COVID_day25_SL

Whitney Mgbara

4/15/2020

Load Libraries

Import case data and covariates

```
# Case data
day25 <- read_excel("Data/day25.xls")
day25$popland <- day25$Population/(day25$LandArea)

# Covariate Data
analytic_data2020 <- read_csv("Data/analytic_data2020.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   `State Abbreviation` = col_character(),
##   Name = col_character(),
##   `Poor or fair health numerator` = col_logical(),
##   `Poor or fair health denominator` = col_logical(),
##   `Poor physical health days numerator` = col_logical(),
##   `Poor physical health days denominator` = col_logical(),
##   `Poor mental health days numerator` = col_logical(),
##   `Poor mental health days denominator` = col_logical(),
##   `Adult smoking numerator` = col_logical(),
##   `Adult smoking denominator` = col_logical(),
##   `Adult obesity denominator` = col_logical(),
##   `Food environment index CI low` = col_logical(),
##   `Food environment index CI high` = col_logical(),
##   `Physical inactivity denominator` = col_logical(),
##   `Access to exercise opportunities CI low` = col_logical(),
##   `Access to exercise opportunities CI high` = col_logical(),
##   `Excessive drinking numerator` = col_logical(),
##   `Excessive drinking denominator` = col_logical(),
##   `Sexually transmitted infections CI low` = col_logical(),
##   `Sexually transmitted infections CI high` = col_logical()
##   # ... with 213 more columns
## )
```

```
## See spec(...) for full column specifications.
```

```
## Warning: 5342 parsing failures.
##  row                              col           expected      actual                          file
## 3097 Communicable disease raw value   1/0/T/F/TRUE/FALSE 923.1628069 'Data/analytic_data2020.csv'
## 3097 Communicable disease numerator   1/0/T/F/TRUE/FALSE 53348       'Data/analytic_data2020.csv'
## 3097 Communicable disease denominator 1/0/T/F/TRUE/FALSE 5778829     'Data/analytic_data2020.csv'
## 3097 Cancer incidence raw value       1/0/T/F/TRUE/FALSE 466.9       'Data/analytic_data2020.csv'
```

```
## 3097 Cancer incidence numerator       1/0/T/F/TRUE/FALSE 160800      'Data/analytic_data2020.csv'
## .... ............................... .................. ........... ............................
## See problems(...) for more details.
```

```r
colnames(analytic_data2020)[which(names(analytic_data2020) == "State Abbreviation")] <- "State"

eco_vars_of_int <- c("State",
                     "Name",
                     "Poor or fair health raw value",
                     "Adult smoking raw value",
                     "Food environment index raw value",
                     "Physical inactivity raw value",
                     "Excessive drinking raw value",
                     "Sexually transmitted infections raw value",
                     "Primary care physicians raw value",
                     "Flu vaccinations raw value",
                     "High school graduation raw value",
                     "Unemployment raw value",
                     "Air pollution - particulate matter raw value",
                     "Drinking water violations raw value",
                     "Diabetes prevalence raw value",
                     "HIV prevalence raw value",
                     "Food insecurity raw value",
                     "Drug overdose deaths raw value",
                     "Median household income raw value",
                     "% below 18 years of age raw value",
                     "% 65 and older raw value",
                     "% Hispanic raw value",
                     "% Females raw value", "% Rural raw value",
                     "Adult obesity raw value",
                     "Income inequality raw value",
                     "Uninsured adults raw value")

eco_health_covars <- names(analytic_data2020) %in% eco_vars_of_int

covars <- analytic_data2020[eco_health_covars]

# Merge case data with covarites
day25_covars <- merge(day25, covars, by = c("Name", "State"))

day25_covars$log_pop_dense <- log(day25_covars$Population/(day25_covars$LandArea))
```

