Solution to multiple regression example

Chaeeun Shin

02/27/2024

# Install packages

library(MASS)  
library(tidyverse)

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'stringr' was built under R version 4.2.3

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.4.4 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.3 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::select() masks MASS::select()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(caret)

## Loading required package: lattice  
##   
## Attaching package: 'caret'  
##   
## The following object is masked from 'package:purrr':  
##   
## lift

library(glmnet)

## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## Loaded glmnet 4.1-8

# Question 1

setwd("/Users/chaeeunshin/Desktop/AMS 580")  
data1 <- read.csv('math.csv', sep = ';')  
data1 <- subset(data1, select = - c(G1,G2))  
set.seed(123)  
training.samples <- data1$G3 %>% createDataPartition(p = 0.75, list = FALSE)  
train.data <- data1[training.samples, ]  
test.data <- data1[-training.samples, ]

# Question 2

fit <- lm(G3~., data = train.data)  
fit\_step <- stepAIC(fit, k = log(nrow(train.data)), trace = 1)

## Start: AIC=1034.07  
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +   
## Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +   
## failures + schoolsup + famsup + paid + activities + nursery +   
## higher + internet + romantic + famrel + freetime + goout +   
## Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - Fjob 4 41.387 4499.2 1014.0  
## - Mjob 4 97.705 4555.6 1017.7  
## - reason 3 27.452 4485.3 1018.8  
## - guardian 2 4.311 4462.2 1023.0  
## - traveltime 1 0.045 4457.9 1028.4  
## - Pstatus 1 0.082 4457.9 1028.4  
## - activities 1 5.276 4463.1 1028.7  
## - school 1 7.353 4465.2 1028.9  
## - Fedu 1 8.054 4465.9 1028.9  
## - address 1 9.158 4467.0 1029.0  
## - paid 1 11.233 4469.1 1029.1  
## - freetime 1 15.892 4473.7 1029.4  
## - internet 1 15.967 4473.8 1029.4  
## - Dalc 1 19.126 4477.0 1029.7  
## - higher 1 19.372 4477.2 1029.7  
## - nursery 1 20.591 4478.4 1029.8  
## - Walc 1 21.954 4479.8 1029.8  
## - goout 1 26.597 4484.4 1030.2  
## - Medu 1 32.380 4490.2 1030.5  
## - health 1 34.929 4492.8 1030.7  
## - famrel 1 34.994 4492.8 1030.7  
## - studytime 1 37.209 4495.1 1030.8  
## - famsup 1 41.951 4499.8 1031.2  
## - sex 1 43.454 4501.3 1031.3  
## - age 1 51.687 4509.5 1031.8  
## - schoolsup 1 53.305 4511.2 1031.9  
## - famsize 1 66.075 4523.9 1032.8  
## - absences 1 74.544 4532.4 1033.3  
## <none> 4457.8 1034.1  
## - romantic 1 109.816 4567.7 1035.6  
## - failures 1 254.495 4712.3 1044.9  
##   
## Step: AIC=1014.04  
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +   
## Fedu + Mjob + reason + guardian + traveltime + studytime +   
## failures + schoolsup + famsup + paid + activities + nursery +   
## higher + internet + romantic + famrel + freetime + goout +   
## Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - Mjob 4 90.713 4589.9 997.20  
## - reason 3 37.315 4536.5 999.41  
## - guardian 2 6.620 4505.9 1003.08  
## - traveltime 1 0.004 4499.2 1008.34  
## - Pstatus 1 0.014 4499.2 1008.34  
## - Fedu 1 1.541 4500.8 1008.44  
## - activities 1 8.044 4507.3 1008.87  
## - school 1 8.559 4507.8 1008.91  
## - paid 1 8.889 4508.1 1008.93  
## - address 1 11.199 4510.4 1009.08  
## - internet 1 11.208 4510.4 1009.08  
## - Dalc 1 12.244 4511.5 1009.15  
## - Walc 1 12.458 4511.7 1009.16  
## - freetime 1 13.421 4512.7 1009.23  
## - higher 1 17.188 4516.4 1009.48  
## - nursery 1 19.765 4519.0 1009.65  
## - goout 1 22.359 4521.6 1009.82  
