# PM 591 – Machine Learning for the Health Sciences.

Haoran Zhang

Due 4/4/2022

### Conceptual

The goal of this exercise is to illustrate the grouping properties of Ridge and Elastic-net compared to Lasso Compute the full path for of solutions the Lasso, Ridge and elastic net on the data generated below using `glmnet` to generate a full path of solutions:

```
# Simulates Features
set.seed(520)

n = 50
x1 = rnorm(n)
x2 = x1
x3 = rnorm(n)
X = cbind(x1, x2, x3)

# Simulates error/noise
e = rnorm(n, sd = 0.1)

#Simulates outcome y
y = 1 + 2*x1 + e
```

```
lambda_grid = 10^seq(2,-2,length=n)
ridge = glmnet(X, y, family='gaussian', alpha=0, standardize=TRUE, lambda=lambda_grid)
ridge_coef = coef(ridge)
round(ridge_coef[, 42:50], 2)
```

```
## 4 x 9 sparse Matrix of class "dgCMatrix"
##                s41   s42   s43   s44   s45   s46   s47   s48  s49
## (Intercept)  0.99  0.99  0.99  0.99  0.99  0.99  0.99  0.99 0.99
## x1           0.99  0.99  0.99  1.00  1.00  1.00  1.00  1.00 1.00
## x2           0.97  0.97  0.97  0.97  0.97  0.98  0.98  0.98 0.98
## x3          -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 -0.01 0.00
```

```
lasso = glmnet(X, y, family = "gaussian", alpha = 1)
lasso_coef = coef(lasso)
round(lasso_coef[, 42:50], 2)
```

```
## 4 x 9 sparse Matrix of class "dgCMatrix"
##              s41  s42  s43  s44  s45  s46  s47  s48  s49
## (Intercept) 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99
```

```
## x1          1.94 1.94 1.95 1.95 1.95 1.96 1.96 1.96 1.96
## x2          .    0.00 .    .    .    .    .    .    0.00
## x3          .    .    .    .    .    .    .    .    .
```

```
enet = glmnet(X, y, family = "gaussian", alpha = 0.5)
enet_coef = coef(enet)
round(enet_coef[, 42:50], 2)
```

```
## 4 x 9 sparse Matrix of class "dgCMatrix"
##               s41  s42  s43  s44  s45  s46  s47  s48  s49
## (Intercept) 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99 0.99
## x1          0.97 0.98 0.98 0.98 0.99 0.99 0.99 0.99 0.99
## x2          0.95 0.95 0.95 0.95 0.95 0.95 0.96 0.96 0.96
## x3          .    .    .    .    .    .    .    .    .
```

Comment on your results regarding grouping effects of the estimated coefficients.
The estimated coefficients of x1 and x2 are similar as the regression coefficients of a group of highly correlated variables tend to be equal, while elastic net regression captures the complexity range between the intercept-only model and the standard linear regression model with all features

**Analysis**

**Exercise 1** You will build a model to predict psa levels using PCA linear regression using the PSA prostate data

i. Load the mlr3 library and the prostate data

```
# read in data
prostate <- read.csv("prostate.csv")
```

ii. Specify the regression task and the base linear regression learner. Note: we will not split the data into training and testing because of the modest sample size. Explain whether this invalidates any prediction model we develop and why in practice we always want a test set.

```
# create PSA task
psa.tsk <- as_task_regr(prostate, target = "lpsa", id = "PSA Prediction")

# create basic linear regression learner
lm.lrn <- lrn("regr.lm")
```

PCA is an unsupervised method (i.e. does not use outcome/labels) thus won't be invalidated. However, it can invalidate base linear regression as it needs test data to evaluate how well the model performs.

iii. Create a new learner adding a PCA preprocessing step to the base learner. In `mlr3` parlance this is called a pipeline operator, `%>>%`. This becomes a new 'composite' learner that can be used just like any of the standard learners we used before. In particular if K-fold CV is used, both the PCA and the linear regression will be used for training on each set of K-1 folds and prediction on the K-th fold.

```
# create PCA step
pca <- po("pca")

# combines linear regression and PCA into a single learner
pca_lm.lrn <- pca %>>% lm.lrn
```

iv. Rather than fixing it as in the lecture, we will treat the number of principal components `pca.rank.` as a tuning parameter. Specify `pca.rank.` as a an integer tuning parameter ranging from 1 to the number of features in the PSA data

```
ps <- ps(pca.rank. = p_int(lower = 1, length(psa.tsk$feature_names)))
```

v. Create a control object for hyperparameter tuning with grid search.

```
ctrl <- tnr("grid_search", resolution = 8)
# resolution is the number of points in the grid of values for the tuning parameter. Since there are 8 p
```

