

Milestone

2025-04-09

Setup

Install packages and data

Installing important packages for the project and downloading the dataset

```
install.packages("survey")
```

```
##  
## The downloaded binary packages are in  
## /var/folders/4r/kdq7r6r51dx66ggpy4zs5cfh0000gn/T//RtmpeIW3bG/downloaded_packages
```

```
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(srvyr)
```

```
##  
## Attaching package: 'srvyr'  
##  
## The following object is masked from 'package:stats':  
##  
## filter
```

```
library(gt)  
library(tinytex)  
library(naniar)  
library(ggthemes)  
library(glmnet)
```

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

```
##
```

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

```
##
```

```
## The following objects are masked from 'package:tidyr':
```

```
##
```

```
## expand, pack, unpack
```

```
##
```

```

## Loaded glmnet 4.1-8
library(olsrr)

##
## Attaching package: 'olsrr'
##
## The following object is masked from 'package:datasets':
##
##     rivers
library(car)

## Loading required package: carData
##
## Attaching package: 'car'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following object is masked from 'package:purrr':
##
##     some
library(survey)

## Loading required package: grid
## Loading required package: survival
##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##     dotchart
library(gt)
library(tidyr)
library(caret)

## Loading required package: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:survival':
##
##     cluster
##
## The following object is masked from 'package:purrr':
##
##     lift
dhs <- read_csv("~/Documents/GitHub/bios/453/bios453/IAIR7EFL.csv")

## Warning: One or more parsing issues, call `problems()` on your data frame for details,
## e.g.:
##   dat <- vroom(...)
##   problems(dat)

```

```
## Rows: 724115 Columns: 423
## -- Column specification -----
## Delimiter: ","
## chr (1): caseid
## dbl (277): v001, v002, v003, v004, v005, v012, v015, v020, v021, v022, v023,...
## lgl (145): midx_4, midx_5, midx_6, m3a_4, m3a_5, m3a_6, m2n_2, m2n_3, m2n_4,...
##
## 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.
```

Description of data management utilized and cleaning

As part of the data management process for this project, I significantly reduced its size to include only variables that I believe are relevant for my research question. This includes all of the relevant variables, in addition to each confounder listed in my concept map (both primary and secondary). Once I had each of these variables selected, I went through and ensured that they were all formatted correctly for the types of regression I wanted to run during my analysis. I also made sure that certain items were filtered out. For example, some continuous reportings like birth weight in kg clearly contained severe misinputs. Since I am also only concerned with it as a numerical variable, I removed some of the categorical sections on the upper end like not weighed at birth, don't know, or missing. Additionally, I chose to adhere to UNICEF's validity threshold: $250g \leq \text{birthweight} \leq 5,500g$ for a more realistic analysis. I also chose to keep variables like time pregnant and age numeric because that makes sense given what I am interested in exploring. Binary variables like whether or not someone took antiparasitics were limited to only response values, as I am not interested in those who couldn't answer the question. This leaves us with a hefty 153,582 observations for analysis. Finally, I included a scaled weight column for women's sample weights due to the nature of my analysis (centering around women's reproductive health). I scaled it by 1,000,000 because DHS weights are stored as six digit strings, and scaling is necessary to prevent serious distortions, getting the weights into their correct format.

```
dhs_clean <- dhs %>%
  select(v005, m60_1, v481, m18_1, m19_1, m17_1, v457, m15_1, v228, v208, v190a, v190, v149, v131, v024)
  rename(
    ids = v021,
    strata = v023,
    wweight = v005,
    paradrug = m60_1,
    insured = v481,
    sizechild = m18_1,
    bwkg = m19_1,
    csect = m17_1,
    anemia = v457,
    delivplace = m15_1,
    termpreg = v228,
    bpast5 = v208,
    wind_urbrur = v190a,
    wind = v190,
    educ = v149,
    ethnic = v131,
    state = v024,
    smokes = v463aa,
    married = v501,
    age = v447a,
  ) %>%
  filter(
    paradrug != 8,
```

```

paradrug!=9,
bwkg <5501,
bwkg > 249,
bwkg != 9996,
bwkg != 9998,
bwkg != 9999,
sizechild!=8,
insured != 9,
sizechild != 9,
csect != 9,
anemia != 9,
termpreg != 9,
educ != 9,
smokes != 9,
married != 9,
age != 99,
) %>%
mutate(
  wswweight = wswweight/1000000,
  paradrug = as.factor(paradrug),
  insured = as.factor(insured),
  sizechild = factor(sizechild,
    levels = c("very large" = "1",
               "larger than average" = "2",
               "average" = "3",
               "smaller than average" = "4",
               "very small" = "5")),
  csect = factor(csect),
  anemia = factor(anemia,
    levels = c("severe" = "1",
               "moderate" = "2",
               "mild" = "3",
               "not anemic" = "4")),
  delivplace = factor(delivplace),
  termpreg = factor(termpreg),
  bpast5 = factor(bpast5),
  wind = factor(wind,
    levels = c("poorest" = "1",
               "poorer" = "2",
               "middle" = "3",
               "richer" = "4",
               "richest" = "5")),
  educ = factor(educ,
    levels = c("none" = "0",
               "incomplete primary" = "1",
               "complete primary" = "2",
               "incomplete secondary" = "3",
               "complete secondary" = "4",
               "higher" = "5")),
  ethnic = factor(ethnic,
    levels = c("caste" = "991",
               "tribe" = "992",
               "no caste/tribe" = "993",

