ACU
AUSTRALIAN CATHOLIC UNIVERSITY

PUBH620 Biostatistics

Week 5

Pearson's Correlation and Simple Linear Regression

Lecture notes by Dr Brandon Cheong

ELECTRONIC WARNING NOTICE

In this lecture…

(clicking the links below will direct you to the topic page)

Topic Learning Objectives (TLOs)

- Understand the basis for correlation and regression techniques, their interpretation and limitations
- Introduce the notion of causation to examine practical examples and outline measurement issues

6.1 Pearson's Correlation

Pearson's correlation test is used to determine the strength and direction of a relationship between two continuous variables.

- Also known as Pearson's product-moment correlation coefficient.
- Test statistic is denoted by "r", test can also be called Pearson's r.

In this unit, we will look at:

- Bivariate correlation, used to measure the linear relationship between two continuous variables.

5.1.1 Introduction to Pearson's correlation test

r is an important parameter in Pearson's correlation. It is the test statistic which gives us the strength and direction of the relationship between two continuous variables.

r can only take values ranging between -1 and 1.

If values are:

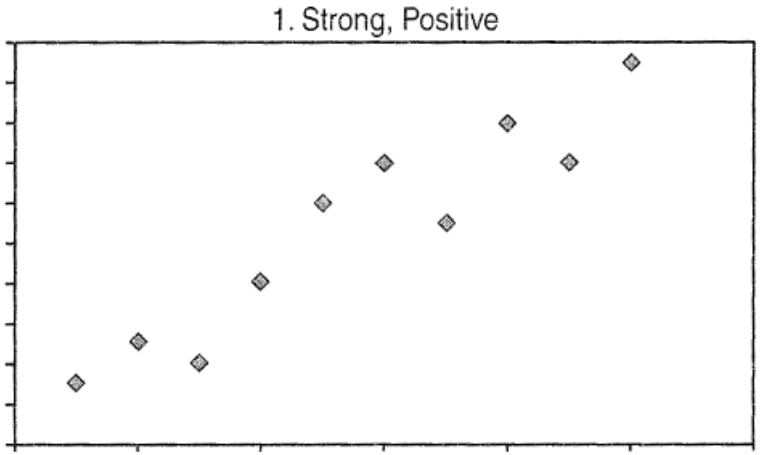
- Positive – then an increase in one variable's values will also result in an increase in the other variable's values.
- Zero – then no linear relationship exists between the two variables.
- Negative – then an increase in one variable's value will result in a decrease in the other variable's values.

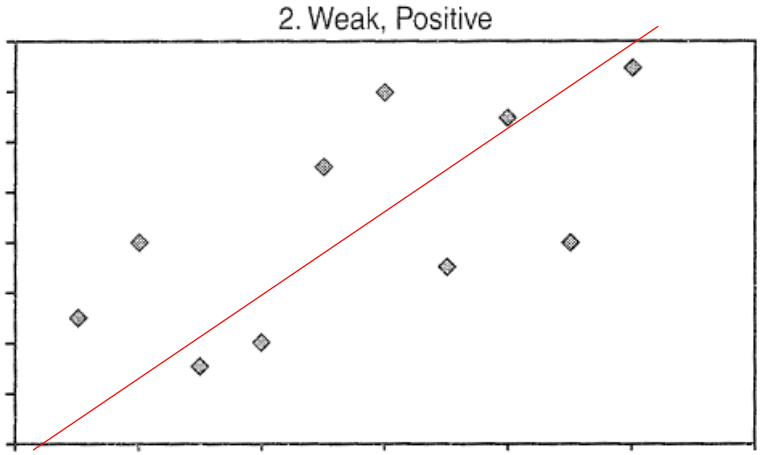
Magnitude of value indicates strength of the correlation:

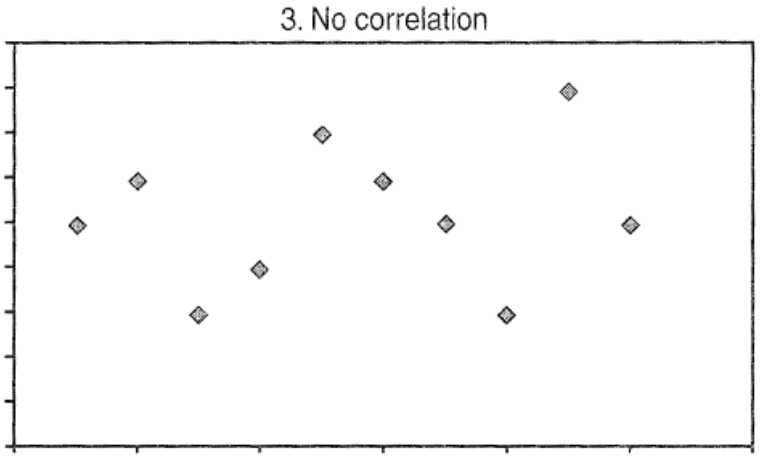
Eg. (in practice, you do not need to write the "+")

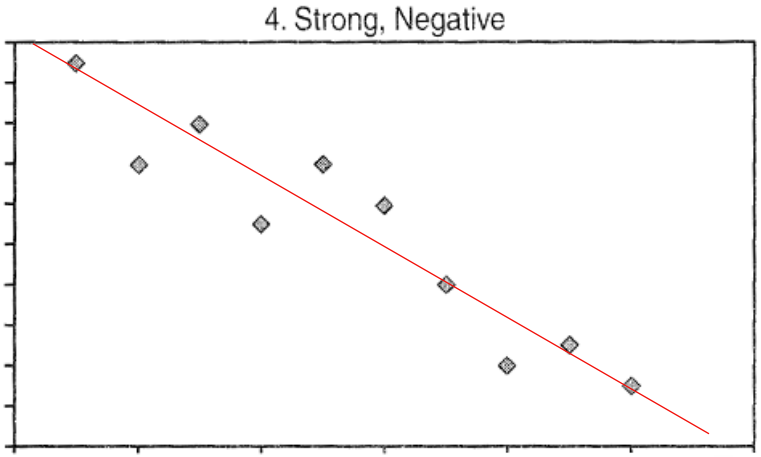
- + 0.1: weak positive correlation
- + 0.9: strong positive correlation
- - 0.1: weak negative correlation
- + 1.0: perfect positive correlation

5.1.1 Introduction to Pearson's correlation test









5.1.1 Introduction to Pearson's correlation test

Eg. Is there an association between gestational age at birth and birth weight?

Gestational age: time between estimated conception and birth

Average gestational age: 40 weeks

Less than 37 weeks: preterm

37 – 38 weeks: early term

39 – 40 weeks: full term

41 – 41 weeks: late term

More than 42 weeks: post-term

What sort of relationship do we expect to see here?

5.1.1 Introduction to Pearson's correlation test

What sort of relationship?

Positive? Negative? None?

Need to test this but first develop hypotheses.

H_0: there is no linear relationship between gestational age and birth weight.

H_1: there is linear relationship between gestational age and birth weight.

Gestational Age (weeks)	Birth weight (grams)
29.3	1410
34.9	1833
36.1	2050
35.7	3172
37.5	2601
38.2	2588
38.3	2835
38.5	3412
38.6	2009
39.6	3250
39.6	3704
40.1	2750
40.3	3217
41.0	3201
39.9	3210
41.2	3552
42.2	3736

5.1.2 Assumptions of Pearson's correlation test

Pearson's r has four assumptions:

- Independence – Each participant (baby) should be independent of each other and cannot participate more than once.
- Normality – Each variable should be approximately normally distributed.
- Linearity – There should be a linear relationship between the two variables.
- Homoscedasticity – Both variables should have approximately equal variances.

Assumption 1 is satisfied at the data collection stage. Assumptions 2, 3 and 4 can be done using SPSS.

5.1.2 Assumptions of Pearson's correlation test

Setting up the data in SPSS

Normality assumption

Run a KS test.

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distribution of GestationalAge is normal with mean 38.29 and standard deviation 3.052.	One-Sample Kolmogorov-Smirnov Test	.090[1]	Retain the null hypothesis.
2	The distribution of BirthWeight is normal with mean 2,854.71 and standard deviation 687.832.	One-Sample Kolmogorov-Smirnov Test	.051[1]	Retain the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

[1] Lilliefors Corrected

PearsonsExample.sav [DataSet0] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Ut

	GestationalAge	BirthWeight
1	29.30	1410.00
2	34.90	1833.00
3	36.10	2050.00
4	35.70	3172.00
5	37.50	2601.00
6	38.20	2588.00
7	38.30	2835.00
8	38.50	3412.00
9	38.60	2009.00
10	39.60	3250.00
11	39.60	3704.00
12	40.10	2750.00
13	40.30	3217.00
14	41.00	3201.00
15	39.90	3210.00
16	41.20	3552.00
17	42.20	3736.00
18		

According to the KS test, both variables are normal.

