EC282: Introduction to Econometrics

Onur Altındağ

Last update: 2020-10-30

Contents

Syllabus

Course description

This course aims to introduce 21st century econometric analysis to business students. It provides tools to infer meaningful information from data using descriptive and regression analyses. In the first half of the semester, we will review the basic statistics used in econometrics and introduce mechanics of univariate and multivariate regressions. In the second half, we focus on causal interpretation of regression results, measures of fit, choice of functional form, multicollinearity and issues related to overfitting a prediction model and how to fix them.

At the end of the semester, I expect you to be familiar with **R** and **RStudio** interface, basic data manipulation, obtaining and interpreting sample statistics, conduct meaningful regression analysis and prediction. Importantly, I expect you to have a clear understanding of the distinction between correlation and causation, and the conditions in which the former implies the latter.

Knowledge and Skills

- Compute and interpret the descriptive statistics of a sample.
- Understand the statistical uncertainty, construct and interpret the confidence intervals.
- Conduct hypothesis testing, interpret the test statistic and the results of a statistical test.
- Construct a multivariate regression model, empirically estimate the model and interpret the results.
- Basic understanding of the randomized controlled trials and the causal inference.
- Choose and modify the functional form of a relationship between the output and the input variables.
- Interpret the regression coefficients on models with interaction variables.
- Understand the concept of overfitting and the difference between in-sample and out-of-sample performance of a prediction model.

Perspectives

- Learn how to conduct a regression analysis, understands its limitation in inferring a causal relationship, generalize its results, and the prediction of an outcome that is unknown to the researcher.
- Understand the regression diagnostics to choose the most appropriate definition of predictors, outcome, functional form, and regressor.

Class Information

Contact

- **Instructor**: Onur Altındağ
- **Course web**: https://ronuraltindag.github.io/metrics/
- **Personal web** : www.onuraltindag.info
- **Office**: AAC 181 - *currently NA due to pandemic*
- **Email**: `oaltindag@bentley.edu`

Office hours

Please go to my calendar and book a virtual office hour to meet me (30 minutes maximum). Email me if you need to talk to me urgently or there is no availability on my calendar.

Textbook

- Stock, J. H., & Watson, M. W. Introduction to Econometrics, any edition. Pearson Education Limited. – **required**
- E-book: https://www.econometrics-with-r.org/ – **optional**

Important Dates

- Weekly homework assignments: Indicated on this web page, subject to change depending on our pace.
- Midterm exam due: **Oct 15, 2020, midnight**
- Empirical project progress report due: **Nov 1, 2020, midnight**
- Empirical project final : **Dec 1, midnight**
- Final exam: **Dec 17, midnight**

Important note: All exams and the final version of the empirical project should be submitted through Black Board. All homework assignments and the "progress report" on your project should be posted under the designated **GitHub** issue.

Evaluation

- Midterm exam: 20%
- Empirical project: 20%
- Final exam: 30%
- Assignments: 20%
- Participation: 10%

Software and Collaborative Work

- **R** and **RStudio**: I assume that you have a basic familiarity with or expect your effort to gain familiarity throughout the semester. The instructions installation, some basic rules and best practices on coding are on this web page. Keep in mind that this course is **not** designed to teach you R and more than anything, the best way to learn programming is to actually work on assigned problems.
- **Github**: To create a collaborative and interactive teaching environment, you need to create an account on GitHub using your Bentley email address and accept the project invitation that you will receive from me for EC282. You will only use the very basic tools on GitHub, mainly issues tab to share your empirical work in progress, ask questions about the homework assignments, post an answer, and learn R from me and your peers through sharing your code.

Grading

- **High-stake assessments**
  - **Midterm** + **Final**: Constitute half of your final grade. I will post the exams on **Black Board** and you will have **24 hours** to work and submit. **DO NOT** try to submit the exams on last minute as the system will close after the deadline and I will not accept it. You **MUST** attend the midterm and the final as there will

be no make-up exams. The midterm and the final are **both** cumulative. If you miss or are likely to miss the **midterm** due to an emergency, please contact me as soon as possible. You will need to provide supporting documentation/verification of your absence. I will re-weight your final exam if you have a valid excuse. If you miss the final exam due to an emergency, you will receive an **incomplete** for this course. **DO NOT** take this class if you know that you will not be able to attend the final exam.

