



The University of Sydney
Australia

Wanna build a statistical model? No worries!

SAS macros to streamline model building

Navneet Dhand

The Faculty of Veterinary Science

Model building: a challenging job

- You have to be:

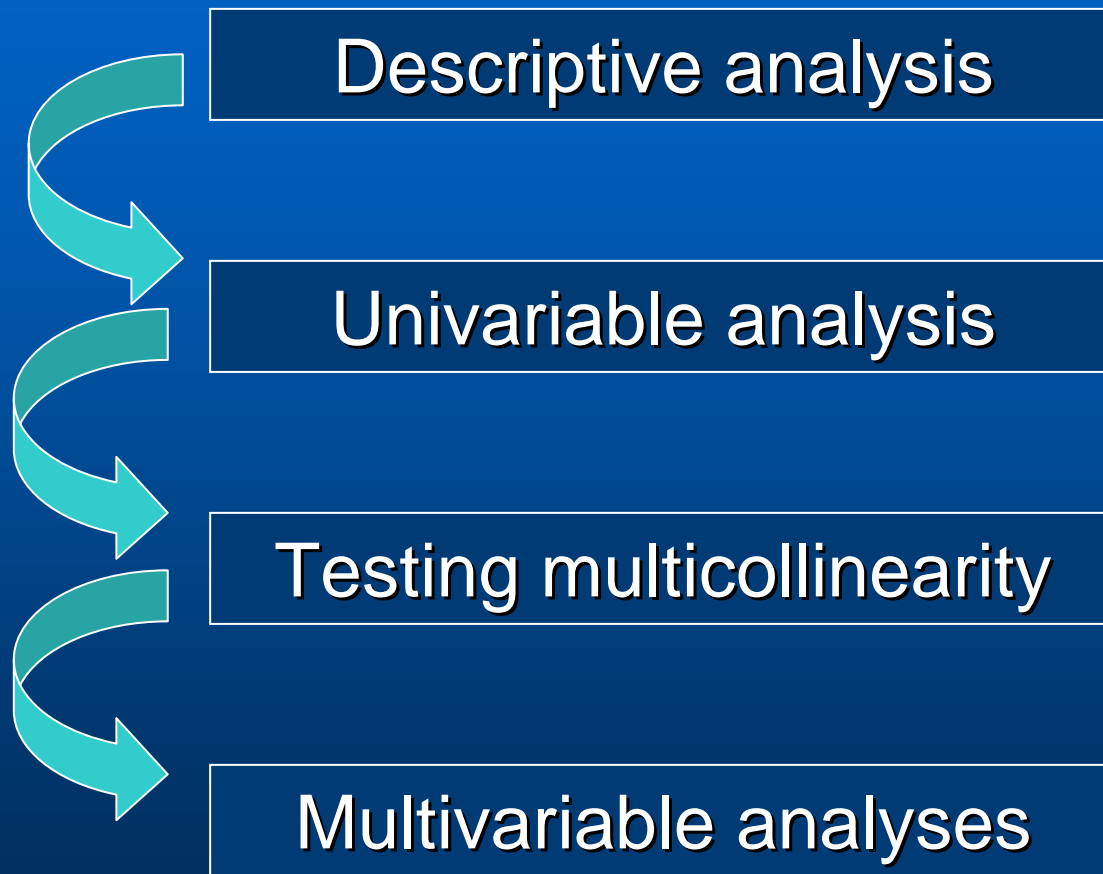
- careful
- attentive
- systematic
- thorough
- alert
- Meticulous



- The Macros:

- make the process significantly less laborious;
- are huge time savers.

Model building process



Model building process

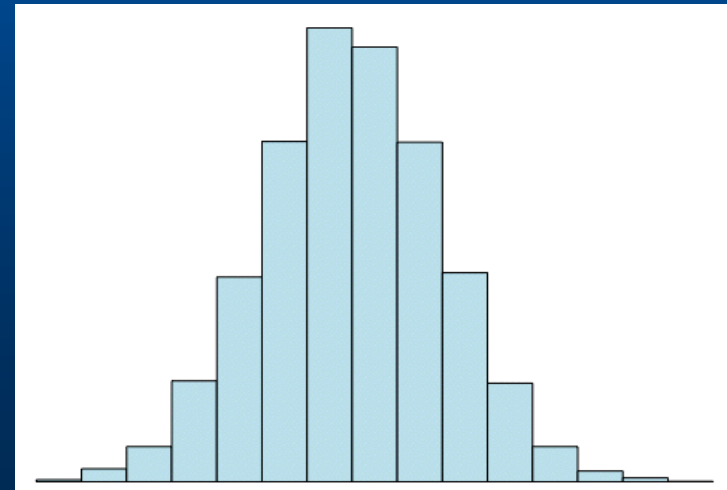
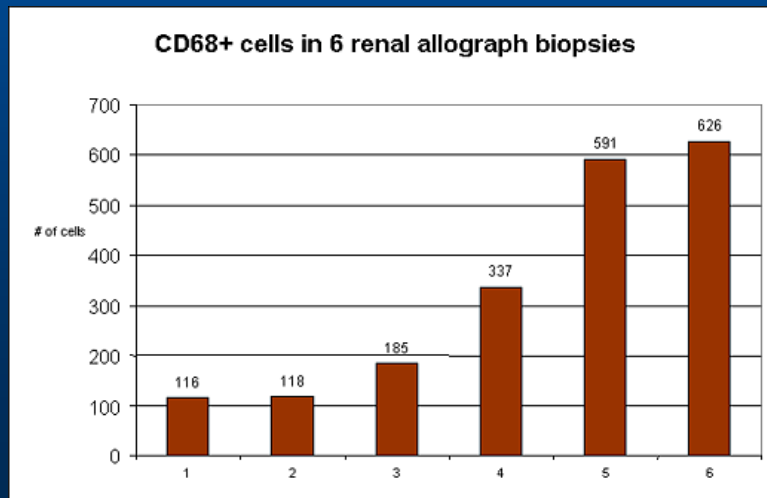
Descriptive analysis

