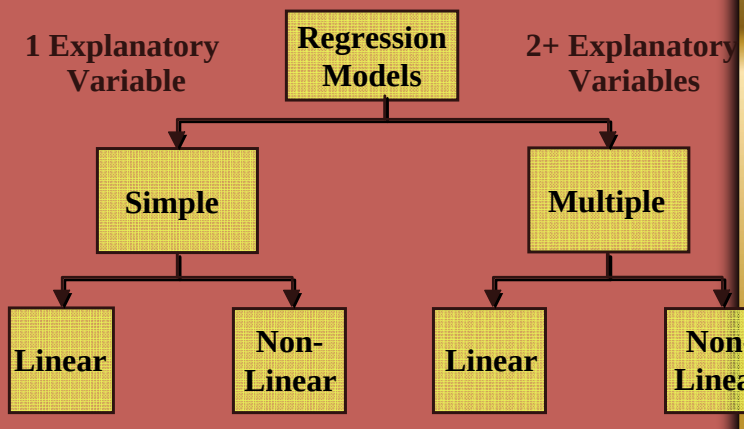
Multiple Regression

1

Types of Regression Models



رگرسیون چند گانه
Multiple Regression

- رگرسیون چند گانه، تعمیم رگرسیون ساده خطی است.
- در رگرسیون چند گانه، ارتباط چند متغیر مستقل با یک متغیر پاسخ، مدل بندی و تحلیل می شود.
- شکل کلی مدل رگرسیون چند گانه به صورت زیر است:

intercept — slopes — Error

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Dependent (response) variable

Independent (explanatory) variables

رگرسیون چند گانه
Multiple Regression

- برآورد ضرایب مدل و مقادیر p آن ها با استفاده از نرم افزار به دست می آید.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

۳

مثال: وزن و فشاخون سیستولیک ده مرد جوان

ردیف	وزن	کلسترول	فشارخون
۱	۵۱	۱۶۲	۱۰۸
۲	۵۳	۱۵۸	۱۱۱
۳	۵۶	۱۵۷	۱۱۵
۴	۵۶	۱۵۵	۱۱۶
۵	۵۸	۱۵۶	۱۱۷
۶	۶۰	۱۵۴	۱۲۰
۷	۵۸	۱۶۹	۱۲۴
۸	۶۱	۱۸۱	۱۲۷
۹	۵۹	۱۷۴	۱۲۲
۱۰	۵۶	۱۸۰	۱۲۱

رگرسیون خطی ساده

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	21.752	17.196		1.265	.241
	weight	1.696	.302	.893	5.610	.001

a. Dependent Variable: systolic bp

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	58.966	24.716		2.386	.044
	cholestrol	.359	.150	.647	2.397	.043

a. Dependent Variable: systolic bp

رگرسیون خطی با دو متغیر مستقل

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.547	10.939		-.141	.892
	cholestrol	.225	.052	.405	4.356	.003
	weight	1.454	.177	.765	8.222	.000

a. Dependent Variable: systolic bp

مثال داده های واقعی

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	117.126	2.964		39.522	.000
	FBS	.057	.029	.124	1.970	.050

a. Dependent Variable: SYSTOLIC BP

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	85.161	4.834		17.616	.000
	FBS	.030	.026	.065	1.145	.253
	AGE	.722	.091	.450	7.909	.000

a. Dependent Variable: SYSTOLIC BP

Multiple Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

β_0 = is the intercept (the value of y when all $x_i = 0$)

β_j = regression coefficient or slope for variable x_j: the **change in y per unit change in x_j**, holding all other x variables constant (all else equal…)

ε = normally distributed independent random error with mean=0, standard deviation=σ

k = number of independent variables; df =n–k–1

مثال: در مطالعه ای، ارتباط سن، شاخص توده بدنی، نژاد و جنس با کلسترول در قالب یک مدل رگرسیون چند گانه بررسی شد. نتایج به صورت زیر به دست آمده است:

Independent Variables	B	Std. Error	t	Sig.
Intercept	-8.3748	1.0338	-8.101	0.0000
Age	0.1603	0.0140	11.442	0.0000
BMI	1.3710	0.0372	36.809	0.0000
Race	-0.9161	0.4005	-2.287	0.0225
Sex	-10.2746	0.3638	-28.242	0.0000

