ANAGHA UPPAL - MW Section - Homework 4

Simple Linear Regression 1

Due Sept 22 by 330pm

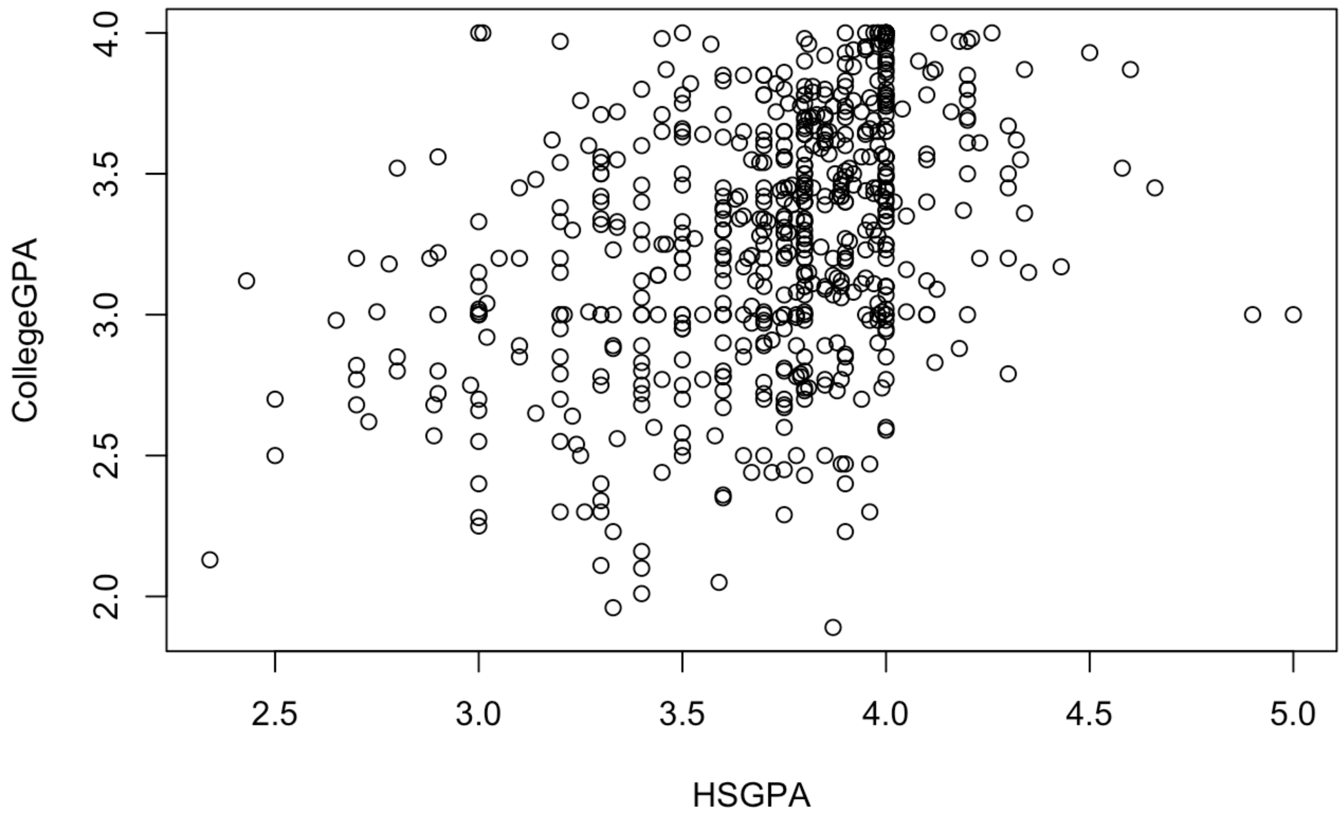
Run `library(regclass)` to load up the data and functions we'll use, then use the `data` function to load in the `EDUCATION` dataset. Run `?EDUCATION` to read a little about this dataset (comment out this line when the document is knitted).

```
#Code for loading up `EDUCATION`
data("EDUCATION")
#?EDUCATION
```

Education analytics is concerned with figuring out what factors are associated with "college success" (as gauged by GPA), and more importantly *how* the factors are associated. For this homework you will be fitting simple linear regressions predicting `CollegeGPA` (students' eventual college GPAs) from information known about them when they applied to UT.

