Does a student's GPA rise or fall as the student moves from freshman year to senior year?

Kate Dao

Load Libraries

```
library(dplyr)
library(data.table)
library(ggplot2)
library(kableExtra)
library(haven)
library(car)
rm(list=ls())
```

Load the datasset

```
df <- read_stata("/Users/katedao/Desktop/college.dta")
```

a.

(i) What is the average GPA in the sample?

```
average_gpa <- mean(df$gpa, na.rm = TRUE)
print(average_gpa)
```

```
## [1] 3.238869
```

(ii) How many men are in the sample? How many women?

```
total_men <- sum(df$male == 1, na.rm = TRUE)
print(total_men)
```

```
## [1] 3467
```

```
total_women <- sum(df$male == 0, na.rm = TRUE)
print(total_women)
```

```
## [1] 6423
```

(iii) What share of students are members of a fraternity or sorority?

```
frat_share <- mean(df$fraternity, na.rm = TRUE)
print(frat_share * 100)
```

```
## [1] 12.32558
```

(iv) What share of students work during the school year?

```
work_share <- mean(df$work, na.rm = TRUE)
print(work_share)
```

```
## [1] 0.6471183
```

(v) What share of students report that they used marijuana in the past 30 days?

```
marijuana_share <- mean(df$marijuana, na.rm = TRUE)
print(marijuana_share * 100)
```

```
## [1] 16.39029
```

b. How many students did not report their GPA? Is this non-reporting likely to be random? Why or why not?

```
na_gpa <- sum(is.na(df$gpa))
print(na_gpa)
```

```
## [1] 144
```

The non-reporting of GPA is unlikely to be random because students with lower academic performance may intentionally leave it blank, while high-achieving students are more likely to report their GPA.

c. Estimate the regression of GPA on male and work. Interpret the regression coefficients (including the intercept).

```
model <- lm(gpa ~ male + work, data = df)
summary(model)
```

```
##
## Call:
## lm(formula = gpa ~ male + work, data = df)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61588 -0.28258  0.06195  0.39535  0.83251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.28258    0.01062 309.194   <2e-16 ***
## male        -0.10386    0.01190  -8.730   <2e-16 ***
## work        -0.01123    0.01189  -0.945    0.345
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5594 on 9743 degrees of freedom
##   (144 observations deleted due to missingness)
## Multiple R-squared:  0.00778,    Adjusted R-squared:  0.007576
## F-statistic:  38.2 on 2 and 9743 DF,  p-value: < 2.2e-16
```

The intercept 3.28258 represents the predicted GPA for a female student who does not work. The coefficient for male is -0.10386, meaning that male students have lower GPA than female students on average.The coefficient for work is -0.01123, meaning that each additional hour of work per week is associated with a slight decrease in GPA.

d. Estimate the regression of GPA on freshman, sophomore, junior and senior for men only. Are all regression coefficients reported in the results? Explain what happened. What is the solution?

The regression failed to compute valid coefficients and returned NaN values for all estimates. This suggests an issue with the dataset.

Not all regression coefficients are reported in the results. The issue is caused by perfect multicollinearity in the dataset. The variables freshman, sophomore, junior, and senior represent mutually exclusive categories where each student belongs to exactly one category. Since one of these variables can be perfectly predicted by the others, OLS cannot estimate all four coefficients simultaneously.

To correctly estimate the regression of GPA on freshman, sophomore, junior, and senior for men only, we must be drop one category to fix the issue.

e. Estimate the regression of GPA on sophomore, junior and senior for men only. What is the interpretation of all coefficients in this regression (including the intercept)?

```r
df_male <- subset(df, male == 1)

df_male <- na.omit(df_male[, c("gpa", "sophomore", "junior", "senior")])

model <- lm(gpa ~ sophomore + junior + senior, data = df_male)

summary(model)
```

```
##
## Call:
```

```
## lm(formula = gpa ~ sophomore + junior + senior, data = df_male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58672 -0.44183  0.07988  0.48726  0.89147
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.10853    0.02019 153.990  < 2e-16 ***
## sophomore    0.03514    0.02853   1.232  0.21811
## junior       0.07091    0.02751   2.577  0.00999 **
## senior       0.14489    0.02844   5.095 3.67e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.577 on 3415 degrees of freedom
## Multiple R-squared:  0.008289,   Adjusted R-squared:  0.007418
## F-statistic: 9.515 on 3 and 3415 DF,  p-value: 2.951e-06
```

