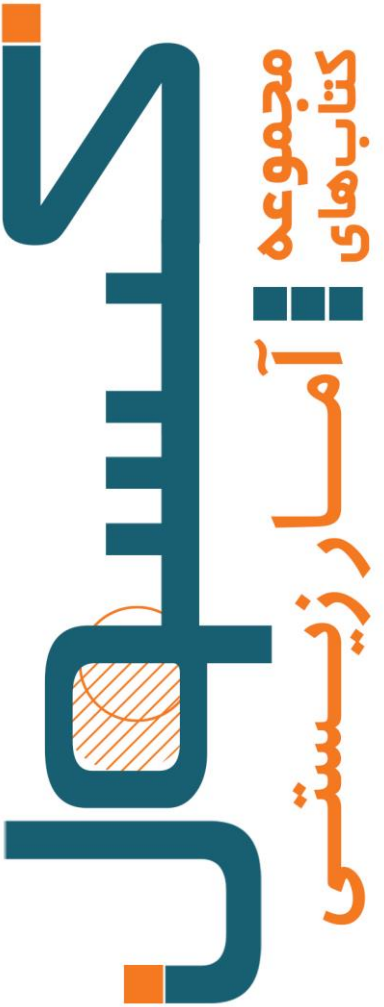

گردآورنده

پریسا ذاکری

تقدیم به

خدایی که آفرید
جهان را، انسان را، عقل را، علم را، معرفت را، عشق را
و به کسانی که عشقشان را در وجودم دمید.
گاهی بیاییم و احوالشان را بپرسیم،

تقدیم به

پدرم به استواری کوه،
مادرم به زلالی چشمه،
خواهرم به صمیمیت باران،
که مشوق دوران کودکی و جوانیم بودند،
و

تقدیم به

دانشجویانی که تمایل به شرکت در کنکور ارشد و دکتری آمار زیستی را دارند و مشوق من برای تهیه این کتاب بودند.

سپاسگزاری

سپاس بی‌کران پروردگار یکتا را که هستی‌مان بخشید و به طریق علم و دانش رهنمونمان شد و به همنشینی رهروان علم و دانش مفتخرمان نمود و خوشه‌چینی از علم و معرفت را روزیمان ساخت. همچنین سپاس ویژه از سرکار خانم دکتر منیره ملکی که پیشنهاد نوشتن این کتاب را به من دادند و به خاطر یاری‌ها و راهنمایی‌های بی‌چشمداشت ایشان بسیاری از سختی‌های این راه برایم آسان‌تر شد.

پریسا ذاکری
خردادماه ۱۳۹۸

فهرست مطالب

فصل اول: روش های آماری ۷

۱-۱ آمار توصیفی ۷

۱-۲ رگرسیون ۱۴

۱- ۳ آنالیز واریانس ۲۶

۱-۴ نمونه‌گیری ۲۸

۱-۵ طرح و آزمایش ۴۰

۱-۶ روش‌های ناپارامتری ۴۲

فصل دوم: آمار نظری ٤٥

۲-۱ آمار ریاضی ۴۵

۲-۲ انواع توزیع‌ها و روابط بین آن‌ها ۵۴

۲-۳ نامساوی‌ها ۶۸

۲-۴ تابع مولد گشتاور ۶۹

۲-۵ احتمالات ۷۲

۲-۶ برآورد فاصله‌ای ۷۵

۲-۷ امیدریاضی و واریانس و کوواریانس ۸۲

فصل سوم: ریاضیات ۸۹

۳-۱ تابع ۸۹

۳-۲- حد ۹۴

۳-۳ مشتق ۹۷

۳-۴ کاربرد مشتق ۹۹

۳-۵ انتگرال ۱۰۱

۳-۶ دنباله و سری ۱۰۳

۳-۷ ماتریس و دترمینان ۱۰۵

۳-۸ صفحه و خط ۱۰۶

۳-۹ توابع چندمتغیره ۱۰۸

۳-۱۰ مجموعه‌ها ۱۱۰

فصل (۱)

۱-۱- آمار توصیفی

در یک پژوهش جهت بررسی و توصیف ویژگی‌های عمومی پاسخ‌دهندگان از روش‌های موجود در آمار توصیفی مانند جداول توزیع فراوانی، در صد فراوانی، درصد فراوانی تجمعی و میانگین استفاده می‌گردد. بنابراین هدف آمار توصیفی یا descriptive محاسبه پارامترهای جامعه با استفاده از سرشماری تمامی عناصر جامعه است.