## - Medu 1 29.220 4528.5 1010.27  
## - famrel 1 32.787 4532.0 1010.50  
## - health 1 32.908 4532.1 1010.51  
## - sex 1 39.533 4538.8 1010.95  
## - studytime 1 40.155 4539.4 1010.99  
## - famsup 1 40.789 4540.0 1011.03  
## - schoolsup 1 44.419 4543.7 1011.27  
## - age 1 50.844 4550.1 1011.69  
## - famsize 1 58.276 4557.5 1012.17  
## - absences 1 70.580 4569.8 1012.98  
## <none> 4499.2 1014.04  
## - romantic 1 101.250 4600.5 1014.97  
## - failures 1 247.187 4746.4 1024.28  
##   
## Step: AIC=997.2  
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +   
## Fedu + reason + guardian + traveltime + studytime + failures +   
## schoolsup + famsup + paid + activities + nursery + higher +   
## internet + romantic + famrel + freetime + goout + Dalc +   
## Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - reason 3 67.742 4657.7 984.47  
## - guardian 2 9.081 4599.0 986.39  
## - traveltime 1 0.127 4590.1 991.51  
## - Pstatus 1 0.705 4590.7 991.55  
## - Fedu 1 2.664 4592.6 991.67  
## - school 1 7.446 4597.4 991.98  
## - activities 1 8.718 4598.7 992.07  
## - paid 1 9.325 4599.3 992.10  
## - freetime 1 11.680 4601.6 992.26  
## - goout 1 13.269 4603.2 992.36  
## - address 1 13.898 4603.8 992.40  
## - Walc 1 16.346 4606.3 992.56  
## - internet 1 16.583 4606.5 992.57  
## - nursery 1 19.575 4609.5 992.77  
## - Dalc 1 22.564 4612.5 992.96  
## - higher 1 23.022 4613.0 992.99  
## - health 1 27.256 4617.2 993.26  
## - famrel 1 32.023 4622.0 993.57  
## - sex 1 32.258 4622.2 993.59  
## - famsup 1 35.123 4625.1 993.77  
## - studytime 1 36.898 4626.8 993.89  
## - age 1 47.197 4637.1 994.55  
## - schoolsup 1 50.192 4640.1 994.74  
## - Medu 1 51.533 4641.5 994.83  
## - absences 1 62.739 4652.7 995.55  
## - famsize 1 65.170 4655.1 995.70  
## <none> 4589.9 997.20  
## - romantic 1 106.790 4696.7 998.35  
## - failures 1 230.587 4820.5 1006.11  
##   
## Step: AIC=984.47  
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +   
## Fedu + guardian + traveltime + studytime + failures + schoolsup +   
## famsup + paid + activities + nursery + higher + internet +   
## romantic + famrel + freetime + goout + Dalc + Walc + health +   
## absences  
##   
## Df Sum of Sq RSS AIC  
## - guardian 2 9.785 4667.5 973.70  
## - Pstatus 1 0.146 4657.8 978.78  
## - traveltime 1 0.368 4658.1 978.80  
## - Fedu 1 3.678 4661.4 979.01  
## - school 1 6.475 4664.2 979.19  
## - activities 1 6.959 4664.6 979.22  
## - address 1 10.186 4667.9 979.43  
## - paid 1 12.148 4669.8 979.55  
## - higher 1 13.122 4670.8 979.61  
## - internet 1 14.679 4672.4 979.71  
## - freetime 1 15.091 4672.8 979.74  
## - goout 1 18.097 4675.8 979.93  
## - Dalc 1 18.681 4676.4 979.97  
## - Walc 1 19.883 4677.6 980.04  
## - nursery 1 21.917 4679.6 980.17  
## - sex 1 22.507 4680.2 980.21  
## - famrel 1 31.653 4689.3 980.79  
## - studytime 1 35.681 4693.4 981.05  
## - health 1 36.394 4694.1 981.09  
## - famsup 1 40.659 4698.3 981.36  
## - schoolsup 1 45.873 4703.6 981.70  
## - age 1 47.479 4705.2 981.80  
## - famsize 1 62.488 4720.2 982.75  
## - absences 1 64.142 4721.8 982.85  
## - Medu 1 66.728 4724.4 983.01  
## <none> 4657.7 984.47  
## - romantic 1 103.281 4761.0 985.31  
## - failures 1 239.621 4897.3 993.72  
##   
## Step: AIC=973.7  
## G3 ~ school + sex + age + address + famsize + Pstatus + Medu +   
## Fedu + traveltime + studytime + failures + schoolsup + famsup +   
## paid + activities + nursery + higher + internet + romantic +   
## famrel + freetime + goout + Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - Pstatus 1 0.064 4667.5 968.01  
