آموزش کاربردی GLM
در SAS ، SPSS و Minitab


http://www.DaneshAmari.ir

نویسنده : سید جمال میرکمالی
گروه دانش آماری – فروردین ۸۶

روال GLM[1]

روال GLM در SAS از شیوه کمترین مربع خطا برای برازش مدل های خطی معمولی استفاده می کند. از جمله روش های آماری که در روال GLM موجود است می توان به رگرسیون ، تحلیل واریانس ، تحلیل کواریانس ، تحلیل واریانس چند متغیره و همبستگی جزیی اشاره کرد.

روال GLM با مدل هایی سر و کار دارد که یک یا چند متغیر پیوسته ی وابسته را به یک یا چند متغیر مستقل مرتبط می کنند. متغیر های مستقل هم می توانند پیوسته باشند و هم می توانند از نوع طبقه ای باشند که مشاهدات را به چند گروه گسسته تقسیم کنند. روال GLM می تواند در تحلیل های گوناگونی مورد استفاده قرار بگیرند:

- رگرسیون ساده
- رگرسیون چندگانه
- تحلیل واریانس(ANOVA) – بخصوص برای داده های نا متعادل
- تحلیل کواریانس
- مدل رویه پاسخ
- رگرسیون وزنی
- رگرسیون چند جمله ای
- همبستگی جزیی
- تحلیل واریانس چند متغیره (MANOVA)
- تحلیل واریانس اندازه های مکرر[2]

پیش فرض های آماری در بکارگیری روال GLM

فرض اساسی برای بکارگیری شیوه کمترین مربعات خطا در مدل خطی عمومی این است که مقادیر مشاهده شده متغیر وابسته را بتوان بصورت حاصل جمع دو جزء نوشت : یکی مولفه ثابت $\mathbf{x}'\boldsymbol{\beta}$ ، که یک ترکیب خطی از ضرایب مستقل است ؛ دیگری عامل تصادفی اخلال، یا مولفه خطای $\boldsymbol{\varepsilon}$:

$$\mathbf{y} = \mathbf{x}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

---

[1] General Linear Model
[2] repeated measures analysis of variance

باید خطا های هر مشاهده با خطای دیگر مشاهدات نا همبسته باشد و واریانس آن ثابت باشد. در این صورت برای این مدل می توان نوشت :

$$E(\mathbf{Y}) = \mathbf{X\beta} \quad , \quad Var(\mathbf{Y}) = \sigma^2 \mathbf{I}$$

که در آن $\mathbf{Y}$ بردار متغیرهای وابسته است و $\mathbf{X}$ ماتریس ضرایب مستقل است و $\mathbf{I}$ ماتریس یکه و σ^2 واریانس مشترک خطا ها است. تحت فرضیاتی که تا کنون مطرح شد ، شیوه کمترین مربعات خطا برآورد هایی برای پارامتر ارائه می دهد که نا اریب و دارای کمترین واریانس در بین مدل های خطی می باشد. در صورتی که خطا ها توزیع نرمال داشته باشند ، برآورد های کمترین مربعات MLE خواهند بود.

تذکر: همه $P\text{-}Value$ ها و حدود اطمینان که توسط روال GLM محاسبه می شوند شرط نرمال بودن را برای اعتبار عینی نیاز دارند. گر چه این ها در بیشتر موارد تقریب های خوبی ارائه می دهند.

مثال: رنگ آستر هواپیما روی سطح آلومینیوم به دو روش اجرا می شود: روش فروبری و روش افشاندن. رنگ آستر به منظور بهبود چسبندگی رنگ اعمال می شود. مسئول گروه مهندسی تولید علاقمند است بداند تفاوت در سه نوع آستر موجب تفاوت در میزان چسبندگی می گردد یا خیر. یک طرح عاملی برای بررسی تاثیر نوع آستر و شیوه اجرا تهیه شده است. برای هر ترکیب از نوع آستر و شیوه اجرا سه نمونه رنگ شده ، سپس رنگ نهایی زده شده و میزان چسبندگی آن اندازه گیری شده است. داده های اندازه گیری شده در جدول زیر آورده شده است:

نوع آستر (*Primer*)	نوع اجرا (*Method*)	
	فروبری	افشاندن
۱	۴.۰ , ۴.۵ , ۴.۳	۵.۴ , ۴.۹ , ۵.۶
۲	۵.۶ , ۴.۹ , ۵.۴	۵.۸ , ۶.۱ , ۶.۳
۳	۳.۸ , ۳.۷ , ۴.۰	۵.۵ , ۵.۰ , ۵.۰

در این حالت مدل عبارت است از :

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk} \qquad \begin{cases} i = ۱,۲ \\ j = ۱,۲,۳ \\ k = ۱,۲,۳ \end{cases}$$

که در آن τ_i اثر اجرای i ام ، β_j اثر آستر j ام ، $(\tau\beta)_{ij}$ اثر متقابل آستر j ام و اجرای i ام ، و ε_{ijk} خطای تصادفی حاصل از تکرار ها است.

