Econometrics Economics/ Introduction to Econometrics
Problem Set 1 (Probability and Statistics Concepts)

def = by definition
CT = Conditioning Theorem
LIE = Law of Iterated Expectation

Question 1: Expectation and Variance

Consider the equation $y = 3 - x_1 + 2x_2 + e$ with $x_1 \sim Ber(0.5)$, $x_2 \sim N(3,4)$, and $e \sim N(0,1)$. Also assume that $\text{Corr}[x_1, x_2] = 0.5$ and $e \perp (x_1, x_2)$. Let $\mathbf{x} \equiv (1, x_1, x_2)'$.

(a) Compute $\mathrm{E}[y]$ and $\mathrm{Var}[y]$.

$$\begin{aligned}
\mathrm{E}[y] &\overset{\text{def}}{=} \mathrm{E}[3 - x_1 + 2x_2 + e] \\
&= 3 - \mathrm{E}[x_1] + 2\mathrm{E}[x_2] + \mathrm{E}[e] \\
&= 3 - 0.5 + 2 * 3 + 0 \\
&= 8.5
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[y] &\overset{\text{def}}{=} \mathrm{Var}[3 - x_1 + 2x_2 + e] \\
&\overset{e\perp(x_1,x_2)}{=} \mathrm{Var}[x_1] - 4\mathrm{Cov}[x_1, x_2] + 4\mathrm{Var}[x_2] + \mathrm{Var}[e] \\
&= \mathrm{Var}[x_1] - 4\mathrm{Corr}[x_1, x_2]\sqrt{\mathrm{Var}[x_1]\,\mathrm{Var}[x_2]} + 4\mathrm{Var}[x_2] + \mathrm{Var}[e] \\
&= 0.5 * (1 - 0.5) - 4 * 0.5 * \sqrt{0.5 * (1 - 0.5) * 4} + 4 * 4 + 1 \\
&= 15.25
\end{aligned}$$

(b) Compute $\mathrm{E}[\mathbf{x}'\mathbf{x}]$, $\mathrm{E}[\mathbf{x}\mathbf{x}']$ and $\mathrm{E}[\mathbf{x}e]$.

$$\begin{aligned}
\mathrm{E}\left[\mathbf{x}'\mathbf{x}\right] &\overset{\text{def}}{=} \mathrm{E}\left[\begin{pmatrix} 1 & x_1 & x_2 \end{pmatrix}\begin{pmatrix} 1 \\ x_1 \\ x_2 \end{pmatrix}\right] \\
&= \mathrm{E}\left[1 + x_1^2 + x_2^2\right] \\
&= 1 + \mathrm{Var}[x_1] + \mathrm{E}[x_1]^2 + \mathrm{Var}[x_2] + \mathrm{E}[x_2]^2 \\
&= 1 + 0.5 * (1 - 0.5) + 0.5^2 + 4 + 3^2 \\
&= 14.5
\end{aligned}$$

$$
\begin{aligned}
\mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] &\overset{\text{def}}{=} \mathrm{E}\left[\begin{pmatrix}1\\ x_1\\ x_2\end{pmatrix}\begin{pmatrix}1 & x_1 & x_2\end{pmatrix}\right]\\
&= \mathrm{E}\left[\begin{pmatrix}1 & x_1 & x_2\\ x_1 & x_1^2 & x_1x_2\\ x_2 & x_1x_2 & x_2^2\end{pmatrix}\right]\\
&= \begin{pmatrix}1 & \mathrm{E}\left[x_1\right] & \mathrm{E}\left[x_2\right]\\ \mathrm{E}\left[x_1\right] & \mathrm{E}\left[x_1^2\right] & \mathrm{Cov}\left[x_1, x_2\right] + \mathrm{E}\left[x_1\right]\mathrm{E}\left[x_2\right]\\ \mathrm{E}\left[x_2\right] & \mathrm{Cov}\left[x_1, x_2\right] + \mathrm{E}\left[x_1\right]\mathrm{E}\left[x_2\right] & \mathrm{E}\left[x_2^2\right]\end{pmatrix}\\
&= \begin{pmatrix}1 & 0.5 & 3\\ 0.5 & 0.5 & 2\\ 3 & 2 & 13\end{pmatrix}
\end{aligned}
$$

