Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Aufgabensammlung / Problem Sets
MS00 & MV04

Contents

Introduction to R

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`

First step:
Generate an R-script and name it `my_problemset1.R`. Add all your solutions of the following tasks to this R-script.

Task 1
Warm-up and vectors

a) Use R as a pocket calculator to compute the following numbers:
   i. $\sqrt{798}$
   ii. $|-9|$
   iii. 10^3
   iv. $\sqrt{\log(43)}$

b) Generate the variables
   i. $u = 10.5$
   ii. $v = 2$
   iii. $w = 3 \cdot u + v$
   iv. $x = 3 \cdot (u + v)$
   v. $y = (3 \cdot u) + v$
   vi. $z = e^{2.5+v}$

c) Generate the vectors
   i. $\mathbf{a} = [1 \quad 1 \quad 1 \quad 1 \quad 1]'$
   ii. $\mathbf{b} = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10]'$
   iii. $\mathbf{c} = [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1]'$

d) Generate the vector

$$\mathbf{d} = \frac{1}{e^{28} + \log(9)} \cdot [7 \quad 1.5 \quad 3 \quad 4 \quad 9]'$$

   i. Which command delivers the third element of the vector **d**?
   ii. Which command delivers the first two elements of the vector **d**?
   iii. Sort the vector **d** in increasing order and extract the minimum value.
   iv. Delete the last element of the unsorted vector **d**.

Task 2

Matrices

a) Generate the following two matrices:

$$\mathbf{A} = \begin{bmatrix} 34 & 2 & 1 \\ 78 & 32 & 13 \\ 40 & 23 & 68 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 10 & 9 & 5 \\ 5 & 3 & 2.5 \\ 2 & 1 & 1 \end{bmatrix}$$

b) Extract the first column of **A** and store it in a variable `a1`.

c) Extract the second row of **B** and store it in a variable `b2`.

d) Which command delivers the number of all elements in matrix **B**?

e) Which command delivers the number of elements in the third column of **B**?

f) Calculate the matrix product of **A** and **B**.

g) Calculate the transpose of the matrix **A**.

h) Calculate the inverses of **A** and **B**.

i) Generate the vector

$$\mathbf{v} = [8 \quad 2 \quad 4]'$$

and compute the matrix products $\mathbf{v}' \cdot \mathbf{A}$ and $\mathbf{A} \cdot \mathbf{v}$.

Task 3

Working with data.frames

The following table contains data of 10 babies.
The first column corresponds to the observation numbers, the second column to the birth weights (measured in ounces) of the babies and the third one to the average number of cigarettes, which their mothers have smoked per day during pregnancy:

Observation number	Birth weight	Number of cigarettes smoked
1	109	0
2	129	6
3	104	10
4	119	20
5	115	40
6	86	0
7	139	3
8	116	30
9	126	15
10	89	40

a) Use the table above to generate the following two vectors:
   1. `bwght_vec` contains the values of the birth weights
   2. `cigs_vec` contains the number of cigarettes smoked

b) Match the vectors `bwght_vec` and `cigs_vec` to a matrix named `mat`, which contains the birth weight as first column and the number of cigarettes smoked as second column.

c) Transform the matrix `mat` into a data.frame and name it `data`.

d) Extract the birth weight variable from the data.frame and name it `bwght`. Extract the cigarettes variable from the data.frame and name it `cigs`.

e) Use the `summary()` command to produce descriptive statistics of the variables `bwght` and `cigs`.

f) How many cigarettes have all women smoked together during their pregnancy?

g) What is the minimum, the maximum and the average birth weight?

