داده‌های پِـرت

همیشه باید داده‌هایی (اطلاعاتی) که وارد برنامه‌هایی مانند اِکسل یا SPSS می‌کنیم را بررسی و بازبینی کنیم. همواره احتمال دارد که در داده‌ها با مقادیر غیرعادی مواجه شویم. موارد غیرعادی می‌تواند شامل مقادیر تعریف نشده و مقادیر پِرت (دور افتاده) باشد. همواره قبل از انجام هرگونه تحلیل آماری بر روی داده‌ها، باید چاره‌ای در مورد مقادیر پرت بیندیشیم.

افرادی که اندازه‌های انتهایی یا غیرمعمول در یک متغیر واحد (تک متغیری) یا در ترکیبی از متغیرها (چندمتغیری) دارند، دور افتاده یا پرت نامیده می‌شوند. داده‌های پرت اغلب سه یا بیش از سه واحد انحراف معیار (SD ±۳) از میانگین مربوط به خودشان فاصله دارند که از مشکلات احتمالی در ابزار اندازه‌گیری، شیوه ثبت یا ضبط پاسخ‌ها یا عضویت شرکت‌کنندگان در جامعه‌ای که فرض می‌شود از آن نمونه‌گیری شده است، ناشی می‌شود. حضور داده‌های پرت می‌تواند نتایج تحلیل را به گونه‌ای نامطلوب تحت تأثیر قرار دهد(تحریف کند). به همین دلیل بیشتر متخصصان پیشنهاد می‌کنند که اندازه‌های پرت قبل از تحلیل داده‌ها باید حذف شوند(میزر و دیگران،۱۳۹۱:۲۳۶).

انواع داده‌های پرت

داده‌های پرت را می‌توان در دو دسته داده‌های پرت تک متغیری و داده‌های پرت چندمتغیری تقسیم کرد:

۱) داده‌های پرت تک متغیری

داده‌های پرت تک متغیری مربوط به یک متغیر می‌شوند. به عنوان مثال وقتی که در یک پژوهش دانشجویی در زمینه میزان رضایت مردم از عملکرد شهرداری تهران؛ ما در متغیر سن افراد با عدد ۱۵۰ روبرو می‌شویم!، به احتمال زیاد با داده پرت مواجه شده ایم. چرا که می‌دانیم احتمال وجود فردی با چنین سن و سالی بسیار بعید است! و یا وقتی که در متغیر درآمد، شخصی درآمد ماهانه خود را از یک کار تمام وقت ۲۵ هزار تومان اعلام می‌کند و یا وقتی که در پاسخ سوالی که از فرد می‌پرسیم تا چه اندازه به آینده امیدوار است و او باید میزان رضایت خود را از عدد ۱ (به معنای خیلی کم) تا عدد ۵ (به معنی خیلی زیاد) اعلام کند، در فایل داده‌ها با عدد ۶ روبرو می شویم (به دلیل اشتباه در ورود داده)، همگی نشان از وجود داده‌های پرت تک متغیری دارد که نخست باید آن‌ها را شناسایی کرد و سپس در مورد آن‌ها چاره‌ای اندیشید.

البته زمانی که با متغیرهای کیفی (اسمی و ترتیبی) سروکار داریم گاهی با مقادیری در داده‌ها روبرو می‌شویم که داده پرت محسوب نمی‌شوند اما مقادیری هستند که به اشتباه وارد شده‌اند و باید حذف شوند. مثلا در متغیر جنس، اگر ما زنان را با کد ۱ و مردان را با کد ۲ تعریف کرده باشیم و در این‌حال با عدد ۱.۵ در داده‌ها مواجه شویم؛ با داده پرت مواجه نیستیم اما با داده‌های اشتباه مواجه شده‌ایم (به دلیل اشتباه پاسخگو در پاسخ به سوال یا اشتباه در ورود داده) و باید آن‌ها را شناسایی کرده و حذف یا اصلاح نماییم.

