آموزش رگرسیون کاربردی در SAS

گروه دانش آماری

http://MiladZaman005.blogfa.com

نویسنده : سید جمال میرکمالی

اسفند ۸۵

رگرسیون در واقع یافتن رابطه بین یک متغیر و سایر متغیر ها می باشد 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. ساده ترین نوع رگرسیون ، رگرسیون خطی ساده می باشد که به بررسی مدل $Y = \beta_0 + \beta_1 X$ می پردازد. بدین منظور یک نمونه تصادفی زوج های $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ تهیه می شود و بر اساس این مشاهدات بهترین خط $Y = \beta_0 + \beta_1 X$ را که میزان خطای آن کم باشد ، ارائه می شود. به مثال زیر توجه کنید.

فرض کنید بخواهیم وزن بچه ها را بر اساس طول قد آنها پیش بینی کنیم. برای این منظور وزن و قد ۱۹ بچه مدرسه ای را اندازه گیری می کنیم و β_0 و β_1 را در مدل $Weight = \beta_0 + \beta_1 Height + \varepsilon$ برآورد می کنیم.

توجه کنید که در این مدل $Weight$ متغیر پاسخ مدل، $Height$ متغیر پیشگوی مدل، β_0 و β_1 پارامتر های مجهول مدل و ε مولفه خطای مدل است.

اکنون داده ها را در SAS وارد می کنیم

```
data class;
   input Name $ Height Weight Age;
   datalines;
   Alfred  69.0 112.5 14
   Alice   56.5  84.0 13
   Barbara 65.3  98.0 13
   Carol   62.8 102.5 14
   Henry   63.5 102.5 14
   James   57.3  83.0 12
   Jane    59.8  84.5 12
   Janet   62.5 112.5 15
   Jeffrey 62.5  84.0 13
   John    59.0  99.5 12
   Joyce   51.3  50.5 11
   Judy    64.3  90.0 14
   Louise  56.3  77.0 12
   Mary    66.5 112.0 15
   Philip  72.0 150.0 16
   Robert  64.8 128.0 12
   Ronald  67.0 133.0 15
   Thomas  57.5  85.0 11
   William 66.5 112.0 15
     ;
```

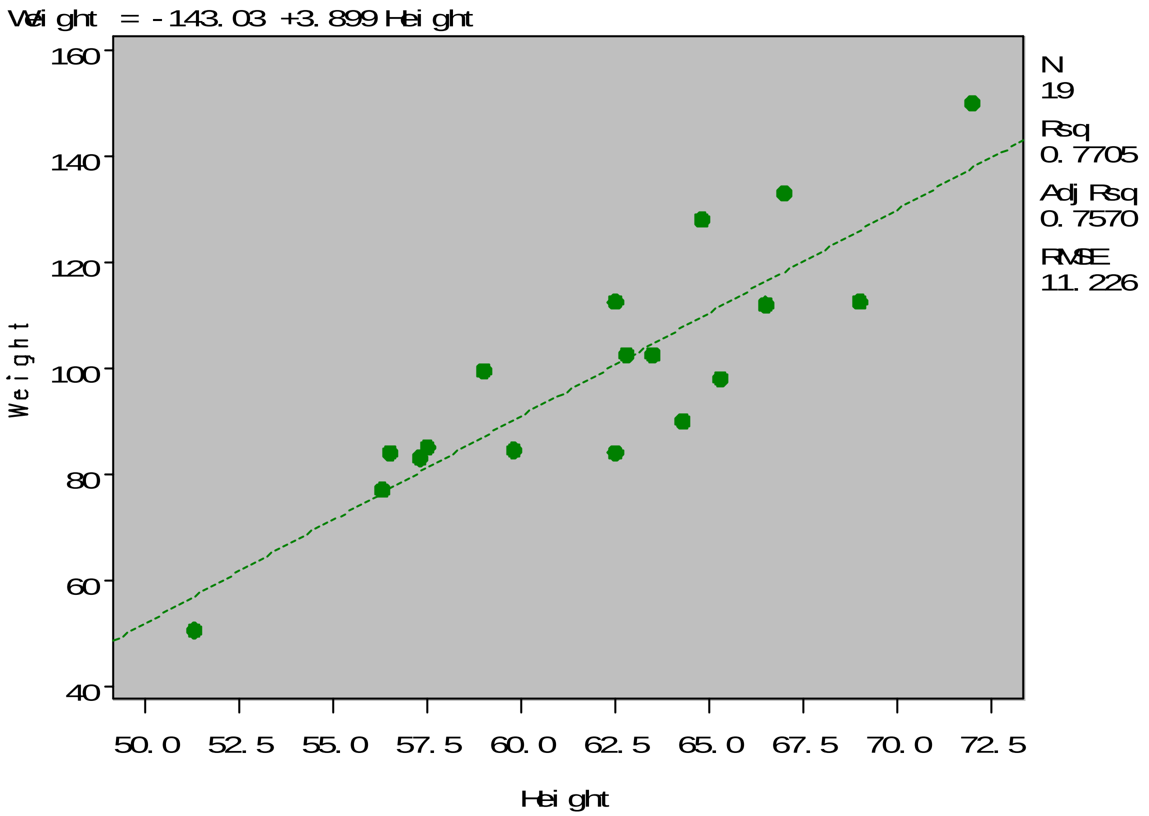
اکنون بر داده های فوق، تحلیل رگرسیونی را اعمال کرده و نمودار خط برازش داده شده را نمایش می دهیم:

```
   symbol1 v=dot c=Green height=3pct;
   proc reg;
      model Weight=Height;
      plot Weight*Height/cframe=ligr;
   run;
```

