# HW1 Num2

*Aaron Coates*

*4/17/2019*

```
setwd("~/Documents/GitHub/MMSS_311_2")
wid <- read.csv("/Users/aaroncoates/Downloads/widget_data.csv")

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(glmnet)
```

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

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
library(broom)
library(aod)
```

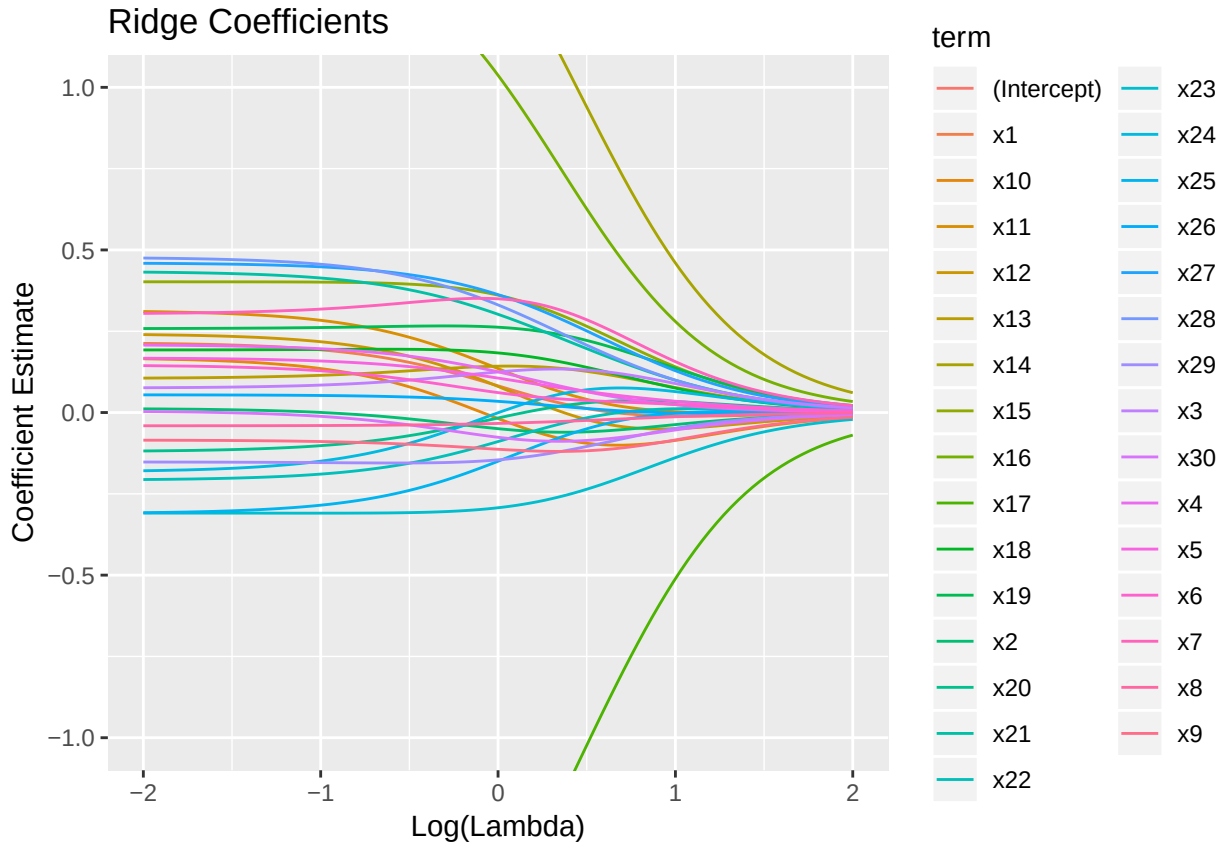
First, I plot the dependent variable, widget performance.


I perform the ridge regression, with lambda ranging from 1/100 to 100.

