Transformations

Merlise Clyde

Readings: Gelman & Hill Ch 2-4, ALR 8-9

Assumptions of Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$$

- Model Linear in X_j but X_j could be a transformation of the original variables

Assumptions of Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$$

- Model Linear in X_j but X_j could be a transformation of the original variables
- $\epsilon_i \sim N(0, \sigma^2)$

Assumptions of Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$$

- Model Linear in X_j but X_j could be a transformation of the original variables
- $\epsilon_i \sim N(0, \sigma^2)$
- $Y_i \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip}, \sigma^2)$

Assumptions of Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$$

- Model Linear in X_j but X_j could be a transformation of the original variables
- $\epsilon_i \sim N(0, \sigma^2)$
- $Y_i \overset{\text{ind}}{\sim} \text{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip}, \sigma^2)$
    - correct mean function

Assumptions of Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$$

- Model Linear in X_j but X_j could be a transformation of the original variables
- $\epsilon_i \sim N(0, \sigma^2)$
- $Y_i \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip}, \sigma^2)$
    - correct mean function
    - constant variance

Assumptions of Linear Regression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip} + \epsilon_i$$

- Model Linear in X_j but X_j could be a transformation of the original variables
- $\epsilon_i \sim N(0, \sigma^2)$
- $Y_i \overset{\text{ind}}{\sim} \mathsf{N}(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots \beta_p X_{ip}, \sigma^2)$
  - correct mean function
  - constant variance
  - independent Normal errors

Animals

Read in Animal data from MASS. The data set contains measurements on body weight (kg) and brain weight (g) on 28 animals.

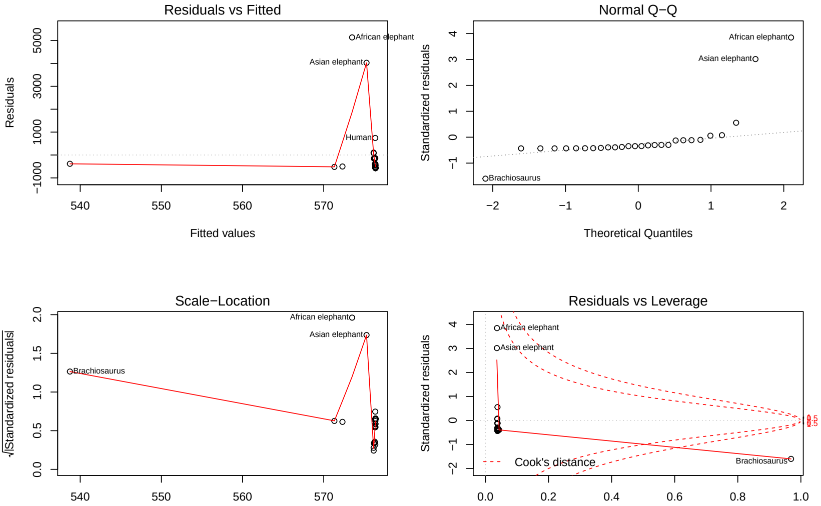
Let's try to predict brain weight from body weight.

```r
library(MASS)
data(Animals)
brain.lm = lm(brain ~ body, data=Animals)
```

