36-401 Homework 1

Due: Friday September 10 at 3:00

Submit the homework on gradescope in Canvas

To receive full credit, you must show all steps in your work. Each question is work 20 points.

Searching for solutions on the web is forbidden.

1. Let Y be a continuous random variable with density $p(y)$. Assume that $p(y) > 0$ for all y. Let m be our prediction of Y. Suppose that our prediction error is $R(m) = \mathbb{E}(|Y - m|) = \int |y - m|\, p(y)dy$. Show that $R(m)$ is minimized by choosing m to be the median ν. Recall that the median ν is defined by

$$P(Y < \nu) = P(Y > \nu) = \frac{1}{2}.$$

   Hint: Show that, if $m \neq \nu$ then $\int |y - m|\, p(y)dy \geq \int |y - \nu|\, p(y)dy$. Consider two cases: (i) $m < \nu$ and (ii) $m > \nu$. In case (i) break the integral into three pieces: $(-\infty, m]$, $[m, \nu]$, (ν, ∞). For case (ii) use $(-\infty, \nu]$, $[\nu, m]$, (m, ∞).

2. If $\widehat{\theta}$ is an estimate of a parameter θ, the bias is defined to be $\mathbb{E}[\widehat{\theta}] - \theta$. The mean squared error (MSE) is

$$\text{MSE} = \mathbb{E}[(\widehat{\theta} - \theta)^2].$$

   Show that

$$\text{MSE} = \text{bias}^2 + \text{Var}[\widehat{\theta}].$$

3. Let $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$. Suppose that σ^2 is known but μ is unkown.

   (a) Suppose we use $\overline{Y}_n = n^{-1} \sum_{i=1}^n Y_i$ to predict Y_{n+1}. Find the prediction error

$$\mathbb{E}[(Y_{n+1} - \overline{Y}_n)^2].$$

   (b) Find c such that

$$P(Y_{n+1} \in C) = 1 - \alpha$$

   where $C = [\overline{Y} - c, \overline{Y} + c]$.

4. Let

$$Y = 2X + \epsilon$$

   where $\epsilon \sim N(0, 1)$ and $X \sim N(0, 1)$. Assume that X and ϵ are independent.

   (a) Find the mean and variance of Y.

   (b) Find $\mathbb{E}[Y^2]$.

   (c) Find $\mathbb{E}[Y|X = x]$.

   (d) Find $\mathbb{E}[Y^3]$.

   (e) Find $\text{Cov}(\epsilon, \epsilon^2)$. Are ϵ and ϵ^2 independent?

5. For this question, hand in your R code and output. Make sure it is in an easy to read form. Load the wages dataset:

```
install.packages("wooldridge")  #### You only need to do this command once
library(wooldridge)
attach(wage1)
names(wage1)
str(wage1)
head(wage1)
```

(a) Plot histograms of `wage`, `educ`, `exper`, `tenure`.

(b) Find the mean and standard deviation of each of the 24 variables. You might want to use the command `apply`

(c) Draw a scatterplot of `wage` versus `exper` with `exper` on the x-axis. Add the regression line to the plot. Do you think the line is a good summary of the relationship between the two variables?

(d) What is the mean and intercept of the regression line? What is a 90 percent confidence interval for the slope?