Homework1.2

Antonio Scognamiglio

09/11/2023

1 Descrizione dataset

Il dataset in esame contiene informazioni sul prezzo di vendita degli immobili a Milano. In particolare contiene 513 osservazioni di 17 variabili, di cui 7 numeriche e 10 dummy. Le variabili sono:

- price_k: prezzo dell'immobile (in 1000€)
- m2: numero di m^2
- rooms: numero di stanze
- bathrooms: numero di bagni
- floor: piano a cui si trova l'immobile
- total_floors: piani totali che compongono l'edificio
- year_of_build: anno di costruzione
- elevator: presenza (1) o assenza (0) di un'ascensore
- heating_centralized: riscaldamento centralizzato (1) oppure autonomo (0)
- Aplus_A: appartenenza (1) o meno (0) alle classi energetiche A+/A
- B_C: appartenenza (1) o meno (0) alle classi energetiche B/C
- D_E: appartenenza (1) o meno (0) alle classi energetiche D/E
- to_be_restructured: immobile da ristrutturare (1, altrimenti 0)
- new: immobile appena costruito (1, altrimenti 0)
- refurbished: immobile rinnovato (1, altrimenti 0)
- heating_air: presenza di riscaldamento ad aria (1, altrimenti 0)
- heating_floor: presenza di riscaldamento nel pavimento (1, altrimenti 0)

2 Modelli di regressione lineare

Si vuole prevedere il prezzo dell'immobile tramite una regressione lineare usando le altre variabili. Impostando il modello si nota che non tutte le covariate sono significative, come si vede dal seguente summary:

```
##
## Call:
## lm(formula = data$price_k ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1259.62  -161.96   -15.22   125.30  1350.12
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3797.4439   704.7959   5.388 1.10e-07 ***
## rooms                -29.0625    23.9939  -1.211 0.226379
## m2                     6.7692     0.4924  13.746  < 2e-16 ***
```

```
## bathrooms             85.7431    30.3003   2.830 0.004847 **
## floor                 30.1443     6.8206   4.420 1.22e-05 ***
## total_floors         -20.6466     6.3178  -3.268 0.001158 **
## year_of_build         -2.1244     0.3631  -5.850 8.91e-09 ***
## elevator             175.4778    44.2829   3.963 8.50e-05 ***
## heating_centralized   58.5779    34.2578   1.710 0.087907 .
## Aplus_A               55.9930    67.9383   0.824 0.410237
## B_C                   72.2368    62.8336   1.150 0.250842
## D_E                   14.8161    33.8980   0.437 0.662244
## to_be_restructured   -22.3575    50.5684  -0.442 0.658593
## new                   93.5877    62.7157   1.492 0.136269
## refurbished           90.7895    34.6475   2.620 0.009053 **
## heating_air          202.8732    62.9335   3.224 0.001349 **
## heating_floor        139.3643    40.7329   3.421 0.000674 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302.9 on 496 degrees of freedom
## Multiple R-squared:  0.7196, Adjusted R-squared:  0.7105
## F-statistic: 79.55 on 16 and 496 DF,  p-value: < 2.2e-16
```

Infatti ben sei covariate non sono significative nemmeno a livello 0.1, il che fa pensare che o ci siano dei problemi di collinearità, oppure semplicemente certi predittori non siano rilevanti nella determinazione del prezzo degli immobili. I VIF del modello scartano però la prima ipotesi, essendo tutti molto bassi:

```
##               rooms                  m2           bathrooms               floor
##              3.8392              4.8314              2.9148              1.3958
##        total_floors       year_of_build            elevator heating_centralized
##              1.4821              1.8350              1.2449              1.2092
##             Aplus_A                 B_C                 D_E  to_be_restructured
##              3.8939              1.2910              1.1904              1.3023
##                 new         refurbished         heating_air       heating_floor
##              3.6550              1.6251              1.1810              1.8157
```

Per selezionare le variabili maggiormente significative verranno usate diverse procedure e verranno poi confrontate in termini di potere predittivo e MSE. A tal fine si dividono i dati in un training set, composto dal 70% delle osservazioni (selezionate casualmente, ma uguali per ogni procedura), e in un test set composto dalle rimanenti. Inoltre, poiché lo scopo è quello di selezionare le covariate più significative, è necessario standardizzare le variabili.

