Executive summary

Models with interactions and without was developed to predict the development by relying on lasso logistic regression. Interactions significantly reduce cross-validation mean deviance.

Introduction and data

The data present results of laboratory tests and subsequent monitoring whether the patient develops heart disease or not.

Attribute Information:

-- 1. age
-- 2. sex
-- 3. chest pain type (4 values)
-- 4. resting blood pressure
-- 5. serum cholestoral in mg/dl
-- 6. fasting blood sugar > 120 mg/dl
-- 7. resting electrocardiographic results (values 0,1,2)
-- 8. maximum heart rate achieved
-- 9. exercise induced angina
-- 10. oldpeak = ST depression induced by exercise relative to rest
-- 11. the slope of the peak exercise ST segment
-- 12. number of major vessels (0-3) colored by flourosopy
-- 13. thal: 3 = normal; 6 = fixed defect; 7 = reversable defect

Attributes types

Real: 1,4,5,8,10,12
Ordered:11,
Binary: 2,6,9
Nominal:7,3,13

Variable to be predicted

Absence (1) or presence (2) of heart disease

Dataset is small, of 270 instances. This present certain challenge as set validation technique can't be used.

Analysis

Simple Linear Regression

Logistic regression was performed on the data. The statistically significant terms are sex, chest pain at the highest level (4 as compared to all lower levels), blood pressure, slope of the peak exercise ST segment at the level 2 as compared to level 1 and 3, thal content at the level 7 as compared to 3 and 6, vessels number with 2 vessel systems more prone to disease.

AIC based stepwise selection confirms the result. Both models chisq is ~0 for overall model testing.

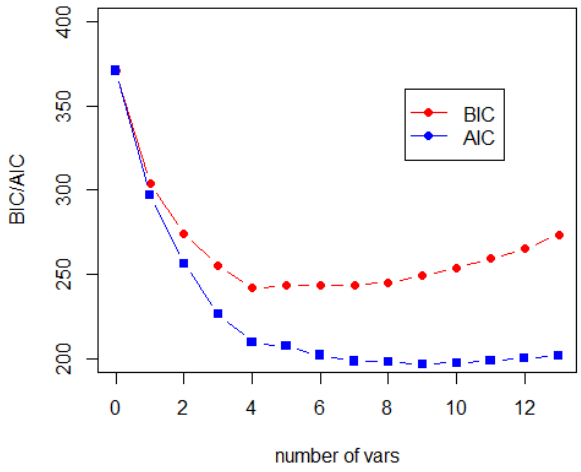
For the direct and step models subsequently variable with p-values above 5% were dropped and leave-one-out cross validation was performed on all models (Table 1). Tenfold cross validation is unstable because of the small sample. The best model is the Step-linear-dropped model by CV error rate, while step-linear for CV mean deviance. Coefficients are in Appendix A.

Model	CV error	AUC	CV mean deviance
Linear regression	0.1242613	0.9426	0.9285503
Step linear	0.1100258	0.9408	0.7420147
Linear p>5% dropped	0.1120733	0.9281	0.9285503
Step linear p>5% dropped	0.1091101	0.932	0.9285503

Appendix A presents R codes.

Best subset search by `bestglm` package

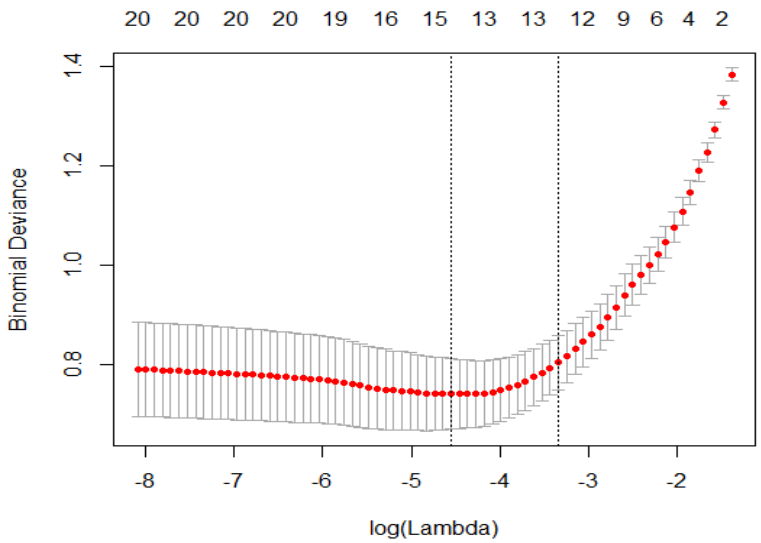
The search was done both by Bayes and Aikaike information criteria. BIC tended to keep smaller number of variables then AIC. However both models do not beat the CV errors of the simple search done through `Step.` AIC flattened from 4 variable which was favored by BIC.



