BIOST 544A Course Project: WHO Life Expectancy

My-Anh Doan and Katie Denecke

2022-12-13

1. Introduction

This publicly available dataset originates from the Global Health Observatory (GHO) data repository under the World Health Organization (WHO). The dataset contains data on life expectancy, health factors, and economic data (collected by the United Nation website) from 2000 to 2015 for 193 countries. There is a total of 22 attributes in this dataset that can be divided into broad categories related to immunization, mortality, economical, and social factors. The dataset aims to determine significant factors that affect life expectancy.

1.1 Variable description

For this course project, we will only look at five predictor variables in relation to life expectancy: alcohol consumption, HepB immunization, polio immunization, diphtheria immunization, and HIV/AIDS death.

Factor	Variable	Type	Description
Life expectancy	Response	Quantitative	Life expectancy in years
Alcohol	Predictor	Quantitative	Alcohol consumption (liters of pure alcohol; per capita)
Hepatitis B	Predictor	Quantitative	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)

Factor	Variable	Type	Description
Polio	Predictor	Quantitative	Polio immunization coverage among 1-year-olds (%)
Diphtheria	Predictor	Quantitative	TDAP immunization coverage among 1-year-olds (%)
HIV/AIDS	Predictor	Quantitative	Deaths per 1000 live births (0-4 years) caused by HIV/AIDS

1.2 Analysis questions and aims

The dataset aims to answer the following questions:

Among the five predictors chosen above, which ones are actually significant to life expectancy?

1.3 General analysis approach

We will use regression tools and techniques to determine the significant predictors of life expectancy.

2. Data preparation

```
library(tidyverse)
library(janitor) # Data cleaning package

# Load data
data <- read.csv("./data/Life Expectancy Data.csv")

data <- data %>%
  clean_names() %>%
  select(life_expectancy, alcohol, hepatitis_b, polio, diphtheria, hiv_aids) %>%
  mutate_all(as.numeric)

str(data)
```

```
## 'data.frame':    2938 obs. of  6 variables:
##  $ life_expectancy: num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ alcohol        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ hepatitis_b    : num  65 62 64 67 68 66 63 64 63 64 ...
##  $ polio          : num  6 58 62 67 68 66 63 64 63 58 ...
##  $ diphtheria     : num  65 62 64 67 68 66 63 64 63 58 ...
##  $ hiv_aids       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
```

