به نام خداوند جان و خرد

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

استنباط آماری

تمرین شماره ۴ – نسخه ۱ - بدون سوالات برنامه‌نویسی

نام و نام خانوادگی: علی خرم فر

شماره دانشجویی: ۸۱۰۱۰۲۱۲۹

بهمن‌ماه ۱۴۰۲

فهرست مطالب

۱_ پاسخ مسئله شماره ۱

برنامه نویسی تحویل بعد از امتحان

۲_پاسخ مسئله شماره ۲

در این مسئله هدف مقایسه ۴برند مختلف باتری است و مطلوب سوال وجود اختلاف در میانگین طول عمر باتری‌ها در هر گروه است. هدف بررسی واریانس داخلی و خارجی است تا بتوان حدود میانگین هرکدام را متوجه شد و با همدیگر مقایسه کرد. با توجه به اینکه میانگین بیش از ۲ گروه مقایسه می‌شود، پس از تست ANOVA باید استفاده کرد.

در این تست تاثیر متغیر Categorical بر Numerical بررسی می‌شود. که در این سوال هدف مقایسه بین برندهای مختلف باتری است. برای این تست داده‌ها باید از توزیع نرمال باشند که در این سوال این مورد رعایت شده‌است. همچنین باید واریانس گروه‌ها حدودا برابر باشد که این شرط نیز برقرار است. برندها نیز مستقل از هم هستند.

فرض صفر بر این اساس است که میانگین طول عمر تمامی برندها یکسان است و فرض مقابل نیز عکس فرض صفر خواهدبود. یعنی فرض یک بر این اساس است که حداقل دو گروه میانگین متفاوتی دارند.

$H0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

$H1: atleast\ 2\ groups\ have\ diffrent\ means$

برای این تست باید test statistics زیر که بر پایه توزیع F است محاسبه شود. در این آماره نسبت اختلاف بین گروهی به اختلاف درون گروهی محاسبه می‌شود. به این منظور Sum Squarel اختلاف‌ها محاسبه شده و تقسیم بر درجه آزادی آن‌ها می‌شود.

$$F = \frac{\frac{SS_B}{I-1}}{\frac{SS_W}{I(J-1)}}$$

در این مثال I که برابر تعداد برندهاست برابر است با ۴ و J که تعداد نمونه هر برند است برابر با ۵ است.

$dfb = I - 1 = 3$

$dfw = I(J-1) = 16$

برای تشکل جدول *ANOVA* باید *SST* و *SSb* محاسبه شده و سپس از روی این مقادیر *SSw* محاسبه شود.

	DF	Sum Sq	Mean Sq	F value	P(>F)
Between Groups	I-1	SSb			
Within Groups	I(J-1)	SSw			
		SST			

حال به محاسبه پارامترهای مورد نظر پرداخته می‌شود:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

که در این فرمول y_i برابر با مقدار مشاهده شده و $\bar{y}$ برابر میانگین کل است.

$$SSb = \sum_{j=1}^{n} nj\left(\bar{y_j} - \bar{y}\right)^2$$

که در فرمول بالا $\bar{y_j}$ برابر با میانگین گروه و $\bar{y}$ برابر با میانگین کل است.

$$SSw = SST - SSb$$

در این مسئله مقادیر میانگین برندها به شرح زیر است:

$\bar{y}total = 30.7$	D	C	B	A
$\bar{y}\ of\ brand$	28.6	29.8	30.8	33.6

$$\begin{aligned} SST = {} & (42-30.7)^2 + (30-30.7)^2 + (39-30.7)^2 + (28-30.7)^2 + (29-30.7)^2 \\ & + (28-30.7)^2 + (36-30.7)^2 + (31-30.7)^2 + (37-30.7)^2 \\ & + (27-30.7)^2 + (24-30.7)^2 + (36-30.7)^2 + (28-30.7)^2 \\ & + (28-30.7)^2 + (33-30.7)^2 + (20-30.7)^2 + (32-30.7)^2 \\ & + (38-30.7)^2 + (28-30.7)^2 + (25-30.7)^2 = 598.2 \end{aligned}$$

$$SSb = 5(28.6-30.7)^2 + 5(29.8-30.7)^2 + 5(30.8-30.7)^2 + 5(33.6-30.7)^2 = 68.2$$

$$SSw = SST - SSb = 598.2 - 68.2 = 530$$

حال با مقادیر بالا توزیع f مربوطه تشکیل می‌شود:

$$F = \frac{\frac{SS_B}{I-1}}{\frac{SS_W}{I(J-1)}} = \frac{\frac{68.2}{3}}{\frac{530}{16}} = 0.686$$

که توزیع بالا درجه آزادی ۳ و ۱۶ دارد. مقدار احتمال توزیع در این نقطه برابر است با 0.57358

در این مسئله مقدار significance level برابر با 0.05 است. برای توزیع f با درجه آزادی بالا مقدار برابر با 3.23 دارد. که مقدار بدست آمده از f از این مقدار کمتر است.

