
بسم الله الرحمن الرحيم


# مدرس: سرکار خانم دکتر سیده سولماز طالبی

# رگرسیون لجستیک

جلسه ششم

# مروری بر رگرسیون خطی

- نکته:

این رگرسیون زمانی که متغیر پاسخ کمی باشد مورد استفاده قرار میگیرد.

اگر متغیر پاسخ شما زمان تا رخدادن یک پیامد باشد، اگر چه کمی است نمیتوان از این رگرسیون استفاده کرد.

تفسیر ضرایب:

$$Y= b0 + b1\ x1$$

به ازای یک واحد افزایش x1، y به میزان b1 واحد تغییر میکند.

# مثال ۱:

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .471[a] | .222 | .214 | .425 |

a. Predictors: (Constant), Metastases

b. Dependent Variable: Stage

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 1.774 | .054 | | 32.865 | .000 |
| | Metastases | .463 | .088 | .471 | 5.283 | .000 |

a. Dependent Variable: Stage

میزان همبستگی متغیر متاستاز داشتن با متغیر stage بیماری ۴۷. میباشد. این رگرسیون ۲۲. از کل تغییرات را تبیین میکند.

معادله رگرسیون به صورت زیر است:

Y=1.77 + .46 metastases

# مثال ۱:

- معادله رگرسیون به صورت زیر است:

Y=1.77 + .46 metastases

تفسیر:

متاستاز داشتن یک متغیر اسمی است که کدهای آن: ۰= متاستاز نداشتن و ۱= متاستاز داشتن میباشد. با جایگذاری کد مورد نظر در معادله رگرسیون، مقدار پیش بینی متغیر وابسته بدست می آید.

مثال: stage بیماری برای یک خانم با متاستاز چقدر پیش بینی میشود؟

Y=1.77 + .46 (1)=2.23

Stage بیماری تقریبا ۲ میباشد.

# مثال ۲:

- حال قصد داریم stage بیماری را با توجه به متغیرهای گرید بیماری، متاستاز داشتن و تعداد غدد درگیر پیش بینی کنیم.

**Model Summary[b]**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .483[a] | .233 | .209 | .426 |

a. Predictors: (Constant), tedad ghodade dargir, Metastases, Grade

b. Dependent Variable: Stage

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | 1.618 | .159 | | 10.169 | .000 |
| | Grade | .056 | .076 | .069 | .736 | .464 |
| | Metastases | .466 | .089 | .474 | 5.223 | .000 |
| | tedad ghodade dargir | .007 | .010 | .066 | .702 | .484 |

a. Dependent Variable: Stage

# مثال ۲:

- میزان همبستگی متغیرهای مستقل و وابسته تقریبا .5 است که مقدار مناسبی را نشان میدهد. این خط رگرسیون مقدار .2 از کل تغییرات متغیر وابسته را تبیین میکند که مقدار کمی است. شاید با وارد کردن متغیر های دیگری این مقدار بهتر شود. معادله خط رگرسیون به صورت زیر است:

Y=1.62 + .47 metastases

همانطور که ملاحظه میفرمایید ضرایب این مدل نیز بسیار شبیه به مدل قبلی میباشد.

# تمرین:

هدف پیش بینی هزینه های درمان برای افراد میباشد. به این منظور متغیرهای سن، فشار خون، سیگاری بودن، کلسترول و میزان فعالیت بدنی را وارد رگرسیون کرده و خروجی های زیر را بدست آورده ایم.
با دانستن اینکه

سن متغیر کمی
فشار خون کد ۰= کم فشار، ۱= نرمال و ۲ = پرفشار
سیگار کد ۰= غیر سیگاری و ۱= سیگاری
کلسترول کد ۰= نرمال و ۱= کلسترول بالا
فعالیت بدنی کد ۰= ندارد و ۱=دارد

معادله خط رگرسیون را نوشته و تفسیر کنید.

میزان هزینه های درمان برای یک شخص ۶۰ساله سیگاری با فشار خون نرمال، کلسترول بالا و عدم فعالیت بدنی را محاسبه کنید.

