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

Exam in Panel and Evaluation Methods
Summer Term 2021

Remarks:

Number of tasks:
- The exam consists of 4 problems.

Grading:
- The total number of points is 60. The number of points for each task is given in parentheses. It corresponds approximately to the recommended time spent on solving the task (in minutes).

Important:
- Answers in German will be graded as well.
- If relevant information (necessary to solve a task) is missing, make a plausible assumption for the missing item and briefly explain it in your answer.
- Most sub-questions can be solved independent of each other. Even if you are not able to answer a certain subquestion, you nevertheless might be able to answer the following parts of the question.
- Whole sentences in your answers are not necessary, but your line of arguments should be accessible, clear and precise! Please stick to the point and avoid general discussions that are only loosely related to the question.

Problem 1 (17 points)

In 2006, Scotland introduced a smoking ban in restaurants and public places. In England, smoking in these places was not restricted. You use a data set of individuals living in Scotland and England in the years 2005 and 2007 with the following variables:

$health_{it}$	self-assessed health status on a scale from 0 [extremely bad] to 10 [extremely good] of an individual i in year t
$Scot_{it}$	=1 if individual i in year t lives in Scotland; =0 otherwise
$after_{it}$	=1 if year is 2007; =0 otherwise

1.1 Calculate the effect of the smoking ban on health using a Difference-in-Differences (DiD) approach based on the following sample means of *health*: (1.5 points)

	2005	2007
if $Scot = 0$	5.25	5.50
if $Scot = 1$	6.30	6.75

1.2 Write down a regression model to estimate the causal effect of the smoking ban on health with a DiD estimation. (2 points)

1.3 State and verbally explain the central assumption that has to hold to identify the causal effect of the smoking ban using the DiD method. Give an example in which the assumption would be violated. (4 points)

1.4 Briefly state (without abbreviation) what kind of treatment effect is identified by a consistent DiD estimation. (1 point)

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)

1.6 Alternatively, you estimate the causal effect of the smoking ban on health using propensity score matching based on a vector of characteristics (age, gender, education). Briefly describe the idea and the procedure of this method. (3.5 points)

1.7 Define the common support (overlap) requirement. (2 points)

Problem 2 (14 points)

You use a panel data set for ten countries for years 1970-2020. The following variables are available:

$crimes_{it}$	=	number of crimes per capita in country i and year t
gdp_{it}	=	gross domestic product (in US \$) in country i and year t
$educ_{it}$	=	average level of education (measured in years) in country i and year t
$area_{it}$	=	surface area (measured in square kilometers) of country i in year t

You estimate the following equation:

$$ln(crimes_{it}) = \beta_0 + \beta_1 ln(gdp_{it}) + \beta_2 educ_{it} + \beta_3 area_{it} + \epsilon_{it}$$

Within and random-effects estimations yield the following results (standard errors in parentheses):

	(1) Within	(2) Random-Effects
Ln GDP	0.016	0.023
	(0.003)	(0.002)
Education	-0.019	-0.024
	(0.015)	(0.015)
Area	.	-0.958
	(.)	(0.799)
Constant	-1.720	-0.760
	(0.534)	(0.438)

2.1 Calculate the size of the estimation sample assuming that the panel is balanced. (1 point)

2.2 Interpret the within estimator for β_1 statistically and economically. (2 points)

2.3 Explain briefly why there is no estimated coefficient for *area* from the within estimation. (2 points)

2.4 You perform a Hausman test to check if the within or a random-effects estimation should be preferred. The test yields a p-value of 0.15. Explain briefly 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)

2.5 Consider a two-period panel and the following general model: $y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$. Using the relevant transformed equations, show that the within and first-difference estimator yield identical results in a two-period panel. (5 points)

Problem 3 (12 points)

You want to analyze how industrial robots affect wages and wage inequality. Using a sample of workers in the industrial sector, you regress a worker's log hourly wage on a dummy variable (Robots) which takes on the value 1 if the worker is employed in a firm which uses industrial robots in the production and 0 otherwise. The following table shows the estimated coefficient of an OLS estimation in column (1) and the estimates of a quantile regression in columns (2)-(6).

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	Quantiles of log hourly wage				
		10%	25%	50%	75%	90%
Robots	0.024	-0.041	-0.035	0.011	0.021	0.068

3.1 Interpret the estimated coefficient of the OLS estimation and the quantile regression at the 25% quantile. (3 points)

3.2 State one advantage of the quantile regression relative to the OLS regression. (1 point)

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 What do the above tabulated results of the quantile regression suggest regarding the association between robots and wage inequality? Explain. (2 points)

Problem 4 (17 points)

You want to estimate the effect of education on life satisfaction. You have a cross-sectional dataset including the following information on 2800 individuals:

$satisf_i$	self-assessed life satisfaction on a scale from 0 [extremely unhappy] to 10 [extremely happy] of individual i.
$educ_i$	years of education of individual i.
age_i	age of individual i in years.
$age2_i$	squared age of individual i in years.
$dist_i$	distance of the place of birth of individual i to the next university (in kilometers).

You estimate the following model using ordinary least squares (OLS):

$$satisf_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age2_i + e_i$$

4.1 Name and briefly explain a potential source of endogeneity for education. (2 points)

4.2 Assume that $dist_i$, age_i, and $age2_i$ are exogenous. The (Durbin-Wu-) Hausman test can be used to check whether the variable $educ_i$ is actually endogenous. Sketch the test procedure including the necessary estimation equations. (5 points)

4.3 Provide the necessary estimation equations for a two-stage-least-squares estimation (2SLS) where you instrument the variable $educ_i$ using the variable $dist_i$. (3 points)

4.4 State the moment condition that needs to be replaced using the instrument for an IV estimation of β_1. (2 points)

4.5 Name and explain the two main conditions that a valid instrumental variable must fulfill. (3 points)

4.6 Name two problems that arise when instruments are only weakly correlated with the endogenous variable. (2 points)