Homework 1 - Model Selection

Sicily Xie 54385315

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**.

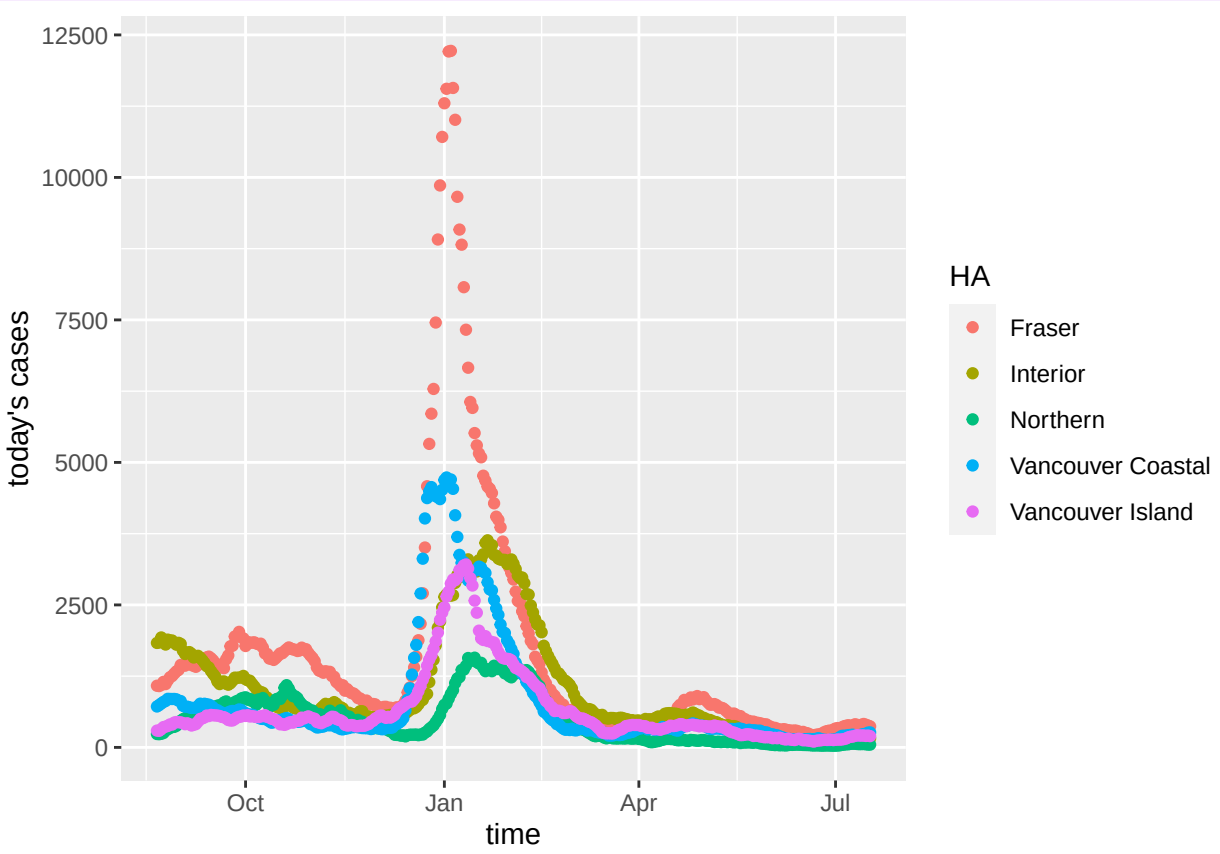
There are 2 BC Covid datasets we need for this assignment: `bccovid_train` and `bccovid_test`. Both have 6 variables: HA, date, cases+14, cases+0, cases-7, cases-14. `cases+14` is the response: it is the total number of cases observed in the week ending 14 days after `date`. The other `cases` variables 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.

```
# ?bccovid
# Stat406::bccovid_train
# bccovid_train %>% colnames()
qplot(
  data = bccovid_train,
  x =  `date`,
  y = `cases+0`,
  color = `HA`,
  xlab = "time",
  ylab = "today's cases"
)
```