شناسایی داده‌های پرت تک متغیری

برای شناسایی داده‌های پرت تک متغیری باید از **جدول فراوانی و نمودار جعبه‌ای** استفاده کرد. از جدول فراوانی برای شناسایی داده‌های پرت در متغیرهای اسمی و ترتیبی استفاده می‌کنیم و از نمودار جعبه‌ای برای شناسایی داده‌های پرت

در متغیرهای فاصله‌ای/نسبی. البته از جدول فراوانی هم می‌توان برای شناسایی داده‌های پرت در متغیرهای فاصله-ای/نسبی استفاده کرد ولی نمودار جعبه‌ای برتری دارد و آسان‌تر است.

الف) جداول فراوانی

از جدول فراوانی برای کشف مقادیر پرت تک متغیری در متغیرهای اسمی و ترتیبی استفاده می‌کنیم. متغیرهایی مثل جنس، وضعیت تاهل، قومیت، تحصیلات و درآمد (هر دو به صورت چندگزینه‌ای و ترتیبی سنجیده شده باشند، مثلا تحصیلات در قالب سوالات دیپلم، فوق دیپلم، لیسانس و... سنجیده شده باشد) و یا تمام سوالاتی که در قالب طیف لیکرت سنجیده شده باشند. یعنی سوالاتی که پاسخ‌های آنان معمولا ۳ تا ۷ گزینه دارد و پاسخ‌هایی مثل کاملا موافقم تا کاملا مخالفم، اصلا تا همیشه و خیلی‌کم تا خیلی‌زیاد را در بر می‌گیرد. همچنین اگر متغیری فاصله‌ای/نسبی داشته باشیم که تعداد طبقات آن محدود (مثلا حدود ۱۰ طبقه) باشد، می‌توانیم از جدول فراوانی استفاده کنیم.

