

Learning Guide Unit 3

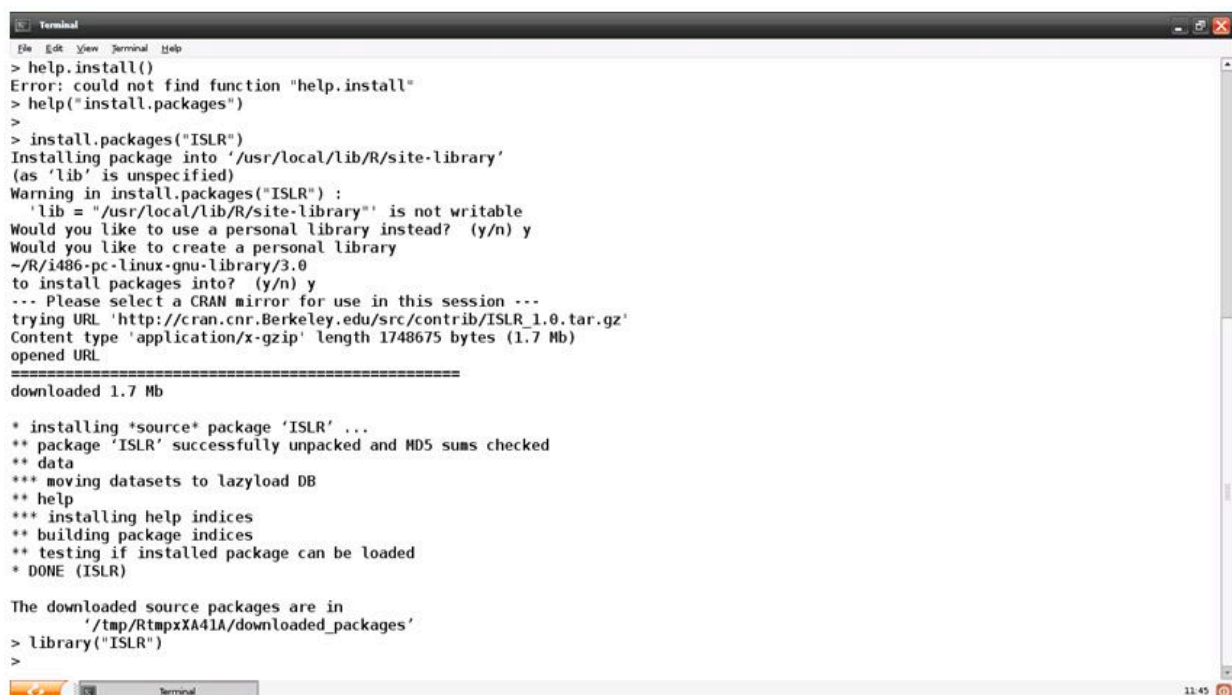
Programming Assignment

Follow the lab instructions in section 3.6 in the textbook for linear regression. If you are using an instance of R that is installed on your local computer then you may need to install either the MASS package or the ISLR package. A video tutorial has been provided in this unit to help you understand the installation procedure for instances of R installed on windows systems.

If you are using the Virtual Computing Lab, you may also be required install the packages. To install a package, simply issue the following command from the R command prompt. Note that this command installs the ISLR package but you can specify any package that is required.

```
> install.packages("ISLR")
```

An example of the installation process is shown in the following screen shot. When prompted to create a personal library you should respond "Y".



```

Terminal
File Edit View Terminal Help
> help.install()
Error: could not find function "help.install"
> help("install.packages")
>
> install.packages("ISLR")
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
Warning in install.packages("ISLR") :
  'lib = "/usr/local/lib/R/site-library"' is not writable
Would you like to use a personal library instead? (y/n) y
Would you like to create a personal library
~/R/i486-pc-linux-gnu-library/3.0
to install packages into? (y/n) y
--- Please select a CRAN mirror for use in this session ---
trying URL 'http://cran.cnr.Berkeley.edu/src/contrib/ISLR_1.0.tar.gz'
Content type 'application/x-gzip' length 1748675 bytes (1.7 Mb)
opened URL
=====
downloaded 1.7 Mb

* installing *source* package 'ISLR' ...
** package 'ISLR' successfully unpacked and MD5 sums checked
** data
*** moving datasets to lazyload DB
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (ISLR)

The downloaded source packages are in
  '/tmp/RtmpXA41A/downloaded_packages'
> library("ISLR")
>

```

As you follow the instructions for the Lab you will recognize that the `lm.fit` command is a function that has been developed to generate the linear regression model. The function can accommodate a single or multiple predictor variables meaning that it can do both single linear regression or multiple linear regression.

The output provided by the `lm.fit` command provides us with the 'fit' data to complete the regression model. You will recall from your reading that linear regression models take the form of a line which is expressed as:

$$y = mx + b \text{ or as it is detailed in the textbook: } Y \approx \beta_0 + \beta_1 X$$

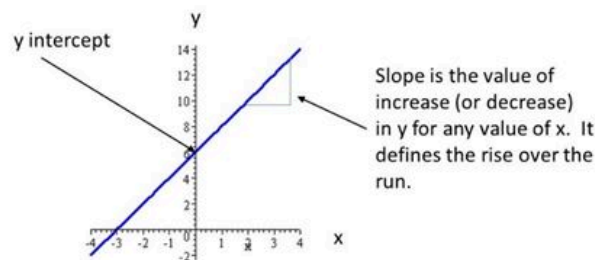
Where:

β_0 - Is the y intercept or the point where the fitted regression line intersects with the y axis when the value of X is 0.

