THE UNIVERSITY OF TEXAS AT AUSTIN
McCOMBS SCHOOL OF BUSINESS

Multiple regression

Lecture 15

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

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

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

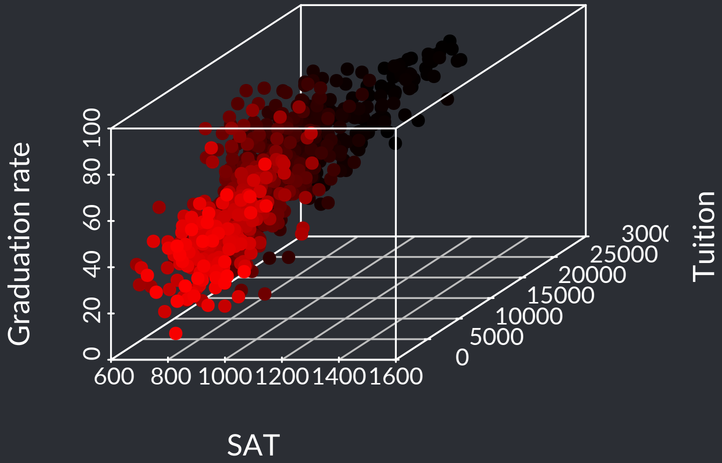
The multiple regression prediction equation is:

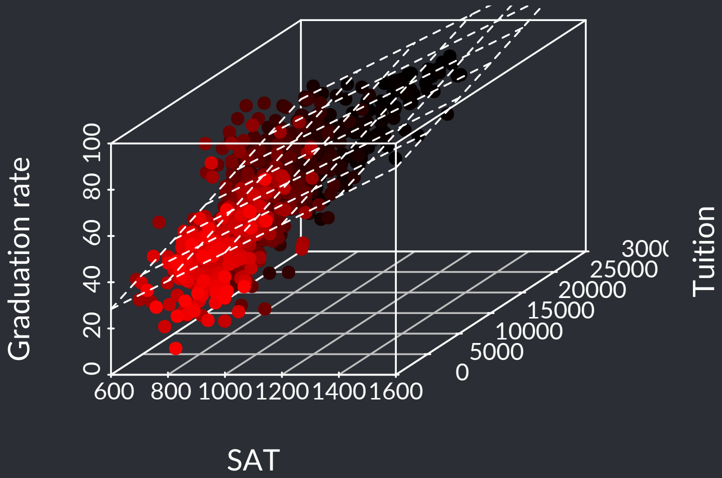
$$\widehat{\text{Graduation rate}} = -8.3246 + 0.0611(\text{SAT}) + 0.0012(\text{tuition})$$

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
Tuition


Graduation rate
0
20
40
60
80
100
600
800
1000
1200
1400
1600
SAT
0
5000
10000
15000
20000
25000
300(
Tuition

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 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 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.

Multiple regression assumptions

We need (the same!) four things to be true for statistical inference (i.e., hypothesis tests, *p*-values, confidence intervals) to work for multiple regression:

Multiple regression assumptions

We need (the same!) four things to be true for statistical inference (i.e., hypothesis tests, p-values, confidence intervals) to work for multiple regression:

1. The errors are independent.
2. Y is a linear function of the X's (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 1: Independence of errors

Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case.

Assumption 1: Independence of errors

Independence means that knowing the error (over-/under-prediction by the regression line) for one case doesn't tell you anything about the error for another case.

Since each college is completely separate, there is no reason to think the errors are not independent.

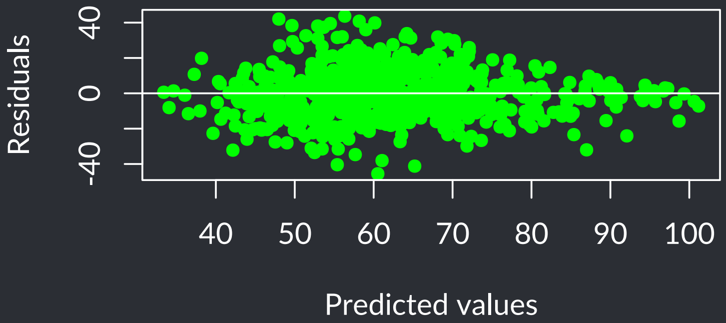
Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors).
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 2: Linearity

