

Stepwise-Regression.R

Shraddha Somani

```
require(car)

## Loading required package: car

head(Salaries)

##           rank discipline yrs.since.phd yrs.service sex salary
## 1         Prof          B             19           18 Male 139750
## 2         Prof          B             20           16 Male 173200
## 3   AsstProf          B              4              3 Male  79750
## 4         Prof          B             45           39 Male 115000
## 5         Prof          B             40           41 Male 141500
## 6 AssocProf          B              6              6 Male  97000

Salaries$rownames = rownames(Salaries)
Salaries$rownames = NULL
Salaries.old = Salaries
Salaries = na.omit(Salaries)
levels(Salaries$discipline) = c("Theoretical", "Applied")
some(Salaries)

##           rank discipline yrs.since.phd yrs.service sex salary
## 51         Prof      Applied             28           28 Male 126621
## 52         Prof      Applied             12           11 Male 108875
## 65   AsstProf      Applied              4              3 Male  68404
## 106        Prof Theoretical             31           28 Male 113543
## 123        Prof Theoretical             24           22 Male  97262
## 138        Prof Theoretical             17           14 Male 105668
## 168        Prof      Applied             18           19 Male 130664
## 193        Prof      Applied             19           18 Male 122100
## 226        Prof Theoretical             20           20 Male 122400
## 234        Prof Theoretical             36           19 Female 117555

# Perform stepwise regression strating with the full model using all the
# predictors of salary.
g1 = lm(salary ~ ., data = Salaries)
summary(g1)

##
## Call:
## lm(formula = salary ~ ., data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65248 -13211  -1775   10384  99592
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    65955.2     4588.6  14.374 < 2e-16 ***
## rankAssocProf  12907.6     4145.3   3.114  0.00198 **
## rankProf       45066.0     4237.5  10.635 < 2e-16 ***
## disciplineApplied 14417.6     2342.9   6.154 1.88e-09 ***
## yrs.since.phd    535.1       241.0   2.220  0.02698 *
## yrs.service     -489.5       211.9  -2.310  0.02143 *
## sexMale         4783.5      3858.7   1.240  0.21584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22540 on 390 degrees of freedom
## Multiple R-squared:  0.4547, Adjusted R-squared:  0.4463
## F-statistic: 54.2 on 6 and 390 DF,  p-value: < 2.2e-16

# - We eliminate Sex, since it has the highest p-value.
# - Next we re-run the model without this variable.
g2 = update(g1, . ~ . - sex)
summary(g2)

##
## Call:
## lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,
##     data = Salaries)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -65244 -13498  -1455    9638   99682
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    69869.0     3332.1  20.968 < 2e-16 ***
## rankAssocProf  12831.5     4147.7   3.094  0.00212 **
## rankProf       45287.7     4236.7  10.689 < 2e-16 ***
## disciplineApplied 14505.2     2343.4   6.190 1.52e-09 ***
## yrs.since.phd    534.6       241.2   2.217  0.02720 *
## yrs.service     -476.7       211.8  -2.250  0.02497 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22550 on 391 degrees of freedom
## Multiple R-squared:  0.4525, Adjusted R-squared:  0.4455
## F-statistic: 64.64 on 5 and 391 DF,  p-value: < 2.2e-16

# - There are no more variables with p-values higher than 0.05.
# - The final model is stored in the object g2.
# - Stepwise Regression
g3 = step(g1)
```

```
## Start: AIC=7965.19
## salary ~ rank + discipline + yrs.since.phd + yrs.service + sex
##
##           Df Sum of Sq      RSS      AIC
## - sex      1 7.8068e+08 1.9890e+11 7964.8
## <none>      1.9812e+11 7965.2
## - yrs.since.phd 1 2.5041e+09 2.0062e+11 7968.2
## - yrs.service  1 2.7100e+09 2.0083e+11 7968.6
## - discipline   1 1.9237e+10 2.1735e+11 8000.0
## - rank         2 6.9508e+10 2.6762e+11 8080.6
##
## Step: AIC=7964.75
## salary ~ rank + discipline + yrs.since.phd + yrs.service
##
##           Df Sum of Sq      RSS      AIC
## <none>      1.9890e+11 7964.8
## - yrs.since.phd 1 2.5001e+09 2.0140e+11 7967.7
## - yrs.service  1 2.5763e+09 2.0147e+11 7967.9
## - discipline   1 1.9489e+10 2.1839e+11 7999.9
## - rank         2 7.0679e+10 2.6958e+11 8081.5

# Compare the coefficients of the stepwise model and the full model.
coefs = compareCoefs(g1, g2, se = FALSE)

##
## Call:
## 1: lm(formula = salary ~ ., data = Salaries)
## 2: lm(formula = salary ~ rank + discipline + yrs.since.phd +
##    yrs.service, data = Salaries)
##           Est. 1 Est. 2
## (Intercept)   65955  69869
## rankAssocProf  12908  12832
## rankProf      45066  45288
## disciplineApplied 14418 14505
## yrs.since.phd    535    535
## yrs.service     -490   -477
## sexMale         4784

colnames(coefs) = c("g.OLS", "g.OLS.step")
coefs

##           g.OLS g.OLS.step
## (Intercept)  65955.2324 69869.0110
## rankAssocProf 12907.5879 12831.5375
## rankProf      45065.9987 45287.6890
## disciplineApplied 14417.6256 14505.1514
## yrs.since.phd    535.0583  534.6313
## yrs.service     -489.5157 -476.7179
## sexMale        4783.4928      NA
```

```

g.OLS = g1
g.OLS.step = g2

# Which variable of variables did stepwise drop from the full model?
# Answer:
# - Variable sex was dropped by stepwise from the full model

# Perform a cross-validation of the stepwise model.
require(DAAG)

## Loading required package: DAAG

## Loading required package: lattice

##
## Attaching package: 'DAAG'

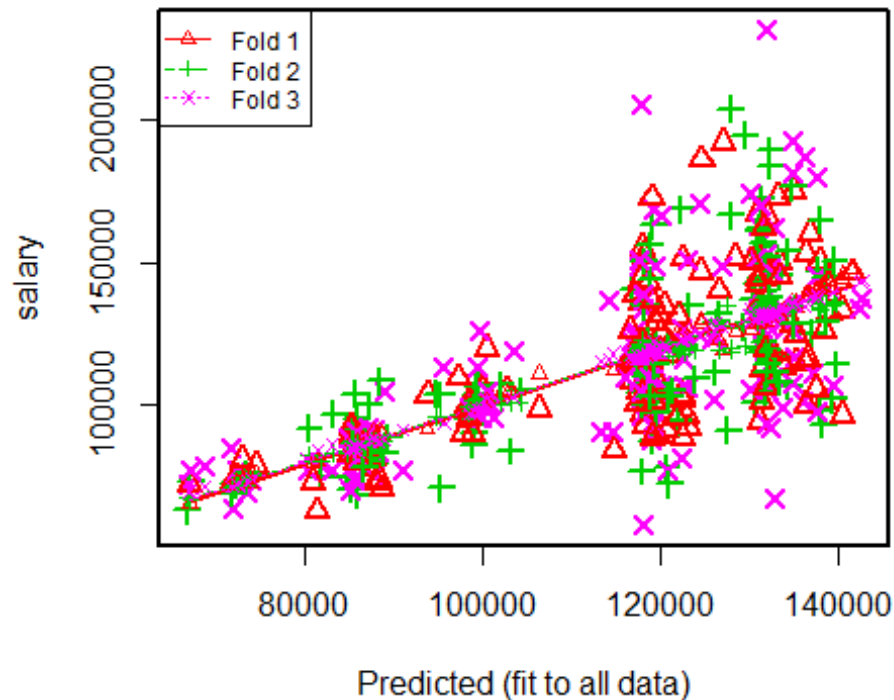
## The following object is masked from 'package:car':
##
##      vif

CVlm(data = Salaries, form.lm = g.OLS, printit = F)

## Warning in CVlm(data = Salaries, form.lm = g.OLS, printit = F):
##
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate

```

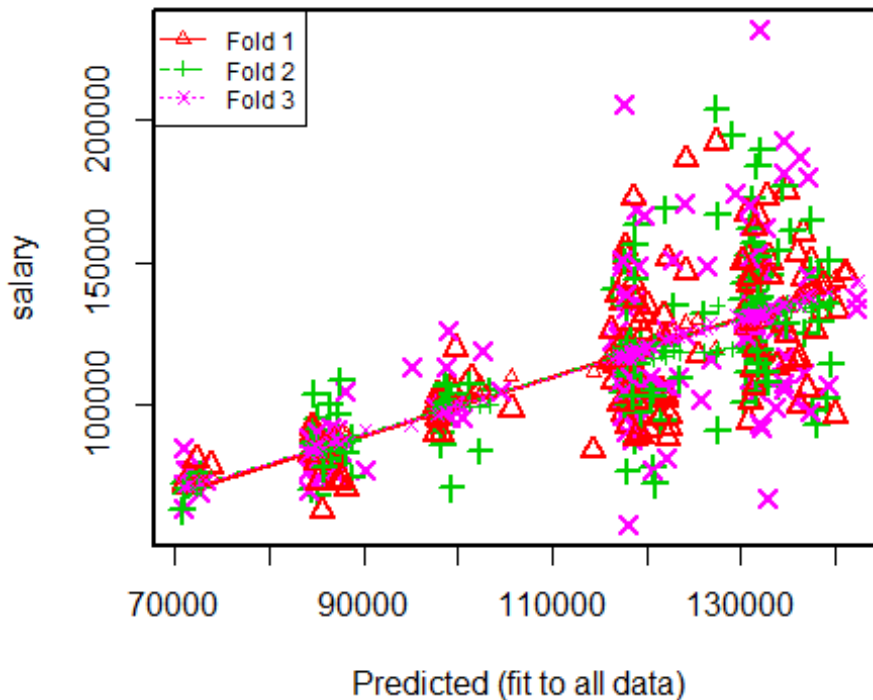
Small symbols show cross-validation predicted values



```
CVlm(data = Salaries, form.lm = g.OLS.step, printit = F)
```

```
## Warning in CVlm(data = Salaries, form.lm = g.OLS.step, printit = F):  
##  
## As there is >1 explanatory variable, cross-validation  
## predicted values for a fold are not a linear function  
## of corresponding overall predicted values. Lines that  
## are shown for the different folds are approximate
```

