Class Activity 11

Courtney Hodge

2024-08-02

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.3     ✔ readr     2.1.4
## ✔ forcats   1.0.0     ✔ stringr   1.5.0
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.0
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

```
life_data <- read.csv("C:\\Users\\hodge\\Desktop\\UVA_Coding_Folder\\Statistics-6021\\expectanc
y.csv")
```

```
df <-select(life_data,Life.expectancy, Status, Adult.Mortality,
infant.deaths,HIV.AIDS,BMI, GDP,Schooling)%>%
na.omit()
```

1

```
model <- lm(Life.expectancy~., data = df)
#summary(model)

aic <- MASS::stepAIC(model, direction = "both", Trace = F)
```

```
## Start:  AIC=367.69
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths + 
##     HIV.AIDS + BMI + GDP + Schooling
## 
##                   Df Sum of Sq    RSS    AIC
## - BMI              1      0.21 1550.8 365.71
## - infant.deaths    1      3.80 1554.3 366.06
## - GDP              1     16.53 1567.1 367.29
## <none>                         1550.5 367.69
## - Status           1     25.11 1575.7 368.12
## - HIV.AIDS         1    154.47 1705.0 380.03
## - Adult.Mortality  1    572.31 2122.9 413.13
## - Schooling        1   1046.90 2597.5 443.60
## 
## Step:  AIC=365.71
## Life.expectancy ~ Status + Adult.Mortality + infant.deaths + 
##     HIV.AIDS + GDP + Schooling
## 
##                   Df Sum of Sq    RSS    AIC
## - infant.deaths    1      4.00 1554.8 364.10
## - GDP              1     17.17 1567.9 365.38
## <none>                         1550.8 365.71
## - Status           1     24.94 1575.7 366.12
## + BMI              1      0.21 1550.5 367.69
## - HIV.AIDS         1    154.42 1705.2 378.05
## - Adult.Mortality  1    577.89 2128.7 411.54
## - Schooling        1   1324.88 2875.6 456.96
## 
## Step:  AIC=364.1
## Life.expectancy ~ Status + Adult.Mortality + HIV.AIDS + GDP + 
##     Schooling
## 
##                   Df Sum of Sq    RSS    AIC
## - GDP              1     17.32 1572.1 363.77
## <none>                         1554.8 364.10
## - Status           1     24.52 1579.3 364.46
## + infant.deaths    1      4.00 1550.8 365.71
## + BMI              1      0.42 1554.3 366.06
## - HIV.AIDS         1    152.02 1706.8 376.19
## - Adult.Mortality  1    591.69 2146.5 410.80
## - Schooling        1   1378.83 2933.6 457.97
## 
## Step:  AIC=363.77
## Life.expectancy ~ Status + Adult.Mortality + HIV.AIDS + Schooling
## 
##                   Df Sum of Sq    RSS    AIC
## <none>                         1572.1 363.77
## + GDP              1     17.32 1554.8 364.10
## - Status           1     31.21 1603.3 364.74
## + infant.deaths    1      4.15 1567.9 365.38
## + BMI              1      1.23 1570.8 365.66
## - HIV.AIDS         1    146.69 1718.8 375.24
## - Adult.Mortality  1    630.36 2202.4 412.69
## - Schooling        1   1553.18 3125.3 465.53
```

```
summary(aic)
```

```
## 
## Call:
## lm(formula = Life.expectancy ~ Status + Adult.Mortality + HIV.AIDS + 
##     Schooling, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.9300 -2.0243  0.3127  2.1598 10.3146 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      57.871590   2.304972  25.107  < 2e-16 ***
## StatusDeveloping -1.443373   0.847760  -1.703 0.090776 .  
## Adult.Mortality  -0.029506   0.003856  -7.651 2.48e-12 ***
## HIV.AIDS         -0.912691   0.247281  -3.691 0.000315 ***
## Schooling         1.536868   0.127964  12.010  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.281 on 146 degrees of freedom
## Multiple R-squared:  0.8418, Adjusted R-squared:  0.8375 
## F-statistic: 194.2 on 4 and 146 DF,  p-value: < 2.2e-16
```

