فصل ششم: عیب شناسی در رگرسیون چندگانه و انجام تبدیلات بر روی آن

در این فصل روشهایی را برای آزمون فرضیات اساسی در مدل رگرسیون چندگانه و نهایتاً عیب شناسی در این مدل ارائه خواهیم نمود. همچنین در صورت امکان روشهایی را برای غلبه بر این مشکلات (نحوه از بین بردن واریانس خطاها و همبستگی بین متغیرها) پیشنهاد خواهیم نمود.

۶.۱ عیب شناسی در رگرسیون چندگانه

در هنگام برازش یک مدل رگرسیون چندگانه برای بررسی برقراری فرضیات اساسی توجه به نکات زیر ضروری است:

۱- مشخص کنیم که آیا مدل برازش شده به داده‌ها معنیدار است یا خیر (برازش کافی به داده‌ها دارد یا خیر) ابزار اصلی جهت رسیدن به این هدف و بررسی برقراری فرضیات رسم نمودار مانده‌های استاندارد و مقادیر برازش شده است. همچنین رسم نمودارهای پراکنش حاشیه‌ای نیز در انجام این امر بسیار مهم هستند.

۲- مشخص نمودن اینکه کدام یک از نقاط دارای تأثیر زیادی بر برازش مدل نهایت آمده دارند. (این کار می‌تواند به کمک متلاً لوریج ها انجام گردد).

۳- مشخص نمودن داده‌های پرت در صورت وجود. یعنی داده‌هایی که از خط کلی مدل برازش شده و سایر داده‌ها تبعیت نمی‌کنند.

۴- ارزیابی تأثیر هرکدام از متغیرهای پیشگو به تنهایی بر روی متغیر پاسخ یا استفاده از نمودارهای متغیر اضافه شده (قسمت ۶.۲.۳).

۵- ارزیابی مقدار همبستگی بین متغیرهای پیشگو، استفاده از معیارهای معرفی شده بعنوان مثال عوامل تورم واریانس.

۶- آزمون اینکه آیا فرضیه ثابت بودن واریانس خطاها برقرار است یا خیر. اگر خیر چگونه می‌توان بر این مشکل غلبه نمود.

۷- اگر داده های زمانی داریم یعنی داده ها در طول زمان جمع آوری شده اند بررسی اینکه آیا این داده ها بررسی زمان همبستگی دارند یا خیر.

یادآوری رگرسیون چندگانه در فرم ماتریسی: فرض کنید $Y_{n\times 1}$ بردار متغیر پاسخ و $X_{n\times(p+1)}$ ماتریس طرح باشد. یعنی

$$Y=\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \quad , \quad X=\begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}$$

براساس ماتریس کای فوق، بردار پارامترها و خطا بصورت زیر قابل تعریفند:

$$\beta=(\beta_0\ \beta_1\ \cdots\ \beta_p)' \quad , \quad \varepsilon=(\varepsilon_1\ \varepsilon_2\ \cdots\ \varepsilon_n)'$$

بنابراین می توان نوشت:

$$Y=X\beta+\varepsilon \quad , \quad \varepsilon\sim(0,\sigma^2 I)$$

$$\hat{\beta}=(X'X)^{-1}X'Y \Rightarrow \hat{Y}=X\hat{\beta}=\underbrace{X(X'X)^{-1}X'}_{H}Y = HY \quad , H=X(X'X)^{-1}X'$$

$$e=Y-\hat{Y}=Y-HY=(I-H)Y$$

۱.۱.۶ لوریج ها در رگرسیون چندگانه: همانطور که دیدیم از مدل رگرسیون خطی ساده، لوریج ها معیاری برای نشان دادن تأثیر مشاهده ام بر روی مدل برازش شده بودند. در مدل رگرسیون چندگانه نیز این مقادیر همین نقش را داشته و بصورت زیر تعریف می شوند:

$$\hat{Y}=HY \quad , \hat{Y}_i=\left(\begin{matrix}\text{سطر } i \text{ ام ماتریس}\\ H\end{matrix}\right)\times Y=\sum_{j=1}^{n} h_{ij}Y_j = h_{ii}Y_i+\sum_{\substack{j=1\\ j\neq i}}^{n} h_{ij}Y_j$$

بنابراین می توان نوشت:

$$h_{ii}=x_i(X'X)^{-1}x_i' \quad , h_{ij}=x_i(X'X)^{-1}x_j'$$

h_{ii} تاثیر مشاهده ام بر مشاهده ام

$h_{ii}>2\frac{p+1}{n}$ → تاثیر گذار

بطوریکه x_i سطر i ام ماتریس طرح یعنی $(x_{i1}\ x_{i2}\ \cdots\ x_{ip})$ و x_i' سطر i ام ماتریس $(x_{i1}\ x_{i2}\ \cdots\ x_{ip})$ است.

۲.۱.۶ بررسی خواص مانده ها در رگرسیون چندگانه: بردار مانده ها در مدل رگرسیون چندگانه بصورت زیر تعریف می شود:

۱۸

$$e = Y - \hat{Y} = (I-H)Y \Rightarrow E(e) = E(Y-\hat{Y}) = E[(I-H)Y]$$

$$\Rightarrow E(e) = (I-H)E(Y) = (I-H)X\beta = X\beta - HX\beta = 0$$

$$Var(e) = Var[(I-H)Y] = \sigma^2(I-H)I(I-H)' = \sigma^2(I-H)$$

زیرا H یک ماتریس متقارن و $(I-H)$ خودتوان است.

$H = H'$

$$(I-H)(I-H^2) = I - 2H + H^2 = I - 2H + H = I - H$$

همچنین: $cov(e,\hat{Y}) = 0$ ؛ $cov(e,Y) = \sigma^2(I-H)$

بنابراین می‌توان فاندهای استاندارد را بصورت زیر تعریف نمود:

$$Var(e_i) = \sigma^2(1-h_{ii}) \Rightarrow r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} \quad , \quad S = \sqrt{MSE} = \sqrt{\frac{SSE}{n-(p+1)}}$$

چنانچه $|r_i| \geq 2$ آنگاه می‌توان گفت که مشاهده iام یک داده پرت است و در غیر اینصورت پرت نیست. باید توجه داشت که در صورتی یک مشاهده پرت خواهد بود که مدل برازش شده معتبر باشد و در غیر اینصورت نباید برچسب داده پرت بر آن زد. همچنین ممکن است برای یک مجموعه (بویژه وقتی n کوچک است) از داده‌ها $r_i = -1.8$ شود که در مقایسه با سایر داده‌ها آن داده یک داده پرت به نظر آید، بنابراین قبل از حذف یک داده بعنوان داده پرت، بهترین کار رسم نمودار پراکنش داده‌ها و بررسی چند معیار بطور همزمان برای پرت بودن داده‌هاست که در ادامه معرفی خواهند شد.

