Influence & Transformations

Merlise Clyde

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

Assumptions of Linear Regression

why linear in X

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

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

$Y \mid X, \beta, \sigma^2$

linearity is not in x but transform of x

is Normal

Animals

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

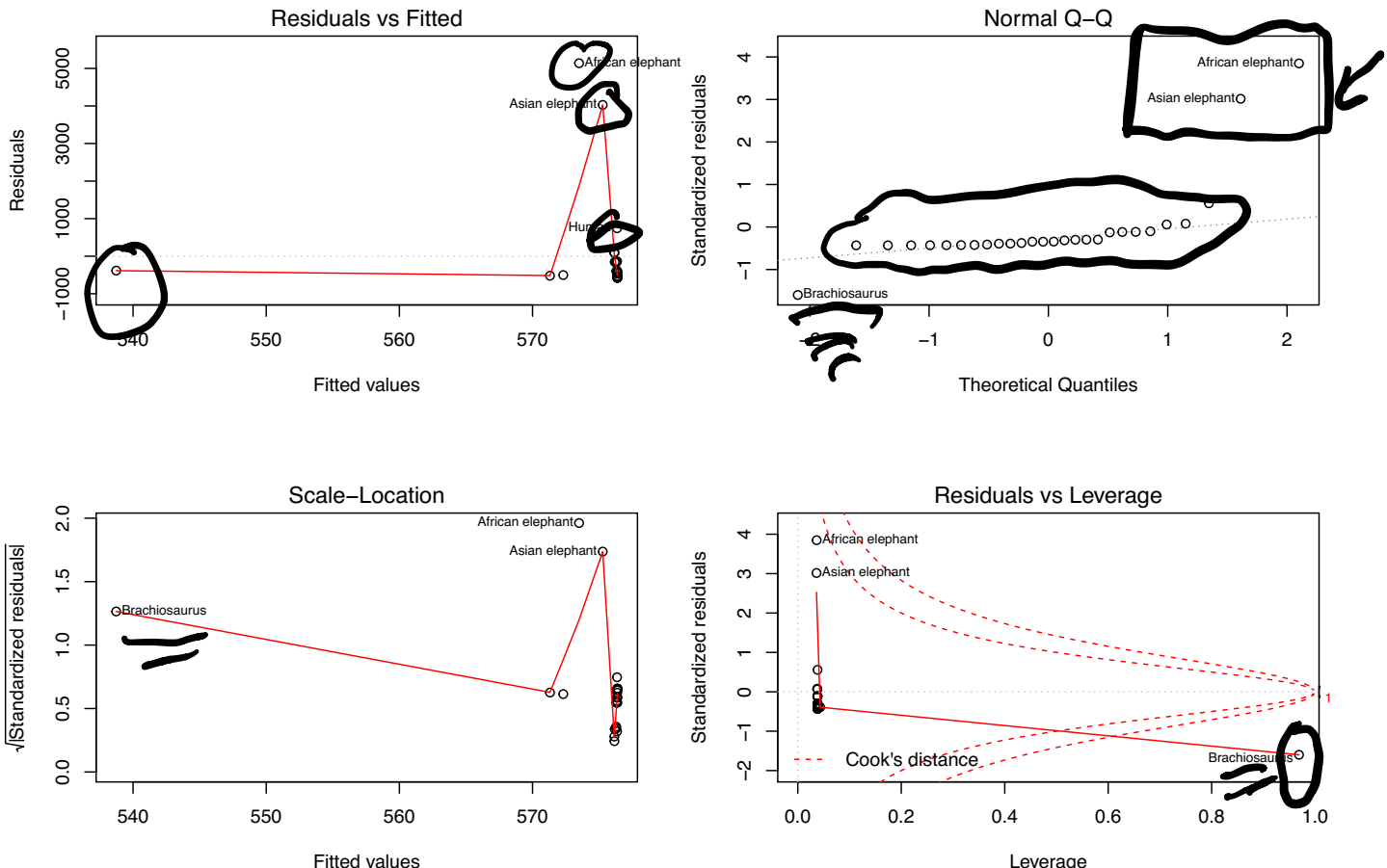
Let's try to predict brain weight from body weight.

```
data(Animals, package="MASS")
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

- 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"
```

- Use functions from the CAR package (Companion to Applied Regression) for Bonferroni correction

```
car::outlierTest(brain.lm)
```

```
##                  rstudent unadjusted p-value Bonferroni
## African elephant 5.751645         5.4098e-06   0.0001514
## Asian elephant   3.667458         1.1576e-03   0.0324120
```

