Problem Set 1:Getting Started Key

Claire Duquennois

Group Member 1: Your name here

Group Member 2: Your name here

Group Member 3: Your name here

In additions, to the instructions given on the problem set, you may find the "GitHub instructions" course document helpful.

1 Preliminary: Set up your GitHub, RStudio Cloud and Git.

Git and GitHub are often used in conjunction but are really two distinct things. Git is basically a version control software that is based on your local computer. GitHub is a web service where you have an account that allows you to do version control via the website, which becomes particularly interesting when you are working on collaborative projects and want to avoid the issue of how to synchronize and manage changes to a document when multiple people are working on the same document simultaneously. In theory you could use only one or the other but many people use them in conjunction.

Since in this class we are primarily working on RStudio Cloud (rather than on your local machine), I will walk you through the RStudio Cloud to GitHub integration. I will also provide information on using Git so you know how to sync files between github and your local machine as well, if you choose to do so. Conceptually the relationship between Github and RStudio Cloud is very much the same as how you would use GitHub with a local folder on your computer, though the exact steps involved in setup are a bit different.

To get a big picture overview of exactly what you can do with Git and Git hub, I would recommend taking the time to watch the Git and GitHub for Poets series of youtube videos https://www.youtube.com/playlist?list=PLRqwX-V7Uu6ZF9C0YMKuns9sLDzK6zoiV.

1.1 Getting Started with RStudio Cloud

I am working on the assumption that you have already been working in RStudio Cloud and do not need instruction on this.

1.2 Getting Started with GitHub

This is quite straightforward:

1. Go to https://github.com/.
2. Click Sign up.
3. Choose a username, an email that GitHub will use to contact you and set you password
4. Respond to the invitation email to activate your account.

1.3 Getting Started with Git (Optional)

Git is often used in conjunction with GitHub as a way to work on GitHub projects locally offline and then reintegrate your work back into a GitHub project. While there is a visual interface you can use to run Git, most programmers seem to prefer to run it directly by typing in Git commands through a terminal prompt.

On MACs: You can type Git commands directly into your Terminal (you can find it in Application/Utilities).

On Windows: You will want to install Git Bash. (See the following link for a video of the installation and some examples of commands https://www.youtube.com/watch?v=J_Clau1bYco.)

1. Go to git-scm.com and download the appropriate .exe file for your operating system.
2. Run the .exe file and install Git. In the installation, the only default you may want to change is to select the option to "Use Git and optional Unix tools from the Windows Command Prompt".

3. Once installed, launch Git Bash and you will see a terminal window in which you can execute Git commands.

You can use the terminal window to navigate through the folders in your computer's directory and then, using the proper commands, you will be able to use this terminal to push and pull materials from your computer onto github and vice versa. This will allow you to work on a github project on your local machine without having to be constantly online.

2 Individual Homework Tasks:

The following list of tasks are designed to get you started with GitHub and RStudio Cloud and get you set up with the course resources. Everyone in the group should complete these tasks individually.

2.1 Clone the course repository to your RStudio Cloud

All of the lecture materials for this course are available in a repository on GitHub. Navigate to https://github.com/claireduq/MQE_Causal to view the course repository. This is where you will find the most up-to-date content for the course.

Step by step instructions are detailed below to help you do the following:

1. Fork the course repository to your GitHub account
2. Clone this forked repository to your RStudio Cloud account
3. Upload the course data to RStudio Cloud
4. Knit an Rmarkdown file

2.1.1 Fork the course repository to your GitHub account

Let's copy the course repo onto your own GitHub page so that you can freely work on it.

To do this you want to fork (ie copy) the entire course repo

1. Navigate into the course repository
2. Click the **Fork** button in the upper right corner

You should now have a copy of the repo on your account.

