

MEDDB 5502, Module 14, review

Topics to be covered, 1 of 2

- What you will learn
 - 01 Linear regression, analysis of variance
 - 02 Linear regression with multiple independent variables
 - 03 Analysis of covariance
 - 04 Multi-factor analysis of variance
 - 05 Dimension reduction
 - 06 Logistic regression
 - 07 Diagnostic tests

Topics to be covered, 2 of 2

- What you will learn
 - 08 Survival analysis
 - 09 Meta-analysis
 - 10 Dark side of data science
 - 11 Hierarchical models
 - 12 Longitudinal data
 - 13 Bayesian statistics

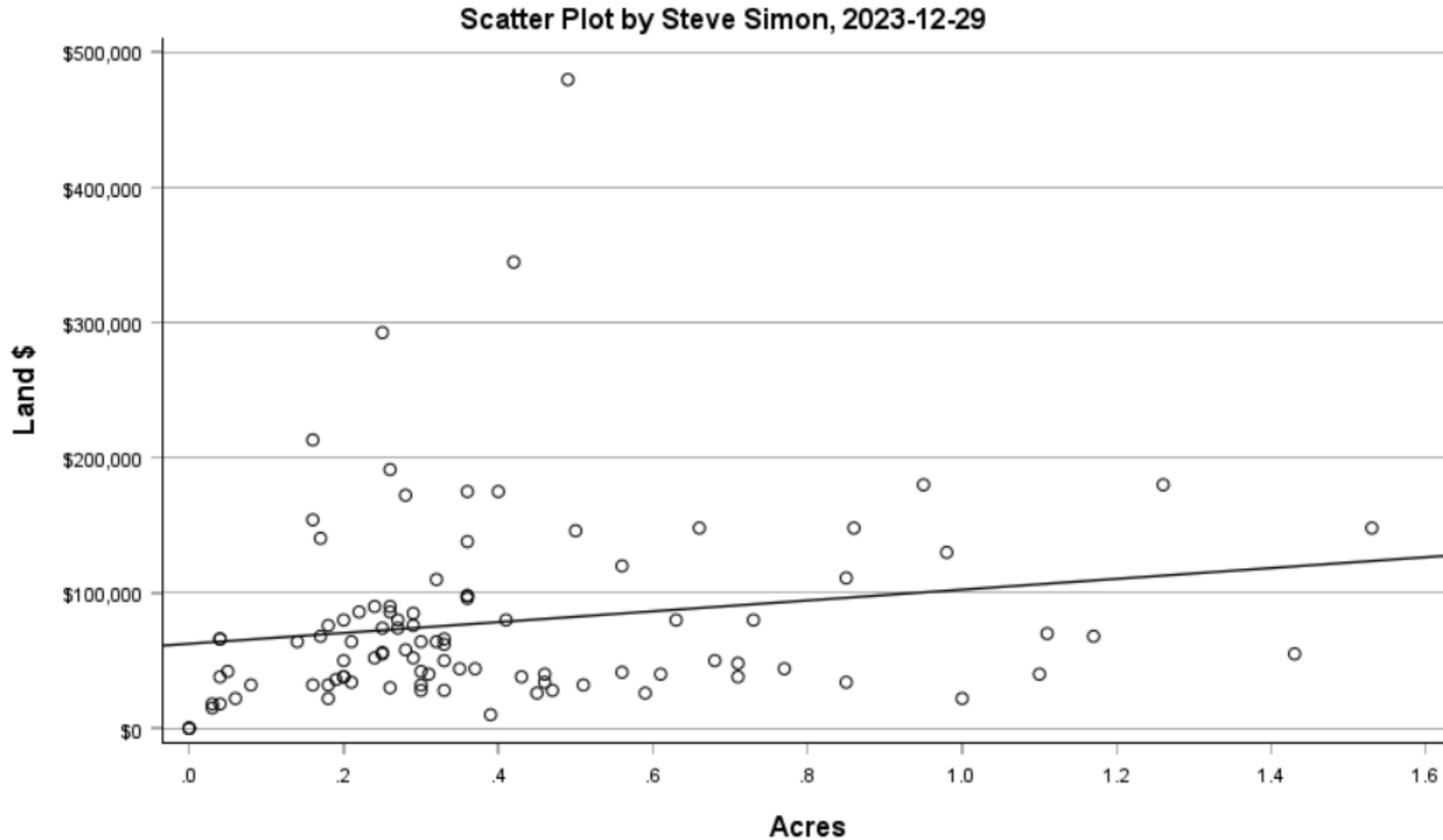
Module 01, Review

- Simple linear regression
- One factor analysis of variance

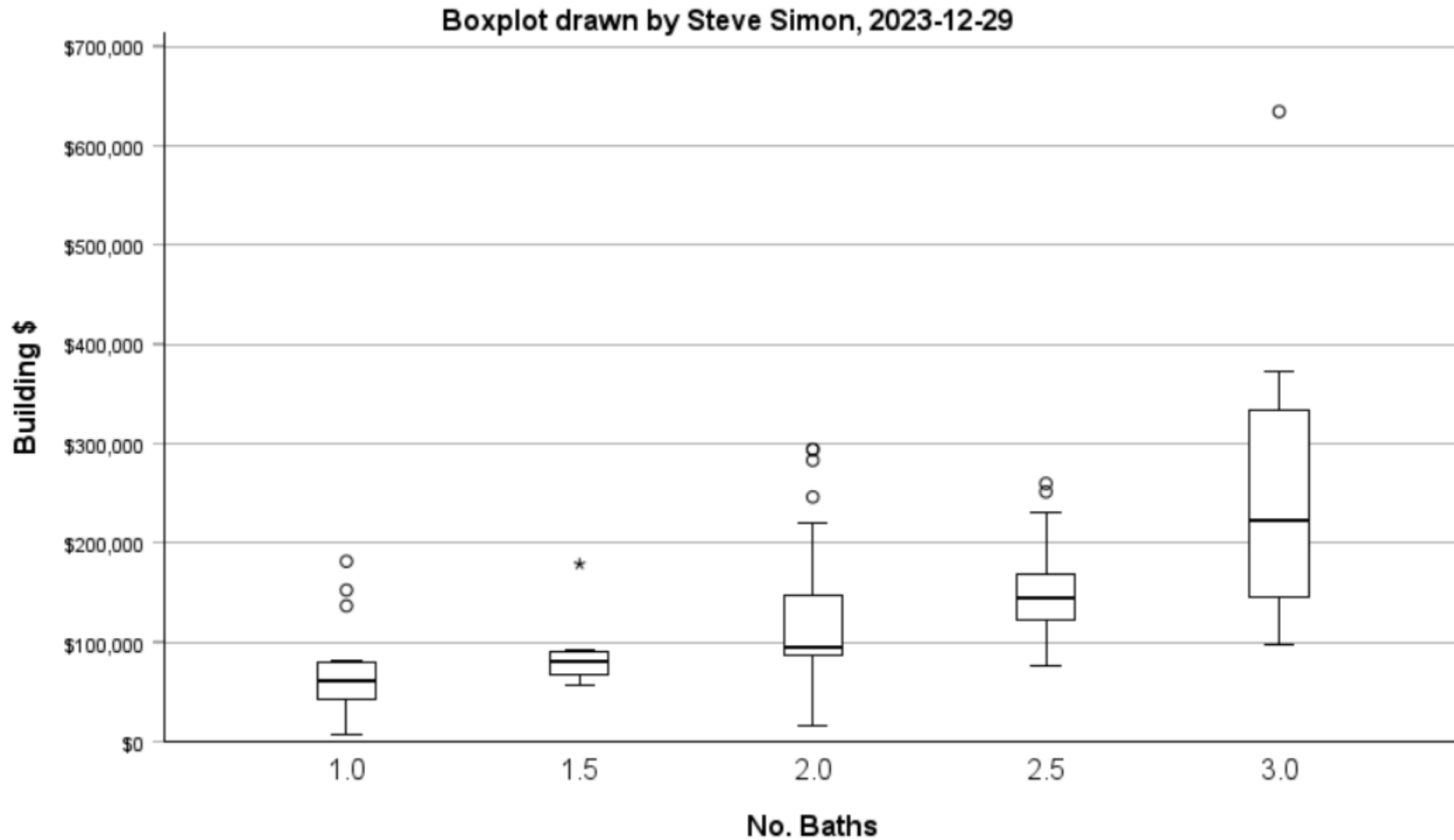
Speaker notes

You started out slowly with a review of how to run a simple linear regression model and a one factor analysis of variance.

Module 01, SPSS scatterplot



Module 01, SPSS boxplot



Module 01, SPSS calculation of R Square

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.177 ^a	.031	.021	71009.335

a. Predictors: (Constant), Acres

b. Dependent Variable: Land \$

Module 01, SPSS ANOVA table

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	15868011734	1	15868011734	3.147	.079 ^b
	Residual	4.891E+11	97	5042325655.1		
	Total	5.050E+11	98			

a. Dependent Variable: Land \$

b. Predictors: (Constant), Acres

Module 01, SPSS linear regression coefficients

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	62462.865	11520.974		5.422	<.001
	Acres	39999.858	22548.239	.177	1.774	.079

a. Dependent Variable: Land \$

Break #1

- What you have learned
 - 01 Linear regression, analysis of variance
- What's coming next
 - 02 Linear regression with multiple independent variables

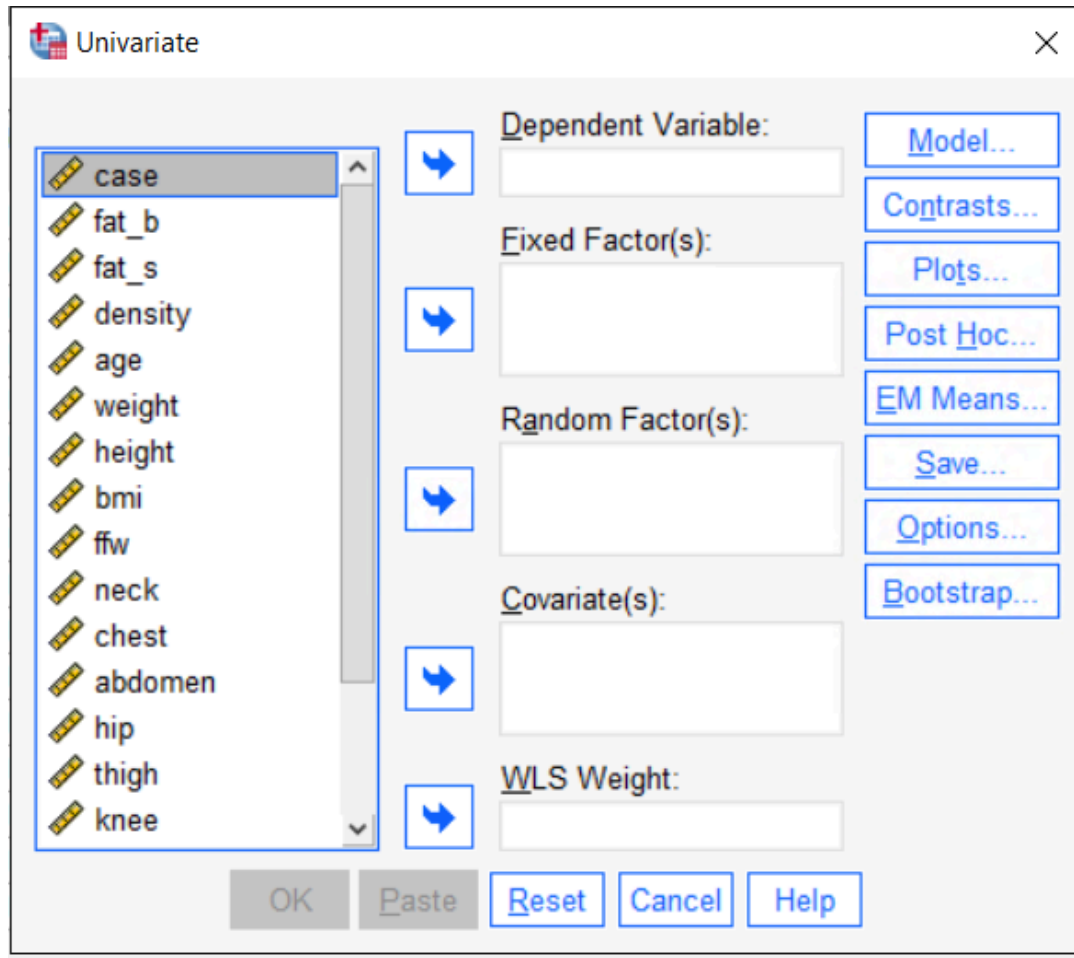
Module 02, Linear regression with multiple independent variables

- Analysis of variance table
 - R-squared
 - Partial F tests
- Stepwise regression
- Interpretation
- Collinearity
- Mediation

Module 02, Checking assumptions

- Non-normality
 - Q-Q plot of residuals
- Lack of independence
 - Assessed qualitatively
- Unequal variances, Non-linearity
 - Residual scatterplot

Module 02, SPSS dialog box for the general linear model



Module 02, SPSS computation of R-squared

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.836 ^a	.700	.696	4.2741

a. Predictors: (Constant), hip, chest, abdomen

- $10,548.480 / 15,079.017 = 0.70$

Speaker notes

You can compute R-squared from the analysis of variance table. The sum of squares for regression is 10,548.480 and the sum of squares total is 15,079.017. When you are calculating, please use the maximum precision for any intermediate calculation, but then round aggressively with the final result.

Often you will see R-squared displayed as a percentage instead of a proportion (e.g., 70% instead of 0.70).

Module 02, SPSS computation of change in R-squared

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.836 ^a	.700	.696	4.2741

a. Predictors: (Constant), hip, chest, abdomen

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.833 ^a	.693	.691	4.3110

a. Predictors: (Constant), hip, abdomen

- $Partial R^2 = 0.700 - 0.693 = 0.007$

Speaker notes

The regression model with chest, abdomen, and hip accounts for 70% of the variation, but a model with just abdomen and hip does almost as well, accounting for 69.3% of the variation. The difference, 0.7% is the amount of additional variation accounted for when you add chest to a model that already included abdomen and hip.

That seems to contradict the earlier finding with the large negative test statistic and the small p-value. But actually, what it is saying is that although there is sufficient statistical evidence to conclude that chest circumference has a real impact, that impact is small. You will see this a lot, especially with datasets with very large sample sizes. You can often achieve statistical significance for an individual variable, but the practical impact can still be negligible.

Module 02, SPSS computation of partial F-test

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9984.086	1	9984.086	489.903	<.001 ^b
	Residual	5094.931	250	20.380		
	Total	15079.017	251			
2	Regression	10548.480	3	3516.160	192.473	<.001 ^c
	Residual	4530.537	248	18.268		
	Total	15079.017	251			

a. Dependent Variable: fat_b

b. Predictors: (Constant), abdomen

c. Predictors: (Constant), abdomen, hip, chest

Speaker notes

You can place abdomen in the first block of variables. Then place hips and chest in the second block. SPSS will calculate the analysis of variance table for the reduced model (only abdomen) and the full model (abdomen, hips, and chest).

Module 02, SPSS computation of full regression model

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.229	6.214		1.163	.246
	neck	-.582	.209	-.183	-2.790	.006
	chest	-.091	.085	-.099	-1.063	.289
	abdomen	.960	.072	1.336	13.414	<.001
	hip	-.391	.113	-.362	-3.473	<.001
	thigh	.134	.125	.091	1.070	.286
	knee	-.094	.212	-.029	-.443	.658
	ankle	.004	.203	.001	.021	.983
	biceps	.111	.159	.043	.699	.485
	forearm	.345	.186	.090	1.857	.064
	wrist	-1.353	.471	-.163	-2.871	.004

a. Dependent Variable: fat_b

Speaker notes

Here's a model with all the circumference measurements included.

Module 02, SPSS computation of collinearity statistics

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	7.229	6.214		1.163	.246		
	neck	-.582	.209	-.183	-2.790	.006	.257	3.893
	chest	-.091	.085	-.099	-1.063	.289	.127	7.855
	abdomen	.960	.072	1.336	13.414	<.001	.111	9.022
	hip	-.391	.113	-.362	-3.473	<.001	.101	9.869
	thigh	.134	.125	.091	1.070	.286	.154	6.513
	knee	-.094	.212	-.029	-.443	.658	.252	3.973
	ankle	.004	.203	.001	.021	.983	.557	1.796
	biceps	.111	.159	.043	.699	.485	.286	3.500
	forearm	.345	.186	.090	1.857	.064	.470	2.128
	wrist	-1.353	.471	-.163	-2.871	.004	.341	2.933

a. Dependent Variable: fat_b

Module 2, What is mediation?

Speaker notes

The image and quote from your Andy Field textbook is about as good as anything. A mediator is a third variable which partially or totally explains the relationship between two variables. You'll see more of this next week's module, but I wanted to introduce it here, because it is a very important concept.

Module 2, SPSS assessment of mediation

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10.955	1.736		6.312	<.001
	decade	1.779	.372	.289	4.776	<.001

a. Dependent Variable: fat_b

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-36.515	2.470		-14.785	<.001
	decade	.660	.229	.107	2.884	.004
	abdomen	.567	.027	.789	21.187	<.001

a. Dependent Variable: fat_b

Speaker notes

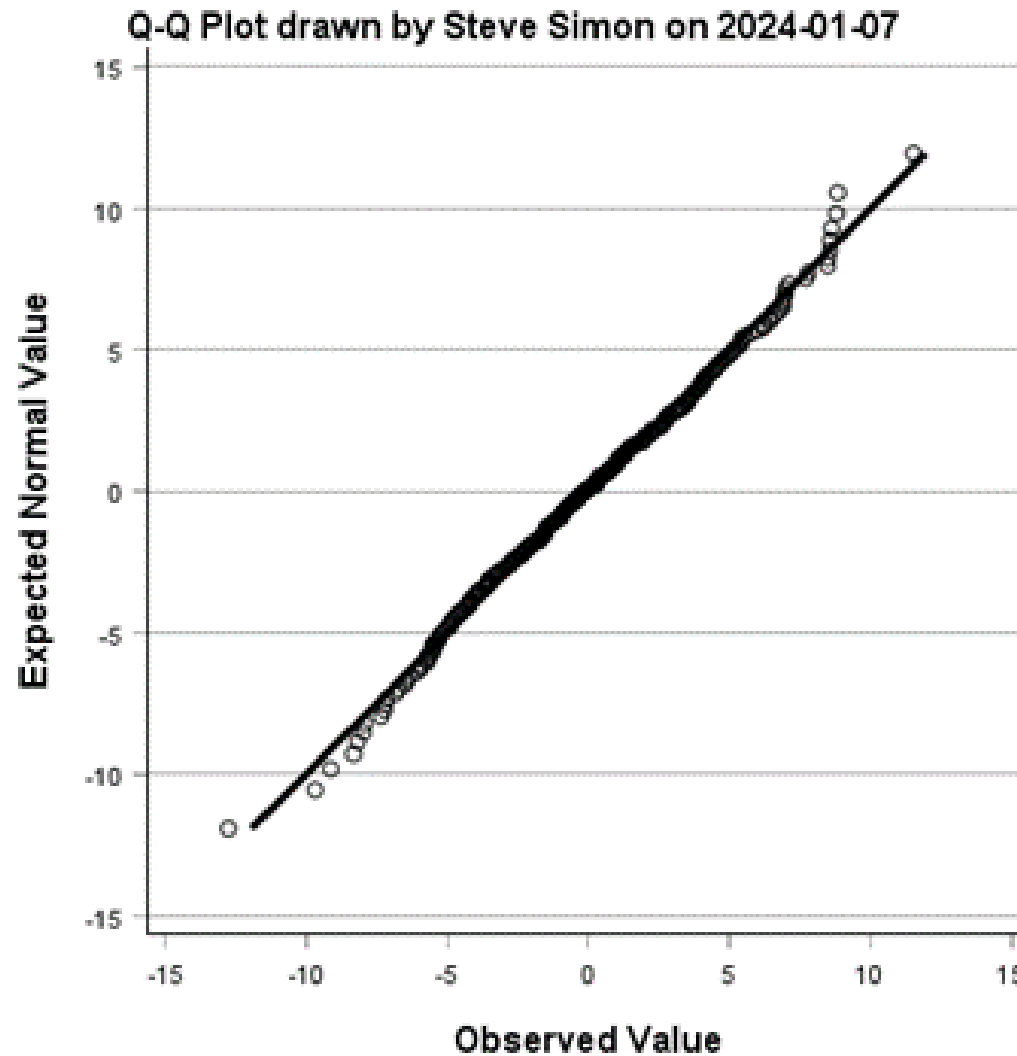
There's a relationship between age and percentage body fat. It is a bit easier to follow if you convert age in years to age in decades. The relationship between decades of age and body fat is that you add about 1.8% to your body fat every decade on average.