Il modello di regressione lineare con variabili standardizzate fittato al solo training set presenta il seguente summary:

```
##
## Call:
## lm(formula = train$price_k ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1027.32  -175.69   -16.62   124.51  1326.23
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          768.7731    16.3920  46.899  < 2e-16 ***
## rooms                 -6.5200    31.4934  -0.207 0.836112
## m2                   384.8617    34.3670  11.199  < 2e-16 ***
## bathrooms             71.6240    27.0811   2.645 0.008550 **
## floor                 80.1474    18.9248   4.235 2.94e-05 ***
```

```
## total_floors         -47.1785    19.1329  -2.466 0.014158 *
## year_of_build       -129.9516    26.0806  -4.983 9.96e-07 ***
## elevator              65.3001    17.7029   3.689 0.000262 ***
## heating_centralized   16.2703    17.3630   0.937 0.349384
## Aplus_A               51.2641    31.1343   1.647 0.100567
## B_C                   20.4069    19.9738   1.022 0.307650
## D_E                   -6.2822    18.0442  -0.348 0.727935
## to_be_restructured     0.6902    18.3790   0.038 0.970066
## new                   53.9623    30.6207   1.762 0.078912 .
## refurbished           52.8313    21.1534   2.498 0.012974 *
## heating_air           51.8195    17.5711   2.949 0.003405 **
## heating_floor         46.0257    22.9197   2.008 0.045413 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 308.3 on 343 degrees of freedom
## Multiple R-squared:  0.7373, Adjusted R-squared:  0.725
## F-statistic: 60.16 on 16 and 343 DF,  p-value: < 2.2e-16
```

2.1 Procedura stepwise

La procedura stepwise "both" è una combinazione di quelle forward e backward. Essa dipende da due parametri: penter e premoval. La procedura parte da un modello privo di predittori ed a ogni passo aggiunge la variabile più significativa (con pvalue minore di penter) oppure rimuove una variabile il cui pvalue supera premoval, in seguito all'aggiunta di altre variabili. Vengono usati i valori standard dei parametri: penter=0.1 e premoval=0.3.

```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                            Added/                   Adj.
## Step      Variable        Removed     R-Square    R-Square     C(p)         AIC         RMSE
## ----------------------------------------------------------------------------------------------
##    1         m2           addition       0.660       0.659    88.3510    5229.8435    343.4993
##    2      elevator        addition       0.674       0.672    71.7020    5216.4082    336.6844
##    3      bathrooms       addition       0.686       0.683    58.5440    5205.3562    331.1001
##    4        floor         addition       0.692       0.689    52.0400    5199.8207    328.1140
##    5     heating_air      addition       0.698       0.693    46.7920    5195.2712    325.6018
##    6    year_of_build     addition       0.703       0.698    42.2770    5191.2813    323.3613
##    7       Aplus_A        addition       0.719       0.713    22.9220    5172.9161    314.7893
##    8    heating_floor     addition       0.723       0.717    19.0020    5169.0606    312.6841
##    9    total_floors      addition       0.728       0.721    15.5310    5165.5632    310.7487
##   10     refurbished      addition       0.732       0.724    12.2440    5162.1687    308.8708
##   11         new          addition       0.735       0.727     9.4350    5159.1918    307.1835
## ----------------------------------------------------------------------------------------------
```

Dunque la procedura stepwise seleziona le seguenti 11 variabili: m2, elevator, bathrooms, floor, heating_air, year_of_build, Aplus_A, heating_floor, total_floors, refurbished e new. Si nota che le variabili sono state solo aggiunte e mai rimosse.

Il potere predittivo del modello di regressione lineare ottenuto, testato sul test set, è dato da:

```
## [1] 0.6316474
```