جدول آمار توصیفی

X	f_i	r_i	F_i
	۲	۰/۱	۲
	۲۰	۱	۲۲

X: متغیر

f_i: فراوانی مطلق یا تعداد مشاهده شده هر متغیر

r_i : $\frac{f_i}{n}$: فراوانی نسبی

$F_i = \Sigma f_i$: فراوانی تجمعی

دامنه $R = X_{max} - X_{min}$

فرمولی جهت یافتن تعداد دسته‌ها: $k \simeq ۱+۳/۳\ log_{۱۰}^{n}$

یا k تعداد دسته‌ها $\Rightarrow ۲^k \le n \le ۲^{k+۱}$

$\frac{R}{K}$: طول دسته

X_i : مرکز دسته

بازه دسته‌ها به صورت $(a,b]$ نشان داده می‌شود و دسته آخر به صورت $[a,b]$

پارامترهای مرکزی

شاخص‌های مرکزی به مقدار یا عددی که به بهترین شکل ویژگی گروه را به‌عنوان یک کل نشان می‌دهد اشاره می‌کنند.

مترادف شاخص مرکزی نمره‌ی متوسط یا معدل (average) است؛ بنابراین شاخص‌های مرکزی یا گرایش به مرکز شاخص‌هایی هستند که با استفاده از آن‌ها مجموعه‌ای از داده‌ها در یک مقدار یا عدد که نماینده‌ی آن مجموعه است خلاصه می‌شود. نما و میانه و میانگین (mean,median,mode) سه شاخص مرکزی هستند.

۱- میانگین: مشهورترین و معتبرترین شاخص گرایش مرکزی، میانگین است. میانگین معدل حسابی گروهی از نمرات است که از طریق جمع کردن تمام نمرات و تقسیم حاصل جمع بر تعداد کل نمرات به دست می‌آید. میانگین پایاترین و دقیق‌ترین شاخص مرکزی محسوب می‌شود. ۳ نوع میانگین برای ۲ نوع داده‌های خام و دسته‌بندی شده وجود دارد که عبارتند از:

انواع میانگین	داده‌های خام	داده‌های دسته‌بندی شده
حسابی	$\bar{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	$\bar{X}=\frac{\sum_{i=1}^{k}f_iX_i}{n}$
هندسی	$G=\sqrt[n]{\prod_{i=1}^{n}X_i}$	$G=\sqrt[n]{\prod_{i=1}^{k}X_i^{f_i}}$
توافقی	$H=\frac{n}{\sum_{i=1}^{n}\frac{1}{X_i}}$	$H=\frac{n}{\sum_{i=1}^{k}\frac{f_i}{X_i}}$

نکات ریز:

- در حالت کلی داریم: $H \le G \le \bar{X}$ اما اگر داده‌ها یکسان باشد حالت تساوی رخ می‌دهد، یعنی هر ۳ میانگین دارای مقادیر یکسان می‌شوند.
- همواره داریم: $G^2=\bar{X}.H$
- برای نسبت‌ها، نرخ‌ها، درصد رشد ⇐ بهترین میانگین ⇐ میانگین هندسی می‌باشد.

- وقتی بخواهیم مشاهدات دارای مقیاس ترکیبی مثل کیلومتر بر سرعت یا دور بر ثانیه باشند ⇐ بهترین میانگین ⇐ میانگین توافقی می‌باشد.
- میانه همواره بین مد و میانگین قرار دارد.
- در یک نمونه تصادفی دوتایی ⇐ میانه و میانگین با هم برابرند پس واریانس آنها نیز با هم برابر است.

۲- میانه (median): میانه نقطه‌ی وسط یک توزیع فراوانی است. هنگامی که نمرات توزیع فراوانی به ترتیب از بالاترین به پایین‌ترین نمره مرتب شود، نصف نمرات بالای میانه و نصف نمرات پایین میانه قرارمی گیرد. بنابراین، میانه ۵۰ درصد است. یعنی عددی است که توزیع نمره‌ها را به دو قسمت مساوی تقسیم می‌کند. طریقه به دست آوردن میانه به این صورت می‌باشد که ابتد کل داده‌ها را که تعداد آن برابر n است به‌صورت صعودی مرتب کنید، سپس:

- اگر تعداد داده‌ها فرد بود میانه برابر است با داده‌ی وسط، یعنی:

$$X_{\left(\frac{n+1}{2}\right)}$$

- اگر تعداد داده‌ها زوج بود میانه برابر است با میانگین دو داده‌ی وسط، یعنی:

$$\frac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n+1}{2}\right)}}{2}$$

روش یافتن میانه برای داده‌های دسته‌بندی شده:

$$\text{Median} : L_m + \frac{\left(\frac{n}{2} - F_{m-1}\right)}{f_m} w_m$$

که در این فرمول دسته شامل میانه، به نخستین دسته‌ای گفته می‌شود که فراوانی تجمعی آن از نصف تعداد داده‌ها بیشتر باشد.