The intercept 3.10853 is the predicted GPA for a freshman male student. Sophomore is associated with a 0.03514 increase in GPA compared to freshmen.However, this difference is not statistically significant, p = 0.21811, meaning it could be due to random variation.Juniors have a 0.07001 higher GPA than freshmen, on average. This effect is statistically significant (p = 0.00999), suggesting that juniors tend to perform better than freshmen. Seniors have a 0.1449 higher GPA than freshmen, on average. This effect is highly significant ($p < 0.001$), meaning seniors consistently perform better than freshmen.

f. For the regression in (e), test the null hypothesis that there is no difference in the GPA of sophomores and juniors. What is the number of restrictions q for this test?

To test the null hypothesis:

$$H_0 : \beta_{sophomore} = \beta_{junior}$$
$$H_A : \beta_{sophomore} \neq \beta_{junior}$$

```
df_male <- subset(df, male == 1)

df_male <- na.omit(df_male[, c("gpa", "sophomore", "junior", "senior")])

model <- lm(gpa ~ sophomore + junior + senior, data = df_male)

f_test <- linearHypothesis(model, "sophomore = junior")

print(f_test)
```

```
## Linear hypothesis test
##
## Hypothesis:
## sophomore - junior = 0
##
## Model 1: restricted model
```

```
## Model 2: gpa ~ sophomore + junior + senior
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   3416 1137.5
## 2   3415 1136.9  1   0.56336 1.6922 0.1934
```

In the f-test for testing whether there is no difference in GPA between sophomores and juniors. This means we have one restriction in our test because we are testing a single equality constraint.q represents the number of independent restrictions placed on the regression model. Since we are only testing one equation, the number of restrictions is 1.

g. For the regression in (e), test the null hypothesis that the coefficients on sophomore, junior and senior are all zero, against the alternative that at least one coefficient is nonzero. State clearly the significance level of the test you are using and the critical value of the F statistic for this test.

We want to test the null hypothesis:

$$H_0 : \beta_{sophomore} = \beta_{junior} = \beta_{senior} = 0$$

against the alternative that at least one of these coefficients is nonzero.

```
df_male <- na.omit(df_male[, c("gpa", "sophomore", "junior", "senior")])

model <- lm(gpa ~ sophomore + junior + senior, data = df_male)

f_test <- linearHypothesis(model, c("sophomore = 0", "junior = 0", "senior = 0"))

print(f_test)
```

```
## Linear hypothesis test
##
## Hypothesis:
## sophomore = 0
## junior = 0
## senior = 0
##
## Model 1: restricted model
## Model 2: gpa ~ sophomore + junior + senior
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3418 1146.4
## 2   3415 1136.9  3    9.5031 9.5147 2.951e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Significance Level:

The standard significance level used for hypothesis testing is: 0.05 (5%). This means we are conducting the test at a 95% confidence level. If the p-value from the F-test is less than 0.05, we reject the null hypothesis and conclude that at least one of the coefficients is statistically significant.

- Critical Value of the F-Statistic:

```
alpha <- 0.05

df_numerator <- 3
df_denominator <- model$df.residual

critical_value <- qf(1 - alpha, df_numerator, df_denominator)

print(critical_value)
```

```
## [1] 2.60751
```

h. Estimate the regression of GPA on age, sophomore, junior and senior for men only. What happens to the statistical significance of the coefficients in this regression? Explain why this is the case.

