رگرسیون

برای اندازه گیری رابطه بین دو یا چند متغیر، شاخص های متعددی وجود دارد که همه این شاخص ها میزان رابطه بین متغیرها را تنها با یک مقدار به نام ضریب همبستگی نشان می دهد. اما سکه رابطه دورو 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 می کند. از آنجا که ممکن است رابطه بین متغیر مستقل و وابسته را به هر صورت ممکن نوشت، مجموعه ای از روش ها 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 می دهد، معادله رگرسیون می گویند. اگر بتوان الگوی همبستگی را به صورت یک معادله خط نوشت، به آن معادله رگرسیون خطی می گویند و در غیر این صورت به آن معادله رگرسیون غیرخطی می گویند.

در رگرسیون هدف آن است که با استفاده از معادله رگرسیون و به کمک یک نمونه تصادفی و بعضی روش های آماری، رفتار متغیر وابسته را با آگاهی از مقادیر و مشخصات متغیرهای مستقل پیش بینی کنیم. در واقع به دنبال رابطه ای به فرم زیر هستیم:

$$Y = f(X_1, X_2, \dots, X_k)$$

به Y متغیر پاسخ Response می گویند که می تواند یک متغیر کمی یا پیوسته باشد.

به $X_1,X_2,\dots,X_k$ متغیر های مستقل یا متغیر های پیش بینی Predictor گفته می شود که می توانند کمی یا کیفی باشند.

ما بدنبال رابطه خطی Y با متغیرهای مستقل $X_1,X_2,\dots,X_k$ هستیم، یعنی:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

$\beta_0, \beta_1, \dots, \beta_k$ ضرایب مدل رگرسیونی هستند که پارامترهای نامعلومی هستند و باید برآورد شوند.

مثال: اگر بخواهیم میزان فشار خون را از روی میزان نمک خون و چربی خون پیش بینی کنیم باید $\beta_0, \beta_1, \beta_2$ را برآورد کنیم:

میزان فشار خون =Y　　میزان نمک خون=X_1　　میزان چربی خون=X_2

$$Y=\beta_0+\beta_1X_1+\beta_2X_2$$

رگرسیون خطی ساده

رگرسیون خطی، ممکن است ساده یا چندگانه 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 به طور 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 توان 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 باقیمانده (Residuals) می گویند.

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

$$Y'=b_0+b_1x$$

در این معادله Y' مقدار برآورد شده ، b_1شیب خط رگرسیونی یا ضریب رگرسیون وb_0 را عرض از مبدا 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:

$$Y'=b_0+b_1x_1+b_2x_2+\dots+b_kx_k$$

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

چرا می گوییم مدل رگرسیون خطی؟

منظور از مدل رگرسیون خطی این است که مدل نسبت به پارامترها خطی باشد. بعبارتی ضرایب مدل نباید بصورت توانی یا توابعی از یکدیگر باشند.

مثال: مدل های زیر خطی اند:

$$Y = \beta_0 + \beta_1 X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 \log(X) + \varepsilon$$

$$Y = \beta_1 e^X + \varepsilon$$

مثال: مدلهای زیر خطی نیستند:

$$Y = \beta_0 + \beta_0^2 X + \varepsilon$$

$$Y = \beta_0 . \beta_1 e^X + \varepsilon$$

نکته: مدل هایی که قابل تبدیل به یک مدل خطی باشند، ذاتا خطی گویند.

مثال: مدل های ذاتا خطی:

$$Y = \beta_0 + \log(\beta_0) X + \varepsilon \rightarrow (\beta_2 := \log(\beta_1)) \rightarrow Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 e^{\beta_1 X} + \varepsilon \rightarrow \log(Y) = \log(\beta_0) + \beta_1 X + \varepsilon \rightarrow (\beta_2 := \log(\beta_0)) \rightarrow \log(Y) = \beta_2 + \beta_1 X + \varepsilon$$

شرایط رگرسیون خطی ساده

در رگرسیون خطی ساده باید شرایط زیر برقرار باشد:

۱- میانگین (امیدریاضی) خطاها صفر باشد. یعنی $E(\varepsilon_i) = 0$

۲- واریانس خطا ها ثابت باشد ، به عبارت دیگر $var(\varepsilon_i) = \sigma^2$

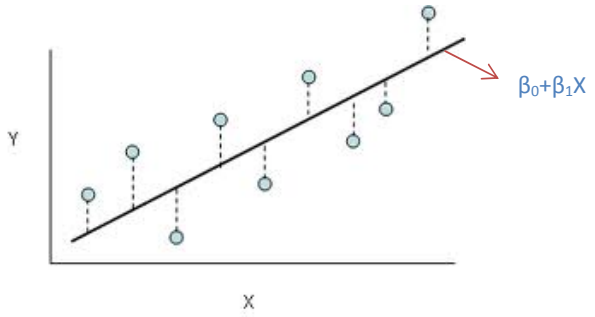
اگر فرض های ۱ و ۲ برقرار باشند به این معنا است که توزیع خطاها نرمال است.

۳- بین خطاها همبستگی وجود نداشته باشد. به عبارت دیگر $cov(\varepsilon_i, \varepsilon_j) = 0$

۴- متغیر وابسته داراری توزیع نرمال باشد.

۵- متغیر های مستقل دارای هم خطی نباشند. یعنی بین متغیرهای مستقل همبستگی معناداری وجود نداشته باشد.

روش کمترین مربعات خطا (Least Square Methods)



شکل ۱

در این روش β_0 و β_1 را طوری پیدا می کنیم که مجموع توان دوم خطاهای ε_i کمینه شود:

$$S(\beta_0, \beta_1) = \sum_{i=1,..,n} (y_i - \beta_0 - \beta_1 x_i)^2$$

پس β_0 و β_1 را طوری پیدا می کنیم که Sکمینه گردد:

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} | \beta_0 = \hat{\beta_0}, \beta_1 = \hat{\beta_1} = 0 \rightarrow \sum_{i=1}^{n} -2(y_i - \beta_0 - \beta_1 x_i)^2 = 0$$

$$\rightarrow \sum_{i=1}^{n} y_i - n\hat{\beta_0} - \hat{\beta_1} \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} | \beta_1 = \hat{\beta_1}, \beta_0 = \hat{\beta_0} = 0 \rightarrow \sum_{i=1}^{n} -2x_i (y_i - \hat{\beta_0} - \hat{\beta_1} x_i) = 0$$

$$\rightarrow \sum_{i=1}^{n} x_i y_i - \hat{\beta_0} \sum_{i=1}^{n} x_i - \hat{\beta_1} \sum_{i=1}^{n} x_i^2 = 0 \rightarrow \sum_{i=1}^{n} x_i y_i - (\bar{y} - \hat{\beta_1}\bar{x}) \sum_{i=1}^{n} x_i - \hat{\beta_1} \sum_{i=1}^{n} x_i^2 = 0 \rightarrow \hat{\beta_1}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - x)^2} = \frac{S_{xy}}{S_{xx}}$$

خواص برآوردگر های کمترین مربعات

۱) برآوردگرهای کمترین مربعات β_0 و β_1 ترکیب خطی از مشاهدات y_i هستند:

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})y_i}{S_{xx}} = \sum_{i=1}^{n} a_i y_i \qquad , a_i = \frac{(x_i-\bar{x})}{S_{xx}}$$

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x} = \frac{1}{n}\sum_{i=1}^{n} y_i - \bar{x}\sum_{i=1}^{n} a_i y_i = \sum_{i=1}^{n}(\frac{1}{n} - \bar{x}\,a_i)y_i = \sum_{i=1}^{n} b_i y_i$$

۲) این برآوردگرها ناریب هستند:

$$E(\widehat{\beta_1}) = E\left(\sum_{i=1}^{n} a_i y_i\right) = \sum_{i=1}^{n} a_i E(y_i) = \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 x_i)$$
$$= \beta_0 \sum_{i=1}^{n} \frac{x_i-\bar{x}}{S_{xx}} + \beta_1 \sum_{i=1}^{n} \frac{x_i-\bar{x}}{S_{xx}} x_i = 0 + \beta_1 = \beta_1$$

$$E(\widehat{\beta_0}) = E\left(\sum_{i=1}^{n} b_i y_i\right) = \sum_{i=1}^{n} b_i E(y_i) = \sum_{i=1}^{n} b_i(\beta_0 + \beta_1 x_i)$$
$$= \beta_0 \sum_{i=1}^{n} \frac{1}{n} - \bar{x}\frac{x_i-\bar{x}}{S_{xx}} + \beta_1 \sum_{i=1}^{n} (\frac{1}{n} - \bar{x}\frac{x_i-\bar{x}}{S_{xx}})x_i = \beta_0 + 0 = \beta_0$$

۳) واریانس برآوردگرها برابرست با:

$$var(\widehat{\beta_1}) = var\left(\sum_{i=1}^{n} a_i y_i\right) = \sum_{i=1}^{n} a_i^2 var(y_i) + 2\sum_{i<j=1} \sum cov(a_i y_i, a_j y_j) = \sum_{i=1}^{n} a_i^2 var(y_i)$$
$$= \sigma^2 \sum_{i=1}^{n} \frac{(x_i-\bar{x}^2)}{S_{xx}^2} = \sigma^2 \frac{S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

$$var(\widehat{\beta_0}) = var(\bar{y} - \widehat{\beta_1}\bar{x}) = var\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + var(\widehat{\beta_1}\bar{x}) - 2cov(\bar{y}, \widehat{\beta_1}\bar{x}) = \frac{\sigma^2}{n} + \bar{x}^2\frac{\sigma^2}{S_{xx}} + 0$$
$$= \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})$$

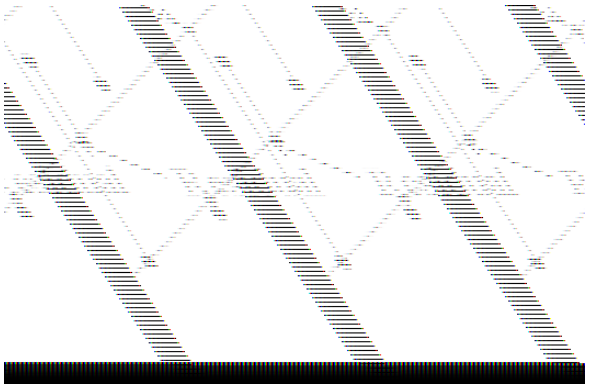
۴) بنا به قضیه گاوس- کارکوف در مدل $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ تحت سه فرض

$$E(\varepsilon_i) = 0 \qquad , var(\varepsilon_i) = \sigma^2 \qquad , cov(\varepsilon_i, \varepsilon_j) = 0 \qquad i \neq j$$

در بین برآوردگر های ناریب β_0 و β_1 که بصورت ترکیب خطی از y_i ها هستند، برآوردگرهای کمترین مربعات β_0 و β_1 کمترین واریانس را دارند یعنی کاراترین هستند .

باقیمانده ها(Residuals)

اگر در معادله خط رگرسیون که بر اساس داده های مطالعه آن را به عنوان مناسب ترین خط برازش شده بدست آورده اید، مقادیر مختلف متغیرx را قرار دهید، برای متغیر وابستهy مقادیری به دست خواهید آورد که به اندازه های مشاهده شده متفاوت خواهند بود. به عبارتی نقاط مشاهده شده بر نقاط برآوردشده آن یعنی Y' منطبق نخواهد 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 مدلهای رگرسیون از اهمیت ویژه ای برخوردار است زیرا از آنها می توان به عنوان شاخصی برای صحت برآورد معادله خط رگرسیون استفاده کرد. در واقع اگر فرضیات مدل رگرسیون برقرار نباشد، با استفاده از باقیمانده ها (e_i) می توانیم نسبت به برقرار نبودن آنها اطمینان حاصل کنیم.

خواص باقیمانده ها:

$$\sum_{i=1,\dots,n} e_i = 0$$

$$\sum_{i=1,\dots,n} e_i x_i = 0$$

$$\sum_{i=1,\dots,n} e_i y_i = 0$$

برآورد σ^2 :

۱) برای برآورد σ^2 کافی است برای حداقل یک سطح x چند مشاهده برای y داشته باشیم. در اینصورت واریانس آنها برآوردی برای σ^2 است.

۲) واریانس باقیمانده ها برآوردی برای σ^2 است.

توجه:

مجموع مربعات خطا Error sum of square: SSE

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

میانگین مربعات Mean Square: MS

$$MS = \frac{SS}{d.f(SS)}$$

میانگین مربعات خطا Error Mean Square: MSE

$$MSE = \frac{SSE}{d.f(SSE)} \qquad MSE = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

آزمون فرضیه راجع به β_0وβ_1

در صورتی که فرض H_1 قبول شود، بدین معنی است که دست کم یکی از bها در پیش بینی y موثر است. برای مشخص کردن آنها، باید تک تک وجود ضرایب b_jها را آزمود، این آزمون به کمک ملاک t استودنت صورت می گیرد، بنابراین هر یک از ضرایب معادله ی رگرسیون را می توان از فرمول $t_j = \frac{b_j}{\sqrt{V_{jj}}}$ مورد آزمون قرار داد. البته t_j دارای توزیع t استودنت با درجه آزادی $(n-m-1)$ میباشد، که درآن V_{jj} واریانس متغیر x_j است که در سطر j ام در ستون j ام ماتریس واریانس و کواریانس است، و باید در محاسبه ی t_j مقادیر b_jها را مانند d در صورت t استودنت به صورت قدرمطلق در نظر گرفت. پس برای آزمون هر یک از b_j ها کافی است که t_j را حساب کرده و با مقایسه ی t جدول حساب کرد. پس از آزمون b_j ها هر x_j را که b_j متناظر با آن معنی دار نشده است حذف می کنیم ؛ به سخن دیگر هر کدام از t ها معنی دار نشدند، نتیجه می گیریم که b های آنها تقریبا صفر بوده و x های آنها در برآورد y موثر نیستند. فقط کافی است که معادله ی رگرسیون را بر اساس b های مخالف صفر محاسبه نماییم. در واقع، b های معنی دار نشان می دهند که متغیر آنها با y همبستگی دارند، و در پیش بینی آن موثر می باشند.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad ,i = 1, \dots, n$$

فرض می کنیم y_i ها توزیع نرمال داشته باشند بنابراین:

$$if \quad \varepsilon_i \sim N(0, \sigma^2) \rightarrow y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i y_i \sim N(\beta_1, \frac{\sigma^2}{S_{xx}})$$

$$\hat{\beta}_0 = \sum_{i=1}^{n} b_i y_i \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}))$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x}^2)^2}{S_{xx}}\right))$$

فرض کنید بخواهیم آزمون زیر را انجام دهیم:

$$H_0: \beta_1 = \beta_1 *$$

$$H_1: \beta_1 \neq \beta_1 *$$

$$Z_0 = \frac{\hat{\beta}_1 - \hat{\beta}_1^*}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

فرض اولیه می شود $\rightarrow |Z_0| > Z_{\alpha/2}$

$$C.I(\beta_1) = \hat{\beta}_1 \pm Z_{u/2}\sqrt{\frac{\sigma^2}{S_{xx}}}$$

حال فرض کنید بخواهیم آزمون زیر را انجام بدهیم

$$H_0: \beta_0 = \beta_0 *$$

$$H_1: \beta_0 \neq \beta_0 *$$

$$Z_0 = \frac{\hat{\beta}_0 - \hat{\beta}_0^*}{\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}} \sim N(0,1)$$

فرض اولیه می شود $\rightarrow |Z_0| > Z_{\alpha/2}$

$$C.I(\beta_0) = \hat{\beta}_0 \pm Z_{\alpha/2}.\sigma\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

آزمون فرضیه و فاصله اطمینان برای ضرائب رگرسیونی در حالتی که σ^2 نامعلوم است

فرض کنید بخواهیم آزمون زیر را انجام دهیم:

$$H_0: \beta_1 = \beta_1^* \qquad vs \qquad H_1: \beta_1 \neq \beta_1^*$$

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \rightarrow \frac{(n-1)MSE}{\sigma^2} \sim \chi^2_{n-1} \overset{H_0}{\Rightarrow} \frac{\hat{\beta}_1 - \beta_1^*}{\frac{\sigma}{\sqrt{S_{xx}}}} \sim N(0,1) \rightarrow t_0 = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\frac{MSE}{S_{xx}}}} \sim t_{n-2}$$

$$|t_0| > t_{\alpha/2,n-2}$$

$$C.I(\beta_1) = \hat{\beta}_1 \pm t_{\alpha/2,n-2}\sqrt{\frac{MSE}{S_{xx}}}$$

حالت خاص آزمون: همیشه این آزمون را انجام می دهیم (آزمون معناداری رگرسیون)

$$H_0: \beta_1 = 0 \qquad vs \qquad H_1: \beta_1 \neq 0$$

$$y = \beta_0 + \beta_1 x + \varepsilon_i$$

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\frac{MSE}{S_{xx}}}}$$

اگر فرض H_0 رد نشود در اینصورت دو حالت دارد:

۱. x و y هیچ همبستگی ندارند.

۲. x و y رابطه غیرخطی دارند.

اگر فرض H_0 رد شود در اینصورت دو حالت دارد:

۱. x و y رابطه خطی ندارند.

۲. x و y رابطه غیرخطی(نسبت به x) دارند.

فرض کنید بخواهیم آزمون زیر را انجام دهیم:

$$H_0: \beta_0 = \beta_0^* \qquad vs \qquad H_1: \beta_0 \neq \beta_0^*$$

$$t_0 = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \sim t_{n-2}$$

$$|t_0| > t_{\alpha/2,n-2}$$

$$C.I(\beta_0) = \hat{\beta}_0 \pm t_{\alpha/2,n-2}\sqrt{MSE(\frac{1}{S_{xx}} + \frac{\bar{x}^2}{S_{xx}})}$$

حالت خاص آزمون به شکل زیر است:

$$H_0: \beta_0 = 0 \qquad vs \qquad H_1: \beta_0 \neq 0$$

$$t_0 = \frac{\hat{\beta}_0}{\sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}})}} = \frac{\hat{\beta}_0}{S.e.(\hat{\beta}_0)}$$

اگر فرض H_0 رد نشود در این صورت دو حالت دارد:

$\beta_0 = 0$.۱

۲. x و y رابطه غیر خطی دارند.

اگر فرض H_0 رد شود در این صورت دو حالت دارد:

$\beta_0 \neq 0$.۱

۲. x و yرابطه غیر خطی دارند.

تخمین (پیش بینی) مشاهدات جدید

می دانیم:

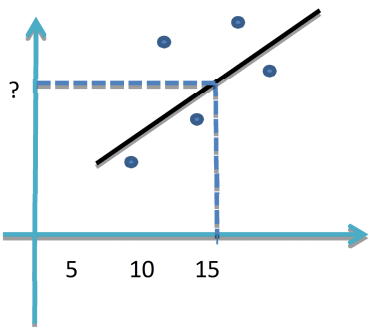
$$\hat{E}(Y|X = x_0) = \hat{Y}_{x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

حال فرض کنیم بخواهیم بر اساس خط رگرسیون مقدار متغیر پاسخ Yرا به ازای $X = x^*$ تخمین بزنیم؛ یا برای حالتی که هیچ اطلاعی راجع به متغیر پاسخ نداریم آنرا تخمین بزنیم.

مثال:

میزان محصول = Y

میزان کود = X



X=15=x* ; $\hat{Y}_{x_0} = ?$

فرض کنیم $Y_{x*}\sim N(\beta_0+\beta_1 x_i,\sigma^2)$ از $Y_1,Y_2,\dots,Y_n\sim N(\beta_0+\beta_1 x_i,\sigma^2)$ مستقل باشد. حال می خواهیم برآورد نقطه ای و فاصله ای به ازای $X=x^*$ برای مشاهده جدید Y_x بدست آوریم.

برآورد نقطه ای Y_x به صورت زیر است:

$$\hat{Y}_{x*}=\hat{\beta}_0+\hat{\beta}_1 x^*$$

Y_x از $Y_1,Y_2,\dots,Y_n$ مستقل است بنابراین از هر ترکیب خطی شان نیز مستقل است. از طرفی برآوردگر های β_0وβ_1 هردو ترکیب خطی از Y_i ها هستند. بنابراین از برآورد Y_x نیز مستقل است:

$$E\left(Y_{x*}-\hat{Y}_{x*}\right)=0 \qquad (1)$$

$$var\left(Y_{x*}-\hat{Y}_{x*}\right)=var(Y_{x*})+var\left(\hat{Y}_{x*}\right)-2cov\left(Y_{x*},\hat{Y}_{x*}\right)=$$

$$\sigma^2+\sigma^2\left(\frac{1}{n}+\frac{(x^*-\bar{x})}{S_{xx}}\right)+0=\sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})}{S_{xx}}\right) \qquad (2)$$

$$\xrightarrow{(1),(2)} (Y_{x*}-\hat{Y}_{x*})\sim N\left(0,\sigma^2\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})}{S_{xx}}\right)\right.$$

$$\frac{Y_{x*}-\hat{Y}_{x*}}{\sigma\sqrt{\left(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}\right)}}\sim N(0,1) \qquad (3)$$

$$\frac{(n-2)MSE}{\sigma^2}\sim \chi^2_{n-2} \qquad (4)$$

$$\xrightarrow{(3),(4)} \frac{Y_{x*}-\hat{Y}_{x*}}{\sqrt{MSE(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}}\sim t_{n-2}$$

$$P\left(-t_{\frac{\alpha}{2},n-2}\leq \frac{Y_{x*}-\hat{Y}_{x*}}{\sqrt{MSE(1+\frac{1}{n}+\frac{(x^*-\bar{x})^2}{S_{xx}}}}\leq t_{\frac{\alpha}{2},n-2}\right)=1-\alpha/2$$

طول این فاصله اطمینان بزرگتر از حالتی است که برای برآورد میانگین پاسخ ها بدست می آوریم. چون واریانس آن بزرگتر است.

جدول آنالیز واریانس

جدول آنالیز واریانس یا جدولANOVA برای بررسی و تجزیه و تحلیل واریانس یا واریانس Yکل بکار می رود. همچنین برای آزمون معنی داری رگرسیون نیز مورد استفاده قرار می گیرد.

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) =$$

$$= SSE + SSR + 0 = SSE + SSR$$

$$Total\ sum\ of\ squares: SST = S_{yy}$$

$$Error\ sum\ of\ squares: SSE$$

$$\text{Regression sum of squares: SSR}$$

واریانس کل = تغییراتی که به وسیله عامل xبیان می شود + تغییراتی که بوسیله x بیان نمی شود

ثابت می شود که SSR و SSE از هم مستقل اند پس:

$$SSR = SSE + SSR$$
$$d.f(SST) = d.f(SSE) + d.f(SSR)$$
$$(n-1) = (n-2) + (1)$$
$$MSR = \frac{SSR}{1} \quad , \quad MSE = \frac{SSE}{n-2}$$

تحت فرض H_0 آزمون معنی داری رگرسیون داریم:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$
$$\frac{MSR}{\sigma^2} \sim \chi^2_1 \qquad *$$
$$\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2} \qquad **$$
$$* \ \& \ ** \rightarrow F_0 = \frac{MSR}{MSE} \sim F_{1,N-2}$$

فرضH_0 رد می شود اگر $F_0 \geq F_{1,N-2,\alpha}$ باشد.

ANOVA:

Source of variance	SS	d.f	MS	F_0
Regression	$\hat{\beta}_1 S_{xy}$	1	MSR	$\frac{MSR}{MSE}$
Error	$S_{yy} - \hat{\beta}_1 S_{xy}$	n-2	MSE	
Total	S_{yy}	n-1		

آماره جدول برای آزمون معنی داری رگرسیون بکار می رود.

کوواریانس و ضریب همبستگی

با مطالعه حالت ساده ی رابطه بین یک متغیر پاسخ Y و یک متغیر پیشگوی X به طرح موضوع می پردازیم . کوواریانس و ضریب همبستگی را به عنوان اندازه های جهت و قوت رابطه ی خطی بین دو متغیر مورد بحث قرار می دهیم . سپس الگوی رگرسیون خطی ساده را فرمول بندی کرده و نتایج ریاضی کلیدی را بدون محاسبات ریاضی ارائه نموده ولی آن را به وسیله ی مثالهای عددی تشریح می کنیم.

فرض کنید مشاهداتی از ۸ واحد شامل یک متغیر وابسته یا پاسخ Y و یک متغیر توضیحی X داریم. مشاهدات معمولا به صورت جدول ۱ ثبت می شوند. ما می خواهیم جهت و قوت رابطه بین X و Y را اندازه گیری کنیم . این دو اندازه را کوواریانس و ضریب همبستگی می نامیم که در زیر توسعه داده می شوند.

جدول ۱. نماد های مربوط به داده هایی که در رگرسیون ساده و همبستگی به کار می روند.

شماره مشاهده	پاسخ	پیشگو
	Y	X
1	y_1	x_1
2	y_2	x_2
…	…	…
N	y_n	x_n

در نمودار پراکنش Y در مقابل X ، یک خط عمودی در $\bar{x}$ و یک خط افقی در $\bar{y}$ به طوری که در شکل ۱ نشان داده شده رسم می کنیم که در آن $\bar{X}=\frac{\sum_{i=1}^{n} x_i}{n}$ و $\bar{Y}=\frac{\sum_{i=1}^{n} y_i}{n}$ به ترتیب میانگین نمونه ی X و Y هستند. این دو خط نمودار را به چهار مربع تقسیم می کنند. برای هر نقطه i در نمودار کمیّتهای زیر را محاسبه می کنیم:

- $y_i - \bar{y}$، انحراف هر مشاهده y_i از میانگین متغیر پاسخ
- $x_i - \bar{x}$، انحراف هر مشاهده x_i از میانگین متغیر پیشگو
- $(x_i - \bar{x}).(y_i - \bar{y})$ حاصل ضرب دو کمیت بالا یعنی.

اگر رابطه خطی بین X و Y مثبت باشد (وقتی X زیاد می شود Y نیز زیاد شود) آن گاه در ربعهای اوّل و سوّم نقاط بیشتری نسبت به ربعهای دوّم و چهارم وجود دارد. برعکس اگر رابطه بین X و Y منفی باشد (با افزایش X، Y کاهش پیدا کند) آن گاه نقاط بیشتری در ربعهای دوّم و چهارم نسبت به ربعهای اوّل و سوّم وجود دارد. بنابراین علامت کمیت $\operatorname{cov}(y,x)=\frac{\sum_{i=1}^{n}(y_i-\bar{Y})(x_i-\bar{X})}{n-1}$ که کواریانس بین X و Y نامیده می شود جهت رابطه خطی بین X و Y را نشان می دهد. اگر

$cov(x,y) < 0$ در آن صورت این رابطه منفی خواهد بود. متأسفانه $cov(x,y)$ در مورد قوت وشدت این رابطه چیز زیادی در اختیار ما قرار نمی دهد زیرا تحت تأثیر تغییرات در واحدهای اندازه گیری قرار می گیرد. برای مثال اگر Y ویا X را به جای دلار را به جای دلار بر حسب هزار دلار گزارش کنیم دو مقدار مختلف به دست خواهیم آورد. برای رفع این عیب کوواریانس داده ها را قبل از محاسبه کوواریانس استاندارد می کنیم. برای استاندارد کردن داده های Y ابتدا میانگین را از هر مشاهده کم می کنیم و سپس آن را بر انحراف معیار تقسیم می کنیم یعنی $z_i = \frac{y_i - \bar{Y}}{s_y}$ را محاسبه می کنیم که در آن $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{Y})}{n-1}}$ انحراف معیار نمونه Y است. می توان نشان داد که متغیر استاندارد Z دارای میانگین صفر و انحراف معیار یک است. کوواریانس بین داده های استاندارد XوY را ضریب همبستگی بین XوY می نامند و به صورت زیر داده می شود.

$$\text{cov}(y,x) = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{y_i - \bar{Y}}{s_y})(\frac{x_i - \bar{X}}{s_x})$$

فرمولهای معادلی برای ضریب همبستگی عبارت اند از :

$$\text{co}\,r(y,x) = \frac{\text{cov}(y,x)}{s_y s_x} = \frac{\Sigma(y_i - \bar{Y})(x_i - \bar{X})}{\sqrt{\Sigma(y_i - \bar{Y})^2 \Sigma(x_i - \bar{X})^2}}$$

بدین ترتیب corr(Y,X) را به صورت کوواریانس بین مقادیر استاندارد شده یا نسبت کوواریانس به انحرافهای معیار دو متغیر می توان تفسیر کرد. از رابطه بالا می توان دید که ضریب همبستگی متقارن است یعنی corr(X,Y)=corr(Y,X).

بر خلاف Corr(Y,X), Cov(Y,X) نسبت به مقیاس پایاست یعنی اگر واحدهای اندازه گیری تغییر کند تغییر نمی کند. علاوه بر این Corr(Y,X) در نامساوی زیر صدق می کند.

$$-1 \le cor(y,x) \le 1$$

این خواص Corr(Y,X) آن را کمیتی مفید برای اندازه گیری جهت و قوت رابطه بین XوY می سازد. اندازه Cor(Y,X) قوت رابطه خطی بین XوY اندازه می گیرد. هر چه Cor(Y,X) به ۱ یا ۱- نزدیکتر باشد رابطه بین XوY قویتر خواهد بود. علامت Cor(Y,X) جهت رابطه خطی بین Yو X را مشخص می کند. یعنی $0 \succ$ Cor(Y,X) نتیجه می دهد که Yو X رابطه مثبتی با هم دارند. برعکس $\prec$ Cor(Y,X)۰ نتیجه می دهد که Yو X رابطه ای منفی دارند.

در عین حال توجه کنید که Cor(Y.X)=۰ الزاماً بدین معنی نیست که XوY رابطه ای ندارند. تنها نتیجه می شود که آنها رابطه ای خطی ندارند زیرا ضریب همبستگی فقط روابط خطی را اندازه می گیرد. به بیان دیگر Cor(Y,X) هنوز می تواند صفر باشد هرگاه X,Y رابطه ای غیرخطی داشته باشند. برای مثال X و Y در جدول ۲ یک رابطه غیر خطی دقیق $Y = 50 - X^2$ را دارند ضمن

اینکه $corr(x,y) = 0$. علاوه بر این مانند بسیاری از آماره های خلاصه شده دیگر Corr(Y,X) می تواند به طور اساسی تحت تأثیر یک یا چند نقطه دور افتاده در داده ها واقع شود.

جدول ۲ یک مجموعه داده که Cor(x,y)= ۰ ولی یک رابطه غیر خطی دقیق بین Yو X وجود دارد.

X	Y	X	Y	X	Y
۳-	41	-2	46	-7	1
4	34	-1	49	-6	14
5	25	0	50	-5	25
6	14	1	49	-4	34
7	1	2	46	-3	41

برای تأکید این نکته، انسکومبت (1973) چهار مجموعه داده ساخته است که به چهار تایی انسکومب معروف است و هر یک طرحی مجزا داشته ولی مجموعه آماره های خلاصه شده یکسان دارند (مثلاً مقدار ضریب همبستگی یکسان).داده ها و نمودار ها در جدول ۳ دوباره تولید شده اند .

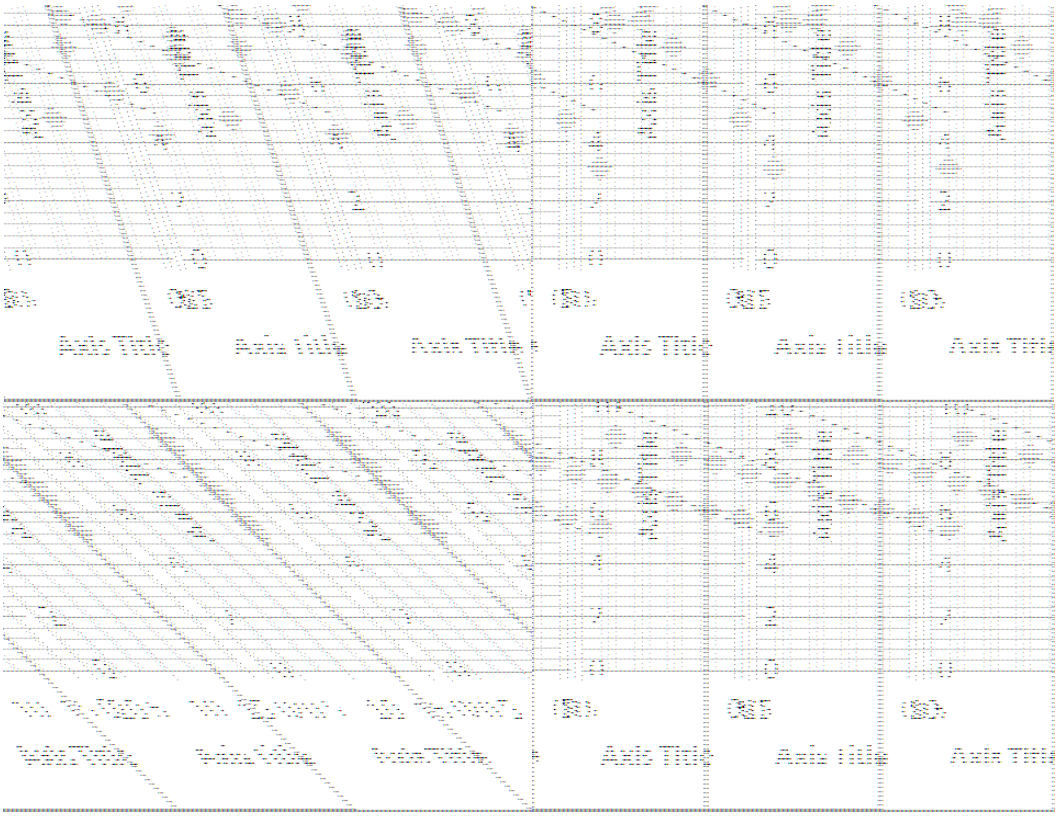
یک تحلیل مبتنی بر امتحان آماره های اختصاری ، مانند ضریب همیستگی نمی تواند اختلافهای موجود در طرح ها را کشف کند.

جدول ۳ چهارتایی انسکومبت: چهار مجموعه داده که مقادیر آماره های اختصاری یکسانی دارد.

X_4	Y_4	X_3	Y_3	X_2	Y_2	X_1	Y_1
8	6/58	10	7/46	10	9/14	10	**80/4**
8	5/76	8	6/77	8	8/14	8	**6/95**
8	7/71	13	12/74	13	8/74	13	**7/58**
8	8/84	9	7/11	9	8/77	9	**8/81**
8	8/47	11	7/81	11	9/26	11	**8/83**
8	7/04	14	8/84	14	8/10	14	**9/96**
8	5/25	6	6/08	6	6/13	6	**7/24**
19	12/50	4	5/39	4	3/10	4	**4/26**
8	5/56	12	8/15	12	9/13	12	**10/84**
8	7/91	7	6/42	7	7/26	7	**4/82**
8	6/89	5	5/73	5	4/74	5	**5/68**

با توجه به شکل این داده ها معلوم می شود که فقط اولین مجموعه ای که نمودار آن در (a)داده شده را با یک الگوی خطی می توان بیان کرد. نمودار (b) مجموعه داده دوّم که مشخصاً غیرخطی است را نشان می دهد که با یک تابع درجه دوّم بهتر برازش می شود . نمودار (c) نشان می دهد مجموعه داده سوم نقطه ای دارد که شیب و عرض از مبدأ خط برازش شده را تغییر می دهد. نمودار (d) نشان می دهد که چهارمین مجموعه داده برای برازش خطی مناسب نیست و خط برازش شده در اصل با یک مشاهده فرین تعیین می شود. بنابراین امتحان نمودار پراکندگی Y در مقابل X قبل از تفسیر مقدار عددی Cor(Y,X)دارای اهمیّت است.

(a) (b)



(d) (c)

نمایش هندسی ضریب همبستگی

ضریب همبستگی دو متغیر عبارت است از مقدار کوسینوس زاویه ای که دو متغیر تشکیل می دهند . پس همبستگی بین دو متغیر را می توان یه طور هندسی با دو بردار نمایش داد ، البته برداری خطی است دارای جهت و طول . اگر بردار طول واحدی داشته باشد ، آنگاه همان گونه که اشاره شد، ضریب همبستگی کوسینوس زاویه ی بین دو بردار است. همبستگی های بین چند متغیر را می توان برای هر چند بردار نمایش داد. پس همبستگی بین متغیر ها به وسیله ی زوایای برداری هر متغیر نشان داده می شود ، مثلاً اگر سه متغیر x_1 ، x_2 و yداشته باشیم ، چنانچه زاویه ی بین ox_1 و ox_2 ۳۰ درجه باشد ، آنگاه :

$$r_{12} = \cos(ox_1, ox_2) = \cos(30) = \frac{\sqrt{3}}{2} = 0/87 .$$

همچنین اگر زاویه ی بین ox_2و oy 30 درجه و زاویه ی oy و ox_1 ۶۰ درجه باشد ، در این صورت نیز r_{2y}و r_{1y}عبارتند از

$$r_{2y} = \cos(ox_2, oy) = \cos 30 = 0/87, r_{1y} \cos(ox_1, oy) = \cos 60 = \frac{1}{2} = 0/5 :$$

ضریب تعیین:

SST=SSE+SSR

تغییرات= میزان تغییراتی که توسط عامل X بیان میشود(SSR)+ میزان تغییراتی که توسط X بیان نمی شود(SSE)

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad ; \quad 0 \leq R^2 \leq 1$$

$R^2 = 0.98$ یعنی 98 درصد تغییرات Y توسط خط رگرسیون بیان می شود.

یکی از معیارهای بررسی خوبی برازش R^2 می باشد و هرچه به یک نزدیکتر باشد می تواند دلیل بر خوب بودن برازش باشد.

البته R^2 وقتی معتبر است که رابطه خطی باشد.

اندازه گیری و مقیاس سازی

فرض کنید α یکی از افراد جمعیت π و t یک ویژگی مورد مطالعه می باشد. مثلا سیب های یک باغ را جمعیت π و α را یکی از سیب ها در نظر بگیرید. ویژگی t برای α ممکن است وزن سیب باشد، که امری است کمی، یا تردی سیب باشد که امری است کیفی. منظور از اندازه گیری ویژگی t در α، اطلاق یک عدد حقیقی x، به α طبق ضابطه ای مشخص می باشد. به سخنی دیگر x تابعی از α به صورت $x = f(\alpha)$ است. این تابع یا x را مقیاس و طرز تعیین عدد x را اندازه گیری یا مقیاس سازی می گویند. اگر t وزن سیب باشد می توان عدد x را با مقیاس وزن مثلاً بر حسب گرم با تقریب لازم پیدا کرد. ولی اگر t تردی سیب باشد، مقیاسی شناخته شده برای اندازگیری و پیدا کردن x در دست نیست.

با این حال سیب شناسان می توانند منظور از تردی سیب را تریف نمایند و مقیاسی برای تردی سیب معرفی کنند. باید توجه داشت که نوع عددی که نماینده وزن است با عددی که نماینده تردی است با هم تفائت دارند . مثلاً می توان گفت که وزن یک سیب دو برابر وزن سیب دیگر است ولی این بیان در مورد تردی مفهومی ندارد.

استیونز [1] استاد روانشناسی دانشگاه هاروارد آمریکا، در مقاله بنیادی ۱۹۴۶ خود چهار نوع مقیاس معرفی کرده است. این مقیاس ها در حدود نیم قرن است که در بعضی از کتابهای آمار بیان می شوند و با این که در عصر آمار و کامپیوتر بارها مورد انتقاد قرار گرفته اند، هنوز دارای اهمیت هستند. برای اطلاع بیشتر در این باره به کتابنامه رجوع کنید. اینک مقیاس های استیونز را به کوتاهی شرح می دهیم.

مقیاس های استیونز

الف: مقیاس اسمی: هرگاه مقیاس x، که معمولاً یک عدد طبیعی است، تنها برای شناسایی افراد یا چیزها یا مکانها به کار می رود، آن را یک مقیاس اسمی می نامند. مثلاً اگر کارگران یک کارخانه از شهرهای تهران، اصفهان، شیراز و کرمان باشند و

[1] S.s. stevens

به ترتیب آنها را با اعداد ۱و۲و۳و۴ مشخص کنیم، این اعداد صرفاً می گویند که هر کارگر از کدام شهر است. کارگری که برچسب ۳ دارد از شیراز و کارگری که برچسب ۴ دارد از کرمان است. توجه کنید که این عداد را نمی توان برای مقایسه یا چهار عمل اصلی حساب به کار برد.

مقیاس اسمی x را با هر تابع یک به یک می توان به مقیاس اسمی y تبدیل کرد بی آنکه شناسایی تغییر نماید. مثلاً تابع یک به یکِ $y = x + 10$ برچسب های ۱،۲،۳،۴ را به ترتیب بر برچسب های ۱۱،۱۲،۱۳،۱۴ تبدیل می کند.

ب: مقیاس ترتیبی : هر گاه مقیاس x ، که یک عدد حقیقی است، برتری را بیان کند، آن را یک مقیاس ترتیبی می نامند. به سخنی دیگر اگر $\alpha_۴$ از نظر ویژگی t بر $\alpha_۱$ برتری داشته باشد و ویژگی t را برای $\alpha_۱$ و $\alpha_۴$ با اعداد $x_۱$ و $x_۴$ نشان دهیم باید داشته باشیم $x_2 > x_1$. مثلاً اگر مهندس یک کارخانه کارگران را از نظر مهارت با اعداد ۱ و ۲و ۳و ۴ مشخص کند، کارگر شماره ۴ از کارگر شماره ۲ ماهرتر است. ولی نمی توان گفت که دو برابر او مهارت دارد ؛ این گونه اعداد را می توان تنها برای مقایسه به کار برد و نمی توان با آنها چهار عمل اصلی انجام داد.

مقیاس ترتیبی x را می توان با یک تابع صعودی به مقیاس ترتیبی y تبدیل کرد، بی آن که برتری یا ترتیب مختل گردد. مثلاً تابع $y = x^3$ اعداد ۱، ۲، ۳ و ۴ را به ۱، ۸، ۲۷ و ۶۴ تبدیل می کند و با اعداد اخیر هم کارگر چهارم از کارگر دوم ماهر تر است.

ج : مقیاس فاصله ای : هر گاه مقیاس x ، که یک عدد حقیقی است، نسبت دو تفاضل یا دو فاصله را حفظ کند، آن را مقیاس فاصله ای می نامند. به سخنی دیگر اگر با این مقیاس برای $\alpha_۱$ و $\alpha_۲$ و $\alpha_۳$ و $\alpha_۴$ ویژگی t با اعداد $x_۱$ و $x_۲$ و $x_۳$ و $x_۴$ اندازه گیری شود، باید نسبت $\frac{(x_۴ - x_۳)}{(x_۲ - x_۱)}$ ثابت بماند و به واحد های اندازه گیری بستگی نداشته باشد. مثلاً فرض کنید ویژگی t گرمای ۴ جسم باشد و با مقیاس سانتی گراید داشته باشیم :

$$x_۱ = ۱۰ \quad و \quad x_۲ = ۱۵ \quad و \quad x_۳ = ۲۰ \quad و \quad x_۴ = ۴۵$$

ملاحضه می شود که :

$$\frac{(x_۴ - x_۳)}{(x_۲ - x_۱)} = ۵$$

اینک با مقیاس فارنهایت طبق فرمول $y = \frac{۹x}{۵} + ۳۲$ داریم :

$$y_۱ = ۵۰ \quad و \quad y_۲ = ۵۹ \quad و \quad y_۳ = ۶۸ \quad و \quad y_۴ = ۱۱۳$$

باز هم ملاحظه می شود که :

$$\frac{(y_4 - y_3)}{(y_3 - y_2)} = 5$$

در مثال بالا نکته های زیر قابل توجه می باشد:

۱. اگر با مقیاس سانتی گراد جسم α دارای درجه حرارت صفر باشد، با مقیاس فارنهایت این جسم دارای درجه حرارت ۳۲ است. بنابراین با مقیاس فاصله ای صفر معنی هیچ نمی دهد و صرفاً جنبه قراردادی دارد.

