THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple regression 1

Lecture 7

STA 371G

Why do some colleges have higher graduation rates than others?

- What factors do you think impact the graduation rate of a college?

Why do some colleges have higher graduation rates than others?

- What factors do you think impact the graduation rate of a college?
- It seems like there is no *one* factor that dominates—it is probably true that to make a good prediction we need to put a lot of variables together, so simple regression is likely not sufficient.

Why do some colleges have higher graduation rates than others?

- What factors do you think impact the graduation rate of a college?
- It seems like there is no *one* factor that dominates—it is probably true that to make a good prediction we need to put a lot of variables together, so simple regression is likely not sufficient.
- Multiple regression allows us to build on simple regression by predicting one Y variable using multiple X variables.

The colleges data set

Today's data set is a sample of 1302 colleges with various factors about the colleges, including SAT scores, student/faculty ratios, tuition rates, acceptance rates, etc.

A quick data clean

Many colleges have no SAT scores reported, so let's ignore those colleges (to enable a fair comparison) and also remove colleges with an obviously incorrect graduation rate of > 100%:

```
> my.sample <- subset(colleges,
+   !is.na(Average.combined.SAT) & Graduation.rate <= 100)
```

SAT scores and (in-state) tuition were the two best single predictors, with R^2 values of 0.353 and 0.325, respectively. Can we combine these together and get an R^2 that is better than either predictor would produce on its own?

Using multiple predictors to predict graduation rate

The simple regression models were:

$$Y_i = \beta_0 + \beta_1(\text{SAT}) + \epsilon_i$$

and

$$Y_i = \beta_0 + \beta_1(\text{tuition}) + \epsilon_i.$$

The multiple regression model is

$$Y_i = \beta_0 + \beta_1(\text{tuition}) + \beta_2(\text{SAT}) + \epsilon_i.$$

```
> model <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition, data=my.
> summary(model)


Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition,
    data = my.sample)

Residuals:
   Min     1Q Median     3Q    Max
-45.53  -9.18   0.05   8.70  43.66

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -8.324646   4.370828    -1.9    0.057 .
Average.combined.SAT  0.061122   0.004888    12.5   <2e-16 ***
In.state.tuition      0.001249   0.000111    11.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7 on 709 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.447,Adjusted R-squared:  0.445
F-statistic:  286 on 2 and 709 DF,  p-value: <2e-16
```

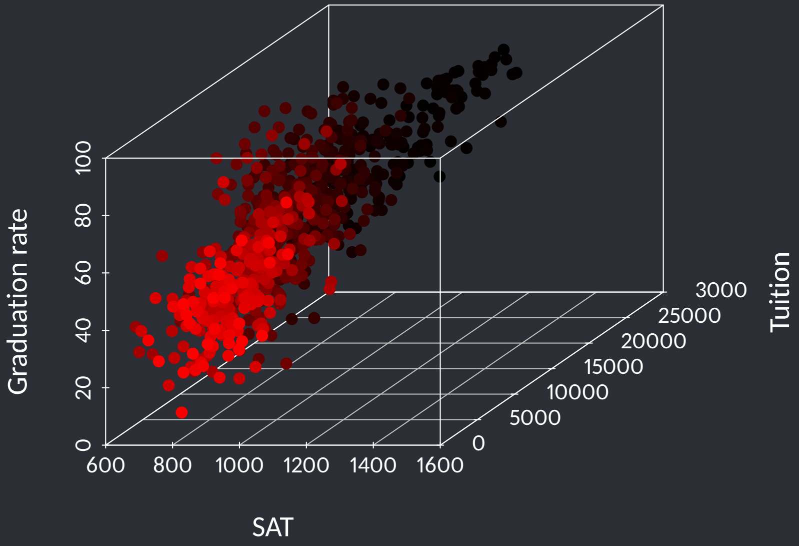
The multiple regression prediction equation is:

$$\widehat{\text{Graduation}}\text{ rate} = -8.3246 + 0.0611(\text{SAT}) + 0.0012(\text{tuition})$$

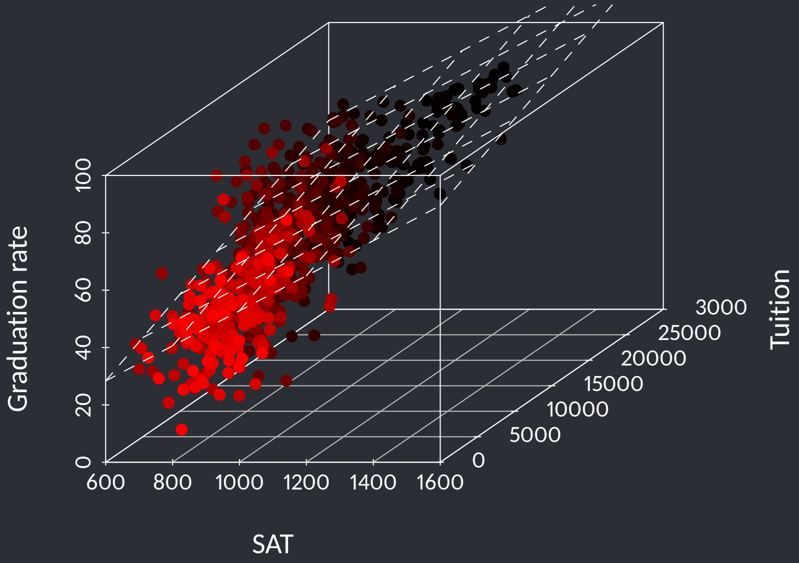
The multiple regression prediction equation is:

$$\widehat{\text{Graduation rate}} = -8.3246 + 0.0611(\text{SAT}) + 0.0012(\text{tuition})$$

We can use this to make predictions like we would for a simple regression!



