Overview

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/
Last revised January 8, 2015

Linear regression model (Population)

$$Y_j = \alpha + \beta_1 X_{1j} + \beta_2 X_{2j} + ... + \beta_k X_{kj} + \varepsilon_j = \alpha + \sum_{i=1}^{k} \beta_i X_{ij} + \varepsilon_j = E(Y_j | X) + \varepsilon_j$$

β_i = partial slope coefficient (also called partial regression coefficient, metric coefficient, or just regression coefficient). It represents the change in E(Y) associated with a one-unit increase in X_i when all other IVs are held constant.

α = the intercept. Geometrically, it represents the value of E(Y) where the regression surface (or plane) crosses the Y axis. Substantively, it is the expected value of Y when all the IVs equal 0.

ε = the deviation of the value Y_j from the mean value of the distribution given X. This error term may be conceived as representing (1) the effects on Y of variables not explicitly included in the equation, and (2) a residual random element in the dependent variable.

Parameter estimation (Sample). In most situations, we are not in a position to determine the population parameters directly. Instead, we must estimate their values from a finite sample from the population. The sample regression model is written as

$$Y_j = a + b_1 X_{1j} + b_2 X_{2j} + ... + b_k X_{kj} + e_j = a + \sum_{i=1}^{k} b_i X_{ij} + e_j = \hat{Y}_j + e_j$$

where a is the sample estimate of α and b_k is the sample estimate of β_k.

Hypothesis testing. We use the sample to test hypotheses about the population parameters. Some typical hypotheses:

H_0: $\beta_1 = 0$ H_A: $\beta_1 <> 0$	One of the Betas = 0. Use a T test or an F test.
H_0: $\beta_1 = \beta_2 = \beta_3 ... = \beta_k = 0$ H_A: At least one beta <> 0	Test of whether any Betas differ from 0. Equivalent to a test of $R^2 = 0$. F test
H_0: $\beta_2 = \beta_3 ... = \beta_k = 0$ H_A: At least one of the above betas <> 0	Test whether a subset of the betas = 0. Same as a test of $R^2_{constrained} = R^2_{unconstrained}$. F test
H_0: $\beta_1 = \beta_2$ H_A: $\beta_1 <> \beta_2$	Test whether two betas are =. F test, or maybe a t-test.
H_0: $\beta^{(1)} = \beta^{(2)}$ H_A: $\beta^{(1)} <> \beta^{(2)}$	Test whether the betas are the same in two different populations, e.g. is the model for men the same as the model for women? F test

Assumptions and violations of assumptions

Assumptions concerning correct model specification. *One of the most critical assumptions you make is that the model is correctly specified, i.e. it really is the case that* $Y_j = \alpha + \sum_{i=1}^{k} \beta_i X_{ij} + \varepsilon_j$ *. Incorrect model specification can result in biased parameter estimates and violations of other assumptions. The following assumptions all follow from the requirement that the model be correctly specified. A good theory, and an accurate understanding of what a model actually means and implies, can help to avoid model misspecification. Statistical techniques can sometimes help you detect and correct specification problems.*

Assumption	Possible violations
All relevant variables are included. Irrelevant variables are excluded.	• Relevant variables are excluded, i.e. there are other variables which also affect Y. Result can be biased parameter estimates • Irrelevant variables are included. Result can be higher than necessary standard errors.
The effect of the IVs is linear. (NOTE: It isn't enough just to have identified which IVs are related to the DV. You must correctly specify how they are related, e.g. the relationship is linear or nonlinear.)	• The effects of the IVs are not linear, e.g. there might be a curvilinear rather than linear relationship between variables. • Variables are not measured at the interval level, e.g. religion (1 = Catholic, 2 = Protestant, 3 = Other.)
Subgroup differences do not exist — or else have been incorporated into the model.	• Two or more different populations have been mixed together. The model for one population is different than the model for the others. It may be that different variables are important. Or, the same variables may be relevant, but the strength and/or direction of their effects differs. For example, a model that explains men's earnings may be different than the model that explains women's. There should either be separate models for each group, or the model should include interaction terms that can account for group differences.
There is no perfect collinearity — no independent variable is perfectly linearly related to one or more of the other independent variables in the model. (Multicollinearity)	• Improper use of dummy variables • One of the IVs has been computed from the others • Less extreme, but still problematic, is the case where IVs are highly but not perfectly correlated with each other. Even though the model may be correctly specified, a high degree of multicollinearity can make it difficult to detect significant relationships.

