**Санкт-Петербургский филиал федерального государственного   
автономного образовательного учреждения высшего образования   
"Национальный исследовательский университет**

**"Высшая школа экономики"**

**Рабочая программа дисциплины**

**Количественные методы в экономике и менеджменте /**

**«Quantitative Methods in Economics and Management»**

для направления 38.06.01 «Экономика»

подготовки научно-педагогических кадров в аспирантуре,

профили: «Экономика и управление народным хозяйством

(управление инновациями, логистика)», «Математические и инструментальные методы в экономике», «Мировая экономика», «Экономика, организация и управление предприятиями, отраслями, комплексами – сфера услуг»

Разработчик программы

Антипов Е.А., к.э.н., доцент департамента менеджмента, eantipov@hse.ru

Согласована Академическим советом Аспирантской школы по экономике

«20» сентября 2016 г., протокол № 26

Согласована Академическим советом Аспирантской школы по менеджменту

«24» октября 2016 г., протокол № 18

Санкт-Петербург - 2016

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения разработчика программы.*

**Аннотация**

|  |  |  |  |
| --- | --- | --- | --- |
| Название дисциплины | **Количественные методы в экономике и менеджменте** | | |
| Образовательная программа | Направление подготовки 38.06.01  Профили: «Экономика и управление народным хозяйством (управление инновациями, логистика)», «Математические и инструментальные методы в экономике», «Мировая экономика», «Экономика, организация и управление предприятиями, отраслями, комплексами – сфера услуг» | | |
| Тип дисциплины | По выбору | | |
| Требования к уровню знаний студентов, необходимых для освоения дисциплины (пререквизиты) | Формальные пререквизиты отсутствуют, но желательны начальные знания статистики и/или эконометрики и высшей математики. | | |
| Объем з.е. | 4 | | |
| Объем в часах | Аудиторная работа | Самостоятельная работа | Всего |
| 48 | 104 | 152 |
| Краткое описание курса | Курс направлен на обучение аспирантов статистическому анализу в пакете R для проведения количественного анализа данных в их прикладных и академических исследованиях. Курс требует лишь базовых знаний математики. Аспиранты узнают, как проводить эмпирический анализ в несколько этапов, где несколько моделей сравниваются по своим статистическим свойствам и практической пользе. Курс также охватывает визуализацию данных – неотъемлемую часть разведочного анализа данных и построения моделей. Этот интенсивный курс включает не только базовые количественные методы, но и продвинутые концепции (структурные модели и анализ транзакций), которые могут оказаться новыми и полезными даже опытным аналитикам. Курс отражает как традиционный, так и байесовский подход к статистическому анализу. | | |
| Образовательные результаты по дисциплине | По завершении обучения по данному курсу аспиранты должны уметь:   * Разрабатывать и применять новые исследовательские методы (ОПК-2) * Решать экономические и управленческие задачи, использую лучшие практики анализа данных с использованием современных вычислительных средств (ПК-2) * Связывать экономическую теорию и статистику для решения реальных задач, с которыми сталкиваются в профессиональной деятельности экономисты и менеджеры (ПК-3) | | |
| Краткое содержание дисциплины | 1. Обзор языка R  2. Описательный анализ  3. Взаимосвязи между количественными переменными  4. Сравнение групп: таблицы и графики  5. Сравнение групп: статистические тесты  6. Линейные модели  7. Снижение размерности  8. Обобщенные линейные модели  9. Подтверждающий факторный анализ и моделирование структурными уравнениями  10. Сегментация: кластеризация и классификация  11. Ассоциативные правила для анализа корзин покупок  12. Моделирование выбора | | |
| Образовательные технологии | * Компьютерный практикум * Решение кейсов | | |
| Формы контроля | Формы контроля:   * Эмпирические задачи, решаемые на занятиях * Домашнее задание: тест * Экзамен: 75-минутный финальный тест на все темы   Формулы оценивания:   * Накопленная оценка= 0.5\*средняя оценка за тесты+0.5\*домашнее задание * Итоговая оценка=0.7\*Накопленная оценка+0.3\*Экзамен | | |
| Литература | Основная   1. R for Marketing Research and Analytics/ Chris Chapman, Elea McDonnell Feit. Springer-Verlag, Switzerland, 2015 (freely available through HSE’s electronic resources for HSE students and staff https://link.springer.com/content/pdf/10.1007%2F978-3-319-14436-8.pdf (Online Digital Library "SpringereBooks”) 2. Cowles K.M. (2013) Applied Bayesian Statistics: With R and OpenBUGS Examples. Springer. https://link.springer.com/book/10.1007%2F978-1-4614-5696-4 (Online Digital Library Springer Books)   Дополнительная   1. Dayal W. (2015) An Introduction to R for Quantitative Economics. New Delhi : Springer. <https://link.springer.com/book/10.1007%2F978-81-322-2340-5> (Online Digital Library "Springer eBooks”) | | |
| Преподаватель | к.э.н., доцент департамента менеджмента Антипов Евгений Александрович | | |