F_{m-1} : فراوانی تجمعی دسته قبل از دسته شامل میانه

w_m : طول دسته شامل میانه

f_m : فراوانی مطلق دسته شامل میانه

L_m : کران پایین دسته شامل میانه

۲- صدک‌ها Q_p:

- روش یافتن در داده‌های خام:

$$Q_P = X_{(m)} \Leftarrow m \simeq \frac{np}{۱۰۰}$$

- روش یافتن در داده‌های دسته‌بندی شده:

تعریف: دسته شامل Q_P به نخستین دسته‌ای گویند که فراوانی تجمعی آن از $\frac{np}{۱۰۰}$ بیشتر باشد.

$$Q_P = L_P + \frac{\left(\frac{np}{۱۰۰} - F_{p-۱}\right)}{f_p} w_p$$

$F_{p-۱}$: فراوانی تجمعی دسته قبل از دسته شامل صدک p

f_p : فراوانی مطلق دسته شامل صدک p

L_p : کران پایین دسته شامل صدک p

نکته: میانه همان «صدک پنجاهم»، دهک پنجم و چارک دوم است.

۴- مد:

$$Mod = lm + \frac{d_۱}{d_۱ + d_۲} w_m$$

تعریف: دسته‌ای که فراوانی مطلق ماکزیمم دارد را مد گویند.

$d_۱$: اختلاف فراوانی مطلق دسته‌ی مد از دسته قبلی

$d_۲$: اختلاف فراوانی مطلق دسته مد از دسته بعدی

w_m : طول دسته مد

l_m : کران پایین دسته مد

پارامترهای پراکندگی

شاخص‌های پراکندگی، میزان پراکندگی (تغییرات) مقادیر هر متغیر را در اطراف میانگین نشان می‌دهند. زمانی که در شاخص‌های گرایش به مرکز، داده‌ها را در یک اندازه واحد خلاصه کنیم، طبیعتاً بخشی از جزئیات و اطلاعات حذف خواهد شد. ازاین‌رو باید به دنبال شاخص‌هایی برای اندازه‌گیری تفاوت موردها در یک متغیر باشیم تا از نحوه و میزان پراکندگی و تغییر داده‌ها در اطراف میانگین (یا میانه و مد) مطلع شویم.

۱- دامنه تغییرات: اختلاف بین بزرگ‌ترین داده از کوچک‌ترین داده در یک توزیع مشخص، دامنه تغییرات آن توزیع نام دارد. دامنه را با علامت R نشان می‌دهند. برای به دست آوردن دامنه تغییرات، بزرگ‌ترین داده (عدد) را منهای

کوچک‌ترین داده می‌کنیم. مثلاً دامنه تغییرات اعداد ۱، ۲ و ۵ عدد ۴ است که اگر عدد ۵ (بزرگ‌ترین داده) را از عدد ۱ (کوچک‌ترین داده) کم کنیم، عدد ۴ که معادل دامنه تغییرات است به دست می‌آید.

دامنه $R = X_{max} - X_{min}$

۲- واریانس: واریانس یک شاخص پراکندگی است که از طریق محاسبه‌ی انحراف نمره‌ها از میانگین محاسبه می‌شود. زیرا با مجذور کردن انحراف نمره‌ها از میانگین یا مجموع مجذور انحراف نمره‌ها از میانگین تقسیم بر تعداد نمره‌ها به دست می‌آید.

$$var(x) = \sum \frac{(x_i - \bar{x})^2}{n-1}$$

$$\sigma^2 = \overline{X^2} - \bar{X}^2$$

- واریانس برای داده‌ها دسته‌بندی شده عبارتست از: (k=تعداد دسته‌ها)

جامعه $\Rightarrow \sigma^2 = \dfrac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{n}$

نمونه $\Rightarrow S_n^2 = \dfrac{1}{n-1} \sum_{i=1}^{k} f_i (x_i - \bar{x})^2$

۳- ضریب تغییرات: در نظریه احتمال و آمار ضریب تغییرات (به انگلیسی: coefficient of variation، مخفف:CV) یک معیار بهنجار است که برای اندازه‌گیری توزیع داده‌های آماری به کار می‌رود. ضریب تغییرات (CV) معیاری است که تغییرات نسبی را اندازه‌گیری می‌کند. (و هرچه مقدار آن کمتر باشد بهتر است.) این معیار از تقسیم انحراف معیار بر میانگین مطابق فرمول زیر به دست می‌آید:

$CV = \dfrac{S}{\bar{X}} \rightarrow$ واحد ندارد

نکته ریز: ضریب تغییرات با تغییر مبدأ و مقیاس تغییر نمی‌کند.

