رگرسیون

برای اندازه گیری رابطه بین دو یا چند متغیر، شاخص های متعددي وجود دارد که همه این شاخص ها میزان رابطه بین متغیرها را تنها با یک مقدار به نام ضریب همبستگی نشان میدهند. اما سکه رابطه دو رو دارد، یک روی آن مقدار همبستگی بین دو متغیر است و روی دیگر آن، استفاده از این رابطه و یافتن معادله ای است که این رابطه را تبیین می‌کند. از آنجا که ممکن است رابطه بین متغیر مستقل و وابسته را به هر صورت ممکن نوشت، مجموعه ای از روشها نیز وجود دارند که می توان به کمک آنها یک معادله ریاضی بین متغیرها تعریف کرد و به کمک آنها مقادیر متغیر وابسته را از روی متغیر یا متغیرهای مستقل، پیش بینی کرد. در صورتی که رابطه بین متغیرها معنی دار باشد، می توان آن را با الگوهای ریاضی بیان کرد. معمولا چنین الگویی ممکن است از نوع خطی یا غیر خطی باشد. به معادله ای که رابطه بین دو متغیر مستقل و وابسته را نشان میدهد، معادله رگرسیون می گویند. اگر بتوان الگوی همبستگی را به صورت یک معادله خط نوشت، به آن معادله رگرسیون خطي می گویند و در غیر این صورت به آن معادله رگرسیون غیر خطی می گویند.

در رگرسیون هدف آن است که با استفاده از معادله رگرسیون و به کمک يك نمونه تصادفي و بعضی روشهای آماری، رفتار متغیر وابسته را با آگاهی از مقادیر و مشخصات متغیرهای مستقل، پیش بینی کنیم.

رگرسیون خطی ساده

رگرسیون خطی، ممکن است ساده یا چند گانه باشد. رگرسیون خطی ساده شامل یک متغیر وابسته و یک متغیر مستقل است ولی رگرسیون خطی چندگانه به ارزیابی رابطه یک متغیر وابسته با چند متغیر مستقل می پردازد.

در رگرسیون خطی ساده ابزاری که به خوبی می تواند الگوی همبستگی دو متغیر را به لحاظ بصری نمایان کند، نمودار پراکنش است. وقتی در یک نمودار پراکنش نقاط بطور تقریبی در امتداد یکدیگر قرار می گیرد ولی دقیقاً روی یک خط واقع نیستند، می توان خطی را فرض کرد که از میان نقاط طوری بگذرد که بیشتر از هر خط دیگری به نقاط نزدیکتر باشد. چنین خطی را به عنوان خط رگرسیون می شناسیم. می توانیم از این خط رگرسیون مثلا جهت تخمین میزان محصول در سال جاری با توجه به میزان بارندگی های اخیر، استفاده کرد. برای اینکار با گذاشتن مقدار بارندگی در معادله خط رگرسیون، میزان محصول پیش بینی خواهد شد. البته ممکن است این مقدار با مقدار واقعی قدری تفاوت داشته باشد. به این تفاوت ها مقادیر باقی مانده (Residual) می گویند.

در رگرسیون خطی ساده اگر y را متغیر وابسته و x را متغیر مستقل در نظر بگیریم، می توان معادله خط رگرسیون را به صورت زیر نوشت:

$$y' = b_0 + b_1 x$$

در این معادله y' مقدار برآورد شده y ، b_1 شیب خط رگرسیون یا ضریب رگرسیون و b_0 را عرض از مبداء خط یا ثابت رگرسیون می گویند.

در رگرسیون خطی چندگانه مقادیر یک متغیر وابسته مانند y از روی مقادیر دو یا چند متغیر مستقل دیگر برآورد می شود. معادله رگرسیون خطی چندگانه را می توان به صورت کلی زیر نوشت:

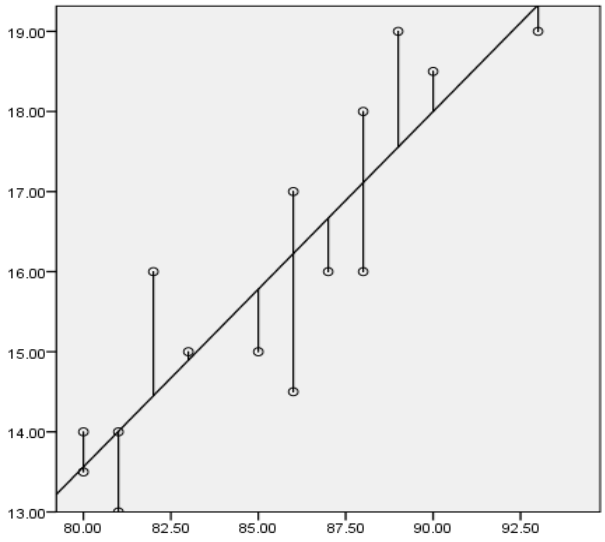
$$y' = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

در این معادله پارامترهای $b_1, b_2, \ldots, b_k$ ضرایب رگرسیون جزئی و b_0 مقدار ثابت رگرسیون است. این معادله را به عنوان معادله رگرسیون خطی چندگانه y بر اساس $x_1, x_2, \ldots, x_k$ می شناسیم.

باقیمانه ها

اگر در معادله خط رگرسیون که بر اساس داده های مطالعه آن را به عنوان مناسب ترین خط برازش شده به دست آورده اید، مقادیر مختلف متغیر x را قرار دهید، برای متغیر وابسته (y) مقادیری به دست خواهید آورد که با اندازه

های مشاهده شده y متفاوت خواهند بود. به عبارتی نقاط مشاهده شده y بر نقاط برآورد شده آن یعنی y' منطبق نخواهند بود. این به آن مفهوم است که نقاط مربوط به داده ها دقیقا بر روی خط مستقیم یا صفحه و فوق صفحه ای که توسط معادله خط رگرسیون مشخص شده است، نمی افتد. این اختلاف $e_i = (y - y')$ در متغیر پیش بینی شده را به عنوان باقیمانده یا خطا می شناسیم.



بررسی این مقادیر در مدل های رگرسیون از اهمیت ویژه ای برخوردار است زیرا از آنها می توان به عنوان شاخصی برای صحت برآورد معادله خط رگرسیون استفاده کرد.

شرایط رگرسیون خطی ساده

در رگرسیون خطی ساده باید شرایط زیر برقرار باشد:

۱- میانگین (امید ریاضی) خطاها صفر باشد. یعنی: $E(e_i) = ۰$

۲- واریانس خطاها ثابت باشد. به عبارت دیگر $v(e_i) = \sigma^۲$

اگر فرض های ۱و ۲ بر قرار باشند به این معنی است که توزیع خطاها نرمال است.

۳- بین خطاها، همبستگی وجود نداشته باشد. به عبارت دیگر $COV(e_i, e_j) = ۰$.

٤- متغیر وابسته (y) دارای توزیع نرمال باشد.

٥- متغیرهاي مستقل دارای هم خطی نباشند. یعنی بین متغیرهای مستقل همبستگی معنا داری وجود نداشته باشد.

ضریب همبستگی چندگانه

وقتی شما در معادله خط رگرسیون مقادیر مشاهده شده x را قرار میدهید، به ازائ هر x یک مقدار برای y به دست خواهید آورد. همبستگی این مقدار با مقادیر مشاهده شده y در نمونه به عنوان شاخصی برای کارایی رگرسیون در پیش بینی y مورد استفاده قرار می گیرد.همبستگی بین y و y' را به عنوان ضریب همبستگی چندگانه می شناسیم و آن را با R نشان میدهیم.

در رگرسیون خطی ساده ضریب همبستگی چندگانه همان قدر مطلق ضریب همبستگی پیرسن بین متغیر مستقل و وابسته است که همواره عددی مثبت خواهد بود.

روند رگرسیون خطی ساده

فرض کنید می خواهید رابطه بین نمره امتحان ورودی (متغیر مستقل vorud) و معدل کل هنگام فارغ التحصیلی یک دانشجو (متغیر وابسته khoruj) را بررسی کنید و از این رابطه یک الگوی خطی بسازید و سپس از این الگو برای دانشجویی که با یک نمره ورودی خاص پذیرفته شده است، معدل پایان تحصیلات او را پیش بینی نمایید. در ابتدا لازم است شما مطمئن شوید همبستگی بین این دو متغیر معنی دار است سپس به سراغ یک معادله خط رگرسیون برای این دو متغیر باشید.

