Simple Linear Regression: First Year College GPA

Solutions

Predicting Success in College

During the college admissions process, many factors are considered to assess whether an applicant will be successful in college. Specifically, these factors are assumed to be good determinants of how an applicant will perform in the first year of college (assessed by GPA). Common metrics used to predict first year GPA include high school GPA and SAT score. In small groups, you will explore the relationships between first year GPA (units: first year GPA points) and high school GPA (units: high school GPA points), and first year GPA and SAT score (units: SAT points). The data set you will be using is available through the openintro library.

```
## load packages
library(ggplot2)
library(openintro)
```

```
## Please visit openintro.org for free statistics materials
```

```
##
## Attaching package: 'openintro'
```

```
## The following object is masked from 'package:ggplot2':
##
##     diamonds
```

```
## The following objects are masked from 'package:datasets':
##
##     cars, trees
```

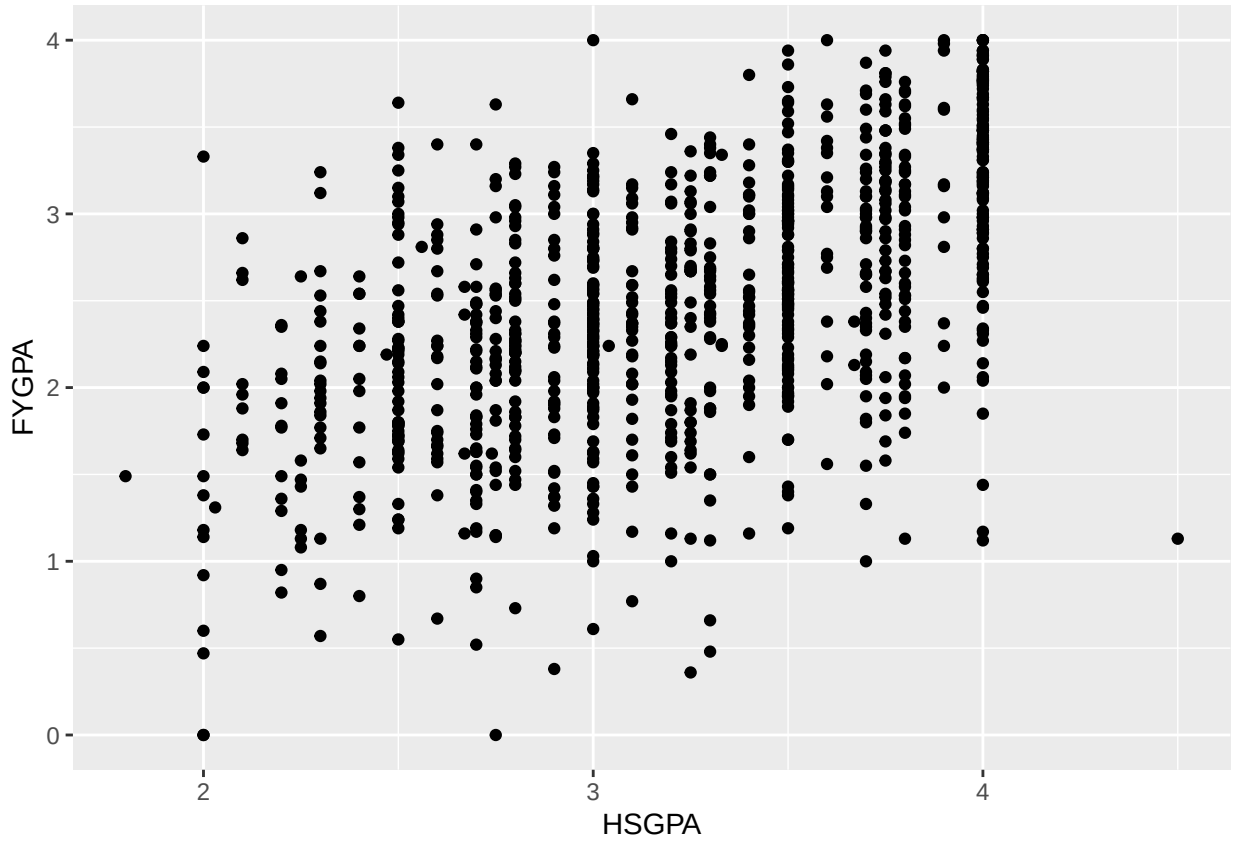
```
### The name of the dataset you need is satGPA. This will load with openintro.
head(satGPA)
```

```
##   sex SATV SATM SATSum HSGPA FYGPA
## 1   1   65   62    127  3.40  3.18
## 2   2   58   64    122  4.00  3.33
## 3   2   56   60    116  3.75  3.25
## 4   1   42   53     95  3.75  2.42
## 5   1   55   52    107  4.00  2.63
## 6   2   55   56    111  4.00  2.91
```

Model the relationship between first year college GPA and high school GPA.

