problem_set_2b.R

Aaryan Agarwal

2023-10-20

```
# PROBLEM SET 2B
# This assignment is done in a group of three
# Prajwal Kaushal, Sumukha Sharma, Aaryan Agarwal

##############################################################################
#Question 1: Load, Prepare, and summarize the data
##############################################################################
library(hdm)
```

```
## Warning: package 'hdm' was built under R version 4.2.3
```

```
# This command loads data contained in a R-package.
data(cps2012)
?cps2012
```

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

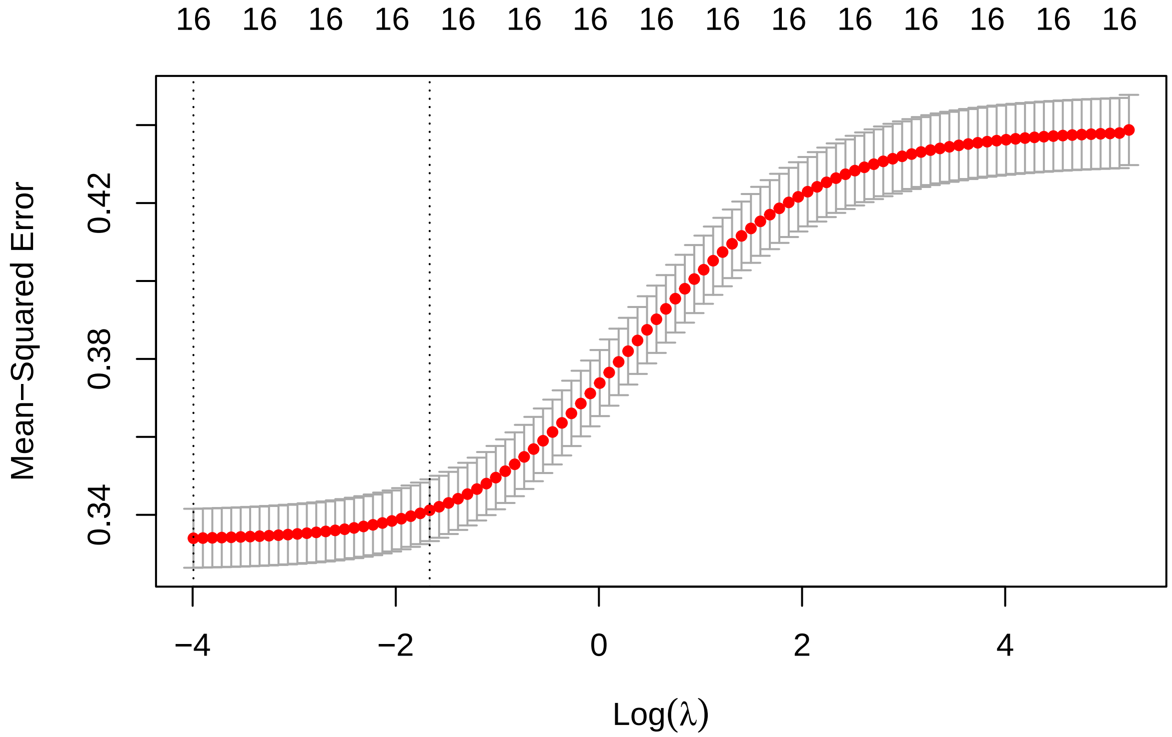
```
# Construct a regressor matrix for use in the different models.
x <- model.matrix( ~ -1 + female + widowed + divorced + separated + nevermarried +
                    hsd08+hsd911+ hsg+cg+ad+mw+so+we+exp1+exp2+exp3, data=cps2012)
dim(x)
```

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

```
y <- cps2012$lnw


##############################################################################
#Question 2: Apply Ridge Regression With CV
##############################################################################

library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.2.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 4.2.2

## Loaded glmnet 4.1-8
```

```r
# Ridge regression
cv.ridge <- cv.glmnet(x, y, alpha=0, nfolds=10)
plot(cv.ridge)
```



```r
# Optimal lambda
optimal_lambda_ridge <- cv.ridge$lambda.min
cat("Optimal Lambda for Ridge:", optimal_lambda_ridge, "\n")
```

```
## Optimal Lambda for Ridge: 0.01845918
```

