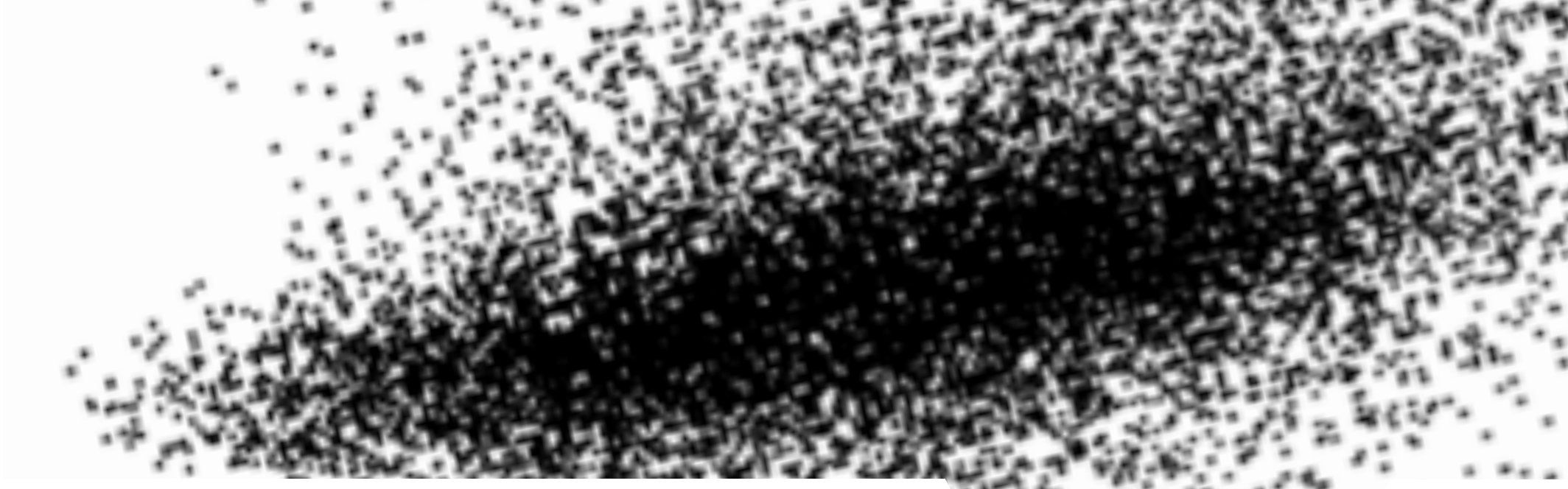


ST 117

5. Regression

WARWICK

Lecture 23/24
(Week 8)

Example: Mammals
Regression diagnostics
Model fit
Data transformations
More examples

R Data Set: Mammals

Brain and body weights for 62 different land mammals.

```
> library(MASS)
> mammals
```

	body	brain
Arctic fox	3.385	44.50
Owl monkey	0.480	15.50
Mountain beaver	1.350	8.10
Cow	465.000	423.00
Grey wolf	36.330	119.50
Goat	27.660	115.00
Roe deer	14.830	98.20
Guinea pig	1.040	5.50
Verbet	4.190	58.00
Chinchilla	0.425	6.40
Ground squirrel	0.101	4.00
Arctic ground squirrel	0.920	5.70

R Data Set: Mammals

Type in ?mammals. You might need to upload the MASS libary first.

Brain and Body Weights for 62 Species of Land Mammals

Description

A data frame with average brain and body weights for 62 species of land mammals.

Usage

```
mammals
```

Format

`body`

body weight in kg.

`brain`

brain weight in g.

`name`

Common name of species. (Rock hyrax-a = *Heterohyrax brucci*, Rock hyrax-b = *Procavia habessinic.*.)

R Data Set: Mammals (Data Format)

```
> library(MASS)
> dim(mammals)
[1] 62  2
> str(mammals)
'data.frame':         62 obs. of  2 variables:
 $ body : num  3.38 0.48 1.35 465 36.33 ...
 $ brain: num  44.5 15.5 8.1 423 119.5 ...
> head(mammals)
                    body brain
Arctic fox         3.385  44.5
Owl monkey         0.480  15.5
Mountain beaver    1.350   8.1
Cow              465.000 423.0
Grey wolf         36.330 119.5
Goat              27.660 115.0
```

Attaching datasets

Allows us to access directly variables of a dataset.

```
> body[1:4] #want to access the first 4 body weights
Error in body[1:4] : object of type 'closure' is not subsettable
> mammals$body[1:4]
[1]   3.385   0.480   1.350 465.000
>
> attach(mammals)
> body[1:4]
[1]   3.385   0.480   1.350 465.000
```

(This is a data preparation step for convenience)

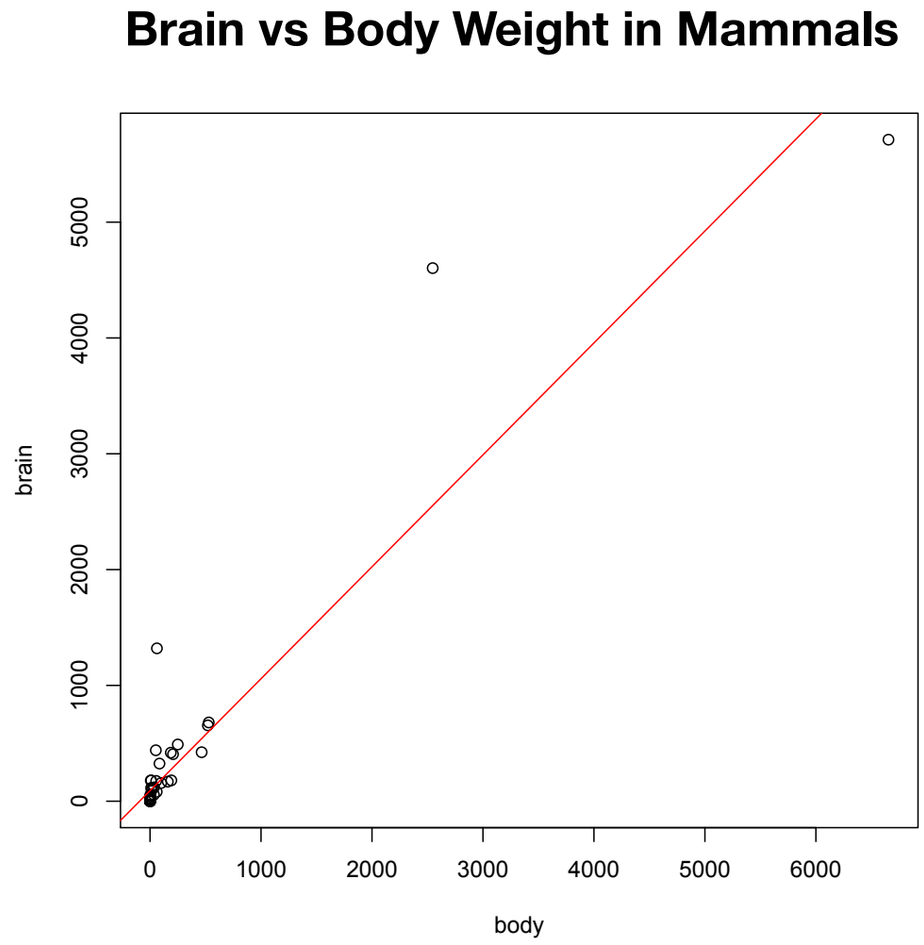
Logical Subscript Practice

```
> mammals[,mammals[,1]>100]
Error in `[.data.frame`(mammals, , mammals[, 1] > 100) :
  undefined columns selected
> mammals[mammals[,1]>100,]
                  body brain
Cow               465.0   423
Asian elephant   2547.0  4603
Donkey            187.1   419
Horse             521.0   655
Giraffe           529.0   680
Gorilla           207.0   406
African elephant 6654.0  5712
Okapi             250.0   490
Pig               192.0   180
Brazilian tapir   160.0   169
>
> #mammals[body>100,]
> #would give the same output
```

Scatter Graph For Mammals Data

Do we have a linear relationship between the body and brain weights in mammals?

```
plot(body, brain, xlab="Body weight (kg)", ylab="Brain weight (kg)",
     main="Brain vs body weights in mammals")
```



Hard to see what is going on in the lower left corner! Too much over plotting. Could make extra plot of just those values on a different scale.

Regression Line

How do we fit a regression line? Try `lm` function. Type ?lm to get:

Fitting Linear Models

Description

`lm` is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).

Usage

```
lm(formula, data, subset, weights, na.action,
   method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
   singular.ok = TRUE, contrasts = NULL, offset, ...)
```

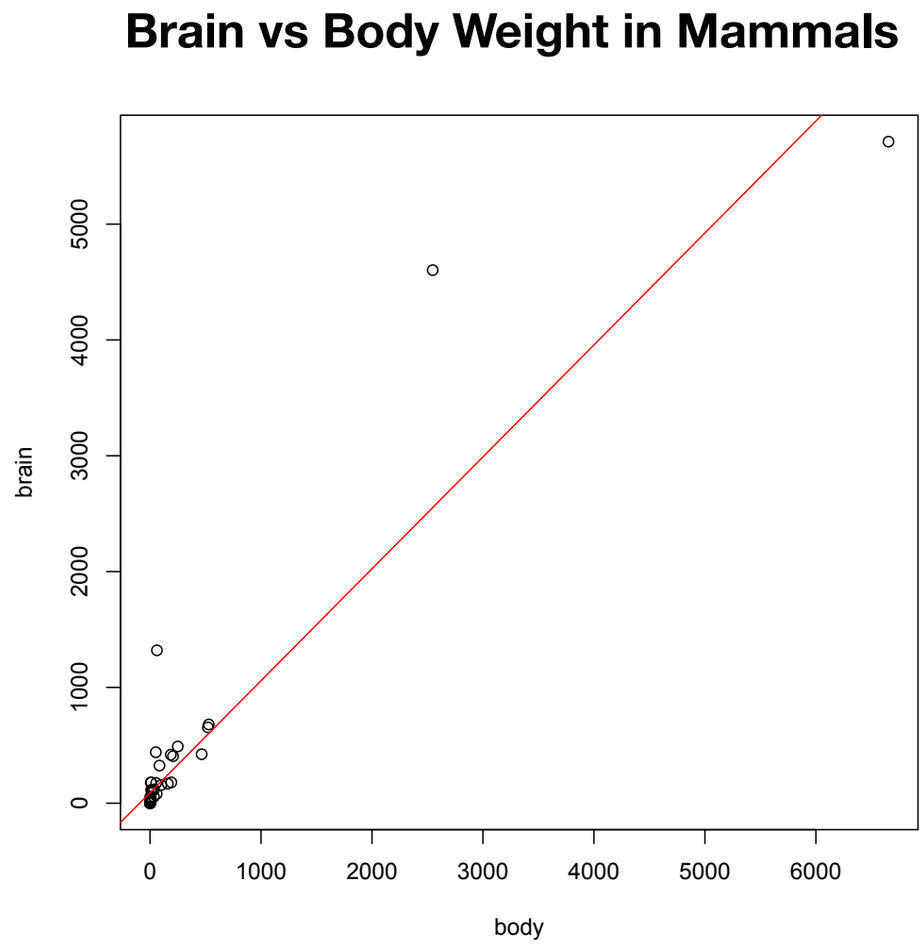
Arguments

`formula`	an object of class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted. The details of model specification are given under 'Details'.

A typical model has the form `response` $\sim$ `terms`, where `response` is the numeric response vector and `terms` is a series of terms which specifies a linear predictor for response. We will use lm(brain $\sim$ body).

Using `lm`

```
plot(body, brain)
Regression<-lm(brain~body)
abline(Regression)
```



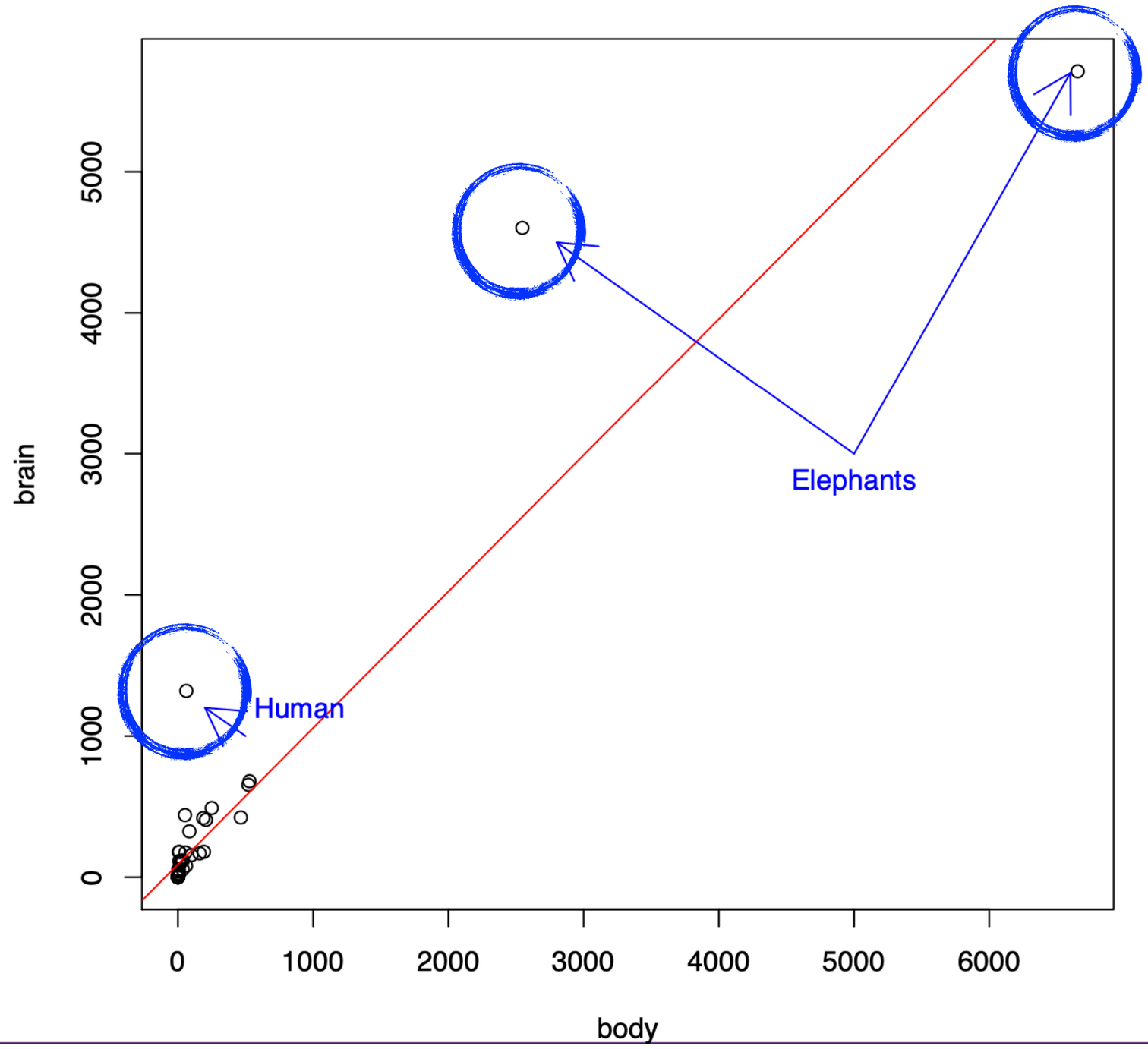
Regression Line Formula

Finding the coefficients of the regression line is not complicated.

```
> Regression

Call:
lm(formula = brain ~ body)

Coefficients:
(Intercept)         body
    91.0044       0.9665
```

$$brain = 91.0044 + 0.9665 \times body$$

Outliers?


brain
0
1000
2000
3000
4000
5000
0
1000
2000
3000
4000
5000
6000
body
Elephants
Human

Dealing With Outliers

Should unusual observations be included in the fitting?

