

Midterm project, group 6

Ruoyuan Qian

1 Introduction

1.1 Background

In the dataset, the response variable is the SalePrice (\$) of residential homes in Ames, Iowa. And there are 1460 observations with 79 explanatory variables describing a variety of aspects of these housings. Among explanatory variables, there are 37 numeric variables and 43 categorical variables.

1.2 Objective

In this report, the focus is among seven methods (multiple linear regression, k-nearest neighbors algorithm (k-NN), ridge regression, lasso regression, elastic regression, principal component regression (PCR), partial least squares regression (PLS)), which one is the best to predict sale price of house for the particular dataset and which predictors are most influential for the response SalePrice.

1.3 Exploratory data analysis (EDA)

1.3.1 Data cleaning

Missing value is checked for each predictor and predictors with the number of missing value greater than 500 are excluded from the dataset. At the meantime, predictors with many zero or near-zero observations are removed as well. Finally, NA's are dropped from the remaining data. When all works are done, there are 55 predictors in total including 28 numeric predictors and 27 categorical predictors.

1.3.2 Visualization

The distribution of response SalePrice (\$) is checked (Fig. 1), as we can see, it is continuous variable with a right skewed shape. Since all methods in report do not need normal distribution assumption, so the original value of response can be used in model fitting.

Scatter plots are checked for numeric variables (Fig. 2), bar plots are shown for categorical variables (Fig. 3). Correlations between predictors are visualized by heat plot (Fig. 4).

2 Methods

The data for all methods is the same to keep the ability of comparison. Data is scaled and standardized in model fitting. There are generally two types of methods: linear methods and non-linear methods.

2.1 Linear Methods

2.1.1 Multiple linear regression (MLR)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

β' s are estimated by least square estimation:

$$RSS = \sum_{i=1}^n (y_i - \hat{y}_i)$$

, where $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}_1 + \dots + \hat{\beta}_p \hat{X}_p$. Although MLR requires a Gaussian error for any inference including prediction interval, the report is focus on the ability of prediction instead of inference, so we don't have to do transformation for response variable.

2.1.2 Ridge regression

The ridge coefficient estimation is the minimum of the loss function:

$$\min(RSS + \lambda \sum_{j=1}^p \beta_j^2)$$

All coefficients will shrink when λ increases, but none of them will shrink to zero. So ridge regression will remain all predictors in the final model. In R, $\alpha = 0$ is fixed, and a range of λ is implemented to find the best tuning parameter with the criteria of smallest MSE through cross validation (CV).

2.1.3 LASSO regression

The LASSO coefficient estimation is the minimum of the loss function:

$$\min(RSS + \lambda \sum_{j=1}^p |\beta_j|)$$

All coefficients will shrink to zero when λ is large enough. LASSO regression will remain a subset of predictors in the final model. In R, $\alpha = 1$ is fixed, selection of best tuning parameter is similar to ridge.

2.1.4 Elastic regression

The elastic coefficient estimation is the minimum of the loss function:

$$\min(RSS + \lambda_1 \sum_{j=1}^n \beta_j^2 + \lambda_2 \sum_{j=1}^n |\beta_j|)$$

All coefficients will not shrink to exact zero when λ increases. In R, α can be changed in $[0,1]$, so combinations of a range of λ and α is implemented to find the best combination of tuning parameters with the criteria of smallest MSE through cross validation (CV).

2.1.5 Principal component regression (PCR)

It includes two steps, dimension reduction and regression.

$$Z_m = \sum_{j=1}^p \phi_{mj} X_j$$

$$y_i = \theta_0 + \sum_{m=1}^M \theta_m z_{im}$$

2.2 Non-linear methods

2.2.1 K-nearest neighbors algorithm (k-NN)

The k nearest points are used to fit the line.

$$\hat{f}(x_0) = Ave(y|x \in N(x_0)) = \sum_{i=1}^n w(x_0, x_i) y_i$$

Where $w(x, x_i) = I(x_i \in N_k(x))/K$

2.2.2 Generalized additive model (GAM)

It allows flexible non-linearities in several variables based on their own scatter plot or degree of freedom (DF), if the points are not linear shaped or the DF is greater than 1, then a non-linear term should be considered.

$$g[E(y|X)] = \beta_0 + f_1(X_1) + \dots + f_p(X_p)$$

2.2.3 Multivariate Adaptive Regression Spline (MARS)

It is a piecewise linear model while the cut points are selected by algorithm, and then the hinge functions can be written as $(h(x - c), h(c - x))$.

