لونا شجاع الاسلامی - تکلیف سری چهارم
۸۱۰۱۰۰۲۹۴ - استنباط آماری

① درست یا غلط؟ توضیح دهید.

ⓐ غلط. به طور یقینی نمی‌توان در این باره نظر داد چون اگر dataset جدید outlierها بیشتری داشته باشد در آن صورت خطای train می‌تواند بیشتر شود؛ پس میانگین خطای train به مقادیری که با آن train می‌شود وابسته است.

ⓑ درست. طبق تعریف داریم $e_i = Y_i - X_i^T \hat{\beta}$ رابطه‌ی residualها

با استفاده از regression خطی از رابطه‌ی (*) تخمین زده می‌شود:

$$(*)\begin{cases} \hat{\beta} = (X^T X)^{-1} X^T Y \\ X = [X_i^T]_{N \times n} \\ Y = [Y_i]_{N \times 1} \end{cases}$$

n: (تعداد predictorها) — N: تعداد داده‌ها

(از راست در $X^T X$)

$$X^T X \hat{\beta} = X^T Y \rightarrow X^T (Y - X\hat{\beta}) = 0 \Longrightarrow Y - X\hat{\beta} = 0$$

$$\Longrightarrow \sum_{i=1}^{N} Y_i - X_i^T \hat{\beta} = \sum_{i}^{N} e_i = 0$$ ✓

پس در این روش همیشه میانگین residualها صفر است.

ⓒ غلط. استانداردسازی (features) فقط به مدل‌های regression در بهینه‌سازی کمک می‌کند و انجام آن پیش از train یک مدل logistic regression الزامی نیست.

ⓓ درست. $\rho = 0$ (correlation coefficient) نشان‌دهنده‌ی رابطه‌ی خطی میان دو متغیر است که در صورت صفر شدن نفی ارتباط خطی بین دو متغیر وجود ندارد اما ممکن است که بین دو متغیر (رابطه‌ای غیرخطی) وجود داشته باشد.

ⓔ غلط. تابع هزینه در ridge regression:

$$J = e^T e + \lambda \beta^T \beta$$

داریم:

اگر predictorها نتوانند خطای SS ع را کاهش دهند با اضافه شدن ترم regularization، مقدار ضریب predictor کاهش یافته و تأثیرش در تخمین کم می‌شود. اگر λ را تا (∞) افزایش دهیم درست است که برخی از ضرایب موجود در رابطه به صفر نزدیک می‌شوند (صفر جزئی $\rightarrow 0$) اما به طور مطلق به صفر نمی‌رسند.

ⓕ غلط. $R^2 = 0 \longrightarrow$ رابطه‌ی خطی بین دو متغیر وجود ندارد. $\not\Longrightarrow$ هیچ رابطه‌ای (مانند رابطه‌ی غیرخطی) بین دو متغیر وجود ندارد. پس محتمل است رابطه‌ی دیگری (غیرخطی) بین دو متغیر وجود داشته باشد.

ⓖ غلط. $R^2 = 0.791 \Longrightarrow |R| \simeq 0.89$ (در صورت سؤال گفته 0.88) جهت رابطه‌ی خطی دچار مشکل است یعنی ما با R^2 فقط می‌توانیم بگوییم رابطه‌ی خطی داریم یا نه.

ادامه‌ی سؤال ①

و) و بنابراین می‌توانیم بدون توجه به جهت بگوییم قدرت رابطه‌ی خطی چقدر است. بین وجود ترم مثبت (رابطه‌ی خطی مثبت)

با توجه به این موضوع دقیق نیست.

ⓗ درست.

طبق تعریف برای اسلاید residual ها داریم:

$e_i = y_i - \hat{y}_i$

y_i → مقدار مشاهده شده (observed)

$\hat{y}_i$ → مقدار تخمین (reggresion value)

پس طبق تعریف اسلاید residual ها در حقیقت فاصله‌ی مقادیر واقعی و تخمینی اند ✓

---

d) $Y = A_0 + A_1 x$ (طبق گراف داده شده)

x → adornment

Y → evaluation score

	estimate	std.error	T-value	$P_r\{>H1\}$
intercept	4.010 (a_0)	0.025	157.21	0
adornment	0.1325	0.032	4.141	0

