A Handbook of Statistical Analyses
Using R — 2nd Edition

Brian S. Everitt and Torsten Hothorn
CHAPTER 6

Simple and Multiple Linear Regression: How Old is the Universe and Cloud Seeding

6.1 Introduction

We give the relative velocity and the distance of 24 galaxies, according to measurements made using the Hubble Space Telescope – the data are contained in the gamair package accompanying ?, see Table 6.1. Velocities are assessed by measuring the Doppler red shift in the spectrum of light observed from the galaxies concerned, although some correction for velocity components is required. Distances are measured using the known relationship between the period of Cepheid variable stars and their luminosity. How can these data be used to estimate the age of the universe? Here we shall show how this can be done using simple linear regression.

Table 6.1: hubble data. Distance and velocity for 24 galaxies.

<table>
<thead>
<tr>
<th>galaxy</th>
<th>velocity</th>
<th>distance</th>
<th>galaxy</th>
<th>velocity</th>
<th>distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>NGC0300</td>
<td>133</td>
<td>2.00</td>
<td>NGC3621</td>
<td>609</td>
<td>6.64</td>
</tr>
<tr>
<td>NGC0925</td>
<td>664</td>
<td>9.16</td>
<td>NGC4321</td>
<td>1433</td>
<td>15.21</td>
</tr>
<tr>
<td>NGC1326</td>
<td>1794</td>
<td>16.14</td>
<td>NGC4414</td>
<td>619</td>
<td>17.70</td>
</tr>
<tr>
<td>NGC1365</td>
<td>1594</td>
<td>17.95</td>
<td>NGC4496A</td>
<td>1424</td>
<td>14.86</td>
</tr>
<tr>
<td>NGC1425</td>
<td>1473</td>
<td>21.88</td>
<td>NGC4548</td>
<td>1384</td>
<td>16.22</td>
</tr>
<tr>
<td>NGC2403</td>
<td>278</td>
<td>3.22</td>
<td>NGC4535</td>
<td>1444</td>
<td>15.78</td>
</tr>
<tr>
<td>NGC2541</td>
<td>714</td>
<td>11.22</td>
<td>NGC4536</td>
<td>1423</td>
<td>14.93</td>
</tr>
<tr>
<td>NGC2690</td>
<td>882</td>
<td>11.75</td>
<td>NGC4639</td>
<td>1403</td>
<td>21.98</td>
</tr>
<tr>
<td>NGC3031</td>
<td>80</td>
<td>3.63</td>
<td>NGC4725</td>
<td>1103</td>
<td>12.36</td>
</tr>
<tr>
<td>NGC3198</td>
<td>772</td>
<td>13.80</td>
<td>IC4182</td>
<td>318</td>
<td>4.49</td>
</tr>
<tr>
<td>NGC351</td>
<td>642</td>
<td>10.00</td>
<td>NGC5253</td>
<td>232</td>
<td>3.15</td>
</tr>
<tr>
<td>NGC368</td>
<td>768</td>
<td>10.52</td>
<td>NGC7331</td>
<td>999</td>
<td>14.72</td>
</tr>
</tbody>
</table>

Weather modification, or cloud seeding, is the treatment of individual clouds or storm systems with various inorganic and organic materials in the hope of achieving an increase in rainfall. Introduction of such material into a cloud that contains supercooled water, that is, liquid water colder than zero degrees Celsius, has the aim of inducing freezing, with the consequent ice particles growing at the expense of liquid droplets and becoming heavy enough to fall as rain from clouds that otherwise would produce none.

The data shown in Table 6.2 were collected in the summer of 1975 from an experiment to investigate the use of massive amounts of silver iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall (15). In the experiment, which was conducted in an area of Florida, 24 days were judged suitable for seeding on the basis that a measured suitability criterion, denoted $S-N_e$, was not less than 1.5. Here $S$ is the ‘seedability’, the difference between the maximum height of a cloud if seeded and the same cloud if not seeded predicted by a suitable cloud model, and $N_e$ is the number of hours between...
1300 and 1600 G.M.T. with 10 centimetre echoes in the target; this quantity biases the decision for experimentation against naturally rainy days. Consequently, optimal days for seeding are those on which seedability is large and the natural rainfall early in the day is small.

On suitable days, a decision was taken at random as to whether to seed or not. For each day the following variables were measured:

- **seeding**: a factor indicating whether seeding action occurred (yes or no),
- **time**: number of days after the first day of the experiment,
- **cloudcover**: the percentage cloud cover in the experimental area, measured using radar,
- **prewetness**: the total rainfall in the target area one hour before seeding (in cubic metres $\times 10^7$),
- **echomotion**: a factor showing whether the radar echo was moving or stationary,
- **rainfall**: the amount of rain in cubic metres $\times 10^7$,
- **sne**: suitability criterion, see above.

The objective in analysing these data is to see how rainfall is related to the explanatory variables and, in particular, to determine the effectiveness of seeding. The method to be used is *multiple linear regression*.