استفاده از باقیمانده‌ها و باقیمانده‌های استاندارد برای بررسی مدل: به زبان ساده، مدل رگرسیونی چندگانه یک مدل معتبر برای داده‌هاست هرگاه میانگین شرطی Y، شرط X یک تابع خطی از X باشد و واریانس شرطی $Var(Y|X=x) = \sigma^2$ ثابت باشد.

$$E(Y|X=x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

بنابراین یک مدل در صورتی معتبر است که نمودار باقیمانده‌های استاندارد در مقابل هریک از متغیرهای پیشگو یا هر ترکیب خطی از آنها دارای دو خاصیت زیر باشد:

۱- چون در صورت برازش یک مدل صحیح $E(e_i) = 0$ است، لذا باید نمودار پراکنش فوق یک طرح

تصادفی حول محور x ها را نشان دهد.

۲- یک تغییر پذیری ثابت حول محور x ها وجود داشته باشد. (چون در این مدل عرض ثابت است)

بنابراین می‌توان گفت که اگر در نمودارهای پراکنش متغیر هرگونه طرحی وجود داشته باشد، کول کول غیر معتبر بودن مدل برازش شده را نتیجه گرفت.

آشکارات

~~در مدل رگرسیون چندگانه اگر هر دو شرط زیر برقرار نباشد، نمودار مانده‌ها حاوی اطلاعات مستقیم است~~ ~~که برای این سوال که کجا مدل برازش شده~~

نمودارهای مربوط به متغیر افزوده: به کمک این نمودارها بطور شهودی می‌توان به تاثیر هریک از متغیرهای پیشگو در توجیه متغیر پاسخ پی برد.

ابتدا مدل $Y = X\beta + \varepsilon$ ، $Var(\varepsilon) = \sigma^2 I$ (*) در نظر بگیرید. فرض کنید می‌خواهیم وقتی که متغیر پیشگوی جدید Z به مدل فوق اضافه می‌شود را بررسی کنیم، یعنی مدل زیر را در نظر بگیریم:

$Z_{n\times 1}$ ، $(p+1)\times 1$ ، $n\times(p+1)$

$$Y = X\beta + Z\alpha + \varepsilon \quad (**) \qquad Z = (z_1, z_2, \dots, z_n)' \ , \ \alpha \in \mathbb{R}$$

بنابراین ما ابتدا تاثیر جزئی متغیر جدید Z در Y را بدون در نظر گرفتن تاثیر سایر p متغیر اصلی X بر Y بررسی می‌کنیم. برای رسیدن به این هدف کافی است مانده‌های حاصل از رگرسیون $Y = X\beta + \varepsilon$ را بر روی مانده‌های حاصل از مدل $Z = X\delta + \varepsilon$ رگرسیون کنیم.

به عبارت دیگر باید $e_{YX} = Y - \hat{Y} = (I - H_X)Y$ را بعنوان متغیر وابسته جدید بر روی متغیر مستقل جدید $e_{ZX} = Z - \hat{Z} = (I - H_X)Z$ بعنوان رگرسیون نمائیم بطوریکه

$H_X = X(X'X)^{-1}X'$

بنابراین می‌توان گفت که نمودار متغیر افزوده در واقع تاثیر متغیرات Y را که توسط X قابل توجیه نبوده و پس از حذف اثرات آن توسط Z توجیه پذیر است.

می‌توان نوشت:

$$e_{yx} = Y - \hat{Y} = (I - H_x)\, Y$$

$$e_{zx} = Z - \hat{Z} = (I - H_x)\, Z$$

$$Y = X\beta + Z\alpha + \varepsilon \xrightarrow{\times (I-H_x)} (I-H_x)Y = (I-H_x)X\beta + (I-H_x)Z\alpha + (I-H_x)\varepsilon$$

$$\Rightarrow e_{yx} = (X - X(X'X)^{-1}X'X)\beta + e_{zx}\alpha + \varepsilon^* \quad (\varepsilon^* = (I-H_x)\varepsilon)$$

$$\Rightarrow e_{yx} = 0 + e_{zx}\alpha + \varepsilon^* \quad \text{(مدل نمودار متغیر افزوده)}$$

$$\Rightarrow e_{yx} = e_{zx}\alpha + \varepsilon^*$$

اگر $\hat{\alpha}_{AVP}$ مقدار برآورد α در مدل فوق و $\hat{\alpha}_{LS}$ برآورد α در مدل $Y = X\beta + Z\alpha + \varepsilon$ باشد، می‌توان ثابت نمود که $\hat{\alpha}_{LS} = \hat{\alpha}_{AVP}$. این حالت سر متغیره تعدیل تعریف می‌شود.

ثابت شود یعنی مدل $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ را در نظر گرفته و با اضافه نمودن متغیر Z به مدل مقدار α را از دو روش برآورد و درستی قضیه فوق را تصدیق نمایید.

$$Y = \beta_0 + \beta_1 X + \alpha Z + \varepsilon$$

حال اگر مدل (**) صحیح باشد آنگاه نمودار متغیر افزوده باید تقاطع را تولید نماید که بطور تصادفی حول خط با شیب $\hat{\alpha}_{LS}$ که از مبدأ مختصات می‌گذرد پراکنده می‌شوند.

