

中級ミクロデータサイエンス Problem Set 1.

氏名：五十嵐大和

学籍番号：2125701

Preparation : Installing some packages

用いるいくつかの package を必要に応じて、インストールする

```
install.packages("tidyverse")
install.packages("here")
install.packages("openxlsx")
install.packages("stringr")
```

(a) semester_dummy_tidy

1. 生データを読み込みなさい (semester_dummy_1.csv, semester_dummy_2.csv)
2. semester_dummy_1.csv については、1 行目を列名としなさい
3. 2 つのデータを適切に結合しなさい
4. 'Y' 列を削除しなさい

【コード】

```
# 1, 2 について
read_semester_data <- function(data_name, skip_option){
  name <- paste0(data_name, ".csv")
  path <- here::here("raw_data", "semester_dummy", name)
  if(file.exists(path))
    data <- readr::read_csv(path, skip = skip_option)
  else
    data <- readr::read_csv(name, skip = skip_option)
```

```

    return(data)
  }
  semester_data_1 <- read_semester_data("semester_data_1", skip_option = 1)
  semester_data_2 <- read_semester_data("semester_data_2", skip_option = 0)

  # 3, 4 について
  names(semester_data_2) <- names(semester_data_1)
  semester_dummy_tidy <- dplyr::bind_rows(semester_data_1, semester_data_2) |>
    dplyr::select(-Y)

```

【コードの説明】

(1, 2 について) `read_semester_data` という関数を作って、データを読み込む。この際、`skip_option` を設けることで、一行目を列名にするか否かを選択できるようにする。(GitHub 等でファイル構造を作った一つの場合と、特定のファイル構造がない場合のどちらにも対応できるように、`if` で場合分けをした。)

(3, 4 について) `names` 関数で、それぞれの列名をそろえたうえで、`dplyr::bind_rows` を用いて結合した。また、`dplyr::select` を用いて、'Y' 列を削除した。

(b) gradrate_tidy

1. 生データを読み込み、適切に結合しなさい
2. 女子学生の 4 年卒業率に 0.01 をかけて、0 から 1 のスケールに変更しなさい

【コード】

```

read_gradrate_tidy <- function(data_name){
  name <- paste0(data_name, " ", ".xlsx")
  path <- here::here("raw_data", "outcome", name)
  if(file.exists(path))
    data <- openxlsx::read.xlsx(path) |>
      dplyr::mutate_all(as.double)
  else
    data <- openxlsx::read.xlsx(name)|>
      dplyr::mutate_all(as.double)
  return(data)
}

years <- c(1991:1993, 1995:2016)
gradrate_tidy <- purrr::map(years, read_gradrate_tidy)|>

```

```
dplyr::bind_rows() |>
dplyr::mutate(women_gradrate_4yr = 0.01*women_gradrate_4yr)
```

【コードの説明】

(a) と同様に、`read_gradrate_tidy` という関数を作り、データを読み込む。生データに含まれる年度に関して、`purrr::map` と `dplyr::bind_rows` を用いて結合した。また、`dplyr::mutate` を用いて、`women_gradrate_4yr` の列を 0.01 倍した。

(c) covariates_tidy

1. 生データを読み込みなさい (`covariates.xlsx`)
2. 'university_id' という列名を 'unitid' に変更しなさい
3. 'unitid' に含まれる 'aaaa' という文字を削除しなさい
4. 'category' 列に含まれる 'instatetuition', 'costs', 'faculty', 'white_cohortsize' を別の列として追加しなさい (wide 型に変更しなさい)

【コード】

```
read_covariates_tidy <- function(data_name) {
  name <- paste0(data_name, ".xlsx")
  path <- here::here("raw_data", "covariates", name)
  if(file.exists(path))
    data <- openxlsx::read.xlsx(path)
  else
    data <- openxlsx::read.xlsx(name)
  return(data)
}

covariates_tidy <- read_covariates_tidy("covariates") |>
  dplyr::rename(unitid = university_id) |>
  dplyr::mutate(unitid = stringr::str_replace(unitid, "aaaa", "")) |>
  tidyr::pivot_wider(names_from = "category",
                    values_from = "value")
```

【コードの説明】

(a), (b) と同様に、`read_covariates_tidy` という関数を作り、データを読み込む。`dplyr::rename` で列名の変更、`dplyr::mutate` と `stringr::str_replace` を用いて文字の削除、`tidyr::pivot_wider` を用いて wide 型に変更した。

(d) gradrate_ready

1. 男女合計の 4 年卒業率と男子学生の 4 年卒業率を計算し、新たな列として追加しなさい
2. 計算した卒業率を有効数字 3 桁に調整しなさい
3. 卒業率に欠損値が含まれている行を削除しなさい

【コード】

```
gradrate_ready <- dplyr::mutate(gradrate_tidy, total_gradrate_4yr = tot4yrgrads/totcohortsize) |>
  dplyr::mutate(man_gradrate_4yr = m_4yrgrads/m_cohortsize) |>
  round(digits = 3) |>
  na.omit()
```

【コードの説明】

`dplyr::mutate` を用いて、男女合計と男子学生それぞれの 4 年での卒業生数/学生数を計算した。また、`round` 関数で四捨五入、`na.omit` で欠損値含む行を削除した。

(e) covariates_ready

1. 'outcome' や 'semester_dummy' に含まれる年を調べ、'covariates' データの期間を他のデータに揃えなさい
2. 'outcome_data' に含まれる 'unitid' を特定し、'covariates' に含まれる 'unitid' を 'outcome' データに揃えなさい

