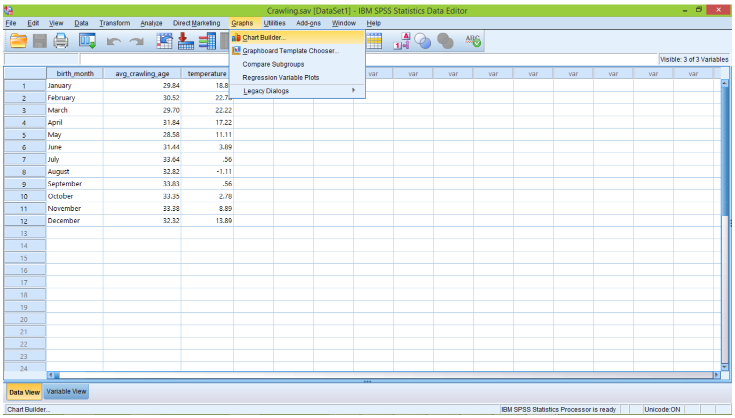
Linear Regression in SPSS

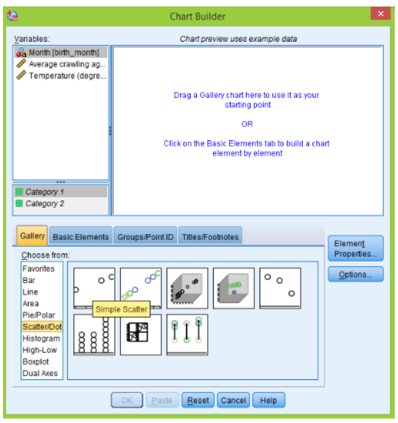
In this document we describe how to perform a simple linear regression in SPSS. We show how to get coefficients of a regression line, test for significance of the slope, find R^2 statistic, make transformations to variables and much more. We also show how to make plot of the data with regression as well as plotting residuals.
For this document we need 'Skeleton', 'Life Expectancy', 'Crawling', 'CFC11' and 'Coffee Shop' data sets. It is assumed that you have managed to upload all these data into SPSS (please refer to 'Data sets import in SPSS' document for detailed explanation).

Introduction

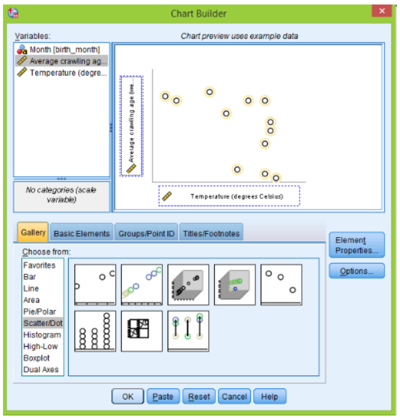
We start with the 'Babies Crawling' data set and we want to investigate the relationship between the temperature and the average crawling age in weeks. Once you open the original data file you will notice that the temperature is in Fahrenheit but we want to have it in Celsius, to make this small transformation please refer to the 'Data sets import in SPSS' document. Hence we assume that the temperature is in Celsius. Next let's make a scatter plot of the 'Average crawling age' versus 'Temperature', go to **Graphs** > **Chart Builder**



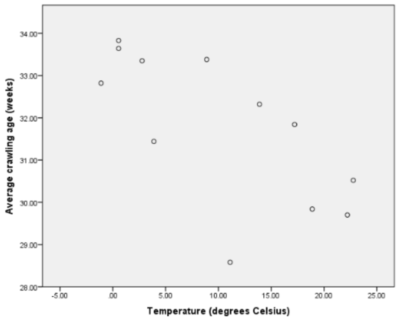
Choose **Scatter/Dot** > **double click on Simple Scatter**



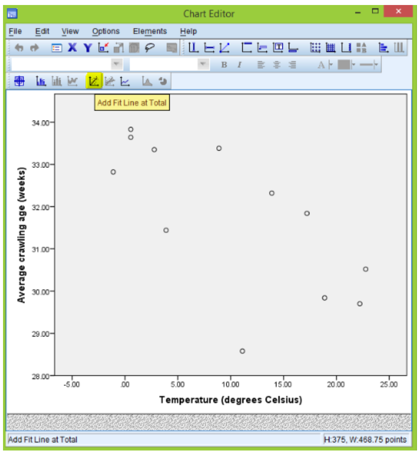
Then drag 'Temperature' to the horizontal axis and 'Average crawling age' to the vertical axis:



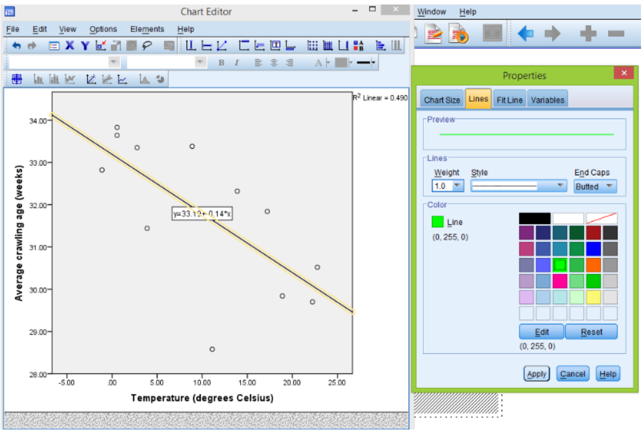
Click **OK** to get the following scatterplot:



