Week 4 In-class Assignment

Felix Ho

2024-09-30

List the names of the variables.

```
finaldata <- read.csv(here("data", "mergealldata.csv"), header = TRUE)
names(finaldata)
```

```
 [1] "country_name" "ISO"          "region"       "year"         "gdp1000"
 [6] "OECD"         "OECD2023"     "popdens"      "urban"        "agedep"
[11] "male_edu"     "temp"         "rainfall1000" "matmor"       "infmor"
[16] "neomor"       "un5mor"       "drought"      "earthquake"   "totdeath"
[21] "armcon"
```

The main exposure variable is armed conflict. As per the paper, there are 10 covariates, country and year fixed effects, and conflict lagged by 1 year. Match my variables to those from the paper.

Corresponding to Table 2 in the paper:

- armcon = armed conflict (binary) variable lagged by 1 year

10 covariates:

- gdp1000 = GDP per capita in US dollars (unit is scaled up by 1,000)
- OECD = OECD member
- popdens = population density represents the % of the population living in a density of >1,000 people/km^2
- urban = urban residence represents the % of the population living in urban areas
- agedep = age dependency ratio represents the proportion of dependents (aged $<$ 15 years or $>$ 64 years) per 100 working-age individuals
- male_edu = male education expressed as years per capita (age-standardised)

- temp = temperature in degrees Celsius and is the mean population-weighted annual temperature
- rainfall1000 = mean population-weighted annual rainfall in mm per year (scaled down by 1,000)
- earthquake = earthquake binary variable (absence or presence)
- drought = drought binary variable (absence or presence)

Primary outcomes:

- matmor = maternal mortality rate
- un5mor = under-5 mortality rate
- infmor = infant mortality rate
- neomor = neonatal mortality rate

Note:

- totdeath = total number of battle related deaths

Determine the classes of the variables.

```r
glimpse(finaldata)
```

```
Rows: 3,720
Columns: 21
$ country_name <chr> "Afghanistan", "Afghanistan", "Afghanistan", "Afghanistan~
$ ISO          <chr> "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "AFG", "~
$ region       <chr> "Southern Asia", "Southern Asia", "Southern Asia", "South~
$ year         <int> 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 200~
$ gdp1000      <dbl> NA, NA, 0.1835328, 0.2004626, 0.2216576, 0.2550551, 0.274~
$ OECD         <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ OECD2023     <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
$ popdens      <dbl> 14.13654, 14.23156, 14.32270, 14.40691, 15.21947, 15.3361~
$ urban        <dbl> 16.25324, 16.25661, 16.42654, 16.60701, 16.71367, 16.8509~
$ agedep       <dbl> 108.34663, 108.98989, 109.34716, 109.44753, 109.28682, 10~
$ male_edu     <dbl> 2.762086, 2.856936, 2.954241, 3.054121, 3.156706, 3.26213~
$ temp         <dbl> 12.69959, 12.85570, 12.71081, 12.16592, 13.04643, 12.2314~
$ rainfall1000 <dbl> 0.2763704, 0.2793079, 0.3805710, 0.4288939, 0.3754336, 0.~
$ matmor       <int> 1450, 1390, 1300, 1240, 1180, 1140, 1120, 1090, 1030, 993~
$ infmor       <dbl> 90.5, 87.9, 85.3, 82.7, 80.0, 77.3, 74.6, 71.9, 69.2, 66.~
$ neomor       <dbl> 60.9, 59.7, 58.5, 57.2, 55.9, 54.6, 53.2, 51.7, 50.3, 48.~
```

```
$ un5mor       <dbl> 129.2, 125.2, 121.1, 116.9, 112.6, 108.4, 104.1, 99.9, 95~
$ drought      <int> 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, ~
$ earthquake   <int> 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, ~
$ totdeath     <int> 5065, 5394, 5553, 1157, 944, 817, 1711, 4982, 7020, 5660,~
$ armcon       <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
```

Not clear what OECD2023 stands for. How is it different from OECD?

Look for duplicated rows.

```r
get_dupes(finaldata)
```

```
No variable names specified - using all columns.
```

```
No duplicate combinations found of: country_name, ISO, region, year, gdp1000, OECD, OECD2023
```

```
 [1] country_name ISO          region       year         gdp1000
 [6] OECD         OECD2023     popdens      urban        agedep
[11] male_edu     temp         rainfall1000 matmor       infmor
[16] neomor       un5mor       drought      earthquake   totdeath
[21] armcon       dupe_count
<0 rows> (or 0-length row.names)
```

There are no duplicated rows.