- Yes, to keep data as they really are.
- No, they might be misleading and regression is too sensitive to outliers.

Look closely at *reasons* for unusual observation:

- Error?
- Different measurement method?
- Too isolated from other values in predictor variable? (Potentially *high leverage*).

Outlier Definition

Recall: The residual e_i is the difference between the true value of y_i and its predictive value $\hat{y}_i$

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha} - \hat{\beta} x_i$$

Observation with response which is unusual relative to the fitted value (i.e. with large absolute residual). But what is "large"?

Scale residual by its standard deviation. Normal errors imply normal residuals, which can be standardised:

$$\underline{e}_i = \frac{e_i}{\sigma_i}$$

Different softwares and methodologies may differ, but flag values outside $(-2, 2)$, or sometimes outside $(-3, 3)$.

Influential Points

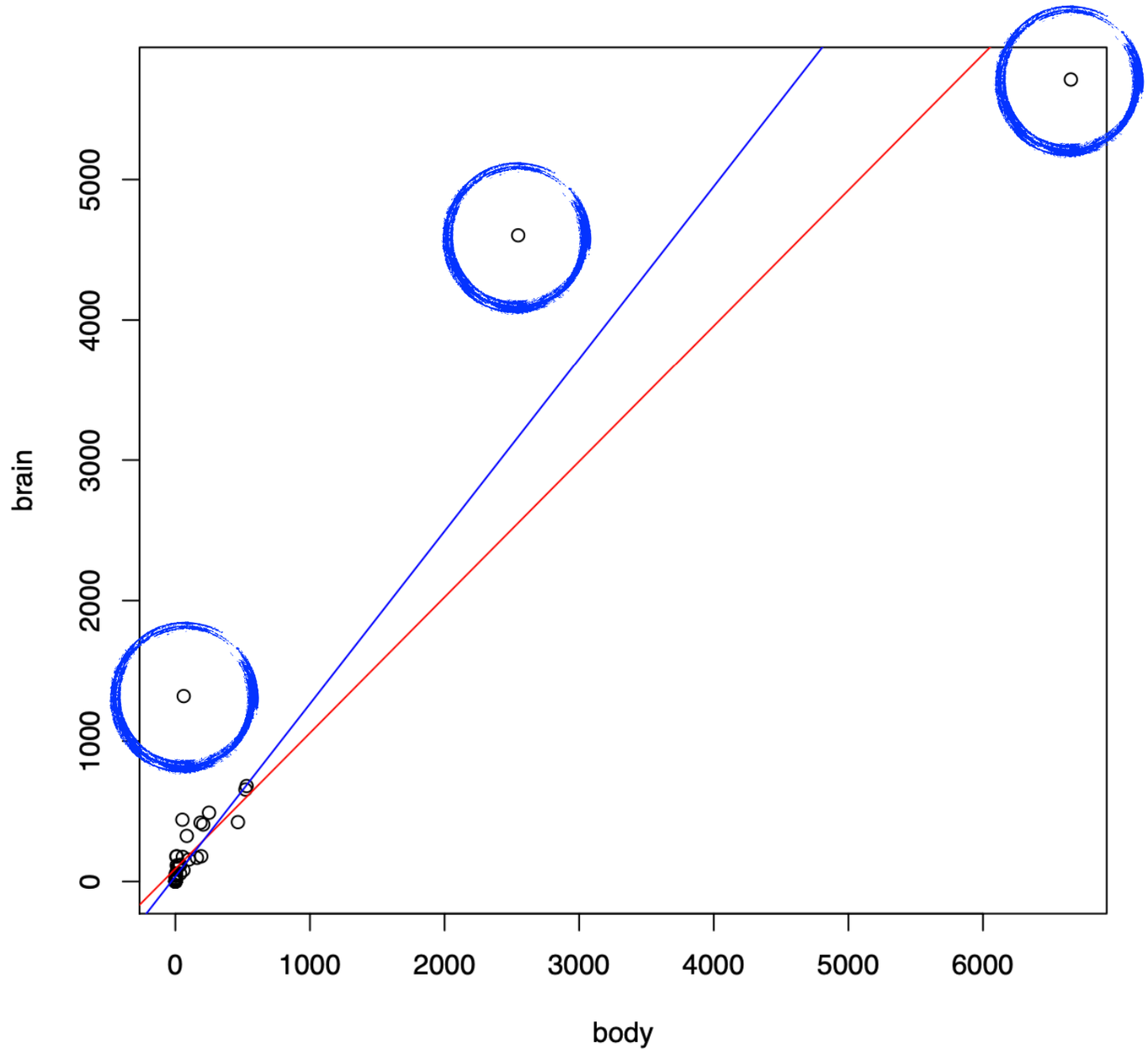
Compare fits with and without influential points.

Old model

Brain = 91.004 + 0.967 x body

New model

Brain = 36.572 + 1.228 x body



Associated code

```
> stand.res<-Regression$residuals/sd(Regression$residuals)
> which(abs(stand.res)>2)
19 32 33
19 32 33
>
> mammals[c(19,32,33),]
                 body brain
Asian elephant   2547  4603
Human              62  1320
African elephant 6654  5712
> mammals2=mammals[-c(19,32,33),]
> dim(mammals2)
[1] 59  2
```

Associated code

```
> reg2<-lm(brain~body, data=mammals2)
> reg2

Call:
lm(formula = brain ~ body, data = mammals2)

Coefficients:
(Intercept)         body
     36.572        1.228
```

```
> plot(body, brain)
> abline(Regression, col="red")
> abline(reg2, col="blue")
> plot(mammals2$body, mammals2$brain)
> abline(reg2, col="red")
```

Residual plots

After defining a linear model `m`, if we type `plot(m)`, `R` gives us four plots (might have to press `ENTER` to make each new plot appear):

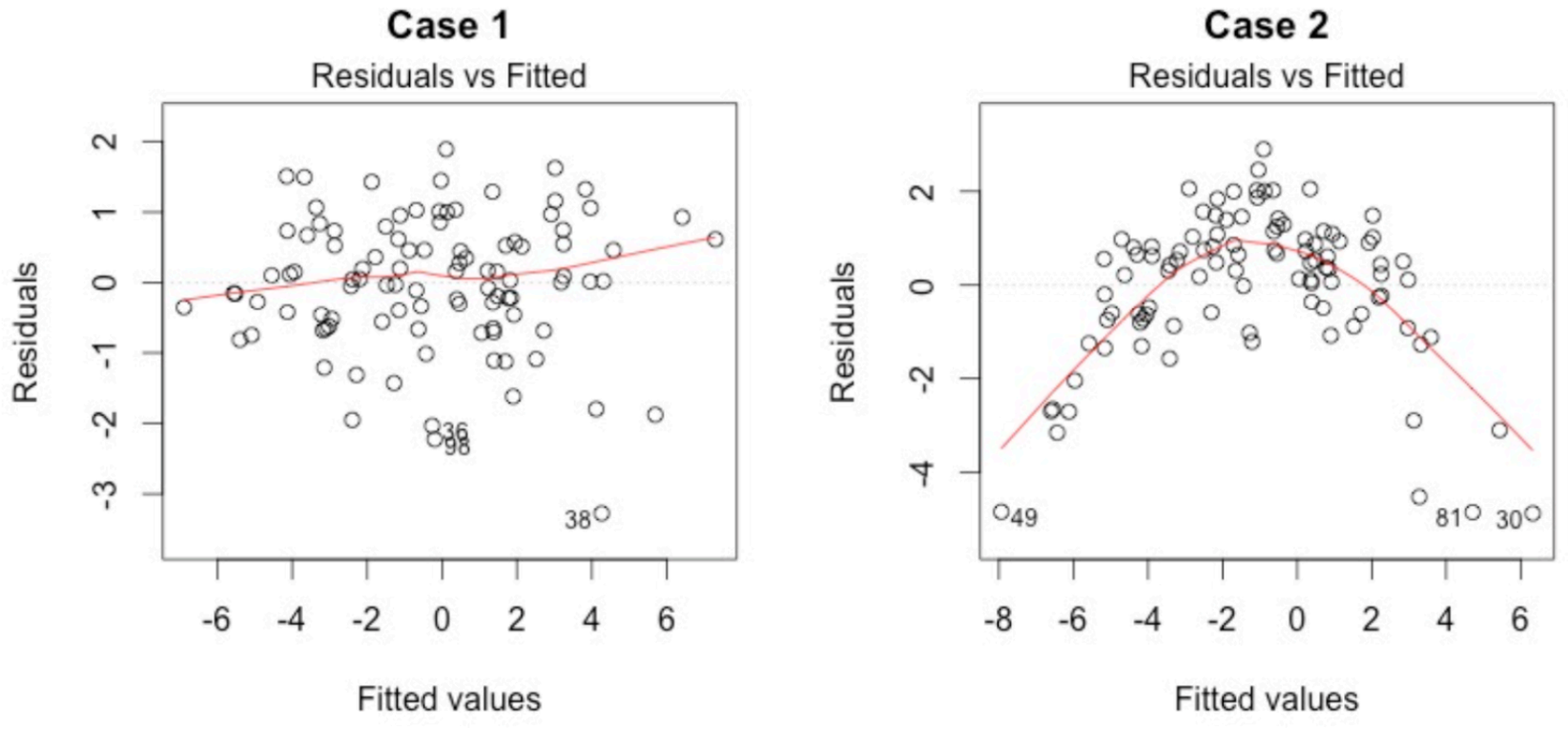
1. Residuals vs Fitted Values plot: This plot shows if residuals have non-linear patterns.
2. Normal Q-Q plot: This plot shows if residuals are normally distributed.
3. Scale-Location plot: This plot shows if residuals are spread equally along the ranges of predictors.
4. Residuals vs Leverage plot: This plot helps us to find influential cases.

Residual plots

After defining a linear model `m`, if we type `plot(m)`, `R` gives us four plots (might have to press `ENTER` to make each new plot appear):

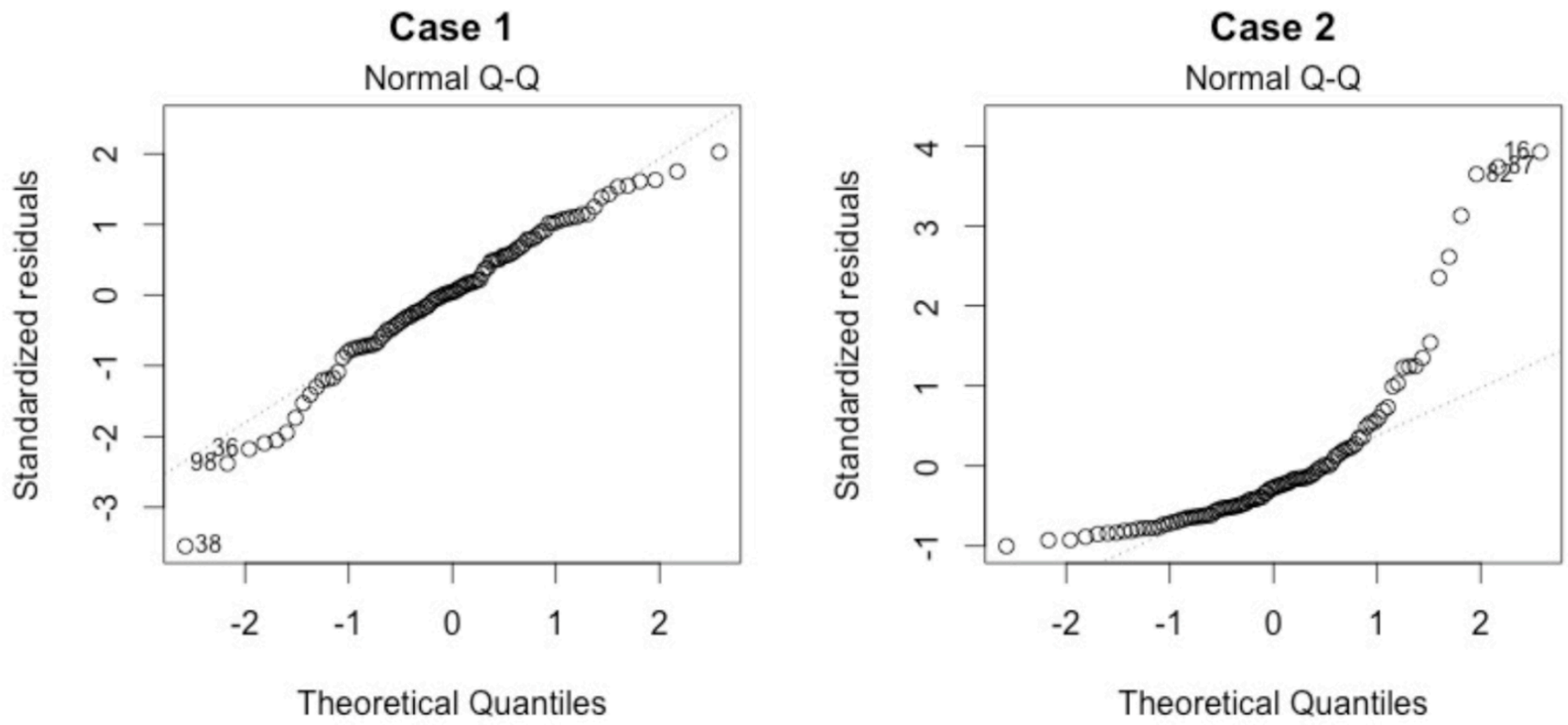
1. Residuals vs Fitted Values plot: This plot shows if residuals have non-linear patterns.
2. Normal Q-Q plot: This plot shows if residuals are normally distributed.
3. Scale-Location plot: This plot shows if residuals are spread equally along the ranges of predictors.
4. Residuals vs Leverage plot: This plot helps us to find influential cases.

