اقتصاد سنجی

فصل چهارم: گسترش مدل رگرسیون ساده

گروه اقتصاد
دانشگاه الزهرا

حمید کردبچه

Hamidkurdbacheh@yahoo.com

هدف فصل:

این فصل به مرور موضوعات در رگرسیون دو متغیره می پردازد که در فصول قبل مورد توجه قرار نگرفته اند.

مفاهیم و نتایج حاصل در این فصل قابل تعمیم به رگرسیون چند متغیره نیز است.

مطالب فصل

- ✓پیش بینی
- ✓رگرسیون از مبداء مختصات
- ✓ مقیاس داده و اندازه تخمینها
- ✓ ضرایب بتا

منابع فصل:

گجراتی فصل ٦ ، هیل فصل ٤ ، وولدریج فصل ٦

پیش بینی (*Prediction*)

مدل رگرسیون مفروض $\mathrm{Y} = \beta_0 + \beta_1 X + u$

یکی از کاربردهای مهم این تخمین پیش بینی تغییرات یا مقادیر آینده متغیر تابع است.
میخواهیم ارزش Y را بر اساس مقدار معین X_0 پیش بینی کنیم (Y_0)

$$Y_0 = \beta_0 + \beta_1 X_0 + u_0$$

فروض:

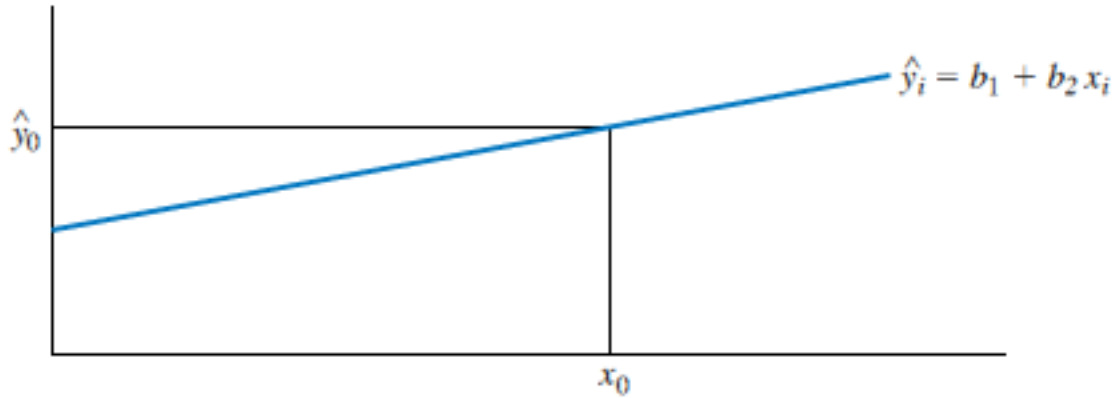
$$E(Y_0) = \beta_0 + \beta_1 X_0$$
$$E(u_0) = 0$$
$$E(u_0)^2 = \sigma^2$$
$$E(u_0, u_i) = 0$$

پیش بيني كننده حداقل مربعات (Least square predictor)

با قرار دادن مقدار مورد نظر Xدر برازش حداقل مربعات رگرسیون خواهیم داشت

$$\hat{Y}_0 = b_0 + b_1 X_0$$

مفهوم این عبارت چیست؟



مثال: از تخمین مدل دستمزد و تحصیلات ملاحظه نمودیم

اگر X=۱۸ باشد $\hat{Y}$ چقدر است.

$$\hat{Y}_i = 14695 + 60.21 X_i$$

$$\hat{Y}_0 = 14695 + 60.21(18) = 2792.1$$

خطای پیش بینی (Forecast error)

پیش بینی انجام شده چقدر خوب است؟

پاسخ: محاسبه خطای پیش بینی که هم ارز مفهوم باقیمانده حداقل مربعات است

$$f = Y_0 - \hat{Y}_0 = (\beta_0 + \beta_1 X_0 + u_0) - (b_0 + b_1 X_0)$$

مقدار مطلوب خطای پیش بینی چقدر است؟

امید ریاضی آن صفر است

$$E(f) = \beta_0 + \beta_1 X_0 + E(u_0) \quad E(b_0) + E(b_1) X_0$$
$$= 0$$

معنای این رابطه چیست؟

بهترین پیش بینی کننده خطی یا BLUP (Best linear unbiased predictor) Y_0 از $\hat{Y}_0$ است

پیش بینی فاصله ای (Prediction Interval)

سوال: آیا می توان با استفاده از پیش بینی نقطه ای $\hat{Y}_0$ فاصله اطمینانی برای مقدار واقعی Y پیدا نمود.

