12 Multiple Regression

Chapter Preview

In Chapter 11, we used simple linear regression to predict one quantitative variable (y) using one other quantitative variable (x). Multiple regression utilizes more than one predictor variable to model the response variable.

12.1: Multiple Regression Model

Multiple regression (MR) extends simple linear regression (one predictor and one response variable) to allow for multiple predictor variables.

Population Models for SLR and MR:

SLR: $y = \beta_0 + \beta_1 x_1 + \epsilon$

MR: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon$

Estimated Regression Equations for SLR and MR:

SLR: $\hat{y} = b_0 + b_1 x_1$

MR: $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$

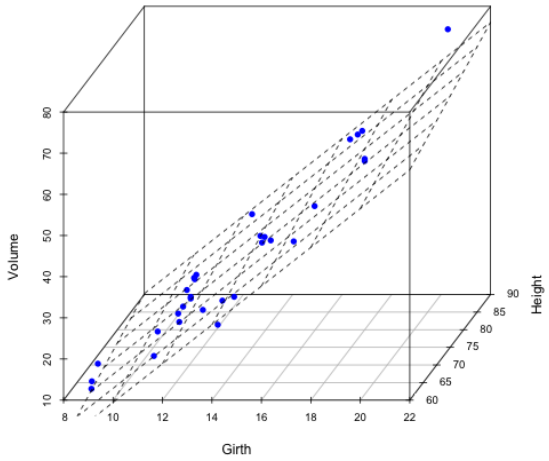
Assumptions for MR:

1. **Errors:** these are independent (and hence uncorrelated) and identically distributed normal random variables with mean 0 and variance σ^2.

2. **Linearity:** the mean of the response variables is a linear combination of the parameters and the predictor variables.

Note: Estimated regression coefficients can be calculated using matrix methods from Linear Algebra, however, we will focus on using R to do these calculations.