Small symbols show cross-validation predicted values



```
# Compare the two models using the mse's from the cross-validations with
# number of folds equal to 3. Which model gives the better mse?
oldpar = par(mfrow = c(1, 2))
mse.g.OLS = CVlm(data = Salaries, form.lm = g.OLS, m = 3, main = "Prediction
Plot: g.OLS")
```

```
## Analysis of Variance Table
```

```
##
```

```
## Response: salary
```

```
##          Df    Sum Sq  Mean Sq F value Pr(>F)
## rank      2 1.43e+11  7.16e+10   140.98 <2e-16 ***
## discipline 1 1.84e+10  1.84e+10    36.28  4e-09 ***
## yrs.since.phd 1 1.66e+08  1.66e+08     0.33  0.568
## yrs.service  1 2.58e+09  2.58e+09     5.07  0.025 *
## sex         1 7.81e+08  7.81e+08     1.54  0.216
## Residuals   390 1.98e+11  5.08e+08
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Warning in CVlm(data = Salaries, form.lm = g.OLS, m = 3, main =
"Prediction Plot: g.OLS"):
```

```
##
```

```
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate
```

```

##
## fold 1
## Observations in test set: 132
##      2      4      7     11     14     18     21     28     30
## Predicted 133091 135209 135015 100569 86226 119493 119664 86363 132727
## cvpred    132796 134436 135361 101653 85438 119327 120221 85426 132826
## salary    173200 115000 175000 119800 78000 103450 89565 82379 118223
## CV residual 40404 -19436 39639 18147 -7438 -15877 -30656 -3047 -14603
##      31      38      40      41      43      45      46      48
## Predicted 138965 86943.7 98474 141550 138408 131088 136256 128444
## cvpred    143209 86286.2 98193 147536 140529 129328 137983 126137
## salary    132261 86373.0 100938 146500 101299 94384 114778 151768
## CV residual -10948 86.8 2745 -1036 -39230 -34944 -23205 25631
##      49      57      61      64      66      71      73      75      76
## Predicted 126577 132249 98963 93781 98963 138339 136438 133945 87968
## cvpred    122659 131049 99061 91539 99061 142352 137968 133633 88018
## salary    140096 117704 90304 103613 100522 126320 100131 113398 73266
## CV residual 17437 -13345 -8757 12074 1461 -16032 -37837 -20235 -14752
##      77      79      80      88      92      93      94      96     101
## Predicted 137850 86272 87388 85247 99988 99988 131953 87297 140457
## cvpred    141485 85434 87158 83702 100793 100793 129256 87165 143992
## salary    150480 86100 84240 88400 105128 105631 166024 84000 133217
## CV residual 8995 666 -2918 4698 4335 4838 36768 -3165 -10775
##      103     108     109     110     112     115     121     123     124
## Predicted 136336 85081 87756 122032 85980 117442 116442 117877 81470
## cvpred    138885 86525 90845 123658 87358 119721 115947 117645 80689
## salary    153303 82600 81500 131205 82100 105000 115313 97262 62884
## CV residual 14418 -3925 -9345 7547 -5258 -14721 -634 -20383 -17805
##      127     130     131     134     136     140     145     147     149
## Predicted 118059 71900 85126 67071 117695 118230 131452 85339 131974
## cvpred    117629 72030 86521 65387 117660 118524 129298 83694 131294
## salary    155500 73000 83001 72500 136000 152664 112696 92000 144651
## CV residual 37871 970 -3520 7113 18340 34140 -16602 8306 13357
##      152     163     164     165     166     167     172     182     183
## Predicted 85339 106409 85293 85691 137542 130951 131088 137406 98428
## cvpred    83694 111160 83698 84574 140602 129340 129328 140613 98197
## salary    92000 98510 89942 88795 105890 167284 151575 141136 100000
## CV residual 8306 -12650 6244 4221 -34712 37944 22247 523 1803
##      184     185     195     197     198     201     203     206     213
## Predicted 133365 131270 97802 85339 85339 85339 136882 140480 134822
## cvpred    132773 129313 93706 83694 83694 83694 138839 145808 136286
## salary    150000 101000 90000 92700 92000 92700 160400 96545 128400
## CV residual 17227 -28313 -3706 9006 8306 9006 21561 -49263 -7886
##      218     219     220     232     234     237     240     242     244
## Predicted 102811 97345 131793 81026 120983 116943 123034 117706 116852
## cvpred    104191 97602 131087 79817 123968 115905 127209 116750 115913
## salary    105350 109650 119500 73300 117555 114000 96200 122875 108200
## CV residual 1159 12048 -11587 -6517 -6413 -1905 -31009 6125 -7713
##      245     247     249     252     258     260     266     267     269
## Predicted 114712 120598 119527 122590 88439 118241 120381 117763 119414

```

```

## cvpred      112457 121961 120233 126338  90787 117614 121070 115837 118425
## salary      84273  91100 128800 102000  74000  92550 134800 143940  89650
## CV residual -28184 -30861  8567 -24338 -16787 -25064 13730  28103 -28775
##              271    279    282    286    297    300    301    309    311
## Predicted   118697 119835 122385 85878 116624  88633 119049 73414 123079
## cvpred      117576 121116 124537 84640 115932  89862 118455 74629 127205
## salary      143250 107100 103600 81800 126300  70700  88600 74000  92050
## CV residual  25674 -14016 -20937 -2840  10368 -19162 -29855  -629 -35155
##              312    314    315    316    319    322    327    330    331
## Predicted   118890 117399 124639 86898 130951 98474 135186 131270 127069
## cvpred      120286 115867 129801 86290 129340 98193 136255 129313 119668
## salary      108100 100351 146800 84716 134550 95642 124714 134778 192253
## CV residual -12186 -15516 16999 -1574  5210 -2551 -11541  5465  72585
##              332    341    344    345    347    349    350    353    356
## Predicted   136518 131793 131634 130211 138943 85828 130962 130917 133319
## cvpred      138870 131087 129283 126675 141393 84562 128430 128434 132777
## salary      116518 106231 162150 150376 142023 80139 144309 142500 145028
## CV residual -22352 -24856  32867  23701  630 -4423  15879  14066  12251
##              357    360    361    363    366    367    369    374    376
## Predicted   122442 74666 117911 117171 119232 118935 119846 119129 118059
## cvpred      123623 76342 118551 115886 118440 120282 120206 119357 117629
## salary      88709  78785 121946 138771 101036 115435 131950 136660 103649
## CV residual -34914  2443  3395  22885 -17404 -4847  11744  17303 -13980
##              377    382    385    390    392    396    397
## Predicted   72389 118992 119971 124650 122556 121838 73061
## cvpred      72898 119369 121104 128891 125432 124583 73750
## salary      74856 172505 125192 186023 151292  95329 81035
## CV residual  1958  53136  4088  57132  25860 -29254  7285
##
## Sum of squares = 5.71e+10    Mean square = 4.33e+08    n = 132
##
## fold 2
## Observations in test set: 133
##              1    6    10    17    25    27    32    33    37
## Predicted   131577 98337 126259 130598 81902 123273 87923 132772 131714
## cvpred      133583 99809 128819 133622 80511 119953 87422 132775 133958
## salary      139750 97000 129000 101000 74830 134885 79916 117256 155750
## CV residual  6167 -2809  181 -32622 -5681 14932 -7506 -15519  21792
##              39    42    47    50    53    54    55    56    58
## Predicted   136985 99111 137918  85202 84790 134378 99100 103107  98963
## cvpred      136158 101934 135995  86769 83205 133092 100540 100635 100165
## salary      125196  93418  98193  70768 74692 106639 103760  83900  90215
## CV residual -10962 -8516 -37802 -16001 -8513 -26453  3220 -16735 -9950
##              65    69    70    72    81    82    83    86    89
## Predicted   85828 126213 127581 132272 139523 131486 132203 131395 131361
## cvpred      87124 128694 134882 136851 136311 133333 133939 133083 134352
## salary      68404 111512  91412 146856 150743 135585 144640 132825 172272
## CV residual -18720 -17182 -43470  10005 14432  2252  10701  -258  37920
##              90    91    99    104    107    111    113    114    116
## Predicted   99453 83276 139421 129287 85616 118674 71854 117490 122635