Look at the residual plot:

```r
plot(predict(model), residuals(model), col="green",
  xlab="Predicted values", ylab="Residuals", pch=16)
abline(h=0)
```

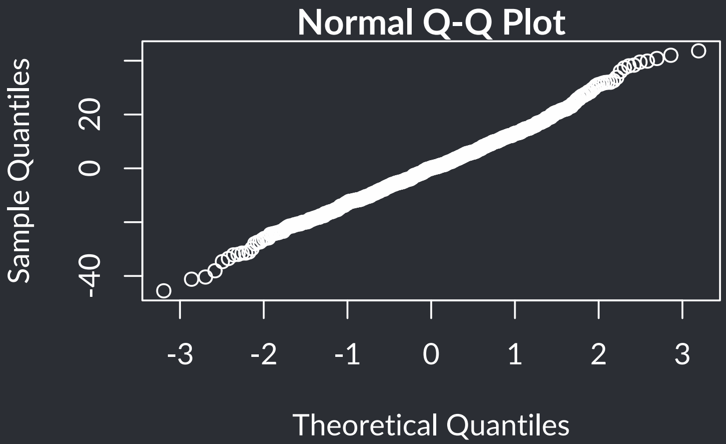


Multiple regression assumptions

1. The errors are independent. $\checkmark$
2. Y is a linear function of the X's (except for the errors). $\checkmark$
3. The errors are normally distributed.
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 3: Normality of residuals

```
qqnorm(residuals(model))
```



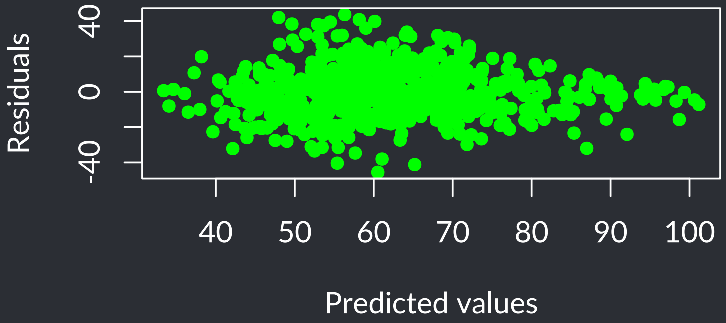
Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity").

Assumption 4: Homoscedasticity

Look at the residual plot:

```r
plot(predict(model), residuals(model), col="green",
  xlab="Predicted values", ylab="Residuals", pch=16)
```



Multiple regression assumptions

1. The errors are independent. ✓
2. Y is a linear function of the X's (except for the errors). ✓
3. The errors are normally distributed. ✓
4. The variance of Y is the same for any value of X ("homoscedasticity"). ¯_(ツ)_/¯

Since one of the assumptions is not completely satisfied, we'll proceed with caution—i.e., take the *p*-values and confidence intervals with a grain of salt. (We could try and fix the problem with a transformation, or by building different models for different subsets of the data, but let's just live with it for now.)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with *k* predictor variables:

- $R^2 = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)
- $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all coefficients are 0 except the intercept)

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\text{cor}(\hat{Y}, Y) = 0$ (in the population)
- $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all coefficients are 0 except the intercept)
- The model has no predictive power

The overall null hypothesis for a regression model

The following are equivalent ways to express the overall null hypothesis with k predictor variables:

- $R^2 = 0$ (in the population)
- $\mathrm{cor}(\hat{Y}, Y) = 0$ (in the population)
- $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ (i.e., all coefficients are 0 except the intercept)
- The model has no predictive power
- Predictions from this model are no better than predicting $\overline{Y}$ for every case

We should always test the overall null hypothesis for a model first. If we can't reject the overall null hypothesis, there's no reason to interpret the model further.

We should always test the overall null hypothesis for a model first. If we can't reject the overall null hypothesis, there's no reason to interpret the model further.

In this model, the overall p-value is very small, so we reject the overall null hypothesis and conclude that yes, we have statistical significance and that this model does have some predictive power.

