Multiple Linear Regression and the Best Model to Estimate Life Expectancy

Code and Output

Create a regression model with all 6 independent variables

```
> ## Regression the whole dataset all_variables_last5yrs
>
> # Create a regression model with all 6 predictor variables
> mod_wholeds_6p <- lm(Life_Expectancy ~ Income_per_person + Food_Supply + Improved_Water + All_Cancer_Death + Total_Health_Spending_Per_Person + Government_Health
_Spending_Per_Person, data = all_variables_last5yrs)
> summary(mod_wholeds_6p)

Call:
lm(formula = Life_Expectancy ~ Income_per_person + Food_Supply +
    Improved_Water + All_Cancer_Death + Total_Health_Spending_Per_Person +
    Government_Health_Spending_Per_Person, data = all_variables_last5yrs)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5427  -1.4681   0.4288   2.2560  11.2301

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                          35.4504098  1.9905463  17.809  < 2e-16 ***
Income_per_person                     0.0001567  0.0000781   2.006  0.04562 *
Food_Supply                           0.0023252  0.0008308   2.799  0.00542 **
Improved_Water                        0.3530365  0.0258351  13.665  < 2e-16 ***
All_Cancer_Death                     -0.0344841  0.0086496  -3.987 8.21e-05 ***
Total_Health_Spending_Per_Person     -0.0008483  0.0009298  -0.912  0.36220
Government_Health_Spending_Per_Person  0.0009117  0.0013630   0.669  0.50405
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.466 on 339 degrees of freedom
Multiple R-squared:  0.6952,	Adjusted R-squared:  0.6898
F-statistic: 128.9 on 6 and 339 DF,  p-value: < 2.2e-16

>
> SSE=sum(mod_wholeds_6p$residuals^2)
> SSE
[1] 6761.735
>
```

Use different methods to select useful independent variables

Method 1: Create Best Subsets Regression

```
> ## Use different methods to select useful predictor variables
> # Create Best Subsets Regression
>
> ols_best_subset(mod_wholeds_6p)
```



```
> ## Use different methods to select useful predictor variables
> # Create Best Subsets Regression
>
> ols_best_subset(mod_wholeds_6p)
```



```
                                                    Best Subsets Regression
------------------------------------------------------------------------------------------------------------------------------------
Model Index    Predictors
------------------------------------------------------------------------------------------------------------------------------------
     1         Improved_Water
     2         Income_per_person Improved_Water
     3         Income_per_person Improved_Water All_Cancer_Death
     4         Income_per_person Food_Supply Improved_Water All_Cancer_Death
     5         Income_per_person Food_Supply Improved_Water All_Cancer_Death Total_Health_Spending_Per_Person
     6         Income_per_person Food_Supply Improved_Water All_Cancer_Death Total_Health_Spending_Per_Person Government_Health_Spending_Per_Person
------------------------------------------------------------------------------------------------------------------------------------

                                                  Subsets Regression Summary
------------------------------------------------------------------------------------------------------------------------
                       Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC         SBC        MSEP       FPE       HSP       APC
------------------------------------------------------------------------------------------------------------------------
  1        0.6319      0.6308      0.6258    67.4453    2081.7496    1099.1763    2093.2890    23.8789    23.8781    0.0692    0.3724
  2        0.6769      0.6750      0.6707    19.3722    2038.6157    1056.4843    2054.0015    21.0812    21.0794    0.0611    0.3288
  3        0.6878      0.6851      0.6801     9.1855    2028.6789    1046.7461    2047.9111    20.4857    20.4826    0.0594    0.3195
  4        0.6944      0.6909      0.6841     3.8415    2023.2815    1041.5566    2046.3601    20.1703    20.1656    0.0585    0.3145
  5        0.6948      0.6903      0.6834     5.4474    2024.8800    1043.2054    2051.8051    20.2658    20.2590    0.0587    0.3160
  6        0.6952      0.6898      0.6834     7.0000    2026.4237    1044.8065    2057.1952    20.3589    20.3497    0.0590    0.3174
------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
 MSEP: Estimated error of prediction, assuming multivariate normality
 FPE: Final Prediction Error
 HSP: Hocking's Sp
 APC: Amemiya Prediction Criteria
```

