China0430

Xiaolan

2023-05-02

In this report, we will use cross validation to compare the models.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
regdat <- readRDS("~/Documents/GitHub/ChinaAg/From Hao/data/dongbei/regdat.rds") %>%
  mutate(ex_revenue = yield_lag * ex_price + subsidy_lag,
         tcost = yield_lag * cost,
         rcratio = (yield_lag * ex_price + subsidy_lag)/(yield_lag * cost))
preddat <- readRDS("~/Documents/GitHub/ChinaAg/From Hao/data/dongbei/preddat.rds") %>%
  mutate(ex_revenue = yield_lag * ex_price + subsidy_lag,
         tcost = yield_lag * cost,
         rcratio = (yield_lag * ex_price + subsidy_lag)/(yield_lag * cost))

# combine two datasets

alldat <- rbind(regdat, preddat) %>% arrange(region, crop, year)

alldat <- alldat %>%
    mutate(ex_revenue = yield_lag * ex_price + subsidy_lag,
         tcost = yield_lag * cost,
         rcratio = (yield_lag * ex_price + subsidy_lag)/(yield_lag * cost))

alldat <- alldat %>%
    mutate(share_wheat = alldat %>% filter(crop == "wheat") %>% "$"(share) %>%
           matrix(., nrow = 21) %>% # use 21 here since we have 21 years
           kronecker(rep(1, 4), .) %>% as.vector()) %>%
  mutate(y2 = log(share/(other + share_wheat)))   # get new y2 since now we combine other

alldat <- alldat %>%
```

```
mutate(pfex_corn = alldat %>% filter(crop == "corn") %>% "$"(profit_ex) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       pfex_rice = alldat %>% filter(crop == "rice") %>% "$"(profit_ex) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       pfex_soy = alldat %>% filter(crop == "soybean") %>% "$"(profit_ex) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       pfex_wheat = alldat %>% filter(crop == "wheat") %>% "$"(profit_ex) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),

       pftr_corn = alldat %>% filter(crop == "corn") %>% "$"(profit_true_lag) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       pftr_rice = alldat %>% filter(crop == "rice") %>% "$"(profit_true_lag) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       pftr_soy = alldat %>% filter(crop == "soybean") %>% "$"(profit_true_lag) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       pftr_wheat = alldat %>% filter(crop == "wheat") %>% "$"(profit_true_lag) %>%
         matrix(., nrow = 21) %>% # "+"(400) %>% log() %>%
         kronecker(rep(1, 4), .) %>% as.vector(),

       revex_corn = alldat %>% filter(crop == "corn") %>% "$"(ex_revenue) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       revex_rice = alldat %>% filter(crop == "rice") %>% "$"(ex_revenue) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       revex_soy = alldat %>% filter(crop == "soybean") %>% "$"(ex_revenue) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       revex_wheat = alldat %>% filter(crop == "wheat") %>% "$"(ex_revenue) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),

       tcost_corn = alldat %>% filter(crop == "corn") %>% "$"(tcost) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       tcost_rice = alldat %>% filter(crop == "rice") %>% "$"(tcost) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       tcost_soy = alldat %>% filter(crop == "soybean") %>% "$"(tcost) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector(),
       tcost_wheat = alldat %>% filter(crop == "wheat") %>% "$"(tcost) %>%
         matrix(., nrow = 21) %>%
         kronecker(rep(1, 4), .) %>% as.vector())%>%
mutate(rr_corn = revex_corn/revex_soy,
       rr_rice = revex_rice/revex_soy,
```

```
        rc_corn = tcost_corn/tcost_soy,
        rc_rice = tcost_rice/tcost_soy)
```

Start the cross validation:

```
RMSE1_list <- NULL
RMSE2_list <- NULL
RMSE3_list <- NULL
RMSE4_list <- NULL
RMSE5_list <- NULL
for(i in 2001:2021)
{
  # Step 1: select regdat and preddat
  regdat <- alldat %>% filter(year != i)
  preddat <- alldat %>% filter(year == i)

  # Step 2: get the true share
  share_true <- preddat$share %>% matrix(., nrow = 4) %>% "["(-4,)

  # Step 3: run all the models

  # model 1
  lm_corn_pf <- lm(y2 ~ 0 + pfex_corn + pfex_rice + pfex_soy  + region,
              data = regdat %>% filter(crop == "corn"))

  lm_rice_pf <- lm(y2 ~ 0 + pfex_corn + pfex_rice + pfex_soy  + region,
                data = regdat %>% filter(crop == "rice"))

  lm_soy_pf <- lm(y2 ~ 0 + pfex_corn + pfex_rice + pfex_soy  + region,
                data = regdat %>% filter(crop == "soybean"))

