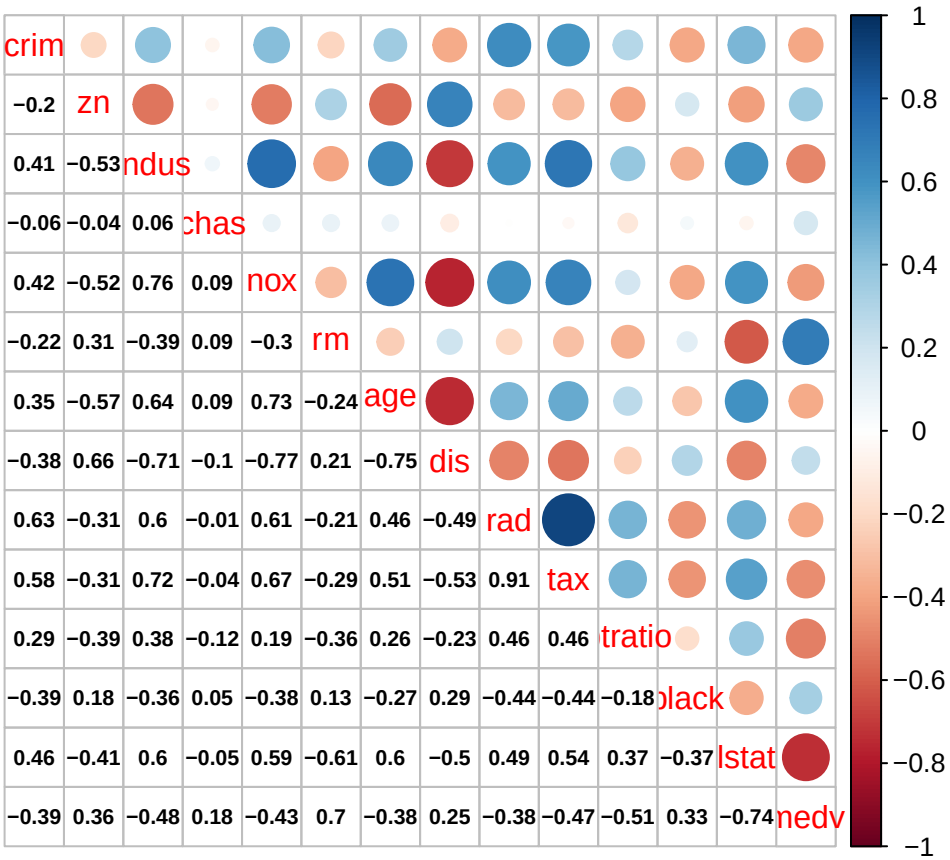
Empirical Methods for Applied Micro

Problem Set 3

Alberto Cappello

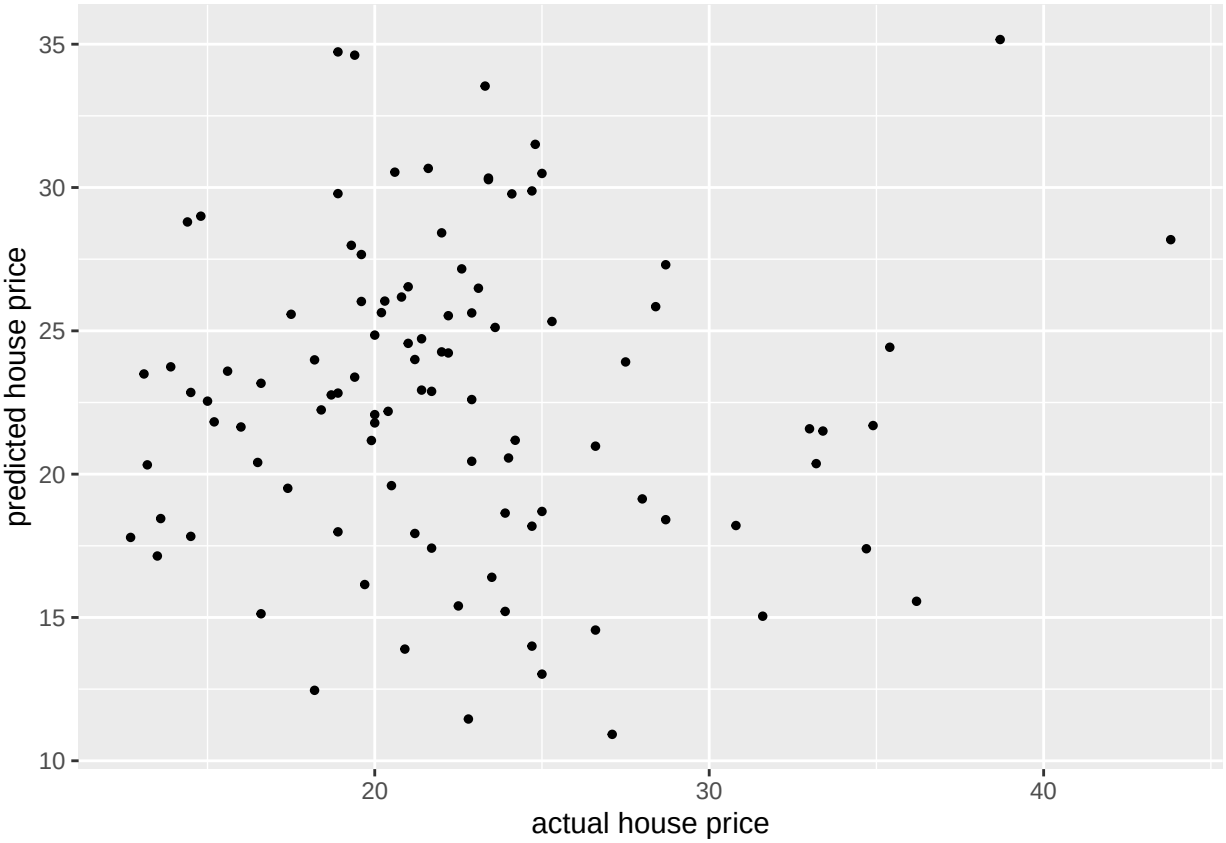
2/13/2021

1. How correlated are the variables?



Several variable like *indus*, *nox*,*stat* and *tax* show high degree of correlation with most of the other variables. Therefore if we run a simple linear regression many coefficients may be poorly identified because of the high multicollinearity across variables.