Categorical Variables

- Frequency tables
- Bar charts

Continuous variables

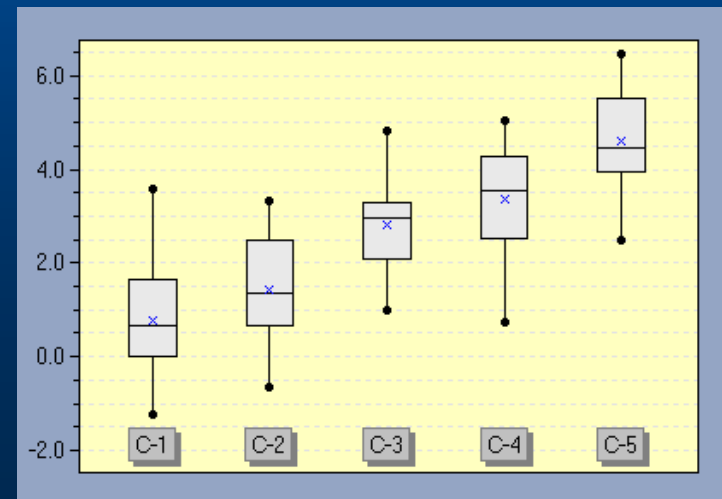
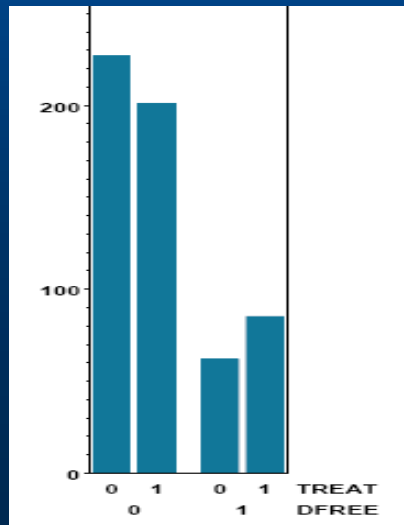
- Means, quantiles and variances
- Histograms



Model building process

Univariable analysis

- Test unconditional associations with the outcome
- Contingency tables and bar-charts of the categorical variables with the outcome
- Quantiles and box-plots of the continuous variables with the outcome



Model building process

Test Collinearity

- Associations between the explanatory variables
- Spearman rank correlation coefficient for the ordinal variables
- Chi-square and P-values for the nominal variables
- Pearson or Spearman rank correlation coefficient for the continuous variables

Model building process

Multivariable Analyses

- Associations after adjusting for other variables
- Automatic methods ■ Manual methods



Manual methods preferred- but are very time intensive

Model building process

Descriptive analysis



Tables

Univariable analysis



Tables

Testing multicollinearity



Tables

Multivariable analyses

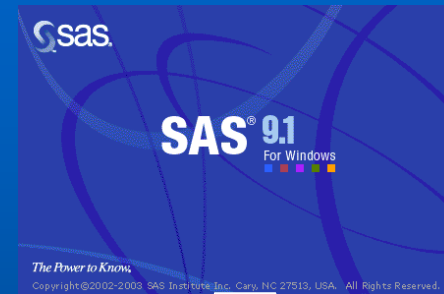


Tables



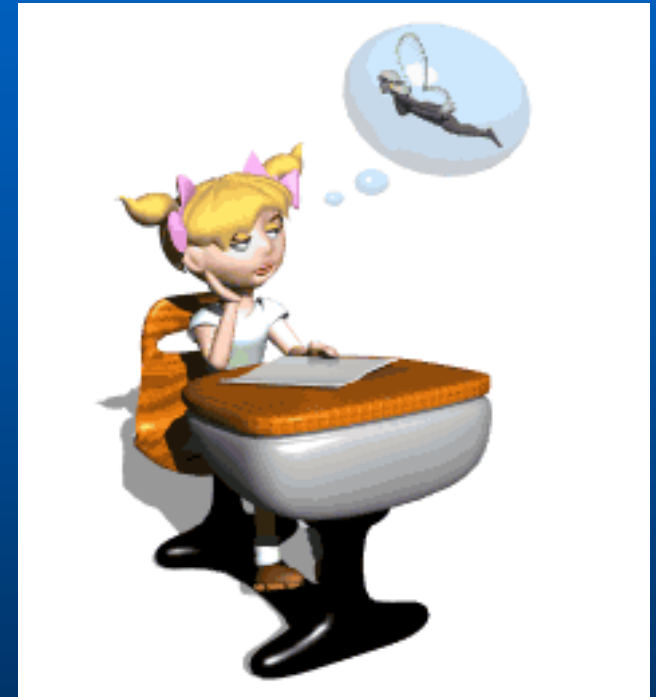
Creation of tables

- Need to present the results in tabulated forms for reports and publications
- MS Excel - preferred software for tabulating
- Copying of desired results from SAS to Excel
 - time consuming
 - great chances of incurring errors
 - repetitive and boring!



A dream...

- What if:
 - all the analyses are conducted automatically!
 - desired results from SAS are automatically exported to Excel!!
 - tables are automatically created for publications!!!



Is this possible?

From dreams to reality...

Yes, now this is possible!

