Simple template for R Markdown

for Advanced Methods for Regression and Classification

Prof. Peter Filzmoser

01.10.2024

```r
data(College,package="ISLR")
str(College)
```

```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

```r
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad        Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books          Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
```

```
##   Median :4200    Median : 500.0    Median :1200    Median : 75.00
##   Mean   :4358    Mean   : 549.4    Mean   :1341    Mean   : 72.66
##   3rd Qu.:5050    3rd Qu.: 600.0    3rd Qu.:1700    3rd Qu.: 85.00
##   Max.   :8124    Max.   :2340.0    Max.   :6800    Max.   :103.00
##     Terminal        S.F.Ratio       perc.alumni        Expend
##   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##   Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##   Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##   Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##     Grad.Rate
##   Min.   : 10.00
##   1st Qu.: 53.00
##   Median : 65.00
##   Mean   : 65.46
##   3rd Qu.: 78.00
##   Max.   :118.00
```

Our goal is to find a linear regression model which allows to predict the variable Apps, i.e. the number of applications received, using the remaining variables except of the variables Accept and Enroll.

For the following tasks, split the data randomly into training and test data (about 2/3 and 1/3), build the model with the training data, and evaluate the model using the RMSE as a criterion.

split the data into training and test data:

```r
set.seed(123)
n <- nrow(College)
train <- sample(1:n, n/3)
test <- -train
train.data <- College[train,]
test.data <- College[test,]
```

creating the model

```r
model <- lm(Apps ~ . - Accept - Enroll, data=train.data)
summary(model)
```

```
##
## Call:
## lm(formula = Apps ~ . - Accept - Enroll, data = train.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3588.2  -593.5   -65.8   563.9  5865.3
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.419e+03  9.156e+02  -1.549  0.12259
## PrivateYes  -5.619e+02  3.150e+02  -1.784  0.07569 .
## Top10perc    2.568e+01  1.105e+01   2.324  0.02093 *
## Top25perc   -1.627e+01  9.200e+00  -1.769  0.07817 .
## F.Undergrad  6.115e-01  2.864e-02  21.351  < 2e-16 ***
## P.Undergrad -1.957e-01  7.431e-02  -2.633  0.00900 **
## Outstate     8.123e-02  3.879e-02   2.094  0.03730 *
## Room.Board   3.788e-01  9.180e-02   4.126 5.07e-05 ***
```

```
## Books          2.474e-01  4.881e-01   0.507  0.61274
## Personal      -1.658e-01  1.265e-01  -1.311  0.19105
## PhD           -4.674e+00  1.105e+01  -0.423  0.67280
## Terminal      -1.590e+01  1.197e+01  -1.329  0.18522
## S.F.Ratio      2.238e+01  3.025e+01   0.740  0.46017
## perc.alumni  -2.019e+01  8.397e+00  -2.404  0.01697 *
## Expend         9.331e-02  2.811e-02   3.320  0.00104 **
## Grad.Rate      1.868e+01  6.886e+00   2.713  0.00714 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1224 on 243 degrees of freedom
## Multiple R-squared:  0.8394, Adjusted R-squared:  0.8295
## F-statistic: 84.67 on 15 and 243 DF,  p-value: < 2.2e-16
```

predict the number of applications for the test data:

```
pred <- predict(model, newdata=test.data)
```

calculate the RMSE:

```
rmse <- sqrt(mean((test.data$Apps - pred)^2))
rmse
```

```
## [1] 2258.293
```