A Predictive Model for Life Expectancy

Rachel Tsong and Adina Zhang

April 7, 2019

Introduction

Motivation

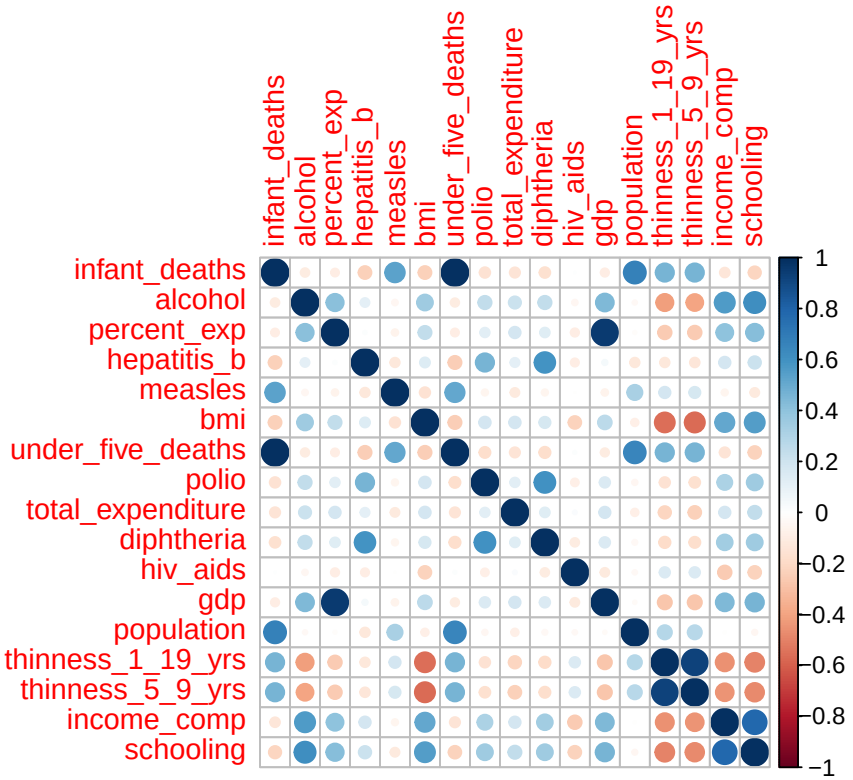
The dataset for our analysis was compiled by the Global Health Observatory (GHO), a branch of the World Health Organization (WHO). The repository contains data collected from 2000 to 2015 from 193 countries about factors relating to life expectancy. These factors include mortality rates of adults as well as infants and children, economic factors such as GDP and government health expenditures, disease related data such as incidence of measles and HIV, and health related variables such as BMI and alcohol consumption. Our analysis aims to answer the following questions:

- Which covariates are most predictive of life expectancy?
- What modeling method most accurately predicts life expectancy?

The results of analyses such as ours can be used by governments to direct resources and health care expenditures on the variables that are most strongly associated with life expectancy in order to improve the population's longevity and livelihood.