Graduation rate
0
20
40
60
80
100
600
800
1000
1200
1400
1600
SAT
0
5000
10000
15000
20000
25000
3000
Tuition


Graduation rate
0
20
40
60
80
100
SAT
600
800
1000
1200
1400
1600
Tuition
0
5000
10000
15000
20000
25000
3000

Interpreting the coefficients: intercept

Let's interpret the intercept coefficient of -8.3246:

- The predicted graduation rate when the average SAT score is 0 and the in-state tuition is \$0 is -8.3246.

Interpreting the coefficients: intercept

Let’s interpret the intercept coefficient of -8.3246:

- The predicted graduation rate when the average SAT score is 0 and the in-state tuition is \$0 is -8.3246.
- This is not a meaningful number on its own in this case, since there will never be a school with those particular predictor values! (The intercept might be interpretable for other models.)

Interpreting the coefficients: SAT

Let’s interpret the SAT coefficient of 0.0611:

- Holding tuition constant, each additional SAT score point increases our predicted graduation rate by 0.0611 percentage points.

Interpreting the coefficients: SAT

Let's interpret the SAT coefficient of 0.0611:

- Holding tuition constant, each additional SAT score point increases our predicted graduation rate by 0.0611 percentage points.
- Among colleges that have the same tuition, an increase in SAT of 1 point would result in a predicted graduation rate that is 0.0611 percentage points higher.

Interpreting the coefficients: SAT

Let's interpret the SAT coefficient of 0.0611:

- Holding tuition constant, each additional SAT score point increases our predicted graduation rate by 0.0611 percentage points.
- Among colleges that have the same tuition, an increase in SAT of 1 point would result in a predicted graduation rate that is 0.0611 percentage points higher.
- If we compared two colleges that have the same tuition but differ in average SAT scores by 1 point, the college with the higher SAT score would be predicted to have a graduation rate that is 0.0611 percentage points higher.

Interpreting the coefficients: tuition

Let's interpret the tuition coefficient of 0.0012:

- Holding SAT constant, each additional dollar of in-state tuition increases our predicted graduation rate by 0.0012 percentage points.

Interpreting the coefficients: tuition

Let’s interpret the tuition coefficient of 0.0012:

- Holding SAT constant, each additional dollar of in-state tuition increases our predicted graduation rate by 0.0012 percentage points.
- Among colleges that have the same average SAT scores, an increase in tuition of $1 would result in a predicted graduation rate that is 0.0012 percentage points higher.

Interpreting the coefficients: tuition

Let's interpret the tuition coefficient of 0.0012:

- Holding SAT constant, each additional dollar of in-state tuition increases our predicted graduation rate by 0.0012 percentage points.
- Among colleges that have the same average SAT scores, an increase in tuition of $1 would result in a predicted graduation rate that is 0.0012 percentage points higher.
- If we compared two colleges that have the same average SAT scores but differ in their tuition by $1, the college with the higher tuition would be predicted to have a graduation rate that is 0.0012 percentage points higher.

What’s the difference?!

- The difference between “the predicted effect of a 1-point increase in SAT score” and “the predicted effect of a 1-point increase in SAT score, holding tuition constant” really are two different things.

What's the difference?!

- The difference between "the predicted effect of a 1-point increase in SAT score" and "the predicted effect of a 1-point increase in SAT score, holding tuition constant" really are two different things.
- The relationship between X_1 and Y may change when we control for (i.e., add to the model) another predictor X_2.

R^2

- R^2 has a similar meaning as in simple regression: how much of the variation in the response variable (Y) are explained by the predictor variables (X's) together?

R^2

- R^2 has a similar meaning as in simple regression: how much of the variation in the response variable (Y) are explained by the predictor variables (X's) together?
- Another way to think about R^2 is that

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)},$$

i.e., it represents how much variance in Y the model predicts.

