Sociology 740 John Fox

Lecture Notes

8. Diagnosing and Correcting Nonlinearity and Other Ills

1. Introduction

- The first three topics of the lecture take up the problems of
  - non-normally distributed errors
  - non-constant error variance
  - nonlinearity.
  - The treatment here stresses simple graphical methods for detecting these problems, along with transformations of the data to correct problems that are detected.
- Subsequent topics (time permitting) describe tests of non-constant error variance and nonlinearity for discrete explanatory variables; and diagnostic methods based upon imbedding the usual linear model in a more general nonlinear model incorporating transformations as parameters.

2. Goals:

- To introduce simple methods for detecting non-normality, non-constant error variance, and nonlinearity.
- To show how these problems can often be corrected by transformation and other approaches.
- To demonstrate the application of the method of maximum likelihood to regression diagnostics (time permitting).

3. Example: The SLID Data

- To illustrate the methods described in this lecture, I will primarily use data from the 1994 wave of Statistics Canada’s Survey of Labour and Income Dynamics (SLID).
- The SLID data set that I use includes 3997 employed individuals who were between 16 and 65 years of age and who resided in Ontario.
- Regressing the composite hourly wage rate on a dummy variable for sex (code 1 for males), education (in years), and age (also in years) produces the following results:

$$\begin{array}{rcllll}
\widehat{\text{Wages}} & = & -8.124 & + \; 3.474 \times \text{Male} & & + \; 0.2613 \times \text{Age} \\
 & & (0.599) & \quad (0.2070) & & \quad (0.0087) \\
 & & & + \; 0.9296 \times \text{Education} & & \\
 & & & \quad (0.0343) & & \\
R^2 & = & .3074 & & &
\end{array}$$

4. Non-Normally Distributed Errors

▶ The assumption of normally distributed errors is almost always arbitrary, but the central-limit theorem assures that inference based on the least-squares estimator is approximately valid. Why should we be concerned about non-normal errors?

- Although the *validity* of least-squares estimation is robust the *efficiency* of least squares is not: The least-squares estimator is maximally efficient among unbiased estimators when the errors are normal. For heavy-tailed errors, the efficiency of least-squares estimation decreases markedly.
- Highly skewed error distributions, aside from their propensity to generate outliers in the direction of the skew, compromise the interpretation of the least-squares fit as a conditional typical value of Y.

- A multimodal error distribution suggests the omission of one or more discrete explanatory variables that divide the data naturally into groups.

▶ Quantile-comparison plots are useful for examining the distribution of the residuals, which are estimates of the errors.

- We compare the sample distribution of the studentized residuals, E_i^*, with the quantiles of the unit-normal distribution, $N(0,1)$, or with those of the t-distribution for $n-k-2$ degrees of freedom.
- Even if the model is correct, the studentized residuals are not an *independent* random sample from t_{n-k-2}. Correlations among the residuals depend upon the configuration of the X-values, but they are generally negligible unless the sample size is small.
- At the cost of some computation, it is possible to adjust for the dependencies among the residuals in interpreting a quantile-comparison plot.

- The quantile-comparison plot is effective in displaying the tail behavior of the residuals: Outliers, skewness, heavy tails, or light tails all show up clearly.
- Other univariate graphical displays, such as histograms and density estimates, effectively supplement the quantile-comparison plot.
- Figure 1 shows a t quantile-comparison plot and a density estimate for the studentized residuals from the SLID regression.
  - The distribution of the studentized residuals is positively skewed and there may be more than one mode.
  - The positive skew in the residual distribution can be corrected by transforming the *response variable* down the ladder of powers, in this case using logs, producing the residual distribution shown in Figure 2.
    - The resulting residual distribution has a slight negative skew, but I preferred the log transformation to the 1/3 power for interpretability.
    - Note that the residual distribution is heavy-tailed and possibly bimodal.

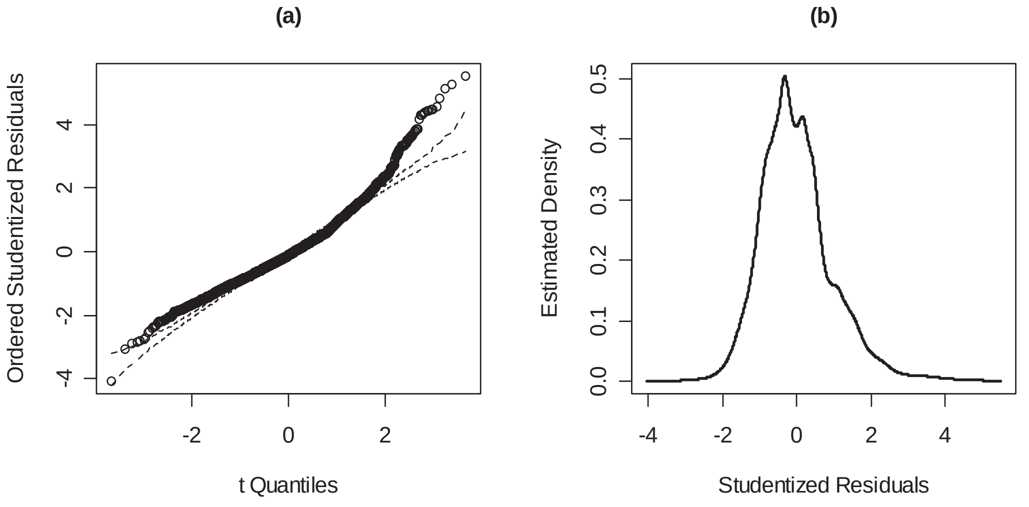

Figure 1. (a) Quantile-comparison plot with point-wise 95-percent simulated confidence envelope and (b) adaptive kernel-density estimate for the studentized residuals from the SLID regression.

