Identifying Socioeconomic Determinants of Healthcare Quality in US

By Erin Kim, Isabella Hu, Kiela King, Ethan Lau, Michael Xu

Table of Contents

01. Intro

Background Info, Question, Variables, Hypotheses

02. Data Prep

Data Cleaning/Merging

03. EDA

Scatterplots, Histograms, Heatmaps

04. Analysis

Linear Regression, Logistic Regression, Random Forest, Neural Network

05. Conclusion

Interpretation of Results, Significance

The Problem

- Healthcare disparity within US
  - High costs
  - Access to care
  - Quality of care

Research Question

What are the most significant socioeconomic factors that influence the quality of healthcare in the US?

Data Preparation

Independent Variables

Percentage of State Pop. Uninsured

Medicare Estimates by State Residence - Personal Health Care (Millions)

Medicaid Estimates by State Residence - Personal Health Care(Millions)

Health Care Expenditures Per Capita

GDP by State

Unemployment Rate by State

Poverty Rate by State

High School Graduation Rate by State

Population with high Medical Burden Cost (%)

Average Median Income per Household

Gini Coefficient by State (Income Inequality)

Dependent Variable

Life Expectancy by State (years)

Data Sources

CMS.gov

KFF

Data Cleaning/Merging

Thank you Tim!

Training Data (2016-2018)

year	state	Longevity	uninsured	Medicare	Medicaid	cost pc	GDP	Unemployment	poverty	Grad Rate	Income	gini	burden
2016	Alabama	75.38	9.1	10485	4818	0.28616	207368400000	5.7	17.1	89	46257	0.4847	0.28616
2016	Alaska	78.06	14	869	1860	0.18287	50727700000	6.7	9.9	78	76440	0.4081	0.18287
2016	Arizona	79.47	10	12276	9455	0.22943	313081400000	5.3	16.4	78	53558	0.4713	0.22943
2016	Arkansas	75.84	7.9	6003	5709	0.24664	119152400000	3.8	17.2	88	44334	0.4719	0.24664
2016	California	80.88	7.3	71628	77965	0.17715	2569634000000	5.4	14.3	83	67739	0.4899	0.17715
2016	Colorado	80.18	7.5	7791	7408	0.22756	329911700000	2.7	11	81	65685	0.4586	0.22756
2016	Connecticut	80.8	4.9	7814	7600	0.20171	263670300000	4.6	9.8	88	73433	0.4945	0.20171
2016	Delaware	78.64	5.7	2203	1658	0.18941	69355000000	4.7	11.7	87	61757	0.4522	0.18941
2016	Florida	79.6	12.5	1117	2660	0.23375	953352500000	4.8	14.7	82	50860	0.4852	0.23375
2016	Georgia	77.37	12.9	51599	18261	0.22502	547546700000	5.3	16	82	53559	0.4813	0.22502
2016	Hawaii	81.25	3.5	17010	8981	0.14557	83914400000	2.7	9.3	85	74511	0.442	0.14557
2016	Idaho	79.15	10.1	2269	1884	0.25492	68837400000	3.5	14.4	80	51807	0.4503	0.25492
2016	Illinois	79.09	6.5	2659	1632	0.23385	807043200000	5.4	13	87	60960	0.481	0.23385
2016	Indiana	77.22	8.1	23982	15912	0.26492	340500700000	4	14.1	84	52314	0.4527	0.26492
2016	Iowa	79.52	4.3	12975	9548	0.23125	181011400000	3.4	11.8	91	56247	0.4451	0.23125
2016	Kansas	78.51	8.7	5607	4123	0.23888	160451000000	4	12.1	87	54935	0.455	0.23888
2016	Kentucky	75.79	5.1	5129	2746	0.20978	196484900000	5.1	18.5	90	46659	0.4813	0.20978
2016	Louisiana	75.64	10.3	9348	8061	0.20805	227090700000	5.9	20.2	81	45146	0.499	0.20805
2016	Maine	78.99	8	9769	8313	0.22334	60254400000	3.5	12.5	87	53079	0.4519	0.22334
2016	Maryland	79.23	6.1	3049	2432	0.18756	387733400000	4.2	9.7	87	78945	0.4499	0.18756
2016	Massachusetts	80.37	2.5	11813	9849	0.19077	514637600000	3.9	10.4	88	75297	0.4786	0.19077
2016	Michigan	78	5.4	15342	15958	0.22569	490264000000	5	15	80	52492	0.4695	0.22569
2016	Minnesota	80.79	4.1	22080	15316	0.24394	344060800000	3.9	9.9	83	65599	0.4496	0.24394
2016	Mississippi	74.75	11.8	9628	10121	0.22404	107291400000	5.7	20.8	83	41754	0.4828	0.22404
2016	Missouri	77.44	8.9	6477	4831	0.23631	300914500000	4.3	14	88	51746	0.4646	0.23631
2016	Montana	78.86	8.1	12431	9500	0.2413	45490900000	4.2	13.3	86	50027	0.4667	0.2413
2016	Nebraska	79.44	8.6	1791	1439	0.25115	118145900000	3.1	11.4	89	56927	0.4477	0.25115
2016	Nevada	78.14	11.4	3316	1830	0.2099	151840400000	5.4	13.8	81	55180	0.4577	0.2099
2016	New Hampshire	79.92	5.9	5382	3084	0.21812	79090100000	2.9	7.3	89	70936	0.4304	0.21812
2016	New Jersey	80.16	8	2681	1812	0.19676	575501000000	4.7	10.4	91	76126	0.4813	0.19676

