1 A lucky strike [11 Pts]

We have discovered rich stores of the valuable and rare mineral unobtanium. There are two different types of unobtanium: unobtanium-A, which is stable, and unobtanium-B, which is highly unstable. We carry out an experiment by starting a reaction that has no effect on unobtanium-A, but causes unobtanium-B to rapidly degrade with a "half-life" of γ milliseconds (ms) as soon as the reaction starts. That is, only half of the initial amount of unobtanium-B remains after γ ms, only one quarter remains after 2γ ms, and so on. If β_0 is the initial amount of unobtanium-A, and β_1 is the initial amount of unobtanium-B, then the total amount of unobtanium remaining after x milliseconds is given by $f_\theta(x)$ the formula:

$$f_\theta(x) = \beta_0 + \beta_1 2^{-x/\gamma}$$

Here $\theta = (\beta_0, \beta_1, \gamma)$, encoding all three model parameters.

We start the reaction, and at ten-millisecond intervals we take a noisy measurement of the total amount of unobtanium remaining. This gives us a data set with five data points, shown in the table below. x is the time since the reaction began, measured in milliseconds (ms), and y is the measured amount of unobtanium remaining x ms after the reaction begins, measured in milligrams (mg).

x	y
0	19
10	13
20	10
30	5
40	5

Unobtanium reaction data

For easy reference, the first four negative powers of 2 are given below:

$$2^{-1} = \frac{1}{2} = 0.5 \qquad 2^{-3} = \frac{1}{8} = 0.125$$

$$2^{-2} = \frac{1}{4} = 0.25 \qquad 2^{-4} = \frac{1}{16} = 0.0625$$

(a) [3 Pts] From briefly inspecting the data, we come up with a rough guess that the parameters are close to $\beta_0 = 4$, $\beta_1 = 16$, and $\gamma = 10$. **What is the MSE** for these parameters, in units of mg^2? Remember that the mean squared error is defined as

$$\text{MSE}(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where n is the number of observations, and $\hat{y}_i = f_\theta(x_i)$ is the prediction we would have made for y_i if we used the parameters $\theta = (\beta_0, \beta_1, \gamma)$.

○ 0.2

○ 0.6

○ 1

○ **1.4**

○ 2

You will get full credit for choosing the right answer, but you can show your work in the box below if you want to get partial credit in case your answer is wrong:

Solution: The predictions would be $\hat{y} = (20, 12, 8, 6, 5)$, so the residuals are $e = (-1, 1, 2, -1, 0)$. The mean squared error is $(1 + 1 + 4 + 1 + 0)/5 = 7/5 = 1.4\,\text{mg}^2$.

(b) [3 Pts] Because our formula is not a linear model, we cannot use our closed-form expression for the MSE-minimizing model parameters. Instead, we could use gradient descent to search for the optimal model parameters, starting at our guess $\beta_0^{(0)} = 4, \beta_1^{(0)} = 16$, and $\gamma^{(0)} = 10$. **What will be the value** of $\beta_0^{(1)}$, the intercept parameter after one step of gradient descent? Assume we use learning rate $\alpha = 0.5$.

Hint: You will save some time if you don't calculate more derivatives than you need to.

Solution: Again, the residuals are $e = (-1, 1, 2, -1, 0)$, and we can calculate the gradient by taking partial derivatives:

$$\begin{aligned}\frac{\partial}{\partial \beta_0} \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 e^{-\beta_2 x_i})^2 &= \frac{-2}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 e^{-\beta_2 x_i}) \\ &= \frac{-2}{n} \sum_{i=1}^{n} e_i \\ &= \frac{-2}{5}\end{aligned}$$

If we use the learning rate 0.5, then $\beta_0^{(1)} = \beta_0^{(0)} - 0.5(-\frac{2}{5}) = 4.2.$

(c) [2 Pts] Now suppose that, before we fit our model, we learn that another research group has discovered that the half-life of unstable unobtanium is exactly ten milliseconds ($\gamma = 10$). Now, instead of estimating γ, we can just plug in $\gamma = 10$. Better still, we don't have to do gradient descent anymore; we can estimate $\beta = (\beta_0, \beta_1)$ using linear regression.

