آموزش رگرسیون لوجستیک در SAS و Minitab

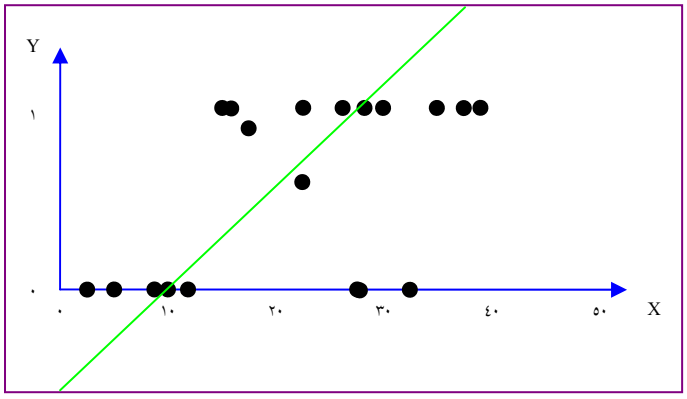
نویسنده

سید جمال میرکمالی

گروه دانش آماری– اسفند ۱۳۸۵

مقدمه

می دانیم رگرسیون به بررسی رابطه بین یک متغیر و سایر متغیر ها می پردازد ، بعبارتی یافتن رابطه $Y = f(\mathbf{X})$ موضوع رگرسیون می باشد. دراینجا Y متغیر پاسخ (یا متغیر وابسته) و $\mathbf{X}$ بردار متغیر های پیشگو (یا مستقل) هستند. گاهی اوقات متغیر پاسخ از نوع طبقه ای می باشد. بعنوان مثال فرض کنید می خواهیم به بررسی تاثیر مقدار مصرف یک دارو در درمان یک بیماری بپردازیم. برای این منظور یک نمونه از بیماران را طی یک دوره مشخص، تحت درمان با داروی فوق الذکر قرار می دهیم. فرض کنیم (x_i, y_i) وضعیت بیمار i ام را نشان دهد بطوریکه $y_i = ۱$ نشاندهنده این باشد که فرد i ام با مصرف مقدار x_i از دارو ، بهبود یافته است و $y_i = ۰$ نشاندهنده این باشد که فرد i ام با مصرف مقدار x_i از دارو ، بهبود نیافته است . در این صورت متغیر Y تنها دو مقدار ۰,۱ می گیرد. اگر نمودار پراکنش داده های مربوط به چنین تحقیقی را رسم کنیم، احتمالا چنین شکلی خواهد داشت:



مشاهده می کنید که خط رگرسیونی که برازش داده شده است برای این داده ها مناسب نمی باشد ، چراکه مقادیری غیر از ۰,۱ اختیار می کند. مثلا این خط برای شخصی که ۴۵ واحد از

دارو مصرف می کند مقدار $y = ۲$ را برآورد می کند که هیچ تعبیری برای آن وجود ندارد. در این موارد از رگرسیون لوژستیک استفاده می شود.

رابطه رگرسیون لوژستیک و رگرسیون ساده

رگرسیون لوژستیک اغلب برای بررسی رابطه بین یک متغیر پاسخ گسسته و سایر متغیر های پیشگو بکار می رود. در مدل های پاسخ دوتایی ، متغیر Y دو مقدار مثلا ۱,۲می گیرد . فرض کنید $\mathbf{x}$ یک بردار از متغیر های پیشگو باشد و $\pi = \Pr(Y = ۱ | \mathbf{x})$ احتمال رخداد $y = ۱$تحت شرایط $\mathbf{x}$ باشد. در این صورت مدل خطی لوجیت به شکل زیر تعریف می شود:

$$\text{logit}(\pi) = \log\left(\frac{\pi}{۱-\pi}\right) = \beta_۰ + \beta_۱ x_۱ + \ldots + \beta_k x_k$$

اکنون مدل فوق یک مدل رگرسیون چندگانه ساده است. پس می توان تکنیک های رگرسیون ساده را برای برآورد $\text{logit}(\pi)$ بکار ببریم. اگر مدل برآورد شده به صورت زیر باشد:

$$\hat{\text{logit}}(\pi) = \hat{\beta}_۰ + \hat{\beta}_۱ x_۱ + \ldots + \hat{\beta}_k x_k$$

در این صورت با محاسبه ساده ای خواهیم داشت:

$$\hat{\pi} = \frac{e^{\hat{\beta}_۰ + \hat{\beta}_۱ x_۱ + \ldots + \hat{\beta}_k x_k}}{۱ + e^{\hat{\beta}_۰ + \hat{\beta}_۱ x_۱ + \ldots + \hat{\beta}_k x_k}}$$

از عبارت فوق مشخص می شود که $\hat{\pi}$ بین صفر و یک است پس شرایط احتمال را دارا است.