Appendix B gives the details.

Lasso regression

Lasso regression was performed. The binomial deviance vs λ is fairly flat. It does however select mostly same variables. Deviance is substantially higher than step selection.



Coefficients are in Appendix C.

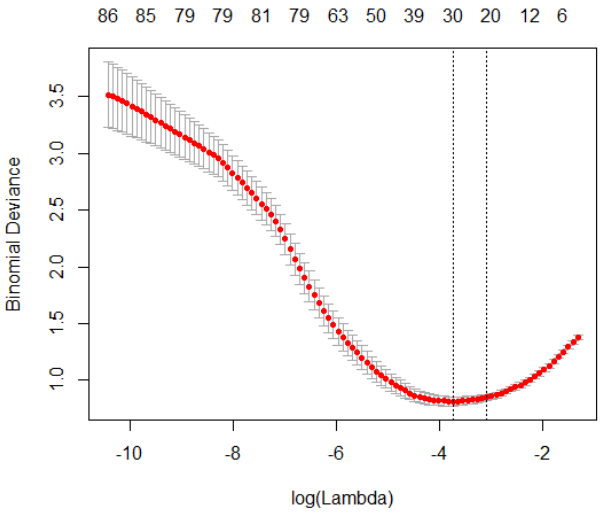
Interactions in the system

The number of possible cross couplings is 200 which is of the same order as the available sample of 270. However this can be handled with lasso selection method. We create a dataframe with all possible interaction and perform lasso regression. Selected variables are fed to `glm()` and `step()` function is applied. This leads to dramatic reduction of variable number and AIC = 154.7, showing significant improvement. Further output was fed again to lasso regression resulting to another substantial shrinkage and mean CV deviance of 0.6603861 which is markedly less than simple regression indicating significance of the interactions.

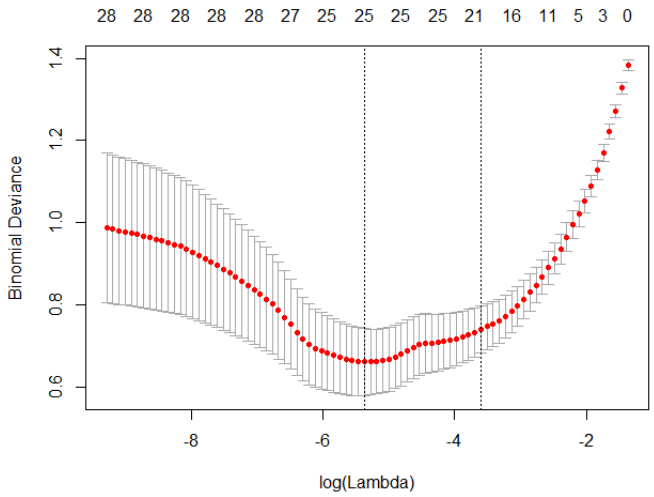
This is the final interaction model:

```
LogOdds = -1.032 +0.002 bloodpressure-0.012 maxheartrate+0.002 `sex1:cholesterol`+0.973 `sex1:vessels1`+0.02 `sex1:vessels2`+0.009 `chestpain4:bloodpressure`+0.19 `chestpain4:EKG2`+0.423 `chestpain4:STslope3`+0.611 `chestpain4:vessels1`+1.847 `chestpain2:vessels3`+0.902 `chestpain3:thal6`+0.232 `chestpain2:thal7`+0.008 `bloodpressure:thal7`+0.218 `bloodsugar1:vessels2`+0.876 `EKG2:vessels2`+1.122 `EKG2:vessels3`+0.252 `EKG2:thal7`+0.005 `maxheartrate:vessels2`+0.625 `STdepression:STslope2`+0.105 `STdepression:thal7`
```

Appendix D lists R outputs.



Lasso for initial shrinkage of 200 term regression



Lasso for further shrinkage after `step` function application

Appendix A. Linear Models.