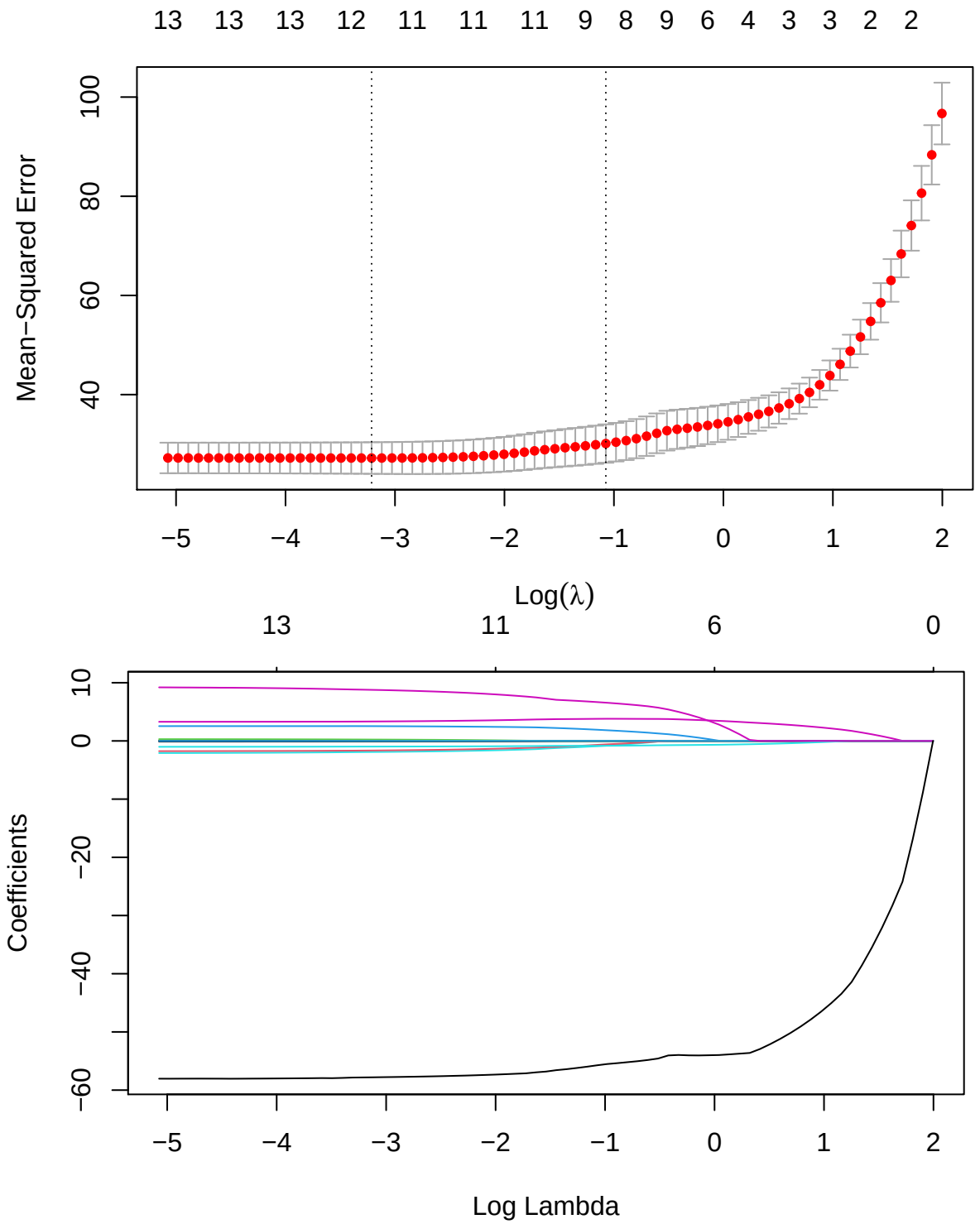
2. Estimate the original HR



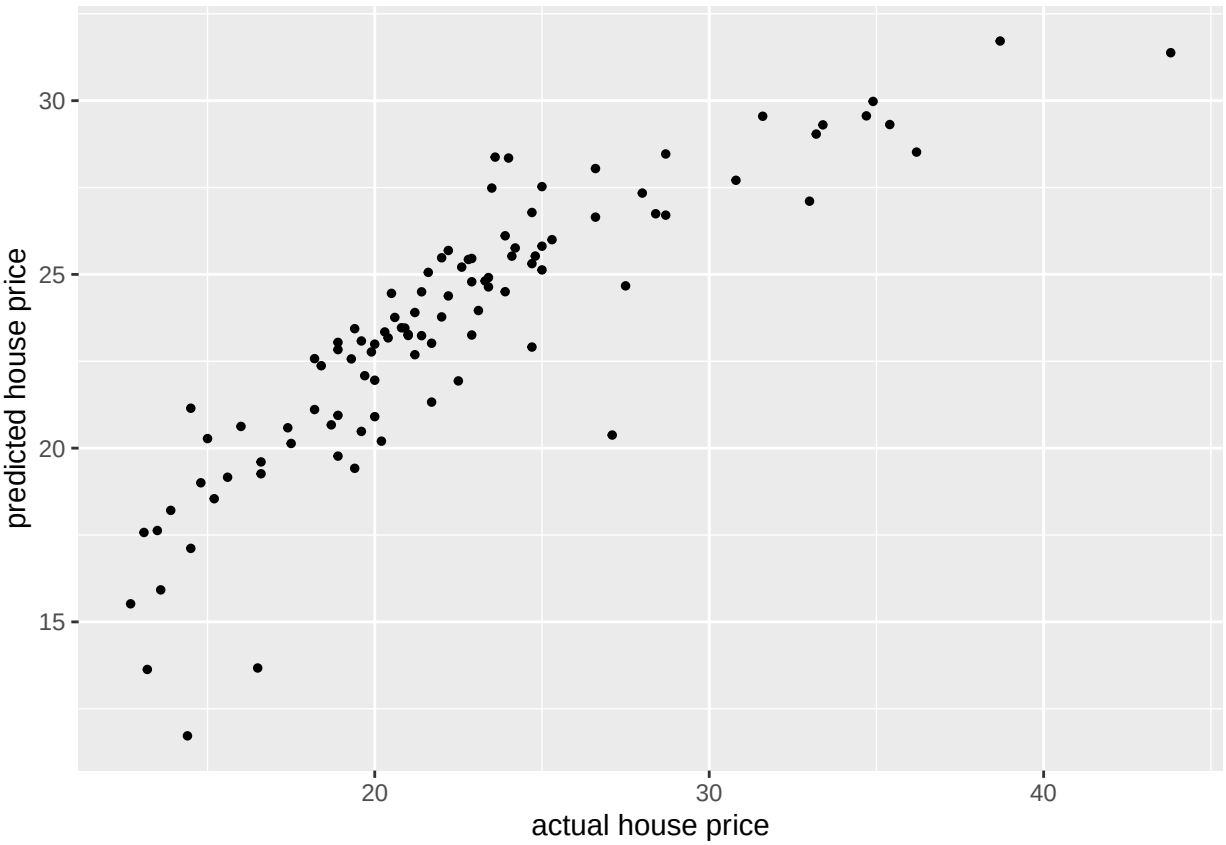
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.24	0.17	24.88	0.00
crim	-0.01	0.00	-7.37	0.00
zn	0.00	0.00	2.22	0.03
indus	0.00	0.00	1.29	0.20
chas	0.10	0.04	2.62	0.01
nox	-0.01	0.00	-5.04	0.00
rm	0.01	0.00	4.58	0.00
age	-0.00	0.00	-0.17	0.87
dis	-0.05	0.01	-5.35	0.00
rad	0.01	0.00	4.75	0.00
tax	-0.00	0.00	-3.91	0.00
ptratio	-0.04	0.01	-5.95	0.00
black	0.41	0.11	3.54	0.00
lstat	-3.01	0.23	-13.14	0.00

As one can see from the table above most of the estimated coefficients are close to zero apart from *lstat* which is quite large and statistically significant coefficient. Not surprisingly *lstat* is highly correlated with all the other variables in our dataset, so probably the coefficient is poorly estimated due to the multicollinearity problem.

