Regression analysis (PC-exercise)

Sebastian Jentschke

Agenda

- Regression analysis
  - ***Checking requirements***
  - Correlation to Linear regression (one predictor)
  - Linear regression (multiple predictors)
  - Methods for adding predictors
  - Assessing the quality of your model
  - ***Assignment***
- Logistic regression (binary)
  - Introduction
  - ***Assignment***

Exam Anxiety.sav

IV: Exam; DV: Revise, Anxiety

Check for normality, linearity, and multi-collinearity

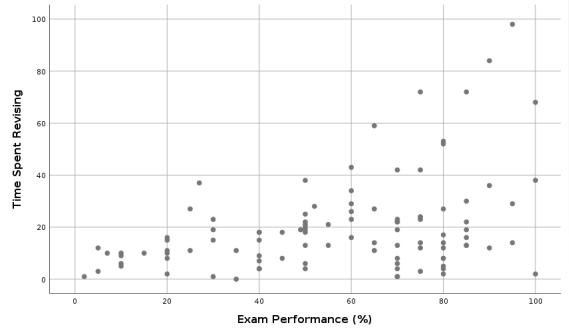
Remove outiers (10 – 15 mins)

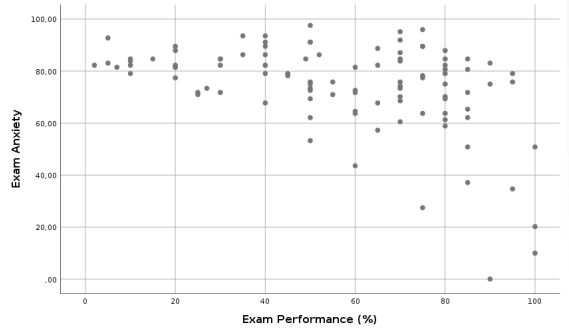
From correlation to regression

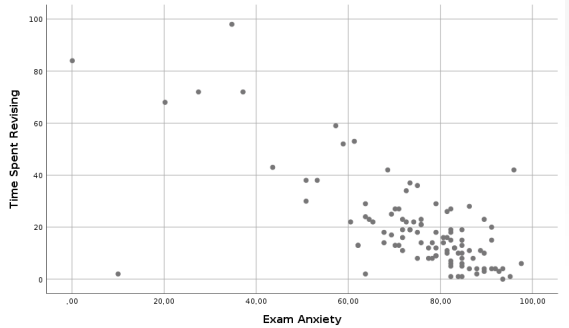
- **check normality:**
  → Analyze → Descriptive statistics → Explore…
  tick «Normality plots» under «Plots»; assess «Tests of Normality»
- **check multicollinearity:**
  → Analyze → Correlate → Bivariate; enter the three variables «Revise», «Exam», and «Anxiety»
  all vars. correlate substantially, esp. Revise and Anxiety
  → possibly only include one in the model; check in any case (using hierarchical regression whether adding both improves the prediction)

From correlation to regression

- **check linearity:**
  → Graphs → Legacy dialogs → Scatter / dot…, select «Simple Scatter», click define; in the window that opens, click «Revise» into «Y axis» and «Exam» into «X axis»; copy the syntax either using «Paste», duplicate it twice and create all three possible variable combinations
  check deviations from linearity in the scatter plots







From correlation to regression

Option 1 – Box-Whisker-plots → manually de-select extreme outliers (stars in the Box-Whisker-plots):

- create a new variable (e.g., selSbj) using → Transform → Create variable… use selSbj as «Target variable» and «1» under «Numeric expression»
- manually set participants that are outliers (stars) in the Box-Whisker-plots to «0»
- use → Data → Select cases and choose «selSbj» under «Select filter variable»

From correlation to regression

Option 1 – Box-Whisker-plots (contd.)

- repeat → Analyze → Descriptive statistics → Explore… and → Analyze → Correlate → Bivariate…
  check the results:
  - how do the «Tests of Normality» change?
  - how do the correlations change?
- create another selection variable and further remove the less extreme outliers (circles) and re-check the analyses
- BEFORE you deselect participants on the basis of that they might be outliers, ask yourself whether the outliers could be genuine (it might be uncommon but valid to be 2.04 m tall)

From correlation to regression

Option 2 – z-scores (use the table with desriptive stats.):

- create a new variable (e.g., selSbjZ)
- re-arrange the table using the Pivoting trays
- calculate M +/- 3.3 * SD (manually or in Excel)

Option 3 – multivariate outliers (Mahalanobis)

- create a new variable (e.g., selSbjM)
- Analyze → Regression → Linear; «Save»-button, select Distances – Mahalanobis; crit. $\chi^2 = 16.266$

Questions?
Comments?

From correlation to regression

→ Analyze → Correlate → Bivariate…
→ Analyze → Regression → Linear (Revise → Exam)

Correlations

		Time Spent Revising	Exam Performance (%)	Exam Anxiety
Time Spent Revising	Pearson Correlation	1	,322**	-,620**
	Sig. (2-tailed)		,001	,000
	N	97	97	97
Exam Performance (%)	Pearson Correlation	,322**	1	-,317**
	Sig. (2-tailed)	,001		,002
	N	97	97	97
Exam Anxiety	Pearson Correlation	-,620**	-,317**	1
	Sig. (2-tailed)	,000	,002	
	N	97	97	97

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

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,322[a]	,104	,094	11,716

a. Predictors: (Constant), Exam Performance (%)

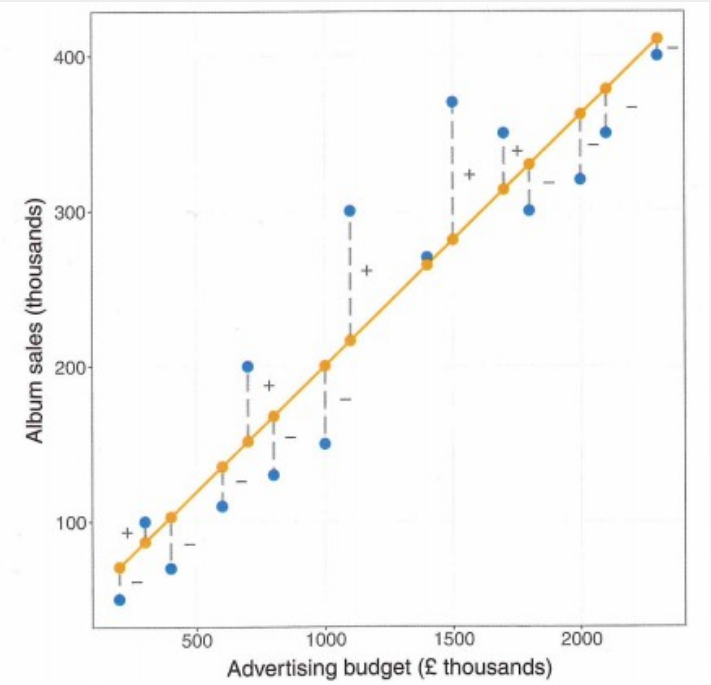
b. Dependent Variable: Time Spent Revising

