The University of British Columbia

Data Science 570 Predictive Modelling

Lab Assignment 2

The TA will demonstrate exercises 1 through 6. You are expected to submit answers to exercise 7. Instructions: Please use a png, pdf or html file for submission.

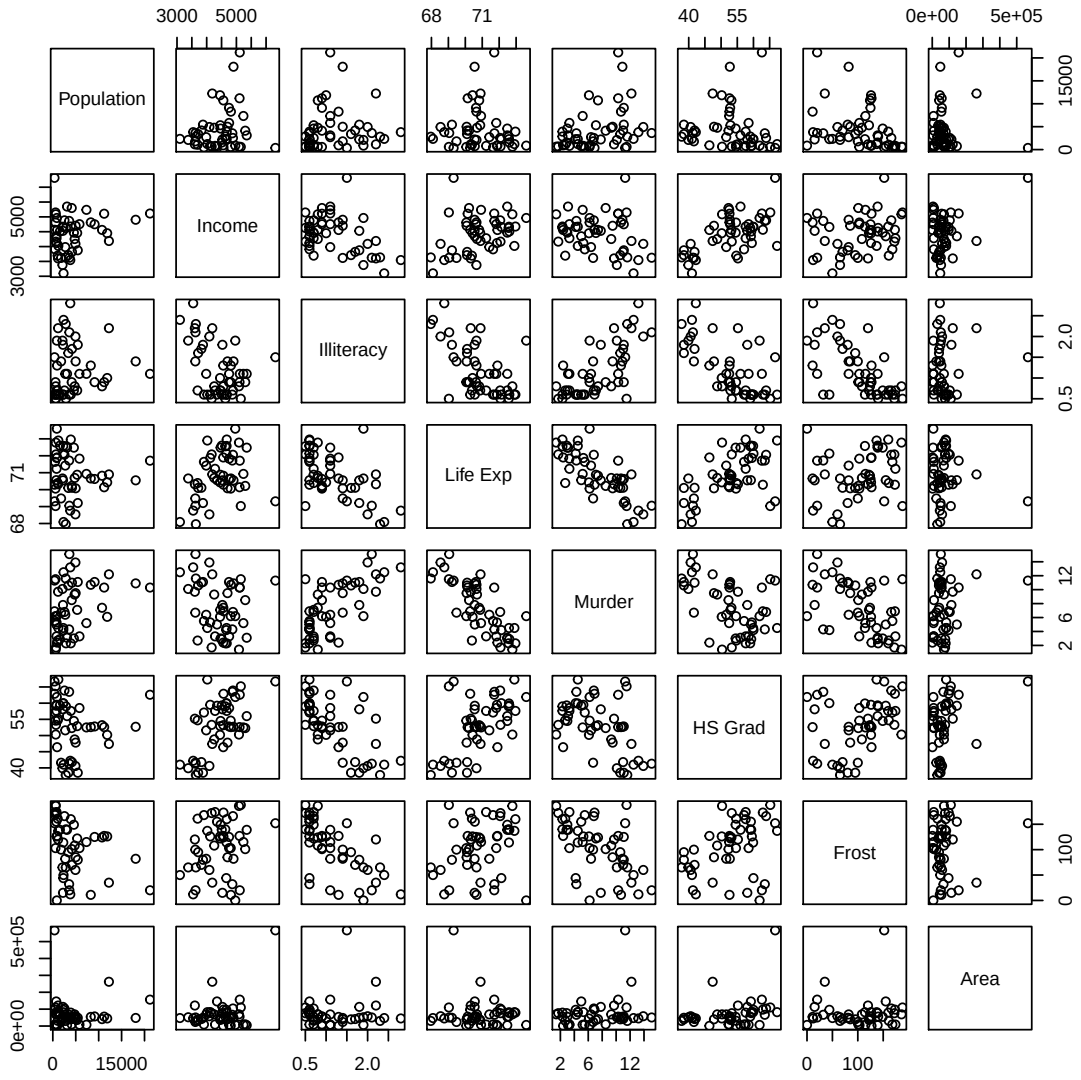
1. Life Expectancy Data

   Consider the data set concerning the Life Expectancy rates for all 50 states from the 1970's. For more information on this data set type:

```
?state.x77
```

```
## starting httpd help server ...  done
```

```
pairs(state.x77)
```



