Modeling Longitudinal Data

Dr. Maria Tackett

12.04.19

STA 210

Click for PDF of slides

Announcements

- Project write up and presentation **due Dec 10 at 11:59p**
- Project presentations on Dec 11
  - Lab 01L: 9a - 10:30a
  - Lab 02L: 10:30a - 12p
- Exam 2 extra credit:
  - 90% response rate on course eval: +1 pt on Exam 02 grades
- Thursday's lab: Project office hours
- Office hours next week: Monday, 12/9 at 1:30p - 3p in Old Chem 118B

US college graduation rates

What factors most effect graduation rates at US colleges?

Response variable:

- **`rate`**: graduation rate, i.e. number of degrees awarded per 100 students enrolled

Predictor variables:

- **`year02`**: number of years since 2002
- **`faculty`**: mean number of full-time faculty in 2002 - 2009
- **`tuition`**: mean yearly tuition between 2002 and 2009

```r
college <- read_csv("data/colleges.csv") %>%
  filter(rate < 100) %>%
  mutate(year02 = year - 2002)
```

college data

```
## # A tibble: 11 x 5
##    instname                                    year02 faculty tuition  r
##    <chr>                                        <dbl>   <dbl>   <dbl> <d
##  1 University of North Carolina at Chapel Hill      3    6.55    10.3  2
##  2 University of North Carolina at Chapel Hill      4    6.55    10.3  2
##  3 University of North Carolina at Chapel Hill      5    6.55    10.3  2
##  4 University of North Carolina at Chapel Hill      6    6.55    10.3  2
##  5 University of North Carolina at Chapel Hill      7    6.55    10.3  2
##  6 Duke University                                  2    5.17    25.3  3
##  7 Duke University                                  3    5.17    25.3  2
##  8 Duke University                                  4    5.17    25.3  2
##  9 Duke University                                  5    5.17    25.3  2
## 10 Duke University                                  6    5.17    25.3  2
## 11 Duke University                                  7    5.17    25.3  2
```

What makes this model different?

- Goals:
    - Understand how the number of faculty members and tuition affects a college's graduation rate
    - How the graduation rate has changed over time
- There are multiple observations for each college (so multiple regression not appropriate)
- The are only a few time points and there's data on multiple colleges (so time series model not appropriate)
- We will use a multilevel model to model the relationship between `faculty`, `tuition` and `rate`.

Multilevel Model

We will fit a two-level model that includes the following model components:

- **Level One**: include time and any other predictors that change within a college over the time period in the data (`year02`)
  - The effects in this component are .vocab[random effects]
  - Typically not interested in drawing inferences about specific levels
- **Level Two**: includes predictors that differ between colleges but that remain the same within a college over the time period in the data (`faculty` and `tuition`)
  - The effects in this component are .vocab[fixed effects]]
  - Typicall the effects we wish to drawing inferences about

Modeling Approach

Approach: Start with simple, preliminary models to establish a baseline that can be used to evaluate more complex models. Work toward the final model by adding predictors and checking model assumptions at each step. We can take the following steps:

1. Exploratory data analysis

2. Fit unconditional means model - model with no predictors

3. Fit unconditional growth model - add time

4. Fit "final" model with time and predictors

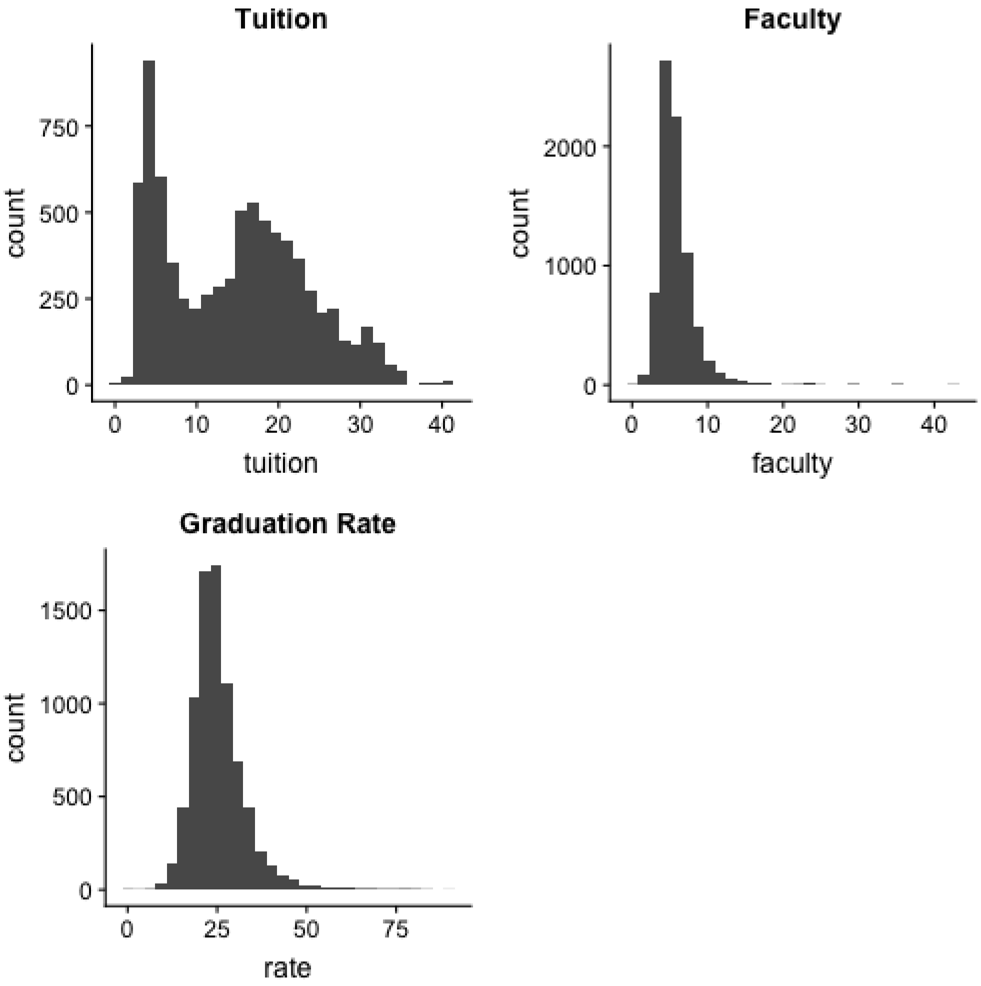
1. Exploratory Data Analysis

- Given the longitudinal structure of the data, we have observations at different time points for each college in the data set.

- When we do EDA, in addition to an univariate analysis of each variable, we want to look at the following:

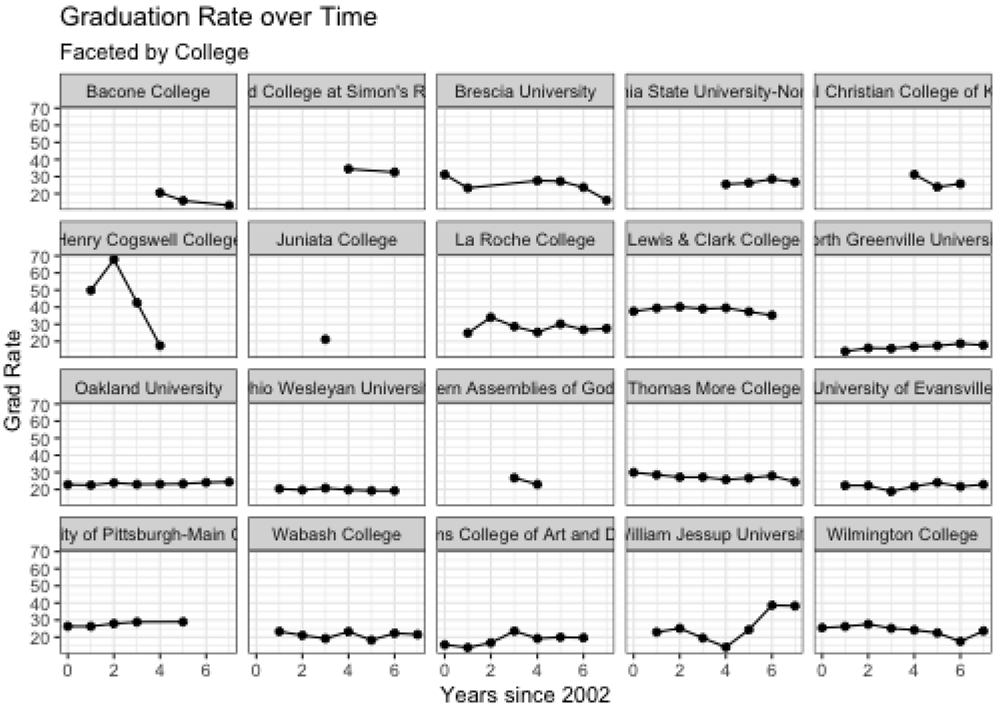
  - **within college**: changes over time within a school
  - **between college**: effects of school-specific predictors (e.g. `faculty`)