۲. با مقیاس اول گرمای α_4 سه برابر گرمای α_3 است ، ولی با مقیاس دوم چنین نیست . بنابراین در اندازه گیری با مقیاس فاصله ای، نسبت محفوظ نمی ماند.

۳. با هر دو مقیاس تفاضل گرمای α_3 از α_4، پنج برابر تفاضل گرمای α_2 از α_1 است. بنابر این با مقیاس فاصله ای، نسبت فاصله ها با تغییر مقیاس فاصله ای تغییر نمی کند.

مقیاس فاصله ای x را می توان با تابع خط $y = ax + b$ با فرض $\alpha > 0$، به مقیاس فاصله ای y تبدیل کرد بی آنکه نسبت دو تفاضل تغییر کند.

مقیاس فاصله ای معمولاً در روانشناسی، تعلیم و تربیت، جامعه شناسی و فیزیک به ویژه در اموری مانند هوش، حافظه، ارزیابی آزمونها، حرارت و ... به کار می رود. این نوع مقیاس سازی به مهارت و تجربه فراوان در رشته مورد بحث بستگی دارد و گاهی دستخوش مجادله می باشد.

د : مقیاس نسبتی : هرگاه مقیاس x که یک عدد حقیقی است، نسبت را حفظ کند، آن را یک مقیاس نسبتی می نامند. به سخنی دیگر اگر با این مقیاس برای دو جسم α_1 و α_2 ویژگی t، مثلاً وزن با x_1 و x_2 اندازه گیری شود، باید $\frac{x_1}{x_2}$ ثابت بماند و به واحد اندازه گیری بستگی نداشته باشد . اگر وزن α_1 و α_2 با مقیاس گرم ۶۰۰۰ و ۲۰۰۰ باشد، با مقیاس کیلوگرم ۶ و ۲ است. ملاحظه می شود که با هر دو مقیاس وزن α_1 سه برابر وزن α_2 است.

مقیاس نسبتی برای امور کمی مانند وزن، طول، سطح، مقدار حرارت به کار می رود و عالی ترین نوع مقیاس است. با اعداد حاصل از این مقیاس می توان چهار عمل اصلی حساب را انجام داد. مقیاس نسبتی x را می توان با تابع خط $y = ax$ با فرض $\alpha > 0$ به مقیاس نسبتی y تبدیل کرد، بی آنکه نسبت تغییر کند.

متغیرهای نشانگر برای حذف فصلی بودن

با به کار بردن یک متغیر فصلی دیگر الگو را به صورت زیر توسعه داده می شود:

$$S_t = \beta_0 + \beta_1 PDI_t + \beta_2 Z_t + \varepsilon_t \qquad (*)$$

که در آن Z_t متغیر صفر - یک است که ذیلا شرح داده شده و β_2 پارامتری است که اثر فصلی را اندازه گیری می کند. توجه کنید که الگوی (*) را می توان با دو الگو (یکی برای فصولی که هوا سرد است ، $Z_t = 1$ و دیگری برای فصول گرم $Z_t = 0$) ارائه کرد:

فصل زمستان : $S_t = (\beta_0 + \beta_2) + \beta_1 PDI_t + \varepsilon_t$

فصل تابستان : $S_t = \beta_0 + \beta_1 PDI_t + \varepsilon_t$

بدین ترتیب این الگو ها فرض هایی را نشان می دهند که فروش را می توان با یک تابع خطی PDI تقریب نمود، یک خط برای فصل زمستان و یک خط برای فصل تابستان. این دو خط موازی اند یعنی اثر جانبی تغییرات در PDI در هر دو فصل یکسان است. سطح فروش که با عرض از مبدا منعکس می شد در هر فصل فرق می کند .

فصلی بودن :

مجموعه داده ای که در این جا به عنوان یک مثال از آن استفاده می شود به داده های فروش اسکی منسوب است که از کتاب وب سایت می توان به دست آورد . این داده ها دو متغیر را شامل می شوند : فروش S به میلیون برای کارخانه ای که اسکی ها و ابزار وابسته به آن را برای سالهای ۱۹۷۳-۱۹۶۴ تولید می کند و درآمدهای مصرف شخصی (اندازه جمعی پتانسیل خرید) PDI . هر یک از این متغیر ها به طور فصلی اندازه گیری می شود . این الگو معادله ای است که S را به PDI مربوط می سازد.

$$S_t = \beta_0 + \beta_1 . PDI_t + \varepsilon_t$$

که در آن S_t فروش به میلیون در دوره tام بوده و PDI_t درآمد مصرف شخصی مربوطه است. رویکرد ما در اینجا این فرض است که یک اثر فصلی روی فروش که بر مبنای فصلی تعیین می شود وجود دارد . برای اندازه گیری این اثر می توانیم متغیرهای نشانگر را جهت مشخص کردن فصلی بودن تعریف کنیم . چون چهار فصل داریم لذا سه متغیر نشانگر Z_1 و Z_2 و Z_3 را تعریف می کنیم که :

$$Z_{t1} = \begin{cases} 1 & \text{اگر دوره } t\text{ام اولین فصل باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

$$Z_{t2} = \begin{cases} 1 & \text{اگر دوره } t\text{ام دومین فصل باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

$$Z_{t3} = \begin{cases} 1 & \text{اگر دوره } t\text{ام سومین فصل باشد} \\ 0 & \text{در غیر این صورت} \end{cases}$$

تحلیل و تفسیر این مجموعه داده ها را به خوانندگان واگذار می کنیم. مولف این داده ها را تحلیل نموده و متوجه می شود که در واقع تنها دو فصل وجود دارد. منتا (۱۹۸۶) را برای بحث بیشتر استفاده از متغیر نشانگر برای تحلیل فصلی بودن ملاحظه کنید.

ردیف	تاریخ	فروش	PDI	ردیف	تاریخ	فروش	PDI
۱	$Q_1/۶۴$	۳۷	۱۰۹	۲۱	$Q_1/۶۹$	۴۴/۹	۱۵۳
۲	$Q_2/۶۴$	۳۳/۵	۱۱۵	۲۲	$Q_2/۶۹$	۴۱/۶	۱۵۶
۳	$Q_3/۶۴$	۳۰/۸	۱۱۳	۲۳	$Q_3/۶۹$	۴۴	۱۶۰
۴	$Q_4/۶۴$	۳۷/۹	۱۱۶	۲۴	$Q_4/۶۹$	۴۸/۱	۱۶۳
۵	$Q_1/۶۵$	۳۷/۴	۱۱۸	۲۵	$Q_1/۷۰$	۴۹/۷	۱۶۶
۶	$Q_2/۶۵$	۳۱/۶	۱۲۰	۲۶	$Q_2/۷۰$	۴۳/۹	۱۷۱
۷	$Q_3/۶۵$	۳۴	۱۲۲	۲۷	$Q_3/۷۰$	۴۱/۶	۱۷۴
۸	$Q_4/۶۵$	۳۸/۱	۱۲۴	۲۸	$Q_4/۷۰$	۵۱	۱۷۵
۹	$Q_1/۶۶$	۴۰	۱۲۶	۲۹	$Q_1/۷۱$	۵۲	۱۸۰
۱۰	$Q_2/۶۶$	۳۵	۱۲۸	۳۰	$Q_2/۷۱$	۴۶/۲	۱۸۴
۱۱	$Q_3/۶۶$	۳۴/۹	۱۳۰	۳۱	$Q_3/۷۱$	۴۷/۱	۱۸۷
۱۲	$Q_4/۶۶$	۴۰/۲	۱۳۲	۳۲	$Q_4/۷۱$	۵۲/۷	۱۸۹
۱۳	$Q_1/۶۷$	۴۱/۹	۱۳۳	۳۳	$Q_1/۷۲$	۵۲/۲	۱۹۱
۱۴	$Q_2/۶۷$	۳۴/۷	۱۳۵	۳۴	$Q_2/۷۲$	۴۷	۱۹۳
۱۵	$Q_3/۶۷$	۳۸/۸	۱۳۸	۳۵	$Q_3/۷۲$	۴۷/۸	۱۹۴
۱۶	$Q_4/۶۷$	۴۳/۷	۱۴۰	۳۶	$Q_4/۷۲$	۵۲/۸	۱۹۶
۱۷	$Q_1/۶۸$	۴۴/۲	۱۴۳	۳۷	$Q_1/۷۳$	۵۴/۱	۱۹۹
۱۸	$Q_2/۶۸$	۴۰/۴	۱۴۷	۳۸	$Q_2/۷۳$	۴۹/۵	۲۰۱
۱۹	$Q_3/۶۸$	۳۸/۴	۱۴۸	۳۹	$Q_3/۷۳$	۴۹/۵	۲۰۲
۲۰	$Q_4/۶۸$	۴۵/۴	۱۵۱	۴۰	$Q_4/۷۳$	۵۴/۳	۲۰۴

کمترین توان های دوم موزون (WLS)

وقتی تمام خطاها ناهمبسته و دارای واریانس های مختلف اند Σ ماتریس قطری است و لذا مساله کمترین مربعات موزون است. واریانس ها معمولا با استفاده از وزن های $w_i = \cdot$ بیان می شوند .

مشاهدات با وزن های غیر یکسان در رگرسیون چندگانه همانند رگرسیون ساده می تواند روی دهد. ماتریس قطری W که عناصر روی قطر آن همان وزن های w_i هستند را بدین شکل تعریف می کنیم.

$$W = \begin{bmatrix} W_i & & \\ & \ddots & \\ & & W_n \end{bmatrix}$$

آنگاه با برآوردگردهای کمترین توان های دوم موزون مربوط به ضرایب رگرسیون عبارتند از:

$$\underline{b} = (X' W X)^{-1} X' W Y \underline{Y}$$

هنگامی که واریانس های مربوط به خطاها یعنی σ_1^2ها برابر نیستند، وزن های w_i را به گونه ای برمی گزینند که به طور معکوس با σ_1^2 متناسب باشند به طوری که $\sigma_1^2 = \sigma^2 / w_i$شود، آنگاه ماتریس واریانس- کوواریانس برآورد شده ی ضرایب رگرسیون عبارت است از:

$$est.V(\underline{b}) = (MSE_w)(X'WX)^{-1}$$

که در آن MSE_wبه صورت زیر تعریف می شود:

$$MSE_w = \frac{\sum_{i=1}^{n} w_i (Y_i - \hat{Y}_i)^2}{n-p}$$

مشابه با رگرسیون خطی ساده اغلب ممکن است وزن های w_i ها از یک رابطه اساسی مشخصی به دست آیند، برای مثال ممکن است مشخص باشد که واریانس خطاها یعنی σ_1^2 متناسب با توان دوم سطح k مین متغیر مستقل باشد. بنابراین آنگاه وزن ها عبارتند از:

$$w_i = \frac{1}{x_{ik}^2}$$

مثال : با استفاده از روش ماتریسی، مدل $Y = \beta_0 + \beta_1 x + \varepsilon$ را به داده های زیر برازش دهید و b_0 و b_1 را به دست آورید. نمودار پراکنش و خط برازش را ترسیم کنید. مقادیر $\hat{Y}$ و مانده ها را با یک رقم اعشار به دست آورید. جدول تجزیه و تحلیل واریانس را تشکیل داده و آزمون عدم پردازش را انجام دهید. ماتریس واریانس- کوواریانس پارامترها و ماتریس $V(\hat{Y})$ را هنگامی که x=65 است، پیدا کرده و برای $E(Y|x=65)$ فاصله ی اطمینان ۹۵ درصد را تشکیل دهید.

X	۳۰	۴۰	۵۰	۸۰	۳۰	۴۰	۶۰	۷۰	۷۰	۷۰	۳۰	۸۰	۷۰	۷۰
Y	۱۳	۱۷	۲۰	۲۹	۱۲	۱۵	۲۲	۲۵	۲۳	۲۷	۱۵	۲۷	۲۴	۲۶

$$X'X = \begin{bmatrix} 14 & 790 \\ 790 & 49300 \end{bmatrix}, \quad X'\underline{Y} = \begin{bmatrix} 295 \\ 18030 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} \frac{49300}{66100} & -\frac{790}{66100} \\ -\frac{790}{66100} & \frac{14}{66100} \end{bmatrix}, \quad \underline{b} = \begin{bmatrix} 4.5 \\ 0.29 \end{bmatrix}$$

پس:

$$\hat{Y} = 4.5 + 0.29x$$

$$\underline{b}'X'\underline{Y} = 6621.514372\ ,\ \underline{Y}'\underline{Y} = 6641$$
$$\frac{1}{n}\underline{Y}'11'\underline{Y} = \frac{1}{14}(295)^2 = 6216.071429$$
$$SS_T = 6641 - 6216.071429 = 424.928571$$
$$SS_R = 6621.514372 - 6216.071429 = 405.442943$$
$$SS_E = 6641 - 6621.514372 = 19.485628$$

متمایز x_i	مقادیر Y	میانگین	SS	d.f.
30	13 12 15	13/33	4/6667	۲
40	17 15	16	2	1
50	20	20	0	0
60	22	22	0	0
70	25 23 27 24 26	25	10	4
80	29 27	28	2	1
		18/6667		8

$$SS_{pe} = 18.6667\ ,\ SS_I = 19.485628 - 18.6667 = 0.818928$$

بنابراین جدول AVONA چنین است:

منبع	SS	d.f.	MS	F
رگرسیون	405/442943	1	405/442943	249/6877
مانده	19/485628	12	1/6238	
عدم برازش	0/818928	4	0/20.4732	0/0877
خطای محض	18/6667	8	2/3333	
جمع	424/928571	13		

پس $\beta_1 \neq 0$ است زیرا آزمون معنی دار است، ولی در مورد عدم برازش چون آزمون معنی دار نیست، پس مدل مناسب است.

$$estV(\underline{b}) = MSE(X'X)^{-1} = \begin{bmatrix} 1.21109 & -0.01941 \\ -0.01941 & 0.000344 \end{bmatrix}$$

برای زمانی که x=65 باشد آنگاه $X_h = \begin{bmatrix} 1 \\ 65 \end{bmatrix}$ است و

$$estV(\hat{Y}_h) = X_h^t \left[MSE(X'X)^{-1} \right] X_h = 0.14119$$
$$\hat{Y}_h = 4.5 + 0.29 \times 65 = 23.35, t_{0.025}(12) = 2.179$$
$$23.35 \pm 2.179\sqrt{0.14119} = (22.53123, 24.16876)$$

رگرسیون چندگانه

چند همخطی و اثرهای آن

در تجزیه تحلیل رگرسیون چندگانه، اغلب با ماهیت و معنی‌دار بودن رابطه‌های بین متغیرهای مستقل و متغیرهای وابسته، سروکار داریم. در این زمینه پرسش‌هایی که معمولاً مطرح می‌شود، عبارتند از:

الف: اهمیت نسبی اثرهای متغیرهای مستقل گوناگون چیست؟

ب: بزرگی اثر یک متغیر مستقل داده شده روی متغیر وابسته چقدر است؟

ج:آیا به دلیل اینکه یک متغیر مستقل دارای اثر کم و یا بدون اثر روی متغیر وابسته است، می‌توان آن متغیر مستقل را از مدل حذف کرد؟

د: آیا هر متغیر مستقل دیگری که تا کنون در مدل حضور نداشته برای تأثیر احتمالی آن باید مورد بررسی قرار گیرد؟

برای این پرسش‌ها پاسخ نسبتا ساده‌ای می‌تواند وجود داشته باشد. اگر متغیرهای مستقلی که در مدل قرار دارند (۱) در بین خودشان ناهمبسته باشند و(۲) با هر متغیر مستقل دیگری که با متغیر وابسته در ارتباط هستند، ولی از مدل حذف شده باشند نیز ناهمبسته باشند جواب های نسبتا ساده ای می توان به این سوالات داد. متأسفانه در بیشتر موقعیت‌های غیر تجربی رشته‌هایی مانند مدیریت، اقتصاد، علوم اجتماعی و علوم زیست‌شناسی متغیرهای مستقل به همبستگی بین یکدیگر تمایل دارند و یا به همبستگی با متغیرهای مستقل دیگری که با متغیر وابسته در ارتباط بوده، ولی در مدل نیستند، تمایل دارند.

هنگامی که متغیرهای مستقل بین خودشان همبسته هستند، می‌گوییم بین آنها همبستگی داخلی[۲] یا چند همخطی وجود دارد. البته، واژه‌ی چند همخطی معمولا در مواردی مورد استفاده قرار می‌گیرد که همبستگی بین متغیرهای مستقل بسیار بزرگ و یا حتی

[2] Intercorrelation

کامل باشد. اینک گونه‌ای از مسائل که در آنها متغیرهای مستقل دارای روابط داخلی هستند و به وسیله‌ی چند همخطی تبیین می‌شود را واکاوی می‌کنیم. هرچند ابتدا موقعیتی را مورد بحث قرار می‌دهیم که متغیرهای مستقل ناهمسته هستند.

مثال ۱ : برای بررسی اثر شمار افراد یک گروه عملی (x_1) و مبلغی که گروه عملی به عنوان جایزه دریافت می‌کند(x_2)، بر روی امتیازهای کسب شده به وسیله‌ی گروه (Y)، داده‌های زیر جمع‌آوری شده‌اند. نشان دهید که (x_1) و (x_2) ناهمبسته هستند، آنگاه یک معادله‌ی رگرسیون خطی چندگانه به این داده‌ها برازش دهید و جدول ANOVA را تشکیل دهید. سپس با حذف متغیر مستقل(x_2) از مدل، معادله‌ی رگرسیون خطی ساده را برازش و جدول ANOVA را تشکیل دهید. همین کار را با حذف متغیر مستقل x_1 از مدل، انجام دهید.

x_1	۴	۴	۴	۴	۶	۶	۶	۶
x_2	۲	۲	۳	۳	۲	۲	۳	۳
Y	۴۲	۳۹	۴۸	۵۱	۴۹	۵۳	۶۱	۶۰

برای محاسبه ضریب همبستگی و همچنین برازش دادن خط رگرسیون چندگانه، آگاهی‌های زیر را از داده‌ها خارج می‌کنیم

$$n = 8,\ \sum_{i=1}^{8} x_{i1} = 40\ ,\sum_{i=1}^{8} x_{i1}^{2} = 208\ ,\ \sum_{i=1}^{8} x_{i2} = 20\ ,\sum_{i=1}^{8} x_{i2}^{2} = 52\ ,\ \sum_{i=1}^{8} x_{i1}x_{i2} = 100$$

$$\sum_{i=1}^{n} Y_i = ۴۰۳\ ,\qquad \sum_{i=1}^{n} Y_i^{۲} = ۲۰۷۲۱\ ,\sum_{i=1}^{n} x_{i۱}Y_i = ۲۰۵۸\ ,\sum_{i=1}^{n} x_{i۲}Y_i = ۱۰۲۶$$

پس:

$$r_{12} = \frac{\sum_{i=1}^{n} x_{i1}x_{i2} - \frac{(\sum_{i=1}^{n} x_{i1})(\sum_{i=1}^{n} x_{i2})}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_{i1}^{2} - \frac{(\sum_{i=1}^{n} x_{i1})^{2}}{n}\right)\left(\sum_{i=1}^{n} x_{i2}^{2} - \frac{(\sum_{i=1}^{8} x_{i2})^{2}}{n}\right)}} = \frac{100 - \frac{40\times 20}{8}}{\sqrt{\left(208 - \frac{(40)^2}{8}\right)\left(52 - \frac{(20)^2}{8}\right)}}$$

پس x_1 و x_2 ناهمبسته هستند. با استفاده از نماد ماتریسی داریم:

$$X'X = \begin{bmatrix} ۸ & ۴۰ & ۲۰ \\ ۴۰ & ۲۰۸ & ۱۰۰ \\ ۲۰ & ۱۰۰ & ۵۲ \end{bmatrix},\ (X'X)^{-۱} = \begin{bmatrix} ۶.۳۷۵ & -۰.۶۲۵ & -۰.۱۲۵ \\ -۰.۶۲۵ & ۰.۱۲۵ & ۰.۰۰۰ \\ -۱.۲۵ & ۰.۰۰۰ & ۰.۵۰۰ \end{bmatrix}$$

$$X'\underline{Y} = \begin{bmatrix} ۴۰۳ \\ ۲۰۵۸ \\ ۱۰۲۶ \end{bmatrix},\ \underline{b} = (X'X)^{-۱}X'\underline{Y} = \begin{bmatrix} ۰.۳۷۵ \\ ۵.۳۷۵ \\ ۹.۲۵۰ \end{bmatrix},\ \underline{Y}'\underline{Y} = ۲۰۷۲۱$$

$$\underline{b}X'\underline{Y} = ۲۰۷۰۳.۳۷۵,\ \frac{۱}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y} = ۲۰۳۰۱.۱۲۵$$

$$SS_T = \underline{Y}'\underline{Y} - \frac{1}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y} = ۲۰۷۲۱ - ۲۰۳۰۱.۱۲۵ = ۴۱۹.۸۷۵$$

$$SS_R(b_1, b_2) = bX'Y - \frac{1}{n}Y'\mathbb{1}\mathbb{1}'\underline{Y} = ۲۰۷۰۳.۳۷۵ - ۲۰۳۰۱.۱۲۵ = ۴۰۲.۲۵۰$$

$$SS_E(b_1, b_2) = \underline{Y}'\underline{Y} - \underline{b}X'\underline{Y} = ۲۰۷۲۱ - ۲۰۷۰۳.۳۷۵ = ۱۷.۶۲۵$$

$$\hat{Y} = ۰.۳۷۵ + ۵.۳۷۵x_1 + ۹.۲۵۰x_2$$

از علامت‌های $SS_R(b_1, b_2)$ و $SS_E(b_1, b_2)$ به عنوان مجموع توان‌های دوم رگرسیون و مجموع توان‌های دوم خطا استفاده کرده‌ایم برای اینکه به صراحت وجود متغیرهای مستقل x_1 و x_2 را در مدل نشان دهیم. جدول ANOVA چنین است.

منبع	ss	df	MS
رگرسیون	$SS_R(b_1, b_2)$ =402.250	2	$\mathbf{MS_R(b_1, b_2) = 201.125}$
خطا	$SS_E(b_1, b_2) = ۱۷.۶۲۵$	5	$\mathbf{MSE(b_1, b_2)}$ =3.525
جمع	$SS_T - ۴۱۹.۸۷۵$	7	

چنانچه x_2 را از مدل حذف کنیم، داریم:

$$X'X - \begin{bmatrix} ۸ & ۴۰ \\ ۴۰ & ۲۰۸ \end{bmatrix}, \quad (X'X)^{-1} - \begin{bmatrix} ۳.۲۵ & -۰.۶۲۵ \\ -۰.۶۲۵ & ۰.۱۲۵ \end{bmatrix}$$

$$X'\underline{Y} = \begin{bmatrix} ۴۰۳ \\ ۲۰۵۸ \end{bmatrix}, \underline{b} = \begin{bmatrix} ۲۳.۵ \\ ۵.۳۷۵ \end{bmatrix}, \underline{b}X'\underline{Y} = ۲۰۵۳۲.۲۵$$

و چون بردار $\underline{Y}$ تغییر نکرده است، پس مقادیر $\underline{Y}'\underline{Y}$ و $\frac{1}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y}$ همان مقادیر قبلی هستند. بنابراین:

$$\hat{Y} = ۲۳.۵ + ۵.۳۷۵x_1$$

$$SS_R(b_1) = \underline{b}X'\underline{Y} - \frac{1}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y} = ۲۰۵۳۲.۲۵ - ۲۰۳۰۱.۱۲۵ = ۲۳۱.۱۲۵$$

$$SS_E(b_1) = \underline{Y}'\underline{Y} - \underline{b}X'\underline{Y} = ۲۰۷۲۱ - ۲۰۵۳۲.۲۵ = ۱۸۸.۷۵ \ , \ SS_T = ۴۱۹.۸۷۵$$

و جدول ANOVA چنین است.

منبع	ss	df	MS
رگرسیون	$SS_R(b_1)$ =231.125	1	$MS_R(b_1) = 231.125$
خطا	$SS_E(b_0) = ۱۸۸.۷۵۰$	6	$MSE(b_1)$ =31.458
جمع	$SS_T = ۴۱۹.۸۷۵$	7	

و چنانچه x_1 را از مدل حذف کنیم، داریم:

$$X'X = \begin{bmatrix} ۸ & ۲۰ \\ ۲۰ & ۵۲ \end{bmatrix} \quad \hat{Y} = ۲۷.۲۵ + ۹.۲۵\, x_2 \quad \underline{b} = \begin{bmatrix} ۲۷.۲۵ \\ ۹.۲۵۰ \end{bmatrix}, X'\underline{Y} = \begin{bmatrix} ۴۰۳ \\ ۱۰۲۶ \end{bmatrix}, (X'X)^{-1} = \begin{bmatrix} ۳.۲۵ & -۱.۲۵ \\ -۱.۲۵ & ۰.۵۰ \end{bmatrix},$$

$$SS_R(b_2) = ۲۰۴۷۲.۲۵ - ۲۰۳۰۱.۱۲۵ = ۱۷۱.۱۲۵ \quad ,SS_T = ۴۱۹.۸۷۵ \quad b'X'\underline{Y} = ۲۰۴۷۲.۲۵$$

$$SS_E(b_2) = ۲۰۷۲۱ - ۲۰۴۷۲.۲۵ = ۲۴۸.۷۵$$

و جدول ANOVA چنین است:

منبع	ss	df	MS
رگرسیون	$SS_R(b_2)$ =171.125	1	$MS_R(b_2) = 171.125$
خطا	$SS_E(b_2) = ۲۴۸.۷۵$	6	$MSE(b_2)$ =41.458
جمع	$SS_T = ۴۱۹.۸۷۵$	7	

از علامت $SS_R(b_1)$ به عنوان مجموع توان‌های دوم رگرسیون برای نشان دادن اینکه فقط متغیر مستقل x_0 در مدل وجود دارد استفاده شده است، علامت $SS_R(b_2)$ نیز به گونه‌ی مشابه مورد استفاده قرار گرفته است برای بیان اینکه در مدل فقط متغیر مستقل x_2 وجود دارد.

نکته‌ی جالب توجهی در مثال ۱ قابل مشاهده است اینکه ضرائب رگرسیون برای x_1 و x_2 در رگرسیون ساده و رگرسیون چندگانه یکسان هستند، بدون توجه به اینکه آیا متغیر مستقل دیگر یا هر دو متغیر مستقل در مدل وجود دارند. این یک نتیجه‌ی مهم متغیرهای مستقل ناهمبسته می‌باشد.

بنابراین، اگر متغیرهای مستقل، ناهمبسته باشند آنگاه اثرهای نسبت داده شده به آنها توسط رگرسیون خطی مرتبه‌ی اول یکسان هستند، بدون توجه به اینکه کدام یک از متغیرهای مستقل در مدل وجود دارند. در صورتی که امکان‌پذیر باشد، این دلیلی قانع کننده برای کنترل کردن آزمایش‌هاست زیرا کنترل کردن آزمایش اجازه خواهد داد که متغیرهای مستقل ناهمبسته شوند.

نکته‌ی جالب توجه دیگر مثال ۱ ، مجموع توان‌های دوم خطاست. ملاحظه می‌شود زمانی که x_1 و x_2 در مدل باشند، مجموع توان‌های دوم خطاها عبارت است از $SS_E(b_1, b_2) = ۱۷.۶۲۵$. و هنگامی که فقط x_1 در مدل وجود دارد، مجموع توان‌های دوم

خطا عبارت است از: $SS_E(b_1) = ۱۸۸.۷۵۰$. از آنجا که تغییرات Y هنگامی که فقط x_1 در مدل وجود دارد، برابر است با 188.75 و وقتی که دو متغیر مستقل x_1 و x_2 در مدل وجود دارند، این تغییرات برابر است با 17.625 است بنابراین، ممکن است که اختلاف آنها را یعنی

$$SS_E(b_1) - SS_E(b_1, b_2) = ۱۸۸.۷۵۰ - ۱۷.۶۲۵ = ۱۷۱.۱۲۵$$

به عنوان اثر متغیر مستقل x_2 تلقی کنیم. این تفاضل را با علامت $SS_R(b_1|b_2)$ نشان می‌دهیم.

پس:

$$SS_R(b_1|b_2) = SS_E(b_1) - SS_E(b_1, b_2)$$

همچنین هنگامی که تابع رگرسیون خطی را برازش می‌دهیم که در آن فقط متغیر مستقل x_2 وجود دارد، یک اندازه‌ی تلخیص شده از تغییرات Y منسوب به x_2 به نام $SS_R(b_2)$ خواهیم داشت که در مثال ۱-۳ این مقدار171.125= $SS_R(b_2)$ است که با $SS_R(b_1|b_2)$ برابر است. دلیل این موضوع ناهمبسته بودن متغیرهای مستقل x_1 و x_2 است. بحث مشابهی را برای متغیر مستقل x_1 خواهیم داشت، یعنی:

$$SS_R(b_1|b_2) = SS_E(b_1) - SS_E(b_1, b_2)$$

که برای مثال داریم:

$$SS_R(b_1|b_2) = 248.750 - 17.625 = 231.125 = SS_R(b_1)$$

دوباره تأکید می‌کنیم که دلیل تساوی $SS_R(b_1|b_2)$ و $SS_R(b_1)$ ناهمبسته بودن متغیرهای مستقل x_1 و x_2 است.

روی هم رفته زمانی که دو متغیر مستقل ناهمبسته باشند، آنگاه سهم کناری یک متغیر مستقل حاصل از تلخیص مجموع توان‌های دوم، دقیقا یکسان است با زمانی که متغیرهای مستقل دیگر در مدل وجود داشته باشد یا فقط آن متغیر مستقل به تنهایی در مدل وجود داشته باشد.

برای نشان دادن اینکه وقتی x_1 و x_2 ناهمبسته هستند، ضریب رگرسیون مربوط به متغیر x_1 تغییر نمی‌کند زمانی که متغیر مستقل x_2 به مدل رگرسیونی افزوده می‌شود، به فرمول b_1 در مدل رگرسیون چندگانه با دو متغیر مستقل x_1 و x_2 که از معادله‌های نرمال و تعریف ضریب همبستگی ناشی می‌شود، توجه کنید.

$$b_1 = \frac{\frac{\sum_{i=1}^{n}(X_{i1}-\bar{X}_1)(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_{i1}-\bar{X}_i)^2} - \left[\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{\sum_{i=1}^{n}(X_{i1}-\bar{X}_i)^2}\right]^{\frac{1}{2}} r_{Yx_2} r_{x_1x_2}}{1 \quad r^2_{x_1x_2}}$$

چون x_1 و x_2 ناهمبسته هستند، پس مقدار $r_{x_1x_2}$ صفر است و بنابراین با جایگزین کردن در معادله قبلی داریم:

$$b_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{X}_1)(Y_i-\bar{Y})}{\sum_{i=1}^{n}(x_{i1}-\bar{X}_i)^2}$$

که همان برآورد ضریب زاویه‌ی رگرسیون ساده‌ی Y نسبت به x_1 است.

در برخی موارد متغیرهای مستقل، همبسته هستند، مثلا در بررسی ارتباط خطی بین هزینه‌ی مواد غذایی و متغیرهای مستقل درآمد خانوار، پس انداز خانوار و سن رئیس خانوار، این متغیرهای مستقل بین یکدیگر همبسته هستند. افزون بر این ممکن است متغیرهای مستقل با متغیرهای مستقل دیگری که در مدل وجود ندارند، ولی در ارتباط با متغیر وابسته است، همبسته باشند. برای مثال اندازه‌ی خانوار در مثال پیش. اینک بررسی اثرهای چند همخطی، یعنی متغیرهای مستقل که به شدت همبسته هستند، بر حسب ضرایب رگرسیون و مجموع توان‌های دوم رگرسیون مورد بحث قرار می‌دهیم.

اثر چند همخطی روی ضرایب رگرسیون

این بخش را نیز مانند قبل با یک مثال تشریح و سپس نتیجه‌گیری کلی را بیان خواهیم کرد.

مثال ۲ : به منظور بررسی ارتباط بین چربی بدن(Y)، ضخامت پوست ماهیچه‌ی بازو(x_1) و محیط ران(x_2)، یک نمونه ۲۰ نفری از خانم‌های سالم بین ۲۵ تا ۳۴ ساله را گزینش و داده‌های زیر به دست آمده‌اندو نخست ضریب همبستگی خطی بین متغیرهای مستقل x_1 و x_2 را به دست آورید دوم یک معادله‌ی رگرسیون خطی چندگانه به این داده‌ها برازش دهید و جدول ANOVA را تشکیل دهید. آنگاه همین کار را با حذف x_1 از مدل تکرار کنید.

x_1	۱۸.۷	۱۹.۷	۱۴.۶	۲۹.۵	۲۷.۷	۳۰.۲
x_2	۴۶.۵	۴۴.۲	۴۲.۷	۵۴.۴	۵۵.۳	۵۸.۶
Y	۱۱.۷	۱۷.۸	۱۲.۸	۲۳.۹	۲۲.۶	۲۵.۴

۲۲.۷	۲۵.۲	۱۹.۵	۲۴.۷	۳۰.۷	۲۹.۸	۱۹.۱
۴۸.۲	۵۱.۰	۴۳.۵	۴۹.۸	۵۱.۹	۵۴.۳	۴۲.۲
۱۴.۸	۲۱.۱	۱۱.۹	۲۲.۸	۱۸.۷	۲۰.۱	۱۲.۹

۲۵.۶	۳۱.۴	۲۷.۹	۲۲.۱	۲۵.۵	۳۱.۱	۳۰.۴
۵۳.۹	۵۸.۵	۵۲.۱	۴۹.۹	۵۳.۵	۵۶.۶	۵۶.۶
۲۱.۷	۲۷.۱	۲۵.۴	۲۱.۳	۱۹.۳	۲۵.۴	۲۵.۴

ابتدا اطلاعات زیر را از داده‌ها به دست می‌آوریم:

$$n = ۲۰,\ \sum_{i=۱}^{۲۰} x_{i۱} = ۵۰۶.۱\ ,\sum_{i=۱}^{۲۰} x_{i۱}^{۲} = ۱۳۲۸۶.۲۹\ ,\ \sum_{i=۱}^{۲۰} x_{i۲} = ۱۰۲۳.۴\ ,\sum_{i=۱}^{۲۰} x_{i۲}^{۲} = ۵۲۸۸۸\ ,\ \sum_{i=۱}^{۲۰} x_{i۱}x_{i۲} = ۲۶۳۵۸.۶۹$$

$$\sum_{i=۱}^{۲۰} Y_i = ۴۰۳.۹\ ,\ \sum_{i=۱}^{۲۰} Y_i^{۲} = ۸۶۵۲.۱۵\ ,\sum_{i=۱}^{۲۰} x_{i۱}Y_i = ۱۰۶۳۱.۶۵\ ,\sum_{i=۱}^{۲۰} x_{i۲}Y_i = ۲۱۱۱۳.۵۰$$

$$X'X = \begin{bmatrix} ۲۰ & ۵۰۶.۱ & ۱۰۲۳.۴ \\ ۵۰۶.۱ & ۱۳۲۸۶.۲۹ & ۲۶۳۵۸.۶۹ \\ ۱۰۲۳.۴ & ۲۶۳۵۸.۶۹ & ۵۲۸۸۸ \end{bmatrix}\ ,,\ (X'X)^{-۱}= \begin{bmatrix} ۱۰.۸۰۷۶۱۰۴۸ & ۰.۲۸۵۵۶۷۹۷۰۶ & -۰.۳۵۱۴۵۴۱۳۲۸ \\ ۰.۲۸۵۵۶۷۹۷۰۶ & ۰.۰۱۴۲۳۶۱۶۸۵۸ & -۰.۰۱۲۶۲۰۹۵۴۰۱ \\ -۰.۳۵۱۴۵۴۱۳۲۸ & -۰.۰۱۲۶۲۰۹۵۴۰۱ & ۰.۰۱۳۱۰۹۷۷۸۶۶ \end{bmatrix}$$

$$X'\underline{Y} = \begin{bmatrix} ۴۰۳.۹ \\ ۱۰۶۳۱.۶۵ \\ ۲۱۱۱۳.۵۰ \end{bmatrix},$$

$$r_{x_۱x_۲} = \frac{۲۶۳۵۸.۶۹ - \frac{۵۰۶.۱۰ \times ۱۰۲۳.۴}{۲۰}}{\sqrt{۴۷۹.۴۲۹۵ \times ۵۲۰.۶۲۲}} = ۰.۹۲۳۸$$

$$\underline{b} = (X'X)^{-۱}X'\underline{Y} = \begin{bmatrix} -۱۹.۱۷۴۲۴۵۶۴ \\ ۰.۲۲۲۳۵۲۵۹۱۱ \\ ۰.۶۵۹۴۲۱۷۹۶۵ \end{bmatrix},\underline{Y}'\underline{Y} = ۸۶۵۲.۱۵\ ,\hat{Y} = -۱۹.۱۷۴۲ + ۰.۲۲۲۴x_۱ + ۰.۶۵۹۴x_۲$$

$$\underline{b}X'\underline{Y} = ۸۵۴۲.۱۹۹۲۱,\frac{۱}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y} = ۸۱۵۶.۷۶۰۵$$

$$SS_T - \underline{Y}'\underline{Y} - \frac{۱}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y} - ۸۶۵۲.۱۵ - ۸۱۵۶.۷۶۰۵ - ۴۹۵.۳۸۹۵$$

$$SS_R(b_۱,b_۲) = \underline{b}X'\underline{Y} - \frac{۱}{n}\underline{Y}'\mathbb{1}\mathbb{1}'\underline{Y} = ۸۵۴۲.۱۹۹۲۱ - ۸۱۵۶.۷۶۰۵ = ۳۸۵.۴۳۸۷۱$$

$$SS_E(b_۱,b_۲) = \underline{Y}'\underline{Y} - \underline{b}X'\underline{Y} = ۸۶۵۲.۱۵ - ۸۵۴۲.۱۹۹۲۱ = ۱۰۹.۹۵۰۷۹$$

منبع	ss	df	MS
رگرسیون	$SS_R(b_1, b_2)$ =385.438 71	2	$\mathbf{MS_R(b_1, b_2) = 192.719355}$
خطا	$SS_E(b_1, b_2) = ۱۰۹.۹۵۰۷۹$	17	$\mathbf{MSE(b_1, b_2)}$ =6.467693529
جمع	$SS_T = ۴۹۵.۳۸۹۵$	19	

$$X'X = \begin{bmatrix} ۲۰ & ۵۰۶.۱ \\ ۵۰۶.۱ & ۱۳۲۸۶.۲۹ \end{bmatrix} \quad , \quad X'Y = \begin{bmatrix} ۴۰۳.۹ \\ ۱۰۶۳۱.۶۵ \end{bmatrix} \quad , \quad (X'X)^{-1} = \begin{bmatrix} ۱.۳۸۵۶۳۵۴۲۷ & -۰.۰۵۲۷۸۱۴۸۲ \\ -۰.۰۵۲۷۸۱۴۸۲ & ۰.۰۰۲۰۸۵۸۱۲۴۰۸ \end{bmatrix}$$

$$\underline{b} = \begin{bmatrix} -۱.۴۹۶۰۹۴۱۴ \\ ۰.۸۵۷۱۸۲۵۷ \end{bmatrix}, \hat{Y} = -۱.۴۹۶۱ + ۰.۸۵۷۲x_1$$

و چون بردار $\underline{Y}$ تغییری نکرده است، پس :

$$SS_T = \underline{Y}'\underline{Y} - \frac{1}{n}\underline{Y}'\mathbb{1}'\underline{Y} = ۸۵۰۸.۹۹۲۶۴۷ - ۸۱۵۶.۷۶۰۵ = ۴۹۵.۳۸۹۵$$

$$SS_R(b_1) = \underline{b}X'\underline{Y} - \frac{1}{n}\underline{Y}'\mathbb{1}'\underline{Y} = ۸۵۰۸.۹۹۲۶۴۷ - ۸۱۵۶.۷۶۰۵ = ۳۵۲.۲۳۲۱۴۷۱$$

$$, SS_E(b_1) = \underline{Y}'\underline{Y} - \underline{b}X'\underline{Y} = ۸۶۵۲.۱۵ - ۸۵۰۸.۹۹۲۶۴۷ = ۱۴۳.۱۵۷۳۵۳$$

و جدول ANOVA چنین است.

منبع	ss	df	MS
رگرسیون	$SS_R(b_1)$ =352.2321471	1	$\mathbf{MS_R(b_1) - 352.2321471}$
خطا	$SS_E(b_1) = ۱۴۳.۱۵۷۳۵۳$	18	$\mathbf{MSE(b_1)}$ =7.953186278
جمع	$SS_T = ۴۹۵.۳۸۹۵$	19	

چنانچه x_1 را از مدل حذف کنیم، داریم:

$$X'X = \begin{bmatrix} ۲۰ & ۱۰۲۳.۴ \\ ۱۰۲۳.۴ & ۵۲۸۸۸ \end{bmatrix} \quad , \quad X'Y = \begin{bmatrix} ۴۰۳.۹ \\ ۲۱۱۱۳.۵۰ \end{bmatrix} \quad , \quad (X'X)^{-1} = \begin{bmatrix} ۵.۰۷۹۳۰۸۹۸ & -۰.۰۹۸۲۸۶۲۸ \\ -۰.۰۹۸۲۸۶۲۸ & ۰.۰۰۱۹۲۰۷۷۹۳۷۵ \end{bmatrix}$$

$$\underline{b} = \begin{bmatrix} -۲۳.۶۳۴۴۷۵۷ \\ ۰.۸۵۶۵۵۳۸۹۲۴ \end{bmatrix}, \hat{Y} = -۲۳.۶۳۴۴۷۵۷ + ۰.۸۵۶۵۵۳۸۹۲۴x_2, \underline{b}X'\underline{Y} = ۸۵۳۸.۸۸۵۸۶۳$$

و چون بردار $\underline{Y}$ تغییری نکرده است، پس، SS_T تغییر نمی‌کند و داریم:

$$SS_R(b_2) = \underline{b}X'\underline{Y} - \frac{1}{n}\underline{Y}'\mathbb{1}'\underline{Y} = ۸۵۳۸.۸۸۵۸۶۳ - ۸۱۵۶.۷۶۰۵ = ۳۸۲.۱۲۵۳۶۲۸$$

$$SS_E(b_2) = \underline{Y}'\underline{Y} - \underline{b}\underline{X}'\underline{Y} = ۸۶۵۲.۱۵ - ۸۵۳۸.۸۸۵۸۶۳ = ۱۱۳.۲۶۴۳۳۷$$

و جدول ANOVA چنین است

منبع	ss	df	MS
رگرسیون	$SS_R(b_2)$ =382.1253628	1	$\mathbf{b_2) = 382.1253628}$
خطا	$SS_E(b_2) = ۱۱۳.۲۶۴۳۳۷$	18	$\mathbf{MSE(b_2)}$ =6.292443167
جمع	$SS_T = ۴۹۵.۳۸۹۵$	19	

ملاحظه می‌شود که ضریب رگرسیون متغیرهای x_1 برای زمانی که x_2 در مدل وجود دارند و زمانی که وجود دارد، یکسان نیست. بنابراین، در اینجا اثر نسبت داده شده به x_1 به وسیله‌ی تابع پاسخ برازش شده، بر حسب اینکه آیا متغیر x_1 یا هر دو متغیر مستقل x_1 و x_2 در مدل وجود داشته باشند، مختلف است. دلیل این اختلاف آن است که متغیرهای مستقل x_1 و x_2 ، به شدت همبسته هستند. همین موضوع در مورد متغیر مستقل x_2 نیز قابل مشاهده است.