Assumptions required for parameter estimation and hypothesis testing with OLS (Ordinary Least Squares). *If the following assumptions are not valid, then OLS (Ordinary Least Squares) estimation can result in biased parameter estimates, inflated standard errors, or both. Model misspecifications (see above) may lead to such violations. In other cases, it may just be that these assumptions do not hold in the real world. Alternative estimation techniques, or "tricks" with OLS, may be necessary.*

Assumption	Possible violations
$E(\varepsilon_j) = 0$ for all j.	• Fortunately, violation of this assumption only affects the estimation of the intercept
$V(\varepsilon_j) = \sigma^2$ for all j. That is, the variance of the error term is constant. (Homoskedasticity)	• Larger values of Y have larger errors, e.g. the error terms for those who make $100 a week tend to be smaller than the error term for those who make $1,000 a week. • Dependent variable is dichotomous.
$COV(\varepsilon_j, \varepsilon_h) = 0$ for j ≠ h. That is, the error terms are uncorrelated, there is no autocorrelation.	• Cases have not been selected independently of each other; for example, both husbands and their wives are included as separate cases. • Data are longitudinal; same subjects are interviewed at multiple points in time
$COV(X_i, \varepsilon) = 0$ for all i. That is, each independent variable is uncorrelated with the error term.	• Relevant variables are excluded from the equation • There is reciprocal causation — X is a cause of Y, and Y is a cause of X. For example, two friends may each influence each other. Put another way, the model is nonrecursive.
For each set of values for the k independent variables, ε_j is normally distributed.	• This assumption is often violated. However, this assumption is necessary *only* for tests of statistical significance; its violation will have no effect on the estimation of the parameters of the regression model. Further, the assumption is really only critical in small samples. The central limit theorem tells us that, in large samples, even if the error term is not normally distributed in the population, the sampling distribution of the partial slope coefficients will be normally distributed.

Assumptions concerning the data and the sample. *The following assumptions are related to data collection and measurement. These issues are also often discussed in research methods classes. Violations of these assumptions typically result in biased parameter estimates. Statistical solutions are sometimes possible, but in other cases the main "cure" is to collect better data.*

Assumption	Possible violations
All variables are measured without error	• Variables suffer from random measurement error. Reported values are too high or too low, perhaps because of faulty recall or sloppy measurement. • Variables suffer from systematic error. There may be a consistent upward or downward bias; the variable may measure something other than what was intended. For example, student reports about how good a teacher is may actually reflect how tough the teacher is.
The data are a random and representative sample of the larger population	• The sampling frame is biased, e.g. names pulled from a phone book will not include those without phones or those with unlisted numbers. • Not all those selected responded, e.g. less educated people did not respond; the people who responded were exceptionally interested in the topic; some of those selected were never reached • Those who responded did not answer all questions, e.g. some may feel some questions are too sensitive, some questions may be too hard or too unclear for everyone to answer. Whenever data are missing, a key concern is whether it is missing on a random bias, or is there something systematically different about the excluded information.
Weighting of cases is correct	• Most statistical techniques are taught based on the assumption that simple random sampling was used. However, many data sets often use more complicated sampling schemes. Cases have differing probabilities of selection and/or have been selected using stratification or clustering techniques. Failure to take this into account can lead to biased parameter estimates and/or incorrect standard errors.

Course Outline. The course will address how you can prevent, detect, or deal with violations of the above assumptions. It will also show how to develop and test causal models using statistical techniques. These are not independent subjects; as we'll see, violations of assumptions sometimes occur because of characteristics of the data, but they also sometimes occur because models are mis-specified.