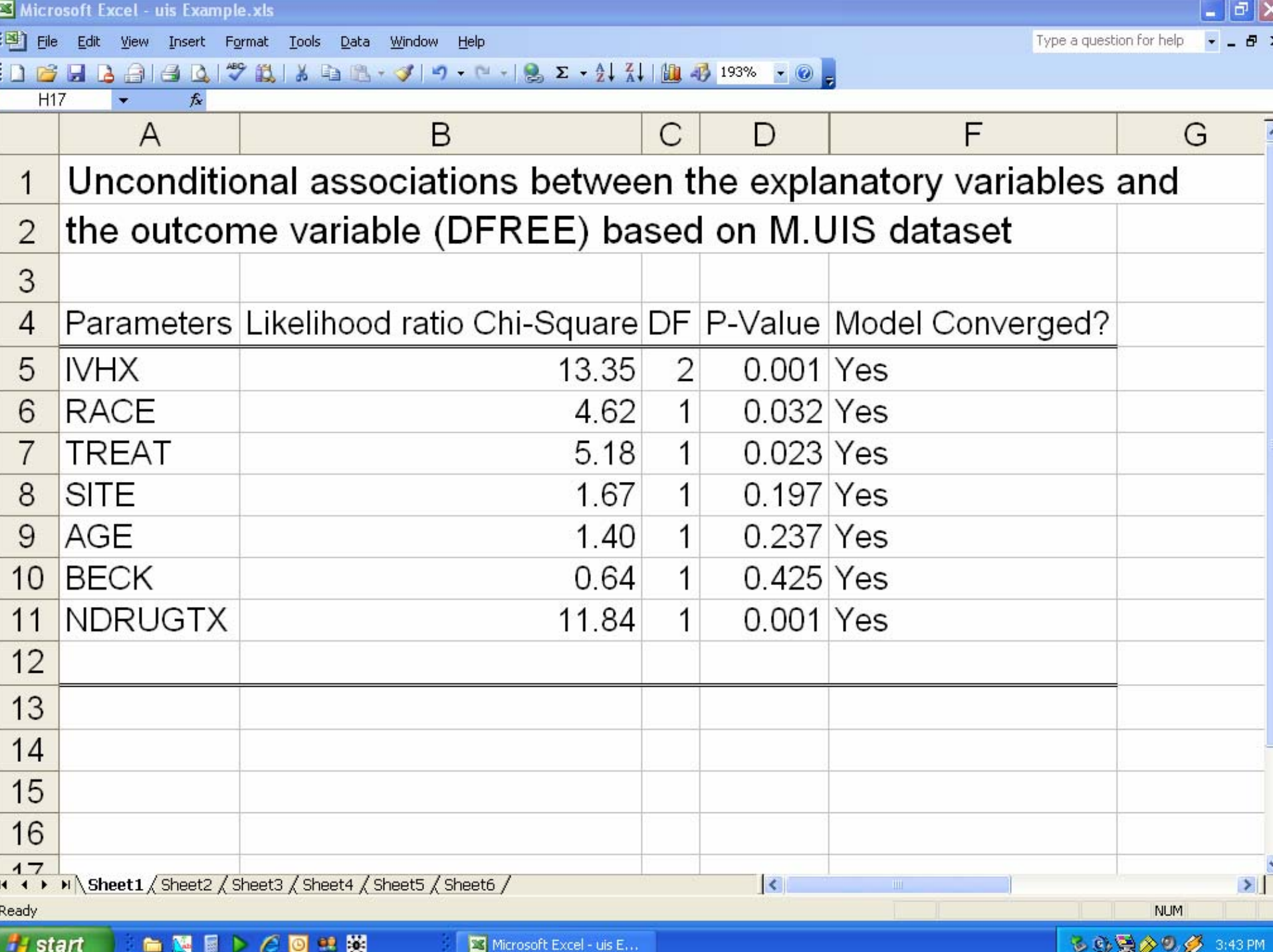
- Created 6 SAS macros

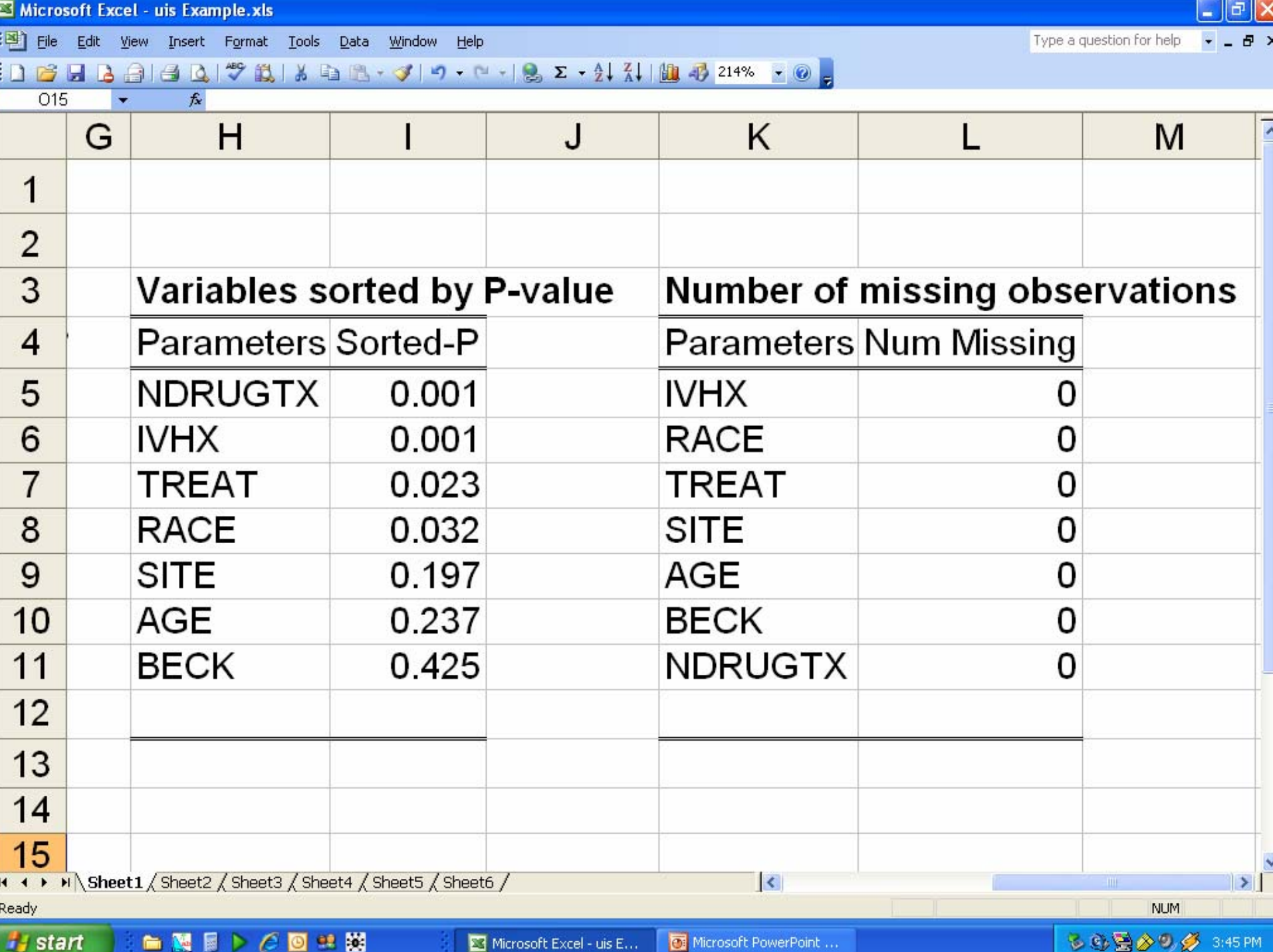
- UniLogistic
 - MultiLogistic
 - UniGLM
 - MultiGLM
 - MultiMixed
 - MultiGlimmix
- } Logistic regression
- } Linear regression
- } Mixed Models