Import Google mobility data

```r
# Data 1
grocery_pharmacy <- read_csv("Data/Mobility/google-mobility-us-groceryAndPharmacy.csv")

df.1 <-
  grocery_pharmacy %>%
  gather(key = date, value = value, -State)
df.1$type <- "grocery_pharmacy"


# Data 2
parks <- read_csv("Data/Mobility/google-mobility-us-parks.csv")
```

```
df.2 <-
  parks %>%
  gather(key = date, value = value, -State)
df.2$type <- "parks"


# Data 3
residential <- read_csv("Data/Mobility/google-mobility-us-residential.csv")

df.3 <-
  residential %>%
  gather(key = date, value = value, -State)
df.3$type <- "residential"


# Data 4
retailAndRecreation <- read_csv("Data/Mobility/google-mobility-us-retailAndRecreation.csv")

df.4 <-
  retailAndRecreation %>%
  gather(key = date, value = value, -State)
df.4$type <- "retailAndRecreation"


# Data 5
transitStations <- read_csv("Data/Mobility/google-mobility-us-transitStations.csv")

df.5 <-
  transitStations %>%
  gather(key = date, value = value, -State)
df.5$type <- "transitStations"


# Data 6
workplaces <- read_csv("Data/Mobility/google-mobility-us-workplaces.csv")

df.6 <-
  workplaces %>%
  gather(key = date, value = value, -State)
df.6$type <- "workplaces"


mobility.data <-
  left_join(df.1, df.2, by = c("State", "date", "value", "type")) %>%
  left_join(df.3) %>%
  left_join(df.4) %>%
  left_join(df.5) %>%
  left_join(df.6)

mobility_data_long <- rbind(df.1, df.2,df.3,df.4,df.5,df.6)
mobility_data_wide <- spread(mobility_data_long, date, value)
```

