به نام خدا

دانشگاه تهران

دانشکدگان فنی

دانشکده مهندسی برق و کامپیوتر

درس استنباط آماری

تمرین چهارم

محمد ناصری

۸۱۰۱۰۰۴۸۶

خرداد ماه ۱۴۰۱

سوال اول

الف) عبارت نادرست است. خطا توسط رابطه $\frac{SSE}{DF}$ بدست می‌آید که اگر مقدار SSE را ثابت و پایدار در نظر بگیریم، میزان خطا به درجه آزادی وابستگی دارد. از طرفی برای درجه آزادی داریم که $DF = |Sample| - |Parameters|$ (که |X| نشان دهنده اندازه است). پس با این حساب هرچه اندازه نمونه بزرگتر شود مقدار dfنیز بزرگتر خواهد شد و به همین روال مقدار خطا کوچکتر خواهد شد.

ب) عبارت درست است. در تحلیل رگرسیون، تفاوت بین مقدار مشاهده شده متغیر وابسته و مقدار پیش بینی شده($y - \hat{y}$) residuals نامیده می شود. هم مجموع و هم میانگین residuals ها برابر با صفر است.

ج) عبارت نادرست است. عمل استانداردسازی ویژگی‌ها باعث میشود تا ویژگی‌ها دارای میانگین برابر صفر بشوند. از طرفی استانداردسازی برای رگرسیون لاجستیک مورد نیاز نیست و از این تکنیک برای کمک به الگوریتم‌های بهینه سازی استفاده میشود.

د) عبارت درست است. اگر ضریب همبستگی دو متغیر صفر باشد، هیچ رابطه خطی بین متغیرها وجود ندارد. با این حال، این فقط برای یک رابطه خطی است. این امکان وجود دارد که متغیرها رابطه منحنی قوی داشته باشند.

ه) عبارت نادرست است. با افزایش مقدار آلفا در واقع مقدار جریمه در حال افزایش است و در نتیجه ضرایب نیز کاهش پیدا میکنند تا جایی که به صفر نزدیک میشوند اما به دلیل اینکه ما در مخرج عبارت lambda را داریم هیچ موقع عبارت صفر مطلق نخواهد شد.

و) عبارت نادرست است. ضریب r^2 و r نشاندهنده قدرت ارتباط خطی میباشند. ولی همانند قسمت «د» . اگر ضریب همبستگی دو متغیر صفر باشد، هیچ رابطه خطی بین متغیرها وجود ندارد. با این حال، این فقط برای یک رابطه خطی است. این امکان وجود دارد که متغیرها رابطه منحنی قوی داشته باشند.

ز) عبارت نادرست است. در حالیکه مقدار R^2 مثبت است این امکان وجود دارد که مقدار R منفی باشد.

ح) عبارت درست است. همانطور که در قسمت‌های قبل توضیح داده شد داریم $Redisual = y - \hat{y}$ و در نمودار نیز این مقدار برابر فاصله نقطه تا خط میباشد.

سوال دوم

	estimate	Std.error	T_value	Pr(>|t|)
Intercept	4.010	0.025	157.21	0
adornment	0.1325	0.032	4.14	0

الف) طبق تعریف میدانیم که در هر حالتی، خط رگرسیون همیشه از میانگین های X و Y می گذرد. این بدان معنی است که، صرف نظر از مقدار شیب، زمانی که X در میانگین خود است، Y نیز همینطور است. فلذا داریم که خط رگرسیون از دو نقطه زیر عبور میکند:

$$\begin{bmatrix} 0 \\ 4.010 \end{bmatrix}, \begin{bmatrix} -0.0883 \\ 3.9983 \end{bmatrix}$$

فلذا برای مقدار شیب خواهیم داشت:

$$Slope = \frac{y2 - y1}{x2 - x1} = \frac{3.9983 - 4.010}{-0.0881 - 0} = 0.1325$$

با توجه به مقدار بالا برای t_value خواهیم داشت:

$$t_{value} = \frac{(m - m_0)}{SE} = \frac{0.1325}{0.032} = 4.14$$

ب) برای تست پیش رو داریم که:

$$H_0: m = 0$$
$$H_a: m > 0$$

همانطور که از مقدار Pr(>|t|) قابل مشاهده است فرض صفر رد شده و میان دو متغیر تست ارتباط مثبت کوچکی وجود خواهد داشت.