```


## cvpred	100145	83138	134668	128953	84959	118233	71563	120435	118203
## salary	107008	97032	102235	127512	82099	112429	72500	104279	120806
## CV residual	6863	13894	-32433	-1441	-2860	-5804	937	-16156	2603
##	119	120	126	129	132	135	138	141	142
## Predicted	72389	67162	120712	118241	117866	122294	118047	87221	86777
## cvpred	71668	67154	122462	119771	122828	119991	117877	85276	85420
## salary	72500	73500	78162	113278	76840	168635	105668	100102	81500
## CV residual	832	6346	-44300	-6493	-45988	48644	-12209	14826	-3920
##	146	148	150	154	156	159	160	162	171
## Predicted	131498	131452	85828	94806	131179	98337	130416	134833	85384
## cvpred	134727	134602	87124	95862	133852	99809	133122	134342	87269
## salary	119015	156938	95079	103994	118971	95408	137167	176500	91227
## CV residual	-15712	22336	7955	8132	-14881	-4401	4045	42158	3958
##	173	175	176	179	180	181	187	189	190
## Predicted	138271	104223	132966	137816	80509	130723	95341	99339	134298
## cvpred	135600	100971	134669	134351	82360	132603	95967	102558	134236
## salary	93164	105000	111751	147349	92000	142467	103750	106300	153750
## CV residual	-42436	4029	-22918	12998	9640	9864	7783	3742	19514
##	193	194	199	200	202	204	205	208	212
## Predicted	131577	98929	132260	139785	132044	130997	137941	131042	131999
## cvpred	133583	101434	135457	135667	136227	133352	133332	133477	136102
## salary	122100	86250	189409	114500	119700	152500	165000	120000	111350
## CV residual	-11483	-15184	53952	-21167	-16527	19148	31668	-13477	-24752
##	216	223	224	225	238	241	243	248	250
## Predicted	133399	102083	138624	119004	66764	71945	118025	118230	127895
## cvpred	133131	100548	135206	120501	67423	71813	120540	118377	119009
## salary	145350	107150	129600	87800	63100	69200	102600	101100	204000
## CV residual	12219	6602	-5606	-32701	-4323	-2613	-17940	-17277	84991
##	253	256	257	261	262	263	268	270	272
## Predicted	126517	86151	116807	89430	117854	119664	118400	119334	129466
## cvpred	119317	85064	118560	88611	121434	119588	117483	117319	120595
## salary	132000	83000	140300	88600	107550	121200	104350	103700	194800
## CV residual	12683	-2064	21740	-11	-13884	1612	-13133	-13619	74205
##	274	278	280	284	287	289	290	293	296
## Predicted	73061	119175	117900	118833	117569	118104	72572	132266	119584
## cvpred	72149	119607	121559	121396	119290	119396	72168	120104	120732
## salary	74000	163200	100600	155865	115800	150500	74000	183800	97150
## CV residual	1851	43593	-20959	34469	-3490	31104	1832	63696	-23582
##	299	304	305	306	307	308	313	317	321
## Predicted	120973	121337	118879	118287	72526	118685	122021	95296	131805
## cvpred	121818	117367	121521	119896	72043	119627	119241	95842	134208
## salary	72300	105260	144050	111350	74500	122500	94350	71065	104428
## CV residual	-49518	-12107	22529	-8546	2457	2873	-24891	-24777	-29780
##	324	325	326	328	334	338	339	343	346
## Predicted	130938	131099	87479	131907	138578	131304	135106	131953	133353
## cvpred	129395	134996	87566	135852	135081	132833	135091	135977	133006
## salary	161101	162221	84500	151650	145098	145000	128464	114596	107986
## CV residual	31706	27225	-3066	15798	10017	12167	-6627	-21381	-25020
##	348	354	355	358	368	370	371	372	375
## Predicted	134936	132203	88947	119539	88507	118287	86686	118321	120950

```

## cvpred      135986 133939 87508 120607 84718 119896 85170 118627 119030
## salary      128250 138000 83600 107309 108413 134690 78182 110515 103275
## CV residual -7736 4061 -3908 -13298 23695 14794 -6988 -8112 -15755
##            380 381 383 386 388 391 393
## Predicted   85616 73551 85479 119425 123979 127906 118776
## cvpred      84959 72130 84584 117569 119164 120403 119877
## salary      104121 75996 86895 114330 109305 166605 103106
## CV residual 19162 3866 2311 -3239 -9859 46202 -16771
##
## Sum of squares = 6.83e+10    Mean square = 5.14e+08    n = 133
##
## fold 3
## Observations in test set: 132
##            3      5      8      9     12     13     15     16     19
## Predicted   85828 131555 132272 131668 87923 85202 132112 135175 124343
## cvpred      85755 130168 130715 131519 88093 85190 132119 136328 124436
## salary      79750 141500 147765 119250 79800 77700 104800 117150 124750
## CV residual -6005 11332 17050 -12269 -8293 -7490 -27319 -19178 314
##            20     22     23     24     26     29     34     35     36
## Predicted   114266 119892 119311 119345 123125 91042 86318 81534 83048
## cvpred      116045 119125 118551 119231 123987 91604 86346 84474 86239
## salary      137000 102580 93904 113068 106294 77000 80225 80225 77000
## CV residual 20955 -16545 -24647 -6163 -17693 -14604 -6121 -4249 -9239
##            44     51     52     59     60     62     63     67     68
## Predicted   131953 131498 131258 99009 87968 85782 132704 136119 133774
## cvpred      130777 130866 131599 97573 88084 85764 132004 136830 133168
## salary      231545 126621 108875 100135 75044 75243 109785 101000 99418
## CV residual 100768 -4245 -22724 2562 -13040 -10521 -22219 -35830 -33750
##            74     78     84     85     87     95     97     98     100
## Predicted   132306 134833 87388 125724 131907 131668 101286 133751 139432
## cvpred      131395 135021 87510 128501 130786 131519 99875 134545 139617
## salary      92391 193000 88825 122960 152708 123683 95611 129676 106689
## CV residual -39004 57979 1315 -5541 21922 -7836 -4264 -4869 -32928
##            102    105    106    117    118    122    125    128    133
## Predicted   133945 88382 118685 117661 119049 117262 117877 67025 80297
## cvpred      133821 87014 117987 116813 117916 116205 117458 69867 81667
## salary      126933 83850 113543 148500 117515 124309 96614 72500 77500
## CV residual -6888 -3164 -4444 31687 -401 8104 -20844 2633 -4167
##            137    139    143    144    151    153    155    157    158
## Predicted   117512 85570 120586 85293 131839 137053 87297 95673 85691
## cvpred      117529 84129 121049 85172 132172 138021 87528 93417 85781
## salary      108262 73877 106608 89942 128148 111168 92000 113341 88000
## CV residual -9267 -10252 -14441 4770 -4024 -26853 4472 19924 2219
##            161    168    169    170    174    177    178    186    188
## Predicted   85247 130553 99407 134788 131133 99988 100614 133194 134958
## cvpred      85181 130363 98181 135030 130937 98755 99319 132595 135683
## salary      89516 130664 101210 181257 134185 95436 100944 134000 107500
## CV residual 4335 301 3029 46227 3248 -3319 1625 1405 -28183
##            191    192    196    207    209    210    211    214    215
## Predicted   137588 142461 99453 132795 87923 131133 85828 134833 103460

