STAT 704 --- Checking Model Assumptions

• Recall we assumed the following in our model:

(1) The regression relationship between the response and the predictor(s) specified in the model is appropriate
(2) The errors have mean zero
(3) The errors have constant variance
(4) The errors are normally distributed
(5) The errors are independent

• We cannot observe the true errors $\varepsilon_1, \ldots, \varepsilon_n$, but we can observe the residuals $e_1, \ldots, e_n$.
• Assumptions typically checked via a combination of plots and formal tests.

<u>Assumption (1)</u>: Sometimes checked by a scatterplot of Y against X (or a scatterplot matrix of $Y, X_1, \ldots, X_k$).
• We look for patterns other than that specified.

• More generally, we can examine a <u>residual plot</u>:

• Look for non-random (especially curved) pattern in residual plot, indicating a violation of Assumption (1).

<u>Remedies</u>: • Choose different functional form of model.
• Use transformation of X variable(s).

• In multiple regression, separate plots of residuals against each predictor can be useful for determining which X variable may need transforming.

• A formal "lack of fit test" is available (see Section 3.7), but it requires replicate observations at one or more levels of the *X* variables (often not applicable when one or more predictors are continuous).

<u>Assumption (2)</u>: not checked separately → residuals have mean zero by definition.

<u>Assumption (3)</u>: Often the most worrisome assumption.

• Violation indicated by "megaphone" or "funnel" shape in residual plot:

<u>Remedy</u>: Transform the *Y* variable:
Use $Y_i* = \sqrt{Y_i}$ **or** $Y_i* = \ln(Y_i)$

<u>Advanced method</u>: Weighted Least Squares (we will see in Chapter 11)

• Formal tests for nonconstant error variance are available:

<u>Breusch-Pagan Test</u>: Tests whether the error variance increases or decreases linearly with the predictor(s).
• H_0 specifies that the error variance is constant.
• Requires large sample.
• Assumes errors are normally distributed.

<u>Brown-Forsythe Test</u>: • Robust to non-normal errors.
• Requires user to break data into groups and test for constancy of error variance across groups.
• Not natural for data with continuous predictors.

• Graphical methods have the advantage of checking for <u>general violations</u>, not just violations of a specific type.

<u>Assumption (4)</u>: Graphical approach: Look at normal Q-Q plot of residuals.
• Violation indicated by severely curved Q-Q plot.

<u>Remedies</u>: • Transformations of Y and/or X.
• Nonparametric methods.

<u>Formal test for error non-normality</u>:
• The Shapiro-Wilk test (implemented in R and SAS) tests for normality.

• Test based on the correlation between the ordered residuals and their expected values when the errors are normal.

Example (Studio data):

<u>Note</u>: With large sample sizes, the normality assumption is not critical.
<u>Note</u>: The formal test will not indicate the <u>type of departure</u> from normality.

Assumption (5): Typically only a concern when the data are gathered over time.
• Violation indicated by a pattern in the residuals plotted against time.
Remedies: • Include a time variable as a predictor.
• Use time series methods.

Transformations of Variables (Section 3.9):

• Some violations of our model assumptions may be alleviated by working with transformed data.

• If the only problem is a nonlinear relationship between *Y* and the *X*'s, a transformation of one or more *X*'s is preferred.

Possible:

• See diagrams in Figure 3.13, p. 130.

• If there is evidence of nonnormality or nonconstant error variance, a transformation of *Y* (and possibly also *X*) is often useful.

Examples:

• If the error variance is nonconstant but linear relationship is fine, then only transforming *Y* may disturb the linearity. May need to transform *X* also.

• The Box-Cox procedure provides an automatic way to determine the optimal transformation of the type:

<u>Note</u>: When working with transformed data, predictions and interpretations of regression coefficients are all in terms of the <u>transformed</u> variables.

• To state conclusions in terms of the <u>original</u> variables, we typically need to do a <u>reverse</u> transformation.

<u>Example</u> (surgical unit data):

<u>Extra Sums of Squares and Related F-tests</u>

• "Extra Sums of Squares" can be defined as the difference in SSE between a model with "a few" predictors and a model with those predictors, <u>plus some others</u>.

