Regression Assignment One

Terry Liu 刘骏杰 l630005038
Section: Dr.Hua-Jun YE

2018 年 10 月 1 日

Question One: Suppose you are a consultant to the local tourism authority and the CEO of the authority would like to know whether a family's annual expenditure on recreation is related to their annual income. In addition, if there is a relationship, he would like you to build a statistical model which quantifies the relationship between the two variables. A data set consisting of a random sample of 8 families, collected last year, is available to help you with the assessment.

Y:Expenditure($1k)	X: Income($1k)
2.35	52.0
4.95	66.0
3.10	44.5
.50	37.7
5.11	73.5
3.10	37.5
2.90	56.7
1.75	35.6

(a) Fit a linear regression $y = \beta_0 + \beta_1 x + \epsilon$ on the data. Denote b_0 and b_1 as the least square point estimations of β_0 and β_1. Calculate SS_{xx}, SS_{yy}, and SS_{xy}, and then calculate b_0 and b_1.
(b) Give an interpretation of each: b_0 and b_1

Answer: (a) This is the R code:

```
x <- c(52.0,66.0,44.5,37.7,73.5,37.5,56.7,35.6)
y <- c(2.35,4.95,3.10,2.50,5.11,3.10,2.90,1.75)

sum_xs <- sum(x^2)
sum_1 <- (sum(x))^2
SS_xx <- sum_xs - (b / length(x))

sum_xy <- sum(x * y)
sum_2 <- sum(x) * sum(y)
SS_xy <- sum_xy - (sum_2 / 8)

b_1 <- SS_xy / SS_xx
b_0 <- mean(y) - b_1 * mean(x)

sum_ys <- sum(y^2)
sum_3 <- (sum(y))^2
SS_yy <- sum_ys - (sum_3 / 8)

# > SS_xx
# [1] 21448.17
# > SS_yy
# [1] 10.1324
# > SS_xy
# [1] 100.395
```

(b) We can get the $y = -0,39495 + 0,07167187x + \epsilon$ When $x = 60$,

$y = 3.905362$

So, we can get these information: if a family's income is \$60,000, the expenditure is about \$3910 .

Question Two: Consider the regression model

$$y = \beta_0 + \beta_1 + \epsilon$$

to fit a data set $\{(x_i, y_i)\}$, that is

$$_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1, \ldots n$$

Answer the following questions: (a) Give an explanation for $y, x, _0$ and $_1$ in model (1).

(b) For obtaining the least square estimates and performing the hypothesis test, what are the assumptions for model (2)?

(c) Prove that the point $\overline{x}, \overline{y}$ is on the regression line $\hat{y} = \hat{\beta_0} + beta_1 x$, where$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$. $\hat{\beta_0}$ and $\hat{\beta_1}$ are the estimates of $beta_0$ and β_1, respectively.

Answer:

(a) x is the independent variable and y is the dependent variable;

β_0 is a peremiter that not influenced by the x but by other variable;

β_1 means that if x increase 1, then the value may increase y values.

(b) We have to assume that $\epsilon_1, \epsilon_2, \ldots, \epsilon_i$ are i.i.d, and $\epsilon\ N(0, \sigma^2)$

Question Three Answer the following questions:

(a) Who proposed the term "regression?" What is the regression phenomenon?

(b) Suppose we want to estimate β_0 and β_1 in model (1) (see question 2). Explain in words what are the least squares estimate, L_1 - norm estimate and robust estimate respectively, according to the following formulas:

L_1 - norm estimate

$$Q_1 = \sum_{i=1}^{n} |e_i| = \sum_{i=1}^{n} \left| y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right| \qquad (1.3)$$

Least squares estimate

$$Q_2 = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 \qquad (1.4)$$

Robust estimate

$$Q_r = \sum_{i=1}^{n} f(e_i), \qquad f(x) = \begin{cases} x^2, & |x| \le k \\ k^2, & |x| > k \end{cases}, \quad k \text{ given.} \qquad (1.5)$$

where $e_i = y_i - \hat{y}_i$, $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Answer:

(a) The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).

(b) The method of least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

$\mathrm{L}_1 -$*norm estimate*: First we have to sum of the absolute values of residuals at all the data points and then we use $\mathrm{L}_1 -$ *norm estimate*: to minimize them.

Question Four: Let $x_1, \ldots, x_n$ be n numbers.

