E392: Problem Set 5

Predicting Wages in the US

Spring 2018

Due: March 7th 2018

Please work on the following questions and hand in your solutions in groups of at most 2 students. You are asked to answer all questions, but we will only select 2 (sub)questions randomly to grade.

The problem is prediction of wages in the US. To that end, you collect US census data from the CPS in the year 2012. The dependent variable is the logarithm of the wage. All other variables denote some other socio-economic characteristics, e.g. marital status, education, and experience. The data can be found in the package "hdm" (`install.packages("hdm")`) under the name `cps2012`. First, consider the 16 predictors female + widowed + divorced + separated + nevermarried + hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3.

Question 1: Load and prepare the data.

```r
library(hdm)
data(cps2012)
help(cps2012)
```

```
## starting httpd help server ... done
```

```r
summary(cps2012)
```

```
##       year           lnw             female          widowed
##  Min.   :2012   Min.   :-7.470   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:2012   1st Qu.: 2.408   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :2012   Median : 2.775   Median :0.0000   Median :0.000000
##  Mean   :2012   Mean   : 2.797   Mean   :0.4288   Mean   :0.007975
##  3rd Qu.:2012   3rd Qu.: 3.182   3rd Qu.:1.0000   3rd Qu.:0.000000
##  Max.   :2012   Max.   : 5.971   Max.   :1.0000   Max.   :1.000000
##     divorced        separated        nevermarried        hsd08
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000000
##  Mean   :0.1134   Mean   :0.0166   Mean   :0.1563   Mean   :0.004107
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.000000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000
##      hsd911            hsg               cg                ad
```

```
##  Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.00000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.02218   Mean   :0.2473   Mean   :0.2834   Mean   :0.1558
##  3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##        mw               so               we              exp1
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:11.50
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :19.00
##  Mean   :0.2916   Mean   :0.2828   Mean   :0.1996   Mean   :18.76
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:26.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :43.50
##       exp2             exp3             exp4             weight
##  Min.   : 0.000   Min.   : 0.000   Min.   :  0.000   Min.   : 106.8
##  1st Qu.: 1.323   1st Qu.: 1.521   1st Qu.:  1.749   1st Qu.: 654.2
##  Median : 3.610   Median : 6.859   Median : 13.032   Median :1472.1
##  Mean   : 4.287   Mean   :10.876   Mean   : 29.409   Mean   :1513.8
##  3rd Qu.: 6.760   3rd Qu.:17.576   3rd Qu.: 45.698   3rd Qu.:1966.6
##  Max.   :18.922   Max.   :82.313   Max.   :358.061   Max.   :6444.1
##   married            ne                sc
##  Mode :logical   Mode :logical   Mode :logical
##  FALSE:8599      FALSE:22618     FALSE:20826
##  TRUE :20618     TRUE :6599      TRUE :8391
##
##
##
```

```r
x <- model.matrix( ~ -1 + female + widowed + divorced + separated + nevermarried +
hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3, data=cps2012)
dim(x)
```

```
## [1] 29217    16
```

```r
y <- cps2012$lnw
```

Question 2: Apply Ridge Regression with CV

Apply ridge regression to the previous dataset for the the default grid of values of lambda. Plot the MSE as a function of lambda. Then, select the optimal lambda by cross-validation. How many variables are used in the Ridge fit? Why the test MSE for Ridge is often smaller than for OLS when lambda is not zero? What is the optimal value of lambda? Is unrestricted OLS optimal here, in a test MSE sense?

