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

Exam in Panel and Evaluation Methods
Winter Term 2019/2020

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.
- In case that 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 independently from 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!

Problem 1 (22 Punkte)

You would like to estimate the effect of education on fertility. You have a cross-sectional dataset including the following information on 3,000 women:

$fert_i$	Number of children of woman i.
$educ_i$	Years of education of woman i.
age_i	Woman i's age (in years)
ref_i	=1 if woman was affected by a reform raising the minimum school leaving age from 14 to 15; =0 otherwise.
$dist_i$	Distance of the place of birth of woman i to the next university (in kilometers).

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

$$fert_i = \beta_1 + \beta_2 educ_i + \beta_3 age_i + e_i$$

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

1.2 Assume that age_i, ref_i and $dist_i$ 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)

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

1.4 State the moment condition that needs to be replaced using the instrument(s) for an IV estimation of β_2. (2 points)

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

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

1.7 Describe briefly how can you test and decide whether an IV regression using ref_i and $dist_i$ as instruments for $educ_i$ is affected by the weak instrument problem. (2 points)

1.8 Consider the case where you use the variables ref_i and $dist_i$ as instruments for $educ_i$. Explain whether you can use the test for overidentifying restrictions to test the validity of the instruments. Explain the idea behind this test. (3 points)

Problem 2 (20 points)

You would like to estimate the impact of subsidies for research and development (R&D) on firms' investment spending. You have a cross-section of 2000 Italian firms who applied for a subsidy program. Firms' project proposals were graded and those projects that scored above a threshold were subsidized (otherwise they were rejected). The data set contains the following variables:

$investment_i$	investment of firm i (natural logarithm).
$subsidized_i$	=1 if firm i's project was subsidized; =0 otherwise.
$score_i$	number of points the proposal scored, 0-100.

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

$$investment_i = \beta_1 + \beta_2 subsidized_i + \epsilon_i$$

2.1 Your estimates yield, among others, the following results: $\hat{\beta}_2 = 0.2$, with an estimated standard error $se(\hat{\beta}_2) = 0.05$. Interpret the coefficient in terms of economic and statistical significance. (2 points)

2.2 Assume that in addition to the variables above, you also know the threshold of points necessary to have the project subsidized. Your colleague suggests to use a regression discontinuity (RD) design exploiting this information. Briefly describe the main idea behind RD designs intuitively. (1 point)

2.3 What is the difference between sharp and fuzzy RD designs? (3 points)

2.4 Name and explain two disadvantages of using an RD design. (4 points)

2.5 Suppose that in this setting, a sharp RD design is appropriate. What is the key identifying assumption for this research design to estimate the causal effect of subsidies on investment? (2 points)

2.6 Write down your estimation equation with an RD design. Define new variables if necessary. (2 points)

2.7 Explain whether you could improve your estimation strategy by matching on the projects having the same score. (1 point)

2.8 You receive additional data on the number of workers in these firms. Describe one possibility how to use this data to show the validity of your estimation strategy. (2 points)

2.9 You receive additional data on the same firms containing the same variables from the time period before the subsidy program. Because of the disadvantages of using RDD, you think about additionally implementing a difference-in-differences approach. What is the main identifying assumption for the validity of a difference-in-differences approach? (1 point)

2.10 Explain how a regression kink design differs from a regression discontinuity design. (2 points)

Problem 3 (10.5 points)

A research project investigates the effect of maternal participation in preventive medical check-ups during pregnancy ($checkup_i$, dummy variable equal to one if mother participated and zero otherwise) on her child's birth weight ($birthweight_i$, in kilograms). The researchers estimate a following model, which additionally conditions on maternal age at birth (age_i, in years).

$$birthweight_i = \beta_0 + \beta_1 checkup_i + \beta_2 age_i + \epsilon_i$$

3.1 The following table shows the estimates from the conditional quantile regressions for the variable ($checkup_i$):

q=0.05	q=0.10	q=0.50	q=0.90	q=0.95
0.34	0.29	0.21	0.09	0.02

Interpret the estimate for q=0,10 economically. (2 points)

3.2 The researchers consider an OLS estimation instead of using a quantile regression for the 50th percentile. Describe the minimization problem of the two estimators. Make sure to name the type of the loss function that is being minimized. (3 points)

3.3 What can we learn from a comparison of the coefficients at the 10th and 90th quantile? (2 points)

3.4 Assume that the true value of maternal age is w_i but instead, you only observe the variable $age_i = w_i + u_i$, where u_i is a classical measurement error. Name and briefly describe the consequences of the measurement error for the OLS estimation of the coefficient for age_i. (1.5 points)

3.5 Define *noise-to-signal-ratio* and briefly explain its relevance for the OLS estimator. (2 points)

Problem 4 (8 points)

Consider a panel data model with individual specific effects:

$$y_{it} = \mu_i + x'_{it}\beta + \epsilon_{it}$$

where μ_i are individual-specific effects. Assume that the $x_{it} \sim iid$ across i, $E[\epsilon_{it}|\mu_i, x_{i1}, ..., x_{iT}] = 0$ for all time periods $t \leq T$ and that $E[\epsilon_i\epsilon'_i|\mu_i, x_{i1}, ..., x_{iT}] = \sigma^2 I_T$, where $\epsilon_i = (\epsilon_{i1}, \epsilon_{i2}, ..., \epsilon_{iT})'$.

4.1 What assumptions on the relationship between the individual-specific effects μ_i and the regressors $x_{i1}, ..., x_{iT}$ need to be imposed for a fixed-effects estimator to deliver consistent estimates of β? Briefly explain. (1.5 points)

4.2 Explain why the fixed-effects estimator is particularly prone to attenuation bias through classical measurement error. (2.5 points)

4.3 Your colleague argues that you could also estimate the individual-specific parameters μ_i consistently using a least squares dummy variables model. Is your colleague right? Briefly explain your answer. (2 points)

4.4 Are there transformations other than the within transformation (i.e., fixed-effects) one could use to eliminate the individual-specific effects above? Name and explain the approach briefly. (2p)