فلذا t_value بدست آمده با با مقدار threshold که برابر ۱.۹۶ است مقایسه میکنیم و فرض صفر رد شده و شیب مثبت داریم

ج) برای بازه اطمینان داریم:

$$CI = b_1 \pm t_{df}.SE = b_1 \pm t_{n-2}SE$$

تابع qt() در R یک احتمال و درجات آزادی را به عنوان آرگومان می پذیرد. مقدار برگشتی فراخوانی تابع، امتیاز t مورد نیاز برای درجات آزادی مشخص شده است تا احتمال مشخص شده، مساحت زیر آن منحنی t Student و در سمت چپ امتیاز t باشد. از این تابع برای بدست آوردن مقدار t استفاده میکنیم:

Qt(0.025, df=461)

حال داریم:

$$0.1325 \pm (-1.91 * 0.032) = (0.07138, 0.1936)$$

با توجه به مقادیر بدست آمده میتوانیم بگوییم که ما با مرز ۹۵ درصد اطمینان داریم که با افزایش هر واحد در میزان adornment معلم میزان ارزیابی رضایتمندی بطور متوسط بین 0.07138 تا 0.193 افزایش پیدا کند. اگر بخواهیم تنها بر اساس بازه ی اطمینان بدست آمده در مورد نتایج صحبت کنیم میتوان نتیجه گرفت که فرضیه 0 رد میشود چون slope=0 در این بازه قرار ندارد. بنابراین میتوانیم بگوییم که میزان شیب یک مقدار مثبت غیر صفر است.

سوال سوم

	estimate	std.error	t-value	pr(>|t|)
(intercept)	-80.41	14.35	-5.60	0.0
smoke	0.44	0.03	15.26	0.0
exercise hours	-3.33	1.13	-2.95	0.0033
age	-0.01	0.09	-0.10	0.9170
height	1.15	0.21	5.63	0.0
weight	0.05	0.03	1.99	0.0471
water consumption	-8.40	0.95	-8.81	0.0

الف) برای بدست آوردن معادله خط رگرسیون داریم :

$$\eta = -80.41 + \ 0.44\beta_1 - 3.33\beta_2 - 0.01\beta_3 + 1.15\beta_4 + 0.05\beta_5 - 8.4\beta_6$$

ب) با توجه به مقادیر جدول باید انتظار داشته باشیم با افزایش میزان سیگار کشیدن به ازای هر واحد فشار خون افراد به میزان ۰.۴۴ افزایش داشته باشد ان در صورتیست که در مقابل انتظار داریم با افزایش میزان مصرف آب به ازای هر واحد فشار خون به مقدار ۸.۴ کاهش داشته باشد. همانطور که واضح است تاثیر مصرف آب بر فشار خون بیشتر است.

ج)R-squared (R2) یک اندازه گیری آماری است که نشان دهنده نسبت واریانس برای یک متغیر وابسته است که توسط یک متغیر مستقل یا متغیرهایی در یک مدل رگرسیونی تعریف شده است.

$$R^2 = 1 - \frac{Unexplained\ Variation}{Total\ Variation} = 1 - \frac{RSS}{TSS} = 1 - \left(\frac{249.28}{332.57}\right) = 0.2504$$

R-Squared فقط همانطور که در یک مدل رگرسیون خطی ساده با یک متغیر توضیحی در نظر گرفته شده است کار می کند. با یک رگرسیون چندگانه متشکل از چندین متغیر مستقل، R-Squared باید تنظیم شود.

Adjusted R-squared قدرت توصیفی مدل های رگرسیون را که شامل تعداد متنوعی از پیش بینی کننده ها هستند، مقایسه می کند. هر پیش‌بینی‌کننده‌ای که به یک مدل اضافه می‌شود، R-squared را افزایش می‌دهد و هرگز آن را کاهش نمی‌دهد. بنابراین، به نظر می‌رسد مدلی با عبارت‌های بیشتر، فقط به خاطر این واقعیت که عبارت‌های بیشتری دارد، تناسب بهتری دارد، در حالی که adjusted R squared، اضافه شدن متغیرها را جبران می‌کند و تنها در صورتی افزایش می‌یابد که عبارت جدید، مدل را بالاتر از آنچه که می‌توانست افزایش دهد، افزایش می‌یابد. با احتمال به دست می آید و زمانی کاهش می یابد که یک پیش بینی کننده مدل را کمتر از آنچه که به طور تصادفی پیش بینی می شود، افزایش دهد.