vi. Perform the tuning

```
set.seed(202)

# resampling method
cv5 <- rsmp("cv", folds = 5)

# create the autotuner
pca_lm.lrn = AutoTuner$new(
  learner = pca_lm.lrn,
  resampling = cv5,
  measure = msr("regr.rsq"),
  search_space = ps,
  terminator = trm("evals", n_evals = 10), # stop after 10 iterations
  tuner = ctrl
)

# complete the tuning
lgr::get_logger("mlr3")$set_threshold("warn")
pca_lm.lrn$train(psa.tsk)
```

```
## INFO  [14:17:39.081] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
## INFO  [14:17:39.124] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:39.832] [bbotk] Result of batch 1:
## INFO  [14:17:39.834] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:39.834] [bbotk]          8 0.5226138        0      0             0.54
## INFO  [14:17:39.834] [bbotk]                                   uhash
## INFO  [14:17:39.834] [bbotk]  9dbf39e2-88dc-44d4-8aa2-f3b0905000f1
## INFO  [14:17:39.836] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:40.409] [bbotk] Result of batch 2:
## INFO  [14:17:40.411] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:40.411] [bbotk]          4 0.3597818        0      0             0.47
## INFO  [14:17:40.411] [bbotk]                                   uhash
## INFO  [14:17:40.411] [bbotk]  b2d53a46-a9b8-4af2-aaa0-15d9247a7923
```

```
## INFO  [14:17:40.412] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:40.992] [bbotk] Result of batch 3:
## INFO  [14:17:40.994] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:40.994] [bbotk]          2 0.0765287        0      0             0.48
## INFO  [14:17:40.994] [bbotk]                                  uhash
## INFO  [14:17:40.994] [bbotk]  6004a60f-8ff8-477a-8934-8e8ac982e5a4
## INFO  [14:17:40.995] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:41.561] [bbotk] Result of batch 4:
## INFO  [14:17:41.563] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:41.563] [bbotk]          1 0.08762075        0      0             0.45
## INFO  [14:17:41.563] [bbotk]                                  uhash
## INFO  [14:17:41.563] [bbotk]  98c6adc5-6319-42dc-adcc-c54a44ad0879
## INFO  [14:17:41.564] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:42.277] [bbotk] Result of batch 5:
## INFO  [14:17:42.279] [bbotk]  pca.rank. regr.rsq warnings errors runtime_learners
## INFO  [14:17:42.279] [bbotk]          3  0.19538        0      0             0.62
## INFO  [14:17:42.279] [bbotk]                                  uhash
## INFO  [14:17:42.279] [bbotk]  44c1b95a-d1e6-4e3b-bd4a-76b50f19eee3
## INFO  [14:17:42.280] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:42.839] [bbotk] Result of batch 6:
## INFO  [14:17:42.840] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:42.840] [bbotk]          6 0.4559056        0      0             0.48
## INFO  [14:17:42.840] [bbotk]                                  uhash
## INFO  [14:17:42.840] [bbotk]  8705fe3e-8dc6-4507-a1c0-a99a8366f0ed
## INFO  [14:17:42.842] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:43.398] [bbotk] Result of batch 7:
## INFO  [14:17:43.400] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:43.400] [bbotk]          7 0.4741076        0      0             0.46
## INFO  [14:17:43.400] [bbotk]                                  uhash
## INFO  [14:17:43.400] [bbotk]  0565648c-13db-4e4b-92d2-1f2b104726e2
## INFO  [14:17:43.402] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:43.966] [bbotk] Result of batch 8:
## INFO  [14:17:43.968] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:43.968] [bbotk]          5 0.4407682        0      0             0.42
## INFO  [14:17:43.968] [bbotk]                                  uhash
## INFO  [14:17:43.968] [bbotk]  38b8263e-ac5e-4b5a-8806-2c925ea03237
## INFO  [14:17:43.973] [bbotk] Finished optimizing after 8 evaluation(s)
## INFO  [14:17:43.974] [bbotk] Result:
## INFO  [14:17:43.975] [bbotk]  pca.rank. learner_param_vals  x_domain  regr.rsq
## INFO  [14:17:43.975] [bbotk]          8          <list[1]> <list[1]> 0.5226138
```

vii. How many principal components are selected? Does preprocessing by PCA help in this case?
8 principal components are selected. Preprocessing by PCA doesn't help as we keep all principal components and don't reduce the number of dimensions.

