Multiple Regression

Learning Centre

TABLE OF CONTENTS

01 What is Multiple Regression?

02 Standard Multiple Regression (SMR) & A Worked Example on SPSS

03 Hierarchical Multiple Regression (HMR) & A Worked Example on SPSS

What is Multiple Regression?

JAMES COOK UNIVERSITY SINGAPORE

- Multiple Regression (MR) is a statistical analysis used to examine the relationship between multiple independent variables (IVs), and a dependent variable (DV)
- The IVs are also known as predictor variables, while the DV is also called the criterion variable
- In other words, a multiple regression answers the question: which IVs predict the DV?
- However, MR cannot always imply causation

Standard Multiple Regression (SMR)

Example

*In SMR, all IVs are placed into the model at the same time!

**The sample size of 30 was used only for illustration purposes; an actual study would require a larger sample size!

A researcher is interested in finding out if scores from 3 different assignments can predict final exam scores

The researcher then invited 30 participants who had enrolled into a module last semester to complete a survey asking for:

1) Scores from each assignment
2) Score from the final exam

Data credit
https://college.cengage.com/mathematics/brase/understandable_statistics/7e/students/datasets/mlr/frames/frame.html

Location of SPSS Data Files

Example SPSS data for practice are available on LearnJCU:

Log in to LearnJCU -> Organisations -> Learning Centre JCU Singapore -> Statistics Support -> Statistics Resources -> SPSS Data for Practice

Assumptions Testing

JAMES COOK UNIVERSITY SINGAPORE

01 Univariate Outliers

Cases with extreme values on single variables

02 Multivariate Outliers

Cases with extreme values on multiple variables

03 Normality

Ensuring that the data is normally distributed

04 Normality, Linearity, Homoscedasticity of Residuals

Ensuring that the differences between observed and predicted values of the DV are normally distributed

