آموزش

رگرسیون لجستیک

با SPSS

تدوین: رامین کریمی
www.kharazmi-statistics.ir

۱) مقدمه

رگرسیون لجستیک و تحلیل تشخیصی روندهای آماری هستند که **برای پیش بینی عضویت گروهی به کار می روند نه پیش بینی نمره**. هر دو روند به ما امکان می دهند تا یک متغیر وابسته طبقه ای را بر اساس تعدادی متغیر پیش‌بین یا مستقل پیش بینی کنیم. این متغیرهای مستقل معمولا پیوسته هستند، اما رگرسیون لوجستیک از عهده متغیرهای مستقل طبقه ای برمی آید.به طور کلی رگرسیون لجستیک در مقایسه با تحلیل تشخیصی می تواند در گستره وسیع تری از موقعیت ها به کار رود. **تفسیر نتایج رگرسیون لجستیک نیز آسان تر از تفسیر نتایج تحلیل تشخیصی است**(بریس و همکاران، ۱۳۹۱:۳۹۴).

مثلا زمانی که بخواهیم نمره افراد (میزان درآمد) افراد را بر حسب سابقه شغلی و میزان تحصیلاتی که دارند پیش بینی کنیم از روش رگرسیون چند متغیره استفاده می کنیم، اما زمانی که بخواهیم پیش بینی کنیم که هر فرد برحسب سابقه شغلی و میزان تحصیلاتی که دارد در کدام گروه درآمدی (درآمد بالا یا درآمدپایین) قرار می گیرد، از روش رگرسیون لجستیک بهره می گیریم. از رگرسیون لجستیک برای پیش بینی عضویت طبقه برای مواردی که از قبل عضویت در آن مشخص نیست استفاده می کنیم.

در تعریفی دیگر، زمانی که متغیر وابسته در سطح اسمی است و متغیرهای مستقل، هم ترتیبی/اسمی و هم فاصله ای هستند، روش های رگرسیون خطی معمولی و تحلیل تشخیصی، مقدار برآوردها را کمتر از مقدار واقعی نشان می دهند، در این وضعیت از رگرسیون لجستیک استفاده می شود.

رگرسیون لجستیک نسبت به تحلیل تشخیصی ارجحیت دارد و مهم‌ترین دلیل آن است که درتحلیل تشخیصی گاهی اوقات احتمال وقوع یک پدیده خارج از طیف ۰ تا ۱ قرار می گیرد و متغیرهای پیش بین نیز باید دارای توزیع نرمال چندمتغیره باشند. در حالی که در رگرسیون لجستیک، **احتمال وقوع یک پدیده در داخل ۰ تا ۱ قرار دارد و رعایت پیش فرض نرمال بودن متغیرهای پیش بین لازم نیست**(سرمد، ۱۳۸۴:۳۳).

در مجموع زمانی از رگرسیون لجستیک استفاده می کنیم که شرایط زیر برقرار باشد:

۱) متغیر وابسته اسمی (دو یا چندوجهی)
۲) متغیرهای مستقل هم ترتیبی و هم فاصله ای
۳) نرمال نبودن توزیع متغیرهای پیش بین.

نکته: زمانی که پیش فرض های نرمال بودن چندمتغیره و برابری ماتریس های واریانس و کوورایانس تامین شدند، و نیز متغیرهای مستقل همه در سطح سنجش فاصله ای/نسبی باشند؛ در آن صورت پیشنهاد می شود که از روش تحلیل تشخیصی به جای روش رگرسیون لجستیک استفاده کنیم.

نکته: دونوع رگرسیون لجستیک وجود دارد: رگرسیون لجستیک اسمی دو وجهی و رگرسیون لجستیک اسمی چندوجهی. یعنی بسته به این که متغیر وابسته دارای چندوجه (طبقه) باشد، از رگرسیون لجستیک مناسب با آن بهره می گیریم.

