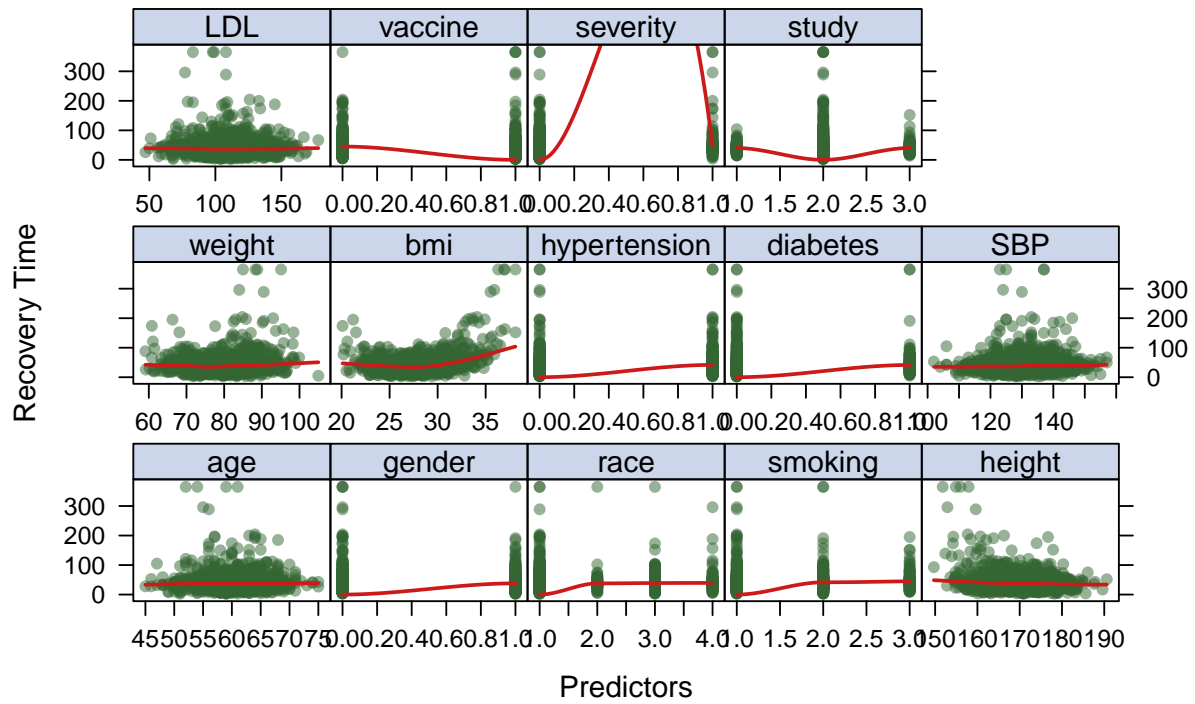


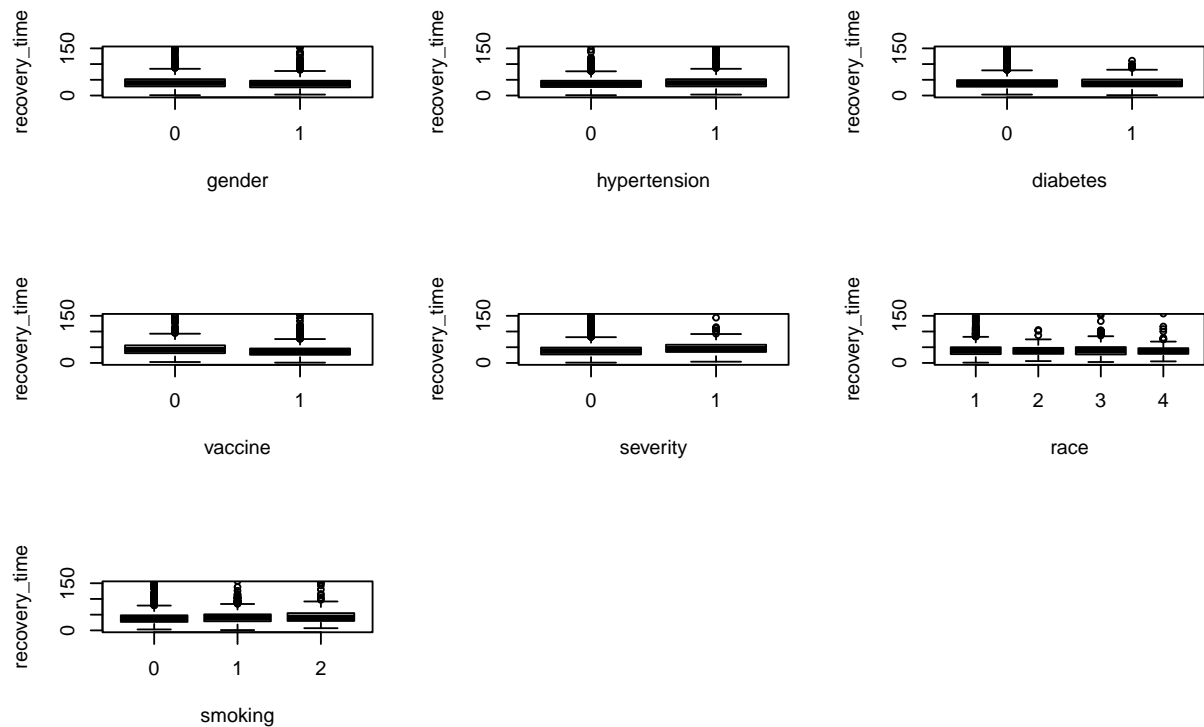
ds2\_midterm

Ruihan Zhang

2023-03-28

**Figure 1. the relationship between predictors and recovery time**





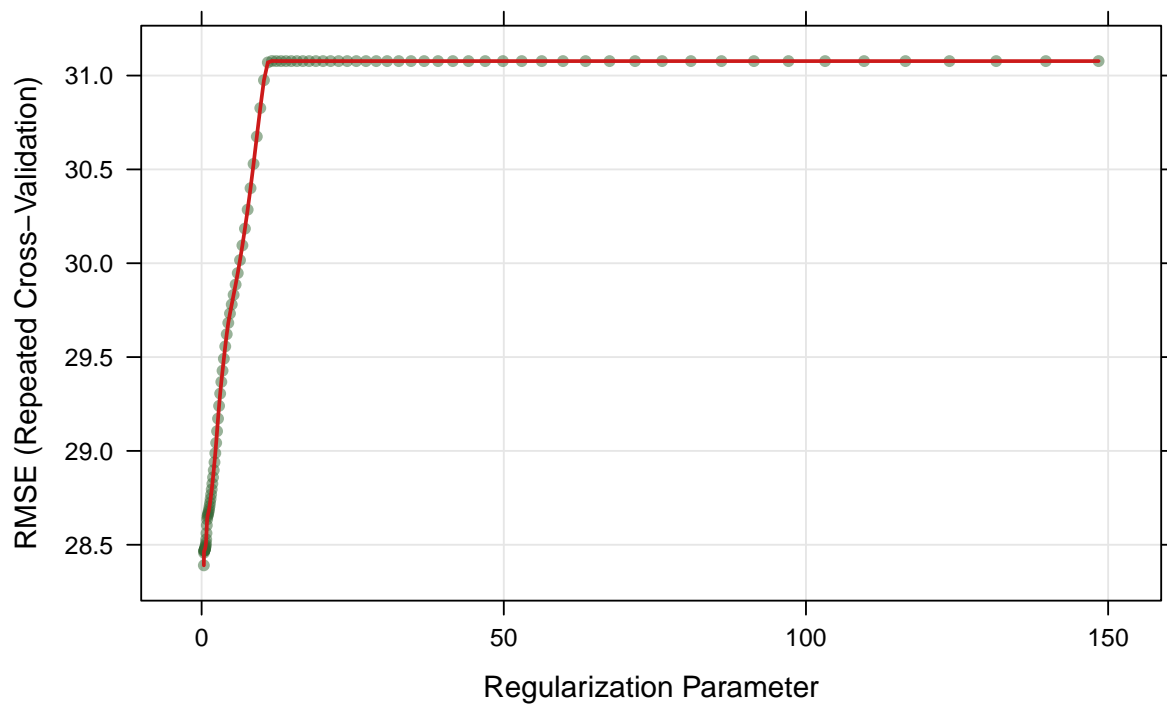
```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -67.936 -13.844  -1.397   10.958  214.844
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.525e+03  1.866e+02 -18.894  < 2e-16 ***
## age          -8.418e-02  1.760e-01  -0.478  0.63260
## gender       -6.721e+00  1.386e+00  -4.849  1.38e-06 ***
## race2        -4.587e+00  