Example: The "trees" dataset in R provides measurements of the girth, height, and volume of 31 timber in 31 felled black cherry trees. We can fit a basic linear regression model with height and girth as predictor variables in R as follows.




```
data(trees)
mod = lm(Volume~Height+Girth,data=trees)
summary(mod)

##
## Call:
## lm(formula = Volume ~ Height + Girth, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

F-test for Overall Model (Model Utility Test)

We want to test if any of our predictor variables ($x_i's$) are linearly related to the response variable (y).

If we have p predictor variables, then our hypotheses will be:

$$\boxed{F_{test} = \frac{MSR}{MSE}, df_1 = p, df_2 = n - p - 1}$$

This will only tell us if any of our predictors are linearly related to the response, not which ones.

Example: Using the output from the previous page, complete an F-test for overall model for the cherry trees dataset.

t-test for Individual Predictors

If we reject H_o in our F-test, then we conclude that one or more of our predictor variables is linearly related to our response variable.

We can complete t-tests for each predictor to determine which are useful/significant. For example, we can test an individual predictor with the following hypothesis test:

$$\boxed{t_{test} = \frac{b_i}{s_{b_i}}, df = n - p - 1}$$

Rule of Thumb: If β_i is not significantly different from zero (p-value $\geq \alpha$), drop that predictor from the model.

Example: Using the output from the previous page, complete t-tests for individual predictors (height and diameter).

Interpreting the model estimates (b_i)

Example: Interpret b_2 from the cherry trees example in the context of the problem.

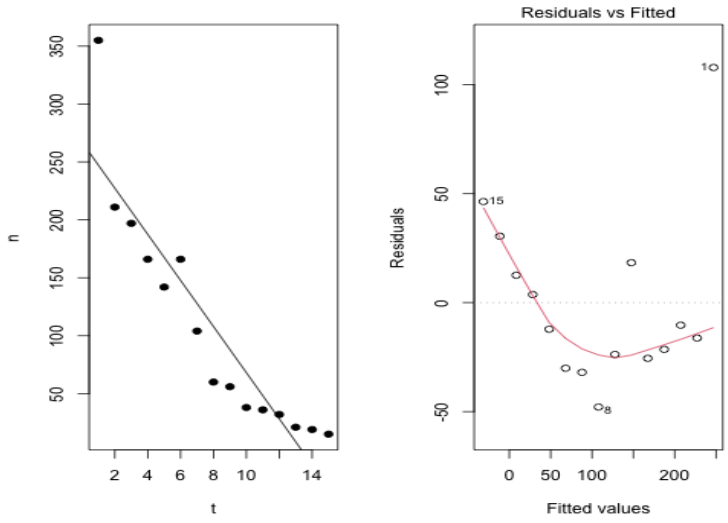
12.2: Variable Transformations

Sometimes a nonlinear regression function is appropriate which may require us to transform the data. An example of such a nonlinearity is given below.

Transforming Response Variable