3. Initial exploration

3.1 Handling missing data in predictor variables

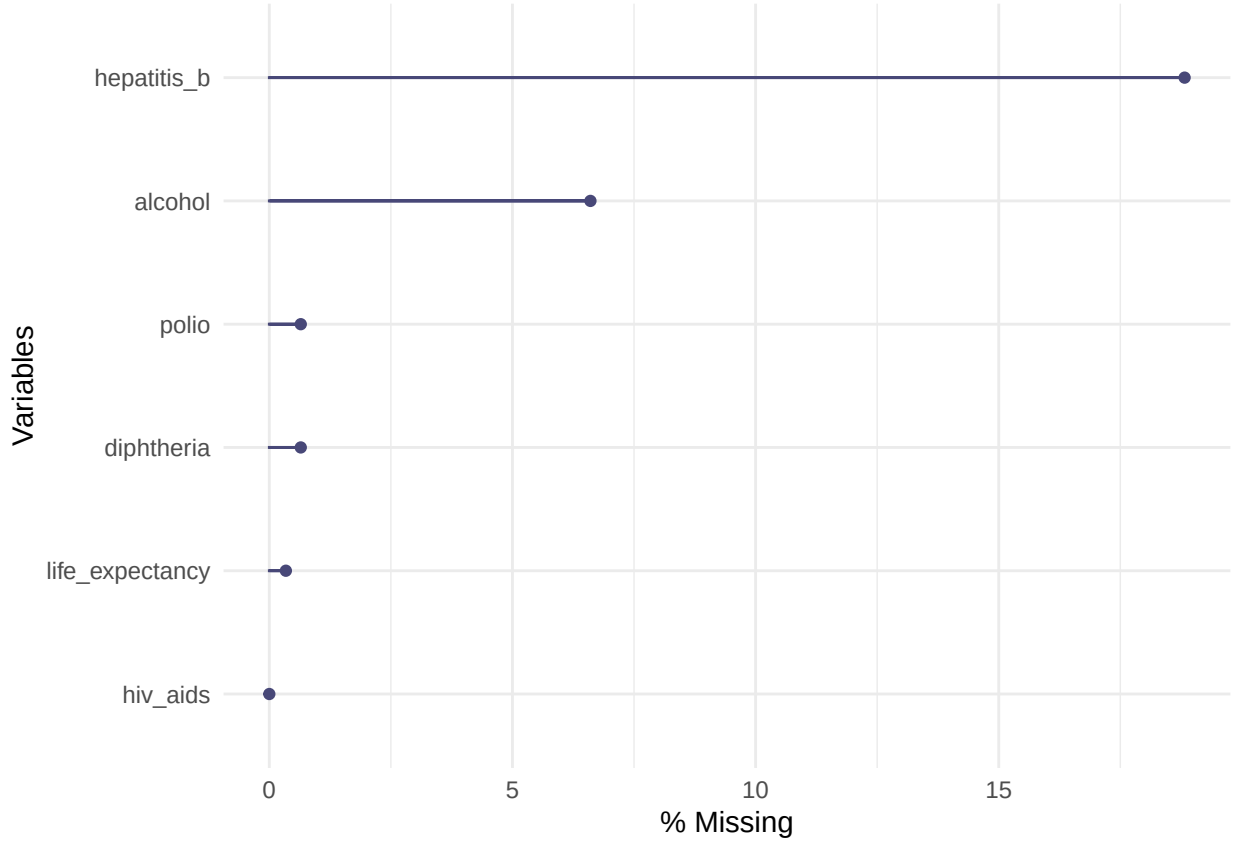
```r
library(naniar)
library(missForest)

# Check for any observations/rows with missing data
sum(complete.cases(data) == FALSE) / nrow(data)    # proportion of missing data
```

```
## [1] 0.2494894
```

```r
missing_vals <- which(complete.cases(data) == FALSE)    # rows with missing data

# Visualization of missing values
gg_miss_var(data, show_pct = TRUE)
```




```r
# Impute missing values in covariates
set.seed(1001)
data_imputed_covars <- missForest(xmis = subset(data, select = -c(life_expectancy)),
                                  maxiter = 50)$ximp

data_imputed <- cbind(data$life_expectancy, data_imputed_covars)
colnames(data_imputed)[1] = "life_expectancy"
```

3.2 Train-test split dataset

```
set.seed(1)
rand_sample <- sample(1:nrow(data_imputed),
                      size = 0.7 * nrow(data_imputed),
                      replace = FALSE)

training_data <- data_imputed[rand_sample, ]
test_data <- data_imputed[-rand_sample, ]
```