Residuals vs fitted



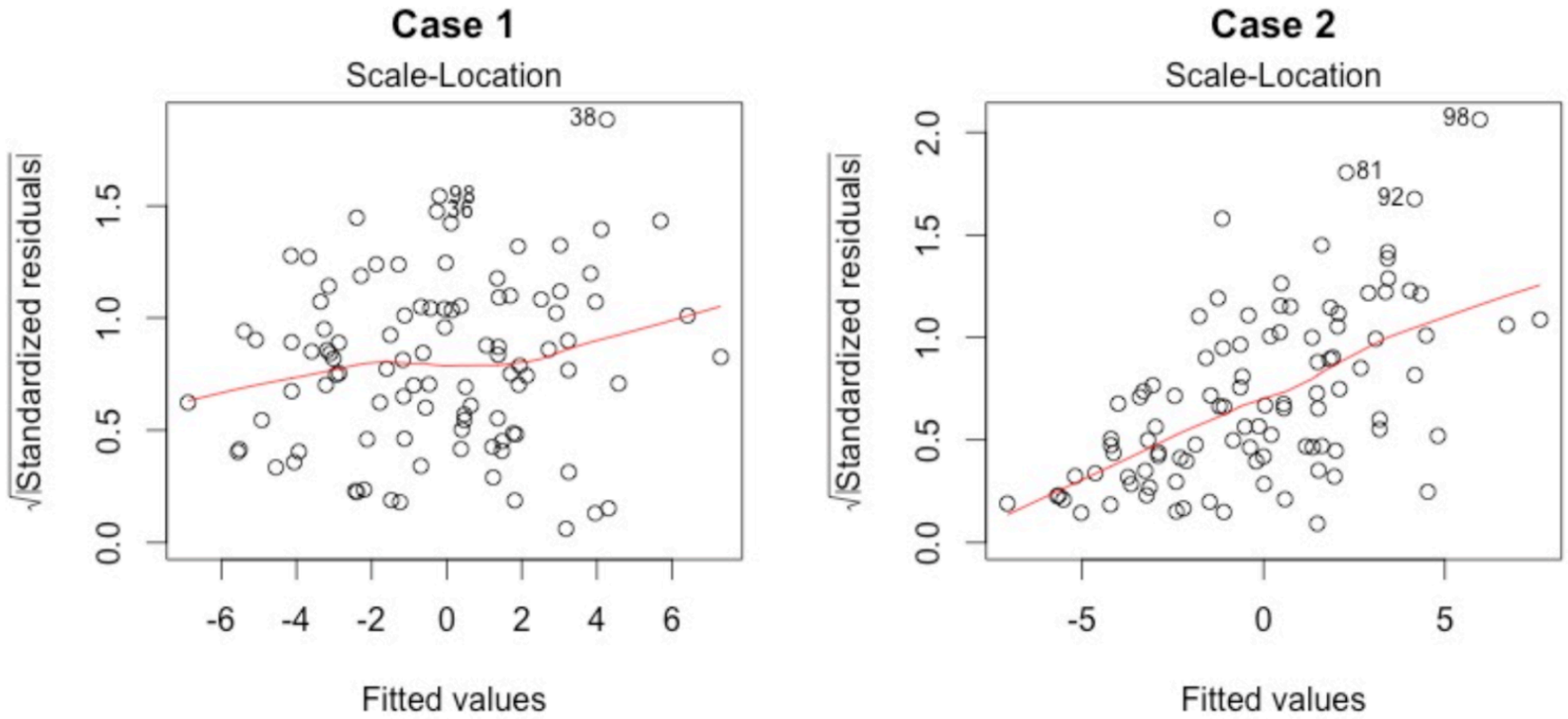
There could be a non-linear relationship between predictor variables and an outcome variable and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships.

Normal Q-Q



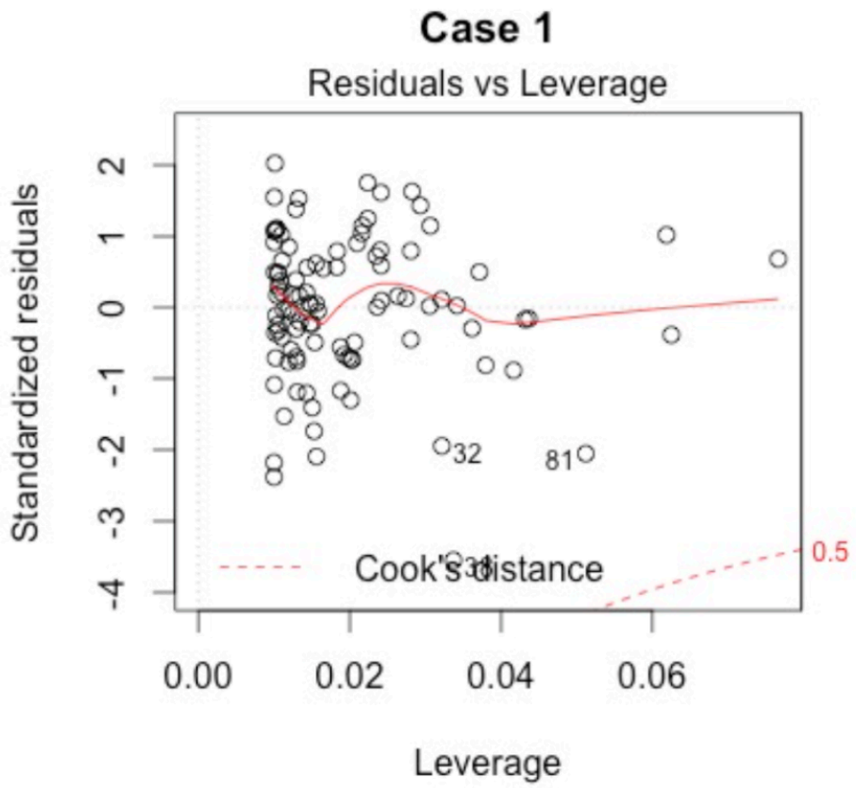
Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line.

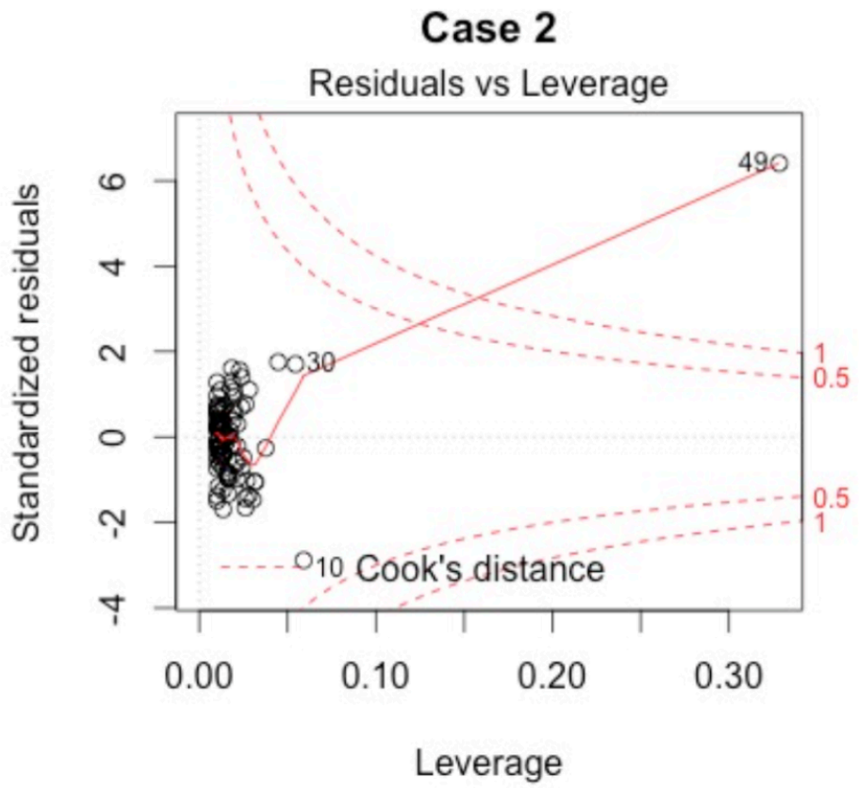
Scale-Location



Also called Spread-Location plot. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points.

Residuals vs Leverage





Unlike the other plots, this time patterns are not relevant. We watch out for values outside the red dashed lines. When cases have high Cook's distance scores, the cases are influential to the regression results.

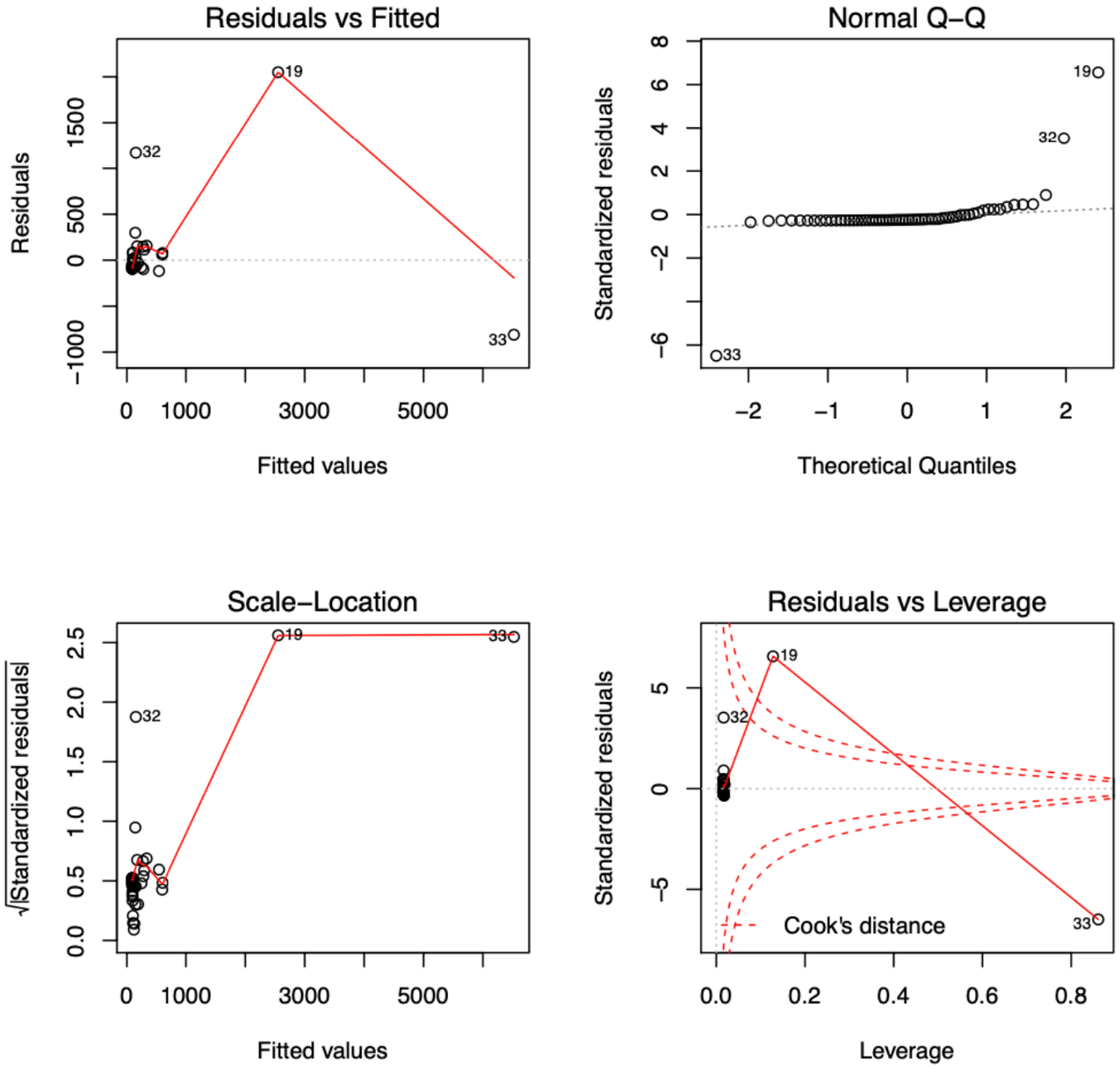
What to do if we identify an issue?

These diagnostic plots are not a strict "go" or "stop" sign. It can tell you several things about the data.
You may want to rethink your model and hypotheses. You may want to:

- Transform variables
- Add new variables in the model
- Remove a few influential points
- Need better or different data collection methods, because of systematic bias in the data
- Possibly other things.

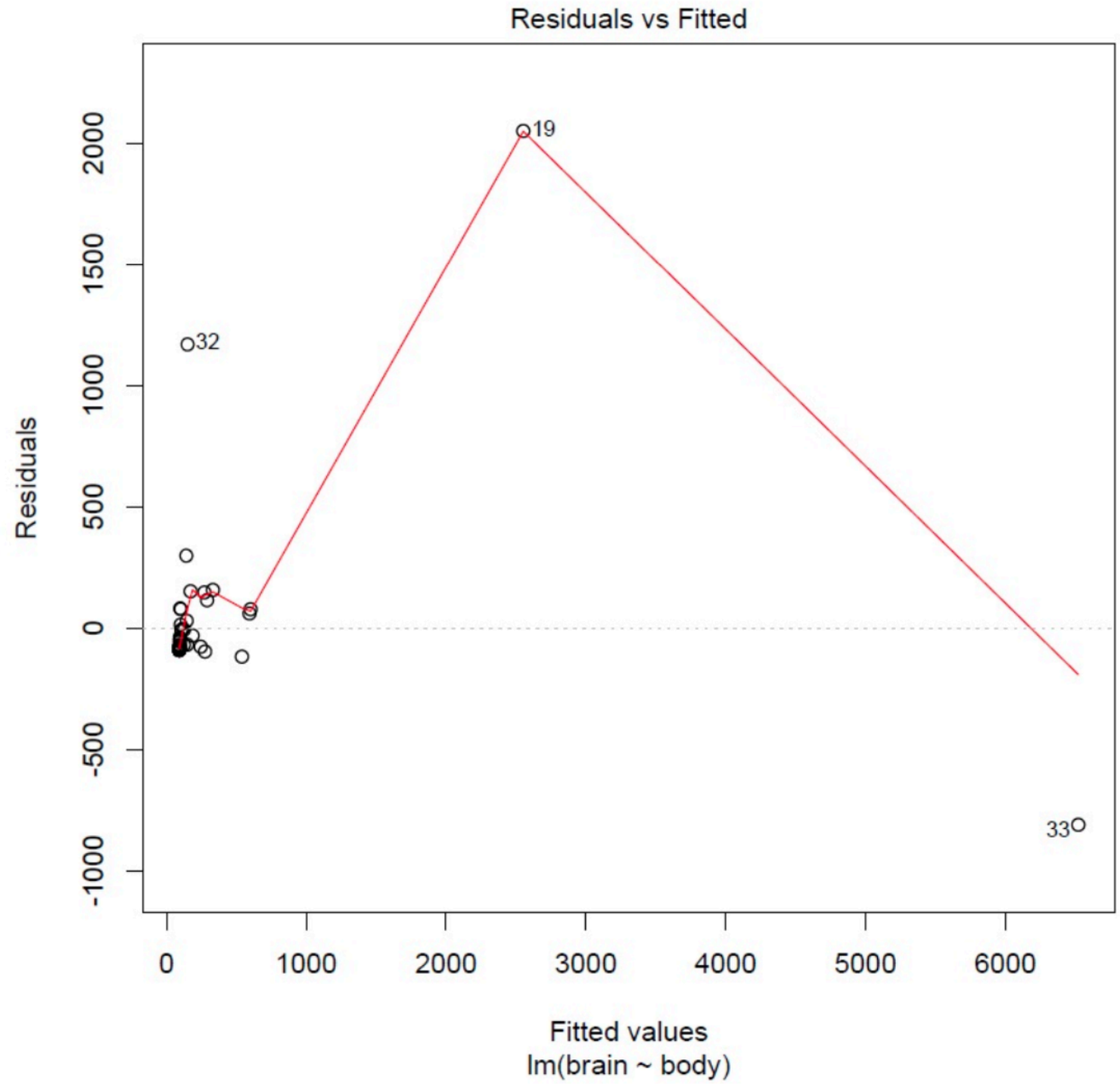
Residual Diagnostics For Mammal Data Cont.

```
par(mfrow=c(2,2))
plot(Regression, which=c(1:3,5))
```