3.202e+00  -1.433  0.15214
## race3        -2.286e+00  1.751e+00  -1.306  0.19189
## race4         3.600e-03  2.367e+00   0.002  0.99879
## smoking1      4.008e+00  1.543e+00   2.597  0.00949 **
## smoking2      7.161e+00  2.503e+00   2.861  0.00429 **
## height       2.060e+01  1.100e+00  18.723  < 2e-16 ***
## weight      -2.239e+01  1.162e+00 -19.275  < 2e-16 ***
## bmi          6.677e+01  3.303e+00  20.212  < 2e-16 ***
## hypertension  1.838e+00  2.307e+00   0.797  0.42565
## diabetes     1.562e+00  1.947e+00   0.803  0.42232
## SBP          1.048e-01  1.491e-01   0.703  0.48218
## LDL         -1.018e-02  3.601e-02  -0.283  0.77741
## vaccine      -8.737e+00  1.415e+00  -6.175  8.67e-10 ***
## severity      9.505e+00  2.314e+00   4.108  4.22e-05 ***
## studyB        4.580e+00  1.739e+00   2.634  0.00854 **
```

```

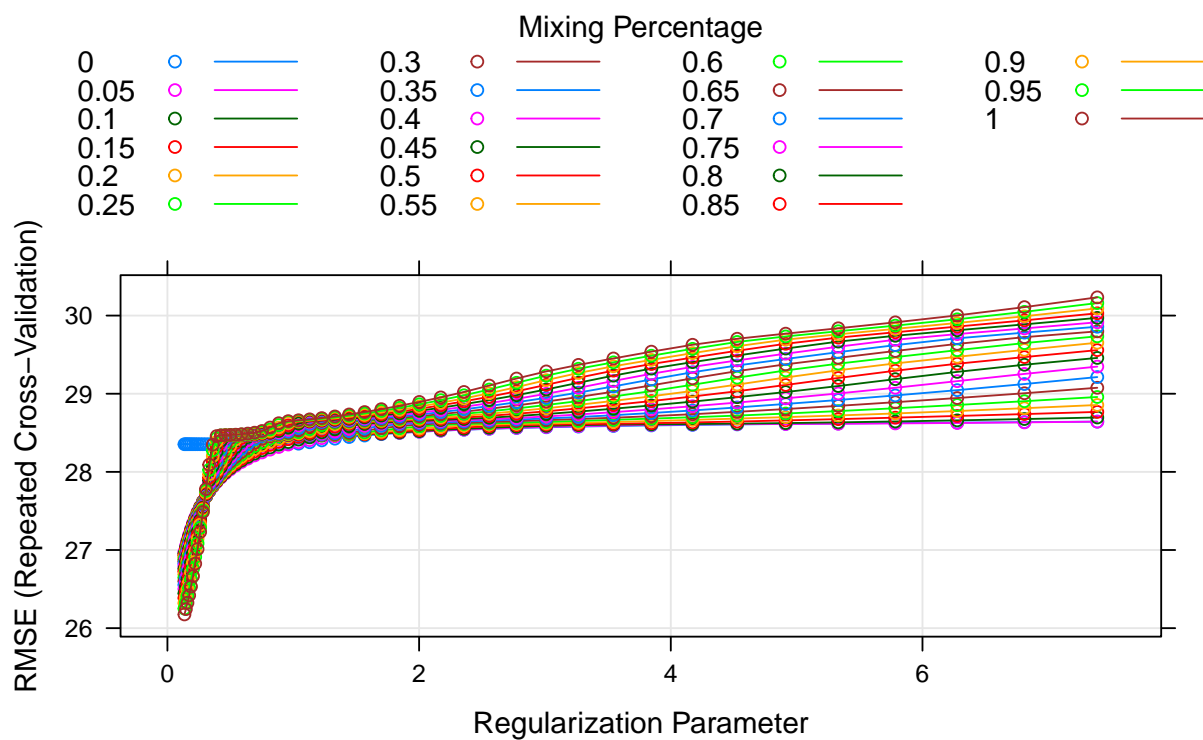
## studyC          3.436e-01  2.183e+00   0.157  0.87492
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.77 on 1383 degrees of freedom
## Multiple R-squared:  0.3552, Adjusted R-squared:  0.3468
## F-statistic: 42.32 on 18 and 1383 DF,  p-value: < 2.2e-16
## [1] 20.59645

