WEEK 14

SIMPLE REGRESSION ANALYSIS

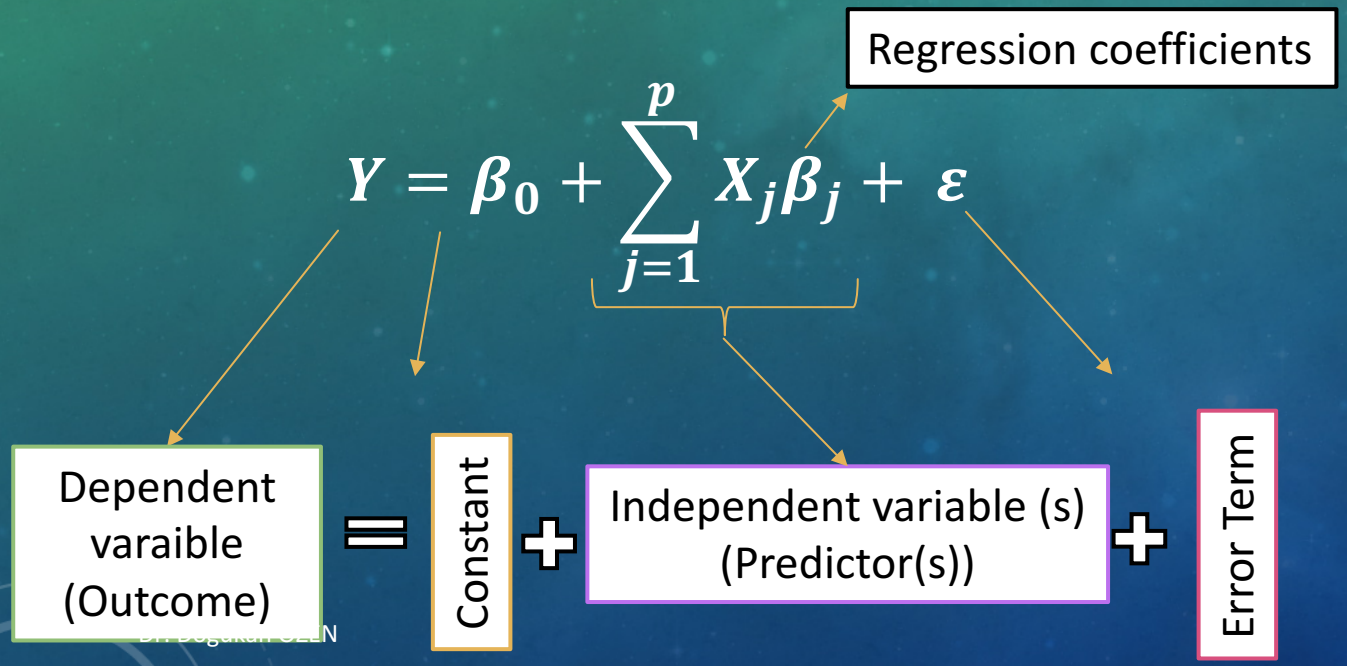
LINEAR REGRESSION

- In **linear correlation** we are concerned with determining whether there is a linear relationship between two numerical variables, and with measuring the degree of that relationship.

- **In linear regression** we describe the linear relationship between the two variables by determining the mathematical equation that relates the variables.
  - We often use this equation to predict the value of one variable (called the outcome, dependent or response variable) from a value of the other variable (called the explanatory, independent or predictor variable)

AIM:

- To determine a mathematical equation that relates the variables
- predict the value of the outcome (or dependent variable) from a value of the other variable(s) (independent variables)



Simple regression

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_i X_{i+} \varepsilon$$

ASSUMPTIONS

Variable type	All predictor variables should be continuous or categorical
	Outcome variable should be continuous
No multicollinearity	The predictor variables should not correlate too highly
Homoscedasticity	The residuals at each level of the predictor should have the same variance
Independent errors (Autocorrelation)	For any two observations, the residual terms should be uncorrelated (*It is tested by Durbin Watson test*)
Normally distributed errors	Residuals in the model should be random and normally distibuted with a mean of 0.
Linearity	The outcome variable should be linearly related to any predictors

