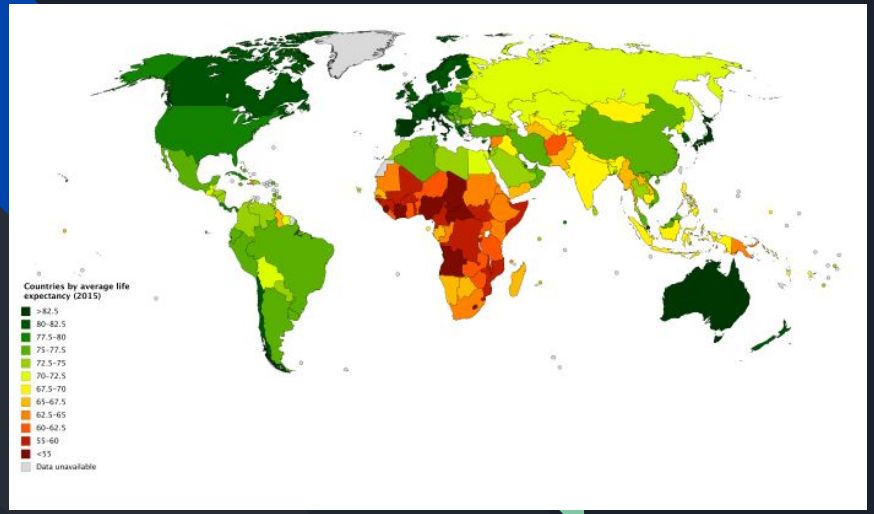


“Don’t get AIDS”: Making inferences about **life expectancy**

World Health Organization

By Aaron Chumsky, Alex Jordan, Tianke Li, Sam Ressin, and Litian Zhou

Problem Overview

- **Problem:** What factors contribute most to a nation's average life expectancy?
- **Response:** Life expectancy (measured in average years)
- Goal of the project: two pronged approach:
  - Individuals
  - Governments
- Why this problem right now?

Our 

- 193 countries, years 2000-2015
- 21 predictors, 1 response (life expectancy)
- Missing data: only in certain variables
  - Full dataset: 2,938 observations
  - Without missing data: 1,649 observations
- Categories of predictors:
  - Mortality: deaths of age groups
  - Public health: vaccination rates, health, diseases
  - Socio-economic factors: dev’t status, population

Our methods

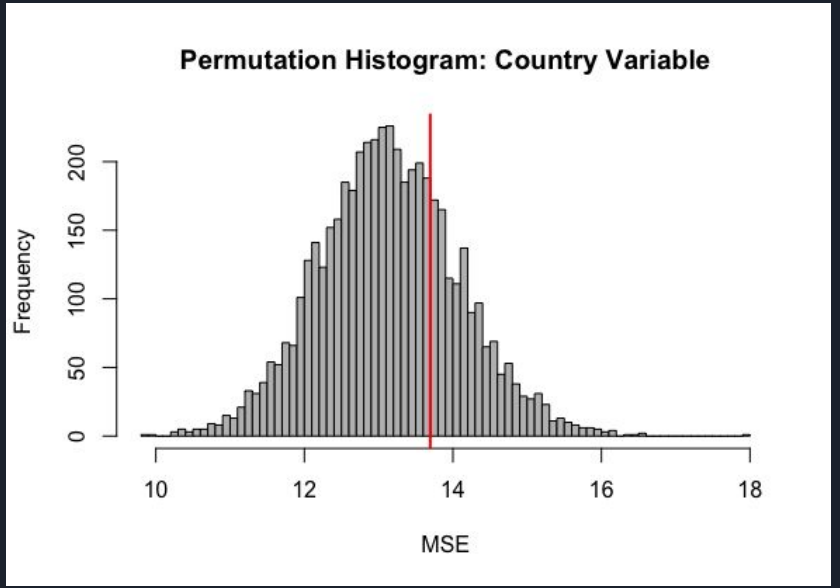
- Permutation
- Linear Regression
- Variable Selection
- Regularization
- Dimensionality Reduction
- Non-linear Regression
- Tree Based Methods