②

$\hat{y} = a_0 + a_1 x$

$a_1 = \frac{S_y}{S_x} R$

$a_0 = \bar{y} - a_1 \bar{x}$

$\bar{x} = -0.0833$

$\bar{y} = 3.9983$

داریم → $a_1 = 0.1325$

$T = \frac{a_1 - 0}{S_{Ea_1}} = \frac{0.1325}{0.032} = 4.141$

b) مطابق درس داریم:

$$\begin{cases} H_0: a_1 = 0 \\ H_1: a_1 > 0 \end{cases} \quad T = 4.141 \quad df = n-2 = 461$$

یک آزمون t test

$$P_{value} = Pr\{t > 4.141\} < 0.05 \Rightarrow$$ فرض صفر رد می‌شود $(\alpha = 0.05)$

یعنی از لحاظ آماری شواهد کافی proof برای فرض دیگر وجود دارد (رابطه‌ی بین score و evaluation, adornment مثبت است.)

c) $95\% \rightarrow \alpha = 0.05 \quad t^*_{461} = 1.96$

$$ME = a_1 \pm 1.96 s_{a_1} \rightarrow (0.06978, 0.1952)$$

بازه a_1

با توجه به بازه به دست آمده برای a_1 فرض صفر رد می‌شود. در اینجا نتایج hypothesis test با نتایج confidence interval یکسان بود و هر دو روش فرض صفر را رد کردند.

a)

$$\hat{Y} = a_0 + a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5 + a_6x_6$$

$Y \triangleq$ blood pressure (bp)

		estimate	stderror	tvalue	Pr{>|t|}
	(intercept)	-80.41	14.53	-5.6	0.0
x_1	Smoke	0.44	0.03	15.26	0.0
x_2	Exercise hours	-3.33	1.13	-2.95	0.0033
x_3	age	-0.01	0.09	-0.1	0.9170
x_4	height	1.15	0.21	5.63	0.0
x_5	weight	0.05	0.03	1.99	0.0471
x_6	water consumption	-8.40	0.95	-8.81	0.0

$$\hat{Y} = -80.41 + 0.44x_1 - 3.33x_2 - 0.01x_3 + 1.15x_4 + 0.05x_5 - 8.4x_6$$

داریم

b) با توجه به Pvalue هر یک از متغیرهای آزمون گرها (predictors) داریم:

for smoke $\rightarrow$ Pvalue $\simeq 0 \rightarrow$ یعنی یک عضو هم در تخمینات ما فرض است و اگر یک عضو دیگر به نظر گرفته شود با افزایش هر یک واحد حجم smoke داده فشار خون به طور متوسط 0.44 افزایش می‌یابد. (طبق مدل تخمینی)

for water consumption $\rightarrow$ Pvalue $\simeq 0 \rightarrow$ یک عضو هم در تخمینات ما فرض است و اگر یک عضو دیگر ثابت بماند با افزایش هر یک واحد حجم مصرف آب فشار خون 8.40 - واحد کاهش می‌یابد. (مطابق مدل)

بازه‌ی اطمینان 95٪

$$t^*_{1229} = 1.962 \rightarrow a_1 \pm t^*_{1229} \cdot s_{a_1} = (0.3811, 0.4989)$$

$\alpha = 0.5$

$df = n-k-1 = 1229$

با احتمال 95٪، a_1 بین بازه مقداری گیرد و رابطه مثبت bp دارد. ⊕

$$a_6 \pm t^*_{1229} s_{a_6} = (-10.264, -6.536)$$

$\alpha = 0.5$

$df = 1229$

با احتمال 95٪، a_6 بین این بازه می‌رود پس رابطه منفی دارد. ⊖

c)

$$R^2 = \frac{SSG}{SST} = 1 - \frac{SSR}{SST} = 1 - \frac{249.28}{332.57} = 1 - 0.75 = 0.25$$

$$R^2_{adj} = 1 - \left(\frac{SSR}{SST} \times \frac{n-1}{n-k-1}\right) = 1 - \left(0.75 \times \frac{1235}{1229}\right) \simeq 0.247$$

④ تست ANOVA

اختلاف میانگین سه گروه را با هم مقایسه می‌کنیم: (good, bad, medium)

	Df	Sum sq	Mean sq	F value	Pr(>F)
class	2	1144	572		0.002326
Residual	[redacted] 425	39518	93		

a) داریم:
$$\begin{cases} H_0 : \mu_1 = \mu_2 = \mu_3 & \text{میانگین همه گروه‌ها برابر} \\ H_1 : \text{دست کم یک گروه میانگین متفاوتی دارند (با دو گروه دیگر)} \end{cases}$$

b) $df_c = K-1 = 3-1 = 2$ (class df_c)

