Honwork 2, task 2

Marat Habibullin

27.10.2014

Part 1: Private universities model

```
library(MASS)
library(lattice)
library(latticeExtra)
```

```
## Loading required package: RColorBrewer
```

```
library(e1071)

df <- read.csv2(file = "input-data/I.csv")
df <- subset(df, select = c(PPIND, NEW10, FULLTIME, IN_STATE, ROOM, ADD_FEE,
                            PH_D, GRADUAT, SAL_FULL, NUM_FULL))
df$PPIND <- factor(df$PPIND, labels = c("Public", "Private"))
df <- na.exclude(df)
df.priv <- subset(df, PPIND == "Private")

# all predictors with log
fit2 <- lm(NEW10 ~ FULLTIME + log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
           log(SAL_FULL) + PH_D + GRADUAT + NUM_FULL, data = df.priv)
summary(fit2)
```

```
##
## Call:
## lm(formula = NEW10 ~ FULLTIME + log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
##     log(SAL_FULL) + PH_D + GRADUAT + NUM_FULL, data = df.priv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.363  -8.451   1.984   7.871  23.657
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.875e+02  2.136e+02  -1.346  0.18955
## FULLTIME      -1.650e-04  9.139e-04  -0.181  0.85807
## log(IN_STATE)  1.138e+01  9.276e+00   1.227  0.23037
## log(ROOM)     -1.498e+01  1.025e+01  -1.462  0.15524
## log(ADD_FEE)  -4.071e+00  3.287e+00  -1.239  0.22616
## log(SAL_FULL)  4.255e+01  3.105e+01   1.370  0.18188
## PH_D           1.912e-01  5.197e-01   0.368  0.71578
## GRADUAT        8.897e-01  2.451e-01   3.631  0.00117 **
## NUM_FULL       3.442e-02  2.596e-02   1.326  0.19603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.64 on 27 degrees of freedom
```

```
## Multiple R-squared:  0.7865, Adjusted R-squared:  0.7233
## F-statistic: 12.43 on 8 and 27 DF,  p-value: 2.812e-07
```

Following the same reason as it was discussed in class we have removed AVRCOMB predictor in advance. Next let's try to simplify our model manually. We can start with the least significant predictors - FULLTIME and PH_D:

```
# manual removing
fit2.manual <- lm(NEW10 ~ log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
                    log(SAL_FULL) + GRADUAT + NUM_FULL, data = df.priv)
summary(fit2.manual)
```

```
##
## Call:
## lm(formula = NEW10 ~ log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
##     log(SAL_FULL) + GRADUAT + NUM_FULL, data = df.priv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.0604  -8.6451   0.8188   7.6399  24.3483
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -320.54341  189.28621  -1.693   0.1011
## log(IN_STATE)   13.69677    7.49452   1.828   0.0779 .
## log(ROOM)      -15.54037    9.84592  -1.578   0.1253
## log(ADD_FEE)    -3.68206    3.03277  -1.214   0.2345
## log(SAL_FULL)   46.50999   28.79141   1.615   0.1171
## GRADUAT          0.94127    0.20188   4.662 6.48e-05 ***
## NUM_FULL         0.02987    0.02071   1.442   0.1600
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.24 on 29 degrees of freedom
## Multiple R-squared:  0.7849, Adjusted R-squared:  0.7404
## F-statistic: 17.64 on 6 and 29 DF,  p-value: 1.718e-08
```