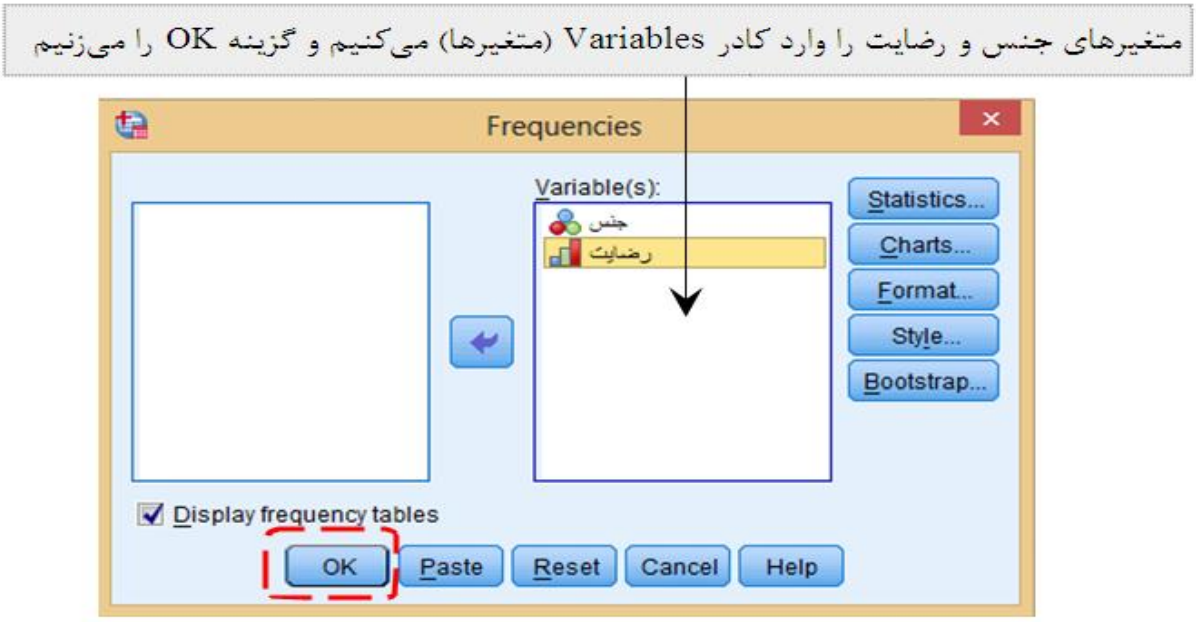
مثـــال

در یک پژوهش (فرضی) از دانشجویان دختر و پسر دانشگاه شهید بهشتی خواسته شد تا میزان رضایت خودشان از عملکرد ریاست دانشگاه را اعلام کنند. بر این اساس از دانشجویان تعدادی سوال پرسیده شد که دو سوال آن عبارت بود از جنس دانشجویان و میزان رضایت‌شان از عملکرد ریاست دانشگاه. جنس دانشجویان شامل دو جنس (دختر کد ۱، و پسر کد ۲) و میزان رضایت در طیف لیکرت ۵ گزینه‌ای (خیلی کم کد ۱، کم کد ۲ ، متوسط کد ۳، زیاد کد ۴ و خیلی زیاد کد ۵) سنجیده شد. همان‌طور که مشاهده می‌شود ما هنگام ورود اطلاعات مربوط به جنس افراد به دانشجویان دختر کد یا عدد ۱ و به دانشجویان پسر کد ۲ داده‌ایم و در فایل داده‌ها و خروجی (برون‌داد) مربوط به آن، تنها باید عدد ۱ و عدد ۲ مشاهده کنیم. در مورد متغیر میزان رضایت هم تنها باید اعداد ۱، ۲، ۳، ۴ و ۵ را مشاهده کنیم و نباید اعداد دیگری را (مثلا ۶، ۱.۵، ۲۰) مشاهده کنیم.

اجـــرا:

دستور فراوانی را اجرا می‌کنیم:

Analyze ---> Descriptive Statistics ---> Frequencies



نتـــایج:

نتایج جدول فراوانی دو متغیر جنس و میزان رضایت در ادامه ارائه شده است. در جدول فراوانی جنس افراد مقادیر پرت مشاهده نمی‌شود، چرا که تنها دو کد یا طبقه ۱ و ۲ (دختر و پسر) وجود دارند. توجه شود که داده‌های گمشده (Missing) جزء داده‌های پرت به حساب نمی‌آیند. ما در فایل داده‌ها مقادیر گمشده را با عدد ۹ نشان داده‌ایم و در فایل خروجی اعداد گمشده با عدد ۹ ظاهر شده‌اند. به غیر از اعداد ۱ و ۲ و مقادیر گمشده، عدد دیگری در فایل خروجی جنس دانشجویان دیده نمی‌شود و بدین معناست که در متغیر جنس دانشجویان داده پرت وجود ندارد.

اما در متغیر میزان رضایت ما با اعدادی غیر از ۱ ، ۲، ۳، ۴ و ۵ مواجه‌ایم و این اعداد مقادیر گمشده هم نیستند و نشان می‌دهد که دو مقدار پرت در داده‌ها وجود دارد (۱.۳ و ۲۲) که باید در فایل داده‌ها شناسایی و حذف شود. چون پاسخگویان تنها می‌توانستند یکی از اعداد ۱، ۲، ۳، ۴ و ۵ را انتخاب کنند در نتیجه اعداد دیگری که وجود دارند (۱.۳ و ۲۲) مقادیر پرت حساب می‌شوند و باید از تحلیل حذف شوند.

لازم به ذکر است که عدد ۱.۳ داده پرت به حساب نمی‌آید و یک داده غیرعادی و تعریف نشده است. در این‌جا به جهت آسان‌تر شدن آموزش، داده‌های غیرعادی و تعریف نشده در ارتباط با متغیرهای اسمی و ترتیبی را داده پرت به حساب آورده‌ایم.

