با سلام و احترام،

فصل بیستم و چهارم: رگرسیون خطی (Linear Regression)

رگرسیون خطی برای مدل کردن مقدار یک متغیر کمی وابسته که بر رابطه خطی‌اش با یک یا چند پیشگو بنا شده است به کار می‌رود.

رگرسیون خطی یکی از تکنیک‌های پیچیده آماری برای داده‌هایی است که معمولاً در سطح سنجش فاصله‌ای می‌باشند. رگرسیون خطی به دو صورت رگرسیون خطی ساده و رگرسیون خطی چند متغیره مطرح می‌گردد. رگرسیون خطی ساده به پیش‌بینی مقدار یک متغیر وابسته بر اساس مقدار یک متغیر مستقل می‌پردازد. اما رگرسیون چند متغیره روشی است برای تحلیل مشارکت جمعی و فردی دو یا چند متغیر مستقل (xi) در تغییرات یک متغیر وابسته (y). از آنجا که وظیفه اساسی علم، پیش‌بینی و تبیین پدیده‌ها است بنابراین در تحقیقاتی که بر پیش‌بینی یا تبیین ناظرند، تحلیل رگرسیون می‌تواند نقش بارزی ایفا کند. در این نوع تحقیقات محقق می‌کوشد تا بر اساس اطلاع از یک یا چند متغیر مستقل، به یک معادله رگرسیونی دست یابد و از آن برای پیش‌بینی مقادیر متغیر وابسته استفاده کند.

مدل رگرسیون خـــــــــطی

مدل رگرسیون خطی فرض می‌کند که یک رابطه خطی (یا خط مستقیم) بین متغیر وابسته و هر پیشگو وجود دارد. این رابطه در فرمول زیر توضیح داده شده است.

$$y_i = b_0 + b_j x_{ij} + \ldots + b_p x_{ip} + e_i$$

که در آن

y_i: مقدار مورد iام متغیر کمی وابسته است.

p: تعداد پیشگوها می‌باشد.

b_j: مقدار ضریب jام است، p و ... و j =0

x_{ij}: مقدار مورد iام از پیشگوی jام می‌باشد.

e_i: خطای در مقدار مشاهده شده برای مورد iام است.

مدل خطی است زیرا با افزایش مقدار پیشگوی jام با یک واحد باعث افزایش مقدار وابسته واحدهای b_j می‌شود. توجه کنید که b_0 عرض از مبدأ است، که وقتی مقدار هر پیشگو برابر صفر می‌شود، b_0 مقدار مدل پیشگوی متغیر وابسته می‌باشد.

رگرسیون خطی ساده زمانی مورد استفاده قرار می‌گیرد که یک متغیر وابسته و یک متغیر مستقل داریم. از طرفی، مقیاس هر دو متغیر (هم وابسته و هم مستقل) در سطح سنجش حداقل فاصله‌ای است. بنابراین در رگرسیون دو متغیر ساده مقادیر یک متغیر (متغیر وابسته یا y) از روی مقادیر متغیر دیگر (متغیر مستقل یا x) به کمک یک معادله خط (خط مستقیم) برآورد می‌شود. یعنی معادله فوق به شکل زیر در می‌آید:

$$y = b_0 + b_1(x_1)$$

به منظور آزمایش فرضیه‌های مربوط به مقادیر پارامترهای مدل، مدل رگرسیون خطی نیز فرضیات زیر را در نظر می‌گیرد:

- عبارت خطا یک توزیع نرمال با میانگین صفر دارد
- واریانس عبارت خطا در سرتاسر موارد ثابت می‌باشد و از متغیرها در مدل مستقل است. یک عبارت خطا با واریانس غیر ثابت را heteroscedastic می‌نامند.
- مقدار عبارت خطا برای یک مورد داده شده مستقل از مقادیر متغیرها در مدل و مستقل از مقادیر عبارت خطا برای موارد دیگر می‌باشد.

استفاده از رگرسیون خطی برای پیشگویی زمان‌های پرداخت کاری

شرکتی یک خط تولید دارد که نیازمند به یک مرحله پرداخت کاری در فرآیند ساخت می‌باشد. برای برنامه‌ریزی زمان تولید، زمان‌های پرداخت کاری ۵۹ محصول، به همراه نوع محصول و اندازه‌های مرتبط با آن، ثبت شده است.