نتیجه مهمی که از بحث بالا می‌توان گرفت عبارت است از: هنگامی که متغیرهای مستقل همبسته هستند، ضریب رگرسیون هر متغیر مستقلی به حضور و یا عدم حضور متغیر مستقل دیگر وابسته است. بنابراین، یک ضریب رگرسیون، اثر ذاتی متغیر مستقل مشخصی روی متغیر وابسته را بازتاب نمی‌کند، ولی زمانی که تعدادی متغیرهای مستقل همبسته‌ی دیگر در مدل وجود داشته باشد، فقط اثری کناری(حاشیه‌ای) یا اثری جزئی را بازتاب می‌کند.

مرکزی کردن و مقیاس بندی

تمام شاخص های همخطی که تاکنون بحث شده اند را با استفاده از محاسبات رگرسیون استاندارد می توان به دست آورد. برای تحلیل همخطی یک راه دیگری هم وجود داردکه لازمه آن محاسباتی است که معمولا در بسته های نرم افزار رگرسیونی استاندارد لحاظ نشده است. این تحلیل از حقیقتی پیروی می کند که هرالگوی رگرسیون خطی را می توان بر حسب مجموعه ای از متغیرهای پیشگویی متعامد دوباره آغاز کرد. این متغیرهای جدید بصورت ترکیبات خطی متغیرهای پیشگوی اولیه به دست می آیند که به عنوان مولفه های اصلی مجموعه متغیرهای پیشگویی در نظر گرفته می شوند (سبر ۱۹۸۴، جانسون و ویچرن۱۹۹۲)

برای توسعه روش مولفه های اصلی ابتدا لازم است متغیرها را مرکزی و یا مقیاس بندی کنیم . عمدتا با الگوهای رگرسیونی به شکل:

$$Y = \beta_0 + \beta_1 X_1 + ... \beta_P X_P + \varepsilon \quad *$$

سروکار داریم که الگوهایی با جمله ثابت β_0 هستند. ولی با وضعیت هایی نیز برخورد کرده ایم که در آن برازش الگوی بدون عرض از مبدا:

$$Y = \beta_1 X_1 + ... \beta_P X_P + \varepsilon \quad **$$

ضروری است . وقتی با الگوهای با جمله ثابت سروکار داریم بهتر است متغیرها را مرکزی و مقیاس بندی کنیم اما وقتی با یک الگوی بدون عرض از مبدا سروکار داریم تنها مقیاس بندی متغیرها لازم است.

مرکزی کردن و مقیاس بندی در الگوهای با عرض از مبدا

اگر الگویی مانند الگوی(*)که شامل عرض از مبدا است را برازش می کنیم باید متغیرها را مرکزی و مقیاس بندی کنیم. یک متغیر مرکزی شده از کم کردن میانگین تمام مشاهدات به دست می آید. برای مثال متغیر پاسخ مرکزی شده($Y - \overline{y}$) و j امین متغیر پیشگو مرکزی شده ($X_j - \overline{x_j}$) است. میانگین یک متغیر مرکزی شده برابر صفر است. متغیرهای مرکزی شده را مقیاس بندی نیز می توان کرد. معمولا دو نوع مقیاس بندی لازم می شود:

مقیاس بندی طول واحد و استاندارد کردن. مقیاس بندی طول واحد متغیر پاسخ y و متغیر پیشگوی j ام یعنی X_j به طریق زیر به دست می آید:

$$Z = \frac{Y - \overline{y}}{L_y}$$

$$\overline{Z_j} = \frac{X_j - \overline{X_j}}{L_j} \quad \text{j=1,2...p}$$

که در آن $\overline{y}$ میانگین y و $\overline{x_j}$ میانگین x_j است و

$$L_y = \sqrt{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \quad L_j = \sqrt{\sum_{i=1}^{n} (x_{ij} - \overline{x_j})^2}, \quad \text{j=1,...p}$$

کمیت L_y طول متغیر مرکزی شده $Y - \overline{y}$ گفته می شود زیرا این کمیت اندازه یا بزرگی مشاهدات را اندازه می گیرد. بطور مشابه L_j طول متغیر $X_j - \overline{x_j}$ را اندازه می گیرد. متغیرهای $\overline{Z_y}$ و $\overline{Z_j}$ دارای میانگین های صفر و طول های واحد است. لذا این نوع مقیاس بندی را مقیاس بندی طول واحد می نامند. علاوه بر این ها مقیاس بندی طول واحد دارای خواص زیر است:

$$Cor(X_j, X_k) = \sum_{i=1}^{n} Z_{ij} Z_{ik}$$

یعنی ضریب همبستگی بین متغیرهای اولیه X_j و X_K را بصورت مجموع حاصل ضربهای صورت های مقیاس بندی شده Z_K و Z_j به سهولت می توان محاسبه نمود وقتی است که متغیرهای پیشگو ناهمبسته باشد.

نوع دوم مقیاس بندی استاندارد کردن است که بصورت زیر تعریف می شود.

$$Y^{*} = \frac{Y - \overline{y}}{S_y}$$

که در آن:

$$X^{*}_{j} = \frac{X_j - \overline{X_j}}{s_j} \quad \text{j=1,2...p}$$

$$s_j = \sqrt{\frac{\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2}{n-1}},\ s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}},\ \text{j=1,...p}$$

به ترتیب انحراف های معیار متغیر پاسخ و متغیرهای پیشگوی j ام است متغیرهای استاندارد شده Y^{*} و X^{*}_{j} در دارای میانگین های صفر و انحراف های معیار واحد اند چون همبستگی ها تحت تاثیر انتقال یا تبدیل داده ها قرار نمی گیرند لذا سرو کار داشتن با مقیاس بندی طول واحد یا صورت های استانداردشده متغیرها هم کافی و هم مناسب است . واریانس ها و کواریانس های P متغیر $X_1....X_P$ را می توان بصورت یک ماتریس نشان داد این ماتریس را ماتریس واریانس-کواریانس می نامند. اعضای روی نظر اصلی که از گوشه چپ بالا تا گوشه راست پایین ادامه دارد را اعضای قطری می نامند. اعضای قطری یک ماتریس واریانس-کواریانس واریانس اعضا بوده و اعضای خارجی قطر اصلی کواریانس ها هستند.

مقیاس بندی الگوهای بدون عرض از مبدا

اگر الگویی بدون عرض از مبدا مانند الگوی(***) را برازش می کنیم داده ها را مرکزی نمی کنیم زیرا مرکزی کردن دارای اثر لحاظ کردن یک جمله ثابت در الگو است این را به شکل زیر می توان دید:

$$Y - \overline{y} = \beta_1(X_1 - \overline{x}) + ...\beta_P(X_P - \overline{x_P}) + \varepsilon$$

که با جابجا کردن جملات الگو نتیجه می شود:

$$Y = \overline{y} - (\beta_1\overline{x} + ...\beta_P\overline{x_P}) + \beta_1 X_1 + ...\beta_P X_P + \varepsilon = \beta_0 + \beta_1 X_1 + ...\beta_P X_P + \varepsilon \quad *$$

که در آن گرچه یک جمله ثابت به شکلی صریح ظاهر نمی شود ولی آن را در (*) به وضوح می توان دید بدین ترتیب وقتی الگوهای بدون عرض از مبدا را بررسی می کنیم فقط باید داده ها را مقیاس بندی کنیم. متغیر های مقیاس بندی شده را بصورت زیر تعریف می کنیم:

$$Z_Y = \frac{Y}{L_y}$$

$$Z_j = \frac{X_j}{L_j} \quad j=1,2...p$$

که در آن

$$L_y = \sqrt{\sum_{i=1}^{n}(y_i)^2}, \quad L_j = \sqrt{\sum_{i=1}^{n}(x_{ij})^2}, \quad j=1,...p$$

در این جا باید متذکر شد که مرکزی کردن(در صورت مناسب بودن) و یا مقیاس بندی را بدون اینکه به کلیت خللی وارد شود را می توان انجام داد زیرا ضرایب رگرسیون متغیرهای اولیه را میتوان از ضرایب رگرسیون متغیرهای تبدیل یلفته دوباره به دست آورد برای مثال اگر یک الگوی رگرسیون به داده های مرکزی شده برازش کنیم ضرایب رگرسیون به دست آمده یعنی $\beta_1....\beta_P$ همانند برآوردهایی است که از برازش الگو به داده های اولیه به دست می آیند وقتی از داده های مرکزی شده استفاده می کنیم برآورد جمله ثابت همیشه صفر است براورد جمله ثابت برای یک الگویی با عرض از مبدا می توان به شکل زیر بدست آورد:

$$\beta_0 = \overline{y} - (\beta_1 \overline{x} + ...\beta_P \overline{x_P})$$

در عین حال مقیاس بندی مقادیر ضرایب رگرسیون برآورد شده را تغییر می دهد برای مثال رابطه بین برآوردهای $\beta_1....\beta_P$ که از داده های اولیه بدست میاید و برآوردهایی که با بکار بردن داده های استاندارد شده حاصل می شود عبارت اند از:

$$\beta_j = (S_y / S_j)\hat{\theta}_j \qquad j=1,2,3....p$$
$$\beta_0 = \overline{y} - \sum_{j=1}^{5} \beta_j \overline{X}_j$$

که در آن β_j و θ_j امین برآورد ضرایب رگرسیون است که به ترتیب از داده های اولیه و استاندارد شده استفاده شده است. اگر از مقیاس بندی ظول واحد به جای استاندارد کردن استفاده کنیم فرمول های مشابهی به دست می آیند در باقیمانده این فصل و فصل ۱۰ به کررات از متغیرهای مرکزی شده و مقیاس بندی شده استفاده می کنیم.

فرم ماتریسی رگرسیون خطی ساده و چندگانه

در تجزیه و تحلیل های آماری و ریاضی، جبر ماتریس ها به طور گسترده ای مورد استفاده قرار می گیرد. در واقع روش ماتریسی تجزیه و تحلیل های رگرسیون خطی چندگانه یک ضرورت است، زیرا اجازه می دهد که دستگاه های معادله و آرایه های بزرگ داده ها به‌صورت فشرده نمایش داده شوند و به گونه کارآمدی بکار گرفته شوند.

نحوه محاسبه ماتریس واریانس- کوواریانس بردار $\underline{b}$

ماترسی واریانس- کوواریانس بردار تصادفی $\underline{b}$ به‌صورت

$$V(\underline{b}) = E\left[\left(\underline{b} - E(\underline{b})\right)\left(\underline{b} - E(\underline{b})\right)'\right]$$

تعریف می‌شود که با توجه به تعریف بردار $\underline{b}$ داریم:

$$v(\underline{b}) = E\left[\begin{bmatrix} b_0 - B_0 \\ b_1 - B_1 \end{bmatrix} [b_0 \quad B_0 \quad b_1 \quad B_1]\right]$$

$$= \begin{bmatrix} (b_0 - B_0)^2 & (b_0 - B_0)(b_1 - B_1) \\ (b_0 - B_0)(b_1 - B_1) & (b_1 - B_1)^2 \end{bmatrix}$$

$$= \begin{bmatrix} v(b_0) & cov(b_0, b_1) \\ cov(b_0, b_1) & v(b_1) \end{bmatrix}$$

بنابراین با استفاده از فرمول‌های قبلی داریم:

$$v(\underline{b}) = \begin{bmatrix} \sigma^2 \sum_{i=1}^{n} x_i^2 & -\bar{x}\sigma^2 \\ n\sum_{i=1}^{n}(x_i - \bar{x})^2 & \sum_{i=1}^{n}(x_i - \bar{x})^2 \\ -\bar{x}\sigma^2 & \sigma^2 \\ \overline{\sum_{i=1}^{n}(x_i - \bar{x})^2} & \overline{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{bmatrix}$$

که با توجه به تعریف ماتریس $(x'x)^{-1}$ خواهیم داشت:

$$v(\underline{b}) = \sigma^2 (x'x)^{-1}$$

که این نتیجه بسیار مهم و جالب است و باید آن را به خاطر بسپاریم. چنانچه عدم برازش وجود نداشته باشد، آنگاه در صورت مجهول بودن σ^2 از برآورد نقطه ای نااریب آن یعنیMSE که پیشتر معرفی شده است، استفاده می‌کنیم و اگر عدم برازش وجود داشته باشد و σ^2 مجهول باشد، آنگاه از برآورد آن به نام میانگین توان‌های دوم محض، یعنی MS_{pe} استفاده می‌کنیم. با این وجود ترتیب ماتریس واریانس- کوواریانس برآورد شده عبارت است از:

$$est.\, v(\underline{b}) = MSE(x'x)^{-1}$$

اینک فرض کنید X_K یک مقدار داده شده‌ی متغیر مستقل x باشد، آنگاه همانگونه که پیشتر بیان شد، پیش بینی میانگین Y برای این مقدار x داده شده عبارت است از:

$$\hat{Y}_k = b_0 + b_1 x_k$$

چنانچه $x_h = \begin{bmatrix} 1 \\ x_k \end{bmatrix}$ را تعریف کنیم، آنگاه :

$$\hat{Y}_k = [1 \quad x_k] \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = x_h' \underline{b} = \underline{b}' x_h$$

آنگاه واریانس، $\hat{Y}_k$ عبارت است از:

$$v(\hat{Y}_k) = v(x_h'\underline{b}) = x_h' v(\underline{b}) x_h$$

$$v(\hat{Y}_k) = \sigma^2 x_h'(x'x)^{-1}x_h \quad \text{و یا}$$

که در صورت مجهول بودن σ^2 واریانس برآورده شده، $\hat{Y}_k$ چنین است:

$$\text{est.}\, v(\hat{Y}_k) = (MSE)x_h'(x'x)^{-1}x_h$$

صورت ماتریسی ضریب تعیین چنین است:

$$R^2 = \frac{\underline{b}'x'\underline{Y} - \frac{1}{n}\underline{Y}'ll'\underline{Y}}{\underline{Y}'\underline{Y} - \frac{1}{n}\underline{Y}'ll'\underline{Y}}$$

مثال: جدول ANOVA را برای مثال ۱ و با استفاده از روش ماتریسی تشکیل دهید. همچنین ماتریس واریانس کوواریانس $\underline{b}$ را به دست آورده و سپس واریانس $\hat{Y}_4$ را برآورد کنید.

$$\underline{Y}=\begin{bmatrix}5\\13\\16\\23\\33\\38\\40\end{bmatrix}; X=\begin{bmatrix}1&1\\1&2\\1&3\\1&4\\1&5\\1&6\\1&7\end{bmatrix}, X'X=\begin{bmatrix}7&28\\28&140\end{bmatrix}$$

$$(x'x)^{-1}=\begin{bmatrix}\frac{5}{7}&\frac{-1}{7}\\\frac{-1}{7}&\frac{1}{28}\end{bmatrix}; X'\underline{Y}=\begin{bmatrix}168\\844\end{bmatrix}; \underline{Y}'\underline{Y}=5112$$

$$b=(x'x)^{-1}X'Y=\begin{bmatrix}-0.571\\6.143\end{bmatrix}$$

$$\underline{b}X'\underline{Y}=5088.596, \frac{1}{n}\underline{Y}'ll'\underline{Y}=4032$$

$$SS_R=5088.596-4032=1056.596$$

$$SS_E=\underline{Y}'\underline{Y}-\underline{b}X'\underline{Y}=5112-5088.596=23.404$$

منبع	SS	d.f	MS
رگرسیون	1056.596	1	1056.596
مانده	23.404	5	4.6808
جمع	1080	6	

$$V(b) = \sigma^2(x'x)^{-1} = \sigma^2\begin{bmatrix} \frac{5}{7} & \frac{-1}{7} \\ \frac{-1}{7} & \frac{1}{28} \end{bmatrix}$$

$$\text{est.v}(\underline{b}) = (MSE)(x'x)^{-1} = 4.6808\begin{bmatrix} \frac{5}{7} & \frac{-1}{7} \\ \frac{-1}{7} & \frac{1}{28} \end{bmatrix} = \begin{bmatrix} 3.343 & -0.669 \\ -0.669 & 0.167 \end{bmatrix}$$

$$\text{est.s.e}(b_1)=\sqrt{0.167} = 0.4086 \; ; \; \text{est.s.e}(b_.) = \sqrt{3.343} = 1.8284$$

که همان پاسخ‌هایی هستند که در ابتدای فصل به دست آمده‌اند.

$$E[Y|x = 4] = \hat{Y}_4 \quad ; x_h = \begin{bmatrix} 1 \\ 4 \end{bmatrix}$$

$$\hat{Y}_4 = x_h'\underline{b} = [1 \quad 4]\begin{bmatrix} -0.571 \\ 6.143 \end{bmatrix} = 24$$

$$=[1 \quad 4]\begin{bmatrix} 3.343 & -0.669 \\ -0.669 & 0.167 \end{bmatrix} = 0.663 = x_h' \; \text{est.v}(\underline{b})x_h \; \text{est.v}(\hat{Y}_k) = (MSE)x_h'(x'x)^{-1}x_h$$

کمترین توان‌های دوم موزون

روش کمترین توان‌های دوم موزون می‌تواند با روش ماتریسی به صورت فشرده بیان شود. فرض کنید ماتریس W یک ماتریس قطری باشد که عناصر روی قطر آن وزن‌های w_i باشند، یعنی

$$W = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix}$$

آنگاه معادله‌های نرمال کمترین توان‌های دوم موزون (۲-۱۲-۱) به صورت زیر می‌توان نوشته شود:

$$x'wx\underline{b} = x'w\underline{Y}$$

با فرض معکوس‌پذیری ماتریس $x'wx$ برآوردهای کمترین توان‌های دوم موزون عبارتند از:

$$\underline{b} = (x'wx)^{-1}x'w\underline{Y}$$

دقت شود که اگر W=I باشد، آنگاه همان معادله‌های کمترین توان‌های دوم ناموزون به دست می‌آیند. ماتریس واریانس- کوواریانس برآوردگرهای کمترین توان‌های دوم موزون عبارت است از:

$$\mathbf{V}(\underline{\mathbf{b}}) = \sigma^2(\mathbf{x}'\mathbf{w}\mathbf{x})^{-1}$$

و ماتریس واریانس- کوواریانس برآورد شده‌ی بردار $\underline{\mathbf{b}}$ چنین است:

$$\mathbf{est.v}(\underline{\mathbf{b}}) = (\mathbf{MSE_w})(\mathbf{x}'\mathbf{w}\mathbf{x})^{-1}$$

که در آن

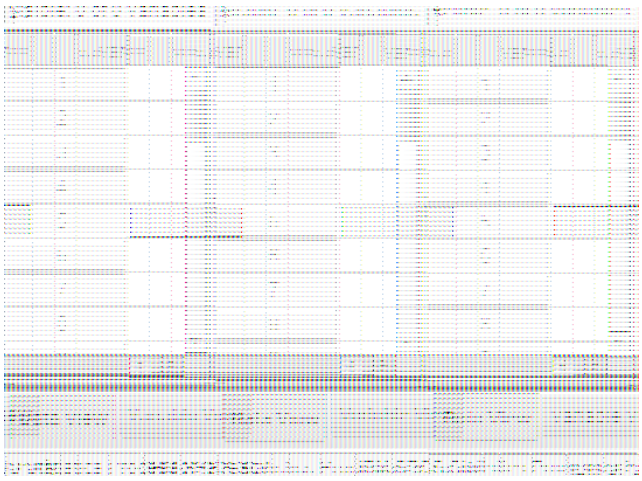
$$\mathbf{MSE_w} = \frac{\sum_{i=1}^{n} w_i(Y_i - \hat{Y}_i)^2}{n-2}$$

SPSS

SPSS حروف اول کلمات عبارت Statistical Package for the Social Science به معنای بسته ی نرم افزار آماری برای علوم اجتماعی، برنامه ی شناخته شده و پرکاربردی است که برای تحلیل آماری طراحی شده است و با تسلط بر آن قادر خواهیم بود پژوهش های پیچیده ای را انجام دهیم که بدون این نرم افزار امکان انجام آنها بسیار دشوار و یا حتی غیر ممکن می نماید.

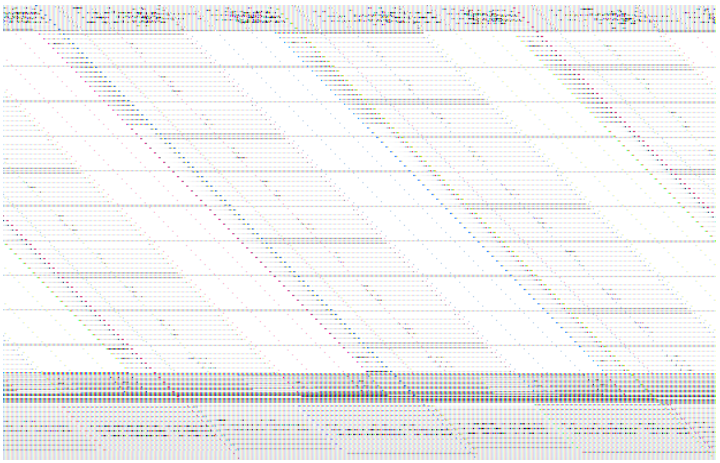
شروع SPSS و تعریف متغیرها

بااجرای برنامه ی SPSS، پنجره ی Data editor مطابق شکل ۱ باز می شود.



شکل ۱

این پنجره دارای دو زبانه ی Data View و Variable View می باشد. در زبانه ی Data View هر سطر نشان دهنده ی یک مورد و هر ستون نشان دهنده ی یک متغیر است. تعریف متغیرها در زبانه ی Variable View انجام می گیرد.



شکل ۲

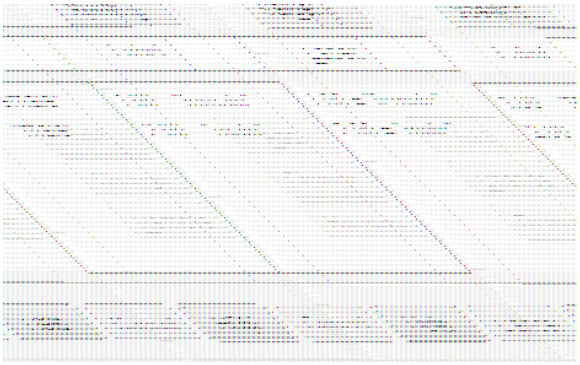
در این پنجره همان گونه که در شکل ۲ می بینید ویژگی های مختلف هر متغیر را می توان تعریف کرد. در این پنجره هر سطر نشان دهنده ی یک متغیر و ستون ها ویژگی های آن متغیر می باشد. در اولین ستون نام متغیر و در ستون بعدی نوع آن را تعریف می کنیم. نوع متغیر همان گونه که در شکل ۳ مشاهده می کنید می تواند حالات مختلف عددی، رشته ای و حتی تاریخ و واحد پولی باشد.

شکل ۱

شکل ۱

توصیه می شود برای انواع متغیرهای اسمی نیز نوع عددی "Numeric" را انتخاب کرد تا اگر در مراحل بعدی نیاز به کدگذاری پیش آمد با مشکلی مواجه نشویم.

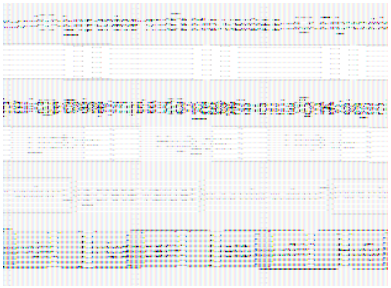
در دو ستون بعدی عرض متغیر و تعداد ارقام اعشاری را تعیین می کنید و در ستون پنجم که Label نام دارد می توانید از عبارتی که تداعی کننده ی نام و نوع داده های این متغیر است به عنوان بر چسب برای آن استفاده کنید. این برچسب ها هنگام نمایش نتایج در نمودارها و جداول دیده می شوند و مشاهده ی آنها به خصوص زمانی که تعداد متغیرها زیاد است و اسم متغیرها خلاصه شده بسیار مفید می باشد. ستون بعدی زبانه ی Variable View تحت عنوان Values زمانی استفاده می شود که برای متغیرهای اسمی و یا ترتیبی بخواهیم از برچسب های مقداری استفاده کنیم. برای مثال در متغیر جنسیت همان گونه که در شکل ۴ مشاهده می کنید با کلیک بر روی خانه ی مربوط به این متغیر در ستون Values پنجره ی Value Labels باز خواهد شد. در این پنجره می توانیم مقدار عددی ۲ را به مردان و ۱ را به زنان نسبت دهیم. پس از هر بار باید دکمه ی Add را کلیک کنیم و نتیجه را در کادر جلوی آن



شکل ۴

مشاهده نماییم و پس از کلیک بر روی ok به صفحه ی Variable View باز خواهیم گشت.

ستون بعدی به نام Missing زمانی مورد استفاده قرار می گیرد که در میان داده ها، داده های مبهم و ناخوانا داشته باشیم و یا در بعضی سوالات، شرکت کنندگان از پاسخ به سوال امتناع کرده باشند و یا شخص پژوهشگر مایل باشد که داده های به خصوصی را در تحلیل آماری خود در نظر نگیرد. پس از کلیک بردکمه ی موجود در خانه ی Missing پنجره ی Missing Values مطابق شکل ۵ باز خواهد شد. در صورت کلیک بر روی اولین دایره مقدار از دست رفته ای نخواهید داشت. دومین دایره اجازه ی داشتن سه مقدار از دست رفته را به شما می دهد ودایره ی سوم به شما این امکان را می دهدکه یک مقدار از دست رفته و یک بازه که حد پایین و بالای آن را مشخص می کنید برای مقادیر Missing خود تعریف نمایید.

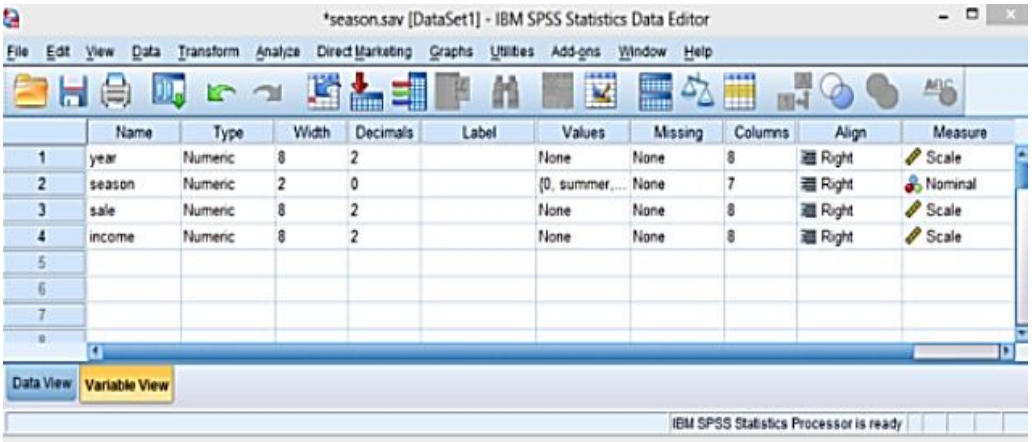


شکل ۵

در ستون بعدی با عنوان Columns عرض ستونی را که قرار است متغیر در زبانه ی Data View اشغال کند تعریف می کنیم. در ستون Align این امکان را داریم که مکان قرارگیری اطلاعات موجود در خانه ی هر متغیر را تعیین نماییم. در ستون Measure مقیاس اندازه گیری متغیر را از بین گزینه های اسمی (Nominal)، ترتیبی (Ordinal) و Scale برای متغیرهای فاصله ای و نسبی انتخاب می کنیم.

پس از انجام این مراحل به زبانه ی Data View رفته و مشاهده می کنیم که نام متغیرهای تعریف شده در ستون های این زبانه قرار گرفته است. اکنون باید مقادیر مربوط به موارد مختلف هر متغیر را وارد کرد. این کار هم به صورت دستی ویا به کمک دستورهای copy/paste از فایل های برنامه های دیگری نظیر اکسل امکان پذیر است.

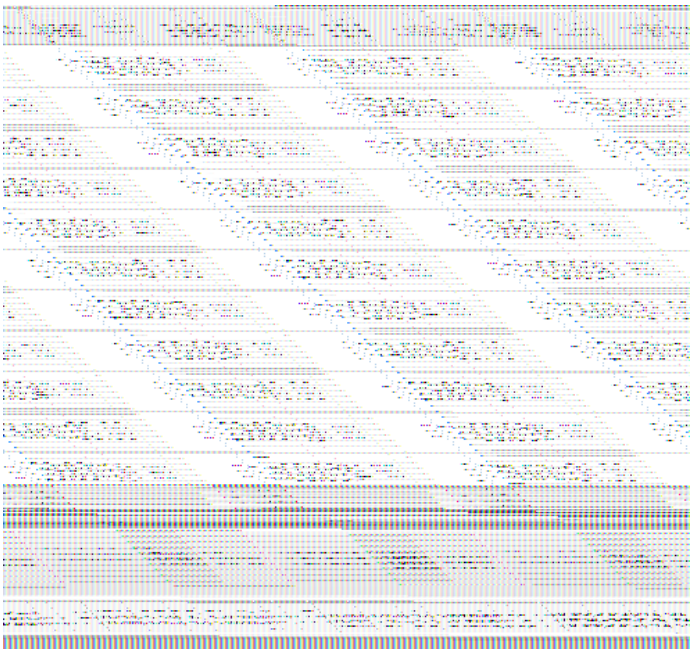
شکل ۶ زبانه ی Variable View برای مثالی که آن را فصل ها نامیده ایم نشان می دهد.



شکل۶

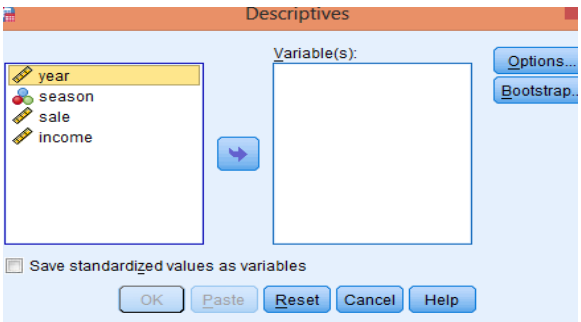
همان گونه که مشاهده می کنید چهار متغیر این مثال، سال، فصل، میزان فروش و درآمد می باشند که از این متغیرها، متغیر فصل را اسمی و بقیه را Scale تعریف کرده ایم. برای متغیر فصل در ستون values مقدار ۰ را به فصل های تابستان و پاییز و مقدار ۱ را به فصل های بهار و زمستان اختصاص داده ایم.

شکل ۷ زبانه ی Data View را برای مثال فصل نشان می دهد مقادیر عددی مربوط به هر متغیر از فایل اکسل copy/paste شده اند.



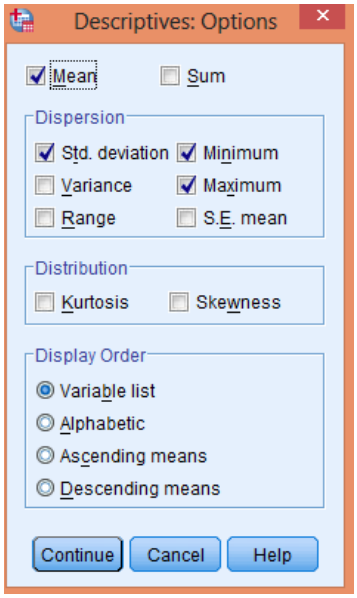
آمار توصیفی در SPSS

برای بدست آوردن شاخص های رایج در آمار توصیفی نظیر شاخص های گرایش به مرکز (میانگین، میانه، مد...) و شاخص های پراکندگی (انحراف استاندارد، واریانس، چارکها و ...) چندین راه وجود دارد. آمار توصیفی اغلب به عنوان output اختیاری از بخش آمار استنباطی قابل دستیابی است. اما دستورهایی هم وجود دارند که اختصاصاً برای به دست آوردن شاخص های آمار توصیفی طراحی شده اند. از جمله ی این دستورها از Analyze در بالای صفحه شروع شده سپس با انتخاب Descriptive Statistics ادامه می یابد. پس از آن اگر می خواهیم برای یک متغیر کیفی شاخص های مربوط به فراوانی و نمودارهای مرتبط را به دست آوریم باید



گزینه ی Frequencies را انتخاب نماییم و در صورتی که متغیر مورد نظر کمی باشد گزینه ی Descriptive را انتخاب کرده و با پنجره ای نظیر شکل ۹ روبرو می شویم در این پنجره باید متغیرهایی که مایلیم شاخص های آماری آنها را محاسبه کنیم به پنجره ی Variable(s) منتقل نماییم. شکل۹

درصورتی که مربع کوچک پایین پنجره را تیک بزنیم متغیرهای مورد نظر را استاندارد کرده و در ستون جدیدی ذخیره می نماید. با کلیک بر روی Options پنجره ی شکل ۱۰ باز شده و به ما این امکان را می دهد تا شخص های آماری مورد نظر را انتخاب نماییم.



در بالای پنجره میانگین و مجموع، در کادر بعدی شاخص های پراکندگی و در کادر میانی شاخص های کشیدگی و چولگی قرار دارند.

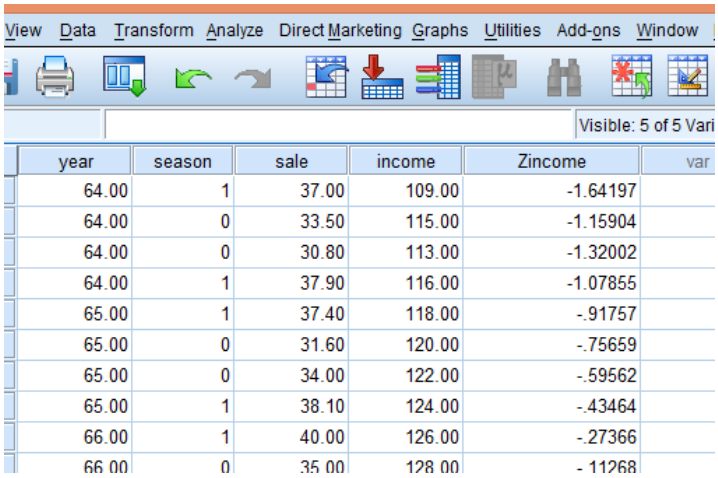
به عنوان مثال اگر مسیر فوق را برای متغیر income از مثال فصل ها انجام دهیم در صفحه ی جدیدی که SPSS تحت عنوان output باز می کند و در آن نتایج و نمودارها و جدول ها را نمایش می دهد جدولی مطابق شکل ۱۱ مشاهده می کنیم.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
income	20	109.00	151.00	129.4000	12.42409	154.358
Valid N (listwise)	20					

شکل۱۱

در این جدول شاخص های آمار توصیفی که انتخاب کرده ایم به نمایش درآمده اند. اگر به صفحه ی Data editor مثال فصلها باز گردیم متغیر جدیدی تحت عنوان Zincome مشاهده می کنیم که همان مقادیر استاندارد شده ی متغیر income می باشد.



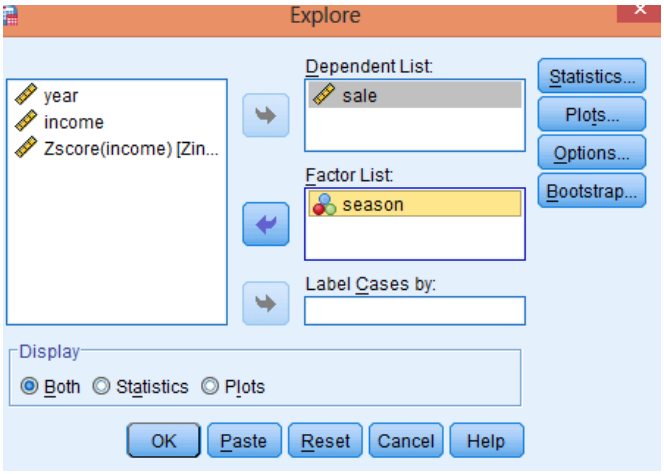
year	season	sale	income	Zincome	var
64.00	1	37.00	109.00	-1.64197	
64.00	0	33.50	115.00	-1.15904	
64.00	0	30.80	113.00	-1.32002	
64.00	1	37.90	116.00	-1.07855	
65.00	1	37.40	118.00	-.91757	
65.00	0	31.60	120.00	-.75659	
65.00	0	34.00	122.00	-.59562	
65.00	1	38.10	124.00	-.43464	
66.00	1	40.00	126.00	-.27366	
66.00	0	35.00	128.00	-.11268	

شکل ۱۲

این مطلب در شکل ۱۲ نشان داده شده است. پس برای به دست آوردن مقادیر استاندارد شده ی یک متغیر از گزینه ی Analyze، Descriptive Statistics را انتخاب کرده و سپس Descriptive را می زنیم و در پنجره ی باز شده گزینه ی Save standard values را تیک می زنیم.

دستور Explore

این دستور که از مسیر Analyze وسپس descriptive statistics وسپس explore قابل دسترسی می باشد به ما این امکان را می دهد که برای گروه های مختلف موجود در یک متغیر به طور جداگانه آمار توصیفی را به دست آوریم. برای مثال اگر بخواهیم در مثال فصل شاخص های آماری مربوط به میزان فروش را در فصل های مختلف که به دو دسته ی گرم و سرد تقسیم شده اند به دست



آوریم باید بر روی Explore کلیک کنیم.

در پنجره ای که مطابق شکل ۱۴ باز می شود متغیر وابسته را میزان فروش (sale) انتخاب کرده و متغیر گروه بندی شده را فصل (season) انتخاب می نماییم. در گوشه ی بالای سمت راست پنجره گزینه های Statistics و Plots وجود دارد که می توان در آنها شاخص های آماری و نمودارهای مورد نظر را انتخاب نمود. نهایتاً جدولی مطابق شکل ۱۵ در صفحه ی output به نمایش در می آید که ویژگی های آماری متغیر فروش را براساس فصل های مختلف سال «گرم و سرد» جداگانه محاسبه کرده است.

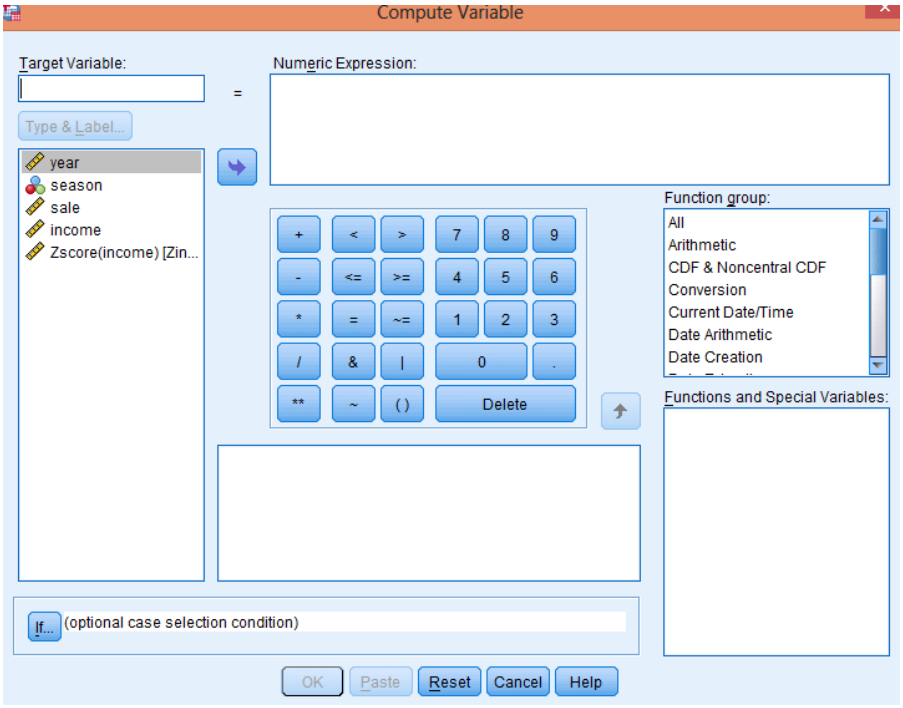
Descriptives

	season			Statistic	Std. Error
sale	summer,fall	Mean		35.2100	.98471
		95% Confidence Interval for Mean	Lower Bound	32.9824	
			Upper Bound	37.4376	
		5% Trimmed Mean		35.1667	
		Median		34.8000	
		Variance		9.697	
		Std. Deviation		3.11393	
		Minimum		30.80	
		Maximum		40.40	
		Range		9.60	
		Interquartile Range		5.48	
		Skewness		.362	.687
		Kurtosis		-.732	1.334
	sprig,winter	Mean		39.5800	.93188
		95% Confidence Interval for Mean	Lower Bound	37.4719	
			Upper Bound	41.6881	
		5% Trimmed Mean		39.5556	
		Median		39.0500	
		Variance		8.684	
		Std. Deviation		2.94686	
		Minimum		35.40	
		Maximum		44.20	
		Range		8.80	
		Interquartile Range		5.05	
		Skewness		.395	.687
		Kurtosis		-.988	1.334

شکل۱۵

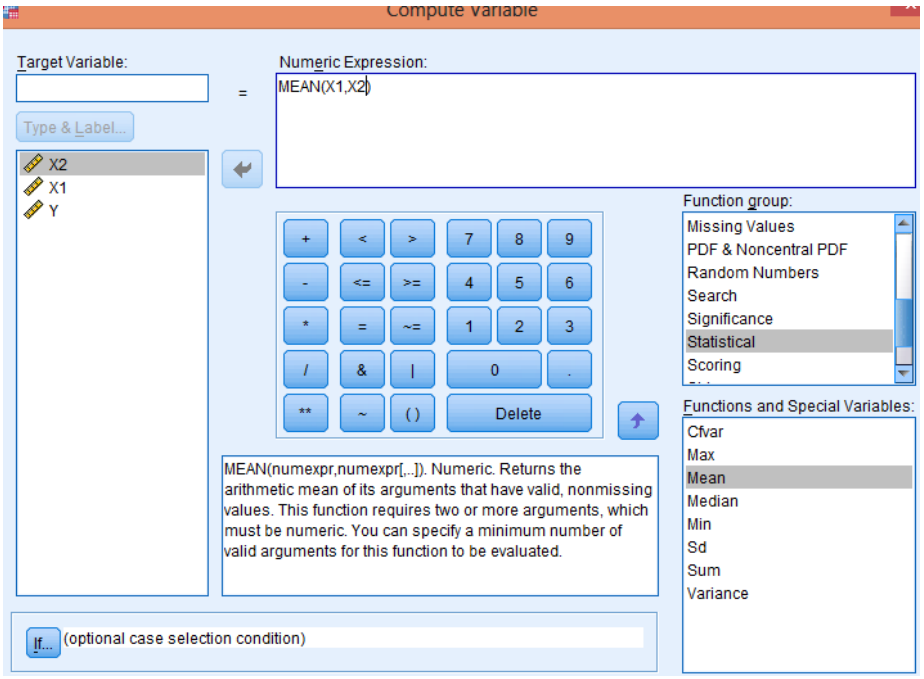
محاسبه ی متغیرهای جدید

گاهی ممکن است لازم باشد که بر پایه ی مقادیر متغیرهای موجود متغیرهای جدیدی محاسبه شود. برای این کار روی کلمه ی Transform در نوار منو کلیک کرده و پس از آن گزینه ی compute را انتخاب می نماییم.



