Residuals, Part II

Key terms

- External Studentization
- Outliers
- Added Variable Plot — Partial Regression Plot
- Partial Residual Plot — Component Plus Residual Plot

Key ideas/results

1. An *external* estimate of σ comes from refitting the model without observation i. Amazingly, it has an easy formula:
$$s_{-i} = \sqrt{\frac{(n-p)s^2 - e_i^2/(1-h_{ii})}{n-p-1}}$$

2. Externally Studentized Residuals
$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}}.$$
Ordinary residuals standardized with s_{-i}. Also known as R-Student.

3. Residual Taxonomy

Names	Definition	Distribution
Ordinary	$e_i = y_i - \hat{y}_i$	$\mathcal{N}(0, \sigma^2(1-h_{ii}))$
PRESS	$e_{i,-i} = e_i/(1-h_{ii})$	$\mathcal{N}(0, \sigma^2/(1-h_{ii}))$
Studentized Internally Studentized	$r_i = e_i/(s\sqrt{1-h_{ii}})$	$\approx\approx t_{n-p}$
Externally Studentized R-Student	$t_i = e_i/(s_{-i}\sqrt{1-h_{ii}})$	t_{n-p-1}

4. Outliers are unusually large observations, due to an unmodeled shift or an (unmodeled) increase in variance.

5. Outliers are not *necessarily* bad points; they simply are not consistent with your model. They may posses valuable information about the inadequacies of your model.

PRESS Residuals & Studentized Residuals

Recall that the PRESS residual has a easy computation form

$$\text{PRESS}_i = \frac{e_i}{1 - h_{ii}}.$$

It's easy to show that this has variance $\sigma^2/(1 - h_{ii})$, and hence a standardized PRESS residual is

$$\frac{\text{PRESS}_i}{s/\sqrt{1 - h_{ii}}} = \frac{e_i/(1 - h_{ii})}{s/\sqrt{1 - h_{ii}}} = \frac{e_i}{s\sqrt{1 - h_{ii}}} = r_i$$

When we standardize a PRESS residual we get the studentized residual! This is very informative. We understand the PRESS residual to be the residual at $\boldsymbol{x}_i$ if we had omitted $(y_i, \boldsymbol{x}_i)$ from the model. However, after adjusting for it's variance, we get the same thing as a studentized residual.

Hence the standardized residual can be interpreted as a standardized PRESS residual.

Internal vs External Studentization

The PRESS residuals remove the impact of point $(y_i, \boldsymbol{x}_i)$ on the fit at $\boldsymbol{x}_i$. But the studentized residual $r_i = e_i/(s\sqrt{1 - h_{ii}})$ can be corrupted by point i by way of s; a large outlier will inflate the residual mean square, and hence s.

Hence it advantageous to define an estimate of σ that is "external" to the ith observation. We define s_{-i} to be the standard deviation of the residual mean squared when the ith observation is omitted. Miraculously, it has a simple definition (see above); that is, we don't have to re-fit n models to obtain the n s_{-i}'s.

The *externally studentized* residual is a usual residual residual standardized using s_{-i} instead of the usual "internal" estimate s:

$$t_i = \frac{e_i}{s_{-i}\sqrt{1 - h_{ii}}}.$$

This is also know as *R-student*.

Another advantage of R-student is that, under the usual Normality assumptions, t_i exactly follows a t_{n-p-1} distribution. This will be important below, where we consider outlier detection.

Outlier Detection

When we observe an unusually large residual, we may suspect it is an outlier. There are two explanations, either there is an unmodeled shift or there is an increase in variance. These can be written as hypotheses:

$$\mathcal{H}_0 : \mathsf{E}(\epsilon_i) = 0 \quad \mathcal{H}_A : \mathsf{E}(\epsilon_i) \neq 0$$

and

$$\mathcal{H}_0 : \mathsf{Var}(\epsilon_i) = \sigma^2 \quad \mathcal{H}_A : \mathsf{Var}(\epsilon_i) \neq \sigma^2$$

Because the R-Student residual follows a t_{n-p-1} distribution (under both of these null hypothesis, corresponding to the model being correct), we can simply assess the magnitude of t_i to determine if point i is an outlier.

If you look at all n t_i values, don't forget that you are implicitly performing n hypothesis tests, and you must appropriately adjust any α thresholds with Bonferroni.

What to do with Outliers

If we find an outlier, we can't just toss it out. In fact, usually when you toss out all the apparent outliers and re-fit the model, new observations will appear to be outliers!

If an observation appears to be outlier, all the statistician can do is look for stupid errors, coding errors, etc. For example, if you find, say, a negative age or an 11 foot human, you can be fairly confident that this represents a typo. If it is not correctable, then you are justified in removing the observation.

Generally, however, all a statistician can do is mark a point as "suspect", and communicate this in a report. If the outliers are particularly bad you might perform two analyses, one with the outliers and one without. The validity of the reduced dataset rests solely on the outliers not being representative of the greater population of interest. In any case, if you alter the dataset in anyway (e.g. drop an observation, correct a typo), you must must clearly document this in your analysis report.

Diagnostic Plots

Residual diagnosis consists of plotting $\boldsymbol{e}$ versus anything. If the model assumptions are valid, no matter how we plot the residuals the should appear homogeneous... mean zero and with constant variance, that is, equally spread above and below zero and constant in spread.

Below are four types of diagnostic plots. For the first two, you will surely want to replace $\boldsymbol{e}$ with R-Student, but the latter two do not allow this generalization.

Names	Definition	Strengths & Weaknesses
Residuals vs Fitted	$\boldsymbol{e}$ vs $\hat{y}$	Good for checking homoscedasticity & general lack of fit.
Residuals vs Predictor	$\boldsymbol{e}$ vs X_j	Good for checking homoscedasticity & finding nonlinearities; but doesn't account for other predictors.
Partial Regression Plot Added Variable Plot	$\boldsymbol{e}_{y\|X_{-j}}$ vs $\boldsymbol{e}_{X_j\|X_{-j}}$	Accounts for other predictors; good for finding outliers, but not so good for identifying nonlinearities.
Component Plus Residual Plot Partial Residual Plot	$\boldsymbol{e} + x_j\hat{\beta}_j$ vs X_j	Good for identifying nonlinearities.
QQ-Normal Plot	$e_{(i)}$ vs $\mathsf{E}(z_{(i)})$	Only used for checking normality

Note the notation for Partial Regression–Added Variable plots: $\boldsymbol{e}_{y|X_{-j}}$ are the residuals from regression y on all regressors but j, and $\boldsymbol{e}_{x_j|X_{-j}}$ are the residuals from regressing X_j on all of the other predictors.