Now fit a standard linear model, and add the line to the plot in red. Note that for markdown, we need to re-plot X and Y in any new code chunk when attempting to add lines. We have seen a few model

ts for this data in class, namely:

```
data(state)
# renames the rows to 2-letter abbreviation of state names
statedata <- data.frame(state.x77,row.names=state.abb)
attach(statedata)
fit1 <- lm(Life.Exp~Illiteracy)
fit2 <- lm(Life.Exp~Illiteracy+Murder+HS.Grad+Frost)
```

fit1 is a simple linear regression model, that is, it only uses a single predictor variable. fit2 is an example of a multiple linear regression model since it uses more than one predictor variable. Let's take a look at the full model which uses all possible predictor variables in our data (i.e. Population, Income, Illiteracy, Life.Exp, Murder, HS.Grad, Frost, Area) to predict the response variable Life.Exp.

```
fit <- lm(Life.Exp~., data = statedata)
summary(fit)

##
## Call:
## lm(formula = Life.Exp ~ ., data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population   5.180e-05  2.919e-05   1.775   0.0832 .
## Income      -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## HS.Grad      4.893e-02  2.332e-02   2.098   0.0420 *
## Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362,Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

2. Adjusted R^2 and BIC

   The function regsubsets() in the library "leaps" can be used for regression subset selection. Thereafter, one can view the ranked models according to different scoring criteria

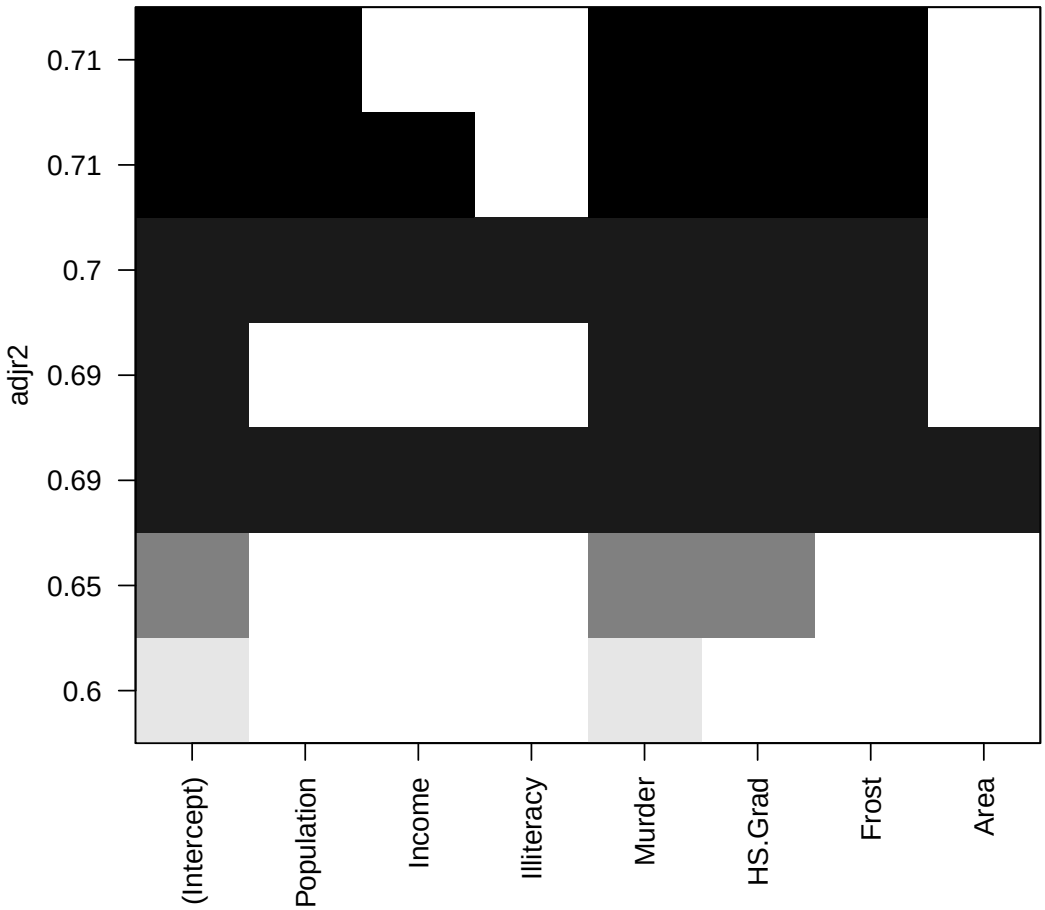
by plotting the results of regsubsets(). Before using the function for the first time you will need to install the library using the R GUI. Alternatively, you can use the command install.packages("leaps") to install it.

