RP 4

Mason Mazurek

12/3/2020

Forward Stepwise

```
# Fit an empty model with only the response
FitStart <-lm(W ~ 1, mydata)
# Fit a full model with all predictors
FitAll <-lm(W ~ League+ERA+SV+IP+H+R+ER+HR+BB+SO+AVG+AB+RBI+CS+OBP+SLG,mydata)
# Run the stepwise regression with forward selection based on the AIC criterion
step(FitStart,direction="forward", scope =formula(FitAll))
```

```
## Start:  AIC=730.42
## W ~ 1
##
##            Df Sum of Sq     RSS    AIC
## + R         1   10925.8  8353.2 606.96
## + ERA       1   10662.4  8616.5 611.62
## + AVG       1   10613.4  8665.5 612.47
## + ER        1   10545.1  8733.8 613.65
## + H         1    9856.2  9422.8 625.04
## + SV        1    7372.9 11906.0 660.12
## + OBP       1    6751.1 12527.9 667.76
## + RBI       1    5822.6 13456.4 678.49
## + SO        1    5674.0 13605.0 680.13
## + IP        1    4832.7 14446.2 689.13
## + SLG       1    3640.6 15638.3 701.03
## + BB        1    3456.4 15822.6 702.78
## + HR        1    2941.7 16337.2 707.59
## + CS        1     445.5 18833.4 728.91
## + AB        1     331.9 18947.0 729.82
## <none>                  19278.9 730.42
## + League    1      57.7 19221.3 731.97
##
## Step:  AIC=606.96
## W ~ R
##
##            Df Sum of Sq    RSS    AIC
## + RBI       1    5703.4 2649.8 436.74
## + SLG       1    5213.2 3140.0 462.20
## + OBP       1    4138.1 4215.0 506.37
## + AB        1    1073.2 7280.0 588.34
## + SV        1    1029.9 7323.3 589.23
## + SO      1     955.3 7397.9 590.75
## + AVG     1     901.3 7451.9 591.84
## + HR      1     825.5 7527.7 593.36
## + H       1     615.4 7737.7 597.48
## + League  1     290.3 8062.9 603.66
## + IP      1     197.9 8155.3 605.37
## + CS      1     165.1 8188.1 605.97
## <none>                8353.2 606.96
## + BB      1      50.1 8303.1 608.06
## + ER      1      48.4 8304.8 608.09
## + ERA     1       0.5 8352.7 608.96
##
## Step:  AIC=436.74
## W ~ R + RBI
##
##          Df Sum of Sq    RSS    AIC
## + SV      1    940.40 1709.4 372.99
## + IP      1     82.44 2567.4 434.00
## + SLG     1     56.04 2593.8 435.53
## <none>                2649.8 436.74
## + ERA     1     16.05 2633.8 437.83
## + BB      1     10.76 2639.1 438.13
## + SO      1      9.38 2640.4 438.21
## + CS      1      6.37 2643.4 438.38
## + AVG     1      5.85 2643.9 438.41
## + OBP     1      4.82 2645.0 438.47
## + League  1      4.24 2645.6 438.50
## + H       1      0.96 2648.8 438.69
## + ER      1      0.58 2649.2 438.71
## + HR      1      0.41 2649.4 438.72
## + AB      1      0.01 2649.8 438.74
##
## Step:  AIC=372.99
## W ~ R + RBI + SV
##
##          Df Sum of Sq    RSS    AIC
## + IP      1    56.098 1653.3 369.98
## + AVG     1    33.605 1675.8 372.01
## + ERA     1    30.737 1678.7 372.27
## <none>                1709.4 372.99
## + H       1    20.131 1689.3 373.21
## + ER      1    10.287 1699.1 374.08
## + HR      1     7.474 1701.9 374.33
## + BB      1     7.430 1702.0 374.34
## + SO      1     5.305 1704.1 374.52
## + League  1     4.744 1704.7 374.57
## + AB      1     2.103 1707.3 374.80
## + SLG     1     1.391 1708.0 374.87
## + CS      1     0.823 1708.6 374.92
## + OBP     1     0.656 1708.7 374.93
##
## Step:  AIC=369.98
## W ~ R + RBI + SV + IP
##
```

```
##           Df Sum of Sq    RSS    AIC
## + H       1   28.6960 1624.6 369.36
## + AVG     1   25.1295 1628.2 369.69
## <none>                1653.3 369.98
## + AB      1   13.1831 1640.1 370.78
## + League  1    9.9003 1643.4 371.08
## + BB      1    3.7544 1649.5 371.64
## + ER      1    3.1775 1650.1 371.70
## + ERA     1    2.5692 1650.7 371.75
## + HR      1    1.7332 1651.6 371.83
## + SLG     1    1.5918 1651.7 371.84
## + OBP     1    0.8891 1652.4 371.90
## + SO      1    0.7513 1652.5 371.92
## + CS      1    0.3752 1652.9 371.95
##
## Step:  AIC=369.36
## W ~ R + RBI + SV + IP + H
##
##           Df Sum of Sq    RSS    AIC
## <none>                1624.6 369.36
## + HR      1   14.7638 1609.8 369.99
## + League  1   13.5073 1611.1 370.11
## + SO      1    5.0054 1619.6 370.90
## + AB      1    3.9123 1620.7 371.00
## + ER      1    2.9934 1621.6 371.08
## + AVG     1    2.9194 1621.7 371.09
## + ERA     1    2.3434 1622.3 371.14
## + OBP     1    2.3269 1622.3 371.14
## + SLG     1    1.2318 1623.4 371.24
## + BB      1    0.5924 1624.0 371.30
## + CS      1    0.3826 1624.2 371.32
```

```
##
## Call:
## lm(formula = W ~ R + RBI + SV + IP + H, data = mydata)
##
## Coefficients:
## (Intercept)            R          RBI           SV           IP            H
##   -20.61047     -0.06799      0.08662      0.41163      0.06002     -0.00911
```

