generalized_results

Madison Thompson

2025-12-09

Running OLS model for Incidence

```
# imports
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()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(dplyr)
library(tidyr)
library(broom)

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

# names(data)

# declare variables
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")

# pool indicators
data_i_long <- data %>%
  pivot_longer(
    cols = all_of(incidence_outcomes),
    names_to = "Cancer_Type",
    values_to = "Incidence_Rate"
  ) %>%
  drop_na(Incidence_Rate)
```

```r
# head(data_i_long)

# run OLS for incidence
model_pooled_inc_ols <- lm(
  as.formula(paste("Incidence_Rate ~", paste(predictors, collapse=" + "))),
  data=data_i_long
)

summary(model_pooled_inc_ols)
```

```
##
## Call:
## lm(formula = as.formula(paste("Incidence_Rate ~", paste(predictors,
##     collapse = " + "))), data = data_i_long)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -80.14 -57.16 -32.46  78.71 137.06
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   103.06948    4.07443  25.297
## College_Graduation                             -8.83101    8.40580  -1.051
## Uninsured_Adults                               -0.81069    4.99294  -0.162
## Health_Care_Workforce_._Primary_Care_Physicians   2.16936    5.70498   0.380
## Adult_Smoking                                   0.05161    8.48870   0.006
## Poverty                                         1.51539   10.14308   0.149
## Food_Insecurity                                 1.40609   11.53270   0.122
## Transportation                                  0.85174    7.03563   0.121
## Air_Pollution                                   4.26337    4.38678   0.972
##                                                Pr(>|t|)
## (Intercept)                                     <2e-16 ***
## College_Graduation                               0.294
## Uninsured_Adults                                 0.871
## Health_Care_Workforce_._Primary_Care_Physicians    0.704
## Adult_Smoking                                    0.995
## Poverty                                          0.881
## Food_Insecurity                                  0.903
## Transportation                                   0.904
## Air_Pollution                                    0.332
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 69.46 on 283 degrees of freedom
## Multiple R-squared:  0.0217, Adjusted R-squared:  -0.005957
## F-statistic: 0.7846 on 8 and 283 DF,  p-value: 0.6166
```

```r
# getting statistical measures
pooled_results <- tibble(
  R2 = summary(model_pooled_inc_ols)$r.squared,
  Adj_R2 = summary(model_pooled_inc_ols)$adj.r.squared,
  AIC = AIC(model_pooled_inc_ols)
)
```

```
pooled_results
```

```
## # A tibble: 1 x 3
##       R2   Adj_R2   AIC
##    <dbl>    <dbl> <dbl>
## 1 0.0217 -0.00596 3316.
```

```
broom::tidy(model_pooled_inc_ols)
```

```
## # A tibble: 9 x 5
##   term                                    estimate std.error statistic  p.value
##   <chr>                                      <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                               103.       4.07   25.3     1.29e-74
## 2 College_Graduation                         -8.83     8.41   -1.05    2.94e- 1
## 3 Uninsured_Adults                           -0.811    4.99   -0.162   8.71e- 1
## 4 Health_Care_Workforce_._Primary_Care_Ph~    2.17     5.70    0.380   7.04e- 1
## 5 Adult_Smoking                               0.0516   8.49    0.00608 9.95e- 1
## 6 Poverty                                     1.52    10.1     0.149   8.81e- 1
## 7 Food_Insecurity                             1.41    11.5     0.122   9.03e- 1
## 8 Transportation                              0.852    7.04    0.121   9.04e- 1
## 9 Air_Pollution                               4.26     4.39    0.972   3.32e- 1
```