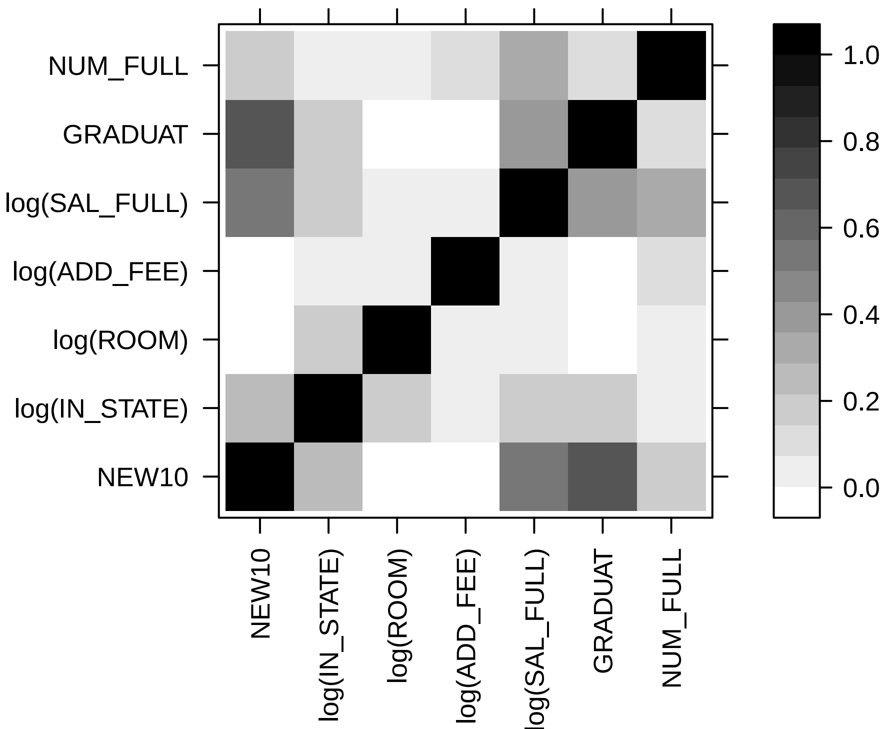
```
AIC(fit2.manual)
```

```
## [1] 290.7169
```

```
tune(lm, fit2.manual$call$formula, data = df.priv,
     tunecontrol = tune.control(sampling = "cross", cross = 36))
```

```
##
## Error estimation of 'lm' using leave-one-out: 208.5229
```

```
levelplot(cor(fit2.manual$model)^2, par.settings = list(regions = list(col = colorRampPalette(grey(1:0)
          scales = list(x = list(rot = 90)), xlab = "", ylab = "")
```



SAL_FULL correlates with GRADUAT and NUM_FULL and logically it is semantically very strange predictor to be significant. Let's remove it:

```
fit2.manual <- lm(NEW10 ~ log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
                                    GRADUAT + NUM_FULL, data = df.priv)
summary(fit2.manual)
```

```
##
## Call:
## lm(formula = NEW10 ~ log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
##     GRADUAT + NUM_FULL, data = df.priv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.957  -7.935   1.179   6.899  27.834
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -45.67356   85.12122  -0.537   0.5955
## log(IN_STATE)  16.34138    7.50716   2.177   0.0375 *
## log(ROOM)     -17.24994   10.04804  -1.717   0.0963 .
## log(ADD_FEE)   -3.14403    3.09423  -1.016   0.3177
## GRADUAT         1.11018    0.17727   6.263  6.7e-07 ***
## NUM_FULL        0.04737    0.01812   2.615   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.56 on 30 degrees of freedom
## Multiple R-squared:  0.7655, Adjusted R-squared:  0.7265
## F-statistic: 19.59 on 5 and 30 DF,  p-value: 1.207e-08
```

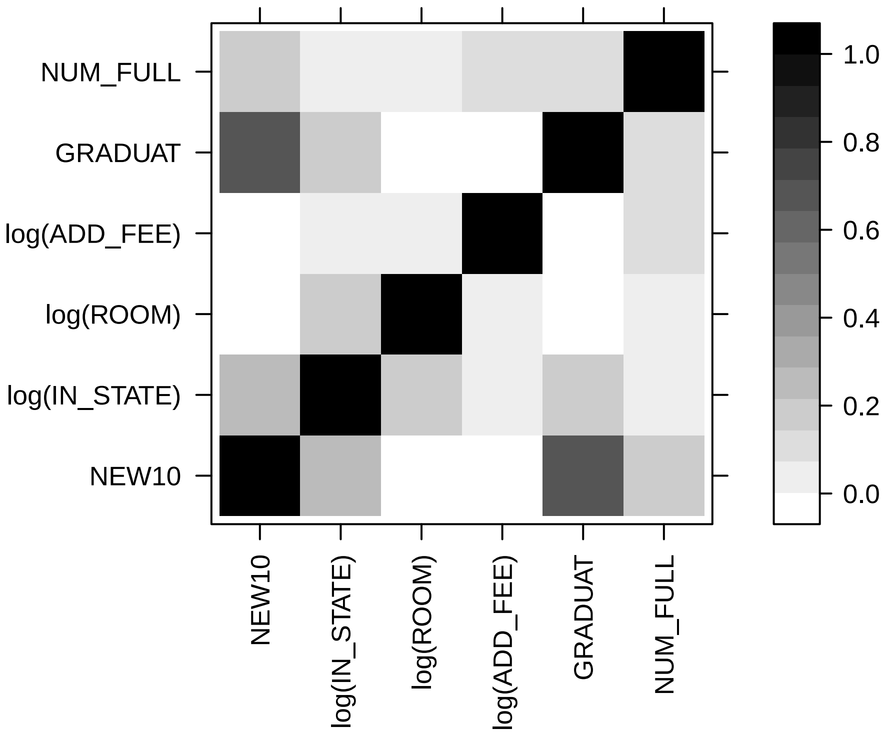
```
AIC(fit2.manual)
```

```
## [1] 291.8188
```

```
tune(lm, fit2.manual$call$formula, data = df.priv,
     tunecontrol = tune.control(sampling = "cross", cross = 36))
```

```
##
## Error estimation of 'lm' using leave-one-out: 199.6208
```

```
levelplot(cor(fit2.manual$model)^2, par.settings = list(regions = list(col = colorRampPalette(grey(1:0)
          scales = list(x = list(rot = 90)), xlab = "", ylab = "")
```



