به نام خداوند بخشنده مهربان

روش‌های آماری و اقتصادسنجی

در تحلیل و مدل‌سازی داده‌های حمل و نقلی

محمد مهدی بشارتی

besharati@iut.ac.ir

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

مقدمه

رگرسیون خطی ساده

رگرسیون خطی چند متغیره

پیش‌فرض‌های رگرسیون خطی

معیارهای نیکویی برازش

تبدیلات (دستکاری متغیرهای مدل)

مقدمه

- رابطه میان سرعت تردد خودروها در یک معبر با سرعت مجاز در آن معبر،
- رابطه میان تعداد سفرهای تولیدشده در یک ناحیه ترافیکی با تعداد خانوارهای ساکن در آن ناحیه
- **رگرسیون خطی:** یک نوع خاص از انواع رابطه‌های موجود بین متغیرها
- منظور از مدل‌های رگرسیون خطی، مدل‌هایی است که **از لحاظ پارامتر،** خطی باشند و یا بتوان آن‌ها را به مدل‌هایی خطی از لحاظ پارامتر تبدیل کرد.
- این مدل‌ها از لحاظ متغیرها ممکن است خطی باشند یا نباشند.
- اگر متغیرها به‌صورت غیرخطی باشند، معمولاً می‌توان به کمک تبدیل‌هایی، آن‌ها را به‌صورت متغیرهای خطی وارد مدل کرد.

مقدمه

مدرس: محمدمهدی بشارتی

- روابط مورد مطالعه در این فصل، از جنس روابط **همبستگی آماری** است؛ نه روابط **علت و معلولی**.
- **یعنی چه؟**
- **مثال:** تعداد خانوارهای ساکن در یک ناحیه تنها ممکن است یک **پیش‌بینی کننده** مناسب برای تعداد سفرهای تولیدشده از آن ناحیه باشد.
- بدین معنا که اگر بخواهید تعداد سفرهای تولیدشده در یک ناحیه را برآورد کنید و تنها در مورد تعداد خانوارهای ساکن در آن ناحیه، اطلاعاتی در دست داشته باشید؛ این مدل می‌تواند به شما کمک کند متوسط تعداد سفرهای تولیدشده را با استفاده از تعداد خانوارهای ساکن در آن ناحیه پیش‌بینی کنید.
- اما این بدین معنا نیست که شما می‌توانید لزوماً با دست‌کاری مستقیم یک متغیر (مثلاً تعداد خانوارهای ساکن یک ناحیه)، مقدار متغیر دیگر (مثلاً تعداد سفرهای تولیـدشده از آن ناحیه) را تغییر دهید.
- درصورتی‌که تحلیل‌گر بخواهد با دست‌کاری یک متغیر، مقدار متغیر دیگر را پیش‌بینی کند؛ به یک رابطه علت و معلولی میان این دو متغیر نیاز دارد.
- **مثال از رابطه علت و معلولی؟**
- اندازه حرکت (مومنتم) یک خودرو مستقیماً به جرم آن بستگی دارد و با افزایش جرم، بدون شک اندازه حرکت نیز افزایش می‌یابد.

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

مقدمه

رگرسیون خطی ساده

رگرسیون خطی چند متغیره

پیش‌فرض‌های رگرسیون خطی

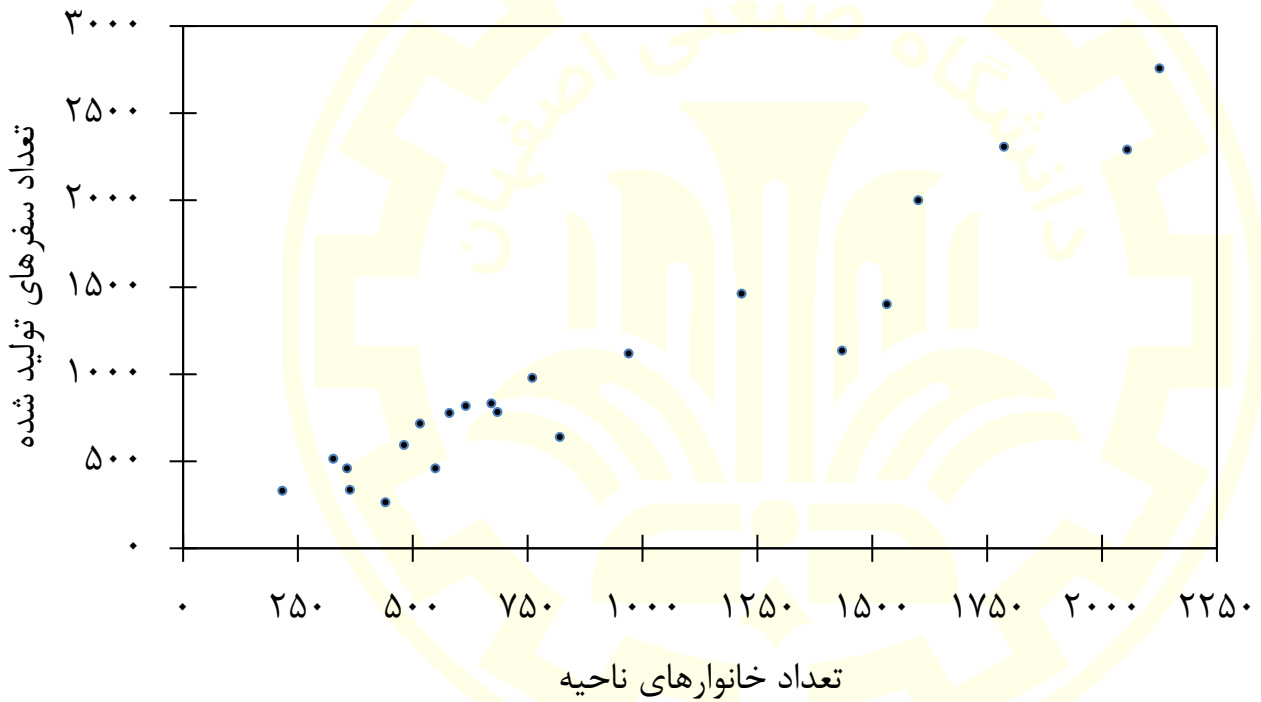
معیارهای نیکویی برازش

تبدیلات (دستکاری متغیرهای مدل)

رگرسیون خطی ساده (بررسی خطی‌بودن رابطه)

❑ بررسی خطی بودن رابطه از طریق ترسیم نمودار پراکنشی

پیش از ساخت مدل رگرسیون خطی، لازم است ابتدا به کمک رسم نمودار پراکنشی و انجام آزمون‌های آماری از خطی بودن رابطه میان متغیر(های) توصیفی و متغیر وابسته اطمینان حاصل گردد.



رگرسیون خطی ساده (بررسی خطی‌بودن رابطه)

مدرس: محمدمهدی بشارتی

❑ ضریب همبستگی پیرسون

برای بررسی میزان و نوع همبستگی خطی بین دو متغیر، به‌صورت کمّی از آزمون‌های ضریب همبستگی استفاده می‌شود.

فرض کنید تعدادی داده زوجی به‌صورت (X_i, Y_i) که $i=1,2,...,n$ را در اختیار داریم. ضریب همبستگی پیرسون، r، معیاری از شدتِ خطی بودن رابطه دو متغیر به دست می‌دهد. نام دیگر این معیار، ضریب همبستگی نمونه است که با رابطه زیر تعریف می‌شود.

$$r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

باید به خاطر داشت که $r=0$، به معنای عدم وجود رابطه **خطی** میان دو متغیر است.

امـا این بدان معنـا نیست که لزوماً هیچ رابطه‌ای بین دو متغیر وجود ندارد.

بلکه ممکن است دو متغیر موردنظر با یکدیگر رابطه غیرخطی داشته باشند.

ترسیم نمودار پراکنشی می‌تواند به تحلیل‌گر برای شناسایی روابط غیرخطی نیز کمک کند.

به‌علاوه، تأکید می‌شود که ضریب همبستگی، یک شاخص آماری است و بزرگ بودنِ مقدار ضریب همبستگی میان دو متغیر به معنای وجود رابطه علت و معلولی بین آن دو متغیر نیست.

رگرسیون خطی ساده (بررسی خطی‌بودن رابطه)

❑ ضریب همبستگی پیرسون

بررسی معنادار بودن مقدار ضریب همبستگی به کمک آزمون فرضیه

می‌خواهیم بدانیم آیا مقدار ضریب همبستگی (در جامعه) تفاوت معناداری با صفر دارد یا خیر. $H_0: \rho=0$ و $H_a: \rho\neq0$

اگر یک نمونه تصادفی با اندازه n از جامعه انتخاب شود، خطای استاندارد ضریب همبستگی از رابطه زیر قابل‌محاسبه است

$$S.E.(r)=\sqrt{1-r^2/n-2}$$

آماره آزمون به صورت زیر تعریف می‌شود و از توزیع t با n -۲ درجه آزادی پیروی می‌کند.

$$t_{H_0}=\frac{r}{\sqrt{1-r^2/n-2}}=r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} \quad \sim t_{n-2}$$

بنابراین، درصورتی‌که قدر مطلق مقدار آماره آزمون ($|t_{H_0}|$)، بیشتر از مقدار بحرانی ($t_{1-\alpha/2,\, n-2}$) باشد؛ آنگاه فرضیه صفر رد می‌شود.

پس از رد شدن فرضیه صفر در سطح معناداری α، با (α-۱)٪ اطمینان می‌توان گفت میان دو متغیر موردنظر همبستگی خطی معناداری وجود دارد.

رگرسیون خطی ساده (برآورد مدل رگرسیون)

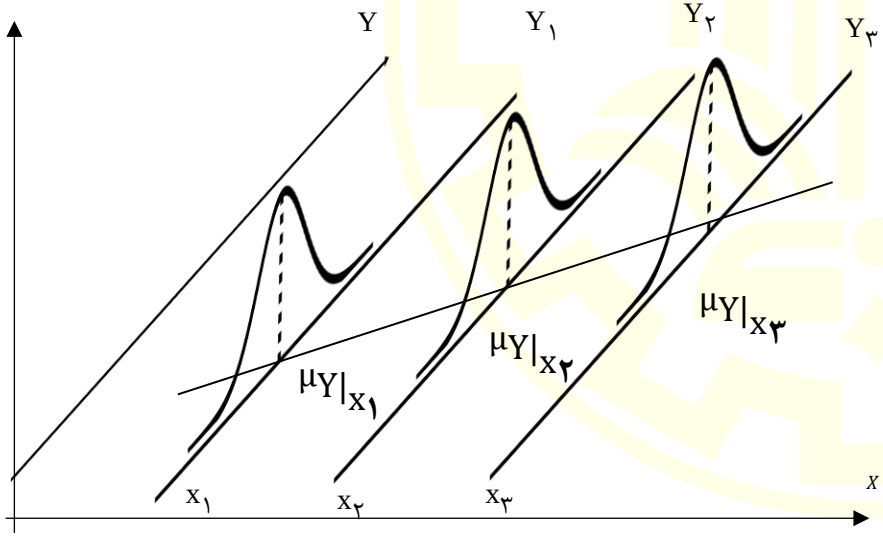
❑ برآورد مدل رگرسیون

در هر بار نمونه‌گیری از دنیای واقعی، ممکن است به ازای یک مقدار x معین، مقادیر y متفاوتی مشاهده شود.

به‌عبارت‌دیگر، y_i در هر زوج داده (x_i , y_i)، یک مقدار از متغیر تصادفی Y_i است.

بنابراین می‌توان گفت $Y|x_i$ یک متغیر تصادفی Y_i به ازای مقدار معین x_i می‌باشد.

نمودار: توزیع‌های **فرضی** مقادیر Y_i به ازای هر مقدارِ معین x_i



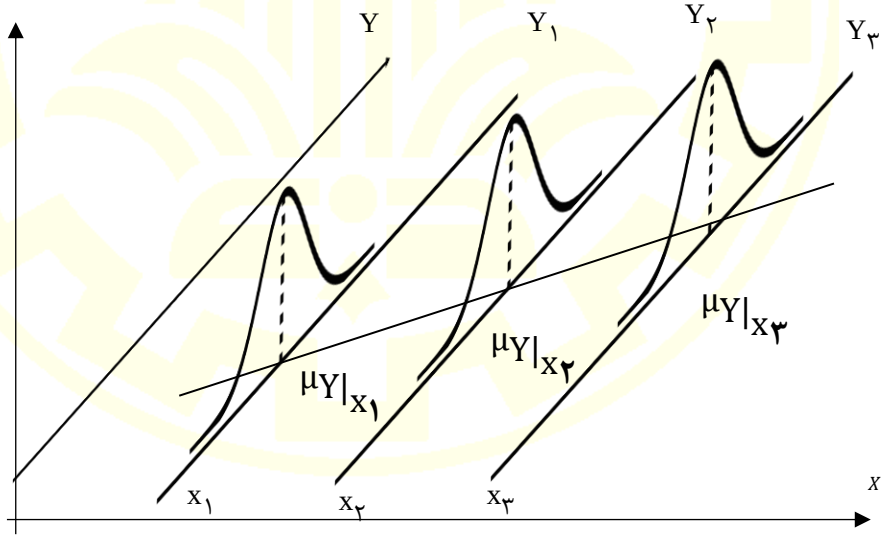
رگرسیون خطی ساده (برآورد مدل رگرسیون)

❑ برآورد مدل رگرسیون

درواقع، یکی از پیش‌فرض‌های تحلیل رگرسیون خطی ساده آن است که توزیع مقادیر Y_i به ازای هر مقدار معین x_i، از یک توزیع نرمال با میانگین $\mu_{Y|x}=a+bx_i$ پیروی می‌کند که واریانس این توزیع‌ها برای همه مقادیر x، مقدار ثابتی (σ^2) است.

با داشتن تابع این توزیع، «رگرسیون Y بر روی X» عبارت است از تعیین امید ریاضی متغیر تصادفی Y به شرطی که متغیر X مقدار x را اختیار کند؛

$$E(Y|X=x)=\mu_{Y|x}=a+bx_i$$



رگرسیون خطی ساده (برآورد مدل رگرسیون)

❑ برآورد مدل رگرسیون

درواقع، یکی از پیش‌فرض‌های تحلیل رگرسیون خطی ساده آن است که توزیع مقادیر Y_i به ازای هر مقدار معین x_i، از یک توزیع نرمال با میانگین $\mu_{Y|x}=a+bx_i$ پیروی می‌کند که واریانس این توزیع‌ها برای همه مقادیر x، مقدار ثابتی (σ^2) است.

با داشتن تابع این توزیع، «رگرسیون Y بر روی X» عبارت است از تعیین امید ریاضی متغیر تصادفی Y به شرطی که متغیر X مقدار x را اختیار کند؛

$$E(Y|X=x)=\mu_{Y|x}=a+bx_i$$

به عبارت‌دیگر، پیش‌بینی می‌شود که اگر X مقدار x را اختیار کند، آنگاه مقدار متغیر Y، به‌طور متوسط برابر با $E(Y|X=x)$ خواهد شد. بنـابراین، می‌توان گفت منحنی رگرسیـون، خطی است که مقـادیر امید ریاضی شرطی متغیر وابسته را به ازای مقادیر مختلف متغیر(های) پیش‌بین به هم متصل می‌کند.

یک حالت ساده و کاربردی از رگرسیون، حالتی است که رگرسیون Y بر روی X به‌صورت تابعی **خطی** از ضرایب منحنی باشد؛

$$E(Y|X=x)=\mu_{Y|x}=a+bx$$

رگرسیون خطی ساده (برآورد مدل رگرسیون)

مدرس: محمدمهدی بشارتی

a و b ، مقادیر ثابتی از خط رگرسیون واقعی (پارامتر جامعه) هستند؛ و به دست آوردن مقادیر واقعی آن‌ها غیرممکن است.

در روند برآورد مدل رگرسیون، با جمع‌آوری یک نمونه از زوج مقادیر (x_i , y_i)، برآوردهای نقطه‌ای از a و b به دست می‌آوریم.

این برآوردها را به ترتیب، با $\hat{a}$ و $\hat{b}$ نمایش می‌دهند.

به‌طور خلاصه منظور از «ساخت مدل رگرسیون خطی»، برآورد کردن پارامترهای مدل (در اینجا a و b) می‌باشد؛

$$\bar{y}_X = \hat{a} + \hat{b}x$$

پس از جمع‌آوری نمونه به‌صورت زوج مقادیر (x_i , y_i)، مشاهده می‌کنیم که همه زوج‌ها بر روی یک خط مستقیم قرار نمی‌گیرند.

بی‌نهایت خط مستقیم را می‌توان از میان این داده‌ها عبور داد.

سؤال: کدام‌یک از این خطوط می‌تواند بهترین برآورد را از خط رگرسیون جامعه ارائه کند؟

یعنی به دنبال بهترین برآوردگرها برای $\hat{a}$ و $\hat{b}$ هستیم.

یکی از رایج‌ترین روش‌ها برای شناسایی بهترین خط از میان خط‌های کاندید (یعنی پیدا کردن بهترین برآوردگرها برای $\hat{a}$ و $\hat{b}$)، روش «**حداقل مربعات معمولی خطا**» می‌باشد.

Ordinary Least Squares (OLS)

برای تشریح این روش، ابتدا لازم است به بررسی مفهوم خطای مدل پرداخته شود.

رگرسیون خطی ساده (برآورد مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ برآورد مدل رگرسیون

معادله

$$E(Y|X=x)=\mu_{Y|x}=a+bx$$

فقط آن بخشی از تغییرات Y را توضیح می‌دهد که متاثر از X است (**بخش معین** مدل).

حال باید تاثیر سایر عوامل را که تصادفی و غیرقابل کنترل است را نیز وارد مدل کنیم.

متغیر تصادفی Y_i را به‌صورت زیر در نظر می‌گیریم:

$$Y_i= a+bx_i+E_i$$

E_i، متغیر تصادفی پراکندگی مقادیر Y_i حول $\mu_{Y|x}$ می‌باشد و به آن **خطای تصادفی** می‌گویند (معمولاً به صورت u_i نمایش داده می‌شود).

پیش‌فرض: E_i یک متغیر تصادفی است که به ازای هریک از مقادیر x، دارای توزیع نرمال با میانگین صفر و واریانس σ^2 است.

بدین ترتیب، هر یک از مشاهدات در جامعه (x_i , y_i)، در رابطه زیر صدق خواهد کرد.

$$y_i= a+bx_i+\varepsilon_i$$

از سوی دیگر، برای یک نمونه تصادفی، هر یک از مشاهدات (x_i , y_i) در رابطه زیر صدق خواهند کرد.

$$y_i= \hat{a}+\hat{b}x_i+e_i$$

e_i، مقدار باقیمانده نامیده می‌شود.