- **Low-stake assessments**
  - **Empirical project:** a short empirical essay that you will develop throughout the semester and execute by using the tools that you learn in this class. On Github, I will suggest some topics that are mainly in my area of interest: health, migration, inequality, political economy, etc. If you don't like any of those, you can also choose your own. You will need to find data, clean and organize it, conduct a small scale econometrics analysis to answer an interesting empirical question. The final output should be between 1500-2500 words and 3-4 graphs/tables combined, and submitted through BlackBoard. Your interaction with me and your peers is a significant part of your grade for preparing the empirical assignment. So the process matters as much as the final output. Here is a blog post that I wrote in spring 2020 which was covered by New York Times, Financial Times, Euro News, El Pais, and in addition to many Turkish media outlets. While I don't expect you to conduct an analysis like a seasoned econometrician, I do want you to do provide mini empirical-investigation and report your findings in a similar format.
  - **Bi-weekly homework assignments:** The homework assignments are posted on this web page with the deadlines. They can be completed in groups of **maximum two** students but each person should post separate answers under the designated **Github** issue.You need to post your work on **GitHub** before Monday meeting after the deadline. I will go over these problems in the classroom and randomly ask students to "help" me with the assignment. If what you know substantially contradicts with what you posted online, I will change your grade to zero and you will receive a significant penalty on your participation grade. In other words, do the best you can with these assignments, work consistently, do not free ride on your friends, and do not cheat. The data sets that each of you will receive are different so I will not tolerate if I see aby copied/pasted answers on GitHub.
  - **Collaborative participation to GitHub and classroom discussions:** You must sign up for a free account on GitHub. Github is an eco-system for web development and version control using Git. You will only need to use the `issues` tab through either creating an issue to ask or answer a question on your or your peer's empirical analysis, homework assignment, or anything related to econometric analysis. I expect you to actively participate to the discussion on GitHub as it will determine your participation grade. Both asking and answering a question in a meaningful way contributes to your participation grade. Your interaction on **Github** through receiving feedback from me and your friends on your empirical project is also part of your empirical assignment grade. To sum, you expect you to actively participate to the online community discussions on GitHub. I will do my best to facilitate the discussion yet I need your active support to make this environment useful for all.

Academic Integrity

Learning is a privilege that demands responsibility. At Bentley, students and faculty are members of an academic community that supports integrity both inside and outside the classroom. The expectation at Bentley is that students will take advantage of the opportunity for intellectual development and, in doing so, will conduct themselves in a manner consistent with the standards of academic integrity. When these standards are violated or compromised, individuals and the entire Bentley community suffer. Students who engage in acts of academic dishonesty not only face university censure but also may harm their future educational and employment opportunities. In other words, don't bring unauthorized materials into exams, don't plagiarize someone else's work, and make sure that your collaborations are conducted in accordance with university and course policy.

All students have access to Bentley's academic integrity policy on Blackboard (via the Academic Integrity course page) and the Undergraduate Student Handbook/Graduate Catalog. The best way to avoid a problem is to consult with your instructor before taking any action that might constitute a violation.

Diversity Inclusion and Support

Statement of Diversity and Inclusion

My goal in this class is to create a teaching environment that is inclusive for all of the members of our small community independent of their race, gender, age, disability status, and political or religious views. Our differences strengthen our ability for perspective taking, being critical about our default believes,and enhance learning.

I will try to reach this goal within my best capacity by respect and professionalism in our class-related engagements and I anticipate students to do the same. These standards of appropriate conduct are well summarized by Bentley's Core Values in our institution's mission statement. If you feel that I or anyone in this class has acted outside these values, please come to me so that we may discuss your experience. If you do not feel comfortable coming to me with your concerns, I encourage you to speak with someone in the Office of Academic Advising: 781.891.2803, academic_services@bentley.edu, Jennison 336.