Linear Regression

```
> linearmodel=glm(disease~., datsepfac, family = binomial)
> summary(linearmodel)

Call:
glm(formula = disease ~ ., family = binomial, data = datsepfac)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9284  -0.4382  -0.1170   0.2969   2.9516

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.686960   3.287776  -2.338 0.019385 *
age           -0.025110   0.026610  -0.944 0.345362
sex1           1.898998   0.606029   3.134 0.001727 **
chestpain2     1.741171   0.945881   1.841 0.065652 .
chestpain3     0.784877   0.792325   0.991 0.321881
chestpain4     2.748658   0.812980   3.381 0.000722 ***
bloodpressure  0.031110   0.012799   2.431 0.015074 *
cholesterol    0.006557   0.004300   1.525 0.127297
bloodsugar1   -0.376047   0.606920  -0.620 0.535522
EKG1           0.803613   3.561836   0.226 0.821499
EKG2           0.676840   0.425719   1.590 0.111863
maxheartrate  -0.020480   0.012182  -1.681 0.092736 .
angina1        0.534717   0.467726   1.143 0.252944
STdepression   0.476061   0.252840   1.883 0.059720 .
```

```
STslope2       1.113087   0.514352   2.164 0.030460 *
STslope3       0.128387   1.061333   0.121 0.903716
vessels1       2.152088   0.559653   3.845 0.000120 ***
vessels2       3.100493   0.807546   3.839 0.000123 ***
vessels3       2.164689   0.926071   2.337 0.019413 *
thal6         -0.318858   0.865164  -0.369 0.712461
thal7          1.468745   0.465422   3.156 0.001601 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.96  on 269  degrees of freedom
Residual deviance: 161.66  on 249  degrees of freedom
AIC: 203.66

Number of Fisher Scoring iterations: 6
```

Stepwise selection of relevant coefficients

```
> summary(linearmodelstep)

Call:
glm(formula = disease ~ sex1 + chestpain2 + chestpain4 + bloodpressure +
    cholesterol + EKG2 + maxheartrate + STdepression + STslope2 +
    vessels1 + vessels2 + vessels3 + thal7, family = binomial,
    data = datsepfac)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8348  -0.4829  -0.1128   0.3429   3.0375

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.731546   2.650368  -2.917 0.003532 **
sex1           1.732441   0.558984   3.099 0.001940 **
chestpain2     1.195658   0.698246   1.712 0.086828 .
chestpain4     2.357371   0.496127   4.752 2.02e-06 ***
bloodpressure  0.025370   0.011581   2.191 0.028483 *
cholesterol    0.005870   0.004063   1.445 0.148572
EKG2           0.615120   0.413251   1.488 0.136622
maxheartrate  -0.018225   0.010774  -1.692 0.090719 .
STdepression   0.525689   0.230162   2.284 0.022372 *
STslope2       1.027106   0.461723   2.225 0.026114 *
vessels1       2.045740   0.531070   3.852 0.000117 ***
vessels2       2.628448   0.735790   3.572 0.000354 ***
vessels3       1.994607   0.911307   2.189 0.028616 *
thal7          1.586373   0.436562   3.634 0.000279 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.96  on 269  degrees of freedom
Residual deviance: 165.30  on 256  degrees of freedom
AIC: 193.3
```

```
Number of Fisher Scoring iterations: 6
```

Linear model with p>5% dropped

```
> summary(linearmodeldropped)

Call:
glm(formula = disease ~ ., family = binomial, data = datlinearmodeldropped)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8543  -0.4439  -0.1466   0.4298   2.7945

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.53923    1.66454  -5.130 2.90e-07 ***
sex1          1.44621    0.47407   3.051  0.00228 **
chestpain4    2.16324    0.39579   5.466 4.61e-08 ***
bloodpressure 0.02971    0.01058   2.809  0.00497 **
STslope2      1.54905    0.39108   3.961 7.47e-05 ***
vessels1      2.01483    0.48701   4.137 3.52e-05 ***
vessels2      2.74066    0.67366   4.068 4.74e-05 ***
vessels3      2.18445    0.83078   2.629  0.00855 **
thal7         1.61926    0.40268   4.021 5.79e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.96  on 269  degrees of freedom
Residual deviance: 182.72  on 261  degrees of freedom
AIC: 200.72

Number of Fisher Scoring iterations: 6
```