Permutation on Country

5000 permutations of three variables to test for significance:

- Country: 0.270 = p > alpha = 0.05
  - Original Test MSE: 13.78
- Year: [1]p = 0.278, [2]p = 0.337)
  - Original Test MSE: 13.55
- Status: [1]p = 0.305, [2]p = 0.306)
  - Original Test MSE: 13.69

MSE of linear model, permuting Country variable

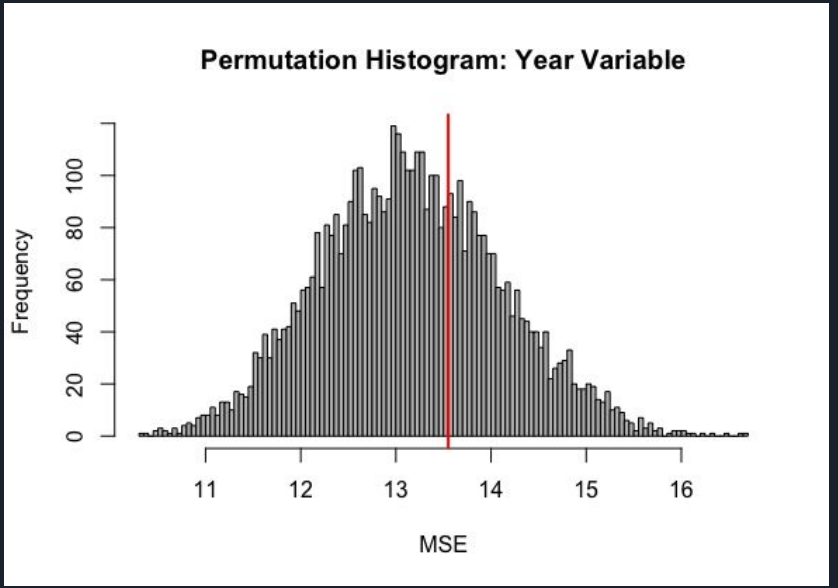


Permutation on Year

5000 permutations of three variables to test for significance:

- Country: 0.270 = p > alpha = 0.05
  - Original Test MSE: 13.78
- Year: [1]p = 0.278, [2]p = 0.337)
  - Original Test MSE: 13.55
- Status: [1]p = 0.305, [2]p = 0.306)
  - Original Test MSE: 13.69

MSE of linear model, permuting Country variable



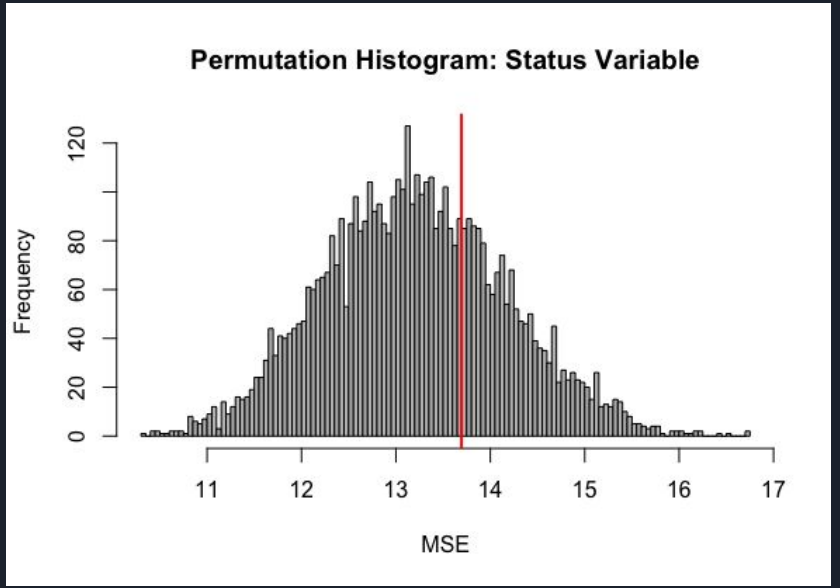
Permutation on Country Status

(Developed or developing)