```
#install.packages("leaps")
library(leaps)
leaps=regsubsets(Life.Exp~.,data=statedata )
reg_summary = summary(leaps)
reg_summary$rsq

## [1] 0.6097201 0.6628461 0.7126624 0.7360328 0.7361014 0.7361440 0.7361563

reg_summary$adjr2

## [1] 0.6015893 0.6484991 0.6939230 0.7125690 0.7061129 0.6993268 0.6921823
```

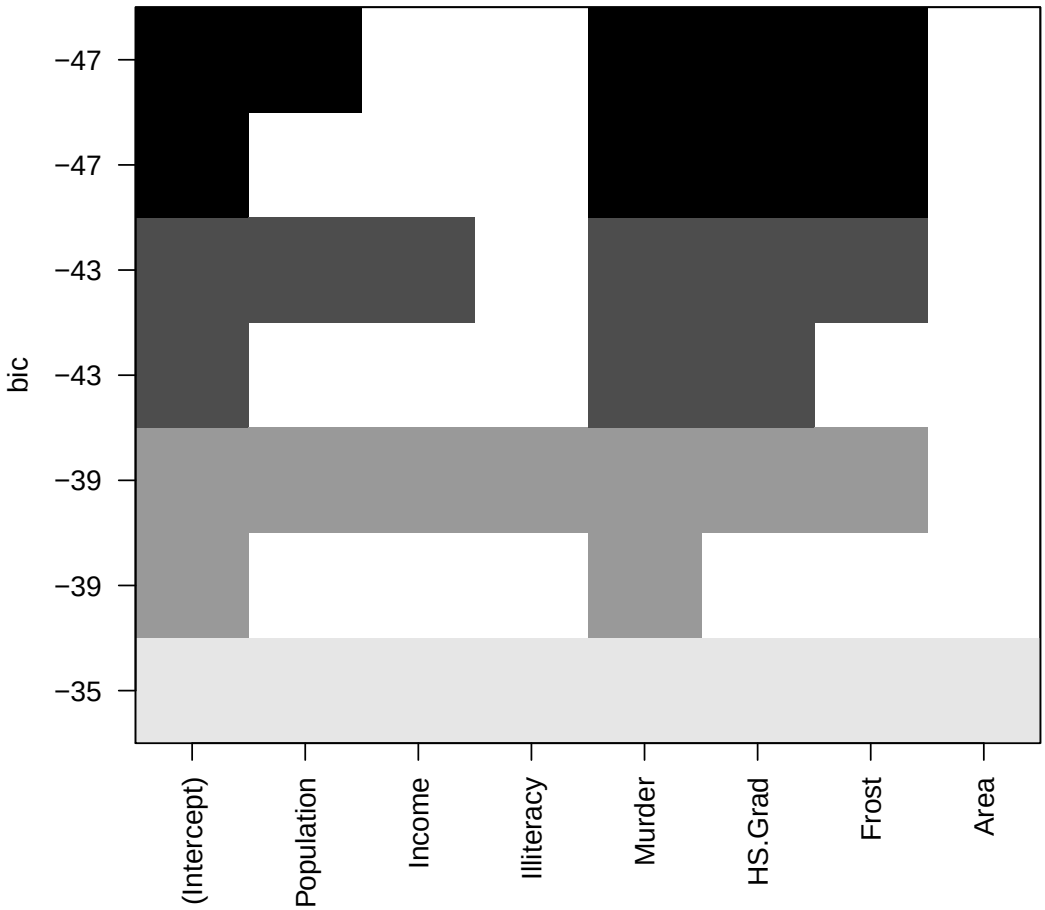
We see that the R^2 statistic increases from 61% when only one variable is included in the model to almost 74% when all variables are included. As expected, the R^2 statistic increases monotonically as more variables are included.

Plotting adjusted R^2 and BIC for all of the models at once will help us decide which model to select.