جنس

		Frequency فراوانی	Percent درصد فراوانی	Valid Percent درصد معتبر	Cumulative Percent درصد تجمعی
Valid	مرد	133	66.5	66.8	66.8
	زن	66	33.0	33.2	100.0
	Total	199	99.5	100.0	
Missing	9	1	.5		
Total		200	100.0		

میزان رضایت

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	94	47.0	47.0	47.0
	1.3	1	.5	.5	47.5
	2	48	24.0	24.0	71.5
	3	37	18.5	18.5	90.0
	4	12	6.0	6.0	96.0
	5	7	3.5	3.5	99.5
	22	1	.5	.5	100.0
	Total	200	100.0	100.0	

ب) نمودار جعبه‌ای

اگر متغیرهایی که سنجیدیم از نوع متغیرهای فاصله‌ای/نسبی باشند هم می‌توان از جداول فراوانی استفاده کرد و هم از نمودار جعبه ای. البته پیشنهاد می‌شود از نمودار جعبه‌ای استفاده شود زیرا در نمودار جعبه‌ای داده‌های پرت در داخل خود نمودار مشخص می‌شود و شماره موردی (پاسخگویی) که دارای داده پرت در فایل داده‌هاست در نمودار مشخص می‌شود. همچنین مبنای انتخاب داده پرت در نمودار جعبه‌ای، داشتن فاصله‌ای به اندازه حداقل ۳± واحد انحراف استاندارد با میانگین است که به صورت خودکار توسط برنامه محاسبه می‌شود و در نمودار جعبه‌ای نشان داده می‌شود.

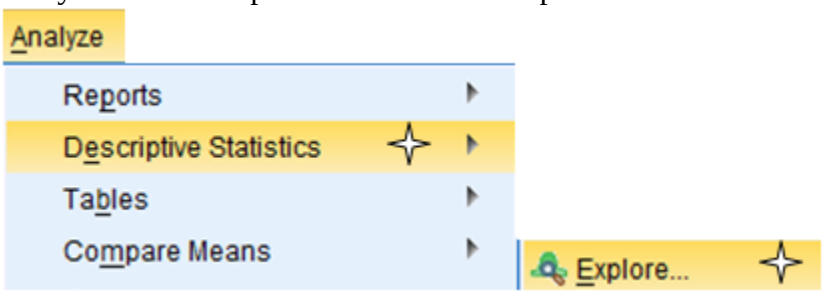
مثـــال

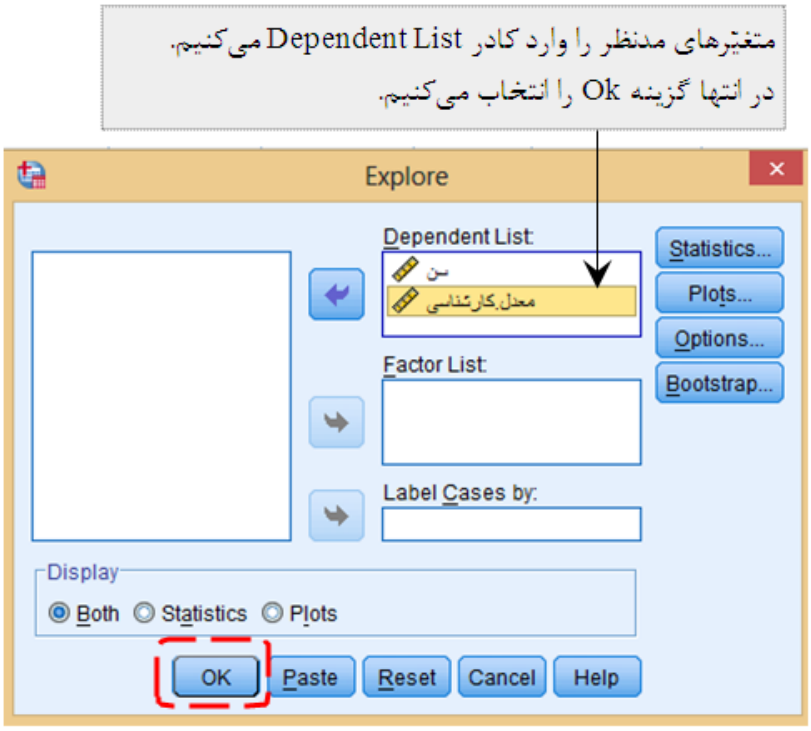
در پرسشنامه رضایت دانشجویان علاوه بر سوالات جنس و میزان رضایت، سن و معدل کارشناسی دانشجویان هم پرسیده شد. سن و معدل دانشجویان در سطح سنجش فاصله‌ای/نسبی سنجیده شد. دو نمودار جعبه‌ای سن و معدل دانشجویان در ادامه آورده شده است.