شاخص‌های دیگری که اگر بدانید مفید است:

- میان چارکی: این شاخص پایاتر از دامنه‌ی تغییرات است و مساوی است با نصف تفاضل بین چارک سوم و اول:

$$Q' = \frac{Q_1 + Q_3}{2}$$

- میانگین انحراف‌ها: $MAD = 1/n \sum_{i=1}^{K} f_i \,|\, x_i - \bar{x} \,|$

- ضریب چولگی: m_3 گشتاور سوم $\Rightarrow S=\frac{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n}}{\sigma^3}=\frac{m_3}{\sigma^3}$
- ضریب برجستگی (کشیدگی) $=\frac{m_4}{S^4}$
- دامنه‌ی میان چارکی: $IQR=Q_3-Q_1$
- محاسبه‌ی چولگی از چندک‌ها: $s_k=\frac{(Q_3-Q_2)-(Q_2-Q_1)}{Q_3-Q_1}$

نکات ریز ضریب چولگی

- اگر مقدار چولگی مثبت باشد یعنی چولگی به سمت راست داریم و اگر مقدار چولگی منفی باشد یعنی چولگی به سمت چپ داریم.
- اگر $|S|\le 0/1$ ← نزدیک به نرمال
- اگر $|S|\ge 0/5$ ← انحراف قابل ملاحظه با نرمال دارد.
- همواره برای چولگی داریم: $S_{ax+b}=\frac{a^3}{|a|^3}S_X$
- در جوامع متقارن مانند نرمال $\Leftarrow S=0\Leftarrow$ یعنی میانگین و میانه و مد بر هم منطبق هستند.
- اگر مقدار ضریب کشیدگی مثبت باشد یعنی شکل نمودار برجسته است و اگر مقدار ضریب کشیدگی منفی باشد یعنی شکل نمودار پهن می‌باشد.

مقیاس

قبل از هرگونه تجزیه‌وتحلیل آماری لازم است که مقیاس اندازه‌گیری متغیرهای موردنظر را مشخص نماییم.
انواع مقیاس اندازه‌گیری عبارتند از، اسمی، ترتیبی، فاصله‌ای و نسبتی که توضیح هرکدام به شرح زیراست:

الف) متغیرهای کیفی

۱- مقیاس اسمی: متغیرهایی هستند که جنبه کیفی یک صفت را در نظر گرفته. بدین معنا که کدهایی که به پاسخ اختصاص داده می‌شوند اولویتی بر یکدیگر ندارند.
مثال: متغیر نوع رنگ: سفید: ۱، سیاه: ۲، زرد: ۳

مثال: جنسیت: خانم: ۱، آقا: ۲

۲- مقیاس ترتیبی: متغیرهایی هستند که در آنها، مقادیر مختلف نمراتی که به پاسخ‌ها اختصاص داده می‌شوند بیان شدت و ضعف آن صفت می‌باشد.

مثال: متغیر میزان رضایت: ۱: ناراضی، ۲: متوسط، ۳: راضی

ب) متغیرهای کمی

۱- مقیاس فاصله‌ای: متغیرهایی هستند با درجات مساوی که امکان اندازه‌گیری یک متغیر را به ما می‌دهند.

مثال: متغیرهای سن، میزان، میزان درآمد.

۲- مقیاس نسبی: برای اندازه‌گیری متغیرهایی که مبدأ سنجش آن‌ها صفر مطلق است.

مثال: متغیر دمای آب

نکته: اغلب این دو مقیاس را یکی در نظر می‌گیرند.

