hw_01

lisa liubovich

2024-01-24

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

R Basics

```
cdi <- read_csv(file = "https://dcgerard.github.io/stat_415_615/data/cdi.csv")
```

```
## Rows: 440 Columns: 17
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (3): county, state, region
## dbl (14): id, area, pop, percent_18_34, percent_65, physicians, beds, crimes...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
cdi <- cdi %>%
  mutate(log_capita_income = log(capita_income))
```

```
cdi %>%
  summarise(mean_area = mean(area),
            sd_are = sd(area))
```

```
## # A tibble: 1 x 2
##   mean_area sd_are
##       <dbl>  <dbl>
## 1     1041.  1550.
```

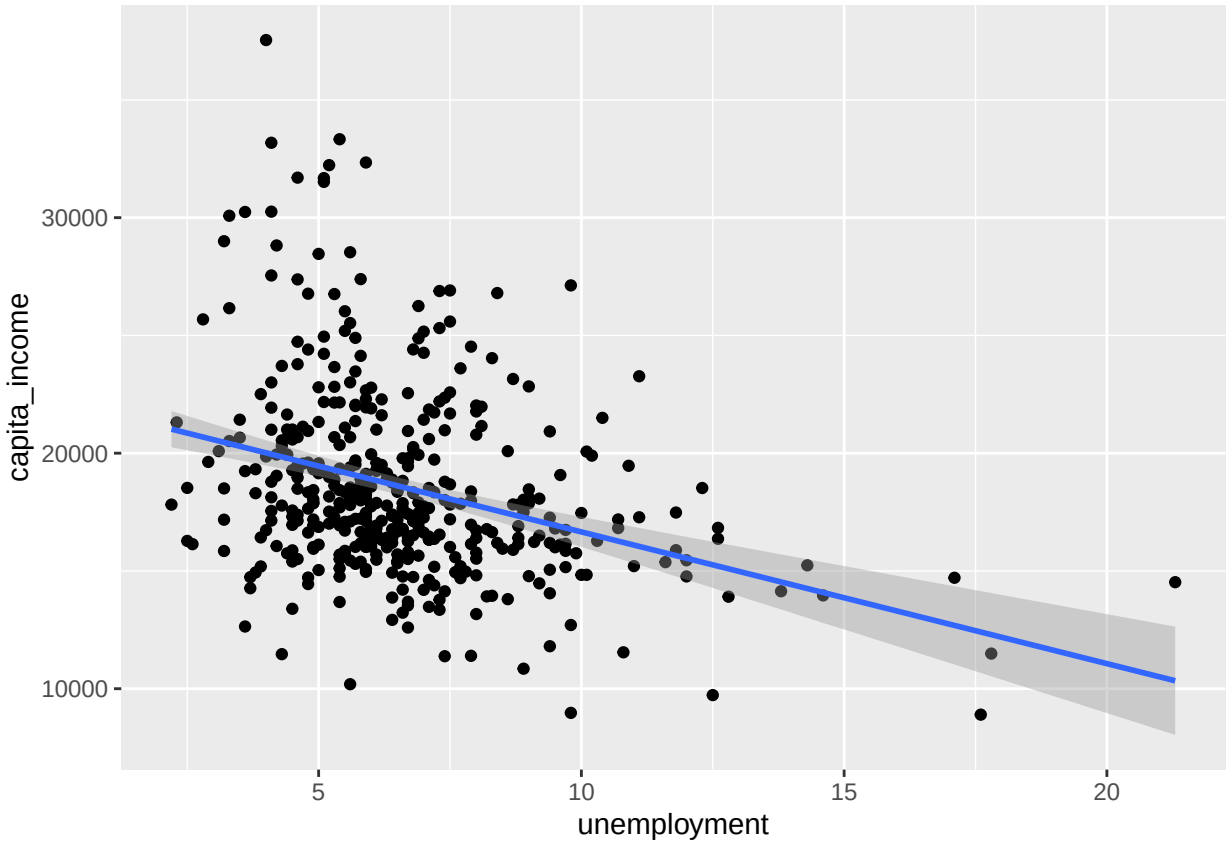
```r
cdi <- cdi %>%
  rename(population = pop)
```

```r
cdi %>%
  filter(state == "DE")
```

```
## # A tibble: 2 x 18
##      id county  state  area population percent_18_34 percent_65 physicians  beds
##   <dbl> <chr>   <chr> <dbl>      <dbl>          <dbl>      <dbl>      <dbl> <dbl>
## 1   109 New_Ca~ DE      426     441946           30.6       11.4       1038  1488
## 2   397 Kent    DE      591     110993           29.7       10.3        123   193
## # i 9 more variables: crimes <dbl>, high_school <dbl>, bachelors <dbl>,
## #   poverty <dbl>, unemployment <dbl>, capita_income <dbl>, total_income <dbl>,
## #   region <chr>, log_capita_income <dbl>
```