پس باتوجه به اینکه مقدار احتمال f از مقدار 0.05 بیشتر است فرض صفر رد نمی‌شود و می‌توان نتیجه گرفت که میانگین‌ها اختلاف قابل ملاحظه‌ای با یکدیگر ندارند.

	DF	*Sum Sq*	*Mean Sq*	*F value*	*P(>F)*
Between Groups	*3*	68.2	*22.7*	*0.686*	*0.57358*
Within Groups	*16*	530	*33.125*		
Total		598.2			

۳_ پاسخ مسئله شماره ۳

در این مسئله یک جدول ANOVA برای مقایسه بین داروهای مختلفی که فاکتور کاهش فشار خون روی آن‌ها آزمایش شده نمایش داده شده‌است. پس ۳ گروه در این مسئله بررسی می‌شود. اختلافی که بین آن‌ها وجود دارد دو نوع است. بین گروهی و درون گروهی. که درون گروهی به عوامل متعددی بستگی دارد از جمله سن و سابقه بیماری. حال به بررسی مطلوب مسئله پرداخته می‌شود.

۳-۱_ پاسخ قسمت ۱

چارچوب آزمون فرض به نحوی است که فرض صفر بیان می‌کند که اختلافی بین میانگین تاثیر هر ۳ دارو وجود ندارد. و فرض مقابل آن بیان می‌کند که حداقل دو گروه از داروها با یکدیگر متفاوت هستند.

$H0: \mu_1 = \mu_2 = \mu_3$

$H1: atleast\ 2\ groups\ have\ diffrent\ means$

۲-۳_ پاسخ قسمت ۲

باتوجه به اینکه مقدار احتمال توزیع f بدست آمده از مقدار 0.05 کم‌تر است پس فرض صفر رد می‌شود. همچنین باتوجه به اینکه مقدار f برای ۰.۰۵ و درجه آزادی ۲ و ۳۹ برابر با ۳.۲۳ است و مقدار بدست آمده در جدول ۳.۳۳ است هم می‌توان فرض صفر را رد کرد.

این موضوع نشان می‌دهد که میانگین داروها برابر نبوده و حداقل ۲ دارو با یکدیگر تفاوت دارند.

۳-۳_ پاسخ قسمت ۳

باتوجه به اینکه تست ANOVA مشخص نمی‌کند که کدام دو گروه باهم متفاوت هستند پس می‌توان از Multiple Comparison استفاده کرد. یکی از این روش‌ها روش Bonferroni است که مشکل مربوط به تجمع significance level ها را در مقایسه‌های چندتایی حل می‌کند. در این روش ابتدا بازه اطمینان برای مقایسه میانگین گروه‌ها مشخص می‌شود:

$$(\overline{Y_{\iota 1}} - \overline{Y_{\iota 2}}) \pm s_p \frac{t_{I(J-1)}(\alpha^*)}{J}$$

که α^* همان مقدار تصحیح شده توسط روش Bonferroni است:

$$K = \frac{I(I-1)}{2} = 3$$

$$\alpha^* = \frac{\alpha}{K} = \frac{0.25}{3} = 0.083 \rightarrow t_{123}(0.083) = 1.394$$

مقدار تخمین انحراف معیار $pool$شده نیز از رابطه زیر بدست می‌آید:

$$SS_w = 3740.43$$

$$s_p = \sqrt{\frac{SS_w}{I(J-1)}} = \sqrt{\frac{3740.43}{123}} = 5.51$$

$$(\overline{Y_{\iota 1}} - \overline{Y_{\iota 2}}) \pm 5.51 \frac{1.394}{42} \rightarrow (\overline{Y_{\iota 1}} - \overline{Y_{\iota 2}}) \pm 0.182$$

یعنی میانگین‌هایی که اختلاف بیش از مقدار بالا دارند، به طور قابل ملاحظه‌ای میانگین متفاوت دارند. در این مسئله داروها همگی دو به دو باهم اختلافی بیش از مقدار بالا داشته پس فرض صفر رد شده و نتیجه می‌شود که اثربخشی آن‌ها به صورت قابل ملاحظه‌ای با یکدیگر متفاوت است.

