Exam MSB 112

14 December, 2021

General instructions

- Please read all instructions carefully and follow them exactly.
- Before you start with the exam, download all files from **Inspera** and save them into a folder on your computer. Open **R-Studio** subsequently and make sure that its *working directory* is set to the folder to which you have saved the data. Make sure that your *R-script* is located in the same folder. Always work in the script!
- Load all necessary packages upon start-up.
- When you are ask to execute commands in **R-Studio**, copy the commands from the exam file into your script and execute them directly in the order specified.
- You have 180 minutes to complete the exam.
- This is an open book exam, you are allowed to use all documents and notes from the course. You can also search for answers on the internet.
- It is not allowed to communicate in any way with any other person but the instructor during the exam. Not complying with this results in an immediate failure and exclusion from the exam.
- The exam has 36 multiple choice questions.
- There is always only one answer correct and wrong answers will give zero points.
- Record your answers in the *Excel answer* file (*Student_Answers.xlsx*) you download from Inspera. The *Excel answer* file is password protected and will only open in read-only. That is OK, you can't overwrite the file. You will have to save and work with a copy of it. Just try to save the file and you will be ask to create a copy. Do that and just work with the copy in the following.
- In the *Excel answer* file, select your answer from the drop down list in the second column in the same row in which the section is listed. If you don't want to answer a question leave it blank, or select blank from the drop down.
- Save the *Excel answer* file once you have completed the Exam. (It is a good idea to write down all your answers in a separate file before and copy them into the answer Excel file at the end of the exam.)
- Enter the version number of your exam that is shown in the lower right corner of the first page of your exam in the second column of the last line in the *Excel answer* file.
- To complete the exam upload the *Excel answer* file to **Inspera**
- In all your interpretations consider a level of significance of 0.05.
- In case your results are not identical to the ones given in the multiple choices, this might be due to differences in rounding. Select the answer closest to your values.

Version 83

Some packages you will likely need.

- library(sf)
- library(tidyverse)
- library(openxlsx)
- library(ggcorrplot)
- library(stargazer)
- library(olsrr)
- library(rms)
- library(lavaan)
- library(tidyquant)
- library(tseries)
- library(sweep)
- library(lmtest)
- library(sandwich)
- library(plm)
- library(pder)

Part 1

You are lucky as you have won a diary farm in a lottery. Unfortunately, you are not much of a farmer (yet) and you still have to learn a lot. Fortunately, you are excellently trained in data analytics, which will allow you to gain some first knowledge into how the dairy farm works. The data set (*Farms.xlsx*) can be found on *Inspera* and should be downloaded to your computer. It includes information on different farms' total milk production in liter (*MILK*), the number of cows (*COWS*), the size of its land in square acres (*LAND*), and the number of employees (*LABOR*). Load it into **R-Studio** and assign it to the object *farm.data*. Afterwards, copy and execute **ALL** of the following commands in *R-Studio* directly in this order:

set.seed(83)

farm.ids <- farm.data %>% pull(FARM) %>% unique() %>% sample(size=100)

my.pdata <- farm.data %>% filter(FARM %in% farm.ids)

From now on, only work with the *my.pdata* object, as your primary data object. Do not continue working with *Exam.data*. You can check if you are using the right data, if your first 6 lines of the data look like the one in the table below.

FARM	YEAR	COWS	LAND	MILK	LABOR	FEED
1	93	15.3	8.0	73647	2	33435.74
1	94	18.1	8.0	91260	2	36869.04
1	95	17.8	8.0	118498	2	54153.59
1	96	17.3	8.0	111454	2	50711.57
1	97	17.1	7.0	110419	2	51013.58
1	98	19.5	7.2	131197	2	59038.65

Question 1

What is the difference between panel data and time series data?

Choice	Answer text
A	Panel data have several entities/units over time.
B	Panel data have many entities/units observed at one moment in time.
C	Panel data have shorter time periods.
D	Panel data have one entity/unit over time.

Question 2

Use the **my.pdata** data set on diary farms. What is the average milk production per cow in the year 96?

Choice	Answer text
A	The avarage milk production per cow is: 5884.389 liter.
B	The avarage milk production per cow is: 5862.9094 liter.
C	The avarage milk production per cow is: 5695.1163 liter.
D	The avarage milk production per cow is: 5551.4823 liter.

Question 3

You want to know why the average milk production per cow differs between farms. More precisely, you want to assess if a farm's milk production per cow depends on the size of farms (*LAND*), its labour input (*LABOR*) or its number of cows (*COWS*). What is the appropriate regression model to use?

Choice	Answer text
A	Because the p-value of the poolability test is: 0.3797 and the p-value of the Hausman test is 0.02122 a pooled panel regression model will be used.
B	Because the p-value of the poolability test is: 1.0000 and the p-value of the Hausman test is 0.001306 a random effects panel regression model will be used.
C	Because the p-value of the poolability test is: 0.3797 and the p-value of the Hausman test is 0.02122 a random effects regression model will be used.
D	Because the p-value of the poolability test is: 0.3797 and the p-value of the Hausman test is 0.02122 a fixed effects panel regression model will be used.
E	Because the p-value of the poolability test is: 1.0000 and the p-value of the Hausman test is 0.001306 a pooled regression model will be used.
F	Because the p-value of the poolability test is: 1.0000 and the p-value of the Hausman test is 0.001306 a fixed effects panel regression model will be used.

