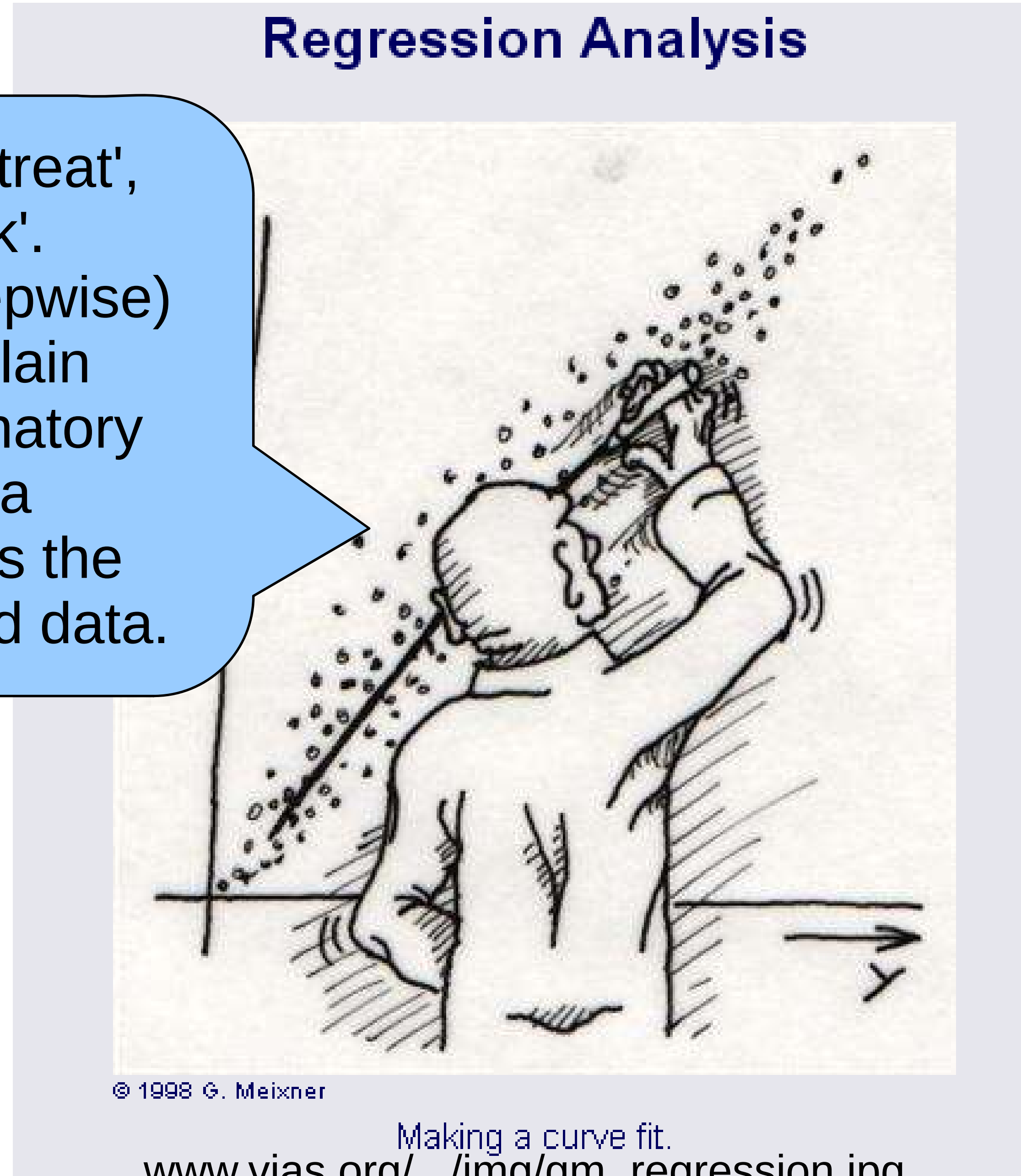
Chapter 5: Regression

'Regression' (latin) means 'retreat', 'going back to', 'stepping back'.
In a 'regression' we try to (stepwise) retreat from our data and explain them with one or more explanatory predictor variables. We draw a 'regression line' that serves as the (linear) model of our observed data.



www.vias.org/.../img/gm_regression.jpg

Correlation vs. regression

- Correlation
- In a correlation, we look at the relationship between two variables without knowing the direction of causality

- Regression
- In a regression, we try to predict the outcome of one variable from one or more predictor variables. Thus, the direction of causality can be established.
- 1 predictor=simple regression
- >1 predictor=multiple regression

Correlation vs. regression

Correlation

For a correlation you do not need to know anything about the possible relation between the two variables

Many variables correlate with each other for unknown reasons

Correlation underlies regression but is descriptive only

Regression

For a regression you do want to find out about those relations between variables, in particular, whether one 'causes' the other.

Therefore, an unambiguous causal template has to be established between the causer and the causee before the analysis!

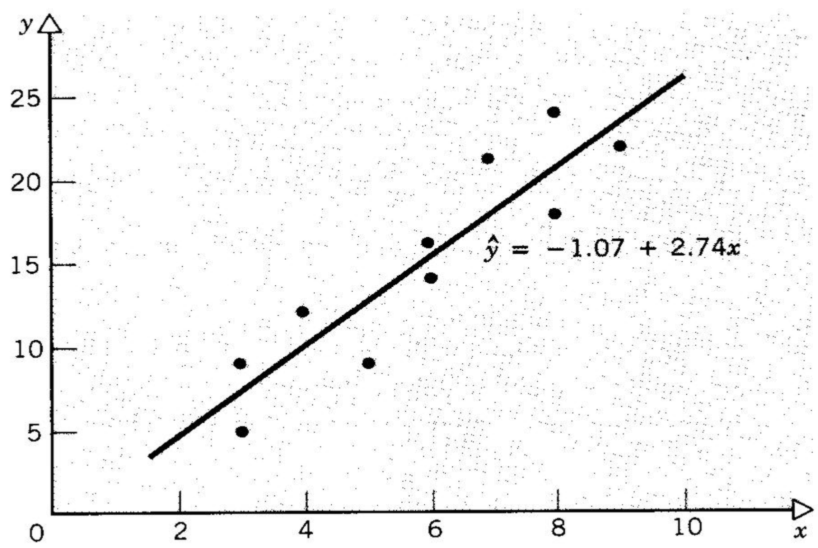
This template is inferential.

Regression is THE statistical method underlying ALL inferential statistics (t-test, ANOVA, etc.). All that follows is a variation of regression.

Linear regression
Independent and dependent variables

In a regression, the predictor variables are labelled **'independent' variables**. They predict the outcome variable labelled '**dependent' variable.**

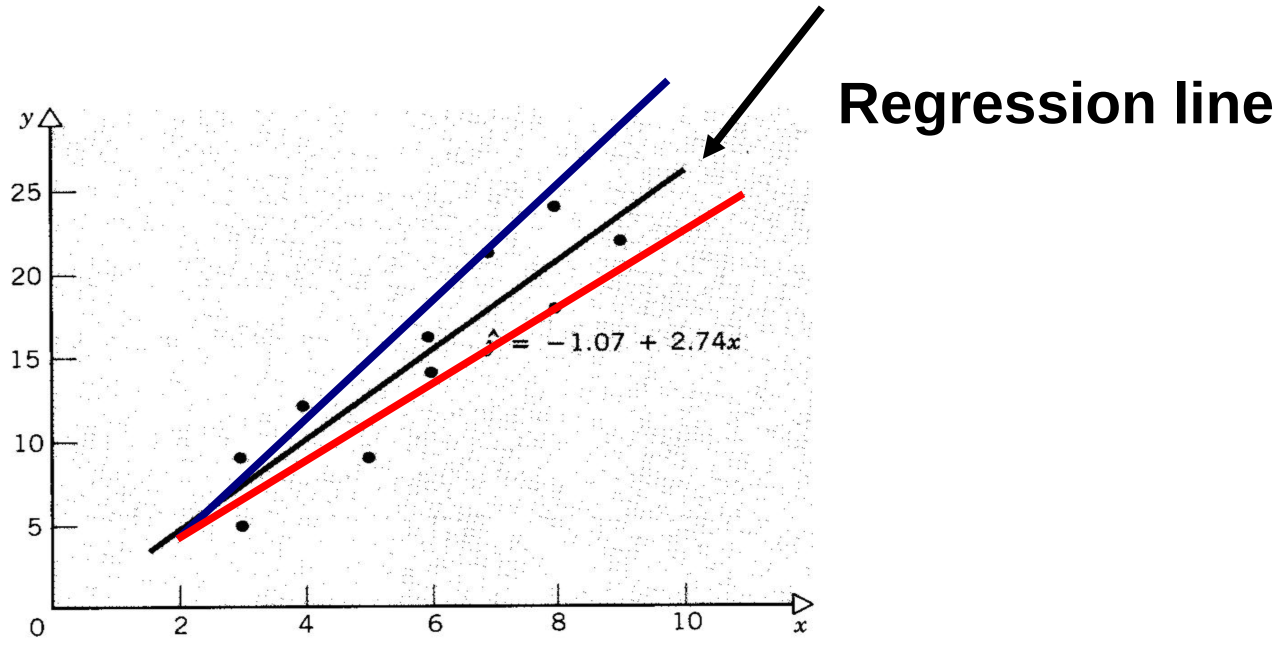
A regression in SPSS is always a **linear** regression, i.e., a **straight line** represents the data as a **model**.



http://snobear.colorado.edu/Markw/SnowHydro/ERAN/regression.jp

Method of least squares

In order to know which line to choose as the best model of a given data cloud, the method of least squares is used. We select the line for which the sum of all squared deviations (SS) of all data points is lowest. This line is labelled '**line of best fit**', or '**regression line**'.



Simple regression
Regression coefficients

In mathematics, a **coefficient** is a constant multiplicative factor of a certain object. For example, the coefficient in $9x^2$ is 9.
http://en.wikipedia.org/wiki/Coefficient

The linear regression equation (5.2) is:

$$Y_i = (b_0 + b_1 X_i) + \varepsilon_i$$

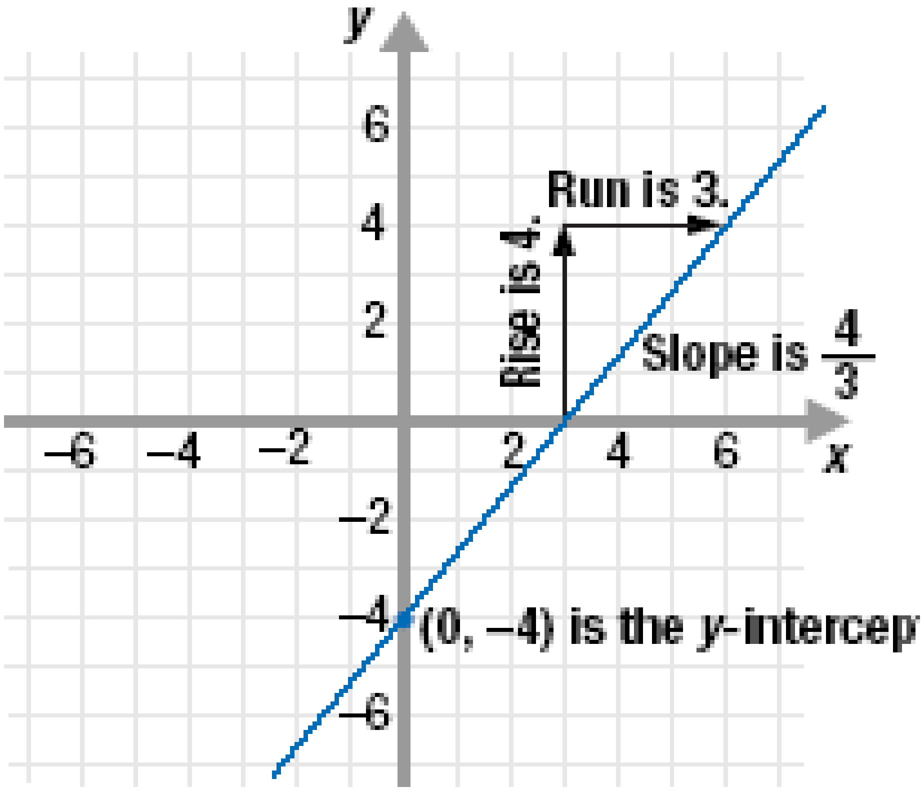
Y_i = outcome we want to predict

$$\left.\begin{array}{l} b_0 = \text{intercept of the regression line} \\ b_1 = \text{slope of the regression line} \end{array}\right\} \text{regression coefficients}$$

X_i = Score of subject$_i$ on the predictor variable

ε_i = residual term, error

Slope/gradient and intercept



- Slope/gradient: steepness of the line; neg or pos
- Intercept: where the line crosses the y-axis

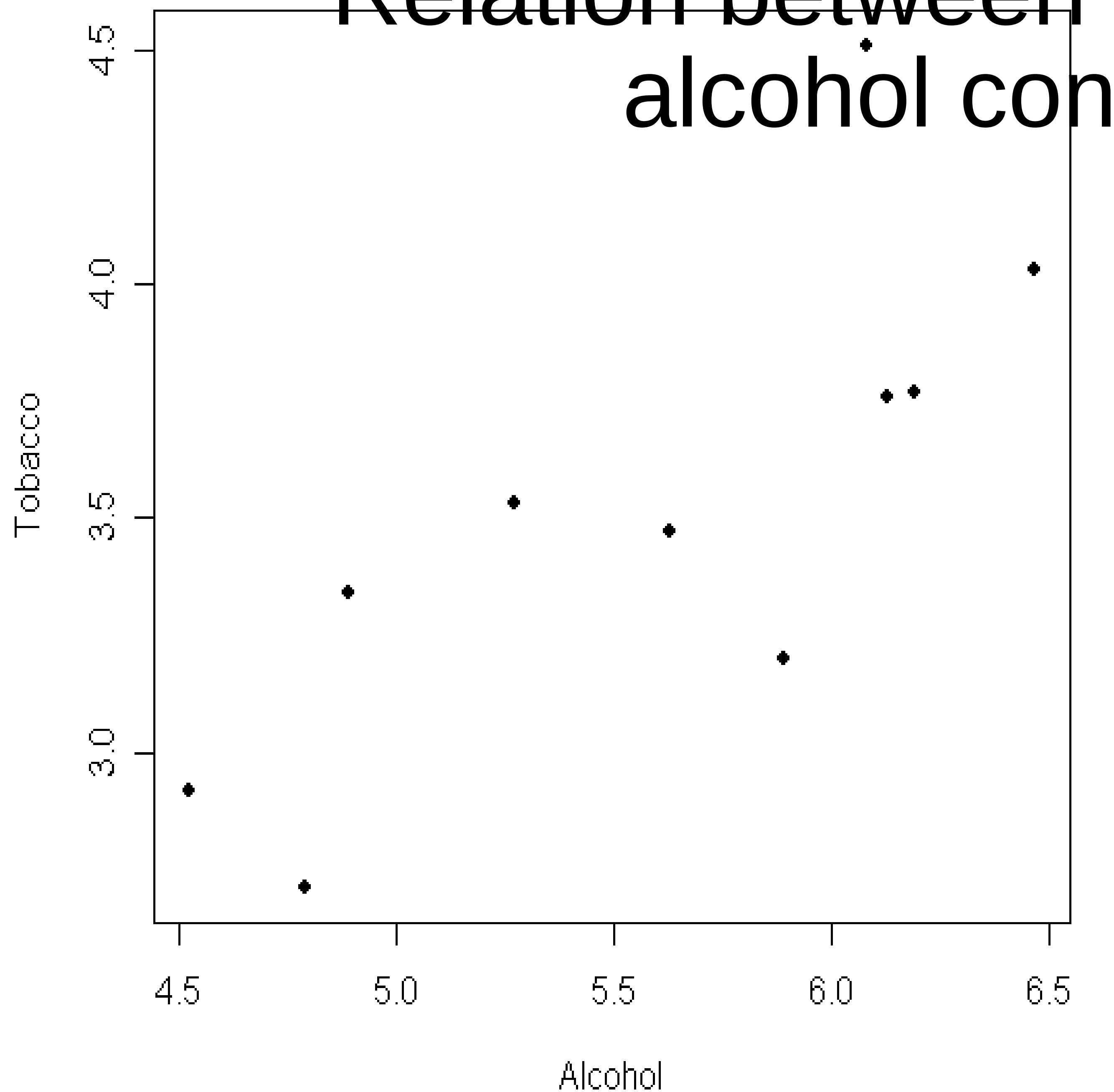
$$\color{blue}{Y_i} = (-\,4 + 1.33\color{green}{X_i}) + \color{purple}{\varepsilon_i}$$

http://algebra-tutoring.com/slope-intercept-form-equation-lines-1-gifs/slope-52.gif

'goodness-of-fit'

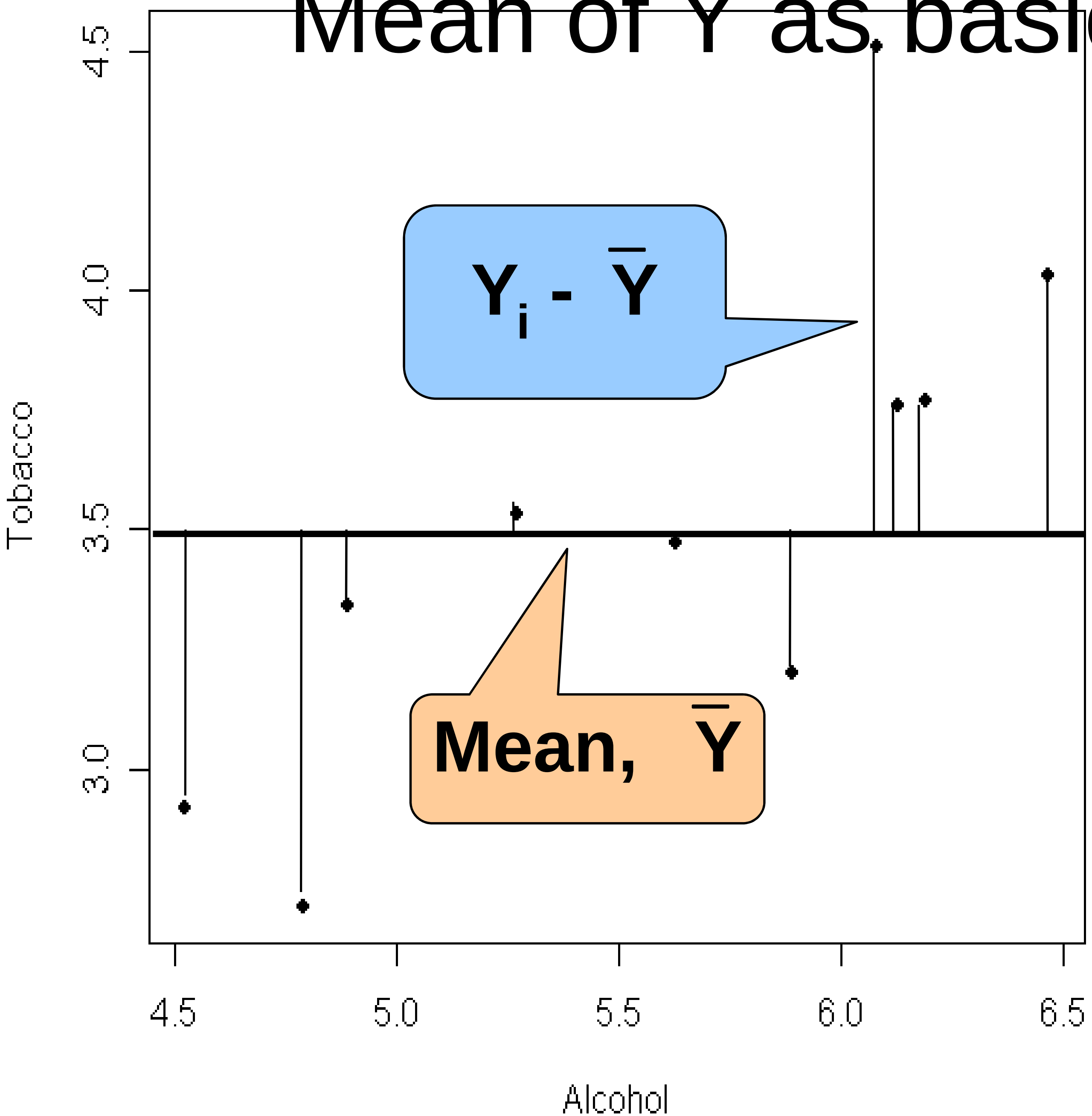
The line of best fit (regression line) is compared with the most basic model. The former should be significantly better than the latter. The most basic model is the mean of the data.