۲) پیش فرض ها و ملاحظات

۱-سطح سنجش: متغیرهای مسقل می توانند هم در سطح کمی (فاصله ای/نسبی) و هم در سطح کیفی (اسمی/ترتیبی) طبقه بندی شده باشند. اما چنان‌چه یک یا چندمتغیر مستقل در سطح اسمی/ترتیبی بودند، حتماً باید ابتدا این متغیرها را به متغیرهای تصنعی تبدیل کنیم (یعنی کدهای ۰ و ۱). البته در روش رگرسیون لجستیک، کادری به نام ... Categorical وجود داردکه با انتخاب و اجرای آن، متغیرهای ترتیبی به طور خودکار به متغیرهای تصنعی تبدیل می شوند. بنابراین، نیازی به کدگذاری مجدد آن ها نیست.

۲- توزیع نرمال: لزوم تبعیت متغیرهای مستقل از توزیع نرمال ضروری نیست. اما چنان چه این متغیرها دارای توزیع نرمال چندمتغیره باشند، در آن صورت برازش مدل بهتر خواهد بود.

۳- چندهم‌خطی: چندهم‌خطی نبودن متغیرهای مستقل، از دیگر مفروضات رگرسیون لجستیک می باشد. چرا که در صورت چندهم خطی بودن این متغیرها، برآوردها دارای اریب بوده و خطاهای استاندارد نیز نوسان زیادی خواهند داشت.

۴- حجم نمونه: اگرچه در ادبیات مربوط به رگرسیون لجستیک، قواعد خاصی برای حجم نمونه و نیز حداقل تعداد نمونه تعداد متغیر مستقل (پیش بین) پیشنهاد نشده است، اما برخی نویسندگان در حوزه آمار چندمتغیره، حداقل حجم نمونه برای یک تحلیل رگرسیون لجستیک خوب را ۱۰۰ نفر و برخی نیز ۵۰ نفر عنوان کرده اند(حبیب پور و صفری، ۱۳۸۸: ۷۱۶-۷۱۵).

۳) مثال

اطلاعات مربوط به سابقه شغلی و میزان تحصیلات کارمندان اداره پست را جمع آوری کرده ایم (داده ها فرضی است) و قصد داریم این را بیازماییم که آیا می توان با استفاده از سابقه شغلی و تحصیلات کارمندان، پیش بینی کنیم که آنان در چه گروه درآمدی (درآمد بالا یا درآمد پایین) قرار می گیرند.

به بیان دیگر قصد داریم عواملی را که بر میزان درآمد کارمندان موثر هستند را شناسایی کنیم، و با استفاده از یک سری متغیرها (سابقه و تحصیلات) احتمال درآمد بالا یا درآمد پایین داشتن کارمندان را بررسی کنیم.

نحوه سنجش و کدگذاری متغیرهای مستقل و وابسته در جدول ۱ آمده است.