## - traveltime 1 0.741 4668.2 968.05  
## - Fedu 1 3.881 4671.4 968.25  
## - activities 1 6.633 4674.1 968.43  
## - school 1 8.291 4675.8 968.53  
## - address 1 9.709 4677.2 968.63  
## - higher 1 10.318 4677.8 968.66  
## - paid 1 10.472 4677.9 968.67  
## - freetime 1 13.671 4681.1 968.88  
## - internet 1 16.231 4683.7 969.04  
## - nursery 1 17.863 4685.3 969.14  
## - goout 1 17.994 4685.5 969.15  
## - Dalc 1 18.655 4686.1 969.19  
## - Walc 1 22.208 4689.7 969.42  
## - sex 1 23.557 4691.0 969.51  
## - famrel 1 29.950 4697.4 969.91  
## - health 1 32.610 4700.1 970.08  
## - studytime 1 36.001 4703.5 970.30  
## - famsup 1 41.609 4709.1 970.65  
## - schoolsup 1 45.532 4713.0 970.90  
## - absences 1 57.759 4725.2 971.67  
## - famsize 1 61.606 4729.1 971.91  
## - Medu 1 66.154 4733.6 972.20  
## - age 1 66.396 4733.9 972.22  
## <none> 4667.5 973.70  
## - romantic 1 106.326 4773.8 974.72  
## - failures 1 288.196 4955.7 985.86  
##   
## Step: AIC=968.01  
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + traveltime +   
## studytime + failures + schoolsup + famsup + paid + activities +   
## nursery + higher + internet + romantic + famrel + freetime +   
## goout + Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - traveltime 1 0.777 4668.3 962.36  
## - Fedu 1 3.827 4671.4 962.56  
## - activities 1 6.886 4674.4 962.75  
## - school 1 8.280 4675.8 962.84  
## - address 1 9.697 4677.2 962.93  
## - higher 1 10.333 4677.9 962.97  
## - paid 1 10.409 4677.9 962.98  
## - freetime 1 13.673 4681.2 963.18  
## - internet 1 16.192 4683.7 963.34  
## - nursery 1 17.855 4685.4 963.45  
## - goout 1 18.085 4685.6 963.47  
## - Dalc 1 18.595 4686.1 963.50  
## - Walc 1 22.208 4689.7 963.73  
## - sex 1 23.513 4691.1 963.81  
## - famrel 1 30.221 4697.8 964.24  
## - health 1 32.822 4700.4 964.40  
## - studytime 1 36.093 4703.6 964.61  
## - famsup 1 41.546 4709.1 964.95  
## - schoolsup 1 45.496 4713.0 965.20  
## - absences 1 59.337 4726.9 966.08  
## - famsize 1 63.213 4730.8 966.32  
## - Medu 1 66.566 4734.1 966.53  
## - age 1 66.584 4734.1 966.53  
## <none> 4667.5 968.01  
## - romantic 1 106.274 4773.8 969.02  
## - failures 1 288.151 4955.7 980.16  
##   
## Step: AIC=962.36  
## G3 ~ school + sex + age + address + famsize + Medu + Fedu + studytime +   
## failures + schoolsup + famsup + paid + activities + nursery +   
## higher + internet + romantic + famrel + freetime + goout +   
## Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - Fedu 1 3.505 4671.8 956.89  
## - activities 1 7.111 4675.4 957.12  
## - school 1 7.648 4676.0 957.15  
## - higher 1 10.593 4678.9 957.34  
## - paid 1 10.837 4679.2 957.36  
## - address 1 12.423 4680.7 957.46  
## - freetime 1 14.508 4682.8 957.59  
## - internet 1 16.559 4684.9 957.72  
## - nursery 1 18.223 4686.5 957.83  
## - goout 1 18.374 4686.7 957.84  
## - Dalc 1 19.867 4688.2 957.93  
## - Walc 1 22.344 4690.7 958.09  
## - sex 1 23.459 4691.8 958.16  
## - famrel 1 29.829 4698.1 958.56  
## - health 1 32.534 4700.8 958.73  
## - studytime 1 37.255 4705.6 959.03  
## - famsup 1 43.715 4712.0 959.44  
## - schoolsup 1 46.150 4714.5 959.60  
## - absences 1 59.413 4727.7 960.43  
## - famsize 1 62.691 4731.0 960.64  
## - age 1 65.898 4734.2 960.84  
## - Medu 1 67.769 4736.1 960.96  
## <none> 4668.3 962.36  
## - romantic 1 108.614 4776.9 963.52  
## - failures 1 287.427 4955.7 974.47  
##   
## Step: AIC=956.89  
## G3 ~ school + sex + age + address + famsize + Medu + studytime +   
## failures + schoolsup + famsup + paid + activities + nursery +   
## higher + internet + romantic + famrel + freetime + goout +   
## Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - school 1 7.228 4679.0 951.65  
