HW 03 - Student Math Performance

Due: Thursday, Mar 05 at 11:59pm

Nathan Kim

Mar 5

Packages

```
library(tidyverse)
library(broom)
```

Data

```
math <- read_csv("data/math_performance.csv")
```

Exercises

Exercise 1

```
math <- math %>%
  mutate(
    medu = factor(medu),
    fedu = factor(fedu),
    traveltime = factor(traveltime),
    studytime = factor(studytime),
    famrel = factor(famrel),
    freetime = factor(freetime),
    goout = factor(goout),
    dalc = factor(dalc),
    walc = factor(walc),
    health = factor(health)
)
```

Exercise 2

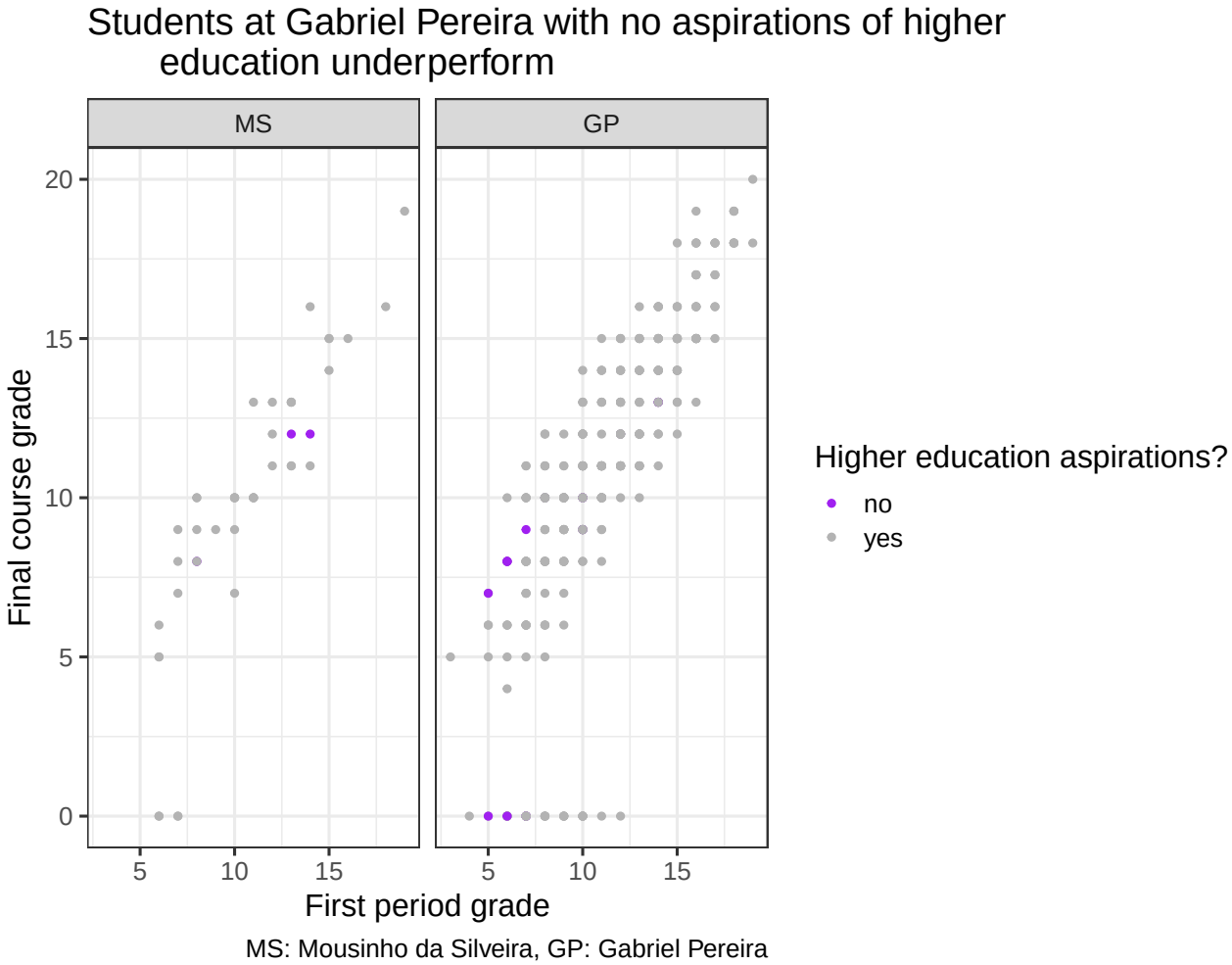
```
math <- math %>%
  mutate_if(is.character, factor)
```