۴-۳_ پاسخ قسمت ۴

همانطور که در قسمت قبلی هم گفته‌شد، این روش باعث می‌شود که مشکل تجمع significance levelها در multiple comparison حل شود و مشخص می‌کند که کدام دو گروه با یکدیگر متفاوت بوده که باعث رد فرض صفر شده است. که این مورد در تست ANOVA قابل دسترس نبود. همچنین این تست FWER را کنترل می‌کند و خطای نوع ۱ کاهش می‌یابد.

از مشکلات این تست می‌توان به این موارد اشاره کرد که احتمال خطای نوع ۲ در این تست افزایش پیدا می‌کند. همچنین کارایی این رویکرد زمانی که تعداد مقایسه‌ها زیاد باشد پایین می‌آید. همچنین فرض این تست مستقل بودن تست‌ها از یکدیگر است.

۴_پاسخ مسئله شماره ۴

در این مسئله Pvalue های مختلف از مقایسه‌های چندتایی بین گروه‌ها به صورت زیر ارائه شده:

P-values : 0.361, 0.387, 0.005, 0.009, 0.022, 0.051, 0.101, 0.019

۴-۱_ پاسخ قسمت ۱

در این قسمت از مسئله برای کنترل FDR از روش BH استفاده می‌شود. در این روش ابتدا مقادیر Pvalue مرتب می‌شوند. سپس رنک بزرگترین pvalue که در رابطه $P(r) \leq \frac{qr}{n}$ صدق کند پیدا می‌کنیم. q برابر با pvalue اصلاح شده‌است. سپس فرض صفرهایی که مقدار $pvalue$ رنک شده آن‌ها کمتر از r است رد می‌شوند. به این صورت مقدار FDP که برابر با نسبت فرض صفرهای به اشتباه رد شده به کل فرض صفرهای رد شده است کنترل می‌شود.

تعداد $pvalue$ها پس از مرتب سازی که کمتر از α هستند برابر است با $R(\alpha)$

همچنین Pvalue امr برابر است با

$$\alpha = P(r)$$

حد بالای تست‌هایی که به اشتباه رد شده‌اند:

$$\frac{P(r) * n}{r} < q \rightarrow P(r) < \frac{q.r}{n}$$

پس برای هرکدام از آن‌ها مقدار q که همان pvalue اصلاح‌شده است محاسبه می‌شود.

Pvalue های مرتب شده:

0.005,0.009,0.019,0.022,0.051,0.101,0.361,0.387

در این مسئله ۸ تست انجام شده که n=8 است. مقدار control value نیز برابر با ۵ درصد است و مشاهده می‌شود که در مقادیر بالا ۴ مورد از آن کمتر است. حال مشاهده می‌کنیم که روش BH چه تغییری در مقادیر بالا انجام می‌دهد.

$R(\alpha) = 4$

$q = 0.05$

$Adjusted\ p = q.\frac{r}{n}$

$BH\ Critial\ Value = \frac{q * r}{n}$

8. $\frac{q * r}{n} = \frac{0.05 * 8}{8} = 0.05 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.387 * 8}{8} = 0.387$

7. $\frac{q * r}{n} = \frac{0.05 * 7}{8} = 0.043 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.361 * 8}{7} = 0.412$

6. $\frac{q * r}{n} = \frac{0.05 * 6}{8} = 0.037 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.101 * 8}{6} = 0.135$

5. $\frac{q * r}{n} = \frac{0.05 * 5}{8} = 0.031 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.051 * 8}{5} = 0.102$

4. $\frac{q * r}{n} = \frac{0.05 * 4}{8} = 0.025 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.022 * 8}{4} = 0.044$

3. $\frac{q * r}{n} = \frac{0.05 * 3}{8} = 0.018 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.019 * 8}{3} = 0.051$

