Math 362: Mathematical Statistics II

Le Chen
le.chen@emory.edu
chenle02@gmail.com

Emory University
Atlanta, GA

Last updated on Spring 2021
Last compiled on January 15, 2023

2021 Spring

Chapter 11. Regression

Chapter 11. Regression

	Indep. variables			Dependent variables		
Sample 1	x_{11}	$\cdots$	x_{1m}	y_{11}	$\cdots$	y_{1d}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Sample n	x_{n1}	$\cdots$	x_{nm}	y_{n1}	$\cdots$	y_{nd}

$$Y_{ij} = \sum_{k=1}^{m} \beta_{kj} X_{ik} + \epsilon_{ij}, \quad 1 \le i \le n, 1 \le j \le d, \ \epsilon_{ij} \ i.i.d. \sim N(0, \sigma^2).$$

$m = d = 1$	(Simple) linear regression
$m \ge 2$	Multiple linear regression
$d \ge 2$	Multivariate linear regression

1. Overdetermined system: $Y = XB$.

2. The least square solutions are (provided that $X^T X$ is nonsignular)

$$B = (X^T X)^{-1} X^T Y$$

E.g. Broadway shows[1]

```
1  > # This is an example of multimple regression.
2  > # Dataset is explained here:
3  > # https://dasl.datadescription.com/datafile/broadway-shows/?_sfm_methods=Multiple+
        Regression&_sfm_cases=4+59943&sort_order=title+asc
4  >
5  > # Read data from the URL link
6  > library(data.table)
7  > mydat <- fread('https://dasl.datadescription.com/download/data/3087')
8  [100*] Downloaded 965 bytes...
9  > head(mydat)
10    Season Gross($M) Attendance Playing weeks New Productions Mean ticket Pct.sold
          LogGross
11 1:  1984    209      7.26     1078            33    28.78788 0.04714286
        2.320146
12 2:  1985    190      6.54     1041            34    29.05199 0.04397695
        2.278754
13 3:  1986    208      7.04     1039            41    29.54546 0.04743022
        2.318063
14 4:  1987    253      8.14     1113            30    31.08108 0.05119497
        2.403120
15 5:  1988    262      7.96     1108            33    32.91457 0.05028881
        2.418301
16 6:  1989    282      8.04     1070            39    35.07463 0.05259813
        2.450249
```

---

[1] https://dasl.datadescription.com/datafile/broadway-shows/?_sfm_methods=Multiple+Regression&_sfm_cases=4+59943&sort_order=title+asc

```
> # Multiple Linear Regression Example with intercept
> fit  <- lm('Gross($M)' ~ Season + Attendance + 'Playing weeks' + 'New Productions' + 'Mean
    ticket' + 'Pct.sold' + LogGross, data=mydat)
> summary(fit) # show results

Call:
lm(formula = 'Gross($M)' ~ Season + Attendance + 'Playing weeks' +
    'New Productions' + 'Mean ticket' + Pct.sold + LogGross,
    data = mydat)

Residuals:
    Min     1Q  Median     3Q     Max
-31.925 -5.756 -0.055   7.172  14.040

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)        -2.053e+04 7.348e+03   -2.795 0.00983 **
Season              1.132e+01 3.829e+00    2.957 0.00670 **
Attendance          9.745e+01 3.537e+01    2.755 0.01079 *
'Playing weeks'     4.566e-02 3.084e-01    0.148 0.88348
'New Productions'  -9.560e-01 5.982e-01   -1.598 0.12255
'Mean ticket'       1.680e+01 8.306e-01   20.221 < 2e-16 *
Pct.sold            1.779e+03 6.811e+03    0.261 0.79604
LogGross           -1.301e+03 1.610e+02   -8.085 1.94e-08 *
---
Signif. codes:  0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.61 on 25 degrees of freedom
Multiple R-squared: 0.9994,   Adjusted R-squared: 0.9992
F-statistic:  6068 on 7 and 25 DF,  p-value: < 2.2e-16
```