Bias Incident Response

The Bias Incident Response Team (BIRT) provides students affected by bias or bias-related incidents with access to appropriate resources. Where appropriate, BIRT assists the University in its response to situations that may impact the overall campus climate related to diversity and inclusion. Working closely with appropriate students, faculty, committees, organizations, and staff, BIRT plays an educational role in fostering an inclusive campus community and supporting targeted individuals when bias or bias-related incidents occur. More information about BIRT and how to file a bias incident report can be found at: https://www.bentley.edu/offices/student-affairs/birt

Disability Services

Bentley University abides by Section 504 of the Rehabilitation Act of 1973 and the Americans with Disabilities Act of 1990 which stipulate no student shall be denied the benefits of an education solely by reason of a disability. If you have a hidden or visible disability which may require classroom accommodations, please call (if you are a residential student or on online student) Disability Services within the first 4 weeks of the semester to schedule an appointment. Disability Services is located in the Office of Academic Services (JEN 336, 781.891.2004). Disability Services is responsible for managing accommodations and services for all students with disabilities.

The Undergraduate Academic Services (UAS) Peer Tutoring program offers online one-on-one and small group tutoring services for students who have worked with their instructors and made use of the Learning Centers, but still require additional academic support. The program goal is to help those students in true need who are willing to take responsibility for their own learning. Please reach out to me if you need more information.

LEAF Tutoring Student Center

This is a new center and I will post more information here when I hear from them.

Online Attendance

Zoom Protocol and Online Attendance

Students **must** join classes through their Bentley Zoom account. Go to bentley.zoom.us and enter the course meeting number to join the session.

I expect you to attend class with a functioning microphone and camera. Cameras should be on to effectively engage in class and participate throughout the course. If you have an impediment to keeping your camera on, please let me know so that we can work to arrive at a mutually agreeable solution.

You are expected to be able to access all electronic course materials. It is your responsibility to review the course syllabus as soon as possible to determine what resources or materials I expect you to use in the course. If you are a student in an international location that may limit access to certain internet resources, please let me know immediately so you can find a solution.

Students are expected to attend classes synchronously despite potential time zone hurdles. Solely watching recorded classes is not deemed to be acceptable course participation or completion. Course recordings are for the benefit of students who miss an occasional class or would like to watch the recording for further edification of materials. Class recordings that are posted to BB are for the sole purpose of this course. Disseminating any portion of this video in any manner is strictly prohibited.

Lecture videos

I will record and post the lecture videos on Black Board.

Lecture Notes

During the online lectures, I will use my iPad as a white board to teach. I will post these notes on this web page throughout the semester as well as the notes from the previous semester.

- Here are my hand-written lecture notes from the previous time that I thought this course.
- Here are the lecture notes from the current semester.

Tentative Schedule

The key readings from Stock and Watson are indicated for each week.

Week 1-2
Aug 31, Sep 3, Sep 10

- Introduction to the course, logistics, syllabus, expectations and pap-talk.
- Review of Probability (Chapter 2)
  - Random sampling and the Distribution of the Sample Average
  - Large-Sample Approximations to Sampling Distributions

Week 3
Sep 14, Sep 17

- Review of Statistics (Chapter 3)
  - Hypothesis Tests Concerning the Population Mean
  - Confidence Intervals for the Population Mean

Week 4
Sep 21, Sep 24

- Review of Statistics (Chapter 3)
  - Comparing Means from Different Populations
  - Scatterplots, the Sample Covariance, and the Sample Correlation

Week 5
Sep 28, Oct 1

- Linear Regression with One Regressor (Chapter 4)
  - The Linear Regression Model
  - Estimating the Coefficients of the Linear Regression Model
  - Measures of Fit and Prediction Accuracy