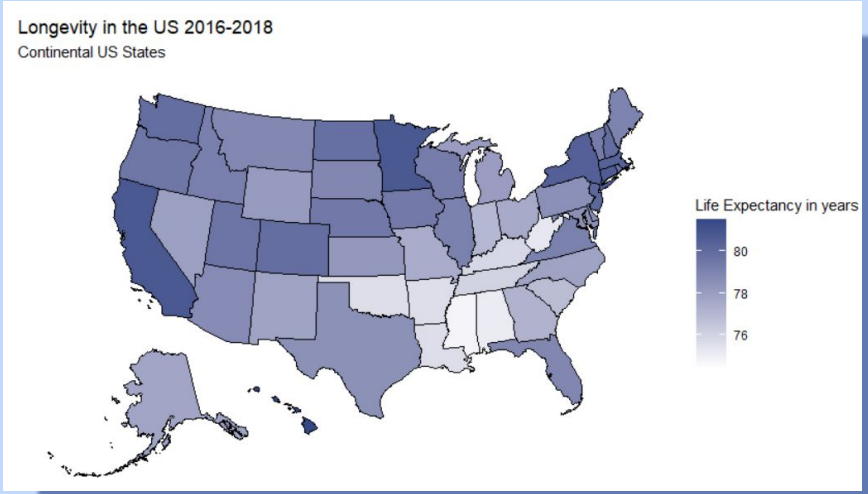
Validation Data (2019)

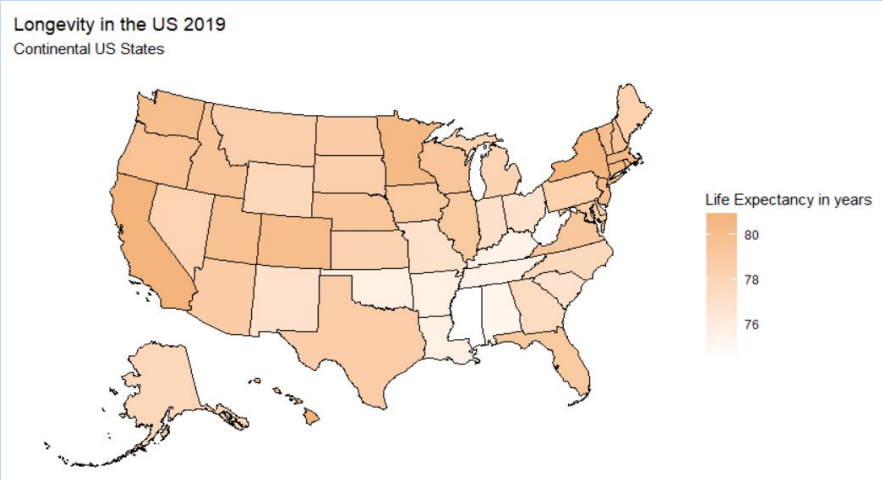
year	state	Longevity	uninsured	Medicare	Medicaid	cost pc	GDP	Unemployment	poverty	Grad Rate	Income	gini	burden
2019	Alabama	75.2	9.7	12497	5649	0.22049	231561900000	3.1	15.5	91	51734	0.4741	0.22049
2019	Alaska	77.7	12.2	1105	2127	0.1501	54728200000	5.3	10.1	79	75463	0.4376	0.1501
2019	Arizona	78.8	11.3	15231	11464	0.18642	372393500000	4.7	13.5	77	62055	0.4591	0.18642
2019	Arkansas	75.7	9.1	7200	6111	0.18512	131578300000	3.5	16.2	88	48952	0.475	0.18512
2019	California	80.9	7.7	84590	78339	0.14466	3042694100000	4.2	11.8	84	80440	0.4666	0.14466
2019	Colorado	80	8	9672	9255	0.17911	394534700000	2.8	9.3	82	77127	0.4548	0.17911
2019	Connecticut	80.3	5.9	8932	8172	0.1804	288536300000	3.7	10	90	78833	0.5024	0.1804
2019	Delaware	78.1	6.6	2622	1994	0.1884	76923700000	3.6	11.3	81	70176	0.4509	0.1884
2019	Florida	79	13.2	1249	2724	0.17001	1111392200000	3	12.7	90	59227	0.4808	0.17001
2019	Georgia	77.4	13.4	61804	19948	0.20114	637992500000	3.5	13.3	84	61980	0.4795	0.20114
2019	Hawaii	80.9	4.2	21050	9951	0.12202	91912700000	2.3	9.3	86	83102	0.4397	0.12202
2019	Idaho	79.5	10.8	2688	1903	0.243	82847200000	2.8	11.2	80	60999	0.4337	0.243
2019	Illinois	79	7.4	3319	2010	0.19183	889540100000	3.6	11.5	86	69187	0.48	0.19183
2019	Indiana	77	8.7	27700	17151	0.2154	381132300000	3.3	11.9	91	57603	0.4584	0.2154
2019	Iowa	79	5	15285	11586	0.18297	195708700000	2.7	11.2	90	61691	0.4422	0.18297
2019	Kansas	78.2	9.2	6712	4746	0.2004	176564400000	3.1	11.4	88	62087	0.45	0.2004
2019	Kentucky	75.5	6.4	6132	3139	0.18493	218494700000	4	16.3	91	52295	0.4764	0.18493
2019	Louisiana	75.7	8.9	10974	9144	0.18493	255481900000	4.8	19	82	51073	0.4978	0.18493
2019	Maine	78.3	8	11353	10435	0.22782	68690400000	3.1	10.9	86	58924	0.449	0.22782
2019	Maryland	78.5	6	3641	2879	0.17556	420371300000	3.2	9	87	86738	0.4558	0.17556
2019	Massachusetts	80.4	3	13685	10861	0.17904	589904300000	2.9	9.4	90	85843	0.4803	0.17904
2019	Michigan	78	5.8	17817	15953	0.17211	532217000000	3.7	13	81	59584	0.4634	0.17211
2019	Minnesota	80.4	4.9	24980	16530	0.20039	383956600000	3.4	9	83	74593	0.4434	0.20039
2019	Mississippi	74.4	13	11368	11527	0.21287	114234400000	5.5	19.6	88	45792	0.4896	0.21287
2019	Missouri	76.9	10	7628	4836	0.21698	332485800000	3.4	12.9	90	57409	0.4633	0.21698
2019	Montana	78.4	8.3	14644	10074	0.22836	51925400000	3.5	12.6	86	57153	0.4597	0.22836
2019	Nebraska	79.2	8.3	2253	1849	0.22819	131866900000	3.1	9.9	88	63229	0.44	0.22819
2019	Nevada	78	11.4	3978	1966	0.19154	182186300000	4.2	12.5	81	63276	0.471	0.19154
2019	New Hampshire	79.4	6.3	6700	3821	0.22691	87338200000	2.6	7.3	87	77933	0.4406	0.22691
2019	New Jersey	80.1	7.9	3211	1843	0.17838	637610400000	3.9	9.2	89	85751	0.4782	0.17838