```

```

## cvpred      138604 143146  98173 131986 88093 130937 85755 135021 102884
## salary      180000 133700 113600 162200 91300 163200 91000 126200 118700
## CV residual  41396 -9446  15427 30214 3207 32263 5245 -8821 15816
##              217    221    222    226    227    228    229    230    231
## Predicted    132864 131179 137634 116716 71854 85035 118981 118070 113321
## cvpred      133345 130928 138595 116311 71729 83547 119302 116734 115542
## salary      146000 170000 145200 122400 63900 70000 88175 133900 91000
## CV residual  12655 39072  6605  6089 -7829 -13547 -31127 17166 -24542
##              233    235    236    239    246    251    254    255    259
## Predicted    126836 73551 122465 120837 114733 117581 67116 122576 74086
## cvpred      127383 73458 122742 119627 118013 116143 69849 126782 74040
## salary      148750 69700 81700 77202 90450 109000 77500 116450 73800
## CV residual  21367 -3758 -41042 -42425 -27563 -7143 7651 -10332 -240
##              264    265    273    275    276    277    281    283    285
## Predicted    117217 118469 72389 68767 119288 120131 118070 118127 84990
## cvpred      116213 117343 72311 71587 119929 118392 116734 116036 83556
## salary      126000 99000 73000 78500 93000 107200 136500 57800 88650
## CV residual  9787 -18343  689  6913 -26929 -11192 19766 -58236 5094
##              288    291    292    294    295    298    302    303    310
## Predicted    71809 130035 119220 89042 122544 119516 122419 124423 120233
## cvpred      71738 131565 118569 88258 123413 119884 122751 125107 120431
## salary      85000 174500 168500 104800 107300 148800 127100 170500 166800
## CV residual  13262 42935 49931 16542 -16113 28916 4349 45393 46369
##              318    320    323    329    333    335    336    337    340
## Predicted    132807 131441 99635 98519 130175 100509 130883 137691 142677
## cvpred      131297 131563 98137 96981 133812 102715 129612 137897 143791
## salary      67559 135027 126431 99247 105450 104542 151445 98053 137317
## CV residual -63738 3464 28294 2266 -28362 1827 21833 -39844 -6474
##              342    351    352    359    362    364    365    373    378
## Predicted    126213 136199 131953 119150 115985 86026 117763 118799 72481
## cvpred      129092 134755 130777 122645 119142 84041 116107 119337 72294
## salary      124312 186960 93519 109954 109646 81285 205500 109707 77081
## CV residual -4780 52205 -37258 -12691 -9496 -2756 89393 -9630 4787
##              379    384    387    389    394    395
## Predicted    117535 117809 118104 118514 123091 126039
## cvpred      116151 116098 117413 117334 123307 126165
## salary      150680 105000 139219 119450 150564 101738
## CV residual  34529 -11098 21806 2116 27257 -24427
##
## Sum of squares = 8.24e+10    Mean square = 6.24e+08    n = 132
##
## Overall (Sum over all 132 folds)
##      ms
## 5.23e+08

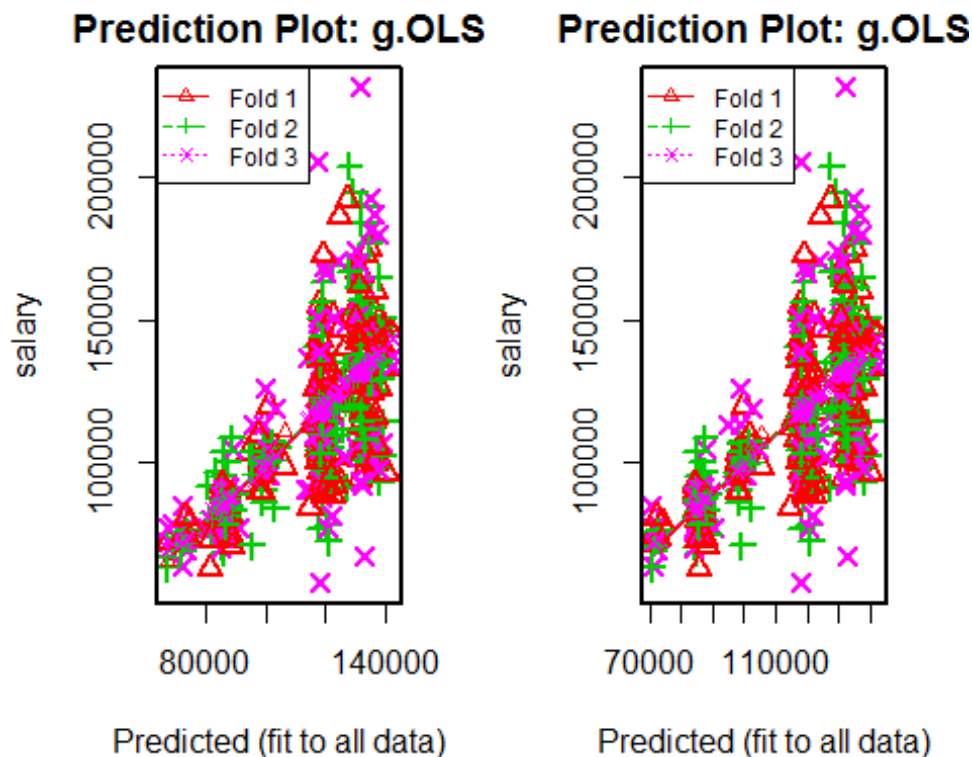
mse.g.OLS.step = CVlm(data = Salaries, form.lm = g.OLS.step, m = 3, main =
"Prediction Plot: g.OLS")

## Analysis of Variance Table
##