**Course Syllabus**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Title of the course | | **Quantitative Methods in Economics and Management** | | | | | | |
| Title of the Academic Programme | | 38.06.01 Economics  Economics and National Economy Management( Innovative Economics, Logictics), Mathematical and Instrumental Methods of Economics; The World Economy, Economics and National Economy Management (Services Sector) | | | | | | |
| Type of the course | | Optional | | | | | | |
| Prerequisites | | Desirable, but not mandatory:   * Introductory Statistics and/or Econometrics * Calculus | | | | | | |
| ECTS workload | | 4 | | | | | | |
| Total indicative study hours | | Directed Study | | Self-directed study | | | Total | |
| 48 | | 104 | | | 152 | |
| Course Overview | | This course is designed to train postgraduate students to use R for quantitative data analysis in their applied and academic research. We presume only basic statistics knowledge and use a minimum of mathematics sufficient for using quantitative methods as a tool for an economic research. Students will learn how to conduct empirical analyses in successive steps where multiple models are compared for statistical strength and practical utility. Visualization skills are taught as an integral part of data exploration and model building. This intensive course contains not only basic quantitative research methods, but also advanced concepts like structural models and transaction analysis that may be new and useful even for experienced analysts. The course reflects both traditional (“frequentist”) and Bayesian approach to statistics. | | | | | | |
| Intended Learning Outcomes (ILO) | | Upon completion of the course students will be able to (competency code is given in brackets):   * Develop and apply new research methods (ОПК-2) * Solve economic and managerial problems using best practices of data analysis using modern computational tools (ПК-2) * Link economic theory and statistics to solve real-world problems economists and managers face (ПК-3) | | | | | | |
| Teaching and Learning Methods | | * Every week a 1-2 hour introductory tutorial is given to familiarize students with the topic * A set of case studies every week is solved in class * 90% of time is allocated to practicing R programming skills | | | | | | |
| Content and Structure of the Course | | | | | | | | |
| **№** | **Topic / Course Chapter** | | **Total** | | **Directed Study** | | | **Self-directed Study** |
| **Lectures** | **Tutorials** | |
| 1 | An Overview of the R Language | | 13 | |  | 4 | | 9 |
| 2 | Descriptive analysis | | 13 | |  | 4 | | 9 |
| 3 | Relationships Between Continuous Variables | | 13 | |  | 4 | | 9 |
| 4 | Comparing Groups: Tables and Visualizations | | 13 | |  | 4 | | 9 |
| 5 | Comparing Groups: Statistical Tests | | 13 | |  | 4 | | 9 |
| 6 | Identifying Drivers of Outcomes: Linear Models | | 13 | |  | 4 | | 9 |
| 7 | Reducing Data Complexity | | 13 | |  | 4 | | 9 |
| 8 | Additional Linear Modeling Topics. Logistic Regression. Hierarchical Linear Models | | 13 | |  | 4 | | 9 |
| 9 | Confirmatory Factor Analysis and Structural  Equation Modeling | | 13 | |  | 4 | | 9 |
| 10 | Segmentation: Clustering and Classification | | 13 | |  | 4 | | 9 |
| 11 | Association Rules for Market Basket Analysis | | 13 | |  | 4 | | 9 |
| 12 | Choice Modeling | | 9 | |  | 4 | | 5 |
| **Total study hours** | | | **152** | |  | **48** | | **104** |
| Indicative Assessment Methods and Strategy | | **Assessment methods:**  **Empirical case studies solved in class:** tests given at classroom every week. Each problem set consists of 2-5 problems.  **Homework:** test based on a problem set of 2 problems.  **Exam:** Final test (duration: 75-minutes) covering all topics  **Assessment strategy:**  **Cumulative grade (before exam)=** 0.5\*Average grade across all problem sets (rounded to the nearest integer)+0.5\*Homework    **Final grade**=0.7\*Cumulative grade+0.3\*Exam | | | | | | |
| Readings / Indicative Learning Resources | | Mandatory   1. R for Marketing Research and Analytics/ Chris Chapman, Elea McDonnell Feit. Springer-Verlag, Switzerland, 2015 (freely available through HSE’s electronic resources for HSE students and staff https://link.springer.com/content/pdf/10.1007%2F978-3-319-14436-8.pdf (Online Digital Library "SpringereBooks”) 2. Cowles K.M. (2013) Applied Bayesian Statistics: With R and OpenBUGS Examples. Springer. https://link.springer.com/book/10.1007%2F978-1-4614-5696-4 (Online Digital Library Springer Books)   Optional   1. Dayal W. (2015) An Introduction to R for Quantitative Economics. New Delhi : Springer.   <https://link.springer.com/book/10.1007%2F978-81-322-2340-5> (Online Digital Library "Springer eBooks”) | | | | | | |
| Indicative Self- Study Strategies | | **Type** | | | | **+/–** | | **Hours** |
| Reading for seminars / tutorials (lecture materials, mandatory and optional resources) | | | | - | |  |
| Assignments for seminars / tutorials / labs | | | | + | | 34 |
| E-learning / distance learning (MOOC / LMS) | | | | + | | 34 |
| Fieldwork | | | | - | |  |
| Project work | | | | - | |  |
| Other (please specify) | | | | - | |  |
| Preparation for the exam | | | | + | | 36 |
| Academic Support for the Course | | Academic support for the course is provided via LMS, where students can find guidelines and recommendations for self-study and sample questions for exam preparation. The exam is also conducted using LMS testing functionality. | | | | | | |
| Facilities, Equipment and Software | | * VNC tool for sharing teacher’s screen with students * R package and RStudio environment   (latest versions are available from the following pages:  https://www.rstudio.com/products/rstudio/download/  https://cran.r-project.org/mirrors.html)   * Student resources are available from the course’s LMS page | | | | | | |
| Special conditions for organization of learning process for students with special needs | | The following types of comprehension of learning information (including e-learning and distance learning) can be offered to students with disabilities (by their written request) in accordance with their individual psychophysical characteristics:   1. *for persons with vision disorders:* a printed text in enlarged font; an electronic document; audios (transferring of learning materials into the audio); an individual advising with an assistance of a sign language interpreter; individual assignments and advising. 2. *for persons with hearing disorders: a* printed text; an electronic document; video materials with subtitles; an individual advising with an assistance of a sign language interpreter; individual assignments and advising. 3. *for persons with muscle-skeleton disorders: a* printed text; an electronic document; audios; individual assignments and advising. | | | | | | |
| Course Instructor | | Evgeny A. Antipov, PhD, Associate Professor, Department of Management | | | | | | |