```

```

    "dont know" = "998")),
  smokes = factor(smokes)
)

```

Missing data report

According to a missing data summary, there are no missing observations within the cleaned dataset. After removing “dont know” values from majority of the variables of interest (due to their being irrelevant) we no longer have any data that are not available and a final cleaned dataset of 153582 observations.

```

miss_var_summary(dhs_clean) %>%
  arrange(desc(pct_miss))

```

```

## # A tibble: 20 x 3
##   variable    n_miss pct_miss
##   <chr>      <int>   <num>
## 1 wweight         0         0
## 2 paradrug        0         0
## 3 insured         0         0
## 4 sizechild       0         0
## 5 bwkg           0         0
## 6 csect          0         0
## 7 anemia         0         0
## 8 delivplace      0         0
## 9 termpreg       0         0
## 10 bpast5         0         0
## 11 wind_urbrur    0         0
## 12 wind          0         0
## 13 educ          0         0
## 14 ethnic         0         0
## 15 state         0         0
## 16 smokes        0         0
## 17 married       0         0
## 18 age           0         0
## 19 ids           0         0
## 20 strata        0         0

```

Concept map

Summary table of characteristics

```

# categorical summary
cat_summary <- dhs_clean %>%
  select(paradrug, insured, educ) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "category") %>%
  group_by(variable) %>%
  mutate(total = n()) %>%
  group_by(variable, category) %>%
  summarise(
    count = n(),
    percent = paste0(round(100 * count / first(total), 1), "%"),
    .groups = "drop"
  ) %>%
  arrange(variable, category) %>%

