STP 598: Homework 3

Antonio Campbell, Sinta Sulistyo, Atta Ullah, Penny Wu

3/9/2021

1. Problem 1. Basic Optimization, MLE for IID Poisson Data

Suppose y_i is a count then a very common model is to assume the Poisson disttribuion:

$$P(Y = y \mid \lambda) = \frac{e^{-\lambda}\,\lambda^{y}}{y!},\ y = 0, 1, 2, \ldots$$

Given $Y_i \sim Poisson(\lambda)$ iid, (that is, $Y_i = y_i$), what is the MLE of λ?

If the random variables are iid, then the joint distribution is the product of the marginal distributions. Let $\mathbf{y}$ be the sample $y_1, \ldots, y_n$

$$L(\lambda|\mathbf{y}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\,\lambda^{y_i}}{y_i!}.$$

We take the log of the likelihood so that we can maximize easily.

$$\log(L(\lambda|\mathbf{y})) = \sum_{i=1}^{n} \log\left(\frac{e^{-\lambda}\,\lambda^{y_i}}{y_i!}\right) = -n\lambda - \sum \log(y_i!) + \sum (y_i)\log(\lambda).$$

Differentiate with respect to lambda and set to zero to maximize:

$$\frac{\delta \log(L(\lambda|\mathbf{y}))}{\delta d\lambda} = -n + \frac{\sum y_i}{\lambda} = 0 \implies \hat{\lambda} = \frac{\sum y_i}{n}$$

.

The MLE for the parameter of the *poisson* distribution is the sample mean.

Problem 2. Constrained Optimization, Minimum Variance Portfolio

Suppose we are considering investing in p stocks where the uncertain return on the ith stock is denoted by R_i, $i = 1, 2, \ldots, p$. Let $R = (R_1, R_2, \ldots, R_p)^{'}$. A portfolio is a given by $w = (w_1, w_2, \ldots, w_p)^{'}$ where w_i is the fraction of wealth invested in asset i. The $\{w_i\}$ must satisfy $\sum w_i = 1$. The return on the portfolio is then

$$P = w'R = \sum w_i R_i.$$

We want to find the global **minimum variance portfolio**:

$$\min_{w} Var(P),\ \text{ subject to} \sum w_i = 1.$$

If we let $\iota = (1, 1, \ldots, 1)^{'}$, the vector of ones, and $Var(R) = \Sigma$ then our problem is

$$\min_{w} w^{'}\Sigma w\ \text{ subject to } w^{'}\iota = 1.$$

Find the global minimum variance portfolio in terms of Σ and ι.

If we know that $\sum w_i = 1$ we can easily say $\sum w_i - 1 = 0$ or $w^{'}\iota - 1 = 0$. We use the Karush-Kuhn-Tucker conditions to minimize. By this we take the derivative with respect to w,

$$\frac{\delta L(\lambda, w)}{\delta w} = \frac{\delta(w^{'}\Sigma w)}{\delta w} + \frac{\delta\lambda(w^{'}\iota - 1)}{\delta w} = 2\Sigma w + \lambda\iota$$

$$\implies \hat{w} = -\lambda\Sigma^{-1}\iota/2.$$

To express this solution in terms of Σ, ι we solve for λ. Premultiply by $\iota^{'}$ and then solve. Since we have that $\iota^{'}w = 1$,

$$\iota^{'}w = -\lambda\iota^{'}\Sigma^{-1}\iota/2 \implies \lambda = -2/(\iota^{'}\Sigma^{-1}\iota).$$

We plug this into our solution of $\hat{w}$ to have:

$$\hat{w} = \frac{-\lambda\Sigma^{-1}\iota/2}{-2/\iota'\Sigma^{-1}\iota} = \frac{\lambda\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota}$$

$\hat{w}$ is the w such that $Var(w^{'}R) = Var(P)$ is minimized.

Problem 3. Polynomial Regression

A basic idea in nonlinear regression is to use polynomial terms.

With one x variable, this means we consider the models:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \epsilon_i.$$

Using the simple used cars data (with $n = 1,000$) with Y= price and x=mileage, find the best choice of p. * Use BIC to pick p * use an out-of-sample criteria to pick p

Fit your chosen polynomial mode using all the data and plot the fit on top of the data. Do you like it? Also plot the fits for a p that is "way to big". What is wrong with it?

Solution

Impriting libraries and loading dataset

```
##   price trim isOneOwner mileage year  color displacement     fuel region
## 1  2988  320          f  193296 1995  Black          3.2 Gasoline    SoA
## 2  6595  320          f  129948 1995  other          3.2 Gasoline    Mid
## 3  7993  320          f  140428 1997  White          3.2 Gasoline    Mid
## 4  5995  420          f  113622 1999 Silver          4.2 Gasoline    Mid
## 5  3000  420          f  167673 1999 Silver          4.2 Gasoline    SoA
## 6  7400  430          f   82419 2002  White          4.3 Gasoline    Mid
##   soundSystem wheelType
## 1        unsp     Alloy
## 2     Premium     Alloy
## 3        Bose     Alloy
## 4        unsp     Alloy
## 5        unsp     Alloy
## 6        Bose     Alloy
```

We only need the price and mileage and various powers of mileage. We create a new dataset with the first column as the price of the following rest of the columns are the mileage with exponent 1 through 9.

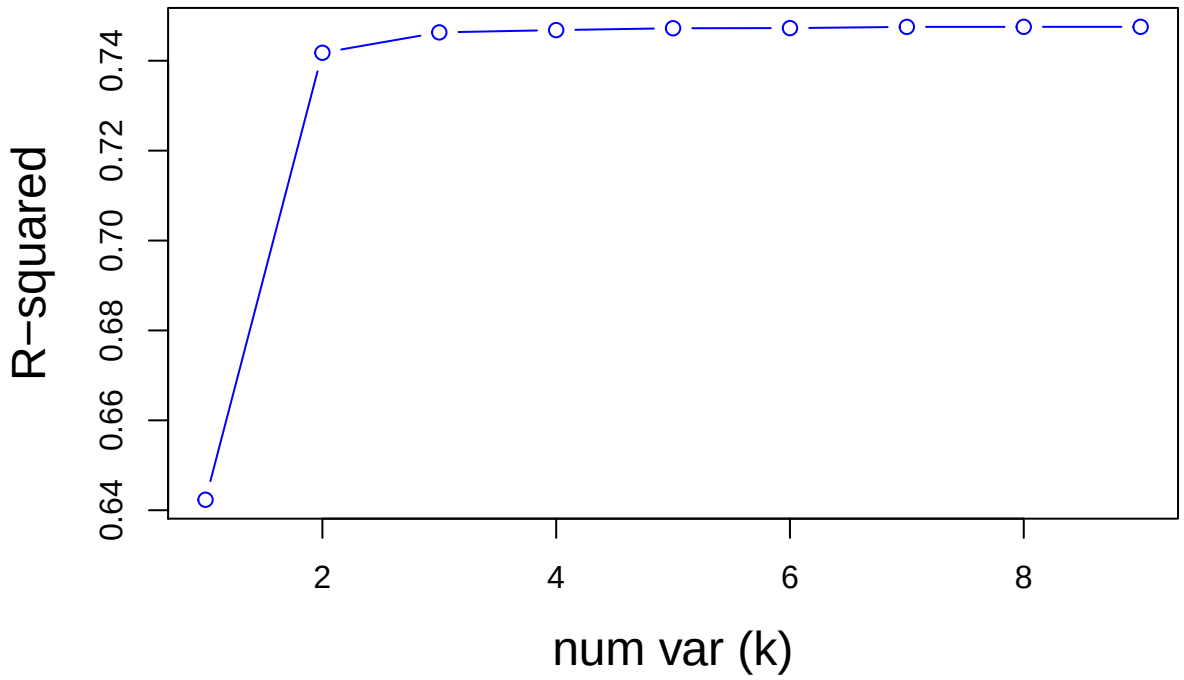
```
##   price mileage    mileage2     mileage3     mileage4     mileage5     mileage6
## 1  2988  193296 37363343616 7.222185e+15 1.396019e+21 2.698450e+26 5.215995e+31
## 2  6595  129948 16886482704 2.194365e+15 2.851533e+20 3.705510e+25 4.815236e+30
## 3  7993  140428 19720023184 2.769243e+15 3.888793e+20 5.460954e+25 7.668709e+30
##       mileage7     mileage8     mileage9
## 1 1.008231e+37 1.948870e+42 3.767088e+47
## 2 6.257303e+35 8.131240e+40 1.056638e+46
## 3 1.076901e+36 1.512271e+41 2.123652e+46
```