پس به یاد داشته باشید:

نوع مقیاس	نوع متغیر	ترتیب	فواصل مساوی	مبدأ صفر مطلق	شاخص پراکندگی	اندازه شاخص مرکزی
اسمی	کیفی	−	−	−		مد یا نما
ترتیبی		+	−	−	دامنه	میانه
فاصله‌ای	کمی	+	+	−	انحراف معیار، واریانس	میانگین حسابی
نسبی		+	+	+		میانگین حسابی و هندسی

الف) اسمی ← گروه خونی

ب) ترتیبی ← کیفیت کالا (درجه ۱، درجه ۲) ← ۴ عمل اصلی را ندارد.

پ) فاصله‌ای ← (+ −) هوش، حافظه،

ت) نسبتی ← ۴ اصل (+ × ÷ −) دارد.

۱-۲- رگرسیون

در مدل‌های آماری، تحلیل رگرسیون یک فرایند آماری برای تخمین روابط بین متغیرها می‌باشد. این روش شامل تکنیک‌های زیادی برای مدل‌سازی و تحلیل متغیرهای خاص و منحصربه‌فرد، با تمرکز بر رابطه بین متغیر وابسته و یک یا چند متغیر مستقل، می‌باشد. تحلیل رگرسیون خصوصاً کمک می‌کند در فهم اینکه چگونه مقدار متغیر وابسته با تغییر هرکدام از متغیرهای مستقل و با ثابت بودن دیگر متغیرهای مستقل تغییر می‌کند.

نکات طلایی که در مدل رگرسیون ساده خطی باید به یاد داشته باشید:

- $e_i = Y_i - \hat{Y}_i \Rightarrow \begin{matrix}\Sigma e_i = 0 \\ \Sigma x_i e_i = 0\end{matrix} \Rightarrow \Sigma e_i^2 \neq 0$
- $S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \Sigma x_i^2 - n\bar{x}^2$
- $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \Sigma y_i^2 - n\bar{y}^2$
- $S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum (x_i - \bar{x}) y_i$
  $= \sum (y_i - \bar{y}) x_i$
  $\hat{\sigma}^2 = S^2 = \frac{SSE}{n-2} = \frac{S_{YY} - \beta S_{XY}}{n-2}$
- $A = \hat{\alpha} = \bar{y} - \beta\bar{x} \quad , \ \beta = \hat{\beta} = \frac{S_{XY}}{S_{XX}}$

که A,B از هم مستقل نیستند و هر دو دارای توزیع نرمال می‌باشند.

$$Var(\beta) = \frac{\sigma^2}{S_{XX}} \qquad , \ Var(A) = \frac{\sum x_i^2}{nS_{XX}}\sigma^2$$

مدل رگرسیون خطی $Y_i = \alpha + \beta x_i + e_i \sim N(\alpha + \beta X_i, \sigma^2)$

$SSE = S_{YY} - \beta S_{XY}$

$$\frac{SSE}{\sigma^2} \sim \chi^2_{n-2}$$

$$\frac{SSR}{\sigma^2} \sim \chi^2_{n-2}$$

فرمول‌هایی که برای رگرسیون ساده خطی باید به خاطر بسپارید:

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

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

$$Y_i = \beta_O + B_1 X_i + E_i$$

$$\hat{\beta}_o = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X = \bar{Y} + \hat{\beta}_1 (X - \bar{X})$$

نکات طلایی

- همواره خط رگرسیون که پارامترهای آن به روش حداقل مربعات به دست آمده از نقطه $(\bar{Y}, \bar{X})$ می‌گذرد.
- همواره داریم: $\Sigma \hat{E}_i = 0$

$$\Sigma X_i \hat{E}_i = 0 \Rightarrow \Sigma \hat{Y}_i \hat{E}_i = 0 \Rightarrow \Sigma \hat{Y}_i^2 = \Sigma Y_1 \hat{Y}_i$$

- برای مجموع مربعات خطا در رگرسیون داریم:

مجموع مربعات خطا $\Rightarrow SSE = S_{yy} - \beta_1^2 S_{xx}$

قوانین مهم

- $\hat{\beta}_0$ دارای توزیع نرمال با میانگین β_0 و واریانس $\sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$ می‌باشد.
- $\hat{\beta}_1$ دارای توزیع نرمال با میانگین β_1 و واریانس $\frac{\sigma^2}{S_{XX}}$ می‌باشد.
- برآوردگر ناریب σ^2 به‌صورت $\frac{SSE}{n-2}$ است که آن را با MSE نشان می‌دهند.
- همواره داریم:

$$cov\left(\hat{\beta}_0, \hat{\beta}_1\right) = \frac{-\bar{X}}{S_{XX}}\sigma^2$$