Relation between tobacco and alcohol consume



http://images.google.de/imgres?imgurl=http://math.uprm.edu/~wrolke/esma3102/graphs/rssfig2.png&imgrefurl=http://math.uprm.edu/~wrolke/esma3102/rss.htm&h=552&w=553&sz=4&hl=de&start=23&tbnid=eY0TWAtPXf0_ZM:&tbnh=133&tbnw=133&prev=/images%3Fq%3Dsum%2Bof%2Bsquares%26start%3D21%26svnum%3D10%26hl%3Dde%26lr%3D%26sa%3DN

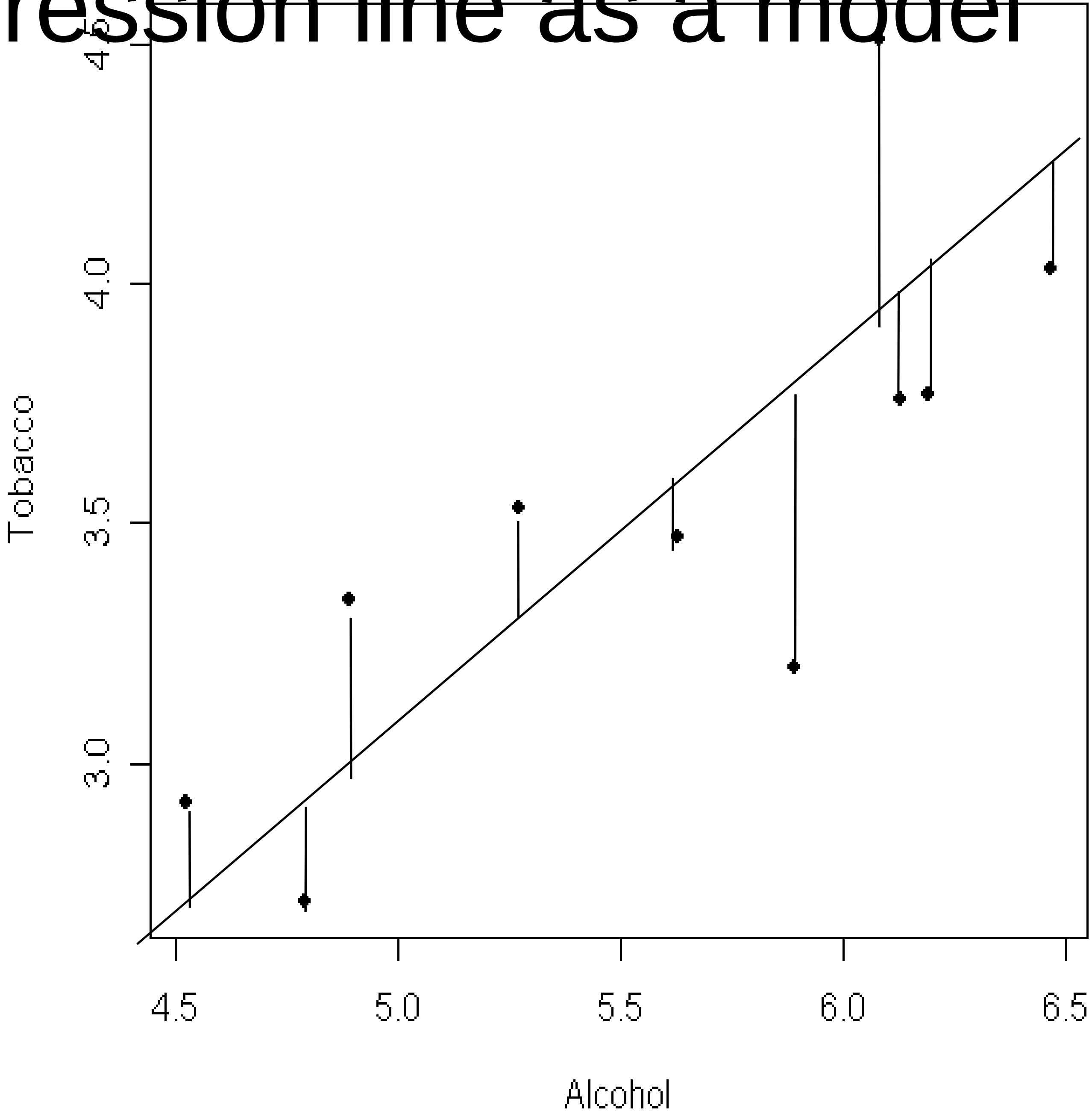
Mean of Y as basic model



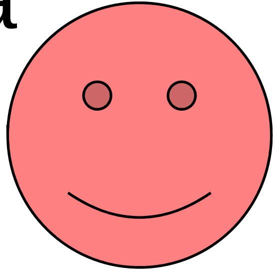
The summed squared differences between observed values and the mean, SST, are big, hence the mean is not a good model of the data

Sum of squares total: SS_T

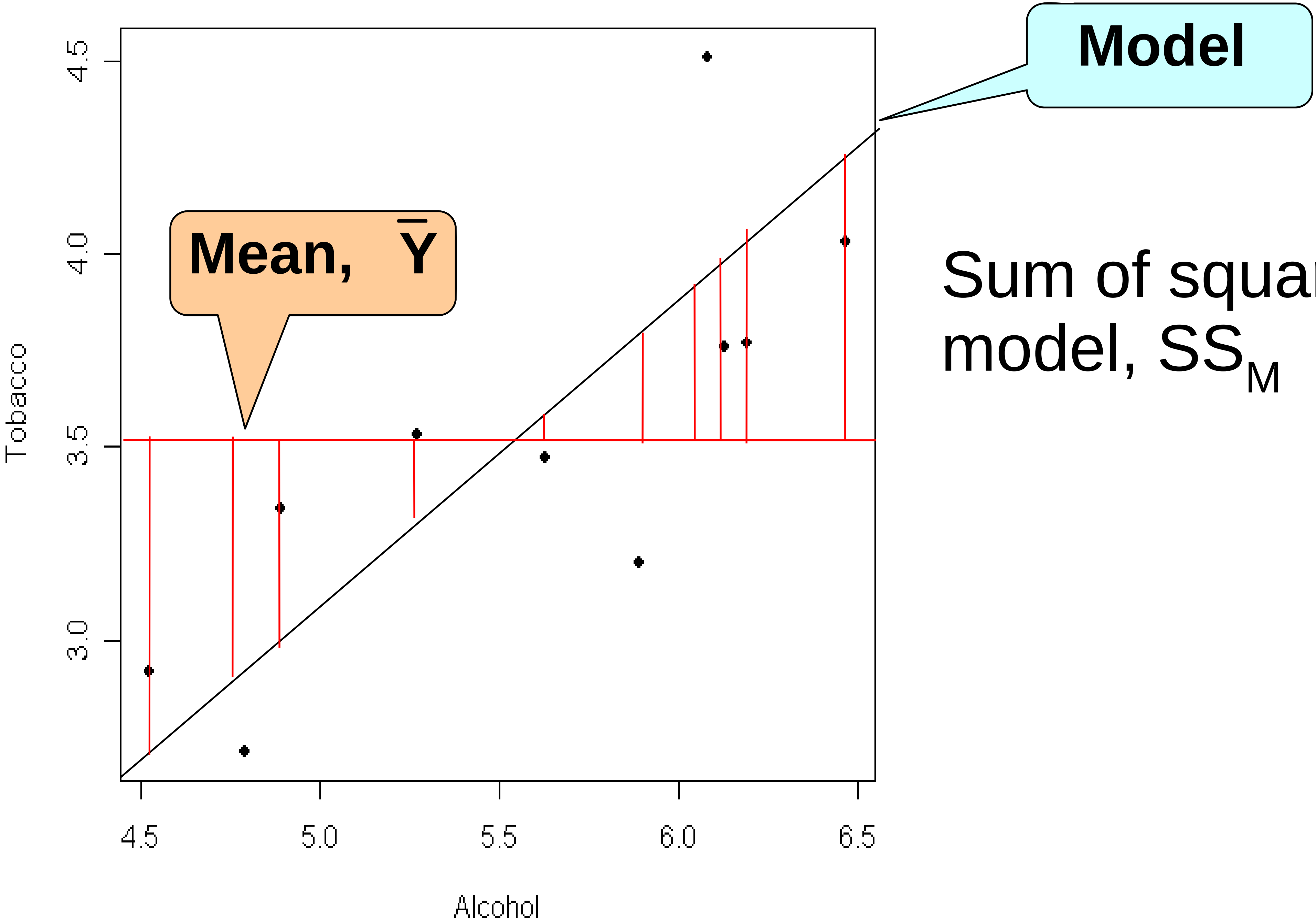
Regression line as a model



The summed squared differences between observed values and the regression line, SS_R, are smaller, hence this regression line is a much better model of the data



sum of squares residual SS_R



Sum of squares model, SS_M

SS_M: sum of squared differences between the mean of Y and the regresion line (as our model)

Comparing the basic model and the regression model: R^2

The improvement by the regression model can be expressed by dividing the sum of squares of the regression model SS_M by the sum of squares of the basic model SS_T:

$$R^2 = \frac{SS_M}{SS_T}$$

The basic comparison in statistics is always to compare the amount of variance that our model can explain with the total amount of variation there is. If the model is good it can explain a significant proportion of this overall variance.

This is the same measure as the R^2 in chapter 4 on correlation. Take the square root of R^2 and you have the Pearson correlation coefficient r!

Comparing the basic model and the regression model: F-Test

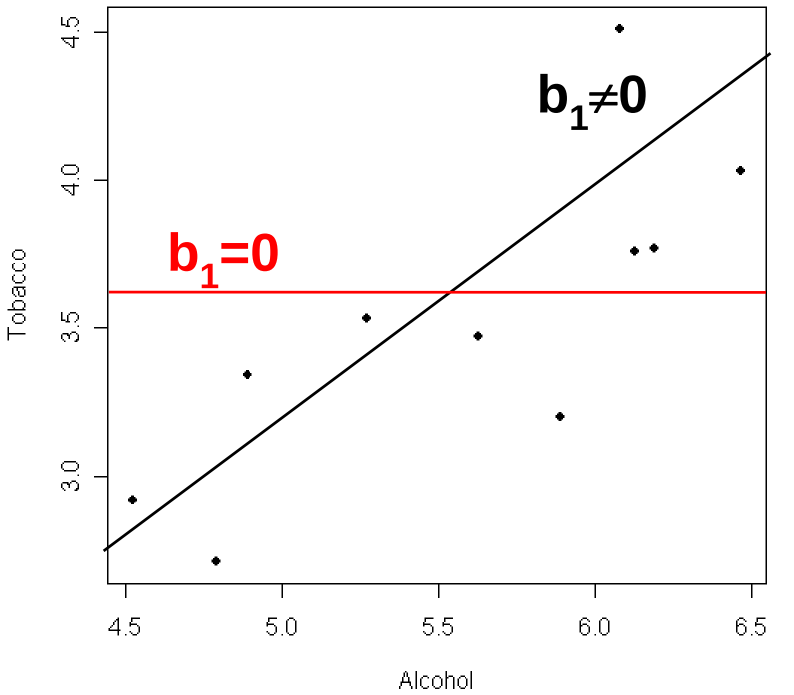
In the F-Test, the ratio of the improvement due to the model SS_M and the difference between the model and the observed data, SS_R, is calculated.
We take the mean sum of squares, or mean squares, MS, for the model, MS_M, and the observed data, MS_R:

$$F = \frac{MS_M}{MS_R}$$

The F-ratio should be high (since the model should have improved the prediction considerably, as expressed in MS_M). MS_R, the difference between the model and the observed data (the residual), should be small.

The coefficient of a predictor

The coefficient of the predictor X is b_1. B_1 indicates the gradient/slope of the regression line. It says how much Y changes when X is changed one unit. In a good model, b_1 should always be different from 0, since the slope is either positive or negative.
Only a bad model, i.e., the basic model of the mean, has a slope of 0.



If $b_1=0$, this means:

- A change in one unit of the predictor X does not change the predicted variable Y
- The gradient of the regression line is 0.

T-Test of the coefficient of the predictor

A good predictor variable should have a b1 that is different from 0 (the regression coefficient of the basic model, the mean). Whether this difference is significant, can be tested by a *t*-test.
The b of the expected values (0-Hypothesis, i.e., 0) is subtracted from the b of the observed values and divided by the standard error of b.

$t = b_{observed} - b_{expected}$ ____Since $b_{expeted}$=0

____

________ SE_b

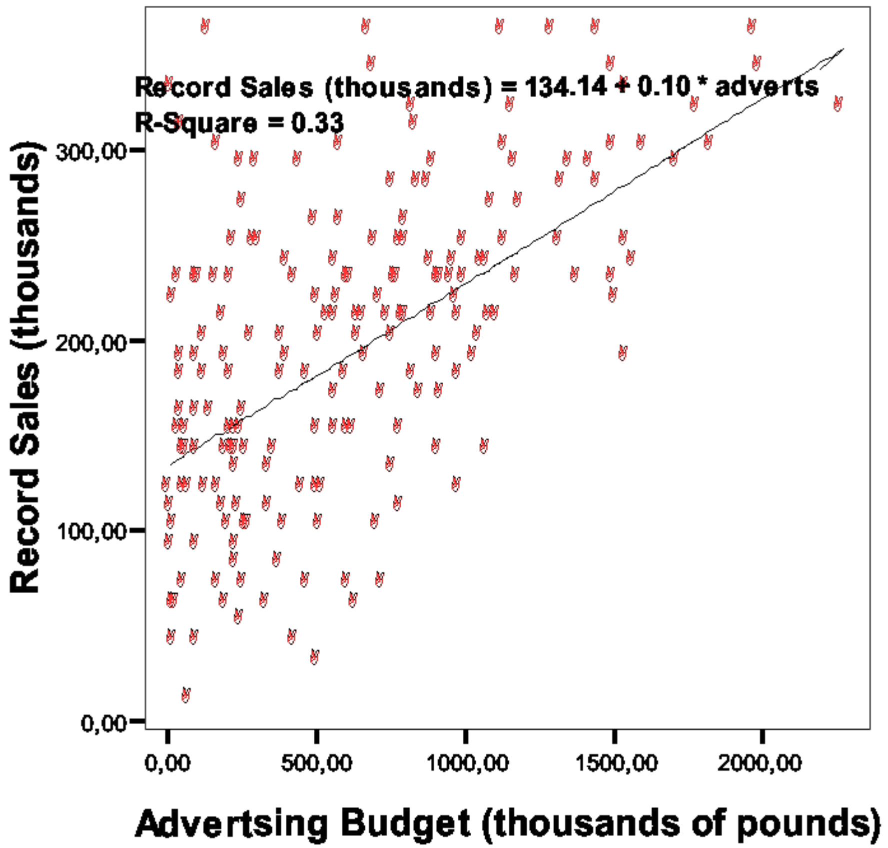
t= $b_{observed}$ _______ *t* should be * different from 0.