viii. Use now benchmark to automate the comparison between PCA regression and standard linear regression on the prostate data

```
design = design = benchmark_grid(
 tasks = psa.tsk,
 learners = list(lm.lrn, pca_lm.lrn),
 resampling = rsmp("cv", folds = 5)
)
```

```
psa_benchmark = benchmark(design)
```

```
## INFO  [14:17:44.142] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
## INFO  [14:17:44.144] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:44.710] [bbotk] Result of batch 1:
## INFO  [14:17:44.711] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:44.711] [bbotk]          6 0.5191904        0      0             0.45
## INFO  [14:17:44.711] [bbotk]                                  uhash
## INFO  [14:17:44.711] [bbotk]  3c974731-1e93-4ea1-831a-1e87079d86bc
## INFO  [14:17:44.712] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:45.269] [bbotk] Result of batch 2:
## INFO  [14:17:45.270] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:45.270] [bbotk]          4 0.3510956        0      0             0.48
## INFO  [14:17:45.270] [bbotk]                                  uhash
## INFO  [14:17:45.270] [bbotk]  67c2acab-f213-4aea-8371-17a3ad451700
## INFO  [14:17:45.271] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:45.845] [bbotk] Result of batch 3:
## INFO  [14:17:45.847] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:45.847] [bbotk]          3 0.1722028        0      0             0.44
## INFO  [14:17:45.847] [bbotk]                                  uhash
## INFO  [14:17:45.847] [bbotk]  da4079c3-bebe-4f1b-8677-461bf87bea27
## INFO  [14:17:45.848] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:46.476] [bbotk] Result of batch 4:
## INFO  [14:17:46.478] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:46.478] [bbotk]          8 0.5464672        0      0             0.53
## INFO  [14:17:46.478] [bbotk]                                  uhash
## INFO  [14:17:46.478] [bbotk]  1034d26d-78e7-4083-861a-4ce52581208d
## INFO  [14:17:46.479] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:47.045] [bbotk] Result of batch 5:
## INFO  [14:17:47.047] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:47.047] [bbotk]          2 0.06328942        0      0             0.46
## INFO  [14:17:47.047] [bbotk]                                  uhash
## INFO  [14:17:47.047] [bbotk]  b7f3f41b-3abb-4763-9d5c-3b19dd6396f6
## INFO  [14:17:47.048] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:47.589] [bbotk] Result of batch 6:
## INFO  [14:17:47.590] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:47.590] [bbotk]          1 0.08388087        0      0             0.48
## INFO  [14:17:47.590] [bbotk]                                  uhash
## INFO  [14:17:47.590] [bbotk]  44d45de8-7e55-4fdf-9034-514fc88ba7ae
## INFO  [14:17:47.597] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:48.160] [bbotk] Result of batch 7:
## INFO  [14:17:48.162] [bbotk]  pca.rank. regr.rsq warnings errors runtime_learners
## INFO  [14:17:48.162] [bbotk]          7 0.535237        0      0            0.46
## INFO  [14:17:48.162] [bbotk]                                  uhash
## INFO  [14:17:48.162] [bbotk]  e4198dc6-d678-4dc8-9dad-116a2da7efa0
## INFO  [14:17:48.163] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:48.723] [bbotk] Result of batch 8:
## INFO  [14:17:48.725] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:48.725] [bbotk]          5 0.5237577        0      0             0.48
## INFO  [14:17:48.725] [bbotk]                                  uhash
## INFO  [14:17:48.725] [bbotk]  3c6daaa2-098e-4216-b22f-b2f224700fef
## INFO  [14:17:48.731] [bbotk] Finished optimizing after 8 evaluation(s)
## INFO  [14:17:48.731] [bbotk] Result:
```

```
## INFO  [14:17:48.732] [bbotk]  pca.rank. learner_param_vals  x_domain  regr.rsq
## INFO  [14:17:48.732] [bbotk]          8          <list[1]> <list[1]> 0.5464672
## INFO  [14:17:48.901] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
## INFO  [14:17:48.903] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:49.459] [bbotk] Result of batch 1:
## INFO  [14:17:49.460] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:49.460] [bbotk]          4 0.4048833        0      0             0.49
## INFO  [14:17:49.460] [bbotk]                                    uhash
## INFO  [14:17:49.460] [bbotk]  ac326b3f-f428-46bc-a1b7-4042aa8866c1
## INFO  [14:17:49.461] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:50.001] [bbotk] Result of batch 2:
## INFO  [14:17:50.003] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:50.003] [bbotk]          2 0.1017573        0      0             0.41
## INFO  [14:17:50.003] [bbotk]                                    uhash
## INFO  [14:17:50.003] [bbotk]  64efd196-f820-42bb-8485-0df725466f25
## INFO  [14:17:50.004] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:50.561] [bbotk] Result of batch 3:
## INFO  [14:17:50.563] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:50.563] [bbotk]          6 0.4898604        0      0             0.49
## INFO  [14:17:50.563] [bbotk]                                    uhash
## INFO  [14:17:50.563] [bbotk]  e8dee14b-285d-4294-9120-7379095e0b12
## INFO  [14:17:50.564] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:51.112] [bbotk] Result of batch 4:
## INFO  [14:17:51.114] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:51.114] [bbotk]          7 0.5451187        0      0             0.47
## INFO  [14:17:51.114] [bbotk]                                    uhash
## INFO  [14:17:51.114] [bbotk]  58407d76-310e-41ba-8282-aec668b6e65b
## INFO  [14:17:51.115] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:51.692] [bbotk] Result of batch 5:
## INFO  [14:17:51.694] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:51.694] [bbotk]          1 0.1383498        0      0             0.48
## INFO  [14:17:51.694] [bbotk]                                    uhash
## INFO  [14:17:51.694] [bbotk]  b8e67127-ef0d-4900-aaa8-450a890841f9
## INFO  [14:17:51.696] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:52.278] [bbotk] Result of batch 6:
## INFO  [14:17:52.280] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:52.280] [bbotk]          5 0.4590287        0      