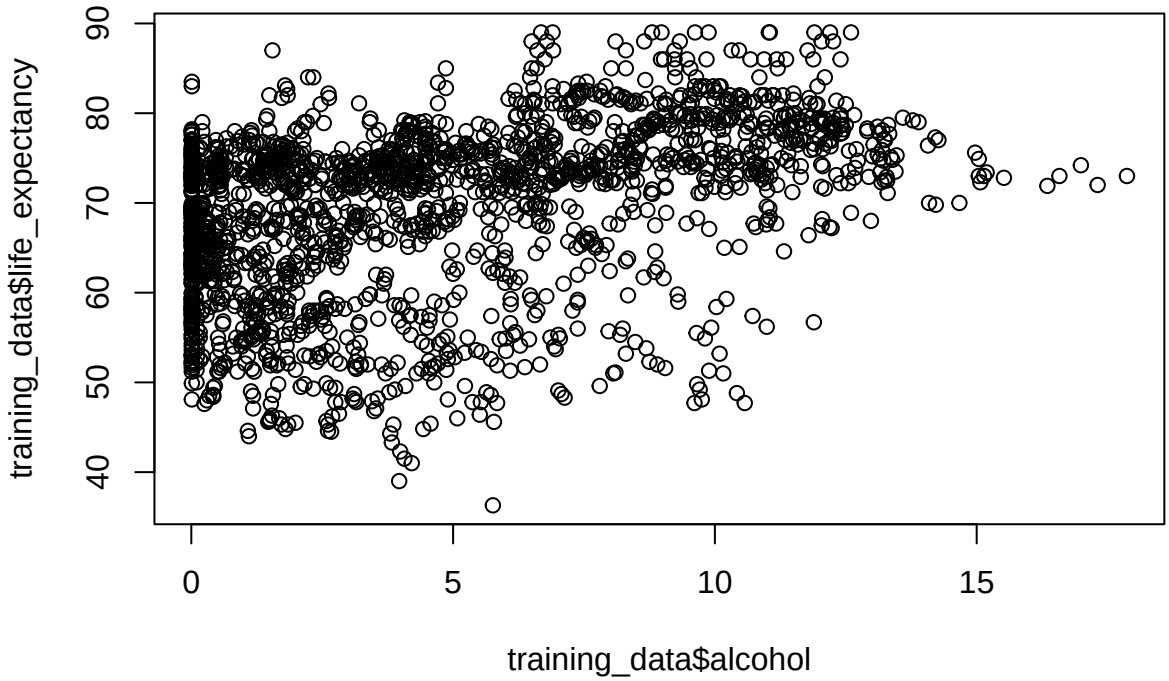
4. Main data analysis

4.1 Simple linear regression models

First, we look at the individual predictors and their effect on life expectancy.

4.1.1 Alcohol

```
plot(x = training_data$alcohol, y = training_data$life_expectancy)
```



```r
# Fit model to training data
alc_model_lrm <- glm(as.factor(life_expectancy) ~ alcohol, family = "binomial",
                     data=training_data)
summary(alc_model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ alcohol, family = "binomial",
##     data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8950   0.0273   0.0299   0.0339   0.0477
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.96777    1.68290   4.735  2.2e-06 ***
## alcohol     -0.06646    0.23500  -0.283    0.777
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17.248  on 2046  degrees of freedom
## Residual deviance: 17.170  on 2045  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 21.17
##
## Number of Fisher Scoring iterations: 10
```

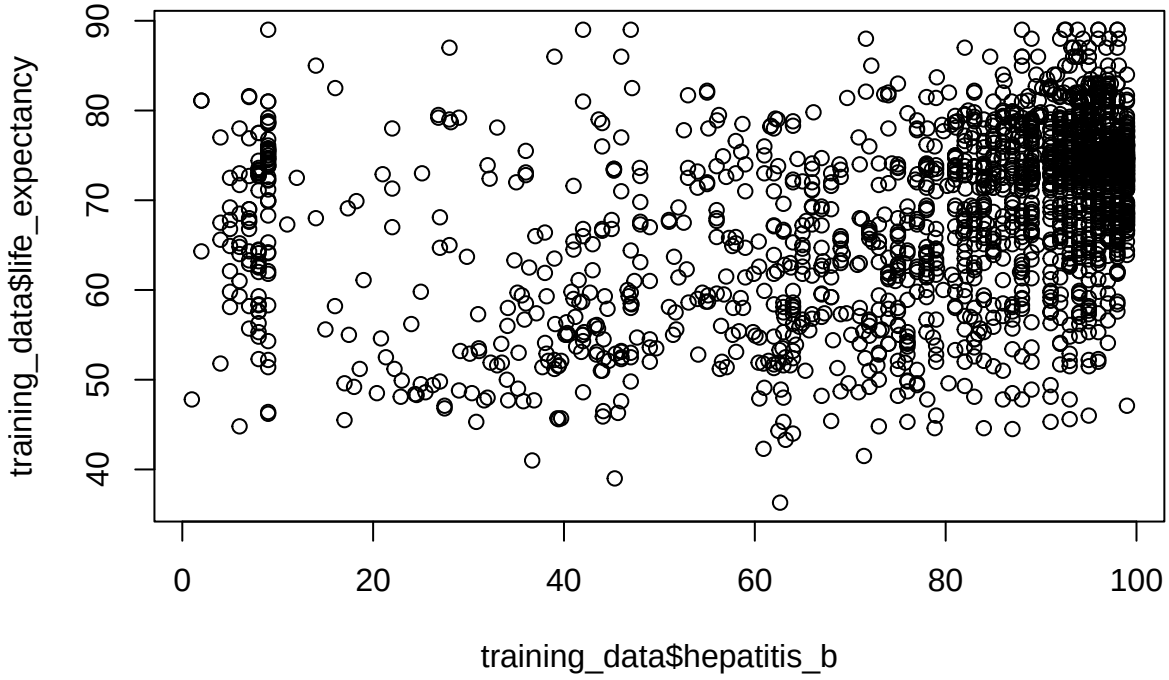
```r
# Use fitted model on test data
model_lrm <- glm(as.factor(life_expectancy) ~ alcohol,  family = "binomial",
                 data=test_data)
summary(model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ alcohol, family = "binomial",
##     data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5421   0.0166   0.0363   0.0646   0.0790
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.7651     1.2269   4.699 2.61e-06 ***
## alcohol       0.4291     0.6003   0.715    0.475
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 15.561  on 880  degrees of freedom
## Residual deviance: 14.545  on 879  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 18.545
##
## Number of Fisher Scoring iterations: 11
```