## - activities 1 7.979 4679.8 951.70  
## - higher 1 10.431 4682.3 951.86  
## - paid 1 12.085 4683.9 951.96  
## - address 1 13.386 4685.2 952.04  
## - freetime 1 15.227 4687.0 952.16  
## - internet 1 16.041 4687.9 952.21  
## - goout 1 19.358 4691.2 952.42  
## - Dalc 1 19.553 4691.4 952.44  
## - nursery 1 19.828 4691.6 952.45  
## - Walc 1 22.149 4694.0 952.60  
## - sex 1 24.876 4696.7 952.77  
## - famrel 1 29.988 4701.8 953.10  
## - health 1 34.198 4706.0 953.37  
## - studytime 1 39.144 4711.0 953.68  
## - schoolsup 1 46.635 4718.5 954.15  
## - famsup 1 47.391 4719.2 954.20  
## - absences 1 61.972 4733.8 955.12  
## - famsize 1 64.452 4736.3 955.27  
## - age 1 64.557 4736.4 955.28  
## - Medu 1 77.401 4749.2 956.09  
## <none> 4671.8 956.89  
## - romantic 1 109.334 4781.2 958.09  
## - failures 1 284.237 4956.1 968.79  
##   
## Step: AIC=951.65  
## G3 ~ sex + age + address + famsize + Medu + studytime + failures +   
## schoolsup + famsup + paid + activities + nursery + higher +   
## internet + romantic + famrel + freetime + goout + Dalc +   
## Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - activities 1 9.353 4688.4 946.55  
## - address 1 9.530 4688.6 946.56  
## - higher 1 10.887 4689.9 946.65  
## - paid 1 12.918 4692.0 946.78  
## - internet 1 16.202 4695.2 946.99  
## - freetime 1 16.364 4695.4 947.00  
## - Dalc 1 17.739 4696.8 947.08  
## - goout 1 19.874 4698.9 947.22  
## - nursery 1 20.514 4699.6 947.26  
## - Walc 1 20.767 4699.8 947.28  
## - sex 1 23.336 4702.4 947.44  
## - famrel 1 28.011 4707.1 947.73  
## - studytime 1 35.392 4714.4 948.20  
## - health 1 35.418 4714.5 948.20  
## - schoolsup 1 47.592 4726.6 948.97  
## - famsup 1 53.342 4732.4 949.33  
## - absences 1 56.334 4735.4 949.52  
## - age 1 57.329 4736.4 949.58  
## - famsize 1 67.750 4746.8 950.24  
## - Medu 1 77.169 4756.2 950.83  
## <none> 4679.0 951.65  
## - romantic 1 108.686 4787.7 952.80  
## - failures 1 287.868 4966.9 963.75  
##   
## Step: AIC=946.55  
## G3 ~ sex + age + address + famsize + Medu + studytime + failures +   
## schoolsup + famsup + paid + nursery + higher + internet +   
## romantic + famrel + freetime + goout + Dalc + Walc + health +   
## absences  
##   
## Df Sum of Sq RSS AIC  
## - higher 1 9.674 4698.1 941.47  
## - address 1 10.791 4699.2 941.54  
## - paid 1 13.422 4701.8 941.71  
## - freetime 1 14.524 4702.9 941.78  
## - internet 1 15.410 4703.8 941.83  
## - Dalc 1 16.596 4705.0 941.91  
## - nursery 1 19.731 4708.1 942.10  
## - Walc 1 20.576 4709.0 942.16  
## - sex 1 21.070 4709.5 942.19  
## - goout 1 21.959 4710.4 942.25  
## - famrel 1 29.880 4718.3 942.75  
## - studytime 1 30.073 4718.5 942.76  
## - health 1 37.868 4726.3 943.25  
## - famsup 1 50.546 4738.9 944.05  
## - schoolsup 1 50.669 4739.1 944.06  
## - age 1 54.697 4743.1 944.31  
## - absences 1 58.566 4747.0 944.55  
## - famsize 1 64.537 4752.9 944.93  
## - Medu 1 71.932 4760.3 945.39  
## <none> 4688.4 946.55  
## - romantic 1 108.989 4797.4 947.70  
## - failures 1 292.019 4980.4 958.86  
##   
## Step: AIC=941.47  
## G3 ~ sex + age + address + famsize + Medu + studytime + failures +   
## schoolsup + famsup + paid + nursery + internet + romantic +   
## famrel + freetime + goout + Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - address 1 10.177 4708.3 936.42  
## - internet 1 14.050 4712.1 936.66  
## - freetime 1 15.496 4713.6 936.75  
## - paid 1 16.179 4714.3 936.80  
## - Dalc 1 16.243 4714.3 