The Statistical Sleuth in R: Chapter 8

Kate Aloisio   Ruobing Zhang   Nicholas J. Horton[*]

September 30, 2013

Contents

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Second Edition of the *Statistical Sleuth* (2002) by Fred Ramsey and Dan Schafer. More information about the book can be found at `http://www.proaxis.com/~panorama/home.htm`. This file as well as the associated knitr reproducible analysis source file can be found at `http://www.amherst.edu/~nhorton/sleuth`.

This work leverages initiatives undertaken by Project MOSAIC (`http://www.mosaic-web.org`), an NSF-funded effort to improve the teaching of statistics, calculus, science and computing in the undergraduate curriculum. In particular, we utilize the mosaic package, which was written to simplify the use of R for introductory statistics courses. A short summary of the R needed to teach introductory statistics can be found in the mosaic package vignette (`http://cran.r-project.org/web/packages/mosaic/vignettes/MinimalR.pdf`).

To use a package within R, it must be installed (one time), and loaded (each session). The package can be installed using the following command:

```
> install.packages("mosaic")  # note the quotation marks
```

Once this is installed, it can be loaded by running the command:

```
> require(mosaic)
```

This needs to be done once per session.

In addition the data files for the *Sleuth* case studies can be accessed by installing the Sleuth2 package.

```
> install.packages("Sleuth2")  # note the quotation marks
```

[*] Department of Mathematics, Amherst College, nhorton@amherst.edu

```
> require(Sleuth2)
```

We also set some options to improve legibility of graphs and output.

```
> trellis.par.set(theme = col.mosaic())  # get a better color scheme for lattice
> options(digits = 4)
```

The specific goal of this document is to demonstrate how to calculate the quantities described in Chapter 8: A Closer Look at Assumptions for Simple Linear Regression using R.

2 Island Area and Number of Species

What is the relationship between the area of islands and the number of animal and plant species living on them? This is the question addressed in case study 8.1 in the *Sleuth*.

2.1 Summary statistics and graphical display

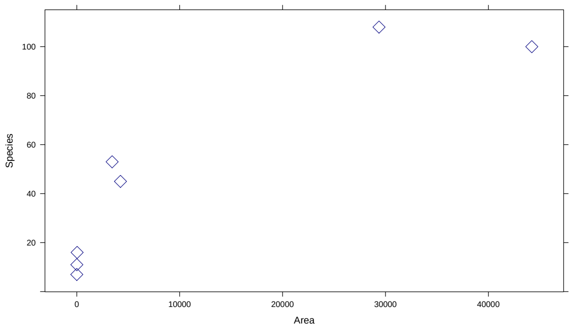
We begin by reading the data and summarizing the variables.

```
> case0801

             Area Species
Cuba        44218     100
Hispaniola  29371     108
Jamaica      4244      45
Puerto Rico  3435      53
Montserrat     32      16
Saba            5      11
Redonda         1       7

> summary(case0801)

      Area          Species     
 Min.   :    1   Min.   :  7.0  
 1st Qu.:   18   1st Qu.: 13.5  
 Median : 3435   Median : 45.0  
 Mean   :11615   Mean   : 48.6  
 3rd Qu.:16808   3rd Qu.: 76.5  
 Max.   :44218   Max.   :108.0  
```

A total of 7 islands are included in this data as displayed in Display 8.1 (page 207).

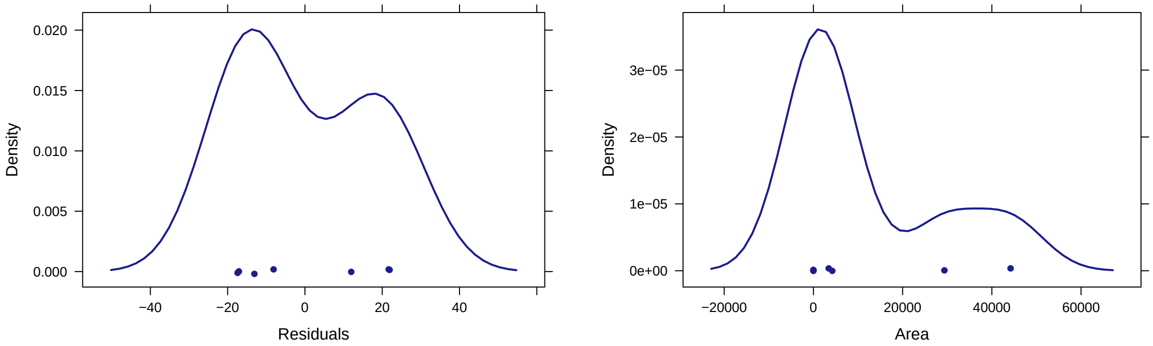
We can then observe the relationship between the area and the number of species for these islands with a scatterplot, akin to the top figure in Display 8.2 (page 208).

```
> xyplot(Species ~ Area, pch = 23, cex = 2, data = case0801)
```