# تمرین:

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .425[a] | .181 | .180 | 15.42655 |

a. Predictors: (Constant), Physically active, Cholesterol, Smoker, Age in years, Blood pressure

**Coefficients[a]**

| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | | |
| 1 | (Constant) | -13.848 | 1.254 | | -11.046 | .000 |
| | Age in years | .468 | .019 | .244 | 25.039 | .000 |
| | Blood pressure | 2.886 | .278 | .102 | 10.374 | .000 |
| | Smoker | 7.218 | .405 | .174 | 17.840 | .000 |
| | Cholesterol | 6.041 | .341 | .173 | 17.739 | .000 |
| | Physically active | -6.615 | .335 | -.194 | -19.762 | .000 |

a. Dependent Variable: Treatment costs

# پاسخ:

میزان همبستگی بین متغیرهای مستقل و وابسته ۴. میباشد و این خط ۱۸ درصد از کل تغییرات را تبیین میکند. معادله رگرسیون به صورت زیر است:

Y= - 13.84+.46  age +2.88 BP +7.22 smoke +6.04 CHOL -6.61 ACTIVE

با افزایش هر سال عمر به شرط ثابت بودن تمام متغیرها، تقریبا ۵. واحد به هزینه های درمان اضافه میشود.
ابتلا به فشار خون به شرط ثابت بودن تمام متغیرها تقریبا ۳ واحد به هزینه های درمان اضافه میکند.
داشتن فعالیت بدنی به شرط ثابت بودن سایر متغیرها، تقریبا ۷ واحد از هزینه های درمان میکاهد.

Y= -13.84 + .46 (60) + 2.88 (1) + 7.22 (1) + 6.04 (1) -6.61 (0)= 29.9

# رگرسیون لجستیک

- برای رگرسیون لجستیک متغیر وابسته حتما باید اسمی دو حالتی یا چند حالتی باشد. متغیرهای مستقل میتوانند هر نوعی باشند.

$$\text{Log } (p/1-p)=\text{logit}(p)= b0 + b1\ x1$$

تفسیر:

یک واحد افزایش x1، نسبت بخت را به میزان exp(b1) تغییر میدهد.

به exp(b1) نسبت بخت و به exp (b0) شیوع اولیه میگویند.

نکته:

در برخی مقالات نسبت بخت را به اشتباه به صورت نسبت خطر تفسیر میکنند. این جابجایی در تفاسیر زمانی که شیوع پیامد کم( زیر ۱۵ درصد ) باشد ایراد ندارد. در غیر این صورت نمیتوان نسبت بخت را به صورت نسبت خطر تفسیر کرد.

# مثال ۱:

- میخواهیم احتمال ابتلا به دیابت را در نظر گرفتن چگونگی فشار خون افراد پیش بین کنیم.
- برای این منظور یک رگرسیون لجستیک برازش میدهیم. خروجی ها به صورت زیر هستند.

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| No | 0 |
| Yes | 1 |

**Categorical Variables Codings**

| | | Frequency | Parameter coding (1) | Parameter coding (2) |
|---|---|---|---|---|
| Blood pressure | Hypotension | 1207 | .000 | .000 |
| | Normal | 6134 | 1.000 | .000 |
| | Hypertension | 2659 | .000 | 1.000 |

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) Lower | 95% C.I.for EXP(B) Upper |
|---|---|---|---|---|---|---|---|---|---|
| Step 1[a] | Blood pressure | | | 317.840 | 2 | .000 | | | |
| | Blood pressure(1) | -.336 | .137 | 6.039 | 1 | .014 | .715 | .547 | .934 |
| | Blood pressure(2) | 1.093 | .133 | 67.907 | 1 | .000 | 2.982 | 2.300 | 3.867 |
| | Constant | -2.758 | .122 | 514.899 | 1 | .000 | .063 | | |

a. Variable(s) entered on step 1: Blood pressure.

# مثال ۱:

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| No | 0 |
| Yes | 1 |

این خروجی کدهای متغیر وابسته را نشان میدهد. با توجه به این جدول ما متوجه میشویم که دیابت داشتن کد ۱ را دارد. در نتیجه مدل ما نسبت بخت ابتلا به دیابت را برآورد میکند.

# مثال ۱:

**Categorical Variables Codings**

| | | Frequency | Parameter coding (1) | Parameter coding (2) |
|---|---|---|---|---|
| Blood pressure | Hypotension | 1207 | .000 | .000 |
| | Normal | 6134 | 1.000 | .000 |
| | Hypertension | 2659 | .000 | 1.000 |

این خروجی کدهای متغیر مستقل را معرفی میکند. در واقع با توجه به این جدول متوجه گروه مرجع میشویم. مثلا در این مثال گروه کم فشار گروه مرجع میباشد و تمامی تفاسیر نسبت به این گروه ارائه میگردد. این متغیر را به عنوان یک متغیر اسمی در نظر گرفته ایم بنابراین برای حالت نرمال و پرفشار برآوردهای مختلفی ارائه میشود. اگر به عنوان متغیر رتبه ای وارد شود یک برآورد برای متغیر فشار خون محاسبه میشد.

