Linear Regression

Sebastian Heucke

9/30/2019

Contents

Linear Regression

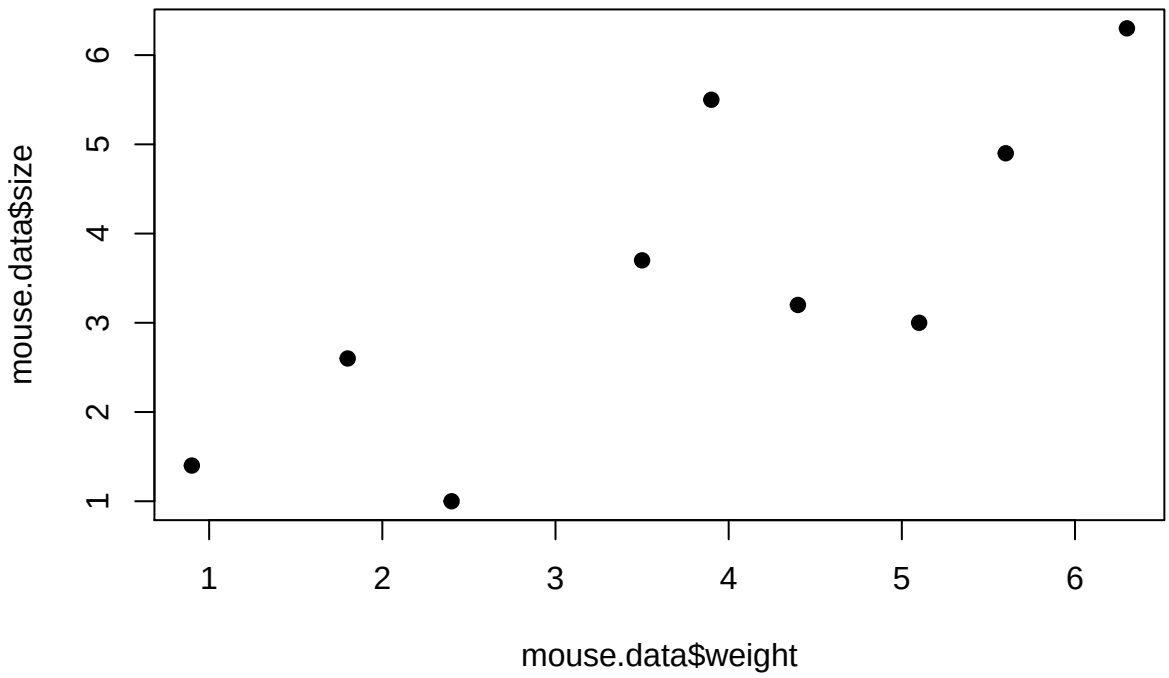
This tutorial walks you through a simple linear regression it is a markdown version of the https://www.youtube.com/watch?v=u1cc1r_Y7M0&list=PLblh5JKOoLUJJpBNfk8_YadPwDTO2SCbx&index=5&t=0s video.

```
# Here's the data from the example:
mouse.data <- data.frame(
  weight=c(0.9, 1.8, 2.4, 3.5, 3.9, 4.4, 5.1, 5.6, 6.3),
  size=c(1.4, 2.6, 1.0, 3.7, 5.5, 3.2, 3.0, 4.9, 6.3))

mouse.data # print the data to the screen in a nice format
```

```
##   weight size
## 1    0.9  1.4
## 2    1.8  2.6
## 3    2.4  1.0
## 4    3.5  3.7
## 5    3.9  5.5
## 6    4.4  3.2
## 7    5.1  3.0
## 8    5.6  4.9
## 9    6.3  6.3
```

```
## plot a x/y scatter plot with the data
plot(mouse.data$weight, mouse.data$size)
```



The formula for the linear regression follows:

$Y_{values} = Y_{intercept} + slope \times X_{values}$ $size = Y_{intercept} + slope \times weight$

The **lm()** function calculates the least squares estimates for intercept and slope.

```
# create a "linear model" - that is, do the regression
mouse.regression <- lm(size ~ weight, data=mouse.data)
# generate a summary of the regression
summary(mouse.regression)
```