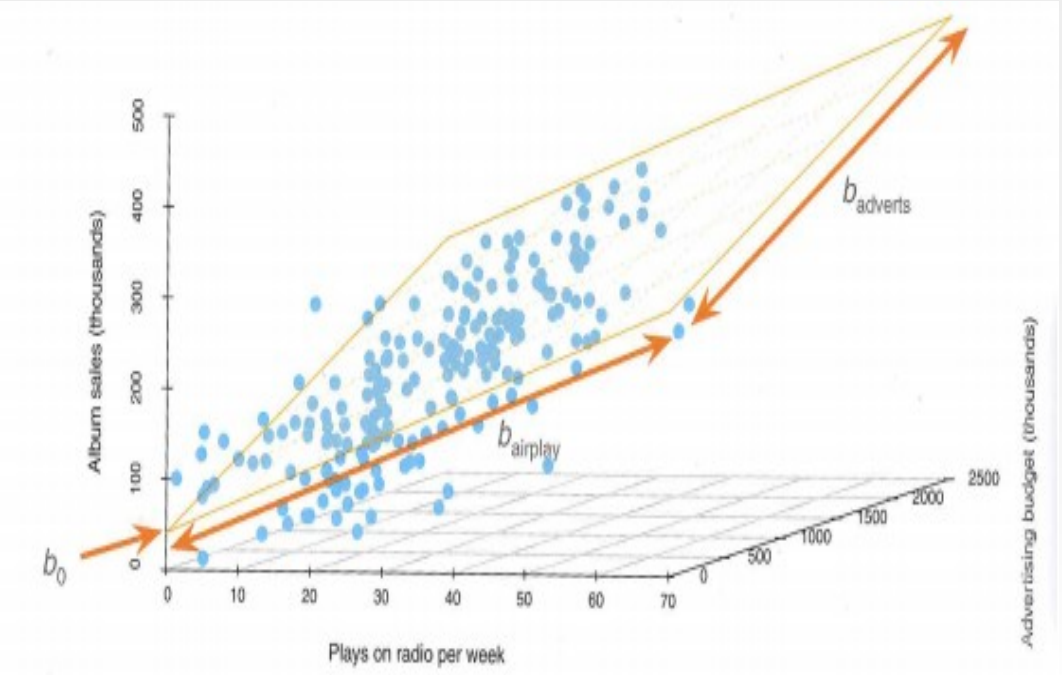
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	8,414	2,850		2,953	,004
	Exam Performance (%)	,158	,048	,322	3,316	,001

a. Dependent Variable: Time Spent Revising

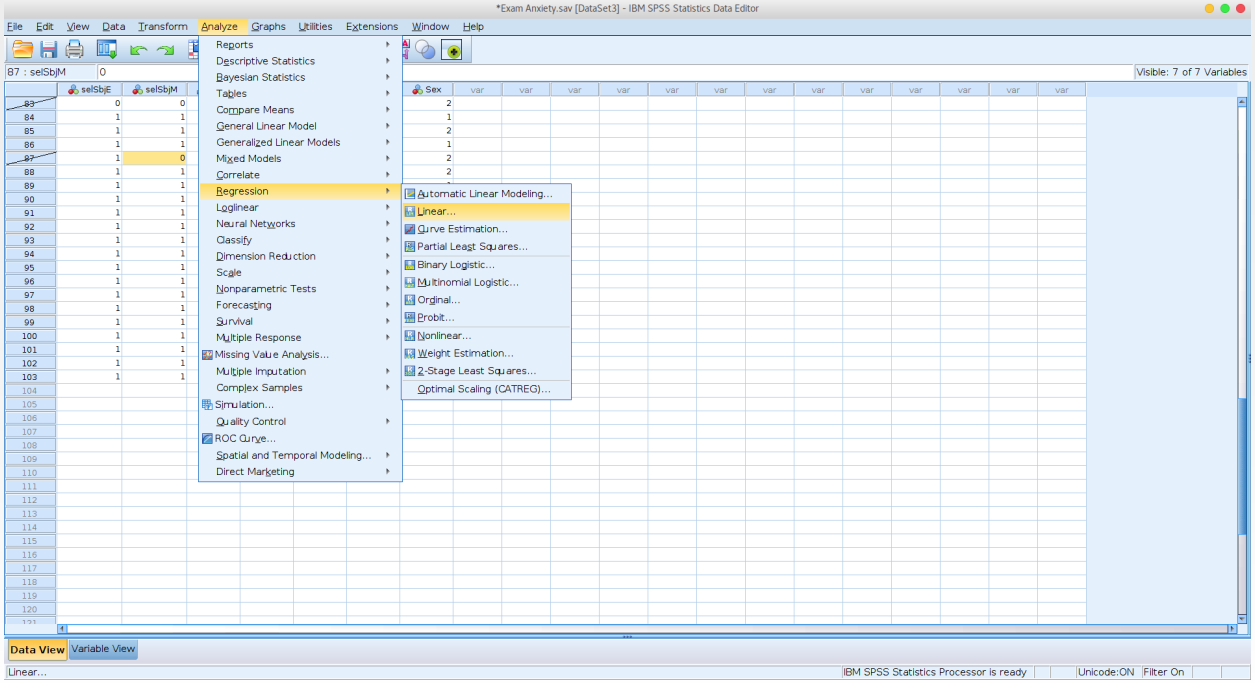
Regression: From uni- to multivar.