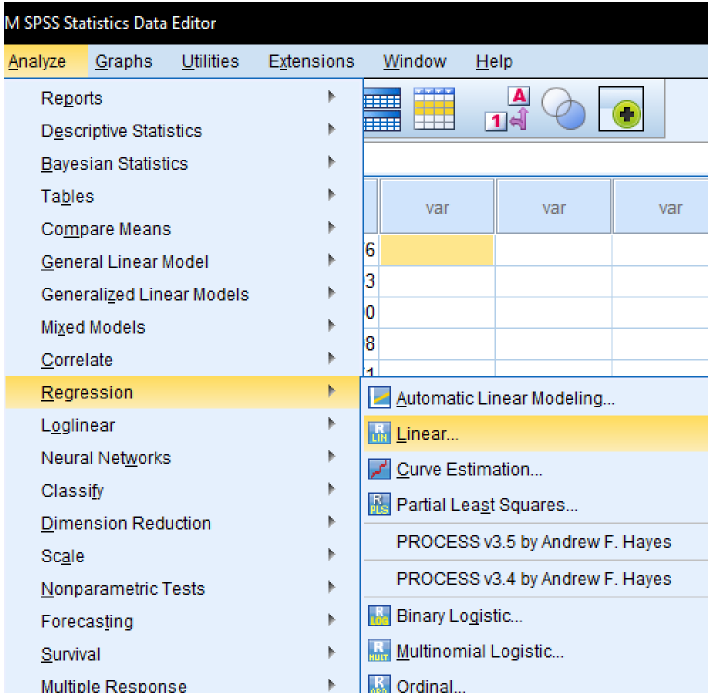
05 Multicollinearity

Ensuring that none of the predictor variables are too correlated

1. Univariate Outliers

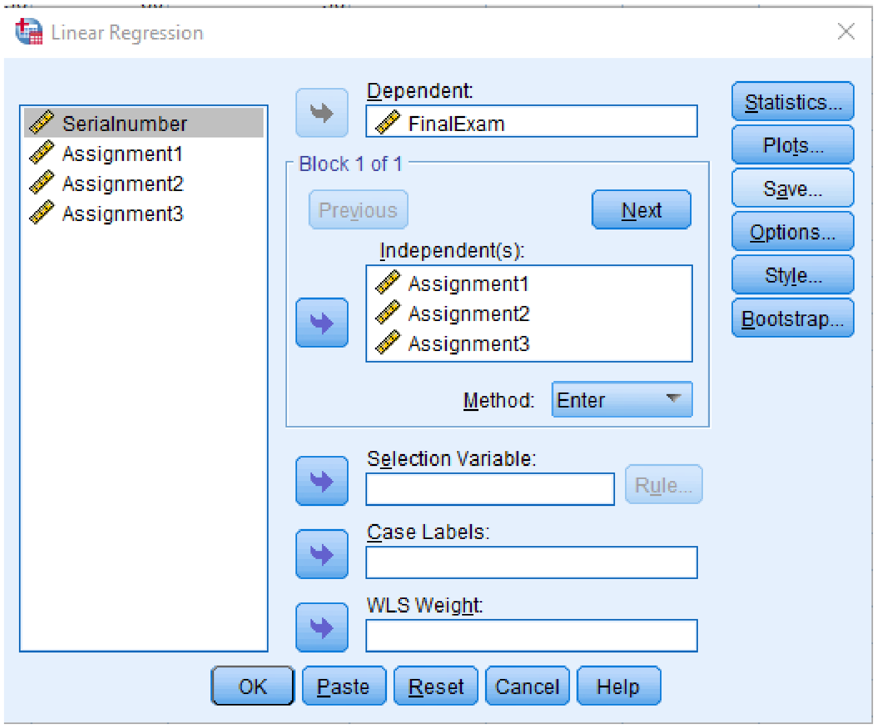
One way to test this assumption is to use <u>Cook's distances</u>

- Go to Analyze -> Regression -> Linear



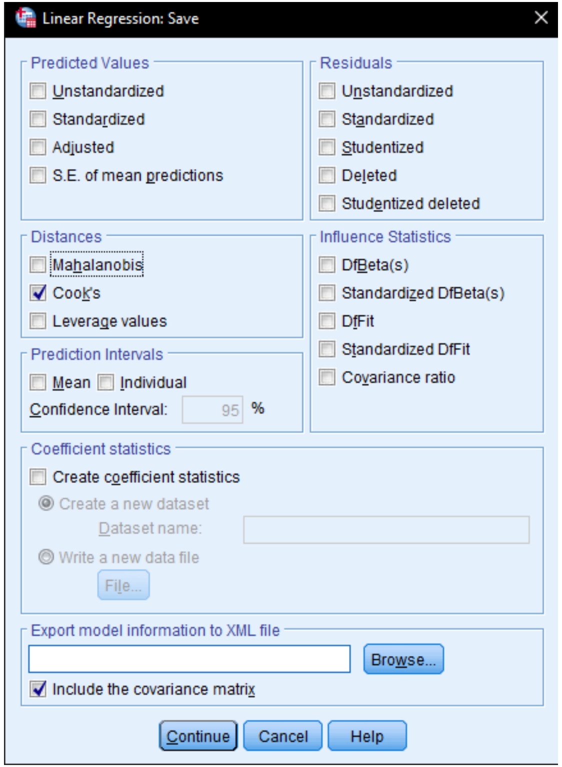
1. Univariate Outliers

- Move ‘FinalExam’ into Dependent, and the 3 assignments into Independent(s)
- Click on ‘Save’



1. Univariate Outliers

- Select *Cook's*
- Click continue



1. Univariate Outliers

JAMES COOK UNIVERSITY SINGAPORE

Note that by selecting Cook's Distance, SPSS will create a new variable for it in your *dataset*

Serialnumber	Assignment1	Assignment2	Assignment3	FinalExam	COO_1
1	73	80	75	76	.01243
2	89	88	93	93	.00595
3	89	91	90	90	.01000
4	94	98	100	98	.12056
5	77	70	75	73	.00120
6	65	61	70	66	.00396
7	69	74	77	75	.02715
8	55	56	60	58	.00001
9	81	79	90	88	.01488
10	75	70	88	82	.00426
11	69	70	73	71	.01178
12	70	65	74	71	.00016
13	93	95	91	92	.02682
14	79	80	73	76	.00215
15	70	73	78	74	.04349
16	90	89	96	96	.05356
17	73	75	68	70	.00489
18	80	80	80	79	.00966
19	86	92	86	89	.00080
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

