تخمین معادلات رگرسیون

(R)

تدوین: مرکز تحلیل آماری خوارزمی

***

www.kharazmi-statistics.ir

مرکز آماری خوارزمی

مقدمه:

نرم افزار R در رشته های مختلف از جمله در تحلیل های آماری و اقتصادی کاربرد فراوانی دارد. همچنین از این نرم افزار در تحلیل های اقتصاد سنجی و رگرسیونی استفاده می شود. در این متن آموزش فراخوانی داده ها از منابع خارجی و چگونگی تخمین رگرسیون حداقل مربعات در این نرم افزار آورده شده است.

فراخوانی داده ها از منابع خارجی

نرم افزار R تنها قادر به خواندن فایلهایی با پسوند txt و csv است و قادر به خواندن فایل هایی با فرمت اکسل با پسوند xls یا xlsx نیست. در نتیجه این فایل ها را باید به فرمت هایی که نرم افزار قادر به خواندن آنها است در بیاورید و برای فراخوانی داده ها از فرمت های txt باید از دستور read.table و با فرمت csv از دستور read.csv استفاده نمایید.

برای فراخوانی داده هایی ذخیره شده از فرمت txt دستور زیر را در صفحه ی دستورات تایپ نمایید.

```
> s<- read.table("C:/users/taheri/desktop/ghad.txt")
```

نام دلخواه

آدرسی که داده ها در آن ذخیره شده اند.

به نحوه ی نوشتن آدرس داده ها دقت نمایید. برای مشاهده آدرس دقیق داده می تواند با راست کلیک و از قسمت properties آدرس فایل را کپی نمایید. (لطفا به علامت ها که در نوشته ی بالا است دقت نمایید و در قرار دادن علائم / ، : ، " " دقیق باشید.)

```
> s<- read.table("C:/users/taheri/desktop/ghad.txt")
> s
   V1  V2 V3 V4
1  21 165 50  6
2  12 160 60  7
3  43 180 90  7
4  51 140 49  9
5  32 180 60  7
6  15 169 70  8
7  32 158 65  5
8  12 130 56  7
9  15 149 64  8
10 18 169 81  7
11 39 171 84  6
12 44 155 70  8
13 29 163 56  5
14 31 179 83  9
```

پس از نوشتن دستور فراخوانی در خط بعد اسم فایلی که در نرم افزار انتخاب کرده اید را نوشته و کلید enter را بزنید. تمام داده ها در صفحه ی نرم افزار نمایش داده می شود.

گاهی ممکن است یک یا چند سطر اول فایل های شما توضیحات و یا اسامی متغیرهایی نوشته شده باشد. در

این حالت با دستور skip=n از نرم افزار خواسته می شود تا سطر nام نادیده گرفته شود. دستور به صورت زیر نوشته می شود.

```
>x<- read.table("C:/users/taheri/desktop/q.txt",skip=1)
```

یعنی سطر اول داده ها در نظر گرفته نشود.

نرم افزار به صورت پیش فرض اسامی را با حرف V اندیس دار نمایش می دهد. برای تغییر این اسامی از فرمان زیر استفاده نمایید.

```
>names(s)<-c("sen","ghad","vazn","factor.salamat")
```

```
> names(s)<-c("sen","ghad","vazn","faktor.salamati")
> s
   sen ghad vazn faktor.salamati
1   21  165   50               6
2   12  160   60               7
3   43  180   90               7
4   51  140   49               9
5   32  180   60               7
6   15  169   70               8
7   32  158   65               5
8   12  130   56               7
9   15  149   64               8
10  18  169   81               7
11  39  171   84               6
12  44  155   70               8
13  29  163   56               5
14  31  179   83               9
```

همانطور که در قبل دیدید فایل s دارای ۴ ستون است. این فایل شامل داده های از سن، وزن، قد و فاکتور سلامتی است که می خواهیم مدل رگرسیونی با توجه به این متغیر ها را بدست آوریم.