```
df_male <- subset(df, male == 1)

df_male <- na.omit(df_male[, c("gpa", "age", "sophomore", "junior", "senior")])

model <- lm(gpa ~ age + sophomore + junior + senior, data = df_male)

summary(model)
```

```
##
## Call:
## lm(formula = gpa ~ age + sophomore + junior + senior, data = df_male)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.61488 -0.44215  0.08411  0.49217  0.91275
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.903715   0.126469  22.960  < 2e-16 ***
## age         0.010796   0.006581   1.640  0.10099
## sophomore   0.021988   0.029630   0.742  0.45809
## junior      0.044092   0.031995   1.378  0.16826
## senior      0.107955   0.036264   2.977  0.00293 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5769 on 3414 degrees of freedom
## Multiple R-squared:  0.00907,    Adjusted R-squared:  0.007909
## F-statistic: 7.812 on 4 and 3414 DF,  p-value: 2.898e-06
```

When we add "age" as an additional independent variable to the regression of GPA on class year (sophomore, junior, senior) for men only, we may observe changes in the statistical significance of the coefficients. The key reason is multicollinearity, meaning that age and class Year (sophomore, junior, senior) are highly correlated.

i. Estimate the same regression as in part (e), but now do this for all respondents (male and female). Include controls for male, work, marijuana, lightdrinker, moddrinker and heavy drinker. Calculate the predicted GPA for a male senior who works, is a moderate drinker and has not smoked marijuana in the past 30 days.

```
df_clean <- na.omit(df[, c("gpa", "sophomore", "junior", "senior", "male", "work", "marijuana", "lightdr

model <- lm(gpa ~ sophomore + junior + senior + male + work + marijuana + lightdrinker + moddrinker + he

summary(model)
```

```
##
## Call:
## lm(formula = gpa ~ sophomore + junior + senior + male + work +
##     marijuana + lightdrinker + moddrinker + heavydrinker, data = df_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75810 -0.33835  0.04664  0.42588  1.16512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.28713    0.01650 199.238  < 2e-16 ***
## sophomore     0.06627    0.01631   4.064 4.87e-05 ***
## junior        0.09059    0.01585   5.717 1.12e-08 ***
## senior        0.17226    0.01627  10.590  < 2e-16 ***
## male         -0.09273    0.01188  -7.809 6.37e-15 ***
## work         -0.03459    0.01196  -2.893  0.00382 **
## marijuana    -0.08909    0.01621  -5.496 3.97e-08 ***
## lightdrinker -0.05647    0.01433  -3.940 8.19e-05 ***
## moddrinker   -0.11258    0.01736  -6.486 9.22e-11 ***
## heavydrinker -0.23584    0.03572  -6.603 4.23e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5525 on 9736 degrees of freedom
## Multiple R-squared:  0.0328, Adjusted R-squared:  0.0319
## F-statistic: 36.68 on 9 and 9736 DF,  p-value: < 2.2e-16
```

```
new_data <- data.frame(
  sophomore = 0,   # Not a sophomore
  junior = 0,      # Not a junior
  senior = 1,      # Is a senior
  male = 1,        # Is male
  work = 1,        # Works
  marijuana = 0,   # Has not smoked marijuana
  lightdrinker = 0, # Not a light drinker
  moddrinker = 1,  # Is a moderate drinker
  heavydrinker = 0 # Not a heavy drinker
)
```

```r
predicted_gpa <- predict(model, new_data)

print(predicted_gpa)
```

```
##       1
## 3.21949
```

j. Using the regression in part (i), test the hypothesis (at the 5% significance level) that freshmen and sophomores have the same GPA on average, holding constant gender, work, type of drinking and marijuana use.

```r
f_test <- linearHypothesis(model, "sophomore = 0")

print(f_test)
```

```
## Linear hypothesis test
##
## Hypothesis:
## sophomore = 0
##
## Model 1: restricted model
## Model 2: gpa ~ sophomore + junior + senior + male + work + marijuana +
##     lightdrinker + moddrinker + heavydrinker
##
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   9737 2977.5
## 2   9736 2972.5  1    5.0419 16.514 4.867e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

k. What is the adjusted R2 for the regression in part (i)? What does this adjusted R2 tell you about the fit of this regression? Does it indicate that omitted variable bias is likely to be a problem?

```r
summary(model)$adj.r.squared
```

```
## [1] 0.03190279
```

As the adjusted R2 < 0.5, the model does not explain much variation in GPA. A low adjusted R2 could suggest that important variables affecting GPA (such as study habits, SAT scores, parental education, etc.) are missing.