If you would like to rename the repo:

1. Open the repo
2. Go to the settings tab
3. Type in a new name and click Rename.

2.1.2 Clone this forked repository to your RStudio Cloud account

So you can make modifications to the course code, experiment with the simulations we will be running, try different specification etc. you need to create your own copy of the course materials in your RStudio Cloud account. To do this:

1. Click on the New project button and select New project from Github in the drop down options.
2. Copy the URL of the MQE_Causal repository in the space provided

RStudio Cloud should now open a new project populated with the course files

2.1.3 Upload the course data to your RStudio Cloud account

The GitHub platform is designed to keep track of changes to code/projects. It is not designed to host large amounts of data. If you try and upload a large data file to github it will throw an error. Thus the data files that are used in the course materials are not hosted on github and will need to be uploaded to your project folder separately.

1. Go to the course Canvas page.
2. Copy the files called `cps_clean.csv` and `IA_MI_merge040504.dta` and the mod7data folder to your local machine.
3. In RStudio Cloud click on the `upload` button in the file tab of the SE panel of the work space.
4. Upload the two data files and the folder directly to your RStudio Cloud project (don't put them in a separate folder).

2.1.4 Run an Rmarkdown file

Course notes and slides are written in Rmarkdown (.Rmd). Your homework assignments will also be produced using Rmarkdown. Rmarkdown is a type of R file that allows the combination of R code, code output and text. For the moment, we just want to check that you have the setup needed to run the course files.

5. Some of the data sets used in lecture are large and will require a fair bit of computing power if you want them to run in a reasonable amount of time. To increase your RAM on RStudio Cloud:
    a. click on the settings button at the top of your browser window
    b. select resources
    c. increase the RAM to 16 GB
    d. Click apply changes (which will re-initialize your session)
6. Make sure your browser allows pop-up windows from rstudio.cloud
7. Open any of the .Rmd files and click the `Knit` button in the NW panel of your work space. Make sure to install any packages when prompted. A pop-up window should appear with the compiled lecture notes or slides. (Note: if the .Rmd file uses the large data in IA_MI_merge040504.dta it will take a few minutes to compile.)

2.1.5 Set up syncronizing with claireduq/MQE_Causal

To make sure the course folder on your RStudio Cloud reflects the most recent changes made in an upstream repo (my course repo that you originally forked from) you need to configure the folder on your computer so that you can easily `fetch` any changes in the upstream repo:

1. Open the terminal tab in the SW panel of your RStudio Cloud project
2. List the current configured remote repository for your project. Type `git remote -v` after `/cloud/project$`: `/cloud/project$ git remote -v`

    ```
    > origin  https://github.com/YOUR_USERNAME/YOUR_FORK.git (fetch)
    ```

    ```
    > origin  https://github.com/YOUR_USERNAME/YOUR_FORK.git (push)
    ```

    (This is telling you where your coputer will get any online changes from (fetch) and where it will send any local changes to (push). Right now, it is probably set to fetch changes from your repo and push changes to your repo. We want to change this so that it fetches updates from the my course repo (not yours).)

3. We want to change this. Specify a new remote upstream repo (my course repo) that will be synced with your fork

   ```
   /cloud/project$ git remote add upstream https://github.com/claireduq/MQE_Causal.git
   ```

4. Verify the new upstream repository you've specified for your fork.

   ```
   /cloud/project$ git remote -v
   ```

   ```
   > origin  https://github.com/YOUR_USERNAME/YOUR_FORK.git (fetch)
   ```

   ```
   > origin  https://github.com/YOUR_USERNAME/YOUR_FORK.git (push)
   ```

   ```
   > upstream  https://github.com/claireduq/MQE_Causal_Inf.git (fetch)
   ```

   ```
   > upstream  https://github.com/claireduq/MQE_Causal_Inf.git (push)
   ```

5. You can now `fetch` changes to the course repo and have them reflected on your local machine with the command

   ```
   /cloud/project$ git fetch upstream
   ```

6. Check out your fork's local master branch.

   ```
   $ git checkout master
   ```

7. Merge the changes from upstream/master course repo into your local master branch. This brings your fork's master branch into sync with the upstream repository, without losing your local changes.

