Boston University Questrom School of Business
MF 793 – Fall 2018

Eric Jacquier

Homework 2 Solutions

Due Monday October 8th in class

Problems turned in after the beginning of class have a notch deduction. Problems turned in after class get a zero.

- Do the Problem Set in groups of two
- Answer on a paper copy in class, no electronic submission.
- To get a check, you need to answer all the questions including the discussion questions.
- The report must be very clear. Follow exactly the format outlined in the empirical analysis with Tables and Figures numbered as indicated.

Problem 1

Finish the proof that the Median minimizes the Mean Absolute Error of prediction. Be sure to use Leibniz rule properly. For credit the proof must be done by hand in the space below:

$$MAE = \int_{-\infty}^{+\infty} |\theta^* - \theta| p(\theta)\, d\theta \quad \text{where } \theta^* \text{ is a location estimator for } \tilde{\theta}$$

$$= \int_{-\infty}^{\theta^*} (\theta^* - \theta) p(\theta)\, d\theta + \int_{\theta^*}^{\infty} (\theta - \theta^*) p(\theta)\, d\theta$$

$$\frac{\partial MAE}{\partial \theta^*} = \int_{-\infty}^{\theta^*} p(\theta)\, d\theta + [\theta^* - \theta]_{\theta^*} + \int_{\theta^*}^{\infty} -p(\theta)\, d\theta + [\theta - \theta^*]_{\theta^*}$$

$$= \int_{-\infty}^{\theta^*} p(\theta)\, d\theta - \int_{\theta^*}^{\infty} p(\theta)\, d\theta + \emptyset + \emptyset \quad [1]$$

$[1] = 0$ iff $\theta^* = \text{Median}(\theta)$

Problem 2: Simulated and theoretical results

The marginal of X is: p(X) ~ Uniform (-2,2). The conditional of p(Y|X) is p(Y)~ Uniform(X,2).

a) *Simulate 10,000 draws of X, and then Y|X for each draw of X. Then use the density and persp commands to plot.*
   The marginal univariate density of y: p(y) density command – Figure 1
   The bivariate density of (X,Y): p(x,y) persp command – Figure 2.

b) *From your simulated data, what is your estimate of E(Y) ?*

c) *Write the theoretical conditional density p(Y|X), write the theoretical joint density p(X,Y).*

d) *Compute $E_Y(Y|X)$. Then use the iterated expectation rule to compute E(Y) as $E_X(E_Y(Y|X))$.*

e) *Compute the marginal density p(Y) by integration. Then use p(Y) to compute E(Y). Hope you find the same result as in d)!*

a) p(x,y) = p(y|x) p(x). We simulate a draw from p(x,y) by drawing from p(x) and then from p(y|x).