It appears that the relationship with the observed values may not be linear. In addition, we need to verify the normality assumption for the residuals. Here we also consider a transformation for the predictor (`Area`).

```
> densityplot(~residuals(lm(Species ~ Area, data = case0801)), xlab = "Residuals")
> densityplot(~Area, data = case0801)
```

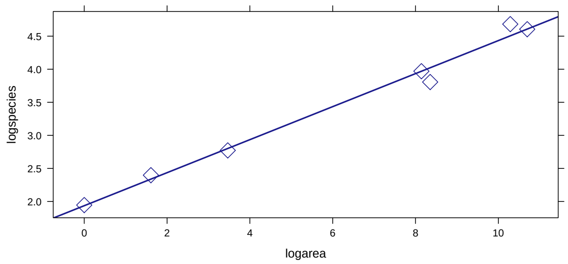


Since neither of these appear to be approximately normal, both the predictor and outcome variables are log-transformed (as suggested by the author).

```
> case0801 = transform(case0801, logarea = log(Area))
> case0801 = transform(case0801, logspecies = log(Species))
```

Then we can create a log-log-scatterplot for these two variables, akin to the bottom figure in Display 8.2 (page 208).

```
> xyplot(logspecies ~ logarea, type = c("p", "r"), pch = 23, cex = 2, data = case0801)
```



2.2 Simple Linear Model

We first fit the model for $\mu\{\log(\text{Species})|\log(\text{Area})\} = \beta_0 + \beta_1 * \log(\text{Area})$.

```
> lm1 = lm(logspecies ~ logarea, data = case0801)
> summary(lm1)


Call:
lm(formula = logspecies ~ logarea, data = case0801)

Residuals:
       Cuba  Hispaniola     Jamaica Puerto Rico  Montserrat        Saba 
  -0.002136    0.176975   -0.215487    0.000947   -0.029244    0.059543 
    Redonda 
   0.009402 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1.9365     0.0881    22.0  3.6e-06 ***
logarea       0.2497     0.0121    20.6  5.0e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.128 on 5 degrees of freedom
Multiple R-squared:  0.988,Adjusted R-squared:  0.986 
F-statistic:  425 on 1 and 5 DF,  p-value: 4.96e-06
```

Thus our estimated equation becomes, $\hat{\mu}\{\log(\text{Species})|\log(\text{Area})\} = 1.94 + 0.25 * \log(\text{Area})$.

Next we calculate the 95% confidence interval for the estimates, note that the `logarea` 95% confidence interval is interpreted in the "Summary of Statistical Findings" on page 207:

```
> confint(lm1)

             2.5 % 97.5 %
(Intercept) 1.7100 2.1631
logarea     0.2186 0.2808
```

To interpret this log-log model the *Sleuth* notes that if $\hat{\mu}\{\log(\mathrm{Y})|\log(\mathrm{X})\} = \beta_0 + \beta_1 * \log(\mathrm{X})$ then $\mathrm{Median}\{\mathrm{Y}|\mathrm{X}\} = \exp(\beta_0)X^{\beta_1}$ (page 216). For this example the researchers are interested in a doubling effect (2^{β_1}). Therefore to obtain the 95% confidence interval for the multiplicative factor in the median we used the following code:

```
> 2^confint(lm1)

            2.5 % 97.5 %
(Intercept) 3.272  4.479
logarea     1.164  1.215
```

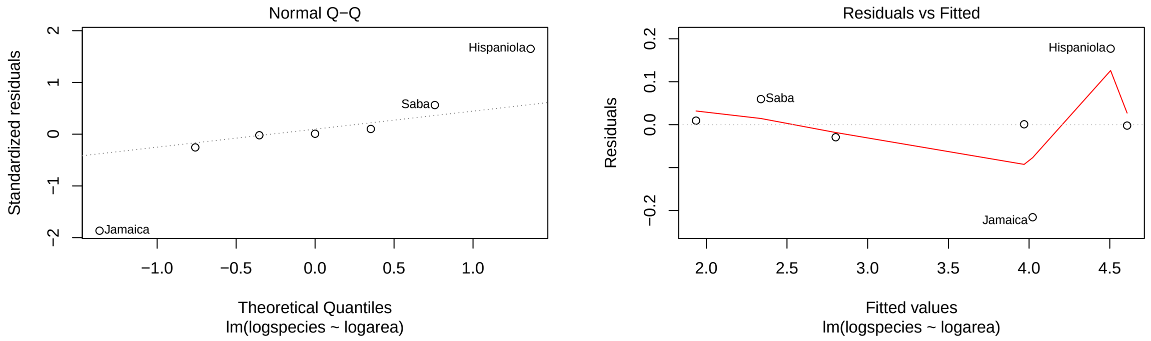
Thus for this model the estimated median number of species is 1.19 ($2^{0.25}$) with a 95% confidence interval between (1.16, 1.21). These match the numbers found on page 216.