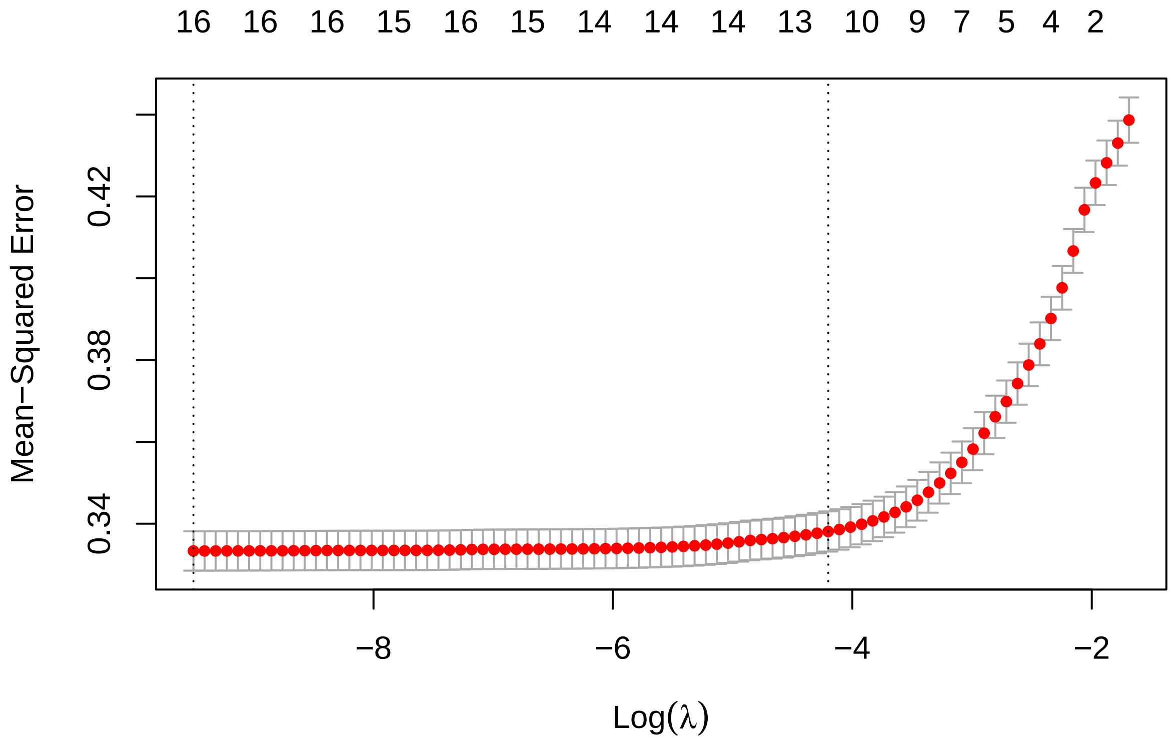
```r
# mse
ridge_optimal_mse <- cv.ridge$cvm[cv.ridge$lambda == optimal_lambda_ridge]
cat("Test MSE for Optimal Ridge:", ridge_optimal_mse, "\n")
```

```
## Test MSE for Optimal Ridge: 0.333992
```

```r
# Number of variables
cat("Number of Variables in Ridge Regression:", length(coef(cv.ridge, s = optimal_lambda_ridge)), "\n")
```

```
## Number of Variables in Ridge Regression: 17
```

```r
# length(coef(cv.ridge, s = optimal_lambda_ridge))
# essentially counts the number of coefficients, including the intercept,
# therefore it is 17 (16 predictors + 1 intercept).
# so there are 16 predictors.

# Ridge regression will include all predictors in the model;
# it shrinks the coefficients but doesn't set any to zero.
# So, the number of predictors remains the same as the original dataset, which is 16.

# Ridge regression adds a penalty to the size of coefficients,
# which can prevent overfitting, especially when predictors are correlated.
# This regularization can lead to better out-of-sample predictions,
# reducing the test MSE compared to OLS, which doesn't have this penalty.

# When lambda is zero, the penalty term disappears, and Ridge regression becomes
# equivalent to OLS regression. That is, there's no regularization or shrinkage applied
# to the coefficients.

# If the test MSE at the optimal lambda for Ridge regression is lower than the MSE
# for the Ridge regression at lambda = 0 (which is equivalent to OLS), then unrestricted OLS
# is not optimal in terms of test MSE.


###############################################################################