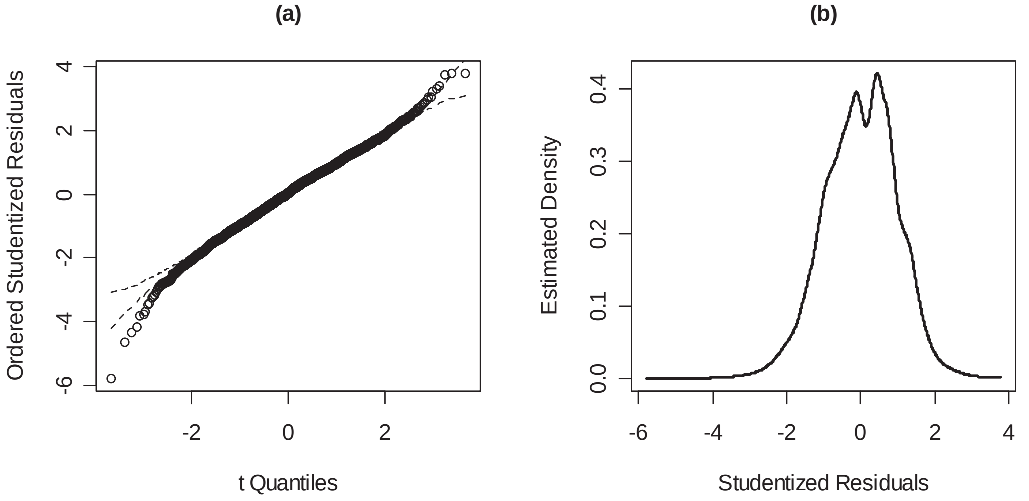

Figure 2. (a) Quantile-comparison plot, and (b) adaptive kernel-density estimate for the studentized residuals from the SLID regression with wages log-transformed.

5. Non-Constant Error Variance

- Although the least-squares estimator is unbiased and consistent even when the error variance is not constant, its efficiency is impaired, and the usual formulas for coefficient standard errors are inaccurate.
  - Non-constant error variance is sometimes termed 'heteroscedasticity.'
- Because the regression surface is k-dimensional, and imbedded in a space of $k+1$ dimensions, it is generally impractical to assess the assumption of constant error variance by direct graphical examination of the data.
- It is common for error variance to increase as the expectation of Y grows larger, or there may be a systematic relationship between error variance and a particular X.
  - The former situation can often be detected by plotting residuals against fitted values;
  - the latter by plotting residuals against each X.

- Plotting residuals against Y (as opposed to $\widehat{Y}$) is generally unsatisfactory, because the plot will be 'tilted'
  - There is a built-in linear correlation between Y and E, since $Y = \widehat{Y} + E$.
  - The least-squares fit insures that the correlation between $\widehat{Y}$ and E is zero, producing a plot that is much easier to examine for evidence of non-constant spread.
- Because the *residuals* have unequal variances even when the variance of the *errors* is constant, it is preferable to plot studentized residuals against fitted values.
- It often helps to plot $|E_i^*|$ or E_i^{*2} against $\widehat{Y}$.
- It is also possible to adapt Tukey's spread-level plot (as long as all of the fitted values are positive), graphing log absolute studentized residuals against log fitted values.

▶ Figure 3 shows a plot of studentized residuals against fitted values and a spread-level plot for the SLID regression.

- The increasing spread with increasing $\widehat{Y}$ suggests moving Y down the ladder of powers to stabilize the variance.
- The slope of the line in the spread-level plot is $b = 0.9994$, suggesting the transformation $p = 1 - 0.9994 = 0.0006 \approx 0$ (i.e., the log transformation).
- After log-transforming Y, the diagnostic plots look much better (Figure 4).

▶ There are alternatives to transformation for dealing with non-constant error variance.

- Weighted-least-squares (WLS) regression, for example, can be used, down-weighting observations that have high variance.
- It is also possible to correct the estimated standard errors of the ordinary least squares (OLS) estimates for non-constant spread.

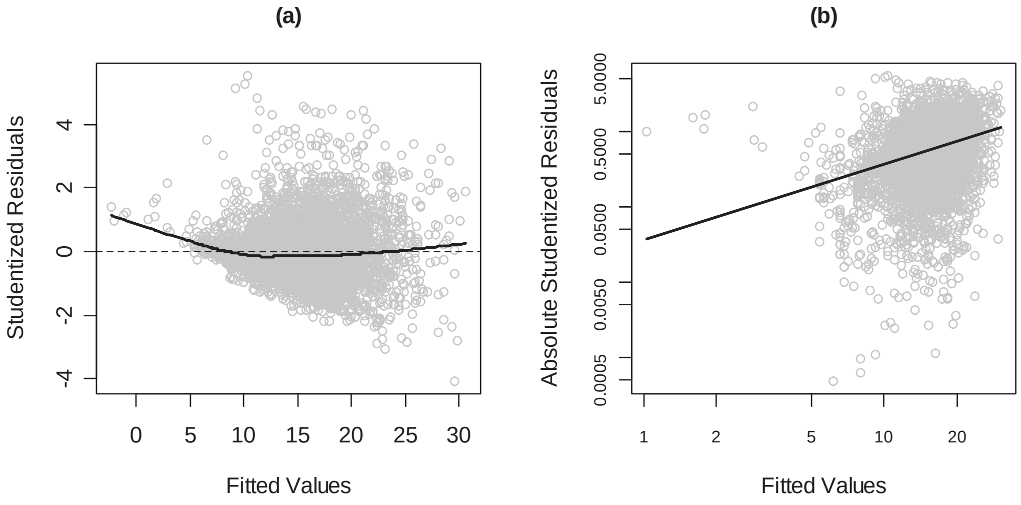

Figure 3. (a) Studentized residuals vs. fitted values, and (b) spread-level plot for the SLID regression. A few observations with $\widehat{Y} \leq 0$ were removed from (b), and the line is fit by robust regression.

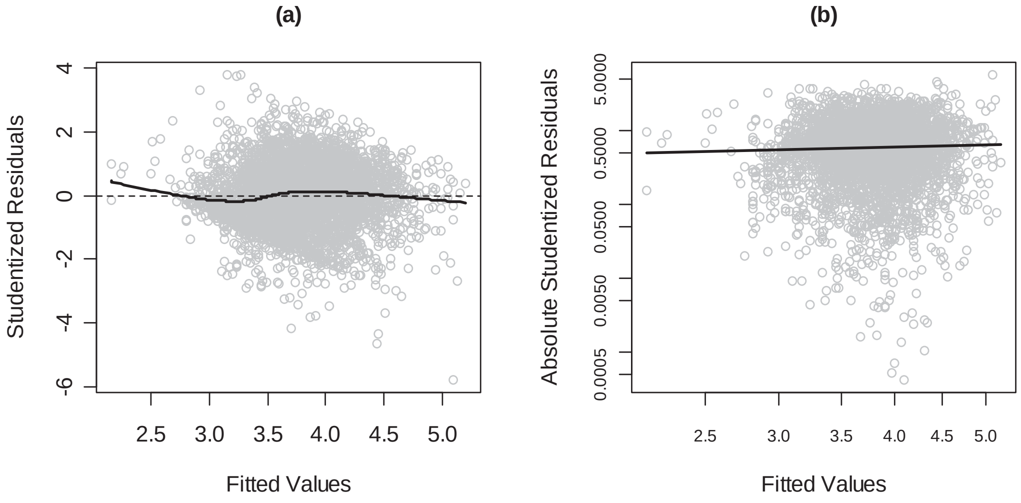

Figure 4. (a) Studentized residuals versus fitted values, and (b) spread--level plot for the SLID regression after log-transforming wages.

