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!

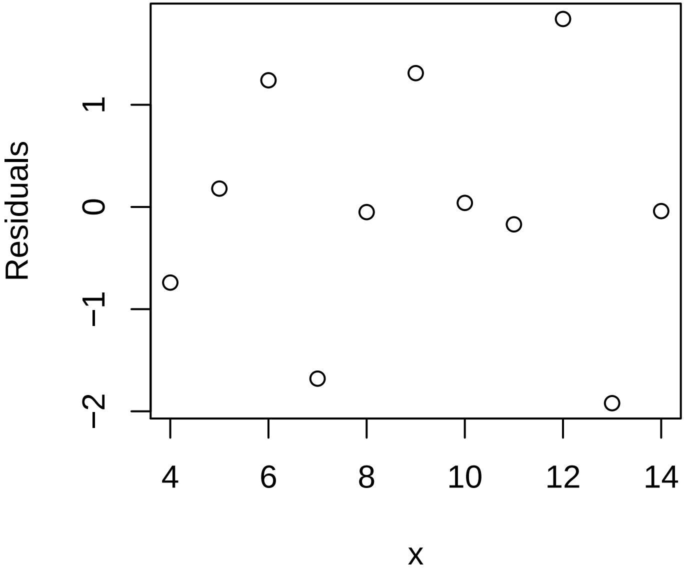
Diagnostics help identify violations of your model assumptions.

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. $Var(Y|X = x_i) = Var(e_i) = \sigma^2$

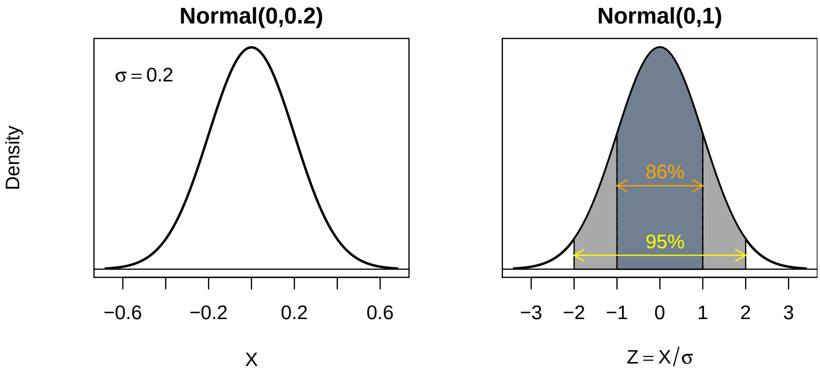
Many problems lead to **outliers** and **high leverage** points.

Residuals $\widehat{e}_i = y_i - \widehat{y}_i \approx e_i$



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 σ!):



Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

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

Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

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

Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

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

Leverage & "Hat" values (h_{ij})

Observe that

$$\widehat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

where

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

We call h_{ii} the **leverage** of the i^{th} data point.

Note $\overline{h_{ii}} = \frac{2}{n}$. 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

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

and these values show up in many places!

- $\text{Var}(\widehat{y_i}) = \sigma^2 h_{ii}$
- Alternative definition: $h_{ij} = \frac{cov(\widehat{y_i}, y_j)}{var(y_j)}$
- Residuals, in matrix notation: $\mathbf{r} = (I - \mathbf{H})\mathbf{y}$
- Properties: $\mathbf{H}$ is symmetric, $\mathbf{H}^2 = \mathbf{H}$, $\mathbf{HX} = \mathbf{X}$
- Similar $\mathbf{H}$ 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.* http://www.stat.ucla.edu/~cocteau/stat201b/handout/hat.pdf

