Precept 9: Regression Diagnostics

Soc 500: Applied Social Statistics

Simone Zhang[1]

Princeton University

November 2016

[1] Includes material from Matt Blackwell.

Today's Agenda

- Introducing `dplyr` for data cleaning and manipulation
- Studentized residuals
- Non-linearity and generalized additive models
- Identifying extreme values
    - Three types of extreme values
    - Leverage, Cook's distance
- Robust estimation

Split-Apply-Combine[2]

Data analysis using Split-Apply-Combine strategy:

- break up large problem into smaller, more manageable pieces
  - ex: cleaning data, sub-group analysis
- operate on each piece independently
  - ex: summary statistics, model estimation
- put the pieces back togther
  - ex: plotting results, table of aggregate statistics,

`dplyr` and `ggplot()` are both based around the split-apply-combine concept.

[2]Wickham, Hadley. "The split-apply-combine strategy for data analysis." Journal of Statistical Software 40.1 (2011): 1-29.

dplyr Cheat Sheet

dplyr cheatsheet: `https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf`

Learning about distribution of errors through residuals

- Assumption is about **unobserved** $\mathbf{u} = \mathbf{y} - \mathbf{X}\beta$
- We can only **observe** residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\beta}$
- If **distribution of residuals** $\approx$ **distribution of errors**, we could check residuals
- But this is actually **not true**—the distribution of the residuals is complicated

To understand the relationship between residuals and errors, we need to derive the distribution of the residuals.

Hat matrix

- Define matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$

$$
\begin{aligned}
\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\beta} \\
&= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
&\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{y}
\end{aligned}
$$

- $\mathbf{H}$ is the **hat matrix** because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

  - $\mathbf{H}$ is an $n \times n$ symmetric matrix

Relating the residuals to the errors

$$\begin{aligned}\widehat{\mathbf{u}} &= (\mathbf{I}-\mathbf{H})(y)\\ &= (\mathbf{I}-\mathbf{H})(\mathbf{X}\beta+\mathbf{u})\\ &= (\mathbf{I}-\mathbf{H})\mathbf{X}\beta+(\mathbf{I}-\mathbf{H})\mathbf{u}\\ &= \mathbf{I}\mathbf{X}\beta-\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta+(\mathbf{I}-\mathbf{H})\mathbf{u}\\ &= \mathbf{X}\beta-\mathbf{X}\beta+(\mathbf{I}-\mathbf{H})\mathbf{u}\\ &= (\mathbf{I}-\mathbf{H})\mathbf{u}\end{aligned}$$

- Residuals $\widehat{\mathbf{u}}$ are a linear function of the errors, $\mathbf{u}$
- For instance,

$$\widehat{u}_1 = (1-h_{11})u_1 - \sum_{i=2}^{n} h_{1i}u_i$$

- Note that the residual is a function of all of the errors

Distribution of the residuals

$$\mathbb{E}[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})\mathbb{E}[\mathbf{u}] = \mathbf{0}$$

$$\text{Var}[\hat{\mathbf{u}}] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

The variance of the ith residual $\hat{u}_i$ is $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, where h_{ii} is the ith diagonal element of the matrix $\mathbf{H}$ (called the **hat value**).

Distribution of the Residuals

Notice in contrast to the unobserved errors, the estimated residuals

1. are not independent (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance. The variance of the residuals varies across data points $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

These properties can obscure the true patterns in the error distribution, and thus are inconvenient for our diagnostics.

Standardized Residuals

Let's address the second problem (unequal variances) by standardizing $\hat{u}_i$, i.e., dividing by their estimated standard deviations.

This produces **standardized** (or "internally studentized") **residuals**:

$$\hat{u}'_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}^2$ is our usual estimate of the error variance.
The standardized residuals are still not ideal, since the numerator and denominator of $\hat{u}'_i$ are not independent. This makes the distribution of $\hat{u}'_i$ nonstandard.

Studentized residuals

If we remove observation i from the estimation of σ, then we can eliminate the dependence and the result will have a standard distribution.