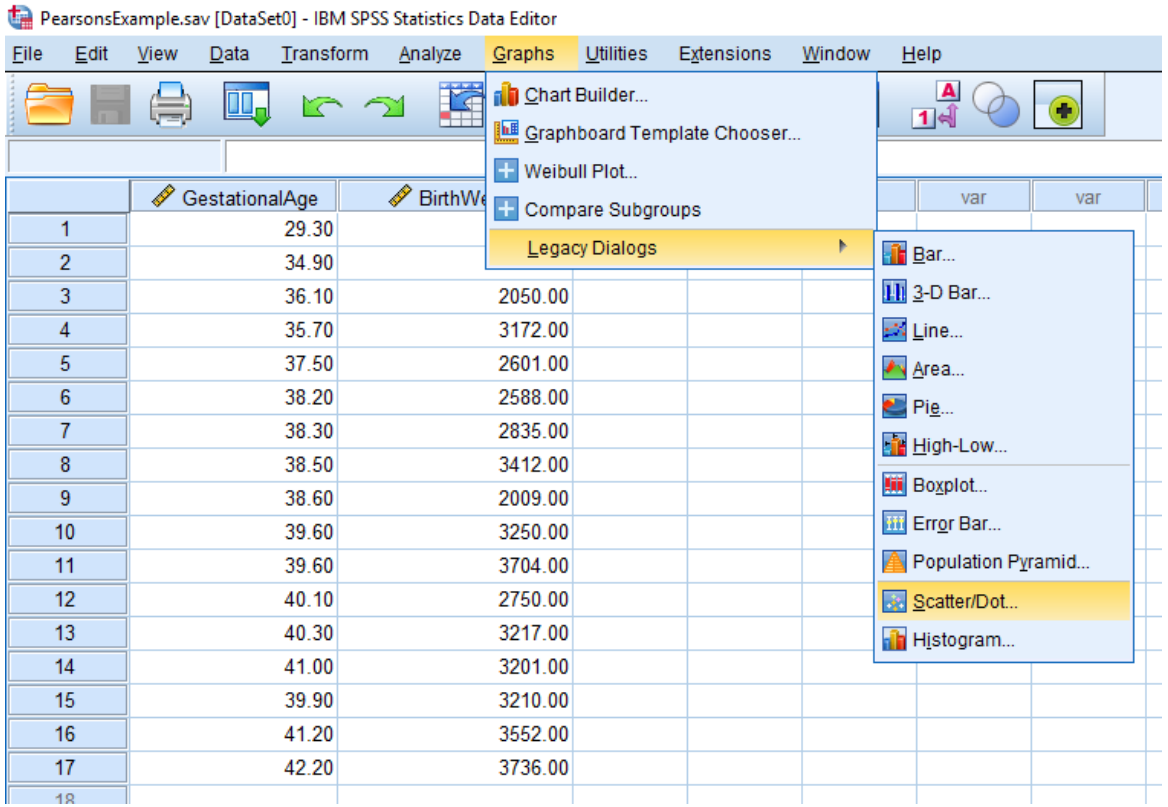
5.1.2 Assumptions of Pearson's correlation test

Assumptions for linearity and homoscedasticity

<Graphs>

<Legacy Dialogs>

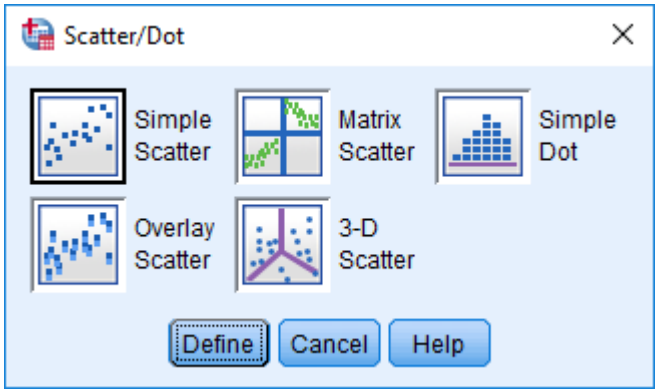
<Scatter/Dot…>



5.1.2 Assumptions of Pearson's correlation test

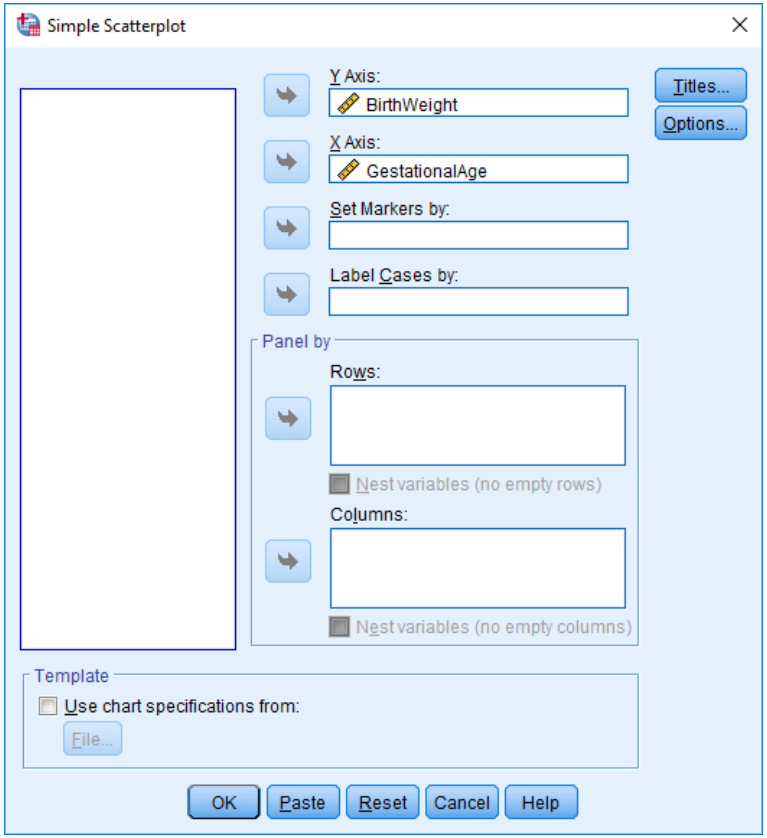
Assumptions for linearity and **homoscedasticity**

Choose Simple Scatter and click Define



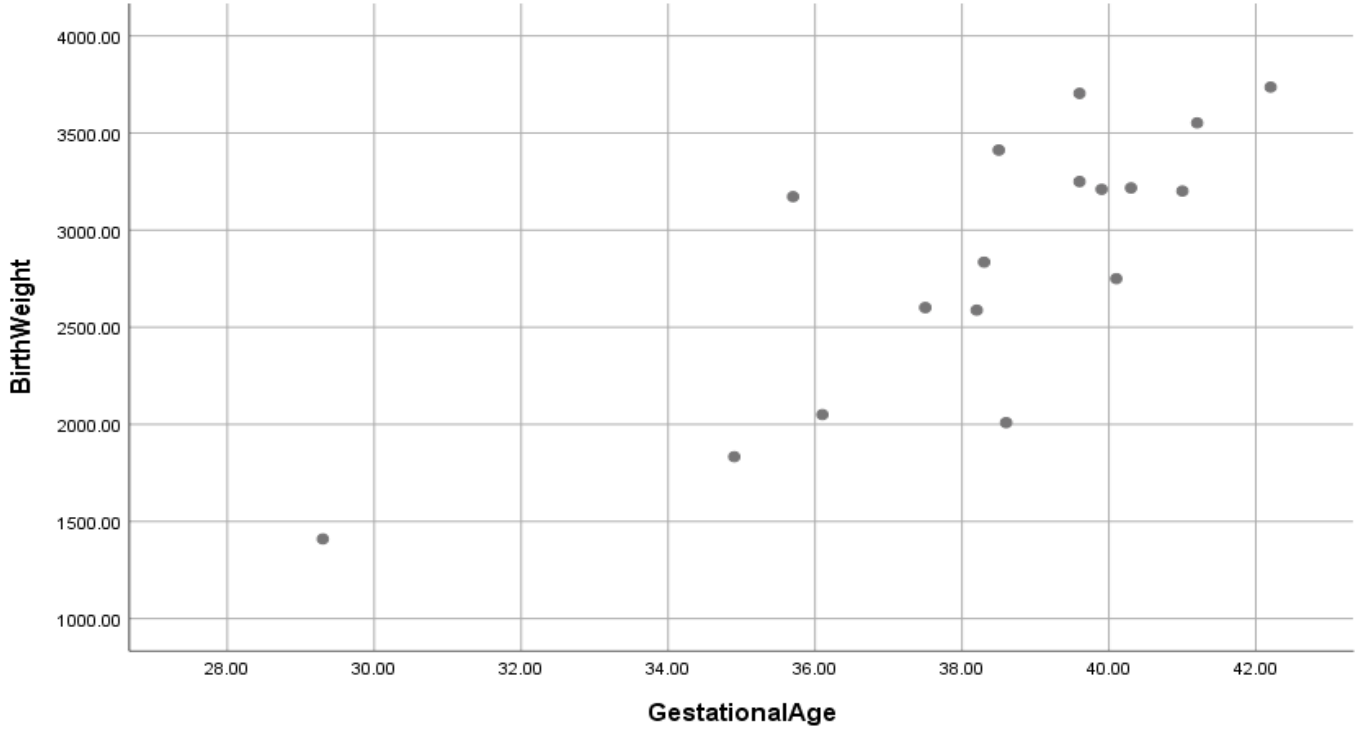
GestationalAge as x-axis
BirthWeight as y-axis