Diagnostic Plots

```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



Outliers

- Issues with multiple comparisons if we compare each p-value to $\alpha = 0.05$

Outliers

- Issues with multiple comparisons if we compare each p-value to $\alpha = 0.05$
- Bonferroni compares p-values to α/n

Outliers

- Issues with multiple comparisons if we compare each p-value to $\alpha = 0.05$
- Bonferroni compares p-values to α/n
- Other corrections for False Discoveries possible (STA 532)

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true
- The **familywise error rate** (FWER) is the probability of rejecting at least one true H_i (making at least one type I error).

$$\begin{aligned} \text{FWER} &= P\left\{\bigcup_{i=1}^{n_0}\left(p_i \le \frac{\alpha}{n}\right)\right\} \le \sum_{i=1}^{n_0}\left\{P\left(p_i \le \frac{\alpha}{n}\right)\right\} \le n_0\frac{\alpha}{n} \le n\frac{\alpha}{n} \\ &= \alpha \end{aligned}$$

Bonferonni Correction & Multiple Testing

- $H_1, \ldots, H_n$ are a family of hypotheses and $p_1, \ldots, p_n$ their corresponding p-values
- n_0 of the n are true
- The **familywise error rate** (FWER) is the probability of rejecting at least one true H_i (making at least one type I error).

$$\begin{aligned} \text{FWER} &= P\left\{\bigcup_{i=1}^{n_0}\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq \sum_{i=1}^{n_0}\left\{P\left(p_i \leq \frac{\alpha}{n}\right)\right\} \leq n_0\frac{\alpha}{n} \leq n\frac{\alpha}{n} \\ &= \alpha \end{aligned}$$

- This does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true.

Bonferonni Correction & Multiple Testing

- $H_1, \dots, H_n$ are a family of hypotheses and $p_1, \dots, p_n$ their corresponding p-values
- n_0 of the n are true
- The **familywise error rate** (FWER) is the probability of rejecting at least one true H_i (making at least one type I error).

$$\begin{aligned}\text{FWER} &= P\left\{\bigcup_{i=1}^{n_0}\left(p_i \le \frac{\alpha}{n}\right)\right\} \le \sum_{i=1}^{n_0}\left\{P\left(p_i \le \frac{\alpha}{n}\right)\right\} \le n_0\frac{\alpha}{n} \le n\frac{\alpha}{n} \\ &= \alpha\end{aligned}$$

- This does not require any assumptions about dependence among the p-values or about how many of the null hypotheses are true.
- Link https://en.wikipedia.org/wiki/Bonferroni_correction

Outliers

- Flag outliers after Bonferroni Correction $p_i < \alpha/n$

```
pval = 2*(1 - pt(abs(rstudent(brain.lm)), brain.lm$df -1))
rownames(Animals)[pval <  .05/nrow(Animals)]
```

```
## [1] "Asian elephant"   "African elephant"
```

Cook's Distance

Measure of influence of case i on predictions

$$D_i = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{(i)}\|^2}{\hat{\sigma}^2\, p}$$

after removing the ith case

Cook's Distance

Measure of influence of case i on predictions

$$D_i = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{(i)}\|^2}{\hat{\sigma}^2\, p}$$

after removing the ith case

Easier way to calculate

$$D_i = \frac{e_i^2}{\hat{\sigma}^2 p} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right],$$

Cook's Distance

Measure of influence of case i on predictions

$$D_i = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{(i)}\|^2}{\hat{\sigma}^2\, p}$$

after removing the ith case

Easier way to calculate

$$D_i = \frac{e_i^2}{\hat{\sigma}^2 p}\left[\frac{h_{ii}}{(1 - h_{ii})^2}\right],$$

$$D_i = \frac{r_{ii}}{p}\frac{h_{ii}}{1 - h_{ii}}$$

Cook's Distance

Measure of influence of case i on predictions

$$D_i = \frac{\|\mathbf{Y} - \hat{\mathbf{Y}}_{(i)}\|^2}{\hat{\sigma}^2 \, p}$$

after removing the ith case

Easier way to calculate

$$D_i = \frac{e_i^2}{\hat{\sigma}^2 p} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right],$$

$$D_i = \frac{r_{ii}}{p} \frac{h_{ii}}{1 - h_{ii}}$$

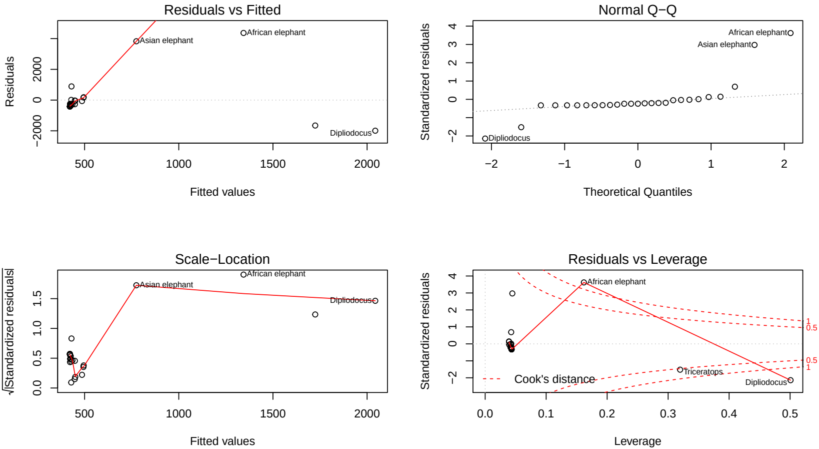
Flag cases where $D_i > 1$ or $D_i > 4/n$