3. LASSO Regression



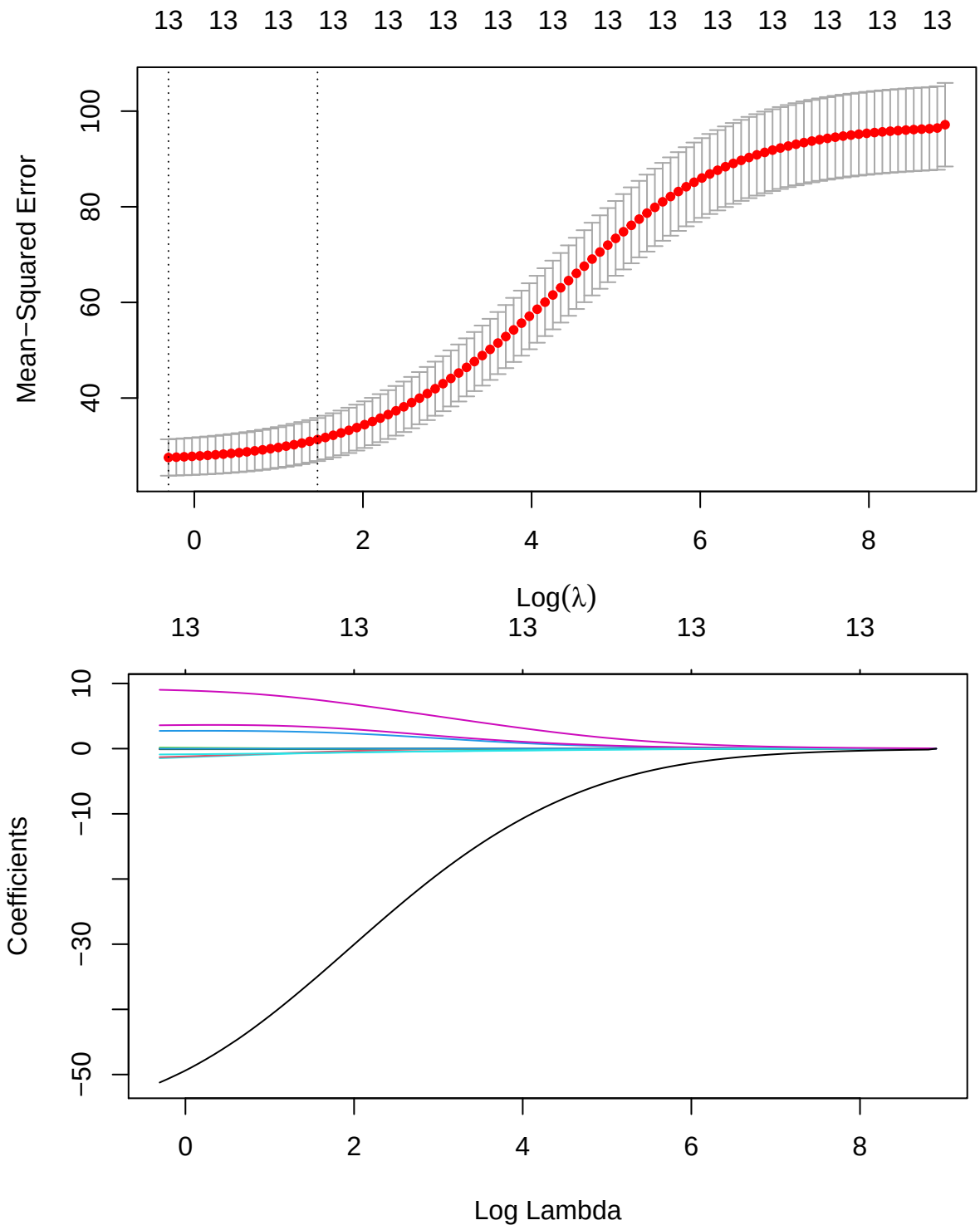
As one ca see from the plots above under the optimal value of $\lambda = 0.875$ (i.e. the maximum value of lambda among those value such that the estimated MSE is within 1 std from the minimum MSE) 8 out of 13 coefficients are shrunk to zero.



