Stat 443 - Fall 2017

Assignment #2

(due Tuesday, October 17, at 1:00 pm)

You do not need to submit solutions to Problems 3 and 5, as they will not be marked.

1. Suppose that a future outcome that we want to predict is described by a random variable Y with the following probability mass function:

$$P(Y=-2)=p_{-2},\ P(Y=-1)=p_{-1},\ \ P(Y=0)=\frac{1}{3},\ P(Y=1)=\frac{1}{6},\ \ \text{and}\ \ P(Y=2)=\frac{1}{6}.$$

If possible, give examples of distributions of the above form (that is, you have to specify the numbers p_{-2} and p_{-1}) for which:

(i) $E(Y) < \text{median}(Y)$

(ii) $E(Y) = \text{median}(Y)$

(iii) $E(Y) > \text{median}(Y)$.

In each case briefly justify your answer. Based on your findings, which predictor (mean or median) you believe is more suitable when the distribution of Y is asymmetric and more extreme values may occur?

2. Suppose that a random variable Z is defined in the following way:

$$Z := Z_1 + Z_2,$$

where Z_1 and Z_2 are independent and $Z_1 \sim N(0, 9/10)$, $Z_2 \sim N(0, 1/10)$.

(i) - find the best prediction of Z in the MS-sense. What is the variance of the prediction error? Show your calculations.

(ii) - the same as in (i) above but now you can assume that we know that $Z_1 = 0.1$.

3. Suppose that a random variable Y follows a discrete distribution according to

$$P(Y = y_i) = p_i, \quad i = 1, 2, \ldots, K,$$

where we assume that the probabilities $p_1, \ldots, p_K$ are all different.
We want to find the best predictor of Y using the zero-one loss function L_{01} defined in class. Show that the value f that minimizes

$$E[L_{01}(Y, f)]$$

is given by

$$f = \text{Arg}\{\max_{i \in 1,\ldots,K} p_i\}.$$

4. In order to predict a random variable Y we want to use another variable X. We assume that

$$P(Y = y_i|X = x) = p_i(x), \quad i = 1, \ldots, K,$$

where for each x the probabilities $p_1(x), \ldots, p_K(x)$ are all different. We also assume that X is a discrete random variable with the following probability mass function

$$P(X = x) = m(x), \quad x \in S_X := \{1, \ldots, M\}.$$

When we use the zero-one loss function, the best predictor of Y in terms of X is defined as the function f that minimizes

$$E[L_{01}(Y, f(X))]. \tag{1}$$

Show that f that minimizes (1) is given by the Bayes classifier

$$\hat{f}(x) = \text{Arg}\{\max_{j \in 1,\ldots,K} p_j(x)\}.$$

Hint: use the law of total expectation

$$E[L_{01}(Y, f(X))] = \sum_{x \in S_X} E[L_{01}(Y, f(x))|X = x]m(x),$$

and then Problem 3.

5. Obtain the historical data from **mp3_y.txt**, which consists of the total number Y_t of a new type of MP3 player sold each month at a particular chain of stores for 72 consecutive months $t = 1, \ldots, 72$. The goal is to use the historical data $Y_1, \ldots, Y_{72}$ to predict player sales in future months.

(a) Provide a scatter plot of player sales vs. time, and use this to develop a linear model, using *only an appropriate subset* of the historical data, which would be useful for forecasting future sales. Give reasons for your choice of subset.

(b) Perform a least squares fit on your model (using either **lm** or **lsfit**) and provide a summary of the output. Provide a scatter plot of the data (restricted to your chosen subset) with the least squares (regression) line superimposed.

(c) Do *residual analysis*: plot the residual vs. time, provide a QQ-plot to check for normality, and an autocorrelation plot (**acf**) to look for correlation between the residuals. Discuss your results: do i.i.d. $N(0, \sigma^2)$ errors seem reasonable here?

(d) Discuss the overall results of your regression analysis *assuming the OLSA holds* (even if you rejected them in the previous question). Does the model appear useful in explaining the average changes in player sales over the subset of time that you are using? What is the predictive power of the model?

(e) Again assuming that OLSA holds, write down your parameter estimates with 95% confidence intervals.

(f) Based on OLS assumptions, write down the general prediction equation (i.e. the estimated mean function) and predict player sales for months 73 and 78. Write down a 95% prediction interval for your predictions for the months 73 and 78.

6. The monthly revenue of a telephone company for 11 consecutive years (beginning January 1986 and ending December 1996) is contained in the file **telephone_y.txt**.

(a) Provide a plot of the data vs. time, and develop a suitable multivariate causal model, using all the data, for forecasting the time series. You may wish to transform the data first and to include both the trend and seasonal components in your model. *Explicitly write down the model you choose, defining all parameters and variables.* We will call this Model I. (If you transform your data, include a plot of the transformed data.)

(b) Perform least squares on Model I (using either **lm** or **lsfit**). You may wish to "borrow" some of the columns of the reduced design matrix **hotel_x** we used in class to build your design matrix. Provide a copy of the output from your procedure.

(c) *Assuming* the OLS assumptions are true, discuss the overall usefulness of Model I in (i) explaining and (ii) predicting the behavior of telephone revenues (that is, discuss the F-test and the R^2-values). Discuss the usefulness of the individual variables. Are there any variables you might consider dropping from the model? (why or why not?)

(d) Perform a *residual analysis* to assess the validity of the OLS assumptions. Turn in the residual plot ϵ_t^* vs. t, the sample **acf** plot, and the sample QQ-plot. Discuss each plot individually. What is your conclusion?

(e) Another possibility for building a forecast model for this time series is to throw away some of the initial historical data. Construct a multi-variable causal model using *only* the last 5 years of data, which includes a trend component that is *quadratic* in t (why?) as well as seasonal components (it should not be necessary to transform the data). *Explicitly write down this model, defining all parameters and variables.* We will call this Model II.

(f) Carry out part (b) for Model II.

(g) Carry out part (c) for Model II.

(h) Carry out part (d) for Model II.

(i) Assuming the OLS assumptions are true for Model II, construct 95% prediction intervals for telephone revenues for March of 1997 and 1998.

(j) The data for 1997 is stored in **telephone_future.txt**. Compute the SOS of forecast errors for each of Model I and Model II (for definitions, see the first set of slides we used in class). Which model performs best?