View the key summary statistics of numeric variables and the number of NA's for the variables.

```r
summary(finaldata)
```

```
 country_name           ISO               region               year     
 Length:3720        Length:3720        Length:3720        Min.   :2000  
 Class :character   Class :character   Class :character   1st Qu.:2005  
 Mode  :character   Mode  :character   Mode  :character   Median :2010  
                                                          Mean   :2010  
                                                          3rd Qu.:2014  
                                                          Max.   :2019  
                                                                        
    gdp1000              OECD          OECD2023          popdens      
 Min.   :  0.1105   Min.   :0.000   Min.   :0.0000   Min.   :  0.00  
 1st Qu.:  1.2383   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:14.79  
```

```
 Median :  4.0719   Median :0.000   Median :0.0000   Median :27.52  
 Mean   : 11.4917   Mean   :0.171   Mean   :0.1882   Mean   :30.57  
 3rd Qu.: 13.1531   3rd Qu.:0.000   3rd Qu.:0.0000   3rd Qu.:40.72  
 Max.   :123.6787   Max.   :1.000   Max.   :1.0000   Max.   :99.86  
 NA's   :62                                          NA's   :20     
     urban            agedep          male_edu           temp       
 Min.   : 0.1025   Min.   : 16.17   Min.   : 1.067   Min.   :-2.405  
 1st Qu.:17.2872   1st Qu.: 47.94   1st Qu.: 5.904   1st Qu.:12.928  
 Median :30.2535   Median : 55.51   Median : 8.368   Median :21.958  
 Mean   :30.6948   Mean   : 61.94   Mean   : 8.258   Mean   :19.625  
 3rd Qu.:41.6558   3rd Qu.: 77.11   3rd Qu.:10.849   3rd Qu.:25.869  
 Max.   :93.4135   Max.   :111.48   Max.   :14.441   Max.   :29.676  
 NA's   :20                         NA's   :20       NA's   :20      
  rainfall1000        matmor           infmor           neomor     
 Min.   :0.01993   Min.   :   2.0   Min.   :  1.60   Min.   : 0.80  
 1st Qu.:0.59146   1st Qu.:  17.0   1st Qu.:  7.60   1st Qu.: 4.90  
 Median :1.01288   Median :  66.0   Median : 18.90   Median :12.10  
 Mean   :1.20216   Mean   : 210.6   Mean   : 28.90   Mean   :16.18  
 3rd Qu.:1.68706   3rd Qu.: 299.8   3rd Qu.: 44.52   3rd Qu.:25.32  
 Max.   :4.71081   Max.   :2480.0   Max.   :138.10   Max.   :60.90  
 NA's   :20        NA's   :426      NA's   :20       NA's   :20     
     un5mor          drought          earthquake         totdeath      
 Min.   :  2.00   Min.   :0.00000   Min.   :0.00000   Min.   :    0.0  
 1st Qu.:  9.00   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:    0.0  
 Median : 22.20   Median :0.00000   Median :0.00000   Median :    0.0  
 Mean   : 40.50   Mean   :0.08737   Mean   :0.08333   Mean   :  361.1  
 3rd Qu.: 61.33   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:    2.0  
 Max.   :224.90   Max.   :1.00000   Max.   :1.00000   Max.   :78644.0  
 NA's   :20                                                            
     armcon      
 Min.   :0.0000  
 1st Qu.:0.0000  
 Median :0.0000  
 Mean   :0.1892  
 3rd Qu.:0.0000  
 Max.   :1.0000  
```

The median of gdp1000 (4.0719) appears to be far from the mean (11.4917). The distribution of gdp1000 may be positively skewed. The median of matmor (66.0) appears to be far from the mean (210.6). The distribution of matmor may be positively skewed. The median of infmor (18.90) appears to be far from the mean (28.90). The distribution of infmor may be positively skewed. The median of un5mor (22.20) appears to be far from the mean (40.50). The distribution of un5mor may be positively skewed. There are a lot of NA's for matmor (426).

```
table(finaldata$OECD)
```

```
   0    1
3084  636
```

OECD is a binary variable. Maybe 0 and 1 represents nonmember and member of OECD, respectively?

Focus on countries with high matmor.

```
highmatmor <- finaldata %>%
  select(country_name, year, matmor) %>%
  arrange(desc(matmor))
highmatmor[1:20,]
```

```
   country_name year matmor
1  Sierra Leone 2000   2480
2  Sierra Leone 2001   2250
3  Sierra Leone 2002   2080
4  Sierra Leone 2003   1960
5  Sierra Leone 2004   1850
6  Sierra Leone 2005   1760
7   South Sudan 2000   1730
8   South Sudan 2001   1690
9  Sierra Leone 2006   1680
10  South Sudan 2002   1660
11 Sierra Leone 2007   1610
12  South Sudan 2003   1610
13  South Sudan 2004   1550
14 Sierra Leone 2008   1530
15  South Sudan 2005   1480
16  Afghanistan 2000   1450
17 Sierra Leone 2009   1450
18         Chad 2000   1420
19         Chad 2001   1410
20  South Sudan 2006   1410
```

The countries with high matmor do not appear to be developed countries, which make sense.