Task 4

Working with data formats and data.frames

a) Load the data set `babies.csv` and store it as data.frame `babydata`.

b) Extract the variable `cigs` from the data.frame and name it `cigarettes`.

c) Generate a new variable `cigarettes2` which is the square of `cigarettes`.

d) Add `cigarettes2` to the data.frame `babydata` as the new variable `cigs2`.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 2 - Simple Linear Regression

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`

Task 1

Implementation and Interpretation

Load the data set `cars.csv`. The data set contains the variables

price	selling price of a car (in dollars)
age	age of a car (in years)

a) Estimate the regression model

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{age}_i + u_i$$

using the command `lm()` and store the regression object in the variable `reg`. Afterwards extract the coefficients, fitted values and residuals from `reg`.

b) Reproduce the results of a) by implementing the corresponding formulas in R (without using `lm()`).

c) Interpret the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

d) Predict the price of a 5 year old car.

Task 2

Units of Measurement and Functional Forms

Load the data set `ceosalary.csv`. It contains two variables:

salary	the annual salary of a CEO (in dollars)
sales	the annual sales of a firm (in dollars)

a) Estimate the following regression model:

$$\text{salary}_i = \beta_0 + \beta_1 \cdot \text{sales}_i + u_i \tag{1}$$

and interpret the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.

For the following sub tasks we introduce two new variables:

salary2	the annual salary of a CEO (in million dollars)
sales2	the annual sales of a firm (in million dollars)

b) Suppose now that we want to estimate the model

$$\text{salary2}_i = \beta_0 + \beta_1 \cdot \text{sales2}_i + u_i \ . \tag{2}$$

Compute the corresponding coefficients without estimating them.

c) Interpret the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ of model (2).

d) Estimate the following regression model:

$$\log(\text{salary}_i) = \beta_0 + \beta_1 \cdot \log(\text{sales}_i) + u_i \tag{3}$$

and interpret the coefficient $\hat{\beta}_1$.

e) Show that the slope coefficients of model (3) and the following model

$$\log(\text{salary2}_i) = \beta_0 + \beta_1 \cdot \log(\text{sales2}_i) + u_i \tag{4}$$

are identical.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 3 - Multiple Linear Regression

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`

Task 1

Implementation and Interpretation

Load the data set `houseprices81NA.csv`. It contains data of houses sold in 1981:

price	selling price of a house (in dollars)
area	area of a house (in square footage)
rooms	the number of rooms of a house

Some observations of the data set are missing (labeled with `NA`).

a) Use the built-in command `lm()` to estimate the regression model:

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{rooms}_i + u_i \tag{1}$$

and interpret the coefficients.

b) Use the built-in command `lm()` to estimate the regression model:

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{rooms}_i + \beta_2 \cdot \text{area}_i + u_i \tag{2}$$

and interpret the coefficients.

c) The built-in command `lm()` deletes the missing observations from the variables that are used in the estimation. Produce a detailed regression output and determine how many observations have been deleted due to missingness.

d) Inspect the variables used in model (2). Apply the nested command `which(is.na())` to each of the variables to determine which observations are missing.
Hint: The nested commands `which(is.na())` returns the indices of the missing observations.

e) Determine the degrees of freedom of model (2).

f) Determine and interpret the R^2 of model (2).

g) Reproduce the results of b) and f) by implementing the corresponding formulas in R (without using `lm()`).

Task 2

Goodness of fit

Your fellow students Arthur and Winston discuss about the goodness of fit of the following regression models

$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{rooms}_i + u_i \tag{1}$$
$$\text{price}_i = \beta_0 + \beta_1 \cdot \text{rooms}_i + \beta_2 \cdot \text{area}_i + u_i \tag{2}$$
$$\log(\text{price})_i = \beta_0 + \beta_1 \cdot \text{rooms}_i + u_i \tag{3}$$
$$\log(\text{price})_i = \beta_0 + \beta_1 \cdot \text{rooms}_i + \beta_2 \cdot \text{area}_i + u_i \tag{4}$$

Winston states that one should always choose the model with the highest R^2. Is Winston right? Explain briefly!

Problem Set 4 - Omitted Variables

Task 1

Zero Conditional Mean Assumption

Consider the following regression model:

$$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{ability}_i + u_i \ , \tag{1}$$

where wage denotes the annual wage (in dollars) and educ denotes the number of years spend on education, and ability is the general ability. Assume that MLR.1–3 hold for model (1).

a) State the formal zero conditional mean assumption under which the OLS estimator $\hat{\boldsymbol{\beta}}$ of model (1) is unbiased.