The question is, how does this happen? Does some of the muscle turn directly into fat, or do you add fat through an expanding girth?

When you look at a model with both decade of age and abdomen circumference, the effect of age is cut markedly. Instead of adding 1.8% body fat on average, you add about 0.7% body fat on average. The rest of the body fat comes from the "love handles" that you develop as you age.

So gaining some fat is inevitable, even if you can still fit into those size 32 jeans that you wore as a young man. But much of the gain in percentage body fat comes from moving from size 32 over time to size 42.

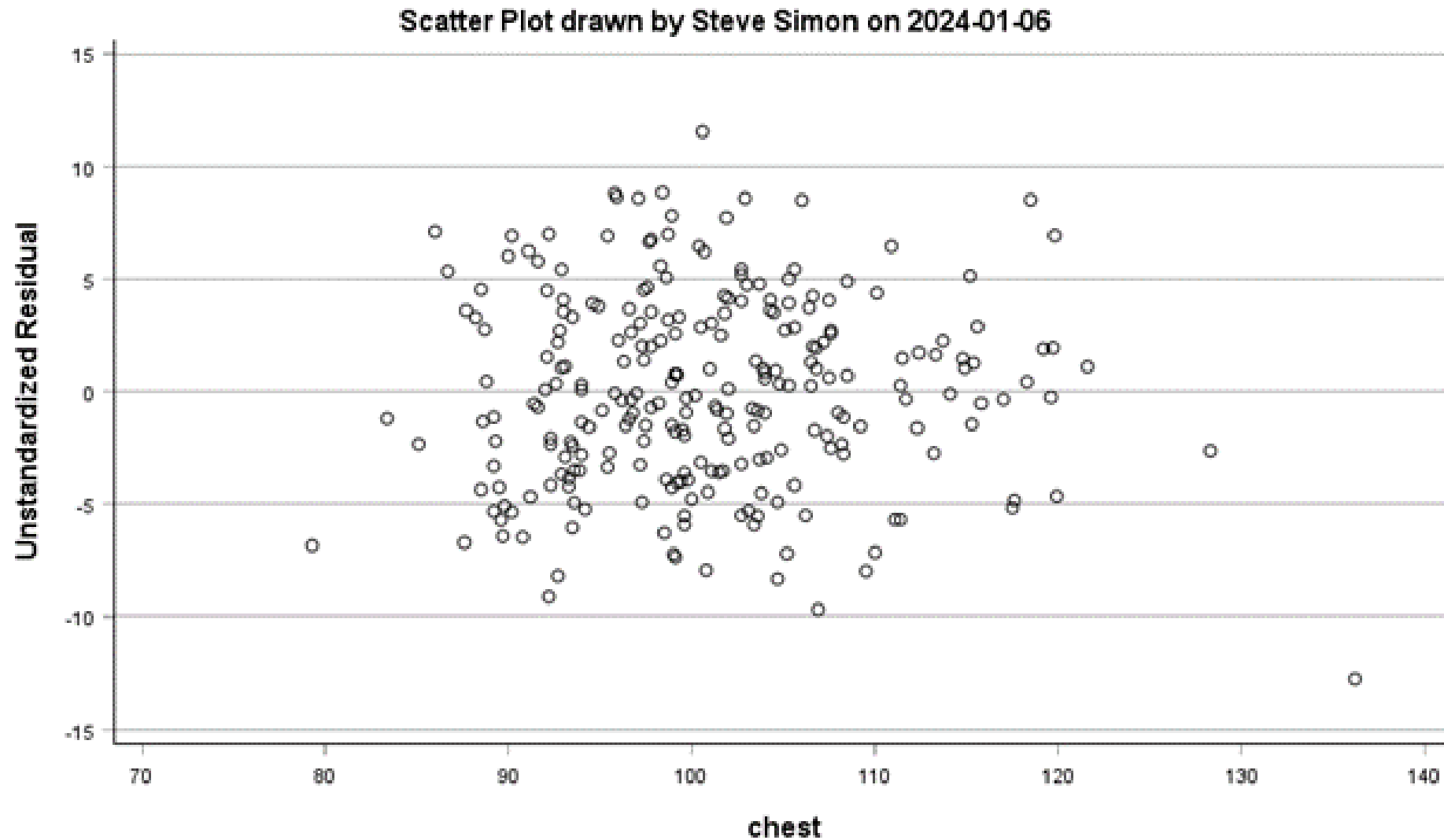
Module 02, SPSS Q-Q plot



Speaker notes

This is the QQ plot of residuals. It looks like a straight line, indicating that the normality assumption is reasonable.

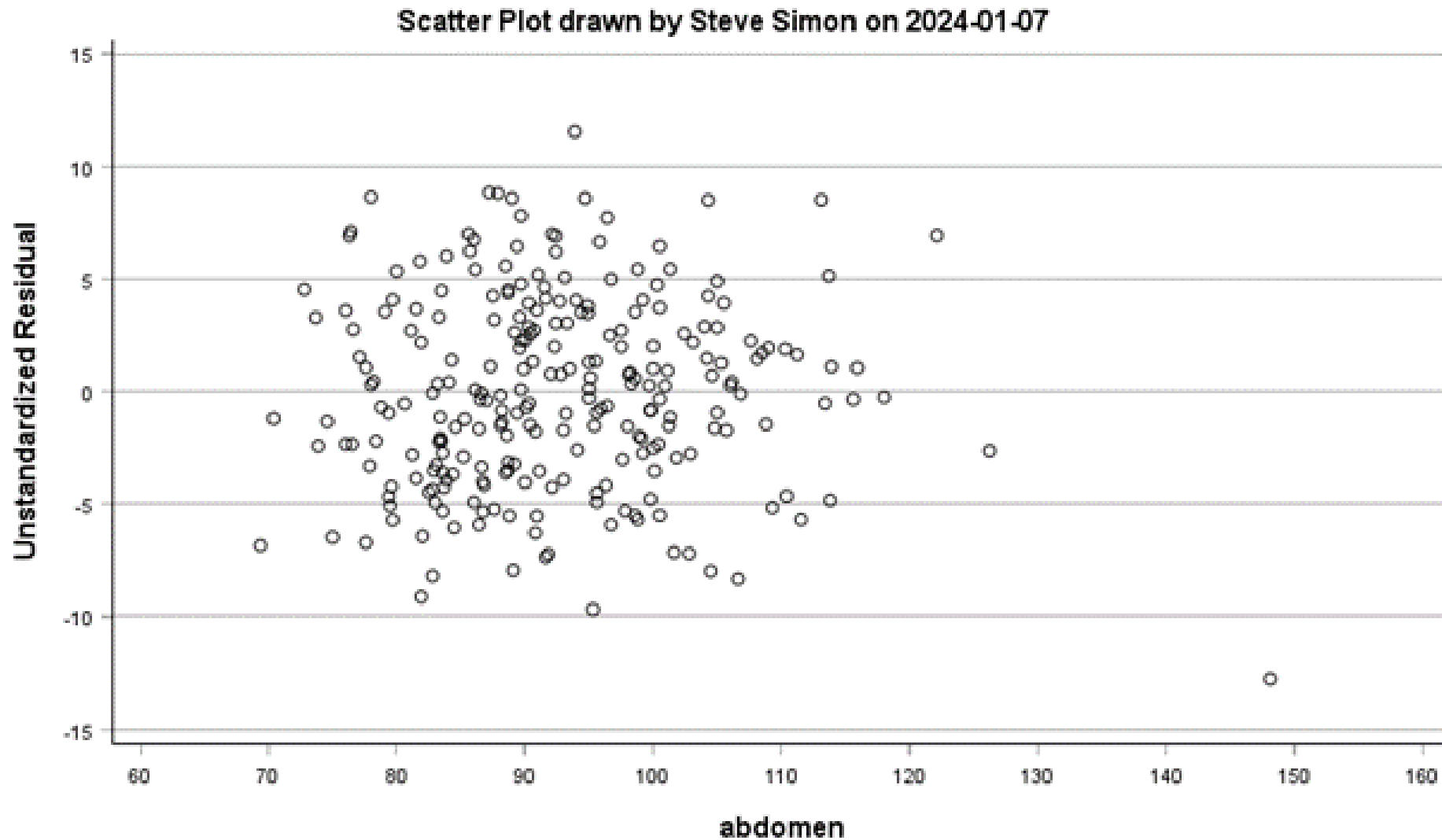
Module 02, SPSS Scatterplot, 1 of 4



Speaker notes

There is no apparent problem with unequal variances or non-linearity with respect to chest circumference ...

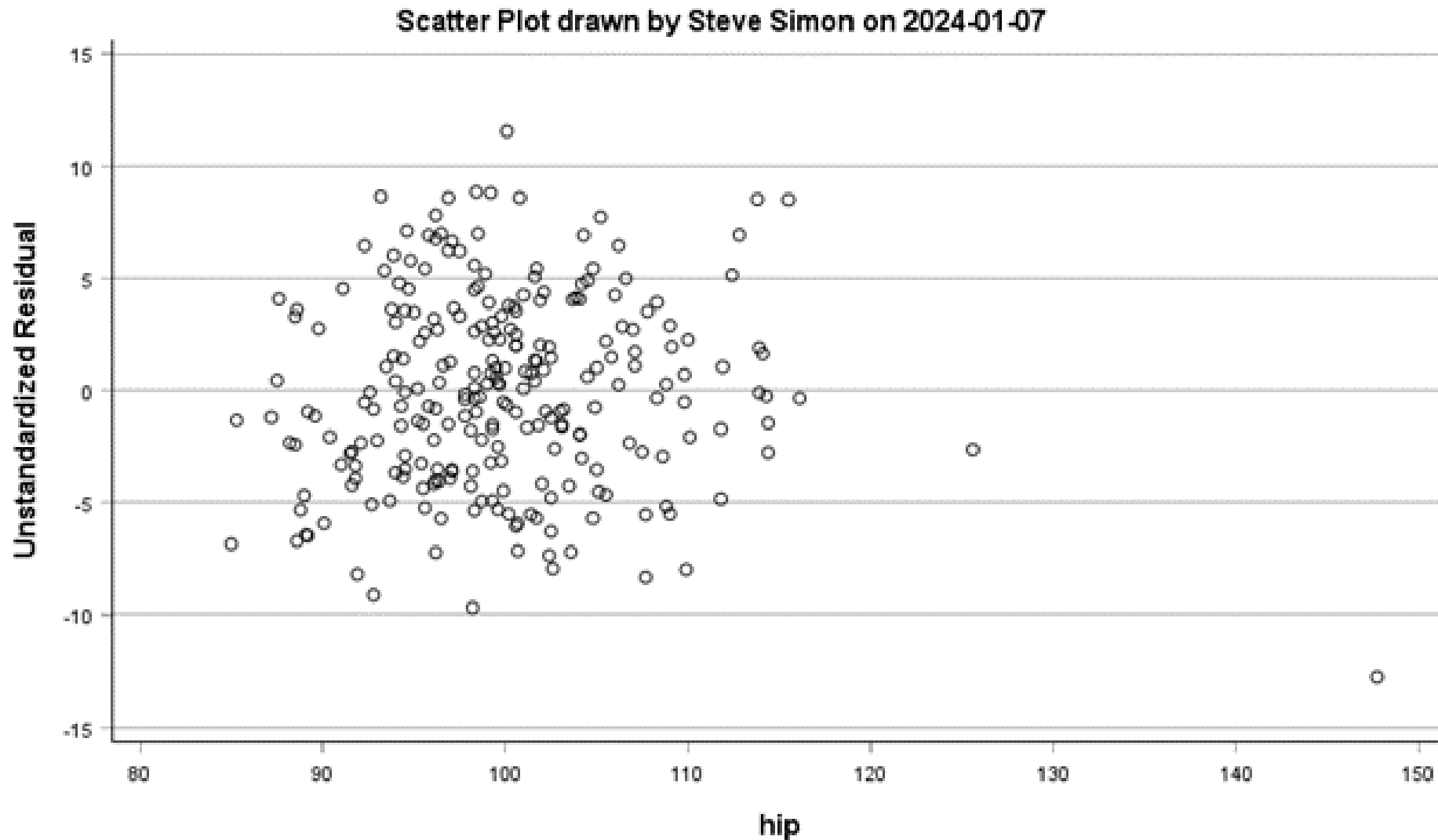
Module 02, SPSS Scatterplot, 2 of 4



Speaker notes

... or abdomen circumference ...

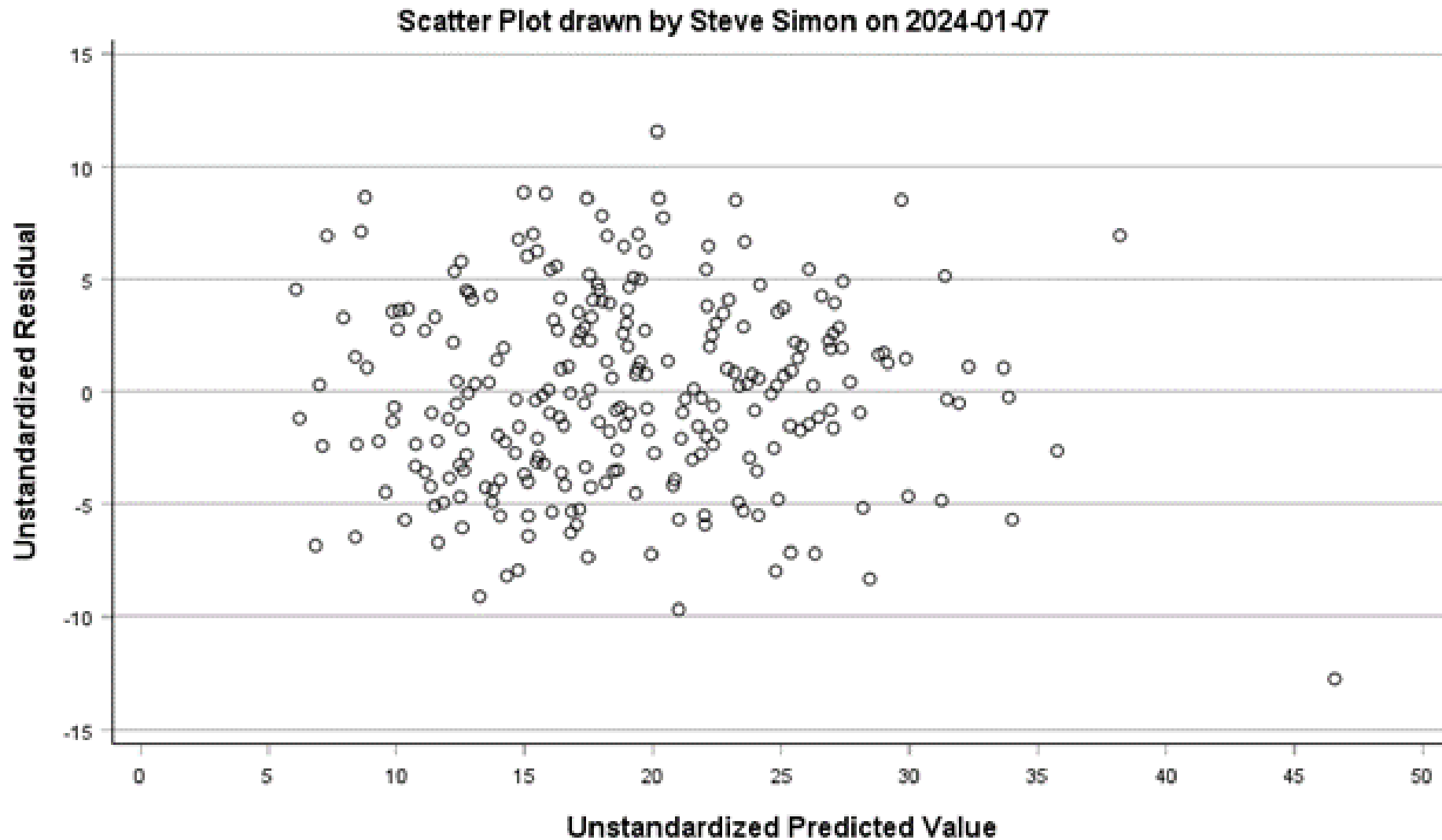
Module 02, SPSS Scatterplot, 3 of 4



Speaker notes

... or hip circumference.

Module 02, SPSS Scatterplot, 4 of 4



Speaker notes

While three variables is not too many, here is what the plot of residuals versus predicted values looks like. Notice that it has pretty much the same features as the three individual plots.

Break #2

- What you have learned
 - 02 Linear regression with multiple independent variables
- What's coming next
 - 03 Analysis of covariance

Module 03, Analysis of covariance

- Confounding/covariate imbalance
- Interpretation
- Interactions

Module 03, Checking assumptions

- Non-normality
 - Q-Q plot of residuals
- Lack of independence
 - Assessed qualitatively
- Unequal variances, Non-linearity
 - Residual scatterplots

Module 03, SPSS calculation of unadjusted estimates

Parameter Estimates

Dependent Variable: price_in_thousands

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	144.678	6.115	23.658	<.001	132.564	156.791
[cust=No]	-49.926	6.973	-7.160	<.001	-63.737	-36.114
[cust=Yes]	0 ^a

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

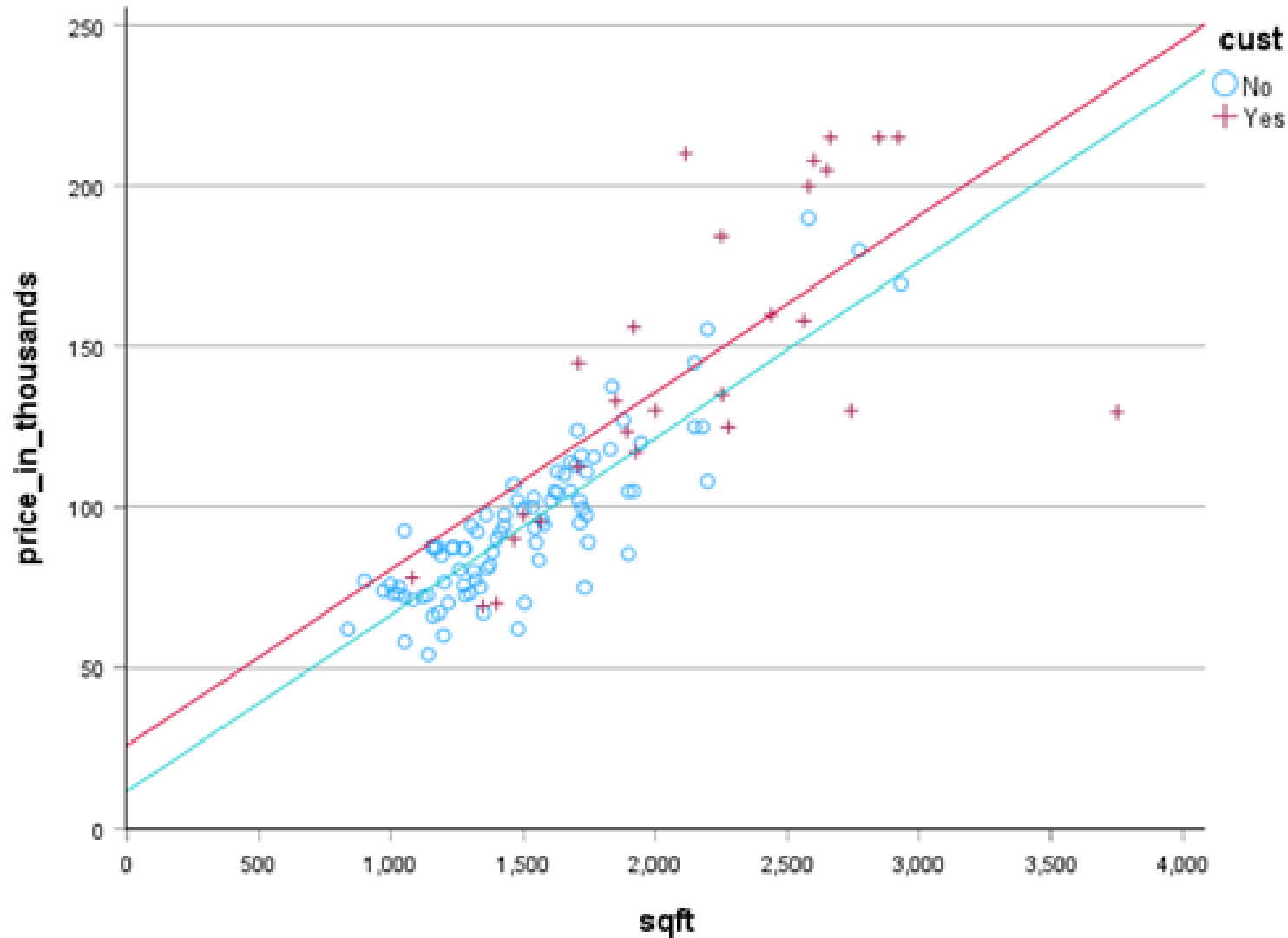
Module 03, SPSS calculation of adjusted estimates