### 6.2 Simple Linear Regression

### 6.3 Multiple Linear Regression

#### 6.3.1 Regression Diagnostics

### 6.4 Analysis Using R

#### 6.4.1 Estimating the Age of the Universe

Prior to applying a simple regression to the data it will be useful to look at a plot to assess their major features. The R code given in Figure 6.1 produces a scatterplot of velocity and distance. The diagram shows a clear, strong relationship between velocity and distance. The next step is to fit a simple linear regression model to the data, but in this case the nature of the data requires a model without intercept because if distance is zero so is relative speed. So the model to be fitted to these data is

$$\text{velocity} = \beta_1 \text{distance} + \varepsilon.$$  

This is essentially what astronomers call Hubble’s Law and $\beta_1$ is known as Hubble’s constant; $\beta_1^{-1}$ gives an approximate age of the universe.

To fit this model we are estimating $\beta_1$ using formula (??). Although this operation is rather easy

```r
R> sum(hubble$distance * hubble$velocity) / 
+ sum(hubble$distance^2)
```
R> plot(velocity ~ distance, data = hubble)

![Scatterplot of velocity and distance.](image)

**Figure 6.1** Scatterplot of velocity and distance.

It is more convenient to apply R’s linear modelling function

R> hmod <- lm(velocity ~ distance - 1, data = hubble)

Note that the model formula specifies a model without intercept. We can now extract the estimated model coefficients via

R> coef(hmod)

```
distance
76.6
```

and add this estimated regression line to the scatterplot; the result is shown in Figure 6.2. In addition, we produce a scatterplot of the residuals $y_i - \hat{y}_i$ against fitted values $\hat{y}_i$ to assess the quality of the model fit. It seems that for higher distance values the variance of velocity increases; however, we
are interested in only the estimated parameter \( \hat{\beta}_1 \) which remains valid under variance heterogeneity (in contrast to \( t \)-tests and associated \( p \)-values).

Now we can use the estimated value of \( \hat{\beta}_1 \) to find an approximate value for the age of the universe. The Hubble constant itself has units of \( \text{km} \times \text{sec}^{-1} \times \text{Mpc}^{-1} \). A mega-parsec (Mpc) is \( 3.09 \times 10^{19} \text{km} \), so we need to divide the estimated value of \( \hat{\beta}_1 \) by this amount in order to obtain Hubble’s constant with units of \( \text{sec}^{-1} \). The approximate age of the universe in seconds will then be the inverse of this calculation. Carrying out the necessary computations

\[
\text{Mpc} \leftarrow 3.09 \times 10^{19} \\
ysec \leftarrow 60^2 \times 24 \times 365.25 \\
\text{Mpcyear} \leftarrow \text{Mpc} / ysec \\
1 / (\text{coef(hmod)} / \text{Mpcyear})
\]

gives an estimated age of roughly 12.8 billion years.
6.4.2 Cloud Seeding

Again, a graphical display highlighting the most important aspects of the data will be helpful. Here we will construct boxplots of the rainfall in each category of the dichotomous explanatory variables and scatterplots of rainfall against each of the continuous explanatory variables.

Both the boxplots (Figure 6.3) and the scatterplots (Figure 6.4) show some evidence of outliers. The row names of the extreme observations in the clouds data.frame can be identified via

```
> rownames(clouds)[clouds$rainfall %in% c(bxpseeding$out, + bxpecho$out)]
[1] "1" "15"
```

where bxpseeding and bxpecho are variables created by boxplot in Figure 6.3. Now we shall not remove these observations but bear in mind during the modelling process that they may cause problems.

In this example it is sensible to assume that the effect of some of the other explanatory variables is modified by seeding and therefore consider a model that includes seeding as covariate and, furthermore, allows interaction terms for seeding with each of the covariates except time. This model can be described by the formula

```
> clouds_formula <- rainfall ~ seeding + + seeding:(sne + cloudcover + prewetness + echomotion) + + time
```

and the design matrix $X^\star$ can be computed via

```
> Xstar <- model.matrix(clouds_formula, data = clouds)
```

By default, treatment contrasts have been applied to the dummy codings of the factors seeding and echomotion as can be seen from the inspection of the contrasts attribute of the model matrix

```
> attr(Xstar, "contrasts")
$seeding
[1] "contr.treatment"
$echomotion
[1] "contr.treatment"
```

The default contrasts can be changed via the contrasts.arg argument to model.matrix or the contrasts argument to the fitting function, for example lm or aov as shown in Chapter 5.

However, such internals are hidden and performed by high-level model-fitting functions such as lm which will be used to fit the linear model defined by the formula clouds_formula:

```
> clouds_lm <- lm(clouds_formula, data = clouds)
> class(clouds_lm)
[1] "lm"
```
ANALYSIS USING R

R> data("clouds", package = "HSAUR2")
R> layout(matrix(1:2, nrow = 2))
R> bxpseeding <- boxplot(rainfall ~ seeding, data = clouds,
+   ylab = "Rainfall", xlab = "Seeding")
R> bxpecho <- boxplot(rainfall ~ echomotion, data = clouds,
+   ylab = "Rainfall", xlab = "Echo Motion")

Figure 6.3 Boxplots of rainfall.
The results of the model fitting is an object of class \texttt{lm} for which a \texttt{summary} method showing the conventional regression analysis output is available. The output in Figure 6.5 shows the estimates $\hat{\beta}$ with corresponding standard errors and $t$-statistics as well as the $F$-statistic with associated $p$-value.