مثال:

برای انجام روند رگرسیون خطی ساده، داده های جدول زیر که نمره ورودی و معدل پایان تحصیلات مربوط به یک نمونه ۱۵ تایی از دانشجویان است، در نظر بگیرید.

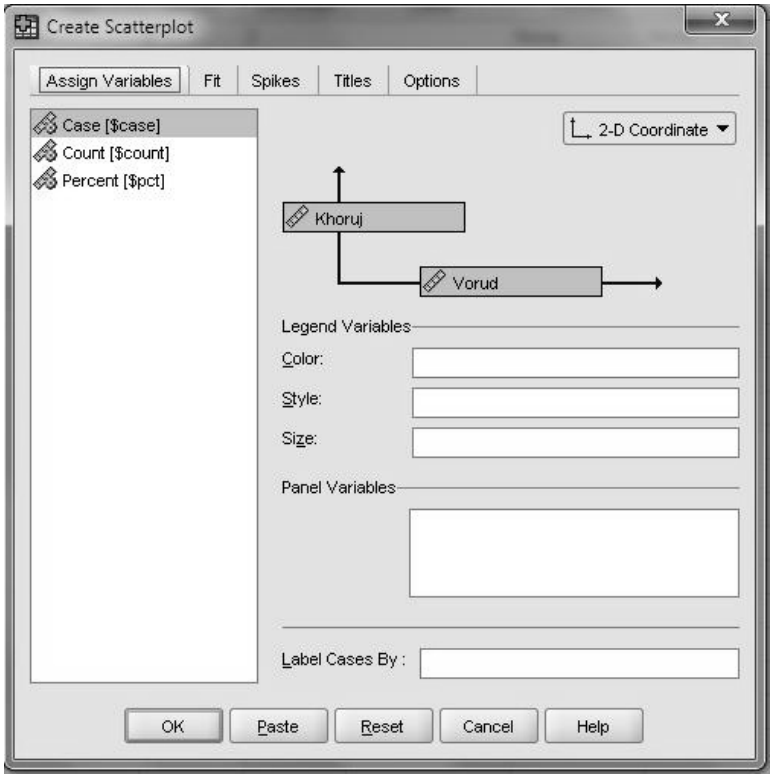
vorud	۸۰	۸۰	۸۱	۸۱	۸۲	۸۳	۸۵	۸۶	۸۶	۸۷	۸۸	۸۸	۸۹	۹۰	۹۳
khoruj	۱۴	۱۳.۵	۱۳	۱۴	۱۴	۱۶	۱۵	۱۵	۱۷	۱۶	۱۶	۱۸	۱۹	۱۸.۵	۱۹

بررسی های اولیه

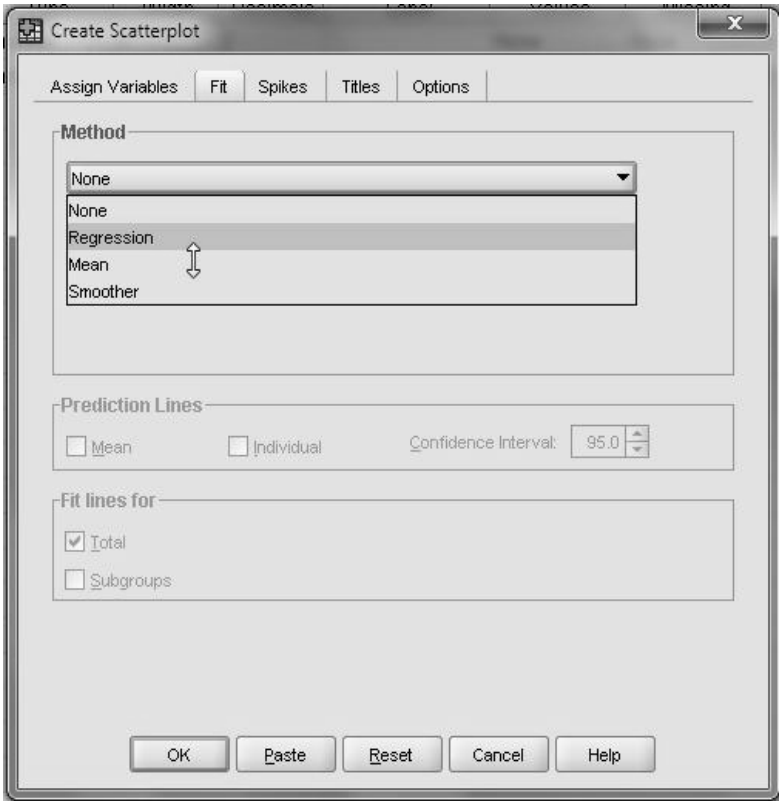
ابتدا داده ها را به همان شکلی که در جدول بالا آمده اند و با دو متغیر نمره ورودی (vorud) و معدل پایان تحصیلات (khoruj) به spss وارد کنید. همیشه قبل از انجام روند رگرسیون، رسم نمودار پراکنش داده ها و مشاهده الگوی همبستگی آنها مفید است. برای این کار از منوی اصلی مسیر زیر را دنبال کنید:

Graphs/Legacy Dialogs/Interactive/Scatterplots…

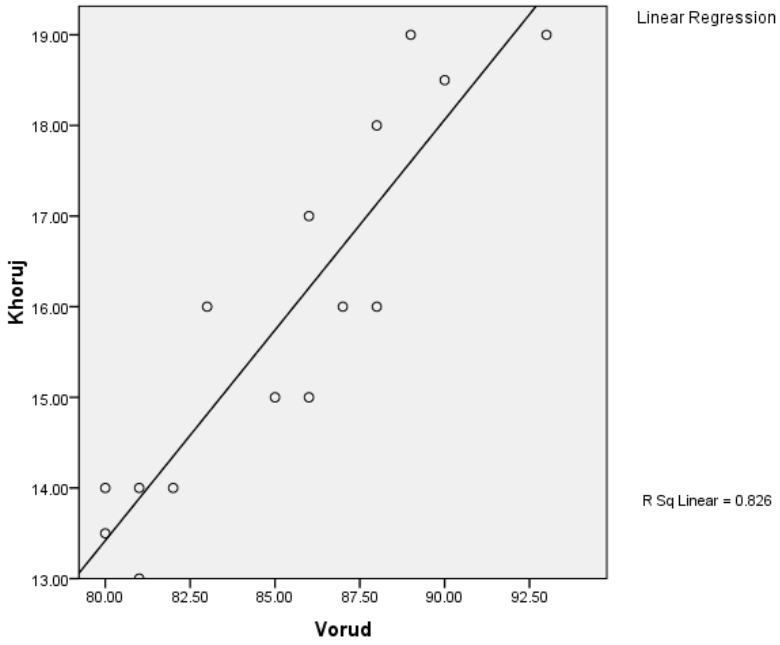
- در کادر مکالمه Create Scatterplots و در برگه Assig Variable متغیر نمره ورودی (Vorud) را به محور افقی و متغیر معدل پایان تحصیلات (Khoruj) را به محور عمودی منتقل کنید.



- در برگه fit و در کشویی Method گزینه Regression را انتخاب کنید.
- در همین برگه بررسی کنید گزینه include costant in equation برای وجود مقدار ثابت در معادله رگرسیون، انتخاب شده باشد.