```

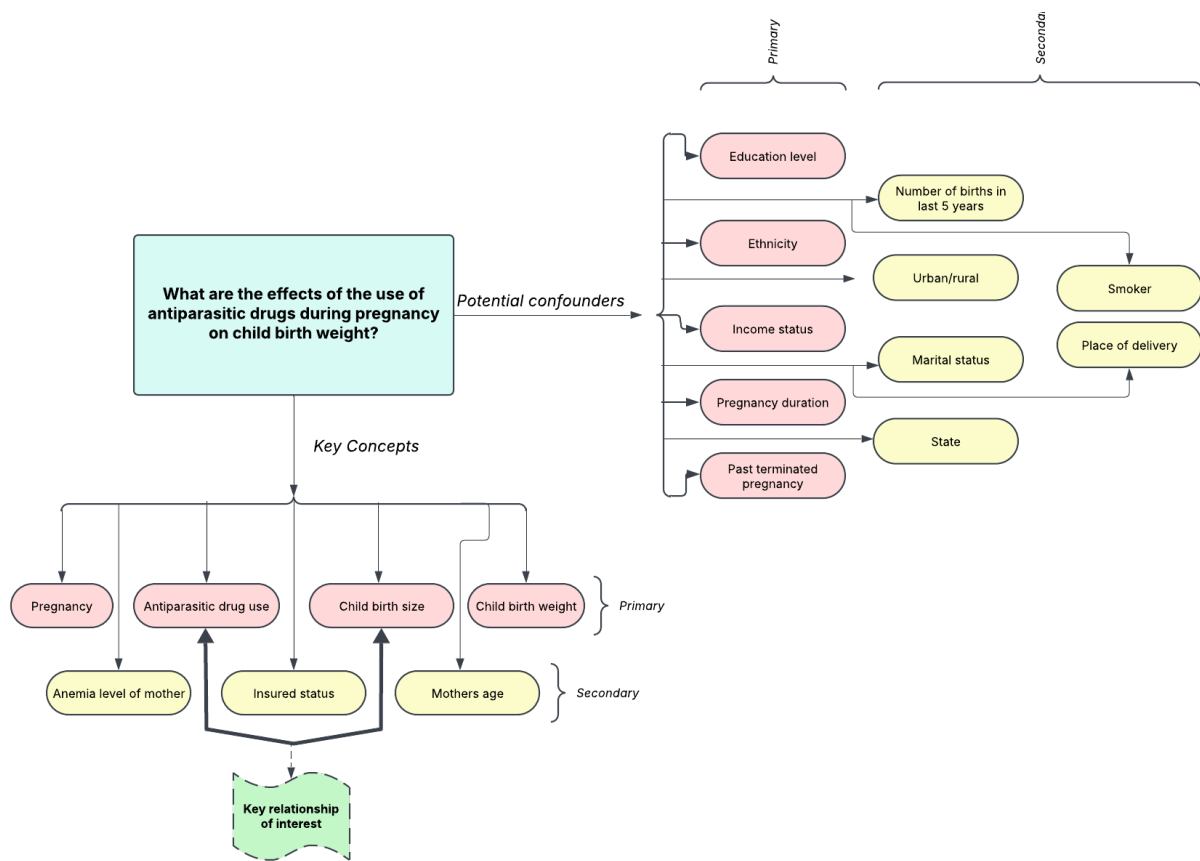


Figure 1: Concept Map

```

group_by(variable) %>%
mutate(row_id = row_number()) %>%
ungroup()

# summary for numerical variables
num_summary <- dhs_clean %>%
  summarise(across(c(age, bwkg),
    list(mean = ~round(mean(., na.rm = TRUE), 2),
          median = ~round(median(., na.rm = TRUE), 2),
          sd = ~round(sd(., na.rm = TRUE), 2)),
    .names = "{.col}_{.fn}")) %>%
  pivot_longer(everything(), names_to = c("variable", "statistic"), names_sep = "_") %>%
  pivot_wider(names_from = statistic, values_from = value) %>%
  mutate(row_id = 1)

# merge and gt table
bind_rows(
  cat_summary,
  num_summary %>% mutate(category = NA, count = NA, percent = NA)
) %>%
  arrange(factor(variable, levels = c("paradrug", "insured", "educ", "age", "bwkg"))) %>%
  gt(groupname_col = "variable") %>%
  tab_header(
    title = "Summary Statistics for DHS Dataset",
  ) %>%
  cols_label(
    category = "Category",
    count = "Count",
    percent = "Percentage",
    mean = "Mean",
    median = "Median",
    sd = "Std Dev"
  ) %>%
  fmt_number(
    columns = c(mean, median, sd),
    decimals = 2
  ) %>%
  cols_align(
    align = "center",
    columns = c(category, count, percent, mean, median, sd)
  ) %>%
  tab_style(
    style = cell_text(weight = "bold"),
    locations = cells_column_labels()
  ) %>%
  tab_style(
    style = cell_fill(color = "#C5DECD"),
    locations = cells_body(
      rows = variable %in% c("paradrug", "insured", "educ")
    )
  ) %>%
  fmt_missing(columns = everything(), missing_text = "--") %>%
  cols_hide(columns = c(row_id))%>%

```

Summary Statistics for DHS Dataset

Category	Count	Percentage	Mean	Median	Std Dev
paradrug					
0	104771	68.2%	-	-	-
1	48811	31.8%	-	-	-
insured					
0	109570	71.3%	-	-	-
1	44012	28.7%	-	-	-
educ					
0	28084	18.3%	-	-	-
1	18238	11.9%	-	-	-
3	80116	52.2%	-	-	-
4	3000	2%	-	-	-
5	24144	15.7%	-	-	-
age					
-	-	-	27.38	27.00	5.12
bwkg					
-	-	-	2,816.83	2,900.00	554.07

```
tab_options(table.background.color = "#F1F7ED")
```

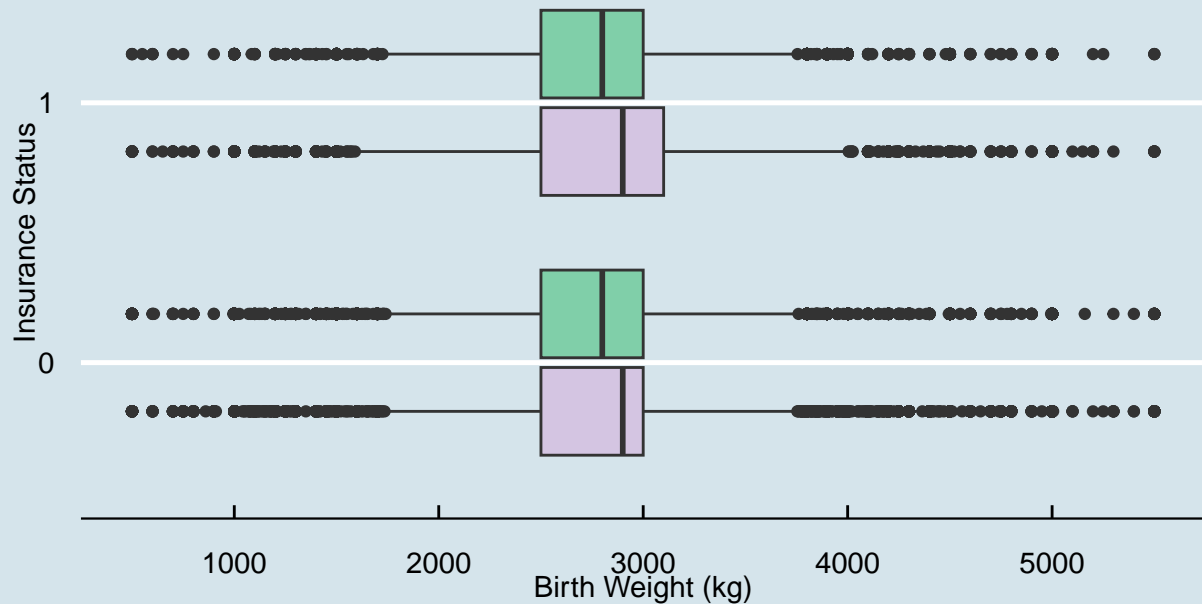
```
## Warning: Since gt v0.6.0 `fmt_missing()` is deprecated and will soon be removed.
## i Use `sub_missing()` instead.
## This warning is displayed once every 8 hours.
```

Graphical representation of characteristics

```
# boxplot comparing birthweights by antiparasitic use and insurance status
ggplot(data = dhs_clean, mapping = aes(y = insured, x = bwkg, fill = paradrug)) +
  geom_boxplot() +
  theme_economist() +
  scale_fill_manual(values = c("#D4C5E2", "#80CFA9")) +
  labs(title = "Birthweight (kg) by Antiparasitics and Insured Status",
       y = "Insurance Status",
       x = "Birth Weight (kg)",
       fill = "Antiparasitics While Pregnant")
```