Exploratory Data Analysis (EDA)

Heat Maps

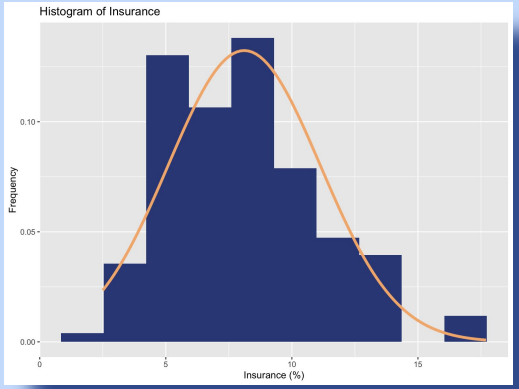
- Highest avg. life expectancy (2016-2018): California, Minnesota, and New York
- Lowest avg. life expectancy (2016-2018): Mississippi, West Virginia, and Alabama



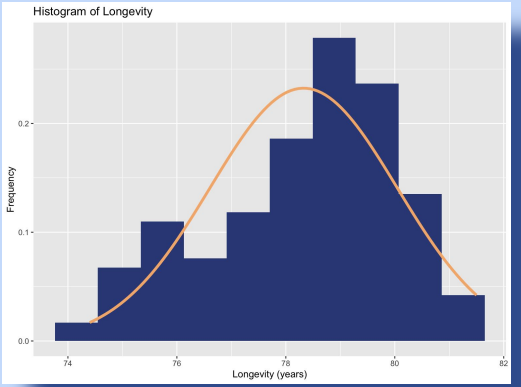


- Highest avg. life expectancy (2019): California, Minnesota, and New York
- Lowest avg. life expectancy (2019): Mississippi, West Virginia, and Alabama

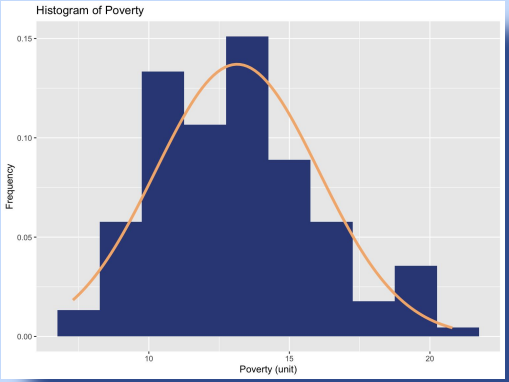
EDA: Distribution of predictors



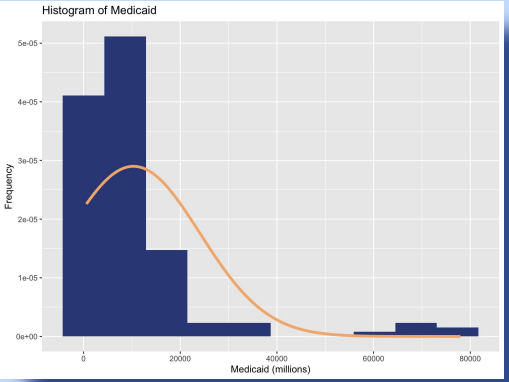
Normal



Normal



Normal



NOT Normal + Outliers

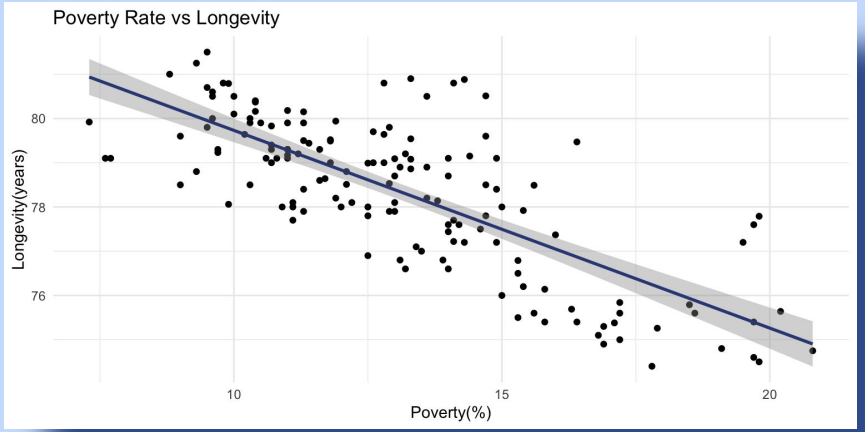
Scatter Plots

Poverty Rate vs Longevity

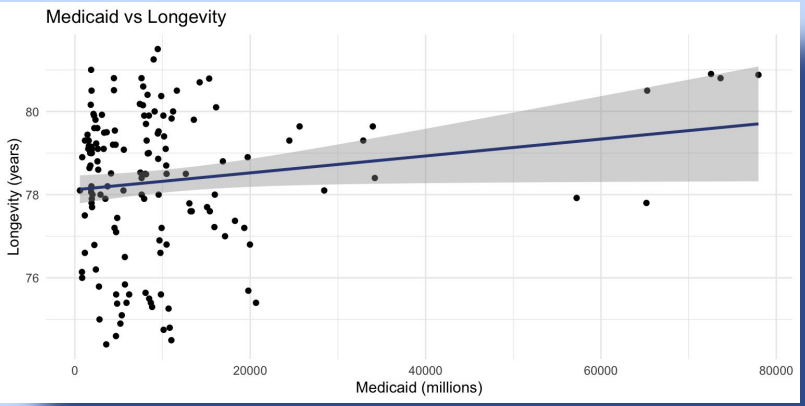
R-squared: 0.574

P-value: < 2e-16

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  84.1989     0.4254   197.9   <2e-16 ***
poverty      -0.4469     0.0316   -14.1   <2e-16 ***
```