اجـــرا:

دستور اجرای نمودار جعبه‌ای در دستور Explore است. مراحل زیر را دنبال می‌کنیم:

Analyze ---> Descriptive Statistics ---> Explore



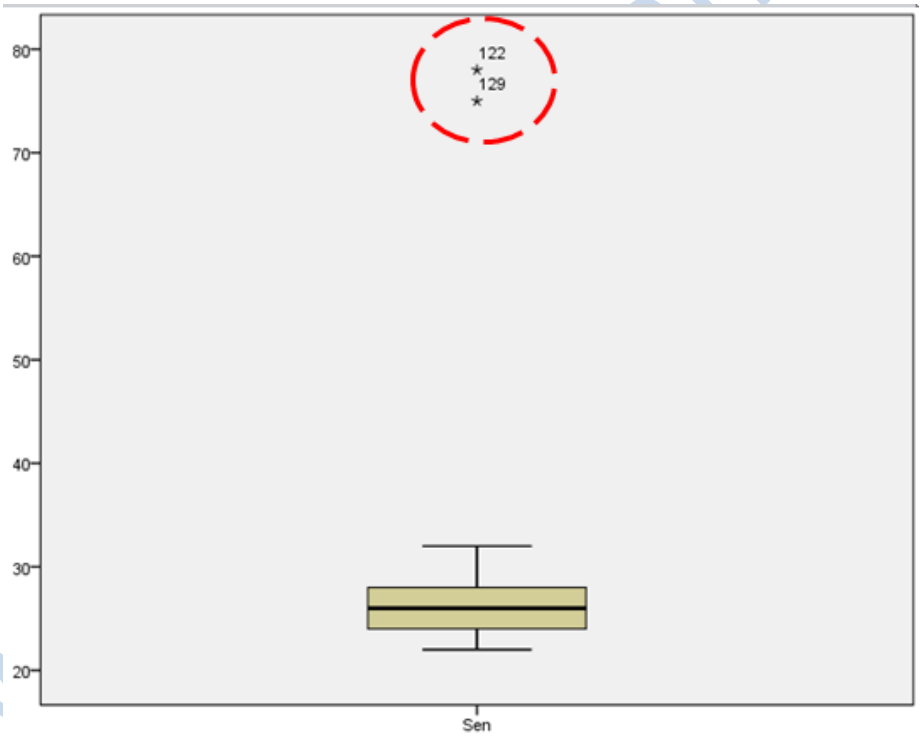


نتـــایج:

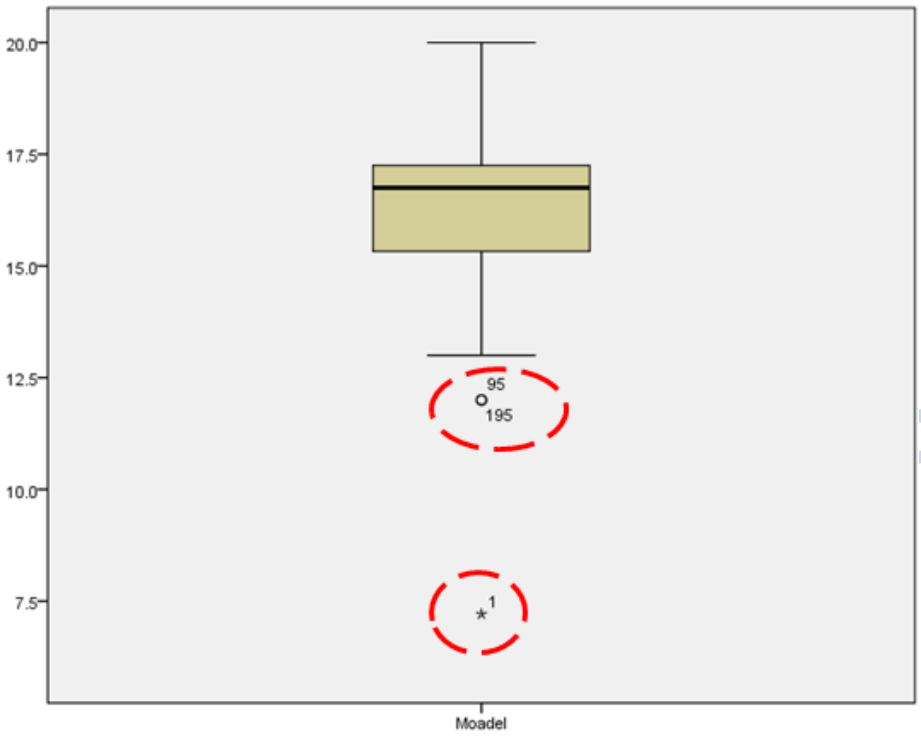
نمودار سن افراد نشان می‌دهد دو مورد دارای مقادیر پرت هستند. شماره این دو مورد در نمودار با علامت ستاره یا دایره کوچک مشخص شده است. افراد شماره ۱۲۲ و ۱۲۹ دارای مقادیر پرت هستند و به احتمال زیاد باید از داده‌ها حذف شوند. در فایل داده‌ها، افراد (پرسشنامه‌های) شماره ۱۲۲ و ۱۲۹ را پیدا کرده و سن آنان را نگاه می‌کنیم. سن این افراد

به ترتیب ۷۸ و ۷۵ سال است. با توجه به این که احتمال وجود دانشجویانی با چنین سنی بسیار بعید است سن این افراد باید از تحلیل حذف شود.

در متغیر معدل سه داده پرت وجود دارد که شامل موردهای ۹۵، ۱۹۵ و ۱ می‌شود. معدل این افراد به ترتیب ۱۲، ۱۱.۲۵ و ۵.۲۱ است. تصمیم در مورد حذف این داده‌ها با پژوهش‌گر است. معدل این افراد حداقل ۳± واحد انحراف استاندارد با معدل کل دانشجویان فاصله دارد. معدل کل دانشجویان برابر با ۱۶.۵۰ و انحراف استاندارد ۱.۷۵ است. به نظر می‌رسد با این که معدل‌های ۱۲ و ۱۱.۲۵، معدل پایینی است و احتمال این که دانشجویی چنین معدل‌هایی داشته باشد اندک است؛ اما همیشه چنین دانشجویانی وجود داشته اند و فرض وجود چنین معدل‌هایی محتمل است. بنابراین موردهای ۹۵ و ۱۹۵ باقی مانده و حذف نمی‌شوند. ولی مورد ۱ که دارای معدل ۵.۲۱ است حذف می‌شود، چون دارای معدل خیلی پایینی است. به احتمال زیاد این فرد معدل خود را به عمد اشتباه اعلام کرده است و یا این که در هنگام ورود داده‌ها از پرسشنامه به برنامه اشتباهی صورت گرفته است.



شکل ۲-۳- نمودار جعبه‌ای متغیر سن



شکل ۲-۴- نمودار جعبه‌ای معدل

۲) داده‌های پرت چند متغیری

بعد از بررسی داده‌ها برای شناسایی داده‌های پرت تک متغیری، سنجش داده‌های پرت چند متغیری انجام می‌شود. یک روش عینی برای ارزیابی وجود داده‌های پرت چندمتغیری محاسبه **فاصله مهالانوبیس** هر فرد است. آماره فاصله مهالانوبیس یعنی D^2، «فاصله» چندمتغیری بین هر فرد و میانگین چندمتغیری گروه را (که کانون نامیده می‌شود) اندازه‌گیری می‌کند.

