درس‌نامه پنجم

بنابراین $\hat{\beta}$ دارای یک توزیع گوسی است . تخمین σ_ε^2 را می توانیم از روی رابطه زیر داشته باشیم :

$$\hat{\sigma_\varepsilon^2} = \frac{1}{N-P-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

که می توان نشان داد که توزیع $\hat{\sigma_\varepsilon^2}$ یک توزیع chi-squared می باشد یعنی :

$$(N-P-1)\hat{\sigma^2} \sim \sigma^2\chi^2_{N-P-1}$$

همچنین میتوان نشان داد که $\hat{\beta}$ و $\hat{\sigma^2}$ از دیدگاه آماری مستقل از یکدیگر می باشند.

آزمون فرض

اگر بخواهیم $\beta_j = 0$ را تست نماییم مقدار زیر را محاسبه می کنیم :

$$t_{n-p-1} = \frac{\hat{\beta}_j}{\hat{\sigma}_\varepsilon\sqrt{v_j}}$$

مقدار فوق دارای توزیع t با درجه آزادی $n-p-1$ می باشد که در آن v_j ، j-امین مقدار قطری $(X^TX)^{-1}$ می‌باشد. اگر بحواهیم به طور گروهی آزمون

$$\beta_j = \beta_k = \cdots = 0$$

را مورد ارزیابی قرار دهیم آنگاه یکبار مدل را با مقادیر صفر فوق یاد می گیریم و سپس با تمامی مقدار ها. در حالت اول اگر میزان مجموع مربعات باقیمانده را RSS_0 و در حالت دوم RSS_1 نمایش دهیم آنگاه :

$$F = \frac{(RSS_0 - RSS_1)/p_1 - p_0}{RSS_1/n-p_1-1}$$

دارای توزیع F با پارامترهای $(p_1-p_0, n-p_1-1)$ می باشد.

متعامد سازی

با توجه به انکه تجزیه $X = QR$ از طریق گرم_اشمیت و دیگر روش ها قابل محاسبه است که در آن $Q^TQ = I$ و R یک ماتریس بالا مثلثی است خواهیم داشت :

$$\hat{\beta} = (R^TR)^{-1}R^TQ^Ty = R^{-1}Q^Ty$$

اگر ماتریس Q را به صورت

$$Q = [q_0 q_1 \cdots q_p]$$

و ماتریس R^{-1} را به صورت

$$R = \begin{bmatrix} \frac{r_{00}}{r_{10}} & & & \\ & r_{11} & & \\ & & \ddots & \\ & & & \\ & & & \end{bmatrix}$$

نمایش دهیم ، رابطه فوق نشان می دهد:

$$\begin{aligned} r_{00}\hat{\beta}_0 &= q_0^Ty \\ r_{10}\hat{\beta}_0 + r_{11}\hat{\beta}_1 &= q_1^Ty \\ r_{20}\hat{\beta}_0 + r_{21}\hat{\beta}_1 + r_{22}\hat{\beta}_2 &= q_2^Ty \\ &\vdots \end{aligned}$$

رابطه فوق را می توان به این صورت دید که رابطه اول در حقیقت یک رگرسیون ساده و یک بعدی برای داده های $x_0 = (1, \cdots, 1)$ و $y = (y_1, \cdots, y_n)$ می باشد . $\hat{\beta}_0$ از روی این مسئله به دست می آید. سپس برای $\hat{\beta}_1$ بایستی مسئله رگرسیون ساده دوم را حل نماییم و الی آخر.

نکات مهم :

۱. اکر یکی از بعد ها داده های دسته بند داشت می بایست آنرا کسر نماییم .به این منظور داده های دو تایی مانند داشتن حساب و نداشتن حساب را یا با $(1, 0)$ و یا $(1, -1)$ کد می نماییم . اگر تعداد دسته ها بیشتر بود می توانیم چند ستون اضافه کنیم و به ازای وجود هر کدام از آن دسته ها عدد ۱ و در غیر این صورت صفر قرار می دهیم .

