The Statistical Sleuth in R:
Chapter 11

Kate Aloisio   Ruobing Zhang   Nicholas J. Horton[*]

September 16, 2012

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.math.smith.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:

---

[*]Department of Mathematics and Statistics, Smith College, nhorton@smith.edu

```
> 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
```

```
> 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 = 3, show.signif.stars = FALSE)
```

The specific goal of this document is to demonstrate how to calculate the quantities described in Chapter 11: Model Checking and Refinement using R.

2   Alcohol metabolism in men and women

How do men and women metabolise alcohol? This is the question addressed in case study 11.1 in the *Sleuth*.

2.1   Data coding, summary statistics and graphical display

We begin by reading the data and summarizing the variables.

```
> summary(case1101)

   Subject         Metabol         Gastric          Sex
 Min.   : 1.0   Min.   : 0.10   Min.   :0.80   Female:18
 1st Qu.: 8.8   1st Qu.: 0.60   1st Qu.:1.20   Male  :14
 Median :16.5   Median : 1.70   Median :1.60
 Mean   :16.5   Mean   : 2.42   Mean   :1.86
 3rd Qu.:24.2   3rd Qu.: 2.93   3rd Qu.:2.20
 Max.   :32.0   Max.   :12.30   Max.   :5.20
          Alcohol
 Alcoholic    : 8
 Non-alcoholic:24
```
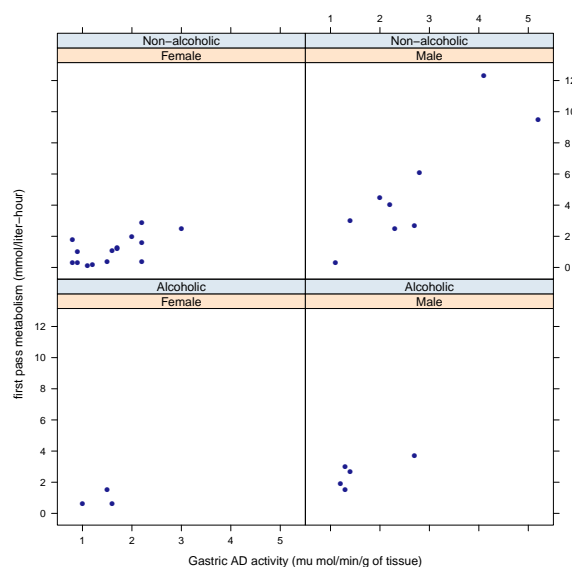
A total of 32 volunteers were included in this data. There were 18 females and 14 males. As recorded in Display 9.2 (page 237 of the *Sleuth*).

The following is a graphical display of the variables akin to Display 11.2 (page 306).

```
> xyplot(Metabol ~ Gastric | Sex + Alcohol, data = case1101, auto.key = TRUE,
+      xlab = "Gastric AD activity (mu mol/min/g of tissue)", ylab = "first pass metabolism (mmo
```


2.2   Multiple regression

First we can fit a full model for estimating *metabolism* given a subjects *gastric AD activity*, whether they are *alcoholic* and *gender*. This first model is summarized on page 315 (Display 11.9).

```
> lm1 = lm(Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex + Sex * Alcohol +
+      Gastric * Alcohol + Gastric * Sex * Alcohol, data = case1101)
> summary(lm1)


Call:
lm(formula = Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex +
    Sex * Alcohol + Gastric * Alcohol + Gastric * Sex * Alcohol,
    data = case1101)

Residuals:
   Min     1Q Median     3Q    Max
-2.429 -0.619 -0.047  0.515  3.652

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                        0.106      3.839    0.03     0.98
```

```
Gastric                               0.581      2.759     0.21      0.84
SexMale                               0.786      4.187     0.19      0.85
AlcoholNon-alcoholic                 -0.300      3.939    -0.08      0.94
Gastric:SexMale                       0.475      2.933     0.16      0.87
SexMale:AlcoholNon-alcoholic         -2.252      4.394    -0.51      0.61
Gastric:AlcoholNon-alcoholic          0.260      2.807     0.09      0.93
Gastric:SexMale:AlcoholNon-alcoholic  1.199      2.998     0.40      0.69


Residual standard error: 1.25 on 24 degrees of freedom
Multiple R-squared: 0.828,Adjusted R-squared: 0.777
F-statistic: 16.5 on 7 and 24 DF,  p-value: 9.35e-08
```