R^2

- R^2 has a similar meaning as in simple regression: how much of the variation in the response variable (Y) are explained by the predictor variables (X's) together?
- Another way to think about R^2 is that

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)},$$

  i.e., it represents how much variance in Y the model predicts.
- R^2 always increases when you add more variables, even if you add variables that have no real relationship with Y.

```
> model1 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition,
+              data=my.sample)
> summary(model1)


Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition,
    data = my.sample)

Residuals:
      Min         1Q    Median         3Q        Max
-45.52572  -9.18156   0.05085   8.70420  43.66097

Coefficients:
                         Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.324645625  4.370827909 -1.90459 0.057238 .
Average.combined.SAT  0.061122082  0.004887825 12.50496  < 2e-16 ***
In.state.tuition      0.001248638  0.000111119 11.23692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7466 on 709 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.44686,Adjusted R-squared:   0.4453
F-statistic: 286.387 on 2 and 709 DF,  p-value: < 2.22e-16
```

```
> Random.numbers <- rnorm(nrow(my.sample))
> model2 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition
+                 + Random.numbers, data=my.sample)
> summary(model2)


Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition +
    Random.numbers, data = my.sample)

Residuals:
      Min         1Q    Median         3Q        Max
-45.59477  -9.13473   0.06836   8.75583  43.74968

Coefficients:
                        Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.433559857  4.378188630 -1.92627 0.054471 .
Average.combined.SAT  0.061244215  0.004896088 12.50881  < 2e-16 ***
In.state.tuition      0.001248531  0.000111177 11.23012  < 2e-16 ***
Random.numbers        0.277098090  0.537299499  0.51572 0.606208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7537 on 708 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.447068,Adjusted R-squared:  0.444725
F-statistic: 190.816 on 3 and 708 DF,  p-value: < 2.22e-16
```

```
> Random.numbers <- rnorm(nrow(my.sample))
> model2 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition
+                + Average.math.SAT, data=my.sample)
> summary(model2)



Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition +
    Average.math.SAT, data = my.sample)

Residuals:
      Min         1Q     Median         3Q        Max
-45.27189  -9.06503   0.03009   8.64981  43.89591

Coefficients:
                         Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.144252350  4.434188943 -1.83669 0.066675 .
Average.combined.SAT  0.054229967  0.023519872  2.30571 0.021416 *
In.state.tuition      0.001256312  0.000115918 10.83790  < 2e-16 ***
Average.math.SAT      0.012667133  0.041953872  0.30193 0.762794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7492 on 706 degrees of freedom
  (21 observations deleted due to missingness)
Multiple R-squared:  0.447693,Adjusted R-squared:  0.445346
F-statistic: 190.758 on 3 and 706 DF,  p-value: < 2.22e-16
```

Adjusted R^2

- There are many, many possible models (think of how many combinations of predictors there are!) so we need some criterion to determine which model is best.

Adjusted R^2

- There are many, many possible models (think of how many combinations of predictors there are!) so we need some criterion to determine which model is best.
- R^2 is not good because adding even a variable of random numbers increases R^2.

Adjusted R^2

- There are many, many possible models (think of how many combinations of predictors there are!) so we need some criterion to determine which model is best.
- R^2 is not good because adding even a variable of random numbers increases R^2.
- Adjusted R^2 makes an adjustment to R^2 by adding a penalty for each variable added (in this example, adjusted R^2 went down even though R^2 increased).

```
> model1 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition,
+              data=my.sample)
> summary(model1)


Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition,
    data = my.sample)

Residuals:
      Min         1Q    Median         3Q        Max
-45.52572  -9.18156   0.05085   8.70420  43.66097

Coefficients:
                         Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.324645625  4.370827909 -1.90459 0.057238 .
Average.combined.SAT  0.061122082  0.004887825 12.50496  < 2e-16 ***
In.state.tuition      0.001248638  0.000111119 11.23692  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7466 on 709 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.44686,Adjusted R-squared:   0.4453
F-statistic: 286.387 on 2 and 709 DF,  p-value: < 2.22e-16
```

```
> Random.numbers <- rnorm(nrow(my.sample))
> model2 <- lm(Graduation.rate ~ Average.combined.SAT + In.state.tuition
+                + Random.numbers, data=my.sample)
> summary(model2)



Call:
lm(formula = Graduation.rate ~ Average.combined.SAT + In.state.tuition +
    Random.numbers, data = my.sample)

Residuals:
      Min         1Q    Median        3Q       Max
-45.59477  -9.13473   0.06836   8.75583  43.74968

Coefficients:
                         Estimate   Std. Error  t value Pr(>|t|)
(Intercept)          -8.433559857  4.378188630 -1.92627 0.054471 .
Average.combined.SAT  0.061244215  0.004896088 12.50881  < 2e-16 ***
In.state.tuition      0.001248531  0.000111177 11.23012  < 2e-16 ***
Random.numbers        0.277098090  0.537299499  0.51572 0.606208
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.7537 on 708 degrees of freedom
  (19 observations deleted due to missingness)
Multiple R-squared:  0.447068,Adjusted R-squared:  0.444725
F-statistic: 190.816 on 3 and 708 DF,  p-value: < 2.22e-16
```

Which model is the best?

- In general, we want to select the model that is the most parsimonious, that is, the model that has the best combination of being simple with a high R^2.

Which model is the best?

- In general, we want to select the model that is the most parsimonious, that is, the model that has the best combination of being simple with a high R^2.
- This is easier said than done—using Adjusted R^2 is not enough. We'll come back to this next week!