## 19 x 1 sparse Matrix of class "dgCMatrix"
##              s1
## (Intercept) -93.205765308
## age          .
## gender       -5.919136420
## race2        -0.010213828
## race3        -1.777671736
## race4        .
## smoking1     1.948592247
## smoking2     6.302942295
## height       0.247367973
## weight      -0.890672253
## bmi          5.573001483
## hypertension 0.275670589
## diabetes     1.355426214
## SBP          0.105877149
## LDL         -0.001277377
## vaccine     -8.352237305
## severity    10.552981481
## studyB       5.646347276
## studyC      .
## [1] 24.77985
## [1] 0

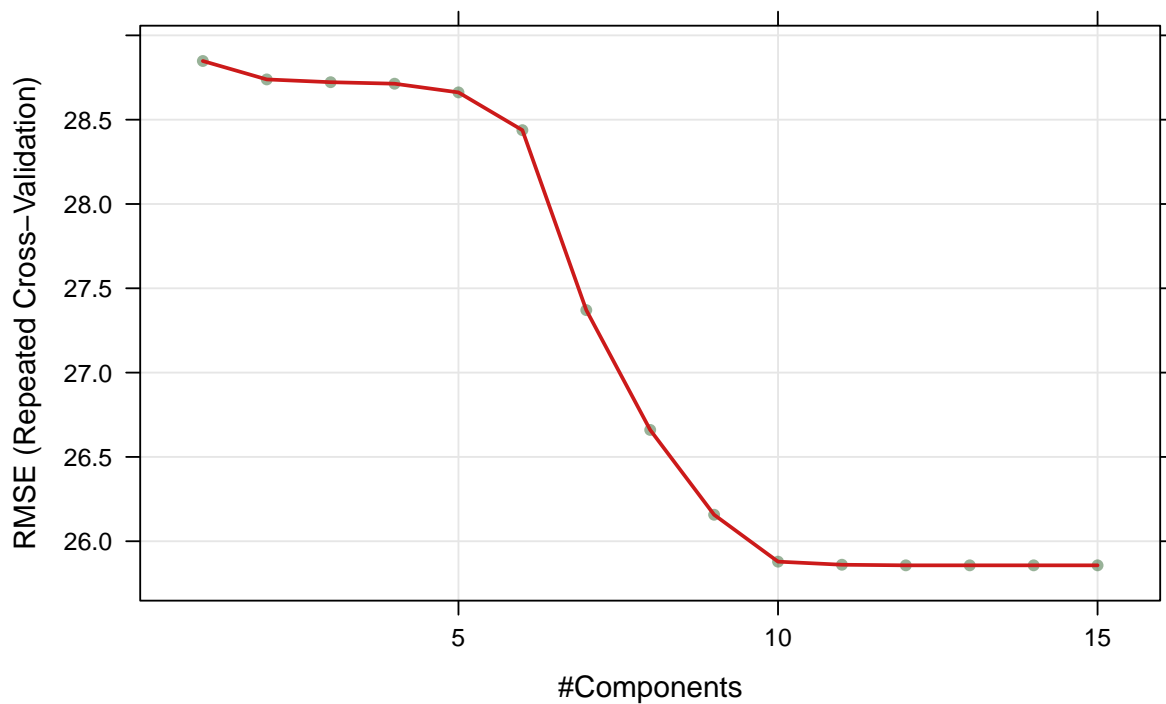
```



```
##      alpha    lambda
## 1001      1 0.1353353
## [1] 20.05926
```



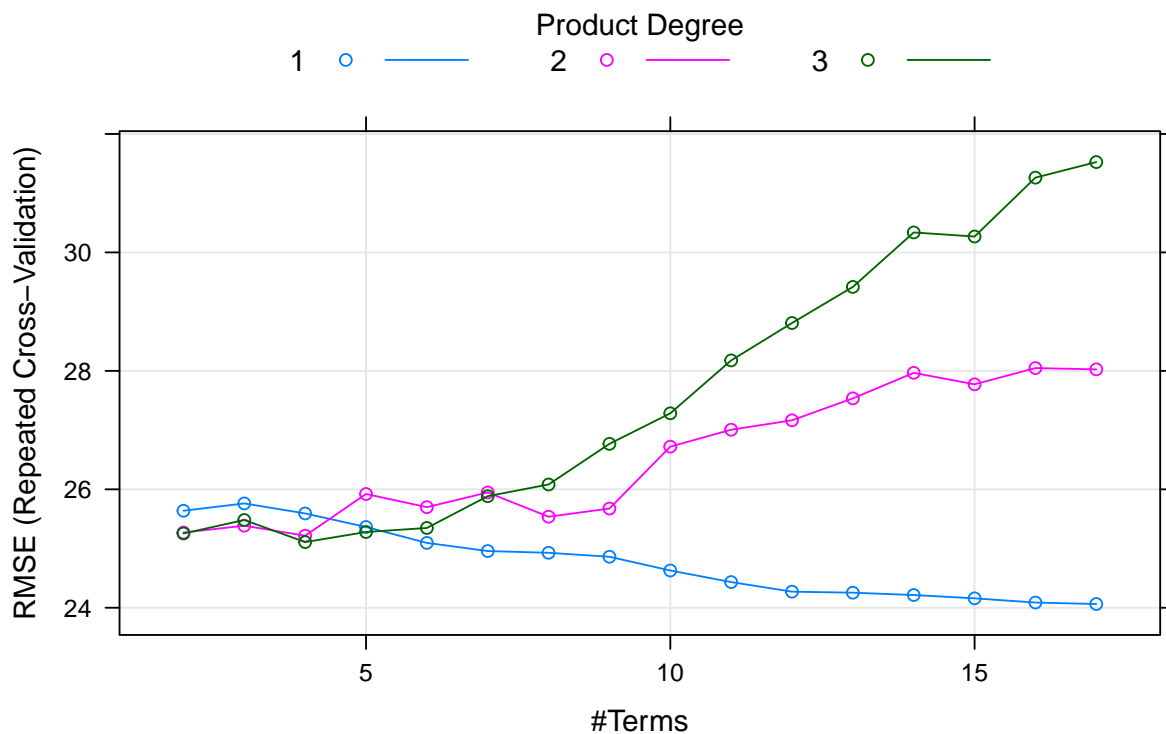
```
## Data:      X dimension: 1402 18
## Y dimension: 1402 1
## Fit method: oscorespls
## Number of components considered: 12
## TRAINING: % variance explained
##           1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps
## X           9.148  17.36  26.52  32.78  38.19  44.15  47.09
## .outcome    16.443  17.59  17.79  18.03  18.58  20.31  27.87
##           8 comps 9 comps 10 comps 11 comps 12 comps
## X           50.86  54.19  57.48  62.98  68.02
## .outcome    32.53  34.50  35.47  35.52  35.52
## [1] 20.59837
```



	nprune	degree
	16	1

```
##      (Intercept)      h(bmi-31.7)      h(31.7-bmi)      h(bmi-35.1)      vaccine
##      -163.699095      -31.641543      34.176354      57.982767      -8.484611
## h(height-159.3) h(159.3-height) h(90.6-weight)      h(bmi-29.4)      gender
##      -14.976026      15.911043      -4.815058      12.373579      -5.847346
##      severity      h(bmi-22.8)      studyB h(height-156.1)      smoking2
##      9.038245      26.313690      4.432291      12.770125      7.568910
##      smoking1 h(weight-67.3)
##      3.473229      -2.285645
## [1] 19.31587
## Call: earth(x=matrix[1402,18], y=c(23,25,37,10,4...), keepxy=TRUE, degree=1,
##      nprune=17)
##
##      coefficients
## (Intercept)      -163.699095
## gender          -5.847346
## smoking1        3.473229
## smoking2        7.568910
## vaccine         -8.484611
## severity        9.038245
## studyB          4.432291
## h(height-156.1) 12.770125
## h(159.3-height) 15.911043
## h(height-159.3) -14.976026
## h(weight-67.3)  -2.285645
```