```
grid=10^seq(2,-2,length=100)
ridge <- glmnet(x = as.matrix(wid[, -1]), y = wid$y, alpha=0, lambda=grid)
```

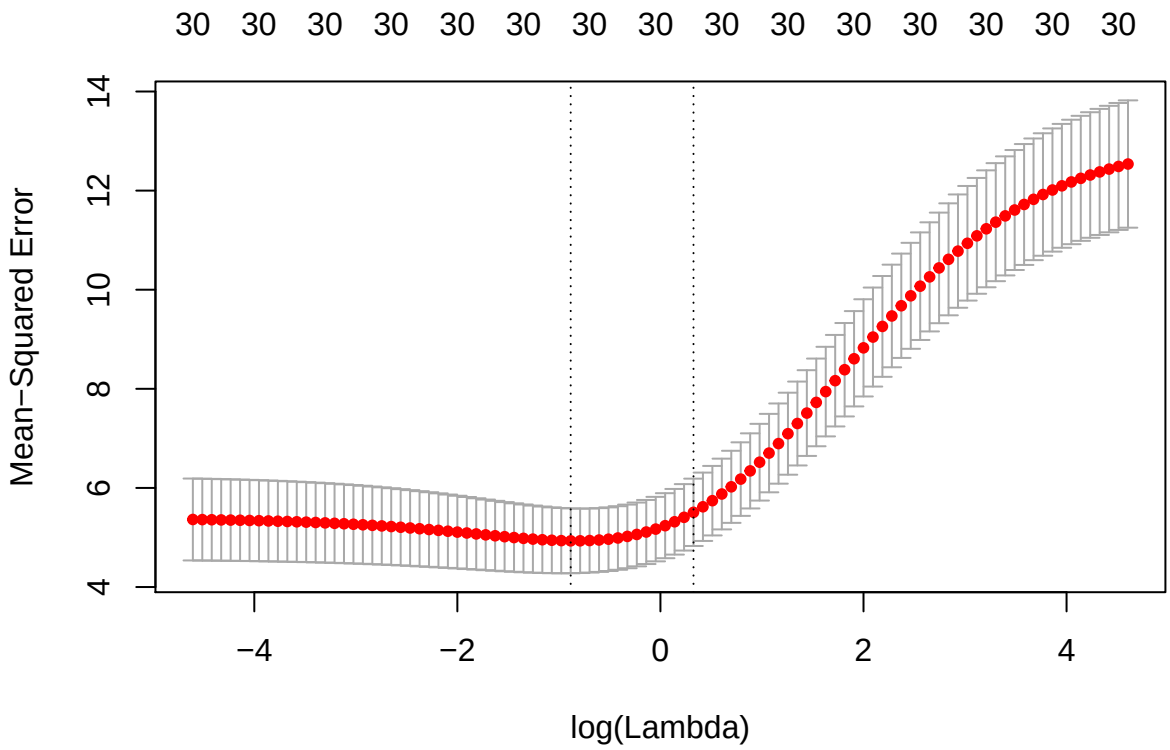
```
ridge1 <- tidy(ridge)
```

Now, I will plot the coefficients as lambda changes.




Now, I will crossvalidate the data and plot the results.

```
crossval <- cv.glmnet(x = as.matrix(wid[, -1]), y = wid$y, alpha=0, lambda=grid)
```



The results below show the lambda value that minimizes MSE. The coefficients are given.