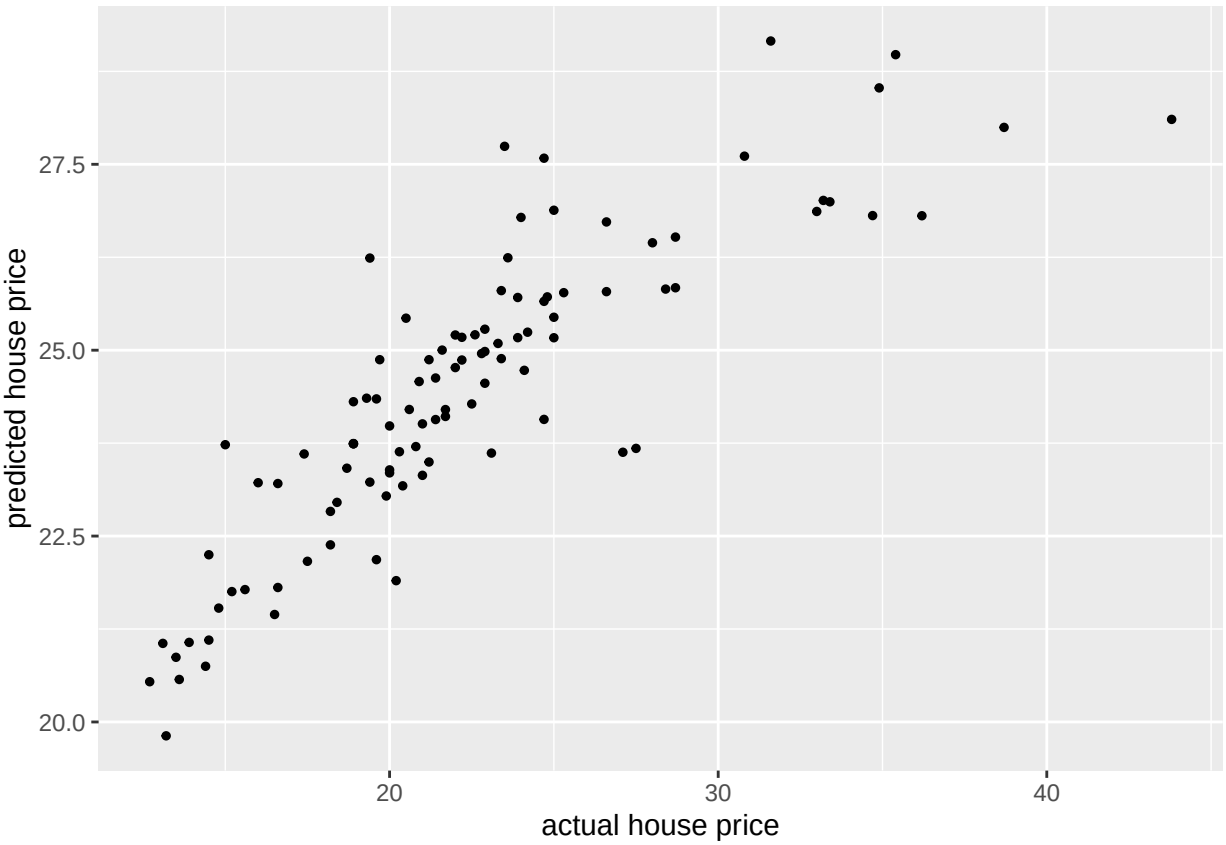
	HR	LASSO
(Intercept)	4.24	18.79
crim	-0.01	0.00
zn	0.00	0.00
indus	0.00	0.00
chas	0.10	0.00
nox	-0.01	0.00
rm	0.01	2.61
age	-0.00	0.00
dis	-0.05	0.00
rad	0.01	0.00
tax	-0.00	0.00
ptratio	-0.04	-0.34
black	0.41	0.00
lstat	-3.01	-49.07

From the table above we notice that there are 3 variables that are the most relevant to estimate the dependent variable: *lstat*, *rm* and *ptratio*. Simple linear regression in the HR model was underestimating the impact of these variables. LASSO regression allow us to largely improve our predictions on the test set as we can observe from the scatterplot above.