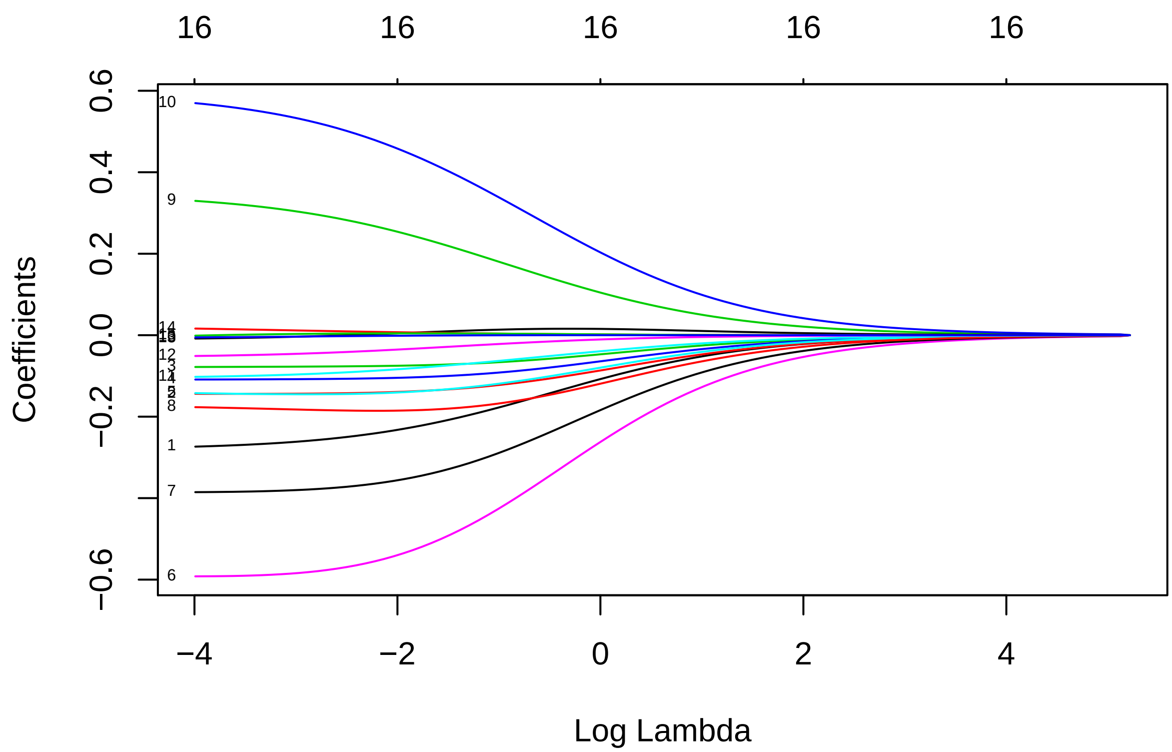
```r
library(glmnet)
```

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

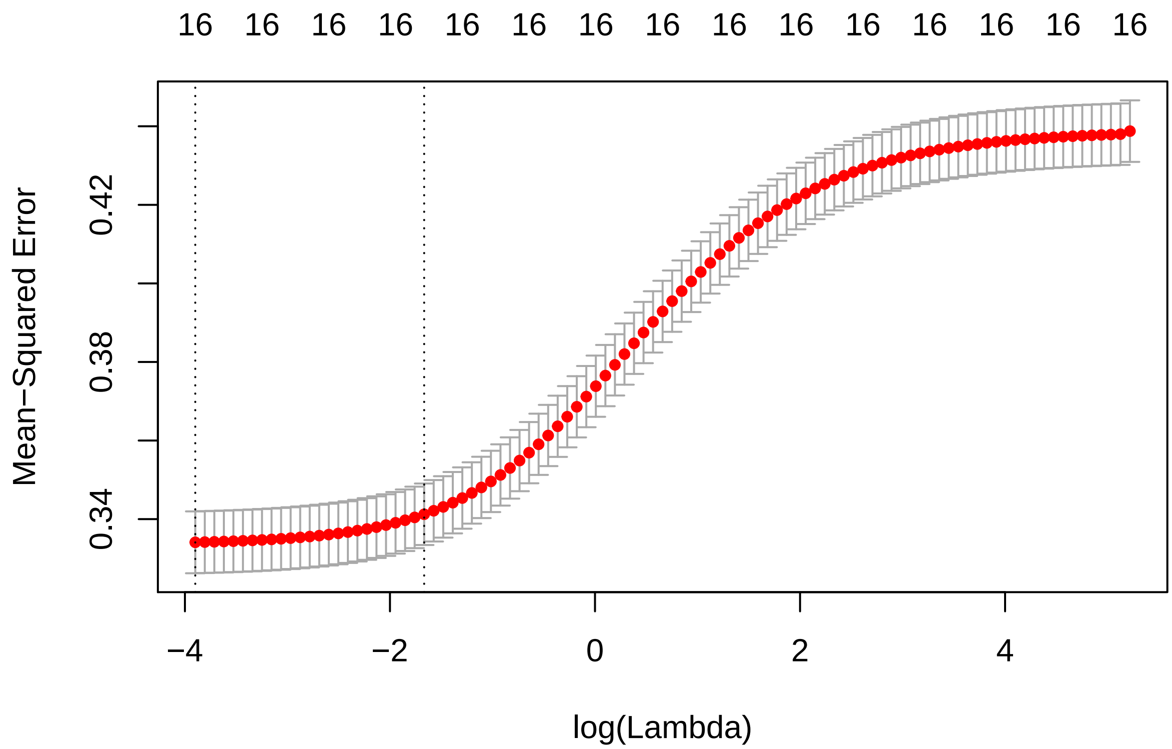
```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-13
```

```r
ridge.mod=glmnet(x,y,alpha=0)
plot(ridge.mod,xvar="lambda",label = "TRUE")
```




```r
cv.ridge=cv.glmnet(x,y,alpha=0)
plot(cv.ridge)
```




```r
cv.ridge$lambda.min
```

```
## [1] 0.02025893
```

```r
cv.OLS=cv.glmnet(x,y,alpha=0,lambda=c(0,cv.ridge$lambda.min))
MSEOLS=cv.OLS$cvm[1]
MSERidgeshort=cv.OLS$cvm[2]
MSEOLS
```

```
## [1] 0.3339791
```

```r
MSERidgeshort
```