Plot of R-Squared Error

```
##   (Intercept) mileage mileage2 mileage3 mileage4 mileage5 mileage6 mileage7
## 1        TRUE    TRUE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 2        TRUE    TRUE     TRUE    FALSE    FALSE    FALSE    FALSE    FALSE
## 3        TRUE    TRUE     TRUE    FALSE     TRUE    FALSE    FALSE    FALSE
## 4        TRUE    TRUE    FALSE     TRUE     TRUE    FALSE     TRUE    FALSE
## 5        TRUE    TRUE    FALSE    FALSE     TRUE    FALSE     TRUE    FALSE
## 6        TRUE    TRUE     TRUE     TRUE    FALSE    FALSE    FALSE     TRUE
## 7        TRUE   FALSE     TRUE     TRUE     TRUE     TRUE     TRUE    FALSE
## 8        TRUE    TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
## 9        TRUE    TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
##   mileage8 mileage9
## 1    FALSE    FALSE
## 2    FALSE    FALSE
## 3    FALSE    FALSE
## 4    FALSE    FALSE
## 5     TRUE     TRUE
## 6     TRUE     TRUE
## 7     TRUE     TRUE
## 8    FALSE     TRUE
## 9     TRUE     TRUE
```

```
##   (Intercept) mileage mileage2 mileage3 mileage4 mileage5 mileage6 mileage7
## 1        TRUE    TRUE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 2        TRUE    TRUE     TRUE    FALSE    FALSE    FALSE    FALSE    FALSE
## 3        TRUE    TRUE     TRUE    FALSE     TRUE    FALSE    FALSE    FALSE
## 4        TRUE    TRUE    FALSE     TRUE     TRUE    FALSE     TRUE    FALSE
## 5        TRUE    TRUE    FALSE    FALSE     TRUE    FALSE     TRUE    FALSE
## 6        TRUE    TRUE     TRUE     TRUE    FALSE    FALSE    FALSE     TRUE
## 7        TRUE   FALSE     TRUE     TRUE     TRUE     TRUE     TRUE    FALSE
## 8        TRUE    TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
## 9        TRUE    TRUE     TRUE     TRUE     TRUE     TRUE     TRUE     TRUE
##   mileage8 mileage9
## 1    FALSE    FALSE
## 2    FALSE    FALSE
## 3    FALSE    FALSE
## 4    FALSE    FALSE
## 5     TRUE     TRUE
## 6     TRUE     TRUE
## 7     TRUE     TRUE
## 8    FALSE     TRUE
```

```
## 9     TRUE     TRUE
```



As expected, r-sqaured error is monotonically increasing as the number of variables increase.

The following function gives the coefficients for any choice of variables. For instance, if we happen to choose the best subset with 5 varaibles, we use

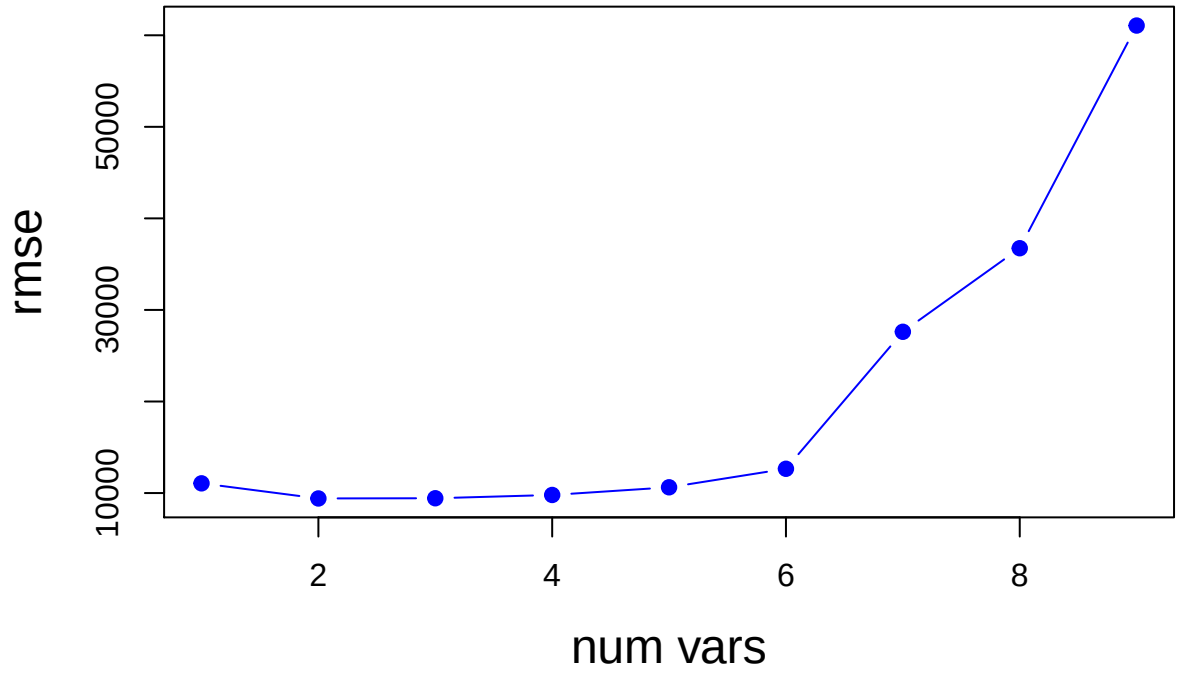
```
##   (Intercept) mileage mileage2 mileage3 mileage4 mileage5 mileage6 mileage7
## 1        TRUE    TRUE    FALSE    FALSE    FALSE    FALSE    FALSE    FALSE
## 2        TRUE    TRUE     TRUE    FALSE    FALSE    FALSE    FALSE    FALSE
## 3        TRUE    TRUE     TRUE    FALSE     TRUE    FALSE    FALSE    FALSE
## 4        TRUE    TRUE    FALSE     TRUE     TRUE    FALSE     TRUE    FALSE
## 5        TRUE    TRUE    FALSE    FALSE     TRUE    FALSE     TRUE    FALSE
##   mileage8 mileage9
## 1    FALSE    FALSE
## 2    FALSE    FALSE
## 3    FALSE    FALSE
## 4    FALSE    FALSE
## 5     TRUE     TRUE
```

```
##   (Intercept)       mileage      mileage2      mileage4
##  7.176771e+04 -8.152826e-01  2.771420e-06 -8.457424e-18
```