After running the forward stepwise selection function, we have the following model as its output: W = -20.61047-0.06799(R)+0.08662(RBI)+0.41163(SV)+0.06002(IP)-0.00911(H)

Backwards Stepwise

```
# Run the stepwise regression with forward selection based on the AIC criterion
step(FitAll,direction="backward", scope =formula(FitStart))
```

```
## Start:  AIC=383.56
## W ~ League + ERA + SV + IP + H + R + ER + HR + BB + SO + AVG +
##     AB + RBI + CS + OBP + SLG
```

```
##
##          Df Sum of Sq    RSS    AIC
## - SO      1      0.04 1542.3 381.56
## - BB      1      1.04 1543.3 381.66
## - CS      1      1.49 1543.8 381.70
## - SLG     1      5.16 1547.4 382.06
## - AVG     1      7.46 1549.7 382.28
## - ER      1      9.49 1551.8 382.48
## - ERA     1      9.66 1551.9 382.49
## - OBP     1      9.79 1552.1 382.51
## - H       1     13.54 1555.8 382.87
## <none>                1542.3 383.56
## - AB      1     23.93 1566.2 383.87
## - HR      1     26.56 1568.8 384.12
## - IP      1     32.27 1574.5 384.66
## - League  1     38.30 1580.6 385.24
## - R       1     46.25 1588.5 385.99
## - RBI     1    359.11 1901.4 412.96
## - SV      1    741.51 2283.8 440.44
##
## Step:  AIC=381.56
## W ~ League + ERA + SV + IP + H + R + ER + HR + BB + AVG + AB +
##     RBI + CS + OBP + SLG
##
##          Df Sum of Sq    RSS    AIC
## - BB      1      1.01 1543.3 379.66
## - CS      1      1.62 1543.9 379.72
## - SLG     1      5.52 1547.8 380.10
## - AVG     1      8.20 1550.5 380.36
## - ER      1      9.47 1551.8 380.48
## - ERA     1      9.63 1552.0 380.49
## - OBP     1      9.87 1552.2 380.52
## - H       1     14.47 1556.8 380.96
## <none>                1542.3 381.56
## - AB      1     24.25 1566.6 381.90
## - HR      1     26.85 1569.2 382.15
## - IP      1     32.54 1574.9 382.69
## - League  1     40.59 1582.9 383.46
## - R       1     46.66 1589.0 384.03
## - RBI     1    362.28 1904.6 411.21
## - SV      1    742.00 2284.3 438.48
##
## Step:  AIC=379.66
## W ~ League + ERA + SV + IP + H + R + ER + HR + AVG + AB + RBI +
##     CS + OBP + SLG
##
##          Df Sum of Sq    RSS    AIC
## - CS      1      1.82 1545.2 377.84
## - SLG     1      5.98 1549.3 378.24
## - AVG     1      7.21 1550.5 378.36
## - ER      1      8.94 1552.3 378.53
## - ERA     1      9.07 1552.4 378.54
## - OBP     1     10.62 1554.0 378.69
## - H       1     13.68 1557.0 378.98
## <none>                1543.3 379.66
## - AB      1     25.33 1568.7 380.10
## - HR      1     26.03 1569.4 380.17
## - IP      1     31.53 1574.9 380.69
## - League  1     43.53 1586.9 381.83
## - R       1     51.44 1594.8 382.58
## - RBI     1    362.27 1905.6 409.29
## - SV      1    744.18 2287.5 436.69
##
## Step:  AIC=377.84
## W ~ League + ERA + SV + IP + H + R + ER + HR + AVG + AB + RBI +
##     OBP + SLG
##
##          Df Sum of Sq    RSS    AIC
## - SLG     1      5.50 1550.7 376.37
## - AVG     1      6.25 1551.4 376.44
## - ER      1      8.60 1553.7 376.67
## - ERA     1      8.67 1553.8 376.68
## - OBP     1      9.12 1554.3 376.72
## - H       1     12.45 1557.6 377.04
## <none>                1545.2 377.84
## - AB      1     23.58 1568.7 378.11
## - HR      1     24.86 1570.0 378.23
## - IP      1     30.09 1575.2 378.73
## - League  1     42.69 1587.8 379.92
## - R       1     51.20 1596.3 380.73
## - RBI     1    383.25 1928.4 409.07
## - SV      1    745.71 2290.8 434.91
##
## Step:  AIC=376.37
## W ~ League + ERA + SV + IP + H + R + ER + HR + AVG + AB + RBI +
##     OBP
##
##          Df Sum of Sq    RSS    AIC
## - AVG     1      7.30 1558.0 375.07
## - ER      1      7.61 1558.2 375.10
## - ERA     1      