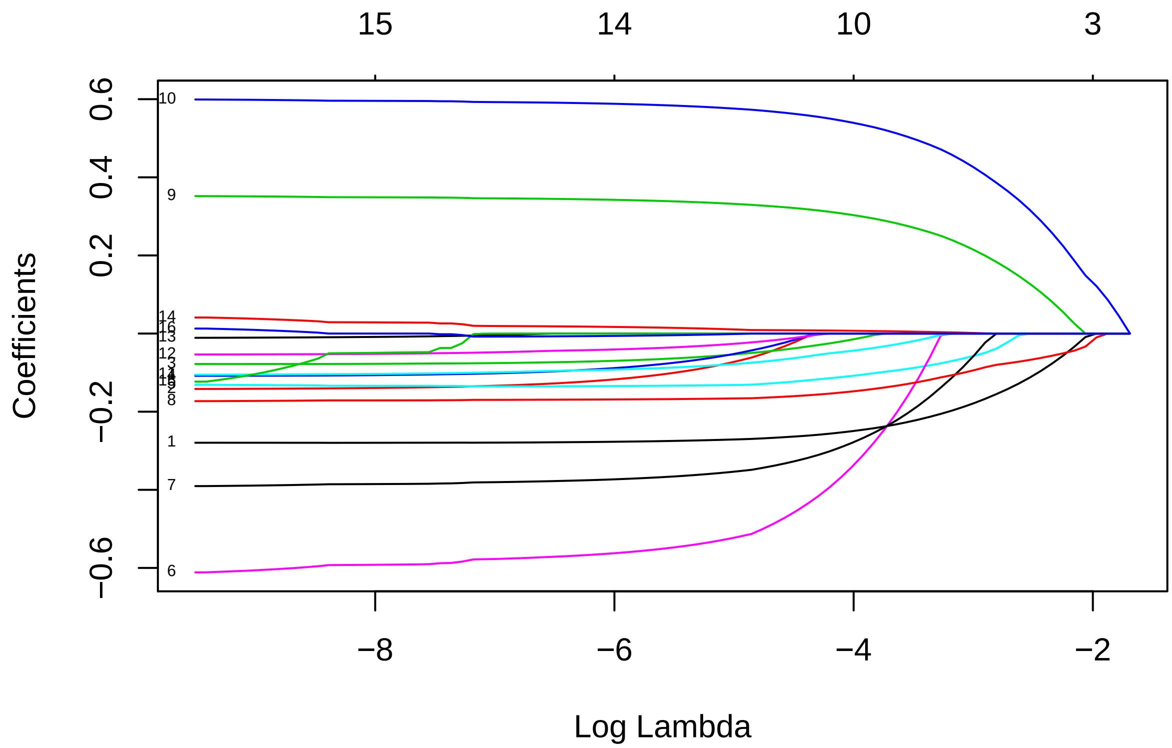
```
## [1] 0.33311
```

Ridge always includes all the variables (here 16). The test MSE for Ridge is different from OLS because it leads to different estimators. The optimal value of lambda is 0.02. Ridge reduces the variance of the estimator by skrinkage. The test MSE for Ridge is smaller than for OLS, so OLS is not optimal, although the difference is not substantial. In this application there is no motivation to do Ridge over OLS.