EXAMPLE

	bodyweight	headlength	chestdepth	chestwidth	bodylength	heightatwithers	heightatrum
1	54,60	24,50	32,50	19,50	68,00	66,00	67,0
2	56,80	21,50	32,00	19,50	67,50	65,00	65,0
3	50,00	21,00	33,00	17,50	69,00	65,00	64,0
4	54,70	23,00	32,00	21,00	70,50	65,50	63,5
5	60,20	21,00	34,00	21,50	69,00	64,50	63,5
6	44,30	20,50	29,00	19,00	65,50	63,00	63,5
7	48,60	22,00	31,50	21,00	62,00	62,00	63,0
8	57,00	22,50	34,00	20,00	70,00	65,00	63,0
9	54,80	22,00	31,00	21,00	71,00	64,00	63,0
10	57,40	21,00	32,00	18,00	70,50	65,50	63,0
11	62,10	21,50	34,00	21,00	68,00	66,00	63,0
12	53,00	24,00	35,00	20,00	66,00	66,00	63,0
13	48,50	22,00	31,50	21,00	62,00	62,00	63,0
14	50,60	24,00	33,00	22,00	66,50	64,50	62,5
15	51,30	18,50	31,50	21,50	69,50	64,00	62,5
16	46,40	21,00	30,00	18,50	69,00	65,50	62,5
17	45,60	20,50	32,50	19,00	66,50	63,50	62,5
18	61,70	22,50	31,50	20,50	70,50	64,00	62,5

A researcher wants to determine a mathematical equation that predicts bodyweight from some body measurements (eg. Headlength, chestdepth, chestwidth, bodylength, withersheight, rumpheight). What would be the model?