Next let's remove the least significant predictor - ADD_FEE:

```
fit2.manual <- lm(NEW10 ~ log(IN_STATE) + log(ROOM) +
                    GRADUAT + NUM_FULL, data = df.priv)
summary(fit2.manual)
```

```
##
## Call:
## lm(formula = NEW10 ~ log(IN_STATE) + log(ROOM) + GRADUAT + NUM_FULL,
##     data = df.priv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.366  -8.566   1.002   7.356  30.377
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -77.08700   79.34991  -0.971   0.3388
## log(IN_STATE)  17.09063    7.47476   2.286   0.0292 *
## log(ROOM)     -16.37263   10.01612  -1.635   0.1122
## GRADUAT         1.11249    0.17735   6.273 5.66e-07 ***
## NUM_FULL        0.04187    0.01730   2.420   0.0216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.57 on 31 degrees of freedom
## Multiple R-squared:  0.7575, Adjusted R-squared:  0.7262
## F-statistic: 24.21 on 4 and 31 DF,  p-value: 3.706e-09
```

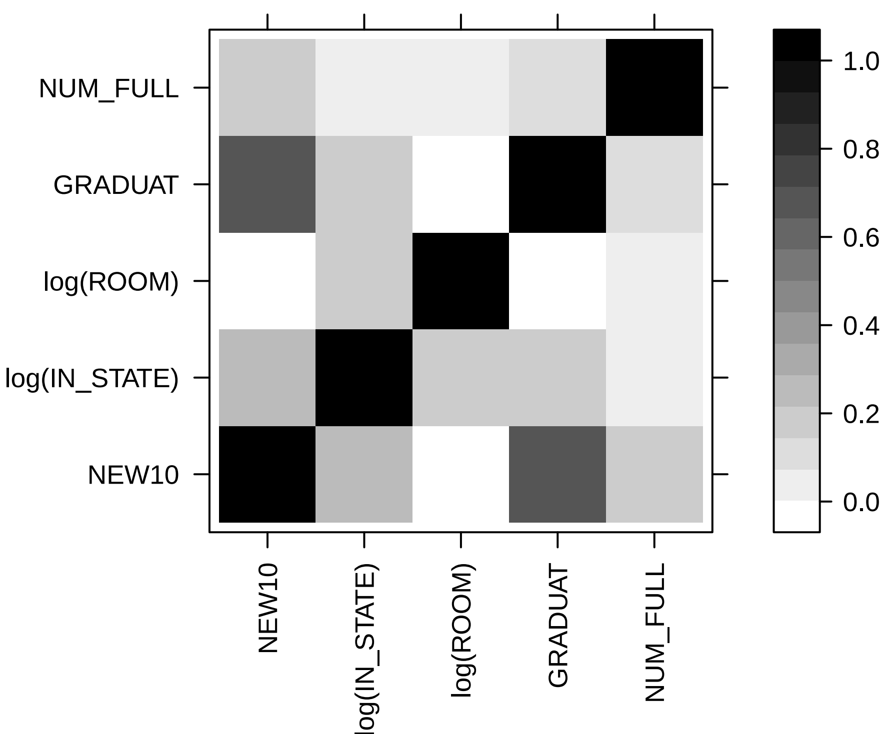
```
AIC(fit2.manual)
```

```
## [1] 291.0369
```

```
tune(lm, fit2.manual$call$formula, data = df.priv,
     tunecontrol = tune.control(sampling = "cross", cross = 36))
```

```
##
## Error estimation of 'lm' using leave-one-out: 190.2668
```

```
levelplot(cor(fit2.manual$model)^2, par.settings = list(regions = list(col = colorRampPalette(grey(1:0))
          scales = list(x = list(rot = 90)), xlab = "", ylab = "")
```