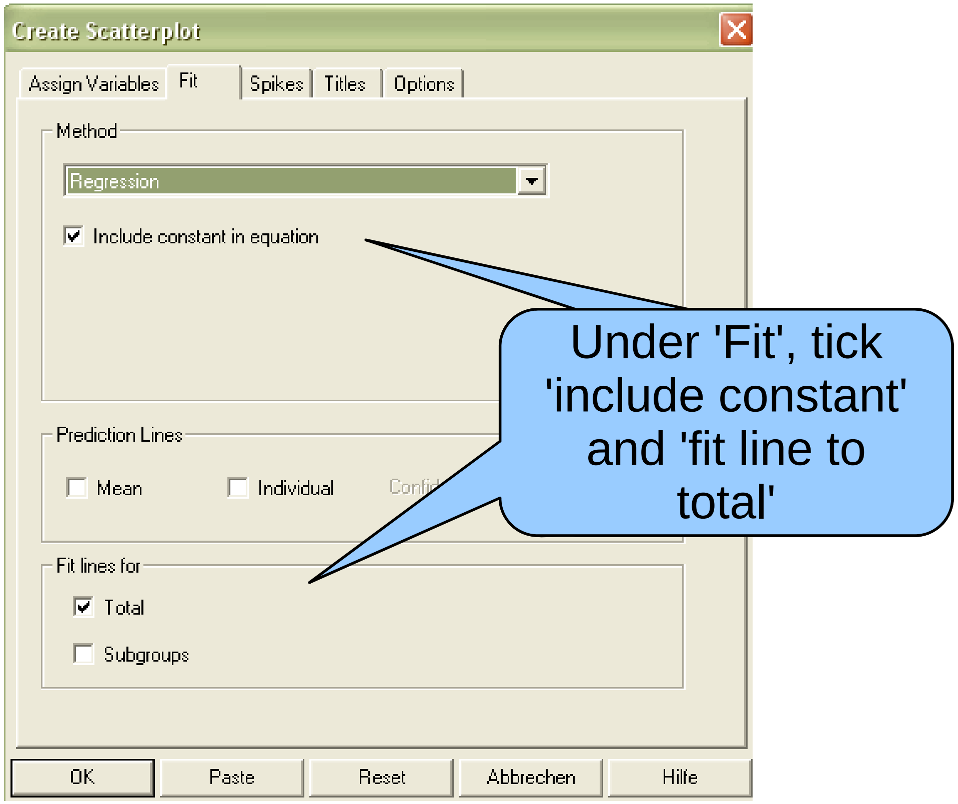
SE_b

Simple regression on SPSS
(using the Record1.sav data)

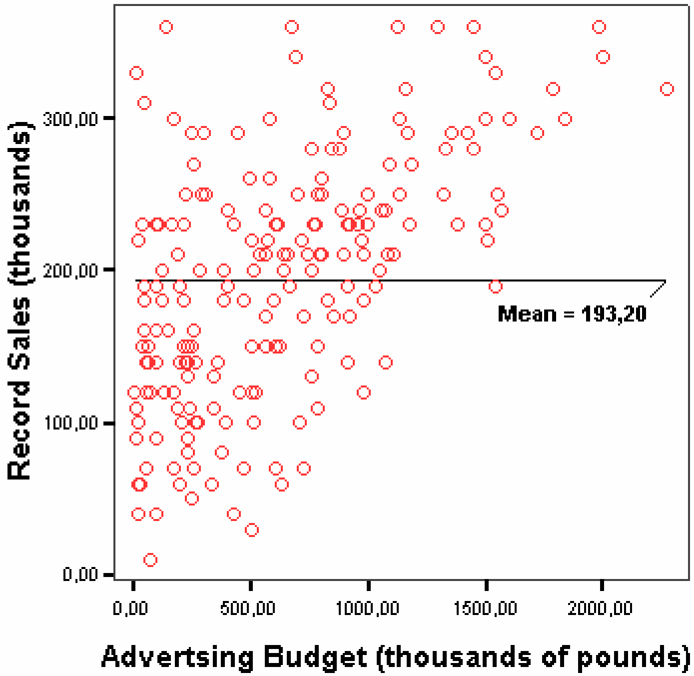
Descriptive glance: Scatterplot of the correlation between advertisement and record sales

Graphs --> Interactive --> Scatterplot

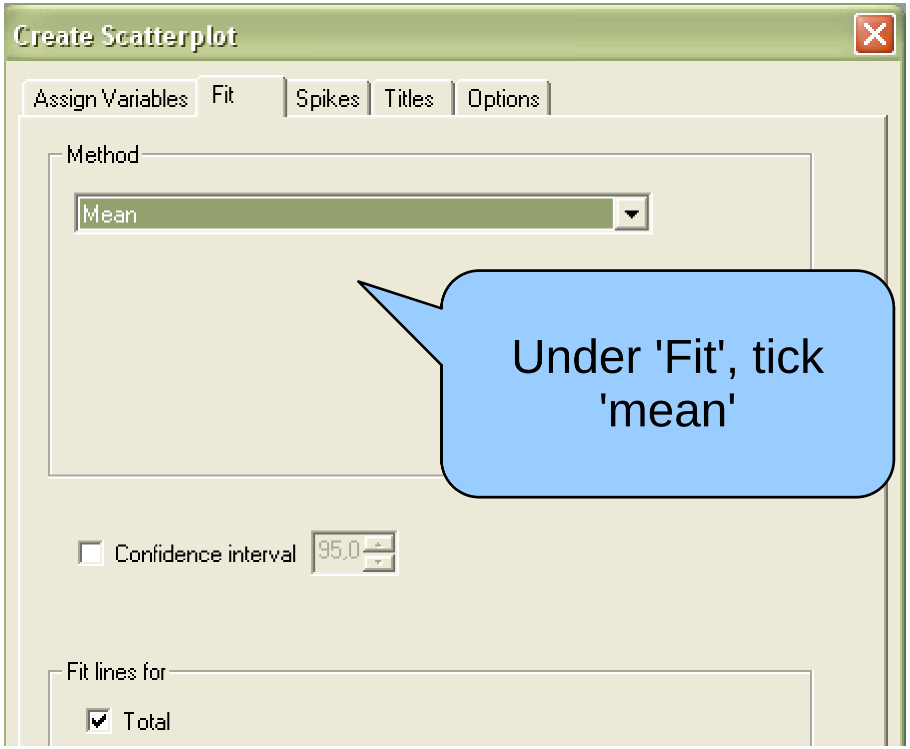


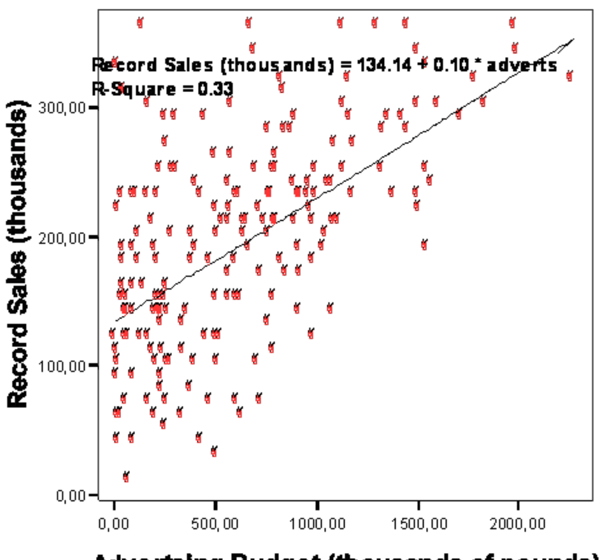


Comparing the mean and the regression model
(using the Record1.sav data)



Graphs --> Interactive --> Scatterplot





Linear Regression

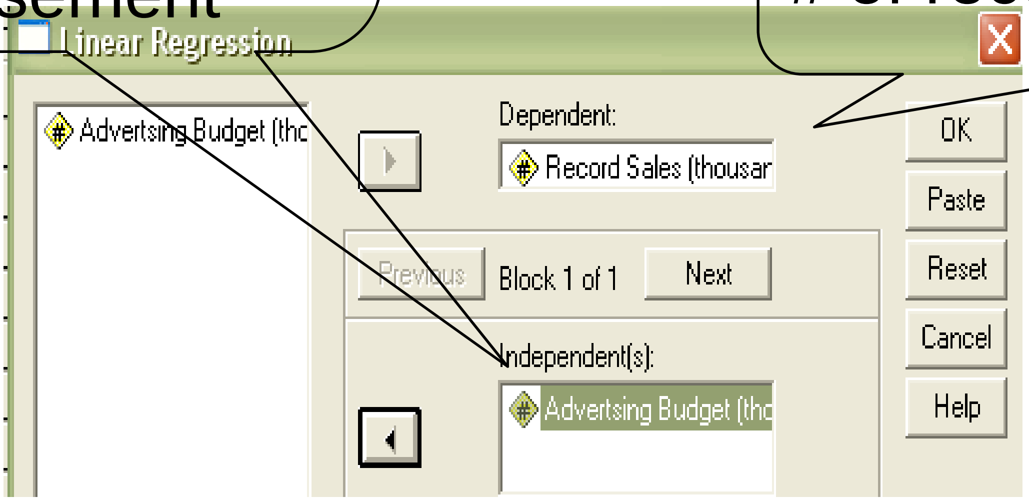
--> The regression line is quite different from the mean

Simple regression on SPSS
(using the Record1.sav data)

Analyze --> Regression --> Linear

Predictor:
How much money
(in 1000)
you spend on
advertisement

What you want to predict:
# of records (in 1000) sold



Output of simple regression on SPSS
(using the Record1.sav data)

Analyze --> Regress --> Linear

R is the simple Pearson correlation between 'advertisement' and 'records sold'

R^2 is the amount of explained variance

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,578[a]	,335	,331	65,9914

a. Predictors: (Constant), ADVERTS Advertsing Budget (thousands of pounds)

R^2= 33% of the total variance can be explained by the predictor 'advertisement'.

66% of the variance cannot be explained.

ANOVA for the SS_M (F-test): advertisement predicts sales significantly

SS_M

SS_R

MS_M

$F = MS_M/MS_R$

$= 433687{,}833/4354{,}87$

$= 99{,}587$

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687,8	1	433687,833	99,587	,000[a]
	Residual	862264,2	198	4354,870		
	Total	1295952	199			

a. Predictors: (Constant), ADVERTS Advertising Budget (thousands of pounds)

b. Dependent Variable: SALES Record Sales (thousands)

SS_T
sum of squares total

MS_R

Regression coefficients b0, b1

b0 intercept
where regres-
sion line
crosses Y axis
When no money
is spent (X=0),
134,140 records are
sold

b1 gradient
If predictor X
is increased
by 1 unit (1000, then
96,12 extra
records will
be sold

t= B/SE_B
134,14/7,537=
17,799

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134,140	7,537		17,799	,000
	ADVERTS Advertising Budget (thousands of pounds)	9,612E-02	,010	,578	9,979	,000

a. Dependent Variable: SALES Record Sales (thousands)

=.09612

A closer look at the t-values

Coefficients[a]

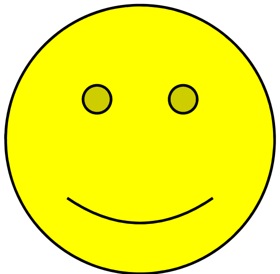
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134,140	7,537		17,799	,000
	ADVERTS Advertsing Budget (thousands of pounds)	9,612E-02	,010	,578	9,979	,000

a. Dependent Variable: SALES Record Sales (thousands)

The equation for computing the t-value is **$t = B/SE_B$**

For the constant: 134,14/7,537=17,799
For ADVERTS: B=0.09612/.010 should result in 9.612, however, t= 9.979

What's wrong? Nothing, this is a rounding error. If you double-click on the output table "Coefficients", a more exact number will be shown:
9.612E-02 = 0,09612448597388
.010 = 0,0096236621523
If you re-compute the equation with these numbers, the result is correct:
0,09612448597388/ 0,0096236621523 = 9.979

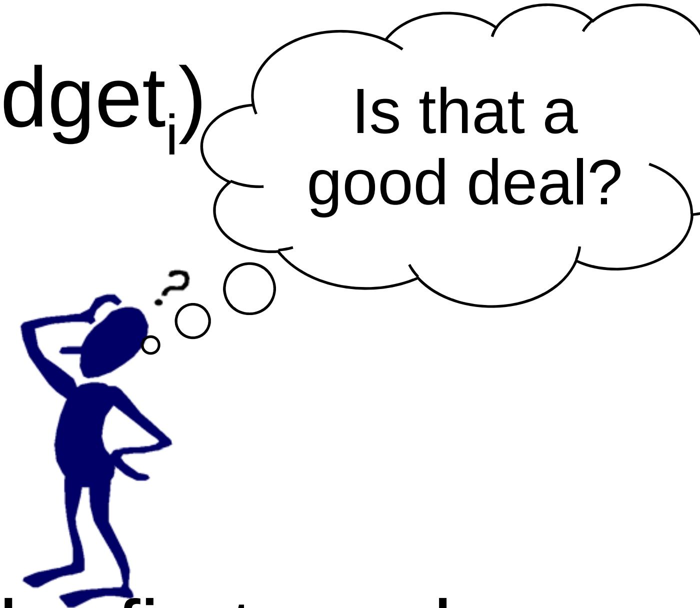


Using the model for Prediction

Imagine the record company wants to spend 100,000 £ for advertisement.
Using Equation 5.2, we can fit in the values of b0 and b1:

$$\mathbf{Y_i = (b_0 + b_1 X_i)}$$

$= 134.14 + (.09612 \times \text{Advertising Budget}_i)$



Expl: If 100,000 £ are spent on ads,

$134.14 + (.09612 \times 100) = 143.75$

144,000 records should be sold on the first week.

http://image.informatik.htw-aalen.de/Thierauf/Knobelaufgaben/Sommer03/zweifel.png

Multiple regression

In a multiple regression, we predict the outcome of a dependent variable Y by a linear combination of >1 independent predictor variables X_i

$\text{Outcome}_i = (\text{Model}_i) + \text{error}_i$

Every variable has its own coefficient: b_1, b_2,...,b_n

(5.9) $$\mathbf{Y_i = (b_0 + b_1X_1 + b_2X_2 + \ldots + b_nX_n) + \varepsilon_i}$$

b_1X_1 = 1st predictor variable with its coefficient
b_2X_2 = 2nd predictor variable with its coefficient, etc.
ε_i = residual term

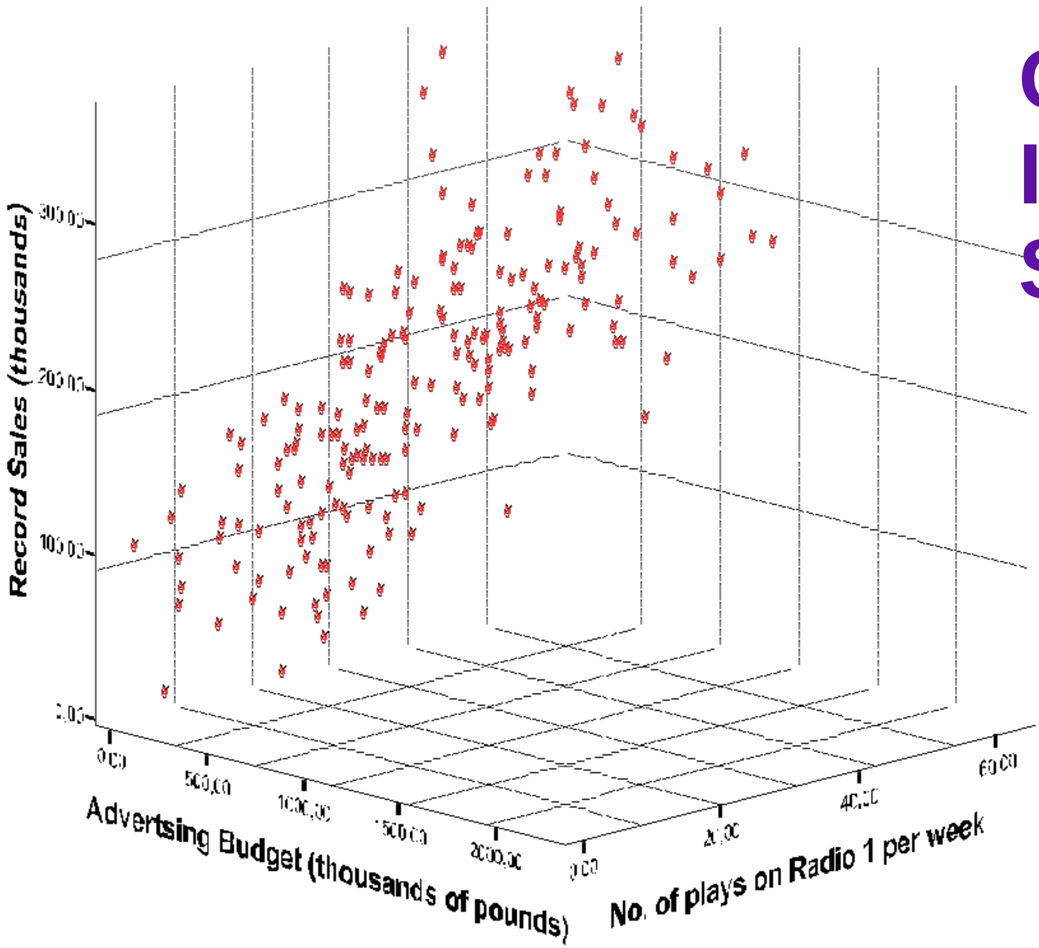
Multiple Regression on SPSS
using file record2.sav

We want to predict record sales (Y) by two predictors:
X1 = advertisement budget
X2 = number of plays on Radio 1

$$\textbf{Record Sales}_i = b_0 + b_1\textbf{Ad}_i + b_2\textbf{Play}_i + \varepsilon_i$$

Instead of a regression line, a regression plane (2 dimensions) is now fitted to the data (3 dimensions)

3D-Scatterplot of the relation between record sale (Y) and advertisement budget (X1) No of plays on Radio 1/week (X2)

