

5. SPSS procedures for linear regression

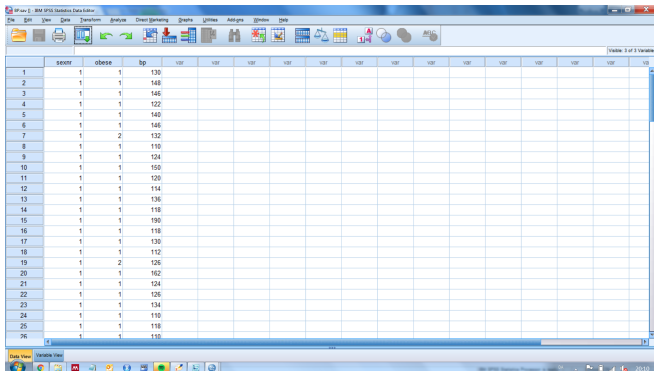
Karl B Christensen

<http://publicifsv.sund.ku.dk/~kach/SPSS>

- Scatter plots
- Correlation
- Simple linear regression
- Residual plots
- Histogram, Probability plot, Box plot
- Data example: obesity score and blood pressure
 - <http://publicifsv.sund.ku.dk/~kach/SPSS/bp.sav>
 - <http://publicifsv.sund.ku.dk/~kach/SPSS/bp.txt>
 - <http://publicifsv.sund.ku.dk/~kach/SPSS/bp.xlsx>

Data example: obesity score and blood pressure

```
GET FILE='p:\BP.sav'.
```

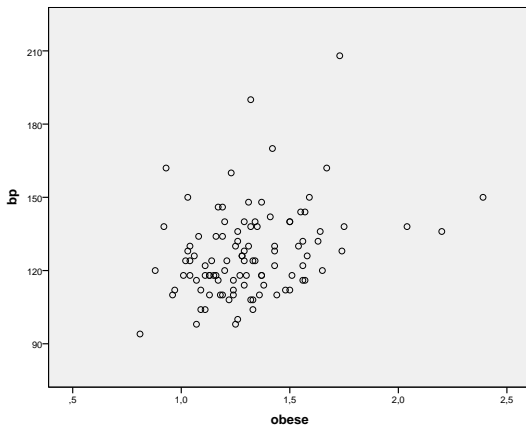


	sevr	obese	bp	V01	V02	V03	V04	V05	V06	V07	V08	V09	V10	V11	V12
1	1	1	130												
2	1	1	148												
3	1	1	146												
4	1	1	122												
5	1	1	140												
6	1	1	146												
7	1	2	132												
8	1	1	110												
9	1	1	124												
10	1	1	150												
11	1	1	120												
12	1	1	114												
13	1	1	136												
14	1	1	118												
15	1	1	190												
16	1	1	118												
17	1	1	130												
18	1	1	112												
19	1	2	126												
20	1	1	162												
21	1	1	124												
22	1	1	126												
23	1	1	134												
24	1	1	110												
25	1	1	118												
26	1	1	110												

Scatter plot

http://publicifsv.sund.ku.dk/~kach/SPSS/F5_gif1.gif

```
GRAPH  
  /SCATTERPLOT(BIVAR)=obese WITH bp  
  /MISSING=LISTWISE.
```

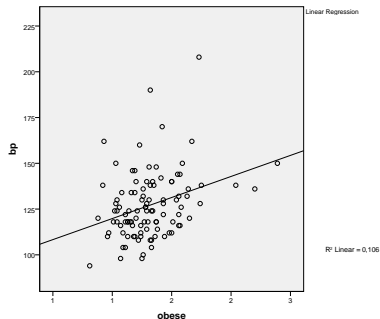


Scatter plot with regression line

Double click in graph to open 'chart editor' - choose

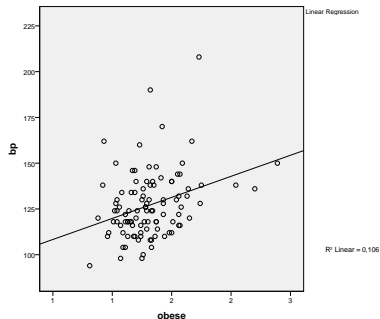
'Elements'/'Fit line at total'.

Point and click to get a scatter plot with a regression line added.