5000 permutations of three variables to test for significance:

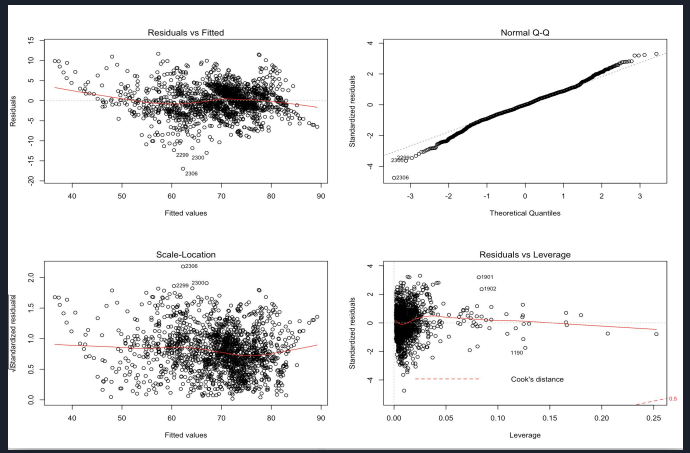
- Country: 0.270 = p > alpha = 0.05
  - Original Test MSE: 13.78
- Year: [1]p = 0.278, [2]p = 0.337)
  - Original Test MSE: 13.55
- Status: [1]p = 0.305, [2]p = 0.306)
  - Original Test MSE: 13.69

MSE of linear model,
permuting Country variable



Full Linear Model

- Basic linear model
- Contains 19 predictors
- R^2=83.56%
- 8 variables are not t-test significant
  - Each of them is significant individually
- Serve as a benchmark




```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.348e+01  7.375e-01  72.515  < 2e-16 ***
Status                          9.684e-01  3.379e-01   2.865  0.00422 **
Adult.Mortality                -1.663e-02  9.494e-04 -17.517  < 2e-16 ***
infant.deaths                   9.350e-02  1.065e-02   8.777  < 2e-16 ***
Alcohol                        -9.140e-02  3.316e-02  -2.756  0.00592 **
percentage.expenditure          3.673e-04  1.801e-04   2.040  0.04156 *
Hepatitis.B                    -6.525e-03  4.449e-03  -1.467  0.14265
Measles                        -7.865e-06  1.079e-05  -0.729  0.46597
BMI                             3.376e-02  5.998e-03   5.628 2.15e-08 ***
under.five.deaths              -7.035e-02  7.711e-03  -9.123  < 2e-16 ***
Polio                           7.935e-03  5.152e-03   1.540  0.12370
Total.expenditure               7.586e-02  4.067e-02   1.865  0.06236 .
Diphtheria                      1.490e-02  5.928e-03   2.513  0.01205 *
HIV.AIDS                       -4.370e-01  1.784e-02 -24.490  < 2e-16 ***
GDP                             8.738e-06  2.837e-05   0.308  0.75813
Population                     -6.425e-10  1.749e-09  -0.367  0.71337
thinness..1.19.years           -1.238e-02  5.300e-02  -0.234  0.81527
thinness.5.9.years             -4.798e-02  5.231e-02  -0.917  0.35917
Income.composition.of.resources 9.817e+00  8.321e-01  11.797  < 2e-16 ***
Schooling                       8.665e-01  5.940e-02  14.587  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Stepwise Selection

- Forward Selection chooses 13 predictor model
  - R^2=83.51%
- Backward and bidirectional Selection produce chooses 15 predictor model
  - R^2=83.55%

```
Life.expectancy ~ Adult.Mortality + infant.deaths + percentage.expenditure +
    BMI + under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources +
    Schooling
```

