lasso_results

Madison Thompson

2025-12-10

LASSO for loop that adds all data to one table to be reported in the paper

```
# 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(glmnet)
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loaded glmnet 4.1-10
```

```
library(broom)

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

# set predictors, same for all models, pruned by LASSO
predictors <- c("College_Graduation","Uninsured_Adults",
                "Health_Care_Workforce_._Primary_Care_Physicians",
                "Adult_Smoking","Poverty","Food_Insecurity",
                "Transportation","Air_Pollution")

# all 6 cancer outcomes, selected for in for loop
```

```r
lasso_outcomes <- c("BreastRate_I", "BreastRate_M",
                    "LungRate_I", "LungRate_M",
                    "ColorectalRate_I", "ColorectalRate_M")

# storage
lasso_results_list <- list()
coef_list <- list()

set.seed(123)

# loop to run lasso for each outcome and synthesize in one table
for (outcome in lasso_outcomes) {

  # prepare data for each model
  model_df <- data %>%
    select(all_of(c(outcome, predictors))) %>%
    drop_na()

  x <- as.matrix(model_df[, predictors])
  y <- model_df[[outcome]]

  # fit lasso w/10-fold cv
  cvfit <- cv.glmnet(x, y, alpha = 1, nfolds = 10)
  best_lambda <- cvfit$lambda.min

  final_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)

  # coefficients
  coef_mat <- as.matrix(coef(final_model))
  coef_df <- tibble(
    Outcome = outcome,
    term = rownames(coef_mat),
    estimate = as.numeric(coef_mat)
  )

  coef_list[[outcome]] <- coef_df

  # get cv-mse
  cv_mse <- min(cvfit$cvm)

  # calculate r-squared
  tss <- var(y) * (length(y) - 1)
  r2_cv <- 1 - (cv_mse / var(y))

# Store the cv r-squared
r2 <- r2_cv

  cv_mse <- min(cvfit$cvm)

  # retained predictors
  retained <- coef_df %>%
    filter(term != "(Intercept)" & estimate != 0) %>%
    pull(term)
```

```
  # save summary row
  lasso_results_list[[outcome]] <- tibble(
    Outcome = outcome,
    Lambda_min = best_lambda,
    R2 = r2,
    CV_MSE = cv_mse,
    N_retained = length(retained),
    Retained_Predictors = ifelse(
      length(retained) == 0,
