ФМИ ПС 2018 - Домашна работа 4

Янис Василев, ianis@ivasilev.net, спец. Статистика, ф.н. 128

5 юни 2018

Задача 1

Mr. Bump иска да изучи зависимостта между цената и обема на продажбите на мляко. За тази цел той използва данни, които е събирал в продължение на 10 седмици. Те са представени в таблицата по-долу.

Номер на седмицата	Количество продадено мляко Y (хиляди галони)	Цена на един галон X (долари)
1	10	1.3
2	6	2.0
3	5	1.7
4	12	1.5
5	10	1.6
6	15	1.2
7	5	1.6
8	12	1.4
9	17	1.0
10	20	1.1

Лесно се вижда, че има обратна линейна зависимост между променливата Y и X. Може да се направи извод, че при нарастване на цената обемът на продажбите намалява.

Mr. Bump се нуждае от количествена мярка за получената зависимост. За тази цел той ще използва извадъчната корелация:

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

Необходимо е да направи следните изчисления:

№	Y	X	XY	X^2	Y^2
1	10	1.3	13.0	1.69	100
2	6	2.0	12.0	4.00	36
3	5	1.7	8.5	2.89	25
4	12	1.5	18.0	2.25	144
5	10	1.6	16.0	2.56	100
6	15	1.2	18.0	1.44	225
7	5	1.6	8.0	2.56	25
8	12	1.4	16.8	1.96	144
9	17	1.0	17.0	1.00	289
10	20	1.1	22.0	1.21	400
Сума	112	14.40	149.3	21.56	1488

а) Покажете, че $\hat{\rho}(X, Y)$ = -0.86.

б) Покажете на Mr. Bump, че регресионните коефициенти на регресионното уравнение

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

се оценяват по МНМК като

$$\hat{\beta}_1 = -14.54$$
$$\hat{\beta}_0 = 32.14$$

и следователно регресионното уравнение се записва във вида

$$\hat{Y} = 32.14 - 14.54X.$$

Каква е връзката между оценката $\hat{\beta}_1$ и коефициента на корелация $\hat{\rho}$?

в) Обяснете на Mr. Bump смисъла на регресионните коефициенти.

г) Mr. Bump не може да разбере какво количество мляко ще бъде „продадено“, ако то се раздава безплатно, т.е. ако цената за един галон е 0. Помогнете му.

д) Покажете, че стандартната грешка на оценката е $\sigma = 2.72$.

е) Покажете, че стандартната грешка на прогнозата в точката $X = 1.63$ е 2.91 и следователно 95%-ният доверителен интервал за прогнозата в тази точка е (1.73, 15.5).

ж) Намерете стандартните грешки на оценките на регресионните коефициенти.

з) Покажете на Mr. Bump как се получават стойностите в ANOVA таблицата

Source	SS	DF	Mean squares
Regr.	SSR = 174.19	1	MSR = 174.19
Error	SSE = 59.41	8	MSE = 7.43
Total	SST = 233.60	9	

Пресметнете коефициента на детерминация R^2. Дайте обяснение на получения резултат.

и) Запишете хипотезите в F и t критериите. Намерете стойностите на F и t статистиките. Каква е връзката между тях? Какъв извод ще направите от тези тестове, ако нивото на съгласие е $\alpha = 0.01$?

й) Каква е връзката между F-статистиката и коефициента на детерминация?

Решение. а) Използваме данните от последната таблица и получаваме

$$\hat{\rho}(X,Y) = \frac{10 \times 149,3 - 14,4 \times 112}{\sqrt{10 \times 21.56 - (14,4)^2}\sqrt{10 \times 1488 - (112)^2}} = \\ = \frac{1493.0 - 1612.8}{\sqrt{215.6 - 207.4}\sqrt{14880 - 12544}} = -\frac{119.8}{\sqrt{8.2}\sqrt{2336}} \approx -0.862.$$

б) Ще изведем оценки за β_0 и β_1. Полагаме

$$\hat{Y}_i = \beta_0 + \beta_1 X$$
$$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X)^2.$$

Ще намерим стойности на β_0 и β_1, които да минимизират SSE. За целта нулираме частните производни на SSE по оценяваните параметри:

$$\frac{\partial SSE}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \iff n\beta_0 = \sum_{i=1}^{n} Y_i - \beta_1 \sum_{i=1}^{n} X_i$$
$$\beta_0 = \overline{Y}_n - \beta_1 \overline{X}_n,$$