Example: A colony of bacteria is exposed to X-rays and the number of surviving bacteria, n, at time t are recorded. Let's examine the data, the SLR line of best fit, and a residual plot.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_t	355	211	197	166	142	166	104	60	56	38	36	32	21	19	15



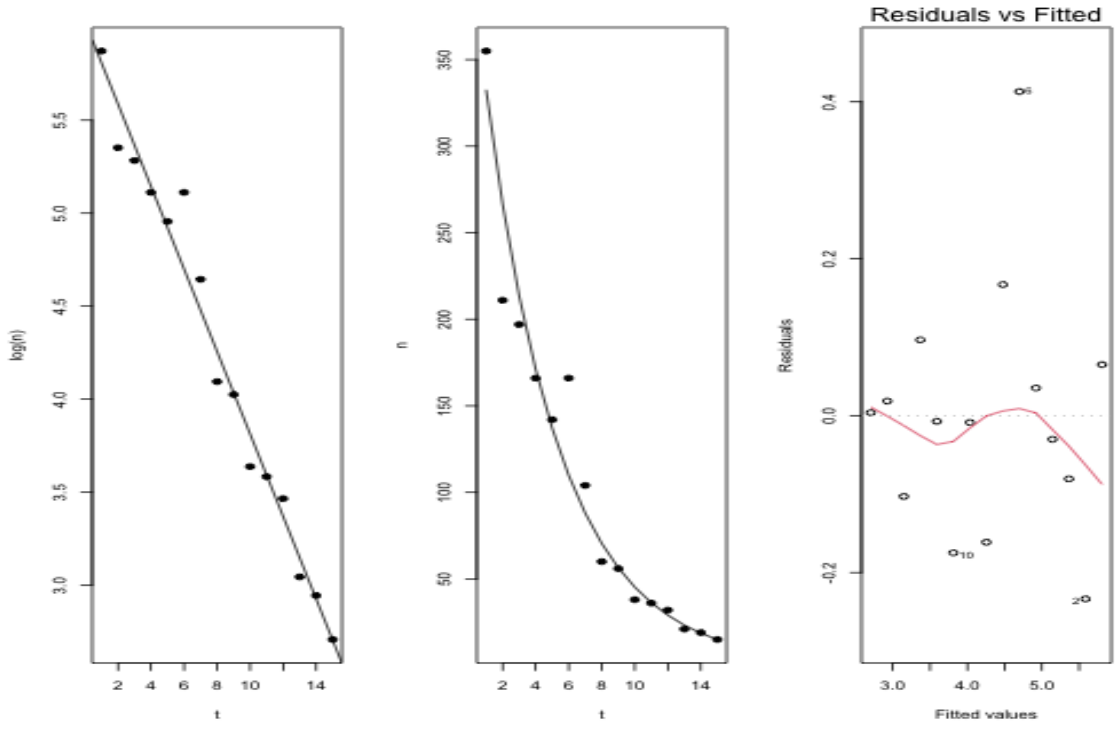
By inspection, the line of best fit does not seem appropriate. Additionally, the residual plot suggests that the data has a nonlinearity.

Instead, we could consider the following model: $n_t = n_0 \cdot e^{\beta t}$.

If we take the natural log of each side of this model, we get: $ln(n_t) = ln(n_0) + ln(e^{\beta t}) = \alpha + \beta t$.

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$ln(n_t)$	5.87	5.35	5.28	5.11	4.96	5.11	4.64	4.09	4.03	3.64	3.58	3.47	3.04	2.94	2.71

After making this transformation, we can refit the regression model with a much more appropriate model.



A regression model is said to be intrinsically linear if, by means of a transformation on Y or X_i, it can be reduced to a linear regression model, $y' = \beta_0 + \beta_1 x' + \epsilon'$.

Examples:

1. $y = \alpha e^{\beta x} \cdot \epsilon \rightarrow ln(y) = y' = \beta_0 + \beta_1 x' + \epsilon'$, where $x' = x$, $\beta_0 = ln(\alpha)$, $\beta_1 = \beta$, and $\epsilon' = ln(\epsilon)$

2. $y = \alpha x^{\beta} \cdot \epsilon \rightarrow log(y) = y' = \beta_0 + \beta_1 x' + \epsilon'$, where $x' = log(x)$, $\beta_0 = log(x) + \epsilon$, and $\epsilon' = log(\epsilon)$

Transforming Predictor Variables

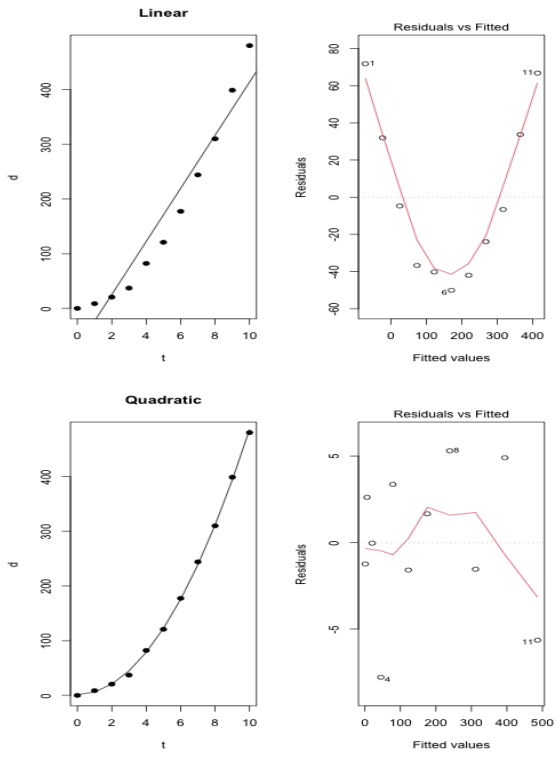
We can also transform our predictor variables, X_i, if a nonlinear model is appropriate. Consider the following example with a quadratic relationship.

Example: From Newtonian Physics, acceleration of an object in free fall drops a distance that is proportional to the square of the elapsed time. Consider the following dataset in which an object's dropped distance as a function of time and some measurement error is present.

t	0	1	2	3	4	5	6	7	8	9	10
d_t	0	8.7	20.6	37.1	82.2	120.9	177.5	244.2	310.1	399.0	480.6

First, we will fit the model $d = b_0 + b_1 t$ and then compare this model with a quadratic model, $d = b_0 + b_1 t^2$.

