Introduction to Econometrics & R Programming

Final Exam

CPES 2 - Fall 2023

Exercise 1: Coin flips (/7,5)

Consider two coins, both with a value on each side. Coin A has the value 0 on one side and the value 100 on the other. Coin B has the value 100 on one side and the value 200 on the other. Denote x the outcome of flipping the two coins and summing the values of the visible sides. Consider that both coins are balanced, such that each of their side has a 50% chance of being drawn.
On considère deux pièces, chacune ayant une valeur sur chaque face. La pièce A a la valeur 0 sur une face et 100 sur l'autre. La pièce B a la valeur 100 sur une face et 200 sur l'autre. On note x la somme des faces visibles après le lancer des deux pièces. On considère que les pièces sont équilibrées, telles que chaque face a une probabilité d'être tirée égale à 50 %.

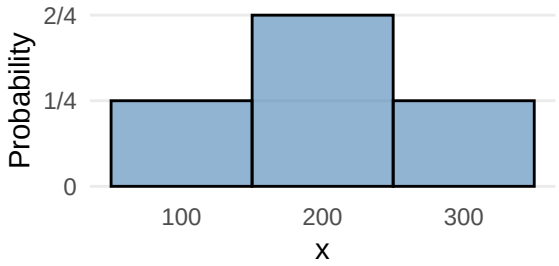
Question 1

a) Is the distribution of x left-skewed, right-skewed, or symmetrical? Justify your answer. /1
La distribution de x est-elle étalée à gauche, à droite, ou symétrique ? Justifiez.

The possible outcomes are the following:

- $0 + 100 = 100$
- $0 + 200 = 200$
- $100 + 100 = 200$
- $100 + 200 = 300$

The distribution of x is thus symmetrical as follows:



b) Compute the first and second theoretical moments of the distribution of x. /2
Calculez le premier et le deuxième moment théorique de la distribution de x.

The first theoretical moment is the expected value:

$$E[x] = \sum_i p_i x_i = (0.25 \times 100) + (0.5 \times 200) + (0.25 \times 300) = 25 + 100 + 75 = 200$$

The second theoretical moment is the variance:

$$\begin{aligned} Var[X] = E[(X - E[X])^2] &= 0.25 \times (100 - 200)^2 + 0.5 \times (200 - 200)^2 + 0.25 \times (300 - 200)^2 \\ &= 0.25 \times 10,000 + 0.5 \times 0 + 0.25 \times 10,000 \\ &= 5,000 \end{aligned}$$

Question 2

You flip both coins 200 times and gather the outcome every time. The resulting distribution has a mean of 185 and a variance of 5,000. This mean seems a bit far from its expected value for the coins to actually be balanced, but can you reject this hypothesis at the 95% confidence level? /2
Vous lancez les deux pièces 200 fois et documentez le résultat à chaque fois. La distribution qui en résulte a une moyenne de 185 et une variance de 5 000. Cette moyenne semble un peu loin de son espérance pour que les pièces soient effectivement équilibrées, mais pouvez-vous rejeter cette hypothèse au seuil de confiance de 95 % ?

We can compute the 95% confidence around 185 as follows:

$$185 \pm t_{99,97.5} \times \frac{SD(X)}{\sqrt{n}}$$

$$185 \pm t_{99,97.5} \times \frac{\sqrt{5,000}}{\sqrt{200}}$$

$$185 \pm t_{99,97.5} \times \sqrt{25}$$

$$185 \pm t_{99,97.5} \times 5$$

We can approximate $t_{99,97.5}$ by 2:

$$185 \pm 2 \times 5$$

and get the 95% confidence interval: $[175, 195]$.

200 being outside this confidence interval, we can reject with 95% confidence that the coins are actually balanced.