UniLogistic Macro

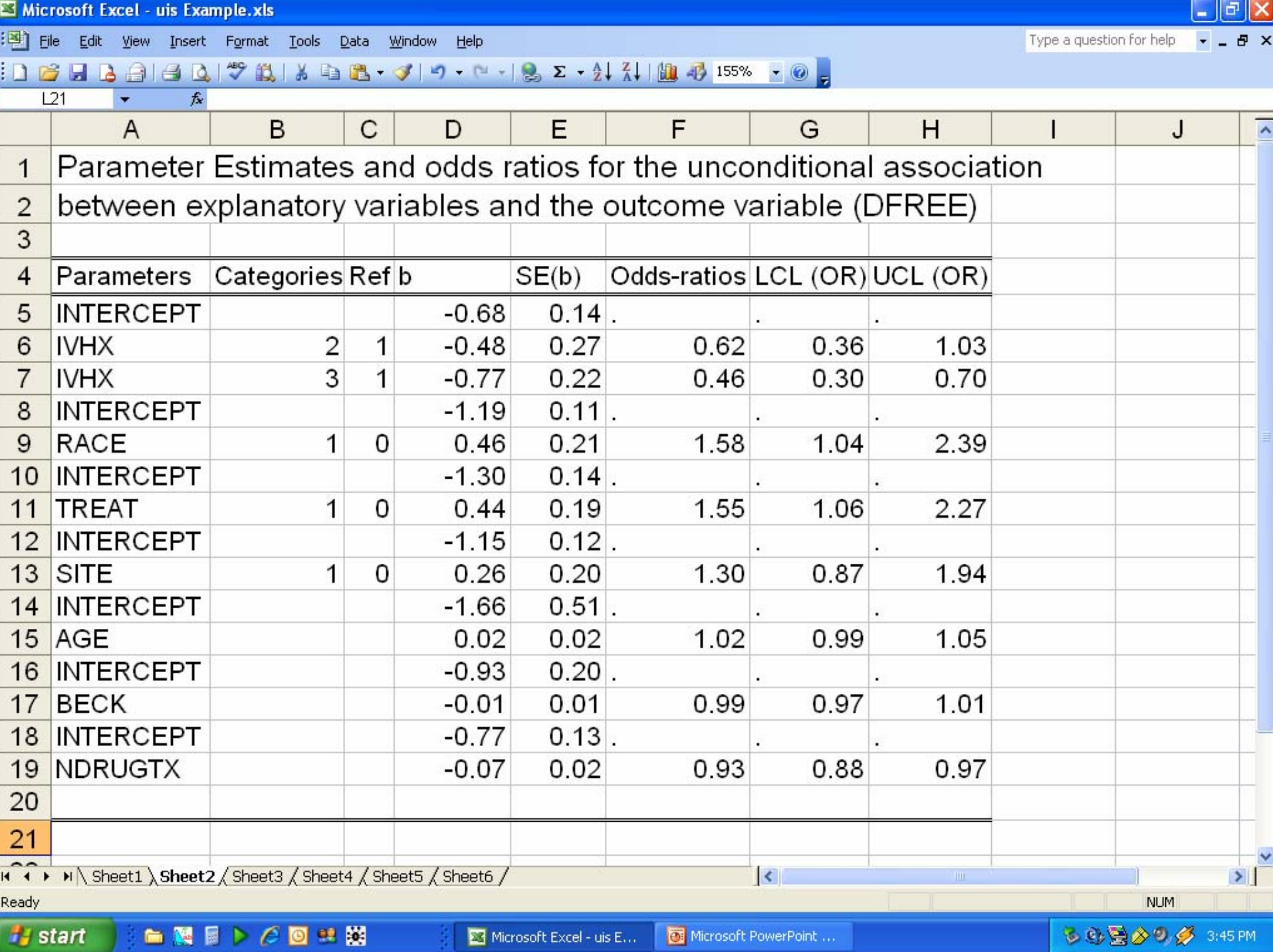
- Conducts descriptive and univariable logistic regression analyses
- Presents results in well formatted tables in MS Excel
- Creates PDF files for graphics and links them to the Excel file

All this with the single click of a mouse!

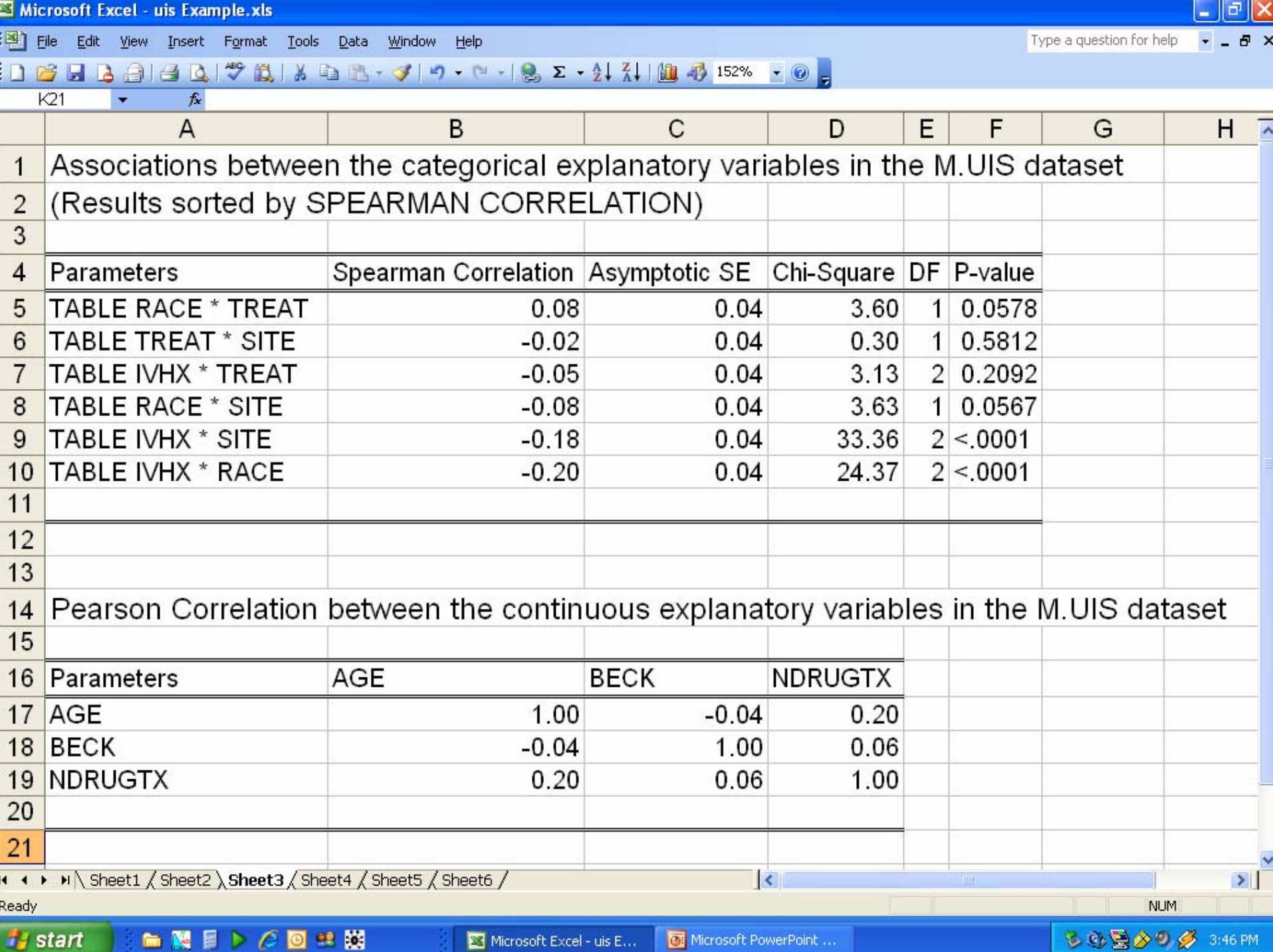




	G	H	I	J	K	L	M
1							
2							
3		Variables sorted by P-value			Number of missing observations		
4		Parameters	Sorted-P		Parameters	Num Missing	
5		NDRUGTX	0.001		IVHX	0	
6		IVHX	0.001		RACE	0	
7		TREAT	0.023		TREAT	0	
8		RACE	0.032		SITE	0	
9		SITE	0.197		AGE	0	
10		AGE	0.237		BECK	0	
11		BECK	0.425		NDRUGTX	0	
12							
13							
14							
15							