1. Univariate Outliers

Look at the ***maximum*** Cook's Distance

- If it is less than 1, there is no univariate outlier

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	57.99	100.18	80.28	10.076	30
Std. Predicted Value	-2.213	1.974	.000	1.000	30
Standard Error of Predicted Value	.373	.755	.567	.114	30
Adjusted Predicted Value	57.98	100.63	80.28	10.100	30
Residual	-2.238	5.354	.000	1.498	30
Std. Residual	-1.414	3.383	.000	.947	30
Stud. Residual	-1.512	3.654	.002	1.013	30
Deleted Residual	-2.634	6.247	.008	1.716	30
Stud. Deleted Residual	-1.553	5.139	.054	1.221	30
Mahal. Distance	.648	5.640	2.900	1.526	30
Cook's Distance	.000	.557	.036	.102	30
Centered Leverage Value	.022	.194	.100	.053	30

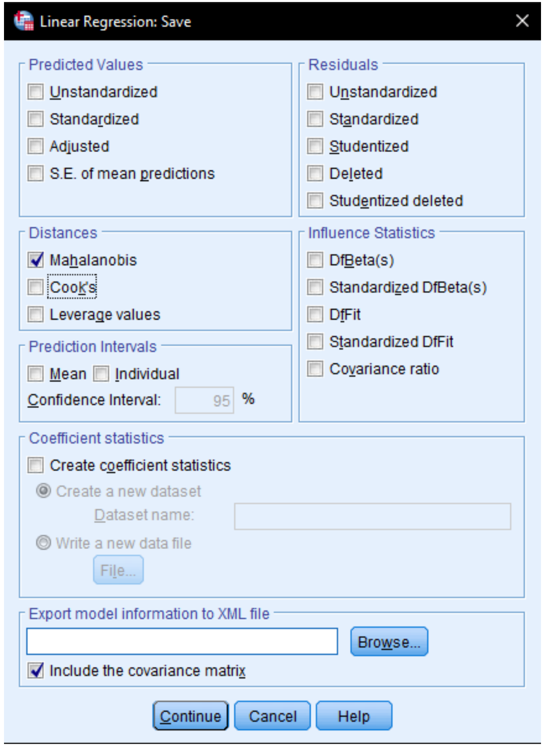
a. Dependent Variable: FinalExam

2. Multivariate Outliers

Multivariate outliers are identified using <u>Mahalonobis Distances</u>

Follow the same steps as univariate outliers… except this time, select the *Mahalanobis*

As mentioned before, selecting this option creates a new variable in the *dataset*



2. Multivariate Outliers

Look at the ***maximum*** Mahalanobis Distance

The ***maximum*** value should be lesser than the critical Chi-square value (from the Chi-square table)

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	57.99	100.18	80.28	10.076	30
Std. Predicted Value	-2.213	1.974	.000	1.000	30
Standard Error of Predicted Value	.373	.755	.567	.114	30
Adjusted Predicted Value	57.98	100.63	80.28	10.100	30
Residual	-2.238	5.354	.000	1.498	30
Std. Residual	-1.414	3.383	.000	.947	30
Stud. Residual	-1.512	3.654	.002	1.013	30
Deleted Residual	-2.634	6.247	.008	1.716	30
Stud. Deleted Residual	-1.553	5.139	.054	1.221	30
Mahal. Distance	.648	5.640	2.900	1.526	30
Cook's Distance	.000	.557	.036	.102	30
Centered Leverage Value	.022	.194	.100	.053	30

a. Dependent Variable: FinalExam

2. Multivariate Outliers

JAMES COOK UNIVERSITY SINGAPORE

- Our degrees of freedom (*df*) is 3 (*df* is a number of IVs), and the alpha is set at .001, giving us a critical value of 16.266
- Since the observed maximum mahalanobis distance is 5.64, which is smaller than 16.266, there is no multivariate outlier

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458

You can easily find this table on the Internet!

Hmmm…but how to deal with outliers or extreme values if any?