Question 4

Which of these models is random effects panel regression model?

A

term	estimate	std.error	statistic	p.value
(Intercept)	4357.60908	154.429848	28.217402	0.0000000
LAND	-45.94347	11.982038	-3.834362	0.0001393
LABOR	351.45678	101.331810	3.468376	0.0005614
COWS	62.90682	7.837991	8.025885	0.0000000

B

term	estimate	std.error	statistic	p.value
(Intercept)	4416.81172	248.12291	17.8009025	0.0000000
LAND	-20.13308	10.90659	-1.8459556	0.0648987
LABOR	26.21998	136.48161	0.1921137	0.8476532
COWS	69.50702	7.61700	9.1252492	0.0000000

C

term	estimate	std.error	statistic	p.value
LAND	-9.551023	11.74064	-0.8135012	0.4163203
LABOR	-151.010333	178.30610	-0.8469163	0.3974494
COWS	74.944810	8.71998	8.5946084	0.0000000

Question 5

You run a number of models with different specifications explaining the amount of milk produced per cow by farms. What models are presented in the table?

	Dependent variable:		
	MILK_COW		
	Model 1	Model 2	Model 3
LAND	−9.551	−20.133*	−45.943***
	(11.741)	(10.907)	(11.982)
LABOR	−151.010	26.220	351.457***
	(178.306)	(136.482)	(101.332)
COWS	74.945***	69.507***	62.907***
	(8.720)	(7.617)	(7.838)
Constant		4,416.812***	4,357.609***
		(248.123)	(154.430)
Observations	600	600	600
R^2	0.149	0.157	0.232
Adjusted R^2	−0.026	0.153	0.228
F Statistic	28.895*** (df = 3; 497)	111.161***	60.056*** (df = 3; 596)
Note:			*p<0.1; **p<0.05; ***p<0.01

Choice	Answer text
A	Model 1 = Random effects panel regression; Model 2 = Fixed effects panel regression; Model 3 = Pooling regression
B	Model 1 = Pooling regression; Model 2 = Fixed effects panel regression; Model 3 = Random effects regression
C	Model 1 = Pooling regression; Model 2 = Random effects panel regression; Model 3 = Fixed effects regression
D	Model 1 = Fixed effects panel regression; Model 2 = Random effects panel regression; Model 3 = Pooling regression

Part 2

You are interested in financial analysis in general and in currencies as well as exchange rates in particular. You are given a data set containing the exchange rate between British pounds and US dolllar (GBPUSD), the short-term interest rates in the two countries (GBP3M and USD3M) and the interest rate difference (I_DIFF=USD3M-GBP3M). The data set (*GBPUSD_2021.xlsx*) can be found on *Inspera* and should be downloaded to your computer. Load it into **R-Studio** and assign it to the object *gbpusd*. Afterwards, copy and execute **ALL** of the following commands in *R-Studio* directly after in this order:

set.seed(83)

set.num<-sample(1:500,size=1)

gbpusd <- gbpusd %>% arrange(DATE)

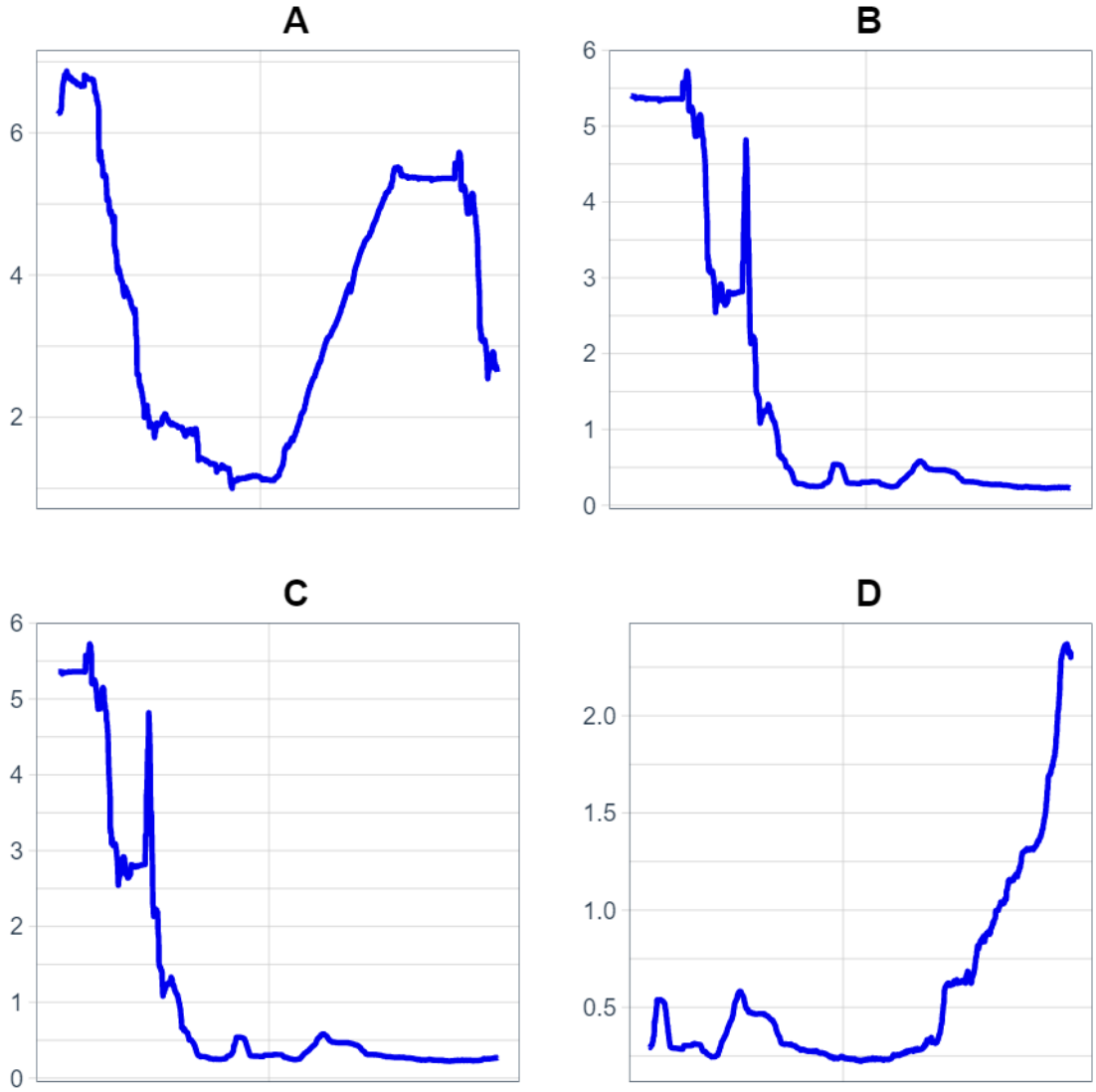
my.gbpusd <- gbpusd %>% slice(set.num:(set.num+1999))