където $\overline{G(X_n, Y_n)} = \frac{1}{n}\sum_{i=1}^{n} G(X_i, Y_i)$ са осреднени суми на някаква трансформация на X и Y.

$$\frac{\partial SSE}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0 \iff \beta_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i - \beta_0 \sum_{i=1}^{n} X_i$$
$$\beta_1 \overline{X_n^2} = \overline{X_n Y_n} - \beta_0 \overline{X}_n$$
$$\beta_1 \overline{X_n^2} = \overline{X_n Y_n} - \overline{X}_n \overline{Y}_n + \beta_1 \overline{X}_n^2$$
$$\beta_1 = \frac{\overline{X_n Y_n} - \overline{X}_n \overline{Y}_n}{\overline{X_n^2} - \overline{X}_n^2}$$

От последното лесно пресмятаме

$$\hat{\beta}_1 = \frac{10 \times 149,3 - 14,4 \times 112}{10 \times 21.56 - (14.4)^2} \approx -14.54,$$
$$\hat{\beta}_0 = \frac{112 - 14.4\beta_1}{10} \approx 32.14.$$

Извадъчният коефициент на корелация може да се запише с осреднени суми по следния начин:

$$\hat{\rho}(X,Y) = \frac{n\sum_{i=1}^n X_iY_i - \sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{\sqrt{n\sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2}\sqrt{n\sum_{i=1}^n Y_i^2 - (\sum_{i=1}^n Y_i)^2}} = \frac{\overline{X_nY_n} - \overline{X}_n\overline{Y}_n}{\sqrt{\overline{X_n^2} - \overline{X}_n^2}\sqrt{\overline{Y_n^2} - \overline{Y}_n^2}},$$

откъдето лесно се вижда, че

$$\hat{\rho}(X,Y) = \beta_1 \frac{\sqrt{\overline{X_n^2} - \overline{X}_n^2}}{\sqrt{\overline{Y_n^2} - \overline{Y}_n^2}} = \beta_1 \frac{\hat{\sigma}_X}{\hat{\sigma}_Y},$$

където $\hat{\sigma}_X$ и $\hat{\sigma}_Y$ са непоправените извадкови дисперсии на $X_1, \ldots, X_n$ и $Y_1, \ldots, Y_n$.

в) Ако имаме линеен модел от вида

$$Y = \beta_0 + \sum_{i=1}^m \beta_i X^{(i)},$$

то регресионните коефициенти $\beta_1, \ldots, \beta_m$ са оценки за линейното влияние на съответните предикторни променливи $X^{(1)}, \ldots, X^{(n)}$ върху Y, докато β_0 е оценка за влиянието на неизвестни предикторни променливи извън модела.

В случая имаме силна линейна зависимост между X и Y, за което свидетелства относително високата абсолютна стойност на $\beta_1 = -14.54$ (за което по-лесно се съди по пропорционалният на тази стойност коефициент на корелация $\hat{\rho}(X,Y) = -0.862$), но имаме и немалко „неизвестно“ влияние върху Y, което се вижда от $\beta_0 = 32.14$.

г) Тъй като членът $\beta_1 X$ се анулира, когато $X = 0$, цената на млякото за един галон ще бъде равно на $\beta_0 = 32.14$.

д) Използваме това, че MSE е неизместена оценка за σ^2 и получаваме $\sigma = \sqrt{MSE} = \sqrt{\frac{SSE}{n-2}} \approx \sqrt{\frac{59.42}{8}} \approx 2.73$.

е) Използваме вече намерената оценка за $\sigma \approx 2.73$, за да намерим $\mathbb{D}\,\hat{\varepsilon}$, където $\hat{\varepsilon} = Y - \hat{Y}$ (в точката $X = 1.63$, която не е сред входните данни), а $\hat{Y} \approx 8.44$. И така,

$$\mathbb{D}\,\hat{\varepsilon} = \sigma^2 \left(\frac{n+1}{n} + \frac{(X - \overline{X}_n)^2}{\overline{X_n^2} - \overline{X}_n^2} \right) \approx (2.73)^2 \times 1.14,$$