Scatter plot with regression line

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(obese) TYPE=SCALE  
/Y=VAR(bp) TYPE=SCALE  
/FITLINE METHOD=REGRESSION LINEAR  
LINE=total  
/SCATTER COINCIDENT=NONE.
```

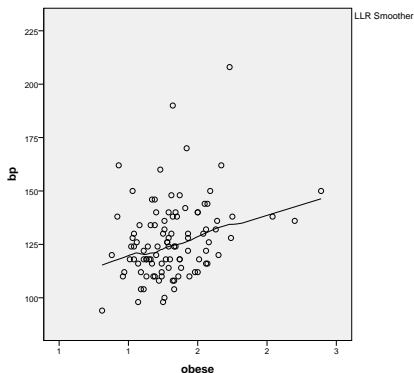


Scatter plot with regression line

- Double-click in graph. Point and click. Cannot 'paste'!
- Write syntax

Scatter plot with smooth curve

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(obese) TYPE=SCALE  
/Y=VAR(bp) TYPE=SCALE  
/COORDINATE=VERTICAL  
/FITLINE METHOD=LLR EPANECHNIKOV BANDWIDTH=CONSTRAINED LINE=total  
/YLENGTH=5.2  
/X1LENGTH=6.5  
/SCATTER COINCIDENT=NONE.
```



Is blood pressure related to obesity ? We compute the correlation

- Default is the parametric correlation, based on the bivariate normal distribution: (the Pearson correlation).
- The Spearman correlation based on ranks is the most commonly used *non parametric* correlation
- The Kendall correlation is an alternative rank correlation based on number of concordant and discordant pairs

Calculated as

$$r = r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Measures the strength of the (linear) association between two variables

- values are between -1 and 1
- 0 corresponds to independence
- -1 and 1 correspond to perfect linearity

http://publicifsv.sund.ku.dk/~kach/SPSS/F5_gif2.gif

tick marks

- Pearson
- Spearman
- Kendalls tau b

```
CORRELATIONS
/VARIABLES=obese bp
/PRINT=TWOTAIL SIG
/MISSING=PAIRWISE.
NONPAR CORR
/VARIABLES=obese bp
/PRINT=BOTH TWOTAIL SIG
/MISSING=PAIRWISE.
```

Computing correlations in SPSS: Output

Correlations

		obese	bp
obese	Pearson Correlation	1	,326
	Sig. (2-tailed)		,001
	N	102	102
bp	Pearson Correlation	,326	1
	Sig. (2-tailed)	,001	
	N	102	102

Correlations

			obese	bp
Kendall's tau_b	obese	Correlation Coefficient	1,000	,213
		Sig. (2-tailed)	.	,002
		N	102	102
	bp	Correlation Coefficient	,213	1,000
		Sig. (2-tailed)	,002	.
		N	102	102
Spearman's rho	obese	Correlation Coefficient	1,000	,304
		Sig. (2-tailed)	.	,002
		N	102	102
	bp	Correlation Coefficient	,304	1,000
		Sig. (2-tailed)	,002	.
		N	102	102

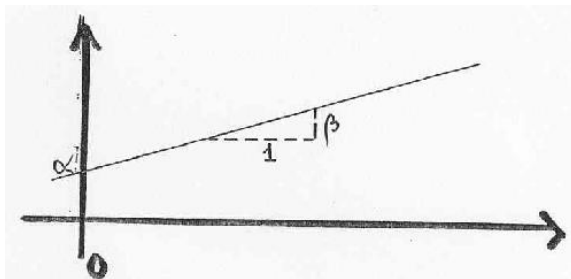
- Y Response variable, outcome variable, dependent variable (blood pressure bp)
- X : Explanatory variable, independent variable, covariate (obesity score obese)

Bivariate observations of X and Y for n individuals

$$(x_i, y_i), \quad i = 1, \dots, n$$

Linear regression

The equation for a straight line: $Y = \alpha + \beta X$



- α : intercept, intersection with Y -axis (at $X=0$)
The expected bp for an individual with obesity score 0 (often an illegal extrapolation).
- β : slope, regression coefficient
Expected difference in bp for two individuals with a difference of one unit in obese.
Often the parameter of interest.