  # model 2
  lm_corn_rev <- lm(y2 ~ 0 + revex_corn + revex_rice + revex_soy   + region,
              data = regdat %>% filter(crop == "corn"))

  lm_rice_rev <- lm(y2 ~ 0 + revex_corn + revex_rice + revex_soy  + region,
                data = regdat %>% filter(crop == "rice"))

  lm_soy_rev <- lm(y2 ~ 0 + revex_corn + revex_rice + revex_soy  + region,
                data = regdat %>% filter(crop == "soybean"))

  # model 3

  lm_corn_rr <- lm(y2 ~ 0 + rr_corn + rr_rice + region, # rc_corn + rc_rice + rc_soy +
              data = regdat %>% filter(crop == "corn"))

  lm_rice_rr <- lm(y2 ~ 0 + rr_corn + rr_rice + region,
                data = regdat %>% filter(crop == "rice"))

  lm_soy_rr <- lm(y2 ~ 0 +  rr_corn + rr_rice +  region,
                data = regdat %>% filter(crop == "soybean"))

  # model 4
```

```r
lm_corn_pftr <- lm(y2 ~ 0 + pftr_corn + pftr_rice + pftr_soy + region,
            data = regdat %>% filter(crop == "corn"))

lm_rice_pftr <- lm(y2 ~ 0 + pftr_corn + pftr_rice + pftr_soy + region,
             data = regdat %>% filter(crop == "rice"))

lm_soy_pftr <- lm(y2 ~ 0 + pftr_corn + pftr_rice + pftr_soy +  region,
             data = regdat %>% filter(crop == "soybean"))

# model 5
lm_corn_rrrc <- lm(y2 ~ 0 + rr_corn + rr_rice + rc_corn + rc_rice + region, # rc_corn + rc_rice + rc_
            data = regdat %>% filter(crop == "corn"))

lm_rice_rrrc <- lm(y2 ~ 0 + rr_corn + rr_rice + rc_corn + rc_rice + region,
             data = regdat %>% filter(crop == "rice"))

lm_soy_rrrc <- lm(y2 ~ 0 +  rr_corn + rr_rice + rc_corn + rc_rice + region,
             data = regdat %>% filter(crop == "soybean"))

# Step 4: calculate the predict share:

# model 1
pd_corn_pf <- predict(lm_corn_pf, newdata = preddat %>% filter(crop == "corn"))
pd_rice_pf <- predict(lm_rice_pf, newdata = preddat %>% filter(crop == "rice"))
pd_soy_pf <- predict(lm_soy_pf, newdata = preddat %>% filter(crop == "soybean"))

y_pred_pf <- rbind(pd_corn_pf, pd_rice_pf, pd_soy_pf)

share_est_pf <- round(sweep(exp(y_pred_pf), MARGIN = 2, FUN = "/", STATS =
                            colSums(exp(y_pred_pf)) + 1) * 100, 2)

# model 2
pd_corn_rev <- predict(lm_corn_rev, newdata = preddat %>% filter(crop == "corn"))
pd_rice_rev <- predict(lm_rice_rev, newdata = preddat %>% filter(crop == "rice"))
pd_soy_rev <- predict(lm_soy_rev, newdata = preddat %>% filter(crop == "soybean"))

y_pred_rev <- rbind(pd_corn_rev, pd_rice_rev, pd_soy_rev)

share_est_rev <- round(sweep(exp(y_pred_rev), MARGIN = 2, FUN = "/", STATS =
                             colSums(exp(y_pred_rev)) + 1) * 100, 2)

# model 3:
pd_corn_rr<- predict(lm_corn_rr, newdata = preddat %>% filter(crop == "corn"))
pd_rice_rr <- predict(lm_rice_rr, newdata = preddat %>% filter(crop == "rice"))
pd_soy_rr <- predict(lm_soy_rr, newdata = preddat %>% filter(crop == "soybean"))

y_pred_rr <- rbind(pd_corn_rr, pd_rice_rr, pd_soy_rr)

share_est_rr <- round(sweep(exp(y_pred_rr), MARGIN = 2, FUN = "/", STATS =
                            colSums(exp(y_pred_rr)) + 1) * 100, 2)