Suppose that ability is not observable and we consider the underspecified model:

$$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \epsilon_i \ . \tag{2}$$

b) State the formal zero conditional mean assumption under which the OLS estimator $\hat{\boldsymbol{\beta}}$ of model (2) is unbiased.

c) Suppose that MLR.4 holds for model (1). Is the assumption in b) likely to be fulfilled? Explain briefly!

Task 2

Consequences of Omitted Variables in a Specific Sample

Consider the following regression models:

$$\text{bwght}_i = \beta_0 + \beta_1 \cdot \text{cigs}_i + u_i \ , \tag{1}$$

$$\text{bwght}_i = \beta_0 + \beta_1 \cdot \text{cigs}_i + \beta_2 \cdot \text{faminc}_i + \epsilon_i \ , \tag{2}$$

where bwght denotes the birthweight of a baby in ounces, cigs measures, how many cigarettes a mother has smoked per day during pregnancy and faminc is the family income (in $1,000$ dollars). Further assume that MLR.1–4 hold for model (2).

a) Do you expect an unchanged, higher or lower coefficient of the variable *cigs* in model (1) compared to the one in model (2)? Explain briefly!

b) Use the R-Output on the next page to reconstruct the slope coefficient of model (1).

R-Output - Task 2:

```
> reg1 <- lm(bwght ~ cigs, data = data)
> reg2 <- lm(bwght ~ cigs + faminc, data = data)
>
> coef(reg1)[1]
119.7719
>
> coef(reg1)[2]
-0.5137721
>
> coef(reg2)[1]
116.9741
>
> coef(reg2)[2]
-0.4634075
>
> coef(reg2)[3]
0.09276474
>
> lm(cigs ~ faminc, data = data)

Call:
lm(formula = cigs ~ faminc, data = data)

Coefficients:
(Intercept)       faminc
    3.68811     -0.05515

> lm(faminc ~ cigs, data = data)

Call:
lm(formula = faminc ~ cigs, data = data)

Coefficients:
(Intercept)         cigs
    30.1598      -0.5429
```

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 5 - Multicollinearity & Standard Errors

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`

Task 1

Multicollinearity

Load the data set `collinear.csv`. It contains the variables `y, x1, x2, x3`. Inspect the following regression model

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + u_i$$

for evidence on multicollinearity. In doing so, use the following rule of thumb: there is a multicollinearity problem if $R_j^2 > 0.9$ for any $j = 1, 2, 3$.

Task 2

Standard Errors

Load the data set `rental.csv`. It contains data of different cities:

rent	average rent
pop	city population
avginc	per capita income
pctstu	percent of students in city population ($0 - 100$ %)

We are interested in the following regression model

$$\log(\text{rent}_i) = \beta_0 + \beta_1 \cdot \log(\text{pop}_i) + \beta_2 \cdot \log(\text{avginc}_i) + \beta_3 \cdot \text{pctstu}_i + u_i \ . \tag{1}$$

Assume that MLR.1–5 hold for model (1).

a) Estimate model (1) and produce a detailed regression output.

b) Reproduce the standard error $\text{se}(\hat{\beta}_1)$ shown in the regression output of a).

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 6 - t-Test

In this week we need the following packages:

Package Name	Commands
rio	import()

Task 1

One- and Two-sided t-tests

The R-Output provided below shows the detailed regression output for the following model

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + u_i$$

Use the R-Output to answer the questions. Please note that some values in the R-Output have been replaced by xxxx.

