Evaluating Assumptions Using the Data

Before looking at the data, you should always attempt to assess the validity of the assumptions using your subject matter knowledge about the data-generating process. The data values can't tell you everything, and often, you can be misled by over-interpreting artifacts of your data set. Recall also that the assumptions do not refer to the data; they refer to the process that generated the data, so even when you do look at the data, you have to imagine data that was not observed that could have been produced by the model, or you have to simulate data from the assumed form of the model to arrive at the correct understanding and conclusion. There is no one canned analysis that can apply to the analysis of a given data set, so don't think that all the answers lie in the data - there is no substitute for careful thinking that involves a clear understanding of the meaning of statistical model as a producer of data.

That having been said, this document concerns the investigation of the assumptions using the data. But whenever possible, you should use your intellect, aided by simulation if necessary, in addition to using the observed data. And it bears repeating that the assumptions all refer to the data-generating process; and that the assumptions do not refer to the actual observed data. So the data shed light on the validity of the assumptions only insofar as they reflect the underlying data-generating process. In particular, data with small samples sizes do not represent the underlying process very accurately, so it is difficult to assess validity of assumptions when you have a small sample size.

I have broken this document into "Graphical Methods" and "Testing Methods." The graphical methods are informal, and require some judgment, as well as care in producing clear graphs, to make proper interpretation. The general consensus in the statistical and scientific communities is that graphical methods are preferred to testing methods, because the graphical methods are more understandable, and they answer the relevant question more directly. Since all assumptions are violated to a degree, the relevant question is not "are the assumption violated?" but rather "are the assumptions so badly violated that alternative methods are needed?"

However, testing methods have been used traditionally. People like to think that there is something "objective" about them, and it is possible that an unenlightened referee will suggest that you use them before agreeing that you paper is publishable. Therefore, we need to discuss the testing methods as well. But the paradigms are shifting, driven by improved understanding of statistical theory and by data mining applications where large data sets are used. But testing methods are becoming less popular than descriptive data-analytic and subjective methods for assumption checking. In fact, testing methods in general have been soundly criticized in all areas of science, not only for assumption checking, but also for testing research hypotheses. Instead, many science journals are moving towards reporting "effect sizes" rather than results of hypothesis tests to assess the validity of research hypotheses. An "effect size" is a measure of the "size" of the deviation, rather than the statistical significance. Effect sizes are more useful to judge the practical significance of a result than are the p-values that result from statistical tests, since p-values only judge statistical significance.

On the other hand, chance variation is real, and has an effect. If you do see something in the graphs that suggests a deviation from the assumption, your first question should be “is this
difference explainable by chance alone?" The \( p \)-value from the test helps you answer that question.

**Benefits of graphics:**
- Transparency: the graph shows exactly what is going on. The statistical tests tend to hide this information.
- "Practical importance" and/or "economic significance" are easily determined using graphical methods, but not by statistical tests.
- Unlike statistical tests of assumptions, larger sample sizes always point you closer to the best answer when you use well-chosen graphs.

**Concerns with graphics:**
- Interpreting graphs requires practice, judgment, and some knowledge of statistics
- Producing good graphs requires some skill and practice (see e.g., Ed Tufte's classic book)

**Benefits of tests:**
- Objectivity. The test provides an objective measure of what is explainable by chance variation alone.
- History. Some science journals have historically required them (although the paradigm has shifted), so have to know how about tests to read the literature.

**Concerns with tests:**
- All assumptions are null hypotheses. You can't prove nulls are true.
- With small sample sizes, all tests have low power. Thus you will likely fail to reject the null, and thus may be inclined to report that "the assumptions are valid" just because you don't have enough data to detect deviations from the assumptions.
- With large sample sizes (e.g. data mining applications), even small but practically unimportant deviations will be flagged as "statistically significant." The statistical conclusion will be correct, i.e., the assumption tested is violated, but the deviation might not be large enough to require remedial action. (Review the effect of sample size on the power of statistical tests. Review also the distinction between "statistical significance" and "practical significance").
- The hypothesis testing methods aren't perfectly "objective." Is random generation a valid assumption? Usually not. Nevertheless, the random generation assumption is the basis for all statistical inference, including tests for validity of assumptions. I don't mind making this assumption, but it is NOT purely objective - it always contains a subjective element. The objective elements of this assumption are (i) that data are variable, (ii) that probability models produce variable data, and (iii) that a model is good if the data it produces match the real process data. However, it is clear that random generation does not produce the real data; rather, this is our mental model for reality. As mental model, it necessarily contains elements of subjectivity. You might even argue that statistical graphs are *more objective* than tests, because they require no assumptions whatsoever.