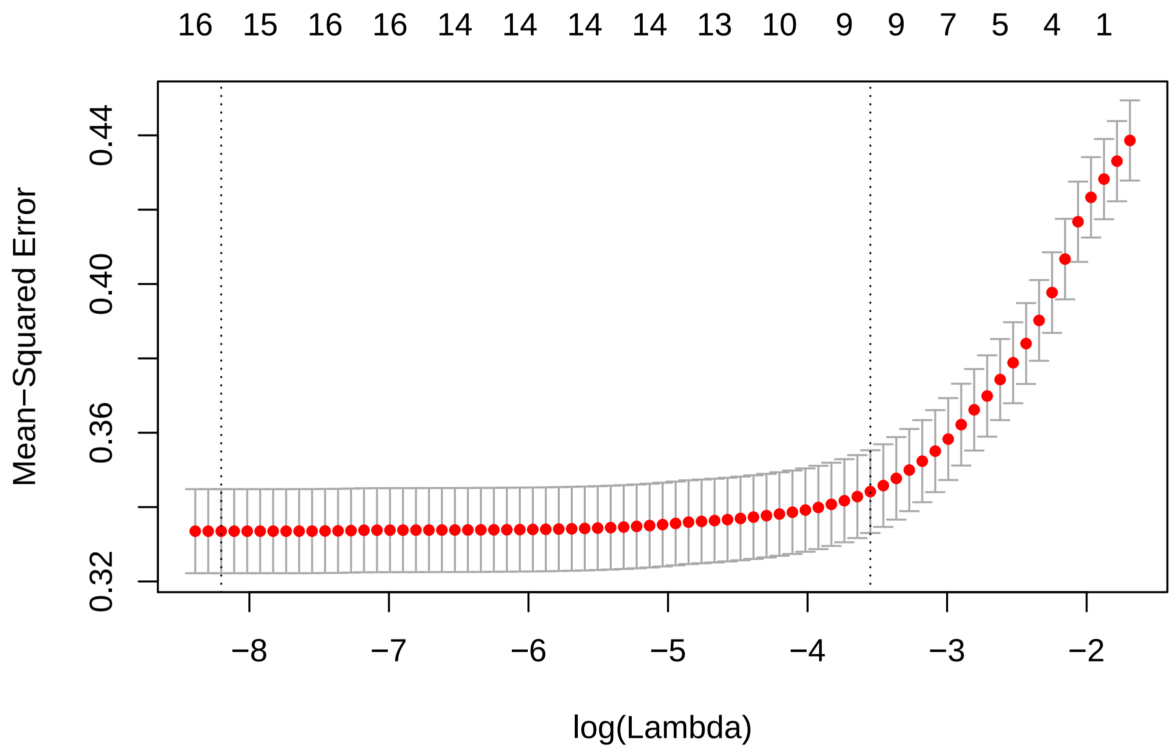
Question 3: Apply Lasso with CV

Apply Lasso regression to the previous dataset for the the default grid of values of lambda. Plot the MSE as a function of lambda. Then, select the optimal lambda by cross-validation. What is the optimal lambda? How many variables are used in the optimal Lasso fit? What are their coefficients? Is there a big difference here between Ridge and Lasso (in terms of test MSE)? Which method of prediction would you choose and why? Is gender an important factor in the prediction model? Interpret the coefficient of female.

```
lasso.mod=glmnet(x,y)  # default is alpha=1, Lasso
plot(lasso.mod,xvar="lambda",label = "TRUE")
```




```
cv.lasso=cv.glmnet(x,y)
plot(cv.lasso)
```




```
cv.ridge$lambda.min
```

```
## [1] 0.02025893
```

```
MSELassoshort=min(cv.lasso$cvm)
MSELassoshort
```

```
## [1] 0.3335186
```

```
coef(cv.lasso)
```

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                        1
## (Intercept)   2.698946572
## female       -0.226635194
## widowed       .
## divorced      .
## separated     .
## nevermarried -0.090441853
## hsd08        -0.160885871
## hsd911       -0.204266421
## hsg          -0.129086493
## cg            0.275341305
```