откъдето получаваме стандартната грешка $\sqrt{\mathbb{D}\,\hat{\varepsilon}} \approx 2.73\sqrt{1.14} \approx 2.91$. Съответният доверителен интервал е

$$8.44 \pm 2.73 \times \mathrm{t}(8)_{0.975} \approx (1.71, 15.16).$$

ж) Първо намираме дисперсиите на β_1 и β_0:

$$\mathbb{D}\,\beta_1 = \mathbb{D}\,\frac{\overline{X_nY_n} - \overline{X}_n\overline{Y}_n}{\overline{X_n^2} - \overline{X}_n^2} = \frac{1}{(\overline{X_n^2} - \overline{X}_n^2)^2}\frac{1}{n^2}\left(\sum_{i=1}^n X_n\,\mathbb{D}\,Y - n\overline{X_n}^2\,\mathbb{D}\,Y \right) = \frac{n^{-1}\sigma^2}{\overline{X_n^2} - \overline{X}_n^2},$$

$$\mathbb{D}\,\beta_0 = \mathbb{D}\left(\overline{Y}_n - \beta_1\overline{X}_n\right) = n^{-1}\sigma^2 + \overline{X}_n^2\,\mathbb{D}\,\beta_1.$$

Като заместим с конкретни числа (използваме намерената оценка за дисперсията), получаваме $\mathbb{D}\,\beta_1 \approx 9.01$ и $\mathbb{D}\,\beta_0 \approx 19.44$, а стандартното отклонение е съответно 3.00 и 4.41.

з) Стойностите в ANOVA таблицата, представляващи суми от квадрати на отклонения, се получават по следния начин:

$$\begin{aligned}
SST &= \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 \in \chi^2(n-1)\\
SSE &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \in \chi^2(n-(k+1))\\
SSR &= \sum_{i=1}^{n}(\overline{Y}_n - \hat{Y}_i)^2 \in \chi^2(k)\\
MSE &= \frac{SSE}{n-(k+1)}\\
MSR &= \frac{SSR}{k},
\end{aligned}$$

където k е броят предикторни променливи (в случая $k = 1$). За сумите е изпълнена следната връзка:

$$\begin{aligned}
SST = \sum_{i=1}^{n}(Y_i - \overline{Y}_n)^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y}_n)^2 =\\
&= \sum_{i=1}^{n}\left[(Y_i - \hat{Y}_i)^2 + 2(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y}_n) + (\hat{Y}_i - \overline{Y}_n)^2\right] =\\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_n)^2 + 2\beta\sum_{i=1}^{n}(Y_i - \overline{Y}_n + \overline{Y}_n - \hat{Y}_i)(\hat{X}_i - \overline{X}_n) =\\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_n)^2 + 2\beta\left[\sum_{i=1}^{n}(Y_i - \overline{Y}_n)(\hat{X}_i - \overline{X}_n) - \beta_1\sum_{i=1}^{n}(X_i - X_n)^2\right] =\\
&= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y}_n)^2 = SSE + SSR.
\end{aligned}$$

Коефициентът на детерминация се дефинира като

$$R^2 \coloneqq \frac{SSR}{SST}.$$

В нашия случай $R^2 \approx 0.75$. Неформално, коефициентът на детерминация ни казва каква част от дисперсията на отклика можем да предскажем от предикторите.

и) И двата критерия се основават на факта, че сумите от квадратите SST, SSE и SSR имат χ^2 разпределение. Имаме хипотезите

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

и критериите

$$F = \frac{MSR}{MSE} < 11.26 \approx \mathrm{F}(k, n-(k+1))_{1-\alpha}$$
$$|T| = \frac{|\beta_1|}{\sqrt{MSE}} < 3.36 \approx \mathrm{t}(n-(k+1))_{1-\alpha/2}$$

С нашите данни получаваме $F = 23.44$ и $T = -5.33$, следователно имаме основание да приемем алтернативната хипотеза, че коефициентът β_1 е различен от 0 и че X оказва влияние върху Y.

й) F-статистиката и коефициента на детерминация имат следната очевидната връзка

$$\frac{1}{F} = \frac{MSE}{MSR} = \frac{k}{n-(k+1)}\frac{SSE}{SSR} = \frac{k}{n-(k+1)}\frac{SST-SSR}{SSR} = \frac{k}{n-(k+1)}\left(\frac{1}{R^2}-1\right).$$