```r
rownames(Animals)[cooks.distance(brain.lm) > 1]
```

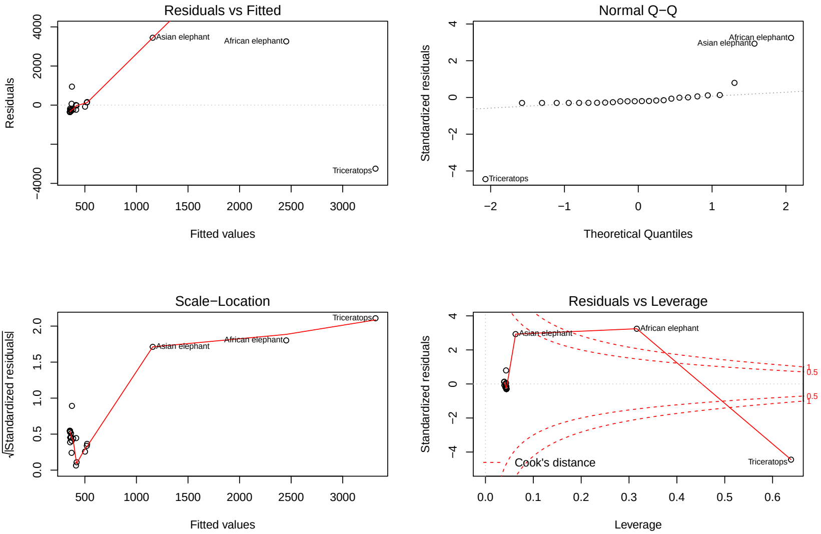
```
## [1] "Brachiosaurus"
```

Remove Influential Point & Refit

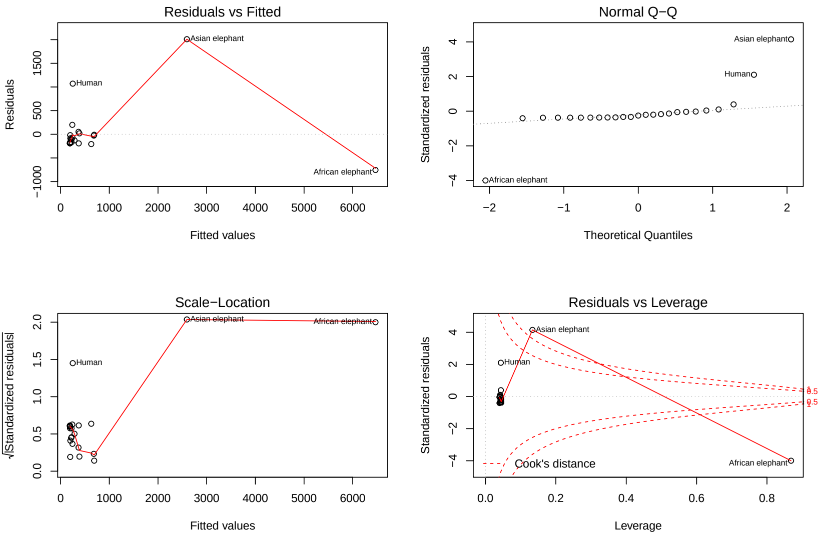
```
brain2.lm = lm(brain ~ body, data=Animals,
                 subset = !cooks.distance(brain.lm)>1)
par(mfrow=c(2,2)); plot(brain2.lm)
```



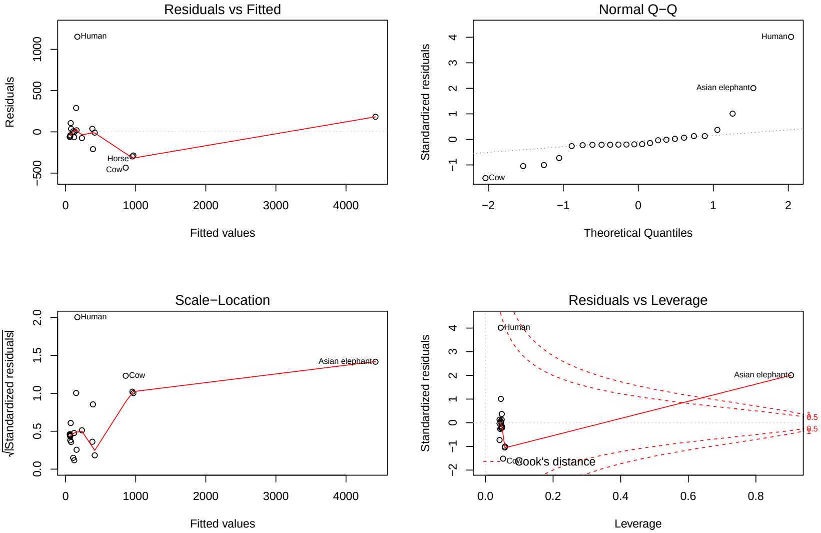
Keep removing points?



And another one bites the dust



and another one

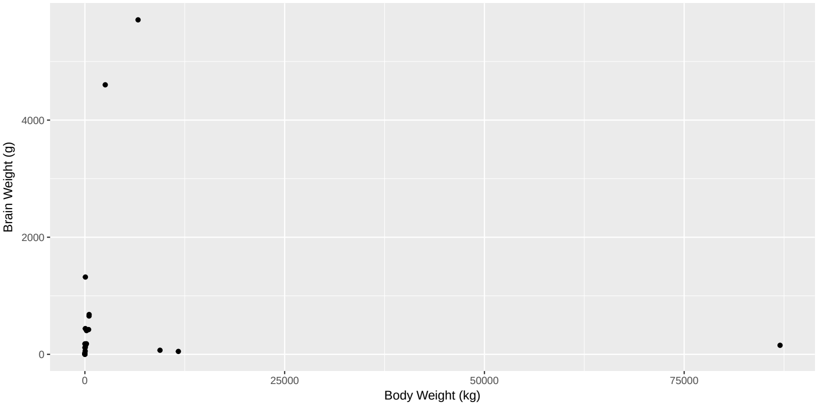


And they just keep coming!

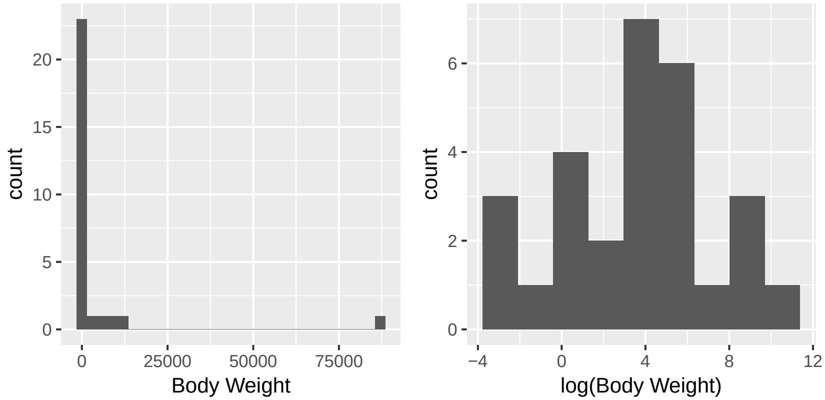
Figure 1: Walt Disney Fantasia

Plot of Original Data (what you should always do first!)