- Non-constant error variance is a serious problem only when it is relatively extreme — say when the magnitude (i.e., the standard deviation) of the errors varies by more than a factor of about three — that is, when the largest error variance is more than about 10 times the smallest (although there are cases where this simple rule fails to offer sufficient protection).

6. Nonlinearity

- The assumption that the average error, $E(\varepsilon)$, is everywhere zero implies that the specified regression surface accurately reflects the dependency of Y on the X's.
  - The term ‘nonlinearity’ is therefore not used in the narrow sense here, although it includes the possibility that a partial relationship assumed to be linear is in fact nonlinear.
  - If, for example, two explanatory variables specified to have additive effects instead interact, then the average error is not zero for all combinations of X-values.
  - If nonlinearity, in the broad sense, is slight, then the fitted model can be a useful approximation even though the regression surface $E(Y|X_1, ...X_k)$ is not captured precisely.
  - In other instances, however, the model can be seriously misleading.

- ▶ The regression surface is generally high dimensional, even after accounting for regressors (such as dummy variables, interactions, and polynomial terms) that are functions of a smaller number of fundamental explanatory variables.
  - As in the case of non-constant error variance, it is necessary to focus on particular patterns of departure from linearity.
  - The graphical diagnostics discussed in this section are two-dimensional projections of the $(k+1)$-dimensional point-cloud of observations $\{Y_i, X_{i1}, ..., X_{ik}\}$.

6.1 Component+Residual Plots

- ▶ Although it is useful in multiple regression to plot Y against each X, these plots can be misleading, because our interest centers on the *partial* relationship between Y and each X, controlling for the other X's, not on the *marginal* relationship between Y and an individual X, ignoring the other X's.
- ▶ Plotting residuals or studentized residuals against each X is frequently helpful for detecting departures from linearity.
  - As Figure 5 illustrates, however, residual plots cannot distinguish between monotone and non-monotone nonlinearity.
    - – The distinction is important because monotone nonlinearity frequently can be 'corrected' by simple transformations.
    - – Case (a) might be modeled by $Y = \alpha + \beta\sqrt{X} + \varepsilon$.

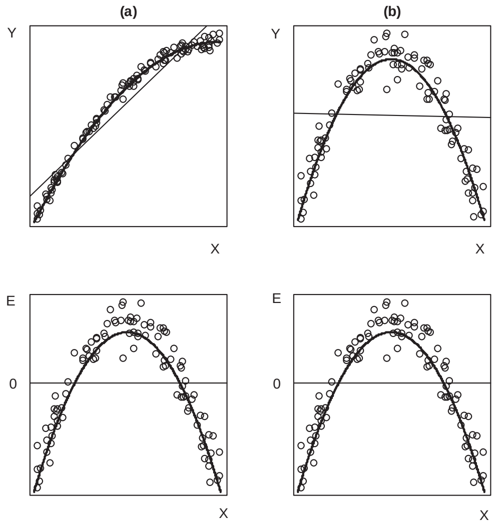

Figure 5. The residual plots of E versus X (bottom) are identical, even though the regression of Y on X in (a) is monotone while that in (b) is non-monotone.

– Case (b) cannot be linearized by a power transformation of X, and might instead be dealt with by the quadratic regression, $Y = \alpha + \beta_1 X + \beta_2 X^2 + \varepsilon$.

▶ Added-variable plots, introduced previously for detecting influential data, can reveal nonlinearity and suggest whether a relationship is monotone.

- These plots are not always useful for locating a transformation, however: The added-variable plot adjusts X_j for the other X's, but it is the unadjusted X_j that is transformed in respecifying the model.

▶ *Component+residual plots*, also called *partial-residual plots* (as opposed to partial-regression = added-variable plots) are often an effective alternative.

- Component+residual plots are not as suitable as added-variable plots for revealing leverage and influence.
- The partial residual for the jth explanatory variable is

$$E_i^{(j)} = E_i + B_j X_{ij}$$

- In words, add back the linear component of the partial relationship between Y and X_j to the least-squares residuals, which may include an unmodeled nonlinear component.
- Then plot $E^{(j)}$ versus X_j.
- By construction, the multiple-regression coefficient B_j is the slope of the simple linear regression of $E^{(j)}$ on X_j, but nonlinearity may be apparent in the plot as well.

▶ The component+residual plots in Figure 6 are for age and education in the SLID regression, using log-wages as the response.

- Both plots look nonlinear:
  - It is not entirely clear whether the partial relationship of log wages to age is monotone, simply tending to level off at the higher ages, or whether it is non-monotone, turning back down at the far right.

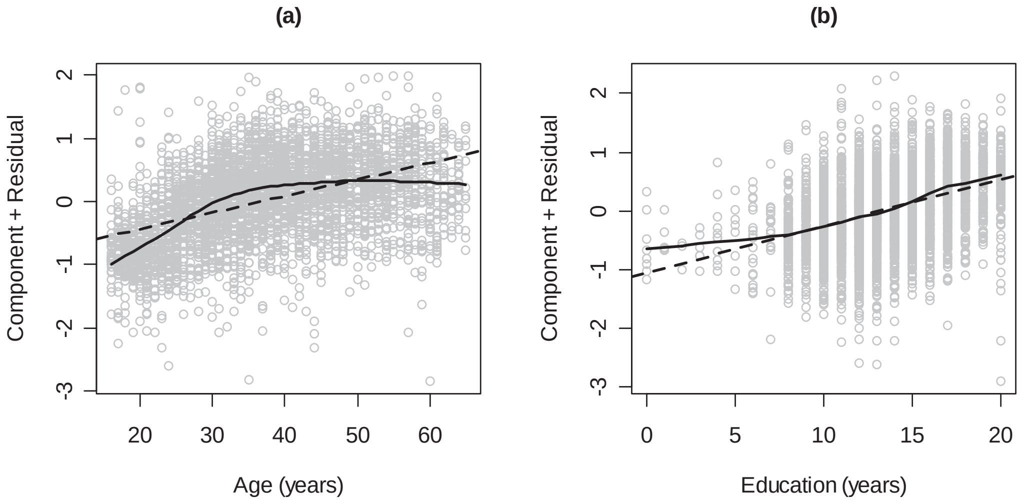

Figure 6. Component-plus-residual plots for age and education in the SLID regression of log wages on these variables and sex. A lowess smooth (span = 0.4) and least-squares line is shown on each graph.