هر فرد با استفاده از توزیع مجذورکای با سطح آلفای دقیق ۰۰۱/. ارزیابی می‌شود. افرادی را که به این آستانه معنی‌داری می‌رسندمی‌توان به عنوان موارد پرت چندمتغیری تلقی کرد و به احتمال باید از نمونه حذف شود (میزر و دیگران،۱۳۹۱:۱۰۶). چنانچه در سطح آلفای ۰۰۱/. (سطح معنی‌داری کمتر از ۰۰۱/.) مقدار مجذورکای به دست آمده معنی‌دار باشد، نشان می‌دهد مورد یا فرد موردنظر دارای داده‌پرت چندمتغیری (در آن تعداد متغیرها) است.

مثـــال

در پژوهشی که بر روی کارمندان یک اداره دولتی انجام شد میزان رضایت شغلی، استرس شغلی و تعهد شغلی آنان سنجیده شد. سوالات در قالب طیف لیکرت طرح شد. می‌خواهیم وجود داده پرت چندمتغیری را در متغیرهای فوق بسنجیم.

☑نکته: دستور شناسایی داده‌های پرت چندمتغیری در فرمان رگرسیون خطّی است.

اجـــرا:

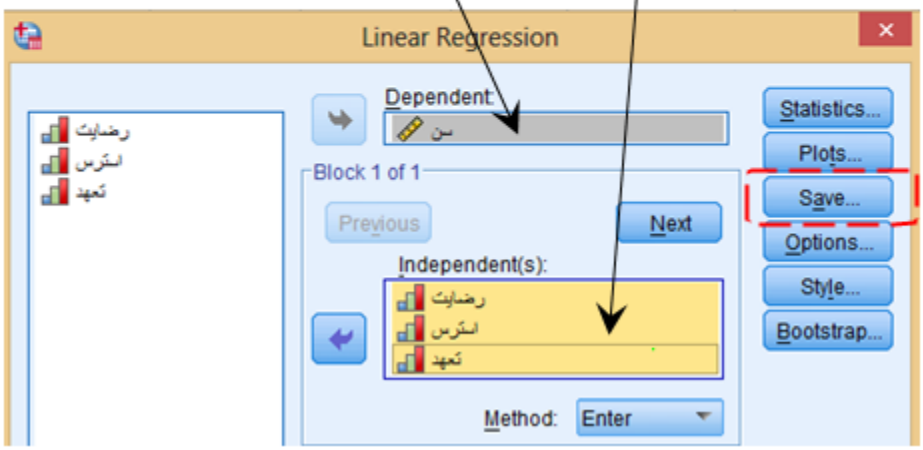
دستور زیر را اجرا می‌کنیم:

Analyze ---> Regression ---> Linear

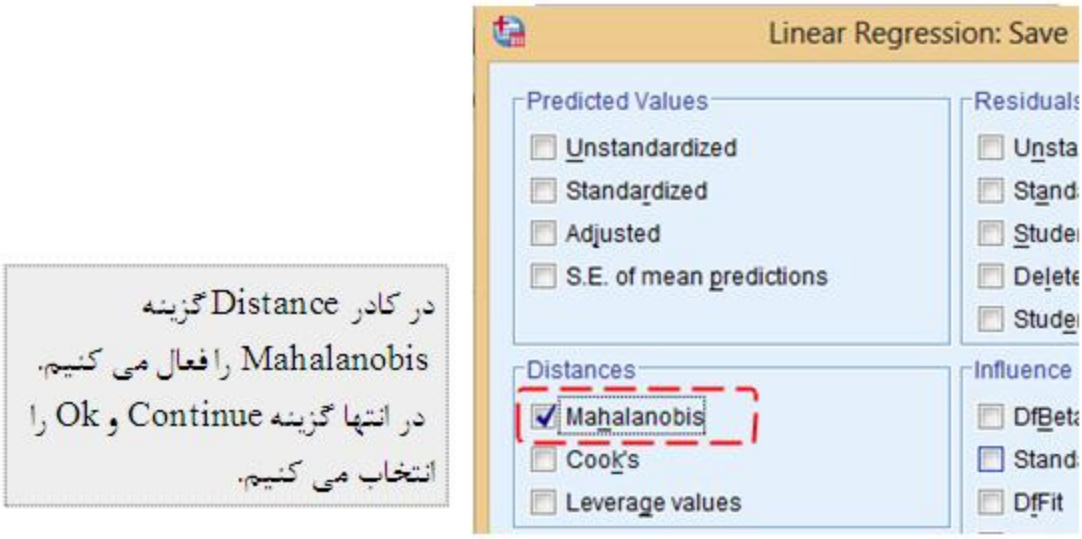
سه متغیر رضایت شغلی، استرس شغلی و تعهد شغلی را وارد کادر Independent می‌کنیم.