936.80  
## - sex 1 17.353 4715.4 936.87  
## - Walc 1 19.344 4717.4 937.00  
## - goout 1 20.839 4718.9 937.09  
## - nursery 1 21.074 4719.1 937.10  
## - famrel 1 28.753 4726.8 937.59  
## - studytime 1 32.684 4730.8 937.84  
## - health 1 35.971 4734.0 938.04  
## - famsup 1 49.822 4747.9 938.91  
## - schoolsup 1 50.848 4748.9 938.98  
## - absences 1 55.275 4753.3 939.26  
## - age 1 61.064 4759.1 939.62  
## - famsize 1 67.757 4765.8 940.04  
## - Medu 1 78.972 4777.0 940.74  
## <none> 4698.1 941.47  
## - romantic 1 116.967 4815.0 943.10  
## - failures 1 313.684 5011.8 955.03  
##   
## Step: AIC=936.42  
## G3 ~ sex + age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + paid + nursery + internet + romantic + famrel +   
## freetime + goout + Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - paid 1 15.44 4723.7 931.69  
## - freetime 1 15.75 4724.0 931.71  
## - sex 1 15.76 4724.0 931.71  
## - Dalc 1 17.11 4725.4 931.80  
## - Walc 1 17.31 4725.6 931.81  
## - goout 1 18.05 4726.3 931.86  
## - nursery 1 20.25 4728.5 932.00  
## - internet 1 21.33 4729.6 932.07  
## - famrel 1 28.37 4736.6 932.51  
## - studytime 1 29.42 4737.7 932.57  
## - health 1 35.00 4743.3 932.93  
## - famsup 1 48.53 4756.8 933.77  
## - schoolsup 1 51.08 4759.3 933.93  
## - absences 1 53.39 4761.6 934.08  
## - age 1 67.83 4776.1 934.98  
## - famsize 1 74.73 4783.0 935.41  
## - Medu 1 81.13 4789.4 935.81  
## <none> 4708.3 936.42  
## - romantic 1 117.73 4826.0 938.08  
## - failures 1 319.20 5027.5 950.27  
##   
## Step: AIC=931.69  
## G3 ~ sex + age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + nursery + internet + romantic + famrel + freetime +   
## goout + Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - sex 1 13.50 4737.2 926.85  
## - freetime 1 14.23 4737.9 926.89  
## - Dalc 1 14.61 4738.3 926.92  
## - goout 1 18.04 4741.7 927.13  
## - nursery 1 18.17 4741.9 927.14  
## - Walc 1 19.51 4743.2 927.23  
## - internet 1 25.80 4749.5 927.62  
## - famrel 1 30.07 4753.8 927.89  
## - studytime 1 34.14 4757.8 928.14  
## - famsup 1 36.61 4760.3 928.30  
## - health 1 37.50 4761.2 928.35  
## - absences 1 53.67 4777.4 929.36  
## - schoolsup 1 57.98 4781.7 929.63  
## - age 1 68.81 4792.5 930.31  
## - famsize 1 75.07 4798.8 930.70  
## - Medu 1 82.27 4806.0 931.14  
## <none> 4723.7 931.69  
## - romantic 1 122.05 4845.7 933.60  
## - failures 1 342.67 5066.4 946.87  
##   
## Step: AIC=926.85  
## G3 ~ age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + nursery + internet + romantic + famrel + freetime +   
## goout + Dalc + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - Dalc 1 11.94 4749.1 921.90  
## - nursery 1 18.16 4755.4 922.29  
## - freetime 1 18.75 4755.9 922.33  
## - goout 1 19.16 4756.4 922.35  
## - Walc 1 24.59 4761.8 922.69  
## - studytime 1 26.22 4763.4 922.80  
## - internet 1 27.28 4764.5 922.86  
## - famrel 1 34.17 4771.4 923.29  
## - health 1 35.80 4773.0 923.39  
## - famsup 1 41.93 4779.1 923.78  
## - absences 1 51.31 4788.5 924.36  
## - schoolsup 1 67.16 4804.4 925.35  
## - age 1 72.51 4809.7 925.68  
## - famsize 1 80.03 4817.2 926.14  
## <none> 4737.2 926.85  
## - Medu 1 95.46 4832.7 927.10  
## - romantic 1 139.96 4877.2 929.83  
## - failures 1 336.01 5073.2 941.57  
##   
## Step: AIC=921.9  
## G3 ~ age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + nursery + internet + romantic + famrel + freetime +   
## goout + Walc + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - Walc 1 13.09 4762.2 917.02  
## - freetime 1 14.80 4763.9 917.13  
## - nursery 1 16.53 4765.7 917.24  