Birthweight (kg) by Antiparasitics and Insured Status

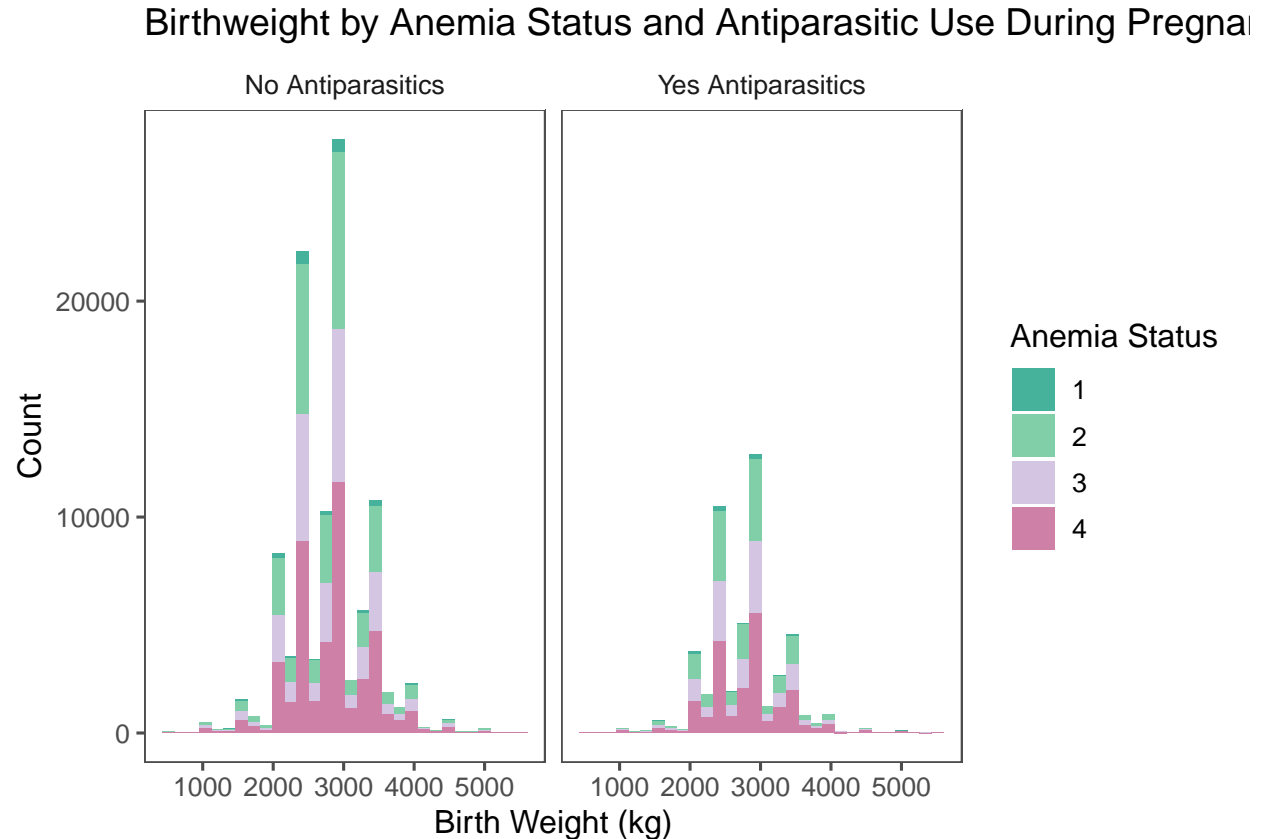
Antiparasitics While Pregnant 0 1



```
# make labels for facetwrap
paradrug_labels <- c("0" = "No Antiparasitics", "1" = "Yes Antiparasitics")

# barplot comparing birthweight by anemia status and antiparasitic use
ggplot(dhs_clean, aes(x = bwkg, fill = factor(anemia))) +
  geom_histogram() +
  facet_wrap(~paradrug, labeller = as_labeller(paradrug_labels)) +
  scale_fill_manual(
    values = c("#46B29B", "#80CFA9", "#D4C5E2", "#CF80A6"),
    name = "Anemia Status") +
  labs(title = "Birthweight by Anemia Status and Antiparasitic Use During Pregnancy", x = "Birth Weight (kg)") +
  theme_few()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Summary and interpretation of characteristics

Based on the characteristics observed in the table and graphic, a few things stand out. To start, given the fact that the core questions involves antiparasitic drug use, its important to point out that majority of people to not use antiparasitics during pregnancy. Additionally, less than 30% of people have some form of health insurance which may act as a major confounder when trying to predict birth weight. Majority of people surveyed have completed primary school (around 52.2%) but only 2% have some form of higher education. As for age, most of those who participated in the survey were in their late 20s, with a median age of 27 (SD of 5.12). The median birth weight was 2900g, which is less than the international average of roughly 3300g. Based on the graphic, it is evident that on average, children born to mothers who received antiparasitics during pregnancy weighed less at birth than those of mothers who did not. There is also a much larger interquartile range of birthweights for children born to mothers who were insured but did not receive antiparasitics during pregnancy than any other combined category. The Lowest average birth weights were among children born to mothers who were both insured and received antiparasitics during pregnancy. It is possible that due to the large discrepancy between number of insured vs uninsured, this relationship is due to noise, but it is interesting nonetheless. Overall, the spread is relatively consistent across categories.

Variable selection for statistical analysis and assumption testing