با مثال زیر نحوه بکارگیری رگرسیون لوژستیک را در SAS نشان می دهیم.

این داده ها بر گرفته از کتاب The Analysis of Binary Data (کاکس و اسنل) است

. این داده ها در مورد تاثیر زمان حرارت و زمان فروبردن در آب بر عدم آمادگی شمش ها برای نورد می باشد.

Heat	Soak	r	n
۷	۱,۰	۰	۱۰
۱٤	۱,۰	۰	۳۱
۲۷	۱,۰	۱	٥٦
٥۱	۱,۰	۳	۱۳
۷	۱,۷	۰	۱۷
۱٤	۱,۷	۰	٤۳
۲۷	۱,۷	٤	٤٤
٥۱	۱,۷	۰	۱
۷	۲,۲	۰	۷
۱٤	۲,۲	۲	۳۳
۲۷	۲,۲	۰	۲۱
٥۱	۲,۲	۰	۱
۷	۲,۸	۰	۱۲
۱٤	۲,۸	۰	۳۱
۲۷	۲,۸	۱	۲۲
٥۱	٤,۰	۰	۱
۷	٤,۰	۰	۹
۱٤	٤,۰	۰	۱۹
۲۷	٤,۰	۱	۱٦

ثبت آزمایش این طور است : در مرحله اول ۱۰ شمش را به اندازه ۷ واحد زمان حرارت می دهیم ، سپس آنها را به اندازه ۱ واحد زمان در آب فرو می بریم. این ۱۰ شمش را آزمایش می کنیم تا ببینیم چندتا از آنها برای نورد آماده نیستند. مشاهده شده است که همه شمش ها برای نورد آماده هستند. در مرحله بعد ۳۱ شمش را مورد بررسی قرار داده ایم. به همین منوال داده ها ثبت گردیده اند.

اکنون داده ها را وارد کرده و دستور تحلیل رگرسیون لوژستیک را صادر می کنیم:

```
data ingots;
   input Heat Soak r n @@;
   datalines;
    ۷     ۱,۰    ۰     ۱۰
   ۱٤     ۱,۰    ۰     ۳۱
   ۲۷     ۱,۰    ۱     ٥٦
   ٥۱     ۱,۰    ۳     ۱۳
    ۷     ۱,۷    ۰     ۱۷
   ۱٤     ۱,۷    ۰     ٤۳
   ۲۷     ۱,۷    ٤     ٤٤
   ٥۱     ۱,۷    ۰      ۱
    ۷     ۲,۲    ۰      ۷
   ۱٤     ۲,۲    ۲     ۳۳
   ۲۷     ۲,۲    ۰     ۲۱
   ٥۱     ۲,۲    ۰      ۱
    ۷     ۲,۸    ۰     ۱۲
   ۱٤     ۲,۸    ۰     ۳۱
   ۲۷     ۲,۸    ۱     ۲۲
   ٥۱     ٤,۰    ۰      ۱
    ۷     ٤,۰    ۰      ۹
   ۱٤     ٤,۰    ۰     ۱۹
   ۲۷     ٤,۰    ۱     ۱٦
;
proc logistic data=ingots;
   model r/n=Heat Soak;
run;
```

خروجی چنین است:

The SAS System 09:00 Wednesday, February 28, 2007 1

The LOGISTIC Procedure

Model Information

Data Set	WORK.INGOTS
Response Variable (Events)	r
Response Variable (Trials)	n
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	19
Number of Observations Used	19
Sum of Frequencies Read	387
Sum of Frequencies Used	387

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	12
2	Nonevent	375

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	108.988	101.346
SC	112.947	113.221
-2 Log L	106.988	95.346

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.6428	2	0.0030
Score	15.1091	2	0.0005
Wald	13.0315	2	0.0015

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.5592	1.1197	24.6503	<.0001
Heat	1	0.0820	0.0237	11.9454	0.0005
Soak	1	0.0568	0.3312	0.0294	0.8639

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Heat	1.085	1.036	1.137
Soak	1.058	0.553	2.026

Association of Predicted Probabilities and Observed Responses

Percent Concordant	64.4	Somers' D	0.460
Percent Discordant	18.4	Gamma	0.555
Percent Tied	17.2	Tau-a	0.028
Pairs	4500	c	0.730

بیان خروجی

- در قسمت Model Information اطلاعاتی در مورد مدل ارائه شده است.

