THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 21
Leverage in Regression

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Leverage During Regression

- During a regression, some data points have more leverage than others
  - Leverage points = data with an extreme value of the predictor variable (x)
- Like outliers, high leverage data can have outsized influence on the regression results
  - We'll define "influence" more exactly later

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Defining Leverage

- Leverage can be quantified by looking at the distances between x's (accounting for correlation)
  - In the matrix formulation of linear regression, this is given by the diagonal elements of the "hat" matrix $\boldsymbol{H}$ (also called the projection matrix): $\hat{\boldsymbol{Y}} = \boldsymbol{HY}$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ (from Lecture 6)

$$h_{ii} \equiv \frac{cov(\hat{y}_i, y_i)}{var(y_i)} \qquad 0 \le h_{ii} \le 1$$

Note that $\sum_{i=1}^{n} h_{ii} = p$ so that $\bar{h} = \frac{p}{n}$ ← Number of parameters in the model

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Defining Leverage

- For the case of only one predictor variable,

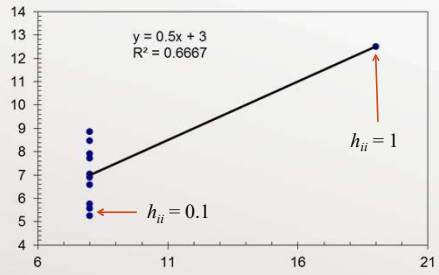
$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{1}{n} + \frac{1}{n-1}\left(\frac{x_i - \bar{x}}{s_x}\right)^2$$

- For two or more predictor variables (multiple regression), we use matrix math to calculate the hat matrix and its diagonal elements

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

From the Anscombe Problems

(recall Lecture 8)



THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Looking at Residuals

- Our model is

$$y = f(\boldsymbol{x}, \boldsymbol{\beta}) + \varepsilon, \quad \varepsilon_i \sim N(0, \sigma_\varepsilon) \text{ iid}$$

- But when we estimate the β_k's with b_k's, the resulting residuals e_i do not have a constant variance
  - High leverage points have lower variance

Residual Variance

- For $\varepsilon \sim N(0, \sigma_\varepsilon)$, our true residuals have a constant variance σ_ε^2, with unbiased estimator

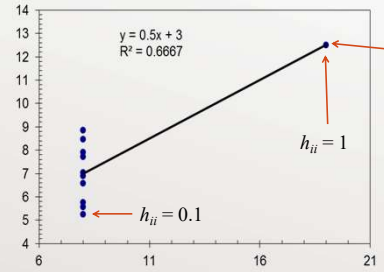
$$s_\varepsilon^2 = \frac{1}{n-p}\sum_{i=1}^{n} \varepsilon_i^2$$

- But the variance of each *fit residual* e_i is

$$var(e_i) = \sigma_\varepsilon^2(1 - h_{ii})$$

$$SE(e_i) = s_e\sqrt{1 - h_{ii}}$$

From the Anscombe Problems



This residual has zero variance!

The fit always passes exactly through this point.

Studentized Residuals

- When doing statistical tests on residuals (Grubbs' test, skewness, etc.) one must studentize the residuals first

$$isr_i = \frac{e_i}{SE(e_i)} = \frac{e_i}{s_e\sqrt{1 - h_{ii}}}$$

Called "internally studentized residual" or "standardized residual"

- Since the SE of the residual varies, tests should always be done on the studentized residuals, not the raw residuals

Studentized Residuals

- Internally Studentized Residual: all data are included in the calculation

This distribution is complicated → $isr_i = \frac{e_i}{SE(e_i)} = \frac{e_i}{s_e\sqrt{1 - h_{ii}}}$

- Externally Studentized Residual: the i^{th} data point is excluded from the calculation of s_e

t-distributed with DF = n – p – 1 [for $\varepsilon \sim N(0, \sigma_\varepsilon)$] → $esr_i = isr_i\sqrt{\frac{n-p-1}{n-p-isr_i^2}}$

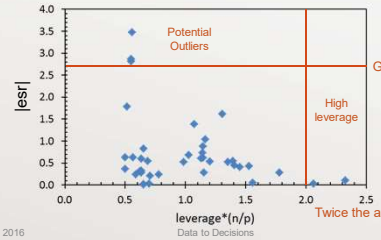
Also called "studentized deleted residual"

Testing the Residuals

- Perform statistical tests on the esr (externally studentized residuals) only
  - QQ Plots
  - Moment testing
  - Outlier detection/rejection (Grubbs, etc.)
  - More to come …
- This is because the esr has a simple sampling distribution: Student's t

Williams Graph

- To see the possibility for both outliers and high-leverage data, plot the externally studentized residual (esr) versus the normalized leverage ($h_{ii}/\bar{h}$) for each data point



Grubbs Critical T

Twice the average leverage

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Lecture 21: What have we learned?

- Define leverage
- What is the role of the hat matrix in determining leverage?
- What is the difference between internally and externally studentized residuals?
- How should residuals be studentized for statistical testing?
- Know how to create and interpret the Williams Graph