From now on, only work with the *my.gbpusd* object, as your primary data object. You can check if you are using the right data, if your first 6 lines of the data look like the one in the table below.

DATE	GBPUSD	GBP3M	USD3M	I_DIFF
2000-04-05	1.5900	6.31031	6.27125	-0.03906
2000-04-06	1.5800	6.31250	6.27125	-0.04125
2000-04-07	1.5830	6.28016	6.28000	-0.00016
2000-04-10	1.5828	6.27766	6.28000	0.00234
2000-04-11	1.5873	6.27875	6.28000	0.00125
2000-04-12	1.5882	6.27750	6.28375	0.00625

Question 6

Use the **my.gbpusd** data set on currency exchange rates. You are asked to investigate whether the British pound/US dollar exchange rate is related to the interest rate difference between the two countries. Which of the following plots represents the short-term interest rate in the USA (*USD3M*)?



Question 7

Before you run your analysis to investigate the relation between the exchange rate and the interest rate difference, you need to check whether the variables are stationary or not. You run the adf test. Which of the following p-value pairs is correct?

Choice	Answer text
A	0.6599 and 0.6682
B	0.0639 and 0.99
C	0.4294 and 0.99
D	0.7452 and 0.4992

Question 8

How do you interpret the results of the previous question? In all your interpretations consider a level of significance of 0.05.

Choice	Answer text
A	The variables are non-stationary and must be transformed.
B	The variables are stationary and must be transformed.
C	The variables are non-stationary and can be used as is in the regression model.
D	The variables are stationary and can be used as is in the regression model.
E	No correct answer is given.

Question 9

Run your chosen regression model in R based on what you now know. What do you find?

Choice	Answer text
A	The coefficient is -0.0075, which implies that when the exchange rate goes down the interest rate difference goes up.
B	The adjusted R2 is high.
C	The coefficient is -0.0242, which implies that when the exchange rate goes down the interest rate difference goes up.
D	The coefficient is -0.0242, which implies that when the exchange rate goes up the interest rate difference goes up.
E	The coefficient is -0.0075, which implies that when the exchange rate goes up the interest rate difference goes up.

Question 10

Correct your results for heteroscedasticity and autocorrelation (coeftest/vcovHAC). Do your results change?

Choice	Answer text
A	Yes, the coefficient is no longer significant.
B	Yes, the Adjusted R2 goes down.
C	Yes, the Adjusted R2 goes up.
D	No, the coefficient is still significant.

Part 3

The good news first: you are rich! At least in the following set-up. You have a budget of maximal 100,000,000 NOK to buy ONE property in New York City for yourself. While you feel good about yourself, you decide to first acquire some deeper knowledge about the property market in NYC using skills that you have acquired in your *Data Analytics and Research Methods* course during your studies. In addition to these skills, you can also exploit a detailed data set about thousands of properties in The Big Apple. The data set (*NYC_data.xlsx*) can be found on *Inspera* and should be downloaded to your computer. Load it into **R-Studio** and assign it to the object *nyc.data*. Afterwards, copy and execute **ALL** of the following commands in *R-Studio* directly and in this order:

set.seed(83)

my.data <- nyc.data %>% sample_n(size=2000)

From now on, only work with the *my.data* object, as your primary data object. Do not continue working with *Exam.data*. You can check if you are using the right data, if your first 6 lines of the data look like the one in the table below.