```
plot(leaps, scale = "adjr2")
```




```
plot(leaps, scale = "bic")
```



Each row in this graph represents a model; the shaded rectangles in the columns indicate the variables included in the given model. The numbers on the left margin are the values of adjusted R squared/BIC Criterion; not that the axis is not quantitative but is ordered. The darkness of the shading simply represents the ordering of the adjusted R squared values. In the example above, the best model with intercept and Population, Murder, HS.Grad, and Frost has the largest adjusted R^2= 0.713. The BIC method gives the same best model.

3. AIC

We can apply the AIC to the state data. The function does not evaluate the AIC for all possible models but uses a search method that compares models sequentially. Thus it bears some comparison to the Mixed Selection (stepwise method) described below but with the advantage that no dubious p-values are used.

```
step(fit)

## Start:  AIC=-22.18
```

```
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
##     Frost + Area
##
##              Df Sum of Sq    RSS     AIC
## - Area        1    0.0011 23.298 -24.182
## - Income      1    0.0044 23.302 -24.175
## - Illiteracy  1    0.0047 23.302 -24.174
## <none>                    23.297 -22.185
## - Population  1    1.7472 25.044 -20.569
## - Frost       1    1.8466 25.144 -20.371
## - HS.Grad     1    2.4413 25.738 -19.202
## - Murder      1   23.1411 46.438  10.305
##
## Step:  AIC=-24.18
## Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +
##     Frost
##
##              Df Sum of Sq    RSS     AIC
## - Illiteracy  1    0.0038 23.302 -26.174
## - Income      1    0.0059 23.304 -26.170
## <none>                    23.298 -24.182
## - Population  1    1.7599 25.058 -22.541
## - Frost       1    2.0488 25.347 -21.968
## - HS.Grad     1    2.9804 26.279 -20.163
## - Murder      1   26.2721 49.570  11.569
##
## Step:  AIC=-26.17
## Life.Exp ~ Population + Income + Murder + HS.Grad + Frost
##
##              Df Sum of Sq    RSS     AIC
## - Income      1     0.006 23.308 -28.161
## <none>                    23.302 -26.174
## - Population  1     1.887 25.189 -24.280
## - Frost       1     3.037 26.339 -22.048
## - HS.Grad     1     3.495 26.797 -21.187
## - Murder      1    34.739 58.041  17.456
##
## Step:  AIC=-28.16
## Life.Exp ~ Population + Murder + HS.Grad + Frost
##
##              Df Sum of Sq    RSS     AIC
## <none>                    23.308 -28.161
## - Population  1     2.064 25.372 -25.920
## - Frost       1     3.122 26.430 -23.877
## - HS.Grad     1     5.112 28.420 -20.246
## - Murder      1    34.816 58.124  15.528
##
## Call:
```

