p8106_hw1_yg2625

Yue Gu

March 3, 2019

Import data

```
train_data = read.csv("./data/solubility_train.csv") %>%
  janitor::clean_names()
test_data = read.csv("./data/solubility_test.csv") %>%
  janitor::clean_names()
```

(a) Fit a linear model using least squares on the training data and calculate the mean square error using the test data.

Fit linear model on the training data

```
fit_lm_tr = lm(solubility ~ .  data = train_data)
summary(fit_lm_tr)
```

```
##
## Call:
## lm(formula = solubility ~ ., data = train_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75620 -0.28304  0.01165  0.30030  1.54887
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.431e+00  2.162e+00   1.124 0.261303
## fp001               3.594e-01  3.185e-01   1.128 0.259635
## fp002               1.456e-01  2.637e-01   0.552 0.580960
## fp003              -3.969e-02  1.314e-01  -0.302 0.762617
## fp004              -3.049e-01  1.371e-01  -2.223 0.026520 *
## fp005               2.837e+00  9.598e-01   2.956 0.003223 **
## fp006              -6.886e-02  2.041e-01  -0.337 0.735917
## fp007               4.044e-02  1.152e-01   0.351 0.725643
## fp008               1.121e-01  1.636e-01   0.685 0.493331
## fp009              -8.242e-01  8.395e-01  -0.982 0.326536
## fp010               4.193e-01  3.136e-01   1.337 0.181579
## fp011               5.158e-02  2.198e-01   0.235 0.814503
## fp012              -1.346e-02  1.611e-01  -0.084 0.933452
## fp013              -4.519e-01  5.473e-01  -0.826 0.409311
## fp014               3.281e-01  4.550e-01   0.721 0.471044
## fp015              -1.839e-01  1.521e-01  -1.209 0.226971
## fp016              -1.367e-01  1.548e-01  -0.883 0.377340
## fp017              -1.704e-01  1.386e-01  -1.230 0.219187
## fp018              -3.824e-01  2.388e-01  -1.602 0.109655
## fp019              -3.131e-01  3.863e-01  -0.811 0.417862
```

```
## fp020              2.072e-01  2.135e-01   0.971 0.332078
## fp021             -5.956e-02  2.632e-01  -0.226 0.821060
## fp022              2.336e-01  3.456e-01   0.676 0.499180
## fp023             -3.193e-01  1.909e-01  -1.672 0.094866 .
## fp024             -4.272e-01  2.827e-01  -1.511 0.131162
## fp025              4.376e-01  4.538e-01   0.964 0.335184
## fp026              2.068e-01  2.564e-01   0.806 0.420273
## fp027              2.424e-01  2.429e-01   0.998 0.318594
## fp028              1.070e-01  1.200e-01   0.892 0.372547
## fp029             -9.857e-02  2.199e-01  -0.448 0.654163
## fp030             -2.361e-01  2.468e-01  -0.957 0.339048
## fp031              8.690e-02  1.346e-01   0.646 0.518754
## fp032             -1.204e+00  7.772e-01  -1.550 0.121628
## fp033              5.766e-01  4.236e-01   1.361 0.173882
## fp034             -1.794e-01  2.618e-01  -0.685 0.493486
## fp035             -2.140e-01  1.704e-01  -1.256 0.209605
## fp036              7.701e-02  1.657e-01   0.465 0.642133
## fp037              1.098e-01  1.725e-01   0.636 0.524693
## fp038              2.721e-01  1.888e-01   1.441 0.150030
## fp039              2.011e-02  2.888e-01   0.070 0.944491
## fp040              5.477e-01  1.890e-01   2.898 0.003873 **
## fp041             -4.265e-01  3.004e-01  -1.420 0.156143
## fp042             -9.901e-01  7.078e-01  -1.399 0.162294
## fp043             -3.725e-02  2.096e-01  -0.178 0.859011
## fp044             -3.860e-01  2.184e-01  -1.768 0.077562 .
## fp045              2.120e-01  1.299e-01   1.631 0.103238
## fp046             -3.504e-02  2.733e-01  -0.128 0.898010
## fp047             -1.675e-02  1.414e-01  -0.118 0.905775
## fp048              2.610e-01  2.434e-01   1.073 0.283810
## fp049              1.241e-01  1.971e-01   0.630 0.529036
## fp050              9.087e-03  1.410e-01   0.064 0.948648
## fp051              1.050e-01  2.014e-01   0.521 0.602210
## fp052             -4.569e-01  2.482e-01  -1.841 0.066029 .
## fp053              2.994e-01  2.466e-01   1.214 0.225129
## fp054              2.734e-02  1.829e-01   0.149 0.881229
## fp055             -3.662e-01  1.970e-01  -1.858 0.063530 .
## fp056             -2.961e-01  2.979e-01  -0.994 0.320541
## fp057             -1.002e-01  1.379e-01  -0.727 0.467703
## fp058              3.100e-01  8.074e-01   0.384 0.701129
## fp059             -1.615e-01  1.690e-01  -0.956 0.339514
## fp060              2.350e-01  1.474e-01   1.595 0.111209
## fp061             -6.365e-01  1.440e-01  -4.421 1.13e-05 ***
## fp062             -5.224e-01  2.961e-01  -1.764 0.078078 .
## fp063             -2.001e+00  1.287e+00  -1.554 0.120553
## fp064              2.549e-01  1.221e-01   2.087 0.037207 *
## fp065             -2.844e-01  1.197e-01  -2.377 0.017714 *
## fp066              2.093e-01  1.264e-01   1.655 0.098301 .
## fp067             -1.406e-01  1.540e-01  -0.913 0.361631
## fp068              4.964e-01  2.028e-01   2.447 0.014630 *
## fp069              1.324e-01  8.824e-02   1.501 0.133885
## fp070              3.453e-03  8.088e-02   0.043 0.965963
## fp071              1.474e-01  1.237e-01   1.192 0.233775
## fp072             -9.773e-01  2.763e-01  -3.537 0.000431 ***
## fp073             -4.671e-01  2.072e-01  -2.254 0.024474 *
```

```
## fp074              1.793e-01  1.206e-01   1.487 0.137566
## fp075              1.231e-01  1.035e-01   1.188 0.235034
## fp076              5.166e-01  1.704e-01   3.031 0.002525 **
## fp077              1.644e-01  1.236e-01   1.331 0.183739
## fp078             -3.715e-01  1.588e-01  -2.339 0.019608 *
## fp079              4.254e-01  1.881e-01   2.262 0.023992 *
## fp080              3.101e-01  1.554e-01   1.996 0.046340 *
## fp081             -3.208e-01  1.117e-01  -2.873 0.004192 **
## fp082              1.243e-01  9.524e-02   1.305 0.192379
## fp083             -6.916e-01  2.134e-01  -3.241 0.001248 **
## fp084              3.626e-01  2.381e-01   1.523 0.128171
## fp085             -3.310e-01  1.428e-01  -2.317 0.020785 *
## fp086              1.169e-02  9.774e-02   0.120 0.904834
## fp087              4.559e-02  2.797e-01   0.163 0.870568
## fp088              2.416e-01  9.959e-02   2.425 0.015534 *
## fp089              5.999e-01  2.320e-01   2.586 0.009915 **
## fp090             -2.450e-02  1.154e-01  -0.212 0.831930
## fp091             -2.858e-01  3.185e-01  -0.897 0.369847
## fp092              2.665e-01  2.069e-01   1.288 0.198156
## fp093              1.974e-01  1.087e-01   1.816 0.069803 .
## fp094             -1.991e-01  1.441e-01  -1.381 0.167707
## fp095             -1.403e-01  1.124e-01  -1.248 0.212449
## fp096             -5.024e-01  1.459e-01  -3.445 0.000605 ***
## fp097             -2.635e-01  1.666e-01  -1.582 0.114020
## fp098             -2.865e-01  1.633e-01  -1.754 0.079863 .
## fp099              2.592e-01  2.568e-01   1.009 0.313136
## fp100             -4.008e-01  3.034e-01  -1.321 0.186949
## fp101             -1.760e-01  3.019e-01  -0.583 0.560147
## fp102              2.445e-01  3.449e-01   0.709 0.478579
## fp103             -1.493e-01  9.148e-02  -1.632 0.103176
## fp104             -1.428e-01  1.176e-01  -1.214 0.225238
## fp105             -6.912e-02  1.395e-01  -0.495 0.620482
## fp106              1.128e-01  1.288e-01   0.876 0.381495
## fp107              2.778e+00  8.247e-01   3.369 0.000796 ***
## fp108              8.836e-03  1.852e-01   0.048 0.961970
## fp109              8.200e-01  2.267e-01   3.617 0.000319 ***
## fp110              3.680e-01  3.311e-01   1.111 0.266811
## fp111             -5.565e-01  1.420e-01  -3.918 9.80e-05 ***
## fp112             -1.079e-01  2.705e-01  -0.399 0.690108
## fp113              1.511e-01  9.481e-02   1.594 0.111478
## fp114             -1.201e-01  1.891e-01  -0.635 0.525628
## fp115             -1.896e-01  1.405e-01  -1.349 0.177736
## fp116              7.778e-03  1.897e-01   0.041 0.967300
## fp117              2.583e-01  1.779e-01   1.452 0.147070
## fp118             -1.964e-01  1.230e-01  -1.596 0.110940
## fp119              7.515e-01  2.630e-01   2.857 0.004402 **
## fp120             -1.814e-01  1.794e-01  -1.011 0.312362
## fp121             -4.731e-02  3.957e-01  -0.120 0.904866
## fp122              1.048e-01  1.041e-01   1.007 0.314268
## fp123              3.926e-02  1.765e-01   0.222 0.824066
## fp124              1.235e-01  1.705e-01   0.724 0.469243
## fp125             -2.633e-04  1.151e-01  -0.002 0.998175
## fp126             -2.782e-01  1.177e-01  -2.363 0.018373 *
## fp127             -6.123e-01  1.739e-01  -3.521 0.000457 ***
```

```
## fp128            -5.424e-01  1.932e-01  -2.807 0.005136 **
## fp129            -6.731e-02  2.243e-01  -0.300 0.764167
## fp130            -1.034e+00  4.106e-01  -2.518 0.012009 *
## fp131             2.158e-01  1.617e-01   1.335 0.182405
## fp132            -1.976e-01  2.382e-01  -0.830 0.406998
## fp133            -1.573e-01  1.217e-01  -1.293 0.196319
## fp134             2.496e+00  1.196e+00   2.086 0.037310 *
## fp135             1.818e-01  1.319e-01   1.379 0.168460
## fp136            -7.763e-02  3.131e-01  -0.248 0.804237
## fp137            -4.613e-02  2.978e-01  -0.155 0.876947
## fp138            -9.392e-02  1.906e-01  -0.493 0.622251
## fp139             7.659e-02  4.063e-01   0.189 0.850517
## fp140             3.145e-01  2.149e-01   1.463 0.143784
## fp141             2.219e-01  2.765e-01   0.802 0.422532
## fp142             6.272e-01  1.488e-01   4.214 2.83e-05 ***
## fp143             9.981e-01  2.929e-01   3.407 0.000692 ***
## fp144             2.207e-01  2.839e-01   0.777 0.437195
## fp145            -1.146e-01  1.188e-01  -0.964 0.335169
## fp146            -2.324e-01  2.086e-01  -1.114 0.265716
## fp147             1.502e-01  1.228e-01   1.223 0.221703
## fp148            -1.600e-01  1.319e-01  -1.213 0.225560
## fp149             1.172e-01  1.650e-01   0.710 0.477770
## fp150             9.046e-02  1.577e-01   0.574 0.566368
## fp151             2.899e-01  3.120e-01   0.929 0.353202
## fp152            -2.544e-01  2.990e-01  -0.851 0.395087
## fp153            -3.765e-01  2.773e-01  -1.358 0.175029
## fp154            -1.027e+00  2.033e-01  -5.054 5.50e-07 ***
## fp155             4.888e-01  2.916e-01   1.676 0.094163 .
## fp156            -3.602e-02  3.636e-01  -0.099 0.921109
## fp157            -4.715e-01  2.468e-01  -1.910 0.056505 .
## fp158             1.669e-02  1.925e-01   0.087 0.930943
## fp159             1.800e-01  2.432e-01   0.740 0.459378
## fp160             1.525e-02  2.177e-01   0.070 0.944155
## fp161            -2.440e-01  1.433e-01  -1.703 0.089063 .
## fp162             4.910e-02  1.859e-01   0.264 0.791710
## fp163             4.785e-01  3.121e-01   1.533 0.125659
## fp164             5.096e-01  1.899e-01   2.684 0.007446 **
## fp165             5.793e-01  2.146e-01   2.700 0.007103 **
## fp166            -6.582e-02  2.185e-01  -0.301 0.763293
## fp167            -6.044e-01  2.515e-01  -2.403 0.016502 *
## fp168            -1.187e-01  1.872e-01  -0.634 0.526173
## fp169            -1.705e-01  8.312e-02  -2.051 0.040650 *
## fp170            -7.902e-02  1.560e-01  -0.506 0.612745
## fp171             4.651e-01  1.186e-01   3.922 9.64e-05 ***
## fp172            -4.426e-01  2.440e-01  -1.814 0.070120 .
## fp173             4.243e-01  1.657e-01   2.561 0.010634 *
## fp174            -1.010e-01  2.098e-01  -0.481 0.630311
## fp175            -4.657e-02  2.481e-01  -0.188 0.851136
## fp176             9.736e-01  2.644e-01   3.682 0.000249 ***
## fp177             1.386e-01  2.393e-01   0.579 0.562538
## fp178             6.497e-02  2.079e-01   0.313 0.754691
## fp179            -3.415e-02  2.232e-01  -0.153 0.878437
## fp180            -7.905e-01  5.523e-01  -1.431 0.152839
## fp181             4.925e-01  3.218e-01   1.531 0.126309
```

```
## fp182              -1.124e-01  1.310e-01  -0.858 0.391384
## fp183               2.998e-01  7.143e-01   0.420 0.674836
## fp184               4.876e-01  1.580e-01   3.087 0.002103 **
## fp185              -3.778e-01  2.037e-01  -1.854 0.064108 .
## fp186              -3.654e-01  1.953e-01  -1.871 0.061710 .
## fp187               4.457e-01  2.682e-01   1.662 0.097015 .
## fp188               1.475e-01  1.258e-01   1.172 0.241519
## fp189              -1.984e-02  3.468e-01  -0.057 0.954384
## fp190               2.629e-01  3.018e-01   0.871 0.383981
## fp191               2.799e-01  1.465e-01   1.911 0.056388 .
## fp192              -2.404e-01  2.751e-01  -0.874 0.382534
## fp193               1.502e-01  1.494e-01   1.005 0.315159
## fp194               8.029e-01  6.379e-01   1.259 0.208566
## fp195               5.967e-02  3.435e-01   0.174 0.862158
## fp196               1.091e-02  2.544e-01   0.043 0.965812
## fp197              -3.736e-02  1.569e-01  -0.238 0.811793
## fp198               1.896e-01  2.665e-01   0.712 0.476893
## fp199              -9.932e-02  1.797e-01  -0.553 0.580702
## fp200              -6.421e-02  2.161e-01  -0.297 0.766462
## fp201              -4.838e-01  1.980e-01  -2.444 0.014771 *
## fp202               5.664e-01  1.869e-01   3.031 0.002527 **
## fp203               2.586e-01  6.447e-01   0.401 0.688462
## fp204              -1.371e-01  2.543e-01  -0.539 0.590008
## fp205               7.177e-02  1.561e-01   0.460 0.645857
## fp206              -6.769e-02  1.860e-01  -0.364 0.716094
## fp207              -5.538e-03  2.060e-01  -0.027 0.978560
## fp208              -5.338e-01  6.324e-01  -0.844 0.398925
## mol_weight         -1.232e+00  2.296e-01  -5.365 1.09e-07 ***
## num_atoms          -1.478e+01  3.473e+00  -4.257 2.35e-05 ***
## num_non_h_atoms     1.795e+01  3.166e+00   5.670 2.07e-08 ***
## num_bonds           9.843e+00  2.681e+00   3.671 0.000260 ***
## num_non_h_bonds    -1.030e+01  1.793e+00  -5.746 1.35e-08 ***
## num_mult_bonds      2.107e-01  1.754e-01   1.201 0.229990
## num_rot_bonds      -5.213e-01  1.334e-01  -3.908 0.000102 ***
## num_dbl_bonds      -7.492e-01  3.163e-01  -2.369 0.018111 *
## num_aromatic_bonds -2.364e+00  6.232e-01  -3.794 0.000161 ***
## num_hydrogen        8.347e-01  1.880e-01   4.439 1.04e-05 ***
## num_carbon          1.730e-02  3.763e-01   0.046 0.963335
## num_nitrogen        6.125e+00  3.045e+00   2.011 0.044645 *
## num_oxygen          2.389e+00  4.523e-01   5.283 1.69e-07 ***
## num_sulfer         -8.508e+00  3.619e+00  -2.351 0.018994 *
## num_chlorine       -7.449e+00  1.989e+00  -3.744 0.000195 ***
## num_halogen         1.408e+00  2.109e+00   0.668 0.504615
## num_rings           1.276e+00  6.716e-01   1.901 0.057731 .
## hydrophilic_factor  1.099e-02  1.137e-01   0.097 0.922998
## surface_area1       8.825e-02  6.058e-02   1.457 0.145643
## surface_area2       9.555e-02  5.615e-02   1.702 0.089208 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5524 on 722 degrees of freedom
## Multiple R-squared:  0.9446, Adjusted R-squared:  0.9271
## F-statistic: 54.03 on 228 and 722 DF,  p-value: < 2.2e-16
```