```
## LASSO
# Create predictor matrix (exclude intercept and outcome variable)
x <- model.matrix(bwkg ~ paradrug + insured + csect + anemia + delivplace +
  termpreg + bpast5 + wind_urbrur + wind + educ + ethnic +
  state + smokes + married + age,
  data = dhs_clean)[, -1] # Remove intercept column
```

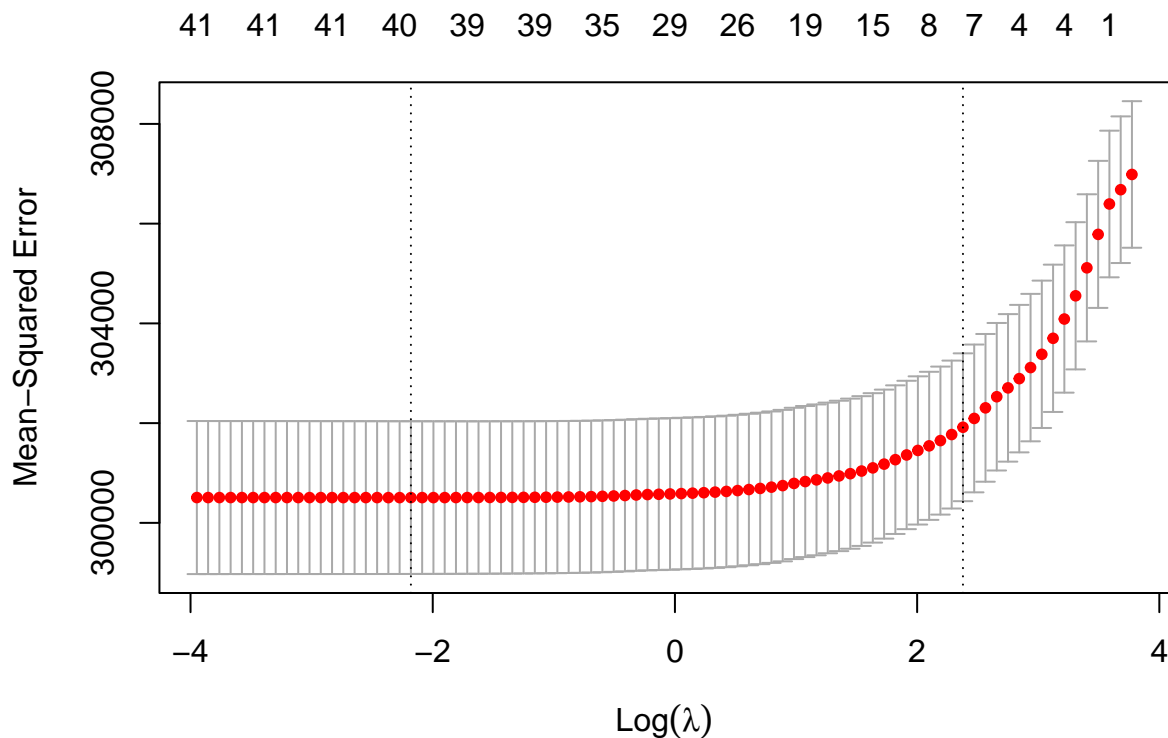
```

# Outcome variable
y <- dhs_clean$bwkg

# Cross-validate to find optimal lambda (penalty parameter)
set.seed(123) # For reproducibility
lasso_model <- cv.glmnet(x, y, alpha = 1) # alpha=1 for LASSO

# Plot cross-validation error
plot(lasso_model)

```



```

# Coefficients at lambda.min (retains more variables)
coef(lasso_model, s = "lambda.min")

```

```

## 43 x 1 sparse Matrix of class "dgCMatrix"
##              s1
## (Intercept) 2453.62676242
## paradrug1   -14.45031861
## insured1     2.78115360
## csect1       21.22229514
## anemia2      41.52196936
## anemia3      54.15828788
## anemia4      67.69394321
## delivplace12 13.82230763
## delivplace13  9.22314701
## delivplace21 20.12884041
## delivplace22 15.89831366

```

```
## delivplace23 16.29798578
## delivplace24 17.10129444
## delivplace25 22.08062354
## delivplace26 11.11465591
## delivplace27 35.16148399
## delivplace31 7.73316354
## delivplace32 .
## delivplace33 51.52247256
## delivplace96 31.79118017
## termpreg1 0.11128850
## bpast52 -3.94368315
## bpast53 -37.73086027
## bpast54 -60.13109914
## bpast55 109.30143226
## wind_urbrur 6.00574549
## wind2 35.31232295
## wind3 53.38908578
## wind4 56.56319494
## wind5 50.18093174
## educ1 16.57187002
## educ2 .
## educ3 55.58945073
## educ4 69.44232237
## educ5 110.24243606
## ethnic992 144.45099655
## ethnic993 34.87722706
## ethnic998 3.89257727
## state 0.04242206
## smokes1 179.61847673
## smokes2 49.37416028
## married -5.29778897
## age 6.28193847
```

```
# Coefficients at lambda.1se (simpler model)
coef(lasso_model, s = "lambda.1se")
```

```
## 43 x 1 sparse Matrix of class "dgCMatrix"
## s1
## (Intercept) 2630.793251
## paradrug1 .
## insured1 .
## csect1 7.575041
## anemia2 .
## anemia3 .
## anemia4 6.461215
## delivplace12 .
## delivplace13 .
## delivplace21 .
## delivplace22 .
## delivplace23 .
## delivplace24 .
## delivplace25 .
## delivplace26 .
## delivplace27 .
## delivplace31 .
```