```
Life.expectancy ~ Schooling + HIV.AIDS + Adult.Mortality + Income.composition.of.resources +
    percentage.expenditure + BMI + Diphtheria + under.five.deaths +
    infant.deaths + Status + Alcohol + thinness.5.9.years + Total.expenditure
```

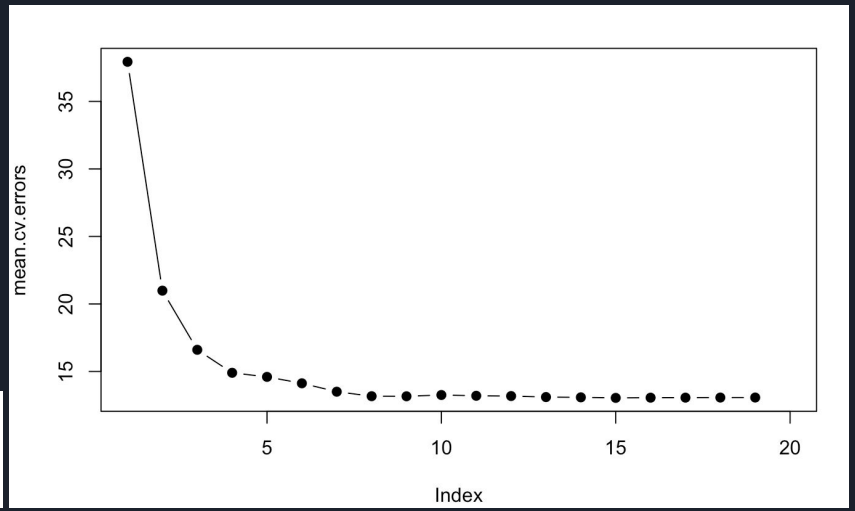
Differences:
Hepatitis.B & Polio

Best Subset Selection

- Adjusted R^2 chooses the model with 15 predictors
  - Same model as backward/bidirectional stepwise selection
- Mallow's Cp chooses the model with 13 predictors
  - Same model as forward stepwise selection
- BIC chooses the model with 9 Predictors
  - R^2=83.32%

