PROBLEMS

Due Friday. Jan. 27, 2023 at 4pm

1. Suppose we are given the following data:

x_1	0	1	0	1
x_2	0	0	1	1
y	2	4	4	5

   (a) Write an equation describing the linear model for y in terms of x_1 and x_2.

   (b) Given the data, compute the least-squares fit of the model parameters using a short Python program, and show the details of your implementation.

   (c) Compute the R^2 coefficient of determination for the least-squares linear model.

2. Suppose that an insurance company wants to predict the lifespan of prospective clients. For 20000 individuals, they have data on blood pressure, resting heart rate, body mass index (BMI), and gender, all taken when they were 40 years old. They also have the age at which they died.

   (a) What is the target variable y? If $\{y_i\}$ is part of the training data, what does i represent? What is the range of i?

   (b) Suppose that they first consider only blood pressure, heart rate, and BMI. From this data, describe how they could construct a linear model to predict the target.

   (c) Now suppose that they want to also incorporate the gender, which is either male or female. Describe how the linear model would change. (Note that there are various options for how to do this. Whichever one you choose, explain your reasoning.)

3. In spectroscopy, material properties are inferred by analyzing the spectrum of electromagnetic radiation when the material is stimulated. This can be formulated as follows: Suppose we are given samples of the radiation waveform y_k for $k = 0, \ldots, N-1$, and we wish to fit a model of the form

$$y_k \approx \sum_{\ell=1}^{L} a_\ell \cos(\Omega_\ell k) + b_\ell \sin(\Omega_\ell k), \tag{1}$$

   where L are the number of tones present, Ω_ℓ are the tonal frequencies, and a_ℓ and b_ℓ are the coefficients. (By using sin and cos terms, we can model an arbitrary phase for each sinusoid.)

   (a) Show that if the frequencies Ω_ℓ are known, then one can solve for the coefficients a_ℓ and b_ℓ using linear regression. Specifically, rewrite the model (1) as $\mathbf{y} \approx \mathbf{A}\boldsymbol{\beta}$ for appropriate $\mathbf{y}$, $\mathbf{A}$, and $\boldsymbol{\beta}$. Then describe exactly how one would fit the coefficients a_ℓ and b_ℓ from this model.

(b) Now suppose that the frequencies Ω_ℓ are unknown. If we had to solve for the parameters a_ℓ, b_ℓ, and Ω_ℓ, would this be a linear regression problem? Explain your answer.

4. An audio engineer wants to model the acoustics of a room. For this purpose, she plays a digital audio waveform u_t at times $t = 0, 1, \ldots, T-1$ through a speaker, and records the echoed response y_t at times $t = 0, 1, \ldots, T-1$ with a microphone. The measurements are made at some sampling rate, say once every 100 microseconds. The engineer then wants to fit a model of the form

$$y_t \approx \sum_{j=1}^{M} a_j y_{t-j} + \sum_{k=0}^{N} b_k u_{t-k} \tag{2}$$

by choosing coefficients a_j and b_k. This is known as an *auto-regressive moving average (ARMA)* model or an *infinite-impulse response (IIR) filter*.

(a) Construct a vector $\boldsymbol{\beta}$ containing the unknown parameters. How many unknown parameters are there?

(b) Construct the corresponding matrix $\mathbf{A}$ and target vector $\mathbf{y}$ so that the model (2) can be written as

$$\mathbf{y} \approx \mathbf{A}\boldsymbol{\beta}$$

and so that the parameters $\boldsymbol{\beta}$ can be fitted via LS as follows.

$$\arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{A}\boldsymbol{\beta}\|^2.$$

To simplify your answer, you can assume that $T-1 \geq M \geq N$. Hint: your matrix $\mathbf{A}$ should contain entries from both the y_t and u_t sequences, but only for $t = 0, 1, \ldots, T-1$. We do not know the values of y_t and u_t for other t, and we do not want to make any assumptions about these values!