Week 6
Oct 5, Oct 8

- Linear Regression with One Regressor (Chapter 4)
  - The Least Squares Assumptions for Causal Inference
  - The Sampling Distributions of the OLS Estimators

Week 7
Oct 12, Oct 15

- Wrap up, Midterm Review and **Midterm exam**
- Deadline to post your progress report on GitHub. One or two paragraphs and a table/figure that summarizes what you have done so far

Week 8
Oct 19, Oct 22

- Regression with a Single Regressor: Hypothesis Tests and Confidence Intervals (Chapter 5)
  - Testing Hypotheses About One of the Regression Coefficients
  - Confidence Intervals for a Regression Coefficient
  - Regression when X is a Binary Variable

Week 9
Oct 26, Oct 29

- Linear Regression with Multiple Regressors (Chapter 6)
  - Omitted Variable Bias
  - The Multiple Regression Model
  - The OLS Estimator in Multiple Regression

Week 10
Nov 2, Nov 5

- Linear Regression with Multiple Regressors (Chapter 6) - Measures of Fit in Multiple Regression - The Least Squares Assumptions for Causal Inference in Multiple Regression

Week 11
Nov 9, Nov 12

- Linear Regression with Multiple Regressors (Chapter 6) - The Distribution of OLS Estimators in Multiple Regression
  - Multicollinearity
  - Control Variables and Conditional Mean Independence

Week 12
Nov 16, Nov 19

- Hypothesis Tests and Confidence Intervals in Multiple Regression (Chapter 7)
  - Hypothesis Tests and Confidence Intervals for a Single Coefficients
  - Tests of Joint Hypotheses
  - Testing Single Restrictions Involving Multiple Coefficients

Week 13
Nov 23

- Hypothesis Tests and Confidence Intervals in Multiple Regression (Chapter 7)
  - Model Specification for Multiple Regression

Week 14
Nov 30, Dec 3

- Nonlinear Regression Functions (Chapter 8)
  - A General Strategy for Modeling Nonlinear Regression Functions
  - Nonlinear Functions of a Single Independent Variable

Week 15
Dec 7, Dec 10

- Nonlinear Regression Functions (Chapter 8)
  - Interaction Between Independent Variables

Week 16
Dec 14, Dec 17
- Wrap up, last remarks, **Final Exam**.

Chapter 1

General Rules and Principles

1.1 Installing R and RStudio

Here are the instructions for installing R and RStudio on your Windows or Mac desktop. Skip the third part and do not install "SDSFoundations Package".

1.2 Basic rules and best practices

All files should exist in a local folder that syncs to a cloud-storage service. No file you ever work on should be at risk of being lost if your computer ceases to function or be in your possesion. NEVER place any file on "downloads" or "desktop" folders.

Get a free cloud-storage service with a desktop application that syncs to a cloud-storage service. I like the Dropbox desktop app but feel free to choose any other service. You don't need a lot of space so free version of any desktop cloud app would work. Under the Dropbox folder, create a designated folder for this course such as EC282.

All subfolders under EC282 and files in them should have unique and descriptive construction: DON'T use spaces in file or folder names. Here is an example of a folder structure that might work for a student in this class:

```
EC282


  Course_docs
        SyllabusEC282.pdf
        LectureNotes.pdf
  Assignments
      Assignment1
           dataset1name.Rda
           Lastname_Firstname_Assignment1_EC282.R
      Assignment2
           dataset2name.Rda
           Lastname_Firstname_Assignment1_EC282.R
      Assignment3
           dataset3name.Rda
           Lastname_Firstname_Assignment3_EC282.R
      Assignment4
           dataset4name.Rda
           Lastname_Firstname_Assignment4_EC282.R
           ...
  Exams
```

```
      Midterm1
            Midterm1Review.pdf
            Midterm1Review_myanswers.docx
|       |   ...
```

1.3 Header

At the beginning of any R script, you should have a standard header that you use across all scripts that clears the workspace, loads/installs packages as necessary, sets the working directory, etc. Here is an example that you can copy paste to the header of any script that you use:

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())

################################################################################
```

When coding, use relative references to files. Typically, any script will begin looking for files in the working directory. At any time you can type `getwd()` on your Rstudio console to see the current working directory. The header above automatically sets the working directory to the folder that the R script is included. For example, if you are working on `Lastname_Firstname_Assignment1_EC282.R` script and need to load file `dataset1name.Rda` into an object, then you would simply run:

```
load(dataset1name.Rda)
```

However, if you were working in the same .R file, and needed to access `dataset2name.Rda`, you would need to point the program to a directory outside the current working directory – so, you go up one level, over one folder, and look there:

```
load(../Assignment2/dataset2name.Rda)
```

When learning R, the most important skill that you need to acquire is to be able to **google** your problem. There is probably not a single R question that you have yet has not been answered on Stack Overflow.

Chapter 2

Homework Assignments

2.1 How to submit homework assignments

You can use a snipping tool to cut and paste the relevant output and figures from **RStudio** to the issue on **GitHub** under the assignment name. If you would like to have a more elegant looking output, I encourage you to learn more the `stargazer` package that transforms your analysis into publishable formats.

2.2 Assignment 1

Deadline: Sep 13, 2020, Midnight

Source: Stock and Watson, 4^{th} Edition, Exercise 2.1
Data Description: The data set includes the joint probability distribution of `age` and average hourly earnings `ahe` for 25- to 34-year-old full-time workers in 2015 with an education level that exceeds a high school diploma, i.e. $P(age = x, ahe = y)$

Questions

a. Compute the marginal distribution of age, i.e $P(age = x)$ in a new data set. **b.** Compute the mean of `ahe` for each value of `age`, i.e. compute $E[ahe|age = 25]$, $E[ahe|age = 26]$, etc. and plot these conditional expected values of `ahe` against `age` Are they related? Briefly comment.
c. Compute the variance of `ahe`.
d. Compute the covariance between `age` and `ahe`.
e. Compute the correlation between `age` and `ahe`.
f. Relate your answers in (d) and (e) to the plot that you constructed in (a)

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment1` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))
```

```
# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
###############################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/aieodljjthbdhks/Age_HourlyEarnings.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf, col_names = FALSE, skip=1)[2:30,1:11] %>%
  rename(ahe = 1) %>%
  mutate(ahe = as.numeric(ahe)) %>%
  gather(key = "age", value="jointp", -c("ahe")) %>%
  mutate(age =  as.numeric(gsub(".*?([0-9]+).*", "\\1", age)) + 23)



head(df1)

#CONDUCT THE ANALYSIS BELOW
```

2.3 Assignment 2

Deadline: Sep 27, 2020, Midnight

Source: Source: Stock and Watson, 4^{th} Edition, Exercise 3.1

Data description: You can find the data description here.

Questions

a. In 2015, the value of the Consumer Price Index (CPI) was 237.0. In 1996, the value of the CPI was 156.9. Create a new variable in your data frame that expressed all earnings in real 2015 dollars. Use this variable to answer the next questions.

b. Construct a 95% confidence interval for the mean of `ahe` for high school graduates in 1996.

c. Construct a 95% confidence interval for the mean of `ahe` for high school graduates in 2015.

d. Construct a 95% confidence interval for the mean of `ahe` for college graduates in 1996.

e. Construct a 95% confidence interval for the mean of `ahe` for college graduates in 2015.

f. Did the inflation adjusted wages of high school graduates increase from 1996 to 2015? Use statistical inference to answer.

g. Did the inflation adjusted wages of collage graduates increase from 1996 to 2015? Use statistical inference to answer.

h. Did the gap between earnings of college and high school graduates increase? Use statistical inference to answer.