Method 2: Stepwise Forward Regression

```
> # Create Stepwise Forward Regression
>
> ols_step_forward(mod_wholeds_6p)
We are selecting variables based on p value...
1 variable(s) added....
1 variable(s) added...
1 variable(s) added...
1 variable(s) added...
No more variables satisfy the condition of penter: 0.3
Forward Selection Method

Candidate Terms:

1 . Income_per_person
2 . Food_Supply
3 . Improved_Water
4 . All_Cancer_Death
5 . Total_Health_Spending_Per_Person
6 . Government_Health_Spending_Per_Person

---------------------------------------------------------------------------------
                                Selection Summary
---------------------------------------------------------------------------------
        Variable                        Adj.
Step         Entered         R-Square    R-Square     C(p)         AIC        RMSE
---------------------------------------------------------------------------------
   1    Improved_Water          0.6319      0.6308    67.4453    2081.7496    4.8725
   2    Income_per_person       0.6769      0.6750    19.3722    2038.6157    4.5715
   3    All_Cancer_Death        0.6878      0.6851     9.1855    2028.6789    4.4998
   4    Food_Supply             0.6944      0.6909     3.8415    2023.2815    4.4585
---------------------------------------------------------------------------------
```

Method 3: Random Forest

```
> # Create Random forest
>
> output_forest_wholeds <- randomForest(Life_Expectancy ~ Income_per_person + Food_Supply + Improved_Water + All_Cancer_Death + Total_Health_Spending_Per_Person +
Government_Health_Spending_Per_Person, data = all_variables_last5yrs, importance=TRUE)
>
> # View the Random forest results
>
> print(output_forest_wholeds)

Call:
 randomForest(formula = Life_Expectancy ~ Income_per_person +      Food_Supply + Improved_Water + All_Cancer_Death + Total_Health_Spending_Per_Person +      Govern
ment_Health_Spending_Per_Person, data = all_variables_last5yrs,      importance = TRUE)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 2

          Mean of squared residuals: 13.93028
                    % Var explained: 78.27
>
> # Importance of each predictor (Random forest)
>
> print(importance(output_forest_wholeds))
                                  %IncMSE IncNodePurity
Income_per_person                17.81619     4956.8886
Food_Supply                      11.24182     1683.3101
Improved_Water                   20.44714     5244.6988
All_Cancer_Death                 17.00173      991.8082
Total_Health_Spending_Per_Person 20.50593     4763.1213
Government_Health_Spending_Per_Person 20.81433     4142.9649
>
> # Pseudo R-Squared
>
> mean(output_forest_wholeds$rsq)
[1] 0.778114
>
```

10-fold Cross-Validation

10-fold Cross-validation for all 6 independent variables:
Income, improved drinking water source, food supply, cancer death, total health spending, and government health spending

```
> ## 10-fold Cross-Validation for all 6 variables
>
> # Change the dataset to data frame to prevent an error when running cv.lm() function
> all_variables_last5yrs_df <- as.data.frame(all_variables_last5yrs)
>
> # Remove unnecessary columns for cv.lm () function
> all_variables_last5yrs_df <- select(all_variables_last5yrs_df, Income_per_person, Life_Expectancy, Improved_Water, Food_Supply, Total_Health_Spending_Per_Person,
 Government_Health_Spending_Per_Person, All_Cancer_Death)
>
> # cv.lm(all_variables_last5yrs_df, form.lm = formula(Life_Expectancy ~ .), m=10)
>
> # Set seed for reproducibility
> data(all_variables_last5yrs_df)
```

```
data set ◆all_variables_last5yrs_df◆ not found
> set.seed(121)
>
> # Fit linear regression model
> mod_10fold_6p <- train(Life_Expectancy ~ ., all_variables_last5yrs_df, method = "lm", trControl = trainControl(method = "cv", number = 10, verboseIter = TRUE))
+ Fold01: intercept=TRUE
- Fold01: intercept=TRUE
+ Fold02: intercept=TRUE
- Fold02: intercept=TRUE
+ Fold03: intercept=TRUE
- Fold03: intercept=TRUE
+ Fold04: intercept=TRUE
- Fold04: intercept=TRUE
+ Fold05: intercept=TRUE
- Fold05: intercept=TRUE
+ Fold06: intercept=TRUE
- Fold06: intercept=TRUE
+ Fold07: intercept=TRUE
- Fold07: intercept=TRUE
+ Fold08: intercept=TRUE
- Fold08: intercept=TRUE
+ Fold09: intercept=TRUE
- Fold09: intercept=TRUE
+ Fold10: intercept=TRUE
- Fold10: intercept=TRUE
Aggregating results
Fitting final model on full training set
> mod_10fold_6p
Linear Regression

346 samples
  6 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 313, 312, 310, 311, 310, 312, ...
Resampling results:

  RMSE     Rsquared   MAE
  4.45503  0.6873986  3.086018

Tuning parameter 'intercept' was held constant at a value of TRUE
```