## - goout 1 17.72 4766.8 917.31  
## - studytime 1 26.15 4775.3 917.84  
## - internet 1 26.36 4775.5 917.85  
## - health 1 36.47 4785.6 918.48  
## - famrel 1 37.07 4786.2 918.52  
## - famsup 1 43.91 4793.0 918.95  
## - absences 1 50.61 4799.7 919.36  
## - schoolsup 1 73.30 4822.4 920.77  
## - famsize 1 77.91 4827.0 921.05  
## - age 1 79.69 4828.8 921.16  
## - Medu 1 91.26 4840.4 921.88  
## <none> 4749.1 921.90  
## - romantic 1 140.73 4889.9 924.91  
## - failures 1 343.43 5092.6 937.01  
##   
## Step: AIC=917.02  
## G3 ~ age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + nursery + internet + romantic + famrel + freetime +   
## goout + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - goout 1 8.82 4771.0 911.88  
## - freetime 1 15.16 4777.4 912.27  
## - studytime 1 20.34 4782.6 912.60  
## - nursery 1 20.45 4782.7 912.60  
## - internet 1 26.73 4788.9 912.99  
## - famrel 1 31.73 4793.9 913.31  
## - health 1 31.82 4794.0 913.31  
## - famsup 1 45.51 4807.7 914.16  
## - absences 1 53.73 4815.9 914.67  
## - age 1 77.91 4840.1 916.16  
## - schoolsup 1 78.78 4841.0 916.22  
## - famsize 1 82.65 4844.9 916.45  
## - Medu 1 89.31 4851.5 916.86  
## <none> 4762.2 917.02  
## - romantic 1 145.52 4907.7 920.30  
## - failures 1 343.02 5105.2 932.05  
##   
## Step: AIC=911.88  
## G3 ~ age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + nursery + internet + romantic + famrel + freetime +   
## health + absences  
##   
## Df Sum of Sq RSS AIC  
## - freetime 1 10.08 4781.1 906.81  
## - studytime 1 19.93 4791.0 907.42  
## - nursery 1 20.88 4791.9 907.48  
## - internet 1 26.67 4797.7 907.84  
## - health 1 31.51 4802.5 908.14  
## - famrel 1 32.62 4803.7 908.21  
## - famsup 1 44.83 4815.9 908.97  
## - absences 1 52.96 4824.0 909.47  
## - schoolsup 1 80.25 4851.3 911.15  
## - famsize 1 82.23 4853.3 911.27  
## - age 1 83.01 4854.0 911.32  
## - Medu 1 84.76 4855.8 911.43  
## <none> 4771.0 911.88  
## - romantic 1 143.67 4914.7 915.02  
## - failures 1 359.32 5130.4 927.82  
##   
## Step: AIC=906.81  
## G3 ~ age + famsize + Medu + studytime + failures + schoolsup +   
## famsup + nursery + internet + romantic + famrel + health +   
## absences  
##   
## Df Sum of Sq RSS AIC  
## - studytime 1 15.69 4796.8 902.09  
## - nursery 1 20.01 4801.1 902.36  
## - health 1 29.23 4810.3 902.93  
## - internet 1 31.42 4812.5 903.06  
## - famrel 1 39.56 4820.7 903.57  
## - famsup 1 44.66 4825.8 903.88  
## - absences 1 47.82 4828.9 904.08  
## - schoolsup 1 79.28 4860.4 906.01  
## - age 1 79.50 4860.6 906.03  
## - famsize 1 80.89 4862.0 906.11  
## - Medu 1 86.17 4867.3 906.43  
## <none> 4781.1 906.81  
## - romantic 1 143.13 4924.2 909.90  
## - failures 1 353.33 5134.4 922.36  
##   
## Step: AIC=902.09  
## G3 ~ age + famsize + Medu + failures + schoolsup + famsup + nursery +   
## internet + romantic + famrel + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - nursery 1 17.69 4814.5 897.49  
## - internet 1 31.09 4827.9 898.32  
## - health 1 33.77 4830.6 898.48  
## - famsup 1 39.78 4836.6 898.85  
## - famrel 1 43.59 4840.4 899.09  
## - absences 1 43.82 4840.6 899.10  
## - age 1 74.67 4871.5 900.99  
## - schoolsup 1 75.89 4872.7 901.07  
## - famsize 1 76.83 4873.6 901.13  
## - Medu 1 83.89 4880.7 901.56  
## <none> 4796.8 902.09  
## - romantic 1 136.24 4933.0 904.74  
## - failures 1 396.48 5193.3 920.06  
##   
## Step: AIC=897.49  
## G3 ~ age + famsize + Medu + failures + schoolsup + famsup + internet +   
## romantic + famrel + health + absences  
##   
## Df Sum of Sq RSS AIC  
## - health 1 30.41 4844.9 