Cook's Distance

Measure of influence of case i on predictions

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

after removing the ith case

Cook's Distance

Measure of influence of case i on predictions

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

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

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

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

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

after removing the ith case

Easier way to calculate

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

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

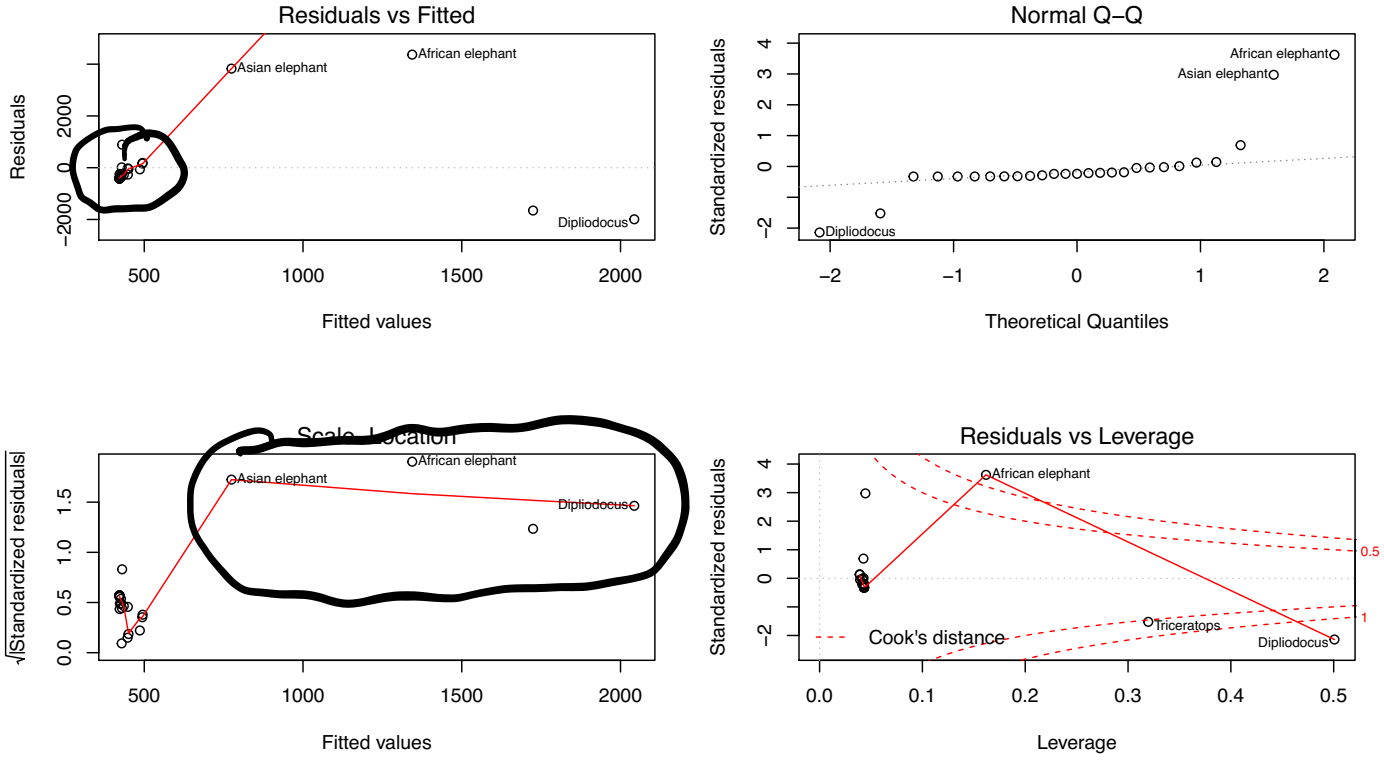
Flag cases where $D_i > 1$ or $D_i > 4/n$