```

```
## Response: salary
##              Df    Sum Sq  Mean Sq  F value  Pr(>F)
## rank          2 1.43e+11  7.16e+10   140.79 <2e-16 ***
## discipline     1 1.84e+10  1.84e+10    36.23  4e-09 ***
## yrs.since.phd  1 1.66e+08  1.66e+08     0.33  0.569
## yrs.service     1 2.58e+09  2.58e+09     5.06  0.025 *
## Residuals     391 1.99e+11  5.09e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## Warning in CVlm(data = Salaries, form.lm = g.OLS.step, m = 3, main =
"Prediction Plot: g.OLS"):
##
## As there is >1 explanatory variable, cross-validation
## predicted values for a fold are not a linear function
## of corresponding overall predicted values. Lines that
## are shown for the different folds are approximate
```



```
##
## fold 1
## Observations in test set: 132
##           2      4      7     11     14     18     21     28     30
## Predicted 132727 135128 134736 99808 85443 119264 119336 85617 132264
## cvpred    132137 134404 134919 100698 83888 119060 119755 83955 131957
## salary     173200 115000 175000 119800 78000 103450  89565 82379 118223
## CV residual 41063 -19404  40081  19102 -5888 -15610 -30190 -1576 -13734
##           31     38     40     41     43     45     46     48     49
```

## Predicted	138448	86210	97727	141005	138176	130762	135877	132901	131110
## cvpred	142364	84853	97222	146692	140257	128706	137363	132205	128841
## salary	132261	86373	100938	146500	101299	94384	114778	151768	140096
## CV residual	-10103	1520	3716	-192	-38958	-34322	-22585	19563	11255
##	57	61	64	66	71	73	75	76	77
## Predicted	131947	98204	97843	98204	137797	136109	133667	87221	137320
## cvpred	130500	98074	97267	98074	141444	137453	133169	86580	140592
## salary	117704	90304	103613	100522	126320	100131	113398	73266	150480
## CV residual	-12796	-7770	6346	2448	-15124	-37322	-19771	-13314	9888
##	79	80	88	92	93	94	96	101	103
## Predicted	85501	86629	84490	99215	99215	131863	86513	140198	135832
## cvpred	83910	85682	82184	99801	99801	129134	85637	143711	138013
## salary	86100	84240	88400	105128	105631	166024	84000	133217	153303
## CV residual	2190	-1442	6216	5327	5830	36890	-1637	-10494	15290
##	108	109	110	112	115	121	123	124	127
## Predicted	84233	86906	121764	85231	121572	115967	117500	85579	117732
## cvpred	85466	89840	123366	86521	125292	115111	117040	86656	117130
## salary	82600	81500	131205	82100	105000	115313	97262	62884	155500
## CV residual	-2866	-8340	7839	-4421	-20292	202	-19778	-23772	38370
##	130	131	134	136	140	145	147	149	152
## Predicted	71054	84291	70996	117268	117803	131226	84606	136514	84606
## cvpred	70450	85488	70428	116950	117825	128886	82229	137611	82229
## salary	73000	83001	72500	136000	152664	112696	92000	144651	92000
## CV residual	2550	-2487	2072	19050	34839	-16190	9771	7040	9771
##	163	164	165	166	167	172	182	183	184
## Predicted	105631	84548	84909	137075	130588	130762	136902	97669	133074
## cvpred	110297	82206	83013	139829	128638	128706	139762	97199	132272
## salary	98510	89942	88795	105890	167284	151575	141136	100000	150000
## CV residual	-11787	7736	5782	-33939	38646	22869	1374	2801	17728
##	185	195	197	198	201	203	206	213	218
## Predicted	130994	97602	84606	84606	84606	136527	139936	134344	102222
## cvpred	128796	93838	82229	82229	82229	138283	144943	135433	103637
## salary	101000	90000	92700	92000	92700	160400	96545	128400	105350
## CV residual	-27796	-3838	10471	9771	10471	22117	-48398	-7033	1713
##	219	220	232	234	237	240	242	244	245
## Predicted	101354	131368	85160	125346	116605	122454	117429	116489	114350
## cvpred	103299	130275	85826	130093	115358	126302	116346	115313	111815
## salary	109650	119500	73300	117555	114000	96200	122875	108200	84273
## CV residual	6351	-10775	-12526	-12538	-1358	-30102	6529	-7113	-27542
##	247	249	252	258	260	266	267	269	271
## Predicted	120231	119162	122036	87775	117963	120102	117647	119309	118542
## cvpred	121436	119687	125472	90177	117220	120719	115764	118410	117446
## salary	91100	128800	102000	74000	92550	134800	143940	89650	143250
## CV residual	-30336	9113	-23472	-16177	-24670	14081	28176	-28760	25804
##	279	282	286	297	300	301	309	311	312
## Predicted	119407	122067	85538	116199	88167	118845	72542	122512	118351
## cvpred	120449	124151	84640	115201	89663	118230	73030	126325	119372
## salary	107100	103600	81800	126300	70700	88600	74000	92050	108100
## CV residual	-13349	-20551	-2840	11099	-18963	-29630	970	-34275	-11272
##	314	315	316	319	322	327	330	331	332

```

## Predicted 117184 124058 86152 130588 97727 134808 130994 127256 136064
## cvpred 115584 128926 84830 128638 97222 135614 128796 120007 138103
## salary 100351 146800 84716 134550 95642 124714 134778 192253 116518
## CV residual -15233 17874 -114 5912 -1580 -10900 5982 72246 -21585
## 341 344 345 347 349 350 353 356 357
## Predicted 131368 131457 130085 138710 85083 130749 130691 133017 122285
## cvpred 130275 128976 126442 141132 83081 128034 128011 132250 123569
## salary 106231 162150 150376 142023 80139 144309 142500 145028 88709
## CV residual -24044 33174 23934 891 -2942 16275 14489 12778 -34860
## 360 361 363 366 367 369 374 376 377
## Predicted 73843 117398 116894 119077 118409 119567 118801 117732 71531
## cvpred 74869 117667 115471 118320 119394 119845 118880 117130 71303
## salary 78785 121946 138771 101036 115435 131950 136660 103649 74856
## CV residual 3916 4279 23300 -17284 -3959 12105 17780 -13481 3553
## 382 385 390 392 396 397
## Predicted 118627 119581 124219 122138 121372 72239
## cvpred 118812 120517 128322 124845 123880 72245
## salary 172505 125192 186023 151292 95329 81035
## CV residual 53693 4675 57701 26447 -28551 8790
##
## Sum of squares = 5.68e+10 Mean square = 4.3e+08 n = 132
##
## fold 2
## Observations in test set: 133
## 1 6 10 17 25 27 32 33 37
## Predicted 131239 97553 130704 130285 85837 122904 87163 132322 131413
## cvpred 133355 99182 133259 133435 84232 119387 86904 132421 133763
## salary 139750 97000 129000 101000 74830 134885 79916 117256 155750
## CV residual 6395 -2182 -4259 -32435 -9402 15498 -6988 -15165 21987
## 39 42 47 50 53 54 55 56 58
## Predicted 136803 98538 137699 84432 88825 133925 98377 102307 98204
## cvpred 135983 101491 135767 86289 87328 132708 99957 99909 99550
## salary 125196 93418 98193 70768 74692 106639 103760 83900 90215
## CV residual -10787 -8073 -37574 -15521 -12636 -26069 3803 -16009 -9335
## 65 69 70 72 81 82 83 86 89
## Predicted 85083 130646 127470 132268 139303 131123 131889 131007 131110
## cvpred 86657 133124 134937 136926 136055 133084 133723 132812 134210
## salary 68404 111512 91412 146856 150743 135585 144640 132825 172272
## CV residual -18253 -21612 -43525 9930 14688 2501 10917 13 38062
## 90 91 99 104 107 111 113 114 116
## Predicted 98680 87337 139027 133680 84768 118222 70996 117299 122093
## cvpred 99510 87312 134249 133291 84041 117670 70780 120139 117486
## salary 107008 97032 102235 127512 82099 112429 72500 104279 120806
## CV residual 7498 9720 -32014 -5779 -1942 -5241 1720 -15860 3320
## 119 120 126 129 132 135 138 141 142
## Predicted 71531 71112 120668 117963 117923 121951 117571 86372 85953
## cvpred 70876 71052 122247 119380 122759 119467 117303 84328 84504
## salary 72500 73500 78162 113278 76840 168635 105668 100102 81500
## CV residual 1624 2448 -44085 -6102 -45919 49168 -11635 15774 -3004
## 146 148 150 154 156 159 160 162 171