> Since the p-val for StatusDeveloping is < 0.05, we are going to run the model again without StatusDeveloping.

```
model2 <- lm(Life.expectancy~.-Status, data = df)
```

```
aic2 <- MASS::stepAIC(model2, direction = "both", Trace = F)
```

```
## Start:  AIC=368.12
## Life.expectancy ~ (Status + Adult.Mortality + infant.deaths + 
##     HIV.AIDS + BMI + GDP + Schooling) - Status
## 
##                   Df Sum of Sq    RSS    AIC
## - BMI              1      0.04 1575.7 366.12
## - infant.deaths    1      3.62 1579.3 366.46
## <none>                         1575.7 368.12
## - GDP              1     23.87 1599.5 368.39
## - HIV.AIDS         1    144.34 1720.0 379.35
## - Adult.Mortality  1    601.49 2177.2 414.94
## - Schooling        1   1519.74 3095.4 468.08
## 
## Step:  AIC=366.12
## Life.expectancy ~ Adult.Mortality + infant.deaths + HIV.AIDS + 
##     GDP + Schooling
## 
##                   Df Sum of Sq    RSS    AIC
## - infant.deaths    1      3.58 1579.3 364.46
## <none>                         1575.7 366.12
## - GDP              1     23.91 1599.6 366.40
## + BMI              1      0.04 1575.7 368.12
## - HIV.AIDS         1    144.32 1720.0 377.35
## - Adult.Mortality  1    603.74 2179.4 413.10
## - Schooling        1   1875.99 3451.7 482.53
## 
## Step:  AIC=364.46
## Life.expectancy ~ Adult.Mortality + HIV.AIDS + GDP + Schooling
## 
##                   Df Sum of Sq    RSS    AIC
## <none>                         1579.3 364.46
## - GDP              1     24.02 1603.3 364.74
## + infant.deaths    1      3.58 1575.7 366.12
## + BMI              1      0.00 1579.3 366.46
## - HIV.AIDS         1    142.18 1721.5 375.48
## - Adult.Mortality  1    617.15 2196.4 412.27
## - Schooling        1   1951.60 3530.9 483.96
```

```
summary(aic2)
```

```
## 
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + HIV.AIDS + GDP + 
##     Schooling, data = df)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -10.0225  -1.8229   0.3517   1.8076  10.3422 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      5.574e+01  1.774e+00  31.419  < 2e-16 ***
## Adult.Mortality -2.935e-02  3.886e-03  -7.553 4.26e-12 ***
## HIV.AIDS        -8.962e-01  2.472e-01  -3.625 0.000398 ***
## GDP              3.846e-05  2.581e-05   1.490 0.138367    
## Schooling        1.586e+00  1.180e-01  13.432  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.289 on 146 degrees of freedom
## Multiple R-squared:  0.8411, Adjusted R-squared:  0.8367 
## F-statistic: 193.2 on 4 and 146 DF,  p-value: < 2.2e-16
```

> based on the step aic result above, a “good” model would be

```
model2 <- lm(Life.expectancy~Adult.Mortality + HIV.AIDS + GDP + 
    Schooling, data = df)
```

> the adjusted R^2 of our model is 0.8367

```
car::vif(model2)
```

```
## Adult.Mortality        HIV.AIDS             GDP       Schooling 
##        1.977966        1.731472        1.282495        1.564588
```

> since the VIFs for each predictor is under 10, we can feel good about this model.

2

a

```
library(glmnet)
```

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

```
## 
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
## 
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-8
```

```
design_matrix <- model.matrix(Life.expectancy~0+., data = df)
#View(design_matrix)

response_var <- df$Life.expectancy

ridgemodel <- glmnet(x = design_matrix, y = response_var, alpha = 0) #specifies that we are doing Ridge Regression!!!!!
```

