v02

# 0) Env Init

# 00) Data Init

# Load data  
student\_data <- read.csv("rss/DATA\_RAW/student-por.csv", sep = ";")  
  
# Select numeric columns  
numeric\_vars <- select\_if(student\_data, is.numeric)  
  
# Count and proportion for each categorical variable  
categorical\_vars <- select\_if(student\_data, is.character)

# 1) Overview

## —- 1-1)Basic Overview

# Dataset summary  
str(student\_data)

'data.frame': 649 obs. of 33 variables:  
 $ school : chr "GP" "GP" "GP" "GP" ...  
 $ sex : chr "F" "F" "F" "F" ...  
 $ age : int 18 17 15 15 16 16 16 17 15 15 ...  
 $ address : chr "U" "U" "U" "U" ...  
 $ famsize : chr "GT3" "GT3" "LE3" "GT3" ...  
 $ Pstatus : chr "A" "T" "T" "T" ...  
 $ Medu : int 4 1 1 4 3 4 2 4 3 3 ...  
 $ Fedu : int 4 1 1 2 3 3 2 4 2 4 ...  
 $ Mjob : chr "at\_home" "at\_home" "at\_home" "health" ...  
 $ Fjob : chr "teacher" "other" "other" "services" ...  
 $ reason : chr "course" "course" "other" "home" ...  
 $ guardian : chr "mother" "father" "mother" "mother" ...  
 $ traveltime: int 2 1 1 1 1 1 1 2 1 1 ...  
 $ studytime : int 2 2 2 3 2 2 2 2 2 2 ...  
 $ failures : int 0 0 0 0 0 0 0 0 0 0 ...  
 $ schoolsup : chr "yes" "no" "yes" "no" ...  
 $ famsup : chr "no" "yes" "no" "yes" ...  
 $ paid : chr "no" "no" "no" "no" ...  
 $ activities: chr "no" "no" "no" "yes" ...  
 $ nursery : chr "yes" "no" "yes" "yes" ...  
 $ higher : chr "yes" "yes" "yes" "yes" ...  
 $ internet : chr "no" "yes" "yes" "yes" ...  
 $ romantic : chr "no" "no" "no" "yes" ...  
 $ famrel : int 4 5 4 3 4 5 4 4 4 5 ...  
 $ freetime : int 3 3 3 2 3 4 4 1 2 5 ...  
 $ goout : int 4 3 2 2 2 2 4 4 2 1 ...  
 $ Dalc : int 1 1 2 1 1 1 1 1 1 1 ...  
 $ Walc : int 1 1 3 1 2 2 1 1 1 1 ...  
 $ health : int 3 3 3 5 5 5 3 1 1 5 ...  
 $ absences : int 4 2 6 0 0 6 0 2 0 0 ...  
 $ G1 : int 0 9 12 14 11 12 13 10 15 12 ...  
 $ G2 : int 11 11 13 14 13 12 12 13 16 12 ...  
 $ G3 : int 11 11 12 14 13 13 13 13 17 13 ...

summary(student\_data)

school sex age address   
 Length:649 Length:649 Min. :15.00 Length:649   
 Class :character Class :character 1st Qu.:16.00 Class :character   
 Mode :character Mode :character Median :17.00 Mode :character   
 Mean :16.74   
 3rd Qu.:18.00   
 Max. :22.00   
 famsize Pstatus Medu Fedu   
 Length:649 Length:649 Min. :0.000 Min. :0.000   
 Class :character Class :character 1st Qu.:2.000 1st Qu.:1.000   
 Mode :character Mode :character Median :2.000 Median :2.000   
 Mean :2.515 Mean :2.307   
 3rd Qu.:4.000 3rd Qu.:3.000   
 Max. :4.000 Max. :4.000   
 Mjob Fjob reason guardian   
 Length:649 Length:649 Length:649 Length:649   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
 traveltime studytime failures schoolsup   
 Min. :1.000 Min. :1.000 Min. :0.0000 Length:649   
 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:0.0000 Class :character   
 Median :1.000 Median :2.000 Median :0.0000 Mode :character   
 Mean :1.569 Mean :1.931 Mean :0.2219   
 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:0.0000   
 Max. :4.000 Max. :4.000 Max. :3.0000   
 famsup paid activities nursery   
 Length:649 Length:649 Length:649 Length:649   
 Class :character Class :character Class :character Class :character   
 Mode :character Mode :character Mode :character Mode :character   
   
   
   
 higher internet romantic famrel   
 Length:649 Length:649 Length:649 Min. :1.000   
 Class :character Class :character Class :character 1st Qu.:4.000   
 Mode :character Mode :character Mode :character Median :4.000   
 Mean :3.931   
 3rd Qu.:5.000   
 Max. :5.000   
 freetime goout Dalc Walc health   
 Min. :1.00 Min. :1.000 Min. :1.000 Min. :1.00 Min. :1.000   
 1st Qu.:3.00 1st Qu.:2.000 1st Qu.:1.000 1st Qu.:1.00 1st Qu.:2.000   
 Median :3.00 Median :3.000 Median :1.000 Median :2.00 Median :4.000   
 Mean :3.18 Mean :3.185 Mean :1.502 Mean :2.28 Mean :3.536   
 3rd Qu.:4.00 3rd Qu.:4.000 3rd Qu.:2.000 3rd Qu.:3.00 3rd Qu.:5.000   
 Max. :5.00 Max. :5.000 Max. :5.000 Max. :5.00 Max. :5.000   
 absences G1 G2 G3   
 Min. : 0.000 Min. : 0.0 Min. : 0.00 Min. : 0.00   
 1st Qu.: 0.000 1st Qu.:10.0 1st Qu.:10.00 1st Qu.:10.00   
 Median : 2.000 Median :11.0 Median :11.00 Median :12.00   
 Mean : 3.659 Mean :11.4 Mean :11.57 Mean :11.91   
 3rd Qu.: 6.000 3rd Qu.:13.0 3rd Qu.:13.00 3rd Qu.:14.00   
 Max. :32.000 Max. :19.0 Max. :19.00 Max. :19.00

# Dimensions and missing values  
dim(student\_data)

[1] 649 33

colSums(is.na(student\_data))

