Simple Linear Regression

Yelsh Gebreselassie

2022-09-21

Contents

In this document, we do basic descriptive and regression analysis to understand what variables best predict freshman GPA.

1. First install/ load the R packages we need

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

gpa.data <- read_csv("data/satgpa.csv")
attach(gpa.data)
```

Exploratory questions

How well do SAT scores correlate with freshman GPA?

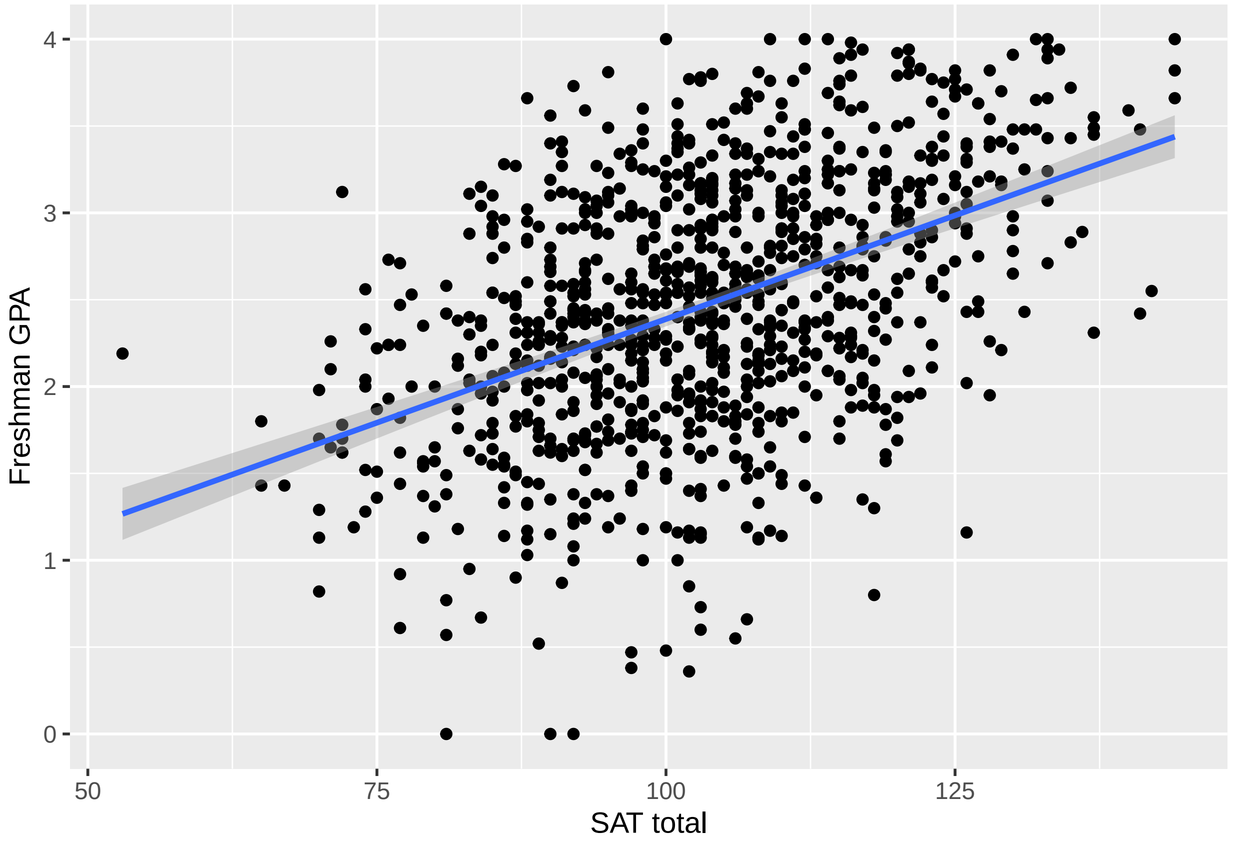
```
cor(gpa_fy, sat_total)
```

```
## [1] 0.460281
```

A correlation of -1 means perfect negative correlation. A correlation of 0 means, no correlation between the two. And a correlation of 1 means perfect positive correlation. The above result shows a positive correlation between SAT scores and freshman GPA.But it is not very strong, meaning close to 1.