JAMES COOK UNIVERSITY SINGAPORE

1. Re-check your data entry. Check if they are measurement errors (e.g., out-of-range values). Before re-running all tests of assumptions:
    - Correct the errors
    - Leave the errors as missing
    - Remove the observation with the errors
    - Replace the errors/wrong values with e.g., mean, the largest valid value, or multiple implication
2. For genuine outliers, consider keeping or removing

Dealing with outliers or extreme values

JAMES COOK UNIVERSITY SINGAPORE

If you want to *keep* outliers (okay for simple regression):

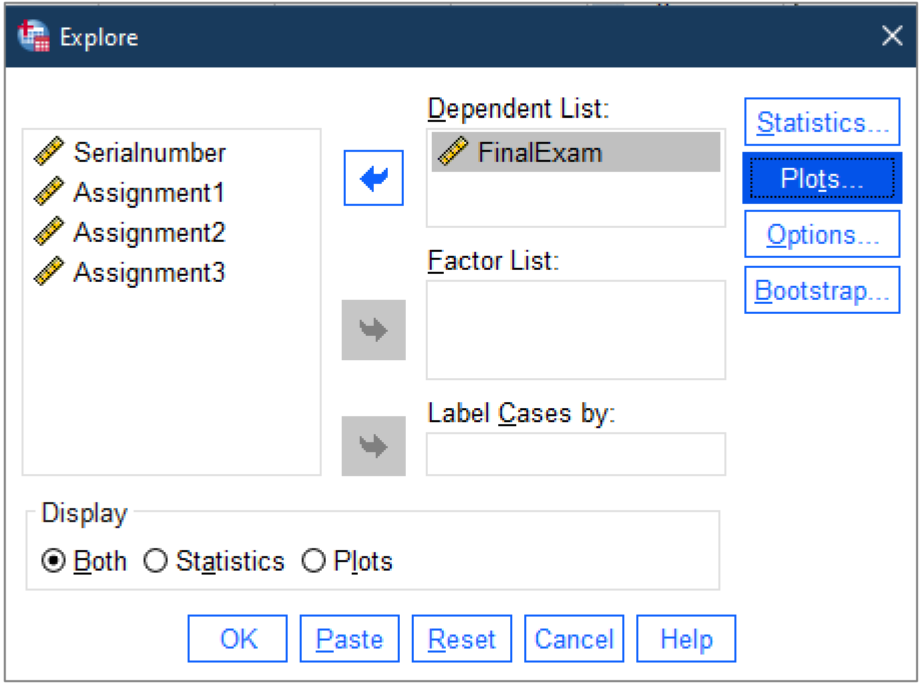
- Transform the DV, or
- Run the linear regression with and without the outlier. If there are no appreciable differences in the results, then keep the outlier and report

Consider removing genuine extreme values.

3. Normality

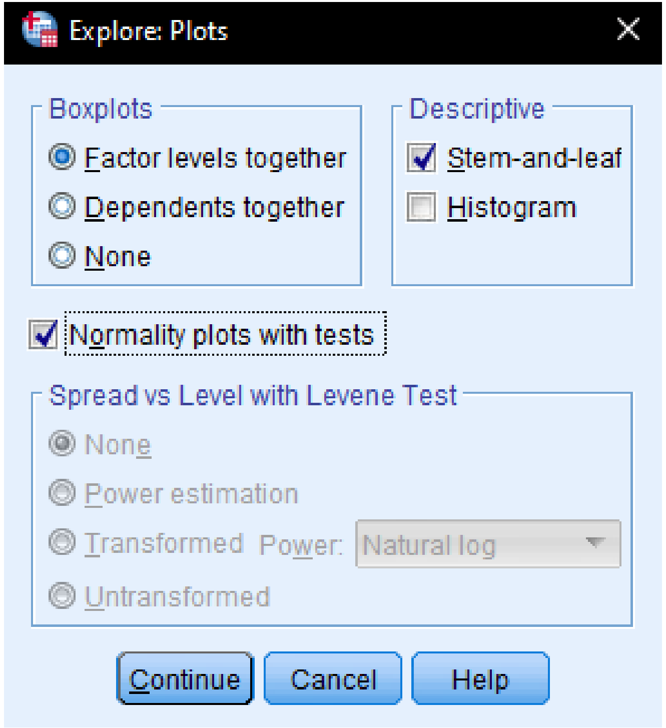
To test the assumption of normality, we can use the Shapiro-Wilk test

- Go to Analyze -> Descriptive Statistics -> Explore



3. Normality

- Click on Plots
- Select *Normality plots with tests*
- Continue and OK!



