The Statistical Sleuth in R: Chapter 7

Ruobing Zhang Kate Aloisio 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 = 4)
```

The specific goal of this document is to demonstrate how to calculate the quantities described in Chapter 7: Simple Linear Regression: A Model for the Mean using R.

2 The Big Bang

Is there relation between distance and radial velocity among extra-galactic nebulae? This is the question addressed in case study 7.1 in the *Sleuth*.

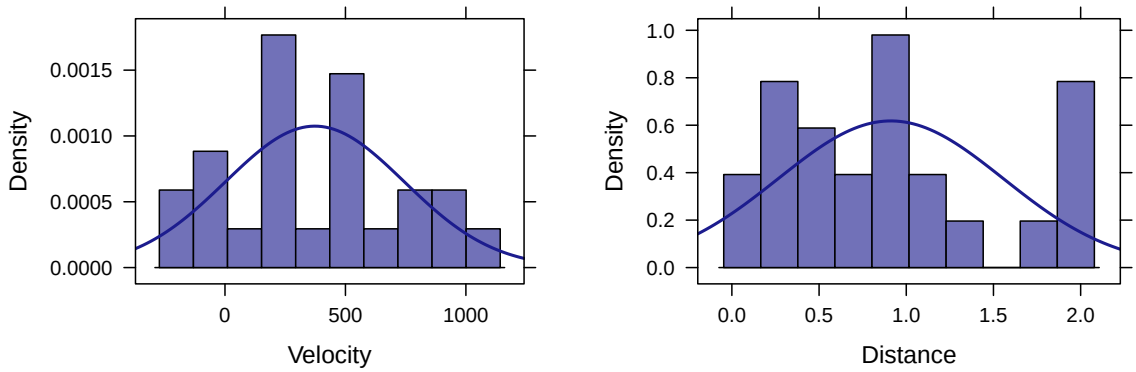
2.1 Summary statistics and graphical display

We begin by reading the data and summarizing the variables.

```
> summary(case0701)

    Velocity       Distance
 Min.   :-220   Min.   :0.032
 1st Qu.: 165   1st Qu.:0.406
 Median : 295   Median :0.900
 Mean   : 373   Mean   :0.911
 3rd Qu.: 538   3rd Qu.:1.175
 Max.   :1090   Max.   :2.000
```

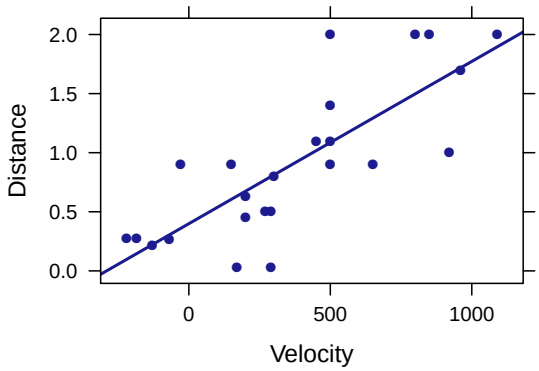
A total of 24 nebulae are included in this data.

```
> histogram(~Velocity, type = "density", density = TRUE, nint = 10, data = case0701)
> histogram(~Distance, type = "density", density = TRUE, nint = 10, data = case0701)
```



The density plots show that the distributions for the two variables are fairly symmetric, but more uniform than normally distributed.

```
> xyplot(Distance ~ Velocity, type = c("p", "r"), data = case0701)
```



The scatterplot is displayed on page 175 of the *Sleuth*. It indicates that there is a linear statistical relationship between distance and velocity.

2.2 The simple linear regression model

The following code presents the results interpreted on page 184 of the *Sleuth*.

```
> lm1 = lm(Distance ~ Velocity, data = case0701)
> summary(lm1)


Call:
lm(formula = Distance ~ Velocity, data = case0701)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.7632 -0.2352 -0.0088  0.2072  0.9144

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.399098   0.118470    3.37   0.0028 **
Velocity    0.001373   0.000227    6.04  4.5e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.405 on 22 degrees of freedom
Multiple R-squared: 0.624,Adjusted R-squared: 0.606
F-statistic: 36.4 on 1 and 22 DF,  p-value: 4.48e-06
```

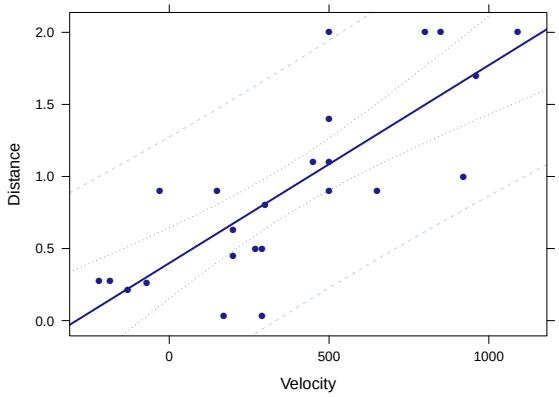
The estimated parameter for the intercept is 0.3991 megaparsecs and the estimated parameter for velocity is 0.0014 megaparsecs/(km/sec). The estimated mean function is $\hat{\mu}$ (distance|velocity) = 0.3991 + 0.0014 * velocity. The estimate of residual standard error is 0.405 megaparsecs with 22 degrees of freedom. These results are also presented by Display 7.9 (page 185).