از رگرسیون خطی برای تعیین زمان پرداخت کاری که با اندازه محصول می‌تواند پیشگویی شود استفاده نمایید.

قبل از اجرای رگرسیون، باید یک نمودار پراکنش از زمان پرداخت کاری نسبت به اندازه محصول تهیه کنید تا مشخص شود که آیا مدل خطی برای این متغیرها منطقی است.

ایجاد نمودار پراکنش متغیر وابسته نسبت به مستقل

۱. برای ایجاد یک نمودار پراکنش از متغیر diam بر حسب time، مسیر زیر را از منوی اصلی برگزینید:

Graphs >Legacy Dialogs> Scatter / Dot

۲. دکمه Define را کلیک کنید.

۳. Time را به عنوان متغیر y و diam را به عنوان متغیر x برگزینید.

٤. Ok را کلیک کنید. در نتیجه نمودار پراکنش ایجاد می‌شود.

٥. برا دیدن بهتر خطی که روی نقاط این نمودار قرار می‌گیرد، با دو بار کلیک کردن نمودار، آن را فعال نمایید.

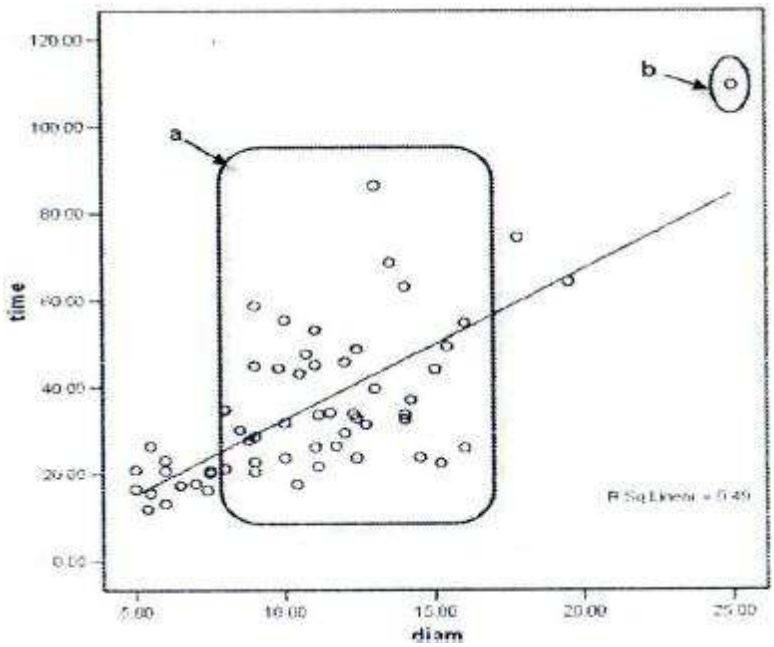
٦. نقطه‌ای را در Chart Editor برگزینید.

۷. آیکن Add Fit Line را کلیک کنید و سپس پنجره Chart Editor را ببندید.

نمودار پراکنش نتیجه با خط رگرسیون مناسب ظاهر می‌گردد.

a- تغییر پذیری زمان پرداخت کاری طوری ظاهر می‌شود که با افزایش اندازه، زیاد می‌شود.

b- نقطه موجود در بالا و سمت راست نمودار ممکن است تأثیر بیش از حدی در وضعیت خط رگرسیون بگذارد.



آغـــــــــــاز تحلیل

۱. برای اجرای یک تحلیل رگرسیون خطی، مسیر Analyze > Regression > Linear را از منوی اصلی برگزینید.
۲. Time را به عنوان متغیر وابسته برگزینید.
۳. diam را به عنوان متغیر مستقل انتخاب کنید.
٤. type را به عنوان متغیر عنوان‌گذاری مورد (Case Labels) انتخاب نمایید.
٥. دکمه Plots را کلیک کنید.
٦. *SDRESID را به عنوان متغیر y و *ZPRED را به عنوان متغیر x برگزینید.
۷. گزینه‌های Histogram و Normal Probability Plot را فعال کنید.
۸. دکمه Continue را کلیک کنید.
۹. دکمه Save را در کادر محاوره Linear Regression کلیک کنید.
۱۰. در مجموعه Predicted Values گزینه Standardized را فعال نمایید.
۱۱. در مجموعه Residuals گزینه Standardized را فعال کنید.
۱۲. گزینه‌های Cook's و Leverage Values را در مجموعه Distances فعال نمایید.
۱۳. دکمه Continue را کلیک کنید.
۱٤. دکمه Ok را در کادر محاوره Linear Regression کلیک کنید.