مثال: بررسی رابطه در متوسط قیمت در یک رستوران جهت در شهرک

در این مطالعه متغیرهای بصورت زیر تعریف می‌شوند:

Y: قیمت یک شام برحسب دلار

X_1: غذا

X_2: اکوراسیون

X_3: سرویس

X_4: متغیر توضیحی ۱ برای شرق و ۰ برای غرب

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

پراکنش Y در مقابل

در شکل 6.9 نمودار ~~[illegible]~~ هرکدام از متغیرهای مستقل رسم شده. در این شکل تأثیر هرکدام از متغیرهای مستقل روی Y نمایش داده شده است. برای بررسی اثر هرکدام از متغیرهای مستقل به تنهایی بر روی متغیر Y، نمودارهای افزوده برای هرکدام از متغیرهای پیشگو در شکل 6.1 رسم شده اند. همانطور که دیده می شود بغیر از متغیر مستقل سردوس سایر متغیرها دارای شیب خطی یکسان (مخالف صفر) هستند. بنابراین می توان گفت که متغیر سردوس دارای تأثیر اندکی در متغیر وابسته است.

2.6 تبدیلات

در این بخش تبدیلات را بررسی خواهیم نمود که به منظور زیر بر روی متغیر(ها) اعمال می گردند:

1- غلبه بر شکل غیر خطی بودن

2- غلبه بر شکل ثابت نبودن واریانس

1.2.6 استفاده از تبدیلات برای غلبه بر غیرخطی بودن

مانند فصل سوم، در این بخش نیز دو روش کلی برای یافتن تبدیل مناسب معرفی خواهند شد که عبارتند از:

1- نمودار پاسخ معکوس (Inverse Response Plot)

2- روش باکس-کاکس

در صورت موفقیت می توانیم از تبدیلات استفاده نماییم که عبارتند از:

(a) اعمال تبدیل تنها بر روی متغیر پاسخ

(b) اعمال تبدیل بر روی متغیر یا متغیرهای مستقل

(c) اعمال تبدیل بر روی متغیرهای پاسخ و مستقل.

استفاده از تبدیل به روش رگرسیون معکوس تنها روی متغیر پاسخ:

فرض کنید مدل صحیح رگرسیون به شکل زیر باشد: $Y = g(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)$

که در آن $g(.)$ یک تابع یکنوا معلوم است. مدل فوق با استفاده از تابع معکوس $g(.)$ می‌تواند به راحتی به مدل رگرسیون خطی زیر تبدیل گردد:

$$g^{-1}(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

بعنوان مثال: $Y = \exp\{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\} \Rightarrow \log(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$

مثال: در این مثال می‌خواهیم متغیر Y (Defective) را بر روی X_1 (Temp) و X_2 (density) و X_3 (Rate) رگرسیون بگیریم. داده‌ها در جدول 6.1 آمده است. ابتدا، مدل زیر شروع می‌نماییم:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

شکل 6.12 نمودار مانده‌گی استاندارد را در مقابل هرکدام متغیرهای مستقل و برازش شده نشان می‌دهد. همه این اشکال اشاره به یک طرح غیر تصادفی دارند. در شکل 6.13 نمودار پراکنش $\hat{Y}$ در مقابل Y رسم شده که با استفاده از آن می‌توان به ضعف یک خط راست به این نقاط پی برد. بنابراین مدل خطی برای داده‌ها مناسب نبوده، و باید به دنبال یک تبدیل مناسب باشیم.

بدین منظور از روش رسم پاسخ معکوس استفاده می‌نماییم. فرض می‌کنیم مدل مناسب بصورت زیر باشد:

$$Y = g(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon) \quad \text{یا} \quad g^{-1}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

با رسم مقادیر $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3$ در مقابل Y در شکل 6.14 درمی‌یابیم که تبدیل مناسب $g^{-1}(Y) = Y^{0.44}$ می‌باشد. بنابراین مدل مناسب برای این داده‌ها عبارتست از:

$$Y^{0.44} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

اعمال تبدیل بر روی متغیر پاسخ با استفاده از تکنیک روش باکس-کاکس که همانند گذشته به تکنیک خانواده بهبود یافته تبدیلات توانی برروی Y بصورت زیر تعریف خواهند شد:

$$\psi_M(\gamma,\lambda) = \psi_S(\gamma,\lambda)\cdot gm(\gamma)^{1-\lambda} = \begin{cases} gm(\gamma)^{1-\lambda}\cdot \dfrac{\gamma^{\lambda}-1}{\lambda} & ,\ \lambda\neq 0 \\ gm(\gamma)\log(\gamma) & ,\ \lambda\neq 0 \end{cases}$$

این روش بر این اساس است که در واقع برای برخی مقادیر λ، $\psi_M(\gamma,\lambda)$ دارای توزیع نرمال تقریبی است. این روش مقدار بهینه λ را، با ماکزیمم سازی تابع درستنمایی برآورد می نماید.

مثال: بررسی مثال گذشته، ژوکرد بالکس- کالکس: در شکل 6.15، تابع درستنمایی $\psi_M(\gamma,\lambda)$ در مقابل λ رسم شده است. بر این اساس، مقدار ماکزیمم کننده λ برابر با 0.145 خواهد بود که به جواب روش نمودار معکوس بسیار نزدیک است. در ادامه، نمودار پراکنش Y_6 در مقابل سه متغیر مستقل مربوطه رسم شده است و مشاهده می شود که رابطه آنها همه به یک رابطه خطی نزدیکتر شده است. در شکل 6.17، نمودار پراکنش مانده های استاندارد در مقابل متغیرهای مستقل و مقادیر برازش شده رسم شده که بهمراه آنها برای مدل تبدیل یافته مناسب بودن مدل جدید را تأکید دارند. در شکل 6.18، نمودار پراکنش $\sqrt{Y_6}$ در مقابل $\hat{Y}_6$ مقادیر برازش شده آن رسم شده است. با توجه به اینکه خط برازش شده به این نقاط به نسبت از ربع اول نزدیکتر گذشته، بهتر است لذا می توان به نیکویی مدل برازش شده پی برد. خروجی مربوط به مدل جدید در صفحه 175 آمده است. با استفاده از این خروجی می توان فهمید که متغیر مستقل X_3 تأثیر معنی داری بر مدل تغییرات Y_6 نداشته و ضریب آن برابر صفر است. این مطلب با رسم نمودارهای متغیر افزوده در شکل 6.20 به وضوح فهمیده می شود (برای اطمینان بیشتر زیرا X_3 با سایر متغیرهای مستقل براساس نمودارهای پراکنش اولیه رابطه خطی دارد و لذا بهتر است اثر خروجی آن بر متغیر وابسته به تنهایی سنجیده گردد. همچنین از صورت وجود همخطی بین متغیرهای مستقل، آماره های t-استودنت مربوطه هر کدام معنی داری ضرایب آنها دیگر قابل اعتماد نمی باشد.)