۲. ورودی های جدید را می توانیم بر پایه ورودی های اولیه تولید کنیم و بدین ترتیب پیچیدگی را بالا ببریم، در این شرایط هنوز به خاطر وجود ضرایب خطی با یک رگرسیون خطی مواجه هستیم. مثلا:

$$\hat{y} = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_1x_2 + \beta_4x_1^2 + \beta_5x_1^3$$

مشکلات احتمالی

۱. **رابطه واقعی ممکن است غیر خطی باشد.**

اگر رابطه واقعی غیر خطی باشد آنگاه

$$e = y - \hat{y} = f(x) - < x, \hat{\beta} > + \varepsilon$$

بنابراین e و x مستقل نمی شوند و در این صورت اگر رابطه e و x را بکشیم یک رابطه را مشاهده می کنیم. اگر x چند بعدی بود آنگاه می توانیم از $\hat{y} =< x, \hat{\beta} >$ استفاده نماییم و رابطه آن را با e بررسی کنیم . به هر حال در حالت ایده آل e فقط با ε رابطه دارد .

۲. **واریانس خطا ممکن است ثابت نشود.**

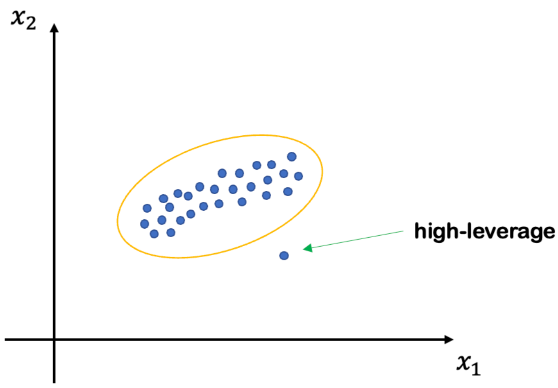
ممکن است فرض آنکه ε_i ها مستقل و یا نا همبسته می باشند درست نباشد. در چنین شرایطی استنتاج‌ها می توانند اشتباه باشند. مثلا در داده های سری زمانی ممکن است نقاط نزدیک دارای خطاهای اندازه گیری همبسته باشند. برای بررسی این قضیه می توان پس از رگرسیون خطا را برحسب زمان ترسیم نمود و اگر رابطه ای در خطا دیده شد آن را یافت . البته در مسائل غیر سری زمان یافتن انها مشکل است.

۳. **وجود نمونه‌های ناهنجار**

نمونه‌های ناهنجار، نمونه‌هایی هستند که از مدل نیامده‌اند و احتمالا به خطا وارد داده شده‌اند. برای از بین بردن آن می‌توانیم مقدار $\frac{e_i}{\hat{\mathrm{SE}}(e)}$ را برای تمامی iــ‌ها بدست اوریم. مقایسه مقدار فوق با یک آستانه می‌تواند نشان دهد که مقدار ناهنجار است یا خیر.

۴. **نقاط High-leverage**

به نقاطی اطلاق می‌گردد که ورودی x نامعقولی نسبت به بقیه نقاط داشته باشند.



این نقاط تاثیر زیادی بر یادگیری می‌گذارند و از بین بردن آن‌ها می‌تواند تاثیر زیادی بر کیفیت داشته باشد. یکی از راه‌های یافتن آن‌ها محاسبه آماره زیر است:

$$h_i = \frac{۱}{n} + \frac{(x_i - \bar{x})^۲}{\sum_{i'=۱}^{n} (x_{i'} - \bar{x})^۲}$$

می‌دانیم $\frac{1}{n} \leqslant h_i \leqslant 1$ بوده و $\frac{\sum h_i}{n} = \frac{p+1}{n}$. بنابراین اگر h_i با مقدار خیلی بزرگتر از $\frac{p+1}{n}$ یافت شد، می‌توان آن را حذف نمود.

۵. Collinearity

ممکن است تعدادی از ستون‌های داده با هم رابطه داشته باشند. یعنی همبستگی خطی بین آن‌ها دیده شود. در این شرایط نمی‌توان تاثیر این دو ستون را از یکدیگر تفکیک نمود. اگر به ماتریس همبستگی نگاه کنیم، مقادیری که مقدار بزرگ دارند می‌توانند بیانگر چنین ستون‌هایی باشند. اما اگر بخواهیم تاثیر چندگانه را بیابیم از VIF (variance inflation factor) استفاده می‌نماییم.

$$\mathrm{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{x_j|x_j}}$$

یعنی از بقیه پارامترها x_j را تخمین بزنیم.