Chair for Statistics and Empirical Economics
Prof. Regina T. Riphahn, Ph.D.

Exam in Panel and Evaluation Methods
Summer Term 2024

Remarks:

Grading:

- The exam consists of four problems.
- The total number of points is 60. The number of points for each problem is given in parentheses. It corresponds approximately to the recommended time spent on solving the problem (in minutes).

Important:

- Answers in German will be graded as well.
- If relevant information (necessary to solve a problem) is missing, make a plausible assumption for the missing item and briefly explain it in your answer.
- Whole sentences in your answers are not necessary, but your line of arguments should be clear and precise!

Problem 1: **[14.5 Points]**

You are using ordinary least squares (OLS) to analyze the effect of the introduction of free public transportation on the usage of public transport. You have survey data from individuals living in Luxembourg and Belgium for the years 2019 and 2021. In 2020, Luxembourg made public transport free, while Belgium did not. For tasks 1.1 - 1.3, ignore the presence of COVID-19 and its consequences in these years. Your dataset includes the following variables:

$usage_{it}$ Self-reported frequency of public transport usage by individual i in year t (number of trips per month).

$Luxembourg_{it}$ =1 if individual i in year t lives in Luxembourg; =0 otherwise.

$post_{it}$ =1 if year is 2021; =0 otherwise

1.1 Write down a regression model to estimate the effect of free public transportation on public transport usage with a Difference-in-Differences (DiD) estimation. Which variable's coefficient reflects the estimated treatment effect? What do the other explanatory variables control for? (5 points)

1.2 Define the causal effect using only conditional expectations. (3 points)

1.3 Calculate the estimated effect of the free public transportation on support for environmental policies using a Difference-in-Differences (DiD) approach based on the following sample means of *usage*: (1.5 points)

	2019	2021
if $Luxembourg = 0$	5	6
if $Luxembourg = 1$	4	8

1.4 Explain how the presence of COVID-19 could influence the identification of the causal effect using the DiD method. (2 points)

1.5 Verbally define the stable unit treatment value assumption (SUTVA). Briefly explain one reason why this assumption might not hold in this specific case. (3 points)

Problem 2: **[15.5 Points]**

You want to study the link between international rail journeys in Europe and the sense of European identity among young people. Your dataset contains a variable measuring the sense of European identity of individual i ($identity_i$). In addition, the dataset contains a dummy variable which is 1 if individual i travelled internationally by train ($train_i$). You estimate the following model, using OLS:

$$identity_i = \beta_0 + \beta_1 train_i + u_i$$

2.1 Use an example to explain why there could be omitted variable bias in this estimation. (2 points)

2.2 The DiscoverEU program runs a lottery and offers the winners the opportunity to explore Europe with free Interrail tickets. Suppose the variable $lottery_i$ is a potential instrument for $train_i$. The variable takes on the value 1 if individual i won an Interrail ticket via a randomized lottery. Explain the two conditions that a valid instrument needs to fulfill and how to test these conditions if possible. (4 points)

2.3 Verbally explain the steps to obtain the instrumental variable estimate in this case. (3 points)

Suppose Switzerland also offers a program to explore Europe, but it uses a different selection regime. It grants free Interrail tickets to every individual who is 6,570 days old or younger. The variables in your dataset are the same as before plus the variable age_i. This variable measures the age in days of individual i. You evaluate the effects of this program on the sense of European identity using a parametric regression discontinuity design (RDD).

2.4 Briefly describe your approach and its intuition using the data and establish the relevant estimation equation. If necessary, define new variables. (3 points)

2.5 Give a brief definition of *running variable*. Which of the variables in your data set is the running variable in this specific example? (1.5 points)

2.6 State the main identifying assumption of your RDD approach to identify the causal effect of train tickets on the sense of European identity. (2 points)

Problem 3: [17 Points]

At a large U.S. university, students could apply for a counseling programme in which they would meet regularly with advanced students to receive support. Among all applicants, 50% of students were randomly selected to participate in the programme. All selected students participated in the programme. You want to evaluate the effect of the program on students' final grade (ranging from 4=best to 1=worst). Using data on all applicants, you regress a student's final grade on a dummy variable (prog) which takes on the value 1 if the student participated and 0 otherwise. The following table shows the estimated coefficient b(prog) of an OLS estimation in column (1) and the estimates of quantile regressions in columns (2)-(6). The table also shows the respective p-values p(prog).

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS			Quantiles of final grade		
		10%	25%	50%	75%	90%
b(prog)	0.143	0.006	0.053	0.234	0.265	0.411
p(prog)	0.012	0.014	0.011	0.015	0.018	0.006

3.1 Interpret the estimated coefficient value and the statistical significance of the OLS estimation and the quantile regression at the 25% quantile. (4 points)

3.2 State two advantages of the quantile regression relative to the OLS regression. (2 points)

3.3 What kind of minimization problem is solved by OLS and quantile regressions, respectively? Give a verbal description and explain how the prediction errors are weighted in quantile regressions. (3 points)

3.4 Explain the connection between least absolute deviation (LAD) estimation and quantile regression. Formally state the loss functions of the LAD-estimator and the OLS-estimator. (3 points)

3.5 The counselling programme was introduced with the goal of reducing inequalities in students' final grades. Was the program successful in this regard? Briefly explain your answer. (2 points)

3.6 The standard errors of the above quantile regression are computed by bootstrapping. Explain verbally the bootstrap method to estimate standard errors. (3 points)

Problem 4: **[13 Points]**

You want to analyze the impact of national healthcare expenditure (x_{it}) on child mortality rates (y_{it}), using a balanced panel of 112 countries. The dataset includes annual observations from 2016 to 2019.

4.1 Explain briefly how a least-squares-dummy-variables (LSDV) estimation would be implemented in this case. How many parameters would be estimated? (2 Points)

4.2 You perform a Hausman test to check whether the within or random-effects estimation should be preferred. The test yields a p-value of 0.19. Briefly explain the main idea of the test, state the null and alternative hypothesis, the number of degrees of freedom, and the test decision. Briefly explain the implications of the test result for the choice of the preferred model. (4 Points)

4.3 Consider the following dynamic model for the regression of healthcare expenditure on child mortality rates:

$$y_{it} = x'_{it}\beta + \gamma_1 y_{i,t-1} + \alpha_i + \varepsilon_{it}$$

Set up the estimation model for the Anderson-Hsiao (AH) estimator in this case and describe the instrumentation the estimator uses. How many observations are used in the final estimation? (3 Points)

4.4 Explain why a first-order autoregressive process (AR(1)-process) in ε_{it} would be a problem for the AH estimator. (2 Points)

4.5 Name and explain one disadvantage of using uncorrected standard errors compared to cluster-robust standard errors and a consequence for your inference. (2 Points)