- estimate residual variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\mathbf{u}'\mathbf{u} - u_i^2/(1 - h_{ii})}{n - k - 2}$$

- Use this i-free estimate to standardize, which creates the **studentized residuals**:

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1 - h_{ii}}}$$

- If the errors are Normal, the studentized residuals follow a t distribution with $(n - k - 2)$ degrees of freedom.
- Deviations from $t \implies$ violation of Normality

Generalized Additive Models (GAM)

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

For GAMs, we maintain additivity, but instead of imposing linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot),$ and $s_3(\cdot)$ are smooth functions that are estimated from the data:

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

Generalized Additive Models (GAM)

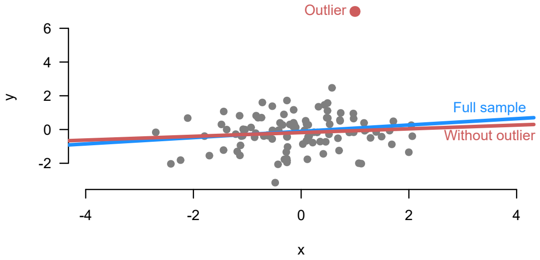
$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- They do NOT give you a set of regression parameters $\hat{\beta}$. Instead one obtains a graphical summary of how $E[Y|X_{,}X_2,...,X_k]$ varies with

Three types of extreme values

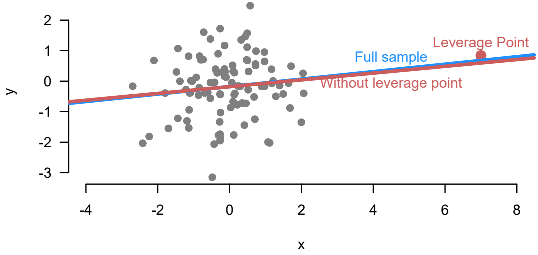
1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

Outlier definition



- An **outlier** is a data point with very large regression errors, u_i
- Very **distant** from the rest of the data **in the y-dimension**
- Increases standard errors (by increasing $\widehat{\sigma}^2$)
- No bias if typical in the x's

Leverage point definition



- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution
- Decrease SEs (more X variation)
- No bias if typical in y dimension

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix **H**:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

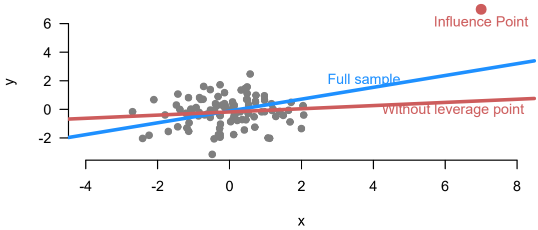
H is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

Therefore,

- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since **H** depends only on **X**)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the **hat values** or **leverages** and a single subscript notation is used: $h_i = h_{ii}$
- Intuitively, the hat values measure how far a unit's vector of characteristics $\mathbf{x}_i$ is from the vector of means of **X**
- **Rule of thumb**: examine hat values greater than $2(k+1)/n$

Influence points



- An **influence point** is one that is both an **outlier** (extreme in X) and a **leverage point** (extreme in Y).
- Causes the regression line to move toward it (bias?)

Detecting Influence Points/Bad Leverage Points

- **Influence Points**:
  Influence on coefficients = Leverage × Outlyingness

- More formally: Measure the change that occurs in the slope estimates when an observation is removed from the data set. Let

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \quad i = 1, \ldots, n, \;\; j = 0, \ldots, k$$

  where $\hat{\beta}_{j(-i)}$ is the estimate of the jth coefficient from the same regression once observation i has been removed from the data set.

- D_{ij} is called the **DFbeta**, which measures the **influence** of observation i on the estimated coefficient for the jth explanatory variable.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^* = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^* is called **DFbetaS**.

- $D_{ij}^* > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.
- $D_{ij}^* < 0$ implies that removing observation i increases the estimate of $\beta_j \rightarrow$ obs i has a negative influence on β_j.
- Values of $|D_{ij}^*| > 2/\sqrt{n}$ are an indication of high influence.
- In R: `dfbetas(model)`

Summarizing Influence across All Coefficients

- Leverage tells us how much one data point affects a **single coefficient**.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is **Cook's distance**:

$$D_i = \frac{\hat{u}_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

  where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.
  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i
  - In R, `cooks.distance(model)`
  - $D > 4/(n-k-1)$ is commonly considered large
- The **influence plot**: the studentized residuals plotted against the hat values, size of points proportional to Cook's distance.

Questions?