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

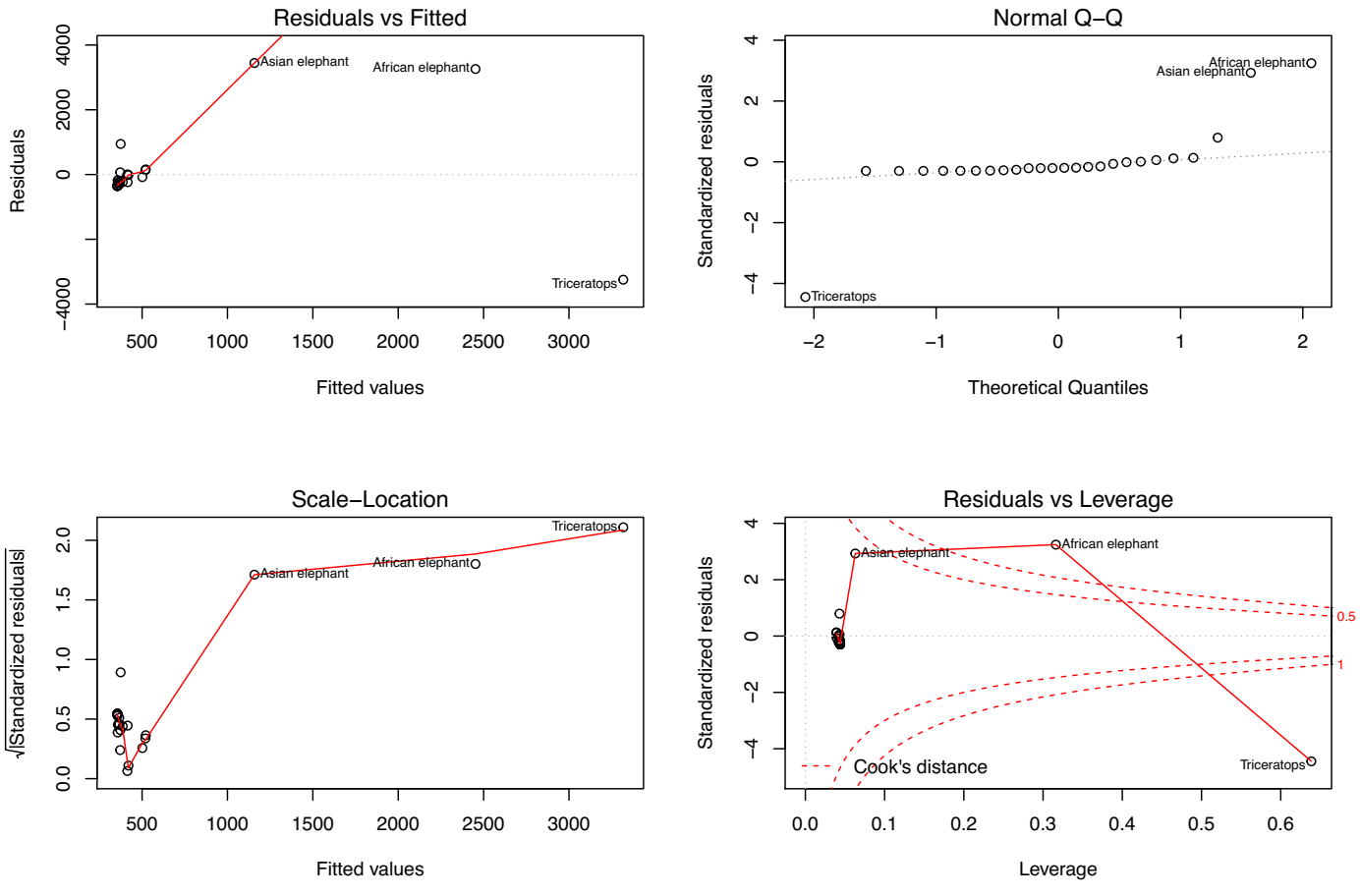
```
## [1] "Brachiosaurus"
```

Remove Influential Point & Refit

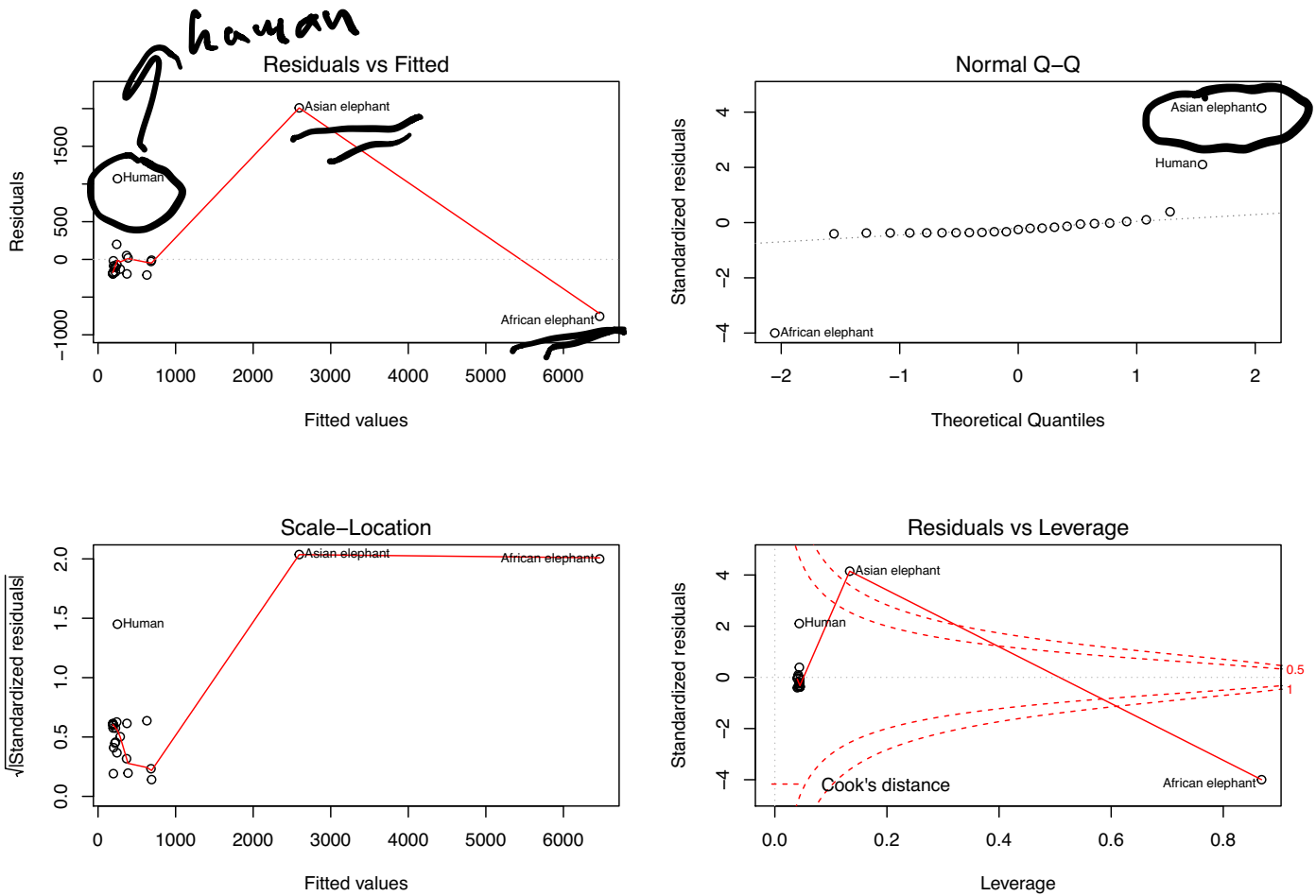
```
brain2.lm = lm(brain ~ body, data=Animals,
                 subset = !cooks.distance(brain.lm)>1)
par(mfrow=c(2,2)); plot(brain2.lm)
```



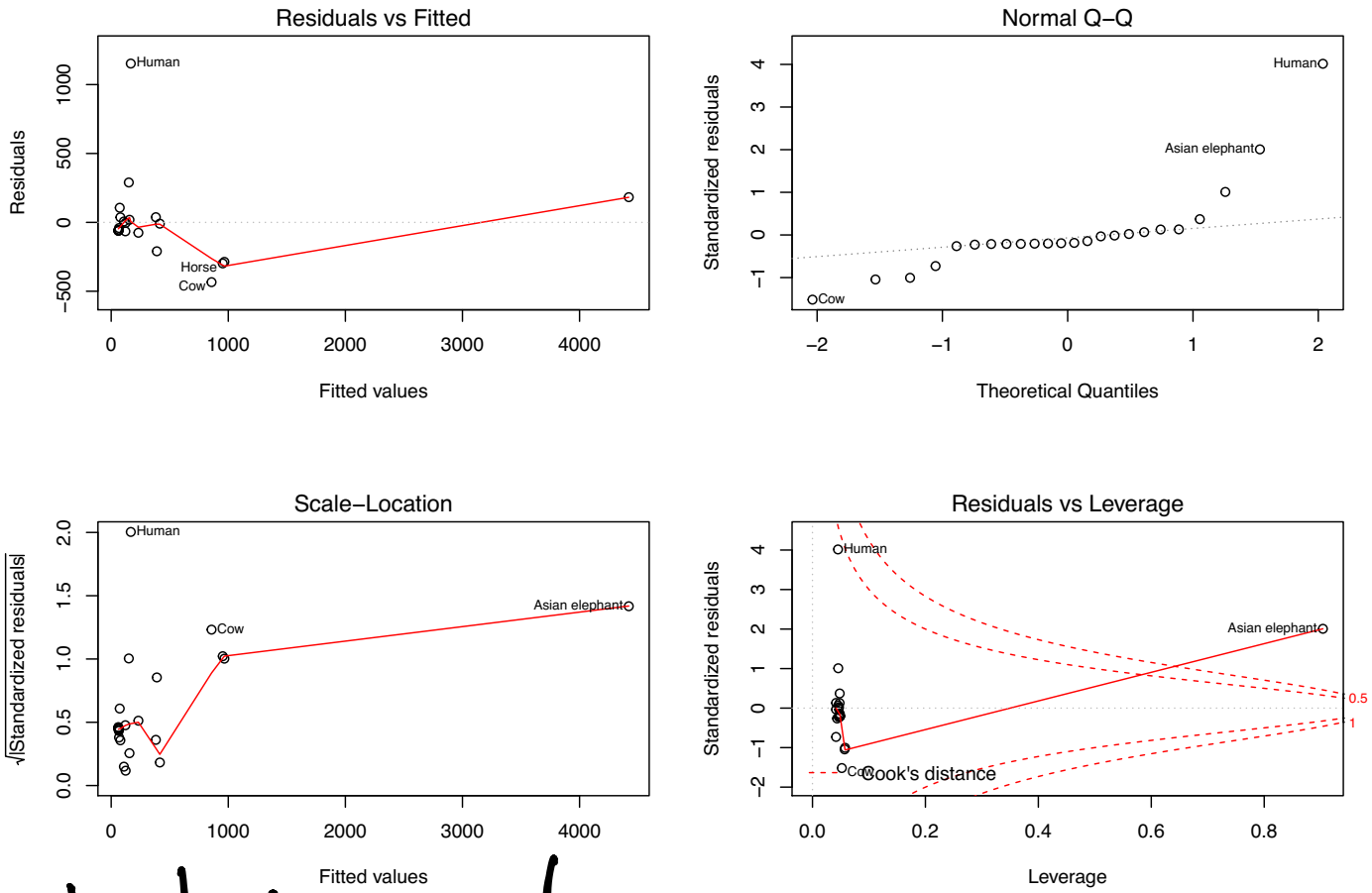
Keep removing points?