Exercise 3

```
math <- math %>%
  mutate(school = fct_relevel(factor(school), "MS"))
```

Exercise 4

```
ggplot(data = math, mapping = aes(x = g1, y = g3, color = higher)) +
  geom_point() +
  scale_color_manual(values = c("purple", "grey70")) +
  facet_grid(.~ school) +
  theme_bw(base_size = 16) +
  labs(title = "Students at Gabriel Pereira with no aspirations of higher
       education underperform", x = "First period grade",
       y = "Final course grade",
       caption = "MS: Mousinho da Silveira, GP: Gabriel Pereira",
       color = "Higher education aspirations?")
```



Exercise 5

```
math %>%
  group_by(school) %>%
  count(school, g3) %>%
  mutate(final_failure_rate = n/sum(n)) %>%
  slice(1) %>%
  select(final_failure_rate)
```

```
# A tibble: 2 x 2
# Groups:   school [2]
  school final_failure_rate
  <fct>               <dbl>
1 MS                 0.0870
2 GP                 0.0974
```

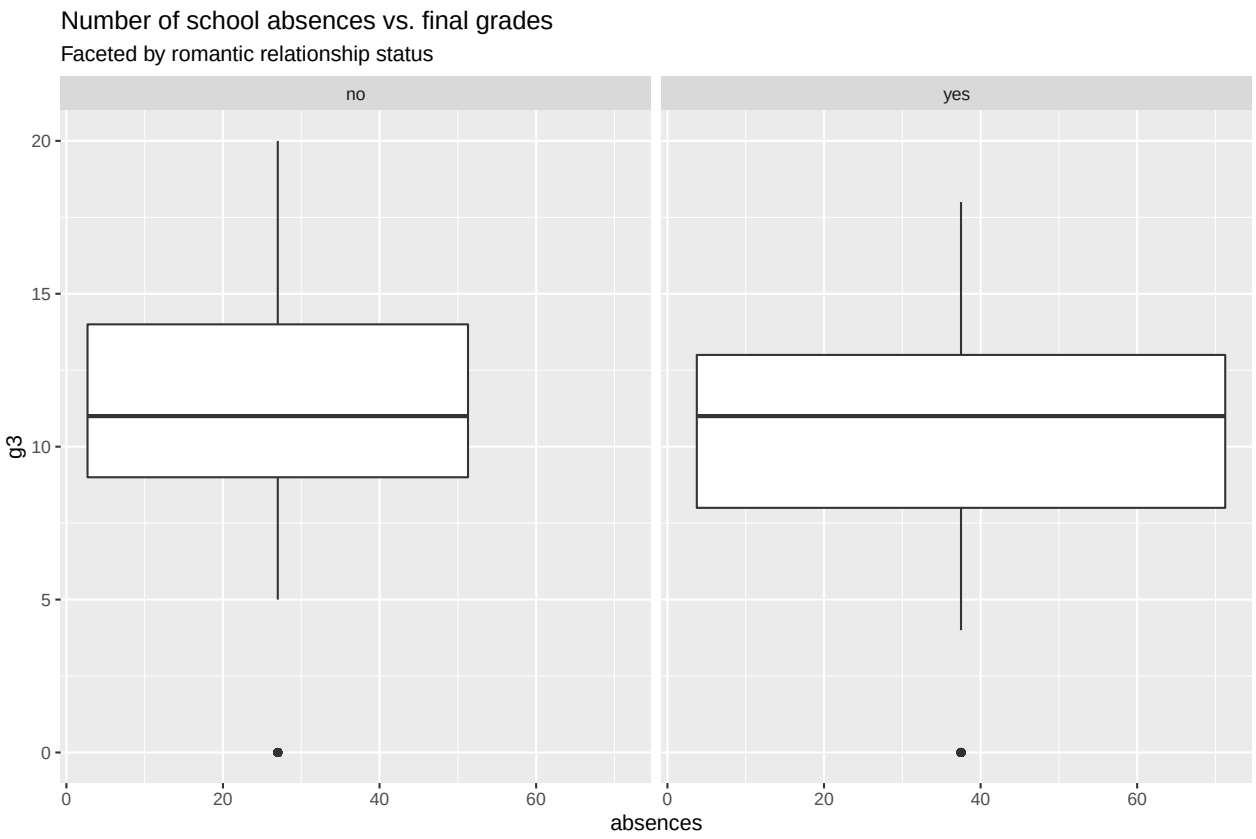
The final grade failure rate at Pereira High School is 9.74% whereas the final grade failure rate at Mousinho da Silveria High School is 8.70%