id	bedrooms	floor	appointments	price	latitude	longitude	property_type	rating	accuracy	cleanness	accessibility	infrastructure	amenities	services	rooms	assemdist	district	house
504322	1	1	18	1460000	40.73034	-74.00056	Entire home/apt	4.94	4.94	4.78	4.94	5.00	5.00	4.89	10	66	District_6	1
33964014	1	1	18	770000	40.76628	-73.91082	Private room	4.82	4.76	4.82	4.82	4.82	4.82	4.82	6	36	District_3	0
2115809	1	1	1	600000	40.66625	-73.98840	Private room	5.00	5.00	5.00	5.00	5.00	5.00	5.00	10	52	District_5	0
9350645	2	2	3	980000	40.70908	-73.94255	Entire home/apt	4.67	5.00	4.33	4.67	5.00	4.67	5.00	6	53	District_5	1
48583184	2	3	1	1200000	40.66715	-73.94990	Entire home/apt	5.00	5.00	5.00	5.00	5.00	5.00	5.00	8	43	District_4	1
22382300	1	4	2	1900000	40.73076	-74.00288	Entire home/apt	5.00	5.00	5.00	5.00	5.00	5.00	4.50	7	66	District_6	1

Question 11

Use the **my.data** data set on properties in New York City. Your are interested in exploring the differences between houses and all other types of properties in New York City. To do so, you employ a regression analysis using the variable *house* as dependent variable. You are interested in how houses compare to other properties in terms of the number of rooms (*rooms*), the number of floors it has or at which it is located (*floor*), their price (*price*), and the number of appointments that have been made to take a look at it (*appointments*). You employ a regression analysis to get empirical insights. What empirical model do you employ to get some first insights?

Choice	Answer text
A	A binary logistic regression because the dependent variable is nominal in nature and has just two values.
B	An ordinary least squares regression because the dependent variable is numeric in nature.
C	A binary logistic regression because the maximum value of the dependent variable is one.
D	A multinominal logistic regression because there are multiple explanatory variables and the dependent variable is nominal in nature.
E	A multiple linear regression because there are multiple explanatory variables.

Question 12

Which of the following statements is correct with respect to your analysis?

Choice	Answer text
A	The model has the potential to give informative insights because the value of the R2 is 0.1554.
B	The model has the potential to give informative insights because the value of the LR-test is 445.402***.
C	The model has the potential to give informative insights because the value of the R2 is 0.267.
D	The model has the potential to give informative insights because multiple coefficients are significant.