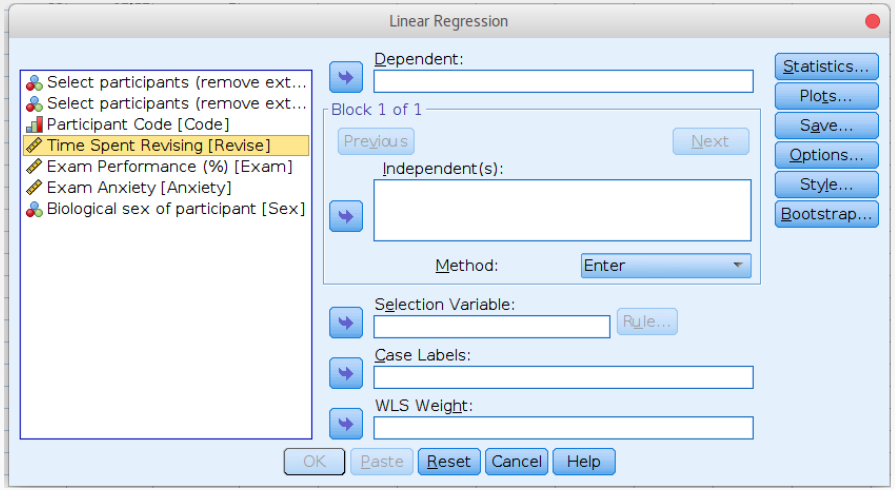


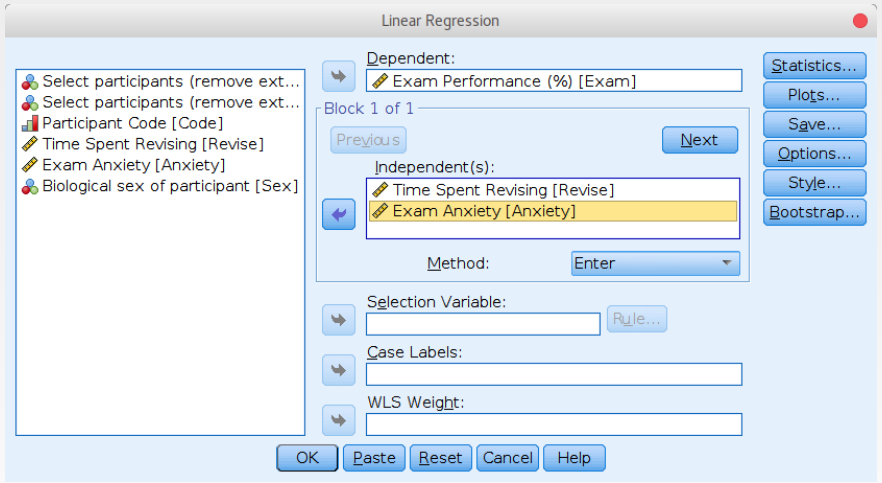
Questions?
Comments?

Regression: From uni- to multivar.



Regression: From uni- to multivar.





Regression: From uni- to multivar.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,322[a]	,104	,094	23,928

a. Predictors: (Constant), Time Spent Revising

b. Dependent Variable: Exam Performance (%)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6294,045	1	6294,045	10,993	,001[b]
	Residual	54393,997	95	572,568		
	Total	60688,041	96			

a. Dependent Variable: Exam Performance (%)

b. Predictors: (Constant), Time Spent Revising

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	43,272	4,157		10,410	,000
	Time Spent Revising	,658	,198	,322	3,316	,001

a. Dependent Variable: Exam Performance (%)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,355[a]	,126	,107	23,755

a. Predictors: (Constant), Exam Anxiety, Time Spent Revising

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7641,955	2	3820,977	6,771	,002[b]
	Residual	53046,086	94	564,320		
	Total	60688,041	96			

a. Dependent Variable: Exam Performance (%)

b. Predictors: (Constant), Exam Anxiety, Time Spent Revising

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	81,471	25,059		3,251	,002
	Time Spent Revising	,417	,251	,204	1,662	,100
	Exam Anxiety	-,440	,284	-,190	-1,545	,126

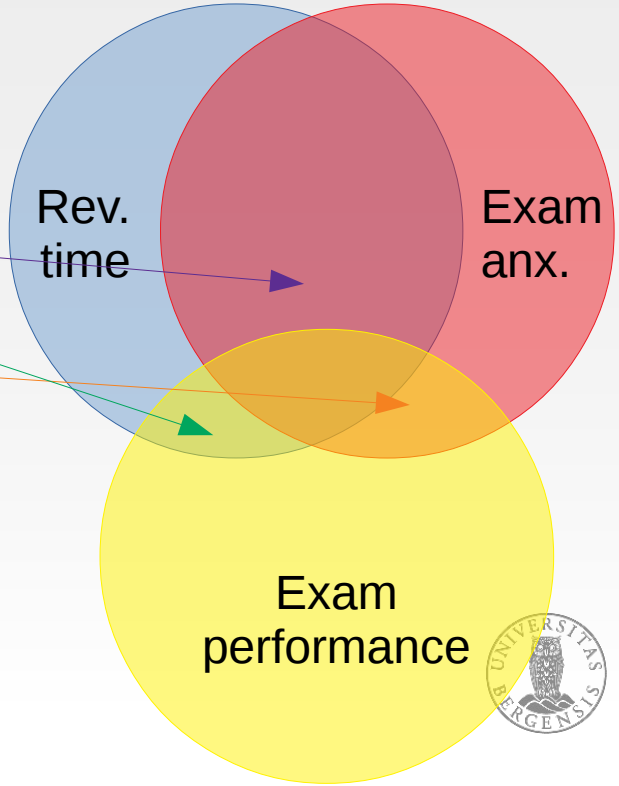
a. Dependent Variable: Exam Performance (%)

Regression: From uni- to multivar.