b

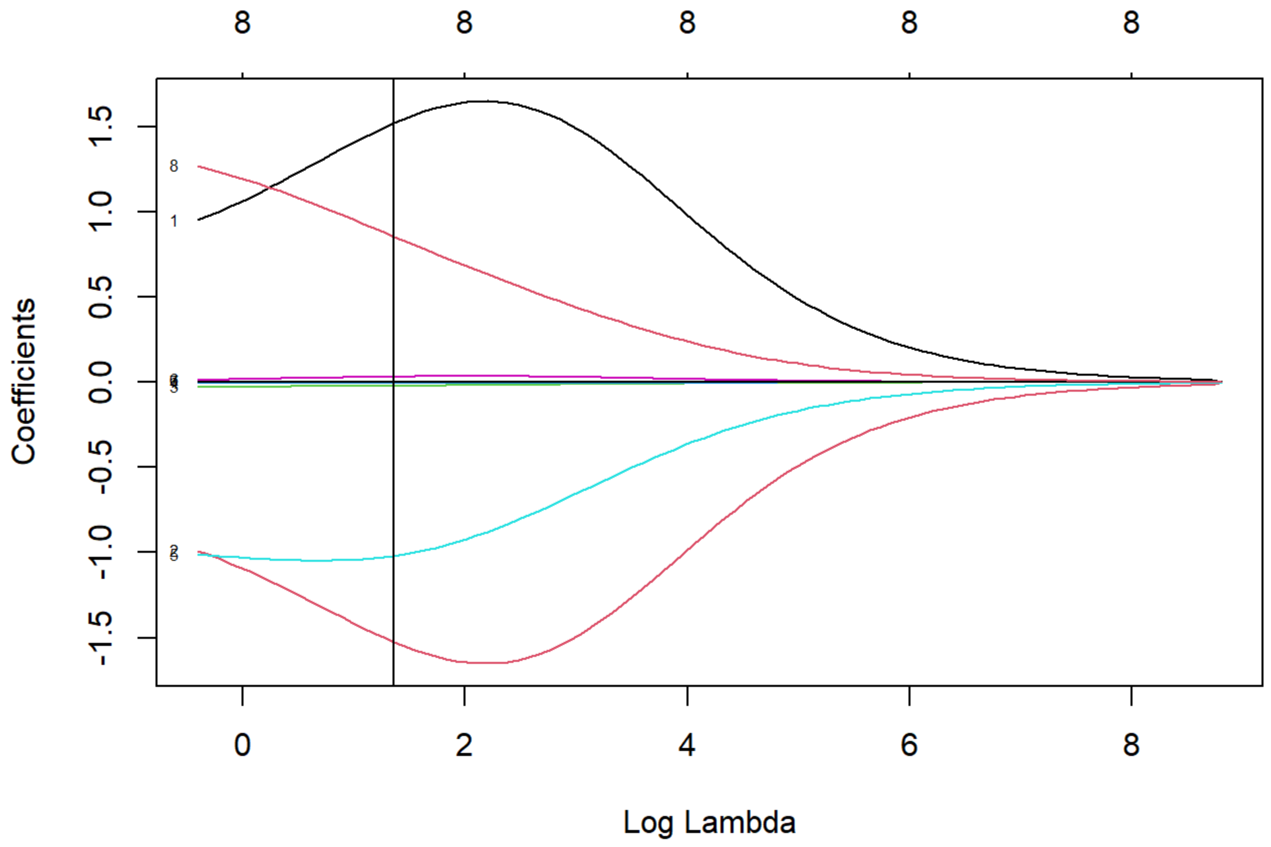
```
kcvglmnet <- cv.glmnet(x = design_matrix, y = response_var, alpha = 0, nfolds = 10) #typically, you want to do more than 2

kcvglmnet$lambda.1se
```

```
## [1] 3.903553
```

c

```
plot(ridgemodel, label = T, xvar = "lambda") + abline(v = log(kcvglmnet$lambda.1se))
```



```
## integer(0)
```

2d

> Compared to my model in Question 1, my ridge regression model found that the status (developed and devoping), HIV.AIDS, and schooling predictors were best for predicting Life.expectancy. In Question 1, the linear model found with step aic that Adult.Mortality + HIV.AIDS + GDP + Schooling were best, instead choosing Adult.Mortality and GDP over status.