Parameter Estimates

Dependent Variable: price_in_thousands

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	11.413	6.550	1.743	.084	-1.562	24.389
i_custom	14.286	5.103	2.800	.006	4.177	24.395
sqft	.055	.004	13.428	<.001	.047	.064

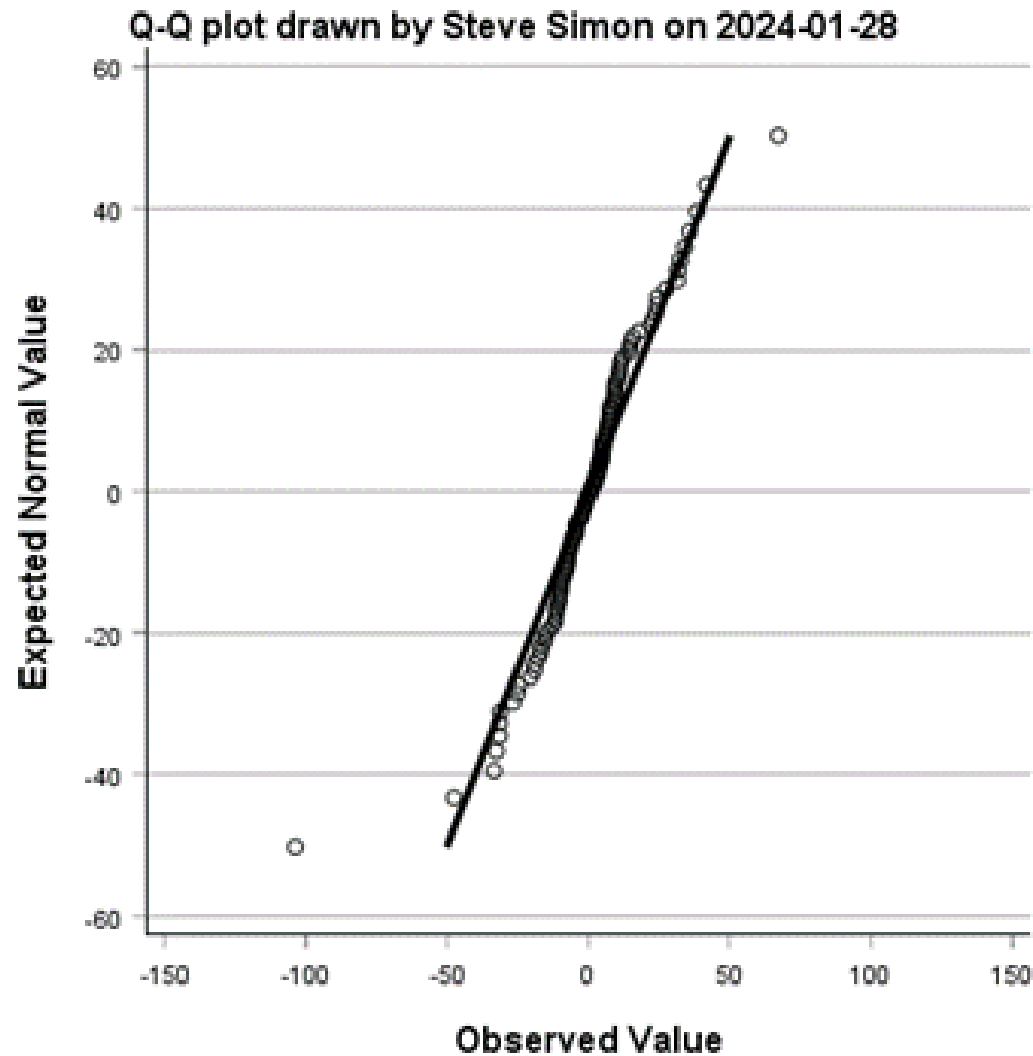
Module 03, SPSS visualization, 1 of 2



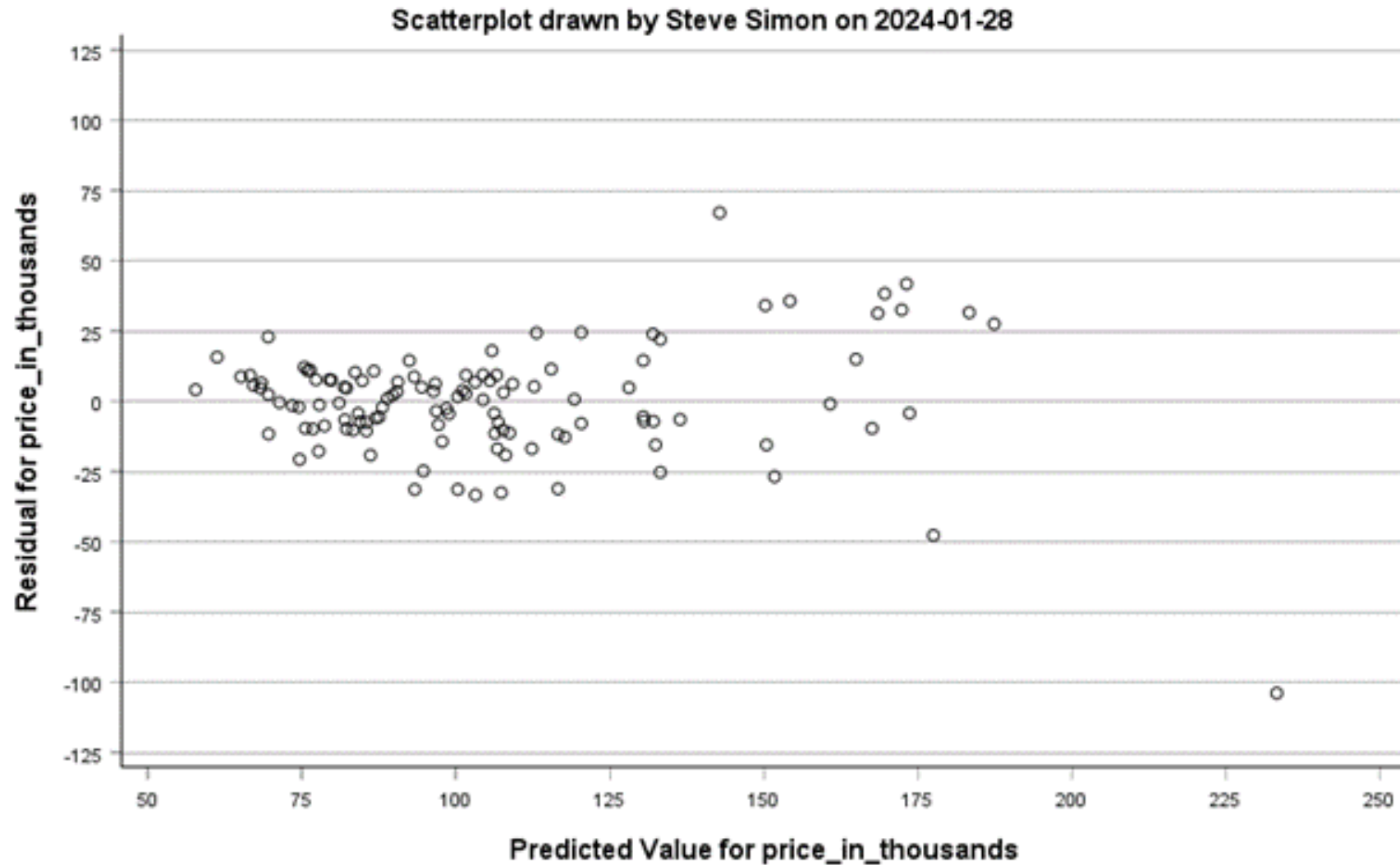
Module 03, SPSS visualization, 2 of 2



Module 03, SPSS Q-Q plot



Module 03, SPSS scatterplot



Module 03, SPSS interaction test

Parameter Estimates

Dependent Variable: price_in_thousands

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	10.119	8.360	1.210	.229	-6.443	26.681
i_custom	18.275	16.710	1.094	.276	-14.831	51.381
sqft	.056	.005	10.460	<.001	.046	.067
i_custom * sqft	-.002	.008	-.251	.802	-.019	.015

Break #3

- What you have learned
 - 03 Analysis of covariance
- What's coming next
 - 04 Multi-factor analysis of variance

Module 04, Multi-factor analysis of variance

- Tukey post hoc test
- Interaction

Module 04, Checking assumptions

- Non-normality
 - Q-Q plot of residuals
- Lack of independence
 - Assessed qualitatively
- Unequal variances
 - Boxplots

Module 04, SPSS crosstabulation

Month * Moon Crosstabulation

Count

		Moon			
		After	Before	During	Total
Month	APR	1	1	1	3
	AUG	1	1	1	3
	DEC	1	1	1	3
	FEB	1	1	1	3
	JAN	1	1	1	3
	JUL	1	1	1	3
	JUN	1	1	1	3
	MAR	1	1	1	3
	MAY	1	1	1	3
	NOV	1	1	1	3
	OCT	1	1	1	3
	SEP	1	1	1	3
Total		12	12	12	36

Speaker notes

This table shows how there is one observation for each combination of month and moon phase.

Module 04, SPSS analysis of variance table

Tests of Between-Subjects Effects

Dependent Variable: Admission

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	497.097 ^a	13	38.238	6.581	<.001
Intercept	5124.174	1	5124.174	881.961	<.001
Moon	41.514	2	20.757	3.573	.045
Month	455.583	11	41.417	7.129	<.001
Error	127.819	22	5.810		
Total	5749.090	36			
Corrected Total	624.916	35			

a. R Squared = .795 (Adjusted R Squared = .675)

Speaker notes

This is the analysis of variance table. There is one less degree of freedom than the number of categories for each categorical predictor variable. There is a statistically significant difference between the twelve months and a borderline significant difference between the three moon phases.

This differs from the single factor analysis of variance because adding in month as a categorical predictor removed a lot of variation. You are now able to account for almost 80% of the variation in admission rates. Without month in the model, you accounted for less than 7% of the variation.

Module 04, SPSS removing irrelevant rows

Tests of Between-Subjects Effects

Dependent Variable: Admission

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	497.097 ^a	13	38.238	6.581	<.001
Moon	41.514	2	20.757	3.573	.045
Month	455.583	11	41.417	7.129	<.001
Error	127.819	22	5.810		
Corrected Total	624.916	35			

a. R Squared = .795 (Adjusted R Squared = .675)

Speaker notes

The rows corresponding to the intercept and the total (uncorrected total) are not needed.

Module 04, SPSS parameter estimates

Parameter Estimates

Dependent Variable: Admission

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	11.253	1.503	7.486	<.001	8.135	14.370
[Moon=After]	-1.958	.984	-1.990	.059	-3.999	.082
[Moon=Before]	-2.500	.984	-2.541	.019	-4.541	-.459
[Moon=During]	0 ^a
[Month=APR]	8.967	1.968	4.556	<.001	4.885	13.048
[Month=AUG]	-4.033	1.968	-2.049	.053	-8.115	.048
[Month=DEC]	-1.400	1.968	-.711	.484	-5.482	2.682
[Month=FEB]	2.900	1.968	1.474	.155	-1.182	6.982
[Month=JAN]	1.000	1.968	.508	.616	-3.082	5.082
[Month=JUL]	7.000	1.968	3.557	.002	2.918	11.082
[Month=JUN]	3.067	1.968	1.558	.133	-1.015	7.148
[Month=MAR]	4.533	1.968	2.303	.031	.452	8.615
[Month=MAY]	4.567	1.968	2.320	.030	.485	8.648
[Month=NOV]	-.333	1.968	-.169	.867	-4.415	3.748
[Month=OCT]	-.300	1.968	-.152	.880	-4.382	3.782
[Month=SEP]	0 ^a

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

Speaker notes

The intercept represents the average admission rate during a full moon when the month is September. The two slope terms show how much lower the average admission rates are before and after a full moon, respectively, compared to during a full moon, holding month constant.

Module 04, SPSS Tukey test

Multiple Comparisons

Dependent Variable: Admission

Tukey HSD

(I) Moon	(J) Moon	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
After	Before	.542	.9840	.847	-1.930	3.014
	During	-1.958	.9840	.138	-4.430	.514
Before	After	-.542	.9840	.847	-3.014	1.930
	During	-2.500*	.9840	.047	-4.972	-.028
During	After	1.958	.9840	.138	-.514	4.430
	Before	2.500*	.9840	.047	.028	4.972

Based on observed means.

The error term is Mean Square(Error) = 5.810.

*. The mean difference is significant at the .05 level.

Admission

Tukey HSD^{a,b}

Moon	N	Subset	
		1	2
Before	12	10.917	
After	12	11.458	11.458
During	12		13.417
Sig.		.847	.138

Means for groups in homogeneous subsets are displayed.

Based on observed means.

The error term is Mean Square(Error) = 5.810.

a. Uses Harmonic Mean Sample Size = 12.000.

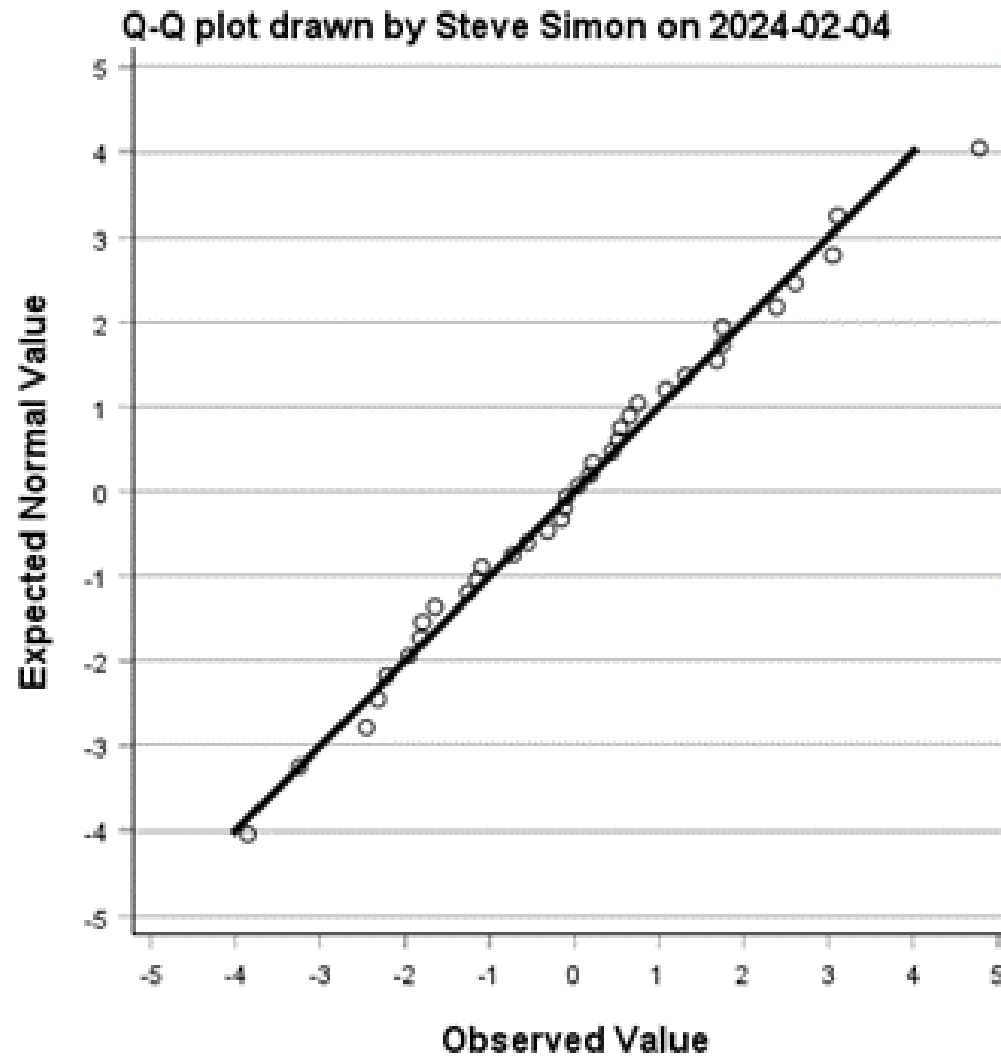
b. Alpha = .05.

Speaker notes

Use the Tukey posthoc test because the sample sizes are equal across the moon phases. The results are a bit ambiguous because before and after are not statistically different, after and during are not statistically different but before and during are statistically different. This is probably due to a lack of precision and an extra year's worth of data would help quite a bit.

The analogy I use is travel time. My wife and I live in Leawood. Our son lives in Lee's Summit. A repair shop we all use is in Olathe. It is not far from Leawood to Olathe. It is not far from Leawood to Lee's Summit. But it is far from Lee's Summit to Olathe.

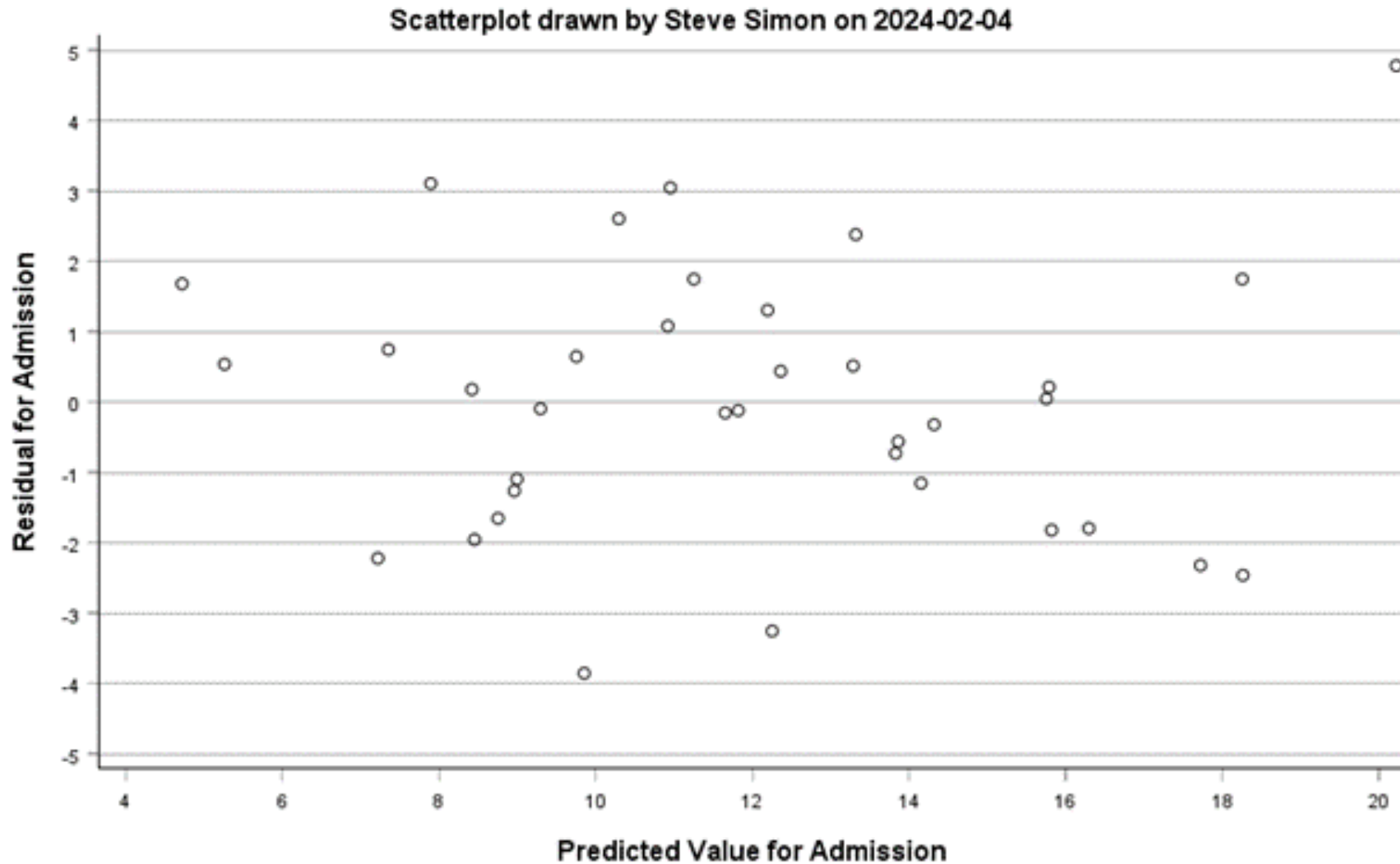
Module 04, SPSS Q-Q plot



Speaker notes

The residuals from the full moon analysis of variance model appear to be normally distributed.