انجام تبدیلات بطور توأم بروی متغیرهای وابسته و مستقل:

در حالتیکه توزیع متغیرهای مستقل و پاسخ همگی دارای چولگی بود، و نیازبه تبدیل برای نزدیک (آشنا) شدن داده‌ها، می‌توانیم از یکی از دو رویکرد زیر استفاده نمائیم.

رویکرد ۱ این روش ترکیبی از روش باکس-کاکس چند متغیره و روش پاسخ مطلوب (غوطه‌ور) است. در این روش ابتدا، با استفاده از روش باکس-کاکس چند متغیره، تبدیل مناسب را برای متغیرهای مستقل یافته و سپس به کمک روش پاسخ مطلوب (غوطه‌ور) تبدیل مناسب را برای متغیر پاسخ پیشنهاد می‌دهیم.

$$(x_1, x_2, \dots, x_p) \xrightarrow[\text{متغیره}]{\text{باکس-کاکس چند}} \psi_M(x_1, \lambda_{x_1}), \psi_M(x_2, \lambda_{x_2}), \dots, (\psi_M(x_p, \lambda_{x_p}))$$

$$y = g\left(\beta_0 + \beta_1 \psi_M(x_1, \lambda_{x_1}) + \dots + \beta_p \psi_M(x_p, \lambda_{x_p})\right)$$

$$\psi_M(x_1, \lambda_{x_1}) \quad \psi_M(x_p$$

$$\text{plot}(y, \hat{y}) \longrightarrow \text{یافتن بهترین } \lambda_y \text{ ممکن} \qquad (\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p)$$

$$y^{\lambda_y} = \beta_0 + \beta_1 \psi_M(x_1, \lambda_{x_1}) + \dots + \beta_p \psi_M(x_p, \lambda_{x_p}) + \varepsilon$$

رویکرد ۲ در این روش با استفاده از متد باکس-کاکس چند متغیره تبدیلات مناسب برای $y, x_1, x_2, \dots, x_p$ را بدست می‌آوریم.

مثال: (سود مجله) یک تحلیلگر علاقه‌مند به فهم ارتباط بین سود حاصل از فروش یک مجله و آگهی‌های آن است. او داده‌هایش را براساس بررسی ۳۰۰ مجله در ایالات متحده جمع و متغیرهای وابسته و مستقل را به شرح زیر تعریف نموده است:

y : سود حاصل از آگهی

x_1 : تعداد صفحاتی که در آن آگهی (تبلیغات) می‌شود

x_2 : درآمد حاصل از اشتراک مشترکین

x_3: درآمد حاصل از رکدها

در شکل ۶.۲۱ نمودار پراکنش متغیرها رسم شده که نشان دهنده وجود چولگی در کلیه متغیرهاست. همچنین به نظر می رسد که بین متغیرهای مستقل رابطه خطی وجود داشته باشد. بنابراین می توان به اعمال تبدیل روی متغیرهای مستقل و وابسته روی آورد.

الف) استفاده از سه اعمال تبدیلات روی متغیرها با استفاده از رویکرد ۱:

براین اساس ابتدا تبدیل مناسب برای متغیرهای مستقل را با استفاده از بالکس - کاکس چند متغیره بدست می آوریم. بنابر خروجی صفحه ۱۷۷ می توان گفت که λ مربوط به هر سه متغیر مستقل تقریباً برابر صفر است. پس، تبدیل مناسب یک تبدیل لگاریتمی است. بنابراین:

$$Y = g(\beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \beta_3 \log(X_3) + \varepsilon) \quad (1)$$

در شکل ۶.۲۲، رسم نمودار معکوس Y در مقابل $\hat{Y}$ آورده ایم که تبدیل مناسب برای Y به ازای $\lambda_y = -0.25$ حاصل می شود. بنابراین مدل نهایی بصورت زیر خواهد بود:

$$Y^{-0.25} = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \beta_3 \log(X_3) + \varepsilon \quad (2)$$

اما با توجه به شکل $\lambda = 0$ نیز برای مقادیر کوچک و متوسط متغیر Y، برازش خوبی به داده ها دارد و چون متغیرهای مستقل با وابسته یکی است، لذا تبدیل جانشین دیگری می تواند $\log(Y)$ باشد و مدل بصورت زیر در نظر گرفته شود:

$$\log(Y) = \beta_0 + \beta_1 \log(X_1) + \cdots + \beta_3 \log(X_3) + \varepsilon \quad (3)$$

اما به هر حال برای هرکدام از دو تبدیل فوق بهتر است آنرا که جدول آنالیز واریانس قابل قبول تری را می دهد، به عنوان مدل نهایی در نظر بگیریم.

ب) اعمال تبدیل بر روی متغیرها با استفاده از رویکرد ۲: براساس خروجی جدول ۲ صفحه ۱۷۹، می توان نتیجه گرفت که روش بالکس - کاکس برای کلیه متغیرها تبدیل لگاریتمی را بعنوان یک تبدیل مناسب معرفی می نماید. در شکل ۶.۲۳، نمودار پراکنش ماتریسی برای همه متغیرها رسم شده. همانطور که از این شکل برمی آید، متغیرهای تبدیل یافته رابطه خطی معنی داری با هم دارند.

مانده‌ها را استاندارد در مقابل متغیرهای مستقل جدید و ۶.۲۵

در شکل ۶.۲۴، نمودارهای عیب‌شناسی رگرسیون برای مدل لگاریتمی رسم شده است که همگی شاخص‌های طرح تصادفی هستند. بنابراین مدل فوق مدل مناسبی برای داده‌ها می‌باشد.

همچنین در شکل ۶.۲۵، نمودار $\hat{Y}$ در مقابل $\hat{Y}$ برای متغیرهای تبدیل یافته رسم شده و چون نقاط تقریباً بر روی نیمساز ربع اول و سوم واقعند، لذا مدل برازش کافی به داده‌ها دارد.

در شکل ۶.۲۶ نمودارهای عیب‌شناسی رگرسیون برای مدل جدید رسم شده‌اند (مدل ۳) که همگی دلالت بر مناسب بودن مدل مورد بررسی دارند. البته نمودار پائین دست چپ این شکل کمی نشان‌دهنده خطاهای که دارای روند افزایشی و سپس کاهشی است. همچنین در نمودار پائین دست راست نقاط چندان عمودی نشده‌اند و مرز لورج‌ها یعنی $\frac{2(p+1)}{n} = 0.109$ دقیقاً نشان‌دهنده مرکزیت مانده‌های استاندارد است.

