.000	75.032	21.301	264.298
	[ed=2]	2.678	.562	22.730	1	.000	14.554	4.840	43.763
	[ed=3]	2.126	.569	13.952	1	.000	8.381	2.747	25.573
	[ed=4]	1.049	.569	3.399	1	.065	2.855	.936	8.708
	[ed=5]	Op			0				
	address	.000	.011	.001	1	.980	1.000	.979	1.021
	employ	.009	.011	.682	1	.409	1.009	.988	1.031

a. The reference category is: Total service.

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

ORDINAL REGRESSION

Using Ordinal Regression to Build a Credit Scoring Model

$g(\Pr(Y \le i \mid X)) = \alpha_i + \beta' X, \quad i = 1, \dots, k$

Link function. The link function is a transformation of the cumulative probabilities that allows estimation of the model. Five link functions are available, summarized in the following table.

Function	Form	Typical application		
Logit	log(ξ/(1-ξ))	Evenly distributed categories		
Complementary log-log	ementary log-log log(-log(1-٤j)) Higher categories more probab			
Negative log-log	-log(-log(٤))	Lower categories more probable		
Probit	Ф ⁻¹ (£)	Latent variable is normally distributed		
Cauchit (inverse Cauchy)	tan(π(ξ=0.5))	Latent variable has many extreme values		

E.g.: german_credit.sav

A creditor wants to be able to determine whether an applicant is a good credit risk, given various financial and personal characteristics. From their customer database, the creditor (dependent) variable is account status, with five ordinal levels: no debt history, no current debt, debt payments current, debt payments past due, and critical account. Potential predictors consist of various financial and personal characteristics of applicants, including age, number of credits at the bank, housing type, checking account status, and so on.

ORDINAL REGRESSION

Analyze	🔂 Ordinal Regression: Options 🛛 🔀
Regression Ordinal	Maximum iterations: 100 Maximum step-halving: 5
Image: construct status Image: construct status	Magimum step-halving: 5 Log-likelihood convergence: 0 Qonfidence interval: 95 Qutput Singularity tolerance: 0.0000001 Singularity tolerance: 0.0000001 Location Link: Cauchit Scale Continue Cancel Help Print teration history for every step(s) Print teration history for every step(s) Summary statistics Predicted category Predicted category probability Actual category probability Asymptotic covariance of parameter estimates Print Log-Likelihood Cell information Information Test of parallel lines Excluding multinomial constant Continue Cancel

ORDINAL REGRESSION

		Model Fitti	ing Information		
	Model	-2 Log Likelihood	Chi-Square	df	Sig.
+	Intercept Only	2249.888			
	Final	1790.028	459.860	9	.000
	Link function:	Cauchit.			

Goodness-of-Fit								
	Chi-Square	df	Sig.					
Pearson	3467.625	3131	.000					
Deviance	1690.392	3131	1.000					
Link function: Cauchit.								

Pseudo R-Square

Cox and Snell	.369
Nagelkerke	.407
McFadden	.194

Link function: Cauchit.

			Paran	neter Estima	tes			
							95% Confid	ence Interval
		Estimate	Std. Error	Wald	df	Siq.	Lower Bound	Upper Boun
Threshold	[chist = 1.00]	-9.356	1.432	42.689	1	.000	-12.163	-6.54
	[chist = 2.00]	-5.232	.989	28.010	1	.000	-7.169	-3.29
	[chist = 3.00]	552	.933	.350	1	.554	-2.382	1.27
	[chist = 4.00]	.432	.929	.216	1	.642	-1.389	2.25
Location	age	.016	.008	3.393	1	.065	.000	.03
	duration	013	.007	3.012	1	.083	028	.00
	[numcred=1.00]	-2.616	.729	12.867	1	.000	-4.046	-1.18
	[numcred=2.00]	.817	.702	1.353	1	.245	560	2.19
	[numcred=3.00]	2.002	.940	4.533	1	.033	.159	3.84
	[numcred=4.00]	0ª			0			
	[othnstal=1.00]	-1.257	.237	28.113	1	.000	-1.721	79
	[othnstal=2.00]	-1.031	.355	8.424	1	.004	-1.727	33
	[othnstal=3.00]	0ª			0			
	[housng=1.00]	275	.377	.533	1	.465	-1.014	.46
	[housng=2.00]	.049	.320	.024	1	.878	577	.67
	[housng=3.00]	0ª			0			

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

LOGISTIC MODELS

- Hays, W. L. 1981. Statistics, 3rd ed. New York: Holt, Rinehart, and Winston.
- McCullagh, P., and J. A. Nelder. 1989. Generalized Linear Models, 2nd ed. London: Chapman & Hall.
- Cox, D. R., and E. J. Snell. 1989. The Analysis of Binary Data, 2nd ed. London: Chapman and Hall.
- Hosmer, D. W., and S. Lemeshow. 2000. Applied Logistic Regression, 2nd ed. New York: John Wiley and Sons.
- Kleinbaum, D. G. 1994. Logistic Regression: A Self-Learning Text. New York: Springer-Verlag.
- Norusis, M. 2004. SPSS 13.0 Advanced Statistical Procedures Companion. Upper Saddle-River, N.J.: Prentice Hall, Inc..

END

Advanced Models

- 30 May Friday 10am-12pm
- Training Room @ Library Level 5
- Curve Estimation
- Non-linear Regression
- Survival Analysis
- Linear Mixed Model
- Time-series Data Analysis