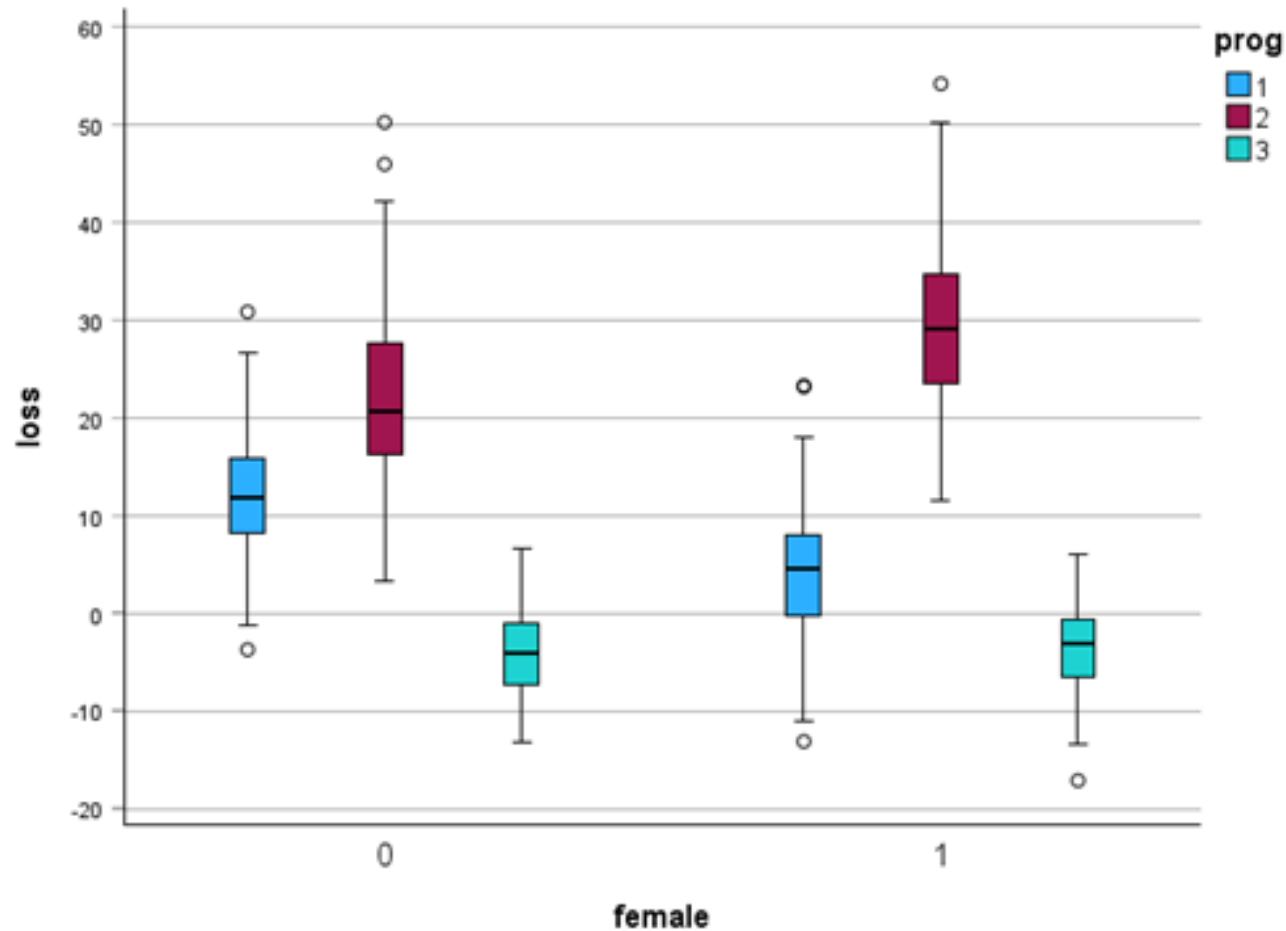
Module 04, SPSS scatterplot



Speaker notes

There is no evidence of unequal variances.

Module 04, SPSS, Box plots of exercise data



Speaker notes

Here is a clustered boxplot of weight loss in a study of two exercise interventions: jogging and swimming, and a control intervention, reading. This study shows a clear interaction between gender and exercise. Both jogging and swimming provide greater weight loss than reading (no big surprise). Swimming appears to be better than jogging. But it is a lot better for females and only a little bit better for males.

Module 04, SPSS, Mean values for the interaction

Report

loss

female	prog	Mean	N	Std. Deviation
0	1	11.7720	150	5.98881
	2	22.5224	150	8.59176
	3	-3.9556	150	4.14022
	Total	10.1129	450	12.67177
1	1	4.2887	150	6.88500
	2	29.1177	150	7.99691
	3	-3.6201	150	4.09571
	Total	9.9287	450	15.41056
Total	1	8.0304	300	7.45267
	2	25.8200	300	8.91992
	3	-3.7879	300	4.11456
	Total	10.0208	900	14.10024

Speaker notes

In this table of means, notice that men lose about 11 pounds on the jogging program, and 22 pounds on the swimming program. So swimming is better. For women, the losses are about 4 pounds on average with jogging and 30 pounds on swimming. The extra benefits of swimming are so much larger in females than in males.

Module 04, SPSS, Analysis of variance table for interaction model

Tests of Between-Subjects Effects

Dependent Variable: loss

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	140748.038 ^a	5	28149.608	662.463	<.001
Intercept	90375.473	1	90375.473	2126.864	<.001
female	7.634	1	7.634	.180	.672
prog	133277.206	2	66638.603	1568.249	<.001
female * prog	7463.198	2	3731.599	87.818	<.001
Error	37988.163	894	42.492		
Total	269111.674	900			
Corrected Total	178736.201	899			

a. R Squared = .787 (Adjusted R Squared = .786)

Speaker notes

The analysis of variance table shows a large F ratio for the interaction between exercise program and sex.

Module 04, SPSS, Parameter estimates for the interaction model

Parameter Estimates

Dependent Variable: loss

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-3.620	.532	-6.802	<.001	-4.665	-2.576
[female=0]	-.335	.753	-.446	.656	-1.813	1.142
[female=1]	0 ^a
[prog=1]	7.909	.753	10.507	<.001	6.432	9.386
[prog=2]	32.738	.753	43.494	<.001	31.261	34.215
[prog=3]	0 ^a
[female=0] * [prog=1]	7.819	1.064	7.345	<.001	5.730	9.908
[female=0] * [prog=2]	-6.260	1.064	-5.881	<.001	-8.349	-4.171
[female=0] * [prog=3]	0 ^a
[female=1] * [prog=1]	0 ^a
[female=1] * [prog=2]	0 ^a
[female=1] * [prog=3]	0 ^a

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

Speaker notes

The intercept is the estimated average weight loss when all the indicator variables are equal to zero. That means the two reference categories: females for gender and reading for the exercise group. If you join a book club expect to gain more than three pounds.

The estimate for $\text{female}=0$ is the how much different the weight loss is in the reading group when you change gender from female to male. It's a small change and not statistically significant.

The estimate for $\text{group}=1$ measures how much difference you see in the average weight loss in the jogging program compared to the reading program in the reference category (female). There is almost an 8 pound difference and this is statistically significant.

The estimate for $\text{group}=2$ measures how much difference you see in the average weight loss in the swimming program compared to the reading program in the reference category (female). There is an astounding 32 plus pound difference between swimming and reading, but again for the reference category of females.

The estimate for $\text{female}=0 * \text{prog}=1$ shows how much different the benefit of jogging over reading is for men compared to women. Men did pretty well on the jogging program, losing 11 pounds instead of gaining more than 3 pounds. This is more impressive than for women who lost 4 pounds instead of gaining 3.

The estimate for $\text{female}=0 * \text{prog}=2$ shows how much different the benefit of swimming is for men than for women. Both groups are better off swimming than reading, men losing 22 pounds versus a 3 pound gain. But the falls short of the benefit of swimming over reading for women a 29 pound loss versus a 3 pound gain.

Both of the parameters associated with the interaction are statistically significant.

Bottom line is that the benefits of jogging over reading and the benefits of swimming over reading are not the same for men and women.

Module 04, SPSS, Interaction plot, 1 of 2

Speaker notes

An interaction plot draws a multiple lines connecting means. In this graph the top line shows the mean weight loss in the swimming program with the men's mean on the left and the women's mean on the right. The middle line shows the mean weight loss in the jogging program, again with the men's mean on the left and the women's on the right. The bottom line shows the mean weight loss (actually a weight gain!) in the reading program. Clearly the swimming program is the best, a bit better for women than men because the line slopes upward. The jogging program is second best, a bit worse for women than for men. The reading program is worst and the benefit, if you could call it that being about the same for men and women (the lines are flat).

It is a lack of parallelism that is the hallmark of an interaction.

Module 04, SPSS, Interaction plot, 2 of 2

Speaker notes

You could draw the interaction plot differently with a line for women (red) and a line for men (blue). The interpretation is about the same, perhaps, but the emphasis is different. The superior weight loss for men in the jogging program and the superior weight loss for women in the swimming program is emphasized by the crossing lines. Both the men's and women's lines almost touching for the reading program emphasizes the equivalent results for the two genders.

Which plot is better? I like the first one, but would not complain if you liked the second one better. It depends on what story you want to emphasize. The first graph emphasizes the difference between the diets a bit more strongly and the second emphasizes the differences between the genders a bit more strongly.

Module 04, When you can't estimate an interaction

- Special case, $n=1$
 - Only one observation for categorical combination

Speaker notes

There is a special case where you have two categorical independent variables and you cannot estimate an interaction. If you have $n=1$, exactly one observation for each combination of your two categorical variables, then you don't have enough degrees of freedom to estimate an interaction and still be able to test whether that interaction is statistically significant.

It's sort of like that old joke I told about married life (it's okay but you lose a degree of freedom). Interactions cause an even bigger loss of degrees of freedom and in the case with only one observation per combination of categories, you lose enough degrees of freedom that it is not marriage, it being in prison.

Module 04, SPSS, Example, full moon study, 1 of 2

Month * Moon Crosstabulation

Count

		Moon			
		After	Before	During	Total
Month	APR	1	1	1	3
	AUG	1	1	1	3
	DEC	1	1	1	3
	FEB	1	1	1	3
	JAN	1	1	1	3
	JUL	1	1	1	3
	JUN	1	1	1	3
	MAR	1	1	1	3
	MAY	1	1	1	3
	NOV	1	1	1	3
	OCT	1	1	1	3
	SEP	1	1	1	3
Total		12	12	12	36

Speaker notes

Here is an example where you only have one observation for each combination of categories.

Module 04, SPSS, Example, full moon study, 2 of 2

Tests of Between-Subjects Effects

Dependent Variable: Admission

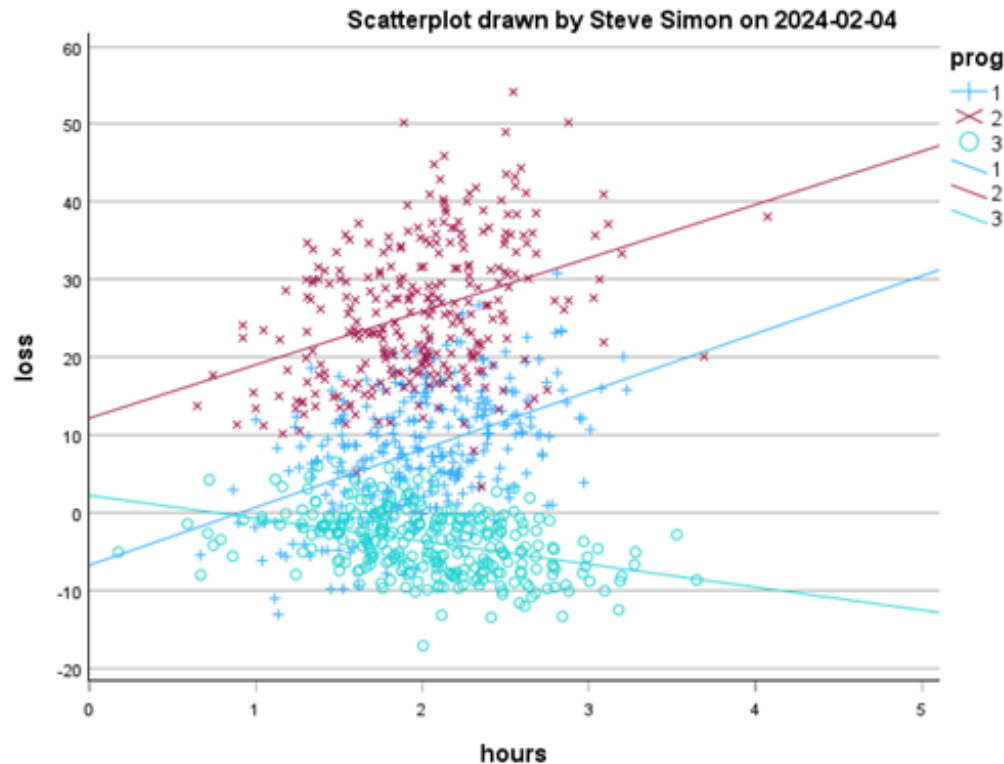
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Moon	41.514	2	20.757	.	.
Month	455.583	11	41.417	.	.
Moon * Month	127.819	22	5.810	.	.
Error	.000	0	.		
Corrected Total	624.916	35			

Speaker notes

You lose two degrees of freedom for moon (3 phases: before, during, and after). You lose 11 degrees of freedom for month (12 months -1). The interaction has 2 times 11 or 22 degrees of freedom. You only started with 35 degrees of freedom. Subtract 2, 11, and 22, and you are left with zero degrees of freedom for error.

If you find yourself in this situation, just state that no test for interaction was possible in your methods section and highlight this as a weakness of your study in the discussion section.

Module 04, SPSS, Interaction between exercise program and hours spent exercising



Speaker notes

This plot shows a marked interaction. If you looked just at prog 1 (jogging program) and prog 2 (swimming program), there is no interaction. The more hours you spent on the program the more weight you lost.

The reading program, however, has a different slope. Instead of increasing steadily, the more time you spent reading the less weight that you lost. You might even gain some weight if you followed the reading program religiously.

Module 04, SPSS, Testing for interaction in analysis of covariance

Tests of Between-Subjects Effects

Dependent Variable: loss

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	140940.743 ^a	5	28188.149	666.752	<.001
Intercept	326.802	1	326.802	7.730	.006
prog	2901.898	2	1450.949	34.320	<.001
hours	3116.205	1	3116.205	73.710	<.001
prog * hours	5319.048	2	2659.524	62.907	<.001
Error	37795.457	894	42.277		
Total	269111.674	900			
Corrected Total	178736.201	899			

a. R Squared = .789 (Adjusted R Squared = .787)

Speaker notes

This table shows a statistically significant interaction. The F ratio is large and the p-value is small.

Module 04, SPSS, Table with irrelevant rows removed

Tests of Between-Subjects Effects

Dependent Variable: loss

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	140940.743 ^a	5	28188.149	666.752	<.001
prog	2901.898	2	1450.949	34.320	<.001
hours	3116.205	1	3116.205	73.710	<.001
prog * hours	5319.048	2	2659.524	62.907	<.001
Error	37795.457	894	42.277		
Corrected Total	178736.201	899			

a. R Squared = .789 (Adjusted R Squared = .787)

Speaker notes

Remove the rows associated with the intercept and the (uncorrected) total.

Module 04, SPSS, Parameter estimates

Parameter Estimates

Dependent Variable: loss

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	2.216	1.486	1.491	.136	-.700	5.133
[prog=1]	-8.997	2.216	-4.060	<.001	-13.346	-4.648
[prog=2]	9.933	2.177	4.562	<.001	5.660	14.205
[prog=3]	0 ^a
hours	-2.956	.708	-4.176	<.001	-4.346	-1.567
[prog=1] * hours	10.409	1.072	9.708	<.001	8.304	12.513
[prog=2] * hours	9.830	1.051	9.349	<.001	7.767	11.894
[prog=3] * hours	0 ^a

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

Speaker notes

The first three coefficients are a bit of an extrapolation because they represent patients who spend zero hours on the proposed intervention.

The intercept is the estimated average weight loss in the reference category (group=3 or reading program) when the number of hours devoted to the program is zero. In this case it is a weight gain of 2 pounds.

The coefficient for group=1 is how much better the weight loss is when you switch from the reading program to the jogging program, but still put in zero hours.

The coefficient for group=2 is how much better the weight loss is on average when you switch from the reading program to the jogging program, but still put in zero hours.

The coefficient for hours is how much better the weight loss is when you add an extra hour of effort and you are in the reading program. Actually, it is how much worse. Each extra hour of reading and you gain an extra three pounds on average.

The coefficient for hours*group=1 shows the change in slope when you switch from the reading program to the jogging program. Each hour invested is 10 pounds better (a loss of 7 pounds per hour invested instead of a gain of 3 pounds per hour invested) when you switch from the reading program to the jogging program.

The interpretation for hours*group=2 is similar It shows the change in slope when you switch from the reading program to the swimming program. Each hour invested is 10 pounds better (a loss of 7 pounds per hour invested instead of a gain of 3 pounds per hour invested) when you switch from the reading program to the swimming program.