با استفاده از نرم افزار SAS داده ها را تحلیل می کنیم:

```
data Aircraft;
 input Primer Method Y;
 datalines;
 1 1 4.0
 1 1 4.5
 1 1 4.3
 1 2 5.4
 1 2 4.9
 1 2 5.6
 2 1 5.6
 2 1 4.9
 2 1 5.4
```

```
 2 2 5.8
 2 2 6.1
 2 2 6.3
 3 1 3.8
 3 1 3.7
 3 1 4.0
 3 2 5.5
 3 2 5.0
 3 2 5.0
 ;
proc glm data=Aircraft;
 class Primer Method;
 model Y=Primer Method Primer*Method;
 output out=rcheck p=yhat r=resid;
run;
```

```
                                The SAS System
                              The GLM Procedure
                            Class Level Information
                         Class         Levels    Values
                         Primer             3    1 2 3
                         Method             2    1 2

                        Number of Observations Read          18
                        Number of Observations Used          18


Dependent Variable: Y
                                        Sum of
Source                      DF         Squares     Mean Square    F Value    Pr > F
Model                        5      9.73111111      1.94622222      23.67    <.0001
Error                       12      0.98666667      0.08222222
Corrected Total             17     10.71777778

                  R-Square     Coeff Var      Root MSE        Y Mean
                  0.907941      5.747656      0.286744      4.988889

Source                      DF       Type I SS     Mean Square    F Value    Pr > F
Primer                       2      4.58111111      2.29055556      27.86    <.0001
Method                       1      4.90888889      4.90888889      59.70    <.0001
Primer*Method                2      0.24111111      0.12055556       1.47    0.2693

Source                      DF      Type III SS     Mean Square    F Value    Pr > F
Primer                       2      4.58111111      2.29055556      27.86    <.0001
Method                       1      4.90888889      4.90888889      59.70    <.0001
Primer*Method                2      0.24111111      0.12055556       1.47    0.2693
```

بیان خروجی:

- در خطوط اول خروجی اطلاعات متغیر های طبقه ای و تعداد مشاهدات ، ارائه شده است.
- در خط بعدی متغیر پاسخ یعنی Y معرفی شده است.
- در جدول تحلیل واریانس بعد ، مدل و خطا بعنوان منابع تغییرات معرفی شده اند. مقابل هر یک از آنها به ترتیب درجه آزادی ، مجموع مربعات ، میانگین مربعات داده شده است. مقدار آماره F که از تقسیم میانگین مربعات مدل به میانگین مربعات خطا بدست می آید برای آزمون فرضیه زیر بکار می رود:

$$\begin{cases} H_0: & \text{مدل معنادار نیست} \\ H_1: & \text{مدل معنادار است} \end{cases}$$

در سطح اطمینان ۹۵% چون مقدار $P\text{-}Value$ از ۰.۰۵ کمتر است لذا فرض صفر رد می شود ، بعبارتی مدل $y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$ معنادار است. بطور کلی مدل ، میزان معناداری از تغییرات را توجیه می کند.

- جداول بعدی تحلیل واریانس به ترتیب بر مبنای مجموعه مربع نوع I و III هستند. در این جدول عوامل نوع آستر و شیوه اجرا و اثر متقابل آنها بعنوان منابع تغییرات ارائه شده اند. مقابل هر یک از آنها به ترتیب