Running Ridge model for Incidence

```
# imports
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-10
```

```
# clean again
ridge_df <- data_i_long %>%
  drop_na(Incidence_Rate, all_of(predictors))

# prepare data for glmnet
x <- model.matrix(
  ~ College_Graduation + Uninsured_Adults +
    Health_Care_Workforce_._Primary_Care_Physicians +
    Adult_Smoking + Poverty + Food_Insecurity +
    Transportation + Air_Pollution,
  data = ridge_df
)[, -1]
```

```r
y <- ridge_df$Incidence_Rate

# run ridge regression with cross-validation to find optimal lambda
set.seed(123)
cv_ridge <- cv.glmnet(x, y, alpha = 0, nfolds = 10)

# get the best lambda
best_lambda <- cv_ridge$lambda.min

# fit the final ridge model
ridge_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)

# get coefficients
ridge_coefficients <- as.matrix(coef(ridge_model))
#ridge_coefficients

ridge_coefficients_tidy <- tibble(
  term = rownames(ridge_coefficients),
  estimate = as.numeric(ridge_coefficients)
)
ridge_coefficients_tidy
```

```
## # A tibble: 9 x 2
##   term                                              estimate
##   <chr>                                                <dbl>
## 1 (Intercept)                                        103.
## 2 College_Graduation                                  -0.825
## 3 Uninsured_Adults                                     0.232
## 4 Health_Care_Workforce_._Primary_Care_Physicians     -0.280
## 5 Adult_Smoking                                        0.672
## 6 Poverty                                              0.573
## 7 Food_Insecurity                                      0.693
## 8 Transportation                                       0.578
## 9 Air_Pollution                                        0.421
```

```r
# getting statistical measures
R2 <- ridge_model$dev.ratio
p_df <- ridge_model$df
n_obs <- length(y)
Adj_R2 <- 1 - (1 - R2) * ((n_obs - 1) / (n_obs - p_df - 1))

# get AIC
TSS <- sum((y - mean(y))^2)
RSS <- TSS * (1 - R2)
AIC_approx <- n_obs * log(RSS / n_obs) + 2 * p_df

# results in tibble
ridge_results <- tibble(
  R2 = R2,
  Adj_R2 = Adj_R2,
  AIC = AIC_approx
)
```

```
ridge_results
```

```
## # A tibble: 1 x 3
##        R2  Adj_R2   AIC
##     <dbl>   <dbl> <dbl>
## 1 0.00976 -0.0182 2487.
```

Running OLS model for Mortality

```
# imports
library(tidyverse)
library(dplyr)
library(tidyr)
library(broom)

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

# names(data)

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

mortality_outcomes <- c("BreastRate_M", "LungRate_M", "ColorectalRate_M")

# pool indicators
data_m_long <- data %>%
  pivot_longer(
    cols = all_of(mortality_outcomes),
    names_to = "Cancer_Type",
    values_to = "Mortality_Rate"
  ) %>%
  drop_na(Mortality_Rate)

# head(data_m_long)

# run OLS for incidence
model_pooled_mor_ols <- lm(
  as.formula(paste("Mortality_Rate ~", paste(predictors, collapse=" + "))),
  data=data_m_long
)

summary(model_pooled_mor_ols)
```

```
##
## Call:
## lm(formula = as.formula(paste("Mortality_Rate ~", paste(predictors,
##     collapse = " + "))), data = data_m_long)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.630 -11.608  -4.911  12.165  28.098
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                     26.5875     0.8629  30.811
## College_Graduation                              -3.2423     1.7576  -1.845
## Uninsured_Adults                                -0.5915     1.0926  -0.541
## Health_Care_Workforce_._Primary_Care_Physicians  0.5719     1.2163   0.470
## Adult_Smoking                                    0.5745     1.8092   0.318
## Poverty                                          0.1223     2.1489   0.057
## Food_Insecurity                                  1.1138     2.4658   0.452
## Transportation                                  -0.7391     1.5234  -0.485
## Air_Pollution                                   -0.2698     0.9273  -0.291
##                                                Pr(>|t|)
## (Intercept)                                      <2e-16 ***
## College_Graduation                               0.0664 .
## Uninsured_Adults                                 0.5888
## Health_Care_Workforce_._Primary_Care_Physicians  0.6387
## Adult_Smoking                                    0.7511
## Poverty                                          0.9547
## Food_Insecurity                                  0.6519
## Transportation                                   0.6280
## Air_Pollution                                    0.7714
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.06 on 225 degrees of freedom
## Multiple R-squared:  0.07212,    Adjusted R-squared:  0.03913
## F-statistic: 2.186 on 8 and 225 DF,  p-value: 0.02944
```

```
# getting statistical measures
pooled_results <- tibble(
  R2 = summary(model_pooled_mor_ols)$r.squared,
  Adj_R2 = summary(model_pooled_mor_ols)$adj.r.squared,
  AIC = AIC(model_pooled_mor_ols)
)

pooled_results
```

```
## # A tibble: 1 x 3
##       R2 Adj_R2   AIC
##    <dbl>  <dbl> <dbl>
## 1 0.0721 0.0391 1877.
```

```
broom::tidy(model_pooled_mor_ols)
```

```
## # A tibble: 9 x 5
##   term                                  estimate std.error statistic  p.value
##   <chr>                                    <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)                             26.6       0.863   30.8    1.10e-82
## 2 College_Graduation                      -3.24      1.76    -1.84   6.64e- 2
```

```
## 3 Uninsured_Adults                              -0.591     1.09    -0.541  5.89e- 1
## 4 Health_Care_Workforce_._Primary_Care_Ph~      0.572     1.22     0.470  6.39e- 1
## 5 Adult_Smoking                                 0.574     1.81     0.318  7.51e- 1
## 6 Poverty                                       0.122     2.15     0.0569 9.55e- 1
## 7 Food_Insecurity                               1.11      2.47     0.452  6.52e- 1
## 8 Transportation                               -0.739     1.52    -0.485  6.28e- 1
## 9 Air_Pollution                                -0.270     0.927   -0.291  7.71e- 1
```

