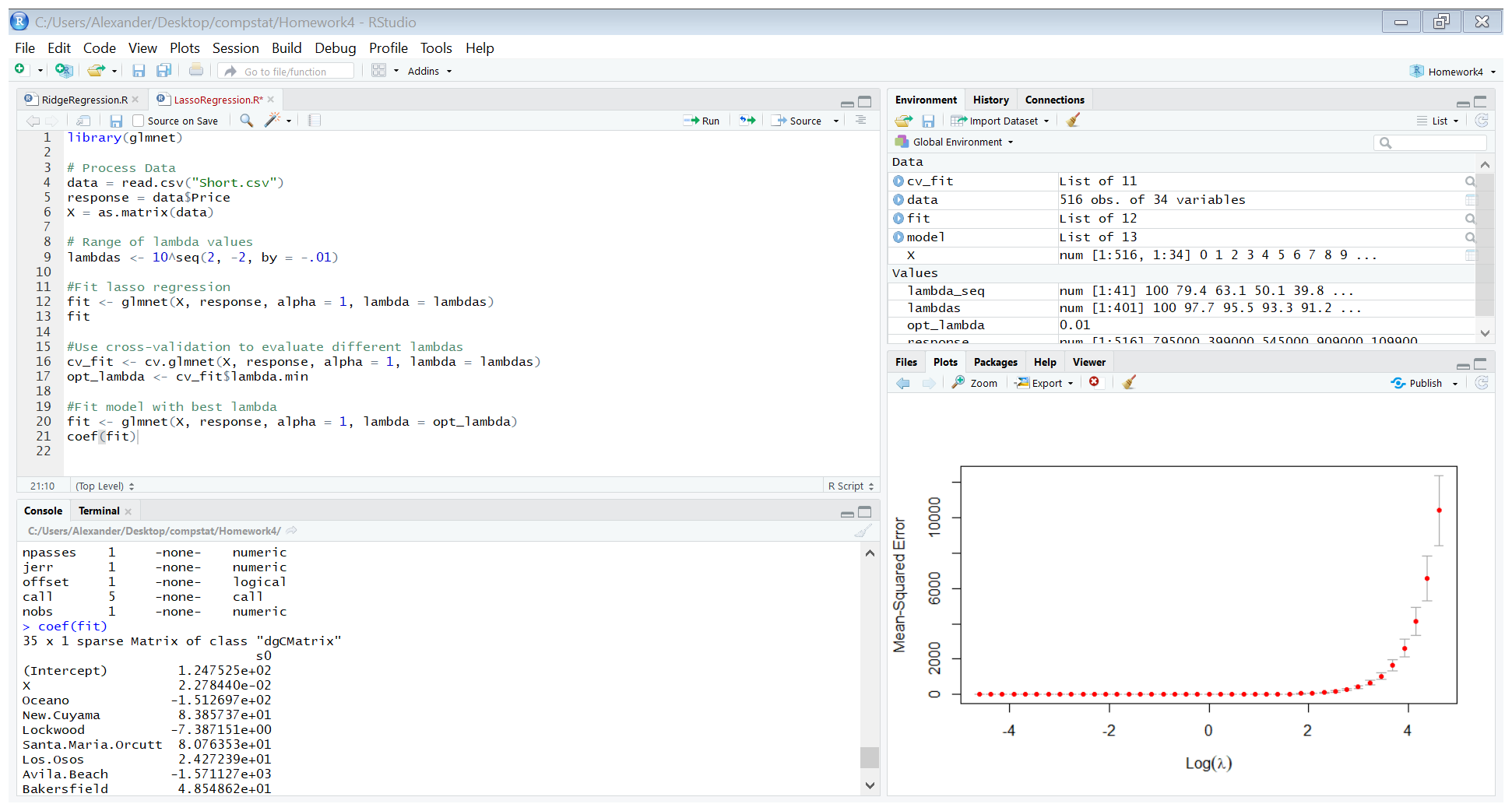