- در قسمت Response Profile تعداد کل رخداد ها داده شده است. در اینجا جمعا ۱۲ مورد از شمش ها برای نورد آماده نبوده اند.
- در قسمت Testing Global Null Hypothesis: BETA=۰ آزمون معنا داری رگرسیون انجام می شود به این ترتیب که

$$\begin{cases} H_0 : \text{رگرسیون معنا دار نیست} \\ H_1 : \text{رگرسیون معنادار است} \end{cases}$$

در سطح اطمینان ۹۵ درصد ، از آنجا که P-Value این آزمون از ۰,۰۵ کمتر است لذا رگرسیون معنا دار است.

- در قسمت Analysis of Maximum Likelihood Estimates متغیرهای مدل و عرض از مبدا Intercept فهرست شده اند و مقابل هر یک درجه آزادی ، برآورد پارامتر، انحراف معیار ، آماره والد و P-Value نوشته شده است.

این P-Value ها مربوط به آزمون های زیر است:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases} \quad j = 0,1,2$$

در سطح اطمینان ۹۵ درصد نمی توان گفت که $\beta_0 = 0, \beta_1 = 0$ اما می توان گفت که $\beta_2 = 0$ است. بنابراین متغیر زمان فرو بردن در آب تاثیری در آمادگی شمش برای نورد ندارد.

- در قسمت Odds Ratio Estimates برآورد های نقطه ای و فاصله ای برای نسبت بخت ها ارائه شده است. این مقادیر اینطور محاسبه شده اند:

$$e^{\hat{\beta}_1} = e^{0.0820} = 1.085$$
$$e^{\hat{\beta}_2} = e^{0.0568} = 1.058$$

لازم به ذکر است که اگر نسبت بخت ها $\theta = ۱$ باشد به این معنی است که بین متغیر پاسخ و متغیر پیشگو همخوانی وجود ندارد. اگر $\theta > ۱$ در این صورت بخت موفقیت بیشتر از بخت شکست است و اگر $\theta < ۱$ آنگاه بخت شکست بیش از بخت موفقیت است.

- در قسمت Association of Predicted Probabilities and Observed Responses معیار هایی برای ارزیابی قابلیت پیش بینی مدل ارائه می کند. هر چه درصد هماهنگی (Percent Concordant) بیشتر باشد مدل مناسب تر است.

با وجود اینکه این مدل هنوز نیاز به بررسی های بیشتری دارد ، فعلا این مدل را می پذیریم. حال بیایید احتمال این را که در یک آزمایش با $Heat = ۷$، $Soak = ۱$ شمش برای نورد آماده نباشد ، محاسبه کنیم.

$$\widehat{\text{logit}}(\pi) = -۵.۵۵۹۲ + ۰.۰۸۲ \times Heat + ۰.۰۵۶۸ \times Soak = -۴.۹۲۸۴$$

$$\hat{\pi} = \Pr(Y = ۱ | \mathbf{x}) = \frac{e^{-۴.۹۲۸۴۰}}{۱ + e^{-۴.۹۲۸۴۰}} = ۰.۰۰۷۲$$

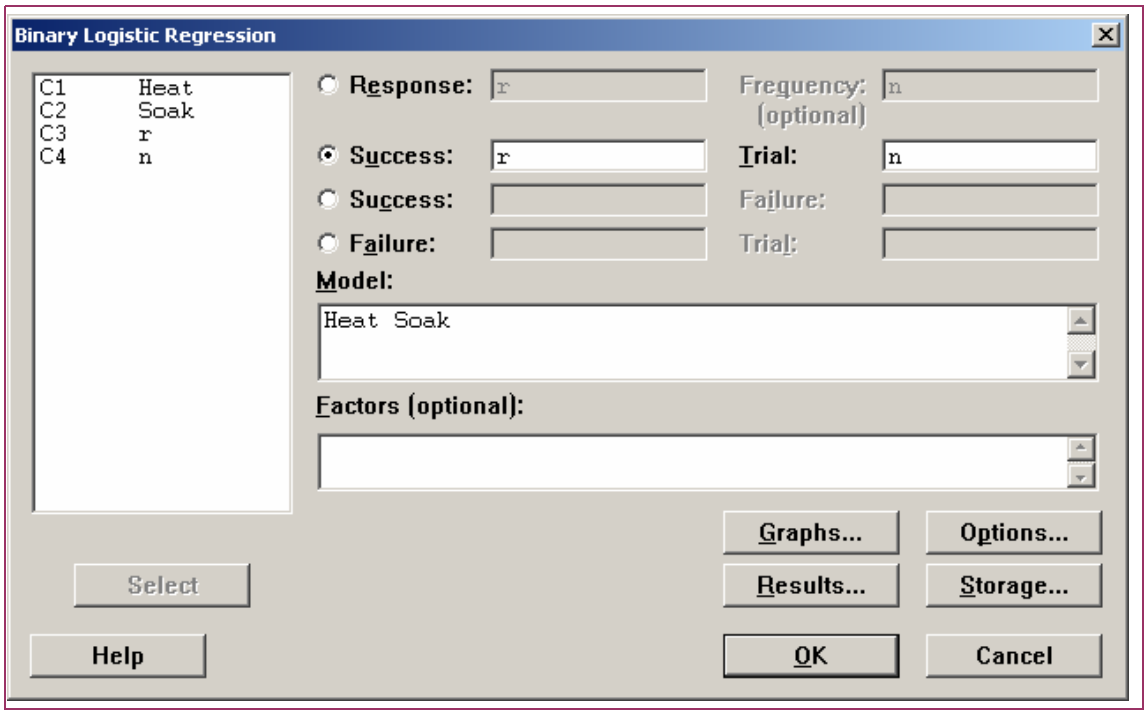
یعنی اگر ۱۰۰۰ شمش را تحت این شرایط قرار دهیم، به طور متوسط ۷ شمش برای نورد آماده نیستند.