Looking at the correlation matrix plot one can notice ROOM and IN_STATE corellate. Moreover in real world private universities' students have rich parents so they don't bother much about money. Let's remove both predictors:

```r
fit2.manual <- lm(NEW10 ~
                    GRADUAT + NUM_FULL, data = df.priv)
summary(fit2.manual)
```

```
##
## Call:
## lm(formula = NEW10 ~ GRADUAT + NUM_FULL, data = df.priv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.336  -7.373   1.898   7.854  35.263
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -59.67788   13.29516  -4.489 8.25e-05 ***
## GRADUAT       1.29373    0.16769   7.715 6.93e-09 ***
## NUM_FULL      0.04101    0.01812   2.263   0.0304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.26 on 33 degrees of freedom
## Multiple R-squared:  0.7128, Adjusted R-squared:  0.6954
## F-statistic: 40.95 on 2 and 33 DF,  p-value: 1.149e-09
```

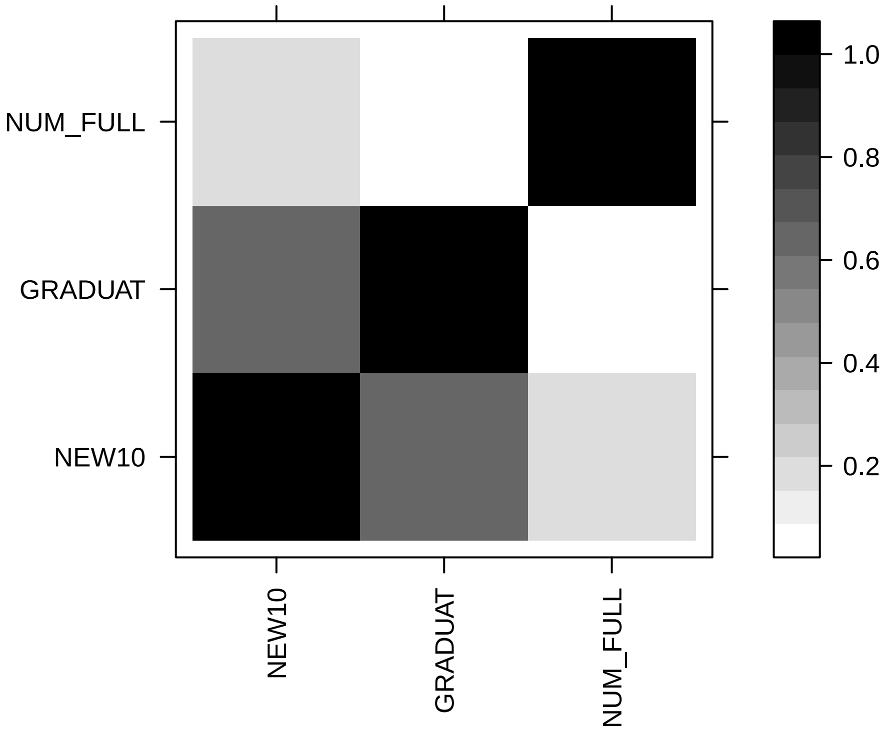
```r
AIC(fit2.manual)
```

```
## [1] 293.1246
```

```r
tune(lm, fit2.manual$call$formula, data = df.priv,
     tunecontrol = tune.control(sampling = "cross", cross = 36))
```

```
##
## Error estimation of 'lm' using leave-one-out: 185.3042
```

```r
levelplot(cor(fit2.manual$model)^2, par.settings = list(regions = list(col = colorRampPalette(grey(1:0)
          scales = list(x = list(rot = 90)), xlab = "", ylab = "")
```



Here we stop with manual removing. One can notice the AIC values have been increasing within model simplification, not much but nevertheless. On the other hand cross validation test error has been decreasing a little bit faster than AIC. The situation with AIC is kind of strange but we prefer to rely on cross validation here =).