2.3 Assessment of Assumptions

First we will have to assume independence from the information given. As seen in the above density plots, the observations for each variable were not normally distributed, once we performed a log transformation the distribution of the values became more approximately normal.

Next we can check for linearity and equal variance.

```
> plot(lm1, which = 2)
> plot(lm1, which = 1)
```



3 Breakdown Times for Insulating Fluid Under Different Voltages

How does the distribution of breakdown time depend on voltage? This is the question addressed in case study 8.2 in the *Sleuth*.

3.1 Summary statistics and graphical display

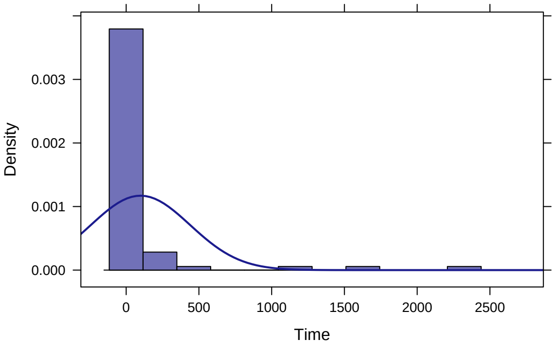
We begin by reading the data and summarizing the variables.

```
> summary(case0802)

      Time             Voltage          Group
 Min.   :   0.1   Min.   :26.0   Group 1: 3
 1st Qu.:   1.6   1st Qu.:31.5   Group 2: 5
 Median :   6.9   Median :34.0   Group 3:11
 Mean   :  98.6   Mean   :33.1   Group 4:15
 3rd Qu.:  38.4   3rd Qu.:36.0   Group 5:19
 Max.   :2323.7   Max.   :38.0   Group 6:15
                                 Group 7: 8
```

A total of 76 samples of insulating fluids are included in this data. Each sample was placed in one of 7 groups representing different degrees of voltage. Each group varried in sample size as shown in Display 8.2 (page 209).

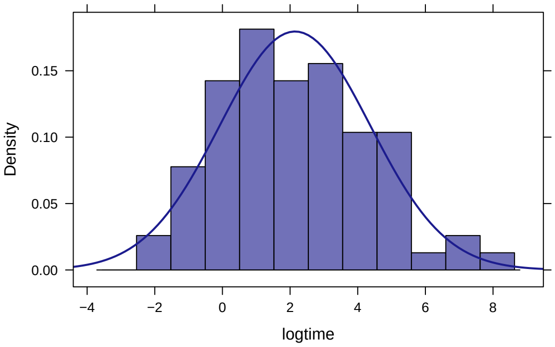
Before we can fit the simple linear regression model we need to assess the assumption of normality through density plots.

```
> histogram(~Time, type = "density", density = TRUE, nint = 10, data = case0802)
```



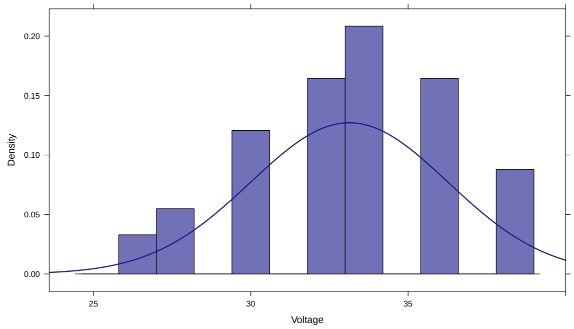
It appears that the distribution of `Time` is highly skewed with a long right tail. Therefore one possible transformation would be to take the log of the `Time` observations.

```
> case0802$logtime = with(case0802, log(Time))
> histogram(~logtime, type = "density", density = TRUE, nint = 10, data = case0802)
```



Now the observations are approximately normally distributed.

