STAT 308 – Section 14.3

Background Information

Recall from Chapter 5, we said that our simple linear model has the form

$$Y = \mu_{Y|X} + \epsilon = \beta_0 + \beta_1 X + \epsilon,$$

where the errors/residuals $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We said that the least squares regression equation is

$$\hat{Y}_x = \hat{\beta}_0 + \hat{\beta}_1 x,$$

and for $i = 1, \ldots, n$,

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

Therefore, we can define our observed residuals as

$$\hat{\epsilon}_i = y_i - \hat{y}_i.$$

These observed residuals will be important to us in determining which observations may be influential points in our data analysis.

Important Definitions

Influential Observations: Observations that *may* influence the estimation of the least squares regression line.

Example

Recall the blood pressure dataset. Determine visually if you think their might be any influential observations.

Leave One Out Regression: A new regression line where one observation is intentionally left out.

The leave one out least squares regression when observation i is left out for $i = 1, \ldots, n$ gives regression estimates $\hat{\beta}_{0,(-i)}$ and $\hat{\beta}_{1,(-i)}$ and an estimate of the regression variance s^2_{-i}. These estimates will help us to identify which points have the highest influence on the original least squares regression line

Leverage: (h_i) A measure of the extremeness of the observed explanatory variables to their means

More formally, h_i is the geometric distance from each observation x_i to its mean $\bar{x}$ scaled so that each leverage value is between 0 and 1.

On their own, leverages cannot determine which observations are influential, but they are used to determine methods for which they can help determine them

Standardized Residuals: $\hat{z}_i$, the observed residuals scaled by the estimate of the regression standard deviation

$$\hat{z}_i = \frac{\hat{\epsilon}_i}{s_{Y|X}}$$

Studentized Residuals: $\hat{r}_i$, the standardized residual scaled by a factor related to the leverage

$$\hat{r}_i = \frac{\hat{z}_i}{\sqrt{1-h_i}} = \frac{\hat{\epsilon}_i}{s_{Y|X}\sqrt{1-h_i}}$$

Example

Plot the studentized residuals for the blood pressure dataset against the observed values of age as well as the predictions from the least squares regression line $\hat{y}_i$ for $i = 1, \ldots, n$

How do we determine numerically which points are influential observations?

Recall from Chapter 5, we said that

$$\hat{z}_i = \frac{Y_i - \hat{Y}_i}{S_{Y|X}} \sim t_{df=n-2}.$$

Similarly, we can say that

$$\hat{r}_i \sim t_{df=n-2-1}$$

Why is it $n-2-1$ degrees of freedom?

So, the influential points according to studentized residuals are points where $|\hat{r}_i| > t_{1-\frac{\alpha}{2},n-2-1}$ where $t_{1-\frac{\alpha}{2},n-2-1}$ is the $(1-\frac{\alpha}{2})^{th}$ quantile from a t-distribution with $n-2-1$ degrees of freedom.

Example

Let's see if there are any influential observations for the blood pressure dataset with $\alpha = 0.01$.

Cook's Distance: A measure of how much the regression coefficients change when an observation is deleted.

$$d_i = (\hat{\beta}_0 - \hat{\beta}_{0,(-i)})^2 + (\hat{\beta}_1 - \hat{\beta}_{1,(-i)})^2$$

$= r_i \left(\frac{1}{2}\right) \left(\frac{h_i}{1-h_i}\right)$

Note that, Cook's distance may be large because x_i is large relative to its mean (i.e. high leverage h_i) or because the studentized residual, $\hat{r}_i$, is high. Typically, $d_i > 1$ means that observation may warrant additional analysis, but choosing to remove an observation cannot be determined by d_i on its own.