$df_t = n-1 = 428-1 = 427 \rightarrow df_r = 427-2 = 425$

$MSG = \frac{SSG}{df_c} = 572 \rightarrow SSG = 1144$

$MSE = \frac{SSE}{df_e} = \frac{39518}{425} \simeq 93$

$P_r\{F > 6.15\} = 0.002326 < 0.01$ پس H_0 رد می‌شود

و H_1 درست است یعنی دست کم بین دو گروه تفاوت دیده می‌شود.

c) $\alpha^* \triangleq \frac{\alpha}{K}$ (significance level) ← دلیل کاهش خطای نوع یک

تعداد t-test ها جفت گروه‌ها (K=3) $\rightarrow \alpha^* = \frac{0.01}{3} \simeq 0.003$

$df = 425 \rightarrow t^*_{df} = 2.95$

t-test بین دو گروه good و Medium

داریم: $SE = \sqrt{\frac{MSE}{n_1} + \frac{MSE}{n_2}}$ ($MSE = 93$, $n_1 = 200$, $n_2 = 150$) $\rightarrow SE \simeq 1.042$

$$\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 \neq 0 \end{cases}$$

$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} SE = (4.926, 11.074)$ → پس چون احتمال ۹۹٪ درباره صفر نداریم فرض صفر H_0 رد می‌شود یعنی بین دو گروه تفاوت داریم $(\mu_1 \neq \mu_2)$

⑤

b) اگر همه مشاهدات از یک گروه مثلن به دست بیاید، مدل خطی معمولی مناسب است ولی اگر یک data structure داشته باشیم که مشاهدات در هر گروهی یا مشاهدات درون یک زیر گروه دیگر از داخل جامعه بستگی داشته باشد در این حالت از random effect استفاده می‌کنیم. از random effect زمانی استفاده می‌شود که تعداد زیادی سطح (block) داده در هر سطح مختلف داریم؛ اگر تعداد سطوح بیشتر شود از Multi-level Model استفاده می‌کنیم. در random effect فرض می‌کنیم که predictor، توزیع نرمال دارد.

a) دقت نتایج، مفروضات تست‌ها بستگی زیادی دارد. یکی از این مفروضات که در همه تست‌ها هست فرض استقلال داده‌هاست. وجود کوچکترین وابستگی میان داده‌ها باعث وجود اریبی در نتایج تست می‌شود. در این حالت داده‌های غیرمستقل، paired sample یا repeated measures data — تست در این موارد فقط زوج مشاهدات می‌توانند مستقل باشند چون بر اساس ساختار داده به صورت جفت است. cross-classified data و nested data از موارد ساختارهای پیچیده‌ای هستند که شرط استقلال را ندارند.

6) الف) فرمول چولگی (skewness) ← معیاری برای بخش تقارن داده

$$\tilde{\mu}_3 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N_s^{\,3}}$$

معیار kurtosis در واقع heavy-tailed / light-tailed بودن داده نسبت توزیع نرمال را نشان می‌دهد.

$$\tilde{\mu}_4 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^4}{N_s^{\,4}}$$

---

7)

a) $\bar{x} = \frac{43 + 59 + \dots + 21}{8} = 34$, $\sigma^2 = \frac{(43 - 34)^2 + \dots}{8}$

$\frac{1498}{8} = \boxed{187.25} \longrightarrow sd = \sigma = \sqrt{187.25} = \boxed{13.6839}$

---

b) بله. در bootstrapping، جهت سوجود را با جای‌گذاری باز نمونه‌برداری می‌کنیم که در این جا چون این روش اطلاعات جدیدی ایجاد نمی‌کند، اگر در تمام داده‌های ضد (بالایی) وجود داشته باشد (زمان‌ها دارای صفر باشند) [illegible] bootstrapping کمک نمی‌کند.

---

c) با توجه به داده‌ها انتظار این است که به سمت راست متمایل باشد چون داده‌های کوچک باعث می‌شوند که بطور شدیدی داده‌ها کوچک‌تر نمونه شوند به علاوه اینکه bootstrap ویژگی‌های توزیع را حفظ می‌کنند توزیع داده‌ی اصلی را تمایل به راست (چولگی به راست) داریم که بعد از bootstrap هم حفظ می‌شود.

---

d) زمان خریدنان (در صف بودن و...) که از نظر توزیع نیز شکل آن تقریباً مانند توزیع uniform است.

---

e) $CI: \bar{x} + 1.66 \times 4.85 = 34 \pm 8.05 = (25.95, 42.05)$