4.1.2 HepB Immunization

```
plot(x = training_data$hepatitis_b, y = training_data$life_expectancy)
```




```
# Fit model to training data
hepB_model_lrm <- glm(as.factor(life_expectancy) ~ hepatitis_b, family = "binomial",
                data=training_data)
summary(hepB_model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ hepatitis_b, family = "binomial",
##     data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

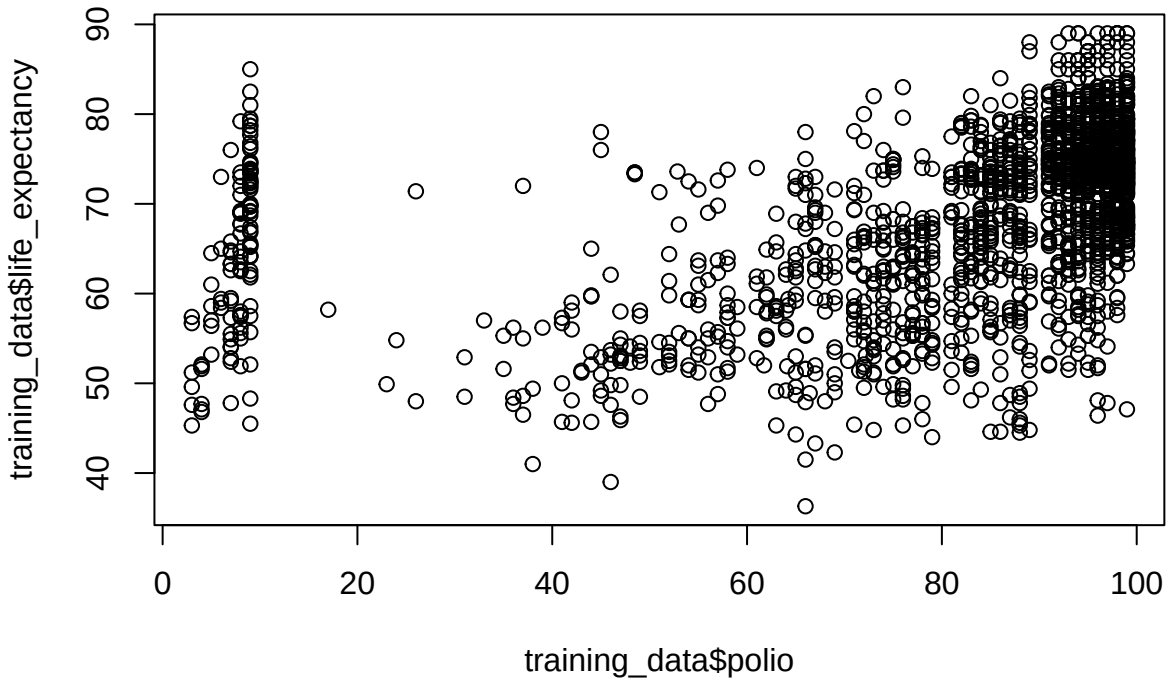
```
## -3.8659   0.0251   0.0267   0.0313   0.0582
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.36352    2.12546   2.994  0.00275 **
## hepatitis_b  0.01769    0.02993   0.591  0.55432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17.248  on 2046  degrees of freedom
## Residual deviance: 16.945  on 2045  degrees of freedom
##   (9 observations deleted due to missingness)
## AIC: 20.945
##
## Number of Fisher Scoring iterations: 10
```

```
# Use fitted model on test data
model_lrm <- glm(as.factor(life_expectancy) ~ hepatitis_b,  family = "binomial",
                 data=test_data)
summary(model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ hepatitis_b, family = "binomial",
##     data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6764   0.0440   0.0452   0.0492   0.0633
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 6.195392   2.708473   2.287   0.0222 *
## hepatitis_b 0.007764   0.034792   0.223   0.8234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.561  on 880  degrees of freedom
## Residual deviance: 15.515  on 879  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 19.515
##
## Number of Fisher Scoring iterations: 9
```

4.1.3 Polio Immunization