function to do rmse for k in $1:p$

Do validation approach several times

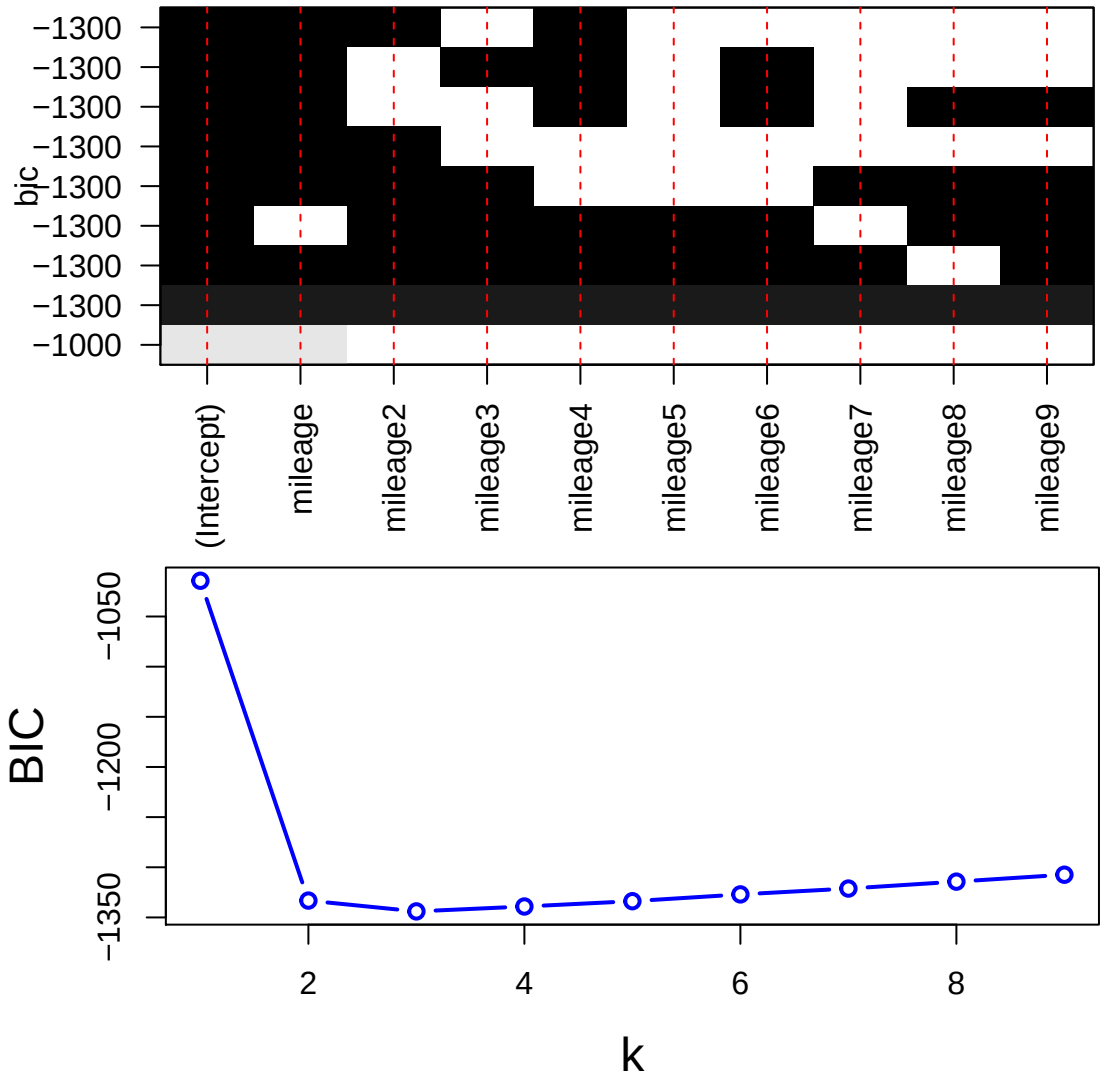
Plot results of repeated train/val



Fit using number of vars chosen by train/validation and all the data.

```
##
## Call:
## lm(formula = price ~ ., data = ddfsub)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44202  -5320      4   5662  29557
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.177e+04  1.141e+03  62.875  < 2e-16 ***
## mileage     -8.153e-01  3.386e-02 -24.077  < 2e-16 ***
## mileage2     2.771e-06  2.353e-07  11.776  < 2e-16 ***
## mileage4    -8.457e-18  2.006e-18  -4.216 2.71e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9248 on 996 degrees of freedom
## Multiple R-squared:  0.7463, Adjusted R-squared:  0.7455
## F-statistic: 976.7 on 3 and 996 DF,  p-value: < 2.2e-16
```

BIC



Ploting the model with the best coice of number of variables.

Using BIC and Cross validation.

By eyeballing we see that the best choice of p for both is 3.

```
##   (Intercept)       mileage      mileage2      mileage4
##  7.176771e+04 -8.152826e-01  2.771420e-06 -8.457424e-18
```