Let's now try simplify initial model with stepAIC:

```
# removing with stepAIC
fit2.aic <- stepAIC(fit2)
```

```
## Start:  AIC=190.28
## NEW10 ~ FULLTIME + log(IN_STATE) + log(ROOM) + log(ADD_FEE) +
##     log(SAL_FULL) + PH_D + GRADUAT + NUM_FULL
##
##                 Df Sum of Sq    RSS    AIC
## - FULLTIME       1      5.21 4317.0 188.32
## - PH_D           1     21.62 4333.4 188.46
## - log(IN_STATE)  1    240.47 4552.3 190.24
## - log(ADD_FEE)   1    244.99 4556.8 190.27
## <none>                       4311.8 190.28
## - NUM_FULL       1    280.68 4592.5 190.55
## - log(SAL_FULL)  1    299.86 4611.7 190.70
## - log(ROOM)      1    341.42 4653.2 191.03
## - GRADUAT        1   2105.05 6416.8 202.59
##
## Step:  AIC=188.32
## NEW10 ~ log(IN_STATE) + log(ROOM) + log(ADD_FEE) + log(SAL_FULL) +
##     PH_D + GRADUAT + NUM_FULL
##
##                 Df Sum of Sq    RSS    AIC
## - PH_D           1     27.50 4344.5 186.55
```

```
## - log(ADD_FEE)   1    242.42 4559.4 188.29
## <none>                       4317.0 188.32
## - log(IN_STATE)  1    272.37 4589.4 188.53
## - log(SAL_FULL)  1    323.45 4640.5 188.93
## - NUM_FULL       1    337.81 4654.8 189.04
## - log(ROOM)      1    342.44 4659.4 189.07
## - GRADUAT        1   2100.07 6417.1 200.59
##
## Step:  AIC=186.55
## NEW10 ~ log(IN_STATE) + log(ROOM) + log(ADD_FEE) + log(SAL_FULL) +
##     GRADUAT + NUM_FULL
##
##                 Df Sum of Sq    RSS    AIC
## - log(ADD_FEE)   1     220.8 4565.3 186.34
## <none>                       4344.5 186.55
## - NUM_FULL       1     311.6 4656.1 187.05
## - log(ROOM)      1     373.2 4717.7 187.52
## - log(SAL_FULL)  1     390.9 4735.4 187.66
## - log(IN_STATE)  1     500.4 4844.9 188.48
## - GRADUAT        1    3256.7 7601.2 204.69
##
## Step:  AIC=186.34
## NEW10 ~ log(IN_STATE) + log(ROOM) + log(SAL_FULL) + GRADUAT +
##     NUM_FULL
##
##                 Df Sum of Sq    RSS    AIC
## - NUM_FULL       1     226.2 4791.5 186.08
## <none>                       4565.3 186.34
## - log(SAL_FULL)  1     333.1 4898.4 186.87
## - log(ROOM)      1     334.2 4899.5 186.88
## - log(IN_STATE)  1     591.2 5156.5 188.72
## - GRADUAT        1    3388.2 7953.5 204.32
##
## Step:  AIC=186.08
## NEW10 ~ log(IN_STATE) + log(ROOM) + log(SAL_FULL) + GRADUAT
##
##                 Df Sum of Sq    RSS    AIC
## - log(ROOM)      1     264.9 5056.4 186.02
## <none>                       4791.5 186.08
## - log(IN_STATE)  1     522.3 5313.8 187.80
## - log(SAL_FULL)  1    1032.7 5824.2 191.10
## - GRADUAT        1    3242.5 8034.0 202.69
##
## Step:  AIC=186.02
## NEW10 ~ log(IN_STATE) + log(SAL_FULL) + GRADUAT
##
##                 Df Sum of Sq    RSS    AIC
## <none>                       5056.4 186.02
## - log(IN_STATE)  1     303.7 5360.1 186.12
## - log(SAL_FULL)  1    1080.6 6136.9 190.99
## - GRADUAT        1    3560.5 8616.9 203.21
```

```
summary(fit2.aic)
```

```
##
## Call:
## lm(formula = NEW10 ~ log(IN_STATE) + log(SAL_FULL) + GRADUAT,
##     data = df.priv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.065  -7.876  -1.223   7.994  28.061
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -534.9545   152.3402  -3.512  0.00135 **
## log(IN_STATE)    9.4317     6.8031   1.386  0.17522
## log(SAL_FULL)   63.6849    24.3528   2.615  0.01349 *
## GRADUAT          0.9677     0.2038   4.747 4.14e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.57 on 32 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7262
## F-statistic: 31.94 on 3 and 32 DF,  p-value: 9.61e-10
```