#Question 3: Apply Lasso Regression With CV
###############################################################################

# Lasso regression
cv.lasso <- cv.glmnet(x, y, alpha=1, nfolds=10)
plot(cv.lasso)
```



```
# Optimal lambda
optimal_lambda_lasso <- cv.lasso$lambda.min
cat("Optimal Lambda for Lasso:", optimal_lambda_lasso, "\n")
```

```
## Optimal Lambda for Lasso: 7.452004e-05
```

```
# Coefficients at optimal lambda
coefficients_lasso <- coef(cv.lasso, s = optimal_lambda_lasso)

# Variables used
non_zero_coef <- coefficients_lasso[coefficients_lasso != 0]
```

```
## <sparse>[ <logic> ]: .M.sub.i.logical() maybe inefficient
```

```
cat("Number of Variables in Optimal Lasso Fit:", length(non_zero_coef) - 1, "\n")  # Subtracting interc
```

```
## Number of Variables in Optimal Lasso Fit: 16
```

```
print(non_zero_coef)
```

```
##  [1]  2.47808273 -0.27952186 -0.14187613 -0.07797162 -0.10819852 -0.13125124
##  [7] -0.61142487 -0.39049171 -0.17297117  0.35183459  0.59916083 -0.10540074
## [13] -0.05369308 -0.01088479  0.04101563 -0.12316500  0.01275232
```

```r
lasso_optimal_mse <- cv.lasso$cvm[cv.lasso$lambda == optimal_lambda_lasso]
cat("Test MSE for Optimal Lasso:", lasso_optimal_mse, "\n")
```

```
## Test MSE for Optimal Lasso: 0.3333426
```

```r
# Difference in MSE between Ridge and Lasso
difference_mse <- lasso_optimal_mse - ridge_optimal_mse
cat("Difference in MSE between Ridge and Lasso:", difference_mse, "\n")
```

```
## Difference in MSE between Ridge and Lasso: -0.0006493848
```