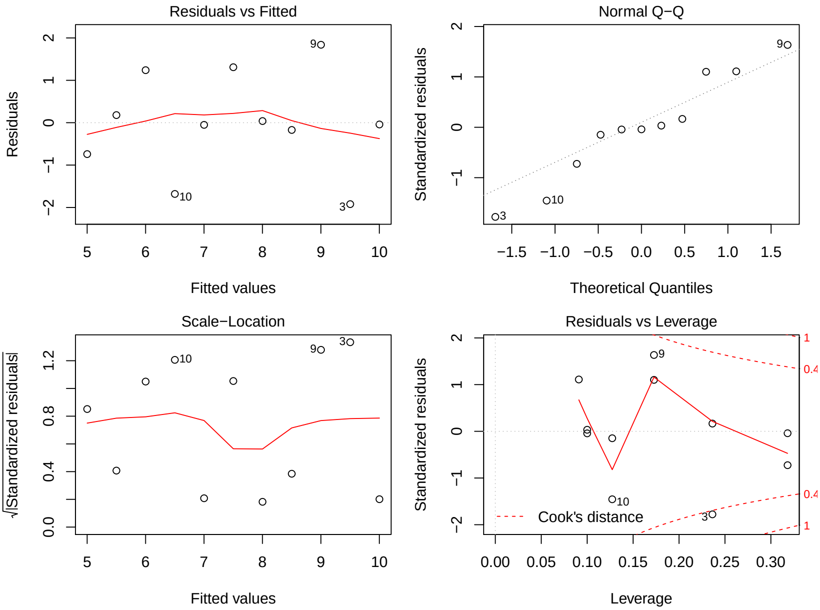
Standardized Residuals

Recall $e_i/\sigma \sim$ Normal(0,1), **BUT** we don't know σ!

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

$$r_i = \frac{\widehat{e_i}}{S\sqrt{1 - h_{ii}}}$$

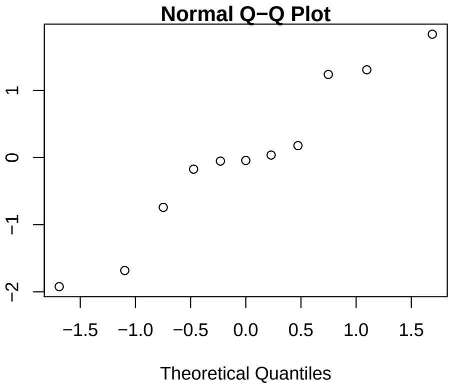
These can be more informative to look at than residual plots, especially if high leverage points exist.

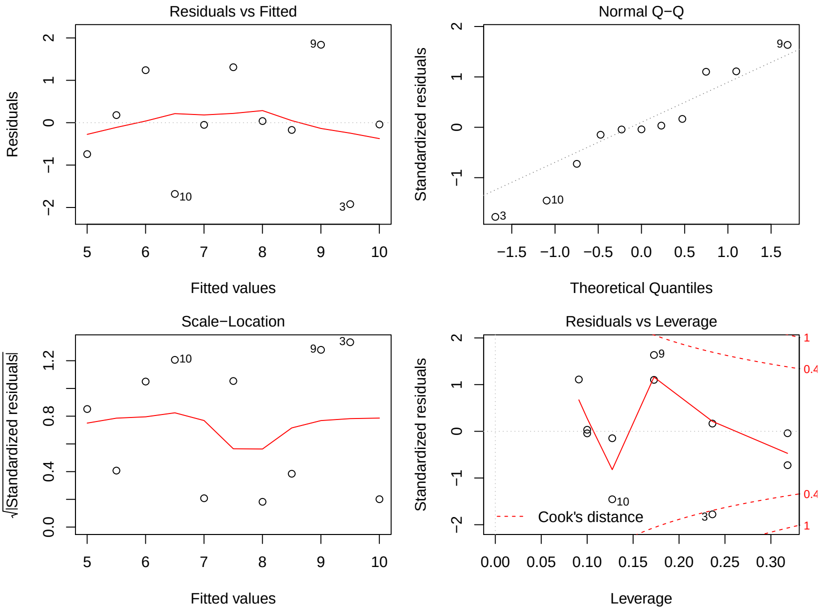

Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage

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()`.

```
qqnorm(fit1$residuals)
```




Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
0.4
1

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 $\widehat{y_{j(i)}}$ is the estimate of $\widehat{y}_j$ with the j^{th} data point removed...

Cook's Distance:

$$D_i = \frac{\sum_{i=1}^{n}(\widehat{y_{j(i)}} - \widehat{y}_j)^2}{2S^2} = \cdots = \frac{r_i^2}{2}\frac{h_{ii}}{1-h_{ii}}$$

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.