DS-GA 1002 - Homework 5

Eric Niblock

October 3nd, 2020

1. (Short questions) Justify all your answers mathematically.

(a) For any random variable $\tilde{a}$, can $E^2(\tilde{a})$ be smaller than $E(\tilde{a}^2)$?

It is true that there exists a random variable $\tilde{a}$ such that $E^2(\tilde{a}) < E(\tilde{a}^2)$. Let,

$$\tilde{a} \in A = \{0, 0.5, 1\} \tag{1}$$

Each with equal probability. Then, it follows that,

$$E(\tilde{a}^2) = \sum_{a \in A} a^2 p_{\tilde{a}}(a) = \frac{1}{3}(0 + 0.25 + 1) = \frac{5}{12} \tag{2}$$

$$E^2(\tilde{a}) = \left(\sum_{a \in A} a p_{\tilde{a}}(a) \right)^2 = \left(\frac{1}{2} \right)^2 = \frac{1}{4} \tag{3}$$

So, this shows that there does exist a random variable $\tilde{a}$ such that $E^2(\tilde{a}) < E(\tilde{a}^2)$.

(b) If the median of $\tilde{a}$ equals m, what is the median of $\tilde{a} + b$?

If m is the median then we have the cdf such that,

$$F_{\tilde{a}}(m) = P(\tilde{a} \leq m) = \frac{1}{2} \tag{4}$$

By the definition of median. Then we know that,

$$\frac{1}{2} = P(\tilde{a} \leq m) = P(\tilde{a} + b \leq m + b) = F_{\tilde{a}+b}(m + b) \tag{5}$$

Which, expressed in words, means that the distribution $\tilde{a}+b$ must have a median of $m + b$.

(c) If $\tilde{a}$ and $\tilde{b}$ have the same distribution and are independent, is it true that $E(\tilde{a}\tilde{b}) = E^2(\tilde{a})$?

Yes, this is true. Since $\tilde{a}$ and $\tilde{b}$ have the same distribution and are independent, we can simply write $E(\tilde{a}\tilde{b}) = E(\tilde{a})E(\tilde{b})$. Furthermore, since $\tilde{a} = \tilde{b}$, then $E(\tilde{a}) = E(\tilde{b})$, and,

$$E(\tilde{a})E(\tilde{b}) = E(\tilde{a})E(\tilde{a}) = E^2(\tilde{a}) \tag{6}$$

(d) A teacher of a class of n children asks their parents to leave a present under the Christmas tree in the classroom. The day after, each child picks a present at random. What is the expected number of children that end up getting the present bought by their own parents? (Hint: Define a random variable I_i that is equal to one when kid i gets the present bought by their own parents, and to zero otherwise.)

As suggested, we define I_i as follows,

$$I_i = \begin{cases} 1, & \textit{child } i \textit{ gets their own parents' gift} \\ 0, & \textit{otherwise} \end{cases} \tag{7}$$

It is clear that,

$$E(I_i) = P(I_i) = \frac{1}{n} \tag{8}$$

Then, by the linearity of expectations, we have,

$$\sum_{i=1}^{n} E[I_i] = n\frac{1}{n} = 1 \tag{9}$$

So, for n children, we would expect 1 child to receive their own gift.

2. (Pasta and rice) You are hired by the management of a restaurant to model its stock probabilistically. You talk to the cook and she says:

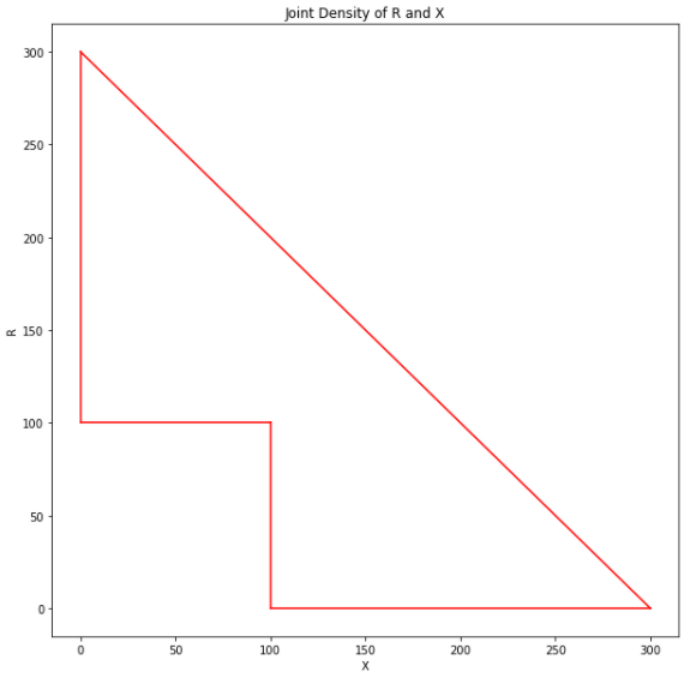
We cook pasta and rice. We always make sure that we have at least 100 lb of pasta or 100 lb of rice (if there is at least 100 lb of pasta, for example, we could have no rice at all);the logic being that we have daily specials and we want to be able to feed a lot of people with the same dish. However we never have more than 300 lb of rice or of pasta because we have no space to store it (we are able to store 300 lb of rice and 300 lb of pasta at the same time).

You decide to model the quantity of pasta as a random variable $\tilde{x}$ and the quantity of rice as a random variable $\tilde{r}$. As you have no information beyond what you have heard, you assume that their joint pdf is constant (within the restrictions that you deduce from talking to the cook).

(a) Draw the joint pdf of $\tilde{x}$ and $\tilde{r}$.

We have the following conditional probabilities given the information,

$$\begin{aligned} f_{\tilde{x}|\tilde{r}}(x|r<100) &= \mathcal{U}(100, 300-r) \\ f_{\tilde{x}|\tilde{r}}(x|100\leq r\leq 300) &= \mathcal{U}(0, 300-r) \\ f_{\tilde{r}|\tilde{x}}(r|x<100) &= \mathcal{U}(100, 300-x) \\ f_{\tilde{r}|\tilde{x}}(r|100\leq x\leq 300) &= \mathcal{U}(0, 300-x) \end{aligned} \tag{10}$$



Where inside the figure, $f_{\tilde{x},\tilde{r}}(x,r) = {}^{1}/_{35000}$, and outside the figure $f_{\tilde{x},\tilde{r}}(x,r) = 0$.

(b) Are $\tilde{x}$ and $\tilde{r}$ uncorrelated? Justify your answer.

We calculate the covariance of the random variables $\tilde{x}$ and $\tilde{r}$ as follows,

$$Cov(\tilde{x},\tilde{r}) = E(\tilde{x}\tilde{r}) - E(\tilde{x})E(\tilde{r}) \tag{11}$$

So, we have that,

$$E(\tilde{x}\tilde{r}) = \int_0^{100}\int_{100}^{300-x} xr f_{\tilde{x},\tilde{r}}(x,r)\,drdx + \int_{100}^{300}\int_0^{300-x} xr f_{\tilde{x},\tilde{r}}(x,r)\,drdx \tag{12}$$

$$E(\tilde{r}) = \int_0^{100}\int_{100}^{300-x} r f_{\tilde{x},\tilde{r}}(x,r)\,drdx + \int_{100}^{300}\int_0^{300-x} r f_{\tilde{x},\tilde{r}}(x,r)\,drdx \tag{13}$$

$$E(\tilde{x}) = \int_0^{100}\int_{100}^{300-x} x f_{\tilde{x},\tilde{r}}(x,r)\,drdx + \int_{100}^{300}\int_0^{300-x} x f_{\tilde{x},\tilde{r}}(x,r)\,drdx \tag{14}$$

The result of these calculations leads to a non-zero covariance (of value -4132.65), which implies a non-zero correlation (of value -0.7839). As such, the variables are correlated.