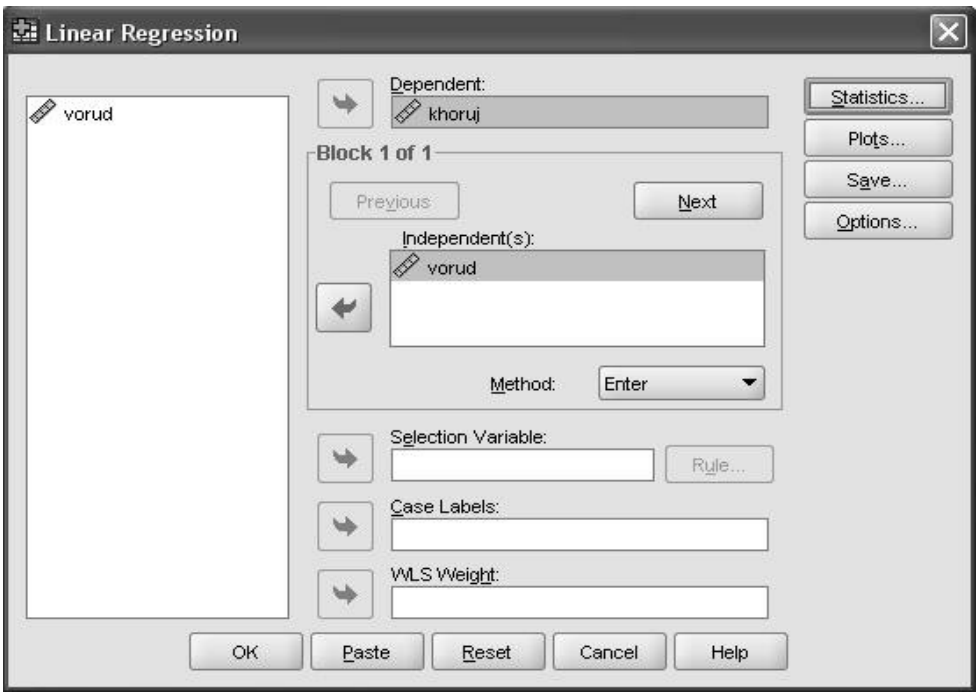
نتیجه نموداري به صورت زیر خواهد بود.



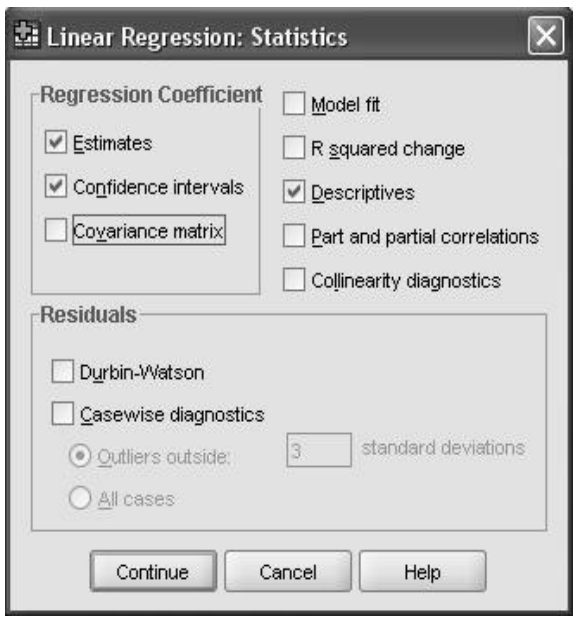
با توجه به شکل بالا و ضریب همبستگی رگرسیون می توان به طور مناسبی یک معدله خط برای رابطه بین دو متغیر برازش داد. بنا براین برای به دست آوردن چنین معادله ای در روند رگرسیون خطی مراحل زیر را دنبال کنید:

- کادر محاوره رگرسیون خطی را از مسیر زیر باز کنید:

Analyze/Regression/Linier



- متغیر وابسته معدل پایان تحصیلات (khoruj) را به کادر Dependent List و متغیر مستقل نمره امتحان ورودی (vorud) را به کادر Independent Variable وارد کنید.
- با انتخاب گزینه Statistics در کادر مکالمه آن باز شده:

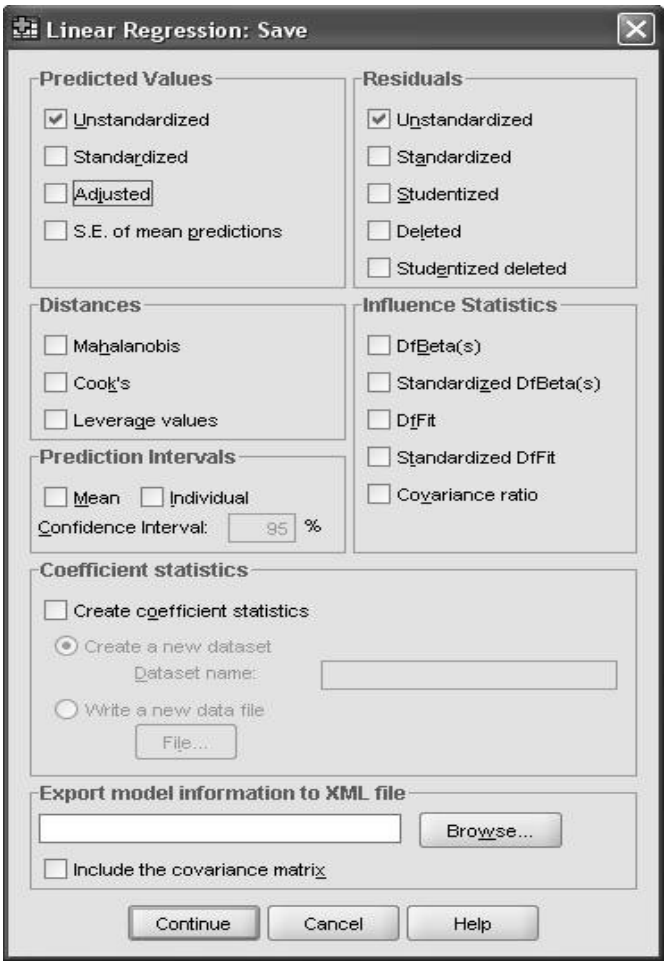


۱- گزینه های Estimates را برای محاسبه برآوردها انتخاب کنید.

۲- گزینه Confidence Interval را جهت به دست آوردن فواصل اطمینان برای برآوردها انتخاب کنید.
۳- گزینه Descriptive را برای محاسبه بعضی آماره های ضروری انتخاب کنید.

- در ادامه Continue را کلیک کنید و به کادر محاوره اصلی باز گردید.
- از گزینه Plots برای رسم نمودار رگرسیون بر اساس مقادیر پیش بینی شده مقادیر برآورد شده استفاده می شود. همچنین می توان هیستوگرام فراوانی را برای باقیمانده ها رسم کرده وآنرا با منحنی توزیع نرمال مقایسه کرد.

- با انتخاب گزینه Save در کادر محاوره Linier Regression می توانی طیف وسیعی از مقادیر پیش بینی شده و باقیمانده ها و آماره های مفیدی برای تشخیص وضعیت رگرسیون را در حالت های مختلف محاسبه و در فایل داده ها به عنوان یک متغیر جدید ذخیره کنید.



ما در این بخش به محاسبه مقادیر پیش بینی شده و باقیمانده ها اکتفا می کنیم. برای این منظور در بخش Predicted Values گزینه Unstandardized و در بخش Residual نیز گزینه Unstandardized را انتخاب کنید و Continue را کلیک کنید.

- برای اجرای این رگرسیون گزینه Ok را کلیک کنید و نتیجه را به صورت زیر مشاهده نمایید.

Descriptive Statistics

	Mean	Std. Deviation	N
khoruj	۱۶.۰۰۰۰	۱.۹۷۳۰۳	۱۵
vorud	۸۵.۲۶۶۷	۳.۹۹۰۴۶	۱۵

Correlations

		khoruj	vorud
Pearson Correlation	khoruj	۱.۰۰۰	.۹۰۷
	vorud	.۹۰۷	۱.۰۰۰
Sig. (۱-tailed)	khoruj	.	.۰۰۰
	vorud	.۰۰۰	.
N	khoruj	۱۵	۱۵
	vorud	۱۵	۱۵

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
۱	vorud[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: khoruj

Coefficients[a]

