Tutorial 3

YANG YANG

The Australian National University

Week 4, 2017

Overview

Leverage

The diagonal elements of H are a measure of the *influence* of each data point. It turns out that the i^{th} diagonal element of H is:

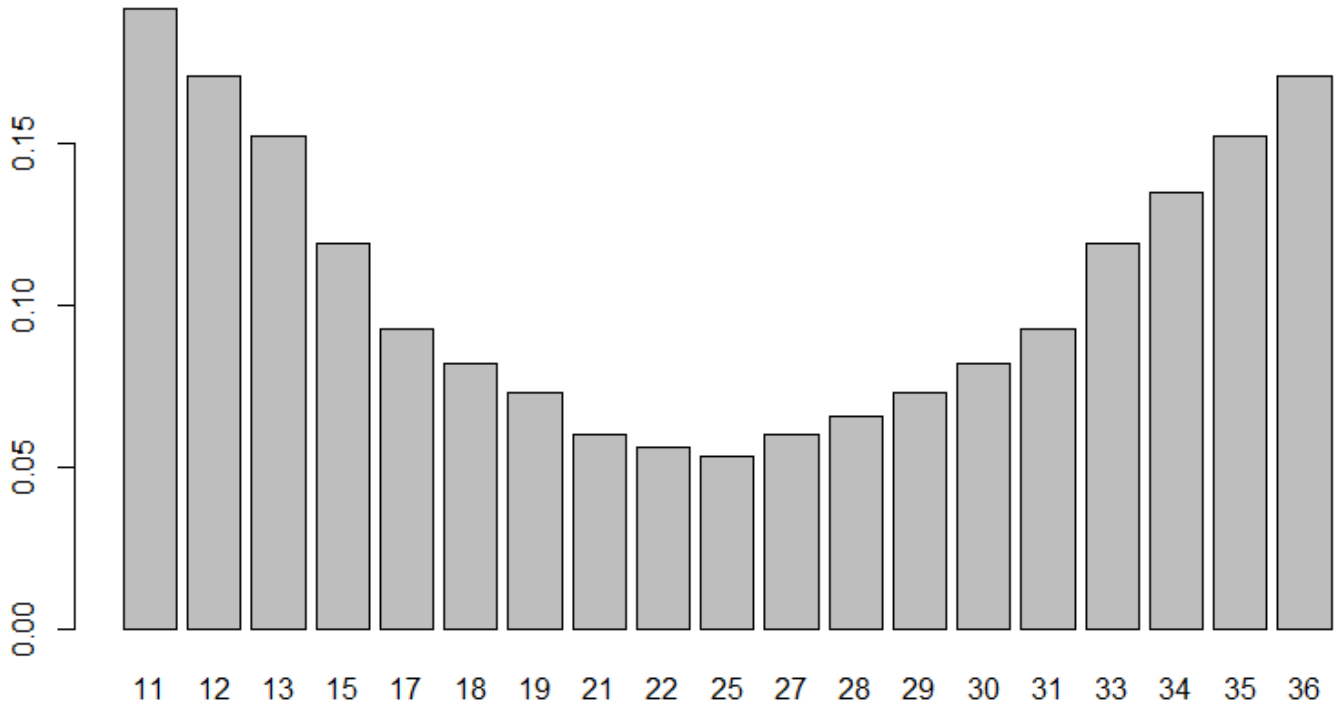
$$h_{ii} = \frac{\sum_{j=1}^{n}(x_j - x_i)^2}{nS_{xx}},$$

and the value h_{ii} is referred to as the *leverage* of the i^{th} data point. The leverage h_{ii} quantifies how far away the i^{th} x value is from the rest of the x values. If the i^{th} x value is far away, the leverage h_{ii} will be large; and otherwise not.

We see that this is a measure of how far from the main body of the data each point is in the horizontal (or predictor) direction.

The leverage h_{ii} is a number between 0 and 1

Leverage Bar Plot



- barplot(hat(X), names.arg $= \cdots$)

Leverage and Influence

A point is highly influential if its removal from the dataset causes a dramatic change in the estimated parameters of the regression line.

One useful way of flagging the potentially influential data points is through the use of the *leverages*, h_{ii}

In order to see whether points with high leverage truly are influential, we examine how the fitted regression line would change once that flagged point is deleted from the data set.

Question 1

(a) Show that the hat matrix, H, and the matrix $I - H$ are projections.

SOLUTION: Since $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, we have

$$H^T = \{X(X^TX)^{-1}X^T\}^T = (X^T)^T\{(X^TX)^{-1}\}^TX^T = X\{(X^TX)^T\}^{-1}X^T = X(X^TX)^{-1}X^T = H$$

$$H^2 = X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T = X(X^TX)^{-1}X^T = H$$

$$(I - H)^T = I^T - H^T = I - H$$

$$(I - H)^2 = (I - H)(I - H) = II - HI - IH + HH = I - H - H + H = I - H$$

Question 4 (a)

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2$$

$$= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y})$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)\hat{Y}_i - \bar{Y}\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} e_i\hat{Y}_i - \bar{Y}\sum_{i=1}^{n} e_i$$

$$= \sum_{i=1}^{n} e_i\hat{Y}_i = \sum_{i=1}^{n} e_i(b_0 + b_1 x_i) = b_0\sum_{i=1}^{n} e_i + b_1\sum_{i=1}^{n} x_i e_i = 0$$

Since $\sum_{i=1}^{n} e_i$ and $\sum_{i=1}^{n} x_i e_i$ are both zero as noted in the hint.

Question 4 (b)

$$
\begin{aligned}
R^2 &= \frac{1}{SS_{Total}} \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = \frac{1}{SS_{Total}} \sum_{i=1}^{n} (b_0 + b_1 x_i - \bar{Y})^2 \\
&= \frac{1}{SS_{Total}} \sum_{i=1}^{n} (\bar{Y} - b_1 \bar{x} + b_1 x_i - \bar{Y})^2 = \frac{1}{SS_{Total}} \sum_{i=1}^{n} b_1^2 (x_i - \bar{x})^2 \\
&= \frac{1}{SS_{Total}} b_1^2 S_{xx} = \frac{S_{xx}}{SS_{Total}} \left(\frac{S_{xy}}{S_{xx}}\right)^2 = \frac{S_{xy}^2}{S_{xx} \cdot SS_{Total}} = \left(\frac{S_{xy}}{\sqrt{S_{xx} \cdot SS_{Total}}}\right)^2 = r^2
\end{aligned}
$$

Question 5

$$
\begin{aligned}
E(MSR) &= E\left\{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2\right\} = E\left\{\sum_{i=1}^{n}(b_0 + b_1 x_i - \bar{Y})^2\right\} = E\left\{\sum_{i=1}^{n}(\bar{Y} - b_1\bar{x} + b_1 x_i - \bar{Y})^2\right\} \\
&= E\left\{\sum_{i=1}^{n}(\bar{Y} - b_1\bar{x} + b_1 x_i - \bar{Y})^2\right\} = E\left\{b_1^2\sum_{i=1}^{n}(x_i - \bar{x})^2\right\} = S_{xx}E(b_1^2) \\
&= S_{xx}[Var(b_1) + \{E(b_1)\}^2] = S_{xx}\left(\frac{\sigma^2}{S_{xx}} + \beta_1^2\right) = \sigma^2 + \beta_1^2 S_{xx}
\end{aligned}
$$

- Here we use a property of variance: $E(x^2) = Var(x) + [E(x)]^2$.