# مثال ۱:

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1[a] | Blood pressure | | | 317.840 | 2 | .000 | | | |
| | Blood pressure(1) | -.336 | .137 | 6.039 | 1 | .014 | .715 | .547 | .934 |
| | Blood pressure(2) | 1.093 | .133 | 67.907 | 1 | .000 | 2.982 | 2.300 | 3.867 |
| | Constant | -2.758 | .122 | 514.899 | 1 | .000 | .063 | | |

a. Variable(s) entered on step 1: Blood pressure.

این خروجی ضرایب متغیرها را ارائه میدهد. با توجه به این جدول معادله رگرسیون به صورت زیر میباشد.

Logit (p)= -2.75 - .33 normal pressure + 1.09 hypertension

ستون Exp(B) نسبت بخت را برای هر متغیر ارائه میدهد. اگر مقدار کمتر از ۱ باشد، آن متغیر پیشگیری کننده و اگر بیشتر از ۱ باشد عامل خطر است.

# مثال ۲:

اینک میخواهیم احتمال ابتلا به دیابت را با توجه به متغیرهای رده سنی، جنس، فشار خون، کلسترول و فعالیت بدنی بدست بیاوریم.

متغیر رده سنی: ۱=۴۵-۵۴، ۲=۵۵-۶۴، ۳= ۶۵ – ۷۴ و ۴= بیشتر از ۷۴
متغیر جنس: ۰= مرد و ۱= زن
متغیر فشار خون: ۰=کم فشار، ۱= فشار نرمال و ۲= پر فشار
متغیر کلسترول: ۰= نرمال و ۱= بالا
متغیر فعالیت فیزیکی: ۰= ندارد و ۱= دارد.

در این مدل، سن به عنوان یک متغیر رتبه ایی وارد مدل شده است. وارد کردن متغیرها با توجه به نظر محقق است. در این حالت محقق ادعا میکند که اثر افزایش رده سنی از ۱ به ۲ همانند اثر افزایش رده سنی ۲ به ۳ است. به همین دلیل یک ضریب برآورد میشود و به جای متغیر کدهای مختلف مینشیند. اگر این تساوی اثرات برقرار نباشد ( مانند فشار خون) یعنی اثر افزایش کد از ۰ به ۱ با اثر افزایش کد از ۱ به ۲ یکسان نباشد باید برای هر رده یک ضریب محاسبه گردد.

# مثال ۲:

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| No | 0 |
| Yes | 1 |

**Categorical Variables Codings**

| | | Frequency | Parameter coding (1) | Parameter coding (2) |
|---|---|---|---|---|
| Blood pressure | Hypotension | 1207 | .000 | .000 |
| | Normal | 6134 | 1.000 | .000 |
| | Hypertension | 2659 | .000 | 1.000 |
| Physically active | No | 5051 | .000 | |
| | Yes | 4949 | 1.000 | |
| Cholesterol | Normal | 5934 | .000 | |
| | High | 4066 | 1.000 | |
| Gender | Male | 5029 | .000 | |
| | Female | 4971 | 1.000 | |

جدول اول: کدهای متغیر وابسته را مشخص میکند.

جدول دوم: گروه های مرجع را برای متغیرهای اسمی نشان میدهد.

# مثال ۲:

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | Lower | Upper |
| Step 1[a] | Age category | .121 | .041 | 8.630 | 1 | .003 | 1.129 | 1.041 | 1.224 |
| | Gender(1) | .066 | .077 | .731 | 1 | .393 | 1.068 | .918 | 1.242 |
| | Blood pressure | | | 301.979 | 2 | .000 | | | |
| | Blood pressure(1) | -.318 | .137 | 5.393 | 1 | .020 | .728 | .556 | .952 |
| | Blood pressure(2) | 1.088 | .133 | 66.708 | 1 | .000 | 2.968 | 2.286 | 3.853 |
| | Cholesterol(1) | .103 | .078 | 1.748 | 1 | .186 | 1.108 | .952 | 1.291 |
| | Physically active(1) | -.131 | .078 | 2.801 | 1 | .094 | .877 | .752 | 1.023 |
| | Constant | -3.062 | .172 | 316.479 | 1 | .000 | .047 | | |

a. Variable(s) entered on step 1: Age category, Gender, Blood pressure, Cholesterol, Physically active.