```
Life.expectancy ~ Adult.Mortality + infant.deaths + percentage.expenditure +
    BMI + under.five.deaths + Diphtheria + HIV.AIDS + Income.composition.of.resources +
    Schooling
```

10-fold Cross Validation Test Error across all Predictor Levels



Regularization

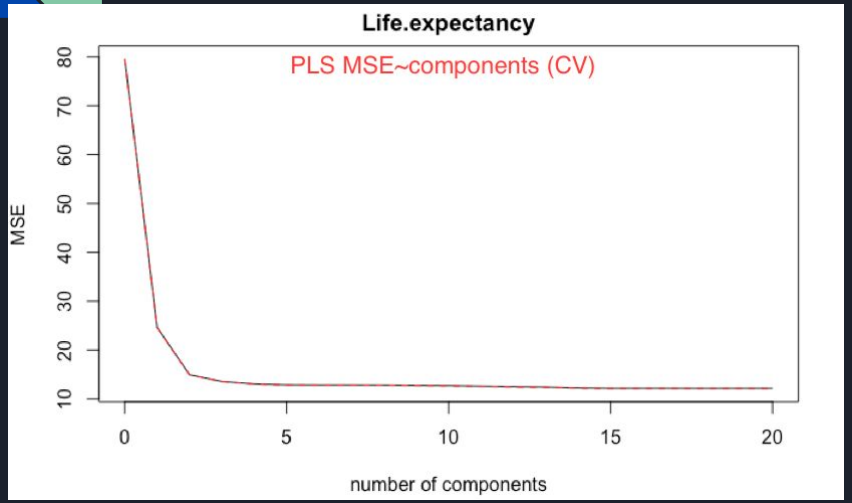
LASSO regression

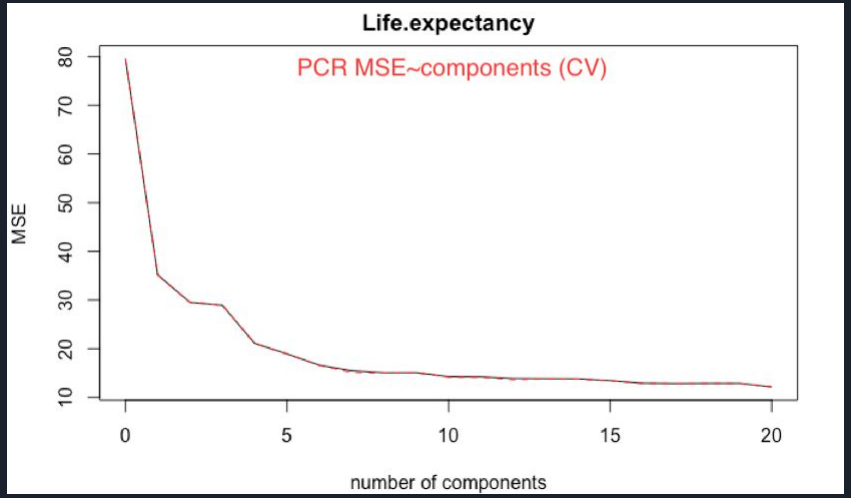
- 10-fold cross validation
- lambda=.0021
- Cannot reduce any predictor
- Perform very similar to linear regression (full model)

Ridge regression

- 10-fold cross validation
- lambda=.0010
- Perform very similar to linear regression (full model)

Dimensionality Reduction (PLS & PCR)





To balance the interpretability and model accuracy, I chose:

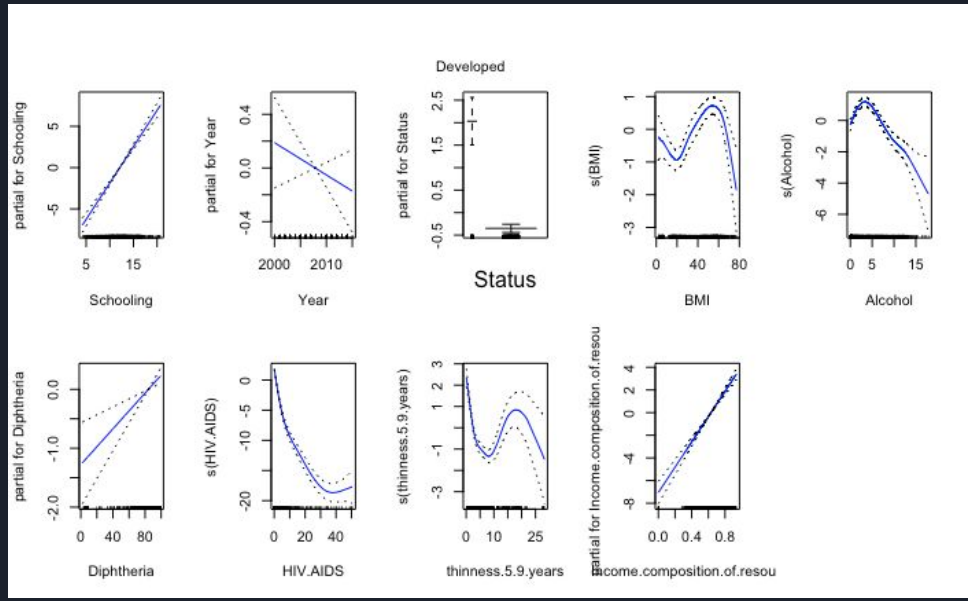
- M=5 in PLS
- M=10 in PCR

Non-linear model - GAM

- Fit 10 CV-chosen variables
- ANOVA table:
  - Only one has relatively low significance
    - Government expenditure on health
    - (p = 0.001)

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)	
1	1647	60009				
2	1646	59985	1.0000	24.5	0.149053	
3	1645	59160	1.0000	824.7	< 2.2e-16	***
4	1641	51395	4.0001	7765.8	< 2.2e-16	***
5	1637	48722	4.0001	2672.9	< 2.2e-16	***
6	1636	48597	1.0000	125.1	0.001098	**
7	1635	47673	1.0000	923.1	< 2.2e-16	***
8	1631	22994	4.0002	24679.9	< 2.2e-16	***