```
model = lm(Longevity~Medicaid, data=data)
summary(model)
```



Medicaid vs Longevity

R-squared: 0.0265

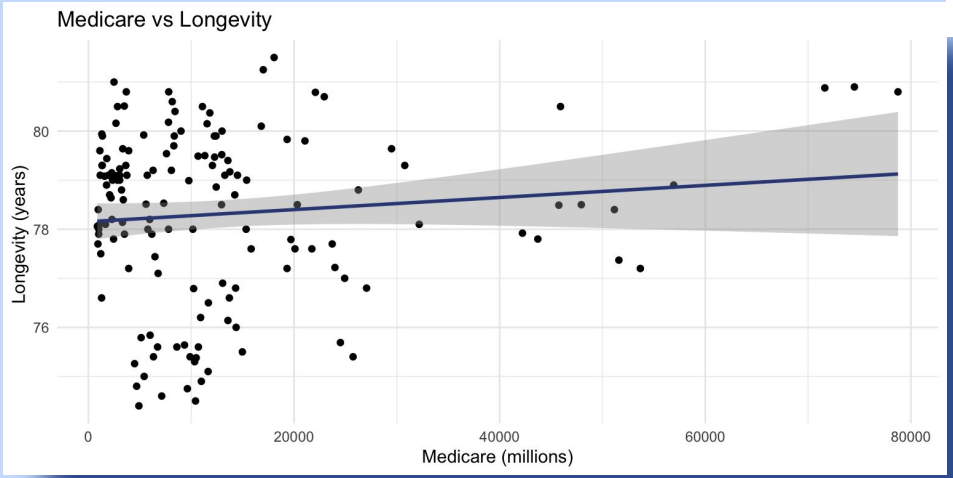
P-value: 0.047

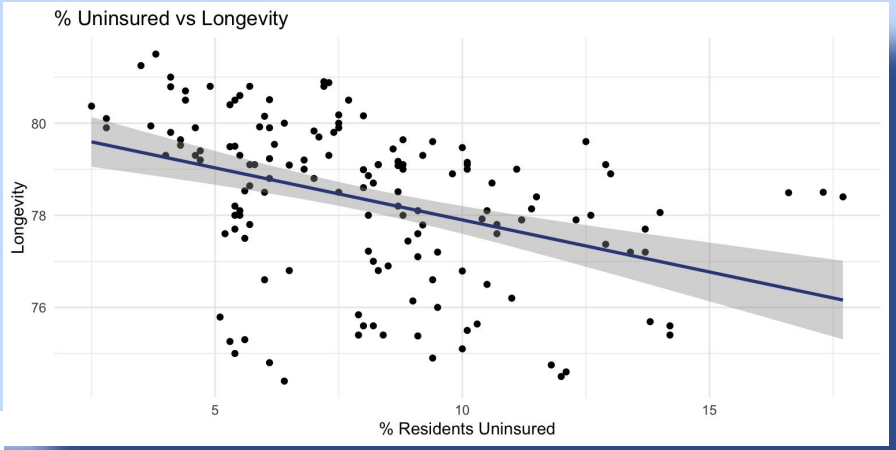
Scatter Plots

Medicare vs Longevity

R-squared: 0.0102

P-value: 0.22





% Uninsured vs Longevity

R-squared: 0.157

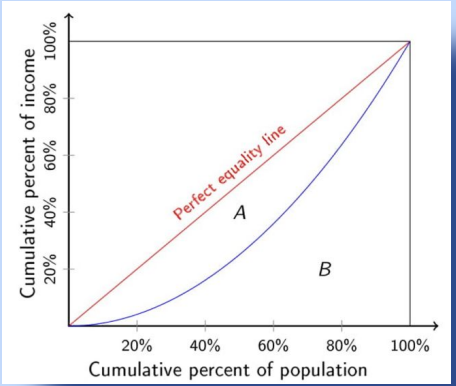
P-value: 5e-07

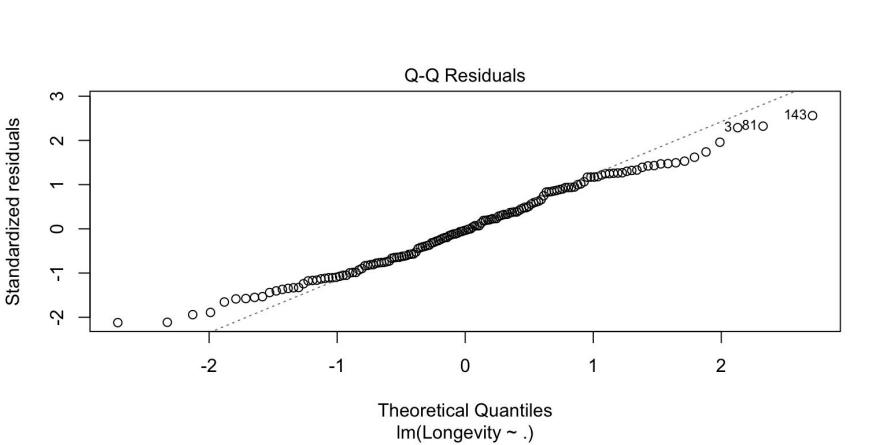
Models

Regression

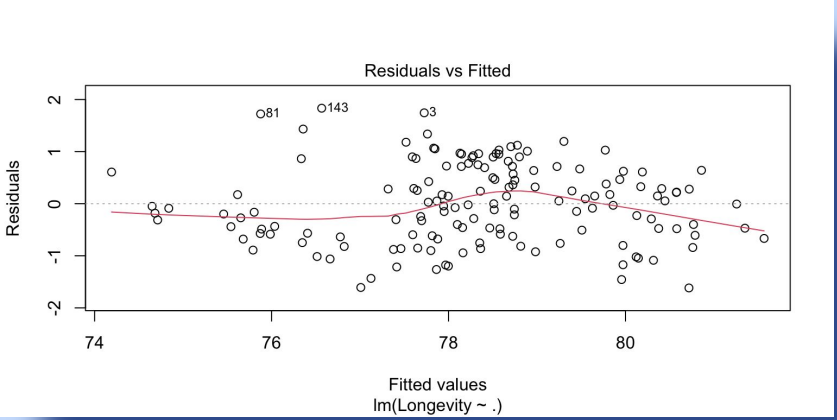
```
   Min     1Q Median     3Q    Max
-1.617 -0.586 -0.026  0.620  1.834