0             0.52
## INFO  [14:17:52.280] [bbotk]                                    uhash
## INFO  [14:17:52.280] [bbotk]  93a2993d-1b51-49c4-9266-39024dbf98cf
## INFO  [14:17:52.281] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:52.850] [bbotk] Result of batch 7:
## INFO  [14:17:52.852] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:52.852] [bbotk]          8 0.5283399        0      0             0.49
## INFO  [14:17:52.852] [bbotk]                                    uhash
## INFO  [14:17:52.852] [bbotk]  c87f4835-5779-4c77-a357-1c3181542922
## INFO  [14:17:52.853] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:53.427] [bbotk] Result of batch 8:
## INFO  [14:17:53.429] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:53.429] [bbotk]          3 0.2261845        0      0             0.45
## INFO  [14:17:53.429] [bbotk]                                    uhash
## INFO  [14:17:53.429] [bbotk]  956515c9-22f8-437b-b77c-eb6e4f2d3dea
## INFO  [14:17:53.434] [bbotk] Finished optimizing after 8 evaluation(s)
## INFO  [14:17:53.434] [bbotk] Result:
## INFO  [14:17:53.435] [bbotk]  pca.rank. learner_param_vals  x_domain  regr.rsq
```

```
## INFO  [14:17:53.435] [bbotk]          7          <list[1]> <list[1]> 0.5451187
## INFO  [14:17:53.625] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
## INFO  [14:17:53.628] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:54.202] [bbotk] Result of batch 1:
## INFO  [14:17:54.203] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:54.203] [bbotk]          8 0.5293737        0      0             0.45
## INFO  [14:17:54.203] [bbotk]                                   uhash
## INFO  [14:17:54.203] [bbotk]  34e730d8-9aaf-4257-a65c-66c278dc0bae
## INFO  [14:17:54.205] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:54.778] [bbotk] Result of batch 2:
## INFO  [14:17:54.780] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:54.780] [bbotk]          5 0.4155771        0      0             0.47
## INFO  [14:17:54.780] [bbotk]                                   uhash
## INFO  [14:17:54.780] [bbotk]  cd950d47-3718-4dd0-8bef-6b0b55636e6b
## INFO  [14:17:54.781] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:55.341] [bbotk] Result of batch 3:
## INFO  [14:17:55.343] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:55.343] [bbotk]          1 0.08248712        0      0              0.5
## INFO  [14:17:55.343] [bbotk]                                   uhash
## INFO  [14:17:55.343] [bbotk]  11bac912-1c7e-4ead-93f5-76b772768861
## INFO  [14:17:55.344] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:55.948] [bbotk] Result of batch 4:
## INFO  [14:17:55.950] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:55.950] [bbotk]          4 0.3479046        0      0              0.5
## INFO  [14:17:55.950] [bbotk]                                   uhash
## INFO  [14:17:55.950] [bbotk]  4105254e-5857-4f69-a100-71ad5aec8436
## INFO  [14:17:55.951] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:56.542] [bbotk] Result of batch 5:
## INFO  [14:17:56.544] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:56.544] [bbotk]          2 0.02319306        0      0             0.49
## INFO  [14:17:56.544] [bbotk]                                   uhash
## INFO  [14:17:56.544] [bbotk]  668d0726-bf59-4f5d-8c6e-3c90494e4555
## INFO  [14:17:56.545] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:57.127] [bbotk] Result of batch 6:
## INFO  [14:17:57.128] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:57.128] [bbotk]          6 0.4191296        0      0             0.48
## INFO  [14:17:57.128] [bbotk]                                   uhash
## INFO  [14:17:57.128] [bbotk]  0f00de67-8a6a-4a7d-8adc-6542c0c7c65c
## INFO  [14:17:57.129] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:57.884] [bbotk] Result of batch 7:
## INFO  [14:17:57.885] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:57.885] [bbotk]          3 0.06394472        0      0             0.51
## INFO  [14:17:57.885] [bbotk]                                   uhash
## INFO  [14:17:57.885] [bbotk]  01219aae-1069-44b6-a726-ca6712b64da0
## INFO  [14:17:57.886] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:58.466] [bbotk] Result of batch 8:
## INFO  [14:17:58.467] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:58.467] [bbotk]          7 0.4627449        0      0             0.49
## INFO  [14:17:58.467] [bbotk]                                   uhash
## INFO  [14:17:58.467] [bbotk]  54bf9191-d539-4dff-830f-3aa95d6aa625
## INFO  [14:17:58.473] [bbotk] Finished optimizing after 8 evaluation(s)
## INFO  [14:17:58.474] [bbotk] Result:
## INFO  [14:17:58.475] [bbotk]  pca.rank. learner_param_vals  x_domain  regr.rsq
## INFO  [14:17:58.475] [bbotk]          8          <list[1]> <list[1]> 0.5293737
```

```
## INFO  [14:17:58.664] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
## INFO  [14:17:58.667] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:59.211] [bbotk] Result of batch 1:
## INFO  [14:17:59.212] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:17:59.212] [bbotk]          5 0.3305024        0      0             0.45
## INFO  [14:17:59.212] [bbotk]                                   uhash
## INFO  [14:17:59.212] [bbotk]  289ca74b-6c2c-4d09-a47c-9dbda2220dcc
## INFO  [14:17:59.214] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:17:59.766] [bbotk] Result of batch 2:
## INFO  [14:17:59.768] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:17:59.768] [bbotk]          4 0.09648935        0      0             0.41
## INFO  [14:17:59.768] [bbotk]                                   uhash
## INFO  [14:17:59.768] [bbotk]  c8b932a0-a878-4155-9d5a-2aee1673a864
## INFO  [14:17:59.769] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:00.338] [bbotk] Result of batch 3:
## INFO  [14:18:00.340] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:00.340] [bbotk]          8 0.3480951        0      0             0.49
## INFO  [14:18:00.340] [bbotk]                                   uhash
## INFO  [14:18:00.340] [bbotk]  dc3a5924-c128-43a9-a29d-21436e2bbd01
## INFO  [14:18:00.341] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:00.870] [bbotk] Result of