Click OK



5.1.2 Assumptions of Pearson's correlation test

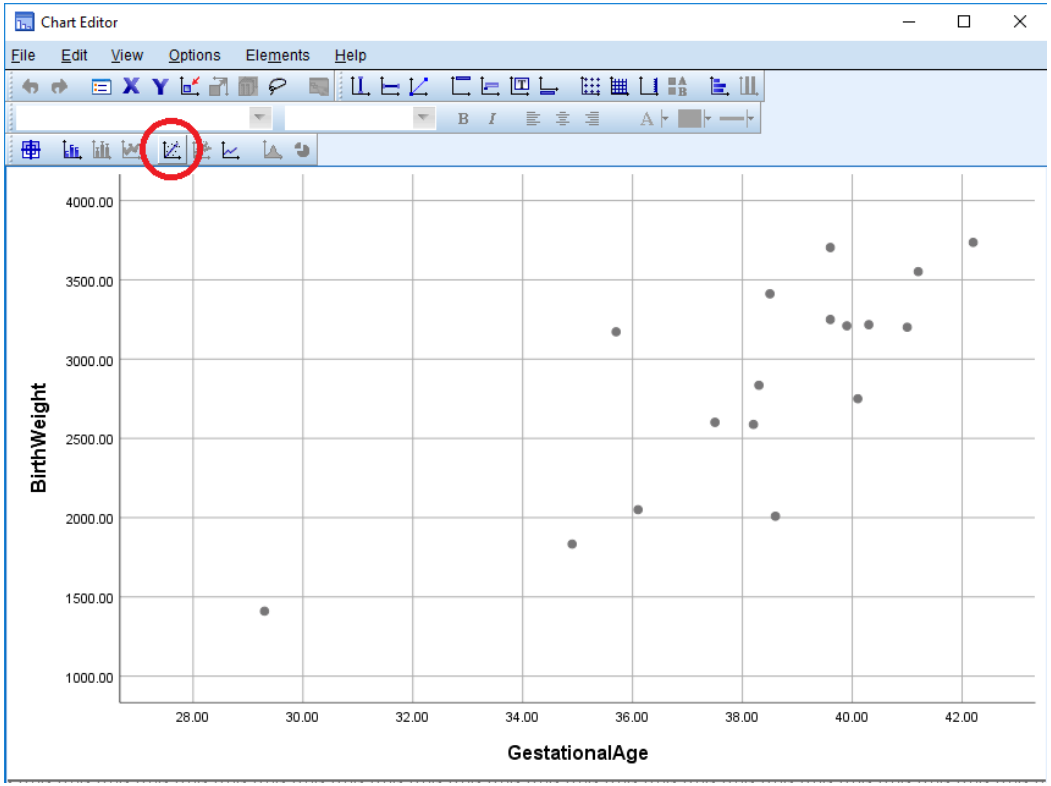
Assumptions for linearity and homoscedasticity



5.1.2 Assumptions of Pearson's correlation test

Assumptions for linearity and homoscedasticity

Add a line of best fit – double click plot and click the button that is circled in red.

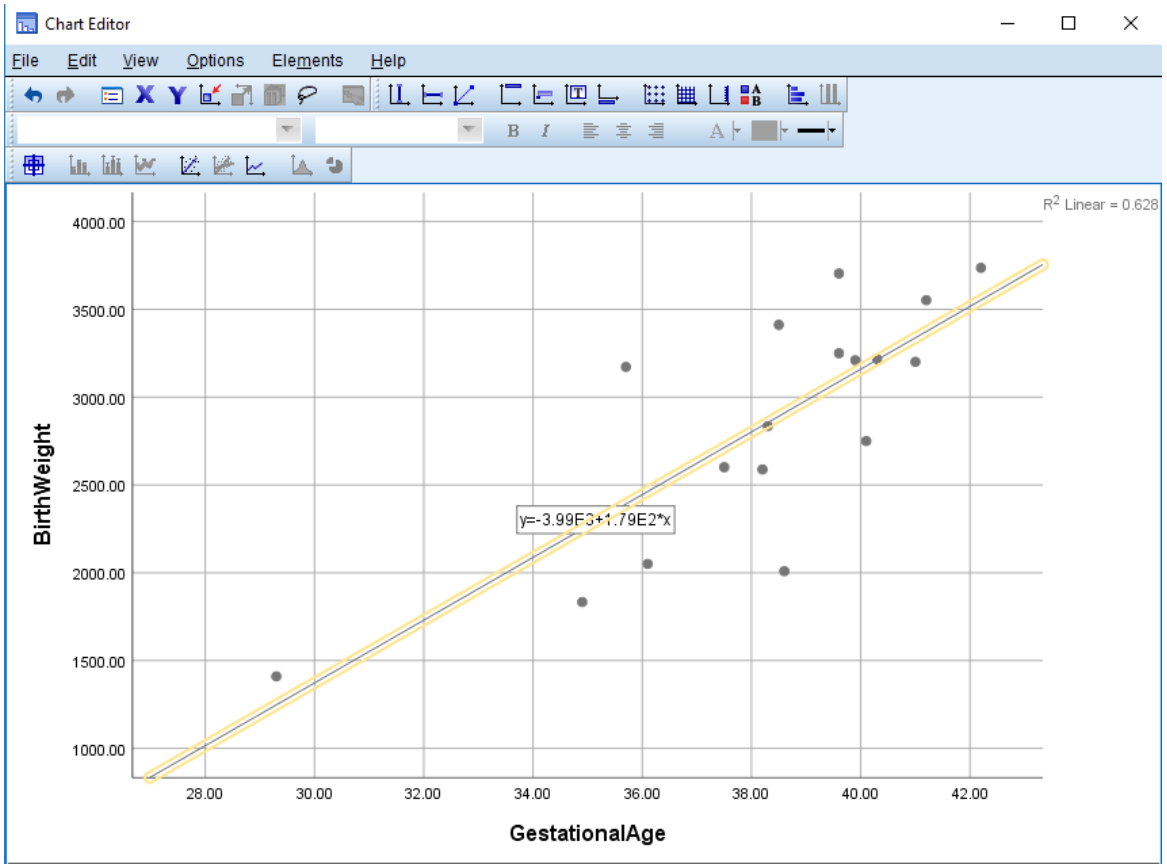


5.1.2 Assumptions of Pearson's correlation test

Assumptions for linearity and homoscedasticity

SPSS will give you a line of best fit, the equation of the line and a r^2 value. The r^2 value is known as the effect size and gives an indication of how well the data fits the linear model.