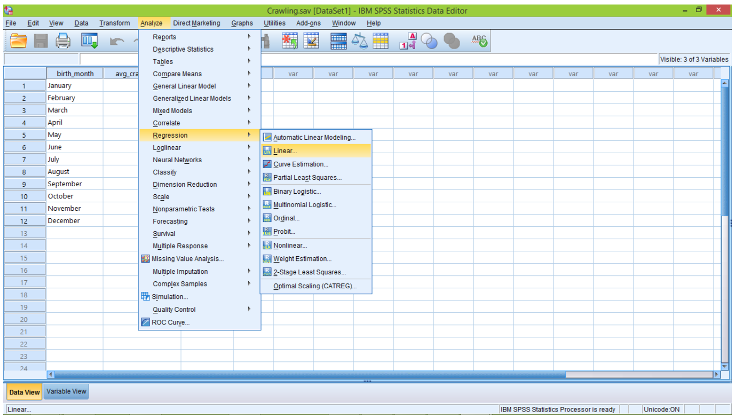
It seems that linear regression model can be appropriate in this case and we want to plot these data with the regression line. It is very simple to do that in SPSS. Just **double click on the last plot** to open ‘Chart Editor’ and then click on the symbol with two axis and diagonal line:



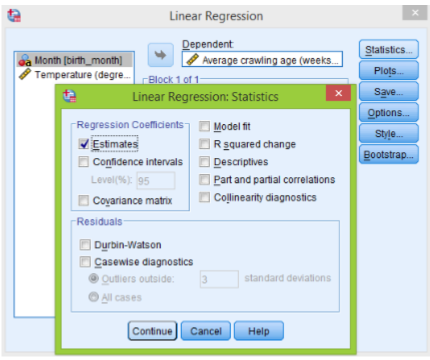
Immediately the regression line appears with regression coefficients. In the ‘Properties’ window select ‘Lines’ to change visual characteristics of the line:



The regression coefficients are given in the plot, however there is another way to get them. We will use this approach very frequently in the next sections: first go to **Analyze** > **Regression** > **Linear**



Move 'Average crawling age' to 'Dependent' section using arrow and similarly 'Temperature' to 'Independent' variables. Then click on 'Statistics', and select 'Estimates' and deselect other options:



Click **Continue then OK** and we get the following table:

Regression

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Temperature (degrees Celsius)[b]	.	Enter

a. Dependent Variable: Average crawling age (weeks)

b. All requested variables entered.

Coefficients[a]

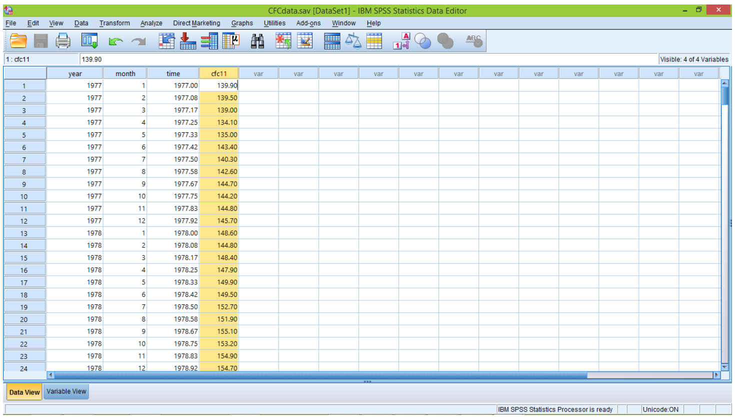
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	33.190	.596		55.716	.000
	Temperature (degrees Celsius)	-.140	.045	-.700	-3.097	.011

a. Dependent Variable: Average crawling age (weeks)

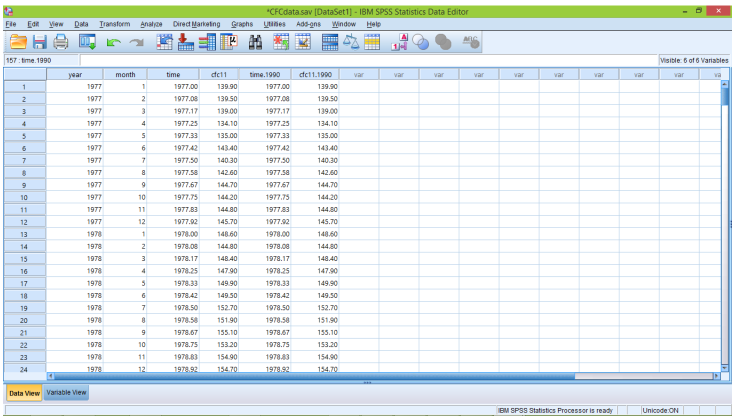
Under 'Unstandardized coefficients' we see that b_0 (intercept) is 33.190 and b_1 (slope) is -0.140. Hence we produce exactly the same coefficients as on the plot.

Some caution

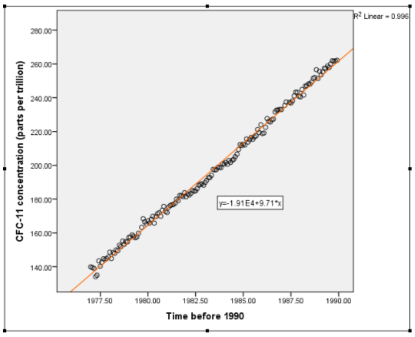
We start this section with the 'CFC' data set.



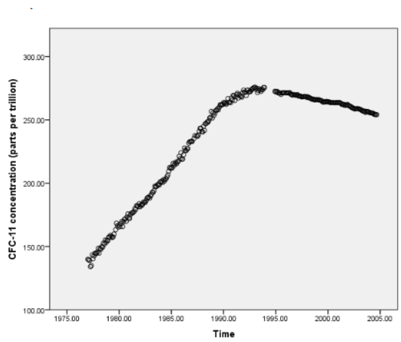
The goal is to investigate the relationship between 'time' and 'cfc11'. These data contain information till 2005 but we first want to analyse relationship before 1990, which are observations from 1 to 156. To do that we copy 'time' and 'cfc11' variables to new columns and call them 'time.1990' and 'cfc.1990' and then manually delete all the observations below 156th positon.