Dataset› BW_Regression.sav

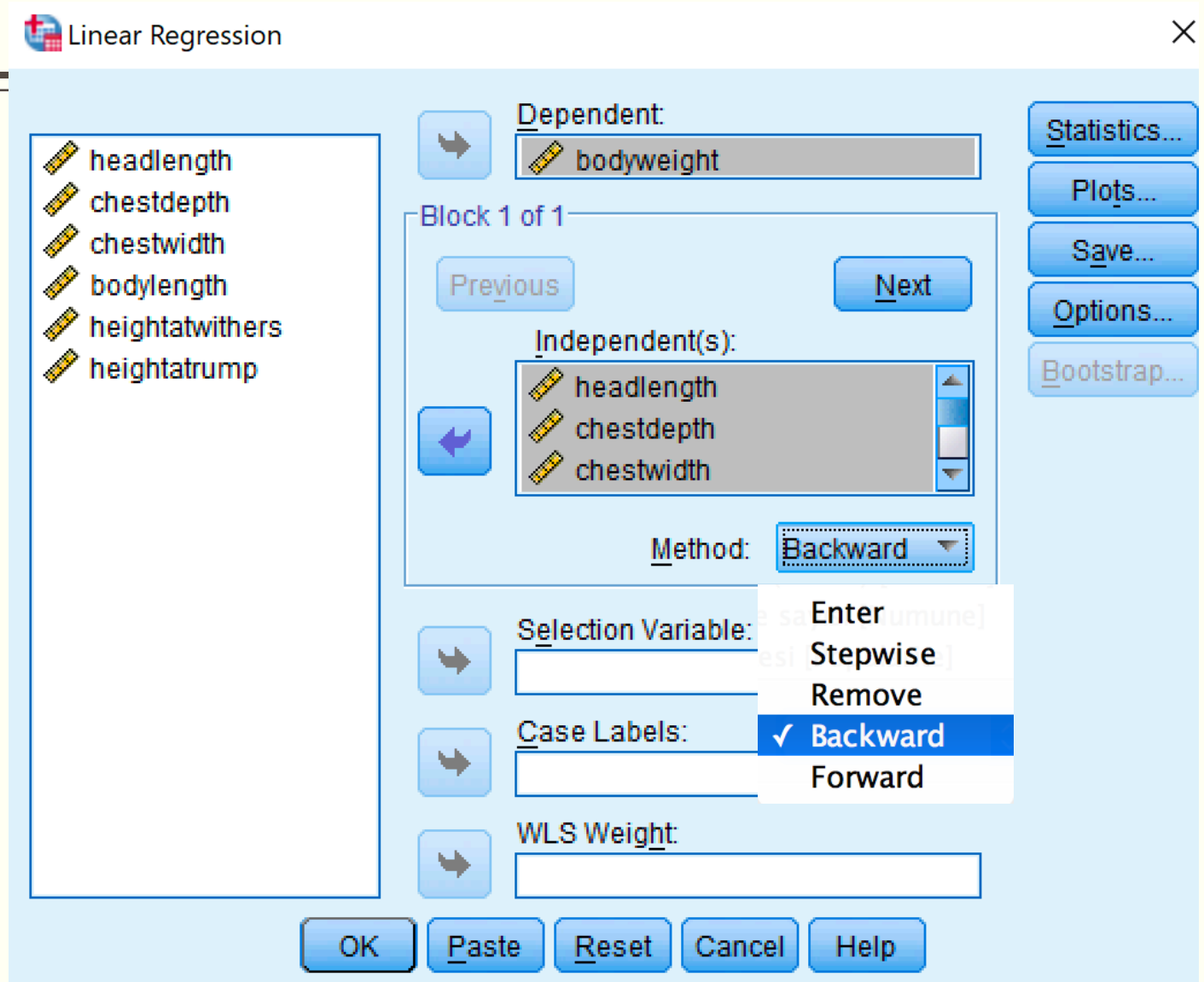
Data analysis

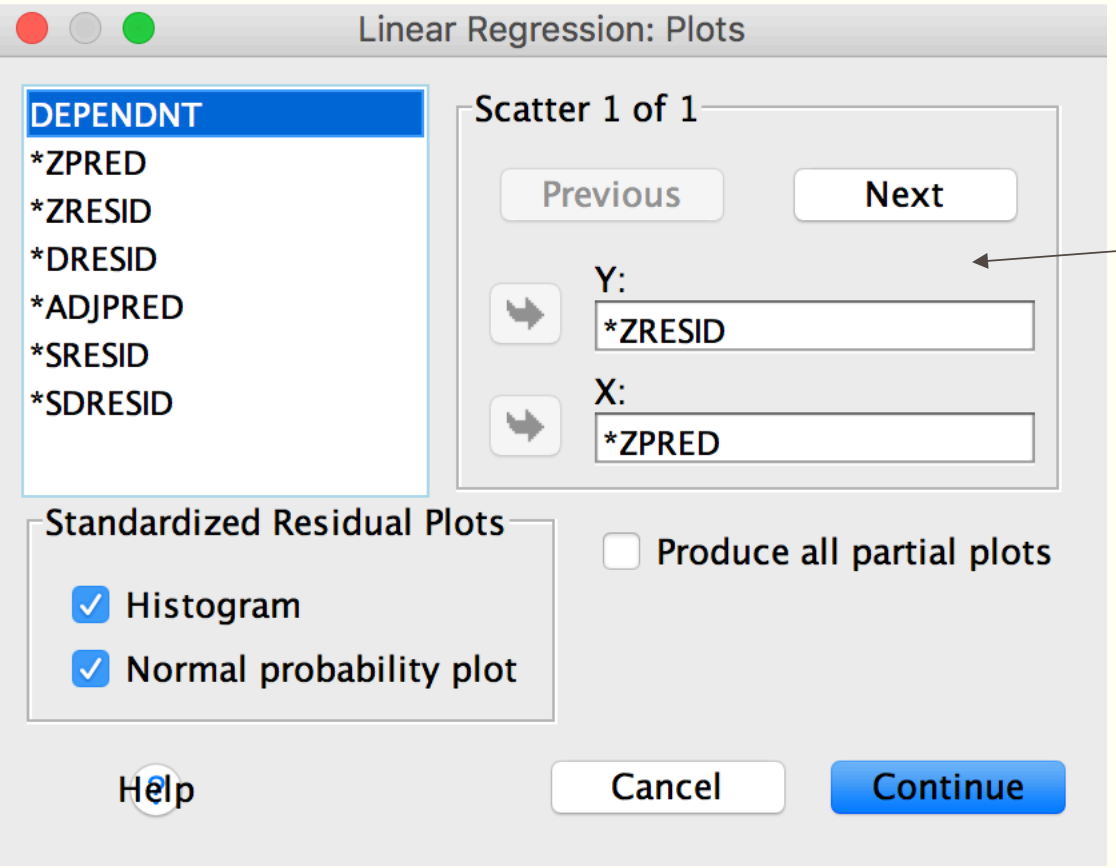
Analyze > Regression > Linear Regression

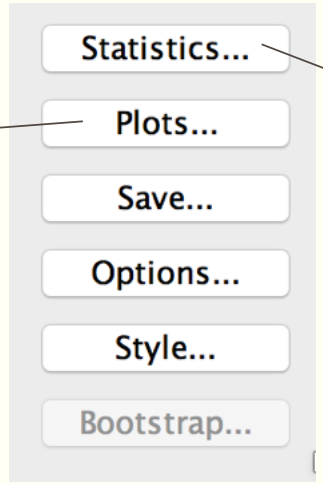
What should be the method for modelling?