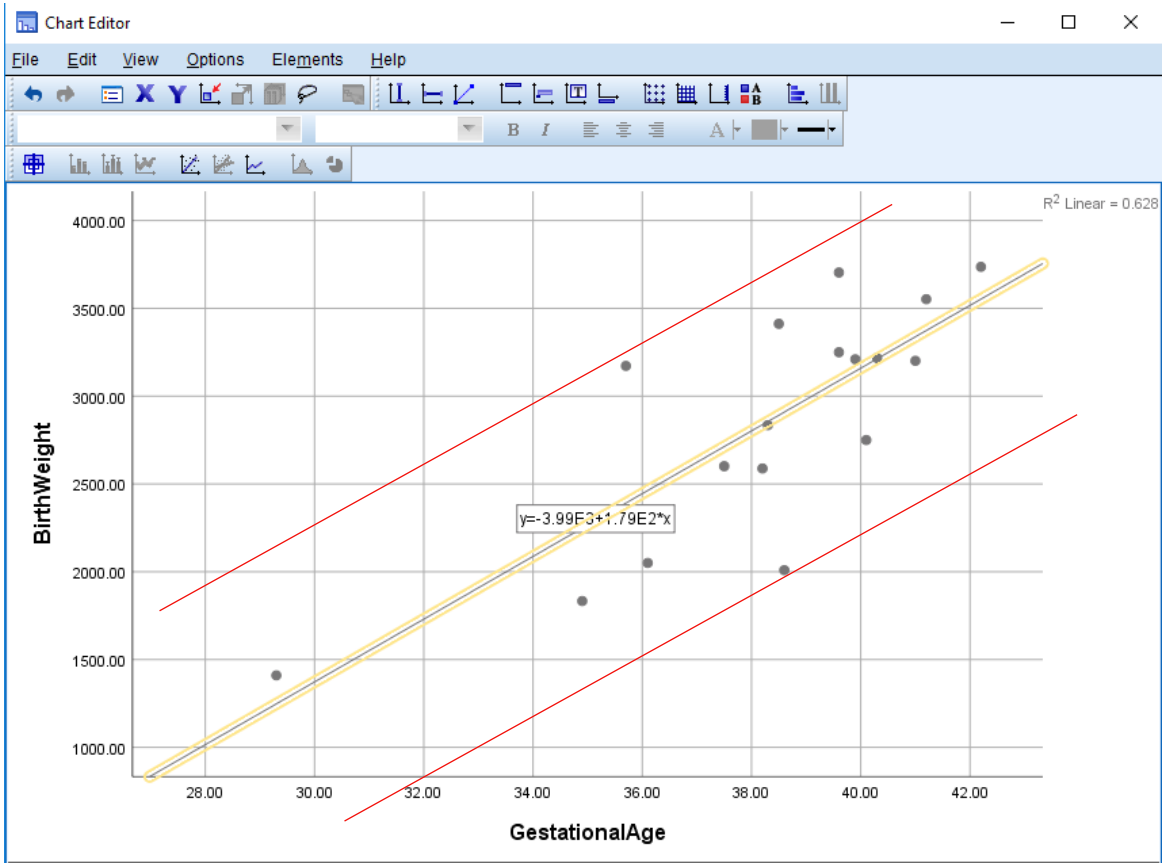
In fact, taking the square root of the r^2 value gives us the test statistic r!



5.1.2 Assumptions of Pearson's correlation test

Assumptions for linearity and homoscedasticity

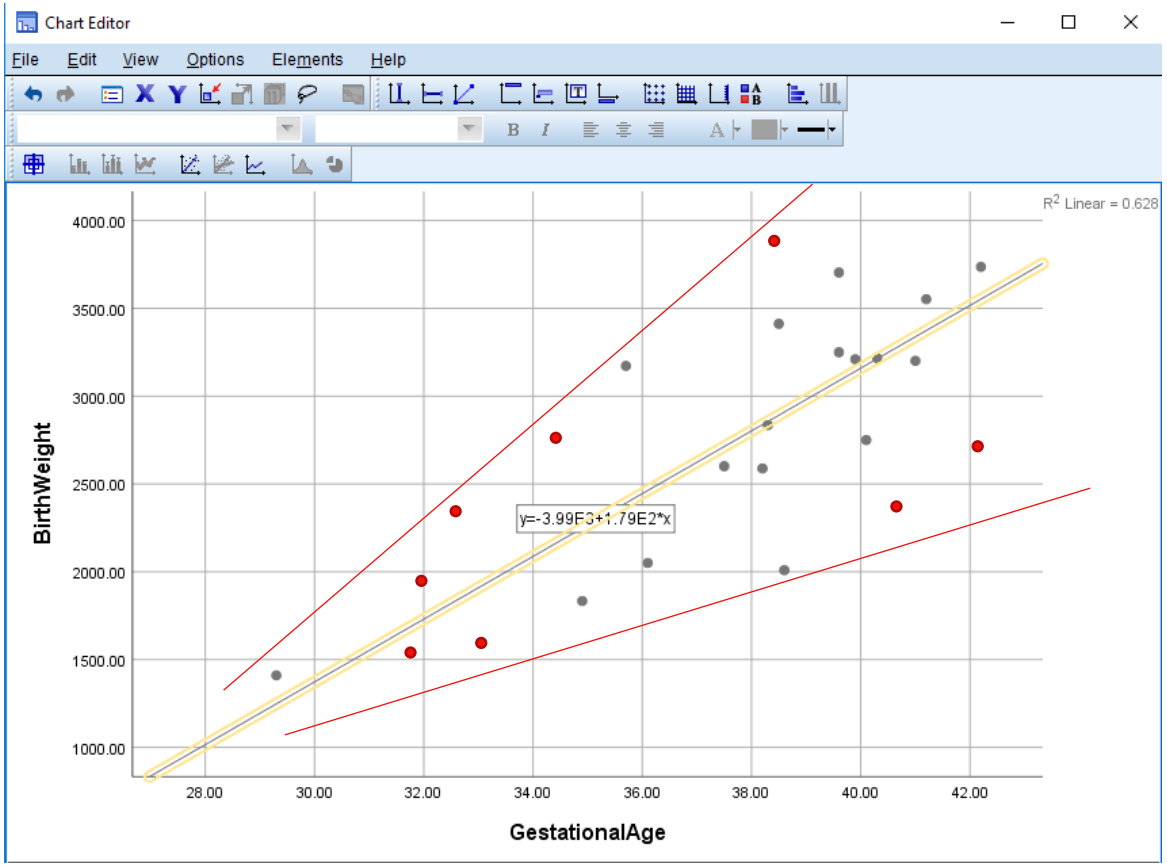
From observation we can see that that the linear plot has approximately equal variance (spread). This is usually done by drawing in lines that are parallel to the line of best fit. This can be rather subjective.



5.1.2 Assumptions of Pearson's correlation test

Assumptions for linearity and homoscedasticity

For example if there were more data points (red dots)..this might be viewed as a heteroscedastic test. In this case…we would need to use a non-parametric test called the Spearman's rho or Kendall's tau (Week 11). This further emphasises the importance of having a large sample size when answering a research question!



5.1.3 Calculating the test statistic r for Pearson's correlation test

The formula to calculate the Pearson's r is:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Where x and y represents each variable and n = the number of pairs of data. As we are looking at correlation only, we are only interested in the relationship between the two variables so it doesn't matter which variable is x and which is y as long as you keep them consistent.

$\sum$ is sigma and represents "sum of".

5.1.3 Calculating the test statistic r for Pearson's correlation test

Assign x for gestational age and y for birth weight.

Create 3 more columns for xy, x^2 and y^2.

x (weeks)	y (grams)
29.3	1410
34.9	1833
36.1	2050
35.7	3172
37.5	2601
38.2	2588
38.3	2835
38.5	3412
38.6	2009
39.6	3250
39.6	3704
40.1	2750
40.3	3217
41.0	3201
39.9	3210
41.2	3552
42.2	3736

5.1.3 Calculating the test statistic r for Pearson's correlation test

Use Excel!

x (weeks)	y (grams)	xy	x^2	y^2
29.3	1410	41313	858.49	1988100
34.9	1833	63971.7	1218.01	3359889
36.1	2050	74005	1303.21	4202500
35.7	3172	113240.4	1274.49	10061584
37.5	2601	97537.5	1406.25	6765201
38.2	2588	98861.6	1459.24	6697744
38.3	2835	108580.5	1466.89	8037225
38.5	3412	131362	1482.25	11641744
38.6	2009	77547.4	1489.96	4036081
39.6	3250	128700	1568.16	10562500
39.6	3704	146678.4	1568.16	13719616
40.1	2750	110275	1608.01	7562500
40.3	3217	129645.1	1624.09	10349089
41.0	3201	131241	1681	10246401
39.9	3210	128079	1592.01	10304100
41.2	3552	146342.4	1697.44	12616704
42.2	3736	157659.2	1780.84	13957696
$\sum$ 651	48530	1885039	25078.5	146108674

5.1.3 Calculating the test statistic r for Pearson's correlation test

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

Use Excel!

$$r = \frac{17(1885039) - (651)(48530)}{\sqrt{[17(25078.5) - (651)^2][17(146108674) - (48530)^2]}}$$