```
> fitted(lm1)

      1       2       3       4       5       6       7       8       9
0.63250 0.79725 0.22062 0.30299 0.14511 0.09705 0.67369 0.79725 0.76979
     10      11      12      13      14      15      16      17      18
0.67369 0.81098 0.35791 1.29151 0.60504 1.08557 1.66220 1.01692 1.08557
     19      20      21      22      23      24
1.08557 1.71712 1.08557 1.56609 1.49745 1.89560

> resid(lm1)^2

        1         2         3         4         5         6         7
3.606e-01 5.826e-01 4.378e-05 1.599e-03 1.687e-02 3.167e-02 5.004e-02
        8         9        10        11        12        13        14
8.836e-02 7.279e-02 1.908e-03 1.205e-04 2.939e-01 1.533e-01 8.700e-02
       15        16        17        18        19        20        21
3.443e-02 4.385e-01 6.902e-03 2.083e-04 9.887e-02 2.930e-04 8.362e-01
       22        23        24
1.883e-01 2.526e-01 1.090e-02

> sum(resid(lm1)^2)

[1] 3.608

> sum(resid(lm1)^2)/sum((fitted(lm1) - mean(~Distance, data = case0701))^2)

[1] 0.6038
```

Display 7.8 (page 184) shows the list of fitted values and residuals for this model. The sum of all the squared residuals is 3.608 and R-squared is 0.6038.

We can also display 95% confidence bands for the model line and the predicted values, the following graph is akin to Display 7.11 (page 189).

```
> xyplot(Distance ~ Velocity, panel = panel.lmbands, data = case0701)
```



2.3 Inferential Tools

First, we test β_0 (the intercept). From the previous summary, we know that the two-sided p-value for the intercept is 0.0028. This p-value is small enough for us to reject the null hypothesis that the estimated parameter for the intercept equals 0 (page 186).

Next we want to examine β_1. The current β_1 for $\hat{\mu}\,(\mathrm{Y}|\mathrm{X}) = \beta_0 + \beta_1 * \mathrm{X}$ is 0.0014, and we want to get the β_1 for $\hat{\mu}\,(\mathrm{Y}|\mathrm{X}) = \beta_1 * \mathrm{X}$, a model with no intercept (page 186).

```
> # linear regression with no intercept
> lm2 = lm(Distance ~ Velocity - 1, data = case0701)
> summary(lm2)


Call:
lm(formula = Distance ~ Velocity - 1, data = case0701)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7681 -0.0694  0.2293  0.4629  1.0391

Coefficients:
         Estimate Std. Error t value Pr(>|t|)
Velocity 0.001922   0.000191    10.1  6.9e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.488 on 23 degrees of freedom
Multiple R-squared: 0.815,Adjusted R-squared: 0.807
F-statistic:  101 on 1 and 23 DF,  p-value: 6.87e-10


> confint(lm2)

            2.5 %   97.5 %
Velocity 0.001526 0.002317
```

Without the intercept, the new estimate for β_1 is 0.0019 megaparsec-second/km. The standard error is 1.91×10^{-4} megaparsecs with 23 degrees of freedom. The 95% confidence interval is (0.0015, 0.0023). Because 1 megaparsec-second/km = 979.8 billion years, the confidence interval could be written as 1.5 to 2.27 billion years, and the best estimate is 1.88 billion years (page 186).

3 Meat Processing and pH

Is there a relationship between postmortem muscle pH and time after slaughter? This is the question addressed in case study 7.2 in the *Sleuth*.

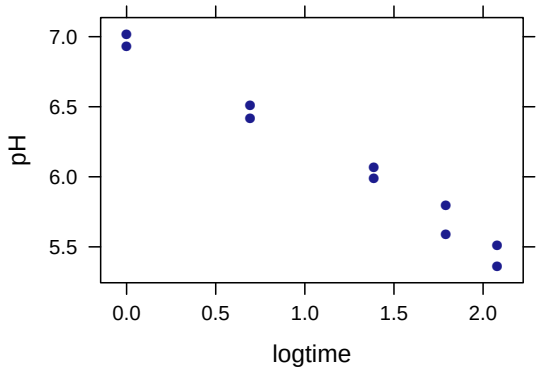
3.1 Summary statistics and graphical display

We begin by reading the data and summarizing the variables.