Running Ridge model for Mortality

```
# imports
library(glmnet)

# clean again
ridge_df_m <- data_m_long %>%
  drop_na(Mortality_Rate, all_of(predictors))

# prepare data for glmnet
x <- model.matrix(
  ~ College_Graduation + Uninsured_Adults +
    Health_Care_Workforce_._Primary_Care_Physicians +
    Adult_Smoking + Poverty + Food_Insecurity +
    Transportation + Air_Pollution,
  data = ridge_df_m
)[, -1]

y <- ridge_df_m$Mortality_Rate

# run ridge regression with cross-validation to find optimal lambda
set.seed(123)
cv_ridge <- cv.glmnet(x, y, alpha = 0, nfolds = 10)

# get the best lambda
best_lambda <- cv_ridge$lambda.min

# fit the final ridge model
ridge_model <- glmnet(x, y, alpha = 0, lambda = best_lambda)

# get coefficients
ridge_coefficients <- as.matrix(coef(ridge_model))
#ridge_coefficients

ridge_coefficients_tidy <- tibble(
  term = rownames(ridge_coefficients),
  estimate = as.numeric(ridge_coefficients)
)
ridge_coefficients_tidy
```

```
## # A tibble: 9 x 2
##   term                                          estimate
##   <chr>                                            <dbl>
## 1 (Intercept)                                     26.5
## 2 College_Graduation                              -1.01
```

```
## 3 Uninsured_Adults                                      0.0846
## 4 Health_Care_Workforce_._Primary_Care_Physicians      -0.323
## 5 Adult_Smoking                                         0.565
## 6 Poverty                                               0.312
## 7 Food_Insecurity                                       0.544
## 8 Transportation                                        0.146
## 9 Air_Pollution                                        -0.167
```

```r
# getting statistical measures
R2 <- ridge_model$dev.ratio
p_df <- ridge_model$df
n_obs <- length(y)
Adj_R2 <- 1 - (1 - R2) * ((n_obs - 1) / (n_obs - p_df - 1))

# get AIC
TSS <- sum((y - mean(y))^2)
RSS <- TSS * (1 - R2)
AIC_approx <- n_obs * log(RSS / n_obs) + 2 * p_df

# results in tibble
ridge_results <- tibble(
  R2 = R2,
  Adj_R2 = Adj_R2,
  AIC = AIC_approx
)

ridge_results
```

```
## # A tibble: 1 x 3
##       R2 Adj_R2   AIC
##    <dbl>  <dbl> <dbl>
## 1 0.0584 0.0249 1213.
```