9	1627	21220	3.9998	1773.8	< 2.2e-16	***
10	1626	19099	1.0000	2120.8	< 2.2e-16	***

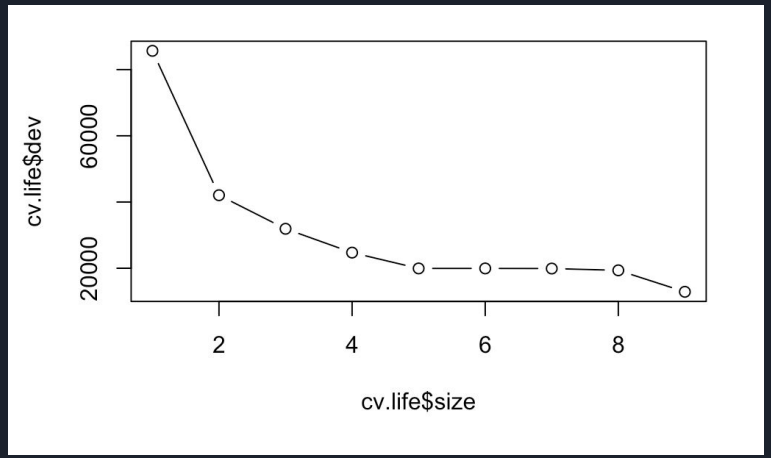
- Interesting interpretations:
  - Scholars live longer
  - The rich live longer
  - BMI - very flexible
  - Alcohol - happy little peak
  - Immunization (DTP3 vaccine) works!



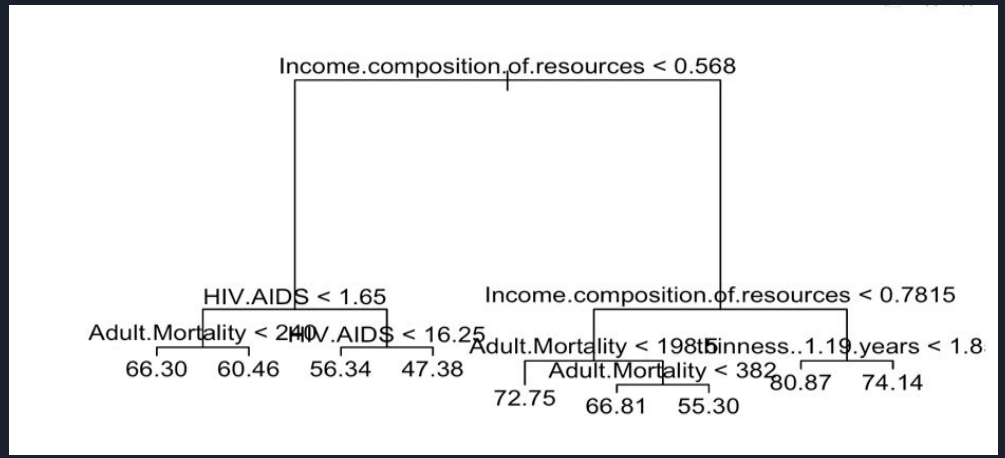
Decision Trees

- Four variables used
- Cross Validation chose nine splits for optimal pruning
- Test MSE around 9.7

Cross Validation for Pruning



Decision Tree Output



Bagging & Random Forest

- For Bagging:
  - 500 bootstrapped trees
  - Test MSE of 3.6
  - 94.76% variance explained
- For Random Forest
  - Mtry of 6
  - Test MSE of 3.67

Variable Importance Table

	%IncMSE	IncNodePurity
Status	6.162644	159.9761
Adult.Mortality	37.367787	16348.5791
infant.deaths	16.910423	1046.0087
Alcohol	22.483156	1244.8038
percentage.expenditure	15.970165	1312.9603
Hepatitis.B	14.081147	379.2195
Measles	14.607869	460.9453
BMI	18.676947	3757.2410
under.five.deaths	18.518176	1553.6357
Polio	13.592255	475.1583
Total.expenditure	24.701756	905.7842
Diphtheria	12.028144	503.3254
HIV.AIDS	31.390865	17664.0541
GDP	13.863620	1505.0965
Population	12.642476	421.1627
thinness..1.19.years	19.164170	2740.6658
thinness.5.9.years	22.209109	3484.1148
Income.composition.of.resources	32.122142	23368.5433