1. Make a scatterplot to examine the relationship between `CollegeGPA` and `HSGPA` . Be sure to assign the roles of y and x correctly. A linear regression model is only appropriate when the overall trajectory of the stream of points is well-described by a straight line. Does it appear "linear enough" or is there "obvious curvature"?

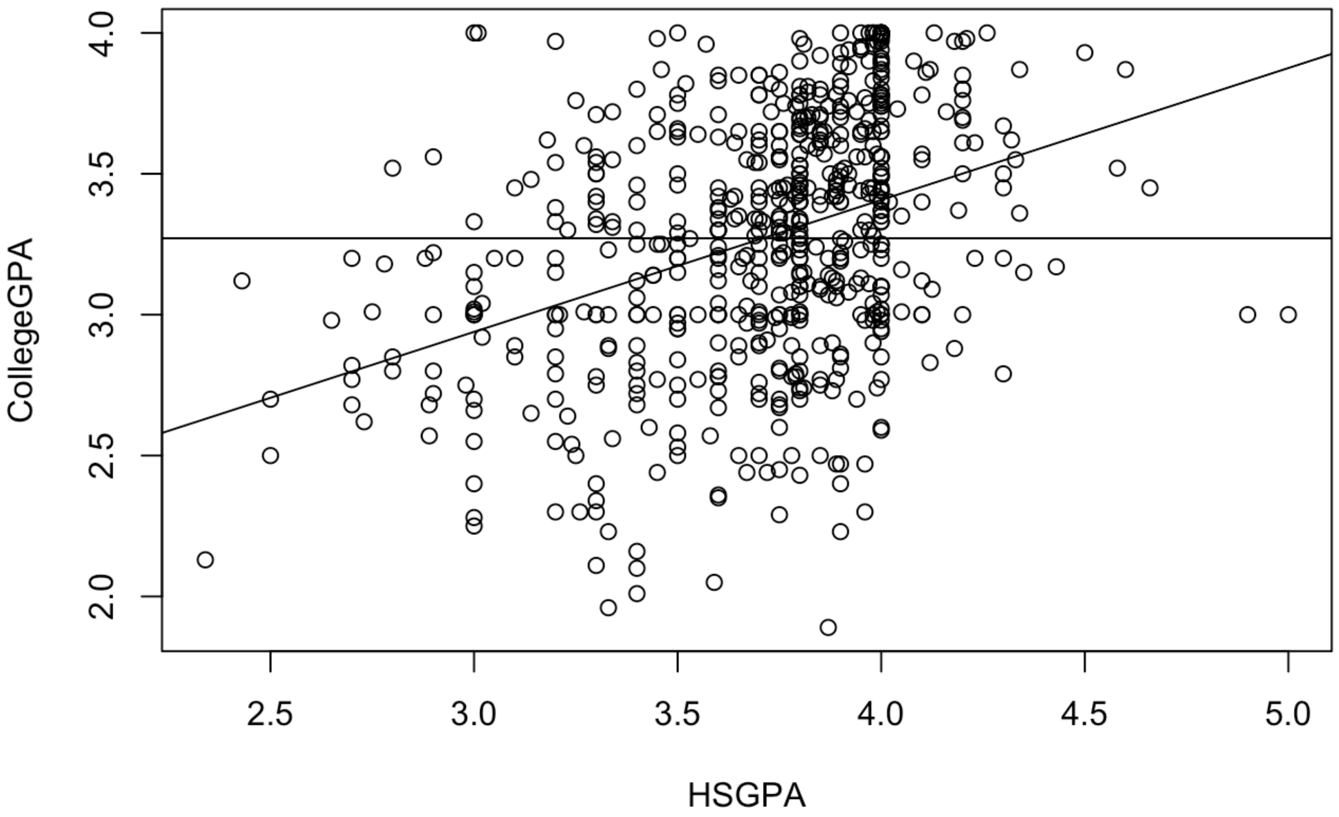
```
#R code for making the plot
plot(CollegeGPA~HSGPA,data=EDUCATION)
```



Response: I think that the graph looks fairly liner with obvious outliers, but can be described by a straight line.