Question 3

You gathered the outcome of each coin flip in a `csv` file that you import in R as follows:
Vous avez documenté le résultat de chaque tirage dans un fichier `csv` *que vous importez ainsi :*

```
flips <- read.csv("C:/User/data/flips.csv")
str(flips)
```

```
'data.frame':   201 obs. of  1 variable:
 $ x: num  200 200 100 100 200 200 200 300 100 200 ...
```

When trying to compute the mean, you obtain the following outcome:
Lorsque vous tentez de calculer la moyenne, vous obtenez le résultat suivant :

```
mean(flips$x)
```

```
[1] NA
```

What is the problem and where does it seem to come from? /1,5
Quel est le problème et d'où semble-t'il venir ?

The function `mean()` returns `NA` when there is an `NA`, i.e., a missing value, in the input vector.

As the dataset contains 201 rows while there were actually 200 draws, there is probably an extra empty line at the end of the csv which is imported as an NA.

Question 4

You then run the following code to compute descriptive statistics on the variable but you obtain the following error.
Vous exécutez ensuite le code suivant pour calculer des statistiques descriptives mais vous obtenez l'erreur suivante.

```
flips %>%
  summarise(mean = mean(x),
            median = median(x),
            sd = sd(x))
```

```
Error in flips %>% summarise(mean = mean(x), median = median(x), sd = sd(x)) :
could not find function "%>%"
```

How would you solve this issue? /1
Comment régleriez-vous ce problème ?

Here, R can't find the pipe operator `%>%`. Indeed, this function is not included in base R, it comes from the `dplyr` package. To solve this problem, the `dplyr` package must be loaded with the `library` function. This requires the package to already be installed on the computer.

Exercise 2: Intergenerational income mobility (/12,5)

In Economics, the study of intergenerational income mobility generally consists in characterizing the joint distribution of individuals' income and the income of their parents. To do so, it is common to rank individuals from those who earn the lowest incomes to those who earn the highest incomes, and to divide this income distribution in 100 groups of 1% of the population. These groups are called *percentile income ranks*. Denote p^p the percentile rank of parents in their income distribution, from 1 to 100, and p^c the percentile rank attained by their children once they are adults, from 1 to 100. One way to measure intergenerational income mobility is to estimate the following regression:
En économie, l'étude de la mobilité intergénérationelle consiste généralement en la caractéristation de la distribution jointe entre le revenu des individus et le revenu de leurs parents. Pour ce faire, il est commun d'ordonner les individus de celui dont les revenus sont les plus faibles à celui dont les revenus sont les plus élevés, et de diviser cette distribution des revenus en 100 groupes de 1 % de la population. Ces groupes peuvent être considérés comme des rangs en centiles de revenus. On note p^p le rang en centile des parents dans leur distribution des revenus, de 1 à 100, et p^c le rang en centile atteint par leurs enfants à l'âge adulte, de 1 à 100. Une façon de mesurer la mobilité intergénérationnelle de revenus est d'estimer la régression suivante :

$$p_i^c = \alpha + \beta \times p_i^p + \varepsilon_i$$

Question 1

a) Using the income tax records of more than 40 million individuals in the United States, Chetty et al. (2014) estimated β to be equal to 0.34. Interpret this coefficient. /1
À partir des données fiscales de plus de 40 millions d'individus aux États-Unis, Chetty et al. (2014) ont estimé la valeur de β à 0,34. Interprétez ce coefficient.

An increase of 10 percentile ranks in the parents income distribution is associated with an increase of 3.4 percentile ranks in the income distribution for their children on average.

b) Using the same data, it can be shown that $Cor(p^p, p^c)$ is also equal to 0.34. Why is this not a coincidence? /1
À partir des mêmes données, on peut montrer que $Cor(p^p, p^c)$ vaut aussi 0,34. Pourquoi est-ce que cela n'est pas une coïncidence ?

The correlation coefficient is equal to the regression coefficient multiplied by the ratio of the standard deviations of x and y as follows:

$$\beta = Cor(x, y) \times \frac{SD(y)}{SD(x)}$$

As x and y are both defined as percentile ranks, they both follow the same uniform distribution between 1 and 100, and thus have the same standard deviation. In this specific case, the correlation coefficient and the regression coefficient are thus equal.