Calculate the mean square error using the test data

```
pred_lm_tr = predict(fit_lm_tr, test_data)
mse_test = mean((pred_lm_tr - test_data$solubility)^2);mse_test
```

```
## [1] 0.5558898
```

Hence, the MSE using test data is 0.5558898.

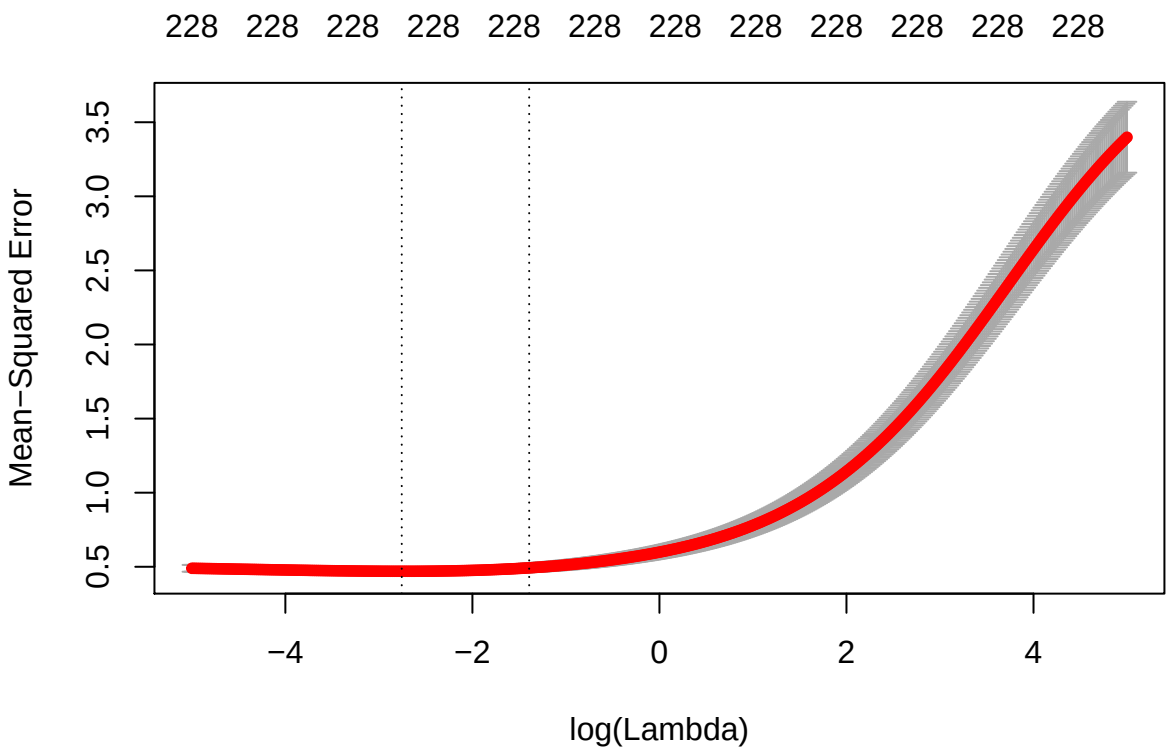
(b) Fit a ridge regression model on the training data, with lambda chosen by cross-validation. Report the test error.