می توان نشان داد که

$$Var(f) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right)$$

در عمل تخمین σ^2 استفاده می شود یعنی

$$\hat{Var}(f) = \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right)$$

مفروض به این که خطای پیش بینی فروض کلاسیک را تامین می کنند، در این صورت مقدار استاندارد شده آن توزیع نرمال استاندارد خواهد داشت. یعنی

$$\frac{f - E(f)}{s_e(f)} = \frac{Y_0 - \hat{Y}_0}{s(f)} \sim Z(0,1)$$

در صورت استفاده از تخمین زیگما ۲ برای

$$\frac{f - E(f)}{se(f)} = \frac{Y_0 - \hat{Y}_0}{se(f)} \sim t(0,1)$$

بر این اساس چگونه می توان فاصله اطمینان Y_0 را ساخت؟

با محاسبه مقدار t بر اساس سطح معنا داری مورد نظر و درجه آزادی (n-2) و با استفاده از مقدار استاندارد شده خطای پیش بینی می توان نوشت:

$$\hat{Y}_0 \pm t_c se(f)$$

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} se(Y_0) < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} se(Y_0)$$

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right) < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right)$$

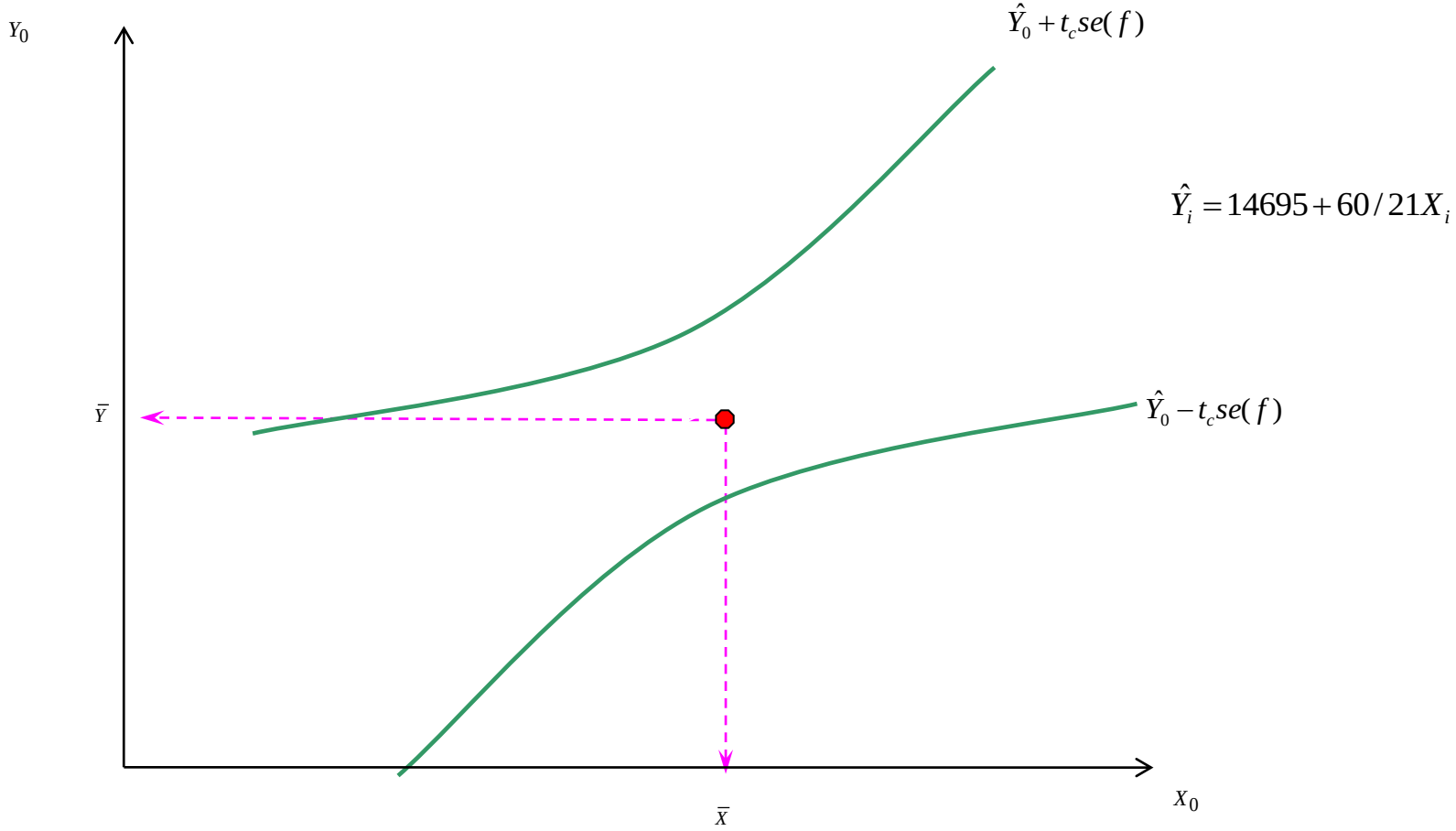
$$\hat{\beta}_0 + \hat{\beta}_1 X_0 - t_{\frac{\alpha}{2}} s_e(\hat{Y}_0) < \beta_0 + \beta_1 X_0 \leq \hat{\beta}_0 + \hat{\beta}_1 X_0 + t_{\frac{\alpha}{2}} s_e(\hat{Y}_0)$$

