بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

بسمه تعالی

مشخصات پایان نامه تحصیلی دانشجویان

دانشگاه فردوسی مشهد

عنوان پایان نامه : تکنیک های پیشرفته رگرسیون

نویسنده: احسان زمان زاده	**استاد راهنما: دکتر ناصررضا ارقامی**
دانشکده : علوم ریاضی	**گروه: آمار**
رشته تحصیلی : آمار ریاضی	**تاریخ دفاع: ۸۷/۶/۱۶**
مقطع تحصیلی : کارشناسی ○ کارشناسی ارشد ● دکتری ○	**تعداد صفحات : ۲۱۳ صفحه**

چکیده :

اینکه یک مدل رگرسیون به مجموعه ایی از متغیرها برازش خوبی نداشته باشد اما به تبدیلاتی از آن ها برازش خوبی داشته باشد، مساله ایی غیر معمول نیست. باکس و کاکس (۱۹۶۴) به بررسی تبدیلات متغیر پاسخ، با تاکید بر روی تبدیلات توانی، که به تبدیلات باکس-کاکس معروف گشت، پرداختند. هر چند از این روش در عمل بسیار مفید است اما روش باکس-کاکس دارای این محدودیت است که تنها به یک خانواده از تبدیلات پارامتری محدود شده است و به تبدیل متغیر های پیش‌بین نمی پردازد.

اما در سال های اخیر با در دسترس قرار گرفتن رایانه های سریع و ارزان قیمت، تکنیک های ناپارامتری برازش منحنی از مقبولیت خاصی برخوردار گشته اند. برآورد تبدیلات بهینه یکی از روش های اصلی هموارسازی است که دارای فرضیات بسیار کمی روی شکل منحنی رگرسیون می باشد. این روش هموارسازی را می توان به طور کلی به دو قسمت تقسیم کرد. دسته اول روش هایی هستند که تنها تبدیل روی متغیرهای پیش بین انجام مـی شـود و دسته دوم، آن هایی هستند که هم تبدیل روی متغیر پاسخ و هم روی متغیرهای پیش بین انجام می گردد.

در این رساله به معرفی و بیان و بررسی خواص هر دو دسته از این روش ها خواهیم پرداخت.

از دسته اول، بیشتر روی روش‌هایی همانند رگرسیون اسپلاین، رگرسیون خطی موضعی، هموار کننده برتر، LOESS و نسخه استوار آن یعنی LOEWSS تمرکز خواهیم کرد.

و در دسته دوم، بر روی دو الگوریتم کارآمد ، یعنی الگوریتم امید ریاضی شرطی جایگزین (ACE) و الگوریتم جمع پذیری و ثبات واریانس (AVAS) تمرکز خواهیم کرد همچینن با تمرکز بیشتر بر روی این دو الگوریتم، به برخی از ناهنجاری ها و معایب آن ها، هنگامی که از آن ها بدون دقت به عنوان ابزاری در رگرسیون استفاده کنیم، خواهیم پرداخت.

واژه های کلیدی : اسپلاین هموارساز، رگرسیون خطی موضعی، هموارکننده برتر، LOESS، LOEWSS، الگوریتم امید ریاضی شرطی جایگزین، الگوریتم جمع پذیری و ثبات واریانس	**امضای استاد راهنما:** **تاریخ**

Equation Section 1

فصل اول

رگرسیون خطی

۱-۱ رگرسیون خطی ساده

۱-۱-۱ برازش مدل رگرسیون خطی ساده

چنانچه اشاره کردیم، تحلیل رگرسیون یک روش آماری برای بررسی و الگو بندی رابطه بین متغیرهاست. در ساده ترین حالت رگرسیون خطی فرض می شود که متغیرها بوسیله یک معادله خط راست در ارتباط اند، لذا با مروری از خط راست شروع می کنیم. شکل کلی یک خط راست بصورت $Y=\beta_۰+\beta_۱X$ است که در آن $\beta_۰$ را عرض از مبدا و $\beta_۱$ را شیب خط می نامند. در این قسمت به مطالعه حالت ساده رابطه بین یک متغیر پاسخ Y و یک متغیر پیش‌بین X می پردازیم. این الگوی رگرسیون خطی ساده عبارت است از

$$Y_i=\beta_۰+\beta_۱X_i+\varepsilon_i \qquad i=۱,۲,\ldots,n \tag{۱.۱}$$

که در آن ε یک جمله خطاست که متغیری است تصادفی و غیر قابل مشاهده. در ضمن $E(\varepsilon)=۰$ و $Var(\varepsilon)=\sigma^۲$ (ثابت) فرض می شود و علاوه بر آن ε ها ناهمبسته فرض می شوند.