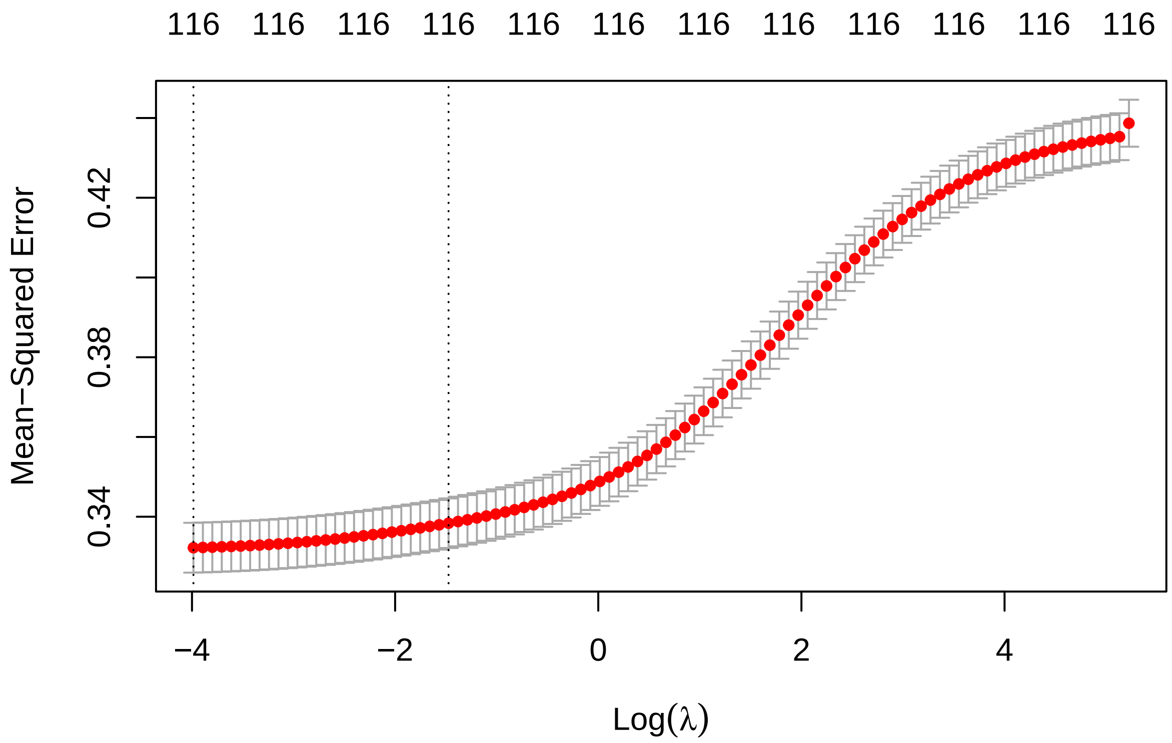
```r
# for this specific dataset and with the chosen predictor variables,
# the Lasso regression model generalizes slightly better to new,
# unseen data than the Ridge regression model. However, it's worth noting that
# the difference is quite small.
# In practical terms, the models might offer similar predictive performance.

# Since the dependent variable is the logarithm of the hourly wage,
# the coefficient for female indicates that being female, on average,
# is associated with a decrease in the logged hourly wage of about 0.28 units
# compared to the baseline (which is male).
# Which means, being a female is associated with a wage that is exp(-0.27952186)
# or roughly 75.6% of the wage for males, holding all else constant.
# This coefficient suggests a wage gap where females earn, on average, less than males
# in this dataset.

################################################################################
#Question 4: Using a 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
```

```r
# Safety check: Exclude all constant variables.
X <- X[,which(apply(X, 2, var)!=0)]
dim(X)
```