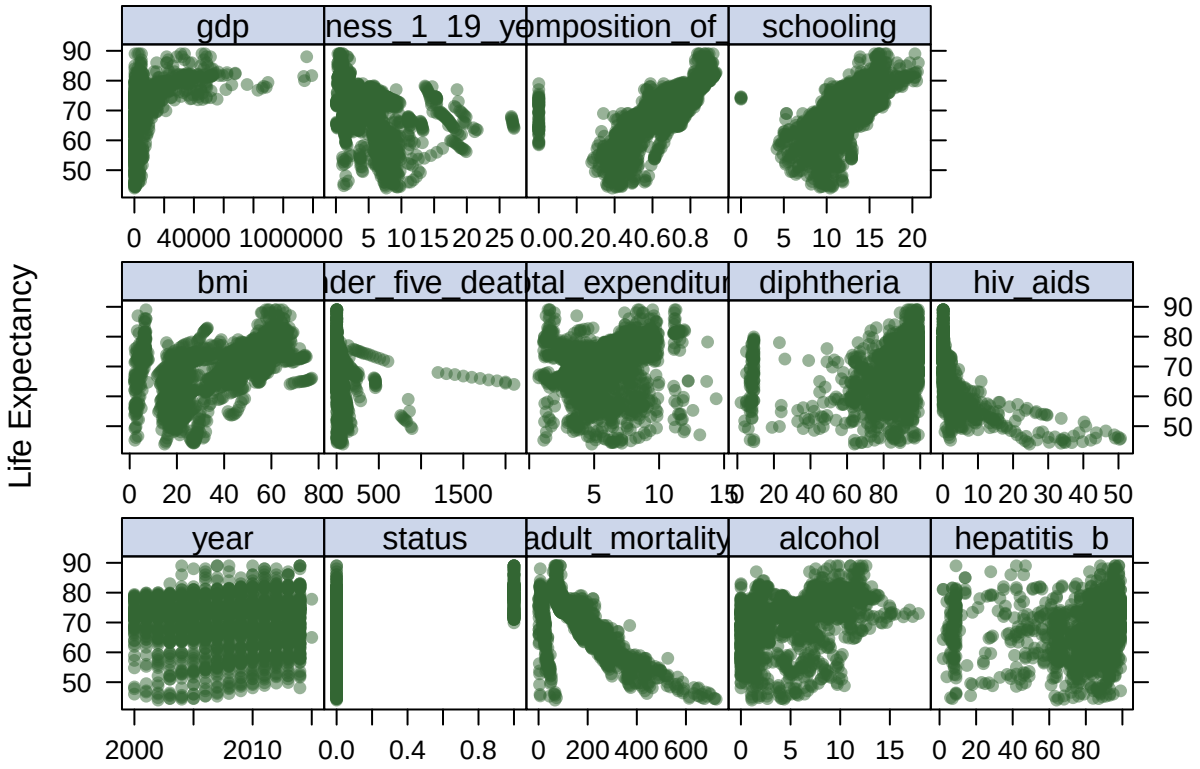
Data Cleaning

The original data set has 22 variables and 2938 observations from 193 countries between 2000 and 2015. A correlation matrix was used to assess the predictors to determine if there were any highly correlated variables. In order to avoid multicollinearity and to create a more parsimonious model, highly correlated variables were removed. Among the variables removed were infant deaths, percentage expenditure, measles, polio, and thinness for children ages 5 to 9 years. Population was also removed from the dataset because it is not an appropriate standardized measurement for life expectancy. Without adjusting population for area it is not a good comparative measurement. One dichotomous variable in our dataset that describes development status was recoded into an integer with groups 0 and 1. Any entries with NAs were omitted. The final analysis dataset has 15 variables with 1853 observations.



EDA

In exploratory data analysis, one of the main goals was to characterize the relationship of predictors with life expectancy. From the scatterplot in the figure below, we were able to conclude that several predictors potentially have a non-linear relationship with life expectancy. These predictors include GDP, thinness of children ages 1-19, BMI, total expenditure, diphtheria, HIV/AIDS, adult mortality, alcohol, and hepatitis B.



Model Building

Method

The data was split into training and test sets through 10-fold cross validation. Four models were fit using the training data including least squares, Lasso, GAM, and MARS. Tuning parameters for Lasso and MARS were chosen through cross validation. To compare model results, the fitted models were used for prediction on the teting sets and RMSE was calculated for comparison.