3

a

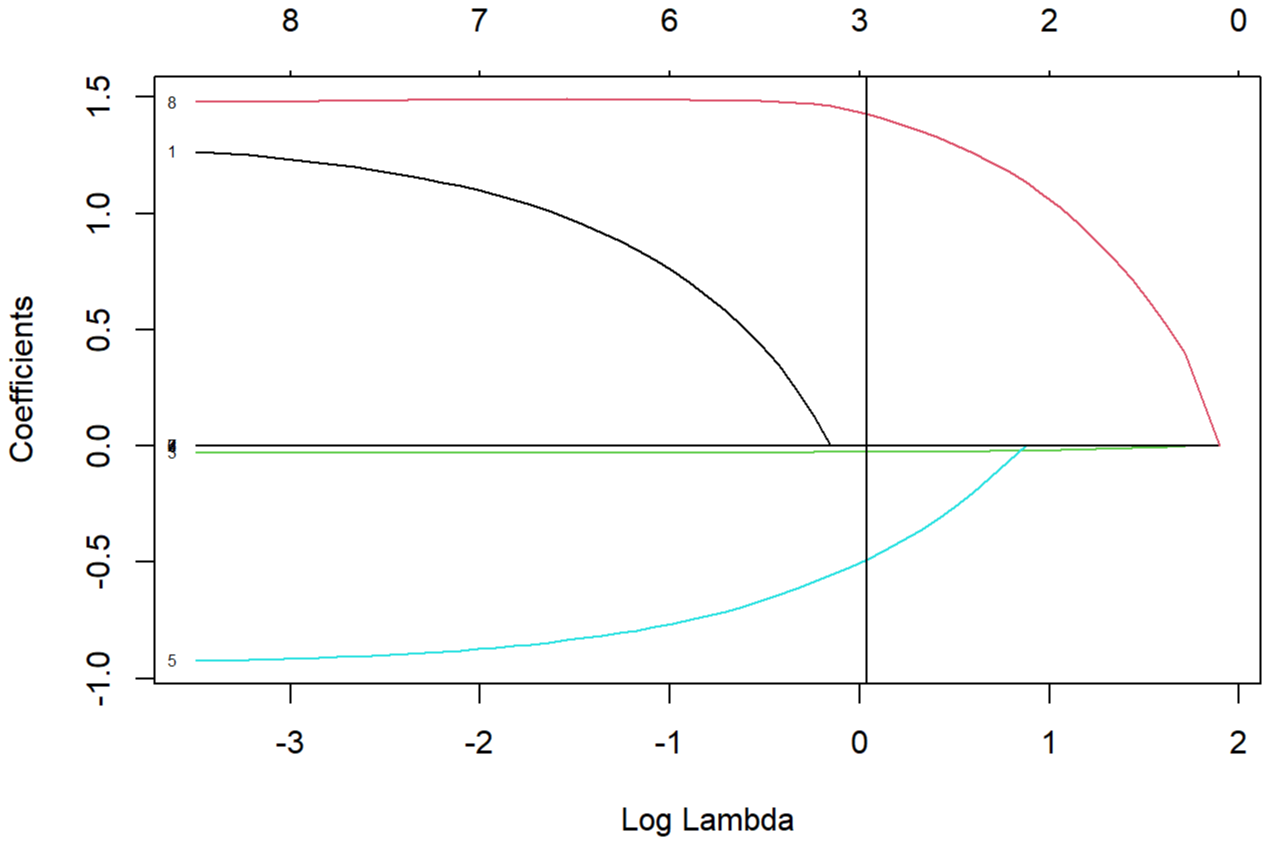
```
lassomodel <- glmnet(x = design_matrix, y = response_var, alpha = 1) #lass regression model
```

b

```
kcvglmnet <- cv.glmnet(x = design_matrix, y = response_var, alpha = 1, nfolds = 10)
```

c

```
plot(lassomodel, label = T, xvar = "lambda") + abline(v = log(kcvglmnet$lambda.1se))
```



```
## integer(0)
```

d

> the lasso model picked HIV.AIDS, and Schooling as the predictors for predicting Life.expectancy.