Checking Assumptions

Merlise Clyde

STA721 Linear Models

Duke University

November 10, 2014

Outline

Topics

- Normality
- Brain Weights and Body Mass
- Box-Cox

Readings: Christensen Chapter 13

Linear Model

Linear Model:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$$

Assumptions:

$$\begin{aligned} \boldsymbol{\mu} \in C(\mathbf{X}) \quad &\Leftrightarrow \quad \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} \\ \boldsymbol{\epsilon} \quad &\sim \quad \mathsf{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n) \end{aligned}$$

What could go wrong?

- Wrong mean for a case or cases (outliers)
- Cases that influence the mean (influential cases)
- Both of the above
- Wrong distribution for $\boldsymbol{\epsilon}$

Normality

Recall

$$\begin{aligned} \mathbf{e} &= (\mathbf{I} - \mathsf{P}_{\mathbf{X}})\mathbf{Y} \\ &= (\mathbf{I} - \mathsf{P}_{\mathbf{X}})(\mathbf{X}\hat{\boldsymbol{\beta}} + \boldsymbol{\epsilon}) \\ &= (\mathbf{I} - \mathsf{P}_{\mathbf{X}})\boldsymbol{\epsilon} \end{aligned}$$

$$e_i = \epsilon_i - \sum_{j=1}^{n} h_{ij}\epsilon_j$$

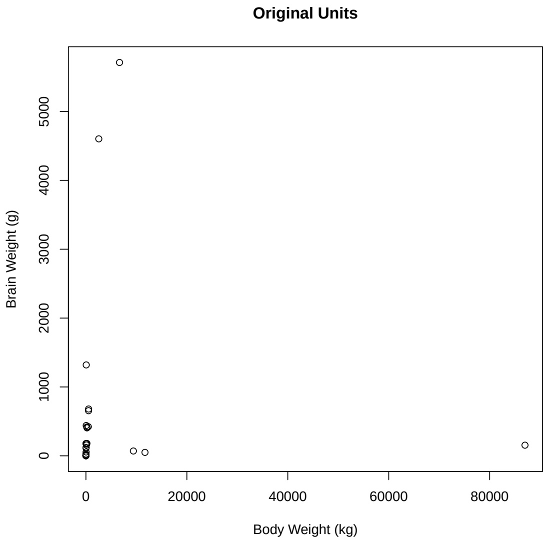
Lyapunov CLT implies that residuals will be approximately normal (even for modest n), if the errors are not normal

"Supernormality of residuals"

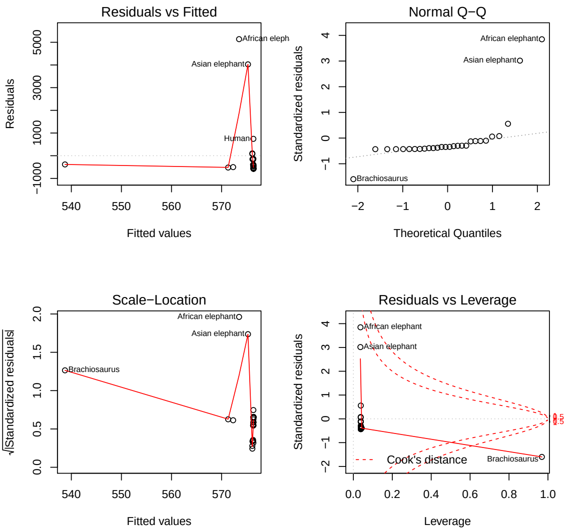
Q-Q Plots

- Order e_i: $e_{(1)} \leq e_{(2)} \ldots \leq e_{(n)}$ sample order statistics or sample quantiles
- Let $z_{(1)} \leq z_{(2)} \ldots z_{(n)}$ denote the expected order statistics of a sample of size n from a standard normal distribution "theoretical quantiles"
- If the e_i are normal then $\mathrm{E}[e_{(i)}] = \sigma z_{(i)}$
- Expect that points in a scatter plot of $e_{(i)}$ and $z_{(i)}$ should be on a straight line.
- Judgment call - use simulations to gain experience!

Animal Example



Residual Plots



Box-Cox Transformation

Box and Cox (1964) suggested a family of power transformations for $Y > 0$

$$U(\mathbf{Y}, \lambda) = Y^{(\lambda)} = \begin{cases} \frac{(Y^{\lambda} - 1)}{\lambda} & \lambda \neq 0 \\ \log(Y) & \lambda = 0 \end{cases}$$

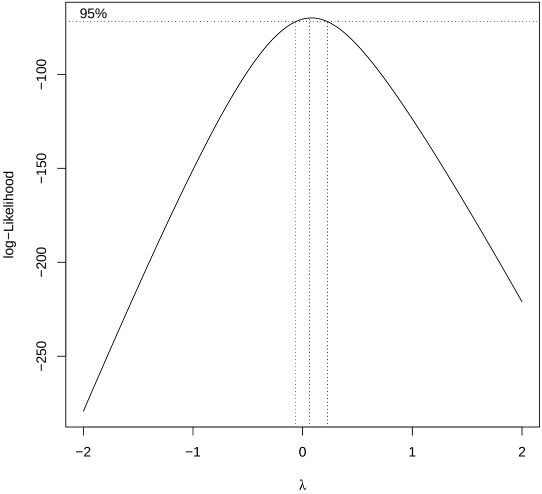
- Estimate λ by maximum Likelihood

$$\mathcal{L}(\lambda, \boldsymbol{\beta}, \sigma^2) \propto \prod f(y_i \mid \lambda, \boldsymbol{\beta}, \sigma^2)$$