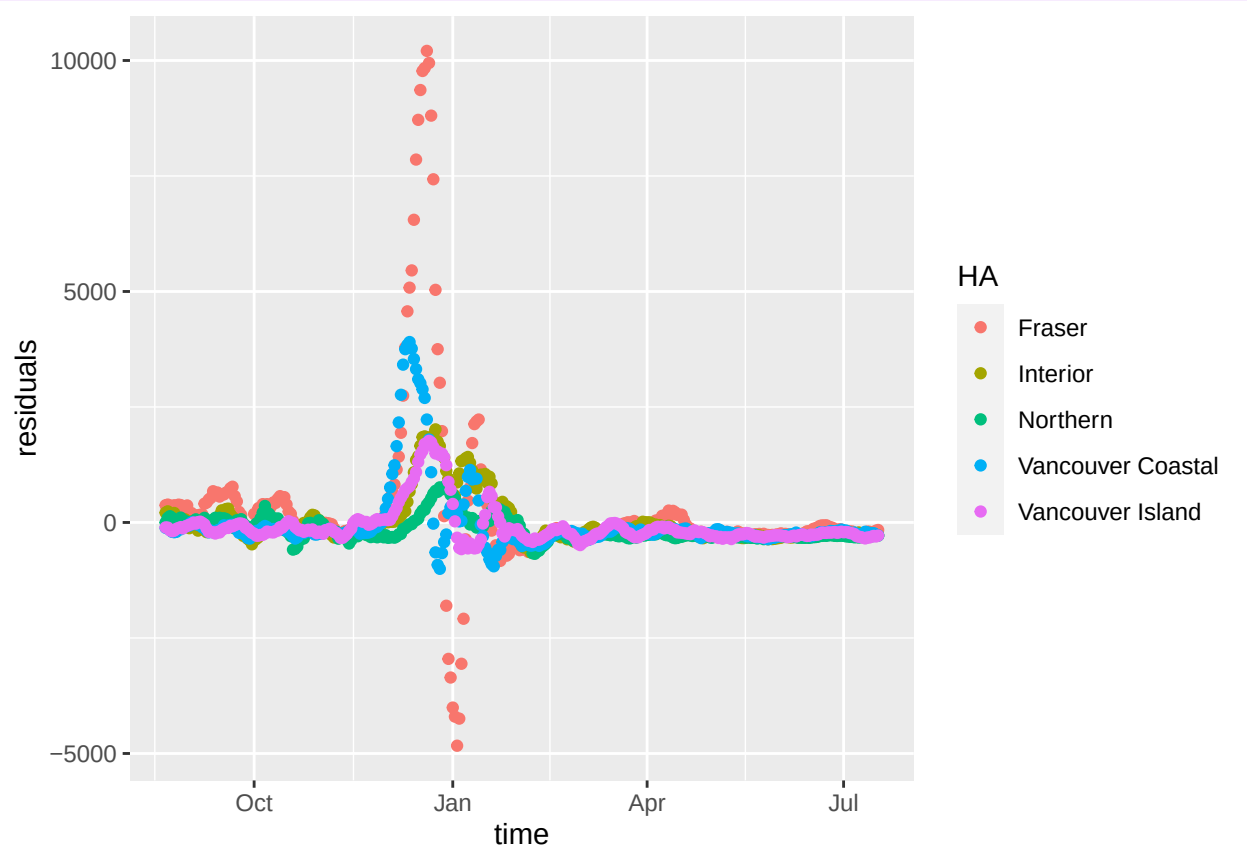
b. The similarity across health authorities is that they are shaped like a bell, being a normal distribution. The difference across health authorities is that the highest point on the curve is different, so that they have different mean, mode and median of the data collected. The cases tend to be 0 over time.

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` ``.