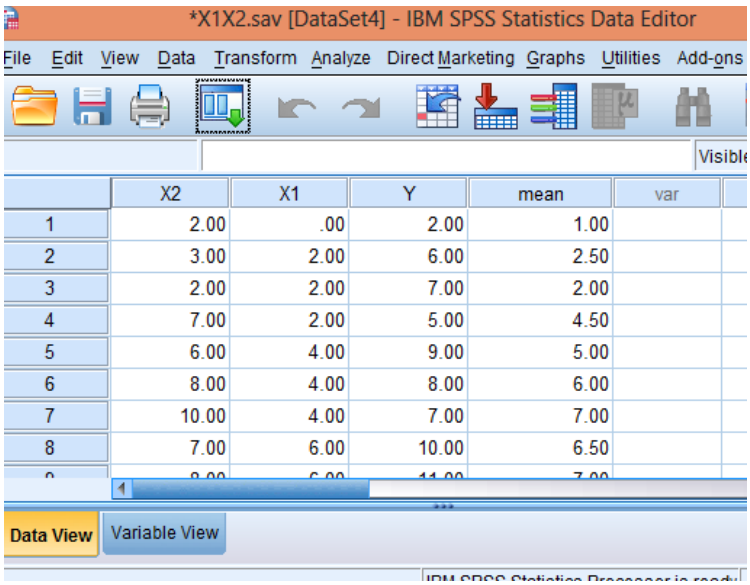
شکل۱۷

پنجره ی جدیدی مطابق شکل ۱۷ باز می شود که در آن می توان متغیر جدید و نحوه ی محاسبه ی آن را تعریف کرد. نام متغیر جدید در قسمت Target Variable وارد شده و نوع و بر چسب آن را نیز می توان در قسمت پایین نام تعریف کرد. در سمت چپ تحت عنوان Function group انواع مختلف توابع ریاضی که SPSS توانایی محاسبه با آنها را دارد مشاهده می کنیم.



شکل۱۸

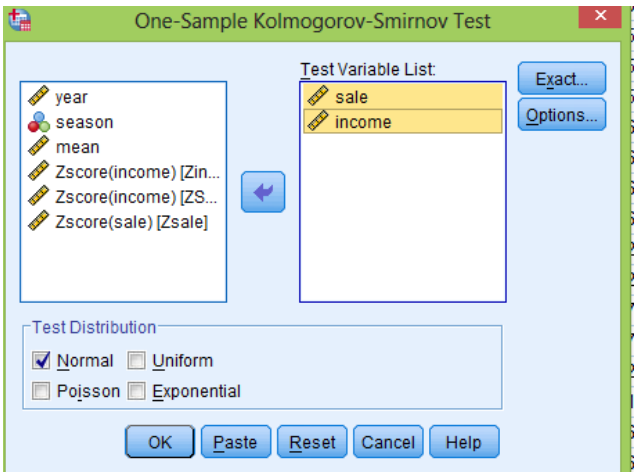
برای مثال اگر گزینه ی Statistical را انتخاب کنیم در کادر پایینی انواع عملگرهای آماری موجود مشاهده می شود. اگر مایل باشیم میانگین دو متغیر x_1 و x_2 را به عنوان متغیر جدیدی با نام Mean محاسبه کنیم کافی است مطابق شکل ۱۸ پس از انتخاب تابع Mean و انتقال آن به محدوده ی Numeric Expression متغیرهای x_1 و x_2 را به پرانتز جلوی این تابع منتقل کرده و دکمه ی ok را بزنیم. اکنون در پنجره ی Data editor متغیر جدید Mean که میانگین متغیرهای x_1 و x_2 می باشدمطابق شکل ۱۹ به نمایش درآمده است.



ضریب همبستگی

اگر مایل باشیم رابطه ی دو متغیر را مورد بررسی قرار دهیم می توانیم از انواع مختلف ضریب همبستگی استفاده کنیم. ضرایب همبستگی مربوط به داده های کمی پیرسون و اسپیرمن می باشد که توسط SPSS قابل انجام است.

برای داده های پارامتریک از ضریب همبستگی پیرسون و برای داده های غیر پارامتریک از ضریب همبستگی اسپیرمن استفاده می کنیم. بنابراین پیش از انتخاب نوع آزمون همبستگی باید نرمال بودن توزیع متغیرها را بررسی کرد. در صورتی که یک یا هر دو متغیر دارای توزیع نرمال نبود باید از ضریب همبستگی اسپیرمن استفاده نمود. در SPSS روش رایج برای بررسی نرمال بودن توزیع متغیرها آزمون کولموگروف اسمیرنف می باشد. این آزمون از طریق منوی Analyze وسپس Nonparametric tests وسپس legacy dialogsوانتخاب 1-sample K-S قابل دستیابی می باشد. برای مثال فرض کنید که در مثال فصل ها می خواهیم ضریب همبستگی متغیرهای income و sale را به دست آوریم.



شکل ۲۰

ابتدا باید نرمال بودن آنها را بررسی کنیم. بعد از طی مسیر ذکرشده و انتخاب متغیرهای sale و income مانند شکل ۲۰ جدولی مطابق با شکل ۲۱ در پنجره ی output به نمایش در می آید. در سطر آخر این جدول میزان خطای نوع اول مربوط به هر دو متغیر به نمایش درآمده که چون هر دو مقدار از ۰/۰۵ بیشتر است فرض صفر هر دو متغیر «فرض نرمال بودن توزیع ها» رد نمی شود. در صورتی که مقدار Sig متغیری از ۰/۰۵ کوچکتر می شد فرض صفر آن متغیر «فرض نرمال بودن توزیع» رد می شد. یعنی متغیر دارای توزیع غیر نرمال بود و برای سنجش همبستگی آن با متغیر دیگر باید از ضریب همبستگی اسپیرمن استفاده می شد.

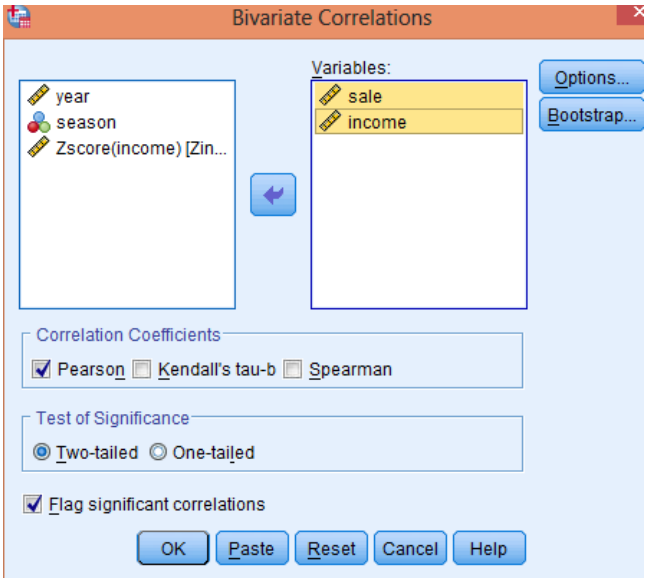
One-Sample Kolmogorov-Smirnov Test

		sale	income
N		20	20
Normal Parameters[a,b]	Mean	37.3950	129.4000
	Std. Deviation	3.70568	12.42409
Most Extreme Differences	Absolute	.105	.075
	Positive	.105	.075
	Negative	-.059	-.072
Kolmogorov-Smirnov Z		.469	.337
Asymp. Sig. (2-tailed)		.980	1.000

a. Test distribution is Normal.

b. Calculated from data.

پس از تعیین نرمال یا غیر نرمال بودن توزیع ها برای به دست آوردن ضریب همبستگی مناسب از منوی Analyze گزینه ی Correlate و سپس Bivariate را مطابق شکل ۲۲ انتخاب کرده و در پنجره ی پدیدار شده متغیرها و نوع ضریب همبستگی مورد نظر را انتخاب کرده و با یک ماتریس همبستگی مشابه شکل ۲۳ در پنجره ی output روبرو می شویم. ضریب همبستگی پیرسون بین متغیرهای income و sale برابر با ۰/۵۰۶ می باشد و از آنجا که مقدار خطای نوع اول این آزمون از ۰/۰۵ کوچکتر باشد (Sig=۰/۰۲۳) فرض صفر آن مبنی بر تصادفی و شانسی بودن این همبستگی رد شده و این همبستگی از لحاظ آماری معنادار است. در صورتی که تعداد متغیرهای بیشتری را جهت آزمون همبستگی انتخاب کنیم ماتریس همبستگی حاصل بزرگتر خواهد شد. مراحل و خروجی های مربوط به ضریب همبستگی اسپیرمن نیز مشابه مطالب گفته شده ی قبلی می باشد.



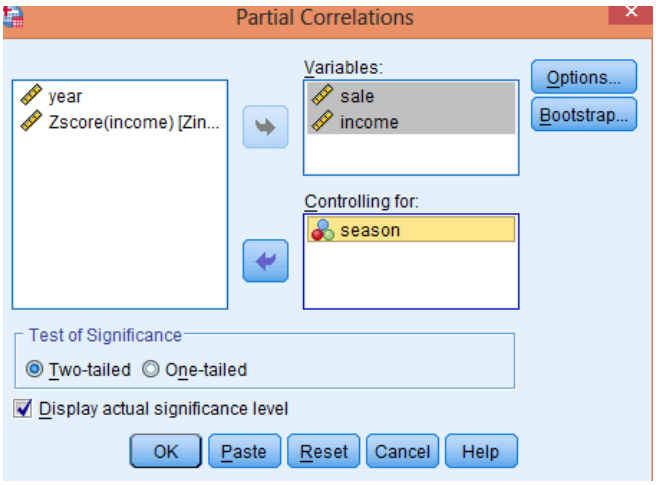
شکل ۲۲

Correlations

		sale	income
sale	Pearson Correlation	1	.506*
	Sig. (2-tailed)		.023
	N	20	20
income	Pearson Correlation	.506*	1
	Sig. (2-tailed)	.023	
	N	20	20

شکل۲۳

نوع دیگری از ضریب همبستگی که SPSS قادر به محاسبه ی آن می باشد، ضریب همبستگی تفکیکی "Partial" است که میزان همبستگی دو متغیر کمی را در حضور متغیر سوم اندازه گیری می کند. برای مثال اگر بخواهیم میزان همبستگی متغیرهای income و sale را در حضور عامل تأثیر گذاری همچون متغیر Season بررسی کنیم کافی است نوع همبستگی را به جای Bivariateنوع Partial انتخاب نماییم و در پنجره ی ظاهر شده مطابق شکل ۲۵ عامل تأثیر گذار را در قسمت Controlling for متغیر Season انتخاب کنیم.



شکل۲۵

پس از آن با جدولی مشابه شکل ۲۶ در پنجره ی output روبرو می شویم که مقدار همبستگی متغیرهای Sale و income را در حضور عامل Season برابر با ۰/۶۴۸ نشان می دهد، از آنجا که Sig =۰/۰۳ می باشد این همبستگی معنادار است. ضمن اینکه با مقایسه ی این ضریب و ضریب به دست آمده بدون حضور عامل Season «۰/۵۰۶» می توان نتیجه گیری کرد که وجود عامل Season موجب شدیدتر شدن همبستگی دو متغیر income و Sale شده است.

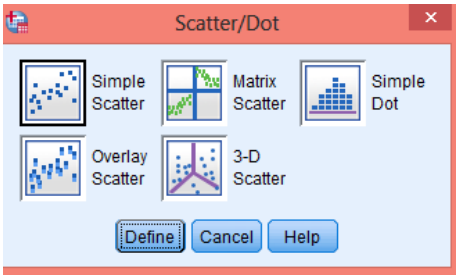
Correlations

Control Variables			sale	income
season	sale	Correlation	1.000	.648
		Significance (2-tailed)	.	.003
		df	0	17
	income	Correlation	.648	1.000
		Significance (2-tailed)	.003	.
		df	17	0

شکل۲۵

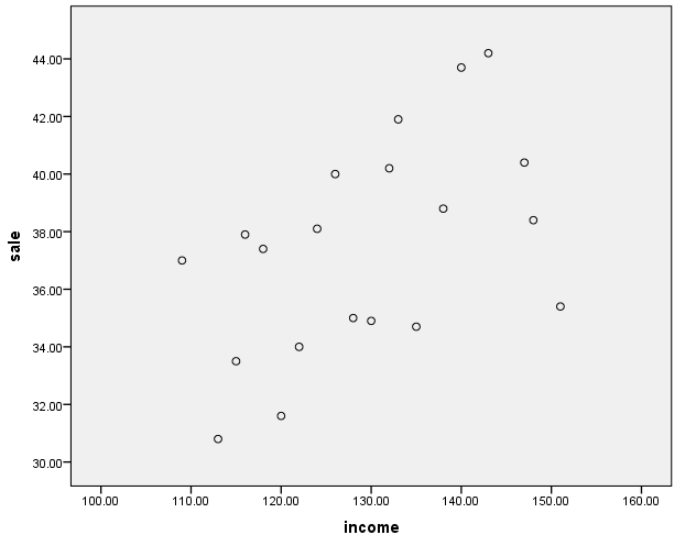
نمودارها در SPSS

این نرم افزار توانایی رسم انواع مختلفی از نمودارکه از مسیر Graphs در نوار منو قابل دسترسی میباشند را دارد. برای مثال اگر پس از طی مسیر Graphs/legacy dialogنمودار Scatter را انتخاب کنیم می توان نمودار پراکنش متغیرهای مورد نظر را رسم کرده و تا حدی به میزان همبستگی بین آنها پی برد.



شکل ۲۷

در شکل ۲۷ نوع نمودار Scatter/dot مورد نظر را انتخاب کرده و با دکمه ی define متغیرهای هر محور را مشخص می نماییم.اگر تعداد متغیرهایی که می خواهیم نمودار پراکنش آنها را رسم کنیم زیاد باشد می توان از مدل Matrix Scatter استفاده کرد. برای نمونه نمودار Simple Scatter را برای متغیر Sale روی محور yها و متغیر income روی محور xها رسم کرده و با نموداری مشابه شکل ۲۸ مواجه می شویم.

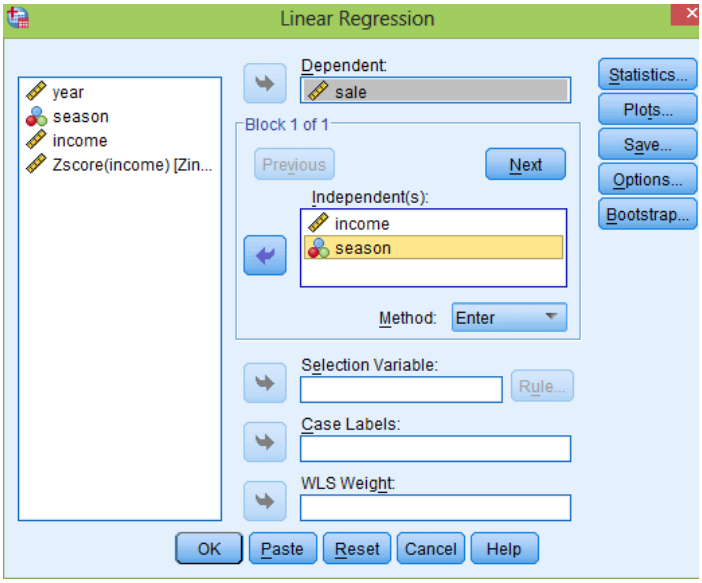


از نحوه ی پراکندگی داده ها در نمودار پراکنش تا حدودی می توان به میزان همبستگی آنها پی برد. اگر بتوان خط مستقیمی را طوری رسم کرد که اکثر نقطه ها نزدیک به آن باشند می توان حدس زد که دو متغیر به شیوه ی خطی با هم رابطه دارند.

رگرسیون

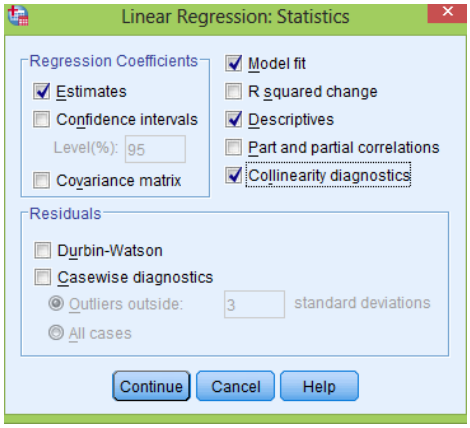
هر گاه لازم باشد جهت و شدت تأثیر یک یا چند متغیر مستقل را بر یک یا چند متغیر وابسته بررسی کنیم و نهایتاً به معادله ای برای پیش بینی متغیر وابسته برسیم باید از آزمون رگرسیون استفاده کنیم. اگر تنها یک متغیر وابسته داشته باشیم و رابطه ی بین متغیر وابسته و متغیرهای مستقل از نوع خطی باشد و خطاهای اندازه گیری دارای توزیع نرمال با میانگین صفر و واریانس ۱ باشند می توانیم از رگرسیون چندگانه خطی استفاده کنیم که مسیر رسیدن به آن Analyze/Regression/Linear است. پس از طی این مسیر پنجره ای مطابق با شکل ۳۰ باز می شود، که در قسمت بالا متغیر وابسته و در پایین متغیرهای مستقل را تعریف می کنیم. برای نمونه در مثال فصل ها میزان فروش (Sale) را به عنوان متغیر وابسته و درآمد (income) و فصل (Season) را به عنوان متغیرهای مستقل تعریف می نماییم. در پایین متغیرهای مستقل گزینه ی Method را مشاهده می کنیم که به صورت پیش فرض Enter در آن نوشته شده است. در SPSS روش های مختلفی برای ورود متغیرهای مستقل به معادله وجود دارد که از آن جمله می توان به روش های هم زمان "Enter"

گام به گام "Stepwise"، جلو رونده "Forward" و عقب رونده "Backward" اشاره کرد.



شکل ۳۰

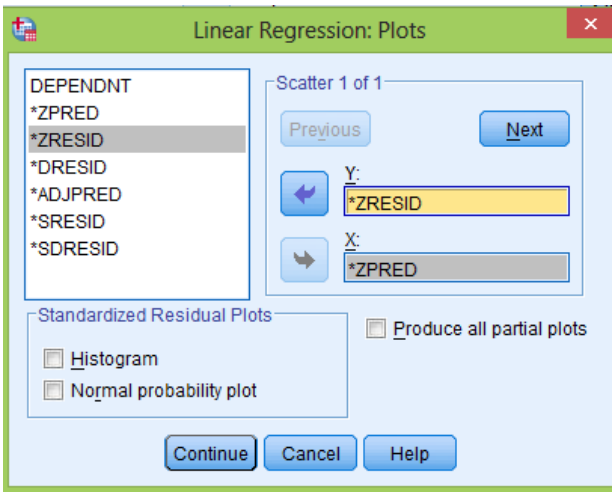
در روش Enter تمام متغیرهای مستقل هم زمان وارد معادله می شوند. زمانی که تعداد متغیرهای مستقل کم باشد و یا پژوهشگر هیچ گونه نظری در مورد اهمیت بیشتر برخی از متغیرهای مستقل بر سایرین نداشته باشد بهتر است از روش Enter استفاده شود.



شکل ۳۱

در گوشه ی بالا سمت راست گزینه های Statistics و Plots به چشم می خورند که با کلیک بر روی آنها پنجره های مشابه شکل های ۳۱ و ۳۲ ظاهر می شوند. در پنجره ی Statistics می توان با تیک زدن گزینه ی Descriptive ویژگی های آمار توصیفی را به دست آورد. همچنین گزینه ی Collinearity diagnostics متغیرهای مستقل را از لحاظ مشکل هم خطی بررسی می کند،در پنجره

ی Plots می توانیم نمودارهای مفیدی همچون نمودار باقی مانده های استاندارد شده "ZResid" در برابر مقدار پیش بینی شده استاندارد را" ZPred"را درخواست نماییم.



شکل ۳۲

در اولین قسمت خروجی که در شکل ۳۳ به نمایش درآمده است جداول مربوط به آمار توصیفی و سپس ماتریس همبستگی متغیرهای موجود در رگرسیون را مشاهده می کنیم.

Descriptive Statistics

	Mean	Std. Deviation	N
sale	37.3950	3.70568	20
income	129.4000	12.42409	20
season	.50	.513	20

Correlations

		sale	income	season
Pearson Correlation	sale	1.000	.506	.605
	income	.506	1.000	-.017
	season	.605	-.017	1.000
Sig. (1-tailed)	sale	.	.011	.002
	income	.011	.	.472
	season	.002	.472	.
N	sale	20	20	20
	income	20	20	20
	season	20	20	20

شکل ۳۳

در شکل ۳۴ بالاترین جدول که جدول مهمی می باشد مقادیر R و R^2 را به نمایش می گذارد. R ضریب همبستگی بین مقدار واقعی متغیر وابسته و مقدار پیش بینی شده ی آن است. R^2 یا ضریب تعیین که مقدار آن در مثال فصل ها ۰/۶۳ شده نشان می دهد که ۶۳ درصد از تغییرات ایجاد شده در متغیر وابسته توسط متغیرهای مستقل موجود در این مدل رگرسیون قابل توضیح می باشند. R^2 اصلاح شده که مقدار آن کمی کوچکتر از R^2 است و تعداد متغیرهای مستقل و تعداد مشاهدات را لحاظ می کند نشان دهنده ی توان مدل رگرسیون می باشد. هر چقدر مقدار R^2 اصلاح شده بیشتر باشد توان مدل رگرسیون در پیش بینی متغیر وابسته بالاتر خواهد بود.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.795[a]	.632	.589	2.37555

a. Predictors: (Constant), season, income

b. Dependent Variable: sale

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	164.974	2	82.487	14.617	.000[b]
	Residual	95.935	17	5.643		
	Total	260.910	19			

a. Dependent Variable: sale

b. Predictors: (Constant), season, income

شکل ۳۴

در جدول ANOVA شکل ۳۴ به بررسی معناداری و مقبولیت مدل رگرسیون پرداخته می شود. چنانچه در این جدول مقدار عددی Sig کمتر از ۰/۰۵ باشد می توان نتیجه گرفت که مدل رگرسیون به طرز معناداری قابلیت پیش بینی متغیر وابسته را دارد.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	15.258	5.735		2.660	.016		
	income	.154	.044	.516	3.509	.003	1.000	1.000
	season	4.432	1.063	.613	4.171	.001	1.000	1.000

a. Dependent Variable: sale

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	income	season
1	1	2.612	1.000	.00	.00	.05
	2	.383	2.610	.00	.00	.94
	3	.004	24.554	1.00	1.00	.00

a. Dependent Variable: sale

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	32.6544	42.9361	37.3950	2.94667	20
Residual	-7.53610	2.51128	.00000	2.24705	20
Std. Predicted Value	-1.609	1.880	.000	1.000	20
Std. Residual	-3.172	1.057	.000	.946	20

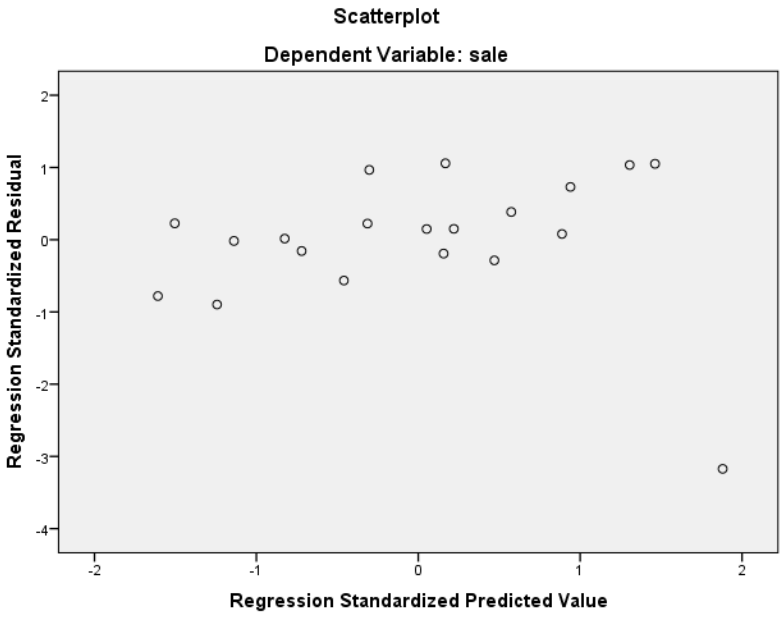
a. Dependent Variable: sale

شکل۳۵

شکل ۳۵ در قسمت بالا جدول ضرایب رگرسیون را نشان می دهد. در ستون B به ترتیب مقدار ثابت B_0 و سپس ضرایب متغیرهای مستقل موجود در مدل رگرسیون را مشاهده می کنیم. ستون Beta مقادیر استاندارد شده ی ضرایب معادله رگرسیون می باشند. مقدار Beta نشان دهنده ی تغییری است که در متغیر وابسته ایجاد می شود اگر در متغیر مستقل به اندازه ی یک انحراف استاندارد تغییر ایجاد شود. برای مثال اگر در متغیر income به اندازه ی یک انحراف استاندارد افزایش صورت گیرد در متغیر وابسته Sale به اندازه ی ۰/۵۱۶ انحراف استاندارد افزایش صورت می گیرد. از آنجا که ضریب بتا استاندارد شده است می توان بین مقادیر بتا برای متغیرهای مستقل مقایسه انجام داد. برای نمونه در مثال فصل ها از آنجا که ضریب بتا برای متغیر Season بزرگتر از ضریب بتا برای متغیر income می باشد می توان نتیجه گرفت سهم متغیر Season در پیش بینی متغیر وابسته ی Sale بیشتر از سهم متغیر income می باشد. مقادیر ستون Sig معناداری ضرایب تأثیر مدل رگرسیون را نشان می دهند. در این مدل چون همه ی مقادیر sig کمتر از ۰/۰۵ می باشند نشان می دهد که تأثیر متغیرهای مستقل بر متغیر وابسته تصادفی نبوده و از لحاظ آماری معنادار است. دو ستون آخر مربوط

به مساله هم خطی میان متغیرهای مستقل است. مقادیر موجود در ستون VIF معکوسی مقادیر ستون Tolerance می باشد. اگر برای متغیری مقدار Tolerance کوچکتر از R^2-۱ شود احتمالاً مشکل هم خطی با سایر متغیرهای مستقل پیدا خواهد کرد. که در این مثال این مسأله به چشم نمی خورد. جدول بعدی شکل ۳۵ نیز مربوط به هم خطی متغیرهای مستقل می باشد. در این جدول مقادیر Eigen Value به چشم می خورد. هر گاه مقدار ویژه ای بسیار کوچک و نزدیک به صفر باشد باعث می شود Condition Index مربوط به آن بسیار بزرگ گردد. این حالت زمانی به وجود می آید که مشکل هم خطی بین متغیرهای مستقل وجود داشته باشد و تغییرات جزیی و ناچیزی در متغیرهای مستقل به تغییرات بسیار بزرگی در براورد ضرایب معادله رگرسیون منتهی شود.

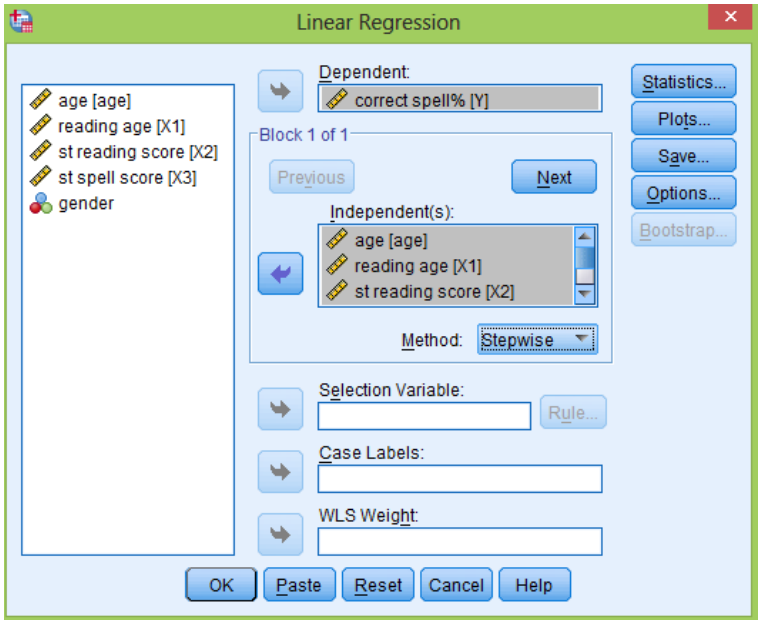
پایین ترین جدول شکل ۳۵ آماره های مربوط به مقادیر پیش بینی شده و باقیمانده ها را در حالت عادی و استاندارد شده به نمایش می گذارد. همان گونه که در فرضیات مدل رگرسیون خطی هم ذکر شد باقیمانده ها دارای توزیع نرمال با میانگین صفر و واریانس ۱ می باشند. در شکل ۳۶ نمودار مقادیر استاندارد شده ی باقی مانده ها در برابر مقادیر استاندارد شده ی پیش بینی شده برای متغیر وابسته نشان داده می شود. در این مثال نقطه ها پراکندگی تصادفی حول و حوش مقدار صفر دارند و این نشان می دهد که دارای توزیع نرمال با واریانس ثابت و میانگین صفر می باشند. هر گاه در این نمودار الگو یا روندی مشاهده شود مثلاً با افزایش مقدار y مقدار x هم افزایش یابد نشان دهنده ی این است که فرض های رگرسیون چند گانه ی خطی محقق نشده و باید از سایر مدل های رگرسیون استفاده کرد.



شکل ۳۶

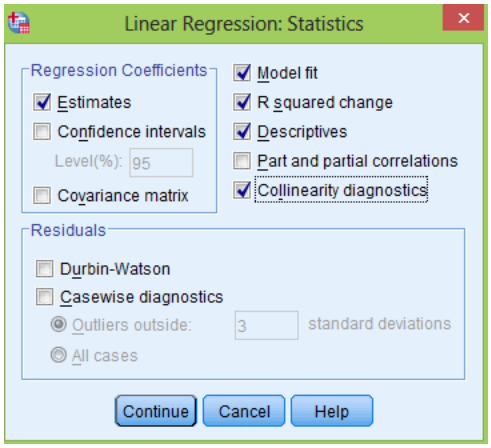
انجام رگرسیون به روش گام به گام (Stepwise)

پیش از این گفتیم که در روش Enter تمام متغیرهای مستقل همزمان وارد مدل رگرسیون می شوند. در روش Stepwise در هر مرحله یکی از متغیرهای مستقل بر مبنای شدت تأثیر و معناداری وارد معادله می شود و این کار تا ورود آخرین متغیری که تأثیر معناداری بر ضریب تعیین R^2 داشته باشد ادامه می یابد. این روش زمانی مفید است که تعداد متغیرهای مستقل بسیار زیاد باشد و پژوهشگر مایل باشد کمترین تعداد متغیرهای مستقل تأثیر گذار را در پیش بینی استفاده کند. برای مثال فرض کنید با ۵ متغیر مستقل سن و جنس و x_1 و x_2 و x_3 و متغیر وابسته ی y مطابق شکل ۳۷ می خواهیم ضرایب معادله رگرسیون را به دست آوریم.



شکل ۳۷

بعد از وارد کردن متغیرهای مستقل و وابسته روش را Stepwise انتخاب کرده و در کادر Statistics در کنار عبارت R square change هم تیک می زنیم (شکل ۳۸)



شکل۳۸

سایر موارد و از جمله نمودارهای مورد نظر را مانند روش Enter و مثال فصل ها انجام می دهیم. اولین جدول خروجی همان شاخص های آمار توصیفی و جدول دوم ماتریس همبستگی متغیرها می باشدکه قبلاً توضیح داده شد. (شکل های ۳۹ و ۴۰)

Descriptive Statistics

	Mean	Std. Deviation	N
correct spell%	56.1500	22.43886	20
age	86.7500	4.19116	20
reading age	82.8500	17.91875	20
st reading score	98.5000	17.59635	20
st spell score	108.5500	9.58329	20
gender	.5000	.51299	20

شکل۳۹

		correct spell%	age	reading age	st reading score	st spell score	gender
Pearson Correlation	correct spell%	1.000	.193	.307	.707	.767	-.281
	age	.193	1.000	-.030	-.296	-.184	-.233
	reading age	.307	-.030	1.000	.432	.186	.089
	st reading score	.707	-.296	.432	1.000	.631	-.140
	st spell score	.767	-.184	.186	.631	1.000	-.091
	gender	-.281	-.233	.089	-.140	-.091	1.000
Sig. (1-tailed)	correct spell%	.	.207	.094	.000	.000	.115
	age	.207	.	.450	.103	.219	.162
	reading age	.094	.450	.	.028	.217	.355
	st reading score	.000	.103	.028	.	.001	.278
	st spell score	.000	.219	.217	.001	.	.351
	gender	.115	.162	.355	.278	.351	.
N	correct spell%	20	20	20	20	20	20
	age	20	20	20	20	20	20
	reading age	20	20	20	20	20	20
	st reading score	20	20	20	20	20	20
	st spell score	20	20	20	20	20	20
	gender	20	20	20	20	20	20

شکل۴۰

جدول بعدی (شکل ۴۱) ترتیب ورود متغیرهای وابسته به مدل رگرسیون را نشان می دهد. در این مثال در هر مرحله یک متغیر جدید وارد

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	st spell score	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	age	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	st reading score	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: correct spell%

شکل ۴۱

شده و هیچ متغیری خارج نشده است. SPSS هر زمان که یک متغیر جدید وارد می شود یک مدل جدید می سازد. در مدل اول متغیر x_3 یا st spell score و در مدل دوم متغیر سن و در مدل سوم متغیر x_2 یا st reading score وارد شده اند.

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.767[a]	.588	.565	14.80181	.588	25.664	1	18	.000
2	.839[b]	.703	.669	12.91731	.116	6.635	1	17	.020
3	.920[c]	.847	.818	9.56357	.144	15.014	1	16	.001

a. Predictors: (Constant), st spell score

b. Predictors: (Constant), st spell score, age

c. Predictors: (Constant), st spell score, age, st reading score

d. Dependent Variable: correct spell%

شکل ۴۲

شکل ۴۲ جدول بعدی را نشان می دهد که مدل های رگرسیون انجام شده را با هم مقایسه می کند. ستون R Square change نشان می دهد که بعد از ورود متغیر سن ضریب تعیین مدل رگرسیون ۰/۱۱۶ افزایش داشته و بعد از ورود متغیر x_2 مجددا به مقدار۰.۱۴۴ افزایش یافته است. یعنی ورود متغیرهای سن و x_2 توان مدل رگرسیون را بهبود بخشیده اند. مقایسه مقادیر R^2 اصلاح شده نیز مؤید همین مطلب می باشد که با مدل رگرسیون شامل سه متغیر مستقل x_2 و x_3 می توان ۸۴ درصد ازتغییرات متغیر وابسته ی y را پیش بینی نمود. تحلیل جدول ANOVA موجود در شکل ۴۳ نیز بیانگر این مطلب است که با توجه به مقادیر Sig کمتر از ۰/۰۵ هر سه مدل رگرسیون استفاده شده از لحاظ آماری معنا دارند.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5622.864	1	5622.864	25.664	.000[b]
	Residual	3943.686	18	219.094		
	Total	9566.550	19			
2	Regression	6729.982	2	3364.991	20.167	.000[c]
	Residual	2836.568	17	166.857		
	Total	9566.550	19			
3	Regression	8103.161	3	2701.054	29.532	.000[d]
	Residual	1463.389	16	91.462		
	Total	9566.550	19			

a. Dependent Variable: correct spell%

b. Predictors: (Constant), st spell score

c. Predictors: (Constant), st spell score, age

d. Predictors: (Constant), st spell score, age, st reading score

شکل۴۳

شکل ۴۴ جدول ضرائب رگرسیون در سه مدل مختلف را نشان می دهد که تفسیر آن مشابه روش Enter است و پیشتر گفته شد.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-138.707	38.606		-3.593	.002		
	st spell score	1.795	.354	.767	5.066	.000	1.000	1.000
2	(Constant)	-315.611	76.496		-4.126	.001		
	st spell score	1.944	.315	.830	6.180	.000	.966	1.035
	age	1.853	.719	.346	2.576	.020	.966	1.035
3	(Constant)	-346.005	57.176		-6.052	.000		
	st spell score	1.243	.295	.531	4.212	.001	.602	1.660
	age	2.354	.548	.440	4.295	.001	.912	1.096
	st reading score	.641	.165	.502	3.875	.001	.569	1.758

a. Dependent Variable: correct spell%

شکل۴۴

شکل ۴۵ اطلاعات آماری متغیرهایی را نشان می دهد که از هر کدام از مدل ها حذف شده اند.

Excluded Variables[a]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
						Tolerance	VIF	Minimum Tolerance
1	age	.346[b]	2.576	.020	.530	.966	1.035	.966
	reading age	.170[b]	1.112	.282	.260	.966	1.036	.966
	st reading score	.371[b]	2.069	.054	.448	.602	1.660	.602
	gender	-.213[b]	-1.445	.167	-.331	.992	1.008	.992
2	reading age	.169[c]	1.278	.219	.304	.966	1.036	.934
	st reading score	.502[c]	3.875	.001	.696	.569	1.758	.569
	gender	-.135[c]	-.983	.340	-.239	.927	1.078	.904
3	reading age	.005[d]	.046	.964	.012	.790	1.266	.465
	gender	-.067[d]	-.639	.532	-.163	.898	1.113	.551

a. Dependent Variable: correct spell%

b. Predictors in the Model: (Constant), st spell score

c. Predictors in the Model: (Constant), st spell score, age

d. Predictors in the Model: (Constant), st spell score, age, st reading score

شکل۴۵

ملاحظه می شود که در مدل ۱ چهار متغیر ,در مدل ۲ سه متغیر و در مدل ۳ دو متغیر خارج از معادله ی رگرسیون باقی مانده اند.

شکل ۴۶ مسائل مربوط به هم خطی در مدل های ۳ گانه را به طور جداگانه بررسی کرده که تحلیل آن پیشتر گفته شد.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				(Constant)	st spell score	age	st reading score
1	1	1.996	1.000	.00	.00		
	2	.004	23.285	1.00	1.00		
2	1	2.993	1.000	.00	.00	.00	
	2	.006	22.335	.01	.75	.11	
	3	.001	58.490	.99	.25	.89	
3	1	3.974	1.000	.00	.00	.00	.00
	2	.022	13.470	.01	.00	.02	.50
	3	.004	33.581	.01	.92	.08	.48
	4	.001	68.072	.98	.08	.90	.02

a. Dependent Variable: correct spell%

شکل ۴۷ اطلاعات مربوط به باقی مانده ها و مقادیر پیش بینی شده است که در روش Enter هم توضیح داده شد.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	22.4802	95.6434	56.1500	20.65144	20
Residual	-13.67290	15.49786	.00000	8.77613	20
Std. Predicted Value	-1.630	1.912	.000	1.000	20
Std. Residual	-1.430	1.621	.000	.918	20

a. Dependent Variable: correct spell%

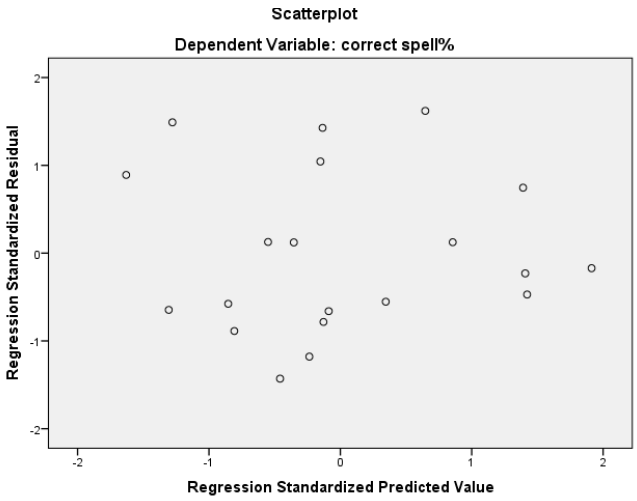
شکل۴۷

در شکل ۴۸ هم نمودار پراکنش باقیمانده ها در برابر مقادیر پیش بینی شده ی حاصل از مدل سوم که کاملترین مدل بود به نمایش درآمده است. پراکندگی در این نمودار کاملاً تصادفی بوده و بیانگر این مطلب است که مدل رگرسیون خطی انتخاب شده مناسب بوده است.

روش های Forward و Backward

در روش پیش رونده متغیرها یکی یکی براساس توان پیش بینی متغیر وابسته وارد مدل شده و اثر اضافه شدن آنها به مدل بررسی می-شود و متغیرهایی که به صورت معناداری توان مدل را افزایش نمی دهند حذف می گردند.

در روش پس رونده تمام متغیرها را به مدل وارد کرده و متغیری که حذف آن کمترین کاهش معنی‌دار را در توان مدل رگرسیون دارد از معادله حذف می‌شود و این کار تا حذف آخرین متغیر ضعیف ادامه می‌یابد. خروجی‌ها و تفسیر آنها در این دو روش مشابه با روش های گفته شده ی قبلی می باشد.



شکل۴۸

تعمیم رگرسیون خطی

رگرسیون خطی حوزه وسیع و پیچیده ای از الگوها را برای جامهٔ عمل پوشیدن به نیازهای بسیاری از تحلیل کنندگان فراهم می نماید. نمی توان انتظار داشت که رگرسیون خطی برای تمام مسائل مناسب باشد. بعضی اوقات پاسخ و پیش بینی ها از طریق تابع غیر خطی معلومی باهم رابطه دارند و هنگامی که بتوان خطاها را جمعی و نرمال در نظر گرفت نتیجه اش رگرسیون غیر خطی خواهد بود و اختلاف آن با رگرسیون خطی فقط در این است که پاسخ به صورت تابعی خطی از پارامترها تغییر می کند.

