

CHAPTER 3: DIAGNOSTICS AND REMEDIAL MEASURES

3.1 diagnostics for predictor variable: plots of X to detect outlying X values that could influence the appropriateness of fitted regression function

3.2 residuals $e_i = Y_i - \hat{Y}_i$ ('observed errors')

- not independent, but for large n the dependency is small
- recall $\sum X_i e_i = 0$
- average of residuals $\bar{e} = \sum e_i / n = 0$
- $MSE = \sum e_i^2 / (n - 2)$
- semistudentized residuals $e_i^* = \frac{e_i}{\sqrt{MSE}}$

departures from model to be studied by residuals

- nonlinearity
- nonconstant error variance
- dependent error terms
- outliers that throw off fit
- nonnormality of error terms
- omission of important predictors

3.3 diagnostics for residuals

departures

nonlinearity

nonconstancy of error variance

outliers

nonindependence of error terms

nonnormality of errors
(note: check other departures first.)

omission of important predictor variables

residual plots

e_i vs. X or e_i vs. \hat{Y}_i

e_i vs. X or e_i vs. \hat{Y}_i

$|e_i|$ or e_i^2 vs. X

e_i vs. X or e_i vs. \hat{Y}_i

boxplot, stemplot, dotplot e_i or e_i^*

e_i vs. time

histogram or boxplot of e_i

normal probability plot

$e_{(i)}$ vs est. $E(e_{(i)} | H_0)$

H_0 : normal errors

est. $E(e_{(i)} | H_0) = \sqrt{MSE} z_{(\frac{k-.375}{n+.25})}$

e_i vs. new X (see p109)

- Note prototype plots in text

3.4 overview of test involving residuals

- runs test for randomness
- Durbin-Watson test (later in text)
- rank correlation of $|e_i|$ and X_i
- Levene test
- Breusch-Pagan test
- test involving fitting line without suspect outlier (later in text)
- normality tests: Lilliefors
goodness-of-fit
correlation test using Table B.*
(e_i vs nscores)

3.5 correlation test for normality

In Minitab, suppose residuals are in c10.

```
MTB> nscores c10 put in c11
```

```
MTB> name c10 'residual' c11 'resnscore'
```

```
MTB> plot c10 c11
```

```
MTB> corr c10 c11
```

- compare correlation coefficient to value in Table B.*

- if corr. coeff. is too small, reject H_0 of normality

3.6 Brown-Forsythe (modified Levene) test (for constancy of error variance)

- 1) divide e_i into two groups based on small X_i and large X_i
to obtain e_{i1} and e_{i2} groups of size n_1 and n_2 , respectively
- 2) calculate medians of each group: \tilde{e}_1 and \tilde{e}_2
- 3) calculate absolute deviations for each group:
 $d_{i1} = |e_{i1} - \tilde{e}_1|$ and $d_{i2} = |e_{i2} - \tilde{e}_2|$
- 4) perform a two-sample t test using

$$t_L^* = \frac{\bar{d}_1 - \bar{d}_2}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where $s^2 = \frac{\sum (d_{i1} - \bar{d}_1)^2 + \sum (d_{i2} - \bar{d}_2)^2}{n - 2}$

note: under H_0 (equal variances), $t_L^* \sim t(n-2)$

- 5) reject H_0 if $|t_L^*| > t_{1-\alpha/2, n-2}$

COMMENTS (read text). 1) If dataset contains many cases, divide cases into 3 or 4 groups according to X levels, then do a two sample t test on the d_i 's from the two extreme groups.

3.6 Breusch-Pagan Test (aka Cook-Weisberg Test)

- large sample test
- assumes error terms are independent, normally distributed and that variance of error terms, σ_i^2 , satisfies

$$\ln \sigma_i^2 = \gamma_0 + \gamma_1 X_i$$

- constant error variance implies $\gamma_1 = 0$ (Ho for test)
- test statistic

$$X_{BP}^2 = \frac{SSR^*/2}{SSE^2/n^2},$$

where SSR^* is regression sum of squares when regressing e^2 on X .