3 Results

3.1 Model comparison

For LASSO, plot of MSE across a sequence of λ is made (Fig. 5), and the best λ is $e^{7.035}$. For ridge model, plot of MSE across a sequence of λ (Fig. 6) shows that the best tuning parameter is $\lambda = e^{10.10}$. As for elastic model, 750 combinations of α and λ are checked (Fig. 7), the best pair is $\alpha = 0, \lambda = e^{10.10}$, which is the same as ridge, so in the model comparison, only ridge regression will be presented. For PCR, 54 principle components (PC) are tested, and the best number of principle component (PC) is 26 through MSE (Fig. 8).

For k-NN model, after testing a sequence of k from 5 to 43, the best tuning parameter is equal to 11. For GAM model, “train” function is implemented and 20 out of 54 predictors are tested for non-linear relation to response (Fig. 9). For MARS model, 8 cut points are used to fit the model (Fig. 10).

All models are compared through MSE (Tab. 1). The best model is the one with the smallest MSE, which is MARS. K-nn method obtain the largest MSE among the all.

3.2 Model interpretation

4 Limitation

Figures and Tables

Table 1 MSE of all methods through cross validation

column	mean	sd	median	min	max	range
Lm	37505.04	13940.946	33542.84	22794.36	72849.29	50054.93
LASSO	36543.47	12986.560	33204.86	22150.00	69494.01	47344.00
Ridge	36278.35	12238.432	32908.06	21886.61	66422.74	44536.14
PCR	36483.11	12114.243	32675.95	23044.04	67262.97	44218.94
Knn	38064.71	8516.354	37404.23	25491.27	58379.22	32887.94
GAM	35926.26	23224.696	27382.22	19331.35	136719.62	117388.26
MARS	33038.51	9329.652	31375.96	21072.25	58224.52	37152.27

Table 2 Hinge functions and their Coefficients

	Coefficient
(Intercept)	3.141179e+05
h(OverallQual-7)	4.077041e+04
h(7-OverallQual)	-9.109677e+03
h(2945-GrLivArea)	-5.752370e+01
h(X2ndFlrSF-1360)	3.718084e+02
h(YearBuilt-2007)	1.563304e+04
h(2007-YearBuilt)	-5.721088e+02
h(BsmtFinSF1-763)	4.820646e+01
h(763-BsmtFinSF1)	-1.379454e+01
h(LotArea-20431)	4.618026e-01
h(20431-LotArea)	-1.953780e+00
h(7-OverallCond)	-1.093203e+04
h(X1stFlrSF-2121)	-3.447610e+02
h(TotalBsmtSF-1626)	7.145797e+01
h(1626-TotalBsmtSF)	-1.678924e+01
h(SaleCondition-4)	1.813065e+04
h(4-SaleCondition)	3.657226e+03

Figure 1 Distridution of response (SalePrice)

