مثال رگرسيون مرحله‌اي:

در مراحل توسعه مدل رگرسیونی ارتباط میزان فشار خون با وضعیت جسمانی و نحوه زندگی افراد، اطلاعات 11 نفر شامل سن، وزن و میزان خواب شبانه آنها جمع‌آوری و حالات مختلف رگرسیون با فرض این که میزان فشار خون متغیر وابسته باشد به صورت جداول صفحه بعد ارایه شده است. اگر ضرایب خطای میزان پذیرش برای ورود و خروج متغیرهای مستقل (αE و αR) 0/1 در نظر گرفته شوند، با استفاده از روش رگرسیون مرحله‌ای مدل نهایی را ارایه نمایید.

مرحله اول توسعه مدلهای رگرسیون یک متغیره (ساده) به شرح زیر می‌باشد:

Variable	Coefficient	t Stat	P-value
Intercept	58.7055	9.0983	7.81E-06
Age	1.4632	**14.2995**	1.71E-07
Intercept	1.1416	0.0923	0.928459
Weight	0.7638	12.0896	7.23E-07
Intercept	180.0275	8.3488	1.57E-05
Sleep	-4.5110	-1.4118	**0.19163**

از نتایج مشخص است که متغیر میزان خواب نمی‌تواند وارد مدل شود و از بین دو متغیر کوچکترین مقدار P-value یا بزرگترین مقدار قدر مطلق آماره t مربوط به متغیر سن می‌باشد. بنابراین متغیر سن به عنوان اولین متغیر ورودی انتخاب و مدل رگرسیون با متغیرهای مستقل (Age, Weight) و (Age, Sleep) توسعه داده می‌شود.

Variable	Coefficient	t Stat	P-value
Intercept	30.9941	2.5950	0.031865
Age	0.8614	3.4702	0.00844
Weight	0.3349	**2.5627**	0.033508
Variable	**Coefficient**	**t Stat**	**P-value**
Intercept	60.7308	5.4835	0.000585
Age	1.4516	12.1818	1.91E-06
Sleep	-0.1960	-0.2321	**0.82228**

از بررسی مدلهای دو متغیره می‌توان نتیجه گرفت که مدل انتخاب می‌گردد و به عنوان مدل نهایی باید متغیر میزان خواب به آن اضافه و بررسی گردد.

Variable	Coefficient	t Stat	P-value
Intercept	31.6866	2.1074	0.073065
Age	0.8598	3.2341	0.014369
Weight	0.3338	2.3822	0.048725
Sleep	-0.0586	-0.0871	**0.933032**

با بررسی مدل نهایی مشخص می‌گردد متغیر میزان خواب نمی‌تواند به مدل اضافه گردد. بنابراین مدل نهایی با متغیرهای سن و وزن به شرح زیر مدل نهایی خواهد بود.

$$\boldsymbol{Y = 30.994 + 0.8614Age + 0.3349Weight}$$

Text provided by Abbas Mahmoudabadi using Excel on April 14, 2018. Data is available at net.

Stepwise Regression (An example)

The researchers were interested in learning how the composition of the cement affected the heat evolved during the hardening of the cement. Therefore, they measured and recorded the following data on 13 batches of cement:

Response y: Heat evolved in calories during hardening of cement on a per gram basis

Predictor x1: % of tricalcium aluminate

Predictor x2: % of tricalcium silicate

Predictor x3: % of tetracalcium alumino ferrite

Predictor x4: % of dicalcium silicate

X1	X2	X3	X4	Y	X1	X2	X3	X4	Y
7	26	6	60	78.5	1	31	22	44	72.5
1	29	15	52	74.3	2	54	18	22	93.1
11	56	8	20	104.3	21	47	4	26	115.9
11	31	8	47	87.6	1	40	23	34	83.8
7	52	6	33	95.9	11	66	9	12	113.3
11	55	9	22	109.2	10	68	8	12	109.4
3	71	17	6	102.7					

Specify an Alpha-to-Enter significance level. $\alpha_E = 0.15$.
Specify an Alpha-to-Remove significance level. $\alpha_R = 0.15$.

