OLS LINEAR REGRESSION CHECKLIST

By Daniel Whitten | Questions? d.j.whitten@fsw.leidenuniv.nl

Use this checklist to ensure you do not forget any steps or assumptions while performing your OLS Regression analysis. For each step, use the test in the "Check" column to check compliance. If the statement in that column is false, see the "Solutions" column for methods of correcting it.

Please note, this is **not a substitute for your own study** and reference to appropriate course material, but a reminder of what to check the course material for. If you find any discrepancy between this checklist and the course material, please **always follow the course material**.

Find your main variables

✅	Assumption	Check	Solution
	y is caused by x	Theory and logic suggest so	Find other variables
		Note: You can check Pearson CORRELATIONS coefficients, but remember correlation ≠ causation	
	y is continuous	Frequency table shows equally spaced numbers on a scale with > 2 possible values	Use different variable or different analysis
		Note: If y is binary (0 or 1), you might be able to use logistic regression	
	x is continuous, binary, or categorical	Frequency table shows equally spaced numbers on a scale (continuous/binary) or discrete unordered categories (categorical)	Use different variable
		Note: If x is ordinal, you might be able to treat as continuous, depending on the case	

Find other variables and prepare them for analysis

✅	Assumption	Check	Solution
	No Omitted Variable Bias	No other variable could account for the variation in y	Include relevant control variables
		Note: Assess effectiveness by running a multiple regression introducing control variables at a second step and check the model fit	
	Values on an appropriate scale	Frequency tables show values as you want them for analysis	Recode variables as needed
		Note: This will vary for every variable and model, do what is logical (eg. 1,2 => 0,1; 1-100 => 0-1; 12 categories => 5 categories)	
	Independent variables are continuous or binary	Values for all variables are equally spaced numbers on a scale	Create dummy variables for categorical variables
		Note: Create 1 separate binary (1/0 – true/false) variable for EACH category; 1 is **excluded** from the regression (baseline/reference category)	
	Effect of x is homogenous	The slope of x does not vary significantly at different levels of z	Run test regressions with interaction term and calculate slopes of x at z values
		Note: Interaction terms are X*Z, and are included alongside X & Z; if the moderation is not effective, they can be removed	
	Errors are independently distributed	The data is not clustered in any way (time, geography, etc.)	Run separate regressions
			Lag variables (time series data)
			Weight variables
			Introduce fixed effects
			Use multi-level model
		Note: Most of these solutions don't have to be demonstrated for the exam in this course, but you should have a sense of what they are and why	
		Note: For time series data, see the related "No Autocorrelation" check below	
	n > k	Number of valid cases > Number of independent variables	More data or fewer variables

		Note: Get the number of valid cases from `DESCRIPTIVES` using `/MISSING LISTWISE.`	
	Variation in x	No variable has the same value for every case	Different variables or more data
		Note: Get this information from `FREQUENCIES`; no value should have 100% of cases for any variable	