3. Normality

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Assignment1	.087	30	.200*	.975	30	.679
Assignment2	.088	30	.200*	.981	30	.840
Assignment3	.124	30	.200*	.959	30	.283
FinalExam	.127	30	.200*	.974	30	.648

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

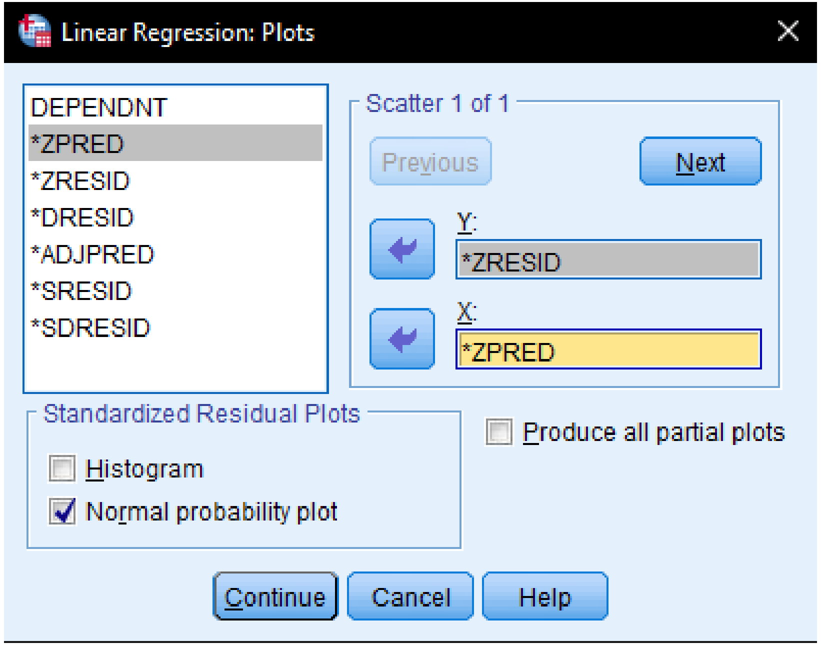
a. Lilliefors Significance Correction

- We focus on the *Sig.* value of the Shapiro-Wilk test of the DV. To assume the normality, we are looking for a non-significant Shapiro-Wilk statistic ($p > .05$)
- Hence, in this example, we conclude that the assumption of normality was met

4. Normality, Homoscedasticity of Residuals, and Linearity

Go to Analyze -> Regression -> Linear -> Plots

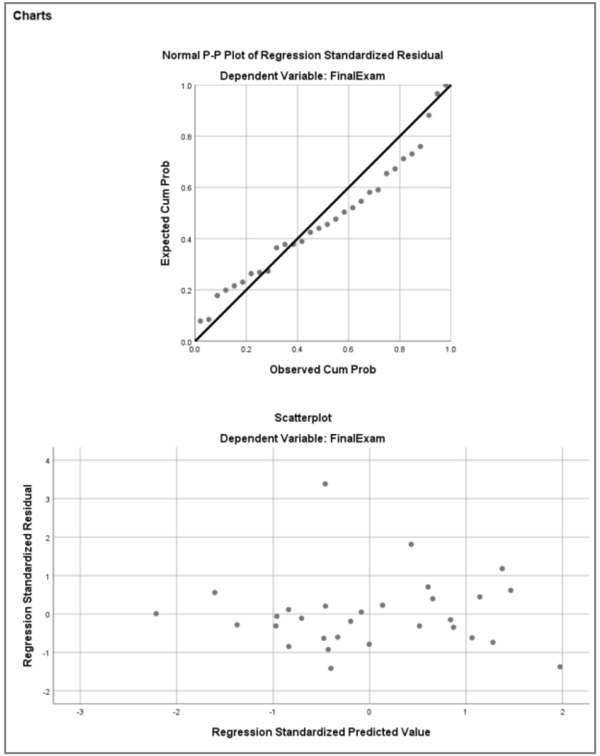
- Move ‘ZRESID’ into Y
- Move ‘ZPRED’ into X
- Select ‘Normal probability plot’
- Continue, and OK!



4. Normality, Homoscedasticity of Residuals, and Linearity

JAMES COOK UNIVERSITY SINGAPORE

- For the upper chart, if the data points are aligned with the diagonal straight line, the residuals are normally distributed.
- For the bottom chart, we are looking for equal spreading of data points across the X axis
- Taken together, if both charts look like the ones we have on the right, we conclude that the assumptions for normality and homoscedasticity of residuals are not violated.