If we construct the right matrix $\mathbb{X}$, we will get a closed form for our estimate:

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y},$$

where $\mathbb{Y}$ is the column vector (or $n \times 1$ matrix) representing the response. Which of the following choices is the correct matrix $\mathbb{X}$?

○ $\begin{array}{|c|}\hline 0 \\ 10 \\ 20 \\ 30 \\ 40 \\ \hline\end{array}$

○ $\begin{array}{|c|c|}\hline 1 & 0 \\ 1 & 10 \\ 1 & 20 \\ 1 & 30 \\ 1 & 40 \\ \hline\end{array}$

○ $\begin{array}{|c|c|c|}\hline 1 & 0 & 0 \\ 1 & 10 & 100 \\ 1 & 20 & 400 \\ 1 & 30 & 900 \\ 1 & 40 & 1600 \\ \hline\end{array}$

○ $\begin{array}{|c|c|}\hline \mathbf{1} & \mathbf{1} \\ \mathbf{1} & \mathbf{0.5} \\ \mathbf{1} & \mathbf{0.25} \\ \mathbf{1} & \mathbf{0.125} \\ \mathbf{1} & \mathbf{0.0625} \\ \hline\end{array}$

○ $\begin{array}{|c|c|}\hline 0 & 1 \\ 10 & 0.5 \\ 20 & 0.25 \\ 30 & 0.125 \\ 40 & 0.0625 \\ \hline\end{array}$

○ $\begin{array}{|c|c|c|}\hline 1 & 0 & 1 \\ 1 & 10 & 0.5 \\ 1 & 20 & 0.25 \\ 1 & 30 & 0.125 \\ 1 & 40 & 0.0625 \\ \hline\end{array}$

Solution: With $\gamma = 10$, our model is $f_\theta(x) = \beta_0 + \beta_1 2^{-x/10}$. Our design matrix would therefore have two columns: one bias/intercept column of all ones, one column with our feature $2^{-x/10}$: $\{2^0, 2^{-1}, \ldots, 2^{-4}\}$.

(d) [3 Pts] Continue to assume that we know $\gamma = 10$ so we are only estimating β_0 and β_1. Even before we estimate the model, **what can we be sure will be true** about the fitted values $\hat{y}$ and the residuals $e = y - \hat{y}$ for the estimated model? **Select all that apply.**

Note: In the following options, x represents time in milliseconds.

□ If we plot the residuals e against the time variable x, we will not see any clear pattern in the plot.

□ **The residuals will add up to zero.**

□ The residuals will be uncorrelated with x.

□ The residuals will be highly correlated with the fitted values $\hat{y}$.

□ The residuals in the left half of the plot will not be as spread out as the ones in the right half of the plot.

□ **The fitted values $\hat{y}$ will have smaller variance than the original y values.**

Solution: Again, our model is $f_\theta(x) = \beta_0 + \beta_1 2^{-x/10}$. Note the only feature is $2^{-x/10}$ and the time x itself is not a feature.

Option A is wrong because x is not a feature; we can't be sure if there will be a pattern in the residual plot. Even if x is a feature, there will still be patterns in the residual plot if the relationship between the response and the predictor is not linear.

Option B is a fact from OLS.

Option C is wrong. It is a fact that in OLS the residuals are uncorrelated with the predictor variables (proved in HW6), but here the predictor variable is not x (it's $2^{-x/10}$) so we can't be sure about the correlation between the residuals and x.

Option D is wrong. Since the fitted values $\hat{y}$ are in the column space of the design matrix, the residuals will be orthogonal to the fitted values and therefore uncorrelated.

Option E is wrong. We can't be sure of the shape and spread of the residual plot before fitting the model.

Option F is correct. We showed that the variance of the original y values are the sum of the variance of the fitted values $\hat{y}$ and the variance of the residuals ($\sigma_y^2 = \sigma_{\hat{y}}^2 + \sigma_e^2$). Therefore, $\sigma_y^2 > \sigma_{\hat{y}}^2$.