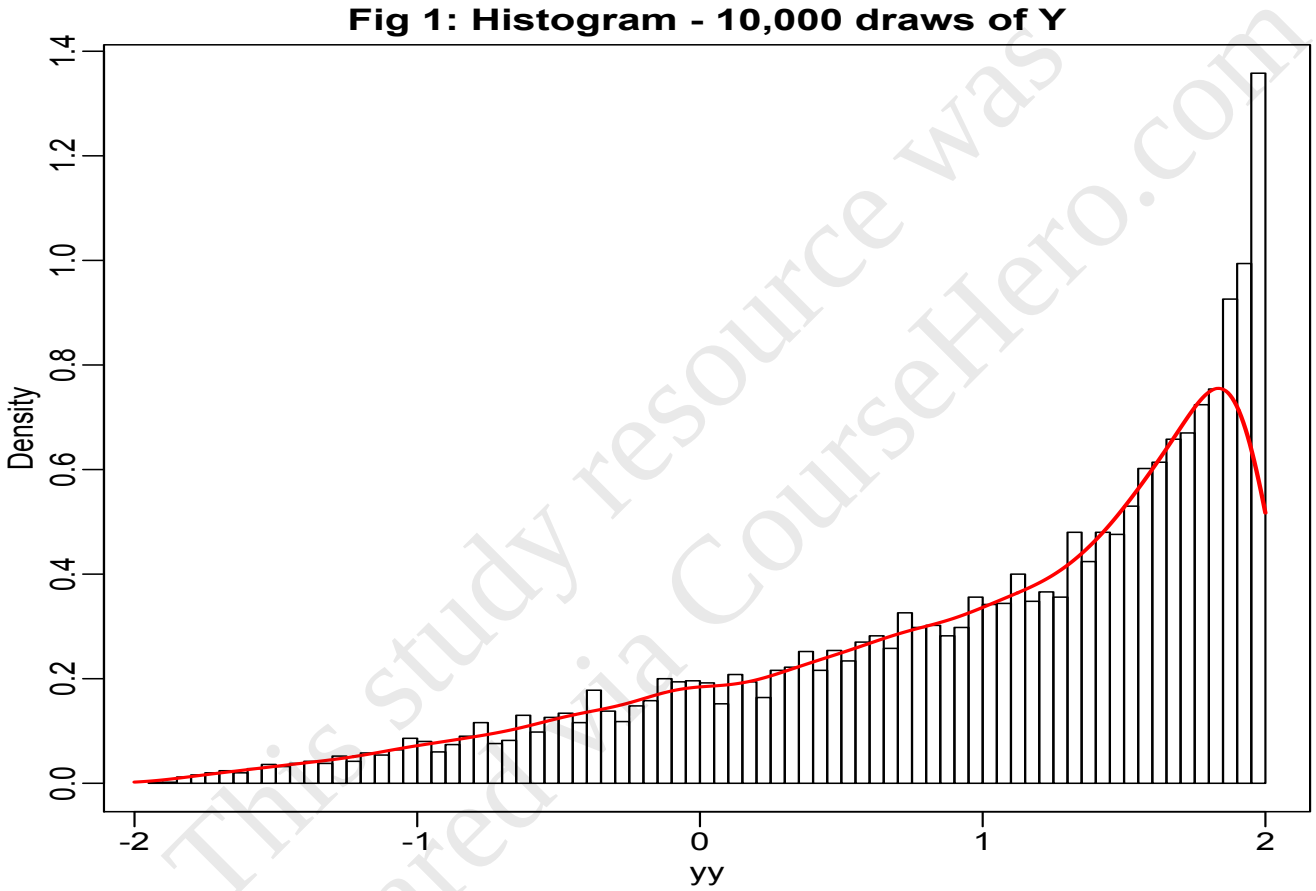
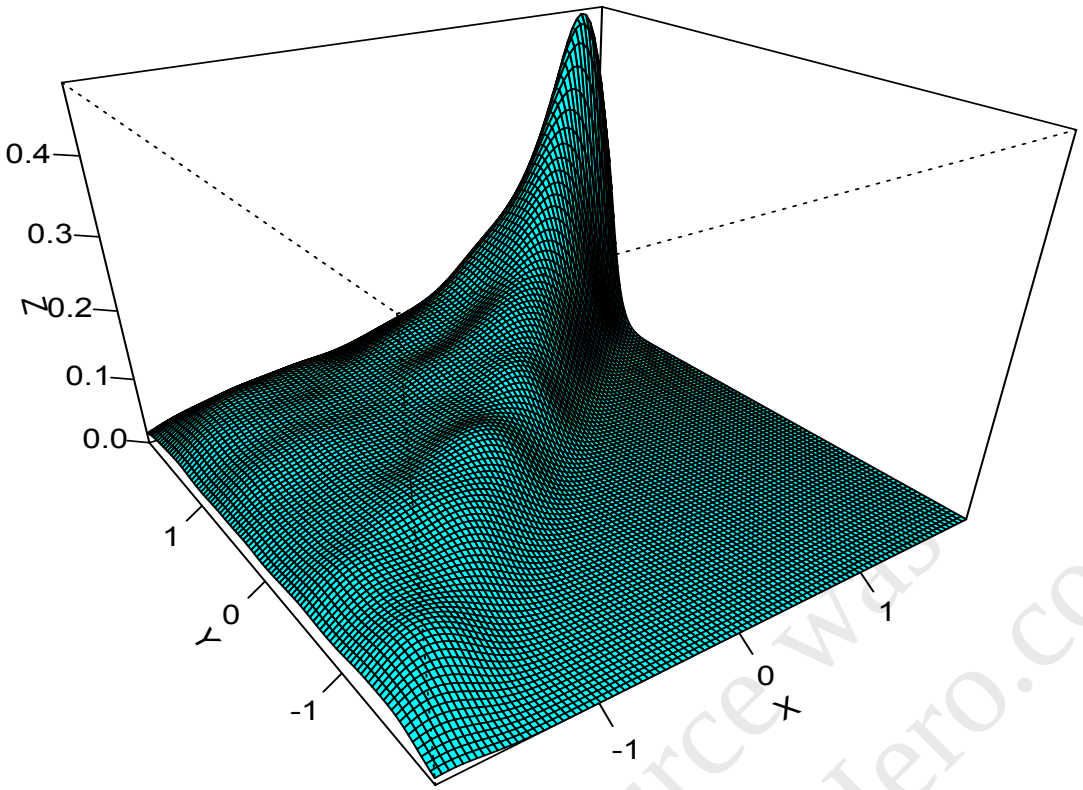