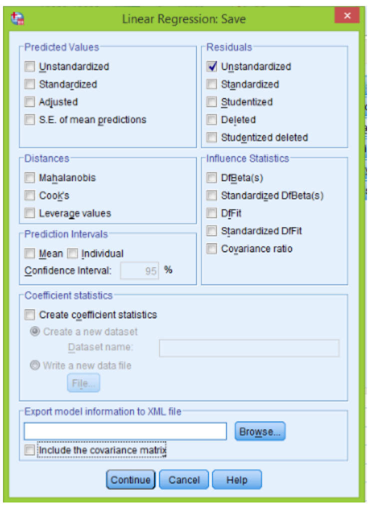
Now we create a scatterplot with linear regression line for these two variables ('cfc11.1990' is the response while 'time.1990' is the predictor)



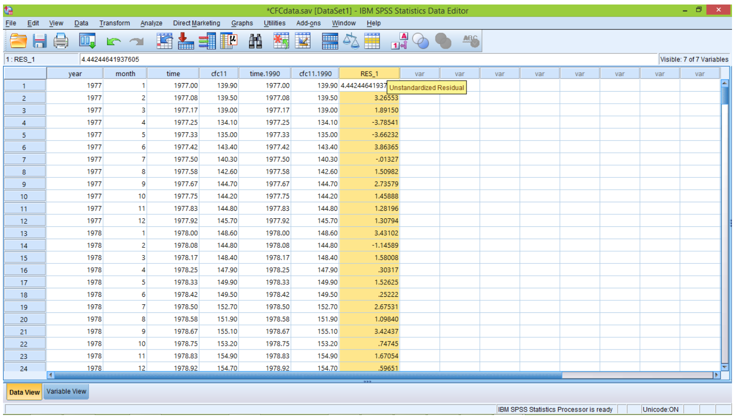
It seems that linear regression fits quite well. Now lets plot all the data points ('cfc11' versus 'time'):



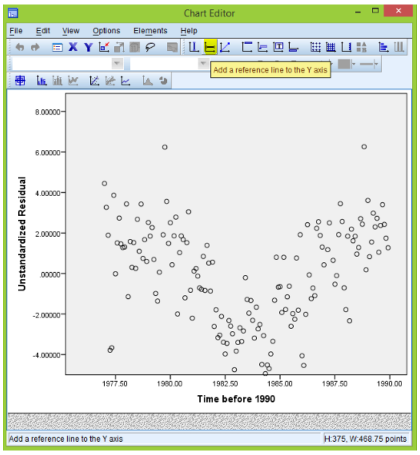
We clearly see that the linear model is not appropriate here. A very important part in checking whether a linear regression is appropriate or not is to plot residuals versus independent variable. First got to **Analyze** > **Regression** > **Linear**, then move 'cfc11.1990' to dependent section and 'time.1990' to independent one, then click on 'Save' button and under 'Residuals' select 'Unstandardized':



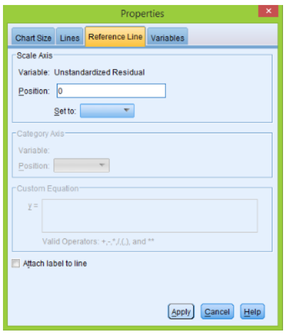
Click **Continue then OK**, that procedure produces a new column of residuals:



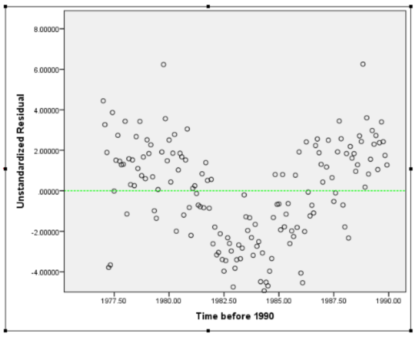
Then make a scatterplot of these residuals versus ‘time.1990’ as explained earlier, **double click on the plot** to open ‘Chart Editor’ and click on the symbol with horizontal line to add reference line to the plot:



In the ‘Properties’ window enter ‘0’ position

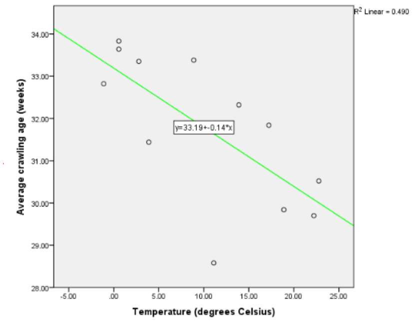


To finish click **Apply** and close the ‘Chart Editor’ to get the residual plot:

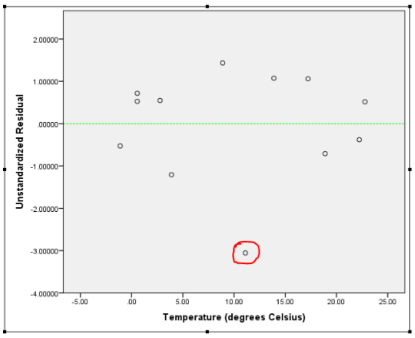


This plot also shows that there is a problem with simple regression since we observe some pattern in the residual plot.