**Annex 1**

**Course content:**

1. *An Overview of the R Language.*

Vectors, data frames, subsetting, new variable creation.

1. *Descriptive analysis*

Summary statistics, boxplots, barplots.

1. *Relationships Between Continuous Variables.*

Pearson, Spearman and Kendall Correlations. Scatterplots and scatterplot matrices.

1. *Comparing Groups: Tables and Visualizations*

Function “aggregate”. Visualization package “lattice”. Group comparison plots.

1. *Comparing Groups: Statistical Tests*

Parametric and nonparametric tests of means. Mood’s test of medians. Normality tests.

1. *Identifying Drivers of Outcomes: Linear Models*

Simple linear regression. Multiple linear regression. Interaction terms. Dummy variables. Regression diagnostics.

1. *Reducing Data Complexity*

Exploratory Factor Analysis. Principal Component Method.

1. *Additional Linear Modeling Topics. Logistic Regression. Hierarchical Linear Models*

Generalized linear models. Modeling binary outcomes. Multilevel regression models. Random-coefficient models.

1. *Confirmatory Factor Analysis and Structural Equation Modeling*

Introduction to “lavaan” package for CFA and SEM.

1. *Segmentation: Clustering and Classification*

Actionable segmentations using clustering and classification techniques. Hierarchical clustering. K-means clustering. Classification trees. Random Forests.

1. *Association Rules for Market Basket Analysis*

Market Basket Analysis with arules package using transaction data in wide and long forms.

1. *Choice Modeling*

Discrete choice modeling using a real-world example. Data collection. Data entry. Data processing using multinomial logit model. Interpretation.