β_1 - Is the slope of the line

X - is the input value of X

Y - is the predicted value



Follow the instructions for the lab. In it the predicted value is medv and the predictor value is lstat. When you run the lm.fit command it will output the Intercept and the β_1 value that is multiplied with the predictor value in our equation.

In the example in our textbook, the intercept will be 34.55 and the lstat value will be -0.95. This fits into our line equation as follows:

$$y = 34.44 + (-0.95 * X)$$

EXERCISES: Multiple Linear Regression Exercises

The following measurements have been obtained in a study:

| No. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|
| y | 1.45 | 1.00 | 0.01 | 0.81 | 1.55 | 0.95 | 0.45 | 1.14 | 0.74 | 0.90 | 1.41 | 0.01 | 0.09 |
| x1 | 0.58 | 0.88 | 0.29 | 0.20 | 0.58 | 0.28 | 0.08 | 0.41 | 0.22 | 0.35 | 0.59 | 0.22 | 0.28 |
| x2 | 0.71 | 0.13 | 0.79 | 0.20 | 0.58 | 0.92 | 0.01 | 0.60 | 0.70 | 0.73 | 0.13 | 0.98 | 0.27 |

| No. | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 |
|-----|------|------|------|------|------|------|------|------|------|------|------|------|
| y | 0.68 | 1.39 | 1.53 | 0.91 | 1.49 | 1.38 | 1.73 | 1.11 | 1.08 | 0.66 | 0.69 | 1.98 |
| x1 | 0.12 | 0.65 | 0.70 | 0.30 | 0.70 | 0.39 | 0.72 | 0.46 | 0.81 | 0.04 | 0.20 | 0.95 |
| x2 | 0.21 | 0.88 | 0.30 | 0.15 | 0.09 | 0.17 | 0.25 | 0.30 | 0.32 | 0.82 | 0.98 | 0.00 |

It is expected that the response variable y can be described by the independent variables x_1 and x_2 . This imply that the parameters of the following model should be estimated and tested.

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

a) Calculate the parameter estimates (β_0 , β_1 , β_2 and σ^2), in addition find the usual 95% confidence intervals for β_0 , β_1 , and β_2 .

You can copy the following lines to R to load the data:

```
D <- data.frame(
  x1=c(0.58, 0.86, 0.29, 0.20, 0.56, 0.28, 0.08, 0.41, 0.22, 0.35,
    0.59, 0.22, 0.26, 0.12, 0.65, 0.70, 0.30, 0.70, 0.39, 0.72,
    0.45, 0.81, 0.04, 0.20, 0.95),
  x2=c(0.71, 0.13, 0.79, 0.20, 0.56, 0.92, 0.01, 0.60, 0.70, 0.73,
    0.13, 0.96, 0.27, 0.21, 0.88, 0.30, 0.15, 0.09, 0.17, 0.25,
    0.30, 0.32, 0.82, 0.98, 0.00),
  y=c(1.45, 1.93, 0.81, 0.61, 1.55, 0.95, 0.45, 1.14, 0.74, 0.98,
    1.41, 0.81, 0.89, 0.68, 1.39, 1.53, 0.91, 1.49, 1.38, 1.73,
```

1.11, 1.68, 0.66, 0.69, 1.98)

)

b) Still using confidence level $\alpha = 0.05$ reduce the model if appropriate.

c) Carry out a residual analysis to check that the model assumptions are fulfilled.

d) Make a plot of the fitted line and 95% confidence and prediction intervals of the line for $x_1 \in [0, 1]$ (it is assumed that the model was reduced above).

MLR simulation exercise

The following measurements have been obtained in a study:

| Nr. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-----|------|-------|-------|------|-------|-------|-------|-------|
| y | 9.29 | 12.67 | 12.42 | 0.38 | 20.77 | 9.52 | 2.38 | 7.46 |
| x1 | 1.00 | 2.00 | 3.00 | 4.00 | 5.00 | 6.00 | 7.00 | 8.00 |
| x2 | 4.00 | 12.00 | 16.00 | 8.00 | 32.00 | 24.00 | 20.00 | 28.00 |

a) Plot the observed values of y as a function of x_1 and x_2 . Does it seem reasonable that either x_1 or x_2 can describe the variation in y?

You may copy the following lines into R to load the data

```
D <- data.frame(
  y=c(9.29,12.67,12.42,0.38,20.77,9.52,2.38,7.46),
  x1=c(1.00,2.00,3.00,4.00,5.00,6.00,7.00,8.00),
  x2=c(4.00,12.00,16.00,8.00,32.00,24.00,20.00,28.00)
)
```

b) Estimate the parameters for the two models

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

And

$$Y_i = \beta_0 + \beta_1 x_{2,i} + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

and report the 95% confidence intervals for the parameters. Are any of the parameters significantly different from zero on a 95% confidence level?