Overview of Regression Assumptions and Diagnostics

Assumptions

Statistical assumptions are determined by the mathematical implications for each statistic, and they set the guideposts within which we might expect our sample estimates to be biased or our significance tests to be accurate. Violations of assumptions therefore should be taken seriously and investigated, but they do not necessarily always indicate that the statistical test will be inaccurate. The complication is that it is almost never possible to know for certain if an assumption has been violated and it is often a judgement call by the researcher on whether or not a violation has occurred or is serious.

You will likely find that the wording of and lists of regression assumptions provided in regression texts tends to vary, but here is my summary.

Linearity. Regression is a summary of the relationship between X and Y that uses a straight line. Therefore, the estimate of that relationship holds only to the extent that there is a consistent increase or decrease in Y as X increases. There might be a relationship (even a perfect one) between the two variables that is not linear, or some of the relationship may be of a linear form and some of it may be a nonlinear form (e.g., quadratic shape). The correlation coefficient and the slope can only be accurate about the linear portion of the relationship.

Normal distribution of residuals. For t-tests and ANOVA, we discussed that there is an assumption that the dependent variable is normally distributed in the population. More accurately, ordinary least squares regression (and therefore the statistical tests it subsumes), assume that the residuals, or the conditional values of the dependent variable, are normally distributed. In other words, after we use predictors to account for the variance in Y in the regression model, the assumption states that the residual values are normally distributed in the population. As we know from the central limit theorem, we can often have population distributions that are fairly nonnormal without any serious impact on the statistical tests, because the sampling distribution will tend to be normal. This is true for regression analysis as well (see Box & Watson, 1962; Lumley, Diehr, Emmerson, & Chen, 2002, for example). It is very important to be aware that this assumption only pertains to the dependent variable Y not the independent variables (predictors), X. There is no assumption about the independent variable distribution. In fact, it is perfectly reasonable even to have a binary independent variable, which is not at all normally distributed.

Residuals are independently and identically distributed (i.i.d). This assumption combines two points. The first is that residuals, or observations, are independent of one another and thus are not correlated. Generally, we tend to take for granted that this is true, but we need to be aware of some circumstances in which it may not be true. When couples, family members, students in classrooms, or employees within companies, are treated as separate cases in a data set, we may be violating this assumption, because they will tend to have similar values within units (e.g., similar IQ scores within a family) and are therefore correlated, or not independent. The second part of the i.i.d. assumption is that the distributions of Y for each value of X (i.e., the conditional distribution, $Y|X$) are the same for all values of X. This assumption is usually stated in terms of having the same variance at each value of X, or homogeneity of variance (heteroscedasticity or constant variance)—what we know in ANOVA terms as equal group variance. The assumption is broader in that it states that the distributions are identical, so the shape, rather than just the variance is the same.

No measurement error in the independent variables. As we discussed last term, measurement error tends to attenuate (decrease the magnitude) of the correlation coefficient. For unstandardized slopes, that pertains to the independent (predictor variables), because X appears in the denominator of the slope equation. Measurement error increases the variance of X and therefore it decreases the magnitude of the

slope. Standardized coefficients are impacted by measurement error in either X or Y, having the same attenuating effect. In multiple regression, the attenuating effect of measurement error impacts the correlations among the independent variables (r_{12}) and the correlation among covariates (other independent variables and the dependent variable (r_{y2}). The consequence is that covariates may not be adequately controlled to the extent that they have measurement error. The assumption is sometimes stated in terms of the requirement that X is "fixed" (in contrast with X being "random") which indicates that each time X is measured the value will be the same as long as the true score is the same.

Diagnostics

There are two basic approaches to exploring whether assumptions violations may be a serious concern. One is by inspecting numeric indexes and one is by inspecting graphs of the data. Below are some common indexes used for quantifying the extent to which violations of assumptions might have occurred, and I will illustrate some graphical methods in subsequent handouts. Keep in mind that the point of most of these indices is not to determine whether there is a "statistically significant violation" of the assumption, because we cannot determine for certain whether a violation has occurred or whether it is severe enough for us to be concerned. There are, however, some recommended cutoffs for their values that have been suggested by authors based on simulation studies (see "Summary of Regression Diagnostics" handout for details).

Normal distribution of residuals. Most of the diagnostics relevant to the normality assumption are focused on identifying outliers. Outliers cannot be strictly defined and identifying an outlier may not necessarily mean that it should or can be legitimately removed. Outlier diagnostics generally identify whether an observation is an outlier on X (which are not really relevant to the assumption violation), an outlier on Y, an outlier on both, or an influential data point, meaning that the presence of the outlier affects the regression estimates.

There are some significance tests for whether a distribution is likely to be normal in the population, and these can be applied to residuals from a regression model. The difficulty with interpreting these tests is that nonnormal distributions tend to be less problematic with larger sample sizes than with smaller sample sizes (think about the impact of outliers, for example), and statistical tests may lack sufficient power with small samples and identify small effects as significant with large samples.

Although researchers may often use a ratio of skewness or kurtosis to their estimated standard errors, comparing the ratio to a z- or t-distribution for significance, this is probably not the best test. There are a number alternative methods that have been proposed (see D'Agostino, 1986 and DeCarlo, 1997 for reviews). DeCarlo has some macros that will compute some of the tests, http://www.columbia.edu/~ld208/. There are several other tests that examine whether the distribution conforms to a theoretically normal distribution, such as the Wilks-Shapiro test (Shapiro & Wilk, 1965) or the Looney-Gulledge test (Looney & Gulledge, 1985). None of these tests can really distinguish whether there are serious assumption violations or not, only significant ones.