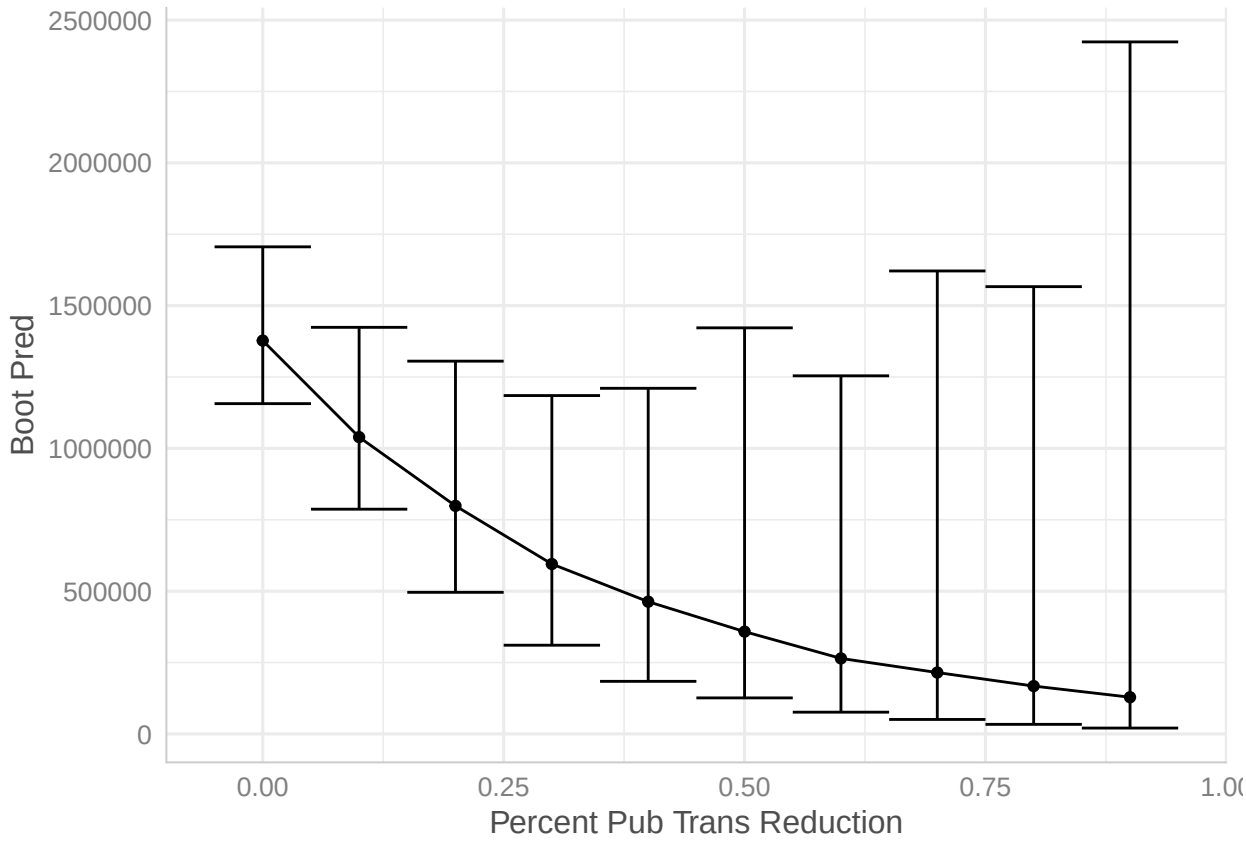
Run Simple Triple Interaction Model and get Marginal Effects for Transportation

```r
## set up bootstrap CI function
 bootstrapCI <- function(model, perc, boot_pred_data) {
  nr <- nrow(model$data)
  data <- model$data
  new_data <- data[sample(1:nr, size = nr, replace = TRUE), ]
  up <- update(model, data = new_data)

  #norm <- sum(predict(up, newdata = new_data, type = 'response'), na.rm = TRUE)

  boot_pred_data[,"log_Pub_Trans"] <- boot_pred_data$log_Pub_Trans - (boot_pred_data$log_Pub_Trans*perc
  perc_red <- sum(predict(up, newdata = boot_pred_data, type = 'response'), na.rm=TRUE)

  return(perc_red)
 }

#set up percentiles for trans reduction
percents <- c(0.0, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90)

#log transformation because the data is skewed forday 25 only
day25$log_pop_dense <- log(day25$popland+1)
day25$log_GDP <- log(day25$GDP+1)
day25$log_Pub_Trans <- log(day25$CommutingByPublicTransportation + 1)

#log transformation because the data is skewed forday 25 covars - 8 regions reduced compared to day25 d

day25_covars$log_pop_dense <- log(day25_covars$popland+1)
day25_covars$log_GDP <- log(day25_covars$GDP+1)
day25_covars$log_Pub_Trans <- log(day25_covars$CommutingByPublicTransportation + 1)


#models
est.pois_model <- glm(ConfirmedCasesDay25~ log_pop_dense+log_GDP*log_Pub_Trans +
                        log(Population) +
                        `% Females raw value` +
                        `% 65 and older raw value` +
                        `Adult obesity raw value`+
                        `Physical inactivity raw value` +
                        `Unemployment raw value` +
                        `Income inequality raw value` +
                        `Poor or fair health raw value`+
                        `Uninsured adults raw value` +
                        `Adult smoking raw value`,
                       family = "poisson",
                       data = day25_covars,
                       offset(log(Population)))

##too many parameters I think here to get convergence based on starting likelihood value, could use Pois
# est2.nb_model <- glm.nb(ConfirmedCasesDay25~ log_pop_dense*log_GDP*log_Pub_Trans,
#                         data = day25,
#                         offset(log(Population)))

model <- est.pois_model
```