این مراحل، یک رگرسیون خطی را برای زمان پرداخت کاری بر حسب اندازه ایجاد می‌نماید. نمودارهای تشخیصی باقی‌مانده‌های استاندارد شده با مقادیر مدل پیشگو مورد نیاز هستند، و مقادیر مختلف برای آزمون تشخیصی بیشتر ذخیره شده‌اند.

جدول زیر ضرائب رگرسیون خطی را نشان می‌دهد داده‌های این جدول بیان می‌کند که زمان پرداخت مورد انتظار برابر **3.475***DIAM-**1.955** می‌باشد. اگر شرکت بخواهد یک قابلمه ۱۵ اینچی را بسازد، زمان پرداخت کاری برابر **3.457***15-**1.955**=**49.**9، یا حدود **50** دقیقه می‌باشد.

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-1.955	5.402		-.362	.719
	DIAM	3.457	.467	.700	7.407	.000

بـــــــــــــــررسی همواری مدل

جدول ANOVA مقبولیت مدل را از منظر آماری بررسی می‌کند. در این جدول نتایج مربوط به ضرائب تأثیر رگرسیونی متغیر مستقل بر متغیر وابسته را نشان می‌دهد.

اولین آماره‌ای که در این جدول می‌بینید عدد ثابت (Constant) است که همان عرض از مبدأ می‌باشد و میزان متغیر وابسته را بدون دخالت متغیرهای مستقل نشان می‌دهد.

دو نوع ضریب تأثیر رگرسیونی داریم. اول، ضرائب تأثیر رگرسیونی استاندارد نشده B یا Unstandardized coefficient و دوم ضریب تأثیر رگرسیونی استاندارد شده بتا (β) یا Standardozed coefficient. مطابق این توضیحات مدل برآورد شده به صورت زیر می‌باشد:

زمان پرداخت کاری =**-1/955 + 3/475** (قطر قطعه)

اما از آنجا که در تحلیل رگرسیون، مقیاس اغلب متغیرهای مستقل، از واحدهای متفاوتی تشکیل یافته، بنابراین به راحتی نمی‌توان به مقایسه سهم هر متغیر مستقل در تبیین تغییرات یا واریانس متغیر وابسته پرداخت. به همین دلیل، ضرایب رگرسیونی استاندارد شده (β) به ما کمک می‌کنند تا سهم نسبی هر متغیر مستقل در تبیین تغییرات متغیر وابسته زمان پرداخت کاری را تعیین نمائیم. یعنی هرچه مقدار ضریب بتای یک متغیر بزرگتر باشد، نقش آن در پیش‌بینی تغییرات متغیر وابسته بیشتر است. از این رو به پژوهشگر پیشنهاد می‌شود که در تفسیر نتایج تأثیر رگرسیونی بر اساس ضرائب آن، به جای ضرائب رگرسیونی استاندارد نشده به ضرایب رگرسیونی استاندارد شده استناد کنیم.

آماره t اهمیت نسبی حضور هر متغیر مستقل در مدل را نشان می‌دهد. برای اینکه تشخیص دهیم کدام متغیرها تأثیر آماری معنی‌داری بر متغیر وابسته داشتند، می‌توانیم به مقدار t نگاه کنیم. معمولاً هر گاه قدر مطلق مقدار این آماره برای متغیری بزرگتر از عدد **2.33** باشد، سطح خطای آن نیز کوچکتر از **0.01** یا **0.05** بوده و در نتیجه خواهیم گفت که متغیر مورد نظر تأثیر آماری معنی‌داری در تبیین تغییرات متغیر وابسته داشته است.

سطر Regression اطلاعات راجع به تغییر مدل شما را نشان می‌دهد (a). سطر Residual اطلاعات راجع به تغییر که در مدل شما به حساب نمی‌آید نشان می‌دهد (b).

هر چه مقدار مجموع مربعات باقیمانده کوچکتر از مجموع مربعات رگرسیون باشد نشان دهنده قدرت تبیین بالای مدل در توضیح تغییرات متغیر وابسته است. بر عکس، هر چه مقدار باقیمانده به میزان بیشتری از رگرسیون بزرگتر باشد، نشان می‌دهد که نقش مدل در تبیین تغییرات متغیر وابسته ضعیف است. در چنین حالتی، باید متغیرهای مستقل ضعیف‌تر را از مدل حذف کرد و متغیرهای مستقل دیگری را که نقش بیشتری در تبیین تغییرات متغیر وابسته دارند، وارد مدل کرد.