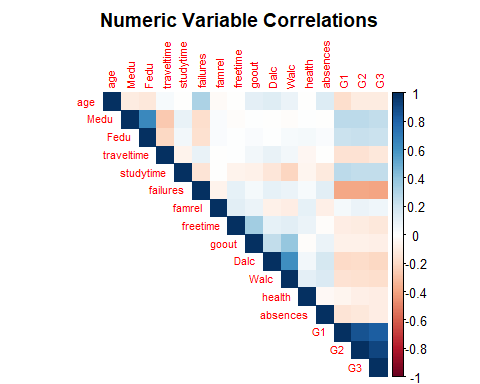
school sex age address famsize Pstatus Medu   
 0 0 0 0 0 0 0   
 Fedu Mjob Fjob reason guardian traveltime studytime   
 0 0 0 0 0 0 0   
 failures schoolsup famsup paid activities nursery higher   
 0 0 0 0 0 0 0   
 internet romantic famrel freetime goout Dalc Walc   
 0 0 0 0 0 0 0   
 health absences G1 G2 G3   
 0 0 0 0 0

## —- 1-2) Numeric Variables: Centrality, Spread, Skewness

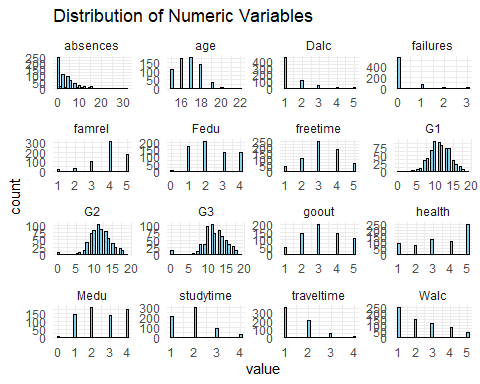
# Descriptive statistics for numeric variables  
psych::describe(numeric\_vars)

vars n mean sd median trimmed mad min max range skew kurtosis  
age 1 649 16.74 1.22 17 16.70 1.48 15 22 7 0.41 0.05  
Medu 2 649 2.51 1.13 2 2.53 1.48 0 4 4 -0.03 -1.27  
Fedu 3 649 2.31 1.10 2 2.27 1.48 0 4 4 0.21 -1.12  
traveltime 4 649 1.57 0.75 1 1.43 0.00 1 4 3 1.24 1.08  
studytime 5 649 1.93 0.83 2 1.85 1.48 1 4 3 0.70 0.02  
failures 6 649 0.22 0.59 0 0.07 0.00 0 3 3 3.08 9.70  
famrel 7 649 3.93 0.96 4 4.05 1.48 1 5 4 -1.10 1.32  
freetime 8 649 3.18 1.05 3 3.19 1.48 1 5 4 -0.18 -0.41  
goout 9 649 3.18 1.18 3 3.20 1.48 1 5 4 -0.01 -0.87  
Dalc 10 649 1.50 0.92 1 1.28 0.00 1 5 4 2.13 4.28  
Walc 11 649 2.28 1.28 2 2.14 1.48 1 5 4 0.63 -0.78  
health 12 649 3.54 1.45 4 3.67 1.48 1 5 4 -0.50 -1.13  
absences 13 649 3.66 4.64 2 2.80 2.97 0 32 32 2.01 5.70  
G1 14 649 11.40 2.75 11 11.38 2.97 0 19 19 0.00 0.02  
G2 15 649 11.57 2.91 11 11.56 2.97 0 19 19 -0.36 1.63  
G3 16 649 11.91 3.23 12 12.04 2.97 0 19 19 -0.91 2.66  
 se  
age 0.05  
Medu 0.04  
Fedu 0.04  
traveltime 0.03  
studytime 0.03  
failures 0.02  
famrel 0.04  
freetime 0.04  
goout 0.05  
Dalc 0.04  
Walc 0.05  
health 0.06  
absences 0.18  
G1 0.11  
G2 0.11  
G3 0.13

# Correlation matrix  
cor\_mat <- cor(numeric\_vars)  
corrplot::corrplot(cor\_mat, method = "color", type = "upper",  
 tl.cex = 0.7, title = "Numeric Variable Correlations",  
 mar = c(0, 0, 2, 0))



# Distribution inspection  
numeric\_vars %>%  
 pivot\_longer(everything()) %>%  
 ggplot(aes(x = value)) +  
 geom\_histogram(bins = 30, fill = "skyblue", color = "black") +  
 facet\_wrap(~ name, scales = "free") +  
 theme\_minimal() +  
 labs(title = "Distribution of Numeric Variables")

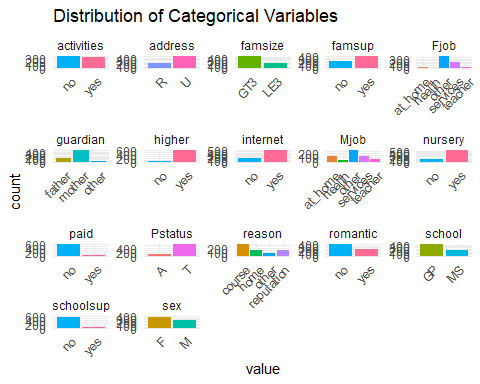


## —- 1-3) Categorical Variables: Balance and Association

# Summary of categorical distributions  
categorical\_summary <- categorical\_vars %>%  
 summarise\_all(~n\_distinct(.)) %>%  
 pivot\_longer(everything(), names\_to = "variable", values\_to = "unique\_levels")  
  
categorical\_summary

# A tibble: 17 × 2  
 variable unique\_levels  
 <chr> <int>  
 1 school 2  
 2 sex 2  
 3 address 2  
 4 famsize 2  
 5 Pstatus 2  
 6 Mjob 5  
 7 Fjob 5  
 8 reason 4  
 9 guardian 3  
10 schoolsup 2  
11 famsup 2  
12 paid 2  
13 activities 2  
14 nursery 2  
15 higher 2  
16 internet 2  
17 romantic 2

# Frequency plots  
categorical\_vars %>%  
 pivot\_longer(everything()) %>%  
 ggplot(aes(x = value, fill = value)) +  
 geom\_bar(show.legend = FALSE) +  
 facet\_wrap(~ name, scales = "free") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 labs(title = "Distribution of Categorical Variables")