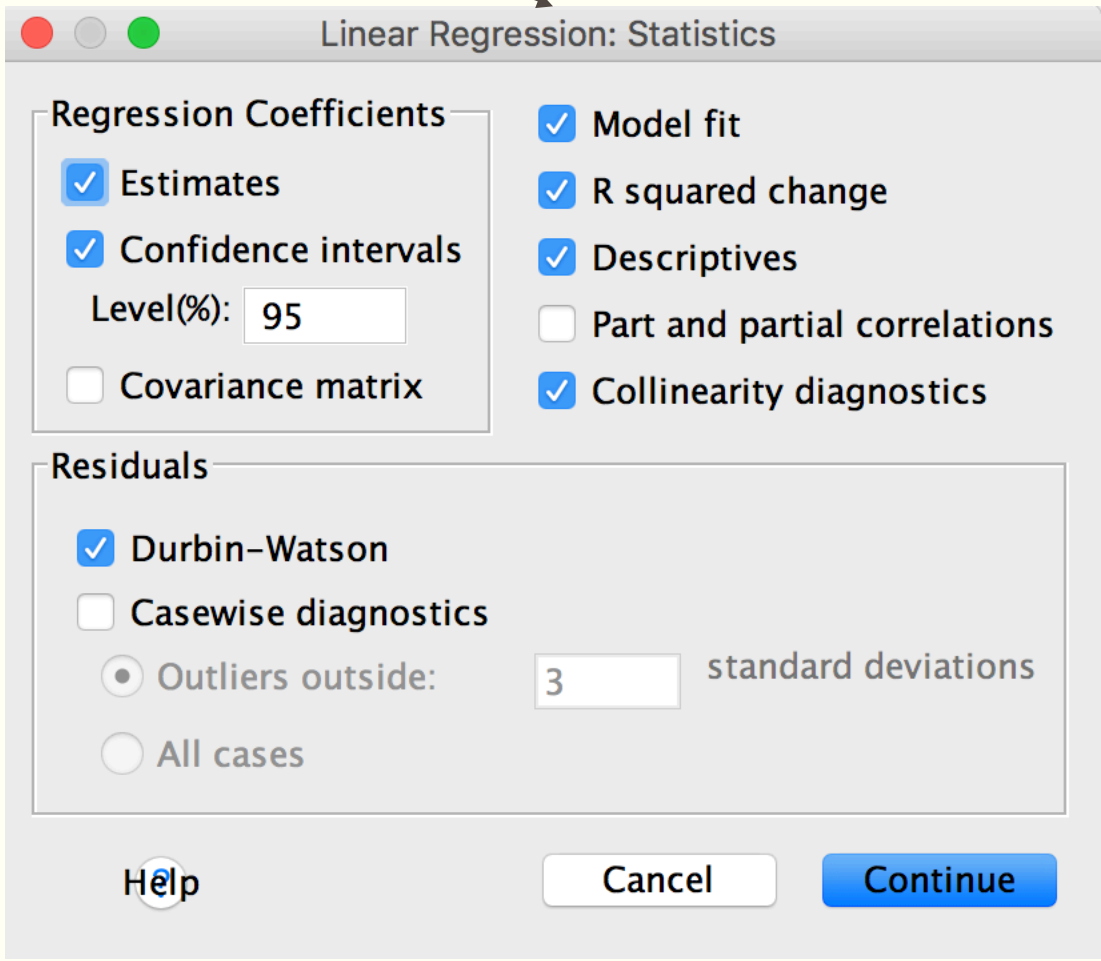
literature?

Enter: all predictors are forced into the model simultaneously.
Forward: an initial model is defined that contains only the constant. Then computer adds next best predictor that has highest simple correlation with the outcome, and so on..
Stepwise: Same as the forward method, except that each time a predictor is added to the equation, a removal test is made of the least useful predictor.
Backward: The opposite of forward method.




Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Scatter 1 of 1
Previous
Next
Y:
*ZRESID
X:
*ZPRED
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
Help
Cancel
Continue


Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...


Linear Regression: Statistics
Regression Coefficients
Estimates
Confidence intervals
Level(%): 95
Covariance matrix
Model fit
R squared change
Descriptives
Part and partial correlations
Collinearity diagnostics
Residuals
Durbin-Watson
Casewise diagnostics
Outliers outside: 3 standard deviations
All cases
Help
Cancel
Continue

Output:

Descriptive Statistics

	Mean	Std. Deviation	N
bodyweight	47,9988	4,94456	250
headlength	19,7720	1,94588	250
chestdepth	30,6055	1,70549	250
chestwidth	18,8786	1,50109	250
bodylength	65,4968	3,01354	250
heightatwithers	61,5152	2,03227	250
heightatrump	59,3536	1,99137	250

Correlations

		bodyweight	headlength	chestdepth	chestwidth	bodylength	heightatwithers	heightatrump
Pearson Correlation	bodyweight	1,000	,366	,469	,433	,262	,353	,290
	headlength	,366	1,000	,477	,367	,300	,359	,399
	chestdepth	,469	,477	1,000	,539	,511	,601	,395
	chestwidth	,433	,367	,539	1,000	,494	,312	,397
	bodylength	,262	,300	,511	,494	1,000	,422	,317
	heightatwithers	,353	,359	,601	,312	,422	1,000	,665
	heightatrump	,290	,399	,395	,397	,317	,665	1,000
Sig. (1-tailed)	bodyweight	.	,000	,000	,000	,000	,000	,000
	headlength	,000	.	,000	,000	,000	,000	,000
	chestdepth	,000	,000	.	,000	,000	,000	,000
	chestwidth	,000	,000	,000	.	,000	,000	,000
	bodylength	,000	,000	,000	,000	.	,000	,000
	heightatwithers	,000	,000	,000	,000	,000	.	,000
	heightatrump	,000	,000	,000	,000	,000	,000	.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	heightatrump, bodylength, headlength, chestwidth, chestdepth, heightatwithers[b]	.	Enter
2	.	heightatrump	Backward (criterion: Probability of F-to-remove >= ,100).
3	.	bodylength	Backward (criterion: Probability of F-to-remove >= ,100).
4	.	heightatwithers	Backward (criterion: Probability of F-to-remove >= ,100).

a. Dependent Variable: bodyweight

b. All requested variables entered.