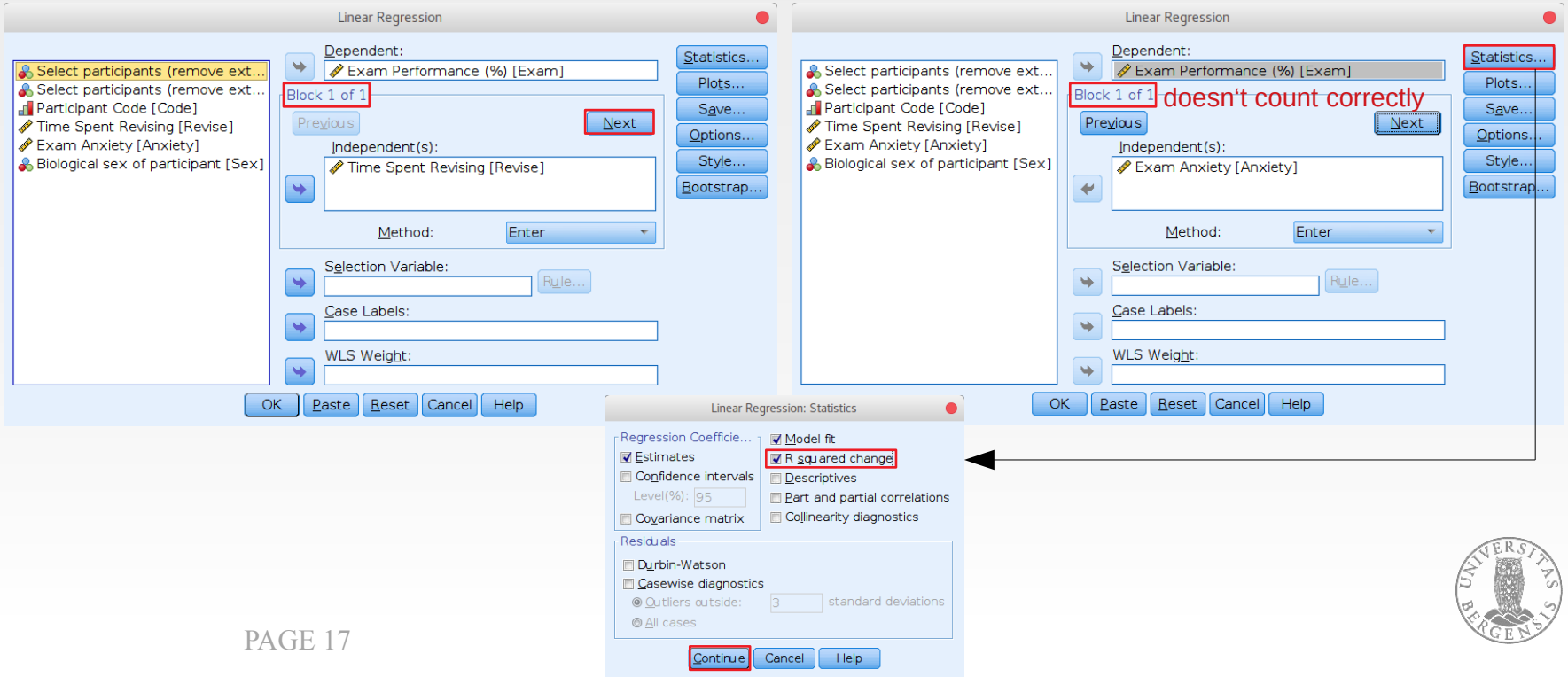
Correlations

		Time Spent Revising	Exam Performance (%)	Exam Anxiety
Time Spent Revising	Pearson Correlation	1	,322**	-,620**
	Sig. (2-tailed)		,001	,000
	N	97	97	97
Exam Performance (%)	Pearson Correlation	,322**	1	-,317**
	Sig. (2-tailed)	,001		,002
	N	97	97	97
Exam Anxiety	Pearson Correlation	-,620**	-,317**	1
	Sig. (2-tailed)	,000	,002	
	N	97	97	97

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



Regression: Entering predictors



Regression: Entering predictors

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,237[a]	,056	,045	24,616	,056	5,103	1	86	,026
2	,290[b]	,084	,063	24,388	,028	2,609	1	85	,110

a. Predictors: (Constant), Time Spent Revising

b. Predictors: (Constant), Time Spent Revising, Exam Anxiety

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3092,020	1	3092,020	5,103	,026[b]
	Residual	52109,878	86	605,929		
	Total	55201,898	87			
2	Regression	4643,996	2	2321,998	3,904	,024[c]
	Residual	50557,902	85	594,799		
	Total	55201,898	87			

a. Dependent Variable: Exam Performance (%)

b. Predictors: (Constant), Time Spent Revising

c. Predictors: (Constant), Time Spent Revising, Exam Anxiety

Coefficients[a]

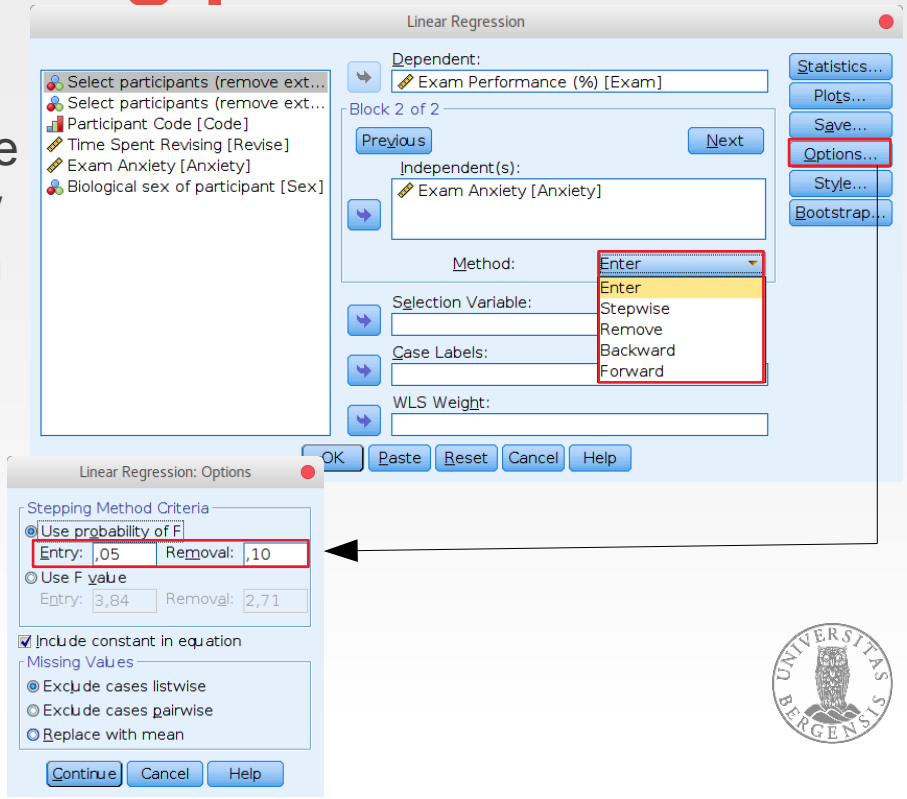
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42,735	5,044		8,472	,000
	Time Spent Revising	,684	,303	,237	2,259	,026
2	(Constant)	92,430	31,168		2,966	,004
	Time Spent Revising	,393	,350	,136	1,122	,265
	Exam Anxiety	-,574	,355	-,196	-1,615	,110

a. Dependent Variable: Exam Performance (%)

Regression: Entering predictors

Methods for entering:

- «Enter»: Enter one new variable
- «Stepwise»: Enter multiple new variables (one step at a time an in order of explained variance) according to F-probability
- «Remove»: Remove one var.
- «Backward»: Remove one var. according to F-probability
- «Forward»: enter one variable according to F-probability