4. Normality, Homoscedasticity of Residuals, and Linearity

The assumption of linearity can be checked by conducting a Pearson's correlation analysis or graph a scatterplot.

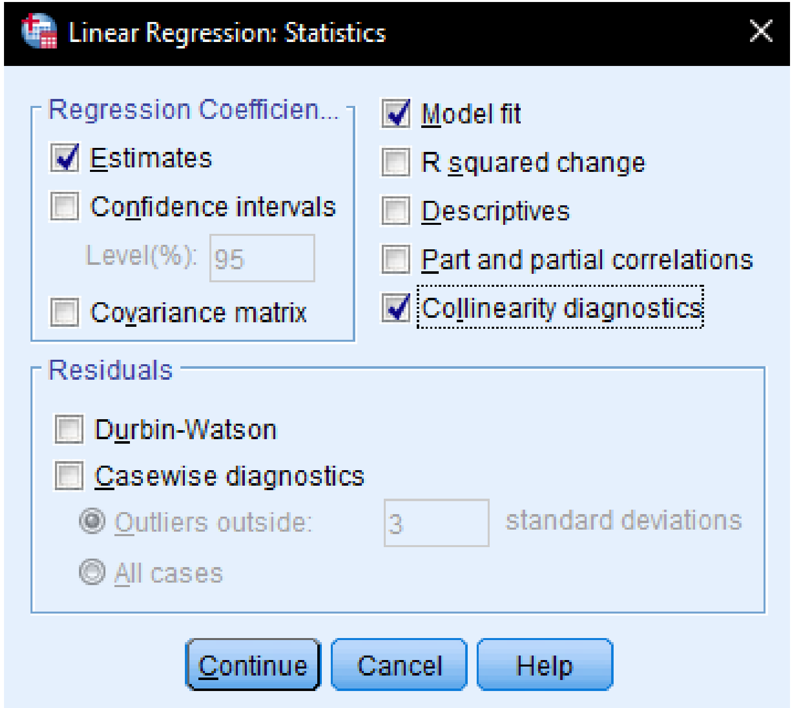
*Check out how to run correlation analysis in the **Correlation** slides (JCUS Learning Centre website -> Statistics and Mathematics Support)

5. Multicollinearity

Analyze -> Regression -> Linear -> Statistics

- Select *Estimates* and *Model fit*
- Select *Collinearity diagnostics*
- Continue, and OK!

*SMR is also conducted using these steps



5. Multicollinearity

JAMES COOK UNIVERSITY SINGAPORE

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	Collinearity Statistics Tolerance	Collinearity Statistics VIF
1	(Constant)	-2.624	2.451		-1.071	.294		
	Assignment1	.005	.095	.004	.049	.961	.101	9.938
	Assignment2	.380	.071	.395	5.310	.000	.151	6.638
	Assignment3	.652	.056	.647	11.590	.000	.267	3.749

a. Dependent Variable: FinalExam

To determine if there is multicollinearity among IVs, look at the *Tolerance* and *VIF.*

Tolerance should be > .1, and VIF should be below 10.

In this example, the assumption for multicollinearity has not been violated.

Standard Multiple Regression (SMR)

*Look at how to conduct SMR in Slide 22

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-2.624	2.451		-1.071	.294		
	Assignment1	.005	.095	.004	.049	.961	.101	9.938
	Assignment2	.380	.071	.395	5.310	.000	.151	6.638
	Assignment3	.652	.056	.647	11.590	.000	.267	3.749

a. Dependent Variable: FinalExam

Assignment 1 has a *p* value of .961, while Assignments 2 and 3 both have *p* values of < .001. We then conclude that only Assignments 2 and 3 are significant predictors of final exam scores

Coefficients tell us which is a 'better' predictor. Assignment 3 has the highest value, thus it can be taken as the 'best' predictor.

Results Write-up

An example write-up can be found on page 198 in

Allen, P., Bennett, K., & Heritage, B. (2019). ***SPSS Statistics: A Practical Guide*** **(4th ed.). Cengage Learning.**

Hierarchical Multiple Regression (HMR)

Example

In HMR, IVs are added into the model cumulatively! It is commonly used to account for control variables.

