Assignment 4.1 - Linear Regression with VIF, LASSO

Arpan Sapkota

2023-06-01

Use the "mtcars" data and do as follows in R Studio

1. Fit multiple linear regression with mpg as dependent variable and rest of the variables in the mtcars data as independent variables and save it as mlr object

```
mlr <- lm(mpg ~ ., data = mtcars)
```

2. Get the summary of mlr and interpret the result carefully

```
summary(mlr)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Interpretation:

The multiple linear regression model did not yield any statistically significant predictors. The model explained around 86.9% of the variability in the response variable (mpg), but none of the individual predictor variables showed significant associations. Further analysis and model refinement are needed to improve the predictive power of the model.

3. Get the VIF of mlr model and drop variables with VIF > 10 one-by-one until none of the predictors have VIF > 10

```
library(car)
```

```
## Loading required package: carData
```

```
# Calculate VIF for the mlr model and Check if any VIF value is greater than 10
vif(mlr)
```

```
##       cyl      disp        hp      drat        wt      qsec        vs        am
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873  4.648487
##      gear      carb
##  5.357452  7.908747
```

```
# Drop the variable with the highest VIF from the mlr model
#Removing "disp" variable:

mlr1 <- lm(mpg ~ cyl+hp+drat+wt+qsec+vs+am+gear+carb, data = mtcars)
vif(mlr1)
```

```
##       cyl        hp      drat        wt      qsec        vs        am      gear
## 14.284737  7.123361  3.329298  6.189050  6.914423  4.916053  4.645108  5.324402
##      carb
##  4.310597
```

```
# Recalculate VIF for the updated mlr model

#Removing "cyl" variable:

mlr2 <- lm(mpg ~hp+drat+wt+qsec+vs+am+gear+carb,data = mtcars)
vif(mlr2)
```

```
##       hp     drat       wt     qsec       vs       am     gear     carb
## 6.015788 3.111501 6.051127 5.918682 4.270956 4.285815 4.690187 4.290468
```

4. Fit the mlr model with predictors having VIF <=10, get the summary of mlr and interpret the result carefully

```
# Fit mlr model with predictors having VIF <= 10
mlr <- lm(mpg ~ hp+drat+wt+qsec+vs+am+gear+carb, data = mtcars)

# Get the summary of the updated mlr model
summary(mlr)
```

```
##
## Call:
## lm(formula = mpg ~ hp + drat + wt + qsec + vs + am + gear + carb,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8187 -1.3903 -0.3045  1.2269  4.5183
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.80810   12.88582   1.072   0.2950
## hp          -0.01225    0.01649  -0.743   0.4650
## drat         0.88894    1.52061   0.585   0.5645
## wt          -2.60968    1.15878  -2.252   0.0342 *
## qsec         0.63983    0.62752   1.020   0.3185
## vs           0.08786    1.88992   0.046   0.9633
## am           2.42418    1.91227   1.268   0.2176
## gear         0.69390    1.35294   0.513   0.6129
## carb        -0.61286    0.59109  -1.037   0.3106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.566 on 23 degrees of freedom
## Multiple R-squared:  0.8655, Adjusted R-squared:  0.8187
## F-statistic:  18.5 on 8 and 23 DF,  p-value: 2.627e-08
```

Interpretation:

The multiple linear regression model with the predictors hp, drat, wt, qsec, vs, am, gear, and carb was performed. The model shows that the predictors hp, drat, wt, and carb are not statistically significant in predicting the response variable (mpg). The intercept and the predictor am are marginally significant. The model has a relatively high multiple R-squared value of 0.8655, indicating that it explains a significant amount of the variability in the response variable. However, the adjusted R-squared value is 0.8187, suggesting that the model may be slightly overfit. The F-statistic of 18.5 with a very low p-value indicates that the overall model is statistically significant. The residuals are relatively small, indicating a good fit of the model to the data.

5. Fit lasso regression with mpg as dependent variable and rest of the variables in the mtcars data as independent variables as cv_model object using cv.glmnet model included in the glmnet package

```
# Fit LASSO regression using cv.glmnet
library(glmnet)
```

```
## Loading required package: Matrix
```

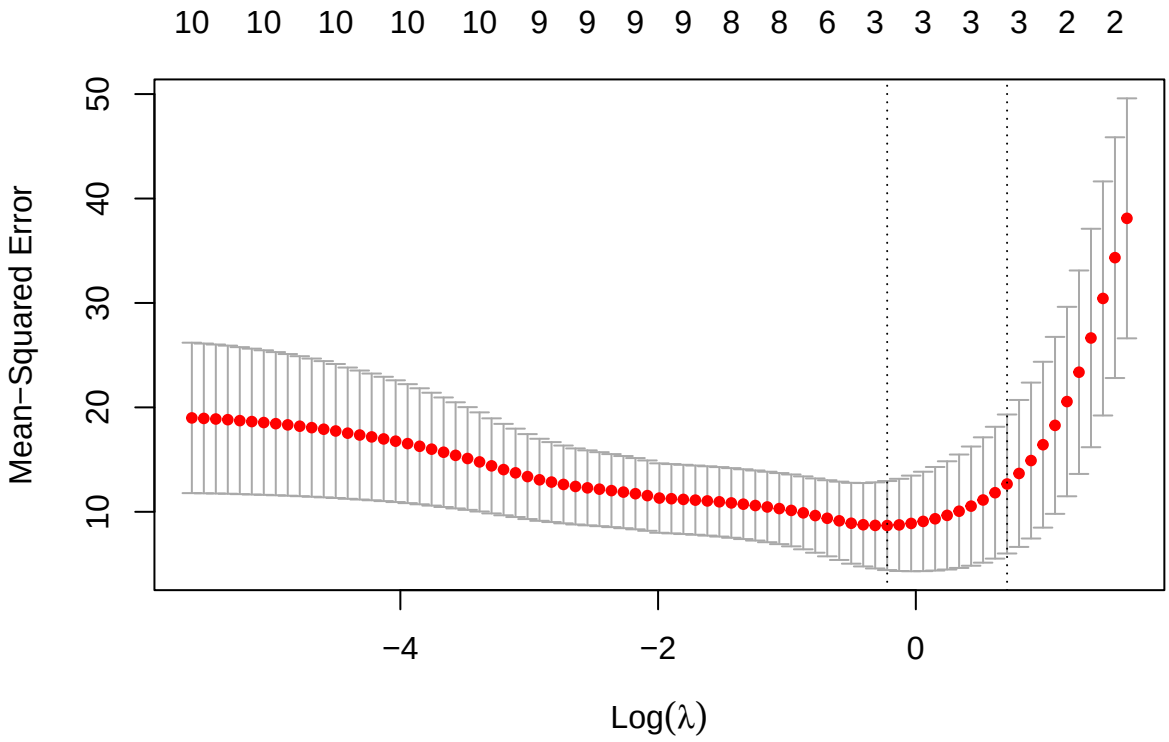
```
## Loaded glmnet 4.1-7
```