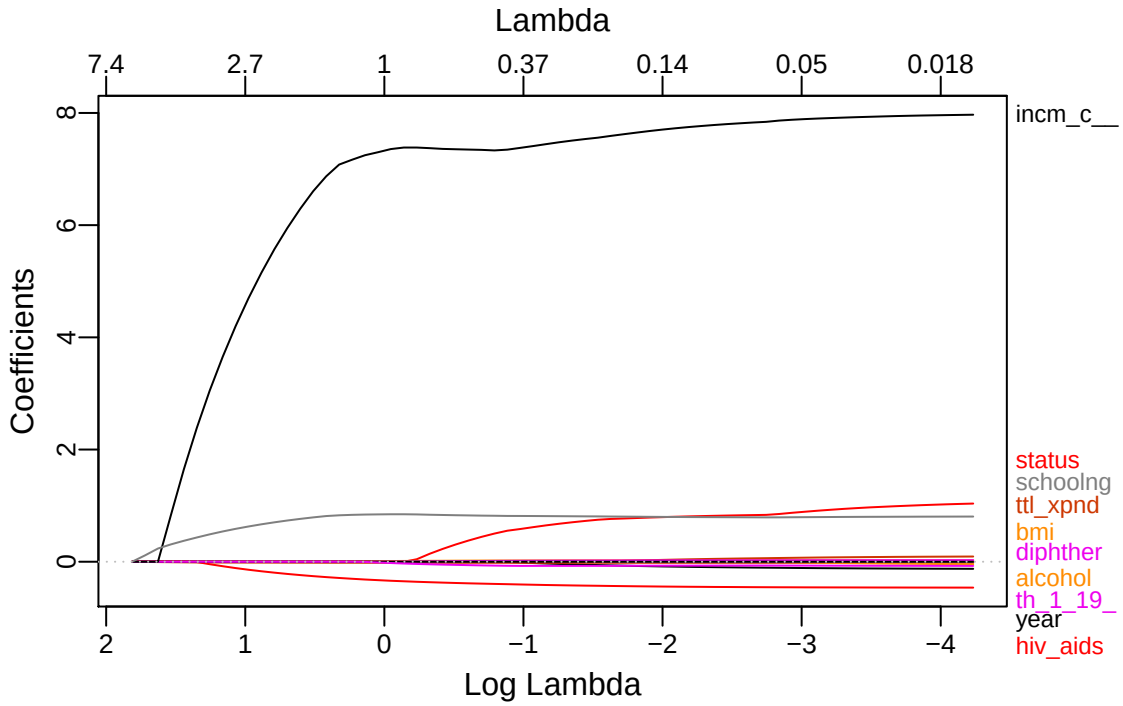
Variables

The reduced dataset contained 14 possible predictors for `life_expectancy` in years: `year`, `status` (a binary variable with 0 = developing and 1 = developed), `adult_mortality` (number of deaths per 1000 people aged 15-60), `alcohol` (average liters of alcohol consumed per capita), `hepatitis_b` (percent of 1 year olds immunized against Hep B), `bmi`, `under_five_deaths` (number of under 5 deaths per 1000), `total_expenditure` (percentage of government spending on health), `diphtheria` (percent of 1 year olds immunized againt diphtheria), `hiv_aids` (deaths from HIV/AIDS per 1000 among 0-4 year olds), `gdp`, `thinness_1_19_years` (prevalence of thinness among 1-19 year olds), `income_composition_of_resources` (ratio of income per capita and the human development index), and `schooling` (number of years).

Linear Models

First, we fit ordinary least squares regression and Lasso regression. We chose these because they are widely used and easily interpretable, but because it seemed as if there were many non-linear relationships we did not necessarily expect these models to fit the data the best. However, they provided baseline RMSEs that we could compare other models to. The assumptions for these types of regressions are linearity in parameters, homoscedasticity, and errors that are normal and uncorrelated. In our least squares model fit all

predictors were significant at the 5% level except alcohol, hepatitis B, and under five deaths. The covariates that were more predictive of life expectancy (meaning their coefficients were large and the p-value for the estimates were small) were adult mortality and income composition of resources. For the Lasso fit, an optimal lambda was chosen by cross-validation. Only hepatitis B was shrunk to zero. The figure below shows that income composition of resources were most predictive of life expectancy as it did not shrink. After 10-fold cross-validation was performed on the test sets, mean RMSEs for least squares and Lasso were 3.73 and 3.74, respectively. These models, though easily interpetable, lack flexibility as they cannot account for non-linear variables. In the next sections, we discuss two more models that have greater flexibility and might capture the true relationship more accurately.



