Вероятности и статистика с R

Асен Чорбаджиев

January 8, 2018

1 Метод на най-малките квадрати (MLS)

Нека са дадени измервания на даден признак y с размер n и средно $\overline{y}$. Стандартна статистическа задача е да се оцени и прогнозира y посредством други измерени един или повече признаци x, наречени предиктори. Тогава решението се търси във вид (за един признак):

$$y = b_0^\star + b_1^\star x \tag{1}$$

където $b_0^\star$ и $b_1^\star$ се наричат съответно intercept(транслация) и slope(наклон). Решението за коефициентите $b_0^\star$ и $b_1^\star$ се намира посредсвом минимализацията на:

$$SS(b_0^\star, b_1^\star) = \sum_{i=1}^{n} (y_i - (b_0^\star + b_1^\star x))^2 \tag{2}$$

което става чрез решаването на системата диференциални уравнения:

$$\frac{\partial}{\partial b_0^\star} SS(b_0^\star, b_1^\star) = 0 \tag{3}$$

$$\frac{\partial}{\partial b_1^\star} SS(b_0^\star, b_1^\star) = 0 \tag{4}$$

Когато моделът е съставен само от един предиктор решението горната система е fitted regression line:

$$b_1 = r(s_y/s_x) \tag{5}$$

$$b_0 = \overline{y} - b_1 \overline{x} \tag{6}$$

където с s_y,s_x са означени стандартните отклонения на y и x, а с r корелацията мужду тях.
За пресмятане на коефицентите на регресията на линейни модели в R се използва функцията lm(y x). Графично кривата на MLE fit се чертае с abline(), на вече съществуваща графика с plot().

2 Условия за линейни регресионни модели

Основни признаци определящи един регресионен модел - **наблюдения** и **грешка** :

Table 1: Извадка от наблюдения

F1.	$Ey_i = \beta_0 + \beta_1 x_i$
F2.	$x_1 ... x_n$ - не са стохастични променливи.
F3.	Var $y_i = \sigma^2$
F4.	$\{y_i\}$ са независими случани променливи.
F5.	$\{y_i\}$ са нормално разпределени.

Грешката на модела е дефинирана като $y_i - (\beta_0 + \beta_1 x_i)$. За нея са верни следните предположения:

Table 2: Извадка от наблюдения

E1.	$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
E2.	$x_1 ... x_n$ - не са стохастични променливи.
E3.	$E\epsilon = 0$; Var $\epsilon_i = \sigma^2$
E4.	$\{\epsilon_i\}$ са независими случани променливи.
E5.	$\{\epsilon_i\}$ са нормално разпределени.

3 Линейна регресия

3.1 ANOVA

Квадратичната разлика $y_i - \overline{y}$ е основа за за измерване на размаха на данните.Тогава сумата:

$$Total\_SS = \sum_{i=1}^{n} (y_i - \overline{y})^2 \tag{7}$$

представлява общата вариация. Нека тогава имаме специално знание за предиктора x. Тогава с регресионната крива за всяко наблюдение имаме оценка fitted value $\hat{y} = b_0 + b_1 x_1$. Също така разликата $y_i - \hat{y}_i$ представлява грешката на оценката. По този начин успехът на един регресионен модел е свързан с това оценката $\hat{y}$ да бъде по-акуратна от $\overline{y}$. Алгебрично това се представя като "Отклонението без знание за x" = "отклонение със знание

за x" + "отклонение обяснено от x":

$$y_i - \overline{y} = y_i - \hat{y}_i + \hat{y}_i - \overline{y} \tag{8}$$

Тогава уравнението за $Total\_SS = Error\_SS + Regression\_SS$, където SS отговаря за Sum of Squares, което след преобразувания и отчитайки, че $2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$ е равно на нула за цялото множество имаме:

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 \tag{9}$$

В термините на линейната регресия тази връзка се показва чрез coefficint of determination R-square (R^2):

$$R^2 = \frac{Regression\_SS}{Total\_SS} \tag{10}$$

Стойностите които приема са между 0 и 1.Така когато регресионната пасва идеално на модела, т.е. $Error\_SS = 0$ имаме $R^2 = 1$. Обратно, когато $R^2 = 0$ моделът не предоставя информация за y.