```
plot(x = training_data$polio, y = training_data$life_expectancy)
```




```
# Fit model to training data
polio_model_lrm <- glm(as.factor(life_expectancy) ~ polio, family = "binomial",
                 data=training_data)
summary(polio_model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ polio, family = "binomial",
##     data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8591   0.0254   0.0264   0.0305   0.0624
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.18327    2.15280   2.872  0.00408 **
## polio        0.01913    0.02887   0.662  0.50769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17.248  on 2046  degrees of freedom
## Residual deviance: 16.892  on 2045  degrees of freedom
##   (9 observations deleted due to missingness)
```

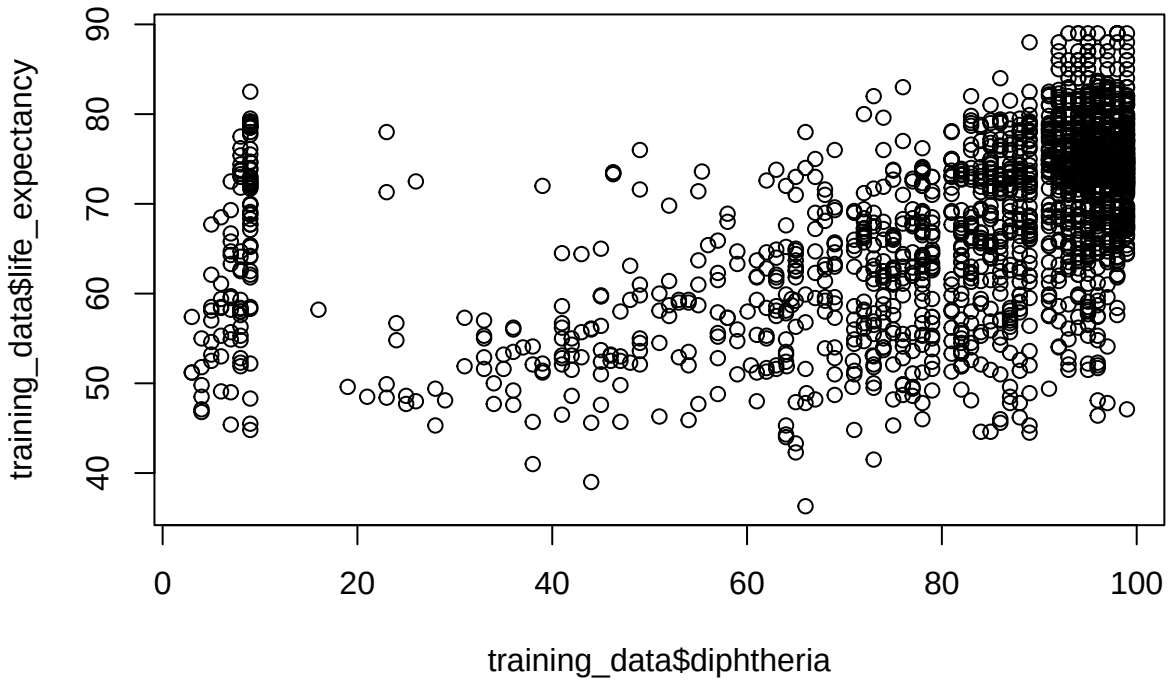
```
## AIC: 20.892
##
## Number of Fisher Scoring iterations: 10
```

```
# Use fitted model on test data
model_lrm <- glm(as.factor(life_expectancy) ~ polio,  family = "binomial",
                 data=test_data)
summary(model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ polio, family = "binomial",
##     data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6653   0.0423   0.0434   0.0483   0.0766
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.79296    2.58149   2.244   0.0248 *
## polio        0.01264    0.03259   0.388   0.6980
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.561  on 880  degrees of freedom
## Residual deviance: 15.433  on 879  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 19.433
##
## Number of Fisher Scoring iterations: 9
```

4.1.4 Diphtheria Immunization