7.67 1558.3 375.11
## - OBP     1      7.85 1558.5 375.13
## - H       1     13.90 1564.5 375.71
## <none>                1550.7 376.37
## - AB      1     21.04 1571.7 376.39
## - HR      1     22.02 1572.7 376.48
## - IP      1     27.45 1578.1 377.00
## - League  1     41.99 1592.6 378.38
## - R       1     50.73 1601.4 379.20
## - SV      1    869.96 2420.6 441.17
## - RBI     1   1166.47 2717.1 458.50
##
## Step:  AIC=375.07
## W ~ League + ERA + SV + IP + H + R + ER + HR + AB + RBI + OBP
##
##          Df Sum of Sq    RSS    AIC
## - ER      1      7.67 1565.6 373.81
## - ERA     1      7.90 1565.8 373.83
## - OBP     1      8.73 1566.7 373.91
## - AB      1     19.60 1577.5 374.95
## <none>                1558.0 375.07
## - IP      1     22.06 1580.0 375.18
## - HR      1     29.58 1587.5 375.89
## - League  1     34.87 1592.8 376.39
## - H       1     42.23 1600.2 377.09
## - R       1     54.77 1612.7 378.26
## - SV      1    889.71 2447.7 440.84
## - RBI     1   1167.31 2725.3 456.95
##
## Step:  AIC=373.81
## W ~ League + ERA + SV + IP + H + R + HR + AB + RBI + OBP
##
##          Df Sum of Sq    RSS    AIC
## - ERA     1      0.33 1566.0 371.84
## - OBP     1      8.49 1574.1 372.62
## - AB      1     18.13 1583.8 373.54
## <none>                1565.6 373.81
## - HR      1     28.38 1594.0 374.50
## - League  1     34.23 1599.8 375.05
## - H       1     40.96 1606.6 375.68
## - R       1     54.57 1620.2 376.95
## - IP      1     56.57 1622.2 377.14
## - SV      1    900.07 2465.7 439.94
## - RBI     1   1167.51 2733.1 455.38
##
## Step:  AIC=371.84
## W ~ League + SV + IP + H + R + HR + AB + RBI + OBP
##
##          Df Sum of Sq    RSS    AIC
## - OBP     1      8.34 1574.3 370.64
## - AB      1     18.33 1584.3 371.59
## <none>                1566.0 371.84
## - HR      1     29.32 1595.3 372.62
## - League  1     34.00 1600.0 373.06
## - H       1     40.68 1606.6 373.69
## - IP      1     78.72 1644.7 377.20
## - R       1    348.07 1914.0 399.95
## - SV      1    912.56 2478.5 438.72
## - RBI     1   1168.61 2734.6 453.46
##
## Step:  AIC=370.64
## W ~ League + SV + IP + H + R + HR + AB + RBI
##
##          Df Sum of Sq    RSS    AIC
## - AB      1     14.50 1588.8 370.01
## <none>                1574.3 370.64
## - League  1     27.84 1602.1 371.27
## - HR      1     30.12 1604.4 371.48
## - H       1     38.92 1613.2 372.30
## - IP      1     72.37 1646.7 375.38
## - R       1    377.61 1951.9 400.89
## - SV      1    927.39 2501.7 438.11
## - RBI     1   3129.93 4704.2 532.84
##
## Step:  AIC=370.01
## W ~ League + SV + IP + H + R + HR + RBI
##
##          Df Sum of Sq    RSS    AIC
## - League  1      21.1 1609.8 369.99
## <none>                1588.8 370.01
## - HR      1      22.3 1611.1 370.11
## - H       1      51.3 1640.1 372.78
## - IP      1      60.1 1648.9 373.58
## - R       1     423.7 2012.5 403.47
## - SV      1     962.3 2551.1 439.05
## - RBI     1    4210.2 5799.0 562.22
##
## Step:  AIC=369.99
## W ~ SV + IP + H + R + HR + RBI
##
##        Df Sum of Sq    RSS    AIC
## - HR    1      14.8 1624.6 369.36
## <none>              1609.8 369.99
## - H     1      41.7 1651.6 371.83
## - IP    1      53.5 1663.4 372.89
## - R     1     478.1 2088.0 407.00
## - SV    1     953.2 2563.1 437.75
## - RBI   1    4422.9 6032.7 566.15
##
## Step:  AIC=369.36
## W ~ SV + IP + H + R + RBI
##
##        Df Sum of Sq    RSS    AIC
## <none>              1624.6 369.36
## - H     1      28.7 1653.3 369.98
## - IP    1      64.7 1689.3 373.21
## - SV    1     938.5 2563.1 435.75
## - R     1    1062.3 2686.9 442.83
## - RBI   1    4683.2 6307.8 570.84
```

