Outlier Detection

Model: $$Y_i \mid \beta, \sigma^2 \overset{ind}{\sim} N(x_i^T\beta, \sigma^2)$$

Residual Analysis: $$\frac{Y_i - x_i^T\beta}{\sigma} \Big| \beta, \sigma^2 \overset{iid}{\sim} N(0,1)$$

Need to estimate $\hat{e}_i = \dfrac{Y_i - x_i^T\hat{\beta}}{\hat{\sigma}} \underset{\substack{\uparrow \\ \text{if good estimates}}}{\dot{\sim}} N(0,1)$

Should be rare to see $|\hat{e}_i| \geq 2$ so detect outliers with the rule

$$\left| \frac{Y_i - x_i^T\hat{\beta}}{\hat{\sigma}} \right| > K_\alpha \quad \text{for} \quad K_\alpha = -\Phi(\alpha/2)$$

In the Bayesian paradigm $\beta \mid Y, \sigma^2 \sim N(\hat{\beta}, \Sigma_\beta)$

so $Y_i - x_i^T\beta \mid Y, \sigma^2 \sim N(Y_i - x_i^T\hat{\beta},\ x_i^T \Sigma_\beta x_i)$

Let's assume for simplicity $P(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$

$$\beta \mid \sigma^2, Y \sim N(\hat{\beta}_{MLE}, \sigma^2 (X^TX)^{-1})$$

$$\sigma^2 \mid Y \sim \text{Inv-}\chi^2\left(n-p, \frac{RSS(\hat{\beta}_{MLE})}{n-p}\right) \equiv \text{Inv-}\chi^2(n-p, \hat{\sigma}^2_{MLE})$$

Aside $\beta | y \sim t_{n-p}\left(\hat{\beta}_{MLE}, \hat{\sigma}^2 (X^TX)^{-1}\right)$

With this, it makes sense to discuss

$$y_i - x_i^T\beta \mid \sigma^2, y \sim N\left(y_i - x_i^T\hat{\beta}_{MLE}, \sigma^2 \underbrace{x_i^T(X^TX)^{-1}x_i}_{H_{ii}}\right)$$

Aside $y_i - x_i^T\beta | y \sim t_{n-p}\left(y_i - x_i^T\hat{\beta}, \hat{\sigma}^2 h_{ii}\right)$

For large n, $\sigma^2 | y$ concentrates around $\hat{\sigma}^2 = \dfrac{RSS(\hat{\beta}_{MLE})}{n-p}$ and

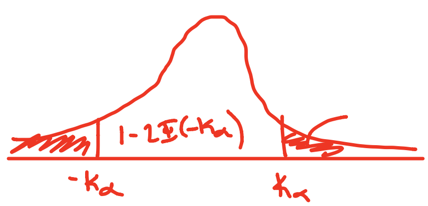
$$\boxed{\frac{y_i - x_i^T\beta}{\sigma} \,\Big|\, y \;\dot\sim\; N\left(\frac{y_i - x_i^T\hat{\beta}}{\hat{\sigma}}, h_{ii}\right)}$$ Posterior Residual dist.

With this we can calculate $\mathbb{P}\left(\left|\frac{y_i - x_i^T\beta}{\sigma}\right| \leq k_\alpha \,\Big|\, y\right)$ will do this numerically.

We will need a multiple test correction.

Let $p(k_\alpha)$ = prob of being an outlier the we want

$$\mathbb{P}(\text{no outliers}) = p(k_\alpha)^n \equiv 1 - \alpha$$



$$p(k_\alpha) = 1 - 2\Phi(-k_\alpha) = (1-\alpha)^{1/2}$$

$$\Rightarrow \boxed{K_\alpha = -\Phi^{-1}\left(\tfrac{1}{2}\left(1-(1-\alpha)^{1/2}\right)\right)}$$

Rmk: Bonferroni Correction.

Finally, we can use this for model checking.

$$Y - X\beta \mid \sigma^2, Y \sim N\left(Y - X\hat{\beta},\ \sigma^2 \underbrace{X(X^TX)^{-1}X^T}_{H}\right)$$

H may be rank deficient; $X = QR$, $\hat{\beta} = (X^TX)^{-1}X^TY$

$$= (R^TQ^TQR)^{-1}R^TQ^TY$$

$$= R^{-1}R^{-T}R^TQ^TY$$

$$\Rightarrow \boxed{R\hat{\beta} = Q^TY}$$

numerically stable to find $\hat{\beta}$

$$Q^T(Y - X\beta) \mid \sigma^2, Y \sim N\left(Q^TY - Q^TQR\hat{\beta},\ \sigma^2 Q^TQQ^TQ\right)$$

$$\equiv N(0, \sigma^2 I_p)$$

$$\Rightarrow \frac{1}{\sigma^2} Q^T(Y - X\beta) \mid Y \sim N(0, I_p)$$

Suggests use of a QQ-Plot on standardized residuals