R-Output:

```
> reg <- lm(y ~ x1 + x2, data = data)
> summary(reg)

Call:
lm(formula = y ~ x1 + x2, data = data)

Residuals:
Min       1Q   Median       3Q      Max
-31.1487  -6.7355  -0.1422   6.9815  31.0247

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 122.0063     5.9687  20.441  < 2e-16 ***
x1            2.4297     1.1930   2.037     xxxx xxxx
x2            1.1613       xxxx   4.157 3.39e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.987 on 1751 degrees of freedom
Multiple R-squared:  0.01216,       Adjusted R-squared:  0.01103
F-statistic: 10.77 on 2 and 1751 DF,  p-value: 2.238e-05
```

a) Test the two-sided hypothesis H_0: $\beta_2 = 0$ at the significance level $\alpha = 5\%$.

b) Sketch the t-distribution corresponding to the hypothesis test in a). Further add the following components: the test statistic, the critical value, the significance level, and the p-value.

c) Reconstruct the missing p-value in the R-Output corresponding to β_1.
Hint: The command `pt(t, df)` gives the probability that $x \leq t$, where $x \sim t_{df}$ with df denoting the degrees of freedom.

d) Test the hypothesis H_0: $\beta_1 = 0$ vs. H_1: $\beta_1 > 0$ at the significance level $\alpha = 1\%$. Derive the test decision based on the p-value that you have reconstructed in c).

e) Test the hypothesis H_0: $\beta_1 = 2$ vs. H_1: $\beta_1 > 2$ at the significance level $\alpha = 5\%$.

f) Sketch the t-distribution corresponding to the hypothesis test in e). Further add the following components: the test statistic, the critical value, the significance level and the p-value.

Task 2

Application using the t-test

Reconsider the regression model from last week:

$$\log(\text{rent}_i) = \beta_0 + \beta_1 \cdot \log(\text{pop}_i) + \beta_2 \cdot \log(\text{avginc}_i) + \beta_3 \cdot \text{pctstu}_i + u_i$$

where the variables are defined as follows:

rent	average rent
pop	city population
avginc	per capita income
pctstu	percent of students in city population (0 – 100 %)

a) Is β_3 significantly different from zero? Use a significance level $\alpha = 5\%$.

b) Is the relative increase in rent lower than the relative increase in per capita income? Use an appropriate hypothesis test to underpin your answer and assume a significance level $\alpha = 5\%$.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 7 - F-Test

In this week we need the following packages:

Package Name	**Commands**
`rio`	`import()`
`car`	`linearHypothesis()`

F-Test

Remember the F-statistic formula from the lecture

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \cdot \frac{n-k-1}{q} \tag{1}$$

However, this formula is only valid if the dependent variable is unaffected by the restriction. The general F-statistic formula is

$$F = \frac{\text{SSR}_r - \text{SSR}_{ur}}{\text{SSR}_{ur}} \cdot \frac{n-k-1}{q} \tag{2}$$

where SSR_r and SSR_{ur} denote the sum of squared residuals ($\sum_{i=1}^{n} \hat{u}_i^2$) of the restricted and unrestricted model.

Task 1

F- and T-tests

Load the data set `ftest.csv`. We are interested in the following regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + u_i \ .$$

a) Test for overall significance at a significance level $\alpha = 5\%$.

b) Test the null hypothesis H_0: $\beta_2 = \beta_3 = 0, \beta_4 = 1$ at a significance level $\alpha = 1\%$ by using `linearHypothesis()`.

c) Reproduce the test statistic and the p-value in b) without using `linearHypothesis()`.
Hint: The command `pf(F, q, df)` gives the probability that $x \leq F$, where $x \sim F_{q,df}$ with q denoting the restrictions, and df denoting the degrees of freedom.

d) Use the F-test to test the null hypothesis H_0: "*x_2 and x_3 have the same effect on y*" at a significance level $\alpha = 5\%$ using `linearHypothesis()`.

e) Reproduce the test statistic and the p-value in d) without using `linearHypothesis()`.

f) Re-arrange the model such that you can test the hypothesis in d) with a t-test.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 8 - Functional Forms

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`
`car`	`linearHypothesis()`

Task 1

Quadratics

Load the data set `wage1.csv`. It contains the following variables:

wage	the hourly wage of a worker (in dollars)
educ	the education of a worker (in years)
exper	the labor force experience of a worker (in years)