- The partial relationship of log wages to education is clearly monotone, and the departure from linearity is not great—except at the lowest levels of education, where data are sparse; we should be able to linearize this partial relationship by moving education *up* the ladder of powers, because the bulge points to the right.
- Trial and error experimentation suggests that the quadratic specification for age works better, producing the following fit to the data:

$$\begin{aligned}\log_2 \widehat{\text{Wages}} &= \underset{(0.0834)}{0.5725} + \underset{(0.0180)}{0.3195} \times \text{Male} + \underset{(0.0046)}{0.1198} \times \text{Age} \\ &\quad - \underset{(0.000059)}{0.001230} \times \text{Age}^2 + \underset{(0.000113)}{0.002605} \times \text{Education}^2 \\ R^2 &= .3892\end{aligned}$$

- We can take two approaches to constructing component+residual plots for this respecified model:

1. We can plot partial residuals for each of age and education against the corresponding explanatory variable. In the case of age, the partial residuals are computed as
$$E_i^{(\text{Age})} = 0.1198 \times \text{Age}_i - 0.001230 \times \text{Age}_i^2 + E_i$$
and for education,
$$E_i^{(\text{Education})} = 0.002605 \times \text{Education}_i^2 + E_i$$
See the upper panels of Figure 7; the solid lines are the *partial fits* (i.e., the components) for the two explanatory variables,
$$\widehat{Y}_i^{(\text{Age})} = 0.1198 \times \text{Age}_i - 0.001230 \times \text{Age}_i^2$$
$$\widehat{Y}_i^{(\text{Education})} = 0.002605 \times \text{Education}_i^2$$
2. We can plot the partial residuals against the partial fits. See the two lower panels of Figure 7.

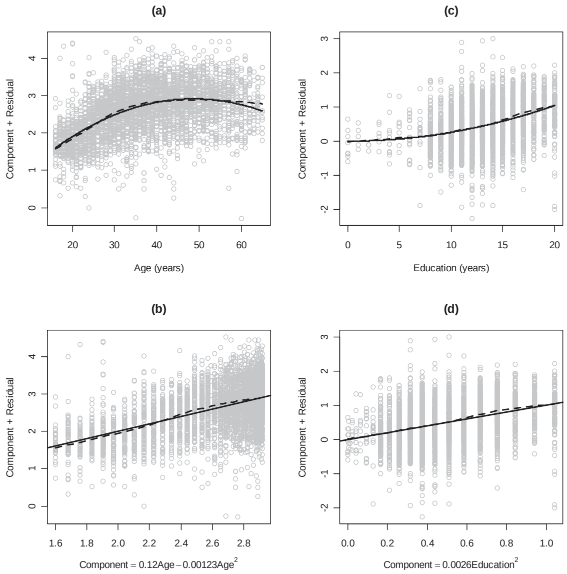

Figure 7. Component-plus-residual plots for age [panels (a) and (b)] and education [panels (c) and (d)] in the respecified model fit to the SLID data.

- Interpretation of the respecified SLID regression model is complicated by the transformation of the response ($\log_2$wages), the transformation of education, and the use of a quadratic for age.
  - The coefficient of the dummy variable for sex, 0.3195, implies that at fixed levels of age and education, men on average earn $2^{0.3195} = 1.25$ times (i.e., 25 percent more) than women.
  - Effect displays for the partial relationship of wages to age and education are shown in Figure 8.

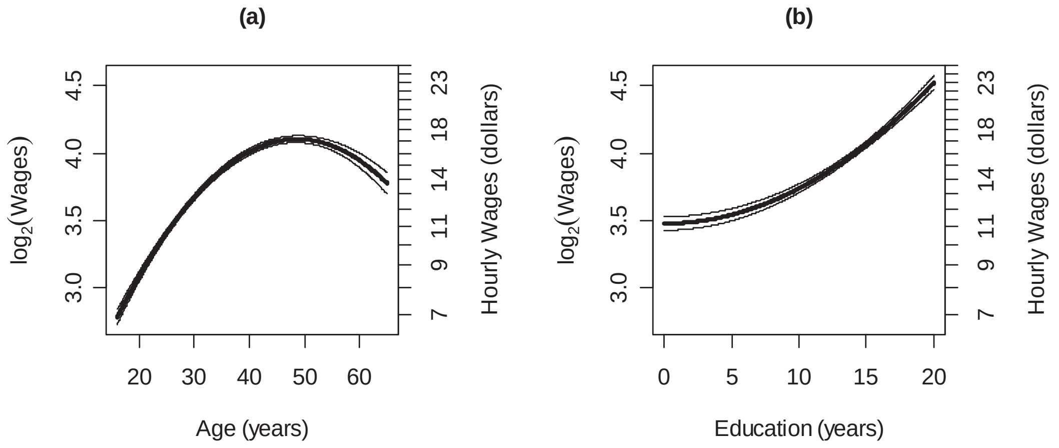

Figure 8. Effect displays for age and education in the regression of log wages on a quadratic in age, the square of education, and sex. The lighter lines give 95-percent point-wise confidence envelopes around the fits.

6.2 Regression Splines

▶ *Regression splines* are an alternative to polynomial regression (such as a quadratic regression) that offer several advantages.

- Polynomial are ‘non-local’ in the sense that data in one region can influence the fit in another region; regression splines, in contrast, provide a local fit to the data.
- Regression splines are similar to nonparametric regression in that they will follow the ‘trend’ in the data, rather than requiring that we specify in advance the pattern of the partial relationship between Y and an X.
- Even though they behave similarly to nonparametric regression, regression splines are fully parametric.

- Regression splines work by dividing the range of each X into a prespecified number of regions, at points called knots, which, for example, can be placed at quantiles such as quartiles or quintiles (to divide the range of an X into four or five regions).
  - Within each region a polynomial regression (typically a cubic regression, with X, X^2, and X^3 terms) is performed, but the various fitted polynomials are constrained to join smoothly at the knots.
  - Because of the constraints, cubic regression splines use only slightly more parameters than there are knots.
- A disadvantage of regression splines — even in comparison to polynomial regression — is that the individual regression coefficients for the regression-spline components are essentially uninterpretable, and thus we have to examine the fit graphically, for example in effect displays.