- برآوردگری کاراتر است که ⟵ واریانس کمتری داشته باشد.
- برای مقادیر مشاهده شده y_i همواره داریم:

$$y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

- برای مقادیر ei همواره داریم:

$$var(E_i) = var(e_i) = \sigma^2\left[1 - \frac{1}{n} - \frac{(X_i - \bar{X})^2}{S_{XX}}\right]$$

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

فاصله اطمینان برای پارامترهای رگرسیون

فاصله اطمینان برای شیب خط رگرسیون یعنی β_1:

$$T = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{MSE / S_{xx}}}$$

فاصله اطمینان $100(1-\alpha)\%$

$$\left(\hat{\beta}_1 - t_{\left(n-2,1-\frac{\alpha}{2}\right)} \sqrt{\frac{MSE}{S_{xx}}}, \hat{\beta}_1 + t_{\left(n-2,1-\frac{\alpha}{2}\right)} \sqrt{\frac{MSE}{S_{xx}}} \right)$$

فاصله اطمینان برای عرض از مبدأ (β_0):

$$T^{(*)} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)}} \sim t_{(n-2)}$$

فاصله اطمینان $100(1-\alpha)\%$:

$$\Rightarrow \left(\hat{\beta}_0 - t_{\left(n-2,1-\frac{\alpha}{2}\right)} \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)}, \right.$$

$$\left. \hat{\beta}_0 + t_{\left(n-2,1-\frac{\alpha}{2}\right)} \sqrt{MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)} \right)$$

۱- آزمون فرض برای شیب خط رگرسیونی:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

آماره آزمون برای این فرض عبارتست از:

$$T = \frac{\hat{\beta}_1}{\sqrt{MSE / S_{xx}}}$$

ناحیه رد برای فرض‌های مختلف H_1:

رد $H_0 \Leftarrow T > t_{(n-2,1-\alpha)} \Leftarrow H_1 : B_1 > 0$

رد $H_0 \Leftarrow T < -t_{(n-2,1-\alpha)} \Leftarrow H_1 : B_1 < 0$

رد $H_0 \Leftarrow |T| > t_{\left(n-2,1-\frac{\alpha}{2}\right)} \Leftarrow H_1 : B_1 \neq 0$

۲- حال اگر فرض صفر به این صورت باشد که:

$$\begin{cases} H_0 : \beta_1 = \beta_1^{(0)} \\ H_1 : \beta_1 ? \beta_1^{(0)} \end{cases}$$

دقیقاً تمام ناحیه‌های رد برای فرض‌های مختلف H_1 مثل بالاست تنها با این تفاوت که آماره آزمون عبارتست از:

$$T = \frac{\hat{\beta}_1 - \beta_1^{(0)}}{\sqrt{MSE / S_{xx}}}$$

آزمون فرض برای β_0 مانند همان β_1 است فقط $T^{(*)}$ را باید محاسبه کنیم:

$$T^{(*)} = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{MSE\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{xx}}\right)}}$$

قوانین مهم

اگر فرض $B_1 = 0$ پذیرفتیم یعنی شیب خط صفر است $\Leftarrow$ x در مدل نیست $\Leftarrow$ مقدار X تأثیری در پیشگوئی y ندارد $\Leftarrow$ بهترین برآوردگر y همان $\bar{Y}$ است $\Leftarrow$ ارتباط خطی بین X و Y نیست. (اما ممکن است ارتباط بین X و Y در درجات بالاتر وجود داشته باشد.) اما اگر $B_1 \neq 0$ پذیرفتیم $\leftarrow$ X در مدل هست $\leftarrow$ بین X و Y ارتباط خطی وجود دارد.

مدل بدون عرض از مبدأ $(\beta_0 = 0)$

$$Y_i = B_1 X_i + E_i$$

$$\hat{\beta}_1 = \frac{\Sigma X_i Y_i}{\Sigma X_i^2} \qquad \sigma^2 = MSE = \frac{SSE}{n-1} \quad , \quad Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\Sigma X_i^2}$$

که در این حالت اگر بخواهیم فاصله اطمینان برای β_1 بسازیم، داریم:

$$\left(\hat{\beta}_1 - t_{\left(n-1,1-\frac{\alpha}{2}\right)}\sqrt{\frac{MSE}{\Sigma X_i^2}}, \hat{\beta}_1 + t_{\left(n-1,1-\frac{\alpha}{2}\right)}\sqrt{\frac{MSE}{\Sigma X_i^2}}\right)$$

