PSTAT 126: Homework Assignment #1 Solutions

Problem 1: Prove the identity:

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

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

Solution: By definition,

$$\text{Var}(aZ + bZ') = E[\big((aZ + bZ') - E[aZ + bZ']\big)^2].$$

Hence,

$$\begin{aligned}\text{Var}(aZ + bZ') &= E[\big((aZ - E[aZ]) + (bZ' - E[bZ'])\big)^2] \\ &= E[(aZ - E[aZ])^2] + E[(aZ' - E[aZ'])^2] \\ &+ E[2(aZ - E[aZ])(bZ' - E[bZ'])].\end{aligned}$$

Problem 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.

Solution: For (a), the answer is no; pairwise independence does not imply mutual independence. For a counterexample in the discrete random variable context, consider 3 random variables *X,Y*, and *Z*, each only taking the values 0 or 1. Suppose their 3-way joint probability mass function is described by

$$f_{X,Y,Z}(0,0,0) = \tfrac{1}{4}, f_{X,Y,Z}(1,0,1) = \tfrac{1}{4}, f_{X,Y,Z}(0,1,1) = \tfrac{1}{4}, f_{X,Y,Z}(1,1,0) = \tfrac{1}{4}$$

(noting as an aside that the notation $p_{X,Y,Z}$ is more standard than $f_{X,Y,Z}$ for the pmf of a discrete random variable). Then, it follows that

$$f_X(0) = f_Y(0) = f_Z(0) = \tfrac{1}{2}$$
$$f_X(1) = f_Y(1) = f_Z(1) = \tfrac{1}{2}.$$

But,

$$f_{X,Y}(0,0) = f_{X,Y}(0,1) = f_{X,Y}(1,0) = f_{X,Y}(1,1) = 1/4.$$

Since each joint probability is the product of the corresponding individual marginals, we have pairwise independence. Yet, it is clear that we do not have

$$f_{X,Y,Z}(x, y, z) = f_X(x)f_Y(y)f_Z(z)$$

for any choice of the variables x, y, z.

(counterexample courtesy Wikipedia)

For (b), the answer is that mutual independence does imply pairwise independence. We will first establish this for the case of three discrete random variables *X,Y,* and *Z*; the proof for *N* r.v.'s, *N* an integer with *N*>3 is very similar. So, regarding $f_{X,Y,Z}(x,y,z) = p_{X,Y,Z}(x,y,z)$ as the probability mass function (now using the more standard notation $p_{X,Y,Z}(x,y,z)$ for the probability mass function in the discrete random variable case), we have that

$$p_{X,Y}(x,y) = \sum_z p_{X,Y,Z}(x,y,z)$$

$$= \sum_z p_X(x)p_Y(y)p_Z(z) \quad \text{(by mutual independence)}$$
$$= p_X(x)p_Y(y)\sum_z p_Z(z)$$
$$= p_X(x)p_Y(y),$$

giving the result. The case for three continuous r.v.'s is analogous:

$$f_{X,Y}(x,y) = \int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)dz.$$

$$= \int_{-\infty}^{\infty} f_X(x)f_Y(y)f_Z(z)dz \quad \text{(by mutual independence)}$$

$$= f_X(x)f_Y(y)\int_{-\infty}^{\infty} f_Z(z)dz$$

$$= f_X(x)f_Y(y).$$

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

Solution: The answer is yes. As discussed in the lecture videos, consider

$$\boldsymbol{E}[Y|X] = c_0 + c_1\varphi_1 + \cdots + c_M\varphi_M,$$

where the φ_i are fixed basis functions ("regressors") and the c_i are scalar coefficient parameters to be determined by the regression procedure, where we are here referring to basis functions in the sense of linear algebra (vector space basis functions, that is). We say that $\boldsymbol{E}[Y|X]$ is a linear combination of the φ_i.

If the functions are, say, simply of the form $\varphi_1(x) = x, \varphi_i(x) = 0$ for all $i > 1$, then we have the case of Simple Linear Regression and the regression function $\boldsymbol{E}[Y|X]$ will indeed by a straight line. However, if we choose instead, for example, $\varphi_i(x) = x^i$ for all i, the mean function will then be described by the graph of a (smooth) curve. Similarly, taking $\varphi_i(x) = \sin(x^i)$ results of course in a curved regression line as well. Still, all of these are examples of ``linear'' regression, because $\boldsymbol{E}[Y|X]$ will still be a linear combination of the regressor functions.