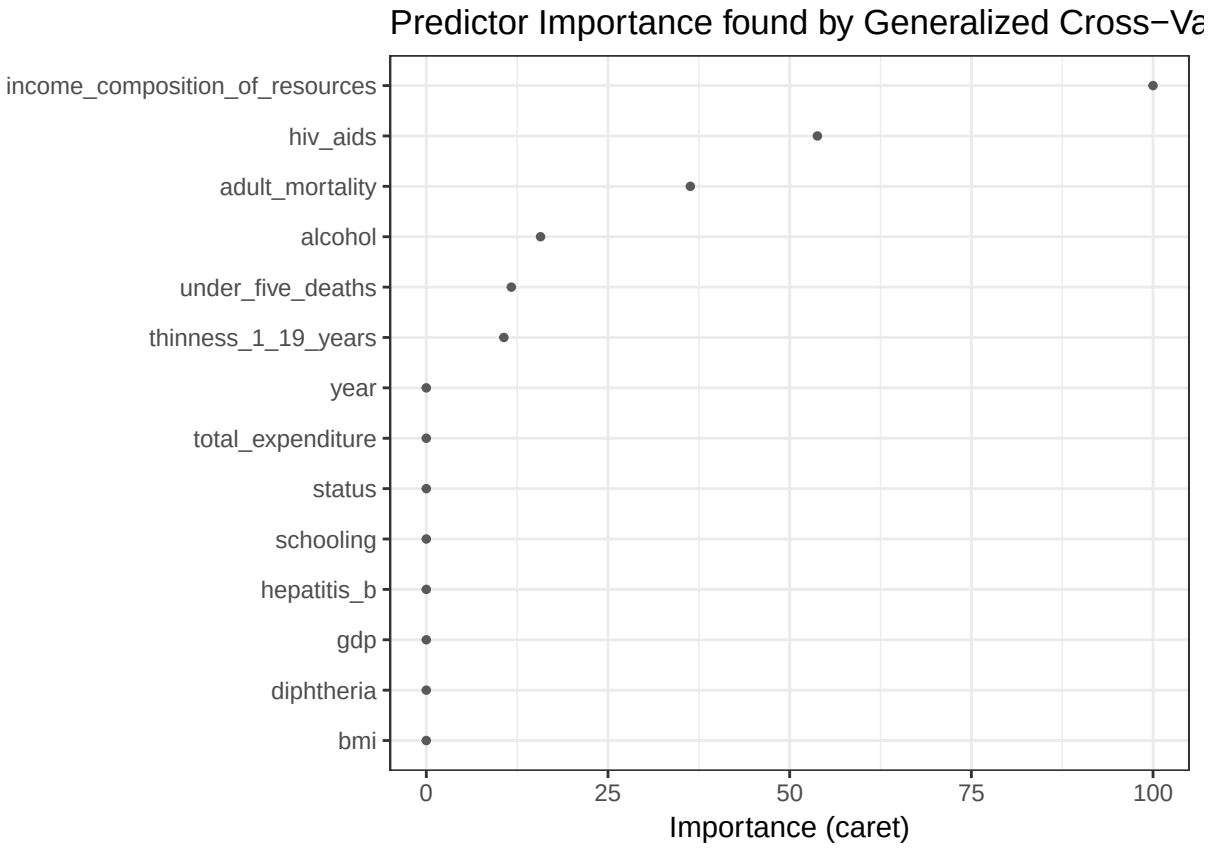
GAM

We fit a generalized additive model (GAM) using all the predictors with smoothing splines on GDP, thinness of children ages 1-19, BMI, total expenditure, diphtheria, HIV/AIDS, adult mortality, alcohol, and hepatitis B. These variables were chosen because the correlation plots appeared highly non-linear. Since we used the `mgcv::gam()` function, we did not have to specify the degrees of freedom for the splines. This model assumes additivity, thus important interaction terms can be missed by this method; however, it is advantageous to linear regression techniques because there can be both linear and non-linear