In this task we consider the regression model:

$$\text{wage}_i = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{exper}_i + \beta_3 \cdot \text{exper}_i^2 + u_i$$

We are interested in the quadratic relationship between wage and exper.

a) Compute the estimated partial effect of exper on wage.

b) Compute and interpret the estimated partial effect of exper on wage at the mean of exper.

c) Compute the turnaround value (maximum/minimum point) of the the quadratic relationship between wage and exper. Is it a minimum or maximum? Further sketch the estimated quadratic relationship between wage and exper and the corresponding partial effect.

d) Use R to determine the number of observations for which exper exceeds the turnaround value.
Hint: Suppose `x` *is an arbitrary numerical vector. The expression* `x > 2` *returns a binary vector of length* `x`*, with i-th element equal to* `TRUE` $(= 1)$ *if the i-th element of* `x` *is larger than* 2*, else* `FALSE` $(= 0)$*.*

e) Is the effect of exper on wage (evaluated at the mean of exper) significantly different from zero? Use a F-test to underpin your answer and assume a significance level $\alpha = 5\%$.

f) Derive the corresponding restricted model of the F-test in e).

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 9 - Binary & Qualitative Information

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`
`car`	`linearHypothesis()`

Task 1

Dummy Variables

Load the data set `wage1.csv`. It contains the following variables:

wage	hourly wage (in dollars)
educ	education (in years)
exper	experience (in years)
married	= 1 if person is married, = 0 otherwise.
female	= 1 if person is female, = 0 otherwise.

a) Consider the model:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{exper}_i + \beta_3 \cdot \text{married}_i + \beta_4 \cdot \text{unmarried}_i + u_i\ ,$$

where *unmarried* is a dummy variable (= 1 if person is unmarried, = 0 otherwise). What might be problematic with this specific model specification? Explain briefly!

b) Estimate the model:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{exper}_i + \beta_3 \cdot \text{married}_i + \beta_4 \cdot \text{female}_i + u_i\ ,$$

and interpret the coefficients $\hat{\beta}_1$ and $\hat{\beta}_3$.

c) Does the regression model

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \cdot \text{educ}_i + \beta_2 \cdot \text{exper}_i + u_i$$

differ across the two groups married and unmarried persons? Conduct an appropriate test at a significance level $\alpha = 5\%$.

Task 2

Multiple Discrete Categories

Load the artificial data set `exam.csv`. It contains the following variables of the last econometrics exam:

exam	achieved exam points
studytime	average weekly study time for econometrics during the lecture period (in hours)
attend	$= 1$ if frequently attended the tutorial, $= 0$ otherwise.
multi	a categorical variable specifying subject of study and existence of previous knowledge in econometrics

More precisely, multi consists of the following four categories:

- *vwleco* (subject of study: economics, previous knowledge in econometrics: yes)
- *vwlnoeco* (subject of study: economics, previous knowledge in econometrics: no)
- *bwleco* (subject of study: business administration, previous knowledge in econometrics: yes)
- *bwlnoeco* (subject of study: business administration, previous knowledge in econometrics: no)

a) Regress exam on multi, studytime, and attend. Interpret the intercept and the coefficients corresponding to the categories bwlnoeco and vwleco of the variable multi.

b) Predict the exam points of an economics student who has previous knowledge in econometrics, frequently attended the tutorial and on average studied 4 hours a week for econometrics.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 10 - Heteroscedasticity

In this week we need the following packages:

Package Name	**Commands**
`rio`	`import()`
`car`	`linearHypothesis()`
`lmtest`	`bptest()`

Task 1

Breusch-Pagan & White test

Load the data set `heteroscedasticity.csv` and consider the following model:

$$y_i = \beta_0 + \beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + \beta_3 \cdot x_{i3} + u_i \tag{1}$$

In this task we want to inspect model (1) for the presence of heteroscedasticity. We assume a significance level $\alpha = 5\%$ in all sub tasks.