   ```
   $ git merge upstream/master
   ```

   Notes:

   - Git may send you to a weird window asking you to write a commit message. You can escape this with `Shift+:` followed by Q.

You can repeat steps 5-7 anytime you want to make sure your local folder is up-to-date with my repository.

2.1.6 Optional: Clone and sync this forked repository to your local computer

If you want to have access to the course files when you are offline, you will need to clone (ie copy) the files in this repo to your local machine:

1. Make an empty folder named, for example, `MQE_Causal` where you will put the files you are going to pull off GitHub (preferably in Dropbox or someplace where it will be backed up).
2. Open your command terminal (or Git Bash on Windows).
3. Navigate via your command terminal to the new folder by typing

   ```
   $ cd /c/Users/ *******/Dropbox/MQE_Causal
   ```

   followed by enter.

   Note: the directory path will be different for you! To easily get your directory path, you can simply drag and drop the folder into your terminal window.

4. Check that you are in the right folder by typing

   ```
   $ pwd
   ```

   and you should see the directory path to the folder you want to be using.

5. Suppose your GitHub user id is `myuserid` and your repo you would like to clone is named `MQE_Causal_Inf`. In your github repo, click the green code button and copy the link it gives you. Most likely it will look like `https://github.com/myuserid/MQE_Causal.git` (Notice, the url is just the web address with .git added to the end.)

6. In the terminal type `git clone` and copy the url it gave you

   ```
   $ git clone https://github.com/myuserid/MQE_Causal.git
   ```

   followed by enter.

You should now find that the folder you created has a copy of all the files that are in the course repo.

As with RStudio Cloud, you can use Git to set up synchronizing between an upstream repo and your local folder

1. Go to the terminal (or Gitbash) window and navigate to the git folder by typing

   ```
   $ cd /c/Users/ *******/filepath
   ```

2. The remaining commands are the same as above.

2.2 Propose an edit to a course document

There are two ways to do this. You could do this directly as a **pull request** on the GitHub platform, or you could write the edit in RStudio Cloud, push it to your forked repository, and then submit it as a **pull request** to my repository. Though more complicated, we will practice the second method because ultimately you will be coding in RStudio Cloud and then merging your changes into your group project repository so we want to practice these steps here.

In the course repository there is a document called `GitHub_names.Rmd`. I would like you to add your first name and username to this document by:

1. Making the edits on to the document in RStudio Cloud
2. Pushing your changes to your GitHub repository
3. Submitting a pull request to my repository asking me to accept your edits.

2.2.1 Making the edits in RStudio Cloud

Open the `GitHub_names.Rmd` file in the folder you cloned to your work space. Add your First name and GitHub user name to the list and knit the document.

2.2.2 Pushing your changes to your GitHub repository

We now want to have these edits reflected in your GitHub repository.

1. At this point you want to check if GitHub knows who you are. Open the terminal tab in the SW panel of your RStudio Cloud project. Type

   ```
   /cloud/project$ git config --list
   ```

   and it will spit out a bunch on information. Look and see if your GitHub username and the email you used to sign on to GitHub with are listed. If they are all is well. If they are not listed type

   ```
   /cloud/project$ git config --global user.name "janedoe"
   ```

   ```
   /cloud/project$ git config --global user.email janedoe@pitt.edu
   ```

   Make sure you set these to your github user name and the email you used to sign up with GitHub.