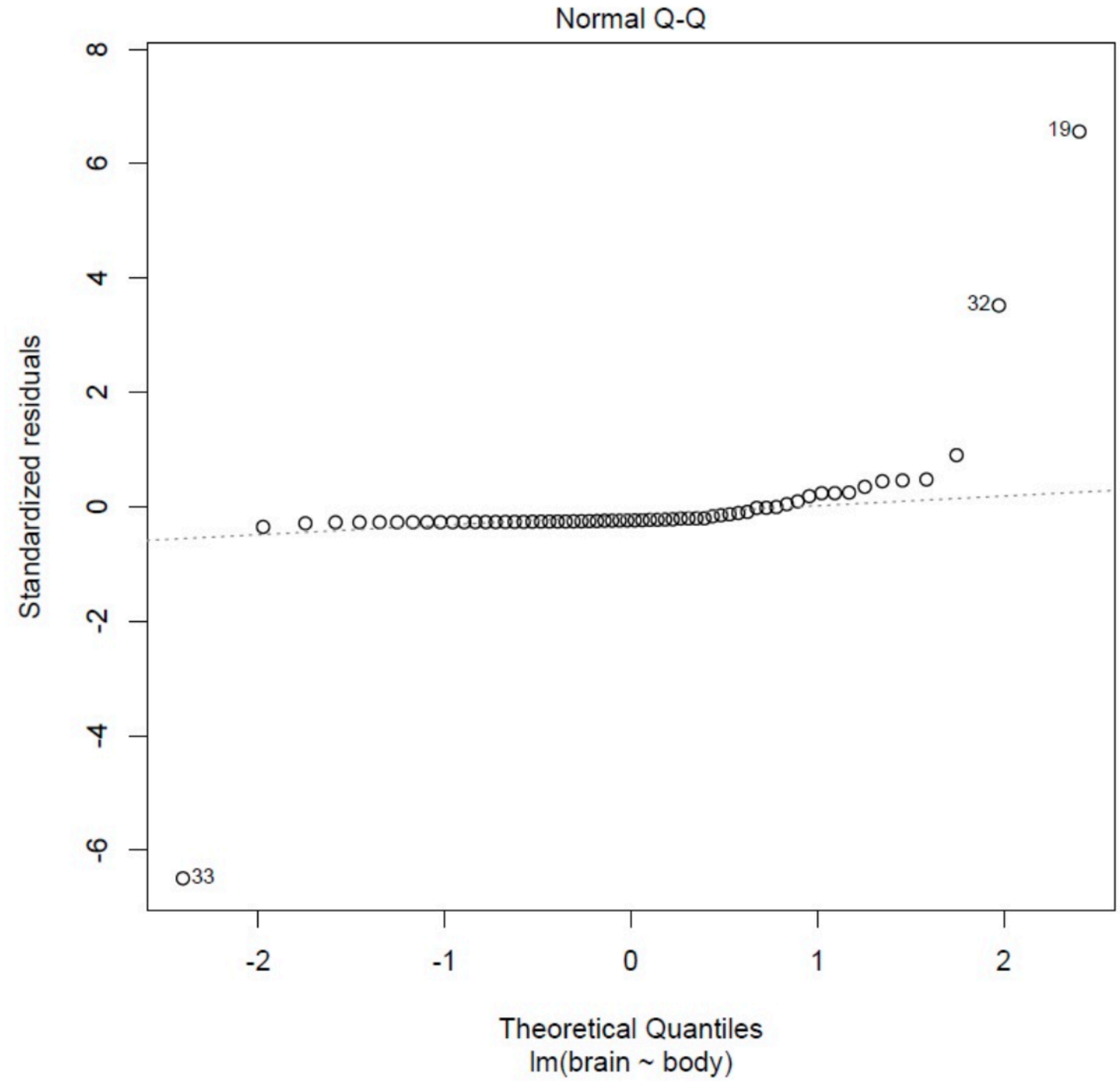


Residual Diagnostics For Mammal Data Cont.

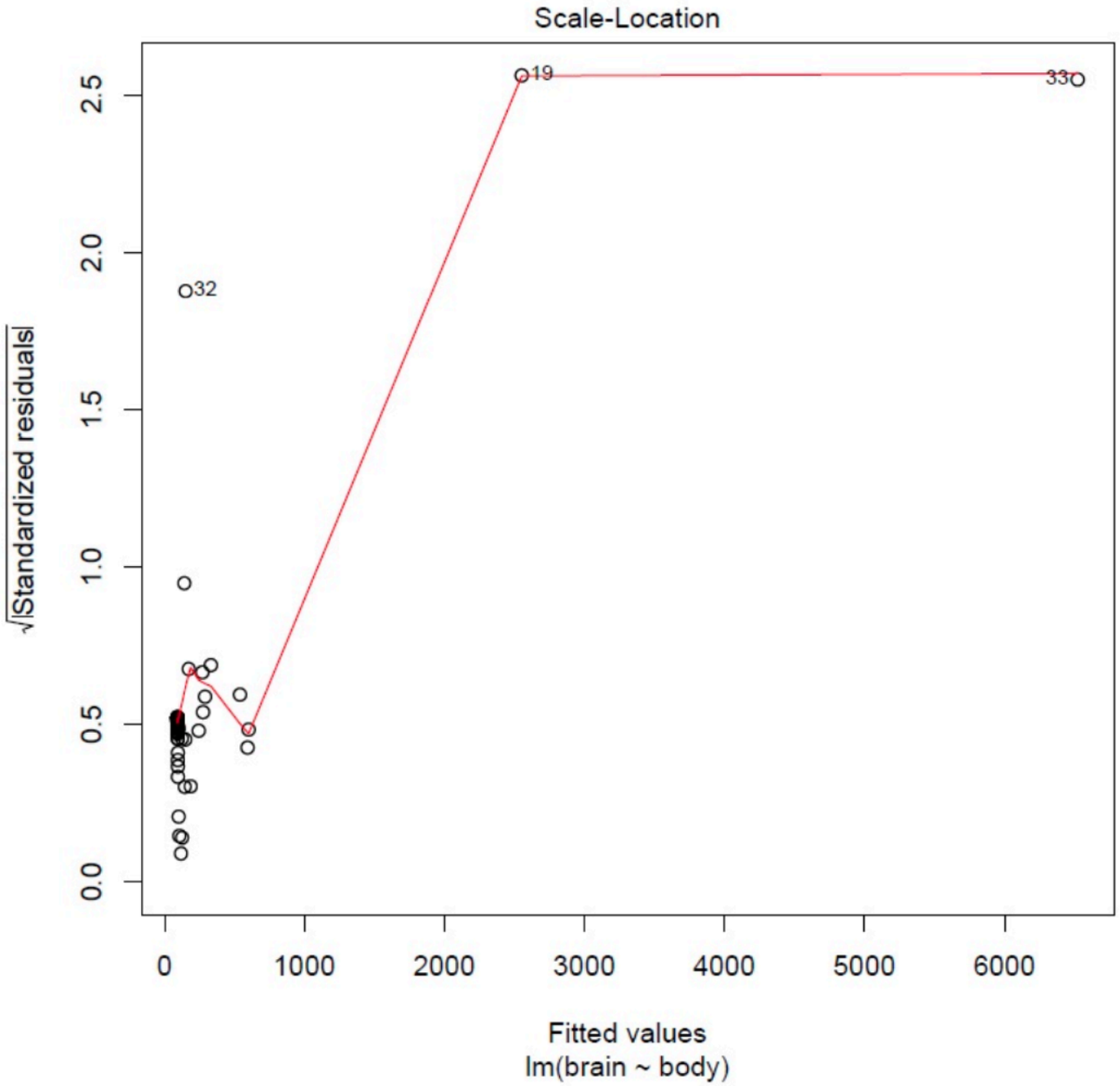
```
> plot(L)
Waiting to confirm page change...
```



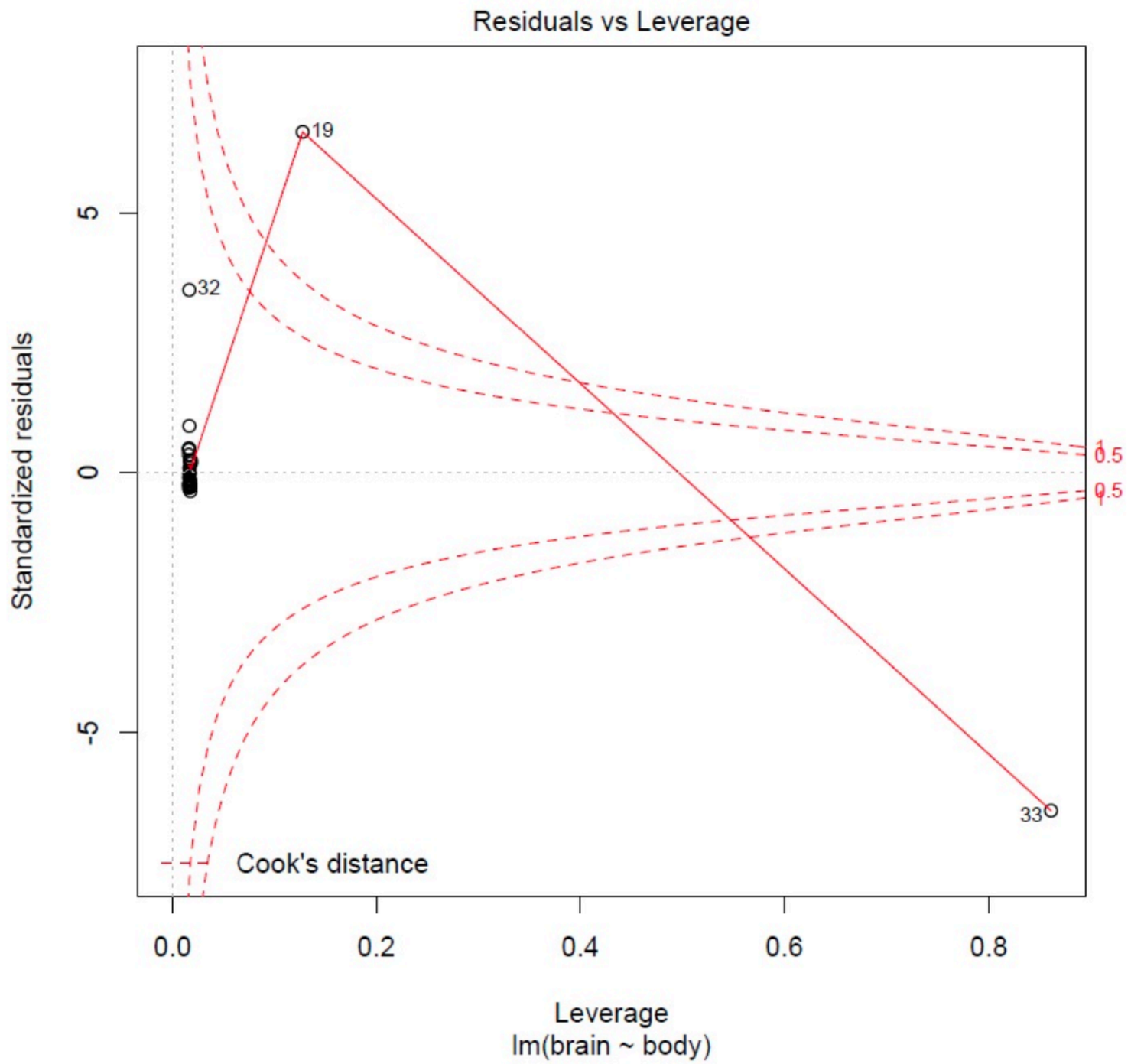
Residual Diagnostics For Mammal Data Cont.



Residual Diagnostics For Mammal Data Cont.



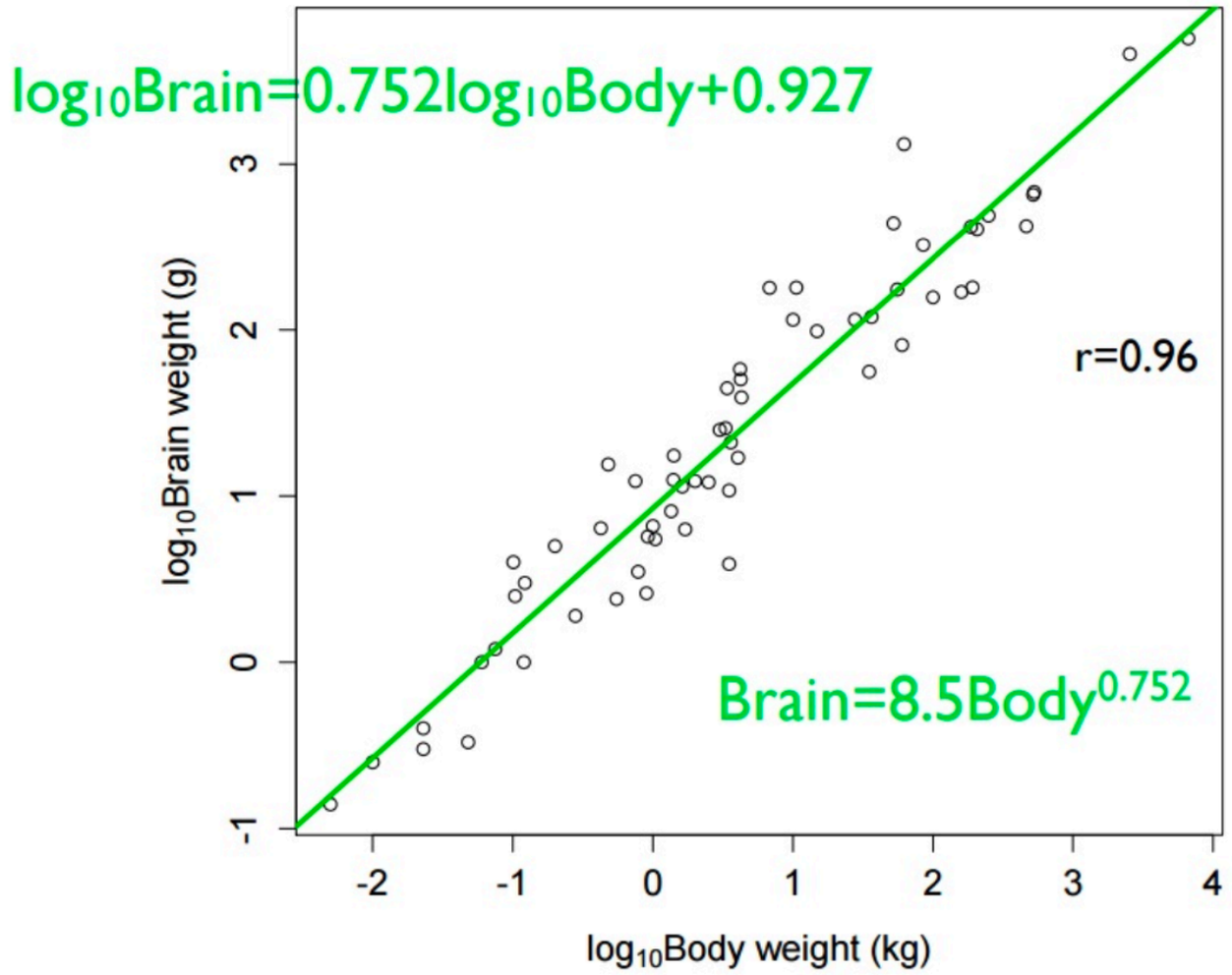
Residual Diagnostics For Mammal Data Cont.