10-fold Cross-Validation for 4 independent variables:
Income, improved drinking water source, food supply, and cancer death

```
> ## 10-fold Cross-Validation for 4 stronger predictor variables based on Best Subsets Regression and Stepwise Forward Regression
>
> # Remove unnecessary columns for cv.lm () function
> all_variables_last5yrs_df_4p_iwfc <- select(all_variables_last5yrs_df, Income_per_person, Life_Expectancy, Improved_Water, Food_Supply, All_Cancer_Death)
>
> # cv.lm(all_variables_last5yrs_df_4p_iwfc, form.lm = formula(Life_Expectancy ~ .), m=10)
>
> # Set seed for reproducibility
> data(all_variables_last5yrs_df_4p_iwfc)
data set ◆all_variables_last5yrs_df_4p_iwfc◆ not found
> set.seed(122)
>
> # Fit linear regression model
> mod_10fold_4p_iwfc <- train(Life_Expectancy ~ ., all_variables_last5yrs_df_4p_iwfc, method = "lm", trControl = trainControl(method = "cv", number = 10, verboseIt
er = TRUE))
+ Fold01: intercept=TRUE
- Fold01: intercept=TRUE
+ Fold02: intercept=TRUE
- Fold02: intercept=TRUE
+ Fold03: intercept=TRUE
- Fold03: intercept=TRUE
+ Fold04: intercept=TRUE
- Fold04: intercept=TRUE
+ Fold05: intercept=TRUE
- Fold05: intercept=TRUE
+ Fold06: intercept=TRUE
- Fold06: intercept=TRUE
+ Fold07: intercept=TRUE
- Fold07: intercept=TRUE
+ Fold08: intercept=TRUE
- Fold08: intercept=TRUE
+ Fold09: intercept=TRUE
- Fold09: intercept=TRUE
+ Fold10: intercept=TRUE
- Fold10: intercept=TRUE
Aggregating results
Fitting final model on full training set
> mod_10fold_4p_iwfc
Linear Regression

346 samples
  4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 311, 312, 312, 311, 312, 311, ...
Resampling results:

  RMSE      Rsquared   MAE
  4.415488  0.7078506  3.069383

Tuning parameter 'intercept' was held constant at a value of TRUE
```

10-fold Cross-Validation for 4 independent variables:
Income, improved drinking water source, total health spending, and government health spending

```
>
> ## 10-fold Cross-Validation for 4 stronger predictor variables based on Random Forest
>
> # Remove unnecessary columns for cv.lm () function
> all_variables_last5yrs_df_4p_iwtg <- select(all_variables_last5yrs_df, Income_per_person, Life_Expectancy, Improved_Water, Total_Health_Spending_Per_Person, Gove
rnment_Health_Spending_Per_Person)
>
> # cv.lm(all_variables_last5yrs_df_4p_iwtg, form.lm = formula(Life_Expectancy ~ .), m=10)
>
> data(all_variables_last5yrs_df_4p_iwtg)
data set ?all_variables_last5yrs_df_4p_iwtg? not found
> set.seed(123)
>
> # Fit linear regression model
> mod_10fold_4p_iwtg <- train(Life_Expectancy ~ ., all_variables_last5yrs_df_4p_iwtg, method = "lm", trControl = trainControl(method = "cv", number = 10, verboseIt
er = TRUE))
+ Fold01: intercept=TRUE
- Fold01: intercept=TRUE
+ Fold02: intercept=TRUE
- Fold02: intercept=TRUE
+ Fold03: intercept=TRUE
- Fold03: intercept=TRUE
+ Fold04: intercept=TRUE
- Fold04: intercept=TRUE
+ Fold05: intercept=TRUE
- Fold05: intercept=TRUE
+ Fold06: intercept=TRUE
- Fold06: intercept=TRUE
+ Fold07: intercept=TRUE
- Fold07: intercept=TRUE
+ Fold08: intercept=TRUE
- Fold08: intercept=TRUE
+ Fold09: intercept=TRUE
- Fold09: intercept=TRUE
+ Fold10: intercept=TRUE
- Fold10: intercept=TRUE
Aggregating results
Fitting final model on full training set
> mod_10fold_4p_iwtg
Linear Regression

346 samples
  4 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 312, 312, 311, 311, 312, 311, ...
Resampling results:

  RMSE      Rsquared   MAE
  4.508128  0.6825604  3.121801

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Create a regression model with independent variables based on Random Forest

Independent variables: Income, improved drinking water source, total health spending, and government health spending

