

The Statistical Sleuth in R:

Chapter 7

Linda Loi Ruobing Zhang Kate Aloisio Nicholas J. Horton*

January 2, 2016

Contents

1	Introduction	1
2	The Big Bang	2
2.1	Summary statistics and graphical display	2
2.2	The simple linear regression model	3
2.3	Inferential Tools	5
3	Meat Processing and pH	6
3.1	Summary statistics and graphical display	6
3.2	The simple linear regression model	7
3.3	Inferential Tools	8

1 Introduction

This document is intended to help describe how to undertake analyses introduced as examples in the Third Edition of the *Statistical Sleuth* (2013) 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/sleuth3>.

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 **Sleuth3** package.

```
> install.packages('Sleuth3') # note the quotation marks
```

```
> require(Sleuth3)
```

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

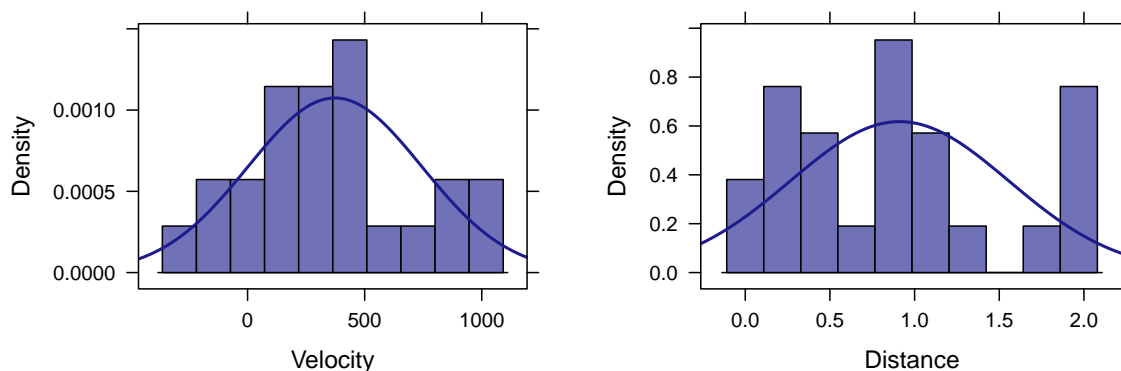
We begin by reading the data and summarizing the variables.

```
> summary(case0701)
```

Velocity		Distance	
Min.	:-220	Min.	:0.030
1st Qu.:	165	1st Qu.:	0.407
Median :	295	Median :	0.900
Mean :	373	Mean :	0.911
3rd Qu.:	538	3rd Qu.:	1.175
Max.	:1090	Max.	:2.000

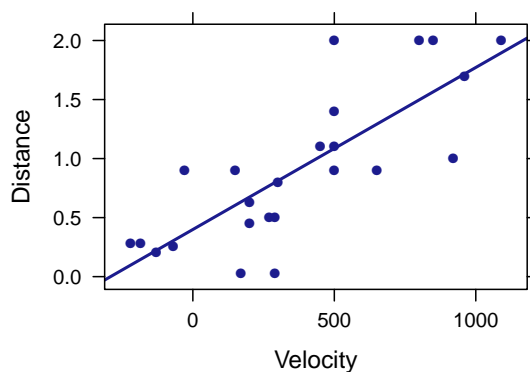
A total of 24 nebulae are included in this data.

```
> 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 177 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 186 of the *Sleuth*.

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

Call:

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

Residuals:

Min	1Q	Median	3Q	Max
-0.7672	-0.2352	-0.0108	0.2108	0.9146

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.399170	0.118666	3.36	0.0028
Velocity	0.001372	0.000228	6.02	4.6e-06

Residual standard error: 0.406 on 22 degrees of freedom

Multiple R-squared: 0.623, Adjusted R-squared: 0.605

F-statistic: 36.3 on 1 and 22 DF, p-value: 4.61e-06

The estimated parameter for the intercept is 0.3992 megaparsecs and the estimated parameter for velocity is 0.0014 megaparsecs/(km/sec). The estimated mean function is $\hat{\mu}$ (distance|velocity) = 0.3992 + 0.0014 * velocity. The estimate of residual standard error is 0.4056 megaparsecs with 22 degrees of freedom. These results are also presented by Display 7.9 (page 187).