رگرسیون خطی ساده (برآورد مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ برآورد مدل رگرسیون

به‌عبارت‌دیگر، اگر مقـادیر y مشاهده‌شده در نمونه را به شکل y_i و مقادیر y پیش‌بینی‌شده توسط مدل به ازای یک x را به‌صورت $\bar{y}_{x_i}$ نشان دهیم؛ واضح است که مقادیر پیش‌بینی‌شده توسط مدل، ممکن است با مقادیر واقعی مشاهده‌شده در نمونه تفاوت داشته باشد (چرا؟). بنابراین، رابطه میان مقادیر $\bar{y}_{x_i}$ و y_i را به‌صورت رابطه زیر نمایش می‌دهیم.

$$y_i = \bar{y}_{x_i} + e_i$$

$\bar{y}_{x_i}$، برآوردی از $E(Y|x_i)$ و e_i تفاوت میان مقادیر y مشـاهده‌شده در نمونه و برآوردشده توسط مدل است.

بنابراین، تغییرات Y به دو دسته تقسیم می‌شود؛

۱- تغییراتی که متأثر از X است (و آن را قابل کنترل و غیرتصادفی می‌دانیم). به آن **جزء معین** یا **غیرتصادفی** می‌گویند.

۲- تغییراتی که ناشی از سایر عوامل تصادفی، پنهان و یا غیرقابل کنترل است. به آن **جزء تصادفی**، **جزء اخلال**، **جزء اختلال** یا **جمله خطا** می‌گویند.

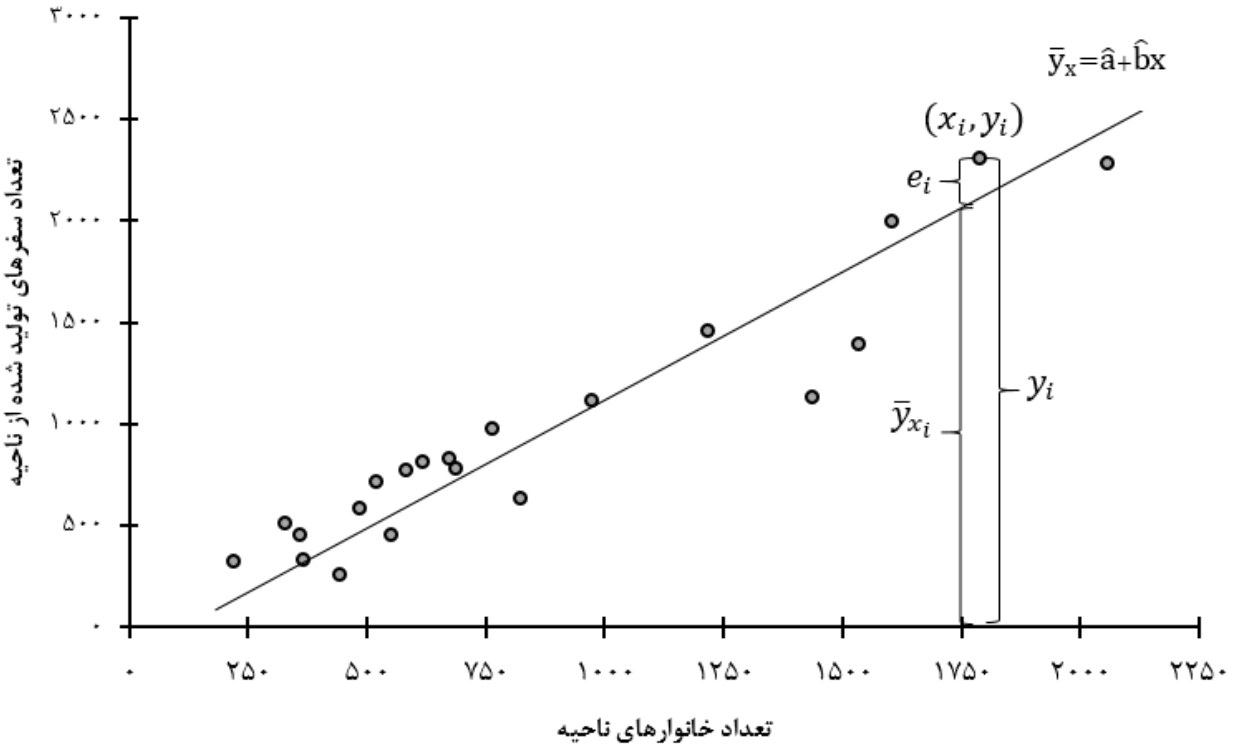
رگرسیون خطی ساده (برآورد مدل رگرسیون)

❑ برآورد مدل رگرسیون

به‌عبارت‌دیگر، می‌توان گفت e_i انحراف قـائمِ مقدار $\bar{y}_{x_i}$ از Y_i است.

یکی از معیارهای مناسب برای برآورد پارامترهای مدل رگرسیون آن است که خطاها را حداقل کنیم.

یعنی خط رگرسیون به گونه‌ای از میان نقاط مشاهدات عبور کند که کمترین خطا ایجاد شود.



براساس این ایده می‌توانیم برآوردگرهایی را برای پارامترهای مدل بسازیم که به آن‌ها،

برآوردگرهای حداقل مربعات معمولی ($\widehat{\theta}_{OLS}$)

می‌گویند.

رگرسیون خطی ساده (برآورد مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ روش حداقل مجموع مربعات خطا

فرض کنید تعدادی داده زوجی به‌صورت (x_i, y_i) داریم که $i=1,2,...,n$ و

می‌خواهیم با استفاده از این مشاهدات، برآوردی از خط رگرسیون جامعه ارائه کنیم.

در روش حداقل مجموع مربعات خطا، پارامترهای خط رگرسیون به‌گونه‌ای برآورد می‌شوند که «مجموع مربعات باقیمانده‌ها» از خط برآوردشده (رابطه زیر)، حداقل شود.

$$SSE=\sum_{i=1}^{n}(e_i)^2=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$$

به‌عبارت‌دیگر، خط رگرسیون حداقل مربعات، خطی است که کمترین مقدار مجموع مربعات فواصل عمودی بین مقادیر مشاهده‌شده در نمونه و خط رگرسیون را دارا باشد.

رگرسیون خطی ساده (برآورد مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ روش حداقل مجموع مربعات خطا

بدین ترتیب، اگر از رابطه مجموع مربعات باقیمانده‌ها نسبت به $\hat{a}$ و $\hat{b}$ مشتق جزئی گرفته و این مشتق‌ها برابر با صفر قرار دهیم؛ پس از حل معـادلات نسبت به $\hat{a}$ و $\hat{b}$، روابط زیر برای برآورد شیب (b) و عرض از مبدأ (a) خط رگرسیون به دست خواهد آمد؛

$$\hat{b}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

$$\hat{a}=\bar{Y}-\hat{b}\bar{X}$$

- ✓ این روابط در واقع **برآوردگرهای پارامترهای مدل رگرسیون خطی ساده** است.
- ✓ در ادبیات علم آمار با عنوان «بهترین برآوردگر خطی ناریب (Best Linear Unbiased Estimator (B.L.U.E.))» شناخته می‌شود.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ معناداری مدل رگرسیون خطی

خط رگرسیونی که با روش حداقل مربعات خطا برآورد شده باشد، خطی خواهد بود که کمترین مقدار مجموع مربعات فواصل عمودی نسبت به مقادیر مشاهده‌شده در نمونه را داراست.

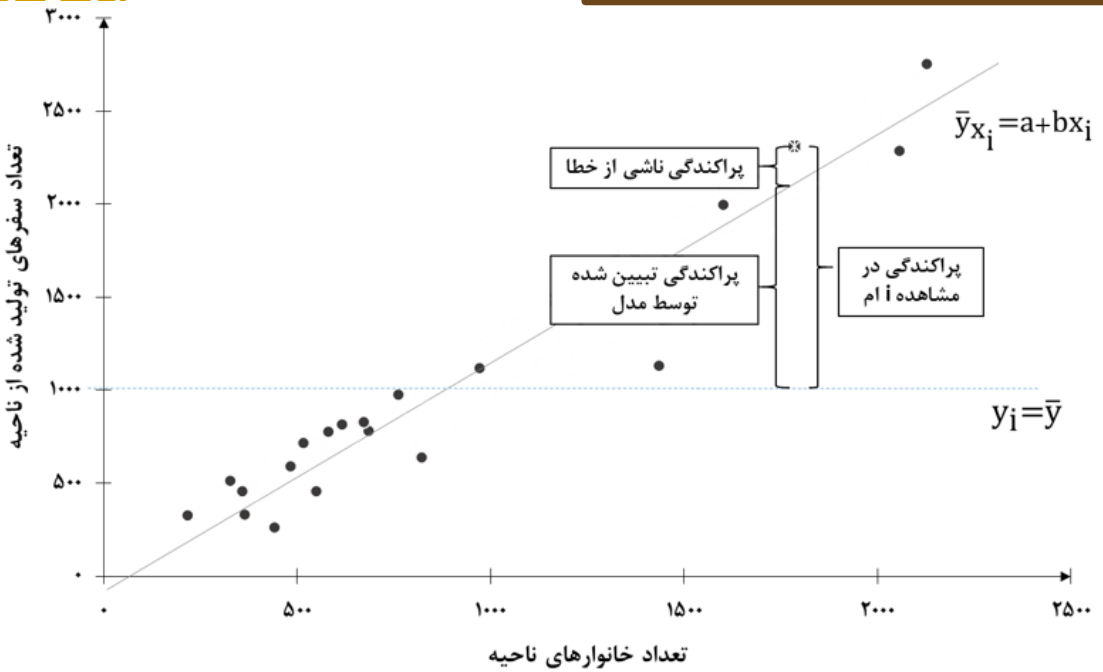
سوال: آیا این بدان معناست که این برازش، لزوماً از نظر آماری معنادار است؟

- خیر.
- برای مثال ممکن است اصولاً رابطه میان متغیر وابسته و متغیر(های) توصیفی، یک رابطه غیرخطی بوده و مدل رگرسیون خطی، مدل مناسبی برای نمایش دادن این رابطه نباشد.
- بنابراین، در گام بعد، لازم است به این سؤال پاسخ دهیم که میزان نیکویی برازش خط برآورده‌شده بر داده‌ها چقدر است؟
- برای این منظور، از آزمون‌ها و شاخص‌های آماری استفاده می‌شود.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری مدل رگرسیون خطی



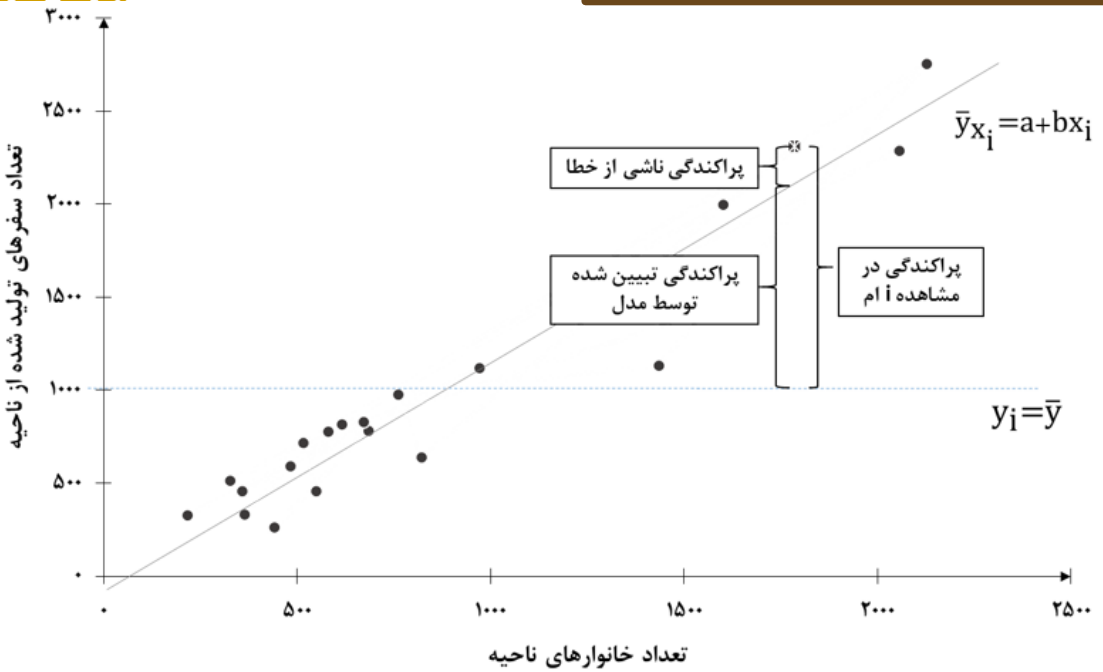
دو خط مستقیم را در نظر بگیرید.

خط اول، خطی است که به‌صورت افقی بوده و در آن مقدار متغیر وابسته به ازای تمام مقادیر متغیر توصیفی، برابر با میانگین مقادیر y مشاهده‌شده است ($y_i=\bar{y}=a$).

خط دوم همان مدل رگرسیون $y_i=\bar{y}_{x_i}=a+bx_i$ است؛ که به ازای هر مقدار متغیر توصیفی، مقدار متغیر وابسته را بر اساس مدل ساخته‌شده، پیش‌بینی می‌کند.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

❑ آزمون معناداری مدل رگرسیون خطی



به‌طور خلاصه، در این مرحله می‌خواهیم به این سؤال پاسخ دهیم که آیا تفاوت معناداری میان توانایی مدل اول (خط میانگین) و مدل دوم (خط رگرسیون) در پیش‌بینی مقادیر y وجود دارد یا خیر.

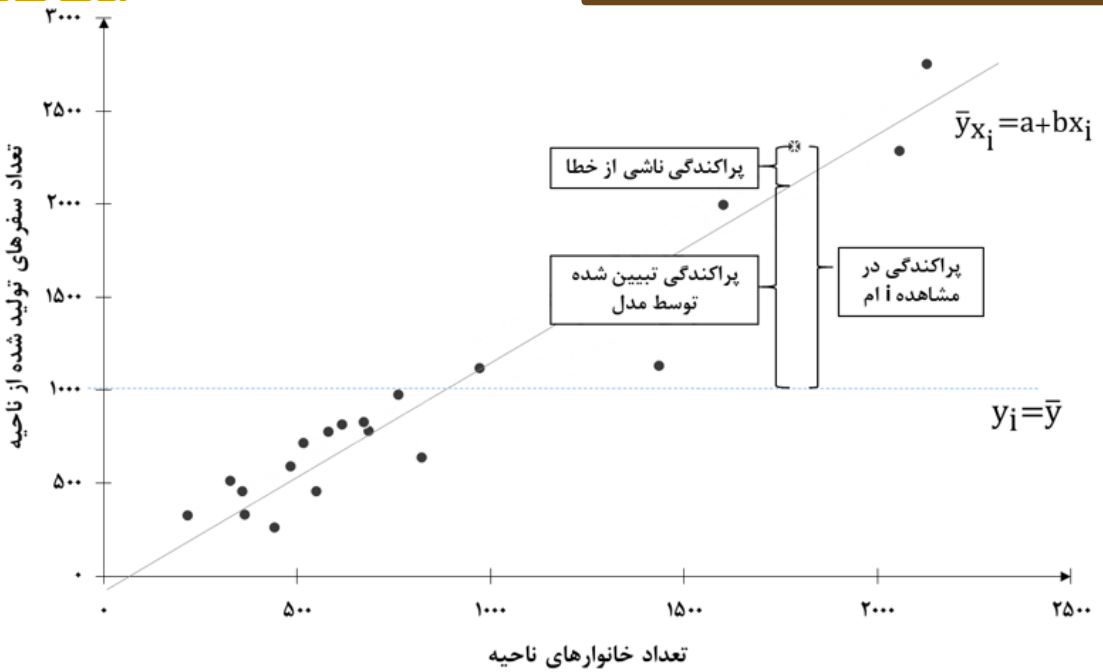
برای پاسخ به این سؤال، از **آزمون تحلیل واریانس** استفاده کرده و فرضیه صفر را به این صورت تعریف می‌کنیم که مدل رگرسیون، معنادار نیست (تفاوت معناداری با مدل ساده میانگین ندارد).

به‌عبارت‌دیگر، فرضیه صفر در آزمون معناداری رگرسیون خطی ساده به‌صورت $H_0: b = 0$ تبیین می‌شود.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری مدل رگرسیون خطی



پراکندگی هر مشاهده y_i نسبت به میانگین داده‌ها را می‌توان به دو جزء تقسیم کرد.

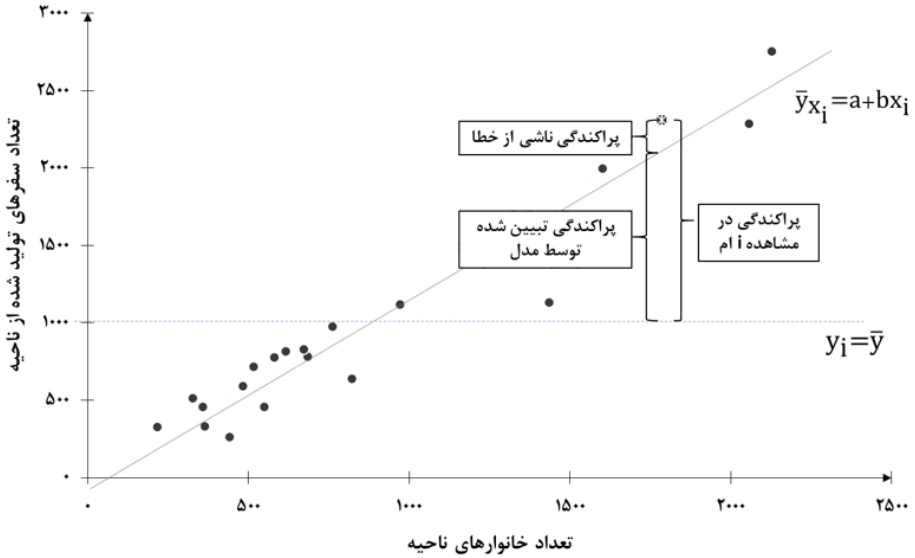
جزء اول که به آن پراکندگی تبیین‌شده توسط مدل رگرسیون گفته می‌شود، شامل تفاوت بین مقدار «y پیش‌بینی‌شده توسط **مدل رگرسیون**» با مقدار «y پیش‌بینی‌شده توسط **مدل میانگین**» است.

جزء دوم که به آن پراکندگی ناشی از خطا و یا پراکندگی تبیین‌نشده می‌گویند، شامل تفاوت بین مقدار «y پیش‌بینی‌شده توسط **مدل رگرسیون**» با مقدار «y **مشاهده‌شده**» می‌باشد.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری مدل رگرسیون خطی



به‌عبارت‌دیگر، مجموع مربعات انحراف‌های مشاهده‌شده در مقادیر Y از میانگین داده‌ها (SS_T) را به‌صورت زیر به دو بخش، شامل مجموع مربعات انحراف‌های تبیین نشده (SS_E) و مجموع مربعات انحرافات تبیین‌شده توسط مدل (SS_M) تقسیم می‌کنیم

$$SS_T=SS_E+SS_M=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{y})^2$$

هرچه مجموع مربعات انحرافات تبیین‌شده توسط مدل ($(\hat{Y}_i-\bar{Y})^2$) سهم بیشتری از مجموع مربعات کل انحرافات را به خود اختصاص دهد؛ بدین معناست که مدل رگرسیون، توانایی بیشتری در پیش‌بینی پراکندگی مقادیر Y به ازای X دارد. به‌عبارت‌دیگر، رابطه رگرسیونِ معناداری میان X و Y وجود دارد.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری مدل رگرسیون خطی

منظور از «معنادار بودن» مدل رگرسیون آن است که توانایی مدل رگرسیون در پیش‌بینی مقادیر Y به ازای مقادیر X به‌طور معناداری بیشتر از یک مدل ساده است که به ازای تمامی مقادیر X، مقدار میانگینِ Y را به‌عنوان مقدار متغیر وابسته، پیش‌بینی می‌کند.