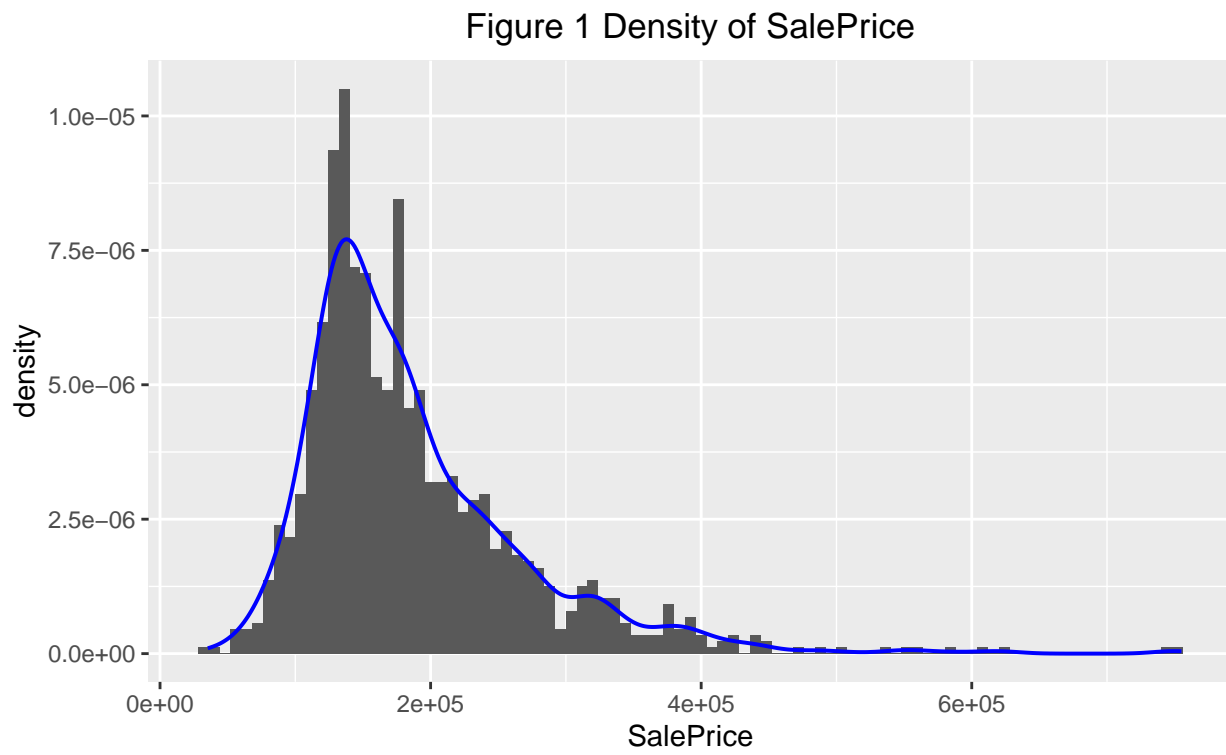


Figure 2 Scatter plots of continuous predictors against SalePrice