Exercise 6

Based on the analysis in exercise 5, Gabriel Pareira's failure rate (0.09742120) is 0.01046468 higher than that of Mounhiso da Silveira (0.08695652). Because Gabriel Pareira has a total of 9 students who failed compared to Mounhiso da Silveria that had 2 students that failed(based on the graphic from exercise 4), Gabriel Pareira seems to have a drastically higher failure rate than Mousinho da Silveria. However, this is misleading because according to the anlysis, Gabriel Pareira only has a higher failure rate than Mousinho da Silveria by 0.01046468, which is almost negligible. Gabriel Pereira only seems to have a drastically higher failure rate because it has more total students. Therefore, despite Gabriel Pareira seeming to have a higher failure rate based on the graph, it is not as drastic as it may initially seem.

Exercise 7

```
ggplot(math, mapping = aes(x = absences, y = g3)) +
  labs(title = "Number of school absences vs. final grades",
       subtitle = "Faceted by romantic relationship status") +
  facet_grid(. ~ romantic) +
  geom_boxplot(base_size = 20)
```



Exercise 8

```
passed_students <- math %>%
  filter(g3 > "0")
```

Exercise 9

```
lm1 <- lm(g3 ~ g1 + school, data = passed_students)
  tidy(lm1) %>%
    select(term, estimate)
```

```
# A tibble: 3 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)    0.966
2 g1             0.887
3 schoolGP       0.638
```

```
 glance(lm1) %>%
  pull(r.squared)
```

```
[1] 0.7993813
```

The linear model from this set is: Final Course Grade = 0.966 + 0.887(g1) + 0.638(schoolGP).

This linear model shows that while holding other variables constant, for each unit increase in first period grade, the final course grade is expected to increase by 0.887. For Gabriel Pereira, when holding other variables constant, the final course grade is expected to increase by 0.638. The value of R squared is 0.80. This means that roughly 80% of the variability in the final course grade can be explained by the student's first period grade and by their school.

The intercept, 0.966 signifies that students at the Mousinho da Silveria with a first period grade of 0 are expected to attain a final grade of 0.966.

Exercise 10

```
lm2 <- lm(g3 ~ g1 + school + famsize, data = passed_students)
tidy(lm2) %>%
select(term, estimate)
```

```
# A tibble: 4 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)  0.968
2 g1           0.887
3 schoolGP     0.637
4 famsizeLE3  -0.00513
```

The linear model for this dataset can be given by: Final Course Grade = 0.968 + 0.887(g1) + 0.637(schoolGP) - 0.005(famsizeLE3).

```
glance(lm2) %>%
  pull(r.squared)
```

```
[1] 0.7993818
```

This linear model explores that when holding other variables constant, for each unit increase in the first period grade, the final course grade is expected to increase by 0.887. For Gabriel Pereira, while holding other variables constant, the final course grade is expected to increase by 0.637. Lastly, when holding other variables constant, if the family size is less than or equal to 3, the course grade is expected to decrease, on average, by 0.005. The value of R squared is 0.80. This means that roughly 80% of the variability in the final course grade can be explained by the student's first period grade, if their school is Gabriel Pereira, and if their family size is less than or equal to three.