حال دکمه را می زنیم تا خروجی های زیر حاصل شود:

```
                                  The SAS System     09:00 Tuesday, February 27, 2007   1

                                 The REG Procedure
                                   Model: MODEL1
                             Dependent Variable: Weight

                        Number of Observations Read          19
                        Number of Observations Used          19

                                Analysis of Variance

                                          Sum of           Mean
     Source                   DF         Squares         Square    F Value    Pr > F

     Model                     1      7193.24912     7193.24912      57.08    <.0001
     Error                    17      2142.48772      126.02869
     Corrected Total          18      9335.73684


                  Root MSE             11.22625    R-Square     0.7705
                  Dependent Mean      100.02632    Adj R-Sq     0.7570
                  Coeff Var            11.22330


                                Parameter Estimates

                                Parameter       Standard
          Variable     DF        Estimate          Error    t Value    Pr > |t|

          Intercept     1      -143.02692       32.27459      -4.43      0.0004
          Height        1         3.89903        0.51609       7.55      <.0001
```

http://MiladZaman005.Blogfa.com



همانطور که مشاهده می کنید برآورد β_0 و β_1 با $b_0 = -۱۴۳.۰$ و $b_1 = ۳.۹$ ارائه شده است. بنابراین رابطه بین طول قد و وزن به صورت $Weight = -۱۴۳.۰ + ۳.۹ \times Height$ داده می شود.

رگرسیون اغلب برای اکتشاف روابط تجربی بکار می رود ؛ مانند رابطه بین وزن و قد در این مثال. به این نکته توجه کنید که در این مثال ، طول قد علت اندازه وزن نمی باشد. در مثال فوق برای برآورد β_0 و β_1 روشی موسوم به کمترین مربعات خطا (LSE) بکار رفته است. در این روش β_0 و β_1 طوری برآورد می گردند که مربع فاصله نقاط از خط برازش داده شده مینیمم گردد.