Fit ridge regression model on the training data

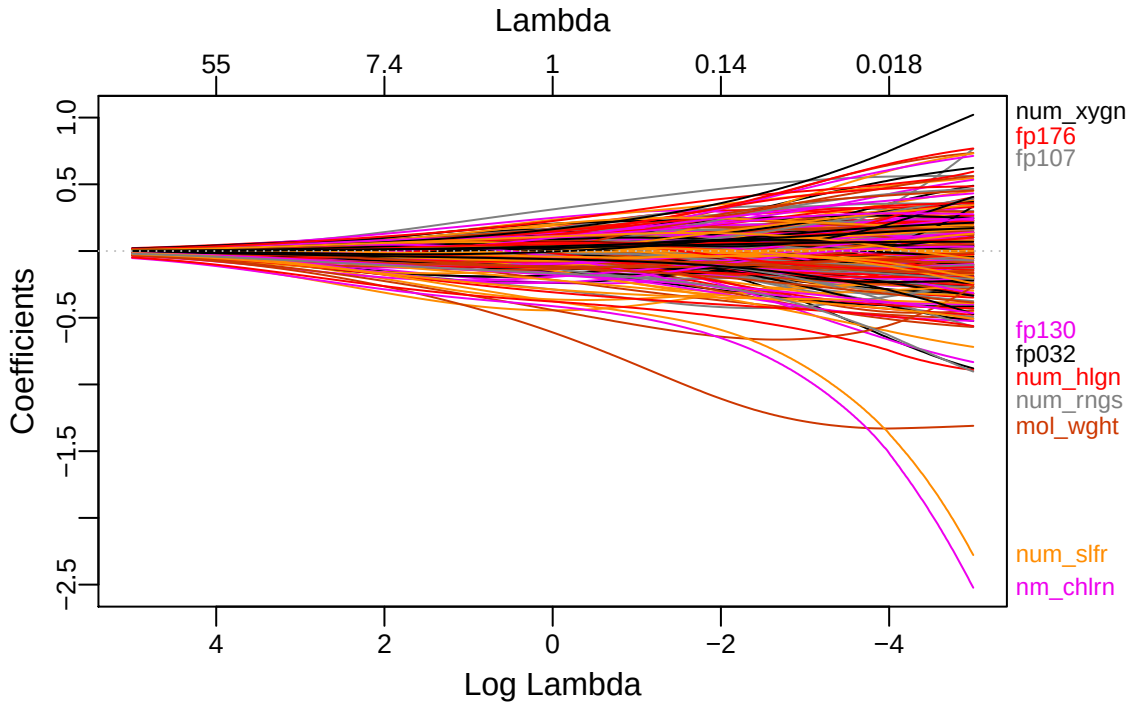
```
set.seed(1)
train_data = na.omit(train_data)
x = model.matrix(solubility ~ ., train_data)[, -1]
y = train_data$solubility

ridge_mod = glmnet(x, y, alpha = 0, lambda = exp(seq(-5, 5, length = 500)))
mat_coef = coef(ridge_mod)
dim(mat_coef)
```

```
## [1] 229 500
```

```
# Cross-validation

cv_ridge = cv.glmnet(x, y,
                     alpha = 0,
                     lambda = exp(seq(-5, 5, length = 500)),
                     type.measure = "mse")
plot(cv_ridge)
```




```
# Trace plot
plot_glmnet(ridge_mod, xvar = "rlambda")
```




```
# Predict response in final model
best_lambda = cv_ridge$lambda.min; best_lambda
```

```
## [1] 0.06357652
```

```
pred_resp_ridge = predict(ridge_mod, newx = model.matrix(solubility ~ ., test_data)[, -1], s = best_lamb
```

```
##                1
## 1     0.48941521
## 2     0.11954505
## 3    -0.52518669
## 4     0.77464701
## 5     0.04274718
## 6     1.48402389
## 7     0.59571708
## 8     0.44460328
## 9     0.54569703
## 10   -0.60731789
## 11   -0.48700381
## 12   -1.44341132
## 13    0.21797590
## 14   -0.11690294
## 15   -0.81545601
## 16   -0.79566115
## 17   -0.27915690
## 18    0.13265181
## 19    0.51922903
```

```
## 20  -0.89994256
## 21   0.44029365
## 22  -0.21268701
## 23  -0.63829771
## 24  -0.51901906
## 25  -1.08864725
## 26  -0.18314230
## 27  -0.56384190
## 28   1.07456060
## 29  -1.86607371
## 30  -1.10395379
## 31  -2.37247429
## 32  -0.84980627
## 33  -0.74129458
## 34  -0.03779135
## 35  -0.02825838
## 36  -1.11852580
## 37   0.43085907
## 38  -0.71691138
## 39   0.08813257
## 40  -0.89310196
## 41  -0.68545769
## 42  -2.18700925
## 43  -1.13793031
## 44  -0.01042009
## 45  -1.66567979
## 46  -0.92660733
## 47  -0.75395714
## 48  -0.95143961
## 49  -0.80900176
## 50  -0.84629344
## 51  -0.03264756
## 52  -1.48423869
## 53  -1.52318887
## 54  -1.40850897
## 55  -1.87907968
## 56  -1.35914058
## 57  -1.27721138
## 58  -1.57046057
## 59  -1.84108872
## 60  -1.83503232
## 61  -0.99716791
## 62  -1.36087808
## 63  -1.19420759
## 64  -1.22891005
## 65  -1.25554163
## 66  -1.89322519
## 67  -2.85054804
## 68  -1.96868495
## 69  -1.44190738
## 70  -2.94780259
## 71  -2.13755350
## 72  -2.66113105
## 73  -1.80064123
```

```
## 74  -3.12660197
## 75  -1.98654846
## 76  -2.72387560
## 77  -2.60029610
## 78  -2.05552587
## 79  -1.78215781
## 80  -2.28297819
## 81  -1.13716258
## 82  -1.68051428
## 83  -2.26955622
## 84  -2.03210987
## 85  -1.08270271
## 86  -1.51388770
## 87  -3.14204785
## 88  -2.30561691
## 89  -2.11917211
## 90  -2.30037402
## 91  -1.77356976
## 92  -2.17811736
## 93  -2.53008405
## 94  -1.67788704
## 95  -0.53651206
## 96  -2.37770516
## 97  -1.85407864
## 98  -2.26567660
## 99  -1.54192051
## 100 -2.05835822
## 101 -2.03866195
## 102 -2.20922029
## 103 -2.07149231
## 104 -2.39072395
## 105 -2.72033818
## 106 -2.09414166
## 107 -2.35198752
## 108 -2.45410779
## 109 -3.13351643
## 110 -3.25558941
## 111 -2.71511337
## 112 -3.14981973
## 113 -3.08925295
## 114 -3.15065451
## 115 -2.71951923
## 116 -2.59466440
## 117 -2.81059608
## 118 -2.46183277
## 119 -2.91891499
## 120 -2.88247881
## 121 -2.38688889
## 122 -1.29834060
## 123 -3.63586052
## 124 -2.94562278
## 125 -2.88418911
## 126 -2.82374107
## 127 -3.57255917
```

```
## 128 -3.08366205
## 129 -3.55755934
## 130 -2.85078624
## 131 -3.04463093
## 132 -3.60424596
## 133 -2.41257058
## 134 -3.67114895
## 135 -2.45902207
## 136 -3.43962909
## 137 -2.54587753
## 138 -2.97778025
## 139 -2.85511753
## 140 -2.34432697
## 141 -2.81707549
## 142 -2.05016993
## 143 -3.53629135
## 144 -2.69722233
## 145 -3.14285931
## 146 -3.47116421
## 147 -2.68979001
## 148 -3.18042276
## 149 -3.51139914
## 150 -3.69230125
## 151 -1.96357403
## 152 -3.17037960
## 153 -2.44653595
## 154 -3.75563118
## 155 -2.97684129
## 156 -3.12417833
## 157 -4.40618079
## 158 -5.02368960
## 159 -3.92560741
## 160 -4.26936374
## 161 -5.50305006
## 162 -4.15812854
## 163 -3.28325040
## 164 -4.65626306
## 165 -4.95892461
## 166 -3.52366733
## 167 -4.67558097
## 168 -4.07368733
## 169 -4.79058878
## 170 -4.58018946
## 171 -3.75860798
## 172 -3.71369287
## 173 -3.54226281
## 174 -4.88979263
## 175 -4.86750008
## 176 -4.06452857
## 177 -3.93597182
## 178 -4.70307544
## 179 -4.40601781
## 180 -3.11256171
## 181 -4.78132235
```

```
## 182 -3.77441273
## 183 -4.68527076
## 184 -4.40275841
## 185 -3.94210697
## 186 -3.82904815
## 187 -4.63957456
## 188 -4.97156400
## 189 -6.05367785
## 190 -5.89975318
## 191 -4.37176830
## 192 -2.91465402
## 193 -4.44010826
## 194 -4.79651145
## 195 -4.49098343
## 196 -4.47585737
## 197 -5.66239763
## 198 -4.42807364
## 199 -4.97029227
## 200 -5.29421195
## 201 -7.29429220
## 202 -6.50768715
## 203 -6.26488542
## 204 -6.75542021
## 205 -5.84476628
## 206 -5.83540962
## 207 -5.58453735
## 208 -5.82794716
## 209 -6.90394506
## 210 -6.76302191
## 211 -7.18299243
## 212 -7.00430738
## 213 -7.66745579
## 214 -7.89481367
## 215 -8.52332168
## 216 -7.61442629
## 217 -0.01366852
## 218  0.38937754
## 219  0.28550555
## 220 -0.11459005
## 221 -1.15880055
## 222 -0.57573048
## 223 -0.93596334
## 224 -0.95082734
## 225 -2.21123389
## 226 -0.88089380
## 227 -0.89351239
## 228 -0.97272349
## 229 -0.48208605
## 230 -1.80088228
## 231 -1.46454020
## 232 -1.44302635
## 233 -0.85529946
## 234 -0.02754922
## 235 -1.46207290
```

```
## 236 -1.01273990
## 237 -3.41718984
## 238 -1.64343007
## 239 -1.52318887
## 240 -1.36438755
## 241 -0.56666305
## 242 -2.01897841
## 243 -1.58005890
## 244 -2.33300479
## 245 -1.39276093
## 246 -0.54979646
## 247 -1.43740364
## 248 -1.12469677
## 249 -1.00943133
## 250 -2.14121154
## 251 -1.89580656
## 252 -2.13041435
## 253 -2.50686885
## 254 -3.52778688
## 255 -2.51171168
## 256 -0.98861031
## 257 -1.64396301
## 258 -1.66656398
## 259 -4.30314364
## 260 -1.44570576
## 261 -2.29107242
## 262 -2.37374424
## 263 -2.78872951
## 264 -3.13725459
## 265 -2.38182752
## 266 -1.46152156
## 267 -2.51672098
## 268 -2.15853794
## 269 -2.72878600
## 270 -2.90411823
## 271 -2.71296132
## 272 -3.45424533
## 273 -3.40008484
## 274 -3.68383973
## 275 -2.88411645
## 276 -3.56298902
## 277 -3.58225018
## 278 -2.71323720
## 279 -3.65103730
## 280 -3.15989742
## 281 -2.36488303
## 282 -3.89056335
## 283 -3.61175742
## 284 -3.83600792
## 285 -3.99397284
## 286 -4.34697930
## 287 -3.36126947
## 288 -3.04944539
## 289 -4.17946081
```

```
## 290 -5.00436278
## 291 -4.45073684
## 292 -4.33700076
## 293 -3.24983082
## 294 -4.60340212
## 295 -4.26427089
## 296 -4.01078363
## 297 -4.17190024
## 298 -3.92192875
## 299 -5.03244870
## 300 -5.66312875
## 301 -5.43425708
## 302 -5.33282348
## 303 -6.41473381
## 304 -5.53616260
## 305 -5.68721419
## 306 -6.92660532
## 307 -7.42465406
## 308 -8.11361513
## 309 -8.13628064
## 310 -8.45427791
## 311 -8.91294552
## 312 -7.16826696
## 313 -2.03394224
## 314 -2.64630312
## 315 -4.65836828
## 316 -4.33980890
```

```
# MSE
mse_ridge = mean((pred_resp_ridge - test_data$solubility)^2); mse_ridge
```