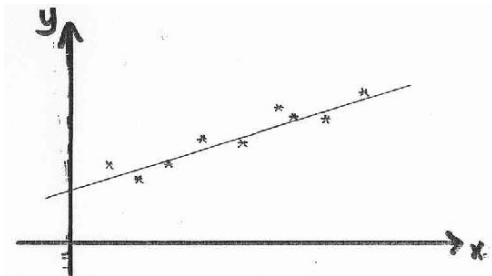
simple linear regression

Model

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \text{ independent}$$

Estimation - least squares: α and β minimizing

$$\sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$



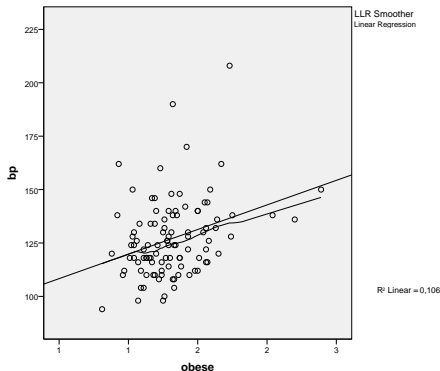
Assumptions in linear regression

- Linearity in the mean value
- Independence between *error terms* ε_i
- Normally distributed error terms, $\varepsilon_i \sim N(0, \sigma^2)$
- Variance homogeneity, that is, identical variances for all ε_i 's

First assumption is the most important. If the assumptions underlying the linear regression model are not met, we cannot trust the results.

Add curve and line to scatter plot

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(obese) TYPE=SCALE  
/Y=VAR(bp) TYPE=SCALE  
/COORDINATE=VERTICAL  
/FITLINE METHOD=LLR EPANECHNIKOV BANDWIDTH=CONSTRAINED LINE=total  
/FITLINE METHOD=REGRESSION LINEAR  
/YLENGTH=5.2  
/X1LENGTH=6.5  
/SCATTER COINCIDENT=NONE.
```



Menu

Analyze/Regression/Linear

```
REGRESSION  
  /MISSING LISTWISE  
  /STATISTICS COEFF OUTS CI(95)  
  /CRITERIA=PIN(.05) POUT(.10)  
  /NOORIGIN  
  /DEPENDENT bp  
  /METHOD=ENTER obese.
```

Linear regression using SPSS: output

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
	B	Std. Error	Beta			Lower Bound	Upper Bound
1 (Constant)	96,818	8,920		10,855	,000	79,122	114,514
obese	23,001	6,667	,326	3,450	,001	9,774	36,229

a. Dependent Variable: bp

Linear regression using SPSS: output

Estimated relation:

$$bp = 96.81793 + 23.00135 \times obese$$

Interpretation: A difference of 1 unit in obesity score corresponds to an expected difference of 23.00135 units in blood pressure.

Variance around the regression line

Variance around the regression line σ^2 is estimated as

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Add ANOVA:

```
REGRESSION  
  /MISSING LISTWISE  
  /STATISTICS COEFF OUTS CI(95) ANOVA  
  /CRITERIA=PIN(.05) POUT(.10)  
  /NOORIGIN  
  /DEPENDENT bp  
  /METHOD=ENTER obese.
```

output

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3552,419	1	3552,419	11,903	,001 ^b
	Residual	29845,541	100	298,455		
	Total	33397,961	101			

a. Dependent Variable: bp

b. Predictors: (Constant), obese

Variance around the regression line

Variance around the regression line σ^2 is estimated as

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

and found in output as Mean Square Residual: value is 298.45541. Standard deviation around the regression line

$$s = \sqrt{s^2} = 17.27586$$

has the same unit as outcome and is easier to interpret.

- SPSS does not report this

Residuals

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\alpha} + \hat{\beta}x_i)$$

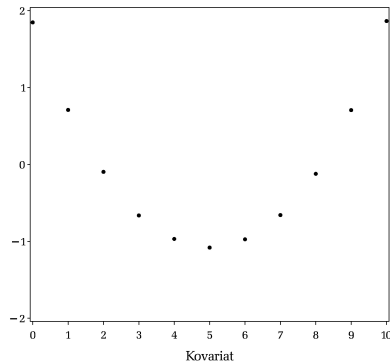
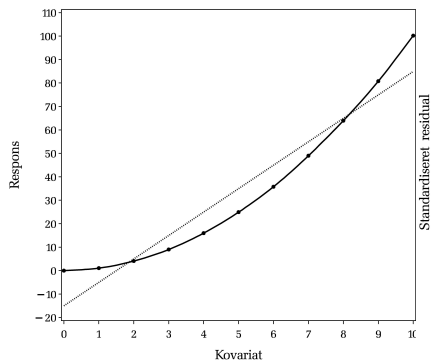
should be plotted against:

- 1 the explanatory variable x_i – to check linearity
- 2 the fitted values \hat{y}_i – to check variance homogeneity (and normality)
- 3 '*normal scores*' i.e. probability plot – to check normality

First two should give impression random scatter, while the probability plot ought to show a straight line.

Residual plots and linearity

Look for \cup or \cap forms



- Ordinary residuals = model deviations

$$\hat{\varepsilon}_i = y_i - \hat{y}_i.$$

- Standardized residuals.

(Residuals where current observation is not used in estimation of corresponding line)

Choose

Analyze/Regression/Linear

and click on Plots to save residuals and generate some plots

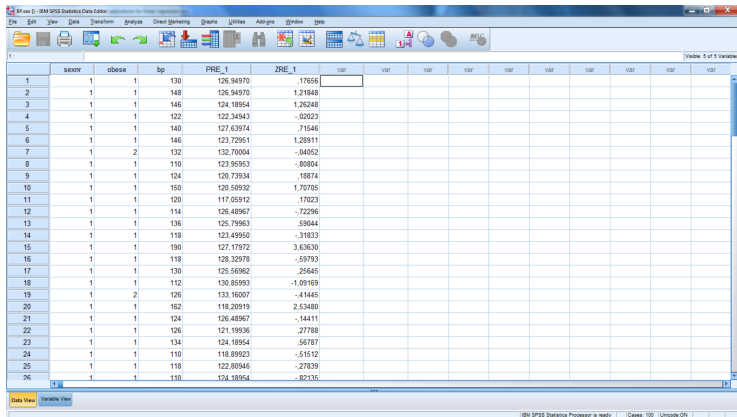
Save residuals and predicted values

Syntax: add /SAVE PRED ZRESID

```
REGRESSION  
  /MISSING LISTWISE  
  /STATISTICS COEFF OUTS CI(95)  
  /CRITERIA=PIN(.05) POUT(.10)  
  /NOORIGIN  
  /DEPENDENT bp  
  /METHOD=ENTER obese  
  /SAVE PRED ZRESID.
```

or click on Save to save residuals.

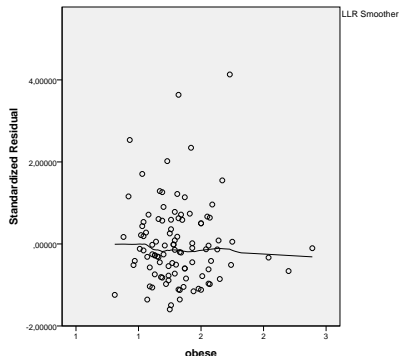
Save residuals and predicted values



	sexrv	obese	bp	PRE_1	ZRE_1	V0F	V0F	V0F	V0F	V0F	V0F	V0F	V0F	V0F	V0F
1	1	1	130	126.94970	1.7656										
2	1	1	148	126.94970	1.21948										
3	1	1	146	124.18954	1.26248										
4	1	1	122	122.34943	-.02023										
5	1	1	140	127.63974	.71546										
6	1	1	146	123.72951	1.28911										
7	1	2	132	132.70004	-.04052										
8	1	1	110	123.95953	-.80804										
9	1	1	124	120.73934	.18874										
10	1	1	150	120.50932	1.70705										
11	1	1	120	117.05912	.17023										
12	1	1	114	126.48967	-.72296										
13	1	1	136	125.79963	.59044										
14	1	1	118	123.49950	-.31833										
15	1	1	190	127.17972	3.63630										
16	1	1	118	128.32978	-.59793										
17	1	1	130	125.56962	.25645										
18	1	1	112	130.85993	-1.09169										
19	1	2	126	133.16007	-.41445										
20	1	1	162	118.20919	2.53480										
21	1	1	124	126.48967	-.14411										
22	1	1	126	121.19936	.27788										
23	1	1	134	124.18954	.56787										
24	1	1	110	118.89923	-.51512										
25	1	1	118	122.80946	-.27839										
26	1	1	130	124.18954	-.82135										

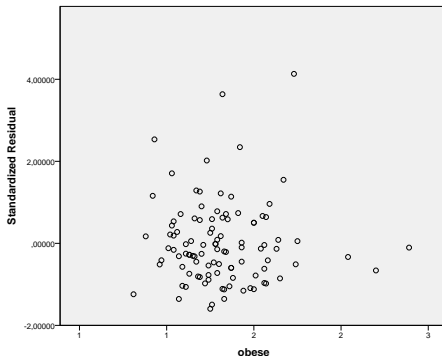
Plot residuals and values of obese

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(obese) TYPE=SCALE  
/Y=VAR(ZRE_1) TYPE=SCALE  
/COORDINATE=VERTICAL  
/FITLINE METHOD=LLR EPANECHNIKOV BANDWIDTH=CONSTRAINED LINE=total  
/YLENGTH=5.2  
/X1LENGTH=6.5  
/SCATTER COINCIDENT=NONE.
```