Step linear model with p>5% dropped

```
> summary(linearmodelsd)

Call:
glm(formula = disease ~ ., family = binomial, data = datlinearmodelstepdroppe
d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.6445  -0.4671  -0.1566   0.4145   2.8717

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -8.13594    1.68004  -4.843 1.28e-06 ***
sex1          1.32941    0.48860   2.721 0.006512 **
chestpain4    2.17297    0.40780   5.329 9.90e-08 ***
bloodpressure 0.02500    0.01082   2.310 0.020887 *
STdepression  0.48492    0.20129   2.409 0.015992 *
```

```
STslope2       1.22341    0.41733   2.932 0.003373 **
vessels1       2.10388    0.49115   4.284 1.84e-05 ***
vessels2       2.49669    0.70614   3.536 0.000407 ***
vessels3       2.22326    0.86718   2.564 0.010354 *
thal7          1.47570    0.41043   3.595 0.000324 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.96  on 269  degrees of freedom
Residual deviance: 176.27  on 260  degrees of freedom
AIC: 196.27

Number of Fisher Scoring iterations: 6
```

CV deviance

```
> cv.glm(datsepfac, linearmodelsd, K=270, cost = function(y, yhat) (-2)*sum(y
*log(yhat)+(1-y)*log(1-yhat)))$delta
[1]  0.9285503 15.2910575
> cv.glm(datsepfac, linearmodel, K=270, cost = function(y, yhat) (-2)*sum(y*l
og(yhat)+(1-y)*log(1-yhat)))$delta
[1] 0.9285503 0.6854669
> cv.glm(datsepfac, linearmodelstep, K=270, cost = function(y, yhat) (-2)*sum
(y*log(yhat)+(1-y)*log(1-yhat)))$delta
[1] 0.7420147 0.6727551
> cv.glm(datsepfac, linearmodeldropped, K=270, cost = function(y, yhat) (-2)*
sum(y*log(yhat)+(1-y)*log(1-yhat)))$delta
[1]  0.9285503 21.7463905
```

Appendix B. Linear Best Subset Models.

AIC

```
> ttest$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)

Coefficients:
 (Intercept)    chestpain2    chestpain3    chestpain4  STdepression      vessels1
    -4.06797       0.85346       0.05852       2.30101       0.83783       2.08728
    vessels2      vessels3         thal6         thal7
     1.97869       2.56197       0.93401       1.97397

Degrees of Freedom: 269 Total (i.e. Null);  260 Residual
Null Deviance:       371
Residual Deviance: 191.5  AIC: 211.5
```

BIC

```
> ttestAIC$BestModel

Call:  glm(formula = y ~ ., family = family, data = Xi, weights = weights)
```

```
Coefficients:
  (Intercept)           sex1     chestpain2     chestpain3     chestpain4  bloodpressu
re
    -8.341128       1.938019       1.571628       0.598611       2.817274       0.0275
39
  cholesterol   maxheartrate   STdepression       STslope2       STslope3       vessel
s1
     0.006962      -0.019367       0.504231       1.139405       0.180206       2.0659
99
     vessels2       vessels3          thal6          thal7
     2.716824       1.922220      -0.454992       1.399752

Degrees of Freedom: 269 Total (i.e. Null);  254 Residual
Null Deviance:       371
Residual Deviance: 166.7  AIC: 198.7
```

Appendix C. Lasso Regression

```
> predict(linearlasso, type="coefficients", s=cvlasso$lambda.min)
21 x 1 sparse Matrix of class "dgCMatrix"
                            1
(Intercept)   -4.766851238
age            .
sex1           1.166277814
chestpain2     0.197492452
chestpain3     .
chestpain4     1.575719176
bloodpressure  0.015314691
cholesterol    0.003302964
bloodsugar1   -0.013187863
EKG1           .
EKG2           0.375985906
maxheartrate  -0.014037453
angina1        0.409356922
STdepression   0.365009637
STslope2       0.694077679
STslope3       .
vessels1       1.397648031
vessels2       1.906221033
vessels3       1.390049372
thal6          .
thal7          1.284240498
```