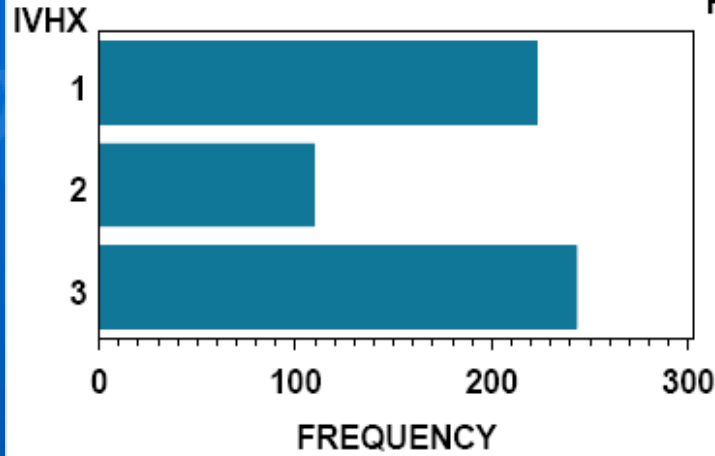
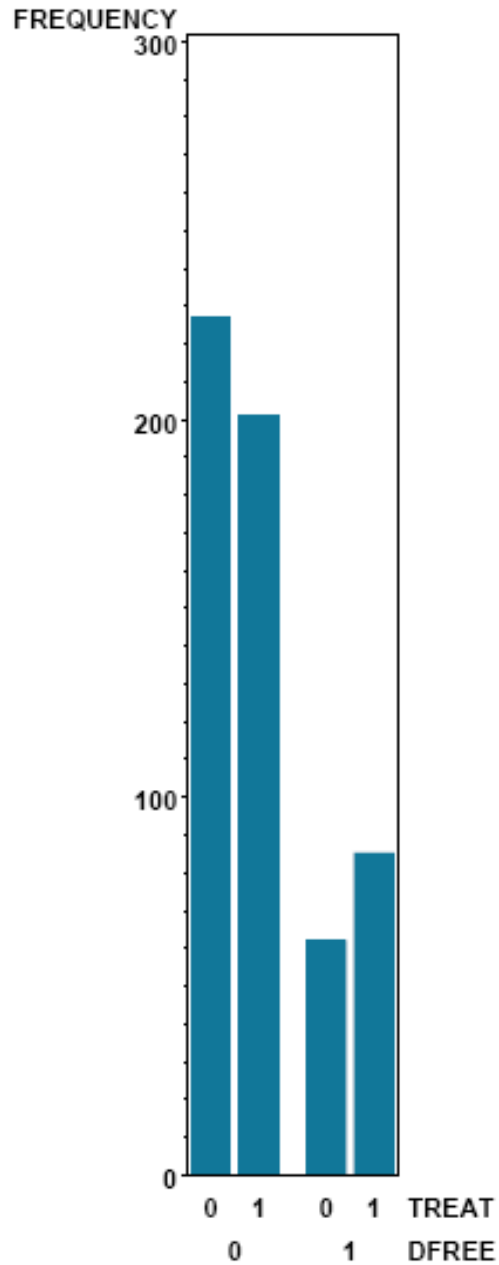


	A	B	C	D	E	F	G	H	I	J
1	Parameter Estimates and odds ratios for the unconditional association									
2	between explanatory variables and the outcome variable (DFREE)									
3										
4	Parameters	Categories	Ref	b	SE(b)	Odds-ratios	LCL (OR)	UCL (OR)		
5	INTERCEPT			-0.68	0.14	.	.	.		
6	IVHX	2	1	-0.48	0.27	0.62	0.36	1.03		
7	IVHX	3	1	-0.77	0.22	0.46	0.30	0.70		
8	INTERCEPT			-1.19	0.11	.	.	.		
9	RACE	1	0	0.46	0.21	1.58	1.04	2.39		
10	INTERCEPT			-1.30	0.14	.	.	.		
11	TREAT	1	0	0.44	0.19	1.55	1.06	2.27		
12	INTERCEPT			-1.15	0.12	.	.	.		
13	SITE	1	0	0.26	0.20	1.30	0.87	1.94		
14	INTERCEPT			-1.66	0.51	.	.	.		
15	AGE			0.02	0.02	1.02	0.99	1.05		
16	INTERCEPT			-0.93	0.20	.	.	.		
17	BECK			-0.01	0.01	0.99	0.97	1.01		
18	INTERCEPT			-0.77	0.13	.	.	.		
19	NDRUGTX			-0.07	0.02	0.93	0.88	0.97		
20										
21										