```
## delivplace32 .
## delivplace33 .
## delivplace96 .
## termpreg1 .
## bpast52 .
## bpast53 .
## bpast54 .
## bpast55 .
## wind_urbrur 15.024979
## wind2 .
## wind3 .
## wind4 .
## wind5 .
## educ1 .
## educ2 .
## educ3 11.549754
## educ4 .
## educ5 56.124309
## ethnic992 105.435788
## ethnic993 .
## ethnic998 .
## state .
## smokes1 .
## smokes2 .
## married .
## age 3.925296
```

Predictors with non-zero coefficients are deemed important for explaining bwkg. It appears as though

Check for multicollinearity

```
vifmod <- lm(data = dhs_clean, bwkg~paradrug+age+ethnic+educ+bpast5+csect+anemia)
vif_values <- vif(vifmod) # Calculate VIF values
print(vif_values) # Display VIF values
```

```
##          GVIF Df GVIF^(1/(2*Df))
## paradrug 1.003222 1      1.001609
## age      1.070778 1      1.034784
## ethnic   1.030956 3      1.005094
## educ     1.139842 4      1.016496
## bpast5   1.034103 4      1.004201
## csect    1.070456 1      1.034628
## anemia   1.010921 3      1.001812
```

VIF are all aprox =1 implying low to no multicollinearity

Regressions and interpretation

several single stratas, merge singles for comparison
options(survey.lonely.psu = "adjust")

```
single_strata <- c(1023, 3241, 3261, 57221, 57222, 62921, 83823, 84022, 84122, 84223, 84421, 84621, 84622)
```

```
dhs_clean <- dhs_clean %>%
  mutate(strata = ifelse(strata %in% single_strata, 1024, strata))
```



```

## educ5          130.4074      8.0575  16.185 < 2e-16 ***
## bpast52        -6.0065     10.0405  -0.598 0.549697
## bpast53       -43.5147     20.7325  -2.099 0.035838 *
## bpast54       -78.0202     69.9704  -1.115 0.264841
## bpast55       248.6160     10.0055  24.848 < 2e-16 ***
## csect1         14.8893      5.1067   2.916 0.003552 **
## anemia2        60.8170     17.4858   3.478 0.000506 ***
## anemia3        67.5114     17.6470   3.826 0.000131 ***
## anemia4        70.8712     17.4808   4.054 5.04e-05 ***
## paradrug1:anemia2 -26.0589   34.3415  -0.759 0.447969
## paradrug1:anemia3 -23.4416   34.4628  -0.680 0.496383
## paradrug1:anemia4 -26.4609   34.1033  -0.776 0.437813
## educ1:bpast52     9.0219    17.1192   0.527 0.598195
## educ3:bpast52    -5.5710    11.7924  -0.472 0.636628
## educ4:bpast52    16.7582    37.0742   0.452 0.651260
## educ5:bpast52   -28.5304    15.7938  -1.806 0.070862 .
## educ1:bpast53   -36.0340    36.9172  -0.976 0.329037
## educ3:bpast53    33.4801    29.0108   1.154 0.248488
## educ4:bpast53   -9.0230    84.8786  -0.106 0.915341
## educ5:bpast53  -114.5483    52.5950  -2.178 0.029420 *
## educ1:bpast54   212.1435   144.9045   1.464 0.143200
## educ3:bpast54   -27.8951   118.9913  -0.234 0.814653
## educ4:bpast54   334.3582    77.5989   4.309 1.65e-05 ***
## educ5:bpast54  -615.4513   232.7468  -2.644 0.008191 **
## educ3:bpast55  -419.0672    11.6558 -35.954 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 305744.5)
##
## Number of Fisher Scoring iterations: 2
## Refined core model
coremodel <- sdesign %>%
  svyglm(
    formula = bwkg ~ paradrug + age + ethnic + educ +
      bpast5 + csect + anemia + educ*bpast5,
    family = gaussian()
  )
summary(coremodel)

##
## Call:
## svyglm(formula = bwkg ~ paradrug + age + ethnic + educ + bpast5 +
##       csect + anemia + educ * bpast5, design = ., family = gaussian())
##
## Survey design:
## Called via srvyr
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2541.542    19.683 129.127 < 2e-16 ***
## paradrug1      7.101      4.027   1.763 0.077898 .
## age           5.223      0.425  12.288 < 2e-16 ***
## ethnic992     -4.523      7.103  -0.637 0.524288

```

