Model Adequacy Checking – Part III

Johns Hopkins Engineering

625.461 Statistical Models and Regression

Module 7 – Lecture 7D

Residual Plots: Normal Plots

Gross nonnormality is potentially more serious as t or F statistics, confidence and prediction intervals depend on the normality assumption. If the errors come from a distribution with thicker or heavier tails than the normal, the LS fit may be sensitive to a small subset of the data. Heavy-tailed errors often generate outliers that "pull" the LS fit too much in their direction.

Normal Plots

Let $t_{[1]} < t_{[2]} < \ldots < t_{[n]}$ be the externally studentized residuals ranked in increasing order.

Plot $t_{[i]}$ against the cumulative probability,

$$P_i = (i - \tfrac{1}{2})/n, \; i = 1, 2, \ldots, n,$$

on the normal probability plot. If the normality holds, the resulting points should be approximately on a straight line. Emphasis on the central values (e.g., the 0.33 and 0.67 cumulative probability points)

Normal Plots

Sometimes normal probability plots are constructed by plotting the ranked residuals $t_{[i]}$ against the "expected normal value" $\Phi^{-1}[(i-\frac{1}{2})/n]$, where Φ denotes the standard normal cumulative distribution.

Normal Plots

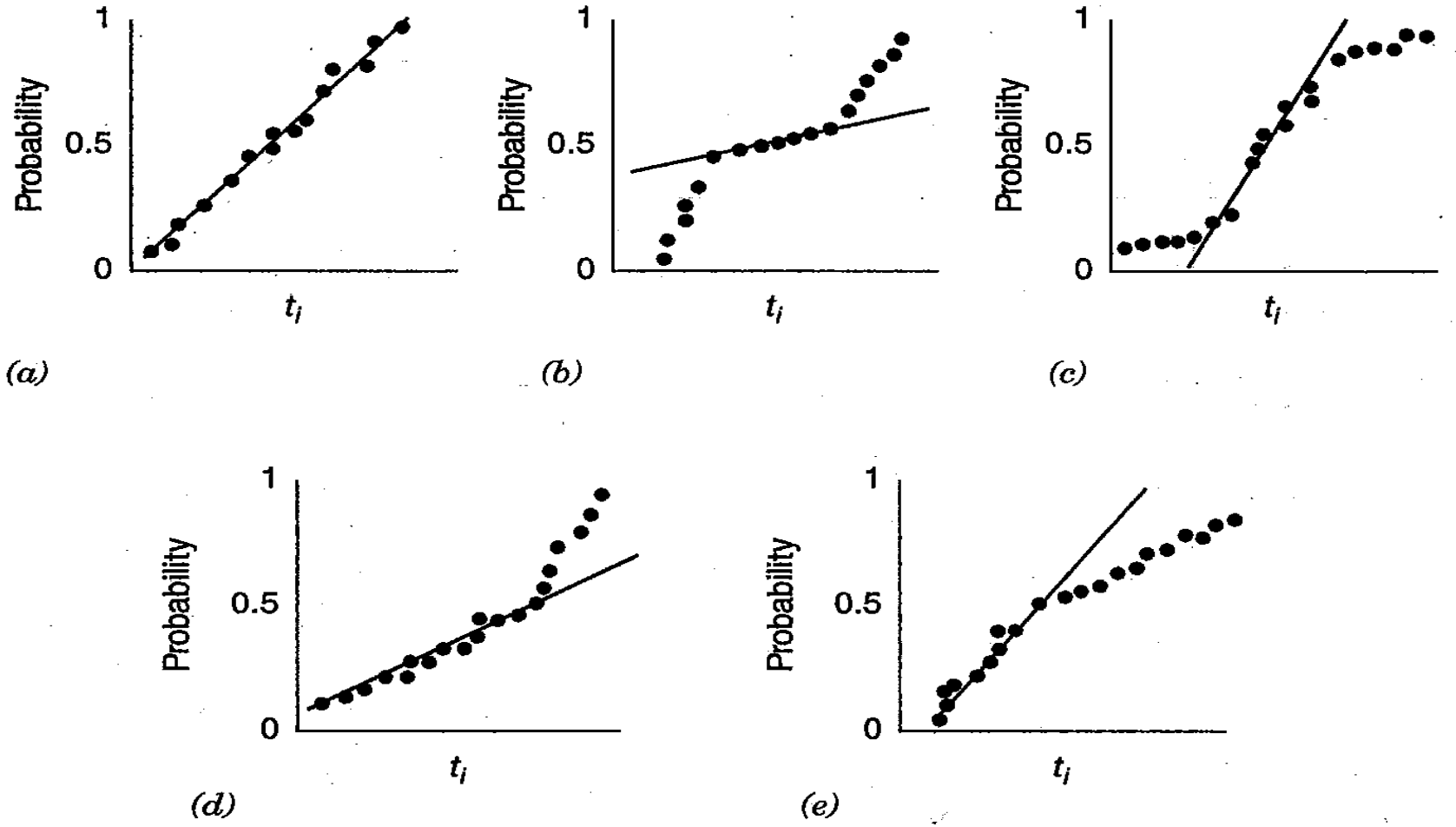


Figure 4.3 Normal probability plots: (*a*) ideal; (*b*) light-tailed distribution; (*c*) heavy-tailed distribution; (*d*) positive skew; (*e*) negative skew.