1. EDA: Univariate analysis

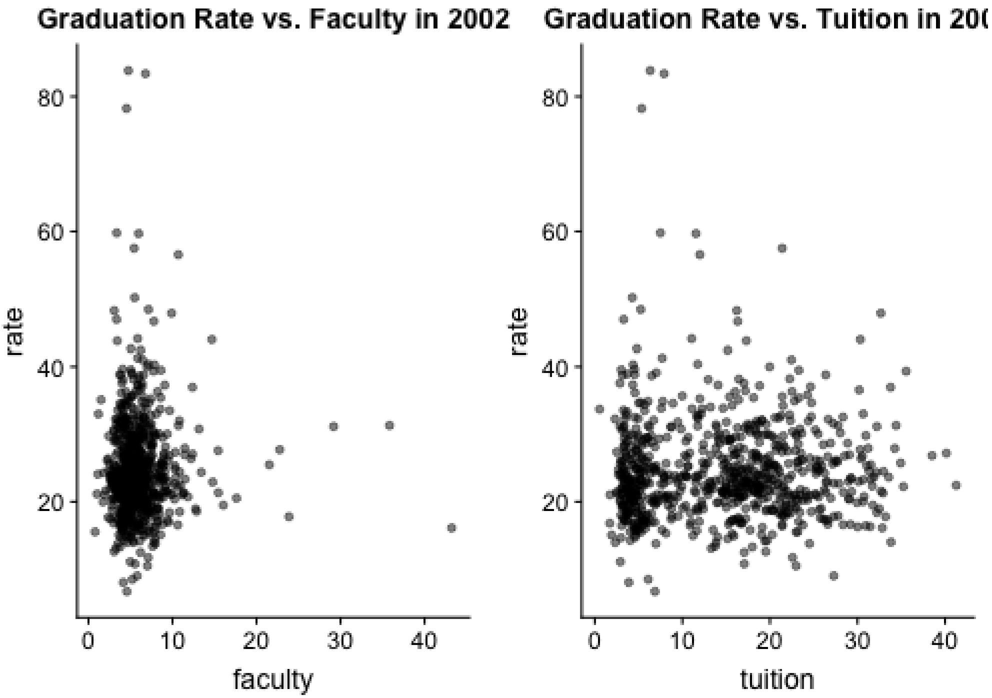


1. EDA: Graduation Rate by Year

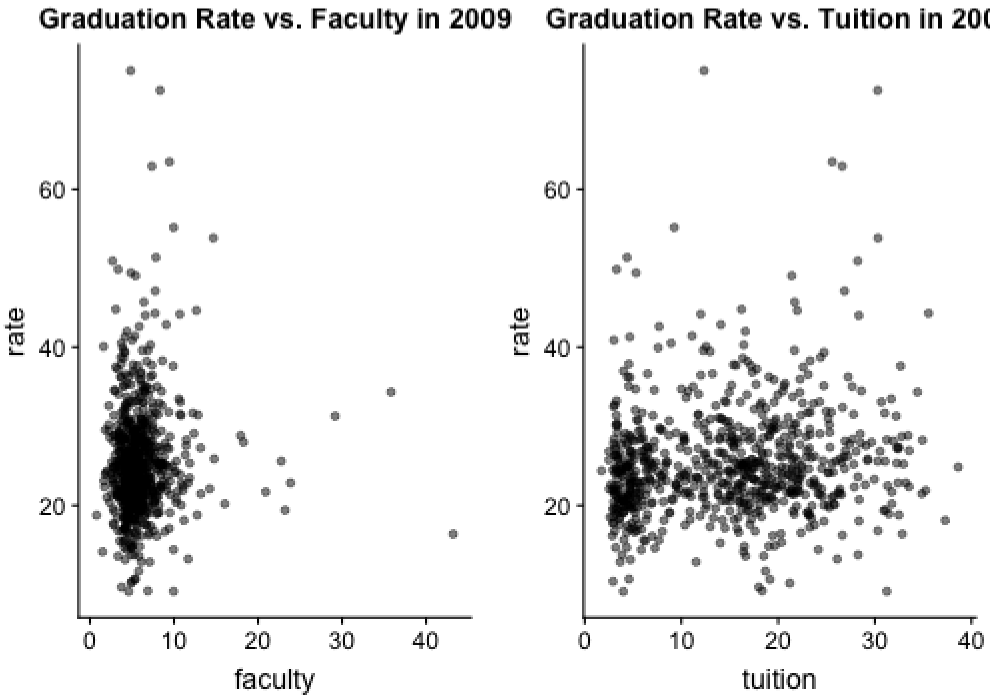
- Let's look at the rate over time for 20 randomly selected colleges



1. EDA: Bivariate Analysis in 2002



1. EDA: Bivariate Analysis in 2009



2. Unconditional Means Model

- In an **unconditional means model**, there are no predictors at any level
- The goal of this model is to compare variability with colleges and variability between colleges

Let Y_{ij} be the graduation rate of college i in year j

$$Y_{ij} = \alpha_0 + u_i + \epsilon_{ij}$$

$$u_i \sim N(0, \sigma_u^2) \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

- σ_u^2: variability between colleges
- σ^2: variability within a college

2. Unconditional Means Model

We can fit the unconditional means model using `lmer` function in the lme4 package.

```
library(lme4)

model_0 <- lmer(rate ~ 1 + (1|instname), data = college)
summary(model_0)
```