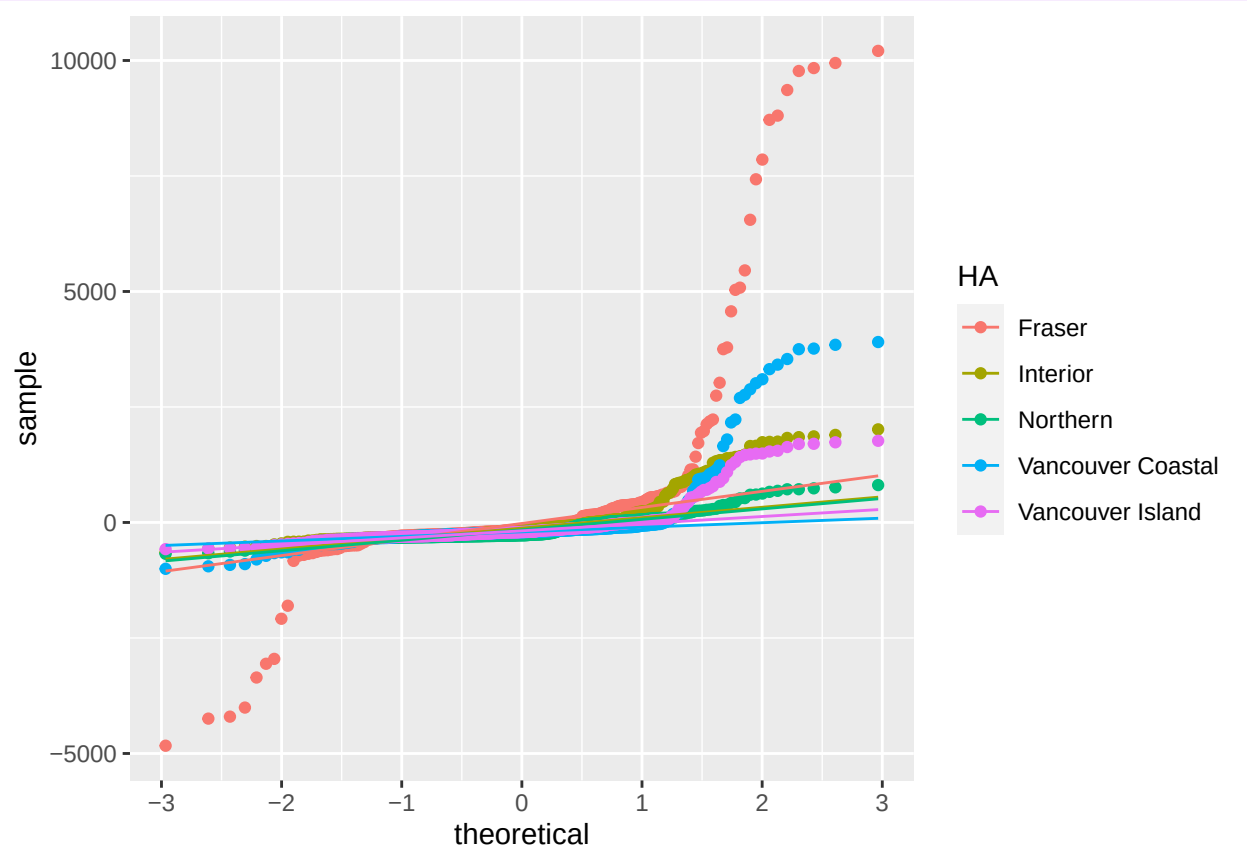
a.

```
x <- as.factor(bccovid_train$date)
lm <- lm(`cases+14`~`cases+0`+`cases-7`+`cases-14`, data = bccovid_train)
p1 <- predict(lm,newdata = bccovid_train)
res <- bccovid_train$`cases+14`- p1
qplot(
  data = bccovid_train,
  x = date,
  y = res,
  color = `HA`,
  xlab = "time",
  ylab = "residuals"
)
```



b.

```
ggplot(bccovid_train, aes(sample = res, colour = HA)) +
  stat_qq() +
  stat_qq_line() +
  labs(x = "theoretical" , y = "sample")
```



The plots seem not reasonable. Since the values in some sections of the plot differ locally from an overall linear trend so that the plot does not follow the normal distribution.

c.

```
lm <- lm(`cases+14`~`cases+0`+`cases-7`+`cases-14`, data = bccovid_train)
lm
```

```
##
## Call:
## lm(formula = `cases+14` ~ `cases+0` + `cases-7` + `cases-14`,
##     data = bccovid_train)
##
## Coefficients:
## (Intercept)    `cases+0`    `cases-7`   `cases-14`
##    325.4594       1.0816      -0.6069       0.1454
```

The slope coefficient on cases+0 is 1.0816. This coefficient means that as one person reported on this day, the cases+14 would increase by 1.0816. The interpretation makes sense since case+0 and case+14 should have a positive relationship.

d.

```
LOOCV <- function(data, estimator, predictor, error_fun) {
n <- nrow(data)
fold.label <- (1:n)
errors <- double(n)
for (fold in 1:n) {
test.rows <- fold.label == 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)
}

est <- function(dataset) lm(`cases+14` ~ `cases+0`+`cases-7`+`cases-14`,
                            data = dataset)
pred <- function(mod, dataset) predict(mod, newdata = dataset)
error_fun <- function(testdata) mutate(testdata, errs = (`cases+14` - .preds)^2) %>%
pull(errs) %>% mean()
LOOCV(bccovid_train,est,pred,error_fun)
```

```
## [1] 909099.9
```

The leave-one-out CV is 909099.9.

e. This model seems not appropriate it is because the residuals over time plot shows some certain pattern and the QQ-plot does not follow the normal distribution.