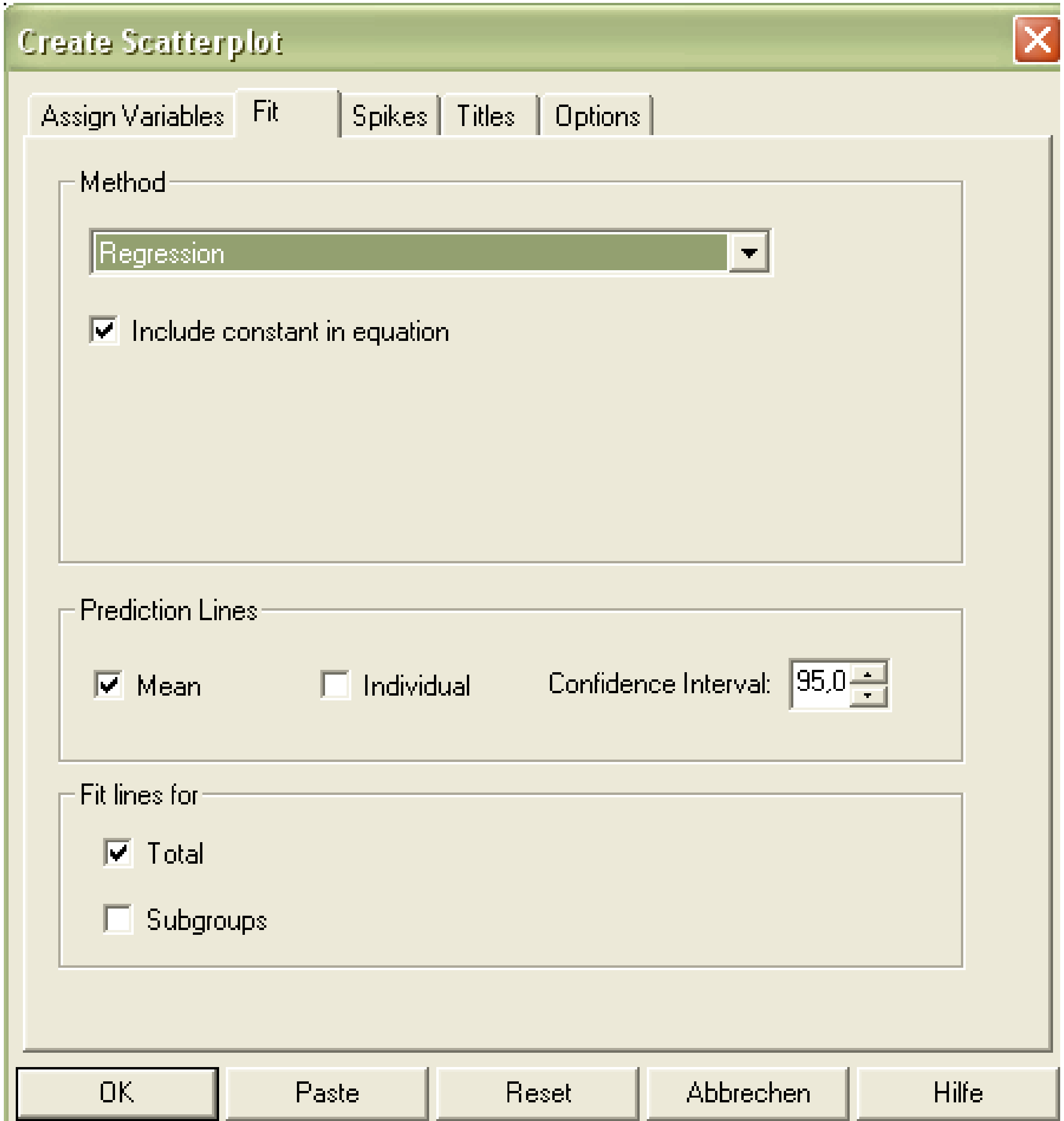


Graphs --> Interactive --> Scatterplot --> 3D

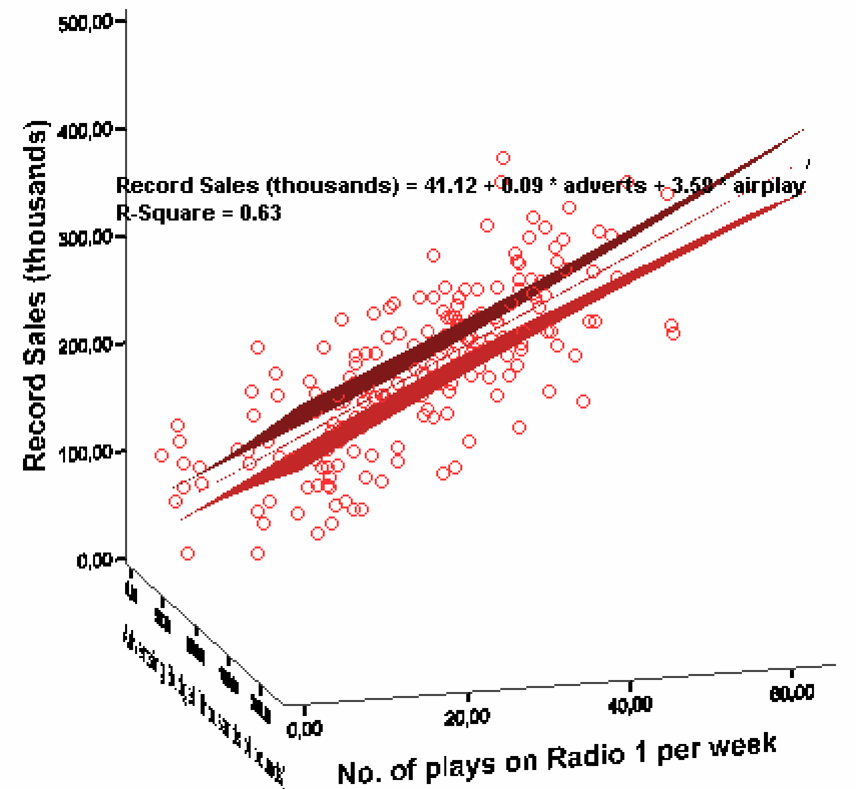
Multiple regression with 2 Variables can be visualized as a 3D-scatterplot. More variables cannot be accomodated visually.

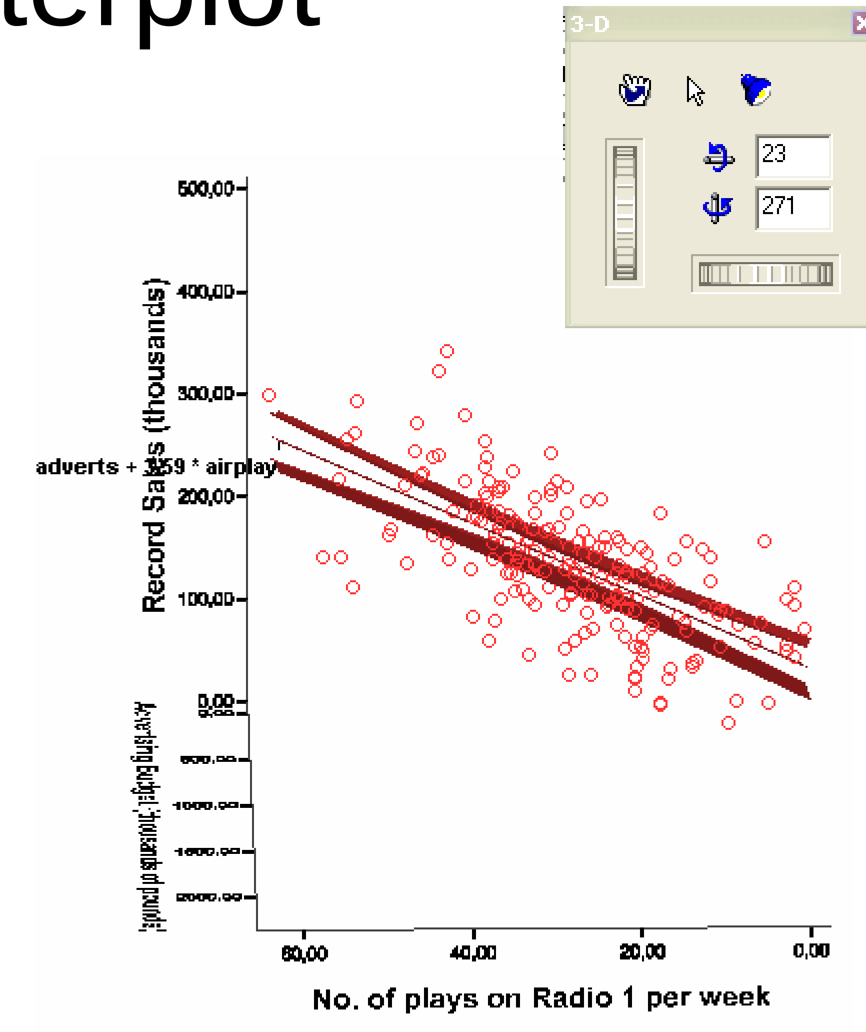
Regression planes and confidence intervals of multiple regression

Under the menu 'Fit', specify the following options



3-D-scatterplot





If adjusted appropriately, you can see the regression plain and the confidence plains almost like lines

The regression plains are chosen as to cover most of the data points in the three-dimensional data cloud

Sum of squares, R, R^2

The terms we encountered for simple regression, $\mathbf{SS_T}$, $\mathbf{SS_R}$, $\mathbf{SS_M}$, still mean the same, but are more complicated to compute now.

Instead of the simple correlational coefficient R, we use a multiple correlation coefficient **Multiple R**.

Multiple R is the correlation between the predicted and observed values of the outcome. As in simple R, Multiple R, should be great.
Multiple R^2 is a measure of the explained variance of Y by the predictor variables X_1-X_n.

Methods of regression

The predictors of the model should be selected carefully, e.g., based on past research or theoretically well motivated.

- **Hierarchical method (ordered entry)**: first, known predictors are entered, then new ones, either blockwise (all together) or stepwise
- **Forced entry ('enter')**: All predictors are forced into the model simultaneously
- **Stepwise methods: *Forward***: Predictors are introduced one by one, according to their predictive power. ***Stepwise***: Same as Forward + a removal test. ***Backward***: Predictors are judged against a removal criterion and eliminated accordingly.

How to choose one's predictors

- Based on the theoretical literature, choose predictors in their order of importance. Do not choose too many
- Run an initial multiple regression
- Eliminate useless predictors
- Take ca. n=15 subjects per predictor

Evaluating the model

1. The model must fit the data sample
2. The model should generalize beyond the sample

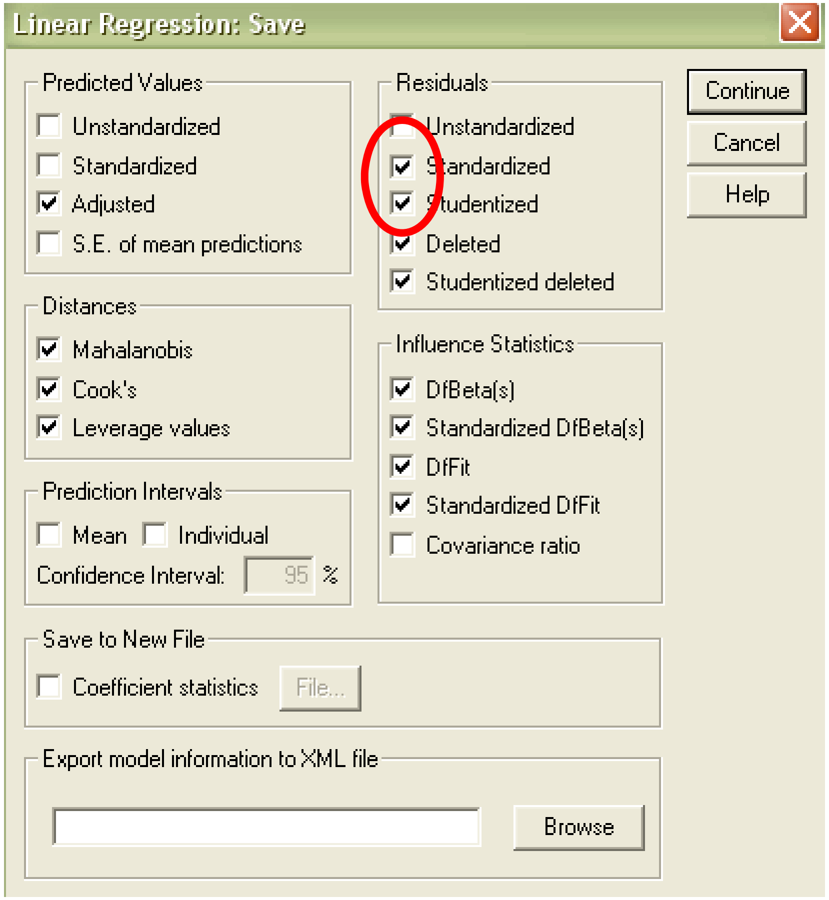
Evaluating the model - diagnostics

1. Fitting the observed data:
- Check for outliers which bias the model and enlarge the residual
- Look at standardized residuals (z-scores): If > 1% are lying outside the margins of +/- 2.58, the model is poor.
- Look at studentized residuals: (unstandardized residuals/ SD that varies point by point.) Yields a more exact estimate of error variance.

Note: SPSS adds the computed scores into new columns in the data file.

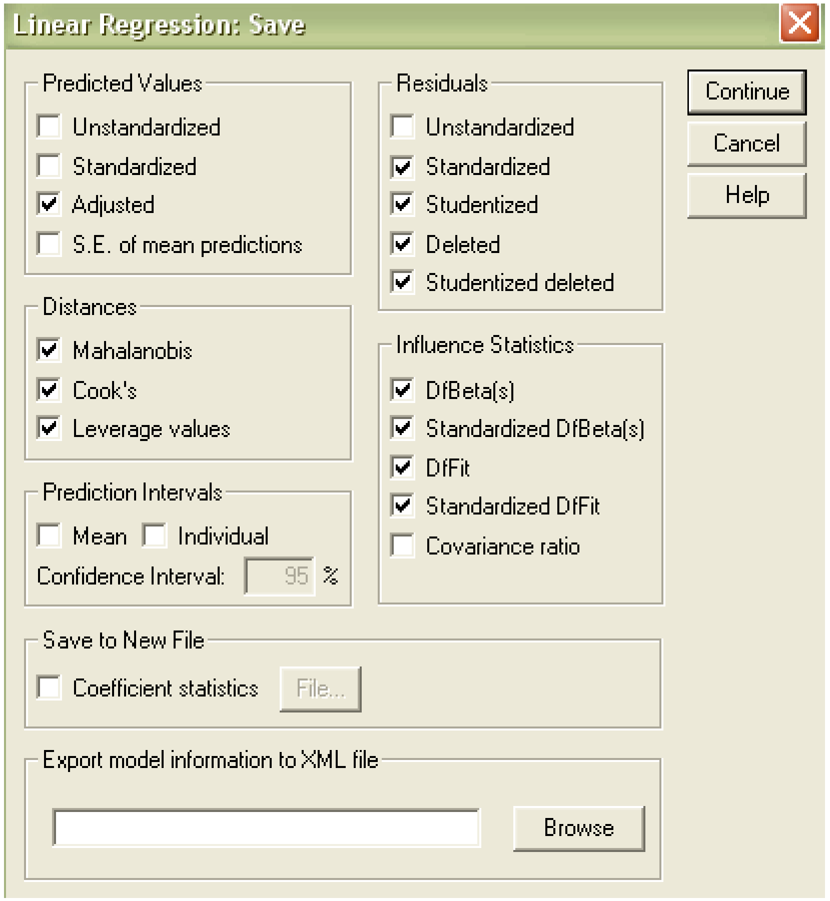
Analyze --> Regression --> Linear

Under 'Save', specify:



Evaluating the model - diagnostics - continued

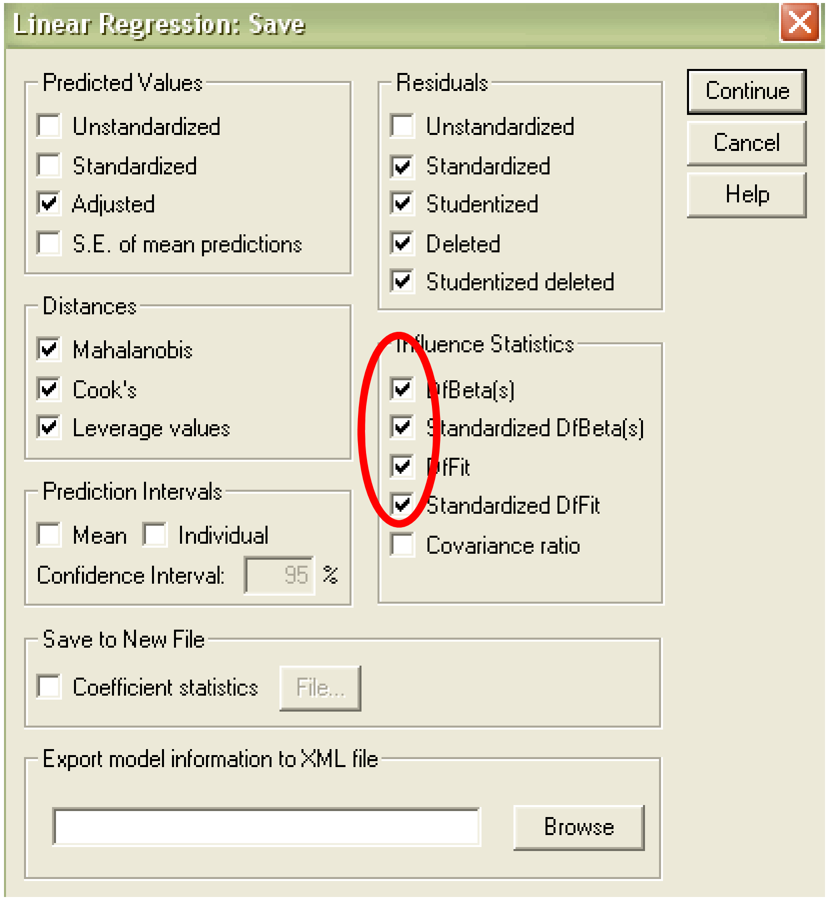
- Identify influential cases and see how the model changes if they are excluded.



This is done by running the regression without that particular case and then use the new model to predict the value of the just excluded case (its 'adjusted predicted value'). If the case is similar to all other cases, its 'adjusted predicted value' will not differ much from its predicted value, given the model including it.