2. Grad. Rates: Unconditional Means Model

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rate ~ 1 + (1 | instname)
##    Data: college
## 
## REML criterion at convergence: 45218.7
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -8.2836 -0.3820 -0.0227  0.3498 17.4261 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  instname (Intercept) 53.485   7.313   
##  Residual              9.938   3.152   
## Number of obs: 7928, groups:  instname, 1337
## 
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  25.4904     0.2043   124.8
```

- Coefficients
  - $\hat{\alpha}_0 = 25.490$: mean graduation rate across all colleges
  - $\hat{\sigma}^2 = 9.938$: variance in within-school deviations between individual rate and college mean across all years
  - $\hat{\sigma}_u^2 = 53.485$: variance in the between-college deviations between the college means and the overall mean across all colleges and all years
- Intraclass correlation

$$\hat{\rho} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \hat{\sigma}^2} = \frac{53.485}{53.485 + 9.938} = 0.843$$

About 84.3% of the total variation in graduation rates can be attributed to the difference among schools rather than the change over time within schools. We can also say the average correlation for any two responses from the same college is about 0.843.

3. Unconditional growth model

- In an unconditional growth model, time is added to the random effects (level one) model but no predictors in the fixed effects (level two) model
- The goal of this model is to determine how much of the within-school variability in graduation rate can be attributed to changes over time
- We can think of this as building individual models for the change in graduation rate over time for each college
  - We assume the same form of the relationship between `rate` and `year` for every college

Let Y_{ij} be the `rate` for college i in year j

$$Y_{ij} = a_i + b_i \times \text{year02}_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

3. Unconditional growth model

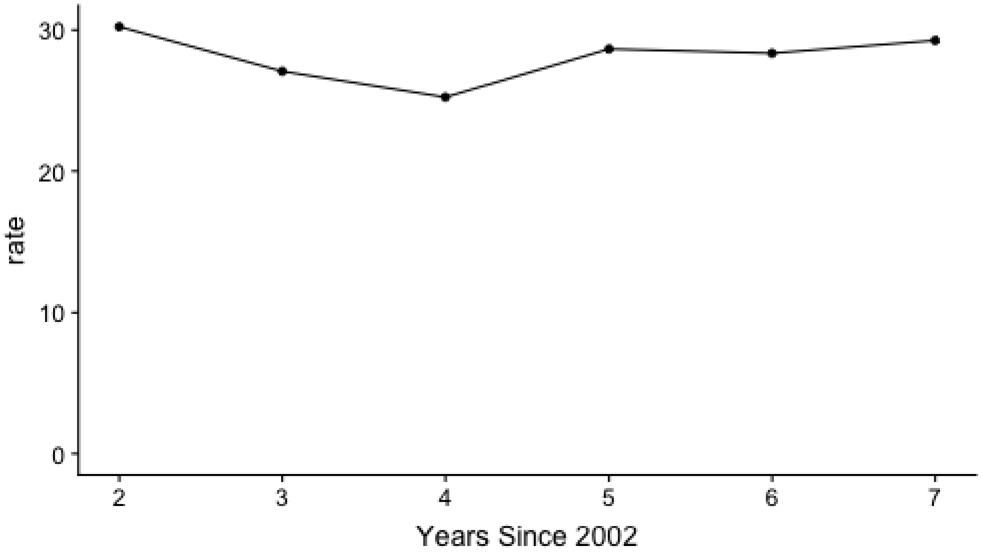
Let Y_{ij} be the `rate` for college i in year j

$$Y_{ij} = a_i + b_i \text{year02}_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} \sim N(0, \sigma^2)$$

- a_i: expected graduation rate for college i at time 0
- b_i: slope for college i, i.e. the rate of change in graduation rate for college i over the time period
- ϵ_{ij}: deviation in college i's expected and actual graduation rate at time j
    - σ^2 is the variability in the ϵ_{ij}'s

Duke: Graduation rate over time.



STA 210

3. Unconditional growth model

We will let a_i and b_i vary by college, so we can fit Level Two models that incorporate college-level variables to estimate these values

Let Y_{ij} be the graduation rate for college i in year j

Level One

$$Y_{ij} = a_i + b_i \text{year02}_{ij} + \epsilon_{ij}$$

Level Two

$$\begin{aligned} a_i &= \alpha_0 + u_i \\ b_i &= \beta_0 + v_i \end{aligned}$$

where $\epsilon_{ij} \sim N(0, \sigma^2)$ and

3. Unconditional growth model

- α_0: mean graduation rate for all colleges in 2002
- β_0: mean yearly change in graduation rate for all colleges during the time period
- σ^2: within-school variability
- σ_u^2: variability between colleges in the 2002 graduation rates
- σ_v^2: variability in the rate of change in the graduation rate (i.e. the slopes) 2002 - 2009
- σ_u^2 and σ_v^2 make up the between-school variability
- ρ_{uv}: Correlation between a college's graduation rate in 2002 and the rate of change of the graduation rate 2002 - 2009

Graduation rate: Unconditional growth model