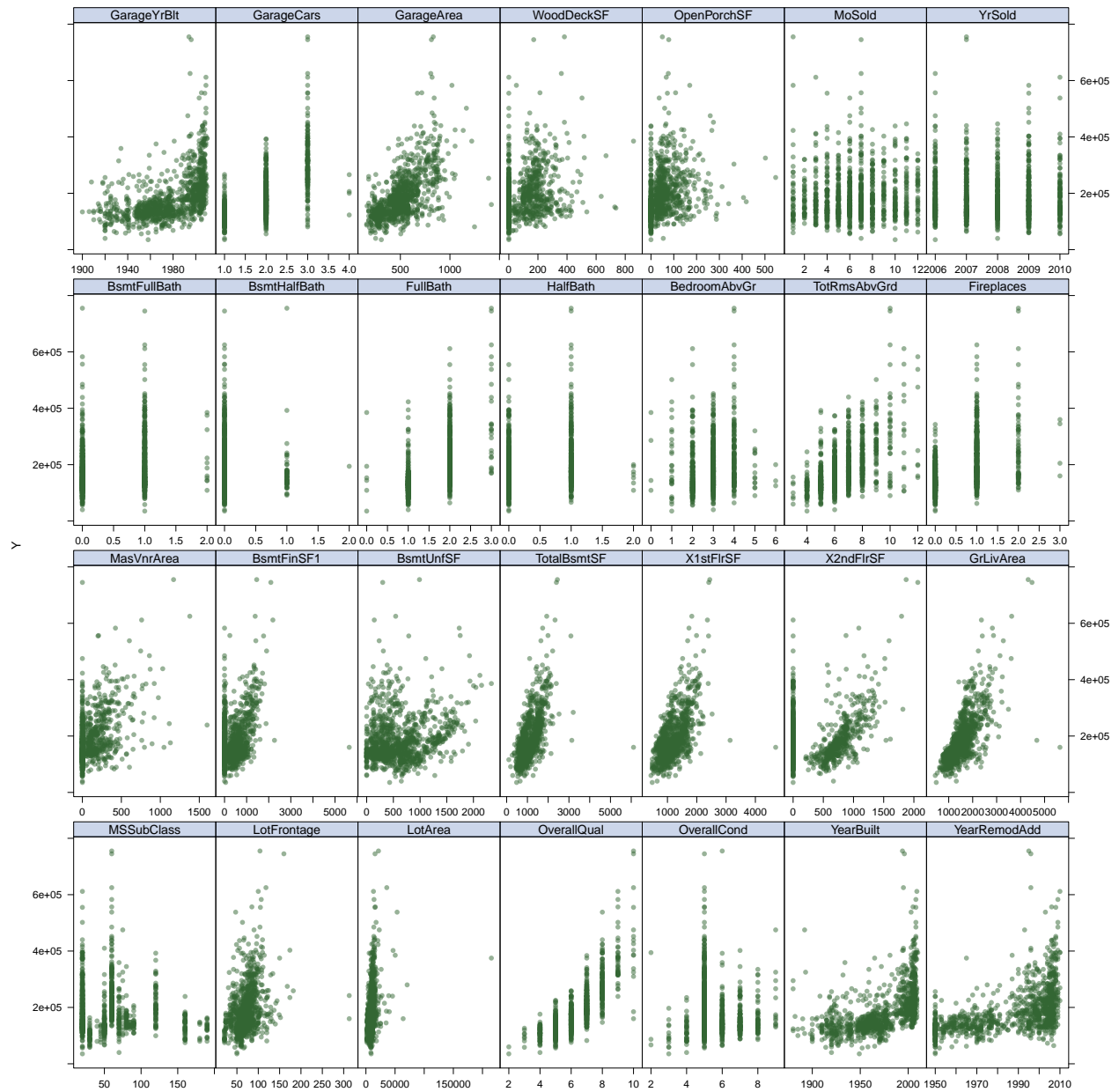


Figure 3 Bar plots of categorical predictors