And another one bites the dust



and another one



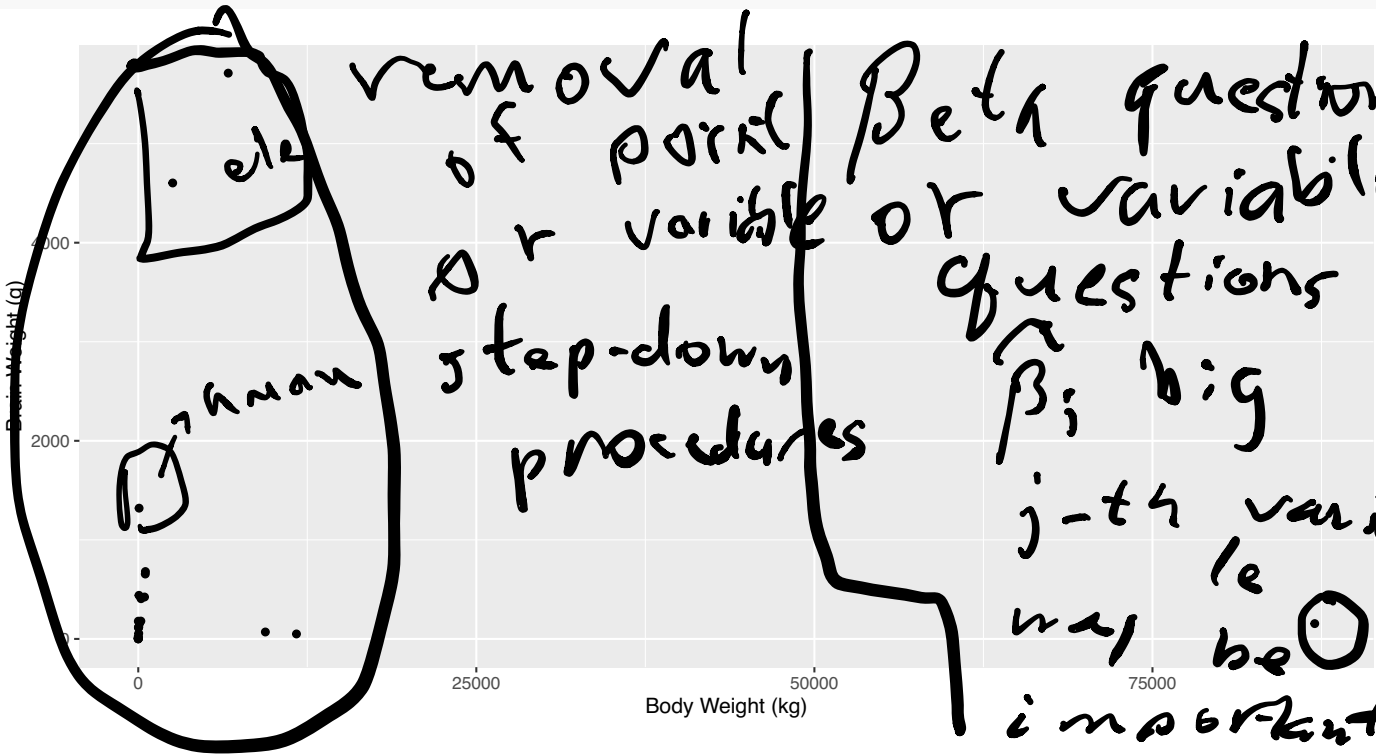
what is strange about

a sample $(y_i, x_i) \rightarrow$ questions about sample influence

And they just keep coming!