2. Check the status of your edits. Type

   ```
   /cloud/project$ git status
   ```

   You should see:

   - if you have modified files with edits that have not yet been commited.
   - if you have any new files that do not exist in the GitHub repo

3. If you generated a .pdf document when you knit the .Rmd file, delete it. This means you only have to

3. If you have new files type

   `/cloud/project$ git add .`

   to add all new files to GitHub

4. To prepare your modifications to be committed, type

   `/cloud/project$ git commit -a -m "Jane Doe's changes"`

   where the -a argument tells git to commit all the modifications and -m adds a message to you commit (a good practice when working on jointly edited files). Now if you type

   `/cloud/project$ git status`

   you will see that the files you changed are ready to be pushed to github so that they are reflected there.

5. Now you can type

   `/cloud/project$ git push origin master` or just `/cloud/project$ git push`

   which basically tells git to push your changes from your remote "origin" computer to the "master" branch on github.

6. You can now look at your repository on github and check that the edits you made appear in your files and the files history.

Note: There are also ways to do this the point and click way in the Git tab in the NE panel.

BUT WHAT ABOUT THE DATA FILES!?! If you loaded the data sets to your work space then how come the data files didn't also get copied onto you GitHub page? If you click on the .gitignore file, you will see that I specified two file types (.dta and .csv) and a folder (mod7data/) that are never to be included in the synchronizing process. If we work with other large files I will add these as well. Without this, you would have been faced with an error.

2.2.3 Submitting a pull request to my repository asking me to accept your edits.

Now that your edits appear in your GitHub repository, you need to propose those changes to the owner of the original repo (me). You will do this via a pull request. Note: This will generate a pull request for ALL of the commits (changes) you made to any file in your forked repo since you originally forked it.

1. Navigate to your forked repo
2. Click the **Pull requests** tab.
3. Click the green **New pull request** button
4. Comparing change: you will now be shown a page that shows you how your edits fit into the existing repo. Check to see the status of your edits:

   a. If it says "Able to merge":

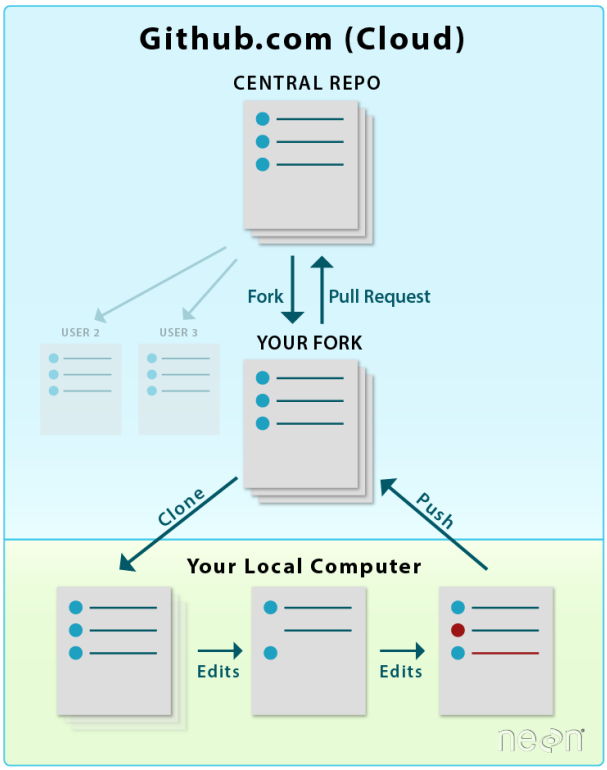
   - Click the green button **Create pull request**
   - In the pull request window, add a description/name for your edits in your pull request
   - Click **Create pull request** at the bottom of the page You have now sent me a pull request. I will review you edits and decide if I want to integrate them into my file.

   b. If it says "Can't automatically merge":

   - This means that your edits conflict with some other changes in the file. You can still create a pull request but I may need to make some modifications before integrating your changes.
   - Click the green button **Create pull request**
   - add a description/name for your edits

- Click the green **Create pull request** button at the bottom of the page You have now sent me a pull request. I will need to review your edits, probably have to make some changes since there is a merge conflict, and decide if I want to integrate them into my file.

