Homework #1 PSTAT 126

1) Prove the identity:

$$\mathrm{Var}(aZ + bZ') = a^2\mathrm{Var}(Z) + b^2\mathrm{Var}(Z') + 2ab\mathrm{Cov}(Z, Z'),$$

for any random variables Z, Z' and real numbers a, b, where $\mathrm{Var}(Z) = E[(Z - E[Z])^2]$ denotes the variance, and $\mathrm{Cov}(Z, Z')$=$E[(Z - E[Z])(Z' - E[Z'])]$ denotes the covariance.

2) We say that a collection of N random variables $Z_1, \dots, Z_N$ (we will assume here that these r.v.'s are either all discrete or all continuous) is mutually independent if we have
$$f_{Z_1,\dots,Z_N}(z_1, \dots, z_N) = f_{Z_1}(z_1) \dots f_{Z_N}(z_N),$$
for all admissible choices of the $z_1, \dots, z_N$, where the $f_{Z_1}, \dots, f_{Z_N}$, and $f_{Z_1,\dots,Z_N}$ are probability density functions if $Z_1, \dots, Z_N$ are continuous r.v.'s and are probability mass functions if the $Z_1, \dots, Z_N$ are discrete r.v.'s. Moreover, these N variables $Z_1, \dots, Z_N$ are instead said to be pairwise independent if any two variables chosen from this collection of N r.v.'s are independent.
   a) Does pairwise independence imply mutual independence? Prove or disprove.
   b) Does mutual independence imply pairwise independence? Prove or disprove.

3) Can linear regression be used to model dynamics for which the regression function is not necessarily described by a straight line? Explain.

4) In the context of Simple Linear Regression, show that the residuals satisfy

$$\sum_{n=1}^{N} e_n = 0.$$

5) For any predictor variable X and response Y, explain what of interest the equation

$$\boldsymbol{E}[(f(X) - Y)^2] = \boldsymbol{E}[(f(X) - \boldsymbol{E}[Y|X])^2] + \boldsymbol{E}[(Y - \boldsymbol{E}[Y|X])^2],$$

which holds for any suitable function f, may imply concerning regression analysis.

6) Solve the least squares optimization (minimization) problem

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\alpha_0,\alpha_1)\epsilon\mathbb{R}^2} \sum_{n=1}^{N} \left(y_n - (\alpha_0 + \alpha_1 x_n)\right)^2$$

for the numerical values $\hat{\beta}_0$, $\hat{\beta}_1$ for the particular case of the three specific data points (0,1), (1,0), and (1,1) (so in this case $N = 3$), without explicitly appealing to the general solution formulas given in the course slides. Show your reasoning/computations, and give reasoning as to why the solution found really is in fact a (local) minimum. (Hint: Use calculus.)

7) In R, use the lm( ) function to solve a Simple Linear Regression model with FAMI (Familiarity with law) as the predictor variable and WRIT (Sound written rulings) as the response, using the USJudgeRatings dataset ("built-in" with R). What are the estimates generated for the intercept and slope ($\hat{\beta}_0$ and $\hat{\beta}_1$, respectively)? Plot the data graphically as well, including graphing the corresponding estimate for the mean function (regression line). In your answer, include only the numerical estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$, the R code used to generate these estimates, and also the graphical plot described in the previous sentence. Do not include the code used to generate the plot.