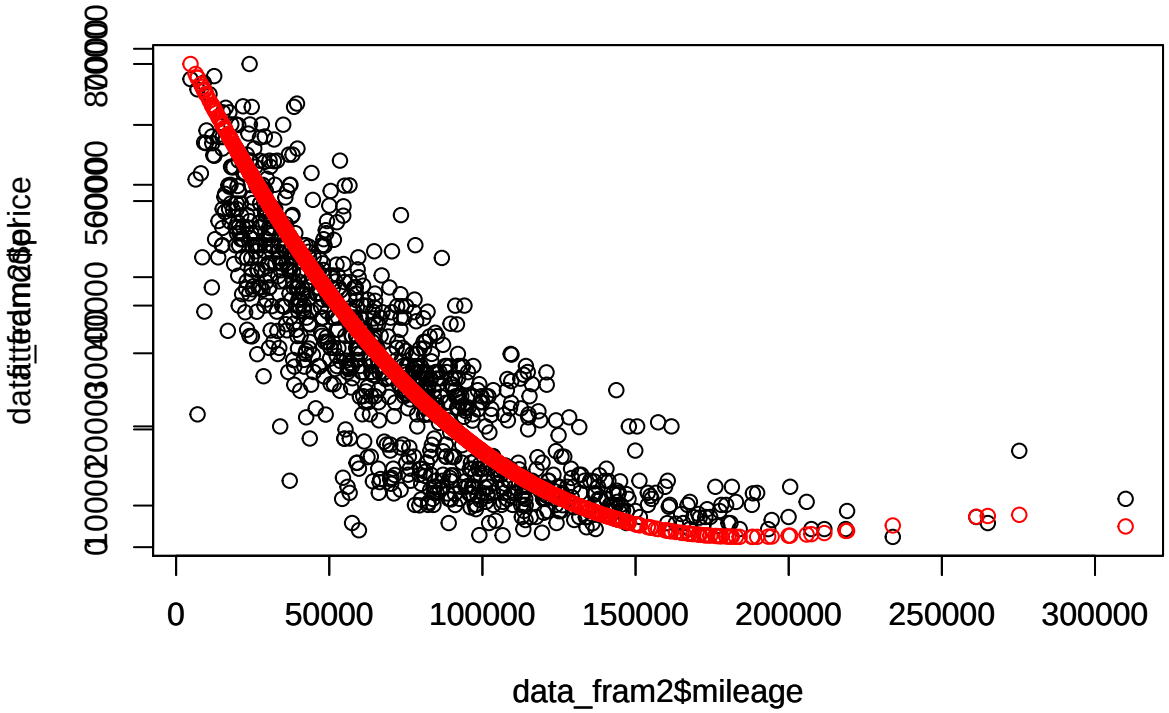
Thus the best polynomila would be

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_4 x_i^4$$

Where

$$\beta_0 = 7.176771 \times 10^{04} \beta_1 = -8.152826 \times 10^{-01} \beta_2 = 2.771420 \times 10^{-06} \beta_4 = -8.457424 \times 10^{-18}$$

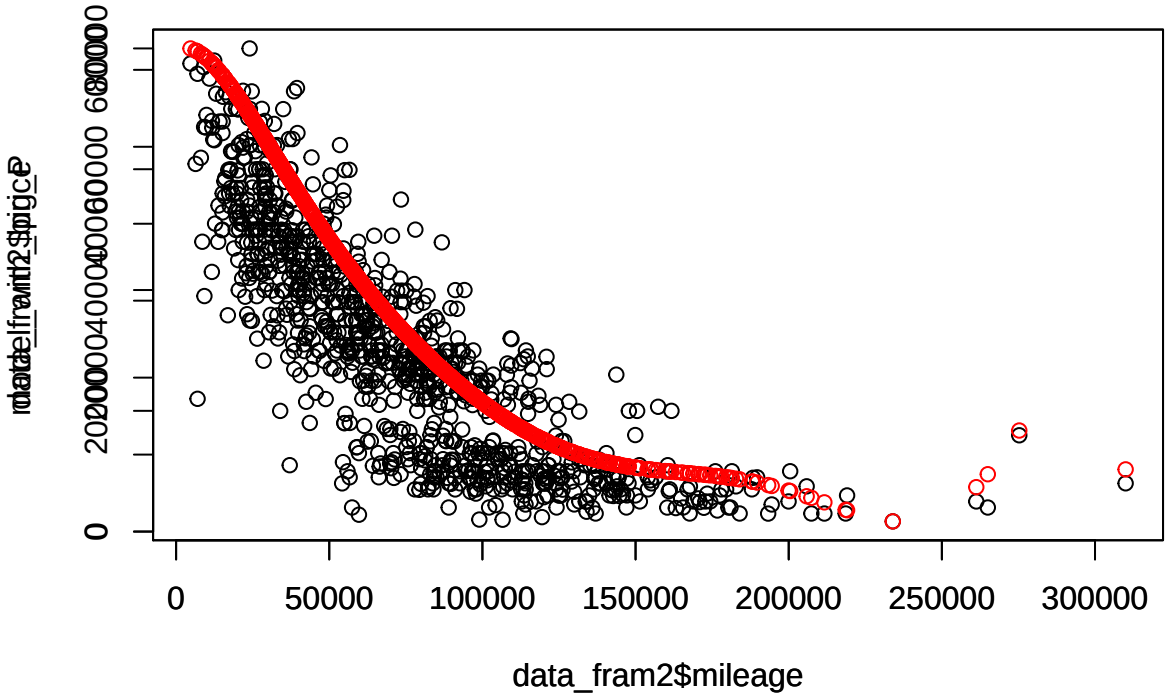
We plot this polynomial with data



This looks good except for the mileage from 150000 to 20000 is a bit off.

Using a "way to big" p.

Since in our data frame we have used the maximum degree 9, we will fit a polynomial of degree 9.



This one is not as good as the previous and slight off for most of the data but does capture the points near the boundary vary well.

Problem 4.Regularized Regression

Let's try ridge and LASSO on the car price data. The cars data has 20 thousand observations and 11 variables. Many of the x variables are categorial so we need to dummy them up.

```
## [1] 20063    11
```

```
##   price trim isOneOwner mileage year  color displacement     fuel region
## 1  2988  320          f  193296 1995  Black          3.2 Gasoline    SoA
## 2  6595  320          f  129948 1995  other          3.2 Gasoline    Mid
## 3  7993  320          f  140428 1997  White          3.2 Gasoline    Mid
## 4  5995  420          f  113622 1999 Silver          4.2 Gasoline    Mid
## 5  3000  420          f  167673 1999 Silver          4.2 Gasoline    SoA
## 6  7400  430          f   82419 2002  White          4.3 Gasoline    Mid
##   soundSystem wheelType
## 1        unsp     Alloy
## 2     Premium     Alloy
## 3        Bose     Alloy
## 4        unsp     Alloy
## 5        unsp     Alloy
## 6        Bose     Alloy
```