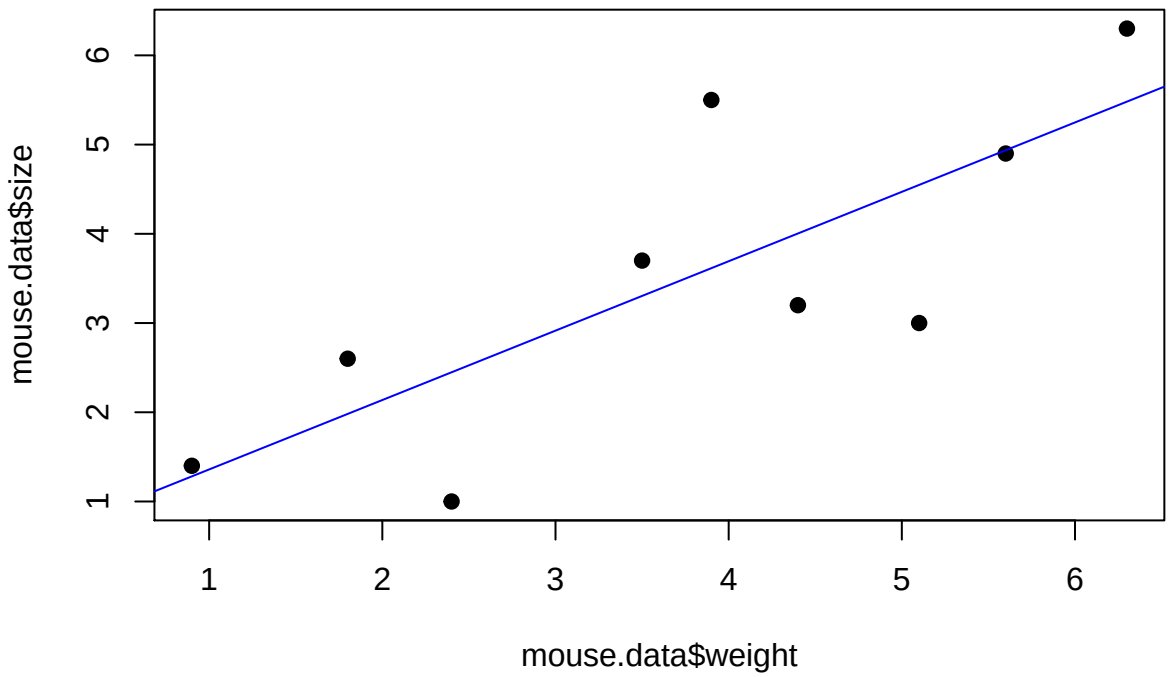
```
##
## Call:
## lm(formula = size ~ weight, data = mouse.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5482 -0.8037  0.1186  0.6186  1.8852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.5813     0.9647   0.603   0.5658
## weight        0.7778     0.2334   3.332   0.0126 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.19 on 7 degrees of freedom
## Multiple R-squared:  0.6133, Adjusted R-squared:  0.558
```

```
## F-statistic:  11.1 on 1 and 7 DF,  p-value: 0.01256
```

The first line just prints out the original call to the **lm()** function. After that is a summary of the residuals (the distamce from the data to the fitted line). Ideally, they should be symmetrically distributed around the line. Median should be close to zero and distance of min/max equal. The next section tells us about the least-squares estimates for the fitted line. The standard error of the estimates and the "t value" are both provided to show you how the p-values were calculated. Lastly, the p-values for the estimated parameters. Generally speaking, we are usually not intersted in the intercept, so it doesn't matter what its p-value is. However, we want the p-value for "weight" to be < 0.05. That is, we want it to be statistically significant. A significant p-value for weight means that it will give us a reliable guess of mouse size. I you were unable to read the actual p-value, but could for some reason, see the star to its right, then these codes would give you a sense of what the p-value was. The next line is the square root of the denominator in the equation for F. Multiple R-squared is just R^2. It means that weight can explain 61% of the variation in size. This is good! Generally speaking the "Adjusted R-squared" is the R^2 scaled by the number of parameters in the model. The F-statistic tells if the R^2 is signgificant or not. The p-value says that weight gives us a reliable estimate for size.

Add the regression line.

```
plot(mouse.data$weight, mouse.data$size)
# add the regression line to our x/y scatter plot
abline(mouse.regression, col="blue")
```



Session information

```
## R version 3.6.1 (2019-07-05)
## Platform: x86_64-pc-linux-gnu (64-bit)
```

```
## Running under: Ubuntu 18.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] knitr_1.23     devtools_2.1.0 usethis_1.5.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.1        magrittr_1.5      pkgload_1.0.2
##  [4] R6_2.4.0          rlang_0.4.0       stringr_1.4.0
##  [7] tools_3.6.1       pkgbuild_1.0.5    xfun_0.8
## [10] sessioninfo_1.1.1 cli_1.1.0         withr_2.1.2
## [13] remotes_2.1.0     htmltools_0.3.6   rprojroot_1.3-2
## [16] yaml_2.2.0        digest_0.6.20     assertthat_0.2.1
## [19] crayon_1.3.4      processx_3.4.1    callr_3.3.1
## [22] fs_1.3.1          ps_1.3.0          testthat_2.1.1
## [25] memoise_1.1.0     glue_1.3.1        evaluate_0.14
## [28] rmarkdown_1.14    stringi_1.4.3     compiler_3.6.1
## [31] backports_1.1.4   desc_1.2.0        prettyunits_1.0.2
```

The document was processed on 2019-09-30.