```
## ad            0.503838213
## mw           -0.022535368
## so            .
## we            .
## exp1          0.005145807
## exp2          .
## exp3          .
```

The optimal value of lambda is 0.02 and the variables selected are given above. Ridge and Lasso are very similar in terms of test MSE, so we prefer lasso as it has one less variable (a more parsimonius model). Female is one of the predictors selected. The interpretation of the coefficient of female is the gender gap: females make on average 23.7% less than males, everything else constant (holding other exploratory variables fixed).

Question 4: Making the model more flexible. Accounting for gender gaps.

Now you want to predict wages with a more flexible model that allows marginal effects that depend on gender. You would like to analyse the effect of gender and interaction effects of other variables with gender on wage jointly. The dependent variable is the logarithm of the wage. The new design matrix is given below. Repeat the previous analysis with this more flexible model.

```
X <- model.matrix( ~ -1 + female + female:(widowed + divorced + separated + nevermarried
hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3) +
+ (widowed + divorced + separated + nevermarried +
hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3)^2, data=cps2012)
dim(X)
```

```
## [1] 29217   136
```

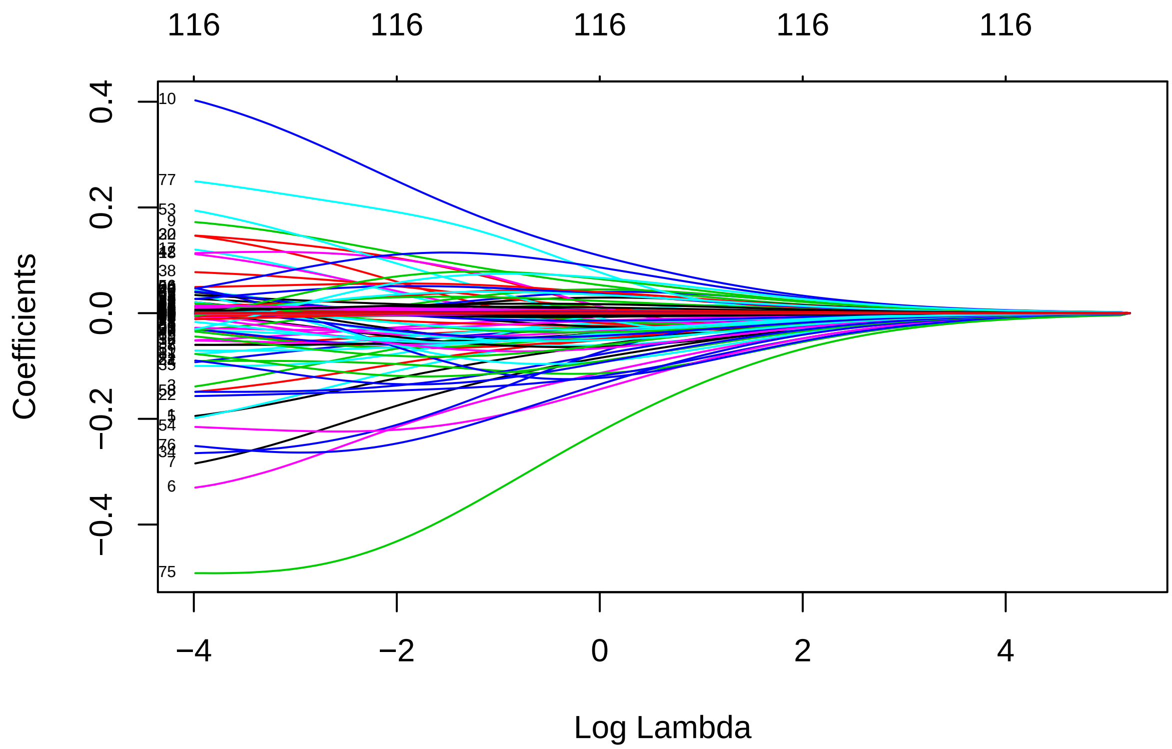
```
X <- X[,which(apply(X, 2, var)!=0)] # exclude all constant variables
dim(X)
```