Module 04, SPSS, Analysis of variance table

Parameter Estimates

Dependent Variable: loss

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	7.799	11.604	.672	.502	-14.975	30.572
hours	-9.376	5.664	-1.655	.098	-20.492	1.740
effort	-.080	.385	-.209	.835	-.835	.675
hours * effort	.393	.188	2.098	.036	.025	.761

Module 04, SPSS, Table of means

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
hours	900	.18	4.07	2.0024	.49454
effort	900	12.95	44.08	29.6592	5.14276
Valid N (listwise)	900				

Module 04, SPSS, Centered analysis

Parameter Estimates

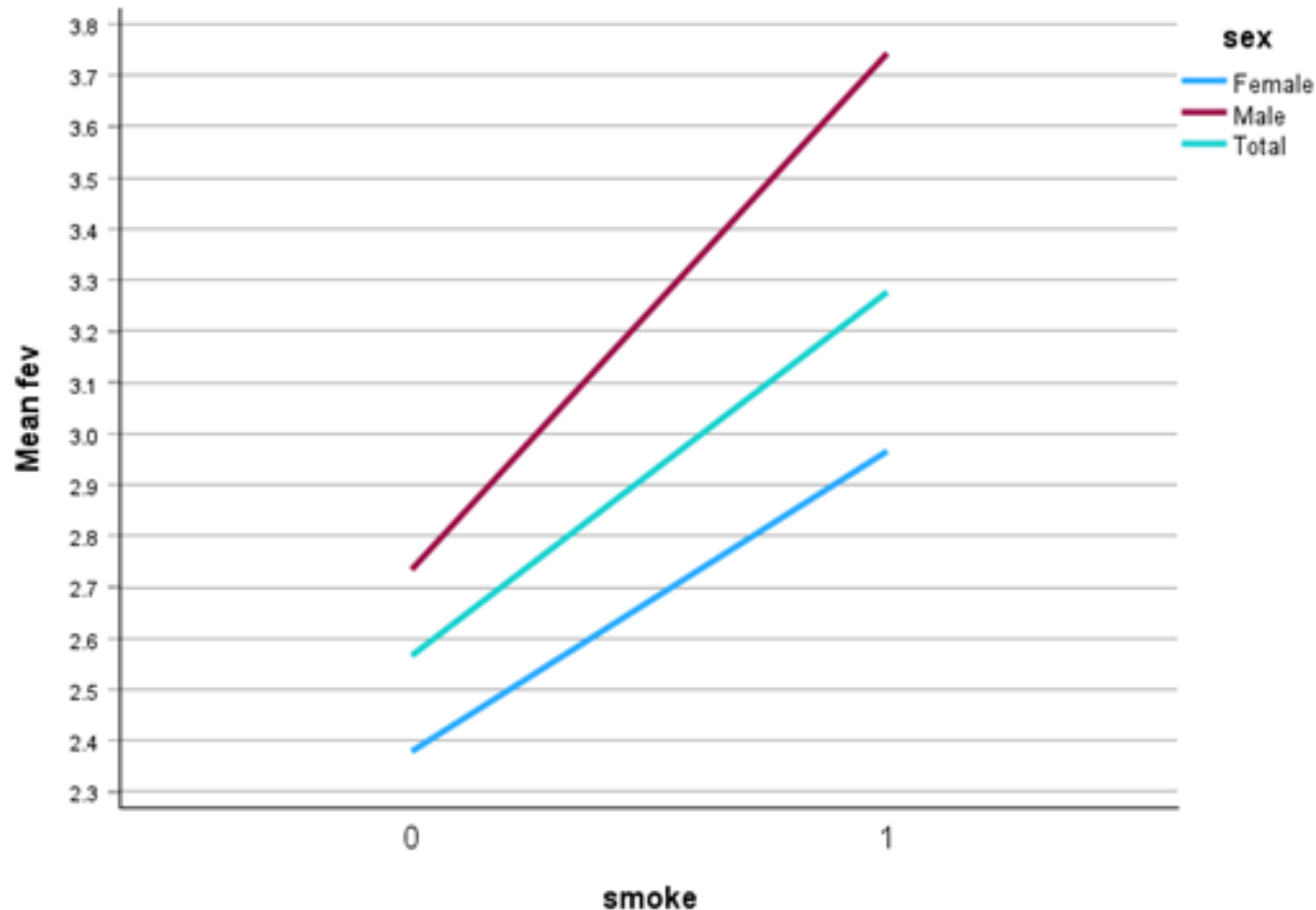
Dependent Variable: loss

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	10.005	.452	22.130	<.001	9.117	10.892
hours_centered	2.291	.915	2.503	.012	.495	4.087
effort_centered	.707	.088	8.042	<.001	.535	.880
hours_centered * effort_centered	.393	.188	2.098	.036	.025	.761

Module 04, Weight loss at various conditions

- hours = 2 (mean), effort = 30 (mean),
 - $\hat{Y} = 10.005$
- hours = 4 (mean+2), effort = 30 (mean),
 - $\hat{y} = 10.005 + 2.291*2 = 14.587$
- hours = 2 (mean), effort = 40 (mean+20)
 - $\hat{Y} = 10.005 + 0.707*20 = 24.145$
- hours = 4 (mean+2), effort = 40 (mean+20)
 - $\hat{Y} = 10.005 + 2.291*2 + 0.707*20 + 0.393*2*20 = 44.447$

Module 04, SPSS, Line plots of means for unbalanced data



Module 04, SPSS, Table of means

Report

fev

sex	smoke	Mean	N	Std. Deviation
Female	0	2.38	279	.64
	1	2.97	39	.42
	Total	2.45	318	.65
Male	0	2.73	310	.97
	1	3.74	26	.89
	Total	2.81	336	1.00
Total	0	2.57	589	.85
	1	3.28	65	.75
	Total	2.64	654	.87

Module 04, SPSS, Table of frequencies and column percentages

sex * smoke Crosstabulation

			smoke		
			0	1	Total
sex	Female	Count	279	39	318
		% within smoke	47.4%	60.0%	48.6%
	Male	Count	310	26	336
		% within smoke	52.6%	40.0%	51.4%
Total	Count		589	65	654
	% within smoke		100.0%	100.0%	100.0%

Break #4

- What you have learned
 - 04 Multi-factor analysis of variance
- What's coming next
 - 05 Dimension reduction

Module 05, Dimension reduction

- Principal components analysis
 - Eigenvectors, Eigenvalues
- Factor analysis
 - Factor rotation

Module 05, SPSS, Correlation matrix, 1 of 3

[illegible]

Module 05, SPSS, Correlation matrix, 2 of 3

Module 05, SPSS, Correlation matrix, 3 of 3

Correlation Matrix					
		To what extent do you feel that physical pain prevents you from doing what you need to do?	How much do you need any medical treatment to function in your daily life?	How much do you enjoy life?	To what extent do you feel your life to be meaningful?
Correlation	To what extent do you feel that physical pain prevents you from doing what you need to do?	1.0	.3	.4	.4
	How much do you need any medical treatment to function in your daily life?	.3	1.0	.2	.0
	How much do you enjoy life?	.4	.2	1.0	.8
	To what extent do you feel your life to be meaningful?	.4	.0	.8	1.0

Module 05, SPSS, Communalities

Communalities

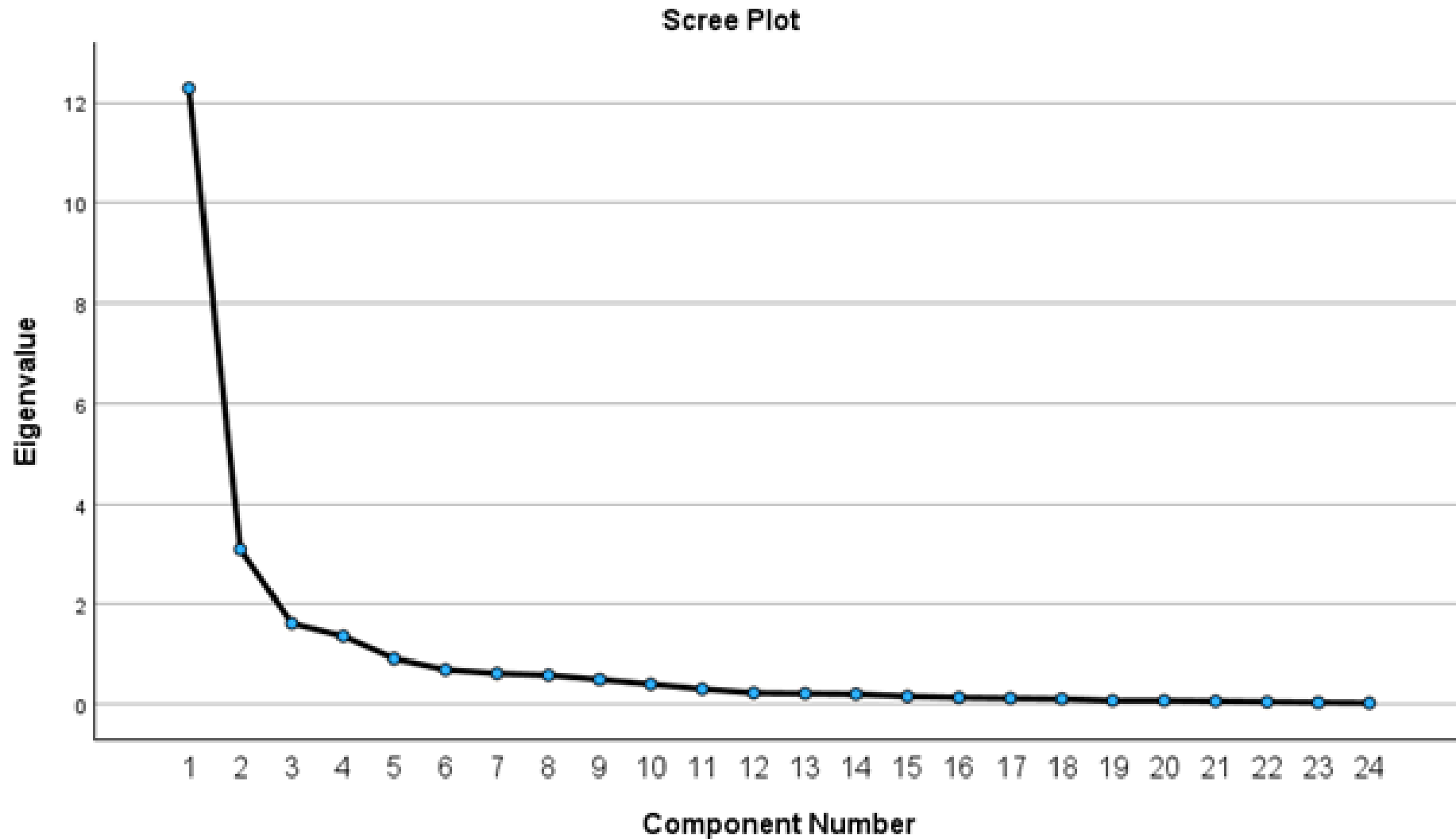
	Initial	Extraction
To what extent do you feel that physical pain prevents you from doing what you need to do?	1.000	.496
How much do you need any medical treatment to function in your daily life?	1.000	.607
How much do you enjoy life?	1.000	.814
To what extent do you feel your life to be meaningful?	1.000	.774
How well are you able to concentrate?	1.000	.797
How safe do you feel in your daily life?	1.000	.764
How healthy is your	1.000	.699

Module 05, SPSS, Eigenvalues

Total Variance Explained

Component	Total	Initial Eigenvalues		Extraction Sums of Squared Loadings		
		% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	12.285	51.185	51.185	12.285	51.185	51.185
2	3.101	12.923	64.108	3.101	12.923	64.108
3	1.623	6.762	70.870	1.623	6.762	70.870
4	1.366	5.691	76.561	1.366	5.691	76.561
5	.918	3.824	80.385			
6	.701	2.919	83.304			
7	.627	2.612	85.915			
8	.592	2.465	88.380			
9	.505	2.103	90.483			
10	.410	1.710	92.193			

Module 05, SPSS, Scree plot

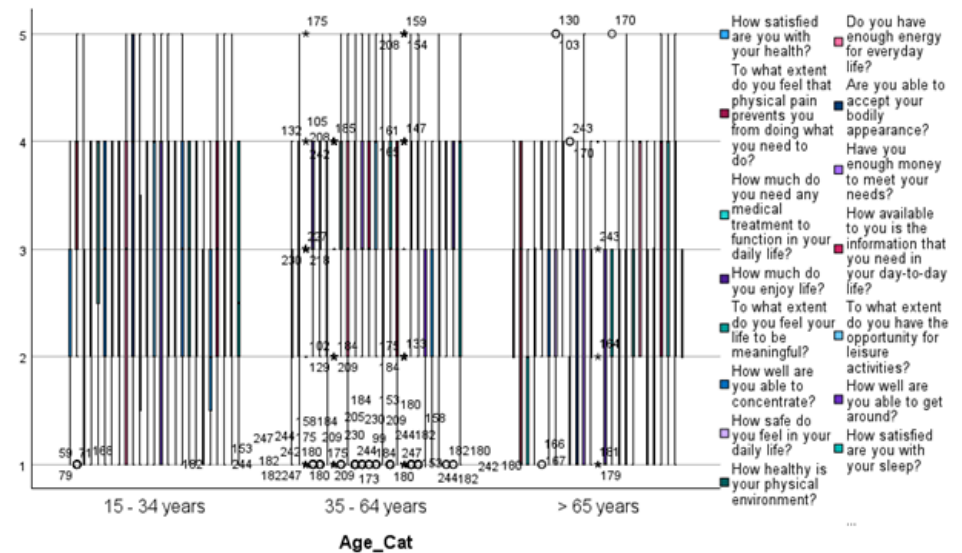
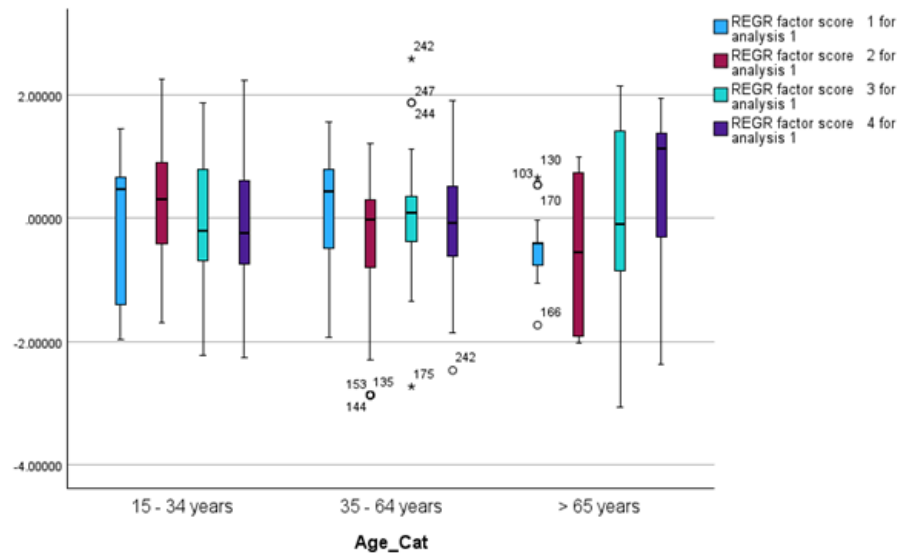


Module 05, SPSS, Component matrix

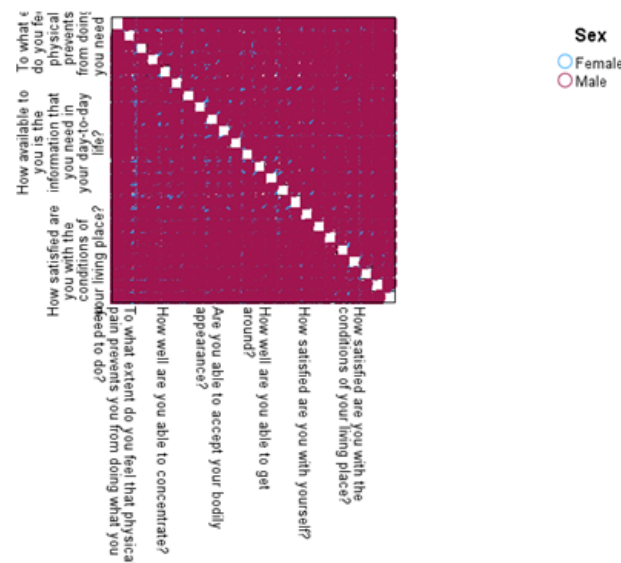
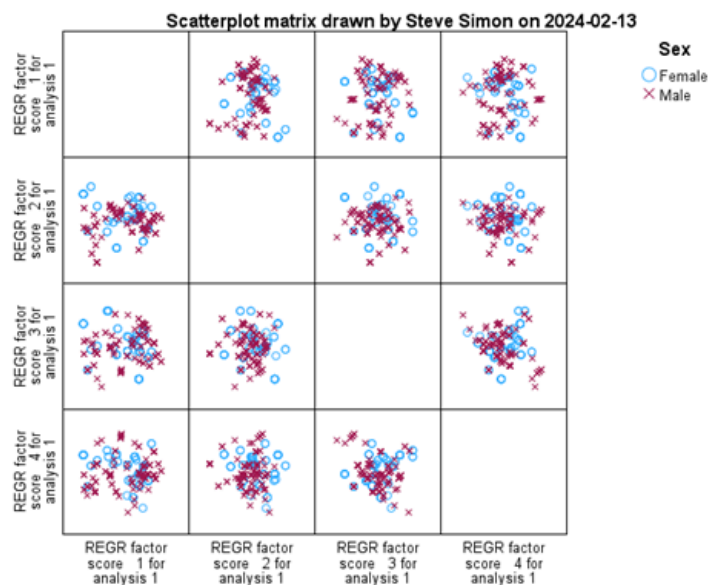
Component Matrix^a

	Component			
	1	2	3	4
To what extent do you feel that physical pain prevents you from doing what you need to do?	.282	.557	.199	.257
How much do you need any medical treatment to function in your daily life?	-.073	.642	-.130	.415
How much do you enjoy life?	.770	.257	.376	.116
To what extent do you feel your life to be meaningful?	.745	.206	.402	.120
How well are you able to concentrate?	.852	.155	.174	.132
How safe do you feel in	.743	.125	.395	.203

Module 05, SPSS, Boxplots of first four principal components



Module 05, SPSS, Scatterplot of first four principal components



Module 05, SPSS, R-squared using four principal components

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.785 ^a	.617	.610	.528

a. Predictors: (Constant), REGR factor score 4 for analysis 1, REGR factor score 3 for analysis 1, REGR factor score 2 for analysis 1, REGR factor score 1 for analysis 1

Module 05, SPSS, R-squared using all 24 variables

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.841 ^a	.707	.676	.482

a. Predictors: (Constant), How often do you have negative feelings such as blue mood, despair, anxiety, depression?, How much do you need any medical treatment to function in your daily life?, To what extent do you feel that physical pain prevents you from doing what you need to do?, How healthy is your physical environment?, Do you have enough energy for everyday life?, To what extent do you have the opportunity for leisure activities?, How satisfied are you with the conditions of your living place?, How satisfied are you with your personal relationships?, How satisfied are you with the support you get from your friends?, How safe do you feel in your daily life?, Have you enough money to meet your needs?, How satisfied are you with your capacity for work?, To what extent do you feel your life to be meaningful?, How satisfied are you with your sex life?, Are you able to accept your bodily appearance?, How satisfied are you with your transport?, How much do you enjoy life?, How satisfied are you with your ability to perform your daily living activities?, How available to you is the information that you need in your day-to-day life?, How satisfied are you with your sleep?, How well are you able to concentrate?, How satisfied are you with yourself?, How satisfied are you with your access to health services?, How well are you able to get around?

Module 05, SPSS, Rotated factor pattern, 1 of 3

Rotated Factor Matrix^a

	Factor			
	1	2	3	4
To what extent do you feel that physical pain prevents you from doing what you need to do?	.278	.351	-.066	-.185
How much do you need any medical treatment to function in your daily life?	.083	.160	-.133	-.510
How much do you enjoy life?	.464	.731	.108	.156
To what extent do you feel your life to be meaningful?	.391	.789	.131	.122
How well are you able to concentrate?	.522	.696	.307	.074
How safe do you feel in	.387	.693	.266	.131

Module 05, SPSS, Rotated factor pattern, 2 of 3

Rotated Factor Matrix^a

	Factor			
	1	2	3	4
How satisfied are you with your ability to perform your daily living activities?	.911	.276	.096	.047
How satisfied are you with your capacity for work?	.843	.274	.139	.077
How well are you able to get around?	.819	.396	.170	.159
How satisfied are you with your sleep?	.813	.293	.196	.058
How satisfied are you with yourself?	.810	.338	.230	.153
Are you able to accept your	.778	.340	.090	.287

Module 05, SPSS, Rotated factor pattern, 3 of 3

Rotated Factor Matrix^a

	Factor			
	1	2	3	4
How satisfied are you with your ability to perform your daily living activities?	.911			
How satisfied are you with your capacity for work?	.843			
How well are you able to get around?	.819			
How satisfied are you with your sleep?	.813			
How satisfied are you with yourself?	.810			
Are you able to accept your	.778			

Break #5

- What you have learned
 - 05 Dimension reduction
- What's coming next
 - 06 Logistic regression

Module 06, Precursors to logistic regression

- Test of two proportions
- Chi-square test of independence
- Odds ratio versus relative risk

Module 06, Logistic regression

- Linear on log odds scale
- Assumptions
 - Independence
 - Linearity

Module 06, SPSS, Confidence interval and test of hypothesis

Independent-Samples Proportions Group Statistics

	Sex	Successes	Trials	Proportion	Asymptotic Standard Error
Did the passenger survive? = Yes	= female	308	462	.667	.022
	= male	142	851	.167	.013

Independent-Samples Proportions Confidence Intervals

	Interval Type	Difference in Proportions	Asymptotic Standard Error	95% Confidence Interval of the Difference	
				Lower	Upper
Did the passenger survive? = Yes	Wald	.500	.025	.450	.550

Independent-Samples Proportions Tests

	Test Type	Difference in Proportions	Asymptotic Standard Error	Z	Significance	
					One-Sided p	Two-Sided p
Did the passenger survive? = Yes	Wald H0	.500	.025	18.222	<.001	<.001

Speaker notes

Here is the output from SPSS. The confidence interval contains only positive values, so you can conclude that the difference in proportions is statistically significant.