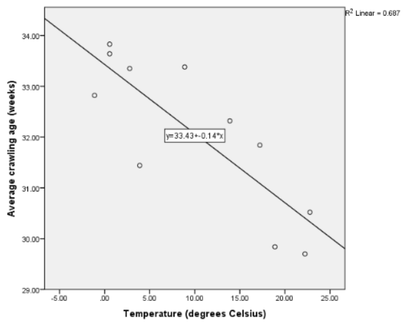
Next we move back to the 'Average crawling age' data. In the last section we have already created the next plot:

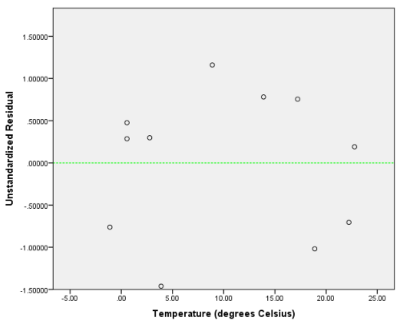


Doing exactly the same procedure as explained above we produce residual plot for this regression model:



Based on the above plot we observe one observation which has lowest residual and might be an influential point. Hence we want to make the analysis again but without this observation. To do that we just delete the fifth observation and then construct the scatterplot with regression line and residual plot for the modified data set:

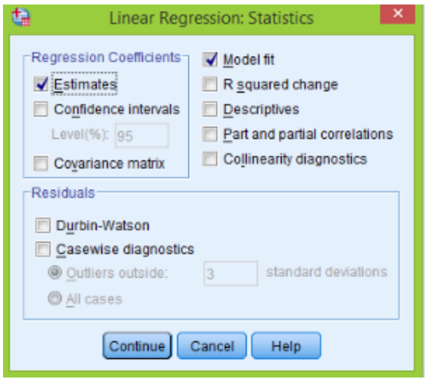




Since the coefficients have not changed by much we cannot say that the removed observation is influential.

The coefficient of determination

An important question in the regression analysis is to find how well a regression line fits the data. One measure of the fit is the coefficient of determination or R^2. Consider first the 'Average crawling age' data. We want to find R^2, sum of squares total, sum of squares regression and sum of squares residuals. It is very easy to find all these statistics in SPSS. As usual go to **Analyze** > **Regression** > **Linear**, choose appropriate dependent and independent variables, then click on the 'Statistics' button. In this window in addition to 'Estimates' also select 'Model fit':



Click **Continue** > **OK** to get the next table:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.700[a]	.490	.439	1.31920

a. Predictors: (Constant), Temperature (degrees Celsius)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16.693	1	16.693	9.592	.011[b]
	Residual	17.403	10	1.740		
	Total	34.096	11			

a. Dependent Variable: Average crawling age (weeks)

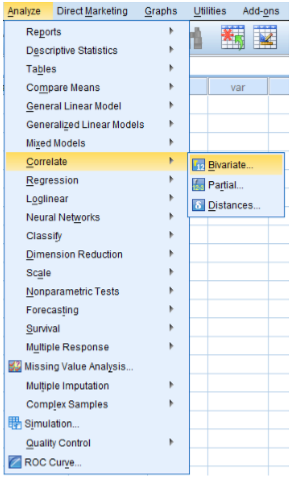
b. Predictors: (Constant), Temperature (degrees Celsius)

Coefficients[a]

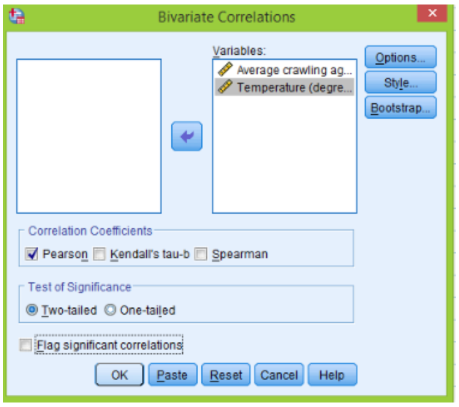
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	33.190	.596		55.716	.000
	Temperature (degrees Celsius)	-.140	.045	-.700	-3.097	.011

a. Dependent Variable: Average crawling age (weeks)

Now in addition to coefficients we have much more information. Under 'Model summary' we see that R^2 is 0.49 also under 'ANOVA' we have all the sums of squares. Also in a simple linear regression, R^2 should be the same as correlation squared. Let's find the correlation between 'Temperature' and 'Average crawling age': go to **Analyze** > **Correlate** > **Bivariate**:



Move these two variables across using arrow, make sure that 'Pearson' correlation is selected:



Click **OK**

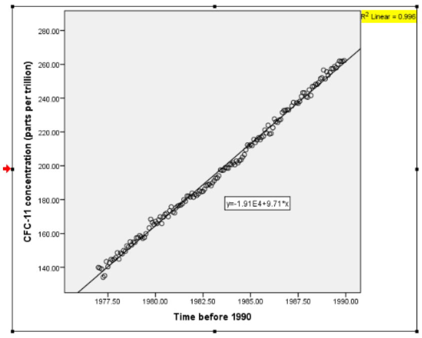
Correlations

Correlations

		Average crawling age (weeks)	Temperature (degrees Celsius)
Average crawling age (weeks)	Pearson Correlation	1	-.700
	Sig. (2-tailed)		.011
	N	12	12
Temperature (degrees Celsius)	Pearson Correlation	-.700	1
	Sig. (2-tailed)	.011	
	N	12	12

This table shows that correlation between two variables is -0.700. If you square this number you get exactly the same R^2.