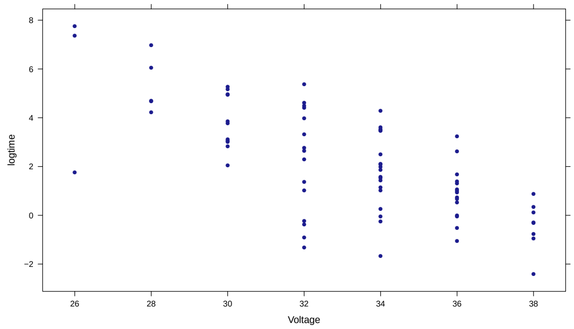
```
> histogram(~Voltage, type = "density", density = TRUE, nint = 10, data = case0802)
```



The distribution of `Voltage` seems to be approximately normal.

Next we can observe the relationship between log(`Time`) and `Voltage` (as in Display 8.4 ,page 210).

```
> xyplot(logtime ~ Voltage, data = case0802)
```



3.2 Simple linear regression models

The model that the researchers want to analyse is $\mu\{\log(\text{Time})|\text{Voltage}\} = \beta_0 + \beta_1 * \text{Voltage}$

```
> lm1 = lm(logtime ~ Voltage, data = case0802)
> summary(lm1)


Call:
lm(formula = logtime ~ Voltage, data = case0802)

Residuals:
   Min     1Q Median     3Q    Max 
-4.029 -0.692  0.037  1.209  2.651 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.9555     1.9100    9.92  3.1e-15 ***
Voltage      -0.5074     0.0574   -8.84  3.3e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.56 on 74 degrees of freedom
Multiple R-squared:  0.514,Adjusted R-squared:  0.507
F-statistic: 78.1 on 1 and 74 DF,  p-value: 3.34e-13
```

Therefore the estimated model is $\hat{\mu}\{\log(\text{Time})|\text{Voltage}\} = 18.96 + (\text{-}0.51)* \log(\text{Area})$. The R^2 for the model is 51.36%, as discussed on page 221.

For the interpretation of the model we first exponentiate the estimated coefficients since the response variable is logged as shown on page 215.

```
> exp(coef(lm1))

(Intercept)     Voltage
  1.707e+08   6.021e-01
```

Thus a 1 kV increase in volatge is associated with a multiplicative change in median breakdown time of 0.6.

Next we can calculate the 95% confidence interval for β_0 and β_1.

```
> confint(lm1)

              2.5 % 97.5 %
(Intercept) 15.1497 22.761
Voltage     -0.6217 -0.393
```

For the interpretation of the model we next need to exponentiate the 95% confidence interval.

```
> exp(confint(lm1))

                2.5 %    97.5 %
(Intercept) 3.797e+06 7.675e+09
Voltage     5.370e-01 6.750e-01
```

Thus the 95% confidence interval for the multiplicative change in median breakdown time is (0.54, 0.68) as interpreted on page 216.

Next we can assess the fit using the Analysis of Variance (ANOVA). The ANOVA results below match those in the top half of Display 8.8 (page 218).

```
> anova(lm1)

Analysis of Variance Table

Response: logtime
          Df Sum Sq Mean Sq F value  Pr(>F)
Voltage    1    190   190.2    78.1 3.3e-13 ***
Residuals 74    180     2.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can then compare this with a model with separate means for each group.

```
> lm2 = lm(logtime ~ as.factor(Voltage), data = case0802)
> summary(lm2)


Call:
lm(formula = logtime ~ as.factor(Voltage), data = case0802)

Residuals:
   Min     1Q Median     3Q    Max
-3.868 -0.819  0.074  1.122  3.143

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)             5.624      0.916    6.14  4.7e-08 ***
as.factor(Voltage)28   -0.294      1.159   -0.25  0.80019
as.factor(Voltage)30   -1.802      1.034   -1.74  0.08571 .
as.factor(Voltage)32   -3.395      1.004   -3.38  0.00118 **
as.factor(Voltage)34   -3.838      0.986   -3.89  0.00023 ***
as.factor(Voltage)36   -4.722      1.004   -4.70  1.3e-05 ***
as.factor(Voltage)38   -6.048      1.074   -5.63  3.6e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.59 on 69 degrees of freedom
Multiple R-squared:  0.531,Adjusted R-squared:  0.49
F-statistic:   13 on 6 and 69 DF,  p-value: 8.87e-10
```

This model has a F-statistic of 13 with a p-value < 0.0001, as shown in the bottom half of Display 8.8 (page 218).

Another way of viewing this model is with the ANOVA.

```
> anova(lm2)

Analysis of Variance Table

Response: logtime
                   Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(Voltage)  6    196    32.7      13 8.9e-10 ***
Residuals          69    174     2.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note that the values for the **Residuals** can also be found in the bottom half of Display 8.8 (page 218).

