regression

2025-12-05

```
# Load packages
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

Reading in the dataset:

```
dat <- read.csv("student-mat.csv", sep = ";")
str(dat)
```

```
## 'data.frame':    395 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 3 0 0 0 0 0 0 0 ...
##  $ schoolsup : chr  "yes" "no" "yes" "no" ...
##  $ famsup    : chr  "no" "yes" "no" "yes" ...
##  $ paid      : chr  "no" "no" "yes" "yes" ...
##  $ 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  6 4 10 2 4 10 0 6 0 0 ...
##  $ G1        : int  5 5 7 15 6 15 12 6 16 14 ...
##  $ G2        : int  6 5 8 14 10 15 12 5 18 15 ...
##  $ G3        : int  6 6 10 15 10 15 11 6 19 15 ...
```

Convert some variables to factors

```
dat <- dat %>%
  mutate(
    sex       = factor(sex),
    school    = factor(school),
    address   = factor(address),
    famsize   = factor(famsize),
    Pstatus   = factor(Pstatus),
    higher    = factor(higher),
    internet  = factor(internet),
    romantic  = factor(romantic),
    studytime = factor(
      studytime,
      levels = 1:4,
      labels = c("low", "moderate", "high", "very high")
    )
  )
```

model 1 - evaluate G3 just from background, support, behavior, etc

```
model1 <- lm(
  G3 ~ sex + age + Medu + Fedu +
    studytime + failures + absences +
    higher + goout + Walc + internet,
  data = dat
)

summary(model1)
```

```
##
## Call:
## lm(formula = G3 ~ sex + age + Medu + Fedu + studytime + failures +
##     absences + higher + goout + Walc + internet, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.2241  -2.0116   0.4896   2.8137   8.7655
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        11.56584    3.42278   3.379 0.000802 ***
## sexM                1.37284    0.47235   2.906 0.003870 **
## age                -0.21051    0.18045  -1.167 0.244091
## Medu                0.49122    0.25664   1.914 0.056365 .
## Fedu               -0.11465    0.25433  -0.451 0.652406
## studytimemoderate   0.07606    0.53510   0.142 0.887040
## studytimehigh       1.30893    0.73093   1.791 0.074124 .
```

```
## studytimevery high  0.55157    0.94906   0.581 0.561468
## failures           -1.78689    0.31490  -5.674 2.76e-08 ***
## absences            0.04123    0.02779   1.484 0.138686
## higheryes           1.64756    1.04281   1.580 0.114955
## goout              -0.45638    0.21340  -2.139 0.033104 *
## Walc                0.09027    0.19390   0.466 0.641807
## internetyes         0.44558    0.58660   0.760 0.447962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.178 on 381 degrees of freedom
## Multiple R-squared:  0.1959, Adjusted R-squared:  0.1685
## F-statistic: 7.141 on 13 and 381 DF,  p-value: 1.74e-12
```

```r
model2 <- lm(G3 ~ G1 + G2, data = dat)
summary(model2)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5713 -0.3888  0.2885  0.9725  3.7089
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.83001    0.33531  -5.458 8.57e-08 ***
## G1           0.15327    0.05618   2.728  0.00665 **
## G2           0.98687    0.04957  19.909  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.937 on 392 degrees of freedom
## Multiple R-squared:  0.8222, Adjusted R-squared:  0.8213
## F-statistic: 906.1 on 2 and 392 DF,  p-value: < 2.2e-16
```

```r
model3_full <- lm(
  G3 ~ G1 + G2 +
    sex + age + Medu + Fedu +
    studytime + failures + absences +
    higher + goout + Walc + internet,
  data = dat
)

summary(model3_full)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2 + sex + age + Medu + Fedu + studytime +
##     failures + absences + higher + goout + Walc + internet, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0225 -0.4694  0.2749  0.9955  3.6924
```