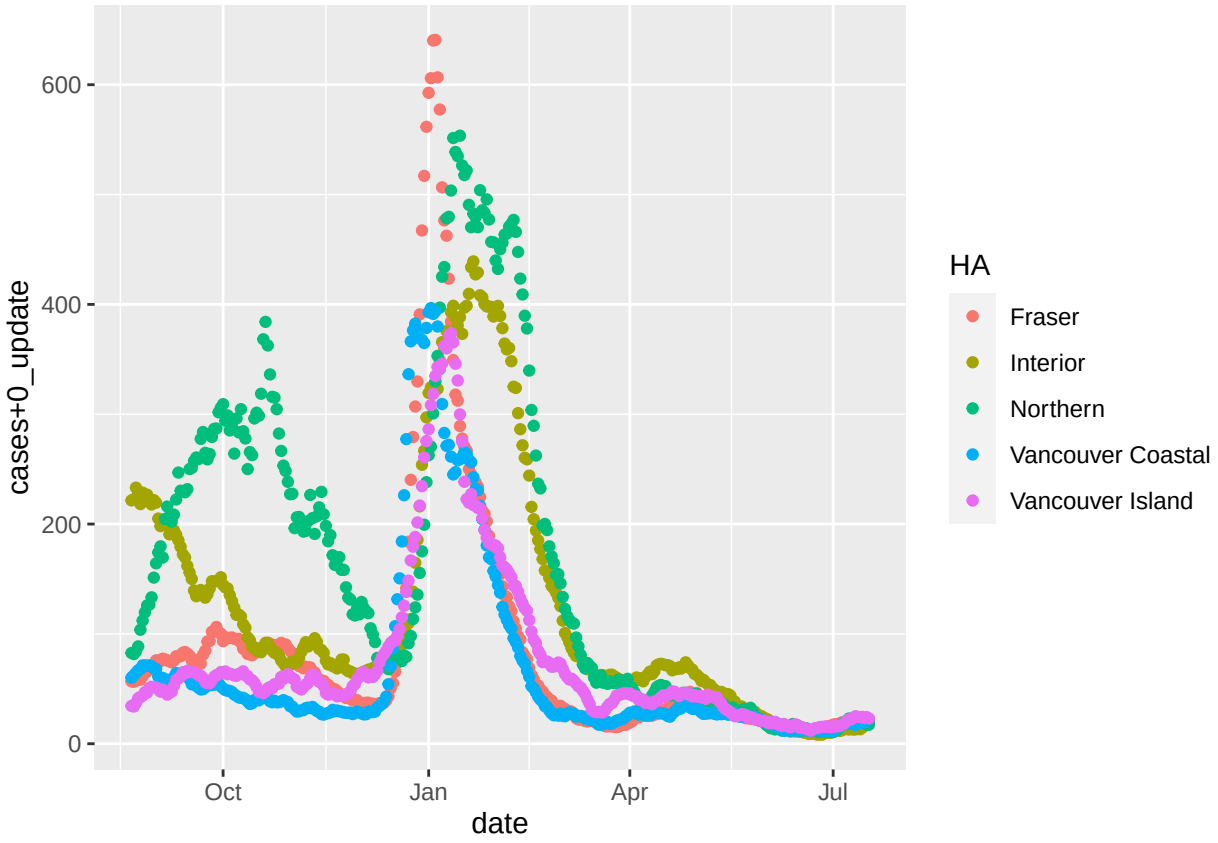
3. Examine the `bcpop` data set. Using an appropriate `*_join()`, add this column to the training data. Now scale all the columns by the population to get "cases per capita". Multiply by 100,000 just to get reasonable looking numbers.
    - Make a plot of today's cases over time scaled by population (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.

```
# ?bcpop
# Stat406::bcpop
# ?scale_y_log10
bccovid_train_new <- bccovid_train %>% left_join(bcpop,by="HA") %>%
  mutate(
    `cases+14_update` = `cases+14` / pop*100000,
    `cases+0_update` = `cases+0` / pop*100000,
    `cases-7_update` = `cases-7` / pop*100000,
    `cases-14_update` = `cases-14` / pop*100000
  ) %>%
select(HA,date,`cases+14_update`,`cases+0_update`,`cases-7_update`,
       `cases-14_update`,pop)


# scale
bccovid_train_scale <-  bccovid_train_new %>%
mutate(`cases+14_log` = log(`cases+14_update`))%>%
mutate(`cases+0_log` = log(`cases+0_update`))%>%
mutate(`cases-14_log` = log(`cases-14_update`))%>%
mutate(`cases-7_log` = log(`cases-7_update`))%>%
select(HA,date,`cases+14_log`,`cases+0_log`,`cases-7_log`,`cases-14_log`,pop)

ggplot(bccovid_train_new, aes(x=date, y=`cases+0_update`,colour=HA)) +
  geom_point()
```




```
ggplot(bccovid_train_new,aes(x=date, y=`cases+0_update`, color=HA)) +
  geom_point() + scale_y_log10()
```