در شکل ۶.۲۷، نمودارهای متغیر افزوده رسم شده که براساس آن می‌توان فهمید که متغیر $\log(X_2)$ تأثیر معنی‌داری روی مدل (۳) ندارد و بهتر است حذف گردد. این مطلب با استفاده از خروجی صفحه ۱۸۴ نیز کاملاً آشکار است.

برآورد درصد تغییرات متغیر وابسته براثر تغییر هر یک از متغیرهای مستقل: مدل رگرسیون زیر را در نظر بگیرید:

$$\log(Y) = \beta_0 + \beta_1 \log(X_1) + \beta_2 \log(X_2) + \varepsilon$$

با ثابت بودن متغیر X_1 می‌توان نوشت:

$$\beta_2 = \frac{\Delta \log(Y)}{\Delta X_2} = \frac{\log(Y_2) - \log(Y_1)}{\Delta X_2} = \frac{\log(Y_2/Y_1)}{\Delta X_2} = \frac{Y_2/Y_1 - 1}{\Delta X_2} = \frac{\%\,\Delta Y}{100\,\Delta X_2}$$

$$\Rightarrow \%\Delta Y = 100\,\Delta X_2 \times \beta_2 \xrightarrow{\Delta X_2 = 1} \boxed{\%\Delta Y = 100\,\beta_2}$$

مثال: برآورد تیراژ یک روزنامه در یکشنبه صفحه ۱۸۴

در این مثال براساس ۸۹ مشاهده در ایالات متحده آمریکا متغیرهای زیر تعریف شده‌اند:

Y: لگاریتم تیراژ در یکشنبه

X_1: لگاریتم تیراژ در سایر روزهای هفته

X_2: متغیر دوحالتی صفر و یک بطوریکه صفر به جز حالتی است که در یک شهر تنها یک روزنامه چاپ

می‌شود و یک برای حالتی است که در آن شهر روزنامهٔ دیگری که حالتی خلاصه دارد نیز علاوه بر روزنامهٔ اصلی چاپ می‌شود.

در شکل ۶.۲۸، نمودار پراکنش Y در مقابل X_1 رسم شده است. ابتدا مدل $\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ برای

در شکل ۶.۲۹ نمودار مانده‌های استاندارد در مقابل متغیرهای مستقل و برازش شده رسم شده است که هیچ نشانگر یک طرح تصادفی هستند. شکل ۶.۴۰ نمودار پراکنش Y را در مقابل $\hat{Y}$ نشان می‌دهد که استفاده از آن می‌توان به نیکویی برازش پی برد. در شکل ۶.۴۱ نمودارهای مربوط به عیب‌یابی رگرسیون رسم شده‌اند که همهٔ آنها اعتبار مدل لگاریتمی را تأیید می‌نمایند. خروجی R مربوط به مدل فوق در صفحهٔ ۱۸۹ آمده که براساس آن می‌توان به معنی‌داری هردو متغیر مستقل در مدل پی برد. براساس آن می‌توان گفت:

الف) با ثابت بودن متغیر X_2، افزایش متغیر پاسخ به ازای یک واحد افزایش X_1 برابر است با ۶۲٪

ب) با ثابت بودن X_1، متغیر پاسخ به ازای یک واحد افزایش X_2 مقدار ۵۳٫۱٪ کاهش می‌یابد.

در شکل ۶.۴۲، نمودارهای متغیرافزوده برای مدل فوق رسم شده‌اند که استفاده از آنها می‌توان به معنی‌داری اثر هرکدام از متغیرهای مستقل بر این مدل پی برد.

در ادامه، قادر خواهیم بود که مقدار متغیر Y را به ازای مقادیر X_1 و X_2 به صورت زیر برآورد و فاصلهٔ اطمینان ۹۵٪ به دست آوریم:

$X_2 = 1 \rightarrow \hat{Y} = 12{,}028$ $\quad (11{,}721 \; , \; 12{,}445)$

$X_2 = 0 \rightarrow \hat{Y} = 12{,}589$ $\quad (12{,}280 \; , \; 12{,}858)$

$X_2 = 1 \rightarrow$ مقدار پیش‌بینی شده $= \exp(12{,}028)$

$X_2 = 0 \rightarrow$ " " " $= \exp(12{,}589)$

هم‌خطی چندگانه: اگر بین دو یا چند متغیر مستقل در مدل رگرسیون چندگانه ما رابطهٔ خطی وجود داشته باشد، گوییم مدل رگرسیون دارای هم‌خطی است. هم‌خطی کامل زمانی رخ می‌دهد که یکی از متغیرهای مستقل یک تابع دقیق از یک یا چند متغیر مستقل دیگر باشد و هم‌خطی ناقص زمانی خواهد رخ داد که این تابع تقریبی باشد. در عمل هم‌خطی کامل رخ نخواهد داد زیرا اگر یکی از متغیرهای مستقل تابعی دقیق از سایر متغیرهای رگرسیون باشد، خطای اندازه‌گیری سبب می‌شود که این رابطهٔ دقیق به یک رابطهٔ تقریبی تبدیل گردد. حال چنانچه هم‌خطی تقریبی اتفاق افتد، آنگاه $|X'X| \to 0$ و لذا ماتریس $(X'X)^{-1}$ وجود نخواهد داشت.

منابع اصلی هم‌خطی:

۱- انتخاب روش نگرداوری داده‌ها

۲- گذاشتن قیدهایی روی مدل

۳- شناسایی مدل

۴- انتخاب مدلی که تعداد متغیرهای مستقل آن بیشتر از تعداد مشاهدات باشد.

در حالت کلی، با زیاد شدن متغیرهای پیشگو، احتمال بوجود آمدن هم‌خطی بین متغیرهای مستقل افزایش می‌یابد.

