Applied Statistics With R
Regression Diagnostics

John Fox

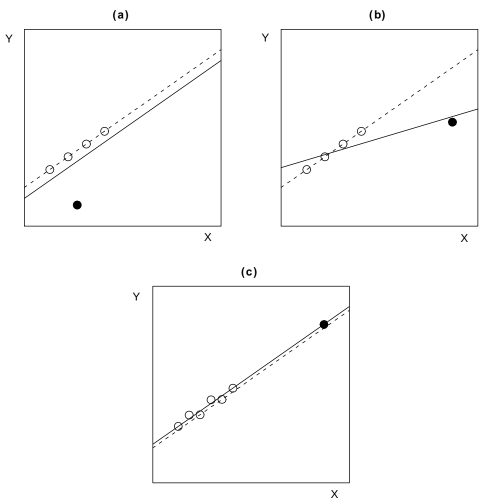
WU Wien May/June 2006

Outline

- Unusual Data
- Non-Normal Errors
- Non-Constant Error Variance
- Nonlinearity
- Collinearity

Unusual Data

Leverage, Outlyingness, and Influence



- (a) Outlier not at a high leverage point and hence not influential.
- (b) Outlier at a high-leverage point and hence influential.
- (c) In-line at a high leverage point and hence not influential.
- Influence on coefficients $=$ Leverage$\times$Outlyingness

Unusual Data

Leverage: Hat-Matrix

- Recall the linear model, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the fitted model, $\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, and the least-squares estimates, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- The least-squares fitted values are therefore a linear function of the observed response:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

- $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the *hat-matrix*, so named because it transforms $\mathbf{y}$ into $\hat{\mathbf{y}}$.
  - The hat matrix is symmetric ($\mathbf{H} = \mathbf{H}'$) and idempotent ($\mathbf{H}^2 = \mathbf{H}$)

Unusual Data

Leverage: Hat-Values

- The diagonal entries of the hat-matrix $h_i \equiv h_{ii}$, called the *hat-values*, are

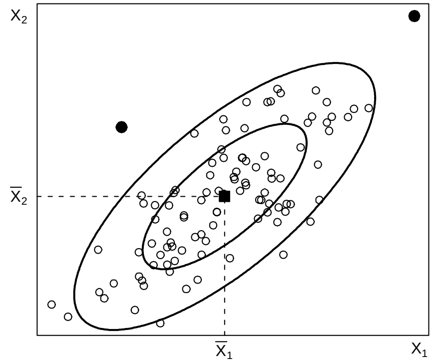
$$h_i = \mathbf{h}_i'\mathbf{h}_i = \sum_{j=1}^{n} h_{ij}^2 = h_i^2 + \sum_{j \neq i} h_{ij}^2$$

  where (because of symmetry) the elements of $\mathbf{h}_i$ comprise both the ith row and the ith column of $\mathbf{H}$.
- This result implies that $0 \leq h_i \leq 1$. If the model matrix $\mathbf{X}$ includes the constant regressor, then $1/n \leq h_i$.
- Because $\mathbf{H}$ is a projection matrix, projecting $\mathbf{y}$ orthogonally onto the $(k+1)$-dimensional subspace spanned by the columns of $\mathbf{X}$, $\sum h_i = k+1$, and thus $\overline{h} = (k+1)/n$.
  - Rough rule-of-thumb: Hat-values exceeding $2\overline{h}$ or $3\overline{h}$ are considered noteworthy.