در شرایط بیش‌برازش، برای R-squared مقدار نادرست به دست می‌آید، حتی زمانی که در مدل واقعاً توانایی پیش‌بینی کاهش یافته است. این مورد در مورد adjusted R-squared صدق نمی کند.

$$Adjusted\ R^2 = 1 - (1 - R^2) * \frac{N - 1}{N - p - 1} = 1 - \left(\frac{(1 - 0.02504)(1236 - 1)}{1236 - 6 - 1}\right) = 7.251$$

سوال چهارم

الف) برای فرض این سوال چون مقایسه سه دسته بندی را داریم میتوانیم بنویسیم:

$$H_0: \mu_1 = \mu_2 = \mu_3$$
$$H_a: at\ least\ one\ group\ mean\ has\ significant\ diffrence$$

ب)

df treatment: k-1

df error: n-k

df total: n-1

MS treatment: SST / df treatment

MS error: SSE / df error

F: MS treatment / MS error

	DF	SSE	MSE	F_Value	Pr(>f)
Class	2	1144	572	6.151	0.002326
Residuals	425	39518	92.983		
	427	41234	664.983		

ج) ؟

سوال هفتم

الف)

Answer:

Standard Deviation	**s = 14.628739**
Variance	s^2 = 214
Count	n = 8
Mean	$\overline{x}$ = 34
Sum of Squares	SS = 1498

Solution

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

$$s = \sqrt{\frac{SS}{n-1}}$$

$$s = \sqrt{\frac{1498}{8-1}}$$

$$s = \sqrt{\frac{1498}{7}}$$

$$s = \sqrt{214}$$

$$s = 14.628739$$

ب) روش بوت استرپ یک تکنیک آماری برای تخمین مقادیر در مورد یک جمعیت با میانگین‌گیری تخمین‌ها از چند نمونه داده کوچک است. نکته مهم این است که نمونه ها با ترسیم مشاهدات از نمونه داده های بزرگ یک به یک و برگرداندن آنها به نمونه داده ها پس از انتخاب ساخته می شوند. این اجازه می دهد تا یک مشاهدات معین در یک نمونه کوچک معین بیش از یک بار گنجانده شود. این رویکرد به نمونه گیری، نمونه برداری با جایگزینی نامیده می شود.

فرآیند ساخت یک نمونه را می توان به صورت زیر خلاصه کرد:

1. اندازه نمونه را انتخاب کنید.
2. در حالی که حجم نمونه کمتر از اندازه انتخاب شده است
   a. به طور تصادفی یک مشاهده از مجموعه داده انتخاب کنید
   b. آن را به نمونه اضافه کنید

روش بوت استرپ را می توان برای تخمین مقداری از جمعیت استفاده کرد. این کار با گرفتن نمونه های کوچک مکرر، محاسبه آمار و گرفتن میانگین آمار محاسبه شده انجام می شود. می توانیم این روش را به صورت زیر خلاصه کنیم:

1. تعدادی نمونه بوت استرپ را برای اجرا انتخاب کنید
2. اندازه نمونه را انتخاب کنید
3. برای هر نمونه بوت استرپ
   a. یک نمونه با جایگزینی با اندازه انتخاب شده بکشید
   b. آمار روی نمونه را محاسبه کنید
4. میانگین آمار نمونه محاسبه شده را محاسبه کنید.

ج) میتوان با توجه به تغریف مساله حدس زد که مقادیر توزیع نرمالی حول میانگین یعنی تقریبا ۳۰ خواهند داشت.