```
> # Create a regression model with 4 stronger predictor variables based on Random Forest
>
> mod_wholeds_4p_iwtg <- lm(Life_Expectancy ~ Income_per_person + Improved_Water + Total_Health_Spending_Per_Person + Government_Health_Spending_Per_Person, data =
 all_variables_last5yrs)
> summary(mod_wholeds_4p_iwtg)

Call:
lm(formula = Life_Expectancy ~ Income_per_person + Improved_Water +
    Total_Health_Spending_Per_Person + Government_Health_Spending_Per_Person,
    data = all_variables_last5yrs)

Residuals:
     Min       1Q   Median       3Q      Max
-26.9355  -1.8196   0.3308   2.6082  13.4135

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      3.516e+01  1.798e+00  19.558   <2e-16 ***
Income_per_person                1.931e-04  7.964e-05   2.425   0.0158 *
Improved_Water                   3.867e-01  2.111e-02  18.319   <2e-16 ***
Total_Health_Spending_Per_Person -5.002e-04  9.401e-04  -0.532   0.5950
Government_Health_Spending_Per_Person  4.542e-04  1.394e-03   0.326   0.7448
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.583 on 341 degrees of freedom
Multiple R-squared:  0.6772,    Adjusted R-squared:  0.6734
F-statistic: 178.8 on 4 and 341 DF,  p-value: < 2.2e-16

>
> SSE=sum(mod_wholeds_4p_iwtg$residuals^2)
> SSE
[1] 7162.108
>
```

Create a regression model with independent variables based on Best Subsets Regression and Stepwise Forward Regression

Independent variables: Income, improved drinking water source, food supply, and cancer death

```
> # Create a regression model with 4 stronger predictor variables based on Best Subsets Regression and Stepwise Forward Regression
>
> mod_wholeds_4p_iwtg <- lm(Life_Expectancy ~ Income_per_person + Food_Supply + Improved_Water + All_Cancer_Death, data = all_variables_last5yrs)
> summary(mod_wholeds_4p_iwfc)

Call:
lm(formula = Life_Expectancy ~ Income_per_person + Food_Supply +
    Improved_Water + All_Cancer_Death, data = all_variables_last5yrs)

Residuals:
     Min       1Q   Median       3Q      Max
-25.6030  -1.6209   0.4656   2.2785  11.3246

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.553e+01  1.984e+00  17.910  < 2e-16 ***
Income_per_person  1.428e-04  2.876e-05   4.965 1.09e-06 ***
Food_Supply        2.206e-03  8.127e-04   2.715  0.00697 **
Improved_Water     3.555e-01  2.527e-02  14.069  < 2e-16 ***
All_Cancer_Death  -3.435e-02  8.595e-03  -3.997 7.86e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.459 on 341 degrees of freedom
Multiple R-squared:  0.6944,    Adjusted R-squared:  0.6909
F-statistic: 193.7 on 4 and 341 DF,  p-value: < 2.2e-16

>
> SSE=sum(mod_wholeds_4p_iwfc$residuals^2)
> SSE
[1] 6778.519
>
```