```
t = 0:10
d = c(0,8.7,20.6,37.1,82.2,120.9,177.5,244.2,310.1,399.0,480.6)
lm1 = lm(d~t)
lm2 = lm(d~I(t^2))
```



By inspection, the quadratic model seems more appropriate in terms of the line of best fit and in terms of the residuals.

12.3: Categorical Variables, Interaction, and Polynomial Regression

Categorical Variables

Categorical variables can be easily incorporated in a regression model. In fact, ANOVA from chapter 10 can be thought of as a regression on a categorical variable.

Example: The following dataset in R was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models).

```
data(mtcars)
head(mtcars,n=4)

##                  mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4       21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag   21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710      22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive 21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
```

To use a categorical variable in a regression model, you need to label it a "factor" variable in R. Let's fit a multiple regression model with miles per gallon as the response variable, weight as a continuous predictor variable, and cylinders (4, 6, or 8) as a categorical variable.

```
mt.lm = lm(mpg~wt+factor(cyl),data=mtcars)
summary(mt.lm)

##
## Call:
## lm(formula = mpg ~ wt + factor(cyl), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5890 -1.2357 -0.5159  1.3845  5.7915
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   33.9908     1.8878  18.006  < 2e-16 ***
## wt            -3.2056     0.7539  -4.252 0.000213 ***
## factor(cyl)6  -4.2556     1.3861  -3.070 0.004718 **
## factor(cyl)8  -6.0709     1.6523  -3.674 0.000999 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.557 on 28 degrees of freedom
## Multiple R-squared:  0.8374,Adjusted R-squared:   0.82
## F-statistic: 48.08 on 3 and 28 DF,  p-value: 3.594e-11
```

Example: For the Motor Trends car example, answer the following questions.

(a) Write the estimated multiple regression equation.

(b) What is the estimated MPG for a car that weighs 2.900 (thousands of pounds) and has 4 cylinders.

(c) What is the estimated difference in MPG between an 8-cylinder car and a 6-cylinder car?

(d) Interpret the estimated regression coefficients in the context of the problem.

Interaction

If the effect of one predictor variable on the response variable depends on other predictor variables, we may want to include an interaction variable.

Example: Let's fit the following model: $MPG = b_0 + b_1 \cdot Weight + b_2 \cdot I(am) + b_3 \cdot Weight \cdot I(am)$, where the "am" variable is 0 for an automatic transmission and 1 is for a manual transmission.

```
lm.int = lm(mpg~wt*factor(am),data=mtcars)
summary(lm.int)

##
## Call:
## lm(formula = mpg ~ wt * factor(am), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6004 -1.5446 -0.5325  0.9012  6.0909
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     31.4161     3.0201  10.402 4.00e-11 ***
## wt              -3.7859     0.7856  -4.819 4.55e-05 ***
## factor(am)1     14.8784     4.2640   3.489  0.00162 **
## wt:factor(am)1  -5.2984     1.4447  -3.667  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.591 on 28 degrees of freedom
## Multiple R-squared:  0.833,Adjusted R-squared:  0.8151
## F-statistic: 46.57 on 3 and 28 DF,  p-value: 5.209e-11
```

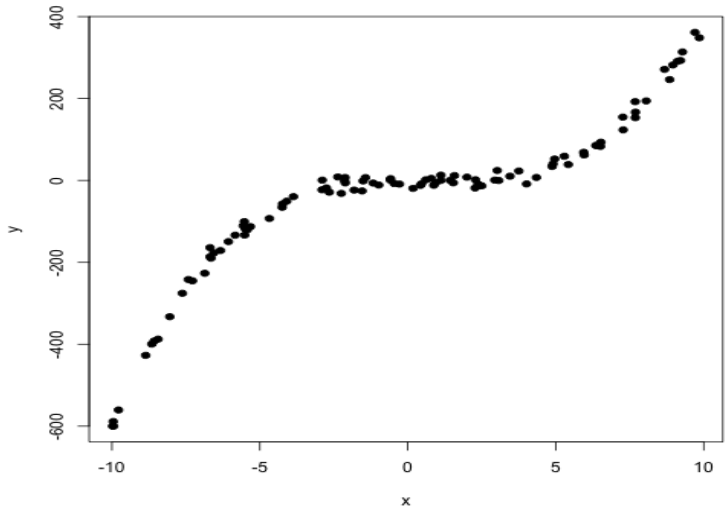
Polynomial regression

There is flexibility in a regression model to fit an arbitrary polynomial to the data, however, we must take care to not "overfit" the data.

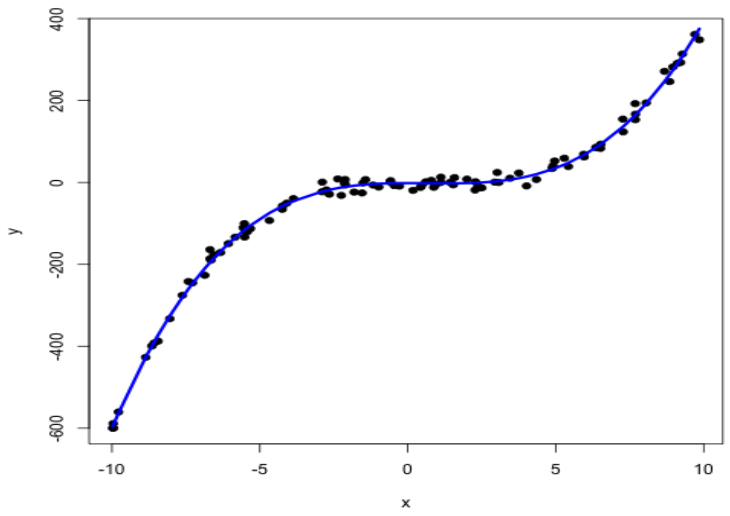
Example: Let's generate a toy dataset based a cubic polynomial and see how successfully we can recover the original function.