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)")
```



removal of point or variable

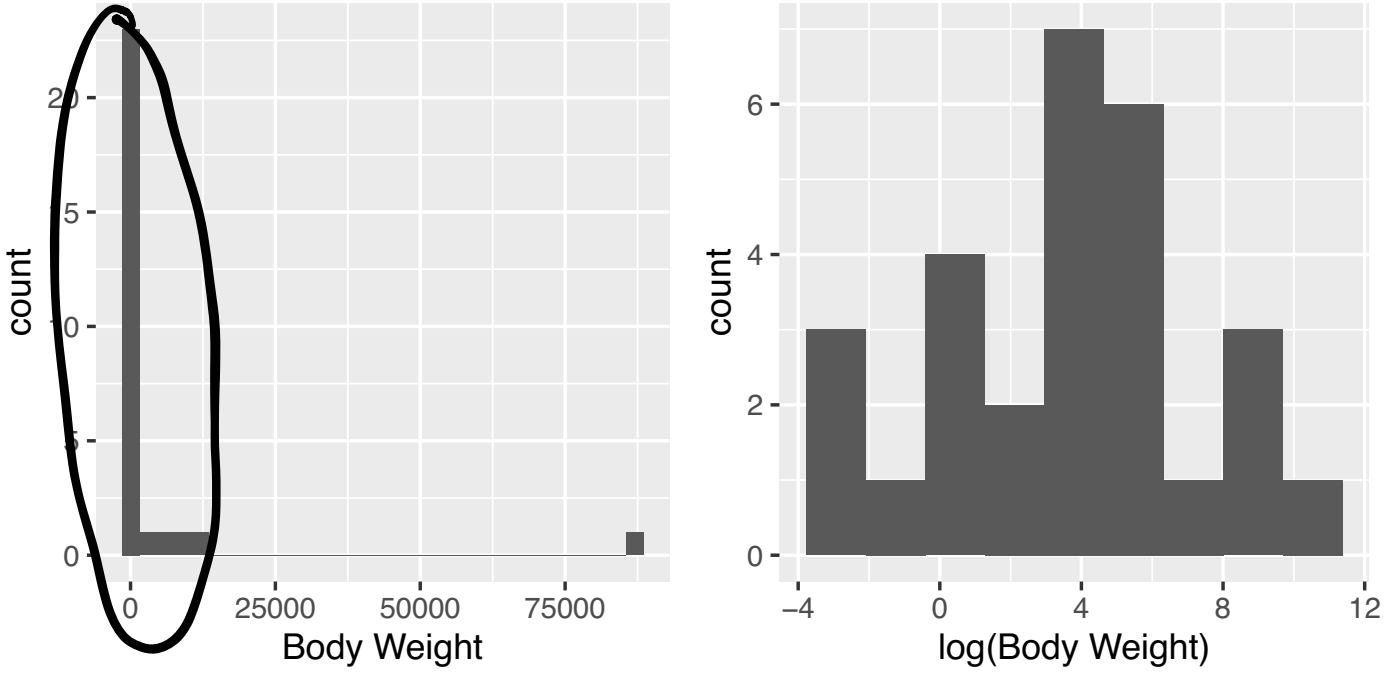
step-down procedures

Beta questions or variable questions

$\hat{\beta}_j$ big

j-th variable may be important

Log Transform



(a)

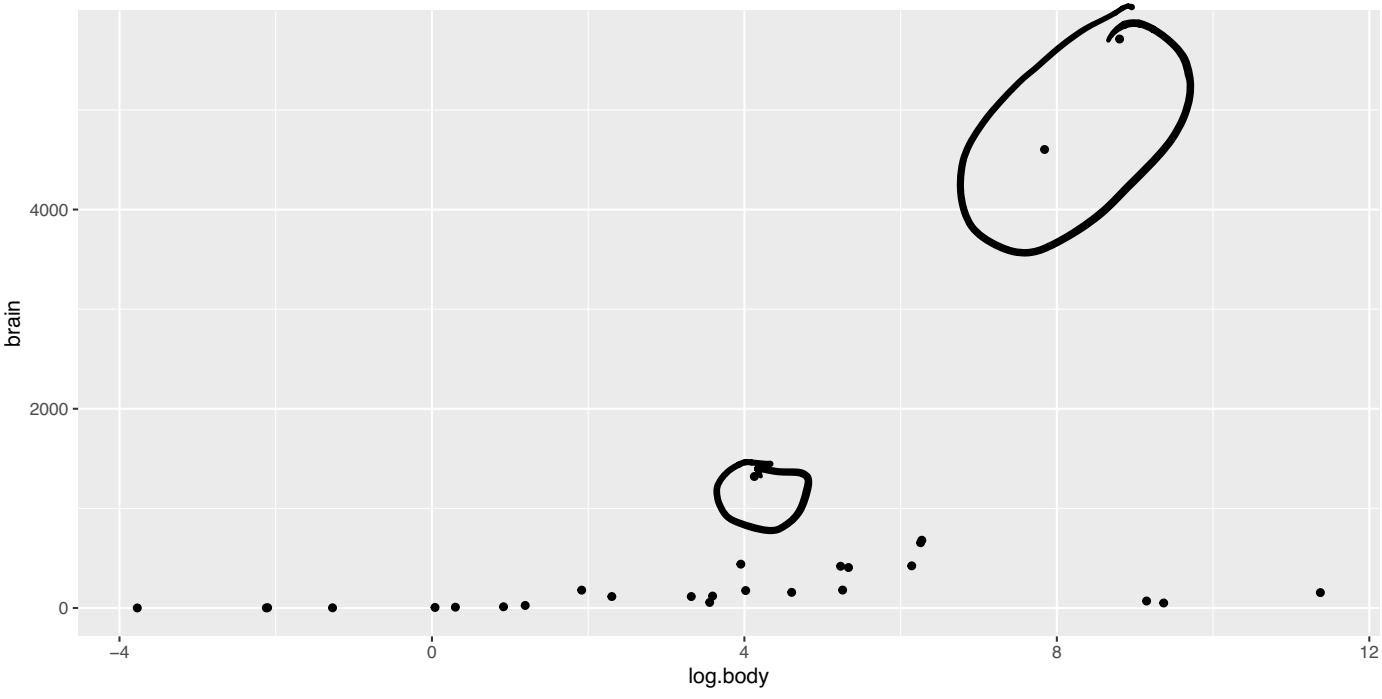
$$bw_i = \beta \log(bw_i) + \varepsilon$$

(b)

$$= \beta\, bw_i + \varepsilon$$

