Data Taming Mid-Trimester Test Reminder

Dongju Ma

2024-07-14

Load the packages and data

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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
```

```
library(inspectdf)
data('midwest')
```

Build a separate linear model

```
# Build a multiple regression linear model by states, with separate lines
mw_sep <- lm(percollege ~ log(poptotal)+state+log(poptotal):state,
             data=midwest)
summary(mw_sep)
```

```
##
## Call:
## lm(formula = percollege ~ log(poptotal) + state + log(poptotal):state,
##     data = midwest)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6472  -3.0337  -0.5885   1.7750  22.9818
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)           -17.3013     4.0095  -4.315 1.98e-05 ***
## log(poptotal)           3.4844     0.3845   9.062  < 2e-16 ***
## stateIN               -10.3181     6.9581  -1.483   0.1388
## stateMI                 5.1710     5.9568   0.868   0.3858
## stateOH               -11.9309     6.9306  -1.721   0.0859 .
## stateWI                -3.5691     6.9893  -0.511   0.6099
## log(poptotal):stateIN   0.7393     0.6636   1.114   0.2659
## log(poptotal):stateMI  -0.5188     0.5630  -0.921   0.3573
## log(poptotal):stateOH   0.6767     0.6371   1.062   0.2888
## log(poptotal):stateWI   0.4082     0.6649   0.614   0.5396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.694 on 427 degrees of freedom
## Multiple R-squared:  0.4496, Adjusted R-squared:  0.438
## F-statistic: 38.76 on 9 and 427 DF,  p-value: < 2.2e-16
```

Separate lines

```
anova(mw_sep)
```

```
## Analysis of Variance Table
##
## Response: percollege
##                       Df Sum Sq Mean Sq  F value    Pr(>F)
## log(poptotal)          1 6022.6  6022.6 273.2999 < 2.2e-16 ***
## state                  4 1547.5   386.9  17.5565 2.375e-13 ***
## log(poptotal):state    4  116.5    29.1   1.3213    0.2612
## Residuals            427 9409.6    22.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The line which has a p-value greater than 0.05 should be separated.

Predict the data

```
# Predict an individual sample of 10000 population county in Ohio
pred_value <- tibble(poptotal = 10000, state = 'OH')
predict(mw_sep, pred_value, interval = 'prediction', level = 0.95)
```

```
##        fit        lwr      upr
## 1 9.092395 -0.3735181 18.55831
```

About skewness

- Right-skewed: the peak is on the left side
- Left-skewed: the peak is on the right side