توضیح: جنس: ۱ =مرد ۰=زن

نژاد: ۱=سفید ۰= سایر

۴

رگرسیون چند گانه
Multiple Regression

Independent Variables	B	Std. Error	t	Sig.
Intercept	-8.37	1.0338	-8.101	0.0000
Age	0.16	0.0140	11.442	0.0000
BMI	1.37	0.0372	36.809	0.0000
Race	-0.91	0.4005	-2.287	0.0225
Sex	-10.27	0.3638	-28.242	0.0000

- مدل رگرسیون عبارت است از:

$$Y = -8.37 + 0.16\,Age + 1.37\,BMI - 0.91\,Race - 10.27\,Sex$$

- مقادیر ضرایب رگرسیون نشان دهنده شدت و جهت همبستگی متغیر های وابسته با متغیر پاسخ است.

۵

Model Building

Model Building

- The key to regression analysis is to properly specify the model.
- We may want to include all the important variables and omit any extraneous ones (variables that don't significantly improve the explanatory power of the model).

Variable Selection Procedures

- *F* Test

To test whether the addition of x_2 to a model involving x_1 (or the deletion of x_2 from a model involving $x_1$and x_2) is statistically significant.

$$F = \frac{(\text{SSE(reduced)-SSE(full)})/\text{number of extra terms}}{\text{MSE(full)}}$$

$$F = \frac{(\text{SSE}(x_1)\text{-SSE}(x_1,x_2))/1}{(\text{SSE}(x_1,x_2))/(n-p-1)}$$

The *p*-value corresponding to the *F* statistic is the criterion used to determine if a variable should be added or deleted

Variable Selection Methods

Three approaches to decide which variable to include in the final model

Forward Selection

Backward Selection

Stepwise Selection

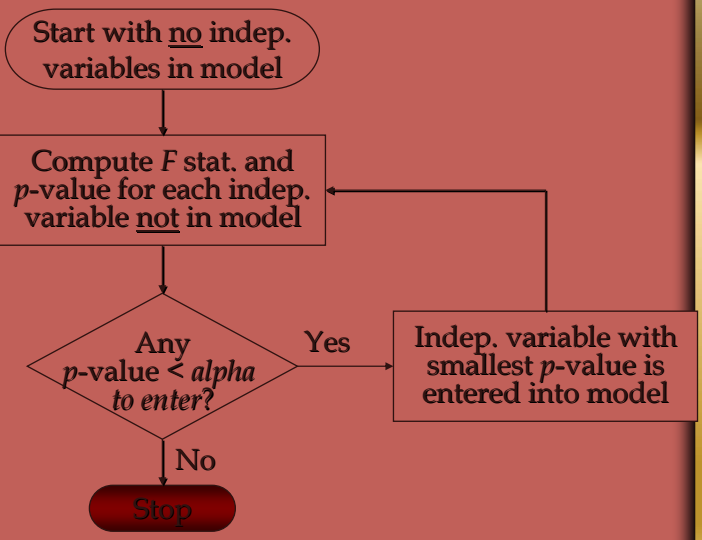
Forward Selection

- Starts with no independent variables.
- Adds the explanatory variable having the lowest p-value.
- Adds variables one at a time as long as a significant reduction in the error sum of squares (SSE) can be achieved.

  (continue until no variable can be added.)

(would have a p-value below the given threshold (e.g., 0.05)).

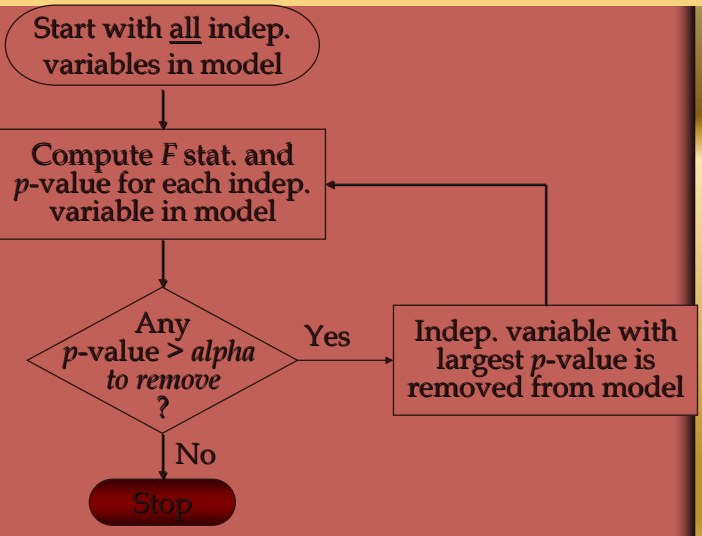
Forward Selection



Backward Elimination

- Begins with a model that includes all the independent variables that the researcher considered.
- Then attempts to delete one variable at a time by determining whether the least significant variable currently in the model can be removed because its *p*-value is less than the user-specified or default value.
- Once a variable has been removed from the model it cannot reenter at a subsequent step.