Figure 4 Heat map for all predictors

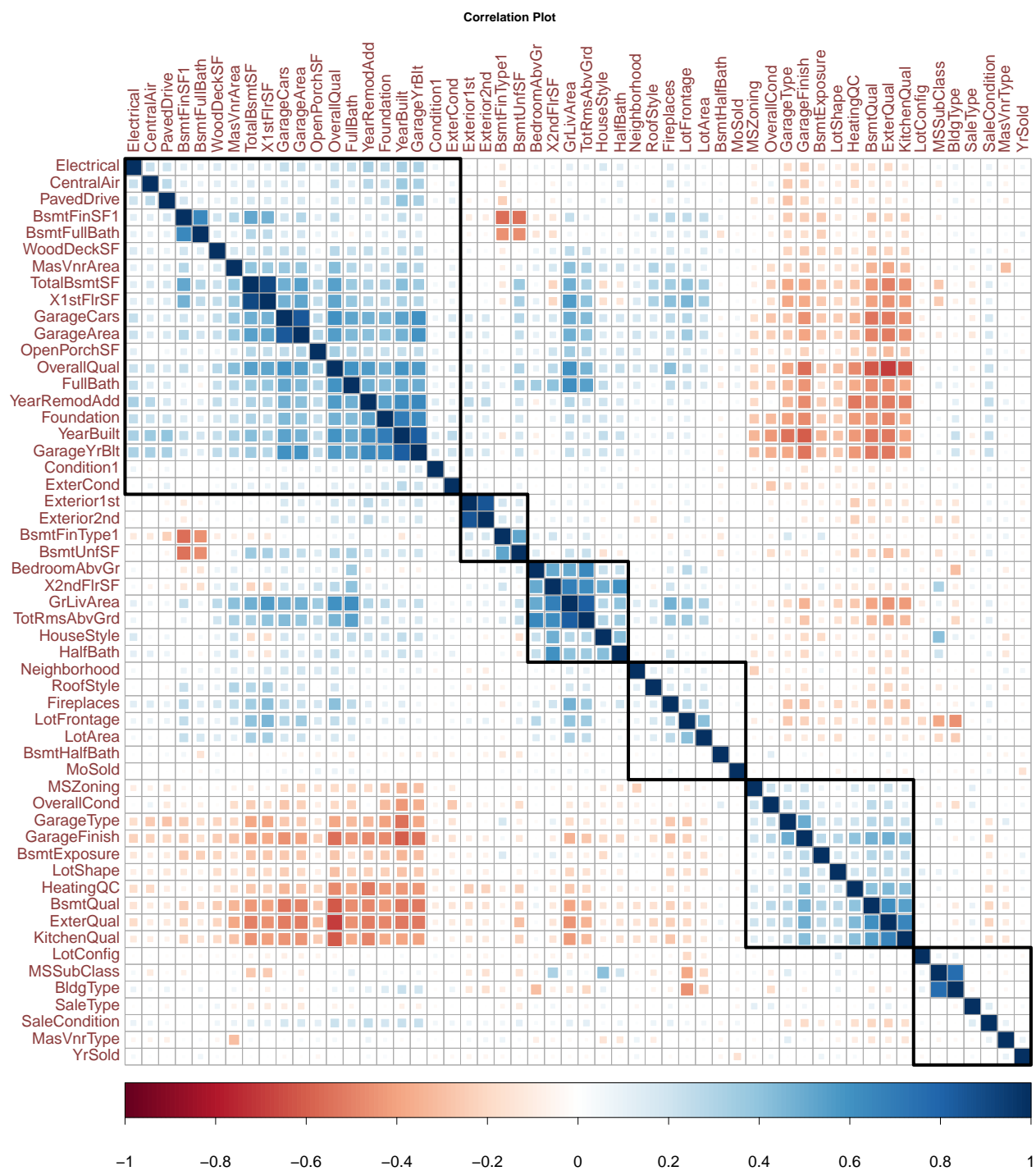


Figure 5 LASSO