Appendix D. Interactions.

```
> summary(linearints)
```

```
After step() function
```

```
Call:
glm(formula = disease ~ bloodpressure + maxheartrate + vessels3 +
    `sex1:chestpain3` + `sex1:cholesterol` + `sex1:vessels1` +
    `sex1:vessels2` + `chestpain4:bloodpressure` + `chestpain3:bloodsugar1` +
```

```
    `chestpain4:EKG2` + `chestpain4:maxheartrate` + `chestpain2:STdepression`
+
    `chestpain4:STslope3` + `chestpain4:vessels1` + `chestpain2:vessels3` +
    `chestpain3:thal6` + `chestpain2:thal7` + `bloodpressure:thal7` +
    `bloodsugar1:vessels2` + `EKG2:STdepression` + `EKG2:vessels2` +
    `EKG2:vessels3` + `EKG2:thal7` + `maxheartrate:vessels2` +
    `maxheartrate:thal7` + `angina1:vessels2` + `STdepression:STslope2` +
    `STdepression:thal7`, family = binomial, data = datinttrunc)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0747  -0.1675  -0.0057   0.0226   3.5346

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.227e+01  3.860e+00  -3.179  0.00148 **
bloodpressure               1.346e-01  3.405e-02   3.953 7.73e-05 ***
maxheartrate               -8.649e-02  2.672e-02  -3.237  0.00121 **
vessels3                   -2.658e+00  1.753e+00  -1.516  0.12945
`sex1:chestpain3`           3.362e+00  1.307e+00   2.573  0.01009 *
`sex1:cholesterol`          5.251e-03  2.983e-03   1.760  0.07835 .
`sex1:vessels1`             2.149e+00  1.136e+00   1.892  0.05851 .
`sex1:vessels2`             7.000e+00  3.681e+00   1.901  0.05724 .
`chestpain4:bloodpressure` -5.410e-02  2.552e-02  -2.120  0.03403 *
`chestpain3:bloodsugar1`   -6.270e+00  2.513e+00  -2.495  0.01258 *
`chestpain4:EKG2`           1.367e+00  8.737e-01   1.565  0.11766
`chestpain4:maxheartrate`   7.288e-02  2.665e-02   2.735  0.00623 **
`chestpain2:STdepression`   3.864e+00  1.181e+00   3.272  0.00107 **
`chestpain4:STslope3`       2.442e+01  2.446e+03   0.010  0.99204
`chestpain4:vessels1`       2.482e+00  1.331e+00   1.865  0.06216 .
`chestpain2:vessels3`       2.205e+01  1.075e+04   0.002  0.99836
`chestpain3:thal6`          2.452e+01  6.680e+03   0.004  0.99707
`chestpain2:thal7`          5.953e+00  2.149e+00   2.770  0.00560 **
`bloodpressure:thal7`      -4.471e-02  2.537e-02  -1.762  0.07803 .
`bloodsugar1:vessels2`      8.643e+00  3.404e+00   2.539  0.01112 *
`EKG2:STdepression`        -1.686e+00  6.036e-01  -2.793  0.00522 **
`EKG2:vessels2`             4.822e+00  2.612e+00   1.846  0.06486 .
`EKG2:vessels3`             1.061e+01  4.828e+00   2.198  0.02793 *
`EKG2:thal7`                2.403e+00  1.161e+00   2.070  0.03850 *
`maxheartrate:vessels2`    -2.968e-02  2.258e-02  -1.314  0.18870
`maxheartrate:thal7`        4.210e-02  2.188e-02   1.924  0.05436 .
`angina1:vessels2`          2.202e+01  1.707e+03   0.013  0.98971
`STdepression:STslope2`     2.645e+00  6.524e-01   4.054 5.04e-05 ***
`STdepression:thal7`        9.407e-01  6.435e-01   1.462  0.14379
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.959  on 269  degrees of freedom
Residual deviance:  96.769  on 241  degrees of freedom
AIC: 154.77

Number of Fisher Scoring iterations: 18
```

Further shrinkage coefficients

```
> predict(linearintslasso, type="coefficients", family="binomial")
29 x 1 sparse Matrix of class "dgCMatrix"
                                      1
(Intercept)                -1.032454282
bloodpressure               0.001661569
maxheartrate               -0.012243279
vessels3                    .
`sex1:chestpain3`           .
`sex1:cholesterol`          0.002438602
`sex1:vessels1`             0.973485565
`sex1:vessels2`             0.019781497
`chestpain4:bloodpressure`  0.008948354
`chestpain3:bloodsugar1`    .
`chestpain4:EKG2`           0.189544418
`chestpain4:maxheartrate`   .
`chestpain2:STdepression`   .
`chestpain4:STslope3`       0.423068884
`chestpain4:vessels1`       0.610768142
`chestpain2:vessels3`       1.846877079
`chestpain3:thal6`          0.902354134
`chestpain2:thal7`          0.232159106
`bloodpressure:thal7`       0.007785827
`bloodsugar1:vessels2`      0.217563373
`EKG2:STdepression`         .
`EKG2:vessels2`             0.876366489
`EKG2:vessels3`             1.121891119
`EKG2:thal7`                0.251625154
`maxheartrate:vessels2`     0.005337359
`maxheartrate:thal7`        .
`angina1:vessels2`          .
`STdepression:STslope2`     0.625124831
`STdepression:thal7`        0.104955756
```