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.62e+03   3.47e+02    4.66  7.3e-06 ***
year        -7.59e-01   1.72e-01   -4.41  2.1e-05 ***
uninsured   -5.19e-02   2.84e-02   -1.83  0.06971 .
Medicare     2.95e-08   8.78e-06    0.00  0.99733
Medicaid     1.10e-05   8.63e-06    1.27  0.20514
cost.pc      8.38e-05   2.94e-05    2.85  0.00502 **
GDP          7.68e-13   2.04e-13    3.76  0.00025 ***
Unemployment -3.98e-01   1.05e-01   -3.78  0.00023 ***
poverty      -4.16e-01   7.05e-02   -5.90  2.7e-08 ***
Grad.Rate    -1.26e-01   1.93e-02   -6.54  1.1e-09 ***
income        1.97e-05   1.75e-05    1.13  0.26057
gini          1.64e+01   5.38e+00    3.04  0.00286 **
burden       -4.00e+00   2.93e+00   -1.36  0.17513
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.785 on 137 degrees of freedom
Multiple R-squared:  0.808,     Adjusted R-squared:  0.791
F-statistic: 47.9 on 12 and 137 DF,  p-value: <2e-16
```

```
fit1 <- lm(Longevity~., data1)
summary(fit1)
```





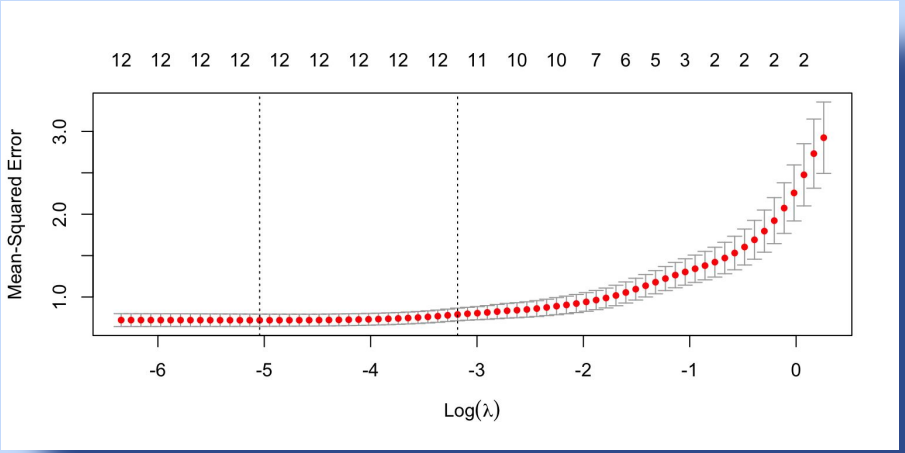
Model Diagnostics



Normality Assumption: Satisfied by Q-Q Plot

Homoscedasticity Assumption: Satisfied by Residual Plot

Linearity Assumption: Satisfied by Residual Plot




```
fit.fl.lambda <- cv.glmnet(as.matrix(X), Y, alpha=1)
names(fit.fl.lambda)
coef.min <- coef(fit.fl.lambda, s="lambda.min")
coef.min <- coef.min[which(coef.min !=0),]
coef.min <- extract.coef(fit.fl.lambda, s="lambda.min")


plot(fit.fl.lambda)

var.min <- rownames(coef.min)[-1]
var.min
```

LASSO

By the graph, minimized MSE: $\ln(\lambda) \approx -5 \rightarrow \lambda \approx 0.006$

```
var.min <- rownames(coef.min)[-1]
```

```
[1] "year"         "uninsured"    "Medicare"     "Medicaid"     "cost.pc"      "GDP"
[7] "Unemployment" "poverty"      "Grad.Rate"    "income"       "gini"         "burden"
```

Backward Elimination: Ist Iteration

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.62e+03   3.47e+02    4.66  7.3e-06 ***
year        -7.59e-01   1.72e-01   -4.41  2.1e-05 ***
uninsured   -5.19e-02   2.84e-02   -1.83  0.06971 .
Medicare     2.95e-08   8.78e-06    0.00  0.99733
Medicaid     1.10e-05   8.63e-06    1.27  0.20514
cost.pc      8.38e-05   2.94e-05    2.85  0.00502 **
GDP          7.68e-13   2.04e-13    3.76  0.00025 ***
Unemployment -3.98e-01   1.05e-01   -3.78  0.00023 ***
poverty      -4.16e-01   7.05e-02   -5.90  2.7e-08 ***
Grad.Rate    -1.26e-01   1.93e-02   -6.54  1.1e-09 ***
income       1.97e-05   1.75e-05    1.13  0.26057
gini         1.64e+01   5.38e+00    3.04  0.00286 **
burden       -4.00e+00   2.93e+00   -1.36  0.17513
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We test at 0.05 significance level.

Highest P-value (‘Medicare’): 0.99733

We eliminate ‘Medicare’ and re-run.

Backward Elimination: 2nd Iteration

```
Anova Table (Type II tests)

Response: Longevity
             Sum Sq  Df F value  Pr(>F)
year           12.1   1   19.82 1.7e-05 ***
uninsured       2.5   1    4.06 0.04593 *
Medicaid        2.8   1    4.56 0.03442 *
cost.pc         5.0   1    8.24 0.00474 **
GDP             8.9   1   14.50 0.00021 ***
Unemployment    9.1   1   14.84 0.00018 ***
poverty        21.5   1   35.13 2.3e-08 ***
Grad.Rate      26.6   1   43.49 8.4e-10 ***
income          0.8   1    1.29 0.25862
gini            6.0   1    9.80 0.00213 **
burden          1.2   1    1.97 0.16246
Residuals      84.4 138
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Highest P-value ('income'): 0.25862

We eliminate 'income' and re-run.