      "None",
      paste(retained, collapse = "; ")
    )
  )
}

# bind all results into one table
lasso_summary_table <- bind_rows(lasso_results_list)

# print the big summary table
lasso_summary_table
```

```
## # A tibble: 6 x 6
##   Outcome          Lambda_min     R2 CV_MSE N_retained Retained_Predictors
##   <chr>                 <dbl>  <dbl>  <dbl>      <int> <chr>
## 1 BreastRate_I         0.358  0.394  379.            6 College_Graduation; Unin~
## 2 BreastRate_M         1.10   0.0880  17.5           1 Food_Insecurity
## 3 LungRate_I           0.133  0.489   83.6           7 College_Graduation; Unin~
## 4 LungRate_M           0.0918 0.469   31.4           7 College_Graduation; Unin~
## 5 ColorectalRate_I     0.142  0.496   31.9           8 College_Graduation; Unin~
## 6 ColorectalRate_M     0.327  0.242    8.16          5 College_Graduation; Adul~
```

```
# getting values for retained vars
all_coefs <- bind_rows(coef_list)

# pivot wide to match paper format
coef_wide <- all_coefs %>%
  filter(term != "(Intercept)") %>%
  mutate(
    Outcome = case_when(
      Outcome == "BreastRate_I" ~ "Breast_Inc",
      Outcome == "BreastRate_M" ~ "Breast_Mort",
      Outcome == "LungRate_I" ~ "Lung_Inc",
      Outcome == "LungRate_M" ~ "Lung_Mort",
      Outcome == "ColorectalRate_I" ~ "Colon_Inc",
      Outcome == "ColorectalRate_M" ~ "Colon_Mort"
    ),
    term = case_when(
      term == "College_Graduation" ~ "College Graduation",
      term == "Uninsured_Adults" ~ "Uninsured Adults",
      term == "Health_Care_Workforce_._Primary_Care_Physicians" ~ "PCPs",
      term == "Adult_Smoking" ~ "Adult Smoking",
      term == "Food_Insecurity" ~ "Food Insecurity",
      term == "Air_Pollution" ~ "Air Pollution",
```

```
      TRUE ~ term
    ),
    estimate = round(estimate, 2)
  ) %>%
  pivot_wider(
    names_from = Outcome,
    values_from = estimate,
    values_fill = 0
  ) %>%
  arrange(term)

# output
coef_wide
```

```