```

## Predicted	131283	131226	85083	98854	130878	97553	130054	134505	84664
## cvpred	134617	134482	86657	99917	133667	99182	132892	134066	86833
## salary	119015	156938	95079	103994	118971	95408	137167	176500	91227
## CV residual	-15602	22456	8422	4077	-14696	-3774	4275	42434	4394
##	173	175	176	179	180	181	187	189	190
## Predicted	138002	103434	132714	137423	84548	130299	99389	98827	133970
## cvpred	135320	100236	134497	133962	86561	132309	100013	102170	133970
## salary	93164	105000	111751	147349	92000	142467	103750	106300	153750
## CV residual	-42156	4764	-22746	13387	5439	10158	3737	4130	19780
##	193	194	199	200	202	204	205	208	212
## Predicted	131239	98306	132108	139490	131978	130646	137436	130704	131920
## cvpred	133355	100948	135392	135336	136247	133124	132835	133259	136111
## salary	122100	86250	189409	114500	119700	152500	165000	120000	111350
## CV residual	-11255	-14698	54017	-20836	-16547	19376	32165	-13259	-24761
##	216	223	224	225	238	241	243	248	250
## Predicted	132972	101296	138305	118788	70751	71112	117834	117803	127324
## cvpred	132788	99853	134872	120155	71363	71052	120235	117846	118173
## salary	145350	107150	129600	87800	63100	69200	102600	101100	204000
## CV residual	12562	7297	-5272	-32355	-8263	-1852	-17635	-16746	85827
##	253	256	257	261	262	263	268	270	272
## Predicted	126010	85302	116431	88889	117763	119336	117874	118770	129030
## cvpred	118564	84136	118102	87915	121225	119124	116855	116640	119858
## salary	132000	83000	140300	88600	107550	121200	104350	103700	194800
## CV residual	13436	-1136	22198	685	-13675	2076	-12505	-12940	74942
##	274	278	280	284	287	289	290	293	296
## Predicted	72239	118859	117821	118716	117255	117790	71762	131717	119380
## cvpred	71379	119164	121361	121145	118877	118972	71419	119211	120386
## salary	74000	163200	100600	155865	115800	150500	74000	183800	97150
## CV residual	2621	44036	-20761	34720	-3077	31528	2581	64589	-23236
##	299	304	305	306	307	308	313	317	321
## Predicted	120855	120735	118774	118021	71705	118382	121603	99331	131528
## cvpred	121528	116616	121281	119516	71283	119204	118653	99877	134034
## salary	72300	105260	144050	111350	74500	122500	94350	71065	104428
## CV residual	-49228	-11356	22769	-8166	3217	3296	-24303	-28812	-29606
##	324	325	326	328	334	338	339	343	346
## Predicted	135342	130923	86744	131805	138247	130891	134852	131863	132914
## cvpred	133714	134929	87080	135840	134737	132540	134881	135975	132652
## salary	161101	162221	84500	151650	145098	145000	128464	114596	107986
## CV residual	27387	27292	-2580	15810	10361	12460	-6417	-21379	-24666
##	348	354	355	358	368	370	371	372	375
## Predicted	134781	131889	88174						

```

## Sum of squares = 6.88e+10      Mean square = 5.17e+08      n = 133
##
## fold 3
## Observations in test set: 132
##      3      5      8      9     12     13     15     16     19
## Predicted  85083 131502 132268 131355 87163 84432 131774 134647 123974
## cvpred     85473 130108 130667 131397 87808 84903 131991 136155 124364
## salary     79750 141500 147765 119250 79800 77700 104800 117150 124750
## CV residual -5723  11392  17098 -12147 -8008 -7203 -27191 -19005   386
##      20     22     23     24     26     29     34     35     36
## Predicted  118845 119625 119033 118930 122570  90255 85559 85559  87047
## cvpred     117882 119076 118500 119147 123870  91314 86062 86062  87820
## salary     137000 102580  93904 113068 106294  77000 80225 80225  77000
## CV residual  19118 -16496 -24596  -6079 -17576 -14314 -5837 -5837 -10820
##      44     51     52     59     60     62     63     67     68
## Predicted  131863 131283 130834  98262 87221  85025 132526 135703 133596
## cvpred     130708 130767 131450  97217 87802  85479 131914 136684 133079
## salary     231545 126621 108875 100135 75044  75243 109785 101000  99418
## CV residual 100837  -4146 -22575   2918 -12758 -10236 -22129 -35684 -33661
##      74     78     84     85     87     95     97     98     100
## Predicted  132166 134505 86629 130170 131805 131355 100574 133275 139187
## cvpred     131314 134896 87226 130244 130714 131397  99529 134385 139513
## salary     92391 193000 88825 122960 152708 123683  95611 129676 106689
## CV residual -38923  58104  1599  -7284  21994  -7714  -3918  -4709 -32824
##      102    105    106    