- I fit a model to the SLID data regressing the $\log_2$ of wages a dummy variable for sex, and ‘natural’ regression splines for age and education, each with four knots, producing the effect displays for age and education in Figure 9.

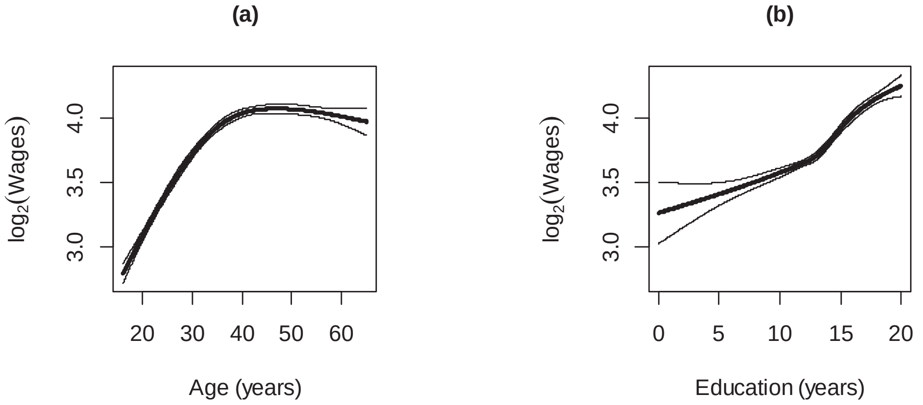

Figure 9. Effect displays for age and education in the regression of log wages on sex and natural regression splines for age and education, each with four knots.

6.3 When Do Component+Residual Plots Work? (time permitting)

- Imagine that the following model accurately describes the data:

$$Y_i = \alpha + f(X_{i1}) + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

  - That is, the partial relationship between Y and X_1 is (potentially) nonlinear, characterized by the function $f(X_1)$, while the other explanatory variables, $X_2, ..., X_k$ enter the model linearly.

- Instead of fitting this model to the data, we fit the 'working model'

$$Y_i = \alpha' + \beta'_1 X_{i1} + \beta'_2 X_{i2} + \cdots + \beta'_k X_{ik} + \varepsilon'_i$$

  and construct a component+residual plot for the working model.

- The partial residuals estimate

$$\varepsilon_i^{(1)} = \beta'_1 X_{i1} + \varepsilon'_i$$

  - What we would really like to estimate, however, is $f(X_{i1}) + \varepsilon_i$, which, apart from random error, will tell us the partial relationship between Y and X_1.

- Cook (1993) shows that $\varepsilon_i^{(1)} = f(X_{i1}) + \varepsilon_i$, as desired, under either of two circumstances:
  - The function $f(X_1)$ is linear.
  - The *other* explanatory variables $X_2, ..., X_k$ are each linearly related to X_1. That is,
  $$E(X_{ij}) = \alpha_{j1} + \beta_{j1}X_{i1} \text{ for } j = 2, ..., k$$
- If there are *nonlinear* relationships between other X's and X_1, then the component+residual plot for X_1 may appear nonlinear even if the true partial regression is linear.
- The second result suggests a practical procedure for improving the chances that component+residual plots will provide accurate evidence of nonlinearity:
  - If possible, transform the explanatory variables to linearize the relationships among them.

- Evidence suggests that weak nonlinearity is not especially problematic, but strong nonlinear relationships among the explanatory variables can invalidate the component+residual plot as a useful diagnostic display.
  - There are more sophisticated versions of component+residual plots that are more robust.

7. Discrete Data (time permitting)

▶ Discrete explanatory and response variables often lead to plots that are difficult to interpret, a problem that can be rectified by 'jittering' the plotted points.

- A discrete *response* variable also violates the assumption that the errors in a linear model are normally distributed.
- Discrete *explanatory* variables, in contrast, are perfectly consistent with the general linear model, which makes no distributional assumptions about the X's, other than independence between the X's and the errors.
- Because it partitions the data into groups, a discrete X (or combination of X's) facilitates straightforward tests of nonlinearity and non-constant error variance.

7.1 Testing for Nonlinearity ('Lack of Fit')

▶ Recall the data on vocabulary and education collected in the U.S. General Social Survey. Years of education in this dataset range between 0 and 20 (see Figure 10). We model the relationship between vocabulary score and education in two ways:

1. Fit a linear regression of vocabulary on education:
$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad \text{(Model 1)}$$
2. Model education with a set of 20 dummy regressors (treating 0 years as the baseline category):
$$Y_i = \alpha' + \gamma_1 D_{i1} + \gamma_2 D_{i2} + \cdots + \gamma_{20} D_{i,20} + \varepsilon'_i \qquad \text{(Model 2)}$$

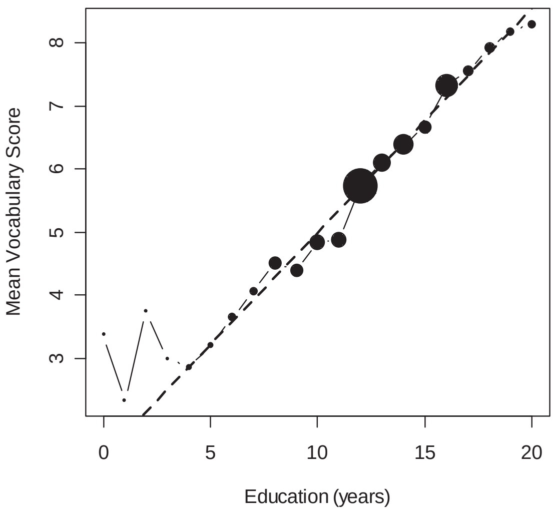

Figure 10. Mean vocabulary score by years of education. The size of the points is proportional to the number of observations. The broken line is the least-squares line.

- Contrasting the two models produces a test for nonlinearity, because the first model, specifying a linear relationship between vocabulary and education, is a special case of the second, which can capture *any* pattern of relationship between $E(Y)$ and X.
  - The resulting incremental F-test for nonlinearity appears in the following ANOVA table:

Source	*SS*	*df*	*F*	*p*
Education (*Model 2*)	26,099	20	374.44	$\ll .0001$
Linear (*Model 1*)	25,340	1	7,270.99	$\ll .0001$
Nonlinear (*“lack of fit”*)	759	19	11.46	$\ll .0001$
Error (*“pure error”*)	75,337	21,617		
Total	101,436	21,637		

- Note that while it is highly statistically significant, the nonlinear component accounts for very little of the variation in vocabulary scores.

▶ The incremental F-test for nonlinearity can easily be extended to a discrete explanatory variable — say X_1 — in a multiple-regression model.