Next we can calculate a number of model diagnostics, including leverage, studentized resids and Cook's distance (pages 319–320).
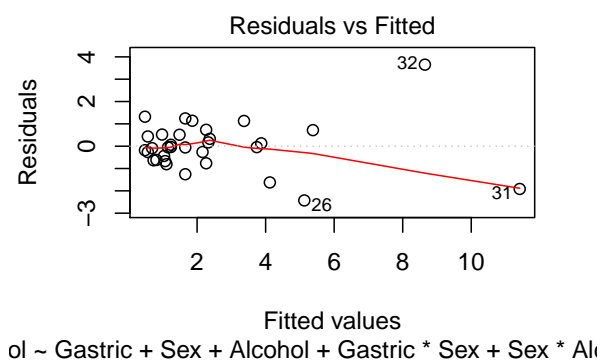
```
> require(MASS)
```

```
> case1101$hat = hatvalues(lm1)
> case1101$studres = studres(lm1)
> case1101$cooks = cooks.distance(lm1)
> case1101[31, ]

   Subject Metabol Gastric  Sex        Alcohol   hat studres cooks
31      31     9.5     5.2 Male Non-alcoholic 0.601   -2.72   1.1
```

The following is a residual plot for the full model akin to Display 11.7 (page 313).

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


Residuals vs Fitted

From these diagnostics it appears that observations 31 and 32 may be influential points. Therefore, we next re-fit the full model excluding these two observations. The following results are found in Display 11.9 and discussed on page 315.

```
> case11012 = case1101[-c(31, 32), ]
> lm2 = lm(Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex + Sex * Alcohol +
+     Gastric * Alcohol + Gastric * Sex * Alcohol, data = case11012)
> summary(lm2)


Call:
lm(formula = Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex +
    Sex * Alcohol + Gastric * Alcohol + Gastric * Sex * Alcohol,
    data = case11012)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8076 -0.5701 -0.0466  0.4976  1.4002

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                            0.106      2.881    0.04     0.97
Gastric                                0.581      2.070    0.28     0.78
SexMale                                0.786      3.141    0.25     0.80
AlcoholNon-alcoholic                  -0.300      2.956   -0.10     0.92
Gastric:SexMale                        0.475      2.201    0.22     0.83
SexMale:AlcoholNon-alcoholic          -1.272      3.467   -0.37     0.72
Gastric:AlcoholNon-alcoholic           0.260      2.106    0.12     0.90
Gastric:SexMale:AlcoholNon-alcoholic   0.606      2.316    0.26     0.80

Residual standard error: 0.941 on 22 degrees of freedom
Multiple R-squared: 0.685,Adjusted R-squared: 0.585
F-statistic: 6.83 on 7 and 22 DF,  p-value: 0.000226
```

2.3   Refining the Model

This section addresses the process of refining the model. We first tested the lack of fit for the removal of `Alcohol` as shown in Display 11.13 (page 322).

```
> lm3 = lm(Metabol ~ Gastric + Sex + Gastric * Sex, data = case11012)
> summary(lm3)


Call:
lm(formula = Metabol ~ Gastric + Sex + Gastric * Sex, data = case11012)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-1.5962 -0.6025 -0.0408  0.4759  1.6473


Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.197      0.586   -0.34    0.739
Gastric          0.837      0.353    2.37    0.026
SexMale          0.267      0.993    0.27    0.790
Gastric:SexMale  0.728      0.539    1.35    0.188


Residual standard error: 0.882 on 26 degrees of freedom
Multiple R-squared: 0.673,Adjusted R-squared: 0.635
F-statistic: 17.8 on 3 and 26 DF,  p-value: 1.71e-06



> anova(lm3, lm2)  # page 322

Analysis of Variance Table

Model 1: Metabol ~ Gastric + Sex + Gastric * Sex
Model 2: Metabol ~ Gastric + Sex + Alcohol + Gastric * Sex + Sex * Alcohol +
    Gastric * Alcohol + Gastric * Sex * Alcohol
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     26 20.2
2     22 19.5  4      0.74 0.21   0.93
```