- درصورتی‌که مقادیر خطای جامعه (ε_i ها) توزیع نرمال داشته باشند (پیش‌فرض)؛ آنگاه، آماره میانگین مجموع مربعات پراکندگی خطا $MS_E = SS_E/(n-2)$ برآوردی از واریانس مشترک داده‌ها (σ^2) خواهد بود.
- درصورتی‌که فرضیه صفر صحیح بوده و رابطه رگرسیون، ***معنادار نباشد***؛ آنگاه میانگین مجموع مربعات پراکندگی مدل $MS_M = SS_M/(1)$ نیز برآورد دیگری از σ^2 خواهد بود.

بر این اساس، درصورتی‌که فرضیه صفر صحیح باشد و مدل رگرسیون خطی تفاوت معناداری با یک مدل ساده میانگین ***نداشته باشد***؛ آنگاه، آماره $F = MS_M/MS_E$ از توزیع $F_{1,n-2}$ پیروی خواهد کرد.

بدین ترتیب، اگر مقدار آماره F، بزرگ‌تر از مقدار بحرانی $F_{\alpha,1,n-2}$ باشد، فرضیه عدم معناداری مدل رگرسیون (در سطح معناداری α) **رد می‌شود**.

به‌علاوه، هرچه پراکندگی بیان‌شده توسط مدل بیشتر باشد، مقدار آماره F بزرگ‌تر بوده و این نشانگر موفق‌تر بودن مدل رگرسیون خطی در پیش‌بینی مقادیر متغیر وابسته است.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری مدل رگرسیون خطی

- درصورتی‌که مقادیر خطای جامعه (ε_i ها) توزیع نرمال داشته باشند (پیش‌فرض)؛ آنگاه، آماره میانگین مجموع مربعات پراکندگی خطا $MS_E = SS_E/(n-2)$ برآوردی از واریانس مشترک داده‌ها (σ^2) خواهد بود.
- درصورتی‌که فرضیه صفر صحیح بوده و رابطه رگرسیون، ***معنادار نباشد***؛ آنگاه میانگین مجموع مربعات پراکندگی مدل $MS_M = SS_M/(1)$ نیز برآورد دیگری از σ^2 خواهد بود.

بر این اساس، درصورتی‌که فرضیه صفر صحیح باشد و مدل رگرسیون خطی تفاوت معناداری با یک مدل ساده میانگین ***نداشته باشد***؛ آنگاه، آماره $F = MS_M/MS_E$ از توزیع $F_{1,n-2}$ پیروی خواهد کرد.

بدین ترتیب، اگر مقدار آماره F، بزرگ‌تر از مقدار بحرانی $F_{\alpha,1,n-2}$ باشد، فرضیه عدم معناداری مدل رگرسیون (در سطح معناداری α) **رد می‌شود**.

جدول ۱۰-۱ **جدول نتیجه آزمون معناداری مدل رگرسیون خطی**

منبع تغییرات	مجموع مربعات	درجه آزادی	میانگین مربعات	آماره F
پراکندگی تبیین‌شده (توسط مدل)	SS_M	۱	MS_M	$\frac{MS_M}{MS_E}$
پراکندگی تبیین‌نشده (باقیمانده‌ها)	SS_E	n-۲	MS_E	
مجموع	SS_T	n-۱		

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری پارامترهای مدل رگرسیون خطی

- پارامترهای برآوردشده در مدل رگرسیون ($\hat{a}$ و $\hat{b}$)، یک برآورد نقطه‌ای از مقادیر مجهول پارامترهای جامعه (a و b) هستند.
- در صورت تکرار نمونه‌گیری، احتمالاً مقادیر جدیدی به‌عنوان برآوردهای نقطه‌ای برای پارامترهای جامعه به دست خواهد آمد.
- پس لازم است با استفاده از این برآوردها، فرضیه‌هایی را در مورد مقادیر واقعی پارامترهای جامعه آزمود.
- **سوال:** آیا شیب و عرض از مبدأ واقعی جامعه، تفاوت معناداری با مقدار صفر دارند؟
- با استفاده از آزمون فرضیه و با داشتن برآوردهایی از شیب و عرض از مبدأ که بر اساس نمونه به‌دست‌آمده، به این سوال پاسخ می‌دهیم.
- فرضیه صفر در این آزمون‌ها، بیان می‌کند که مقدار برآوردشده برای شیب و عرض از مبدأ، **تفاوت معناداری با صفر ندارند** و بنابراین، دلیلی برای گنجاندن پارامتر موردنظر درون مدل رگرسیون وجود ندارد.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری پارامترهای مدل رگرسیون خطی

یادآوری: با توجه به رابطه SS_E، می‌توان نشان داد که امید ریاضی SS_E برابر با $(n-۲)\sigma^۲$ می‌باشد. بنابراین، $SS_E/(n-۲)$ یک برآوردکننده ناریب برای $\sigma^۲$ خواهد بود (رابطه زیر).

$$\hat{\sigma}^۲ = \frac{\sum_{i=۲}^{n}(y_i - \bar{y}_i)^۲}{n-۲}$$

ازآنجاکه $\sigma^۲$ مجهول است؛ به‌جای آن از برآورد ناریبش ($\hat{\sigma}^۲$) در محاسبات استفاده می‌کنیم.

$\hat{\sigma}^۲$ انحراف معیار برآورد است و نشان می‌دهد به طور متوسط، Yهای واقع چقدر از Yهای برآورد شده انحراف دارند.

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری پارامترهای مدل رگرسیون خطی

پیش فرض: ε_iها دارای توزیع نرمال مستقل با میانگین صفر و واریانس مشترک σ^2 می‌باشد.

به‌طور معادل: Y_iها (یا $E(Y|x_i)$) متغیر تصادفی دارای توزیع نرمال مستقل با میانگین $a+bx_i$ و واریانس مشترک σ^2

آنگاه برآوردکننده‌های پارامترهای مدل رگرسیون نیز دارای توزیع نرمال خواهند بود که میانگین و واریانس این پارامترها از روابط ارائه‌شده در جدول زیر قابل‌محاسبه خواهند بود:

جدول ۲-۱۰ **میانگین و واریانس پارامترهای مدل رگرسیون خطی ساده**

برآوردگر	میانگین	واریانس	خطای استاندارد برآورد
$\hat{a}$	a	$\hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right)$	$\hat{\sigma}_{\hat{a}}=\hat{\sigma}\sqrt{\frac{1}{n}+\frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$
$\hat{b}$	b	$\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$	$\hat{\sigma}_{\hat{b}}=\frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}}$

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری پارامترهای مدل رگرسیون خطی

برای بررسی معناداری مقدار عرض از مبدأ، فرضیه صفر $a = 0$ و فرضیه مقابل $a \neq 0$.

آماره زیر درصورتی‌که فرضیه صفر صحیح باشد $(a = 0)$، از توزیع t_{n-p} پیروی می‌کند (p تعداد پارامترهای مدل).

$$t_a = \frac{\hat{a} - a}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}}$$

بنابراین، در مدل رگرسیون خطی ساده اگر شرط $|t_a| > t_{1-\alpha/2,\, n-2}$ برقرار باشد؛ آنگاه فرضیه صفر $(a = 0)$ رد می‌شود. بدین ترتیب می‌توان گفت، مقدار عرض از مبدأ واقعی جامعه تفاوت معناداری با صفر دارد.

- به‌علاوه، **فاصله اطمینانِ (α-۱) درصدی** برای عرض از مبدأ خط رگرسیون (a) به‌صورت زیر قابل‌محاسبه خواهد بود.

$$\hat{a} \pm t_{(1-\alpha/2,\, n-2)}\hat{\sigma}_{\hat{a}}$$

رگرسیون خطی ساده (ارزیابی مدل رگرسیون)

مدرس: محمدمهدی بشارتی

❑ آزمون معناداری پارامترهای مدل رگرسیون خطی

برای انجام آزمون فرضیه در مورد ضریب متغیر توصیفی، فرضیه صفر $b=0$ و فرضیه مقابل $b\neq 0$

آماره زیر درصورتی‌که فرضیه صفر صحیح باشد $(b = 0)$، از توزیع t_{n-p} پیروی می‌کند (p تعداد پارامترهای مدل).

$$t_b = \frac{\text{ضریب رگرسیون}}{\text{خطای استاندارد ضریب رگرسیون}} = \frac{\hat{b} - b}{\hat{\sigma} / \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}}$$

در مدل رگرسیون خطی ساده، اگر شرط $|t_a| > t_{1-\alpha/2,\, n-2}$ برقرار باشد؛ فرضیه صفر $(b = 0)$ رد می‌شود. بدین ترتیب می‌توان گفت مقدار پارامتر b تفاوت معناداری با صفر دارد.

- به‌علاوه، **فاصله اطمینانِ $(1-\alpha)$ درصدی** برای شیب خط رگرسیون ساده (b) به‌صورت زیر قابل‌محاسبه خواهد بود.
- تمام مقادیری که در این بازه قرار می‌گیرند، ازنظر آماری می‌توانند مقادیر قابل‌قبولی برای پارامتر موردنظر (b) باشند.

$$\hat{b} \pm t_{(1-\alpha/2,\, n-2)} \hat{\sigma}_{\hat{b}}$$

رگرسیون خطی ساده

مدرس: محمدمهدی بشارتی

مثال؛

داده‌های مربوط به «تعداد سفرهای تولیدشده از نواحی ترافیکی یک شهر» و «تعداد خانوارهای ساکن در هر ناحیه از آن شهر» را در اختیار داریم. مطلوب است ساخت یک مدل رگرسیون خطی ساده برای پیش‌بینی تعداد سفرهای تولیدشده از هر ناحیه.

جدول ۱۰-۴ نتیجه آزمون تحلیل واریانس برای بررسی معناداری مدل رگرسیون خطی ساده

مدل	مجموع مربعات	درجه آزادی	میانگین مربعات	آماره F	Sig.
رگرسیون	۹۷۴۹۲۴۱	۱	۹۷۴۹۲۴۱	۲۲۶/۸	۰/۰۰۰
باقیمانده‌ها	۸۵۹۶۴۲	۲۰	۴۲۹۸۲		
مجموع	۱۰۶۰۸۸۸۳	۲۱			

جدول ۱۰-۵ جدول نتایج ازمون‌های معناداری پارامترهای مدل رگرسیون خطی

مدل	ضرایب استانداردنشده		ضرایب استانداردشده	t	Sig.	فاصله اطمینان ٪۹۵ برای B	
	B	خطای استاندارد	بتا			حد پایین	حد بالا
عرض از مبدأ	۱۴/۵۸	۸۳/۱۱		۰/۱۷۵	۰/۸۶۳	-۱۵۸/۷۹	۱۸۷/۹۵
تعداد خانوار	۱/۱۶۰	۰/۰۷۷	۰/۹۵۹	۱۵/۰۶	۰/۰۰۰	۰/۹۹۹	۱/۳۲۱

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

مقدمه

رگرسیون خطی ساده

رگرسیون خطی چند متغیره

پیش‌فرض‌های رگرسیون خطی

معیارهای نیکویی برازش

تبدیلات (دستکاری متغیرهای مدل)

رگرسیون خطی چندمتغیره

مدرس: محمدمهدی بشارتی

در مطالعات کاربردی، معمولاً با مدل‌های رگرسیون با بیش از یک متغیر روبرو هستیم. به مدل رگرسیونی که بیش از یک متغیر توصیفی دارد، مدل رگرسیون چندمتغیره می‌گویند

$$Y_i=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+...+\beta_k X_{ki}+\varepsilon_i$$

در این مدل، k متغیر توصیفی وجود دارد که برخی از آن‌ها ممکن است کاملاً مستقل از یکدیگر نبوده و با یکدیگر همبستگی داشته باشند.

Explanatory variable Independent variable

به همین دلیل، بهتر است به‌جای واژه «**متغیر مستقل**» از اصطلاح «**متغیر توصیفی**» برای اشاره به این متغیرها استفاده شود.

درواقع، اصطلاح «متغیر توصیفی»، هدف از استفـاده از این متغیرها را نیز بهتر توضیح می‌دهد.

زیرا تحلیل‌گر تلاش می‌کند در قالب مـدل رگرسیون، از این متغیرها برای توصیف پراکندگی‌های مشاهده‌شده در متغیر وابسته استفاده کند.

علاوه بر این، معادله رگرسیـون خطی چندمتغیره به‌جای یک عرض از مبدأ و یک شیب، دارای یک مقدار ثـابت (متناظر با عرض از مبدأ در رگرسیون خطی ساده) و یک ضریب برای هریک از متغیرهای توصیفی است.

این ضرایب، «**ضرایب رگرسیون جزئی**» یا «**پارامترهای مدل**» نامیده می‌شوند.

رگرسیون خطی چندمتغیره

درواقع، **ضرایب رگرسیون جزئی**، برآوردی از **پارامترهای مجهول جامعه** می‌باشند.

در مدل رگرسیون چندمتغیره، ضریب رگرسیون جزئی برای یک متغیر توصیفی به ما می‌گوید که با یک واحد افزایش در مقدار آن متغیر (و با شرط ثابت ماندن مقدار سایر متغیرهای توصیفی درون مدل)، مقدار متغیر وابسته چقدر تغییر می‌کند.

در تشریح رگرسیون چند متغیره معمولاً از جبر بردارها و ماتریس‌ها استفاده می‌شود.

زیرا استخراج نتایج و تحلیل آن‌ها را ساده‌تر و کوتاه‌تر می‌نماید.

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} \quad , \quad \mathbf{X} = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1K} \\ 1 & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n2} & \cdots & X_{nK} \end{bmatrix}_{n\times K}$$

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1K} \\ 1 & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n2} & \cdots & X_{nK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$\mathbf{y}$ و $\mathbf{u}$ بردارهای ستونی $\mathbf{n \times 1}$

$\mathbf{X}$ ماتریس $\mathbf{n \times k}$

$\boldsymbol{\beta}$ بردار ستونی $\mathbf{k \times 1}$

برآورد ضرایب رگرسیون چندمتغیره (برآوردگرهای OLS)

برآوردگرهای OLS از طریق حداقل نمودن مجموع مجذور خطا (SSE) به دست می‌آیند.

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\mathbf{y}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

$$\boldsymbol{e}' = [e_1 \quad \cdots \quad e_n]$$

$$\text{SSE} = \sum_{i=1}^{n} (e_i)^2 = \boldsymbol{e}'\boldsymbol{e}$$

e بردار ستونی خطاها و e' ترانسپوز آن است.

برای حداقل شدن $e'e$ نسبت به $\hat{\beta}$ مشتق می‌گیریم.

با حل معادلات مربوطه، برآوردگرهای $\hat{\beta}$ با استفاده از روش OLS به صورت زیر به دست می‌آید؛

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

آزمون معناداری مدل رگرسیون چندمتغیره

$$Y_i=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots+\varepsilon_i$$

برای معناداری مدل رگرسیون، می‌خواهیم این فرضیه را آزمون کنیم که هیچ‌کدام از متغیرهای توصیفی درون مدل بر تغییرات Y_i تأثیری ندارند. بنابراین فرضیه صفر را به‌صورت زیر تبیین می‌کنیم:

$$H_0: \beta_1=\beta_2=\ldots=\beta_k=0$$

و فرضیه مقابل (H_a) بیان می‌کند که حداقل یک متغیر توصیفی وجود دارد که در تغییرات Y_i مؤثر است (حداقل یک β_i وجود دارد که برابر با صفر نیست). درصورتی‌که فرضیه صفر صحیح باشد، خواهیم داشت:

$$Y_i = \beta_0+\varepsilon_i$$

که در آن، $\beta_0 = \bar{y}$ است. بنابراین، صحیح بودن فرضیه صفر بدان معناست که تغییرات Y_i صرفاً تصادفی بوده و فقط تابعی از ε_i خواهد بود. همان‌طور که قبلاً اشاره شد، از روش تحلیل واریانس برای آزمون معناداری مدل رگرسیون استفاده می‌شود (در رگرسیون خطی ساده، این آزمون دقیقاً همان آزمون t برای بررسی فرضیه صفر $H_0: b=0$ است). زیرا در مدل رگرسیون ساده، فقط یک متغیر توصیفی وجود دارد. اما در مدل رگرسیون چندمتغیره به دلیل تعدد متغیرهای توصیفی، باید از تحلیل واریانس و آزمون F برای بررسی معناداری مدل رگرسیون استفاده شود.

آزمون معناداری مدل رگرسیون چندمتغیره

مشابه آنچه در مورد رگرسیون خطی ساده گفتیم، می‌توان نشان داد که برای یک مدل رگرسیون با k پارامتر درصورتی‌که فرضیه صفر ($H_0: \beta_1=\beta_2=...=\beta_k=0$) صحیح باشد، آماره زیر از توزیع F پیروی می‌کند.

$$F=\frac{SS_M/(k-1)}{SS_E/(n-k)} \sim F_{(k-1,\, n-k)}$$

بدین ترتیب، اگر مقدار آماره F، بزرگ‌تر از مقدار بحرانی $F_{\alpha,\, k-1,\, n-k}$ باشد، فرضیه عدم معناداری مدل رگرسیون رد می‌شود.

بنابراین، جدول تحلیل واریانس با استفاده از آماره F، این فرضیه صفر را آزمون می‌کند که بین متغیر وابسته و متغیرهای توصیفی رابطه خطی وجود ندارد. این آزمون را آزمون رگرسیون کلی (Overall regression F-test) می‌نامند.

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

مقدمه

رگرسیون خطی ساده

رگرسیون خطی چند متغیره

پیش‌فرض‌های رگرسیون خطی

معیارهای نیکویی برازش

تبدیلات (دستکاری متغیرهای مدل)

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

- تعدادی پیش‌فرض زیربنایی وجود دارد که ساخت مدل و استفاده از روش رگرسیون خطی مستلزم برقراری این پیش-فرض‌ها است.
- تحلیل‌گر برای سنجش اعتبار مدل خود، پس از برازش مدل و انجام آزمون‌های آماری در مورد معناداری مدل و پارامترهای آن، باید برقرار بودن پیش‌فرض‌های زیربنایی رگرسیون خطی را نیز بررسی کند.
- در صورت عدم برقراری این پیش‌فرض‌ها، مدل رگرسیون خطی عادی، مدل مناسبی برای مدل‌سازی پدیده موردنظر نخواهد بود و لازم است تمهیداتی برای رفع مشکل اندیشیده شود.