(c) Are $\tilde{x}$ and $\tilde{r}$ independent? Justify your answer.

The random variables $\tilde{x}$ and $\tilde{r}$ cannot be independent, because if they were, they would have a covariance of zero.

Alternative, it is clear that $\tilde{x}$ and $\tilde{r}$ are not independent if we examine the case of

$$f_{\tilde{x},\tilde{r}}(50,50) = 0 \tag{15}$$

However, it is clear that,

$$f_{\tilde{x}}(50) > 0 \qquad ; \qquad f_{\tilde{r}}(50) > 0 \tag{16}$$

Therefore, there exists at least some values of $\tilde{x}$ and $\tilde{r}$ such that,

$$f_{\tilde{x},\tilde{r}}(x,r) \neq f_{\tilde{x}}(x) f_{\tilde{r}}(r) \tag{17}$$

3. (Law of conditional variance) In this problem we define the conditional variance in a similar way to the conditional expectation.

(a) What is the object $Var(\tilde{b}|\tilde{a} = a)$ (i.e. is it a number, a random variable or a function)? What does it represent?

The object $Var(\tilde{b}|\tilde{a} = a)$ is a number, and not a random variable or function. It represents the variance of random variable $\tilde{b}$ given that we know $\tilde{a} = a$. In other words, it is a measure of the spread of $\tilde{b}$ (the degree to which $\tilde{b}$ differs from $E(\tilde{b})$) when we know that $\tilde{a} = a$.

(b) Setting $h(a) = Var(\tilde{b}|\tilde{a} = a)$ we define the conditional variance as $Var(\tilde{b}|\tilde{a}) = h(\tilde{a})$. What is this object?

It is clear that $h(\tilde{a}) = Var(\tilde{b}|\tilde{a})$ is a random variable, because any function of a random variable is also a random variable.

(c) Prove the law of conditional variance:

$$Var(\tilde{b}) = E\left(Var(\tilde{b}|\tilde{a})\right) + Var\left(E(\tilde{b}|\tilde{a})\right) \tag{18}$$

and describe it in words.

First, it is productive to introduce the definition of variance, and therefore, conditional variance, given by,

$$Var(\tilde{b}) = E(\tilde{b}^2) - E(\tilde{b})^2 \tag{19}$$

$$Var(\tilde{b}|\tilde{a}) = E(\tilde{b}^2|\tilde{a}) - [E(\tilde{b}|\tilde{a})]^2 \tag{20}$$

Then, we proceed by taking expectations of both sides,

$$E[Var(\tilde{b}|\tilde{a})] = E[E(\tilde{b}^2|\tilde{a})] - E\left([E(\tilde{b}|\tilde{a})]^2\right) \tag{21}$$

The law of iterated expectations informs us that,

$$E(\tilde{b}^2) = E[E(\tilde{b}^2|\tilde{a})] \tag{22}$$

So we then yield the following,

$$\begin{aligned} E[Var(\tilde{b}|\tilde{a})] &= E(\tilde{b}^2) - E\left([E(\tilde{b}|\tilde{a})]^2\right) \\ &= E(\tilde{b}^2) - E\left([E(\tilde{b}|\tilde{a})]^2\right) + [E(\tilde{b})]^2 - [E(\tilde{b})]^2 \\ &= \left(E(\tilde{b}^2) - [E(\tilde{b})]^2\right) - \left(E\left([E(\tilde{b}|\tilde{a})]^2\right) - [E(\tilde{b})]^2\right) \\ &= Var(\tilde{b}) - \left(E\left([E(\tilde{b}|\tilde{a})]^2\right) - [E[E(\tilde{b}|\tilde{a})]]^2\right) \\ &= Var(\tilde{b}) - Var[E(\tilde{b}|\tilde{a})] \end{aligned} \tag{23}$$

Rearranging the terms yield the desired result:

$$Var(\tilde{b}) = E\left(Var(\tilde{b}|\tilde{a})\right) + Var\left(E(\tilde{b}|\tilde{a})\right) \tag{24}$$

(d) We model the time at which a runner gets injured (in hours) during a marathon as an exponential random variable with parameter equal to 1 if the runner is under 30 years old and 2 if she is over 30. What is the mean and the standard deviation of the time at which a runner gets injured if 20% of the runners are over 30?