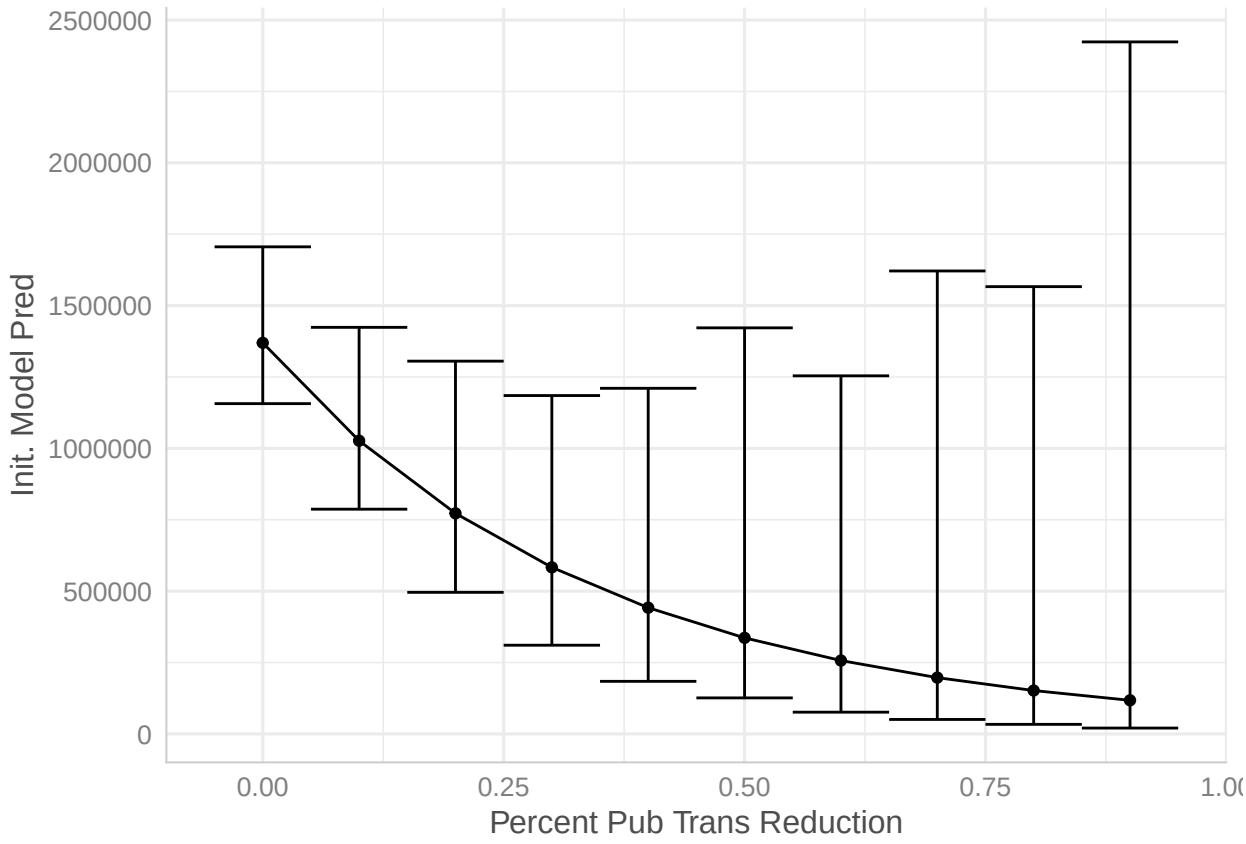
```
results <- as.data.frame(matrix(nrow = length(percents), ncol = 6))
colnames(results) <- c('Percent Pub Trans Reduction','Total Cases', 'Boot Pred', 'Boot Low' , 'Boot High'

for (i in 1:length(percents)) {

  perc <- percents[i]
  data_temp <- day25_covars

  data_temp$log_Pub_Trans <- day25_covars$log_Pub_Trans - (day25_covars$log_Pub_Trans*perc)
  pred_perc_red <- predict(model, newdata = data_temp, type = 'response')

  sum_perc_red <- sum(pred_perc_red, na.rm = TRUE)

  boot <- replicate(1000, bootstrapCI(model = model,
                                      perc =  perc,
                                      boot_pred_data = day25_covars))

  #mean_boot_diff <- mean(boot, na.rm = TRUE)

  CI_boot <- quantile(boot, probs = c(0.025,0.50, 0.975), na.rm = TRUE)

  results[i,1] <- perc
  results[i,2] <- sum_perc_red
  results[i,3] <- CI_boot[[2]]
  results[i,4] <- CI_boot[[1]]
  results[i,5] <- CI_boot[[3]]
  results[i,6] <- sum_perc_red
}
```