جدول ۱. متغیرهای مستقل و وابسته

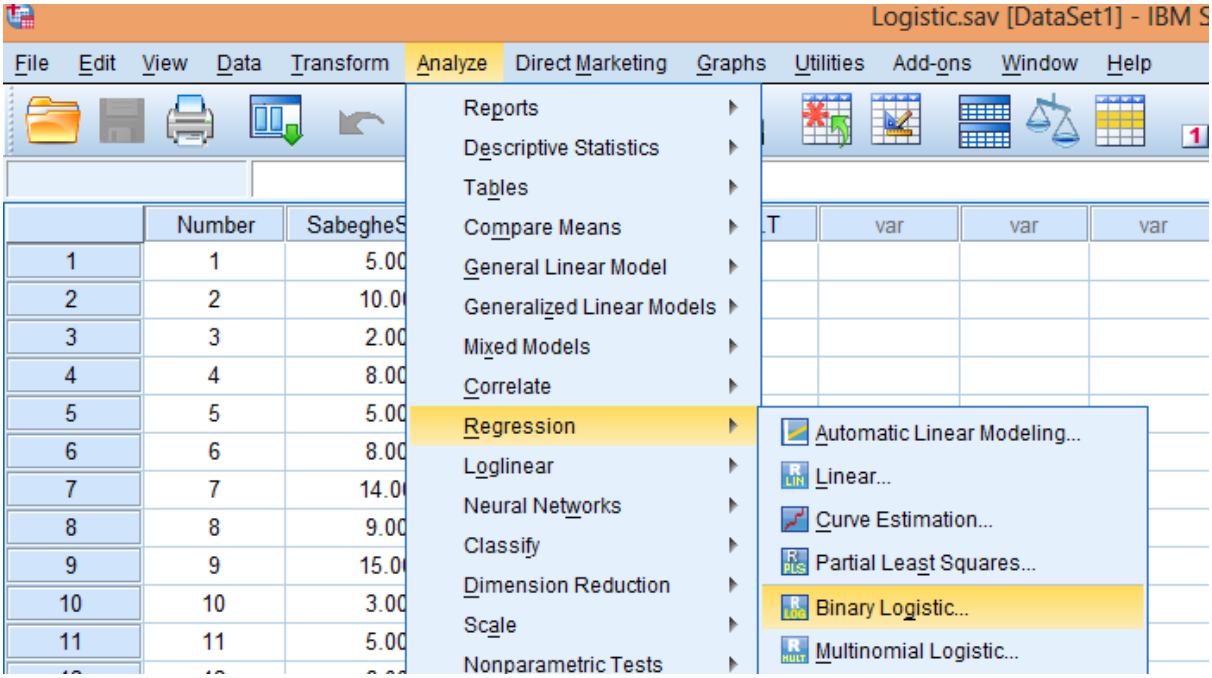
متغیرهای مستقل:	سابقه شغلی: (به سال) میزان تحصیلات : شامل چهار طبقه زیر دیپلم: کد ۱ دیپلم: کد ۲ فوق دیپلم و لیسانس: کد ۳ فوق لیسانس و دکترا: کد ۴
متغیر وابسته:	میزان درآمد: شامل دو طبقه درآمد پایین: کد ۱ درآمد بالا: کد ۲

۴) نحوه اجرای رگرسیون لجستیک در برنامه SPSS

برنامه را باز می کنیم. داده ها در برنامه spss وارد شده اند و کدگذاری متغیرها هم انجام شده است.

نحوه کدگذاری در جدول ۱ نشان داده شده است.

۱- مسیر Analyze ---> Regression ---> Binary Logeistic را دنبال می کنیم.

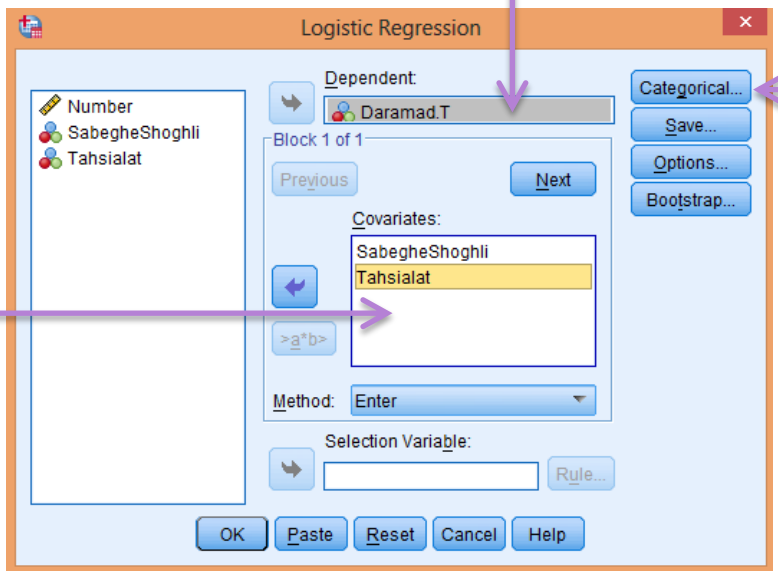


۲- مطابق اشکال زیر دستورات را اجرا می کنیم.

