M14 Activity

Hannah Valenty

2024-08-02

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.1     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ 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 become errors
```

```
df <- read.csv('data/expectancy.csv')
df2 <- select(df, Life.expectancy, Status, Adult.Mortality, infant.deaths,HIV.AIDS,BMI, GDP,Schooling) %>%
na.omit()
head(df2)
```

```
##   Life.expectancy     Status Adult.Mortality infant.deaths HIV.AIDS  BMI
## 1            65.0 Developing             263            62      0.1 19.1
## 2            77.8 Developing              74             0      0.1 58.0
## 3            75.6 Developing              19            21      0.1 59.5
## 4            52.4 Developing             335            66      1.9 23.3
## 5            76.4 Developing              13             0      0.2 47.7
## 6            76.3 Developing             116             8      0.1 62.8
##          GDP Schooling
## 1   584.2592      10.1
## 2  3954.2278      14.2
## 3  4132.7629      14.4
## 4  3695.7937      11.4
## 5 13566.9541      13.9
## 6 13467.1236      17.3
```

Task 1

```
# Model from previous activity
mod1 <- lm(Life.expectancy~Adult.Mortality+HIV.AIDS+Schooling, data = df2)
summary(mod1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Adult.Mortality + HIV.AIDS + Schooling,
##     data = df2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9228 -1.8432  0.1605  2.0037 10.2611
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     55.310506   1.757712  31.467  < 2e-16 ***
## Adult.Mortality -0.030206   0.003859  -7.827 9.03e-13 ***
## HIV.AIDS        -0.870038   0.247593  -3.514 0.000587 ***
## Schooling        1.647743   0.110863  14.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.303 on 147 degrees of freedom
## Multiple R-squared:  0.8387, Adjusted R-squared:  0.8354
## F-statistic: 254.7 on 3 and 147 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value of this model is 0.8354.

Task 2.A – Ridge Regression

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.4.1
```

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

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

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

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

```
X <- model.matrix(Life.expectancy~0+., data=df2)
y <- df2$Life.expectancy
# alpha picks either squared penalty or abs penalty
# model
rmod <- glmnet(x=X,y=y, alpha=0) # ridge
# cross validate within mode
rmodcv <- cv.glmnet(x=X,y=y, alpha=0, nfolds=10, set.seed(1)) # ridge
```

Task 2.B

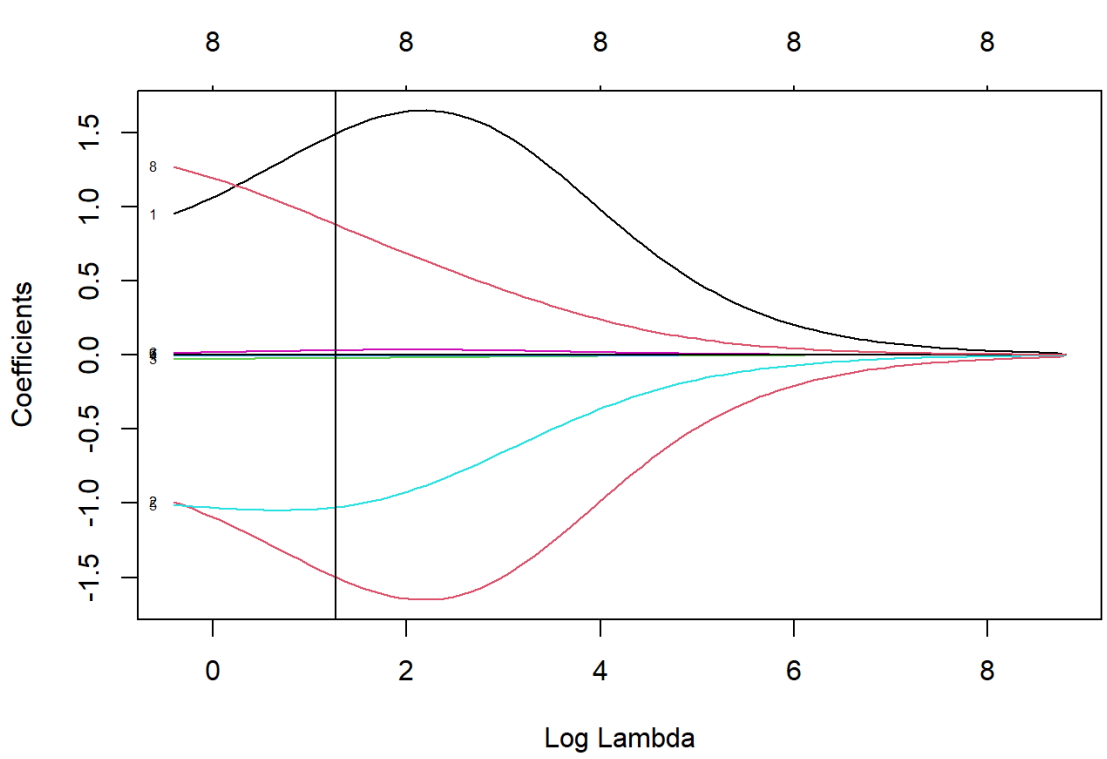
```
rmodcv$lambda.1se
```

