Homework 1 - Model Selection

Instructions

Due 29 September at 11pm

Instructions

- Use the space inside of

```
::: {.solbox data-latex=""}


:::
```

to answer the following questions.

- Do not move this file or copy it and work elsewhere. Work in the same directory.
- Use a new branch named whatever you want. Create it now! Can't come up with something, try here. Make a small change, say by adding your name. Commit and push now!
- Try to Knit your file now. Are there errors? Fix them now, not at 11pm on the due date.
- There MUST be some text between `::: {.solbox}` and the next `:::` or this will fail to Knit.
- If your code or figures run off the .pdf, you'll lose 2 points automatically.

Forecasting COVID in BC

In the course `Stat406` package, you have up-to-date COVID-19 case data for the five health authorities in BC. We're going to play with a few models for making forecasts. Your goal is to **predict 2-week ahead case counts for all 5 health authorities**.

There are 2 BC Covid datasets we need for this assignment: `bccovid_train` and `bccovid_test`. Both have 6 variables: HA, date, cases+14, cases+0, cases-7, cases-14. `cases+14` is the response: it is the total number of cases observed in the week ending 14 days after `date`. The other `cases` variables are features: the total number of cases in the week ending on the date, 1 week earlier and 2 weeks earlier. For complete descriptions, try `?bccovid` after loading the course package.

1. Load the training data.
   a. make a plot of today's cases over time. Color the points by health authority.
   b. Write a few sentences discussing the data. Describe any similarities or differences across health authorities you notice. How do cases seem to "evolve" over time?

   Hint: to use these covariates in `ggplot()`, you have to put them inside "backtick" marks, like `` `cases+0` ``.

a.
b.

2. Estimate a linear model of 2-week ahead cases on the three features using the training data.
    a. Plot the residuals over time (again colored by HA).
    b. Make a QQ-plot of the residuals (colored by HA). Do these plots seem reasonable?
    c. Report the slope coefficient on `cases+0`. Describe what this coefficient means in the context of this problem. Does this interpretation make sense, why or why not?
    d. Calculate leave-one-out CV.
    e. Does this model seem appropriate? Justify your answer.

Hint: to use these covariates in a formula, you have to put them inside "backtick" marks, like `` `cases+0` ``.

a.
b.
c.
d.
e.

3. Examine the bcpop data set. Using an appropriate *_join(), add this column to the training data. Now scale all the columns by the population to get "cases per capita". Multiply by 100,000 just to get reasonable looking numbers.
   - Make a plot of today's cases over time scaled by population (and colored by HA). Furthermore, scale the axis on the log10() scale. (Hint: try ?scale_y_log10). How does this compare to the plot in Question 1? Write 2 complete sentences.
   - Repeat Question 2 with your scaled data after taking log() of all the predictors AND using HA as a categorical predictor.

First bullet.

Repeat Q2 with different data.

a.
b.
c.
d.
e.

4. Plot the fitted values from the two models against each other (fits from the Question 2 on the x-axis, with fits from Question 3 on the y-axis). Be sure to transform them as needed to be comparable. Color the points by HA. What do you notice?

(placeholder text)

5. Calculate the mean absolute error and the root mean squared error of both models. Discuss the results in 2 sentences.

(placeholder text)

6. Can we use the CV scores calculated in the previous two problems to choose between these models? Justify your answer.

(placeholder text)

7. Use the K-fold CV function from class (loaded below) to compare these two models (the first model being the one with `bccovid_train` from Question 2, and the second being the with the log of case rates and the factor for HA from Question 3). Use $K = 20$. In order to do this, you'll need to be sure that the errors are on the same scale.

Use `Cases` as the scale. This means you'll have to specify the model and the error function carefully. a. What are the CV scores for both models? b. Which do you prefer? c. How much more accurate is it?

```r
#' @param data The full data set
#' @param estimator Function. Has 1 argument (some data) and fits a model.
#' @param predictor Function. Has 2 args (the fitted model, the_newdata) and produces predictions
#' @param error_fun Function. Has one arg: the test data, with fits added.
#' @param kfolds Integer. The number of folds.
kfold_cv <- function(data, estimator, predictor,
                     error_fun, kfolds = 5) {
  n <- nrow(data)
  fold.labels <- sample(rep(1:kfolds, length.out = n))
  errors <- double(kfolds)
  for (fold in 1:kfolds) {
    test.rows <- fold.labels == fold
    train <- data[!test.rows, ]
    test <- data[test.rows, ]
    current_model <- estimator(train)
    test$.preds <- predictor(current_model, test)
    errors[fold] <- error_fun(test)
  }
  mean(errors)
}
```

a.
b.
c.

8. Use your preferred model to predict the test set.
    a. Plot the actual values and the predictions, colored by HA. Use lines for the predictions and dots for the observations.
    b. For each HA, how accurate is the prediction on average (what is the mean absolute error on the scale of cases)?
    c. What about overall?
    d. How does this compare to the CV score you computed above?

a.
b.
c.
d.