# Class 26 in-class problems, 18.05, Spring 2022

## Board questions

**Problem 1. Make it fit**

*We are given bivariate data: $(1,3),\ (2,1),\ (4,4)$.*

***(a)** Do (simple) linear regression to find the best fitting line.*

*(i) Give the model for simple linear regression.*

*(ii) Write down the formula for the total squared error.*

*(iii) Use calculus to find the parameters that minimize the total squared error.*

***(b)** Do linear regression to find the best fitting parabola. (Really just set this up and get as far as needing to solve equations to find the coefficients.)*

***(c)** Find the best fitting exponential $y = e^{ax+b}$. (As before, set up the equations but don't solve them.)*

*Hint: take $\ln(y)$ and do simple linear regression.*

***(d)** For data $(x_1, y_1), \dots, (x_n, y_n)$. Set up the linear regression to find the best fitting cubic. Don't try to take derivatives or actually find the formulas for the coefficients.*

**(a) Solution:** Model $y_i = ax_i + b + e_i$, where $a$, $b$ are constants and $e_i$ is random error. We then have the prediction $\hat{y}_i = ax_i + b$.

$$\begin{aligned} \text{Total squared error } = T &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - ax_i - b)^2 \\ &= (3 - a - b)^2 + (1 - 2a - b)^2 + (4 - 4a - b)^2 \end{aligned}$$

Take the partial derivatives and set to 0:

$$\frac{\partial T}{\partial a} = -2(3 - a - b) - 4(1 - 2a - b) - 8(4 - 4a - b) = 0$$
$$\frac{\partial T}{\partial b} = -2(3 - a - b) - 2(1 - 2a - b) - 2(4 - 4a - b) = 0$$

A little arithmetic gives the system of simultaneous linear equations and solution:

$$\begin{aligned} 21a \quad &+7b \quad &= 21 \\ 7a \quad &+3b \quad &= 8 \end{aligned}$$

Solving, we find $a = 1/2,\ b = 3/2$.

So, the least squares best fitting line is $y = \dfrac{1}{2}x + \dfrac{3}{2}$.

**(b) Solution:** Model $y_i = ax_i^2 + bx_i + c + e_i$, where $a$, $b$, $c$ are constants and $e_i$ is random error. So, our prediction is $\hat{y}_i = ax_i^2 + bx_i + c$.

Total squared error:

$$\begin{aligned} T &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum (y_i - ax_i^2 - bx_i - c)^2 \\ &= (3 - a - b - c)^2 + (1 - 4a - 2b - c)^2 + (4 - 16a - 4b - c)^2 \end{aligned}$$

We didn't really expect people to carry this all the way out by hand. If you did you would have found that taking the partial derivatives and setting them equal to 0 gives

$$\frac{\partial T}{\partial a} = -2(3-a-b-c) - 8(1-4a-2b-c) - 32(4-16a-4b-c) = 0$$
$$\frac{\partial T}{\partial b} = -2(3-a-b-c) - 4(1-4a-2b-c) - 8(4-16a-4b-c) = 0$$
$$\frac{\partial T}{\partial c} = -2(3-a-b-c) - 2(2-4a-2b-c) - 2(4-16a-4b-c) = 0$$

We didn't really expect people to go beyond this. If you did: cleaning up the equations gives

$$\begin{array}{llll} 273a & +73b & +21c & = 71 \\ 73a & +21b & +7c & = 21 \\ 21a & +7b & +3c & = 8 \end{array}$$

Solving gives, $a = 1.1667,\ b = -5.5,\ c = 7.3333$.

The least squares best fitting parabola is $y = 1.1667x^2 - 5.5x + 7.3333$.

**(c)** Model $\ln(y_i) = ax_i + b + e_i$, where $a$, $b$ are constants and $e_i$ is random error. So, our prediction is $\hat{y}_i = \mathrm{e}^{ax_i+b}$.