مثال :

$$\hat{Y}_0 = 14695 + 60.21(18) = 2792.1$$

$$Var(\hat{Y}_0) = 146168\left(1 + \frac{1}{10} + \frac{18 - 13.4685}{4506.82}\right) = 161373$$

$$[2792.1 - 1.96(401.7) \le Y_f \le 2792.1 + 1.96(401.7)]$$



دقت پیش بینی

چگونه می توان دقت پیش بینی را افزایش داد؟

با فاصله اطمینان کوچکتر

$$\hat{Y}_0 - t_{\frac{\alpha}{2}} \hat{\sigma}^2 \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}} < Y_0 < \hat{Y}_0 + t_{\frac{\alpha}{2}} \hat{\sigma}^2 \sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}}$$

کاهش انحراف معیار مهمترین عامل برای کاهش فاصله اطمینان است.

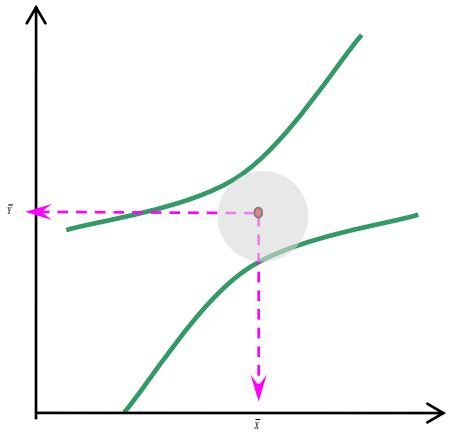
چگونه فاصله اطمینان را کاهش دهیم؟

افزایش حجم نمونه

افزایش پراکندگی مشاهدات از طریق نمونه گیری صحیح

پیش بینی بر اساس مقادیری از متغیر توضیحی

که به میانگین آن نزدیکتر است



مدل رگرسیون از مبداء مختصات RTO
(Regression Through the Origin)

مفهوم عرض از مبدا در مدل رگرسیون

$$Y_i = \beta X_i + u_i \quad \text{if} \quad X_i = 0 \implies Y_i = 0$$

اهمیت وجود عرض از مبدا در مدل رگرسیون

- بی معنا بودن متغیر وابسته صفر
- ایجاد اطمینان از این که فرض میانگین صفر جزء خطا خلل به مدل وارد نمی کند
- هزینه کمتر خطای تصریح