بکارگیری Minitab برای تحلیل رگرسیون لوژستیک

همه این محاسبات را می توان با نرم افزار Minitab نیز انجام داد. برای این منظور جدول داده ها را در Minitab کپی کنید. سپس دستورات زیر را وارد کنید:

```
MTB > BLogistic 'r' 'n' = Heat Soak;
SUBC>    ST;
SUBC>    Logit;
SUBC>    Brief ۲.
```

همچنین می توانید از منوی *Stat > Regression > Binary Logistic Regression* استفاده کنید و کادر باز شده را اینطور پر کنید:



Binary Logistic Regression: r; n versus Heat; Soak

```
Link Function: Logit

Response Information

Variable  Value     Count
r         Success      ۱۲
          Failure     ۳۷۰
n         Total       ۳۸۷
```

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-٥,٥٥٩١٧	١,١١٩٦٩	-٤,٩٦	٠,٠٠٠			
Heat	٠,٠٨٢٠٣٠٨	٠,٠٢٣٧٣٤٤	٣,٤٦	٠,٠٠١	١,٠٩	١,٠٤	١,١٤
Soak	٠,٠٥٦٧٧١٣	٠,٣٣١٢١٢	٠,١٧	٠,٨٦٤	١,٠٦	٠,٥٥	٢,٠٣

Log-Likelihood = -٤٧,٦٧٣

Test that all slopes are zero: G = ١١,٦٤٣, DF = ٢, P-Value = ٠,٠٠٣

Goodness-of-Fit Tests

Method	Chi-Square	DF	P
Pearson	١٣,٥٤٣١	١٦	٠,٦٣٣
Deviance	١٣,٧٥٢٦	١٦	٠,٦١٧
Hosmer-Lemeshow	٧,٣٨١٢	٦	٠,٢٨٧

Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

	Group								
Value	١	٢	٣	٤	٥	٦	٧	٨	Total
Success									
Obs	٠	٠	٠	٢	١	٤	١	٤	١٢
Exp	٠,٣	٠,٥	٠,٦	٠,٩	٢,٣	١,٦	١,٧	٤,١	
Failure									
Obs	٤٦	٤٠	٤٣	٦٢	٧٤	٤٠	٤٢	٢٨	٣٧٥
Exp	٤٥,٧	٣٩,٥	٤٢,٤	٦٣,١	٧٢,٧	٤٢,٤	٤١,٣	٢٧,٩	
Total	٤٦	٤٠	٤٣	٦٤	٧٥	٤٤	٤٣	٣٢	٣٨٧

Measures of Association:
(Between the Response Variable and Predicted Probabilities)

Pairs	Number	Percent	Summary Measures	
Concordant	٢٨٩٨	٦٤,٤	Somers' D	٠,٤٧
Discordant	٧٦٨	١٧,١	Goodman-Kruskal Gamma	٠,٥٨
Ties	٨٣٤	١٨,٥	Kendall's Tau-a	٠,٠٣
Total	٤٥٠٠	١٠٠,٠		

می توانید این خروجی را با خروجی SAS مقایسه کرده و نتایج لازم را به دست آورید.

مراجع:

[۱] میرکمالی، س.ج. (۱۳۸۵)، **آشنایی با رگرسیون لوژستیک – کاربردی**، http://mirkamali.persiangig.com/Regression_Logistic.pdf .

[۲] Cox, D.R., and Snell, E.J. (۱۹۸۹), **The Analysis of Binary Data, Second Edition**, Chapman and Hall, London.

[۳] Hosmer, D.W., and Lemeshow, S. (۲۰۰۰), **Applied logistic regression**, John

[٤] SAS Institute Inc. (۲۰۰۳), **SAS Help and Documentation**.