```
## [1] 3.556772
```

The one standard error lambda for a 10 fold cross validation is 3.556772 for this model.

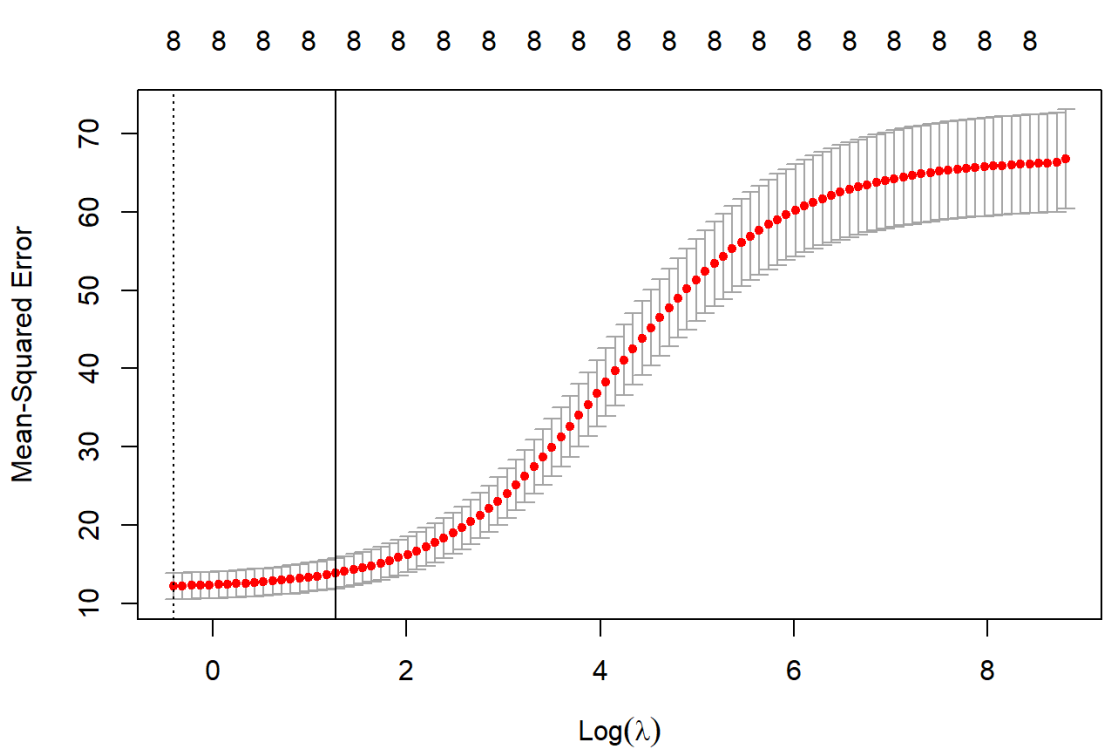
Task 2.C

```
# model plot
plot(rmod, label=T, xvar='lambda')+abline(v=log(rmodcv$lambda.1se))
```




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

```
# CV model plot
plot(rmodcv)+abline(v=log(rmodcv$lambda.1se))
```




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

Task 2.D

```
rmodfit <- glmnet(x=X,y=y, alpha=0, lambda=rmodcv$lambda.min)
rmodfit$dev.ratio
```

```
## [1] 0.8412157
```

Compared to question 1 the ridge regression plot has a slightly higher r-squared value at 0.8412157 compared to the adjusted r-squared of 0.8354 above. The ridge regression keeps all parameters in the model, however it places the most weight on four parameters, with the higher coefficients. These four parameters would be selected when the lambda is a minimum as shown by the vertical line. The model in question 1 only has three parameters with difference from the ridge regression.

Task 3.A – lasso regression

