COLLEGE OF COMPUTIG

DEPARTMENT OF DATA SCIENCE

REGRATION PROGECT

GROUPMEMBERS NAME	**ID**
1. NATENAEL BEKELE	DBU1501407
2. HAFIZE HUSEN	DBU1501241
3. REDIET ESUBALEW	DBU1501704
4. BEREKET GETAW	DBU1501044
5. DAWIT ALEMU	DBU1501117
6. YIFERU MEKONEN	DBU1501562
7. ELBETEL ABEDI	DBU1501145
8. GENET MINDA	DBU1501217

###what is forward selection

Forward selection: is a type of stepwise regression technique used to build a predictive model by iteratively adding predictors to the model. It starts with an empty model (no predictors) and adds one predictor at a time based on a predefined criterion (e.g., p-value, AIC, or R-squared). The goal is to identify the most significant predictors that improve the model's performance.

How Forward Selection Works:

1**. Start with an empty model**: The initial model contains only the intercept (no predictors).

2. **Evaluate all possible predictors**: For each predictor not yet in the model, fit a new model by adding that predictor.

3. **Select the best predictor:** Choose the predictor that improves the model the most based on a criterion (e.g., lowest p-value, highest improvement in R-squared, or lowest AIC).

4. **Add the predictor to the model**: Include the selected predictor in the model.

5**. Repeat:** Continue adding predictors one at a time until no further improvement is achieved or a stopping criterion is met.

Advantages of Forward Selection:

1. **Simplicity**: Easy to implement and understand.
2. **Efficiency**: Works well when there are many predictors, as it reduces the number of models to evaluate.
3. **Interpretability**: Produces a simpler model with fewer predictors, making it easier to interpret.

Disadvantages of Forward Selection:

1. **Greedy Algorithm**: It selects the best predictor at each step but does not reconsider previous choices, which may lead to suboptimal models.
2. **Risk of Overfitting**: If too many predictors are added, the model may overfit the training data.
3. **Dependence on Criteria**: The results depend on the criterion used (e.g., p-value, AIC), which may not always lead to the best model.

When to Use Forward Selection:

- When you have a large number of predictors and want to identify the most important ones.

- When computational efficiency is important, as it evaluates fewer models compared to exhaustive methods like all-subsets regression.

Comparison with Other Methods:

- Backward Elimination: Starts with all predictors and removes the least significant one at each step.

- **Stepwise Regression**: Combines forward selection and backward elimination, allowing predictors to be added or removed at each step.

- **All-Subsets Regression**: Evaluates all possible combinations of predictors, which is computationally expensive but thorough.

###which criterion is best for our dataset

For our dataset (31 observations, 6 predictors), BIC (Bayesian Information Criterion) is likely the best choice because:

- It strongly penalizes model complexity, which is important for small datasets.

- It helps avoid overfitting and ensures the model generalizes well to new data.

---Here is our code--

 # Load the dataset

data <- data.frame(

y = c(6.75, 13.00, 14.75, 12.60, 8.25, 10.67, 7.28, 12.67, 12.58, 20.60, 3.58, 7.00, 26.20, 11.67, 7.67, 12.25, 0.76, 1.35, 1.44, 1.60, 1.10, 0.85, 1.20, 0.56, 0.72, 0.47, 0.33, 0.26, 0.76, 0.80, 2.00),

x1 = c(2.80, 1.40, 1.40, 3.30, 1.70, 2.90, 3.70, 1.70, 0.92, 0.68, 6.00, 4.30, 0.60, 1.80, 6.00, 4.40, 88.00, 62.00, 50.00, 58.00, 90.00, 66.00, 140.00, 240.00, 420.00, 500.00, 180.00, 270.00, 170.00, 98.00, 35.00),

x2 = c(4.68, 5.19, 4.82, 4.85, 4.86, 5.16, 4.82, 4.86, 4.78, 5.16, 4.57, 4.61, 5.07, 4.66, 5.42, 5.01, 4.97, 4.01, 4.96, 5.20, 4.80, 4.98, 5.35, 5.04, 4.80, 4.83, 4.66, 4.67, 4.72, 5.00, 4.70),

x3 = c(4.87, 4.50, 4.73, 4.76, 4.95, 4.45, 5.05, 4.70, 4.84, 4.76, 4.82, 4.65, 5.10, 5.09, 4.41, 4.74, 4.66, 4.72, 4.90, 4.70, 4.60, 4.69, 4.76, 4.80, 4.80, 4.60, 4.72, 4.50, 4.70, 5.07, 4.80),

x4 = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1),

x5 = c(8.4, 6.5, 7.9, 8.3, 8.4, 7.4, 6.8, 8.6, 6.7, 7.7, 7.4, 6.7, 7.5, 8.2, 5.8, 7.1, 6.5, 8.0, 6.8, 8.2, 6.6, 6.4, 7.3, 7.8, 7.4, 6.7, 7.2, 6.3, 6.8, 7.2, 7.7),

x6 = c(4.916, 4.563, 5.321, 4.865, 3.776, 4.397, 4.867, 4.828, 4.865, 4.034, 5.450, 4.853, 4.257, 5.144, 3.718, 4.715, 4.625, 4.977, 4.322, 5.087, 5.971, 4.647, 5.115, 5.939, 5.916, 5.471, 4.602, 5.043, 5.075, 4.334, 5.705)

)

