Linear Regression based on Least Squares Method:

Connecting Calculus, Linear Algebra, and Statistics

Linear regression, one of the most important modeling skills, is almost the first model when analyzing multiple quantitative variables in statistics, and the logic of how it functions could be reasoned by calculus and also linear algebra.

In this performance assessment(PA), we are going to prove the mathematical foundation of linear regression by calculus and linear algebra respectively, and apply regression method on a dataset in Excel.

(14points)Part1 - Theoretical Part: Two Variables

There is a dataset of two variable X and Y with m subjects $(x_1, y_1), \dots, (x_m, y_m)$. We want a best-fit line $y = b_1 x + b_0$ to summarize the relationship between X and Y

1. (1point)If there is just one record (x_1, y_1), do you think it's possible to give the line? Why?
2. (1point)If there are exactly two different records (x_1, y_1), (x_2, y_2), do you think it's possible to give the line? Why?
3. (1point)If there are m $(m > 2)$ different records $(x_1, y_1), \dots, (x_m, y_m)$, do you think it's possible to find a line passing through all the records? Why and when?
4. For the Q1-Q3 above, they can also be understood from linear algebra. If we want all m records on the same line $b_1 x + b_0 = y$, it means we have built a system of m equations and pursue its solution. Please use the matrix form of system of equation $A\boldsymbol{x} = \boldsymbol{b}$, and
    a. (1point)Use $(x_1, y_1), \dots, (x_m, y_m)$, b_1, b_0 and other constant(s) to give the matrix A (m by 2)
    b. (1point)Use $(x_1, y_1), \dots, (x_m, y_m)$, b_1, b_0 and other constant(s) to give the vector $\boldsymbol{x}$ and $\boldsymbol{b}$
    c. (2points)As what we have known in linear algebra, the No. of solutions for $A\boldsymbol{x} = \boldsymbol{b}$ depends on the value of m. When $m = 1$, or $m = 2$, or $m > 2$, give the rank of A and the No. of solutions for $A\boldsymbol{x} = \boldsymbol{b}$ respectively
5. When there are m $(m > 2)$ different records $(x_1, y_1), \dots, (x_m, y_m)$ which are not on the same line, i.e., the most common case in real life, we still want a best-fit line to summarize the relationship between X and Y. That's where we need a criterion to find the only line!
   This criterion is **least squares**: The best-fit line $y = b_1 x + b_0$ is the line (i.e., the best choice for the value of b_1 and b_0 because the line is set when the values of b_1 and b_0 are

chosen) which let minimize the $\sum_{i=1}^{m}\left(y_i-(b_1x_i+b_0)\right)^2$ [1]. Let's compute the values of b_1 and b_0 by Pre-calculus method, calculus, and linear algebra respectively.

a. (2points)Pre-cal Method:

   When we want the best b_1 and b_0 to minimize $\sum_{i=1}^{m}\left(y_i-(b_1x_i+b_0)\right)^2$, (x_1, y_1), …, (x_m, y_m) are all knowns, and b_1 and b_0 are unknown, which means $\sum_{i=1}^{m}\left(y_i-(b_1x_i+b_0)\right)^2$ could be understood as a function of b_1 and b_0, i.e., $f(b_1,b_0)=\sum_{i=1}^{m}\left(y_i-(b_1x_i+b_0)\right)^2$

   i. Set $t_i=y_i-b_1x_i$, prove that $f(b_1,b_0)$ is a quadratic function of b_0 and give the standard form of this quadratic function.[2]

   ii. What value of b_0 will minimize the quadratic function

   iii. Input the value of b_0 you have got in (ii.) into $f(b_1,b_0)=\sum_{i=1}^{m}\left(y_i-(b_1x_i+b_0)\right)^2$ to simplify it as only a function of b_1. Prove that it's also a quadratic function of b_1 and give the standard form of this quadratic function

   iv. What value of b_1 will minimize the quadratic function

Now , the values of b_1 and b_0 are both been set.

b. (2points)Calculus Method:

   From calculus it's easy to know that the minimum of $f(b_1,b_0)=\sum_{i=1}^{m}\left(y_i-(b_1x_i+b_0)\right)^2$ can only be got when the derivatives of f with respect to b_1 and b_0 are both 0.

   i. Take b_0 as one known, give the derivative of f with respect to b_1, write it as $\frac{\partial f}{\partial b_1}$

   ii. Set $\frac{\partial f}{\partial b_1}=0$ and compute the value of b_1

   iii. Take b_1 as one known, give the derivative of f with respect to b_0, write it as $\frac{\partial f}{\partial b_0}$

   iv. Set $\frac{\partial f}{\partial b_0}=0$ and compute the value of b_0

Now , the values of b_1 and b_0 are been set, and should be the same as answer in (a)

c. (3points)Linear algebra method:

   When there are $m>2$ equations in the linear system and these m records are not on the same line, there is no solution for the system $A\boldsymbol{x}=\boldsymbol{b}$**.** Then the only thing we can do is to find an approxiamte solution $\hat{\boldsymbol{x}}$ (so called "least squares solution) to minimize the error $\boldsymbol{e}=\boldsymbol{b}-A\hat{\boldsymbol{x}}$

   i. Explain why the solution $\hat{\boldsymbol{x}}$ for $A^TA\hat{\boldsymbol{x}}=A^T\boldsymbol{b}$ can minimize the error $\boldsymbol{e}=\boldsymbol{b}-A\hat{\boldsymbol{x}}$ based on the knowledge in Section4.2 & 4.3

---

[1] If you would like to know why least squares criterion is commonly used rather than other criterion, you may log into the Desmos class by https://student.desmos.com/?prepopulateCode=m8t5ay and code: M8T5AY. There is an activity named "How to find the best-fit line in linear regression" designed by Ignacio.

[2] Standard form of quadratic function is $f(x)=ax^2+bx+c$

ii. When will the equation $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ have only one solution? Then what kind of dataset $(x_1, y_1), \dots, (x_m, y_m)$ can ensure there is one and only one best-fit line?

iii. Give the only solution $\hat{\boldsymbol{x}}$ when the requirement in (ii.) is met.

Now , the values of b_1 and b_0 are been set. Check them with the answers above.[3]

(8points)Part2 - Theoretical Part: Multiple Variables

There is a dataset of multiple variable $X_1, \dots, X_k$ and Y with m subjects $(x_{11}, \dots, x_{k1}, y_1), \dots, (x_{1,m}, \dots, x_{k,m}, y_m)$. We want a best-fit line $y = b_1x_1 + b_2x_2 + \cdots + b_kx_k + b_0$ to summarize the relationship between $X_1, \dots, X_k$ and Y. Because there are multiple explanatory variables $X_1, \dots, X_k$, linear algebra has its strength. Let's write the system as $A\boldsymbol{x} = \boldsymbol{b}$

1. (1point)Use $(x_{11}, \dots, x_{k1}, y_1), \dots, (x_{1,m}, \dots, x_{k,m}, y_m)$, $b_k, \dots, b_1, b_0$ and other constant(s) to give the matrix A (m by k)
2. (1point)Use $(x_{11}, \dots, x_{k1}, y_1), \dots, (x_{1,m}, \dots, x_{k,m}, y_m)$, $b_k, \dots, b_1, b_0$ and other constant(s) to give the vector $\boldsymbol{x}$ and $\boldsymbol{b}$
3. (2points)As what we have known in linear algebra, the No. of solutions for $A\boldsymbol{x} = \boldsymbol{b}$ depends on the relationship of m and k. When $m < k$, or $m = k$, or $m > k$, give the rank of A and the No. of solutions for $A\boldsymbol{x} = \boldsymbol{b}$ respectively
4. When there are m $(m > k)$ different records $(x_1, y_1), \dots, (x_m, y_m)$ which are not on the same line, i.e., the most common case in real life, we still want a best-fit line to summarize the relationship among $X_1, \dots, X_k$ and Y. That's where we need a criterion to find the only line!
   This criterion is **least squares** too: The best-fit line $y = b_1x_1 + b_2x_2 + \cdots + b_kx_k + b_0$ is the line (i.e., the best choice for the values of $b_k, \dots, b_1$ and b_0 because the line is set when the all values are chosen) which let minimize the $\sum_{i=1}^{m}\left(y_i - (b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + b_0)\right)^2$
   Or to say, there is no solution for the system $A\boldsymbol{x} = \boldsymbol{b}$**.** Then the only thing we can do is to find an approxiamte solution $\hat{\boldsymbol{x}}$ (so called "least squares solution) to minimize the error $\boldsymbol{e} = \boldsymbol{b} - A\hat{\boldsymbol{x}}$
   i. (1point)Explain why the solution $\hat{\boldsymbol{x}}$ for $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ can minimize the error $\boldsymbol{e} = \boldsymbol{b} - A\hat{\boldsymbol{x}}$ based on the knowledge in Section4.2 & 4.3
   ii. (1point)When will the equation $A^TA\hat{\boldsymbol{x}} = A^T\boldsymbol{b}$ have only one solution? Then what kind of dataset $(x_1, y_1), \dots, (x_m, y_m)$ can ensure there is one and only one best-fit line?
   iii. (1point)Give the only solution $\hat{\boldsymbol{x}}$ when the requirement in (ii.) is met.
   iv. (1point)What will happen if there is a hidden relationship bewteen the variable X_1 and X_2: $X_2 = 100X_1$(For example, X_1 is the height in meters, and X_2 is the height in centimeters). Will the linear regression still work?

[3] The inverse of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is $\frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$