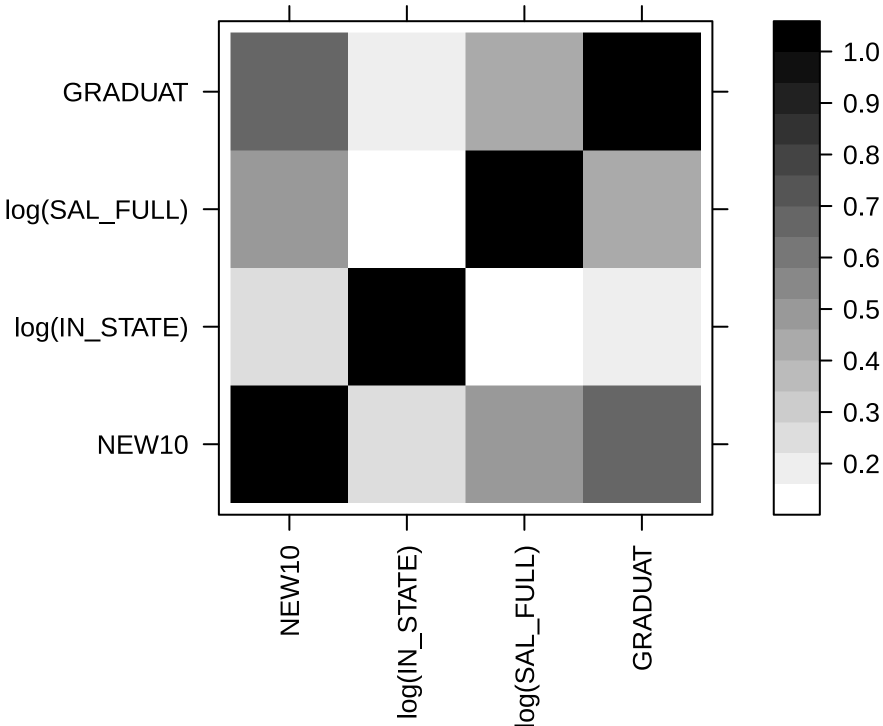
```r
AIC(fit2.aic)
```

```
## [1] 290.1793
```

```r
tune(lm, fit2.aic$call$formula, data = df.priv,
     tunecontrol = tune.control(sampling = "cross", cross = 36))
```

```
##
## Error estimation of 'lm' using leave-one-out: 189.7339
```

```r
levelplot(cor(fit2.aic$model)^2, par.settings = list(regions = list(col = colorRampPalette(grey(1:0)))),
          scales = list(x = list(rot = 90)), xlab = "", ylab = "")
```



We can see that IN_STATE is not significant so let's remove it:

```r
fit2.aic <- update(fit2.aic, . ~ . - log(IN_STATE))
summary(fit2.aic)
```

```
##
## Call:
## lm(formula = NEW10 ~ log(SAL_FULL) + GRADUAT, data = df.priv)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.8820  -6.8687   0.5165   8.0855  29.2902
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -490.0817   150.9274  -3.247  0.00268 **
## log(SAL_FULL)   69.7759    24.2856   2.873  0.00705 **
## GRADUAT          1.0373     0.2003   5.179 1.09e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.74 on 33 degrees of freedom
## Multiple R-squared:  0.7346, Adjusted R-squared:  0.7185
## F-statistic: 45.67 on 2 and 33 DF,  p-value: 3.118e-10
```