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

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

# The provided code created a design matrix X that not only includes the main
# effects of all predictors but also interaction effects of gender with all other
# variables and squared terms for predictors.

# This new design matrix initially has 136 variables.
# After removing constant variables, 116 remain.

# Ridge regression for new X
cv.ridge <- cv.glmnet(X, y, alpha=0, nfolds=10)
plot(cv.ridge)
```



```r
# Optimal lambda
optimal_lambda_ridge <- cv.ridge$lambda.min
cat("Optimal Lambda for Ridge:", optimal_lambda_ridge, "\n")
```

```
## Optimal Lambda for Ridge: 0.01857752
```

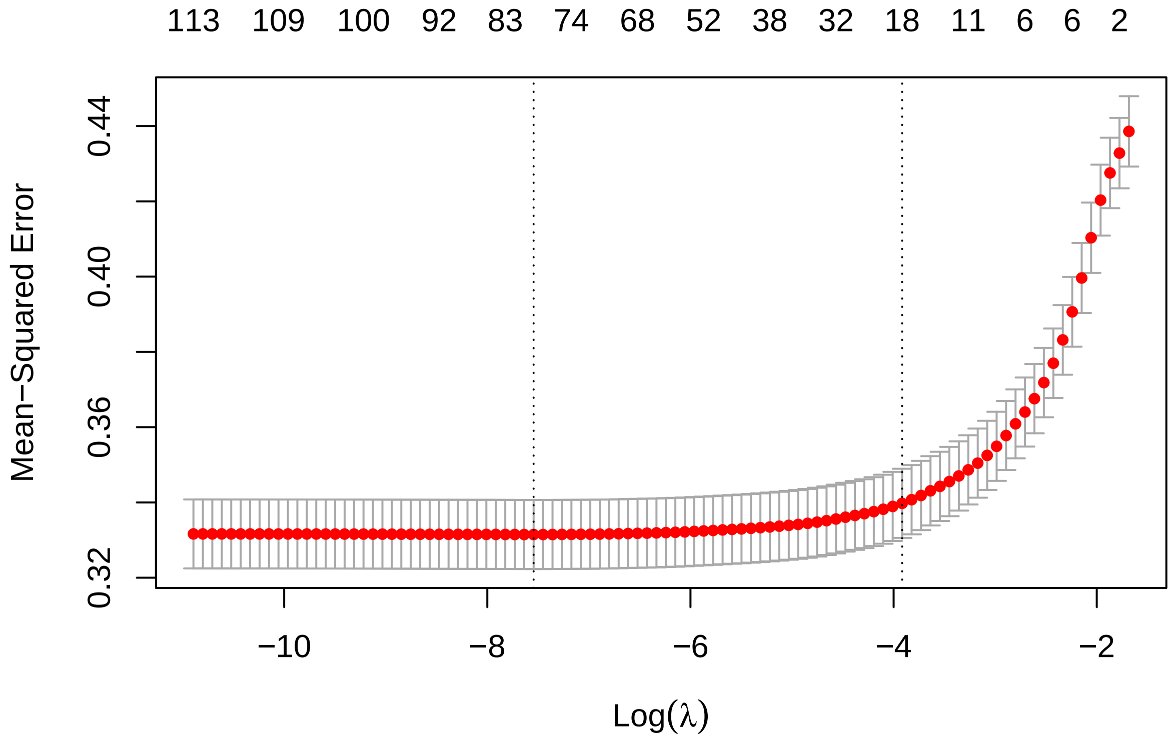
```r
# mse
ridge_flexible_mse <- cv.ridge$cvm[cv.ridge$lambda == optimal_lambda_ridge]
cat("Test MSE for Optimal Ridge:", ridge_flexible_mse, "\n")
```

```
## Test MSE for Optimal Ridge: 0.3322257
```

```r
# Number of variables
cat("Number of Variables in Ridge Regression:", length(coef(cv.ridge, s = optimal_lambda_ridge)), "\n")
```

```
## Number of Variables in Ridge Regression: 117
```

```
# Lasso regression for new X

cv.lasso <- cv.glmnet(X, y, alpha=1, nfolds=10)
plot(cv.lasso)
```



```
# Optimal lambda
optimal_lambda_lasso <- cv.lasso$lambda.min
cat("Optimal Lambda for Lasso:", optimal_lambda_lasso, "\n")
```

```
## Optimal Lambda for Lasso: 0.0005290945
```

```
# Coefficients at optimal lambda
coefficients_lasso <- coef(cv.lasso, s = optimal_lambda_lasso)

# Variables used
non_zero_coef <- coefficients_lasso[coefficients_lasso != 0]
```

```
## <sparse>[ <logic> ]: .M.sub.i.logical() maybe inefficient
```

```
cat("Number of Variables in Optimal Lasso Fit:", length(non_zero_coef) - 1, "\n")  # Subtracting interc
```