(*Note that* $\mathrm{E}\left[\mathbf{x}\mathbf{x}'\right]$ *always produces a symmetric matrix.*)

$$
\begin{aligned}
\mathrm{E}\left[\mathbf{x}e\right] &\overset{\text{def}}{=} \mathrm{E}\left[\begin{pmatrix}1\\ x_1\\ x_2\end{pmatrix} e\right]\\
&= \mathrm{E}\left[\begin{pmatrix}e\\ x_1e\\ x_2e\end{pmatrix}\right]\\
&\overset{e\perp(x_1,x_2)}{=} \begin{pmatrix}\mathrm{E}\left[e\right]\\ \mathrm{E}\left[x_1\right]\mathrm{E}\left[e\right]\\ \mathrm{E}\left[x_2\right]\mathrm{E}\left[e\right]\end{pmatrix}\\
&= \begin{pmatrix}0\\ 0\\ 0\end{pmatrix}
\end{aligned}
$$

(c) Compute $\mathrm{E}\left[y \mid x_1, x_2\right]$.

$$
\begin{aligned}
\mathrm{E}\left[y \mid x_1, x_2\right] &\overset{\text{def}}{=} \mathrm{E}\left[3 - x_1 + 2x_2 + e \mid x_1, x_2\right]\\
&\underset{e\perp(x_1,x_2)}{\overset{\text{CT}}{=}} 3 - x_1 + 2x_2 + \mathrm{E}\left[e\right]\\
&= 3 - x_1 + 2x_2
\end{aligned}
$$

Question 2: Independence

Consider a population with 8 individuals. The following table summarizes all information about the distribution of (y, x_1, x_2):

Individuals	y	x_1	x_2
1	6	1	1
2	6	1	1
3	3	1	1
4	4	1	0
5	5	0	1
6	3	0	0
7	3	0	0
8	0	0	0

Consider a random draw from the population.

(a) For $k_1 \in \{0,1\}$ and $k_2 \in \{0,1\}$, compute the marginal probabilities $\mathrm{P}[x_1 = k_1]$ and $\mathrm{P}[x_2 = k_2]$, the joint probabilities $\mathrm{P}[x_1 = k_1, x_2 = k_2]$, and the conditional probabilities $\mathrm{P}[x_2 = k_2 \mid x_1 = k_1]$.

Marginal Probabilities:

$$\mathrm{P}[x_1 = 1] = \frac{1}{2}, \quad \mathrm{P}[x_1 = 0] = \frac{1}{2}, \quad \mathrm{P}[x_2 = 1] = \frac{1}{2}, \quad \mathrm{P}[x_2 = 0] = \frac{1}{2}.$$

Joint Probabilities:

$$\mathrm{P}[x_1 = 1, x_2 = 1] = \frac{3}{8}, \quad \mathrm{P}[x_1 = 1, x_2 = 0] = \frac{1}{8}, \quad \mathrm{P}[x_1 = 0, x_2 = 1] = \frac{1}{8}, \quad \mathrm{P}[x_1 = 0, x_2 = 0] = \frac{3}{8}.$$

Conditional Probabilities:

$$\mathrm{P}[x_2 = 1 \mid x_1 = 1] = \frac{3}{4}, \quad \mathrm{P}[x_2 = 0 \mid x_1 = 1] = \frac{1}{4}, \quad \mathrm{P}[x_2 = 1 \mid x_1 = 0] = \frac{1}{4}, \quad \mathrm{P}[x_2 = 0 \mid x_1 = 0] = \frac{3}{4}.$$

(b) Based on the results in (a), determine whether x_1 and x_2 are independent.

(*Remark: Recall that two discrete random variables x_1 and x_2 with values in $\mathcal{X}$ are independent, if $\mathrm{P}[x_1 = k_1, x_2 = k_2] = \mathrm{P}[x_1 = k_1]\,\mathrm{P}[x_2 = k_2]$ for each $k_1 \in \mathcal{X}$ and $k_2 \in \mathcal{X}$. Provided $\mathrm{P}[x_1 = k_1] > 0$, this definition is equivalent to $\mathrm{P}[x_2 = k_2 \mid x_1 = k_1] = \mathrm{P}[x_2 = k_2]$ for each $k_1 \in \mathcal{X}$ and $k_2 \in \mathcal{X}$. Alternatively, provided $\mathrm{P}[x_2 = k_2] > 0$, this definition is also equivalent to $\mathrm{P}[x_1 = k_1 \mid x_2 = k_2] = \mathrm{P}[x_1 = k_1]$ for each $k_1 \in \mathcal{X}$ and $k_2 \in \mathcal{X}$.*)