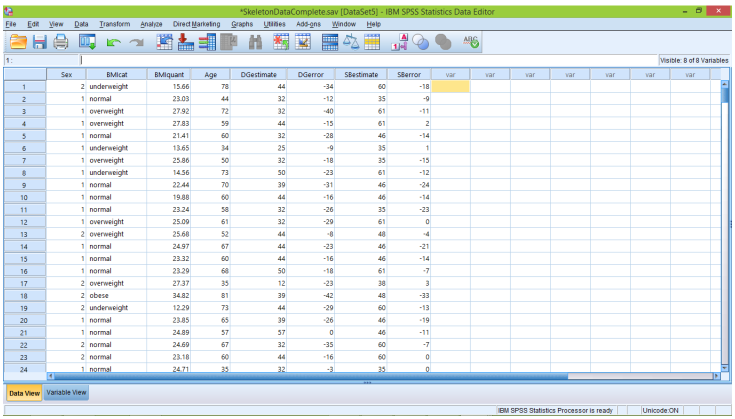
To finish this section let us return to the 'CFC11' data set. We focus on the data before 1990. Even though we know that linear regression is not appropriate for these data, lets get R^2 anyway. We can get it using the above procedure but if we just make a scatterplot with regression line then SPSS shows R^2 on the graph automatically:



Hence 99.6% of variation is explained by this regression line.

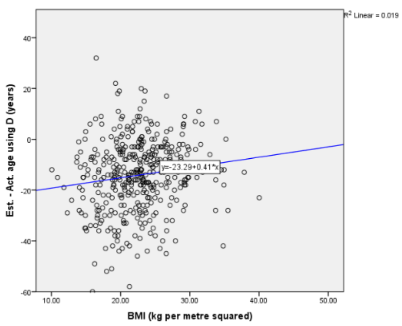
Inference for the slope

In this section we show how to test that a particular estimate (of slope or intersection) is statistically significant or not. We start with 'Skeleton' data.

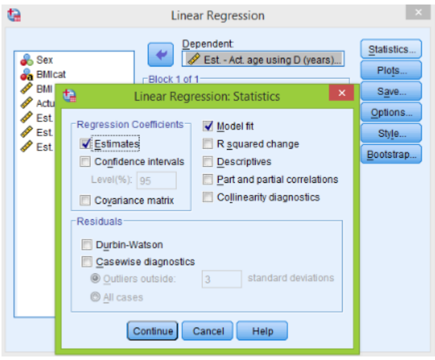


	Sex	BMIcat	BMIquant	Age	DGestimate	DGerror	SBestimate	SBerror
1	2	underweight	15.66	78	44	-34	60	-18
2	1	normal	23.03	44	32	-12	35	-9
3	1	overweight	27.92	72	32	-40	61	-11
4	1	overweight	27.83	59	44	-15	61	2
5	1	normal	21.41	60	32	-28	46	-14
6	1	underweight	13.65	34	25	-9	35	1
7	1	overweight	25.86	50	32	-18	35	-15
8	1	underweight	14.56	73	50	-23	61	-12
9	1	normal	22.44	70	39	-31	46	-24
10	1	normal	19.88	60	44	-16	46	-14
11	1	normal	23.24	58	32	-26	35	-23
12	1	overweight	25.09	61	32	-29	61	0
13	2	overweight	25.68	52	44	-8	48	-4
14	1	normal	24.97	67	44	-23	46	-21
15	1	normal	23.32	60	44	-16	46	-14
16	1	normal	23.29	68	50	-18	61	-7
17	2	overweight	27.37	35	12	-23	38	3
18	2	obese	34.82	81	39	-42	48	-33
19	2	underweight	12.29	73	44	-29	60	-13
20	1	normal	23.85	65	39	-26	46	-19
21	1	normal	24.89	57	57	0	46	-11
22	2	normal	24.69	67	32	-35	60	-7
23	2	normal	23.18	60	44	-16	60	0
24	1	normal	24.71	35	32	-3	35	0

Our response in this analysis would be 'DGerror' variable, and independent variable is 'BMIquant'. Doing the standard procedures we obtain the scatterplot with the regression line:



Hence the slope is 0.41, intercept is -23.29 and R^2 is 0.019. This means that less than 2% of variation of the response is explained by this regression. But we have also a very important question: is the slope statistically significant? Because if it is not, then 'BMIquant' is not important for the prediction of 'DGerror'. We can easily answer this question (and not only for the slope but also for the intercept) if we go to **Analyze** > **Regression** > **Linear**. Move 'DGerror' to dependent window and 'BMIquant' to independent one. Then click on 'Statistics' and make sure that 'Estimates' are selected:



Click **Continue** > **OK** to get the usual table:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.136[a]	.019	.016	14.011

a. Predictors: (Constant), BMI (kg per metre squared)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1483.241	1	1483.241	7.556	.006[b]
	Residual	78131.759	398	196.311		
	Total	79615.000	399			

a. Dependent Variable: Est. - Act. age using D (years)

b. Predictors: (Constant), BMI (kg per metre squared)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-23.288	3.397		-6.855	.000
	BMI (kg per metre squared)	.406	.148	.136	2.749	.006

a. Dependent Variable: Est. - Act. age using D (years)

The p-values are displayed under 'Sig.' title. Hence we note that the p-value for the slope is 0.006 which is quite small and therefore we conclude that 'BMIquant' variable is important for prediction and we should not ignore it.

Now we return to the 'Crawling' data set. We want to check whether temperature really effects the average crawling age or not. Doing similar procedure we get:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.700[a]	.490	.439	1.31920

a. Predictors: (Constant), Temperature (degrees Celsius)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16.693	1	16.693	9.592	.011[b]
	Residual	17.403	10	1.740		
	Total	34.096	11			

a. Dependent Variable: Average crawling age (weeks)

b. Predictors: (Constant), Temperature (degrees Celsius)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	33.190	.596		55.716	.000
	Temperature (degrees Celsius)	-.140	.045	-.700	-3.097	.011

a. Dependent Variable: Average crawling age (weeks)

Based on the output we conclude that the p-value for the slope is 0.011 which can be considered as small and therefore temperature is statistically significant.