Plot cases over time marginally attributed to each reduction level

```
ggplot(results, aes(x=`Percent Pub Trans Reduction`, y=`Boot Pred`)) +
    geom_errorbar(aes(ymin=`Boot Low`, ymax=`Boot High`), width=.1) +
    geom_line() +
    geom_point()
```




```
ggplot(results, aes(x=`Percent Pub Trans Reduction`, y=`Init. Model Pred`)) +
    geom_errorbar(aes(ymin=`Boot Low`, ymax=`Boot High`), width=.1) +
    geom_line() +
    geom_point()
```



Setting Up SuperLearner and TMLE

```
##set up node list

node_list <- list(
  W = c('popland', 'GDP',eco_vars_of_int[-c(1,2)]),
  A = "pub_trans_quantile",
  Y = "ConfirmedCasesDay25"
)

#process missing
processed <- process_missing(day25_covars, node_list)
COVID_data <- processed$data
node_list <- processed$node_list

ate_spec <- tmle_ATE(
  treatment_level = 10,
  control_level = 1
)

sl3_list_learners("continuous")
sl3_list_learners("binomial")
sl3_list_learners("categorical")

# choose y learners
lrnr_mean <- make_learner(Lrnr_mean)
lrnr_xgboost <- make_learner(Lrnr_xgboost)
```

```
Lrnr_glm <- make_learner(Lrnr_glm)
Lrnr_hal9001 <- make_learner(Lrnr_hal9001)

# choose a learners
Lrnr_grf <- make_learner(Lrnr_grf)
#Lrnr_multivariate <- make_learner(Lrnr_multivariate)


# define metalearners appropriate to data types
ls_metalearner <- make_learner(Lrnr_nnls)
#mn_metalearner <- make_learner(Lrnr_solnp, metalearner_linear_multinomial,
                               #loss_loglik_multinomial)

sl_Y <- Lrnr_sl$new(learners = list(lrnr_mean, lrnr_xgboost, Lrnr_glm, Lrnr_hal9001),
                    metalearner = ls_metalearner)

sl_A <- Lrnr_sl$new(learners = list(Lrnr_grf))

learner_list <- list(A = sl_A, Y = sl_Y)
```

```
tmle_fit <- tmle3(ate_spec, COVID_data, node_list, learner_list)
tmle_fit
estimates <- tmle_fit$summary$psi_transformed
estimates
```

```
States <- unique(day25$State)
CV.risk <- matrix(NA, nrow=length(States), ncol=6)

#estimates<- matrix(NA, nrow=500, ncol=4)
ObsData <- day25

for(i in States){

idx <- match(i,States)

validation_data <- ObsData %>% filter(State == i)
training_data <- ObsData %>% filter(State != i)

est1.model <- glm(ConfirmedCasesDay25 ~ popland + GDP + CommutingByPublicTransportation, family = "gaus

est2.model <- glm(ConfirmedCasesDay25 ~ popland + GDP + CommutingByPublicTransportation, family = "pois

est3.model <- glm(ConfirmedCasesDay25 ~ popland *GDP * CommutingByPublicTransportation, family = "gauss

est4.model <- glm(ConfirmedCasesDay25 ~ popland *GDP * CommutingByPublicTransportation, family = "poiss

est5.model <- glm(ConfirmedCasesDay25 ~ popland + GDP , family = "gaussian", data = training_data)

est6.model <- glm(ConfirmedCasesDay25 ~ popland + GDP , family = "poisson", data = training_data)


predict.est1 <- predict(est1.model, newdata = validation_data)
predict.est2 <- predict(est2.model, newdata = validation_data)
predict.est3 <- predict(est3.model, newdata = validation_data)
```

```
predict.est4 <- predict(est4.model, newdata = validation_data)
predict.est5 <- predict(est5.model, newdata = validation_data)
predict.est6 <- predict(est6.model, newdata = validation_data)
#predict.est7 <- predict(est7.model, newdata = validation_data)

l2.hat1 <- mean((validation_data$ConfirmedCasesDay25 - predict.est1)^2)
l2.hat2 <- mean((validation_data$ConfirmedCasesDay25 - predict.est2)^2)
l2.hat3 <- mean((validation_data$ConfirmedCasesDay25 - predict.est3)^2)
l2.hat4 <- mean((validation_data$ConfirmedCasesDay25 - predict.est4)^2)
l2.hat5 <- mean((validation_data$ConfirmedCasesDay25 - predict.est5)^2)
l2.hat6 <- mean((validation_data$ConfirmedCasesDay25 - predict.est6)^2)
#l2.hat7 <- mean((validation_data$ConfirmedCasesDay25 - predict.est7)^2)

CV.risk[idx,] <- c(l2.hat1, l2.hat2, l2.hat3, l2.hat4,l2.hat5,l2.hat6)
}

colnames(CV.risk) <- c("l2.est1", "l2.est2", "l2.est3", "l2.est4", "l2.est5", "l2.est6")
CV.risk <- as.data.frame(CV.risk)

CV.risk
mses <- colMeans(CV.risk, na.rm = TRUE)

match(mses, min(mses))
```

```
Y <- day25$ConfirmedCasesDay25
X <-  subset(day25, select= -ConfirmedCasesDay25)
X <- as.data.frame(X)

Q_lib <- c("SL.mean", "SL.glmnet", "SL.ranger", "SL.rpartPrune", "SL.bayesglm")
g_lib <- c("SL.mean", "SL.glmnet")


vim <- varimpact(Y = Y, data = X, Q.library = Q_lib, g.library = g_lib, family="gaussian" )
vim$results_all

plot_var("Population", vim)
```