تخمین رگرسیونی مدل به روش حداقل مربعات معمولی

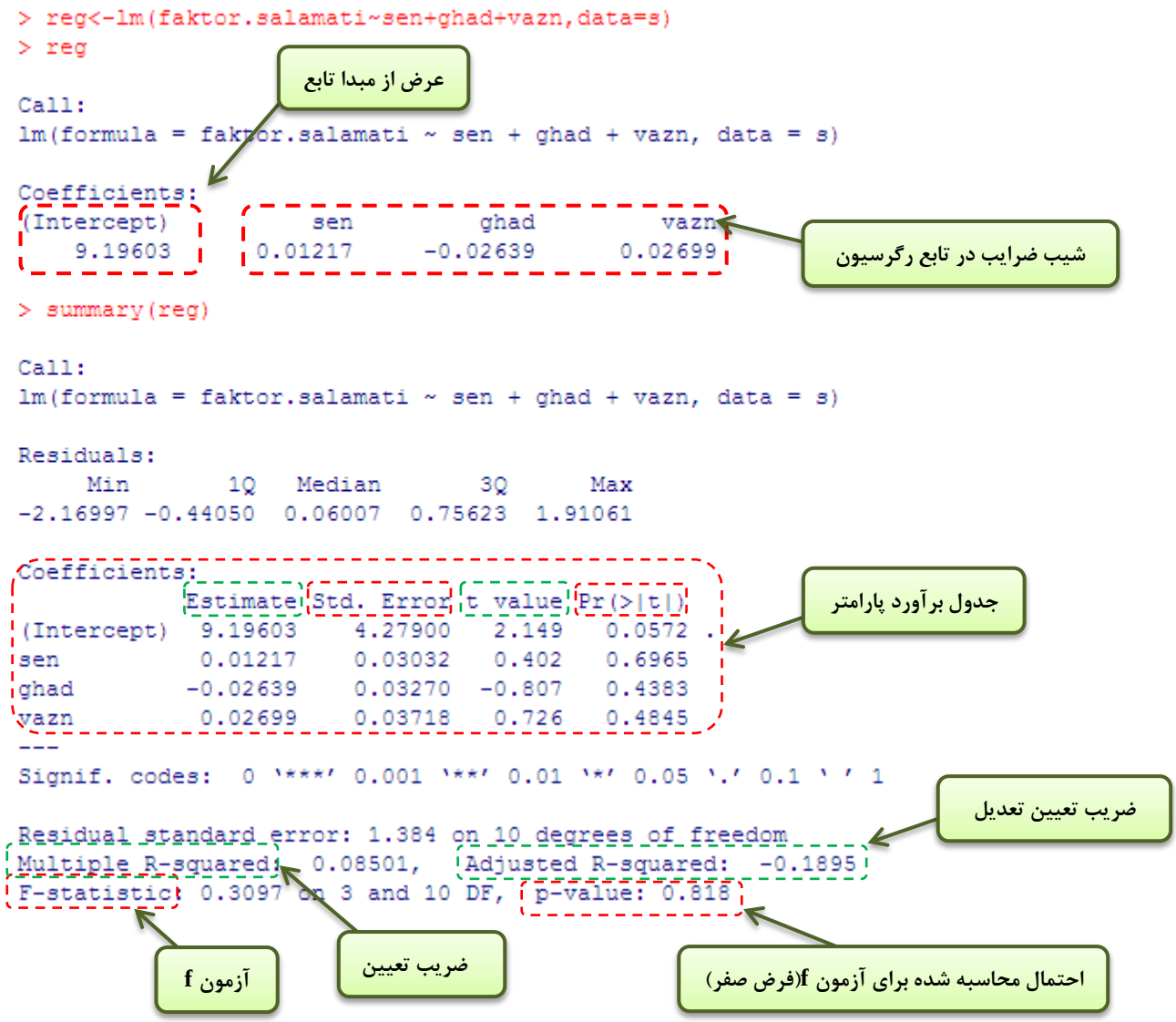
برای تخمین یک مدل رگرسیون متغیر وابسته و مستقل وجود دارد. فرض بر آن است که در مثال بیان شده در بالا برای تابع s ، factor.salamati متغیر وابسته و متغیر های ghad, vazn, sen مستقل هستند. برای تخمین مدل دستور زیر را تایپ نمایید.

```
> reg<-lm(faktor.salamati~sen+ghad+vazn,data=s)
> reg
```

data به این معناست که نرم افزار برای تخمین مدل از کدامیک از داده ها استفاده نماید. عبارت روبروی data نیز اسمی است که برای فایلی که داده ها در آن قرار دارند در نظر گرفته شده است. reg نامی است که برای تخمین رگرسیون به صورت دلخواه نوشته شده است. با تایپ عبارت reg و summary(reg) خروجی مشابه تصویر پایین نمایش داده می شود. عبارت lm برای فراخوانی مدل رگرسیونی است. در نرم افزار R برای