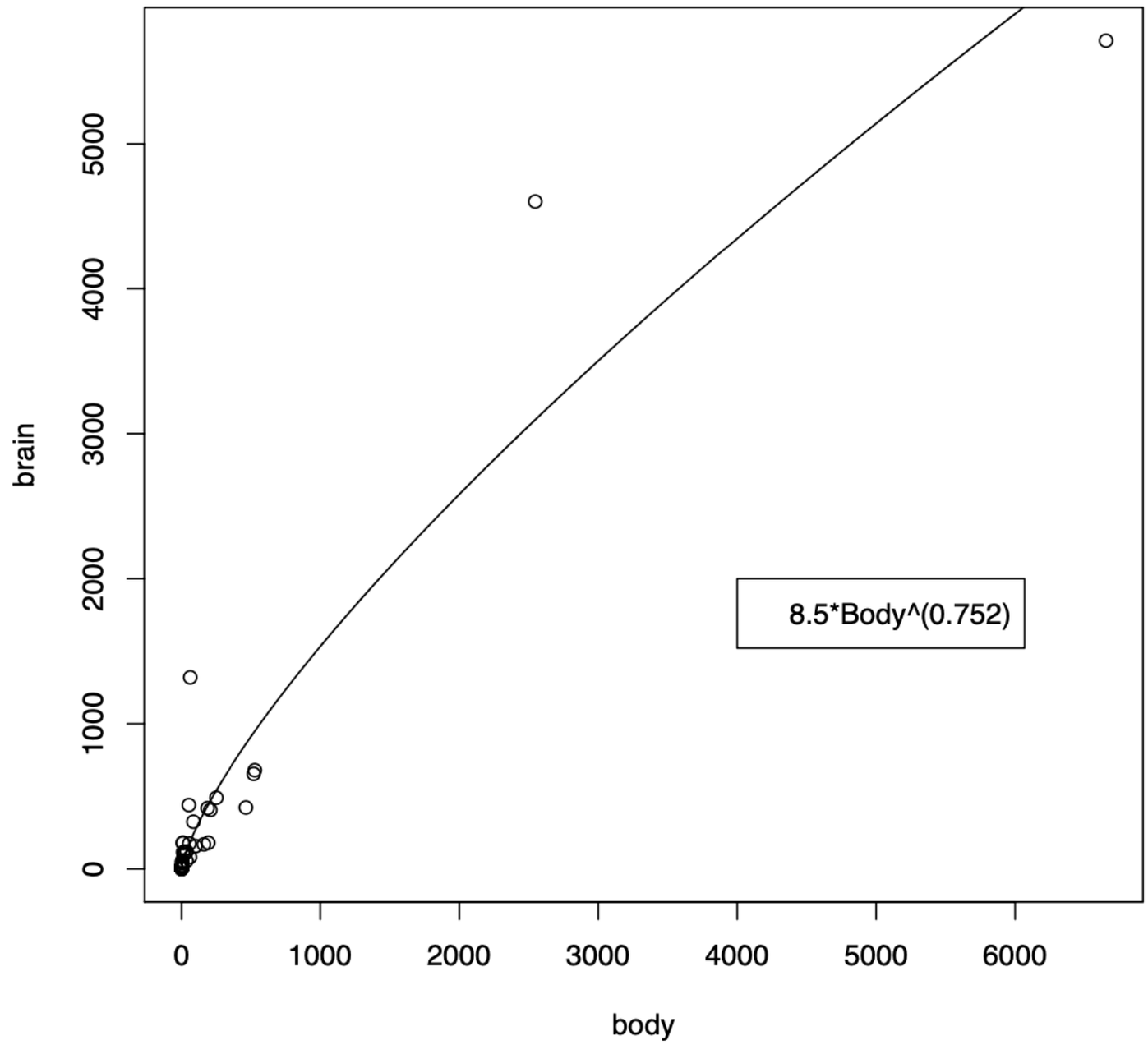
Transforming Mammals Data Cont.

Could try replacing measures with log values of those measures.

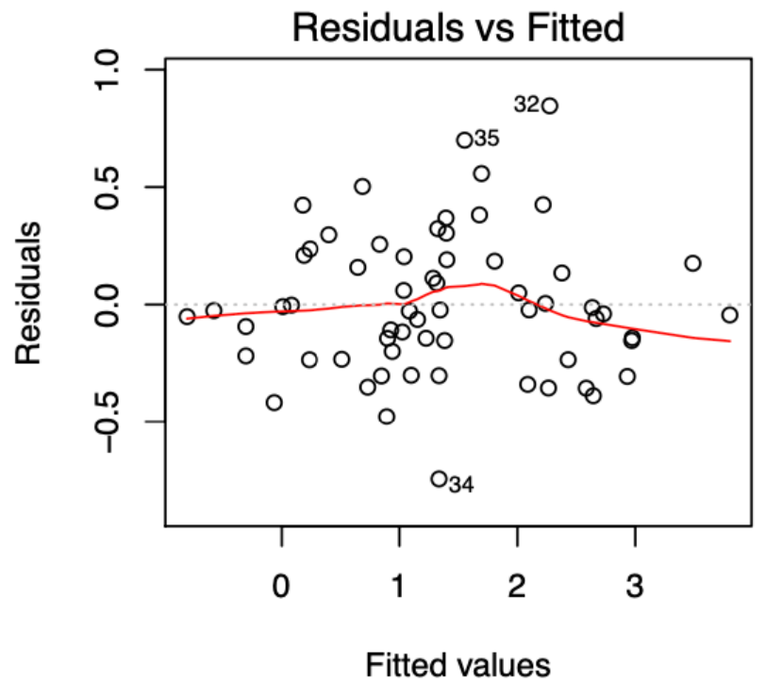
```
> brain.log<-log10(brain)
> body.log<-log10(body)
> plot(body.log, brain.log)
>
> reg2<-lm(brain.log~body.log)
> abline(reg2)
> reg2

Call:
lm(formula = brain.log ~ body.log)

Coefficients:
(Intercept)     body.log
     0.9271       0.7517
```

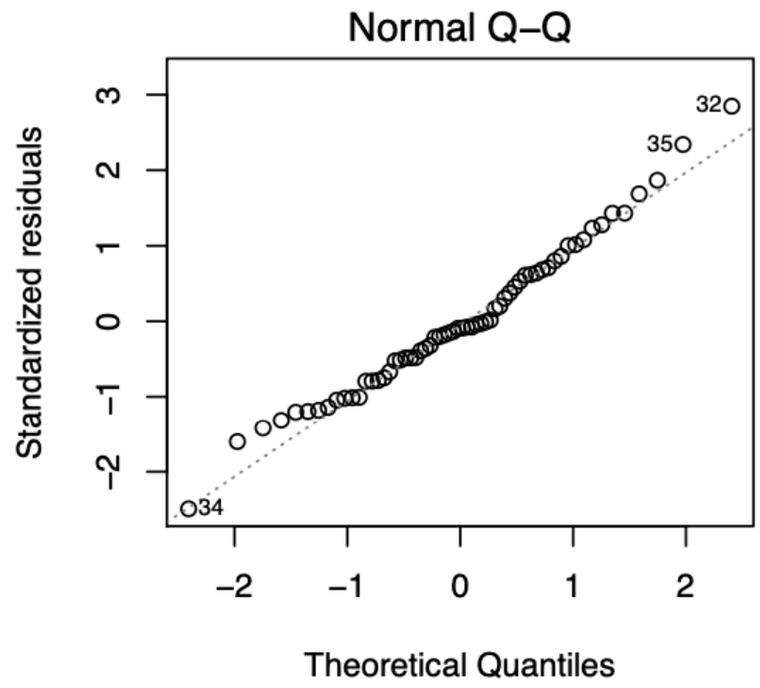


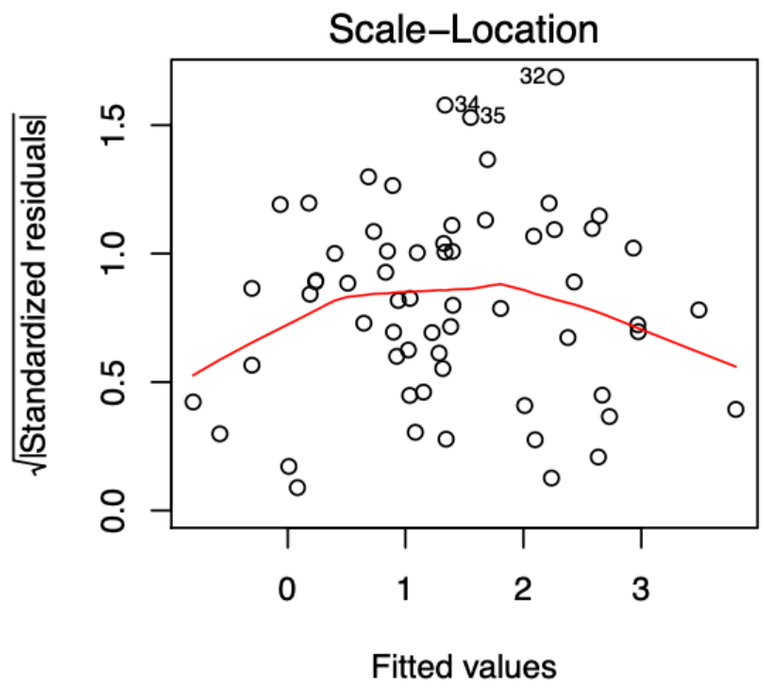
Original scatterplot

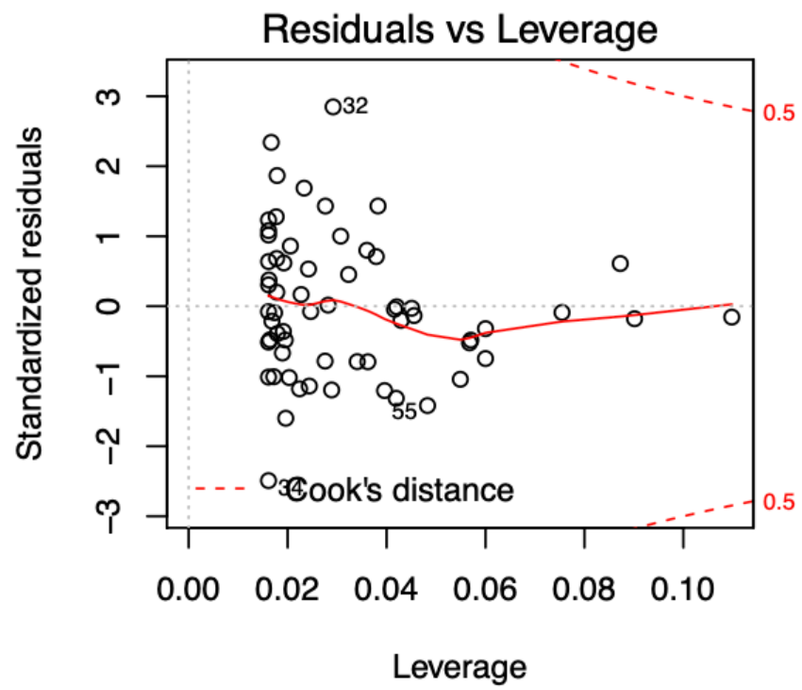


Diagnostic plots of the new model









Cook's Distance

A measure that estimates the influence of a data point when performing a least-squares regression analysis.

Given the data set $(x_1, y_1), \ldots, (x_n, y_n)$:

$\hat{y}_j = \hat{\alpha} + \hat{\beta} x_j$: fitted response value.

$s^2 := \frac{1}{n-2} \sum_{j=1}^{n} (\hat{y}_j - y_j)^2$: mean squared error of the regression model.
(can show this is unbiased estimate of the error variance in the model)

$\hat{y}_j(i) =:$ fitted response value obtained after fitting the model without the i^{th} observation (but including a fitted value for x_i from the new fit).

Then *Cook's Distance* (*Cook's D*) is given by:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_j(i))^2}{2s^2}, \ i = 1, \ldots, n$$

Cook's Distance Interpretation

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_j(i))^2}{2s^2}, \ i = 1, \ldots, n$$

D_i is a measure of how point i is influencing **all** predicted mean values $\hat{y}_j$. It's then normalised.

Large values indicate an influential observation. But how large is too large? The consensus seems to be that a D_i value of more that 1 indicates an influential value, but you may want to look at all values and investigate the ones that stick out from the other.

We might consider removing influential observations or outliers from the analysis if there is justification for doing so in the context of the scientific problem. Otherwise, we could report the analysis with and without the data point.

Cook's D For Mammals Data

We have

$$s^2 = \frac{1}{60}\sum_{i=1}^{62}(Y_i - \hat{Y}_i)^2 = 112037.3$$

Linear model for all points: $y = 0.9965x + 91.0044$.

Linear model for all points but 1st: $y = 0.9663x + 91.8609$. So

$$D_1 = \frac{\sum_{j=1}^{62}(0.9965x_j + 91.0044 - (0.9663x_j + 91.8609))^2}{2 \times 112037.3} = 0.0001934$$

Associated code for "manual" computation

```
s2<-sum((brain-fitted(Regression))^2)/(nrow(mammals)-2)

cooks.distances=vector()
for (i in 1:nrow(mammals)){
        reg.reduced<-lm(brain[-i]~body[-i])
        pred.red<-coef(reg.reduced)[1]+coef(reg.reduced)[2]*body
        cooks.distances[i]=sum((fitted(Regression)-pred.red)^2)/(2*s2)
}
cooks.distances
```

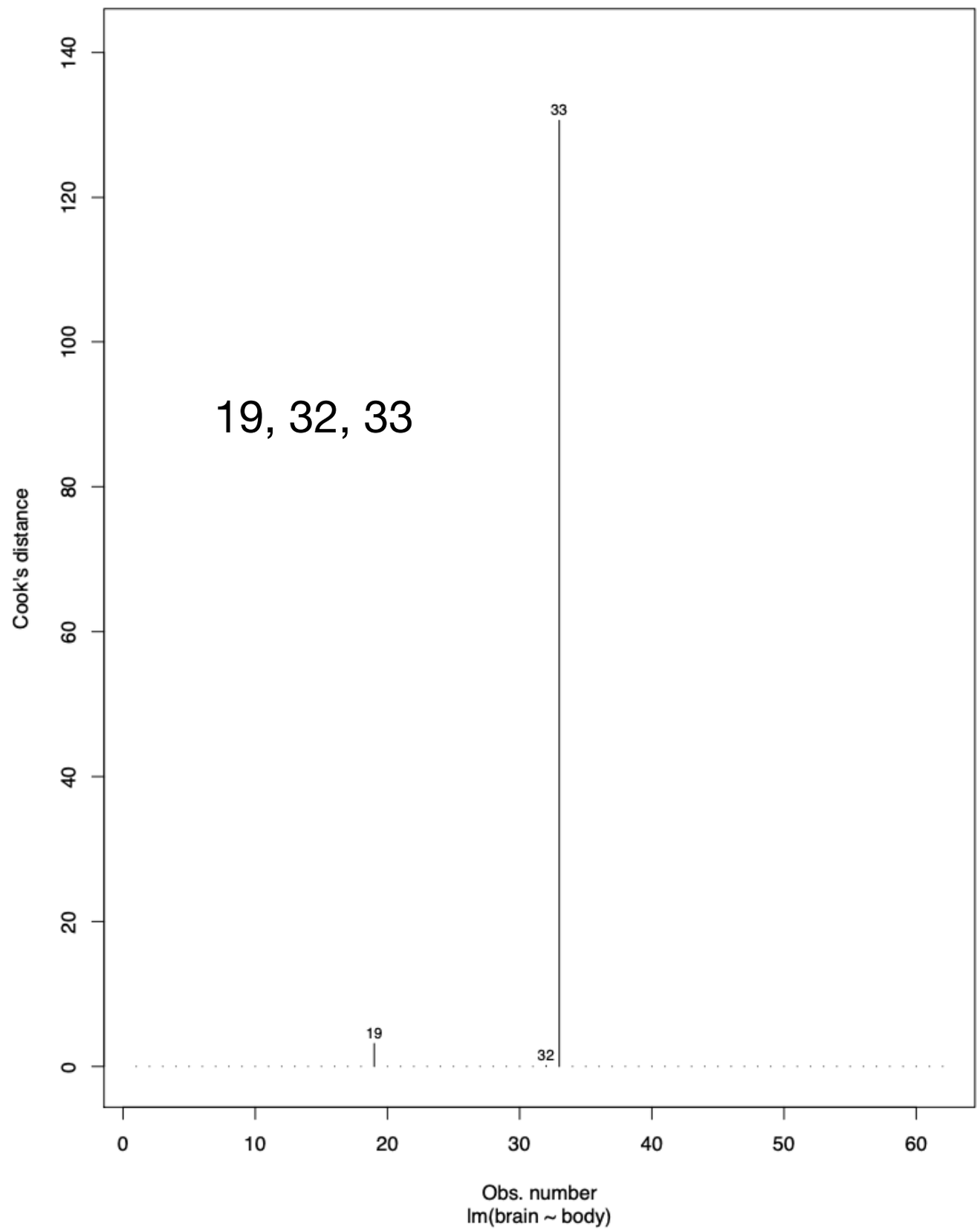
Cook's D For Mammals Data Cont.

```
> cooks.distance(Regression)
            1             2             3             4             5             6
1.933860e-04 4.510931e-04 5.540317e-04 1.119943e-03 3.367734e-06 5.785392e-07
            7             8             9            10            11            12
3.954908e-06 5.848035e-04 1.071240e-04 5.649444e-04 5.931145e-04 5.805735e-04
           13            14            15            16            17            18
5.695214e-04 6.455597e-04 6.341294e-04 5.453056e-04 5.078680e-04 5.824256e-04
           19            20            21            22            23            24
3.156125e+00 6.435294e-04 1.610721e-03 3.085204e-04 6.088278e-04 1.597883e-05
           25            26            27            28            29            30
3.672598e-04 5.808344e-04 4.379110e-04 5.130252e-04 9.823146e-04 1.743314e-03
           31            32            33            34            35            36
4.930657e-04 1.041011e-01 1.306176e+02 6.390499e-04 5.166193e-04 3.645885e-04
           37            38            39            40            41            42
4.737279e-04 6.349275e-04 6.421116e-04 6.440647e-04 4.982377e-04 1.847855e-03
           43            44            45            46            47            48
5.163850e-04 7.034557e-05 7.075959e-05 6.814559e-03 4.768863e-04 6.210010e-04
           49            50            51            52            53            54
3.523865e-04 4.214277e-04 2.442171e-04 6.244191e-04 6.315151e-04 6.070849e-04
           55            56            57            58            59            60
6.434387e-04 6.935297e-04 3.706414e-04 4.376644e-04 6.228221e-04 5.146961e-04
           61            62
6.137580e-04 1.558739e-04
```

```
> cooks.distance(Regression)[c(19,32,33)]
        19          32          33
 3.1561247   0.1041011 130.6176034
```