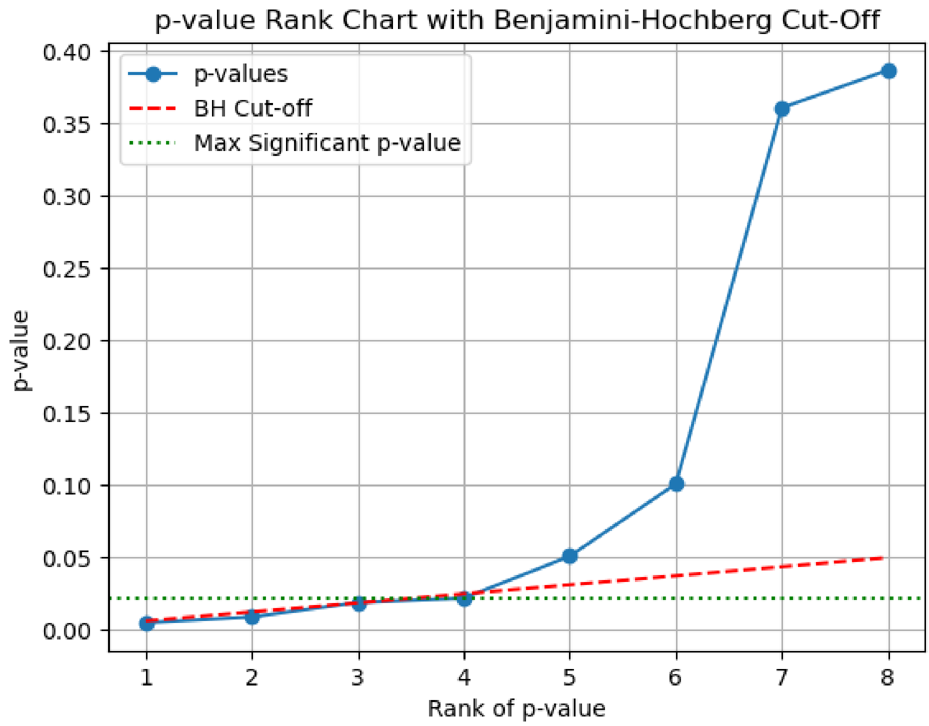
2. $\frac{q * r}{n} = \frac{0.05 * 2}{8} = 0.01 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.009 * 8}{2} = 0.036$

1. $\frac{q * r}{n} = \frac{0.05 * 1}{8} = 0.006 \rightarrow\ adjusted\ pvalue = \frac{P(r)n}{r} = \frac{0.005 * 8}{1} = 0.04$

Pvalue	*0.005*	*0.009*	*0.019*	*0.022*	*0.051*	*0.101*	*0.361*	*0.387*
Rank	1	2	3	4	5	6	7	8
BH Critial Values	0.006	0.01	0.018	0.025	0.031	0.037	0.043	0.05
Adjusted p	0.04	0.036	0.051	0.044	0.102	0.135	0.412	0.387
Reject H0?	1	1	1	1	0	0	0	0

در این مسئله از آنجا که0.025 بزرگ تر از ۰.۰۲۲است پس pvalue هایی که کمتر از ۰.۰۲۲ هستند significance می‌شوند.

۲-۴_ پاسخ قسمت ۲



۳-۴_ پاسخ قسمت ۳

در روش FWER که یک روش محافظه‌کار است، حتی اگر ۱ فرض صفر به اشتباه رد شود، حساسیت وجود دارد. اما اگر لیبرال‌تر عمل شود، در روش FDR فقط نسبت فرض صفرهایی که به اشتباه رد شده به نسبت کل فرض صفرها مد نظر است و از این نظر حساسیت کمتری دارد. برای مثال در FWER با کمک روش Bonferroni مقدار Significance level بسته به تعداد کل آزمون‌ها اصلاح می‌شود.در این روش مقدار خطای نوع ۱ کاهش می‌یابد. این برای زمانی است که قصد بر این است که فرض صفر حفظ شود و رد نشود. ولی در روشی مثل BH که زیرمجموعه FDR است، برای زمانی بهتر است که مبنا بر رد فرض صفر باشد. همچنین این روش برای زمانی که تعداد مقایسه‌ها زیاد باشد مناسب‌تر است.

۵_پاسخ مسئله شماره ۵

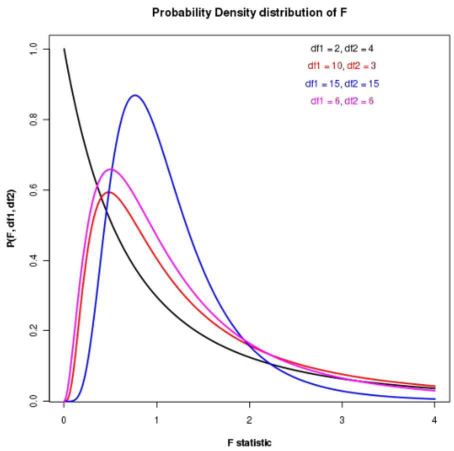
1) If the number of groups increases, then the type 1 error increases in multiple comparisons tests, so the corrected significance level should increase.