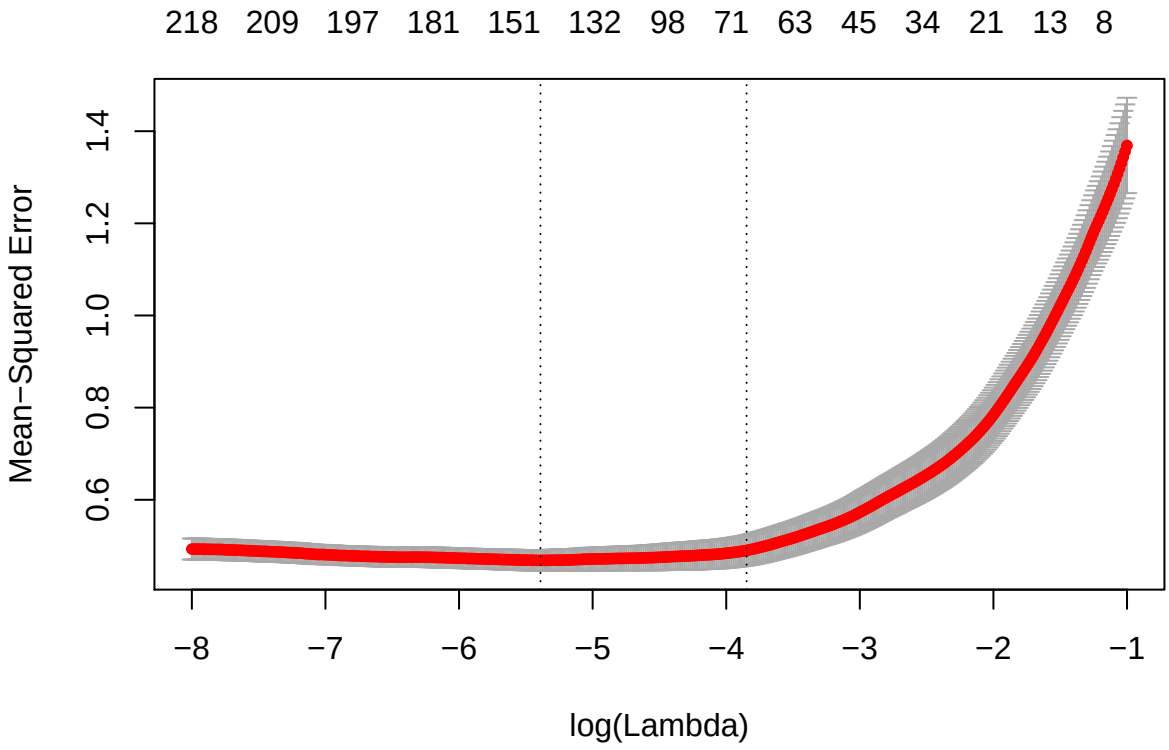
```
## [1] 0.5126573
```

Based on the result, the MSE for ridge regression is 0.5126573.

(c) Fit a lasso model on the training data, with lambda chosen by cross-validation. Report the test error, along with the number of non-zero coefficient estimates.