```r
library(lme4)
model_1 <- lmer(rate ~ year02 + (year02|instname),
                data = college)
summary(model_1)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: rate ~ year02 + (year02 | instname)
##    Data: college
##
## REML criterion at convergence: 44669.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.9247 -0.3424 -0.0177  0.3194 14.9125
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  instname (Intercept) 59.1021  7.6878
##           year02       0.4807  0.6933   -0.31
##  Residual              7.7405  2.7822
## Number of obs: 7928, groups:  instname, 1337
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 24.98437    0.22650 110.307
## year02       0.13751    0.02645   5.198
##
## Correlation of Fixed Effects:
##        (Intr)
## year02 -0.433
```

Understanding the model

What do each of the following values tell you?

- $\hat{\alpha}_0 = 24.98$:
- $\hat{\beta}_0 = 0.14$:
- $\hat{\sigma}^2 = 7.74$:
- $\hat{\sigma}_u^2 = 59.10$:
- $\hat{\sigma}_v^2 = 0.48$:
- $\rho_{uv} = -0.31$:

4. Add predictors

- Do `faculty` and `tuition` affect graduation rates?
- We will add these predictor variables to the fixed effects (level two) model, since they differ by college but don't change within a college in our data

Let Y_{ij} be the `rate` for college i in year j

Level One

$$Y_{ij} = a_i + b_i \times \text{year02}_{ij} + \epsilon_{ij}$$

Level Two

$$a_i = \alpha_0 + \alpha_1 \times \text{faculty}_i + \alpha_2 \times \text{tuition}_i + u_i$$
$$b_i = \beta_0 + \beta_1 \times \text{faculty}_i + \beta_2 \times \text{tuition}_i + v_i$$

4. Add predictors