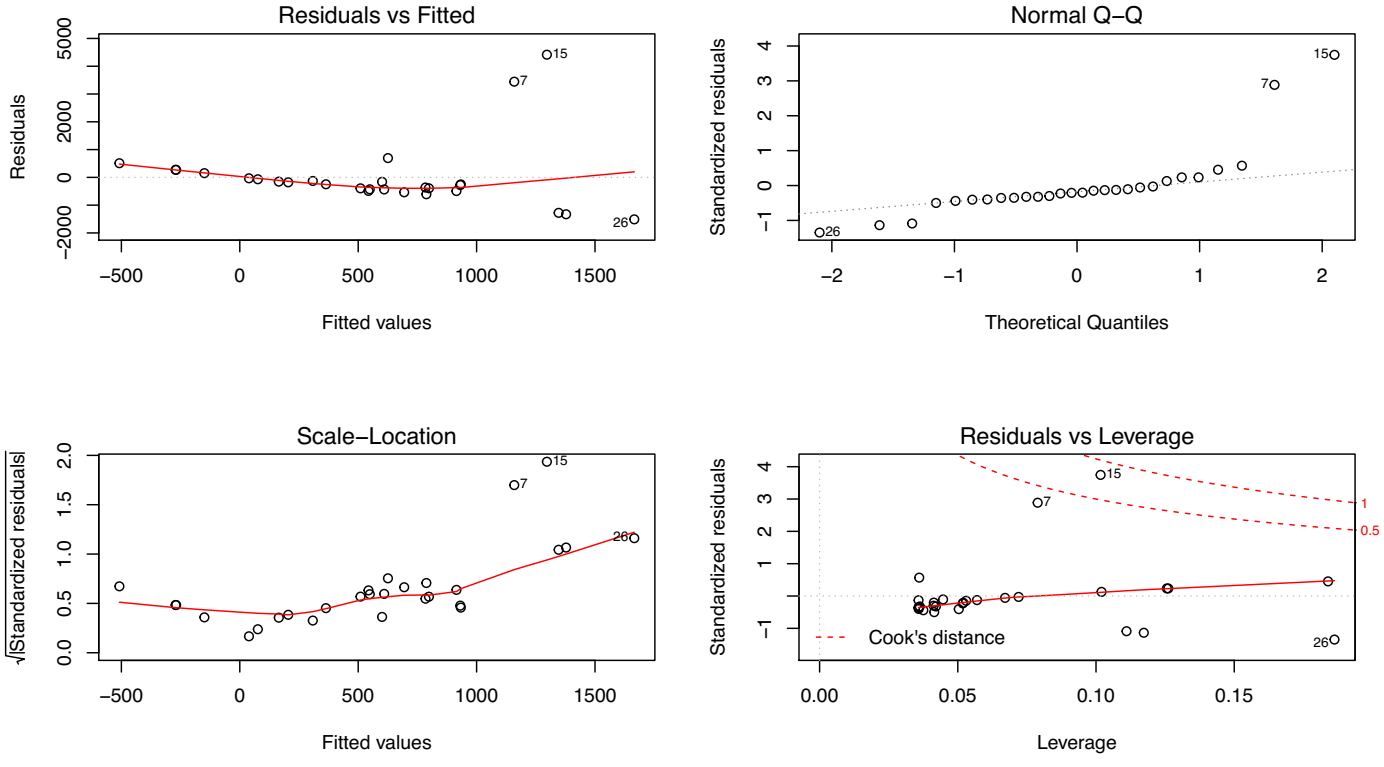
Plot of Transformed Data

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




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

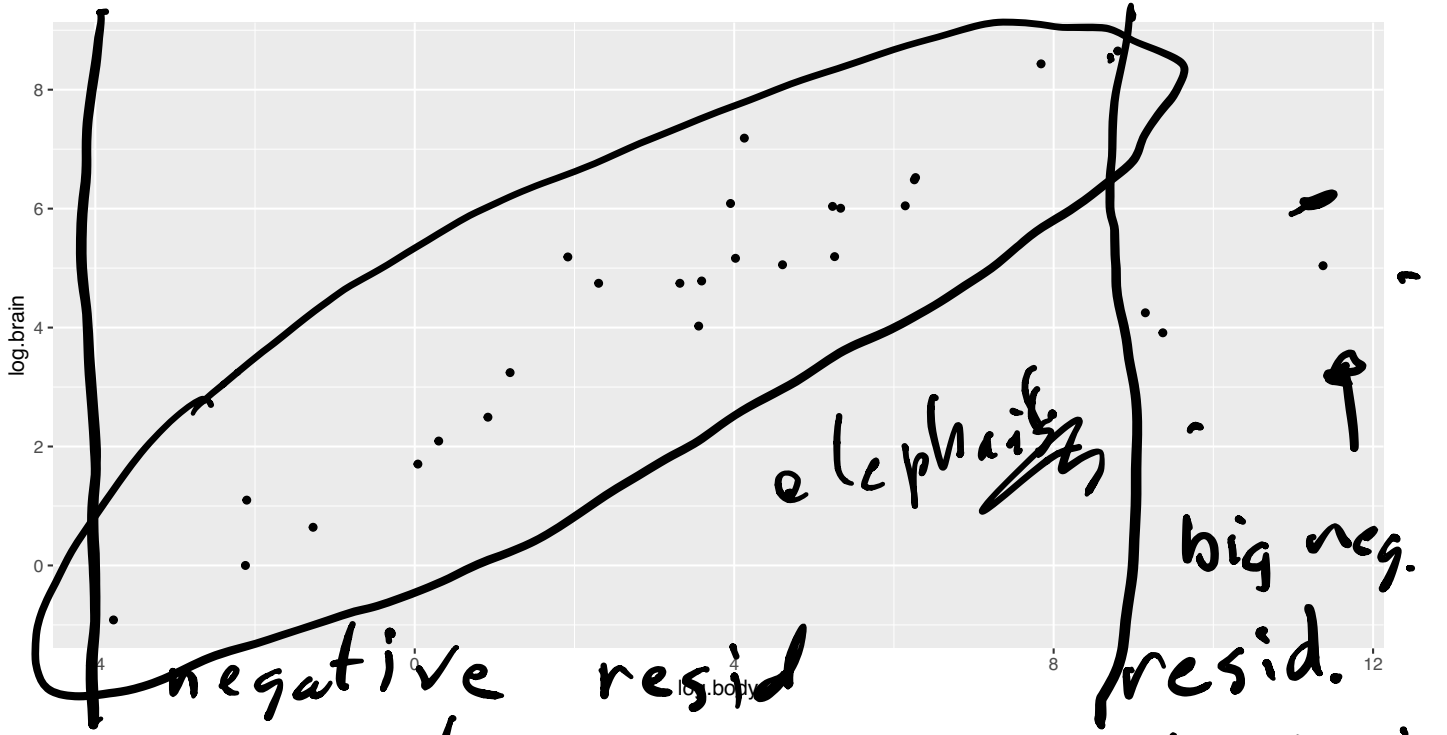
Diagnostics with log(body)



Variance increasing with mean

Try Log-Log

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

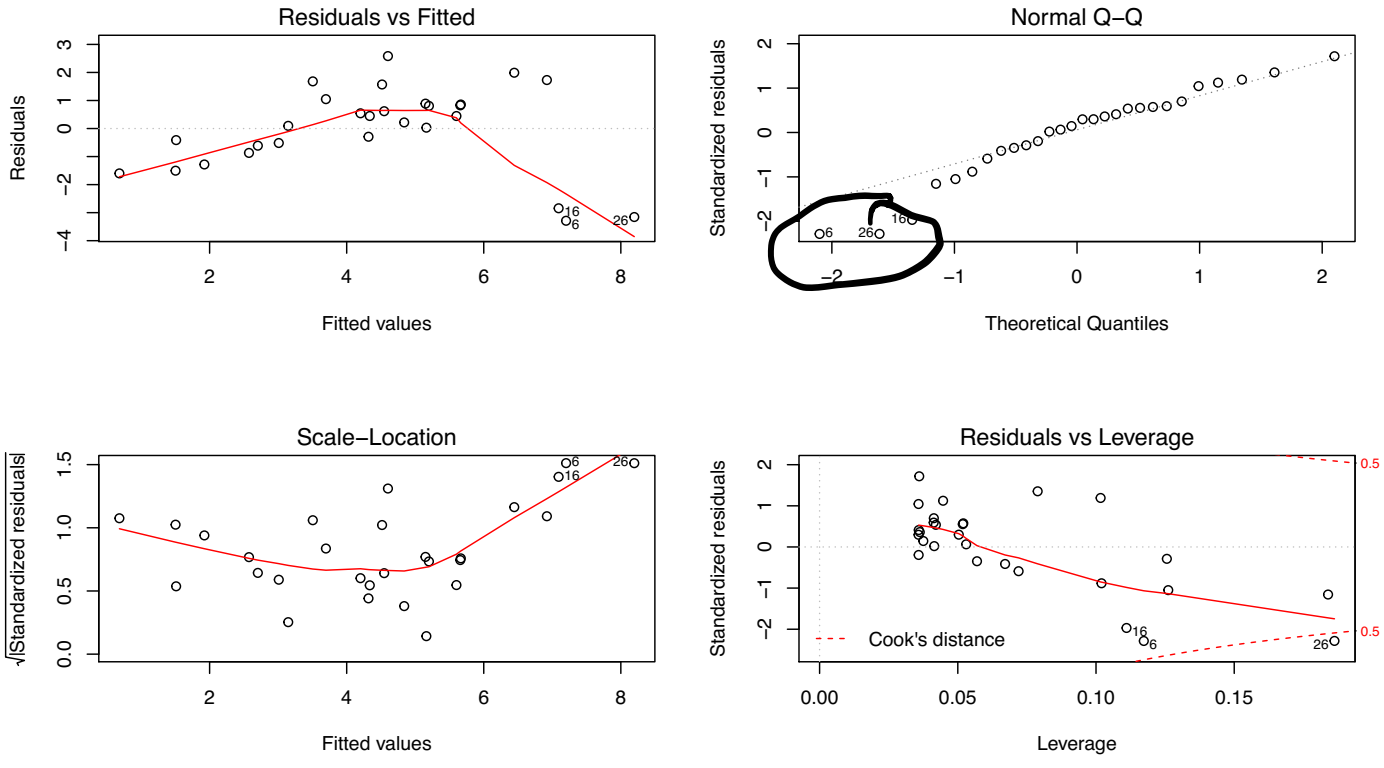