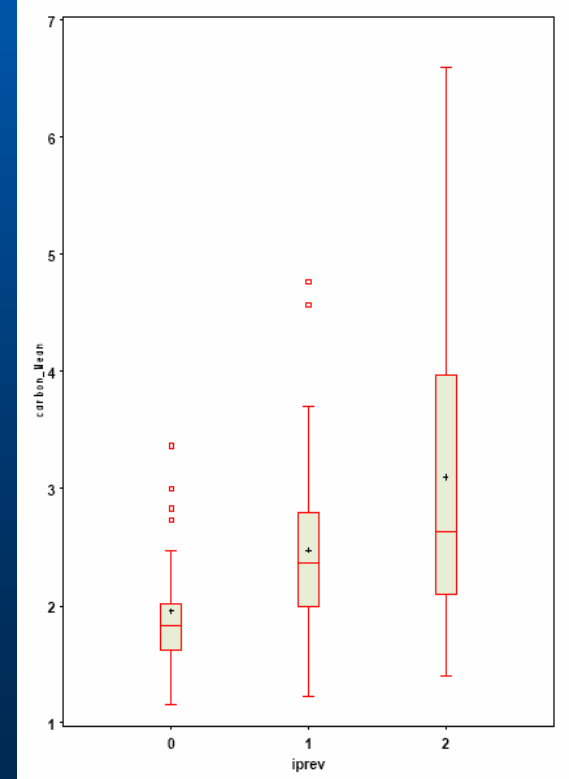
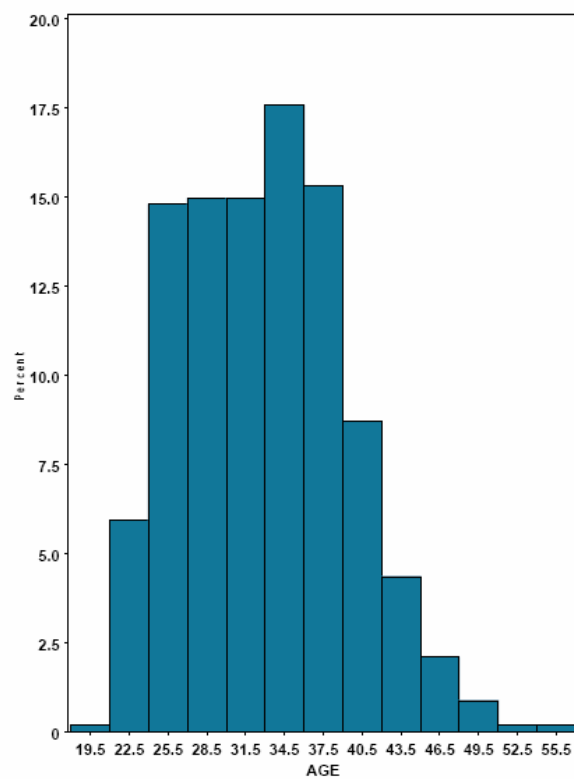


	A	B	C	D	E	F	G	H
1	Associations between the categorical explanatory variables in the M.UIS dataset							
2	(Results sorted by SPEARMAN CORRELATION)							
3								
4	Parameters	Spearman Correlation	Asymptotic SE	Chi-Square	DF	P-value		
5	TABLE RACE * TREAT	0.08	0.04	3.60	1	0.0578		
6	TABLE TREAT * SITE	-0.02	0.04	0.30	1	0.5812		
7	TABLE IVHX * TREAT	-0.05	0.04	3.13	2	0.2092		
8	TABLE RACE * SITE	-0.08	0.04	3.63	1	0.0567		
9	TABLE IVHX * SITE	-0.18	0.04	33.36	2	<.0001		
10	TABLE IVHX * RACE	-0.20	0.04	24.37	2	<.0001		
11								
12								
13								
14	Pearson Correlation between the continuous explanatory variables in the M.UIS dataset							
15								
16	Parameters	AGE	BECK	NDRUGTX				
17	AGE	1.00	-0.04	0.20				
18	BECK	-0.04	1.00	0.06				
19	NDRUGTX	0.20	0.06	1.00				
20								
21								

Bar Diagrams with Outcome



	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
1	223	223	38.78	38.78
2	109	332	18.96	57.74
3	243	575	42.26	100.00



Macro implementation

- How can you obtain all this output?

```
%UNILOGISTIC (  
    dsn = health,                /* dataset name */  
    outcome = pos/total,        /* outcome variable */  
    catvar = cat1 cat2 cat3 cat4, /*categorical variables*/  
    contvar = cont1 cont2 cont3 /* continuous variables */  
)
```

MultiLogistic Macro

- Builds multivariable models by stepwise, forward or backward procedure
- Mimics the **manual** model building process
- User decides the variable to be included or excluded at each step based on the:
 - likelihood ratio chi-square
 - parameter estimates and standard errors
 - AIC and BIC
- Tests interactions
- Creates final table

MultiLogistic Macro

Wald or Likelihood-ratio chi-square?

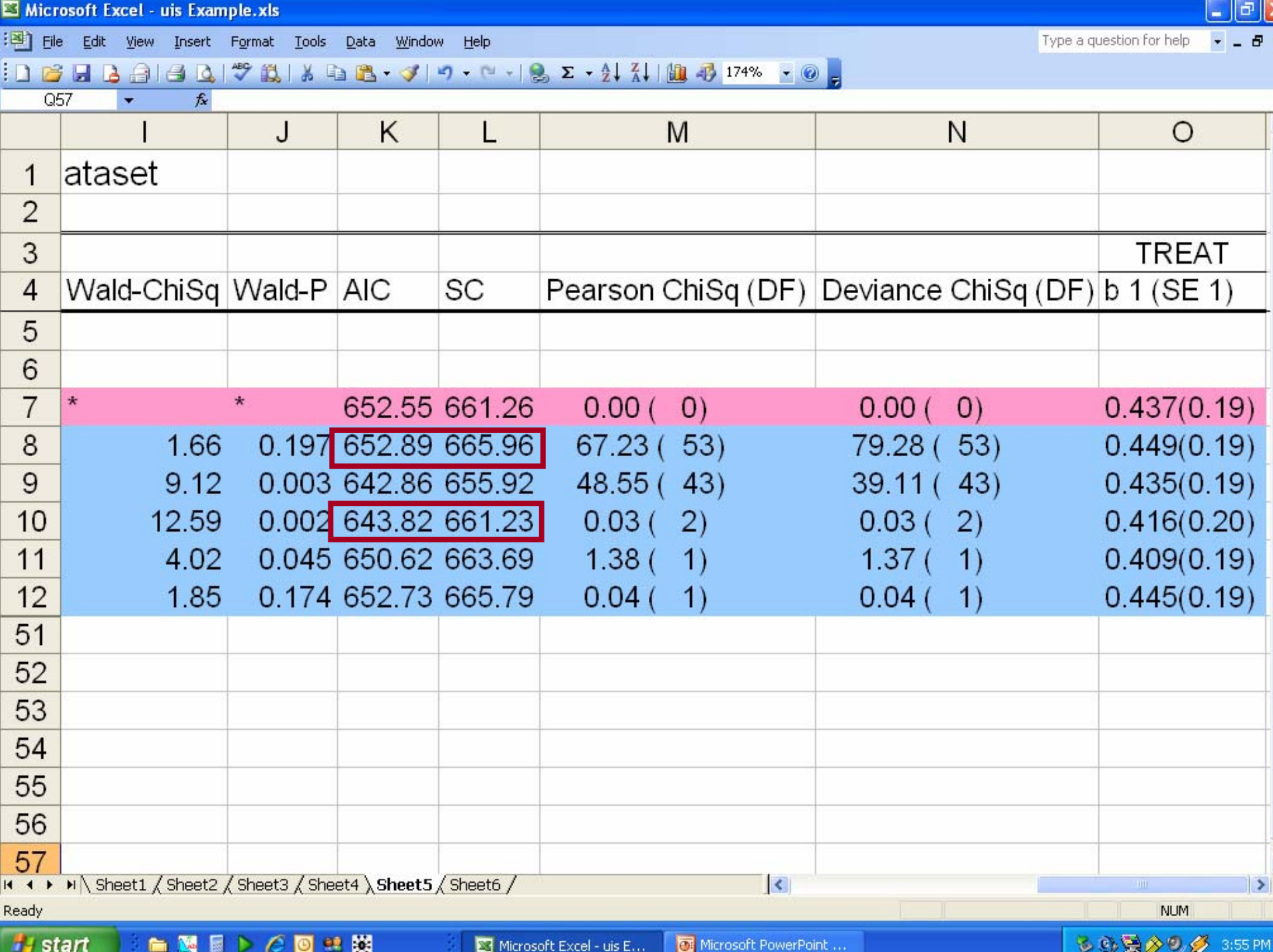
- Wald chi square: automatic, but less robust
- Likelihood ratio: more robust, but needs to be calculated manually based on log-likelihood values
- Most users end up using Wald chi-square instead
- The MultiLogistic macro calculates likelihood-ratio chi-square and P-value for each model