**Аnnex 2**

**Assessment Methods** **and Criteria**

**Assessment Methods**

|  |  |  |  |
| --- | --- | --- | --- |
| **Types of Assessment** | **Forms of Assessment** | **1 Year periods** | |
| **1** | **2** | |
| Formative Assessment | In-class Participation | \* |  | |
| Homework | \* |  | |
| Summative Assessment | Exam | \* |  | |

**Assessment Criteria**

All assessments are graded automatically based on the percentage of correct answers entered by students to LMS. The range of raw scores for each assessment produced by LMS is from 0 to 100%. Percentages are converted to 10-point grades using standard rules: [95%;100%]→10, [85%,95%)→9, [75%,85%)→9.

**Assessments are conducted in LMS where there is a large database of all questions and problems. The examples of each type of assesment are presented below.**

**Empirical case studies solved in class:**

Fill in the gaps in the code.

#import data and install packages

sales<-read.csv(\*---\*)

if(!"dplyr" %in% installed.packages()) install.packages("dplyr")

library(dplyr)

if(!"lubridate" %in% installed.packages()) install.packages("lubridate")

library(lubridate)

#In data frame "sales" convert WEEK\_END\_DATE to date format

#create variable TPR =1 if there was a temporary price reduction and 0 otherwise

#create a categorical (factor) variable PROMOTYPE taking on 4 values: "no promo", "only price promo",

#"price and non-price promo" (i.e. price and at least one of two: feature and display) and "only non-price promo"

#Reminder: in R we use & and | as AND and OR operators respectively

sales<-sales%>%

mutate(WEEK\_END\_DATE=dmy(\*---\*),

TPR=ifelse(PRICE<BASE\_PRICE,1,0),

PROMOTYPE=ifelse(TPR+DISPLAY+FEATURE==0,"no promo",

ifelse(\*---\*, "only price promo",

ifelse(\*---\*,"price and non-price promo",

"only non-price promo"))),

PROMOTYPE=factor(\*---\*))

#A comparison of 4 groups of PROMOTYPE requires each of them to have a sufficient sample size,

#i.e. each type of promotion should be used in a sufficiently large number of weeks

#Now we will try to work out, data on which UPC in which store is sufficient if we want

#to compare sales in weeks with different types of promotions (based on PROMOTYPE)

#For each UPC-store combination (i.e. for each UPC in each store) find

#the number of weeks with each value of PROMOTYPE variable

#the minimum weekly number of units of this UPS sold in this store

#the maximum weekly number of units of this UPS sold in this store

#the total number of units of the UPC in the store

#the number of observations (i.e. weeks or, formally, rows) for this UPC-store combination in the "sales" data frame

upc.store.promo<-sales%>%

group\_by(UPC, STORE\_NUM)%>%

summarise(nopromotion=sum(PROMOTYPE=="no promo"),

onlypricepromo=sum(PROMOTYPE==\*---\*),

priceandnonpricepromo=sum(PROMOTYPE==\*---\*),

onlynonpricepromo=sum(PROMOTYPE==\*---\*),

minunits=min(\*---\*),

maxunits=\*---\*(\*---\*),

sumunits=\*---\*(\*---\*),

n=n())

#Convert upc.store.promo from tibble to the more traditional data.frame to allow for a wider range of manipulations

upc.store.promo<-data.frame(upc.store.promo)

#Look at the top 30 rows of the "upc.store.promo" data frame

#and make sure the data makes sense (e.g. weeks with different types of promotion add up to the total number of weeks)

head(upc.store.promo,\*---\*)

#Now we will select the "best" UPC-store combinations for our analysis and store them in "upc.store.promo.best" tibble created as follows.

#For data frame "upc.store.promo" add the number of weeks of the least widely used type of promotion for each UPC in each store

#and select UPCs and stores that meet the following condition: the least widely used PROMOTYPE was used at least 20 weeks

#AND the total sample size (i.e. the number of observations (weeks)) for this UPC-store pair is at least 156,

#i.e. the UPC was available in the store for all 3 years

upc.store.promo.best<-upc.store.promo%>%

#minimum is usually found by columns and we need to find the minimum of the 4 values

#(nopromotion, onlypricepromo, priceandnonpricepromo and onlynonpricepromo)

#for each UPC-store combination. It is possible with dplyr's rowwise()

rowwise()%>%

mutate(min\_promotype=min(nopromotion, onlypricepromo, priceandnonpricepromo, onlynonpricepromo))%>%

arrange(-min\_promotype)%>%

filter(\*---\*)