رگرسیون غیرخطی

مسئله ساختن الگوی افزایش وزن (y) را بعنوان تابعی از رژیم غذایی به حیوانات جوان (x) را در نظر می گیریم. رابطه بین y و x را چگونه الگوسازی کنیم؟

یک خط راست الگوی رگرسیون خطی ساده احتمالاً به چند دلیل برای این مسئله مناسب نیست. پاسخ احتمالاً از طرف پائین به وسیلهٔ میزان رشدی که بدون تکمیل کردن پیش می آید محدود می شود. شاید پاسخ از سمت بالا به وسیلهٔ ماکزیمم رشد اضافی بیولوژیکی که می تواند به خاطر تکمیل جیرهٔ غذایی باشد محدود شود. خطوط راست نمی توانند این رفتار را داشته باشند زیرا آنها به ازاء هر واحد افزایش در پیش بینی به میزان ثابتی افزایش پیدا می کنند و سرانجام مرزهای پائینی یا بالایی را قطع می کنند. یک منحنی به شکل S که سه پارامتر دارد ممکن است این رفتار را به اندازه کافی بیان کند. این سه پارامتر به مقدار می نیمم یا عرض از مبدا، مقدار اضافه شده که مجانب نامیده می شود و پارامتر سومی که میزان افزایش از می نیمم به ماکزیمم را کنترل می کند مربوط می شوند. الگویی که دارای این خواص است به صورت زیر می باشد.

$$e_i = \theta_1 + \theta_2[1 - exp(\theta_3 x_i)] + e_i \qquad *$$

به ازای $\theta_3 < 0$ اگر xi زیاد رشد کند پاسخ yi بصورت یک مجانب به $\theta_1 + \theta_2$ میل میکند و به ازای xi=0 انتظار داریم yi برابر θ_1 شود. پارامتر سوم θ_3 پارامتر میزان است. معادله* فقط یکی از الگوهای ممکن برای رگرسیون مجانبی است و رگرسیون مجانبی فقط یکی از الگوهای غیر خطی است. معهذا جنبه های مهم نمونه الگوهای غیر خطی را نمایش می دهد:

(۱) تابعی که پاسخ را به پیش بینی ها مربوط می کند تابعی غیرخطی از پارامترهاست. الگوی * نسبت به θ_3 غیرخطی است.

(۲) برخلاف یک الگوی خطی لازم نیست رابطه ای مستقیم بین پیش بینی ها و پارامترها وجود داشته باشد. در * فقط یک پیش بینی ولی سه پارامتر وجود دارد.

(۳) چون پارامتری کردن منحصربفرد نیست بنابراین الگوهای رگرسیون غیرخطی زیادی معادل هم هستند.

برآورد

روش استاندارد برآورد پارامترها در رگرسیون غیرخطی، کمترین مربعات است. برای الگوی غیر خطی کلی

$$y_i = f(x_i;\theta) + e_i$$

که در آن x_i بردار پیش بینی ها ، θ بردار پارامترهای ،f تابعی غیر خطی از θ و $var(e_i) = \sigma^2/w_i$ که w_iمعلوم است، برآوردکننده$\hat{\theta}$ مقداری از θ است که تابع مجموع مربعات باقیمانده موزون را می نیمم کند،

$$RSS(\theta) = \sum_{i=1}^{n} w_i[y_i - f(x_i;\theta)]^2$$

اگر e_i ها مستقل و $N(0,\sigma^2/w_i)$ باشند آنگاه $\hat{\theta}$ برآورد درستنمایی θ است. برآورد درستنمایی ماکزیمم σ^2برابر $\hat{\sigma}^2 = RSS\,(\hat{\theta})/n$ است که اغلب در مخرج به جای n از n-q استفاده می شود.

رگرسیون لجستیک

همان طور که می دانیم، برای انجام تحلیل رگرسیون خطی، متغیر وابسته باید کمی و در سطح سنجش فاصله ای/ نسبی باشد. اما گاهی اوقات اتفاق می افتد که متغیر وابسته تحقیق در مقیاس فاصله ای / نسبی نبوده و مقیاس آن به صورت اسمی (دو وجهی یا چند وجهی) است . حال، سوال این جا است که برای این کار باید چه کرد، در حالی که پیش فرض اساسی تحلیل رگرسیون، مقیاس فاصله ای / نسبی متغیر وابسته است . در چنین حالتی، نرم افزارهایی این امکان را برای ما فراهم کرده است تا بتوانیم عوامل پیش بینی کننده تغییرات یک متغیر اسمی را نیز شناسایی کنیم . این روش، که رگرسیون لجستیک نام دارد، در اواخر دهه 0431 و اویا دهه 0491 به عنوان بدیلی برای روش رگرسیون خطی و همچنین تحلیل تابع تشخیصی مطرح شد . مثلاً یک تحلیل گر مالی ممکن است مایل باشد با توجه به متغیرهای مستقل مانند فروش، هزینه های تولید، سهم بازار و سود، پیش بینی کند که آیا شرکت در سال آینده ورشکست می شود یا خیر؟ چون مقیاس متغیر وابسته اسمی است، روش حداقل مجموع مربعات برای حل این نوع کاربردها، نامناسب است و به جای آن به ناچار باید از روش دیگری مانند رگرسیون لجستیک استفاده کرد . زمانی که متغیر وابسته در سطح اسمی است و متغیرهای مستقل هم ترتیبی و فاصله ای هستند، روش های رگرسیون خطی معمولی و تحلیل تشخیصی، مقدار برآوردها را کم تر از مقدار واقعی نشان می دهند. بنابراین از رگرسیون لجستیک برای تخمین احتمال وقوع یک رویداد خاص استفاده می شود و متغیر وابسته در آن به معنی(نسبت ناجورها است) odds ratio که روش دیگری برای بیان احتمال است .برای مثال، اگر احتمال وقوع یک رویداد مانند آوردن شیر یا خط در پرتاب یک سکه باشد ۰.۵ باشد، نسبت ناجورهای آن $\frac{۰.۵}{۱-۰.۵} = ۱$ و یا به نسبت یک به یک است.

همیشه به فرض داشتن نسبت ناجورهای یک پدیده، می توان احتمال وقوع آن رویداد را محاسبه کرد. یعنی رابطه احتمال و نسبت ناجورها به این صورت قابل تعریف است:

$$\text{احتمال (جور) موفقیت} = \frac{\text{نسبت ناجور}}{\text{۱ + نسبت ناجور}}$$

اگرچه می توان از رگرسیون لجستیک برای دسته بندی های ترتیبی یعنی طبقه بندی های مرتب شده با دو یا چند حالت (مانند سوالات متداول در پرسشنامه ها) نیز استفاده کرد، اما اگر متغیر وابسته مانند اسم چند شهر بزرگ دارای چندین طبقه بندی های نامرتب باشد ، شما می توانید از تحلیل تفکیک کننده یا چند متغیره استفاده کنید. رگرسیون لجستیک، شبیه رگرسیون خطی است با این تفاوت که نحوه محاسبه ضرایب در این دو روش یکسان نمی باشد . بدین معنی که رگرسیون لجستیک، به جای حداقل کردن مجذور خطاها (کاری که رگرسیون خطی انجام می دهد) ، احتمالی را که یک واقعه رخ می دهد، حداکثر می کند. همچنین، در تحلیل رگرسیون خطی، برای آزمون برازش مدل و معنی دار بودن اثر هر متغیر در مدل، به ترتیب از آماره های F و t استفاده می شود، در حالی که در رگرسیون لجستیک ، از آماره های کای اسکوئر(X^2) و والد(wald) استفاده می شود.

رگرسیون لجستیک نسبت به تحلیل تشخیصی نیز ارجحیت دارد و مهم ترین دلیل آن این است که در تحلیل تشخیصی گاهی اوقات احتمال وقوع یک پدیده خارج از طیف ۰ تا ۱ قرار می گیرد و متغیرهای پیش بین باید دارای توزیع نرمال چند متغیره باشند . در حالی که در رگرسیون لجستیک، احتمال وقوع یک پدیده در داخل محدوده ۰ تا ۱ قرار دارد و رعایت پیش فرض نرمال بودن متغیرهای پیش بین لازم نیست.

مدل رگرسیون لجستیک

تحلیل مدل رگرسیون لجستیک به جز در مورد متغیر وابسته، مشابه رگرسیون خطی است که در آن:

y = نسبت ناجورها

ln(y) = لگاریتم طبیعی

$x_1,x_2,\ldots,x_k$ =متغیرهای مستقل

$B_0,B_1,B_2,\ldots,B_k$ = ضرایب متغیرهای مستقل

e = متغیر خطا است

نخست باید با استفاده از تکنیک های آماری موسوم به حداکثر درستنمایی maximum likelihood ضرایب را تخمین زد و سپس معادله رگرسیون لجستیک را بدست آورد:

$$\ln(\hat{y}) = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

مقدار $\hat{y}$ تخمینی از نسبت ناجورها است که از طریق محاسبه $\ln(\hat{y})$ به دست می آید و سرانجام تخمین احتمال وقوع یک رویداد از تخمین نسبت ناجورها به این صورت است:

$$\text{تخمین احتمال وقوع یک رویداد} = \frac{\hat{y}}{\hat{y}+1}$$

نکته : نرم افزارهای آماری متعددی مانند Minitab ، SAS و SPSS می توانند محاسبات لازم را انجام دهند (اکسل مستقیماً فاقد مدول " Modul " اجرای تحلیل رگرسیون لجستیک است.)

نکته : در حالی که دامنه تغییرات نسبت بخت ها بین ۰ تا ۱ نوسان دارد، دامنه تغییرات لوجیت نسبت بخت ها بین $-\infty$ تا $+\infty$ است.

لوجیت Logit

محوری ترین مفهوم ریاضی در رگرسیون لجستیک، لوجیت است . لوجیت به معنای لگاریتم طبیعی (ln) بخت های متغیر وابسته (Y) می باشد که مدل آن، به مدل لوجیت معروف است .ساده ترین مثال از یک لوجیت را می توان در قالب یک جدول توافقی 2*2 مشاهده کرد . در جدول زیر، توزیع متغیر وابسته شرکت در انتخابات(Y) بر اساس یک متغیر مستقل جنسیت(X) آمده است .در این مثال، لوجیت، لگاریتم طبیعی (ln) بخت های Y است که مقدار Y را از روی X پیش بینی می کند.

متغیر وابسته / متغیر مستقل		جنسیت (X) مرد	جنسیت (X) زن
شرکت در انتخابات (Y)	بله	۹۰ (p_1)	۸۰ (p_0)
	خیر	۲۰ ($1-p_1$)	۱۰ ($1-p_0$)

آماره والد (Wald)

در رگرسیون لجستیک، آماره والد معنی دار بودن حضور هر متغیر مستقل در معادله را نشان می دهد .در نتیجه، آماره والد معادل آماره t در رگرسیون خطی است . آزمون والد از رابطه زیر محاسبه می شود که در آن، β_i به معنای بتا وضریب متغیر X_i و S.E.خطای استاندارد آن است .در واقع، آماره والد این فرض صفر را به آزمون می گذارد که مقدار تمامیβ ها برابر است با ۰ .

یعنی میزان تاثیر تمامی متغیرهای مستقل بر متغیر وابسته برابر صفر است .پس، اگر قرار است فرض صفر را رد کنیم، مقدار حداقل یکی از β ها نباید صفر باشد. نکته ای که وجود دارد این است که زمانی که مقدارβ بزرگ باشد، والد اریب پیدا می کند و موقعی که درجه آزادی یک متغیر برابر با عدد ۱ باشد، در آن صورت مقدار آماره والد از جذر نسبت ضریب رگرسیونی(β) آن متغیر به خطای (S.E.) آن بدست می آید.

اما برای متغیرهای ترتیبی، که درجه آزادی آنها همیشه از عدد یک بیشتر است، درجه آزادی آماره والد یک متغیر برابراست با تعداد طبقات آن متغیر منهای عدد یک (N – 1).

بخت ها (Odds)

بخت ها عبارت است از احتمال رخ دادن یک واقعه بر احتمال رخ ندادن آن واقعه . بخت ها از طریق فرمول زیر محاسبه می شوند که در آن، P_1 احتمال رخ دادن یک واقعه و $1 – p_1$ احتمال رخ ندادن آن واقعه است:

برای درک بهتر مفهوم بخت ها، مثال مربوط به مدل لوجیت در صفحات قبل را تکرار می کنیم .در این مثال، بخت شرکت مردان در انتخابات عبارت است از تعداد مردانی که در انتخابات شرکت کرده اند p_1 ، نسبت به تعداد مردانیکه در انتخابات شرکت نکرده اند $1- p_1$ ، یعنی 90/20=4.5 .در گروه زنان نیز، بخت شرکت زنان در انتخابات برابر است با تعداد زنانی که در انتخابات شرکت کرده اند p_0 ، نسبت به تعداد زنانی که در انتخابات شرکت نکرده اند $1 – p_0$ ، یعنی 80/10=8 .

نسبت بخت ها (Odds ratio)

در رگرسیون لجستیک، برای تعیین میزان تاثیر هر متغیر مستقل بر متغیر وابسته، از آماره ای به نام نسبت بخت ها استفاده می شود .نسبت بخت ها، در واقع نسبت دو بخت به همدیگر است و به معنای نسبت احتمال وقوع یک

پیامد با فرض عضویت در گروه اول به احتمال وقوع آن پیامد با فرض عضویت در گروه دوم می باشد .به عبارتی، نسبت بخت ها نشان دهنده یک واحد تغییر در بخت های وقوع یک پیامد به ازای یک واحد تغییر در متغیر مستقل است .از اینرو می توان نسبت بخت ها را معادل β در رگرسیون خطی دانست که بر اساس فرمول زیر و از طریق تقسیم دو بخت بر همدیگر محاسبه می شود:

$$OR = \frac{p_1/1 - p_1}{p_0/1 - p_0}$$

که در آن

P1 = احتمال وقوع یک پیامد با فرض عضویت در گروه اول (۱)

P0 = احتمال وقوع یک پیامد با فرض عضویت در گروه دوم (۰)

برای درک بهتر نحوه محاسبه نسبت بخت ها، مثال مربوط به مدل لوجیت و بخت ها را بار دیگر تکرار می کنیم .در این مثال، برای محاسبه نسبت بخت ها، ابتدا لازم است بخت شرکت در انتخابات در دو گروه مردان و زنان را محاسبه کنیم .در محاسبه

بخت ها، ملاحظه کردیم که بخت شرکت مردان در انتخابات برابر با 4.5 = 90/20 و بخت آن برای زنان برابر با 8=80/10 است .حال، اگر مقدار دو بخت را بر همدیگر تقسیم کنیم، نسبت بخت ها برابر است با 1.78=8/4.5. یعنی در زنان، نسبت شرکت در انتخابات نزدیک به دو برابر مردان است . دقت داشته باشیم که نسبت بخت ها در فرمول با نماد OR و در خروجی SPSSبا نماد Exp(B) مشخص شده است.

• در تفسیر نتایج نسبت بخت ها، باید قواعد زیر را رعایت کنیم:

۱- هر گاه نسبت بخت ها بزرگتر از عدد 0 باشد، تغییر متغیرهای مستقل و وابسته مثبت و هم جهت است .یعنی باافزایش مقدار متغیر مستقل، مقدار متغیر وابسته نیز افزایش می یابد (در این حالت مقدار B نیز مثبت است)

۲- هر گاه نسبت بخت ها کوچکتر از عدد 0 باشد، تغییر متغیرهای مستقل و وابسته منفی و در جهت مخالف هم است (در این حالت مقدار B نیز منفی است)

۳- هر گاه نسبت بخت ها برابر با عدد 0 باشد، متغیر مستقل تاثیر معنی داری بر متغیر وابسته ندارد و مقدار بتا یا اثرآن صفر است.

• نسبت بخت ها را می توانیم به دو شیوه تفسیر کنیم:

۱- در شیوه اول، همان طور که در بالا اشاره شد، بر اساس نسبت تغییر در متغیر وابسته به ازای یک واحد تغییر درمتغیر مستقل تفسیر کنیم .به عنوان مثال، نسبت بخت های ۱.۷۸ در مثال مربوط به شرکت مردان و زنان در
انتخابات نشان می دهد که زنان نزدیک به دو برابر مردان در انتخابات شرکت می کنند

2-در شیوه دوم، می توانیم نسبت بخت ها را به صورت درصد تفسیر کنیم، برای این کار، ابتدا نسبت بخت ها را از عدد ۱ کم و سپس در عدد ۱۰۰ ضرب می کنیم .به عنوان مثال، اگر نسبت بخت های ۱.۷۸ را از عدد ۱ کم کنیم و در عدد ۱۰۰ ضرب کنیم حاصل آن برابر ۷۸ درصد خواهد بود که نشان می دهد با افزایش یک واحد در متغیر جنسیت، بخت شرکت در انتخابات به اندازه ۷۸ درصد افزایش می یابد.

حجم نمونه در رگرسیون لجستیک

اگرچه در ادبیات مربوط به رگرسیون لجستیک، قواعد خاصی برای حجم نمونه و نیز حداقل نسبت تعداد نمونه به تعداد متغیر مستقل پیشنهاد نشده است، اما برخی نویسندگان در حوزه آمار چند متغیره، حداقل حجم نمونه برای یک تحلیل رگرسون لجستیک خوب را ۱۰۰ نفر و برخی نیز ۵۰ نفر عنوان کرده اند .در خصوص حداقل نسبت تعداد نمونه به تعداد متغیر مستقل نیز، به عنوان یک قاعده کلی، حداقل نسبت ۱۰ متغیر مستقل به ۱ نمونه لازم است .اما آن چه مسلم می باشد، این است که هر چه تعداد متغیرهای مستقل بیشتر باشد، حجم نمونه باید بیشتر باشد .ضمن آن که در رگرسیون لجستیک، به حجم نمونه بسیار بیشتر از حجم نمونه در رگرسیون خطی نیاز داریم.

نحوه تعریف متغیرهای طبقه بندی شده (اسمی و ترتیبی) در رگرسیون لجستیک

یکی از مهم ترین مشکلات در اجرای تحلیل رگرسیون لجستیک، وجود متغیرهای ترتیبی است .در هنگام اجرای رگرسیون لجستیک، فرض بر این است که تمامی متغیرهای مستقل در سطح سنجش فاصله ای / نسبی هستند، درحالی که در عمل چنین نیست و برخی از آنها اسمی و ترتیبی نیز هستند .اما از آنجا که در رگرسیون لجستیک با نسبت احتمال وقوع یک پدیده به احتمال عدم وقوع آن پدیده سر وکار دارد، بنابراین متغیرهای مستقل حتما باید به متغیرهای شبه فاصله ای (با دو کد صفر و یک) تبدیل شوند تا بتوانیم نسبت طبقات آن در متغیر وابسته را بررسی کنیم .به همین خاطر، در نرم افزار SPSS در هنگام اجرای دستور رگرسیون لجستیک، از طریق کادر ....Categorical در کادر اصلی دستور، این امکان وجود دارد که متغیرهای طبقه بندی شده (اسمی و ترتیبی) را به صورت تصنعی به متغیرهای فاصله ای تبدیل کنیم.

برای تصنعی کردن متغیرهای اسمی و ترتیبی، باید هر یک از طبقات (گزینه های) آن متغیر به عنوان یک متغیر جداگانه با دو طبقه تعریف شده و به طبقه اول کد ۰ و به طبقه دوم کد ۱ تعلق گیرد .به عنوان مثال، اگر متغیر مورد نظر ما " سطح تحصیلات " است که در طبقات پایین، متوسط و بالا تعریف شده است، باید هر گزینه را به عنوان یک متغیر دو وجهی حساب کرده و به کسانی که آن میزان تحصلات را دارند کد ۱و به کسانی که آن میزان تحصیلات را ندارند کد ۰ تعلق گیرد .یعنی بدین صورت:

متغیر اول) تحصیلات پایین =۱و تحصیلات غیر پایین =۰

متغیر دوم) تحصیلات متوسط =۱و تحصیلات غیر متوسط =۰

همان طور که در طبقه بالا ملاحظه می شود، متغیر تحصیلات در هنگام تبدیل به متغیر تصنعی، فقط در 2 طبقه تعریف شده و طبقه سوم (یعنی تحصیلات بالا) حذف شده است .دلیل این امر آن است که در رگرسیون لجستیک، همانند رگرسیون خطی ، متغیر تصنعی برای طبقه آخر (یعنی بزرگترین کد) تعریف نمی شود و تعداد آن همواره باید یکی کمتر از تعداد طبقات متغیر اصلی باشد (یعنیk-1) این اصل برای اجتناب از مساله تکنیکی چند هم خطی بودن در رگرسیون لجستیک است .طبقه ای که به متغیر تصنعی تبدیل نمی شود، طبقه مرجع Reference category نام دارد که مبنای مقایسه و تقابل با سایر طبقات قرار می گیرد.

موقعی که طبقات متغیر مستقل با طبقات مختلف متغیر وابسته به منظور مقایسه در تقابل قرار می گیرند، در هنگام اجرای کادر ...Categoricalدر دستور رگرسیون لجستیک، امکان انتخاب چندین نوع تقابل وجود دارد:

۱- شاخص(Indicator): در این روش، تقابل ها به صورت عضویت یا عدم عضویت در یک طبقه نشان داده میشوند .طبقه مرجع نیز به صورت یک ردیف در ماتریس تقابل با مقادیر) 1 (نشان داده می شود .این روش، رایج ترین روش انتخاب تقابل ها است که اغلب نیز از روش استفاده می کنیم.

۲- ساده (Simple): در این روش، هر طبقه از متغیر پیش بین (جز طبقه مرجع) با طبقه مرجع متغیر وابسته مقایسه می شوند.

۳- تفاوت (Difference): هر طبقه از متغیر پیش بین (جز طبقه اول) با میانگین اثر طبقات قبلی مقایسه می شود .این روش، به معکوس تقابل های هلمرت (Helmert) نیز معروف است.

۴- هلمرت (Helmert): هر طبقه از متغیر پیش بین (جز طبقه آخر) با میانگین اثر طبقات بعدی مقایسه می شوند (یعنی بر عکس روش تفاوت)

۵- چند جمله ای (Polynomial): در این روش، که به تقابل های چند جمله ای متعامد (Orthogonal polynomial contrasts) نیز معروف است، فرض بر این است که فاصله بین طبقات برابر می باشد .این تقابل ها، فقط برای متغیرهای عددی امکان پذیر هستند.

۶- انحراف (Deviation): هر طبقه از متغیر پیش بین (جز طبقه مرجع) با اثر کل مقایسه می شود.

ارزیابی مدل رگرسیون لجستیک

در تحلیل رگرسیون لجستیک، برای ارزیابی میزان برازش کل مدل، از آزمون نسبت درست نمایی(likelihood ratio) (نسبت درست نمایی در خروجی SPSS با نماد Log likelihood نشان داده میشود) استفاده می شود که آماره آن X^2 می باشد . بنابراین، در اینجا، آماره X^2 معادل آماره F در تحلیل رگرسیون خطی است .هدف آزمون نسبت درست نمایی این است که تفاوت بین احتمال پیش بینی شده حضور یک پاسخگو در یک طبقه و طبقه واقعی او را به حداقل کاهش دهد .برای این منظور، این آزمون ضرایب لجستیک تولید می کند که قادرند پاسخگویان را با دقت هر چه بیشتری در طبقه واقعی خود قرار دهند.

نسبت درست نمایی بر اساس تفاوت در مقدار انحراف ها محاسبه می شود .یعنی انحراف بدون وجود متغیر پیش بین در مدل منهای انحراف با وجود متغیر پیش بین در مدل .به عبراتی روشن تر، در آزمون نسبت درست نمایی، مقدار آماره X^2 یک بار فقط برای عدد ثابت در معادله بدون هیچ متغیر پیش بین (مستقل) و بار دیگر پس از ورود هر متغیر پیشبین به معادله محاسبه می شود.

مقدار انحراف از طریق فرمول زیر محاسبه می شود:

$$D = ۲\Sigma [y_i \ln(\frac{\pi(x_i)}{y_i}) + (y_i) \ln(\frac{i\pi(x_i)}{y_i})$$

بنابراین مقدار تفاضل دو انحراف از همدیگر (D) ، که نسبت درست نمایی بر اساس آن محاسبه می شود برابر است با:

G= X2= D (تغییر با مدل) – (تغییر بدون مدل)

در تفسیر مقدار درست نمایی با استفاده از معنی داری مقدار آماره X2 در سطح خطای کوچک تر از 0/05 ، می توانیم پی ببریم که آیا مدل رگرسیون به خوبی داده ها را برازش می کند یا خیر؟ البته باید توجه داشت که بر خلاف آماره X2 پیرسون در جدول توافقی و همچنین سایر آزمون های مشابه که از آماره X2 استفاده می کنند و در آنها مقدار بالاتر X2 نشان دهنده میزان بیشتر رابطه یا تفاوت است، در آزمون نسبت درستنمایی بر عکس است، یعنی در اینجا هر چه مقدار آماره X2 کوچکتر باشد برازش مدل بهتر است .بر اساس توضیحات می توان چنین نوشت:

۱- برای پی بردن به برازش کل مدل رگرسیون لجستیک، از آماره χ_2 استفاده می کنیم.

۲- برای پی بردن به معنی داری اثر هر متغیر بر متغیر وابسته، از آماره والد استفاده می کنیم.

۳- برای پی بردن به میزان تاثیر هر متغیر بر متغیر وابسته، از آماره Exp(B) استفاده می کنیم که همان نسبت بخت ها است. بنابراین، آماره والد Wald مقدم بر آماره Exp(B) می باشد.

رگرسیون لجستیک اسمی دو وجهی

تحلیل رگرسیون لجستیک اسمی دو وجهی زمانی مورد استفاده قرار می گیرد که متغیر وابسته در سطح اسمی دو وجهی است و بنا داریم وجود یا عدم یک صفت را بر اساس مجموعه ای از متغیرهای مستقل پیش بینی کنیم . بنابراین در رگرسیون لجستیک اسمی دو وجهی ما نمی توانیم همانند رگرسیون خطی چند متغیره، مقدار عددی دقیق یک متغیر وابسته را بر اساس اطلاعاتی که راجع به متغیرهای مستقل داریم ، تعیین کنیم .بلکه در این روش، ما نسبت احتمال (P) سروکار داریم که آن را با کد (۱) نشان می دهند تا کد (۰). مثال های متعددی را می توان بر شمرد که در آن با یک متغیر وابسته اسمی دووجهی سروکار داریم : اینکه چرا برخی کودکان در هنگام تولد می میرند و برخی دیگر زنده می مانند؟ این که چرا برخی شهروندان در انتخابات شرکت می کنند و برخی دیگر شرکت نمی کنند؟ این که چرا برخی دانشجویان موفق به دریافت مدرک تحصیلی می شوند و برخی دیگر موفق نمی شوند؟ این که چرا برخی مردم دچار بیماری های قلبی می شوند و برخی دیگر نمی شوند؟ این که چرا برخی بنگاه ها موفق اند و برخی دیگر موفق نیستند؟ اینها همگی سوالاتی هستند که نرم افزار SPSS این امکان را برای ما فراهم می کند تا بتوانیم عوامل پیش بینی کننده تغییرات یک متغیر اسمی دووجهی را نیز شناسایی کنیم .در واقع، در این نوع رگرسیون لجستیک، می توانیم احتمال وقوع یک واقعه را به طور مستقیم برآورد کنیم.

پیش فرض ها

۱- متغیر وابسته باید حتماً باید در سطح سنجش اسمی دووجهی باشد

۲- متغیرهای مستقل می توانند هم در سطح کمی (فاصله ای / نسبی) و هم در سطح کیفی طبقه بندی شده(اسمی / ترتیبی) باشند .اما چنان چه یک یا چند متغیر مستقل در سطح اسمی / ترتیبی بودند ، حتما باید ابتدا این متغیرها را به متغیرهای تصنعی تبدیل کنیم (یعنی کدهای 0 و۱) .البته در روش رگرسیون لجستیک، کادری به نام ....Categorical وجود دارد که با انتخاب و اجرای آن، متغیرهای ترتیبی به طور خودکار به متغیرهای تصنعی تبدیل می شوند .بنابراین، نیازی به کد گذاری مجدد آنها توسط محقق نیست.

۳- لزوم تبعیت داده های متغیرهای مسقل از توزیع نرمال ضروری نیست .اما چنان چه این متغیرها دارای توزیع نرمال چند متغیره باشند ، در آن صورت برازش مدل بهتر خواهد بود.

۴- چند همخطی نبودن متغیرهای مستقل ، از دیگر مفروضات رگرسیون لجستیک می باشد. چرا که در صورت چند همخطی بودن این متغیرها، برآوردها دارای اریب بوده و خطاهای استاندارد نیز نوسان زیادی خواهد داشت .ترسیم نمودار پراکنش به ما کمک می کند تا از چند همخطی بودن یا نبودن متغیرهای مستقل اطمینان حاصل کنیم.

نکته :چنان چه پیش فرض های نرمال بودن چند متغیره و برابری ماتریس های واریانس و کوواریانس تامین شدند، در آن صورت پیشنهاد می شود که از روش تحلیل تشخیصی ، به جای روش تحلیل لجستیک استفاده کنیم.

مثال: پژوهشگری در صدد است تا مهمترین عوامل موثر بر مرگ و میر نوزادان را شناسایی کند و پی ببرد که مرگ و میر این نوزادان تحت تاثیر چه عواملی است؟ به عبارتی این پژوهشگر قصد دارد تا پی ببرد که آیا با استفاده از یک سری متغیرها می تواند تعیین کند که احتمال زنده ماندن یا فوت نوزادان چقدر است؟

در این تحقیق ۱۰۰ نوزاد تازه متولد شده به عنوان حجم نمونه انتخاب و در آن، وضعیت مرگ و میر نوزادان به عنوان متغیر وابسته ، و 5 متغیر جنسیت ، وضعیت دو قلو بودن ، رتبه تولد، سن مادر و تحصیلات مادر به عنوان متغیرهای مستقل در نظر گرفته شده اند .این پژوهشگر سپس با استفاده از دستور (Transform > Recode) متغیر را به صورت زیر تعریف و کد گذاری کرده است.

متغیر وابسته : متغیر وابسته وضعیت مرگ و میر نوزدان (Mortality) است که در دو طبقه زیر تعریف و دسته بندی می شوند:

طبقه اول : فوت شده(کد ۱) ، طبقه دوم : زنده (کد ۰)

متغیر مستقل:

۱) متغیر جنسیت نوزاد (Gender) در دو طبقه زیر دسته بندی می شود :طبقه اول : مرد (کد ۱)، طبقه دوم:زن (کد ۲)

۲) متغیر وضعیت دو قلو بودن نوزاد (Twin) در دو طبقه تعریف می شود :طبقه اول : یک قلو(کد ۱)، طبقه دوم :دو قلو(کد۲)

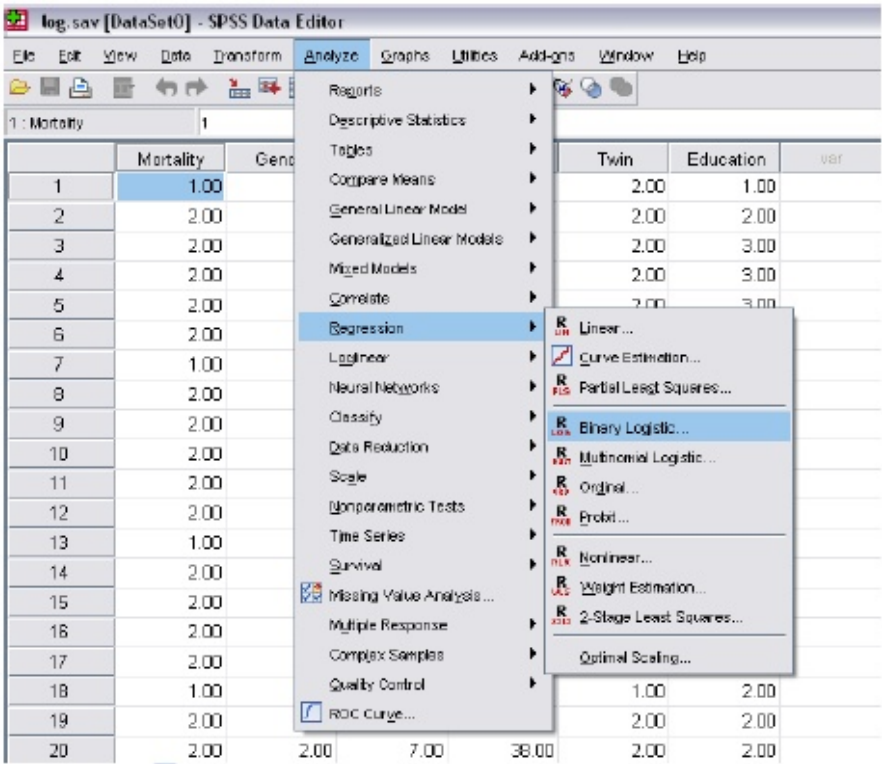
۳) متغیر رتبه تولد نوزاد (LiveRank) در هفت طبقه تعریف شد: ۱ و ۲ و ۳و ۴و ۵و ۶و ۷

۴) متغیر سن مادر (Mether Age) به صورت فاصله ای از ۱۵ سال تا ۳۹ سال تعریف شد.

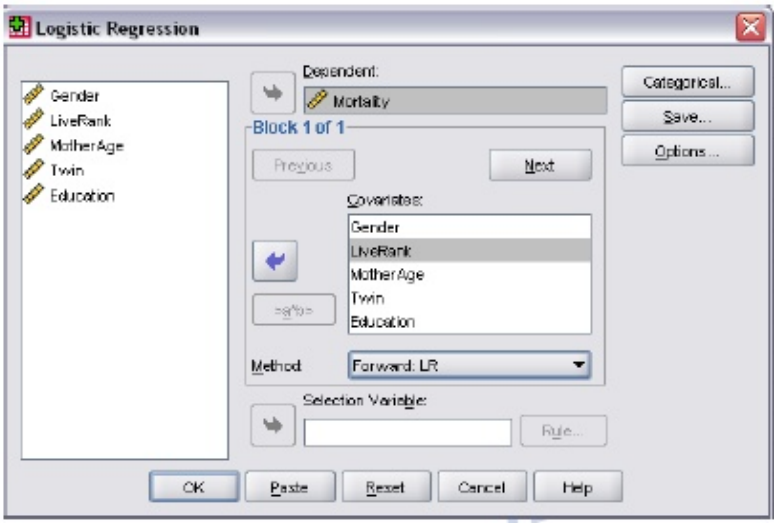
۵) متغیر سطح تحصیلات مادر (Education) در سه طبقه تحصلی تعریف شد :طبقه اول :بی سواد و ابتدائی (کد ۱)، طبقه دوم : راهنمایی و متوسط (کد ۲) ، طبقه سوم :عالیه (کد ۳)

نحوه اجرا در SPSS

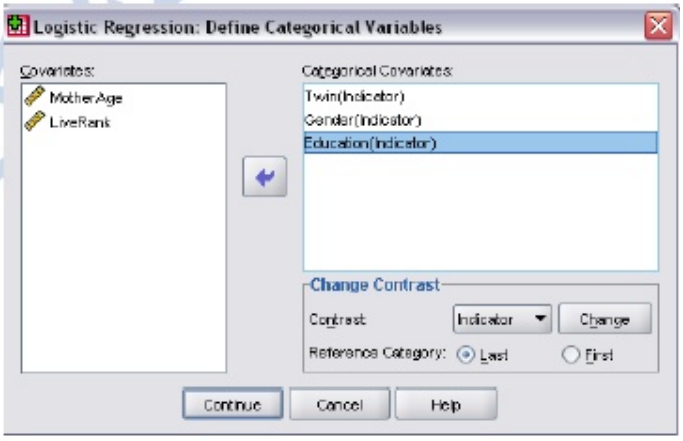
۱–دستور .... Analyze > Regression > Binary Logistic را اجرا می کنیم .



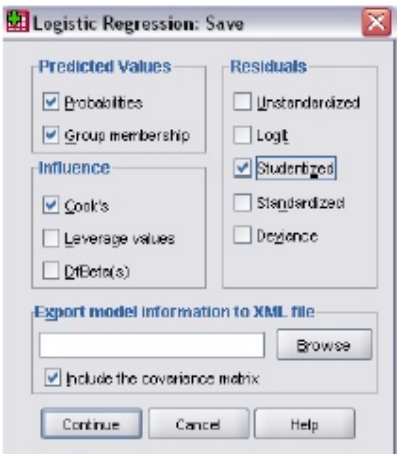
۲– متغیر وابسته (Mortality) را وارد کادر Dependent و متغیرهای مستقل (در اینجا 5 متغیر) را وارد کادر Covariance می کنیم .سپس در کادر Method روش Forward: LR را انتخاب می کنیم.



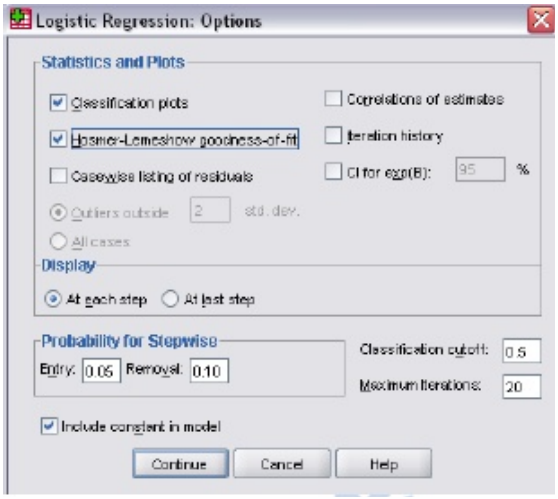
۳- دکمه …Categorical را کلیک می کنیم .در پنجره جدیدی که باز می شود ، متغیرهای طبقه بندی شده (Education , Twin , Gender) را از سمت چپ انتخاب کرده و به سمت راست انتقال می دهیم (اگر در کادر Contrast روشی غیر از Indicatorرا به عنوان روش تقابل و First را به عنوان طبقه مرجع انتخاب کردیم در آن صورت حتما باید بر روی دکمه Changeکلیک کنیم تا تغییرات اعمال شود.)



۴- بر روی دکمه … Save کلیک می کنیم .در پنجره جدیدی که باز می شود گزینه های , Group membership probabilitiesرا در قسمت predicted values ، گزینه Cooks را در قسمت Influence و گزینه Studentized را در قسمت Residualsانتخاب و سپس، بر روی دکمهContinue کلیک می کنیم.



5-دکمه … Options را کلیک کرده و در پنجره جدیدی که باز می شود ، گزینه هایClassification plots و – Hosmer Lemeshow goodness – of – fitرا در قسمت Statistics and plots انتخاب می کنیم .سپس ، دکمه Continue را کلیک می کنیم.



۶- بر روی دکمه Ok در کادر اصلی دستور … Binary Logistic کلیک می کنیم.

جدول زیر به عنوان اولین خروجی رگرسیون لجستیک نشان می دهد که از مجموع ۱۰۰ نفر ، ۹۰ نوزاد (۹۰ درصد) مورد تحلیل قرار گرفته و ۱۰ نوزاد(۱۰درصد) به علت داشتن مقدار گمشده و نامعلوم وارد تحلیل نشده اند (توجه داشته باشید زمانی که مقدار یکی از متغیرهای وابسته یا مستقل برای یک پاسخگو نامعلوم باشد ، رگرسیون لجستیک آن پاسخگو را از تحلیل خارج می کند مانند 01 نوزاد در این مثال.)

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	90	90.0
	Missing Cases	10	10.0
	Total	100	100.0
Unselected Cases		0	.0
Total		100	100.0

a. If weight is in effect, see classification table for the total number of cases.

جدول زیر با عنوان (Dependent Variable Encoding) کدهای اولیه متغیرهای وابسته را به کدهای جدید تغییر می دهد .در تحقیق حاضر، متغیر وابسته مورد نظر، وضعیت مرگ و میر نوزادانان می باشد که از دو طبقه نوزادن فوت شده(با کد ۱) و نوزادان زنده مانده (با کد ۲) تشکیل شده است. اما همان طور که در این جدول می بینید ، دستور رگرسیون لجستیک کدهای (۱) و (۲) را به کدهای (۰) و (۱) تبدیل می کند .یعنی کد(۱) برای گروه نوزادان فوت شده را به کد(۰) و کد (۲) برای گروه نوزدان زنده مانده را به کد (۱) تبدیل می کند .بنابراین، موقعی که ضریب تاثیر (B) یک متغیر مستقل بر متغیر واتبسته مثبت باشد، بدین معنی است که در نتیجه این متغیر مستقل، شاهد افزایش در احتمال زنده ماندن نوزادان خواهیم بود و بر عکس، ضریب تاثیر منفی دلالت بر کاهش احتمال زنده ماندن و در واقع افزایش احتمال فوت نوزادان دارد.

Dependent Variable Encoding

Original Value	Internal Value
1.00	0
2.00	1

جدول زیر اطلاعات توصیفی در خصوص نحوره برخورد با کدهای متغیر کیفی طبقه بندی شده (اسمی و ترتیبی) را ارائه می دهد .در این جدول، در مورد متغیرهای دو وجهی، برای طبقه اول از هر متغیر، کد (۱)و برای طبقه دوم کد(۰) در نظر گرفته شده است .اما در مورد متغیرهای چند وجهی مانند تحصیلات، ابتدا در مقایسه طبقات اول و دوم با هم، برای طبقه اول کد (۱)و برای طبقه دوم کد (۰)را در نظر گرفته می شود .سپس، در مقایسه طبقات دوم و سوم با هم، برای طبقه دوم کد (۱)و برای طبقه سوم کد (۰)را در نظر گرفته است .در مورد سایر طبقات بالای)6نیز، به همین نحو عمل می شود .یعنی هر طبقه با طبقه بالاتر از خود مقایسه و کد گذاری می شود) .

		Frequency	Parameter coding (1)	(2)
Education	1.00	5	1.000	.000
	2.00	55	.000	1.000
	3.00	30	.000	.000
Twin	1.00	10	1.000	
	2.00	80	.000	
Gender	1.00	50	1.000	
	2.00	40	.000	

خروجی تحلیل رگرسیون لجستیک شامل دو بلوک است: **Block 0**؛ این بلوک، خروجی مرحله صفر در رگرسیون لجستیک را نشان می دهد .یعنی مرحله ای که هنوز هیچ داده ای وارد تحلیل نشده است .برای تفسیر نتایج رگرسیون لجستیک، از این بلوک استفاده نمی کنیم. **Block 1** ؛ بلوک اصلی رگرسیون لجستیک است که در هنگام تفسیر نتایج باید آن را گزارش کنیم . این مرحله، نتایج پس از ورود متغیرها به تحلیل را در برمی گیرد .در ادامه، به نتایج هر دو بلوک به تفکیک اشاره می شود .

۱- خروجی بلوک ۰ :

خروجی جدول زیر نتایج مربوط به Block 0 یا تحلیل اولیه را نشان می دهد .در این بلوک، هیچ مرحله ای(گامی) برای ورود داده ها به مدل اجرا نشده است .به همین خاطر، بر چسب (Step 0 مرحله صفر) را در تمامی خروجی های این بلوک مشاهده می کنیم .در این بلوک با استفاده از روش تحلیل رگرسیون گام به گام Forward: LR ، متغیرهای مستقل (در صورت معنی دار بودن) ، به ترتیب مقدار نمره از بالا به پایین ، وارد مدل می شوند .البته در نرم افزارspss زمانی که مقدار نمره متغیرها بیشتر باشد، ورود آن ها به مدل به صورت پیش گزیده و خود به خود انجام می شود.