893.67  
## - internet 1 34.19 4848.7 893.90  
## - absences 1 40.95 4855.4 894.32  
## - famrel 1 41.21 4855.7 894.33  
## - famsup 1 44.03 4858.5 894.50  
## - age 1 67.36 4881.9 895.93  
## - famsize 1 68.21 4882.7 895.98  
## - Medu 1 76.03 4890.5 896.46  
## - schoolsup 1 77.64 4892.1 896.56  
## <none> 4814.5 897.49  
## - romantic 1 139.45 4953.9 900.30  
## - failures 1 390.81 5205.3 915.05  
##   
## Step: AIC=893.67  
## G3 ~ age + famsize + Medu + failures + schoolsup + famsup + internet +   
## romantic + famrel + absences  
##   
## Df Sum of Sq RSS AIC  
## - famrel 1 32.40 4877.3 889.96  
## - absences 1 34.46 4879.4 890.08  
## - internet 1 40.03 4884.9 890.42  
## - famsup 1 47.43 4892.3 890.87  
## - age 1 54.51 4899.4 891.30  
## - schoolsup 1 67.63 4912.5 892.10  
## - famsize 1 70.45 4915.3 892.27  
## - Medu 1 80.10 4925.0 892.86  
## <none> 4844.9 893.67  
## - romantic 1 147.34 4992.2 896.90  
## - failures 1 412.26 5257.2 912.31  
##   
## Step: AIC=889.96  
## G3 ~ age + famsize + Medu + failures + schoolsup + famsup + internet +   
## romantic + absences  
##   
## Df Sum of Sq RSS AIC  
## - absences 1 29.39 4906.7 886.05  
## - internet 1 42.10 4919.4 886.82  
## - famsup 1 47.65 4925.0 887.16  
## - age 1 49.65 4926.9 887.28  
## - schoolsup 1 65.82 4943.1 888.25  
## - famsize 1 73.12 4950.4 888.69  
## - Medu 1 81.49 4958.8 889.20  
## <none> 4877.3 889.96  
## - romantic 1 151.57 5028.9 893.38  
## - failures 1 402.76 5280.1 907.90  
##   
## Step: AIC=886.05  
## G3 ~ age + famsize + Medu + failures + schoolsup + famsup + internet +   
## romantic  
##   
## Df Sum of Sq RSS AIC  
## - age 1 36.07 4942.8 882.54  
## - famsup 1 48.23 4954.9 883.27  
## - internet 1 49.39 4956.1 883.34  
## - schoolsup 1 56.14 4962.8 883.74  
## - famsize 1 71.01 4977.7 884.63  
## <none> 4906.7 886.05  
## - Medu 1 96.15 5002.8 886.14  
## - romantic 1 141.09 5047.8 888.80  
## - failures 1 397.95 5304.6 903.59  
##   
## Step: AIC=882.54  
## G3 ~ famsize + Medu + failures + schoolsup + famsup + internet +   
## romantic  
##   
## Df Sum of Sq RSS AIC  
## - schoolsup 1 38.61 4981.4 879.16  
## - famsup 1 43.39 4986.2 879.44  
## - internet 1 53.16 4995.9 880.03  
## - famsize 1 63.68 5006.4 880.65  
## <none> 4942.8 882.54  
## - Medu 1 112.56 5055.3 883.55  
## - romantic 1 166.38 5109.1 886.70  
## - failures 1 493.04 5435.8 905.17  
##   
## Step: AIC=879.16  
## G3 ~ famsize + Medu + failures + famsup + internet + romantic  
##   
## Df Sum of Sq RSS AIC  
## - internet 1 48.76 5030.1 876.36  
## - famsup 1 54.85 5036.2 876.72  
## - famsize 1 65.56 5046.9 877.36  
## <none> 4981.4 879.16  
## - Medu 1 115.88 5097.3 880.31  
## - romantic 1 159.04 5140.4 882.83  
## - failures 1 505.19 5486.6 902.25  
##   
## Step: AIC=876.36  
## G3 ~ famsize + Medu + failures + famsup + romantic  
##   
## Df Sum of Sq RSS AIC  
## - famsup 1 48.61 5078.8 873.53  
## - famsize 1 66.91 5097.1 874.60  
## <none> 5030.1 876.36  
## - Medu 1 145.22 5175.4 879.15  
## - romantic 1 147.76 5177.9 879.29  
## - failures 1 521.78 5551.9 900.08  
##   
## Step: AIC=873.53  
## G3 ~ famsize + Medu + failures + romantic  
##   
## Df Sum of Sq RSS AIC  
## - famsize 1 78.09 5156.8 872.38  
## <none> 5078.8 873.53  
## - Medu 1 114.79 5193.5 874.50  
## - romantic 1 147.25 5226.0 876.35  
## - failures 1 526.64 5605.4 897.24  
##   
## Step: AIC=872.38  
## G3 ~ Medu + failures + romantic  
##   
## Df Sum of Sq RSS AIC  
## <none> 5156.8 872.38  
## - Medu 1 102.28 5259.1 872.54  
## - romantic 1 136.67 5293.5 874.48  
## - failures 1 545.76 5702.6 896.66