#Convert upc.store.promo.best to the data frame

upc.store.promo.best<-data.frame(\*---\*)

#Obtain store and UPC values from upc.store.best by looking UP STORE\_NUM

#and UPC corresponding to the row with the maximum value of min\_promotype

store\_num<-upc.store.promo.best[which.max(upc.store.promo.best[,"min\_promotype"]),"STORE\_NUM"]

upc<-upc.store.promo.best[which.max(upc.store.promo.best[,"min\_promotype"]),"UPC"]

#Make sure that values of store\_num and upc selected by the code above and stored as values in R Studio's "Environment"

#are the same as what you inferred when you reviewed upc.store.promo.best at the previous step.

#Prepare the dataset containing data on sales of a single product in a single store.

#PROMOTYPE has already been created, so we should only filter the dataset and arrange it

#(not necessary for the current analysis, but desirable)

#Take data frame "sales" and store it to pretzels1 tibble after some preprocessing

pretzels1<-sales%>%

#leave only data on sales of pretzels (соленые крендельки) "SNYDR SOURDOUGH NIBBLERS 16 OZ" (UPC=7797508004)

#sold in store number 6187

filter(\*---\*)%>%

#sort the dataset by week\_end\_date (optional, but makes the dataset look better)

\*---\*(\*---\*)

#Inspect the first 10 rows of pretzels1 data frame

head(\*---\*, \*---\*)

#Now we will explore the distribution of units sold overall and by groups based on PROMOTYPE

#Package tableone for summary statistics

if(!"tableone" %in% installed.packages()) install.packages("tableone")

library(tableone)

#Descriptive stats for units sold for the whole sample

desc.overall<-CreateTableOne(vars = c("UNITS"), data = \*---\*)

summary(desc.overall)

#Descriptive stats by groups based on PROMOTYPE

desc.by.group<-CreateTableOne(vars = c("UNITS"), strata=\*---\*, data = pretzels1)

summary(desc.by.group)

#density plot for the whole sample (unit sales are always non-negative,

#which is why it is a good idea to use option "xlim=c(0,max(pretzels1$UNITS))"

#whenever you find it inappropriate to show the negative part of the X axis)

plot(density(\*---\*), main="Density plot",xlim=c(0,max(pretzels1$UNITS)))

#Compare densities of unit sales distribution by weeks with different promo types

#(unit sales are always non-negative, which is why it is a good idea to use option "xlim=c(0,max(pretzels1$UNITS))"

#whenever you find it inappropriate to show the negative part of the X axis)

if(!"sm" %in% installed.packages()) install.packages("sm")

library(sm)

sm.density.compare(\*---\*, \*---\*, xlab="Units sold", model="equal", lwd=3, xlim=c(0,max(pretzels1$UNITS)))

title(main="Comparison of sales density by promotion type")

legend("topright", levels(pretzels1$PROMOTYPE), fill=1+(1:nlevels(pretzels1$PROMOTYPE)))

#boxplot of UNITS (overall)

boxplot(\*---\*, main="Boxplot",ylab="Units Sold")

#horizontal boxplot of UNITS by PROMOTYPE

#set plot margins (bottom, right, upper, left)

#to account for the fact that we have long category labels

#and change the names of categories so that they the long ones

#are split across two lines using the new line (\n) symbol

par(mar=c(2,8,2,2))

boxplot(\*---\*~\*---\*,

main="Boxplots by promotion type",

horizontal=TRUE,

xlab="Units Sold",

las=2, #las=axes label orientation (2 - perpendicular to the axis)

names=c("price and\n non-price promo",

"only price promo",

"only\n non-price promo",

"no promo"))

#Assess normality of the distribution using Q-Q plots

if(!"\*---\*" %in% installed.packages()) install.packages("\*---\*")

library(\*---\*)

qqPlot(pretzels1$UNITS, main="QQ Plot", ylab="Units Sold")

qqPlot(\*---\*[\*---\*=="no promo",]$UNITS, main="\*---\*", ylab="Units Sold")

qqPlot(\*---\*[\*---\*=="only non-price promo",]$UNITS, main="\*---\*", ylab="Units Sold")

qqPlot(\*---\*[\*---\*=="only price promo",]$UNITS, main="\*---\*", ylab="Units Sold")

qqPlot(\*---\*[\*---\*=="price and non-price promo",]$UNITS, main="\*---\*", ylab="Units Sold")