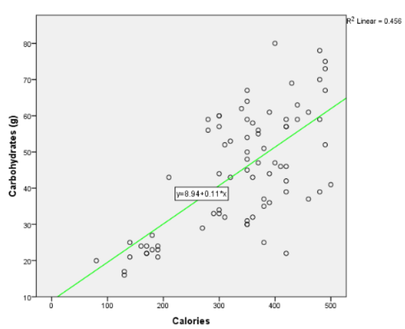
Checking for conditions

In this section we focus on the new data set 'Coffee Shop'. The first 5 observations are shown below

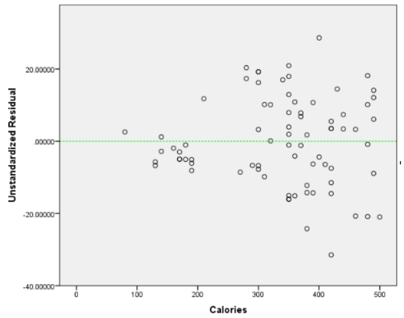
	calories	carb	type
1	350	67	bakery
2	350	64	bakery
3	420	59	bakery
4	490	75	bakery
5	130	17	bakery

In this example we want to fit a linear regression with 'carb' (carbohydrates) as the response and 'calories' variable as the predictor. We get scatterplot with regression line in a usual way:

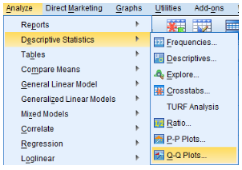
GGraph



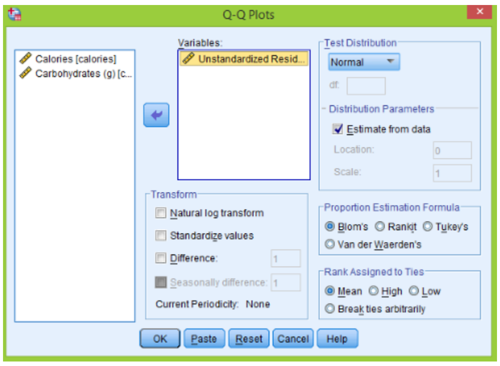
Regression analysis is not complete without the residual plot, so we make residuals versus 'calories' scatterplot with horizontal dashed line: (remember that we get residuals from **Analyze** > **Regression** > **Linear** then click on 'Save' and select 'Unstandardized' to produce column of residuals)



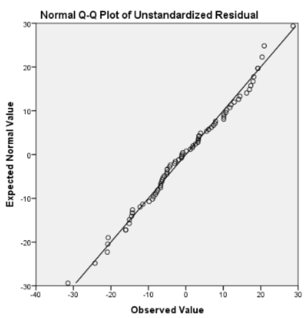
We immediately see a problem. The variance is not constant and increases as the 'calories' increase. Hence it is not appropriate to carry out inference on the slope of the regression line in this case. Lets also make a quantile-quantile plot of the residuals: go to **Analyze** > **Descriptive Statistics** > **Q-Q Plots**



Then move 'Residuals' variable across using arrow:



Click **OK** and two plots appear in the output window but we need only the first one:

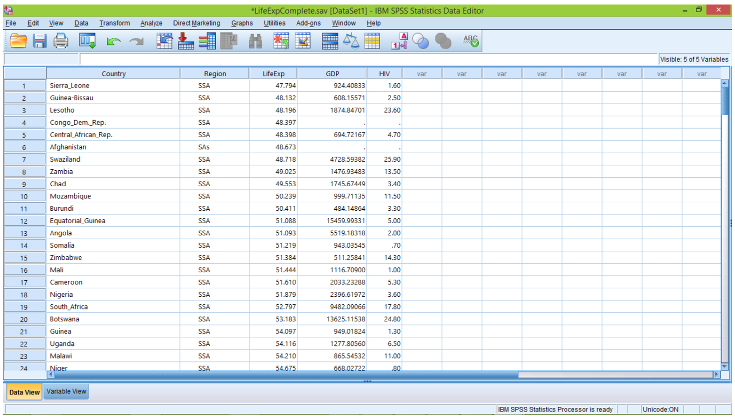


The points on this plot should lie on a straight line (and that would indicate that residuals have normal distribution). In this example the plot is generally straight with some small departure from linearly in the right tail.

Transformations

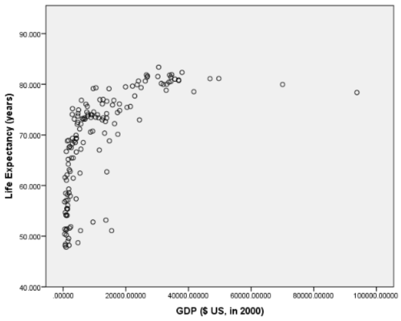
As shown in the last section, linear regression is not appropriate in some cases. In this section we show how to transform predictor and/or response to make linear regression valid. Consider the

‘Life Expectancy’ data set.