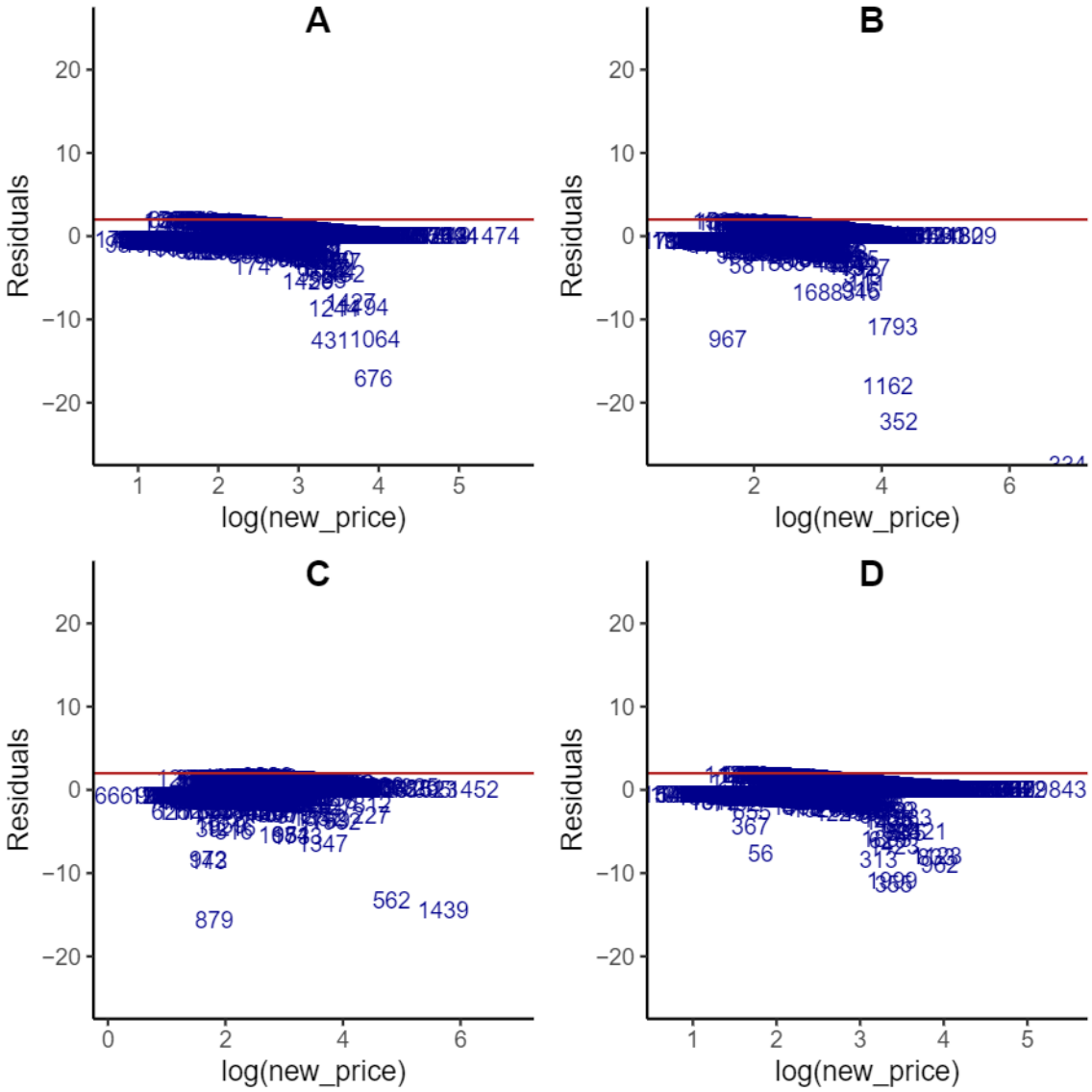
Question 13

You noticed that the values of the variable *price* are very large. To make your results more appealing, you decide to create a new variable (*new_price*) by dividing *price* by `100000` implying that the unit of this variable is now 100000 Dollar. You use this transformed variable instead now. Which of the following statements is correct with respect to your analysis?

Choice	Answer text
A	An increase in the price by 100000 Dollar implies an increases in the odds of the property being a house of 2.0068 percent.
B	An increase in the price by 100000 Dollar implies an increases in the odds of the property being a house of 1.9869 percent.
C	An increase in the price by 100000 Dollar implies an increases in the odds of the property being a house of 12.5024 percent.
D	An increase in the price by 100000 Dollar implies an increases in the odds of the property being a house of 0.0922 percent.
E	An increase in the price by 100000 Dollar implies an increases in the odds of the property being a house of 11.7805 percent.

Question 14

Which of the following plots resambles the residuals of your analysis? Note that the y-axis has been limited to the interval -25 to +25 cutting some observations off.



Question 15

Which of the following statements applies to your analysis?

Choice	Answer text
A	There are 44 potential outliers in the data.
B	There are 40 potential outliers in the data.
C	There are 37 potential outliers in the data.
D	There are 39 potential outliers in the data.

Part 4

Question 16

Use the **my.data** data set on properties in New York City. For each property, you have information on several `soft` characteristics of its neighborhood that are based on a survey. The data features an rating score (*rating*) and assessments of the accuracy of the survey information (*accuracy*), the cleanness of the local environment (*cleanness*), the accessibility (*accessibility*), the quality of the infrastructure (*infrastructure*), the existence of amenities (*amenities*), as well as the access to various services (*services*). Create a data set with all these variables in long-format. The data should have one column with the *id*, one with the type of information from the survey, and one with the corresponding values (long-format). Which of the ones below corresponds to your data?

A

id	type	score
504322	rating	4.94
504322	accuracy	4.94
504322	cleanness	4.78
504322	accessibility	4.94
504322	infrastructure	5.00
504322	amenities	5.00

B

id	type	score
25328184	rating	4.93
25328184	accuracy	4.94
25328184	cleanness	4.96
25328184	accessibility	4.97
25328184	infrastructure	4.99
25328184	amenities	4.73

C

id	type	score
48103216	rating	4
48103216	accuracy	3
48103216	cleanness	5
48103216	accessibility	3
48103216	infrastructure	5
48103216	amenities	5

D

id	type	score
41175544	rating	4.86
41175544	accuracy	4.95
41175544	cleanness	5.00
41175544	accessibility	5.00
41175544	infrastructure	5.00
41175544	amenities	4.86

Question 17

You want to see whether the overall rating is similar to a latent variable formed by the subcategories included in the survey. You decide to define a model that creates an alternative overall rating (*new_rating*) based on the variables *accuracy*, *cleanness*, *accessibility*, *infrastructure*, *amenities*, and *services*. You want to use a **cfa** model as setup. Which of the following codes will accomplish this?

Choice	Answer text
A	new_rating == accuracy + cleanness + accessibility + infrastructure + amenities + services
B	new_rating =~ accuracy + cleanness + accessibility + infrastructure + amenities + services
C	new_rating = accuracy + cleanness + accessibility + infrastructure + amenities + services
D	new_rating ~~ accuracy + cleanness + accessibility + infrastructure + amenities + services

Question 18

Run a factor analysis using the *lavaan* package and calculate the factor loadings. Which ones represent your data?

A

	Value
new_rating=~cleanness	0.9406814
new_rating=~accessibility	0.7811866
new_rating=~infrastructure	0.8384775
new_rating=~amenities	0.6972597
new_rating=~services	0.9952570
new_rating~~new_rating	0.2278906

B

	Value
new_rating=~cleanness	0.9799590
new_rating=~accessibility	0.7719258
new_rating=~infrastructure	0.8611592
new_rating=~amenities	0.6762813
new_rating=~services	0.9571808
new_rating~~new_rating	0.2052516

C

	Value
new_rating=~cleanness	0.9977294
new_rating=~accessibility	0.7577677
new_rating=~infrastructure	0.8765259
new_rating=~amenities	0.6463280
new_rating=~services	0.9958079
new_rating~~new_rating	0.1575510

D

	Value
new_rating=~cleanness	0.9388788
new_rating=~accessibility	0.9267306
new_rating=~infrastructure	0.9115428
new_rating=~amenities	0.5970357
new_rating=~services	1.0588711
new_rating~~new_rating	0.1612533

Question 19

When it comes to the interpretation of the factor loadings, which statement is correct?