(a) Prove:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\boldsymbol{2}} = \boldsymbol{x' D_n X}$$

where $\boldsymbol{x} = (\boldsymbol{x_1, \ldots, x_n})$ and $\boldsymbol{D_n} = \boldsymbol{I_n} = \frac{\boldsymbol{1}}{\boldsymbol{1}}\boldsymbol{_n 1'_n}$, where $\boldsymbol{I_n}$ is the identical matrix of order $\boldsymbol{n}$ and $\boldsymbol{1_n}$ is the n-vertor of one's.

(b) Prove that $\boldsymbol{D_n}$ is a projection matrix with rank $\boldsymbol{n-1}$

(c) Let $\boldsymbol{z} = \boldsymbol{D_n y}$, where y is an n-column vector. Show that the sample mean of **z** is zero and $\boldsymbol{z} = \boldsymbol{D_n z}$.

(d) Prove $\boldsymbol{B} = \frac{\boldsymbol{1}}{\boldsymbol{n}}\boldsymbol{1_n 1'_n}$ is a projection matrix and find its eigenvalues.

Answer:

(a) We can get this equations:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x}) &= \sum_{i=1}^{n}(x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\ &= \sum_{i=1}^{n} x_i^2 - 2x_i \sum_{i=1}^{n} \overline{x} + \sum_{i=1}^{n} \overline{x}^2 \\ &= x'x - n\overline{x}^2 \\ &= x'x - n(\frac{1}{n}1'_n x) \\ &= x'(1_n - \frac{1}{n}1_n 1'_n)x = x'D_n x\end{aligned}$$

(b) Proof:

$$\begin{aligned}D_n D_n &= (I_n - \frac{1}{n}1_n 1'_n)(I_n - \frac{1}{n}1_n 1'_n) \\ &= I_n \frac{2}{n}1_n 1'_n + \frac{1}{n^2}1_n 1'_n 1_n 1'_n \\ D'_n &= (I_n - \frac{1}{n}1_n 1'_n) = I_n - \frac{1}{n}1_n 1'_n = D_n\end{aligned}$$

We can notice that $Rank(A) = Trace(A) = n - \frac{1}{n} \times n = n - 1$
That means, D_n is a projection matrix and its rank is: $n - 1$

(c) The sample mean of z:

$$z = D_n y = \begin{bmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} \end{bmatrix} + \begin{bmatrix} y_1 \\ y_2 \\ \ddots \\ y_n \end{bmatrix} = \begin{bmatrix} (1 - \frac{1}{n})y_1 & () - \frac{1}{n})y_2 & \dots & -\frac{1}{n}y_n \\ -\frac{1}{n}y_1 & (1 - \frac{1}{n})y_2 & \dots & -\frac{1}{n}y_n \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n}y_1 & -\frac{1}{n}y_2 & \dots & 1 - \frac{1}{n}y_n \end{bmatrix}$$

we can get the mean by using sum values of z and divided by n:
$\overline{z} = \frac{(1-n\times\frac{1}{n})y_1 + (1-n\times\frac{1}{n}y_2 + \cdots + (1-n\times\frac{1}{n}y_n)}{n} = 0$

$$D_n z = \begin{bmatrix} (1-\frac{1}{n})y_1 & ()-\frac{1}{n})y_2 & \dots & -\frac{1}{n}y_n \\ -\frac{1}{n}y_1 & (1-\frac{1}{n})y_2 & \dots & -\frac{1}{n}y_n \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n}y_1 & -\frac{1}{n}y_2 & \dots & 1-\frac{1}{n}y_n \end{bmatrix} = z$$

(d)

$$B'B = \frac{1}{n^2}\mathbf{1}_n\mathbf{1}_n'\mathbf{1}_n\mathbf{1}_n = \frac{1}{n^2}(n\mathbf{1}_n\mathbf{1}_n') = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' = B$$

$$B' = \frac{1}{n}(\mathbf{1}_n\mathbf{1}_n')' = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n' = B$$

$$det(B-\lambda I_n) = \begin{bmatrix} \frac{1}{n}-\lambda & \frac{1}{n} & \dots & \frac{1}{n} \\ \frac{1}{n} = \frac{1}{n}-\lambda & \dots & \frac{1}{n} & \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \dots & \frac{1}{n}-\lambda \end{bmatrix}$$

$$\Rightarrow \lambda = \frac{1}{n}$$

It is noticeable that $B = \frac{1}{\mathbf{1}_n}\mathbf{1}_n'$ is the projection matrix and its eignvalue is $\frac{1}{n}$