Define the full model (with all predictors)

full_model <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6, data = data)

summary(full_model)

output:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.0548 -1.8258 -0.1374  1.6179 11.9741 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -32.663676  30.199504  -1.082    0.290    
x1            0.001131   0.008386   0.135    0.894    
x2            4.343137   3.047723   1.425    0.167    
x3            4.272295   4.745619   0.900    0.377    
x4           -9.753643   1.973300  -4.943 4.81e-05 ***
x5            0.950435   1.210822   0.785    0.440    
x6           -0.950855   1.628797  -0.584    0.565    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.055 on 24 degrees of freedom
Multiple R-squared:  0.7113,	Adjusted R-squared:  0.6391 
F-statistic: 9.853 on 6 and 24 DF,  p-value: 1.645e-05
```

Define the null model (intercept-only model)

null_model <- lm(y ~ 1, data = data)

summary(null_model)

output:

```
Call:
lm(formula = y ~ 1, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-6.247 -5.682 -2.927  5.453 19.693 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    6.507      1.212   5.368 8.24e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.749 on 30 degrees of freedom
```

 # Perform forward selection using BIC

forward_model <- step(null_model, scope = list(lower = null_model, upper = full_model), direction = "forward", k = log(nrow(data)))

output:

```
Start:  AIC=120.8
y ~ 1

       Df Sum of Sq     RSS     AIC
+ x4    1    898.57  467.95  91.014
+ x1    1    424.93  941.59 112.689
+ x6    1    245.58 1120.94 118.094
<none>              1366.52 120.801
+ x5    1     99.00 1267.52 121.904
+ x2    1     80.23 1286.29 122.359
+ x3    1     60.38 1306.14 122.834

Step:  AIC=91.01
y ~ x4

       Df Sum of Sq    RSS    AIC
<none>              467.95 91.014
+ x2    1   30.5099 437.44 92.358
+ x3    1   17.7425 450.21 93.250
+ x6    1   14.7853 453.17 93.452
+ x5    1    9.1078 458.85 93.838
+ x1    1    1.9125 466.04 94.321
```

Interpretation of Forward Selection Using BIC

Step 1: Understanding the Forward Selection Process

- The selection starts with the **null model** (intercept-only).
- Variables are added one by one based on the **BIC criterion** (which is approximated here using AIC with k=log(n)).
- The variable with the largest reduction in **Residual Sum of Squares (RSS)** and lowest AIC/BIC is selected at each step.

Step 2: First Selection Step

Starting Model: y~1

```
Start:  AIC=120.8
y ~ 1

       Df Sum of Sq     RSS     AIC
+ x4    1    898.57  467.95  91.014
+ x1    1    424.93  941.59 112.689
+ x6    1    245.58 1120.94 118.094
<none>              1366.52 120.801
+ x5    1     99.00 1267.52 121.904
+ x2    1     80.23 1286.29 122.359
+ x3    1     60.38 1306.14 122.834
```

Explanation:

- **Best predictor to add:** x4 (since it gives the lowest AIC = 91.01).
- The improvement in RSS from 1366.52 to 467.95 is significant.

Other variables were considered, but they resulted in **higher AIC values**, meaning they were less optimal in reducing model complexity while improving fit.

Step 3: Second Selection Step

Current Model: y~x4

```
Step:  AIC=91.01
y ~ x4

       Df Sum of Sq    RSS    AIC
<none>              467.95 91.014
+ x2    1   30.5099 437.44 92.358
+ x3    1   17.7425 450.21 93.250
+ x6    1   14.7853 453.17 93.452
+ x5    1    9.1078 458.85 93.838
+ x1    1    1.9125 466.04 94.321
```

Explanation:

- The algorithm checks whether adding another predictor would further reduce the AIC/BIC significantly.
- None of the remaining predictors **(x1, x2, x3, x5, x6**) provide a large enough improvement.
    - The lowest AIC among them is **x2 (AIC = 92.36),** which is higher than **91.01**, meaning it's not worth adding.
    - Since no additional variable significantly improves the model, the algorithm **stops here**.

Step 4: Interpretation of the Final Model

 # Summarize the final model

summary(forward_model)

output:

```
Call:
lm(formula = y ~ x4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1400 -0.6517 -0.0967  0.7567 14.4800

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.720      1.004  11.670 1.78e-12 ***
x4           -10.773      1.444  -7.462 3.18e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.017 on 29 degrees of freedom
Multiple R-squared:  0.6576,    Adjusted R-squared:  0.6457
F-statistic: 55.69 on 1 and 29 DF,  p-value: 3.181e-08
```

Final Selected Model:

$$y=11.72-10.77x4$$

- **Intercept (11.72):** When x4=0, the predicted value of yyy is **11.72**.
- **Effect of x4 (-10.77):**
  - If x4=1, the predicted y drops to **0.95** (a strong negative effect).
  - This suggests $\mathbf{x_4}$ is a key factor influencing **y**.

Step 5: Why Were Other Variables Excluded?

- **BIC** prefers **simplicity** and penalizes adding more variables unless they significantly improve fit.
- None of the other variables provided enough reduction in **AIC/BIC** to be worth adding.
- $\mathbf{x_4}$ alone explains a significant portion of the variance in **y**.

Step 6: Conclusion

- $\mathbf{x_4}$ is the most significant predictor of **y.**
- Other variables do not improve model fit enough to justify inclusion.
- The final model balances simplicity and predictive power, avoiding overfitting.

- **Final Decision: The best model**

  $y = 11.72 - 10.77\ x_4$