میانگین مربعات (Mean Square) از تقسیم مجموع مربعات هر منبع بر درجه آزادی همان منبع حاصل می‌شود. مقدار F به بررسی مناسب بودن یا نبودن مدل رگرسیونی می‌پردازد. یعنی به بررسی اینکه آیا متغیرهای مستقل قادرند به خوبی تغییرات متغیر وابسته را توضیح دهند یا نه. تشخیص این موضوع، با معنی‌داری مقدار F در سطح خطای کوچکتر یا بزرگتر از **0.05** امکان‌پذیر است.

جمع مربعات رگرسیون و باقی‌مانده تقریباً برابرند، که نشان می‌دهد نیمی از تغییر در زمان پرداخت کاری توسط مدل نشان داده شده است (c). سطح معنی‌داری آماره F کمتر از **0.05** می‌باشد، و این بدان معنی است که تغییر نشان داده شده به وسیله مدل بر اثر اتفاق نیست (d).

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10287.173	1	10287.173	54.865	.000
	Residual	10687.511	57	187.500		
	Total	20974.684	58			

a → Regression, c → Sum of Squares, b → F, d → Sig.

از آنجا که جدول ANOVA یک آزمون سودمند از توانایی مدل در توضیح تأثیر هر متغیر در متغیر وابسته می‌باشد، مستقیماً به شدت رابطه توجه ندارد.

جدول خلاصه مدل شدت رابطه بین مدل و متغیر وابسته را گزارش می‌نماید. در این جدول Model به تعداد مدل تشکیل شده اشاره دارد. پارامتر R یا ضریب همبستگی چندگانه، میزان همبستگی چندگانه بین مجموع متغیرهای مستقل و متغیر وابسته را نشان می‌دهد. این ضریب مقادیر بین صفر تا یک را می‌گیرد. نزدیک بودن به عدد یک نشان از همبستگی قوی بین متغیر مستقل و وابسته و نزدیک صفر بودن نشان از ضعف این همبستگی دارد.

R Square یا مربع ضریب همبستگی چندگانه یا همان ضریب تعیین، میزان تبیین واریانس و تغییرات متغیر وابسته توسط مجموعه متغیرهای مستقل را نشان می‌دهد. مقدار این ضریب بین صفر تا یک می‌باشد. هر چه به یک نزدیکتر باشد به این معنی است که متغیرهای مستقل توانسته‌اند مقدار زیادی از واریانس متغیر وابسته را تبیین نمایند و هر چه مقدار آن به صفر نزدیکتر باشد نشان می‌دهد متغیرهای مستقل نقش کمتری در تبیین واریانس متغیر وابسته دارد.

Adjusted R Square یا ضریب تعیین تعدیل شده که بیشتر برای تفسیر ضریب تعیین از آن استفاده می‌شود. با توجه به وجود برخی ایرادات در ضریب تعیین مثل بیش از اندازه برآورد کردن میزان موفقیت مدل و کمتر در

نظر گرفتن تعداد متغیرهای مستقل و حجم نمونه و به حساب نیاوردن تعداد درجات آزادی، پیشنهاد می‌شود از ضریب تعدیل شده برای تفسیر استفاده شود.

و در نهایت std. Error of Estimate یا خطای استاندارد تخمین که مقدار آن نشان دهنده مقدار قدرت پیش‌بینی معادله رگرسیون است.

R، یعنی ضریب همبستگی چندگانه، همبستگی خطی بین مقادیر مشاهده شده و مقادیر مدل پیشگوی متغیر وابسته می‌باشد. مقدار بزرگ آن یک رابطه قوی را نشان می‌دهد (a). R' Square، یعنی ضریب تعیین، مقدار مربع ضریب همبستگی چندگانه است. این نشان می‌دهد که حدود نیمی از تغییر در زمان با مدل تبیین شده است (b).