آزمون پارامترها در رگرسیون

تحت فرض $H_0 : \beta_1 = 0 \leftarrow F = \frac{MRS}{NSE} \sim F_{(1,n-2)}$

$$SSR = \hat{\beta}_1^2 S_{XX}$$

منبع تغییر	df	SS	MS	F
رگرسیون	1	SSR	MSR	$F_{\circ}=\frac{MSR}{MSE}$
خط	n-2	SSE	MSE	
جمع	n-1	SSY		

$F_{\circ} > F_{(۱,\ n-۲,\ \alpha)} \Rightarrow$ رد $H_{\circ}$

تحلیل ماتریسی رگرسیون خطی ساده

$$Y = XB + E$$

$$B = \left(X'X\right)^{-۱} X'Y$$

$$MSE = \frac{۱}{n-۲} Y' \left[I - \underbrace{X\left(X'X\right)^{-۱} X'}_{H} \right] Y$$

$$\hat{E} = Y - \hat{Y} = Y - X\hat{B} = \left[I - H\right]Y$$

$$X = \begin{bmatrix} ۱ & X_{۱} \\ . & . \\ . & . \\ ۱ & X_{n} \end{bmatrix} \qquad B = \begin{bmatrix} B_{\circ} \\ B_{۱} \end{bmatrix}$$

رگرسیون چندگانه

$$Var\left(e_i\right) = \sigma^{۲}\left(۱ - P_{ii}\right)$$

$$\left(\hat{y}_i\right) = \sigma^{۲} P_{ii}$$

نکات طلایی

- می‌دانیم $\hat{\beta}_{۱} = \frac{S_{XY}}{S_{XX}}$ لذا چنانچه مقادیر x را در n ضرب کنیم. شیب خط، $\frac{۱}{n}$ شیب خطی اولیه است.

فرمول‌هایی که باید به خاطر بسپارید:

$$Cov\left(\bar{Y}, \hat{Y}_i\right) = Var\left(\bar{Y}\right)$$

$$SSY = SSR + SSE$$

$$\sum_{i=۱}^{n}\left(Y_i - \bar{Y}\right)^{۲} = \sum_{i=۱}^{n}\left(\hat{Y}_i - \bar{Y}\right)^{۲} + \sum_{i=۱}^{n}\left(Y_i - \hat{Y}_i\right)^{۲}$$

آخرین فرمول نوشته شده در بالا به ترتیب از چپ نشان‌دهنده‌ی:
میزان تغییرات توضیح داده نشده توسط مدل- میـزان تغییـرات توضـیح داده شده توسط مدل= میزان کل تغییرات

ضریب تعیین

$$\text{نسبت تغییرات} \leftarrow r^2 = \frac{SSR}{SSY} = \frac{SSY - SSE}{SSY} \Rightarrow$$

- بهترین مدل $r^2 = 1$ یعنی $SSE = 0$

$$\text{مدل مناسب نیست} \Leftarrow SSE \cong SSY \leftarrow r^2 \cong 0$$

$$SSR = \hat{\beta}_1^2 S_{XX} = r^2 S_{YY} \text{ و } SSE = SSY - r^2 S_{YY}$$

$$\text{نکته} \Rightarrow \rho_{Y_i, \hat{e}_i} = \sqrt{1 - r_{XY}^2}$$

نکته: آزمون $H_0 : \rho = 0$ در برابر $H_1 : \rho \neq 0$ معادل با آزمـون $H_1 : \beta_1 = 0$ در برابر $H_1 : \beta_1 \neq 0$ می‌باشد.

نکته

برای آزمون: $H_0 : \rho = 0$ داریم $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t$

و برای آزمون $H_0 : \rho = \rho_0$ داریم:

$$Z^* = \frac{\frac{1}{2} ln \frac{1+r}{1-r} - \frac{1}{2} ln \frac{1+\rho_0}{1-\rho_0}}{\frac{1}{\sqrt{n-3}}}$$

- که برای استفاده از ایـن آمـاره آزمـون لازم اسـت x وy دارای توزیـع نرمال باشد.