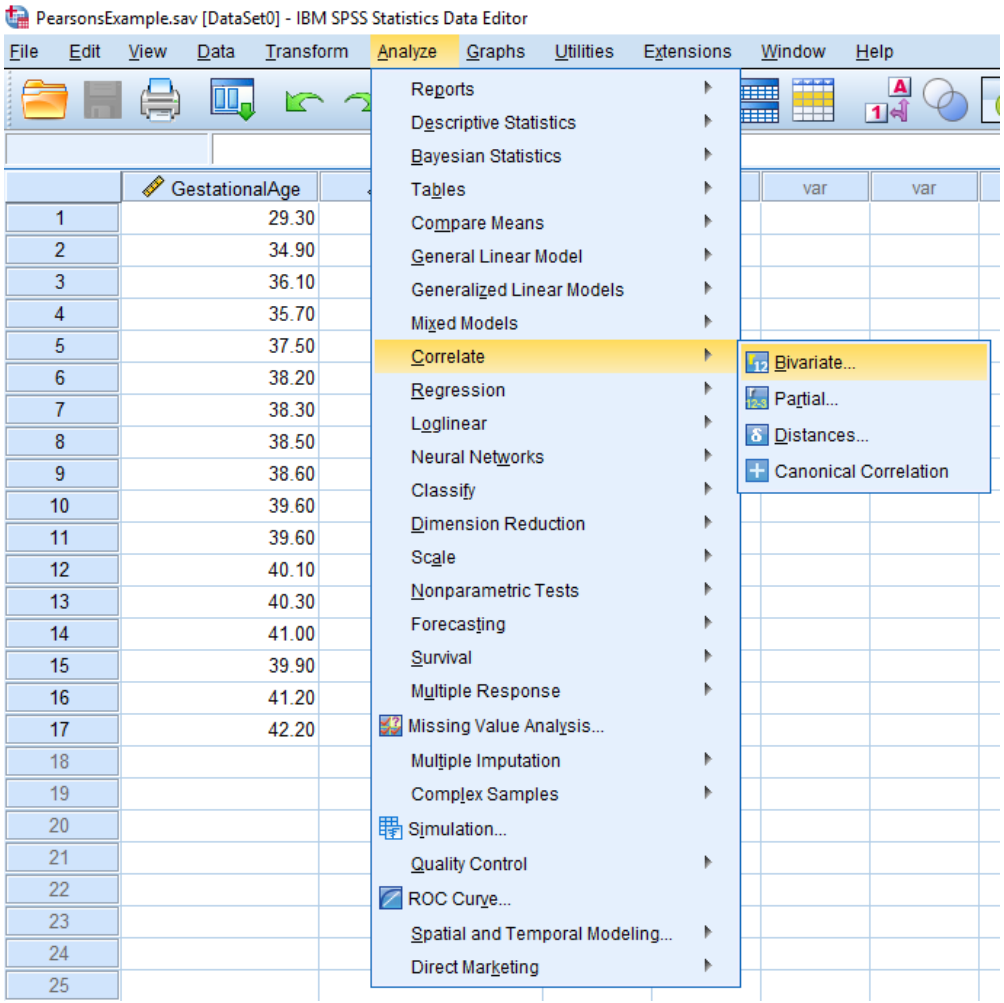
$$r = 0.793$$

5.1.4 Pearson's correlation in SPSS

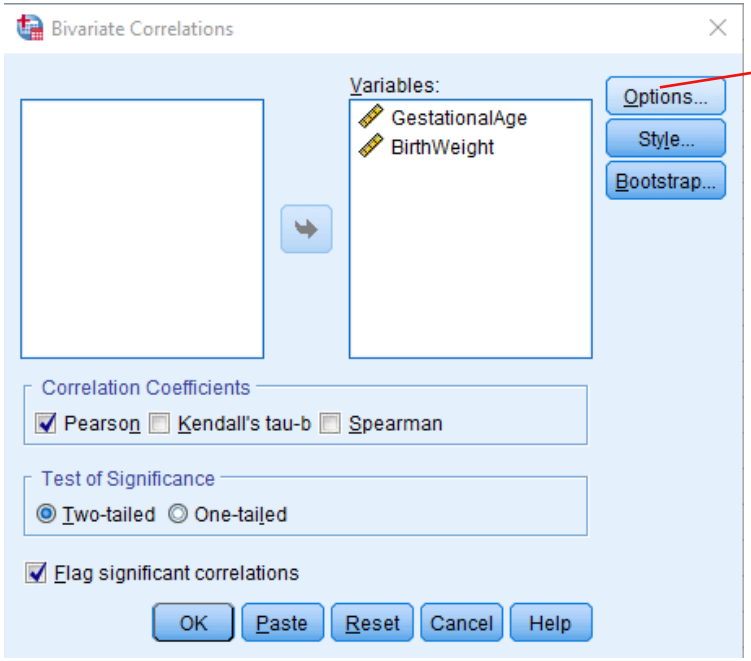
<Analyze>

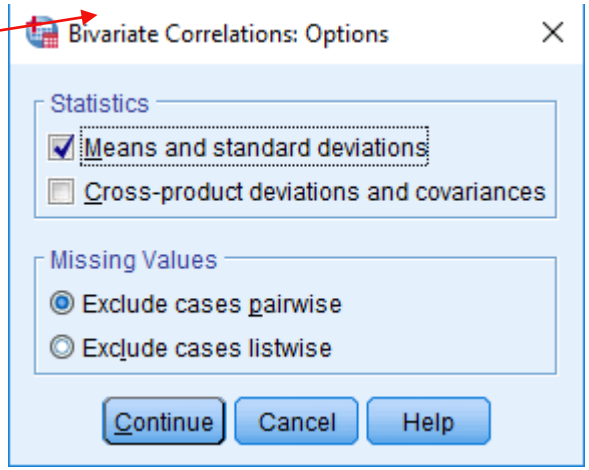
<Correlate>

<Bivariate>



5.1.4 Pearson's correlation in SPSS





Ensure that "Pearson" is checked.

You may also wish for SPSS to include the means and standard deviations.

5.1.4 Pearson's correlation in SPSS

The r test statistic (or correlation coefficient) is 0.793. This represents a strong positive correlation, which is what we expect for these two variables.

Note when classifying correlations, we only use "weak", "moderate" and "strong". A correlation coefficient of 0.01 may be considered "very weak" and a correlation of 1 is considered "perfect" however, in practice…it is rare to get a r of 1 and more common to get 0.9x, which would be considered "very strong". Here $p < 0.001$, hence it is statistically significant and we would reject H_0.

Correlations

		GestationalAge	BirthWeight
GestationalAge	Pearson Correlation	1	.793**
	Sig. (2-tailed)		.000
	N	17	17
BirthWeight	Pearson Correlation	.793**	1
	Sig. (2-tailed)	.000	
	N	17	17

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

5.1.4 Pearson's correlation in SPSS

Just like previous tests we have looked at, we can also determine statistical significance by using a critical r table.

If calculated r > critical r then reject H_0.

df = n – 2 for a Pearson's correlation test.

We see that 0.793 > 0.482 therefore reject H_0. $p < 0.05$ (or 0.001 as we saw from SPSS)

Degrees of Freedom	Alpha level 0.05	0.01
1	0.997	0.999
2	0.950	0.990
3	0.878	0.959
4	0.811	0.917
5	0.754	0.874
6	0.707	0.834
7	0.666	0.798
8	0.632	0.765
9	0.602	0.735
10	0.576	0.708
11	0.553	0.684
12	0.532	0.661
13	0.514	0.641
14	0.497	0.623
15	0.482	0.606
16	0.468	0.590
17	0.456	0.575
18	0.444	0.561
19	0.433	0.549
20	0.423	0.537
21	0.413	0.526
22	0.404	0.515
23	0.396	0.505
24	0.388	0.496
25	0.381	0.487
26	0.374	0.479
27	0.367	0.471
28	0.361	0.463
29	0.355	0.456
30	0.349	0.449

5.1.5 Interpreting the results of Pearson's correlation test

A Pearson's correlation test was used to assess the strength and direction of the linear relationship between gestational age (weeks) and birth weight (grams).

The assumptions of normality, linearity and homoscedasticity were assessed. A Kolmogorov-Smirnov test and a visual inspection of the normal Q-Q plots were used to confirm that both variables were normally distributed. Similarly, visually inspecting a scatterplot of gestational age against birth weight confirmed that the relationship between these variables was linear and homoscedastic.

There was a strong positive correlation between these two variables, $r(15) = .79$, $p < .001$.

Adapted from (Allan et al., 2019)

5.2 Simple Linear Regression

A simple linear regression assesses the mathematical relationship between an outcome (dependent, response) variable and one other predictor variable (independent, explanatory).