【コード】

```
year_gradrate_ready <- tidyr::expand(gradrate_ready, year) |>
  as.data.frame()
year_semester_dummy_tidy <- tidyr::expand(semester_dummy_tidy, year) |>
  as.data.frame()
```

```
unitid_gradrate_ready <- tidyr::expand(gradrate_ready, unitid) |>
  as.data.frame()

covariates_ready <- dplyr::mutate(covariates_tidy, dplyr::across(everything(), as.double, na.rm =
  dplyr::filter(year %in% year_gradrate_ready$year & year %in% year_semester_dummy_tidy$year) |>
  dplyr::filter(unitid %in% unitid_gradrate_ready$unitid)
```

【コードの説明】

tidyr::expand を用いて、異なる年を抽出する。dplyr::filter を用いて、gradrate_ready や semester_dummy_tidy に含まれる年や、gradrate_ready に含まれる 'unitid' を特定し、整理した。

(f) master

1. 結合に用いる変数を考え、semester_dummy_tidy, covariates_ready, gradrate_ready を適切に結合しなさい
2. 白人学生が学生全体に占める割合を計算し、有効数字 3 桁に調整した上で、新たな列として追加しなさい

【コード】

```
master <- dplyr::inner_join(semester_dummy_tidy, gradrate_ready, by = c("unitid", "year")) |>
  dplyr::inner_join(covariates_ready, by = c("unitid", "year")) |>
  dplyr::mutate(rate_for_white_student = white_cohortsize/totcohortsize)
```

【コードの説明】

dplyr::inner_join を用いて、3 つのデータを結合した。dplyr::mutate を用いて、白人学生数/全学生数の率を現した列を追加した。

(補足) それぞれのデータの中身の確認

■(a) semester_dummy_tidy

```
# A tibble: 6 x 5
  unitid instnm semester quarter year
  <dbl> <chr>      <dbl>   <dbl> <dbl>
```

1	100654	ALABAMA A&M UNIVERSITY	1	0	1991
2	100654	ALABAMA A&M UNIVERSITY	1	0	1992
3	100654	ALABAMA A & M UNIVERSITY	1	0	1993
4	100654	ALABAMA A & M UNIVERSITY	1	0	1995
5	100654	ALABAMA A & M UNIVERSITY	1	0	1996
6	100654	ALABAMA A & M UNIVERSITY	1	0	1997

■(b) gradrate_tidy

	unitid	year	totcohortsize	w_cohortsize	m_cohortsize	tot4yrgrads	m_4yrgrads
1	100654	1991	1010	527	483	152	32
2	100663	1991	937	500	437	82	33
3	100751	1991	2511	1348	1163	630	213
4	100858	1991	3024	1496	1528	846	312
5	101435	1991	189	101	88	60	25
6	101480	1991	1100	567	533	109	42

	w_4yrgrads	women_gradrate_4yr
1	120	0.2277
2	49	0.0980
3	417	0.3093
4	534	0.3570
5	35	0.3465
6	67	0.1182

■(c) covariates_tidy

A tibble: 6 x 6

	unitid	year	instatetuition	costs	faculty	white_cohortsize
	<chr>	<chr>	<chr>	<chr>	<chr>	<chr>
1	100654	1987	<NA>	<NA>	<NA>	<NA>
2	100654	1988	<NA>	<NA>	<NA>	<NA>
3	100654	1989	<NA>	<NA>	<NA>	<NA>
4	100654	1990	1248	<NA>	240	<NA>
5	100654	1991	1298	53.121007	223	11
6	100654	1992	1600	52.536624	267	5

■(d) gradrate_ready

	unitid	year	totcohortsize	w_cohortsize	m_cohortsize	tot4yrgrads	m_4yrgrads
1	100654	1991	1010	527	483	152	32
2	100663	1991	937	500	437	82	33
3	100751	1991	2511	1348	1163	630	213
4	100858	1991	3024	1496	1528	846	312
5	101435	1991	189	101	88	60	25
6	101480	1991	1100	567	533	109	42

	w_4yrgrads	women_gradrate_4yr	total_gradrate_4yr	man_gradrate_4yr
1	120	0.228	0.150	0.066
2	49	0.098	0.088	0.076
3	417	0.309	0.251	0.183
4	534	0.357	0.280	0.204
5	35	0.346	0.317	0.284
6	67	0.118	0.099	0.079

■(e) covariates_ready

A tibble: 6 x 6

	unitid	year	instatetuition	costs	faculty	white_cohortsize
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	100654	1991	1298	53.1	223	11
2	100654	1992	1600	52.5	267	5
3	100654	1993	1600	50.4	262	7
4	100654	1995	2022	68.0	298	18
5	100654	1996	2312	66.4	311	17
6	100654	1997	2420	72.5	306	22

■(f) master

A tibble: 6 x 19

	unitid	instnm	semester	quarter	year	totcohortsize	w_cohortsize	m_cohortsize
	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	100654	ALABAMA~	1	0	1991	1010	527	483
2	100654	ALABAMA~	1	0	1992	876	444	432
3	100654	ALABAMA~	1	0	1993	1019	543	476
4	100654	ALABAMA~	1	0	1995	849	434	415
5	100654	ALABAMA~	1	0	1996	716	372	344
6	100654	ALABAMA~	1	0	1997	789	412	377

i 11 more variables: tot4yrgrads <dbl>, m_4yrgrads <dbl>, w_4yrgrads <dbl>,

```
#  women_gradrate_4yr <dbl>, total_gradrate_4yr <dbl>, man_gradrate_4yr <dbl>,  
#  instatetuition <dbl>, costs <dbl>, faculty <dbl>, white_cohortsize <dbl>,  
#  rate_for_white_student <dbl>
```