We define a random variable $\tilde{x}$ such that,

$$\tilde{x} = \begin{cases} 1, & \textit{if the runner is under 30 years old} \\ 2, & \textit{if the runner is over 30 years old} \end{cases} \tag{25}$$

Then we have, $p_{\tilde{x}}(1) = 0.8$ and $p_{\tilde{x}}(2) = 0.2$. Furthermore, if $\tilde{y}$ is an exponential random variable with parameter λ, we have that,

$$E(\tilde{y}|\tilde{x}) = \frac{1}{\lambda} = \frac{1}{\tilde{x}} \tag{26}$$

Since, in this case, the parameter of the distribution is $\tilde{x}$. Furthermore,

$$Var(\tilde{y}|\tilde{x}) = \frac{1}{\lambda^2} = \frac{1}{\tilde{x}^2} \tag{27}$$

So, we have then,

$$E(\tilde{y}) = E\left(E(\tilde{y}|\tilde{x})\right) = 0.8(1) + 0.2(0.5) = 0.9 \tag{28}$$

This is the mean time in hours that the runner gets injured. As for the standard deviation, we have,

$$\begin{aligned} Var(\tilde{y}) &= E\left(Var(\tilde{y}|\tilde{x})\right) + Var\left(E(\tilde{y}|\tilde{x})\right) \\ &= E\left(Var(\tilde{y}|\tilde{x})\right) + \left(E[E(\tilde{y}|\tilde{x})^2] - E(\tilde{y})^2\right) \\ &= (0.8)(1) + (0.2)(\frac{1}{4}) + ((0.8)(1) + (0.2)(\frac{1}{4}) - (0.9)^2) \\ &= 0.89 \end{aligned} \tag{29}$$

Then, we have,

$$\sigma_{\tilde{y}} = \sqrt{Var(\tilde{y})} = 0.9434 \tag{30}$$

This is the standard deviation in the time in hours that the runner gets injured.

4. **(Water salinity and temperature) A quick Google search will tell you that the salinity of water, which is the salt content in water, increases with temperature. This is because water expands at larger temperature and can fit in more molecules, including salt, increasing the salinity (according to Sciencing). In this question, we will use oceanographic data to understand the relationship between salinity and temperature. We will perform our analysis on a cleaned and subsampled version of the data, *bottle.csv*. Please refer to the Kaggle website for any details about the data.**

   (a) **Find the best linear MMSE estimator of salinity with temperature. Plot the line and the scatter plot of data on the same graph. According to the relationship you uncovered here, does water salinity increase with temperature?**

   The plot of the best linear MMSE estimator of salinity was calculated using the following formulas taking $\tilde{s}$ and $\tilde{t}$ to be the random variables associated with the salinity and temperature, respectively,

$$\rho_{\tilde{s},\tilde{t}} = \frac{Cov(D)}{\sigma_{\tilde{s}}\sigma_{\tilde{t}}} = \frac{\sum_{i=1}^{n}(s_i - \mu_{\tilde{s}})((t_i - \mu_{\tilde{t}})}{n\sigma_{\tilde{s}}\sigma_{\tilde{t}}} \tag{31}$$

$$\alpha = \frac{\rho_{\tilde{s},\tilde{t}}\sigma_{\tilde{s}}}{\sigma_{\tilde{t}}} \tag{32}$$

$$\beta = \mu_{\tilde{s}} - \frac{\rho_{\tilde{s},\tilde{t}}\sigma_{\tilde{s}}\mu_{\tilde{t}}}{\sigma_{\tilde{t}}} \tag{33}$$