Next we assessed a model without an intercept which is scientifically plausible as summarized in Display 11.14 (page 323).

```
> lm4 = lm(Metabol ~ Gastric + Gastric:Sex - 1, data = case11012)
> summary(lm4)


Call:
lm(formula = Metabol ~ Gastric + Gastric:Sex - 1, data = case11012)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6171 -0.6075 -0.0262  0.4772  1.6230

Coefficients: (1 not defined because of singularities)
                  Estimate Std. Error t value Pr(>|t|)
Gastric              1.599      0.125   12.80  3.2e-13
Gastric:SexFemale   -0.873      0.174   -5.02  2.6e-05
Gastric:SexMale         NA         NA      NA       NA

Residual standard error: 0.852 on 28 degrees of freedom
```

```
Multiple R-squared: 0.877,Adjusted R-squared: 0.868
F-statistic: 99.9 on 2 and 28 DF,  p-value: 1.8e-13


> anova(lm4, lm3)

Analysis of Variance Table

Model 1: Metabol ~ Gastric + Gastric:Sex - 1
Model 2: Metabol ~ Gastric + Sex + Gastric * Sex
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     28 20.3
2     26 20.2  2     0.094 0.06   0.94
```

Note that the "Summary of Statistical Findings" section (page 306) is based on this final model.

3   Blood brain barrier

Neuroscientists working to better understand the blood brain barrier have infused rats with cells to induce brain tumors. This is the topic addressed in case study 11.2 in the *Sleuth*.

3.1   Data coding and summary statistics

We begin by reading the data, performing transformations where needed and summarizing the variables.

```
> case1102$Y = with(case1102, Brain/Liver)
> case1102$logliver = with(case1102, log(Liver))
> case1102$logbrain = with(case1102, log(Brain))
> case1102$SAC = as.factor(case1102$Time)
> case1102$logy = with(case1102, log(Brain/Liver))
> case1102$logtime = with(case1102, log(Time))
> case1102$Treat = relevel(case1102$Treat, ref = "NS")
> summary(case1102)

     Brain              Liver              Time       Treat         Days
 Min.   :  1334   Min.   :    928   Min.   : 0.5   NS:17   Min.   : 9
 1st Qu.: 19281   1st Qu.:  16210   1st Qu.: 1.1   BD:17   1st Qu.:10
 Median : 32572   Median : 643965   Median : 3.0           Median :10
 Mean   : 39965   Mean   : 668776   Mean   :23.5           Mean   :10
 3rd Qu.: 50654   3rd Qu.:1318557   3rd Qu.:24.0           3rd Qu.:10
 Max.   :123730   Max.   :1790863   Max.   :72.0           Max.   :11
 Sex        Weight          Loss            Tumor           Y
 F:26   Min.   :184    Min.   :-4.90    Min.   : 25    Min.   :0.01
 M: 8   1st Qu.:225    1st Qu.: 1.20    1st Qu.:136    1st Qu.:0.03
        Median :240    Median : 3.95    Median :166    Median :0.12
```

```
        Mean   :242    Mean    : 3.64    Mean    :183    Mean    :1.50
        3rd Qu.:259    3rd Qu.: 5.97     3rd Qu.:223     3rd Qu.:1.95
        Max.   :298    Max.    :12.80    Max.    :484    Max.    :8.55
     logliver          logbrain      SAC         logy           logtime
 Min.   : 6.83    Min.    : 7.20   0.5:9    Min.    :-4.58   Min.    :-0.69
 1st Qu.: 9.69    1st Qu.: 9.86    3   :9   1st Qu.:-3.39    1st Qu.:-0.25
 Median :13.37    Median :10.39    24 :8    Median :-2.13    Median : 1.10
 Mean   :11.61    Mean    :10.23   72 :8    Mean    :-1.39   Mean    : 1.86
 3rd Qu.:14.09    3rd Qu.:10.83             3rd Qu.: 0.67    3rd Qu.: 3.18
 Max.   :14.40    Max.    :11.73            Max.    : 2.15   Max.    : 4.28
```

A total of 34 rats were included in this experiment. Each rat was given either the barrier solution (n = 17) or a normal saline solution (n = 17). Then variables of interest were calculated and are displayed in Display 9.2 (page 237 of the *Sleuth*).