Total squared error:

$$\begin{aligned} T &= \sum(\ln(y_i) - \ln(\hat{y}_i))^2 \\ &= \sum(\ln(y_i) - ax_i - b)^2 \\ &= (\ln(3) - a - b)^2 + (\ln(1) - 2a - b)^2 + (\ln(4) - 4a - b)^2 \end{aligned}$$

Now we can find $a$ and $b$ as before. (Using R: $a = 0.18$, $b = 0.41$)

**(d)** Model $y_i = ax_i i^3 + bx_i^2 + cx_i + d + e_i$, where $a$, $b$, $c$, $d$ are constants and $e_i$ is random error. So, our prediction is $\hat{y}_i = ax_i^3 + bx_i^2 + cx_i + d$.

Total squared error:

$$\begin{aligned} T &= \sum(y_i - \hat{y}_i)^2 \\ &= \sum(y_i - ax_i^3 - bx_i^2 - cx_i - d)^2 \end{aligned}$$

Now we could set the partial derivatives to 0 and solve for $a$, $b$, $c$, $d$.

**Problem 2. Using the formulas plus some theory**

*Bivariate data:* $(1,3),\ (2,1),\ (4,4)$

***(a)*** *Calculate the sample means for $x$ and $y$.*

***(b)*** *Use the formulas to find a best-fit line in the $xy$-plane.*

$$\hat{a} = \frac{s_{xy}}{s_{xx}} \qquad \hat{b} = \overline{y} - \hat{a}\overline{x}$$
$$s_{xy} = \frac{1}{n-1}\sum(x_i - \overline{x})(y_i - \overline{y}) \qquad s_{xx} = \frac{1}{n-1}\sum(x_i - \overline{x})^2.$$

***(c)*** *Show the point $(\overline{x}, \overline{y})$ is always on the fitted line.*

***(d)*** *(For fun later!) Under the assumption $E_i \sim N(0, \sigma^2)$ show that the least squares method is equivalent to finding the MLE for the parameters $(a, b)$.*

*Hint: $f(y_i \,|\, x_i, a, b) \sim N(ax_i + b, \sigma^2)$.*

**(a) Solution:** $\bar{x} = 7/3$, $\bar{y} = 8/3$.

**(b)** $s_{xx} = 7/3, \quad s_{xy} = 7/6$. So,

$$\hat{a} = \frac{s_{xy}}{s_{xx}} = 1/2, \quad \hat{b} = \bar{y} - \hat{a}\bar{x} = 3/2.$$

(The same answer as in the previous problem.)

**(c) Solution:** The formula $\hat{b} = \bar{y} - \hat{a}\bar{x}$ can be changed to $\bar{y} = \hat{a}\bar{x} + \hat{b}$. That is, the point $(\bar{x}, \bar{y})$ is on the line $y = \hat{a}x + \hat{b}$

**(d) Solution:** Our model is $y_i = ax_i + b + E_i$, where the $E_i$ are independent. Since $E_i \sim \mathrm{N}(0, \sigma^2)$ this becomes

$$y_i \sim \mathrm{N}(ax_i + b, \sigma^2)$$

Therefore the likelihood of $y_i$ given $x_i$, $a$ and $b$ is

$$f(y_i \mid x_i, a, b) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(y_i - ax_i - b)^2}{2\sigma^2}}$$

Since the data $y_i$ are independent the likelihood function is just the product of the expression above, i.e. we have to sum exponents

$$\text{likelihood } = f(y_1, \dots, y_n \,|x_1, \dots, x_n, a, b) = \mathrm{e}^{-\frac{\sum_{i=1}^{n}(y_i - ax_i - b)^2}{2\sigma^2}}$$

Since the exponent is negative, the maximum likelihood will happen when the exponent is as close to 0 as possible. That is, when the sum

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2$$

is as small as possible. This is exactly what we were asked to show.

18.05 Introduction to Probability and Statistics
Spring 2022