Fig 2: Estimate of the joint density of (x,y) - 10,000 draws



b) We have 10,0000 draws from y. Let's use the sample mean as an estimate of the true mean.

$\hat{\mu} = \mathbf{0.998 = 1}$

c)

$p(Y|X) = I_{Y>X} / (2-x)$, where I is the indicator function, I=1 if y>x, and y∈[x,2]
$p(X,Y)$ = p(x) p(y|x) = **$0.25 / (2-x)$**

d) **$E_Y(Y|X) = (2+X)/2$**. Since it is uniform on [x,2]

I hope nobody wasted their time computing $\int_X^2 \frac{Y}{2-X} dY$ ☺

By iterated expectation: **$E(Y|X)$** = $E_X[(X+2)/2]$ = 0.5 [$\frac{-2+2}{2} + \frac{2+2}{2}$] **= 1** **That was easy!**

e) $\boldsymbol{p(y)} = \int_y^1 p(x,y)\,dx = \int_{-2}^{y} \frac{1}{4(2-x)} dx = [-Log(2-x)]_{-2}^{Y} = \mathbf{\frac{1}{4}(\log(4) - \log(2-y))}$
Careful to put the correct bound on x to reflect the indicator function: -2< x < y
Indeed, there is no mode at zero, instead p(y) tends to infinity. Is it a problem?

$$\boldsymbol{E(y)} \quad = \int_{-2}^{2} y\, \frac{1}{4}(\log(4) - \log(2-y))\, dy = \frac{\mathbf{1}}{\mathbf{4}} \int_{-2}^{2} y\, (\mathbf{log(4) - log\,(2 - y)})dy$$

= etc.. after some not-so exciting high school calculus … **= 1**

That was not very pleasant. This is where iterated expectation is your friend !

Problem 3: Multivariate change of variable

a) U~N(0,1) and V~N(0,1) are independent normals. Write the joint density of U and V, p(U,V).

$$\mathrm{p}(U,V) = p(U)p(V) = \frac{1}{2\pi} e^{-\frac{u^2+v^2}{2}}, \quad \text{since U and V are independent} \qquad [1]$$

b) $X = a + \sigma_1 U$, *and* $Y = b + \sqrt{1-\rho^2}\,\sigma_2 V + \rho\,\sigma_2 U$. **A classic transformation you need to know**
Compute μ_X, μ_Y, σ_X, σ_Y, *and* $\rho_{X,Y}$,