Evaluating the model - continued

DFBeta: a measure of the influence of a case on the values of b_i.
DFFit: "...difference between the adjusted predicted value and the original predicted value of a particular case." (Field 2005, 729).
Deleted residual: residual based on the adjusted predicted value. "... the difference between the adjusted predicted value for a case and the original observed value for that case." (Field 2005, 728)

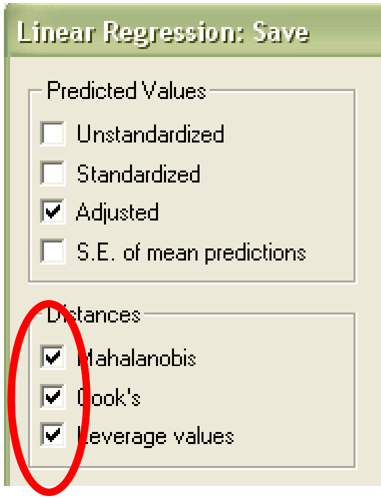


A way of standardizing the deleted residual is to divide it by its SD --> studentized deleted residual.

Evaluating the model
- continued

- Identify influential cases and see how the model changes if they are excluded.

Cook's distance measures the influence of a case on the overall model's ability to predict all cases.
Leverage estimates "the influence of the observed value of the outcome variable over the predicted values." (Field 2005, 736)
Leverage values lie between 0<x>1 and may be used to define cut-off points for excluding influential cases.



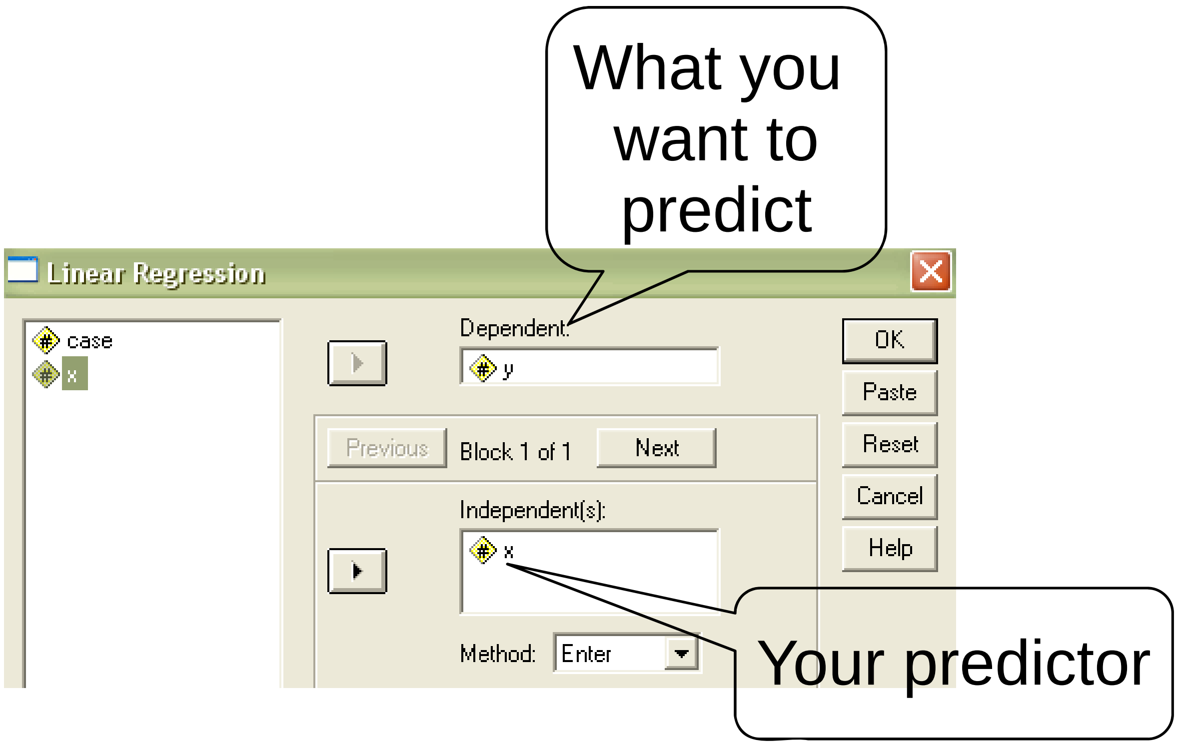
Mahalanobis distances measure the distance of cases from the means of the predictor variables.

Example for using DFBeta as an indicator of an 'influential case'

using file dfbeta.sav

- Run a simple regression with all data (including outlier, case 30):

Analyze --> Regression --> Linear



Example for using DFBeta as an indicator of an 'influential case'

using file dfbeta.sav

- **All data (including outlier**, case 30):
- B0=29; b1= -.90

- **Case 30 removed (with Data --> Select cases --> use filter variable)**
- B0 = 31; b1=-1

→ Both regression coefficients b0 (constant/intercept) and b1 (gradient/slope) changed !

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta
1	(Constant)	29,000	,992	
	X	-,903	,056	-,950

a. Dependent Variable: Y

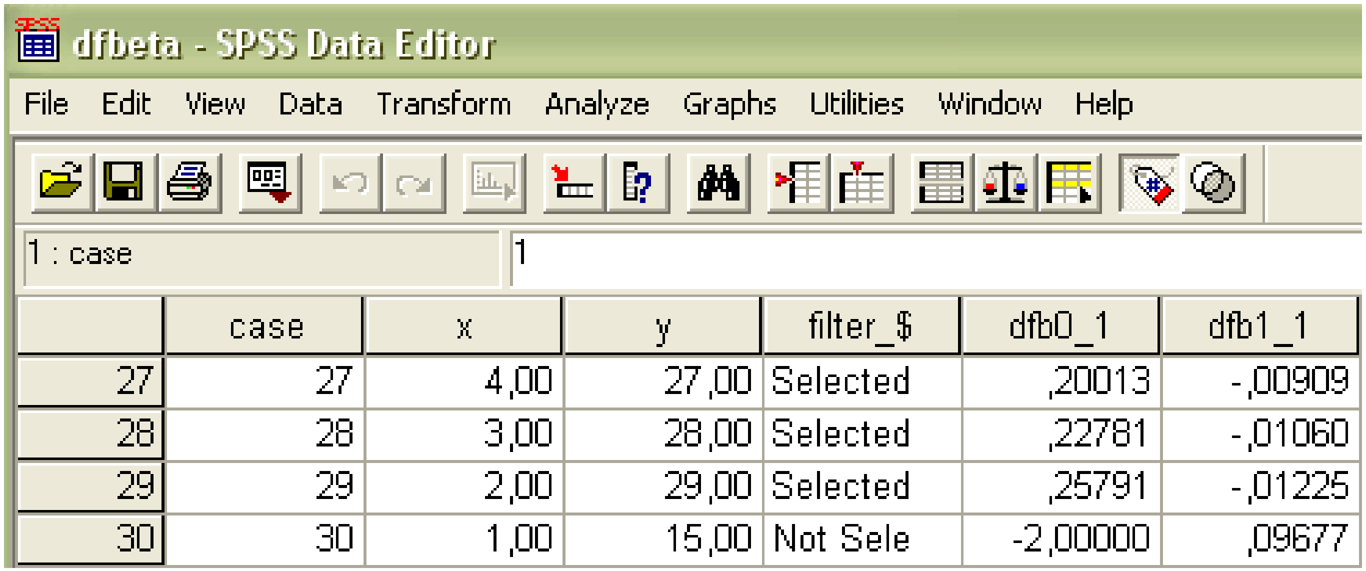
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	31,000	,000		,	,
	X	-1,000	,000	-1,000	,	,

a. Dependent Variable: Y

Example for using DFBeta as an indicator of an 'influential case'

using file dfbeta.sav


dfbeta - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : case 1

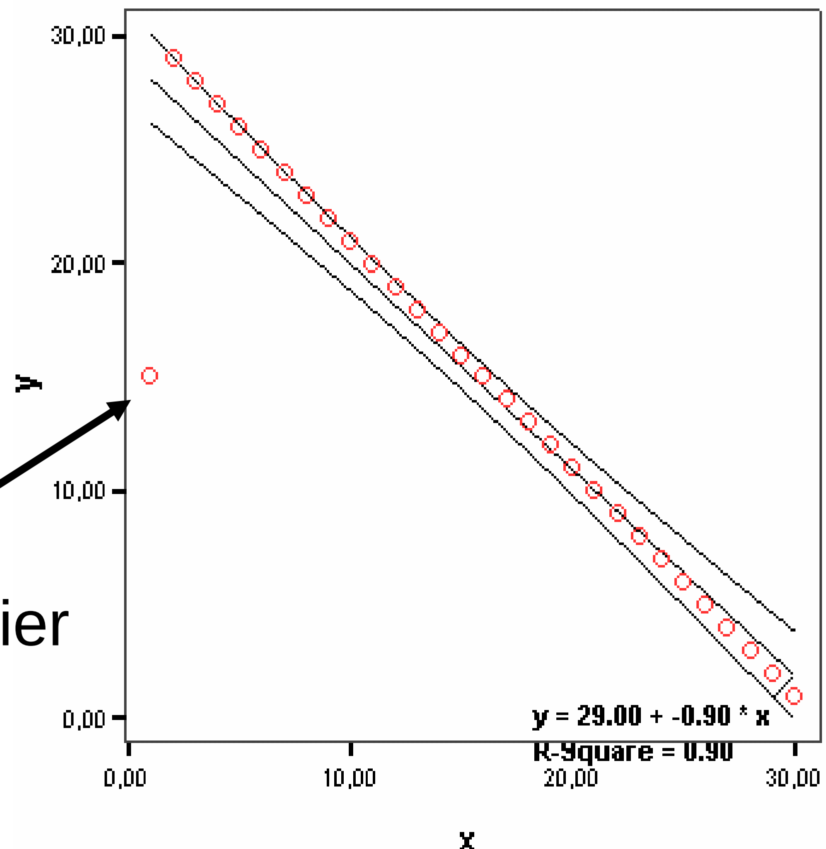
	case	x	y	filter_$	dfb0_1	dfb1_1
27	27	4,00	27,00	Selected	,20013	-,00909
28	28	3,00	28,00	Selected	,22781	-,01060
29	29	2,00	29,00	Selected	,25791	-,01225
30	30	1,00	15,00	Not Sele	-2,00000	,09677

Dfbeta of the constant (dfb0) and of the predictor x (dfb1) are much higher than those of the other cases

Summary of both calculations
Scatterplots for both samples

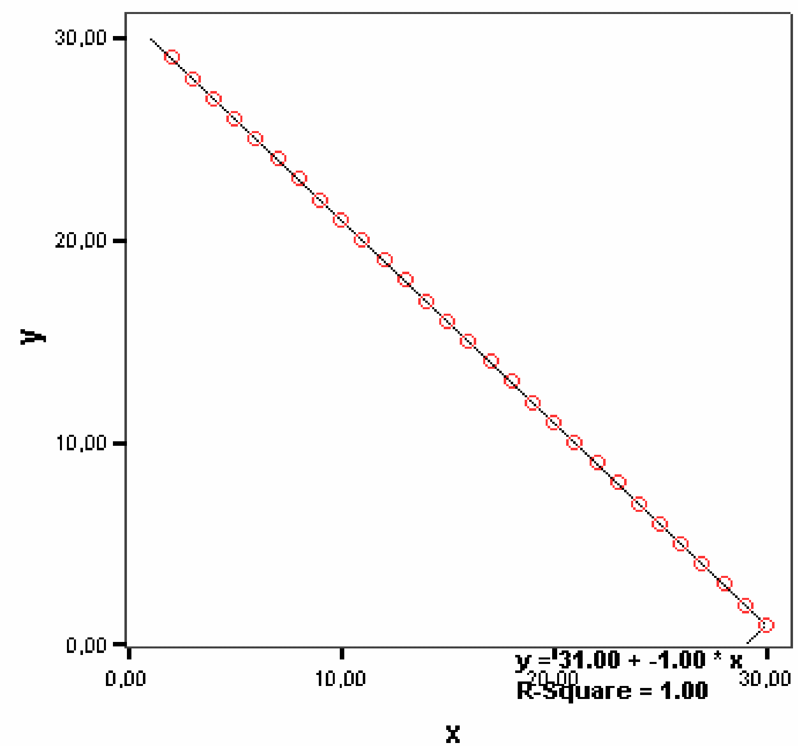
Parameter (b)	+ case 30	- case 30	Difference
Constant (b0)	29.00	31.00	-2.00
Gradient (b1)	-.90	-1	.10
Model	Y=(-.9)X+29	Y=(-1)X+31	
Predicted Y	28.0100		30-1.09

- With case 30:



Outlier

- Without case 30



DFBetas, DFFit, CVR's

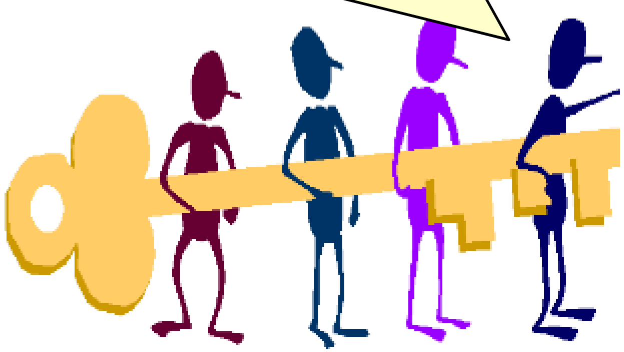
All the following measures measure the difference between a model including and one excluding influential cases:

- Standardized DFBeta: Difference between a **parameter** estimated using all cases and estimated when one case is excluded, e.g. DFBetas of the parameters b_0 and b_1.
- Standardized DFFit: Difference between the predicted value for a **case** in a model including vs. in a model excluding this value.
- Covariance ratio (CVR): measure of whether a case influences the variance of the regression parameters. This ratio should be close to 1.

Help-Window,Topic index 'Linear Regression' Window „Save new variables“

I find it hard to remember what all those influence statistics mean...

Why don't you look them up in the „Help window“ ?



SPSS for Windows

Inhalt | Index | Zurück | Drucken | Optionen

Linear Regression Save

How To | See Also

You can save predicted values, residuals, and other statistics useful for diagnostics. Each selection adds one or more new variables to your active data file.

Predicted Values. Values that the regression model predicts for each case.

Distances. Measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the regression model.

Prediction Intervals. The upper and lower bounds for both mean and individual prediction intervals.

Residuals. The actual value of the dependent variable minus the value predicted by the regression equation.

Influence Statistics. The change in the regression coefficients (DfBeta(s)) and predicted values (DfFit) that results from the exclusion of a particular case. Standardized DfBetas and DfFit values are also available along with the covariance ratio, which is the ratio of the determinant of the covariance matrix with a particular case excluded to the determinant of the covariance matrix with all cases included.

Save to New File. Saves regression coefficients to a file that you specify.

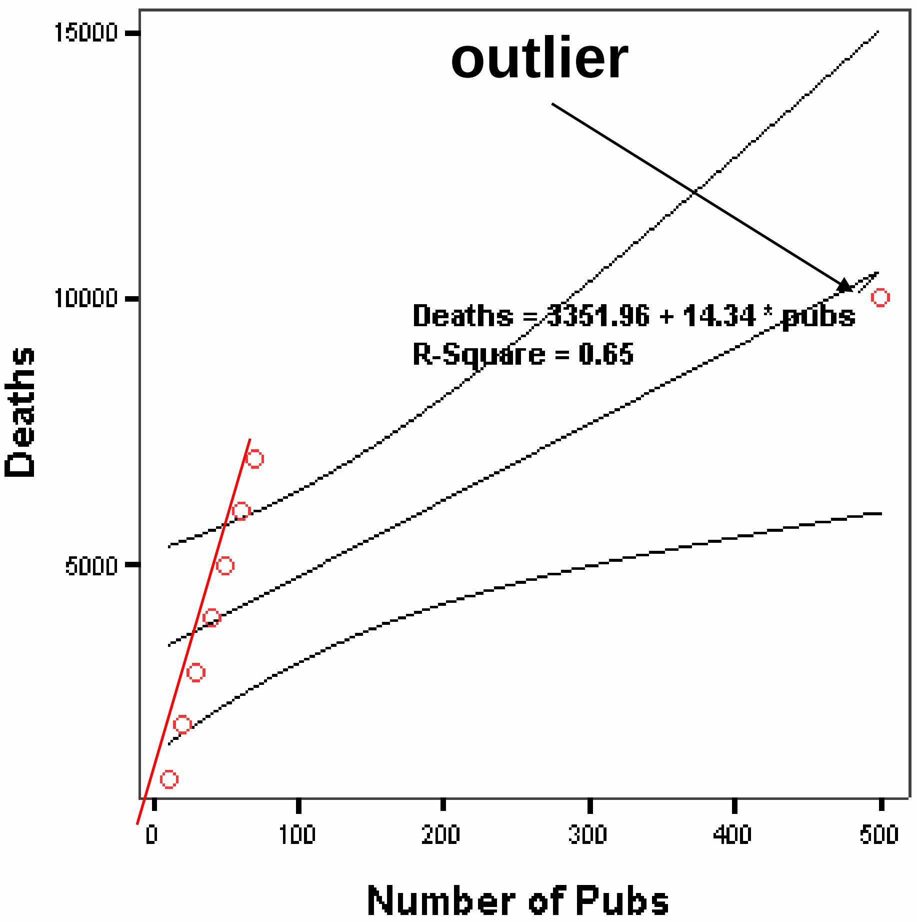
Export model information to XML file. Exports model information to the specified file. SmartScore and future releases of WhatIf? will be able to use this file.

Click **See Also** for descriptions of related dialog boxes and procedures.