یک متغیر دیگر را وارد کادر Dependent می کنیم(سن افراد).

گزینه Save را انتخاب می کنیم.



نکته: اهمیتی ندارد که چه متغیری را به عنوان متغیر وابسته وارد کادر Dependent کنیم. تعریف یک متغیر وابسته تنها جهت اجرای رگرسیون است و غیر از سه متغیر رضایت شغلی، استرس شغلی و تعهد شغلی می‌توانیم هر متغیر دیگری را وارد کادر Dependent کنیم. در کادر Dependent، متغیری غیر از متغیرهایی که می‌خواهیم پرت بودنشان را بسنجیم، قرار می-دهیم.



نتـــایج:

حاصل این دستورات، خروجی رگرسیون و یک متغیر جدید در فایل داده‌ها است. متغیر جدید « Mah_1 » نام دارد که در فایل داده‌ها تشکیل می‌شود و آخرین متغیر است.

برای ارزیابی داده‌های پرت چندمتغیره باید مقادیر به‌دست آمده برای فاصله مهالانوبیس را با توزیع مجذور کای (کای اسکوئر) مقایسه می‌کنیم(جدول توزیع مجذور کای در انتهای برخی کتب آماری یا در سایت‌های آماری موجود است[1]). برای این مقایسه ابتدا درجه آزادی فاصله مهالانوبیس را به دست می‌آوریم که از تفریق تعداد متغیرهای مستقل (وارد شده در کادر Independent رگرسیون) منهای عدد یک به دست می‌آید. در این مثال ما سه متغیر مستقل داریم و در نتیجه درجه آزادی برابر با ۲ است(۳ متغیر مستقل داشتیم که از عدد ۱ کم می شود و عدد به دست آمده درجه آزادی نام دارد که برابر با عدد ۲ است). مقدار مجذور کای متناظر با درجه آزادی ۲، عدد ۱۳.۸۲ است و هر مورد یا پاسخگویی که فاصله مهالانوبیس آن از عدد مذکور بیشتر باشد داده پرت محسوب می‌شود. برای یافتن اعداد بزرگتر از ۱۳.۸۲ در فایل داده‌ها و در متغیر Mah_1 بهتر است بر روی نام متغیر در پنجره Data view کلیک راست کنیم و گزینه Sort Descending را انتخاب کنیم تا داده‌ها از زیاد به کم مرتب شوند. حال کافیست اعداد بزرگتر از ۱۳.۸۲ را پیدا کنیم و سپس مورد یا پاسخگوی مورد نظر را از فایل داده‌ها حذف کنیم.

مقدار مجذورکای برای درجه‌های آزادی ۱ تا ۸ در جدول بعد گزارش شده است. مقادیر مجذورکای ارائه شده، برای سطح آلفای ۰۰۱/. کاربرد دارد.

[1] به عنوان مثال می توانید به سایت www.Kharazmi-Statistics.ir رجوع کنید.

درجه آزادی	۸	۷	۶	۵	۴	۳	۲	۱
مقدار مجذور کای	۲۶.۱۲	۲۴.۳۲	۲۲.۴۶	۲۰.۵۲	۱۸.۴۷	۱۶.۲۷	۱۳.۸۲	۱۰.۸۳

نتایج بدست آمده نشان می‌دهد که بالاترین عدد بدست آمده برابر با ۱۰.۴۴ است که کمتر از مقدار۱۳.۸۲ است و نشان می دهد داده یا موارد پرت چندمتغیری در داده های ما وجود ندارد.

MAH_1	var
10.44309	
10.07063	
9.02496	
9.02496	
8.70621	
8.53228	