آیا مقدار عددی عرض از مبدا معنای خاص دارد

وجود عرض از مبدا همیشه مهم است اما معنای ان به ندرت

کاربرد RTO

چه موقع RTO مناسب است

نفی عرض از مبدا به دلیل ماهیت مدل و داده

تخمین درآمد دائمی

توابع تولید خاص

مدل های رگرسیون خاص

تبدیل مدل با عرض از مبدا OLS برای حل مشکلات ناهمسانی یا خود همبستگی

رگرسیون با استفاده از دادههای استاندارد شده

آیا RTO یک حالت خاص از CLR است؟

تخمین زن OLS

$$e_i = Y_i - \hat{Y}_i$$

$$\sum e_i^2 = \sum (Y_i - \hat{Y}_i)^2$$
$$= \sum (Y_i - \tilde{b}_1 X_i)^2$$

$$\frac{\partial \sum e_i^{۲}}{\partial \tilde{b}_۱} = ۲ \sum (Y_i \quad \tilde{b}_۱ X_i) X_i = ۰$$

$$\Rightarrow \tilde{b}_۱ = \frac{\sum Y_i X_i}{\sum X_۱^{۲}}$$

- با مقایسه b_1 و $\tilde{b}_1$ ملاحظه می شود هرگاه $\bar{X} = 0$ باشد $b_1 = \tilde{b}_1$
- دستور stata: reg y x , noconstant

$\tilde{b}_1$ بدون تورش است تنها اگر $\beta_0 = 0$ باشد.

$$\tilde{b}_1 = \frac{\sum X_i(\beta_1 X_i - u_i)}{\sum X_i^2}$$

$$= \beta_1 - \frac{\sum X_i u_i}{\sum X_i^2}$$

$$\Rightarrow E(\tilde{b}_1) = \beta_1 - \frac{E\left(\sum X_i u_i\right)}{\sum X_i^2} \qquad \Rightarrow E(\tilde{b}_1) = \beta_1$$

- اگر $\beta_0 \neq 0$ باشد در آنصورت $\tilde{b}_1$ با تورش است (اثبات کنید)

واریانس $\tilde{b}_1$

$$Var(\tilde{b}_1) = \frac{\sigma^2}{\sum X_i^2}$$

واریانس σ_u^2 در مدل بدون عرض از مبداء کوچکتر است.

$$\sigma^2 = \frac{\sum e_i^2}{n-1}$$

اگر $\beta_0 = 0$ باشد دیگر در مدل با عرض از مبدا $E(u_i) = 0$ نخواهد بود و بر عکس.

از معادله نرمال با عرض از مبداء داریم.

$$\sum (Y_i - \hat{Y}_i) = 0$$

$$\sum (Y_i - b_0 - b_1 x_i) = 0$$

$$\sum e_i = 0$$

اگر این رابطه درست باشد در اینصورت عدم وجود b_0 در مدل به معنای غیر صفر بودن $\sum e_i$ در مدل بدون عرض از مبداء است.

مقیاس متغیرها و اندازه تخمین ها (Scaling and units of measurement)

با تغییر مقیاس واحدها مثلاً از ریال به ۱۰۰۰ ریال یا از تن به کیلو تخمین ها چگونه تغییر می کنند؟
اثر تغییر مقیاس x_1بر ضرایب b_0 و b_1
می دانیم :

$$b_0 = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

فرض کنید:

$$X_i = \delta X_i$$

$$\Rightarrow b_1^* = b_1 / \delta$$

$$b_0^* = b_0$$

اثر تغییر مقیاس بر $\text{var}(b_1)$ ، t و R^2

$$\text{var}(b_1) = \frac{\sum e_i^2 / n-2}{\sum x_i^2}$$

$$\text{var}(b_1) = \frac{1}{\delta^2} \ \text{var}(b_1)$$

$$s_e(b_1^*) = \frac{1}{\delta} \ s_e(b_1)$$

$$t = \frac{b_1}{s_e(b_1)} \Rightarrow t^* = t$$

$$R^{2*} = \frac{SSE}{SST} = \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

$$\Rightarrow R^{2*} = R^2$$

مثال: رابطه بین سرمایه گذاری خصوصی داخلی ناخالص (GPDI) و تولید ناخالص ملی (GNP)

(ارقام میلیون دلار)

$$GPDI_i = \underset{(76.3)}{-37.002} + \underset{(0.054)}{0.17395} GNP_i$$

$$r^2 = 0.56$$

حال اگر مقیاس متغیر مستقل بر حسب بیلیون دلار (۱۰۰۰ برابر شود)

$$GPDI_i = \underset{(76.3)}{-37.00152} + \underset{0.000054}{0.00017395} GNP$$

$$r^2 = 0.56$$

مشاهده می شود با ۱۰۰۰ برابر شدن مقیاس متغیر مستقل ضریب آن در مدل تخمینی تقسیم بر ۱۰۰۰ شده است.