- Here, we need to contrast the general model
$$Y_i = \alpha + \gamma_1 D_{i1} + \cdots + \gamma_{m-1} D_{i,m-1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$
with the model specifying a linear effect of X_1,
$$Y_i = \alpha \ + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$
where $D_1, ..., D_{m-1}$ are dummy regressors constructed to represent the m categories of X_1.

7.2 Testing for Non-Constant Error Variance

▶ A discrete X (or combination of X's) partitions the data into m groups (as in analysis of variance).

- Let Y_{ij} denote the ith of n_j response-variable scores in group j.
- If the error variance is constant across groups, then the within-group sample variances
$$S_j^2 = \frac{\sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_j)^2}{n_j - 1}$$
should be similar.
  - Tests that examine the S_j^2 directly do not maintain their validity well when the distribution of the errors is non-normal.

- The following simple F-test (called Levene's test) is both robust and powerful:
  - Calculate the values
  $$Z_{ij} \equiv |Y_{ij} - \widetilde{Y}_j|$$
  where $\widetilde{Y}_j$ is the median response-variable value in group j.
  - Then perform a one-way analysis-of-variance of the Z_{ij} over the m groups.
  - If the error variance is not constant across the groups, then the group means $\overline{Z}_j$ will tend to differ, producing a large value of the F-test statistic.

- For the vocabulary data, where education partitions the $21,638$ observations into $m = 21$ groups, $F_0 = 4.26$, with 20 and $21,617$ degrees of freedom, for which $p \ll .0001$. There is, therefore, strong evidence of non-constant spread in vocabulary across the categories of education, though, as revealed in Figure 11, the within-group standard deviations are not very different (discounting the small numbers of individuals with very low levels of education).

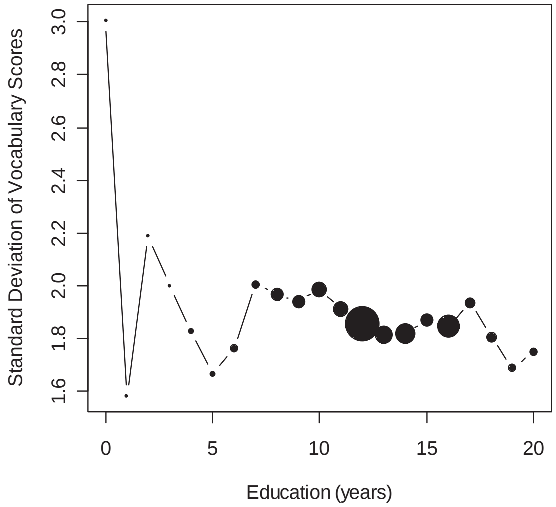

Figure 11. Standard deviation of vocabulary scores by education. The relative size of the points is proportional to the number of observations.

8. Maximum-Likelihood Methods (time permitting)

▶ A statistically sophisticated approach to selecting a transformation of Y or an X is to imbed the linear model in a more general nonlinear model that contains a parameter for the transformation.

- If several variables are potentially to be transformed then there may be several such parameters.

▶ Suppose that the transformation is indexed by a single parameter λ, and that we can write down the likelihood for the model as a function of the transformation parameter and the usual regression parameters: $L(\lambda, \alpha, \beta_1, ..., \beta_k, \sigma_\varepsilon^2)$.

- Maximizing the likelihood yields the maximum-likelihood estimate of λ along with the MLEs of the other parameters.
- Now suppose that $\lambda = \lambda_0$ represents *no* transformation (e.g., $\lambda_0 = 1$ for the power transformation Y^λ).

- A likelihood-ratio test, Wald test, or score test of H_0: $\lambda = \lambda_0$ assesses the evidence that a transformation is required.
- A disadvantage of the likelihood-ratio and Wald tests is that they require finding the MLE, which usually requires iteration.
  - In contrast, the slope of the log-likelihood at λ_0 — on which the score test depends — generally can be assessed or approximated without iteration.
  - Often, the score test can be formulated as the t-statistic for a new regressor, called a *constructed variable*, to be added to the linear model.
  - Moreover, an added-variable plot for the constructed variable then can reveal whether one or a small group of observations is unduly influential in determining the transformation.

8.1 Box-Cox Transformation of Y

▶ Box and Cox suggest power transformation of Y with the object of normalizing the error distribution.

▶ The general Box-Cox model is

$$Y_i^{(\lambda)} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

where the errors ε_i are independently $N(0, \sigma_\varepsilon^2)$, and

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \log_e Y_i & \text{for } \lambda = 0 \end{cases}$$

- Note that all of the Y_i must be positive.

- A simple procedure for finding the MLE is to evaluate the maximized $\log_e L(\alpha, \beta_1, ..., \beta_k, \sigma_\varepsilon^2|\lambda)$, called the *profile log-likelihood*, for a range of values of λ, say between -2 and $+2$.
  - If this range turns out not to contain the maximum of the log-likelihood, then the range can be expanded.
  - To test H_0: $\lambda = 1$, calculate the likelihood-ratio statistic
  $$G_0^2 = -2[\log_e L(\lambda = 1) - \log_e L(\lambda = \widehat{\lambda})]$$
  which is asymptotically distributed as χ^2 with one degree of freedom under H_0.
  - Equivalently, a 95-percent confidence interval for λ includes those values for which
  $$\log_e L(\lambda) > \log_e L(\lambda = \widehat{\lambda}) - 1.92$$
    - The figure 1.92 comes from $1/2 \times \chi^2_{1,.05} = 1/2 \times 1.96^2$.
- Figure 12 shows a plot of the profile log-likelihood against λ for the original SLID regression.

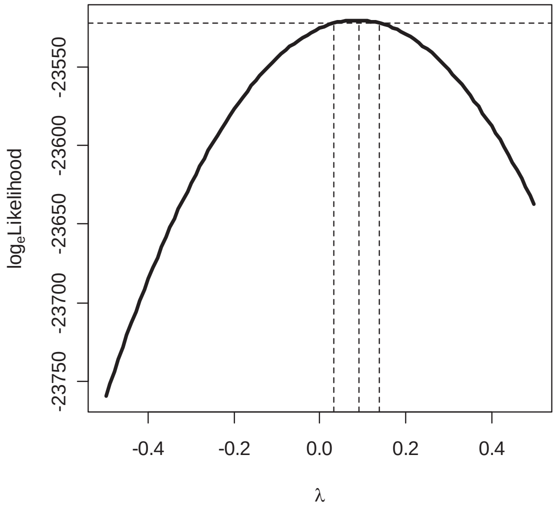

Figure 12. Box-Cox transformations for the SLID regression of wages on sex, age, and education. The profile log-likelihood is plotted against the transformation parameter λ.