batch 4:
## INFO  [14:18:00.871] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:18:00.871] [bbotk]          2 -0.2136195        0      0             0.45
## INFO  [14:18:00.871] [bbotk]                                   uhash
## INFO  [14:18:00.871] [bbotk]  b49a8676-3f6f-46e7-909e-f387e3e05d9e
## INFO  [14:18:00.873] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:01.439] [bbotk] Result of batch 5:
## INFO  [14:18:01.441] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:18:01.441] [bbotk]          3 -0.2032147        0      0              0.5
## INFO  [14:18:01.441] [bbotk]                                   uhash
## INFO  [14:18:01.441] [bbotk]  959cbd4f-c15b-4959-8a31-8af7c87a9fae
## INFO  [14:18:01.442] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:01.999] [bbotk] Result of batch 6:
## INFO  [14:18:02.000] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:02.000] [bbotk]          7 0.2887412        0      0             0.43
## INFO  [14:18:02.000] [bbotk]                                   uhash
## INFO  [14:18:02.000] [bbotk]  bb3697cc-82ea-4644-8b18-25f9c912e162
## INFO  [14:18:02.002] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:02.558] [bbotk] Result of batch 7:
## INFO  [14:18:02.560] [bbotk]  pca.rank.   regr.rsq warnings errors runtime_learners
## INFO  [14:18:02.560] [bbotk]          1 -0.2321828        0      0             0.48
## INFO  [14:18:02.560] [bbotk]                                   uhash
## INFO  [14:18:02.560] [bbotk]  aa23cab4-aabd-4d9e-acc1-6d4c0b54e1fa
## INFO  [14:18:02.561] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:03.118] [bbotk] Result of batch 8:
## INFO  [14:18:03.119] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:03.119] [bbotk]          6 0.3372783        0      0             0.46
## INFO  [14:18:03.119] [bbotk]                                   uhash
## INFO  [14:18:03.119] [bbotk]  0e325ccd-2ed5-4a6d-9f41-8fe844dce37f
## INFO  [14:18:03.125] [bbotk] Finished optimizing after 8 evaluation(s)
## INFO  [14:18:03.126] [bbotk] Result:
## INFO  [14:18:03.127] [bbotk]  pca.rank. learner_param_vals  x_domain  regr.rsq
## INFO  [14:18:03.127] [bbotk]          8          <list[1]> <list[1]> 0.3480951
## INFO  [14:18:03.333] [bbotk] Starting to optimize 1 parameter(s) with '<TunerGridSearch>' and '<Term
```

```
## INFO  [14:18:03.335] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:03.897] [bbotk] Result of batch 1:
## INFO  [14:18:03.899] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:03.899] [bbotk]          4 0.5299629        0      0             0.48
## INFO  [14:18:03.899] [bbotk]                                   uhash
## INFO  [14:18:03.899] [bbotk]  e2b73fb7-4ceb-42eb-bf80-7b6a72267971
## INFO  [14:18:03.900] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:04.453] [bbotk] Result of batch 2:
## INFO  [14:18:04.455] [bbotk]  pca.rank. regr.rsq warnings errors runtime_learners
## INFO  [14:18:04.455] [bbotk]          8 0.606006        0      0             0.49
## INFO  [14:18:04.455] [bbotk]                                   uhash
## INFO  [14:18:04.455] [bbotk]  3ab6ec4f-63c8-49df-a596-27c2602f8b9d
## INFO  [14:18:04.457] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:05.012] [bbotk] Result of batch 3:
## INFO  [14:18:05.013] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:05.013] [bbotk]          5 0.5541408        0      0             0.47
## INFO  [14:18:05.013] [bbotk]                                   uhash
## INFO  [14:18:05.013] [bbotk]  89bc1d8c-8cca-4f79-ad83-cf88774a370d
## INFO  [14:18:05.015] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:05.582] [bbotk] Result of batch 4:
## INFO  [14:18:05.584] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:05.584] [bbotk]          6 0.5490199        0      0             0.46
## INFO  [14:18:05.584] [bbotk]                                   uhash
## INFO  [14:18:05.584] [bbotk]  16fd4076-6e28-4f8c-a859-b345208ea732
## INFO  [14:18:05.585] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:06.152] [bbotk] Result of batch 5:
## INFO  [14:18:06.153] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:06.153] [bbotk]          7 0.5467932        0      0             0.47
## INFO  [14:18:06.153] [bbotk]                                   uhash
## INFO  [14:18:06.153] [bbotk]  74068764-27e8-4f38-94f3-7de6511a8a14
## INFO  [14:18:06.154] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:06.732] [bbotk] Result of batch 6:
## INFO  [14:18:06.733] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:06.733] [bbotk]          3 0.1746811        0      0             0.46
## INFO  [14:18:06.733] [bbotk]                                   uhash
## INFO  [14:18:06.733] [bbotk]  64d89069-3a2e-442a-a3c2-bbbb3e5eb4e1
## INFO  [14:18:06.735] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:07.299] [bbotk] Result of batch 7:
## INFO  [14:18:07.301] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:07.301] [bbotk]          2 0.1167378        0      0             0.45
## INFO  [14:18:07.301] [bbotk]                                   uhash
## INFO  [14:18:07.301] [bbotk]  cc28e813-a4cc-4ba6-b705-50376f5fca28
## INFO  [14:18:07.302] [bbotk] Evaluating 1 configuration(s)
## INFO  [14:18:07.894] [bbotk] Result of batch 8:
## INFO  [14:18:07.896] [bbotk]  pca.rank.  regr.rsq warnings errors runtime_learners
## INFO  [14:18:07.896] [bbotk]          1 0.1287061        0      0             0.47
## INFO  [14:18:07.896] [bbotk]                                   uhash
## INFO  [14:18:07.896] [bbotk]  6a110888-aeb3-43af-8867-746aa6371077
## INFO  [14:18:07.901] [bbotk] Finished optimizing after 8 evaluation(s)
## INFO  [14:18:07.902] [bbotk] Result:
## INFO  [14:18:07.903] [bbotk]  pca.rank. learner_param_vals  x_domain regr.rsq
## INFO  [14:18:07.903] [bbotk]          8          <list[1]> <list[1]> 0.606006
```

```
psa_benchmark$aggregate(msr('regr.rsq'))
```

```
##    nr      resample_result        task_id        learner_id resampling_id iters
## 1:  1 <ResampleResult[22]> PSA Prediction           regr.lm            cv     5
## 2:  2 <ResampleResult[22]> PSA Prediction pca.regr.lm.tuned            cv     5
##     regr.rsq
## 1: 0.5065547
## 2: 0.4891356
```