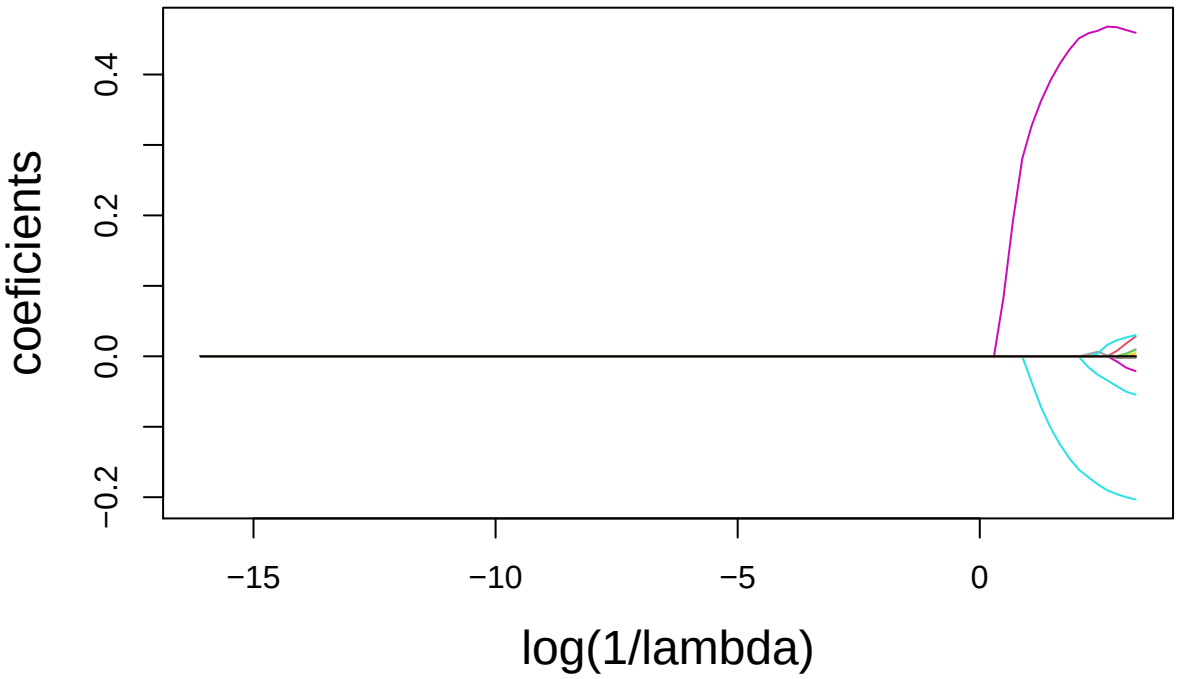
```
##      price            trim       isOneOwner    mileage            year
##  Min.   :  599   550    :11825   f:16594    Min.   :     8   Min.   :1994
##  1st Qu.:13495   430    : 2787   t: 3469    1st Qu.: 39888   1st Qu.:2004
##  Median :29454   500    : 2661              Median : 67187   Median :2007
##  Mean   :30747   63 AMG :  599              Mean   : 73114   Mean   :2007
##  3rd Qu.:43995   600    :  572              3rd Qu.: 98213   3rd Qu.:2010
##  Max.   :79999   55 AMG :  356              Max.   :488525   Max.   :2014
##                  (Other): 1263
##     color       displacement        fuel           region
##  Black :8194   5.5    :9561   Diesel  :  211   SoA    :6492
##  Blue  : 914   4.6    :2794   Gasoline:19632   Pac    :3252
##  Gray  :2168   4.3    :2787   Hybrid  :  220   Mid    :2671
##  other : 961   5      :2661                    WSC    :2493
##  Silver:4353   6.3    : 494                    ENC    :1984
##  unsp  : 832   5.4    : 356                    Mtn    :1006
##  White :2641   (Other):1410                    (Other):2165
##         soundSystem     wheelType
##  Bang Olufsen : 104   Alloy  :11111
##  Bose         :1261   other  :  120
##  Harman Kardon:4278   Premium:  428
##  Premium      :5320   unsp   : 8404
##  unsp         :9100
##
##
```

```
## [1] 20063    48
```

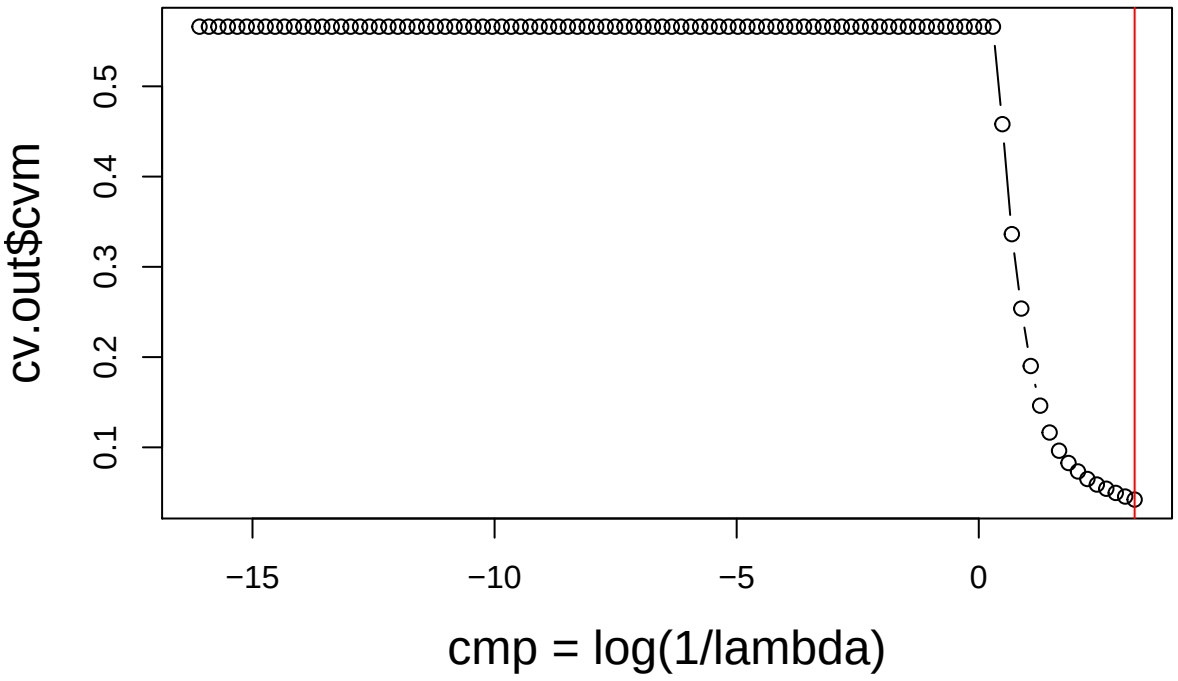
```
##  [1] "trim350"                "trim400"
##  [3] "trim420"                "trim430"
##  [5] "trim500"                "trim55 AMG"
##  [7] "trim550"                "trim600"
##  [9] "trim63 AMG"             "trim65 AMG"
## [11] "isOneOwnert"            "mileage"
## [13] "year"                   "colorBlue"
```

```
## [15] "colorGray"                "colorother"
## [17] "colorSilver"              "colorunsp"
## [19] "colorWhite"               "displacement3.2"
## [21] "displacement3.5"          "displacement3.7"
## [23] "displacement4.2"          "displacement4.3"
## [25] "displacement4.6"          "displacement5"
## [27] "displacement5.4"          "displacement5.5"
## [29] "displacement5.8"          "displacement6"
## [31] "displacement6.3"          "fuelGasoline"
## [33] "fuelHybrid"               "regionESC"
## [35] "regionMid"                "regionMtn"
## [37] "regionNew"                "regionPac"
## [39] "regionSoA"                "regionWNC"
## [41] "regionWSC"                "soundSystemBose"
## [43] "soundSystemHarman Kardon" "soundSystemPremium"
## [45] "soundSystemunsp"          "wheelTypeother"
## [47] "wheelTypePremium"         "wheelTypeunsp"
```

(a). Use the LASSO to relate log of price to the features.