Compared to the plot in Question 1, the new plot in Question 3 fluctuates more. The HA of Northern at the beginning of the date is not as flat as the plot in Question 1.

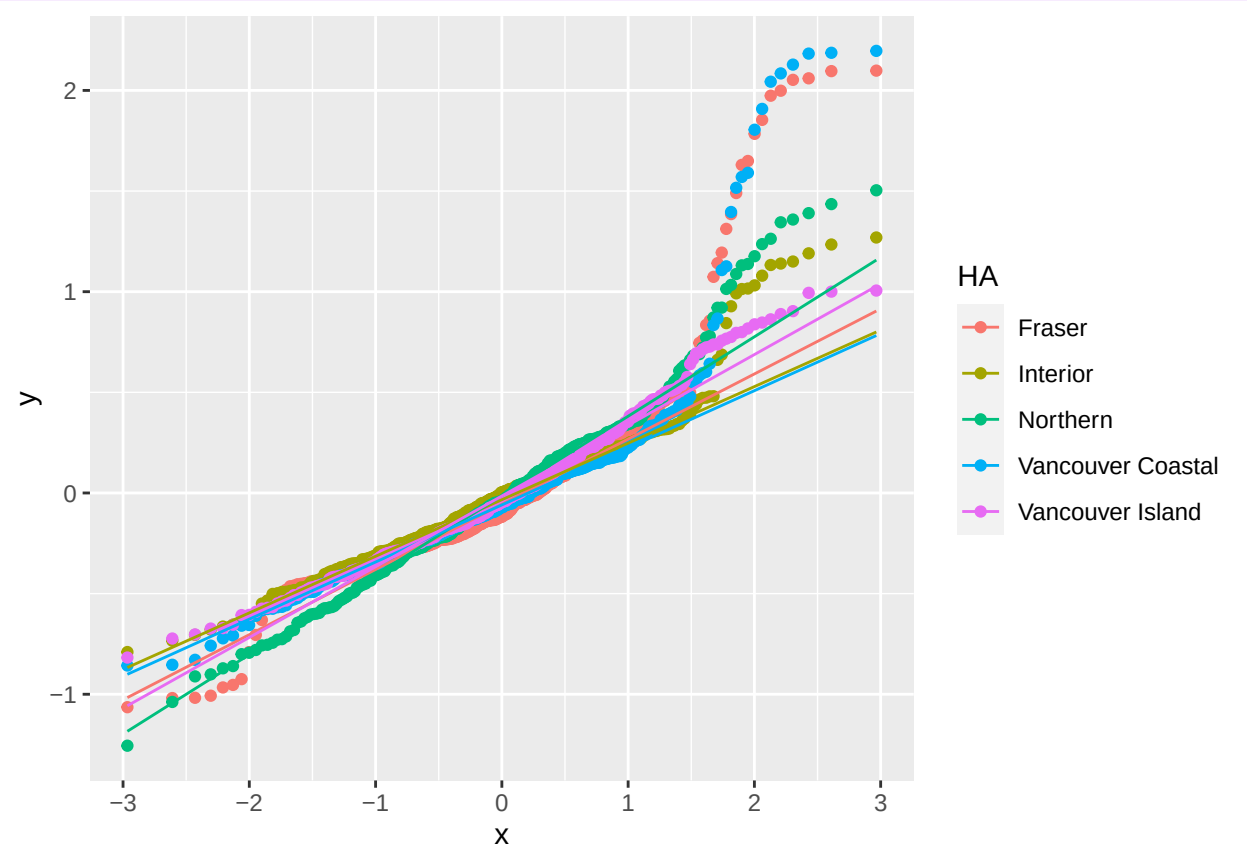
Repeat Q2 with different data.

