Lab 7 Outliers and Influence Revisit

October 29, 2020

Outline

Outline

Outliers

- An outlier is a point with a large residual.
- We use the *mean shift outlier model* to define outliers.
- We assume that the mean function for all other cases is

$$E(Y|X = x_j) = x_j^T \beta,$$

but for case i the mean function is

$$E(Y|X = x_i) = x_i^T \beta + \delta.$$

- The expected response for the ith case is shifted by an amount δ, and a test of $\delta = 0$ is a test for a single outlier in the ith case.

Outline

An Outlier Test

- Suppose that the ith case is suspected to be an outlier.
- Suppose that the ith case is suspected to be an outlier. First, define a new term, say U, with the jth element $u_j = 0$ for $j \neq$i, and the ith element $u_i = 1$. Thus U is a dummy variable that is zero for all cases but the ith.
- Simply compute the regression of the response on both the terms in X and U. That is we fit a new model $Y \sim X + U$. The estimated coefficient for U is the estimate of the mean shift δ.
- The t-statistic for testing $\delta = 0$ against a two-sided alternative is the appropriate test statistic.
- Normally distributed errors are required for this test, and then the test will be distributed as Student t with $n - p - 2$ df.

An Outlier Test

An alternative approach.

- Again suppose that the ith case is suspected to be an outlier.
- Delete the ith case from the data, so $n-1$ cases remain in the reduced data set.
- Using the reduced data set, estimate β and σ^2. Call these estimates $\hat{\beta}_{(i)}$ and $\hat{\sigma}^2_{(i)}$. The estimator $\hat{\sigma}^2_{(i)}$ has $n-p-2$ df.
- For the deleted case, compute the fitted value $\hat{y}_{(i)} = x_i^T \hat{\beta}_{(i)}$. Since the ith case was not used in estimation, y_i and $\hat{y}_{i(i)}$ are independent. The variance of $y_i - \hat{y}_{i(i)}$ is given by

$$Var(y_i - \hat{y}_{i(i)}) = \sigma^2 + \sigma^2 x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i$$

where $X_{(i)}$ is the matrix X with the ith row deleted. This variance is estimated by replacing σ^2 with $\hat{\sigma}^2$.

An Outlier Test

An alternative approach continue.

- Now $E(y_i - \hat{y}_{i(i)}) = \delta$, which is zero under the null hypothesis that case i is not an outlier but nonzero otherwise. Assuming normal errors, a Student t-test of the hypothesis $\delta = 0$ is given by

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i^T (X_{(i)}^T X_{(i)})^{-1} x_i}}$$

  This test has $n - p - 2$ df, and is identical to the t-test suggested in the previous approach.

Computation of t_i

Define an intermediate quantity often called a standardized residual, by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

(Standardized in the sense that $Var[\hat{e}_i] = \sigma^2(1-h_{ii})$, standardized the variance). We can show that

$$t_i = r_i\left(\frac{n-(p+1)-1}{n-(p+1)-r_i^2}\right)^{1/2} = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}.$$

The cool think is that we can compute t_i without ever having to actually delete the observation and re-fit the model.

Outline

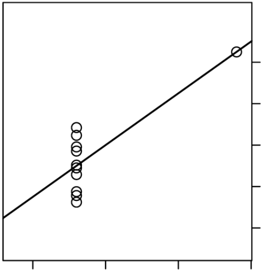
Significance Levels for the Outlier Test

- If the analyst suspects in advance that the *i*th case is an outlier, then t_i should be compared with the central t-distribution with the appropriate number of df.
- Testing the case with the largest value of $|t_i|$ to be an outlier is like performing n significance tests, one for each of n cases.
- The technique we use to find critical values is based on the *Bonferroni* correction, which states that for n tests each of size α, the probability of falsely labeling at least one case as an outlier is no greater than $n\alpha$.
- Choosing the critical value to be the $(\alpha/n) \times 100\%$ point of t will give a significance level of no more than $n \times (\alpha/n) = \alpha$.

Outline

Influence of Cases

- Single cases or small groups of cases can strongly influence the fit of a regression model. Example of anscombe.txt data.



Recall that

$$\hat{Y} = HY$$

where the H is the hat matrix. This means that each $\hat{Y}_i$ is a linear combination of elements of H. In particular, H_{ii} is the contribution of the i^{th} data point to $\hat{Y}_i$. For this reason we call $h_{ii} = H_{ii}$ the *leverage*. leverage.

Outline

Cook's Distance

To get a better idea of how influential the i^{th} data point is, we could ask: how much do the fitted values change if we omit an observation? Let $\hat{Y}^{(-i)}$ be the vector of fitted values when we remove observation i. Then Cook's distance is defined by

$$D_i = \frac{(\hat{Y} - \hat{Y}^{(-i)})^T(\hat{Y} - \hat{Y}^{(-i)})}{(p+1)\hat{\sigma}^2}$$

It turns out that there is a handy formula for computing D_i, namely:

$$D_i = \left(\frac{r_i^2}{p+1}\right)\left(\frac{h_{ii}}{1-hii}\right),$$

This means that the influence of a point is determined by both its residual and its leverage. Often, people interpret $D_i > 1$ as an influential point.

Note that $\hat{Y} = X\hat{\beta}$ and $\hat{Y}^{(-i)} = X\hat{\beta}^{(-i)}$, then we have

$$D_i = \frac{(\hat{\beta}^{(-i)} - \hat{\beta})^T X^T X (\hat{\beta}^{(-i)} - \hat{\beta})}{(p+1)\hat{\sigma}^2}.$$

Diagnostics in practice

We have three ways of looking at whether points are outliers:

- We can look at their leverage, which depends only on the value of the predictors
- We can look at their studentized residuals, either ordinary or cross-validated, which depend on how far they are from the regression line.
- We can look at their Cook's statistics, which say how much removing each point shifts all the fitted values; it depends on the product of leverage and residuals.