Using PCA regression doesn't improve the model comparing the R square.

**Exercise 2** You will build a classifier to predict cancer specific death among breast cancer patients within 5-year of diagnosis based on a subset of 1,000 gene expression features from the Metabric data using ridge, lasso and elastic net logistic regression. (The metabric data contains close to 30,000 gene expression features, here we use a subset to keep computation times reasonable for this in class Lab. In the homework version you will use the full feature set)

i. Load the Metabric data

```
load('metabric.Rdata')
```

ii. Check the dimension of the metabric dataframe using `dim` check the number of deaths using `table` on the binary outcome variable

```
# check dimensions
cat("Dataset Dimensions: \n"); dim(metabric)
```

```
## Dataset Dimensions:
```

```
## [1]  803 1001
```

```
# make sure to factor outcome variable
metabric$y <- factor(metabric$y, labels=c("survive", "die"))

# check number of deaths
cat("Number of deaths: \n"); table(metabric$y)
```

```
## Number of deaths:
```

```
##
## survive     die
##     657     146
```

iii. Create an appropriate mlr3 task

```
metabric.tsk <- as_task_classif(metabric, target = "y", id = "One-year Breast Cancer Mortality")
```

iv. Split the data into training (70%) and test (30%)

```
# specify resampling to have 70/30 training/testing split
holdout.desc <- rsmp("holdout", ratio = 0.7)

# instantiate split
holdout.desc$instantiate(metabric.tsk)

# extract training and testing sets
train <- holdout.desc$train_set(1)
test  <- holdout.desc$test_set(1)
```

v. Create lasso, ridge, and Elastic net learners using "classif.cv_glmnet" (Recall that by specifying `cv.glmnet` as the learner, k-fold (10-fold by default) will be automatically used to tune the lambda penalty parameter. This takes advantage of the fast implementation of cross-validation within the `glmnet` package rather than cross-validating using `mlr3` tools).

```
# LASSO
lasso.lrn <- lrn("classif.cv_glmnet",
                 alpha = 1,
                 type.measure = "auc")

lasso.lrn$predict_type <- "prob"

# Ridge
ridge.lrn <- lrn("classif.cv_glmnet",
                 alpha = 0,
                 type.measure = "auc")

ridge.lrn$predict_type <- "prob"

# Enet
enet.lrn <- lrn("classif.cv_glmnet",
                 alpha = 0.5,
                 type.measure = "auc")

enet.lrn$predict_type <- "prob"
```

vi. Train the models on the training data using CV with an appropriate performance measure (hint: you can check the available measures for your task using `listMeasures`). Extract the cross-validated measure of performance. Why is the CV measure of performance the relevant metric to compare models?

```
# LASSO
lasso.lrn$train(metabric.tsk, row_ids=train)
lasso.lrn$model
```

```
##
## Call:  (if (cv) glmnet::cv.glmnet else glmnet::glmnet)(x = data, y = target,      type.measure = "au
##
## Measure: AUC
##
##      Lambda Index Measure      SE Nonzero
## min 0.03295    26  0.6980 0.03191      39
## 1se 0.07974     7  0.6678 0.03993       5
```

```r
cat("LASSO cross-validated AUC:");max(lasso.lrn$model$cvm)
```

```
## LASSO cross-validated AUC:
```

```
## [1] 0.6979535
```

```r
# Ridge
ridge.lrn$train(metabric.tsk, row_ids=train)
ridge.lrn$model
```

```
##
## Call:  (if (cv) glmnet::cv.glmnet else glmnet::glmnet)(x = data, y = target,      type.measure = "au
##
## Measure: AUC
##
##     Lambda Index Measure      SE Nonzero
## min   4.89    67  0.7663 0.01551    1000
## 1se  96.04     3  0.7610 0.01711    1000
```

```r
cat("Ridge cross-validated AUC:");max(ridge.lrn$model$cvm)
```

```
## Ridge cross-validated AUC:
```

```
## [1] 0.7662756
```

```r
# LASSO
enet.lrn$train(metabric.tsk, row_ids=train)
enet.lrn$model
```

```
##
## Call:  (if (cv) glmnet::cv.glmnet else glmnet::glmnet)(x = data, y = target,      type.measure = "au
##
## Measure: AUC
##
##      Lambda Index Measure      SE Nonzero
## min 0.05731    29  0.7014 0.04161      62
## 1se 0.15222     8  0.6601 0.04326       8
```

```r
cat("Enet cross-validated AUC:");max(lasso.lrn$model$cvm)
```

```
## Enet cross-validated AUC:
```