```
set.seed(2020)
x = sort(runif(n=100,min=-10,max=10))
y = 1/2*x^3 - x^2 - 2 + rnorm(100,sd=10)
toydata = cbind(x,y)
head(toydata)
```

```
##              x         y
## [1,] -9.976245 -599.8586
## [2,] -9.957027 -588.7105
## [3,] -9.948346 -599.9990
## [4,] -9.779192 -560.4769
## [5,] -8.859972 -427.0615
## [6,] -8.652312 -399.5171
```



```
##
## Call:
## lm(formula = y ~ x + I(x^2) + I(x^3))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.9139  -8.4193   0.2316   6.0311  31.3370
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.582774   1.716694  -0.922    0.359
## x            0.002634   0.516494   0.005    0.996
## I(x^2)      -1.013334   0.039545 -25.625   <2e-16 ***
## I(x^3)       0.497069   0.007792  63.792   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.92 on 96 degrees of freedom
## Multiple R-squared:  0.9961,Adjusted R-squared:  0.996
## F-statistic:  8212 on 3 and 96 DF,  p-value: < 2.2e-16
```



12.4: Model and Variable Selection

Returning to the cherry trees example, consider the following regression models:

Model 1: $\widehat{volume} = b_0 + b_1 \cdot Height + b_2 \cdot Diameter$

Model 2: $\widehat{volume} = b_0 + b_1 \cdot Height + b_2 \cdot Diameter^2$

Model 3: $\widehat{volume} = b_0 + b_1 \cdot Height \cdot Diameter^2$

Model 4: $\widehat{volume} = b_1 \cdot Height \cdot Diameter^2$

Model 4 seems most appropriate based on how volume is calculated ($Volume \propto Height \cdot Diameter^2$), but how could we determine this from the data had we not known this fact?

```
lm1 = lm(Volume~Height+Girth,data=trees)
summary(lm1)

##
## Call:
## lm(formula = Volume ~ Height + Girth, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

```r
lm2 = lm(Volume~Height+I(Girth^2),data=trees)
summary(lm2)
```

```
##
## Call:
## lm(formula = Volume ~ Height + I(Girth^2), data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8844 -2.2105  0.1196  2.6134  4.2404
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -27.511603   6.557697  -4.195 0.000248 ***
## Height        0.348809   0.093152   3.744 0.000830 ***
## I(Girth^2)    0.168458   0.006679  25.222  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.799 on 28 degrees of freedom
## Multiple R-squared:  0.9729,Adjusted R-squared:  0.971
## F-statistic: 503.2 on 2 and 28 DF,  p-value: < 2.2e-16
```

```r
lm3 = lm(Volume~Height:I(Girth^2),data=trees)
summary(lm3)
```

```
##
## Call:
## lm(formula = Volume ~ Height:I(Girth^2), data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6195 -1.1002 -0.1656  1.7451  4.1976
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -2.977e-01  9.636e-01  -0.309     0.76
## Height:I(Girth^2)  2.124e-03  5.949e-05  35.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.493 on 29 degrees of freedom
## Multiple R-squared:  0.9778,Adjusted R-squared:  0.977
## F-statistic:  1275 on 1 and 29 DF,  p-value: < 2.2e-16
```

```r
lm4 = lm(Volume~Height:I(Girth^2)-1,data=trees)
summary(lm4)

##
## Call:
## lm(formula = Volume ~ Height:I(Girth^2) - 1, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6696 -1.0832 -0.3341  1.6045  4.2944
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## Height:I(Girth^2) 2.108e-03  2.722e-05   77.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.455 on 30 degrees of freedom
## Multiple R-squared:  0.995,Adjusted R-squared:  0.9949
## F-statistic:  5996 on 1 and 30 DF,  p-value: < 2.2e-16
```