```
> fitted(lm1)
```

1	2	3	4	5	6	7	8	9
0.63248	0.79717	0.22076	0.30310	0.14528	0.09724	0.67365	0.79717	0.76972
10	11	12	13	14	15	16	17	18
0.67365	0.81089	0.35800	1.29124	0.60503	1.08537	1.66179	1.01675	1.08537
19	20	21	22	23	24			
1.08537	1.71668	1.08537	1.56572	1.49710	1.89509			

```
> resid(lm1)^2
```

1	2	3	4	5	6	7
0.3629818	0.5885477	0.0001157	0.0018578	0.0181508	0.0334009	0.0500202
8	9	10	11	12	13	14
0.0883092	0.0727491	0.0019055	0.0001187	0.2937659	0.1530651	0.0870064
15	16	17	18	19	20	21
0.0343636	0.4379599	0.0069299	0.0002139	0.0989894	0.0002783	0.8365403
22	23	24				
0.1886019	0.2529120	0.0110051				

```
> sum(resid(lm1)^2)
```

```
[1] 3.62
```

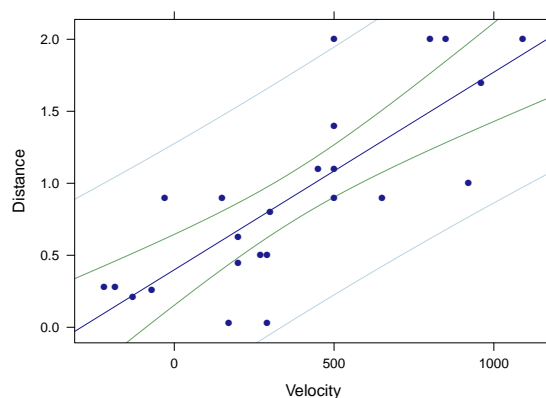
```
> sum(resid(lm1)^2)/sum((fitted(lm1)-mean(~Distance, data=case0701))^2)
```

```
[1] 0.6062
```

Display 7.8 (page 186) shows the list of fitted values and residuals for this model. The sum of all the squared residuals is 3.62 and R-squared is 0.6062.

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 191).

```
> 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 188).

Next we want to examine β_1 . The current β_1 for $\hat{\mu}(Y|X) = \beta_0 + \beta_1 * X$ is 0.0014, and we want to get the β_1 for $\hat{\mu}(Y|X) = \beta_1 * X$, a model with no intercept (page 188).

```
> # 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
-0.7677	-0.0691	0.2295	0.4606	1.0393

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
Velocity	0.001921	0.000191	10	7e-10

Residual standard error: 0.488 on 23 degrees of freedom

Multiple R-squared: 0.814, Adjusted R-squared: 0.806

```
F-statistic: 101 on 1 and 23 DF, p-value: 7.05e-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 1.49 to 2.27 billion years, and the best estimate is 1.88 billion years (page 188).

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

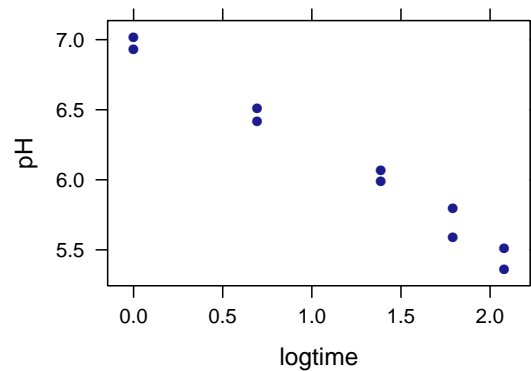
We begin by reading the data and summarizing the variables.

```
> 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 179.

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



The above scatterplot indicates a negative linear relationship between pH and $\log(\text{Time})$.

3.2 The simple linear regression model

We fit a simple linear regression model of pH on $\log(\text{time})$ 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

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 189).

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 189):

```
> 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 193):


```

> 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 in predict.lm(lm3, interval = "prediction"): predictions on current data refer
to _future_ responses

    fit    lwr    upr
5.978 5.778 6.177

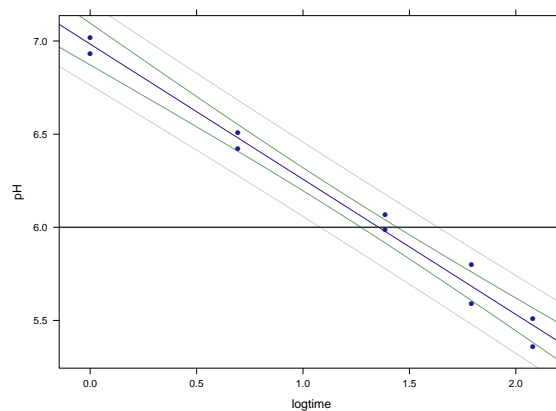
```

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 180).