Cook's distance plot

```
> plot(L,which = 4)
```



What to do if we identify an issue?

These diagnostic plots are not a strict "go" or "stop" sign. It can tell you several things about the data.
You may want to rethink your model and hypotheses. You may want to:

- Transform variables
- Add new variables in the model
- Remove a few influential points
- Need better or different data collection methods, because of systematic bias in the data
- Possibly other things.

Examples for Linear Regression Fit and Diagnostics

Taken from the book by John Rice,
Mathematical Statistics and Data Analysis, Duxbury Press

1. **Yellow dye quantification by chromatography**
2. **Stream depth and flow**
3. **Breast cancer mortality in 301 countries**

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

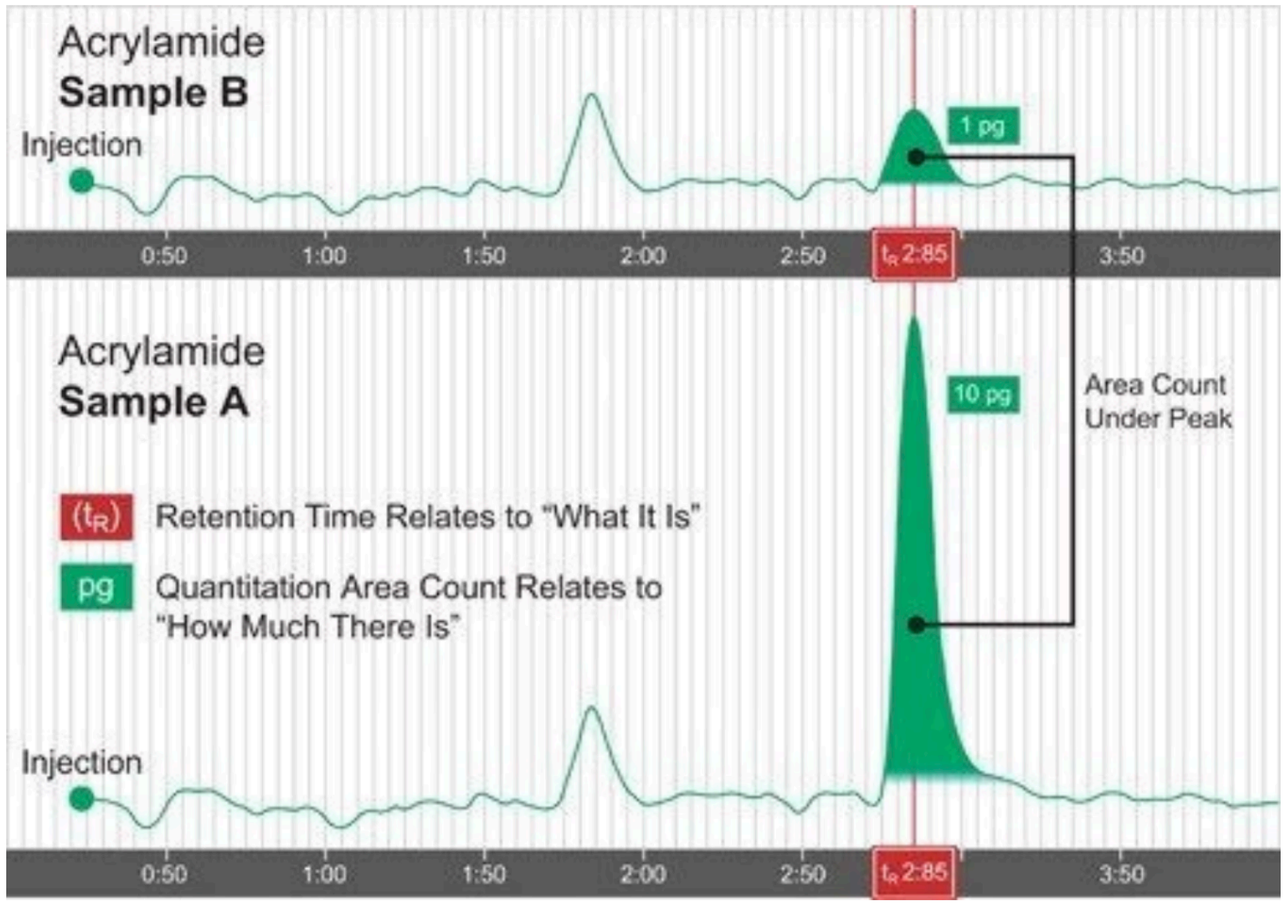


Figure I-2. Identification and Quantitation.

- Three dye compounds are represented by three peaks separated in time in the chromatogram.
- Each elutes at a specific location.
- Is the **area under the peak** linked to relative **amount of the dye**?

John Rice: Mathematical Statistics and Data Analysis, Duxbury Press
Chromatography figure source: https://www.waters.com/nextgen/us/en/education/primers/beginner-s-guide-to-liquid-chromatography/identifying-and-quantitating-compounds.html

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

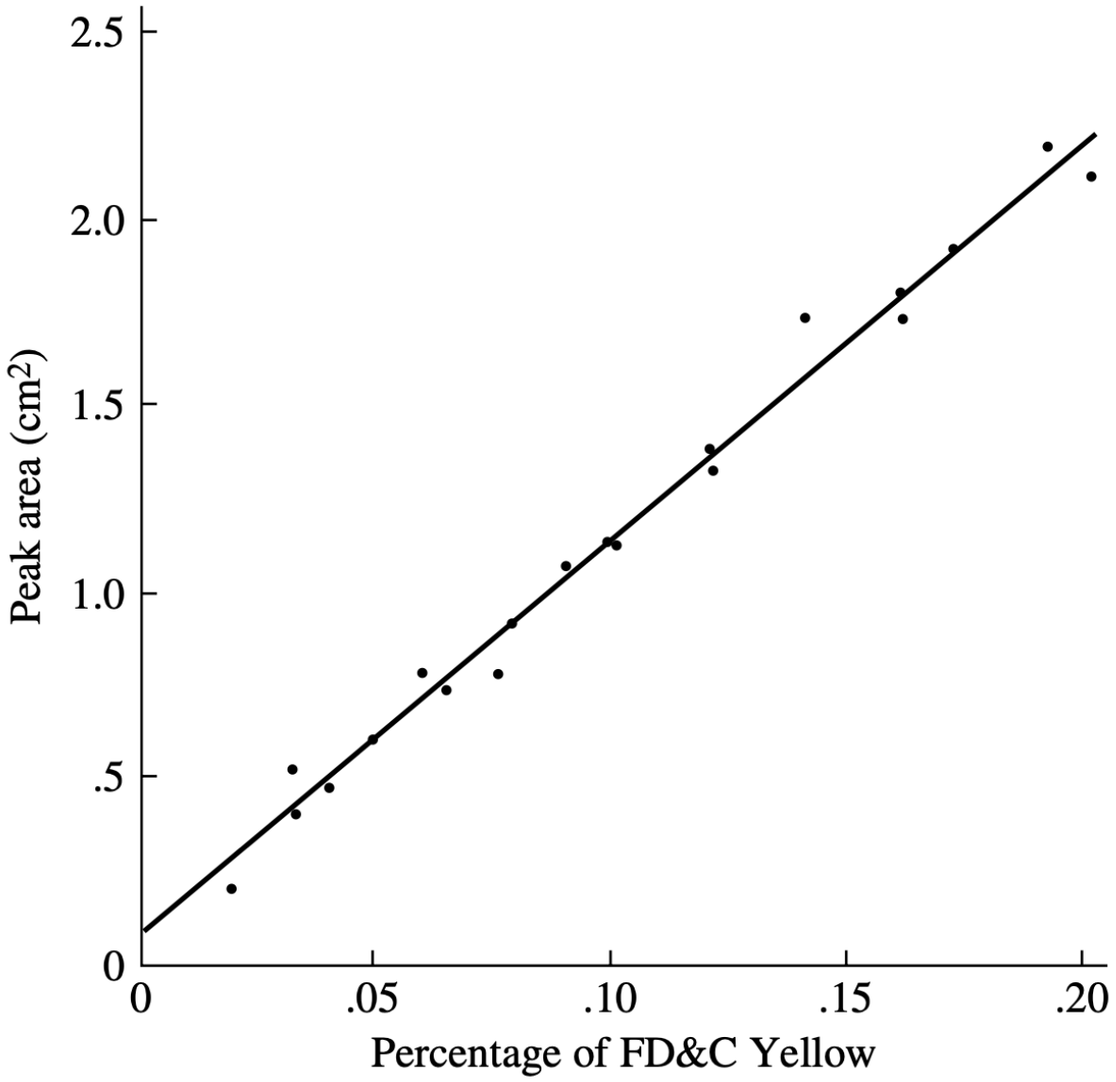


FIGURE **14.2** Data points and the least squares line for the relation of sulfanilic acid peak area to percentage of FD&C Yellow.

1. Curves are often fit to data as part of the process of calibrating instruments. For example, Bailey, Cox, and Springer (1978) discuss a method for measuring the concentrations of food dyes and other substances by high-pressure chromatography. Measurements of the chromatographic peak areas corresponding to sulfanilic acid were taken for several known concentrations of FD&C Yellow No. 5.

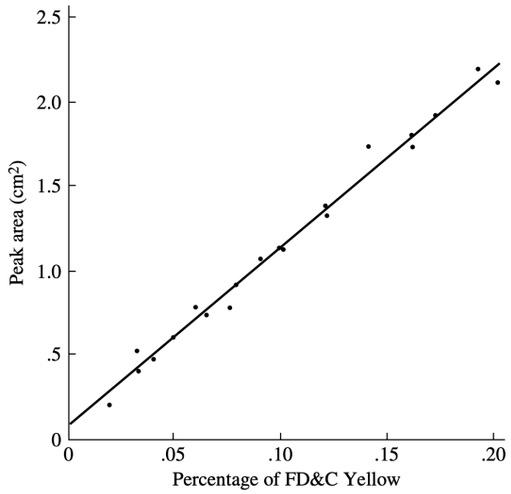


FIGURE **14.2** Data points and the least squares line for the relation of sulfanilic acid peak area to percentage of FD&C Yellow.

Residual plot indicates good fit (no pattern)

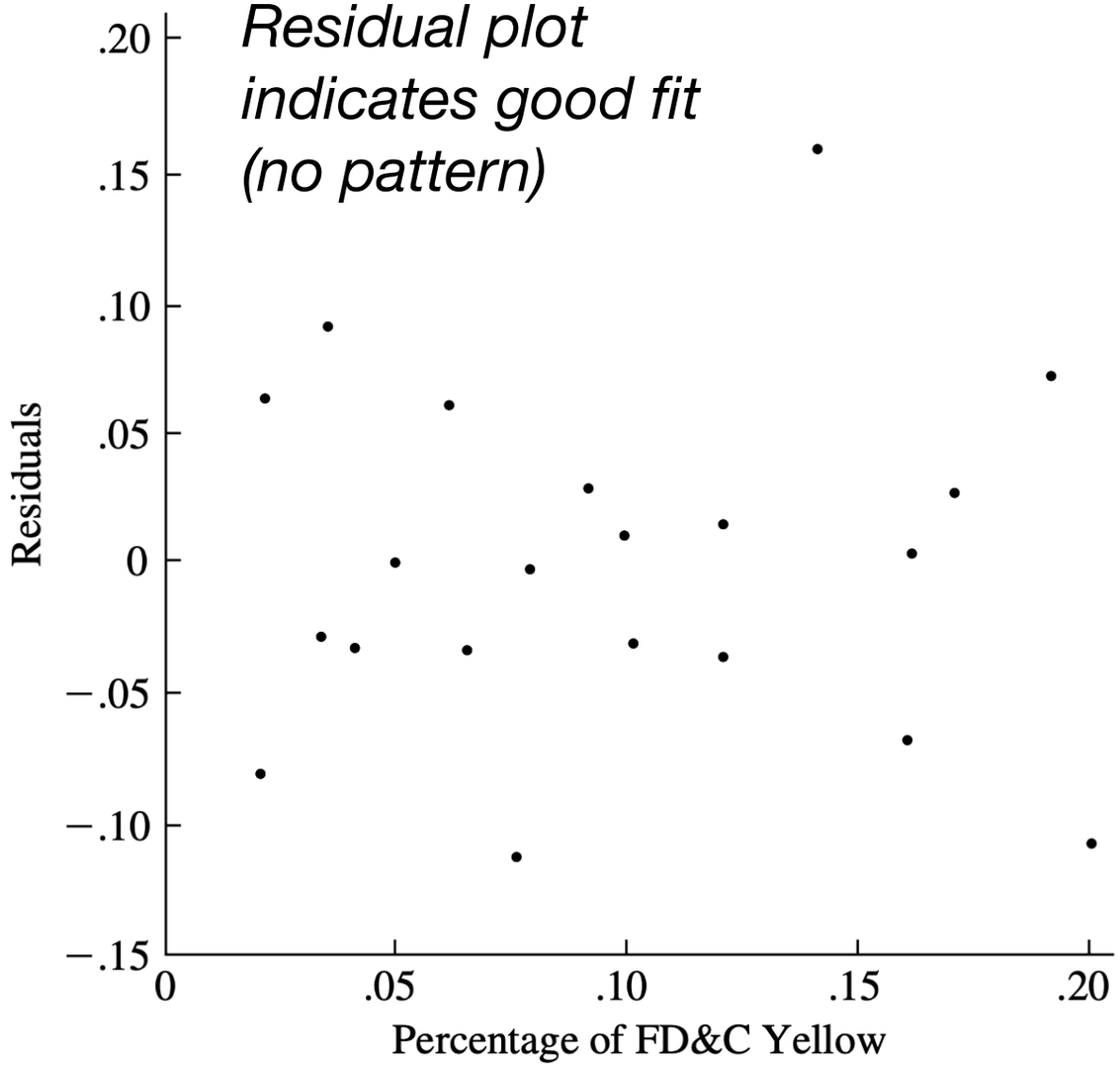


FIGURE **14.5** A plot of residuals for the data on chromatographic peak area.