```
## h(90.6-weight)      -4.815058
## h(bmi-22.8)         26.313690
## h(bmi-29.4)         12.373579
## h(31.7-bmi)         34.176354
## h(bmi-31.7)        -31.641543
## h(bmi-35.1)         57.982767
##
## Selected 17 of 22 terms, and 9 of 18 predictors (nprune=17)
## Termination condition: Reached nk 37
## Importance: bmi, vaccine, height, weight, gender, severity, studyB, ...
## Number of terms at each degree of interaction: 1 16 (additive model)
## GCV 521.9204    RSS 697691    GRSq 0.4868174    RSq 0.5099927
```

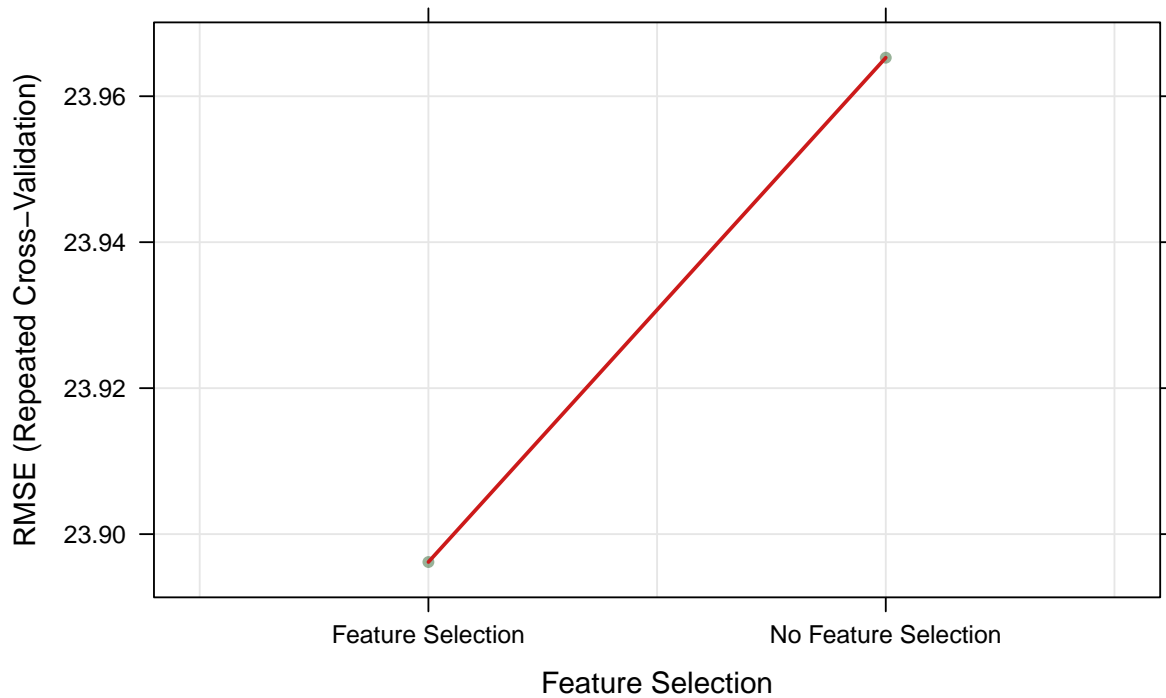


```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## .outcome ~ gender + race2 + race3 + race4 + smoking1 + smoking2 +
##   hypertension + diabetes + vaccine + severity + studyB + studyC +
##   s(age) + s(SBP) + s(LDL) + s(bmi) + s(height) + s(weight)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.7757    1.8958   24.146 < 2e-16 ***