Congratulations! You have now practiced most of the key actions needed to work collaboratively with GitHub. The way group projects are run on this platform is illustrated in the image below. Your group will have a central repository in which you are all collaborators. Each group member (other then the one hosting the central repository) will have their own fork (copy) of the repository on their own GitHub account. This copy will be cloned to your RStudio Cloud work space and then set up so that it syncs with the upstream central repo. When you want to work on your group assignment you will sync your RStudio Cloud folder with the central repo so that any changes your group mates made are reflected on your local machine. You can then do your coding and work on your local machine, and once you know that your code runs, you will push your changes to your fork, and then submit a pull request and merge you change to the central repo. For more detailed step by step instructions, and instructions on some other actions you may need to use, check out the GitHub_instructions document in the course folder.



3 Group Homework Project

3.1 Setup

On GitHub, have one group member generate a private fork of the problem set's project repository for your group named `PS1_name1_name2_name3` where the usernames are the first names of your group members. You will then add collaborators to your repository who will create their own forks of the repository.

3.1.1 Group "leader": To set up a private repo:

1. Log onto you GitHub account
2. Click on the **Repositories** tab

3. Make sure you accepted my email invitation to the private problem set repo
4. For the problem set repo and name you repository "PS1_name1_name2_name3"

Congratulations! You have created your repository! A couple things to note: Private repos will not show up on your main home site. To navigate to them you need to go to the Repository tab.

3.1.2 Group "leader": Adding collaborators

1. Navigate into the private PS1_name1_name2_name3 repo
2. Click on the **Settings** tab
3. Click on the **Manage access** tab in the left menu
4. Click on the green **Invite collaborator** button
5. Enter your teammates username in the popup box, select the correct user and invite them
6. Repeat for all of your teammates, our course TA and I

3.1.3 Group team members: Becoming a collaborator on a private repo

1. Once invited, you will receive an invitation via email to collaborate on the repo (it could be in the updates tab in gmail).
2. Click view the invitation to be sent to the repo
3. Once in the repo, click the green **Accept invitation**
4. You can now view the private repo your leader set up.

3.1.4 Group team members: Fork the private repo and load data

1. generate a fork of the repo (see instructions above).
2. Clone the repo to your RStudio Cloud (see instructions above).
3. Download the problem set data from canvas and load it into your cloud work space. Note: for your code to work when your teammates run it, make sure you save the data in the same location in your work spaces.
4. Work on your contribution to the code and answers in your work space
5. Make sure your edits run by knitting the .Rmd file
6. Once you know the .Rmd file runs, delete the generated .pdf (or other) file(s). This makes it easier to merge changes since you only have to do it in the .Rmd file and not the others as well.
7. Push your changes to your GitHub repo (see instructions above).

If you are a team member:

Submit a pull request to incorporate your contribution into the main project repository **AND** Accept/Merge the edits you proposed your repo into the main project repository (Since you are collaborators on this repo you can authorize and merge pull requests into the main folder, you do not need the repo owner (group leader) to do this.) You can do this by

a. Navigate to your forked repository that contains the changes you want to integrate into the main repo
b. Click on the **pull request** tab
c. Click the green **New pull request** button
d. Click on the green **Create pull request** button
e. Name your pull request
f. Click on the green **Create pull request** button
g. From here there are two possibilities:

- If the pull request is able to Merge:
  - If you want to see the detail of what edits were made, click on pull request name to look over the changes and then navigate back to the main pull request page
  - Click the green **Merge pull request** button
  - Click the green **confirm merge**
  - Your changes have now been added to the main file

