STAT 151A Lecture 25

Henry Liev

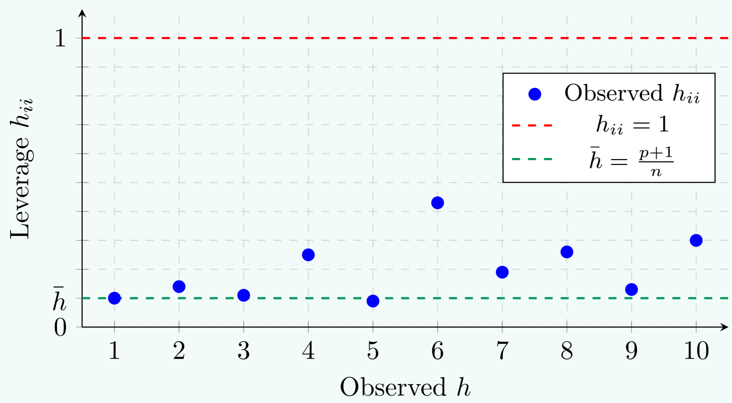
24 October 2025

Remark 0.1 (How to Detect Regression Outliers Quantitatively)

Starting point: residuals $e_1, \ldots, e_n$
Problem: Residuals do not all have the same variance
$y \xrightarrow{\text{under NLM}} \text{Var}(Y|X) = \sigma^2 \mathbb{I}_n$
$\hat{y} = \mathbb{H}y, e = (\mathbb{I} - \mathbb{H})y$
$\text{Var}(\hat{y}) = \sigma^2 \mathbb{H}, \text{Var}(e) = \sigma^2(\mathbb{I} - \mathbb{H}), \text{Var}(e_i) = \hat{\sigma}^2(\mathbb{I} - \mathbb{H})_{j,j} = \hat{\sigma}^2(1 - \mathbb{H}_{j,j})$
Standardized residual
$\tilde{e}_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{j,j}}}, h_{j,j} = (H)_{j,j}$ more comparaible than e_i's
What if we want to test for unusually large e_i or $\tilde{e}_i$
What is the distribution of $\tilde{e}_i$? Hard: numerator and denominator are correlated
Tweak the residuals to make them independent
Studentized residuals $e_i^* = \frac{e_i}{\hat{\sigma}_{(-j)}\sqrt{1-h_{j,j}}} \sim t_{n-p-2}$, where $\hat{\sigma}_{(-j)}$ is the value of $\hat{\sigma}$ from a regression excluding point j
Shortcut: $e_i^* = \tilde{e}_i \sqrt{\frac{n-p-2}{n-p-1-\tilde{e}_i^2}}$
Testing for large studentized residuals
$e_{(1)}^*, \ldots, e_{(n)}^*$ smallest absolute value to largest absolute value
Are any of these values implausibly large? Focus on $e_{\max}^*$ compare to a t_{n-p-2} distribution, two-sided test or one-sided test with absolute values
Correct for multiple testing: Bonferroni correction for n tests $\mathbb{P}_{z \sim t_{n-p-2}}(|Z| \geq |e_{\max}^*|)_j$ compare to $\frac{\alpha}{n}$

Remark 0.2 (Leverage)

How unusual is the x-value at this point?
Number to measure this: h_{ii} = leverage i'th diagonal of the hat matrix
$\mathbb{H} = \mathbf{X}(\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}, \hat{y} = \mathbb{H}y, \hat{y}_i = \hat{h}_{i,*} \cdot \vec{y} = \sum_{j=1}^{n} h_{ij}y_i$
If h_{ij} is large, it tells us that point i palys a big role in determining the fitted value for point j
h_{ii} is a measure of the influence of point i on its own fitted value
Facts about h_{ii}
$\frac{1}{n} \leq h_{ii} \leq 1$ and $\bar{h} = \frac{p+1}{n}$, sample mean of $h_{11}, h_{22}, \ldots, h_{nn}$
SLR: $h_{ii} = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$
Idea generalizes to multiple regression: $\vec{x}$ is a vector mean vector $\bar{x}$
Leverage measures weighted distance of $\vec{x}_i$ from $\bar{x}$ Mahalanobis distance
$\hat{y}_i = \sum_{j=1}^{m} h_{ij}y_i$ and $\mathbb{H}^2 = \mathbb{H} \to h_{ii} = \sum_{i=1}^{n} h_{ii}^2$
This is why $h_{ii} \leq 1$
In practice: Calculate these without $\vec{Y}$
index plot:



Remark 0.3 (Influence)

How much does point j influence the value of $\hat{\beta}$
Basic idea: $\left\|\hat{\beta} - \hat{\beta}_{(-i)}\right\|^2$, but standardize
Instead use Cook's distance: $D_i = \frac{(\hat{\beta}-\hat{\beta}_{(-i)})^\intercal(\mathbf{X}^\intercal\mathbf{X})(\hat{\beta}-\hat{\beta}_{(-i)})}{(p+1)\hat{\sigma}^2}$
Intuition: "divide" by $\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$
More interesting formula $D_i = \frac{(\tilde{e}_i)}{p+1} \cdot \frac{h_{ii}}{1-h_{ii}}$