- برای یک مدل رگرسیون خطی (در برآورد OLS)، فرض می‌شود جمله خطا ε_i، یک متغیر تصادفی **مستقل** است که به ازای هر X_i دارای **توزیع نرمال** با **میانگین صفر** و **واریانس ثابت** σ^2 است.
- بنابراین، درصورتی‌که مدل رگرسیون خطی برای داده‌های موردمطالعه مناسب باشد؛ مقادیر باقیمانده‌ها (e_i) می‌بایست ویژگی‌های مفروض برای ε_i را منعکس کنند.
- به همین دلیل، از مقادیر باقیمانده‌ها برای بررسی پیش‌فرض‌های مدل رگرسیون استفاده می‌شود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

❖ خصوصیات جزء خطا (جزء تصادفی) مدل رگرسیون (به منظور استفاده از روش OLS)

۱- میانگین جملات خطا برابر با صفر است. (Zero mean of disturbances)

$$E(E_i|X_i) = 0$$

۲- واریانس جملات خطا (برای همه مقادیر متغیر توصیفی) همسان است. (Homoscedasticity of disturbances)

$$var(E_i|X_i) = E(E_i^2|X_i) = \sigma^2$$

۳- عدم خودهمبستگی (عدم رابطه میان جزءهای خطا). (Non-Autocorrelation of disturbances)

$$cov(E_k, E_j) = E(E_k E_j) = 0 \qquad k \neq j$$

۴- استقلال جزء تصادفی از جزء غیرتصادفی. (Exogeneous regressors)

$$cov(E_i, X_i) = 0$$

۵- نرمال بودن جملات خطا (Normality of disturbances)

$$E_i \sim N(0, \sigma^2)$$

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۱– صفر بودن میانگین خطاها

این فرض بیان می‌کند که؛

$$E(\varepsilon_i) = 0$$

با توجه به اینکه؛

$$e = y - \hat{y}$$

وقتی امیدریاضی خطاها برابر با صفر باشد، بدان معناست که مقدار برآوردشده به طور متوسط برابر با مقدار واقعی است.

از طرف دیگر، اگر معادله رگرسیون دارای عرض از مبدا باشد، فرض $E(\varepsilon_i) = 0$ نقض نمی‌شود؛ اما اگر رگرسیون فاقد عرض از مبدا باشد، مقدار متوسط خطاها لزوماً صفر نخواهد شد و دارای آثار نامطلوبی خواهد بود. از جمله اینکه؛

الف) R^2 ممکن است منفی شود. یعنی متوسط نمونه ($\bar{Y}$) می‌تواند تغییرات Y را بهتر از متغیرهای توصیفی (مدل)، توضیح دهد.

ب) به طور بالقوه می‌تواند منجر به اریب شدید در برآورد شیب مدل شود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

این فرض بیان می‌کند که واریانس مقادیر Y مشاهده‌شده، حول خط رگرسیون باید به ازای همه مقادیر متغیر توصیفی، ثابت (برابر) باشد.

دو فرض «نرمال‌بودن» و «ثابت‌بودنِ واریانس خطاها» بیان می‌کنند که اگر با ثابت نگه‌داشتن مقدار x، چندین بار نمونه‌گیری انجام شود، آنگاه توزیع مقادیر Yبه ازای هر x، از یک توزیع نرمال پیروی خواهد کرد که واریانس این توزیع‌ها به ازای همه مقادیر x، ثابت است.

به‌عبارت‌دیگر، به ازای هریک از مقادیر x، ε_i یک متغیر تصادفی مستقل بوده که از توزیع نرمال با میانگین صفر و واریانس ثابت σ^2 پیروی می‌کند.

این فرض که واریانس جمله خطا به ازای همه مقادیر متغیر توصیفی، ثابت بوده و مستقل از مشاهدات است؛ با نام هم‌واریانسی (Homoscedasticity) شناخته می‌شود.

این ویژگی به این معناست که تأثیر خالص عدم اطمینان موجود درون مدل (که شامل اثرات پنهان، خطا در اندازه‌گیری‌ها و نیز تصادفی بودن نمونه‌گیری است)، در میان داده‌ها به‌صورت سامانمند نبوده، بلکه به‌صورت تصادفی است.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

نتیجه نقض این فرض آن است که دقت برآورد پارامترهای مدل، کاهش‌یافته و یا به‌عبارت‌دیگر، **کارایی** مدل رگرسیون کاهش می‌یابد.

ویژگی ناهمسانی واریانس، موجب تغییر در ناااریب بودن و سازگار بودن برآوردگرهای OLS نمی‌شود و فقط بر روی کارایی (واریانس) اثر می‌گذارد (یعنی این برآوردگرها دیگر حداقل واریانس را ندارند)

یعنی برآورد واریانس جمله خطا را دچار اریب می‌کند.

و از انجا که واریانس‌ها به درستی برآورد نمی‌شوند، لذا آزمون فرضیه‌هایی که با t و Fو کای‌دو انجام می‌شوند، قابل اطمینان نخواهند بود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

آزمون‌های تشخیص واریانس ناهمسان

۱- آزمون بارتلت.

وقتی قابل استفاده است که مشاهدات تکراری باشند (مثلا داده‌های مربوط به استان‌های کشور طی ۲۰ سال)

یا مدل‌سازی حجم تردد سالانه برای گروه‌های جمعیتی استان‌های کشور طی ۲۰ سال

اگر مشکل ناهمسانی واریانس وجود داشته باشد، مثلا با افزایش جمعیت، واریانس‌ها نیز افزایش می‌یابد.

فرضیه صفر: واریانس‌ها همسان است.

رد فرضیه صفر: مشکل ناهمسانی واریانس وجود دارد.

آماره آزمون درصورت صحت فرضیه صفر، از توزیع کای‌دو با n-1 درجه آزادی پیروی می‌کند (n: تعداد گروه‌هایی است که می‌خواهیم همسانی واریانس آن‌ها را بیازماییم)

پیش‌فرض‌های زیربنایی رگرسیون خطی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

آزمون‌های تشخیص واریانس ناهمسان

۲- آزمون رتبه‌ای اسپیرمن.

برای بررسی رابطه میان واریانس E_i و X_i می‌توان از ضریب همبستگی رتبه‌ای اسپیرمن استفاده نمود.

برای این منظور، مقادیر قدرمطلق باقیمانده‌های مدل(ε_i) را حساب کرده و به همراه X_i ها رتبه‌بندی می‌کنیم.

سپس آزمون ضریب همبستگی رتبه‌ای اسپیرمن را انجام می‌دهیم.

فرضیه صفر: واریانس‌ها همسان است.

رد فرضیه صفر: مشکل ناهمسانی واریانس وجود دارد.

آماره آزمون درصورت صحت فرضیه صفر، از توزیع t با n-2 درجه آزادی پیروی می‌کند.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

آزمون‌های تشخیص واریانس ناهمسان

۳- آزمون بروش-پاگان.

این آزمون به طور کلی در قالب مراحل زیر انجام می‌شود؛

الف) مدل رگرسیون موردنظر با روش OLS برآورد گردیده و مقادیر باقیمانده‌ها (ε_i) محاسبه می‌گردد.

ب) واریانس مدل رگرسیون به صورت $\hat{\sigma}^2 = \frac{\sum_{i=i}^{n} (\varepsilon_i)^2}{n}$ محاسبه می‌شود.

ج) یک مدل رگرسیون با درنظر گرفتن $\frac{(\varepsilon_i)^2}{\hat{\sigma}^2}$ به عنوان متغیر وابسته، با روش OLS برآورد گردیده و مجموع تغییرات توضیح داده‌شده توسط مدل (SS_M) محاسبه می‌شود.

د) آماره آزمون درصورت صحت فرضیه صفر، از توزیع کای‌دو با r-1 درجه آزادی پیروی می‌کند (r: تعداد متغیرهای موجود در مدل مرحله ج)

فرضیه صفر: واریانس‌ها همسان است.

رد فرضیه صفر: مشکل ناهمسانی واریانس وجود دارد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

آزمون‌های تشخیص واریانس ناهمسان

سایر آزمون‌های رایج شامل؛

آزمون وایت

آزمون ضریب لاگرانژ

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

یک روش کیفی برای تشخیص واریانس ناهمسان

بررسی نمودار مقادیر باقیمـانده‌ها در برابر مقـادیر پیش‌بینی‌شده

برای بررسی فرض «همسانی واریانس» می‌توان نمودار مقادیر باقیمـانده‌ها را در برابر مقـادیر پیش‌بینی‌شده ترسیم نمود.

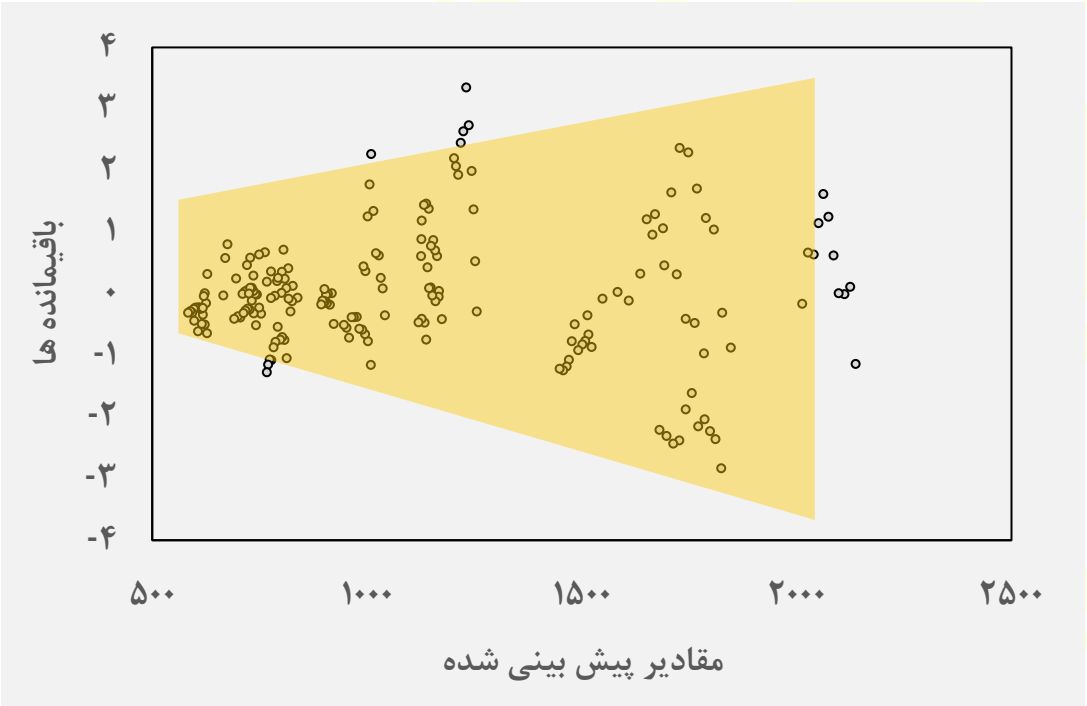
درصورتی‌که با افزایش مقدار متغیر توصیفی، واریـانس متغیر وابسته افزایش یا کاهش یابد، نقاط موجود در این نمودار به**صورت قیفی شکل** درخواهد آمد و این شرایط، حاکی از عدم برقراری فرض ثابت بودن واریانس خواهد بود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

مثال؛

مثال قبلی را دوباره در نظر بگیرید، می‌خواهیم بدانیم آیا پیش‌فرض ثابت بودن واریانس خطاهای مدل برقرار است؟



ناهمسانی واریانس خطاها
(heteroscedasticity of disturbances)
or
(heterogeneity of variance)

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

- هنگامی‌که خطاها به‌صورت ناهمسان (غیرثابت) می‌باشند (خطاها به‌صورت سامانمند در میان مشاهدات تغییر کنند)؛ آنگاه لازم است از سایر روش‌های مدل‌سازی مانند **حداقل مربعات وزن‌دار** و یا **حداقل مربعات تعمیم‌یافته (GLS)** استفاده شود.
- از سوی دیگر، گاهی اوقات ممکن است بتوان با **تغییر مقیاس متغیر وابسته** شرایط مناسبی برای انجام رگرسیون خطی به وجود آورد.
- برای این منظور، اگر واریانس متغیر وابسته با افزایش مقدار متغیر توصیفی، افزایش یابد، ممکن است با **جذر گرفتن** از مقادیر متغیر وابسته بتوان این مشکل را حل کرد.
- به‌طور معکوس، اگر واریانس متغیر وابسته با افزایش مقادیر متغیر توصیفی، کاهش پیدا کند، ممکن است با **لگاریتم گرفتن** از متغیر وابسته بتوان این مشکل را حل کرد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۲– ثابت (همسان) بودن واریانس خطاها (Homoscedastic disturbances)

نوع دیگری از ناهمسانی واریانس وجود دارد که طبق آن، واریانس ε_t تابعی از خطاهای گذشته (ε_{t-i} ها) است.

این نوع مدل‌ها معروف به مدل‌های Autoregressive Conditional Heteroskedastic (ARCH) هستند.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۳– مستقل و تصادفی بودن خطاها (عدم خودهمبستگی)

- یکی از پیش‌فرض‌های زیربنایی رگرسیون آن است که نمونه‌گیری به‌صورت مستقل و تصادفی انجام شده است.
- بدین معنا که مقدار یک مشاهده با مقدار مشاهده دیگر، رابطه‌ای ندارد.
- این فرض گاهی اوقات به این شکل بیان می‌شود که به ازای هریک از x، مقادیر خطا (ε_i) به‌صورت یک متغیر تصادفی و مستقل از هم هستند.
- **نام‌های دیگر:** «عدم خودهمبستگی خطاها»، «ناهمبسته بودن خطاها»، «استقلال خطاها»
- بدین معنا که مقدار خطاها در میان مشاهدات، مستقل از یکدیگر هستند.
- فرض استقلال خطاها برای به دست آوردن یک «فاصله اطمینان» و «بازه پیش‌بینی» معتبر، ضروری است.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۳– مستقل و تصادفی بودن خطاها (عدم خودهمبستگی)

- منظور از استقلال مشاهدات (یا باقیمانده‌ها) این است که احتمال مشاهده هریک از مقادیر درون جامعه، تحت تأثیر مشاهدات قبلی نباشد. این فرض به‌صورت زیر نیز بیان می‌شود:

اگر $i \neq j$، $Cov(\varepsilon_i, \varepsilon_j) = 0$

- خطاها هنگامی همبستگی خواهند داشت که مشاهدات در میان افراد، زمان‌ها و یا مکان‌ها، مستقل نباشند.
- خودهمبستگی مکانی، خودهمبستگی زمانی، خودهمبستگی میان افراد، ....
- **خودهمبستگی شاخصی از میزان همبستگی میان مشاهدات مجاور هم است (مجاورت مکانی، زمانی، ...).**
- مثلا مقدار خطاهای امروز ارتباطی با مقدار خطاهای دیروز ندارد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۳– مستقل و تصادفی بودن خطاها

خودهمبستگی مکانی (Spatial Autocorrelation)

- هنگـامی‌که داده‌ها از نواحی جغرافیایی که ویژگی‌های مشترکی دارند (مثلا مجاور یکدیگر هستند) جمع‌آوری شده باشد؛ ممکن است با مشکل همبستگی مکانی روبرو باشیم.
- برای مثال، درآمد خانوارهـای ساکن در بسیـاری از نواحی شهری، بی‌ارتباط با درآمد خانوارهای ساکن در نواحی مجاور نیست.
- آمار فضایی (Spatial Statistics)، حوزه مطالعاتی خاصی است که موضوع همبستگی مکانی را نیز در تحلیل‌های آماری در نظر می‌گیرد و می‌تواند کاربردهای فراوانی در حوزه برنامه‌ریزی حمل‌ونقل و کاربری زمین داشته باشد.
- آماره آیِ موران (Moran’s I) شاخصی برای بررسی وجود خودهمبستگی مکانی است.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۳– مستقل و تصادفی بودن خطاها

خودهمبستگی زمانی (Serial Correlation)-(Autocorrelation)- (Temporal Correlation)

- مشاهداتی که در طول زمان‌های متوالی انجام‌شده‌اند (داده‌های سری زمانی) رایج‌ترین داده‌هـایی هستند که ممکن است موضوع همبستگی میان مشـاهدات در مورد آن‌ها صدق کند.
- هنگامی‌که خطاهای مربوط به بازه‌های زمانی پشت‌سرهم همبستگی داشته باشند، همبستگی سریالی یا خودهمبستگی جمله خطا وجود دارد.
- برای مثال، احجام ترافیکی که هر ۱۵ دقیقه ثبت می‌شود، معمولاً با یکدیگر همبستگی دارند و یا حجم ترافیک یک ۱۵ دقیقه مشخص در هرروز کاری احتمالاً با روز بعد همبستگی خواهد داشت.
- همچنین، مشاهدات انجام‌شده بر روی یک فرد مشخص در طول زمان نیز ممکن است همبستگی زمانی داشته باشد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۳– مستقل و تصادفی بودن خطاها

خودهمبستگی زمانی (Temporal Correlation) -(Autocorrelation)-(Serial Correlation)

- **یک مثال دیگر:** فرض کنید قرار است با استفاده از داده‌های سری‌زمانی ماهیانه، یک مدل رگرسیون برای مدل کردن وسیله نقلیه-کیلومتر طی شده توسط حمل‌ونقل همگانی با استفاده از متغیرهای تعداد کارگر، سوخت و متغیرهای تولید سفر ساخته شود.
- درصورتی‌که اعتصاب کارگران در یک ماه بر روی وسیله نقلیه-کیلومتر طی شده تأثیر بگذارد، ممکن است بتوان گفت که این تغییر ناگهانی بر روی وسیله نقلیه-کیلومتر طی شده در ماه(های) بعدی نیز تأثیر خواهد گذاشت.
- این پدیده موجب بروز همبستگی سریالی در خطاها خواهد شد.
- همبستگی سریالی ممکن است غیر از پدیده‌های دوره‌ای یا شوک‌ها درنتیجه‌ی مشکلات دیگری ازجمله ناقص بودن مدل (حذف متغیرهای توصیفی مؤثر از درون مدل) یا (فرمول‌بندی اشتباه مدل) Model misspecification باشد.
- این مشکلات در بسیاری از مطالعات حمل‌ونقلی رخ می‌دهد اما به‌طور گسترده‌ای هم در تحقیقات و هم در عمل مورد بی‌توجهی و اغماض قرار می‌گیرد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۳– مستقل و تصادفی بودن خطاها

خودهمبستگی زمانی (Temporal Correlation) -(Autocorrelation)-(Serial Correlation)

- نمودار مناسب برای شناسایی همبستگی سریالی، نموداری است که مقادیر باقیمانده‌ها را در مقابل یک متغیر توالی نمایش می‌دهد.
- این متغیر توالی می‌تواند زمان انجام مشاهدات و یا ترتیب انجام مشاهدات (دوره فعلی و دوره قبلی) باشد.
- درصورتی‌که باقیمانده‌ها همبستگی سریالی نداشته باشند، نمودار باقیمانده‌ها نسبت به توالی انجام مشاهدات باید بدون هیچ الگوی خاصی، حول خط صفر پراکنده شده باشند و بازه تغییرات باقیمانده‌ها به ازای مقادیر مختلف X تفاوت چندانی نداشته باشد.
- اما اگر خطاها همبستگی سریالی داشته باشند، الگوی زمانی آن‌ها؛

۱) دارای روند صعودی یا نزولی است؛ و یا

۲) با نوسانات دوره‌ای (مثلاً ماهیانه) همراه است.