```
> summary(case0702)

      Time            pH
 Min.   :1.0   Min.   :5.36
 1st Qu.:2.0   1st Qu.:5.64
 Median :4.0   Median :6.03
 Mean   :4.2   Mean   :6.12
 3rd Qu.:6.0   3rd Qu.:6.49
 Max.   :8.0   Max.   :7.02
```

A total of 10 steer carcasses are included in this data as shown in Display 7.3, page 117.

```
> logtime = log(case0702$Time)
> xyplot(pH ~ logtime, data = case0702)
```



The above scatterplot indicates a negative linear relationship between pH and log(Time).

3.2 The simple linear regression model

We fit a simple linear regression model of pH on log(time) after slaughter. The estimated mean function will be $\hat{\mu}\,(\text{pH}|\text{logtime}) = \beta_0 + \beta_1 * \log(\text{Time})$.

```
> lm3 = lm(pH ~ logtime, data = case0702)
> summary(lm3)


Call:
lm(formula = pH ~ logtime, data = case0702)

Residuals:
    Min      1Q  Median      3Q     Max
-0.1147 -0.0589  0.0209  0.0361  0.1166

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.9836     0.0485   143.9  6.1e-15 ***
logtime      -0.7257     0.0344   -21.1  2.7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.0823 on 8 degrees of freedom
Multiple R-squared: 0.982,Adjusted R-squared: 0.98
F-statistic:  444 on 1 and 8 DF,  p-value: 2.7e-08


> beta0 = coef(lm3)["(Intercept)"]
> beta0

(Intercept)
      6.984
```

```
> beta1 = coef(lm3)["logtime"]
> beta1

logtime
-0.7257

> sigma = summary(lm3)$sigma
> sigma

[1] 0.08226
```

The $\hat{\beta}_0$ is 6.9836 and the $\hat{\beta}_1$ is -0.7257. The $\hat{\sigma}$ is 0.0823 (page 187).

3.3 Inferential Tools

With the previous information, we can calculate the 95% confidence interval for the estimated mean pH of steers 4 hours after slaughter (Display 7.10, page 187):

```
> mu = beta0 + beta1 * log(4)
> mu

(Intercept)
      5.978

> n = nrow(case0702)
> mean = mean(~logtime, data = case0702)
> sd = sd(~logtime, data = case0702)
> se = sigma * sqrt(1/n + (log(4) - mean)^2/((n - 1) * sd))
> se

[1] 0.0267

> upper = mu + qt(0.975, df = 8) * se
> upper

(Intercept)
      6.039

> lower = mu - qt(0.975, df = 8) * se
> lower

(Intercept)
      5.916
```

Or we can use the following code to get the same result:

```
> predict(lm3, interval = "confidence")[5, ]

  fit   lwr   upr
5.978 5.916 6.040
```

So the 95% confidence interval for estimated mean is (5.92, 6.04).

Next, we can calculate the 95% prediction interval for a steer carcass 4 hours after slaughter (Display 7.12, page 191):

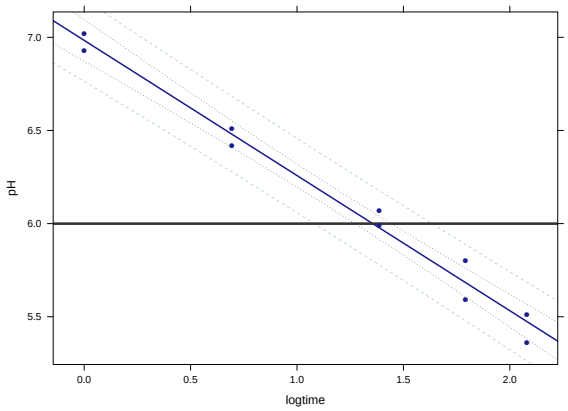
```
> pred = beta0 + beta1 * log(4)
> pred

(Intercept)
      5.978

> predse = sigma * sqrt(1 + 1/n + (log(4) - mean)^2/((n - 1) * sd))
> predse

[1] 0.08648

> predupper = pred + qt(0.975, df = 8) * predse
> predupper

(Intercept)
      6.177

> predlower = pred - qt(0.975, df = 8) * predse
> predlower

(Intercept)
      5.778
```

Or we can use the following code to get the 95% prediction interval for a steer carcass 4 hours after slaughter:

```
> predict(lm3, interval = "prediction")[5, ]

Warning:  Predictions on current data refer to _future_ responses

  fit   lwr   upr
5.978 5.778 6.177
```

So the 95% prediction interval is (5.78, 6.18).

```
> xyplot(pH ~ logtime, abline = (h = 6), data = case0702, panel = panel.lmbands)
```



The 95% prediction band is presented as Display 7.4 (page 178).