117    118    122    125    128    133
## Predicted  133667 87557 118382 117371 118845 117010 117500  70938 84233
## cvpred     133708 86703 117930 116759 117882 116159 117383  71489 83221
## salary     126933 83850 113543 148500 117515 124309  96614  72500  77500
## CV residual  -6775 -2853  -4387  31741  -367   8150 -20769  1011 -5721
##      137    139    143    144    151    153    155    157    158
## Predicted  117037 84710 120071 84548 131426 136599 86513  95040 84909
## cvpred     117430 83809 120941 84891 132026 137866 87238  93088 85491
## salary     108262 73877 106608 89942 128148 111168 92000 113341 88000
## CV residual  -9168 -9932 -14333   5051  -3878 -26698  4762  20253  2509
##      161    168    169    170    174    177    178    186    188
## Predicted  84490 130228  98622 134447 130820  99215  99865 133003 134518
## cvpred     84897 130238  97817 134902 130814  98394  98964 132502 135531
## salary     89516 130664 101210 181257 134185  95436 100944 134000 107500
## CV residual  4619   426   3393  46355   3371 -2958   1980   1498 -28031
##      191    192    196    207    209    210    211    214    215
## Predicted  137133 142163  98680 132642 87163 130820 85083 134505 102610
## cvpred     138449 143030  97811 131902 87808 130814 85473 134896 102505
## salary     180000 133700 113600 162200 91300 163200 91000 126200 118700
## CV residual  41551 -9330  15789  30298  3492  32386  5527  -8696 16195
##      217    221    222    226    227    228    229    230    231
## Predicted  132437 130878 137191 116315 70996  84175 118467 117892 117790
## cvpred     133197 130809 138443 116230 71483  83227 119194 116706 117353
## salary     146000 170000 145200 122400 63900  70000  88175 133900  91000
## CV residual  12803  39191  6757   6170 -7583 -13227 -31019  17194 -26353
##      233    235    236    239    246    251    254    255

```



```

## Predicted    126415 72716 122022 120681 119002 117415 71054 126789
## cvpred      127299 73218 122653 119606 119777 116118 71477 128534
## salary      148750 69700 81700 77202 90450 109000 77500 116450
## CV residual  21451 -3518 -40953 -42404 -29327 -7118 6023 -12084
##              259    264    265    273    275    276    277    281
## Predicted    73250.539 116952 118253 71531 72716 118712 120075 117892
## cvpred      73800.261 116165 117306 72065 73218 119806 118394 116706
## salary      73800.000 126000 99000 73000 78500 93000 107200 136500
## CV residual  -0.261 9835 -18306 935 5282 -26806 -11194 19794
##              283    285    288    291    292    294    295    298    302
## Predicted    118110 84117 70938 129463 118917 88105 121978 119002 121964
## cvpred      116047 83233 71489 131445 118512 87920 123294 119777 122658
## salary      57800 88650 85000 174500 168500 104800 107300 148800 127100
## CV residual  -58247 5417 13511 43055 49988 16880 -15994 29023 4442
##              303    310    318    320    323    329    333    335    336
## Predicted    123929 119768 132803 131065 98912 97785 134518 104503 130793
## cvpred      125005 120335 131249 131426 97788 96629 135531 104222 129544
## salary      170500 166800 67559 135027 126431 99247 105450 104542 151445
## CV residual  45495 46465 -63690 3601 28643 2618 -30081 320 21901
##              337    340    342    351    352    359    362    364    365
## Predicted    137409 142292 130646 136242 131863 123452 120302 85289 117647
## cvpred      137784 143654 130832 134719 130708 124417 120918 83750 116094
## salary      98053 137317 124312 186960 93519 109954 109646 81285 205500
## CV residual  -39731 -6337 -6520 52241 -37189 -14463 -11272 -2465 89406
##              373    378    379    384    387    389    394    395
## Predicted    118235 71647 117357 117705 117790 118311 122673 125693
## cvpred      119218 72053 116124 116088 117353 117300 123223 126099
## salary      109707 77081 150680 105000 139219 119450 150564 101738
## CV residual  -9511 5028 34556 -11088 21866 2150 27341 -24361
##
## Sum of squares = 8.28e+10    Mean square = 6.27e+08    n = 132
##
## Overall (Sum over all 132 folds)
##      ms
## 5.25e+08

df = data.frame(mse.g.OLS = attr(mse.g.OLS, "ms"), mse.g.OLS.step =
attr(mse.g.OLS.step, "ms"))
df

##      mse.g.OLS mse.g.OLS.step
## 1 5.23e+08      5.25e+08

```