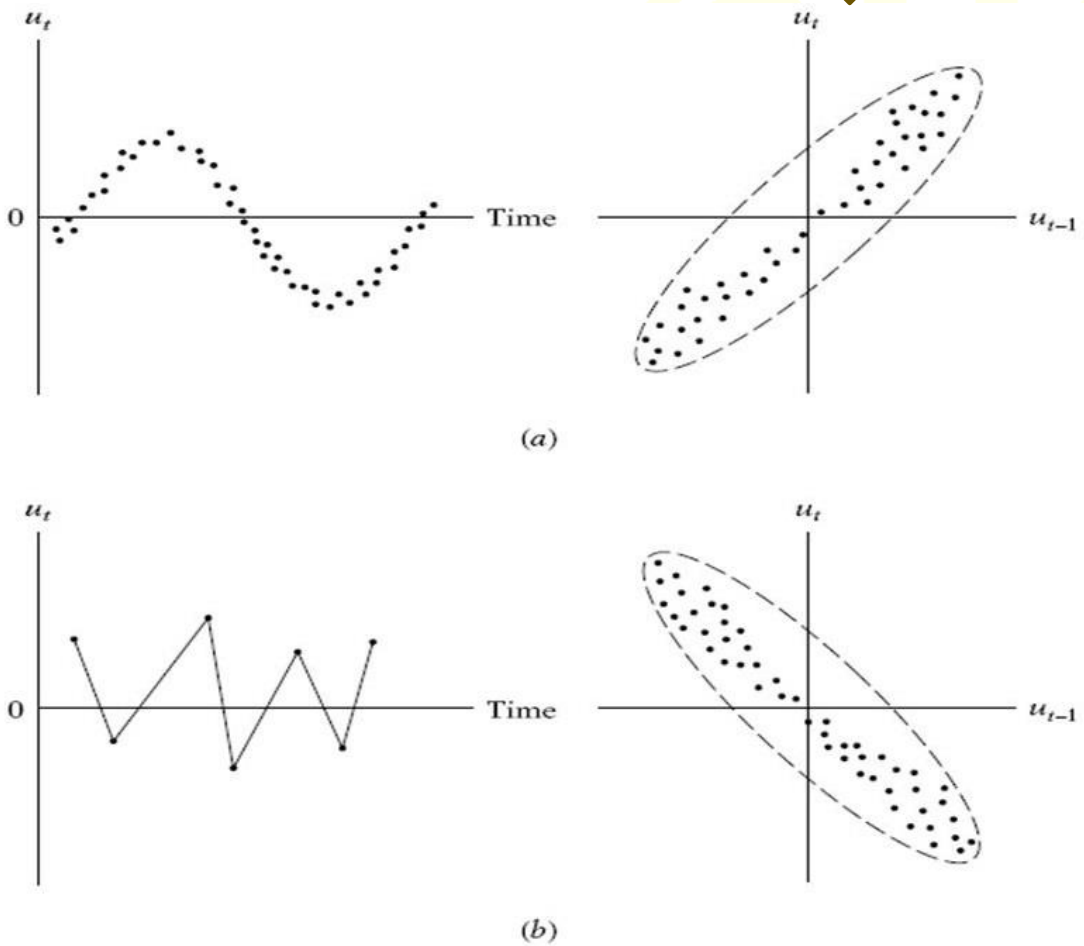
پیش‌فرض‌های زیربنایی رگرسیون خطی

۳– مستقل و تصادفی بودن خطاها

مثالی از خودهمبستگی (همبستگی سریالی) میان داده‌ها

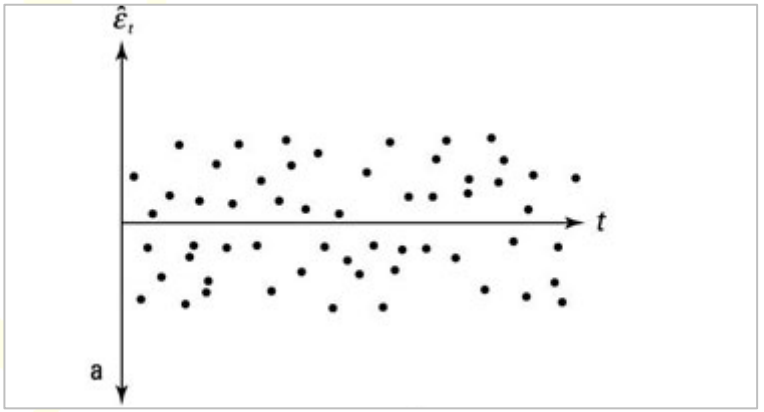
نمودار مقادیر باقیمانده‌ها را در مقابل ترتیب انجام مشاهدات (دوره فعلی و دوره قبلی)

نمودار مقادیر باقیمانده‌ها را در مقابل زمان انجام مشاهدات



(a) Positive and (b) negative autocorrelation.

نمودار شرایط عدم وجود خودهمبستگی



پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۳- مستقل و تصادفی بودن خطاها

خودهمبستگی زمانی

به کمک **آزمون دوربین-واتسون** نیز می‌توان وجود همبستگی سریالی را آزمود.

فرضیه صفر: همبستگی سریالی میان جملات خطا (در دوره فعلی و دوره قبلی) وجود ندارد (خطاها مستقل از هم هستند).

آماره این آزمون بر اساس مقادیر خطا در یک مدل رگرسیون OLS و از رابطه زیر محاسبه می‌شود.

$$d=\frac{\sum_{۲}^{n}(\hat{\varepsilon}_i-\hat{\varepsilon}_{i-۱})^{۲}}{\sum_{۲}^{n}\hat{\varepsilon}_i{}^{۲}}\cong 2(1-\hat{\rho})$$

$\hat{\rho}$: ضریب همبستگی میان جملات خطا در دوره فعلی و دوره قبلی

- درصورتی‌که خطاها مستقل باشند (فرضیه صفر صحیح باشد)، مقدار آماره این آزمون تقریباً برابر با ۲ خواهد بود.
- اما اگر مقدار آماره آزمون کمتر (بیشتر) از ۲ باشد، اطمینان کمتری در مورد عدم وجود همبستگی مثبت (منفی) بین مقادیر خطاها خواهیم داشت.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۳– مستقل و تصادفی بودن خطاها

خودهمبستگی زمانی

به کمک **آزمون دوربین–واتسون** نیز می‌توان وجود همبستگی سریالی را آزمود.

- مقادیر بحرانی آماره d به X_i بستگی دارد و برای داده‌های مختلف، مقدار متفاوتی خواهد داشت (آماره این آزمون از هیچکدام از توزیع‌های شناخته شده پیروی نم‌کند).
- برای رفع این مشکل، دوربین و واتسون حدود بالا (d_U) و پایینی (d_L) را برای مقادیر بحرانی به ازای تعداد متغیرهای توصیفی درون مدل ارائه کردند. مقادیر بحرانی در جداول آماده ارائه می‌شود.
- اما به‌عنوان یک قاعده کلی، درصورتی‌که مقدار آماره آزمون **دوربین–واتسون بین ۱.۵ تا ۲.۵ باشد**؛ می‌توان گفت **شرط استقلال خطاها برقرار است**.

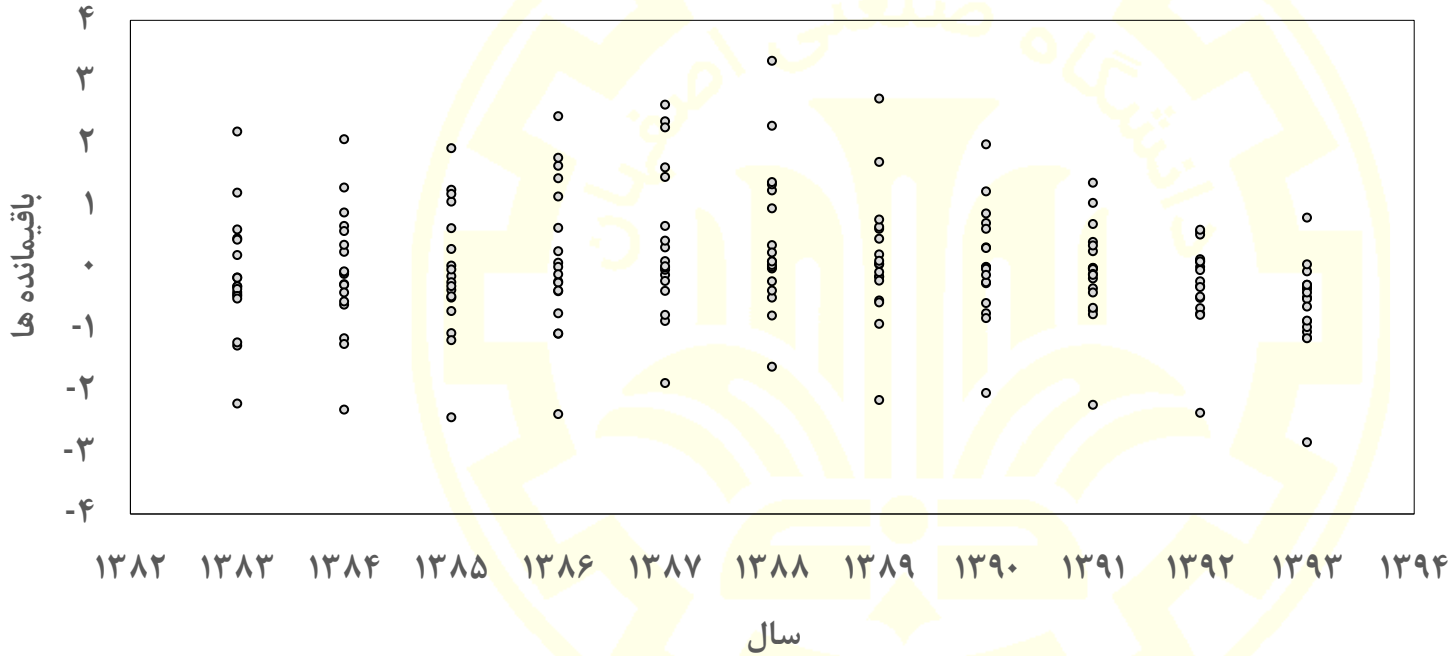
پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۳– مستقل و تصادفی بودن خطاها

مثال؛

مثال قبلی را دوباره در نظر بگیرید، می‌خواهیم بدانیم آیا در میان این داده‌ها، همبستگی سریالی وجود دارد یا خیر؟



جدول ۱۰-۶ خلاصه اطلاعات کلی مربوط به تحلیل رگرسیون (آماره دوربین–واتسون)

R^2	R^2 تعدیل‌شده	خطای استاندارد	آماره دوربین–واتسون
۰/۶۱۳	۰/۶۱۱	۳۴۶/۴۷	۱/۷۹۳

پیش‌فرض‌های زیربنایی رگرسیون خطی

۳– مستقل و تصادفی بودن خطاها

روش‌های لحاظ نمودن خودهمبستگی جملات خطا

در شرایطی که در میان داده‌ها **خودهمبستگی** وجود داشته باشد؛ روش‌های زیر که وجود همبستگی میان خطاها (مشاهدات) را در نظر می‌گیرند، می‌توانند گزینه‌های مناسبی برای تحلیل پدیده موردنظر باشند؛

- **حداقل مربعات تعمیم‌یافته (GLS)،**
- **مدل سری‌های زمانی (مدل‌های پویا، ARIMA، ARIMAX)،**
- **مدل‌های رگرسیون فضایی (Spatial regression)**

پیش‌فرض‌های زیربنایی رگرسیون خطی

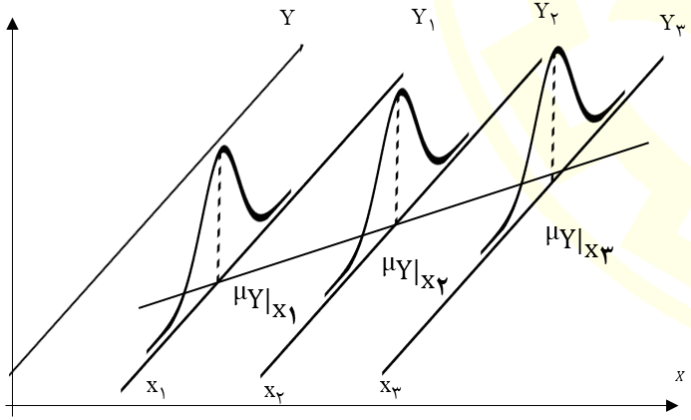
۴– نرمال بودن توزیع خطاها

به ازای هریک از مقادیر x، فرض می‌شود ε_i یک متغیر تصادفی مستقل است که از توزیع نرمال پیروی می‌کند.

درصورتی‌که این فرض را با فرض استقلال خطاها ترکیب نماییم؛ می‌توان نتیجه گرفت که در رگرسیون خطی فرض می‌شود جملات خطا به‌صورت مستقل و با توزیع نرمال مشابه هم (به‌صورت زیر) هستند (Independent and identically distributed (iid)).

$$\varepsilon_i \approx N(0 , \sigma^2)$$

$$Y_i \approx N(bX_i , \sigma^2)$$



- می‌دانیم که مقدار Y مشاهده‌شده به ازای یک x معین، خود یک متغیر تصادفی است و ممکن است در یک نمونه دیگر، به ازای همان مقدار x، مقدار جدیدی برای Y مشاهده شود.
- این فرض بیان می‌کند که توزیع مقادیر متغیر وابسته به ازای هر مقدار معین از متغیر توصیفی از توزیع نرمال پیروی می‌کند.
- به همین ترتیب، این فرض را می‌توان برای باقیمانده‌ها (خطاها) بیان کرد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۴– نرمال بودن توزیع خطاها

- به‌عنوان اولین گام برای بررسی نرمال بودن توزیع مقادیر باقیمانده‌ها، می‌توان از نمودار مستطیلی استفاده کرد.
- از سوی دیگر، می‌توان از نمودار Q-Q نیز برای بررسی نرمال بودن بهره برد.
- درصورتی‌که داده‌ها از توزیع نرمال پیروی کنند، انتظار داریم در نمودار Q-Q، داده‌ها، کم‌وبیش بر روی خط اریب قرار بگیرند.
- انحراف اندک توزیع باقیمانده‌ها از توزیع نرمال تأثیر زیادی بر نتایج تحلیل رگرسیون ندارد.
- تا زمانی که فرض نرمال بودن به‌صورت شدید نقض نشده باشد، نتایج تحلیل رگرسیون قابل‌قبول خواهد بود.
- **توجه:** برقرار نبودن سایر پیش‌فرض‌های رگرسیون خطی (به‌خصوص غیرخطی بودن مدل رگرسیون و یا ثابت نبودن واریانس خطاها) نیز می‌تواند موجب شود که توزیع باقیمانده‌ها، غیر نرمال شود.
- همواره لازم است قبل از نگرانی در مورد غیر نرمال بودن توزیع خطاها، برقرار بودن سایر پیش‌فرض‌های رگرسیون خطی را بررسی کنیم.

پیش‌فرض‌های زیربنایی رگرسیون خطی

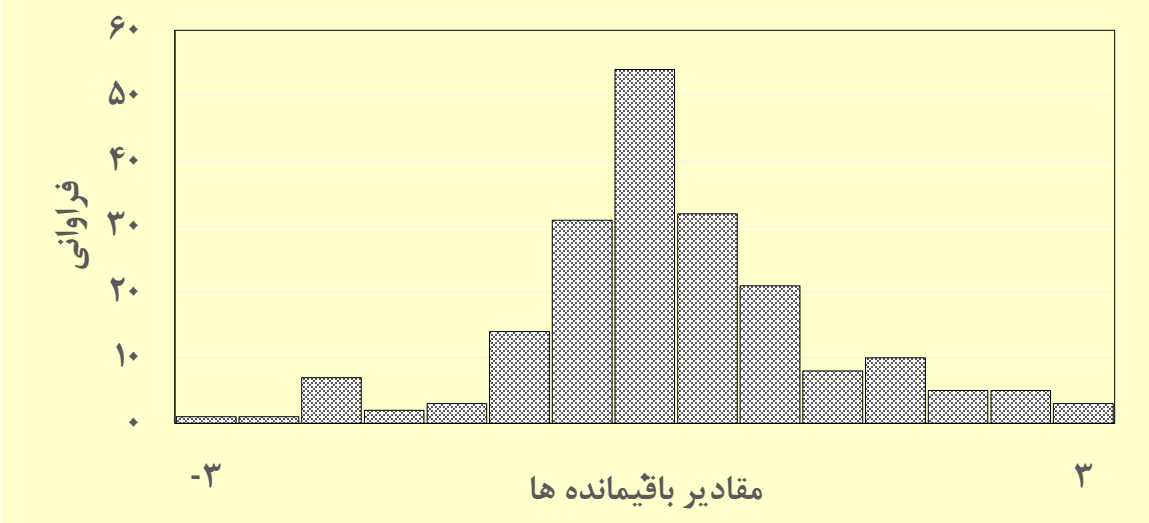
مدرس: محمدمهدی بشارتی

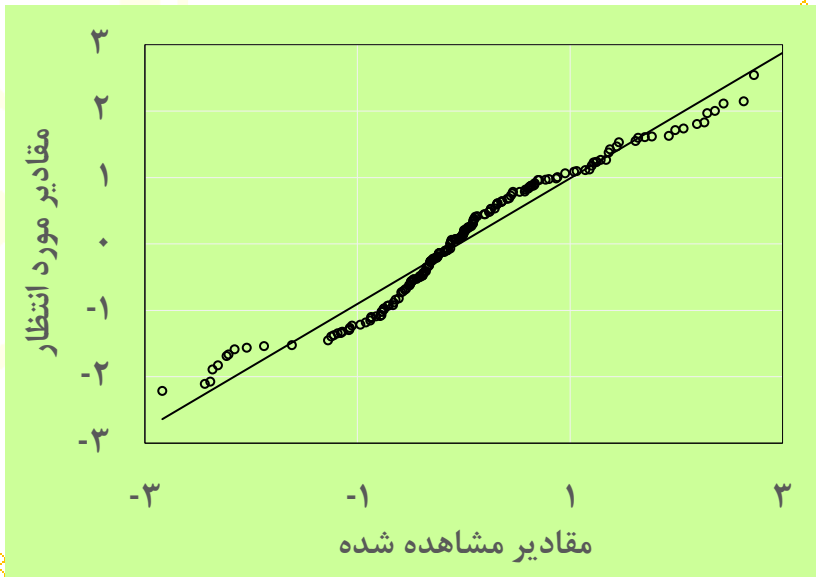
۴– نرمال بودن توزیع خطاها

مثال؛

مثال قبلی را دوباره در نظر بگیرید، می‌خواهیم بدانیم آیا پیش‌فرض «نرمال بودن توزیع خطاها» برای این مدل برقرار است؟

- می‌توان از مقایسه دو شاخص **چولگی** و **کشیدگی** برای **توزیع نرمال استاندارد** و **توزیع جملات خطا** نیز استفاده کرد.





پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۴– نرمال بودن توزیع خطاها

درصورتی‌که فرض نرمال بودن توزیع باقیمانده‌ها برقرار نباشد، می‌توان برای رفع این مشکل از روش‌های زیر استفاده کرد؛

- **تبدیل مقادیر متغیر وابسته**
  - اگر توزیع باقیمانده‌ها **چولگی مثبت** داشته باشد، از تبدیل لگاریتمی و
  - اگر **چولگی منفی** داشته باشد، به‌عنوان تبدیل، می‌توان تمـام مقادیر متغیر وابسته را به **توان دو** رساند.

- **مشاهدات پرت نیز می‌تواند فرض نرمال بودن را نقض کند.**
  - مشاهدات پرت ممکن است به دلیل وقوع پدیده‌های نادر به وجود آید.
  - یکی از راهکارهای حل این پدیده، استفاده از **متغیرهای مجازی (dummy)** است.
  - با استفاده از متغیر مجازی، می‌توان رویداد خاص موردنظر را به عنوان یک متغیر مجازی وارد مدل کرده و از این طریق، اثر آن رویداد خاص بر باقیمانده‌های مدل را تعدیل نمود.
  - برای مثال، گنجاندن متغیر سال ۱۳۸۵ برای لحاظ نمودن اثر هدف‌مندی یارانه‌ها در یک مدل رگرسیون.

پیش‌فرض‌های زیربنایی رگرسیون خطی

۵– ناهمبسته بودن متغیرهای توصیفی با جملات خطا – برون‌زایی (Exogeneity)

برون‌زایی متغیرهای توصیفی بدین معناست که آن متغیرها با جمله خطا همبستگی ندارند.

به‌عبارت‌دیگر، برون‌زایی یک متغیر مستقل بیان می‌کند که مقدار آن متغیر تحت تأثیر عواملی غیر از خود مدل است.

بنابراین، Y به‌طور مستقیم در مقدار یک متغیر برون‌زا تأثیری ندارد. این ویژگی در زبان ریاضی به‌صورت زیر بیان می‌شود.

برای تمامی مقادیر i و j $Cov(X_i, \varepsilon_j)=0$

هنگامی‌که یک متغیر توصیفی مهم، درون‌زا (وابسته به Y) باشد (Endogeneity)؛ آنگاه می‌بایست از روش‌های جایگزین مانند **روش‌های حداقل مربعاتِ دو یا سه مرحله‌ای** (با استفاده از متغیرهای ابزاری) و یا **مدل‌های معادلات ساختاری** برای مدل‌سازی رابطه متغیر وابسته با متغیرهای توصیفی استفاده نمود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۶- پیوسته بودن متغیر وابسته

- در مدل رگرسیون خطی، لازم است متغیر وابسته، ماهیت **پیوسته** داشته باشد.
- منظور از متغیر پیوسته، متغیری است که در **مقیاس‌های فاصله‌ای** یا **نسبتی** اندازه‌گیری شده باشد.
- درصورتیکه این شرایط برقرار نباشد؛ از رویکردهای زیر استفاده می‌شود؛

✓ **رگرسیون داده‌های شمارشی:** متغیرهایی که ماهیت **شمارشی** دارند (مانند فراوانی تصادفات) باید با استفاده از روش‌های رگرسیون داده‌های شمارشی مانند رگرسیون پواسون، رگرسیون دوجمله‌ای منفی و غیره، مدل‌سازی شوند.

✓ **مدل‌سازی متغیرهای گسسته:** مدل‌سازی متغیرهای وابسته‌ای که مقیاس **اسمی** دارند؛ نیازمند رویکردهای مدل-سازی متغیرهای گسسته است.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

۷– عدم وجود همخطی میان متغیرهای توصیفی

متغیرهای توصیفی گنجانده شده درون یک مدل نباید با یکدیگر همبستگی داشته باشند.

«عدم وجود هم‌خطی» یعنی هیچ نوع همبستگی کامل بین متغیرهای توصیفی وجود ندارد.

مسئله هم‌خطی یکی از مباحث مهم در مدل‌های رگرسیون است.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

هم‌خطی چندگانه (multicollinearity)

در تحلیل رگرسیون، انتظار می‌رود متغیرهای توصیفی همبستگی بالایی با متغیر وابسته داشته و درعین‌حال، همبستگی بالایی با یکدیگر نداشته باشند.

در این راستا، «هم‌خطی چندگانه (و یا به اختصار، هم‌خطی)» شرایطی است که در آن حداقل یکی از متغیرهای توصیفی ارتباط نزدیکی (همبستگی کامل و یا خیلی بالا) با یک یا چند متغیر توصیفی دیگر درون مدل داشته باشد.

همخطی یا همبستگی خطی میان دو متغیر بیانگر شدت همبستگی بین آن‌هاست.

انواع هم‌خطی:

۱- وجود **همخطی کامل (perfect collinearity)** . یک متغیر توصیفی ضریبی از متغیر دیگر باشد (به ندرت اتفاق می‌افتد).

۲- عدم **وجود همخطی**. هیچ رابطه خطی بین دو متغیر توصیفی وجود نداشته باشد (کوواریانس و ضریب همبستگی بین دو متغیر صفر باشد).

۳- وجود **همخطی ناقص (imperfect collinearity)** (حالت متداول). میزان همبستگی بین متغیرهای توصیفی، بین صفر و یک باشد. زمانی مشکل به وجود می‌آید که همخطی شدید باشد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

هم‌خطی چندگانه (multicollinearity)

هم‌خطی چندگانه ممکن است به دلایل زیر رخ دهد:

- وجود یک متغیر توصیفی درون مدل که خودش از طریق سایر متغیرهای توصیفی درون مدل محاسبه می‌شود (برای مثال فرض کنید متغیر «تعداد کل افراد شاغل در یک ناحیه ترافیکی» از مجموع سه متغیر «تعداد افراد شاغل در مشاغل صنعتی»، «تعداد افراد شاغل در مشاغل خدماتی» و «تعداد افراد شاغل در مشاغل تولیدی»، تشکیل شده باشد؛ و همه این چهار متغیر به‌طور هم‌زمان درون مدل رگرسیون گنجانده شده باشند).
- گنجاندن دو متغیر یکسان و یا بسیار شبیه به هم درون یک مدل رگرسیون. مثلاً گنجاندن دو متغیر که هر دو برای اندازه‌گیری یک مفهومِ واحد استفاده می‌شوند. برای نمونه، تعداد خودروهای پلاک‌گذاری شده و کیلومتراژ طی شده در خیابان‌های یک شهر، هر دو می‌توانند متغیرهایی برای اندازه‌گیری مفهوم مواجهه در حوزه ایمنی ترافیک باشند. بنابراین، گنجاندن هردو متغیر در یک مدل رگرسیون برای مدل‌سازی ریسک تصادفات، می‌تواند موجب بروز مشکل هم‌خطی گردد.

مثلا اگر دو متغیر «طول راه‌ها برحسب کیلومتر» و «طول راه‌ها برحسب متر» درون مدل باشند، همخطی کامل داریم (**perfect collinearity**).

- در موارد بالا، بیشتر به خطاهایی که ممکن است توسط تحلیل‌گر رخ دهد، اشاره شد. اما درمجموع، گنجاندن متغیرهای توصیفی که همبستگی کامل یا بسیار بالایی با یکدیگر دارند، موجب بروز مشکل هم‌خطی می‌شود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

هم‌خطی چندگانه (multicollinearity)

برخی از مهم‌ترین پیامدهایی که پدیده هم‌خطی به همراه دارد:

1. علیرغم این‌که در هنگام بروز هم‌خطی چندگانه، برآوردهای به‌دست‌آمده از روش حداقل مربعات، کماکان ناااریب می‌باشند؛ اما واریانس این برآوردها ممکن است به‌طور قابل‌توجهی بیشتر از مقدار واقعی باشد. هرچه هم‌خطی بیشتر باشد؛ خطای استاندارد برآورد پارامترهای متغیرهایی که هم‌خطی داشته‌اند، بیشتر خواهد بود. این موضوع موجب کاهش مقدار آماره t و رد نشدن فرضیه صفرِ عدم معناداری پارامترهای مدل می‌گردد. به‌عبارت‌دیگر، در صورت وجود هم‌خطی میـان متغیرهای مدل، ممکن است درحالی‌که متغیرهای توصیفی موردنظر تأثیر بسزایی در پیش‌بینی تغییرات مشاهده‌شده در متغیر وابسته داشته‌اند، فرضیه صفرِ عدم معناداری پارامترهای آن متغیرها را نتوان رد کرد.
2. ممکن است درحالی‌که آزمون F رگرسیون کلی، معنادار شده است؛ آزمون‌های F جزئی یا آزمون‌های t معنادار نشوند.
3. ممکن است **علامت ضرایب متغیرهای درون مدل، غیرمنطقی** شود. این موضوع به‌خصوص هنگامی رخ می‌دهد که دو متغیر که تقریباً یک مفهومِ واحد را تشریح می‌کنند؛ درون مدل گنجانده‌شده و با گرفتن علامت معکوس، اثرات یکدیگر را خنثی می‌کنند. درواقع، هنگامی‌که دو متغیر توصیفی، همبستگی مثبت و بالایی دارند، ضرایب آن‌ها در مدل رگرسیون، همبستگی منفی و بالایی خواهند داشت.
4. ضرایب متغیرها غالباً **غیر پایدار بوده** و به دلیل بالا بودن خطای استاندارد برآورد، نمی‌توان برآوردهای قابل اتکایی از پارامترهای مدل به دست آورد. به این معنا که با یک تغییر کوچک در داده‌های نمونه، ضرایب متغیرها تغییر می‌کنند. درواقع، اگر تحلیل‌گر تأثیر یک پـارامتر را بیش از مقدار واقعی برآورد کند؛ این موضوع موجب می‌شود اثرات پـارامتر دیگری را کمتر از مقدار واقعی برآورد نماید. بنابراین، برآوردهای ارائه‌شده برای پـارامترها، غیر پایدار بوده و از یک نمونه به نمونه دیگر، به‌شدت متفاوت خواهد بود.

به‌علاوه، حذف و یا اضافه نمودن متغیرها به مدل موجب تغییرات چشمگیر در مقدار و علامت ضرایب سایر متغیرها (که قبلاً درون مدل بوده‌اند) می‌گردد.

1. در این شرایط، روندهای خودکار که برای شناسایی و انتخاب بهترین مدل توسط نرم‌افزارها استفاده می‌شوند، مدل‌های مختلفی را به‌عنوان «بهترین مدل» پیشنهاد می‌کنند.

پیش‌فرض‌های زیربنایی رگرسیون خطی

هم‌خطی چندگانه (multicollinearity)

توجه

- هم‌خطی، درجات مختلفی دارد و برای پاسخ به این سؤال که در یک مسئله خاص، پدیده هم‌خطی باعث بروز مشکل جدی شده است یا خیر؛ یک آزمون قطعی وجود ندارد.
- تنها راه برای اطمینان از این‌که هم‌خطی به حداقل رسیده و یا کاملاً حذف شده است؛ طراحی دقیق یک آزمـایش است.
- اما رویکرد طراحی آزمایش‌ها در بسیـاری از مسائل واقعی، قابل‌استفاده نیست.
- زیرا در اکثر مسائل، توانایی کنترل نمودن تمام متغیرهای مؤثر بر پدیده موردنظر را نداریم.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

هم‌خطی چندگانه (multicollinearity)

روش‌های شناسایی مشکل هم‌خطی

- R^2 بالا ولی مقادیر t غیرمعنادار. هنگامی‌که بر اساس آزمون t، هیچ‌یک از ضرایب مدل معنادار نشده باشند (Sig.>۰.۰۵)؛ اما بر اساس آزمون F، مدل رگرسیون به‌طورکلی معنادار شده باشد (Sig.<۰.۰۵). البته باید توجه داشت که ممکن است تنها برخی از ضرایب معنادار نشده باشد؛ اما کماکان، مشکل هم‌خطی وجود داشته باشد.

- هنگامی‌که مدل رگرسیون را با استفاده از نمونه‌های مختلفی می‌سازیم؛ می‌بایست ضرایب مدل، پایدار بماند.

برای بررسی این موضوع، می‌توان از ابتدا، نمونه موجود را به‌صورت تصادفی به دو قسمت تقسیم نموده و پس از ساختن مدل با زیرمجموعه اول، اعتبـار مدل را با استفـاده از زیرمجموعه دوم، سنجید.

درصورتی‌که ضرایب متغیرهای دو مدل به‌طور قابل‌ملاحظه‌ای با یکدیگر متفاوت باشند، می‌توان گفت احتمالاً مشکل هم‌خطی چندگانه وجود دارد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

هم‌خطی چندگانه (multicollinearity)

روش‌های شناسایی مشکل هم‌خطی

- روش دیگر آن است که دوباره با استفاده از داده‌های نمونه قبلی، یک تغییر اندک در مدل قبلی ایجاد کرده و مدل جدیدی بسازید (مثلاً حذف کردن یک متغیر، یا اضافه کردن یک متغیر جدید و یا جایگزین کردن یک متغیر، با متغیر دیگری که همان مفهوم متغیر قبلی را اندازه‌گیری می‌کند). درصورتی‌که حذف یا اضافه نمودن یک متغیر، تغییر قابل‌ملاحظه‌ای در مقادیر ضرایب مدل به وجود بیاورد؛ می‌توان گفت احتمالاً مشکل هم‌خطی چندگانه وجود دارد.

- بررسی همبستگی میان هریک از جفت متغیرهای توصیفی. درصورتی‌که ضریب همبستگی میان دو متغیر توصیفی بیش از ۰.۸ باشد، می‌توان گفت حضور هم‌زمان این دو متغیر درون مدل، موجب بروز مشکل هم‌خطی خواهد شد.

البته باید توجه داشت که ۱) مقدار پیشنهادی ۰.۸ بستگی به مسئله موردمطالعه دارد. به‌علاوه، هنگامی‌که اندازه نمونه، کوچک است؛ می‌توان مقدار حدی ۰.۸ برای شناسایی همبستگی را کاهش داد. ۲) ممکن است یک متغیر توصیفی، از ترکیب خطیِ چند متغیر توصیفی دیگر به دست بیاید؛ اما درعین‌حال، همبستگی بالایی با مقادیر تک‌تک آن‌ها نداشته باشد. بنابراین، وجود همبستگی میان متغیرهای توصیفی لزوماً به معنای وجود پدیده هم‌خطی در مدل نخواهد بود و لازم است به سایر موارد نیز توجه شود.

پیش‌فرض‌های زیربنایی رگرسیون خطی

هم‌خطی چندگانه (multicollinearity)

روش‌های شناسایی مشکل هم‌خطی

یک شاخص مناسب و رایج برای شناسایی و اندازه‌گیری پدیده هم‌خطی چندگانه، «عامل تورم واریانس» و یا شاخص «تُلرانس» است Variance Inflation Factor (VIF)

- برای محاسبه این شاخص، ابتدا هریک از متغیرهای توصیفی X را براساس سایر متغیرهای توصیفی، مدل‌سازی کرده و مقدار R^2 مربوط به آن مدل را محاسبه می‌کنیم.
- سپس، به کمک رابطه زیر، مقدار VIF مربوط به متغیر موردنظر را محاسبه می‌کنیم.

$$VIF_i = \frac{1}{1 - R_i^2}$$

- R^2_i، مقدار R^2 مربوط به مدل رگرسیونی است که i-امین متغیر توصیفی را با سایر متغیرهای توصیفی، پیش‌بینی می‌کند.
- نهایتاً، با استفاده از مقدار به‌دست‌آمده از این شاخص، به شناسایی متغیری که باعث بروز هم‌خطی شده است، می‌پردازیم.

پیش‌فرض‌های زیربنایی رگرسیون خطی

هم‌خطی چندگانه (multicollinearity)

روش‌های شناسایی مشکل هم‌خطی

- بهترین حالت آن است که برای یک متغیر، $VIF_i \cong ۱$ شود که متناظر با $R_i^2 \cong 0$ خواهد بود.
- اما به‌عنوان یک قاعده سرانگشتی، هرگاه، شرط $VIF_i > ۱۰$ برقرار باشد ($R^2_i > ۰.۹$ و یا به‌طور معادل، تلرانس کمتر از ۰.۱ شود)؛ آنگاه می‌توان گفت مشکل هم‌خطی چندگانه به‌طور قابل‌ملاحظه‌ای وجود دارد.
- یک تفسیر مناسب برای VIF آن است که بگوییم وجود هم‌خطی چندگانه موجب می‌گردد که عرض فاصله اطمینان برای ضریب رگرسیون i-امین متغیر، به‌اندازه $\sqrt{VIF_i}$ نسبت به حالتی که متغیرهای توصیفی، کاملاً مستقل از هم باشند؛ متورم شود. به‌عبارت‌دیگر، مقدار $\sqrt{VIF_i}$ نشان می‌دهد که خطای استاندارد برآورد ضریب متغیر i، چقدر بیشتر از حالتی است که متغیر i، با سایر متغیرهای توصیفی درون مدل، همبستگی نداشته باشد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

مدرس: محمدمهدی بشارتی

هم‌خطی چندگانه (multicollinearity)

روش‌های حل مشکل هم‌خطی چندگانه

1. اطمینان از این‌که اشتباه آشکاری در گنجاندن متغیرها درون مدل رگرسیون رخ نداده است.
2. افزایش اندازه نمونه. گاهی اوقات، پدیده هم‌خطی به این دلیل رخ می‌دهد که نمونه‌گیری از یک فضای محدود و معین از کل جامعه آماری جمع‌آوری شده است. بنابراین، هنگامی‌که تلاش شود نمونه‌هایی تصادفی از کل فضای نمونه جمع‌آوری گردد، ممکن است بتوان مشکل هم‌خطی را مرتفع نمود.
3. در شرایطی که متغیرها دارای هم‌خطی جدی هستند، اغلب اوقات حذف یکی از آن‌ها از درون مدل می‌تواند یکی از گزینه‌های پیشِ رو باشد. برای مثال، در حالتی که متغیرهایی که مشکل هم‌خطی دارند، مفهوم مشترکی را توصیف می‌کنند؛ ممکن است بتوان با حذف یکی از متغیرها مشکل هم‌خطی را حل کرد. البته باید توجه داشت که حذف یک متغیر می‌بایست توجیه نظری داشته باشد و حذف یک متغیر که واقعاً تأثیر معناداری در مدل دارد، مجاز نیست. برای مثال، متغیری که دارای رابطه علت و معلولی با متغیر وابسته می‌باشد را نمی‌توان از مدل حذف کرد. توجیه دیگری که در این رابطه مطرح می‌شود، آن است که تنها متغیری درون مدل نگه داشته شود که در عمل، امکان جمع‌آوری داده‌های مربوط به آن وجود داشته باشد.