R-squared value: square of the correlation coefficient – measures how well the equation (below) “fits” the data (0 = not at all; 1 = perfect).

General equation for a simple linear regression:

$$y = b_0 + b_1 x$$

Where b_0 is a constant and b_1 is the coefficient of x.

5.2.1 Introduction to Simple Linear Regression

Let's say we want to look at associations between BMI and cholesterol level. We could use a correlation test for this however, if we wanted to see if BMI predicts cholesterol level then we would need to use a regression analysis.

Eg. BMI as predictor (x) and cholesterol level as dependent (y).

Do a simple linear regression between the two variables (BMI and cholesterol level).

In practice, multiple linear regression is more commonly used as it allows us to use additional predictors such as age, sex, smoking status, behavioural characteristics and other dietary markers.

The null hypothesis for a simple linear regression is that the slope of the regression line is zero ie. There is no relationship between the two variables. This gives us the indication that if there was a linear relationship between the two variables then H_0 is rejected.

5.2.1 Introduction to Simple Linear Regression

Patient ID	BMI	Total cholesterol (mg/dL)	Patient ID	BMI	Total cholesterol (mg/dL)
1	20.5	161	11	28.0	217
2	21.6	180	12	28.5	190
3	22.4	190	13	29.2	228
4	22.8	162	14	31.0	240
5	24.5	199	15	30.4	208
6	25.8	155	16	32.0	207
7	25.7	201	17	31.9	220
8	26.5	175	18	32.0	250
9	27.5	190	19	31.5	260
10	27.6	220	20	31.4	262

5.2.2 Assumptions of simple linear regression

There are four assumptions for simple linear regression:

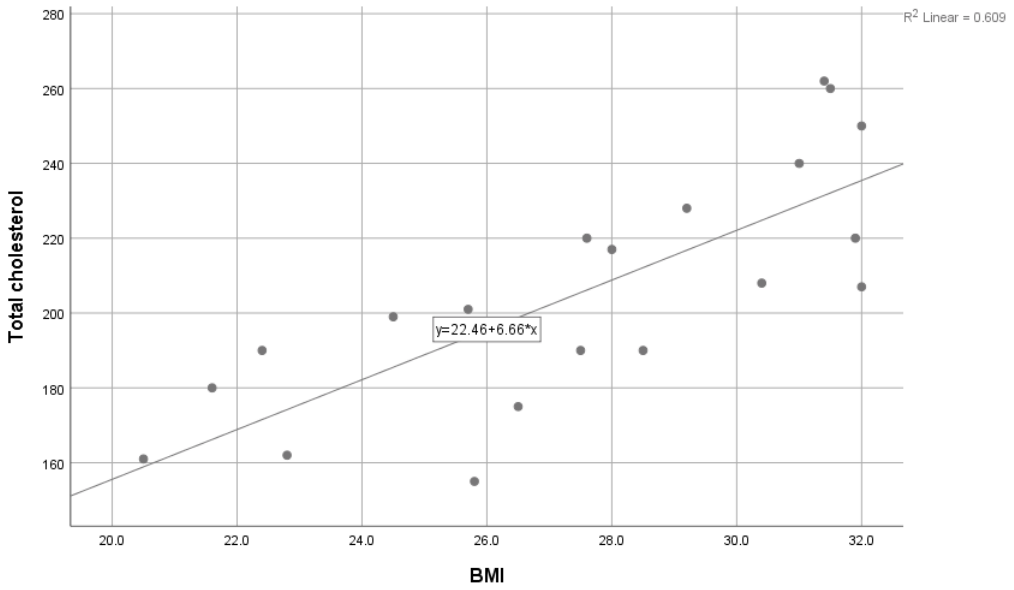
- Linearity – the outcome variable (y) has a linear relationship with the explanatory variable (x) – this can be a rough linear relationship.
- Homoscedasticity – For each value of x, the distribution (variance) of residuals is roughly the same.
- Normality – the outcome variable is normally distributed.
- There are no significant outliers.

5.2.2 Assumptions of simple linear regression

Linearity assumption

To check the assumption of linearity, it is always good to check this using a scatterplot. Remember BMI was the independent variable (explanatory) and hence goes on the x-axis. Total cholesterol is therefore the dependent variable and goes on the y-axis. SPSS or Excel can be used to plot a scatterplot with equation shown.

It is clear to see that the $R^2 = 0.609$, if we take the square root of this value, we get 0.780 and therefore there is a strong positive linear relationship between the two variables. Linearity assumption satisfied.

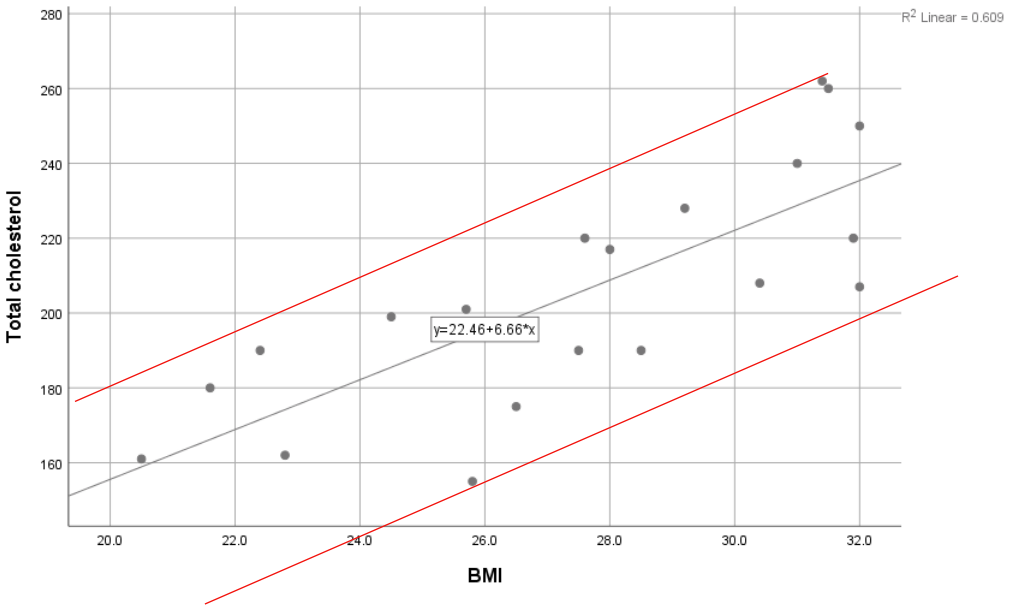


5.2.2 Assumptions of simple linear regression

Homoscedasticity assumption

We can use our existing scatterplot to check the homoscedasticity assumption by drawing in lines parallel lines. If the lines can be made to be parallel while encompassing the data points then the assumption has been met.

This gives us an indication that we have equal variances.



5.2.2 Assumptions of simple linear regression

Normality assumption

We can check the normality assumption using SPSS.

Using KS and SW tests, we see a non-significant result. Hence, our outcome variable is normally distributed.

Can also check the Normal Q-Q and Detrended Normal Q-Q plots.

If most of the points hover mostly around the line of best fit in a Normal Q-Q plot then this is an indication of normality.

If there are roughly and even amount of points above and below the horizontal line then this is an indication of normality.

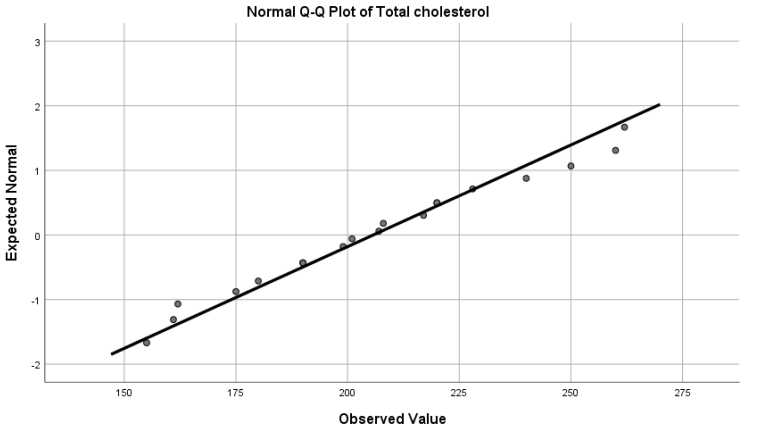
Sometimes the satisfying of an assumption can be subjective (according to the person's visual interpretation). This is why it's important to look at multiple tests and plots for normality.