Plot of Residuals vs. Fitted Values or Regressor Values

Figure 4.5

4.5a: no obvious model deficiency

4.5b: variance of the errors is an increasing function of y

4.5c: often occurs when y is a proportion between 0 and 1

4.5d: nonlinearity

The Delivery Time Data (Ex 4.4, page 141 of Textbook)

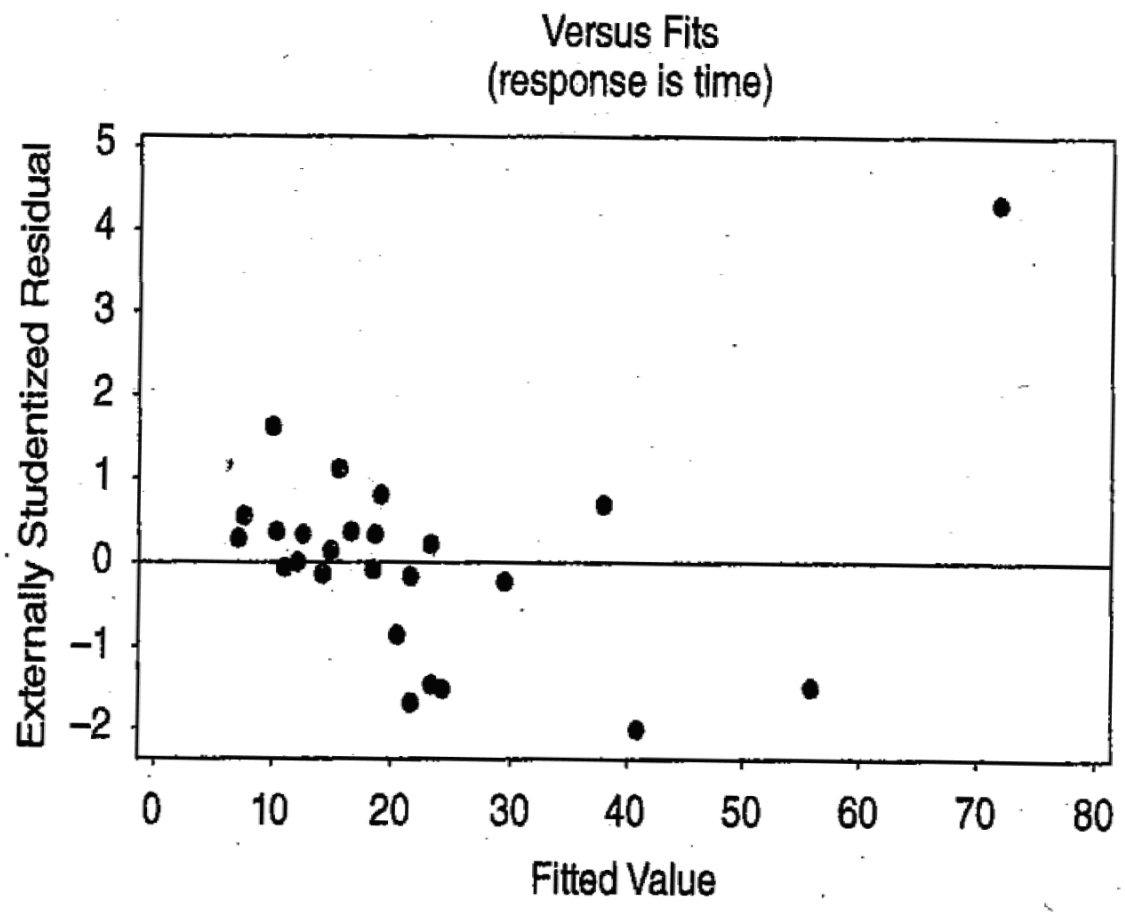
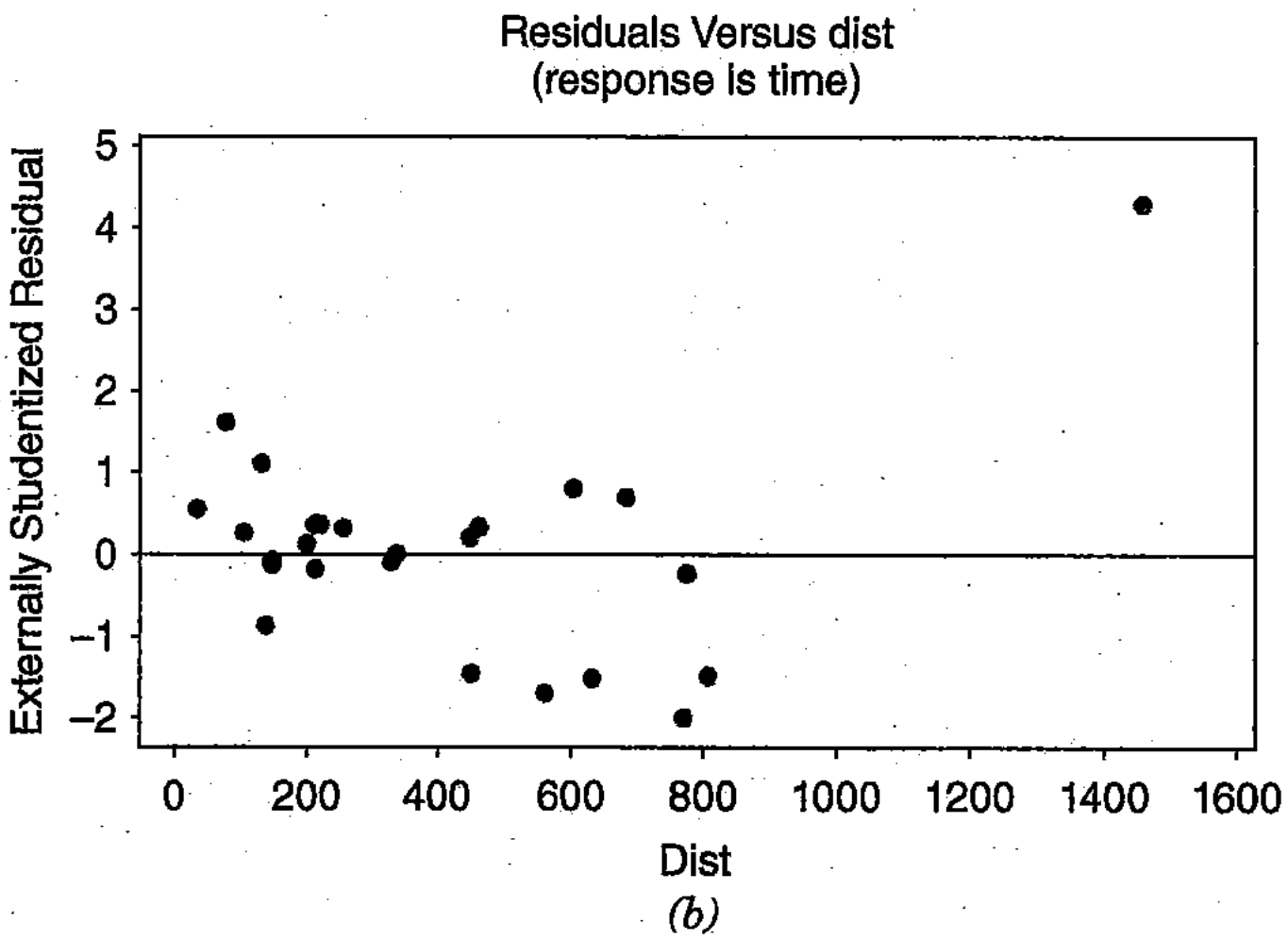