```
## [1] 29217   116
```

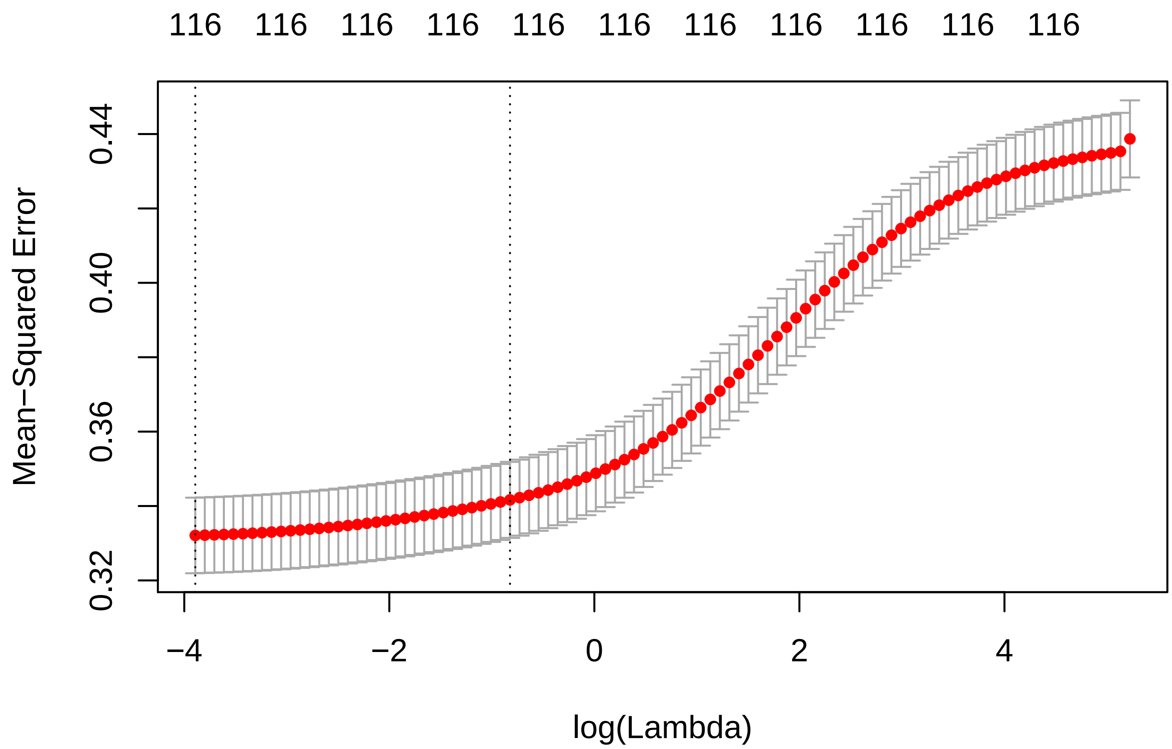
```
index.gender <- grep("female", colnames(X))
```

Ridge regression

```
ridge.mod=glmnet(X,y,alpha=0)
plot(ridge.mod,xvar="lambda",label = "TRUE")
```

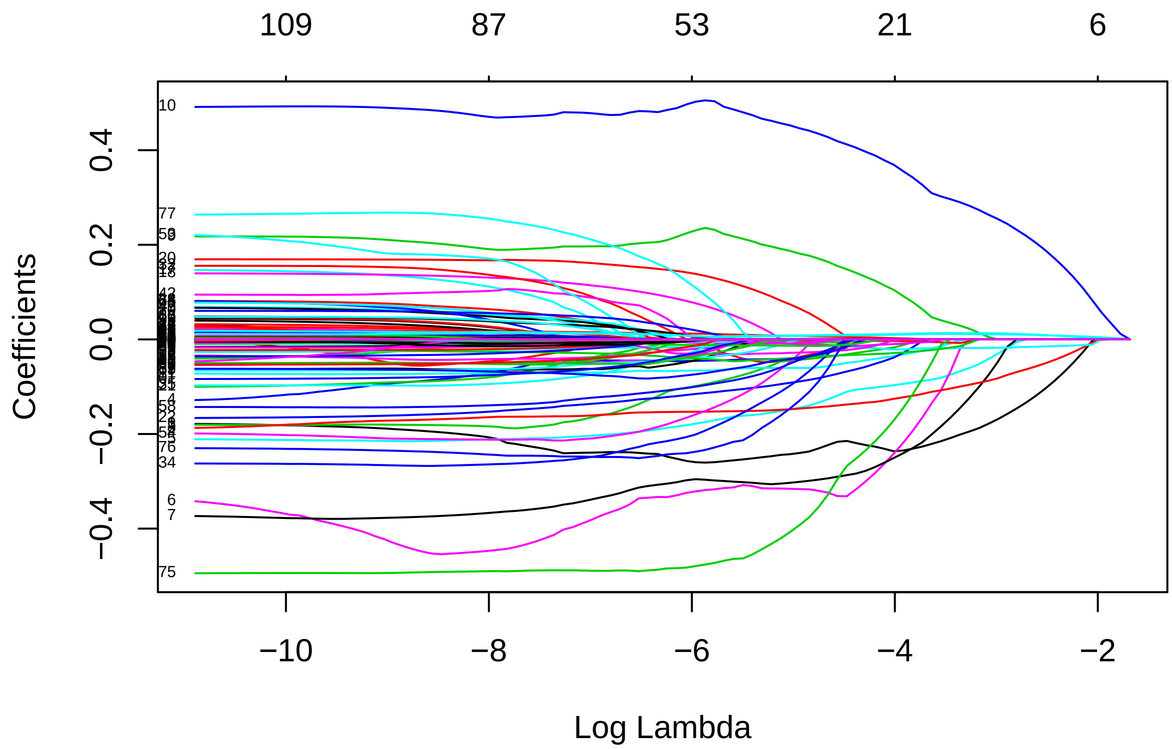