- If there are conflict in the pull request:
  - There will be a warning sign informing you of a conflict.
  - Click the **resolve conflicts** button
  - GitHub will open an editable window that contains your file in which the conflicting part(s) of the file is highlighted by a red bracket. Within these brackets the proposed changes are bracketed by the lines [«««< master] and [=======] and what is currently in your file is bracket by [=======] and [»»»> master]. With all this information, you can edit your file within this window until you are happy with it.
  - Click the **Mark as resolved** button (upper right of the file window)
  - Click the green **commit merge** button
  - Click the **I understand, update master** button in the pop up window
  - Click the green **Merge pull request** button
  - Click the green **Confirm merge** button
  - Your changes have now been added to the main file

Note: If you are the group leader, things could be changing in your repo without you realizing it! Keep an eye on your repo history so you know what happened while you were away. You may also want to be cautious when pulling repo changes to your local machine since there may be changes you didn't know about. (Maybe keep a back up of your own work on your local computer for your records?)

7. Next time you want to work on the assignment, make sure you synchronize your forked repo with the group leaders repo so that you are working on the most up to date version of your group's code (see instructions above). `$git pull` any updates to your work space.

8. Once the main group .Rmd file is finalized, you can compile it to html, pdf or word for submission.

Hints: I would recommend only keeping track of the .Rmd file on GitHub, until the very end when you will produce your final html, pdf or word file. This will make it so that when you deal with any conflicts in pull requests, you only need to edit the .Rmd file, and don't have to worry about the other files.

When you are working on your local machine, do make sure to knit the .Rmd file frequently to make sure your code is compiling correctly and that you can produce the html or pdf files. However when it comes time to push your changes to GitHub, I would recommend deleting all the extra files that R created so that you only push the .Rmd file and only have to worry about merging the .Rmd file when you make pull requests to your group leader.

Now you are set up to work cooperatively on an assignment using GitHub. The remainder of this assignment consists of a simple coding and plotting exercise to practice collaborating with Github.

4 Omitted Variable Bias Simulation (Group excercise)

For this problem we are going to generate simulated data, with known characteristics, so that we can explore how the omission of a key variable biases our estimates.

To do this in R markdown we must first tell R to interpret the text we write as code. We do this by embedding a code *chunk* in the document. The following is a code *chunk* that only consists of a comment. If you are reading this in the .Rmd file, notice the chunk starts and ends with three ‘ marks. The first marks are followed by {r NAME} where NAME is the name of the code chunk (all chunks must have different names). This tells R markdown that the text inside the chunk should be interpreted as R code (not written text). Notice there are also some icons at the top right of the chunk. The first allows you to set the chunk settings (which you can also do by writing in the settings in the {} at the beginning of the chunk). The chunk settings will tell R markdown whether you want the output, the warnings, the code... of the chunk to be visible in the final document. The second icon tells R to run all the code above the current chunk and the last to run the code in the current chunk.

```
#This is a comment, alone in a chunk.
```