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

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

Solution: As discussed on slide 14 of the course slides (also see HW #1, Problem 6 and its solution below), $(\hat{\beta}_0, \hat{\beta}_1)$ must be a critical point of the function *F* on slide 14, that is, both partial derivatives of *F* with respect to both variables α_0 and α_1 must be 0 when evaluated at $(\hat{\beta}_0, \hat{\beta}_1)$. Hence,

$$\begin{aligned} 0 &= \partial_{\alpha_0}\left(\sum_{n=1}^{N} (\alpha_0 + \alpha_1 x_n - y_n)^2\right)\Big|_{(\alpha_0=\hat{\beta}_0, \alpha_1=\hat{\beta}_1)} \\ &= 2\sum_{n=1}^{N}\left(\hat{\beta}_0 + \hat{\beta}_1 x_n - y_n\right) \\ &= 2\sum_{n=1}^{N}(\hat{y}_n - y_n), \end{aligned}$$

which, by definition (see slide 19), implies that the residuals sum to 0 as was to be shown.

Problem 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.

Solution: As discussed on slide 7 of the course slides, the equation in the problem says that that, for any function f, the expectation (average) of the square of the difference between $f(X)$ and Y is equal to the expectation of the square of the difference between f and the mean function $\boldsymbol{E}[Y|X]$, plus a nonnegative term that does not depend on f.

- Since $(f(X) - Y)^2 \geq 0$ for any function f we can minimize the magnitude of the error of approximating f by Y on the left-hand side of the equation by in fact taking $f(X) = \boldsymbol{E}[Y|X]$, since this makes the first term on the right 0.
- This means that the function of *X*=*x* that approximates the behavior of the response *Y* with the smallest error on average is in fact the mean $\boldsymbol{E}[Y|X]$ function itself.
- So, it is the mean function which gives us the "best" representation of the functional relationship between *X* and *Y* in the sense described.

Hence, it is the mean function $\boldsymbol{E}[Y|X]$ that we would like to use regression methods and algorithms to determine or at least closely approximate in order to identify and understand any functional relationship between *X* and *Y.*

Problem 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.)

Solution: With reference to (11) on slide 14 of the course slides, consider, in order to solve the least-squares regression problem of Problem #6 above, that

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

for the 3 specific data points specified in the problem statement and where we employ the notation

$$F(\alpha_0, \alpha_1) = (\alpha_0 - 1)^2 + (\alpha_0 + \alpha_1)^2 + (\alpha_0 + \alpha_1 - 1)^2.$$

To find the minimum point of the function *F*, we need to find its critical point(s). So, differentiate (take partial derivatives) with respect to both α_0 and α_1:

$$0 = \partial_{\alpha_0} F(\alpha_0, \alpha_1)|_{(\alpha_0,\alpha_1)} = 2(3\alpha_0 + 2\alpha_1 - 2)$$
$$0 = \partial_{\alpha_1} F(\alpha_0, \alpha_1)|_{(\alpha_0,\alpha_1)} = 2(2\alpha_0 + 2\alpha_1 - 1).$$

Solving these two equations for α_0 and α_1, we get $\alpha_0 = 1, \alpha_1 = -\frac{1}{2}$. So, $\hat{\beta}_0 = 1, \hat{\beta}_1 = -\frac{1}{2}$.

Now to show that $(\hat{\beta}_0, \hat{\beta}_1) = \left(1, -\frac{1}{2}\right)$ is in fact a (local) minimum, we can use the well-known Second (Partial) Derivative Test. As a reference for this test, see for example either of the websites

https://mathworld.wolfram.com/SecondDerivativeTest.html
https://en.wikipedia.org/wiki/Second_partial_derivative_test

or, presumably, your favorite introductory (multivariable) calculus textbook.