a.

```
lm <- lm(`cases+14_log` ~ `cases+0_log` + `cases-14_log` + `cases-7_log`+ HA,
         data = bccovid_train_scale)
p2 <- predict(lm, newdata = bccovid_train_scale)
res <- bccovid_train_scale$`cases+14_log`- p2
ggplot(bccovid_train_scale,aes(x = date, y = res,color = HA))+ geom_point()
```

b.

```
ggplot(bccovid_train_scale, aes(sample = res,color = HA)) +
stat_qq() +
stat_qq_line()
```



The plots seem reasonable. Since the values of the plot are close to the overall linear trend, although it may have right skewed, it is more relative to the normal distribution.

c.

```
lm <- lm(`cases+14_log` ~ `cases+0_log` + `cases-14_log` + `cases-7_log`+ HA,
         data = bccovid_train_scale)
lm
```

```
##
## Call:
## lm(formula = `cases+14_log` ~ `cases+0_log` + `cases-14_log` +
##     `cases-7_log` + HA, data = bccovid_train_scale)
##
## Coefficients:
##         (Intercept)        `cases+0_log`       `cases-14_log`
##             0.64641              1.58687             -0.14218
##       `cases-7_log`           HAInterior           HANorthern
##             -0.61801              0.06406              0.09715
## HAVancouver Coastal   HAVancouver Island
##             -0.03265              0.01356
```

The slope coefficient on cases+0_log is 1.58687. This coefficient means that as one person reported on this

day, the cases+14_log would increase by 1.58687. The interpretation makes sense since case+0_log and case+14_log should have a positive relationship.

d.

```
est <- function(dataset) lm(`cases+14_log`~`cases+0_log`+`cases-14_log`+
                              `cases-7_log`+HA, data = bccovid_train_scale)
pred <- function(mod, dataset) predict(mod, newdata = dataset)
error_fun <- function(testdata)
  mutate(testdata, errs = (exp(`cases+14_log`)/100000*pop -
                             exp(.preds)/100000*pop)^2) %>%
pull(errs) %>% mean()
LOOCV(bccovid_train_scale,est,pred,error_fun)
```