```

## ethnic993      31.744      9.393   3.380 0.000727 ***
## ethnic998     -9.165     25.717  -0.356 0.721570
## educ1         -3.216      9.235  -0.348 0.727684
## educ3         55.682      6.871   8.104 5.56e-16 ***
## educ4         50.813     17.289   2.939 0.003295 **
## educ5        130.412      8.062   16.176 < 2e-16 ***
## bpast52       -5.962     10.040  -0.594 0.552635
## bpast53      -43.519     20.719  -2.100 0.035703 *
## bpast54      -77.284     70.043  -1.103 0.269871
## bpast55       248.109      9.698   25.583 < 2e-16 ***
## csect1        14.895      5.107   2.917 0.003542 **
## anemia2       52.759     14.992   3.519 0.000434 ***
## anemia3       60.308     15.355   3.928 8.60e-05 ***
## anemia4       62.672     15.026   4.171 3.04e-05 ***
## educ1:bpast52   8.913     17.117   0.521 0.602567
## educ3:bpast52  -5.600     11.791  -0.475 0.634867
## educ4:bpast52  16.669     37.067   0.450 0.652933
## educ5:bpast52 -28.559     15.794  -1.808 0.070585 .
## educ1:bpast53 -35.988     36.905  -0.975 0.329488
## educ3:bpast53  33.447     28.999   1.153 0.248773
## educ4:bpast53  -8.949     84.865  -0.105 0.916024
## educ5:bpast53 -114.414     52.596  -2.175 0.029614 *
## educ1:bpast54  211.021    144.945   1.456 0.145442
## educ3:bpast54  -28.485    119.076  -0.239 0.810936
## educ4:bpast54  333.853     77.610   4.302 1.70e-05 ***
## educ5:bpast54 -617.194    233.008  -2.649 0.008082 **
## educ3:bpast55 -418.220     10.975  -38.108 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 305747.7)
##
## Number of Fisher Scoring iterations: 2
## Unweighted coremodel
coremodelunw <- lm(data = dhs_clean, bwkg~paradrug+age+ethnic+educ+bpast5+csect+anemia+educ*bpast5)
summary(coremodelunw)

##
## Call:
## lm(formula = bwkg ~ paradrug + age + ethnic + educ + bpast5 +
##      csect + anemia + educ * bpast5, data = dhs_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2490.58  -311.38    28.83   271.42  2840.04
##
## Coefficients: (3 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2474.9705    13.1998  187.500 < 2e-16 ***
## paradrug1    -14.5371     3.0116   -4.827 1.39e-06 ***
## age           6.6920     0.2832   23.629 < 2e-16 ***
## ethnic992    132.3587     4.0425   32.741 < 2e-16 ***
## ethnic993     28.4006     6.4062    4.433 9.29e-06 ***
## ethnic998      6.1665    19.1284    0.322 0.747171

```



```
## educ1      29.5576      6.4468      4.585 4.55e-06 ***
## educ3      82.0725      4.7711     17.202 < 2e-16 ***
## educ4     103.1462     12.2017      8.453 < 2e-16 ***
## educ5     153.5773      5.7867     26.540 < 2e-16 ***
## bpast52      0.2704      7.2425      0.037 0.970213
## bpast53     -46.4182     15.0623     -3.082 0.002058 **
## bpast54     -10.5414     65.7140     -0.160 0.872556
## bpast55     268.3289    387.9885      0.692 0.489196
## csect1      27.3449      3.4637      7.895 2.93e-15 ***
## anemia2     52.0177      9.6717      5.378 7.53e-08 ***
## anemia3     65.6903      9.7257      6.754 1.44e-11 ***
## anemia4     81.4989      9.5852      8.503 < 2e-16 ***
## educ1:bpast52 -4.9443     11.5648     -0.428 0.668994
## educ3:bpast52 -3.7783      8.5412     -0.442 0.658231
## educ4:bpast52 -8.1080     25.7163     -0.315 0.752545
## educ5:bpast52 -31.4146     11.6976     -2.686 0.007242 **
## educ1:bpast53 -13.3432     25.9963     -0.513 0.607760
## educ3:bpast53  21.2279     19.8308      1.070 0.284420
## educ4:bpast53 -8.4797     73.3097     -0.116 0.907914
## educ5:bpast53 -77.5505     40.1294     -1.933 0.053298 .
## educ1:bpast54 -12.7683    111.7083     -0.114 0.909000
## educ3:bpast54 -91.8172     92.8624     -0.989 0.322790
## educ4:bpast54  471.6026    323.7171      1.457 0.145164
## educ5:bpast54 -793.8416    217.5708     -3.649 0.000264 ***
## educ1:bpast55      NA          NA          NA          NA
## educ3:bpast55 -431.3054    671.9877     -0.642 0.520981
## educ4:bpast55      NA          NA          NA          NA
## educ5:bpast55      NA          NA          NA          NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 548.7 on 153551 degrees of freedom
## Multiple R-squared:  0.01967,    Adjusted R-squared:  0.01948
## F-statistic: 102.7 on 30 and 153551 DF,  p-value: < 2.2e-16
```

Finding R^2

```
weighted_r2 <- function(model) {
  y <- model.response(model.frame(model))
  w <- weights(model, type = "prior")
  pred <- predict(model)

  # NA handling
  valid <- complete.cases(y, pred, w)
  y <- y[valid]
  pred <- pred[valid]
  w <- w[valid]

  ss_res <- sum(w * (y - pred)^2)
  ss_tot <- sum(w * (y - weighted.mean(y, w))^2)

  1 - (ss_res / ss_tot)
}
```

Usage with survey model

```
coremodel <- svyglm(bwkg ~ paradrug + age,
  design = sdesign,
  family = gaussian())
```

```
weighted_r2(coremodel)
```

```
## [1] 0.002233451
```

```
## This implies that the variabce observed in the model can only explain about 0.22% of that observed i
```

Residual analysis, multicollinearity, and validation

