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

§ 11.1 Introduction

§ 11.4 Covariance and Correlation

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.A Appendix Multiple/Multivariate Linear Regression

§ 11.5 The Bivariate Normal Distribution

Plan

§ 11.1 Introduction

§ 11.4 Covariance and Correlation

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.A Appendix Multiple/Multivariate Linear Regression

§ 11.5 The Bivariate Normal Distribution

Chapter 11. Regression

§ 11.1 Introduction

§ 11.4 Covariance and Correlation

§ 11.2 The Method of Least Squares

§ 11.3 The Linear Model

§ 11.A Appendix Multiple/Multivariate Linear Regression

§ 11.5 The Bivariate Normal Distribution

	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}

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

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

1. Overdetermined system: $Y = XB$.

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

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

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]

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

E.g. Broadway shows[1]

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

```
> # 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
```

```
> # Multiple  Linear  Regression Example without intercept
>  fit2  <– lm(‘Gross($M)‘ ~ Season + Attendance + ‘Playing weeks‘ + ‘New Productions‘ + ‘Mean
         ticket‘ + ‘Pct.sold‘  + LogGross –1, data=mydat)
> summary(fit2) # show results

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

Residuals:
    Min       1Q  Median      3Q     Max
–36.334 –3.758   2.570   6.282  18.324

Coefficients:
                     Estimate Std. Error  t  value Pr(>|t |)
Season                0.62744    0.15089   4.158 0.000309 *
Attendance           91.28669   39.65848   2.302 0.029610 *
‘Playing weeks‘       0.04173    0.34641   0.120 0.905047
‘New Productions‘    –0.74486    0.66658  –1.117 0.274032
‘Mean ticket‘        18.09840    0.77213  23.440  < 2e–16 *
Pct.sold           1369.35407 7649.90823 0.179 0.859323
LogGross           –990.63826 130.72506 –7.578 4.81e–08 *
–––
Signif. codes:  0 ’*’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 11.92 on 26 degrees of freedom
Multiple  R–squared: 0.9998,   Adjusted R–squared: 0.9998
F– statistic :  2.069e+04 on 7 and 26 DF, p–value: < 2.2e–16
```

```
> # Compute the coefficients using the generalized inverse (without intercept)
> library (matlib)
> m <-length(mydat)-1
> M <- data.matrix(mydat, rownames.force = NA)
> n <- nrow(M)
> m <- ncol(M)
> X <- M[1:n,c(1,3:m)]
> Y <- M[1:n,2]
> inv((t(X)*X)) * t(X) * Y
                                 [,1]
Season                     0.62744066
Attendance                91.28668689
Playing weeks              0.04172758
New Productions           -0.74485881
Mean ticket               18.09839993
Pct.sold                1369.35406937
LogGross                -990.63826155
> # Or you can compute the generalized inverse use the package pracma
> library (pracma)
> pinv(X) *Y
               [,1]
 [1,]    0.62744066
 [2,]   91.28668689
 [3,]    0.04172758
 [4,]   -0.74485881
 [5,]   18.09839993
 [6,] 1369.35406890
 [7,] -990.63826154
```