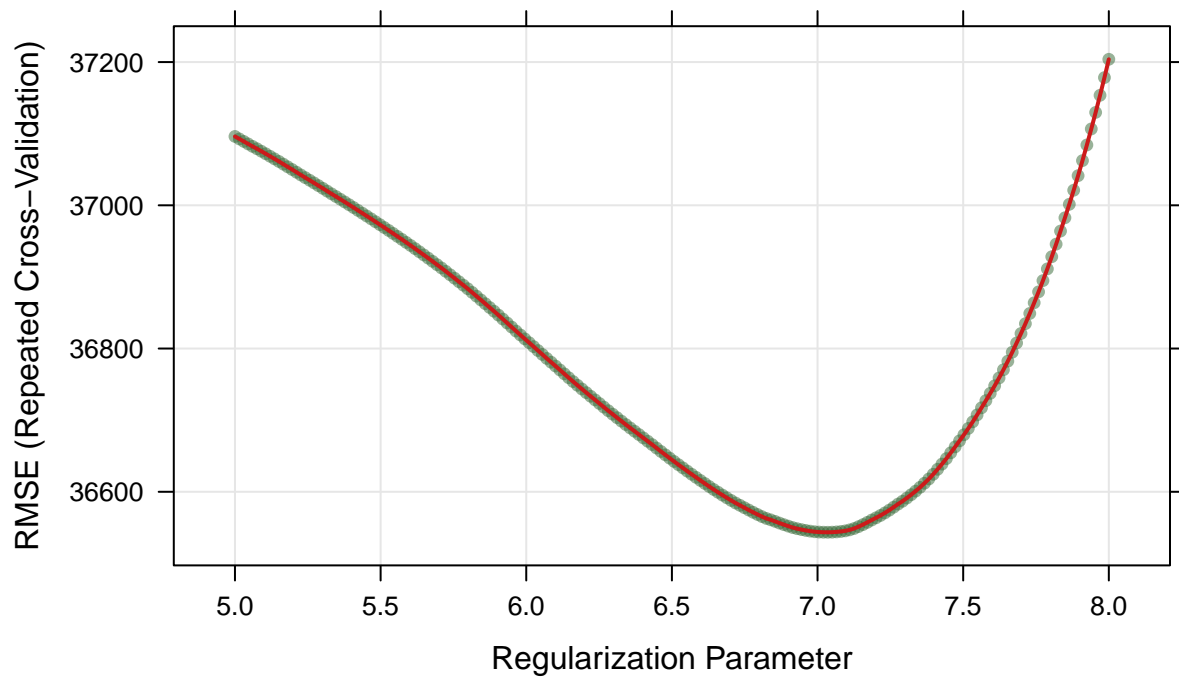


Figure 6 Ridge

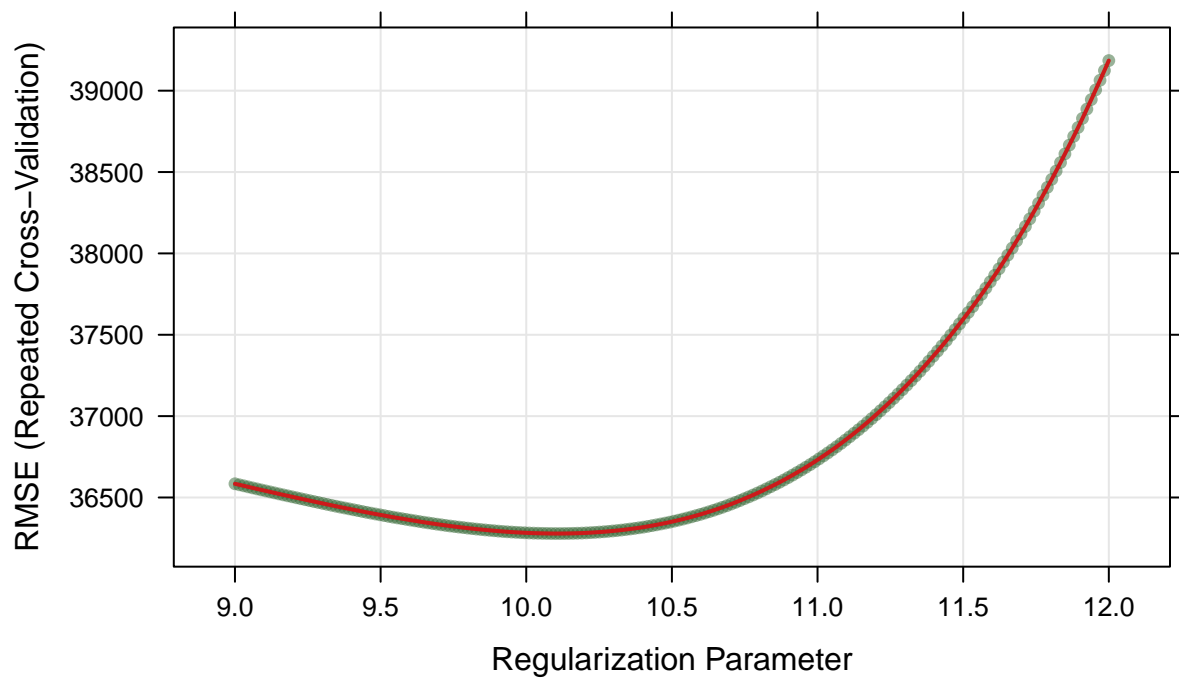