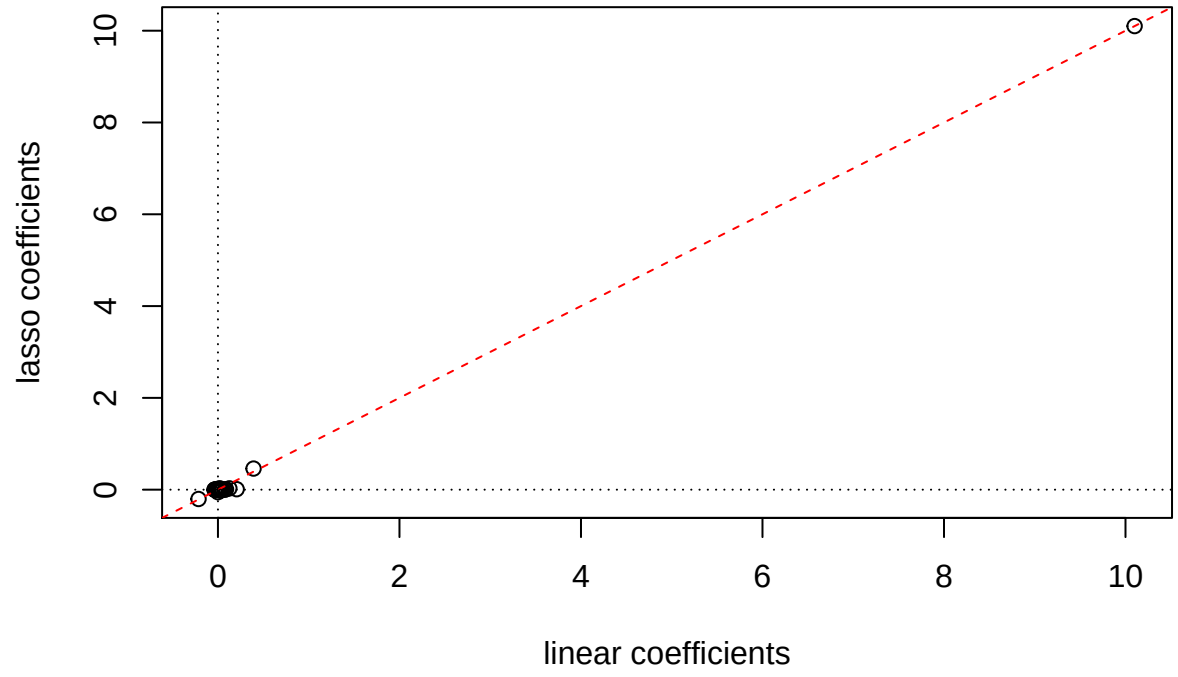


It is clearly seen from the graph above, there ar several variables with zero coefficient, and as lambda gets smaller - meaning that the penalty gets lower - the coefficients increase.




```
## [1] 0.04
```

Based on the cross validation, the best lambda is 0.04.

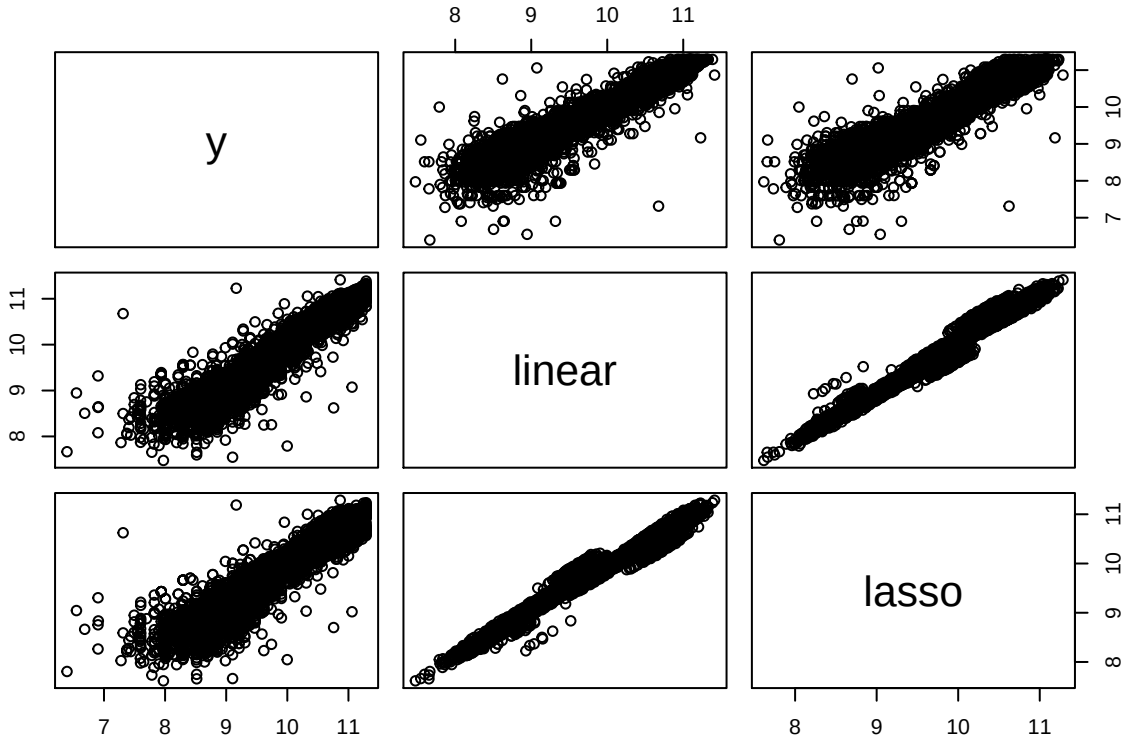


Acknowledging that lambda is relatively small, it is almost similar with linear regression. Hence, the coefficients of lasso and linear regression are very close.

```
##          (Intercept)               trim350               trim400
##         10.100456351          0.000000000          0.000000000
##               trim420               trim430               trim500
##          0.000000000         -0.054435888         -0.021047293
##          trim55 AMG               trim550               trim600
```