**COMMENT ON ORDER OF IMPORTANCE IN EVALUATING ASSUMPTIONS:** The given order is a suggested order but not a hard and fast rule. Later assumptions (particularly uncorrelated errors and normality) use the residuals \( e_i = y_i - \hat{y}_i \), where the predicted values \( \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \) are based on the linear fit. If the assumption of linearity is badly violated, then these estimated residuals will not accurately reflect actual deviations of \( Y \)-values from the means.
of their respective distributions, and are therefore less useful in the subsequent evaluation of the assumptions. Evaluation of these assumptions in this case would best be done by fitting a more appropriate (non-linear) model, and using that model to calculate the predicted values.

On the other hand, one might also argue that outliers should be investigated first, since the predicted values can be badly distorted by outliers. In any event, I will consider the assumptions in the order linearity, constant variance, uncorrelated errors, and normality.

1. LINEARITY:

GRAPHICAL METHODS

1.G.A. Plot the ordinary (x,y) scatterplot. Look for evidence of curvature. NOTE: If you are going to claim there is curvature, such curvature MUST MAKE SENSE IN THE CONTEXT OF THE SUBJECT MATTER OF THE DATA-GENERATING PROCESS. For example, are there boundary constraints that force curvature? Is there a reason that the response should increase for small x, but decrease for large x (for example, as in the case of the product preference vs product complexity, or crop yield vs fertilizer relationships)? Remember, the assumption is about the process that produced these data, not about the data themselves, so don't over-interpret the data. Imagine the forces at work behind the data; i.e., consider the data-generating process when discussing reasons for curvature to exist.

1.G.B. Plot the residual plot (x,e). This is another look at the previous plot, with the e=0 line representing the least squares line in plot 1.G.A. It offers a "magnified" view. Look for upward or downward "U" shape to suggest curvature.

1.G.C. Plot (ˆy,e). In simple regression (i.e., one x), this plot is identical to plot 1.G.B., with the exception that the horizontal scale has been linearly transformed. It looks about the same as plot 1.G.B., except when the estimated slope is negative, in which case the horizontal axis is "reflected" - large values of x map to small values of ˆy and vice versa. Use this plot just like 1.G.B.

In simple regression, this plot offers no advantage over 1.G.B. However, in multiple regression, this plot is invaluable for getting a quick look at the overall model, in that there is just one plot, rather than several.

1.G.D. Use the LOESS smoother on all of the above - look for curvature that makes sense from subject matter standpoint. Don't over-interpret curvature shown in the LOESS plot; remember the LOESS plot is from the data, not from the model. Remember that model produces data, not vice versa. The data results are simply ESTIMATES of the underlying model parameters. In particular, LOESS estimates the underlying function using a type of “local averaging,” so where the x data are more sparse, the estimates of the mean of y are less reliable.
HYPOTHESIS TESTING METHODS

1.T.A. Expand the model to include a quadratic term. Significance of the quadratic term suggests evidence of curvature; insignificance suggests that any evidence of curvature in the observed data is explainable by chance alone. CAVEAT: As with all significance tests, one should not take the result of the test as a "recipe" for model construction. If significant, is the curvature dramatic enough to warrant the additional modeling complexity? Do the predictions differ much, whether you use a model for curvature or the ordinary linear model? Note also that models employing curvature (particularly quadratics) are notoriously poor at the extremes of the x-range(s). So the linear model might again be preferred on this count, even if the curvature is "significant."