Watson (1951) DW: 1-3 - ok

Model Summary[e]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,542[a]	,293	,276	4,20739	,293	16,816	6	243	,000	
2	,541[b]	,293	,279	4,19971	,000	,110	1	243	,740	
3	,538[c]	,289	,278	4,20181	-,004	1,246	1	244	,265	
4	,532[d]	,283	,274	4,21194	-,006	2,187	1	245	,140	1,930

a. Predictors: (Constant), heightatrump, bodylength, headlength, chestwidth, chestdepth, heightatwithers

b. Predictors: (Constant), bodylength, headlength, chestwidth, chestdepth, heightatwithers

c. Predictors: (Constant), headlength, chestwidth, chestdepth, heightatwithers

d. Predictors: (Constant), headlength, chestwidth, chestdepth

e. Dependent Variable: bodyweight

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1786,118	6	297,686	16,816	,000[b]
	Residual	4301,612	243	17,702		
	Total	6087,730	249			
2	Regression	1784,169	5	356,834	20,231	,000[c]
	Residual	4303,561	244	17,638		
	Total	6087,730	249			
3	Regression	1762,196	4	440,549	24,953	,000[d]
	Residual	4325,534	245	17,655		
	Total	6087,730	249			
4	Regression	1723,584	3	574,528	32,385	,000[e]
	Residual	4364,145	246	17,740		
	Total	6087,730	249			

a. Dependent Variable: bodyweight

b. Predictors: (Constant), heightatrump, bodylength, headlength, chestwidth, chestdepth, heightatwithers

c. Predictors: (Constant), bodylength, headlength, chestwidth, chestdepth, heightatwithers

d. Predictors: (Constant), headlength, chestwidth, chestdepth, heightatwithers

e. Predictors: (Constant), headlength, chestwidth, chestdepth

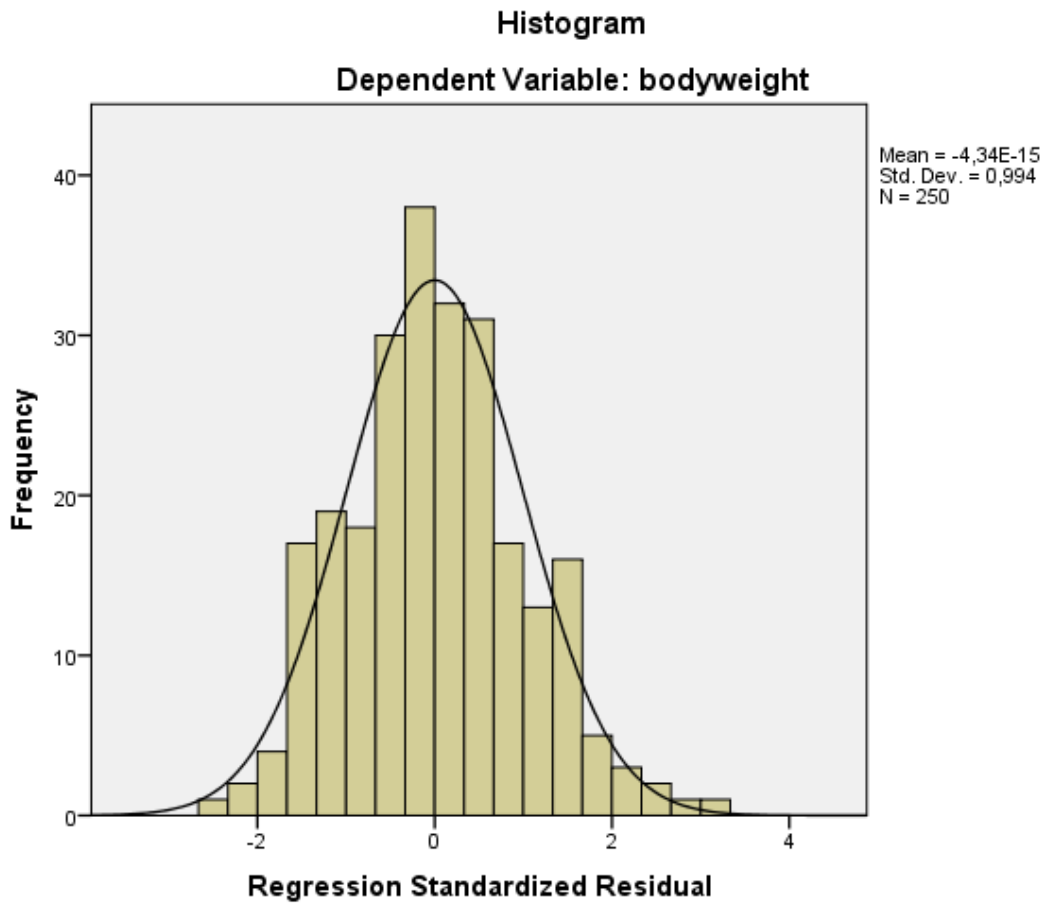
Myers (1990) VIF < 10 !!!
Bowerman ve O'Connell (1990) Tolerance >0.2 !!!