```
#plot(brain ~ body, Animals, log="xy")
```

brain is much tn

body

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.

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

$$\frac{Y^{\lambda}-1}{\lambda}$$

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 $\mathbf{Y}$ first)

Find λ to minimize

$$\lim_{\lambda \to 0} \frac{Y^{\lambda}-1}{\lambda} = \log(Y)$$

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

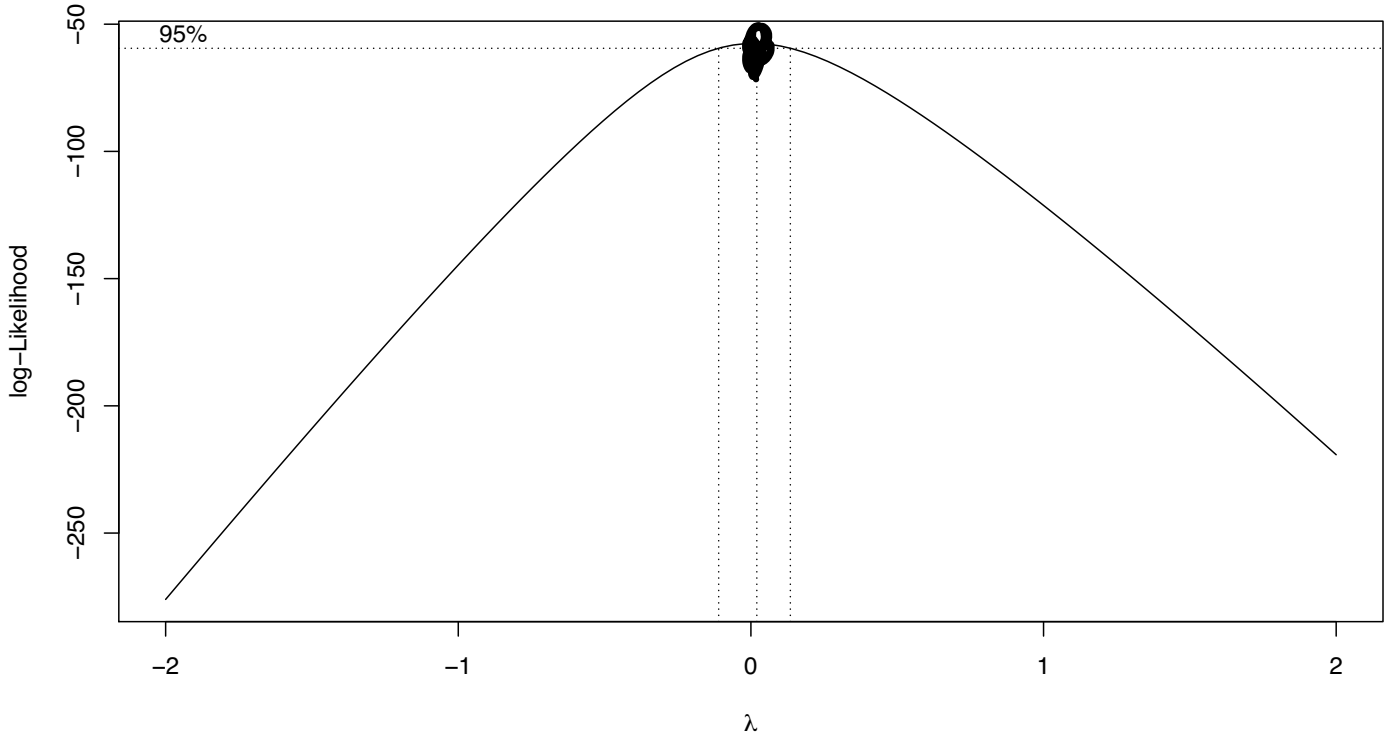
$$\Psi_M(\mathbf{Y}, \lambda) = \begin{cases} (\mathrm{GM}(\mathbf{Y})^{1-\lambda}(\mathbf{Y}^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \mathrm{GM}(\mathbf{Y}) \log(\mathbf{Y}) & \text{if } \lambda = 0 \end{cases}$$

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

boxcox in R: Profile likelihood

$\exp(2\log L) = (\prod_i Y_i)^{2/n}$