```
## Number of Variables in Optimal Lasso Fit: 78
```

```r
print(non_zero_coef)
```

```
##  [1]  2.613287e+00 -2.279558e-01 -3.508805e-02 -1.844256e-01 -5.076617e-02
##  [6] -2.092132e-01 -4.275747e-01 -3.584893e-01 -1.632006e-01  1.919077e-01
## [11]  4.722729e-01 -7.159128e-02 -4.088920e-02  1.638364e-02 -3.717957e-05
## [16] -4.513593e-03  9.059483e-02  1.255438e-01  1.313238e-02  1.670227e-01
## [21] -6.419766e-02 -1.461957e-01 -2.046883e-02  7.860327e-03 -1.006526e-02
## [26] -5.040741e-03 -5.252955e-03  1.354585e-03  1.222451e-01 -2.597985e-01
## [31] -8.813695e-02 -4.937114e-02 -6.362154e-02  5.482176e-02 -3.038139e-03
## [36]  1.022040e-01  1.016307e-02 -1.633908e-04  4.684749e-02  4.277049e-02
## [41]  1.298321e-03  1.414897e-01 -2.113741e-01  7.529357e-03  1.475375e-02
## [46] -5.316187e-02 -1.355404e-01 -6.082792e-02 -1.182389e-02  2.755023e-05
## [51]  2.039321e-02  4.629085e-02  3.374892e-02 -2.720568e-02  5.274204e-02
## [56] -2.149931e-03 -4.884956e-01 -2.459638e-01  2.397755e-01  2.488484e-03
## [61] -5.069295e-02 -7.006582e-02  3.501034e-02  2.828965e-03  1.464603e-02
## [66]  7.377531e-03 -1.851796e-02 -2.576040e-02  1.289222e-02 -7.781601e-03
## [71] -4.632816e-02 -7.074875e-02  1.500384e-02 -1.380197e-02 -1.583595e-03
## [76] -5.952074e-04 -7.675553e-04  1.916314e-04 -3.068724e-05
```

```r
#print(coefficients_lasso)

lasso_flexible_mse <- cv.lasso$cvm[cv.lasso$lambda == optimal_lambda_lasso]
cat("Test MSE for Optimal Lasso:", lasso_flexible_mse, "\n")
```

```
## Test MSE for Optimal Lasso: 0.3314723
```

```r
# Difference in MSE between Ridge and Lasso
difference_mse <- lasso_flexible_mse - ridge_flexible_mse
cat("Difference in MSE between Ridge and Lasso:", difference_mse, "\n")
```

```
## Difference in MSE between Ridge and Lasso: -0.0007533859
```

```r
# Lasso regression has chosen 78 variables out of the total 116,
# implying that lasso selected 38 variables to be not influential
# in predicting wages in the context of other predictors.

index.gender
```

```
##  [1]  1 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31
```

```r
#this shows that there are interaction term present with gender


###############################################################################
#Question 5: What is the most preferred prediction model of all?
###############################################################################