2. The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

Depth	Flow Rate
.34	.636
.29	.319
.28	.734
.42	1.327
.29	.487
.41	.924
.76	7.350
.73	5.890
.46	1.979
.40	1.124

2. The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

A plot of flow rate versus depth suggests that the relation is not linear (Figure 14.6).

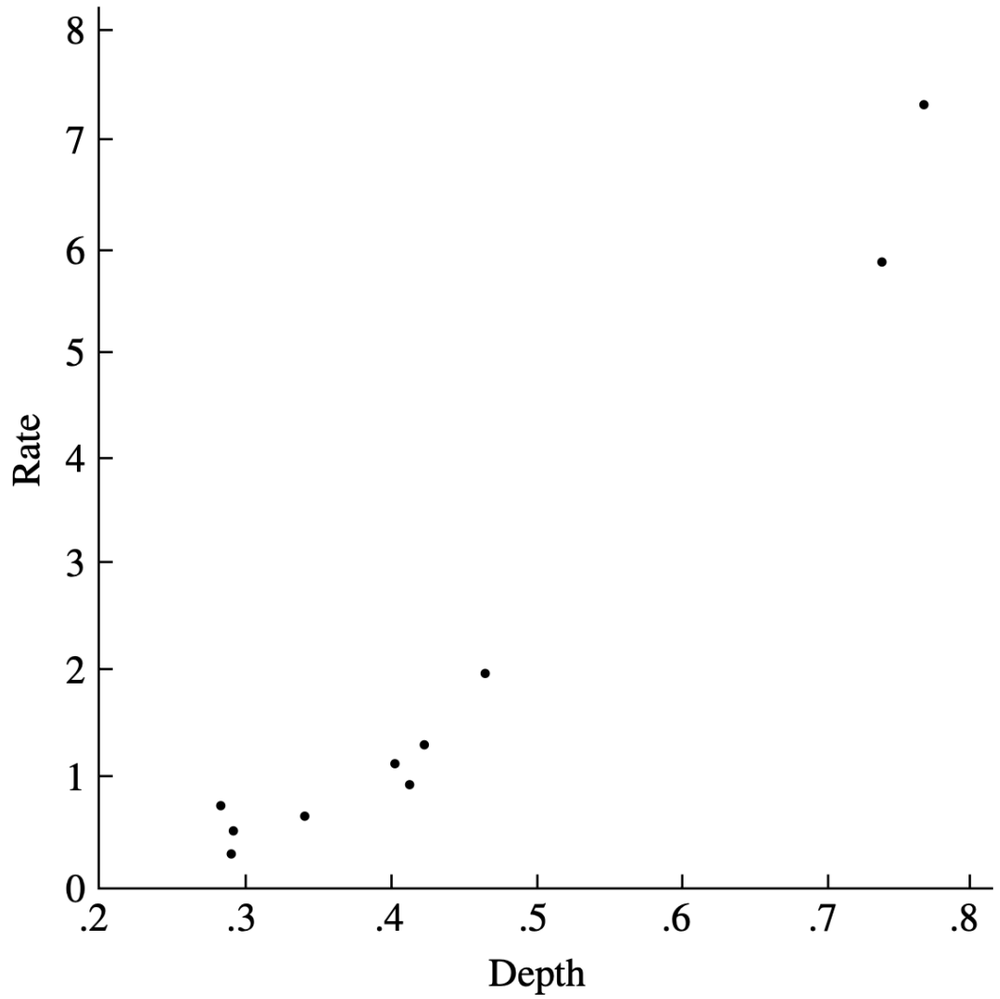


FIGURE **14.6** A plot of flow rate versus stream depth.

2. The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of a stream and the rate of its flow (Ryan, Joiner, and Ryan 1976).

A plot of flow rate versus depth suggests that the relation is not linear (Figure 14.6). This is even more immediately apparent from the bowed shape of the plot of the residuals versus depth (Figure 14.7).

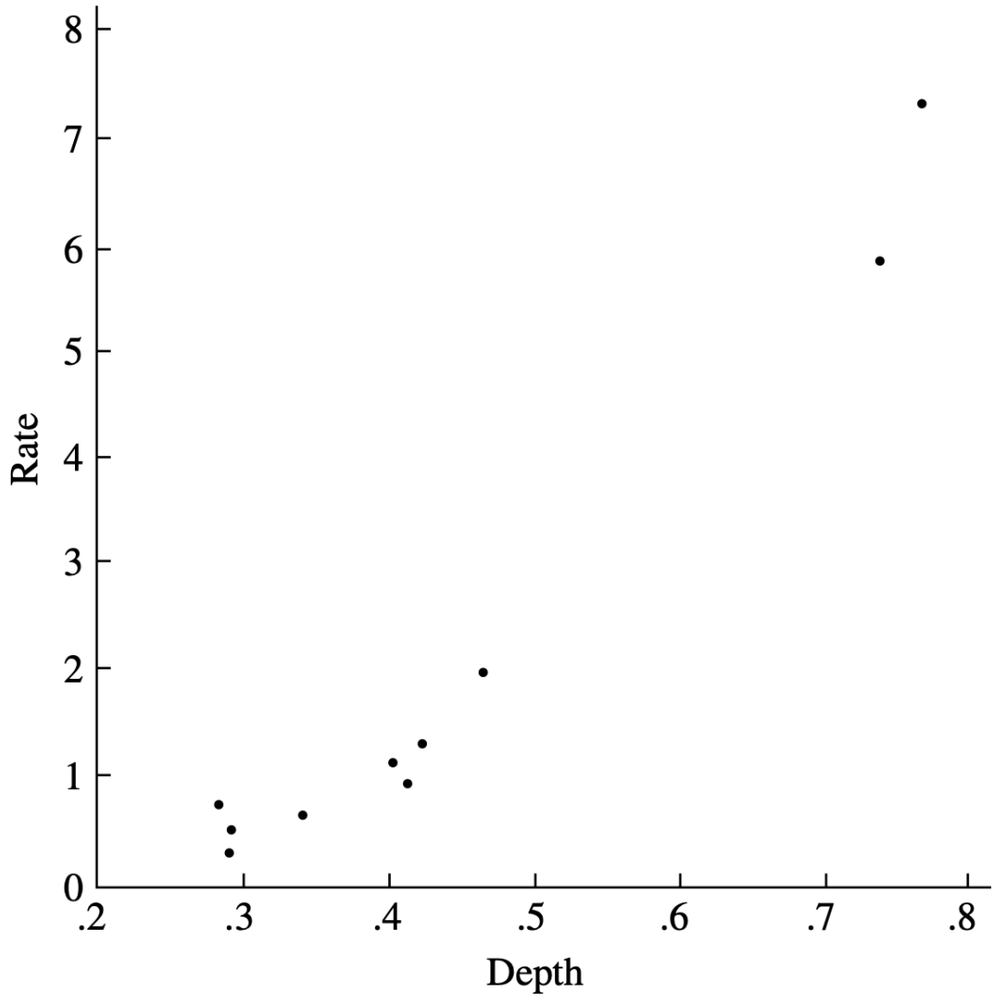


FIGURE **14.6** A plot of flow rate versus stream depth.

Residual plot is imbalanced.

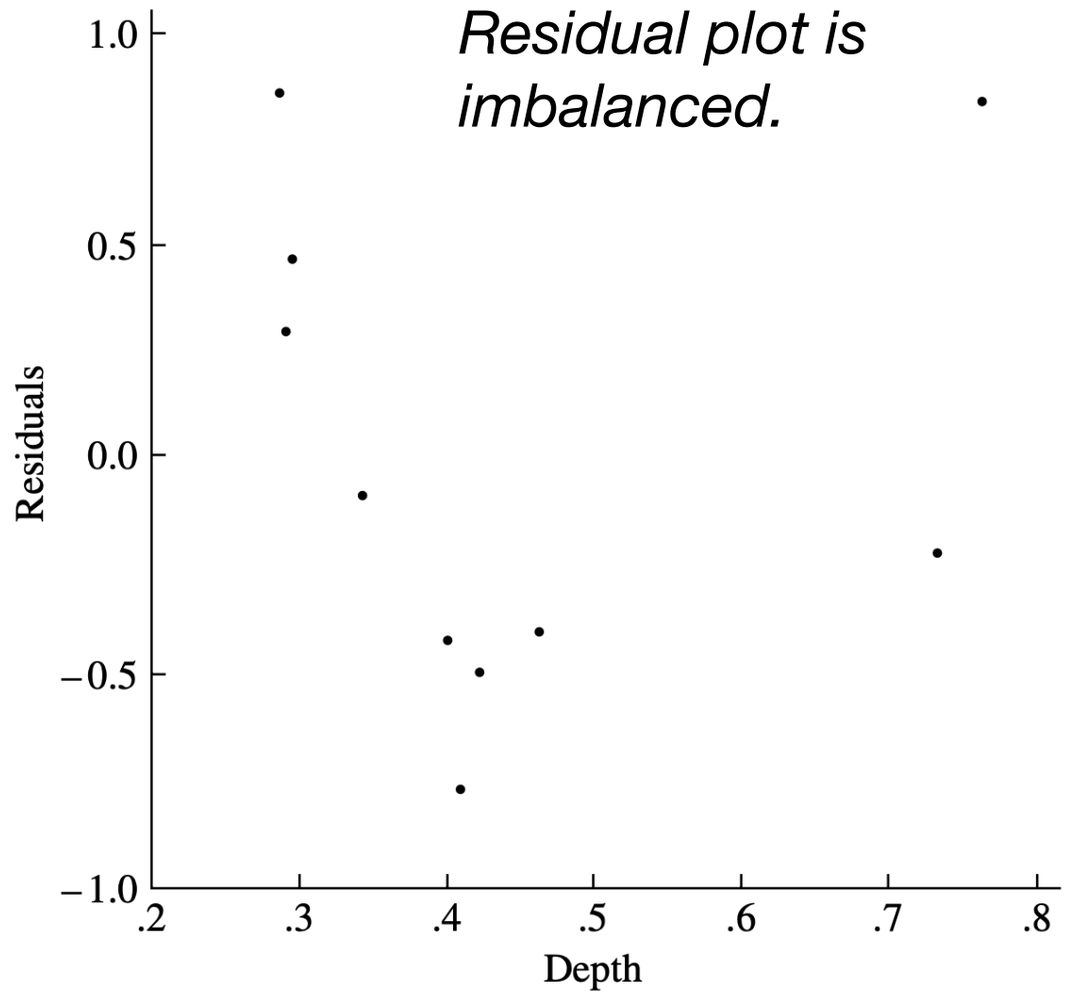


FIGURE **14.7** Residuals from the regression of flow rate on depth.

Use of log transforms: In order to empirically linearize relationships, transformations are frequently employed. Figure 14.8 is a plot of log rate versus log depth, and Figure 14.9 shows the residuals for the corresponding fit. There is no sign of obvious misfit.

Residual plot indicates good fit (no pattern)

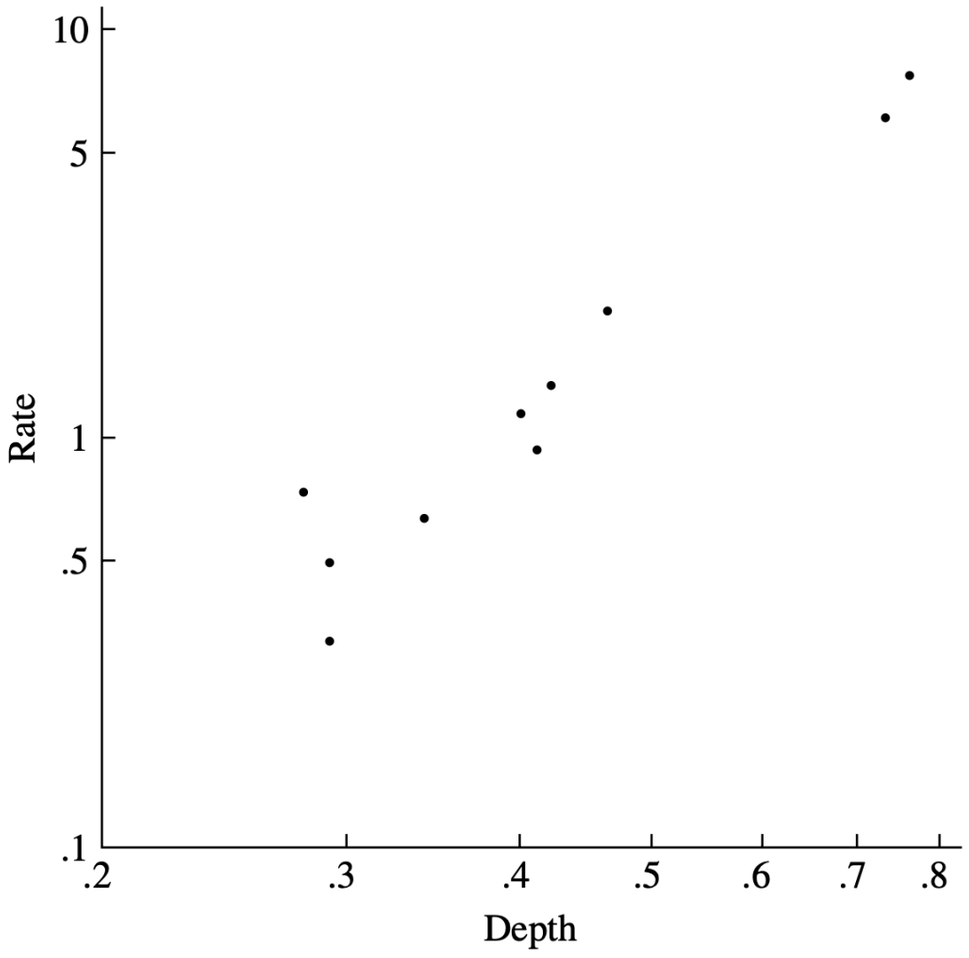


FIGURE **14.8** Plot of log flow rate versus log depth.