```r
cv.ridge=cv.glmnet(X,y,alpha=0)
MSERidgelong=min(cv.ridge$cvm)
plot(cv.ridge)
```

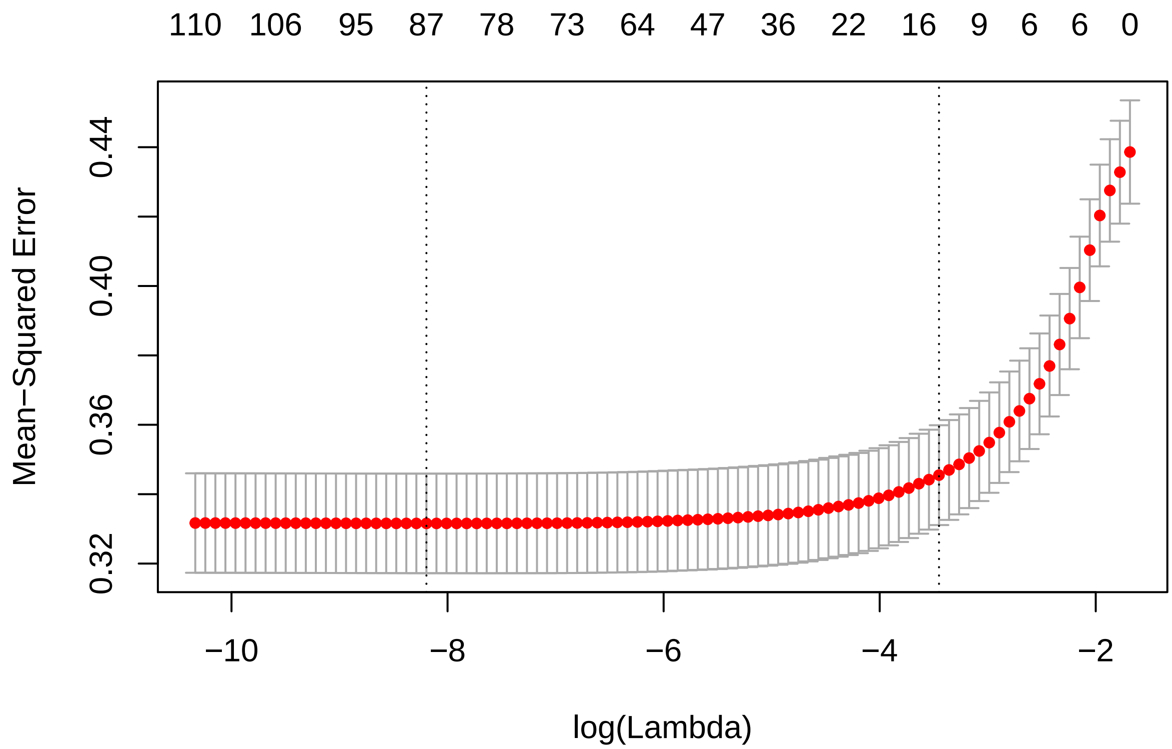


Lasso regression