MultiLogistic code

```
%MULTILOGISTIC (  
  dsn = health, /* Dataset name */  
  outcome = out, /* Outcome variable */  
  catvar = cat1 cat2 cat3 cat4, /*Categorical variables*/  
  contvar = cont1 cont2 cont3, /* Continuous variables */  
  
  selection = stepwise, /* Variable selection */  
  basemodel = cat1 cat3 cat4 cont1 cont3 /* BASE Model */  
)
```

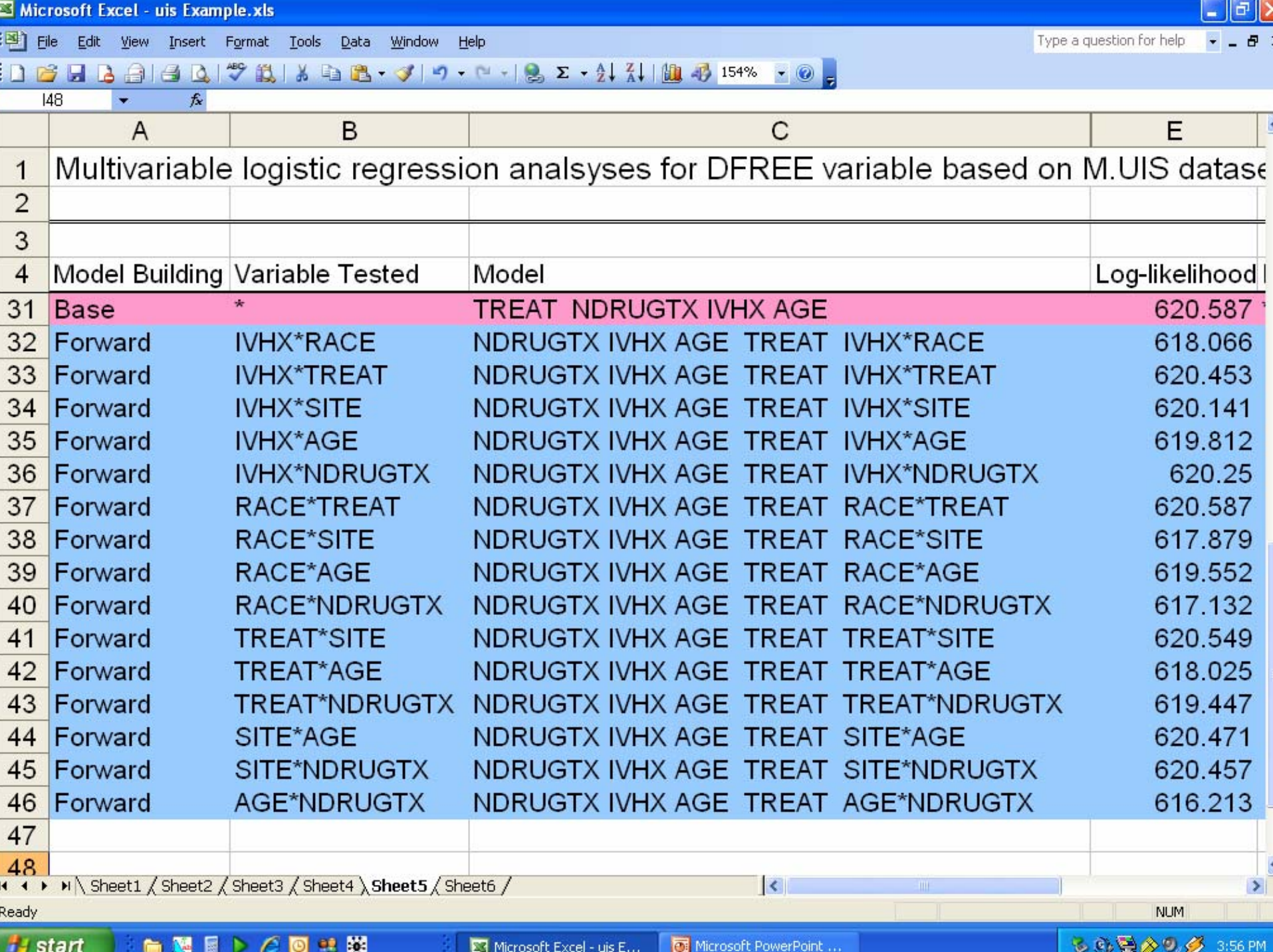
- **Base Model:** The macro compares all the models with the specified base model.
 - **Backward selection:** Specify all the variables in the beginning
 - **Forward and stepwise selection:** Specify only the exposure or the forced variables in the beginning

	A	B	C	E	F	G	H
1	Multivariable logistic regression analyses for DFREE variable based on M.UIS						
2							
3							
4	Model Building	Variable Tested	Model		Log-likelihood	LR Chi-Square	DF P
5							
6							
7	Base	*	TREAT		648.551	*	* *
8	Forward	AGE	TREAT AGE		646.894	1.66	1 0.198
9	Forward	NDRUGTX	TREAT NDRUGTX		636.86	11.69	1 0.001
10	Forward	IVHX	TREAT IVHX		635.815	12.74	2 0.002
11	Forward	RACE	TREAT RACE		644.622	3.93	1 0.047
12	Forward	SITE	TREAT SITE		646.728	1.82	1 0.177
51							
52							
53							
54							
55							
56							
57							
58							