? Click your right mouse button on any item in the dialog box for a description of the item.

http://image.informatik.htw-aalen.de/Thierauf/Knobelaufgaben/Sommer03/zweifel.png

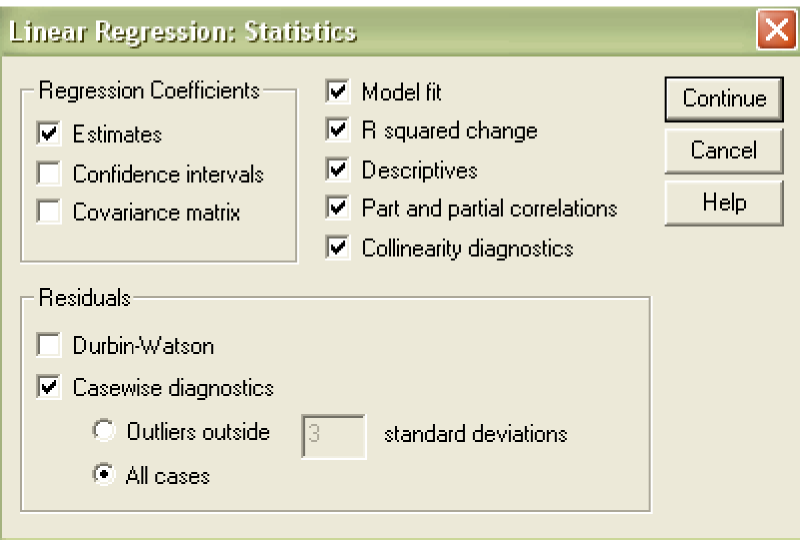
Residuals and influence statistics
(using the file pubs.sav)



Scatterplot of both variables
Graphs --> Interactive --> scatterplot

The correlation between no. of pubs in London districts and deaths with and without the outlier.
Note: The residual for the outlier fitted to the regression line including it is small. However, its influence statistics is huge.

Why? The outlier is the 'City of London' district, where a lot of pubs are but only few residents live. The ones who are drinking in those pubs are visitors, hence, the ratio of deaths of citizens given the overall consumation of alcohol is relatively low.



Case summary: 8 London districts

	St. Res.	Lever	St. DFFIT	St. DFB Interc	St. DFB Pubs
1	-1,34	0,04	-0,74	-0,74	0,37
2	-0,88	0,03	-0,41	-0,41	0,18
3	-0,42	0,02	-0,18	-0,17	0,07
4	0,04	0,02	0,02	0,02	-0,01
5	0,5	0,01	0,2	0,19	-0,06
6	0,96	0,01	0,4	0,38	-0,1
7	1,42	0	0,68	0,63	-0,12
8	-0,28	0,86	-4,60E+008	92676016	-4,30E+008
Total	8	8	8	8	8

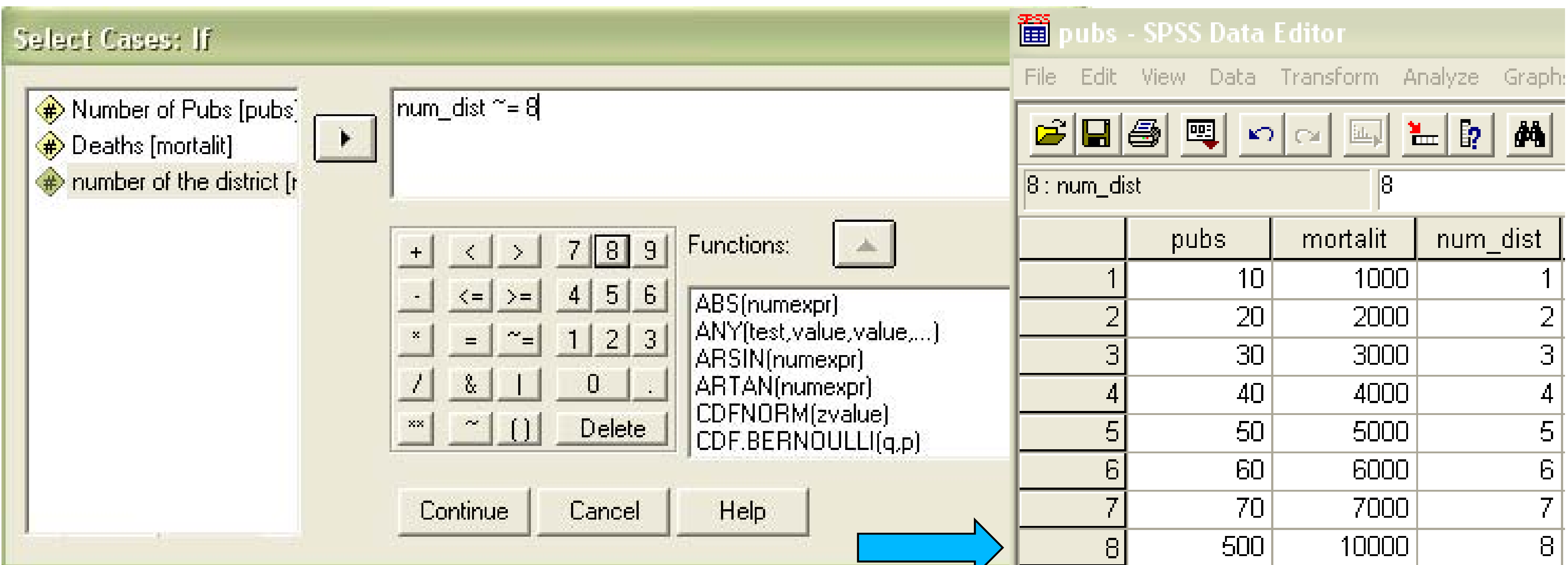
The residual of the outlier #8 is small because it actually sits very close to the regression line

The influence statistics are huge!

Excluding the outlier
(pubs.sav)

If you create a variable “num_dist” (number of the district) in the variables list of the pubs.sav file and simply allocate a number to each district (1-8), you can use this variable to exclude the problematic district #8.

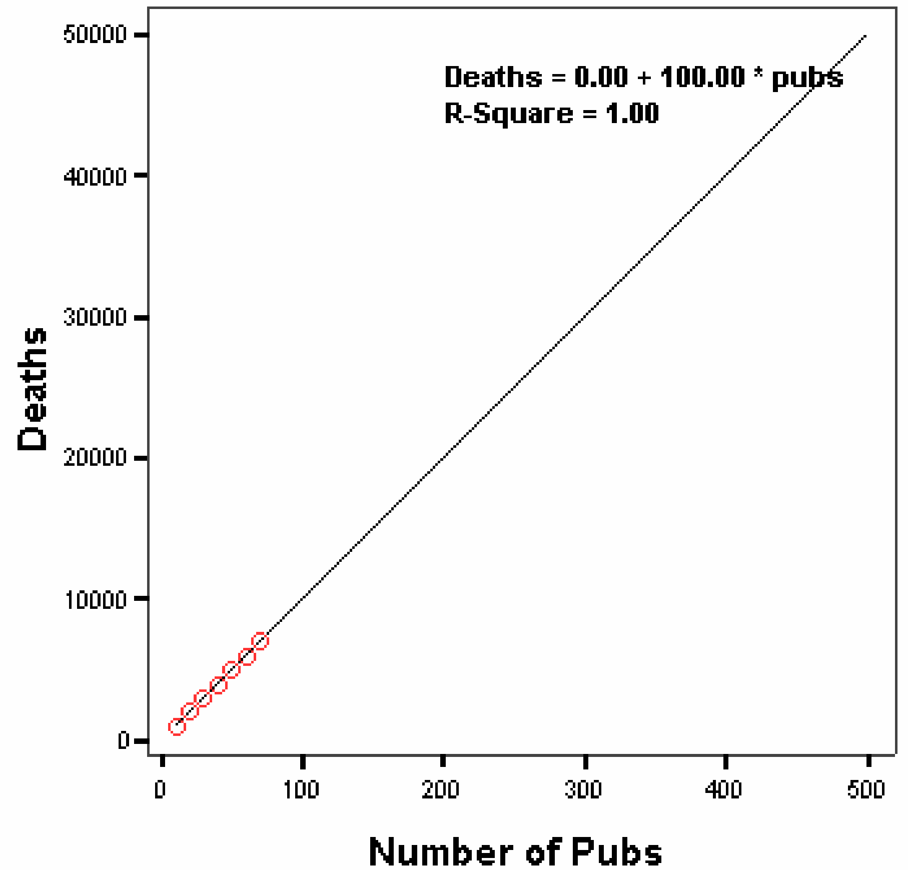
Data → Select cases → If condition is satisfied → num_dist~=8



Excluding the outlier – continued
(pubs.sav)

Look at the scatterplot again now that district # 8 has been excluded:

Graphs → Interactive → Scatterplot



Now the 7 remaining districts all line up perfectly on the (idealized) regression line

Will our sample regression generalize to the population?

If we want to generalize our findings of one sample to the population, we have to check some assumptions:

- Variable types: predictor variables must be quantitative (interval) or categorical (binary); outcome variable must be quantitative, continuous and unbounded (whole range must be instantiated)
- Non-zero variance of predictors
- No perfect correlation between ≥ 2 predictors
- Predictors are uncorrelated to any 'third variable' which was not included in the regression
- All levels of the predictor variables should have same variance

Will our sample regression generalize to the population?
- continued

- Independent errors: The residual terms of any two observations should be uncorrelated (Durbin-Watson Test)
- Residuals should be normally distributed
- All of the values of the outcome variable are independent
- Predictors and outcome have a linear relation

- If these assumptions are not met, we cannot draw valid conclusions from our model!

Two methods for the cross-validation of the model

If our model is generalizable, it should be able to predict the outcome of a different sample.

- ***Adjusted R^2:* R^2 indicates the loss of predictive power (shrinkage) if the model were applied to the population:**

$$adj\ R^2 = 1 - \left\{ \left(\frac{n-1}{n-k-1}\right) \left(\frac{n-2}{n-k-2}\right) \left(\frac{n+1}{n}\right) \right\} (1-R^2)$$

R^2= unadjusted value
n= number of cases
k= number of predictors in the model

- ***Data splitting*: The entire sample is split into two. Regressions are computed and compared for both halves. Nice method but one rarely has so many data.**

Sample size

The required sample size for a regression depends on

- ***The number of predictors k***
- ***The size of the effect***
- ***The size of the statistical power***

e.g.,

large efffect --> n= 80 (for up to 20 predictors)

medium effect --> n=200

small effect --> n=600

(Multi-)Collinearity

If $\geq$ 2 predictors are inter-correlated, we speak of **collinearity**. In the worst case, 2 variables have a correlation of 1. This is bad for a regression, since the regression cannot be computed reliably anymore. This is because the variables become interchangeable.
High collinearity is rare, but some degree of collinearity is always around.

Problems with collinearity:

- It underestimates the variance of a second variable if this variable is strongly intercorrelated with the first variable. It adds little unique variance although – taken for itself – it would explain a lot.
- We can't decide which variable is important, which variable should be included
- The regression coefficients (b-values) become instable.

How to deal with collinearity

SPSS has some collinearity diagnostics:

- Variance inflation factor
- Tolerance statistics
- ...

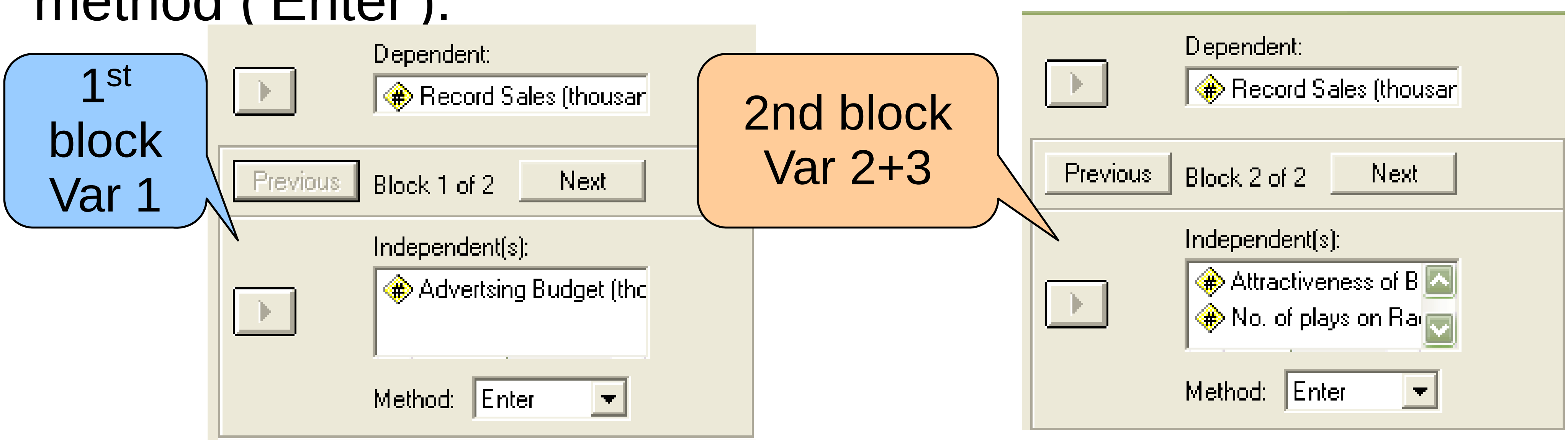
→ in the 'Statistics' window of the 'linear regression' menu

Multiple Regression on SPSS
(using the file Record2.sav)

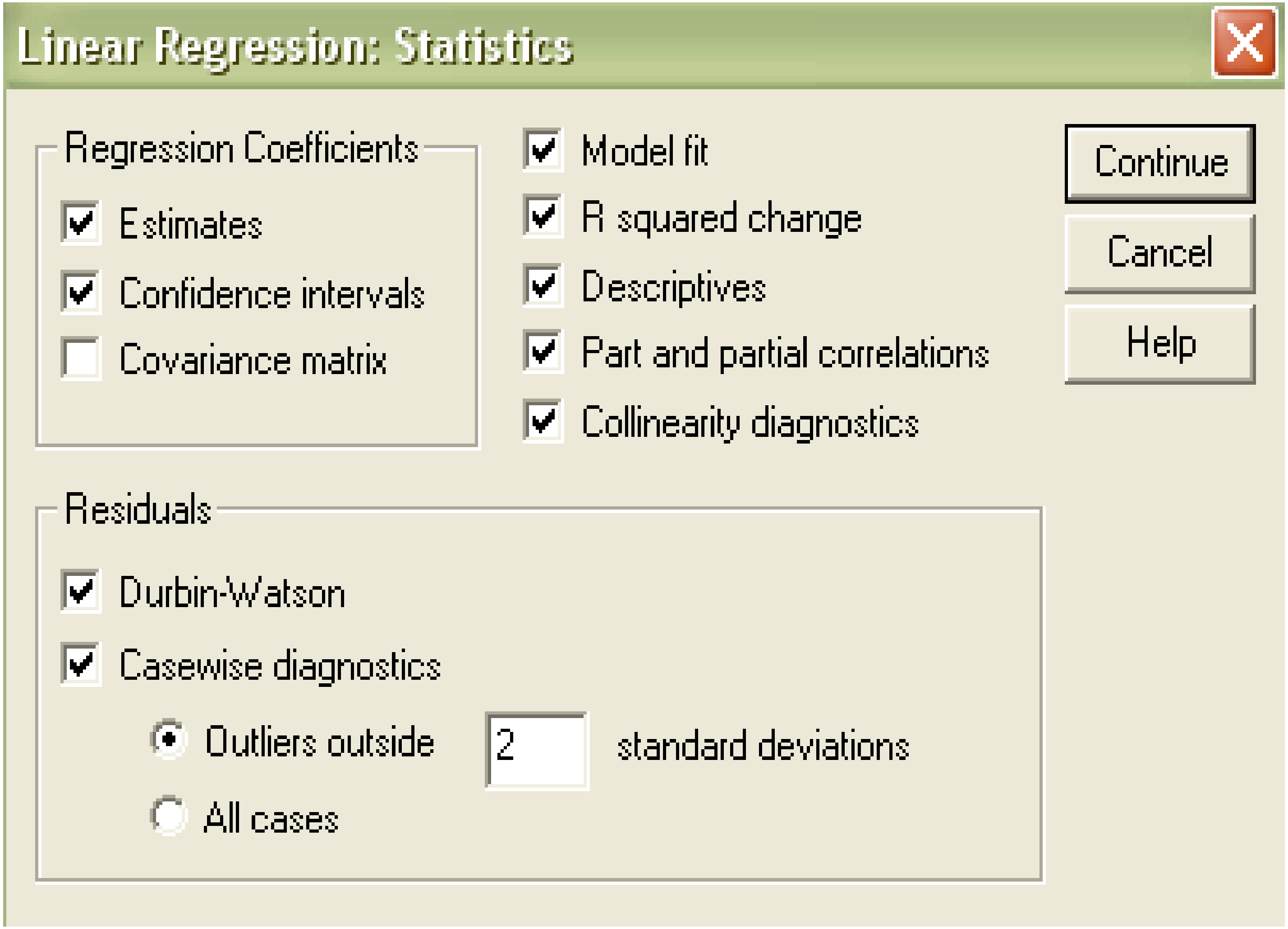
Example: **Predicting the record sales from 3 predictors:**

- ***X1: Advertisement budget,***
- ***X2: times played on radio,***
- ***X3: attractiveness of the band***

Since we know already that money for ads is a predictor, it will be entered into the regression first (1^{st} block), and the 2 new predictors later (2^{nd} block) --> hierarchical method ('Enter').

