3.5.2 → Leverage & Residuals

→ Overview → To better understand what makes an observation unusual or influential, let's discuss some concepts ideas in relation to leverage, residuals & Cook's distance

→ Leverage

→ The leverage of an observation measures its influence in predicting the response. We denote the i^{th} leverage as h_i, which is the i^{th} diagonal entry of the hat matrix H. Recall the hat matrix is

$$H = \begin{bmatrix} h_1 & & \\ & h_2 & \\ & & \ddots \\ & & & h_n \end{bmatrix}$$

→ In the case of SLR, $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \downarrow = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX}$

→ other facts about leverage

→ It is a function of the x_i's but not y

→ The larger h_i is, the more unusual the set of $x_1, \dots, x_p$ values is relative to the other observations

→ $\sum_{i=1}^{n} h_i = p+1$

→ There is nothing inherently bad about an observation w/ a high leverage. The concern is whether $\hat{y}$ is mainly driven by a few observations that dominate a vast amount of the "total leverage" ($p+1$), to the point that $\hat{y}$ would alter drastically without these observations. Our rule of thumb is that the i^{th} observation is a high leverage point if

$$h_i > 3\left(\frac{p+1}{n}\right)$$

→ Residuals

→ recall that a residual is $y - \hat{y}$ (actual − predicted). This means it has the same units as y. Consequently, the value of an extreme residual depends on the unit. For this reason, it is common to analyze a version of the residuals that is unitless. There are two types to discuss:

→ Standardized residuals & studentized residuals

→ Standardized residuals are the residuals divided by an estimate standard error. The i^{th} standardized residual is

$$e_{sta,i} = \frac{e_i}{\sqrt{MSE(1-h_i)}}$$

→ since the standard error is estimated, the standardized residuals are approximate realizations of the standard normal distribution, provided the model is correct

→ Studentized residuals are similar to standardized residuals, except they are divided by a different estimated standard error. The i^{th} (externally) studentized residual is

$$e_{stu,i} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_i)}}$$

→ where $MSE_{(i)}$ is the MSE of the regression that excludes the i^{th} observation from the training data.

→ The studentized residuals are realizations of a t-dist, provided the model is correct. Recall $t \to Z$ as $df \to \infty$

⇒ z 99% $\in [-3, 3]$

$\approx$ 99.7% $\in [-3, 3]$

→ This allows us to choose a sensible cutoff in defining an outlier, regardless of the unit of the response. It is rather common to find the rule of thumb of $>|3|$ to identify outliers. Moreover, both types of residuals are equally effective in determining outliers.

→ Cook's distance

→ If we wish to combine leverage & residuals into a single measure, we can compute either DFITS or Cook's distance for each observation.

$$DFITS_i = e_{sta,i}\sqrt{\frac{h_i}{1-h_i}}$$

The i^{th} Cook's distance is

$$d_i = \frac{DFITS_i^2}{p+1} \downarrow = \frac{e_{sta,i}^2 h_i}{(p+1)(1-h_i)} \downarrow = \frac{e_i^2 h_i}{MSE(p+1)(1-h_i)^2}$$

$\because e_{sta,i} = \frac{e_i}{\sqrt{MSE(1-h_i)}}$ →

→ one rule of thumb is that the i^{th} observation is an influential point if d_i exceeds unity, i.e. $d_i > 1$

→ Assignment

→ Gill Given → $n = 15$, $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ ($p = 3$), $d_6 = 0.0715$, $e_{sta,6} = 0.9032$, $h_{6,6} = ?$

$$d_6 = \frac{DFITS_6^2}{p+1} \downarrow = \frac{e_{sta,6}^2 (h_6)}{(p+1)(1-h_6)}$$

$$0.0715 = \frac{0.9032^2 h_6}{4(1-h_6)}$$

$$\underbrace{\frac{0.0715}{0.9032^2}}_{\to c} = \frac{h_6}{1-h_6}$$

$$c - ch_6 = h_6$$

$$c = h_6(1+c)$$

$$\Rightarrow h_6 = \frac{c}{1+c} \downarrow = 0.17339$$

→ Dai) Given → $p = 12$ (excludes intercept)

$SSE = 618$

$h_{10,10} = 0.35$

$e_{10} = 9.5$

$d_{10} = 5$

→ $n = ?$

→ $d_{10} = \frac{e_{10}^2 h_{10}}{\frac{SSE}{n-(p+1)}(p+1)(1-h_{10})^2}$

$$5 = \frac{9.5^2(0.35)}{\frac{618}{n-(12+1)}(12+1)(1-0.35)^2} \downarrow = \frac{(n-13)\,9.5^2(0.35)}{618(13)(1-0.35)^2}$$

$\Rightarrow n = 550.24 \Rightarrow n \geq 551$

~~3.5.4 → VIF~~

→ overview → one way to measure multicollinearity is the variance inflation factor (VIF). To calculate the VIF for the j^{th} predictor, first run a MLR w/ x_j acting as the response variable, predicted by the rest of the $p-1$ predictors. Denote the coefficient of determination of that regression as R_j^2. Then

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

→ Intuitively, R_j^2 is itself a suitable measure for whether x_j is almost a linear combination of the other predictors. A reason to use VIF instead comes from how $SE(\beta_j)$ can be written as

$$SE(\beta_j) = \sqrt{VIF_j}\sqrt{\frac{MSE}{(n-1)S_{xx}}}$$ → S_{xx} for regression of x_j on remaining $p-1$ vars

→ Since $R_j^2 = 0.8 \Rightarrow VIF_j = 5$, it is suggested that we should be concerned w/ multicollinearity if any of the p VIFs is 5 or greater

→ Another inspection way to detect multicollinearity

→ calculate all β_j's & r_{xy} (simple correlation between the response & each predictor)

→ If signs differ, then this contradiction is possibly the result of multicollinearity (i.e. some interaction from multiple predictors)