- If Ho holds and n is reasonably large, $X_{BP}^2 \sim \chi^2(1)$.

3.7 F Test for Lack of Fit

Official test requires repeat observations (replicates) at one or more X levels.

(If no replicates occur, use Minitab **xlof** subcommand.)

Let c denote number of distinct X levels.

- Lack of Fit Test

Hypotheses:

$$H_0: E\{Y\} = \beta_0 + \beta_1 X_i$$

$$H_a: E\{Y\} \neq \beta_0 + \beta_1 X_i$$

Test statistic: $F_{LOF}^* = \frac{MSLF}{MSPE} \underset{\text{under } H_0}{\sim} F(c-2, n-c)$

Decision rule: Reject H_0 (and conclude lack of fit) if $F_{LOF}^* > F_{1-\alpha, c-2, n-c}$.

- Partitioning SSESSE = SSPE + SSLF

$$\sum_{ij} (Y_{ij} - \hat{Y}_{ij})^2 = \sum_{ij} (Y_{ij} - \bar{Y}_j)^2 + \sum_{ij} (\bar{Y}_j - \hat{Y}_{ij})^2$$

$$j = 1, \dots, c; i = 1, \dots, n_j$$

$$\text{df: } n - 2 = n - c + c - 2$$

$$\text{df: } n - p = n - c + c - p$$

(where $p = \#$ pred. variables)

- F_{LOF}^* can be derived using the general linear test procedure with

full model: $Y_{ij} = \mu_j + \epsilon_{ij}$

reduced model: $Y_{ij} = \beta_0 + \beta_1 X_j + \epsilon_{ij}$

use $SSE(R) = SSE$ and $SSE(F) = SSPE$

- Expected mean squares: $E\{\text{MSLF}\} = \sigma^2 + \sum n_j (\mu_j - (\beta_0 + \beta_1 X_j))^2 / (c - 2)$
 $E\{\text{MSPE}\} = \sigma^2$

- If H_0 accepted, use MSE as estimator of σ^2 since it has larger degrees of freedom; always check for appropriateness of model before inferences drawn; need independent trials with respect to error for replicates (read text).

3.8 Overview of Remedial Measures

- read text

problem

remedial measure

nonlinearity

use a more appropriate model (Chapter 6 and beyond)

use a transformation (see text)

use exploratory analysis (later in text)

nonconstant error variance

use weighted least squares (later in text)

use transformation (see text)

nonindependent error terms

use model that works with correlated error (later in text)

work with first differences (later in text)

nonnormality of error terms

use transformation (see text)

omission of important predictor

use multiple regression (later in text)

outlying observations

discard if bogus (recording error)

use robust estimation procedure (later in text)

3.9 Transformations

- read appropriate section in text

A transformation of X is sometimes called for if the regression pattern exhibited by a scatter plot is a curvilinear band of constant vertical width. See the patterns and recommended transformations in text which include square root, exponential, logarithmic, reciprocal, and squaring transformations.

A transformation of Y is sometimes helpful in linearizing a curvilinear regression relation, especially when nonnormality and/or unequal error variances are suggested by residual diagnostics. See examples of such regression patterns in text.

- The Box-Cox procedure for finding an appropriate power transformation of Y can be done using a MINITAB macro.

The Box-Cox procedure finds a value for λ that makes the transformation

$Y' = Y^\lambda$ (with Y^0 defined to be $\ln Y$) provide a model with a better fit to the data.

The procedure finds a λ that minimizes the SSE when W is regressed on X where

$$W_i = \begin{cases} (Y_i^\lambda - 1) / [\lambda * gm(Y)^{\lambda-1}] & \text{for } \lambda \neq 0 \\ \ln(Y_i) * gm(Y) & \text{for } \lambda = 0 \end{cases}$$

and $gm(Y) = (\prod Y_i)^{1/n}$ denote the geometric mean of the Y_i .