Unusual Data

Leverage: Hat-Values

- *Interpretation*: Observations with large hat-values are multivariate outliers in the X-space.
  - Contours of constant leverage with two X's:



Unusual Data

Regression Outliers: Studentized Residuals

- The least-squares residuals $\mathbf{e} = \{E_i\}$ do not have equal variances even when the errors $\boldsymbol{\epsilon} = \{\epsilon_i\}$ do:

$$V(E_i) = \sigma_\epsilon^2(1 - h_i)$$

  - The *standardized residuals*

$$E_i' = \frac{E_i}{S_E\sqrt{1-h_i}}$$

    are not t-distributed, however.
- The *studentized residuals* follow t-distributions with $n-k-2$ df when the model holds:

$$E_i^* = \frac{E_i}{S_{E(-i)}\sqrt{1-h_i}}$$

  where $S_{E(-i)}$ is the residual standard error computed deleting the ith observation from the regression.

Unusual Data

Regression Outliers: Studentized Residuals

- *Bonferroni outlier test:*
  - Let $E_{\max}^*$ represent the largest of the $|E_i^*|$.
  - Let $p' = \Pr(\ t_{n-k-2} > E_{\max}^*)$.
  - The two-sided Bonferroni p-value for the largest absolute studentized residuals is then $p = 2np'$.

Unusual Data

Influential Observations: DFBETA and DFBETAS

- The impact on the regression coefficients of omitting observation i:

$$\begin{aligned}\mathbf{DFBETA}_i &= -\mathbf{b}_{(-i)} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \frac{E_i}{1-h_i}\end{aligned}$$

- Standardizing each entry of $\mathbf{DFBETA}_i$ by a deleted estimate of the coefficient standard error produces

$$\mathbf{DFBETAS}_{ij} = \frac{\mathbf{DFBETA}_{ij}}{\mathrm{SE}_{(-i)}(B_j)}$$

Unusual Data

Influential Observations: Cook's Distances

- *Cook's distances* summarize the impact on all regression coefficients of deleting obervation i:Cook's D_i is the F-statistic for testing the "hypothesis" that $\boldsymbol{\beta} = \mathbf{b}_{(-i)}$:

$$\begin{aligned}D_i &= \frac{(\mathbf{b}-\mathbf{b}_{(-i)})'\mathbf{X}'\mathbf{X}(\mathbf{b}-\mathbf{b}_{(-i)})}{(k+1)S_E^2} \\ &= \frac{(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(-i)})'(\hat{\mathbf{y}}-\hat{\mathbf{y}}_{(-i)})}{(k+1)S_E^2}\end{aligned}$$

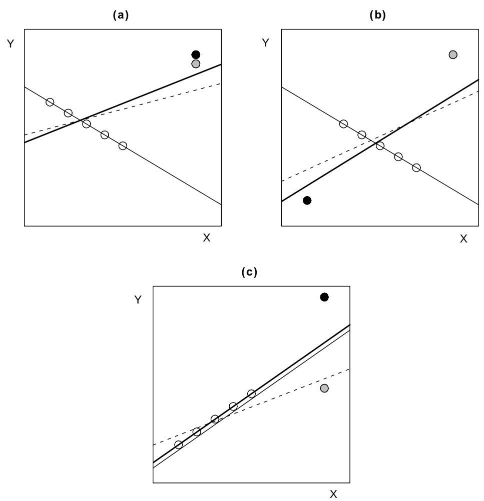
- Cook's D can also be written as

$$\begin{aligned}D_i &= \frac{E_i^2}{S_E^2(k+1)} \times \frac{h_i}{(1-h_i)^2} \\ &= \frac{E_i'^2}{k+1} \times \frac{h_i}{1-h_i}\end{aligned}$$

i.e., outlyingness $\times$ leverage.

Unusual Data

Jointly Influential Data



- *Jointly influential observations* can mask each other's presence, as in (a).
- This can happen even if the points are widely separated, as in (b).
- Points can also offset each other's influence, as in (c).