Choice	Answer text
A	The interpretation of the factor loading is: increasing infrastructure by 1 increases the underlying factor by 0.84%.
B	The interpretation of the factor loading is: accuracy has no effect on amenities.
C	The interpretation of the factor loading is: accuracy has the highest correlation with the underlying factor.
D	The interpretation of the factor loading is: increasing infrastructure by 1 increases the underlying factor by 0.84.

Question 20

Predict the latent variable and show its summary statistics. Which output is correct?

A	
	new_rating
	Min. :-4.82914
	1st Qu.:-0.04449
	Median : 0.09651
	Mean : 0.00000
	3rd Qu.: 0.18517
	Max. : 0.24188

B	
	new_rating
	Min. :-3.63124
	1st Qu.:-0.04715
	Median : 0.10555
	Mean : 0.00000
	3rd Qu.: 0.19482
	Max. : 0.25433

C	
	new_rating
	Min. :-4.91078
	1st Qu.:-0.04166
	Median : 0.11458
	Mean : 0.00000
	3rd Qu.: 0.21063
	Max. : 0.27563

D	
	new_rating
	Min. :-3.87643
	1st Qu.:-0.02966
	Median : 0.12397
	Mean : 0.00000
	3rd Qu.: 0.21836
	Max. : 0.28040

Question 21

Concerning the interpretation of the predicted latent variable, which statement is correct?

Choice	Answer text
A	The third quartile of the predicted latent variable is 0.22 and it means that 75% of the variance is below 0.22.
B	The minimum value of the predicted latent variable is -3.88 and it means that the value is 3.88 standard deviations lower than the mean.
C	The mean value of the predicted latent variable is 0 and it means that the variable follows a standard normal distribution.
D	The maximum value of the predicted latent variable is 0.28 and it means that the value is 0.28 standard deviations lower than the mean.

Question 22

Next, you want to compare your predicted latent variable to the aggregate survey variable in the original data (*rating*). For that, you first need to z-transform the original variable rating. Which are the correct values of the z-transformed variable *rating*?

A	
	Value
Min.	-6.7735813
1st Qu.	-0.1401136
Median	0.2836913
Mean	0.0000000
3rd Qu.	0.5969384
Max.	0.5969384

B	
	Value
Min.	-8.0722108
1st Qu.	-0.1590672
Median	0.2757209
Mean	0.0000000
3rd Qu.	0.6235514
Max.	0.6235514

C	
	Value
Min.	-7.3481926
1st Qu.	-0.1765324
Median	0.2816570
Mean	0.0000000
3rd Qu.	0.6203187
Max.	0.6203187

D	
	Value
Min.	-9.0661521
1st Qu.	-0.1610944
Median	0.2841585
Mean	0.0000000
3rd Qu.	0.6132584
Max.	0.6132584

Question 23

Compare the z-transformed original variable (*rating*) and the predicted latent variable (*new_rating*). Which statement is correct?

Choice	Answer text
A	The predicted latent variable (new_rating) represents the data better than the original ratings variable (rating).
B	The minimum of the predicted latent variable (new_rating) is -6.77, and the minimum of the z-transformed original variable (rating) is -3.88.
C	The two variables are similar because both means are 0.
D	The minimum of the z-transformed original variable (rating) is -6.77, and the minimum of the predicted latent variable (new_rating) is -3.88.

Part 5

Question 24

Use the **my.data** data set on properties in New York City, which includes detailed information on the location of each property. The city is divided into 65 (State Assembly districts) districts whose borders are given in the shapefile *nyc.shp* that you can download from *Inspera* as well. Note the shape file (and its supportive files) is compressed into a *.zip* file, which you will need to uncompress first. The assembly districts are numbered from 23 to 87. The corresponding variable *assemdist* is included in the shapefile and in your data set. To get an overwiew, calculate the number of properties per assembly district and produce a map similar to the one shown below, which presents this number for each district. Using *ggplot2*, you can easily create such a map and add text labels (to show the number of properties per district) with the layer `geom_sf_text(aes(label = Number.of.properties))` function. You can directly pass the number of properties using the *aes()* function and the parameter *label* to it. The visibility of the numbers can be improved setting the option *size* to an adequate value. Which of the following maps represents the distribution of properties in your data?

Question 25

To better understand prices of the available properties, you decide to identify the most important determinants of their price. More precisely, you suspect that the total number of rooms (*rooms*), the number of bedrooms (*bedrooms*), the number of floors (*floor*), the number of people that have already booked an appointment (*appointments*), the type of property (*property_type*), and their location in any of NYC's larger districts (*district*, not the assembly district!) are likely influences. You employ a regression analysis to learn about their relative relevance of differences in prices. What empirical model do you employ to get some first insights?

Choice	Answer text
A	An ordinary least squares regression because the dependent variable is continuous in nature and normally distributed.
B	An ordinary least squares regression because the dependent variable is continuous in nature.
C	An ordinary least squares regression because the dependent variable is normally distributed and the explanatroy variables are continuous in nature.
D	An ordinary least squares regression because the dependent variable is normally distributed.
E	An ordinary least squares regression because the dependent and explanatroy variables are continuous in nature.

Question 26

You run the ordinary linear regression. Which of the following corresponds to your results?