3. Ridge Regression



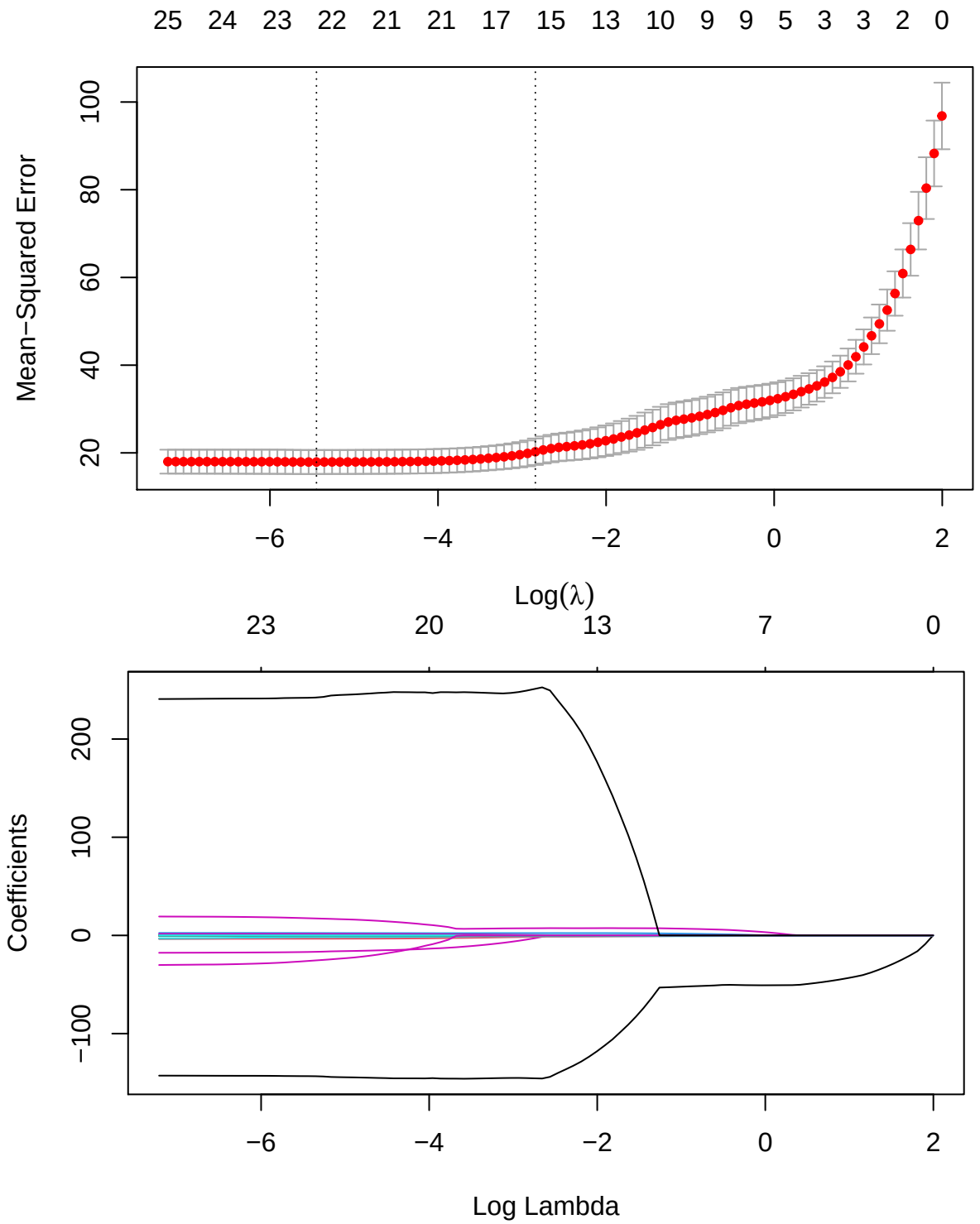
In this case the cross validated value of the penalization parameter is larger than under LASSO, but the estimated MSE is very close to the one obtained with LASSO regression. AS expected most of the coefficient are shrunk towards zero, but not exactly zero.