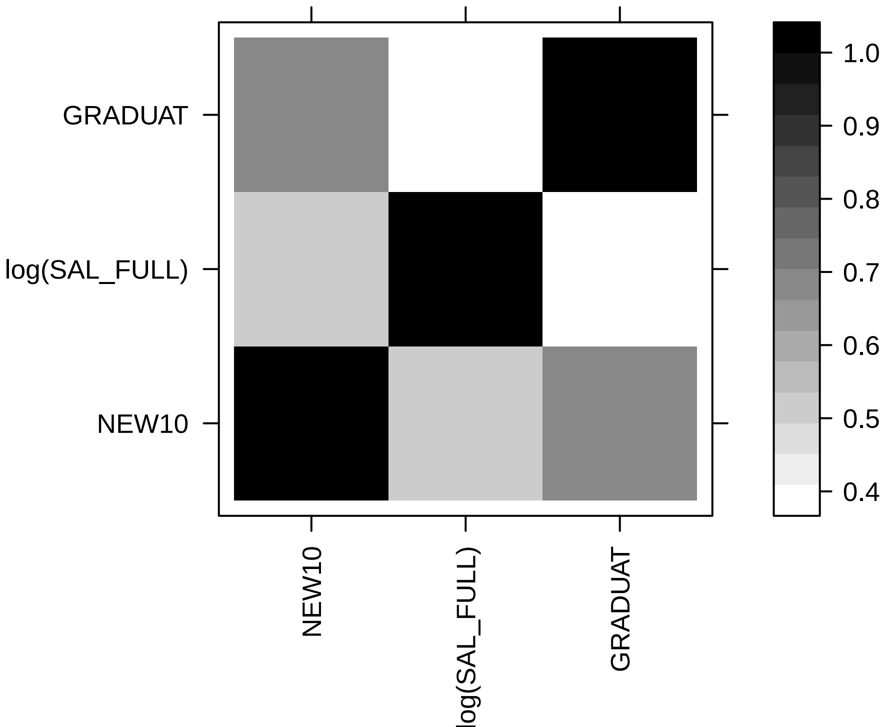
```r
AIC(fit2.aic)
```

```
## [1] 290.2792
```

```r
tune(lm, fit2.aic$call$formula, data = df.priv,
     tunecontrol = tune.control(sampling = "cross", cross = 36))
```

```
##
## Error estimation of 'lm' using leave-one-out: 176.3303
```

```r
levelplot(cor(fit2.aic$model)^2, par.settings = list(regions = list(col = colorRampPalette(grey(1:0)))),
          scales = list(x = list(rot = 90)), xlab = "", ylab = "")
```



Here we stop. As we can see this model is better than one created by manual removing. But this model is strange in fact having SAL_FULL as a significant predictor. We decide to rely on logic and use manual model as the final model for private universities. Moreover it will be very convenient later to merge private and public universities because our model of private unversities is a submodel of public ones.

The model has a pretty clear interpretation: the best newcomer students choosing private university prefer one with higher graduation percentage and higher number of good lecturers (full professors) and don't bother much about money factors due to a rich parents.