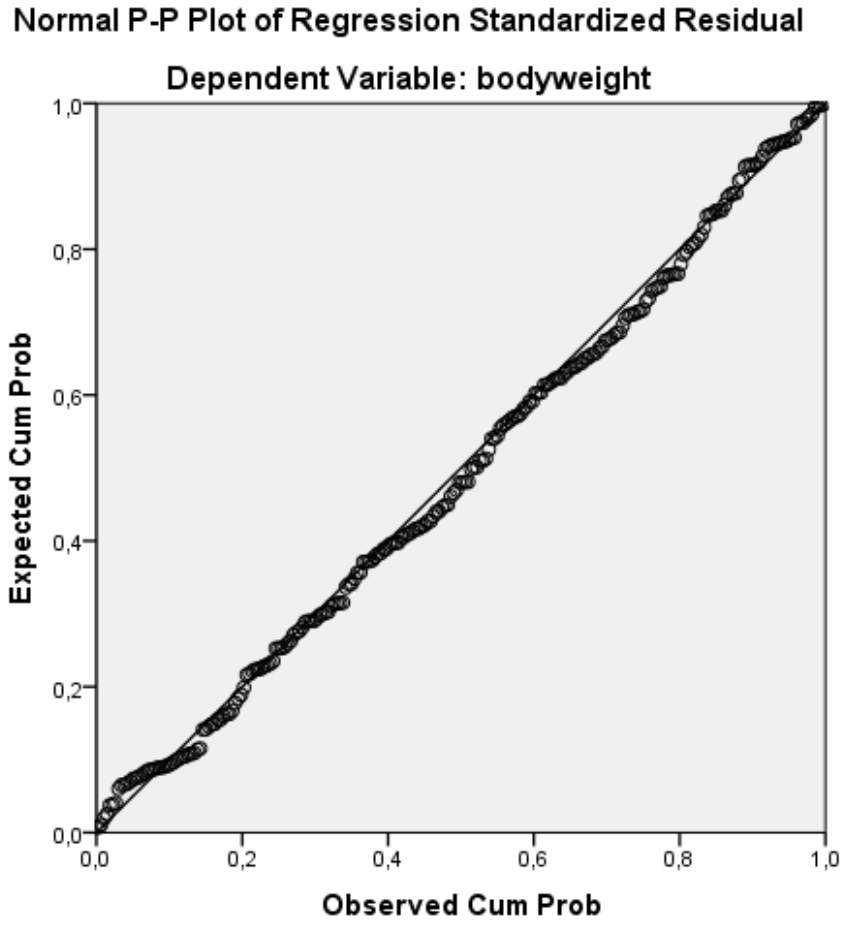
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	-3,610	9,285		-,389	,698	-21,899	14,679		
	headlength	,373	,162	,147	2,299	,022	,053	,693	,712	1,404
	chestdepth	,659	,237	,227	2,779	,006	,192	1,126	,435	2,301
	chestwidth	,867	,233	,263	3,716	,000	,407	1,326	,580	1,724
	bodylength	-,124	,110	-,075	-1,126	,261	-,340	,093	,649	1,542
	heightatwithers	,319	,209	,131	1,523	,129	-,093	,731	,393	2,542
	heightatrump	-,064	,193	-,026	-,332	,740	-,444	,316	,482	2,076
2	(Constant)	-4,791	8,560		-,560	,576	-21,651	12,069		
	headlength	,362	,158	,142	2,285	,023	,050	,673	,747	1,339
	chestdepth	,674	,232	,233	2,905	,004	,217	1,132	,452	2,213
	chestwidth	,845	,224	,257	3,780	,000	,405	1,285	,629	1,590
	bodylength	-,122	,110	-,075	-1,116	,265	-,338	,094	,650	1,539
	heightatwithers	,277	,168	,114	1,655	,099	-,053	,607	,611	1,637
3	(Constant)	-7,765	8,138		-,954	,341	-23,795	8,266		
	headlength	,360	,158	,142	2,271	,024	,048	,671	,747	1,339
	chestdepth	,625	,228	,216	2,741	,007	,176	1,074	,469	2,132
	chestwidth	,769	,213	,234	3,610	,000	,349	1,189	,693	1,443
	heightatwithers	,244	,165	,100	1,479	,140	-,081	,569	,631	1,585
4	(Constant)	1,861	4,898		,380	,704	-7,786	11,508		
	headlength	,385	,158	,151	2,438	,015	,074	,695	,756	1,323
	chestdepth	,791	,199	,273	3,979	,000	,400	1,183	,620	1,614
	chestwidth	,758	,213	,230	3,553	,000	,338	1,179	,694	1,442