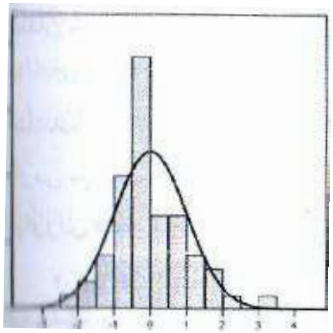
Model	a R	b R Square	Adjusted R Square	Std. Error of the Estimate
1	.700[a]	.490	.482	13.69307

برای بررسی بیشتر درباره همواری مدل، خطای استاندارد برآورد در جدول خلاصه مدل را با انحراف معیار زمان گزارش شده در جدول آماره‌های توصیفی مقایسه نمایید.

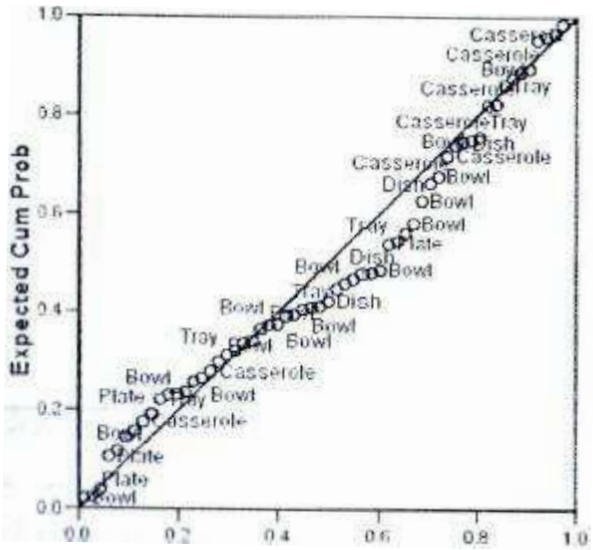
بدون آگاهی قبلی درباره اندازه یک محصول جدید، بهترین حدس شما برای زمان پرداخت کاری حدود **35.8** دقیقه با انحراف معیار **19.0** می‌باشد. با مدل رگرسیون خطی، خطای برآوردهای شما به طور معنی‌داری کمتر می‌باشد (حدود **13.7**).

بـــــــــــــررسی نرمال بودن عبارت خطا

Residual (باقی‌مانده) اختلاف بین مشاهده و مقادیر مدل پیشگوی متغیر وابسته است. باقی‌مانده یک محصول عبارتست از مقدار مشاهده شده عبارت خطا برای آن محصول. هیستوگرام یا نمودار p-p باقی مانده برای بررسی فرض نرمال بودن عبارت خطا مورد استفاده قرار می‌گیرد. شکل هیستوگرام تقریباً باید از شکل منحنی نرمال تبعیت کنند.

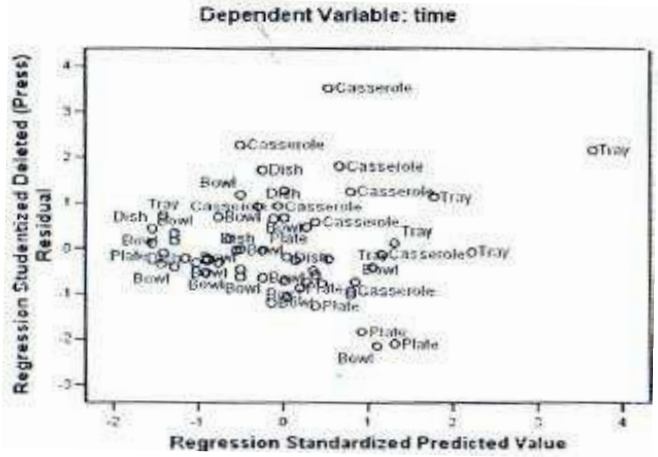


باقی‌مانده‌ها در نمودار p-p باید تابع خط ٤٥ درجه باشند. نه هیستوگرام و نه نمودار p-p فرضیه نرمالیته را نقض نمی‌کنند. در نمودار p-p نقاط روی خط ٤٥ درجه، نشان می‌دهند که احتمال تجمعی مشاهده شده با احتمال تجمعی مورد انتظار یکسان است. در حقیقت هر چقدر تجمع نقاط حول خط ٤٥ درجه بیش‌تر باشد با دقت بیشتری می‌توان متغیر وابسته را پیش‌بینی کرد.



بـــــــــــررسی استقلال عبارت خطا

نمودار باقی‌مانده‌ها نسبت به مقادیر پیشگویی نشان می‌دهد که واریانس خطاها با افزایش زمان پرداخت کاری پیشگویی شده افزایش می‌یابند.