متغیر پاسخ Y یک متغیر تصادفی پیوسته است در حالی که X تصادفی نبوده و توسط تحلیل گر کنترل و با خطای قابل اغماضی اندازه گیری می شود. بنابراین در هر مقدار ممکن X برای Y یک توزیع احتمال وجود دارد. چون X تصادفی نیست، لذا می توان نوشت

$$E(Y)=\beta_0+\beta_1 X$$

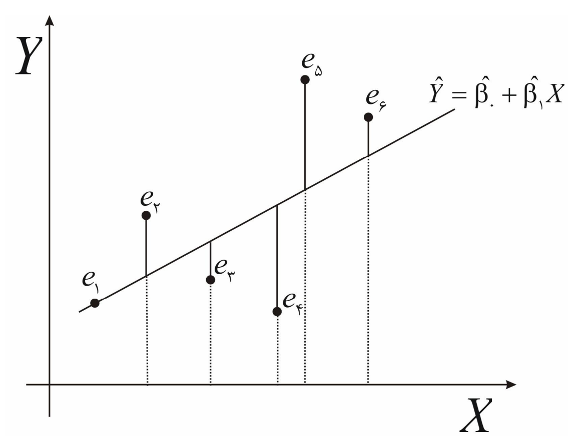
و می گوئیم رگرسیون Y بر X خطی است. پارامترهای β_0 و β_1 را معمولا **ضرایب رگرسیون** می نامند ودر واقع پارامترهایی هستند که باید برآورد شوند.

این ضرایب اغلب تعبیر ساده و مفیدی دارند. β_1 در واقع میزان افزایش در میانگین Y است هنگامی که یک واحد افزایش در متغیر پیش‌بین (X)رخ می دهد. در صورتی که دامنه داده ها شامل $X=0$ باشد در آن صورت β_0 میانگین متغیر پاسخ Y به ازای $X=0$ است در غیر این صورت β_0 هیچگونه تعبیر عملی ندارد.

اولین سوالی که در مسئله رگرسیون خطی ساده مطرح می شود، این است که چگونه با داشتن یک نمونه $Y_1,Y_2,...,Y_n$ همراه با مقادیر متناظر متغیر X یعنی $X_1,X_2,...,X_n$، ضرایب β_0 و β_1 را برآورد کنیم؟

استفاده از روش کمترین مربعات، یک روش ساده و منطقی برای پاسخ دادن به این سوال است.

در این روش بهترین خط راست برازش شده خطی است که مجموع توان‌های دوم خطاها (توان دوم طول قطعه خط‌های عمودی رسم شده در نمودار پراکنش شکل ۱–۱)را مینیمم کند. ایده اصلی در اینجا این است که هر چه انحرافات مقادیر مشاهده شده از این خط کمتر باشد خط برازش شده به داده ها نزدیک تر است. بنابراین در این روش $\hat{\beta}_0$ و $\hat{\beta}_1$ را طوری تعیین می کنیم که مجموع توان‌های دوم خطاها یعنی $e_1^2+e_2^2+...+e_n^2=\sum_{i=1}^{n} e_i^2$ مینیمم شود، که در آن e_i تفاوت بین مقدار واقعی Y_i و $\hat{Y}_i=\hat{\beta}_0+\hat{\beta}_1 X_i$ است.



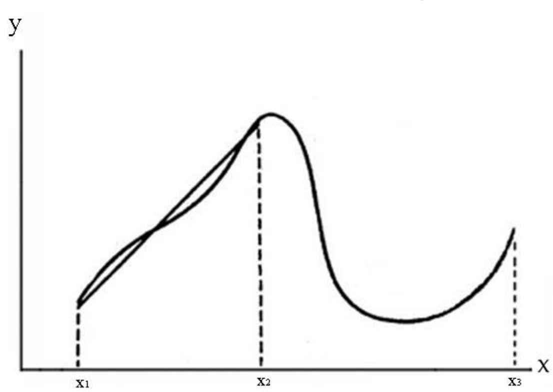
شکل ۱–۱ برازش خط رگرسیون با استفاده از روش کمترین توان‌های دوم

برای بدست آوردن $\hat{\beta}_0$ و $\hat{\beta}_1$، کافی است از $\sum_{i=1}^{n} e_i^2$ نسبت به $\hat{\beta}_0$ و $\hat{\beta}_1$ مشتق گرفته و مساوی صفر قرار دهیم. در این صورت به دست خواهیم آورد:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2} \tag{2.1}$$

۱-۱-۲ خطر برون یابی در رگرسیون

نکته مهمی که باید به آن توجه داشت این است که چون معادلات رگرسیون تنها تقریبی از رابطه واقعی X و Y هستند، لذا این معادلات عموماً فقط در ناحیه ای که مقادیر متغیرهای پیش‌بین مشاهده شده اند معتبر هستند، و نمی توان در خارج از این ناحیه، از این معادلات استفاده کرد. به عنوان مثال فرض کنید که داده های (x, y) در بازه $x_1 \le x \le x_2$ جمع آوری شده باشند. در این بازه معادله رگرسیون خطی یک تقریب خوب از رابطه واقعی میان X و Y است. حال فرض کنید که بخواهیم از این معادله خط رگرسیون برآورد شده برای پیش بینی مقدار Y در بازه $x_2 \le x \le x_3$ استفاده کنیم، واضح است که این معادله خط رگرسیون برآورد شده عملکرد خوبی در این ناحیه از مقادیر x را به دلیل خطای تشخیص نامناسب الگو ندارد. (شکل ۱– ۲).