```
crossval$lambda.min
```

```
## [1] 0.4132012
```

```
coef(crossval, s = "lambda.min")
```

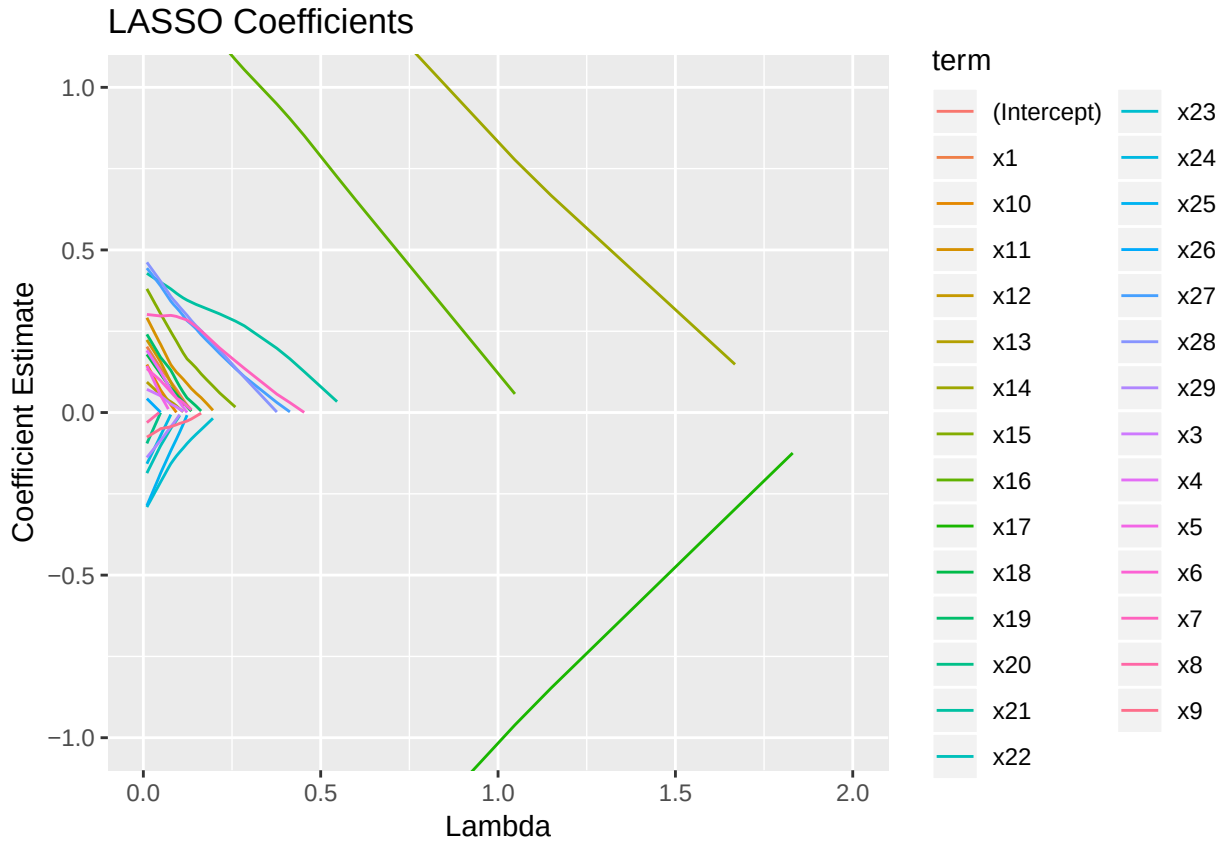
```
## 31 x 1 sparse Matrix of class "dgCMatrix"
##                       1
## (Intercept)  3.55629124
## x1           0.14010399
## x2          -0.02625175
## x3           0.10510766
## x4           0.16464575
## x5           0.13453094
## x6           0.09195510
## x7           0.34359006
## x8          -0.03744489
## x9          -0.10065883
## x10          0.05908487
## x11          0.21258857
## x12          0.15576680
## x13          0.13240286
## x14          1.70685664
## x15          0.39080920
## x16          1.26855954
## x17         -1.87230076
## x18          0.19366222
## x19          0.26614793
## x20         -0.06023155
## x21          0.36377907
## x22         -0.14339337
## x23         -0.30632889
## x24         -0.07525945
## x25         -0.22247347
## x26          0.04557173
## x27          0.41395747
## x28          0.40140857
## x29         -0.15460807
## x30         -0.04665621
```

Now, I will perform the same steps for the LASSO.

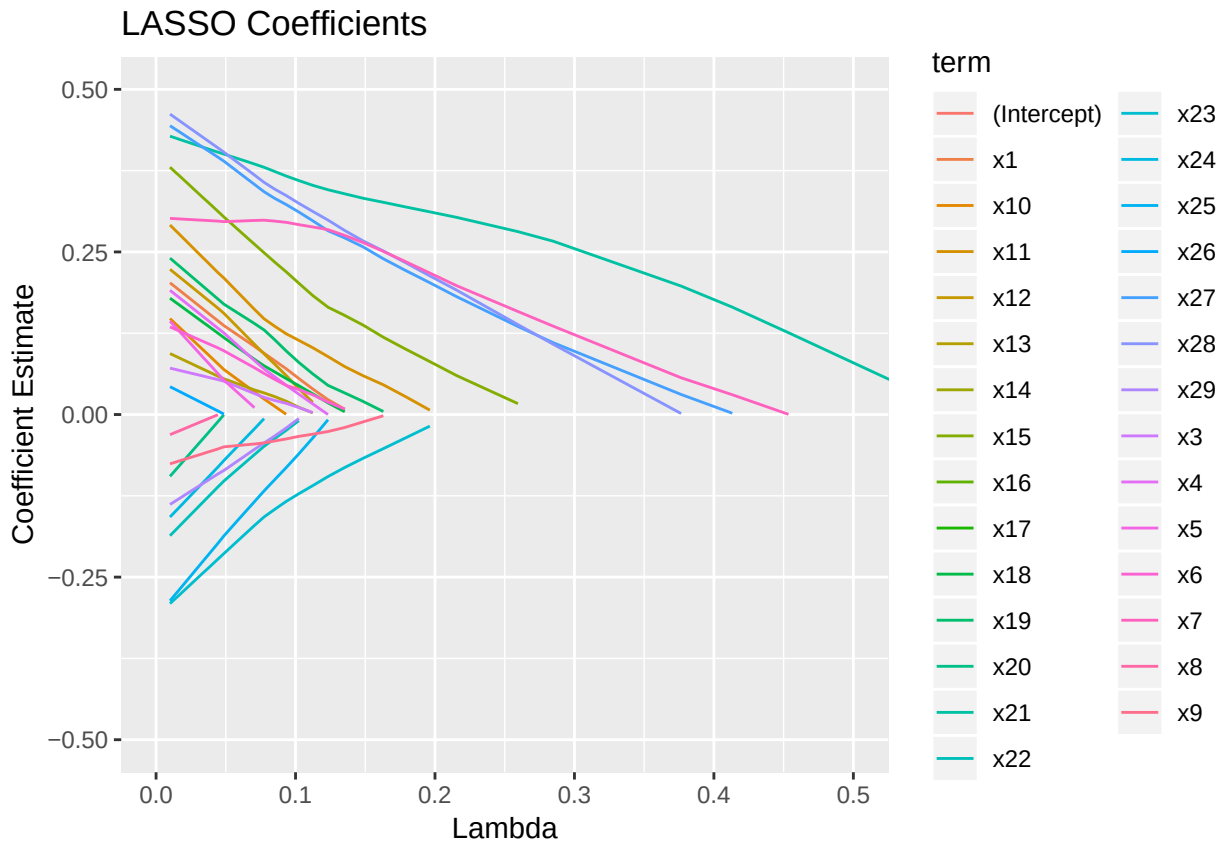
```
grid=10^seq(2,-2,length=100)
lasso <- glmnet(x = as.matrix(wid[, -1]), y = wid$y, alpha=1, lambda=grid)
```

