University of Pennsylvania
BIOL4536 Fall 2023

HW#2 (SOLUTIONS)
(Regression)

Problem 1. (3 points)

Using the formula of expected value for a discrete random variable (on slide #11 of the Regression lecture) prove that if all values in the range of a random variable X are equally likely, then $E[X]$ equals the average of the values in the range of X.

ANSWER: The formula on slide 11 is

$$E[X] = \sum xp(X = x)$$

If all $p(X = x)$ are equal, then the random variable X must have a finite number of values N and it must be that $p(X = x) = 1/N$ for all x in the range of X. Thus,

$$E[X] = \sum xp(X = x) = \sum x/N = \frac{1}{N} \sum x$$

and this last expression is exactly the average of the values of X.

Problem 2. (4 points)

Consider the regression model

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

(i) What component of this model is the source of all randomness?

ANSWER: ϵ is the only random part of the right hand side.

(ii) How many unknown parameters must be specified to completely specify the model?

ANSWER: β_0, β_1 and σ (the standard deviation of ϵ).

(iii) True or False: If $\sigma = 0$ then Y is a function of X?

ANSWER: True, because that means ϵ always equals its mean (which is zero), therefore there is no randomness in the model and Y becomes a function of X.

(iv) What is the form (shape) of the regression curve of the following model?

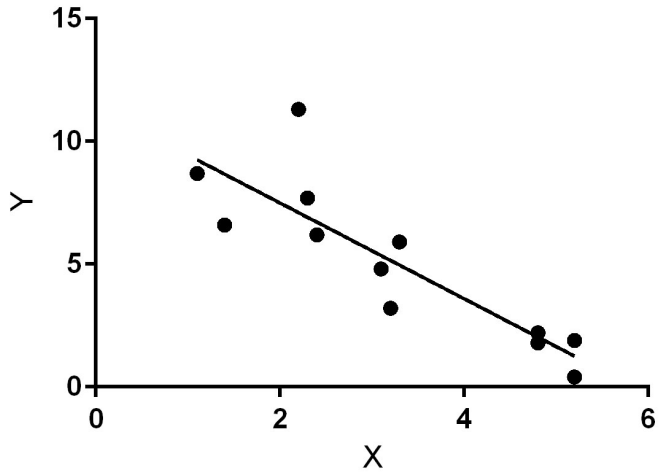
$$Y = \beta_0 + \beta_1 2^X + \epsilon$$

ANSWER: The regression curve is $Y = \beta_0 + \beta_1 2^X$ which is exponential.

Problem 3. (3 points)

Consider this data, two measurements from 12 subjects:

	X	Y
subject #1	1.4	6.6
subject #2	1.1	8.7
subject #3	2.2	11.3
subject #4	2.3	7.7
subject #5	2.4	6.2
subject #6	3.2	3.2
subject #7	3.1	4.8
subject #8	3.3	5.9
subject #9	4.8	2.2
subject #10	4.8	1.8
subject #11	5.2	1.9
subject #12	5.2	0.4



Use this Linear Regression Caclulator:

`https://www.graphpad.com/quickcalcs/linear1/`

(ii) According to our notation, what are the estimates of β_0 and β_1?

ANSWER: $b_0 = 11.4$ and $b_1 = -1.95$

(i) What is the regression model?

ANSWER: The regression model is $Y = 11.4 - 1.95X + \epsilon$.

(iii) Is there a significant regression relation between the variables?

ANSWER: The p-value for $H_0 : \beta_1 = 0$ is 0.0003, which is significant.

(iv) Using the model, what is the estimate of Y when $X = 4$?

ANSWER: Plug in $X = 4$ to the regression equation: $Y = -1.95 \times 4 + 11.4 = 3.6$

(v) Which is definitely true:
The probability that the Y-Intercept is > 15 is
(A) ≤ 0.05 ⟵ **This one, because** 15 **is outside the 95% confidence interval.**
(B) ≥ 0.05