```
## lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
##     data = statedata)
##
## Coefficients:
## (Intercept)   Population       Murder      HS.Grad        Frost
##   7.103e+01    5.014e-05   -3.001e-01    4.658e-02   -5.943e-03
```

Again, The AIC method gives the same best model.

4. Mixed (Stepwise) selection

   Before we learn the finer details, let me again provide a broad overview of the steps involved. First, we start with no predictors in our "stepwise model." Then, at each step along the way we either enter or remove a predictor based on the t-tests for the slope parameters that are obtained. We stop when no more predictors can be justifiably entered or removed from our stepwise model, thereby leading us to a "final model."

   Now, let's make this process a bit more concrete. Here goes: Starting the procedure. The first thing we need to do is set a significance level for deciding when to enter a predictor into the stepwise model. We'll call this the Alpha-to-Enter significance level and will denote it as α_{enter} (or α_{IN}). Of course, we also need to set a significance level for deciding when to remove a predictor from the stepwise model. We'll call this the Alpha-to-Remove significance level and will denote it as α_{remove} (or α_{OUT}). That is, first:

   - Specify an Alpha-to-Enter significance level. This will typically be greater than the usual 0.05 level so that it is not too difficult to enter predictors into the model. We suggest α_{enter}=0.1.
   - Specify an Alpha-to-Remove significance level. This will typically be greater than the usual 0.05 level so that it is not too easy to remove predictors from the model. Again, we suggest α_{remove}=0.1.

   Step 1.
   1. Fit each of the one-predictor models that is, regress y on x1, regress y on x2, ..., and regress y on xk.
   2. Of those predictors whose t-test P-value is less than $\alpha_{enter} = 0.1$, the first predictor put in the stepwise model is the predictor that has the smallest t-test P-value.
   3. If no predictor has a t-test P-value less than $\alpha_{enter} = 0.1$, stop.

   Step 2.
   1. Suppose x1 had the smallest t-test P-value below $\alpha_{enter} = 0.1$ and therefore was deemed the "best" single predictor arising from the the first step.
   2. Now, fit each of the two-predictor models that include x1 as a predictor that is, regress y on x1 and x2, regress y on x1 and x3, ..., and regress y on x1 and xk.
   3. Of those predictors whose t-test P-value is less than $\alpha_{enter} = 0.1$, the second predictor put in the stepwise model is the predictor that has the smallest t-test P-value.
   4. If no predictor has a t-test P-value less than $\alpha_{enter} = 0.1$, stop. The model with the one predictor obtained from the first step is your final model.
   5. But, suppose instead that x2 was deemed the "best" second predictor and it is therefore entered into the stepwise model.

6. Now, since x1 was the first predictor in the model, step back and see if entering x2 into the stepwise model somehow affected the significant of the x1 predictor. That is, check the t-test P-value for testing $\beta_1 = 0$. If the t-test P-value for $\beta_1 = 0$ has become not significant that is, the P-value is greater than α_{remove}=0.1 remove x1 from the stepwise model.

Step 3. 1. Suppose both x1 and x2 made it into the two-predictor stepwise model and remained there. Now, fit each of the three-predictor models that include x1 and x2 as predictors that is, regress y on x1, x2, and x3, regress y on x1, x2, and x4, ..., and regress y on x1, x2, and xk.

2. Of those predictors whose t-test P-value is less than $\alpha_{enter} = 0.1$, the third predictor put in the stepwise model is the predictor that has the smallest t-test P-value.
3. If no predictor has a t-test P-value less than $\alpha_{enter} = 0.1$, stop. The model containing the two predictors obtained from the second step is your final model.
4. But, suppose instead that x3 was deemed the "best" third predictor and it is therefore entered into the stepwise model.
4. Now, since x1 and x2 were the
rst predictors in the model, step back and see if entering x3 into the stepwise model somehow affected the significance of the x1 and x2 predictors. That is, check the t-test P-values for testing
$\beta_1 = 0$ and $\beta_2 = 0$. If the t-test P-value for either
$\beta_=0$ or $\beta_2 = 0$ has become not significant that is, the P-value is greater than $\alpha_{remove} = 0.1$ remove the predictor from the stepwise model.

Stopping the procedure. Continue the steps as described above until adding an additional predictor does not yield a t-test P-value below $\alpha_{enter} = 0.1$.

Let's return to our state data example so we can try out the stepwise procedure as described above. Now, regressing Life.Exp on Population, regressing Life.Exp on Income, regressing Life.Exp on Illiteracy, d regressing Life.Exp on Murder, regressing Life.Exp on HS.Grad, regressing Life.Exp on Frost, and regressing Life.Exp on Area, we obtain:

```
summary(lm(Life.Exp~ Population, data=statedata))$coefficients[,4]

##  (Intercept)   Population
## 7.898980e-78 6.386594e-01

summary(lm(Life.Exp~ Income, data=statedata))$coefficients[,4]

##  (Intercept)       Income
## 1.979131e-43 1.561728e-02

summary(lm(Life.Exp~ Illiteracy, data=statedata))$coefficients[,4]

##  (Intercept)   Illiteracy
## 3.474322e-73 6.969250e-06

summary(lm(Life.Exp~ Murder, data=statedata))$coefficients[,4]
```

```
##  (Intercept)       Murder
## 4.716796e-78 2.260070e-11
```

```
summary(lm(Life.Exp~ HS.Grad, data=statedata))$coefficients[,4]
```

```
##  (Intercept)      HS.Grad
## 9.916827e-48 9.196096e-06
```

```
summary(lm(Life.Exp~ Frost, data=statedata))$coefficients[,4]
```

```
##  (Intercept)        Frost
## 4.326144e-68 6.598740e-02
```

```
summary(lm(Life.Exp~ Area, data=statedata))$coefficients[,4]
```

```
##  (Intercept)         Area
## 3.456070e-79 4.581464e-01
```

The predictor Murder having the smallest t-test P-value it is 2.260070e-11. As a result of the first step, we enter Murder into our stepwise model.

Now, following the step 2, we fit each of the two-predictor models that include Murder as a predictor, obtaining:

```
g1 <- lm(Life.Exp~ Murder, data=statedata)
summary(update(g1,. ~ . +Population))$coefficients[,4]
```

```
##  (Intercept)       Murder   Population
## 1.546745e-77 2.145114e-12 1.636940e-02
```

```
summary(update(g1,. ~ . +Income))$coefficients[,4]
```

```
##  (Intercept)       Murder       Income
## 3.319429e-50 1.222489e-10 6.663619e-02
```

```
summary(update(g1,. ~ . +Illiteracy))$coefficients[,4]
```

```
##  (Intercept)       Murder   Illiteracy
## 1.561385e-75 7.964272e-07 5.429104e-01
```

```
summary(update(g1,. ~ . +HS.Grad))$coefficients[,4]
```

```
##  (Intercept)       Murder      HS.Grad
## 5.909435e-49 2.180515e-08 9.088366e-03
```

```
summary(update(g1,. ~ . +Frost))$coefficients[,4]
```

```
##  (Intercept)       Murder        Frost
## 2.363681e-64 2.054101e-11 3.520523e-02
```

```
summary(update(g1,. ~ . +Area))$coefficients[,4]
```

```
##  (Intercept)       Murder         Area
## 2.728406e-76 3.473539e-11 4.243751e-01
```

The P-value for HS.Grad is smallest. As a result of the second step, we enter HS.Grad into our stepwise model. Now, since Murder was the first predictor in the model, we must step back and see if entering HS.Grad into the stepwise model affected the significance of the Murder predictor. The t-test P-value for testing Murder is 2.180515e-08, and thus smaller than $\alpha remove = 0.1$. Therefore, we proceed to the third step with both Murder and HS.Grad as predictors in our stepwise model.