We can graphically relationships between the variables using a pairs plot.

```
> smallds = case1102[, c("logy", "logbrain", "logliver", "Treat", "SAC")]
> pairs(smallds)
```
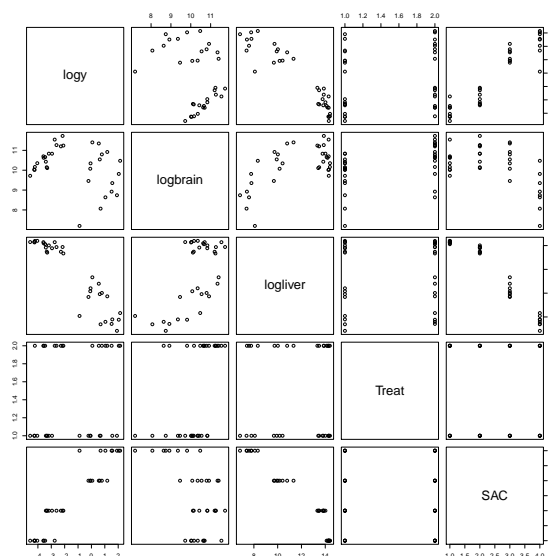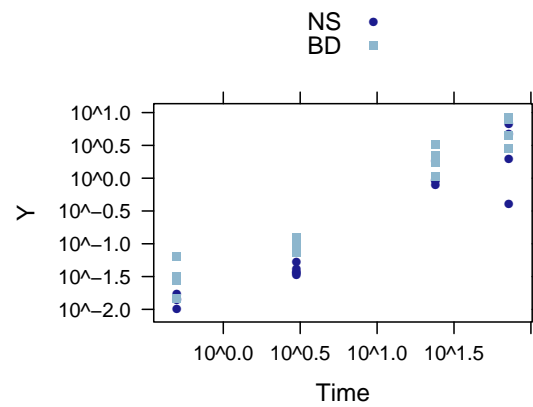

3.2   Graphical presentation

The following displays a scatterplot of log ratio (Y) as a function of log time, akin to Display 11.5 on page 309.
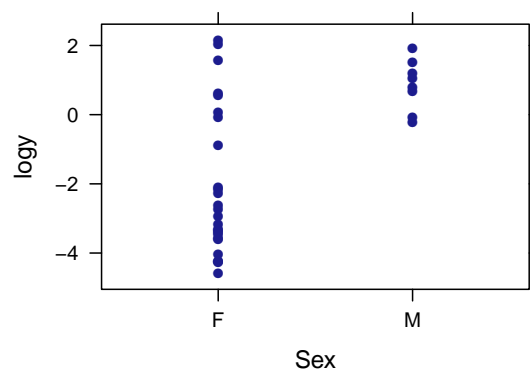
```
> xyplot(Y ~ Time, group = Treat, scales = list(y = list(log = TRUE), x = list(log = TRUE)),
+     auto.key = TRUE, data = case1102)
```
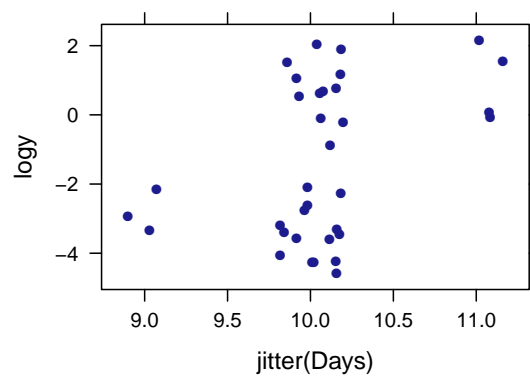
The following graphs are akin to the second and third plots in Display 11.16 on page 326.

```
> xyplot(logy ~ Sex, data = case1102)
```



```
> xyplot(logy ~ jitter(Days), data = case1102)
```