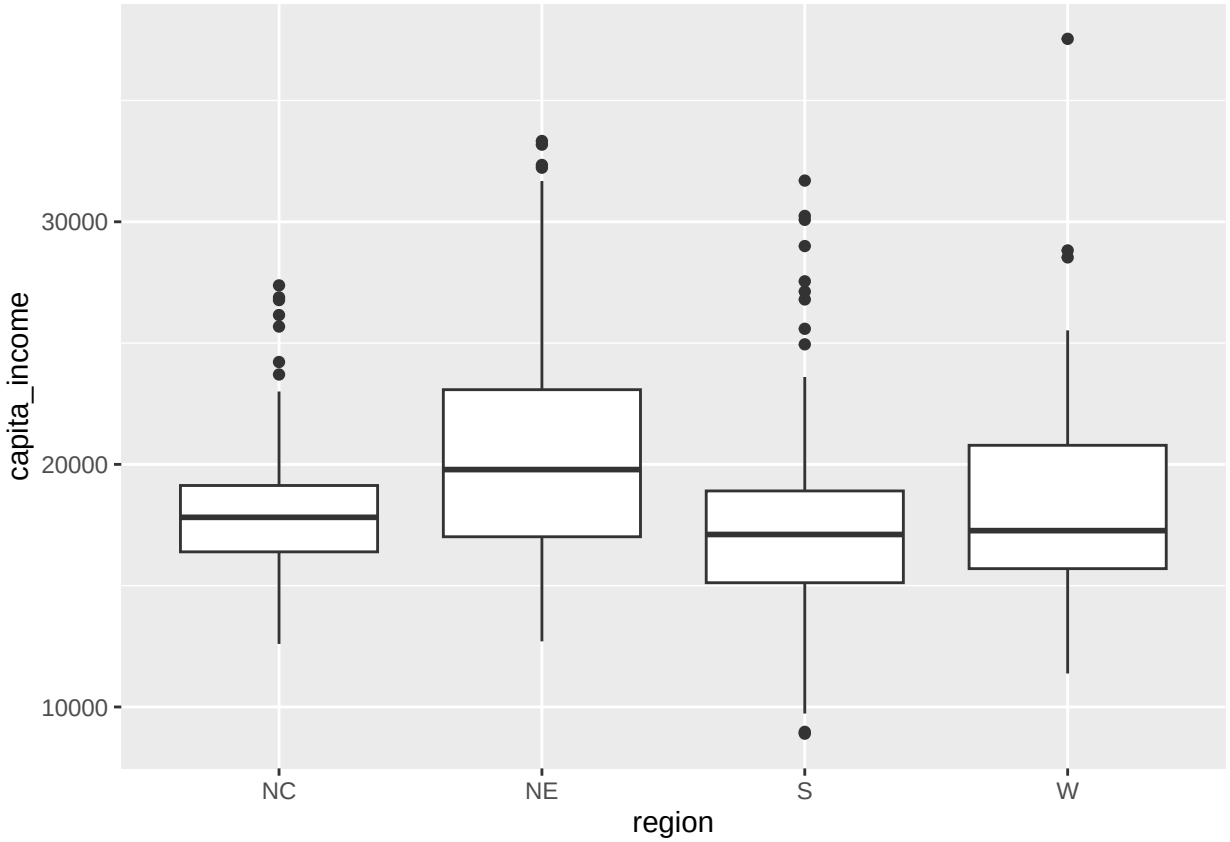
```r
ggplot(cdi, mapping = aes(x = unemployment, y = capita_income)) +
  geom_point() +
  geom_smooth(method = lm)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



There seems to be a weak-moderate negative relationship between unemployment and capita_income. The relationship does not fit a linear relationship very well, so a non linear relationship might be more appropriate.

```
ggplot(cdi, mapping = aes(x = region, y = capita_income)) +
  geom_boxplot()
```



Misc

let:

```
x <- seq(1, 100, by = 3)
```

```
sum(log(x))
```

```
## [1] 122.594
```

```
log(sum(x))
```

```
## [1] 7.448334
```

No, the values are not the same.