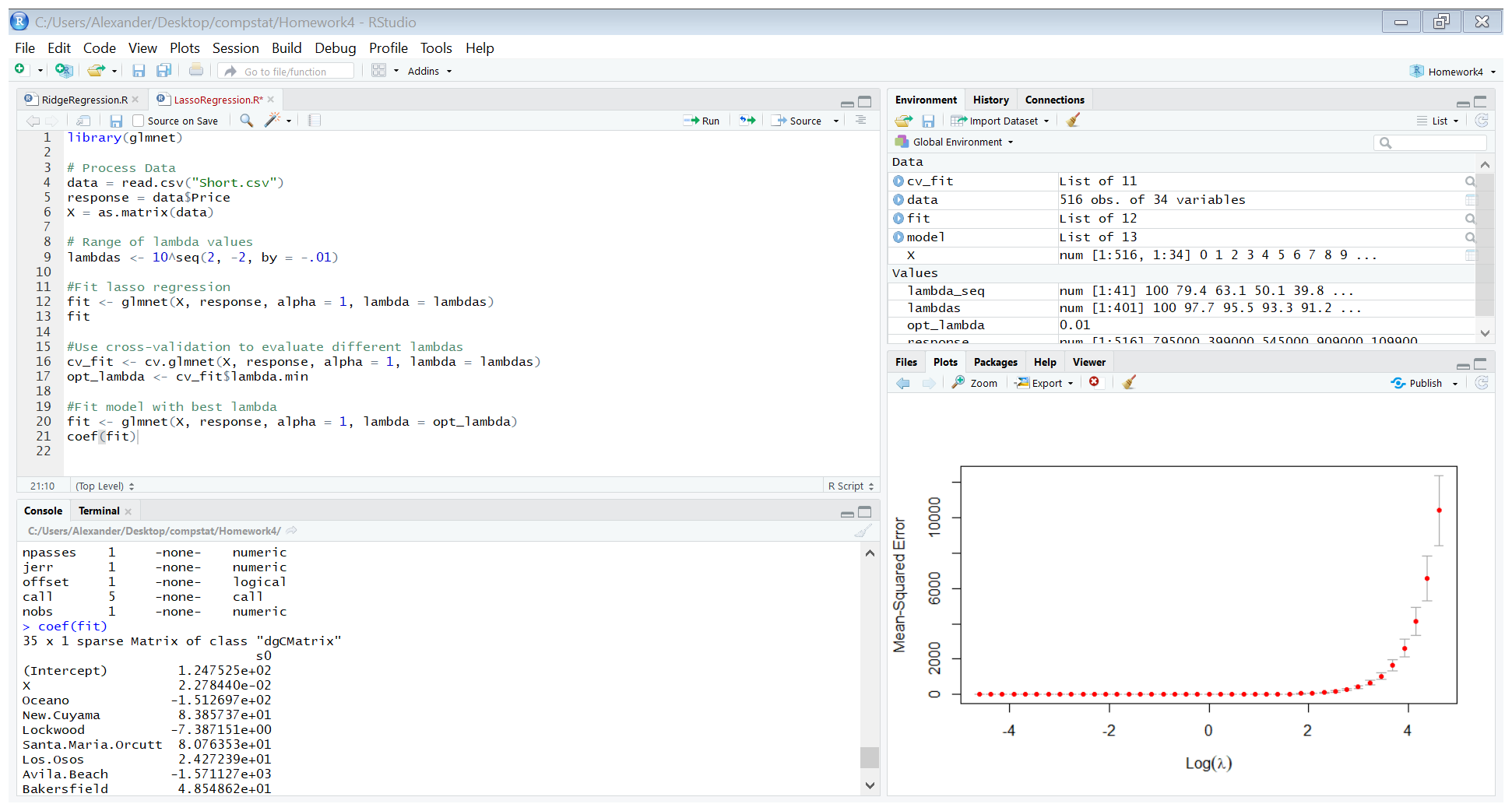
Alexander Bukharin