• Recall: As predictors are added to the model, SSE

<u>Example</u>: Predictors under consideration are $X_1, \ldots, X_8$.

Two possible models:

• Book's notation for this:

• Why important? We can formally test whether a certain set of predictors is useless, <u>in the presence</u> of the other predictors in the model.

<u>Question</u>: Are X_2, X_4, X_7 needed, if the other predictors are in the model?
• We want our model to have "large" SSR and "small" SSE. (Why?)
• If "full" model has much lower SSE than "reduced" model (without X_2, X_4, X_7), then at least one of X_2, X_4, X_7 is needed.

→

To test

use

Reject H_0 if

Example above:

• <u>Note</u>: The tests for <u>individual coefficients</u> are examples of this type of test.
Example:

• **To test about more than one (but not all) coefficients in SAS, use a `TEST` statement in `PROC REG`.**

<u>Example</u> (Body fat data): Y = amount of body fat, X_1 = triceps skinfold thickness, X_2 = thigh circumference, X_3 = midarm circumference. Is the set of X_2, X_3 significantly useful if X_1 is already in the model?

<u>Multicollinearity</u>

<u>Note</u>: In the body fat example, the F-test for testing

was but individual t-tests for each of

"Paradoxical" conclusion:

Reason?

Example: The correlation coefficient between triceps thickness and thigh circumference is

• This condition is known as <u>multicollinearity</u> among the predictors.

• With uncorrelated predictors, the model can show us the individual effect of each predictor on the response.

• When predictors are correlated, it is difficult to separate the effects of each predictor.

Effects of Multicollinearity

(1) The model may still provide a good fit and precise prediction of the response and estimation of the mean response.
(2) Estimated regression coefficients (b_1, b_2, …) will have <u>large variances</u> – leads to the conclusion that individual predictors are <u>not significant</u> although overall F-test may be <u>highly</u> significant.
(3) Concept of "holding all other X variables constant" doesn't make sense in practice.
(4) Signs of estimated regression coefficients may seem "opposite" of intuition.

Detecting Multicollinearity

• For each predictor, say X_j, its Variance Inflation Factor (VIF) is:

For any predictor X_j

<u>Remedies for Multicollinearity</u>

(1) Drop one or more predictors from model.
(2) More advanced methods:

(3) More advanced:

<u>Polynomial Regression</u>

• Used when the relationship between *Y* and the predictor(s) is <u>curvilinear</u>.

<u>Example</u>: Quadratic Regression (one predictor):

• <u>Note</u>: Usually in polynomial regression, all predictors are first <u>centered</u> by subtracting the sample mean (of the predictor values) from each *X*-value. This reduces

• Another option: Use "orthogonal polynomials" which are uncorrelated.

<u>Example</u>: Cubic Regression (one predictor):

• Polynomials of higher order than cubic should rarely be used.

• High-order polynomials may be excessively "wiggly" and erratic for both interpolations and extrapolations.

Example:

Polynomial Regression, More than One Predictor

• In the case of multiple predictors, all cross-product terms must be included in the model.

<u>Example</u>: Quadratic Regression (two predictors):

<u>Notes</u>: (1) These models are all cases of the "general linear model" so we can fit them with least squares as usual.
(2) A model containing a particular term should also contain all terms of lower order:

(3) Extrapolation is particularly dangerous with polynomial models.

Examples:

(4) A common approach is to fit a high-order model and then test (with t-test or F-test) whether a lower-order model is sufficient.

Example:

<u>Interaction Models</u>

• An interaction model is one that includes one or several <u>cross-product</u> terms.

Example (two predictors):

• Question: What is the change in mean response for a one-unit increase in X_1 (holding X_2 fixed)?

Example:

• The marginal effect of X_1 on the mean response

• We may see this phenomenon graphically through <u>interaction plots</u>.

Example:

<u>Notes</u>: Including interaction terms may lead to multicollinearity problems. Possible remedy:

• Including all pairwise cross-product terms can complicate a model greatly.
• We should test whether interactions are significant.

Graphical Check:
• Fit model with no interaction.
• Plot residuals from this model against each potential interaction term separately.
• If plot shows random scatter, that interaction term is probably not needed.

Formal F-test: