به نام خدا

سمینار درس مدل های خطی

stepwise method

ارائه دهندگان :

مینا عزیزی

سمیرا شاه نوشی

دی ماه ۹۷

* مسئله انتخاب متغیر
* اهمیت انتخاب متغیرهای مناسب
* روش های انتخاب متغیر
* (انتخاب پیشرو)Forward selection
* (حذف پسرو)Backward elimination
* (رگرسیون قدم به قدم)Stepwise method
* معایب روش قدم به قدم
* منابع ومراجع

* مسئله انتخاب متغیر(variable selection)

* پیدا کردن زیرمجموعه ای از متغیرهای رگرسیونی مناسب برای مدل، مسئله انتخاب متغیر نامیده می شود.

A **variable selection** method is a way of selecting a particular set of independent variable for use in a regression model.

*اهمیت انتخاب متغیرهای مناسب

انتخاب متغیرهایی با بیشترین تاثیر برروی متغیر پاسخ

*مدل واقعی معمولا بسیار پیچیده،مجهول و ناشناخته است.

*همواره مدل های با تعداد متغیرکم دلخواه تحلیل گران آماری بوده است.زیرا:

*۱- روابط بین متغیر های علمی را به صورت ساده بیان کرده و قابلیت تفسیر ساده تری دارند.

*۲- هزینه های محاسباتی و صرف زمان، اهمیت انتخاب متغیرهای مناسب را بیشتر می کند.

* ۳- چنانچه متغیرهای توضیحی نامناسب وارد مدل شوند باعث افزایش MSE می شود ودرنتیجه:

* کاهش دقت برآورد ضرایب رگرسیونی ودقت پیش بینی

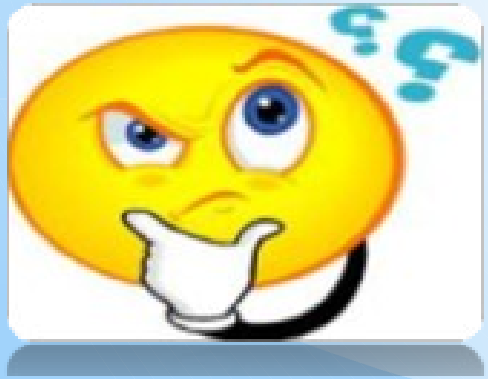
* افزایش طول فواصل اطمینان وفواصل پیش بینی

* کاهش دقت آزمون ها (ممکن است متغیر توضیحی مناسبی به اشتباه حذف شود.)

$Y_{(n \times 1)}$
متغیر پاسخ

$X_1, \dots, X_k$
متغیرهای توضیحی

$$Y = X\beta + \varepsilon$$



فرض می کنیم k متغیر توضیحی و متغیر پاسخ را در اختیار داریم ومی خواهیم مدل مناسبی برازش دهیم.

یکی از راه های انتخاب مدل این است که تمام 2^p مدل ممکن را برازش دهیم (همه رگرسیون های ممکن) سپس با استفاده از معیارهای مناسبت مدل بهترین آن ها راانتخاب کنیم. اما این کار در حالتی که p بزرگ باشد بسیار مشکل است .پس به دنبال روشی برای یافتن مدل مناسب هستیم.

ازجمله افرادی که در این زمینه تلاش کرده اند می توان کاکس وسنل(۱۹۷۴) ،هاکینگ(۱۹۷۲)،میرز(۱۹۹۰) ، هاکینگ ولاموت (۱۹۷۳) و تومپسون(۱۹۷۸) را نام برد.

*روشهای انتخاب متغیر

Tests-based

Penalty-based

Screening-based

*

*Tests-based

*

1) Stepwise(forward & backward)

2)Autometrics

*Forward selection

* این روش با فرض اینکه هیچ متغیر رگرسیونی غیر از عرض ازمبدا در مدل نیست ،آغاز می شود.سپس با هر یک از kمتغیرتوضیحی وعرض ازمبدا به صورت جداگانه مدل رگرسیونی برازش می دهیم و برای هرکدام آزمون صفربودن ضریب رگرسیونی مربوط به آن را انجام می دهیم. درواقع داریم:

* $Y_i=\beta_0+\beta_1 x_{i1}+\epsilon_i$
* $Y_i=\beta_0+\beta_2 x_{i2}+\epsilon_i$
* .
* .
* .
* $Y_i=\beta_0+\beta_k x_{ik}+\epsilon_i$

گام اول

* $Y_i=\beta_0+\beta_1 x_{i1}+\epsilon_i$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-2)}$$

* $Y_i=\beta_0+\beta_2 x_{i2}+\epsilon_i$

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-2)}$$

*

*

*

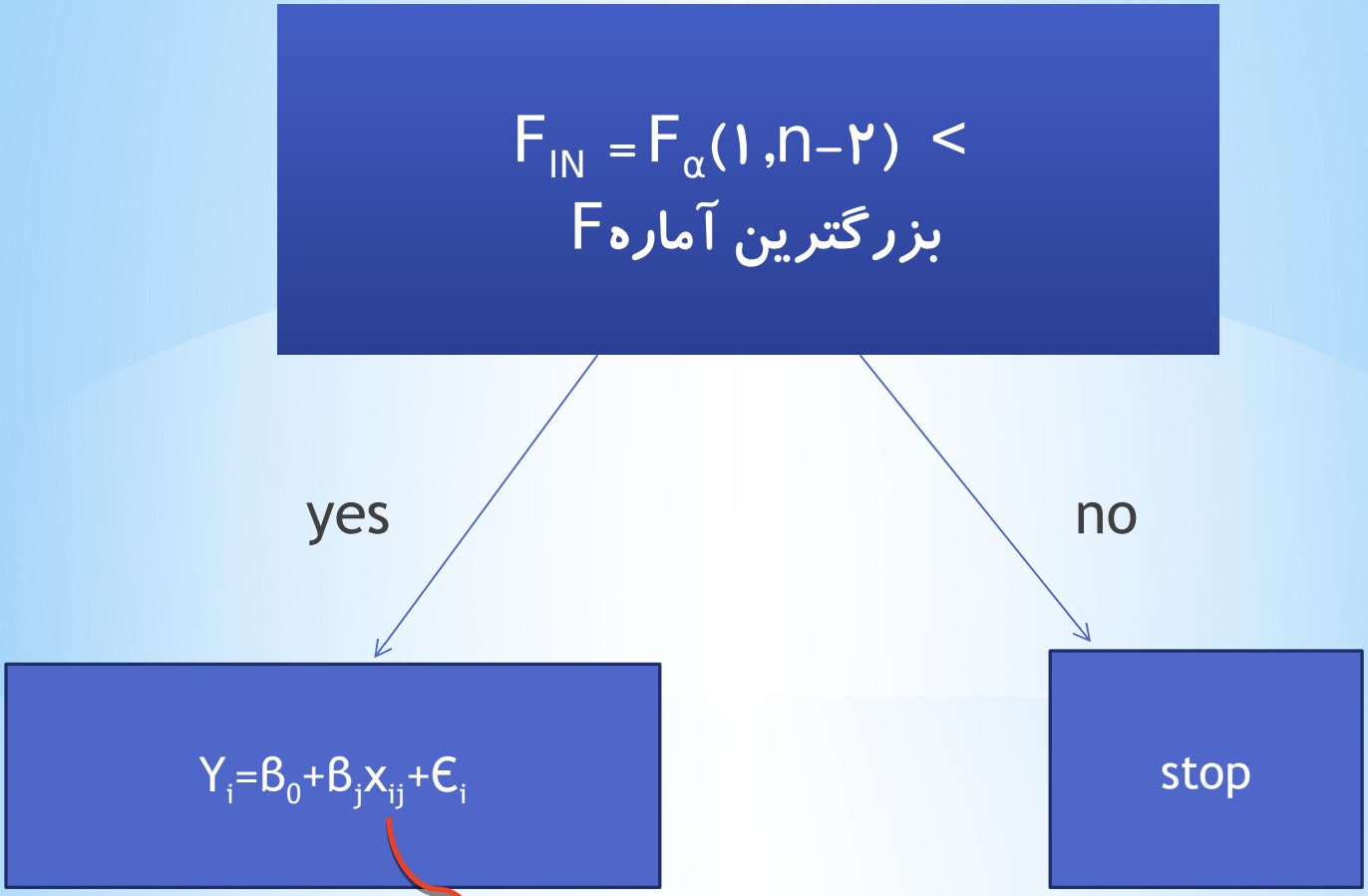
* $Y_i=\beta_0+\beta_k x_{ik}+\epsilon_i$

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-2)}$$

*



* بافرض اینکه بزرگترین آماره F مربوط به آزمون صفربودن ضریب jامین متغیر توضیحی بوده است.

* در روش انتخاب پیشرو $\alpha = 0.25$ توصیه می شود.

گام دوم

$Y_i=\beta_0+\beta_j x_{ij}+\beta_1 x_{i1}+\epsilon i$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-3)}$$

$Y_i=\beta_0+\beta_j x_{ij}+\beta_2 x_{i2}+\epsilon i$

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-3)}$$

.

.

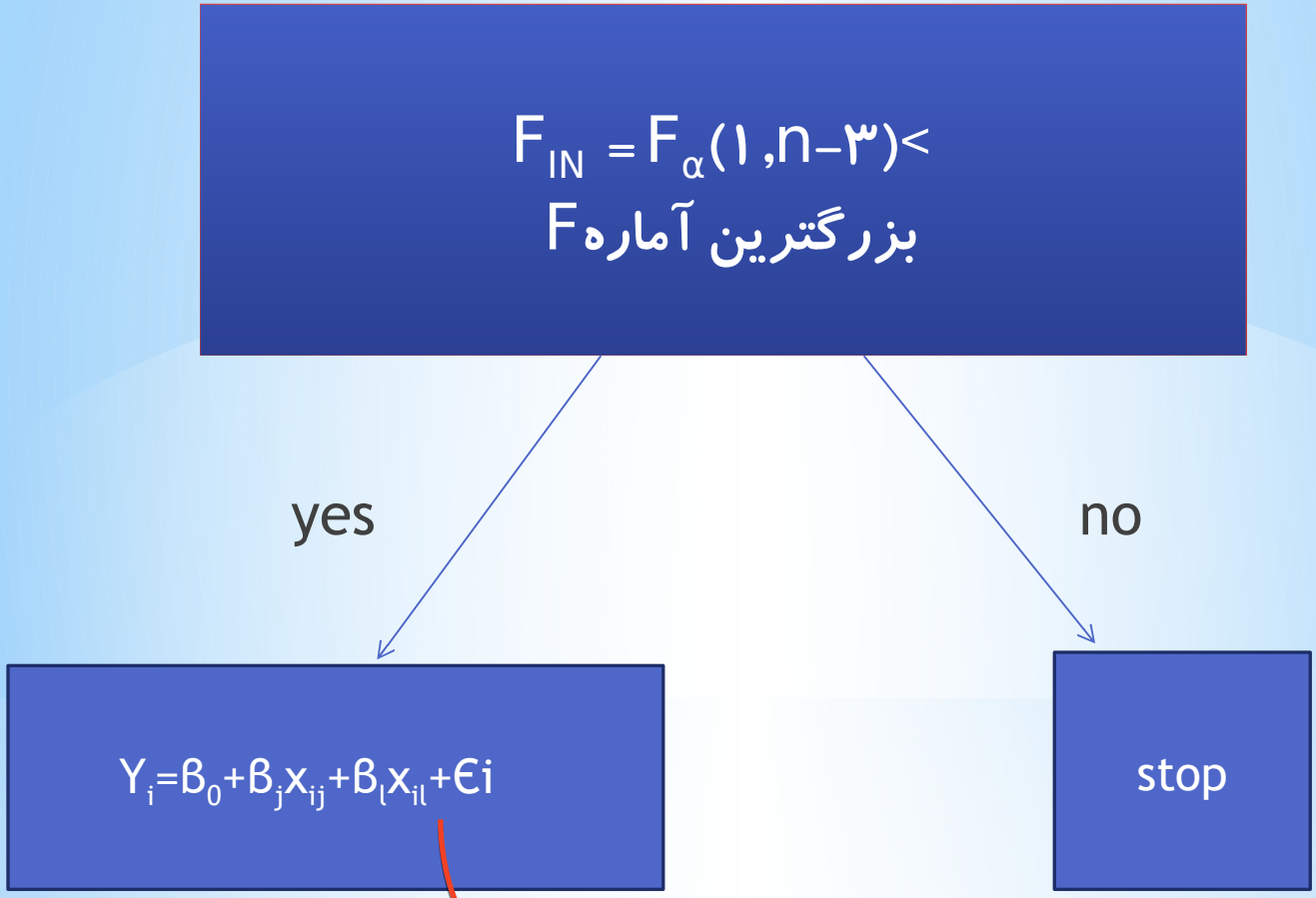
.

$Y_i=\beta_0+\beta_j x_{ij}+\beta_{k-1} x_{i(k-1)}+\epsilon i$

$H_0 : \beta_{k-1} = 0$

$H_1 : \beta_{k-1} \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-3)}$$



*بافرض اینکه بزرگترین آماره Fمربوط به آزمون صفربودن ضریب iامین متغیر توضیحی بوده است.

در این روش در هر گام متغیری که آماره آزمون مربوط به صفر بودن ضریب آن از بقیه **بزرگتر** است (دارای بیشترین همبستگی ساده با متغیر پاسخ) برای **ورود** به مدل کاندید می شود و این روند تا زمانی ادامه می یابد که بزرگترین آماره Fبه دست آمده از F_{IN} تجاوز نکند و یا اینکه همه متغیرهای توضیحی به مدل وارد شده باشند و مدل به دست آمده در گام قبل به عنوان مدل نهایی گزارش می شود.

*

مزایای انتخاب پیشرو

۱- این روش برای مسائلی که در ابتدانیازبه جمع آوری اطلاعات دارند مناسب است.

۲- انجام این روش ساده و قابل فهم است.

معایب انتخاب پیشرو

متغیری که در این روش وارد مدل شده، قابل حذف نیست.

*Backward elimination

* در این روش ابتدا مدلی شامل کلیه k متغیر رگرسیونی نامزد برازش داده می شود، سپس آزمون صفر بودن همه ضرایب رگرسیونی را به صورت مجزا انجام می دهیم.

$$*Y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}+\epsilon i$$

گام اول

$Y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_k x_{ik}+\epsilon i$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

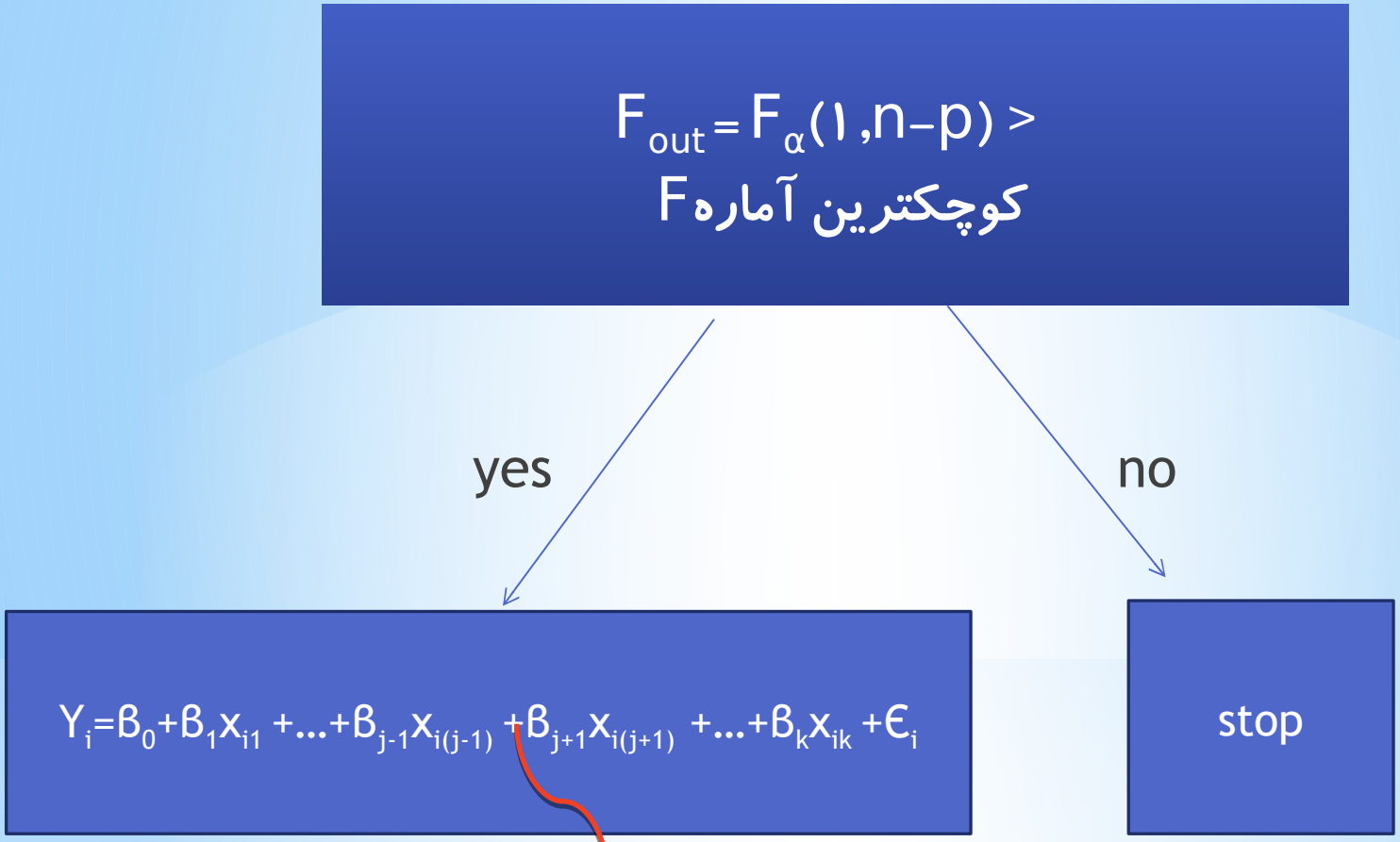
$$F=\frac{(RSS_r-RSS_f)/1}{RSS_f/(n-p)}$$

.
.
.

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$

$$F=\frac{(RSS_r-RSS_f)/1}{RSS_f/(n-p)}$$



* بافرض اینکه کوچکترین آماره Fمربوط به آزمون صفربودن ضریب jامین متغیر توضیحی بوده است.

* در روش حذف پسرو $\alpha=0.1$ توصیه می شود.

گام دوم

$$Y_i=\beta_0+\beta_1 x_{i1}+\ldots+\beta_{j-1}x_{i(j-1)}+\beta_{j+1}x_{i(j+1)}+\ldots+\beta_k x_{ik}+\epsilon_i$$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$F=\frac{(RSS_r-RSS_f)/1}{RSS_f/(n-(p-1))}$$

.

.

.

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$

$$F=\frac{(RSS_r-RSS_f)/1}{RSS_f/(n-(p-1))}$$

$F_{out} = F_{\alpha}(1, n-(p-1)) >$
کوچکترین آماره F

yes

no

$Y_i=\beta_0+\beta_1 x_{i1} +\ldots+\beta_{j-1} x_{i(j-1)} +\beta_{j+1} x_{i(j+1)} +\ldots+\beta_{l-1} x_{i(l-1)}+\beta_{l+1} x_{i(l+1)} +\ldots+\beta_k x_{ik}+\epsilon_i$

stop

*بافرض اینکه کوچکترین آماره F مربوط به آزمون صفربودن ضریب امین متغیر توضیحی بوده است.

* در این روش در هر گام متغیری که آماره آزمون مربوط به صفر بودن ضریب آن از بقیه کوچکتر است برای حذف از مدل کاندید می شود و این روند تا زمانی ادامه می یابد که کوچکترین آماره Fبه دست آمده از F_{out} کوچک تر نباشد و یا اینکه همه متغیرهای توضیحی از مدل حذف شده باشند و مدل به دست آمده در گام قبل به عنوان مدل نهایی گزارش می شود.

مزایای حذف پسرو

۱- این روش برای مسائلی که در آغاز فرضی بنا شده وسپس در جهت اثبات آن باشیم مناسب است.

۲- انجام این روش ساده و قابل فهم است.

معایب حذف پسرو

متغیری که در این روش ازمدل حذف شده، قابل برگشت نیست.

*Stepwise method

* به دلیل وجود معایب ذکر شده در روش انتخاب پیشرو وحذف پسرو ، روش جدیدی به نام **stepwise** ارائه می شود که در واقع ترکیبی از دو روش مذکور می باشد.
* این روش توسط **Efroymson** در سال ۱۹۶۰ ارائه شد.
* روش قدم به قدم با انتخاب پیشرو شروع شده و با حذف پسرو ادامه می یابد ودر واقع رگرسیون قدم به قدم تعدیل انتخاب پیشرو می باشد.

* در رگرسیون قدم به قدم در هرمرحله بعد از وارد شدن متغیر جدید به مدل بررسی می کنیم که آیا **در حضور بقیه متغیرها** به این متغیر احتیاجی هست یا خیر.چرا که ممکن است متغیر رگرسیونی اضافه شده در مرحله قبل به لحاظ ارتباط با متغیرهای رگرسیونی که اکنون در معادله اند، زائد باشد.

* درواقع در این روش با مدل شامل عرض از مبدا شروع کرده و روشی همانند الگوریتم ارائه شده در صفحه بعد را دنبال می کنیم تا به مدل نهایی دست یابیم.

گام اول

$*Y_i=\beta_0+\beta_1 x_{i1}+\epsilon_i$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$H_0 : \beta_2 = 0$

$*Y_i=\beta_0+\beta_2 x_{i2}+\epsilon_i$

$H_1 : \beta_2 \neq 0$

*

*

*

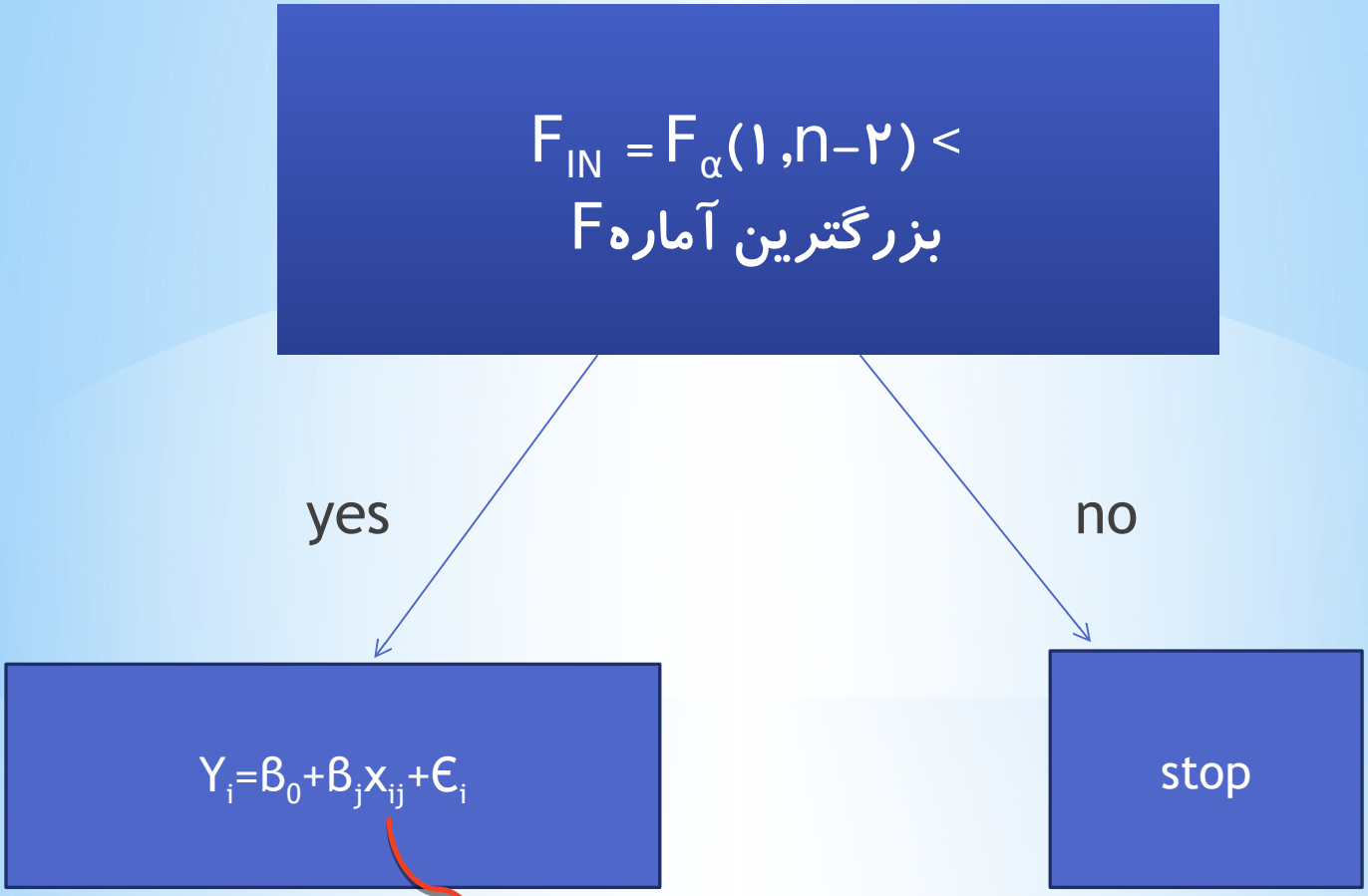
*

$H_0 : \beta_k = 0$

$*Y_i=\beta_0+\beta_k x_{ik}+\epsilon_i$

*

$H_1 : \beta_k \neq 0$



* بافرض اینکه بزرگترین آماره F مربوط به آزمون صفربودن ضریب jامین متغیر توضیحی بوده است.

گام دوم

$Y_i=\beta_0+\beta_j x_{ij}+\beta_1 x_{i1}+\epsilon i$

$H_0 : \beta_1 = 0$

$H_1 : \beta_1 \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-3)}$$

$Y_i=\beta_0+\beta_j x_{ij}+\beta_2 x_{i2}+\epsilon i$

$H_0 : \beta_2 = 0$

$H_1 : \beta_2 \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-3)}$$

.

.

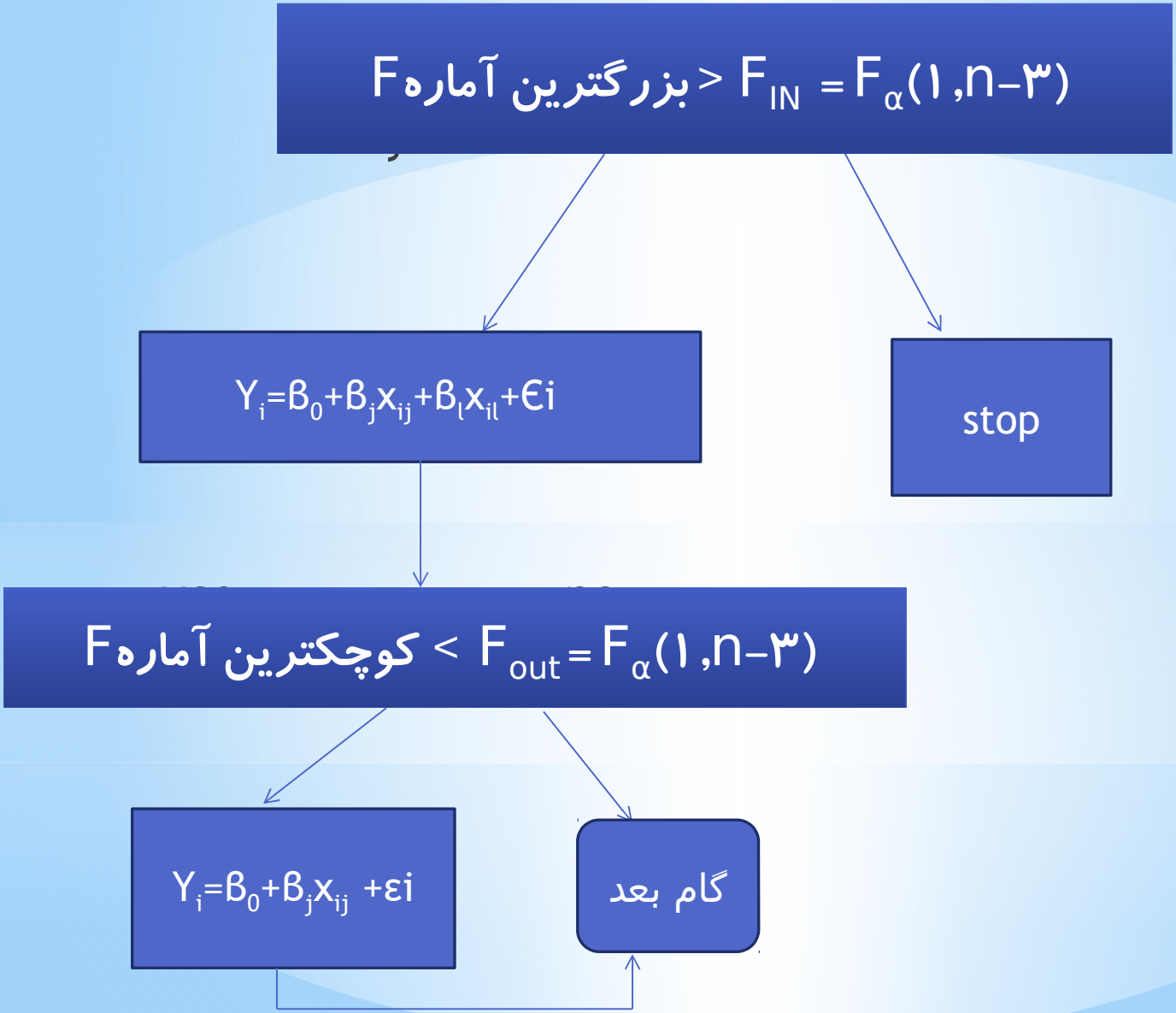
.

$Y_i=\beta_0+\beta_j x_{ij}+\beta_{k-1} x_{i(k-1)}+\epsilon i$

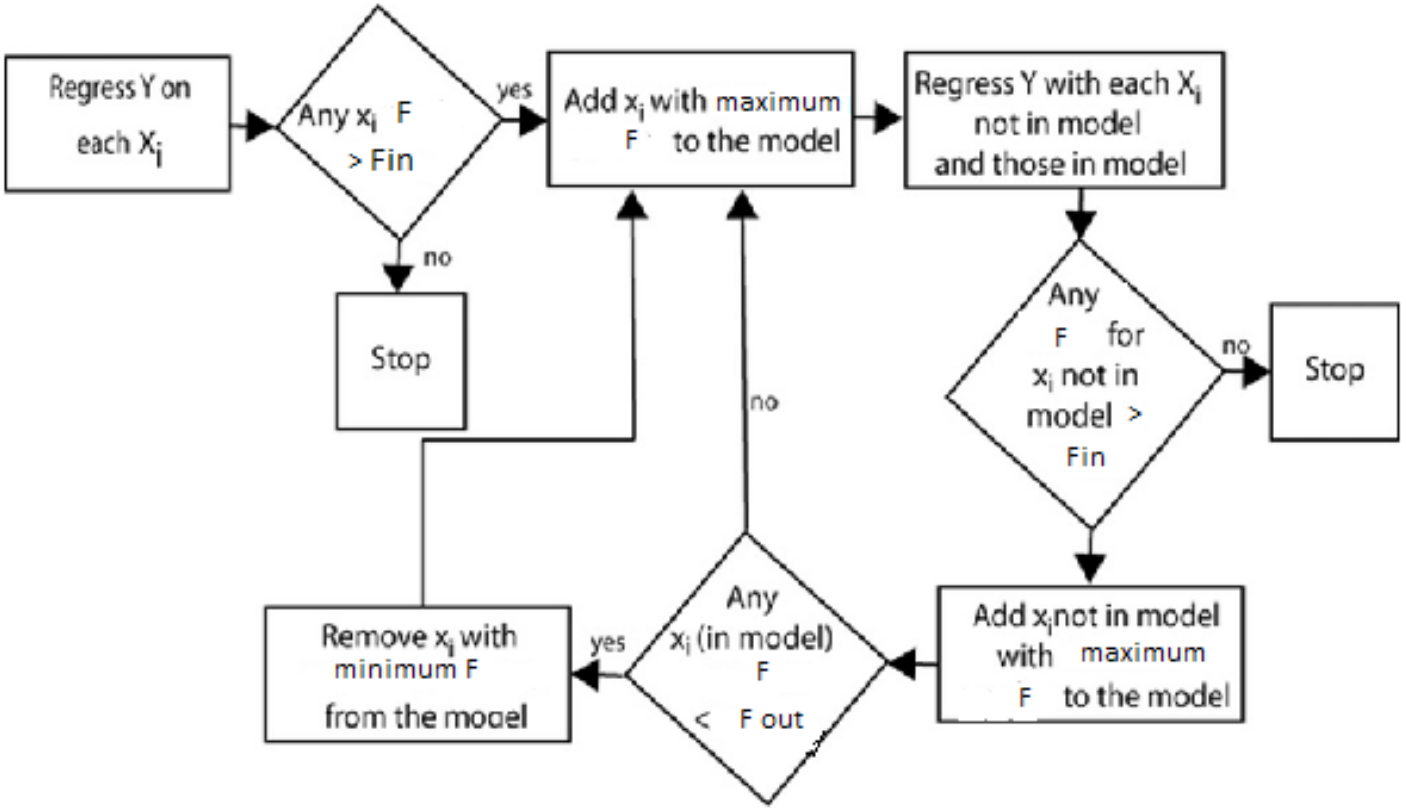
$H_0 : \beta_{k-1} = 0$

$H_1 : \beta_{k-1} \neq 0$

$$F = \frac{(RSS_r - RSS_f)/1}{RSS_f/(n-3)}$$


F بزرگترین آماره > $F_{IN} = F_{\alpha}(1, n-3)$
$Y_i=\beta_0+\beta_j x_{ij}+\beta_l x_{il}+\epsilon i$
stop
F کوچکترین آماره < $F_{out} = F_{\alpha}(1, n-3)$
$Y_i=\beta_0+\beta_j x_{ij}+\varepsilon i$
گام بعد

*الگوریتم روش قدم به قدم :

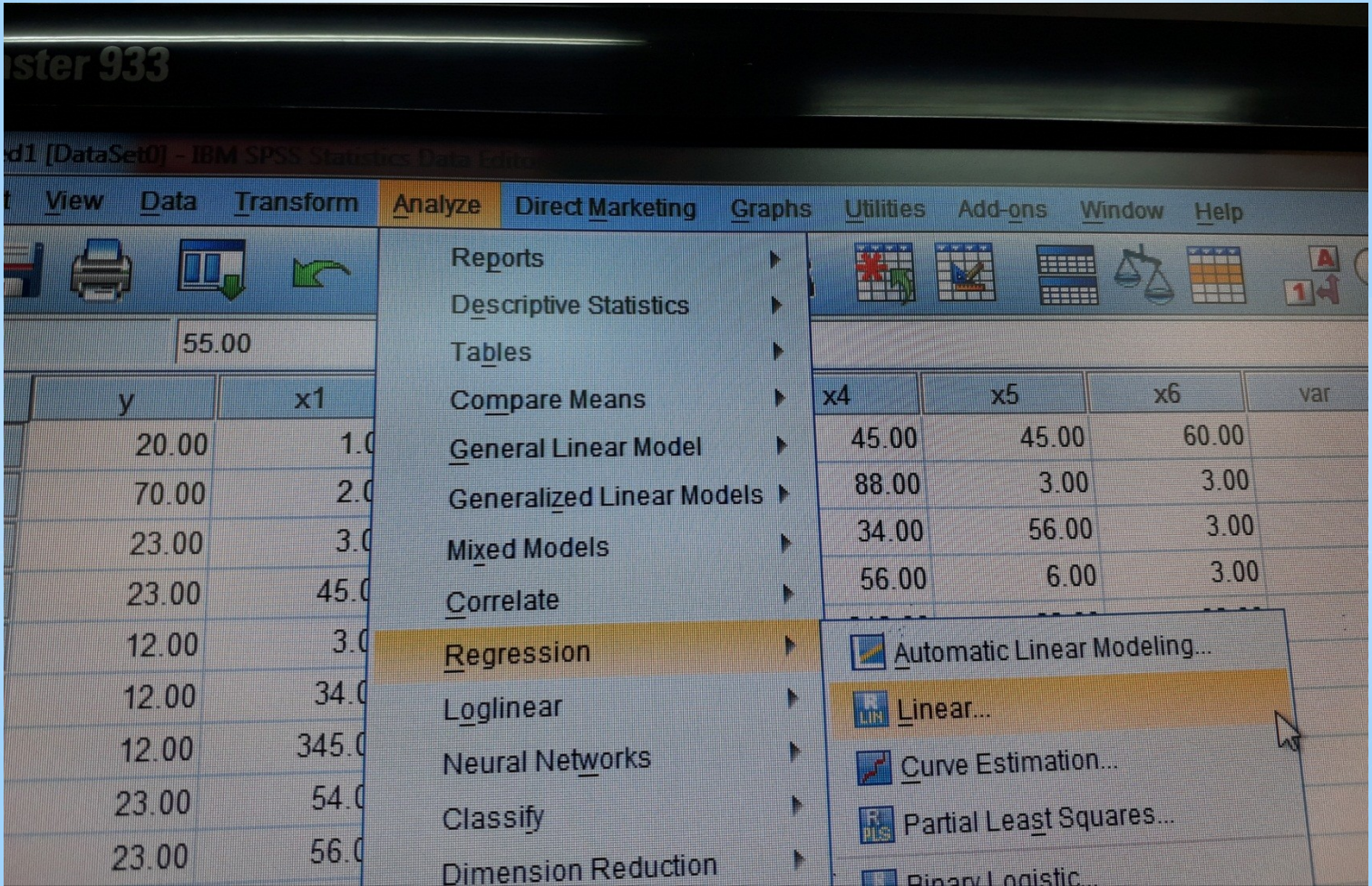


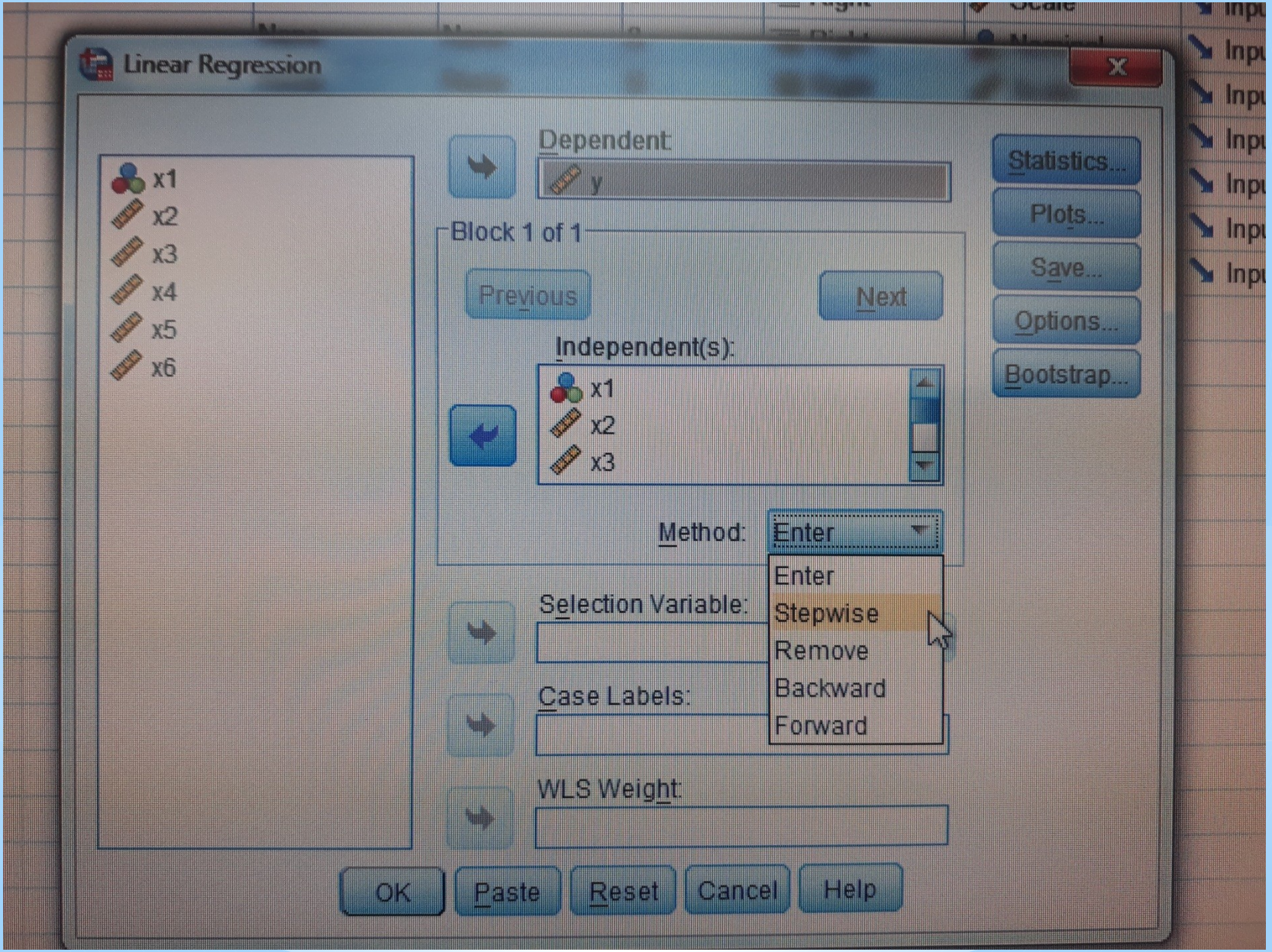
* انتخاب مقادیر برای F_{IN} و F_{out} در حد زیادی یک موضوع ترجیحی تحلیلگراست . می توان مقادیر F_{IN} و F_{out} را طوری انتخاب کرد که همه متغیرهای رگرسیونی به وسیله انتخاب پیشرو وارد ویا به وسیله حذف پسرو کنار گذاشته شوند.
* بعضی تحلیلگران ترجیح می دهند در رگرسیون قدم به قدم
* ۴ = $F_{IN} = F_{out}$ انتخاب کنند
* به طور معمول $F_{IN} < F_{out}$ انتخاب می کنیم ،چرا که به طور نسبی اضافه کردن یک متغیر رگرسیونی نسبت به حذف آن مدل را مشکل تر می کند.

* مثال:

* داده های مربوط به یک نمونه ۵۰ تایی از متغیر پاسخ yو ۶ متغیر توضیحی x۱,...,x۶ را واردنرم افزار spss کرده و به صورت زیر به روش step wise مدل مناسبی می یابیم.

* p=۷ k=۶ n=۵۰




Linear Regression
x1
x2
x3
x4
x5
x6
Dependent:
y
Block 1 of 1
Previous
Next
Independent(s):
x1
x2
x3
Method:
Enter
Enter
Stepwise
Remove
Backward
Forward
Selection Variable:
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Bootstrap...
OK
Paste
Reset
Cancel
Help

Variables Entered/Removed[a]			
Model	Variables Entered	Variables Removed	Method
1	X1	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
a. Dependent Variable: Y			

* خروجی نرم افزار :

$$yi = \beta 0 + \beta 1 xi1 + \varepsilon i$$ *

Coefficients[a]						
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	479.145	40.527		11.823	.000
	X1	.388	.048	.757	8.014	.000
a. Dependent Variable: Y						

$$\tilde{y}_i = 479.145 + .388 xi1$$ *

*معایب رگرسیون قدم به قدم :

*اگرچه این روش بسیار سریع، ساده ،آسان و قابل فهم است اما لزوما بهترین مدل را ارائه نمی دهندو معایبی نیز دارد از جمله اینکه:

*در هر گام بعد از برازش مدل باید بررسی شود:

*۱- برقراری فرضیات مدل

*۲- عدم وجود داده پرت

*۳- عدم وجود مشکل هم خطی چندگانه

*و تکرار عدم بررسی این موارد در هرگام باعث کاهش اعتبار مدل می شود.

*هم چنین در هر گام آزمونی با میزانی از خطا انجام می گردد و در هر مرحله این خطاها تکرار می شود و این امر باعث زیاد شدن خطای مدل نهایی می گردد .

*در مواردی که تعداد متغیرهای توضیحی (p) زیاد باشد گام ها طولانی شده واین روش کارایی چندانی ندارد.

* Harrell(۲۰۰۱) به بررسی معایب این روش پرداخته و موارد زیر را عنوان کرده است:

* ۱- اریبی برآوردگرهای پارامترها بسیار بالاست.
* ۲- مساله همبستگی تشدید پیدا می کند.
* ۳- تخمین پارامترها خیلی دور از صفر خواهد بود.
* ۴- تخمین واریانس برآورد پارامترها دقیق نمی باشد.
* ۵- آزمون فرضها و فواصل اطمینان نیز اشتباه بدست می آیند.

* ۶- مقادیر R^2 بسیار اریب هستند.

* ۷- خطاهای استاندارد براورد پارامترها بسیار کوچک هستند.

* ۸- فواصل اطمینان حول برآوردهای پارامترها باریک هست.

*منابع ومراجع

*1- George A.F. Seber and Alan J. Lee _Linear Regression Analysis_second Edition

*2- Loann Desboulets _ A Review on Variable Selection in Regression Analysis

*3- L. Flom, National Development and Research Institutes, New York, NY David L. Cassell, Design Pathways, Corvallis, OR 4_ Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use_Peter Introduction to Linear Regression Analysis _Montgomery,Douglas C; Peck, Elizabeth

*5- Harrell, F. E. (2001), Regression modeling strategies: With applications to linear models, logistic regression, and survival

*analysis, Springer-Verlag, New York.


با تشکر از توجه شما