Unusual Data

Jointly Influential Data: Added-Variable Plots

- *Added-variable plots* (also called *partial-regression plots*) can often detect jointly influential points.
  - Added-variable plots show leverage and influence on individual regression coefficients.
- To draw the added-variable plot for X_1:
  1. Regress Y on all of the X's except X_1:
  $$Y_i = A^{(1)} + B_2^{(1)}X_{i2} + \cdots + B_k^{(1)}X_{ik} + Y_i^{(1)}$$
  2. Regress X on all of the other X's:
  $$X_{i1} = C^{(1)} + D_2^{(1)}X_{i2} + \cdots + D_k^{(1)}X_{ik} + X_i^{(1)}$$
  3. Plot the residuals $Y_i^{(1)}$ against the residuals $X_i^{(1)}$ to form the added-variable plot
- This procedure is repeated for each regressor, including if desired the constant regressor $\mathbf{x}_0 = \{1\}$.

Unusual Data

Jointly Influential Data: Added-Variable Plots

- The added-variable plot has the following properties:

1. The slope from the least-squares regression of $Y^{(1)}$ on $X^{(1)}$ is the slope B_1 from the full multiple regression.
2. The residuals from the simple regression of $Y^{(1)}$ on $X^{(1)}$ are the same as those from the full regression; that is,
$$Y_i^{(1)} = B_1 X_i^{(1)} + E_i$$
3. The variation of $X^{(1)}$ is the *conditional variation* of X_1 holding the other X's constant.
    - Thus, the standard error of B_1 in the auxiliary simple regression
    $$\text{SE}(B_1) = \frac{S_E}{\sqrt{\sum X_i^{(1)2}}}$$
    is the same as the multiple-regression standard error of B_1.
    - Unless X_1 is uncorrelated with the other X's, its conditional variation is smaller than its *marginal variation* $\sum(X_{i1} - \overline{X}_1)^2$.

Non-Normal Errors

Why Worry?

- The central-limit theorem suggests that the *validity* of least-squares inference is robust with respect to departures from normality, so why worry about non-normal errors?
    - The *efficiency* of least-squares estimation is not robust when the error distribution is heavy-tailed.
    - Least-squares estimates a conditional mean, which is not a reasonable summary of the conditional centre of the distribution of Y when the error distribution is skewed.
    - A multi-modal error distributions suggests the omission of a factor dividing the data into groups.

Non-Normal Errors

Quantile-Comparison Plot of Residuals

- To diagnose non-normal errors we can plot the ordered studentized residuals against the corresponding quantiles of $N(0, 1)$ or t_{n-k-2}.
- Postively skewed residuals can be "corrected" by moving Y down the *ladder of powers and roots*—e.g., (for positive Y) to $\sqrt{Y}$, $\log(Y)$, or Y^{-1}.
    - log is treated as the "0th" power.
- Negatively skewed residuals (less common) can be "corrected" by moving Y up the ladder of powers and roots—e.g., to X^2 or X^3.
- Heavy-tailed residuals can be dealt with by *robust estimation*.

Non-Normal Errors

Parametric-Bootstrap Confidence Envelope

- The studentized residuals are not independent and have a complex joint distribution.

1. Fit the regression model obtaining fitted values $\widehat{Y}_i$ and the estimated standard error S_E.
2. Construct m samples, each consisting of n simulated Y-values; for the jth such sample, $Y_{ij}^s = \widehat{Y}_i + S_E Z_{ij}$, where Z_{ij} is a random draw from the unit-normal distribution.
3. Regress the Y_{ij}^s on the X's in the original sample, obtaining simulated studentized residuals, $E_{1j}^*, E_{2j}^*, \ldots, E_{nj}^*$.
4. Order the studentized residuals for sample j from smallest to largest, $E_{(1)j}^*, E_{(2)j}^*, \ldots, E_{(n)j}^*$.
5. To construct an estimated $(100 - a)\%$ confidence interval for $E_{(i)}^*$, find the $a/2$ and $1 - a/2$ empirical quantiles of the m simulated values $E_{(i)1}^*, E_{(i)2}^*, \ldots, E_{(i)m}^*$.

Non-Constant Error Variance

Why Worry?