Regression: Entering predictors

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,266[a]	,071	,060	24,425	,071	6,530	1	86	,012

a. Predictors: (Constant), Exam Anxiety

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3895,715	1	3895,715	6,530	,012[b]
	Residual	51306,183	86	596,584		
	Total	55201,898	87			

a. Dependent Variable: Exam Performance (%)

b. Predictors: (Constant), Exam Anxiety

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	114,314	24,343		4,696	,000
	Exam Anxiety	-,779	,305	-,266	-2,555	,012

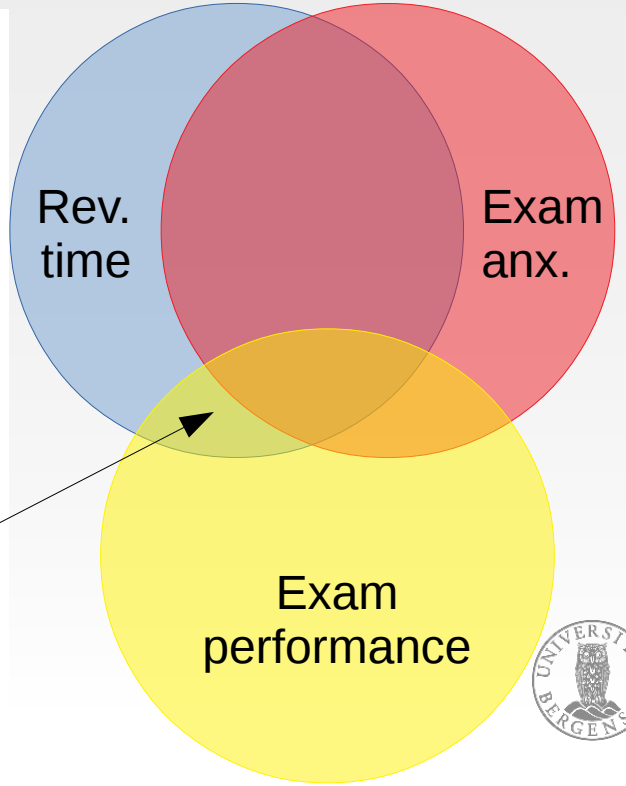
a. Dependent Variable: Exam Performance (%)

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	Time Spent Revising	,136[b]	1,122	,265	,121	,734

a. Dependent Variable: Exam Performance (%)

b. Predictors in the Model: (Constant), Exam Anxiety

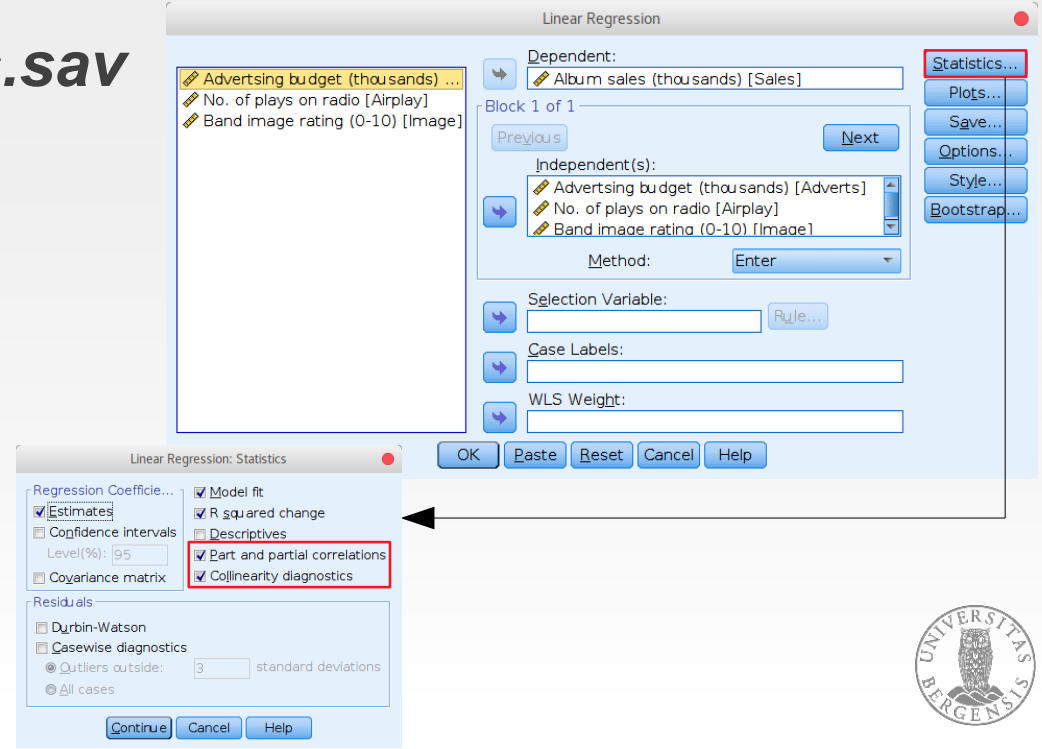


Questions?
Comments?

Use Album Sales.sav Predict Sales from Adverts, Airplay, and Image
(~ 5 – 10 mins)

Regression: Diagnostics

- use ***Album Sales.sav***
- → Analyze
  → Regression
  → Linear…
- click «Statistics» and tick those



Regression: Diagnostics

- collinearity describes a linear association between explanatory variables (i.e. the degree to which one explanatory variable can be predicted by a combination of one or more other explanatory variables)
- tolerance: $1 - R_j^2$ (R_j^2 – what degree of variance of variable j is explained by the other predictor variables)
- variance inflation factor (VIF): 1 / tolerance
- (a) VIF < 5 and tolerance > 0.2; (b) the average of the VIF of all variables should be close to 1

Regression: Diagnostics

- all tolerances are > 0.2, all VIF < 5
- the average VIF is (1.015 + 1.043 + 1.038) / 3 = 1.032 which is close to 1

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-26,613	17,350		-1,534	,127					
	Advertsing budget (thousands)	,085	,007	,511	12,261	,000	,578	,659	,507	,986	1,015
	No. of plays on radio	3,367	,278	,512	12,123	,000	,599	,655	,501	,959	1,043
	Band image rating (0-10)	11,086	2,438	,192	4,548	,000	,326	,309	,188	,963	1,038

a. Dependent Variable: Album sales (thousands)

Regression: Diagnostics

- $R^2 = A + B + C + D + E$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	,815[a]	,665	,660	47,087	,665	129,498	3	196	,000

a. Predictors: (Constant), Band image rating (0-10), Advertsing budget (thousands), No. of plays on radio