Make an appropriate graph to look at the relationship between your two variables. Identify the response variable and the predictor. Verbally describe the relationship between the two variables.

```
ggplot(data = satGPA, aes(x = HSGPA, y = FYGPA)) + geom_point()
```



There appears to be a moderate linear relationship between high school GPA and first year GPA. There is also a potential outlier - a high school GPA of 4.5 was reported, but seems unlikely. This could be a data entry error. If possible, you would want to check with the person who collected the data to see if this is legitimate.

Find the correlation between the response and the predictor. Is is positive or negative?

```
R <- cor(satGPA$HSGPA, satGPA$FYGPA)
R
```

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

The correlation is positive.

Calculate the five summary statistics necessary to estimate β_0 and β_1. Save them as `x_bar`, `y_bar`, `s_x`, `s_y`, and `R`.

```
x_bar <- mean(satGPA$HSGPA)
y_bar <- mean(satGPA$FYGPA)
s_x <- sd(satGPA$HSGPA)
s_y <- sd(satGPA$FYGPA)
## already found R above, so I am not going to repeat that code
```

Then calculate these estimates (i.e. $\hat{\beta}_0$ and $\hat{\beta}_1$). Save them as `b0` and `b1`, respectively. What are the units for each of these estimates (see problem description)?

```
b1 <- s_y/s_x*R
b0 <- y_bar-b1*x_bar

b1
```

```
## [1] 0.7431385
```

```
b0
```

```
## [1] 0.09131887
```

$\hat{\beta}_0 = 0.091$ first year GPA points

$\hat{\beta}_1 = 0.743$ first year GPA points/high school GPA point

Fit the linear model for your problem. Be sure to identify the response and predictor correctly. Be sure to assign this a descriptive name (e.g. `lm_FYGPA_HSGPA`).

```
lm_FYGPA_HSGPA <- lm(data=satGPA, FYGPA ~ HSGPA)
```

Use the summary function to look at a summary of your model fit. Compare the estimates of β_0 and β_1 that you get from this summary to those that you calculated in (c).

```
summary(lm_FYGPA_HSGPA)
```

```
##
## Call:
## lm(formula = FYGPA ~ HSGPA, data = satGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30544 -0.37417  0.03936  0.41912  1.75240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.09132    0.11789   0.775    0.439
## HSGPA        0.74314    0.03635  20.447   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6222 on 998 degrees of freedom
## Multiple R-squared:  0.2952, Adjusted R-squared:  0.2945
## F-statistic: 418.1 on 1 and 998 DF,  p-value: < 2.2e-16
```

Interpret your estimates of intercept and slope in the context of the problem. Be sure to consider units.

Intercept: For a high school GPA of 0, we predict a first year college GPA of 0.091 (first year GPA) points.

Slope: For a 1 point increase in high school GPA, we predict an increase in first year GPA of 0.743 points.

Is it appropriate to interpret the intercept in this case? Why or why not? What about the slope?

It is not appropriate to interpret the intercept, because we do not have any high school GPAs that are 0. Interpretting the intercept in this particular example would be extrapolation and should be avoided.

It is appropriate to interpret the slope. In general, interpretting this is going to be fine.

Is the linearity condition for least squares regression satisfied? To make a residual plot, you will need the following data frame (this will run after you have completed the previous steps of this lab). When you are ready, you can uncomment the relevant lines.

