

A Basic Approach to Linear Regression

R. Hayward

Like other regression techniques, linear regression can be used for 1) evaluating associations (nature and strength of relationships between a dependent variable [Y] and one or multiple independent variables [Xs]) or 2) for prediction (giving the best estimate of the expected value of the dependent variable based upon the observed values of the independent variables). Important things to know about linear regression include:

- 1) The dependent variable (Y) should be continuous (and it is best, but not required, if it has a symmetric distribution).
- 2) The relationships (association) between Y and the Xs (independent variables) are linear and consistent throughout its distribution (the residuals [difference between the predicted and observed values of Y] should be normally distributed around the predicted value with equal variance for all predicted values of Y).
- 3) *Discrimination/Predictiveness* are often summarized using the R^2 (the amount of variation in Y explained by the Xs).
- 4) *Calibration* and fulfillment of assumptions are tested by visual inspection and diagnostic tests on residuals (especially *rvfplot* and *rvpplot*).
- 5) Calibration of a model is particularly important when using a model to make predictions.
- 6) IndVars (Xs) cannot be too highly correlated with each other (begin to worry if you have single correlation > 0.75 or multiple correlations >0.5)

Below, you will find a check list of questions to ask as you plan and do linear regression.

Step 1. Carefully Define the Study Objective/Question, Hypothesis & Conceptual Model.

- a) What is the precise association that you wish to address and how would you state this as a study objective/question?
- b) Why have you selected this study question?
- c) What is your idealized model/design (i.e., what will affect or influence my outcome measure [DepVar] and how)?
- d) What is your study model/design (i.e., which components of my theoretical model are not adequately measured and how important are these omissions?

Step 2. Develop a thorough understanding of your data the variables you wish to put in your regression model?

- a) What does one observation in your dataset represent (one encounter, one person, one person-year, one clinic, etc)? Is this the unit of analysis that makes sense for your data analysis?

- b) Explore the variables in the dataset using *-list -tab -sum,d-* (are they complete, sensible, skewed, etc? Do any Xs have lots of missing values?)
- c) Is your DepVar (Y) heavily skewed? Do any of your IndVars (Xs) have strong (>0.5) correlations with each other?
- d) How many IndVars can I evaluate (limit = 1 IndVar for every 10-20 observations of Y)

Step 3. What are the magnitude and statistical significance of the bivariate associations between the DepVar [Y] and the IndVars [Xs]?

- a) What specific DepVar(s) and IndVar(s) are needed to address my study objective/question?
- b) Always look at the raw data (scatter plots, cross tabs, etc) and do not just look at the p values and summary statistics.

Step 4. What Are the Effect Sizes & Statistical Significance of the Independent (Multivariate) Association(s)?

- a) Have I included all available confounders in the model? Have I included things that may be pathways and not confounders at all?
- b) What happens if I put Xs in one domain at a time (ie., sociodemographics, health status measures, attitudes and behaviors, etc..)?
- c) What is the final sample size of your model and how does that compare to the sample size of Y?
- d) What are the individual associations? What interaction terms should be evaluated (if you have adequate power to look at more variables)?
- e) What is the predictiveness of the model?
- f) Do the diagnostics and plotting of the residuals suggest that the model is well calibrated and meets the assumption of homoscedasticity? If not, consider 1) transforming the DepVar, 2) looking at interaction terms, 3) transforming IndVars.
- g) What does it all mean?

A Check List for Interrogating a Linear (OLS) Regression Model Using Stata

Did you lose sample size?

Does the number of observations in the regression model differ much from the number of observations with a non-missing value for the DepVar? If yes, you need to impute values for one or more of your IndVars.

How predictive (discriminate) is your model?

What is the level of discrimination (ie, predictiveness) of your model? R^2 = amount of variance of the DepVar that is explained by the IndVars in the model.

Do I have a problem with multicollinearity?

Use the *vif* command, and if the VIF for a variable is greater than 5 (or the tolerance $[1/VIF]$ is <0.2), then you should probably do further examination for multicollinearity. If the VIF for a variable is > 5 , try to predict that variable using the other IndVars in regression analysis. If an IndVar can be predicted using the other IndVars with an R^2 greater than 0.80, than that variable (or one of the key predictors of that variable) should be dropped from the analysis.

Use *rvfplot* and *rvpplot* to check the following:

Rvfplot allows you to test several important attributes of your regression model at once. There are statistical tests for calibration and linearity but they are less important than the visual inspect of these graphs.

Is my model well calibrated and are the associations linear?

Type *rvfplot*, *xlab ylab*. Are the residuals distributed roughly equally above and below the zero point throughout the range of predicted values of your DepVar? Is there evidence of non-linearity or heteroscedasticity? If yes, you need to examine, a) whether your DepVar should be transformed (log, sqrt, exponentiate, etc), b) whether some of your IndVars need to be transformed (*rvpplot [IndVar]*, *xlab ylab* can be a good way to check for this), c) whether there are interactions between variables.

Are my residuals normally distributed?

It is not that important that the residuals are normally distributed across zero, however, it is important that there are more residuals that are close to the line than are further from the line and that the pattern is fairly symmetric throughout the distribution.

Are there important Outliers?

When looking at the *rvfplot* output, are there a few variables far above or below the zero point? If so, an effort to check for data entry or coding errors should be perform and if it is suspected that there is an error, this observation should be fixed or dropped. Also, are there a few observation that have much higher or lower observed values of the DepVar (deviate from the pack horizontally on the graph)? If so, these observations should be dropped even if they are valid values, since you do not have enough observations in this distribution of the DepVar to make estimating these values statistically valid.