Continuing, note that

$$\partial^2_{\alpha_0} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)} = 6,\ \partial^2_{\alpha_1} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)} = 4,\ \partial_{\alpha_0}\partial_{\alpha_1} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)} = 4.$$

But, by the Second Derivative Test, if $\partial^2_{\alpha_0} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)} > 0$ and

$$\partial^2_{\alpha_0} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)} \partial^2_{\alpha_1} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)} - \left(\partial_{\alpha_0}\partial_{\alpha_1} F(\alpha_0, \alpha_1)|_{(\hat{\beta}_0,\hat{\beta}_1)}\right)^2 > 0,$$

then the critical point $(\hat{\beta}_0, \hat{\beta}_1)$ will be a (local) minimum. Hence, we conclude that $(\hat{\beta}_0, \hat{\beta}_1) = \left(1, -\frac{1}{2}\right)$ must be a local minimum.

(For the solution to Problem #7 see the R Markdown on the following two pages.)

Solution to Problem #7, Homework Assignment #1

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

Problem 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.

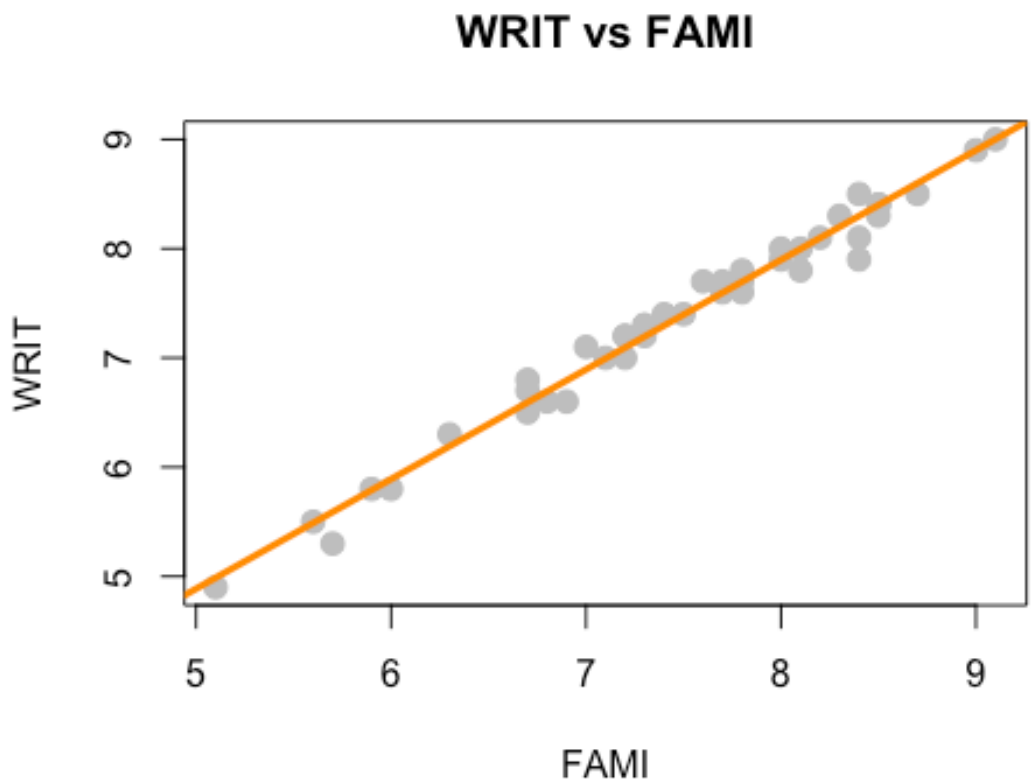
Solution: The R code solves the Simple Linear Regression model as in the problem statement and produces the ouput that immediately follows:

```
    judge_model = lm(WRIT ~ FAMI, data = USJudgeRatings)
    judge_model

##
## Call:
## lm(formula = WRIT ~ FAMI, data = USJudgeRatings)
##
## Coefficients:
## (Intercept)         FAMI
##     -0.1299       1.0034
```

The following R code can then be used to generate the corresponding plot on the page following:

```
plot(WRIT ~ FAMI, data = USJudgeRatings,
     xlab = "FAMI",
     ylab = "WRIT",
     main = "WRIT vs FAMI",
     pch  = 20,
     cex  = 2,
     col  = "grey")
abline(judge_model, lwd = 3, col = "darkorange")
```


WRIT vs FAMI
WRIT
FAMI
5
6
7
8
9
5
6
7
8
9