The *F*-statistic and its associated *p*-value for the lack-of-fit discussion on page 219 can be calculated by comparing the two models with an ANOVA.

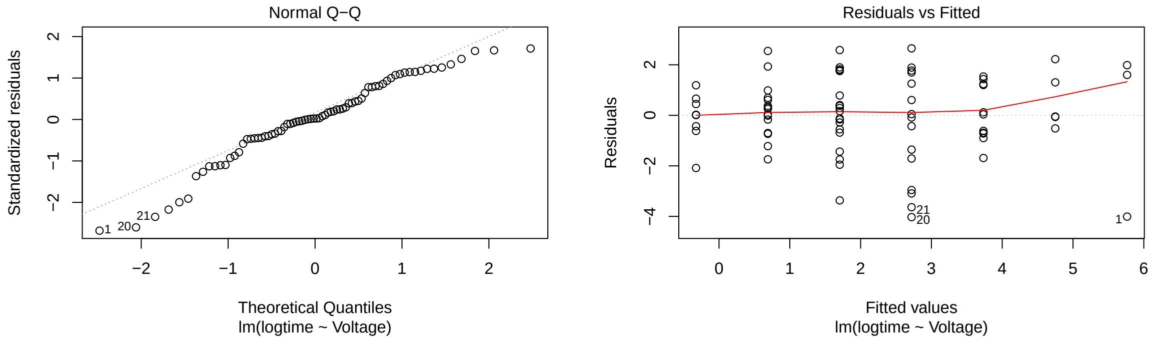
```
> anova(lm1, lm2)

Analysis of Variance Table

Model 1: logtime ~ Voltage
Model 2: logtime ~ as.factor(Voltage)
  Res.Df RSS Df Sum of Sq   F Pr(>F)
1     74 180
2     69 174  5      6.33 0.5   0.77
```

3.3 Assessment of Assumptions

First we will have to assume independence for the information given. As seen in the above density plot the observations for `Time` was not normally distributed, once we preformed a log transformation the distribution of the values became more approximately normal.

Next we can check for linearity (as in Display 8.14, page 225) and equal variance.

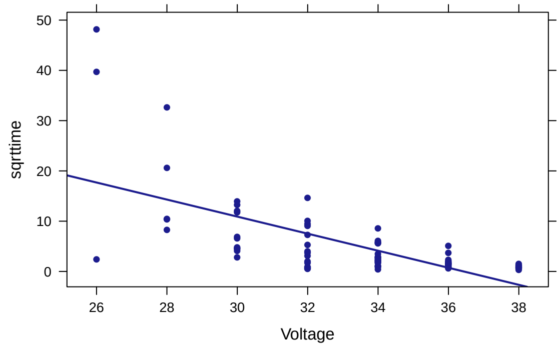
```
> plot(lm1, which = 2)
> plot(lm1, which = 1)
```



3.4 Other transformations

The *Sleuth* also discusses the use of a square root transformation for the breakdown time. The following figure is a scatterplot of the square root of breakdown time versus voltage, akin to the left figure in Display 8.7 (page 215).

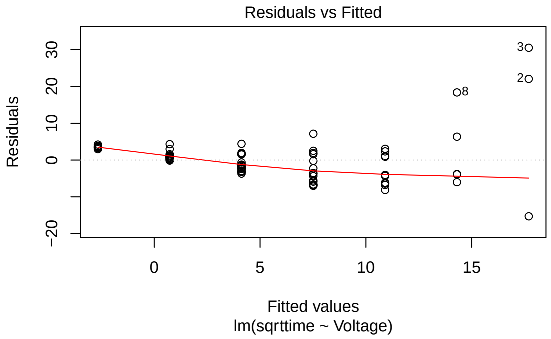
```
> case0802$sqrttime = with(case0802, sqrt(Time))
> xyplot(sqrttime ~ Voltage, type = c("p", "r"), data = case0802)
```



We can assess this transformation by observing the residual plot based on the simple linear regression fit, akin to the right figure in Display 8.7 (page 215).

```
> lm3 = lm(sqrttime ~ Voltage, data = case0802)
> summary(lm3)


Call:
lm(formula = sqrttime ~ Voltage, data = case0802)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.285  -3.711   0.142   2.040  30.514 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.784      7.777    7.94  1.6e-11 ***
Voltage       -1.696      0.234   -7.26  3.3e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.35 on 74 degrees of freedom
Multiple R-squared:  0.416,Adjusted R-squared:  0.408
F-statistic: 52.7 on 1 and 74 DF,  p-value: 3.25e-10

> plot(lm3, which = 1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(sqrttime ~ Voltage)