جدول طبقه بندی زیر نشان می دهد که با اطمینان ۷۷/۸ درصد با استفاده از مجموع ۶ متغیر مستقل در این تحقیق قادریم تغییرات متغیر وابسته مرگ و میر نوزادان را تبیین کنیم.

Classification Table[a,b]

		Predicted		
		Mortality		Percentage
	Observed	1.00	2.00	Correct
Step 0	Mortality 1.00	0	20	.0
	2.00	0	70	100.0
	Overall Percentage			77.8

a. Constant is included in the model.
b. The cut value is .500

در جدول زیر، چون هنوز هیچ متغیری وارد مدل نشده، بنابراین تنها نتایج مربوط به عدد ثابت در مدل آمده است که نسبت آماره والد ان برابر با ۲۴/۴۱۳ و نسبت بخت های آن ۳/۵۰۰ می باشد.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1.253	.254	24.413	1	.000	3.500

در تحقیق حاضر 5 متغیر جنسیت، وضعیت دوقلوبودن، رتبه تولد، سن مادر، و تحصیلات مادر به عنوان متغیرهای پیش بینی کننده احتمال مرگ و میر نوزادان مورد تحلیل قرار گرفته اند .با توجه به نتایج تحلیل اولیه در بلوک ۰، به استثنای متغیرهای رتبه تولد و سن مادر و میزان متوسط تحصیلات با سطح معنی داری بزرگتر از ۰/۰۵

(به ترتیب با سطح معنی داری ۰/۴۹۱ و ۰/۴۲۱ و ۰/۱۴۹) سایر متغیرهای وارد شده در تحلیل رگرسیون قادر به پیش بینی تغییرات متغیر وابسته مرگ و میر نوزادان می باشند.

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Gender(1)	3.937	1	.047
		LiveRank	.475	1	.491
		MotherAge	.648	1	.421
		Twin(1)	39.375	1	.000
		Education	26.883	2	.000
		Education(1)	18.529	1	.000
		Education(2)	2.087	1	.149
	Overall Statistics		71.214	6	.000

۲- خروجی بلوک ۱

مهم ترین خروجی تحلیل رگرسیون لجستیک ، خروجی Block 1 است که تفسیر نتایج رگرسیون لجستیک باید بر اساس این خروجی انجام بگیرد .این خروجی، نتایج رگرسیون لجستیک را به تفکیک در هر مرحله نشان می دهد .در مجموع، کل خروجی رگرسیون لجستیک در بلوک ۱ را می توان به ۴ بخش زیر تقسیم کرد که هر محققی در هنگام گزارش نویسی باید به آنها اشاره کند:

۱- ارزیابی کل مدل

۲- آماره نکوئی برازش

۳- آزمون های آماری مربوط به تاثیر هر متغیر پیش بین (مستقل)

۴- ارزیابی احتمالات پیش بینی شده

اولین بخش از خروجی بلوک ۱، نتایج آزمون اوم نی بوس مربوط به ارزیابی کل مدل رگرسیون لجستیک را نشان می‌دهد .این آزمون به بررسی این موضوع می پردازد که مدل تا چه اندازه قدرت تبیین و کارایی دارد؟ با توجه به نتایج آزمون اوم نی بوس، در مرحله چهارم مدل قابل قبول و در سطح خطای کمتر از ۰/۰۱ معنی دار است.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	35.064	1	.000
	Block	35.064	1	.000
	Model	35.064	1	.000
Step 2	Step	28.888	2	.000
	Block	63.952	3	.000
	Model	63.952	3	.000
Step 3	Step	8.901	1	.003
	Block	72.854	4	.000
	Model	72.854	4	.000
Step 4	Step	22.493	1	.000
	Block	95.347	5	.000
	Model	95.347	5	.000

جدول بعدی نتایج مربوط به دو آماره لگاریتم درستنمایی و ضریب تعیین پزودو (شامل ضریب تعیین کاکس و نل و ضریب تعیین نیجل کرک) را نشان می دهد .این ضرایب، تقریب های ضریب تعیین (R^2) در رگرسیون خطی هستندکه در اینجا در رگرسیون لجستیک استفاده می شود .در رگرسیون لجستیک، چون محاسبه دقیق مقدار ضریب تعیین دشوار هستند ، بنابراین، از مقادیر آماره های فوق برای این کار استفاده می شود تا مشخص گردد که متغیرهای مستقل توانسته اند تا چه میزان از واریانس متغیر وابسته را تبیین کنند .مقادیر آماره های ضریب تعیین پزودو بین ۰ تا ۱ نوسان دارد و هر چه مقدار این آماره ها به عدد ۱ نزدیک تر باشد ، نشان می دهد که نقش متغیرهای مستقل درتبیین واریانس متغیر وابسته زیاد است و بر عکس مقادیر نزدیک به ۰ دلالت بر نقش ضعیف متغیرها در این امر دارد.

در مورد این مثال، ملاحظه می شود که در مرحله چهارم، مقادیر هر دو آماره مربوط به ضریب تعیین پزودو تقریبا بالا (۰/۶۵۳ و ۱) بوده و این نشان می دهد که 5 متغیر مستقل این تحقیق از قدرت تبیین تقریباً بالایی در خصوص واریانس و تغییرات متغیر وابسته به مرگ و میر نوزادان برخوردار هستند .در واقع، این 5 متغیر توانسته اند بین ۶۵/۳ تا ۱۰۰ درصد از تغییرات مرگ و میر نوزادان را تبیین کنند.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	60.283[a]	.323	.494
2	31.395[a]	.509	.779
3	22.493[a]	.555	.849
4	.000[a]	.653	1.000

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

در دستور رگرسیون لجستیک به کمک دو روش می توانیم به میزان برازش مدل با داده ها پی ببریم:

۱)آماره نکوئی برازش هوسمر – لمشو: با توجه به نتیجه حاصل از آزمون هوسمر و لمشو در مرحله چهارم(۰) برازش میزان پیش بینی تغییرات متغیر وابسته در سطح خطای کوچکتر از ۰/۰۱ معنی دار نیست. بدین معنا که مدل تحقیق مناسب نبوده و از برازش لازم برخوردار نیست .یعنی متغیرهای مستقل قادر به پیش بینی تغییرات متغیر وابسته نیست.

۲) نمودارهای باقیمانده :دو نمودار مفیدی که برای این کار وجود دارند ، نمودار تغییر در انحراف در مقابل احتمالات پیش بینی شده می باشند .برای ترسیم این نمودار می توانیم در هنگام اجرای رگرسیون لجستیک،

با انتخاب گزینه Cooks از قسمت Influence و گزینه Deviance از قسمت Residuals در کادر ... Save ، نمودارهای مورد نظر را ترسیم کنیم.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	.000	0	.
2	.000	2	1.000
3	.000	4	1.000
4	.000	6	1.000

مبنای محاسبه مقدار آزمون هوسمر – لمشو که در بالا مشاهده شد ، جدول توافقی زیر می باشد .در این جدول، برای هر تکرار ، فراوانی های مشاهده شده و مورد انتظار پاسخگویان در هر طبقه از متغیر وابسته نشان داده شده است.

Contingency Table for Hosmer and Lemeshow Test

		Mortality = 1.00		Mortality = 2.00		
		Observed	Expected	Observed	Expected	Total
Step 1	1	10	10.000	0	.000	10
	2	10	10.000	70	70.000	80
Step 2	1	5	5.000	0	.000	5
	2	10	10.000	0	.000	10
	3	5	5.000	40	40.000	45
	4	0	.000	30	30.000	30
Step 3	1	10	10.000	0	.000	10
	2	5	5.000	0	.000	5
	3	5	5.000	15	15.000	20
	4	0	.000	25	25.000	25
	5	0	.000	20	20.000	20
	6	0	.000	10	10.000	10
Step 4	1	10	10.000	0	.000	10
	2	10	10.000	0	.000	10
	3	0	.000	10	10.000	10
	4	0	.000	10	10.000	10
	5	0	.000	5	5.000	5
	6	0	.000	10	10.000	10
	7	0	.000	5	5.000	5
	8	0	.000	30	30.000	30

پس از اجرای تحلیل رگرسیون لجستیک ، به دو طریق می توانیم قدرت مدل در تفکیک افراد در طبقات متغیر وابسته را تعیین کنیم؛ ۱) جدول طبقه بندی (Classification Table) و ۲) نمودار طبقه بندی (Classification Plot). در خروجی رگرسیون لجستیک ، ابتدا نتایج جدول طبقه بندی و سپس نتایج نمودار طبقه بندی نشان داده می شود .به همین خاطر ابتدا به نتایج جدول طبقه بندی و سپس نمودار طبقه بندی (که آخرین خروجی است) اشاره می شود. البته باید اشاره داشت که نتایج هر دو خروجی یکسان و مکمل همدیگر هستند.

جدول صفحه بعد که جدول طبقه بندی نام دارد، به ما کمک می کند تا از طریق ترسیم توافقی پاسخ ها در طبقات مشاهده شده و مورد انتظار، عملکرد مدل و قدرت تفکیک افراد در طبقات متغیر وابسته (نوزادان زنده مانده و نوزادان فوت شده) را به پاسخ مورد انتظار در همان طبقه نشان می دهد .این جدول به ما کمک می کند تا میزان عملکرد پیش بینی پذیری مدل را ارزیابی کنیم .در این جدول، خانه های قطری ، تعداد پیش بینی های صحیح را نشان می دهد وخانه های خارج از قطر نیز تعداد پیش بینی های غیر صحیح را نشان می دهند .بر اساس نتایج این جدول می توانیم به میزان صحت و سقم مدل در طبقه بندی افراد پی ببریم .همان طور که در جدول مشاهده می شود ، درصد صحت پیش بینی و طبقه بندی مدل در ۴ مرحله پس از ورود هر متغیر به مدل نشان داده شده است . به عنوان مثال، در مرحله اول که فقط متغیر Twin در مدل بود، دقت طبقه بندی افراد توسط مدل برابر با ۸۸/۹ درصد بود. یعنی دراین مرحله، ۱۰ نفر از نوزادانان فوت شده و ۷۰ نفر از نوزادان زنده مانده به درستی تفکیک شده اند . به عبارتی ، ۱۰ نفر از نوزادان فوت شده به اشتباه در گروه نوزادان زنده مانده و بر عکس صفر تا از نوزادان زنده مانده در

گروه نوزادان فوت شده قرار گرفته اند .در مرحله دوم، صحت طبقه بندی با ورود متغیر تحصیلات مادر به ۹۴/۴ درصد افزایش یافت .در این مرحله نیز ، 5 نفر از نوزادن فوت شده به اشتباه در گروه نوزادان زنده مانده قرار گرفتند .در مرحله سوم، با ورود متغیر سن مادر)، صحت طبقه بندی افراد همان۹۴/۴ درصد باقی ماند . این مقدار صحت طبقه بندی نشان می دهد که با اطمینان ۹۴/۴ درصد می توانیم بگوئیم با استفاده از مجموع 5 متغیر مستقل این تحقیق ، قادریم تغییرات متغیر وابسته مرگ و میر را تبیین کنیم .ضمن آن که در این مرحله، اشتباه طبقه بندی به همان صورت بوده است که 5 نفر از نوزادان فوت شده به اشتباه در گروه نوزادان زنده مانده طبقه بندی شده اند و همچنین ۱۵ نفر از نوزادان فوت شده و ۷۰ نفر از نوزادان زنده مانده به درستی تفکیک شده اند .برای درک بهتر این جدول، می توانیم نتایج آن را با نمودار طبقه بندی (Classification Plot) به عنوان آخرین خروجی تحلیل رگرسیون لجستیک مقایسه کنیم.

Classification Table[a]

Observed		Predicted		
		Mortality		Percentage Correct
		1	2	
Step 1	Mortality 1	10	10	50.0
	2	0	70	100.0
	Overall Percentage			88.9
Step 2	Mortality 1	15	5	75.0
	2	0	70	100.0
	Overall Percentage			94.4
Step 3	Mortality 1	15	5	75.0
	2	0	70	100.0
	Overall Percentage			94.4
Step 4	Mortality 1	20	0	100.0
	2	0	70	100.0
	Overall Percentage			100.0

a. The cut value is .500

جدول بعد با عنوان (Variables in the Equation) ضمن ارائه خلاصه ای از نقش هر متغیر در مدل، نشان می دهد که کدام متغیرها بعد از اجرای رگرسیون لجستیک، در مدل باقی مانده اند .این جدول، مهم ترین جدول در تفسیر نتایج مربوط به معنی داری و میزان تاثیر هر متغیر مستقل بر متغیر وابسته می باشد .در این جدول، چندین آماره مهم وجود دارند که برای تفسیر نتایج آنها، دانستن ماهیت و کارکرد آنها ضروری است:

B: این آماره همان ضریب تاثیر رگرسیونی استاندارد نشده است .در واقع، این آماره همان ضریب برآورد شده ، همراه با خطای استاندارد است.

.S .E: این آماره که Standard Error است، عبارت از خطای استاندارد می باشد.

Wald: آماره والد، مهم ترین آماره برای آزمون معنی داری حضور هر متغیر مستقل در مدل می باشد که می توانیم از طریق سطح معنی داری آن (sig) به این امر پی ببریم .در واقع، آماره والد معادل آماره t در رگرسیون خطی است .در تفسیر نتیجه

آماره والد نیز می گوئیم چنان چه مقدار این آماره برای هر متغیر در سطح خطای کوچکتر از ۰/۰۵ معنی دار باشد، در آن صورت نتیجه می گیریم که وجود آن متغیر در مدل مفید و اثر آن معنی دار است.

Exp(B): این آماره که به نسبت بخت ها معروف است، عبارت می باشد از نسبت احتمال وقوع یک پدیده به احتمال عدم وقوع آن .در واقع، این آماره عبارت است از تغییرات پیش بینی شده در بخت ها به ازای یک واحد افزایش در متغیر مستقل .این نسبت، معادل ضرایب رگرسیون استاندارد شده (Beta)در رگرسیون خطی می باشد که برای تفسیر نتایج تحقیق از آن استفاده می شود .موقعی که نسبت بخت ها کوچکتر از عدد ۱ باشد، در آن صورت می گوئیم که با افزایش مقادیر متغیر مستقل، احتمال وقوع پدیده کاهش می یابد (اثر منفی) و بر عکس، موقعی که نسبت بخت ها بزرگتر از عدد ۱ باشد در آن صورت می گوئیم با افزایش مقادیر متغیر مستقل، احتمال وقوع پدیده افزایش می یابد (اثر مثبت). بنابراین، در تحلیل رگرسیون لجستیک، تاثیر منفی هر متغیر مستقل در مدل را می توان از دو طریق تشخیص داد ۱- از طریق علامت منفی مقدار آماره B ، و ۲- از طریق کوچک تر بودن مقدار Exp(B) از عدد یک.

نکته :یادآور می شویم برای اینکه پی ببریم کدام متغیرها بر متغیر وابسته تاثیر آماری معنی داری دارند، از آمارهwald استفاده می کنیم .اما برای اینکه پی ببریم میزان تاثیر هر یک از این متغیرها بر متغیر وابسته چقدر است، از آماره Exp(B) استفاده می کنیم .بنابراین، آماره wald مقدم بر آماره Exp(B) می باشد.

در مورد این مثال با استناد به جدول (Variable in the Equatio)می توانیم بگوئیم که تمامی متغیرهای مستقل وارد شده در تحلیل رگرسیونی به جز متغیر رتبه تولد ، قادر به پیش بینی تغییرات متغیر وابسته (احتمال فوت یا زنده ماندن نوزادان) هستند و توانایی پیش بینی آنها در سطح خطای کوچکتر از ۰/۰۱ باشد .از طرفی، در میان مجموعه متغیرهای معنی دار باقیمانده در مدل، متغیر سن مادر بیشترین توانایی را در وضعیت مرگ و میر نوزادان دارد.جزئیات دقیق تر تاثیر این متغیرها بر مرگ و میر نوزادان در زیر بیان می گردد:

۱- در مورد متغیر تحصیلات شاهدیم که هم میزان پائین و هم میزان متوسط تحصیلات مادر بر وضعیت مرگ ومیر نوزادان موثر است .بر اساس نتایج جدول، مشاهده می کنیم که این اثرات به صورت منفی است و با افزایش میزان تحصیلات متوسط مادر به تحصیلات بالا احتمال زنده ماندن نوزاد کاهش می یابد .در مورد تحصیلات پائین مادر نیز، شاهد تاثیر معنی دار این سطح از تحصیلات بر احتمال فوت یا زنده ماندن نوزاد هستیم .نسبت بخت های این متغیر برابر با ۰ ولی ضریب B نشان می دهد که این اثر به صورت منفی است یعنی با افزایش میزان تحصیلات پائین مادر به تحصیلات بالا احتمال زنده ماندن نوزادان کاهش می یابد.

۲- سن مادر (Motherage) متغیر دیگری است که بر وضعیت مرگ و میر نوزادان تاثیر دارد .تاثیر این متغیر

مثبت است .یعنی با افزایش سن مادر، احتمال زنده ماندن نوزاد افزایش می یابد و بر عکس، کاهش سن مادر، احتمال زنده ماندن او را افزایش می باید.

۳- در مورد دو متغیر دیگر دوقلو بودن و جنسیت (Gender , Twin)میبینم که با توجه به اینکه نسبت بخت آنها ۰ است و همچنین ضریب B می توان مشاهده کرد که اثرات این دو متغیر هم به صورت منفی است .به عبارت دیگر دوقلو بودن نوزاد، احتمال مرگ او را افزایش می دهد تا نسبت به زمانی که نوزاد یک قلو متولد شود.

همچنین بر اساس نتایج این جدول می توانیم مدل رگرسیون لجستیک را بر اساس مرحله چهارم به صورت زیر نشان دهیم:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta 1x1 + \beta 2x2 + \cdots + \beta kxk$$

مدل رگرسیون = -17.92 (عدد ثابت) – 46.052 (Gender) + 4.599 (Motherage) – 61.959 (Twin) – 63.011 (Low Education) – 36.967 (Medium Education)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Twin(1)	-23.149	1.271E4	.000	1	.999	.000
	Constant	1.946	.338	33.132	1	.000	7.000
Step 2[b]	Twin(1)	-23.282	1.271E4	.000	1	.999	.000
	Education			.000	2	1.000	
	Education(1)	-42.406	1.942E4	.000	1	.998	.000
	Education(2)	-19.123	7.338E3	.000	1	.998	.000
	Constant	21.203	7.338E3	.000	1	.998	1.615E9
Step 3[c]	Gender(1)	-19.475	5.868E3	.000	1	.997	.000
	Twin(1)	-40.865	1.282E4	.000	1	.997	.000
	Education			.000	2	1.000	
	Education(1)	-41.859	1.923E4	.000	1	.998	.000
	Education(2)	-19.557	6.836E3	.000	1	.998	.000
	Constant	40.131	9.009E3	.000	1	.996	2.682E17
Step 4[d]	Gender(1)	-46.052	7.151E3	.000	1	.995	.000
	MotherAge	4.599	658.666	.000	1	.994	99.402
	Twin(1)	-61.959	1.316E4	.000	1	.996	.000
	Education			.000	2	1.000	
	Education(1)	-63.011	1.907E4	.000	1	.997	.000
	Education(2)	-36.967	6.702E3	.000	1	.996	.000
	Constant	-17.921	1.100E4	.000	1	.999	.000

a. Variable(s) entered on step 1: Twin.
b. Variable(s) entered on step 2: Education.
c. Variable(s) entered on step 3: Gender.
d. Variable(s) entered on step 4: MotherAge.

جدول بعدی بر اساس تغییر در آماره لگاریتم درستنمایی، سهم ورود هر متغیر به مدل را در تبیین تغییرات متغیر وابسته در هر ۴ مرحله نشان می دهد . در این جدول، در هر مرحله، چنان چه سهم هر متغیر در مدل بیشتر باشد، ورود آن متغیر به مدل باعث افزایش مقدار آماره لگاریتم درستنمائی می شود .از طرفی، در این جدول، مبنای خروج هر متغیر سطح معنی داری بزرگتر از ۰/۰۱ است .یعنی در هر مرحله، هر متغیر که سطح معنی داری آن از ۰/۰۱ بزرگتر باشد از مدل خارج می شود .به

عنوان مثال، در مرحله چهارم، ملاحظه می شود که ابتدا با ورود متغیر Gender به مدل، میزان تغییر در مقدار آماره لگاریتم درستنمایی برابر با ۲۴/۵۶۲ بود که با ورود متغیر Motherage مقدار این آماره به ۲۲/۴۹۳ کاهش یافت. همینطور با ورود متغیر Twin به مدل به ۵۵/۰۶۵ افزایش و دوباره با ورود متغیرEducation به ۴۰/۸۷۰ کاهش پیدا کرد.

Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	Twin	-47.674	35.064	1	.000
Step 2	Twin	-32.227	33.060	1	.000
	Education	-30.142	28.888	2	.000
Step 3	Gender	-15.697	8.901	1	.003
	Twin	-30.342	38.191	1	.000
	Education	-23.837	25.180	2	.000
Step 4	Gender	-12.263	24.526	1	.000
	MotherAge	-11.247	22.493	1	.000
	Twin	-27.533	55.065	1	.000
	Education	-20.435	40.870	2	.000

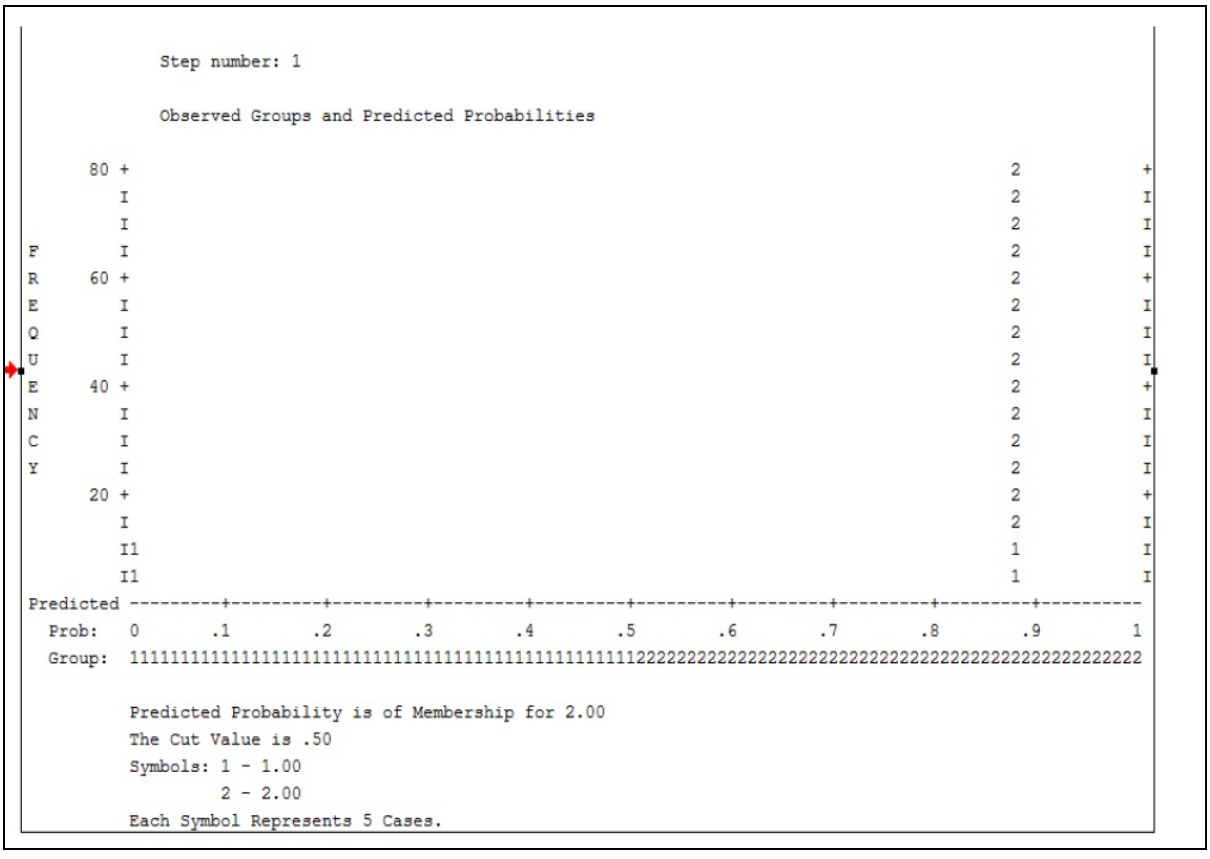
جدول بعدی متغیرهایی را نشان می دهد که در هر مرحله از تحلیل، تاثیر معنی داری بر تغییرات متغیر وابسته نداشته و در نتیجه از مدل خارج شده اند .همچنین، در این جدول، نمره کل آماره ها نیز برای هر مرحله محاسبه می شود و این فرض صفر را به آزمون می گذارد که هر متغیری که در معادله نیامده، دارای ضریب بتای ۰ است .یعنی تاثیری بر تغییرات متغیر مرگ و میر نوزادان ندارد .در این مثال، در مرحله چهارم، سطح معنی داری متغیر LiveRank بزرگتر از ۰/۰۵ یعنی ۱ است، بنابراین تاثیر معنی داری بر مرگ و میر نوزادان ندارد .همچنین، مقدار کل آماره این متغیر در مرحله چهارم برابر با ۱ است.

Variables not in the Equation

			Score	df	Sig.
Step 1	Variables	Gender(1)	8.889	1	.003
		LiveRank	8.717	1	.003
		MotherAge	2.899	1	.089
		Education	39.365	2	.000
		Education(1)	37.333	1	.000
		Education(2)	.181	1	.670
	Overall Statistics		50.324	5	.000
Step 2	Variables	Gender(1)	7.031	1	.008
		LiveRank	5.104	1	.024
		MotherAge	4.113	1	.043
	Overall Statistics		16.004	3	.001
Step 3	Variables	LiveRank	9.231	1	.002
		MotherAge	15.152	1	.000
	Overall Statistics		15.259	2	.000
Step 4	Variables	LiveRank	.000	1	1.000
	Overall Statistics		.000	1	1.000

نمودار صفحه بعد که به نمودار طبقه بندی (Classification Plot) معروف است، تصویر بصری صحت طبقه بندی را در یک نمودار هیستوگرام پشته ای نشان می دهد .همان طور که گفته شد، این نمودار را در کنار جدول طبقه بندی(Classification Table) که قبلاً توضیح داده شد، روش دیگری برای نمایش صحت طبقه بندی نوزادان در دو گروه نوزادان زنده مانده و فوت شده است .این نمودار از دو محور X و Y تشکیل شده است و میزان انطباق احتمالات پیش بینی شده با پیامدها (وقوع یا عدم

وقوع نتیجه) را نشان می دهد .در واقع، با استفاده از این نمودار می توانیم پی ببریم که آیا احتمالات پیش بینی شده بالا با وقوع پیامد، و احتمالات پیش بینی شده پایین با عدم وقوع پیامد هماهنگی دارد یا خیر؟ بنابراین، طبیعی است هر چه این هماهنگی بیشتر باشد صحت طبقه بندی بیشتر است .در این نمودار، محورX احتمال پیش بینی شده و محور Y فراوانی را نشان می دهد . احتمال پیش بینی شده بین ۰ تا ۱ نوسان دارد و هر پاسخگو می تواند مقداری از این احتمال را دریافت کند. مقدار برش یا نقطه تمایز احتمال پیش بینی شده برابر با ۰/۵ و پاسخگویان گروه اول مساوی یا پایین تر از احتمال ۰/۵ و پاسخگویان گروه دوم نیز بالاتر از احتمال ۰/۵ قرار می گیرند .بنابراین، هر چه پاسخگویان گروه اول در سمت چپ محور (یعنی تا احتمال۰/۵) و پاسخگویان گروه دوم در سمت راست محور (یعنی بالاتر از احتمال۰/۵)قرار بگیرند ، صحت طبقه بندی و پیش بینی مدل بیشتر است .



روش تحلیل مولفه های اصلی (PCA)

روش PCA یکی از روش های تحلیل آماری چندمتغیره برای تقلیل بعد داده ها و مقدمه ای بر روش های دیگری نظیر تحلیل عاملی است. از کاربردهای روش تحلیل مؤلفه های اصلی می توان به موارد زیر اشاره کرد:

- ❖ شناسایی نقاط پرت احتمالی در داده های چندمتغیره
- ❖ تقلیل تعداد متغیرهای توضیحی در رگرسیون چندگانه
- ❖ بررسی تک بعدی بودن آیتم ها در تحلیل آیتمی

پیش از شروع به ارائه ی این روش و کاربردهای آن چند مفهوم را یادآوری می کنیم:

۱- ماتریس داده های نمونه:

تعداد اعضای نمونه با n، تعداد متغیرها (شاخص ها) با p، اعضای نمونه با x_{ij} و ماتریس داده های نمونه با X نمایش داده می شوند

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \cdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}=\begin{bmatrix} \underline{X}_1^T \\ \underline{X}_2^T \\ \vdots \\ \underline{X}_n^T \end{bmatrix}$$

در این ماتریس $\mathbf{x_n}^T=(x_{n1} \quad x_{n1} \quad \cdots \quad x_{np})$ سطر n ام ماتریس X می باشد. در اغلب تحلیل های آماری لازم است مقدار n بزرگتر از p باشد.در حالت ایده آل عبارت زیر برقرار است:

$$n - p \geq 30$$

۲- بردار میانگین نمونه:

$$\bar{\underline{X}} = \left(\bar{x}_1 \quad \bar{x}_2 \quad \cdots \quad \bar{x}_p\right) = \left(\frac{1}{n}\sum_{r=1}^{n} x_{r1} \;,\; \frac{1}{n}\sum_{r=1}^{n} x_{r2} \;,\; \frac{1}{n} \;,\ldots\ldots,\; \frac{1}{n}\sum_{r=1}^{n} x_{rp}\right)^T$$

$$= \frac{1}{n}\sum_{r=1}^{n} \mathbf{x}_r$$

۳- ماتریس کواریانس نمونه:

ماتریس کواریانس نمونه را که ماتریسی است $p \times p$ با S نشان می دهند.

$$S=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \cdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{pp} \end{bmatrix} = \frac{1}{n-1}\sum_{r=1}^{n}(\underline{x_r}-\underline{\bar{x}})(\underline{x_r}-\underline{\bar{x}})^{T}$$

کواریانس نمونه بین متغیرهای i ام و j ام به صورت زیر محاسبه می شود:

$$s_{ij}=\frac{1}{n-1}\sum_{r=1}^{n}(x_{ri}-\bar{x_i})(x_{rj}-\bar{x_j})$$

Note: ماتریس S متقارن است. یعنی: $s_{ij} = s_{ji}$

Note: کواریانس یک متغیر با خودش واریانس آن متغیر نامیده می شود. واریانس نمونه ی مربوط به متغیر i ام از طریق زیر قابل محاسبه است:

$$s_{ij}=\frac{1}{n-1}\sum_{r=1}^{n}(x_{ri}-\bar{x_i})^2$$

٤- مقادیر و بردارهای ویژه ی ماتریس:

برای به دست آوردن مقادیر ویژه کافی است معادله ی مشخصه ی ماتریس را برابر صفر قرار دهیم:

$$\mathbf{Det}\ (\lambda \mathbf{I} - \mathbf{A}) = 0$$

بردارهای ویژه را نیز به طریق زیر محاسبه می نماییم:

$$AV_i = \lambda_i V_i \rightarrow (\lambda_i I - A)V_i = 0$$

تعریف تحلیل مؤلفه های اصلی

فرض می کنیم X یک ماتریس $n \times p$ داده های مربوط به متغیرهای X_1 و X_2 و ...و X_p می باشد. ماتریس کواریانس متناظر آن S_x می باشد و p متغیرهای جدید را با Y_1 و Y_2 و ...و Y_p نشان می دهیم. هر یک از این p متغیر جدید ترکیبی تابعی از p متغیر اولیه می باشند که بر اساس مقادیر و بردارهای ویژه ی ماتریس کواریانس داده های اولیه تعریف می شوند.

مراحل به دست آوردن متغیرهای جدید به طور خلاصه در ادامه شرح داده شده است:

۱- به دست آوردن ماتریس کواریانس بر اساس تعریف ذکر شده

۲- به دست آوردن مقادیر و بردارهای بر اساس تعاریف ذکر شده

۳- به دست آوردن مؤلفه های اصلی به صورت زیر:

فرض می کنیم مقادیر λ_1 و λ_2 و ...و λ_p مقادیر ویژه ی مرتب شده ی ماتریس S_x می باشد که $\lambda_p \geq ... \geq \lambda_2 \geq \lambda_1$ می باشد. q_1 و q_2و ... و q_p نیز بردارهای ویژه ی متناسب با این مقادیر می باشند. مؤلفه ی اصلی i ام به صورت زیر تعریف می شود.

$$Y_i = \underline{q_i}^T \underline{X} = q_{i1} X_1 + q_{i2} X_2 + q_{i3} X_3 + \ldots\ldots\ldots\ldots + q_{ip} X_p$$

$$\underline{qi} = (q_{i1} \quad q_{i2} \quad \cdots \quad q_{ip})^T$$

برای مثال: $Y_1 = \underline{q_1}^T \underline{X}$

در صورتیکه برای هریک از n عضو نمونه مقدار هر یک از p مؤلفه ی اصلی را محاسبه کنیم یک ماتریس $n \times p$ به صورت زیر حاصل می شود. که در آن y_{ri} مقدار مؤلفه ی اصلی i ام مربوط به r امین عضو نمونه است و $Y_r^T = (y_{r1} \quad y_{r1} \quad \cdots \quad y_{rp})$ سطر r ام ماتریس Y است.

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \ldots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix} = \begin{bmatrix} \underline{Y_1}^T \\ \underline{Y_2}^T \\ \vdots \\ \underline{Y_n}^T \end{bmatrix}$$

خواص مؤلفه های اصلی

- ✓ ضریب همبستگی دو به دویی مؤلفه های اصلی برابر صفر است. یعنی آنها دو به دو ناهمبسته می باشند.

- ✓ واریانس y_i برابر است با λ_i و انحراف معیار آن $\sqrt{\lambda i}$ می باشد. در نتیجه چون λ_iها از بزرگ به کوچک مرتب شده اند y_i ها نیز به ترتیب بزرگی واریانس مرتب شده اند.
- ✓ جمع واریانس های متغیرهای اولیه برابر است با جمع واریانس های مؤلفه های اصلی:

$$tr(S_x) = tr(S_y)$$

Note: اگر مقیاس اندازه گیری متغیرها تغییر کند، مؤلفه های اصلی نیز تغییر می کنند. درنتیجه لازم است مؤلفه های اصلی را بر اساس ماتریس داده های استاندارد شده به دست آوریم تا این مشکل برطرف گردد. از آنجاییکه ماتریس کوواریانس داده های استاندارد شده برابر با ماتریس همبستگی داده های استاندارد نشده است (یعنی $S_u=R_x$)، با استفاده از داده های استاندارد شده، مؤلفه های اصلی بر اساس مقادیر و بردارهای ویژه ی ماتریس همبستگی محاسبه می شوند.

برای این امـــــر فرض می کنیم λ_1 و λ_2 و ... و λ_p مقادیـــــر ویژه ی ماتریس R_x می باشند به طـوری که $\lambda_1 \geq \lambda_2 \geq .... \geq \lambda_p$ و q_1 و q_2 و ...و q_p بردارهای ویژه ی یکه‌ی متعامد متناظر آنها هستند. در این صورت مؤلفه ی اصلی i ام به صورت زیر تعریف می شود:

$$Y_i = \underline{q_i}^T \underline{U} = q_{i1} U_1 + q_{i2} U_2 + q_{i3} U_3 + ... + q_{iP} U_P$$

$$U_i = \frac{Xi - X}{Si}$$

در حالتی که داده های استاندارد شده به کار روند، از آنجا که واریانس هر یک از p متغیر استاندارد شده برابر ۱ می باشد، واریانس کل برابر است با :

$$Tr(R)=p$$

تصمیم راجع به تعداد مؤلفه های اصلی

در هر کاربردی از تحلیل مؤلفه های اصلی ابتدا باید تصمیم بگیریم چه تعدادی از مؤلفه های اصلی را در تحلیل به کار ببریم. در این رابطه دو قاعده سرانگشتی که رنچر معرفی کرده است و کاربرد فراوانی دارند را ارائه می کنیم:

۱- تعدادی از مؤلفه های اصلی را به کار ببرید که درصد بزرگ مشخصی از کل تغییر پذیری را (مثلاً ۸۰٪) پوشش دهند.

Note: درصد تغییر پذیری که مؤلفه ی اصلی i ام آن را به حساب آورده است برابر است با:

$$۱۰۰\,\lambda_i \Big/ \sum_{i=۱}^{p} \lambda_i \xrightarrow{\text{برای داده های استاندارد شده}} ۱۰۰\,\lambda_i \Big/ p$$

در نتیجه سهم k مؤلفه ی اصلی اول $(k < p)$ از کل تغییرپذیری برابر است با:

$$۱۰۰ \sum_{i=۱}^{k} \lambda_i \Big/ \sum_{i=۱}^{p} \lambda_i \xrightarrow{\text{برای داده های استاندارد شده}} ۱۰۰ \sum_{i=۱}^{k} \lambda_i \Big/ p$$

۲- مؤلفه هایی را به کار ببرید که مقادیر ویژه ی متناظر آنها (یعنی واریانس آنها) بزرگتر از متوسط همه ی مقادیر ویژه باشد. متوسط مقادیر ویژه به صورت زیر محاسبه می گردد:

$$\bar{\lambda} = \sum_{i=1}^{p} \lambda i \Big/ p$$

اگر در تحلیل مؤلفه های اصلی ماتریس همبستگی به کار رفته باشد، در این صورت:

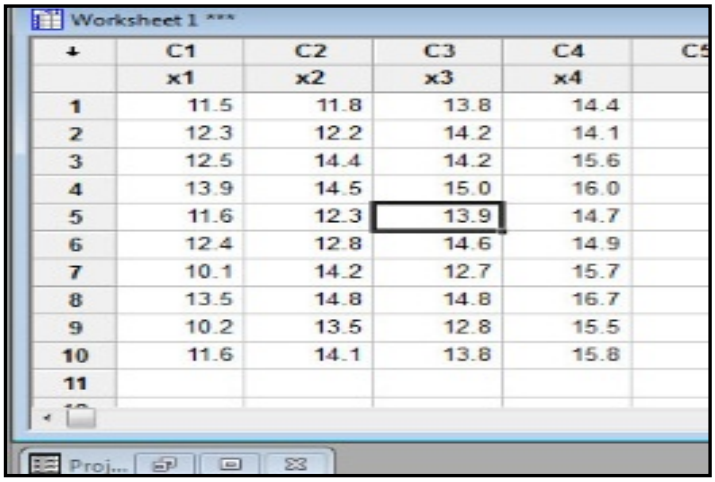
$$\sum_{i=1}^{p} \lambda i = p, \quad \bar{\lambda} = 1$$

فرآیند تحلیل مؤلفه های اصلی بر اساس مبنی توضیح داده شده، با نرم افزارهایی نظیر مینی تب به آسانی قابل انجام می باشد. در زیر مثالی که با این نرم افزار انجام پذیرفته است ارائه می گردد:

مثال : نمرات دروس ریاضی۱(X1)، فارسی۱(X2)، ریاضی۲ (X3)و فارسی۲ (X4) مربوط به ۱۰ دانشجو به شرح زیر موجود است:

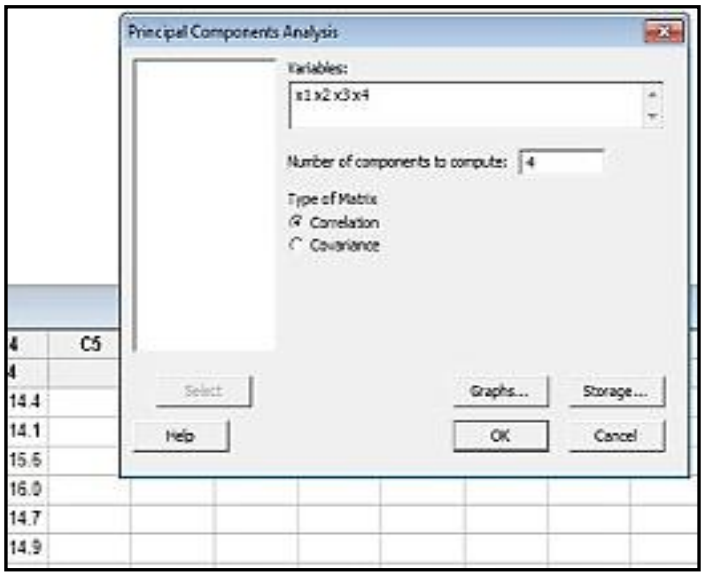
\	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
$X_۱$	۱۱.۵	۱۲.۳	۱۲.۵	۱۳.۹	۱۱.۶	۱۲.۴	۱۰.۱	۱۳.۵	۱۰.۲	۱۱.۶
$X_۲$	۱۱.۸	۱۲.۲	۱۴.۴	۱۴.۵	۱۲.۳	۱۲.۸	۱۴.۲	۱۴.۸	۱۳.۵	۱۴.۱
$X_۳$	۱۳.۸	۱۴.۲	۱۴.۲	۱۵.۰	۱۳.۹	۱۴.۶	۱۲.۷	۱۴.۸	۱۲.۸	۱۳.۸
$X_۴$	۱۴.۴	۱۴.۱	۱۵.۶	۱۶.۰	۱۴.۷	۱۴.۹	۱۵.۷	۱۶.۷	۱۵.۵	۱۵.۸

در محیط نرم افزار مینی تب داده ها را به صورت زیر در پنجره ی worksheet وارد می کنیم:

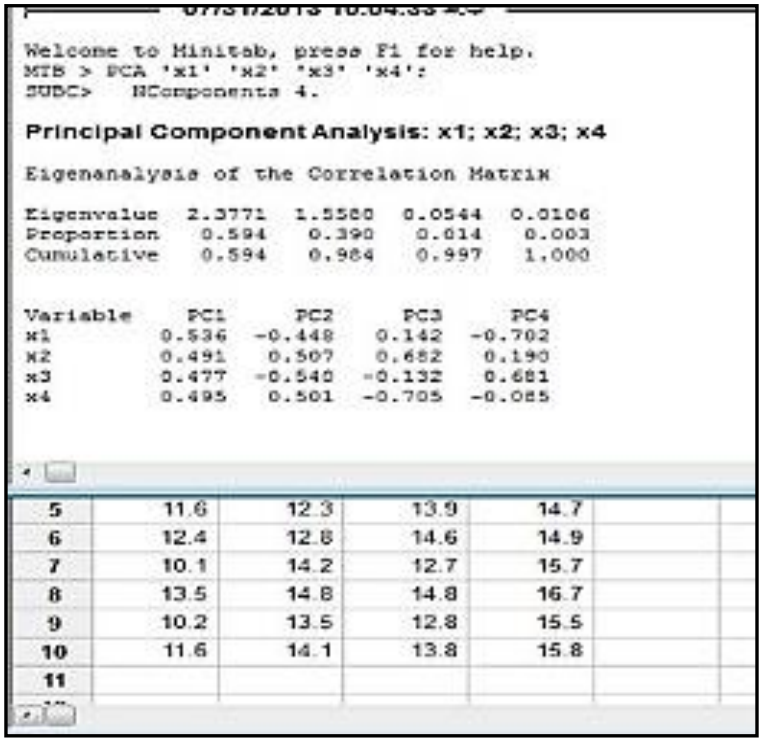

Worksheet 1 ***

	C1	C2	C3	C4	C5
	x1	x2	x3	x4	
1	11.5	11.8	13.8	14.4	
2	12.3	12.2	14.2	14.1	
3	12.5	14.4	14.2	15.6	
4	13.9	14.5	15.0	16.0	
5	11.6	12.3	13.9	14.7	
6	12.4	12.8	14.6	14.9	
7	10.1	14.2	12.7	15.7	
8	13.5	14.8	14.8	16.7	
9	10.2	13.5	12.8	15.5	
10	11.6	14.1	13.8	15.8	
11					

سپس از منوی stat گزینه ی principal components > Multivariate را انتخاب می کنیم و مطابق شکل زیر مراحل را انجام می دهیم:



پس از زدن OK نتایج فرآیند به صورت روبرو نمایش داده می شود:



```
Welcome to Minitab, press F1 for help.
MTB > PCA 'x1' 'x2' 'x3' 'x4';
SUBC>   NComponents 4.

Principal Component Analysis: x1; x2; x3; x4

Eigenanalysis of the Correlation Matrix

Eigenvalue   2.3771   1.5560   0.0544   0.0106
Proportion    0.594    0.390    0.014    0.003
Cumulative    0.594    0.984    0.997    1.000


Variable      PC1      PC2      PC3      PC4
x1          0.536   -0.448    0.142   -0.702
x2          0.491    0.507    0.682    0.190
x3          0.477   -0.540   -0.132    0.681
x4          0.495    0.501   -0.705   -0.085
```

5	11.6	12.3	13.9	14.7
6	12.4	12.8	14.6	14.9
7	10.1	14.2	12.7	15.7
8	13.5	14.8	14.8	16.7
9	10.2	13.5	12.8	15.5
10	11.6	14.1	13.8	15.8
11				

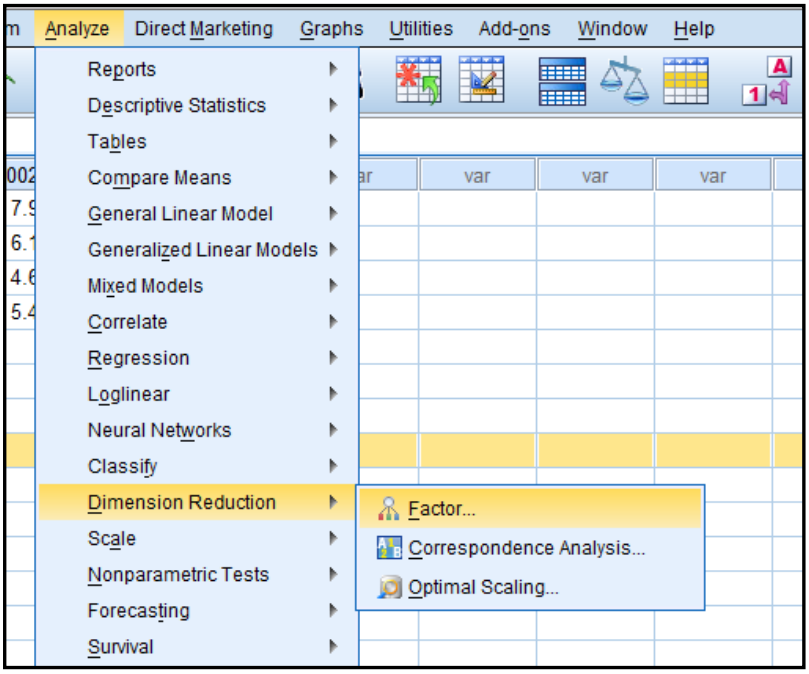
PCA در SPSS:

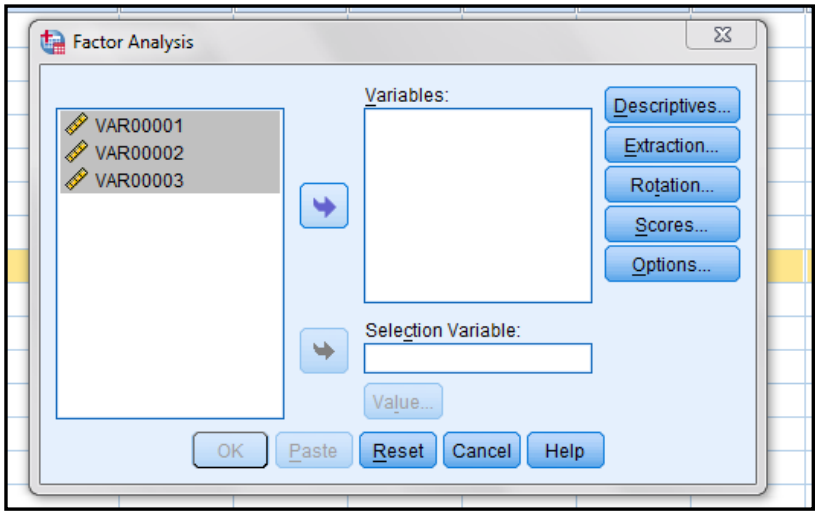
از دیگر نرم افزارهایی که می توان با آن تحلیل مؤلفه های اصلی را انجام داد نرم افزار spss می باشد. در ادامه نحوه ی انجام PCA با این نرم افزار ارائه خواهد شد.