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment2` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl',  'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication


options(scipen=999)
################################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/hbi82scuz9q4k11/CPS96_15.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf) %>%
  mutate(ahe = ahe + rnorm(length(ahe)))



head(df1)

#CONDUCT THE ANALYSIS BELOW
```

2.4 Assignment 3

Deadline: Oct 4, 2020, Midnight

Source: Stock and Watson, 4^{th} Edition, Exercise 4.1

Data description: You can find the data description here.

Questions

a. Construct a scatterplot of `growth` and `tradesshare` with a regression line fit on the top.

b. Look at the data set and find Malta on your graph. Why is Malta an outlier?

c. Using all the observations run a regression of `growth` on `tradeshare`. Interpret the intercept and the slope. Predict the growth rate for a country with a trade share of 0.5 and another with a trade share equal to 1.

d. Estimate the regression without Malta and interpret the coefficients. Should Malta be excluded from the regression? Briefly comment.

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment3` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
################################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/lbk73b0amzfj8px/Growth.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked

df1 <- read_excel(tdf) %>%
  mutate(growth = growth + rnorm(length(growth))/5)



head(df1)

#CONDUCT THE ANALYSIS BELOW
```

2.5 Assignment 4

Deadline: Oct 18, 2020, Midnight

Source: Stock and Watson, 4^{th} Edition, Exercise 5.3

Data description: You can find the data description here.

Questions

a. Run a regression of `birthweight` on `age`. Interpret the coefficient on age. Is the coefficients statistically significant?

b. Estimate the mean and the standard error of birth weight for (i) mother who smoked during the pregnancy and (ii) mother who did not smoke during the pregnancy.

c. Estimate the difference between (i) and (ii). Construct a 95% confidence interval for the difference in the average `birthweight` for smoking and nonsmoking mothers.

d. Run a regression of `birthweight` on on the binary variable `smoker` explain how the estimated intercept, slope related to your previous answers. How about the standard error of $\hat{\beta}_1$?

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment4` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your enviroment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl',  'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
################################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/z8r6hc0r4ytt4f8/birthweight_smoking.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf) %>%
  mutate(birthweight = birthweight + rnorm(length(birthweight)) * 50)



head(df1)

#CONDUCT THE ANALYSIS BELOW
```

2.6 Assignment 5

Deadline: Nov 8, 2020, Midnight

Source: Stock and Watson, 4^{th} Edition, Exercise 6.1

Data description: You can find the data description here.

Questions

a. Regress (i) `birthweight` on `smoker` and (ii) `birthweight` on `smoker`, `alcohol`, and `nprevist` . Compare the estimated coefficient on `smoker` in (i) and (ii). Does the regression suffer from omitted variable bias?

b. Predict the birthweight for a child whose mother smoked during the pregnancy, did not drink alcohol, and had 8 prenatal care visits.

c. Compare the R^2 and adjusted-R^2 from (ii), why are they so similar?

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment5` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
################################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/z8r6hc0r4ytt4f8/birthweight_smoking.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf) %>%
    mutate(birthweight = birthweight + rnorm(length(birthweight)) * 50)



head(df1)

#CONDUCT THE ANALYSIS BELOW
```

2.7 Assignment 6

Deadline: Nov 14, 2020, Midnight

Source: Stock and Watson, 4^{th} Edition, Exercise 7.1

Data description: You can find the data description here.

Questions

a. Regress (i) `birthweight` on `smoker`, `alcohol`,`nprevist`, and `unmarried`. Interpret the coefficient on `unmarried`.

b. Construct a 95% confidence interval on for the coefficient. Is it statistically significant? Is the magnitude of the coefficient large?

c. Looking at this regression, a family advocacy group claims that higher rates of marriage will lead to healthier babies thus one obviuous public policy is to encourage marriage. Do you agree?

d. Consider the data set that you have and briefly discuss what variables can be added to the regression to help to solve question (c).

e. Run the regression with these additional controls. How did the coefficient on marriage has changed with these additional controls.

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment6` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

```
###############################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
###############################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/z8r6hc0r4ytt4f8/birthweight_smoking.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf) %>%
    mutate(birthweight = birthweight + rnorm(length(birthweight)) * 50)



head(df1)

#CONDUCT THE ANALYSIS BELOW
```

2.8 Assignment 7

Deadline: Nov 29, 2020, Midnight

Source: Stock and Watson, 4^{th} Edition, Exercise 8.1

Data description: You can find the data description here.