   And so, the final equation of the line looks like,

$$\hat{s} = \alpha\tilde{t} + \beta \tag{34}$$

   The actual equation of the red line plotted below is therefore,

$$\hat{s} = -0.0587\,\tilde{t} + 34.473 \tag{35}$$

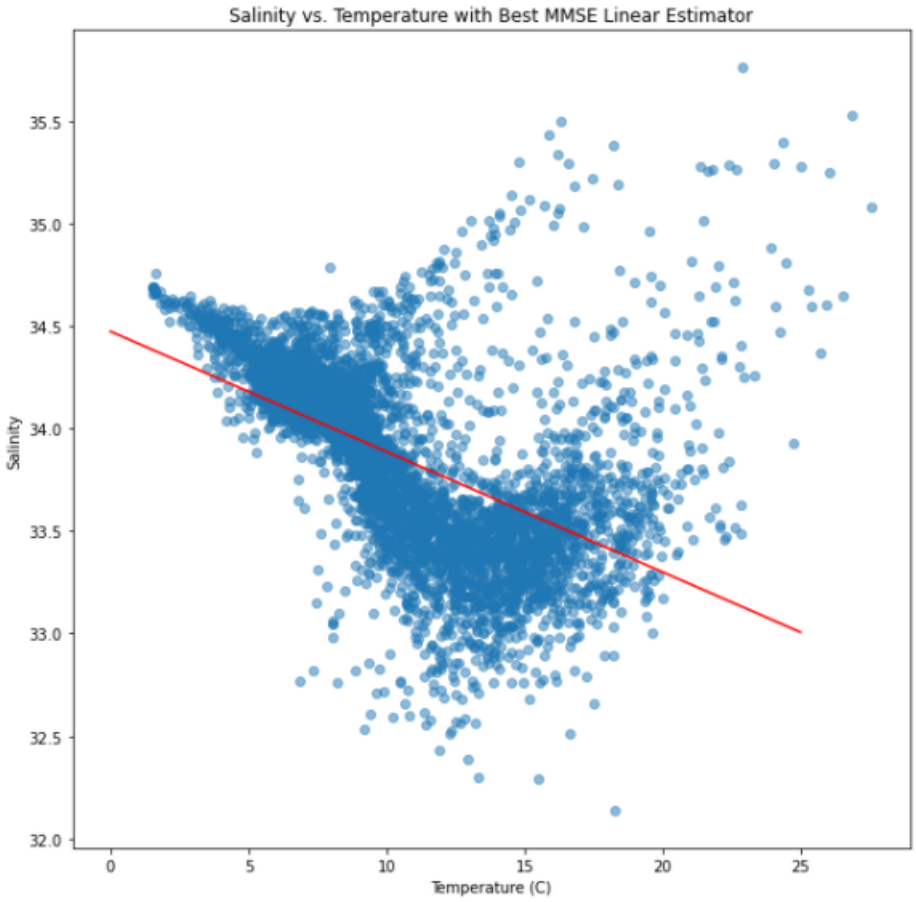
```
T = np.array(data['T_degC'])
S = np.array(data['Salnty'])
uT = np.mean(T)
uS = np.mean(S)
std_S = np.std(S)
std_T = np.std(T)

cov = np.cov(T,S)[0][1]
p = cov/(std_S*std_T)

alpha = p*std_S/std_T

beta = uS - (p*std_S*uT/std_T)
print(alpha, beta)
```

```
-0.05866370354738622 34.473402218284406
```



(b) Plot an estimate of the conditional mean of salinity given the temperature along with the scatter plot of data. What trend do you see from this plot? (Hint: this question closely follows example 5.2).