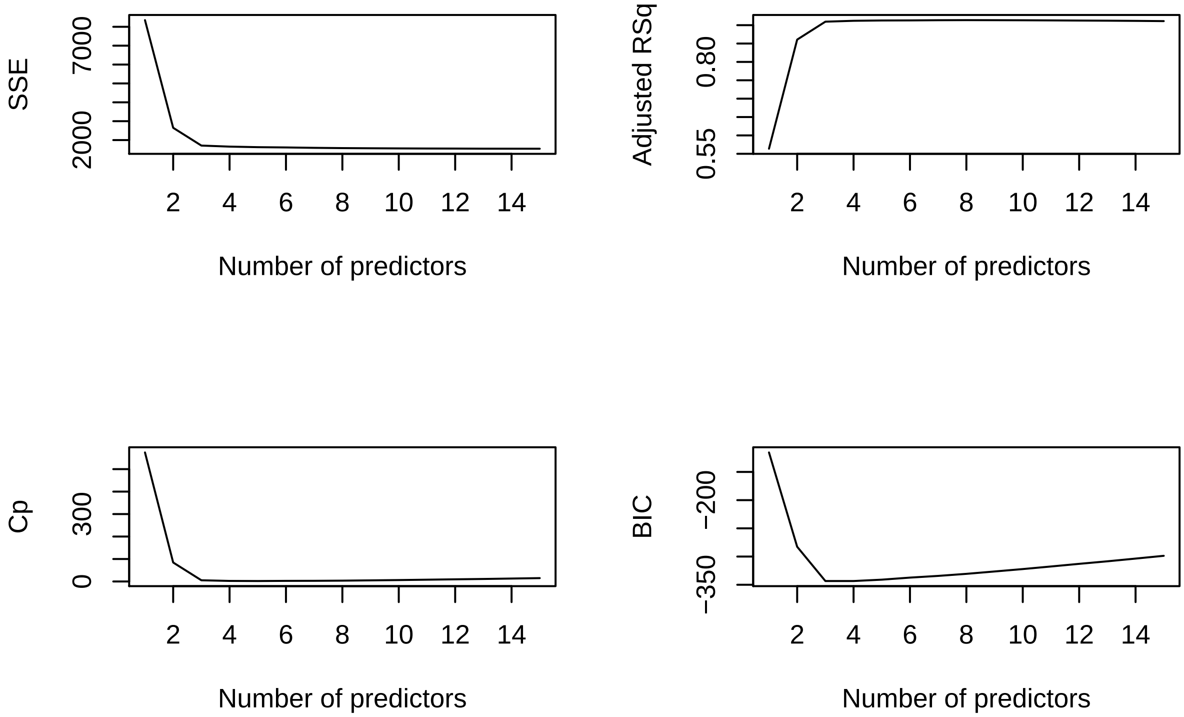
```
##
## Call:
## lm(formula = W ~ SV + IP + H + R + RBI, data = mydata)
##
## Coefficients:
## (Intercept)           SV           IP            H            R          RBI
##   -20.61047      0.41163      0.06002     -0.00911     -0.06799      0.08662
```

After running the backward stepwise selection function, we have the following model as its output: W = -20.61047-0.06799(R)+0.08662(RBI)+0.41163(SV)+0.06002(IP)-0.00911(H)

This is the same model as forward stepwise.