a) Is there evidence for heteroscedasticity? Conduct the Breusch-Pagan test using `bptest()` to underpin your answer.

b) Reproduce the test statistic and the p-value in b) without using `bptest()`. Also compute the corresponding critical value.
Hints:
- *the quantile q of a χ^2_k distribution with k degrees of freedom can be computed using the command `qchisq(q, k)`*
- *the command `pchisq(LM, k)` gives the probability that $x \leq LM$, where $x \sim \chi^2_k$ with k degrees of freedom.*

c) Conduct the F-test version of the Breusch-Pagan test using `linearHypothesis()`.

d) Is there evidence for heteroscedasticity? Conduct the White test using `bptest()` to underpin your answer.

Task 2

Feasible Generalized Least Squares

In this task we reconsider the model and data set of task 1 and estimate the model with the FGLS estimator.

Problem Set 11 - Miscellaneous

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`
`lmtest`	`resettest()`

Task 1

Functional Form Misspecification

Load the data set `ceosal2.csv`. It contains the following variables:

salary	CEO salary (in thousand dollars)
sales	firm sales (in million dollars)
mktval	firms market value (in million dollars)
profmarg	profit of sales (in %)
ceoten	years with company as CEO
comten	years with company

Does the regression model

$$\begin{aligned}\log(\text{salary}_i) = &\beta_0 + \beta_1 \cdot \log(\text{sales}_i) + \beta_2 \cdot \log(\text{mktval}_i) + \beta_3 \cdot \text{profmarg}_i + \\ &\beta_4 \cdot \text{ceoten}_i + \beta_5 \cdot \text{comten}_i + u_i\end{aligned} \tag{1}$$

suffer from functional form misspecification? Use the standard RESET-test to underpin your answer. Assume a significance level $\alpha = 5\%$.

Task 2

Proxy Variables

Suppose we have a data set of several high schools where the following variables are included:

math10	high school students passing a specific math test (in %)
expend	high school expenditure (in dollars per student)
lnchprg	students eligible for the federally funded lunch program (in %)

We are mainly interested in the effect of high school expenditure on students math performance and specify the following model:

$$\text{math10}_i = \beta_0 + \beta_1 \cdot \log(\text{expend}_i) + \beta_2 \cdot \text{poverty}_i + u_i\,, \tag{2}$$

where poverty is the share of students in the direct catchment area living in poverty. Assume that the assumptions MLR.1–4 hold.

a) Which problem might arise if we drop the variable poverty from model (2)? Explain briefly!

b) Suppose we don't observe poverty. Under which conditions and how can lnchprg be used to get a consistent estimate of β_1?

Task 3

Measurement Errors

Consider the following regression model

$$y_i = \beta_0 + \beta_1 \cdot x_i + u_i \ , \tag{3}$$

where MLR.1–5 hold and $\beta_j > 0 \, \forall \, j = 0, 1$.

Unfortunately, we can only obtain a noisy measure of x given by $z_i = x_i + e_i$, such that the model for the observed variable is

$$y_i = \beta_0 + \beta_1 \cdot z_i + v_i \ . \tag{4}$$

Suppose we want to estimate model (4). Given e is a classical measurement error, which statement about the estimator $\hat{\beta}_1$ for model (4) is correct?

- ○ $\hat{\beta}_1$ is unbiased.
- ○ $\hat{\beta}_1$ is consistent.
- ○ $plim \quad \hat{\beta}_1 > \beta_1$.
- ○ $plim \quad \hat{\beta}_1 < \beta_1$.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 12 - Instrumental Variables

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`
`car`	`linearHypothesis()`
`ivreg`	`ivreg()`

Task 1

Instrumental Variables Conditions

Suppose we have a data set of students who wrote the last econometrics exam. The data set contains the following variables:

points	number of points achieved in the exam
pretime	attendance at lecture (in %)
dist	distance between place of residence and university (in km)

We are mainly interested in the effect of the attendance rate on students performance in the econometrics exam. Thus, we specify the following model and assume that assumptions MLR.1–4 hold:

$$\text{points}_i = \beta_0 + \beta_1 \cdot \text{pretime}_i + \gamma \cdot \text{motivation}_i + u_i \ . \tag{1}$$

Unfortunately the variable motivation is not observable.

a) Which problem might arise if we drop the variable motivation from model (1)? Explain briefly!

b) State the two necessary conditions, in formal notation, that distance has to fulfill to get a consistent estimate of β_1.