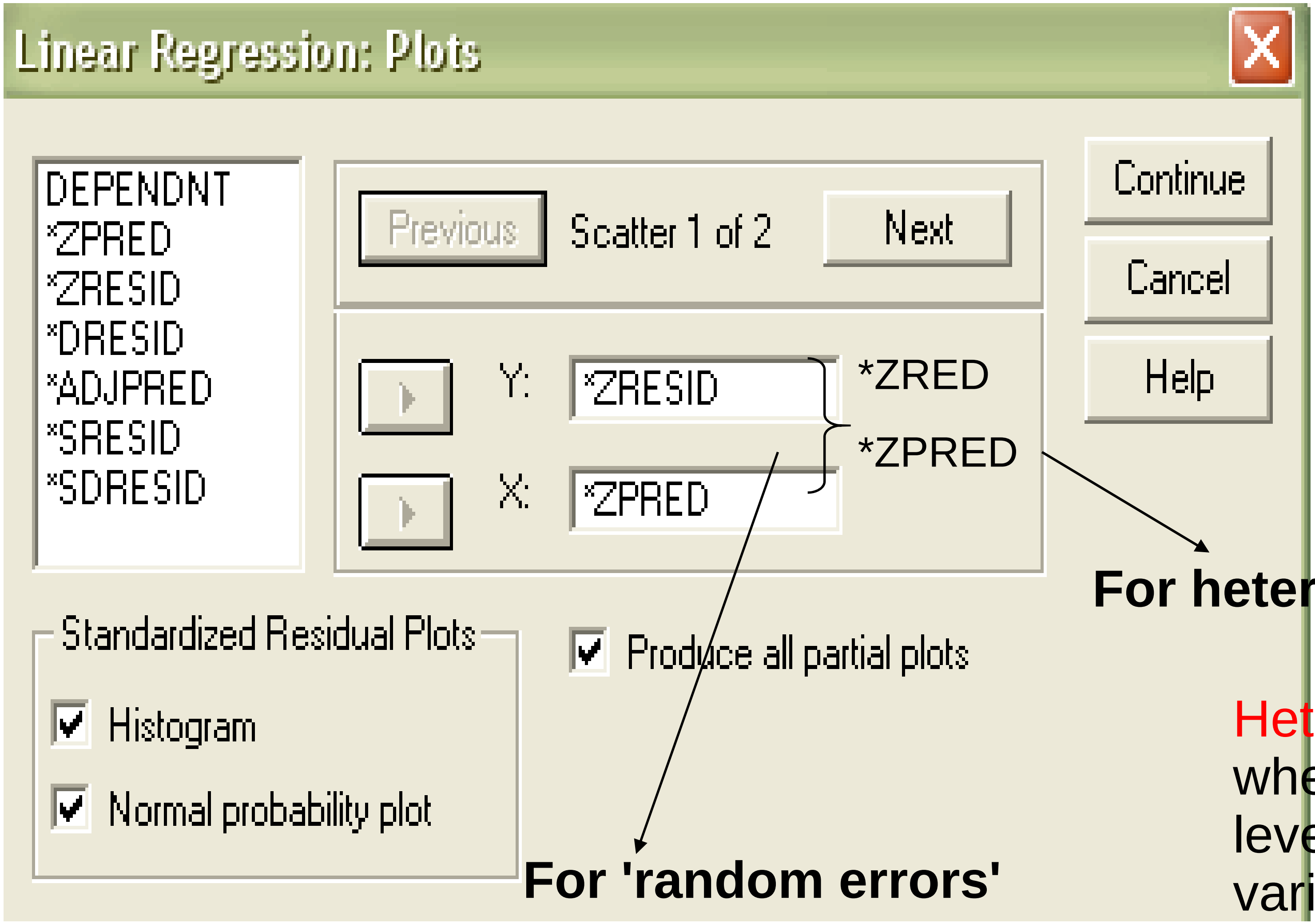


What the „Statistics“ box should look like
Analyze --> Regression --> Linear



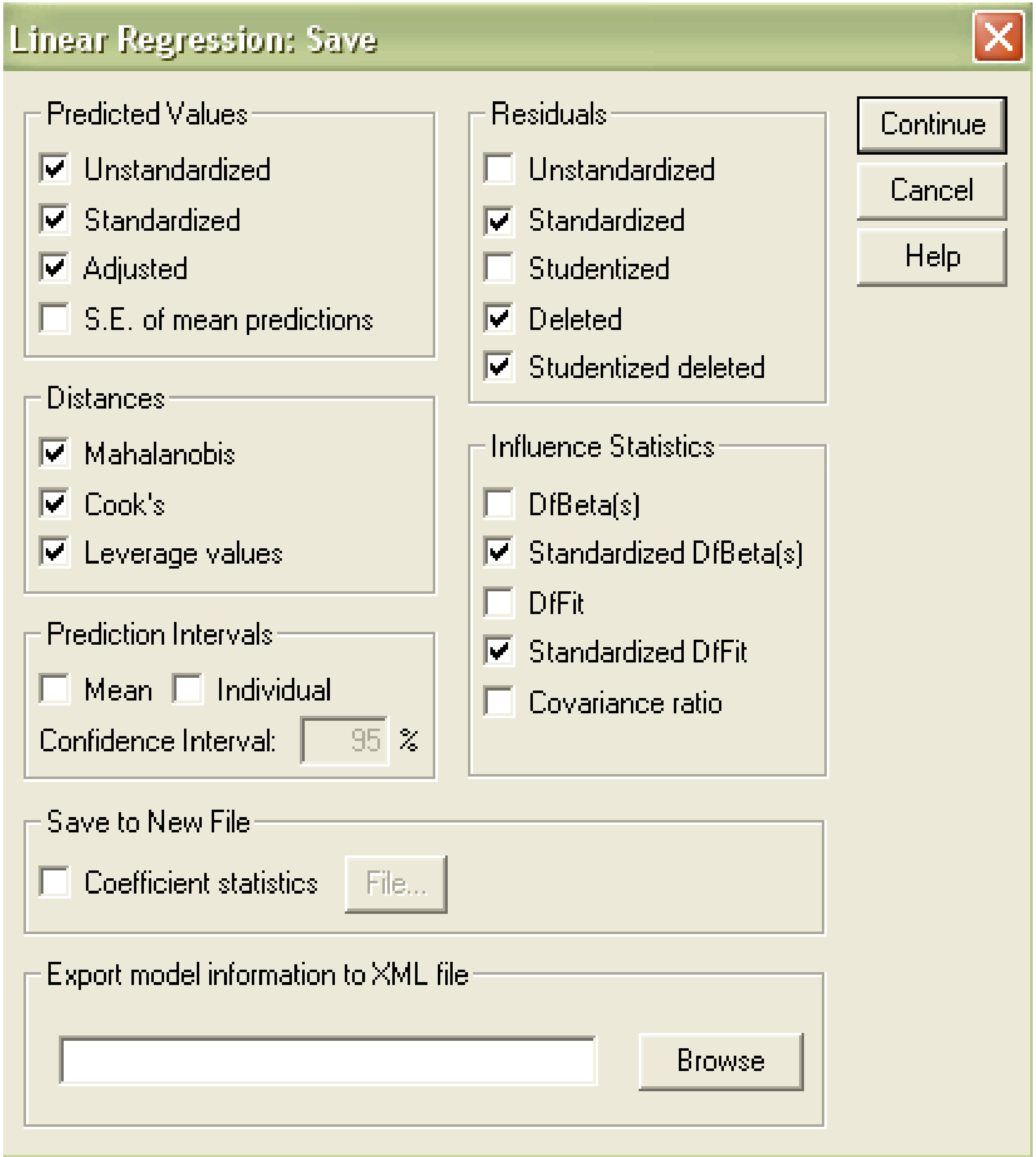
Regression Plots

Plotting *ZRESID (standardized residuals = errors) against *ZPRED (standardized predicted values) helps us determine whether the assumption of random errors and homoscedasticity (equal variances) are met.



Heteroscedasticity occurs when the residuals at each level of the predictor variables have unequal variances.

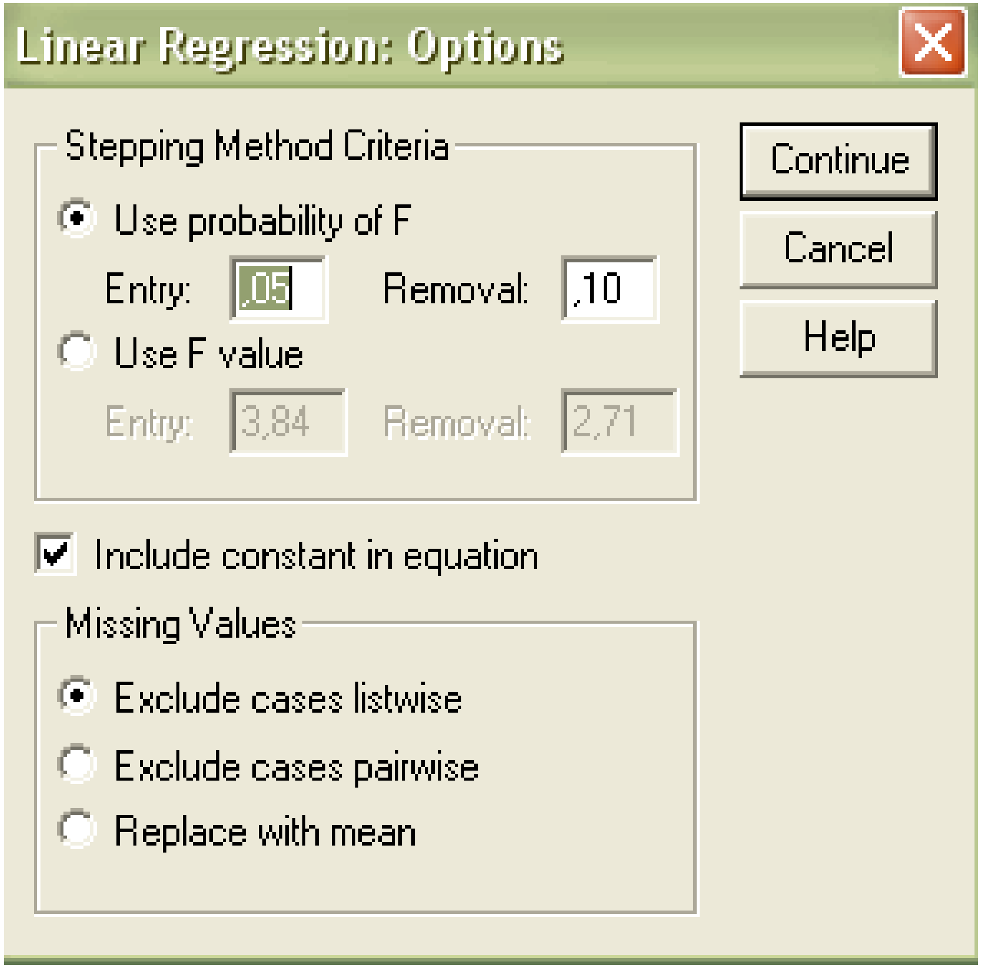
Regression diagnostics



The regression diagnostics are saved in the data file, each as a separate variable in a new column

Options

leave them as they are



Interpreting Multiple Regression

Descriptive Statistics

	Mean	Std. Deviation	N
SALES Record Sales (thousands)	193,2000	80,6990	200
ADVERTS Advertsing Budget (thousands of pounds)	614,4123	485,6552	200
AIRPLAY No. of plays on Radio 1 per week	27,5000	12,2696	200
ATTRACT Attractiveness of Band	6,7700	1,3953	200

The **'Descriptives'** give you a brief summary of the variables

Interpreting Multiple Regression

Correlations

		SALES Record Sales (thousands)	ADVERTS Advertising Budget (thousands of pounds)	AIRPLAY No. of plays on Radio 1 per week	ATTRACT Attractiveness of Band
Pearson Correlation	SALES Record Sales (thousands)	1,000	,578	,599	,326
	ADVERTS Advertising Budget (thousands of pounds)	,578	1,000	,102	,081
	AIRPLAY No. of plays on Radio 1 per week	,599	,102	1,000	,182
	ATTRACT Attractiveness of Band	,326	,081	,182	1,000
Sig. (1-tailed)	SALES Record Sales (thousands)	.	,000	,000	,000
	ADVERTS Advertising Budget (thousands of pounds)	,000	.	,076	,128
	AIRPLAY No. of plays on Radio 1 per week	,000	,076	.	,005
	ATTRACT Attractiveness of Band	,000	,128	,005	.

Pearson correlations R

R of predictors 123 with outcome

R of pred1 with the others

R of pred2 with the other

R of pred3 with the others

Significance levels for all correlations

Correlations: R's between all variables and signif-levels. Pred 2 (plays on radio) is the best predictor.

Predictors should not correlate higher than R>.9 (collinearity)

Summary of model

Only adver-tisement as predic tor

Correlation between predictor(s) and out-come

Change from 0 to .335 (Model 1) and another change of .330 (Model 2)

Degrees of freedom; df1:p-1 df2:N-p-1 (N=sample size; p=# of predictors)

If errors are independent. If value close to 2, then OK

mary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Sta[...]tics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	,578[a]	,335	,331	65,9914	,335	99,587	1	198	,000	
2	,815[b]	,665	,660	47,0873	,330	96,447	2	196	,000	1,950

a. Predictors: (C[...]nstant), AD[...]VERTS Advertsing Budget (thousands of pounds)

b. Predictors: ([...]onstant), [...]VERTS Advertsing Budget (thousands [...]f pounds), ATTRACT Attrac[...]eness of Band, AIRPLAY N[...] of plays [...]Radio 1 per week

c. Dependen[...]ariab[...] Sales (thousand[...]

3 predic tors

Explained variance by the predic-tor(s)

How well the model generalizes. Similar val-ues to R^2 are good. Only 5% shrinkage

F-values for R^2 change

The model(s) bring about a significant change

ANOVA for the model against the basic model (the mean)

SSM

Df equal to # of cases minus 1
200-1=199

Df equal to # of predic- tors

Df equal to # of cases minus # of coefficients (b0,b1)
200-2=198

F-values:
MSM/MSR:
433687.833/4354.87=99.587
287125.806/2217.217=129.498

Significance level

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687,8	1	433687,833	99,587	,000[a]
	Residual	862264,2	198	4354,870		
	Total	1295952	199			
2	Regression	861377,4	3	287125,806	129,498	,000[b]
	Residual	434574,6	196	2217,217		
	Total	1295952	199			

a. Predictors: (Constant), ADVERTS Advertsing Budget (thousands of pounds)

b. Predictors: (Constant), ADVERTS Advertsing Budget (thousands of pounds), ATTRACT Attractiveness of Band, AIRPLAY No. of plays on Radio 1 per week

c. Dependent Variable: SALES Record Sales (thousands)

SSR

SST

Mean squares:
SS/df
433687.8/1=433687.8
862264.2/198=4354.87

Both Model 1 and 2 have improved the prediction significantly, Model 2 (3 predictors) even better than Model 1 (1 predictor)

Model parameters

Record sales increase by .511 SD's when the predictor (ads) changes 1 SD; b1 and b2 have equal 'gains'

With 95% confidence the b-values lie within these boundaries
Tight boundaries are good

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95% Confidence Interval for B Lower Bound	95% Confidence Interval for B Upper Bound	Correlations Zero-order	Correlations Partial	Correlations Part	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	134,140	7,537		17,799	,000	119,28	149,002					
	ADVERTS Advertsing Budget (thousands of pounds)	9,61E-02	,010	,578	9,979	,000	,077	,115	,578	,578	,578	1,000	1,0
2 b0	(Constant)	-26,613	17,350		-1,534	,127	-60,830	7,604					
b1	ADVERTS Advertsing Budget (thousands of pounds)	8,49E-02	,007	,511	12,261	,000 *	,071	,099	,578	,659	,507	,986	1,0
b2	AIRPLAY No. of plays on Radio 1 per week	3,367	,278	,512	12,123	,000	2,820	3,915	,599	,655	,501	,959	1,0
b3	ATTRACT Attractiveness of Band	11,086	2,438	,192	4,548	,000	6,279	15,894	,326	,309	,188	,963	1,0

a. Dependent Variable: SALES Record Sales (thousands)

Model 1= same as in first analysis

Pearson Corr of predictor x outcome controlled for each single other predictor

Pearson Corr of predictor x outcome controlled for all other predictor 'unique relationship'

The 'Coefficients' table tells us the individual contribution of variables to the regression model. The Standardized Beta's tell us the importance of each predictor

Excluded variables

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	VIF	Minimum Tolerance
1	AIRPLAY No. of plays on Radio 1 per week	,546[a]	12,51	,000	,665	,990	1,010	,990
	ATTRACT Attractiveness of Band	,281[a]	5,136	,000	,344	,993	1,007	,993

a. Predictors in the Model: (Constant), ADVERTS Advertsing Budget (thousands of pounds)

b. Dependent Variable: SALES Record Sales (thousands)

What contribution would this predictor have made to a model containing it

SPSS gives a summary of those predictors that were not entered in the Model (here only for Model 1) and evaluates the contribution of the excluded variables.

Regression equation for Model 2 (including all 3 predictor variables)

Model	Model 1= same as in first analysis	Unstandardized Coefficients B	Std. Error
1	(Constant)	134,140	7,537
	ADVERTS Advertsing Budget (thousands of pounds)	9,61E-02	,010
2 b0	(Constant)	-26,613	17,350
b1	ADVERTS Advertsing Budget (thousands of pounds)	8,49E-02	,007
b2	AIRPLAY No. of plays on Radio 1 per week	3,367	,278
b3	ATTRACT Attractiveness of Band	11,086	2,438

$$Sales_i = b0 + b1 Advertising_i + b2 airplay_i + b3 attractiveness_i$$

$$= -26.61 + (0.08 Ad_i) + (3.37 Airplay_i) + (11.09\ Attract_i)$$

Interpretation:

If Ad increaes 1 unit-->sales increase .08 units; if airplay + 1 unit-->sales+3.37; if attract + 1 unit --> sales +11 units, independent of the contributions of the other predictors.