Question 2

For each percentile rank p^p of the parents income distribution, the following graph represents the average percentile income rank $\bar{p}_p^c$ attained by children, separately for France and the United States.
Pour chaque rang en centile p^p de la distribution des revenus des parents, le graphique suivant représente le rang moyen en centile $\bar{p}_p^c$ atteint par les enfants, séparément pour la France et les États-Unis.

a) Draw on the graph the line that would be obtained if the position of children in the income distribution was completely independent from that of their parents. Label this line "A". /0,5
Tracez sur le graphique la ligne qui serait obtenue si la position des enfants dans la distribution des revenus était complétement indépendante de celle de leurs parents. Notez cette ligne "A".

b) Draw on the graph the line that would be obtained if every child had the same position as their parents in the income distribution. Label this line "B". /0,5
Tracez sur le graphique la ligne qui serait obtenue si chaque enfant avait la même position que leurs parents dans la distribution des revenus. Notez cette ligne "B".

c) Do the two lines cross, and if so, at which coordinates? /1
Les deux lignes se croisent-elles, et si oui, à quelles coordonnées ?

The two lines cross at the average of x an y, which are equal, and which are also equal to their median: $(1 + 100)/2 = 50.5$.

So they cross at $(50.5, 50.5)$.

Question 3

a) Which country has the highest intergenerational income mobility in the bottom quartile? Motivate your answer. /0,5
Quel pays a la mobilité intergénérationnelle la plus élevée dans le quartile le plus défavorisé ? Justifiez votre réponse.

In the bottom quartile, the French rank-rank relationship is closer to the "perfect mobility" line A. The slope is slightly less steep than the US relationship, so relative mobility is higher, and expected income ranks are higher for children, so absolute upward mobility is higher as well.

b) Which country has the highest intergenerational income mobility in the top 5%? Motivate your answer. /0,5
Quel pays a la mobilité intergénérationnelle la plus élevée parmi le top 5 % ? Justifiez votre réponse.

In the top 5 parents' income ranks, the US relationship is closer to the "perfect mobility" line A. The slope is less steep than the French relationship, so relative mobility is higher, and children tend to earn relatively less, so they are further away from their parents' position in the income distribution on average.

Question 4

a) Would regressing $\bar{p}_p^c$ on p^p provide a higher R^2 on the US data or on the French data, and why? /1
La régression de $\bar{p}_p^c$ sur p^p produirait-elle un R^2 plus élevé sur les données françaises ou états-uniennes, et pourquoi ?

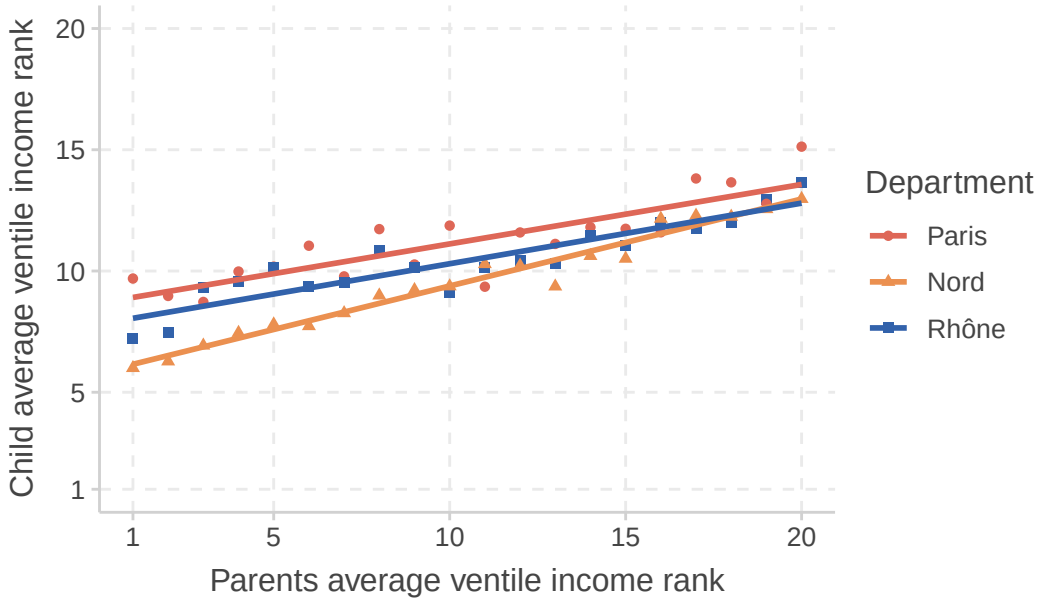
The R^2 is the share of the variance of the dependent variable which is explained by the model. Given that the French relationship is nonlinear, a straight line would not fit the data as well as for the US relationship. Data points would either be further up at the bottom of the parents income distribution, and further down at the top. The R^2 would thus be higher for the US relationship.