You can draw the same conclusion from the p-value, which is less than 0.001.

Module 06, SPSS, Example: Titanic survival by sex

Sex * Did the passenger survive?

Crosstabulation

Count

		Did the passenger survive?		
		No	Yes	Total
Sex	female	154	308	462
	male	709	142	851
Total		863	450	1313

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2- sided)	Exact Sig. (1- sided)
Pearson Chi-Square	332.057 ^a	1	<.001		
Continuity Correction ^b	329.842	1	<.001		
Likelihood Ratio	332.534	1	<.001		
Fisher's Exact Test				<.001	<.001
N of Valid Cases	1313				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 158.34.

b. Computed only for a 2x2 table

- Moderate or large sample size: Pearson Chi-Square
- Small sample size: Fisher's Exact test

Speaker notes

Here is the output from SPSS. Like most other parts of SPSS, the default is to include four different tests. The tests can differ from one another, but in this case, they all tell the same story. To be honest, this is usually the case.

I recommend using the Person Chi-Square if the sample size is moderate or large and Fisher's Exact Test if the sample size is small.

What makes it small is the expected count. If any expected count is less than 5, then you should rely on Fisher's Exact Test.

There is a lot of conflict in the research community about the use of a continuity correction.

Module 06, SPSS, An example of a log odds model with real data, 1 of 3

GA	Actual prob BF
28	$2/6 = 33.3\%$
29	$2/5 = 40.0\%$
30	$7/9 = 77.8\%$
31	$7/9 = 77.8\%$
32	$16/20 = 80.0\%$
33	$14/15 = 93.3\%$

Speaker notes

There are other approaches that also work well for this type of data, such as a probit model, that I won't discuss here. But I did want to show you what the data relating GA and BF really looks like.

Module 06, SPSS, An example of a log odds model with real data, 2 of 3

Speaker notes

I've simplified this data set by removing some of the extreme gestational ages.

The table below shows the predicted log odds, and the calculations needed to transform this estimate back into predicted probabilities.

Module 06, SPSS, An example of a log odds model with real data, 3 of 3

- $\log \text{ odds} = -16.72 + 0.577 \times 30 = 0.59$
- $\text{odds} = \exp(\log \text{ odds}) = 1.8$
- $\text{prob} = \text{odds} / (1 + \text{odds}) = 0.64$

Speaker notes

Let's examine these calculations for GA = 30. The predicted log odds would be the intercept plus the slope times 30.

Convert from log odds to odds by exponentiating.

And finally, convert from odds back into probability.

$$\text{prob} = 1.80 / (1 + 1.80) = 0.643$$

The predicted probability of 64.3% is reasonably close to the true probability (77.8%).

You might also want to take note of the predicted odds. Notice that the ratio of any odds to the odds in the next row is 1.78. For example,

$$3.20 / 1.80 = 1.78$$

$$5.70 / 3.20 = 1.78$$

It's not a coincidence that you get the same value when you exponentiate the slope term in the log odds equation.

$$\exp(0.59) = 1.78$$

This is a general property of the logistic model. The slope term in a logistic regression model represents the log of the odds ratio. This represents the increase (decrease) in risk as the independent variable increases by one unit.

Module 06, SPSS, Categorical variables in a logistic regression model, 1 of 3

Speaker notes

You treat categorical variables in much the same way as you would in a linear regression model. Create indicator variables for each level of your categorical variable and then include all but one of them in your model. The category associated with the omitted variable is the reference category.

How would SPSS handle a variable like Passenger Class, which has three levels

1st, 2nd, 3rd?

Here's a crosstabulation of survival versus passenger class.

Notice that the odds of dying are 0.67 to 1 in 1st class, 1.35 to 1 in 2nd class, and 4.15 to 1 in 3rd class. These are odds in favor of dying. The odds against dying are 1.50 to 1, 0.74 to 1, and 0.24 to 1, respectively.

Module 06, SPSS, Categorical variables in a logistic regression model, 2 of 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	pclass			159.120	2	.000	
	pclass(1)	1.827	.148	152.162	1	.000	6.212
	pclass(2)	1.121	.154	53.267	1	.000	3.069
	Constant	-1.424	.095	225.403	1	.000	.241

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

- $1.50 / 0.24 = 6.212$
- $0.74 / 0.24 = 3.069$

Speaker notes

The odds ratio for the pclass(1) row is 6.212, which is equal to $1.50 / 0.24$. You should interpret this as the odds against dying are 6 times better in first class compared to third class. The odds ratio for the pclass(2) row is 3.069, which equals $0.74 / 0.24$. This tells you that the odds against dying are about 3 times better in second class compared to third class. The Constant row tells you that the odds are 0.241 to 1 in third class.

Module 06, SPSS, Categorical variables in a logistic regression model, 3 of 3

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1	pclass			159.120	2	.000	
	pclass(1)	-.705	.166	18.050	1	.000	.494
	pclass(2)	-1.827	.148	152.162	1	.000	.161
	Constant	.403	.114	12.550	1	.000	1.496

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

- $0.74 / 1.50 = 0.494$
- $0.24 / 1.50 = 0.161$

Speaker notes

If you prefer to do the analysis with each of the other classes being compared back to first class, then select FIRST for reference category.

This produces the following output:

Here the pclass(1) row provides an odds ratio of 0.494 which equals $0.74 / 1.50$. The odds against dying are about half in second class versus first class. The pclass(2) provides an odds ratio of 0.161 (approximately $1/6$) which equals $0.24 / 1.50$. The odds against dying are $1/6$ in third class compared to first class. The Constant row provides an odds of 1.496 to 1 against dying for first class.

Notice that the numbers in parentheses (pclass(1) and pclass(2)) do not necessarily correspond to first and second classes. It depends on how SPSS chooses the indicator variables. How did I know how to interpret the indicator variables and the odds ratios? I wouldn't have known how to do this if I hadn't computed a crosstabulation earlier. It is very important to do a few simple crosstabulations before you run a logistic regression model, because it helps you orient yourself to the data.

Module 06, SPSS, Odds ratios for first class

Sex * Survived Crosstabulation^a

Count

		Survived		Total
		Died	Survived	
Sex	female	9	134	143
	male	120	59	179
Total		129	193	322

a. PClass = 1st

Risk Estimate^a

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex_reversed (male / female)	30.282	14.399	63.685
For cohort Survived = Died	10.652	5.613	20.215
For cohort Survived = Survived	.352	.284	.435
N of Valid Cases	322		

a. PClass = 1st

Module 06, SPSS, Odds ratio for second class

Sex * Survived Crosstabulation^a

Count

		Survived		Total
		Died	Survived	
Sex	female	13	94	107
	male	148	25	173
Total		161	119	280

a. PClass = 2nd

Risk Estimate^a

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex_reversed (male / female)	42.806	20.871	87.794
For cohort Survived = Died	7.041	4.215	11.763
For cohort Survived = Survived	.164	.114	.238
N of Valid Cases	280		

a. PClass = 2nd

Module 06, SPSS, Odds ratio for third class

Sex * Survived Crosstabulation^a

Count

		Survived		Total
		Died	Survived	
Sex	female	132	80	212
	male	441	58	499
Total		573	138	711

a. PClass = 3rd

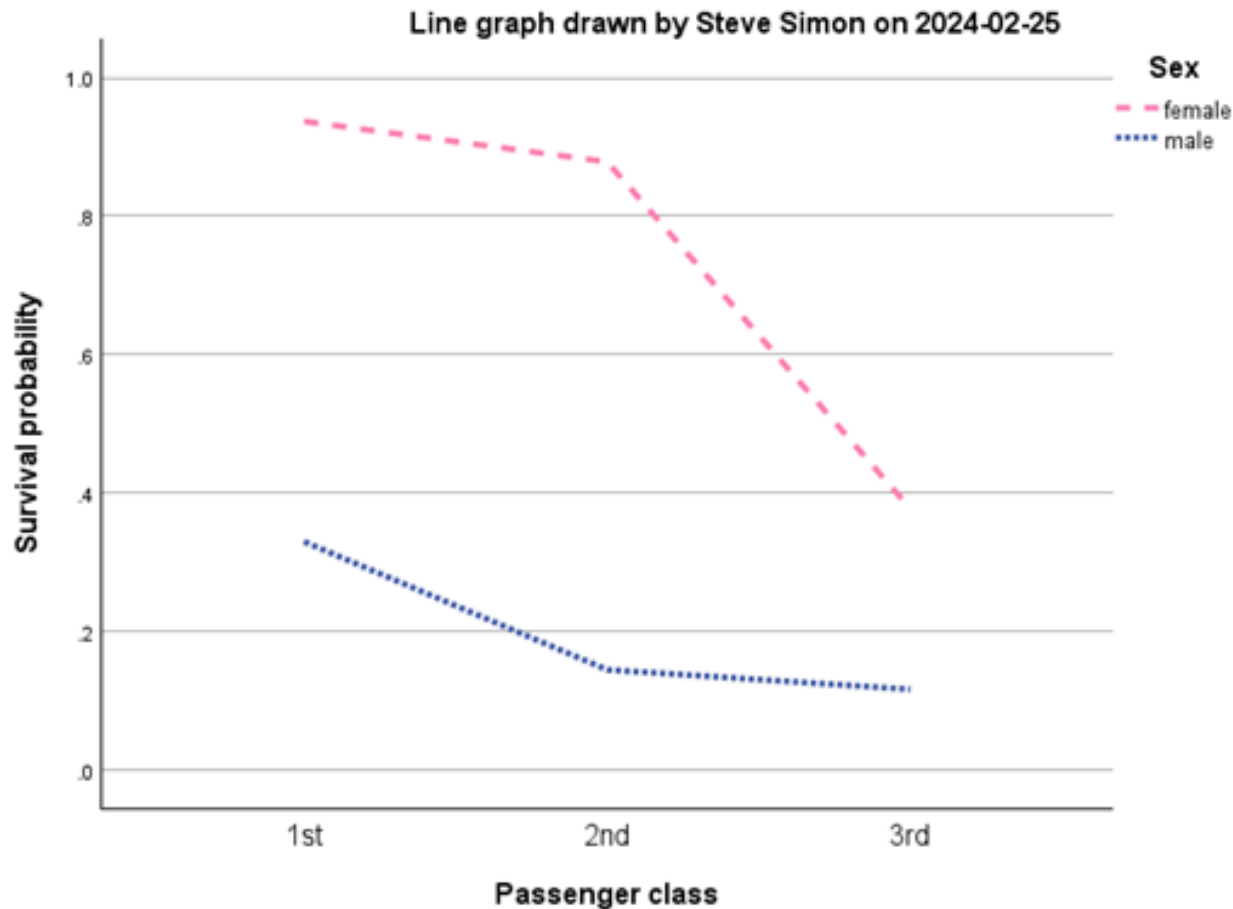
Risk Estimate^a

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for sex_reversed (male / female)	4.608	3.120	6.806
For cohort Survived = Died	1.419	1.272	1.584
For cohort Survived = Survived	.308	.229	.415
N of Valid Cases	711		

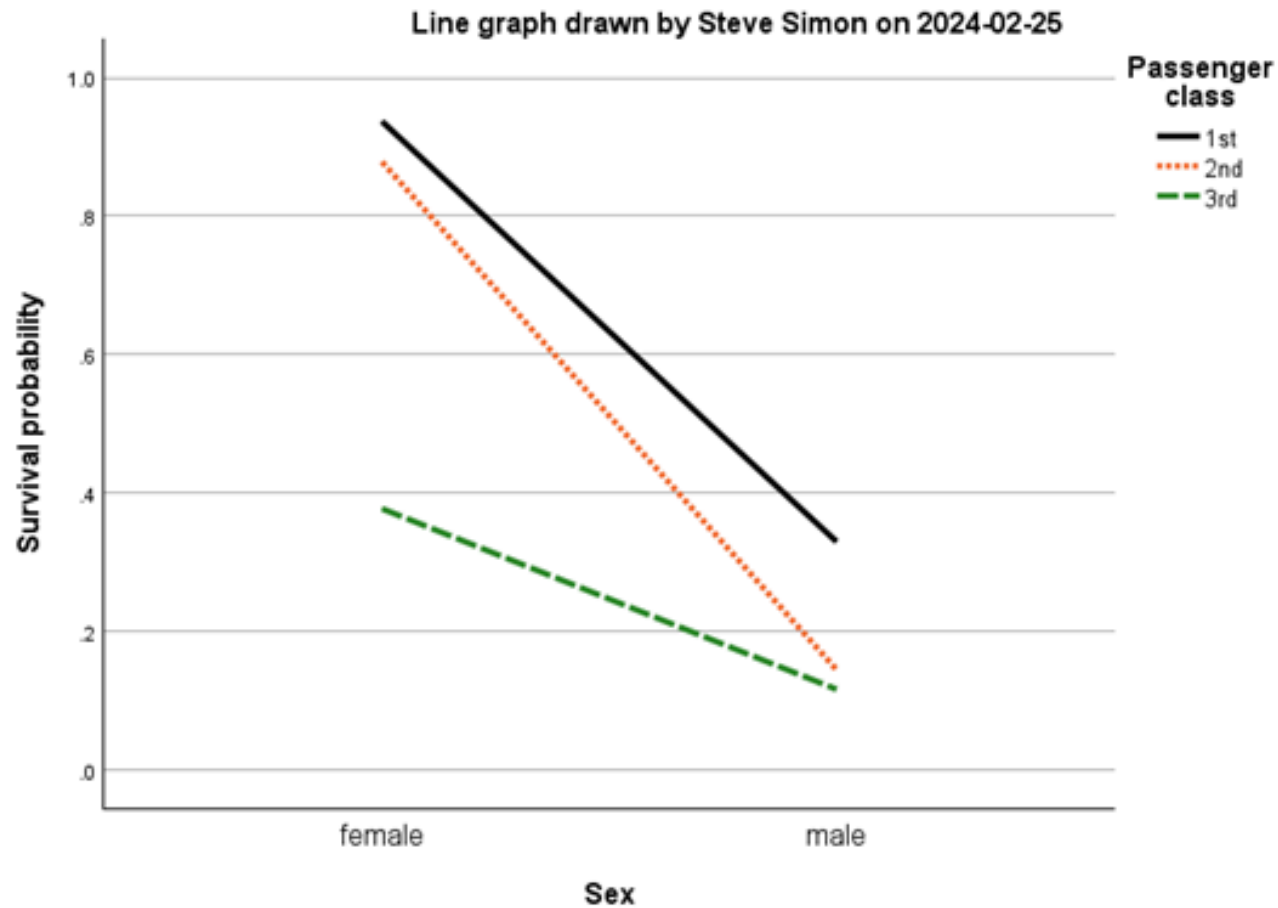
a. PClass = 3rd

Module 06, SPSS, Logistic regression with interaction

Module 06, SPSS, Line plot for interaction, 1 of 2



Module 06, SPSS, Line plot for interaction, 2 of 2



Module 06, SPSS, Creating a binary outcome

Bf six months after discharge * breast_feeding_at_six_months Crosstabulation

Count

		breast_feeding_at_six_months		Total
		.00	1.00	
Bf six months after discharge	Exclus.	0	24	24
	Partial	9	0	9
	None	50	0	50
Total		59	24	83

Module 06, SPSS, Crosstabulation of predictor and outcome

Binary coding -- control=0, treatment=1 * breast_feeding_at_six_months Crosstabulation

			breast_feeding_at_six_months		Total
			0	1	
Binary coding -- control=0, treatment=1	0=Control	Count	33	12	45
		% within Binary coding -- control=0, treatment=1	73.3%	26.7%	100.0%
	1=Treatment	Count	17	21	38
		% within Binary coding -- control=0, treatment=1	44.7%	55.3%	100.0%
Total	Count		50	33	83
	% within Binary coding -- control=0, treatment=1		60.2%	39.8%	100.0%

Module 06, SPSS, Unadjusted odds ratio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	Binary coding -- control=0, treatment=1	1.223	.469	6.795	1	.009	3.397
	Constant	-1.012	.337	9.005	1	.003	.364

a. Variable(s) entered on step 1: Binary coding -- control=0, treatment=1.

Module 06, SPSS, Adjusted odds ratio

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1 ^a	Binary coding -- control=0, treatment=1	1.190	.492	5.836	1	.016	3.286
	Mother's age	.008	.037	.047	1	.828	1.008
	Constant	-1.214	.996	1.488	1	.223	.297

a. Variable(s) entered on step 1: Binary coding -- control=0, treatment=1, Mother's age.