Run the regression and check remaining assumptions

✅	Assumption	Check	Solution
	No multicollinearity	VIF: ▪ most < 5 ▪ all < 10	Remove, replace, or combine variables Acknowledge and live with it
		Note: Get VIF by adding `TOL` to the `/STATISTICS` line	
	No autocorrelation	Not time series data OR Durbin-Watson between 1 & 3	See "Errors are independently distributed" solution above
		Note: Get Durbin-Watson scores with `/RESIDUALS` in your syntax	
		Note: Durbin-Watson tests are only valid if data is sorted chronologically	
	The relationship is linear	Scatterplot of standardized residuals & standardized predicted values has no non-linear patterns	Examine partial plots to find the x variable with a non-linear pattern Then Include polynomial terms (x^2 or x^3) alongside the problem x Replace the problem x with a ln(x) (natural logarithm of x) to see if pattern becomes linear
		Note: Significance of polynomial term indicates if there is significant non-linear trend	
		Note: Direction of x and the polynomial term together show the bend of line	
		Note: Replacing x with ln(x) will change the interpretation (see final section below)	
		Note: Get these scatterplots with these in your syntax: ▪ `/SCATTERPLOT (*ZRESID *ZPRED)` ▪ `/PARTIALPLOT`	
	No heteroskedasticity	Scatterplot of standardized residuals & standardized predicted values has no clear pattern & has even width of residuals	Identify omitted control variables Identify moderation and introduce interaction term Identify non-independent standard errors (clustering) Identify outliers/influential cases Replace the problem x with ln(x) (natural logarithm of x) and see if that fixes the pattern Bootstrap standard errors (not on exam) Standard errors for heteroskedastic models (not on exam)
		Note: Get this scatterplot with `/SCATTERPLOT (*ZRESID *ZPRED)`	
		Note: Replacing x with ln(x) will change the interpretation (see final section below)	
	Errors are normally distributed	Values on P-P plot of standardized residuals are close to the line	Acknowledge and live with it Bootstrapping (not on exam)
		Note: Note: Get the P-P plot by including `/RESIDUALS` in your syntax or checking "Normal probability plot" via menu	

	No outliers and influential cases	No flags from multiple measures of outliers & influential cases, eg.: ▪ No outliers visible on partial plots ▪ Normal distribution of standardized residuals: ○ 0 cases > |3.29| ○ < 1% of cases > |2.58| ○ < 5% of cases > |1.96| ▪ Cook's Distance < 1 for all cases ▪ All DFBetas < |1| for all cases ▪ No suspicious values in the casewise diagnostics table	Investigate cases with flagged values Exclude cases included by mistake Correct typos/mistaken values Run regression with & without questionable cases; no significant change to results suggests no influence Acknowledge and live with it Back to the drawing board
		Note: Get these scatterplots with these in your syntax: ▪ `/SCATTERPLOT (*ZRESID *ZPRED)` ▪ `/PARTIALPLOT`	
		Note: Get casewise diagnostics with: `/CASWISE PLOT OUTLIERS(2)`	
		Note: Get top 10 standardized residuals with: `/RESIDUALS`	
		Note: Save measurements as variables with: `/SAVE [MEASUREMENT NAME] (variable_name)` If you have to run the regression again, go in and delete these saved variables from your dataset, or you will get an error if you try to create them again	
		Note: Count the standardized residuals by creating a binary variable to check if it passes a given threshold or use `AGGREGATE` as per the SPSS Syntax Reference here	
		Note: Check the rules for values via `DESCRIPTIVES` (include all measures, including all DFBetas)	

Interpret your results

✅	Assumption	Check	Solution
	The results model a relationship in the population based off a given sample	All above checks are true, or solutions sufficiently applied	See the checks and solutions above
	b_i is the predicted change in y given 1 unit change in x_i, holding all other variables constant	x is not ln(x) (natural logarithm of x)	$b_i/100$ is the predicted change in y given a 1% change in x_i, holding all other variables constant
		There is no interaction term in the model between x_i and z (moderator)	b_i is the predicted change in y given a 1 unit change in x_i when z=0, holding all other variables constant
		Note: If you have dummy variables, the constant value assumes the (omitted) baseline category is true (1) and all other categories are false (0); each other coefficient for the dummy variables is the change in Y when that value is true (1) and all others are false (0)	

Final Notes

Use the sig. column for p-values, and include `CI` on the `/STATISTICS` line of your regression syntax to get confidence intervals

Take the unstandardized coefficients, standard errors, p-values, R^2 and N values and use them to make your formatted table to present results

To gauge the effects, use the regression equation to calculate the predicted value of y at different levels of x

Keep other values constant:

Continuous variable:	Mean
Binary/Categorical variable:	Mode
Note:	This means 1 for dummy representing the mode value, 0 for all other dummies
Note:	Remember to do the appropriate math on the constants if you have interaction terms, polynomial terms, or logarithms