```
library(ggplot2)
ggplot(Animals, aes(x=body, y=brain)) +
  geom_point() +
  xlab("Body Weight (kg)") + ylab("Brain Weight (g)")
```

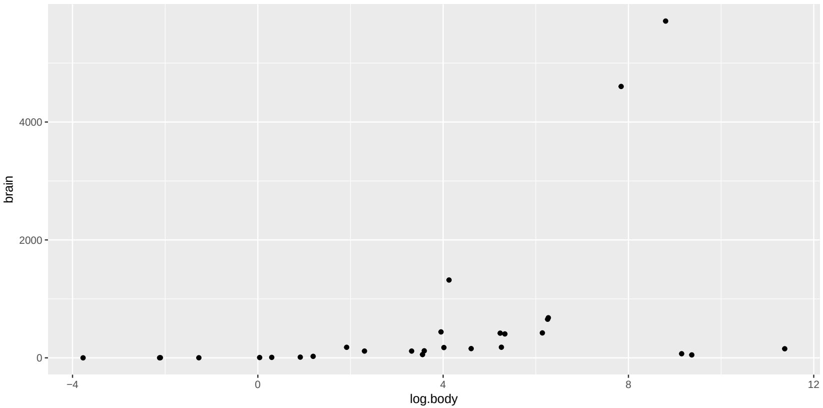


Log Transform



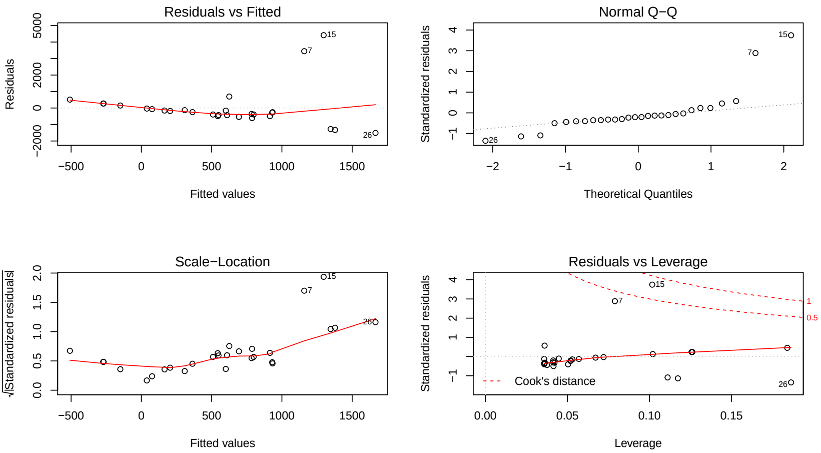
Plot of Transformed Data