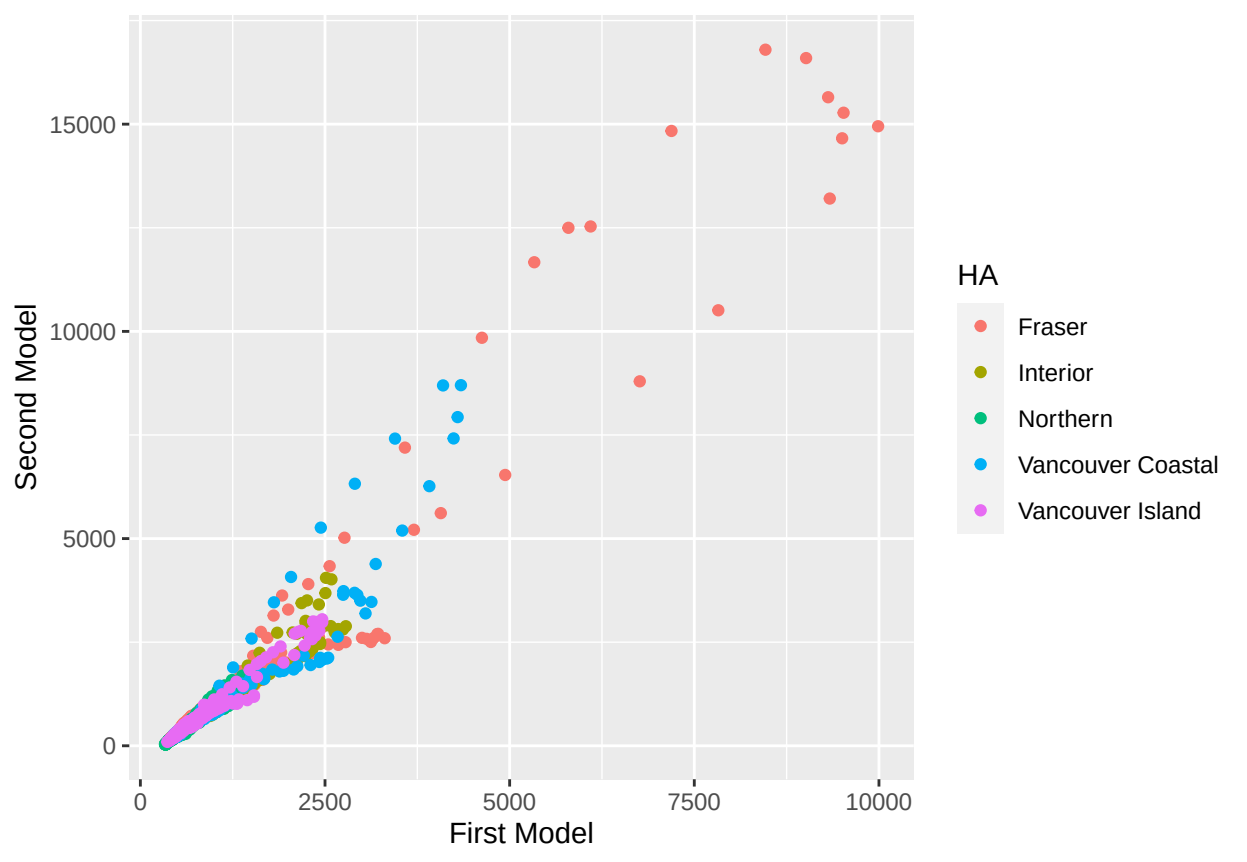
```
## [1] 1165350
```

The leave-one-out CV is 1165350.

e. The model seems appropriate. It is because the QQ-plot is close to the normal distribution and the residuals over time plot have a quite equal variance.

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?

```
a1 <- bccovid_train %>% left_join(bcpop,by = "HA")%>% cbind(p2) %>% cbind(p1)
a2 <- a1%>% mutate(predict = exp(p2)*pop/100000)
ggplot(a2,aes(x = p1, y = predict, color = HA)) + geom_point()+
labs(x = "First Model", y = "Second Model")
```



They are the linear trend. The HA of Fraser is always greater than Vancouver Coastal in both models.

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

```
print(mean(abs(a2$`cases+14` -a2$p1)))
```

```
## [1] 405.2396
```

```
print(mean(abs(a2$`cases+14` -a2$predict)))
```

```
## [1] 337.3383
```

```
print(sqrt(mean((a2$`cases+14` -a2$p1)^2)))
```

```
## [1] 943.4295
```

```
print(sqrt(mean((a2$`cases+14` -a2$predict)^2)))
```

```
## [1] 1079.514
```

For the first model, the mean absolute error is 405.2396; the root mean squared error is 943.4295. For the second model, the mean absolute error is 337.3383; root mean squared error is 1079.514. Therefore, the second model has lower MAE and higher MSE. The second model may experience higher influence from outliers.

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

We can not. It is because the covariance may be considerable due to the variables are not independent. Besides, each held-out set is n = 1, and the variance of the squared error is may large.

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?

```
#' @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)
}
```

```
est <- function(dataset) lm(`cases+14`~`cases+0`+`cases-7`+`cases-14`,
                            data = dataset)
pred <- function(mod, dataset) predict(mod, newdata = dataset)
error_fun <- function(testdata) mutate(testdata, errs =
                                         (`cases+14` - .preds)^2) %>%
pull(errs) %>% mean()
kfold_cv(bccovid_train,est,pred,error_fun,20)
```

```
## [1] 907823
```

```
est.2 <- function(dataset) lm(`cases+14_log`~`cases+0_log`+`cases-7_log`+
                                `cases-14_log`+HA, data = dataset)
error1 <- function(testdata) testdata %>%
mutate(error2 = (exp(`cases+14_log`) * pop / 100000 -exp(.preds) * pop /
                   100000)^2) %>%
pull(error2) %>% mean()
kfold_cv(bccovid_train_scale,est.2,pred,error1,20)
```

```
## [1] 1187994
```

a. The first model is 913059.4; and the second model is 1191354.

b. The second model is preferable.