```
plot(x = training_data$diphtheria, y = training_data$life_expectancy)
```




```
# Fit model to training data
diph_model_lrm <- glm(as.factor(life_expectancy) ~ diphtheria, family = "binomial",
                data=training_data)
summary(diph_model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ diphtheria, family = "binomial",
##     data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8609   0.0255   0.0264   0.0304   0.0614
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  6.21596    2.15852   2.880  0.00398 **
## diphtheria   0.01874    0.02897   0.647  0.51778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17.248  on 2046  degrees of freedom
## Residual deviance: 16.906  on 2045  degrees of freedom
##   (9 observations deleted due to missingness)
```

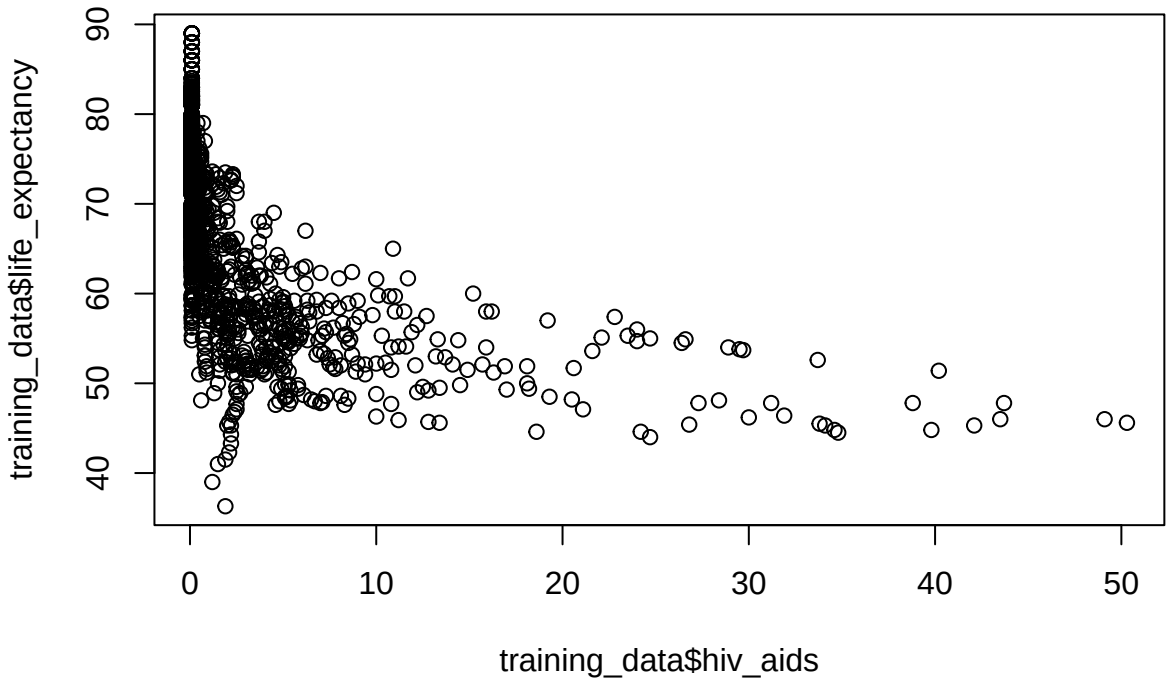
```
## AIC: 20.906
##
## Number of Fisher Scoring iterations: 10
```

```
# Use fitted model on test data
model_lrm <- glm(as.factor(life_expectancy) ~ diphtheria,  family = "binomial",
                 data=test_data)
summary(model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ diphtheria, family = "binomial",
##     data = test_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.6724   0.0436   0.0444   0.0481   0.0697
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 5.998704   2.720704   2.205   0.0275 *
## diphtheria  0.009913   0.033727   0.294   0.7688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.561  on 880  degrees of freedom
## Residual deviance: 15.486  on 879  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 19.486
##
## Number of Fisher Scoring iterations: 9
```

4.1.5 HIV/AIDS Deaths

```
plot(x = training_data$hiv_aids, y = training_data$life_expectancy)
```




```
# Fit model to training data
hiv_model_lrm <- glm(as.factor(life_expectancy) ~ hiv_aids, family = "binomial",
                 data=training_data)
summary(hiv_model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ hiv_aids, family = "binomial",
##     data = training_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.9046   0.0310   0.0310   0.0311   0.0385
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.639063   1.061027   7.200 6.04e-13 ***
## hiv_aids    -0.008642   0.186330  -0.046    0.963
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17.248  on 2046  degrees of freedom
## Residual deviance: 17.246  on 2045  degrees of freedom
##   (9 observations deleted due to missingness)
```

```
## AIC: 21.246
##
## Number of Fisher Scoring iterations: 10
```

```
# Use fitted model on test data
model_lrm <- glm(as.factor(life_expectancy) ~ hiv_aids,  family = "binomial",
                 data=test_data)
summary(model_lrm)
```

```
##
## Call:
## glm(formula = as.factor(life_expectancy) ~ hiv_aids, family = "binomial",
##     data = test_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.15096   0.02893   0.02893   0.03025   0.46512
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  7.78950    1.56389   4.981 6.33e-07 ***
## hiv_aids    -0.11107    0.04808  -2.310   0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.561  on 880  degrees of freedom
## Residual deviance: 11.952  on 879  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: 15.952
##
## Number of Fisher Scoring iterations: 10
```

4.2 Multiple linear regression models

4.2.1 Model fitting