پیش‌فرض‌های زیربنایی رگرسیون خطی

هم‌خطی چندگانه (multicollinearity)

روش‌های حل مشکل هم‌خطی چندگانه

4. استفاده از مدل رگرسیون ریـج (Ridge regression). بر این ایده بنانهاده شده که با اضافه کردن اندکی اریب در برآوردهای پارامترهای مدل، می‌توان مقدار خطـای استـاندارد این برآوردها را به‌طور قابل‌توجهی کـاهش داد. درواقع، در محاسبه برآوردگر ریج، از فرض ناریب بودن برآوردها چشم‌پوشی شده و برآورد اریبی برای پارامتر مدل به دست می‌آید. به‌گونه‌ای که این برآورد، واریانس کمتری نسبت به برآورد به‌دست‌آمده از روش حداقل مربعات داشته باشد. این موضوع موجب می‌شود که برآوردهای به‌دست‌آمده، پایدار بمانند و با حذف یا اضافه نمودن یک متغیر توصیفی جدید به مدل، ضرایب برآورد شده برای سایر متغیرها، تغییر چندانی نکنند.
5. استفاده از روش Principal Component Analysis (PCA) برای تبدیل متغیرهای همبسته به متغیرهای ناهمبسته.
6. گاهی اوقـات، تنها اقـدام ممکن، آن است که وجود هم‌خطی چندگـانه درون مدل را به‌عنـوان یکی از محدودیت‌های مدلِ ساخته-شده، قبول کرده و گزارش کنیم!

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

یادآوری: شکل ماتریسی مدل رگرسیون خطی.

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} \quad , \quad \mathbf{X} = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1K} \\ 1 & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n2} & \cdots & X_{nK} \end{bmatrix}_{n\times K}$$

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{u}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{12} & \cdots & X_{1K} \\ 1 & X_{22} & \cdots & X_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & X_{n2} & \cdots & X_{nK} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

$\mathbf{y}$ و $\mathbf{u}$ بردارهای ستونی $\mathbf{n \times 1}$

$\mathbf{X}$ ماتریس $\mathbf{n \times k}$

$\boldsymbol{\beta}$ بردار ستونی $\mathbf{k \times 1}$

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

ماتریس واریانس-کوواریانس $\mathbf{u}$ را با ماتریس Ω نشان می‌دهیم؛

$$\Omega = \text{var} - \text{cov}(\mathbf{u}|\mathbf{X}) = E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}[u_1 \quad u_2 \quad \cdots \quad u_n] = E\begin{bmatrix} u_1^2 & u_1u_2 & \cdots & u_1u_n \\ u_2u_1 & u_2^2 & \cdots & u_2u_n \\ \vdots & \vdots & \cdots & \vdots \\ u_nu_1 & u_nu_2 & \cdots & u_n^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}$$

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

۱- امید ریاضی جمله خطا برابر با صفر است؛

$$\boldsymbol{E}(\mathbf{u}|\mathbf{X}) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

۲- جملات خطا مستقل بوده و خودهمبستگی ندارند. $\mathbf{Cov\left(\varepsilon_i,\varepsilon_j\right)=0}$ **یعنی** $\boldsymbol{E(uu')=0}$

این فرض موجب می‌شود که ماتریس واریانس-کوواریانس $\mathbf{u}$ تبدیل به یک ماتریس قطری شود (زیرا مولفه‌های غیر قطر اصلی، بیانگر کوواریانس u_i ها است.)

$$\Omega=\begin{bmatrix}\sigma_1^2 & 0 & \cdots & 0\\ 0 & \sigma_2^2 & \cdots & 0\\ \vdots & \vdots & \cdots & \vdots\\ 0 & 0 & \cdots & \sigma_n^2\end{bmatrix}$$

اگر فرض عدم خودهمبستگی جملات خطا برقرار نباشد، ماتریس واریانس-کوواریانس $\mathbf{u}$ به صورت زیر خواهد بود (کوواریانس‌ها غیرصفر خواهند بود)

$$\Omega=\begin{bmatrix}\sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n}\\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n}\\ \vdots & \vdots & \cdots & \vdots\\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2\end{bmatrix}$$

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

۳- واریانس جمله خطا ثابت (همسان) است. $Var(\varepsilon_i)=\sigma^2$

این فرض به همراه فرض قبلی موجب می‌شود که ماتریس واریانس-کوواریانس $\mathbf{u}$ تبدیل به ماتریس زیر شود؛

$$\Omega = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \mathbf{\sigma^2 I_n}$$

که در آن، $E(u_i^2) = \sigma^2$ و $\mathbf{I_n}$ ماتریس یکه n در n است.

اگر فرض همسانی واریانس‌ها برقرار نباشد اما فرض عدم خودهمبستگی برقرار باشد، ماتریس واریانس-کوواریانس $\mathbf{u}$ به صورت زیر خواهد بود.

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

۳- واریانس جمله خطا ثابت (همسان) است. $Var(\varepsilon_i)=\sigma^2$

همان‌طور که قبلا گفته شد، می‌توان از روش GLS برای حل مشکل ناهمسانی واریانس‌ها استفاده کرد.

برآوردگرهای $\hat{\beta}$ با استفاده از روش OLS به صورت زیر است؛

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

در روش GLS، این عبارت تعمیم داده می‌شود و ناهمسانی جملات خطای مدل نیز لحاظ می‌گردد؛

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{GLS}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{Y})$$

که در آن، $\boldsymbol{\Omega}$ ماتریس واریانس-کوواریانس $\mathbf{u}$ به صورت زیر خواهد بود.

$$\Omega = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

پیش‌فرض‌های زیربنایی رگرسیون خطی (به صورت ماتریسی)

۴- جزء تصادفی دارای توزیع نرمال است؛

$$\mathbf{u} \sim N(\mathbf{0}, \boldsymbol{\sigma}^2 \mathbf{I_n})$$

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

مقدمه

رگرسیون خطی ساده

رگرسیون خطی چند متغیره

پیش‌فرض‌های رگرسیون خطی

معیارهای نیکویی برازش

تبدیلات (دستکاری متغیرهای مدل)

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

- هدف از اضافه کردن یک متغیر توصیفی به یک مدل چندمتغیره، آن است که بخش بیشتری از واریانس مشاهده‌شده در مقادیر متغیر وابسته را توضیح دهیم.
- هزینه اضافه کردن یک متغیر جدید به مدل، **کاهش درجات آزادی**، **سخت‌تر شدن یافتن رابطه معنادار در آزمون‌های فرضیه** و **افزایش بازه فواصل اطمینان** می‌باشد.
- بهترین مدل، مدلی است که با کمترین تعداد متغیرهای توصیفی، بیشترین بخش از واریانس متغیر وابسته را پیش‌بینی کند.
  - از کجا بفهمیم که آیا اضافه کردن یک متغیر توصیفی به مدل بر میزان برازش مدل تأثیر معناداری دارد یا خیر؟
  - پنج مـورد از رایج‌ترین معیارهـای مورداستفاده برای بررسی نیکویی برازش مدل‌ها،
    - ✓ R^2،
    - ✓ R^2 تعدیل‌شده،
    - ✓ C_p،
    - ✓ AIC و
    - ✓ BIC

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

ضریب تعیین (R^2)

- ضریب تعیین، نشان می‌دهد که چه مقدار از کل پراکندگی موجود در مقـادیر مشـاهده‌شده متغیر وابسته، توسط مـدل، پیش‌بینی شده است.
- برای رگرسیـون خطی ساده، R^2 برابر است با توان دوم ضریب همبستگی پیرسون ($R^2 r = {}^2$).
- در حالت کلی، R^2 با استفاده از رابطه زیر محاسبه می‌شود.

$$R^2 = \frac{(SS_M)}{(SS_T)} = \frac{\Sigma_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{\Sigma_{i=1}^{n} (Y_i - \bar{Y})^2} = 1 - \frac{SS_E}{SS_T}$$

- SS_M مجموع مربعات مدل و SS_T مجموع مربعات کل در چهارچوب تحلیل واریانس است.
- مقدار R^2 بین صفر و یک قرار می‌گیرد.
- درواقع، از میان کل پراکندگی‌های مشـاهده‌شده در مقادیر Y، هرچه پراکندگی بیشتری را بتوان با خط رگرسیـون توضیح داد، R^2 به ۱ نزدیک‌تر می‌شود؛ و این نشـان‌دهنده آن است که مـدل، برازش بیشتری بر روی داده‌ها داشته است.
- هرچه R^2 به صفر نزدیک‌تر باشد، بدین معناست که مدل توانایی کمتری در پیش‌بینی پراکندگی‌های مشاهده‌شده در Y داشته است.

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

ضریب تعیین (R^2)

- متأسفانه نمی‌توان به‌طور قاطع، یک مقدار مشخص را به‌عنوان مقدار قابل‌قبول برای R^2 ارائه داد.
- مقدار قابل‌قبول برای R^2، به موضوع موردمطالعه بستگی دارد.
- برای مثال، در برخی از حوزه‌های مطالعات حمل‌ونقلی مانند مطالعات اقتصادی، $R^2 \geq 0/6$ را می‌توان به‌عنوان یک مقدار قابل‌قبول در نظر گرفت. درحالی‌که شاید در مدل‌های تولید و جذب سفر، ممکن است بتوان مدل‌هایی با R^2 بیش از ۰/۸۵ نیز ساخت.
- R^2 تنها یک معیار خام برای سنجش میزان برازش مدل بر روی داده‌ها است.
- وجود داده‌های پرت می‌تواند بر روی R^2 تأثیر گذاشته و مدل را بهتر یا بدتر ازآنچه در واقعیت وجود دارد، نشان دهد.
- برای مثال، داده‌های زوجی (x,y) را به‌صورت (۱،۴)، (۲،۶)، (۲،۴)، (۳،۳) در نظر بگیرید، مقدار R^2 برای این داده‌ها، نزدیک به صفر است. اکنون فرض کنید زوج (۱۰۰۰،۱۱۰۰) نیز به داده‌ها اضافه شود. مشاهده می‌شود که R^2 برای این مجموعه داده جدید، تقریباً برابر با ۱ خواهد شد.
- بنابراین همواره می‌بایست مراقب بود که وجود چند داده پرت بر روی استنباط ارائه‌شده در مورد معنادار بودن یا نبودن رابطه میان دو یا چند متغیر تأثیر نگذارد.
- در مواردی که رابطه میان متغیر وابسته و متغیر(های) توصیفی کاملاً غیرخطی است؛ استفاده از R^2 بی‌معنا خواهد بود.

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

ضریب تعیین (R^2)

در هنگام مقایسه R^2 دو مدل، می‌بایست نکات زیر را نیز در نظر داشت:

1) اندازه نمونه‌های استفـاده‌شده برای برآورد هر دو مـدل باهم برابر باشند. برای مثال، اگر پارامترهای یک مدل با ۱۰۰ داده و دیگری با ۱۰۰۰ داده برآورد شده و R^2 مربوط به مدل ۱۰۰۰ تایی کمتر از دیگری باشد؛ نمی‌توان گفت مدل با R^2 بیشتر، لزوماً مدل بهتری است.

2) متغیر وابسته مدل، یکسان باشد. برای مثال، R^2 دو مدل زیر را نمی‌توان با یکدیگر مقایسه کرد:

$$Y_i=\beta_0+\beta_1X_1+\beta_2X_2+...+\beta_pX_p+\varepsilon_i$$

$$Ln(Y_i)=\alpha_0+\alpha_1X_1+\alpha_2X_2+...+\alpha_pX_p+\varepsilon_i$$

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

ضریب تعیین تعدیل شده (R^2 تعدیل‌شده)

- گنجاندن هر متغیر توصیفی (حتی متغیرهـای بی‌ربط به مسئله موردبررسی) در مـدل می‌تواند تا حدودی موجب کاهش مجموع مربعات خطا و یا افزایش R^2 گردد.
- با اضافه شدن یک متغیر توصیفی جدید به یک مدل رگرسیون، مقدار R^2 نیز معمولاً افزایش می‌یابد و یا بدون تغییر باقی می‌ماند.
- بنابراین، افزایش R^2 توجیه کافی برای گنجاندن یک متغیر درون مدل موردنظر نیست.
- باید دید فایده اضافه کردن یک متغیر به مدل، در مقایسه با هزینه تحمیل‌شده به مدل از بابت کاهش درجه آزادی، توجیه‌پذیر است یا خیر؟
- پس تصمیم‌گیری در مورد اضافه کردن یک متغیر به مدل، می‌بایست بر پایه یک تحلیل هزینه-فایده باشد.
- ضریب تعیینی که با تعدیل R^2 نسبت به درجات آزادی، به‌دست‌آمده باشد را «ضریب تعیین تعدیل‌شده» می‌نامند.
- برای یک مدل رگرسیون با k متغیر توصیفی، ضریب تعیین تعدیل‌شده از رابطه زیر به دست می‌آید

$$R^2_{Adjusted} = \frac{\Sigma_{i=1}^{n} e_i^2/n-k}{\Sigma_{i=1}^{n}(Y_i-\bar{Y})^2/n-1} = 1-\left(\frac{n-1}{n-k}\right)(1-R^2)$$

- n تعداد مشاهدات و k تعداد پارامترهایی است که در مدل رگرسیون باید تخمین زد.

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

شاخص C_p

- ایده معیار Cp، ایجاد یک تعادل است میان حداکثر سازی میزان واریانس قابل تشریح y از طریق گنجاندن تمام متغیرهای مناسب و حداقل کردن واریـانس برآوردهـا (حداقل کردن خطـای استاندارد برآورد) از طریق کاهش تعداد پارامترهایی که قرار است برآورد گردند. آماره Cp از رابطه زیر محاسبه می‌شود.

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n - 2p)$$

- n، تعداد مشـاهدات، p تعداد پـارامترهای مـدل، SSE_p، میـانگین مربعات خطا برای مدل موردنظر با p پارامتر و $\hat{\sigma}^2$ بهترین برآورد از خطای واقعی است؛ که معمولاً برابر با کمترین مقدار MS_E در میان 2^k مدلِ رقیب در نظر گرفته می‌شود.
- بهترین مدل، مدلی است که کمترین مقدار Cp را داشته باشد.
- مدلی که همه پارامترهای موردبررسی را درون خود داشته باشد، همواره $C_p = p$ است.
- بنابراین، برای بررسی مدلی که همه پارامترهای موردبررسی را درون خود گنجانده است، نمی‌توان از معیار C_p استفاده کرد.
- در شرایطی که مقدار Cp برای چند مدل مختلف تقریباً باهم برابر باشند؛ می‌توان مدلی که ازلحاظ منطقی، مناسب‌تر بوده، مشکل کمتری در ارتباط با هم‌خطی چندگانه دارد و هزینه اندازه‌گیری متغیرهای آن کمتر است را به‌عنوان بهترین مدل انتخـاب کرد.

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

معیارهای اطلاعات (AIC و BIC)

دو معیـار R_a^2 و Cp برای اضافه کردن متغیرهای توصیفی جدید به مـدل، جریمه‌ای در نظر می‌گیرند.

دو معیار AIC و BIC نیز به‌منظور ایجاد تعادل میان پیچیدگی مدل و افزایش برازش مدل، برای اضافه کردن متغیرهای جدید به مدل، جریمه در نظر می‌گیرند تا از این طریق از افزودن متغیرهـای توصیفی که کمک چندانی به بهبود برازش مدل نمی‌کنند، جلوگیری شود.

این دو معیار عبارت‌اند از معیار اطلاعات آکائیک و معیار اطلاعات بیزین (معیار بیزینِ شوارتز). این دو معیار، بر اساس برآورد حداکثر درستنمایی پارامترهای مدل به دست می‌آیند.

هرچه مقدار این دو معیار برای یک مدل، کمتر باشد، مدل موردنظر مناسب‌تر خواهد بود.

برای حالتی که مدل رگرسیون با روش حداکثر درستنمایی، برآورد شده باشد؛ معیارهای اطلاعات آکائیک و بیزین از روابط زیر محاسبه می‌شود.

$$AIC = ۲ - \ln(\hat{L}) + ۲p$$

$$BIC = ۲ - \ln(\hat{L}) + p.\ln(n)$$

p ، تعداد پارامترهای مدل و $\hat{L}$ مقدار حداکثر تابع درستنمایی برای مدل برآوردشده، می‌باشد.

معیارهای نیکویی برازش (Goodness Of Fit(GOF))

معیارهای اطلاعات (AIC و BIC)

زمانی که مدل رگرسیون با روش حداقل مربعات با فرض توزیع نرمال خطاها، برآورد شده باشد. در این حالت، مقادیر معیارهای اطلاعات آکائیک و بیزین از روابط زیر به دست می‌آیند

$$AIC = n.\ln(SSE) - n.\ln(n) + ۲p$$

$$BIC = n.\ln(SSE) - n.\ln(n) + p.\ln(n)$$

n ، اندازه نمونه و SSE_p، مجموع مربعات خطا برای مدل با p پارامتر است.

به‌علاوه، در این حالت، واریانس برآوردشده نیز به‌عنوان یکی از پارامترهای مدل در نظر گرفته می‌شود. بنابراین، در این حالت، p برابر است با تعداد پارامترهای مدل به‌اضافه یک.

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

مقدمه

رگرسیون خطی ساده

رگرسیون خطی چند متغیره

پیش‌فرض‌های رگرسیون خطی

معیارهای نیکویی برازش

تبدیلات (دستکاری متغیرهای مدل)

دست‌کاری متغیرها در مدل رگرسیون

مدرس: محمدمهدی بشارتی

- تحلیل‌گر به دلایل مختلفی ممکن است تصمیم بگیرد متغیرهای درون مدل رگرسیون را دست‌کاری کند.
- برای مثال، تحلیل‌گر ممکن است بخواهد اهمیت نسبی متغیرهای توصیفی درون مدل را با یکدیگر مقایسه کند.
- برای این منظور می‌توان از «مدل‌های رگرسیون استانداردشده» استفاده نمود.
- در شرایطی که پیش‌فرض‌های مـدل رگرسیون برقرار نباشد، ممکن است با دست‌کاری مقیاس متغیرها به کمک تبدیلات، بتوان فرضیات را ارضـا نمود.
- همچنین، با استفاده از متغیرهای حالت (Indicator variables) می‌توان متغیرهای با مقیاس اسمی یا ترتیبی را درون مدل رگرسیون گنجاند.

دست‌کاری متغیرها در مدل رگرسیون

مدرس: محمدمهدی بشارتی

مدل‌های رگرسیون استانداردشده

بسیاری از اوقات تحلیل‌گر علاقه‌مند است بداند اثرات نسبی هریک از متغیرهای توصیفی بر متغیر وابسته چه قدر است.

یکی از اشتباهات رایج در تحلیل رگرسیون آن است که بزرگی ضریب یک متغیر درون مدل را معادلِ بااهمیت بودن آن متغیر بدانیم.