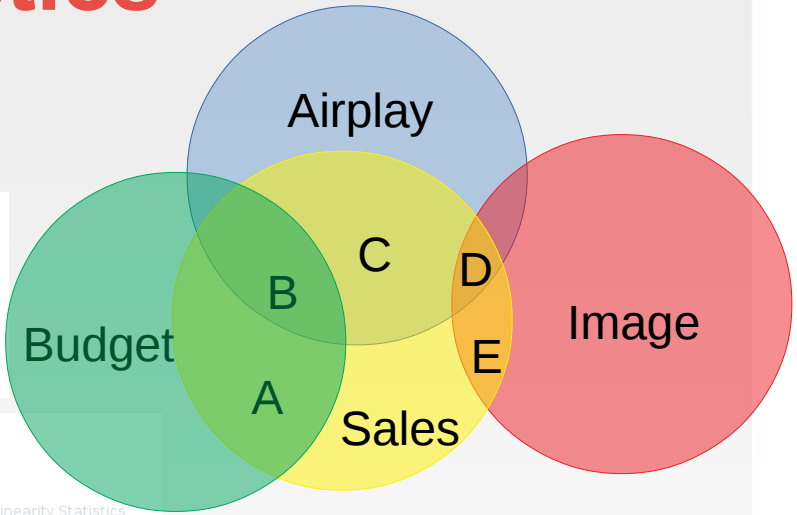
b. Dependent Variable: Album sales (thousands)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-26,613	17,350		-1,534	,127					
	Advertsing budget (thousands)	,085	,007	,511	12,261	,000	,578	,659	,507	,986	1,015
	No. of plays on radio	3,367	,278	,512	12,123	,000	,599	,655	,501	,959	1,043
	Band image rating (0-10)	11,086	2,438	,192	4,548	,000	,326	,309	,188	,963	1,038

a. Dependent Variable: Album sales (thousands)

= A
= C
= E



Regression: Diagnostics

- within the Variance Proportions, for each dimension should only one variable have high loadings

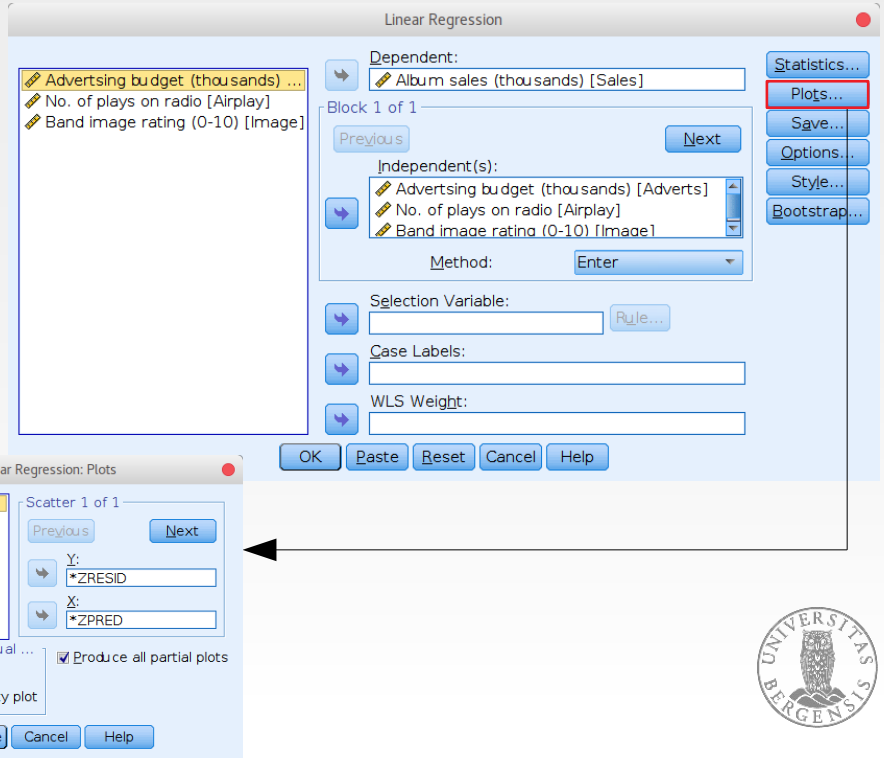
Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Advertsing budget (thousands)	No. of plays on radio	Band image rating (0-10)
1	1	3,562	1,000	,00	,02	,01	,00
	2	,308	3,401	,01	,96	,05	,01
	3	,109	5,704	,05	,02	,93	,07
	4	,020	13,219	,94	,00	,00	,92

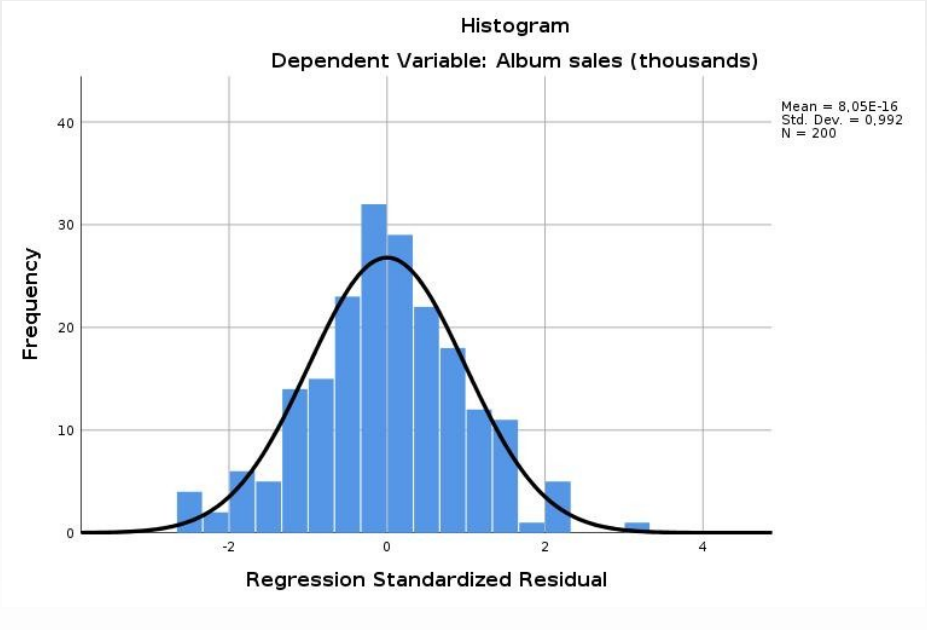
a. Dependent Variable: Album sales (thousands)