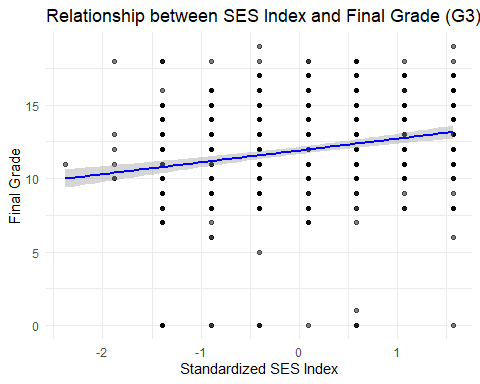


# Chi-square associations with G3 (converted to factor)  
cat\_g3\_assoc <- lapply(names(categorical\_vars), function(var) {  
 tbl <- table(student\_data[[var]], student\_data$G3)  
 chisq <- suppressWarnings(chisq.test(tbl))  
 data.frame(var = var, p\_value = chisq$p.value)  
}) %>% bind\_rows()  
  
cat\_g3\_assoc %>% arrange(p\_value)

var p\_value  
1 higher 1.897542e-15  
2 school 1.019738e-12  
3 reason 4.831586e-04  
4 address 2.394404e-03  
5 Mjob 8.076200e-03  
6 internet 7.964799e-02  
7 Fjob 8.213326e-02  
8 schoolsup 8.309750e-02  
9 sex 1.461145e-01  
10 activities 1.555599e-01  
11 famsup 2.445735e-01  
12 romantic 3.853833e-01  
13 nursery 4.418679e-01  
14 guardian 4.789574e-01  
15 paid 5.913118e-01  
16 famsize 7.745382e-01  
17 Pstatus 9.989803e-01

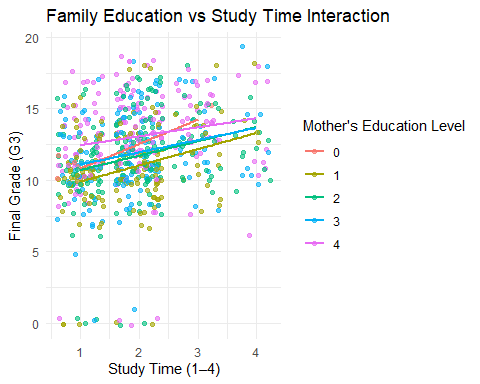
# SES index from parental education  
student\_data <- student\_data %>%  
 mutate(SES\_index = scale(Medu + Fedu))  
  
# Quick check of SES vs G3  
ggplot(student\_data, aes(x = SES\_index, y = G3)) +  
 geom\_point(alpha = 0.5) +  
 geom\_smooth(method = "lm", color = "blue") +  
 theme\_minimal() +  
 labs(title = "Relationship between SES Index and Final Grade (G3)",  
 x = "Standardized SES Index", y = "Final Grade")

`geom\_smooth()` using formula = 'y ~ x'

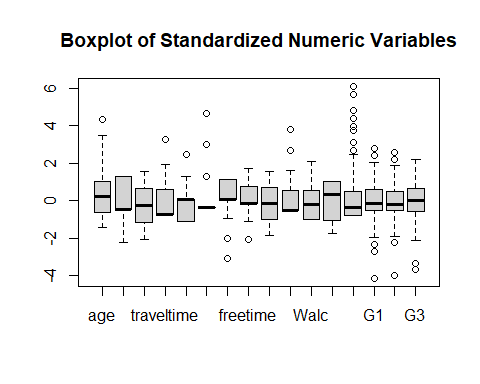


ggplot(student\_data, aes(x = studytime, y = G3, color = as.factor(Medu))) +  
 geom\_jitter(alpha = 0.6) +  
 geom\_smooth(method = "lm", se = FALSE) +  
 theme\_minimal() +  
 labs(title = "Family Education vs Study Time Interaction",  
 x = "Study Time (1–4)", y = "Final Grade (G3)",  
 color = "Mother's Education Level")

`geom\_smooth()` using formula = 'y ~ x'



# Check for outliers  
boxplot(scale(numeric\_vars), main = "Boxplot of Standardized Numeric Variables")



# Z-score summary for extreme values  
apply(scale(numeric\_vars), 2, function(x) sum(abs(x) > 3))

age Medu Fedu traveltime studytime failures famrel   
 3 0 0 16 0 14 22   
 freetime goout Dalc Walc health absences G1   
 0 0 17 0 0 11 1   
 G2 G3   
 7 16

# 2) transformation

## —- 2-0) Packages —-

## ---- 0) Packages ----  
pkgs <- c("dplyr","tidyr","ggplot2","fastDummies","stringr","broom","car","DescTools")  
to\_install <- setdiff(pkgs, rownames(installed.packages()))  
if (length(to\_install)) install.packages(to\_install, quiet = TRUE)  
invisible(lapply(pkgs, library, character.only = TRUE))

Warning: package 'fastDummies' was built under R version 4.4.3

Warning: package 'car' was built under R version 4.4.3

Loading required package: carData

Warning: package 'carData' was built under R version 4.4.3

Attaching package: 'car'

The following object is masked from 'package:psych':  
  
 logit

The following object is masked from 'package:dplyr':  
  
 recode

The following object is masked from 'package:purrr':  
  
 some

Warning: package 'DescTools' was built under R version 4.4.3

Attaching package: 'DescTools'

The following object is masked from 'package:car':  
  
 Recode

The following objects are masked from 'package:caret':  
  
 MAE, RMSE

The following objects are masked from 'package:psych':  
  
 AUC, ICC, SD

set.seed(2025)

## —- 2-1) Base typing: numeric vs categorical —-

## ---- 1) Base typing: numeric vs categorical ----  
# Ensure characters become factors (we keep numeric as is)  
student\_data <- student\_data %>%  
 mutate(across(where(is.character), as.factor))

## —- 2-2) SES features —-

## ---- 2) SES features ----  
# Basic additive SES and standardized version (both kept)  
student\_data <- student\_data %>%  
 mutate(  
 SES\_raw = Medu + Fedu,  
 SES\_z = as.numeric(scale(SES\_raw))  
 )  
  
# Optional: a PCA SES using education + address + Pstatus + parental jobs (kept if useful later)  
ses\_pca\_vars <- c("Medu","Fedu","address","Pstatus","Mjob","Fjob")  
# Convert the categorical subset to dummies for PCA  
ses\_dum <- fastDummies::dummy\_cols(student\_data[ses\_pca\_vars],  
 remove\_first\_dummy = TRUE, remove\_selected\_columns = TRUE)  
ses\_pca <- prcomp(scale(as.matrix(ses\_dum)), center = TRUE, scale. = TRUE)  
student\_data$SES\_pca1 <- as.numeric(scale(ses\_pca$x[,1])) # first PC as another SES proxy