#Best Subsets

```r
# Find the best model for each number of predictors
models <- regsubsets(W ~ League+ERA+SV+IP+H+R+ER+HR+BB+SO+AVG+AB+RBI+CS+OBP+SLG,mydata, nvmax = 15)
models.sum <- summary(models)
# Create four plots within a 2x2 frame to compare the different criteria
par(mfrow = c(2,2))
 # SSE
 plot(models.sum$rss, xlab = "Number of predictors", ylab = "SSE", type = "l")
 # R2
 plot(models.sum$adjr2, xlab = "Number of predictors", ylab = "Adjusted RSq", type = "l")
 # Mallow's Cp
 plot(models.sum$cp, xlab = "Number of predictors", ylab = "Cp", type = "l")
 # BIC
 plot(models.sum$bic, xlab = "Number of predictors", ylab = "BIC", type = "l")
```



Since we are trying to minimize SSE, Cp, and BIC while minimizing Adj R^2, the model with 3 variables seems to be the model that achieves the most maximization while not overfitting many variables.

```r
#Displays the best model for each number of predictors
models.sum$outmat
```

```
##            LeagueNL ERA SV  IP  H   R   ER  HR  BB  SO  AVG AB  RBI CS  OBP SLG
## 1  ( 1 )  " "      " " " " " " " " "*" " " " " " " " " " " " " " " " " " " " " " "
## 2  ( 1 )  " "      " " " " " " " " "*" " " " " " " " " " " " " "*" " " " " " "
## 3  ( 1 )  " "      " " "*" " " " " "*" " " " " " " " " " " " " "*" " " " " " "
## 4  ( 1 )  " "      " " "*" "*" " " "*" " " " " " " " " " " " " "*" " " " " " "
## 5  ( 1 )  " "      " " "*" "*" "*" "*" " " " " " " " " " " " " "*" " " " " " "
## 6  ( 1 )  " "      " " "*" "*" "*" "*" " " "*" " " " " " " " " "*" " " " " " "
## 7  ( 1 )  "*"      " " "*" "*" "*" "*" " " "*" " " " " " " " " "*" " " " " " "
## 8  ( 1 )  "*"      " " "*" "*" "*" "*" " " "*" " " " " " " "*" "*" " " " " " "
## 9  ( 1 )  "*"      " " "*" "*" "*" "*" " " "*" " " " " "*" "*" "*" " " " " " "
## 10  ( 1 ) "*"      " " "*" "*" "*" "*" " " "*" " " " " "*" "*" "*" " " "*" " "
## 11  ( 1 ) "*"      " " "*" "*" "*" "*" " " "*" " " " " "*" "*" "*" " " "*" "*"
## 12  ( 1 ) "*"      "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*" " " "*" " "
## 13  ( 1 ) "*"      "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*" " " "*" "*"
## 14  ( 1 ) "*"      "*" "*" "*" "*" "*" "*" "*" " " " " "*" "*" "*" "*" "*" "*"
## 15  ( 1 ) "*"      "*" "*" "*" "*" "*" "*" "*" "*" " " "*" "*" "*" "*" "*" "*"
```

As such the model selected by best subsets has the variables R, RBI and SV.

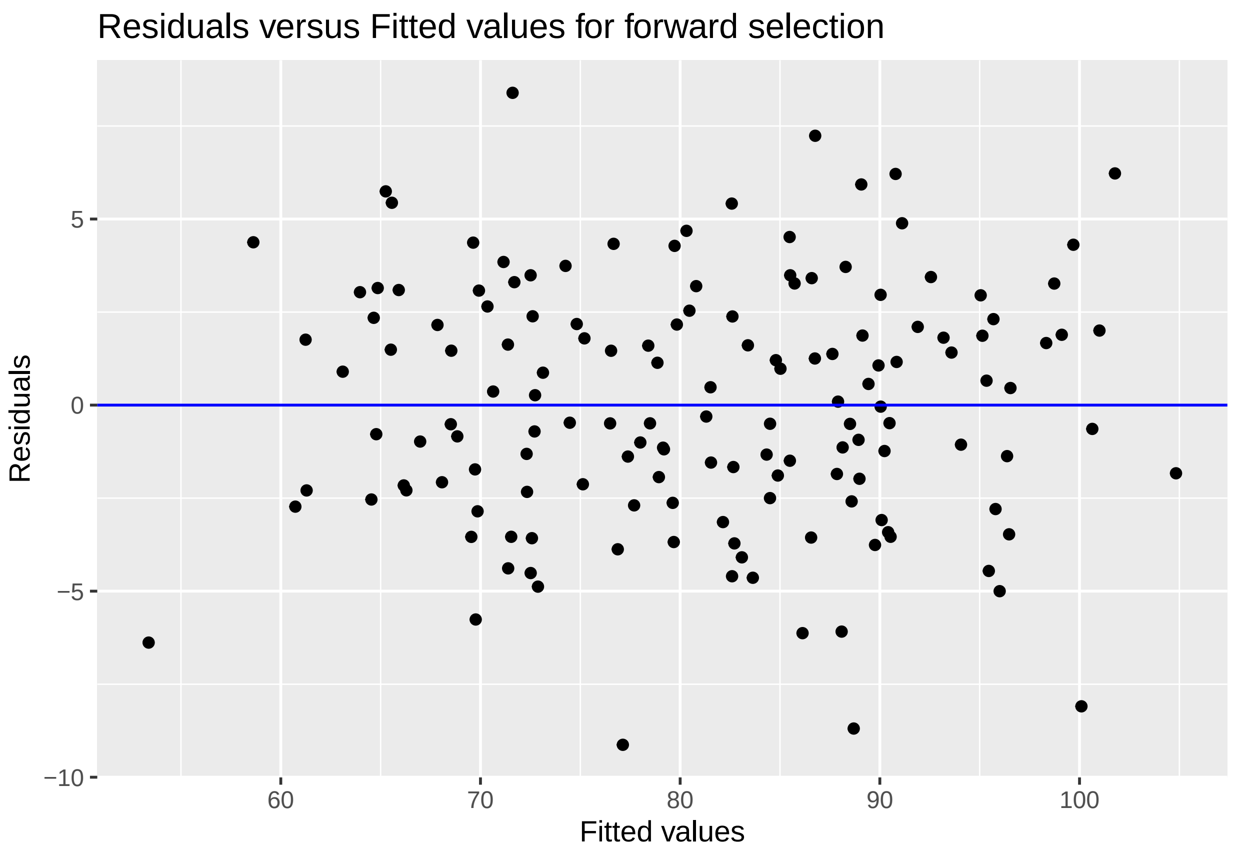
Best Subset Model: W = 61.077 - .0812(R) + .0894(RBI) + .4091(SV)