Schooling	19.408247	7759.3831

Summary of Results

- Lasso and Ridge very similar to OLS
- PCR performs the worst
- Non-linear GAM has lowest test error
- Agreement over important variables

Test MSE

OLS Full	OLS Reduced	Ridge (lambda)	Lasso (lambda)	PCR (# of components)	PLS (# of components)	GAM
13.72	13.95	13.71 (.01)	13.73 (.01)	15.9 (5)	14 (3)	8.85

Test R-Squared

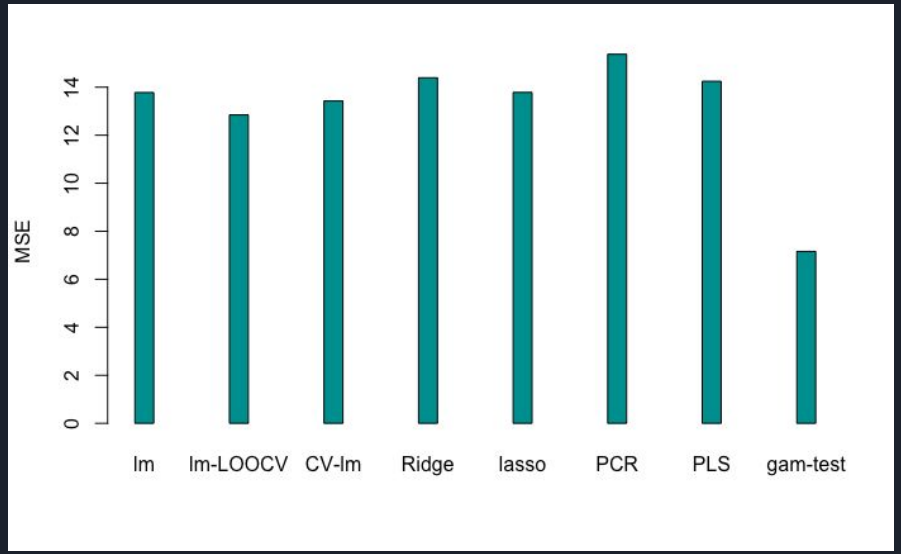
OLS Full	OLS Reduced	Ridge	Lasso	PCR	PLS	GAM
80.4	80.24%	80.6%	80.6%	77.5%	80%	88.5%

Variable Importance

- Important Variables:
  - Percent expenditure on health
  - Deaths from AIDS
  - # of years Schooling
  - Income Composition of Resources
  - BMI
  - **Adult Mortality (**not useful in practice)
    - Removal of this variable only reduced R^2 by 3%

Conclusions

- Most models have similar MSE
- GAM → Best performance
  - Most complex
- Reduced OLS→ Adequate performance
  - Best interpretability
- Dataset and model account for most factors relevant to life expectancy



Implications

To increase life expectancy…

Government Actions

- Increase Education rate
- Address Economic development
- Control spread HIV/AIDS
- Invest more on health

Individual Actions

- Lower BMI
- Take precaution against HIV/AIDS
- Educate themselves
- Get rich

Thank you!

- Keep healthy and live long !!