```
ggplot(data = gpa.data, mapping = aes(x = sat_total, y = gpa_fy)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(y = "Freshman GPA", x = "SAT total")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



The above plot shows the positive correlation between SAT scores and freshman GPA.

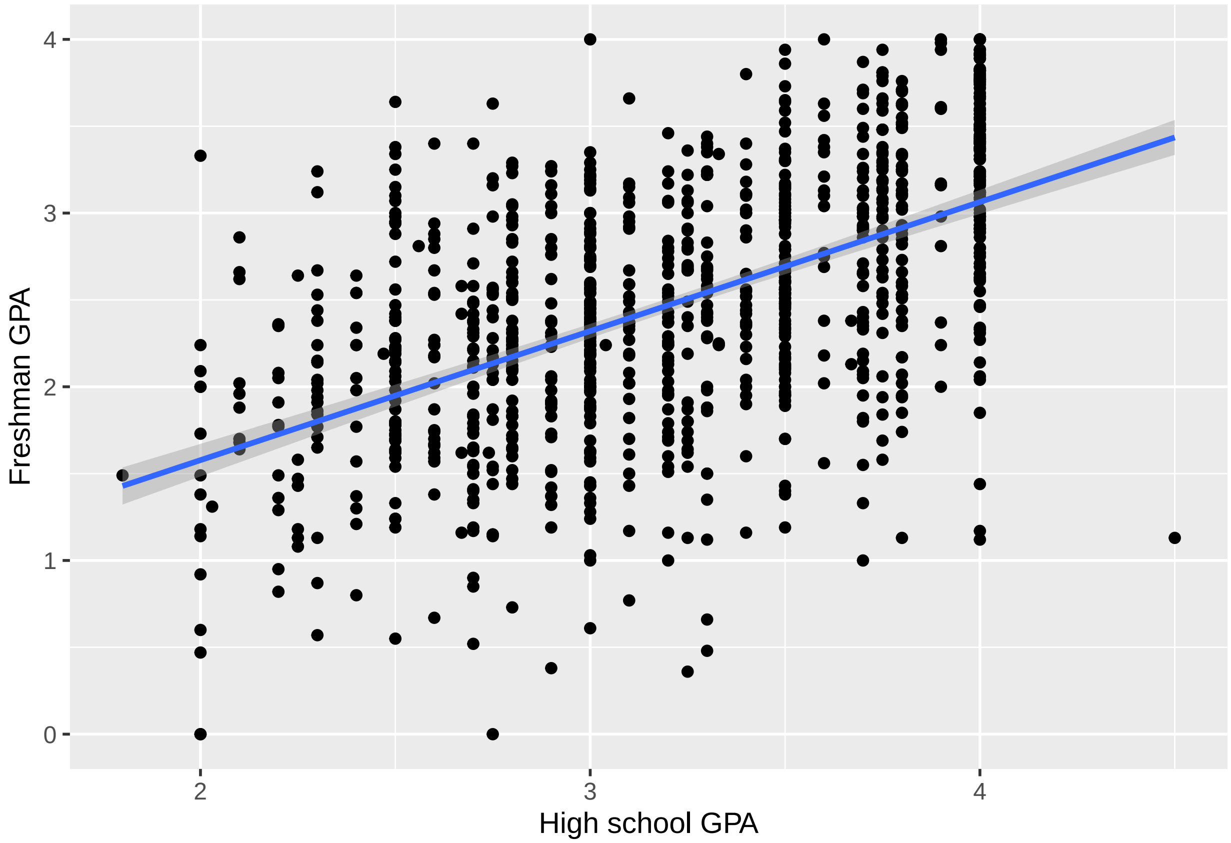
How well do high school GPA correlate with freshman GPA?

```
cor(gpa_fy, gpa_hs)
```

```
## [1] 0.5433535
```

```
ggplot(data = gpa.data, mapping = aes(y = gpa_fy, x = gpa_hs)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(x = "High school GPA", y = "Freshman GPA")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



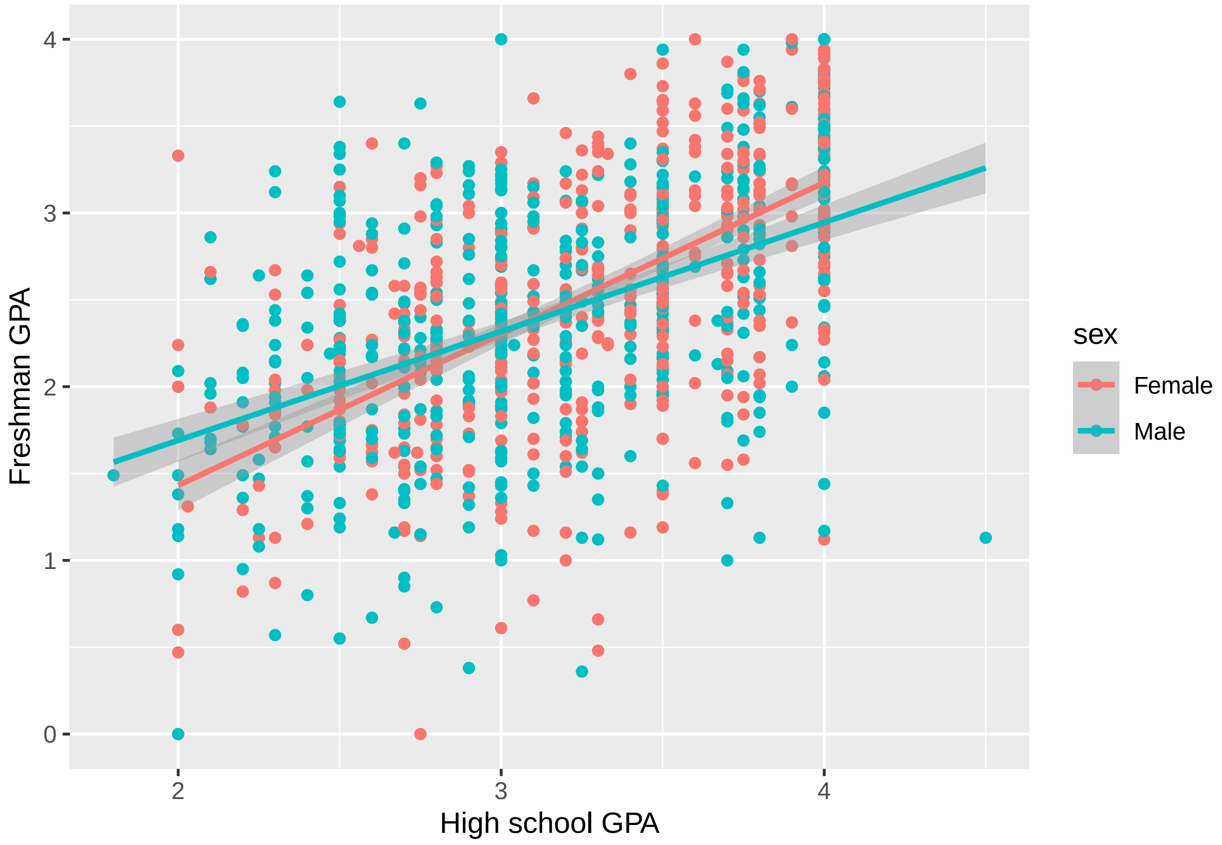
Is the correlation between SAT scores and freshman GPA stronger for men or for women?