Вторият параметър за оценка на регресията се определя от грешките (residuals) e_i, които са равни на:

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_i x_i) \tag{11}$$

Тогава оценката за σ^2 се нарича mean square error (MSE) и е равен на:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i = \frac{Error\_SS}{n-2} \tag{12}$$

А положителното решение на $s = \sqrt{s^2}$ - residual standard deviation. Делението на $n - 2$ степени на свобода е центрирано понеже са необходими поне 2 точки за регресионна права. Тази тип анализ се нарича Analysis of variance (ANOVA), с общо n-1 степени на свобода.
За оценка на надежността (точност) на наклона на кривата се използва стандартното отклонение (standard deviation) на b_1:

$$se(b_1) = \frac{s}{s_x\sqrt{n-1}} \tag{13}$$

Заключенията които следват:
1. По-голямо n означава по-малко b_1.
2. Колкото точките са по-близо до правата толкова b_1 и по-малко.
3. Колкото повече точките x са по-отдалечени, толкова s_x става по-голямо и b_1 и по-малко.

3.2 Статистическа значимост

1. Използване на t-test за проверка на $H_0 : Eb_1 = \beta_1 = 0$ vs $H_\alpha : Eb_1 = \beta_1 \neq 0$. Проверката на хипотезаа става с t-ratio $t(b_1) = b_1/se(b_1)$ с n-2 степени на свобода.
2. $100(1-\alpha)\%$ доверителен интервал за β_1 е равен на:

$$b_i \pm t_{n-2,1-\alpha/2} se(b_i) \tag{14}$$

3.3 Многомерност

Когато за изграждане на оценка на се използват повече от един предиктор, то извадката от данни изглежда така:

$$\begin{Bmatrix} x_{11} & x_{12} & x_{1k} & y_1 \\ x_{21} & x_{22} & x_{2k} & y_2 \\ & & & \\ x_{n1} & x_{n2} & x_{nk} & y_n \end{Bmatrix} \tag{15}$$

а регресионния модел изглежда по следния начин:

$$y = b_0 + b_1 x_1 + ... + b_k x_k \tag{16}$$

В резултат на това SS и вектора b на коефицентите на линейната регресия са:

$$SS(b_0^\star, b_1^\star, ..., b_k^\star) = \sum_{i=1}^{n} (y_i - (b_0^\star + b_1^\star x_{i1} + ... + b_k^\star x_{ik}))^2 \tag{17}$$

$$b = (X'X)^{-1} X'y \tag{18}$$

Оценката на σ^2 MSE става равна на:

$$s^2 = \frac{1}{n-(k+1)} \sum_{1}^{n} (y_i - \hat{y}_i)^2 \tag{19}$$

с $n-(k+1)$ степени на свобода. Общият брой на степени на свобода на $Total\_SS$ става $n-1$.
Стандарната грешка на коефициентите b_j се пресмята:

$$se(b_j) = s\sqrt{(j+1)\text{-ият диагонал на } (X'X)^{-1}} \tag{20}$$

Удобен графически метод за анализ е scatterplotMatrix() от библиотеката "car".

3.4 Пресмятане на ANOVA с R

Често е нужно да се изследват няколко регресионни коефициенти едновременно. Такъв случай е когато имаме $(k+1) \times 1$ вектор $Eb = \beta = (\beta_0, \beta_1, ...\beta_k)$. Тогава тестът на user-selected стойности описани с матрица C, $p \times (k+1)$, за която се очаква за хипотезата H_0 следното равенство да бъде вярно $C\beta = d$, d е вектор с размер $1 \times p$. Такъв тест може да бъде например, $H(0) : \beta_1 = \beta_2$, което е равно на $\beta_1 - \beta_2 = 0$.
Матрично подобен тест се представя с $p \leq k+1$ брой restrictions. Алтернатвната хипотезата, която се тества е $H_\alpha : C\beta \neq d$. Заключението се получава от F-тест пресметнат да статистиката:

$$F - ratio = \frac{(Error\_SS_{reduced}) - (Error\_SS_{full})}{ps^2_{full}} \tag{21}$$

където с *reduced* е означен ограничения модел. Параметрите на F-разпределението са $F(p, n-(k+1))$. Функцията в R се нарича anova().