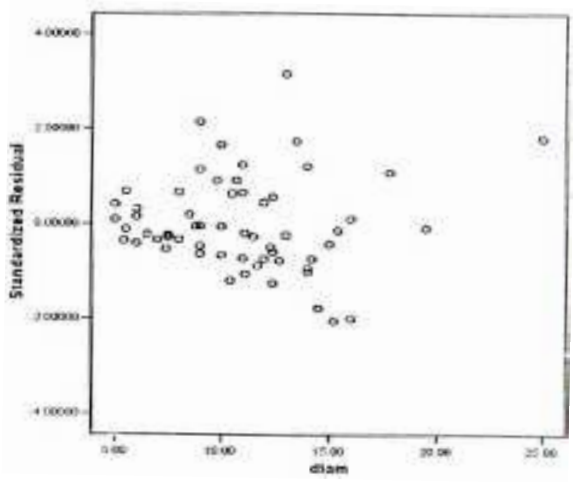


۱. برای بررسی باقی‌مانده‌ها نسبت به ابعاد، کادر محاوره Simple Scatterplot را فراخوانی کنید.

۲. به جای time، گزینه standardized Residual را متغیر y نمایید.

۳. Ok را کلیک کنید.

نمودار باقی‌مانده‌ها نسبت به ابعاد نتایج یکسانی را در بردارد. برای تصحیح Heteroscedasticity (عبارت خطا با واریانس غیر ثابت) در باقی‌مانده‌ها در تحلیل‌های تکمیلی، باید یک متغیر وزن‌دهی را بر اساس معکوس ابعاد محصول تعریف نمایید. استفاده از متغیر وزن‌دهی، تأثیر محصولات با ابعاد بزرگ و متغیر زمان‌های پرداخت کاری خیلی زیاد که ناشی از برآوردهای رگرسیون خیلی دقیق است کاهش می‌دهد.



شناســـــــــــایی نـــــــــقاط مؤثر

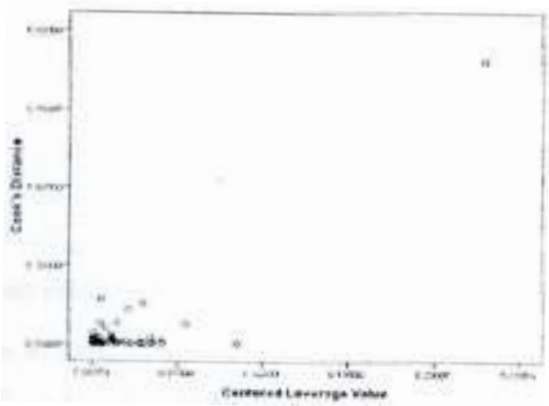
٤. برای بررسی نقاط مؤثر، کادر محاوره Simple Scatterplot را فراخوانی نمایید.

٥. به جای standardized Residual گزینه Cook's Distance را متغیر y نمایید.

٦. به جای diam گزینه Centered Leverage Value را متغیر x کنید.

٧. Type را برای متغیر عنوان‌گذاری مورد (Label Cases by) برگزینید.

٨. Ok را کلیک کنید.



نمودار پراکنش نتیجه، یک نقطه را در سمت راست در فاصله‌ای دور از بقیه نشان می‌دهد.

٩. برای شناسایی نقطه، نمودار را با دو بار کلیک کردن فعال نمایید.

۱۰. آیکن Data Label Mode را کلیک کنید.

۱۱. نقطه را انتخاب نمایید. این نقطه با کلمه Tray معرفی می‌شود.

این مورد دارای اثر اهرمی و تأثیر زیادی می‌باشد. اثر اهرمی آن باعث وزن بالا در محاسبات رگرسیون خطی می‌شود، و تأثیر بالای آن روی شیب و رگرسیون خطی اعمال می‌گردد. شما می‌توانید با یک نقطه مؤثر به کمک یک متغیر وزن‌دهی که نقطه مؤثر را کم وزن می‌کند سر و کار داشته باشید.

خـــــــــــــلاصه

دانستن زمان پرداخت کاری برای هر محصول به شرکت کمک می‌کند تا برنامه زمان‌بندی مناسب‌تری داشته باشد. با استفاده از قابلیت رگرسیون خطی می‌توانید با استفاده از رابطه بین ابعاد محصول و زمان پرداخت کاری برنامه زمان‌بندی را به روز نمایید.

ادامه فصل ۲٤ را در مرجع کاربردی SPSS 20 (38) دنبال نمایید.