## gender       -5.8950    1.2196   -4.833 1.49e-06 ***
## race2        -4.7692    2.8264   -1.687 0.09176 .
```

```

## race3          -2.3028      1.5440   -1.491   0.13606
## race4          -2.2816      2.0921   -1.091   0.27565
## smoking1       3.5670      1.3643    2.614   0.00903 **
## smoking2       7.5238      2.2095    3.405   0.00068 ***
## hypertension   2.0701      1.2218    1.694   0.09044 .
## diabetes        2.1927      1.7190    1.276   0.20233
## vaccine        -8.1553      1.2483   -6.533  9.06e-11 ***
## severity        8.8627      2.0398    4.345  1.50e-05 ***
## studyB          4.3530      1.5327    2.840   0.00458 **
## studyC          0.1153      1.9233    0.060   0.95222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##              edf Ref.df      F  p-value
## s(age)       1.212e-09    9  0.000   0.971
## s(SBP)        2.143e-09    9  0.000   0.976
## s(LDL)        2.499e-09    9  0.000   0.873
## s(bmi)        8.753e+00    9 79.687 < 2e-16 ***
## s(height)     7.240e+00    9  6.053 < 2e-16 ***
## s(weight)     4.320e+00    9  5.018 5.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.497   Deviance explained = 50.9%
## GCV = 523.55   Scale est. = 511.11    n = 1402
## NULL
## [1] 19.36985

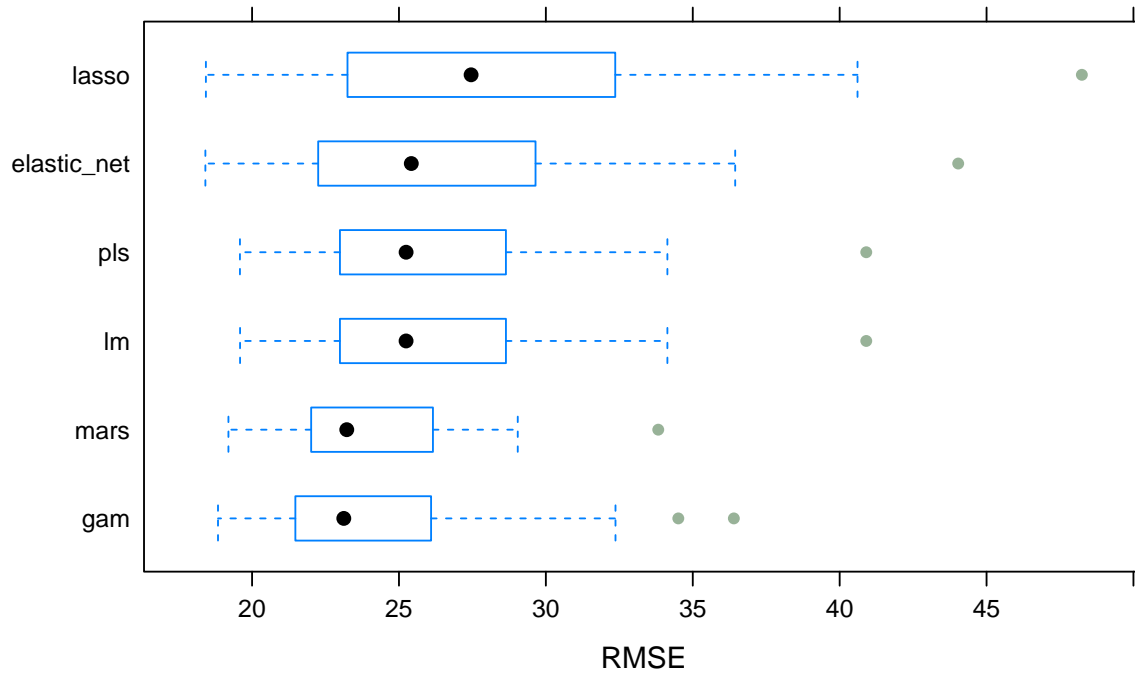
```





```
##
## Call:
## summary.resamples(object = res)
##
## Models: lm, lasso, elastic_net, pls, gam, mars
## Number of resamples: 50
##
## MAE
##           Min. 1st Qu.  Median    Mean 3rd Qu.    Max. NA's
## lm          14.92169 16.79714 17.53143 17.48965 18.20362 20.65141    0
## lasso        14.44659 16.68324 17.68983 17.84015 18.87016 21.82297    0
## elastic_net  14.46166 16.27802 17.01237 17.06807 17.87792 20.75223    0
## pls          14.92145 16.79669 17.53011 17.48984 18.20224 20.65250    0
## gam          13.62327 15.14377 15.84965 16.05984 16.72174 19.57934    0
## mars         14.10529 15.40746 16.13835 16.17264 16.80479 19.07897    0
##
## RMSE
##           Min. 1st Qu.  Median    Mean 3rd Qu.    Max. NA's
## lm          19.59117 23.05018 25.24154 25.85722 28.59408 40.90569    0
## lasso        18.42737 23.38819 27.45523 28.39035 32.31464 48.24480    0
## elastic_net  18.40971 22.26193 25.41912 26.17550 29.58809 44.03522    0
## pls          19.58634 23.05011 25.24227 25.85717 28.59370 40.90569    0
## gam          18.83947 21.49673 23.12193 23.89618 26.02227 36.39843    0
## mars         19.19470 22.04318 23.22292 24.06300 26.12428 33.82569    0
##
## Rsquared
##           Min. 1st Qu.  