Figure 7 Elastic

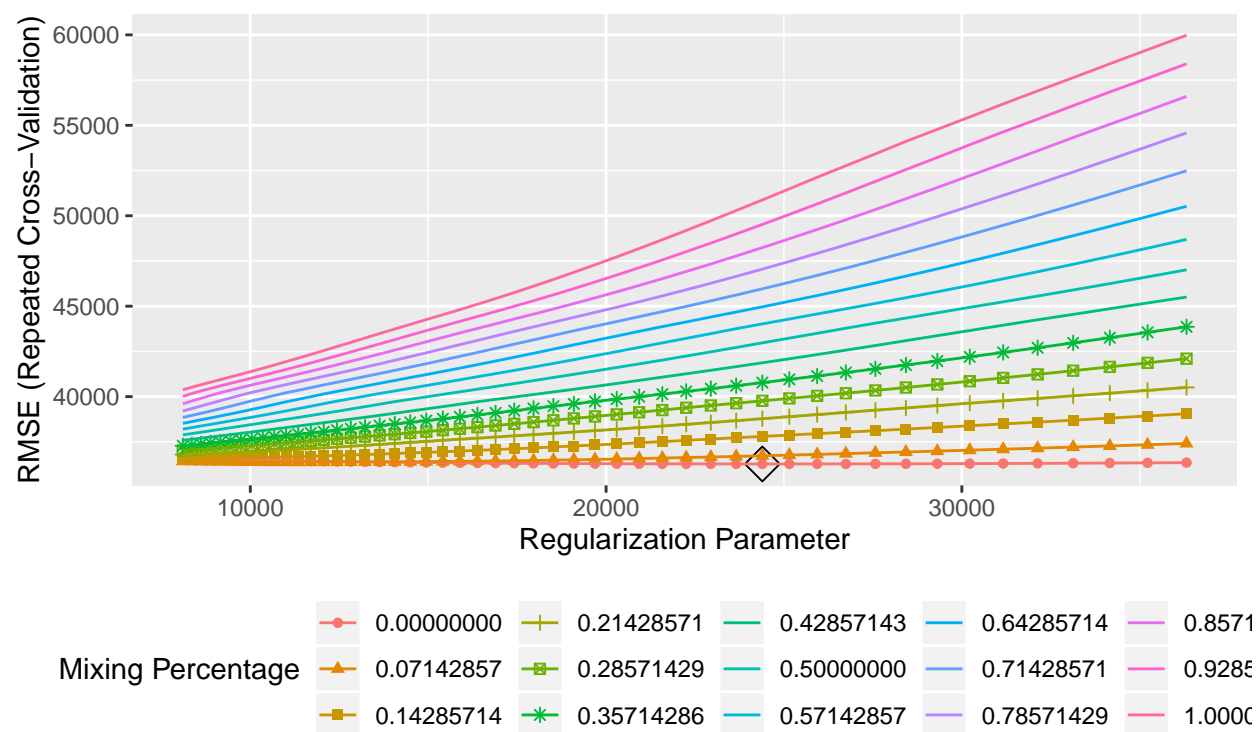


Figure 8 PCR