```
Animals= mutate(Animals, log.body = log(body))
ggplot(Animals, aes(log.body, brain)) + geom_point()
```



```
#plot(brain ~ body, Animals, log="x")
```

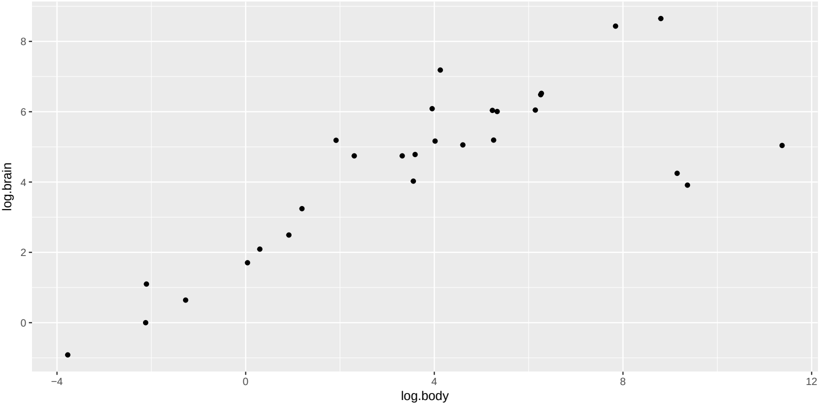
Diagnostics with log(body)



Variance increasing with mean

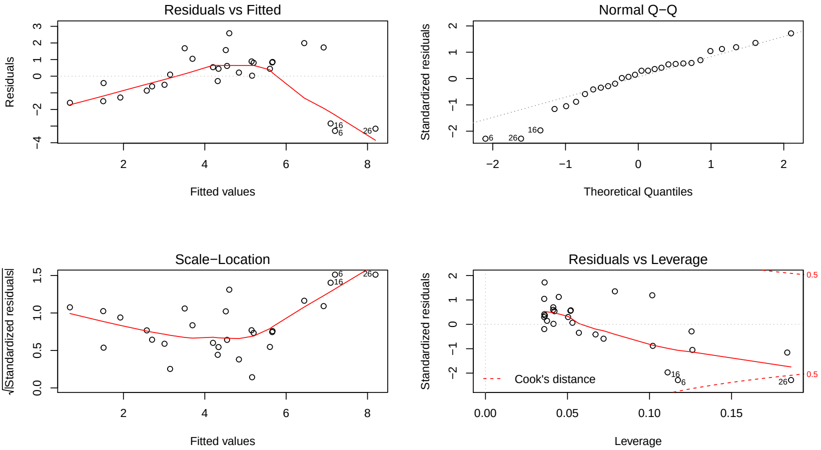
Try Log-Log

```
Animals= mutate(Animals, log.brain= log(brain))
ggplot(Animals, aes(log.body, log.brain)) + geom_point()
```