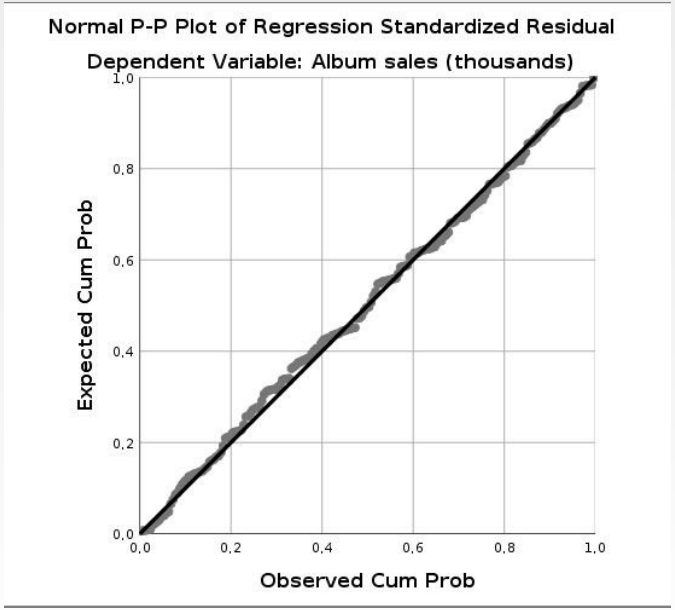
Regression: Diagnostics

- click «Plots»
  (a) tick all options
  (b) click ZPRED to X
  click ZRESID to Y



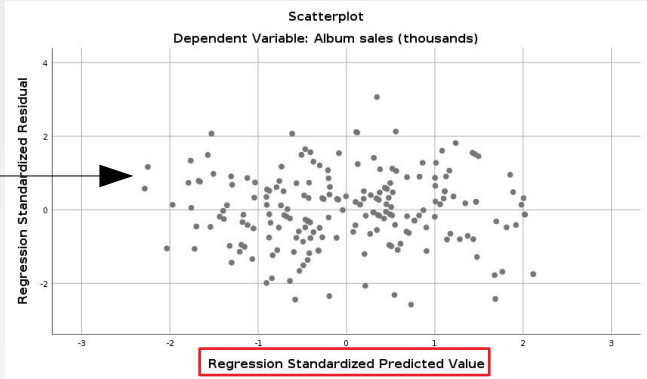
Regression: Diagnostics

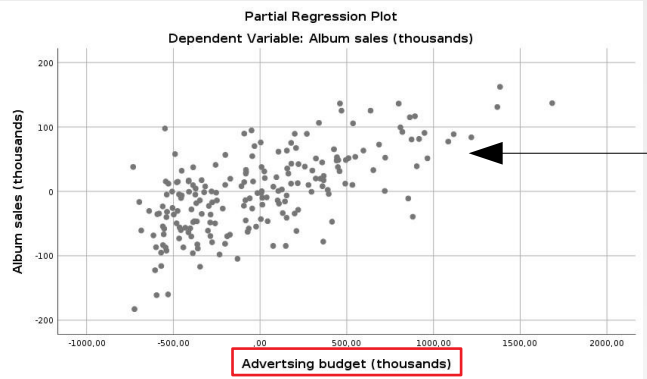




Regression: Diagnostics

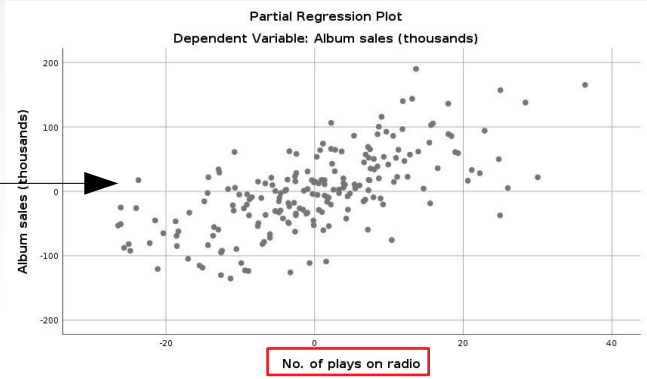
should have no shape ($r \approx 0$)

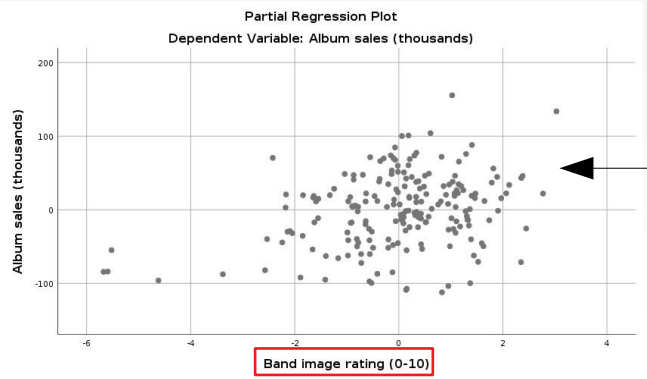




should show a linear trend ($r \neq 0$)

should show a linear trend ($r \neq 0$)





Questions?
Comments?

Logistic Regression

- predicting categorical outcomes from categorical and continous predictors (binary / multinomial)
- log-transform the result of the GLM → probability of the category (e.g., successful treatm.) to occur

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_n X_{ni})}}$$

→ maximum-likelihood-estimation

Logistic Regression

assessing the quality of the model:

$$\text{log-likelihood} = \sum_{i=1}^{N} \left[Y_i \ln\left(P(Y_i)\right) + (1 - Y_i) \ln\left(1 - P(Y_i)\right) \right]$$

deviance = -2 × log-likelihood (-2LL); χ^2-distribtd.

$\chi^2 = (-2LL_{baseline}) - (-2LL_{model}) = 2LL_{model} - 2LL_{baseline}$

$df = k_{model} - k_{baseline}$ (predictors + 1 [intercept])

Logistic Regression

assessing the quality of the model:

- $R = \sqrt{(\chi^2 - 2df) / -2LL_{baseline}}$

  $R^2_{HL} = (2LL_{model} - 2LL_{baseline}) / -2LL_{baseline}$ (Hosmer & Lemeshow)

  $R^2_{CS} = 1 - \exp((2LL_{baseline} - 2LL_{model}) / n)$ (Cox & Snell)
- Wald (assessing signif. of predictors): $z = b / SE_b$
- odds-ratio = P(event) / P(non-event)
  = odds after unit chg. in pred. / orig. odds
  (if > 1: if the pred. increases the prob. of outcome incr.
  if < 1: if the pred. increases the prob. of outcome decr.