Part 2: general model

```r
# general model
contrasts(df$PPIND) <- contr.treatment
contrasts(df$PPIND)
```

```
##         2
## Public  0
## Private 1
```

```r
df$PPIND <- as.factor(df$PPIND)

gm <- lm(formula = NEW10 ~ (log(IN_STATE) + log(ADD_FEE) + GRADUAT + NUM_FULL) * PPIND, data = df)
summary(gm)
```

```
##
## Call:
## lm(formula = NEW10 ~ (log(IN_STATE) + log(ADD_FEE) + GRADUAT +
##     NUM_FULL) * PPIND, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.641  -7.887  -0.425   6.908  47.525
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.169e+02  3.114e+01   3.753 0.000319 ***
## log(IN_STATE)        -1.389e+01  3.625e+00  -3.832 0.000243 ***
## log(ADD_FEE)         -4.357e+00  2.011e+00  -2.167 0.033056 *
## GRADUAT               7.875e-01  1.305e-01   6.034 4.05e-08 ***
## NUM_FULL              3.756e-03  8.242e-03   0.456 0.649785
## PPIND2               -2.559e+02  7.363e+01  -3.475 0.000806 ***
## log(IN_STATE):PPIND2  2.469e+01  7.798e+00   3.166 0.002143 **
## log(ADD_FEE):PPIND2   1.669e+00  3.729e+00   0.448 0.655565
## GRADUAT:PPIND2        3.861e-01  2.196e-01   1.758 0.082341 .
## NUM_FULL:PPIND2       3.995e-02  2.009e-02   1.988 0.049994 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.8 on 85 degrees of freedom
## Multiple R-squared:  0.7323, Adjusted R-squared:  0.704
## F-statistic: 25.84 on 9 and 85 DF,  p-value: < 2.2e-16
```

```r
AIC(gm)
```

```
## [1] 765.3821
```

```r
tune(lm, gm$call$formula, data = df,
     tunecontrol = tune.control(sampling = "cross", cross = 95))
```

```
##
## Error estimation of 'lm' using leave-one-out: 190.6542
```

```r
gm.aic <- stepAIC(gm)
```

```
## Start:  AIC=493.78
## NEW10 ~ (log(IN_STATE) + log(ADD_FEE) + GRADUAT + NUM_FULL) *
##     PPIND
##
##                        Df Sum of Sq   RSS    AIC
## - log(ADD_FEE):PPIND    1     32.81 13953 492.01
```

```
## <none>                               13920 493.78
## - GRADUAT:PPIND        1    506.14 14426 495.18
## - NUM_FULL:PPIND       1    647.43 14567 496.10
## - log(IN_STATE):PPIND  1   1641.96 15562 502.38
##
## Step:  AIC=492.01
## NEW10 ~ log(IN_STATE) + log(ADD_FEE) + GRADUAT + NUM_FULL + PPIND +
##     log(IN_STATE):PPIND + GRADUAT:PPIND + NUM_FULL:PPIND
##
##                       Df Sum of Sq   RSS    AIC
## <none>                               13953 492.01
## - GRADUAT:PPIND        1    553.67 14506 493.70
## - NUM_FULL:PPIND       1    742.60 14695 494.93
## - log(ADD_FEE)         1    855.97 14809 495.66
## - log(IN_STATE):PPIND  1   1611.68 15564 500.39
```

```r
summary(gm.aic)
```

```
##
## Call:
## lm(formula = NEW10 ~ log(IN_STATE) + log(ADD_FEE) + GRADUAT +
##     NUM_FULL + PPIND + log(IN_STATE):PPIND + GRADUAT:PPIND +
##     NUM_FULL:PPIND, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.840  -7.370  -0.553   7.379  47.708
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.129e+02  2.968e+01   3.803 0.000267 ***
## log(IN_STATE)        -1.365e+01  3.570e+00  -3.825 0.000247 ***
## log(ADD_FEE)         -3.871e+00  1.685e+00  -2.297 0.024052 *
## GRADUAT               7.741e-01  1.264e-01   6.122 2.67e-08 ***
## NUM_FULL              3.883e-03  8.199e-03   0.474 0.636944
## PPIND2               -2.417e+02  6.620e+01  -3.652 0.000447 ***
## log(IN_STATE):PPIND2  2.406e+01  7.635e+00   3.152 0.002234 **
## GRADUAT:PPIND2        3.998e-01  2.164e-01   1.847 0.068138 .
## NUM_FULL:PPIND2       4.183e-02  1.955e-02   2.139 0.035235 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.74 on 86 degrees of freedom
## Multiple R-squared:  0.7317, Adjusted R-squared:  0.7068
## F-statistic: 29.32 on 8 and 86 DF,  p-value: < 2.2e-16
```

```r
AIC(gm.aic)
```

```
## [1] 763.6058
```

```r
tune(lm, gm.aic$call$formula, data = df,
     tunecontrol = tune.control(sampling = "cross", cross = 95))
```

```
##
## Error estimation of 'lm' using leave-one-out: 187.6378
```

Finally let's try to interpret general model predictors' coefficients.

First of all we have `-13.6 * log(IN_STATE) + 24 * log(IN_STATE) * PPIND`. That means that in public universities the lower tuition fees are the higher percentage of good newcomer students is and vice versa in private ones. The first thing is pretty intuitive. The second one could be interpreted in the following way: private universities' students being from rich families don't bother much about money moreover high fees may be used as university's quality factor or prestige value.

`log(ADD_FEE)` has a negative influence on a percentage of good newcomers in both cases. So nobody likes to spend additional money even rich people because in this case the reasons of additional fees are not completely understandable (in the case of IN_STATE fees there is a clear reason for this).

`GRADUAT` has a positive influence on the `NEW10` moreover the influence is higher in the case of private universities. The reason of it can be found in the fact there are not only students from wellbeing families in private universities but also there are newcomers who has not very rich parents. This students has a double interest in the graduation percentage because they don't want to spend their money in vain.

`NUM_FULL` has a positive influence and again higher in private universities so may be it is all about prestige value or quality level.

And finally `PPIND` has a negative coefficient so `NEW10` is a little bit higher in public universities (according to our encoding for `PPIND`).