c.

```
accurate <- (1191354 -  913059.4) / 913059.4
accurate * 100
```

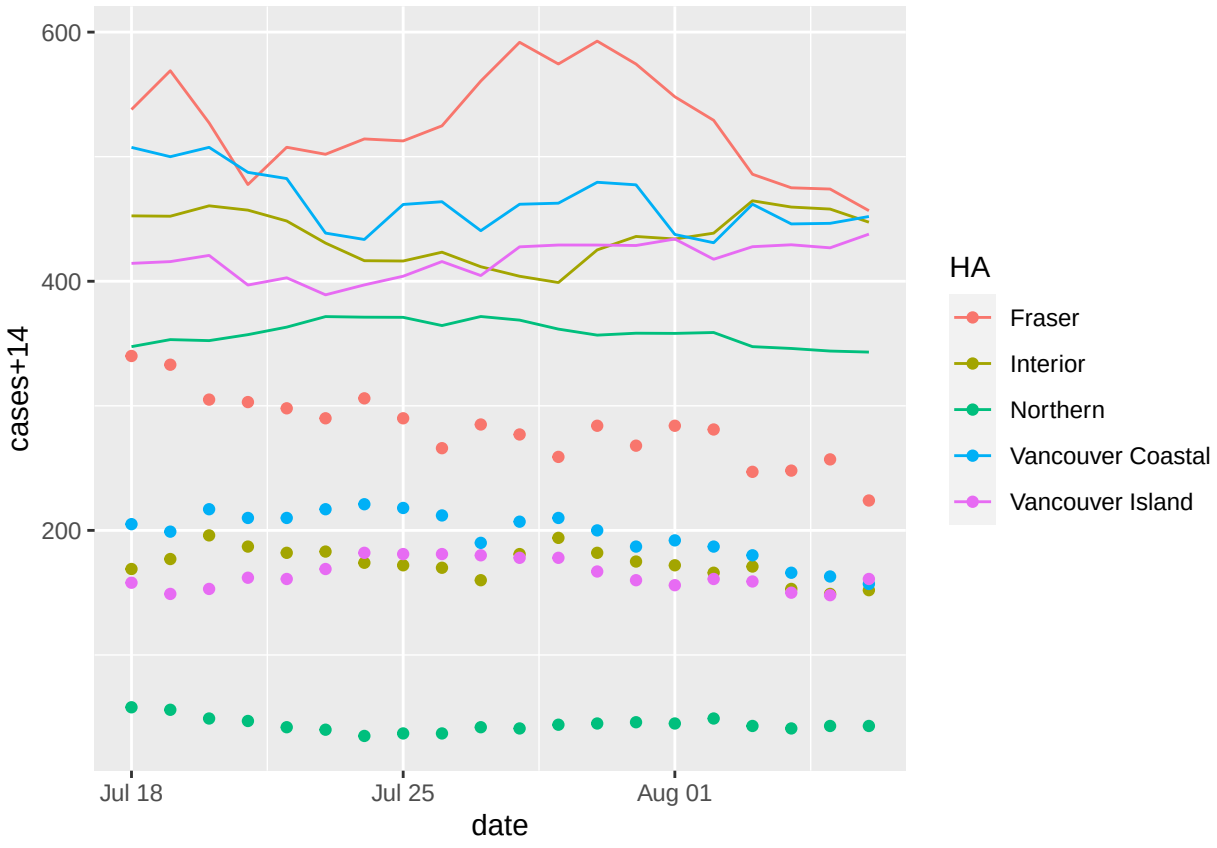
```
## [1] 30.47935
```

30.5% more accurate.

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.

```
bccovid_test$HA <- as.factor(bccovid_test$HA)
lm <- lm(`cases+14`~`cases+0`+`cases-7`+`cases-14`, data = bccovid_train)
p3 <- predict(lm,newdata = bccovid_test)
t <- p3 %>% cbind(bccovid_test)
ggplot(t,aes(x=date, y=`cases+14`, color=HA)) + geom_point() +
  geom_line(aes(x=date, y=p3, color=HA))
```



b.

```
t2 <- t %>% mutate(error = abs(p3 - `cases+14`)) %>%
select(error, HA) %>% group_by(HA) %>%
dplyr::summarise(error_mean = mean(error))
t2
```

```
## # A tibble: 5 x 2
##   HA                error_mean
##   <fct>                  <dbl>
## 1 Fraser                  245.
## 2 Interior                264.
## 3 Northern                314.
## 4 Vancouver Coastal       267.
## 5 Vancouver Island        253.
```

c.

```
overall <- t %>%
mutate(error = abs(p3 - `cases+14`)) %>% select(error, HA) %>%
dplyr::summarise(error_mean = mean(error))
overall
```

```
##   error_mean
## 1   268.3054
```

d.

```
error_fun <- function(testdata) mutate(testdata, errs = abs(`cases+14` - .preds)) %>%
pull(errs) %>% mean()
kfold_cv(bccovid_train,est,pred,error_fun,20)
```

```
## [1] 407.9847
```

The CV score is 407.8004; the mean absolute error is 268.3054, smaller than the previous.