Conversely, if the quadratic term is insignificant, you SHOULD NOT claim that the function is linear. Instead, you have failed to detect evidence of curvilinearity. Recall the difference between results that are explainable by chance alone and those that are explained by chance alone. In reality, in the true process, there is some curvature. Why would nature choose to put the means all exactly on a straight line? Just because we humans like straight lines for their simplicity? No, that is not how nature works. Nature doesn't particularly care whether or not we like straight lines. Curvature is always real in nature; the question is whether it is so pronounced that you have to use a model other than a straight line model.

1.T.B. Group the x data into categories and fit the model using "dummy variables" instead of the x variable in question. This test allows for more general alternatives than the quadratic, i.e., it does not presume that the nonlinearity takes a quadratic form. But this test is usually less powerful than the test using the quadratic term.

2. CONSTANT VARIANCE: (Also called the "homoscedasticity" assumption. Non-constant variance is called "heteroscedasticity.")

GRAPHICAL METHODS

2.G.A. Plot (ŷ,e). Look for changes in the pattern of vertical variance for differing ŷ. The most common appearance in an "outward funnel", indicating increasing variability of y at larger levels of ŷ. Whatever pattern of variability is claimed, this pattern must make sense in the CONTEXT of the SUBJECT MATTER OF THE DATA-GENERATING PROCESS (e.g., physical boundaries on data force smaller variance when the data are closer to the boundary). Do not try to over-interpret. If the pattern is not obvious, and sensible, than there probably is no pattern worth discussing. Recall also that the assumptions refer to conceptual distributions, not specifically the data, so any explanation must involve an extrapolation from the data that you can see to the conceptual distributions that you cannot see.

2.G.B. Plot the absolute values of the residuals. Plot |e| (on the vertical axis) against ŷ (on the horizontal axis). Use the LOESS smoother; this curve estimates the variability of the residuals as a function of ŷ. Theoretically, it should be a flat line. If there is a distinct increasing (or
decreasing) trend in the size of the residuals, this is evidence of heteroscedasticity. However, do not over-interpret. Recall that data are variable, so even if homoscedasticity is true in reality (i.e. the true model is homoscedastic), the curve will not be a perfectly flat line, just due to randomness. The hypothesis test for homoscedasticity will help you to decide whether the observed deviation from a flat line is explainable by randomness alone.

TESTING METHODS

2.T.A. Consider different ranges of the $x$-variable (or variables). Conduct a test of equality of variances for the different subgroups using Bartlett's test.

2.T.B. Test for regression trend in the $|e|$, $\hat{y}$ data set, where $|e|$ is the dependent variable and $\hat{y}$ is the predictor variable.

2.T.C. Use the Breusch-Pagan test.

2.T.D. Model the mean and variance portions of the model simultaneously and then test the null hypothesis that the variance is constant using a full model/ reduced model likelihood ratio test. This can be done using my excel spreadsheet "ML estimation for linear, nonlinear, heteroscedastic, and nonnormal regression models using SOLVER of Microsoft EXCEL" or using canned procedures such as PROC MODEL or PROC MIXED.

General comment: Many tests for variances are very sensitive to the normality assumption—this is an additional problem with tests for variances (in addition to the other general problems about testing listed above.) Because of this sensitivity, you will not be able to state that a rejection of the null hypothesis is due to truly different variances, or simply due to nonnormality when you use one of these classical tests. Investigate "robust tests for homoscedasticity" for alternatives.

3. UNCORRELATED ERRORS

Evaluation of this assumption depends on the type of data you have, whether time-series, spatial, repeated measures, hidden grouping variable, etc. Note: with time-series data, correlated errors are sometimes called "autocorrelated".

GRAPHICAL METHODS

3.G.A. (For time-series data). The $t$ denote the time indicator ($t=1,2,...,T$), and plot $(t,e)$. Look for systematical sinusoidal-type functional patterns to suggest failure of this assumption. A completely random appearance is consistent with uncorrelated errors.