	I	J	K	L	M	N	O
1	ataset						
2							
3							
4	Wald-ChiSq	Wald-P	AIC	SC	Pearson ChiSq (DF)	Deviance ChiSq (DF)	b 1 (SE 1)
5							
6							
7	*	*	652.55	661.26	0.00 (0)	0.00 (0)	0.437(0.19)
8	1.66	0.197	652.89	665.96	67.23 (53)	79.28 (53)	0.449(0.19)
9	9.12	0.003	642.86	655.92	48.55 (43)	39.11 (43)	0.435(0.19)
10	12.59	0.002	643.82	661.23	0.03 (2)	0.03 (2)	0.416(0.20)
11	4.02	0.045	650.62	663.69	1.38 (1)	1.37 (1)	0.409(0.19)
12	1.85	0.174	652.73	665.79	0.04 (1)	0.04 (1)	0.445(0.19)
51							
52							
53							
54							
55							
56							
57							

	A	B	C	E	F	G	H
1	Multivariable logistic regression analysys for DFREE variable based on M.UIS datase						
2							
3							
4	Model Building	Variable Tested	Model		Log-likelihood	LR Chi-Square	DF P
5							
6							
7	Base	*	TREAT		648.551 *		* *
8	Forward	AGE	TREAT AGE		646.894	1.66	1 0.198
9	Forward	NDRUGTX	TREAT NDRUGTX		636.86	11.69	1 0.001
10	Forward	IVHX	TREAT IVHX		635.815	12.74	2 0.002
11	Forward	RACE	TREAT RACE		644.622	3.93	1 0.047
12	Forward	SITE	TREAT SITE		646.728	1.82	1 0.177
13	Base	*	TREAT NDRUGTX		636.86 *		* *
14	Backward	NDRUGTX	TREAT		648.551	11.69	1 0.001
15	Forward	AGE	TREAT NDRUGTX AGE		632.443	4.42	1 0.036
16	Forward	IVHX	TREAT NDRUGTX IVHX		630.05	6.81	2 0.033
17	Forward	RACE	TREAT NDRUGTX RACE		634.01	2.85	1 0.091
18	Forward	SITE	TREAT NDRUGTX SITE		635.942	0.92	1 0.338
53							
54							
55							



	A	B	C	E
1	Multivariable logistic regression analyses for DFREE variable based on M.UIS dataset			
2				
3				
4	Model Building	Variable Tested	Model	Log-likelihood
31	Base	*	TREAT NDRUGTX IVHX AGE	620.587
32	Forward	IVHX*RACE	NDRUGTX IVHX AGE TREAT IVHX*RACE	618.066
33	Forward	IVHX*TREAT	NDRUGTX IVHX AGE TREAT IVHX*TREAT	620.453
34	Forward	IVHX*SITE	NDRUGTX IVHX AGE TREAT IVHX*SITE	620.141
35	Forward	IVHX*AGE	NDRUGTX IVHX AGE TREAT IVHX*AGE	619.812
36	Forward	IVHX*NDRUGTX	NDRUGTX IVHX AGE TREAT IVHX*NDRUGTX	620.25
37	Forward	RACE*TREAT	NDRUGTX IVHX AGE TREAT RACE*TREAT	620.587
38	Forward	RACE*SITE	NDRUGTX IVHX AGE TREAT RACE*SITE	617.879
39	Forward	RACE*AGE	NDRUGTX IVHX AGE TREAT RACE*AGE	619.552
40	Forward	RACE*NDRUGTX	NDRUGTX IVHX AGE TREAT RACE*NDRUGTX	617.132
41	Forward	TREAT*SITE	NDRUGTX IVHX AGE TREAT TREAT*SITE	620.549
42	Forward	TREAT*AGE	NDRUGTX IVHX AGE TREAT TREAT*AGE	618.025
43	Forward	TREAT*NDRUGTX	NDRUGTX IVHX AGE TREAT TREAT*NDRUGTX	619.447
44	Forward	SITE*AGE	NDRUGTX IVHX AGE TREAT SITE*AGE	620.471
45	Forward	SITE*NDRUGTX	NDRUGTX IVHX AGE TREAT SITE*NDRUGTX	620.457
46	Forward	AGE*NDRUGTX	NDRUGTX IVHX AGE TREAT AGE*NDRUGTX	616.213
47				
48				

	A	B	D	E	F	G	H	I	J
1	Final logistic regression model for the outcome variable DFREE based on 575 observation								
2									
3	Pearson Chi-square:	509.92	442						
4	Deviance Chi-square:	469.25	442						
5	Concordant pair %:	66.12	0.66						
6									
7	Parameters	Categories	Ref	b	SE(b)	Odds-ratios	LCL	UCL	P(Likelihood-ratio)
8	INTERCEPT			-1.32	0.73	.	.	.	
9	TREAT	1	0	0.44	0.20	1.55	1.05	2.29	0.028
10	NDRUGTX			-0.36	0.15	.	.	.	0.012
11	IVHX	2	1	-0.56	0.29	0.57	0.32	0.99	0.005
12	IVHX	3	1	-0.78	0.25	0.46	0.28	0.75	0.005
13	AGE			0.02	0.02	.	.	.	0.312
14	NDRUGTX*AGE			0.01	0.00	.	.	.	
15									
16									
17									
18									
19									
20									
21									
22									

Other Macros

- Linear regression
 - UniGLM
 - MultiGLM
- Linear mixed modelling
 - MultiMixed
- Generalised linear mixed modelling
 - MultiGlimmix

Documentation



SAS® Users Group International

San Francisco, CA | March 26-29, 2006

<http://www2.sas.com/proceedings/sugi31/135-31.pdf>

Paper 135-31

Combining the Power of ODS, Data Set Concatenation, and DDE to Output Customized Statistical Results from SAS® to Microsoft Excel

Navneet K. Dhand, The University of Sydney, Camden, NSW, Australia

Jenny-Ann LML. Toribio, The University of Sydney, Camden, NSW, Australia

A research paper is being prepared for the journal
Computational Statistics and Data Analysis

Web site

www.usyd.edu.au

- Web pages are being created to make these macros available on the University of Sydney website
- The web site will include:
 - macro documentation
 - a facility for free download
 - example applications
 - tutorials on model building
 - a glossary
 - frequently asked questions

Likely to be functional by August

Limitations

- Don't have as many options as in the actual statistical software
- Not rigorously tested; might have bugs
- The user needs to have:
 - MS Excel
 - SAS
 - Adobe Reader
- The user is required to have a basic understanding of SAS software

Acknowledgements

- Dr Jenny-Ann Toribio
- Professor Richard Whittington
- Dr Peter Thomson
- Sally Pope



The University of Sydney
Australia



www.listserv.uga.edu

The University of Georgia