b) Would regressing $\bar{p}_p^c$ on a third-order polynomial of p^p provide a higher R^2 on the US data or on the French data, and why? /1
La régression de $\bar{p}_p^c$ sur un polynome d'ordre 3 de p^p produirait-elle un R^2 plus élevé sur les données françaises ou états-uniennes, et pourquoi ?

A third-order polynomial would capture the French relationship much better than a straight line. For the US relationship, it would be slightly better by construction but not much. It would still probably be very close to a straight line and the coefficients associated with x^2 and x^3 would probably be close to 0. The functional form would then be right for both relationships, but one is much less noisy than the other. The US relationship would thus have a higher R^2 because of that.

Question 5

The following graph shows the relationship between parents *ventile* income rank and the average *ventile* income rank of children, separately for those who grew up in Paris, in the Nord department, and in the Rhône department. Ventiles are used instead of centile because of sample size issues. Ventiles are computed based on the national income distribution. Table 2 in Annex provides the underlying values to this graph and additional information on the distribution of the variables p^p and $\bar{p}_p^c$ in these three departments. Table 1 reports the regression results corresponding to this graph.
Le graphique suivant représente la relation entre le rang en vingtile de revenus des parents et le rang moyen en vingtile de leurs enfants, séparément pour ceux qui ont grandi à Paris, dans le département du Nord, et dans le Rhône. Les vingtiles sont utilisés au lieu des centiles pour des raisons de taille d'échantillon. Les vingtiles sont calculés à partir de la distribution nationale des revenus. Le tableau 2 en Annexe documente les valeurs sous-jacentes à ce graphique ainsi que des informations supplémentaires sur la distribution des variables p^p et $\bar{p}_p^c$ dans ces trois départements. Le tableau 1 reporte les résultats de régression correpondants à ce graphique.



a) Interpret the results associated with: /3
Interprétez les résultats associés à :

- Department: Nord
- Parents ventile x Department: Nord
- Parents ventile x Department: Rhône

The coefficient associated with `Department: Nord` indicates that the expected income rank for an individual whose parents would locate at the 0th percentile income rank is 2.87 percentiles lower in the Nord department than in Paris. This coefficient is statistically significantly different from 0 at the 99.9% confidence level.

Table 1

	Child ventile
Parents ventile	0.245 ***
	(0.027)
Department: Nord	-2.870 ***
	(0.457)
Department: Rhône	-0.865
	(0.457)
Parents ventile x Department: Nord	0.113 **
	(0.038)
Parents ventile x Department: Rhône	0.005
	(0.038)
Constant	8.665 ***
	(0.323)
N	60
R2	0.882

*** p < 0.001; ** p < 0.01; * p < 0.05.

The coefficient associated with `Parents ventile x Department: Nord` indicates that the increase in child income rank associated with a 1 percentile increase in parents income rank is 0.113 percentile higher in the Nord department than in Paris, on average. This coefficient is statistically significantly different from 0 at the 99% confidence level.