نکته مهـم: اگر شیب خط رگرسیون Y به X در نمونه‌ای برابر β_1 و شـیب خـط رگرسیون X و Y برابر β_2 باشد، آنگاه r^2 از رابطه‌ی زیر به دست می‌آید:

$$r^2 = \beta_1 \beta_2$$

$$r_{Y,e}^2 = 1 - r_{XY}^2$$

رگرسیون چندگانه: مدل رگرسیونیست که شامل بیش از یک متغیر مستقل باشد.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n K_n$$

جدول رگرسیون: هنگامی که k پارامتر داشته باشیم.

	df	SS	MS	F
رگرسیون	K	SSR	MSR	
خط	n-k-1	SSE	MSE	$F_{\circ}=\frac{MSR}{MSE}$
مجموع	n-1	SSY		

نکات طلایی

- ضریب همبستگی با شیب خط (β_1) هم علامت است.
- در رگرسیون خطی ساده مربع ضریب همبستگی ساده برابر با ضریب تعیین است.
- خط رگرسیونی همواره از $(\bar{x},\bar{y})$ می‌گذرد.
- در توزیع نرمال ۲ متغیره داریم:

$$\text{رگرسیون Y به X} \Rightarrow E(y|x)$$

- اگر εiها (خطاها) از توزیع نرمال باشند آنگاه برای هر زوج $\varepsilon i, \varepsilon j, i \neq j$ از هم مستقلند اما اگر شرط نرمال بودن εiها را نداشتیم، تنها ناهمبسته می‌شوند.

ضریب همبستگی

ضریب همبستگی ابزاری آماری برای تعیین نوع و میزان رابطه خطی یک متغیر کمی با متغیر کمی دیگر است. ضریب همبستگی، یکی از معیارهای مورداستفاده در تعیین همبستگی دو متغیر است. ضریب همبستگی شدت رابطه و همچنین نوع رابطه (مستقیم یا معکوس) را نشان می‌دهد. این ضریب بین ۱ تا ۱- است و در صورت عدم وجود رابطه بین دو متغیر، برابر صفر است:

ضریب همبستگی و استقلال دو معنای متفاوت دارند.

$$\rho=\frac{cov(x,y)}{\sqrt{var(x)var(y)}}$$

آزمون معناداری ρ: وقتی می‌خواهیم آزمون کنیم که آیا $\rho=\circ$ هست یا خیر:

$$\begin{cases}H_{\circ}:\rho=\rho_{\circ}\\ H_{1}:\rho\neq\rho_{\circ}\end{cases}\xrightarrow{\text{ناحیه بحرانی}}|T|=\left|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right|>t_{\frac{\alpha}{2},n-2}$$

که در این صورت $H_{\circ}$ رد می‌شود.

نکات طلایی

- $e_i \sim N(۰, \sigma^۲)$
- $R^۲ = \frac{SSR}{SST} = ۱ - \frac{SSE}{SST}$
- $\hat{Y} - \bar{Y} = \hat{B}_۱(X - \bar{X})$
- $Cov(\bar{Y}, \hat{\boldsymbol{\beta}}_۱) = ۰ \rightarrow$ همبستگی خطی وجود ندارد
- $r = \frac{B_۱ S_x}{S_y} = B_۱ \sqrt{\frac{S_{xx}}{S_{yy}}} = B_۱ \sqrt{\frac{S^۲x}{S^۲y}}$
- $\sum e_i^۲ = SEE$
- $\sigma^۲ = \frac{۱}{n}\sum(X_i - \bar{X})^۲ \qquad S^۲ = \frac{۱}{n-۱}\sum(x_i - \bar{x})^۲$

$$\rho_{ax+b,\, cy+d} = \frac{ac}{|ac|}\rho_{x,y}$$

$r = \pm ۱ \Leftarrow$ شرط لازم و کافی برای این رابطه اینست که تمام نقـاط روی خط باشند.

از طرفـی پـس r و B_1 هـم علامـت بـا هـم هسـتند، اگـر B_1 منفـی باشـد همبستگی نیز منفی شود. $\Rightarrow r = B_۱ \frac{S_x}{S_y}$

$$\hat{\beta}_۱ = r\sqrt{\frac{s^۲y}{s^۲x}} = r\sqrt{\frac{S_{yy}}{S_{xx}}}$$

ضریب همبستگی اسپیرمن:

$$r_s = ۱ - \frac{۶\sum d_i^۲}{n(n^۲ - ۱)},\ d_i \rightarrow \text{اختلاف رتبه‌ها}$$

همواره به یاد داشته باشید:

در مدل ساده خطی $\Rightarrow Y_i = B_۰ + B_۱ x_۱ + \varepsilon_i$

مقدار آماره آزمون F در رگرسیون چندگانه این $\Rightarrow F = \frac{n-k-۱}{k}\frac{R^۲}{۱-R^۲}$ گونه به دست می‌آید.

در مدل رگرسیون خطی ساده $F = (n-۲)\frac{R^۲}{۱-R^۲}$