اثرات هم‌خطی: همانطور که گفته شد، اگر هم‌خطی وجود داشته باشد $|X'X| \simeq 0$ شده و لذا ماتریس $(X'X)^{-1}$ خیلی بزرگ خواهند شد چون $\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$، لذا می‌توان نتیجه گرفت که دقت برآورد کمترین مربعات کاهش یافته و همچنین فواصل اطمینان بدست آمده برای پارامترها پهن خواهند شد. گاهی اوقات ممکن است حتی علامت پارامتر β_p یا $\beta_0, \ldots, \beta_p$ به اشتباه برآورد گردد. همچنین مقادیر آماره‌ی t مربوط به ضرایب رگرسیون بسیار کوچک شده و غیرمعنی‌دار به نظر می‌رسند زیرا $t_j = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 c_{j+1,j+1}}}$ که در آن $c_{j+1,j+1}$ $(j+1)$امین درایهٔ قطری اصلی ماتریس $(X'X)^{-1}$ است.

چگونه می‌توان به وجود همخطی پی برد؟ کاهی اوقات آماره F جدول آنالیز واریانس معنی‌دار است ولی تعداد زیادی از ضرایب رگرسیون متغیرها معنی‌دار نیستند. در چنین مواردی می‌توان به وجود همخطی پی برد ولی بسیاری از مجموعه داده‌هایی که همخطی معنی‌داری دارند این رفتار را از خود نشان نمی‌دهند و لذا این معیار چندان سودمند نمی‌باشد. در چنین مواردی از معیار زیر می‌توانند مورد استفاده قرار گیرند:

Condition number

۱- عدد شرط: این آماره براساس مقادیر ویژه ماتریس $(X'X)$ بصورت زیر تعریف می‌شود:

$$k(X'X) = \sqrt{\frac{\max \lambda_i}{\min \lambda_i}} \qquad i = 1, \dots, p$$

که در آن $\lambda_1, \dots, \lambda_p$ مقادیر ویژه ماتریس $(X'X)$ هستند.

$k(X'X)$ کمتر از ۱۰۰ همخطی وجود ندارد
" بین ۱۰۰ و ۱۰۰۰ همخطی متوسط تا شدید است
" بیشتر از ۱۰۰۰ همخطی شدید است

۲- عامل تورم واریانس: (Variance Inflation Factor) عامل تورم واریانس مربوط به متغیر مستقل jام را با VIF_j نشان داده و بصورت زیر تعریف می‌کنیم:

$$VIF_j = \frac{1}{1 - R_j^2}$$

که در آن R_j^2 ضریب تعیین در مدل رگرسیون X_j بر سایر متغیرهای پیشگو، مستقل می‌باشد.

$$R_j^2 \to 1 \Rightarrow VIF_j \to \infty$$

هرگاه $VIF_j > 5$ $(R_j^2 > 0.8)$ آنگاه می‌توان به وجود همخطی پی برد.

همانطورکه در رگرسیون I دیدیم، در مدل خطی دو متغیره

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

$$Var(\hat{\beta}_j) = \frac{1}{1 - r_{12}^2} \cdot \frac{\sigma^2}{S_{X_j X_j}} \quad , \quad S_{X_j X_j} = \sum_{i=1}^{n} (X_{ij} - \bar{X}_j)^2 \qquad j = 1, 2$$

بنابراین می‌توان نوشت:

$$Var(\hat{\beta}_j) = VIF_j \cdot \frac{\sigma^2}{S_{X_j X_j}} \Rightarrow VIF_j \to \infty \Rightarrow Var(\hat{\beta}_j) \to \infty$$

دقت برآورد کمترین مربعات کم می‌شود.

رابطه اخیر در حالت کلی نیز درست بود. و می‌توان نشان داد: $y=\beta_0+\beta_1 x_1+\cdots+\beta_p x_p+\varepsilon$

$$Var(\hat{\beta}_j)=\frac{1}{1-R_j^2}\cdot\frac{\sigma^2}{S_{x_jx_j}} \qquad ,j=1,\cdots,p$$

$$\Rightarrow Var(\hat{\beta}_j)=VIF_j\cdot\frac{\sigma^2}{S_{x_jx_j}}$$

مثال: (ساخت پل)

در مسأله ساخت پل، متغیرهای زیر قابل بررسی می‌باشند:

Y: زمان ساخت برحسب روز (Time)

X_1: مساحت پل برحسب فوت مربع (DArea)

X_2: هزینه ساخت برحسب ۱۰۰۰ دلار (CCost)

X_3: تعداد طراحان پل (Dwgs)

X_4: طول پل (Length)

X_5: تعداد طاق‌ها (Spans)

ابتدا به کمک روش باکس-کاکس چند متغیره، تبدیل مناسب جهت تبدیل داده‌ها به یک توزیع نرمال چند متغیره و ای در رابطه خطی بین آنها را بدست می‌آوریم. و سپس بهترین ای مناسب را برای کلیه متغیرها برابر صفر در نظر گرفت و در نتیجه مدل زیر حاصل می‌شود:

$$Log(Y)=\beta_0+\beta_1 log(x_1)+\cdots+\beta_5 log(x_5)+\varepsilon$$

شکل ۶.۳۹ و ۶.۴۰ بترتیب قادرند عوارض ایجاد شده متغیرها قبل و بعد از تبدیل نشان دهند.

شکل ۶.۴۱ عوارض باقیمانده‌ها را نشان می‌دهد در مقابل متغیرهای وابسته و مستقل برای مدل تبدیل یافته (راست) می‌دهد. (همه عوارض ثابت نشده طرح تصادفی هستند).

شکل ۶.۴۲ عوارض $\hat{Y}$ در مقابل $\widehat{Log(Y)}$ (راست) می‌دهد. در شکل ۶.۴۳ عوارض مربوط به مجموعه‌های رگرسیون رسم شده‌اند که شکل دلالت بر متغیر بودن مدل فوق دارند. در

ادامه

هنگامی که به جدول آنالیز واریانس بزرگی خوبی می‌بینیم که با وجود اینکه آماره F در سطح بالای معنی دار است، تنها آماره t مربوط به متغیر X3 معنی دار بود، و سایر متغیرها دارای اثر معنی داری بر مدل رگرسیون نمی‌باشند. علاوه بر این علامت مربوط به ضرایب برآورد شده متغیرهای مستقل X1 (مساحت) و X4 (طول پل) متناقض بنظر هر چه سطح (طول) یک پل افزایش یابد زمان ساخت آن کمتر می‌شود که این جمله منطقاً غلط است. همچنین براساس شکل (6.45) (نمودارهای متغیر افزوده) تنها متغیر مستقل X3 دارای اثر معنی دار بر مدل بوده و سایر متغیرها فاقد اثر معنی دار بر مدلند.

