Residual Analysis and Leverage

This study guide covers outliers and some germane topics in residual analysis. Many of the formulas are primarily derived from *Linear Regression Analysis, Fifth Edition* by Montgomery et al. and/or the materials from the STAT 6021 course taught by Dr. Woo at the University of Virginia. It is believed that the materials are well-known equations and concepts in the public domain. If you believe otherwise, please reach out to me through my Github account so that I can correct the material. If not otherwise stated, quotes are from the textbook.

Outliers

An *outlier* is an observation that is very different from the others.

A *leverage point* is a point that is distant from the other points used to fit a model but that may or may not be consistent with the trend of the other points.

An *influence point* is distant but also inconsistent with the trend of the other points.

In other words, removing an influence point will impact the resulting fit of a model more than removing a leverage point.

Types of Residuals

Residuals

Types

Given a fitted model against a set of observations where y_i are the observed responses and $\hat{y}_i$ are the fitted values from the model for the same $\mathbf{x}_i$ predictors, then the residuals are

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

Standardized residuals

The standardized residuals are the

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}}$$

where MS_{Res} are computed as normal:

$$MS_{Res} = \frac{\sum_{i=1}^{n} e_i^2}{n-p} \quad i = 1, 2, \ldots, n.$$

These residuals effectively are scaling the regular residuals using the residual mean square, which is an approximation of the variance of the residuals. The regular residuals have a zero means as a consequence of the fitting method when using least squares.

If the residuals follow a normal distribution, the standardized residuals follow a standard normal distribution.

Studentized residuals

Studentized residuals utilize the standard deviation of individual residuals to perform a more accurate scaling. See p. 131 of the textbook for details.

The studentized residuals can be computed using

$$r_i = \frac{e_i}{\sqrt{MS_{Res}(1-h_{ii})}}, \quad i = 1, 2, \ldots, n.$$

PRESS Residuals

PRESS residuals are computed as

$$e_{(i)} = \frac{e_i}{1-h_{ii}}$$

The variance of a PRESS residual is

$$\mathrm{Var}[e_{(i)}] = \frac{\sigma^2}{1-h_{ii}}$$

Thus the standardized PRESS residual can be computed as

$$\frac{e_i}{\sqrt{\sigma^2(1-h_{ii})}},$$

and this becomes the studentized residual if we use MS_{Res} for the value of σ^2.

Interpretation

- If the PRESS residual for an observation differs greatly from the plain residual, this may indicate a high influence point (p. 135 of the textbook).

The R-Student residuals

Note: In long form, these are the "externally studentized" residuals.

Let us compute an estimate $S^2_{(i)}$ of σ^2 with the ith observation dropped. Then

$$S^2_{(i)} = \frac{(n-p)MS_{Res} - e_i^2/(1-h_{ii})}{n-p-1}.$$

Then you can calculate the R-student residuals as

$$t_i = \frac{e_i}{\sqrt{S^2_{(i)}(1-h_{ii})}}.$$

The book says the following: > It turns out that under the usual regression assumptions, t_i will follow the t_{n-p-1} distribution.

See the book for additional details.

Interpreting the residuals

The book says, "Examining **scaled residuals**, such as the studentized and R-student residuals, is an excellent way to identify potential outliers."

- The book says that if a residual is more than 3-4 standard deviations from the mean, it may be an outlier.
- As a more specific test, on p. 135, the book says that under standard assumptions, the externally studentized residuals t_i should follow the t_{n-p-1} distribution. Therefore, you could compare $|t_i|$ to $t_{(a/2n),n-p-1}$ to look for outliers.

Additional notes

- If the standardized residuals have a large value, then they are probably outliers.
- The studentized residuals should have constant variance $\mathrm{Var}(r_i) = 1$ for a correct model
- Often standardized and studentized residuals convey the same information, although this only happens if the "variance of the residuals stabilizes [e.g.,] for large data sets".
- The book cites several sources for tests. Please read the book.

The hat matrix

According to the book, the leverage can be related to the hat matrix: > "The elements $h_{ij} of the matrix **H** may be interpreted as the amount of **leverage** exerted by the ith observation y_i on the jth fitted value $\hat{y}_j$.

Recall that the hat matrix is

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

The diagonals are thus

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i.$$

Relation to properties of residuals

The residuals are related to the model fit error variable:

$$\mathbf{e} = (\mathbf{I} - \mathbf{H})\varepsilon.$$

The covariance matrix is

$$\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H}).$$

The standard deviation of a residual is

$$\mathrm{Var}(e_i) = \sigma^2(1 - h_{ii})$$

and the covariance of residuals is

$$\mathrm{Cov}(e_i, e_j) = -\sigma^2 h_{ij}.$$

Relation to calculating dropping points

Let $\hat{y}_n^*$ be the fitted value when ignoring the nth data point. Then

$$\hat{y}_n = \hat{y}_n^* + h_{nn}\delta.$$

The textbook also states a form without the hat matrix:

$$\hat{y}_n = \hat{y}_n^* + \left[\frac{1}{n} + \left(\frac{n-1}{n}\right)^2 \frac{(x_n - \bar{x}^*)^2}{S_{xx}}\right].$$

Other properties

The mean over the diagonals is

$$\overline{h_{ii}} = p/n,$$

where p is the number of coefficients ($p = k + 1$) and n is the number of observations.

Measures of influence

Hat matrix as leverage

See above for a general discussion. If $h_{ii} > 2p/n$ for any observation, this is likely a leverage point (p. 213 of the textbook). Note that the main deficiency of this approach is that it uses only the locations vis-a-vis the regressors, and it ignores the response variable and fitted values.

Cook's D

Cook's D is (per the book)

> "...a measure of the squared distance between the least-squares estimate based on all n points $\hat{\boldsymbol{\beta}}$ and the estimate obtained by deleting the ith point, say $\hat{\boldsymbol{\beta}}_{(i)}$."

Matric form

The basic formula is

$$D_i(\mathbf{M}, c) = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' \mathbf{M}(\mathbf{M}, c) = \left(\boldsymbol{\beta}_{(i)} - \boldsymbol{\beta}\right)}{c}, \qquad i = 1, 2, \ldots, n$$

If we set $\mathbf{M}, c$ appropriately, we get a variant of D_i:

$$D_i = D_i(\mathbf{X}'\mathbf{X},\ \ p\, MS_{Res}) = \frac{\left(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}}\right)' \mathbf{X}'\mathbf{X}(\mathbf{M}, c) = \left(\boldsymbol{\beta}_{(i)} - \boldsymbol{\beta}\right)}{p\, MS_{Res}}, \qquad i = 1, 2, \ldots, n$$

Alternate forms

Alternatively:

$$D_i = \frac{r_i^2}{p} \frac{h_{ii}}{1 - h_{ii}}$$

Alternatively:

$$D_i = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{p\,MS_{Res}}$$

Interpretation

The book says that points for which $D_i > 1$ are often taken to be influential.

According to the textbook, it is typical to compare D_i and $F_{0.5,p,n-p}$, although D_i is *not* an F statistic:

> "If [they are equal], then deleting point i would move $\hat{\boldsymbol{\beta}}_{(i)}$ to the boundary of an approximate 50% confidence region for $\boldsymbol{\beta}$ based on the complete data set... The distance D_i is not an F statistic."

$DFBETAS_{j,i}$

Definition

The value can be defined as

$$DFBETAS_{j,i} = \frac{\hat{\boldsymbol{\beta}}_j - \hat{\boldsymbol{\beta}}_{j,(i)}}{\sqrt{S_{(i)}^2 C_{jj}}},$$

where we are dealing with the effect of dropping the ith observation on the jth coefficient.

where C_{jj} is the appropriate utility matrix. See p. 217 for details.

Alternatively, you can define it as

$$DFBETAS_{j,i} = \frac{r_{j,i}}{\sqrt{\mathbf{r}_j' \mathbf{r}}} \frac{t_i}{\sqrt{q - h_{ii}}},$$

where r_j' is the jth row of $(\mathbf{X}'X)^{-1}\mathbf{X}'$ and t_i is the R-student residual.

Interpretation

- Per the book: "$DFBETAS_{j,i}$ indicates how much the regression coefficient $\hat{\beta}_j$ changes, in standard deviation units, if the ith observation were deleted."
- "$DFBETAS_{j,i}$ measures both leverage ... and the effect of large residual", according to the book.
- If $|DFBETAS_{j,i}| > 2/\sqrt{n}$, the ith observation may be a problematic point.

$DFFITS_i$

Definition

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_{(i)}}{\sqrt{S_{(i)}^2 h_{ii}}}, \qquad i = 1, 2, \ldots, n$$

with $y_{(i)}$ having the appropriate interpretation of the fitted value if we drop the ith observation, per the book.
Also per the book, the denominator is a standardization, given the value of $\mathrm{Var}(\hat{y}_i)$.

Alternatively:

$$DFFITS_i = \left(\frac{h_{ii}}{1 - h_{ii}}\right)^{1/2} t_i.$$

Interpretation

- Per the book: "$DFFITS_i$ is the number of standard deviations that the fitted value $\hat{y}_i$ **changes** if observation i is removed."
- Again per the book: "$DFFITS_i$ is affected by both leverage and prediction error"
- Again per the book: "… any observation for which $|DFFITS_i| > 2\sqrt{p/n}$ warrants attention."