```r
library(lme4)
model_2 <- lmer(rate ~ year02 + faculty + tuition  +
                  faculty:year02 + tuition:year02 +
                  (year02|instname), data = college)
summary(model_2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## rate ~ year02 + faculty + tuition + faculty:year02 + tuition:year02 +
##     (year02 | instname)
##    Data: college
##
## REML criterion at convergence: 44689.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -7.9022 -0.3414 -0.0162  0.3172 14.9148
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  instname (Intercept) 58.9803  7.6799
##           year02       0.4843  0.6959   -0.31
##  Residual              7.7326  2.7808
## Number of obs: 7928, groups:  instname, 1337
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)      2.416e+01  4.397e-01  54.934
## year02           2.132e-01  6.323e-02   3.372
## faculty          1.663e-01  6.472e-02   2.569
## tuition         -1.061e-02  2.309e-02  -0.459
## year02:faculty  -1.284e-02  9.308e-03  -1.379
## year02:tuition   1.169e-05  3.209e-03   0.004
```

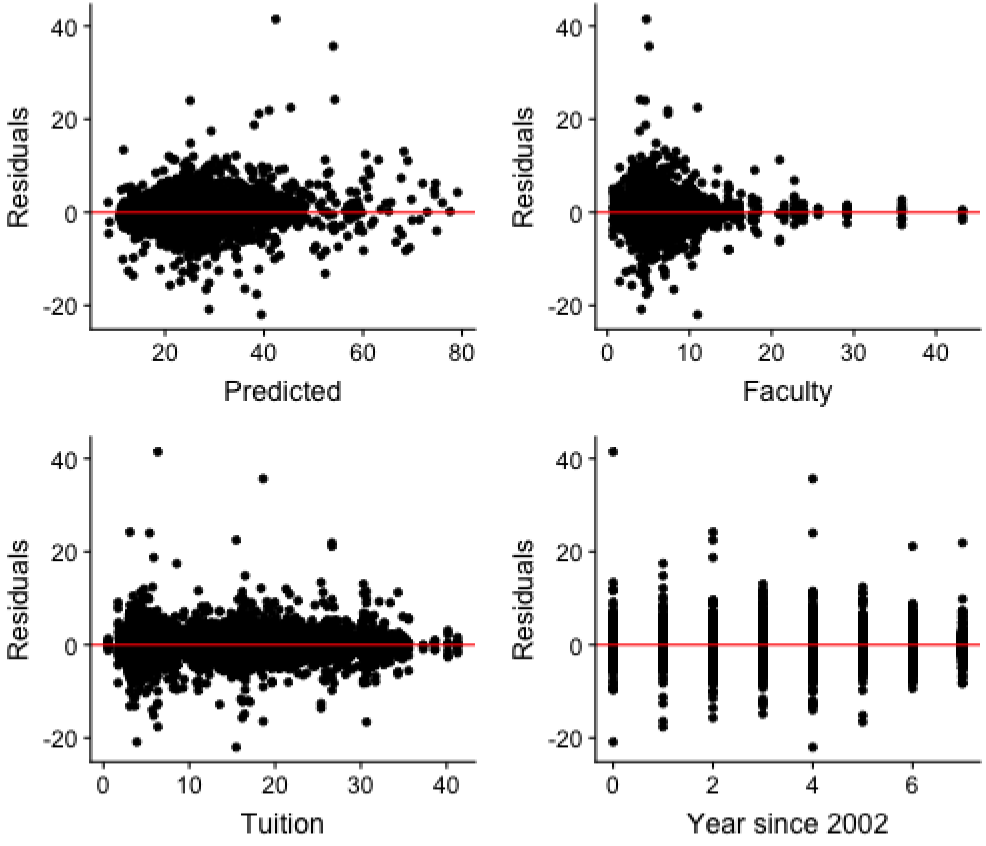
Write out the model

Predicted values

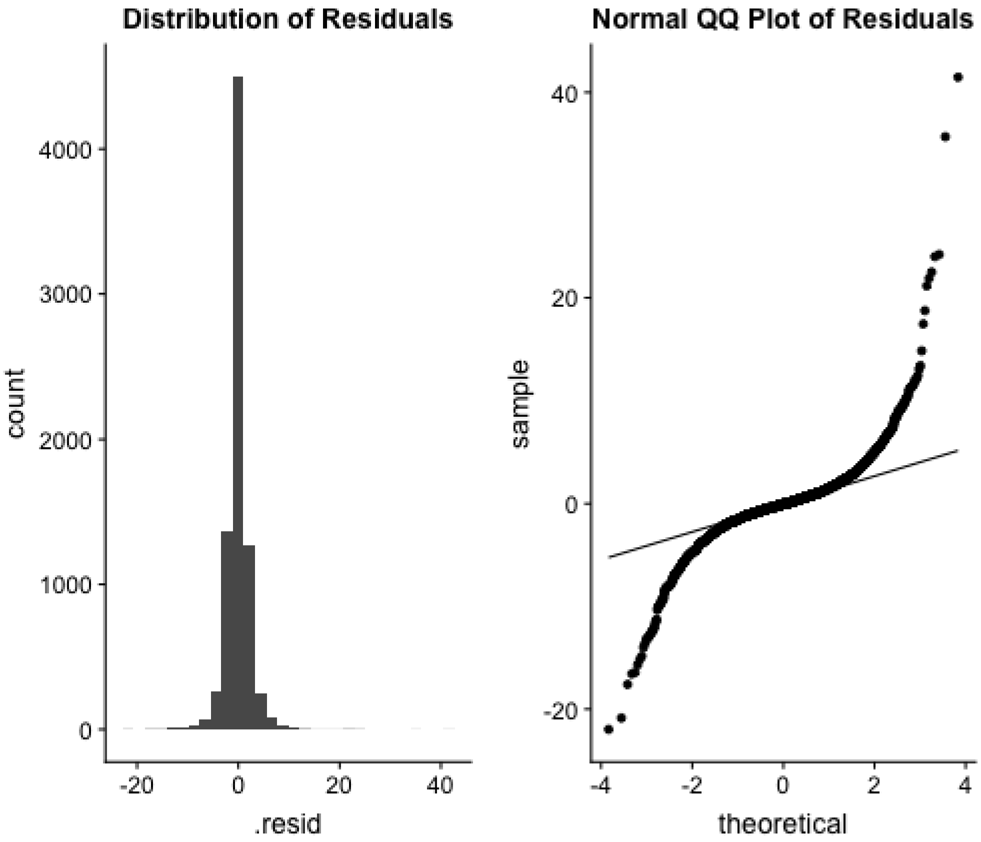
We can use the `augment` function to get predicted values and residuals