هرگاه دو یا چند متغیر مستقل با همبستگی بالا وارد یک مدل رگرسیون شوند، آنها بطور مؤثری اطلاعات تکراری را در مورد متغیر وابسته حمل کرده و این باعث می‌شود که روش کمترین مربعات نتواند اثر جزئی هریک از آنها روی متغیر پاسخ تشخیص دهد. در چنین مواقعی آماره F جدول آنالیز واریانس بزرگ شده، سطح معنی داری آن بسیار کوچک می‌شود (ممکن است) در حالیکه تعداد کمی از ضرایب رگرسیون معنی دار شوند. شکل زیر ایجاد شده در چنین مواقعی، برآورد غلط علامت ضرایب رگرسیون می‌باشد.

بازرسی همبستگی متغیرهای مستقل در صفحه 2.2 آمده است. همانطور که دیده می‌شود بین متغیرهای مستقل همبستگی خطی بالایی وجود دارد. همچنین مقادیر VIF برای متغیرهای مستقل عبارتند از:

X_1	X_2	X_3	X_4	X_5
7.14	8.48	4.31	8.01	4.88

فصل هفتم :

"نحوهٔ انتخاب متغیرهای در مدل"

در این بخش روشهای برای انتخاب بهترین مدل ممکن از میان تمام مدلهای قابل انتخاب، معرفی خواهد شد. مدل رگرسیون چندگانه زیر را در نظر بگیرید :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

متغیرهای مدل

روشهای انتخاب متغیرها به انتخاب بهترین زیر مجموعهٔ ممکن از متغیرهای مستقل کمک می‌کند. از بین این زیر مجموعه‌ها نهایتاً متغیرهای مستقل :

در ادامه معیارهایی جهت ارزیابی زیر مجموعه‌هایی از متغیرهای مستقل ارائه خواهد شد :

۱- معیار R^2_{adj} : همانطور که در رابطه افزایش متغیرهای مستقل بر R^2 اغلب باعث افزایش R^2 می‌شوند ولی این اتفاق در مورد R^2 تصحیح شده که بصورت زیر تعریف می‌شود عمل می‌کند :

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{S_{yy}/(n-1)}$$

می‌توان ثابت نمود که در صورتی افزایش R^2_{adj} با افزودن یک متغیر به مدل اتفاق خواهد افتاد که مقدار F آماره F جزئی مربوط به آن متغیر مستقل از یک بیشتر باشد. در عمل تعداد زیادی متغیرهای مستقل طوری انتخاب می‌شود که منجر به بیشترین آماره R^2_{adj} شوند. می‌توان نشان داد ماکزیمم R^2_{adj} برای متغیرهای مستقلی اتفاق خواهد افتاد که دارای کمترین $\frac{SSE}{n-p-1}$ باشند.

البته چنانچه افزایش یک متغیر مستقل جدید به مدل، باعث افزایش جزئی در R^2_{adj} شود برای تفسیر بهتر نتایج و سادگی بیشتر، بهتر است آن متغیرها در مدل نهایی در نظر نگیریم. بعنوان مثال فرض کنید :

$p=9 \rightarrow R^2_{adj} = .1691$ $\quad$ $p=10 \rightarrow R^2_{adj} = .1694$

$p=8 \rightarrow R^2_{adj} = .1541$

در مدلهای فوق هرچند $p=10$ متغیر مستقل دارای بیشترین مقدار R^2_{adj} است ولی بهتر است که $p=9$ متغیر در نظر گرفته شود زیرا تغییر چندانی در R^2_{adj} ایجاد نشده است.

سه معیار زیر تنها در صورتی قابل استفاده می‌باشند که توزیع متغیر پاسخ پیشگو وابسته نرمال باشد (در صورت نرمال بودن توزیع، آنها تابع درستنمایی برابرند):

$$\log L(\beta_0, \dots, \beta_p, \sigma^2 | Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip})^2$$

$$\Rightarrow \log L(\hat{\beta}_0, \dots, \hat{\beta}_p, \hat{\sigma}^2 | Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} SSE$$

$$\Rightarrow \hat{\sigma}^2_{MLE} = \frac{SSE}{n} \quad , \quad \hat{\sigma}^2_{LS} = S^2 = \frac{SSE}{n-p-1}$$

$$\Rightarrow \log L(\hat{\beta}_0, \dots, \hat{\beta}_p, \hat{\sigma}^2 | Y) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{SSE}{n}\right) - \frac{n}{2}$$

۲- معیار AIC (Akaike's Information criterion):

این معیار براساس لگاریتم تابع درستنمایی تعریف شده و میزان برازش را اندازه می‌گیرد. AIC معیاری جهت اندازه‌گیری الطلاعات بوده و بصورت زیر تعریف می‌شود:

$$AIC = -2\left[\log L(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p, \hat{\sigma}^2 | Y) - (p+2)\right]$$

$$= -2\left[-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{SSE}{n}\right) - \frac{n}{2} - (p+2)\right]$$

$$= n\log\left(\frac{SSE}{n}\right) + 2p + \text{other terms}$$

بطوریکه سایر جملات به نحوه برازش مدل بستگی نداشته و برای همه مدلهای ممکن مشترکند.

بنابراین در نرم‌افزار R، AIC بصورت زیر محاسبه می‌شود:

$$AIC = n\log\left(\frac{SSE}{n}\right) + 2p$$

لازم به توضیح است که بهترین مدل، مدلی است که دارای کمترین AIC باشد.

۳- معیار AIC تصحیح شده (AIC_c): اگر حجم نمونه کوچک و یا تعداد پارامترهای مدل نسبتاً زیاد باشد، بهتر است از AIC_c استفاده شود. اگر $\frac{n}{p+2} \geq 40$ باشد، AIC_c به کار برده شود. این معیار بصورت زیر محاسبه می‌شود:

$$AIC_c = AIC + \frac{2(p+2)(p+3)}{n-p-1} \qquad AIC_c \xrightarrow{n\to\infty} AIC$$

۴- معیار BIC (Bayesian Information Criterion):

معیار اطلاع بیز بصورت زیر محاسبه می‌شود:

$$BIC = -2\log L(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2 | y) + (p+2)\log(n)$$

شبیه به معیار AIC، هرچه BIC کوچکتر، شبه مدل برازش بهتری برای داده‌ها دارد.