Making one-predictor regression models

	Coefficients	t Stat	P-value
Intercept	81.4793442	16.53619	4.07E-09
X1	1.868747684	3.550002	0.004552
	Coefficients	t Stat	P-value
Intercept	57.42368296	6.763219	3.1E-05
X2	0.789124795	4.686214	0.000665
	Coefficients	t Stat	P-value
Intercept	110.2026563	13.86573	2.6E-08
X3	-1.25578125	-2.09843	0.059762
	Coefficients	t Stat	P-value
Intercept	117.5679312	22.34195	1.62E-10
X4	-0.73816181	-4.77478	0.000576

All p-values are less than 0.15; the biggest absolute of t-stats is |-4.77478|, so X4 inters as the first variable to the regression model. Three two-predictor regression models of (X1-X4; X2-X4; X3-X4) should be developed.

	Coefficients	t Stat	P-value
Intercept	103.0973816	48.53963	3.32E-13
X1	1.439958285	10.40307	1.11E-06
X4	-0.61395363	-12.6212	1.81E-07
	Coefficients	t Stat	P-value
Intercept	94.16007475	1.662808	0.127329
X2	0.310904744	0.415312	0.686684
X4	-0.4569419	-0.65657	0.526276
	Coefficients	t Stat	P-value
Intercept	131.2824064	40.08909	2.23E-12
X3	-1.19985118	-6.3478	8.38E-05
X4	-0.72460014	-10.0179	1.56E-06

Predictor X2 can’t be entered to the model. X1 and X3 are candidate to enter but absolute value of t-stat for X1 is the biggest. The second predictor is X1. So, X1 is added to model and two regression models of (X4, X1, X2; and (X4, X1, X3) should be examined.

	Coefficients	t Stat	P-value
Intercept	111.6844054	24.47886	1.52E-09
X1	1.051854159	4.7024	0.001116
X3	-0.41004331	-2.05812	0.069692
X4	-0.64279615	-14.4305	1.58E-07

	Coefficients	t Stat	P-value
Intercept	71.64830697	5.066208	0.000675
X1	1.451937963	12.40998	5.78E-07
X2	0.416109762	2.241844	0.051687
X4	-0.23654022	-1.36501	0.205395

Both predictors X2 and X3 are candidate to enter and X2 is entered (bigger absolute of t-stat), but predictor X4 should leave model because P-value is more than α_R=0.15. Therefore, final predictors are X1 and X2. In this case, two regression models of (X1, X2, X3; X1, X2, X4) should be checked.

	Coefficients	t Stat	P-value
Intercept	48.19363432	12.31533	6.17E-07
X1	1.695890167	8.289537	1.66E-05
X2	0.656914878	14.85083	1.23E-07
X3	0.250017607	1.353561	0.208889

	Coefficients	t Stat	P-value
Intercept	71.64830697	5.066208	0.000675
X1	1.451937963	12.40998	5.78E-07
X2	0.416109762	2.241844	0.051687
X4	-0.23654022	-1.36501	0.205395

As shown, X3 and X4 can't be entered to regression model. So the final remaining predictors are X1 and X2.

Model should be developed again for predictors X1 and X2 as follow:

ANOVA

	df	*SS*	*MS*	*F*	*Sig. F*
Regression	2	2657.85859	1328.929	229.5037	4.41E-09
Residual	10	57.9044832	5.790448		
Total	12	2715.76308			

	Coefficients	*t Stat*	*P-value*
Intercept	52.57734888	22.99796	5.46E-10
X1	1.468305742	12.10465	2.69E-07
X2	0.662250491	14.44236	5.03E-08

Final regression equation: $\boldsymbol{Y = 52.577 + 1.468X_1 + 0.662X_2}$