۱- متغیر وابسته را وارد کادر می‌کنیم

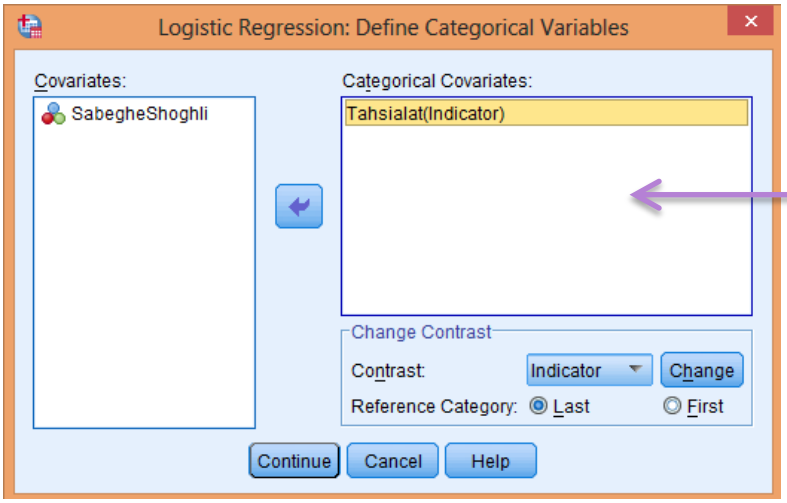
۳- گزینه Categorical را انتخاب می کنیم

۲- متغیرهای مستقل را وارد کادر می کنیم

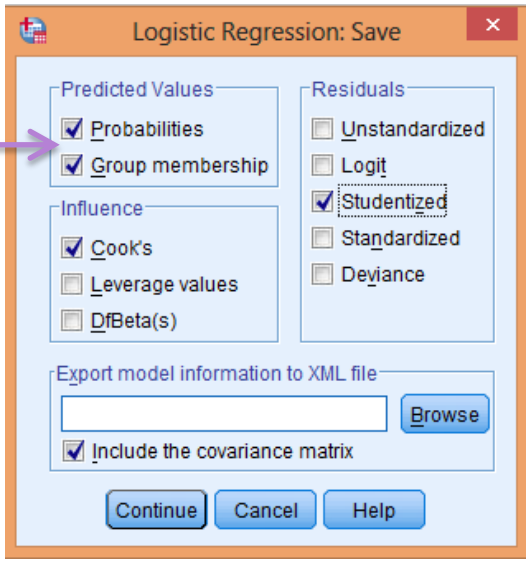


متغیرطبقه‌ای(اسمی‌یاترتیبی)که در اینجا میزان تحصیلات است را وارد کادر می کنیم.

در انتها گزینه Continue را انتخاب می کنیم

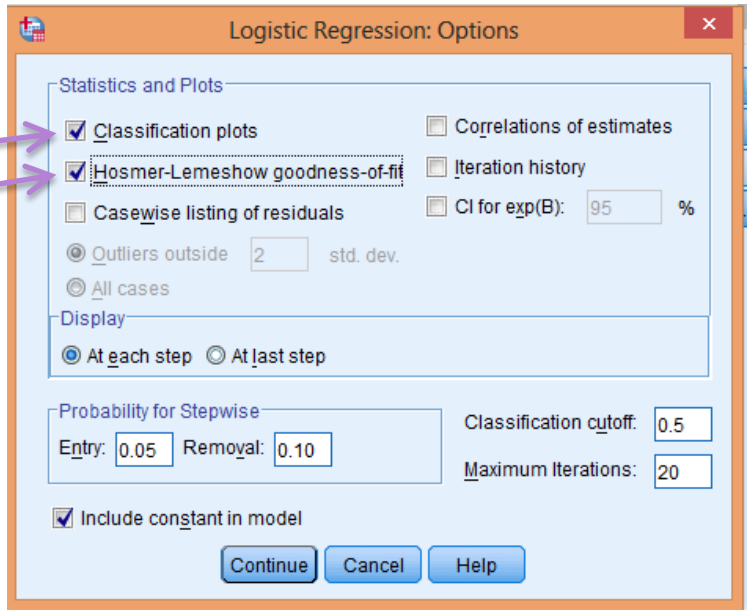


گزینه های زیر را مطابق شکل فعال می‌کنیم:
Probabilities
Group membership
Cook’s
Studentized