د) پارامتر مهم این آزمایش میانگین است که میتوان تخمین آنرا به حدود مقدار ۳۰ در نظر گرفت. همچنین واریانس را نیز میتوان مقدار تقریبی ۲۰ در نظر گرفت

ه) با فرض اینکه توزیع نرمال داریم مقدار $z^{(05/2)}$ را محاسبه میکنیم (۱.۹۵) و به صورت زیر بازه اطمینان محسابه میشود:

$$Std\ Err = 4.85 \rightarrow 4.85 * 1.95 = 9.45$$
$$CI = \bar{X} \pm error\ margin = 34 \pm 9.45$$

توزیع نرمال حول میانگین ۳۴ و واریانس ۹.۴۵

سوال هشتم

الف) در آمار، مقایسه‌های چندگانه بدین معناست که بخواهیم مجموعه‌ای از استنتاج‌های آماری را به‌طور همزمان در نظر بگیریم یا زیرمجموعه‌ای از پارامترها را استنباط کنیم که بر اساس مقادیر مشاهده‌شده انتخاب شده‌اند. ما همیشه علاقه‌ای به مقایسه دو گروه در هر آزمایش نداریم. گاهی اوقات (در عمل، اغلب)، ممکن است مجبور شویم تعیین کنیم که آیا تفاوت هایی بین میانگین

های سه یا چند گروه وجود دارد یا خیر. رایج ترین روش تحلیلی مورد استفاده برای این گونه تعیین ها آنالیز واریانس (ANOVA) است. نتیجه ANOVA اطلاعات دقیقی در مورد تفاوت بین ترکیب های مختلف گروه ها ارائه نمی دهد و در صورت رد شدن فرض صفر که برابری میانگین‌هاست، تنها به ما میگوید که حداقل یکی از ۳ گروه تفاوت معناداری با دیگران دارد.

از تست‌های مقایسه چندگانه برای این گونه موارد استفاده میشود تا بفهمیم کدام یک از حالت‌های ممکن باعث رد شدن فرض صفر در مساله ما شده اند.

ب)

Bonferroni method:

روش بونفرونی یک روش نسبتا محافظه کارانه برای مدیریت تخمین احتمال خطا در آزمایش های چندگانه است. این روش اساساً یک حد بالای قابل اثبات درستی را در مورد احتمال کلی خطای رد نادرست حداقل یک فرضیه در کل مجموعه فرضیه‌های در نظر گرفته ارائه می‌کند. تنظیم Bonferroni برای کنترل نرخ خطای Family-wise (FWER) استفاده می شود. در آمار، نرخ خطای خانوادگی، احتمال انجام یک یا چند کشف نادرست، یا خطای نوع اول در هنگام انجام آزمون فرض‌های متعدد است. با افزایش تعداد فرضیه های آزمایش شده، خطای نوع اول افزایش می یابد. بنابراین significance-level را به تعداد آزمون فرض ها تقسیم می کنیم. به این ترتیب، خطای نوع اول را می توان کاهش داد. به عبارت دیگر، هر چه تعداد فرضیه های مورد آزمایش بیشتر باشد، معیار دقیق تر و احتمال تولید خطاهای نوع اول کمتر می شود. در حالت کلی نامساوی بونفرونی به صورت زیر تعریف میشود:

$$P\left(\bigcap_{i=1}^{g} A_i\right) \geq 1 - \sum_{i=1}^{g} P[\bar{A}_i]\,,$$

Benjamini-Hochberg:

این روش برای کنترل نرخ کشف نادرست (FDR) استفاده می شود. تنظیم نرخ به کنترل مواقعی کمک می کند که مقادیر p کوچک (کمتر از 5٪) به طور تصادفی اتفاق می افتد، که می تواند منجر به رد نادرست فرضیه های صفر واقعی شود. به عبارت دیگر، روش B-H به ما کمک می کند تا از خطاهای نوع اول (مثبت نادرست) اجتناب کنیم.

ج) تصحیح بونفرونی و رویه بنجامینی-هوچبرگ فرض می‌کنند که آزمون‌های فردی مستقل از یکدیگر هستند، مانند زمانی که نمونه A را در مقابل نمونه B، C در مقابل D، E در مقابل F، و غیره مقایسه می‌کنید. اگر نمونه A را با نمونه B، A در مقابل C، A در مقابل D، و غیره مقایسه میکنید، مقایسه ها مستقل نیستند. اگر A بالاتر از B باشد، احتمال زیادی وجود دارد که A نیز بالاتر از C باشد. یکی از مکان‌هایی که این اتفاق می‌افتد، زمانی است که شما در حال انجام مقایسه‌های برنامه‌ریزی نشده میانگین‌ها در ANOVA هستید، که تکنیک‌های متنوع دیگری مانند آزمون Tukey-Kramer برای این دسته کارها ایجاد شده است. یکی دیگر از موارد قابل تامل در رابطه با مقایسه‌های متعدد و غیرمستقل، زمانی است که چندین متغیر را بین گروه‌ها مقایسه می‌کنید و