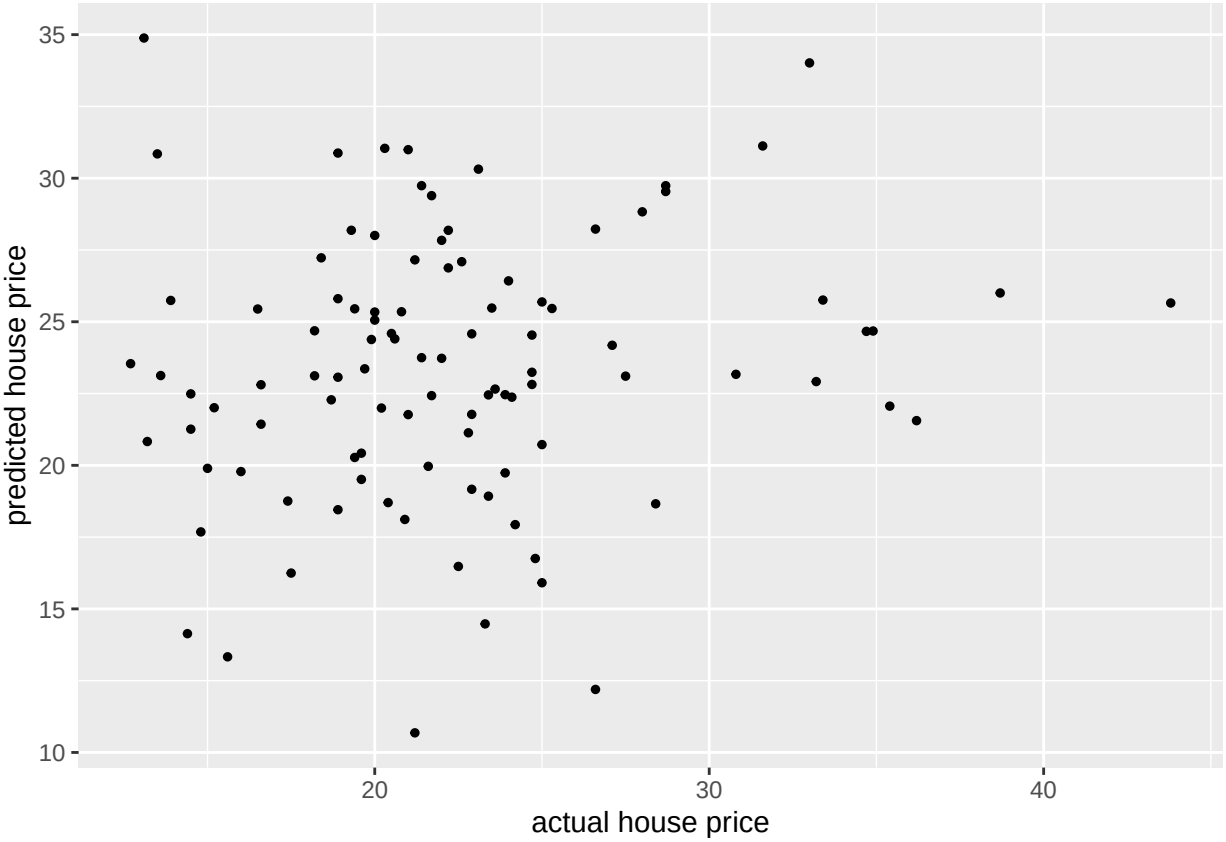
	HR	LASSO	RIDGE
(Intercept)	4.24	18.79	24.85
crim	-0.01	0.00	-0.04
zn	0.00	0.00	0.01
indus	0.00	0.00	-0.07
chas	0.10	0.00	1.05
nox	-0.01	0.00	-0.34
rm	0.01	2.61	1.30
age	-0.00	0.00	-0.01
dis	-0.05	0.00	0.02
rad	0.01	0.00	-0.03
tax	-0.00	0.00	-0.00
ptratio	-0.04	-0.34	-0.29
black	0.41	0.00	3.67
lstat	-3.01	-49.07	-13.01

The estimated coefficients are quite different compared to LASSO. *lstat* is still the variable with the largest coefficient, but under Ridge regression also *black*, *rm* and *chas* are well above zero in absolute value while being zero under LASSO. However, the scatterplot looks very similar to the one obtained with LASSO, so the benefit in the terms of prediction accuracy that we get relaxing the penalization is negligible. (Recall that Ridge tend to shrink coefficients to zero, whereas LASSO perform variable selection).

4. LASSO on expanded data set



In these case at the cross validated λ of LASSO under the extended dataset is much larger than that of LASSO on the original dataset without squared terms. Most of the coefficients are shrunk to zero, but more than one coefficient survive the penalization: *lstat* is the largest in absolute terms, but also *ptratio*, *black* and rm^2 are well above zero.



	LASSO2
intercept	26.75
ptratio	-0.54
black	1.43
lstat	-50.67
rm_sq	0.30

Altough the estimated MSE is slightly lower than the one under original LASSO, the performance on the test set is much worse. This is probably due to the fact that the squared terms are making the multicollinearity problem worse and the selected penalization terms is not enough to shrunk to zero some of the coefficients that actually survive.

5. Models' comparison

If we use the test set MSE to compare the models discussed above, the best model is LASSO.

	lambda	cv mse	test mse
original HR	0.00	0.20	20.09
LASSO	2.20	40.44	11.44
Ridge	40.21	52.86	22.84
LASSO 2	1.26	33.34	48.99