

Regression Diagnostics

There are n observations and p explanatory variables in the current model.

- residual = $y - \hat{y}$
- Standardized residual: $z = \frac{(y - \hat{y})}{\text{RMSE}}$

The distribution of z is approximately standard normal.

- An observation is said to have high leverage if $h > 2\left(\frac{p+1}{n}\right)$

$$F = \frac{\left(h - \frac{1}{n}\right) / p}{(1-h)/(n-p-1)}$$

The distribution of F is an F-distribution with p and $n - p - 1$ degrees of freedom.

- Cook's D: $d = \left(\frac{h}{p+1}\right) \left(\frac{z}{1-h}\right)^2$

An observation is said to have high influence if $d > 1$.

- Studentized residual: $r_s = \frac{z}{\sqrt{(1-h)}}$

The distribution of r_s is approximately a t-distribution with $n - p - 1$ degrees of freedom.

- Variance Inflation Factor: $VIF_i = \frac{1}{(1 - R_i^2)}$ where R_i^2 is the RSquare value when you regress X_i against all the other explanatory variables, $X_1, X_2, \dots, X_{i-1}, X_{i+1}, \dots, X_p$.

There may be severe multicollinearity if $VIF_i > 10$.