Building on example 1, the researcher thinks that other than the 3 assignments that could predict exam scores, sleep could also affect how well a student performs.

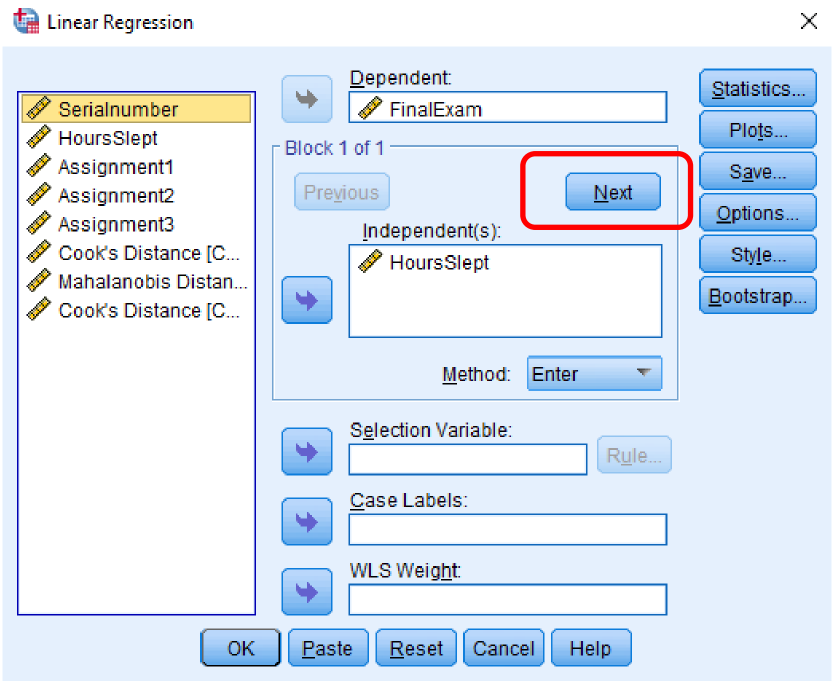
To find out the sole effect of assignments on exam scores, he controlled for this new variable ‘sleeping hours’.

The researcher asks the 30 participants from Example 1 to also provide an average of how many hours of sleep they get in a night.

Hierarchical Multiple Regression (HMR)

Before we begin, note that assumption testing has to be conducted! (look at Example 1)

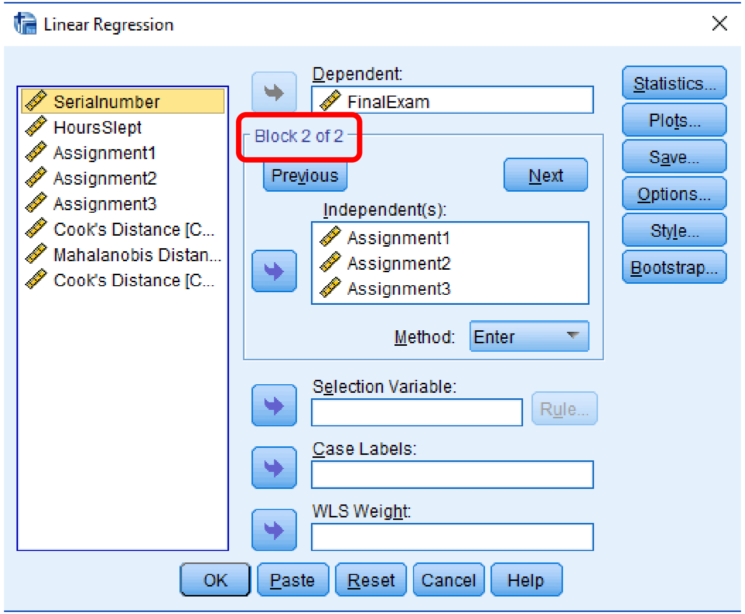
- To conduct a HMR: Go to Analyze -> Regression -> Linear
- Move 'FinalExam' into Dependent, and 'HoursSlept' into Independent(s) (*controlled variables are added in the first block!)
- Then click Next to create another block (see picture) to input our 3 assignments



Hierarchical Multiple Regression (HMR)