غلط - قسمت اول که بیان می‌کند اگر که تعداد گروه‌ها افزایش پیدا کرد، خطای نوع ۱ افزایش پیدا می‌کند صحیح است زیرا که مقدار Significance level جمع شده و تجمیع آن باعث افزایش خطای نوع ۱ می‌شود. ولی قسمت دوم که بیان می‌کند مقداد آن باید زیاد شود اشتباه است و برعکس باید کم شود.

2) If the number of samples increases, the degree of freedom for the residuals also increases.

صحیح است. درجه آزادی به J بستگی داشته که با افزایش نمونه‌ها مقدارش افزایش می‌یابد.

3) The F distribution is a symmetric distribution around the zero mean.



خیر غلط است. توزیع f به سمت راست چولگی دارد. شکل بالا توزیع f را نشان می‌دهد.

4) Using ANOVA test, we can conclude that all means are different from each other

خیر غلط است. در تست ANOVA فقط می‌توان نتیجه گرفت که حداقل ۲ گروه با یکدیگر متفاوت هستند.

5) If the initial hypothesis is rejected in the ANOVA test, the standardized variability between groups is higher than the standardized variability within groups.

تست ANOVA وقتی رد می‌شود که مقدار F زیاد باشد. یعنی احتمال توزیع در آن نقطه کم باشد تا از مقدار Significance level کمتر شده و فرض صفر رد شود. در این حالت یعنی نسبت اختلاف بین گروهی به نسبت درون گروهی زیاد است. یعنی مقدار اختلاف بین گروهی بیشتر از درون گروهی است. هرچند که این موضوع به پارامترهای دیگری نیز بستگی داشته و نمیتوان گفت قطعا صحیح است.

۶_پاسخ مسئله شماره ۶

برنامه نویسی تحویل بعد از امتحان

۷_پاسخ مسئله شماره ۷

۱-۷_ پاسخ قسمت ۱

این سوال درباره یک مطالعه آماری است که باید قواعد مربوط به experimental design در آن رعایت شود تا نتیجه آزمایش به واقعیت نزدیک باشد. هدف این آزمایش بررسی تاثیر فاکتور کاهش وزن در مصرف دانه چیا که در ادامه آن را دارو می‌نامیم، می‌باشد. برای بررسی این فاکتور نمونه‌ها به دو دسته Control و treatment به صورت تصادفی تقسیم شده‌اند. به گروه treatment داروی واقعی داده و به گروه Control دارونما داده می‌شود که از همه جهات شبیه دارو اصلی است به جز دانه چیا که هدف ما بررسی تاثیر این مورد در کاهش وزن است.

۲-۷_ پاسخ قسمت ۲

در این مسئله دارویی که به دو گروه treatment و Control می‌دهند باید از همه نظر شبیه هم باشند به جز فاکتور مورد بررسی مطالعه آماری انجام شده که در این سوال دانه چیا می‌باشد. یعنی در یکی از داروها تمام محلولی که عصاره دانه چیا در آن ترکیب می‌شود مشابه داروی دیگر است که در آن ترکیب نمی‌شود.

گروه treatment : ۲۵ گرم دانه چیا ۲بار در روز گروه control : دارونما

۳-۷_ پاسخ قسمت ۳

بله در این مسئله Blocking انجام شده‌است. blocking یعنی در نمونه‌برداری اولیه نمونه‌های که تفاوت بسیاری با یکدیگر در فاکتور مورد بررسی دارند جدا شوند و سپس از بین آن‌ها گروه‌های treatment و control به صورت جداگانه انتخاب شود. در این مسئله متغیر جنسیت به این منظور انتخاب شده‌است. زیرا چربی متوسط بدن زن و مرد متفاوت بوده و ممکن است باعث خطا در نتیجه‌گیری آزمایش شود.

۴-۷_ پاسخ قسمت ۴

در این مسئله به طور دقیق بیان نشده که به داوطلبان گفته شده که عضو چه گروهی هستند. ولی باتوجه به اینکه به صورت تصادفی بین دو گروه treatment و control تقسیم شده‌اند فرض می‌شود که blinding رعایت شده‌است و افراد نمی‌دانند جز کدام گروه هستند.

۵-۷_ پاسخ قسمت ۵

باتوجه به اینکه در مسئله بیان شده که خود پژوهشگران داوطلبان را به دو گروه تقسیم کرده‌اند، هرچند که این تقسیم تصادفی است ولی می‌دانند که چه فردی عضو چه گروهی شده‌است. پس double-blind در این آزمایش رعایت نشده‌است. اگر که پژوهشگران نمی‌دانند هر فرد عضو کدام گروه است پس رعایت شده است. که در این مسئله این مورد به طور دقیق بیان نشده‌است.