شکل ۱-۲ خطر برون یابی در رگرسیون

۱-۱-۳ اندازه ای از کیفیت برازش خط راست

به محض اینکه خط کمترین توان‌های دوم تعیین شد می خواهیم ارزیابی کنیم که آیا خط برازش شده واقعا در تخمین Y به ما کمک می کند یا خیر و اگر کمک می کند این کمک به چه میزان است. برای این منظور اگر فرض کنیم

انحراف از میانگین X ها $\quad x_i = X_i - \bar{X}$

انحراف از میانگین Y ها $\quad y_i = Y_i - \bar{Y}$

انحراف از میانگین $\hat{Y}$ ها $\quad \hat{y}_i = \hat{Y}_i - \bar{Y}$

می توان نوشت:

$$\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2 = SS_x$$

پراکندگی کل:
$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \left(Y_i - \bar{Y} \right)^2 = SS_y = SS_T$$

پراکندگی تبیین شده:
$$\sum_{i=1}^{n} \hat{y}_i^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y} \right)^2 = SS_{\hat{y}} = SS_R$$

پراکندگی تبیین نشده:
$$\sum_{i=1}^{n} e_i^2 = SS_e = SS_{\mathrm{Res}}$$

با این نمادها در قضیه۱-۱ نشان می دهیم که:

$$SS_T = SS_R + SS_{\mathrm{Res}} \qquad (۳.۱)$$

یکی از منابع پراکندگی SS_y ملاحظه می کنیم که $SS_{\hat{y}}$ است که عامل این پراکندگی صرفا متغیر پیش‌بین X است و چون $\hat{Y}$ یک تابع خطی از X است، لذا $SS_{\hat{y}}$ را می توان قسمتی از پراکندگی Y که به وسیله X قابل توجیه است نامید به همین دلیل $SS_{\hat{y}}$ را پراکندگی بیان شده(تبیین شده)[1] یا نوسانات بیان شده (تبیین شده) متغیر وابسته می نامند. از طرفی با توجه به رابطه (۱-۳) پراکندگی Y به اندازه SS_e بیش از پراکندگی $\hat{Y}$ است، بنابراین SS_e سهمی از پراکندگی Y است که توسط تغییرات X قابل توجیه نمی باشد و معلول اثر متغیرهای دیگری است که بر Y موثرند و در الگو لحاظ نشده اند. به این دلیل SS_e را پراکندگی یا نوسانات بیان نشده (تبیین نشده)[2] می نامیم. مجموع مربعات تبیین شده[3] و مجموع مربعات خطا[4] به ترتیب نام های دیگر $SS_{\hat{y}}$ و SS_e می باشند.

---

[1] Explained variations
[2] Unexplained variations
[3] Explained sum of squares
[4] Error sum of squares

رابطه (۱-۳) به رابطه تجزیه واریانس موسوم است. واضح است که $0 \leq \frac{SS_{\hat{Y}}}{SS_Y} \leq 1$ و می تواند معیار خوبی برای میزان دقت رابطه خطی Y با X باشد و چون مثبت است می توان آن را با R^2_{XY} نشان داد که آن را ضریب تعیین می نامند بنابراین

$$R^2 = \frac{SS_{\hat{Y}}}{SS_y} = \frac{SS_R}{SS_T}$$

$$= \frac{\text{پراکندگی بیان شده به وسیله } X}{\text{پراکندگی کل}}$$

توجه می کنیم که هر چه مقدار ضریب تعیین R^2 به یک نزدیک تر باشد به معنای برازش بهتر خط رگرسیون بوده و هر چه مقدار R^2 به صفر نزدیک باشد دال بر خوب نبودن برازش است.
می توان گفت مقدار ضریب تعیین R^2 درصدی از تغییرات متغیر وابسته Y است که به وسیله متغیر پیش‌بین X تبیین و توجیه می شود.
در مساله ای اگر $R^2 = 0/9$ شود، به این معنی است که ۹۰ درصد از تغییرات متغیر وابسته Y توسط متغیر پیش‌بین X قابل توجیه است. ده درصد باقیمانده در واقع معلول عواملی است که در رگرسیون لحاظ نشده اند. بنابراین در این مثال متغیر پیش‌بین قادر به تبیین تمام تغییرات متغیر وابسته Y نمی باشد. فرض کنید Y (تولید) و X (سرمایه) است و یک رابطه خطی ساده نیز بین این دو متغیر برقرار است و فرض کنید بر اساس مشاهدات مربوط به Y و X مقدار ضریب تعیین برابر $R^2 = 0/95$ شده است. می توان گفت سرمایه ۹۵ درصد از تغییرات تولید را تبیین می کند و پنج درصد باقیمانده معلول متغیرهای دیگری است که در معادله رگرسیون لحاظ نشده اند.