در اغلب موارد متغیرهای توصیفی موجود درون مدل، مقیاس‌های یکسانی نداشته و در حالت عادی، امکان مقایسه اهمیت متغیرهای توصیفی با استفـاده از ضرایب آن متغیرهـا در مـدل، فراهم نیست.

برای مثـال فرض کنید در یک مدل رگرسیون خطی، تعداد سفرهای روزانه یک خانوار به کمک دو متغیر بُعد خانوار و درآمد خانوار مدل شده باشد. واضح است که نمی‌توان مستقیماً با مقایسه پارامترهای برآوردشده برای هریک از این متغیرهـا، به اهمیـت نسبی این دو متغیر در پیش‌بینی مقدار Y پی برد. زیرا واحـدهای اندازه‌گیری دو متغیر بعد خانوار و درآمد خانوار باهم متفاوت است.

در این شرایط، استفاده از یک مدل رگرسیون استانداردشده می‌تواند امکان مقایسه نسبی اهمیت متغیرها را فراهم آورد. منظور از مدل رگرسیون استانداردشده، مدل رگرسیونی است که تمام متغیرهای توصیفی آن بر اساس رابطه زیر استانداردشده باشند.

$$X_i^* = \frac{X_i - \overline{X}_i}{S(X_i)}$$

به کمک این رابطه، به ازای هر متغیر توصیفی، یک متغیر استـانداردشده با میـانگین صفر و واریـانس یک به دست می‌آید.

بدین ترتیب، مقیـاس همه متغیرها، یکسان شده و مقایسه اهمیت نسبی آن‌ها امکان‌پذیر خواهد بود.

دست‌کاری متغیرها در مدل رگرسیون

مدرس: محمدمهدی بشارتی

تبدیلات (Transformations)

ساختن مدل‌های خطی بسیـار ساده‌تر و پرکاربردتر از مدل‌های غیرخطی است.

واحدهای اندازه‌گیری یک متغیر به‌خودی‌خود، هیچ برتری نسبت به هم ندارند؛

گاهی اوقات، با تبدیل کردن مقیـاس اندازه‌گیری یک متغیر، می‌توان یک معـادله غیرخطی را به رابطه خطی تبدیل کرد.

برای مثـال، لگاریتم زمان سفر به‌اندازه خودِ زمان سفر به‌عنوان یک واحد اندازه‌گیری اعتبار دارد.

روش‌های تبدیل داده‌هـا روش‌هـایی هستند که به کمک آن‌هـا، تحلیل‌گر تلاش می‌کند رابطه‌ای که تمایل دارد را در میان داده‌هـا کشف کند.

داده‌هـا ممکن است برای دستیابی به یکی از اهداف زیر، تبدیل شوند:

1. افزایش تقـارن داده‌ها
2. کاهش واریـانس داده‌ها
3. افزایش میزان خطی بودن رابطه میان متغیرها

یکی از پرکاربردترین تبدیل‌ها در مهندسی برنامه‌ریزی حمل‌ونقل، لگاریتم است. لگاریتم زمان سفر، لگاریتم مطلوبیت گزینه‌ها و غیره، مثال‌هایی از این نوع هستند.

دست‌کاری متغیرها در مدل رگرسیون

مدرس: محمدمهدی بشارتی

تبدیلات (Transformations)

- گاهی اوقات دو متغیر به نحوی باهم ارتباط دارند که تبدیل یافته حداقل یکی از آن‌ها منجر به ایجاد رابطه خطی بین آن دو متغیر می‌شود.
- ازاین‌رو یکی از اصلی‌ترین موارد کاربرد تبدیل‌ها، تبدیل نمودن یک مدل غیرخطی برحسب متغیرها، به یک مدل خطی است.
- به‌گونه‌ای که پس از تبدیل داده‌ها، بتوان از روش رگرسیون خطی برای مدل‌سازی رابطه میان متغیرها استفاده کرد.
- بنابراین، هنگامی‌که نمودار پراکنشی حاکی از وجود رابطه «غیرخطی» میان دو متغیر است، ممکن است تبدیل نمودن یکی از متغیرها یا هر دو آن‌ها، به خطی کردن رابطه میان آن‌ها بیانجامد.
- نباید همواره به دنبال بهترین تبدیل برای هر مجموعه داده بود.
- برای مثال، اگر سه مجموعه داده مشابه دارید، به‌جای یافتن تبدیل‌های جداگانه‌ای که برای هریک مناسب است، بهتراست یک تبدیل که برای هر سه مجموعه، نسبتاً مناسب است را انتخاب نمایید.
- درواقع، همواره باید به خاطر داشت که هر مجموعه داده، نمونه‌ای تصادفی از یک جامعه است. بنابراین یک نمونه دیگر از همان جامعه، احتمالاً کمی با نمونه قبلی متفاوت خواهد بود.

دست‌کاری متغیرها در مدل رگرسیون

تبدیلات (Transformations)

مدل لُگ-لُگ (log-log):

فرض کنید رابطه دو متغیر به‌صورت زیر باشد:

$$Y=\alpha X^{\beta} e^{u}$$

در این صورت، با تبدیل لگاریتمی، می‌توان این رابطه را به‌صورت خطی درآورد:

$$Ln(Y)=Ln(\alpha)+\beta Ln(X)+u$$

$$Y^{*}=\alpha^{*}+\beta X^{*}+u$$

این مدل برحسب پارامترها و همچنین برحسب لگاریتم متغیرها، خطی است.

بنابراین، می‌توان با روش OLS پارامترهای آن را برآورد نمود.

دست‌کاری متغیرها در مدل رگرسیون

تبدیلات (Transformations)

مدل لُگ-خطی (Log-Linear):

فرض کنید رابطه دو متغیر به‌صورت زیر باشد:

$$Y=\alpha e^{\beta X+u} = \alpha e^{\beta X} e^{u}$$

در این صورت، با تبدیل لگاریتمی، می‌توان این رابطه را به‌صورت خطی درآورد:

$$Ln(Y)=Ln(\alpha)+\beta X+u$$

$$Y^{*}=\alpha^{*}+\beta X+u$$

این مدل نسبت به Y لگاریتمی و نسبت به X، خطی است.

به همین دلیل، به آن مدل «لگ-خطی (Log-Linear Model)» گفته می‌شود.

دست‌کاری متغیرها در مدل رگرسیون

تبدیلات (Transformations)

در ارتباط با تبدیل مقیاس متغیرهای وابسته و توصیفی توجه به سه نکته زیر ضروری است:

۱. تبدیل متغیر وابسته موجب تغییر توزیع متغیر وابسته (و بالتبع جمله خطا در مدل) می‌گردد.

بنابراین، درصورتی‌که توزیع خطای مدل مطابق با پیش‌فرض‌های موردنظر نباشد، می‌توان از تبدیل متغیر وابسته برای حل این مشکل، کمک گرفت.

۲. با تبدیل متغیر توصیفی می‌توان به خطی‌شدن رابطه میان متغیر توصیفی و وابسته کمک کرد.

به‌علاوه، تبدیل متغیر توصیفی تغییری در توزیع جمله خطا به‌وجود نمی‌آورد.

۳. گاهی اوقات ساختن یک مدل رگرسیون غیرخطی و یا استفاده از روش حداقل مربعات تعمیم‌یافته، می‌تواند جذاب‌تر و آسان‌تر از اعمال مجموعه‌ای از تبدیلات بر روی داده‌های خام برای فراهم آوردن امکان ساخت مدل رگرسیون خطی باشد.

مدل رگرسیون خطی

فصل
دهم

Linear Regression Model

متغیرهای مجازی

متغیرهای مجازی

متغیر مجازی برای توصیف دو (چند) حالت کیفی مورد استفاده قرار می‌گیرد.

$$Y_i = \alpha + \beta D_i + u_i$$

$$D_i = \begin{cases} 1 & \text{حالت مورد نظر} \\ 0 & \text{حالت دیگر} \end{cases}$$

معادله را با روش OLS برآورد می‌کنیم.

چون D فقط دو مقدار دارد، پس فقط دو مقدار برای Y_i به دست می‌آید:

$$D_i = 0 \Longrightarrow Y_i = \alpha$$

$$D_i = 1 \Longrightarrow Y_i = \alpha + \beta$$

بنابراین، تاثیر عامل موردنظر برابر است با β .

اگر تاثیر عامل موردنظر را به صورت درصد بیان کنیم، آنگاه درصد افزایش Y در ازای وقوع رویداد D عبارتست از

$$\frac{\Delta Y_i}{\Delta D_i} = \frac{\beta}{\alpha}$$

متغیرهای مجازی

مدرس: محمدمهدی بشارتی

مثال؛

فرض کنید می‌خواهیم اثر جنسیت بر سرعت تردد در آزادراه‌های یک شهر را بررسی کنیم. بدین منظور متغیر مجازی D_i را تعریف می‌کنیم (D=0 برای زنان و D=1 برای مردان)

اگر در مدل $Y_i = \alpha + \beta D_i + u_i$

$\beta > 0$ و معنادار باشد، آنگاه می‌توان گفت سرعت تردد مردان به طور متوسط بیشتر از زنان است.

- **میزان افزایش سرعت به خاطر جنسیت برابر با β خواهد بود.**
- **درصد افزایش سرعت ناشی از جنسیت برابر با $\frac{\beta}{\alpha}$ خواهد بود.**

متغیرهای مجازی

مدرس: محمدمهدی بشارتی

دام متغیرهای مجازی؛

در استفاده از متغیرهای مجازی باید دقت شود که اگر جامعه را به دو گروه تقسیم می‌کنیم، تنها برای یک گروه می‌توان متغیر مجازی تعریف کرد.

در غیر اینصورت چه اتفاقی می‌افتد؟

در حالت کلی، اگر جامعه به m حالت (گروه) تقسیم می‌شود، بایستی فقط $m-1$ متغیر مجازی تعریف گردد.

البته توجه شود که اگر رگرسیون فاقد عرض از مبدأ است، می‌توان m متغیر مجازی تعریف نمود.

رگرسیون روی متغیرهای مجازی و متغیرهای توصیفی؛

فرض کنید مدل رگرسیون به صورت زیر است؛

$$Y_i = \alpha + \beta X_i + \gamma D_i + u_i$$

$$D_i = \begin{cases} 1 & \text{حالت مورد نظر} \\ 0 & \text{حالت دیگر} \end{cases}$$

بنابراین، نتایج معادله رگرسیون به صورت روبروست؛

$$D_i = 0 \Longrightarrow Y_i = \alpha + \beta X_i$$

$$D_i = 1 \Longrightarrow Y_i = \alpha + \beta X_i + \gamma = (\alpha + \gamma) + \beta X_i$$

در واقع، رویداد موردنظر باعث می‌شود خط رگرسیون جابجا شود.

یعنی عرض از مبدأ معادله تغییر می‌کند.

رگرسیون روی متغیرهای مجازی و متغیرهای توصیفی؛

فرض کنید مدل رگرسیون به صورت زیر است؛

$$Y_i = \alpha + \beta X_i + \theta D_i X_i + u_i$$

$$D_i = \begin{cases} 1 & \text{حالت مورد نظر} \\ 0 & \text{حالت دیگر} \end{cases}$$

بنابراین، نتایج معادله رگرسیون به صورت روبروست؛

$$D_i = 0 \Longrightarrow Y_i = \alpha + \beta X_i$$

$$D_i = 1 \Longrightarrow Y_i = \alpha + (\beta + \theta) X_i$$

در این حالت، رویداد موردنظر باعث می‌شود شیب خط رگرسیون تغییر کند.

متغیرهای مجازی

رگرسیون روی متغیرهای مجازی و متغیرهای توصیفی؛

فرض کنید مدل رگرسیون به صورت زیر است؛

$$Y_i = \alpha + \beta X_i + \gamma D_i + \theta D_i X_i + u_i$$

$$D_i = \begin{cases} 1 & \text{حالت مورد نظر} \\ 0 & \text{حالت دیگر} \end{cases}$$

بنابراین، نتایج معادله رگرسیون به صورت روبروست؛

$$D_i = 0 \Longrightarrow Y_i = \alpha + \beta X_i$$

$$D_i = 1 \Longrightarrow Y_i = (\alpha + \gamma) + (\beta + \theta) X_i$$

در این حالت، رویداد موردنظر باعث می‌شود هم شیب خط رگرسیون و هم عرض از مبدأ تغییر کند.

متغیرهای مجازی

مدرس: محمدمهدی بشارتی

وجود ۲ عامل کیفی در مدل رگرسیون؛

فرض کنید می‌خواهیم اثر جنسیت و لغزنده بودن سطح جاده بر سرعت تردد در آزادراه‌های یک شهر را بررسی کنیم. بدین منظور دو متغیر مجازی برای جنسیت و لغزنده بودن تعریف می‌کنیم)؛

$$Y_i = \alpha + \beta D_{1i} + \gamma D_{2i} + u_i$$

$$D_{1i} = \begin{cases} 1 & \text{مردان} \\ 0 & \text{زنان} \end{cases} \qquad D_{2i} = \begin{cases} 1 & \text{لغزنده} \\ 0 & \text{خشک} \end{cases}$$

چهار حالت به صورت روبرو وجود خواهد داشت؛

$$D_{1i} = D_{2i} = 0 \Longrightarrow Y_i = \alpha$$
$$D_{1i} = 1 \; ; \; D_{2i} = 0 \Longrightarrow Y_i = \alpha + \beta$$
$$D_{1i} = 0 \; ; \; D_{2i} = 1 \Longrightarrow Y_i = \alpha + \gamma$$
$$D_{1i} = D_{2i} = 1 \Longrightarrow Y_i = \alpha + \gamma + \beta$$

پس از برآورد مدل، اگر β و γ معنادار باشند، آنگاه می‌توان گفت جنسیت و لغزنده بودن سطح جاده بر سرعت تردد اثر معناداری دارد.

همچنین، α سرعت تردد را در حالتی نشان می‌دهد که $D_1 = D_2 = 0$ (سرعت راننده زن در وضعیت خشکی راه).

متغیرهای مجازی

وجود ۲ عامل کیفی در مدل رگرسیون؛

فرض کنید می‌خواهیم اثر جنسیت و لغزنده بودن سطح جاده بر سرعت تردد در آزادراه‌های یک شهر را بررسی کنیم. بدین منظور دو متغیر مجازی برای جنسیت و لغزنده بودن تعریف می‌کنیم)؛

چهار حالت به صورت روبرو وجود خواهد داشت؛

$$D_{1i} = D_{2i} = 0 \quad \Longrightarrow \quad Y_i = \alpha$$
$$D_{1i} = 1 \; ; \; D_{2i} = 0 \quad \Longrightarrow \quad Y_i = \alpha + \beta$$
$$D_{1i} = 0 \; ; \; D_{2i} = 1 \quad \Longrightarrow \quad Y_i = \alpha + \gamma$$
$$D_{1i} = D_{2i} = 1 \quad \Longrightarrow \quad Y_i = \alpha + \gamma + \beta$$

β تأثیر جنسیت بر سرعت تردد اثر را نشان می‌دهد؛

۱- تأثیر جنسیت بر سرعت تردد، برای حالتی که «سطح راه خشک باشد»؛

$$D_{1i} = D_{2i} = 0 \quad \Longrightarrow \quad Y_i = \alpha$$
$$D_{1i} = 1 \; ; \; D_{2i} = 0 \quad \Longrightarrow \quad Y_i = \alpha + \beta \quad \Longrightarrow \quad \frac{\Delta Y_i}{Y_i} = \frac{\beta}{\alpha}$$

وجود ۲ عامل کیفی در مدل رگرسیون؛

فرض کنید می‌خواهیم اثر جنسیت و لغزنده بودن سطح جاده بر سرعت تردد در آزادراه‌های یک شهر را بررسی کنیم. بدین منظور دو متغیر مجازی برای جنسیت و لغزنده بودن تعریف می‌کنیم)؛

چهار حالت به صورت روبرو وجود خواهد داشت؛

$$D_{1i} = D_{2i} = 0 \Longrightarrow Y_i = \alpha$$
$$D_{1i} = 1 \; ; \; D_{2i} = 0 \Longrightarrow Y_i = \alpha + \beta$$
$$D_{1i} = 0 \; ; \; D_{2i} = 1 \Longrightarrow Y_i = \alpha + \gamma$$
$$D_{1i} = D_{2i} = 1 \Longrightarrow Y_i = \alpha + \gamma + \beta$$

β تأثیر جنسیت بر سرعت تردد اثر را نشان می‌دهد؛

۲- تأثیر جنسیت بر سرعت تردد، برای حالتی که «سطح راه لغزنده باشد»؛

$$D_{1i} = 0 \; ; \; D_{2i} = 1 \Longrightarrow Y_i = \alpha + \gamma$$
$$D_{1i} = D_{2i} = 1 \Longrightarrow Y_i = \alpha + \gamma + \beta$$

$$\Longrightarrow \frac{\Delta Y_i}{Y_i} = \frac{\beta}{\alpha + \gamma}$$

تأثیر متقابل (اندرکنش) ۲ عامل کیفی در مدل رگرسیون؛

اثر اندرکنشی را نیز به صورت یک متغیر وارد مدل می‌کنیم؛

$$Y_i = \alpha + \beta D_{1i} + \gamma D_{2i} + \theta D_{1i} D_{2i} + u_i$$

چهار حالت به صورت روبرو وجود خواهد داشت؛

$$D_{1i} = D_{2i} = 0 \Rightarrow Y_i = \alpha$$
$$D_{1i} = 1 \; ; \; D_{2i} = 0 \Rightarrow Y_i = \alpha + \beta$$
$$D_{1i} = 0 \; ; \; D_{2i} = 1 \Rightarrow Y_i = \alpha + \gamma$$
$$D_{1i} = D_{2i} = 1 \Rightarrow Y_i = \alpha + \beta + \gamma + \theta$$

پارامتر θ اثر متقابل D_1 و D_2 را بر متغیر وابسته نشان می‌دهد.

مثال: لحاظ نمودن اثر هدفمندی یارانه‌ها درون مدل (به دو دو صورت قابل انجام است)؛

حالت اول: تقسیم به دو دوره قبل و بعد از آن،

$$Y_t = \alpha + \beta X_t + \gamma D_t + \theta D_t X_t + u_t$$

$$D_t = \begin{cases} 1 & t \geq 1385 \\ 0 & t < 1385 \end{cases}$$

بنابراین، دو مدل رگرسیون خواهیم داشت (یکی برای قبل از ۱۳۸۵ و یکی برای پس از آن)؛

$$D_t = 0 \Longrightarrow Y_t = \alpha + \beta X_t + u_t \quad ; \; t < 1385$$

$$D_i = 1 \Longrightarrow Y_t = (\alpha + \gamma) + (\beta + \theta) X_t + u_t \quad ; \; t \geq 1385$$

مثال: لحاظ نمودن اثر هدفمندی یارانه‌ها درون مدل (به دو دو صورت قابل انجام است)؛

حالت دوم: اثر رویداد موردنظر فقط در زمان وقوع آن درنظر گرفته شود،

$$Y_t = \alpha + \beta X_t + \gamma D_t + u_t$$

$$D_t = \begin{cases} 1 & t = 1385 \\ 0 & t \neq 1385 \end{cases}$$