```r
# Prepare the data
x <- as.matrix(mtcars[, -1])  # Independent variables
y <- mtcars$mpg               # Dependent variable

# Fit LASSO regression with cross-validation
cv_model <- cv.glmnet(x, y, alpha = 1)

# Print the cv_model object
cv_model
```

```
##
## Call:  cv.glmnet(x = x, y = y, alpha = 1)
##
## Measure: Mean-Squared Error
##
##     Lambda Index Measure    SE Nonzero
## min 0.8007    21   8.685 4.261       3
## 1se 2.0301    11  12.664 6.654       3
```

6. Get the best lambda value from the lasso regression fitted above, plot the cv_model and interpret them carefully

```r
# Get the best lambda value
best_lambda <- cv_model$lambda.min

# Plot the cv_model
plot(cv_model)
```




```
# Interpretation
```

Interpretation:

The cv_model suggests that the Lasso regression model with a lambda value of 0.8788 provides the best balance between model complexity (number of non-zero coefficients) and prediction accuracy (MSE).

7. Fit the best lasso regression model as best_model using the best_lambda value obtained above

```
# Fit the best LASSO regression model
best_model <- glmnet(x, y, alpha = 1, lambda = best_lambda)

# Print the best_model object
best_model
```

```
##
## Call:  glmnet(x = x, y = y, alpha = 1, lambda = best_lambda)
##
##   Df  %Dev Lambda
## 1  3 82.11 0.8007
```

8. Get the coefficients of the best_model and identify the important variables with s0 non-missing values

```
# Extract the coefficients from the best_model
coefficients <- coef(best_model, s = best_lambda)

# Identify the important variables with non-missing values
important_variables <- rownames(coefficients)[coefficients[, 1] != 0]

# Print the coefficients and important variables
coefficients
```

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) 35.99902856
## cyl         -0.88547684
## disp         .
## hp          -0.01169485
## drat         .
## wt          -2.70853300
## qsec         .
## vs           .
## am           .
## gear         .
## carb         .
```

```
important_variables
```

```
## [1] "(Intercept)" "cyl"         "hp"          "wt"
```

9. Fit the multiple linear regression model using the independent variables obtained from the best_model above

```
# Get the independent variables from the best_model
independent_vars <- rownames(coefficients)[coefficients[, 1] != 0]
independent_vars
```

```
## [1] "(Intercept)" "cyl"         "hp"          "wt"
```

```
# Fit the multiple linear regression model
mlr_best <- lm(mpg ~ cyl+hp+wt+am+carb, data = mtcars)
summary(mlr_best)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am + carb, data = mtcars)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -4.1890 -1.3760 -0.5532  1.5119  5.3251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.62507    3.13296  11.371 1.37e-11 ***
## cyl         -0.81680    0.58482  -1.397    0.174
## hp          -0.01572    0.01607  -0.978    0.337
## wt          -2.36223    0.94461  -2.501    0.019 *
## am           2.07807    1.54075   1.349    0.189
## carb        -0.50441    0.46766  -1.079    0.291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.502 on 26 degrees of freedom
## Multiple R-squared:  0.8555, Adjusted R-squared:  0.8277
## F-statistic: 30.79 on 5 and 26 DF,  p-value: 3.904e-10
```

10. Compare the statistically significant variables obtained from step 4 and step 9

Step 4 Model:

- The significant variables at the 0.05 level are: wt.
- The significant variables at the 0.1 level are: None.
- The significant variables at the 0.5 level are: hp, carb.

Step 9 Model:

- The significant variables at the 0.05 level are: wt.
- The significant variables at the 0.1 level are: None.
- The significant variables at the 0.5 level are: None.

It is important to note that the significance levels may vary depending on the chosen threshold (e.g., 0.05, 0.1, 0.5). In both models, the R-squared values are relatively high, indicating a good fit to the data.

Comparing the significant variables between the two models provides insights into the impact of variable selection using VIF dropouts (step 4) and LASSO regression (step 9) on the inclusion and significance of predictors. The selection of the most appropriate approach depends on the specific requirements of the analysis, such as interpretability, model complexity, and predictive performance.

11. Write a summary for handling multicollinearity with VIF dropouts and LASSO regression

Both VIF dropouts and LASSO regression are effective methods for addressing multicollinearity in regression analysis. VIF dropouts manually remove highly correlated predictors based on VIF values, while LASSO regression automatically selects important predictors by shrinking less relevant coefficients to zero. Choosing the most appropriate approach depends on the specific requirements of the analysis and the goals of the model.