Diagnostics & Remedial Measures for SLR (Ch 3)

Week 6 – Tuesday

Applied Regression Analysis (STAT 757)

Paul J. Hurtado

23 Feb, 2016
Checking Assumptions

Remember: Estimates, confidence intervals, p-values, etc. are all meaningless if you’re using the wrong model!

Diagnostics help identify violations of your model assumptions.

SLR Model Assumptions:

1. All data follow $Y|X = x_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$, hence $E(Y|X = x_i) = \beta_0 + \beta_1 x_i$
2. Normal errors: $e_i \sim N(0, \sigma)$
3. Independent errors $e_i$
4. $\text{Var}(Y|X = x_i) = \text{Var}(e_i) = \sigma^2$

Many problems lead to outliers and high leverage points.
Residuals $\hat{e}_i = y_i - \hat{y}_i \approx e_i$
Standardized Residuals

Recall that $e_i \sim N(0, \sigma)$, which means that standardizing $z_i = e_i/\sigma$ (by dividing by the standard deviation) would yield values that follow a Normal(0,1) distribution (if we knew $\sigma$!):
Leverage & “Hat” values \((h_{ij})\)

Observe that

\[
\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j
\]

where

\[
h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2},
\]

We call \(h_{ii}\) the leverage of the \(i\)th data point.
Leverage & “Hat” values \((h_{ij})\)

Observe that

\[
\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j
\]

where

\[
h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}, \quad \sum_{j=1}^{n} h_{ij} = 1,
\]

We call \(h_{ii}\) the leverage of the \(i\)th data point. Note \(h_{ii} = \frac{2}{n}\). A high leverage point is 2x that mean: 

\(h_{ii} > \frac{4}{n}\).
Leverage & “Hat” values \((h_{ij})\)

Observe that

\[
\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j
\]

where

\[
h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n} (x_k - \bar{x})^2}, \quad \sum_{j=1}^{n} h_{ij} = 1, \quad \text{and} \quad \sum_{i=1}^{n} h_{ii} = 2
\]
Leverage & “Hat” values \((h_{ij})\)

Observe that 

\[
\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j
\]

where 

\[
h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{k=1}^{n}(x_k - \bar{x})^2}, \quad \sum_{j=1}^{n} h_{ij} = 1, \quad \text{and} \quad \sum_{i=1}^{n} h_{ii} = 2
\]

We call \(h_{ii}\) the leverage of the \(i^{th}\) data point.

Note \(\bar{h}_{ii} = \frac{2}{n}\). A high leverage point is 2x that mean: \(h_{ii} > \frac{4}{n}\).
“Hat” values \((h_{ij})\)

**Side Note:** These “hat” values form a matrix \(H\) which gives

\[
\hat{y} = H y
\]

and these values show up in many places!

- \(\text{Var}(\hat{y}_i) = \sigma^2 h_{ii}\)
- Alternative definition: \(h_{ij} = \frac{\text{cov}(\hat{y}_i, y_j)}{\text{var}(y_j)}\)
- Residuals, in matrix notation: \(r = (I - H)y\)
- Properties: \(H\) is symmetric, \(H^2 = H\), \(HX = X\)
- Similar \(H\) matrices for other models may not have all these properties.

Want more? See online resources and publications such as

_Hoaglin and Welsch. 1978. The Hat Matrix in Regression and ANOVA._

Standardized Residuals

Recall $e_i / \sigma \sim \text{Normal}(0, 1)$, **BUT** we don’t know $\sigma$!

Using our estimate, $S$, in it’s place (and some algebra to show that $\text{Var}(\hat{e}_i) = \sigma^2 (1 - h_{ii})$) yields **standardized residuals** $r_i$:

$$r_i = \frac{\hat{e}_i}{S \sqrt{1 - h_{ii}}}$$

These can be more informative to look at than residual plots, especially if high leverage points exist.
Normal Quantile-Quantile Plots

In place of a **Shapiro-Wilk** test, plot Standardized Residuals versus the Expected Values of the Order Statistics for a Normal(0,1) distribution. See `shapiro.test()` & `qqnorm()`.

```r
qqnorm(fit1$residuals)
```
Leave-one-out Diagnostics

Another approach to identifying problem data points (with problematic influence) is to compare estimates with and without them. For example, if $\hat{y}_{j(i)}$ is the estimate of $\hat{y}_j$ with the $j^{th}$ data point removed...

Cook’s Distance:

$$D_i = \frac{\sum_{i=1}^{n}(\hat{y}_{j(i)} - \hat{y}_j)^2}{2S^2} = \ldots = \frac{r_i^2}{2} \frac{h_{ii}}{1 - h_{ii}}$$

Roughly speaking, scrutinize points with $D_i > \frac{4}{n-2}$ or values that deviate markedly from the other distances.
Summary Remark

“Bad” leverage points are **high leverage** points that are also **outliers** – they signal a problem with your model!

The two main approaches to fixing that problem:

1. **Omit the data point** from the data set, or
2. **Redo your analysis using a more appropriate model.**
   This is often the preferred approach.