قضیه ۱-۱ (رابطه تجزیه واریانس)

پراکندگی متغیر پاسخ Y را می توان به دو منبع پراکندگی تبیین شده و پراکندگی تبیین نشده تجزیه کرد یعنی:

$$SS_Y = SS_{\hat{Y}} + SS_e$$

برهان

$$SS_{\hat{Y}} = \sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum_{i=1}^{n}\left((\hat{\beta}_0 + \hat{\beta}_1 X) - \bar{Y}\right)^2$$

$$= \sum_{i=1}^{n}(\bar{Y} - \hat{\beta}_1\bar{X} + \hat{\beta}_1 X - \bar{Y})^2$$

$$=\sum_{i=1}^{n}\hat{\beta}_1^2\left(X_i-\bar{X}\right)^2=\hat{\beta}_1^2\sum_{i=1}^{n}x_i^2=\hat{\beta}_1^2 SS_X$$

از طرفی می توان نوشت:

$$\begin{aligned}\sum_{i=1}^{n}e_i^2&=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{\beta}_0-\hat{\beta}_1X\right)^2\\&=\sum_{i=1}^{n}\left(Y_i-\bar{Y}+\hat{\beta}_1\bar{X}-\hat{\beta}_1X_i\right)^2\\&=\sum_{i=1}^{n}\left(\left(Y_i-\bar{Y}\right)-\hat{\beta}_1\left(X_i-\bar{X}\right)\right)^2\\&=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_1x_i\right)^2=\sum_{i=1}^{n}y_i^2-2\hat{\beta}_1\sum_{i=1}^{n}x_iy_i+\hat{\beta}_1^2\sum_{i=1}^{n}x_i^2\\&=\sum_{i=1}^{n}y_i^2-2\hat{\beta}_1\left(\hat{\beta}_1\sum_{i=1}^{n}x_i^2\right)+\hat{\beta}_1^2\sum_{i=1}^{n}x_i^2\\&=\sum_{i=1}^{n}y_i^2-\hat{\beta}_1^2\sum_{i=1}^{n}x_i^2\end{aligned}$$

در نتیجه

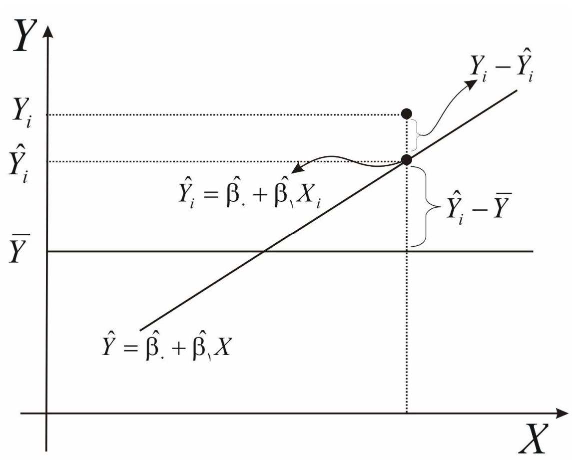
$$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}y_i^2-\sum_{i=1}^{n}\hat{y}_i^2$$

و یا

$$\sum_{i=1}^{n}y_i^2=\sum_{i=1}^{n}\hat{y}_i^2+\sum_{i=1}^{n}e_i^2$$

و حکم ثابت شده است.

در شکل زیر مقادیر Y_i، $\hat{Y}_i$، $y_i=Y_i-\bar{Y}$ ، $\hat{y}_i=\hat{Y}_i-\bar{Y}$ و $e_i=Y_i-\hat{Y}_i$ نشان داده شده اند.



شکل ۱-۳

۱–۲ رگرسیون خطی چندگانه

۱–۲–۱ برازش الگو در رگرسیون چندگانه

فرض کنید قد پسر به قد مادر و قد پدر مرتبط است. الگوی رگرسیون چندگانه ای که این رابطه را بیان می کند، بصورت زیر است:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (۴.۱)$$

که در آن Y قد پسر و X_1 قد پدر و X_2 قد مادر است. این یک الگوی رگرسیون خطی چندگانه با دو متغیر پیش‌بین X_1 و X_2 می باشد. علت آنکه این الگو را خطی می گوئیم، این است که معادله فوق یک تابع خطی از پارامترهای مجهول β_0، β_1 و β_2 می باشد. در صورتی که داده ها شامل $X_1 = X_2 = 0$ باشد، β_0 در واقع میانگین Y، وقتی که $X_1 = X_2 = 0$ است، می باشد. در غیر اینصورت β_0 هیچ گونه تعبیری ندارد. تغییر مورد انتظار در Y را به ازاء یک واحد تغییر در X_1 در صورت ثابت بودن X_2 نشان می دهد و همینطور پارامتر β_2 تغییر مورد انتظار در Y را به ازاء یک واحد تغییر در X_2 در صورت ثابت بودن X_1 نشان می دهد.
در حالت کلی متغیر پاسخ Y ممکن است با k متغیر پیش‌بین رابطه داشته باشد که در این صورت داریم:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \qquad (۵.۱)$$