تفسیر یکسان نتایج

ملاحظه می شود تغییر مقیاس تنها یک عمل زیباسازی نتایج است.زیرا تغییری در نتایج حاصل از رگرسیون یا تفیسر نتایج ایجاد نمیکند.

مدل اول: اگر GNP یک واحد (یک میلیون دلار) افزایش یابد سرمایه گذاری ناخالص داخلی ۰/۱۷٤ واحد افزایش خواهد یافت.

مدل دوم: اگر GNP یک واحد (یعنی یک بیلیون دلار) افزایش یابد سرمایه گذاری ۰/۰۰۰۱۷٤ واحد افزایش خواهد یافت.

آیا دو نتیجه فوق متفاوت اند؟

$$\frac{174}{1000}(1000000) = \frac{174}{1000000}(1000000000)$$

تغییر مقیاس متغیر وابسته

اثر تغییر مقیاس بر ضرایب

$$Y_i^* = \theta Y_i$$

$$\Rightarrow \; b_1^* = \theta b_1$$

$$b_0 = \theta b_1$$

اثر تغییر مقیاس بر سایر نتایج

$$\text{var}(b_0) = \frac{\sum X_i^2}{n\sum x_i^2} \; \frac{\sum e_i^2}{n-2}$$

$$\text{var}(b_0^*) = \frac{\sum x_i^2}{n\sum x_i^2} . \frac{\sum (y_i^* - \hat{y}_i^*)^2}{n-2}$$

$$= \frac{\sum X_i^2}{n\sum x_i^2} . \theta^2 \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

نتایج:

$$\mathrm{var}(b_0^*) = \theta^2 \,\mathrm{var}(b_0)$$

$$s_e(b_0^*) = \theta \; s_e(b_0)$$

با استدلال مشابه

$$s_e(b_1^*) = \theta \; s_e(b_1)$$

$$R^{2*} = R$$

$$t = t^*$$

مثال: در مثال قبل فرض کنید مقیاس متغیر مستقل میلیون دلار و مقیاس متغیر وابسته بیلیون دلار باشد.

$$GPDI_i = \underset{(76300)}{37002} + \underset{(54)}{173/95}$$

$$r^2 = 0/56$$

ملاحظه می شود که با ۱۰۰۰ برابر شدن مقیاس متغیر وابسته ضرایب تخمینی ۱۰۰۰ برابر می شود.

آیا تفسیر نتایج متفاوت خواهد بود؟

اگر مقیاس کلیه متغیرها بیلیون دلار باشد.

$$GPDI = \underset{(76300)}{37002} + \underset{(0/054)}{0.17395}GNP_i$$

$$r^2 = 0.56$$

از تفسیر نتایج در کلیه مثالها چه نتیجه ای می توان گرفت؟

ضرایب بتا beta coefficients

از بخش قبل ملاحظه نمودیم که مقدار عددی ضرایب به مقیاس متغیرها وابسته است.

رگرسیون با استفاده از مقادیر استاندارد شده متغیرها از چنین تاثیری جلوگیری میکند

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

با استاندارد نمودن متغیرها

$$Y_i^* = \frac{Y_i - \bar{Y}}{Se(Y_i)} \qquad X_i^* = \frac{X_i - \bar{X}}{Se(X_i)}$$

$$Y_i^* = \beta_0 + \beta_1 X_i^* + u_i^*$$

می توان نشان داد که در این مدل رگرسیون عرض از مبدا صفر است. چرا؟

$$Y_i^* = \beta_1^* X_i^* + u_i^*$$

β_1^*= ضریب بتا

تفسیر این ضریب چیست؟

مثال:

$$\widehat{\text{GPDI}}_t = -1026.498 + 0.3016\ \text{GDP}_t$$
$$\text{se} = \quad (257.5874) \quad (0.0399) \qquad r^2 = 0.8872$$

$$\widehat{\text{GPDI}}_t^{*} = 0.9387\ \text{GDP}_t^{*}$$
$$\text{se} = (0.1149)$$

مزیت استفاده از داده های استاندارد شده چیست؟

آیا رابطه ای بین ضرایب دو مدل وجود دارد؟

$$\beta_1^{*} = \beta_1 \frac{se(X)}{se(Y)}$$

دستور stata : reg y x, b