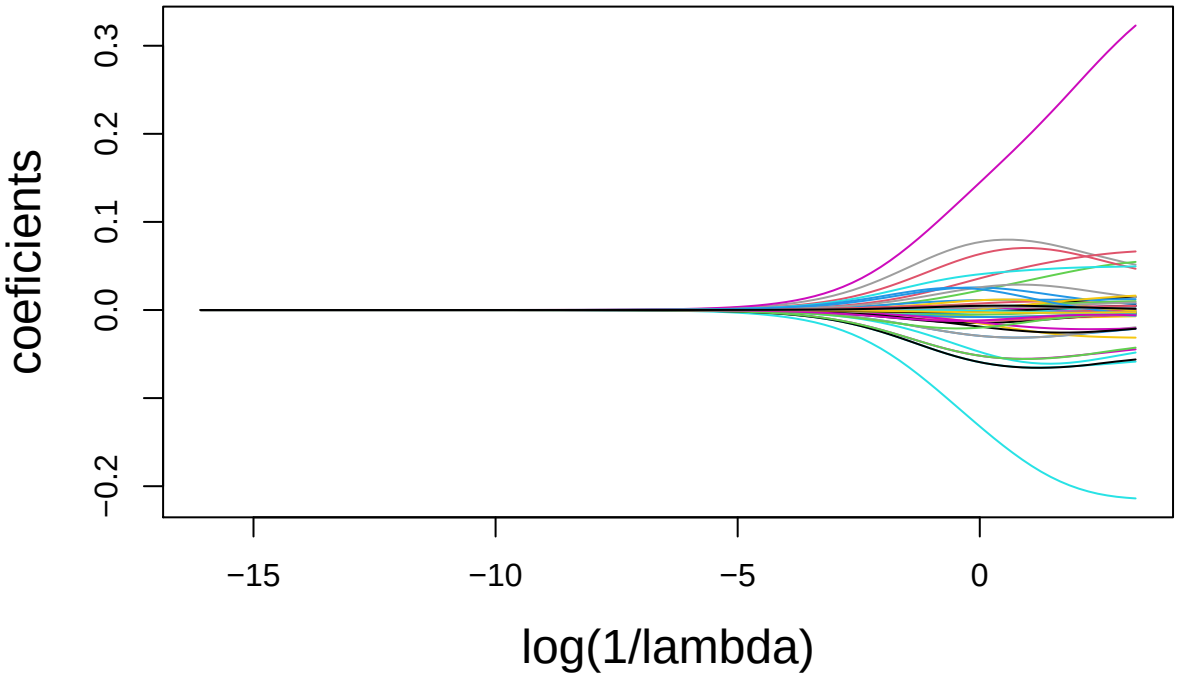
```
##              0.000000000              0.010061012              0.000000000
##               trim63 AMG               trim65 AMG               isOneOwnert
##              0.028151893              0.009197102              0.000000000
##                  mileage                     year                colorBlue
##             -0.203250357              0.459321190              0.000000000
##                colorGray               colorother               colorSilver
##              0.000000000              0.000000000              0.000000000
##                colorunsp               colorWhite          displacement3.2
##              0.000000000              0.000000000              0.000000000
##          displacement3.5          displacement3.7          displacement4.2
##              0.000000000              0.000000000              0.000000000
##          displacement4.3          displacement4.6            displacement5
##             -0.001531371              0.000000000              0.000000000
##          displacement5.4          displacement5.5          displacement5.8
##              0.000000000              0.030198358              0.000000000
##            displacement6          displacement6.3             fuelGasoline
##              0.004345682              0.000000000              0.000000000
##               fuelHybrid                 regionESC                 regionMid
##              0.000000000              0.000000000              0.000000000
##                 regionMtn                 regionNew                 regionPac
##              0.000000000              0.000000000              0.000000000
##                 regionSoA                 regionWNC                 regionWSC
##              0.000000000              0.000000000              0.000000000
##          soundSystemBose soundSystemHarman Kardon       soundSystemPremium
##              0.000000000              0.000000000              0.000000000
##          soundSystemunsp           wheelTypeother          wheelTypePremium
##              0.000000000              0.000000000              0.000000000
##            wheelTypeunsp
##              0.000000000
```

As shown above, some variables have 0 coefficient, such as trim350, trim 400, trim420.

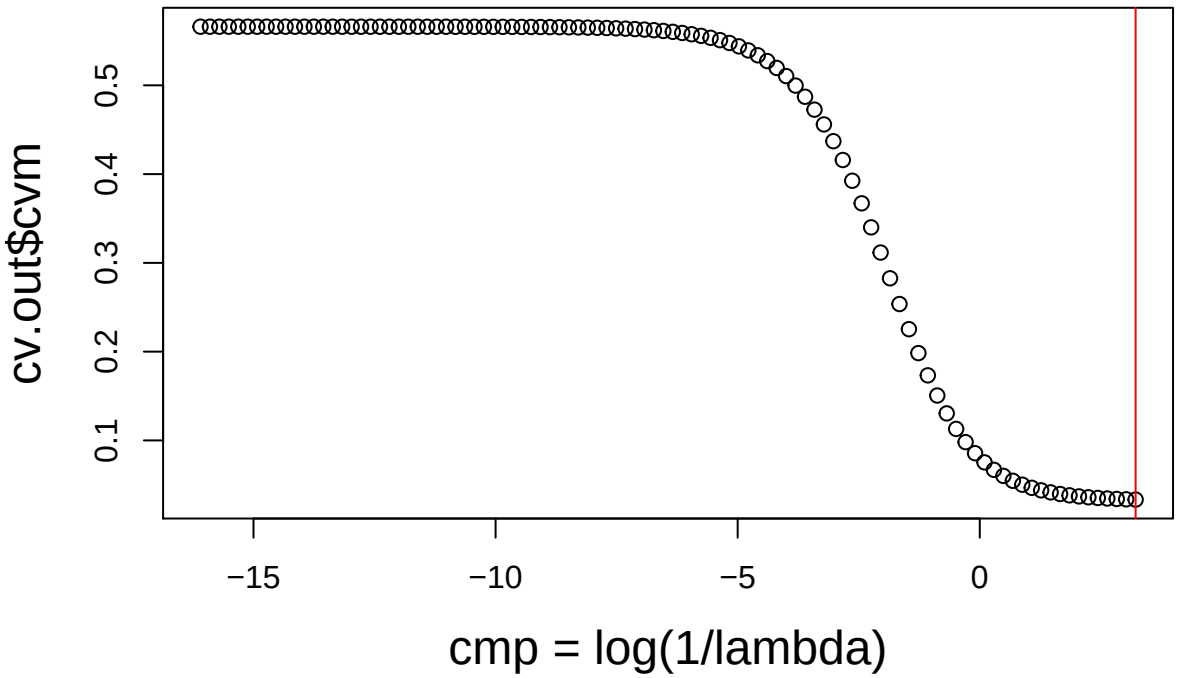


Comparing the fit between linear and lasso with lambda = 0.04, it is pretty close.

(b). Use ridge regression to relate log of price to the features.