- $\mu_X = a$ $\quad \mu_Y = b$ $\quad \sigma_x = \sigma_1$
- $\sigma_Y^2 = \rho^2\sigma_2^2 + (1-\rho^2)\sigma_2^2 + 0 = \sigma_2^2$ $\quad \sigma_Y = \sigma_2$
- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(\sigma_1 U\ ,\ \rho\sigma_2 U + \sqrt{1-\rho^2}\sigma_2 V) = \rho\,\sigma_1\,\sigma_2$ $\quad \rho_{XY} = \rho$

b) Write the inverse transformation: $(U,V) = g^{-1}(X,Y)$

$$U = \quad (X-a)/\sigma_1 \quad + 0 * Y$$
$$V = \quad \frac{-\rho}{\sqrt{1-\rho^2}\sigma_1}(X-a) + \frac{1}{\sqrt{1-\rho^2}\sigma_2}(Y-b)$$

c) Use the change of variable method for a bivariate density to write the joint density p(X,Y) of two correlated zero-mean normals.

Hint: Write the 2x2 matrix of derivatives of the inverse transformation $(U,V) = g^{-1}(X,Y)$. *Compute the absolute value of its determinant (the Jacobian). Then write*

$$p_{XY}(X,Y) = p_{UV}(g^{-1}(X,Y))\ \ |dg^{-1}/d(X,Y)|$$

(U,V) is linear in (X,Y). The 2x2 matrix of derivatives is:

$$\begin{matrix} 1/\sigma_1 & 0 \\ \frac{-\rho}{\sqrt{1-\rho^2}\sigma_1} & \frac{1}{\sqrt{1-\rho^2}\sigma_2} \end{matrix}$$

The determinant is $\mathbf{1 / \sigma_1\sigma_2\sqrt{1-\rho^2}}$

$$P(X,Y) = \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}\,\sigma_1\sigma_2}\,e^{-\frac{\left(\frac{X-a}{\sigma_1}\right)^2+\left(\frac{Y-b}{\sqrt{1-\rho^2}\,\sigma_2}\right)^2+\left(\frac{-\rho(X-a)}{\sqrt{1-\rho^2}\,\sigma_1}\right)^2-\frac{2\rho(X-a)(Y-b)}{(1-\rho^2)\sigma_1\sigma_2}}{2}}$$

$$= \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}\,\sigma_1\sigma_2}\,e^{-\frac{(1-\rho^2)(\frac{X-a}{\sigma_1})^2+\left(\frac{Y-b}{\sigma_2}\right)^2+\left(\frac{-\rho(X-a)}{\sigma_1}\right)^2-\frac{2\rho(X-a)(Y-b)}{\sigma_1\sigma_2}}{2(1-\rho^2)}}$$

$$= \frac{1}{2\pi}\frac{1}{\sqrt{1-\rho^2}\,\sigma_1\sigma_2}\,e^{-\frac{(\frac{X-a}{\sigma_1})^2+\left(\frac{Y-b}{\sigma_2}\right)^2-\frac{2\rho(X-a)(Y-b)}{\sigma_1\sigma_2}}{2(1-\rho^2)}}$$

As a further exercise, make sure you can prove that it can be rewritten as:

$p(X,Y) = \frac{1}{2\pi}\frac{1}{|V|^{1/2}}e^{-\frac{(X-a,Y-b)V^{-1}\binom{X-a}{Y-b}}{2}}$, where V is the covariance matrix of X and Y

Problem 4: The dangers of data mining

The head of your portfolio group decided to start accumulating information on fund performance analysis. Last year she told Bonzo to analyze five randomly chosen mutual funds every month and report on whether and by how much they statistically beat the market over the past 5 years. So, every month Bonzo has collected the last 60 months of returns of 100 largefunds i = 1, 100 and computed the t-statistics for H_0: $\mu_i = \mu_M$.