Variable selection for Housing Data Set

1. We conduct ridge regression in R for each dataset, using the following code



Then we conduct cross validation for each lambda, resulting in the following plot (This plot is for short sales, but we get similar plots for each dataset:



**Short Sale:**

For short sale, we get an optimal lambda of 0.01 (found through cross validation) and the following coefficients on each variable:

(Intercept) 1.116788e+03

X -1.691746e-01

San.Miguel -1.443411e+02

Atascadero -1.550446e+02

Nipomo -2.052037e+01

Santa.Maria.Orcutt -1.471024e+02

Oceano -7.439408e+01

Avila.Beach 1.942708e+03

Guadalupe -5.923402e+02

Templeton -9.985017e+01

Pismo.Beach -1.977123e+03

Los.Osos 7.460002e+01

Bradley -1.565921e+02

Buellton 1.278216e+02

Paso.Robles 7.293118e+01

King.City 4.133076e+02

San.Luis.Obispo 4.525455e+02

Greenfield 2.146475e+02

Arroyo.Grande -3.784938e+01

Santa.Ynez -3.042558e+02

Solvang 8.611794e+01

Lompoc -1.236085e+02

Morro.Bay -1.950852e+01

Santa.Margarita -7.650095e+02

Grover.Beach 5.222503e+01

San.Simeon 2.126226e+02

Soledad 6.112494e+01

Creston 2.001780e+01

Cambria -4.277762e+01

MLS 9.032487e-03

Price 1.007314e+00

Bedrooms 9.082209e+01

Bathrooms 1.943300e+01

Size -1.582079e+00

Price.SQ.Ft -1.288510e+01

**Foreclosure:**

We get an optimal lambda of 0.01 and the following coefficients:

(Intercept) 1.116788e+03

X -1.691746e-01

San.Miguel -1.443411e+02

Atascadero -1.550446e+02

Nipomo -2.052037e+01

Santa.Maria.Orcutt -1.471024e+02

Oceano -7.439408e+01

Avila.Beach 1.942708e+03

Guadalupe -5.923402e+02

Templeton -9.985017e+01

Pismo.Beach -1.977123e+03

Los.Osos 7.460002e+01

Bradley -1.565921e+02

Buellton 1.278216e+02

Paso.Robles 7.293118e+01

King.City 4.133076e+02

San.Luis.Obispo 4.525455e+02

Greenfield 2.146475e+02

Arroyo.Grande -3.784938e+01

Santa.Ynez -3.042558e+02

Solvang 8.611794e+01

Lompoc -1.236085e+02

Morro.Bay -1.950852e+01

Santa.Margarita -7.650095e+02

Grover.Beach 5.222503e+01

San.Simeon 2.126226e+02

Soledad 6.112494e+01

Creston 2.001780e+01

Cambria -4.277762e+01

MLS 9.032487e-03

Price 1.007314e+00

Bedrooms 9.082209e+01

Bathrooms 1.943300e+01

Size -1.582079e+00

Price.SQ.Ft -1.288510e+01

**Regular:**

We get an optimal lambda of 0.01 and regression coefficients of:

(Intercept) -4.968071e+06

X -5.064343e+01

Paso.Robles -1.620031e+02

Cambria 3.216432e+02

Santa.Maria.Orcutt -9.857711e+02

Santa.Maria.Orcutt.1 -5.643062e+02

Los.Osos -1.389067e+02

Nipomo -8.720637e+02

Templeton 2.394432e+02

San.Luis.Obispo -4.516032e+02

Pismo.Beach 5.764155e+02

Lompoc -9.314137e+02

Out.Of.Area -1.200558e+03

Morro.Bay 3.473556e+02

Arroyo.Grande -2.275898e+02

Cayucos 2.083651e+02

Solvang -1.849187e+01

San.Miguel 3.847827e+01

Oceano 4.500061e+03

Bradley 9.881947e+02

Atascadero -1.821692e+02

Grover.Beach 6.670645e+02

Creston 2.801709e+02

Out.Of.Area.1 1.217929e+02

MLS 3.240683e+01

Price 1.002738e+00

Bedrooms 3.312374e+02

Bathrooms 7.507214e+02

Size -1.949226e+00

Price.SQ.Ft -8.022762e+00

1. For short sale, using lambda = 75, we get that the leading factors are price per square foot and size.

For Regular, we get that that the leading factors again are price per square foot and size.

For Foreclosure, we get that Pismo.Beach (the location) and price per square foot are the two leading factors.

For all three attributes, we find that the optimal lambda is 0.01 for all three datasets, selecting from a set of lambdas that is in 10^-2 and 10^2 with step size 0.01.