Plot residuals and values of obese

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(obese) TYPE=SCALE  
/Y=VAR(ZRE_1) TYPE=SCALE  
/COORDINATE=VERTICAL  
/YLENGTH=5.2  
/X1LENGTH=6.5  
/SCATTER COINCIDENT=NONE.
```

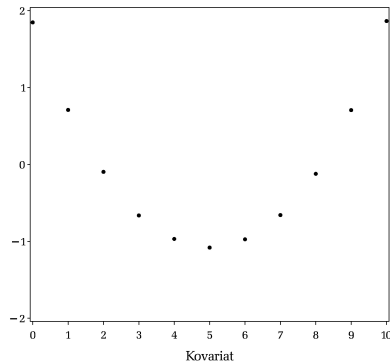
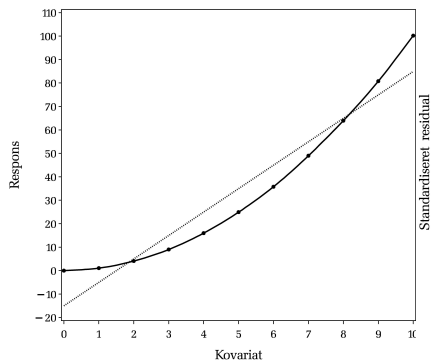


Plot residuals and values of obese

- plot with or without smooth curve
- evaluate if it looks like random scatter
- look for non-linearity
- not standard to include smooth curve
 - but helpful

Residual plots and linearity

Look for \cup or \cap forms

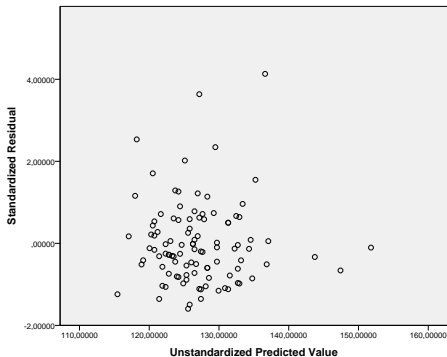


Plot residuals and values of obese

- plot with or without smooth curve
- evaluate if it looks like random scatter
- look for non-linearity
- not standard to include smooth curve
 - but helpful
- no evidence of non-linearity

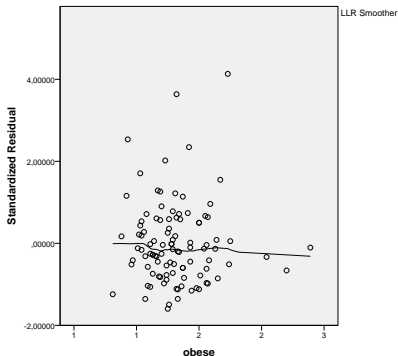
Plot residuals and predicted values

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(PRE_1) TYPE=SCALE  
/Y=VAR(ZRE_1) TYPE=SCALE  
/COORDINATE=VERTICAL  
/YLENGTH=5.2  
/X1LENGTH=6.5  
/SCATTER COINCIDENT=NONE.
```



Plot residuals and predicted values

```
IGRAPH  
/VIEWNAME='Scatterplot'  
/X1=VAR(PRE_1) TYPE=SCALE  
/Y=VAR(ZRE_1) TYPE=SCALE  
/LINE  
/COORDINATE=VERTICAL  
/FITLINE METHOD=REGRESSION LINEAR  
/SCATTER COINCIDENT=NONE.
```