Fit lasso model on the training data

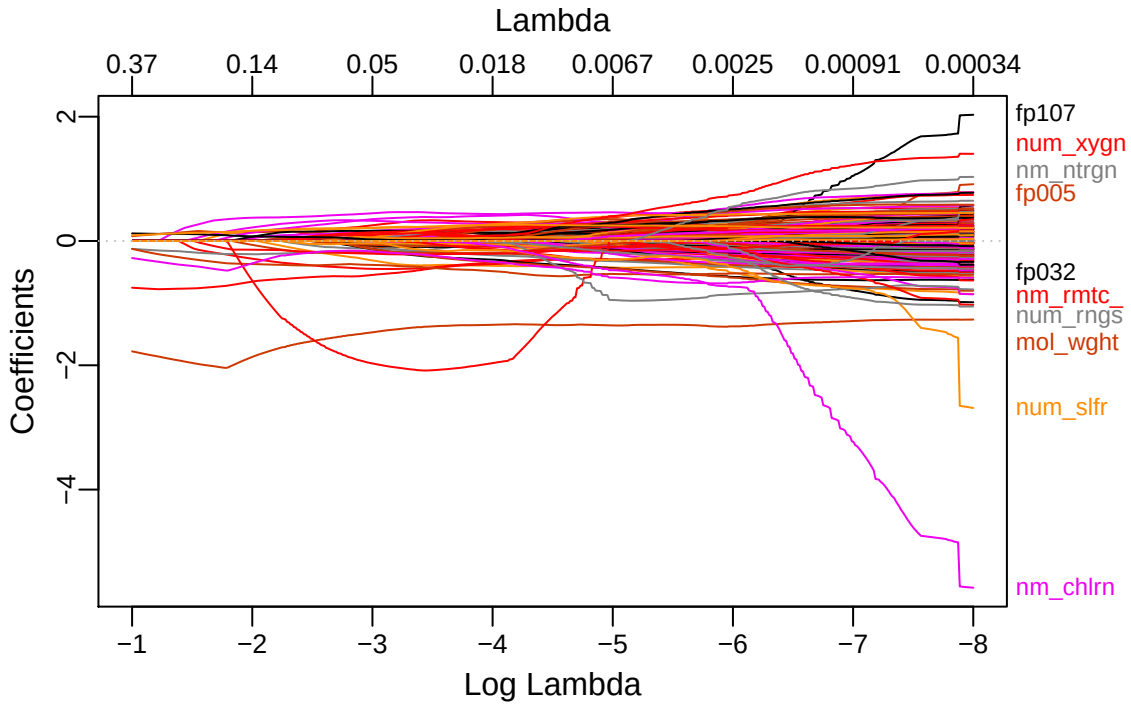
```
set.seed(1)
cv_lasso = cv.glmnet(x, y, alpha = 1, lambda = exp(seq(-8, -1, length = 500)))

# Cross-validation
plot(cv_lasso)
```




```
cv_lasso$lambda.min
```

```
## [1] 0.004558409
```

```
# Trace plot
plot_glmnet(cv_lasso$glmnet.fit)
```




```
# Predict response in the final model
pred_resp_lasso = predict(cv_lasso, newx = model.matrix(solubility ~ ., test_data)[, -1], s = cv_lasso$
```

```
##                1
## 1    0.607503459
## 2    0.161469313
## 3   -0.459656322
## 4    0.728395275
## 5   -0.023925904
## 6    1.501171940
## 7    0.547907832
## 8    0.401744085
## 9    0.499373514
## 10  -0.527721885
## 11  -0.404097416
## 12  -1.519677063
## 13   0.098065939
## 14  -0.075045987
## 15  -0.728119392
## 16  -0.786844859
## 17  -0.279413989
## 18   0.001382503
## 19   0.988889602
## 20  -0.927188301
## 21   0.570639335
## 22  -0.046398361
```

```
## 23  -0.578908669
## 24  -0.311179902
## 25  -0.985776350
## 26  -0.151723788
## 27  -0.593149620
## 28   0.941800730
## 29  -1.934628602
## 30  -1.176433170
## 31  -2.278254172
## 32  -0.840148664
## 33  -0.761857445
## 34  -0.051779328
## 35  -0.112690831
## 36  -1.149212692
## 37   0.580083675
## 38  -0.692117217
## 39   0.038082961
## 40  -0.926465150
## 41  -0.705651019
## 42  -2.058799611
## 43  -1.136180115
## 44   0.021358197
## 45  -1.733333445
## 46  -0.694702533
## 47  -0.591830735
## 48  -0.877099036
## 49  -0.837719211
## 50  -1.033406538
## 51  -0.271270003
## 52  -1.578044612
## 53  -1.566719417
## 54  -1.391127375
## 55  -1.797711329
## 56  -1.305282808
## 57  -1.358399849
## 58  -1.691059540
## 59  -1.949377379
## 60  -1.696162463
## 61  -0.741048196
## 62  -1.295633190
## 63  -0.818979326
## 64  -1.264194253
## 65  -1.425777290
## 66  -1.933577879
## 67  -2.705080533
## 68  -2.026835815
## 69  -1.409286806
## 70  -2.715096141
## 71  -2.111840771
## 72  -2.763229132
## 73  -1.974793907
## 74  -2.944139760
## 75  -2.004093949
## 76  -2.908618804
```

```
## 77  -2.676927617
## 78  -2.077323073
## 79  -1.611319396
## 80  -2.310163885
## 81  -1.061671483
## 82  -1.554429316
## 83  -2.183309686
## 84  -2.085571249
## 85  -1.127567507
## 86  -1.535874696
## 87  -3.002831126
## 88  -2.350784790
## 89  -2.127778527
## 90  -2.333183718
## 91  -1.836233972
## 92  -2.308311270
## 93  -2.663910244
## 94  -1.492580363
## 95  -0.652088672
## 96  -2.562455991
## 97  -1.876553057
## 98  -2.214800553
## 99  -1.519981720
## 100 -2.120005441
## 101 -2.160931799
## 102 -2.155884897
## 103 -2.364156227
## 104 -2.263400024
## 105 -2.804428768
## 106 -2.261709059
## 107 -2.418252054
## 108 -2.503701068
## 109 -3.016921698
## 110 -3.086911589
## 111 -2.827016996
## 112 -2.968212409
## 113 -3.130034484
## 114 -2.832459215
## 115 -2.675056363
## 116 -2.712288654
## 117 -2.853784258
## 118 -2.827001298
## 119 -2.924302584
## 120 -2.776130179
## 121 -2.284710376
## 122 -1.344876450
## 123 -3.375220603
## 124 -2.922380765
## 125 -2.918563238
## 126 -2.910143061
## 127 -3.584378341
## 128 -3.180393670
## 129 -3.479145629
## 130 -3.037560245
```

```
## 131 -3.044449490
## 132 -3.518862586
## 133 -2.422785487
## 134 -3.806415772
## 135 -2.673584785
## 136 -3.588177494
## 137 -2.393775366
## 138 -2.992242690
## 139 -2.553527575
## 140 -2.118017306
## 141 -2.898358008
## 142 -2.138733391
## 143 -3.397383945
## 144 -2.686699036
## 145 -3.168618561
## 146 -3.591477727
## 147 -2.667419395
## 148 -3.291282445
## 149 -3.518668422
## 150 -3.704745795
## 151 -2.151452816
## 152 -3.047337063
## 153 -2.358483175
## 154 -3.791783764
## 155 -3.008058380
## 156 -3.032006410
## 157 -4.385352812
## 158 -4.892594544
## 159 -3.865739840
## 160 -4.048165212
## 161 -5.476500481
## 162 -4.102820065
## 163 -3.184154649
## 164 -4.583354994
## 165 -4.836895121
## 166 -3.481131425
## 167 -4.669700584
## 168 -4.044135477
## 169 -4.691195653
## 170 -4.428211357
## 171 -3.901249604
## 172 -3.529492499
## 173 -3.506076687
## 174 -4.979355404
## 175 -4.667747904
## 176 -4.102414542
## 177 -3.968431612
## 178 -4.561799324
## 179 -4.417707404
## 180 -2.947188866
## 181 -4.816619246
## 182 -3.585725789
## 183 -4.627039864
## 184 -4.439645907
```

```
## 185 -3.902197926
## 186 -3.698985168
## 187 -4.681459009
## 188 -4.537717594
## 189 -6.024041808
## 190 -5.689297036
## 191 -4.278160969
## 192 -2.958129818
## 193 -4.588564944
## 194 -4.793571431
## 195 -4.658946607
## 196 -4.367833006
## 197 -5.658814974
## 198 -4.401064886
## 199 -4.967252332
## 200 -4.981372946
## 201 -7.129863803
## 202 -6.497883505
## 203 -6.332259245
## 204 -6.792185189
## 205 -5.838168042
## 206 -5.931015395
## 207 -5.304535725
## 208 -5.875747490
## 209 -7.079145099
## 210 -6.781264784
## 211 -7.223703581
## 212 -7.135208781
## 213 -7.663672616
## 214 -7.921458457
## 215 -8.393637758
## 216 -7.659293006
## 217 -0.043018856
## 218  0.416193612
## 219  0.533183102
## 220 -0.027584410
## 221 -1.184368700
## 222 -0.490217101
## 223 -0.856800974
## 224 -0.849842443
## 225 -2.257852966
## 226 -0.843276164
## 227 -0.878244191
## 228 -1.125891198
## 229 -0.496539969
## 230 -1.928711069
## 231 -1.296196187
## 232 -1.197259847
## 233 -0.911153845
## 234 -0.032932829
## 235 -1.499068065
## 236 -1.260153031
## 237 -2.803608556
## 238 -1.732179106
```

```
## 239 -1.566719417
## 240 -1.182795054
## 241 -0.554756813
## 242 -2.162108397
## 243 -1.616288594
## 244 -2.082377916
## 245 -1.322335154
## 246 -0.702946027
## 247 -1.443719073
## 248 -1.207500969
## 249 -1.105523955
## 250 -2.156034465
## 251 -2.124846373
## 252 -1.759569773
## 253 -2.688436488
## 254 -3.235853972
## 255 -2.277634398
## 256 -1.056426655
## 257 -1.501512327
## 258 -1.851856414
## 259 -4.187592374
## 260 -1.619770340
## 261 -2.369608949
## 262 -2.328253571
## 263 -2.985534388
## 264 -3.151152699
## 265 -2.328996784
## 266 -1.907767102
## 267 -2.563974431
## 268 -2.229456823
## 269 -2.741320780
## 270 -3.002873020
## 271 -2.747115552
## 272 -3.596832716
## 273 -3.331850092
## 274 -3.549603849
## 275 -2.894477029
## 276 -3.704183155
## 277 -3.643537251
## 278 -2.717723199
## 279 -3.653494086
## 280 -3.216870277
## 281 -2.205456080
## 282 -3.988234359
## 283 -3.419962998
## 284 -3.883757696
## 285 -4.085669685
## 286 -4.041602293
## 287 -3.176811958
## 288 -3.372884521
## 289 -4.322121721
## 290 -5.015731053
## 291 -4.636521472
## 292 -4.316179915
```

```
## 293 -2.991549628
## 294 -4.557501982
## 295 -4.256074032
## 296 -4.121907916
## 297 -4.068954697
## 298 -4.011152116
## 299 -5.008837766
## 300 -5.692589640
## 301 -5.142022345
## 302 -5.362343279
## 303 -6.440900751
## 304 -5.389124184
## 305 -5.763560216
## 306 -6.880909862
## 307 -7.498048357
## 308 -8.167710985
## 309 -8.201669222
## 310 -8.585999423
## 311 -8.938689356
## 312 -7.155254638
## 313 -2.142272870
## 314 -2.528473843
## 315 -4.634193113
## 316 -4.588458037
```

```
# MSE
mse_lasso = mean((pred_resp_lasso - test_data$solubility)^2); mse_lasso
```

```
## [1] 0.4995506
```

```
# Number of non-zero coefficient estimates
dim(as.matrix(predict(cv_lasso, s = "lambda.min", type = "coefficients")@x))
```

```
## [1] 144   1
```

Thus, we know the MSE for lasso model is 0.4995506, and the number of non-zero coefficient estimates is 144.

(d) Fit a PCR model on the training data, with M chosen by cross-validation. Report the test error, along with the value of M selected by cross-validation.