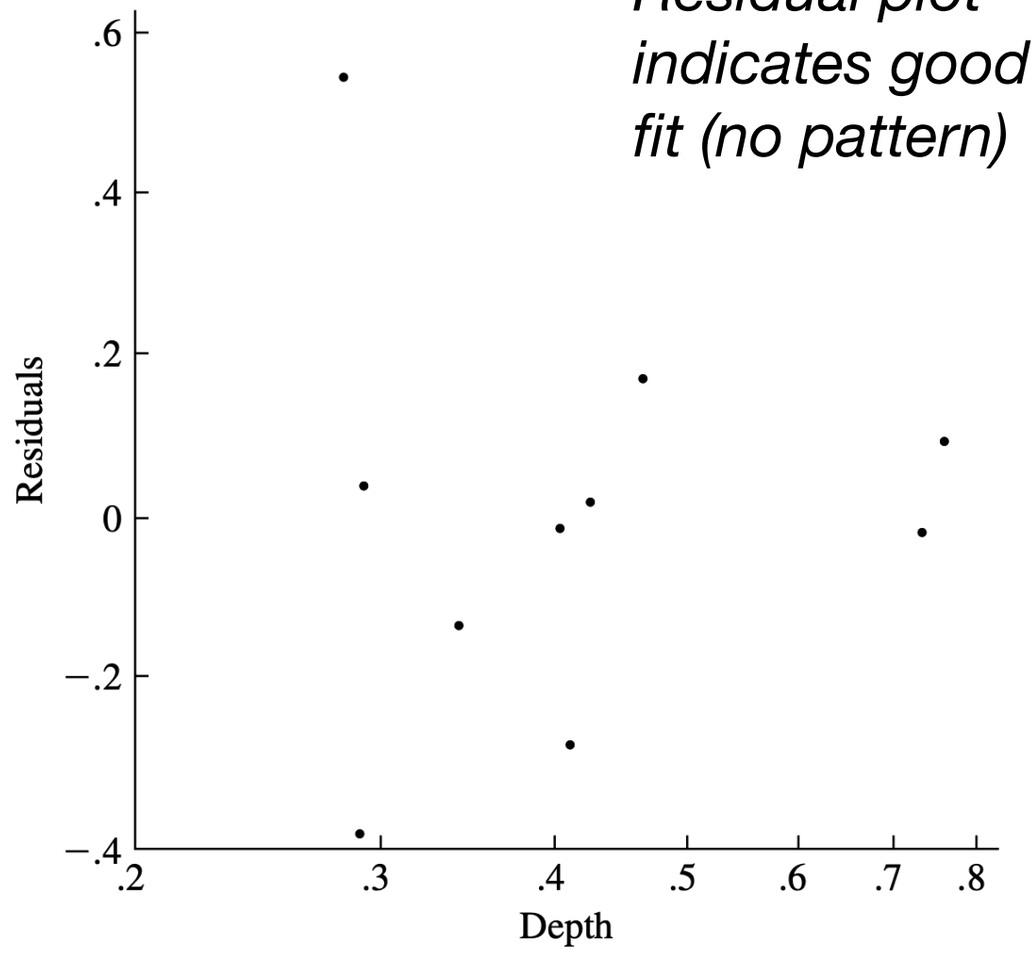


FIGURE **14.9** Residuals from the regression of log flow rate on log depth.

3. Breast cancer mortality

A scatterplot of the number of cases (y) versus population (x) is shown in Figure 14.10. This plot appears to be consistent with the simple model that the number of cases is proportional to the population size, or $y \approx \beta x$. (We will test whether or not the intercept is zero below.) Accordingly, we fit a model with zero intercept by least squares to the data, yielding $\hat{\beta} = 3.559 \times 10^{-3}$. (See Problem 15 at the end of this chapter for fitting a zero intercept model.) Figure 14.11 shows the residuals from the regression of the number of cases on population plotted versus population. Since it is very hard to see what is going on in the left-hand side of this plot, the residuals are plotted versus log population in Figure 14.12, from which it is quite clear that the error variance is not constant but grows with population size.

The residual plot in Figure 14.12 shows no curvature but indicates that the variance is not constant. For counted data, the variability often grows with the mean, and frequently a square root transformation is used in an attempt to stabilize the variance. We therefore fit a model of the form $\sqrt{y} \approx \gamma\sqrt{x}$. Figure 14.13 shows the plot of residuals for this fit. The residual variability is more nearly constant

3. Breast cancer mortality in 301 countries

A scatterplot of the number of cases (y) versus population (x) is shown in Figure 14.10. This plot appears to be consistent with the simple model that the number of cases is proportional to the population size, or $y \approx \beta x$.

Can fit model with
Intercept 0 and
slope 3.56/10000

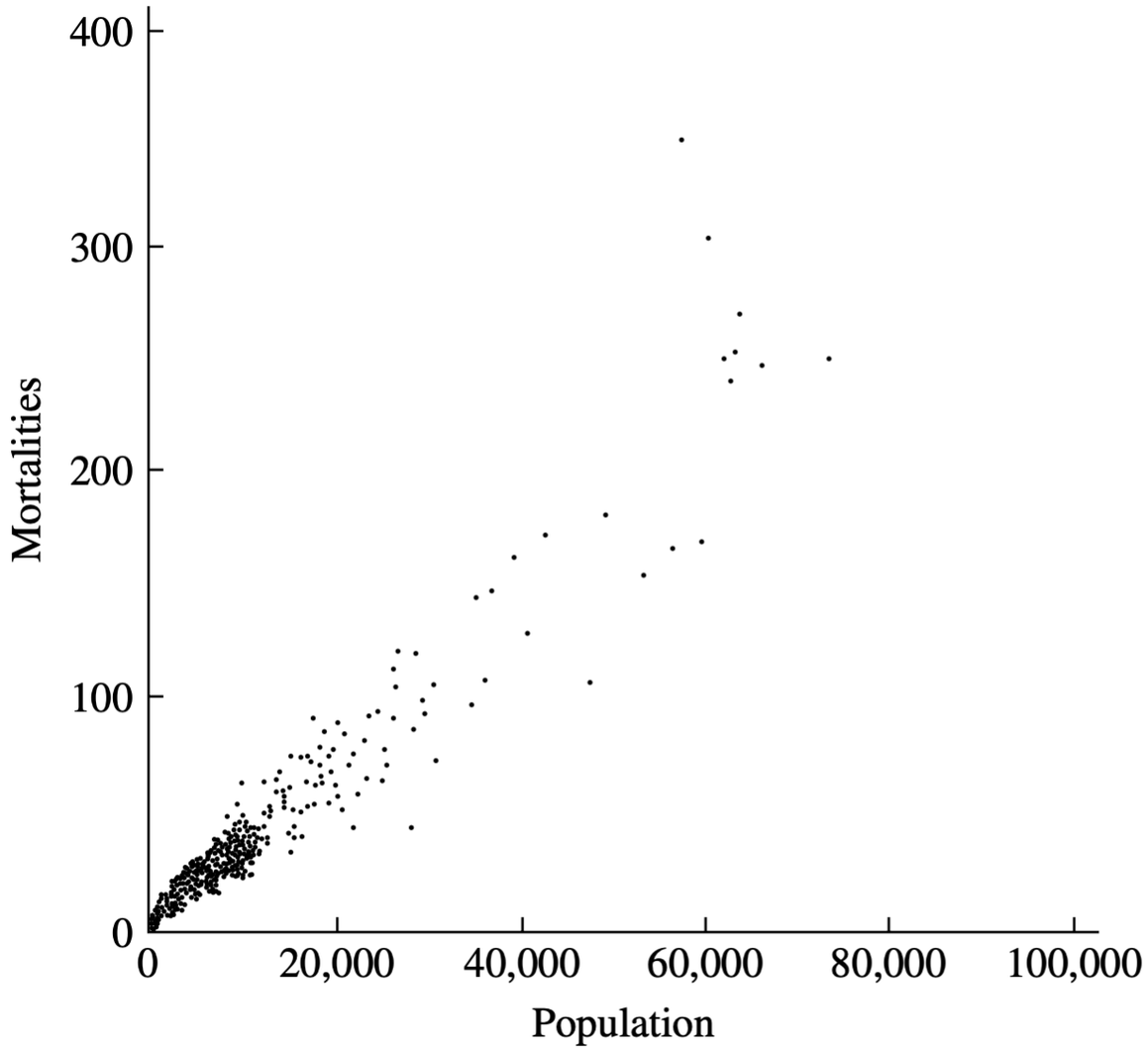


FIGURE **14.10** Scatterplot showing breast cancer mortality versus population.

The residual plot in Figure 14.12 shows no curvature but indicates that the variance is not constant.

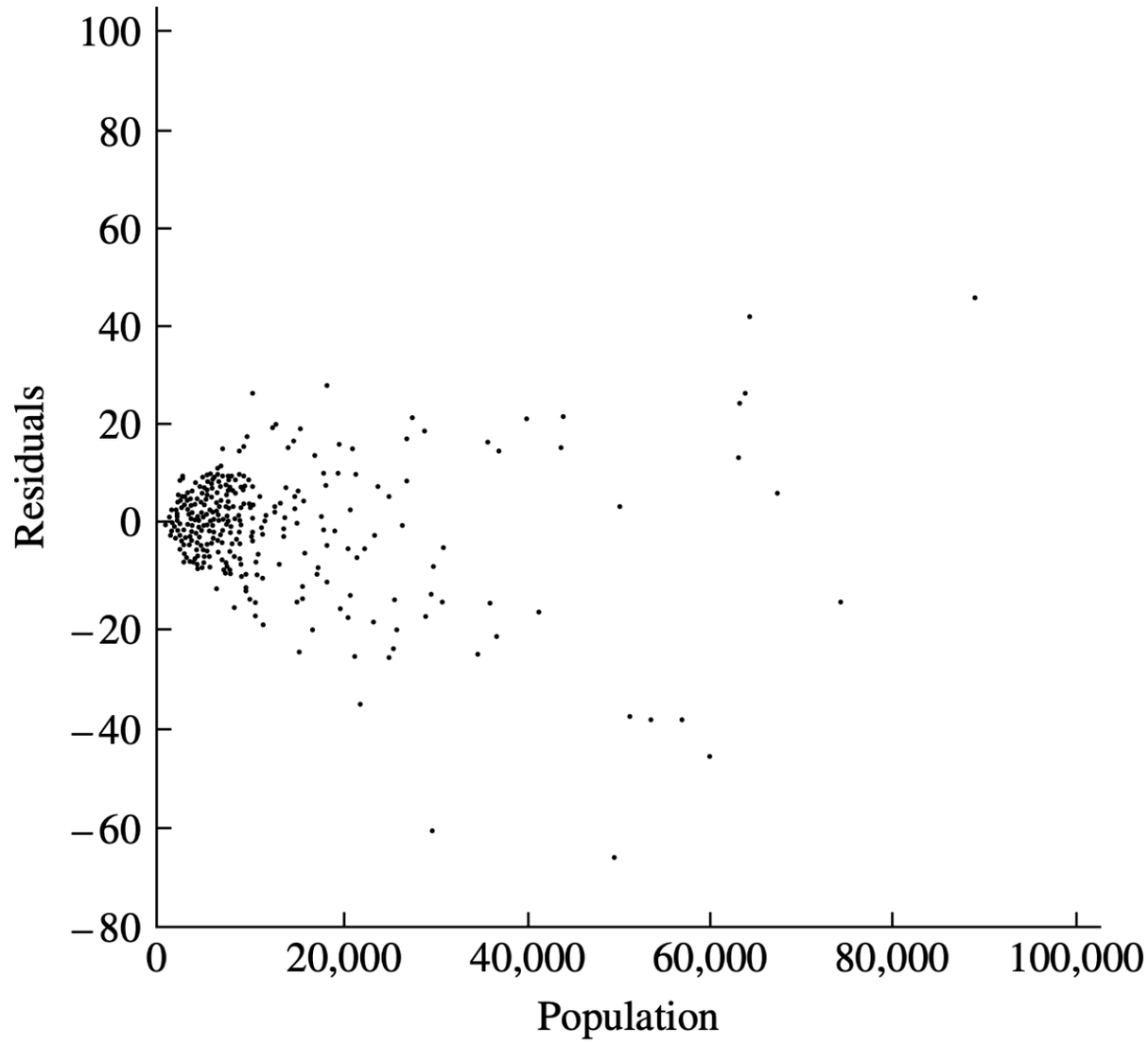
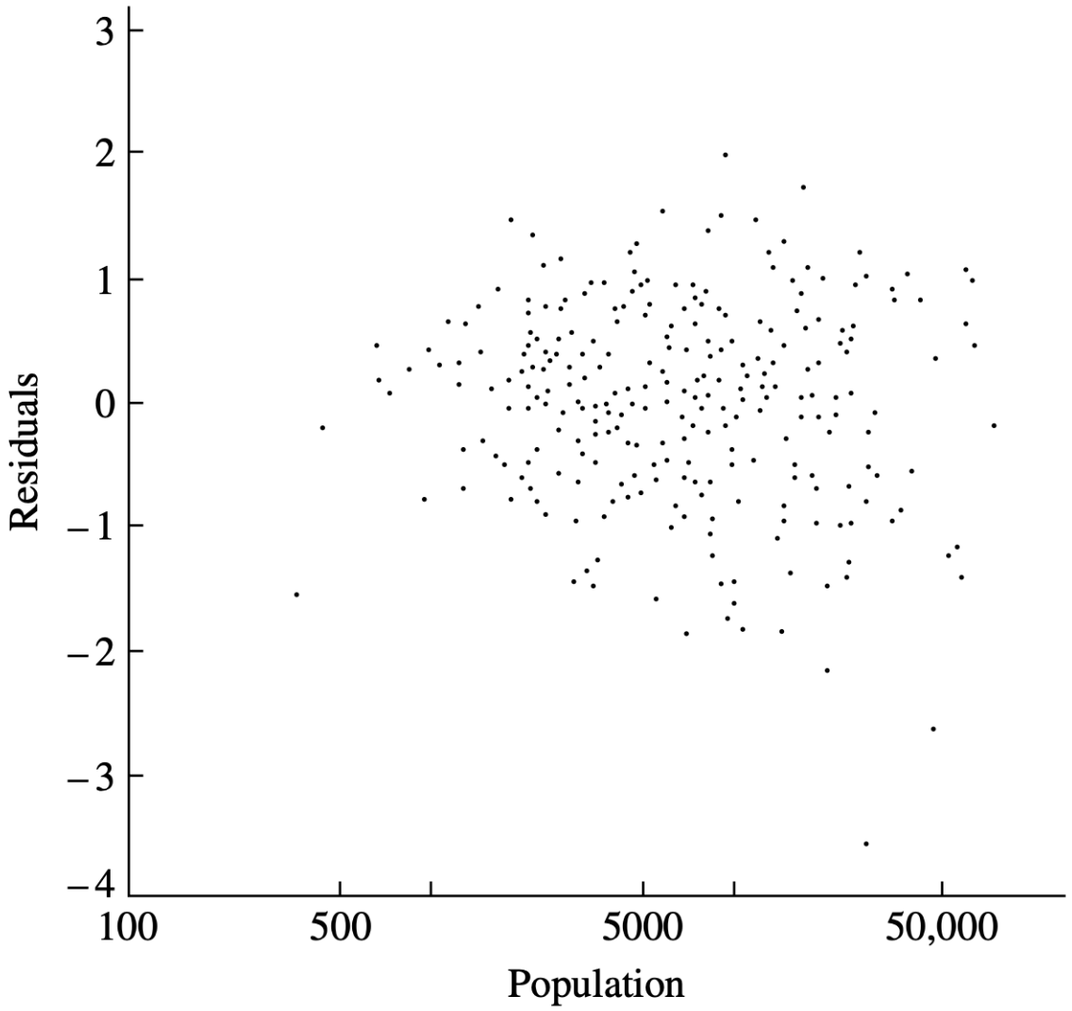


FIGURE **14.11** Residuals from the regression of mortality on population. pulation.

Residual plot has no slope or other trends, but shows increased variance as population grows.

The is called
heteroscedasticity

The residual plot in Figure 14.12 shows no curvature but indicates that the variance is not constant. For counted data, the variability often grows with the mean, and frequently a square root transformation is used in an attempt to stabilize the variance. We therefore fit a model of the form $\sqrt{y} \approx \gamma\sqrt{x}$. Figure 14.13 shows the plot of residuals for this fit. The residual variability is more nearly constant here; β is estimated by the square of the slope, $\hat{\gamma}$, which for this example gives $\tilde{\beta} = \hat{\gamma}^2 = 3.471 \times 10^{-3}$.



Residual plot has not slope or other trends, but shows increased variance as population grows.

FIGURE **14.13** Residuals from the regression of the square root of mortality on the square root of population.