```
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.25742    1.58303   0.794 0.427511
## G1                   0.16100    0.05741   2.804 0.005302 **
## G2                   0.96538    0.05040  19.153  < 2e-16 ***
## sexM                 0.19145    0.21834   0.877 0.381128
## age                 -0.19603    0.08299  -2.362 0.018681 *
## Medu                 0.09254    0.11745   0.788 0.431278
## Fedu                -0.16847    0.11634  -1.448 0.148402
## studytimemoderate   -0.11125    0.24381  -0.456 0.648438
## studytimehigh        0.13159    0.33603   0.392 0.695578
## studytimevery high  -0.78727    0.43380  -1.815 0.070342 .
## failures            -0.25375    0.14943  -1.698 0.090297 .
## absences             0.04264    0.01266   3.368 0.000834 ***
## higheryes            0.24349    0.47667   0.511 0.609776
## goout                0.06398    0.09822   0.651 0.515153
## Walc                 0.06668    0.08887   0.750 0.453523
## internetyes         -0.29108    0.26863  -1.084 0.279252
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.903 on 379 degrees of freedom
## Multiple R-squared:  0.834,  Adjusted R-squared:  0.8274
## F-statistic:   127 on 15 and 379 DF,  p-value: < 2.2e-16
```

```r
model_step <- step(model3_full, direction = "both", trace = FALSE)
summary(model_step)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2 + age + failures + absences + Walc,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9517 -0.4347  0.2737  0.9542  3.7652
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.80205    1.35383   0.592  0.55390
## G1           0.16845    0.05628   2.993  0.00294 **
## G2           0.95808    0.04962  19.309  < 2e-16 ***
## age         -0.17375    0.07963  -2.182  0.02971 *
## failures    -0.21104    0.14266  -1.479  0.13987
## absences     0.03974    0.01224   3.247  0.00127 **
## Walc         0.11149    0.07635   1.460  0.14504
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.901 on 388 degrees of freedom
## Multiple R-squared:  0.8305, Adjusted R-squared:  0.8279
## F-statistic: 316.9 on 6 and 388 DF,  p-value: < 2.2e-16
```

```r
bin_vars <- c("schoolsup","famsup","paid","activities",
              "nursery","higher","internet","romantic")

for (v in bin_vars) {
  cat("\n", v, "\n")
  print(
    dat %>%
      group_by(.data[[v]]) %>%
      summarise(
        n = n(),
        mean_G3 = mean(G3),
        sd_G3   = sd(G3)
      )
  )
}
```

```
##
##  schoolsup
## # A tibble: 2 x 4
##   schoolsup     n mean_G3 sd_G3
##   <chr>     <int>   <dbl> <dbl>
## 1 no          344   10.6   4.77
## 2 yes          51    9.43  2.87
##
##  famsup
## # A tibble: 2 x 4
##   famsup     n mean_G3 sd_G3
##   <chr>  <int>   <dbl> <dbl>
## 1 no       153    10.6  4.64
## 2 yes      242    10.3  4.55
##
##  paid
## # A tibble: 2 x 4
##   paid      n mean_G3 sd_G3
##   <chr> <int>   <dbl> <dbl>
## 1 no      214    9.99  5.13
## 2 yes     181   10.9   3.79
##
##  activities
## # A tibble: 2 x 4
##   activities     n mean_G3 sd_G3
##   <chr>      <int>   <dbl> <dbl>
## 1 no           194    10.3  4.49
## 2 yes          201    10.5  4.68
##
##  nursery
## # A tibble: 2 x 4
##   nursery     n mean_G3 sd_G3
##   <chr>   <int>   <dbl> <dbl>
## 1 no         81    9.95  4.56
## 2 yes       314   10.5   4.59
##
##  higher
## # A tibble: 2 x 4
```

```
##   higher     n mean_G3 sd_G3
##   <fct>  <int>   <dbl> <dbl>
## 1 no        20     6.8  4.83
## 2 yes      375    10.6  4.49
##
##  internet
## # A tibble: 2 x 4
##   internet     n mean_G3 sd_G3
##   <fct>    <int>   <dbl> <dbl>
## 1 no          66    9.41  4.49
## 2 yes        329   10.6   4.58
##
##  romantic
## # A tibble: 2 x 4
##   romantic     n mean_G3 sd_G3
##   <fct>    <int>   <dbl> <dbl>
## 1 no         263   10.8   4.39
## 2 yes        132    9.58  4.86
```