Assumptions

Since we have two models, we must evaluate them based on their assumptions.

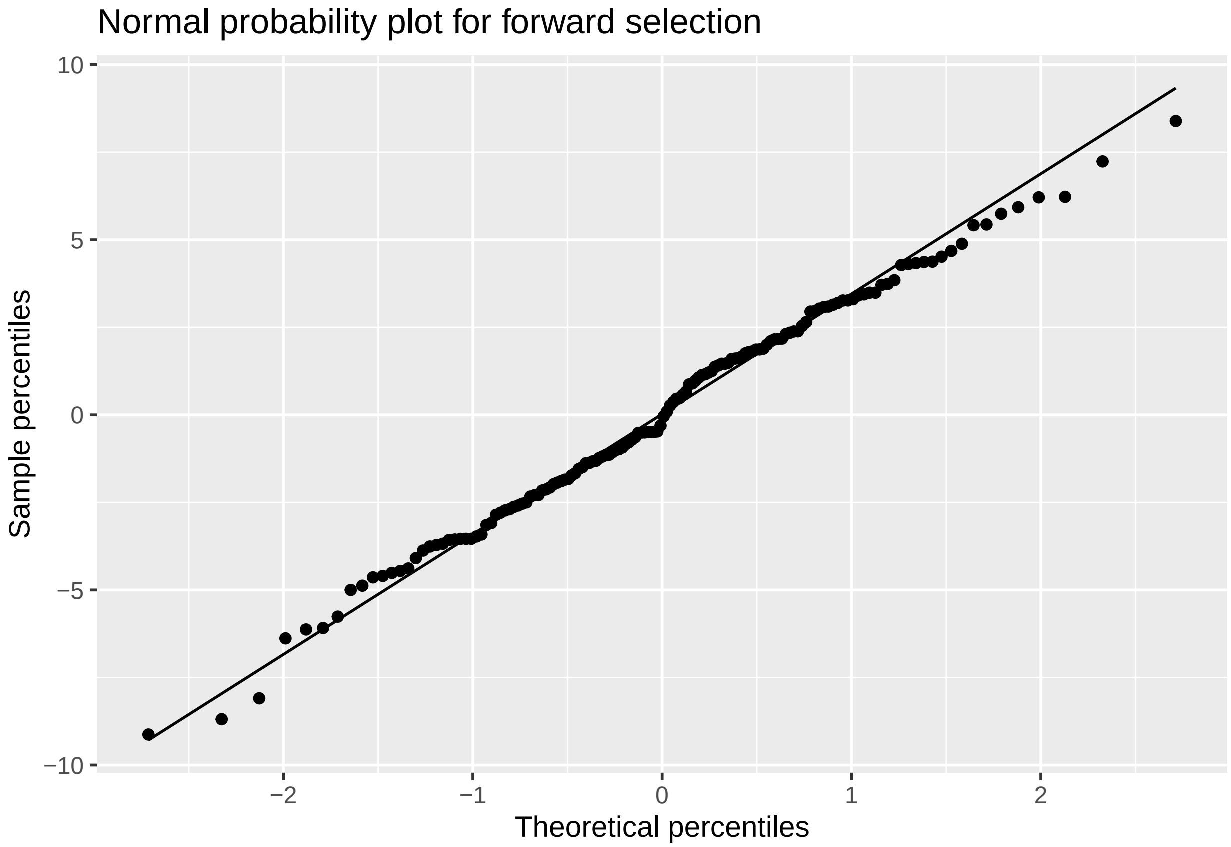
Stepwise Assumption Evaluation

```r
# Fit the model obtained from forward selection (same in this case)
mydata$resids <- residuals(stepwise)
mydata$predicted <- predict(stepwise)
ggplot(mydata, aes(x=predicted, y=resids)) + geom_point() + geom_hline(yintercept=0, color = "blue") +
 labs(title ="Residuals versus Fitted values for forward selection",
      x = "Fitted values", y = "Residuals")
```