- One of the assumptions of the regression model is that the variation of the response around the regression surface—the error variance—is everywhere the same:

$$V(\epsilon) = V(Y|x_1, \ldots, x_k) = \sigma_\epsilon^2$$

- Non-constant error variance is often termed *heteroscedasticity*; constant error variance is termed *homoscedasticity*.
- The least-squares estimator is unbiased and consistent even when the error variance is not constant, but:
  - The *efficiency* of the least-squares estimator is impaired.
  - The usual formulas for coefficient standard errors are inaccurate.
  - Seriousness depends on the degree to which error variances differ, the sample size, and the configuration of X-values.

Non-Constant Error Variance

Dealing With Non-Constant Error Variance

- When the error variance increases systematically with the level of Y, as is often the case, it can often be stabilized by power transformation down the ladder of powers and roots.
  - This pattern can be detected in a plot of residuals (e.g., studentized residuals, E_i^*) against fitted values, $\widehat{Y}_i$.
  - The common heteroscedastic pattern is for the residuals to "fan out" as the fitted values increase.
- If the error variance is known up to a constant of proportionality, then *weighted-least-squares* (*WLS*) estimation can be used in place of *ordinary least-squares* (*OLS*).

Non-Constant Error Variance

Dealing With Non-Constant Error Variance

- If there is an unknown pattern of estimation then the usual coefficient standard errors can be replaced by so-called *White standard errors*—also called *heteroscedasticity-consistant standard errors* or *sandwich estimates*.
- Because the data are in general high-dimensional, it is not possible to check graphically for completely general patterns of non-constant error variance.

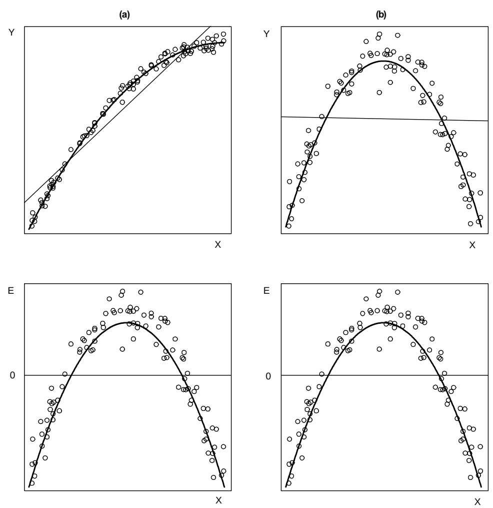
Nonlinearity

What Is It?

- The assumption of linearity in the broad sense is that the average error, $E(\epsilon)$, is everywhere 0
  - This implies that the specified regression surface accurately reflects the dependency of the conditional average value of Y on the X's.
  - Violating the assumption of linearity implies that the model fails to capture the systematic pattern of relationship between the response and explanatory variables.
  - Because the data are high dimensional, it is not generally possible to check graphically for nonlinearity in the broad sense.
- Nonlinearity in the narrow sense is the assumption that the partial relationship between Y and a particular X_j is captured by the term $\beta_j X_j$.

Nonlinearity

Inadequacy of Plotting Residuals Against Each X



- *Monotone nonlinearity*, as at the left, can often be corrected by a power transformation of X (or Y or both): e.g., $\widehat{Y} = A + B\log(X)$.
- *Non-monotone nonlinearity*, as at the right, requires another approach: e.g., $\widehat{Y} = A + B_1X + B_2X^2$.
- The residual plots (at the bottom) do not distinguish the two cases.

Nonlinearity

Component+Residual Plots

- *Component+residual plots* can be used to detect nonlinearity in the narrow sense.
  - These plots are also called *partial-residual plots* (not to be confused with partial-*regression*, i.e., added-variable, plots).
- The *partial residual* for the jth explanatory variable is

$$E_i^{(j)} = E_i + B_jX_{ij}$$

- 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.
  - Nonlinearity may be apparent in the plot as well.
- One such plot is constructed for each (quantitative) X.
- Component+residual plots can be generalized to more complex fits, such as polynomial-regression models, and to models with interactions.