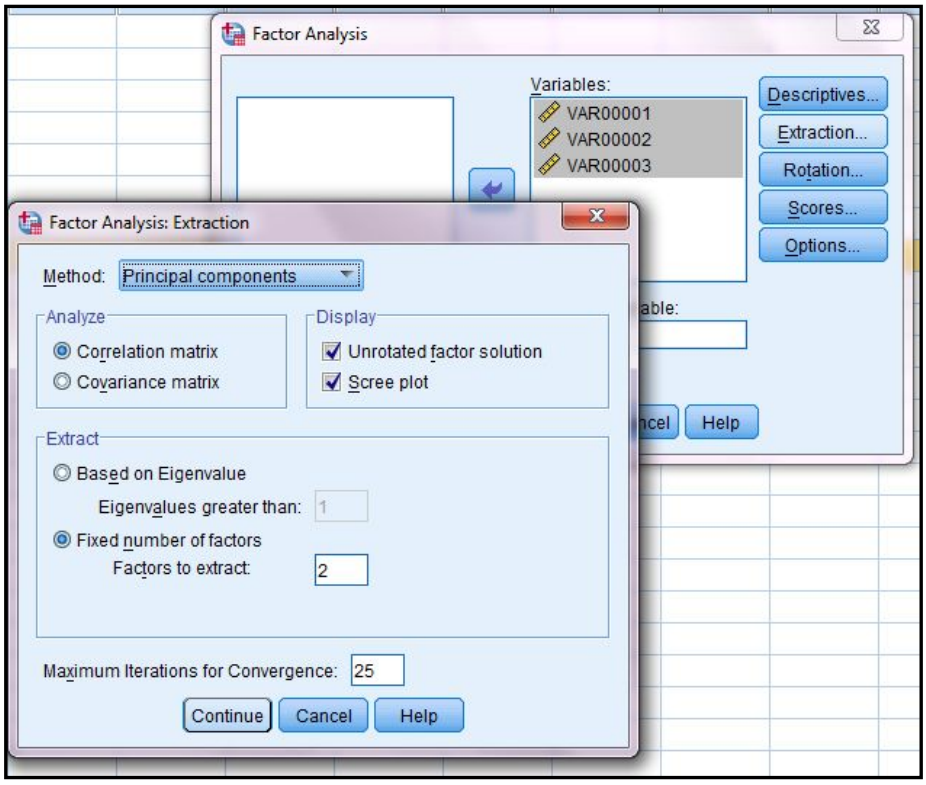
ابتدا داده ها را وارد می کنیم:

VAR00001	VAR00002	VAR00003
1.50	7.90	2.20
2.80	6.10	6.20
.60	4.60	6.30
3.10	5.40	5.30

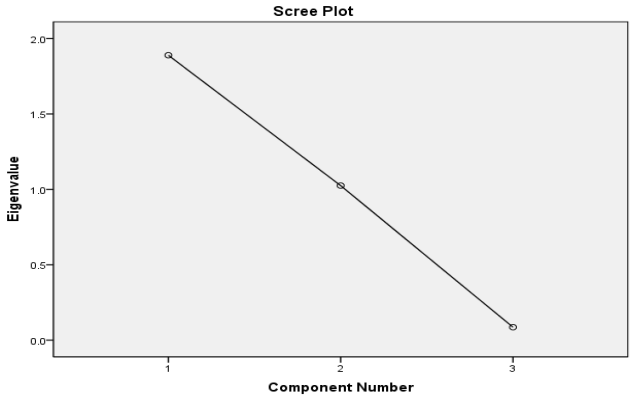
در این مثال فقط سه متغیر داریم و می خواهیم بعد را از ۳ به ۲ کاهش دهیم. حال اگر در مثالی با متغیرهای بیشتر ندانیم که چند عاما اصلی را می خواهیم استخراج کنیم به صورت زیر scree plot را رسم می کنیم. محور عمودی این نمودار مقادیر ویژه را نشان می دهد و محور افقی تعداد مؤلفه ها. با تحلیل این نمودار و با توجه به سیر نزولی مقادیر ویژه می توان آن تعداد از مؤلفه های اصلی را که بهتر می توانند نمایانگر کل داده ها باشند، تخمین زد. این روند در ادامه نشان داده شده است.



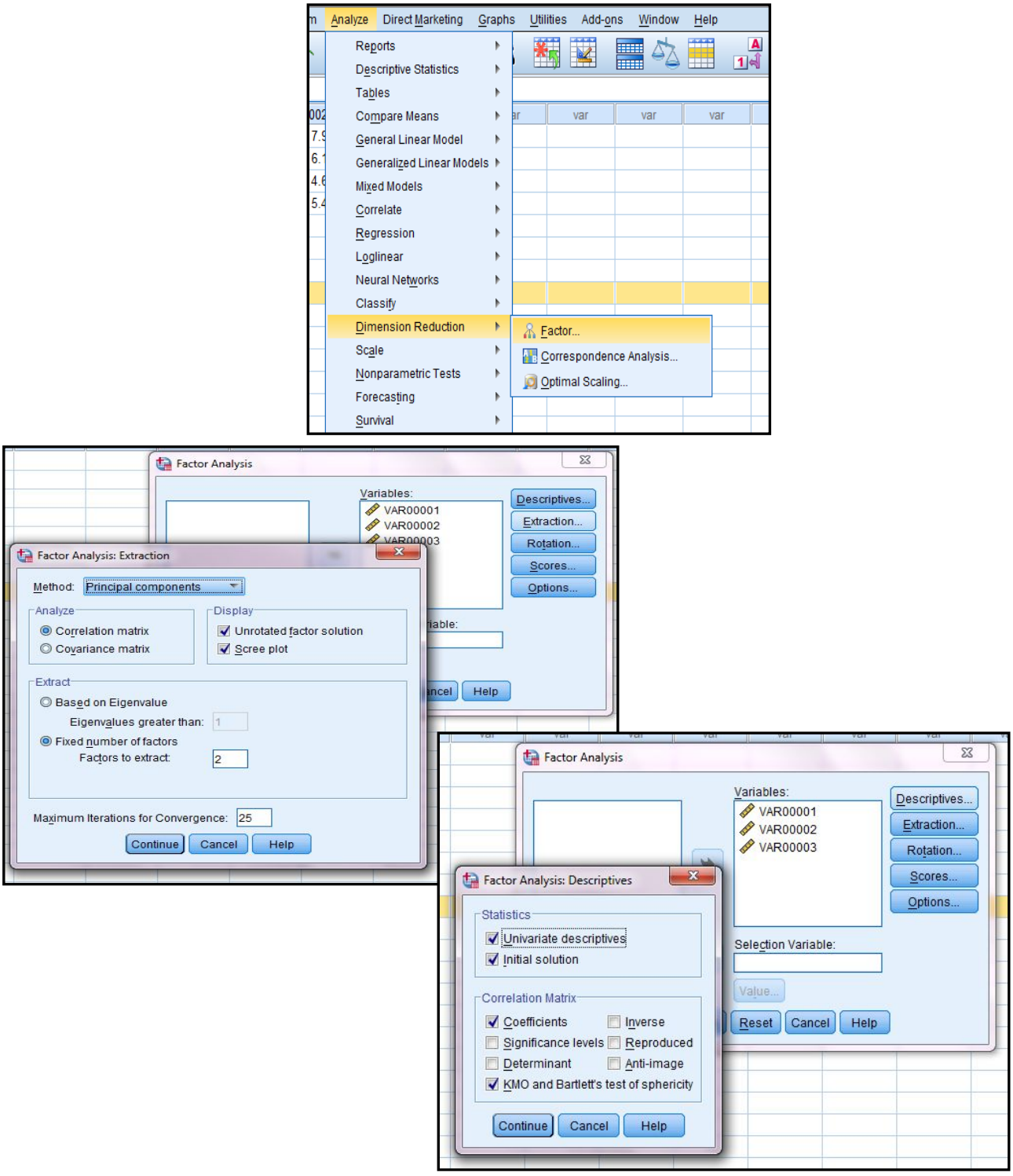

Factor Analysis
VAR00001
VAR00002
VAR00003
Variables:
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Selection Variable:
Value...
OK
Paste
Reset
Cancel
Help

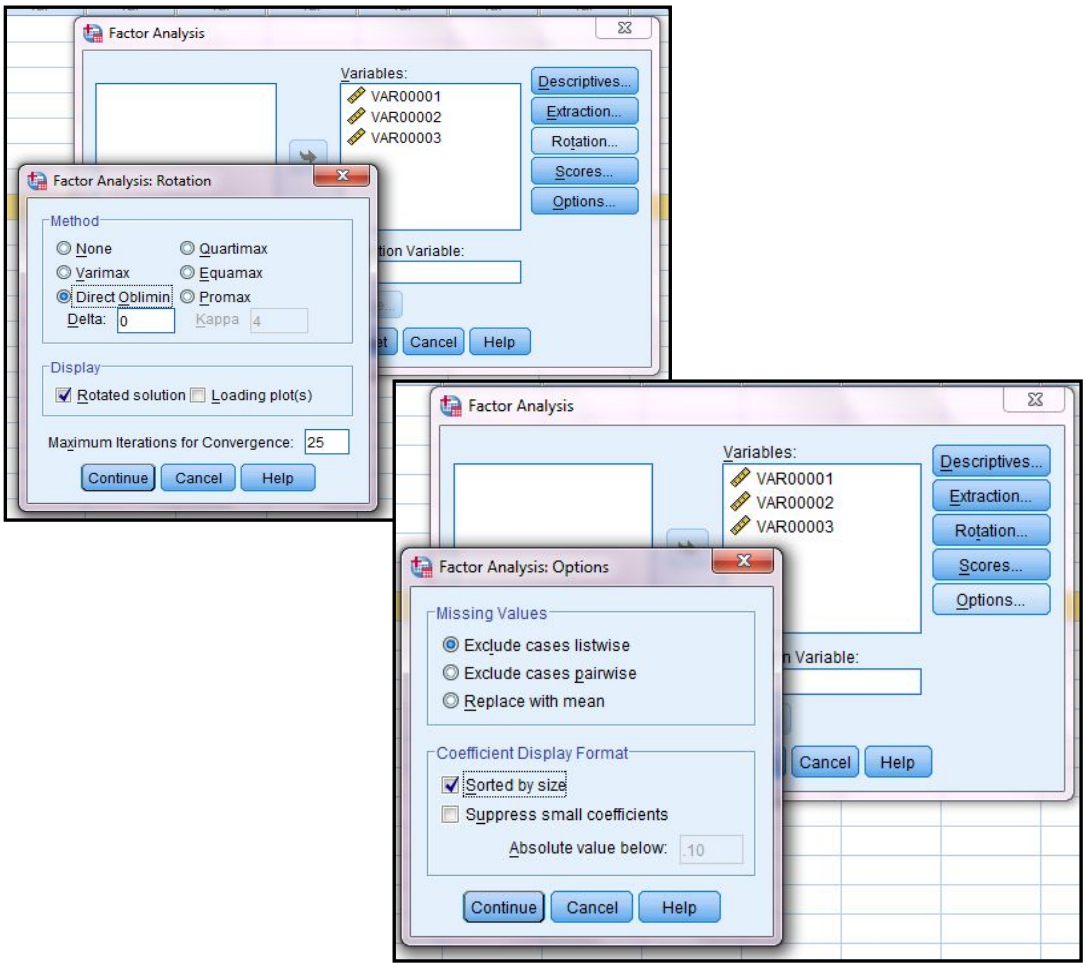

Factor Analysis
Variables:
VAR00001
VAR00002
VAR00003
Descriptives...
Extraction...
Rotation...
Scores...
Options...
Help
Factor Analysis: Extraction
Method: Principal components
Analyze
Correlation matrix
Covariance matrix
Display
Unrotated factor solution
Scree plot
Extract
Based on Eigenvalue
Eigenvalues greater than: 1
Fixed number of factors
Factors to extract: 2
Maximum Iterations for Convergence: 25
Continue
Cancel
Help

پس از انجام این مرحله نمودار را در صفحه ی output مشاهده کرده و آن را تحلیل می نماییم. در این مثال نمودار به صورت زیر می باشد:



متوجه می شویم که دو عامل اول مقادیر ویژه ی بزرگتری دارند. ادامه ی مراحل به صورت زیر می باشد:





پس از طی مراحل فوق نتایج در صفحه ی output به صورت زیر نمایش داده می شوند.

Factor Analysis

[DataSet0]

Descriptive Statistics

	Mean	Std. Deviation	Analysis N
VAR00001	2.0000	1.16333	4
VAR00002	6.0000	1.40712	4
VAR00003	5.0000	1.92007	4

Correlation Matrix

		VAR00001	VAR00002	VAR00003
Correlation	VAR00001	1.000	.088	.130
	VAR00002	.088	1.000	-.888
	VAR00003	.130	-.888	1.000

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.397
Bartlett's Test of Sphericity	Approx. Chi-Square	2.094
	df	3
	Sig.	.553

Communalities

	Initial	Extraction
VAR00001	1.000	.998
VAR00002	1.000	.958
VAR00003	1.000	.958

Extraction Method: Principal Component Analysis.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings[a]
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	1.889	62.977	62.977	1.889	62.977	62.977	1.889
2	1.025	34.161	97.139	1.025	34.161	97.139	1.027
3	.086	2.861	100.000				

Extraction Method: Principal Component Analysis.

a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance.

Component Matrix[a]

	Component	
	1	2
VAR00003	.974	.094
VAR00002	-.969	.143
VAR00001	.047	.998

Extraction Method: Principal Component Analysis.

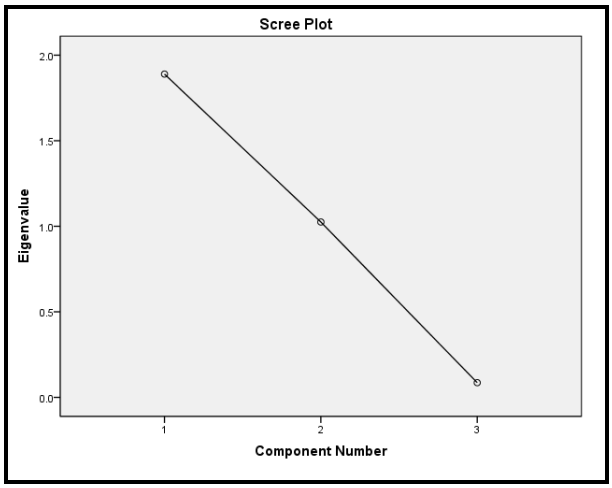
a. 2 components extracted.

Pattern Matrix[a]

	Component	
	1	2
VAR00002	-.974	.118
VAR00003	.969	.120
VAR00001	.000	.999

Extraction Method: Principal Component Analysis.
Rotation Method: Oblimin with Kaiser Normalization.

a. Rotation converged in 3 iterations.



PCA در متلب

```
[COEFF,SCORE] = princomp(X)
[COEFF,SCORE,latent] = princomp(X)
[COEFF,SCORE,latent,tsquare] = princomp(X)
[...] = princomp(X,'econ')
```

توضیح:

دستور COEFF = princomp(X) تحلیل مولفه های اصلی (PCA) را روی ماتریس داده X_{nxp} ، اعمال کرده ضرایب مولفه های اصلی را بر میگرداند. سطرهای ماتریس X مربوط به مشاهدات بوده و ستونها مربوط به متغیرها است . COEFF یک ماتریس $p \times p$ می باشد که هر ستون آن دربردارنده ضرایب مربوط به یک مولفه اصلی می باشد . ستونها به ترتیب کاهش واریانس مولفه ها قرار گرفته اند .

دستور princomp با کاهش میانگین ستونها به X مرکزیت میدهد، اما تغییری در مقیاس ستونها نمی دهد. به منظور اعمال تحلیل مولفه های اصلی با متغیرهای استاندارد ، که بر مبنای ضریب همبستگی است ، از دستور princomp(zscore(X)) استفاده نمایید . به منظور اعمال مستقیم تحلیل مولفه های اصلی روی ماتریس کوواریانس یا همبستگی از دستور pcacov استفاده نمایید .

دستور [COEFF,SCORE] = princomp(X) ، SCORE را برمیگرداند ، امتیازهای مولفه های اصلی که نماینده X در فضای مولفه های اصلی میباشد . سطرهای ماتریس SCORE مربوط به مشاهدات ، و ستونها مربوط به مولفه ها میباشد.

دستور [COEFF,SCORE,latent] = princomp(X) ، latent را بر میگرداند ، برداری که شامل مقادیر ویژه ماتریس کوواریانس X میباشد.

[COEFF,SCORE,latent,tsquare] = princomp(X) ، tsquare را برمیگرداند . که شامل آماره T^2 ی هتلینگ برای هر داده است.

امتیازها مقادیری هستند که از تبدیل داده های اولیه به فضای مولفه های اصلی بوجود آمده اند . مقادیر بردار latent برابر با واریانس ستونهای ماتریس SCORE میباشد . T^2 ی هتلینگ مقیاسی است برای فاصله چند متغیره هر مشاهده از مرکز مجموعه داده.

وقتی n <= p باشد، SCORE(:,n:p) و latent(n:p) قطعاً برابر صفر خواهد بود ، و ستون COEFF(:,n:p) معرف مسیرهایی است که بر X عمودند.

دستور princomp(X,'econ') = [...] تنها درایه هایی از latent که لزوماً صفر نیستند ، و نیز ستونهای متناظر COEFF و SCORE ،که وقتی $n \leq p$ باشد فقط شامل $n - ۱$ تای اول میشود، را بر میگرداند .

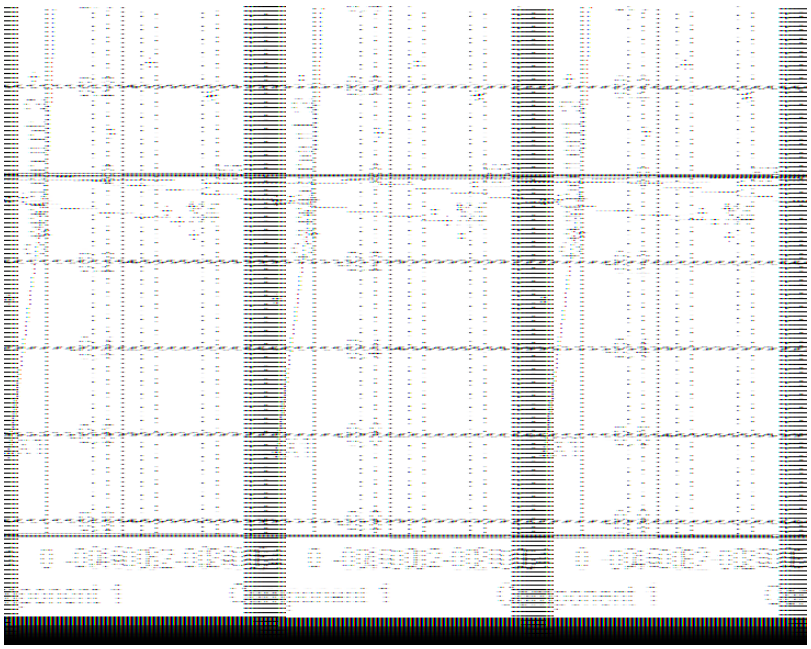
مثال

مولفه های اصلی را برای داده های ingredients در مجموعه داده Hald بدست آورده واریانس هر مولفه محاسبه شود .

```
load hald;
[pc,score,latent,tsquare] = princomp(ingredients);
pc,latent
pc =
  0.0678 -0.6460  0.5673 -0.5062
  0.6785 -0.0200 -0.5440 -0.4933
 -0.0290  0.7553  0.4036 -0.5156
 -0.7309 -0.1085 -0.4684 -0.4844
latent =
 517.7969
  67.4964
  12.4054
   0.2372
```

نمودار و دستور زیر نشان میدهند که ۹۸٪ واریانس مربوط به دو مولفه میباشد:

```
cumsum(latent)./sum(latent)
ans =
     0.86597
     0.97886
      0.9996
           1
biplot(pc(:,1:2),'Scores',score(:,1:2),'VarLabels',...
{'X1' 'X2' 'X3' 'X4'})
```



روش تحلیل آیتمی

تحلیل آیتمی مجموعه ای از روش هایی است که هدف آنها بررسی استاندارد بودن سؤالات چندگزینه ای در آزمون ها در فعالیت های آموزشی است از اهداف تحلیل آیتمی می توان به موارد زیر اشاره کرد:

- ✓ محاسبه ی معیار سازگاری درونی سؤالات
- ✓ محاسبه ی معیار قابلیت ممیزی
- ✓ محاسبه ی معیار دشواری

این تحلیل کاربردهای فراوانی در تحلیل داده های پرسشنامه ای دارد ولی برای داده های غیر پرسشنامه ای نیز کاربرد دارد. زیرا بسیاری از ایده هایی که در داده های پرسشنامه ای استفاده می شوند، در داده های غیر پرسشنامه ای نیز به کار می رود.

یعنی گاهی که لازم است گروهی از متغیرها را با هم ترکیب کنیم تا متغیر پنهانی را اندازه گیری کنیم و به این ترتیب بعد را کاهش دهیم، می توان از روش تحلیل آیتمی استفاده کرد .

در تحلیل آیتمی به دنبال پاسخ به دو سؤال اساسی زیر هستیم:

۱- آیا یک گروه خاص از سؤالات (آیتم ها)، متغیر پنهان واحدی را اندازه گیری می کنند؟

۲- چگونه مقدار متغیر پنهان را با استفاده از پاسخ سؤالات اندازه گیری کنیم؟

برای پاسخ به سؤال اول روش های زیر وجود دارد:

- ✓ استفاده از معیار آلفای کرونباخ
- ✓ بررسی واریانس اولین مؤلفه ی اصلی

و برای پاسخ به سؤال دوم روش های زیر:

- ✓ استفاده از میاگین حسابی
- ✓ استفاده از میاگین وزنی

سؤال ۱ در واقع یک بعدی بودن را بررسی می کند.

NOTE: بررسی تک بعدی بودن یعنی بررسی این موضوع که آیا آیتم های مدنظر یک متغیر پنهان خاص را اندازه گیری می کنند یا خیر.

معیار آلفای کرونباخ

این معیار از مشهورترین معیارها برای پاسخ به سؤال اول می باشد.

الف)معیار آلفای کرونباخ برای داده های استاندارد نشده (داده های خام)

فرض: K سؤال x_1 x_2 ... x_K متغیر پنهان F را اندازه گیری می کنند.

$$Alpha=k/(1-k)[1-tr(s)/total(s)]$$

tr(s)= جمع درایه های قطری ماتریس کواریانس (S)

total(s)= جمع تمام درایه های ماتریس کواریانس (S)

NOTE: می دانیم که کواریانس هر متغیر با خودش برابر واریانس آن متغیر می باشد.

ب)معیار آلفای کرونباخ برای داده های استاندارد شده

مسلماً اگر داده ها دارای مقیاس اندازه گیری متفاوت با هم باشند، می بایست به جای ماتریس کواریانس از ماتریس همبستگی (R) استفاده کرد. متفاوت بودن مقیاس اندازه گیری در داده های پرسشنامه ای بعید است.

Alpha=k/(1-k)[1-k/total(R)]

Tr(R)=K می باشد. زیرا می دانیم در ماتریس همبستگی درایه های روی قطر اصلی برابر ۱ می باشد. چون همبستگی هر متغیر با خودش ۱ می باشد.

به عنوان نمونه همان مثال مطرح شده در بخش قبل را بررسی می کنیم.

مثال : نمرات دروس ریاضی۱(X1)، فارسی۱(X2)، ریاضی۲ (X3) و فارسی۲ (X4) مربوط به ۱۰ دانشجو به شرح زیر موجود است:

	۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰
X۱	۱۱.۵	۱۲.۳	۱۲.۵	۱۳.۹	۱۱.۶	۱۲.۴	۱۰.۱	۱۳.۵	۱۰.۲	۱۱.۶
X۲	۱۱.۸	۱۲.۲	۱۴.۴	۱۴.۵	۱۲.۳	۱۲.۸	۱۴.۲	۱۴.۸	۱۳.۵	۱۴.۱
X۳	۱۳.۸	۱۴.۲	۱۴.۲	۱۵.۰	۱۳.۹	۱۴.۶	۱۲.۷	۱۴.۸	۱۲.۸	۱۳.۸
X۴	۱۴.۴	۱۴.۱	۱۵.۶	۱۶.۰	۱۴.۷	۱۴.۹	۱۵.۷	۱۶.۷	۱۵.۵	۱۵.۸

می خواهیم با ترکیب متغیر x1 و x3 متغیر پنهانی به نام استعداد ریاضی را اندازه گیری کنیم:

	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	X1	11.5	12.3	12.5	13.9	11.6	12.4	10.1	13.5	10.2	11.6
3	X2	11.8	12.2	14.4	14.5	12.3	12.8	14.2	14.8	13.5	14.1
4	X3	13.8	14.2	14.2	15.0	13.9	14.6	12.7	14.8	12.8	13.8
5	X4	14.4	14.1	15.6	16.0	14.7	14.9	15.7	16.7	15.5	15.8
6											
7	X1 bar	11.96		S=	1.529	0.927		R=	1	0.977	
8	X3 bar	13.98			0.927	0.588			0.977	1	
9											
10	بر اساس داده های خام	alpha=	$\frac{2}{2-1}$	$\left(1-\frac{1.529+0.588}{1.529+0.588+2*0.927}\right)$				=	0.934		
11											
12											
13	بر اساس داده های استاندارد شده	alpha=	$\frac{2}{2-1}$	$\left(1-\frac{1+1}{1+1+2*0.977}\right)$				=	0.988		
14											
15											

به منظور بررسی امکان ترکیب دو متغیر x2 و x4 و اندازه گیری متغیر پنهانی به نام استعداد فارسی نیز به همین ترتیب عمل می کنیم:

	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	X1	11.5	12.3	12.5	13.9	11.6	12.4	10.1	13.5	10.2	11.6
3	X2	11.8	12.2	14.4	14.5	12.3	12.8	14.2	14.8	13.5	14.1
4	X3	13.8	14.2	14.2	15.0	13.9	14.6	12.7	14.8	12.8	13.8
5	X4	14.4	14.1	15.6	16.0	14.7	14.9	15.7	16.7	15.5	15.8
6											
7	X2 bar	13.46		S=	1.084	0.748		R=	1	0.947	
8	X4 bar	15.34			0.748	0.574			0.947	1	
9											
10	بر اساس داده های خام	alpha=	$\frac{2}{2-1}$	$\left(1-\frac{1.084+0.574}{1.084+0.574+2*0.748}\right)$			=	0.948			
11											
12											
13	بر اساس داده های استاندارد شده	alpha=	$\frac{2}{2-1}$	$\left(1-\frac{1+1}{1+1+2*0.947}\right)$			=	0.973			
14											

تحلیل معیار آلفای کرونباخ

این معیار به ندرت خارج از فاصله ی (0,1) قرار می گیرد.

اگر بر اساس ماتریس کواریانس در نظر بگیریم:

$$Alpha=k/(1-k)[1-tr(s)/total(s)]$$

$$tr(s)/total(s)=\frac{\text{جمع واریانس ها}}{\text{جمع کواریانس ها+جمع واریانس ها}}$$

بزرگ بودن کواریانس بین دو متغیر بیانگر وجود همبستگی خطی هم جهت است. در نتیجه وقتی جمع کواریانس ها بیشتر باشد، کسر فوق کوچکتر شده و بر اساس فرمول، alpha بزرگتر می شود. پس نتیجه می گیریم که مقادیر بزرگ alpha حاکی از وجود همبستگی خطی هم جهت قوی بین آیتم ها می باشد.

و اگر بر اساس ماتریس همبستگی پیش برویم:

Alpha=k/(1-k)[1-tr(R)/total(R)]

$$tr(s)/total(s)=\frac{K}{K+\text{جمع همبستگی ها}}$$

در این صورت هم هرچه جمع همبستگی ها بیشتر باشد کسر فوق کوچکتر شده و alpha بزرگتر می باشد که این مسئله نشان دهندهٔ این است که آیتم ها هماهنگی قوی با هم داشته و متغیر پنهان واحدی را اندازه گیری می کنند.

قواعد مختلفی برای تعبیر آلفای کرونباخ وجود دارد.از جمله می توان به موارد زیر اشاره کرد:

۱- مقایسه ی alpha با عدد0.7

اگر alpha >0.7 باشد ← سؤالات مطلوب تلقی می شوند. درغیر این صورت متغیر واحدی را اندازه گیری نمی کنند.

۲- تعبیر آلفا به صورت زیر:

$Alpha \geq 0.9$	عالی
$Alpha \geq 0.8$	خوب
$Alpha \geq 0.7$	قابل قبول
$Alpha \geq 0.6$	مشکوک
$Alpha \geq 0.5$	ضعیف
$Alpha < 0.5$	غیر قابل قبول

استفاده از تحلیل مؤلفه های اصلی

روش دیگری که برای بررسی تک بعدی بودن وجود دارد، تحلیل واریانس مؤلفه های اصلی می باشد. به این ترتیب که اگر واریانس اولین مؤلفه ی اصلی در مقایسه با واریانس سایر مؤلفه ها خیلی بزرگ باشد، می توان نتیجه گرفت که آن آیتم ها متغیر واحدی را (که همان مؤلفه ی اصلی اول است) اندازه گیری می کنند.

*اگر در مراحل تحلیل از داده‌های خام استفاده شود← واریانس اولین مؤلفه ی اصلی = بزرگترین مقدار ویژه ی ماتریس کواریانس

* اگر در مراحل تحلیل از داده‌های استاندارد شده استفاده شود← واریانس اولین مؤلفه ی اصلی = بزرگترین مقدار ویژه ی ماتریس همبستگی

برای چگونگی تشخیص بزرگی مقدار ویژه ی مورد نظر قاعده ی سرانگشتی زیر وجود دارد.

فرض: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ ، مقادیر مرتب شده ی ماتریس کواریانس باشند.

میانگین آنها به صورت زیر تعریف می شود:

$$\bar{\lambda} = \sum_{i=1}^{k} \lambda i \,/_{k}$$

اگر $\lambda_i \geq \bar{\lambda}$ باشد، می توان گفت که λ_i به اندازه کافی بزرگ است.

حال برای بررسی تک بعدی بودن کافی است λ_1 بزرگتر یا مساوی $\bar{\lambda}$ باشد و بقیه λ ها کوچکتر از $\bar{\lambda}$ باشند.

اگر داده های استاندارد شده را به کار ببریم $\bar{\lambda} = 1$. درنتیجه برای قبول تک بعدی بودن کافی است که $\lambda_i \geq 1$ بوده و بقیه λ ها کوچکتر از ۱ باشند.

در بخش **استفاده از میانگین وزنی** این مورد با مثال توضیح داده می شود.

استفاده از میانگین حسابی

پس از آنکه متوجه شدیم که آیتم ها متغیر پنهان واحدی را اندازه گیری می کنند (از طریق روش های آلفای کرونباخ و با تحلیل مؤلفه های اصلی)، باید به این سؤال پاسخ داد که حال برای محاسبه ی مقادیر متغیر پنهان چگونه آیتم ها را با هم ترکیب کنیم. یکی از راه ها استفاده از میانگین حسابی می باشد.

برای مثال داده های درس ریاضی و فارسی را دوباره در نظر می گیریم:

فرض می کنیم F1 عامل اول است و استعداد ریاضی را می سنجد و F2 عامل دوم بوده و استعداد فارسی را اندازه می گیرد.

Xi = نمره ی دانشجو در درس i F1=1/2(x1+x3) F2=1/2(x2+x4)

	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	X1	11.5	12.3	12.5	13.9	11.6	12.4	10.1	13.5	10.2	11.6
3	X2	11.8	12.2	14.4	14.5	12.3	12.8	14.2	14.8	13.5	14.1
4	X3	13.8	14.2	14.2	15.0	13.9	14.6	12.7	14.8	12.8	13.8
5	X4	14.4	14.1	15.6	16.0	14.7	14.9	15.7	16.7	15.5	15.8
6											
7		1	2	3	4	5	6	7	8	9	10
8	F1	12.65	13.25	13.35	14.45	12.75	13.50	11.40	14.15	11.50	12.70
9	F2	13.10	13.15	15.00	15.25	13.50	13.85	14.95	15.75	14.50	14.95
10											
11	F1 bar	12.97		S=	0.89	0.19			R=	1	0.23
12	F2 bar	14.4			0.19	0.79				0.23	1

استفاده از میانگین وزنی

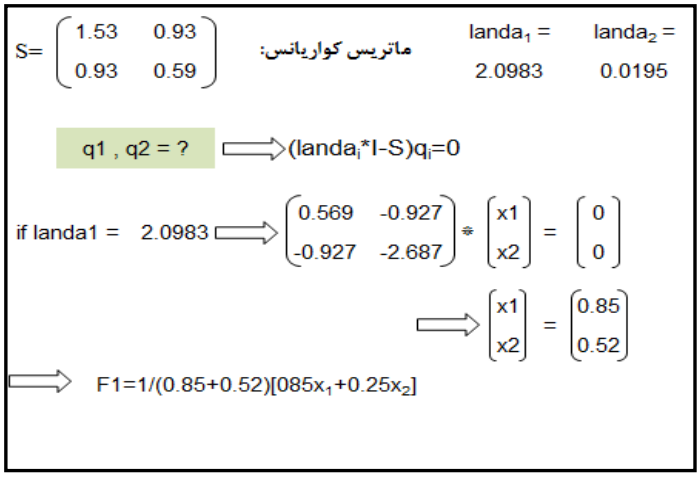
اگر ماهیت تک بعدی بودن آیتم ها پذیرفته شود، برای اندازه گیری متغیر پنهان مورد نظر به جای استفاده از میانگین حسابی که توضیح داده شد می توام از میانگین وزنی استفاده کرد. به این ترتیب که ضرایب اولین مؤلفه های اصلی در حکم وزن به کار گرفته می شوند. در نتیجه بیشترین تغییر پذیری موجود در آیتم ها به حساب می آیند.

همان طور که می دانیم و قبلاً نیز توضیحاتی ارائه شده است، ضرایب مؤلفه های اصلی در واقع همان بردارهای ویژه ی متناظر با مقادیر ویژه ی به دست آمده می باشند.

برای درک بیشتر استفاده از تحلیل مؤلفه های اصلی و کاربرد آن در روش تحلیل آیتمی مثال نمرات درس ریاضی و فارسی را بررسی می کنیم.

	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	X1	11.5	12.3	12.5	13.9	11.6	12.4	10.1	13.5	10.2	11.6
3	X2	11.8	12.2	14.4	14.5	12.3	12.8	14.2	14.8	13.5	14.1
4	X3	13.8	14.2	14.2	15.0	13.9	14.6	12.7	14.8	12.8	13.8
5	X4	14.4	14.1	15.6	16.0	14.7	14.9	15.7	16.7	15.5	15.8
6											
7	S=	1.53	0.93		ماتریس کواریانس:		s-landa*I=	landa-1.53	-0.93		
8		0.93	0.59					-0.93	landa-0.59		
9											
10	s-landa*I=	landa^2-2.12landa+0.041			معادله ی مشخصه:						
11	$landa_1$ =	2.0983									
12	$landa_2$ =	0.0195			⇒ $landa_1$ >landa bar ⇒					آیتم های X1 و X2	
13	landa bar =	1.0589								تک بعدی اند.	

پس از اینکه مشخص شد آیتم ها تک بعدی هستند می بایست وزن ها را به دست آورده و میانگین وزنی بگیریم: (یعنی باید بردارهای ویژه را حساب کنیم.)



همین روند را برای عامل F2 نیز انجام میدهیم و در نهایت به جدول زیر می رسیم:

	A	B	C	D	E	F	G	H	I	J	K
1		1	2	3	4	5	6	7	8	9	10
2	X1	11.5	12.3	12.5	13.9	11.6	12.4	10.1	13.5	10.2	11.6
3	X2	11.8	12.2	14.4	14.5	12.3	12.8	14.2	14.8	13.5	14.1
4	X3	13.8	14.2	14.2	15.0	13.9	14.6	12.7	14.8	12.8	13.8
5	X4	14.4	14.1	15.6	16.0	14.7	14.9	15.7	16.7	15.5	15.8
6											
7		1	2	3	4	5	6	7	8	9	10
8	F1	12.37	13.02	13.15	14.32	12.47	13.24	11.09	13.99	11.19	12.44
9	F2	12.88	12.99	14.90	15.13	13.30	13.68	14.83	15.59	14.33	14.81
10											

شناسایی نقاط پرت احتمالی در داده های چندمتغیره

یک داده پرت مشاهده ای است که بطور غیر عادی یا اتفاقی از وضعیت عمومی داده های تحت آزمایش و نسبت به قاعده ای که براساس آن آنالیز می شوند، انحراف داشته است. بطور ساده تر، داده پرت مقداری است که نسبت به مجموع نمرات توزیع در حد افراط یا تفریط قرار داشته باشد. عمده ترین دلایل پدیدارشدن داده های پرت در یک پژوهش و مطالعه عبارتند از:

۱ -پاسخگو به هر دلیلی قادر به پاسخگویی نباشد.

۲ -پاسخگو منظور درست پرسشگر یا سوال را متوجه نمیگردد و پاسخی غلط می دهد.

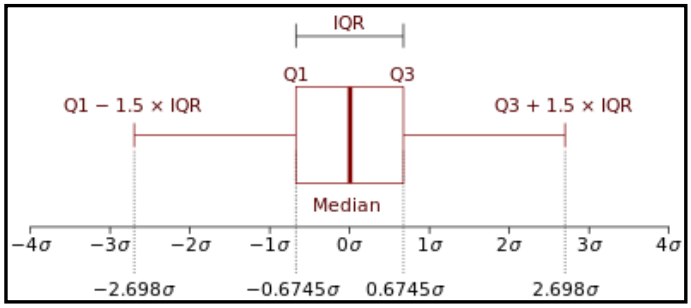
۳ -اشتباه در پرسیدن سوال توسط پرسشگر.