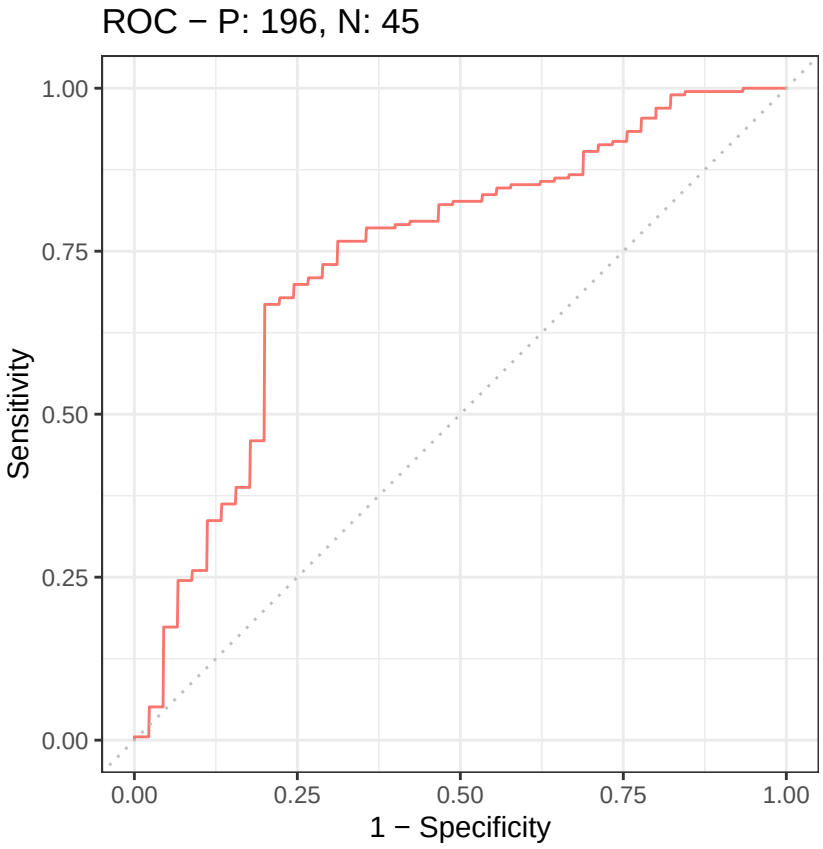
```
## [1] 0.6979535
```

vii. Which method performs best? What does this say about the likely nature of the true relationship between the expression features and the outcome?
Ridge performs best with greatest R-square and AUC.

viii. Report an 'honest' estimate of prediction performance, plot the ROC curve.

```r
#Fill in the ...

ridge.prd <- ridge.lrn$predict(metabric.tsk, row_ids = test)
autoplot(ridge.prd, type= 'roc')
```

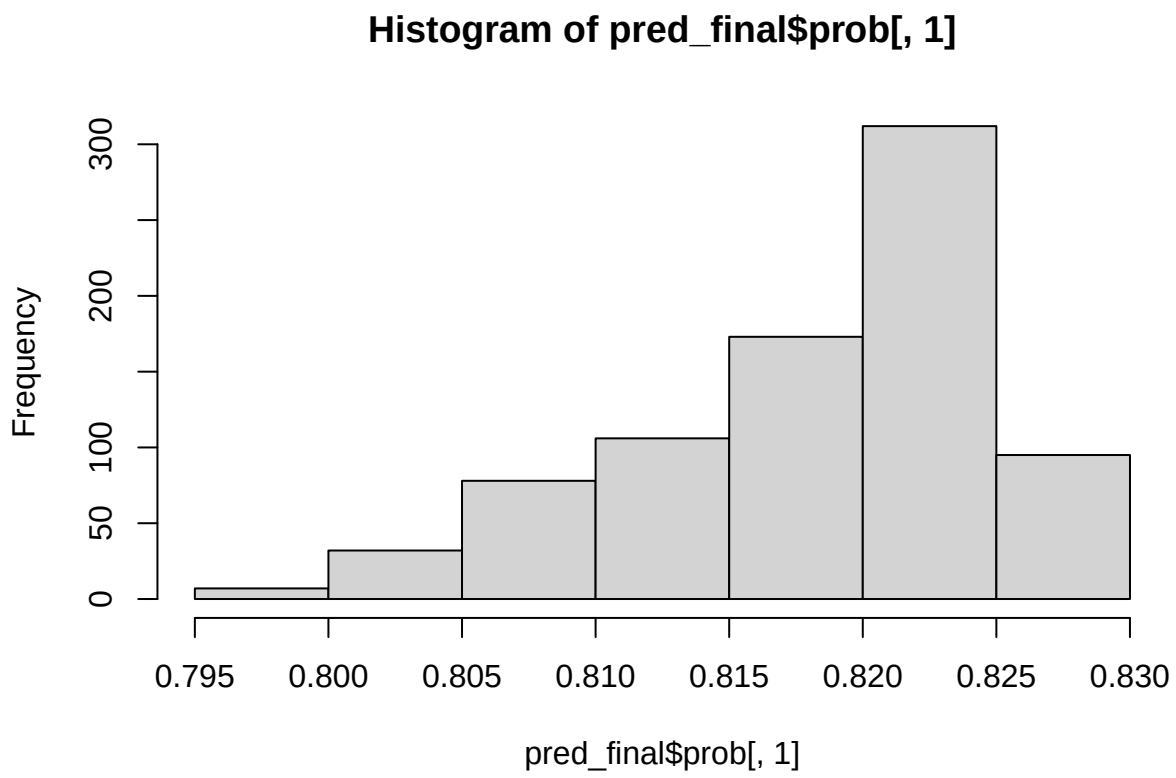



ix. Re-train the best performing method on all the data (training and test). This is the final model you would use to predict death in new women just diagnosed and treated for breast cancer. Why is this ok and why is this better than simply using the model trained on just the training data?

```r
#Fill in the ...
metabric_final = ridge.lrn$train(metabric.tsk)
pred_final <- ridge.lrn$predict(metabric.tsk)
head(pred_final$prob)
```

```
##        survive       die
## [1,] 0.8254399 0.1745601
## [2,] 0.8082788 0.1917212
## [3,] 0.8203934 0.1796066
## [4,] 0.8051121 0.1948879
## [5,] 0.8274138 0.1725862
## [6,] 0.8198981 0.1801019
```