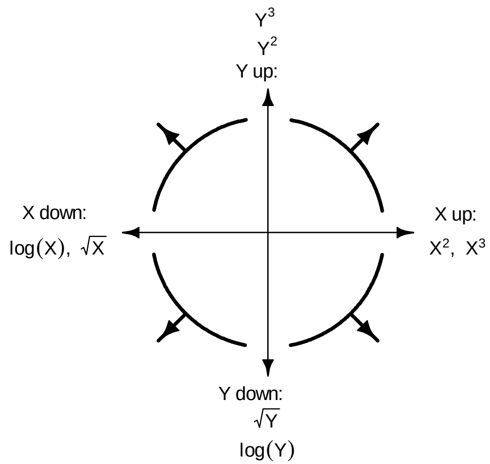
Nonlinearity

What To Do?

- Simple monotone nonlinearity: Transform X (or possibly Y).
- Other strategies:
  - Polynomial regression—quadratic, cubic, etc. (but high-degree polynomials are usually a bad idea).
  - Regression splines.
  - Binning (categorizing) X.
  - Nonparametric regression.

Nonlinearity

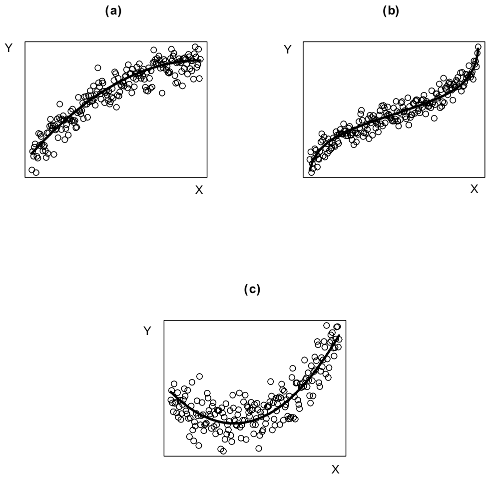
Mosteller and Tukey's "Bulging Rule"



- Follow the direction of the "bulge" to decide whether to move up or down the latter of powers and roots for X (and/or Y).
- In multiple regression, unless there is a common pattern to all of the partial relationships, we generally prefer to transform an X.

Nonlinearity

Simple Monotone Nonlinearity



- The bulging rule works for *simple monotone nonlinearity*, as in (a).
- (b) Monotone but not simple.
- (c) Simple but non-monotone.

Collinearity

Nature of the Problem

- When the explanatory variables in a regression are very highly correlated, the regression coefficients are imprecisely estimated.
- The sampling variance of B_j is

$$V(B_j) = \frac{1}{1 - R_j^2} \times \frac{\sigma_\epsilon^2}{(n-1)S_j^2}$$

  where
  - R_j^2 is the squared multiple correlation for the regression of X_j on the other X's;
  - σ_ϵ^2 is the error variance;
  - n is the sample size;
  - $S_j^2 = \sum(X_{ij} - \overline{X}_j)^2/(n-1)$ is the variance of X_j.
- The formula reveals the sources of imprecision in regression: collinearity but also weak relationships, small samples, and homegenous X's.

Collinearity

Variance-Inflation Factors

- The term $1/(1 - R_j^2)$ is called the *variance-inflation factor* (VIF_j).
- The square-root of the VIF expresses the impact of collinearity on the coefficient standard error and hence on the width of the confidence interval for β_j.
- R_j has to get very large before the precision of estimation is seriously degraded; e.g., for $R_j = .8$,

$$\sqrt{\text{VIF}} = \sqrt{\frac{1}{1 - .8^2}} = 1.67$$

- Variance-inflation factors can be extended to sets of related regressors (e.g., sets of dummy regressors or polynomial regressors) by considering the size of the confidence region for the coefficients.