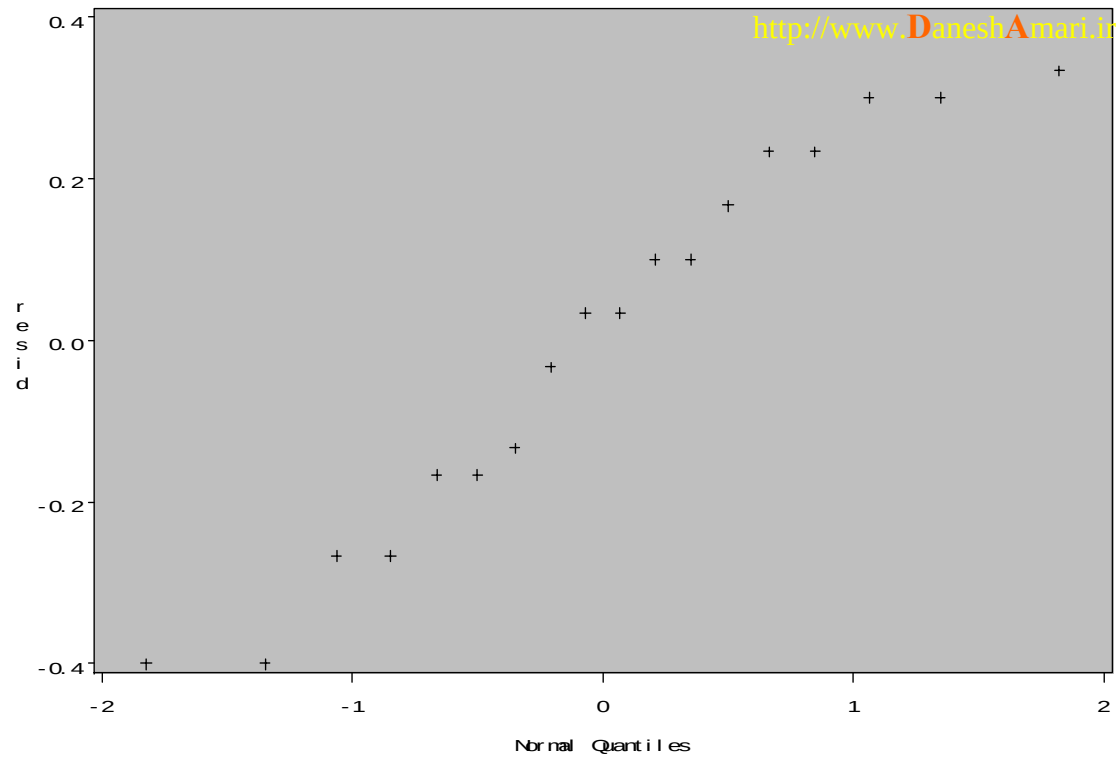
درجه آزادی ، مجموع مربعات ، میانگین مربعات داده شده است. مقدار آماره F که از تقسیم میانگین مربعات عامل به میانگین مربعات خطا بدست می آید برای آزمون فرضیه زیر بکار می رود:

$\begin{cases} H_0: & \text{اثرعامل معنادار نیست} \\ H_1: & \text{اثر عامل معنادار است} \end{cases}$

در سطح اطمینان ۹۵% چون مقدار $P\text{-}Value$ عوامل نوع آستر و شیوه اجرا، از ۰.۰۵ کمتر است لذا فرض صفر برای این عوامل رد می شود ، بعبارتی نوع آستر و شیوه اجرا بر میزان چسبندگی تاثیر معناداری دارند.
اما اثر متقابل این دو عامل در سطح اطمینان ۹۵% معنادار نمی باشد.

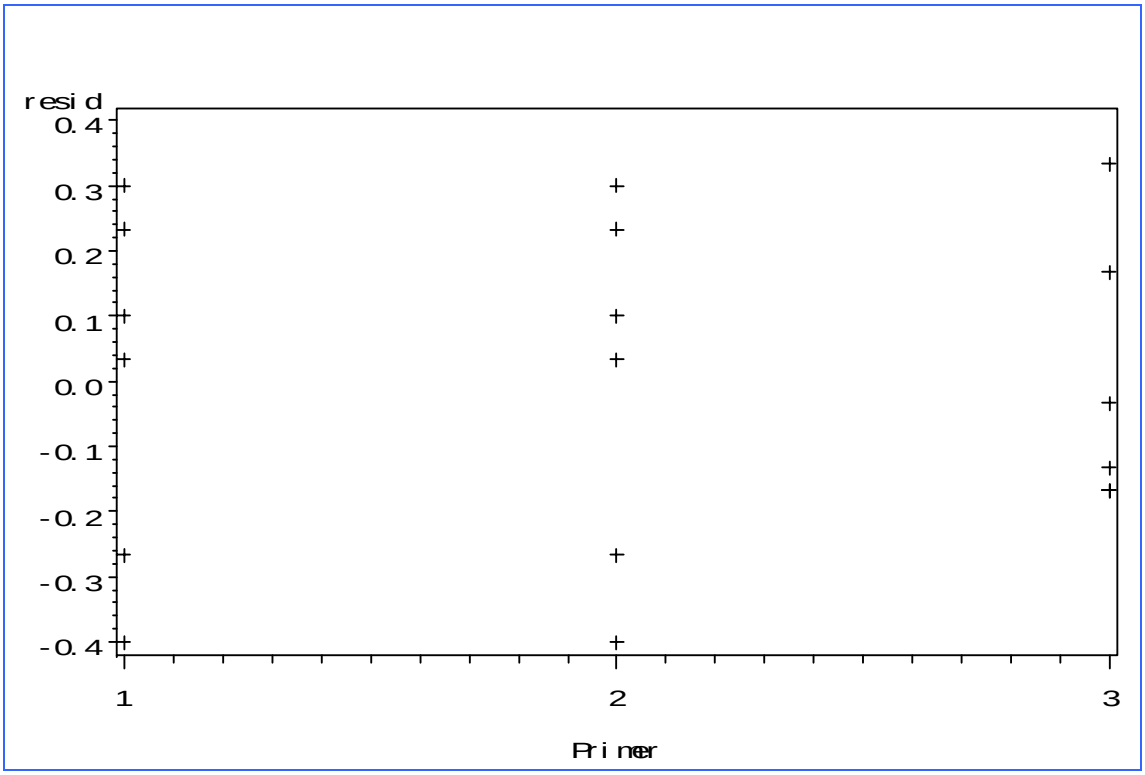
اکنون بایستی معیار های مناسبت مدل بررسی گردد. دستور زیر را برای رسم $QQ\ Plot$ باقیمانده ها وارد می کنیم.

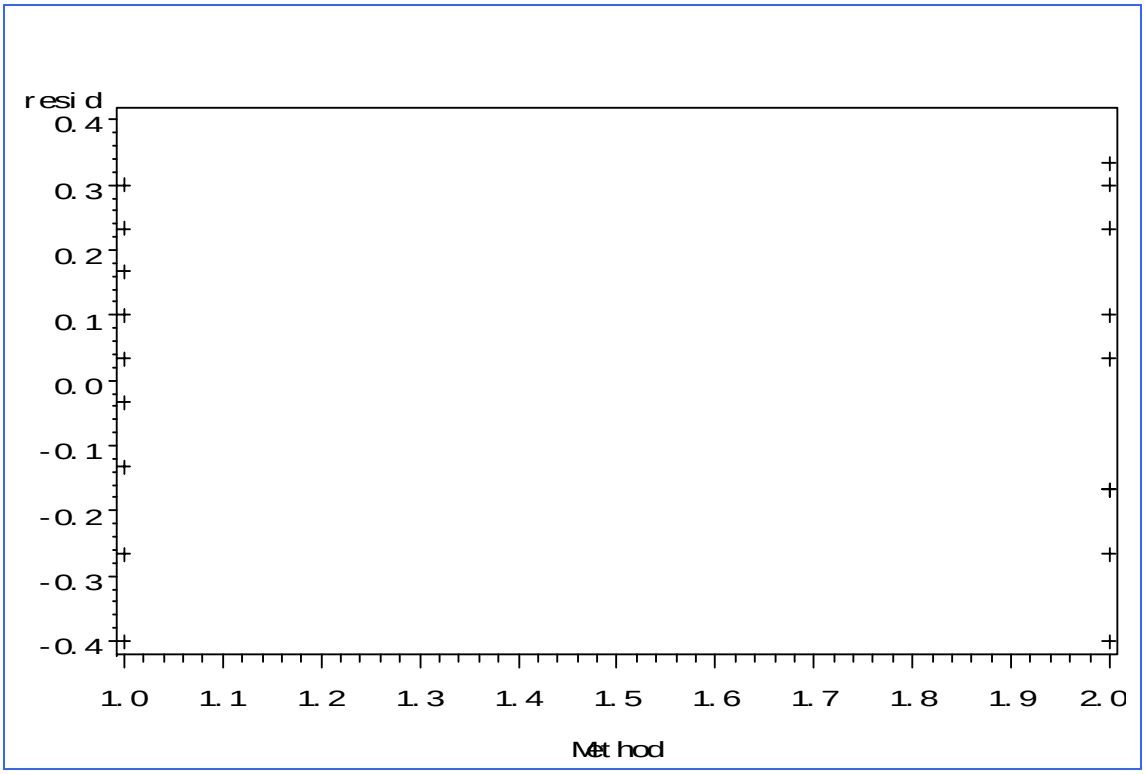
```
proc capability data=rcheck;
  qqplot resid/ cframe = ligr;
run;
```