```
## Q-Q plot for normality.
```

```
# Extract residuals from your svyglm model
```

```
residuals <- residuals(coremodel)
```

```
#plot
```

```
ggplot(data.frame(residuals = residuals), aes(sample = residuals)) +
  stat_qq(color = "#80CFA9") +
  stat_qq_line(color = "#CF80A6", size = 1) +
  theme_minimal() +
  labs(x = "Theoretical quantiles", y = "Ordered Values",
    title = "Normal Q-Q Plot of Residuals") +
  theme_economist()
```

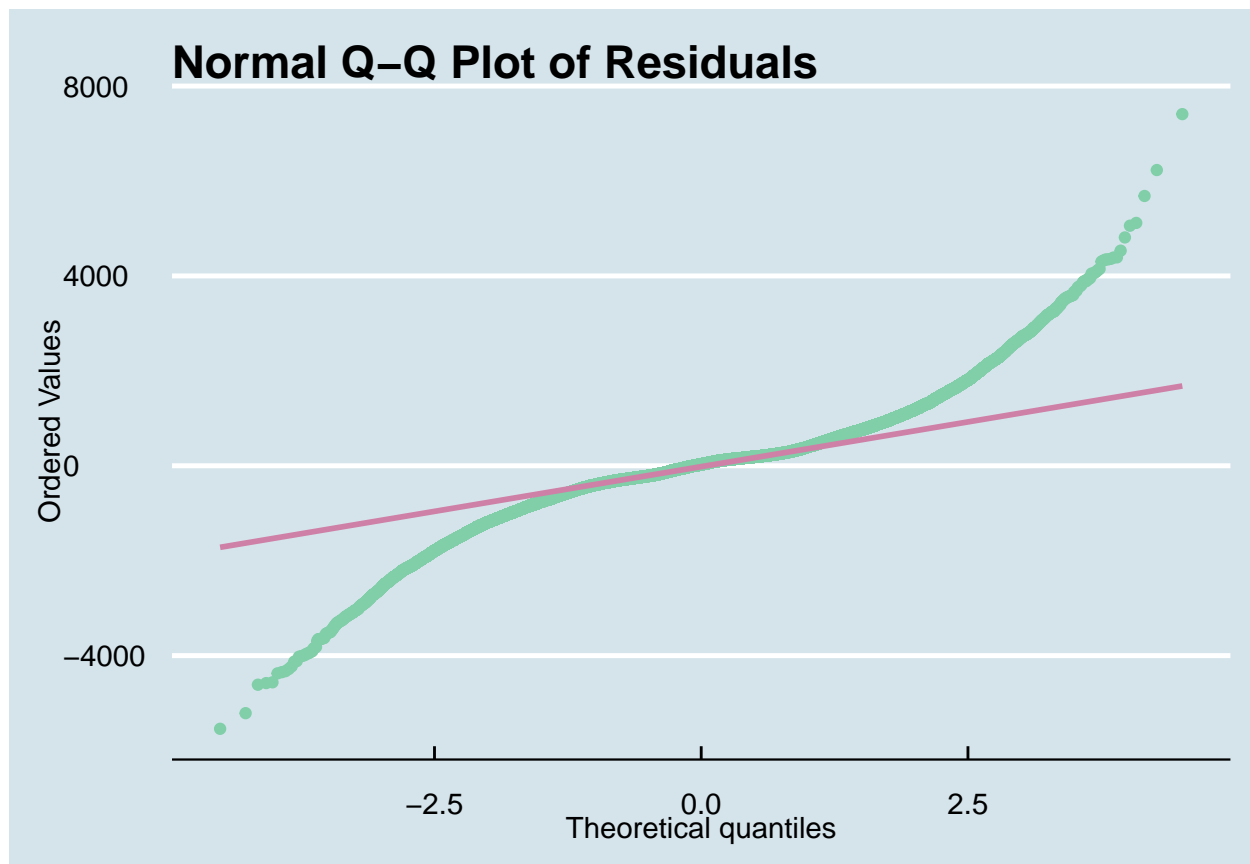
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
```

```
## i Please use `linewidth` instead.
```

```
## This warning is displayed once every 8 hours.
```

```
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
```

```
## generated.
```



The residuals generally follow the diagonal line, suggesting that the residuals are approximately normal.

```
# Perform Durbin-Watson test
durbinWatsonTest(coremodel)
```

```
## lag Autocorrelation D-W Statistic p-value
## 1 0.04293297 1.914134 0
## Alternative hypothesis: rho != 0
```

D-W Statistic = 1.914134: This value is reasonably close to 2 (which indicates no autocorrelation), suggesting that the residuals are approximately independent.

```
## Validation K-Fold
```

```
# Define cross-validation method
```

```
ctrl <- trainControl(method = "cv", number = 5)
```

```
# Train model with k-fold CV
```

```
model_cv <- train(
  bwkg ~ paradrug+age+ethnic+educ+bpast5+csect+anemia+age*paradrug+educ*bpast5, # Include all predictors
  data = dhs_clean,
  method = "lm",
  trControl = ctrl
)
```

```
## Warning in predict.lm(modelFit, newdata): prediction from rank-deficient fit;
## attr(*, "non-estim") has doubtful cases
```

```
# View average RMSE and R² across folds
```

```
print(model_cv$results)
```

##	intercept	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD
## 1	TRUE	548.7155	0.01923701	419.986	3.654693	0.001421575	2.479597

The validation results suggest poor model performance across the board. An RMSE (Root Mean Squared Error, A measure of the average distance between predicted and actual values. Lower values indicate better fit) of ~548.7 kg means that the model's predictions are super inaccurate. The R² of 0.022% indicated that paradrug and other predictors have very minimal explanatory power for bwkg. Low standard deviations (SD) in RMSE/R² across folds indicate the model isn't overfitting, but it's consistently underperforming.