The residuals versus fitted values shows no violation of the equal variance or linearity assumption. The residuals appear to be evenly distributed across the fitted values, there are no trends present in the residuals and there are no obvious outliers.

```
ggplot(mydata, aes(sample = resids)) + stat_qq() + stat_qq_line() +
 labs(title ="Normal probability plot for forward selection",
      x = "Theoretical percentiles", y = "Sample percentiles")
```

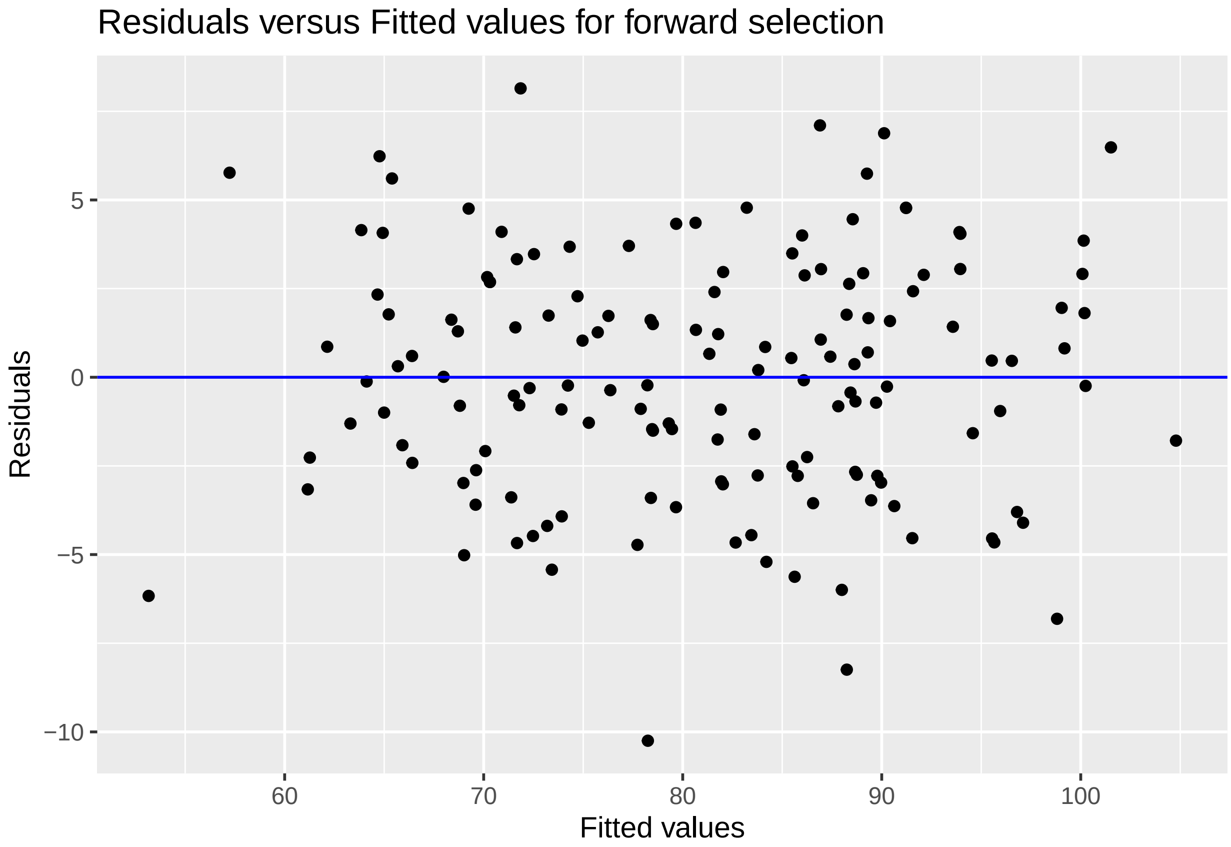


The normal probability plot shows that the normality of residuals assumption appears to not be violated. While there is some fluctuation around the line there are no significant tails or trends that veer too significantly, thus signaling that the residuals are roughly normally distributed.

Overall this model demonstrates that is has the characteristics needed to say that this model appears to fit the data. The diagnostics show no sign for alarm as all assumptions are validated.