3.5 Изграждане на по-добър модел с R

Линейната регресия в R се моделира с функцията lm(). Извеждането на всичките резултати от регресията става чрез summary(lm()). Коефицентите на кривата се достигат чрез coefficients(lm())[index or 'name'] или $lm()\$coefficients$. Достъпът до статистиките на коефицентите на кривата са достъпни чрез coefficients(summary(lm()))[index or 'name']. функцията за достъп до остатъците е residuals().
Когато се анализира регресия първо трябва да се уверите, че са избегнати следните грешки:
1. Зависимост.
2. heteroscedasticity. Различни E3 условия.
3. Ненормално разпределени величини.
4. Outliers. Критерии да отделяне на outliers са:
- графични - boxplot, scatterplot
- residuals със стойности на r извън интервала [-2,2] за 95% достоверност:

$$r = \frac{e_i}{\hat{\sigma(e_i)}} \tag{22}$$

Това се пресмята с функцията rstandard() в R.
5. Връзка между предиктори и девиацията на модела. Residual Analysis. Това става с анализ на residuals. Функцията за тяхното получаване е resid(model). Решението става графично - scatterplot residuals vs predictors. Проблемни са тези остатъци, които са извън общия тренд.

4 Упражнения:

- **Зад. 1.**: Налична е следната таблица с данни: Пресметнете регресионната

Table 3: Dataset 1

i	1	2	3
x_i	2	-6	7
y_i	3	4	6

линия с MLE. Пресметнете r,b_0, b_1. Начертайте графики.

- **Зад. 2. Перфектна корелация**: Налична е следната таблица с данни за квадратичното уравнение $y = x^2$: Пресметнете регресионните

Table 4: Dataset 2

i	1	2	3	4	5
x_i	-2	-1	0	1	2
y_i	4	1	0	1	4

линии с MLE за $y \sim x$ и $y \sim x^2$. Пресметнете r,b_0, b_1. Начертайте графики.

- **Зад. 3.**: Количеството Азотни соли $NsNO_3$, което може да се разтвори в 100г. вода в зависимост от температурата $X(^oC)$ (Таблица 5) : Изчеслете параметрите на линейната регресия и определете интервалите

Table 5:

X	0	4	10	15	21	29	36	51	68
Y	66.7	71	76.3	80.6	85.7	92.9	99.4	113.6	125.1

на параметрите на правата. Използвайте и графични методи.

- **Зад. 4.**: За следната таблица с от Base Points (Таблица 6): Изградете регресионен модел. Проверете за наличие outlier-и. Оценете тяхното влияние. Използвайте и графични методи.

- **Зад. 5.**: Регресионен модел със следните коефиценти (таблица 7): със $s = 1.513$. Да се провери нулевата хипотеза $\beta_4 = \beta_5 = 0$ с $\alpha = 5\%$ при $Error_S S_{reduced} = 630.43$.

- **Зад. 6.**: Инсталирайте пакета "faraway". В него ще намерите таблица "savings". Описание на колоните ще намерите с командата "?savings". Да се определят коефициентите на регресията, да се намери най-добър модел редуцеряне. Да се сравнят моделите с ANOVA. Да

Table 6:

x	1.5	1.7	2.0	2.5	2.5	2.7	2.9	3	3.5	3.4	9.5
y	3	2.5	3.5	3	3.1	3.6	3.2	3.9	4	4	8
x	9.5	3.8	4.2	4.3	4.6	4	5.1	5.1	5.1	5.2	5.5
y	2.5	4.2	4.1	4.8	4.2	5.1	5.1	5.1	5.1	4.8	5.3

Table 7: Dataset 2

Source	Sum of Squares	df	Mean Square
Regression	343.28	5	68.66
Error	615.62	269	2.29
Total	948.9	274	

се пресметнат прогнозни интервали за всяка държава. Процесът да бъде съпроводен с описателни методи.