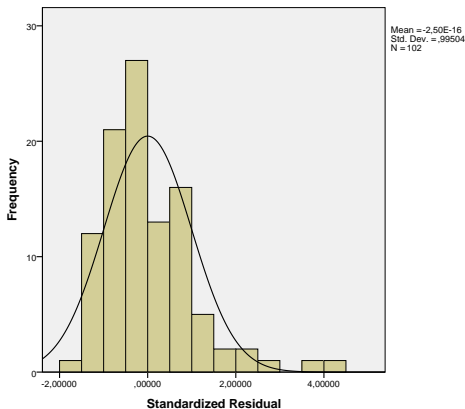
Plot residuals and predicted values

- Look for 'trumpet shape'
- evident here
- larger variance for large bp values
- variance increases

(smooth curve does not really help)

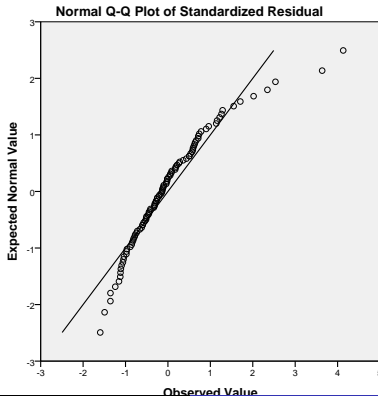
Plot histogram for the residuals

```
GRAPH  
/HISTOGRAM(NORMAL)=ZRE_1.
```



Plot probability plot for the residuals

```
/VARIABLES=ZRE_1  
/NOLOG  
/NOSTANDARDIZE  
/TYPE=Q-Q  
/FRACTION=BLOM  
/TIES=MEAN  
/DIST=NORMAL.
```



Histogram and probability plot for the residuals

The residuals do not have a normal distribution. Transformation

$$y \mapsto \log(y)$$

may be needed. (Transform - Compute Variable..)

```
COMPUTE lpb=LN(bp).  
EXECUTE.
```

Exercise: Regression and graphics I

In this exercise we want to study the effect of age on the SIGF1-level in the Juul data.

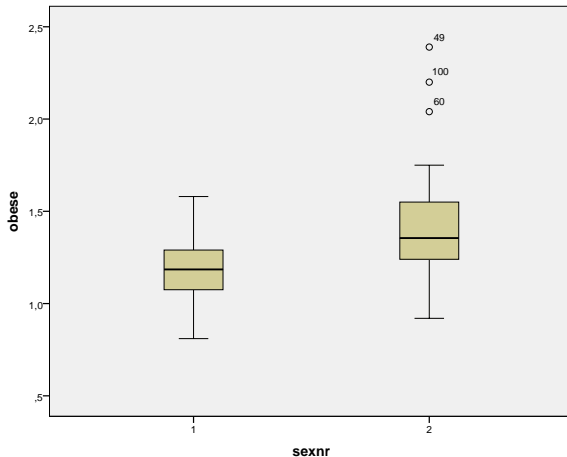
- ① Create a new data set containing only prepubertal children (Tanner stage 1 and age > 5).
- ② Plot the relationship between *SIGF1* and age for prepubertal children. Add regression lines or smooth curves.
- ③ Do a regression analysis of *SIGF1* vs. age for prepubertal children.
- ④ Make a data set with residuals and use these to evaluate model fit.
 - Histogram of residuals - normal distribution ?
 - Scatter plot of (residuals, expected sigf1) - random scatter ?
 - Scatter plot of (residuals, age) -random scatter ?
- ⑤ If the assumptions are not met. Try transforming sigf1

Two different ways:

Graphs - Legacy Dialogs - Boxplot

```
/PLOT=BOXPLOT  
/STATISTICS=NONE  
/NOTOTAL.
```

Box plots



Exercise: Regression and graphics II

- 1 Download the bp.sav data set from the home page and get it into SPSS.
- 2 Use graphical methods to answer the questions
 - Is there an association between obese and sexnr ?
 - Is there an association between bp and sexnr ?
 - Is there an association between obese and bp ?
- 3 Click and point to get a scatter plot. Choose

Set Markers By:

or

```
GRAPH  
  /SCATTERPLOT(BIVAR)=obese WITH bp BY sexnr  
  /MISSING=LISTWISE.
```

to get different plotting symbols for the two genders. Double click in graph choose Elements - Fit line at Subgroups