Question 1.
What are the problems with simply ignoring missing data? Can this cause biases? Does it affect the standard error?

Suppose there is a missing data mechanism $P(M|Y, \phi)$ where $M = 0$ if the data is present, and $M = 1$ is the data is missing. Give the definitions of MAR, MCAR, and NMAR.

Question 2.
Consider a biased coin where the probability of a head is $\theta$ with $0 < \theta < 1$. Let $Y_i$ be a random variable so $Y_i = 1$ means “heads” and $Y_i = 0$ means “tails”. This gives a model $P(Y|\theta) = \theta^{Y_i}(1 - \theta)^{1-Y_i}$.

Describe how to do Maximum Likelihood (ML) to estimate $\theta$ from a sequence $Y_1, ..., Y_n$ of independently identically distributed (i.i.d.) coin tosses. What is the ML estimate of $\theta$ if we get $k$ heads in $n$ tosses?

Question 3.
Now suppose that there is a missing data mechanism $P(M|Y, \phi)$. For example, a friend is tossing a biased coin and is only showing you some of the tosses (i.e. the tosses your friend does not show you are missing data).

Describe how to model this problem by combining the distributions $P(M|Y, \phi)$ and $P(Y|\theta)$, and splitting the data $Y$ into observable $Y_{obs}$ and missing $Y_{mis}$. What simplifications occur if the missing data is MCAR or MAR?

Suppose the missing data mechanism is $P(M = 0|Y, \phi) = \phi$ (with $0 < \phi < 1$). Is this MCAR or MAR? What are the maximum likelihood estimates of $\phi$ for this case? Now suppose the missing data mechanism is $P(M = 0|Y, \phi) = \phi^{Y_i}(1 - \phi)^{1-Y_i}$? Is this MCAR or MAR? How much can you say about the maximum likelihood estimate?

Question 4.
Consider the problem of estimating the distribution of the heights of a population from the sample obtained from army records. Assume that the distribution of height $Y$ is Gaussian with unknown parameters $\theta = (\mu, \sigma)$: $p(Y|\theta) = \frac{1}{\sqrt{2\pi\sigma}}e^{-(Y-\mu)^2/2\sigma^2}$, with $\theta = (\mu, \sigma)$.

Now assume a missing data mechanism: $P(M = 0|Y, \phi) = \phi(Y - c : \lambda)$ where $\phi = (c, \lambda)$ and $\sigma(Y - c : \lambda) = \frac{1}{1+e^{-\lambda(Y-c)}}$. Is this missing data mechanism MAR, MCAR or NMAR?

Give formulas for the best estimates of $\phi, \theta$ in terms of the observed data $Y_{obs}$, (note it is impossible to get closed form solutions for them).
Question 5.
Suppose we observe $X$ and we want to estimate $Y$. We hypothesize that there is a regression relationship $Y = \beta X + e$. Where $e$ is randomly distributed with mean zero and unknown variance $\sigma^2$.

First, describe the least squares formulation of regression. Give the least squares criterion and obtain the solution for the regression coefficient $\beta$ in terms of the $n$ i.i.d. samples $\{(X_i, Y_i): i = 1, ..., n\}$. Give a formula for the estimation of the variance.

Second, describe the Maximum Likelihood (ML) perspective. Define a probability distribution $P(Y|X, \beta, \sigma) = \frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(Y-\beta X)^2}{2\sigma^2}}$. Give the probability of generating the $n$ i.i.d. samples. Perform Maximum Likelihood to estimate $\beta, \sigma$. Show that this reduces to the least squares estimate for $\beta$ and a formula for $\sigma$.

Now suppose that data samples are missing with mechanism $P(M|Y, X, \phi) = P(M|X, \phi)$. Is this MAR, MCAR, or NMAR? How does the solution of ML relate to the solution when there is no missing data?