```
library(car)

# Simple linear regression model using all variables, fitted on training data
lm_all <- lm(life_expectancy ~ ., data = training_data)
summary(lm_all)
```

```
##
## Call:
## lm(formula = life_expectancy ~ ., data = training_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.0836  -3.9441   0.2917   3.9595  20.9967
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.868545   0.582292  94.229  < 2e-16 ***
## alcohol      0.696985   0.035887  19.422  < 2e-16 ***
## hepatitis_b  0.018033   0.007490   2.408   0.0161 *
## polio        0.065527   0.008181   8.010 1.91e-15 ***
## diphtheria   0.072948   0.008880   8.214 3.74e-16 ***
## hiv_aids    -0.930850   0.029589 -31.460  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.297 on 2041 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.5587, Adjusted R-squared:  0.5576
## F-statistic: 516.8 on 5 and 2041 DF,  p-value: < 2.2e-16
```

```
vif(lm_all) # no multicollinearity in our data
```

```
##     alcohol hepatitis_b       polio  diphtheria    hiv_aids
##    1.066125    1.871195    1.897303    2.281632    1.036868
```

```
# Perform stepwise forward selection using AIC as selection criteria on training data
lm_intercept_only <- lm(life_expectancy ~ 1 , data = training_data)
summary(lm_intercept_only)
```

```
##
## Call:
## lm(formula = life_expectancy ~ 1, data = training_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -33.041  -5.891   2.959   6.409  19.659
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  69.3413     0.2093   331.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.467 on 2046 degrees of freedom
##   (9 observations deleted due to missingness)
```

```
forward_AIC <- step(lm_intercept_only, scope = formula(lm_all),
                    direction = "forward", trace = 0, k = 2)
forward_AIC$anova
```

```
##            Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1               NA         NA      2046  183383.18 9203.680
## 2    + hiv_aids -1 56395.9326      2045  126987.25 8453.425
## 3  + diphtheria -1 26113.6073      2044  100873.64 7984.169
## 4     + alcohol -1 16653.4801      2043   84220.16 7616.820
```

```
## 5       + polio -1  3062.6005      2042   81157.56 7542.995
## 6 + hepatitis_b -1   229.8422      2041   80927.72 7539.190
```

```
# Perform stepwise forward selection using BIC as selction criteria on training data
forward_BIC <- step(lm_intercept_only, scope = formula(lm_all),
                    direction = "forward", trace = 0, k = log(nrow(training_data)))
forward_BIC$anova
```

```
##           Step Df  Deviance Resid. Df Resid. Dev      AIC
## 1               NA        NA      2046  183383.18 9209.308
## 2   + hiv_aids -1 56395.933      2045  126987.25 8464.682
## 3 + diphtheria -1 26113.607      2044  100873.64 8001.054
## 4    + alcohol -1 16653.480      2043   84220.16 7639.334
## 5      + polio -1  3062.601      2042   81157.56 7571.138
```

```
forward_AIC$coefficients
```

```
## (Intercept)    hiv_aids  diphtheria     alcohol       polio hepatitis_b
## 54.86854510 -0.93085026  0.07294768  0.69698493  0.06552666  0.01803251
```

```
forward_BIC$coefficients
```

```
## (Intercept)    hiv_aids  diphtheria     alcohol       polio
## 55.11227022 -0.93243972  0.08284081  0.69448882  0.07001033
```

```
anova(forward_BIC, forward_AIC)
```

```
## Analysis of Variance Table
##
## Model 1: life_expectancy ~ hiv_aids + diphtheria + alcohol + polio
## Model 2: life_expectancy ~ hiv_aids + diphtheria + alcohol + polio + hepatitis_b
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1   2042 81158
## 2   2041 80928  1    229.84 5.7966 0.01615 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Adding a polynomial on most significant predictor (based on SLR model p-values)
# Builds upon the stepwise forward selection model using AIC

M2 <- lm(life_expectancy ~ poly(hiv_aids, degree = 4, raw = TRUE) + diphtheria
         + alcohol + polio + hepatitis_b, data = training_data)

anova(forward_AIC, M2)
```

```
## Analysis of Variance Table
##
## Model 1: life_expectancy ~ hiv_aids + diphtheria + alcohol + polio + hepatitis_b
## Model 2: life_expectancy ~ poly(hiv_aids, degree = 4, raw = TRUE) + diphtheria +
##     alcohol + polio + hepatitis_b
```