```
gpa.data %>%
  group_by(sex) %>%
  summarize(correlation = cor(sat_total, gpa_fy))
```

```
## # A tibble: 2 x 2
##   sex    correlation
##   <chr>        <dbl>
## 1 Female       0.493
## 2 Male         0.481
```

```
ggplot(data = gpa.data, mapping = aes(y = gpa_fy, x = gpa_hs, color = sex)) +
  geom_point() +
  geom_smooth(method = lm) +
  labs(x = "High school GPA", y = "Freshman GPA")
```

```
## ‘geom_smooth()‘ using formula ’y ~ x’
```



Models

Do SAT scores predict freshman GPAs?

- X = SAT scores
- Y = Freshman GPA

```
model_simple <- lm(gpa_fy ~ sat_total, data = gpa.data)
tidy(model_simple, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  0.00193   0.152      0.0127 9.90e- 1  -0.296     0.300
## 2 sat_total    0.0239    0.00146   16.4    1.39e-53   0.0210    0.0267
```

```
glance(model_simple)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.212        0.211 0.658    268. 1.39e-53     1  -999. 2005. 2019.    432.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Does a certain type of SAT score have a larger effect on freshman GPAs?

```
model_sat <- lm(gpa_fy ~ sat_verbal + sat_math, data = gpa.data)
tidy(model_sat, conf.int = TRUE)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  0.00737   0.152      0.0484 9.61e- 1  -0.291     0.306
## 2 sat_verbal   0.0254    0.00286    8.88   3.07e-18   0.0198    0.0310
## 3 sat_math     0.0224    0.00279    8.04   2.58e-15   0.0169    0.0279
```

```
glance(model_sat)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.212        0.211 0.658    134. 2.36e-52     2  -999. 2006. 2026.    432.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Do high school GPAs predict freshman GPAs?