۴ -اشتباه پرسشگر در نوشتن پاسخ.

۵ -اشتباه در هنگام ورود اطلاعات به بانک های اطلاعاتی.

۶ -اشتباه در انتخاب نمونه مناسب با طرح و موضوع.

روش تشخیص داده های پرت: برای تشخیص مقادیر پرت از توصیه توکی که در نمودار های جعبه ای بیان گردید، استفاده می شود. برای این کار می بایست ابتدا نمودار جعبه ای رسم گردد. مقادیر بیش از سه برابر IQR (دامنه بین چارکی)بالاتر و پایین تر از گوشه های جعبه، مقادیر پرت شدید هستند. در بعضی از نرم افزارهای آماری با نام extreme شناخته می شوند. مقادیری که بین ۱،۵ تا ۳ برابر IQR بالاتر و پایین تر از گوشه های جعبه، مقادیر پرت جزئی هستند. در بعضی از نرم افزارهای آماری با نام outliers شناخته می شوند.



یکی از راه های پرکاربردی که می توان برای شناسایی داده های پرت مورد استفاده قرار داد تحلیل مؤلفه های اصلی می باشد. همان طور که توضیح داده شد، اولین مؤلفه ی اصلی که پس از محاسبه ی مقادیر ویژه و بردارهای ویژه و انجام محاسبات مربوطه به دست می آید، دارای بیشترین واریانس (λ_{max})است که نشان دهنده ی بالا بودن قابلیت اولین مؤلفه ی الی در شناسایی تغییرات می باشد. (می دانیم که واریانس مؤلفه های اصلی برابر با مقادیر ویژه ی متناظر با آنها می باشد.)در واقع اولین مؤلفه ی اصلی خطی است که امتداد آن منطبق با بیشترین پراکندگی قابل مشاهده در داده های اصلی است. دومین مؤلفه ی اصلی دارای واریانس λ_2 بوده که از نظر مقدار در رتبه ی دوم قرار دارد. مؤلفه های اصلی از مرکز داده های اصلی عبور کرده و دو به دو بر هم عمود می باشند.

راهکاری که با استفاده از تحلیل مؤلفه های اصلی برای شناسایی داده های پرت می توان استفاده نمود، رسم مؤلفه ی اصلی اول در مقابل مؤلفه ی اصلی دوم و در صورت نیاز سایر مؤلفه های اصلی متعاقب می باشد. (رسم مؤلفه ی اصلی اول در مقابل مؤلفه ی اصلی سوم و .....)

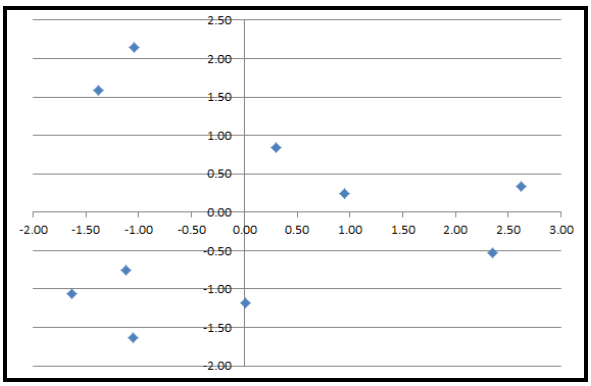
اگر در نمودار رسم شده، داده هایی موجود باشند که از محل تمرکز سایر داده ها به طور قابل ملاحظه ای فاصله داشته باشند، آن داده ها به عنوان داده های مشکوک یا پرت شناسایی می شوند و می توان پرت بودن آن ها را با آزمون G-B مورد بررسی قرار داد.

برای نمونه همان مثال نمرات دروس ریاضی و فارسی را بررسی می کنیم و نمودار مؤلفه ها را رسم مس نماییم. البته خود برخی از نرم افزارها خود قابلیت رسم این نمودار را دارند. مانند نرم افزار مینی تب که در ادامه نمودار آن را نشان خواهیم داد.

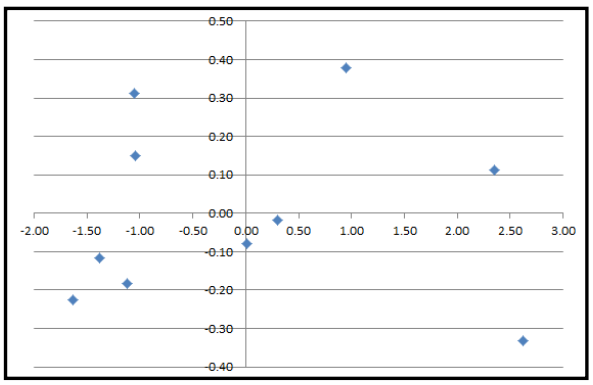
در این مثال پس از انجام محاسبات تحلیل مؤلفه های اصلی با نرم افزار مینی تب، ۴ عامل به ترتیب اولویت به صورت زیر به دست آمده اند:

	A	B	C	D	E	F	G	H	I	J	K
1					داده های اولیه						
2											
3		1	2	3	4	5	6	7	8	9	10
4	X1	11.5	12.3	12.5	13.9	11.6	12.4	10.1	13.5	10.2	11.6
5	X2	11.8	12.2	14.4	14.5	12.3	12.8	14.2	14.8	13.5	14.1
6	X3	13.8	14.2	14.2	15.0	13.9	14.6	12.7	14.8	12.8	13.8
7	X4	14.4	14.1	15.6	16.0	14.7	14.9	15.7	16.7	15.5	15.8
8											
9											
10					مؤلفه های اصلی استخراج شده						
11											
12		1	2	3	4	5	6	7	8	9	10
13	PC1	-1.64	-1.05	0.95	2.35	-1.12	0.01	-1.05	2.62	-1.38	0.30
14	PC2	-1.06	-1.64	0.25	-0.53	-0.75	-1.18	2.14	0.34	1.59	0.84
15	PC3	-0.22	0.31	0.38	0.11	-0.18	-0.08	0.15	-0.33	-0.12	-0.02
16	PC4	-0.09	-0.08	0.02	-0.09	0.00	0.23	0.01	-0.06	-0.06	0.11

نمودار اولین مؤلفه ی اصلی در مقابل دومین مؤلفه ی اصلی و نیز سومین مؤلفه ی اصلی را به صورت زیر رسم می کنیم.



مؤلفه ی اول در مقابل مؤلفه ی دوم



مؤلفه ی اول در مقابل مؤلفه ی سوم

پس با توجه به این گونه نمودارها در صورتی که داده ی پرت یا مشکوکی وجود داشته باشد به راحتی می توان آنها را شناسایی کرد.

تقلیل تعداد متغیرهای توضیحی در رگرسیون چندگانه

رگرسیون چندگانه مدلی است که در آن ارتباط موجود بین چندین متغیر مستقل (توضیحی) با یک متغیر وابسته (پاسخ) سنجیده می شود.

مسئله ی بسیار مهمی که در بحث رگرسیون چندگانه مطرح می باشد، انتخاب متغیرهای مستقل به نحوی است که رگرسیون بیشترین کارایی را داشته باشد. یعنی انتخاب متغیرهای مستقل باید به صورتی انجام پذیرد که از ورود متغیرهایی به مدل که اطلاعات خاصی به ما نمی دهند جلوگیری شود.

برای این منظور روش های مختلفی از جمله مقایسه ی R^2 (ضریب تعیین) و یا مقایسه ی MSE (میانگین مربعات خطا)برای حالت های مختلف وجود دارد. منظور از حالت های مختلف این است که مثلاً اگر n متغیر مستقل وجود دارد ترکیبات مختلف آنها را مورد ملاحظه قرار داده و R^2 یا MSE مدل های ساخته شده با آنها را بررسی کرده و بهترین ترکیب را انتخاب کنیم.

و اما روش دیگری که برای انتخاب متغیرهای مناسب می توانیم مورد توجه قرار دهیم، روش تحلیل مؤلفه های اصلی است.

همان گونه که گفته شد، مؤلفه های اصلی در روش PCA، بر اساس بردارهای ویژه ی ماتریس همبستگی و یا کواریانس داده ها به دست می آیند. اولین مؤلفه ی اصلی، مهمترین مؤلفه است و پراکندگی و تغییرات بیشتری را در نمونه پوشش می دهد. پس برای انتخاب متغیرهای مورد نیاز برای ورود به مدل رگرسیونی می توان از این روش استفاده کرد و آن تعداد از مؤلفه های اصلی مهم تر که دارای واریانس بالاتری می باشند را انتخاب نمود.

برای درک بیشتر موضوع به ارائه ی یک مثال می پردازیم:

فرض می کنیم رئیس بیمارستانی تمایل دارد رابطه ی بین رضایت بیمار(y) را با متغیر های زیر بسنجد. برای این کار یک نمونه ی ۲۳ تایی از بیماران انتخاب شده اند.

X1 = سن بیمار X2 = شدت بیماری X3 = سطح نگرانی

داده های جمع آوری شده به صورت زیر می باشند. برای انجام PCA از نرم افزار مینی تب استفاده می کنیم و نتایج به دست آمده در ادامه ارائه می شوند؛

Principal Component Analysis: x1; x2; x3

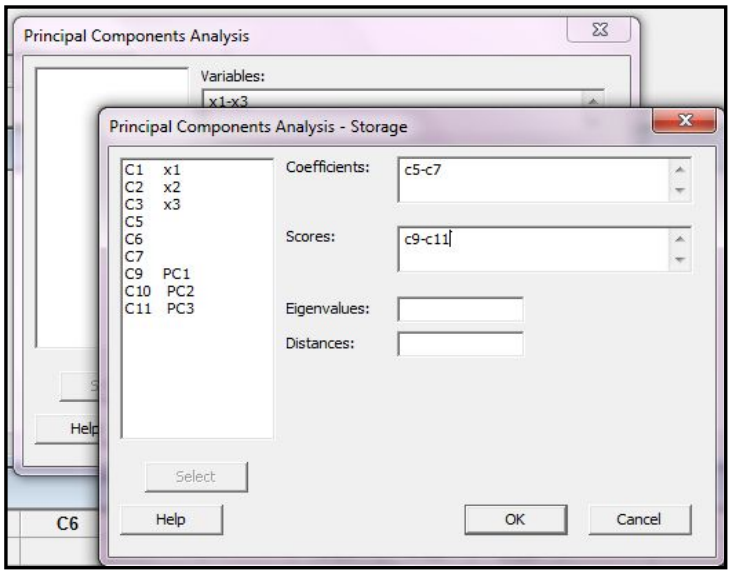
Eigenanalysis of the Correlation Matrix

Eigenvalue	2.1865	0.6090	0.2045
Proportion	0.729	0.203	0.068
Cumulative	0.729	0.932	1.000

Variable	PC1	PC2	PC3
x1	0.498	-0.866	-0.044
x2	0.609	0.386	-0.693
x3	0.617	0.319	0.720

y	x1	x2	x3
48	50	51	2.3
57	36	46	2.3
66	40	48	2.2
70	41	44	1.8
89	28	43	1.8
36	49	54	2.9
46	42	50	2.2
54	45	48	2.4
26	52	62	2.9
77	29	50	2.1
89	29	48	2.4
67	43	53	2.4
47	38	55	2.2
51	34	51	2.3
57	53	54	2.2
66	36	49	2.0
79	33	56	2.5
88	29	46	1.9
60	33	49	2.1
49	55	51	2.4
77	29	52	2.3
52	44	58	2.9
60	43	50	2.3

همانگونه که ملاحظه می شود مقادیر ویژه، نسبت تغییرات و نیز نسبت تغییرات تجمعی به دست آمده است. اعداد قسمت بعد نیز ضرایب مؤلفه های اصلی می باشند که در واقع همان بردار های ویژه ی متناظر آنها هستند. با ملاحظه ی اعداد به دست آمده به این نتیجه می رسیم که دو مؤلفه ی اول که همان "سن بیمار" و "شدت بیماری" می باشند بیشترین تغییرپذیری را در بر گرفته و در نتیجه مدل رگرسیونی می تواند بر مبنای آنها نوشته شود. یعنی متغیر سوم که "سطح نگرانی" می باشد تأثیر زیادی در "رضایت بیمار" که متغیر وابسته ی ما می باشد نمی گذارد. علاوه بر این نتایج، مینی تب قادر به ارائه ی مقدار مؤلفه های اصلی نیز می باشد. برای این کار کافی است در کادر محاوره ای storage محل قرار گرفتن آنها در worksheet را تعیین کنیم. در نتیجه نتایج به صورت زیر نمایش داده می شوند:



C9	C10	C11
PC1	PC2	PC3
0.65019	-1.03914	-0.077183
-0.85951	-0.04158	0.775271
-0.55291	-0.38195	0.205635
-1.85517	-1.25128	-0.122885
-2.75740	-0.00876	0.100177
2.22154	-0.04669	0.881198
-0.16090	-0.41286	-0.116792
0.14759	-0.68331	0.653666
3.49535	0.34098	-0.382739
-1.12909	0.81150	-0.286691
-0.79404	0.95278	0.736131
0.71567	-0.04477	-0.116323
0.28948	0.43019	-0.876473
-0.29143	0.59696	0.005282
1.03510	-1.19045	-0.797725
-1.05741	-0.09605	-0.403610
0.74174	1.34316	-0.296060
-2.08396	0.25436	-0.136255
-1.03084	0.31568	-0.151247
1.14757	-1.44545	0.133947
-0.44854	1.19504	-0.125008
2.47591	0.81180	0.282731
0.10108	-0.41015	0.114955

رگرسیون ستیغی

ابتدا یک مثال با استفاده از اکسل

در این جا یک مثال با حل در اکسل ارائه می‌شود که نشان می‌دهد که رگرسیون ستیغی چگونه اثرات متغیرهای چندگانه بر هم را درمان می‌نماید. ما در این جا یک مجموعه از سه متغیر داریم به نام‌های x_1، x_2 و x_1x_2 که ضرب آن دو است. معادله‌ی به دست آمده به این شکل طراحی شده است:

$$y = 3x_1 + 2x_2 + 0.5\ x_1x_2 + e$$

که در آن e میزان خطا است که دارای یک توزیع نرمال با میانگین صفر و انحراف معیار ۱۰ است.

تابع اکسل مربوطه برای وارد کردن خطا در داده ها را به این صورت طراحی می‌کنیم:

e =norminv(rand(),0,10)

همانطور که روشن است، rand() تابع تولید عدد تصادفی در اکسل است. برای تولید داده برای x_1 هم از x_1= rand() استفاده می‌نماییم. و ۸۰ نمونه برای x_1 به دست می‌آوریم و x_2 را هم به این شکل محاسبه می‌کنیم: x_2= x_1+rand()/100. وقتی که به روش حداقل مربعات رگرسیون مربوط به این داده‌ها را اجرا می‌کنیم به نتایج زیر می‌رسیم:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.990
R Square	0.980
Adjusted R Square	0.979
Standard Error	11.858
Observations	80

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	518,438.1	172,812.7	1,229.0	2.77059E-64
Residual	76	10,686.3	140.6		
Total	79	529,124.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	(2.30)	4.59	(0.50)	0.62	(11.44)	6.83
X1	84.38	466.19	0.18	0.86	(844.12)	1,012.88
X2	(78.54)	466.18	(0.17)	0.87	(1,007.01)	849.94
X1X2	0.46	0.04	10.49	0.00	0.37	0.55

شکل ۱ : نتایج حاصل از OLS

همانطور که دیده می‌شودضرایب به دست آمده از رگرسیون با روش حداقل مربعات نسبت به ضرایب موجود در معادله‌ی اولیه بسیار متفاوت است. مشخص است که این رگرسیون چندگانه تحت تأثیر اثر همخطی و تاثیر متغیرها بر هم قرار گرفته است.

استفاده از رگرسیون ستیغی

رگرسیون ستیغی همچنین با نام‌های رگرسیون تیخونو و نام‌های دیگر شناخته می‌شود. این رگرسیون متغیر جریمه‌ی λ را معرفی و از آن استفاده می‌کند. فرمول مربوطه به این شکل است:

$$\beta(k) = (X^TX + I\lambda)^{-1} X^TY$$

$\hat{\beta}$ ماتریس ضرایب برآورد شده، در این جا برداری از برآوردهای بتای تولید شده توسط رگرسیون ستیغی نامیده می‌شود.

در این مثال X ماتریسی دارای ۸۰ ردیف و ۳ ستون است که ستون‌ها حاوی مقادیر x_1، x_2و x_1x_2 می‌باشد. X^T نیز ماتریس ترانهاده‌ی ماتریس داده‌ها است. بر این اساس حاصل X^TX ماتریس ۳ در ۳ می‌باشد.

ماتریس I یک ماتریس یکانی سه در سه است که به جز قطر اصلی آن همه‌ی درایه‌های آن صفر می‌باشد. و قطر اصلی آن یک است. و λ ثابت ضرب شده در هر یک از ۱ها بر قطر اصلی است که با توجه به X^TX تعیین می‌گردد.

نتیجه در نهایت به یک ماتریس 1در۳ منتهی می‌شود که شامل متغیرهای توضیحی رگرسیون است.

چگونگی انتخاب λ

حال می‌خواهیم ببینیم که عامل λ را چگونه انتخاب نماییم. پارامتر تنظیم بهینه $\alpha(\lambda)$ معمولا ناشناخته است و اغلب با روش ad hoc مشخص می‌گردد. روش استفاده از تفسیر بیزی. روش اصل اختلاف، اعتبار متقاطع، روش منحنی L، حداکثر درستنمایی محدود شده و برآورد خطر پیش‌بینی بدون تبعیض روش‌های محاسبه‌ی آنها می‌باشد.

چگونگی راه اندازی صفحه گسترده

ما از دو تابع ماتریسی اکسل استفاده می‌کنیم : mmult(array1,array2) و minverse(array) بنابراین برای کثال ماتریس X^TX به صورت زیر تشکیل می‌شود:

X'X		
=MMULT(H3:CI5,D3:F82)	=MMULT(H3:CI5,D3:F82)	=MMULT(H3:CI5,D3:F82)
=MMULT(H3:CI5,D3:F82)	=MMULT(H3:CI5,D3:F82)	=MMULT(H3:CI5,D3:F82)
=MMULT(H3:CI5,D3:F82)	=MMULT(H3:CI5,D3:F82)	=MMULT(H3:CI5,D3:F82)

شکل۲: چگونگی به دست آمدن X^TX

در این جا محدوده‌ی H3:CI5 شامل X^T و محدوده‌ی D3:F82 شامل ۸۰ سطر در ۳ ستون ماتریس داده‌ی X می‌باشد. توجه داشته یاشید که برای محاسبه‌ی دترمینان در اکسل : بعد از انتخاب ماتریس‌های ورودی و خروجی و وارد کردن در فرمول باید در انتها کلیدهای Ctrl+Shift+Enter را فشار دهید. تمام عناصر ماتریس نتیجه محاسبه و نمایش داده می‌شود.

λI			$X^TX+\lambda I$			$(X^TX+\lambda I)^{-1}$		
=M11	0	0	=H8+L8	=I8+M8	=J8+N8	=MINVERSE(P8:R10)	=MINVERSE(P8:R10)	=MINVERSE(P8:R10)
0	=M11	0	=H9+L9	=I9+M9	=J9+N9	=MINVERSE(P8:R10)	=MINVERSE(P8:R10)	=MINVERSE(P8:R10)
0	0	=M11	=H10+L10	=I10+M10	=J10+N10	=MINVERSE(P8:R10)	=MINVERSE(P8:R10)	=MINVERSE(P8:R10)
λ	0.25							
			$(X^TX+\lambda I)^{-1}(X'y)$					
			=MMULT(T8:V10,H14:H16)					
			=MMULT(T8:V10,H14:H16)					
			=MMULT(T8:V10,H14:H16)					

شکل۳: چگونگی محاسبات نهایی

λ در سلول M11 است و در این جا مساوی ۰.۲۵ شده است. با توجه به آنچه در صفحه نمایش صفحه گسترده دیده می‌شود ضرایب به دست آمده به صورت زیر است.

$(X^TX+\lambda I)^{-1}(X'y)$
3.08174
2.22167
0.48429

شکل۳: خروجی رگرسیون ستیغی

همانطور از نتایج به دست می‌آید در این روش نتایج تقریبا مساوی فرمول اولیه است که بسیار مناسب می‌شود.

تشریح رگرسیون ستیغی

رگرسیون ستیغی روش برآورد دیگری است که وقتی متغیرهای پیش‌بینی همخطی بالایی دارند مورد استفاده قرار می‌گیرد. راه‌های زیادی برای تعریف و محاسبه‌ی برآوردهای ستیغی وجود دارد. ما روش مرتبط با اثر ستیغی را برگزیده‌ایم. این روش یک رویکرد نموداری است و آن را می‌توان یک تکنیک اکتشافی تلقی کرد. تحلیل ستیغی که از اثر ستیغی استفاده می‌کند یک رویکرد منحصر به فردی را برای مسائل مربوط به کشف و برآورد وقتی همخطی مورد سوءظن است ارائه می‌کند. برآوردگرهای حاصله اریب بوده ولی میانگین توان دوم خطای کمتری نسبت به برآوردگرهای OLS دارد(هورل و کنارد(۱۹۷۰) را ملاحظه کنید).

برآوردهای ستیغی ضرایب رگرسیون را با حل معادلاتی که اندک تفاوتی با معادلات نرمال (ارائه شده در فصل) دارند می‌توان به‌دست آورد. فرض کنید شکل استاندارد الگوی رگرسیون به صورتزیر است:

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \theta_2 \tilde{X}_2 + \ldots + \theta_p \tilde{X}_p + \varepsilon'$$

معادلات برآورد مربوط به ضرایب رگرسیون ستیغی عبارت‌اند از:

$$\begin{aligned}
&(1+k)\,\theta_1 + r_{12}\,\theta_2 + \ldots + r_{1p}\,\theta_p = r_{1y}\\
&r_{21}\,\theta_1 + (1+k)\theta_2 + \ldots + r_{2p}\,\theta_p = r_{2y}\\
&\quad\vdots \qquad\quad \vdots \qquad \vdots \qquad\quad \vdots\\
&r_{p1}\,\theta_1 + r_2\ \theta_2 + \ldots + (1+k)\theta_p = r_{py}
\end{aligned}$$

که در آن r_{ij} همبستگی بین متغیرهای پیشگوی i و j و r_{ij}همبستگی بین متغیر پیشگوی iام و متغیر پاسخ $\tilde{Y}$ است. برآورد ضرایب رگرسیون ستیغی یعنی $\theta_1, \ldots, \theta_p$ جواب دستگاه معادلات (۱۰-۱۶) است. برآورد ضرایب رگرسیون ستیغی را می‌توان به عنوان حاصل مجموعه داده‌هایی که اندکی تغییر یافته‌اند تلقی نمود. برای طرز عمل رسمی در این مورد پیوست این فصل را ملاحظه نمایید.

پارامتر اصلی که رگرسیون ستیغی را از OLS متمایز می‌سازد K است. باید توجه داشت که وقتی k=0 و $\hat{\theta}$ها برآوردهای OLS خواهند بود. پارامتر α را پارامتر اریبی می‌نامند. با افزایش k از صفر، اریبی برآوردها نیز افزایش می‌یابد. از طرفی واریانس کل (مجموع واریانس‌های برآوردهای ضرایب رگرسیون) عبارت است از:

$$\text{واریانس کل (k)} = \sum_{j=1}^{p} var(\hat{\theta}j(k)) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda j}{(\lambda j+k)^2}$$

که تابعی کاهشی نسبت به k است. فرمول(۱۰-۱۷) اثر پارامتر ستیغی روی واریانس کل برآوردهای ضرایب رگرسیون را نشان می‌دهد. اگر K=0 را در (۱۰-۱۷) قرار دهیم خواهیم داشت.

$$\text{واریانس کل(۰)} = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda j}$$

که اثر مقدار ویژه کوچک را روی واریانس کل برآوردهای OLS ضرایب رگرسیون نشان می‌دهد اگر k بدون کران افزایش پیدا کند تمام برآوردهای رگرسیون به صفر میل می‌کنند.

ایده‌ی رگرسیون ستیغی این است که مقداری از k را انتخاب کند که کاهش در واریانس کل با افزایش اریبی زیاد نشود.

آن که پیشتر بیان گردید که مقدار مثبتی برای k وجود دارد که به ازای آن برآوردهای ستیغی نسبت به تغییرات کم در برآورد داده‌ها پایدار خواهند بود. (هورل و کنارد ۱۹۷۰) در عمل با محاسبه $\theta_1, \ldots, \theta_P$ برای مقادیر kی بین ۰ و ۱ و رسم نتایج در مقابل مقدارk را انتخاب می‌کنیم. نمودار حاصل از اثر ستیغی نامیده و برای انتخاب یک مقدار مناسب k مورد استفاده قرار می‌دهیم. در مثال راهنمایی‌هایی را برای انتخاب k ارائه می‌کنیم.

برآورد با روش ستیغی

روشی برای همخطی چندگانه که از تحلیل ستیغی نتیجه می‌شود. با ناپایداری در ضرایب برآورد شده حاصل از تغییرات اندک در برآورد داده‌ها سروکار دارد. ناپایداری را در اثر ستیغی می‌توان مشاهده کرد. اثر ستیغی یک نمودار همزمان ضرایب رگرسیون

$\theta_1, \ldots, \theta_P$ است که در مقابل k برای مقادیر مختلف k نظیر 0.001 ، 0.002 و غیره رسم شده است. شکل ۱۰-۲ اثر ستیغی مربوط به داده‌های واردات است. این نمودار با توجه به جدول ۱۰-۷ رسم شده که این جدول برآورد ضرایب ستیغی را برای ۲۹ مقدار k که از ۰ تا ۱ تغییر می‌کند را دربردارد. به عنوان مثال مقادیر kی نزدیک به انتهای پایینی برد را انتخاب کرده‌ایم. اگر ضرایب برآورد شده نوسانات زیادی را برای مقادیر کوچک k نشان دهند در آن صورت ناپایداری ظاهر شده و احتمالا همخطی چندگانه وجود دارد.

آنچه که از اثر ستیغی با معادل با آن از جدول ۱۰-۷ پیداست این است که مقادیر برآورد شده‌ی ضرایب θ_1 و θ_3 برای مقادیر کوچک k کاملا ناپایدار است. برآورد θ_1 به سرعت از یک مقدار منفی غیر موجه 0.339- به یک مقدار پایدار در حدود 0.43 تغییر می‌کند. برآورد θ_3 از ۱.۳۰۳ شروع شده و در نزدیکی ۰.۵۰ پایدار می‌شود. ضریب $\tilde{X}_2$(STOCK) یعنی θ_2 تحت تأثیر همخطی چندگانه قرار نمی‌گیرد و در حدود ۰.۲۱ پایدار باقی می‌ماند.

جدول ۱۰-۷ برآوردهای ستیغی $\theta_j(k)$ به صورت توابعی از پارامتر ستیغی k برای داده‌های واردات (۱۹۵۹-۱۹۴۹)

$\theta_3(k)$	$\theta_2(k)$	$\theta_1(k)$	k
1.303	0.213	-0.339	0.000
1.080	0.215	-0.117	0.001
0.870	0.217	0.092	0.003
0.768	0.217	0.192	0.005
0.709	0.217	0.251	0.007
0.669	0.217	0.290	0.009
0.654	0.217	0.304	0.010
0.630	0.217	0.328	0.012
0.611	0.217	0.345	0.014
0.597	0.217	0.359	0.016

0.585	0.216	0.370	0.018
0.575	0.216	0.379	0.020
0.567	0.216	0.386	0.022
0.560	0.215	0.392	0.024
0.553	0.215	0.398	0.026
0.548	0.216	0.402	0.028
0.543	0.214	0.406	0.030
0.525	0.213	0.420	0.040
0.513	0.211	0.427	0.050
0.504	0.209	0.432	0.060
0.497	0.207	0.434	0.070
0.491	0.206	0.436	0.080
0.486	0.204	0.436	0.090
0.481	0.202	0.436	0.100
0.450	0.186	0.436	0.200
0.427	0.173	0.411	0.300
0.408	0.161	0.396	0.400
0.391	0.151	0.381	0.500
0.376	0.142	0.376	0.600
0.361	0.135	0.354	0.700
0.348	0.128	0.342	0.800
0.336	0.121	0.330	0.900
0.325	0.115	0.319	1.000

مرحله بعدی در تحلیل ستیغی انتخاب مقدار k و به دست آوردن برآوردهای مربوط به ضرایب رگرسیون است. اگر همخطی چندگانه یک مسئله‌ی جدی باشد، برآوردگرهای ستیغی را با افزایش آهسته‌ی k از صفر به طور عجیبی تغییر خواهد کرد. با افزایش k ضرایب سرانجام پایدار می‌شوند. چون k یک پارامتر اریبی است لذا انتخاب کمترین مقدار k که برای آن پایداری اتفاق می‌افتد. مطلوب خواهد بود زیرا مقدار k مستقیما به رابطه‌ی مقدار اریبی معرفی شده بستگی دارد برای انتخاب k چندین روش پیشنهاد شده است. این روش‌ها موارد زیر را در بر می‌گیرد:

۱.نقطه ثابت

هورل، کنارد و بالدوین(۱۹۷۵) برآورد k را با:

$$k=\frac{p\hat{\sigma}^{2}(0)}{\sum_{j=1}^{p}[\hat{\theta}_j(0)]^2}$$

پیشنهاد کرده‌اند که $\hat{\theta}_p(0),\ldots,\hat{\theta}_1(0)$ برآوردهای کمترین توان‌های دوم $\theta_p,\ldots,\theta_1$ هستند وقتی الگوی (۱۰-۱۵) به داده‌های برازش می‌شوند (یعنی وقتی $k = 0$) و $\hat{\sigma}^2(0)$ میانگین توان دوم خطای مربوطه است.

۲.روش تکراری

هورلد و کنارد (۱۹۷۶) روش تکراری زیر را برای انتخاب k پیشنهاد کرده‌اند: با برآوردهای اولیه‌ای از k (۱۰-۱۹) این روش آغاز می‌شود. این مقدار را k_0 می‌نامیم و سپس مقدار k_1 را محاسبه می‌کنیم:

$$k= \frac{P\sigma^{\wedge}(\cdot)}{\sum_{j=1}^{p}[\theta^{\wedge}j(k\circ)]^{\wedge}}$$

آن‌گاه از k_1 برای محاسبه‌ی k_2 به صورت زیر استفاده می‌کنیم:

$$k= \frac{P\sigma^{\wedge}(\cdot)}{\sum_{j=1}^{p}[\theta^{\wedge}j(k1)]^{\wedge}}$$

این فرآیند را تا اینکه اختلاف بین دو برآورد متوالی k قابل اغماض باشد تکرار می‌کنیم.

۳.اثر ستیغی

رفتار $\theta_j(k)$ به عنوان تابعی از k را به سهولت از اثر ستیغی می‌توان مشاهده کرد. مقدار k انتخاب شده، کوچکترین مقدار است که برای تمام ضرایب $\theta_j(k)$ پایدار می‌شوند. علاوه بر این در مقدار انتخابی k، مجموع توان‌های دوم مانده‌ها باید نزدیک مقدار مینیممش باقی بماند عامل افزایش واریانس، $VIF_j(k)$ نیز باید به کمتر از ۱۰ برسد.(یادآوری می‌کنیم که مقدار ۱ ویژگی یک سیستم متعامد و مقدار کمتر از ۱۰ نشان‌دهنده‌ی یک سیستم غیر همخط یا پایدار است.).

۴.سایر روش‌ها

بسیاری از روش‌های دیگر برآورد k نیز پیشنهاد شده است. به عنوان مثال می‌توانید مارکورت (۱۹۷۰)، مالوز (۱۹۷۳) گلدستین و اسمیس (۱۹۷۴)، مکدونالد و گالارنو (۱۹۷۵)، دمپستر و سایرین و (۱۹۷۷) و وهبا، گولب و هلس (۱۹۷۹) را ملاحظه کنید. در عین حال جذبه اثر ستیغی به ارائه نموداری اثراتی که همخطی روی ضرایب برآورد شده دارد مربوط می‌شود.

برای داده‌های واردات فرمول (۱۰-۱۹) مقدار زیر را حاصل می‌کند.

$$k = \frac{۳\times۰.۰۱۰۱}{(-۰.۳۳۹)^{۲}+(۰.۲۱۴)^{۲}+(۱.۳۰۳)^{۲}} = 0.0164$$

روش تکراری مقادیر k_0=0.0164 ، k_1 = 0.161 ، k_2 = 0.0161 را می‌دهد. بنابراین بعد از دو تکرار به k = 0.0161 همگرا می‌شود. اثر ستیغی شکل ۱۰-۲ (جدول ۱۰-۷ را نیز ملاحظه می‌کنید). برای kای تقریبا ۰.۰۴ پایدار می‌شود. بنابراین سه برآورد k (۰.۰۱۶۴،۰.۰۱۶۱،۰.۰۴) را داریم.

از جدول ۱۰–۷ می‌بینیم که در هر یک از این مقادیر علامت منفی نادرست برآورد θ_1 از بین می‌رود و این ضریب(در ۰.۳۵۹ برای k= 0.016 و در ۰.۴۲ برای k=0.04) پایدار می‌شود. از جدول ۱۰–۸ ملاحظه می کنیم که مجموع توان‌های دوم مانده‌ها (SSE(k)) فقط از ۰.۰۸۱ در k=0 به 0.108 در k=0.016 افزایش یافته و در k= 0.04 به 0.117 افزایش پیدا کرده است.

عوامل افزایش واریانس، $VIF_1(k)$ و $VIF_3(k)$ از 185 به مقادیر بین 1 و 4 کاهش پیدا کرده است.

واضح است که مقادیر k در فاصله‌ی (0.04 و 0.016) رضایت بخش به نظر می‌رسد.

ضرایب برآورد شده از الگو به واحدهای متغیرهای استاندارد شده و اصلی در جدول ۱۰–۹ خلاصه شده‌اند. ضرایب اولیه β_j از ضریب استاندارد θ_j با استفاده از (۱۰–۳) به دست می‌آید. برای مثال β_1 به صورت زیر محاسبه می‌شود.

$$\beta_{1j} = (S_y/S_1)\,\theta_1 = (4.5437/29.9995)(0.4196) = 0.0635$$

بدین ترتیب الگوی حاصل برحسب متغیرهای اولیه برازش شده با روش ستیغی با به کاربردن k=0.04 عبارت است از:

$$IMPORT = -8.5537 + 0.0635 \,.\, DOPROD + 0.5859 .,STOCK + 0.1156 \,.\, CONSUM$$

این معادله یک نمایش موجه‌ای از رابطه را می‌دهد. توجه کنید که معادله‌ی نهایی برای این داده‌ها به خصوص تفاوتی با نتیجه‌ای که از به کاربردن دو مؤلفه‌ی اصلی اول به دست آمد ندارد. (جدول ۱۰–۳ را ملاحظه کنید) گر چه دو روش محاسبه بسیار متفاوت‌اند.

رگرسیون ستیغی: چند تذکر

رگرسیون ستیغی ابزاری را برای قضاوت پایداری یک دسته‌ای از داده برای تحلیل با کمترین توان‌های دوم فراهم می‌سازد در وضعیت‌هایی با همخطی بالا، چنان‌که اشاره شد تغییرات کم در داده‌ها موجب تغییرات زیادی در ضرایب برآورد شده دارد رگرسیون ستیغی این شرط را آشکار می کند.

در این وضعیت‌ها باید با احتیاط از رگرسیون کمترین توان‌های دوم استفاده کرد. رگرسیون ستیغی برآوردهای استوارتری نسبت به برآوردهای کمترین توان‌های دوم برای تغییرات کم در داده‌ها فراهم می کند.

برآوردگرهای ستیغی از این نظر که تحت تأثیر تغییرات اندک در برآورد داده‌ها واقع نمی‌شوند پایدار هستند. به خاطر خاصیت میانگین توان دوم خطای کمتر، مقادیر ضرایب برآورد شده‌ی رگرسیون ستیغی انتظار می‌رود نسبت به برآوردهای OLS به مقادیر واقعی ضرایب رگرسیون نزدیکتر باشند. همچنین پیش‌بینی‌های متغیر پاسخ مربوط به مقادیر متغیرهای پیشگو که در مجموعه‌ی برآورد لحاظ نشده‌اند دقیق‌تر خواهند بود.

برآورد پارامتر اریبی K نسبتا ذهنی است. بسیاری از روش‌های دیگر برآورد K وجود دارند ولی در این مورد توافق عامی که کدام روش بهتر است وجود ندارد. صرف نظر از روش انتخاب برآورد پارامتر ستیغی K، پارامتر برآورد شده تحت تأثیر حضور نقاط دورافتاده در داده‌ها واقع می‌شود.

مجموع توان‌های دوم مانده‌ها، SSE(k) و عوامل افزایش واریانس $VIF_j(k)$ به عنوان تابعی از پازامتر ستیغی k برای داده‌های واردات (۱۹۵۹-۱۹۴۹).

$VIF_3(k)$	$VIF_2(k)$	$VIF_1(k)$	SSE(k)	k
186.00	1.02	186.11	0.081	0.000
98.98	1.01	99.04	0.084	0.001
41.78	1.00	41.80	0.091	0.003
22.99	0.99	23.00	0.096	0.005
14.57	0.99	14.58	0.010	0.007
10.09	0.98	10.09	0.103	0.009
8.60	0.98	8.60	0.104	0.010
6.48	0.98	6.48	0.106	0.012
5.08	0.97	5.08	0.107	0.014
4.10	0.97	4.10	0.108	0.016
3.39	0.97	3.39	0.109	0.018
2.86	0.96	2.86	0.110	0.020
2.45	0.96	5.45	0.111	0.022
2.13	0.95	2.13	0.112	0.024
1.88	0.95	1.88	0.113	0.026
1.67	0.95	1.67	0.113	0.028
1.50	0.94	1.50	0.114	0.030
0.98	0.93	0.98	0.118	0.040
0.72	0.91	0.72	0.120	0.050
0.58	0.89	0.58	0.123	0.060
0.49	0.87	0.49	0.127	0.070
0.43	0.86	0.43	0.131	0.080
0.39	0.84	0.39	0.135	0.090
0.35	0.83	0.35	0.140	0.100
0.24	0.69	0.24	0.205	0.200
0.20	0.59	0.20	0.298	0.300
0.18	0.51	0.18	0.411	0.400
0.17	0.44	0.17	0.536	0.500
0.15	0.39	0.15	0.679	0.600
0.14	0.35	0.14	0.819	0.700
0.13	0.31	0.13	0.967	0.800
0.12	0.28	0.12	1.116	0.900
0.11	0.25	0.11	1.267	1.000

برآوردهای OLS و ستیغی ضرایب رگرسیون مربوط به داده‌های واردات(۱۹۵۹-۱۹۴۹).

ستیغی (k=0)		OLS(K=0)		
ضرایب	ضرایب	ضرایب		
اولیه	استاندارد شده	ضرایب اولیه	استاندارد شده	متغیر
-8.5537	0	-10.13	0	ثابت
0.0635	0.4196	-0.0514	-63393	DOPROD
0.5859	0.2127	0.5869	0.213	STOCK
0.1156	0.5249	0.2868	1.3027	CONSUM
	R2 = 0.988		R2 = 0.992	

مانند روش مؤلفه‌های اصلی، معیار تصمیم‌گیری که بیان کند چه وقت برآوردگرهای ستیغی برتر از برآوردگرهای OLS هستند بستگی به مقادیر ضرایب رگرسیون واقعی در الگو دارد گر چه این مقادیر نمی‌توانند معلوم باشند ما هنوز هم پیشنهاد می‌کنیم در مواردی که همخطی چندگانه فرین مورد سوءظن است، تحلیل ستیغی مفید خواهد بود. ضرایب ستیغی می‌توانند یک تعبیر دیگر داده‌هایی که ممکن است منتهی به درک بهتر فرآیند مورد مطالعه شود را پیشنهاد کنند.

مشکل عملی دیگر رگرسیون ستیغی این است که در نرم‌افزارهای آماری انجام نشده است. اگر یک نرم‌افزار آماری برنامه‌ای را برای رگرسیون ستیغی نداشته باشد، برآوردهای رگرسیون ستیغی را از بسته نرم‌افزار کمترین توان‌های دوم استاندارد با استفاده از مجموعه داده‌هایی که قدری تغیر داده شده‌اند، می‌توان به دست آورد. به خصوص برآوردهای ستیغی ضرایب رگرسیون را از رگرسیون بر $X^*_1, \ldots, X^*_P$ بدون داشتن یک جمله‌ی ثابت در الگو به دست می‌آیند.

خلاصه

هر دو روش برآورد رگرسیون ستیغی و رگرسیون مؤلفه‌های اصلی، اطلاعات بیشتری را درباره‌ی داده‌هایی که تحلیل می‌شوند به ما می‌دهد. دیدیم که مقادیر ویژه ماتریس همبستگی متغیرهای مستقل نقش عمده‌ای را در کشف همخطی چندگانه و تحلیل اثرات آن ایفا می‌کند. برآوردهای رگرسیونی که با این روش‌ها حاصل می‌شود دارای اریبی بوده ولی می‌تواند دقیق‌تر از برآوردهای OLS با توجه به میانگین توان دوم خطا باشد. ارزیابی بهتر بودن دقت برای یک مسأله‌ی خاص امکانپذیر نیست زیرا مقایسه‌ی این دو روش با روش OLS نیاز به دانش مقادیر واقعی ضریب دارد. مع‌هذا وقتی همخطی چندگانه مورد سوءظن است توصیه می‌شود که مجموعه‌ای از برآوردها را علاوه بر برآوردهای OLSمحاسبه کنیم. این برآوردها ممکن است تعبیری از داده‌ها که قبلا در نظر گرفته نشده بود را پیشنهاد کند.

هیچ تأیید نظری قوی برای استفاده از روش‌های مؤلفه‌ی اصلی و رگرسیون ستیغی وجود ندارد. توصیه‌ی ما این است که از این روش‌ها وقتی همخطی چندگانه جدی است به عنوان ابزار تشخیص بصری برای قضاوت پایداری داده‌ها در تحلیل کمترین توان‌های دوم، استفاده می‌شود. وقتی تحلیل مؤلفه‌های اصلی یا تحلیل رگرسیون ستیغی ناپایداری در مجموعه داده خاصی را نشان می‌دهد، تحلیل‌گر ابتدا باید از رگرسیون ستیغی ناپایداری در مجموعه داده خاصی را نشان می‌دهد، تحلیل‌گر ابتدا باید از رگرسیون کمترین توان‌های دوم برای مجموعه خلاصه شده‌ای از متغیرها استفاده نماید. اگر رگرسیون کمترین توان‌های دوم رضایت بخش نبود یعنی VIF های بزرگ، ضرایب با علامت‌های غلط، عدد شرطی بزرگ است در آن صورت فقط باید از رگرسیون مؤلفه‌های اصلی یا رگرسیون ستیغی استفاده شود.