Choice	Answer text
A	The price changes by 10911.37 Dollar, when the property has one additional room, and by 176028.57 Dollar when it is has an additional floor or is located one floor higher.
B	The price changes by 25798.46 Dollar, when the property has one additional room, and by 190093.19 Dollar when it is has an additional floor or is located one floor higher.
C	The price changes by 10911.37 percent, when the property has one additional room, and by 176028.57 percent when it is has an additional floor or is located one floor higher.
D	The price changes by 25798.46 percent, when the property has one additional room, and by 190093.19 percent when it is has an additional floor or is located one floor higher.

Question 27

Looking at the results of your analysis some more, which of the following statements is correct?

Choice	Answer text
A	The model explains 29.140 percent of the variance of the property price.
B	The model explains 28.680 percent of the variance of the property price.
C	The model explains 4.730 percent of the variance of the property price.
D	The model explains 4.100 percent of the variance of the property price.

Question 28

While you find the results very interesting, you suspect that maybe your empirical model might have some issues, which may limit the extent to which you can rely on it. You therefore decide to evaluate it. Which of the following applies to your model?

Choice	Answer text
A	There does not exist substantial linear dependencies among the explanatory variables, as the corresponding test returns a maximum value of 8.2824.
B	There does not exist substantial linear dependencies among the explanatory variables, as the corresponding test returns a maximum value of 7.6689.
C	There exists substantial linear dependencies among the explanatory variables, as the corresponding test returns a maximum value of 7.6689.
D	There exists substantial linear dependencies among the explanatory variables, as the corresponding test returns a maximum value of 8.2824.

Question 29

What is your reaction to the results?

Choice	Answer text
A	You will apply a stepwise regression approach to solve the issue.
B	You will estimate additional models with varying sets of explanatory variables exploring their influence on the regression results. If you don't observe substantial changes in the coefficients and their levels of significances, you will proceed with the initial set of explanatory variables.
C	You will remove all variables with a test value larger than 5 and proceed with the reduced model.
D	You will not do anything as the test does not give any indications of the presence linear dependencies between the explanatory variables.
E	You will estimate additional models with varying sets of explanatory variables exploring which of them are linearly dependent upon each other. You will then remove all of these variables.

Question 30

Independently of the outcomes of the previous questions, use the initial model explaining property prices in the following. For this model, you decide that you need to do some additional test to learn more about its reliability. Which of the following statements is correct?

Choice	Answer text
A	The observations 1599 and 1439 show extreme properties that might distort the estimation of the model.
B	The observations 1235 and 843 show extreme properties that might distort the estimation of the model.
C	The observations 334 and 1280 show extreme properties that might distort the estimation of the model.
D	The observations 1832 and 1626 show extreme properties that might distort the estimation of the model.

Question 31

Independently of the outcomes of the previous questions, use the initial model explaining property prices in the following. For this model, you decide that you need to do some additional test to learn more about its reliability. Which of the following statements is correct?

Choice	Answer text
A	All observations with a test value larger than 0.004 are outlier observations.
B	You will re-estimate the model removing all observations with test values larger than 2 because these are substantially altering your results and are likely outliers.
C	You will remove all observations with a test value larger than 0.002 from your data to obtain the final model.
D	You will re-estimate the model removing all observations with a test value larger than 0.002 from your data and assess if these observations substantially alter your results.

Question 32

Independently of the outcomes of the previous questions, use the initial model explaining property prices in the following. For this model, you decide that you need to do some additional test to learn more about its reliability. Which of the following statements is correct?

Choice	Answer text
A	The tests for coefficients' significances are not reliable because the regression residuals are normally distributed.
B	The tests for coefficients' significances is reliable because the Shapiro-Wilk test statistic is larger than 0.2604 and the p-value is 0.000.
C	The tests for coefficients' significances are not reliable because the Shapiro-Wilk test statistic is larger than 0.0604.
D	The tests for coefficients' significances is reliable because the Shapiro-Wilk test statistic is smaller than 0.2604.
E	The tests for coefficients' significances are not reliable because the regression residuals are not normally distributed.

Question 33

Independently of the outcomes of the previous questions, use the initial model explaining property prices in the following. For this model, you decide that you need to do some additional test to learn more about its reliability. Which of the following statements is correct?

Choice	Answer text
A	One requirement for the test of coefficients' significances is met because the relevant test statistic is 888.592 and the p-value is 0.000.
B	One requirement for the test of coefficients' significances is met because the null-hypothesis of the relevant test is not rejected.
C	One requirement for the test for coefficients' significances is not met because the null-hypothesis of the relevant test is rejected.
D	One requirement for the test of coefficients' significances is not met because the regression residuals are homoskedastic.
E	One requirement for the test of coefficients' significances is met because the relevant test statistic is 7.3713 and the p-value is 0.1741.

Question 34

Independently of the outcomes of the previous questions, use the initial model explaining property prices in the following. For this model, you decide that you need to do some additional test to learn more about its reliability. Which of the following statements is correct?