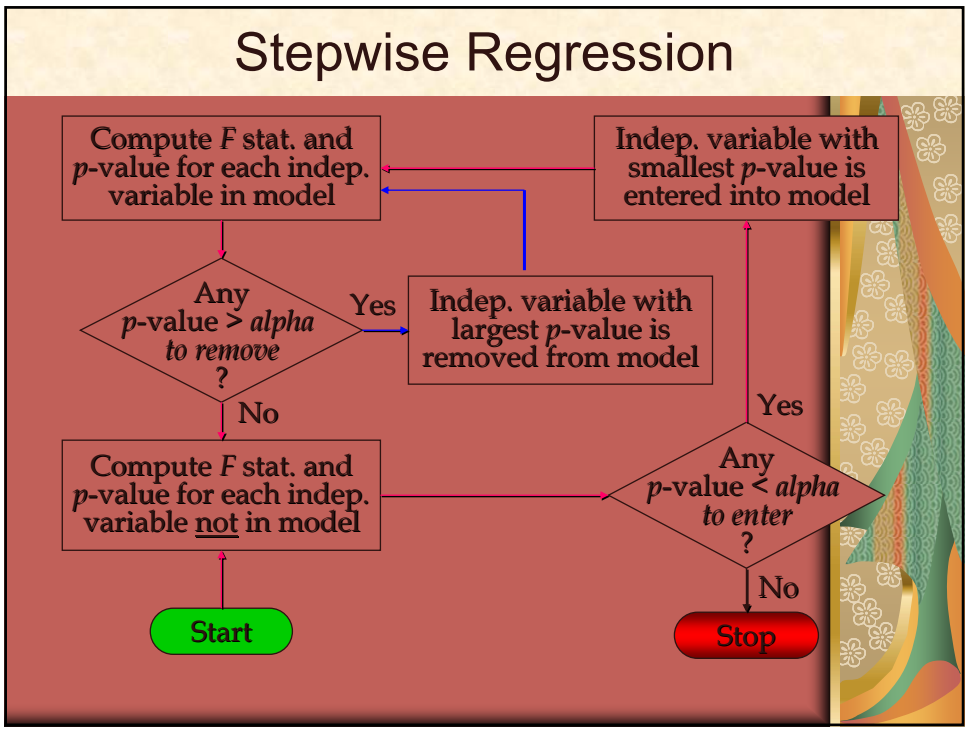
Backward Elimination



Include/Exclude Decisions

Stepwise:

- like the forward procedure, except deletions are considered along the way.


Stepwise Regression
Compute F stat. and p-value for each indep. variable in model
Indep. variable with smallest p-value is entered into model
Any p-value > alpha to remove ?
Yes
Indep. variable with largest p-value is removed from model
No
Yes
Compute F stat. and p-value for each indep. variable not in model
Any p-value < alpha to enter ?
No
Start
Stop


رگرسیون لجستیک
Logistic regression
شرایط
کاربردها

رگرسیون لجستیک
Logistic Regression

- اگر متغیر پاسخ از نوع کیفی دو حالته (Binary) باشد، مدل حاصل، مدل رگرسیون لجستیک نامیده می شود.
- اگر دو حالت متغیر پاسخ را با صفر و یک نشان دهیم و P_X احتمال یک بودن متغیر پاسخ باشد، مدل رگرسیون لجستیک برای k متغی مستقل به صورت زیر خواهد بود:

$$P_X = \frac{1}{1+\exp\left[-\left(b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k\right)\right]}$$

- معمولا اگر $P_X \geq 0.50$ دراین صورت پاسخ یک برای متغیر وابسته برآورد می شود.
- برآورد ضرایب مدل و مقادیر p آزمون به کمک نرم افزار انجام می شود.