۶-۷_ پاسخ قسمت ۶

باتوجه به اینکه نمونه انتخاب شده اندازه کوچکی دارد و نمی‌توان به طور دقیق بیان کرد که آیا این مورد برای جمعیت بالا نیز برقرار است یا خیر. همچنین مشخص نشده که نمونه‌های اولیه آیا اضافه‌وزن دارند یا خیر که این مورد فاکتور مهمی است. پس اگر که هدف تعمیم این نتیجه‌گیری به یک جمعیت با اندازه بزرگ است باید جمعیت هدف نیز از نظر فاکتورهای دیگر که در این مسئله بررسی نشده‌اند شبیه باشند. زیرا که کاهش وزن به عامل‌های بسیاری بستگی دارد و جنسیت فقط یکی از آن‌هاست. پس به طور کلی نمی‌توان نتیجه را به هر جمعیتی تعمیم داد.

۸_پاسخ مسئله شماره ۸

در نمودار scatter ارائه شده در این مسئله دو متغیر وزن و قد در ۵۰۷ نمونه که از نظر فیزیکی فعال هستند مشاهده می‌شود.

۱-۸_ پاسخ قسمت ۱

بدون نیاز به محاسبات، با بررسی کلی نمودار می‌توان مشاهده کرد که رابطه‌ای خطی بین وزن و قد این افراد وجود دارد و این رابطه مثبت است. یعنی با افزایش متغیر وزن، قد نیز افزایش دارد.

۲-۸_ پاسخ قسمت ۲

شکل کلی خط رگرسیون به شکل زیر است:

$$\hat{y} = \beta_0 + \beta_1 x$$

که در رابطه بالا ، $\hat{y}$ متغیر response است که در این سوال وزن به این متغیر تخصیص داده می‌شود.

متغیر x نیز همان explanatory است که در این سوال قد به این متغیر تخصیص داده می‌شود.

متغیر β_0 همان intercept یا عرض از مبدا است که از رابطه زیر بدست می‌آید:

$$\beta_0 = \bar{y} - b_1\bar{x}$$

متغیر β_1 همان slope است که از رابطه زیر بدست می‌آید:

$$\beta_1 = \frac{S_{xy}}{S_{xx}}$$

در جدول ارائه شده در این مسئله مقدار intercept برابر با -105.0113 و مقدار slope برابر با 1.0176 می‌باشد. پس باتوجه به توضیحات بالا خط رگرسیون به صورت زیر است:

$$\hat{y} = \beta_0 + \beta_1 x \rightarrow \hat{y} = -105.0113 + 1.0176x \rightarrow$$

$$Estimated\ Weight = -105.0113 + 1.0176 height$$

۸-۳_ پاسخ قسمت ۳

برای ارزیابی قدرت رگرسیون روی متغیر β_1 آزمون فرض انجام می‌شود تا مشخص شود آیا متغیر explanatory که در اینجا وزن است یک predictor خوب برای متغیر Response که در اینجا قد است می‌باشد یا خیر. آزمون فرض به صورت زیر خواهدبود:

$$H0: \beta_1 = 0 \rightarrow Model\ is\ not\ good$$

$$H1: \beta_1 \neq 0 \rightarrow Model\ is\ good$$

برای بدست آوردن مقدار pvalue ابتدا test statistics زیر باید محاسبه شود.

$$T = \frac{\beta_1}{SE(\beta_1)} = \frac{1.0176}{0.0440} = 23.127$$

که مقدار بالا در جدول نیز مشخص است. درجه آزادی نیز برابر *N-2* می باشد:

$$df = n - 2 = 507 - 2 = 505$$

مقدار pvalue برای مقدار 23.127 و درجه آزادی ۵۰۵ برابر است با:

$$pvalue = 1.541e - 81$$

این مقدار بسیار به صفر نزدیک است. پس فرض صفر رد شده و نتیجه می‌شود که مدل ارائه شده، مدل خوبی برای تخمین قد است.

۸-۴_ پاسخ قسمت ۴

ضریب همبستگی قد و وزن در این مسئله برابر است با:

$$\sigma_{xy} = r = 0.72$$

مقدار *coefficient of determination* بیانگر میزان خوب‌بودن مدل رگرسیون است که از طریق زیر محاسبه می‌شود:

$$R^2 = \frac{SSR}{SST}$$

$$r = \sqrt{R^2} \rightarrow R^2 = r^2 \rightarrow R^2 = 0.72^2 = 0.51$$

این مورد نشان می‌دهد که در ۵۱ درصد تغییرات قد نسبت به وزن توسط این مدل قابل ارائه بوده و پیش‌بینی درستی دارد.۴۹ درصد دیگر تغییرات توسط عواملی است که در این مدل توصیف نشده است.