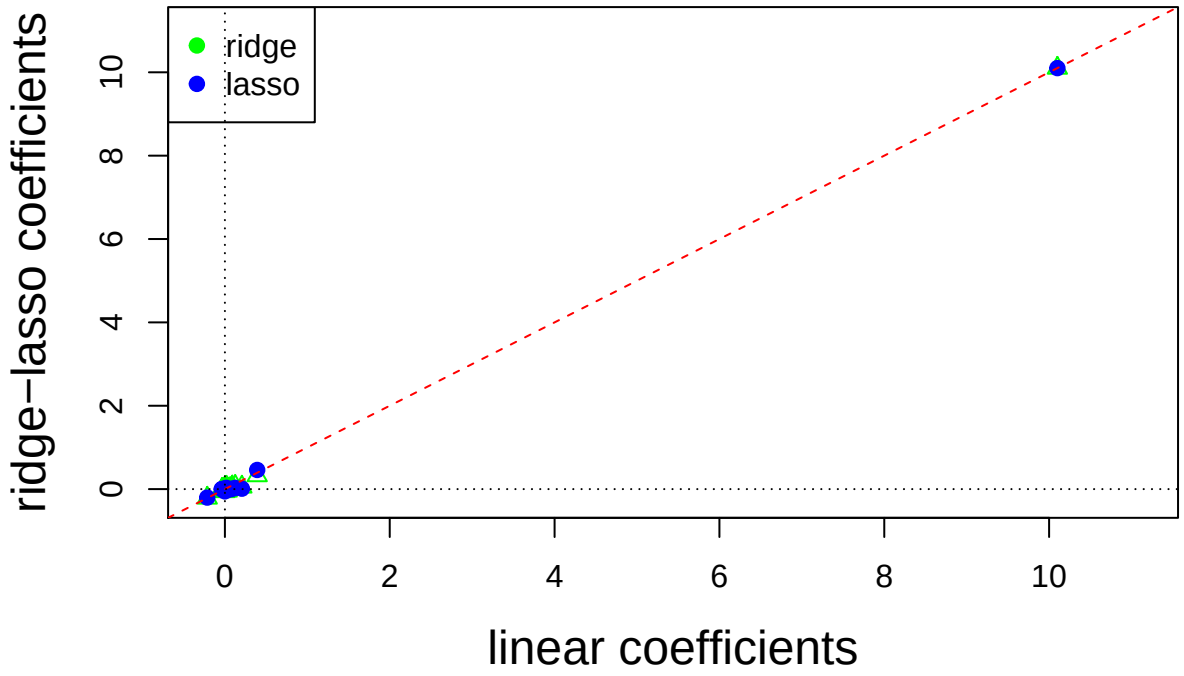


The graph shows the coefficient value of each variables as lambda becomes smaller. Most of the variables' coefficient get bigger as the lambda gets smaller. This makes sense since when lambda is small, the penalty is low.




```
## [1] 0.04
```

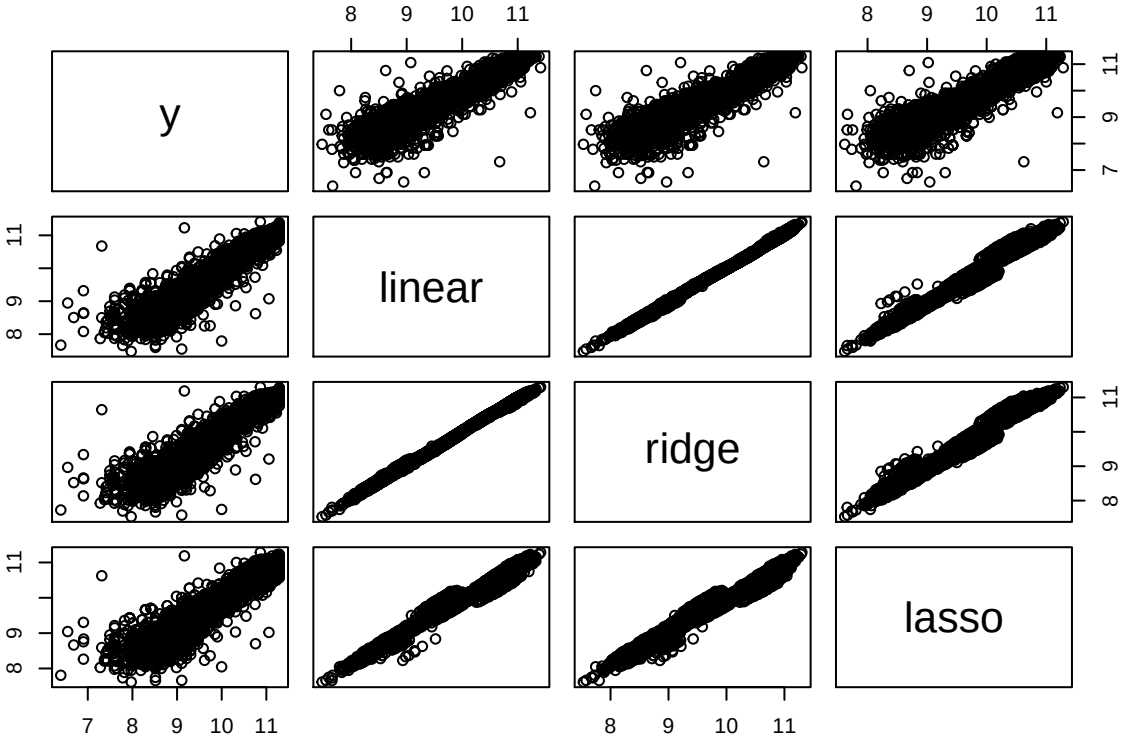
Based on the cross validation, the best lambda for ridge regression is 0.04.




```
##             (Intercept)                   trim350                   trim400
##           10.1004563515             0.0072794311            -0.0018402748
##                 trim420                   trim430                   trim500
##           -0.0213444779            -0.0585559994            -0.0446466927
##              trim55 AMG                   trim550                   trim600
##           -0.0030096369             0.0513904138             0.0147827565
##              trim63 AMG                trim65 AMG                 isOneOwnert
##            0.0664795193             0.0546051416             0.0066718744
##                 mileage                      year                 colorBlue
##           -0.2138628733             0.3229542721            -0.0071490162
##                colorGray                colorother                colorSilver
##           -0.0013495939            -0.0043105774            -0.0055776360
##                colorunsp                colorWhite           displacement3.2
##            0.0001301514             0.0123292204            -0.0482554711
##          displacement3.5          displacement3.7          displacement4.2
##            0.0005477789            -0.0313759698            -0.0196163159
##          displacement4.3          displacement4.6            displacement5
##           -0.0561465475             0.0468653862            -0.0427115911
##          displacement5.4          displacement5.5          displacement5.8
##           -0.0026620986             0.0495116881            -0.0208230717
##            displacement6          displacement6.3             fuelGasoline
##           -0.0007728622             0.0146437690            -0.0211912636
##              fuelHybrid                 regionESC                 regionMid
##           -0.0029915783             0.0100105234            -0.0027844123
##                regionMtn                 regionNew                 regionPac
##            0.0038741043             0.0043104266             0.0162995207
##                regionSoA                 regionWNC                 regionWSC
##            0.0086846989             0.0049501583             0.0046937152
##          soundSystemBose soundSystemHarman Kardon       soundSystemPremium
##           -0.0046959650            -0.0052423940            -0.0067485086
##          soundSystemunsp           wheelTypeother          wheelTypePremium
```

```
##            -0.0060461537            -0.0025251908             0.0018394325
##            wheelTypeunsp
##             0.0014175177
```

Using ridge regression, there is no variable selection since all of the variables' coefficient is not zero.



The plot above shows the comparison between linear, ridge, and lasso.