## # A tibble: 8 x 7
##   term            Breast_Inc Breast_Mort Lung_Inc Lung_Mort Colon_Inc Colon_Mort
##   <chr>                <dbl>       <dbl>    <dbl>     <dbl>     <dbl>      <dbl>
## 1 Adult Smoking         2.43        0        0.83      2.33      1.1        0.12
## 2 Air Pollution        11.0         0        1.55      1.68      0.54      -0.05
## 3 College Gradua~      -9.62        0       -8.34     -5.4      -3.83      -0.89
## 4 Food Insecurity       0           1.06     2.06      0         0.21       0.57
## 5 PCPs                  3.38        0        0.86      0.86      0.03       0
## 6 Poverty               0           0        0        -0.58      0.99       0.12
## 7 Transportation        4.7         0        0.24     -0.65      1.33       0
## 8 Uninsured Adul~       0.35        0       -1.9      -1.53     -0.9        0
```

Plotting data for paper, trying to get a figure that shows incidence v mortality because its the major finding

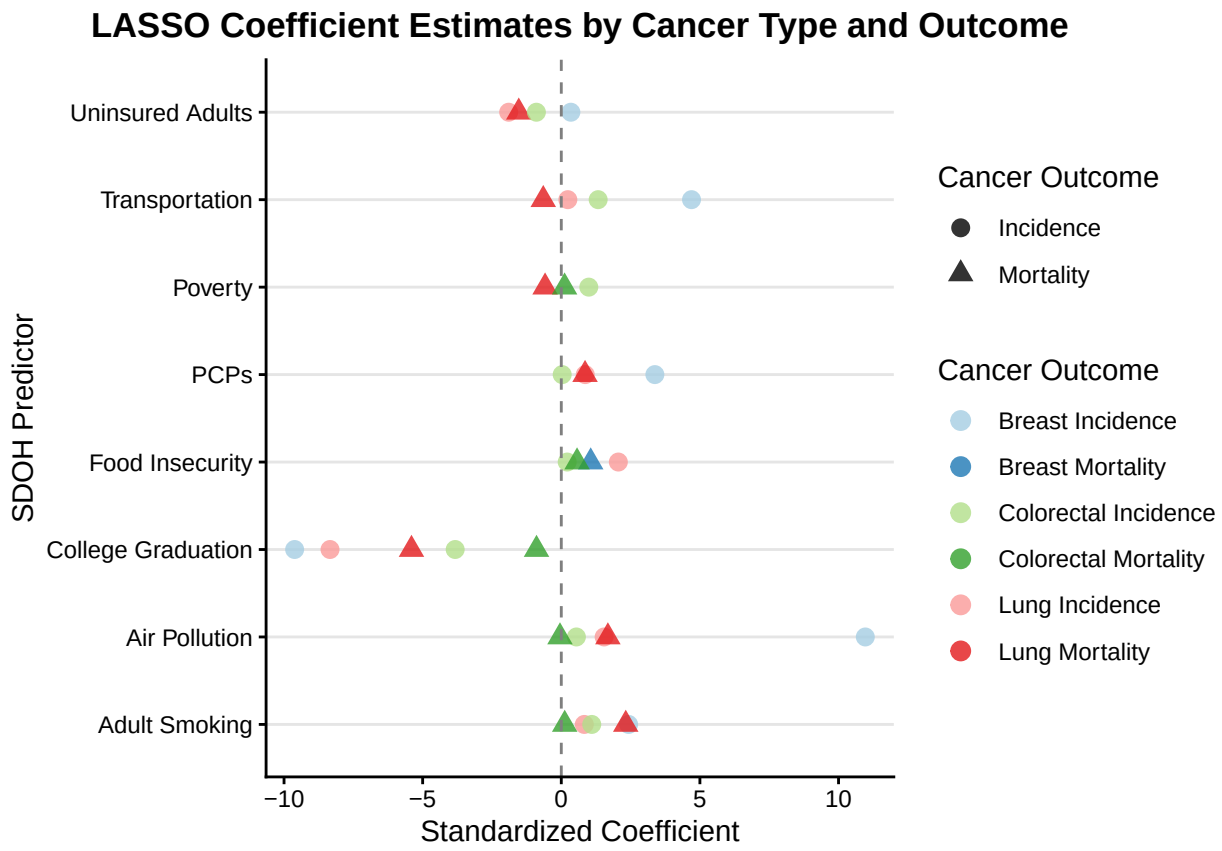
```
# combine all coefs
all_coefs <- bind_rows(coef_list)

# remove non-meaningful variables
coef_plot_data <- all_coefs %>%
  filter(term != "(Intercept)") %>%
  filter(estimate != 0) %>%
  mutate(
    Outcome = case_when(
      Outcome == "BreastRate_I" ~ "Breast Incidence",
      Outcome == "BreastRate_M" ~ "Breast Mortality",
      Outcome == "LungRate_I" ~ "Lung Incidence",
      Outcome == "LungRate_M" ~ "Lung Mortality",
      Outcome == "ColorectalRate_I" ~ "Colorectal Incidence",
      Outcome == "ColorectalRate_M" ~ "Colorectal Mortality"
    ),
    term = case_when(
      term == "College_Graduation" ~ "College Graduation",
      term == "Uninsured_Adults" ~ "Uninsured Adults",
      term == "Health_Care_Workforce_._Primary_Care_Physicians" ~ "PCPs",
      term == "Adult_Smoking" ~ "Adult Smoking",
      term == "Food_Insecurity" ~ "Food Insecurity",
      term == "Air_Pollution" ~ "Air Pollution",
      TRUE ~ term
    )
```

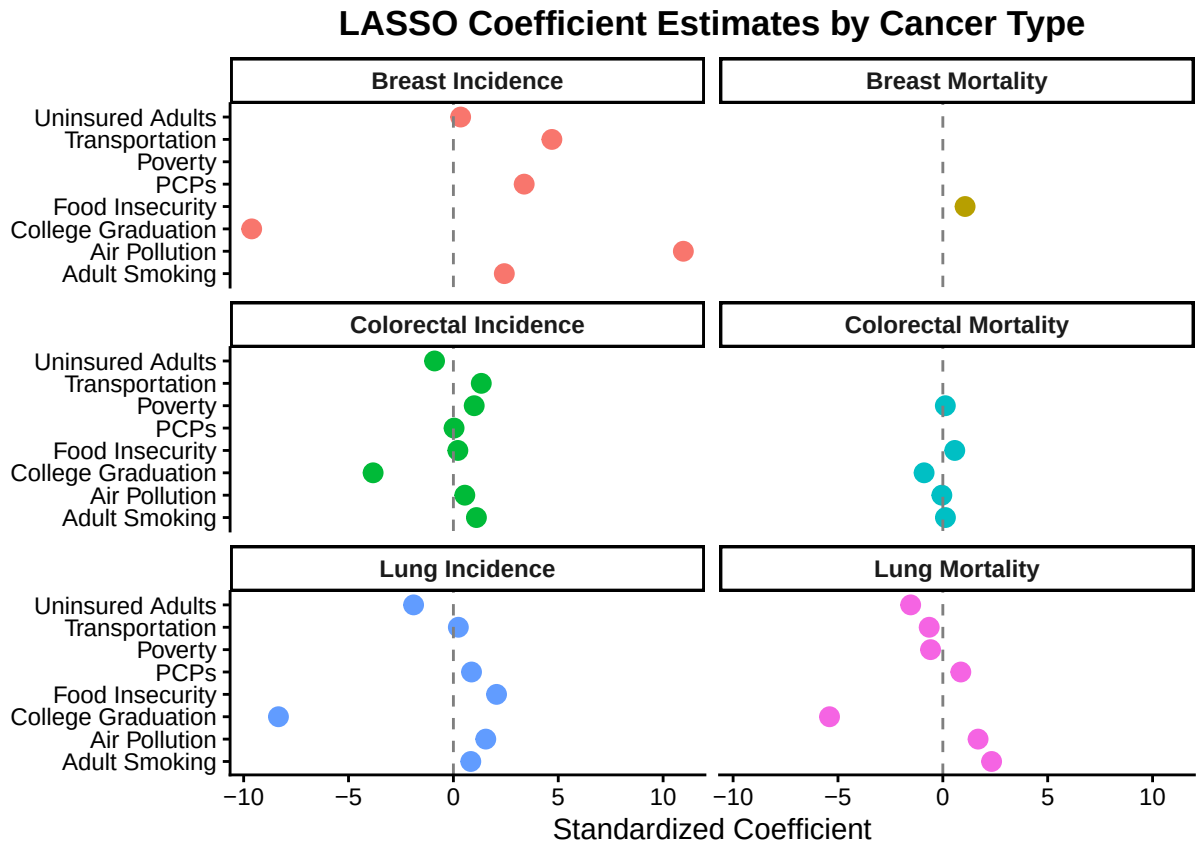
```
  )

# grouping by type to make shapes the same for outcome type
coef_plot_data <- coef_plot_data %>%
  mutate(
    Type = ifelse(grepl("Incidence", Outcome), "Incidence", "Mortality")
  )

# plot
ggplot(coef_plot_data, aes(x = estimate, y = term, color = Outcome, shape = Type)) +
  geom_point(size = 3, alpha = 0.8) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