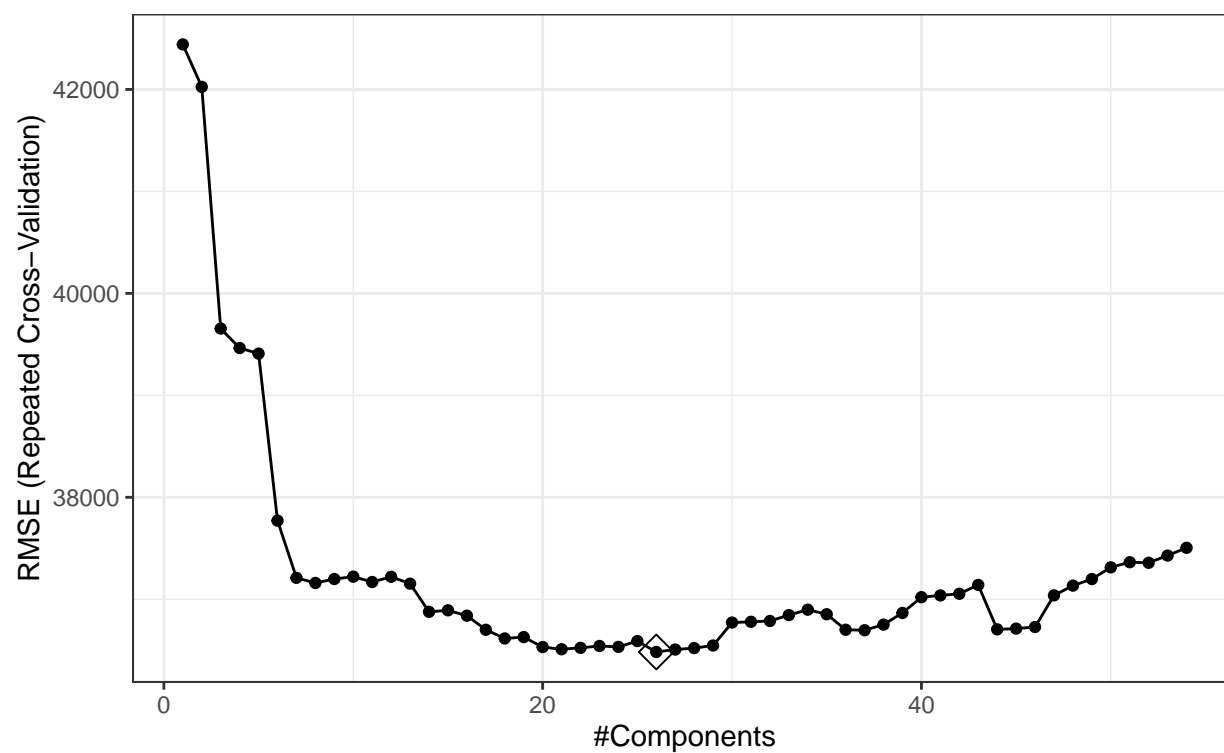


Figure 9 GAM

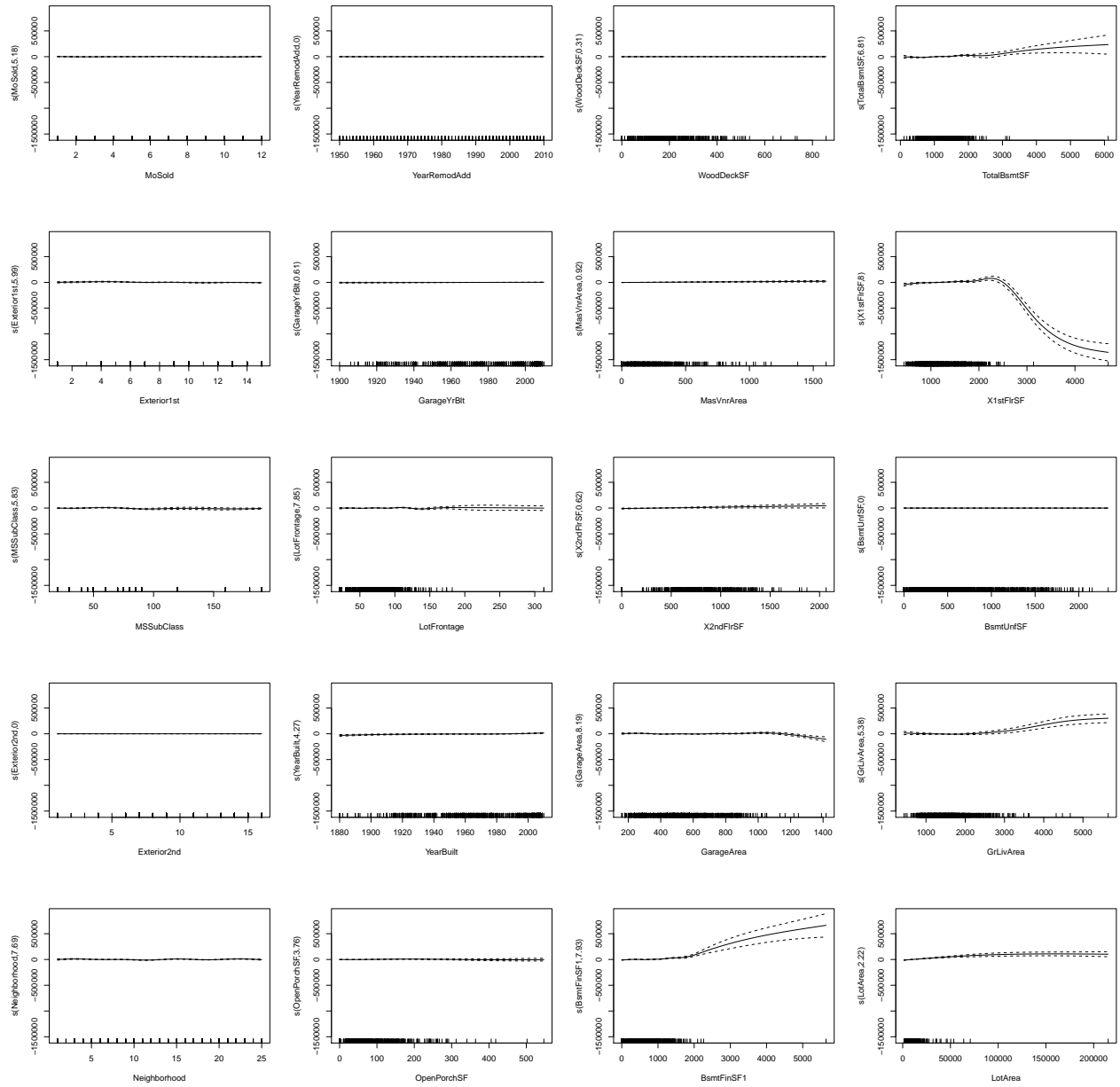


Figure 10 MARS