و این را الگوی رگرسیون خطی چندگانه با k متغیر پیش‌بین می نامند. پارامترهای β_j، $j = 0,\ldots,k$ را ضرایب رگرسیون می نامند. پارامتر β_j، $j = 1,\ldots,k$ تغیر مورد انتظار در متغیر پاسخ Y را به ازاء یک واحد تغیر از X_j نشان می دهد هر گاه تمام متغیرهای پیش‌بین دیگر ثابت تلقی شوند.
توجه داریم که رابطه تابعی واقعی بین Y و متغیرهای پیش‌بین $X_1,\ldots,X_k$ معلوم نیست، ولی روی دامنه ای از متغیرهای پیش‌بین می توانیم الگوی رگرسیون خطی را بعنوان یک تقریب مناسب برای تابع مجهول مورد نظر پیدا کنیم. بنابراین الگوی رگرسیون خطی اغلب به عنوان الگوهای تجربی مورد استفاده قرار می گیرند. می دانیم که مقادیر ضرایب رگرسیون و واریانس خطا یعنی σ^2 معلوم نیستند و بایستی با توجه به داده های نمونه برآورد شوند.
در رگرسیون چندگانه داده های ما بصورت

$$(Y_1, X_{11},\ldots,X_{k1}), (Y_2, X_{12},\ldots,X_{k2}),\ldots,(Y_n, X_{1n},\ldots,X_{kn})$$

بوده و می خواهیم یک رابطه خطی مانند $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k$ را به این داده ها برازش دهیم.

مناسب تر است که وقتی با رگرسیون چندگانه سروکار داریم از نماد ماتریسی استفاده کنیم، لذا الگوی رگرسیون نمونه را بصورت زیر می نویسیم:

$$\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e} \qquad (۶.۱)$$

که در آن:

بردار ضرایب رگرسیون $\hat{\boldsymbol{\beta}}' = \begin{bmatrix} \hat{\beta}_۰ & \hat{\beta}_۱ & \cdots & \hat{\beta}_k \end{bmatrix}_{۱\times(k+۱)}$ بردار مشاهدات متغیر وابسته $\mathbf{Y}' = \begin{bmatrix} Y_۱ & Y_۲ & \cdots & Y_n \end{bmatrix}_{۱\times n}$

ماتریس مشاهدات متغیرهای پیش‌بین (ماتریس طرح) $\mathbf{X} = \begin{bmatrix} ۱ & X_{۱۱} & X_{۱۲} & \cdots & X_{۱k} \\ ۱ & X_{۲۱} & X_{۲۲} & \cdots & X_{۲k} \\ \vdots & \vdots & \vdots & & \vdots \\ ۱ & X_{n۱} & X_{n۲} & \cdots & X_{nk} \end{bmatrix}_{n\times(k+۱)}$

بردار مقادیر پیش بینی شده متغیر وابسته $\hat{\mathbf{Y}}' = \begin{bmatrix} \hat{Y}_۱ & \hat{Y}_۲ & \cdots & \hat{Y}_n \end{bmatrix}_{۱\times n}$ بردار مانده ها (خطا) $\mathbf{e} = \begin{bmatrix} e_۱ \\ e_۲ \\ \vdots \\ e_n \end{bmatrix}_{n\times ۱}$

بنابراین رابطه $\hat{Y}_i = \hat{\beta}_۰ + \hat{\beta}_۱ X_{i۱} + \ldots + \hat{\beta}_k X_{ik}$ ، $i = ۱,۲,\ldots,n$ را می توان بصورت $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ نوشت ضمن اینکه $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ خواهد بود. می توان نوشت:

$$\begin{bmatrix} \hat{Y}_۱ \\ \hat{Y}_۲ \\ \vdots \\ \hat{Y}_n \end{bmatrix}_{n\times ۱} = \begin{bmatrix} ۱ & X_{۱۱} & \cdots & X_{۱k} \\ ۱ & X_{۲۱} & & X_{۲k} \\ \vdots & \vdots & & \vdots \\ ۱ & X_{n۱} & \cdots & X_{nk} \end{bmatrix}_{n\times(k+۱)} \times \begin{bmatrix} \hat{\beta}_۰ \\ \hat{\beta}_۱ \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

در مورد خطاها نیز می توان نوشت:

$$\begin{bmatrix} e_۱ \\ e_۲ \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} Y_۱ \\ Y_۲ \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} \hat{Y}_۱ \\ \hat{Y}_۲ \\ \vdots \\ \hat{Y}_n \end{bmatrix}$$

حال اگر بخواهیم مجموع توان‌های دوم خطاها را مینیمم کنیم، داریم:

$$\sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = \left(\mathbf{Y}-\hat{\mathbf{Y}}\right)'\left(\mathbf{Y}-\hat{\mathbf{Y}}\right) = (\mathbf{Y}'-\hat{\mathbf{Y}}')(\mathbf{Y}-\hat{\mathbf{Y}})$$
$$= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\hat{\mathbf{Y}} - \hat{\mathbf{Y}}'\mathbf{Y} + \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

چون $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}$ یک اسکالر است لذا با ترانهاده خود یعنی $\mathbf{Y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ برابر است لذا خواهیم داشت:

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} e_i^2 = \mathbf{Y}'\mathbf{Y} - 2\mathbf{X}'\mathbf{Y}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

بنابراین

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \hat{\boldsymbol{\beta}}} = 0 - 2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0$$
$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{Y}$$