```
lasso1 <- tidy(lasso)
```

Below, I will plot the coefficient estimates as lambda changes.

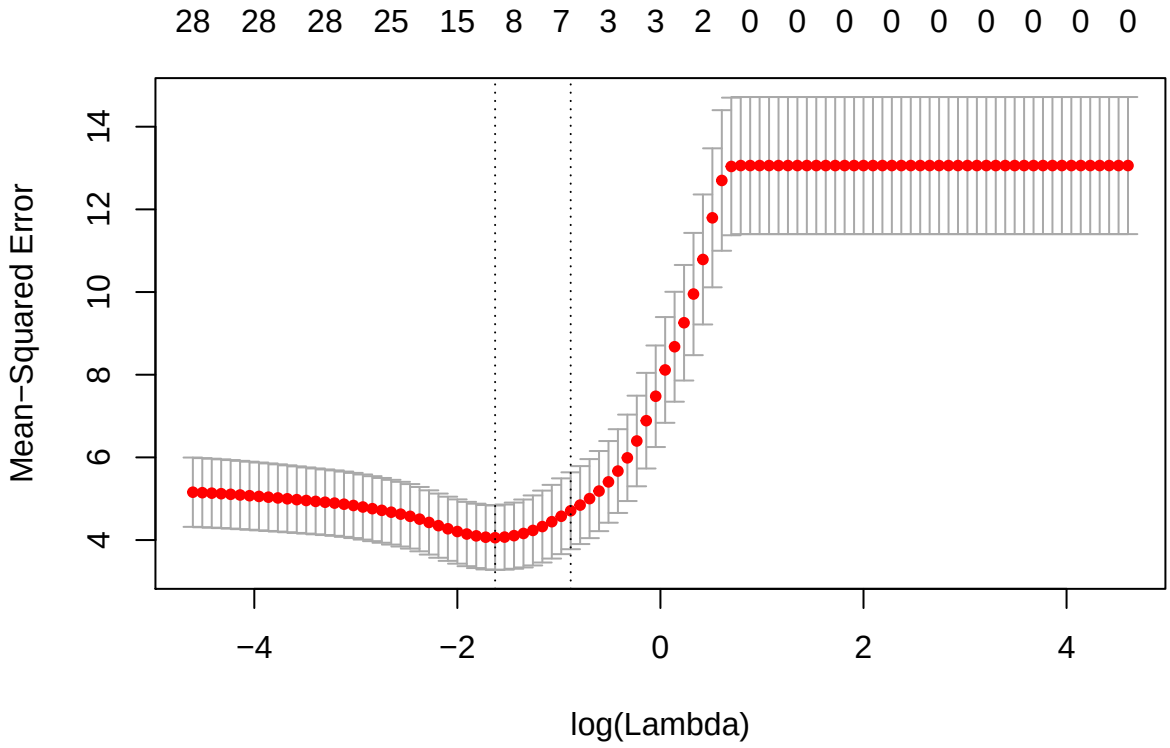



The graph below zooms in on the clustered lines above.




Now, I will crossvalidate and plot the results.