Backward Elimination: 3rd Iteration

```
Anova Table (Type II tests)

Response: Longevity
             Sum Sq  Df F value  Pr(>F)
year           11.9   1   19.44 2.1e-05 ***
uninsured       2.4   1    3.88 0.05089 .
Medicaid        3.3   1    5.31 0.02273 *
cost.pc         5.3   1    8.66 0.00381 **
GDP             9.4   1   15.36 0.00014 ***
Unemployment    8.3   1   13.57 0.00033 ***
poverty       130.1   1  212.26 < 2e-16 ***
Grad.Rate      27.3   1   44.59 5.4e-10 ***
gini            8.4   1   13.67 0.00031 ***
burden          2.4   1    3.92 0.04975 *
Residuals      85.2 139
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Highest P-value (‘uninsured’): 0.05089

We eliminate ‘uninsured’ and re-run.

Final Regression Model

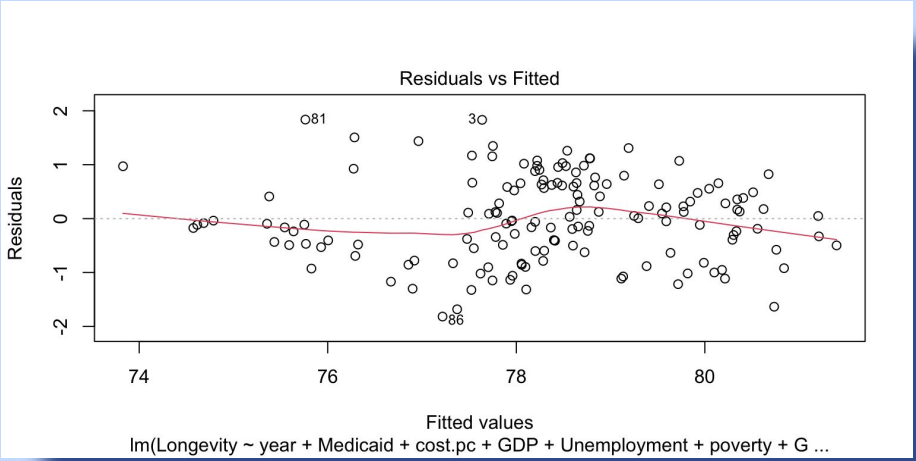
```
Anova Table (Type II tests)

Response: Longevity
             Sum Sq  Df F value  Pr(>F)
year           15.9   1   25.46 1.4e-06 ***
Medicaid        3.2   1    5.08 0.02577 *
cost.pc         6.2   1    9.90 0.00202 **
GDP             8.1   1   12.97 0.00044 ***
Unemployment   10.4   1   16.64 7.6e-05 ***
poverty       150.0   1  239.77 < 2e-16 ***
Grad.Rate      25.4   1   40.63 2.5e-09 ***
gini           10.5   1   16.76 7.1e-05 ***
burden          4.7   1    7.46 0.00711 **
Residuals      87.6 140
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

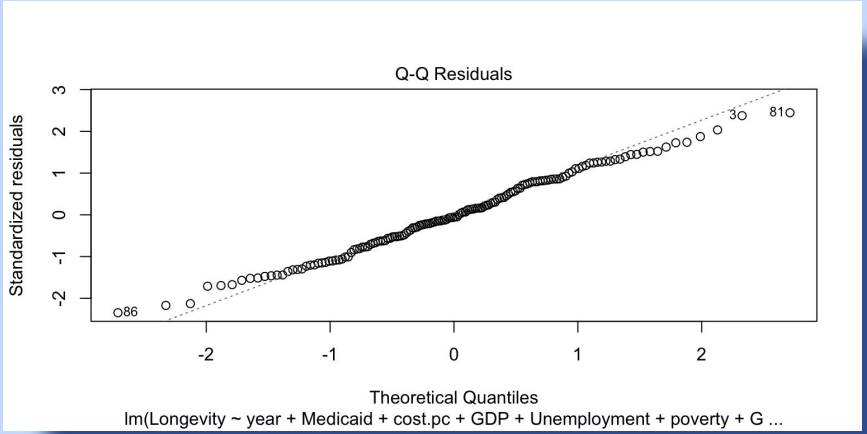
Following variables’ P-values < 0.05:

{ ‘Medicare’, ‘income’, ‘uninsured’}

Thus, we’ve eliminated all necessary variables and have our final factors.

Model Diagnostics





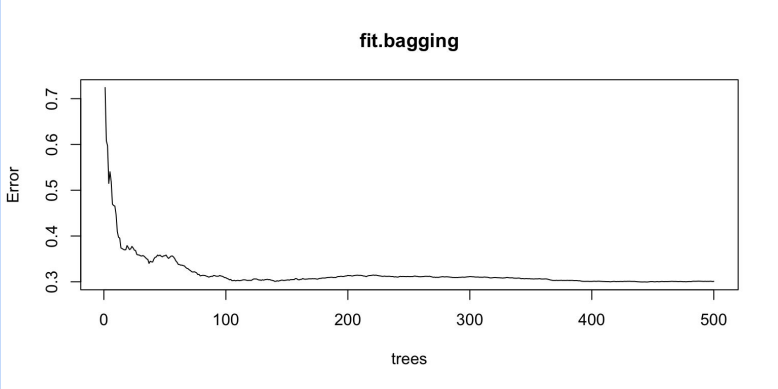
Normality Assumption: Satisfied by Q-Q Plot

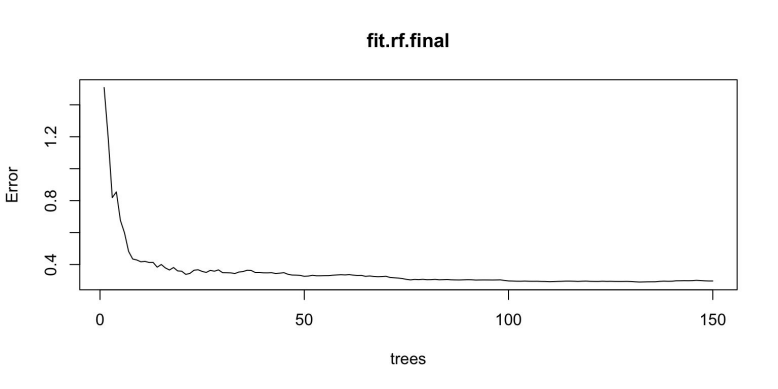
Homoscedasticity Assumption: Satisfied by Residual Plot