pred <- fit\_step %>% predict(test.data)  
data.frame(  
RMSE = RMSE(pred, test.data$G3),  
Rsquare = R2(pred, test.data$G3)  
)

## RMSE Rsquare  
## 1 4.32272 0.147046

# Question 3

library(leaps)  
fit\_bs <- regsubsets(G3~., data = train.data, nvmax = 30)  
result <- summary(fit\_bs)  
which.min(result$bic)

## [1] 3

result$which[3,]

## (Intercept) schoolMS sexM age   
## TRUE FALSE FALSE FALSE   
## addressU famsizeLE3 PstatusT Medu   
## FALSE FALSE FALSE FALSE   
## Fedu Mjobhealth Mjobother Mjobservices   
## FALSE TRUE FALSE FALSE   
## Mjobteacher Fjobhealth Fjobother Fjobservices   
## FALSE FALSE FALSE FALSE   
## Fjobteacher reasonhome reasonother reasonreputation   
## FALSE FALSE FALSE FALSE   
## guardianmother guardianother traveltime studytime   
## FALSE FALSE FALSE FALSE   
## failures schoolsupyes famsupyes paidyes   
## TRUE FALSE FALSE FALSE   
## activitiesyes nurseryyes higheryes internetyes   
## FALSE FALSE FALSE FALSE   
## romanticyes famrel freetime goout   
## TRUE FALSE FALSE FALSE   
## Dalc Walc health absences   
## FALSE FALSE FALSE FALSE

fit\_bs <- lm(G3~Mjob + failures + romantic, data = train.data)  
pred <- fit\_bs %>% predict(test.data)  
data.frame(  
RMSE = RMSE(pred, test.data$G3),  
Rsquare = R2(pred, test.data$G3)  
)

## RMSE Rsquare  
## 1 4.339538 0.1412396

# Question 4

## Two models with different predictors, three predictors each

## BICs: Smaller -> better

mod1 <- lm(G3~Medu + failures + romantic, data = train.data)  
mod2 <- lm(G3~Mjob + failures + romantic, data = train.data)  
BIC(mod1)

## [1] 1723.767

BIC(mod2)

## [1] 1734.683

Based on the BIC values of the fitted model using the training data - the one selected by the Stepwise regression using the BIC criterion has a slughtly smaller BIC value. This is because the Stepwise procedure uses the BIC critrion while the Best Subset procedure here uses the SSE (RSS) as selection criterion. Theoretically, had we used the same model selection criterion, say BIC, for both the stepwise and the best subset procedures, then with the same number of predictors, the best subset model should be the optimal.

According to the RMSE and R2 of the predition for the test data, the slightly better model is the 3-variable stepwie regression model as well, for this example.

We could (a) include interaction terms or (b) transform the dependent variables or (c) consider the randomness of regressors to improve the result.