STAT 509 / ECON 580

HOMEWORK 5

This homework will be due in quiz section on **Friday**, November 3.

## Prediction

1. Suppose that $X$ is a continuous random variable, $F(\cdot)$ is the CDF for $X$ and $m$ is a median for $X$. (Recall that $m$ is a median for $X$ iff $F(m) = 0.5$.) Let

$$I[X \leq m] = \begin{cases} 1 & \text{if } X \leq m \\ 0 & \text{otherwise.} \end{cases}$$

indicate the indicator variable that $X$ is less than or equal to $m$.

   (a) Show that

$$E[|X - m|] = E\left[(1 - 2I\left[X \leq m\right])\, X\right].$$

   *Hint: Directly use the definition of $E[h(X)]$.*

   (b) Further show that

$$E[|X - m|] = B \cdot \mathrm{Cov}(X, I\left[X \leq m\right])$$

   where $B$ is a fixed constant. Report the value of $B$.

   (c) Why does the question refer to 'a' median? If we had supposed that $X$ had support on $\mathbb{R}$, would this be necessary? Briefly explain.

2. Suppose that $X$ is a continuous random variable with support on $\mathbb{R}$. Suppose that the pdf for $X$ is symmetric around a point $t$, so that $f(t - x) = f(t + x)$ for all $x$.

   (a) Find the median of $X$. *Hint: use the fact that the pdf integrates to* 1 *and then split the integral into two pieces.*

   (b) Find the mean of $X$. *Hint: use the fact that $E[X] = t^*$ if and only if $E[X - t^*] = 0$. Again split the integral.*

3. Again let $X$ be a continuous random variable with CDF $F(\cdot)$. Let $a, b \in \mathbb{R}$ with $a < b$, and $P(a < X < b) > 0$.

   (a) Show that

   $$E[|X-b|] = E[|X-a|]+2(a-E[X|a < X < b])P(a < X < b)+(b-a)[C_1+C_2F(b)]$$

   where $C_1$ and $C_2$ are fixed constants. Find $C_1$ and $C_2$.

   *Hint: Break up the integral into three pieces: $x \leq a$; $a < x \leq b$; $b < x$.*

   *Also note that:*

   $$\begin{aligned} E[X|a < X < b]P(a < X < b) &= \int_a^b x \left( \frac{f(x)}{P(a < X < b)} \right) dx \; P(a < X < b) \\ &= \int_a^b x f(x) dx. \end{aligned}$$

   (b) Show that

   $$E[|X-b|] = E[|X-a|]+2(b-E[X|a < X < b])P(a < X < b)+(b-a)[D_1+D_2F(a)]$$

   where $D_1$ and $D_2$ are fixed constants. Find $D_1$ and $D_2$.

   *Hint: Use your answer from (a).*

   (c) State upper and lower bounds on: $E[X|a < X < b]$.

   *Hint: The answers here are simple, but these observations are useful in the next part.*

   (d) Using your answers to (a), (b) and (c), show that if $m$ is a median for $X$ then:

   $$E[|X - m|] \leq E[|X - c|]$$

   with equality if and only if $c$ is also a median for $X$.

   *Hint: consider separately the cases $F(c) < F(m)$ and $F(m) < F(c)$.*

4. Suppose that $X$ and $Y$ are continuous random variables, with support on $\mathbb{R}^2$. Suppose that two researchers, Thelma and Louise, wish to predict $Y$ from $X$ using a function of $X$.

   (a) Thelma wishes to use the function $g(X)$ that minimizes the average squared prediction $E[(Y - g(X))^2]$. What function will Thelma use? *You may justify your answer by quoting results from the Lecture.*

   (b) Louise, however, wishes to use the function $h(X)$ that minimizes the average absolute error $E[|Y - h(X)|]$. What function will Louise choose? Explain your answer. *Hint: Use the law of iterated expectations and Qu.3.*

   (c) Suppose that there is a function $r(x)$ such that the conditional density for $Y$ given $X = x$ is symmetric around $r(x)$, so that for all $x$ and $y$, $f(r(x)-y \mid x) = f(r(x)+y \mid x)$, what can we say about the functions $g(X)$ and $h(X)$ used by Thelma and Louise?

       *Hint: Use Qu.2.*

5. For the California Student Teacher Ratio and Test Score dataset. See:

   `http://www.stat.washington.edu/tsr/s509/examples/caschool.csv`

   Using `R` or a similar package find:

   (a) The best linear predictor of Test Score from Student Teacher Ratio.

   (b) The approximate conditional expectation function E[ Test Score | Student Teacher Ratio ] via binning Student Teacher Ratio.

   (c) The conditional expectation function via the loess smoother E[ Test Score | Student Teacher Ratio ].

   *Construct a scatterplot showing the data together with these three functions. Also provide the code that you used.*

   *See the example code here:*

   `http://www.stat.washington.edu/tsr/s509/examples/edwage.r`

## Iterated expectations and covariances

6. A population consists of two types, *humans* and *replicants*. The proportion of humans is $q$. The height of each type approximately follow normal distributions. Let $N(\mu_H, \sigma_H^2)$ be the distribution of lengths for humans; let $N(\mu_R, \sigma_R^2)$ be the distribution of lengths for replicants.
   (a) Find the mean height of a randomly sampled subject in this population.
   (b) Find the variance of the distribution of height for subjects in this population.

7. Let $X_1$ and $X_2$ be independent random variables, with means $\mu_1$, $\mu_2$, and variances $\sigma_1^2$ and $\sigma_2^2$ respectively. Further, let $S = (X_1 + X_2)/4$ and $T = (X_1 - X_2)/4$. Find:
   (a) $\mathrm{E}[S]$ and $\mathrm{E}[T]$;
   (b) $\mathrm{V}(S)$ and $\mathrm{V}(T)$;
   (c) $\mathrm{Cov}(S, T)$;
   (d) $\mathrm{Cov}(X_1, S)$, $\mathrm{Cov}(X_2, T)$.

## Bayesian Statistics

8. Suppose a medical test has the following characteristics:

$$\begin{aligned} Pr(\text{Test +ve} \mid \text{Patient Diseased}) &= 0.98 \\ Pr(\text{Test -ve} \mid \text{Patient Not Diseased}) &= 0.99 \end{aligned}$$

   (a) Find $Pr(\text{Test -ve} \mid \text{Patient Diseased })$ and $Pr(\text{Test +ve} \mid \text{Patient Not Diseased })$.

   Suppose that 1 in 20,000 people have this disease so

$$Pr(\text{Patient Diseased}) = 0.00005$$

   (b) Compute $Pr(\text{Test +ve})$. *Hint: Find $Pr(\textit{Test +ve, Patient Diseased})$ and $Pr(\textit{Test +ve, Patient Not Diseased})$.*
   (c) Use Bayes' rule to find $Pr(\text{Patient Diseased} \mid \text{Test +ve})$.
   (d) Give an intuitive explanation for the discrepancy between $Pr(\text{Patient Diseased} \mid \text{Test +ve})$ and $Pr(\text{Test +ve} \mid \text{Patient Diseased})$.