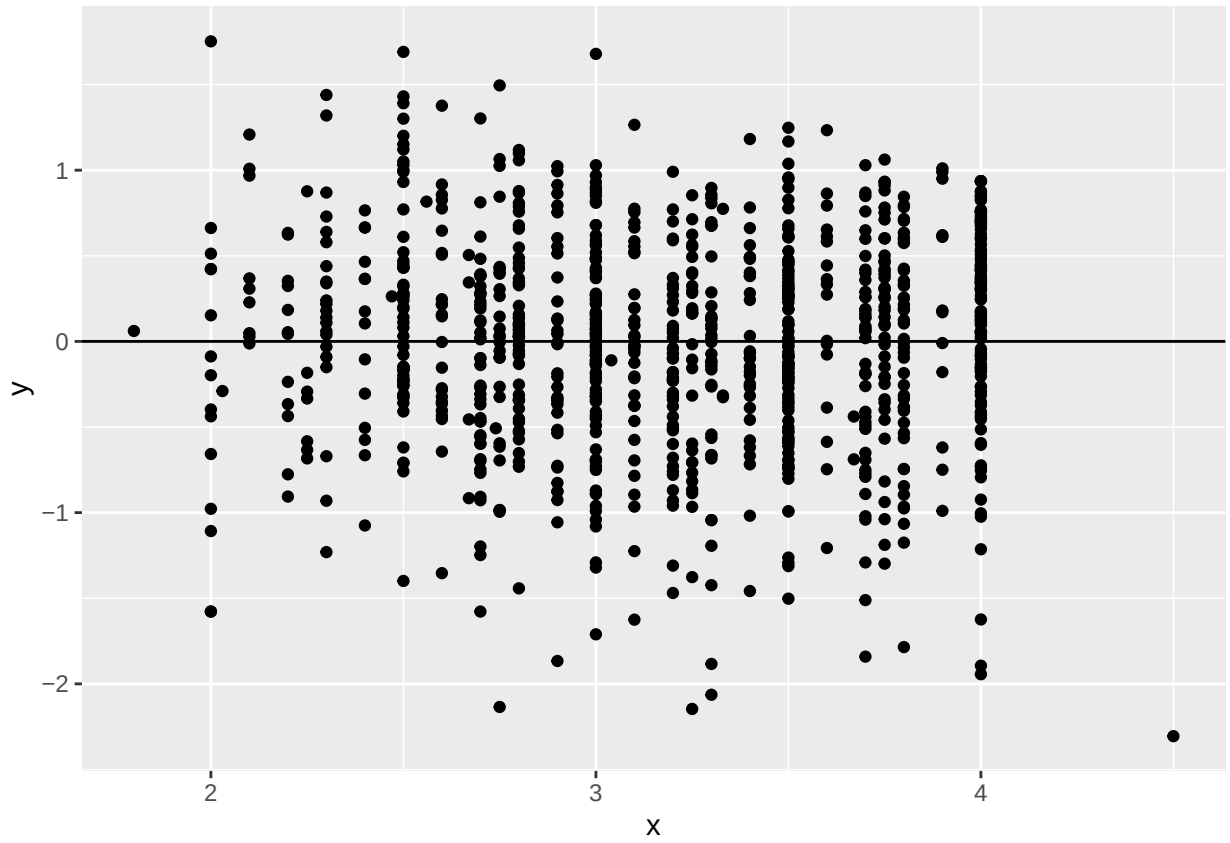
Reminder: the basic anatomy of a plot made in ggplot is as follows:

```
## Skeleton code - should not run anything
(ggplot(data=<name of data frame>,
        aes(x=<variable for x axis>, y=<variable for y axis>,
            color = <variable for color lines>,
            fill= <variable for color area>))
        + geom_<geometry type>()
        + <optional other things like axis labels, ...>)
```

```
## Make data frame for residual plot Uncomment the following two lines when
## you are ready to do this part.

resid_df_m1 <- data.frame(x = satGPA$HSGPA, y = lm_FYGPA_HSGPA$residuals)

## Use ggplot to make residual plot
ggplot(data = resid_df_m1, aes(x = x, y = y)) + geom_point() + geom_hline(yintercept = 0)
```



What is the R^2 for your model? Interpret this value in the context of the problem.

Based on the summary output, we can see that $R^2 = 0.295$. This means that `29.5%` of the variability in FYGPA can be explained by high school GPA.

Model the relationship between first year college GPA and SAT score (SATSum).

Repeat the steps you took in the previous section, this time using SAT score as the predictor.

Solutions would be similar, so I am not including them.