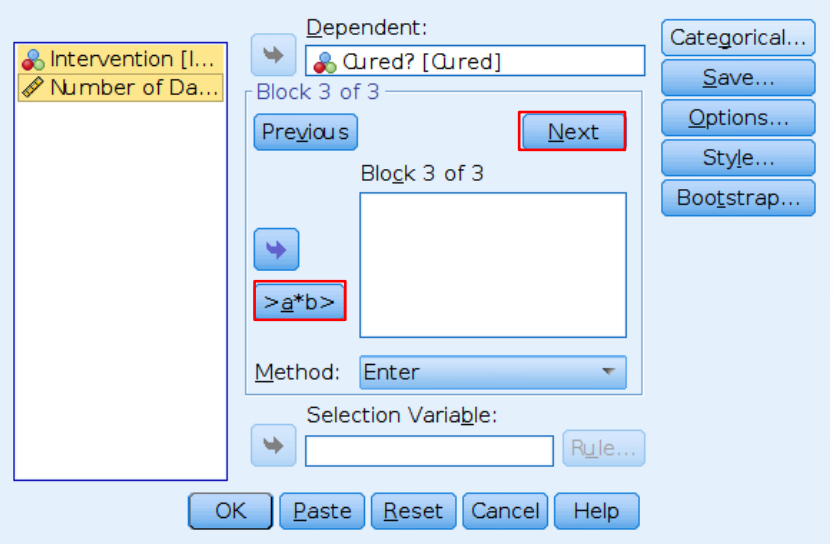
Logistic Regression

assumptions:

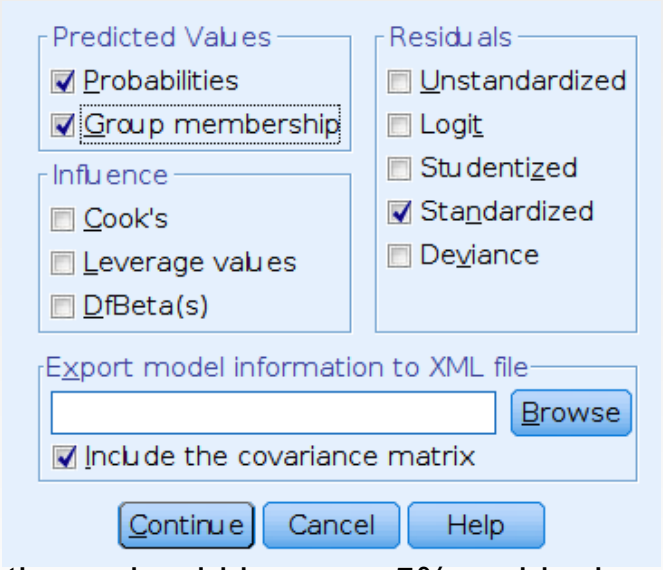
- linear relationship between any (continous) predictor and the logit of the outcome
  (can be tested by testing the significance of the interaction of a predictor with it's log-transform.)
- independence of errors
- threats to convergence: incomplete information
  (not all possible combinations of variables available)
  complete separation (outcome can be perfectly predicted by one or a combination of variables)

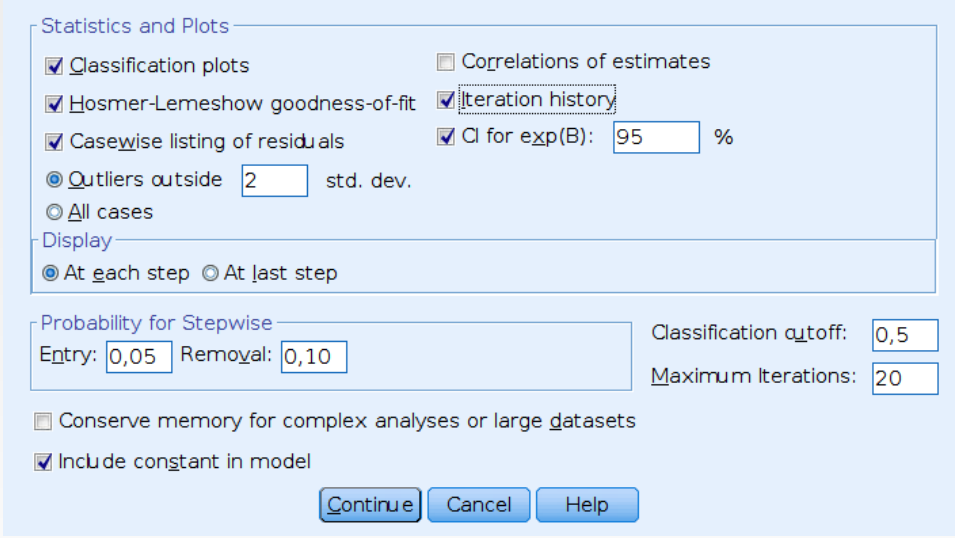
Logistic Regression

- using Eel.sav:
  Analyze → Regression → Binary Logistic
  DV: Cured;
  IV: Intervention [Block 1; press Next] Duration [Block 2];
  Intervention × Duration [Block 3; select both by holding [Ctrl] + Click, press button >a*b>]
  use «Enter» as method



Logistic Regression





there should be max. 5% residuals > 2; max. %1 > 2.5; > 3 is certainly an outlier
Cook's distance > 1: case influences the model; look out for DfBeta > 1
Leverage should be with 2-3 times predictors / N (2 / 113 = 0.018 → check > 0.036)

Logistic Regression

Block 2: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	,002	1	,964
	Block	,002	1	,964
	Model	9,928	2	,007

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	144,156[a]	,084	,113

a. Estimation terminated at iteration number 3 because parameter estimates changed by less than ,001.

Classification Table[a]

			Predicted		
			Cured?		
	Observed		Not Cured	Cured	Percentage Correct
Step 1	Cured?	Not Cured	32	16	66,7
		Cured	24	41	63,1
	Overall Percentage				64,6

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Intervention(1)	1,234	,415	8,854	1	,003	3,433	1,523	7,737
	Number of Days with Problem before Treatment	-,008	,176	,002	1	,964	,992	,703	1,401
	Constant	-,235	1,221	,037	1	,848	,791		

a. Variable(s) entered on step 1: Number of Days with Problem before Treatment.

use Penalty.sav
DV: Scored
IVs: PSWQ, Anxious, Previous
(10 - 15 min)

It’s your turn now!

UNIVERSITY OF BERGEN

Eco-Lighthouse®