$$\mathrm{P}[x_2 = 1 \mid x_1 = 1] = \frac{3}{4} \neq \frac{1}{2} = \mathrm{P}[x_2 = 1],$$

$$\mathrm{P}[x_2 = 0 \mid x_1 = 1] = \frac{1}{4} \neq \frac{1}{2} = \mathrm{P}[x_2 = 0],$$

$$\mathrm{P}[x_2 = 1 \mid x_1 = 0] = \frac{1}{4} \neq \frac{1}{2} = \mathrm{P}[x_2 = 0],$$

$$\mathrm{P}[x_2 = 0 \mid x_1 = 0] = \frac{3}{4} \neq \frac{1}{2} = \mathrm{P}[x_2 = 0].$$

x_1 and x_2 are not independent, since some conditional probabilities is not equal to the marginal probabilities.

(c) Compute $\mathrm{E}[y \mid x_1 = k_1, x_2 = k_2]$. Use this to compute $\mathrm{E}[y]$ by using Law of Iterated Expectation.

$$\begin{aligned}
\mathrm{E}[y \mid x_1 = 1, x_2 = 1] &= \sum_j j\mathrm{P}[y = j \mid x_1 = 1, x_2 = 1] \\
&= 6 * \frac{2}{3} + 3 * \frac{1}{3} \\
&= 5
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[y \mid x_1 = 1, x_2 = 0] &= \sum_j j\mathrm{P}[y = j \mid x_1 = 1, x_2 = 0] \\
&= 4
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[y \mid x_1 = 0, x_2 = 1] &= \sum_j j\mathrm{P}[y = j \mid x_1 = 0, x_2 = 1] \\
&= 5
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[y \mid x_1 = 0, x_2 = 0] &= \sum_j j\mathrm{P}[y = j \mid x_1 = 0, x_2 = 0] \\
&= 3 * \frac{2}{3} + 0 * \frac{1}{3} \\
&= 2
\end{aligned}$$

Compute $\mathrm{E}[y]$:

$$\begin{aligned}
\mathrm{E}[y] &\overset{\text{LIE}}{=} \mathrm{E}[\mathrm{E}[y \mid x_1, x_2]] \\
&= \sum_{k_1, k_2} \mathrm{E}[y \mid x_1 = k_1, x_2 = k_2] \, \mathrm{P}[x_1 = k_1, x_2 = k_2] \\
&= 5 * \frac{3}{8} + 4 * \frac{1}{8} + 5 * \frac{1}{8} + 2 * \frac{3}{8} \\
&= 3.75
\end{aligned}$$

Question 3:

The `cps09mar` dataset contains annual survey data on total salary earnings as well as other self-reported characteristics from 50,742 individuals across the U.S. from 2009. See `cps09mar_description.pdf` on OLAT for the description of variables in `cps09mar`. We obtain the following rounded `average earnings` (measured in 1000 U.S. dollars) according to self-reported gender:

	Men	Women
Education=0 (less than 1st grade)	26	20
Education=6 (5th or 6th grade)	27	20
Education=11 (11th grade, no high school diploma)	35	24

(a) Summarize your conclusions from the numbers above.

We find that the observed average annual earnings for men is higher than that for women. Besides, we find that the observed average gender earnings gap grows with years of education. We also find that the observed average gender earnings gap grows in a nonlinear manner. While it increases from 6 to 7 (= 1 thousand U.S. dollars) from no education to 6 years of education, it increases from 7 to 11 (= 4 thousand U.S. dollars) from 6 years of education to 11 years of education.

(b) Consider the variable `age` in the data. Explain why you think the gender earnings gap might also differ for age (no calculation needed).

`age`: Age is associated with people's working experiences. If career development is different between men and women, this will be reflected in the different average earnings between men and women. Additionally, older individuals were subject to different gender roles that influence the different income developments between men and women.

(c) Can you think of another variable for which the gender earnings gap differs strongly?

`The number of children`: As pregnancy and gender roles lead to women doing more care work (on average and at least in the past), this is reflected in the salary differences between men and women. Thus, the gender earnings gap is relatively small for individuals without children, and relatively large for individuals with children.

(d) Can you think of one for which the gender earnings gap might be smaller?

`Occupation`: Occupations have different mean earnings and some occupations are chosen more often by a specific gender. If occupation is considered, the gender earnings gap is generally lower (but still significant).