```
g2 <- update(g1,. ~ . +HS.Grad)
summary(update(g2,. ~ . +Population))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad   Population
## 3.946928e-49 1.905005e-09 1.117967e-02 1.994926e-02

summary(update(g2,. ~ . +Income))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad       Income
## 1.332621e-46 2.916433e-08 6.050626e-02 6.924184e-01

summary(update(g2,. ~ . +Illiteracy))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad   Illiteracy
## 2.410921e-44 3.632048e-07 8.254485e-03 4.094209e-01

summary(update(g2,. ~ . +Frost))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost
## 5.253889e-49 8.039156e-10 1.950415e-03 6.987727e-03

summary(update(g2,. ~ . +Area))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad         Area
## 2.301479e-45 1.301152e-06 1.102823e-02 5.138632e-01
```

As a result of the third step, we enter Frost into our stepwise model. Now, since Murder and HS.Grad were the first predictors in the model, we must step back and see if entering Frost into the stepwise model affected the significance of the Murder and HS.Grad predictors. The t-test P-value for Murder and HS.Grad is less than $\alpha remove = 0.1$. Therefore, we proceed to the fourth step with Murder, HS.Grad, and Frost as predictors in our stepwise model.

```
g3 <- update(g2,. ~ . +Frost)
summary(update(g3,. ~ . +Population))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost   Population
## 8.612596e-49 1.774520e-10 2.968091e-03 1.801778e-02 5.200514e-02

summary(update(g3,. ~ . +Income))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost       Income
## 7.527625e-47 1.068145e-09 2.643239e-02 6.956301e-03 5.710310e-01
```

```
summary(update(g3,. ~ . +Illiteracy))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost   Illiteracy
## 1.283585e-42 3.501046e-08 1.489583e-02 9.358724e-03 5.823608e-01

summary(update(g3,. ~ . +Area))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost         Area
## 3.920753e-45 5.338424e-08 5.662013e-03 9.403878e-03 8.317269e-01
```

As a result of the fourth step, we enter Population into our stepwise model. The t-test P-value for Murder, HS.Grad, and Frost is less than $\alpha remove = 0.1$. Therefore, we proceed to the fifth step with Murder, HS.Grad, Frost, and Population as predictors in our stepwise model.

```
g4 <- update(g3,. ~ . +Population)
summary(update(g4,. ~ . +Income))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost   Population       Income
## 1.659665e-46 2.907482e-10 1.367027e-02 2.095338e-02 6.566097e-02 9.153104e-01

summary(update(g4,. ~ . +Illiteracy))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost   Population   Illiteracy
## 8.774013e-42 9.573445e-09 9.000725e-03 5.323324e-02 6.510101e-02 9.318143e-01

summary(update(g4,. ~ . +Area))$coefficients[,4]

##  (Intercept)       Murder      HS.Grad        Frost   Population         Area
## 5.812118e-45 1.160570e-08 1.028340e-02 2.108742e-02 5.613296e-02 9.693690e-01
```

Neither of the remaining predictors are eligible for entry into our stepwise model, because each t-test P-value is greater than $\alpha enter = 0.1$. That is, we stop our stepwise regression procedure. Our final regression model, based on the stepwise procedure contains only the predictors Murder, HS.Grad, Frost, and Population.

5. Backward selection

   It starts with a regression model containing all the potential predictors (full models). Then at each stage we remove the predictor with the largest p-value over $\alpha_{remove} = 0.1$.

   We illustrate the backward method. Since Area has the highest p-value, we start by removing it from the model:

```
bmod1 <- lm(Life.Exp~.-Area, data=statedata)
summary(bmod1)

##
## Call:
## lm(formula = Life.Exp ~ . - Area, data = statedata)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.49047 -0.52533 -0.02546  0.57160  1.50374
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.099e+01  1.387e+00  51.165  < 2e-16 ***
## Population   5.188e-05  2.879e-05   1.802   0.0785 .
## Income      -2.444e-05  2.343e-04  -0.104   0.9174
## Illiteracy   2.846e-02  3.416e-01   0.083   0.9340
## Murder      -3.018e-01  4.334e-02  -6.963 1.45e-08 ***
## HS.Grad      4.847e-02  2.067e-02   2.345   0.0237 *
## Frost       -5.776e-03  2.970e-03  -1.945   0.0584 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7361 on 43 degrees of freedom
## Multiple R-squared:  0.7361,Adjusted R-squared:  0.6993
## F-statistic: 19.99 on 6 and 43 DF,  p-value: 5.362e-11
```

The above is equivalent to using the following command:

```
fit <- update(fit, .~. - Area)
```

We will continue our backwards selection method using the update() formula. Since Illiteracy has the highest p-value, we remove it from the model:

```
fit <- update(fit, .~. -Illiteracy)
summary(fit)

##
## Call:
## lm(formula = Life.Exp ~ Population + Income + Murder + HS.Grad +
##     Frost, data = statedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4892 -0.5122 -0.0329  0.5645  1.5166
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.107e+01  1.029e+00  69.067  < 2e-16 ***
## Population   5.115e-05  2.709e-05   1.888   0.0657 .
## Income      -2.477e-05  2.316e-04  -0.107   0.9153
## Murder      -3.000e-01  3.704e-02  -8.099 2.91e-10 ***
## HS.Grad      4.776e-02  1.859e-02   2.569   0.0137 *
## Frost       -5.910e-03  2.468e-03  -2.395   0.0210 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7277 on 44 degrees of freedom
## Multiple R-squared:  0.7361,Adjusted R-squared:  0.7061
## F-statistic: 24.55 on 5 and 44 DF,  p-value: 1.019e-11
```

Since Income has the highest p-value, we remove it from the model:

```
fit <- update(fit, .~. -Income)
summary(fit)

##
## Call:
## lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,
##     data = statedata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.47095 -0.53464 -0.03701  0.57621  1.50683
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.103e+01  9.529e-01  74.542  < 2e-16 ***
## Population   5.014e-05  2.512e-05   1.996  0.05201 .
## Murder      -3.001e-01  3.661e-02  -8.199 1.77e-10 ***
## HS.Grad      4.658e-02  1.483e-02   3.142  0.00297 **
## Frost       -5.943e-03  2.421e-03  -2.455  0.01802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7197 on 45 degrees of freedom
## Multiple R-squared:  0.736,Adjusted R-squared:  0.7126
## F-statistic: 31.37 on 4 and 45 DF,  p-value: 1.696e-12
```

Since Population has the highest p-value, we remove it from the model:

```
fit <- update(fit, .~. -Population)
summary(fit)

##
## Call:
## lm(formula = Life.Exp ~ Murder + HS.Grad + Frost, data = statedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## HS.Grad      0.049949   0.015201   3.286  0.00195 **
## Frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127,Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Since all p-value are below 0.1 (which we have selected to be our threshold), we can stop here.

6. Forward selection

   This procedure works the same way as stepwise regression except that once a predictor is entered into the model with the smallest p-value less than $\alpha_{enter} = 0.1$, it is never removed.

7. Blood pressure data

   Some researchers observed the following data (https://onlinecourses.science.psu.edu/stat501/sites/onli
   les/data/bloodpress.txt) on 20 individuals with high blood pressure:

   - blood pressure (y = BP, in mm Hg)
   - age (x1 = Age, in years)
   - weight (x2 = Weight, in kg)
   - body surface area (x3 = BSA, in sq m)
   - duration of hypertension (x4 = Dur, in years)
   - basal pulse (x5 = Pulse, in beats per minute)
   - stress index (x6 = Stress)

   The researchers were interested in determining if a relationship exists between blood pressure and age, weight, body surface area, duration, pulse rate and/or stress level. Read the data bloodpress.txt into R.

```
#bp=read.table(".../bloodpress.txt", header = TRUE, sep = "" )
```

   7.1. (4 marks) Plot adjusted R^2 and BIC for all of the models. Comment the best model from each method.

   7.2. (4 marks) Perform the stepwise regression procedure.

   7.3. (2 marks) Perform the forward regression procedure.

   7.4. (2 marks) Perform the backward regression procedure.