data_validation

Madison Thompson

2025-12-10

Validating the data used for the models (part G of methods)

```
# imports
library(car)
```

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

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.1     v stringr   1.5.2
## v ggplot2   4.0.0     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.1.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::recode() masks car::recode()
## x purrr::some()   masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(broom)

# load data
data <- read.csv("nc_county_health_data_final.csv")

predictors <- c("College_Graduation","Uninsured_Adults",
                "Health_Care_Workforce_._Primary_Care_Physicians",
                "Adult_Smoking","Poverty","Food_Insecurity",
                "Transportation","Air_Pollution")

incidence_outcomes <- c("BreastRate_I", "LungRate_I", "ColorectalRate_I")

# get pooled incidence data for VIF (multicollinearity, probably a problem)
data_i_long <- data %>%
  pivot_longer(
    cols = all_of(incidence_outcomes),
    names_to = "Cancer_Type",
```

```
    values_to = "Incidence_Rate"
  ) %>%
  drop_na(Incidence_Rate)

# Refit ols model
model_pooled_inc_ols <- lm(
  as.formula(paste("Incidence_Rate ~", paste(predictors, collapse=" + "))),
  data=data_i_long
)

# vif calc
vif_values <- vif(model_pooled_inc_ols)
print("VIF Values:")
```

```
## [1] "VIF Values:"
```

```
print(vif_values)
```

```
##                                College_Graduation
##                                          4.045790
##                                  Uninsured_Adults
##                                          1.466314
## Health_Care_Workforce_._Primary_Care_Physicians
##                                          1.909960
##                                     Adult_Smoking
##                                          4.272940
##                                           Poverty
##                                          6.026509
##                                   Food_Insecurity
##                                          7.715591
##                                    Transportation
##                                          2.983949
##                                     Air_Pollution
##                                          1.074449
```

```
# id any vif > 5, looking for cluster from eda
high_vif <- vif_values[vif_values > 5]
print("Problematic VIF (>5):")
```

```
## [1] "Problematic VIF (>5):"
```

```
print(high_vif)
```

```
##         Poverty Food_Insecurity
##        6.026509        7.715591
```

```
# residual plots, save in case needed for paper (if the data is really chopped)
png("diagnostics_residuals.png", width=800, height=600, res=100)
par(mfrow=c(2,2))
plot(model_pooled_inc_ols)
dev.off()
```

```
## pdf
##   2
```

```
# cooks distance
cooksd <- cooks.distance(model_pooled_inc_ols)
n <- nrow(data_i_long)
influential_threshold <- 4/n

# influential cooks
influential_obs <- which(cooksd > influential_threshold)
influential_counties <- data_i_long[influential_obs, ] %>%
  distinct(County) %>%
  pull(County)

print(paste("Influential observations (Cook's D > 4/n):", length(influential_obs)))
```

```
## [1] "Influential observations (Cook's D > 4/n): 9"
```

```
print("Counties flagged as influential:")
```

```
## [1] "Counties flagged as influential:"
```

```
print(influential_counties)
```

```
## [1] "Duplin"     "Durham"     "Onslow"     "Orange"     "Pasquotank"
## [6] "Robeson"    "Scotland"   "Swain"      "Wake"
```

```
# top 5 (if this is wake/mecklenburg, may be problematic)
top5_influential <- order(cooksd, decreasing=TRUE)[1:5]
influential_data <- tibble(
  Observation = top5_influential,
  County = data_i_long$County[top5_influential],
  Cancer_Type = data_i_long$Cancer_Type[top5_influential],
  Cooks_D = cooksd[top5_influential]
)
print("Top 5 influential observations:")
```

```
## [1] "Top 5 influential observations:"
```

```
print(influential_data)
```

```
## # A tibble: 5 x 4
##   Observation County     Cancer_Type  Cooks_D
##         <int> <chr>      <chr>           <dbl>
## 1         227 Robeson    BreastRate_I  0.0322
## 2         197 Orange     BreastRate_I  0.0266
## 3         203 Pasquotank BreastRate_I  0.0253
## 4         254 Swain      BreastRate_I  0.0193
## 5         242 Scotland   BreastRate_I  0.0175
```

```r
# shapiro-wilk test for normality of residuals
shapiro_test <- shapiro.test(residuals(model_pooled_inc_ols))
print("Shapiro-Wilk test for residual normality:")
```

```
## [1] "Shapiro-Wilk test for residual normality:"
```

```r
print(shapiro_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model_pooled_inc_ols)
## W = 0.80865, p-value < 2.2e-16
```

```r
# compare results of modeling
comparison <- tibble(
  Model = c("Full Data"),
  R2 = c(summary(model_pooled_inc_ols)$r.squared),
  Adj_R2 = c(summary(model_pooled_inc_ols)$adj.r.squared),
  N_obs = c(nrow(data_i_long))
)
print("Model Comparison:")
```

```
## [1] "Model Comparison:"
```

```r
print(comparison)
```

```
## # A tibble: 1 x 4
##   Model         R2   Adj_R2 N_obs
##   <chr>      <dbl>    <dbl> <int>
## 1 Full Data 0.0217 -0.00596   292
```

```r
# coef comparison
coef_comparison <- tibble(
  Predictor = names(coef(model_pooled_inc_ols))[-1],
  Full_Data = coef(model_pooled_inc_ols)[-1],
)

print("Coefficient Comparison:")
```

```
## [1] "Coefficient Comparison:"
```

```r
print(coef_comparison)
```

```
## # A tibble: 8 x 2
##   Predictor                                      Full_Data
##   <chr>                                              <dbl>
## 1 College_Graduation                                -8.83
## 2 Uninsured_Adults                                  -0.811
```

```
## 3 Health_Care_Workforce_._Primary_Care_Physicians    2.17
## 4 Adult_Smoking                                      0.0516
## 5 Poverty                                            1.52
## 6 Food_Insecurity                                    1.41
## 7 Transportation                                     0.852
## 8 Air_Pollution                                      4.26
```

```r
# get summary
diagnostics_summary <- list(
  VIF = vif_values,
  High_VIF = high_vif,
  N_Influential = length(influential_obs),
  Influential_Counties = unique(influential_counties),
  Shapiro_p = shapiro_test$p.value,
  Sensitivity = comparison
)

# show
diagnostics_summary
```

```
## $VIF
##                                 College_Graduation
##                                           4.045790
##                                   Uninsured_Adults
##                                           1.466314
## Health_Care_Workforce_._Primary_Care_Physicians
##                                           1.909960
##                                      Adult_Smoking
##                                           4.272940
##                                            Poverty
##                                           6.026509
##                                     Food_Insecurity
##                                           7.715591
##                                      Transportation
##                                           2.983949
##                                       Air_Pollution
##                                           1.074449
##
## $High_VIF
##         Poverty Food_Insecurity
##        6.026509        7.715591
##
## $N_Influential
## [1] 9
##
## $Influential_Counties
## [1] "Duplin"     "Durham"     "Onslow"     "Orange"     "Pasquotank"
## [6] "Robeson"    "Scotland"   "Swain"      "Wake"
##
## $Shapiro_p
## [1] 2.907968e-18
##
## $Sensitivity
## # A tibble: 1 x 4
```

```
##   Model          R2   Adj_R2 N_obs
##   <chr>       <dbl>    <dbl> <int>
## 1 Full Data 0.0217 -0.00596   292
```