گزینه های زیر را مطابق شکل فعال می‌کنیم:
Classification plots
Hosmer-Lemeshow goodness…

درانتها بر روی گزینه Continue و در کادر اصلی بر روی Ok کلیک می کنیم



خروجی های آزمون رگرسیون لجستیک در ادامه آورده شده است. در ادامه به توضیح جداول مهم می پردازیم و در انتها به تفسیر و نحوه گزارش نتایج خواهیم پرداخت.

۵) خروجی ها و توضیحات

جدول زیر نشان می دهد که حجم نمونه معتبر ۱۱۰ نفر است و داده ناقص یا گمشده ای وجود ندارد. به بیان دیگر ۱۰۰ درصدپاسخگویان که برابر با ۱۱۰ نفر می شوند در پردازش شرکت داده شده اند.

Case Processing Summary

	Unweighted Cases[a]	N	Percent
Selected Cases	**Included in Analysis**	**110**	**100.0**
	Missing Cases	0	.0
	Total	110	100.0
	Unselected Cases	0	.0
	Total	110	100.0

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

کدگذاری مجدد متغیر وابسته: برنامه SPSS به کسانی که درآمد پایین دارند و کد ۱ گرفته بودند کد ۰ می دهد و به کسانی که درآمد بالا دارند و کد ۲ گرفته بودند در آزمون رگرسیون لجستیک کد ۱ می دهد و تحلیل ها را بر اساس کدهای صفر و یک انجام می دهد. این طبقه بندی به صورت خودکار انجام می شود.

Dependent Variable Encoding

Original Value	Internal Value
Daramad-Paeen	**0**
Daramad-Bala	**1**

کدگذاری مجدد متغیر مستقل طبقه ای (تحصیلات): آزمون رگرسیون لجستیک هر طبقه را با طبقه بالاتر از آن مقایسه می کند. به بیان دیگر چون میزان تحصیلات طبقه ای بوده و شامل چهار طبقه می شود در نتیجه به آن ۳ کد اختصاص داده است. شیوه کار بدین صورت است که در مقایسه طبقه اول و دوم با هم، برای طبقه اول کد ۱ و برای طبقه دوم کد ۰ را در نظر گرفته است. سپس در مقایسه طبقه دوم و سوم با هم، برای طبقه دوم کد ۱ و برای طبقه سوم کد ۰ را در نظر می گیرد و الی آخر. برای طبقه آخر کدی در نظر گرفته نمی شود.

لازم به ذکر است که ما در فایل داده های SPSS متغیرهای تحصیلات و درآمد را کدگذاری کرده ایم. این کدگذاری باعث می شود که در فایل خروجی رگرسیون لجستیک به جای عدد، نام‌های طبقات درآمد و تحصیلات گزارش شود و فهم جداول آسان تر شود.

فراوانی هر کدام از طبقات تحصیلی نیز ذکر شده است. که برای مثال افرادی که میزان تحصیلاتشان زیر دیپلم است ۱۹ نفر هستند.

Categorical Variables Codings

		Frequency	Parameter coding		
			(1)	(2)	(3)
Tahsialat	ZireDiplom	**19**	1.000	.000	.000
	Diplom	**31**	.000	1.000	.000
	FoghDiplom.Lisans	**46**	.000	.000	1.000
	FoghLisans.Doctora	**14**	.000	.000	.000

نتایج جدول زیر نشان می دهد که با اطمینان ۵۷.۳ درصد با استفاده از مجموع ۲ متغیر مستقل در این تحقیق یعنی تحصیلات و سابقه شغلی، قادریم تغییرات متغیر وابسته میزان درآمد (بالا و پایین) را تبیین کنیم.