```
crossvallasso <- cv.glmnet(x = as.matrix(wid[, -1]), y = wid$y, alpha=1, lambda=grid)
```



Below is the lambda value that minimizes MSE. The coefficient estimates are given.

```
crossvallasso$lambda.min
```

```
## [1] 0.1963041
```

```
coef(crossvallasso, s = "lambda.min")
```

```
## 31 x 1 sparse Matrix of class "dgCMatrix"
##                          1
## (Intercept)  3.768062409
## x1           .
## x2           .
## x3           .
## x4           .
## x5           .
## x6           .
## x7           0.217553111
## x8           .
## x9           .
## x10          .
## x11          0.006782432
## x12          .
## x13          .
## x14          1.792361518
## x15          0.081407403
## x16          1.164083238
## x17         -1.939493895
## x18          .
## x19          .
## x20          .
## x21          0.311618325
## x22          .
## x23         -0.017812790
## x24          .
## x25          .
## x26          .
## x27          0.202638484
## x28          0.213191110
## x29          .
## x30          .
```

As can be seen, the ridge regression shrinks the coefficient on every independent variable, while the LASSO regression selects the most important independent variables, while the less important coefficients shrink to zero. It seems as though the LASSO regression is the more effective regression technique in this instance. This is because the LASSO helps to eliminate less important variables. This is especially helpful since the question states that we do not know which specific data are relevant for the outcome we are measuring.