۵- معیار C_p مالوس: این معیار نخستین بار توسط مالوس پیشنهاد و بصورت زیر تعریف شد:

$$C_{p^*} = \frac{SSE_{p^*}}{s^2} - (n-2p^*)$$

که در آن SSE_{p^*} مجموع مربعات خطای مدلی با p^* پارامتر از جمله β_0 و s^2 برآوردی از σ^2 یا MSE مدلی با بیشترین تعداد متغیرهای مستقل ممکن است. حال اگر مدلی با p^* پارامتر کفایت کند در اینصورت $E(SSE_{p^*}) = (n-p^*)\sigma^2$ و چون بنا به فرض $E(s^2) = \sigma^2$ است لذا می‌توان نوشت:

$$E(C_{p^*}) \cong \frac{(n-p^*)\sigma^2}{\sigma^2} - (n-2p^*) = p^*$$

بنابراین می‌توان نتیجه گرفت هرگاه، مقدار C_{p^*} به p^* نزدیک شود، مدل مناسب خواهد بود

تصمیم‌گیری در مورد کلاسیک کالکسیون همهٔ زیر مجموعه‌های ممکن از متغیرهای پیشگو:

برای انجام تصمیم‌گیری در مورد متغیرهای مفید در مدل دو راه کلی زیر که کاملاً متفاوت وجود دارد:

۱- همهٔ رگرسیون‌های ممکن: این روش براساس برازش همهٔ 2^p مدل رگرسیون ممکن به متغیرهای و شناسایی زیر مجموعه‌ای از متغیرهای مستقل است که معیار برازش R^2_{adj} را ماکزیمم و یا معیارهای اطلاع را مینیمم نماید. البته باید توجه داشت که معیارهای متفاوت ممکن است منجر به نتایج متفاوتی گردند و هیچ معیاری به تنهایی همیشه بهترین نخواهد بود. لذا باید بیشتر از یک معیار را در نظر بگیریم.

مثال: مجدداً به مسأله ساخت پل برمی‌گردیم. ابتدا مدل شکل تمام متغیرها را در نظر می‌گیریم:

$$\log(Y_1) = \beta_0 + \beta_1 \log(X_1) + \cdots + \beta_5 \log(X_5) + \varepsilon$$

خروجی R مربوط به مدل فوق در صفحه ۲۳۴ کتاب آمده است. همانطور که مشاهده می‌شود با وجود معنی‌دار بودن آماره F جدول آنالیز واریانس، غالب آماره‌های t جزئی + جزئی بغیر از ضریب مربوط به متغیر X_3 معنی‌دار نمی‌باشند. بنابراین تحقیق در مورد انتخاب متغیرهای پیشگو را با محاسبه R^2_{adj} شروع می‌کنیم. در شکل ۷.۱ نمودار R^2_{adj} در مقابل تعداد متغیرهای مستقل در مدل رسم شده است. بعنوان مثال، بهترین زیرمجموعه شامل ۲ متغیر مستقل X_3 و X_5 و بهترین زیرمجموعه شامل ۳ متغیر مستقل X_3، X_5 و X_{20} است. در جدول ۷.۱، مقادیر R^2_{adj}، AIC، AIC_c و BIC برای زیرمجموعه‌های بهینه با حجم‌های یک تا پنج آورده شده‌اند. طبق این جدول، دو معیار R^2_{adj}، AIC بهترین مجموعه ممکن را شامل X_3، X_5 و X_1 پیشنهاد می‌کنند، ولی دو معیار AIC_c، BIC بهترین مجموعه ممکن را شامل X_3 و X_5 در نظر می‌گیرند. با عنایت به آماره‌های این جدول و تفاوت بسیار جزئی آماره‌های R^2_{adj} و AIC برای دو و سه متغیر مستقل بهترین انتخاب ممکن، مدلی شامل X_3 و X_5 بصورت زیر است که خروجی مربوط به آن در صفحه ۲۳۶ آمده است:

$$\log(Y) = \beta_0 + \beta_3 \log(X_3) + \beta_5 \log(X_5) + \varepsilon$$

۲- گزینش به روش گام به گام (Stepwise Selection):

این روش در واقع جنبه اصلاح شده روش رگرسیونی پیشرو است که به ما اجازه می‌دهد در هر مرحله متغیرهای لحاظ شده در مدل در مراحل قبل را دوباره امتحان کنیم. متغیری که در مرحله قبل در مدل وارد شده، ممکن است در مرحله بعدی بخاطر همبستگی آن با سایر متغیرهای مستقل در الگو زائد به نظر برسد. برای بررسی این موضوع در هر مرحله یک آزمون F جزئی برای هر متغیر که در مدل است انجام می‌شود دلو اینکه این متغیر جدیدترین متغیر

وارد شده به مدل رگرسیون، یک به یک و هر مرحله از آنکه چه وقت به مدل وارد شده‌اند. در هر مرحله متغیر با کوچکترین F جزئی با معنی (در صورت وجود) حذف شده و مدل با متغیرهای باقیمانده دوباره برازش می‌شود و Fهای جزئی محاسبه می‌گردند و بطور شبیه به امتحان می‌گردند و این کار ادامه پیدا می‌کند. این فرآیند تا زمانی ادامه می‌یابد که متغیرهای بیشتری به رگرسیون وارد نشوند و یا از مدل حذف نگردند. بنابراین روش گام به گام حداکثر $\frac{R(R+1)}{2}$ مدل از میان تمام 2^R مدل $1+2+\cdots+(R-1)+R=$ ممکن را بررسی می‌نماید. لذا، می‌توان گفت که روش گام به گام لزوماً مدلی را ارائه می‌نماید که بهترین مدل از لحاظ کمترین معیارهای اطلاع گفته شده در قسمت قبل است. به عبارت دیگر نتایج روش گام به گام، نتایج معیارهای اطلاعاتی R^2_{adj} ممکن است یکسان نباشد. با پیاده‌سازی این روش روی داده‌های مثال من، روش پیشرو و پسرو هر دو هیچ کدام مدل ممکن را ندی، متغیرهای X_3، X_5 و X_2 معرفی خواهند نمود.