The intercept, 0.968 signifies that students at the MS school with a first period grade of 0 and a family size larger than three are expected to attain a final grade of 0.968.

Exercise 11

```
 glance(lm1) %>%
  pull(r.squared)
```

```
[1] 0.7993813
```

```
 glance(lm1) %>%
   pull(adj.r.squared)
```

```
[1] 0.7982478
```

```
  glance(lm2) %>%
    pull(r.squared)
```

```
[1] 0.7993818
```

```
  glance(lm2) %>%
   pull(adj.r.squared)
```

```
[1] 0.7976768
```

When comparing the R squared values to the adjusted r squared values for exercises 9 and 10, it both adjusted r squared values are slightly lower than the original r squared values. This occurred because the predictor improved the model less than what was actually predicted by chance.

Exercise 12

```
lm_interact <- lm(g3 ~ g1 * school, data = passed_students)
  tidy(lm_interact) %>%
    select(term, estimate)
```

```
# A tibble: 4 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)   1.12
2 g1            0.873
3 schoolGP      0.466
4 g1:schoolGP   0.0155
```

The linear model for this equation can be given by g3 = 1.12 + 0.873(g1) + 0.466(schoolGP) + 0.016(g1)(schoolGP)

Exercise 13

```
lm_interact <- lm(g3 ~ g1 + absences + school + sex + age + internet + pstatus,
                  data = passed_students)

step_model <- step(lm_interact, direction = c("backward"), results = "hide",
trace = 0)

step_model %>%
tidy() %>%
  select(term, estimate)
```

```
# A tibble: 7 x 2
  term        estimate
  <chr>          <dbl>
1 (Intercept)   4.56
2 g1            0.869
3 absences     -0.0423
4 schoolGP      0.404
5 age          -0.175
6 internetyes   0.427
7 pstatusT     -0.395
```