## —- 2-3) “effort/behavior” transforms (keep originals) —-

## ---- 3) “effort/behavior” transforms (keep originals) ----  
student\_data <- student\_data %>%  
 mutate(  
 absences\_log = log1p(absences),  
 failures\_log = log1p(failures),  
 studytime\_log = log(studytime), # studytime is 1..4 → safe  
 Dalc\_log = log(Dalc),  
 Walc\_log = log(Walc),  
 alcohol\_use = (Dalc + Walc) / 2, # weekly avg  
 failure\_flag = as.integer(failures > 0), # 0/1  
 weekend\_drink = as.integer(Walc >= 3) # 0/1 indicator of high W-end alcohol  
 )

## —- 2-4) Optional Winsorization for heavy tails (kept as \*\_win) —-

## ---- 4) Optional Winsorization for heavy tails (kept as \*\_win) ----  
# We keep winsorized copies; originals remain untouched.  
student\_data <- student\_data %>%  
 mutate(  
 absences\_win = winsorize(absences, p = c(0.01, 0.99)),  
 Dalc\_win = winsorize(Dalc, p = c(0.01, 0.99)),  
 Walc\_win = winsorize(Walc, p = c(0.01, 0.99))  
 )  
  
summary(student\_data[, c("absences","absences\_win","Dalc","Dalc\_win","Walc","Walc\_win")])

absences absences\_win Dalc Dalc\_win   
 Min. : 0.000 Min. : 0.000 Min. :1.000 Min. :1.000   
 1st Qu.: 0.000 1st Qu.: 0.000 1st Qu.:1.000 1st Qu.:1.000   
 Median : 2.000 Median : 2.000 Median :1.000 Median :1.000   
 Mean : 3.659 Mean : 3.613 Mean :1.502 Mean :1.502   
 3rd Qu.: 6.000 3rd Qu.: 6.000 3rd Qu.:2.000 3rd Qu.:2.000   
 Max. :32.000 Max. :21.000 Max. :5.000 Max. :5.000   
 Walc Walc\_win   
 Min. :1.00 Min. :1.00   
 1st Qu.:1.00 1st Qu.:1.00   
 Median :2.00 Median :2.00   
 Mean :2.28 Mean :2.28   
 3rd Qu.:3.00 3rd Qu.:3.00   
 Max. :5.00 Max. :5.00

## —- 2-5) Z-scales for interaction-ready predictors (kept as \*\_z) —-

## ---- 5) Z-scales for interaction-ready predictors (kept as \*\_z) ----  
zscale <- function(x) as.numeric(scale(x))  
student\_data <- student\_data %>%  
 mutate(  
 studytime\_z = zscale(studytime),  
 studytime\_log\_z = zscale(studytime\_log),  
 SES\_raw\_z = zscale(SES\_raw),  
 absences\_log\_z = zscale(absences\_log),  
 failures\_log\_z = zscale(failures\_log),  
 alcohol\_use\_z = zscale(alcohol\_use)  
 )  
  
glimpse(student\_data)

Rows: 649  
Columns: 54  
$ school <fct> GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP, GP…  
$ sex <fct> F, F, F, F, F, M, M, F, M, M, F, F, M, M, M, F, F, F, …  
$ age <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, 15…  
$ address <fct> U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, U, …  
$ famsize <fct> GT3, GT3, LE3, GT3, GT3, LE3, LE3, GT3, LE3, GT3, GT3,…  
$ Pstatus <fct> A, T, T, T, T, T, T, A, A, T, T, T, T, T, A, T, T, T, …  
$ Medu <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3, …  
$ Fedu <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3, …  
$ Mjob <fct> at\_home, at\_home, at\_home, health, other, services, ot…  
$ Fjob <fct> teacher, other, other, services, other, other, other, …  
$ reason <fct> course, course, other, home, home, reputation, home, h…  
$ guardian <fct> mother, father, mother, mother, father, mother, mother…  
$ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3, …  
$ studytime <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2, …  
$ failures <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ schoolsup <fct> yes, no, yes, no, no, no, no, yes, no, no, no, no, no,…  
$ famsup <fct> no, yes, no, yes, yes, yes, no, yes, yes, yes, yes, ye…  
$ paid <fct> no, no, no, no, no, no, no, no, no, no, no, no, no, no…  
$ activities <fct> no, no, no, yes, no, yes, no, no, no, yes, no, yes, ye…  
$ nursery <fct> yes, no, yes, yes, yes, yes, yes, yes, yes, yes, yes, …  
$ higher <fct> yes, yes, yes, yes, yes, yes, yes, yes, yes, yes, yes,…  
$ internet <fct> no, yes, yes, yes, no, yes, yes, no, yes, yes, yes, ye…  
$ romantic <fct> no, no, no, yes, no, no, no, no, no, no, no, no, no, n…  
$ famrel <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5, …  
$ freetime <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3, …  
$ goout <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2, …  
$ Dalc <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
$ Walc <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, …  
$ health <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4, …  
$ absences <int> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2,…  
$ G1 <int> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12, 12, …  
$ G2 <int> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 13, 12…  
$ G3 <int> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, 13…  
$ SES\_index <dbl[,1]> <matrix[26 x 1]>  
$ SES\_raw <int> 8, 2, 2, 6, 6, 7, 4, 8, 5, 7, 8, 3, 8, 7, 4, 8, 8,…  
$ SES\_z <dbl> 1.56737064, -1.39110793, -1.39110793, 0.58121112, 0.58…  
$ SES\_pca1 <dbl> -1.76001951, 0.95015632, 0.95015632, -1.13412429, 0.35…  
$ absences\_log <dbl> 1.609438, 1.098612, 1.945910, 0.000000, 0.000000, 1.94…  
$ failures\_log <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.00…  
$ studytime\_log <dbl> 0.6931472, 0.6931472, 0.6931472, 1.0986123, 0.6931472,…  
$ Dalc\_log <dbl> 0.0000000, 0.0000000, 0.6931472, 0.0000000, 0.0000000,…  
$ Walc\_log <dbl> 0.0000000, 0.0000000, 1.0986123, 0.0000000, 0.6931472,…  
$ alcohol\_use <dbl> 1.0, 1.0, 2.5, 1.0, 1.5, 1.5, 1.0, 1.0, 1.0, 1.0, 1.5,…  
$ failure\_flag <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …  
$ weekend\_drink <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, …  
$ absences\_win <dbl> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, 2,…  
$ Dalc\_win <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, …  
$ Walc\_win <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1, …  
$ studytime\_z <dbl> 0.08358848, 0.08358848, 0.08358848, 1.28912007, 0.0835…  
$ studytime\_log\_z <dbl> 0.2937155, 0.2937155, 0.2937155, 1.2209365, 0.2937155,…  
$ SES\_raw\_z <dbl> 1.56737064, -1.39110793, -1.39110793, 0.58121112, 0.58…  
$ absences\_log\_z <dbl> 0.54010926, 0.01036996, 0.88903962, -1.12891923, -1.12…  
$ failures\_log\_z <dbl> -0.4049795, -0.4049795, -0.4049795, -0.4049795, -0.404…  
$ alcohol\_use\_z <dbl> -0.8947665, -0.8947665, 0.6109469, -0.8947665, -0.3928…