۹_پاسخ مسئله شماره ۹

برای این مسئله جفت داده‌های زیر به عنوان ورودی داده شده‌است.

Xi	2.5	8.7	1.2	7.9	0.8	5.3	4.1	7.4	9.6	0.4
Yi	1.3	3.9	0.6	3.9	0.5	2.4	2.1	3.0	4.4	0.2

۱-۹_ پاسخ قسمت ۱

برای تخمین پارامترهای داده‌شده از فرمول‌های MLE برای مدل خطی استفاده می‌شود:

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

$$\sum X_I = 48.9 \rightarrow \frac{\sum X_I}{N} = \frac{47.9}{10} = 4.79 = \bar{X}$$

$$\sum Y_I = 22.3 \rightarrow \frac{\sum Y_I}{N} = \frac{22.3}{10} = 2.23 = \bar{Y}$$

$$\widehat{\beta}_1 = \frac{\sum(x_i - \bar{X})\,(y_i - \bar{Y})}{\sum(x_i - \bar{X})^2} = 0.441$$

$$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} = 0.113$$

$$\sigma_{unbiased} = \frac{SSE}{N-2} = \frac{\sum(y_i - \bar{Y})^2}{N-2} = 0.03$$

$$Var(\widehat{\beta}_1) = \frac{\sigma^2}{S_{xx}}$$

$$Var(\widehat{\beta}_0) = \frac{\sigma^2 \sum X_I^2}{nS_{xx}}$$

مقادیر بالا توسط کد پیوست شده محاسبه شد که در زیر خروجی کد مشاهده می‌شود:

```
beta_1_hat: 0.4418045499178665
```

```
beta_0_hat: 0.11375620589341962
var_beta_1_hat: 0.00033404403236717 62
var_beta_0_hat: 0.011304384099337608
Estimate of sigma^2 (unbiased): 0.036400444163018 82
```

۲-۹_ پاسخ قسمت ۲

آزمون فرض اول

$Significance\ Level = 0.05$

$H0: \beta_0 = 0.5$

$H1: \beta_0 \neq 0.5$

ابتدا مقدار SE برای B0 محاسبه می‌شود:

$$Var(\widehat{\beta}_0) = \sigma^2 \frac{\sum X_I^2}{nS_{xx}} = 0.0113 \rightarrow SE = \sqrt{0.0113} = 0.106$$

$$T = \frac{(\widehat{\beta}_0) - 0.5}{SE} = \frac{0.113 - 0.5}{0.106} = -3.65$$

مقدار احتمال *Statistics*بالا برای درجه آزادی *n-2* که برابر با ۸ است برابر است با ۰.۰۰۶ که از مقدار *significance level* کمتر است .پس فرض صفر رد می‌شود.

آزمون فرض دوم

خط رگرسیون از مبدا در صفحه XY می گذرد. یعنی عرض از مبدا یاb0 برابر صفر است.

$Significance\ Level = 0.05$

$H0: \beta_0 = 0$

$H1: \beta_0 \neq 0$

ابتدا مقدار SE برای B0 محاسبه می‌شود:

$$Var(\widehat{\beta}_0) = \frac{\sum X_I^2}{nS_{xx}} = 0.0113 \rightarrow SE = \sqrt{0.0113} = 0.106$$

$$T = \frac{(\widehat{\beta}_0) - 0.5}{SE} = \frac{0.113}{0.106} = 1.066$$

مقدار احتمال *Statistics*بالا برای درجه آزادی *n-2* که برابر با ۸ است برابر است با ۰.۳۱ که از مقدار *significance level* بیشتر است .پس فرض صفر رد نمی‌شود.

۱۰_ پاسخ مسئله شماره ۱۰

بازه اطمینان در Simple Regression از فرمول زیر بدست می‌آید:

$$\hat{y} \pm t_{\frac{\alpha}{2},df} SE(\hat{y})$$

از طرفی SE برابر است با:

$$SE(\hat{y}) = \sqrt{\frac{\sigma^2 \sum X_I^2}{nS_{xx}}} = \sqrt{\frac{(x_i - \bar{x})^2 \sum X_I^2}{(n)nS_{xx}}}$$

از فرمول بالا مشخص است که طول آزمون فرض به مقدار $x_i - \bar{x}$ بستگی دارد. وقتی که مقدار این عبارت برابر صفر است که مقدار x_i برابر با $\bar{x}_n$ باشد یعنی شرایطی که مسئله بیان کرده‌است. در این حالت طول بازه اطمینان در کمترین حالت خود خواهدبود.