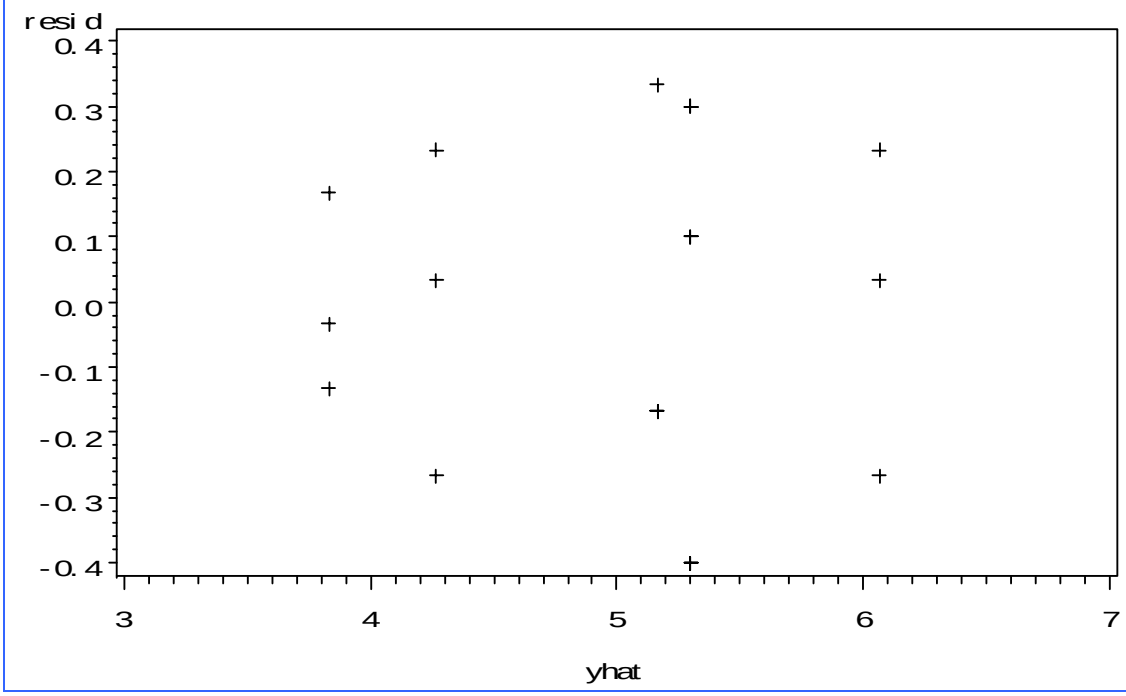


با توجه به این نمودار با کمی اغماض می توان گفت پیش فرض نرمال بودن خطاها برقرار است.

```
proc gplot data=rcheck;
   plot resid*Primer;
proc gplot data=rcheck;
   plot resid*Method;
proc gplot data=rcheck;
   plot resid*yhat;
run;
```


resid
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
1
2
3
Priner


resid
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
Method



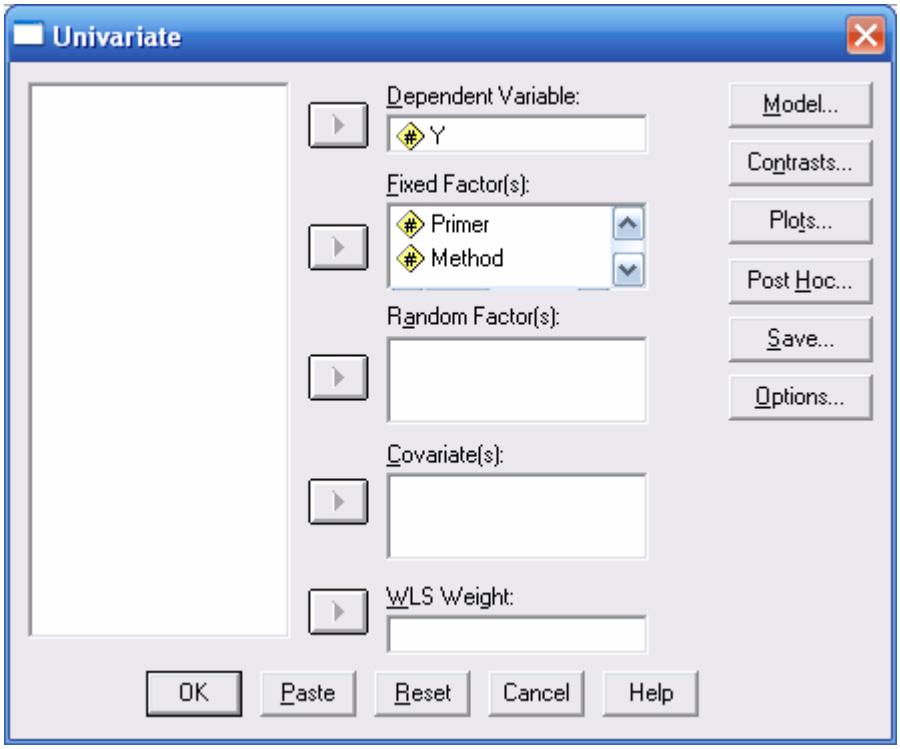
نمودار های فوق هیچ الگوی غیر معمول قابل تشخیصی را نشان نمی دهند. بنابراین نمی توان گفت از پیش فرض ها انحرافی صورت گرفته است.