Questions

a. Using a regression, show the average infant mortality rate `infrate` for for cities with lead pipes and for cities with nonlead pipes. Is there a statistically significant different in the averages?

b. Amount of lead leached from lead pipes depends on the chemistry of the water running through the pipes. Lower the `ph`, more asidic the water is and the more lead is leached. Run a regression of `infrate` on `lead`, `ph`, and the interaction term `lead` *times* `ph`

c. Does `lead` have a statistically significant effect on infant mortality? Explain.

d. Does the effect of `lead` on infant mortality depend on `ph`? Is this dependence statistically significant?

e. Construct a 95% confidence interval for the effect of `lead` on infant mortality when `ph` is 6.5

Header for the R script

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Assignment7` or a similar path that you created for this homework assignment. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

```
################################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
################################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/5nszqnejl7uu9f5/lead_mortality.xlsx?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".xlsx")))
#check if it worked
df1 <- read_excel(tdf) %>%
  mutate(infrate1 = infrate + rnorm(length(infrate))/50)
```

```r
head(df1)

#CONDUCT THE ANALYSIS BELOW
```

Chapter 3

Replication Project

3.1 Instructions

a. Carefully read the study and provide a two paragraph summary that includes:

- Motivation of the study, why is it important?
- Data and empirical approach. How do researchers establish a causal inference? What data and method do they use? What do they estimate?
- Main findings and the implications of the study.

b. Run the following regression for higher and lower quality CVs separately as well as the full sample of CVs. Prepare a regression table of the results and interpret these results within the context of the study.

$call_i = \beta_0 + \beta_1 black_i + u_i$

where *black* is a dummy variable that equals one for distinctly African-American names.

c. Run separate linear regressions using `call` as an outcome and `yearsexp`, `volunteer`, `military`, `email`, `empholes`, `workingschool`, `honors`, `computerskills`, and `specialskills`for CVs with African-American and White names. Prepare a regression table and interpret the results within the context of the study. Write up the results that you find interesting and related to the research question in the study, you don't need to interpret each coefficient.

d. Conduct an additional analysis using the data that you have that offer any interesting additional findings to the reported ones in the published study. This could be a subgroup analysis, a graphical representation of a finding in the study or any empirical exploration that you find interesting and use the tools that we learned in the classroom.

e. Conclude your replication study by 1-2 paragraphs of discussion of the results above. What do these ressults imply? What are the policy implications?

3.2 Files: Data, codebook, study

Start a new R script, copy/paste the header below and save it to `Dropbox\EC282\Empirical_Project` or a similar path that you created for this course. Run the R script and make sure that you have the data `df1` in your environment. Conduct the analysis below the header.

- You can find the data codebook here.
- You can find the original study here

```
##############################################################################
# list the packages we need and loads them, installs them automatically if we don't have them
# add any package that you need to the list
need <- c('glue', 'dplyr','readxl', 'ggplot2','tidyr','AER','scales','mvtnorm',
          'stargazer','httr', 'repmis')

have <- need %in% rownames(installed.packages())
if(any(!have)) install.packages(need[!have])
invisible(lapply(need, library, character.only=T))

# Save the R script to the assignment 1 folder before this
# To set up the working directory
getwd()
setwd(getwd()) #change getwd() here is you need to set a different working directory


#this clears the workspace
rm(list = ls())
#this sets the random number generator seed to my birthday for replication

options(scipen=999)
##############################################################################
#get the data url
df1.url <- 'https://www.dropbox.com/s/oekcvjgy4yfjv0e/lakisha_aer.dta?dl=1'
#download the data
GET(df1.url, write_disk(tdf <- tempfile(fileext = ".dta")))
#check if it worked
df1 <- read_dta(tdf)

head(df1)

#CONDUCT THE ANALYSIS BELOW
```