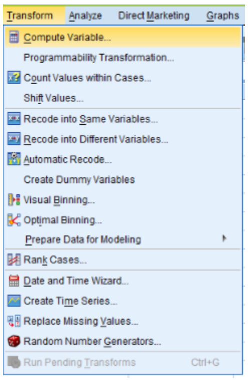


	Country	Region	LifeExp	GDP	HIV
1	Sierra_Leone	SSA	47.794	924.40833	1.60
2	Guinea-Bissau	SSA	48.132	608.15571	2.50
3	Lesotho	SSA	48.196	1874.84701	23.60
4	Congo_Dem._Rep.	SSA	48.397	.	.
5	Central_African_Rep.	SSA	48.398	694.72167	4.70
6	Afghanistan	SAs	48.673	.	.
7	Swaziland	SSA	48.718	4728.59382	25.90
8	Zambia	SSA	49.025	1476.93483	13.50
9	Chad	SSA	49.553	1745.67449	3.40
10	Mozambique	SSA	50.239	999.71135	11.50
11	Burundi	SSA	50.411	484.14864	3.30
12	Equatorial_Guinea	SSA	51.088	15459.99331	5.00
13	Angola	SSA	51.093	5519.18318	2.00
14	Somalia	SSA	51.219	943.03545	.70
15	Zimbabwe	SSA	51.384	511.25841	14.30
16	Mali	SSA	51.444	1116.70900	1.00
17	Cameroon	SSA	51.610	2033.23288	5.30
18	Nigeria	SSA	51.879	2396.61972	3.60
19	South_Africa	SSA	52.797	9482.09066	17.80
20	Botswana	SSA	53.183	13625.11538	24.80
21	Guinea	SSA	54.097	949.01824	1.30
22	Uganda	SSA	54.116	1277.80560	6.50
23	Malawi	SSA	54.210	865.54532	11.00
24	Niger	SSA	54.675	668.02722	.80

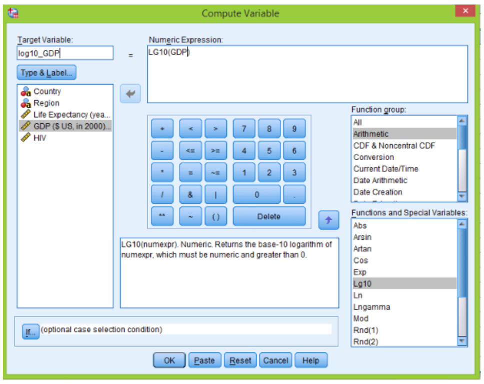
Next we plot ‘LifeExp’ versus ‘GDP’ (using the usual **Chart Builder**):



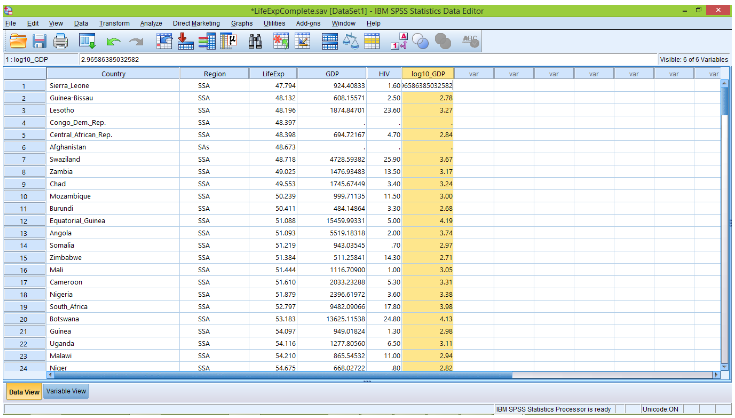
Clearly the relationship is not linear. However we see that ‘GDP’ variable has many small values and several observations are very large. Hence base 10 log transformation may help in this situation. First we want to construct a new variable ‘log10_GDP’ that will store base 10 logs of the original ‘GDP’ variable; go to **Transform** > **Compute Variable**



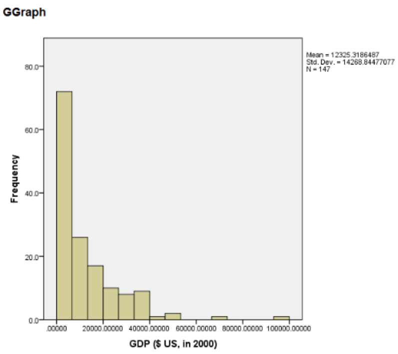
Then we call the target variable ‘log10_GDP’ and enter ‘LG10(GDP)’ which means log base 10 of ‘GDP’ variable:

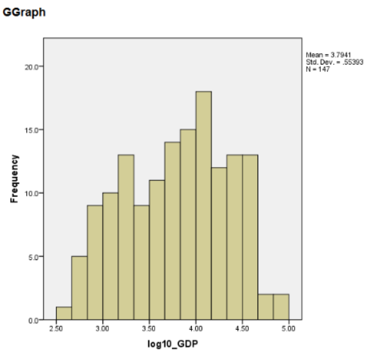


Click **OK**; and the new variable appears

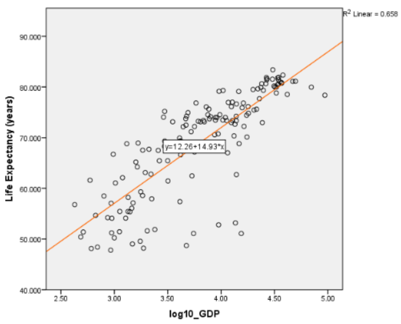


Now we make two histograms. The first one is histogram of the original variable ‘GDP’, second one of the ‘log10_GDP’:



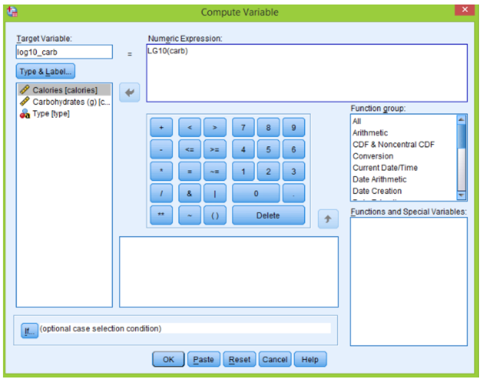


See how the distribution changed from right skewed to almost symmetric. Next lets make a scatterplot of ‘LifeExp’ versus ‘log10_GDP’ with regression line:

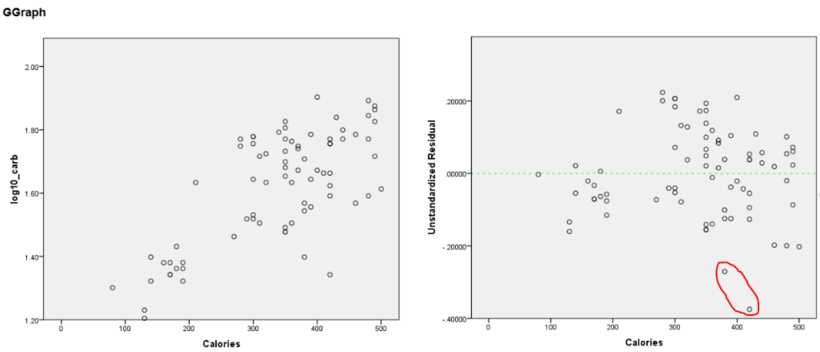


Now the plot is much more linear than the original one and linear regression model seems appropriate in this case.