```
model_hs <- lm(gpa_fy ~ gpa_hs, data = gpa.data)
tidy(model_hs, conf.int = TRUE)
```

```
## # A tibble: 2 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)   0.0913    0.118      0.775 4.39e- 1  -0.140     0.323
## 2 gpa_hs        0.743     0.0363    20.4   6.93e-78    0.672     0.814
```

```
glance(model_hs)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.295        0.295 0.622    418. 6.93e-78     1  -943. 1893. 1908.    386.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Explaining College GPA using SAT scores and sex

```r
model_sat_sex <- lm(gpa_fy ~ sat_total + sex, data = gpa.data)
tidy(model_sat_sex, conf.int = TRUE)
```

```
## # A tibble: 3 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  -0.0269   0.149      -0.181 8.57e- 1  -0.319     0.265
## 2 sat_total     0.0255   0.00145    17.6   1.14e-60   0.0227    0.0284
## 3 sexMale      -0.274    0.0414     -6.62  6.05e-11  -0.355    -0.193
```

```r
glance(model_sat_sex)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.245        0.243 0.644    162. 1.44e-61     2  -978. 1964. 1983.    414.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Explain college GPA using SAT scores, high school GPA and sex

```r
model_final <- lm(gpa_fy ~ sat_verbal + sat_math + gpa_hs + sex,
                  data = gpa.data)
tidy(model_final, conf.int = TRUE)
```

```
## # A tibble: 5 x 7
##   term        estimate std.error statistic  p.value conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
## 1 (Intercept)  -0.835    0.149       -5.62 2.49e- 8  -1.13     -0.544
## 2 sat_verbal    0.0161   0.00263      6.12 1.32e- 9   0.0110    0.0213
## 3 sat_math      0.0155   0.00273      5.68 1.78e- 8   0.0102    0.0209
## 4 gpa_hs        0.545    0.0395      13.8  9.55e-40   0.467     0.623
## 5 sexMale      -0.142    0.0401      -3.54 4.20e- 4  -0.220    -0.0632
```

```r
glance(model_final)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squa~1 sigma stati~2  p.value    df logLik   AIC   BIC devia~3
##       <dbl>        <dbl> <dbl>   <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
## 1     0.367        0.364 0.591    144. 3.98e-97     4  -890. 1792. 1822.    347.
## # ... with 2 more variables: df.residual <int>, nobs <int>, and abbreviated
## #   variable names 1: adj.r.squared, 2: statistic, 3: deviance
```

Which model best predicts freshman GPA?

	Model 1	Model 2	Model 3	Model 4	Model 5
(Intercept)	0.002	0.007	0.091	−0.027	−0.835
	(0.152)	(0.152)	(0.118)	(0.149)	(0.149)
sat_total	0.024			0.026	
	(0.001)			(0.001)	
sat_verbal		0.025			0.016
		(0.003)			(0.003)
sat_math		0.022			0.016
		(0.003)			(0.003)
gpa_hs			0.743		0.545
			(0.036)		(0.040)
sexMale				−0.274	−0.142
				(0.041)	(0.040)
Num.Obs.	1000	1000	1000	1000	1000
R2	0.212	0.212	0.295	0.245	0.367
R2 Adj.	0.211	0.211	0.295	0.243	0.364
AIC	2004.8	2006.4	1893.0	1963.8	1792.2
BIC	2019.5	2026.0	1907.7	1983.4	1821.6
Log.Lik.	−999.382	−999.189	−943.477	−977.904	−890.098
RMSE	0.66	0.66	0.62	0.64	0.59

```
modelsummary(list(model_simple, model_sat, model_hs,
                  model_sat_sex, model_final))
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 2 > 1' in coercion to 'logical(1)'
```