3.G.B. (Time-series again). Plot $(e_{t-1}, e_t)$, where $e_{t-1}$ is the lag 1 residual (i.e., the residual value immediately preceding in time sequence). Any appearance of a linear trend in this plot suggests dependence between the current residual and the immediately preceding residual, a violation of the assumption. A random scatter with no trend is consistent with uncorrelated errors.


3.G.E. (Clustered data, repeated measures). Plot the (Clustering variable, e), using side-by-side boxplots. Do the residuals tend to be systematically larger than zero for some values of the clustering variable, and systematically smaller for other values? This suggests correlation induced by the clustering variable.

TESTING METHODS

3.T.A. (Time series) The Durbin-Watson test is commonly used. This is just a test for first-order autocorrelation, and there are various other tests which are equivalent, and supplied routinely by software that can analyze time series data. For example, the test for "white noise" is a test that the residuals are uncorrelated.

3.T.B. More general time series: Fit a model with time series effects (e.g., AR, ARCH, GARCH) and test their significance.

3.T.C. (Clustered data, repeated measures): fit a covariance structure model using PROC MIXED, PROC GLIMMIX, or other software that fits non-independent covariance structures, and test whether the covariance parameters are zero.

4. NORMALITY

GRAPHICAL METHODS

4.G.A. Look at the (x,y) and (x,e) scatterplots used in testing linearity and constant variance assumptions. Are there any outliers that are indicative of nonnormality? Any evidence of skewness? Are strong discreteness characteristics shown?

4.G.B. Look at the histogram of the residuals. Is it approximately bell-shaped?

4.G.C. Examine the normal quantile-quantile plot of the residuals. Do the values fall along a reasonably straight line? Do the data values fall on distinct lines, indicating discreteness?

TESTING METHODS

4.T.A. Shapiro-Wilk test for normality using the residuals (not the y values). Details on Shapiro-Wilk test from SAS documentation; also a hands-on paper describing how to do it and interpret the results (note that the paper overstates the necessity of normality; specifically Gauss-Markov assumptions are often adequate. But the paper is otherwise ok.)
COMMENT ABOUT NORMALITY EVALUATION AND RESIDUALS. All of the given graphical and testing methods involve the residuals. However, the assumption says that for each X=x, the Y values are normally distributed. The reason we do not use the y's directly is that there are typically not enough y's at any particular X=x to get a good picture of their distribution. So we pool all the residuals together and look at their distribution. A major limitation of this procedure is that obvious non-normal characteristics such as pronounced discreteness (in Y) will be hidden in the plots, and not detected in the tests, when using the residuals, since when residuals all are pooled their values cover more of a continuum than do the actual Y values.

Here is an example that illustrates the concept. The distribution of Y is discrete, taking values 1,2,3,4 and 5, and hence is clearly non-normal. Yet the diagnostic tools of normal q-q plot and the test for normality show little no significant deviation from normality.

data discrete;
do i = 1 to 1000;
   X = rannor(22321)+3;
   Y = round(X + rannor(0),1); if Y>5 then Y=5; if Y<1 then Y=1; /* Y is discrete with values 1,2,3,4,5 */
   output;
end;
proc reg;
   model Y = X;
   plot residual.*nqq./noline nostat nomodel;
run;
%let dataset = discrete;
%let yvar = Y;
%let xvar = X;
proc reg data=&dataset;
   model &yvar = &xvar;
   output out=new r=resid;
run;
proc univariate data=new normal;
   title "Analysis of residuals; &dataset data set";
   title2 "Model uses Y = &yvar, X = &xvar ";
   var resid;
   histogram;
   qqplot;
run; title;

Conclusion: Mainly, you should use your subject matter knowledge about the data-generating process when evaluating all assumptions. After all, the assumptions refer to the process, NOT THE DATA. Use what you know about the data-generating process FIRST, before you even look at the data. For example, if you know that the Y's are highly discrete, then the assumption is violated, period. You don’t need a hypothesis test or normal q-q plot. In fact, if you use such a test and it indicates normality, then it is an obvious Type II error. You will look silly if you report that the distribution is normal in this case.