یا

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{۷.۱}$$

ملاحظه می کنیم که برای بدست آوردن برآورد ضرایب لازم است ابتدا وارون ماتریس $\mathbf{X}'\mathbf{X}$ و سپس بردار $\mathbf{X}'\mathbf{Y}$ را محاسبه نموده در هم ضرب کنیم. اگر متغیرهای پیش‌بین مستقل خطی باشند یعنی هیچ ستون ماتریس $\mathbf{X}$ یک ترکیب خطی از ستون‌های دیگر نباشد در آن صورت $(\mathbf{X}'\mathbf{X})^{-1}$ وجود خواهد داشت. ضمنا بدیهی است که اگر $n < k+1$ باشد آنگاه $\mathbf{X}'\mathbf{X}$ وارون پذیر نخواهد بود، یعنی یک شرط لازم برای وجود برآوردگرهای می‌نیموم مربعات ضرائب رگرسیون خطی این است که حجم نمونه بزرگتر از تعداد متغیرهای پیش‌بین باشد.

ماتریس $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ [5] را ماتریس برازش می نامند و این ماتریس نقش مهمی در تحلیل رگرسیون دارد. بردار مانده های (خطاهای) $\mathbf{e}$ را با استفاده از ماتریس $\mathbf{H}$ بصورت زیر می توان نوشت:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{Y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Y} - \mathbf{H}\mathbf{Y}$$

یا

$$\mathbf{e} = \left(\mathbf{I}-\mathbf{H}\right)\mathbf{Y}$$

با ضرب طرفین رابطه فوق در $\mathbf{X}'$(از سمت چپ) داریم

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$
$$= \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{Y} = 0$$

که معادل است با

[5] Hat Matrix

$$\begin{cases} \sum_{i=1}^{n} e_i = 0 \\ \sum_{i=1}^{n} e_i X_{i1} = 0 \\ \sum_{i=1}^{n} e_i X_{i2} = 0 \\ \vdots \\ \sum_{i=1}^{n} e_i X_{ik} = 0 \end{cases} \qquad (8.1)$$

در رگرسیون چند گانه، نیز همانند رگرسیون ساده، رابطه تجزیه واریانس برقرار است. که آن را در قضیه زیر بیان می کنیم.

قضیه ۱–۲

در رگرسیون چند گانه نیز رابطه تجزیه واریانس برقرار است، یعنی:

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2$$

برهان

می توان نوشت:

$$\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y}\right)^2=\sum_{i=1}^{n}\left(\left(\hat{Y}_i-\bar{Y}\right)+\left(Y_i-\hat{Y}_i\right)\right)^2$$
$$=\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2+\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2+2\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)\left(Y_i-\hat{Y}_i\right)$$

حال اگر نشان دهیم جمله سوم طرف راست تساوی فوق صفر است حکم ثابت مـی شـود. امـا مـی توان نوشت:

$$\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)\left(Y_i-\hat{Y}_i\right)=\sum_{i=1}^{n}e_i\left(\hat{Y}_i-\bar{Y}\right)$$
$$=\sum_{i=1}^{n}e_i\hat{Y}_i-\bar{Y}\sum_{i=1}^{n}e_i=0$$

و برهان کامل است.

۱–۲–۲ ضریب تعیین در رگرسیون چندگانه

مطابق تعریف می دانیم که ضریب تعیین (R^2) حاصل تقسیم پراکندگی بیان شده بـه پراکنـدگی کـل است لذا داریم،

$$R^2 = \frac{\hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2} \tag{9.1}$$

قضیه ۱–۳:

با افزایش متغیرهای پیش‌بین، ضریب تعیین دست کم کاهش نمی یابد.

برهان

فرض کنید ضریب تعیین رگرسیون Y بر k متغیر پیش‌بین $X_1,\dots,X_K$ را با R_1^2 و مجمـوع تـوان‌هـای دوم خطای آن را بـا $\sum_{i=1}^{n} e_{i1}^2$ نـشان دهـیم و ضـریب تعیـین رگرسـیون Y بـر $k+1$ متغیـر پـیش‌بـین $X_1,\dots,X_K,X_{k+1}$ را با R_2^2 و مجموع توان‌های دوم خطای آن را با $\sum_{i=1}^{n} e_{i2}^2$ نشان دهیم، باید ثابت کنـیم که: $R_1^2 \le R_2^2$.

و یا به عبارتی دیگر باید نشان دهیم که:

$$R_1^2 = 1 - \frac{\sum_{i=1}^{n} e_{i1}^2}{\sum_{i=1}^{n} y_i^2} < 1 - \frac{\sum_{i=1}^{n} e_{i2}^2}{\sum_{i=1}^{n} y_i^2}$$

و یا به طور معادل:

$$\sum_{i=1}^{n} e_{i1}^2 \ge \sum_{i=1}^{n} e_{i2}^2$$

اما این رابطه به وضوح برقرار است، زیرا اگر تعریف کنیم:

$$\Phi(\beta_0, \beta_1, \dots, \beta_k, \beta_{k+1}) = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_k X_{ik} - \beta_{k+1} X_{ik+1}\right)^2$$