می توان داده ها را با نرم افزار *SPSS* نیز تحلیل کرد:

: Primer | 1

	Primer	Method	Y	var
1	1	1	4.0	
2	1	1	4.5	
3	1	1	4.3	
4	1	2	5.4	
5	1	2	4.9	
6	1	2	5.6	
7	2	1	5.6	
8	2	1	4.9	
9	2	1	5.4	
10	2	2	5.8	
11	2	2	6.1	
12	2	2	6.3	
13	3	1	3.8	
14	3	1	3.7	
15	3	1	4.0	
16	3	2	5.5	
17	3	2	5.0	
18	3	2	5.0	
19				

از منوی *Analyze > General Linear Model > Univariate* استفاده می کنیم.



Univariate Analysis of Variance

Between-Subjects Factors

		N
Primer	1	6
	2	6
	3	6
Method	1	9
	2	9

Tests of Between-Subjects Effects

Dependent Variable: Y

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	9.731(a)	5	1.946	23.670	.000
Intercept	448.002	1	448.002	5448.676	.000
Primer	4.581	2	2.291	27.858	.000
Method	4.909	1	4.909	59.703	.000
Primer * Method	.241	2	.121	1.466	.269
Error	.987	12	.082		
Total	458.720	18			
Corrected Total	10.718	17			

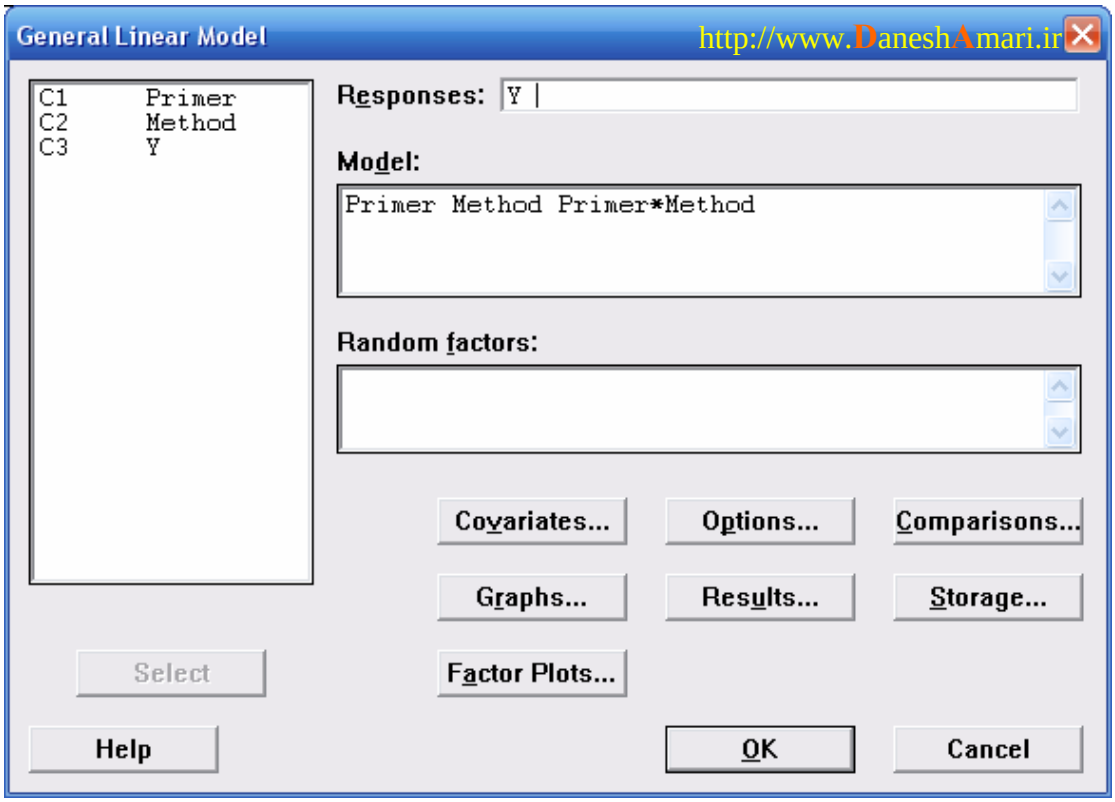
a R Squared = .908 (Adjusted R Squared = .870)