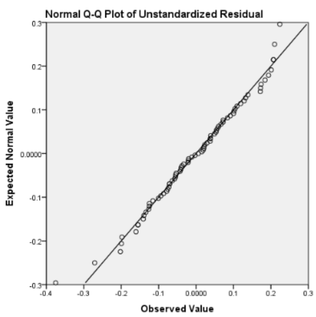
To finish this section we return to the 'Coffee Shop' data set. From the last section we remember that the variance of residuals is not constant and linear regression ('carb' versus 'calories') is not appropriate in this case. To solve this problem we try to make base 10 logarithm transformation of the response. We construct a new transformed variable 'log10_carb'



Then we make a scatterplot of transformed variable versus 'calories' and residual plot:

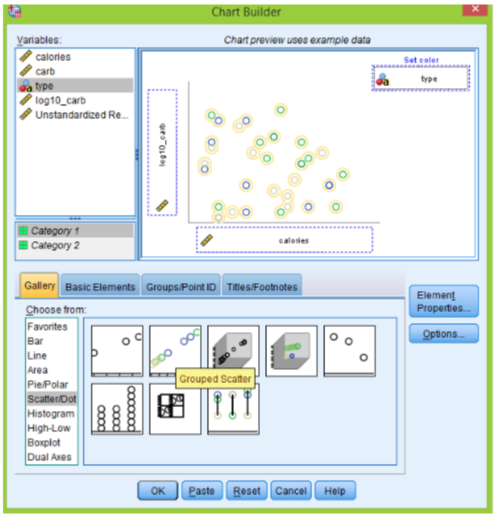


Now the variance is constant and using linear regression is appropriate (there are two large negative residuals which we will discuss later). Lets check the condition that residuals follow a normal distribution using quantile-quantile plot; as explained earlier go to **Analyze** > **Descriptive Statistics** > **Q-Q Plots** and we get the next plot:

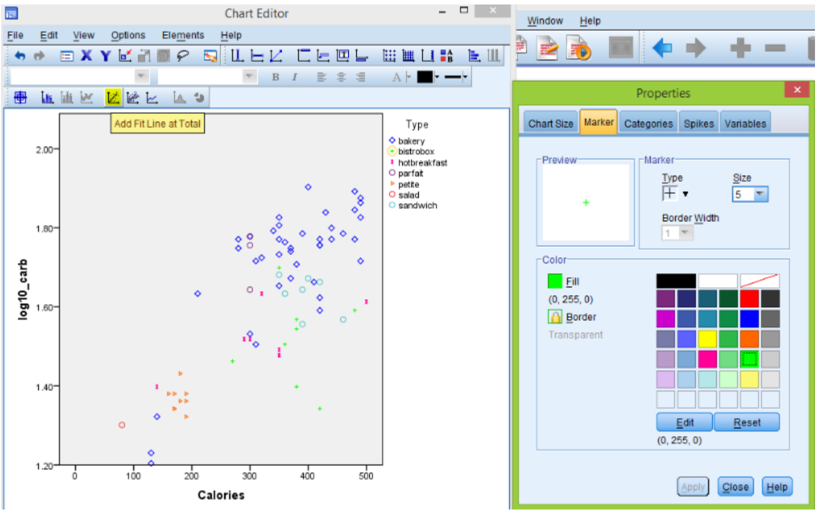


The plot looks straight and therefore we can conclude that normal assumption is satisfied.

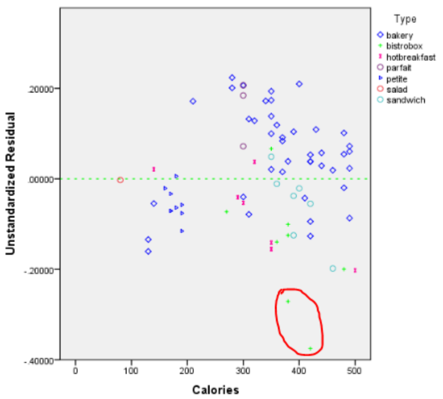
To explain two unusual observations from the residual plot, let's make a scatterplot of 'log10_carb' versus 'calories' but with different symbols and colors corresponding to the 'Type' of the food. Go to **Graphs** > **Chart Builder** > **Scatter/Dot** > **double click on Grouped Scatter** next drag 'calories' to the x-axis, 'log10_crab' to the y-axis and 'type' to 'Set color' section:



Click **OK**, the scatterplot is produced. Double click on the plot to open 'Chart Editor', to change symbols and colors for each 'Type' double click on symbols in the legend and then select 'Marker' from the 'Properties' window. To add regression line for the plot, click on 'Add Fit Line at Total' as usual:



Similar plot we do for the residuals versus 'calories'



We see that these two unusual observations correspond to 'bistrobox' items. Actually almost all the 'bistrobox' food is below the fitted line, therefore it is important to use 'type' variable in the analysis to explain relationship between 'carbohydrates' and 'calories'.