متغیرها در گروه‌ها با یکدیگر همبستگی دارند. به عنوان مثال می‌توان ژن مورد علاقه‌تان را در موش‌ها از بین برد و هر چیزی را که در موش‌های ناک اوت می‌توان در نظر گرفت با موش‌های کنترل مقایسه کرد: طول، وزن، قدرت، سرعت دویدن، مصرف غذا، تولید مدفوع، و غیره. همه این متغیرها احتمالاً همبستگی درون گروهی دارند؛ موش‌هایی که بلندتر هستند احتمالاً وزن بیشتری نیز دارند، قوی‌تر می‌شوند، سریع‌تر می‌دوند، غذای بیشتری می‌خورند و بیشتر مدفوع می‌کنند. برای تجزیه و تحلیل این نوع آزمایش، می توانید از تحلیل واریانس چند متغیره یا manova استفاده کرد

سایر تکنیک‌های پیچیده‌تر، مانند راینر و همکاران. (2003)، برای کنترل نرخ کشف کاذب توسعه داده شده اند که ممکن است در صورت عدم استقلال در داده ها مناسب تر باشد.

د) مقایسه‌های چندگانه را می‌توان با روش هایی مانند بونفرونی و اصلاحات دیگر یا با رویکرد کنترل نرخ کشف نادرستی مانند Benjamini شناخت. اما این رویکردها همیشه مورد نیاز نیستند. سه موقعیت وجود دارد که محاسبات این روش ها مورد نیاز نیست.

1. هنگام تفسیر نتایج به جای محاسبات:

برخی از آماردانان توصیه می کنند که هرگز در هنگام تجزیه و تحلیل داده ها، مقایسه های متعدد را اصلاح نکنید. در عوض تمام مقادیر P و فواصل اطمینان را گزارش کنید و بیان کنید که هیچ اصلاح ریاضی برای مقایسه‌های چندگانه انجام نشده است. این رویکرد مستلزم آن است که همه مقایسه ها گزارش شوند.

2. اگر فقط چند مقایسه برنامه ریزی شده انجام دهید:

در این مواقع:

- ما به جای هر مقایسه ممکن، چند مقایسه علمی معقول را انتخاب کرده ایم
- انتخاب مقایسه هایی که باید انجام شود بخشی از طراحی آزمایشی بود
- پس از مشاهده داده ها نیازی به مقایسه بیشتر نمی بینیم

در این مورد باید یک significance level برای هر مقایسه جداگانه بدون اصلاح برای مقایسه های چندگانه تعیین کنیم.

3. وقتی مقایسه ها مکمل هم هستند:

در تست هایی که نمونه ها به زیر گروه ها تقسیم شده اند و ما به دنبال پاسخ یک سوال برای همه آنها هستیم.

ه)

در ابتدا مقادیر p_value را مرتب کرده و رتبه بندی میکنیم:

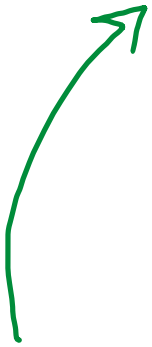
$$0.0008 < 0.009 < 0.165 < 0.396 < 0.450 < 0.641 < 0.781 < 0.90 < 0.993$$

در مرحله بعد هر p_valueرا با مقدار بدست آمده از رابطه benjamini مقایسه میکنیم. این رابطه به صورت زیر تعریف میشود:

$$\left(\frac{i}{m}\right).Q$$

که در آن i رتبه، m تعداد آزمون‌ها و Q مقدار FDR مورد نظر میباشد.

P_value	Rank	Crit_value
0.0008	1	0.0055
0.009	2	0.0111
0.165	3	0.0166
0.396	4	0.0222
0.450	5	0.0277
0.641	6	0.033
0.781	7	0.389
0.9	8	0.044
0.993	9	0.05



پس از آن بزرگتریم مقدار p_value که از مقدار critical خود کوچکتر است و تمامی مقادیر قبل از آن را انتخاب میکنیم.