```
MASS::boxcox(braintransX.lm)
```



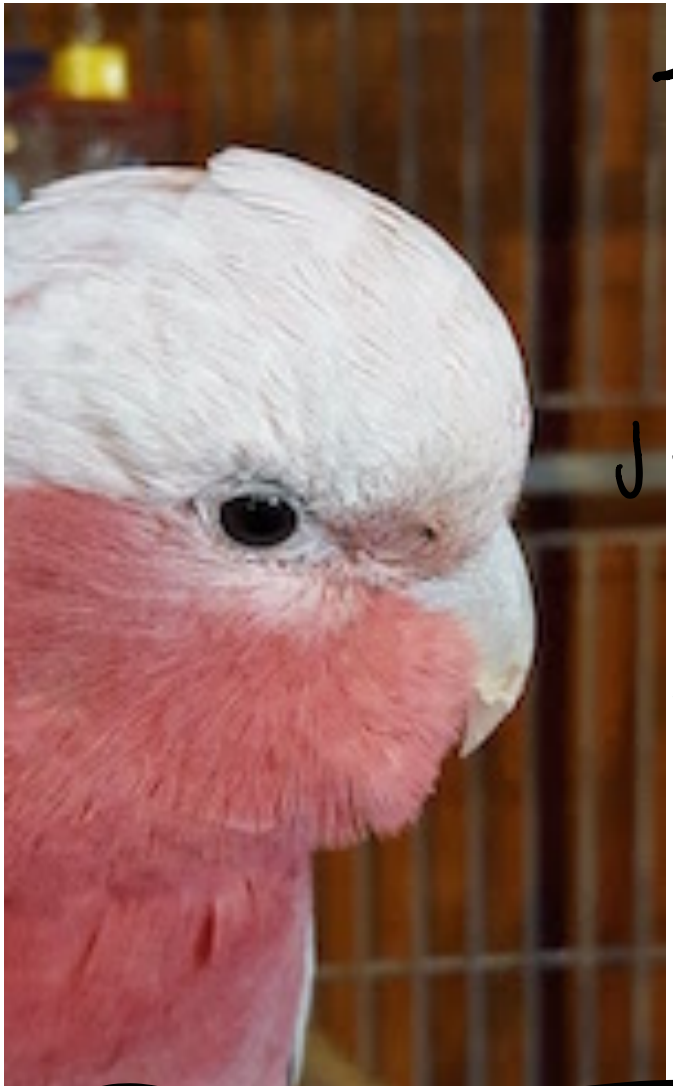
Caveats

- Boxcox transformation depends on choice of transformations of X's
- For choice of X transformation use `boxTidwell` in `library(car)` or for multivariate normality of response and predictors `powerTransform` in the `car` library.
- 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 for removing a case

Next 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?
- after you determine whether dinos can stay or go and refine your model, what about prediction?
- what about model uncertainty?

Check Your Prediction Skills



predict $Y_i \mid X_i = 259$ gr.

y_i

- 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?

Jacobian

$P(y_i \mid x_i)$ is in 95% interval

$f(x)\,dx$

$f(T(x))$

$\boxed{\text{extrapolation vs. interpolation}}$

Pred. modeling

---

mammals + dinosaurs

$$\log y_i = (\log x_i)\beta_m + \varepsilon$$

$$\log y_i = (\log x_i)\beta_d + \varepsilon$$

don't know if (x_i, y_i) is a dino or a mammal?

---

hierarchical model.

$Z_i = 1$ if dino 0 o.w.

$$Z_i \sim \text{Bern}(p)$$

$$Y_i \mid Z_i = 1, \beta_d, X_i, \sigma^2 \sim N(\beta_d \log(x_i), \sigma^2)$$

$$Y_i \mid Z_i = 0, \beta_m, x_i, \sigma^2 \sim N(\beta_m \log(x_i), \sigma^2)$$

what if I don't give you z_i's?

$$\max_{\beta\text{'s}, z_i} \log \text{lik}(Y_i \mid X_i, \beta_d, \beta_m, \sigma^2, z_i)$$

harder problem