در این صورت:

$$\sum_{i=1}^{n} e_{i2}^2 = Min_{\beta_0,\beta_1,\dots,\beta_k,\beta_{k+1}} \Phi(\beta_0, \beta_1, \dots, \beta_k, \beta_{k+1}) \le Min_{\beta_0,\dots,\beta_k} \Phi(\beta_0, \beta_1, \dots, \beta_k, 0) = \sum_{i=1}^{n} e_{i1}^2$$

۱-۲-۳ هم خطی و اثرات آن در رگرسیون چندگانه

غالبا تفسیر و استفاده از الگوی رگرسیون چندگانه به برآورد تک تک ضرایب رگرسیون بستگی دارد. به عنوان مثال برخی از استنباط هایی که در رگرسیون چندگانه اعمال می شود، عبارتست از:

۱- تشخیص اثرات نسبی متغیرهای پیش‌بین
۲- پیش بینی و برآورد
۳- انتخاب یک مجموعه مناسب از متغیرها برای مدل

اگر هیچ رابطه خطی میان متغیرهای پیش‌بین وجود نداشته باشد، این متغیرها متعامد[۶] نامیده می شوند. هنگامی که متغیرهای پیش‌بین متعامد باشند، استنباط هایی همانند استنباط های فوق، به سادگی می توانند اتخاذ شوند. اما متاسفانه در بسیاری از موارد در عمل، متغیرهای پیش‌بین متعامد نیستند. گاهی اوقات، عدم وجود تعامد در متغیرهای پیش‌بین مساله بسیار حادی نیست. اما اگر در یک الگوی رگرسیون متغیرهای پیش‌بین به صورت کامل و یا تقریبا کامل، وابسته خطی باشند، در این صورت ممکن است استنباط بر اساس این الگوی رگرسیون تاحد بسیاری گمراه کننده و پراشتباه باشد. در این هنگام می گوییم که میان متغیرها در الگوی رگرسیون همخطی[۷] وجود دارد.

اگر یکی از متغیرهای پیش‌بین یک تابع دقیق خطی از یک یا چند متغیر پیش‌بین دیگر باشد گوئیم رگرسیون دارای **همخطی کامل**[۸] است. همخطی ناقص وقتی اتفاق می افتد که یکی از متغیرهای پیش‌بین بطور تقریبی یک تابع خطی از یک یا چند متغیر پیش‌بین دیگر باشد. به عبارت دیگر همخطی کامل زمانی رخ می دهد که (لااقل به ازای یک j) $R_j^2=1$ و همخطی ناقص زمانی پیش می آید که (لااقل به ازای یک j) $R_j^2 \approx 1$ ، که R_j^2 ضریب تعیین رگرسیون خطی X_j بر سایر متغیرهای پیش‌بین است. بعنوان مثال فرض کنید که استحکام یک مفتول را به دو متغیر قطر مقطع و محیط مقطع وابسته بدانیم. در این حالت واضح است که متغیر پیش‌بین محیط مقطع تابعی از متغیر پیش‌بین قطر مقطع است و لذا همخطی کامل وجود دارد و این همخطی بنیادی است یعنی مربوط به نمونه نمی باشد. البته توجه داریم که همخطی بنیادی عموما در عمل رخ نمی دهد زیرا همواره با خطای اندازه گیری مواجه هستیم. هم خطی کامل در عمل هنگامی رخ می دهد که کاربر به اشتباه در الگوی رگرسیون یکی از متغیرها را تابعی خطی از متغیر دیگر قرار دهد و یا وقتی تعداد متغیرهای پیش‌بین بیشتر از تعداد مشاهدات باشد نیز مساله همخطی کامل به وجود می آید. با این قبیل الگوها معمولاً در علوم رفتاری و پزشکی مواجه می شویم، زیرا که ممکن است فقط تعداد کمی از واحدهای نمونه وجود داشته باشد و اطلاعات برای تعداد زیادی از

[6] Orthogonal
[7] Collinearity
[8] Exact collinearity

متغیرهای پیش‌بین روی هر آزمودنی گردآوری شود. در این قبیل موارد راه حل این است که بعضی از متغیرهای پیش‌بین را حذف کنیم.

همچنین منابع اصلی همخطی ناقص را می توان به سه دسته تقسیم نمود:

(۱) انتخاب روش گردآوری داده ها

(۲) گذاشتن قیدهایی روی الگو یا در جامعه

(۳) شناسایی الگو

هنگامی که تحلیل گر تنها از یک زیرفضای ناحیه متغیرهای پیش‌بین نمونه گیری می کند روش گردآوری داده ها ممکن است مساله همخطی را به وجود آورد. همخطی حاصل از روش نمونه گیری در الگو یا در جامعه ای که از آن نمونه گرفته شده است ذاتی نیست.
همچنین قیدهایی روی الگو یا در جامعه ای که از آن نمونه گیری شده می تواند همخطی را به وجود آورد. برای مثال فرض کنید در می‌خواهیم اثر میزان شکر (X_1) و میزان آبلیمو (X_2) را روی مقدار شیرینی یک لیوان شربت آبلیمو بررسی کنیم در این صورت همخطی صرف نظر از روش نمونه گیری انتخاب شده وجود خواهد داشت. این قیدها عموما در مسائلی نظیر تولید یک محصول و یا فرآیندهای شیمیایی رخ می دهند، هنگامی که متغیرهای پیش‌بین مولفه‌های تشکیل دهنده متغیر پاسخ هستند.