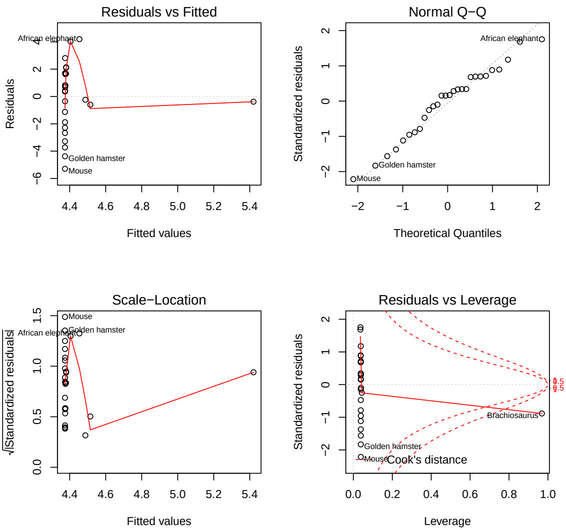
- $U(\mathbf{Y}, \lambda) = Y^{(\lambda)} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$
- Jacobian term is $\prod_i y_i^{\lambda - 1}$ for all λ
- Profile Likelihood based on substituting MLE $\boldsymbol{\beta}$ and σ^2 for each value of λ is

$$\log(\mathcal{L}(\lambda) \propto (\lambda - 1) \sum_i \log(Y_i) - \frac{n}{2} \log(\mathrm{SSE}(\lambda))$$

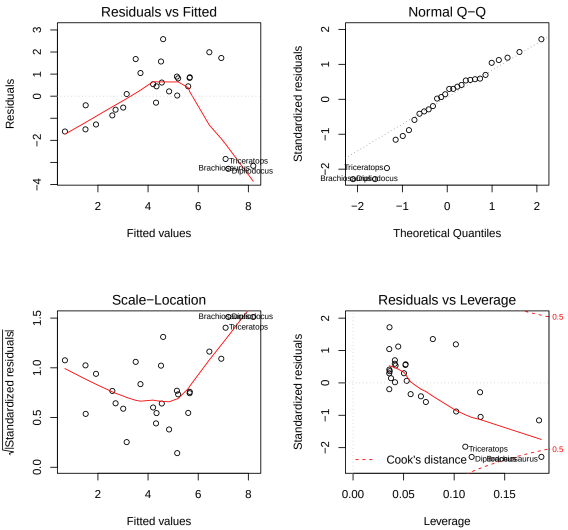
Profile Likelihood



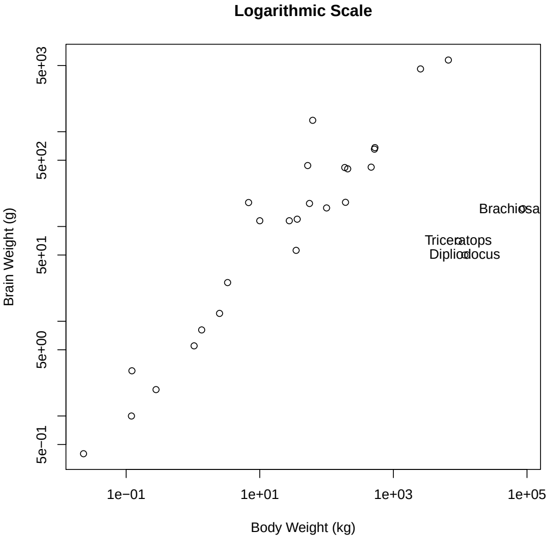
Residuals After Transformation of Response



Residuals After Transformation of Both



Transformed Data



Test that Dinos are Outliers

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	23	12.12				
2	26	60.99	-3	-48.87	30.92	0.0000

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.1504	0.2006	10.72	0.0000
log(body)	0.7523	0.0457	16.45	0.0000
Triceratops	-4.7839	0.7913	-6.05	0.0000
Brachiosaurus	-5.6662	0.8328	-6.80	0.0000
Dipliodocus	-5.2851	0.7949	-6.65	0.0000

Dinosaurs come from a different population from mammals

To Remove or Not?

- For suspicious cases, check data sources for errors
- Check that points are not outliers because of wrong mean function or distributional assumptions
- Investigate need for transformations (use EDA at several stages)
- Influential cases - report results with and without cases (results may change - are differences meaningful?)
- Outlier test - suggests alternative population for the case(s); if not influential may in keep analysis, but will inflate $\hat{\sigma}^2$ and interval estimates
- Document how you handle any case deletions - reproducibility!
- Consider BMA with outliers (See BMA package) to address model uncertainty
- Robust Regression Methods

Variance Stabilizing Transformations

- If $Y - \mu$ (approximately) $N(0, h(\mu))$
- Delta Method implies that

$$g(Y) \dot{\sim} \mathsf{N}(g(\mu), g'(\mu)^2 h(\mu)$$