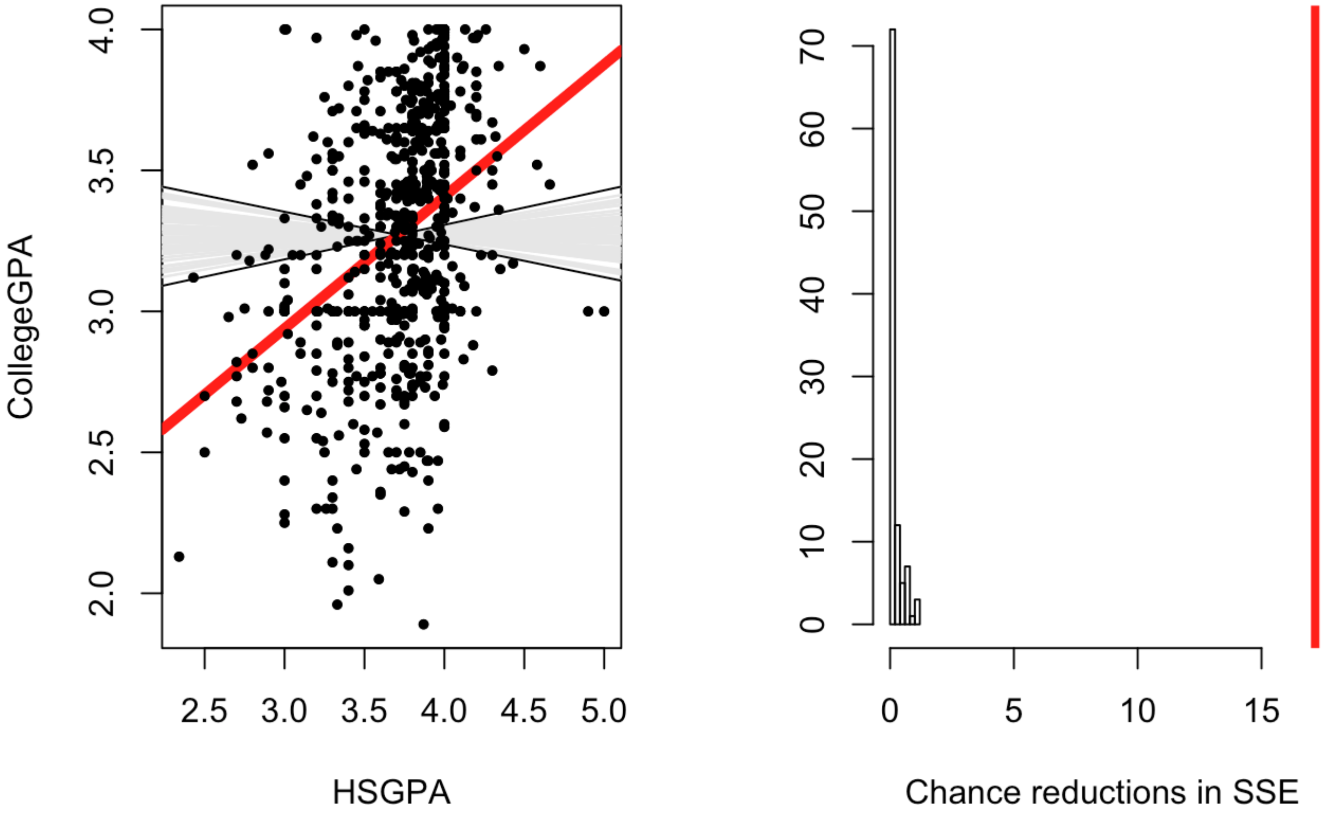
2. Fit the simple linear regression model, naming it `M`. Add the regression line to the plot by running `abline(M)`. Also add a horizontal line at the average value of `CollegeGPA` to represent the "naive model" (code provided, just uncomment it).

```
#R code fitting the model predicting price from bids
plot(CollegeGPA~HSGPA,data=EDUCATION)
M <- lm(CollegeGPA~HSGPA,data=EDUCATION)
#R code adding the line
abline(M)
abline(h=mean(EDUCATION$CollegeGPA))
```



3. We want to determine if the regression model predicting `CollegeGPA` from `HSGPA` fits the data better than a regression model using a variable that contains no inherent information about `CollegeGPA`. We do this by comparing the reduction in the sum of squared errors (compared to the naive model) achieved by our regression with the possible reductions that can occur "by chance", i.e., achieved by regressions using predictors unrelated to `CollegeGPA` via the permutation procedure. Run `possible.regressions(M)`. Comment on why the output indeed suggests that the regression is statistically significant.

```
#R code for possible.regressions
possible.regressions(M)
```



Response: On the left, the regression line looks vastly different from a possible model independent from HSGPA. On the right, the reduction in SSE via the regression is clearly very much higher than a reduction through the possible permutation lines.