```
X <- model.matrix(Life.expectancy~0+., data=df2)
y <- df2$Life.expectancy
# alpha picks either squared penalty or abs penalty
# model
rmod_lasso <- glmnet(x=X,y=y, alpha=1) # lasso
# cross validate within mode
rmodcv_lasso <- cv.glmnet(x=X,y=y, alpha=1, nfolds=10, set.seed(1)) # lasso
```

Task 3.B

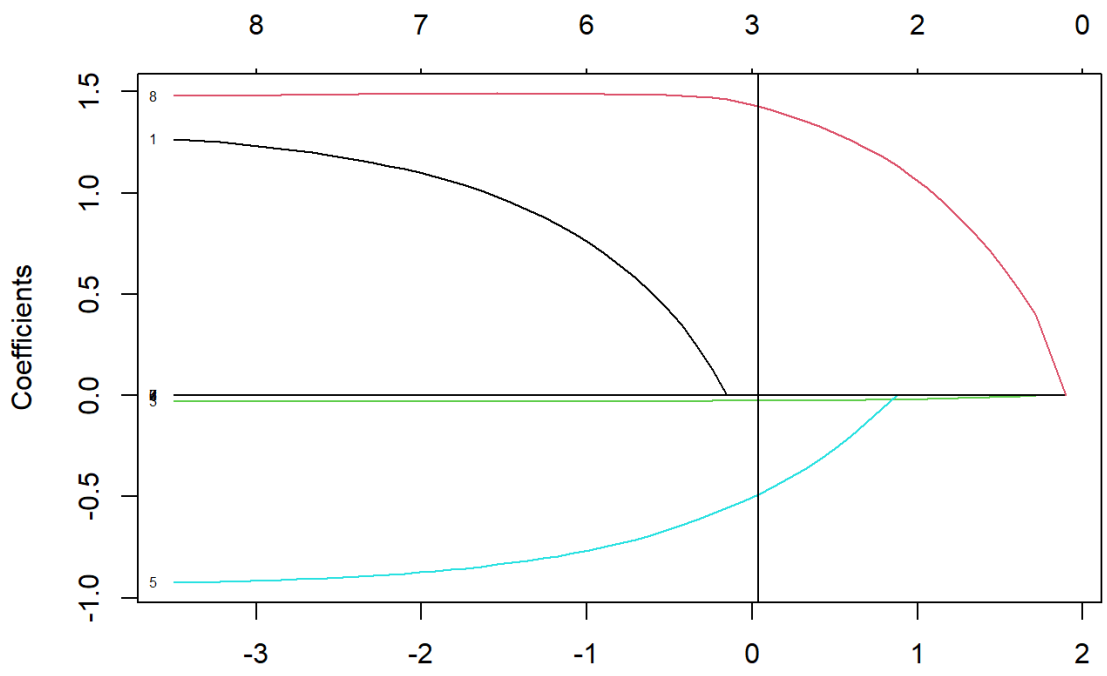
```
rmodcv_lasso$lambda.1se
```

```
## [1] 1.036818
```

The one standard error lambda for a 10 fold cross validation is 1.036818 for this model.

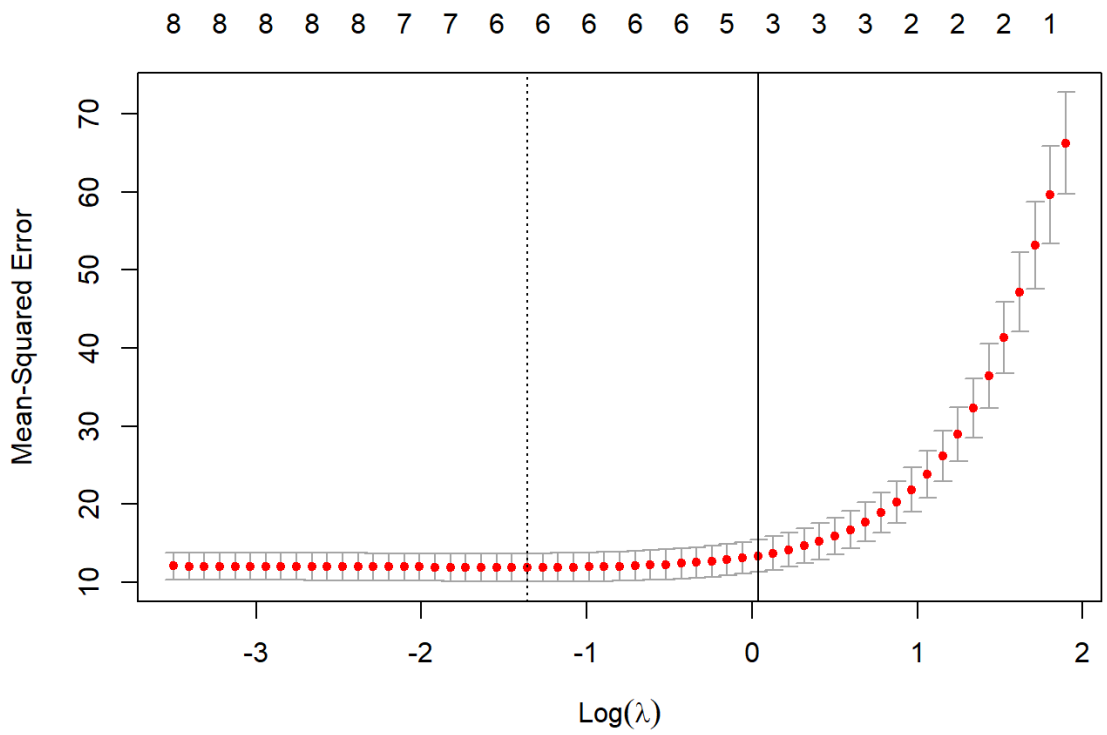
Task 3.C

```
plot(rmod_lasso, label=T, xvar='lambda')+abline(v=log(rmodcv_lasso$lambda.1se))
```




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

```
plot(rmodcv_lasso)+abline(v=log(rmodcv_lasso$lambda.1se))
```




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

Task 3.D

Based on the lambda coefficient plot for the lasso model, it appears the parameters chosen for a 1 standard deviation lambda would correspond to parameters 5 and 8. Which are the variables HIV.AIDS and Schooling. The variable Status could also be considered, as the StatusDeveloped feature has the next highest coefficient and the only other essentially non-zero value.