Many methods are available for extracting components of the fitted model. The estimates $\hat{\beta}$ can be assessed via

\begin{verbatim}
R> betastar <- coef(clouds_lm)
R> betastar

(Intercept)
\end{verbatim}
### R Output of the Linear Model Fit for the clouds Data

```r
R> summary(clouds_lm)

Call:
  lm(formula = clouds_formula, data = clouds)

Residuals:
     Min      1Q  Median       3Q      Max
-2.5300 -1.1500 -0.2700  1.0400  4.3900

Coefficients:
             Estimate   Std. Error t value
(Intercept)  -0.3462      2.7877   -0.12
seedingyes   15.6829      4.4463    3.53
time         -0.0450      0.0251   -1.80
seedingno:sne  0.4198      0.8445    0.50
seedingyes:sne -2.7774      0.9284   -2.99
seedingno:cloudcover  0.3879      0.2179    1.78
seedingyes:cloudcover -0.0984      0.9284   -0.09
seedingno:prewetness  4.1083      3.6010    1.14
seedingyes:prewetness  1.5513      2.6929    0.58
seedingno:echomotionstationary  3.1528      1.9325    1.63
seedingyes:echomotionstationary  2.5906      1.8173    1.43

Pr(>|t|)     
(Intercept)  0.9031
seedingyes  0.0037
          
time         0.0959
seedingno:sne  0.6274
seedingyes:sne  0.0104
seedingno:cloudcover  0.0984
seedingyes:cloudcover  0.3885
seedingno:prewetness  0.7745
seedingyes:prewetness  0.5744
seedingno:echomotionstationary  0.1268
seedingyes:echomotionstationary  0.1776

Residual standard error: 2.2 on 13 degrees of freedom
Multiple R-squared: 0.716,   Adjusted R-squared: 0.497
F-statistic: 3.27 on 10 and 13 DF,  p-value: 0.0243
```

**Figure 6.5** R output of the linear model fit for the clouds data.
and the corresponding covariance matrix $\text{Cov}(\hat{\beta}^\star)$ is available from the `vcov` method:

```r
R> Vbetastar <- vcov(clouds_lm)
```

where the square roots of the diagonal elements are the standard errors as shown in Figure 6.5:

```r
R> sqrt(diag(Vbetastar))

(Intercept)     2.7877
seedinglyes     4.4463
time             0.0251
seedinglyes:nse 0.8445
seedinglyes:nse  0.9284
seedinglyes:cloudcover 0.2179
seedinglyes:cloudcover 0.1103
seedinglyes:prewetness 3.6010
seedinglyes:prewetness  2.6929
seedinglyes:echomotionstationary 1.9325
seedinglyes:echomotionstationary 1.8173
```

In order to investigate the quality of the model fit, we need access to the residuals and the fitted values. The residuals can be found by the `residuals` method and the fitted values of the response from the `fitted` (or `predict`) method:

```r
R> clouds_resid <- residuals(clouds_lm)
R> clouds_fitted <- fitted(clouds_lm)
```

Now the residuals and the fitted values can be used to construct diagnostic plots; for example the residual plot in Figure 6.7 where each observation is labelled by its number (using `textplot` from package `wordclouds`). Observations 1 and 15 give rather large residual values and the data should perhaps be reanalysed after these two observations are removed. The normal probability plot of the residuals shown in Figure 6.8 shows a reasonable agreement between theoretical and sample quantiles, however, observations 1 and 15 are extreme again.
ANALYSIS USING R

R> psymb <- as.numeric(clouds$seeding)
R> plot(rainfall ~ sne, data = clouds, pch = psymb,
+ xlab = "S-Ne criterion")
R> abline(lm(rainfall ~ sne, data = clouds,
+ subset = seeding == "no"))
R> abline(lm(rainfall ~ sne, data = clouds,
+ subset = seeding == "yes"), lty = 2)
R> legend("topright", legend = c("No seeding", "Seeding"),
+ pch = 1:2, lty = 1:2, bty = "n")

![Figure 6.6](image)

Figure 6.6 Regression relationship between S-Ne criterion and rainfall with and without seeding.
Figure 6.7  Plot of residuals against fitted values for clouds seeding data.

An index plot of the Cook’s distances for each observation (and many other plots including those constructed above from using the basic functions) can be found from applying the `plot` method to the object that results from the application of the `lm` function. Figure 6.9 suggests that observations 2 and 18 have undue influence on the estimated regression coefficients, but the two outliers identified previously do not. Again it may be useful to look at the results after these two observations have been removed (see Exercise 6.2).
ANALYSIS USING R

R> qqnorm(clouds_resid, ylab = "Residuals")
R> qline(clouds_resid)

![Normal Q–Q Plot](image)

**Figure 6.8** Normal probability plot of residuals from cloud seeding model `clouds_lm`. 
R> plot(clouds_lm)

Figure 6.9  Index plot of Cook’s distances for cloud seeding data.