Fit PCR model on training data

```
set.seed(1)
pcr_mod = pcr(solubility ~ .,
              data = train_data
              scale = T,
              validation = "CV")
summary(pcr_mod)
```

```
## Data:    X dimension: 951 228
##  Y dimension: 951 1
## Fit method: svdpc
```

```
## Number of components considered: 228
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           2.048    2.041    1.979    1.715    1.605    1.565    1.445
## adjCV        2.048    2.041    1.979    1.714    1.600    1.595    1.444
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       1.292    1.290    1.292     1.267     1.244     1.243     1.244
## adjCV    1.289    1.288    1.291     1.266     1.240     1.242     1.243
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## CV        1.194     1.169     1.111     1.053     1.049     1.034
## adjCV     1.193     1.168     1.107     1.048     1.047     1.032
##        20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## CV        1.012     1.006     1.003    0.9745    0.9736    0.9732
## adjCV     1.010     1.004     1.006    0.9716    0.9716    0.9720
##        26 comps  27 comps  28 comps  29 comps  30 comps  31 comps
## CV       0.9673    0.9613    0.9580    0.9589    0.9588    0.9449
## adjCV    0.9630    0.9594    0.9568    0.9575    0.9585    0.9379
##        32 comps  33 comps  34 comps  35 comps  36 comps  37 comps
## CV       0.9302    0.9188    0.9174    0.9017    0.8915    0.8842
## adjCV    0.9291    0.9156    0.9157    0.8998    0.8890    0.8804
##        38 comps  39 comps  40 comps  41 comps  42 comps  43 comps
## CV       0.8846    0.8801    0.8732    0.8711    0.8716    0.8696
## adjCV    0.8823    0.8799    0.8685    0.8679    0.8713    0.8706
##        44 comps  45 comps  46 comps  47 comps  48 comps  49 comps
## CV       0.8542    0.8492    0.8441    0.8385    0.8398    0.8420
## adjCV    0.8488    0.8445    0.8424    0.8352    0.8375    0.8392
##        50 comps  51 comps  52 comps  53 comps  54 comps  55 comps
## CV       0.8364    0.8334    0.8283    0.8281    0.8279    0.8303
## adjCV    0.8377    0.8312    0.8249    0.8250    0.8255    0.8280
##        56 comps  57 comps  58 comps  59 comps  60 comps  61 comps
## CV       0.8290    0.8276    0.8212    0.8206    0.8244    0.8125
## adjCV    0.8274    0.8273    0.8238    0.8200    0.8219    0.8058
##        62 comps  63 comps  64 comps  65 comps  66 comps  67 comps
## CV       0.8063    0.8092    0.8070    0.8063    0.8088    0.8068
## adjCV    0.8012    0.8051    0.8036    0.8038    0.8065    0.8028
##        68 comps  69 comps  70 comps  71 comps  72 comps  73 comps
## CV       0.8069    0.8065    0.8077    0.8023    0.7965    0.7945
## adjCV    0.8040    0.8032    0.8061    0.8029    0.7908    0.7895
##        74 comps  75 comps  76 comps  77 comps  78 comps  79 comps
## CV       0.7962    0.7943    0.7956    0.7976    0.7960    0.7935
## adjCV    0.7920    0.7906    0.7915    0.7945    0.7929    0.7892
##        80 comps  81 comps  82 comps  83 comps  84 comps  85 comps
## CV       0.7940    0.7951    0.7924    0.7917    0.7916    0.7873
## adjCV    0.7912    0.7918    0.7882    0.7883    0.7884    0.7831
##        86 comps  87 comps  88 comps  89 comps  90 comps  91 comps
## CV       0.7876    0.7887    0.7900    0.7902    0.7899    0.7861
## adjCV    0.7836    0.7850    0.7873    0.7888    0.7874    0.7814
##        92 comps  93 comps  94 comps  95 comps  96 comps  97 comps
## CV       0.7816    0.7783    0.7765    0.7778    0.7789    0.7795
## adjCV    0.7742    0.7714    0.7705    0.7721    0.7737    0.7756
##        98 comps  99 comps  100 comps  101 comps  102 comps  103 comps
## CV       0.7808    0.7781     0.7732     0.7708     0.7728     0.7714
```