همچنین می توان داده ها را با *Minitab* تحلیل کرد.

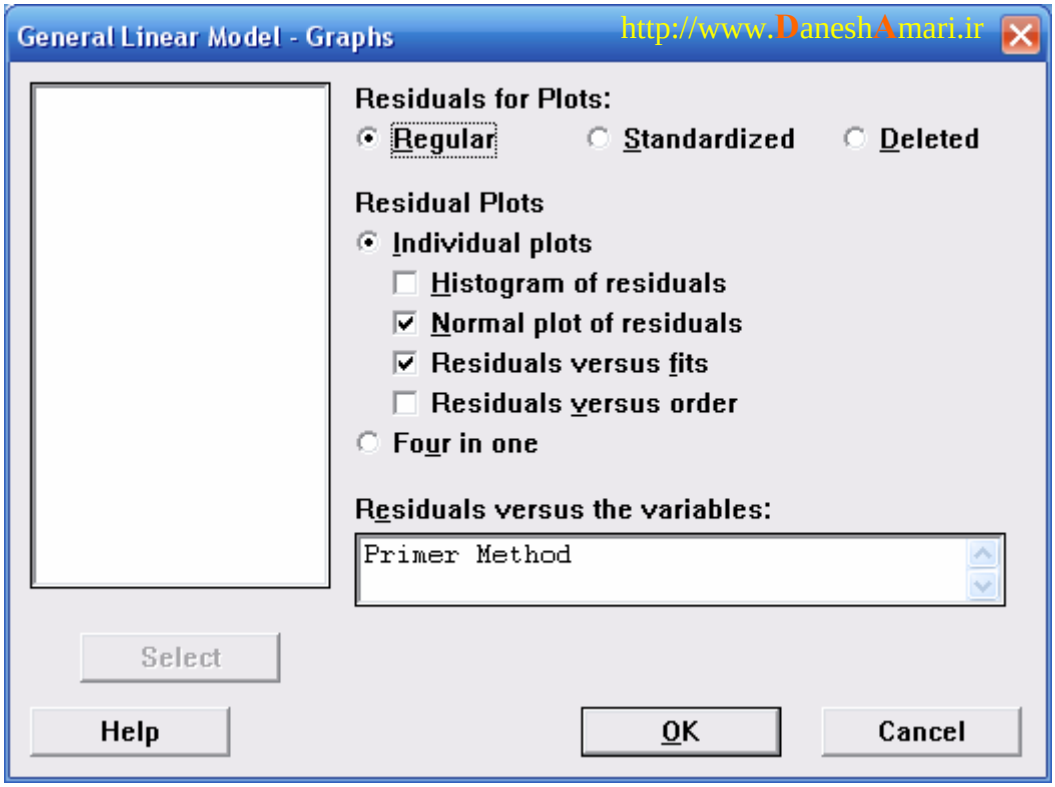
Worksheet 1

	C1	C2	C3
	Primer	Method	Y
1	1	1	4.0
2	1	1	4.5
3	1	1	4.3
4	1	2	5.4
5	1	2	4.9
6	1	2	5.6
7	2	1	5.6
8	2	1	4.9
9	2	1	5.4
10	2	2	5.8
11	2	2	6.1
12	2	2	6.3
13	3	1	3.8
14	3	1	3.7
15	3	1	4.0
16	3	2	5.5
17	3	2	5.0
18	3	2	5.0

برای این منظور از منوی *Stat > ANOVA > General Linear Model* استفاده می کنیم.