Bonzo did not take data analytics during his Finance MSc and he seems a bit confused sometimes, you check whether he is "doing it right". He tells you that "To test the null that a fund has the same return as the market at the 5% significance level, I use 2.00 as the cutoff for my absolute t-statistic with 59 degrees of freedom.", "Boss wants five random funds analyzed every months, so I pick 100 new random large funds and I write a report on 5 of them." Sounds about right but you smell a rat: Every month he computes all the 100 t-statistics and writes a report on the five funds with the highest t-statistics. This has been going a year. Bonzo can do Feynman-Kac in his sleep but he does not get it. Ah these quants who never learnt statistics, they are a danger to the profession! You remember the first article you had to read in your first Data Analytics course at BU and you immediately understands what's wrong.

"Bonzo, you are not testing at the 5% level, you are data-mining, your are p-hacking! All this evidence of performance in your reports is vastly exagerated. You need to correct all these reports."

You conduct a small – and obviously simplified, sampling experiment to illustrate his problem.

To simulate the Null hypothesis, you simulate a sample of 5 years of monthly returns for 100 funds normal returns, all with monthly mean equal to 0.01. For the monthly variance of these random funds, you use the average fund variance from your fund returns data (the average of the 1584 variances). To keep it simple, you simulate uncorrelated returns across funds and time. You then compute the 100 t-statistics. You then save the highest 5 t-statistics, and also save 5 randomly chosen t-statistics. You now have 10 t-statistics. You do this again, … 10,000 times

You have 10,000 simulations of the 10 t-statistics under the null of zero fund performance. To help Bonzo click, fill Table 1 below, with 1) the average of the t-stat, 2) the fraction of the time (with 3 decimals, like 0.025) the

*null is rejected when it is true if using the 2.00 cutoff value – the so-called **type I error** also known as the **size of the test**. Fill the third row 3) with the 97.5% quantiles of the ten t-statistics.*

You can have variations of Table 1, using 5% or other significance level, using one-sided or two-sided tests. You will still learn what happens when one data mines without corrections. You just need to be consistent between the top 3 and bottom 3 rows

Table 1: Simulated Type 1 error – probability of rejection under the H_0

	Highest five (of 100) t-statistics					
	t_1	t_2	t_2	t_4	t_5	
1) Mean	2.59	2.20	1.99	1.84	1.71	
2) Pr(reject H_0|H_0 true)%	92	71	46	24	11	
3) 97.5% of t_i	3.7	2.9	2.6	2.4	2.19	
	Randomly chosen five t-statistics					
	t_1	t_2	t_2	t_4	t_5	Theory
1) Mean	0.015	0.016	-0.011	0.021	-0.013	0
2) Pr(reject H_0|H_0 true)%	2.60	2.75	2.71	2.55	2.45	2.75
3) 97.5% of t_i	2.01	2.04	2.03	2.01	1.99	2.00

In Figure 3 plot the histogram of the fifth highest t-statistic. Make sure to use a large enough nclass, and put a vertical bar at 2.00 and a vertical bar at the 97.5% quantile of the t-statistic.

- *Conclude for Bonzo. What should the mean of the Student-t be under the null hypothesis? What fraction of the time should we expect reject the null when it is correct? Which statistics are distributed following the theoretical distributions, which are not? What is wrong with reporting on the highest five t-stat funds with the usual 97.5% size number of 2.00?*

We do expect the mean of a t-statistic for a randomly chosen fund to be 0 under the null hypothesis. We do expect the 97.5th percentile of a random t-statistic(ν =59) to be 2.00. All this is of course what is confirmed in the last three rows.

But Bonzo's data mining reports different distributions under H_0, namely the largest (and 2nd largest and etc..) of 100 randomly chosen t-statistics(ν=59). All these distributions are of course shifted right, they are not centered at zero and their 97.5th quantile is larger than 2.00.

- *Using the results in Table 1, show Bonzo how he could slightly modify his reports to eliminate his data-mining bias, while still reporting on the top 5 funds every month.*

He could use the 97.5th percentile of the **ordered** t-statistic as threshold value to reject the null. This would take into account the fact that he is reporting on the highest five t-statistics out of 100. Some refinement would be surely needed such as taking incorporating fund cross-correlations in the simulation. That will be for another day.

Figure 3: Distribution of Data-Mined-Student-t under H0