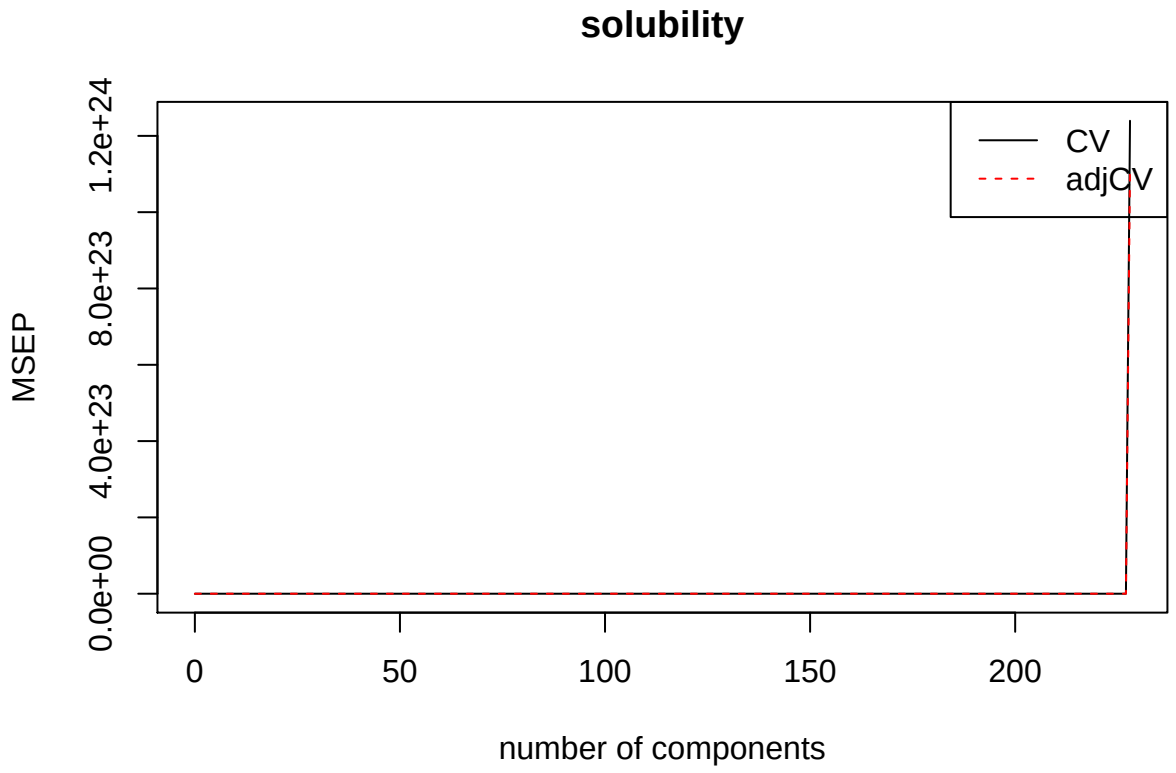
```
## adjCV    0.7772    0.7758     0.7677     0.7662     0.7684     0.7676
##        104 comps  105 comps  106 comps  107 comps  108 comps  109 comps
## CV        0.7698     0.7677     0.7688     0.7701     0.7703     0.7707
## adjCV     0.7631     0.7616     0.7641     0.7663     0.7677     0.7629
##        110 comps  111 comps  112 comps  113 comps  114 comps  115 comps
## CV        0.7700     0.7672     0.7623     0.7604     0.7604     0.7580
## adjCV     0.7642     0.7624     0.7593     0.7556     0.7553     0.7553
##        116 comps  117 comps  118 comps  119 comps  120 comps  121 comps
## CV        0.7588     0.7581     0.7584     0.7571     0.7510     0.7523
## adjCV     0.7554     0.7558     0.7578     0.7550     0.7504     0.7468
##        122 comps  123 comps  124 comps  125 comps  126 comps  127 comps
## CV        0.7460     0.7459     0.7452     0.7433     0.7458     0.7468
## adjCV     0.7427     0.7408     0.7414     0.7362     0.7394     0.7416
##        128 comps  129 comps  130 comps  131 comps  132 comps  133 comps
## CV        0.7459     0.7417     0.7408     0.7363     0.7326     0.7321
## adjCV     0.7423     0.7372     0.7385     0.7275     0.7232     0.7244
##        134 comps  135 comps  136 comps  137 comps  138 comps  139 comps
## CV        0.7285     0.7280     0.7270     0.7275     0.7295     0.7301
## adjCV     0.7226     0.7229     0.7204     0.7202     0.7227     0.7242
##        140 comps  141 comps  142 comps  143 comps  144 comps  145 comps
## CV        0.7244     0.7187     0.7177     0.7191     0.7160     0.7167
## adjCV     0.7186     0.7124     0.7129     0.7166     0.7083     0.7098
##        146 comps  147 comps  148 comps  149 comps  150 comps  151 comps
## CV        0.7160     0.7097     0.7079     0.7075     0.7054     0.7066
## adjCV     0.7086     0.7039     0.7000     0.7003     0.6977     0.6992
##        152 comps  153 comps  154 comps  155 comps  156 comps  157 comps
## CV        0.7087     0.7101     0.7071     0.7089     0.7123     0.7081
## adjCV     0.7020     0.7034     0.7014     0.7014     0.7054     0.6998
##        158 comps  159 comps  160 comps  161 comps  162 comps  163 comps
## CV        0.7071     0.7093     0.7093     0.7081     0.7113     0.7135
## adjCV     0.6985     0.7013     0.7015     0.7001     0.7038     0.7067
##        164 comps  165 comps  166 comps  167 comps  168 comps  169 comps
## CV        0.7140     0.7153     0.7146     0.7126     0.7135     0.7141
## adjCV     0.7059     0.7079     0.7062     0.7047     0.7053     0.7061
##        170 comps  171 comps  172 comps  173 comps  174 comps  175 comps
## CV        0.7176     0.7187     0.7195     0.7184     0.7195     0.7209
## adjCV     0.7093     0.7103     0.7115     0.7108     0.7109     0.7123
##        176 comps  177 comps  178 comps  179 comps  180 comps  181 comps
## CV        0.7215     0.7236     0.7228     0.7217     0.7184     0.7206
## adjCV     0.7119     0.7141     0.7139     0.7124     0.7100     0.7113
##        182 comps  183 comps  184 comps  185 comps  186 comps  187 comps
## CV        0.7254     0.7245     0.7211     0.7227     0.7204     0.7192
## adjCV     0.7165     0.7149     0.7119     0.7137     0.7103     0.7093
##        188 comps  189 comps  190 comps  191 comps  192 comps  193 comps
## CV        0.7216     0.7215     0.7222     0.7253     0.7265     0.7232
## adjCV     0.7118     0.7113     0.7118     0.7151     0.7168     0.7145
##        194 comps  195 comps  196 comps  197 comps  198 comps  199 comps
## CV        0.7222     0.7232     0.7232     0.7261     0.7244     0.7298
## adjCV     0.7128     0.7145     0.7118     0.7153     0.7148     0.7205
##        200 comps  201 comps  202 comps  203 comps  204 comps  205 comps
## CV        0.7341      0.739     0.7360     0.7348     0.7341     0.7351
## adjCV     0.7230      0.728     0.7244     0.7236     0.7236     0.7240
##        206 comps  207 comps  208 comps  209 comps  210 comps  211 comps
## CV        0.7339     0.7403     0.7321     0.7336     0.7369     0.7407
```

```
## adjCV     0.7218     0.7286     0.7199     0.7213     0.7245     0.7282
##        212 comps  213 comps  214 comps  215 comps  216 comps  217 comps
## CV        0.7383     0.7444     0.7445     0.7413     0.7360     0.7409
## adjCV     0.7258     0.7318     0.7315     0.7286     0.7238     0.7270
##        218 comps  219 comps  220 comps  221 comps  222 comps  223 comps
## CV        0.7379     0.7396     0.7400     0.7441     0.7401     0.7440
## adjCV     0.7246     0.7263     0.7269     0.7310     0.7272     0.7302
##        224 comps  225 comps  226 comps  227 comps  228 comps
## CV        0.7403     0.7391     0.7420     0.7401  1.113e+12
## adjCV     0.7273     0.7253     0.7281     0.7278  1.056e+12
##
## TRAINING: % variance explained
##             1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X            12.417   23.083    30.29    34.91    39.27    43.53    46.98
## solubility    0.734    7.182    30.52    39.36    39.52    50.82    60.83
##             8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## X             50.08    53.04     55.46     57.67     59.81     61.72
## solubility    61.00    61.01     62.57     64.10     64.17     64.36
##             14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
## X              63.43     64.82     66.16     67.40     68.58     69.68
## solubility     67.12     68.79     71.69     74.75     74.96     75.59
##             20 comps  21 comps  22 comps  23 comps  24 comps  25 comps
## X              70.73     71.76     72.72     73.64     74.48     75.31
## solubility     76.72     76.96     77.00     78.34     78.40     78.48
##             26 comps  27 comps  28 comps  29 comps  30 comps  31 comps
## X              76.09     76.85     77.57     78.29     78.95     79.59
## solubility     78.97     79.03     79.20     79.42     79.43     80.34
##             32 comps  33 comps  34 comps  35 comps  36 comps  37 comps
## X              80.22     80.81     81.38     81.92     82.46     82.96
## solubility     80.94     81.52     81.53     82.15     82.66     82.96
##             38 comps  39 comps  40 comps  41 comps  42 comps  43 comps
## X              83.45     83.92     84.37     84.82     85.23     85.64
## solubility     82.97     82.98     83.49     83.58     83.61     83.73
##             44 comps  45 comps  46 comps  47 comps  48 comps  49 comps
## X              86.03     86.42     86.78     87.13     87.46     87.77
## solubility     84.41     84.54     84.56     84.82     84.82     84.93
##             50 comps  51 comps  52 comps  53 comps  54 comps  55 comps
## X              88.08     88.39     88.68     88.97     89.25     89.52
## solubility     84.99     85.19     85.46     85.48     85.55     85.57
##             56 comps  57 comps  58 comps  59 comps  60 comps  61 comps
## X              89.77     90.02     90.26     90.51     90.75     90.97
## solubility     85.61     85.62     85.70     85.89     86.12     86.61
##             62 comps  63 comps  64 comps  65 comps  66 comps  67 comps
## X              91.19     91.41     91.62     91.83     92.03     92.23
## solubility     86.65     86.66     86.66     86.67     86.69     86.82
##             68 comps  69 comps  70 comps  71 comps  72 comps  73 comps
## X              92.42     92.60     92.77     92.95     93.12     93.28
## solubility     86.83     86.92     86.94     87.00     87.45     87.48
##             74 comps  75 comps  76 comps  77 comps  78 comps  79 comps
## X              93.44     93.60     93.76     93.91     94.06     94.20
## solubility     87.50     87.51     87.55     87.57     87.62     87.74
##             80 comps  81 comps  82 comps  83 comps  84 comps  85 comps
## X              94.34     94.47     94.61     94.74     94.86     94.99
## solubility     87.76     87.83     87.95     87.95     88.00     88.11
```

```
##              86 comps  87 comps  88 comps  89 comps  90 comps  91 comps
## X               95.11     95.22     95.34     95.45     95.56     95.66
## solubility      88.11     88.13     88.14     88.14     88.23     88.41
##              92 comps  93 comps  94 comps  95 comps  96 comps  97 comps
## X               95.77     95.87     95.97     96.07     96.16     96.26
## solubility      88.60     88.67     88.68     88.71     88.72     88.72
##              98 comps  99 comps  100 comps  101 comps  102 comps  103 comps
## X               96.35     96.44      96.53      96.61      96.70      96.78
## solubility      88.74     88.74      88.94      88.97      89.02      89.12
##             104 comps  105 comps  106 comps  107 comps  108 comps
## X               96.86      96.94      97.02      97.09      97.17
## solubility      89.30      89.33      89.33      89.34      89.39
##             109 comps  110 comps  111 comps  112 comps  113 comps
## X               97.24      97.31      97.38      97.45      97.51
## solubility      89.62      89.64      89.65      89.66      89.77
##             114 comps  115 comps  116 comps  117 comps  118 comps
## X               97.58      97.64      97.70      97.76      97.82
## solubility      89.81      89.81      89.87      89.88      89.88
##             119 comps  120 comps  121 comps  122 comps  123 comps
## X               97.88      97.94      98.00      98.05      98.11
## solubility      90.00      90.06      90.34      90.36      90.44
##             124 comps  125 comps  126 comps  127 comps  128 comps
## X               98.16      98.21      98.26      98.31      98.36
## solubility      90.49      90.67      90.69      90.70      90.70
##             129 comps  130 comps  131 comps  132 comps  133 comps
## X               98.41      98.45      98.50      98.54      98.59
## solubility      90.79      90.79      91.14      91.24      91.25
##             134 comps  135 comps  136 comps  137 comps  138 comps
## X               98.63      98.67      98.71      98.75      98.79
## solubility      91.25      91.25      91.34      91.39      91.41
##             139 comps  140 comps  141 comps  142 comps  143 comps
## X               98.82      98.86      98.89      98.93      98.96
## solubility      91.42      91.49      91.63      91.65      91.65
##             144 comps  145 comps  146 comps  147 comps  148 comps
## X               99.00      99.03      99.06      99.09      99.12
## solubility      91.91      91.91      91.96      91.97      92.06
##             149 comps  150 comps  151 comps  152 comps  153 comps
## X               99.15      99.18      99.20      99.23      99.26
## solubility      92.06      92.12      92.12      92.12      92.15
##             154 comps  155 comps  156 comps  157 comps  158 comps
## X               99.28      99.31      99.33      99.35      99.38
## solubility      92.16      92.26      92.26      92.35      92.37
##             159 comps  160 comps  161 comps  162 comps  163 comps
## X               99.40      99.42      99.44      99.46      99.48
## solubility      92.37      92.37      92.40      92.41      92.41
##             164 comps  165 comps  166 comps  167 comps  168 comps
## X               99.50      99.52      99.54      99.56      99.57
## solubility      92.47      92.47      92.53      92.54      92.55
##             169 comps  170 comps  171 comps  172 comps  173 comps
## X               99.59      99.61      99.62      99.64      99.65
## solubility      92.55      92.57      92.58      92.58      92.58
##             174 comps  175 comps  176 comps  177 comps  178 comps
## X               99.67      99.68       99.7      99.71      99.73
## solubility      92.64      92.64       92.7      92.72      92.72
```