دکمه *Graphs* را کلیک کنید و کادر ظاهر شده را به شکل زیر تنظیم کنید:




```
MTB > GLM 'Y' = Primer Method Primer*Method;
SUBC>   Brief 2 ;
SUBC>   GNormalplot;
SUBC>   GFits;
SUBC>   NoDGraphs;
SUBC>   GVars 'Primer' 'Method';
SUBC>   RType 1 .
```

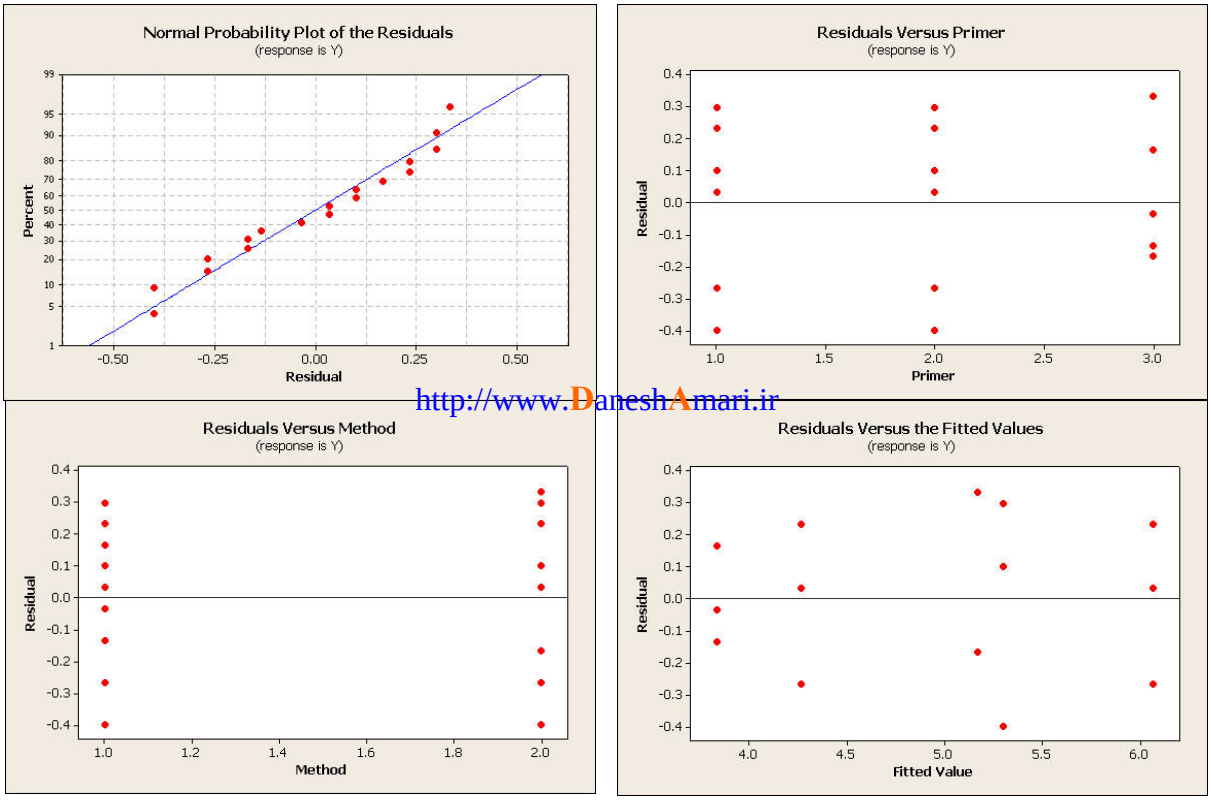
General Linear Model: Y versus Primer, Method

```
Factor  Type   Levels  Values
Primer  fixed       3  1, 2, 3
Method  fixed       2  1, 2


Analysis of Variance for Y, using Adjusted SS for Tests

Source          DF   Seq SS  Adj SS  Adj MS      F      P
Primer           2   4.5811  4.5811  2.2906  27.86  0.000
Method           1   4.9089  4.9089  4.9089  59.70  0.000
Primer*Method    2   0.2411  0.2411  0.1206   1.47  0.269
Error           12   0.9867  0.9867  0.0822
Total           17  10.7178


S = 0.286744   R-Sq = 90.79%   R-Sq(adj) = 86.96%
```



خروجی *SPSS* و *Minitab* را می توان با خروجی *SAS* انطباق داد و نتیجه گیری لازم را بدست آورد.

منابع :

[۱] *SAS Help and Documentation*

[۲] سایت آموزش نرم افزار های آماری *http://statistics.mihanblog.com/Post-15.ASPX*

[3] Montgomery, Douglas C.,George C.Runger., Applied statistics and probability for engineers , 3rd ed.2003 ,John Wiley