فراخوانی توابع از عبارت lm , glm و .... استفاده می شود. با توجه به اعداد بدست آمده معادله ی زیر مدل رگرسیونی بدست آمده است.

$$y = 9.19603 + 0.01217 * sen - 0.02639 * ghad + 0.02699 * vazn$$




```
> reg<-lm(faktor.salamati~sen+ghad+vazn,data=s)
> reg

Call:
lm(formula = faktor.salamati ~ sen + ghad + vazn, data = s)

Coefficients:
(Intercept)          sen         ghad         vazn
    9.19603      0.01217     -0.02639      0.02699

> summary(reg)

Call:
lm(formula = faktor.salamati ~ sen + ghad + vazn, data = s)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16997 -0.44050  0.06007  0.75623  1.91061

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.19603    4.27900   2.149   0.0572 .
sen          0.01217    0.03032   0.402   0.6965
ghad        -0.02639    0.03270  -0.807   0.4383
vazn         0.02699    0.03718   0.726   0.4845
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.384 on 10 degrees of freedom
Multiple R-squared:  0.08501,   Adjusted R-squared:  -0.1895
F-statistic: 0.3097 on 3 and 10 DF,  p-value: 0.818
```

علامت * در کنار عدد احتمال به معنی معنا دار بودن آن ضریب در سطح معنی داری است که در قسمت signif.codes معرفی شده است. در صورتی که در کنار یکی از احتمال ها علامت * قرار داده شده باشد متغیر مورد نظر در نرخ متغیر وابسته آزمون شده بی اثر است.(مثلا با قرار گرفتن ** در کنار احتمال متغیر قد یعنی در سطح ۰.۰۱ متغیر قد در متغیر نرخ سلامتی بی اثر است.)

تحلیل خروجی نرم افزار

در انتهای خروجی نرم افزار آزمونی با آماره ی f فیشر انجام شده است. با توجه به فرض صفر آن چون عدد محاسبه شده بزرگتر از ۰.۰۵ است در نتیجه فرض صفر در سطح ۰.۰۵ رد نمی شود و در حقیقت دلیلی بر رد فرض صفر وجود ندارد. مقدار مربع r یا ضریب تعیین مقدار خوبی برازش را نشان می دهد. مقدار بالاتر بیانگر برازش بهتر است و در مقایسه ی متدهای مختلف و یا معادلات مختلف به کار می رود. با ضرب عدد ضریب تعیین در ۱۰۰ می توان گفت که این مدل می تواند چند درصد از تغییرات متغیر وابسته را تبیین کند. در خروجی بالا مدل بدست آمده تنها ۸ درصد تغییرات متغیر وابسته را تبیین می کند که می توان گفت برازش خوبی نیست. مربع r تعدیل شده که همان ضریب تعیین است که وابسته به درجات آزادی به تعدیل شده است. چنانچه تعداد متغیر های مستقل را افزایش دهید بهتر است که به این آماره تکیه کنید و نه ضریب تعیین عادی. در جدول برآورد پارامترها یا همان ضرایب رگرسیونی این موارد برای هرکدام از ضرایب به چشم می خورد. مقدار برآورد شده(Esimate)، خطای استاندارد(Std.Error) ، مقدار آمارT(t value) ، مقدار احتمال آماره ی Tجهت آزمون فرض صفر بی اثر بودن یا صفر بودن ضرایب($Pr(>|t|)$). با توجه به مقادیر بدست آمده متغیرهای سن، وزن و قد در نرخ فاکتور سلامتی اثرگذار است.

مطالب تکمیلی

اگر مدلی که در حال بررسی است کاربر بخواهد اثر متقابل دو متغیر را نیز بررسی کنید باید از دستور زیر استفاده نماید.