```
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   2041 80928
## 2   2038 54193  3     26735 335.13 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.2.2 Model selection

```
anova(forward_AIC, M2)
```

```
## Analysis of Variance Table
##
## Model 1: life_expectancy ~ hiv_aids + diphtheria + alcohol + polio + hepatitis_b
## Model 2: life_expectancy ~ poly(hiv_aids, degree = 4, raw = TRUE) + diphtheria +
##     alcohol + polio + hepatitis_b
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1   2041 80928
## 2   2038 54193  3     26735 335.13 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(forward_AIC)
```

```
##
## Call:
## lm(formula = life_expectancy ~ hiv_aids + diphtheria + alcohol +
##     polio + hepatitis_b, data = training_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.0836  -3.9441   0.2917   3.9595  20.9967
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 54.868545   0.582292  94.229  < 2e-16 ***
## hiv_aids    -0.930850   0.029589 -31.460  < 2e-16 ***
## diphtheria   0.072948   0.008880   8.214 3.74e-16 ***
## alcohol      0.696985   0.035887  19.422  < 2e-16 ***
## polio        0.065527   0.008181   8.010 1.91e-15 ***
## hepatitis_b  0.018033   0.007490   2.408   0.0161 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.297 on 2041 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.5587, Adjusted R-squared:  0.5576
## F-statistic: 516.8 on 5 and 2041 DF,  p-value: < 2.2e-16
```

```
summary(M2)
```

```
##
## Call:
## lm(formula = life_expectancy ~ poly(hiv_aids, degree = 4, raw = TRUE) +
##     diphtheria + alcohol + polio + hepatitis_b, data = training_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.7906  -3.3572   0.3788   3.3754  15.1053
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                            6.282e+01  5.405e-01 116.227  < 2e-16
## poly(hiv_aids, degree = 4, raw = TRUE)1 -5.141e+00  1.583e-01 -32.469  < 2e-16
## poly(hiv_aids, degree = 4, raw = TRUE)2  4.281e-01  2.177e-02  19.659  < 2e-16
## poly(hiv_aids, degree = 4, raw = TRUE)3 -1.329e-02  8.782e-04 -15.137  < 2e-16
## poly(hiv_aids, degree = 4, raw = TRUE)4  1.326e-04  1.046e-05  12.678  < 2e-16
## diphtheria                             4.515e-02  7.325e-03   6.164 8.53e-10
## alcohol                                6.297e-01  2.950e-02  21.348  < 2e-16
## polio                                  3.801e-02  6.758e-03   5.625 2.11e-08
## hepatitis_b                            7.473e-03  6.157e-03   1.214    0.225
##
## (Intercept)                           ***
## poly(hiv_aids, degree = 4, raw = TRUE)1 ***
## poly(hiv_aids, degree = 4, raw = TRUE)2 ***
## poly(hiv_aids, degree = 4, raw = TRUE)3 ***
## poly(hiv_aids, degree = 4, raw = TRUE)4 ***
## diphtheria                            ***
## alcohol                               ***
## polio                                 ***
## hepatitis_b
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.157 on 2038 degrees of freedom
##   (9 observations deleted due to missingness)
## Multiple R-squared:  0.7045, Adjusted R-squared:  0.7033
## F-statistic: 607.3 on 8 and 2038 DF,  p-value: < 2.2e-16
```

```r
# Use fitted model on test data
test_pred <- predict(M2, newdata = test_data[ , -1])
test_obs_pred <- cbind(test_data$life_expectancy, test_pred)
colnames(test_obs_pred) <- c("Observed", "Predicted")

head(test_obs_pred)
```

```
##    Observed Predicted
## 2      59.9  67.78183
## 6      58.8  68.29644
## 18     77.5  74.03056
## 20     76.9  74.51791
## 21     76.6  74.66275
## 24     75.3  74.81388
```

5. Results and interpretations

From the linear regression model that we fit on the data we determined that alcohol, hepatitis b, polio, diphtheria, and HIV/AIDS are all significant predictors of life expectancy. We came to this conclusion as HIV/AIDS had a p-value of 2e-16, diphtheria had a p-value of 3.74e-16, polio had a p-value of 1.91e-15, hepatitis b had a p value of 0.0161, and alcohol had a p value of 2e-16. All of these p-values are less than 0.05 and thus are statistically significant.