predictors. One disadvantage is that it is difficult to interpret non-linear relationships. In our GAM fit, even when a smoothing spline was applied, hepatitis B was still non-significant at the 5% confidence level. As in the linear models in the previous section income composition was an important predictor of life expectancy. 10-fold cross-validation was performed on testing data, and the mean RMSE was 2.99.

MARS

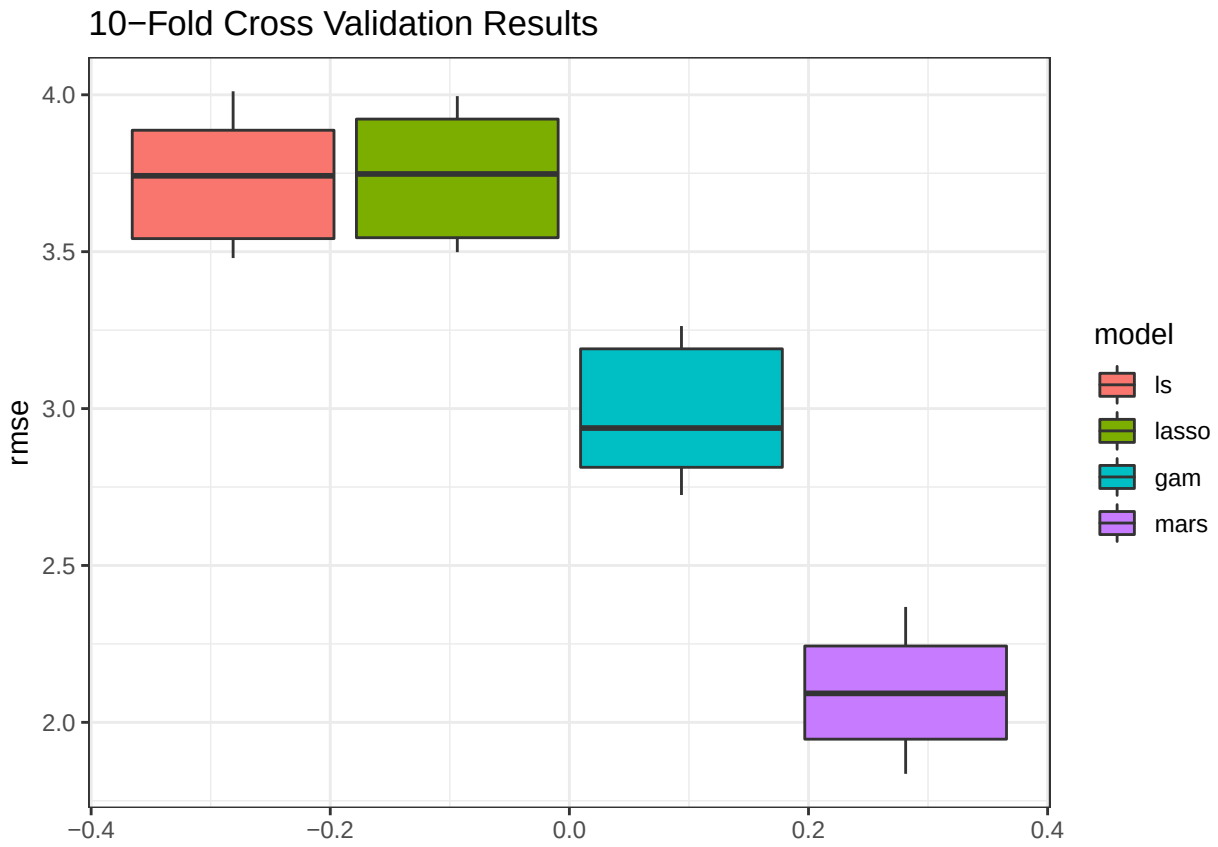
Lastly, we fit a multivariate adaptive regression spline (MARS) from the `earth` package. The optimal maximum degree and number of terms were chosen by cross-validation. The MARS model is advantageous