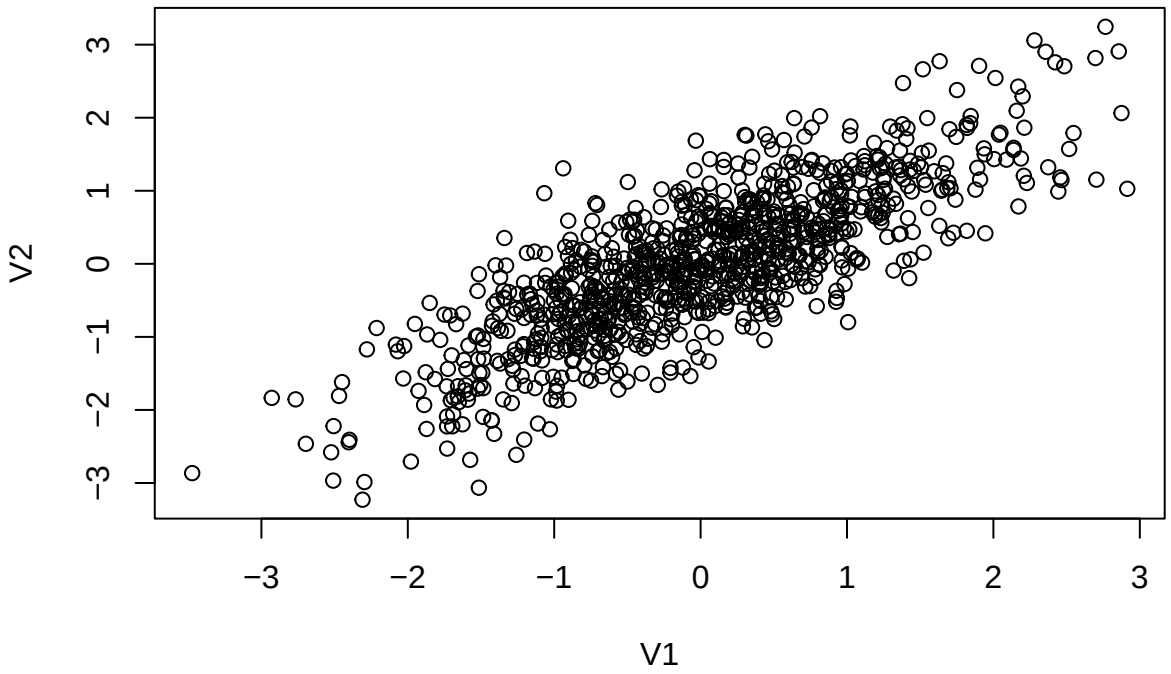
In the following chunk I generate simulated data consisting of two correlated variables. Notice the chunk setting is set to suppress the warning messages that are generated when R loads a package.

```
library(MASS)
library(ggplot2)

out <- as.data.frame(mvrnorm(1000, mu = c(0,0),
                     Sigma = matrix(c(1,0.8,0.8,1), ncol = 2),
                     empirical = TRUE))
cor(out)
```

```
##     V1  V2
## V1 1.0 0.8
## V2 0.8 1.0
```

```
plot(out)
```



Next I generate a randomly distributed error term and I calculate the outcome variable which depends on both V1 and V2 and some noise:

$$Y = \beta_1 V_1 + \beta_2 V_2 + \epsilon$$

```
out$error<-rnorm(1000, mean=0, sd=1)

#The data generating process
B1<-3
B2<-6

out$Y<-out$V1*B1+out$V2*B2+out$error
```

TO DO: For the questions below write the needed code and a written response to the question.

4.1 Question:

Write a chunk in which you regress Y on V_1 and V_2. Are your estimates of β_1 and β_2 biased?

Answer:

4.2 Question:

Write a chunk in which you regress Y on V_1 only. Is your estimate of β_1 biased?

Answer:

4.3 Question:

Generate a new variable Y_{adj} such that $Y_{adj} = Y - \beta_2 * V_2$. Then regress Y_{adj} on V_1. Is your estimate of β_1 biased? Can you explain why/why not?

Answer:

4.4 Question:

The code below generates a scatter plot and regression line for the relationship between V_1 and Y as well as V_1 and Y_{adj}. Submit an improved visualization of this data. Hint: you will need to delete the # to get the code to run

```
#plotted<-ggplot(out, aes(V1, y = value, color = variable)) +
#    geom_point(aes(y = Y, col = "Y")) +
#    geom_point(aes(y = adjY, col = "adjY"))+
#    geom_smooth(method='lm', aes(y = Y, col = "Y"))+
#    geom_smooth(method='lm', aes(y = adjY, col = "Y"))

#plotted
```

Answer:

4.5 Question:

Load the 'cps_clean.csv' dataset (available on Canvas). Regress income on education and interpret the coefficient.

Answer:

4.6 Question:

Add additional control variables to the regression you estimated above. How does this change your interpretation of the coefficient on education? Answer:

5 Submission instructions:

1) Make sure the final version of your assignment is uploaded on GitHub in both html and Rmarkdown format.