Linearity Assumption: Satisfied by Residual

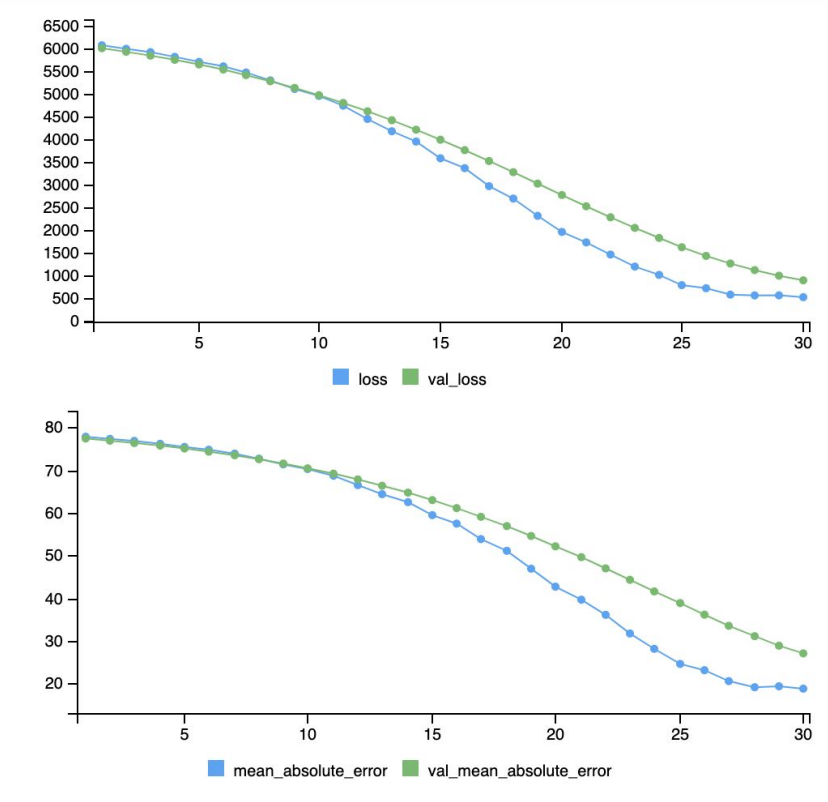
Random Forest

```
fit.bagging <- randomForest(Longevity~., data1, mtry=12, ntree=500)
```





Neural Network




```r
model <- keras_model_sequential() %>%
  layer_dense(units = 128, activation = 'relu', input_shape = ncol(X_train)) %>%
  layer_dropout(rate = 0.4) %>%
  layer_dense(units = 64, activation = 'relu') %>%
  layer_dropout(rate = 0.3) %>%
  layer_dense(units = 1)

# Compile the model
model %>% compile(
  optimizer = 'adam',
  loss = 'mean_squared_error',
  metrics = c('mean_absolute_error')
)

history <- model %>% fit(
  X_train, Y_train,
  epochs = 30,
  batch_size = 32,
  validation_split = 0.2
)
```

Model Comparison

fit1.pred.error <dbl>	final.lm.pred.error <dbl>	rf.pred.error <dbl>	pred_nn.error <dbl>
1.34	1.71	0.624	588

Initial Multiple Regression Model

Multiple Regression Model after Regularization

Random Forest Model

Neural Network Error

Using Mean Squared Error

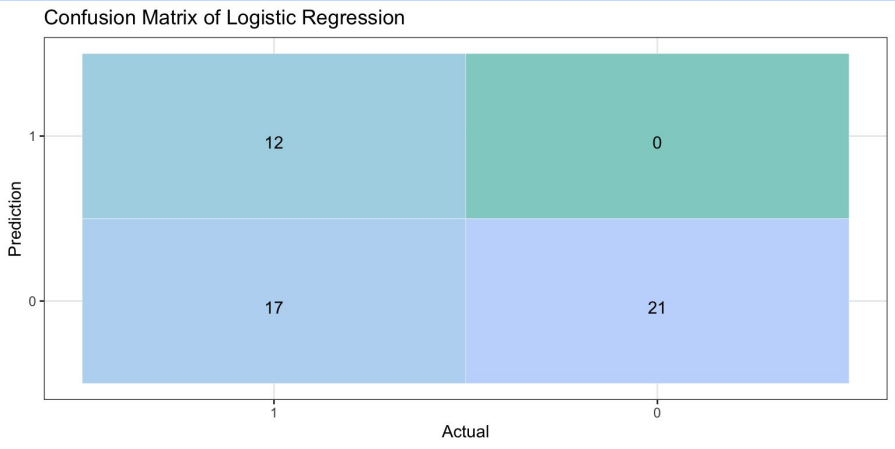
```
fit1.pred.error <- sum((fit1.pred-valid.2019$Longevity)^2)/50
final.lm.pred.error <- sum((final.lm.pred-valid.2019$Longevity)^2)/50
rf.pred.error <- sum((rf.pred-valid.2019$Longevity)^2)/50
pred_nn.error <- sum((pred_nn - Y_val)^2)/50

data.frame(fit1.pred.error, final.lm.pred.error, rf.pred.error, pred_nn.error)
```