Two common indices for outliers on X variables are *leverage* (h_{ii}) and *Mahalanobis distance* (MD). Leverage indicates the distance from the mean of X, and with multiple predictors the collective distance of all of the X variables from their collective mean (a "multivariate" distance measure). The equation below is simplified for the single predictor case,

$$h_{ii} = \frac{1}{n} + \frac{\left(X_i - M_x\right)^2}{\sum x^2}$$

where M_x is the mean of X, and x is the deviation score for X, where $x = X - \overline{X}$. SPSS adjusts this value and reports the "centered" leverage, which is equal to $h^*_{ii} = h_{ii} - (1/n)$.

Mahalanobis distance is another common multivariate distance measure, used widely across many applications, that provides a single number for each case about the distance of a case's X values, taken together, from a central multivariable mean for all cases. Because multiple variables are involved, the equation is typically expressed in matrix notation,[1] but one can more simply consider a case where all of the X variables are uncorrelated. In that instance, Mahalanobis distance is the sum of squared independent z-scores (Darlington, 1990)

$$MD_i = z^2_{x1} + z^2_{x2.1} + \ldots + z^2_{xk.1\ldots k}$$

or, in the independent case, it is the sum of squared deviations of the variables taking into account their variances, MD also can be closely related to leverage, $MD_i = h^*_{ii}(n-1)$.

Outliers on Y are usually defined by the residuals, $e = Y - \hat{Y}$. Residuals are hard to interpret unless they are standardized, however. And as so many quantities in statistics, there are several suggestions about how to standardize them. *Standardized residuals* simply divide by the standard deviation of the residuals. An improvement on this index is to calculate the distance of the observed point, Y_i, from the predicted point when the regression has been re-estimated without using the observed point, $Y_{(i)}$. The i in parentheses (i) subscript indicates that the case is deleted in the computation. *Studentized residuals* use information about the fit of the model, the mean squared residual (MS_{res}, or mean-square error, MSE), in the denominator. The *deleted studentized residual* is more sensitive to outliers, because the potentially problematic point is removed when estimating the predicted point,

$$studentized\ deleted\ residual = \frac{e_i}{\sqrt{MS_{res(i)}\left(1 - h_{ii}\right)}}$$

Influence statistics identify cases with outlying values on X and Y, such as *Cook's distance*, or the effect of removal of the case on the predicted values, such as *dffits*, or the regression coefficients, *dfbetas*.

$$Cook's\ distance = \frac{\sum \hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{\left(MS_{res}\right)\left(h_{ii}\right)}}$$

$$dfbeta = \frac{B_{ki} - B_{k(i)}}{SE_{B_{k(i)}}}$$

where the subscript k is represents a coefficient for particular predictor, such as B_1.

$$dffit = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\sqrt{\left(MS_{res}\right)\left(h_{ii}\right)}}$$

[1] Mahalanobis distance involves deviations of X variables from their means, taking into account all of the variance-covariances among the X variabes, $MD_i = (\mathbf{x} - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})$

The standardized versions of *dfbetas* or *dffits* (not shown here) may be more useful for interpreting magnitude.

Nonconstant variance

There are several tests for constant variance (or homoscedasticity) across the values of X that attempt to assess (at least one aspect of) the i.i.d. assumption. Nonconstant variance can lead to incorrect standard errors, although it does not affect the regression coefficients themselves. A modification of the *Levene's test* for continuous predictors is also known as the *Brown-Forsythe test* (see Box 4.4.1 in Cohen, Cohen, West, & Aiken, 2003). The *Cook-Weisberg* (Cook & Weisberb, 1983) and *White* (1980) tests are alternatives. As with the normality tests, these tests could have insufficient power in small samples and power to detect minor violations in large samples, precisely the opposite of the circumstances when detecting violations may be the most important. There is not an accepted magnitude cutoff for these tests (although see your text, pp. 120, 146 for one common cutoff), so deciding when there is really a serious violation is difficult. Also note that these tests may also be affected by problems other than nonconstant variance, such as nonlinear relationships.

Nonindependent Residuals

Testing for the non-independent residuals, another aspect of the i.i.d assumption, can be done with the *Durbin-Watson test* (Durbin & Watson, 1950,1951), also with similar potential problems with interpreting significance and deciding whether the magnitude is important.

Nonlinear Relations

Nonlinear relationships can be examined through scatterplots of the original variables, or more sensitively, by examining a scatterplot of residuals against X values. Nonlinear relationships can also be tested by modifying the regression model to test for them, an analysis approach which we will discuss soon.

References

Box, G. E., & Watson, G. S. (1962). Robustness to non-normality of regression tests. *Biometrika, 49*(1-2), 93-106.
Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika*, *70*(1), 1-10.
Darlington, R.B. (1990). *Regression and linear models.* New York: McGraw-Hill.
D'agostino, R. B., Belanger, A., & D'Agostino Jr, R. B. (1990). A suggestion for using powerful and informative tests of normality. *The American Statistician*, *44*(4), 316-321.
DeCarlo, L. T. (1997). On the meaning and use of kurtosis. *Psychological methods*, *2*(3), 292.
Durbin, J.; Watson, G. S. (1950). "Testing for Serial Correlation in Least Squares Regression, I". *Biometrika. 37* (3–4): 409–428.
Durbin, J.; Watson, G. S. (1951). "Testing for Serial Correlation in Least Squares Regression, II". Biometrika. 38 (1–2): 159–179.
Looney, S. W., & Gulledge Jr, T. R. (1985). Use of the correlation coefficient with normal probability plots. *The American Statistician*, *39*(1), 75-79.
Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual review of public health*, *23*(1), 151-169.
Shapiro, S. S., and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," Biometrika, 52, 591-611.
White, H. (1980). "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". *Econometrica, 48* (4): 817–838.