Classification Table[a,b]

Observed			Predicted		
			Daramad.T		Percentage Correct
			Daramad-Paeen	Daramad-Bala	
Step 0	Daramad.T	Daramad-Paeen	0	47	.0
		Daramad-Bala	0	63	100.0
	Overall Percentage				**57.3**

a. Constant is included in the model.

b. The cut value is .500

ارزیابی کل مدل

نتایج آزمون اوم نیبوس، ارزیابی کل مدل رگرسیونی لجستیک را نشان می دهد. این آزمون به بررسی این موضوع می پردازد که مدل (نقش تحصیلات و سابقه شغلی در طبقه بندی گروه درآمدی) تا چه اندازه قدرت تبیین و کارایی دارد. با توجه به سطح معنی داری مدل، برازش مدل قابل قبول بوده و در سطح خطای کمتر از ۰۵/. معنی دار است.

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	29.265	4	.000
	Block	29.265	4	.000
	Model	**29.265**	**4**	**.000**

بررسی برازش مدل

ضرایب جدول زیر، تقریب های ضریب تعیین (R^2) در رگرسیون خطی هستند که در این جا در رگرسیون لجستیک استفاده می شوند. در رگرسیون لجستیک، چون محاسبه دقیق مقدار ضریب تعیین دشوار است، 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 بین ۰ تا ۱ نوسان دارد. مقادیر دو آماره برابر با ۲۳۴/. و ۳۱۴/. بدست آمده است و بدین معناست که دو متغیر مستقل توانسته اند بین ۲۳ تا ۳۱ درصد از تغییرات متغیر میزان درآمد پاسخگویان (درآمد بالا یا پایین داشتن) را تبیین کنند.

Model Summary

Step	-2 Log likelihood	**Cox & Snell R Square**	**Nagelkerke R Square**
1	120.892[a]	**.234**	**.314**

a. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

جهت تبیین قدرت مدل در تفکیک افراد در طبقات متغیر وابسته، از نتایج جدول زیر استفاده می کنیم. از نتایج این جدول می توانیم به میزان صحت و درستی در طبقه بندی افراد پی ببریم. دقت کل طبقه بندی افراد برابربا ۷۳.۶ درصد بوده است. این دقت در افراد با درآمد پایین برابر با ۷۰.۲ درصد است که نشان می دهد ۳۳ نفر از کسانی که درآمد پایین داشته اند درست تفکیک شده اند و ۱۴ نفر به اشتباه تفکیک شده اند. همچنین دقت طبقه بندی در کارمندان با درآمد بالا ۷۶.۲ درصد است که در این گروه، ۴۸ نفر از کسانی که درآمد بالا داشته اند به درستی تفکیک شده اند و ۱۵ نفر نیز به اشتباه تفکیک شده اند.

Classification Table[a]

			Predicted		
			Daramad.T		**Percentage Correct**
	Observed		Paeen	Bala	
Step 1	Daramad.T	Paeen	33	14	**70.2**
		Bala	15	48	**76.2**
	Overall Percentage				**73.6**

a. The cut value is .500

جدول بعد مهم ترین جدول در تفسیر نتایج مربوط به معنی داری و میزان تاثیر هر متغیر مستقل بر متغیر وابسته است. در این جدول:

B همان ضریب رگرسیونی استانداردنشده است.

S.E همان خطای استاندارد است.

Wald: آماره والد، مهم ترین آماره برای آزمون معنی داری حضور هر متغیر مستقل در مدل می باشد. آماره والد معادل آماره t در رگرسیون خطی است. مثلا نتایج نشان می دهد که تاثیر متغیر میزان تحصیلات در مجموع معنی دار است اما تاثیر متغیر سابقه شغلی بر میزان درآمد معنی دار نیست (P=./۱۵۸).