Focus on countries with high un5mor.

```
highun5mor <- finaldata %>%
  select(country_name, year, un5mor) %>%
  arrange(desc(un5mor))
highun5mor[1:20,]
```

```
   country_name year un5mor
1         Niger 2000  224.9
2  Sierra Leone 2000  224.9
3  Sierra Leone 2001  219.4
4         Niger 2001  215.2
5  Sierra Leone 2002  213.9
6  Sierra Leone 2003  208.1
7         Niger 2002  204.5
8        Angola 2000  204.4
9         Haiti 2010  203.6
10 Sierra Leone 2004  202.0
11       Angola 2001  198.4
12 Sierra Leone 2005  195.5
13        Niger 2003  193.1
14       Angola 2002  191.5
15      Liberia 2000  189.7
16 Sierra Leone 2006  188.9
17         Mali 2000  187.4
18       Rwanda 2000  185.2
19         Chad 2000  184.0
20       Angola 2003  183.8
```

The countries with high un5mor do not appear to be developed countries, which make sense.

Focus on countries with high infmor.

```
highinfmor <- finaldata %>%
  select(country_name, year, infmor) %>%
  arrange(desc(infmor))
highinfmor[1:20,]
```

```
               country_name year infmor
1              Sierra Leone 2000  138.1
```

```
2               Sierra Leone 2001  135.6
3               Sierra Leone 2002  132.9
4               Sierra Leone 2003  130.2
5                    Liberia 2000  127.9
6               Sierra Leone 2004  127.2
7               Sierra Leone 2005  124.1
8                     Angola 2000  121.5
9               Sierra Leone 2006  120.9
10                   Liberia 2001  119.7
11                    Angola 2001  118.2
12              Sierra Leone 2007  117.6
13                    Angola 2002  114.5
14              Sierra Leone 2008  114.2
15                Mozambique 2000  112.4
16                   Liberia 2002  111.9
17              Sierra Leone 2009  110.6
18                    Angola 2003  110.4
19 Central African Republic 2000  109.9
20                   Nigeria 2000  109.8
```

The countries with high infmor do not appear to be developed countries, which make sense. Sierra Leone stands out with high matmor, un5mor, and infmor.

Focus on countries with high neomor.

```
highneomor <- finaldata %>%
  select(country_name, year, neomor) %>%
  arrange(desc(neomor))
highneomor[1:20,]
```

```
    country_name year neomor
1    Afghanistan 2000   60.9
2    Afghanistan 2001   59.7
3    Afghanistan 2002   58.5
4    Afghanistan 2003   57.2
5       Pakistan 2000   56.8
6    South Sudan 2000   56.0
7    Afghanistan 2004   55.9
8       Pakistan 2001   55.8
9  Guinea-Bissau 2000   55.3
10      Pakistan 2002   54.9
11   South Sudan 2001   54.9
```

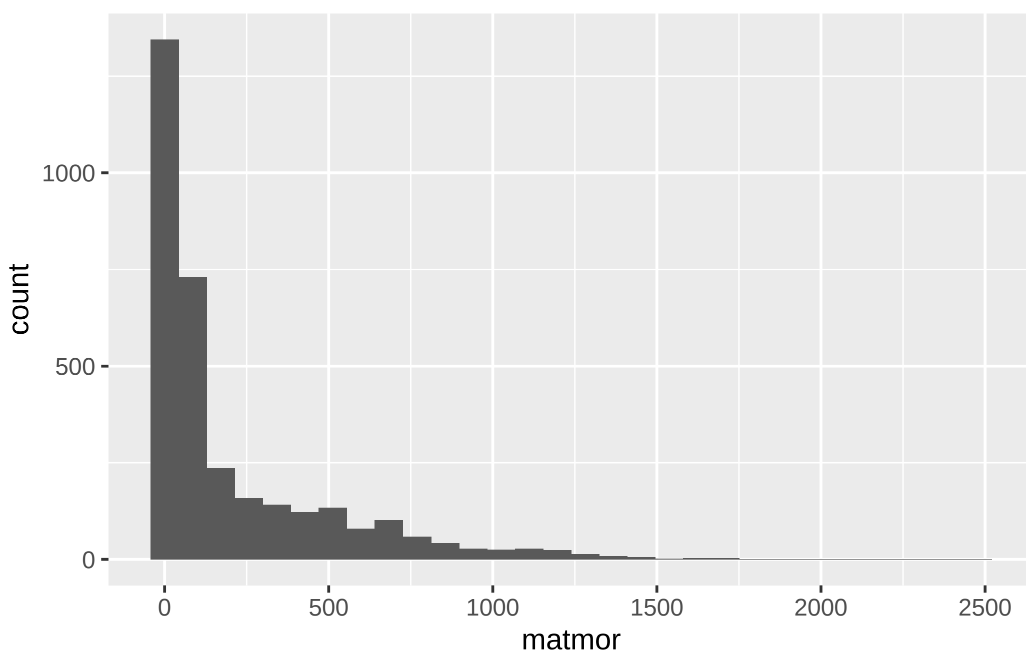
```
12   Afghanistan 2005   54.6
13 Guinea-Bissau 2001   54.2
14      Pakistan 2003   54.0
15   South Sudan 2002   53.5
16 Guinea-Bissau 2002   53.3
17      Pakistan 2004   53.3
18   Afghanistan 2006   53.2
19      Pakistan 2005   52.6
20 Guinea-Bissau 2003   52.5
```