انتخاب الگو نیز سبب همخطی می شود. از مبحث الگوی رگرسیون چندجمله ای می دانیم که وقتی مرتبه چندجمله ای زیاد می شود ماتریس $\mathbf{X'X}$ دچار بدشرطی می شود، بدین معنی که محاسبه وارون ماتریس نادقیق خواهد بود و در برآورد پارامترها خطای قابل ملاحظه ای به وجود می آیند. علاوه بر این اگر دامنه X کوچک باشد افزودن جمله X^2 می تواند همخطی معنی داری را حاصل کند. زیرا ممکن است در یک بازه کوچک، رابطه میان X و X^2 تقریباً خطی باشد.
اغلب وقتی با این وضعیت مواجه می شویم که دو یا چند متغیر پیش‌بین تقریباً وابسته خطی باشند و نگاهداشتن تمام این متغیرها ممکن است موجب همخطی شود. در این موارد از نقطه نظر همخطی انتخاب زیرمجموعه ای از متغیرها معمولاً ترجیح داده می شود.
گاهی اوقات ملاحظه آماره F برای معنی داری رگرسیون همراه با آماره t (یا F جزیی) وجود همخطی را نشان می دهد. اگر آماره F کلی معنی دار باشد(آزمون معنی داری کلی رگرسیون) ولی آماره های t همگی معنی دار نباشند، می تواند دال بر وجود همخطی ناقص باشد. متاسفانه بسیاری از مجموعه داده هایی که همخطی معنی داری دارند این رفتار را نشان نمی دهد و لذا سودمندی این معیار همخطی سوال برانگیز است.

اثرات همخطی

همخطی ناقص باعث می شود که دترمینان ماتریس $\mathbf{X'X}$ بسیار کوچک شود و در نتیجه ماتریس $\mathbf{(X'X)^{-1}}$، دارای اعضایی باشند که از لحاظ قدرمطلق بسیار بزرگ هستند و چون $Var(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbf{X'X}\right)^{-1}$ لذا واریانس های برآوردهای $\beta_1, \beta_2, \ldots, \beta_k$ بسیار بزرگ می شود. یعنی $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ که به طریق کمترین توان‌های دوم از فرمول $\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1}X'Y}$ بدست می آیند برآوردهای دقیقی از ضرایب $\beta_1, \beta_2, \ldots, \beta_k$ نمی باشند. به عبارت دیگر دقت برآورد ضرایب رگرسیون بسیار کم می شود. علاوه بر این $\hat{\boldsymbol{\beta}}$ ناپایدار است. یعنی با تغییر کوچکی در مقادیر متغیرهای پیش‌بین، مقادیر موجود در $\hat{\boldsymbol{\beta}}$ تغییرات زیادی پیدا می کنند و در نتیجه محاسبه $\hat{\boldsymbol{\beta}}$ توسط دو کامپیوتر متفاوت (با دو روش روند کردن متفاوت) ممکن است جواب های کاملا متفاوتی برای $\hat{\boldsymbol{\beta}}$ به دست دهند.

همچنین در صورت وجود هم‌خطی کامل ماتریس $\mathbf{(X'X)^{-1}}$ وارون پذیر نبوده (زیرا این ماتریس تمام رتبه نمی شود.) و در نتیجه برای برآورد ضرایب جواب یکتا وجود نخواهد داشت.

عامل تورم واریانس

یکی از معیارهای رایج و پراستفاده برای تشخیص هم‌خطی در مدل رگرسیون، عامل تورم واریانس است. این معیار به صورت $VIF_j = \frac{1}{1-R_j^2}$ تعریف می شود که در آن R_j^2 ضریب تعیین رگرسیون X_j بر سایر متغیرهای پیش‌بین است.

برای برآوردهای ضرایب مدل رگرسیون $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$، ثابت می شود که:

$$Var\left(\hat{\beta}_j\right) = \frac{\sigma^2}{nS_j^2} VIF_j \ , j \geq 1 \qquad (۱۰.۱)$$

که در آن $\hat{\beta}_j$ برآورد کمترین مربعات β_j و σ^2 واریانس جمله خطا و S_j^2 واریانس نمونه متغیر پیش‌بین jام است.

واضح است که(به ازای مقادیر ثابتی برای S_j^2) هرچه قدر VIF_j بیشتر باشد، مقدار واریانس $\hat{\beta}_j$ بیشتر خواهد شد به عبارت دیگر اگر $R_j^2 \to 1$ در این صورت $VIF_j \to \infty$ و لذا $Var\left(\hat{\beta}_j\right) \to \infty$.

یک مقدار بزرگ R_j^2 این گونه تفسیر می شود که X_j رابطه خطی تقریبا کامل با سایر متغیرهای پیش‌بین دارد و معمولا $VIF_j \geq 10$ به عنوان قابل اغماض نبودن مساله هم‌خطی تفسیر می شود.