VIF and Contingency Table

Bryan Mui UID 506021334

2025-12-02

Setup

```
# Load necessary libraries
library(car)  # For calculating VIF
```

```
## Loading required package: carData
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following object is masked from 'package:car':
##
##     recode

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(knitr)
library(broom)
```

Making an individual data set for each demographic

Copypasta

```
df <- read_csv("df_2024_model.csv")
```

```
## Rows: 832 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): State, Demographic
## dbl (6): Total citizen population, Total registered, Percent registered (Cit...
##
## 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.
```

```
# Define Region Mapping
region_map <- c(
  'Connecticut' = 'Northeast', 'Maine' = 'Northeast', 'Massachusetts' = 'Northeast',
  'New Hampshire' = 'Northeast', 'Rhode Island' = 'Northeast', 'Vermont' = 'Northeast',
  'New Jersey' = 'Northeast', 'New York' = 'Northeast', 'Pennsylvania' = 'Northeast',
  'Illinois' = 'Midwest', 'Indiana' = 'Midwest', 'Michigan' = 'Midwest', 'Ohio' = 'Midwest',
  'Wisconsin' = 'Midwest', 'Iowa' = 'Midwest', 'Kansas' = 'Midwest', 'Minnesota' = 'Midwest',
  'Missouri' = 'Midwest', 'Nebraska' = 'Midwest', 'North Dakota' = 'Midwest',
  'South Dakota' = 'Midwest', 'Delaware' = 'South', 'Florida' = 'South', 'Georgia' = 'South',
  'Maryland' = 'South', 'North Carolina' = 'South', 'South Carolina' = 'South',
  'Virginia' = 'South', 'West Virginia' = 'South', 'Alabama' = 'South', 'Kentucky' = 'South',
  'Mississippi' = 'South', 'Tennessee' = 'South', 'Arkansas' = 'South', 'Louisiana' = 'South',
  'Oklahoma' = 'South', 'Texas' = 'South', 'District Of Columbia' = 'South',
  'Arizona' = 'West', 'Colorado' = 'West', 'Idaho' = 'West', 'Montana' = 'West',
  'Nevada' = 'West', 'New Mexico' = 'West', 'Utah' = 'West', 'Wyoming' = 'West',
  'Alaska' = 'West', 'California' = 'West', 'Hawaii' = 'West', 'Oregon' = 'West',
  'Washington' = 'West'
)

# Apply mapping
df$State_Title <- tools::toTitleCase(tolower(df$State))
df$Region <- region_map[df$State_Title]

# Clean Data (Remove United States aggregate and rows with missing regions)
df_clean <- df %>%
  filter(State_Title != 'United States', !is.na(Region))

# ---------------------------------------------------------
# 2. Contingency Table (LaTeX)
# ---------------------------------------------------------
# Filter out 'Total' for the demographic counts
df_counts <- df_clean %>% filter(Demographic != 'Total')

contingency_table <- table(df_counts$Demographic, df_counts$Region)

print("--- Contingency Table (LaTeX) ---")
```

```
## [1] "--- Contingency Table (LaTeX) ---"
```

```
(kable(contingency_table, format = "latex", caption = "Demographic vs Region Counts"))
```

```
# ---------------------------------------------------------
# 3. Separate VIF Models
```

Table 1: Demographic vs Region Counts

	Midwest	Northeast	South	West
18 to 24 years	12	9	16	13
25 to 34 years	12	9	16	13
35 to 44 years	12	9	16	13
45 to 64 years	12	9	16	13
65 years and over	12	9	16	13
Asian alone	12	9	16	13
Asian alone or in combination	12	9	16	13
Black alone	12	9	16	13
Black alone or in combination	12	9	16	13
Female	12	9	16	13
Hispanic (any race)	12	9	16	13
Male	12	9	16	13
White alone	12	9	16	13
White alone or in combination	12	9	16	13
White non-Hispanic alone	12	9	16	13