Task 2

Demand Estimation

The data set `demand.csv` contains the following variables:

p.butter	price of butter (in EURO per kilo)
p.marg	price of margarine (in EURO per kilo)
q.butter	quantity of butter sold (in thousand kilos)
q.marg	quantity of margarine sold (in thousand kilos)
c.butter	production costs of butter (in EURO per kilo)
c.marg	production costs of margarine (in EURO per kilo)
income	average regional income (in thousand EUROs)

Consider the following demand model for butter:

$$\log(\text{q.butter}_i) = \beta_0 + \beta_1 \cdot \log(\text{p.butter}_i) + \beta_2 \cdot \log(\text{income}_i) + \beta_3 \cdot \log(\text{p.marg}_i) + u_i\ . \tag{2}$$

Model (2) suffers from an endogeneity problem because the price of butter and the price of margarine are correlated with demand shocks that are captured in the error term u.

a) Give an economic interpretation of the parameters β_1, β_2 and β_3.

b) Estimate model (2) by instrumental variable regression. Use the production costs $\log(\text{c.butter})$ and $\log(\text{c.marg})$ as instruments.

c) Are the production costs weak instruments?

d) Is the price elasticity of butter elastic? Use an appropriate hypothesis test to underpin your answer. Assume a significance level $\alpha = 5\%$.

Wirtschaftswissenschaftliche Fakultät
Düsseldorf Institute for Competition Economics (DICE)
Prof. Dr. Jannis Kück

Problem Set 13 - Panel Data Analysis

In this week we need the following packages:

Package Name	Commands
`rio`	`import()`
`plm`	`pdata.frame(), plm()`

Task 1

Panel Data Analysis

Load the data set `weapon.csv`. It is a balanced panel data set of all US states for the years 1977 to 1985 and includes the following variables:

vio	number of violent crimes (per 100,000 inhabitants)
concealed	= 1 if state permits concealed weapons,
	= 0 otherwise.
prisonrate	sentenced prisoners (per 100,000 inhabitants)
density	population (per square mile of land area divided by 1,000)
avginc	real per capita personal income in the state (in thousands dollars)
stateid	state identifier
year	year

In some US states, you are allowed to carry concealed weapons, i. e. other people are not able to see whether you are carrying a weapon or not. We are interested in the effect of the possibility to carry concealed weapons on the number of violent crimes per 100,000 inhabitants and specify the following model:

$$\text{vio}_{it} = \beta_0 + \beta_1 \cdot \text{concealed}_{it} + \beta_2 \cdot \text{prisonrate}_{it} + \beta_3 \cdot \text{density}_{it} + \beta_4 \cdot \text{avginc}_{it} + v_{it}\,. \tag{1}$$

In this task, we will estimate model (1) with pooled OLS, fixed effects, random effects and correlated random effects using the command `plm()`.

a) Prepare the data so that it is in the correct format for plm().
Hint: use `pdata.frame()` and specify its argument `index` to transform the corresponding `data.frame`.

b) Estimate the model using pooled OLS.

In the following sub tasks, we want to exploit the panel structure of our data set and assume that v_{it} is a composite error term: $v_{it} = a_i + u_{it}$, where a_i is a time-constant unobserved effect and u_{it} is the idiosyncratic error term.

c) Estimate the model with the random effects estimator.

d) Estimate the model with the fixed effects estimator. What is the problem?

Now we assume that prisonrate is endogenous due to correlation with the time-constant unobserved effect a.

e) Estimate the model with the correlated random effects estimator.