# model 4:
pd_corn_pftr <- predict(lm_corn_pftr, newdata = preddat %>% filter(crop == "corn"))
```

```
  pd_rice_pftr <- predict(lm_rice_pftr, newdata = preddat %>% filter(crop == "rice"))
  pd_soy_pftr <- predict(lm_soy_pftr, newdata = preddat %>% filter(crop == "soybean"))

  y_pred_pftr <- rbind(pd_corn_pftr, pd_rice_pftr, pd_soy_pftr)

  share_est_pftr <- round(sweep(exp(y_pred_pftr), MARGIN = 2, FUN = "/", STATS =
                                colSums(exp(y_pred_pftr)) + 1) * 100, 2)

  # model 5:
  pd_corn_rrrc <- predict(lm_corn_rrrc, newdata = preddat %>% filter(crop == "corn"))
  pd_rice_rrrc <- predict(lm_rice_rrrc, newdata = preddat %>% filter(crop == "rice"))
  pd_soy_rrrc <- predict(lm_soy_rrrc, newdata = preddat %>% filter(crop == "soybean"))

  y_pred_rrrc <- rbind(pd_corn_rrrc, pd_rice_rrrc, pd_soy_rrrc)

  share_est_rrrc <- round(sweep(exp(y_pred_rrrc), MARGIN = 2, FUN = "/", STATS =
                                colSums(exp(y_pred_rrrc)) + 1) * 100, 2)

  # Step 5: calculate the MSE

  RMSE1 <- rowMeans((share_true - share_est_pf)^2) %>% sqrt()
  RMSE2 <- rowMeans((share_true - share_est_rev)^2) %>% sqrt()
  RMSE3 <- rowMeans((share_true - share_est_rr)^2) %>% sqrt()
  RMSE4 <- rowMeans((share_true - share_est_pftr)^2) %>% sqrt()
  RMSE5 <- rowMeans((share_true - share_est_rrrc)^2) %>% sqrt()

  # Step 6: combine the results
  RMSE1_list <- rbind(RMSE1_list, RMSE1)
  RMSE2_list <- rbind(RMSE2_list, RMSE2)
  RMSE3_list <- rbind(RMSE3_list, RMSE3)
  RMSE4_list <- rbind(RMSE4_list, RMSE4)
  RMSE5_list <- rbind(RMSE5_list, RMSE5)
}

RMSE1_avg <- colMeans(RMSE1_list)
RMSE2_avg <- colMeans(RMSE2_list)
RMSE3_avg <- colMeans(RMSE3_list)
RMSE4_avg <- colMeans(RMSE4_list)
RMSE5_avg <- colMeans(RMSE5_list)
```

The RMSE table:

```
cbind(RMSE1_avg, RMSE2_avg, RMSE3_avg, RMSE4_avg, RMSE5_avg) %>%
  "colnames<-"(paste0("Model", 1:5)) %>%
  "rownames<-"(unique(regdat$crop)[-4]) %>% pander::pander()
```

	Model1	Model2	Model3	Model4	Model5
corn	3.818	4.361	4.56	6.821	4.714
rice	1.51	1.547	1.486	2.028	1.352
soybean	1.961	2.318	2.364	2.909	2.375

The standard deviation of RMSE

```r
RMSE1_sd <- apply(RMSE1_list, MARGIN = 2, FUN = sd)
RMSE2_sd <- apply(RMSE2_list, MARGIN = 2, FUN = sd)
RMSE3_sd <- apply(RMSE3_list, MARGIN = 2, FUN = sd)
RMSE4_sd <- apply(RMSE4_list, MARGIN = 2, FUN = sd)
RMSE5_sd <- apply(RMSE5_list, MARGIN = 2, FUN = sd)

cbind(RMSE1_sd, RMSE2_sd, RMSE3_sd, RMSE4_sd, RMSE5_sd) %>%
  "colnames<-"(paste0("Model", 1:5)) %>%
  "rownames<-"(unique(regdat$crop)[-4]) %>% pander::pander()
```

	Model1	Model2	Model3	Model4	Model5
corn	1.204	2.503	1.545	2.776	1.492
rice	0.5797	0.6596	0.6355	0.7629	0.6763
soybean	1.367	1.568	1.703	1.473	1.796