## —- 2-6) Categorical encoding (dummies) —-

## ---- 6) Categorical encoding (dummies) ----  
cat\_vars <- c("school","sex","address","famsize","Pstatus","Mjob","Fjob","reason",  
 "guardian","schoolsup","famsup","paid","activities","nursery",  
 "higher","internet","romantic")  
  
student\_dum <- fastDummies::dummy\_cols(  
 student\_data,  
 select\_columns = cat\_vars,  
 remove\_selected\_columns = TRUE,  
 remove\_first\_dummy = TRUE # avoids perfect collinearity (baseline kept implicitly)  
)  
  
glimpse(student\_dum)

Rows: 649  
Columns: 63  
$ age <int> 18, 17, 15, 15, 16, 16, 16, 17, 15, 15, 15, 15, 15, …  
$ Medu <int> 4, 1, 1, 4, 3, 4, 2, 4, 3, 3, 4, 2, 4, 4, 2, 4, 4, 3…  
$ Fedu <int> 4, 1, 1, 2, 3, 3, 2, 4, 2, 4, 4, 1, 4, 3, 2, 4, 4, 3…  
$ traveltime <int> 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 3, 1, 2, 1, 1, 1, 3…  
$ studytime <int> 2, 2, 2, 3, 2, 2, 2, 2, 2, 2, 2, 3, 1, 2, 3, 1, 3, 2…  
$ failures <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
$ famrel <int> 4, 5, 4, 3, 4, 5, 4, 4, 4, 5, 3, 5, 4, 5, 4, 4, 3, 5…  
$ freetime <int> 3, 3, 3, 2, 3, 4, 4, 1, 2, 5, 3, 2, 3, 4, 5, 4, 2, 3…  
$ goout <int> 4, 3, 2, 2, 2, 2, 4, 4, 2, 1, 3, 2, 3, 3, 2, 4, 3, 2…  
$ Dalc <int> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
$ Walc <int> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1…  
$ health <int> 3, 3, 3, 5, 5, 5, 3, 1, 1, 5, 2, 4, 5, 3, 3, 2, 2, 4…  
$ absences <int> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, …  
$ G1 <int> 0, 9, 12, 14, 11, 12, 13, 10, 15, 12, 14, 10, 12, 12…  
$ G2 <int> 11, 11, 13, 14, 13, 12, 12, 13, 16, 12, 14, 12, 13, …  
$ G3 <int> 11, 11, 12, 14, 13, 13, 13, 13, 17, 13, 14, 13, 12, …  
$ SES\_index <dbl> 1.56737064, -1.39110793, -1.39110793, 0.58121112, 0.…  
$ SES\_raw <int> 8, 2, 2, 6, 6, 7, 4, 8, 5, 7, 8, 3, 8, 7, 4, 8, 8, 6…  
$ SES\_z <dbl> 1.56737064, -1.39110793, -1.39110793, 0.58121112, 0.…  
$ SES\_pca1 <dbl> -1.76001951, 0.95015632, 0.95015632, -1.13412429, 0.…  
$ absences\_log <dbl> 1.609438, 1.098612, 1.945910, 0.000000, 0.000000, 1.…  
$ failures\_log <dbl> 0.000000, 0.000000, 0.000000, 0.000000, 0.000000, 0.…  
$ studytime\_log <dbl> 0.6931472, 0.6931472, 0.6931472, 1.0986123, 0.693147…  
$ Dalc\_log <dbl> 0.0000000, 0.0000000, 0.6931472, 0.0000000, 0.000000…  
$ Walc\_log <dbl> 0.0000000, 0.0000000, 1.0986123, 0.0000000, 0.693147…  
$ alcohol\_use <dbl> 1.0, 1.0, 2.5, 1.0, 1.5, 1.5, 1.0, 1.0, 1.0, 1.0, 1.…  
$ failure\_flag <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
$ weekend\_drink <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0…  
$ absences\_win <dbl> 4, 2, 6, 0, 0, 6, 0, 2, 0, 0, 2, 0, 0, 0, 0, 6, 10, …  
$ Dalc\_win <dbl> 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
$ Walc\_win <dbl> 1, 1, 3, 1, 2, 2, 1, 1, 1, 1, 2, 1, 3, 2, 1, 2, 2, 1…  
$ studytime\_z <dbl> 0.08358848, 0.08358848, 0.08358848, 1.28912007, 0.08…  
$ studytime\_log\_z <dbl> 0.2937155, 0.2937155, 0.2937155, 1.2209365, 0.293715…  
$ SES\_raw\_z <dbl> 1.56737064, -1.39110793, -1.39110793, 0.58121112, 0.…  
$ absences\_log\_z <dbl> 0.54010926, 0.01036996, 0.88903962, -1.12891923, -1.…  
$ failures\_log\_z <dbl> -0.4049795, -0.4049795, -0.4049795, -0.4049795, -0.4…  
$ alcohol\_use\_z <dbl> -0.8947665, -0.8947665, 0.6109469, -0.8947665, -0.39…  
$ school\_MS <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
$ sex\_M <int> 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0…  
$ address\_U <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
$ famsize\_LE3 <int> 0, 0, 1, 0, 0, 1, 1, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0…  
$ Pstatus\_T <int> 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1…  
$ Mjob\_health <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0…  
$ Mjob\_other <int> 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1…  
$ Mjob\_services <int> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0…  
$ Mjob\_teacher <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0…  
$ Fjob\_health <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0…  
$ Fjob\_other <int> 0, 1, 1, 0, 1, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 0, 1…  
$ Fjob\_services <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0…  
$ Fjob\_teacher <int> 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
$ reason\_home <int> 0, 0, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0…  
$ reason\_other <int> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
$ reason\_reputation <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1…  
$ guardian\_mother <int> 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 0, 1, 1, 1…  
$ guardian\_other <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0…  
$ schoolsup\_yes <int> 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1…  
$ famsup\_yes <int> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
$ paid\_yes <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0…  
$ activities\_yes <int> 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1…  
$ nursery\_yes <int> 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
$ higher\_yes <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…  
$ internet\_yes <int> 0, 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0…  
$ romantic\_yes <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0…