Break #6

- What you have learned
 - 06 Logistic regression
- What's coming next
 - 07 Diagnostic tests

Module 07, Diagnostic tests

- Sensitivity, specificity
 - SpPin, SnNout
- Positive/negative predictive value
- Likelihood ratio
- ROC curve

Break #7

- What you have learned
 - 07 Diagnostic tests
- What's coming next
 - 08 Survival analysis

Module 08, Basic survival analysis

- Censoring
- Kaplan-Meier curve
- Log rank test

Module 08, Cox regression

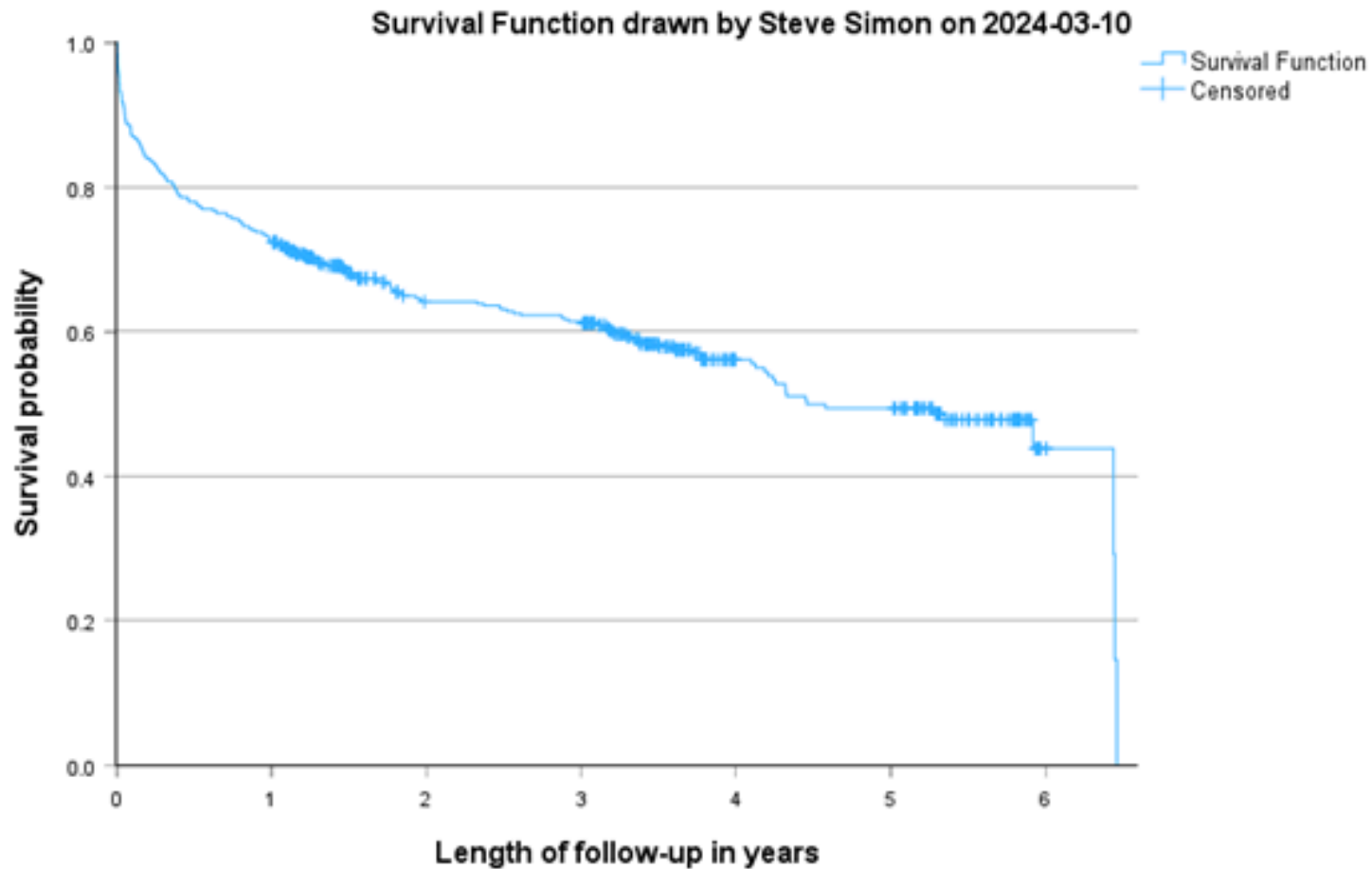
- Define hazard function
 - Increasing/decreasing/constant hazard
 - Hazard ratio
- Assumptions
 - Independence
 - Non-informative censoring

Module 08, SPSS, Event count

fstat

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	285	57.0	57.0	57.0
	1	215	43.0	43.0	100.0
	Total	500	100.0	100.0	

Module 08, SPSS, Overall Kaplan-Meier curve



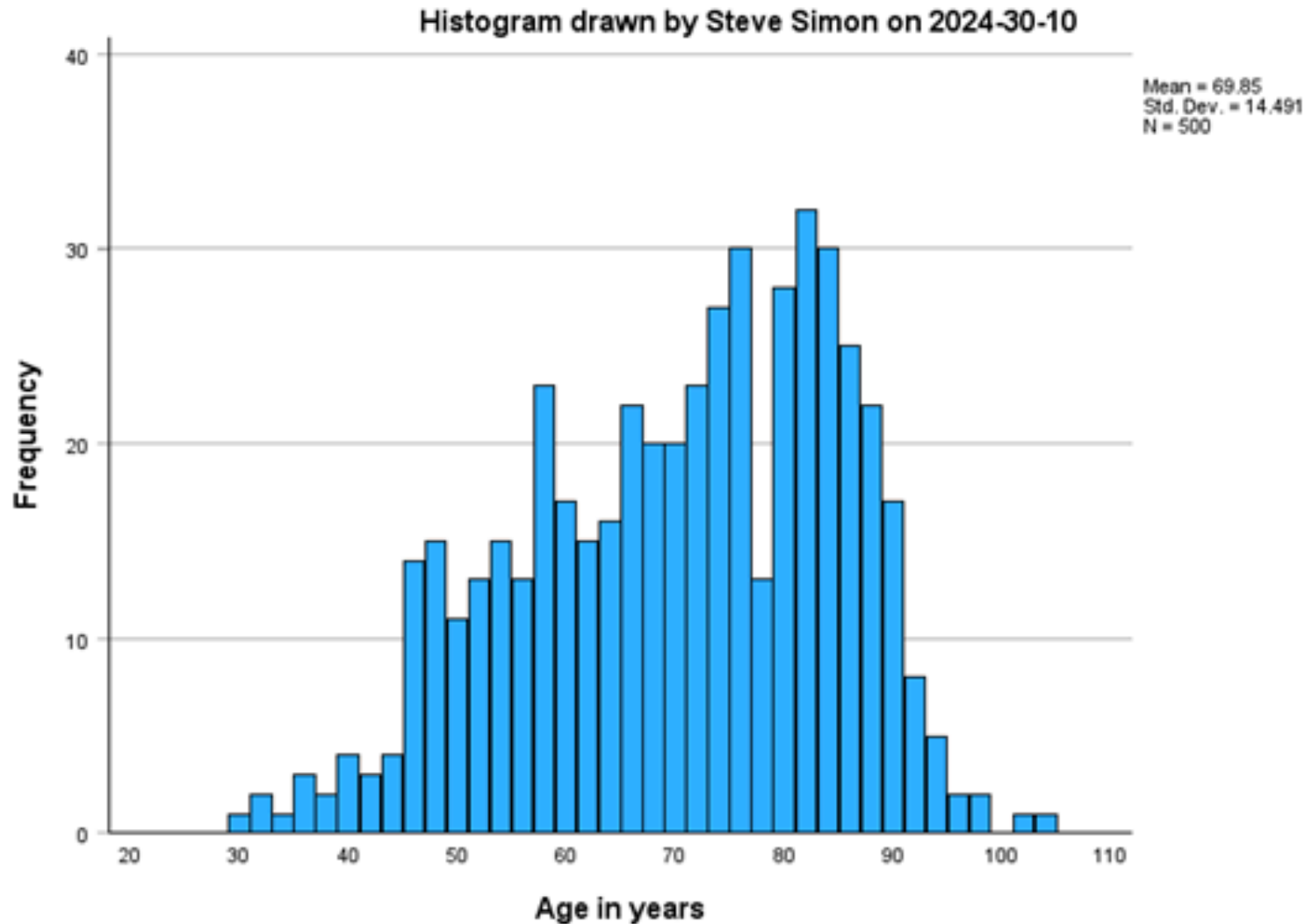
Module 08, SPSS, Event count by gender

gender * fstat Crosstabulation

Count

		fstat		
		0	1	Total
gender	0	189	111	300
	1	96	104	200
Total		285	215	500

Module 08, SPSS, Histogram of ages



Module 08, SPSS, Quality check of age group coding

age * age_group Crosstabulation

Count

		age_group				Total
		1	2	3	4	
age	30	1	0	0	0	1
	32	2	0	0	0	2
	33	1	0	0	0	1
	35	1	0	0	0	1
	36	2	0	0	0	2
	37	2	0	0	0	2
	39	2	0	0	0	2
	40	2	0	0	0	2
	41	2	0	0	0	2
	42	1	0	0	0	1
	43	2	0	0	0	2
	44	2	0	0	0	2
	45	4	0	0	0	4
	46	10	0	0	0	10
	47	4	0	0	0	4
	48	11	0	0	0	11
	49	4	0	0	0	4
	50	7	0	0	0	7
	51	0	4	0	0	4
	52	0	9	0	0	9

Module 08, SPSS, Event count by age group, 1 of 2

age_group * fstat Crosstabulation

Count

		fstat		
		0	1	Total
age_group	1	55	5	60
	2	100	25	125
	3	92	78	170
	4	38	107	145
Total		285	215	500

Module 08, SPSS, Event count by age group, 2 of 2

age_group * fstat Crosstabulation

Count

		fstat		
		0	1	Total
age_group	65 and under	155	30	185
	66 to 80	92	78	170
	81 and above	38	107	145
Total		285	215	500

Module 08, SPSS, Kaplan-Meier analysis by gender, 1 of 3

Case Processing Summary

gender	Total N	N of Events	Censored	
			N	Percent
0	300	111	189	63.0%
1	200	104	96	48.0%
Overall	500	215	285	57.0%

Means and Medians for Survival Time

gender	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
0	1448.506	55.852	1339.035	1557.976	2160.000	.	.	.
1	1260.208	75.267	1112.684	1407.732	1317.000	177.039	970.004	1663.996
Overall	1417.215	48.137	1322.867	1511.562	1627.000	159.555	1314.271	1939.729

a. Estimation is limited to the largest survival time if it is censored.

Module 08, SPSS, Kaplan-Meier analysis by gender, 2 of 3

Percentiles

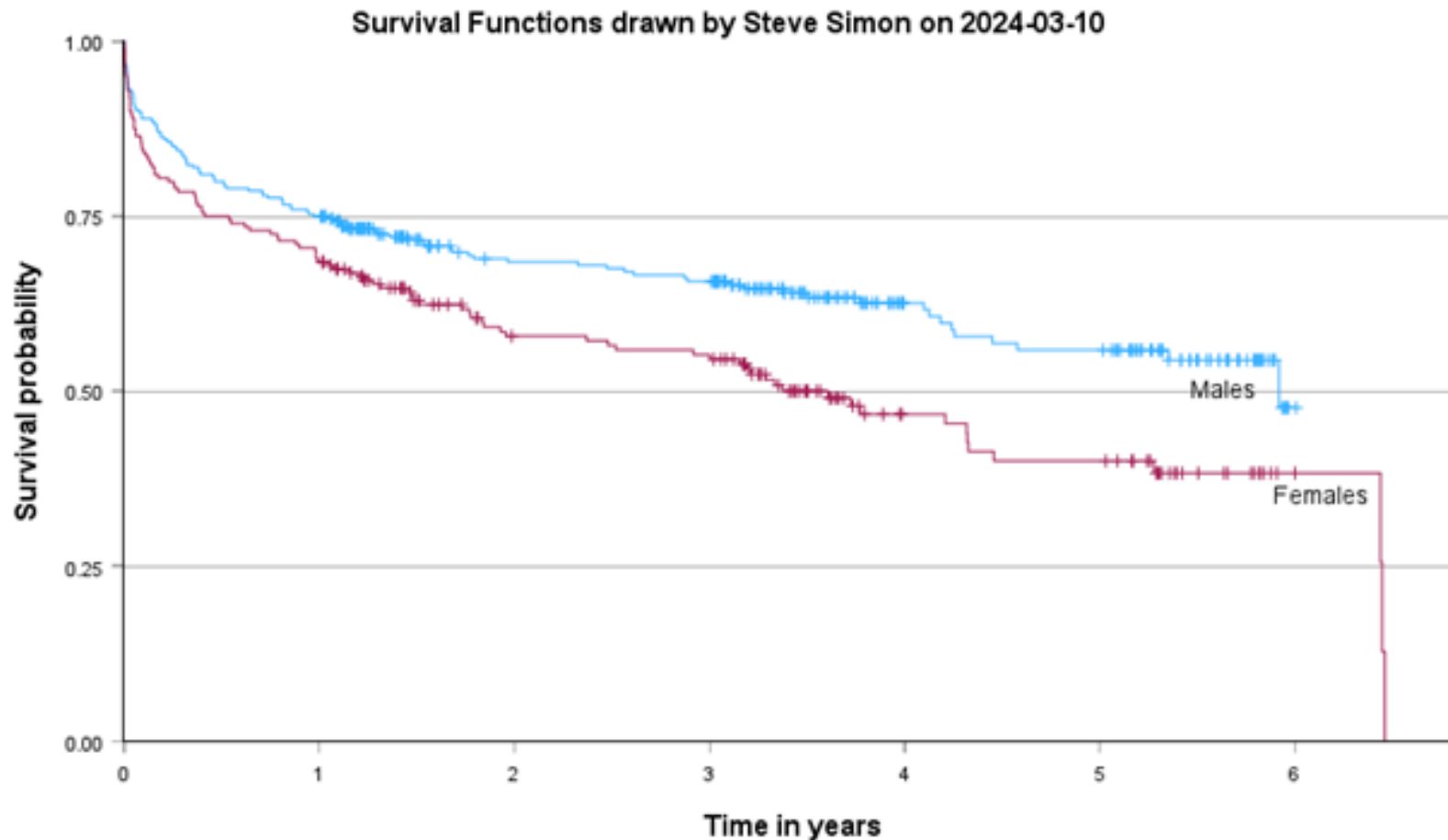
gender	25.0%		50.0%		75.0%	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
0			2160.000	.	354.000	109.695
1	2353.000	176.168	1317.000	177.039	151.000	84.201
Overall	2353.000	79.465	1627.000	159.555	295.000	63.223

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	7.791	1	.005

Test of equality of survival distributions for the different levels of gender.

Module 08, SPSS, Kaplan-Meier analysis by gender, 3 of 3



Module 08, SPSS, Kaplan-Meier analysis by age group, 1 of 3

Case Processing Summary

age_group	Total N	N of Events	Censored	
			N	Percent
65 and under	185	30	155	83.8%
66 to 80	170	78	92	54.1%
81 and above	145	107	38	26.2%
Overall	500	215	285	57.0%

Means and Medians for Survival Time

age_group	Mean ^a				Median			
	Estimate	Std. Error	95% Confidence Interval		Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound			Lower Bound	Upper Bound
65 and under	2005.895	61.543	1885.270	2126.520	2353.000	550.241	1274.527	3431.473
66 to 80	1398.427	79.795	1242.028	1554.825	1624.000	259.869	1114.658	2133.342
81 and above	727.181	71.645	586.757	867.605	343.000	88.116	170.292	515.708
Overall	1417.215	48.137	1322.867	1511.562	1627.000	159.555	1314.271	1939.729

a. Estimation is limited to the largest survival time if it is censored.

Module 08, SPSS, Kaplan-Meier analysis by age group, 2 of 3

Percentiles

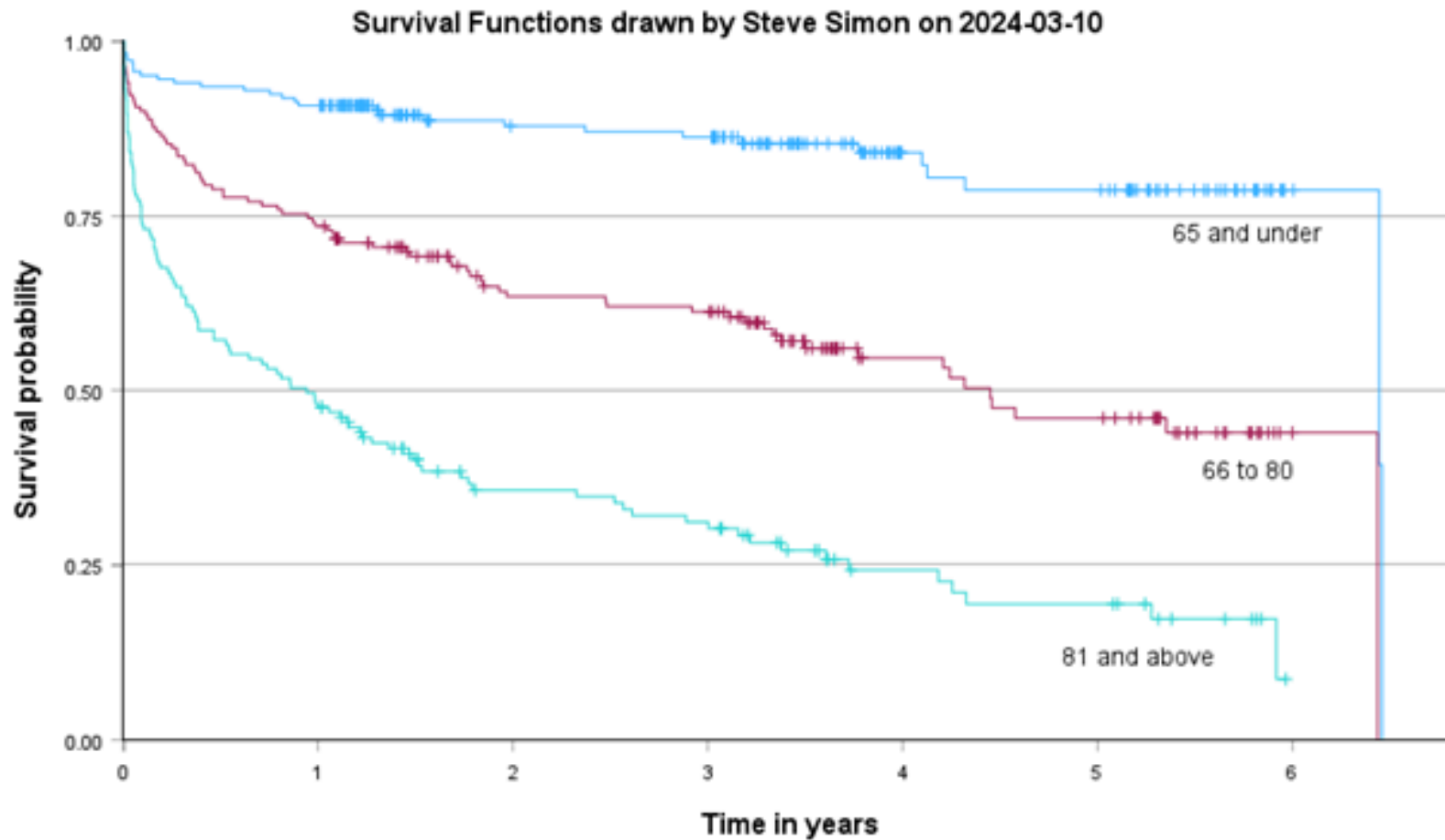
age_group	25.0%		50.0%		75.0%	
	Estimate	Std. Error	Estimate	Std. Error	Estimate	Std. Error
65 and under	2358.000	.	2353.000	550.241	2353.000	574.473
66 to 80	2350.000	.	1624.000	259.869	345.000	126.261
81 and above	1359.000	181.588	343.000	88.116	33.000	12.970
Overall	2353.000	79.465	1627.000	159.555	295.000	63.223

Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	129.254	2	<.001

Test of equality of survival distributions for the different levels of age_group.