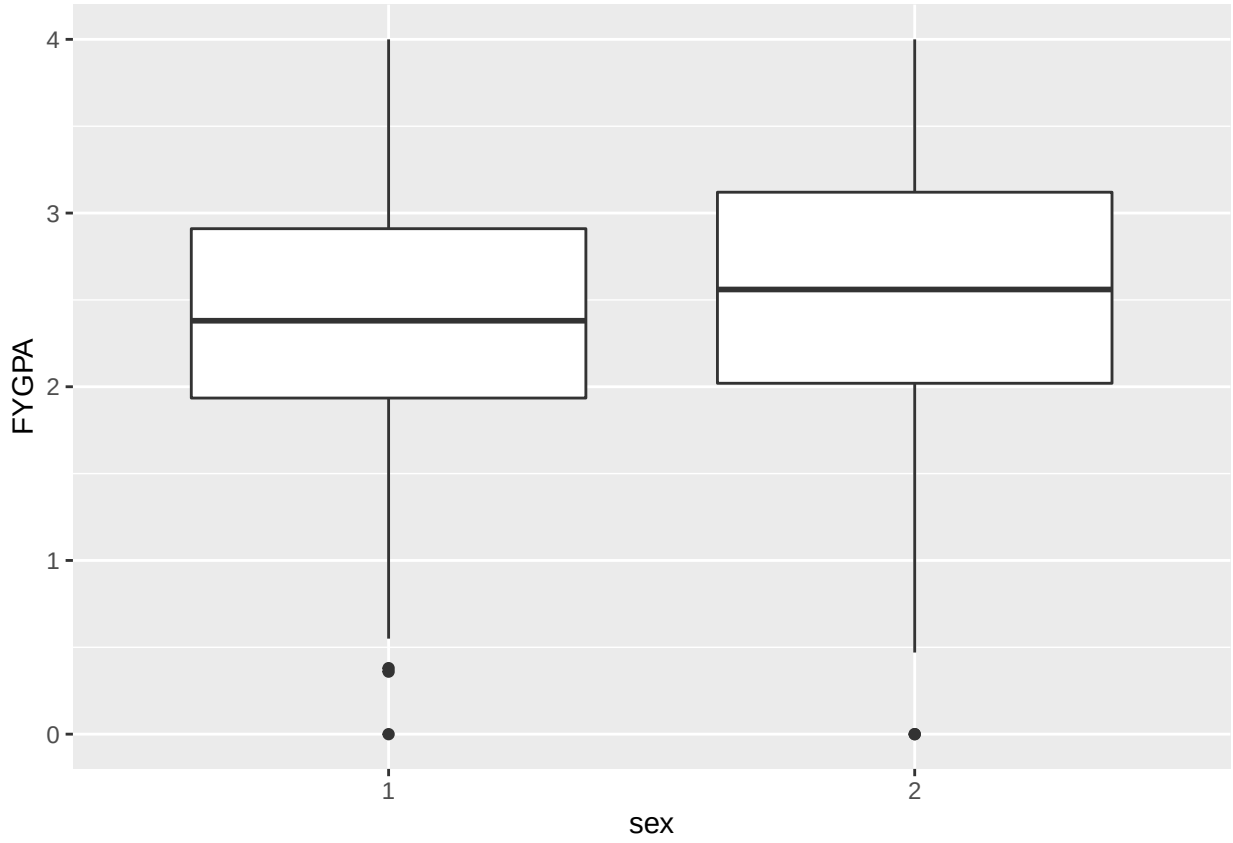
Model the relationship between first year college GPA and sex.

Repeat the steps you took in the previous section, this time using sex as the predictor. You will need to do one additional step before you start:

```
satGPA$sex <- as.factor(satGPA$sex)
```

Make a side-by-side boxplot to show the relationship between FYGPA and sex.

```
## boxplot
ggplot(data=satGPA, aes(x=sex, y=FYGPA)) + geom_boxplot()
```



Write doen the (generic) equation for the linear model. Define what x and y are in the equation.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

, where y=FYGPA and x=indicator variable for x (meaning it will be 0 if sex=1 (male) and 1 if sex=2 (female)). You wouldn't necessarily know this last part before you run the model.

What values can x take on here?

In a model with a categorical predictor with two levels, x will take on 0 and 1, only.

Fit a linear model that uses sex to predict FYGPA; assign this linear model to `lm_FYGPA_sex`.

```
lm_FYGPA_sex <- lm(data=satGPA, FYGPA~sex)
```

Use the summary function on your linear model to print out the model summary details.

```
summary(lm_FYGPA_sex)
```

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

```
## lm(formula = FYGPA ~ sex, data = satGPA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.54459 -0.50459 -0.00459  0.55541  1.60393
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.39607    0.03246  73.807   <2e-16 ***
## sex2         0.14852    0.04666   3.183   0.0015 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7374 on 998 degrees of freedom
## Multiple R-squared:  0.01005,    Adjusted R-squared:  0.009056
## F-statistic: 10.13 on 1 and 998 DF,  p-value: 0.001504
```

What is the estimate $\hat{\beta}_0$? What does this mean?

$\hat{\beta}_0 = 2.396$ FY GPA points. On average, we expect a FYGPA of 2.396 for males.

What is the estimate for $\hat{\beta}_1$? What does this mean?

$\hat{\beta}_1 = 0.149$. This is the expected change in GPA as we move from males to females. In other words, on average, we expect FYGPA for females to be 0.149 points higher for females than for males, based on this model.

What is the R^2? Interpret this value in the context of the problem.

Looking at the summary function output, $R^2 = 0.01$. This means that about 1 percent of the variability in FYGPA can be explained by sex.

Here we have a categorical predictor with two levels (male or female). Is the linearity assumption satisfied? Why or why not?

Linearity is always satisfied for a linear model with a categorical predictor with two levels.