## —- 2-7) Interaction scaffolds (kept; you can expand later) —-

## ---- 7) Interaction scaffolds (kept; you can expand later) ----  
student\_dum <- student\_dum %>%  
 mutate(  
 SESxStudy = SES\_z \* studytime\_z,  
 SESxStudyLog = SES\_z \* studytime\_log\_z  
 )

## —- 2-8) Two modeling frames —-

# (a) Core model frame (no G1/G2)  
drop\_grades <- c("G1","G2")  
core\_keep <- setdiff(names(student\_dum), drop\_grades)  
# base R subsetting; preserves column order and avoids dplyr  
model\_core <- student\_dum[, core\_keep, drop = FALSE]  
  
# sanity check  
stopifnot("G3" %in% names(model\_core))  
str(model\_core[1:5])

'data.frame': 649 obs. of 5 variables:  
 $ age : int 18 17 15 15 16 16 16 17 15 15 ...  
 $ Medu : int 4 1 1 4 3 4 2 4 3 3 ...  
 $ Fedu : int 4 1 1 2 3 3 2 4 2 4 ...  
 $ traveltime: int 2 1 1 1 1 1 1 2 1 1 ...  
 $ studytime : int 2 2 2 3 2 2 2 2 2 2 ...

# (b) Extended frame WITH G1/G2 (if you want it):  
model\_with\_grades <- student\_dum # nothing to drop

# 3) Build a pruned modeling matrix (no leakage; robust to dummies)

## —- 3-1)Recreate the pruned modeling matrix

# Start from core frame (no G1/G2)  
df0 <- model\_core  
  
# Drop twins; keep z-versions (and/or SES\_pca1)  
drop\_families <- c(  
 "studytime","studytime\_log",  
 "absences\_log",  
 "failures\_log",  
 "alcohol\_use",  
 "SES\_raw","SES\_raw\_z","SES\_index"  
)  
df0 <- df0[, setdiff(names(df0), drop\_families), drop = FALSE]  
  
# Optional: drop near-constant factor dummies (stabilises fits)  
if ("higher\_yes" %in% names(df0)) {  
 df0 <- df0[, setdiff(names(df0), "higher\_yes"), drop = FALSE]  
}  
  
# Keep SESxStudy; drop SESxStudyLog if present  
df0 <- df0[, setdiff(names(df0), "SESxStudyLog"), drop = FALSE]  
  
# Build design matrix (no intercept) and prune  
if (!requireNamespace("caret", quietly = TRUE)) install.packages("caret")  
library(caret)  
  
X <- model.matrix(G3 ~ ., data = df0)[, -1, drop = FALSE]  
  
nzv <- caret::nearZeroVar(X)  
if (length(nzv)) X <- X[, -nzv, drop = FALSE]  
  
lc <- caret::findLinearCombos(X)  
if (!is.null(lc$remove)) X <- X[, -lc$remove, drop = FALSE]  
  
# Final modeling data  
df <- data.frame(G3 = df0$G3, X)  
str(df)

'data.frame': 649 obs. of 49 variables:  
 $ G3 : int 11 11 12 14 13 13 13 13 17 13 ...  
 $ age : num 18 17 15 15 16 16 16 17 15 15 ...  
 $ Medu : num 4 1 1 4 3 4 2 4 3 3 ...  
 $ Fedu : num 4 1 1 2 3 3 2 4 2 4 ...  
 $ traveltime : num 2 1 1 1 1 1 1 2 1 1 ...  
 $ failures : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ famrel : num 4 5 4 3 4 5 4 4 4 5 ...  
 $ freetime : num 3 3 3 2 3 4 4 1 2 5 ...  
 $ goout : num 4 3 2 2 2 2 4 4 2 1 ...  
 $ Dalc : num 1 1 2 1 1 1 1 1 1 1 ...  
 $ Walc : num 1 1 3 1 2 2 1 1 1 1 ...  
 $ health : num 3 3 3 5 5 5 3 1 1 5 ...  
 $ absences : num 4 2 6 0 0 6 0 2 0 0 ...  
 $ SES\_z : num 1.567 -1.391 -1.391 0.581 0.581 ...  
 $ SES\_pca1 : num -1.76 0.95 0.95 -1.134 0.354 ...  
 $ Dalc\_log : num 0 0 0.693 0 0 ...  
 $ Walc\_log : num 0 0 1.099 0 0.693 ...  
 $ failure\_flag : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ weekend\_drink : num 0 0 1 0 0 0 0 0 0 0 ...  
 $ absences\_win : num 4 2 6 0 0 6 0 2 0 0 ...  
 $ studytime\_z : num 0.0836 0.0836 0.0836 1.2891 0.0836 ...  
 $ studytime\_log\_z : num 0.294 0.294 0.294 1.221 0.294 ...  
 $ absences\_log\_z : num 0.5401 0.0104 0.889 -1.1289 -1.1289 ...  
 $ failures\_log\_z : num -0.405 -0.405 -0.405 -0.405 -0.405 ...  
 $ school\_MS : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ sex\_M : num 0 0 0 0 0 1 1 0 1 1 ...  
 $ address\_U : num 1 1 1 1 1 1 1 1 1 1 ...  
 $ famsize\_LE3 : num 0 0 1 0 0 1 1 0 1 0 ...  
 $ Pstatus\_T : num 0 1 1 1 1 1 1 0 0 1 ...  
 $ Mjob\_health : num 0 0 0 1 0 0 0 0 0 0 ...  
 $ Mjob\_other : num 0 0 0 0 1 0 1 1 0 1 ...  
 $ Mjob\_services : num 0 0 0 0 0 1 0 0 1 0 ...  
 $ Mjob\_teacher : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ Fjob\_other : num 0 1 1 0 1 1 1 0 1 1 ...  
 $ Fjob\_services : num 0 0 0 1 0 0 0 0 0 0 ...  
 $ Fjob\_teacher : num 1 0 0 0 0 0 0 1 0 0 ...  
 $ reason\_home : num 0 0 0 1 1 0 1 1 1 1 ...  
 $ reason\_other : num 0 0 1 0 0 0 0 0 0 0 ...  
 $ reason\_reputation: num 0 0 0 0 0 1 0 0 0 0 ...  
 $ guardian\_mother : num 1 0 1 1 0 1 1 1 1 1 ...  
 $ guardian\_other : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ schoolsup\_yes : num 1 0 1 0 0 0 0 1 0 0 ...  
 $ famsup\_yes : num 0 1 0 1 1 1 0 1 1 1 ...  
 $ paid\_yes : num 0 0 0 0 0 0 0 0 0 0 ...  
 $ activities\_yes : num 0 0 0 1 0 1 0 0 0 1 ...  
 $ nursery\_yes : num 1 0 1 1 1 1 1 1 1 1 ...  
 $ internet\_yes : num 0 1 1 1 0 1 1 0 1 1 ...  
 $ romantic\_yes : num 0 0 0 1 0 0 0 0 0 0 ...  
 $ SESxStudy : num 0.131 -0.1163 -0.1163 0.7493 0.0486 ...