Module 08, SPSS, Kaplan-Meier analysis by age group, 3 of 3



Module 08, SPSS, Mean ages for men and women

Report

age			
gender	Mean	N	Std. Deviation
Male	66.60	300	14.943
Female	74.72	200	12.301
Total	69.85	500	14.491

Module 08, SPSS, Unadjusted and adjusted Cox regression models for gender

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
gender	.381	.138	7.679	1	.006	1.464	1.118	1.917

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
gender	-.066	.141	.217	1	.641	.937	.711	1.234
age	.067	.006	116.401	1	<.001	1.069	1.056	1.082

Break #8

- What you have learned
 - 08 Survival analysis
- What's coming next
 - 09 Meta-analysis

Module 09, Meta-analysis

- Forest plot
- Publication bias
 - Funnel plot
- Heterogeneity
 - Cochran's Q, I-squared

Module 09, SPSS, Vaccine results, 1 of 6

Effect Size Estimates for Subgroup Analysis

	Effect Size	Std. Error	Z	Sig. (2-tailed)	95% Confidence Interval	
					Lower	Upper
booster	.631	.1046	6.032	<.001	.426	.836
primary	.095	.0421	2.245	.025	.012	.177
Overall	.341	.1146	2.973	.003	.116	.566

Module 09, SPSS, Vaccine results, 2 of 6

Effect Size Estimates for Individual Studies

	ID	Effect Size	Std. Error	Z	Sig. (2-tailed)	95% Confidence Interval		Weight	Weight (%)
						Lower	Upper		
booster	Achrekar	.434	.0777	5.586	<.001	.282	.587	11.108	14.6
	Batra	.814	.0942	8.638	<.001	.629	.998	10.769	14.2
	Yadete	.657	.0456	14.389	<.001	.567	.746	11.619	15.3
primary	Blasio	.100	.1254	.797	.425	-.146	.346	10.029	13.2
	Correa-Rodriguez	.133	.1121	1.190	.234	-.086	.353	10.358	13.6
	Gendler	.148	.0962	1.542	.123	-.040	.337	10.726	14.1
	Gutierrez	.065	.0566	1.148	.251	-.046	.176	11.469	15.1

Module 09, SPSS, Vaccine results, 3 of 6

Test of Homogeneity

	Chi-square (Q statistic)	df	Sig.
booster	10.485	2	.005
primary	.705	3	.872
Overall	104.335	6	<.001

Test of Subgroup Homogeneity

	Chi-square (Q statistic)	df	Sig.
vaccine_type	22.630	1	<.001

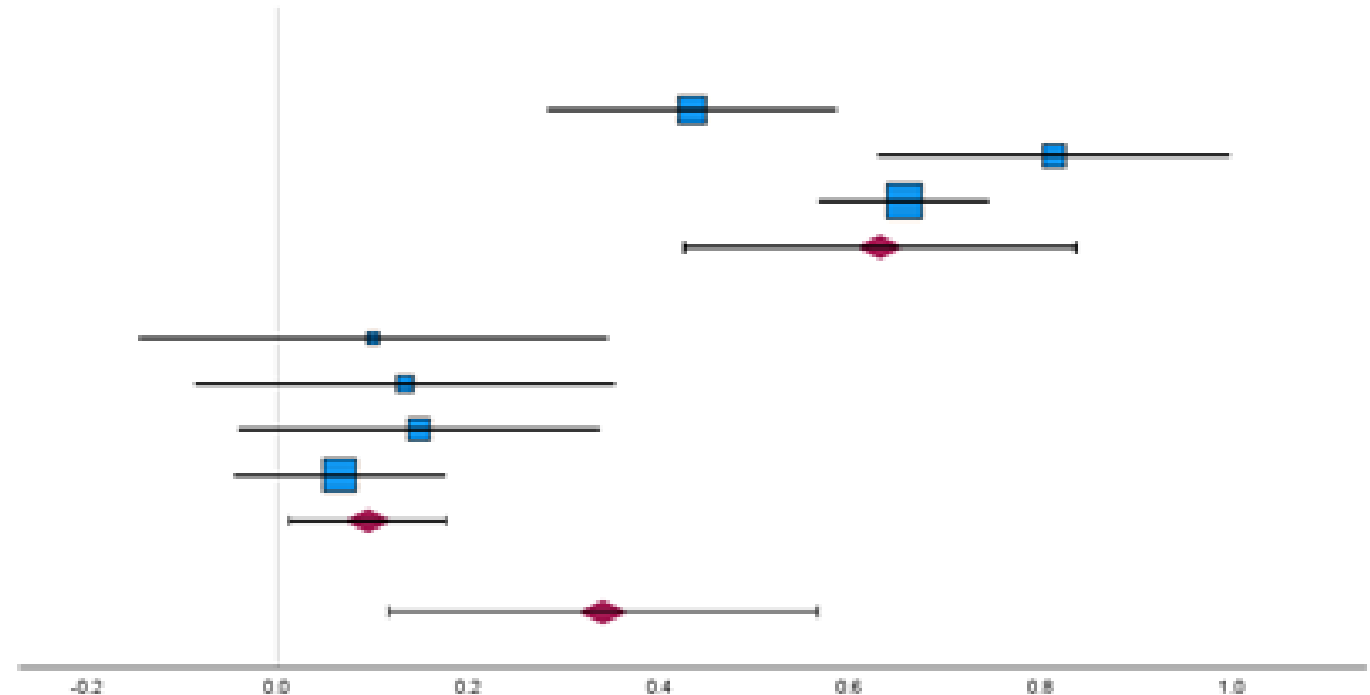
Module 09, SPSS, Vaccine results, 4 of 6

Module 09, SPSS, Vaccine results, 5 of 6

Forest Plot drawn by Steve Simon on 2024-03-17

■ Effect size of each study
◆ Estimated overall effect size
| Confidence interval of effect size
— No-effect value
I Estimated overall confidence interval

vaccine_type	ID	Cohen's d	Weight	Weight (%)
booster	Achredkar	0.43	11.11	14.60
	Batra	0.81	10.77	14.16
	Yadava	0.48	11.60	15.27
	Subgroup Overall	0.63		
primary	Biasio	0.10	10.03	13.18
	Correa-Rodriguez	0.13	10.36	13.61
	Gendler	0.15	10.73	14.10
	Gutierrez	0.07	11.47	15.08
	Subgroup Overall	0.09		
Overall		0.34		



Model: Random-effects model

Test of between-subgroup homogeneity: $Q = 22.63$, $df = 1$, $p\text{-value} = 0.00$

Module 09, SPSS, Vaccine results, 6 of 6

Break #9

- What you have learned
 - 09 Meta-analysis
- What's coming next
 - 10 Dark side of data science

Module 10, Dark side of data science

- Empiricism
- Reification
- Bias in data science

Break #10

- What you have learned
 - 10 Dark side of data science
- What's coming next
 - 11 Hierarchical models

Module 11, Hierarchical models

- Clustered data
- Between and within cluster variation
 - Intraclass correlation

Module 11, Checking assumptions

- Independence
 - Only between clusters
- Normality
 - Within clusters
 - Between clusters

Module 11, SPSS, Variance components, 1 of 2

ANOVA

Source	Type I Sum of Squares	df	Mean Square
Corrected Model	415.044	30	13.835
Intercept	25855.880	1	25855.880
SEX	23.426	1	23.426
GRP	66.368	2	33.184
LITTER	325.250	27	12.046
Error	160.520	149	1.077
Total	26431.444	180	
Corrected Total	575.564	179	

Dependent Variable: WGTP21

Module 11, SPSS, Variance components, 2 of 2

Expected Mean Squares

Source	Variance Component		Quadratic Term
	Var(LITTER)	Var(Error)	
Intercept	6.000	1.000	Intercept, SEX, GRP
SEX	.000	1.000	SEX
GRP	6.000	1.000	GRP
LITTER	6.000	1.000	
Error	.000	1.000	

Dependent Variable: WGTP21

Expected Mean Squares are based on Type I Sums of Squares.

For each source, the expected mean square equals the sum of the coefficients in the cells times the variance components, plus a quadratic term involving effects in the Quadratic Term cell.

Variance Estimates

Component	Estimate
Var(LITTER)	1.828
Var(Error)	1.077

Dependent Variable:

WGTP21

Method: ANOVA (Type I Sum of Squares)

Speaker notes

The F-ratios are $23.426/1.077 = \text{round}(23.426/1.077, 1)$ and $33.184/12.046 = 2.8$.

The intraclass correlation is $1.828/(1.828+1.077) = 0.63$.

Break #11

- What you have learned
 - 11 Hierarchical models
- What's coming next
 - 12 Longitudinal data






Module 12, Longitudinal data

- Random intercepts model
- Random slopes model

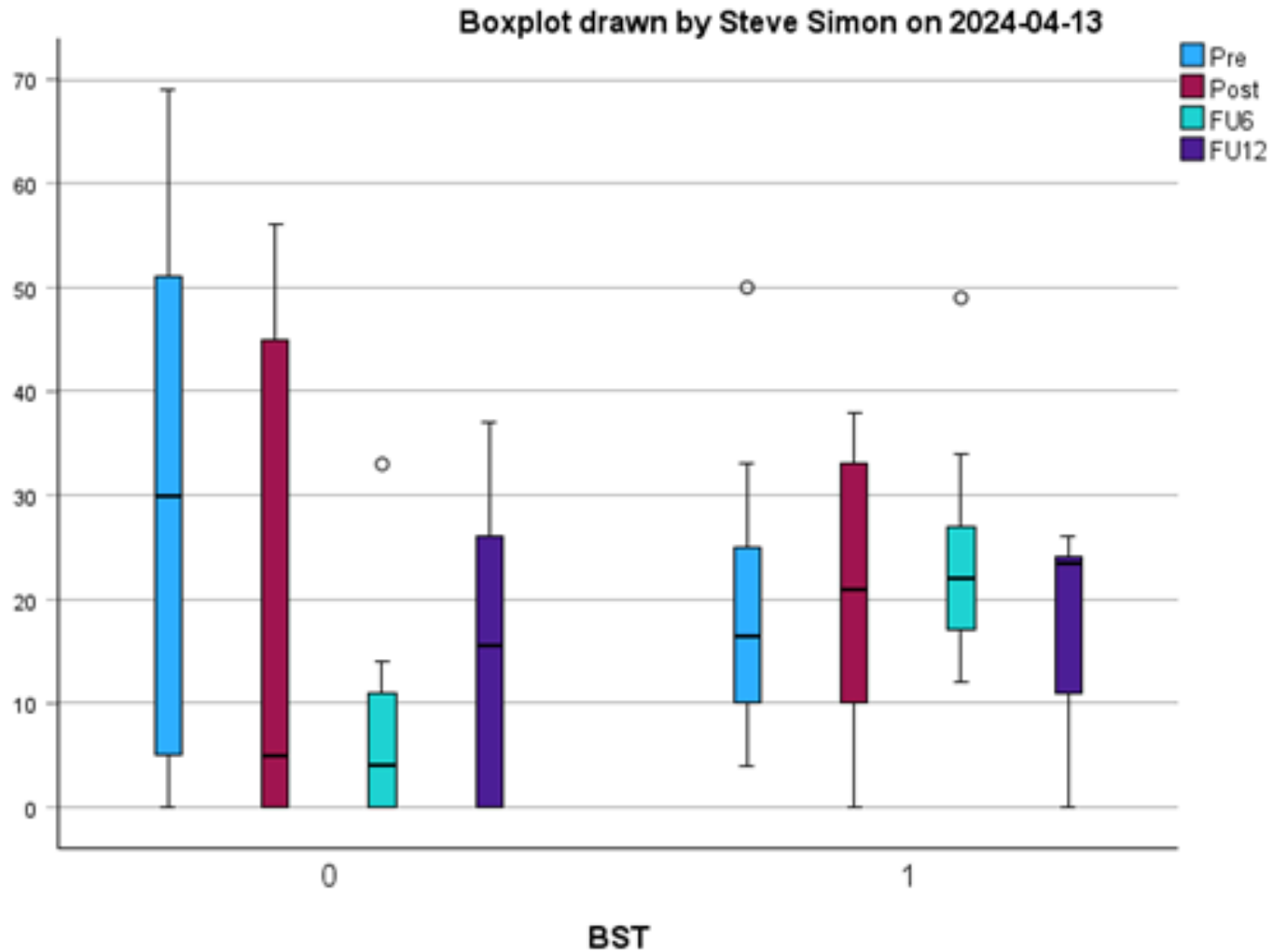
Module 12, Checking assumptions

- Independence
 - Only between subjects
- Normality
 - Residuals
 - Random intercepts/slopes
- Linearity
 - Scatterplot of residuals

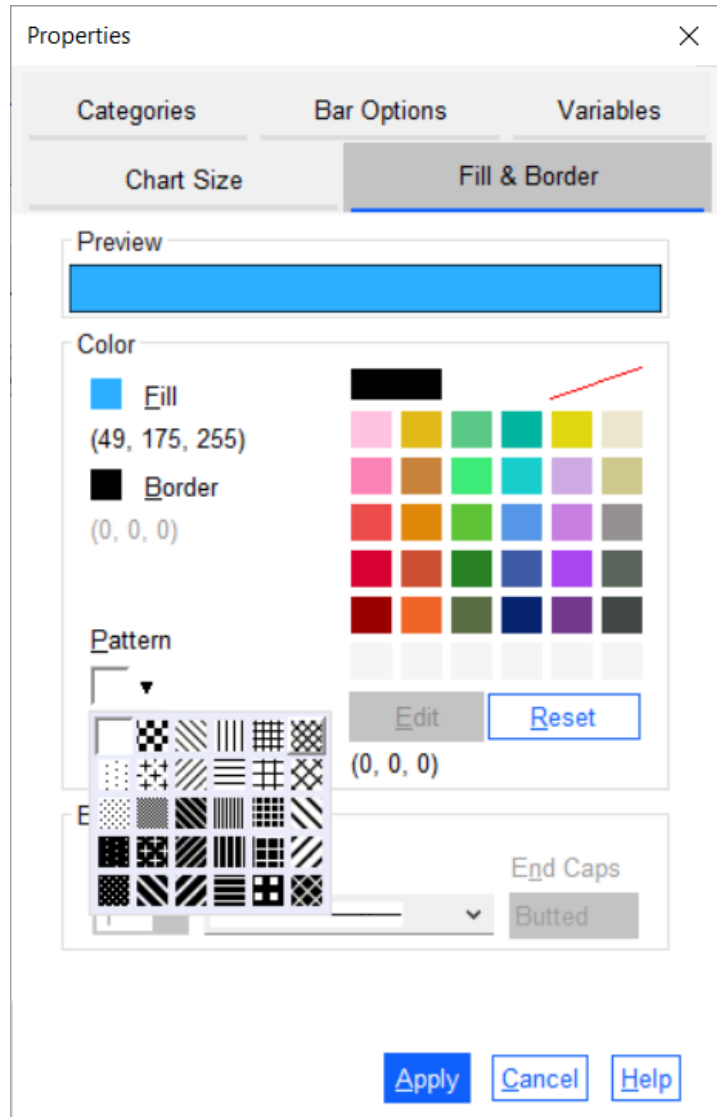
Module 12, SPSS, Wide format

	 BST	 Pre	 Post	 FU6	 FU12	
1	1	7	22	13	14	
2	1	25	10	17	24	
3	1	50	36	49	23	
4	1	16	38	34	24	
5	1	33	25	24	25	
6	1	10	7	23	26	
7	1	13	33	27	24	
8	1	22	20	21	11	
9	1	4	0	12	0	
10	1	17	16	20	10	
11	0	0	0	0	0	
12	0	69	56	14	36	
13	0	5	0	0	5	







Module 12, SPSS, Boxplots



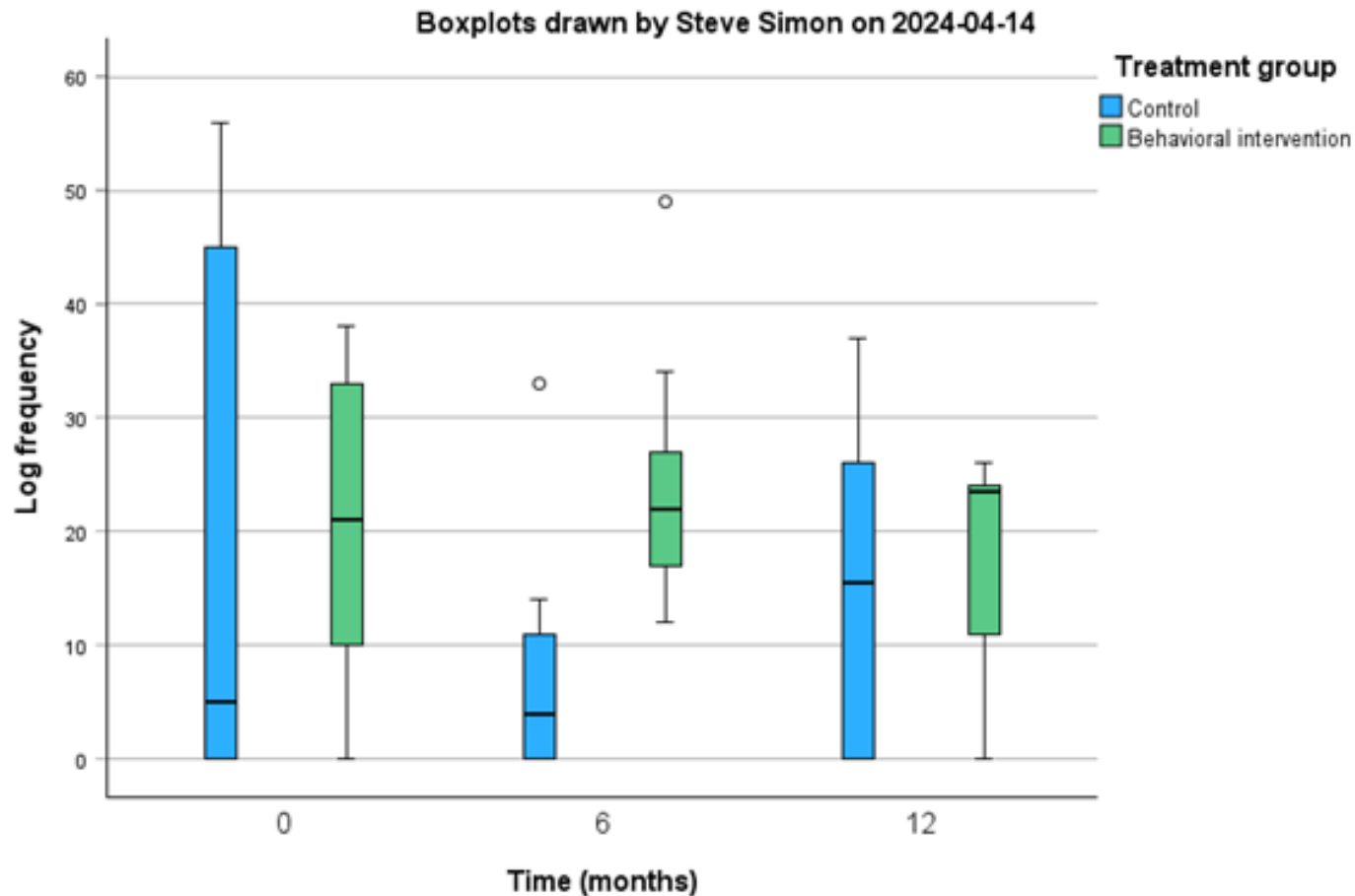
Module 12, SPSS, Colors and patterns



Module 12, SPSS, Tall format

	 id	 BST	 Pre	 Index1	 log_frequency	 time
1	1	1	7	Post	22	0
2	1	1	7	FU6	13	6
3	1	1	7	FU12	14	12
4	2	1	25	Post	10	0
5	2	1	25	FU6	17	6
6	2	1	25	FU12	24	12
7	3	1	50	Post	36	0
8	3	1	50	FU6	49	6
9	3	1	50	FU12	23	12
10	4	1	16	Post	38	0
11	4	1	16	FU6	34	6
12	4	1	16	FU12	24	12
13	5	1	22	Post	25	0

Module 12, SPSS, Alternate clustering of boxplots



Module 12, SPSS, Random intercepts analysis, 1 of 6

Model Dimension^a

		Number of Levels	Covariance Structure	Number of Parameters	Subject Variables
Fixed Effects	Intercept	1		1	
	BST	2		1	
	time	1		1	
Random Effects	Intercept	1	Variance Components	1	id
Residual				1	
Total		5		5	

a. Dependent Variable: log_frequency.

Module 12, SPSS, Random intercepts analysis, 2 of 6

Information Criteria^a

-2 Restricted Log Likelihood	473.80460947
Akaike's Information Criterion (AIC)	477.80460947
Hurvich and Tsai's Criterion (AICC)	478.02683170
Bozdogan's Criterion (CAIC)	483.89071201
Schwarz's Bayesian Criterion (BIC)	481.89071201

The information criteria are displayed in smaller-is-better form.

a. Dependent Variable: log_frequency.

Module 12, SPSS, Random intercepts analysis, 3 of 6

Coefficients of Determination

Pseudo-R Square Measures	Marginal	.064
	Conditional	.395

Module 12, SPSS, Random intercepts analysis, 4 of 6

Type III Tests of Fixed Effects^a

Source	Numerator df	Denominator df	F	Sig.
Intercept	1	38.556	37.935	<.001
BST	1	18.002	2.110	.164
time	1	39.000	.723	.400

a. Dependent Variable: log_frequency.

Module 12, SPSS, Random intercepts analysis, 5 of 6

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Intercept	22.508	3.935	28.633	5.720	<.001	14.455	30.561
[BST=0]	-7.133	4.910	18.002	-1.453	.164	-17.450	3.183
[BST=1]	0 ^b	0
time	-.263	.309	39.000	-.850	.400	-.887	.362

a. Dependent Variable: log_frequency.

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

Module 12, SPSS, Random intercepts analysis, 6 of 6

Estimates of Covariance Parameters^a

Parameter		Estimate	Std. Error
Residual		137.199	31.070
Intercept [subject = id]	Variance	74.827	41.497

a. Dependent Variable: log_frequency.

Break #12

- What you have learned
 - 12 Longitudinal data
- What's coming next
 - 13 Bayesian statistics

Module 13, Bayesian statistics

- Prior
 - Flat or non-informative prior
- Likelihood
- Posterior

Summary, 1 of 2

- What you have learned
 - 01 Linear regression, analysis of variance
 - 02 Linear regression with multiple independent variables
 - 03 Analysis of covariance
 - 04 Multi-factor analysis of variance
 - 05 Dimension reduction
 - 06 Logistic regression
 - 07 Diagnostic tests

Summary, 2 of 2

- What you have learned
 - 08 Survival analysis
 - 09 Meta-analysis
 - 10 Dark side of data science
 - 11 Hierarchical models
 - 12 Longitudinal data
 - 13 Bayesian statistics