Table 2: VIF Results: Race and Region

	GVIF	Df	GVIF^(1/(2*Df))
Demographic	1	7	1
Region	1	3	1

```
# ------------------------------------------------------------

# Define Groups
race_groups <- c('Asian alone', 'Asian alone or in combination', 'Black alone',
                 'Black alone or in combination', 'Hispanic (any race)',
                 'White alone', 'White alone or in combination',
                 'White non-Hispanic alone')
gender_groups <- c('Male', 'Female')

# --- Model A: Race + Region ---
df_race <- df_clean %>% filter(Demographic %in% race_groups)
model_race <- lm(`Voter participation rate` ~ Demographic + Region, data = df_race)
vif_race <- vif(model_race)

print("--- VIF: Race + Region (LaTeX) ---")
```

```
## [1] "--- VIF: Race + Region (LaTeX) ---"
```

```
(kable(as.data.frame(vif_race), format = "latex", caption = "VIF Results: Race and Region"))
```

```
# --- Model B: Gender + Region ---
df_gender <- df_clean %>% filter(Demographic %in% gender_groups)
model_gender <- lm(`Voter participation rate` ~ Demographic + Region, data = df_gender)
vif_gender <- vif(model_gender)

print("--- VIF: Gender + Region (LaTeX) ---")
```

Table 3: VIF Results: Gender and Region

	GVIF	Df	GVIF^(1/(2*Df))
Demographic	1	1	1
Region	1	3	1

Table 4: Regression Results: Region Only

term	estimate	std.error	statistic	p.value
(Intercept)	0.8860797	0.0098217	90.2169835	0.0000000
RegionNortheast	0.0217642	0.0150028	1.4506767	0.1536552
RegionSouth	-0.0183314	0.0129928	-1.4108832	0.1650066
RegionWest	0.0103571	0.0136202	0.7604222	0.4508825

```
## [1] "--- VIF: Gender + Region (LaTeX) ---"
```

```
(kable(as.data.frame(vif_gender), format = "latex", caption = "VIF Results: Gender and Region"))
```

```
# --- Model C: Region Only ---
# Note: VIF requires at least 2 predictors to calculate multicollinearity.
# We will fit the model using the 'Total' rows (State averages) and show the summary.
df_total <- df_clean %>% filter(Demographic == 'Total')
model_region <- lm(`Voter participation rate` ~ Region, data = df_total)

print("--- Model: Region Only (Summary) ---")
```

```
## [1] "--- Model: Region Only (Summary) ---"
```

```
# Using broom::tidy to format model results nicely for kable
if(require(broom)) {
  (kable(tidy(model_region), format = "latex", caption = "Regression Results: Region Only"))
} else {
  print(summary(model_region))
}
```

```
# ------------------------------------------------------------
# 4. Average Participation Tables (LaTeX)
# ------------------------------------------------------------

# Region Average
region_stats <- df_total %>%
  group_by(Region) %>%
  summarise(Average_Participation = mean(`Voter participation rate`, na.rm = TRUE))

(kable(region_stats, format = "latex", caption = "Average Voter Participation by Region"))
```

Table 5: Average Voter Participation by Region

Region	Average_Participation
Midwest	0.8860797
Northeast	0.9078440
South	0.8677484
West	0.8964368

Table 6: Average Voter Participation by Age Group

Demographic	Average_Participation
18 to 24 years	0.8146502
25 to 34 years	0.8270717
35 to 44 years	0.8699542
45 to 64 years	0.9049878
65 years and over	0.9316525

```
# Age Average
age_groups <- c('18 to 24 years', '25 to 34 years', '35 to 44 years', '45 to 64 years', '65 years and o
age_stats <- df_clean %>%
  filter(Demographic %in% age_groups) %>%
  group_by(Demographic) %>%
  summarise(Average_Participation = mean(`Voter participation rate`, na.rm = TRUE))

(kable(age_stats, format = "latex", caption = "Average Voter Participation by Age Group"))
```

```
# Race Average
race_stats <- df_clean %>%
  filter(Demographic %in% race_groups) %>%
  group_by(Demographic) %>%
  summarise(Average_Participation = mean(`Voter participation rate`, na.rm = TRUE))

(kable(race_stats, format = "latex", caption = "Average Voter Participation by Race"))
```

Table 7: Average Voter Participation by Race

Demographic	Average_Participation
Asian alone	0.8520952
Asian alone or in combination	0.8625600
Black alone	0.8425733
Black alone or in combination	0.8493464
Hispanic (any race)	0.8165691
White alone	0.8958723
White alone or in combination	0.8946120
White non-Hispanic alone	0.9033247