Mentre l'MSE è:

```
## [1] 91513.46
```

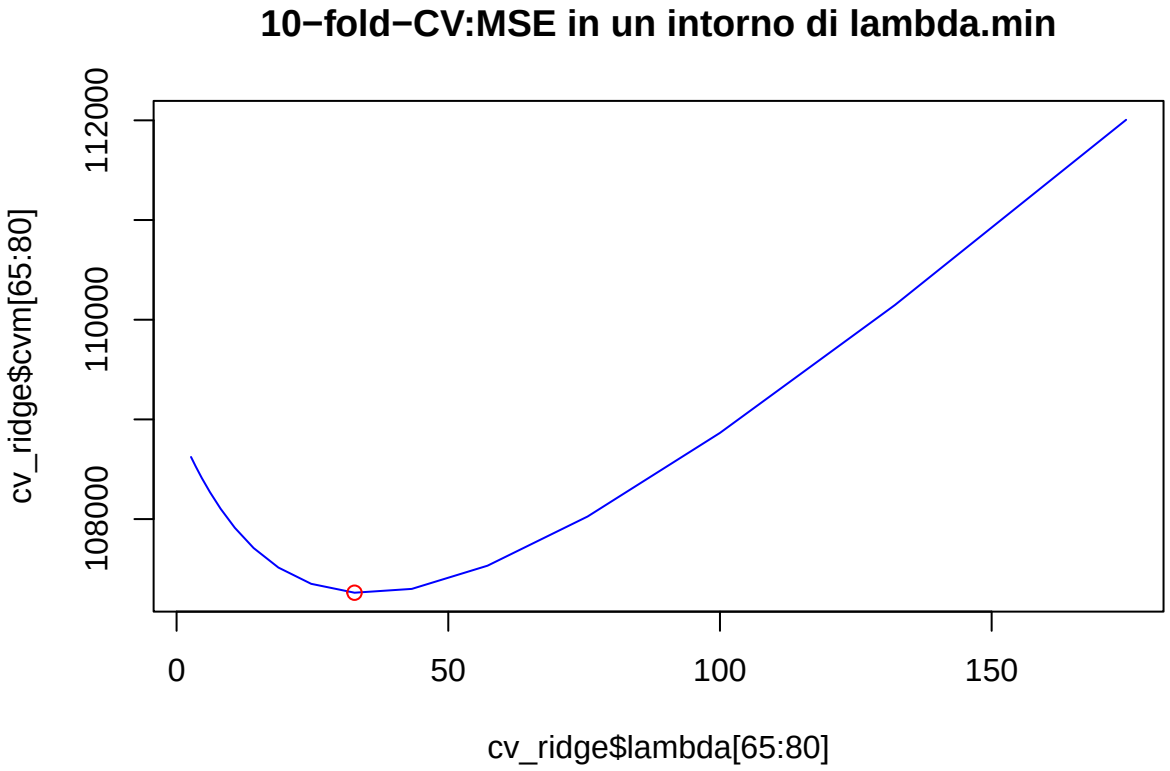
2.2 Ridge regression

Come secondo metodo di selezione delle variabili più significative si usa una ridge regression con la seguente griglia di λ:

```
grid <- 10^seq(10, -2, length=100)
```

Si usa quindi una 10-fold cross validation sul training set per trovare il valore di λ ottimale. La 10-fold CV divide casualmente il training set in 10 folds della stessa dimensione. Ad ogni passo, 9 folds vengono utilizzati per fittare la ridge regression e 1 fold viene utilizzato per calcolare l'MSE del modello trovato (per ogni valore di λ). Infine viene fatta la media dei vari MSE e si seleziona il lambda che la minimizza. Tale λ è dato da:

```
## [1] 32.74549
```



I coefficienti della ridge regression con il valore di λ trovato sono:

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                s1
## (Intercept)          768.430284
## rooms                 37.312757
## m2                   315.392905
## bathrooms             92.569182
## floor                 77.023620
## total_floors         -40.915818
## year_of_build       -119.794747
## elevator              59.634193
## heating_centralized   16.943029
## Aplus_A               46.653021
```

```
## B_C                   22.391211
## D_E                   -6.773579
## to_be_restructured    -3.458450
## new                   45.456719
## refurbished           42.850192
## heating_air           48.522012
## heating_floor         48.226805
```

Dunque la ridge regression suggerisce di costruire un modello di regressione lineare senza le covariate D_E e to_be_restructured. Fittando tale modello al training set si ottiene il seguente summary:

```
##
## Call:
## lm(formula = train$price_k ~ ., data = train[, -c(12, 13)])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1023.3  -171.9   -14.4   124.9  1326.5
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           768.847     16.341  47.050  < 2e-16 ***
## rooms                  -7.537     31.044  -0.243 0.808324
## m2                    384.985     34.184  11.262  < 2e-16 ***
## bathrooms              71.700     26.897   2.666 0.008042 **
## floor                  79.757     18.837   4.234 2.95e-05 ***
## total_floors          -46.856     19.050  -2.460 0.014400 *
## year_of_build        -130.554     25.833  -5.054 7.04e-07 ***
## elevator               64.770     17.582   3.684 0.000267 ***
## heating_centralized    16.810     17.117   0.982 0.326759
## Aplus_A                53.531     30.360   1.763 0.078754 .
## B_C                    21.913     19.443   1.127 0.260516
## new                    52.627     29.768   1.768 0.077963 .
## refurbished            51.239     19.195   2.669 0.007959 **
## heating_air            51.193     17.413   2.940 0.003504 **
## heating_floor          46.672     22.728   2.054 0.040772 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.5 on 345 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7265
## F-statistic: 69.12 on 14 and 345 DF,  p-value: < 2.2e-16
```

Il potere predittivo di tale modello sul test set è:

```
## [1] 0.6369304
```

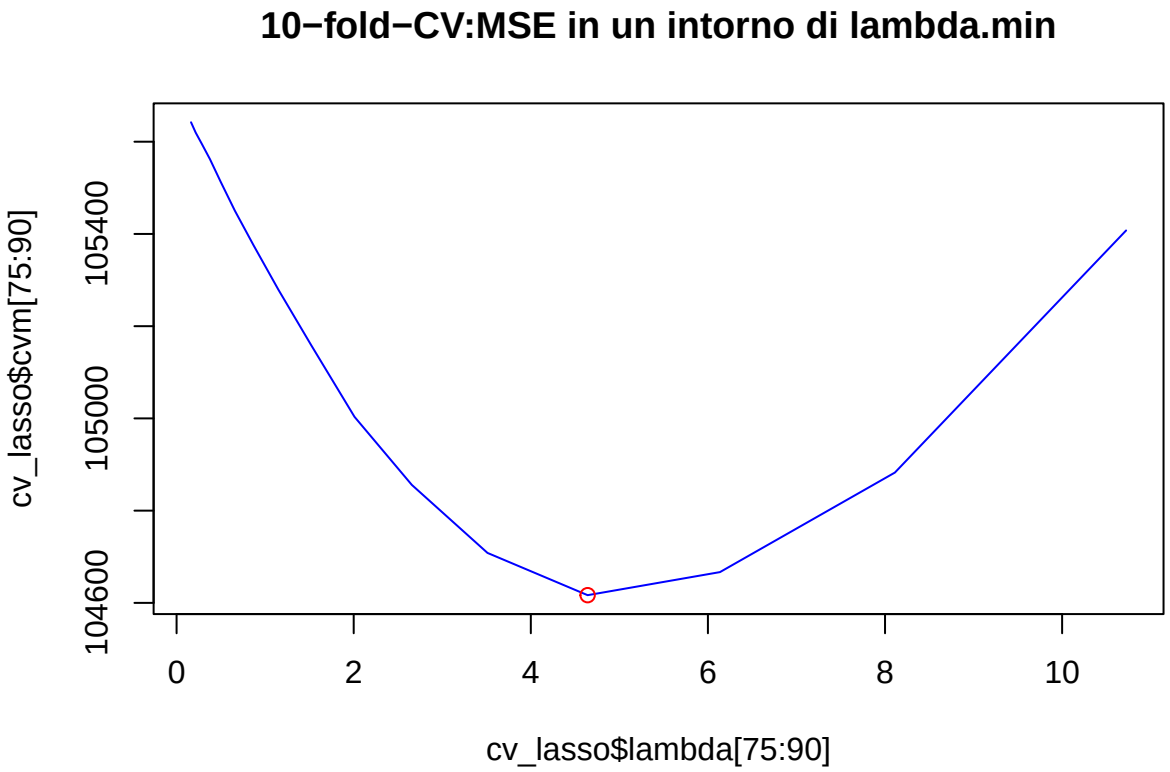
e l'MSE è:

```
## [1] 90200.95
```

2.3 LASSO regression

Come terzo metodo di selezione delle variabili si usa una LASSO regression con la stessa griglia di λ usata per la ridge. Si usa quindi una 10-fold cross validation per trovare il valore di λ ottimale. La procedura è del tutto analoga al caso della ridge regression, cambia solo la funzione di penalizzazione. Il valore di λ ottimale è:

```
## [1] 4.641589
```



I coefficienti della LASSO regression con tale valore di λ sono:

```
## 17 x 1 sparse Matrix of class "dgCMatrix"
##                                s1
## (Intercept)          767.748029
## rooms                  .
## m2                   376.550541
## bathrooms             72.102802
## floor                 70.887800
## total_floors         -36.220631
## year_of_build       -112.611009
## elevator              59.819201
## heating_centralized   10.217421
## Aplus_A               44.621927
## B_C                   17.709368
## D_E                   -1.664252
## to_be_restructured     .
## new                   45.826827
## refurbished           44.197418
## heating_air           46.069850
## heating_floor         42.621951
```

Si nota che le variabili rooms e to_be_restructured hanno coefficiente zero. Si fitta quindi un modello di regressione lineare senza tali variabili:

```
##
```

```
## Call:
## lm(formula = train$price_k ~ ., data = train[, -c(2, 13)])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1025.45  -176.74   -16.02   124.45  1327.27
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          768.708     16.335  47.059  < 2e-16 ***
## m2                   380.154     25.688  14.799  < 2e-16 ***
## bathrooms             70.653     26.551   2.661 0.008155 **
## floor                 80.381     18.837   4.267 2.56e-05 ***
## total_floors         -47.084     19.063  -2.470 0.014001 *
## year_of_build       -130.356     25.826  -5.047 7.26e-07 ***
## elevator              65.212     17.640   3.697 0.000254 ***
## heating_centralized   16.313     17.162   0.951 0.342509
## Aplus_A               51.188     31.038   1.649 0.100020
## B_C                   20.628     19.878   1.038 0.300107
## D_E                   -6.723     17.828  -0.377 0.706341
## new                   54.490     29.700   1.835 0.067413 .
## refurbished           52.890     19.459   2.718 0.006900 **
## heating_air           52.109     17.460   2.984 0.003044 **
## heating_floor         46.333     22.754   2.036 0.042484 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 307.4 on 345 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7266
## F-statistic: 69.14 on 14 and 345 DF,  p-value: < 2.2e-16
```

Il potere predittivo di tale modello sul test set è:

```
## [1] 0.6338095
```

e l'MSE è:

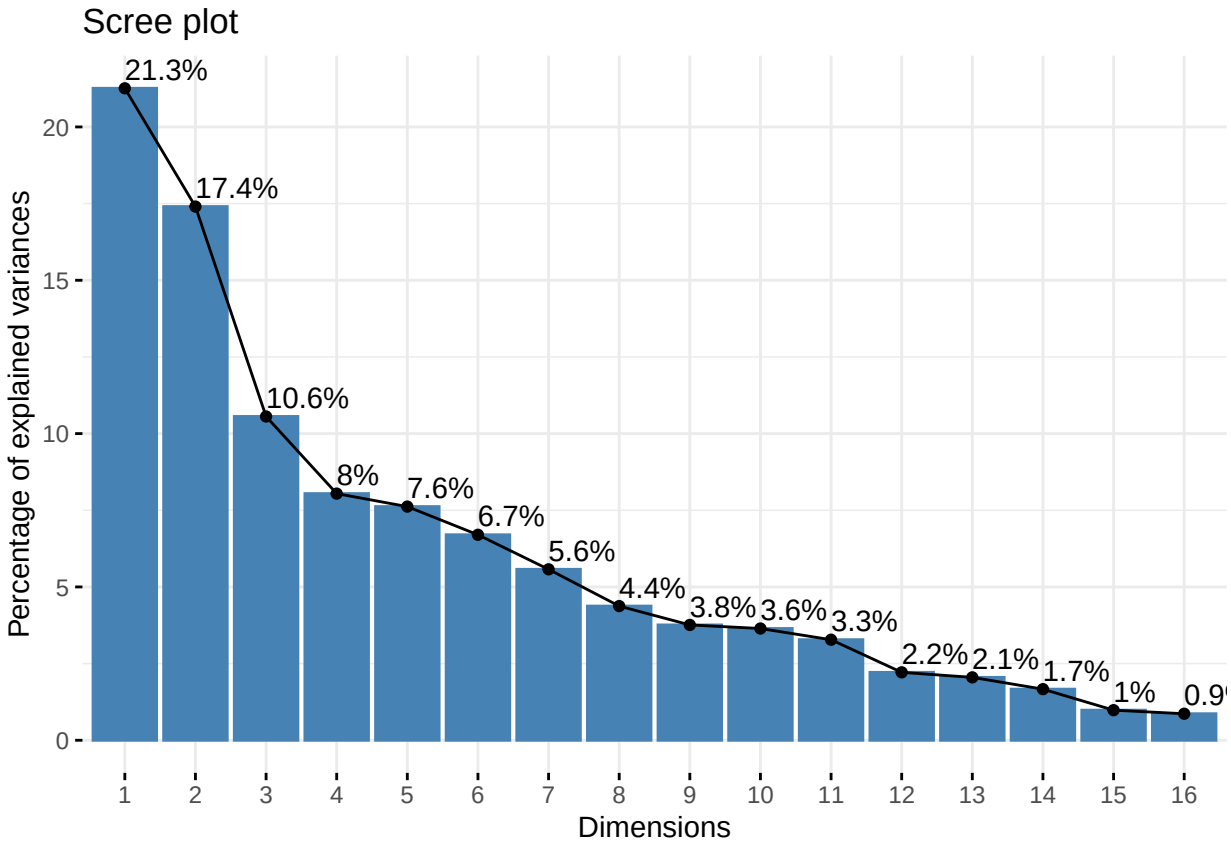
```
## [1] 90976.31
```

2.4 Principal component regression

Come ultimo metodo di selezione delle variabili si usa una principal component regression. Eseguenedo la PCA sul training set, si nota dai seguenti summary e scree plot che le prime 11 componenti principali spiegano il 92.2% della varianza.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     1.8460555 1.6702402 1.3009323 1.13550052 1.10518772
## Proportion of Variance 0.2125976 0.1740311 0.1055792 0.08043473 0.07619756
## Cumulative Proportion  0.2125976 0.3866287 0.4922079 0.57264265 0.64884021
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     1.03638122 0.94531474 0.83751962 0.77647401 0.76433806
## Proportion of Variance 0.06700513 0.05574704 0.04375815 0.03761169 0.03644516
## Cumulative Proportion  0.71584534 0.77159238 0.81535053 0.85296221 0.88940738
```

```
##                            Comp.11    Comp.12    Comp.13   Comp.14     Comp.15
## Standard deviation     0.72500691 0.59586968 0.57346546 0.5168993 0.396700258
## Proportion of Variance 0.03279089 0.02214989 0.02051556 0.0166679 0.009817342
## Cumulative Proportion  0.92219827 0.94434816 0.96486372 0.9815316 0.991348960
##                            Comp.16
## Standard deviation     0.37239143
## Proportion of Variance 0.00865104
## Cumulative Proportion  1.00000000
```



Si imposta quindi un modello di regressione lineare usando le prime 11 componenti. Il summary di questo modello è:

```
##
## Call:
## lm(formula = train_pcr$price_k ~ ., data = train_pcr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -858.07 -203.82  -16.01  155.45 1657.97
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  767.641     17.784  43.165  < 2e-16 ***
## Comp.1        91.256      9.634   9.473  < 2e-16 ***
## Comp.2       257.807     10.648  24.213  < 2e-16 ***
## Comp.3        66.126     13.670   4.837 1.98e-06 ***
## Comp.4       -39.906     15.662  -2.548 0.011263 *
## Comp.5        47.110     16.091   2.928 0.003640 **
```

```
## Comp.6       -17.587     17.160  -1.025 0.306131
## Comp.7       -28.655     18.813  -1.523 0.128630
## Comp.8         7.510     21.234   0.354 0.723787
## Comp.9        88.062     22.904   3.845 0.000143 ***
## Comp.10        2.401     23.267   0.103 0.917857
## Comp.11      -74.643     24.529  -3.043 0.002520 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 337.4 on 348 degrees of freedom
## Multiple R-squared:  0.6807, Adjusted R-squared:  0.6706
## F-statistic: 67.45 on 11 and 348 DF,  p-value: < 2.2e-16
```