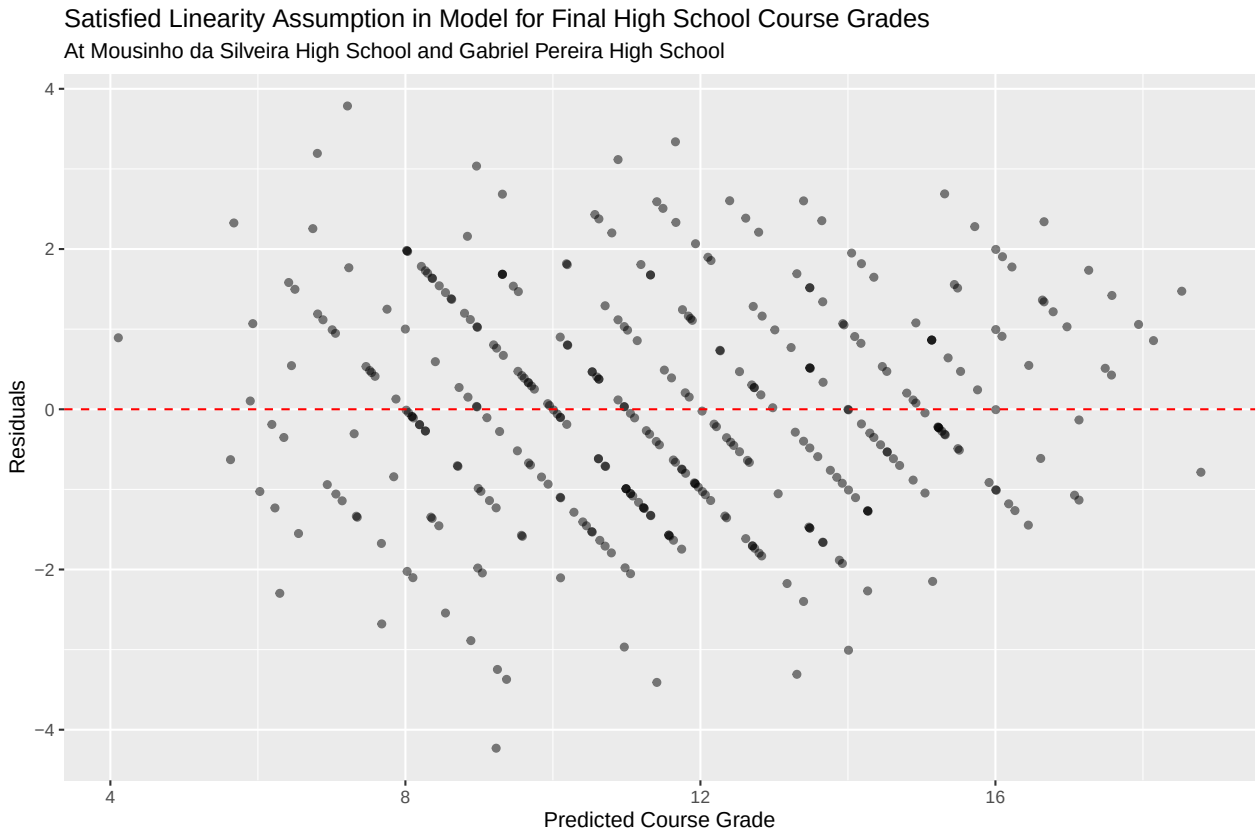
Exercise 14

While holding other variables constant, for each unit increase in the first period grade, the final course grade is expected to increase by 0.869. Furthermore, while holding other variables constant, for each unit increase in absence, the final course grade is expected to decrease on average by a value of 0.042.

Exercise 15

```
step_model_aug <- augment(step_model)

ggplot(step_model_aug, mapping = aes(x = .fitted, y = .resid)) +
  geom_point(alpha = 0.5) +
  geom_hline(yintercept = 0, color = "red", lty = 2) +
  labs(title = "Satisfied Linearity Assumption in Model for Final High School Course Grades",
       subtitle = "At Mousinho da Silveira High School and Gabriel Pereira High School",
       x = "Predicted Course Grade",
       y = "Residuals")
```



Yes, the linearity assumption is fulfilled. The observed residuals are normally distributed, and randomized. The only reason that they appear linear is because the prediction variable is not continuous. Additionally, the points do not favor particularly above or below the y-intercept of 0. They are relatively equal showing there is no favorability between positive and negative.

Exercise 16

```
passed_students %>%
  group_by(school) %>%
  summarize(median= median(g1))
```

```
# A tibble: 2 x 2
  school median
  <fct>   <dbl>
1 MS         11
2 GP         11
```

```
A <- tibble(g1 = 11, age = 18, school = "GP", internet = "yes", pstatus = "T", absences = 12 )
B <- tibble(g1 = 11, age = 17, school = "MS", internet = "yes", pstatus = "A", absences = 11 )
C <- tibble(g1 = 11, age = 18, school = "MS", internet = "no", pstatus = "A", absences = 23 )

augment(step_model, newdata = A) %>%
  select(.fitted)
```

```
# A tibble: 1 x 1
  .fitted
    <dbl>
1    10.9
```

```
augment(step_model, newdata = B) %>%
  select(.fitted)
```

```
# A tibble: 1 x 1
  .fitted
    <dbl>
1    11.1
```

```
augment(step_model, newdata = C) %>%
  select(.fitted)
```

```
# A tibble: 1 x 1
  .fitted
    <dbl>
1    9.99
```

Student A is expected to have a final grade of 10.89.

Student B is expected to have a final grade of 11.10.

Student C is expected to have a final grade of 9.99.