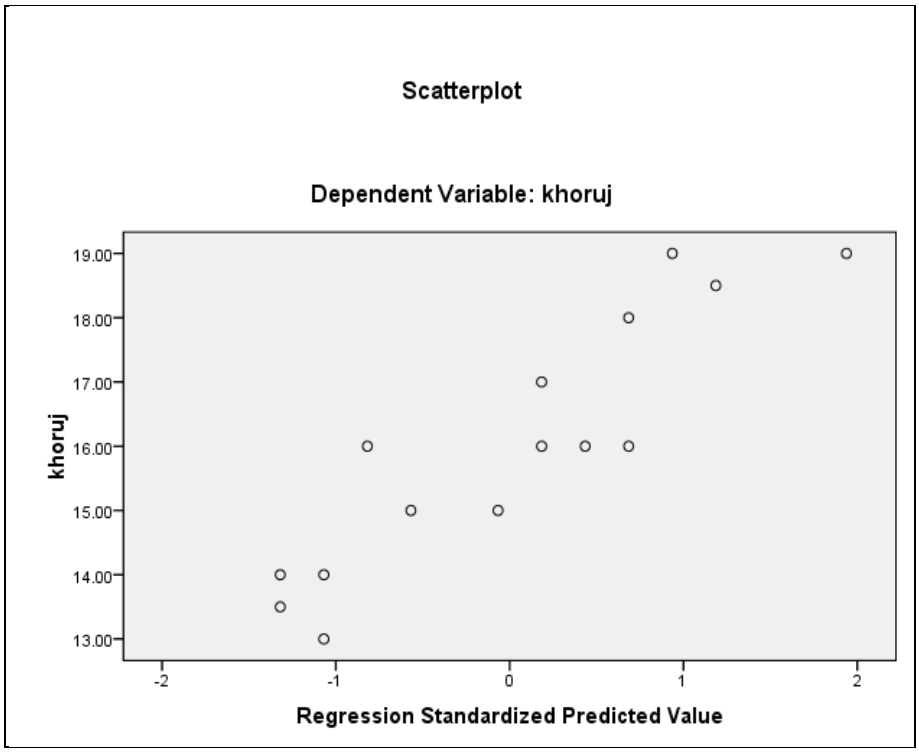
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	۹۵% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
۱	(Constant)	-۲۲.۲۴۸	۴.۹۲۴		-۴.۵۱۹	.۰۰۱	-۳۲.۸۸۴	-۱۱.۶۱۱
	vorud	.۴۴۹	.۰۵۸	.۹۰۷	۷.۷۷۶	.۰۰۰	.۳۲۴	.۵۷۳

a. Dependent Variable: khoruj

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	۱۳.۶۳۷۶	۱۹.۴۶۸۹	۱۶.۰۰۰۰	۱.۷۸۹۹۸	۱۵
Residual	-۱.۲۲۶۰۸	۱.۴۶۵۳۱	.۰۰۰۰۰	.۸۲۹۹۵	۱۵
Std. Predicted Value	-۱.۳۲۰	۱.۹۳۸	.۰۰۰	۱.۰۰۰	۱۵
Std. Residual	-۱.۴۲۴	۱.۷۰۱	.۰۰۰	.۹۶۴	۱۵

a. Dependent Variable: khoruj



اگر به فایل داده ها نگاه کنید می توانید دو متغیر PRE_۱ و RES_۱ که به ترتیب مربوط به مقادیر پیش بینی شده و باقیمانده ها هستند را مشاهده کنید.

در جدول Descriptive Statistics میانگین و انحراف معیار متغیرها را مشاهده می کنید. جدول Correlations ضرایب همبستگی و آزمون مربوط به آن را نمایش میدهد و مقدار P-Value در این جدول نشان از همبستگی بالای دو متغیر دارد. جدول Variables Entered/Removed متغیرهایی را که به معادله رگرسیون وارد و از آن خارج می‌شوند، نمایش میدهد. چون تنها یک متغیرمستقل در معادله رگرسیون وجود دارد، شما تنها متغیر مستقل نمره ورودی آزمون را مشاهده می کنید.

در جدول Coefficients می توانید ضرایب رگرسیون و آزمون های مربوط به آنها را مشاهده کنید. بر اساس ضرایب این جدول می توان معادله خط رگرسیون را به صورت زیر نوشت:

$$\text{khoruj} = -۲۲.۲۴۸ + ۰.۴۴۹(\text{vorud})$$

ستون Beta با مقدار ۰.۹۰۷ نشان از نقش موثر متغیر مستقل در پیش گویی کنندگی معادله رگرسیون دارد.

جدول Residuals Statistics آماره های مربوط به باقیمانده ها و مقادیر پیشگویی شده را نشان میدهد. در نمودار پراکنش ارائه شده شما می توانید مقادیر پیشگویی شده استاندارد را به ازای مقادیر مختلف متغیر مستقل مشاهده کنید.

موضوع رگرسیون غیر خطی را قبلا و به اختصار در همین وبلاگ مطرح کرده ایم.