Il potere predittivo di tale modello sul test set è:

```
## [1] 0.541784
```

dove prima di poter fare predizioni sul test set è stato necessario proiettarlo lungo le prime 11 componenti principali del training set attraverso il comando:

```
test_pcr <- as.data.frame(predict(x_pca, newdata=test))[, 1:11]
```

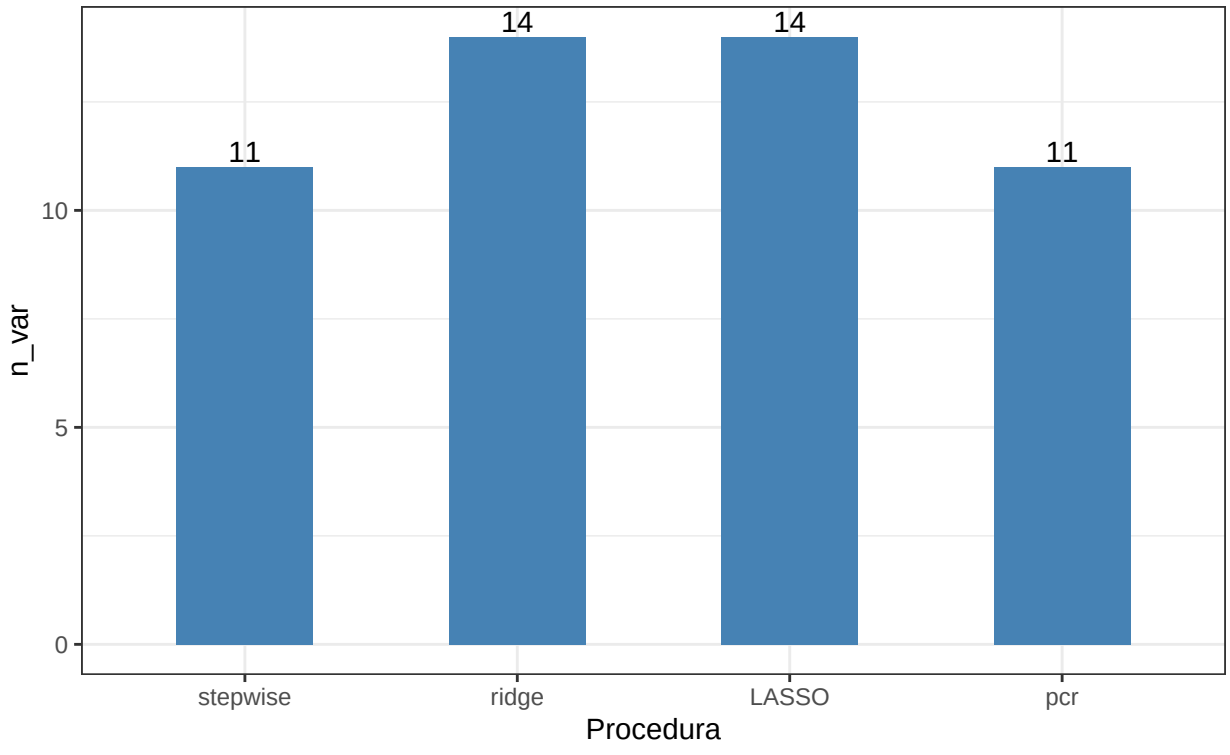
dove `x_pca` è il risultato del comando `princomp` applicato al training set.

L'MSE sul test set è:

```
## [1] 113839.1
```

3 Confronto delle procedure

La seguente figura mostra il numero di variabili selezionate da ogni procedura.



Si osserva che la stepwise e la pcr selezionano un numero minore di covariate rispetto a ridge e LASSO, dunque producono modelli di regressione lineari più semplici. D'altra parte le figure seguenti mostrano che la ridge regression produce il modello con potere predittivo maggiore e MSE minore, mentre la pcr fornisce un modello deludente. Le scarse prestazioni della pcr sono probabilmente dovute alla presenza di ben 10 dummy variables. Il modello migliore dal punto di vista della semplicità è quello ottenuto tramite la procedura stepwise, perché è quello con minor numero di predittori ma con potere predittivo e MSE buoni. Mentre dal punto di vista dell'accuratezza il modello migliore è quello ottenuto tramite la ridge regression.