## —- 3-2) Fit the model and compute VIFs safely

# Fit full model on pruned matrix  
fit\_core3 <- lm(G3 ~ ., data = df)  
  
# Robust VIF (handles aliased terms)  
safe\_vif <- function(model) {  
 aliased <- is.na(coef(model))  
 if (any(aliased)) {  
 keep\_terms <- setdiff(names(aliased)[!aliased], "(Intercept)")  
 message("Dropped ", sum(aliased), " aliased terms before VIF computation.")  
 form <- as.formula(paste("G3 ~", paste(keep\_terms, collapse = " + ")))  
 model <- lm(form, data = model$model)  
 }  
 car::vif(model)  
}  
  
vif\_tbl <- safe\_vif(fit\_core3)

Dropped 1 aliased terms before VIF computation.

sort(vif\_tbl, decreasing = TRUE)[1:25]

failures\_log\_z failures failure\_flag SES\_pca1 absences\_win   
 1810.282273 767.859839 265.799712 118.991578 90.761860   
 absences Walc Walc\_log Dalc\_log Dalc   
 74.719784 34.277303 25.719498 24.870472 24.563832   
studytime\_log\_z studytime\_z Fjob\_other Fedu Medu   
 22.128899 21.558996 15.391763 13.461976 11.950930   
 Mjob\_teacher Mjob\_other absences\_log\_z weekend\_drink Mjob\_health   
 7.875369 6.878167 6.843278 5.134780 3.811830   
 Mjob\_services Fjob\_teacher Fjob\_services address\_U guardian\_other   
 3.380575 3.210994 3.034984 2.525120 1.572256

## —- 3-3) Shortcut for lose the matrix step

fit\_full <- lm(G3 ~ ., data = df0)  
ali <- is.na(coef(fit\_full))  
keep\_terms <- setdiff(names(ali)[!ali], "(Intercept)")  
fit\_reduced <- lm(as.formula(paste("G3 ~", paste(keep\_terms, collapse=" + "))), data = df0)  
safe\_vif(fit\_reduced)

age Medu Fedu traveltime   
 1.428371 8.754850 9.824862 1.308540   
 failures famrel freetime goout   
 767.859839 1.151903 1.284936 1.484333   
 Dalc Walc health absences   
 24.563832 34.277303 1.168194 74.719784   
 SES\_pca1 Dalc\_log Walc\_log failure\_flag   
 74.220826 24.870472 25.719498 265.799712   
 weekend\_drink absences\_win studytime\_z studytime\_log\_z   
 5.134780 90.761860 21.558996 22.128899   
 absences\_log\_z failures\_log\_z school\_MS sex\_M   
 6.843278 1810.282273 1.515016 1.406439   
 address\_U famsize\_LE3 Pstatus\_T Mjob\_health   
 1.969961 1.158858 1.217472 2.779573   
 Mjob\_other Mjob\_services Mjob\_teacher Fjob\_health   
 4.916147 2.962749 6.178462 1.625832   
 Fjob\_other Fjob\_services reason\_home reason\_other   
 12.879290 2.854072 1.322092 1.222987   
reason\_reputation guardian\_mother guardian\_other schoolsup\_yes   
 1.408485 1.374422 1.572256 1.160154   
 famsup\_yes paid\_yes activities\_yes nursery\_yes   
 1.146974 1.115633 1.169570 1.083173   
 internet\_yes romantic\_yes SESxStudy   
 1.267617 1.136394 1.092050

# 4) Forward / Backward stepwise (AIC & BIC)

# Null and Full  
null\_mod <- lm(G3 ~ 1, data = df)  
full\_mod <- lm(G3 ~ ., data = df)  
  
# AIC stepwise (both directions from null → full)  
step\_aic\_both <- stepAIC(null\_mod,  
 scope = list(lower = ~1, upper = formula(full\_mod)),  
 direction = "both", trace = FALSE)  
  
# Backward (AIC) from full  
step\_aic\_back <- stepAIC(full\_mod, direction = "backward", trace = FALSE)  
  
# BIC versions (penalize complexity more): k = log(n)  
n <- nrow(df)  
step\_bic\_both <- stepAIC(null\_mod,  
 scope = list(lower = ~1, upper = formula(full\_mod)),  
 direction = "both", k = log(n), trace = FALSE)  
step\_bic\_back <- stepAIC(full\_mod, direction = "backward", k = log(n), trace = FALSE)  
  
# Compare quick stats  
model\_summ <- function(m) {  
 c(k = length(coef(m))-1,  
 adjR2 = summary(m)$adj.r.squared,  
 AIC = AIC(m), BIC = BIC(m))  
}  
rbind(  
 AIC\_both = model\_summ(step\_aic\_both),  
 AIC\_back = model\_summ(step\_aic\_back),  
 BIC\_both = model\_summ(step\_bic\_both),  
 BIC\_back = model\_summ(step\_bic\_back)  
)

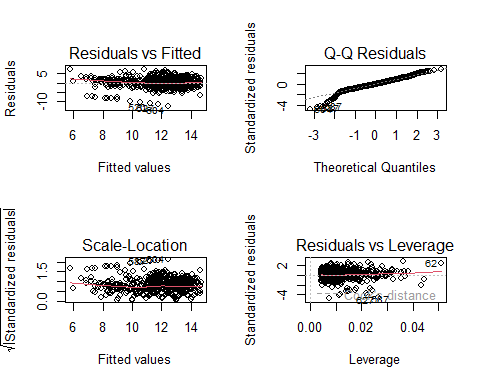
k adjR2 AIC BIC  
AIC\_both 15 0.3262685 3125.422 3201.504  
AIC\_back 15 0.3260703 3125.613 3201.695  
BIC\_both 7 0.3071567 3135.727 3176.005  
BIC\_back 7 0.3053516 3137.415 3177.694