114

```r
> # Compute the coefficients using the generalized inverse (with intercept)
> library (matlib)
> m <-length(mydat)-1
> M <- data.matrix(mydat, rownames.force = NA)
> n <- nrow(M)
> m <- ncol(M)
> X<- cbind(rep(1,n),M[1:n,c(1,3:m)])
> Y <- M[1:n,2]
> inv((t(X)*X)) * t(X) * Y
                         [,1]
                 -2.053451e+04
Season           1.132227e+01
Attendance       9.745043e+01
Playing weeks    4.565847e-02
New Productions -9.560446e-01
Mean ticket      1.679521e+01
Pct.sold         1.779471e+03
LogGross        -1.301463e+03
> # Or you can compute the generalized inverse use the package pracma
> library (pracma)
> pinv(X) *Y
                 [,1]
[1,]  -2.053451e+04
[2,]   1.132227e+01
[3,]   9.745043e+01
[4,]   4.565847e-02
[5,]  -9.560446e-01
[6,]   1.679521e+01
[7,]   1.779471e+03
[8,]  -1.301463e+03
```

```
1  > # Multiple  Linear Regression Example without intercept
2  > fit2  <- lm('Gross($M)' ~ Season + Attendance + 'Playing weeks' + 'New Productions' + 'Mean
       ticket' + 'Pct.sold'  + LogGross –1, data=mydat)
3  > summary(fit2) # show results
4
5  Call :
6  lm(formula = 'Gross($M)' ~ Season + Attendance + 'Playing weeks' +
7      'New Productions' + 'Mean ticket' + Pct.sold + LogGross –
8      1, data = mydat)
9
10 Residuals:
11     Min      1Q  Median      3Q     Max
12 –36.334 –3.758   2.570   6.282  18.324
13
14 Coefficients :
15                    Estimate Std. Error  t value Pr(>|t |)
16 Season              0.62744    0.15089    4.158 0.000309 *
17 Attendance         91.28669   39.65848    2.302 0.029610 *
18 'Playing weeks'     0.04173    0.34641    0.120 0.905047
19 'New Productions'  –0.74486    0.66658   –1.117 0.274032
20 'Mean ticket'      18.09840    0.77213   23.440 < 2e–16 *
21 Pct.sold         1369.35407 7649.90823    0.179 0.859323
22 LogGross         –990.63826  130.72506   –7.578 4.81e–08 *
23 –––
24 Signif . codes:  0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
25
26 Residual standard error: 11.92 on 26 degrees of freedom
27 Multiple  R–squared: 0.9998,  Adjusted R–squared: 0.9998
28 F– statistic : 2.069e+04 on 7 and 26 DF, p–value: < 2.2e–16
```

116

```
 1 > # Compute the coefficients using the generalized inverse (without intercept)
 2 > library(matlib)
 3 > m <-length(mydat)-1
 4 > M <- data.matrix(mydat, rownames.force = NA)
 5 > n <- nrow(M)
 6 > m <- ncol(M)
 7 > X <- M[1:n,c(1,3:m)]
 8 > Y <- M[1:n,2]
 9 > inv((t(X)*X)) * t(X) * Y
10                          [,1]
11 Season            0.62744066
12 Attendance       91.28668689
13 Playing weeks     0.04172758
14 New Productions  -0.74485881
15 Mean ticket      18.09839993
16 Pct.sold       1369.35406937
17 LogGross       -990.63826155
18 > # Or you can compute the generalized inverse use the package pracma
19 > library(pracma)
20 > pinv(X) *Y
21                 [,1]
22 [1,]      0.62744066
23 [2,]     91.28668689
24 [3,]      0.04172758
25 [4,]     -0.74485881
26 [5,]     18.09839993
27 [6,]   1369.35406890
28 [7,]   -990.63826154
```