Longevity as a binary outcome

Logistic Regression

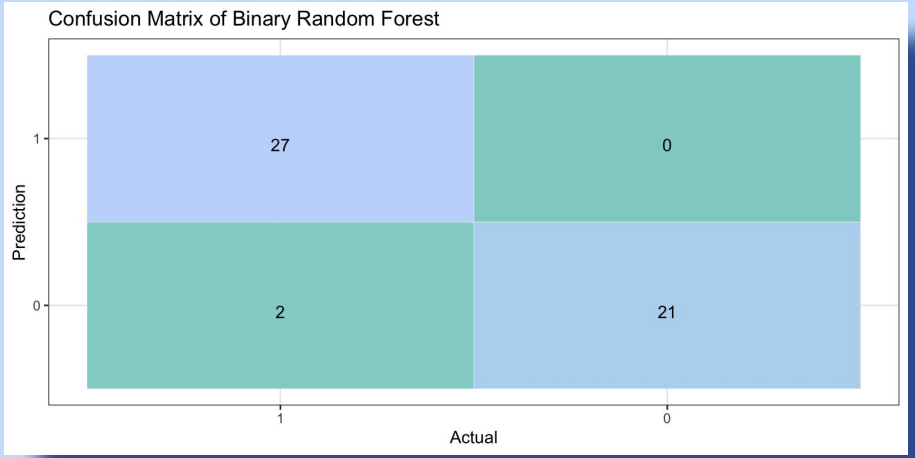
```
data1$Longevity_binary <- ifelse(data1$Longevity > mean(data1$Longevity), 1, 0)
```

```
fit.logistic <- glm(Longevity_binary~., data2, family=binomial)
```

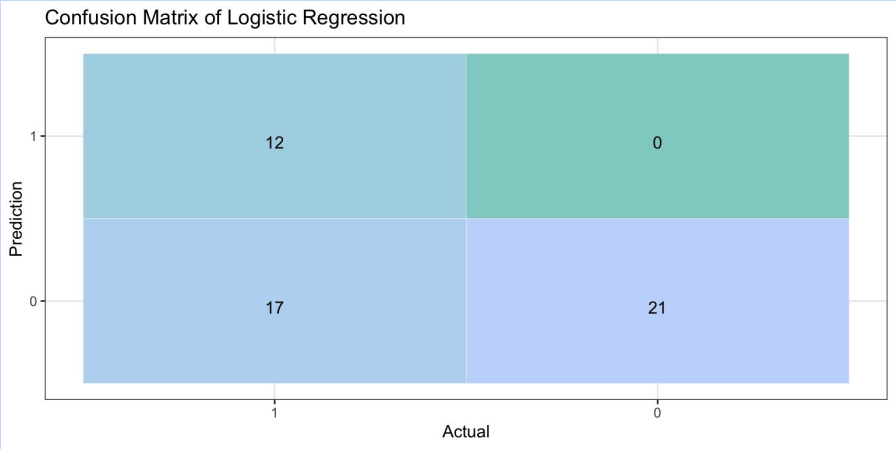


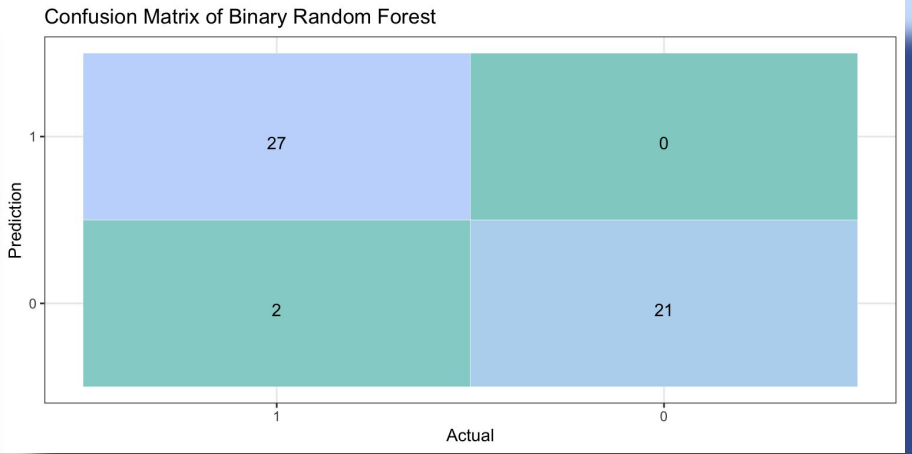
Binary Random Forest

```
fit.bagging.2 <- randomForest(data1$Longevity_binary~., data1, mtry=12, ntree=500)
```



Comparing Logistic with Random Forest





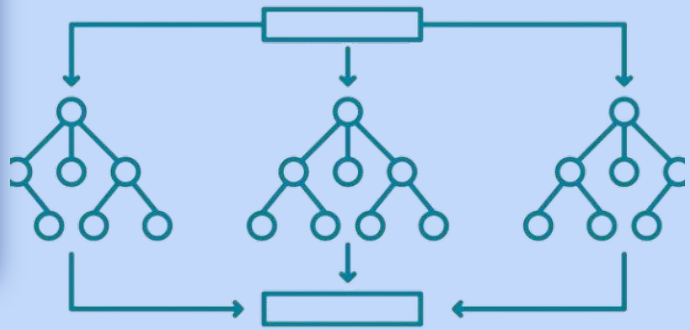
Precision: 1
Recall: 0.414
f1: 0.585

Precision: 1
Recall: 0.931
f1: 0.964

Conclusion

```
Call:
lm(formula = Longevity ~ year + Medicaid + cost.pc + GDP + Unemployment +
    poverty + Grad.Rate + gini + burden, data = data.fl.sub)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8177 -0.5225 -0.0428  0.6137  1.8372

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.78e+03   3.36e+02    5.31  4.3e-07 ***
year         -8.39e-01   1.66e-01   -5.05  1.4e-06 ***
Medicaid      1.16e-05   5.16e-06    2.25  0.02577 *
cost.pc       9.22e-05   2.93e-05    3.15  0.00202 **
GDP           7.21e-13   2.00e-13    3.60  0.00044 ***
Unemployment -3.92e-01   9.62e-02   -4.08  7.6e-05 ***
poverty      -5.03e-01   3.25e-02  -15.48  < 2e-16 ***
Grad.Rate    -1.21e-01   1.90e-02   -6.37  2.5e-09 ***
gini          2.01e+01   4.91e+00    4.09  7.1e-05 ***
burden       -6.90e+00   2.52e+00   -2.73  0.00711 **
```

Random Forest



Significance of Model

Identifies significant factors of healthcare quality which allows:

- Effective resource allocation
- Optimizes healthcare services
- Evidence-based and informed policymaking

Overall improves information on healthcare quality and life expectancy

Special thanks to

Professor Zhao, Tim, Neil, Jeff