because it is highly flexible yet still interpretable. The relationships between covariate and outcome are linear, but the model fits knots to account for non-constant slopes, and the model can take into account interaction terms. In our MARS model, only 6 predictors were kept in the model. The figure below shows the decreasing order of importance of variables starting from income composition, HIV/AIDS, adult mortality, under five deaths, and ending with thinness 1-19 years. Additionally, 1 knot was created for adult mortality, under five deaths, HIV/AIDS, and thinness, indicating that these predictors have a non-constant relationship to life expectancy. The model added 17 interaction terms, suggesting that many of the predictors have effects on the others. After subsetting the data into test and training sets, 10-fold cross validation was performed, and the mean RMSE was found to be 2.09.



Model Comparison

The results of 10-fold cross validation are shown in the figure below. Both GAM and MARS had significantly lower RMSEs than either least squares or lasso regression which performed similarly. Ultimately, MARS fit the best model as it has the lowest RMSE. As discussed previously, the reasons why GAM and MARS perform better is because they allow flexibility for non-parametric variables. MARS outperforms GAM because the method takes into account any interaction terms that exist, and it eliminates any variables not deemed significant to the outcome, thereby creating a more parsimonious model.



Conclusions

From our analysis, it was determined that the best model to predict life expectancy was the MARS method. This was expected because the MARS method offered both flexibility when handling non-parameteric variables and a parsimonious model. Furthermore, the MARS method takes into account interaction terms that may exist which is ignored in the GAM method. When it came to predicting life expectancy and understanding which parameters were most significant in prediction, income composition of resources was consistently important in least squares, Lasso, and MARS. The significance of this variable makes sense as it is a ratio of the gross national income per capita and human development index (HDI). The HDI is a measurement that takes into consideration the life expectancy at birth and amount of schooling. Thus, in order to extend life expectancy, it would be important for a country to devote resources towards increasing income per capita as well as improving health outcomes and increasing schooling years.

Appendix

Data Cleaning

```
# Load dataset
life_analysis = read_csv("./Life Expectancy Data.csv") %>%
  janitor::clean_names() %>%
  select(-country, -population, -infant_deaths, -measles,
         -polio, -thinness_5_9_years, -percentage_expenditure) %>%
  mutate(status = factor(status),
         status = fct_recode(status, "0" = "Developing", "1" = "Developed"),
         status = ifelse(status == "1", 1, 0))
life_analysis = na.omit(life_analysis)
```

Correlation Plot

```
# Correlation plot of only continuous variables
cont = life_exp %>% select(-country, -status)
cont = na.omit(cont) %>%
  rename(income_comp = income_composition_of_resources,
         percent_exp = percentage_expenditure,
         thinness_1_19_yrs = thinness_1_19_years,
         thinness_5_9_yrs = thinness_5_9_years)
x = model.matrix(life_expectancy ~., cont)[,-c(1,3)]
corrplot(cor(x[,-1]))
```

Scatterplot

```
# Scatterplot
x = model.matrix(life_expectancy ~., life_analysis)
y = life_analysis$life_expectancy

theme1 <- trellis.par.get()
theme1$plot.symbol$col <- rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch <- 16
theme1$strip.background$col <- rgb(.0, .2, .6, .2)
trellis.par.set(theme1)
featurePlot(x[,-1], y, plot = "scatter", labels = c("","Life Expectancy"))
```