```
##             179 comps  180 comps  181 comps  182 comps  183 comps
## X               99.74      99.75      99.76      99.77      99.79
## solubility      92.76      92.76      92.84      92.84      92.89
##             184 comps  185 comps  186 comps  187 comps  188 comps
## X                99.8      99.81      99.82      99.83      99.84
## solubility       92.9      92.92      92.98      92.98      92.99
##             189 comps  190 comps  191 comps  192 comps  193 comps
## X               99.85      99.86      99.86      99.87      99.88
## solubility      93.00      93.02      93.02      93.02      93.03
##             194 comps  195 comps  196 comps  197 comps  198 comps
## X               99.89      99.90      99.90      99.91      99.92
## solubility      93.07      93.11      93.24      93.24      93.25
##             199 comps  200 comps  201 comps  202 comps  203 comps
## X               99.92      99.93      99.94      99.94      99.95
## solubility      93.26      93.35      93.35      93.42      93.42
##             204 comps  205 comps  206 comps  207 comps  208 comps
## X               99.95      99.96      99.96      99.97      99.97
## solubility      93.42      93.48      93.57      93.57      93.65
##             209 comps  210 comps  211 comps  212 comps  213 comps
## X               99.97      99.98      99.98      99.98      99.99
## solubility      93.69      93.70      93.70      93.73      93.73
##             214 comps  215 comps  216 comps  217 comps  218 comps
## X               99.99      99.99      99.99      99.99      99.99
## solubility      93.76      93.78      93.82      93.88      93.90
##             219 comps  220 comps  221 comps  222 comps  223 comps
## X              100.00     100.00     100.00     100.00     100.00
## solubility      93.92      93.93      93.94      93.97      94.04
##             224 comps  225 comps  226 comps  227 comps  228 comps
## X              100.00     100.00     100.00     100.00     100.00
## solubility      94.04      94.16      94.16      94.16      94.46
```

```r
# Validation plot
validationplot(pcr_mod, val.type = "MSEP", legendpos = "topright")
```




```
# MSE (choose M = 150 based on the model result for smallest CV error)
pred_resp_pcr = predict(pcr_mod, newdata = test_data, ncomp = 150); pred_resp_pcr
```

```
## , , 150 comps
##
##       solubility
## 1    0.519327856
## 2    0.156261115
## 3   -0.609436215
## 4    0.626046771
## 5   -0.137702477
## 6    1.681472717
## 7    0.826069309
## 8    0.092759942
## 9    0.702830568
## 10  -0.572456453
## 11  -0.641298863
## 12  -1.098876355
## 13   0.322236910
## 14  -0.082801202
## 15  -0.597472730
## 16  -0.833897974
## 17  -0.334222983
## 18   0.185876063
## 19   0.795453856
## 20  -0.886646371
```

```
## 21   0.187922590
## 22   0.009630525
## 23  -0.537465201
## 24  -0.676182920
## 25  -1.242425014
## 26  -0.111183101
## 27  -0.382307525
## 28   1.333448096
## 29  -1.986510591
## 30  -1.102231434
## 31  -1.880493147
## 32  -0.814262512
## 33  -0.633252915
## 34   0.119909499
## 35   0.099762288
## 36  -1.044210221
## 37   0.654175940
## 38  -0.803951446
## 39   0.303323985
## 40  -0.863998715
## 41  -1.158939716
## 42  -2.381458977
## 43  -1.062265126
## 44  -0.236709177
## 45  -1.549432027
## 46  -0.892611764
## 47  -0.724423539
## 48  -0.978206883
## 49  -0.934124045
## 50  -0.587109653
## 51  -0.002073653
## 52  -1.276290525
## 53  -1.442396972
## 54  -1.381956668
## 55  -1.992248566
## 56  -1.379241318
## 57  -1.202231524
## 58  -1.588575809
## 59  -1.880072572
## 60  -2.119314215
## 61  -0.954385323
## 62  -1.240599863
## 63  -1.081938201
## 64  -1.354143745
## 65  -1.220261655
## 66  -1.424815857
## 67  -2.566885518
## 68  -1.977261490
## 69  -1.240262943
## 70  -3.079530930
## 71  -2.057095156
## 72  -2.592709259
## 73  -1.668329289
## 74  -2.949716792
```

```
## 75  -2.156553873
## 76  -2.811551925
## 77  -2.637058599
## 78  -1.960924462
## 79  -1.688299095
## 80  -2.178528311
## 81  -0.877534467
## 82  -1.662330671
## 83  -2.247501644
## 84  -1.969840392
## 85  -0.945820938
## 86  -1.477814865
## 87  -3.085952392
## 88  -2.165696637
## 89  -2.102008568
## 90  -2.189151807
## 91  -1.727234231
## 92  -2.168818037
## 93  -2.554402523
## 94  -1.617676792
## 95  -0.576426134
## 96  -2.575970146
## 97  -1.958111371
## 98  -1.619882227
## 99  -1.414477101
## 100 -2.185213656
## 101 -1.960802022
## 102 -2.378644113
## 103 -1.912819260
## 104 -2.310935218
## 105 -2.784696978
## 106 -1.749318966
## 107 -2.558642254
## 108 -2.359795575
## 109 -3.360098081
## 110 -3.196926247
## 111 -2.673045779
## 112 -2.879759513
## 113 -3.297517350
## 114 -3.068027871
## 115 -2.340937347
## 116 -2.476509172
## 117 -2.926077995
## 118 -2.326124036
## 119 -3.049458764
## 120 -2.912704104
## 121 -2.568763838
## 122 -1.314237824
## 123 -3.681038650
## 124 -3.100228191
## 125 -3.060316241
## 126 -2.612062067
## 127 -3.798263530
## 128 -2.551688796
```

```
## 129 -3.704979546
## 130 -2.462917840
## 131 -3.103068876
## 132 -3.663433988
## 133 -2.521208403
## 134 -3.647750026
## 135 -2.623885718
## 136 -3.549464396
## 137 -2.343279088
## 138 -2.999208532
## 139 -2.954195103
## 140 -2.417624229
## 141 -2.948949155
## 142 -2.233961959
## 143 -3.616915449
## 144 -3.170621873
## 145 -2.995807412
## 146 -3.555202042
## 147 -2.454693504
## 148 -3.369165383
## 149 -3.561244212
## 150 -3.954263742
## 151 -1.163613978
## 152 -3.191526655
## 153 -2.433105610
## 154 -3.785994145
## 155 -3.047543605
## 156 -3.118156274
## 157 -4.463344891
## 158 -4.656315945
## 159 -3.958835952
## 160 -4.335496707
## 161 -5.757682150
## 162 -4.117346722
## 163 -3.298401273
## 164 -4.515649539
## 165 -4.624514096
## 166 -3.550739376
## 167 -4.773412798
## 168 -4.300387076
## 169 -4.976649206
## 170 -4.682692236
## 171 -3.918994814
## 172 -3.989607733
## 173 -3.455927227
## 174 -4.953991318
## 175 -5.107020177
## 176 -4.264436266
## 177 -3.990226708
## 178 -4.707501717
## 179 -4.465738105
## 180 -3.362217234
## 181 -5.080516958
## 182 -3.821079709
```

```
## 183 -4.781281109
## 184 -4.586404174
## 185 -4.013182482
## 186 -3.718557053
## 187 -4.688066078
## 188 -4.467977629
## 189 -5.968023868
## 190 -6.249407302
## 191 -4.157210797
## 192 -3.187872616
## 193 -4.399903535
## 194 -5.026884854
## 195 -4.478264851
## 196 -4.731301748
## 197 -5.817998828
## 198 -4.461090721
## 199 -5.059901209
## 200 -5.022319824
## 201 -7.438987639
## 202 -6.535220515
## 203 -6.253879258
## 204 -6.900110367
## 205 -5.924814417
## 206 -5.816495268
## 207 -5.373337217
## 208 -5.863183475
## 209 -6.857361183
## 210 -6.647760466
## 211 -7.192691705
## 212 -6.886630158
## 213 -7.721164517
## 214 -7.953698743
## 215 -8.655178481
## 216 -7.658784889
## 217 -0.188776929
## 218  0.566365630
## 219 -0.029531516
## 220 -0.174385993
## 221 -0.995692203
## 222 -0.527587809
## 223 -1.009555456
## 224 -1.054477590
## 225 -1.929490440
## 226 -0.817833420
## 227 -0.704342264
## 228 -1.197705683
## 229 -0.712023968
## 230 -1.299915829
## 231 -1.309161174
## 232 -1.704605772
## 233 -0.868698430
## 234  0.087333757
## 235 -1.289354691
## 236 -0.938920399
```

```
## 237 -3.520764135
## 238 -1.937904919
## 239 -1.442396972
## 240 -1.114526691
## 241 -0.453018913
## 242 -2.325450445
## 243 -1.723395549
## 244 -2.655159111
## 245 -1.335941680
## 246 -0.385724841
## 247 -1.535296407
## 248 -1.003762022
## 249 -1.053613568
## 250 -1.967730429
## 251 -1.433371145
## 252 -1.792844155
## 253 -2.358315351
## 254 -3.301859565
## 255 -2.199816624
## 256 -1.124104318
## 257 -2.114777238
## 258 -1.667464115
## 259 -4.539943885
## 260 -1.521276987
## 261 -2.237497111
## 262 -2.183261675
## 263 -2.945268308
## 264 -3.226894634
## 265 -2.515407139
## 266 -1.115567835
## 267 -2.616664141
## 268 -2.422961687
## 269 -2.835733299
## 270 -3.117234036
## 271 -2.853227943
## 272 -3.408109074
## 273 -3.900665685
## 274 -3.703624920
## 275 -2.724015081
## 276 -3.716374294
## 277 -3.519546933
## 278 -2.907870296
## 279 -3.777997127
## 280 -3.183772244
## 281 -2.248459664
## 282 -3.897477009
## 283 -3.694146637
## 284 -3.968700696
## 285 -3.974100154
## 286 -4.432019798
## 287 -3.548164858
## 288 -3.073857639
## 289 -4.195252263
## 290 -5.034266753
```

```
## 291 -5.266144596
## 292 -4.618167463
## 293 -3.317858005
## 294 -4.649295760
## 295 -4.215445793
## 296 -3.943854280
## 297 -4.267087263
## 298 -4.163830091
## 299 -5.024970323
## 300 -5.744636720
## 301 -5.199535964
## 302 -5.530959062
## 303 -6.234901437
## 304 -5.442757921
## 305 -5.665776845
## 306 -6.956228819
## 307 -7.439823260
## 308 -8.176589032
## 309 -8.183838929
## 310 -8.431667841
## 311 -8.842576706
## 312 -7.203360374
## 313 -2.066208420
## 314 -2.301105711
## 315 -4.599633154
## 316 -4.711770103
```

```
mse_pcr = mean((pred_resp_pcr - test_data$solubility)^2); mse_pcr
```

```
## [1] 0.5483713
```

Thus, the mean square error for pcr model is 0.5483713, along with M = 150 which was selected based on its smallest CV error.

(e) Discussion

```
cbind(c("Model", "LS", "Ridge", "Lasso", "PCR"), c("MSE", mse_test, mse_ridge, mse_lasso, mse_pcr)) %>%
  knitr::kable()
```

Model	MSE
LS	0.555889819199859
Ridge	0.512657282223093
Lasso	0.499550611204814
PCR	0.548371340928091

Based on the R result, we observe that Lasso model produced the smallest mean square error(MSE) while Least Squares model produced the highest MSE. Thus, we could conclude that Lasso produced the best model fit among 4 different methods when building models using CV for prediction to solubility of compounds using chmical structures.

Moreover, Ridge, PCR and Lasso all produced model with smaller MSE compared to LS for the reason that these techniques involved regulations and dimension reductions to decrease the variability of the predictors in the model, hence, they produced model with smaller MSE compared to LS.