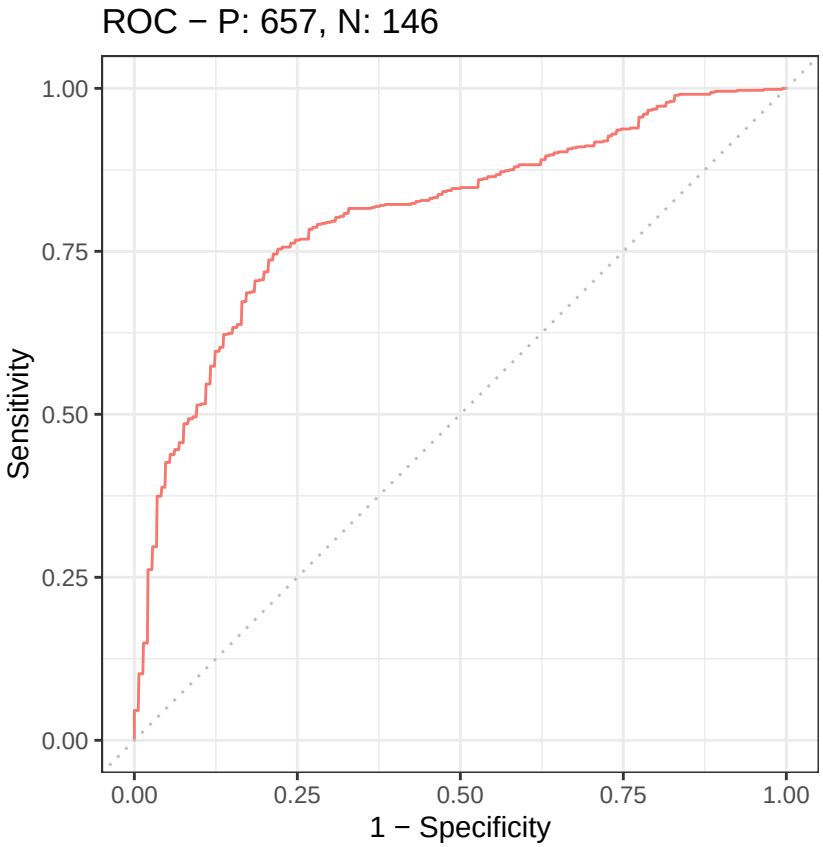
```r
hist(pred_final$prob[,1])
```




```r
p0.5 = median(pred_final$prob[,1])  # median predicted probability
pred_final$set_threshold(p0.5) # change prediction cutoff
pred_final$confusion
```

```
##          truth
## response  survive die
##   survive     383  18
##   die         274 128
```

```r
autoplot(pred_final, type= 'roc')
```




We've already used cross-validation so it's ok to use all the data. A larger training set on all the data can improve the performance of prediction comparing to simply using the model trained on just the training data.

x. The dataset `new_expression_profiles` contains the gene expression levels for 15 women newly diagnosed with breast cancer. Estimate their one-year survival probabilities using the selected model.

```
# read in data
new_expression_profiles <- read.csv("new_expression_profiles.csv", header=T)

# predict in new data
predict_new <- metabric_final$predict_newdata(new_expression_profiles)
predict_new$prob
```

```
##         survive       die
##  [1,] 0.8073033 0.1926967
##  [2,] 0.8092850 0.1907150
##  [3,] 0.8215832 0.1784168
##  [4,] 0.8164474 0.1835526
##  [5,] 0.8138908 0.1861092
##  [6,] 0.8130966 0.1869034
##  [7,] 0.8238124 0.1761876
##  [8,] 0.8190581 0.1809419
##  [9,] 0.8229510 0.1770490
## [10,] 0.8182328 0.1817672
## [11,] 0.8201474 0.1798526
```

```
## [12,] 0.8250576 0.1749424
## [13,] 0.8242306 0.1757694
## [14,] 0.8213745 0.1786255
## [15,] 0.8157173 0.1842827
```

```
predict_new$response
```

```
##  [1] survive survive survive survive survive survive survive survive survive
## [10] survive survive survive survive survive survive
## Levels: survive die
```

xi. Redo the model comparison between lasso, ridge, elastic net using mlr3's `benchmark` function rather than manually

```
design2 = design2 = benchmark_grid(
    tasks = metabric.tsk,
    learners = list(ridge.lrn, lasso.lrn, enet.lrn),
    resampling = rsmp("cv", folds = 5)
)
metabric_benchmark2 = benchmark(design2)
metabric_benchmark2$aggregate(msr('classif.auc'))
```

```
##    nr      resample_result                          task_id        learner_id
## 1:  1 <ResampleResult[22]> One-year Breast Cancer Mortality classif.cv_glmnet
## 2:  2 <ResampleResult[22]> One-year Breast Cancer Mortality classif.cv_glmnet
## 3:  3 <ResampleResult[22]> One-year Breast Cancer Mortality classif.cv_glmnet
##    resampling_id iters classif.auc
## 1:            cv     5   0.7583559
## 2:            cv     5   0.7046439
## 3:            cv     5   0.7137941
```

Ridge regression has the largest AUC thus is the best model.