a. Dependent Variable: bodyweight

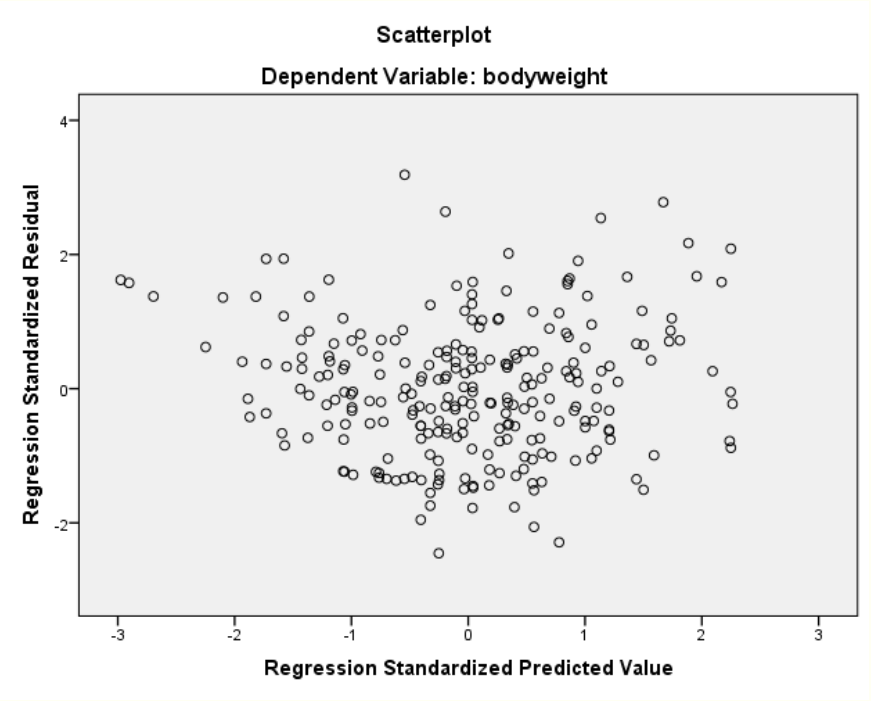
Testing the normality of the residuals...