No Multicollinearity
(In this regression, variables are not closely linearly related)

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	ADVERTS Advertsing Budget (thousands of pounds)	AIRPLAY No. of plays on Radio 1 per week	ATTRACT Attractiveness of Band
1	1	1,785	1,000	,11	,11		
	2	,215	2,883	,89	,89		
2	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	2,039E-02	13,219	,94	,00	,00	,92

a. Dependent Variable: SALES Record Sales (thousands)

Each predictor's variance proportions load highly on a different dimension (Eigenvalue)
--> they are not intercorrelated, hence no collinearity

Casewise diagnostics

Casewise Diagnostics[a]

Case Number		Std. Residual (z-value)	SALES Record Sales (thousands)	Predicted Value	Residual
1	>5%	2,125	330,00	229,9203	100,0797
2		-2,314	120,00	228,9490	-108,9490
10		2,114	300,00	200,4662	99,5338
47		-2,442	40,00	154,9698	-114,9698
52		2,069	190,00	92,5973	97,4027
55		-2,424	190,00	304,1231	-114,1231
61		2,098	300,00	201,1897	98,8103
68		-2,345	70,00	180,4156	-110,4156
100		2,066	250,00	152,7133	97,2867
164	>1%	-2,577	120,00	241,3240	-121,3240
169	>1%	3,061	360,00	215,8675	144,1325
200	>5%	-2,064	110,00	207,2061	-97,2061

a. Dependent Variable: SALES Record Sales (thousands)

The casewise diagnostics lists cases that lie outside the boundaries of 2 SD (in the z-distribution, only 5% should be beyond 1.96 SD and only 1% beyond 2.58

Case 169 deviates most and needs to be followed up

Following up influential cases with „Case summaries“ --> everything OK

No DFBETA's >1 (all OK

Leverage values <.06 (all OK)

Case S

	SDB0_1 Standardized DFBETA Intercept	SDB1_1 Standardized DFBETA ADVERTS	SDB2_1 Standardized DFBETA AIRPLAY	SDB3_1 Standardized DFBETA ATTRACT	SDF_1 Standardized DFFIT	COO_1 Cook's Distance	MAH_1 Mahalanobis Distance	LEV_1 Centered Leverage Value
1	-,31554	-,24235	,15774	,35329	,48929	,05870	8,39591	,04219
2	,01259	-,12637	,00942	-,01868	-,21110	,01089	,59830	,00301
3	-,01256	-,15612	,16772	,00672	,26896	,01776	2,07154	,01041
4	,06645	,19602	,04829	-,17857	-,31469	,02412	2,12475	,01068
5	,35291	-,02881	-,13667	-,26965	,36742	,03316	4,81841	,02421
6	,17427	-,32649	-,02307	-,12435	-,40736	,04042	4,19960	,02110
7	,00082	-,01539	,02793	,02054	,15562	,00595	,06880	,00035
8	-,00281	,21146	-,14766	-,01760	-,30216	,02229	2,13106	,01071
9	,06113	,14523	-,29984	,06766	,35732	,03136	4,53310	,02278
10	,17983	,28988	-,40088	-,11706	-,54029	,07077	6,83538	,03435
11	-,16819	-,25765	,25739	,16968	,46132	,05087	3,14841	,01582
12	,16633	-,04639	,14213	-,25907	-,31985	,02513	3,49043	,01754
Total N	12	12	12	12	12	12	12	12
N	12	12	12	12	12	12	12	12

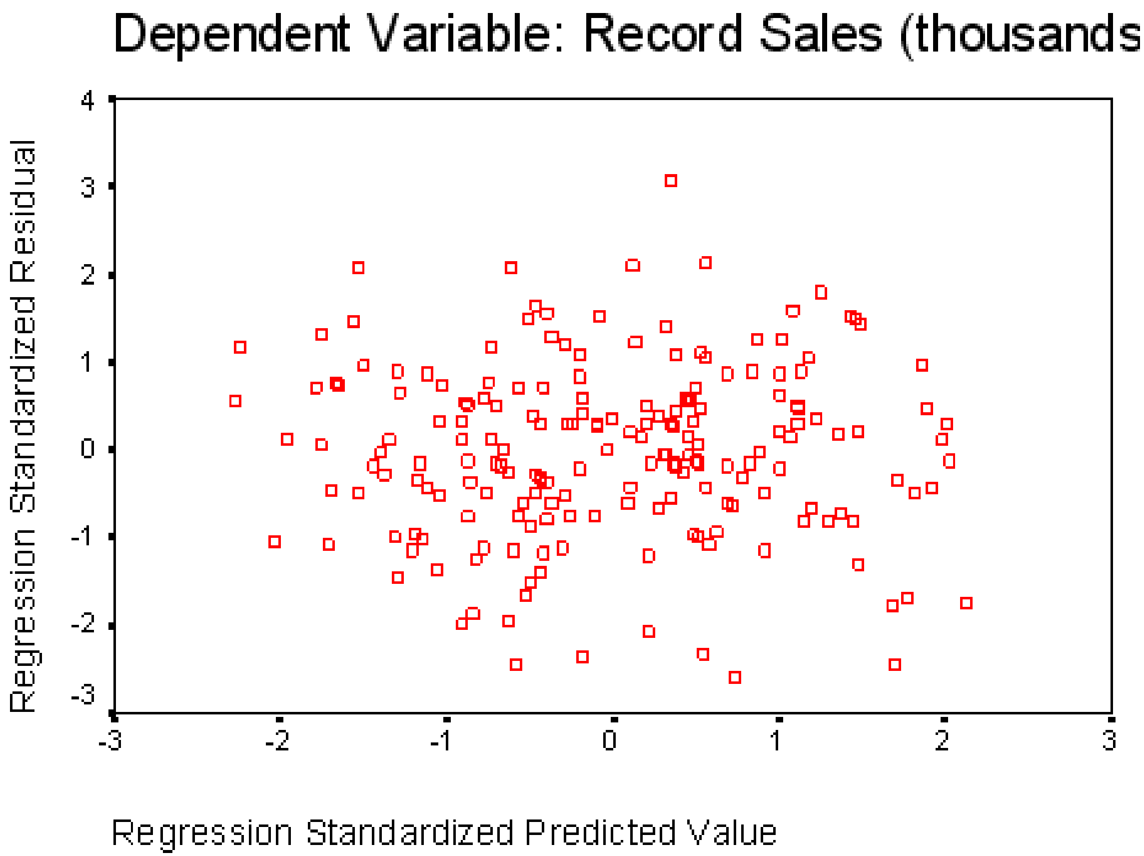
Cook distances <1 (all OK)

Mahalanobis' distances <15 (all OK)

Identify influencing cases by the case summary

- In the standardized residulas, no more than 5% must have values exceeding 2 and 1% exceeding 3.
- Cook's distances >1 might pose a problem
- Leverage (# of predictors + 1/sample size) must not be twice or three times higher
- Mahalanobis distance: cases with >25 in large samples (n=500) and >15 in small samples (n=100) can be problemantic
- Absolute values of DFBeta should not exceed 1
- Determine upper and lower limit of covariance ratio (CVR). Upper limit = 1+3(average leverage); lower limit = 1-3(average leverage).

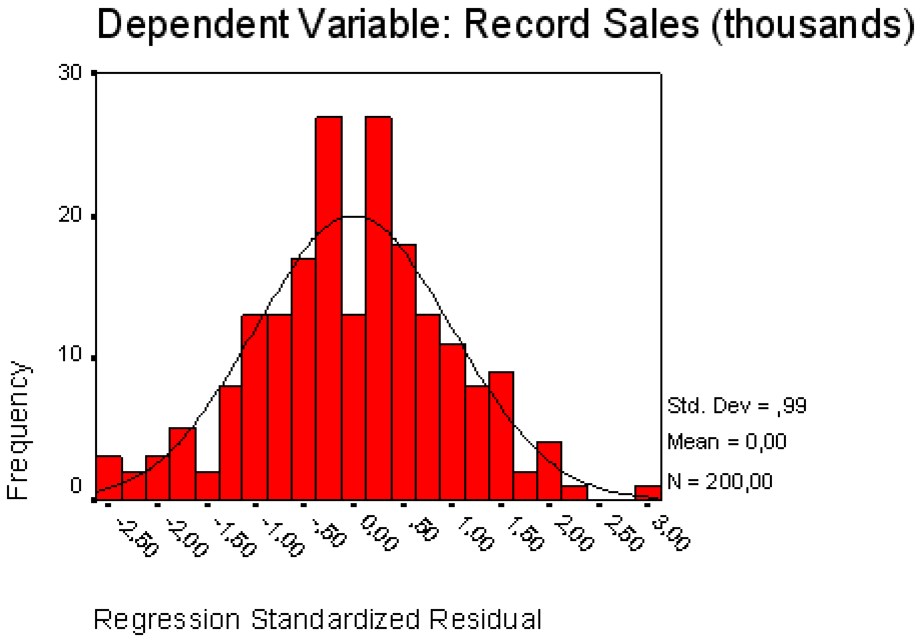
Checking assumptions: Heteroscedasticity

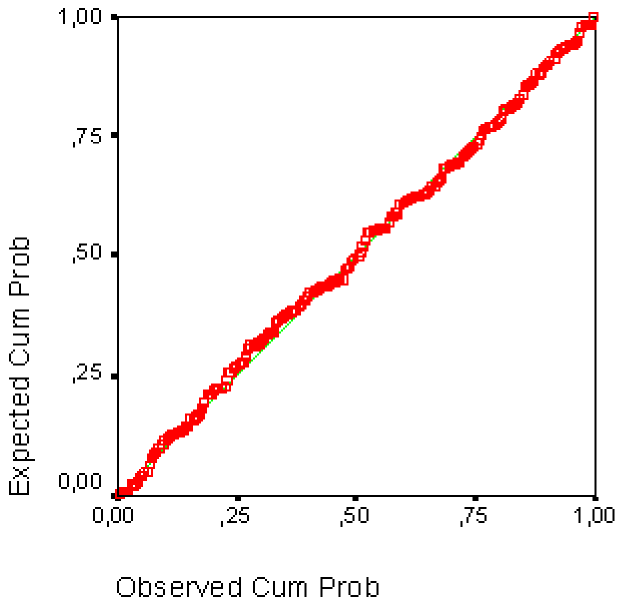


(Heteroscedasticity: residuals (errors) at each level of predictor have different variances). Here variances are equal

Plot of standardized residual *ZRESID/ standardized predicted value *ZPRED
Points are randomly and evently dispersed
--> assumptions of linearity and homoscedasticity are met

Checking assumptions
Normality of residuals





The distribution of the residuals is normal (left hand picture), the observed probabilities correspond to the expected ones (right hand side)

Checking assumptions
Normality of residuals - continued

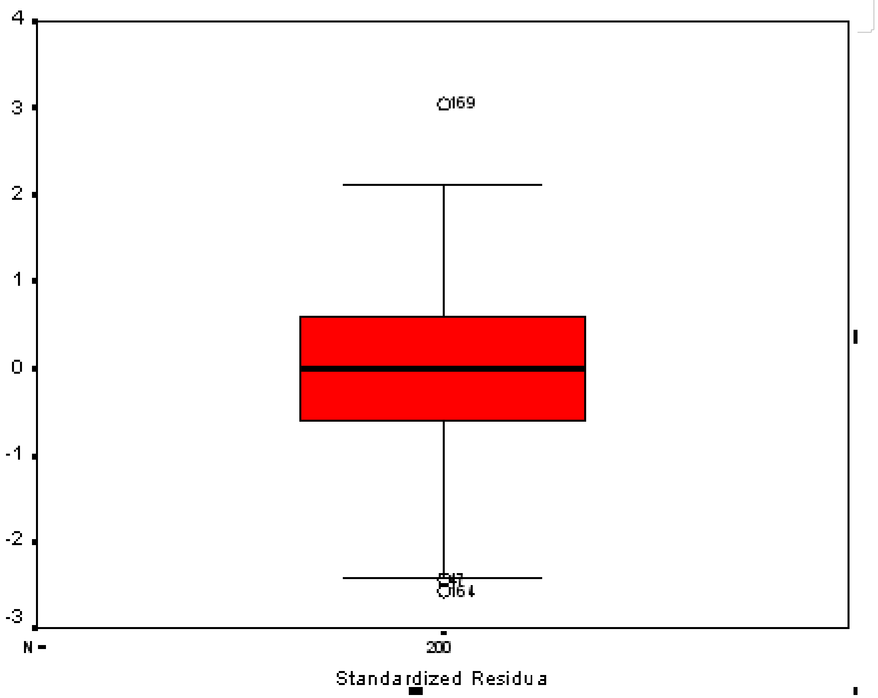
Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
ZRE_1 Standardized Residual	,035	200	,200*

*. This is a lower bound of the true significance.

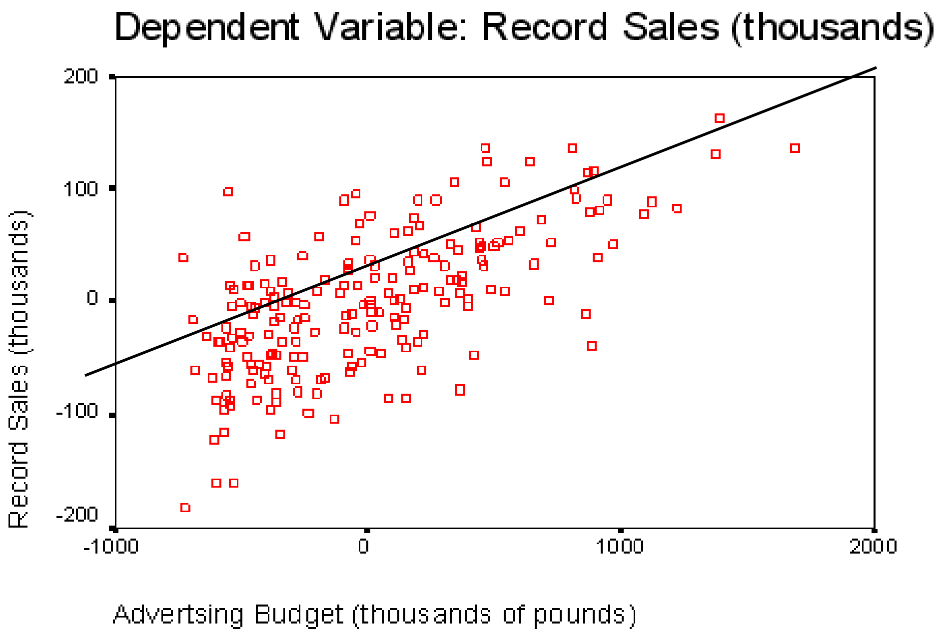
a. Lilliefors Significance Correction

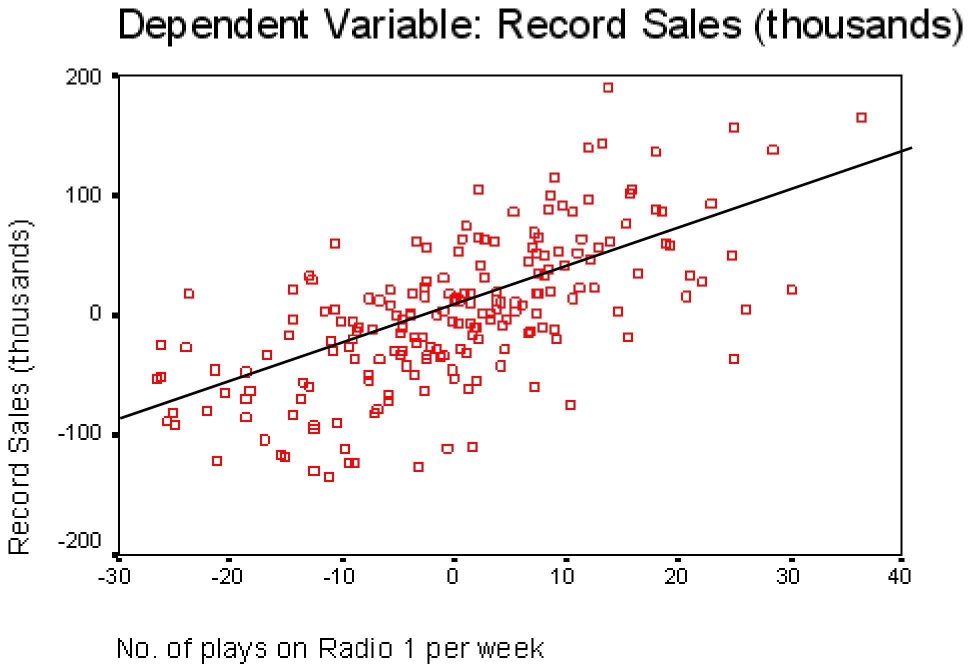
The Kolmogoroff-Smirnov-Test for the standardized residuals is n.s.
--> normal distribution

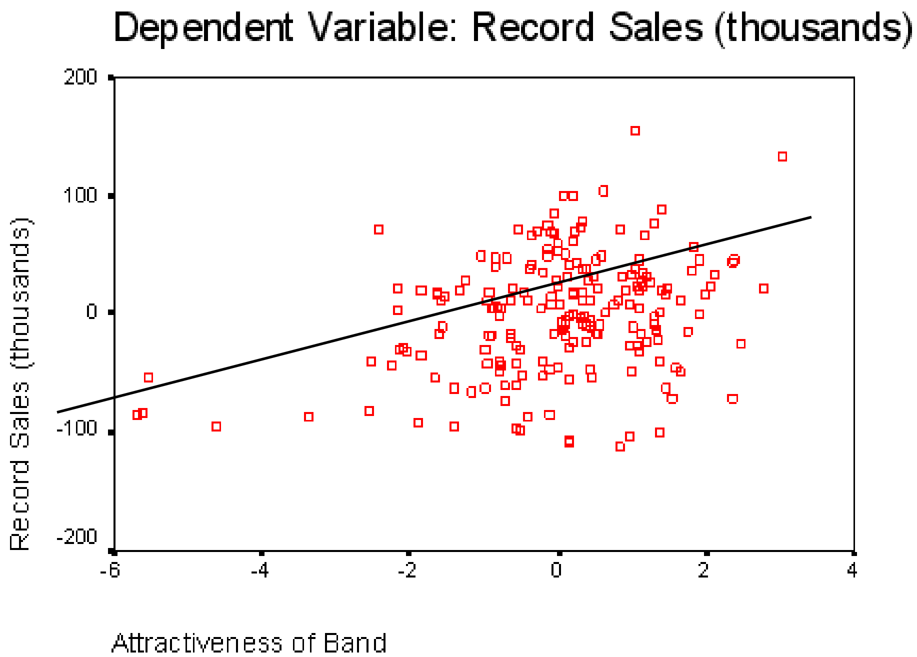


Boxplots, too, show the normality
(note the 3 outliers!)

Checking assumptions Partial Regression Plots







Scatterplots of the residuals of the outcome variable and each of the predictors separately.

No indication of outliers, evenly spaced out cloud of dots (only the residual variance of 'attractiveness of band' seems to be uneven.