Junk Drawer

```
boot_log <- replicate(1000, bootstrapCI(model = est1.log_model, newdata = newdata))

day25$uci_log <- apply(boot_log, MARGIN = 1, FUN = quantile, probs = 0.925, na.rm=TRUE)
day25$lci_log <- apply(boot_log, MARGIN = 1, FUN = quantile, probs = 0.025, na.rm=TRUE)
day25$fit_log <- apply(boot_log, MARGIN = 1, FUN = quantile, probs = 0.5, na.rm=TRUE)

day25$uci <- apply(boot, MARGIN = 1, FUN = quantile, probs = 0.925, na.rm=TRUE)
day25$lci <- apply(boot, MARGIN = 1, FUN = quantile, probs = 0.025, na.rm=TRUE)
day25$fit <- apply(boot, MARGIN = 1, FUN = quantile, probs = 0.5, na.rm=TRUE)



g3_log <- ggplot(day25, aes(x = day25$CommutingByPublicTransportation, y = ConfirmedCasesDay25)) +
  theme_bw() +
  geom_point() +
```

```
  geom_line(aes(y = fit)) +
  geom_ribbon(aes(ymin = lci, ymax = uci), alpha = 0.3)

g3_log

g3 <- ggplot(day25, aes(x = day25$CommutingByPublicTransportation, y = ConfirmedCasesDay25)) +
  theme_bw() +
  geom_point() +
  geom_line(aes(y = fit_log)) +
  geom_ribbon(aes(ymin = lci_log, ymax = uci_log), alpha = 0.3)

g3


#predictions for 10 and 90 quantiles log model
predict.log_est_10 <- predict(est1.log_model, newdata = marginal_data_10, type='response')
predict.log_est_90 <- predict(est1.log_model, newdata = marginal_data_90, type='response')

#predictions for 10 and 90 quantiles
predict.est_10 <- predict(est1.model, newdata = marginal_data_10, type='response')
predict.est_90 <- predict(est1.model, newdata = marginal_data_90, type='response')

log_marg_10_90_diff <- mean(predict.log_est_90 - predict.log_est_10, na.rm = TRUE)
marg_10_90_diff <- mean(predict.est_90 - predict.est_10, na.rm = TRUE)
```

The marginal differences for public transportation at 90 vs. 10 percentile

```
marg_10_90_diff
log_marg_10_90_diff
```