```
model_2_aug <- augment(model_2)
```

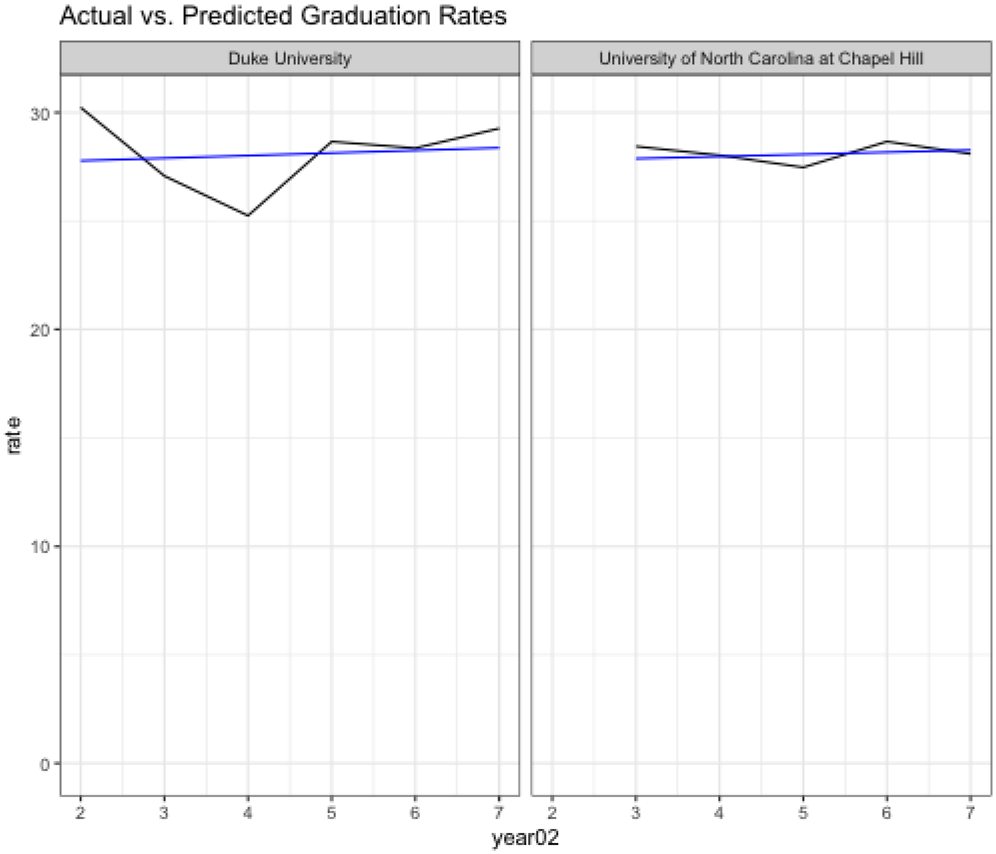
Check Residuals



Residuals


Distribution of Residuals
count
4000
3000
2000
1000
0
-20
0
20
40
.resid
Normal QQ Plot of Residuals
sample
40
20
0
-20
-4
-2
0
2
4
theoretical

Actual vs. Predicted Graduation Rates



References

Broadening Your Statistical Horizons

- "Introduction to Multilevel Models" -
- "Two Level Longitudinal Data"

Congrats on completing STA 210! 😄