---

4. Run `anova` on your model and report:

a. **SSE of the regression model:** 104.243

b. **Reduction in SSE of regression model:** 17.168

c. **p-value of the reduction:** 2.2 e^(-16)

```
#R code running the ANOVA on the model
anova(M)
```

```
## Analysis of Variance Table
##
## Response: CollegeGPA
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## HSGPA       1  17.168 17.1683  99.641 < 2.2e-16 ***
## Residuals 605 104.243  0.1723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Based on the analysis, is `HSGPA` a better predictor of `CollegeGPA` than a random, unrelated quantity? Why or why not?

Response: Absolutely. The p-value is so small that this association could not have occurred by random chance.

6. Run `summary` on the model and report:

a. **The intercept:** 1.534
b. **The slope:** 0.4684
c. **The standard error of the slope:** 0.0469
d. **The RMSE:** 0.4151
e. R^2**:** 0.1414
f. p**-value of the regression** 2e-16

```
#R code getting a summary of the model
summary(M)
```

```
##
## Call:
## lm(formula = CollegeGPA ~ HSGPA, data = EDUCATION)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45672 -0.28533  0.03922  0.32606  1.06079
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.53399    0.17485   8.773   <2e-16 ***
## HSGPA        0.46841    0.04693   9.982   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4151 on 605 degrees of freedom
## Multiple R-squared:  0.1414, Adjusted R-squared:   0.14
## F-statistic: 99.64 on 1 and 605 DF,  p-value: < 2.2e-16
```

7. Briefly explain the difference between the standard error of the slope and the root mean squared error (RMSE) of the regression

Response: RMSE measures scatter in the points. RMSE is the typical size of the residual and will tell you how far most normally are from the line. The standard error tells you for each individual value, how far off the predicted value is from the truth.

8. Give a precise interpretation of the intercept and the slope of the regression model predicting `CollegeGPA` from `HSGPA`.

Intercept: The model predicts that if the HS GPA were at a 0.0, the college GPA of the student would be around 1.534.

Slope: With each per digit increase in HS GPA, the predicted college GPA increases by 0.4686.

9. The analysis has intrigued parents of high schoolers since students that tend to do well in high school tend to do well in college. One parent asks you: “Currently my kid has a HS GPA of 3.2, so his college GPA is predicted to be 3.03. If I can get my kid to study harder and raise his GPA to 3.5 by the time he applies, what does the model about how much his college GPA will increase”?

Comment: The model predicts a student with a HS GPA of 3.5 will have a college GPA of 3.1741. However, I would warn the parent that the model predicts larger trends in GPA association, and these models do not necessarily apply to individual cases, especially on such small scales as GPA increases of 0.3.

10. Provide a 99% confidence interval for the slope to capture the amount of uncertainty we have regarding our estimate.

```
#R code giving a 99% confidence interval for the slope
confint(M, level = 0.99)
```

```
##                 0.5 %    99.5 %
## (Intercept) 1.0821778 1.9857943
## HSGPA       0.3471542 0.5896614
```

11. Fit five more simple linear regressions predicting `CollegeGPA` using `ACT`, `APHours`, `JobHours`, `Height`, and `Weight` as predictors. Which regressions are statistically significant? Which of the variables (including `HSGPA`) is the “best” predictor of `CollegeGPA`. Explain.

Comment regarding which regressions are significant:

Comment regarding which predictor is best:

```
#R code fitting the simple linear regressions.
act <- lm(CollegeGPA~ACT,data=EDUCATION)
summary(act) # = 3.96e-10; statistically significant
ap <- lm(CollegeGPA~APHours,data=EDUCATION)
summary(ap) # = 0.00127; statistically significant
job <- lm(CollegeGPA~JobHours,data=EDUCATION)
summary(job) # = 0.0744; NOT statistically significant

height <- lm(CollegeGPA~Height,data=EDUCATION)
summary(height) # = 5.29e-05; statistically significant??

weight <- lm(CollegeGPA~Weight,data=EDUCATION)
summary(weight) # = 1.69e-0.05; statistically significant

#Note: the eval=FALSE option is active in this chunk, so while the R code is included in the writeup, the results will not (the summary output of the models is not necessary, just the comments on what regressions are significant and what predictor is best).

#Student's ACT score, predictably, seems to be the best predictor of College GPA
```