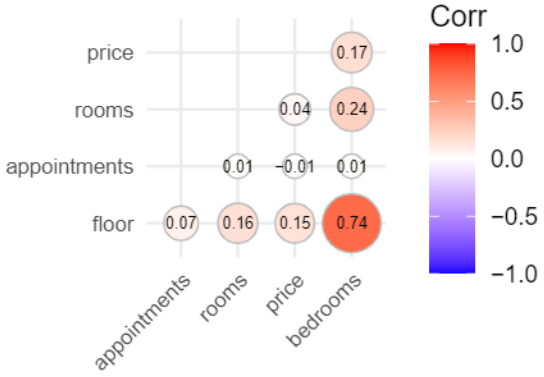
Choice	Answer text
A	The values of the models' diagnostics are for multicollinearity: 8.2824; for the residuals' normality: 0.160411***; for residuals' constant variance: 7.3713***.
B	The values of the models' diagnostics are for multicollinearity: 7.6689; for the residuals' normality: 0.160411***; for residuals' constant variance: 7.3713***.
C	The values of the models' diagnostics are for multicollinearity: 8.2824; for the residuals' normality: 0.6155773***; for residuals' constant variance: 7.3713***.
D	The values of the models' diagnostics are for multicollinearity: 8.2824; for the residuals' normality: 0.6155773***; for residuals' constant variance: 888.592***.

Part 6

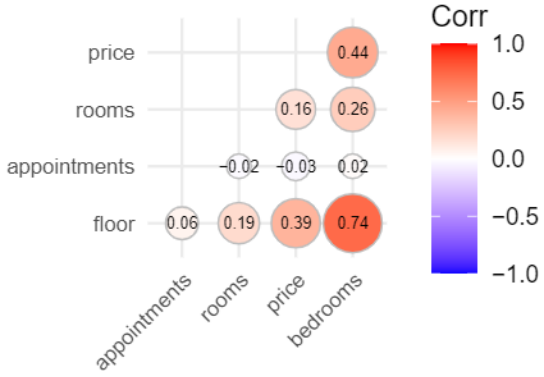
Question 35

Use the **my.data** data set on properties in New York City. Which of the following figures resembles your data?

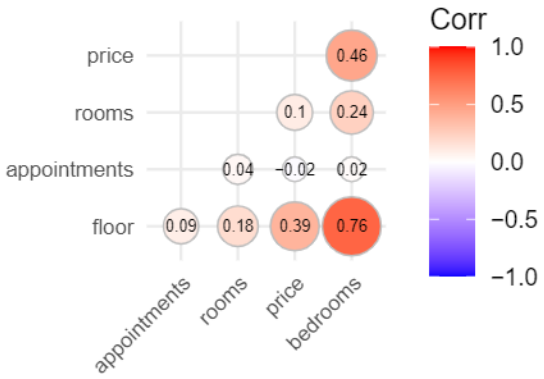
A



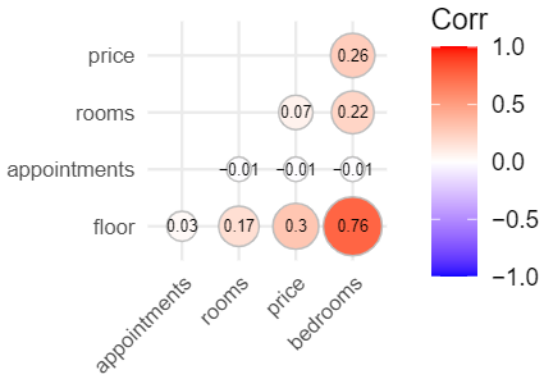
B



C



D



Question 36

Which of the following descriptive tables resembles your data?

A

	nbr.val	min	max	range	median	mean	var	std.dev
floor	2000	1e+00	18	17	1	1.7	1.300000e+00	1.1
appointments	2000	1e+00	670	669	9	31.2	2.923300e+03	54.1
rooms	2000	2e+00	15	13	7	6.8	8.700000e+00	2.9
price	2000	2e+05	29430000	29230000	1090000	1508045.0	2.733655e+12	1653376.7
bedrooms	2000	1e+00	8	7	1	1.3	5.000000e-01	0.7

B

	nbr.val	min	max	range	median	mean	var	std.dev
floor	2000	1	12	11	1	1.7	1.200000e+00	1.1
appointments	2000	1	798	797	10	33.7	3.497300e+03	59.1
rooms	2000	2	14	12	7	6.9	8.500000e+00	2.9
price	2000	110000	99990000	99880000	1130000	1670715.0	1.389324e+13	3727364.1
bedrooms	2000	1	6	5	1	1.3	4.000000e-01	0.7

C

	nbr.val	min	max	range	median	mean	var	std.dev
floor	2000	1	1.00e+01	9	1	1.7	1.300000e+00	1.1
appointments	2000	1	6.03e+02	602	10	32.4	3.143100e+03	56.1
rooms	2000	2	1.60e+01	14	7	6.9	8.500000e+00	2.9
price	2000	190000	1.00e+08	99810000	1100000	1520575.0	7.299464e+12	2701752.1
bedrooms	2000	1	6.00e+00	5	1	1.3	5.000000e-01	0.7

D

	nbr.val	min	max	range	median	mean	var	std.dev
floor	2000	1e+00	11	10	1	1.6	1.100000e+00	1.1
appointments	2000	1e+00	1009	1008	9	32.1	3.522200e+03	59.3
rooms	2000	2e+00	15	13	7	6.8	8.800000e+00	3.0
price	2000	2e+05	24000000	23800000	1150000	1569060.0	2.661365e+12	1631369.0
bedrooms	2000	1e+00	6	5	1	1.3	5.000000e-01	0.7