#formal normality tests (Shapiro-Wilk test)

#whole sample

\*---\*

#for weeks with no promotion

\*---\*

#for weeks with only non-price promotion

\*---\*

#for weeks with only price promotion

\*---\*

#for weeks with price and non-price promotion

\*---\*

# Fill in the gaps (\*---\*) with appropriate code

# Import data and install packages

pretzels1<-read.csv("https://goo.gl/BAwJyd")

# One-sample t-test. Test that sample mean UNIT sales equal their sample median.

t.test (pretzels1$UNITS, mu=median(\*---\*))

# Independent 2-group t-test comparing "price and non-price promo" with "only price promo"

t.test(UNITS~\*---\*, data=\*---\*, subset=(PROMOTYPE==\*---\*)|(PROMOTYPE==\*---\*))

# Mahn-Whitney test comparing "price and non-price promo" with "only price promotion"

wilcox.test(UNITS~\*---\*, data=\*---\*, subset=(PROMOTYPE==\*---\*)|(PROMOTYPE==\*---\*))

# Based on the p-values of the tests conducted above,

# what can you say about the significances of differences in units sold under two types of promotion?

# one-way ANOVA analog of UNITS by PROMOTYPE

summary(aov(\*---\*~\*---\*, data=\*---\*))

# not assuming equal variances

oneway.test(\*---\*~\*---\*, data=\*---\*)

# investigating the association between DISPLAY and TPR (reminder: TPR=temporary price reduction)

# first make sure you use factor variables

pretzels1$TPR=factor(\*---\*, levels=c(\*---\*,\*---\*), labels=c(\*---\*, \*---\*))

pretzels1$DISPLAY=factor(\*---\*, levels=c(\*---\*,\*---\*), labels=c(\*---\*, \*---\*))

# simple contingency table

table1<-table(pretzels1$TPR, pretzels1$DISPLAY)

table1

#alternative way to obtain a somewhat nicer table with headings

table1<-xtabs(~pretzels1$TPR+pretzels1$DISPLAY)

table1

# table of proportions (100% - all observations)

prop.table(table1)

# table of proportions (100% - in rows)

prop.table(table1, margin=\*---\*)

# table of proportions (100% - in columns)

prop.table(table1, margin=\*---\*)

# Chi-square test and its Fisher version.

# Null hypothesis: there is no association between TPR and DISPLAY

chisq.test(\*---\*)

fisher.test(\*---\*)

# what do the following commands do?

table2<-xtabs(~(pretzels1$PRICE>median(pretzels1$PRICE))+(pretzels1$UNITS>median(pretzels1$UNITS)))

prop.table(table2)

prop.table(table2, margin=1)

prop.table(table2, margin=2)

chisq.test(table2)

fisher.test(table2)

# scatterplot of UNITS (make sure it is on the Y axis) vs. PRICE (make sure it is on the X axis).

plot(\*---\*, \*---\*, xlab="PRICE, $", ylab="UNITS")

# correlation tests

# Pearson

cor.test(\*---\*, \*---\*, method="pearson")

# Spearman

cor.test(\*---\*, \*---\*, method="spearman")

# Kendall

cor.test(\*---\*, \*---\*, method="kendall")

**Homework:**

Analyze the link between UNITS and PRICE for the pretzels1 dataset we used before

pretzels1<-read.csv("https://goo.gl/BAwJyd")

following the instructions from the tutorial

<http://uc-r.github.io/linear_regression>

You need only sections related to the simple regression analysis (from the start until words “In the next tutorial we will look at how we can extend a simple linear regression model into a multiple regression”.

* Replication Requirements
* Preparing Our Data
* Model Building
* Assessing Coefficients
* Assessing Model Accuracy
* Assessing Our Model Visually
* Making Predictions

**Don’t forget to install packages they use in the tutorial by running install.packages(“tidyverse”), etc. first, before you start the analysis. If anything does not work on your computer remember to react to warnings and error messages by installing additional packages required.**

**Exam:**

1. Partial (imperfect) multicollinearity:

[ ] means that 2 or more regressors are strongly linearly related

[ ] occurs only in financial studies

[ ] does not lead to biased estimates of slope coefficients

[ ] leads to inefficiency of OLS estimates

[ ] occurs in economic studies

2. VIF in econometrics DOES NOT stand for:

[ ]Variable importance factor (coefficient showing the importance of the dependent variable compared to other regressors)