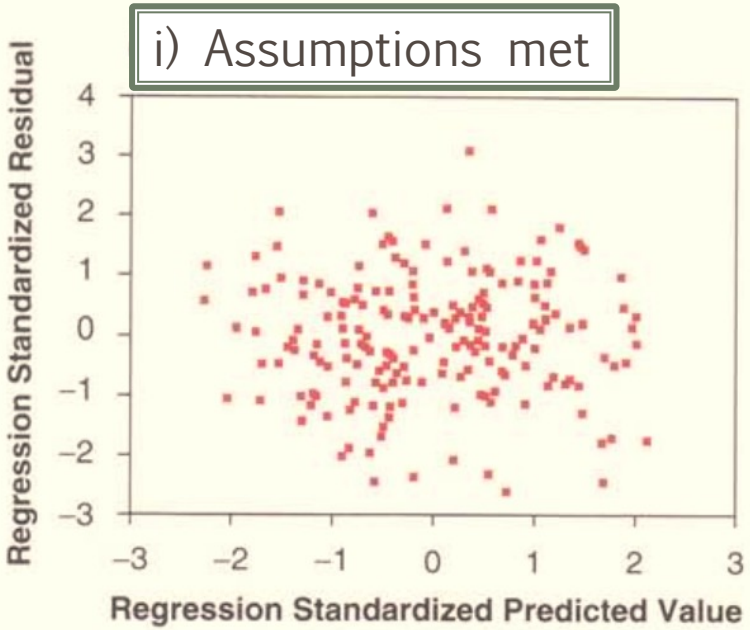


Testing homoscedasticity…

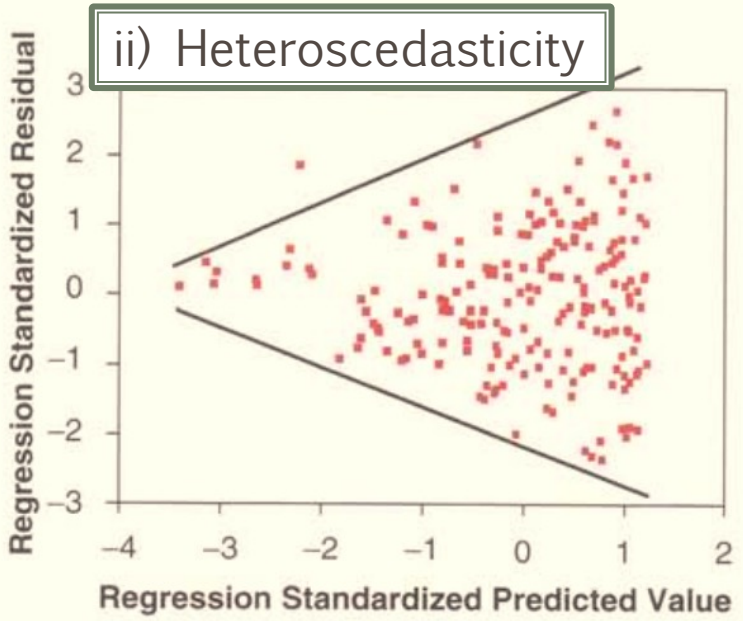
Our example



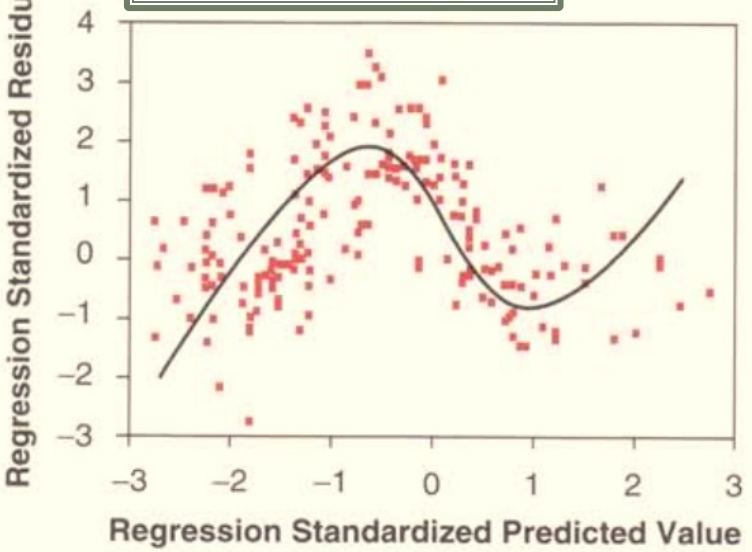
i) Assumptions met



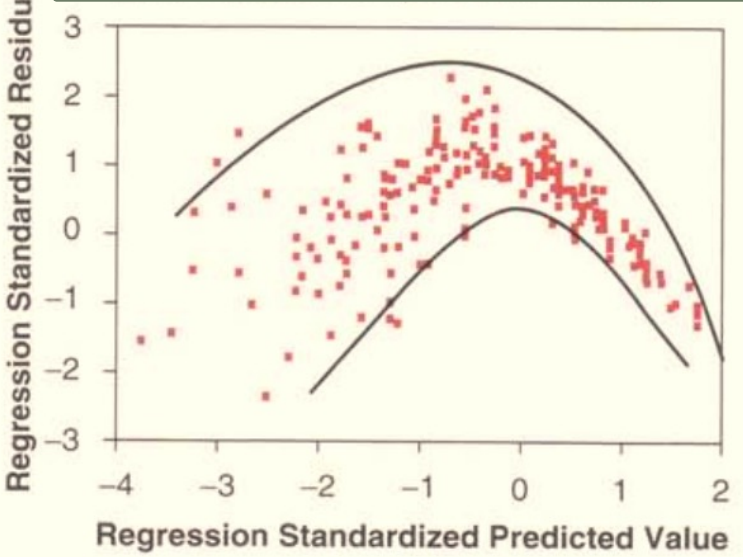
ii) Heteroscedasticity



iii) Non linearity



iv) Heteroscedasticity and non linearity



Reference: Andy Field (2009). Discovering statistics using SPSS, third edition, SAGE Publications, p:248.

Report:

	Unstandardized Coefficients		Standardized Beta	t	P	95% Confidence Interval	
	B	Std. Error				Lower Bound	Upper Bound
(Constant)	1,861	4,898		0,38	0,704	-7,786	11,508
headlength	0,385	0,158	0,151	2,438	0,015	0,074	0,695
chestdepth	0,791	0,199	0,273	3,979	<0.001	0,4	1,183
chestwidth	0,758	0,213	0,23	3,553	<0.001	0,338	1,179

a Dependent Variable: bodyweight

Body weight = 1.861 + 0,385 * Head Length + 0.791 * Chest Depth + 0.758 * Chest Width

What if some of assumptions are violated? What are the possible approaches?

- Bootstrap regression
- Robust regression analysis
- Ridge or Lasso regression
- Regression analysis using factors

Assist Prof. Dr. Doğukan ÖZEN

Ankara University

Faculty of Veterinary Medicine

Department of Biostatistics

ozen@ankara.edu.tr