```
lasso.mod=glmnet(X,y)  # default is alpha=1, Lasso
plot(lasso.mod,xvar="lambda",label = "TRUE")
```




```
cv.lasso=cv.glmnet(X,y)
MSELassolong=min(cv.lasso$cvm)
plot(cv.lasso)
```




```r
coef(cv.lasso)
```

```
## 117 x 1 sparse Matrix of class "dgCMatrix"
##                                1
## (Intercept)          2.787956155
## female              -0.208430736
## widowed              .
## divorced             .
## separated            .
## nevermarried        -0.074508921
## hsd08               -0.067247098
## hsd911              -0.165121222
## hsg                 -0.103832608
## cg                   0.034622499
## ad                   0.296329412
## mw                  -0.008314454
## so                   .
## we                   .
## exp1                 .
## exp2                 .
## exp3                 .
## female:widowed       .
```

```
## female:divorced        .
## female:separated       .
## female:nevermarried    .
## female:hsd08           .
## female:hsd911          .
## female:hsg           -0.018420791
## female:cg              .
## female:ad              .
## female:mw            -0.006845204
## female:so              .
## female:we              .
## female:exp1            .
## female:exp2            .
## female:exp3            .
## widowed:hsd911         .
## widowed:hsg            .
## widowed:cg             .
## widowed:ad             .
## widowed:mw             .
## widowed:so             .
## widowed:we             .
## widowed:exp1           .
## widowed:exp2           .
## widowed:exp3           .
## divorced:hsd08         .
## divorced:hsd911        .
## divorced:hsg           .
## divorced:cg            .
## divorced:ad            .
## divorced:mw            .
## divorced:so            .
## divorced:we            .
## divorced:exp1          .
## divorced:exp2          .
## divorced:exp3          .
## separated:hsd08        .
## separated:hsd911       .
## separated:hsg          .
## separated:cg           .
## separated:ad           .
## separated:mw           .
## separated:so           .
## separated:we           .
## separated:exp1         .
## separated:exp2         .
```

```
## separated:exp3       .
## nevermarried:hsd08   .
## nevermarried:hsd911  .
## nevermarried:hsg     .
## nevermarried:cg      .
## nevermarried:ad      .
## nevermarried:mw     -0.017305134
## nevermarried:so      .
## nevermarried:we      .
## nevermarried:exp1    .
## nevermarried:exp2    .
## nevermarried:exp3    .
## hsd08:mw             .
## hsd08:so             .
## hsd08:we             .
## hsd08:exp1           .
## hsd08:exp2           .
## hsd08:exp3           .
## hsd911:mw            .
## hsd911:so            .
## hsd911:we            .
## hsd911:exp1          .
## hsd911:exp2          .
## hsd911:exp3          .
## hsg:mw               .
## hsg:so               .
## hsg:we               .
## hsg:exp1             .
## hsg:exp2             .
## hsg:exp3             .
## cg:mw                .
## cg:so                .
## cg:we                .
## cg:exp1              0.013322646
## cg:exp2              .
## cg:exp3              .
## ad:mw                .
## ad:so                .
## ad:we                .
## ad:exp1              0.011513603
## ad:exp2              .
## ad:exp3              .
## mw:exp1              .
## mw:exp2              .
## mw:exp3              .
```

```
## so:exp1            .
## so:exp2            .
## so:exp3            .
## we:exp1            .
## we:exp2            .
## we:exp3            .
## exp1:exp2          .
## exp1:exp3          .
## exp2:exp3          .
```

Question 5: What is the preferred prediction method of all?

Do the effect of gender on wages depend on education? That is, are the interactions between gender and education important for prediction with Lasso (are they selected)?

```
MSERidgeshort
```

```
## [1] 0.33311
```

```
MSERidgelong
```

```
## [1] 0.3320909
```

```
MSELassoshort
```

```
## [1] 0.3335186
```

```
MSELassolong
```

```
## [1] 0.3315714
```

The preferred method is the Lasso for the long regression, and Lasso selects the interaction with female and hsg, so the gender gap depends on education. Again, in this application there is not much motivation for Lasso or Ridge in terms of test MSE but lasso leads to a model that is slightly simpler, without sacrificing predicting accurary.