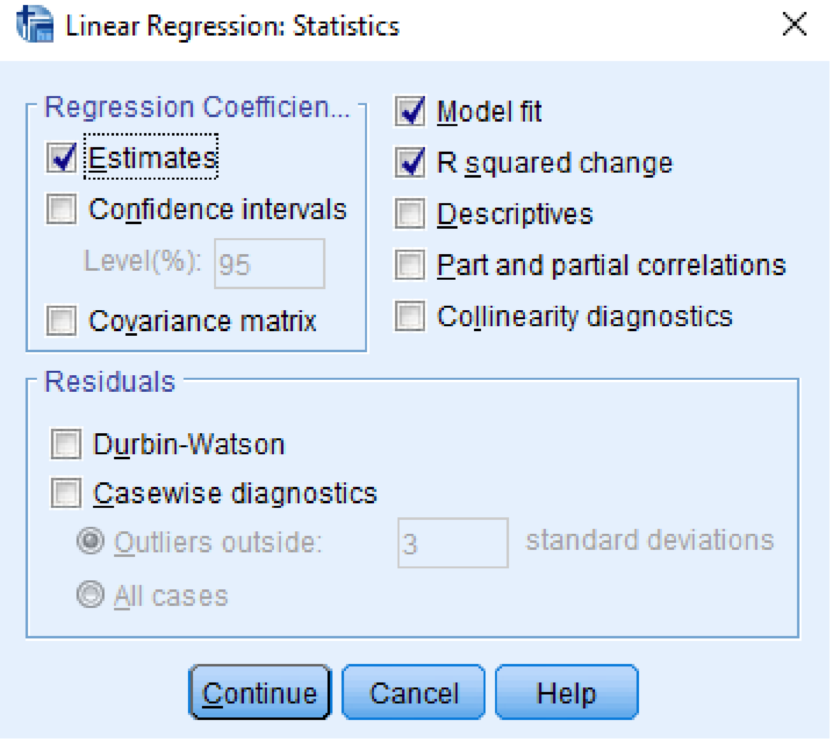
We should now see that it is at block 2 of 2

- Move the main predictors (Assignments 1 - 3) into Independent(s)



Hierarchical Multiple Regression (HMR)

- Click on Statistics
- Select *Estimates*, *Model fit*, and *R squared change*
- Continue, and OK!



Output

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	HoursSlept[b]	.	Enter
2	Assignment3, Assignment2, Assignment1[b]	.	Enter

a. Dependent Variable: FinalExam

b. All requested variables entered.

This table shows us the order in which we entered the variables.

In block 1 (Model 1), we input HoursSlept
In block 2 (Model 2), we entered Assignments 1 – 3.

Output

JAMES COOK UNIVERSITY SINGAPORE

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.406[a]	.165	.135	9.476	.165	5.516	1	28	.026
2	.989[b]	.978	.975	1.613	.814	313.880	3	25	.000

a. Predictors: (Constant), HoursSlept

b. Predictors: (Constant), HoursSlept, Assignment3, Assignment2, Assignment1

In model 1, a number of sleeping hours contributed to 17% of variability in exam scores, $F(1, 28) = 5.52$, $p = .026$

In model 2, the addition of our 3 predictors resulted in an R squared change of .81, $\Delta F(3, 25) = 313.88$, $p < .001$. Model 2 accounted for 98% of variability in exam scores

Output

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	62.704	7.682		8.162	.000
	HoursSlept	2.761	1.176	.406	2.349	.026
2	(Constant)	-2.556	2.524		-1.013	.321
	HoursSlept	-.044	.233	-.006	-.188	.853
	Assignment1	.001	.099	.001	.008	.994
	Assignment2	.385	.079	.400	4.895	.000
	Assignment3	.653	.057	.648	11.350	.000

a. Dependent Variable: FinalExam

Looking at the individual variables, Assignments 2 and 3 are significant predictors of exam scores

Also, notice the change from model 1 to 2. After the addition of the main predictors, the *p* value of sleeping hours had changed from .026 to .853

Results Write-up

An example write-up can be found on page 204 in

Allen, P., Bennett, K., & Heritage, B. (2019). *SPSS Statistics: A Practical Guide* (4th ed.). Cengage Learning.

Any Questions?

learningcentre-singapore@jcu.edu.au