Statistics Module

Homework 2.

Instructions: *This is an optional homework; turning it in or not will not affect your grade. If you do turn it in on time, we will grade it with feedback. Your answers must be handwritten; you can either turn in a hardcopy to Prof. Cowley's office (Freeman 113) or e-mail Prof. Cowley, the TAs, and Razan with a scan. Any use of resources is permitted.*

This homework is due by **Friday, September 19, 11:59pm**.

Problem 1.

Consider the Kaplan-Meier estimator:
$\hat{S}(t) = \prod_{i:t_i \leq t} (1 - \frac{d_i}{n_i})$
Explain the meaning of each of the variables. Then, consider consider time events $\{t_0 = 0, t_1 = 2, t_2 = 3, t_3 = 5, t_4 = 8, t_5 = 10\}$ where 5 other subjects remained at $t_6 = 15$ (a total of 10 subjects).
Plot the estimated survival function.

Problem 2.

Part 1.

You are evaluating your predicted responses $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_N$ versus true responses $y_1, y_2, \ldots, y_N$ for N samples. You use the coefficient of determination:
$R^2_{\text{COD}}(y, \hat{y}) = 1 - \frac{E[(y-\hat{y})^2]}{E[(y-\bar{y})^2]}$ where $\bar{y}$ is the expected value/mean of y.
However, something weird is happening. You are getting a negative R^2?! Let's explore how this could happen. Assume $\bar{y} = 0$ and $V[y] = \sigma^2$.

What is $R^2_{\text{COD}}(y, \hat{y})$ when $\hat{y} = y + 1$?

What is $R^2_{\text{COD}}(y, \hat{y})$ when $\hat{y} = 2y$?

What is $R^2_{\text{COD}}(y, \hat{y})$ when $\hat{y} = -y$?

What do you conclude from these results?

Part 2.
You are wondering if another metric may be more approriate to see how $\hat{y}$ covaries with y. You heard from a friend that Pearson's correlation squared ($\rho^2(y, \hat{y})$) is actually the metric you seek. Let's see how this metric fairs.

What is $\rho^2(y, \hat{y})$ when $\hat{y} = y + 1$?

What is $\rho^2(y, \hat{y})$ when $\hat{y} = 2y$?

What is $\rho^2(y, \hat{y})$ when $\hat{y} = -y$?

Would you use Pearson's correlation squared ρ^2 as your metric? Why or why not?

Problem 3.

You have K feature variables $\mathbf{x} \in \mathcal{R}^K$ with N samples. You want to predict multiple output/target variables $\mathbf{y} \in \mathcal{R}^M$. One option is to train a ridge regression model on each variable individually. Another option is to train a single ridge regression model on all output variables at once...this is called multivariate linear regression.
What are the benefits and disadvantages of either approach? Which one would you choose and why?

Problem 4.

You have K feature variables $\mathbf{x} \in \mathcal{R}^K$ with N samples as well as $y \in \mathcal{R}$ where $(\mathbf{x}_n, y_n)$ were collected together in pairs. You would like to train a ridge regression model's weights β with the best possible hyperparameter λ and then report your evaluated R^2 prediction performance.
Give pseudocode of this procedure and explain its steps.

Challenge Problem 5.

Assume that the data follow a linear model:
$y = X\beta + \epsilon$, where $X : (N \text{ samples} \times K \text{ features})$, $\beta : (K \times 1)$, and $\epsilon \sim N(0, \sigma_\epsilon^2)$
Given data matrix X, we want to know how bad our estimates of β may be.
Compute the expected error $E[(\beta - \hat{\beta})^2 | (X, \mathbf{y})]$, where $\mathbf{y} \in \mathcal{R}^N$ are the responses for N samples. The expected value is over any random variables (hint: ϵ), and should have an expression containing σ_ϵ^2 and the X covariance matrix Σ (hint: substitute in an expression for $\hat{\beta}$ and $\mathbf{y}$). Provide intuition for this final expression.