cat("Ridge MSE:", ridge_optimal_mse, "\n")
```

```
## Ridge MSE: 0.333992
```

```r
cat("Lasso MSE:", lasso_optimal_mse, "\n")
```

```
## Lasso MSE: 0.3333426
```

```r
cat("Flexible Ridge MSE:", ridge_flexible_mse, "\n")
```

```
## Flexible Ridge MSE: 0.3322257
```

```r
cat("Flexible Lasso MSE:", lasso_flexible_mse, "\n")
```

```
## Flexible Lasso MSE: 0.3314723
```

```r
# here we can see that flexible lasso is the best model

print(coefficients_lasso)
```

```
## 117 x 1 sparse Matrix of class "dgCMatrix"
##                                 s1
## (Intercept)           2.613287e+00
## female               -2.279558e-01
## widowed              -3.508805e-02
## divorced             -1.844256e-01
## separated            -5.076617e-02
## nevermarried         -2.092132e-01
## hsd08                -4.275747e-01
## hsd911               -3.584893e-01
## hsg                  -1.632006e-01
## cg                    1.919077e-01
## ad                    4.722729e-01
## mw                   -7.159128e-02
## so                   -4.088920e-02
## we                     .
## exp1                  1.638364e-02
## exp2                 -3.717957e-05
## exp3                 -4.513593e-03
## female:widowed        9.059483e-02
## female:divorced       1.255438e-01
## female:separated      1.313238e-02
## female:nevermarried   1.670227e-01
## female:hsd08         -6.419766e-02
## female:hsd911        -1.461957e-01
## female:hsg           -2.046883e-02
## female:cg             7.860327e-03
## female:ad            -1.006526e-02
## female:mw              .
## female:so            -5.040741e-03
## female:we              .
## female:exp1          -5.252955e-03
## female:exp2            .
## female:exp3           1.354585e-03
## widowed:hsd911        1.222451e-01
## widowed:hsg          .
## widowed:cg          -2.597985e-01
## widowed:ad          -8.813695e-02
## widowed:mw          -4.937114e-02
## widowed:so          -6.362154e-02
## widowed:we           5.482176e-02
## widowed:exp1        -3.038139e-03
## widowed:exp2         .
## widowed:exp3         .
## divorced:hsd08       1.022040e-01
## divorced:hsd911      .
## divorced:hsg         1.016307e-02
## divorced:cg         -1.633908e-04
## divorced:ad          .
## divorced:mw          4.684749e-02
## divorced:so          .
## divorced:we          4.277049e-02
## divorced:exp1        .
## divorced:exp2        .
## divorced:exp3        1.298321e-03
## separated:hsd08      1.414897e-01
## separated:hsd911    -2.113741e-01
## separated:hsg        7.529357e-03
## separated:cg         1.475375e-02
## separated:ad        -5.316187e-02
## separated:mw        -1.355404e-01
## separated:so        -6.082792e-02
## separated:we        -1.182389e-02
## separated:exp1       .
## separated:exp2       2.755023e-05
## separated:exp3       .
## nevermarried:hsd08   2.039321e-02
## nevermarried:hsd911  4.629085e-02
## nevermarried:hsg     .
## nevermarried:cg      3.374892e-02
## nevermarried:ad      .
## nevermarried:mw     -2.720568e-02
## nevermarried:so      5.274204e-02
## nevermarried:we      .
## nevermarried:exp1    .
## nevermarried:exp2    .
## nevermarried:exp3   -2.149931e-03
## hsd08:mw            -4.884956e-01
## hsd08:so            -2.459638e-01
## hsd08:we             2.397755e-01
## hsd08:exp1           .
## hsd08:exp2           .
## hsd08:exp3           2.488484e-03
## hsd911:mw           -5.069295e-02
## hsd911:so           -7.006582e-02
## hsd911:we            3.501034e-02
## hsd911:exp1          .
## hsd911:exp2          .
## hsd911:exp3          2.828965e-03
## hsg:mw                1.464603e-02
## hsg:so                .
## hsg:we                7.377531e-03
## hsg:exp1              .
## hsg:exp2              .
## hsg:exp3              .
## cg:mw                -1.851796e-02
## cg:so                 .
## cg:we                -2.576040e-02
## cg:exp1               1.289222e-02
## cg:exp2               .
## cg:exp3              -7.781601e-03
## ad:mw                -4.632816e-02
## ad:so                 .
## ad:we                -7.074875e-02
## ad:exp1               1.500384e-02
## ad:exp2               .
## ad:exp3              -1.380197e-02
## mw:exp1               .
## mw:exp2               .
## mw:exp3              -1.583595e-03
## so:exp1               .
## so:exp2              -5.952074e-04
## so:exp3              -7.675553e-04
## we:exp1               .
## we:exp2               .
## we:exp3               1.916314e-04
## exp1:exp2            -3.068724e-05
## exp1:exp3             .
## exp2:exp3             .
```

```
# female:hsd08: The interaction between females and education level hsd08 is negative,
# suggesting that females with the hsd08 level of education might earn 0.0784 (or about 7.84%)
# less on the log wage scale compared to the reference group, holding other predictors constant.

# female:hsd911: Females with education level hsd911 have a negative coefficient of 0.1517
# (or about 15.17%), indicating they might earn less compared to the reference group.

# female:hsg: Females with high school graduation (hsg) education level have a negative
# coefficient of 0.0226 (or about 2.26%), suggesting a wage reduction compared to the reference group.

# female:cg: The interaction coefficient for females with the cg level of education
# is positive but quite small, suggesting a very slight increase in wages compared
# to the reference group.

# female:ad: Females with the ad level of education have a negative coefficient of 0.0115
# (or about 1.15%), indicating a slight wage reduction compared to the reference group.

# Conclusion:
# The flexible Lasso regression selected the interaction terms between gender (female)
# and various education levels. The majority of these interactions suggest that females
# with these education levels have a reduction in wages when compared to the reference group
# (which would be males with the same education level.
# The only exception is the cg education level, which indicates a slight wage
# advantage for females.
```