اثر متقابل vazn و sen

```
> reg1<-lm(faktor.salamati~sen+ghad+vazn+sen:vazn,data=s)
```

عبارت sen:vazn اثر متقابل دو متغیر sen و vazn را در محاسبات لحاظ می کند.

```
> reg1

Call:
lm(formula = faktor.salamati ~ sen + ghad + vazn + sen:vazn,
    data = s)

Coefficients:
(Intercept)          sen         ghad         vazn     sen:vazn
   1.470391     0.229278    -0.024106     0.144021    -0.003441
```

در ادامه معادلات رگرسیون در حالت های مختلف و نحوه ی نوشتن در نرم افزار را بیان می کنیم. و پس از آن نمونه هایی از خروجی نرم افزار نمایش داده شده است.

نحوه ی نوشتن در نرم افزار	معادله ی رگرسیون
y~x1+x2 *	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ **
y~x1+x2+x1:x2	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$
y~(x1+x2+x1:x3)^2 یا y~x1+x2+x3+x1:x2+x2:x3+x1:x3+x1:x2:x3	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{23} x_2 x_3 + \beta_{13} x_1 x_3 + \beta_{123} x_1 x_2 x_3 \varepsilon$
y~(x1+x2+x1:x2)^2- x2:x3	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{123} x_1 x_2 x_3 \varepsilon$
y~I(x1+x2)	$y = \beta_0 + \beta_1 (x_1 + x_2) + \varepsilon$
y~x1+I(2*x2)	$y = \beta_0 + \beta_1 x_1 + 2 * \beta_2 x_2 + \varepsilon$
y~x1+x2+poly(x3^2) ***	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + + \beta_3 x_3^2 + \varepsilon$

* به صورت پیش فرض معادله در نظر گرفته شده در R دارای عرض از مبدا است. اگر رابطه فوق را به صورت روبرو بنویسید عرض از مبدا حذف می شود. y~ -1+x1+x2

دقت داشته باشید که عملگر – جمله یا جملاتی که در معادله رگرسیون هستند حذف می کند.

** ε خطای واریانس است که غالبا دارای توزیع نرمال با میانگین صفر و واریانس σ_ε^2 است.

*** برای نوشتن توان های دیگر در معادله می توان با از تابع poly و نوشتن جمله ای مشابه بالا دستور مورد نظر را به نرم افزار داد.

در اینجا چون دو متغیر مستقل q1 وq2 معرفی شده است برای حالت تعاملی، تعامل دو جمله ای و سه جمله ای بیان شده است.

```
> ols<-lm(q4~(q1+q2+q1:q4)^2,data=s)
> ols

Call:
lm(formula = q4 ~ (q1 + q2 + q1:q4)^2, data = s)

Coefficients:
(Intercept)           q1           q2        q1:q2        q4:q1     q4:q1:q2
  8.480e+00   -2.409e-01   -7.015e-03    6.620e-05    1.878e-02    7.445e-05
```

در این معادله یک ضریب برای متغیر های q1 وq2 وجود دارد. و عدد بدست آمده برای ضریب 0.342 است.

```
> P<-lm(q3~I(q1+q2),data=s)
> P

Call:
lm(formula = q3 ~ I(q1 + q2), data = s)

Coefficients:
(Intercept)    I(q1 + q2)
      1.962         0.342
```

متغیر q4 در عدد دو ضرب شده و در معادله قرار گرفته است. دقت نمایید که نوشتن عبارت به صورت q3~q2+2*q4 اشتباه است.

```
> G<-lm(q3~q2+I(2*q4),data=s)
> G

Call:
lm(formula = q3 ~ q2 + I(2 * q4), data = s)

Coefficients:
(Intercept)           q2      I(2 * q4)
   -38.0684       0.5624         0.9869
```

در معادله ی روبرو متغیر q3 به توان دو دارد و در معادله رگرسیون ضریب 0.042314 محاسبه شده است. نوشتن عبارت توان دو به صورت زیر اشتباه است.

q3~q2+q4^2

```
> h<-lm(q4~poly(q3^2)+q1,data=s)
> h

Call:
lm(formula = q4 ~ poly(q3^2) + q1, data = s)

Coefficients:
(Intercept)   poly(q3^2)           q1
    6.76926      0.42314      0.01074
```

منبع:

- ✓ آموزش تخمین رگرسیون به روش حداقل مربعات معمولی در نرم افزارR ، حسین خاندانی، econometrics.blog.ir
- ✓ موسسه فرهنگی دیجیتال بهگامان www.behkaman.ir
- ✓ آشنایی با زبان محاسبات R، سید سعید موسوی ندوشنی، پاییز ۱۳۹۱