Median    Mean 3rd Qu.    Max. NA's
## lm          0.119520680 0.2365440 0.3266709 0.3285216 0.4085749 0.5970516    0
## lasso        0.003666023 0.1352594 0.1871645 0.1856959 0.2323649 0.3296231    0
## elastic_net  0.102361853 0.2560678 0.3266106 0.3125386 0.3777359 0.5246915    0
## pls          0.119510263 0.2365178 0.3266308 0.3285227 0.4085783 0.5970333    0
## gam          0.108667896 0.2926807 0.3566997 0.4136214 0.5908192 0.7489117    0
## mars         0.156098173 0.2989208 0.3764912 0.4229348 0.5628118 0.7415360    0
```

**Figure 2. Model Comparison**



```
## [1] 23.96528 23.89618
## [1] 19.36985
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## recovery_time ~ gender + race + +smoking + hypertension + diabetes +
##      vaccine + severity + study + s(age) + s(SBP) + s(LDL) + s(bmi) +
##      s(height) + s(weight)
##
## Parametric coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.8507    2.0477   22.391  < 2e-16 ***
## gender       -5.8254    1.2283   -4.743  2.33e-06 ***
## race2        -4.7591    2.8307   -1.681  0.092944 .
## race3        -2.3053    1.5474   -1.490  0.136504
## race4        -2.3204    2.0956   -1.107  0.268361
## smoking1      3.5753    1.3667    2.616  0.008992 **
## smoking2      7.5368    2.2135    3.405  0.000681 ***
## hypertension  1.7696    2.0366    0.869  0.385059
## diabetes      2.1977    1.7229    1.276  0.202329
## vaccine      -8.1433    1.2508   -6.510  1.05e-10 ***
## severity      8.9118    2.0454    4.357  1.42e-05 ***
## studyB        4.3804    1.5364    2.851  0.004421 **
## studyC        0.1369    1.9278    0.071  0.943408
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##           edf Ref.df      F  p-value
## s(age)      1.000  1.000  0.115   0.735
## s(SBP)      1.000  1.000  0.058   0.810
## s(LDL)      1.000  1.000  0.018   0.893
## s(bmi)      8.738  8.973 78.931 < 2e-16 ***
## s(height)   7.179  8.173  5.161 2.08e-06 ***
## s(weight)   5.084  6.224  5.775 5.00e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.496   Deviance explained = 50.9%
## GCV = 526.4   Scale est. = 512.51    n = 1402
```