نکته های تفسیر تحلیل رگرسیونی

در اکثر کاربرد های رگرسیون، مدل های رگرسیونی صرفا تقریب های مفیدی ارائه می دهند. واقعیت آنست که اغلب ، گفتن اینکه کدام مدل درست است خیلی پیچیده است. ممکن است بخواهید فرا تر رفته و مدلی با متغیر هایی انتخاب کنید که قابل اندازه گیری و برآورد نباشند ؛ مثلا بخواهید بدانید جهان چگونه کار می کند. در هر صورت ، حتی در مواردی که تئوری ها با کمبود مواجه هستند ، مدلهای رگرسیونی که با دقت ، برای نمونه های بزرگ فرموله شده اند ، می توانند پیش بینی های خوبی ارائه کنند.

آماردانان معمولا کلمه "پیش بینی" را در مفاهیم تخصصی به کار می برند. پیش بینی در این جا به معنی "پیشگویی آینده" یا "غیب گویی" نیست بلکه به معنی حدس زدن پاسخ با توجه به مقادیر مشاهده شده ی متغیر های پیشگو می باشد.

بیان خروجی

- در ابتدای خروجی ، مدل و متغیر پیشگو($Weight$) و تعداد مشاهدات مشخص شده است.
- در جدول جدول آنالیز واریانس منبع تغییرات به صورت $Model$ ، $Error$ ، $Corrected\ Total$ فهرست شده اند. مقابل هر یک از آن ها درجه آزادی و مجموع مربعات مربوطه نوشته شده است. آماره F که در ستون پنجم این جدول داده شده است از تقسیم میانگین مربعات مدل(MSR) بر میانگین مربعات خطا (MSE) حاصل می شود. در ستون ششم جدول $P\text{-}Value$ آزمون معناداری رگرسیون آورده شده است:

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

در سطح اطمینان ۹۵ درصد ، چون $P\text{-}Value$ از ۰,۰۵ کمتر است بنابراین این فرض رد می شود، یعنی رگرسیون معنا دار است بعبارتی این مدل درصد معنا داری از تغییرات داده ها را توجیه می کند. $Root\ MSE$ برآورد انحراف معیار مولفه خطا می باشد. $Dependent\ Mean$ میانگین داده های متغیر پاسخ ($Weight$) است. $Coeff\ Var$ ضریب تغییرات داده ها است که میزان تغییرات داده ها را بدون واحد اندازه گیری بیان می کند. $R\text{-}Square$ و $Adj\ R\text{-}Sq$ به ترتیب ضریب تعیین R^2 و ضریب تعیین اصلاح شده هستند که برای ارزیابی مدل بکار می روند. در این جا $R^2 = ۰.۷۷۰۵$ یعنی ۷۷ درصد تغییرات داده ها توسط متغیر $Height$ توجیه می شود.
- در جدول برآورد پارامتر ، متغیر ها به همراه عرض از مبدا $Intercept$ فهرست شده اند. مقابل هر یک ، درجه آزادی ، برآورد پارامتر ، خطای معیار ، آماره t و $P\text{-}Value$ نوشته شده است. در این جا ضریب $Height$ برابر با **۳.۹** است و همانطور که قبلا گفته شد، مدل برآورد شده به صورت $Weight = -۱۴۳.۰ + ۳.۹ \times Height$ می باشد.

$P\text{-}Value$ های ارائه شده در جدول مربوط به آزمون های زیر است:

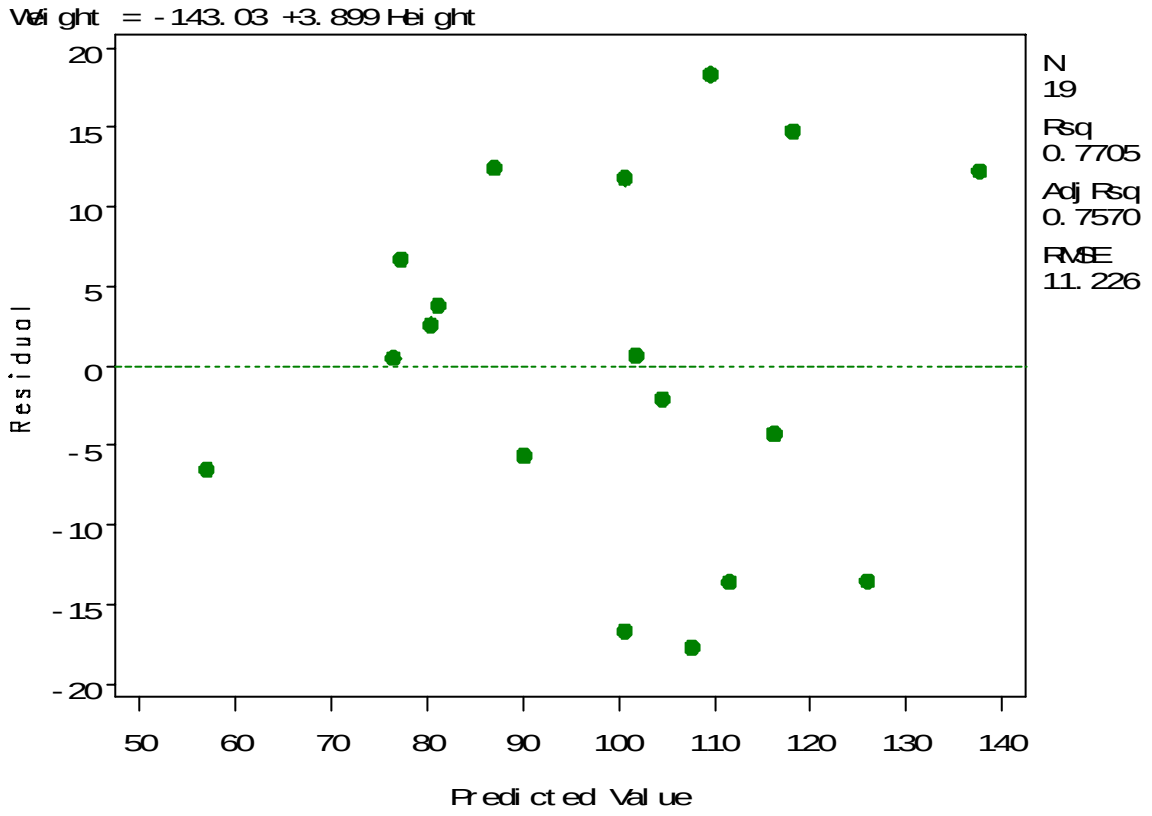
$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases} \quad j = 0,1$$

در سطح اطمینان ۹۵ درصد، با توجه به اینکه مقادیر *P-Value* کمتر از ۰,۰۵ هستند لذا هر دو آزمون رد می شوند، یعنی هر دو ضریب β_0 و β_1 مخالف صفر هستند.

پس از برآورد مدل بایستی معیار های مناسبت مدل بررسی گردد. بدین منظور باید تحلیل باقیمانده را انجام دهیم. اکنون عبارت زیر را در *SAS* اجرا می کنیم:

```
plot r.*p.;
run;
```

http://**MiladZaman005**.Blogfa.com



وجود یک روند در نمودار باقیمانده ها در مقابل مقادیر پیش بینی نشان دهنده عدم ثبات واریانس داده ها می باشد. نمودار فوق یک روند ضعیف در باقیمانده ها نشان می دهد ؛ به نظر می آید هرچه مقادیر پیش بینی افزایش می یابد، باقیمانده ها افزایش می یابد. اما از آنجا که این روند ناچیز است، همین مدل را می پذیریم. توجه کنید که اگر این نمودار شکل قیف داشته باشد ، تبدیلاتی برای ثابت کردن واریانس لازم است. اگر شکلی شبیه دایره داشته باشد احتمالا باید از مدل درجه دو استفاده کنیم.

منبع : *SAS Help and Documentation*

برای اطلاعات بیشتر در خصوص جزییات رگرسیون در *SAS* به آدرس *http://Statisics.Mihanblog.com* مراجعه کنید.

برای مشاهده آموزش رگرسیون خطی به آدرس *http://mirkamali.persiangig.com/Regression_MiladZaman005.pdf* رجوع کنید.