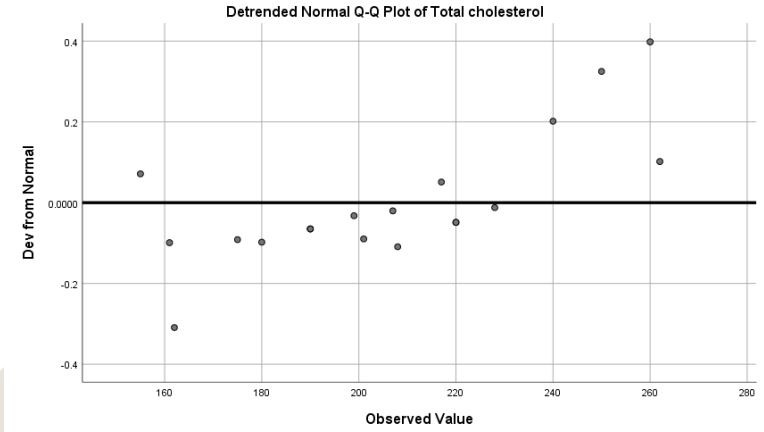
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Total cholesterol	.090	20	.200*	.965	20	.651

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

a. Lilliefors Significance Correction



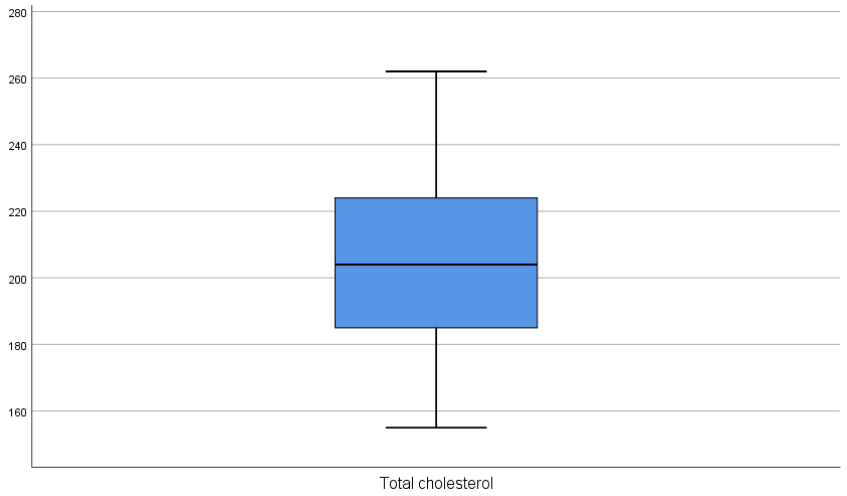


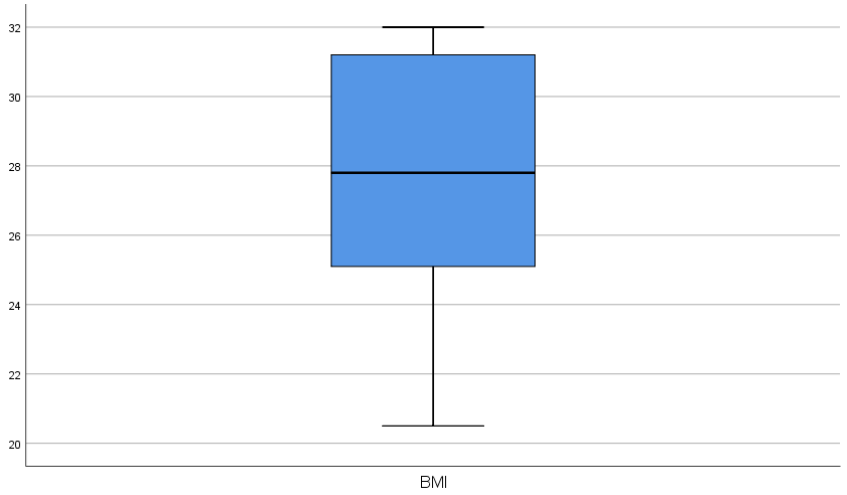
5.2.2 Assumptions of simple linear regression

Outlier assumption

By producing a boxplot of your data, you can see that there are no outliers present. If there were outliers, this would come up as small circles and data file row number (i.e. o[13]). If an asterisk, (i.e. *[27]) shows up instead of a circle, this indicates the outlier is an extreme score.

Here, we can see there are no “circles” or “asterisks” hence there are no (significant) outliers.





5.2.3 Developing the simple linear regression model

Our goal in Simple Linear Regression is to come up with an equation that describes our data.

As we expect a linear model, we should have a linear equation in the form of:

$$y = b_0 + b_1 x$$

We realise that b_0 is intercept and b_1 is the slope. We can calculate both b_0 and b_1 individually using:

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \; and \; b_0 = \bar{y} - b_1 \bar{x}$$

Where:

$$\bar{y} = mean\ value\ of\ the\ y\ variable$$
$$\bar{x} = mean\ value\ of\ the\ x\ variable$$
$$n = number\ of\ pairs\ of\ data$$

After performing a regression in SPSS, we note the effect size R^2.

5.2.3 Developing the simple linear regression model

Work out all parameters in the equation (use Excel):

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

n = 20
$\bar{y} = 205.75$
$\bar{x} = 27.54$

$\sum x = 550.8$

$\sum y = 4115$

$\sum x^2 = 15432.92$

$\sum xy = 115083.4$

Patient ID	BMI	Total cholesterol (mg/dL)	Patient ID	BMI	Total cholesterol (mg/dL)
1	20.5	161	11	28.0	217
2	21.6	180	12	28.5	190
3	22.4	190	13	29.2	228
4	22.8	162	14	31.0	240
5	24.5	199	15	30.4	208
6	25.8	155	16	32.0	207
7	25.7	201	17	31.9	220
8	26.5	175	18	32.0	250
9	27.5	190	19	31.5	260
10	27.6	220	20	31.4	262

5.2.3 Developing the simple linear regression model

Substitute into equations:

$$b_1 = \frac{20(115083.4) - (550.8)(4115)}{20(15432.92) - (550.8)^2}$$

$$b_1 = 6.655$$

$$b_0 = 205.75 - 6.655 \times 27.54$$
$$b_0 = 22.458$$

(When doing calculations make sure that you use Excel referencing to avoid rounding errors!)

Final equation for our model is:

$$y = b_0 + b_1 x$$
$$y = 22.458 + 6.655x$$

More rightly written as:

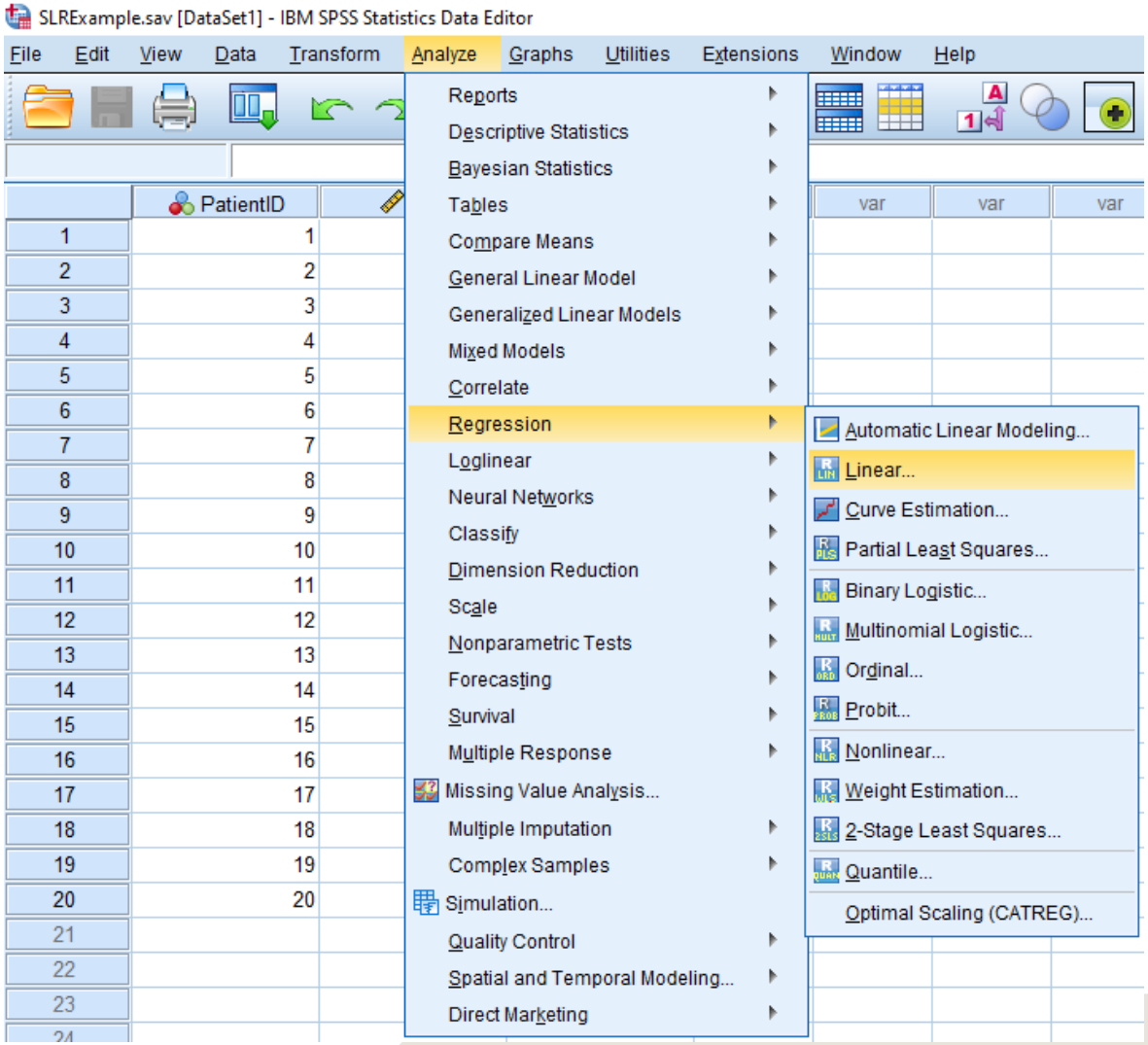
Total cholesterol = 22.458 + 6.655BMI

5.2.4 Simple Linear Regression in SPSS

<Analyze>

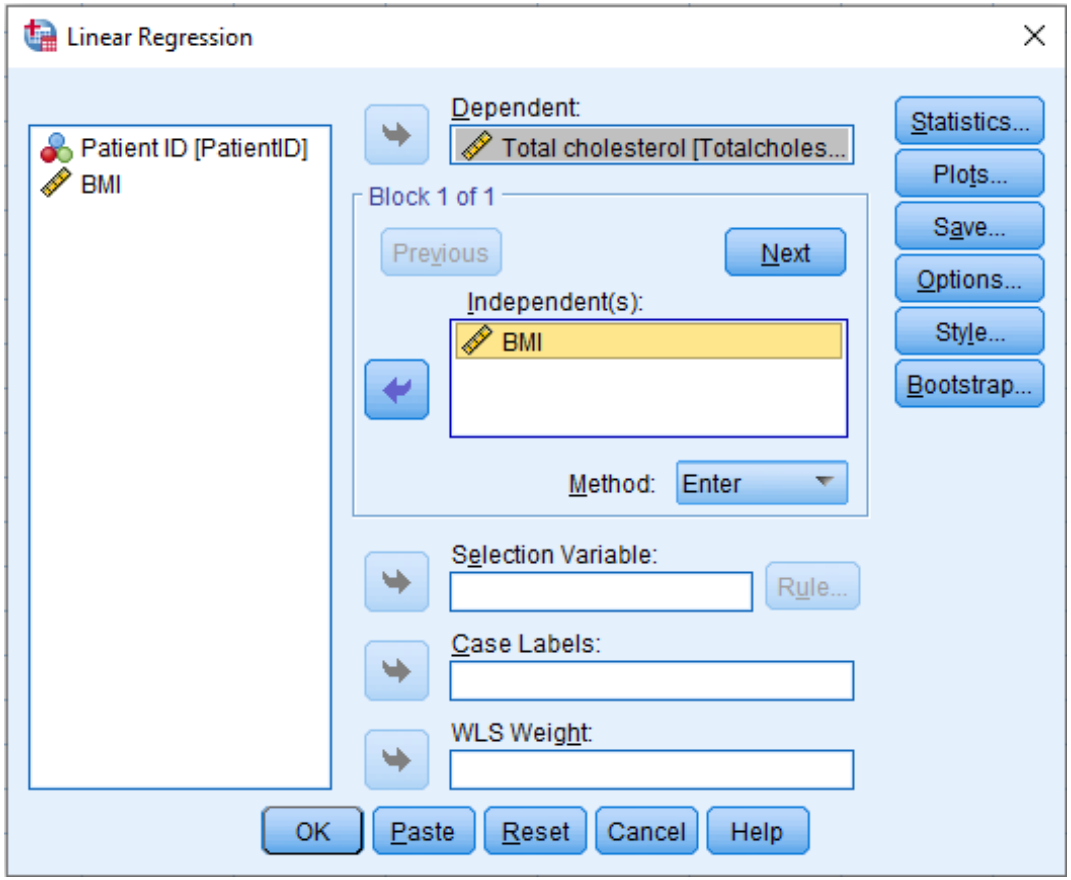
<Regression>

<Linear>



5.2.4 Simple Linear Regression in SPSS

Place in the corresponding dependent and independent variables and click OK.



5.2.4 Simple Linear Regression in SPSS

SPSS gives us four tables after performing a linear regression:

SPSS will also give us a lot of information that we do not necessarily need to report.

Variables Entered/Removed table:

In a simple linear regression, we only have one independent variable and in this case it is BMI. This table also tells us our dependent variable, which is Total cholesterol. The Method is the enter method where all variables are entered into the program at once. This becomes more relevant to us in multiple linear regression.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	BMI[b]	.	Enter

a. Dependent Variable: Total cholesterol

b. All requested variables entered.

5.2.4 Simple Linear Regression in SPSS

Model Summary Table

This table gives us our R-squared value. Remembering that R is our Pearson's correlation coefficient (strength and direction of the linear relationship). Squaring this value gives us the effect size known as R^2. This value indicates how well our data fits a linear model and is best explained by how much a change in dependent variable (Total cholesterol) is related to a change in the independent variable (BMI). In this case we see that 60.9% of the variability in Total cholesterol is due to BMI, which means that 39.1% of the variation is due to other factors (this error term is caused by the deviations between the data points and the line of best fit).

The adjusted R squared value is a value that SPSS uses mainly in multiple linear regression. When there are two or more independent variables, SPSS will compensate for the additional predictors.

In simple linear regression, it is appropriate to quote the R squared value however, in multiple regression, the adjusted R squared is usually more accurate.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.781[a]	.609	.588	20.408

a. Predictors: (Constant), BMI

5.2.4 Simple Linear Regression in SPSS

ANOVA table

The ANOVA table presents results for testing the null hypothesis that the slope of the line is not significantly different to zero (Note: the ANOVA test is also a linear model).

The mean squares result for the model (in the regression line) show variance in the data caused by changes in BMI while the mean squares result for the residual shows variance by the deviations between the line and actual data points (residuals). You can think of these as variance between groups and variance within groups in an ANOVA.

These two mean square values are used to calculate the F test statistic for the null hypothesis. As we can see, the null hypothesis is rejected as p (Sig.) < 0.001 hence we conclude that the slope of the model is significantly different to zero and so our model is actually showing a significant relationship between the two variables.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11689.011	1	11689.011	28.066	.000[b]
	Residual	7496.739	18	416.486		
	Total	19185.750	19			

a. Dependent Variable: Total cholesterol

b. Predictors: (Constant), BMI

5.2.4 Simple Linear Regression in SPSS

Coefficients Table:

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	22.458	34.898		.644	.528
	BMI	6.655	1.256	.781	5.298	.000

a. Dependent Variable: Total cholesterol

The final table is perhaps the most important table for developing our model. This table gives us the coefficients for the constant and independent variable in our equation.

Total cholesterol = 22.458 + 6.655BMI

Just like we had calculated by hand!

We see a lot of information here such as unstandardized and standardized coefficients, test statistics and Sig.

We will cover this in more detail in next week's topic on Multiple Linear Regression.

5.2.5 Interpreting the results of a simple linear regression model

A simple linear regression was used to predict the total cholesterol of each patient from their BMI. The regression equation showed a significant result, $F(1, 18) = 28.07$, $p < .001$, with an effect size of $R^2 = .61$. The regression equation was Total cholesterol = 22.46 + 6.66BMI mg/dL, which indicates that a unit increase in BMI resulted in a 6.66 mg/dL increase in total cholesterol.

5.2.5 Interpreting the results of a simple linear regression model

A simple linear regression was used to predict the total cholesterol of each patient from their BMI. The regression equation showed a significant result, $F(1, 18) = 28.07$, $p < .001$, with an effect size of $R^2 = .61$. The regression equation was Total cholesterol = 22.458 + 6.66BMI mg/dL, which indicates that a unit increase in BMI resulted in a 6.66 mg/dL increase in total cholesterol.