# 5) Diagnostics on chosen model

best\_mod <- step\_bic\_both # or step\_aic\_both / step\_bic\_back / step\_aic\_back  
  
summary(best\_mod)

Call:  
lm(formula = G3 ~ failure\_flag + school\_MS + studytime\_log\_z +   
 Dalc + SES\_z + schoolsup\_yes + sex\_M, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-12.4727 -1.3969 0.0004 1.6816 7.2072   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 13.7884 0.2313 59.617 < 2e-16 \*\*\*  
failure\_flag -3.0627 0.3050 -10.043 < 2e-16 \*\*\*  
school\_MS -1.3127 0.2354 -5.575 3.64e-08 \*\*\*  
studytime\_log\_z 0.4375 0.1119 3.910 0.000102 \*\*\*  
Dalc -0.3741 0.1205 -3.106 0.001982 \*\*   
SES\_z 0.4174 0.1117 3.737 0.000203 \*\*\*  
schoolsup\_yes -1.2627 0.3508 -3.600 0.000343 \*\*\*  
sex\_M -0.6319 0.2330 -2.712 0.006867 \*\*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 2.689 on 641 degrees of freedom  
Multiple R-squared: 0.3146, Adjusted R-squared: 0.3072   
F-statistic: 42.04 on 7 and 641 DF, p-value: < 2.2e-16

# Residual diagnostics  
par(mfrow = c(2,2)); plot(best\_mod); par(mfrow = c(1,1))



# Heteroskedasticity (Breusch–Pagan)  
if (!requireNamespace("lmtest", quietly = TRUE)) install.packages("lmtest")  
if (!requireNamespace("sandwich", quietly = TRUE)) install.packages("sandwich")  
  
bptest(best\_mod)

studentized Breusch-Pagan test  
  
data: best\_mod  
BP = 29.835, df = 7, p-value = 0.0001018

# Robust (HC3) SEs if needed  
coeftest(best\_mod, vcov. = vcovHC(best\_mod, type = "HC3"))

t test of coefficients:  
  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 13.78841 0.23516 58.6342 < 2.2e-16 \*\*\*  
failure\_flag -3.06275 0.34335 -8.9201 < 2.2e-16 \*\*\*  
school\_MS -1.31268 0.27641 -4.7490 2.524e-06 \*\*\*  
studytime\_log\_z 0.43750 0.10581 4.1349 4.024e-05 \*\*\*  
Dalc -0.37414 0.14965 -2.5001 0.0126635 \*   
SES\_z 0.41742 0.12111 3.4466 0.0006048 \*\*\*  
schoolsup\_yes -1.26270 0.34923 -3.6157 0.0003231 \*\*\*  
sex\_M -0.63192 0.24594 -2.5694 0.0104139 \*   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 6) Test the “compensation” hypothesis (SES×StudyTime)

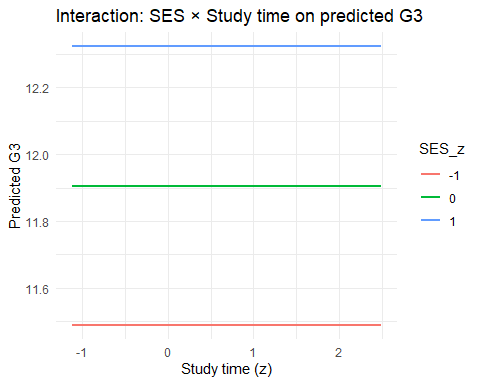
We want to know if high SES reduces the penalty of low study time. Interpret the interaction β(SES×Studytime\_z):

* If β\_interaction > 0 and significant → the slope of Studytime on G3 is steeper at higher SES; students with high SES gain more per unit study time.
* To show compensation at low study time, compute simple slopes at low vs high SES.

If the low-SES curve sits higher than expected at very low study time, or if the gap between low vs high SES narrows at the low end, that’s evidence toward “compensation.” (Statistically you’d confirm by the sign/size of the interaction and regions of significance; for a compact course project, the interaction term + the plot is usually adequate.)

# Simple slopes at SES\_z = {-1 SD, 0, +1 SD}  
# With standardized SES\_z already in df, -1/0/+1 are direct.  
  
# Refit on original df to ensure variables exist in environment  
mod <- best\_mod  
  
# Build a grid for predicted lines across studytime\_z at SES\_z levels  
  
# choose an evenly spaced range for studytime\_z present in df  
st\_rng <- seq(min(df$studytime\_z, na.rm=TRUE), max(df$studytime\_z, na.rm=TRUE), length.out = 50)  
grid <- do.call(expand.grid, c(  
 list(studytime\_z = st\_rng,  
 SES\_z = c(-1, 0, 1)),  
 # Fill other predictors at their sample means or 0 for dummies  
 lapply(df[setdiff(names(df), c("G3","studytime\_z","SES\_z"))], function(x) {  
 if (is.numeric(x)) mean(x, na.rm=TRUE) else 0  
 })  
))  
  
grid$pred <- predict(mod, newdata = grid)  
  
ggplot(grid, aes(x = studytime\_z, y = pred, color = factor(SES\_z))) +  
 geom\_line(size = 1) +  
 labs(x = "Study time (z)", y = "Predicted G3",  
 color = "SES\_z",  
 title = "Interaction: SES × Study time on predicted G3") +  
 theme\_minimal()

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
ℹ Please use `linewidth` instead.



# 7) Clean coefficient table (publishable)

if (!requireNamespace("broom", quietly = TRUE)) install.packages("broom")  
  
tidy(best\_mod) %>%  
 arrange(p.value) %>%  
 mutate(term = stringr::str\_replace\_all(term, ":", " × ")) %>%  
 print(n = Inf)

# A tibble: 8 × 5  
 term estimate std.error statistic p.value  
 <chr> <dbl> <dbl> <dbl> <dbl>  
1 (Intercept) 13.8 0.231 59.6 1.09e-263  
2 failure\_flag -3.06 0.305 -10.0 3.86e- 22  
3 school\_MS -1.31 0.235 -5.58 3.64e- 8  
4 studytime\_log\_z 0.437 0.112 3.91 1.02e- 4  
5 SES\_z 0.417 0.112 3.74 2.03e- 4  
6 schoolsup\_yes -1.26 0.351 -3.60 3.43e- 4  
7 Dalc -0.374 0.120 -3.11 1.98e- 3  
8 sex\_M -0.632 0.233 -2.71 6.87e- 3