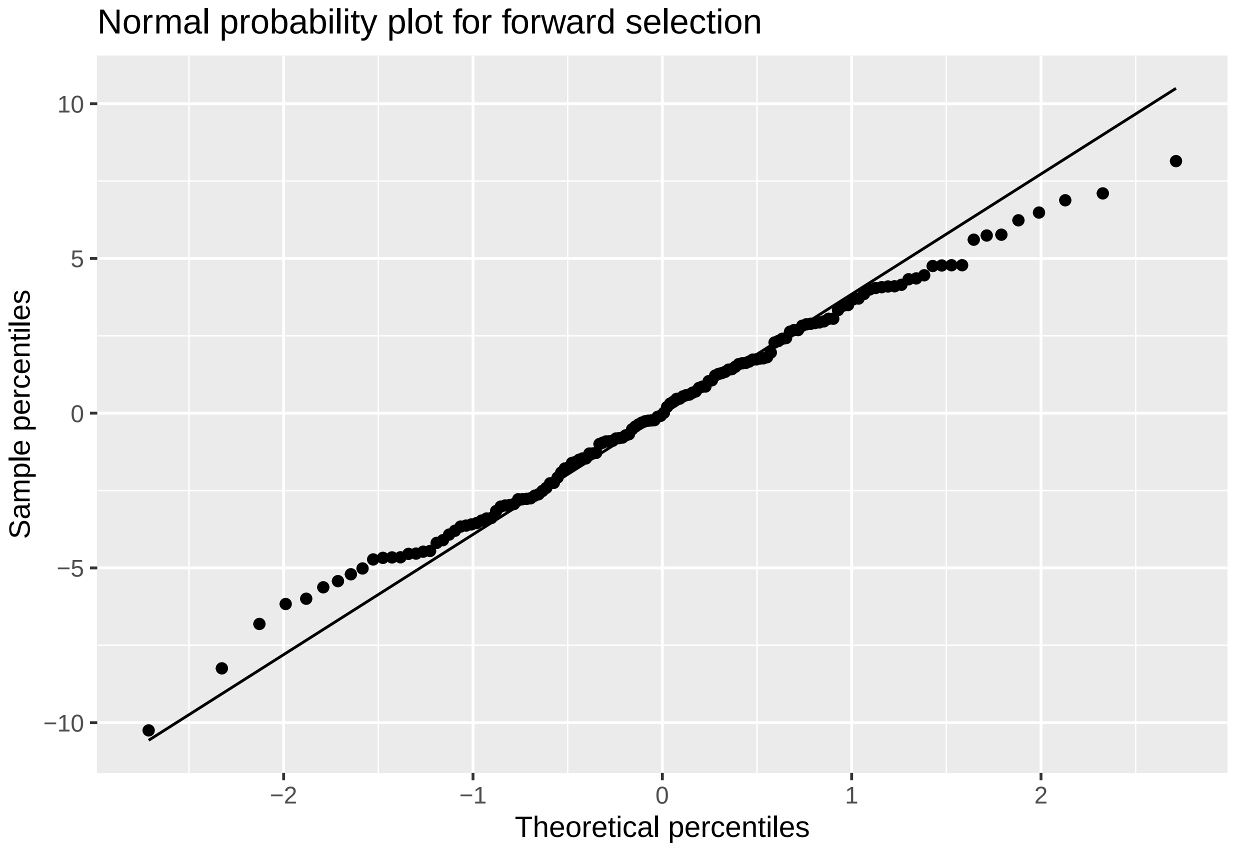
Subsets Assumption Evaluation

```
# Fit the model obtained from forward selection (same in this case)
mydata$resids <- residuals(subset)
mydata$predicted <- predict(subset)
ggplot(mydata, aes(x=predicted, y=resids)) + geom_point() + geom_hline(yintercept=0, color = "blue") +
 labs(title ="Residuals versus Fitted values for forward selection",
      x = "Fitted values", y = "Residuals")
```



The residuals versus fitted values shows no violation of the equal variance or linearity assumption. The residuals appear to be evenly distributed across the fitted values, there are no trends present in the residuals and there are no obvious outliers.

```r
ggplot(mydata, aes(sample = resids)) + stat_qq() + stat_qq_line() +
 labs(title ="Normal probability plot for forward selection",
      x = "Theoretical percentiles", y = "Sample percentiles")
```



The normal probability plot shows that the normality of residuals assumption appears to be violated. There is a slight veer off the line in both tails, albeit very minor.

This model does not violate the linearity or independence assumption but does slightly violate the normality assumption.

Model Selection

Both models have very similar diagnostics but the subset model has a worse normality plot than the stepwise plot. As such, we will select the the stepwise equation as the final model.

Final Model:

$W = -20.61047-0.06799(R)+0.08662(RBI)+0.41163(SV)+0.06002(IP)-0.00911(H)$