3.3 Multiple regression

We first fit a model that reflects the initial investigation. This is the proposed model from page 311.

```
> lm1 = lm(logy ~ SAC + Treat + SAC * Treat + Days + Sex + Weight + Loss + Tumor,
+     data = case1102)
> summary(lm1)


Call:
lm(formula = logy ~ SAC + Treat + SAC * Treat + Days + Sex +
    Weight + Loss + Tumor, data = case1102)

Residuals:
    Min      1Q  Median      3Q     Max
-1.4056 -0.2559  0.0458  0.1957  1.1583

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.836741   3.391046   -1.13    0.271
SAC3           1.015463   0.399578    2.54    0.019
SAC24          4.337135   0.477836    9.08  1.0e-08
SAC72          5.010605   0.454953   11.01  3.5e-10
TreatBD        0.795999   0.378970    2.10    0.048
Days          -0.036987   0.295645   -0.13    0.902
SexM           0.001295   0.373368    0.00    0.997
Weight        -0.000558   0.005330   -0.10    0.918
Loss          -0.059544   0.030422   -1.96    0.064
Tumor          0.001551   0.001226    1.26    0.220
SAC3:TreatBD   0.179831   0.551964    0.33    0.748
SAC24:TreatBD -0.386047   0.585450   -0.66    0.517
SAC72:TreatBD  0.379104   0.569242    0.67    0.513


Residual standard error: 0.564 on 21 degrees of freedom
Multiple R-squared: 0.96,Adjusted R-squared: 0.937
F-statistic: 41.9 on 12 and 21 DF,  p-value: 6.45e-12
```
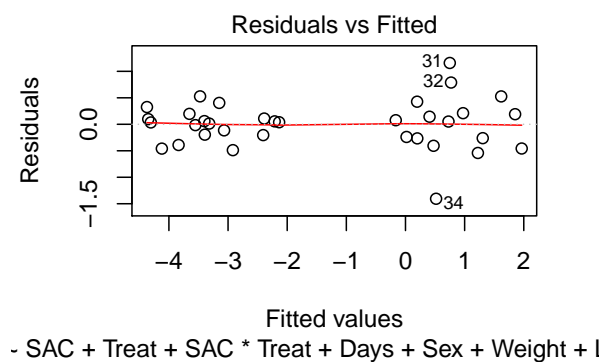
We can then display a residual plot to assess the fit of the above model. This is provided in Display 11.6 (page 312).

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

Residuals vs Fitted

Fitted values
- SAC + Treat + SAC * Treat + Days + Sex + Weight + I

3.4   Refining the model

Lastly, we fit a refined model. These results can be found in Display 11.17 (page 327).

```
> lm2 = lm(logy ~ SAC + Treat, data = case1102)
> summary(lm2)


Call:
lm(formula = logy ~ SAC + Treat, data = case1102)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7402 -0.1755 -0.0178  0.2477  1.0551

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -4.302      0.205  -21.01  < 2e-16
SAC3           1.134      0.252    4.50  0.00010
SAC24          4.257      0.259   16.43  3.1e-16
SAC72          5.154      0.259   19.89  < 2e-16
TreatBD        0.797      0.183    4.35  0.00016

Residual standard error: 0.533 on 29 degrees of freedom
Multiple R-squared: 0.951,Adjusted R-squared: 0.944
F-statistic:  140 on 4 and 29 DF,  p-value: <2e-16


> anova(lm2, lm1)

Analysis of Variance Table

Model 1: logy ~ SAC + Treat
```

```
Model 2: logy ~ SAC + Treat + SAC * Treat + Days + Sex + Weight + Loss +
    Tumor
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     29 8.23
2     21 6.68  8      1.55 0.61   0.76
```