Statistical vs practical significance

- As in simple regression, once we determine that there is statistical significance, we want to then assess whether there is also practical significance.
- For the test of the overall null hypothesis, we look to the value of R^2 in the sample to assess practical significance.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i, for each $i = 1, 2, \ldots, k$.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i, for each $i = 1, 2, \ldots, k$.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i, for each $i = 1, 2, \ldots, k$.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.
- The test statistic for testing the null hypothesis $\beta_i = S$ follows a t-distribution with $n - k - 1$ degrees of freedom:

$$t = \frac{\beta_i - S}{\mathrm{SE}(\beta_i)}$$

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i, for each $i = 1, 2, \dots, k$.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.
- The test statistic for testing the null hypothesis $\beta_i = S$ follows a t-distribution with $n - k - 1$ degrees of freedom:

$$t = \frac{\beta_i - S}{\mathrm{SE}(\beta_i)}$$

- The regression output calculates the p-value for us for testing the null hypotheses $\beta_i = 0$.

Testing individual coefficients

- Next, we want to test $H_0 : \beta_i = 0$ for each of the predictors X_i, for each $i = 1, 2, \ldots, k$.
- This is equivalent to the null hypothesis that X_i has no correlation with Y once the other predictors are held constant.
- The test statistic for testing the null hypothesis $\beta_i = S$ follows a t-distribution with $n - k - 1$ degrees of freedom:

$$t = \frac{\beta_i - S}{\mathrm{SE}(\beta_i)}$$

- The regression output calculates the p-value for us for testing the null hypotheses $\beta_i = 0$.
- If we reject this null hypothesis for a coefficient, we say that X_i is a (statistically) significant predictor of Y in the model.

Testing individual coefficients

If a predictor is not statistically significant, we should:

1. Interpret it as if it were zero.

Testing individual coefficients

If a predictor is not statistically significant, we should:

1. Interpret it as if it were zero.
2. Remove it from the model (unless there are other reasons to keep it), as it does not contribute to predicting Y above and beyond the other predictors.

Residual standard error

- Like with simple regression, the residual standard error s_e is approximately equal to the standard deviation of the residuals.

Residual standard error

- Like with simple regression, the residual standard error s_e is approximately equal to the standard deviation of the residuals.
- Since one of the assumptions of regression is that the residuals are approximately normal, we can conclude that approximately 95% of the residuals will be less than $\pm 2s_e$.

Confidence intervals for coefficents

Confidence intervals for the individual coefficients are found the same way as in simple regression, and interpreted the same way:

```
confint(model)

                          2.5 %   97.5 %
(Intercept)         -16.90596 0.256669
Average.combined.SAT  0.05153 0.070718
In.state.tuition      0.00103 0.001467
```

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

A 95% CI for the graduation rate at the University of California, Merced, which is not in the data set and has an average SAT score of 1100 and in-state tuition of $11,502:

```
predict(model, list(Average.combined.SAT=1100,
                    In.state.tuition=11502),
               interval="prediction")
```

```
    fit   lwr   upr
1 73.27 46.24 100.3
```

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

A 95% CI for the graduation rate at the University of California, Merced, which is not in the data set and has an average SAT score of 1100 and in-state tuition of $11,502:

```
predict(model, list(Average.combined.SAT=1100,
                    In.state.tuition=11502),
               interval="prediction")

    fit   lwr   upr
1 73.27 46.24 100.3
```

Our best guess for UC Merced is 73.27%, with a 95% CI of (46.24%, 100.3%).

Confidence intervals for predictions

We can also put confidence intervals on our predictions for Y.

A 95% CI for the graduation rate at the University of California, Merced, which is not in the data set and has an average SAT score of 1100 and in-state tuition of $11,502:

```
predict(model, list(Average.combined.SAT=1100,
                    In.state.tuition=11502),
               interval="prediction")

    fit   lwr   upr
1 73.27 46.24 100.3
```

Our best guess for UC Merced is 73.27%, with a 95% CI of (46.24%, 100.3%). (It turns out that the actual graduation rate at UC Merced is 64%.)

Confidence intervals for predictions

A 95% CI for average graduation rate among all colleges with an average SAT score of 1100 and in-state tuition of $11,502:

```
predict(model, list(Average.combined.SAT=1100,
                    In.state.tuition=11502),
               interval="confidence")
```

```
      fit      lwr      upr
1 73.2715 71.8091 74.7338
```

Confidence intervals for predictions

A 95% CI for average graduation rate among all colleges with an average SAT score of 1100 and in-state tuition of $11,502:

```
predict(model, list(Average.combined.SAT=1100,
                    In.state.tuition=11502),
              interval="confidence")

      fit     lwr     upr
1 73.2715 71.8091 74.7338
```

As with simple regression, our point estimate is the same, but the confidence interval is much narrower, because it’s easier to estimate a mean than a prediction for a single new case.