  labs(
    title = "LASSO Coefficient Estimates by Cancer Type and Outcome",
    x = "Standardized Coefficient",
    y = "SDOH Predictor",
    color = "Cancer Outcome",
    shape = "Cancer Outcome"
  ) +
  theme_classic() +
  theme(
    legend.position = "right",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.major.y = element_line(color = "gray90"),
    panel.grid.minor = element_blank()
  ) +
  scale_color_brewer(palette = "Paired")
```




```r
# plot each cancer outcome on its own
ggplot(coef_plot_data, aes(x = estimate, y = term, color = Outcome)) +
  geom_point(size = 3) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
  facet_wrap(~ Outcome, ncol = 2) +
  labs(
    title = "LASSO Coefficient Estimates by Cancer Type",
    x = "Standardized Coefficient",
    y = ""
  ) +
  theme_classic() +
  theme(
    legend.position = "none",
    plot.title = element_text(hjust = 0.5, face = "bold"),
    strip.text = element_text(face = "bold")
  )
```




```r
# grouping data
coef_plot_data <- coef_plot_data %>%
  mutate(
    Cancer_Type = case_when(
      str_detect(Outcome, "Breast") ~ "Breast",
      str_detect(Outcome, "Lung") ~ "Lung",
      str_detect(Outcome, "Colorectal") ~ "Colorectal"
    ),
    Outcome_Type = ifelse(str_detect(Outcome, "Incidence"), "Incidence", "Mortality")
  )

# facted plot to spot within-cancer trends
ggplot(coef_plot_data, aes(x = estimate, y = term,
                           color = Outcome_Type, shape = Outcome_Type)) +
  geom_point(size = 3.5, alpha = 0.8) +
  geom_vline(xintercept = 0, linetype = "dashed", color = "gray50") +
  facet_wrap(~ Cancer_Type, ncol = 3) +
  labs(
    title = "LASSO Coefficient Estimates by Cancer Type",
    x = "Standardized Coefficient",
    y = "",
    color = "",
    shape = ""
  ) +
  theme_classic() +
  theme(
```

```
    legend.position = "bottom",
    plot.title = element_text(hjust = 0.5, face = "bold", size = 14),
    strip.text = element_text(face = "bold", size = 11),
    panel.grid.major.y = element_line(color = "gray90"),
    panel.grid.minor = element_blank()
  ) +
  scale_color_manual(values = c("Incidence" = "purple", "Mortality" = "red"))
```