M 374G/384G, Fall 04

DIAGNOSTICS

Questions:

- Does the model fit?
- Is the result unduly influenced by one or a small number of points?

We've discussed some techniques to study these questions. (e.g., ___________________)

Some more:

I. **RESIDUAL PLOTS**

We've looked at these a little in cases with one or two terms, where arc easily generates them ("Remove Linear Trend"). Several types of residual plots can also be made fairly easily when more terms are involved.

Recall the error formulation of the models:

Ordinary Linear:

$$Y|\,\underline{x} = \underline{\eta}^T\underline{u} + e|\,\underline{x}$$

$e|\,\underline{x} \sim N(0, \sigma^2)$, independent of $\underline{x}$

Weighted linear:

$$Y|\,\underline{x}_i = \underline{\eta}^T\underline{u}_i + \frac{e_i}{\sqrt{w_i}}$$

$e_i \sim N(0, \sigma^2)$, independent of i

Recall the Least Squares fits

OLS:

$$\hat{y}_i = \underline{\hat{\eta}}^T\underline{u}_i$$

$$\hat{e}_i = y_i - \hat{y}_i$$

WLS:

$$\hat{y}_i = \underline{\hat{\eta}}^T\underline{u}_i$$

$$\hat{e}_i = \sqrt{w_i}\,(y_i - \hat{y}_i)$$

Intuitively, $\hat{e}_i$ is an estimate of e_i (= y_i - $\underline{\eta}^T\underline{u}_i$ for the ordinary model.) We know E(e_i) and E($\hat{e}_i$) are both zero, so $\hat{e}_i$ is an unbiased estimate of e_i. Thus it seems reasonable to plot the $\hat{e}_i$'s against various things to give some check on whether the model assumptions are reasonable.

However:

Recall: (Section 7.6 of book)

Var($\hat{e}_i|\underline{x}_i$) = $\sigma^2(1 - h_i)$, where h_i is the i^{th} leverage (whereas Var($e_i|\underline{x}_i$) = σ^2)

Thus

- If the h_i's are all small, then Var($\hat{e}_i|\underline{x}_i$) ≈ Var($e_i|\underline{x}_i$), so a plot using the $\hat{e}_i$'s should approximate a plot using the e_i's.
- If the h_i's are all approximately equal, then Var($\hat{e}_i|\underline{x}_i$) is approximately a constant times Var($e_i|\underline{x}_i$), so a plot using the $\hat{e}_i$'s should approximate a rescaled plot using the e_i's
- If the h_i's vary noticeably, then a plot using the $\hat{e}_i$'s will not give a good approximation of a plot using the e_i's. In this case, use instead *studentized residuals* $\dfrac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_i}}$.

Types of Residual Plots (roughly in order of importance)

- Against fitted values $\hat{y}_i$
- Against individual or pairs of predictors
- Against other possible predictors not in the model (especially time, location)
- Against individual terms other than predictors
- Against linear combinations of terms

Suggestions for Residual Plots for Specific Purposes

Checking linearity

Plot against fitted values $\hat{y}_i$. (Like "remove linear trend")

Checking constant variance

Plot against fitted values, predictors, pairs of predictors, other possible predictors.

Checking independence

Plot against other possible predictors

Checking normality

Use a normal probability plot.

Cautions:

- The usual cautions in interpreting normal plots
- Since $\sum \hat{e}_i$ = 0, the $\hat{e}_i$'s are *not* independent.

  (Thus only severe departures from a line should be taken as evidence of non-normality.)

Checking for outliers:

Plot against fitted values, predictors, pairs of predictors

II. **COOK'S DISTANCE**

With 1 or 2 terms, it's relatively easy to spot a potential influential point on the scatterplot and check if the point is influential. Leverage can help pick out x-outliers, which are potentially influential. *Cook's distance* can help check for influence.

The idea:

- Delete the i^{th} case and compute the OLS fit.
- Evaluate it at $\underline{x}_j$, giving $\hat{y}_{(i),j}$ = the fit at $\underline{x}_j$, not using the i^{th} case
- $D_i = \dfrac{1}{k\hat{\sigma}^2}\sum_{j=1}^{n}\left(\hat{y}_{(i),j} - \hat{y}_j\right)^2$ measures the total influence of the i^{th} case. (This can be expressed in terms of the coefficient estimates -- see more details in Section 15.2)

Rules of thumb in using D_i:

- Plot D_i vs case number.
- Examine cases that have relatively large D_i.
- Examine cases with $D_i > 0.5$, and especially cases with $D_i > 1$.
- There is no hypothesis test using D_i.