- The maximum-likelihood estimate of λ is $\widehat{\lambda} = 0.09$, and a 95% confidence interval, marked out by the intersection of the line near the top of the graph with the profile log-likelihood, runs from 0.04 to 0.13.

▶ Atkinson has proposed an approximate score test for the Box-Cox model, based on the constructed variable

$$G_i = Y_i \left[\log_e\left(\frac{Y_i}{\widetilde{Y}}\right) - 1\right]$$

where $\widetilde{Y}$ is the *geometric mean* of Y:

$$\widetilde{Y} \equiv (Y_1 \times Y_2 \times \cdots \times Y_n)^{\frac{1}{n}}$$

- The augmented regression, including the constructed variable, is then

$$Y_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \phi G_i + \varepsilon_i$$

- The t-test of H_0: $\phi = 0$, that is, $t_0 = \widehat{\phi}/\text{SE}(\widehat{\phi})$, assesses the need for a transformation.

- An estimate of λ (though not the MLE) is given by $\widetilde{\lambda} = 1 - \widehat{\phi}$.
- The added-variable plot for the constructed variable G shows influence and leverage on $\widehat{\phi}$, and hence on the choice of λ.
- Atkinson's constructed-variable plot for the interlocking-directorate regression is shown in Figure 13.
  - The coefficient of the constructed variable in the regression is $\widehat{\phi} = 1.454$, with SE$(\widehat{\phi}) = 0.026$, providing overwhelmingly strong evidence of the need to transform Y.
  - The suggested transformation, $\widetilde{\lambda} = 1 - 1.454 = -0.454$, is far from the MLE.

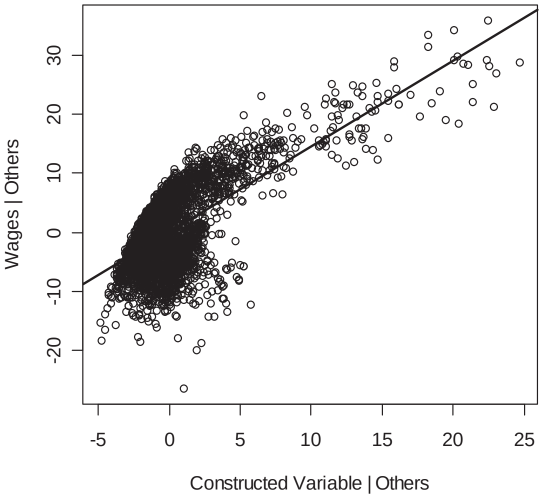

Figure 13. Constructed-variable plot for the Box-Cox transformation of wages in the SLID regression. The least-squares line is shown on the plot.

8.2 Box-Tidwell Transformation of the X's

▶ Now, consider the model

$$Y_i = \alpha + \beta_1 X_{i1}^{\gamma_1} + \cdots + \beta_k X_{ik}^{\gamma_k} + \varepsilon_i$$

where the errors are independently distributed as $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and all of the X_{ij} are positive.

▶ The parameters of this model — $\alpha, \beta_1, ..., \beta_k, \gamma_1, ..., \gamma_k$, and σ_ε^2 — could be estimated by general nonlinear least squares, but Box and Tidwell suggest instead a computationally more efficient procedure that also yields a constructed-variable diagnostic:

1. Regress Y on $X_1, ..., X_k$, obtaining $A, B_1, ..., B_k$.
2. Regress Y on $X_1, ..., X_k$ *and* the constructed variables $X_1 \log_e X_1, ..., X_k \log_e X_k$, obtaining $A', B_1', ..., B_k'$ and $D_1, ..., D_k$.

3. The constructed variable $X_j \log_e X_j$ can be used to assess the need for a transformation of X_j by testing the null hypothesis H_0: $\delta_j = 0$, where δ_j is the population coefficient of $X_j \log_e X_j$ in step 2. Added-variable plots for the constructed variables are useful for assessing leverage and influence on the decision to transform the X's.
4. A preliminary estimate of the transformation parameter γ_j (not the MLE) is given by

$$\widetilde{\gamma}_j = 1 + \frac{D_j}{B_j}$$

- ▶ This procedure can be iterated through steps 1, 2, and 4 until the estimates of the transformation parameters stabilize, yielding the MLEs $\widehat{\gamma}_j$.

- ▶ Consider the SLID regression of log wages on sex, education, and age.
  - The dummy regressor for sex is not a candidate for transformation, of course, but I will consider power transformations of age and education.
    - Recall that we were initially undecided about whether to model the age effect as a quadratic or as a transformation down the ladder of powers and roots.
  - To make power transformations of age more effective, I use a negative start of 15 (recall that age ranges from 16 to 65).
  - The coefficients of $(\text{Age} - 15) \times \log_e(\text{Age} - 15)$ and Education$\times$ $\log_e$Education in the step-2 augmented model are, respectively, $D_{\text{Age}} = -0.04699$ with SE$(D_{\text{Age}}) = 0.00231$, and $D_{\text{Education}} = 0.05612$ with SE$(D_{\text{Education}}) = 0.01254$.
  - Both score tests are statistically significant, but there is much stronger evidence of the need to transform age.

- The first-step estimates of the transformation parameters are

$$\widetilde{\gamma}_{\text{Age}} = 1 + \frac{D_{\text{Age}}}{B_{\text{Age}}} = 1 + \frac{-0.04699}{0.02619} = -0.79$$

$$\widetilde{\gamma}_{\text{Education}} = 1 + \frac{D_{\text{Education}}}{B_{\text{Education}}} = 1 + \frac{0.05612}{0.08061} = 1.69$$

- The fully iterated MLEs of the transformation parameters are $\widehat{\gamma}_{\text{Age}} = 0.051$ and $\widehat{\gamma}_{\text{Education}} = 1.89$ — very close to the log transformation of started-age and the square of education.
- Constructed-variable plots for the transformation of age and education are shown in Figure 14.

