Diagnostic checks of regression models

The model we arrive at when performing regression analysis depends strongly on the quality of the data.

Problems

Multicollinearity: Two or more columns in the $\underline{X}$-matrix are almost linear dependent.

checked by the correlation between regression variables.

Estimated correlation between x_i and x_j is given by

$$r_{ij} = \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}$$

Multicollinearity often occurs when the variation interval of two or more variables are approximately equal.

A good measure of multicollinearity is the variance inflation factor.

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is the multiple determination coefficient when x_j is regressed on the other regression variables.

Important $VIF < 10$. If it occurs, consider.

1. Remove columns
2. Collect more data
3. Respecify the model. Instead of x_1, x_2, x_3, $\frac{x_1 + x_2 + x_3}{3}$ or $x_1 x_2 x_3$ may be used.
4. Center variables $(x_i - \bar{x})$ in polynomial models, or standardize.

5. Use PCA regression or PLS regression, eventually Ridge regression.

Influential observations.

Leverage points. $\hat{y} = \underline{H}\,\underline{y} \Rightarrow \hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j = h_{ii} y_i + \sum_{i \neq j} h_{ij} y_j$

If h_{ii} is large, y_i could have a large influence on $\hat{y}_i$

$\text{rank}(\underline{H}) = k+1 = \sum_{i=1}^{n} h_{ii}$. Average $h_{ii} = \frac{k+1}{n}$. Observations where $h_{ii} > \frac{2(k+1)}{n}$ are said to be leverage points.

~~Cook's distance D_i is~~

Cook's distance

$$(\underline{\hat{y}} - \underline{\hat{y}}_{(-i)})^T (\underline{\hat{y}} - \underline{\hat{y}}_{(-i)}) = (\underline{X}\underline{b} - \underline{X}\underline{b}_{(-i)})^T (\underline{X}\underline{b} - \underline{X}\underline{b}_{(-i)}) = (\underline{b} - \underline{b}_{(-i)})^T \underline{X}^T \underline{X} (\underline{b} - \underline{b}_{(-i)})$$

The $(-i)$ notation means that the i-th set of observations $(y_i, x_{i1}, \dots, x_{ik})$ is taken out.

Cook's distance D_i is defined as

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T \underline{X}^T \underline{X} (\hat{\beta} - \hat{\beta}_{(-i)})}{(k+1)\sigma^2}, \quad i = 1, 2, \dots, n$$

with substituted estimates: $\frac{h_{ii} d_i^2}{(1-h_{ii})(k+1)}$ where $d_i = \frac{e_i}{s\sqrt{1-h_{ii}}}$, $s = \hat{\sigma}_{LS}$

We consider observation $(y_i, x_{1i}, \dots, x_{ki})$ to be influential if $D_i > 1$

Study of residuals

If the model is correct, we expect the residuals to be uncorrelated with mean zero and constant variance.

$$\underline{\hat{\varepsilon}} = \underline{y} - \underline{\hat{y}} = (\underline{I} - \underline{H})\underline{y} = (\underline{I} - \underline{H})\underline{\varepsilon}$$

Thereby $E[\underline{\hat{\varepsilon}}] = (\underline{I} - \underline{H}) E[\underline{\hat{\varepsilon}}] = \underline{0}$

and $Cov[\underline{\hat{\varepsilon}}] = E[(\underline{I} - \underline{H})\underline{\varepsilon}\,\underline{\varepsilon}^T(\underline{I} - \underline{H})] = \sigma^2(\underline{I} - \underline{H})$

which shows that $\hat{\varepsilon}_1, \hat{\varepsilon}_2, \cdots, \hat{\varepsilon}_n$ are not uncorrelated.

$$Var[\hat{\varepsilon}_i] = \sigma^2(1 - h_{ii})$$

Hence the variances of $\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_n$ need not to be equal.

Since $Var\left[\dfrac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}}\right] = 1$, we often use $r_i = \dfrac{e_i}{s\sqrt{1-h_{ii}}}$ in the study of the residuals. These are called standardized residuals. Some books call them studentized or internal studentized.

The residuals $\delta_i = y_i - \hat{y}_{i,(-i)}$ $i = 1, 2, \cdots, n$ where $\hat{y}_{i,-i}$ is the estimator for the i-th fitted value when the observation $(y_i, x_{i1}, \cdots, x_{ik})$ is taken out, are called the Press residuals (Predictive sum of squares). These can be shown to equal $\dfrac{\hat{\varepsilon}_i}{1-h_{ii}}$ and

$$Var(\delta_i) = \frac{Var[\hat{\varepsilon}_i]}{(1-h_{ii})^2} = \frac{\sigma^2(1-h_{ii})}{(1-h_{ii})^2} = \frac{\sigma^2}{1-h_{ii}}$$

Hence $\frac{\delta_i}{sd(\delta_i)} = \frac{\hat{\varepsilon}_i}{(1-h_{ii})\sigma} = \frac{\hat{\varepsilon}_i}{\sigma\sqrt{1-h_{ii}}}$

Assume we estimate σ^2 by $s^2_{(-i)}$ which is the ~~residual~~ estimate for the variance when $(y_i, x_{i1}, \ldots x_{ik})$ is taken out. Then $\delta_i = Y_i - \hat{Y}_{i,-i}$ and $s^2_{(-i)}$ are independent ($\hat{\beta}$ and $\hat{\sigma}^2$ are independent)

and $\frac{\hat{\varepsilon}_i}{s_{(-i)}\sqrt{1-h_{ii}}} \sim t_{n-k-2}$ and can be used to

investigate if the expected value of a residual is different from zero, i.e if we have an outlier

The residuals $\frac{e_i}{s_{(-i)}\sqrt{1-h_{ii}}}$ are called R-studentized or

sometimes External studentized.

Check of residuals provide information about

1. Outliers
2. Heterogeneity of variance
3. Misspecification of models (or if other variables should be included.
4. Normal distribution

The most common residual plots are plot of residuals against fitted values to check for heterogeneity in variance, against other variables to see if they should be included, and normal-plot to check for normally distributed data and outliers.