The coefficient associated with `Parents ventile x Department: Rhône` indicates that the increase in child income rank associated with a 1 percentile increase in parents income rank is 0.005 percentile higher in the Rhône department than in Paris, on average. This coefficient is not statistically significantly different from 0 at the 95% confidence level. Thus, even though the magnitude of the coefficient is slightly larger than 0, it is not possible to conclude that there is any difference in relative mobility between Paris and the Rhône department.

b) What is the ventile-ventile correlation in the Rhône department? /1
Quelle est la corrélation vingtile-vingtile dans le département du Rhône ?

Here the correlation and the regression coefficients are not necessarily equal because the standard deviations of y and x are not necessarily equal. Indeed, while the distributions are uniform at the national level by construction, they are not necessarily uniform in each department.

$$
\begin{aligned}
Cor(x, y) &= \beta \times \frac{SD(x)}{SD(y)} \\
&= (0.245 + 0.005) \times \frac{4}{5} \\
&= 0.25 \times \frac{4}{5} = \frac{1}{5} = 0.2
\end{aligned}
$$

c) Compare the expected ventile rank for children whose parents locate at the 10th ventile in these three departments. /1,5
Comparez l'espérance du rang en vingtile pour les enfants dont les parents sont situés au 10ème vingtile dans ces trois départements.

The expected ventile of children whose parents locate at the 10th ventile can be computed from the regression coefficients in Table 1:

- Paris: $8.665 + 10 \times 0.245 = 8.665 + 2.45 = 11.115$
- Nord: $(8.665 - 2.870) + 10 \times (0.245 + 0.113) = 5.795 + 3.58 = 9.375$
- Rhône: $(8.665 - 0.865) + 10 \times (0.245 + 0.005) = 7.8 + 2.5 = 10.3$

Thus, the expected rank for children whose parents locate at the 10th ventile is higher in the Rhône than in the Nord, and even higher in Paris.

Annex

Table 2

	Paris	Rhône	Nord
Mean child ventile	12.0	10.6	9.2
Child ventile standard deviation	4.8	5.0	5.3
Mean parents ventile	13.8	11.4	8.2
Parents ventile standard deviation	3.7	4.0	5.5
Mean ventile for children from v. 1	9.7	7.2	6.0
Mean ventile for children from v. 2	9.0	7.5	6.3
Mean ventile for children from v. 3	8.7	9.3	6.9
Mean ventile for children from v. 4	10.0	9.6	7.4
Mean ventile for children from v. 5	10.2	10.1	7.8
Mean ventile for children from v. 6	11.0	9.3	7.7
Mean ventile for children from v. 7	9.8	9.5	8.3
Mean ventile for children from v. 8	11.7	10.8	9.0
Mean ventile for children from v. 9	10.3	10.2	9.2
Mean ventile for children from v. 10	11.9	9.1	9.4
Mean ventile for children from v. 11	9.4	10.1	10.3
Mean ventile for children from v. 12	11.6	10.5	10.2
Mean ventile for children from v. 13	11.1	10.3	9.4
Mean ventile for children from v. 14	11.8	11.4	10.6
Mean ventile for children from v. 15	11.7	11.0	10.5
Mean ventile for children from v. 16	11.6	12.0	12.2
Mean ventile for children from v. 17	13.8	11.8	12.3
Mean ventile for children from v. 18	13.7	12.0	12.2
Mean ventile for children from v. 19	12.8	13.0	12.6
Mean ventile for children from v. 20	15.1	13.7	13.0

⚠ NE PAS RETOURNER AVANT D'Y AVOIR ETE INVITE

Déroulé de l'examen

Durée de l'examen :

- Sans tiers-temps : 1h15
- Avec tiers-temps : 1h40

Matériel autorisé :

- Crayon(s)
- Feuille de notes A4 recto manuscrite

Copies :

- Ecrire son nom sur chaque copie
- Lever la main si besoin de copie/brouillon supplémentaire

Langue :

- L'énoncé est en anglais et en français
- Rédiger les réponses soit en anglais soit en français

Le barème est indicatif et peut être sujet à modifications.

Toute sortie est définitive.

Le sujet est à rendre dans la copie double.