The countries with high neomor do not appear to be developed countries, which make sense.

Look at the distribution of matmor.

```
finaldata %>%
  ggplot(aes(x = matmor)) +
  geom_histogram(bins = 30)
```

```
Warning: Removed 426 rows containing non-finite outside the scale range
(`stat_bin()`).
```

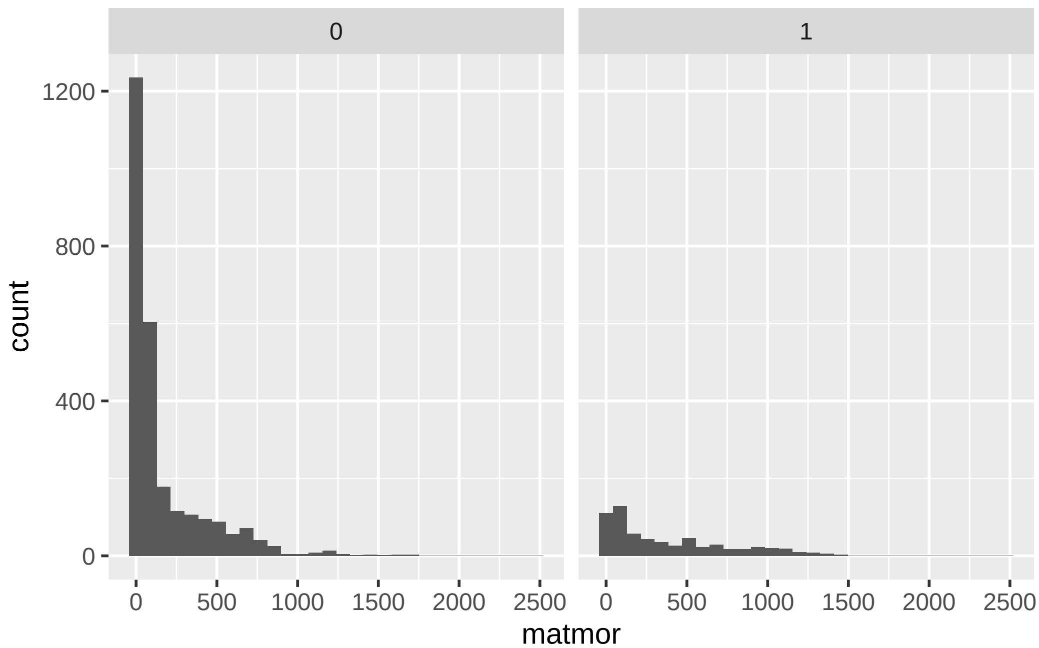


This suggests the presence of outliers: there are a small number of countries with high matmor.

Group by armcon.

```
finaldata %>%
  ggplot() +
  geom_histogram(
    aes(x = matmor),
    bins = 30) +
  facet_wrap(vars(armcon))
```

```
Warning: Removed 426 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Determine the counts in both responses of armcon.

```
table(finaldata$armcon)
```

```
   0    1
3016  704
```

It makes sense that the histogram above has more counts for armcon = 0.

Create a correlation matrix of the four mortality rates.

```r
mortality <- select(finaldata, matmor, un5mor, infmor, neomor)
mortality.nona <- na.omit(mortality)
matrix = cor(mortality.nona)
corrplot(matrix, method = 'number')
```

There is very strong positive correlation among the four mortality rates, which makes sense.