EKT-816: Problem Set 1

There are 160 points available on this problem set, and it will be graded out of 120. That is, there are 40 bonus points available.

Due Date: Wednesday, February 20

Probability Review

1. Consider the following joint distribution:

↓ X Y→	1	2	3	4
0	0.1	0.05	0.025	0.025
1	0.07	0.13	0.04	0.06
2	0.1	0.1	0.25	0.05

   (a) Find the first two moments - the mean and variance - of the marginal distributions of X and Y. Check that the identity $V[X] = E[X^2] - E[X]^2$ holds.

   (b) Confirm that $E[XY] = 3.19$. Use this in combination with your calculations in (a) above to find $\text{cov}(X, Y)$.

   (c) Find the conditional probability mass function $P(Y = y|X = x)$ and use it to calculate the conditional mean function $E[Y|X = x]$ for each possible realization of X.

   (d) Calculate $E[E[Y|X = x]]$. With respect to which distribution is the outer expectation taken? The inner expectation? What do you notice about your answer as compared to the means you calculated in (a) above?

[3 + 2 + 3 + 2 = 10 pts]

2. Consider a population of workers, each with two possible skills: call them s_0 and s_1. Think of s_0 as, say, programming skill and s_1 as language skill, which is useful for becoming e.g. a lawyer. Assume that earnings are determined as follows: there is a prevailing wage $w_0 > 0$ in the tech sector and $w_1 > 0$ in the legal sector. Workers are paid by skill, so more skilled workers earn more, and in particular,

$$Y_0 = w_0 e^{s_0} \quad (1)$$

$$Y_1 = w_1 e^{s_1} \quad (2)$$

   where Y_0 and Y_1 are *potential* earnings in each sector.

Assume that the joint distribution of skills in the population is bivariate normal, with $E[s_0] = E[s_1] = 0$. (Here skills are on a logarithmic scale, so the fact that they can be negative is not a problem.) Let the standard deviation of s_0 be σ_0, the standard deviation of s_1 be σ_1, and let $\text{cov}(s_0, s_1) = \sigma_{01}$.

(a) Each worker has a choice about whether to become a programmer or a lawyer, given his or her endowment of skills (s_0, s_1), but no one can work as both (and no one decides not to work at all). What is a reasonable decision rule for a worker facing this problem?

(b) Suppose workers behave according to the decision rule you wrote down in (a) above. Find an expression for the fraction of workers who become lawyers.

Hint: use the fact that if (X, Y) are jointly normally distributed with mean zero, then

$$Y = \frac{\text{cov}(X, Y)}{V[X]} X + U \quad (3)$$

where U is independent of X, $U \sim N(0, V_U)$, and $V_U = V[Y] - \text{cov}(X, Y)^2/V[X]$.

Another hint: the sum of two normal random variables is also normally distributed.

[5 + 10 = 15 points]

3. Consider the random variable X with density

$$f_X(x) \quad = \quad \begin{cases} \frac{x}{9} & \text{if } x \in [0,3] \\ \frac{6}{9} - \frac{x}{9} & \text{if } x \in (3,6] \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

   (a) Find the CDF of X, $F_X(x)$. Draw the density f_X and the CDF F_X on the same set of axes.
   (b) Find the inverse CDF. (The inverse CDF is the function $x = G(p)$ such that $F_X(G(p)) = p$, i.e. given a probability p, G returns the value x such that $P(X \leq x) = p$.)

[5 + 5 = 10 points]

4. Table 1.2 of Chapter 1 of Deaton (1997), which is included in the assignment repository, contains summary statistics about the joint distribution of income and consumption for Ivorian households over the two years 1985-1986. Use the notation y_{it} for the income of household i in year t, and c_{it} for the consumption of household i in year t.

   (a) What is the correlation between consumption in the two years?

   Hint: use the fact that $V[c_{i,t+1} - c_{i,t}] = V[c_{i,t+1}] + V[c_{i,t}] - 2\text{cov}(c_{i,t+1}, c_{i,t})]$.

   (b) Suppose these data come from a simple random sample of households. Further, suppose there is some measurement error in the income data, so that

$$y_{it} \quad = \quad y^*_{it} + \varepsilon_{it} \tag{5}$$

   where y^*_{it} true income, and ε_{it} is "noise" that is independent of y^*_{it} each period, with $E[\varepsilon_{it}] = 0$ and $V[\varepsilon_{it}] = \sigma^2$.

   However, measurement error may be persistent over time (perhaps because it is driven by negligence on the part of the reporting household), so let ρ be the correlation between $\varepsilon_{i,t+1}$ and ε_{it}.

   Using the notation $\Delta y_{i,t+1} = y_{i,t+1} - y_{it}$ for the observed change in income for household i, and similarly for true income y^*_{it}, how would you calculate the variance in true income growth from panel data on y_{it}, if you knew ρ and σ^2?

   (c) Consider two methods of estimating the mean change in income, $E[\Delta y^*_{i,t+1}]$, over this period: (i) taking two independent cross-sectional surveys, each of size n, and (ii) collecting a two-period panel, also of size n. For method (i), say the households surveyed at time t are labelled $i = 1, \ldots n$ and those surveyed at $t+1$ are labelled $i = n+1, \ldots 2n$. The first method leads to the estimator

$$\overline{\Delta} \quad = \quad \frac{1}{n} \sum_{i=n+1}^{2n} y_{i,t+1} - \frac{1}{n} \sum_{i=1}^{n} y_{i,t} \tag{6}$$

   while the second leads to the estimator

$$\widehat{\Delta} \quad = \quad \frac{1}{n} \sum_{i=1}^{n} \Delta y_{i,t+1} \tag{7}$$

Under what conditions - on σ^2, ρ, and the joint distribution of $(y^*_{i,t+1}, y^*_{i,t})$ - will the first method be more precise than the second? (Here, "precise" means having a lower variance.) Does your answer depend on the sample size? If so, why?

[5 + 5 + 10 = 20 points]

Data Manipulation in R

Filtering, Sorting, and Generating New Variables

5. Do exercise 1 from section 5.2.4 of Wickham and Grolemund (2017).

[5 points]

6. Do exercises 3-4 of section 5.3.1 of Wickham and Grolemund (2017).

[$2 \times 5 = 10$ points]

7. Do exercise 2 of section 5.5.2 of Wickham and Grolemund (2017)

[5 points]

Grouped Summaries and Filters

8. Do exercise 6 of section 5.7.1 of Wickham and Grolemund (2017)

[5 points]

9. Do exercise 4 of section 5.6.7 of Wickham and Grolemund (2017)

[5 points]

Reshaping Data

10. Do exercise 2-4 of section 12.3.3 of Wickham and Grolemund (2017)

[$3 \times 5 = 15$ points]

Counterfactuals

Included in this repository are six short texts. Read each one carefully, and, identify the claims being made. Classify them as either causal claims, or non-causal claims. If they are not causal, what are they - value judgements? A statement of fact? Something else?

For those claims you think are causal, comment on whether the authors present any evidence for their claim. If so, what might be some concerns or alternative interpretations of that evidence? If they do not present any evidence for their claims, what sort of information would you need to decide whether it was true or false?

Please note: You are welcome to have any views you like about these topics, but I will deduct points for extraneous opinion. The point of this exercise is to read the texts carefully and critically, and to be able to recognize causal statements (including implicit ones). It is **not** to discuss the particular issues mentioned in the texts.

[$6 \times 10 = 60$ points]

References

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly. https://r4ds.had.co.nz.