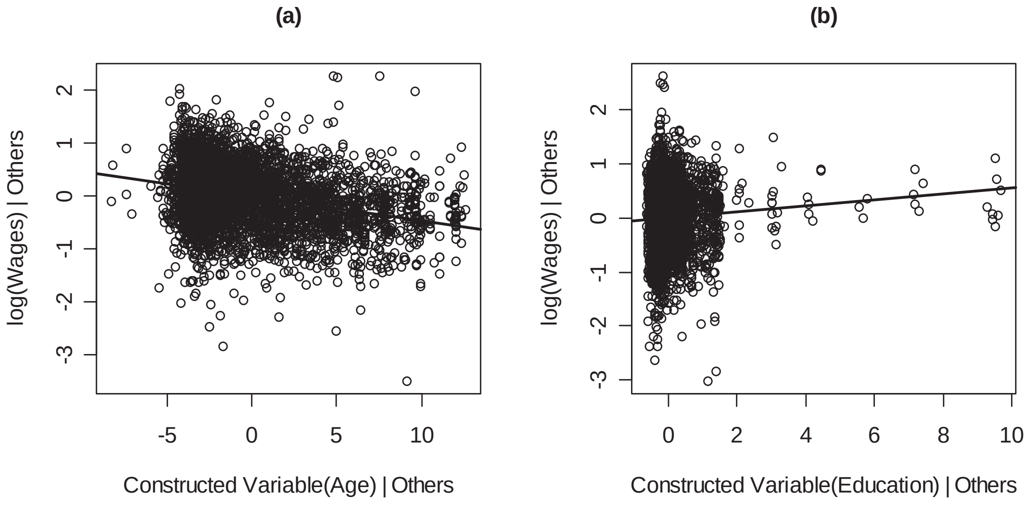

Figure 14. Constructed-variable plots for the Box-Tidwell transformation of (a) age and (b) education in the SLID regression of log wages on sex, age, and education.

8.3 Non-Constant Error Variance Revisited

▶ Breusch and Pagan develop a score test for heteroscedasticity based on the specification:

$$\sigma_i^2 \equiv V(\varepsilon_i) = g(\gamma_0 + \gamma_1 Z_{i1} + \cdots + \gamma_p Z_{ip})$$

where $Z_1, ..., Z_p$ are known variables, and where the function $g(\cdot)$ is quite general.

- The same test was independently derived by Cook and Weisberg.

▶ The score statistic for the hypothesis that the σ_i^2 are all the same, which is equivalent to H_0: $\gamma_1 = \cdots = \gamma_p = 0$, can be formulated as an auxiliary-regression problem.

- Let $U_i \equiv E_i^2 / \widehat{\sigma}_\varepsilon^2$, where $\widehat{\sigma}_\varepsilon^2 = \sum E_i^2 / n$ is the MLE of the error variance. Regress U on the Z's:

$$U_i = \eta_0 + \eta_1 Z_{i1} + \cdots + \eta_p Z_{ip} + \omega_i$$

- Breusch and Pagan show that the score statistic

$$S_0^2 = \frac{\sum(\widehat{U}_i - \overline{U})^2}{2}$$

is asymptotically distributed as χ^2 with p degrees of freedom under the null hypothesis of constant error variance.

- Here, the $\widehat{U}_i$ are fitted values from the regression of U on the Z's, and thus S_0^2 is half the regression sum of squares from the auxiliary regression.

▶ To apply this result, it is necessary to select Z's, the choice of which depends upon the suspected pattern of non-constant error variance.

- Employing $X_1, ..., X_k$ in the auxiliary regression, for example, permits detection of a tendency of the error variance to increase (or decrease) with the values of one or more of the explanatory variables in the main regression.

- Cook and Weisberg suggest regressing U on the fitted values from the main regression (i.e., $U_i = \eta_0 + \eta_1 \widehat{Y}_i + \omega_i$), producing a one-degree-of-freedom score test to detect the common tendency of the error variance to increase with the level of the response variable.
  - Anscombe suggests correcting detected heteroscedasticity by transforming Y to $Y^{(\widetilde{\lambda})}$ with $\widetilde{\lambda} = 1 - 1/2\widehat{\eta}_1\overline{Y}$.

▶ Applied to the initial SLID regression of wages on sex, age, and education, an auxiliary regression of U on $\widehat{Y}$ yields $\widehat{U} = -0.3449 + 0.08652\widehat{Y}$, and $S_0^2 = 567.66/2 = 283.83$ on 1 degree of freedom, for which $p \approx 0$.

- The suggested variance-stabilizing transformation using Anscombe's rule is

$$\widetilde{\lambda} = 1 - \frac{1}{2}(0.08652)(15.545) = 0.33$$

- An auxiliary regression of U on the explanatory variables in the main regression yields $S_0^2 = 579.08/2 = 289.54$ on $k = 3$ degrees of freedom.
  - The score statistic for the more general test is not much larger than that for the regression of U on $\widehat{Y}$, implying that the pattern of non-constant error variance is indeed for the spread of the errors to increase with the level of Y.

9. Summary

- ▶ Heavy-tailed errors threaten the efficiency of least-squares estimation; skewed and multimodal errors compromise the interpretation of the least-squares fit.
  - Non-normality can often be detected by examining the distribution of the least-squares residuals, and frequently can be corrected by transforming the data.
- ▶ It is common for the variance of the errors to increase with the level of the response variable.
  - This pattern of non-constant error variance can often be detected in a plot of residuals against fitted values.

  - Strategies for dealing with non-constant error variance include transformation of the response variable to stabilize the variance; the substitution of weighted-least-squares estimation for ordinary least squares; and the correction of coefficient standard errors for heteroscedasticity.
  - A rough rule of thumb is that non-constant error variance seriously degrades the least-squares estimator only when the ratio of the largest to smallest variance is about 10 or more.
- ▶ Simple forms of nonlinearity can often be detected in component+residual plots.
  - Once detected, nonlinearity can frequently be accommodated by variable transformations or by altering the form of the model (to include a quadratic term in an explanatory variable, for example).
  - Component+residual plots adequately reflect nonlinearity when the explanatory variables are themselves not strongly nonlinearly related.

- Discrete explanatory variables divide the data into groups.
  - A simple incremental F-test for nonlinearity compares the sum of squares accounted for by the linear regression of Y on X with the sum of squares accounted for by differences in the group means.
  - Likewise, tests of non-constant variance can be based upon comparisons of spread in the different groups.
- A statistically sophisticated general approach to selecting a transformation of Y or an X is to imbed the linear-regression model in a more general model that contains a parameter for the transformation.
  - The Box-Cox procedure selects a power transformation of Y to normalize the errors.
  - The Box-Tidwell procedure selects power transformations of the X's to linearize the regression of Y on the X's.
  - In both cases, ‘constructed-variable’ plots help us to decide whether individual observations are unduly influential in determining the transformation parameters.

- Simple score tests are available to determine the need for a transformation and to test for non-constant error variance.