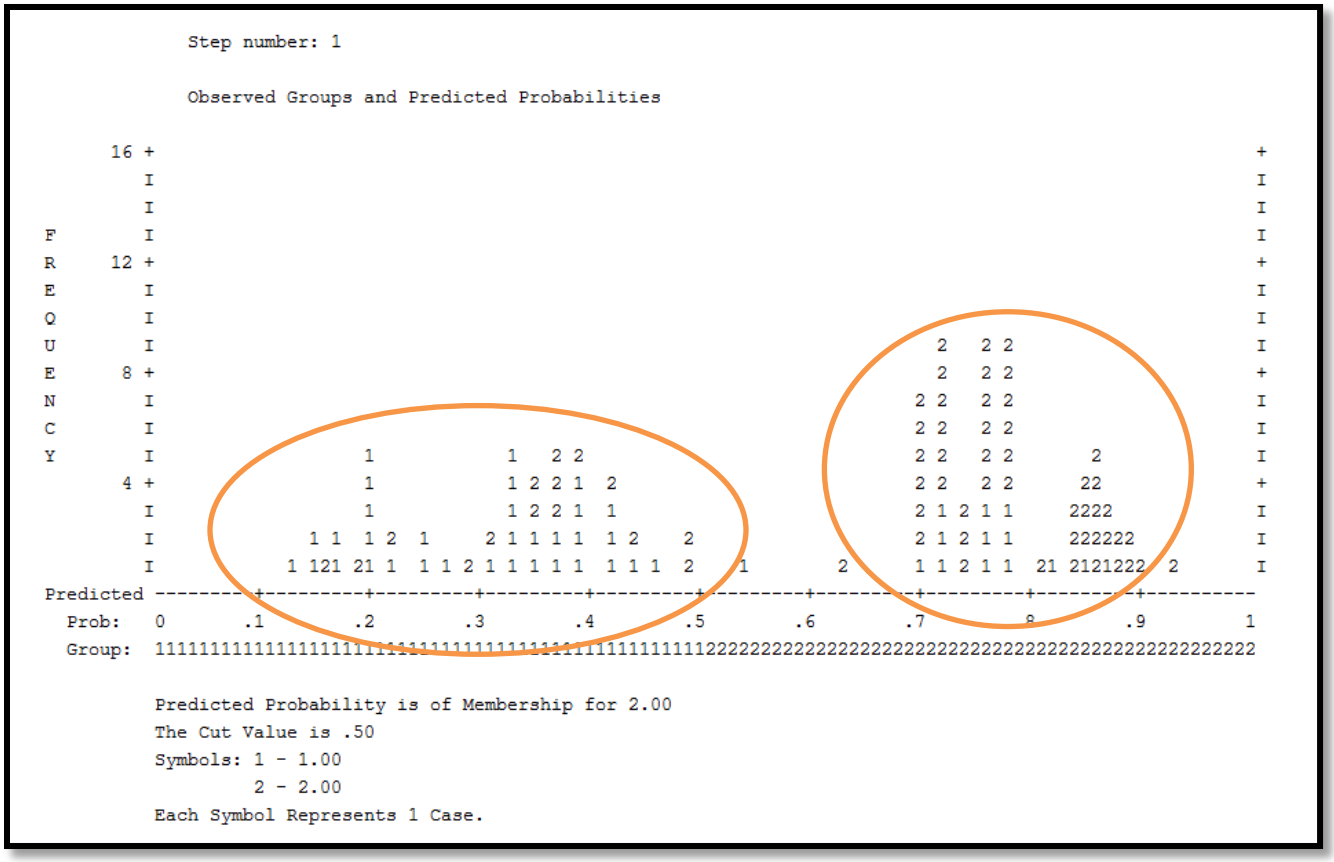
Exp: معادل ضریب رگرسیونی استاندارد شده در رگرسیونی خطی است که برای تفسیر نتایج تحقیق از آن استفاده می شود.