Figure 4.6 Plot of externally studentized residuals versus predicted for the delivery time data.

The Delivery Time Data (Ex 4.4, page 141 of Textbook)



(b)

PRESS Statistic

Use the prediction error sum of squares as a measure of model quality. The PRESS statistic is

$$\text{PRESS} = \sum_{i=1}^{n} [y_i - \hat{y}_{(i)}]^2$$

$$= \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

which is regarded a measure of how well a regression model will perform in predicting new data.

The Delivery Time Data (Ex 4.6, page 151 of Textbook)

See Table 4.1 (p.137). PRESS = 457.4 which is nearly twice as large as the $SS_{\text{Res}} = 233.7$

Almost half of the PRESS is contributed by point 9, a relatively remote point in x space with a moderately large residual. This indicates that the model will unlikely predict new observation with large case volumes and long distances particularly well.

The Delivery Time Data (Ex 4.6, page 151 of Textbook)

R^2 for prediction based on PRESS:

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{SS_T}$$

$$R^2_{\text{prediction}} = 1 - \frac{\text{PRESS}}{SS_T}$$

$$= 1 - \frac{457.4000}{5784.5426}$$

$$= 0.9209$$

The Delivery Time Data (Ex 4.6, page 151 of Textbook)

This model is expected to explain about 92.09% of the variability in predicting new observations, as compared to the approximately 95.96% of the variability in the original data explained by the LS fit.

The predictive capability of the model seems satisfactory, overall. But recall that the individual PRESS residuals indicated that observations that are similar to point 9 may not be predicted well.

JOHNS HOPKINS

WHITING SCHOOL
of ENGINEERING

© The Johns Hopkins University 2017, All Rights Reserved.