Tuning parameters for Lasso and MARS

```
# Tuning parameters

# Tuning lambda for lasso through cross validation
set.seed(123)
cv_lasso = cv.glmnet(x, y, alpha = 1)
best_lambda = cv_lasso$lambda.min
plot_glmnet(cv_lasso$glmnet.fit)

# Tuning for MARS
# Create a tuning grid
hyper_grid = expand.grid(
  degree = 1:3,
  nprune = seq(2, 70, length.out = 20) %>% floor()
  )
```

```r
set.seed(1)
# Cross validated model
tuned_mars = train(
  x = subset(life_analysis, select = -life_expectancy),
  y = life_analysis$life_expectancy,
  method = "earth",
  metric = "RMSE",
  trControl = trainControl(method = "cv", number = 10),
  tuneGrid = hyper_grid
)

# best model
tuned_mars$bestTune

ggplot(tuned_mars)
```

Variable Importance Plot - MARS

```r
# Variable importance plot
plot = vip(tuned_mars,
           num_features = 15,
           bar = FALSE,
           value = "gcv") +
  ggtitle("Predictor Importance found by Generalized Cross-Validation") +
  theme_bw()
plot
```

10-fold cross validation with four modeling methods

```r
# Function to indicate y outcomes
y_data = function(data){
  y = data$life_expectancy
}

# Function to calculate Lasso RMSE from test dataset
lasso_rmse = function(model, x2, y2){
  predictions = model %>% predict(x2) %>% as.vector()
  rmse_lasso = RMSE(predictions, y2)
  return(rmse_lasso)
}

set.seed(281)
# Set up 10-fold cross validation
# Create training and test datasets
cv_df = crossv_kfold(life_analysis, k = 10) %>%
  mutate(train = map(train, as_tibble),
         test = map(test, as_tibble),
         x = map(train, ~model.matrix(life_expectancy ~ ., data = .x)[,-3]),
         x2 = map(test, ~model.matrix(life_expectancy ~ ., data = .x)[,-3]),
         y = map(train, y_data),
         y2 = map(test, y_data))

# Fit four models: Least squares, lasso, GAM, MARS
```

```r
# Tuning parameters already selected
# Calculate RMSE from each model
cv_df = cv_df %>%
  mutate(ls_mod = map(train, ~lm(life_expectancy ~ ., data = .x)),
         lasso_mod = map2(x, y, ~glmnet(x = .x,
                                        y = .y, alpha = 1,
                                        lambda = best_lambda)),
         gam = map(train, ~ gam(life_expectancy ~ s(gdp) +
                                  s(hiv_aids) +
                                  s(alcohol) +
                                  s(bmi) +
                                  s(hepatitis_b) +
                                  s(total_expenditure) +
                                  s(thinness_1_19_years) +
                                  s(under_five_deaths) +
                                  year +
                                  status +
                                  adult_mortality +
                                  income_composition_of_resources +
                                  schooling +
                                  s(diphtheria),
                                data = .x)),
         mars = map(train, ~ earth(life_expectancy ~ . , data = life_analysis,
                                   degree = 2, nprune = 27), data = .x)) %>%
  mutate(rmse_ls = map2_dbl(ls_mod, test, ~rmse(model = .x, data = .y)),
         rmse_lasso = pmap_dbl(list(lasso_mod, x2, y2), lasso_rmse),
         rmse_gam = map2_dbl(gam, test, ~ rmse(model = .x, data = .y)),
         rmse_mars = map2_dbl(mars, test, ~ rmse(model = .x, data = .y)))

# Boxplot of RMSE from all four model fits
cv_df %>%
  select(starts_with("rmse")) %>%
  gather(key = model, value = rmse) %>%
  mutate(model = str_replace(model, "rmse_", ""),
         model = fct_inorder(model)) %>%
  ggplot(aes(y = rmse, fill = model)) +
  geom_boxplot() +
  labs(
    title = "10-Fold Cross Validation Results"
  ) +
  theme_bw()
```