[ ] Variable importance factor (coefficient showing the importance of an explanatory variable compared to other regressors)

[ ] Variance inflation factor (number of times, in which a regression coefficient estimate’s variance is decreased compare to the situation of no correlation among regressors)

[ ] Variance inflation factor (number of times, in which a regression coefficient estimate’s variance is increased compare to the situation of no correlation among regressors)

3. Open file HPRICE\_1.dta in R (hint: use library(foreign)).

1. What is the percentage of houses in colonial style in the sample? ###% (round to integer)

2. What is the average price of houses built in colonial style? ### thousand dollars (round to integer)

3. What is the correlation coefficient between real and assessed price of houses (price and assess)? ### 0.91 (round to 2 decimals)

4. Find the correspondence between variables and their ranks in a list sorted by correlation with house price (in ascending order). Write down variable name (colonial, sqrtf, lotsize or bdrms) corresponding to each of the places.

1 (strongest correlation)

###

2

###

3

###

4 (weakest correlatio)

###

5. Create variable sqrmeter that equals the area of the house in square meters (assuming that 1 foot= exactly 0.3 meters) Be as precise as possible when doing this conversion. What is the average area of houses in the sample? ### square meters (round to integer)

6. Create variable lotsize\_sqm that equals lot size in square meters. Be as precise as possible when doing the conversion from square feet to square meters. What is the average lot area in our sample? ### square meters (round to integer)

7. Build a regression of log price on binary variable colonial. Enter + if the statement is TRUE and – of FALSE:

On average, colonial houses cost 18 higher than non-colonial. ###

Even though R/Stata print just a few decimals, in fact p-values of the F-test for overall significance of the model and of the t-test for the significance of the slope are exactly equal. +

8. Print + if TRUE and - if FALSE:

The elasticity of house price by house area does not depend on whether the area is measured in square meters or square feet ###

9. The elasticity of house price by house area (when using the model with constant elasticity of price by area) is ### (round to 2 decimals)

10. What is the number of houses with 3 bedrooms in the sample? ### What is the number of houses with 4 bedrooms in the sample? ###

11. Print + if TRUE and - if FALSE:

In a regression of the log-transformed house price on the colonial style binary indicator, the coefficient of the regressor is statistically significant at the 5% significance level, but not at the 10% level ###

In a regression of the log-transformed house price on the colonial style binary indicator, the coefficient of the regressor is statistically significant at the 10% significance level, but not at the 9% level ###

4.Under perfect multicollinearity the relationship among explanatory variables is not necessarily functional (i.e. deterministic).

( )False

( )True

5. Spearman’s rank correlation between an explanatory variable x and squared residuals e2 is not useful for diagnosing:

* Standard errors of regression coefficients
* Heteroscedasticity
* Multicollinearity
* Autocorrelation of residuals

6. We study the relationship between coffee consumption and several factors: x1 – coffee brand, x2 – coffee strength (strong, medium, weak), x3 – consumer’s income, x4 – coffee price. The researcher wants to correctly account for nominal and ordinal variable. The following variable(s) are ordinal in the model:

[ ]X1 ← Неверный ответ

[ ]X3

[ ]X4

[ ]X2 ← Правильный ответ

7. Peter Kennedy in his book "A Guide to Econometrics" gives the following empirical rule: you don’t’have to fix multicollineaity if the R2 of your model exceeds the maximum R2 of regressions of an explanatory variable on all other explanatory variables. Thus if R2 of the model is 0.9 then you don’t have to worry about multicollinearity if all VIFs do not exceed ### (if the answer is integer, then enter it as integer, if not – round to 1 decimal).

**Annex 3**

**Recommendations for students:**

1. Students are expected to install R and R Studio on their personal desktops/laptops so as to practice their coding skills.

2. Students should keep in mind that lectures contain not only theoretical material, but also solutions of problems. We do not train for tests on seminars. To successfully master the material attending both lectures and seminars is essential.

3. In order to prepare for the exam, study all problems given on lectures and seminars – exam problems will be similar to them.

4. Most of the coursework is scored automatically in LMS. Students should keep in mind that “.” and “,” can be used interchangeably there, but figures should be rounded as requested in each problem.

5. When students have been given access to DataCamp, they are encouraged to take not only courses they are required to take, but also at least 3 extra courses that will help acquire skills useful for the job market.