۱۱_ پاسخ مسئله شماره ۱۱

۱-۱۱_ پاسخ قسمت ۱

پاسخ سؤال ۱۱)

$X \sim bin(n,p)$ MLE of $p \Rightarrow \hat{p} = \frac{X}{n}$ با مدل‌سازی n تا

پاسخ قسمت ۱)

توزیع دوجمله‌ای $\mathcal{L}_{X=k} = \binom{n}{k} p^k (1-p)^{n-k} \rightarrow L(p) = \prod_{i=1} f_{X_i}(n,p) \Rightarrow L = \prod_{i=1}^{m} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i}$

تابع Likelihood

$$= \prod_{i=1}^{m} \binom{n}{x_i} \prod_{i=1}^{m} p^{x_i} \prod_{i=1}^{m} (1-p)^{n-x_i} = \prod_{i=1}^{m} \binom{n}{x_i} \cancel{\prod_{i=1}^{m}} p^{\sum x_i} \prod (1-p)^{\sum (n-x_i)} = \prod_{i=1}^{m} \binom{n}{x_i} \cancel{\prod_{i=1}^{m}} p^{\sum x_i} \cancel{\prod} (1-p)^{nm-\sum x_i}$$

$$\text{LogLikelihood} = \log(L) = \log\left[\prod_{i=1}^{m} \binom{n}{x_i}\right] + \log\left[p^{\sum x_i}\right] + \log\left[(1-p)^{nm-\sum x_i}\right]$$

$$= \log\left[\prod_{i=1}^{m} \binom{n}{x_i}\right] + \left(\sum x_i\right)\log p + \left(nm - \sum x_i\right)\log(1-p) \longrightarrow$$ حل مشتق نسبت به p و برابر صفر

$$\frac{\partial \ell p}{\partial p} = \frac{1}{p}\sum x_i - \frac{1}{1-p}\left(nm - \sum x_i\right) = 0 \Rightarrow \frac{1}{\hat{p}}\sum x_i = \frac{1}{1-\hat{p}}\left(nm - \sum x_i\right)$$

$$\Rightarrow \frac{nm - \sum x_i}{\sum x_i} = \frac{1-\hat{p}}{\hat{p}} \Rightarrow \frac{nm}{\sum x_i} - 1 = \frac{1}{\hat{p}} - 1 \Rightarrow \hat{p} = \frac{1}{nm}\sum x_i = \frac{\bar{x}_n}{m} \xrightarrow[n=m \text{ برابر}]{\text{فرض اینکه}} \hat{p} = \frac{\bar{x}}{n}$$

۲-۱۱_ پاسخ قسمت ۲

پاسخ قسمت ۲)

Cramer_Rao) Fisher Information ≥? $Var(\theta_0) \geq \frac{1}{n I(\theta_0)}$

$$\frac{\partial^2 L(p)}{\partial p^2} = -\frac{1}{p^2}\sum x_i - \frac{1}{(1-p)^2}\left(nm - \sum x_i\right)$$

$$\log f_X = \log\left(\binom{n}{x} p^x (1-p)^{n-x}\right) = \log\binom{n}{x} + x\log p + (n-x)\log(1-p)$$

مشتق اول) $\frac{\partial \log f}{\partial p} = \frac{x}{p} - \frac{m-x}{1-p} = \boxed{\frac{x-np}{p(1-p)}}$ $\quad \frac{\partial^2 \log f}{\partial p^2} = \frac{(x-np)^2}{p^2(1-p)^2}$ مشتق دوم

$$E\frac{(X-np)^2}{p^2(1-p)^2} = \frac{Var(X)}{p^2(1-p)^2} = \frac{n}{p(1-p)} \to Var \geq \frac{1}{1\cdot\frac{n}{p(1-p)}} = \boxed{\frac{p(1-p)}{n}} \checkmark$$

ثابت

$$Var\left(\frac{x}{n}\right) = \frac{Var(x)}{n^2} = \frac{np(1-p)}{n^2} = \boxed{\frac{p(1-p)}{n}} \checkmark$$

۱۲_ پاسخ مسئله شماره ۱۲

برنامه نویسی تحویل بعد از امتحان

۱۳_ پاسخ مسئله شماره ۱۳

برنامه نویسی تحویل بعد از امتحان