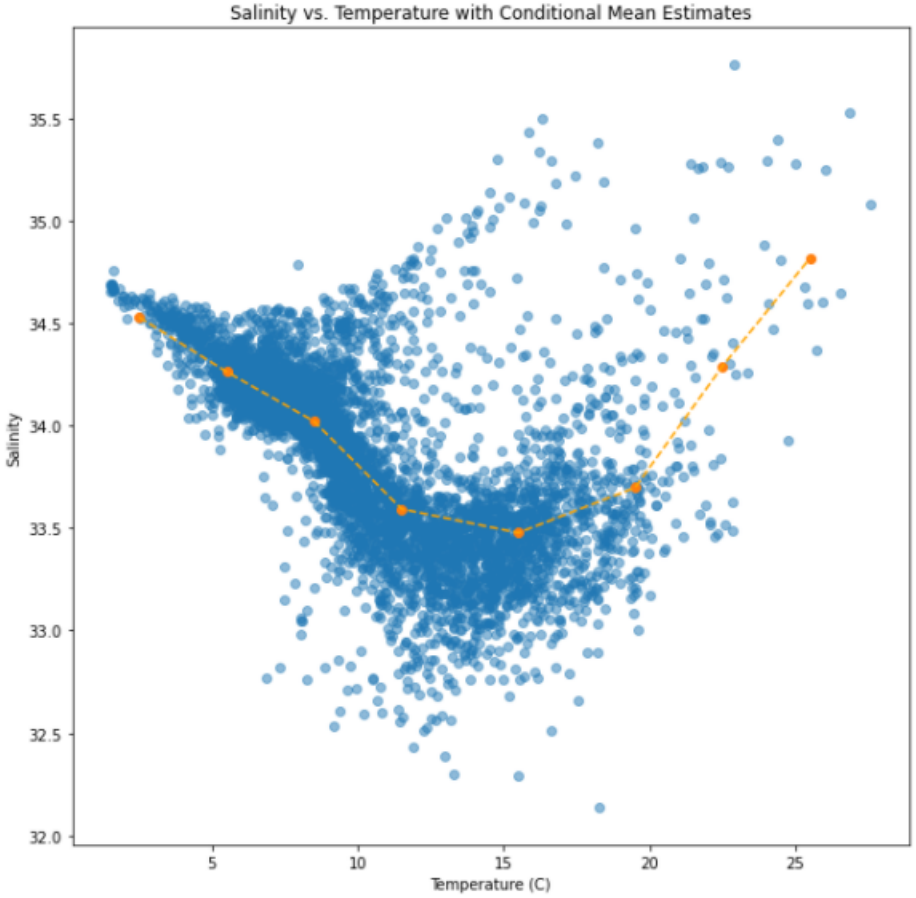
The data was binned into bins with a width of 3 degrees Celsius. The mean of each bin was calculated, therefore producing the conditional mean of salinity given temperature. The trend of the conditional means is almost parabolic, though it should be noted that the latter half of the curve is less reliable due to a lack of data. Observe the following code and plot,

```
bins = [[1,4],[4,7],[7,10],[10,13],[13,18],[18,21],[21,24],[24,27]]

print('Frequency per Bin: ')
binmeans = []
for b in bins:
    print(len(data[(data['T_degC'] < b[1]) & (data['T_degC'] > b[0])]))
    bmean = data[(data['T_degC'] < b[1]) & (data['T_degC'] > b[0])]['Salnty'].mean()
    binmeans.append(bmean)

mpoints = []
for i in bins:
    midpt = ((i[1]-i[0])/2)+i[0]
    mpoints.append(midpt)
```

```
Frequency per Bin:
124
779
1541
1030
1245
204
44
14
```



(c) Are your conditional mean estimates equally reliable at every point? If not, which estimates are more reliable? Please explain your reasoning. (We are not looking for a mathematical answer, you can reason in words)

The conditional mean estimates are certainly not equally reliable at every point. The last two conditional means are composed of only 44 and 14 data points, respectively, where as the third, fourth, and fifth conditional means are composed of over a thousand different data points each. More data is needed in the last regions in order to produce reliable conditional means.

(d) (Not graded for points) Why do you think the trend you find is different from what Sciencing suggests? It is not because of limited data - the full dataset has 810k data points and we still observe the same trend.

There is likely a confounding variable which influences the dependent variable. In other words, the salinity of the water is probably also influenced by other factors (i.e. other substances which dissolve in water, what organisms thrive in warmer environments and use salt, etc.).