جدول ضرایب

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Tahsialat			**22.819**	3	**.000**	
	Tahsialat(1)	-3.289	.968	**11.536**	1	**.001**	**.037**
	Tahsialat(2)	-2.442	.866	**7.957**	1	**.005**	**.087**
	Tahsialat(3)	-.771	.846	**.830**	1	**.362**	**.463**
	Sabeghe	.097	.069	**1.989**	1	**.158**	**1.102**
	Constant	1.210	.864	**1.962**	1	**.161**	**3.352**

a. Variable(s) entered on step 1: Tahsialat, Sabeghe.



نمودار بالا به نمودار طبقه بندی (Classification Plot) معروف است و تصویر بصری از صحت طبقه بندی را در یک نمودار هیستوگرام پشته ای نشان می دهد. این نمودار در کنار جدول طبقه بندی (ClassificationTable) روش دیگری برای نمایش صحت طبقه بندی پاسخگویان است.

این نمودار از دو محور X و Y تشکیل شده است و میزان انطباق احتمالات پیش بینی شده با پیامدها را نشان می دهد. هر علامت ۱ نشان دهنده پاسخگویی است که درآمد پایین دارد و هر علامت ۲ نشان دهنده پاسخگویی است که درآمد بالا دارد.

هرجه پاسخگویان یک گروه در یک سمت نمودار و پاسخگویان دیگر در سمت دیگر نمودار قرار بگیرند، صحت طبقه بندی و پیش بینی مدل بالاتر است. دقیق ترین شیوه طبقه بندی به صورتی است که همه اعداد ۱ (درآمد پایین) در یک سمت نمودار و همه اعداد ۲ (درآمد بالا) در سمت دیگر نمودار قرار بگیرند.

۶) تفسیر و گزارش نتایج:

آزمون تحلیل رگرسیون لجستیک انجام شد که در آن عامل میزان درآمد به عنوان متغیر وابسته، و میزان تحصیلات و سابقه شغلی به عنوان متغیرهای پیش بین (مستقل) انتخاب شدند. در کل ۱۱۰ نفر در تحلیل وارد شدند و مدل کامل Full Model به طور معنی داری پایا بود (P< ./۰۱ ، df =۴ و ۲۹.۲۶۵=مجذور کای).

این مدل بین ۲۳ تا ۳۲ درصد از واریانس میزان درآمد (بالا یا پایین) را تبیین می کند. ۷۰.۲ درصد از از افراد با درآمد پایین درست طبقه بندی شده اند و ۷۶.۲ درصد از پیش بینی ها در مورد افراد با درآمد بالا صحیح بود. در کل ۷۳.۶ درصد از پیش بینی ها درست بود.

{ جدول شماره... (جدول ضرایب)}، ضرایب و آماره Wald و درجات آزادی مربوط و مقادیر احتمال برای هرکدام از متغیرهای پیش‌بین را ارائه می دهد. تاثیر نتایج نشان می دهد که فقط میزان تحصیلات به طور معنی داری درآمد افراد را پیش بینی می کند. جهت این تاثیر منفی است که نتایج نشان می دهد که با افزایش میزان تحصیلات کارمندان، میزان درآمد آنان کاهش می یابد!!

منابع:

- بریس، نیکلا و کمپ، ریچارد و سلنگار، رزمری (۱۳۹۱) تحلیل داده های روانشناسی با برنامه SPSS ، ویرایش سوم، تهران: نشر دوران.
- حبیب پور؛ کرم و صفری شالی، رضا (۱۳۸۸). راهنمای جامع کاربرد SPSS در تحقیقات پیمایشی، تهران : لویه، متفکران.
- سرمد، زهره و همکاران (۱۳۸۲)، روش های تحقیق در علوم رفتاری، چاپ هفتم، تهران: انتشارات آگاه