معادله خط با توجه به جدول بالا به صورت زیر است:

Y= .12 age - .3 normal pressure + 1.08 hypertention

اثر سایر متغیرها به شرط حضور رده سنی و فشار خون معنادار نیست.

کم فشاری به شرط ثابت بودن سایر متغیرها نسبت بخت ابتلا به دیابت را تقریبا ۳. برابر کاهش میدهد.

پر فشاری به شرط ثابت بودن سایر متغیرها نسبت بخت ابتلا به دیابت را تقریبا ۳ برابر افزایش میدهد.

# تمرین:

- برای محاسبه  نسبت بخت ابتلا به سکته قلبی، تصمیم به برازش رگرسیون لجستیک در حضور متغیرهای رده سنی، جنسیت، فشار خون، سیگار و کلسترول گرفته ایم. خروجی ها به صورت زیر است. معادله رگرسیون را نوشته، ضرایب را تفسیر کنید

- متغیر رده سنی: ۱=۴۵-۵۴، ۲=۵۵-۶۴، ۳= ۶۵ – ۷۴ و ۴= بیشتر از ۷۴
- متغیر جنس: ۰= مرد و ۱= زن
- متغیر فشار خون: ۰=کم فشار، ۱= فشار نرمال و ۲= پر فشار
- متغیر سیگار: ۰=غیر سیگاری و ۱= سیگاری
- متغیر کلسترول: ۰= نرمال و ۱= بالا

# تمرین:

**Dependent Variable Encoding**

| Original Value | Internal Value |
|---|---|
| No | 0 |
| Yes | 1 |

**Categorical Variables Codings**

| | | Frequency | Parameter coding (1) |
|---|---|---|---|
| Cholesterol | Normal | 5934 | .000 |
| | High | 4066 | 1.000 |
| Gender | Male | 5029 | .000 |
| | Female | 4971 | 1.000 |

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) | 95% C.I.for EXP(B) Lower | Upper |
|---|---|---|---|---|---|---|---|---|---|
| Step 1[a] | Age category | .619 | .025 | 625.042 | 1 | .000 | 1.857 | 1.769 | 1.949 |
| | Gender(1) | .052 | .044 | 1.416 | 1 | .234 | 1.053 | .967 | 1.147 |
| | Blood pressure | .596 | .037 | 260.653 | 1 | .000 | 1.815 | 1.689 | 1.952 |
| | Smoker | 1.235 | .055 | 513.459 | 1 | .000 | 3.439 | 3.090 | 3.826 |
| | Cholesterol(1) | .701 | .045 | 247.468 | 1 | .000 | 2.016 | 1.848 | 2.200 |
| | Constant | -2.716 | .086 | 1006.733 | 1 | .000 | .066 | | |

a. Variable(s) entered on step 1: Age category, Gender, Blood pressure, Smoker, Cholesterol.

# پاسخ:

دراین تمرین متغیر فشار خون به صورت رتبه ایی وارد شده است.

Logit(p)= - 2.7+ .6 age +.6 BP +1.2 smoke +.7 chol

نسبت بخت برای افزایش هر رده سنی برابر ۱.۸۵ است و این یعنی سن یک عامل خطر برای سکته به حساب میاید و افزایش هر رده سنی ( با توجه به رده بندی داده ها) احتمال سکته را ۸۵درصد بالا میبرد.

نسبت بخت برای افزایش هر رده فشار خون برابر ۱.۸۱ است و این یعنی فشار خون یک عامل خطر برای سکته به حساب میاید و افزایش هر رده فشار خون( با توجه به رده بندی داده ها) احتمال سکته را ۸۵درصد بالا میبرد.

نسبت بخت برای سیگاری بودن برابر ۳.۴۳ است و این یعنی سیگار یک عامل خطر برای سکته به حساب میاید و در افراد سیگاری( کد ۱ = سیگاری بودن) احتمال سکته تقریبا ۳.۵ برابر افراد غیر سیگاری است.

نسبت بخت برای کلسترول بالا برابر ۲.۰۶ است و این یعنی کاسترول بالا یک عامل خطر برای سکته به حساب میاید و در افراد با کلسترول بالا ( کد ۲ = کلسترول بالا ) احتمال سکته تقریبا ۲ برابر افراد با کلسترول نرمال است.