```
#plot(brain ~ body, Animals, log="xy")
```

Diagnostics with log(body) & log(brain)



Optimal Transformation for Normality

The BoxCox procedure can be used to find "best" power transformation λ of Y (for positive Y) for a given set of transformed predictors.

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

Find value of λ that maximizes the likelihood derived from $\Psi(\mathbf{Y}, \lambda) \sim \mathrm{N}(\mathbf{X}\beta_{\lambda}, \sigma^2_{\lambda})$ (need to obtain distribution of **Y** first)

Find λ to minimize

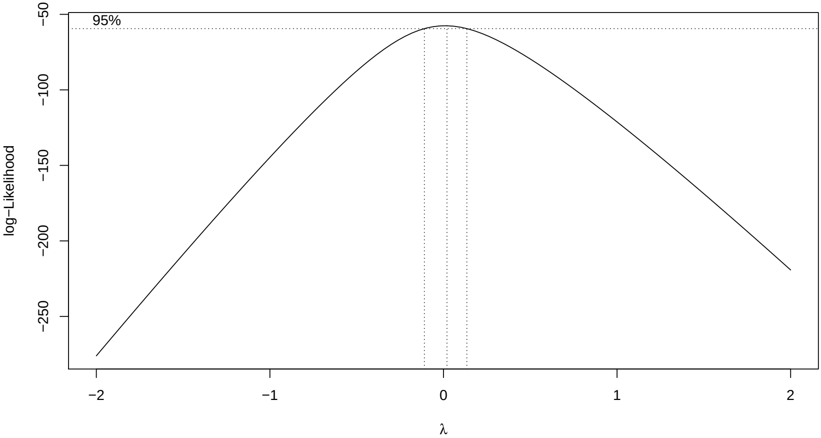
$$\mathrm{RSS}(\lambda) = \|\Psi_M(\mathbf{Y}, \lambda) - \mathbf{X}\hat{\beta}_{\lambda}\|^2$$

$$\Psi_M(\mathbf{Y}, \lambda) = \begin{cases} (\mathrm{GM}(\mathbf{Y})^{1-\lambda}(\mathbf{Y}^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \mathrm{GM}(\mathbf{Y}) \log(\mathbf{Y}) & \text{if } \lambda = 0 \end{cases}$$

where $\mathrm{GM}(\mathbf{Y}) = \exp(\sum \log(Y_i)/n)$ (Geometric mean)

boxcox in R: Profile likelihood

```r
library(MASS)
boxcox(braintransX.lm)
```



Caveats

- Boxcox transformation depends on choice of transformations of X's

Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)`

Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)`
- transformations of X's can reduce leverage values (potential influence)

Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)`
- transformations of X's can reduce leverage values (potential influence)
- if the dynamic range of Y or X is less than 1 or 10 (ie max/min) then transformation may have little effect

Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)`
- transformations of X's can reduce leverage values (potential influence)
- if the dynamic range of Y or X is less than 1 or 10 (ie max/min) then transformation may have little effect
- transformations such as logs may still be useful for interpretability

Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)`
- transformations of X's can reduce leverage values (potential influence)
- if the dynamic range of Y or X is less than 1 or 10 (ie max/min) then transformation may have little effect
- transformations such as logs may still be useful for interpretability
- outliers that are not influential may still affect the estimate of σ and width of confidence/prediction intervals.

Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)`
- transformations of X's can reduce leverage values (potential influence)
- if the dynamic range of Y or X is less than 1 or 10 (ie max/min) then transformation may have little effect
- transformations such as logs may still be useful for interpretability
- outliers that are not influential may still affect the estimate of σ and width of confidence/prediction intervals.
- Reproducibility - describe steps and decisions

Review of Last Class

- In the model with both response and predictor log transformed, are dinosaurs outliers?

Review of Last Class

- In the model with both response and predictor log transformed, are dinosaurs outliers?
- should you test each one individually or as a group; if as a group how do you think you would you do this using lm?

Review of Last Class

- In the model with both response and predictor log transformed, are dinosaurs outliers?
- should you test each one individually or as a group; if as a group how do you think you would you do this using lm?
- do you think your final model is adequate? What else might you change?

Check Your Prediction Skills

After you determine whether dinos can stay or go and refine your model, what about prediction?

- I would like to predict Aria's brain size given her current weight of 259 grams. Give me a prediction and interval estimate.

Check Your Prediction Skills

After you determine whether dinos can stay or go and refine your model, what about prediction?

- I would like to predict Aria's brain size given her current weight of 259 grams. Give me a prediction and interval estimate.
- Is her body weight within the range of the data in Animals or will you be extrapolating? What are the dangers here?

Check Your Prediction Skills

After you determine whether dinos can stay or go and refine your model, what about prediction?

- I would like to predict Aria's brain size given her current weight of 259 grams. Give me a prediction and interval estimate.
- Is her body weight within the range of the data in Animals or will you be extrapolating? What are the dangers here?
- Can you find any data on Rose-Breasted Cockatoo brain sizes? Are the values in the prediction interval?