

Activity — Reproducible Analysis with R

Understanding the data

There are 1682 students in the data. These students have a total of 1193 unique teachers and 760 unique schools.

Let's take a peek at the data:

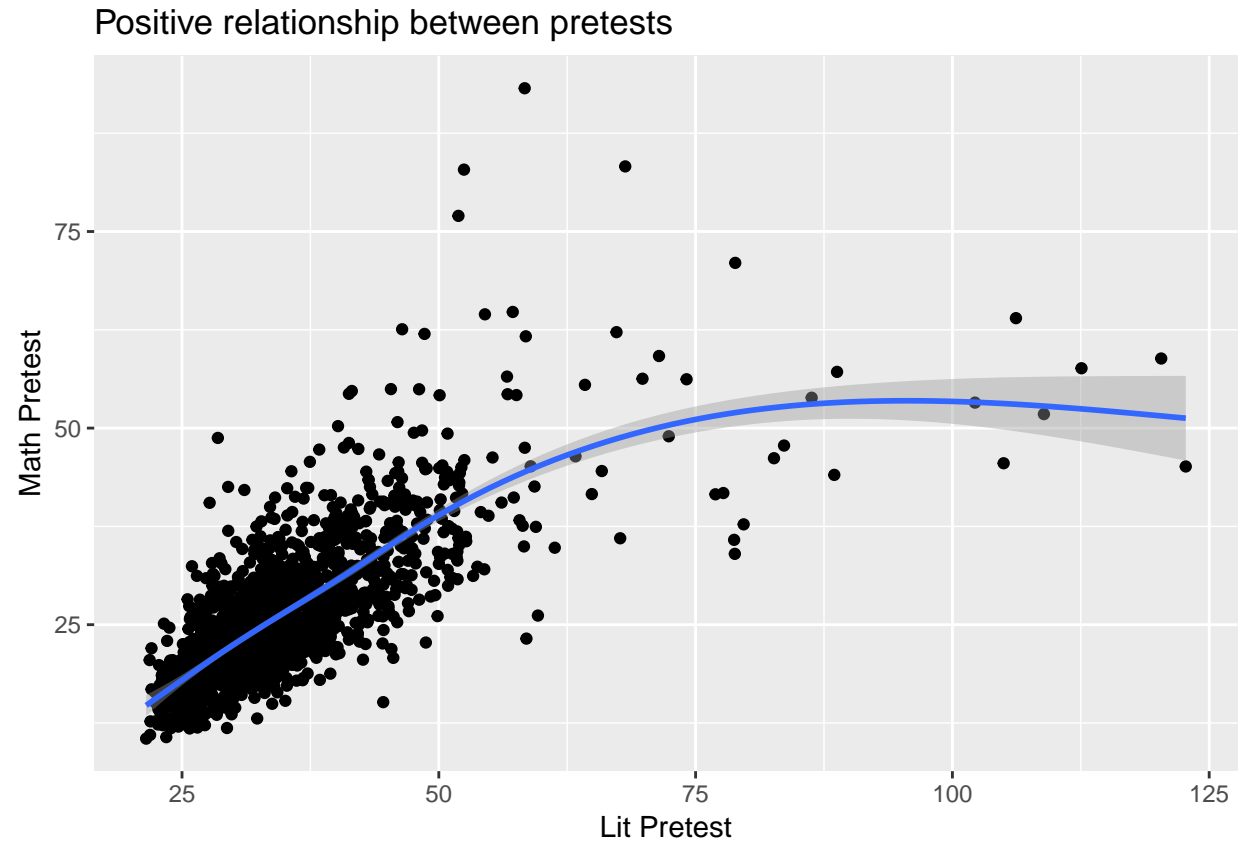
school_id	teach_id	child_id	lit_pretest	math_pretest	lit_posttest	math_posttest	ses	female	black	h
0194	0194T03	0194021C	24.88	12.92	35.07	23.89	-0.33	0	0	
0089	0089T04	0089013C	29.44	20.07	40.35	29.75	0.77	1	0	
0343	0343T01	0343024C	33.35	31.59	67.49	52.89	0.17	1	0	
0282	0282T02	0282017C	29.45	25.65	30.71	29.47	0.98	0	0	
0855	0855T01	0855020C	42.50	34.97	40.71	34.78	0.06	0	0	
0155	0155T01	0155009C	31.67	29.11	49.47	55.58	0.84	0	0	

Let's see how many times each of the top 10 most common schools show up in the data

```
data %>%  
  count(school_id, sort = TRUE) %>%  
  slice(1:10) %>%  
  kable()
```

school_id	n
0404	7
0011	6
0058	6
0105	6
0229	6
0458	6
0626	6
0635	6
0776	6
0808	6

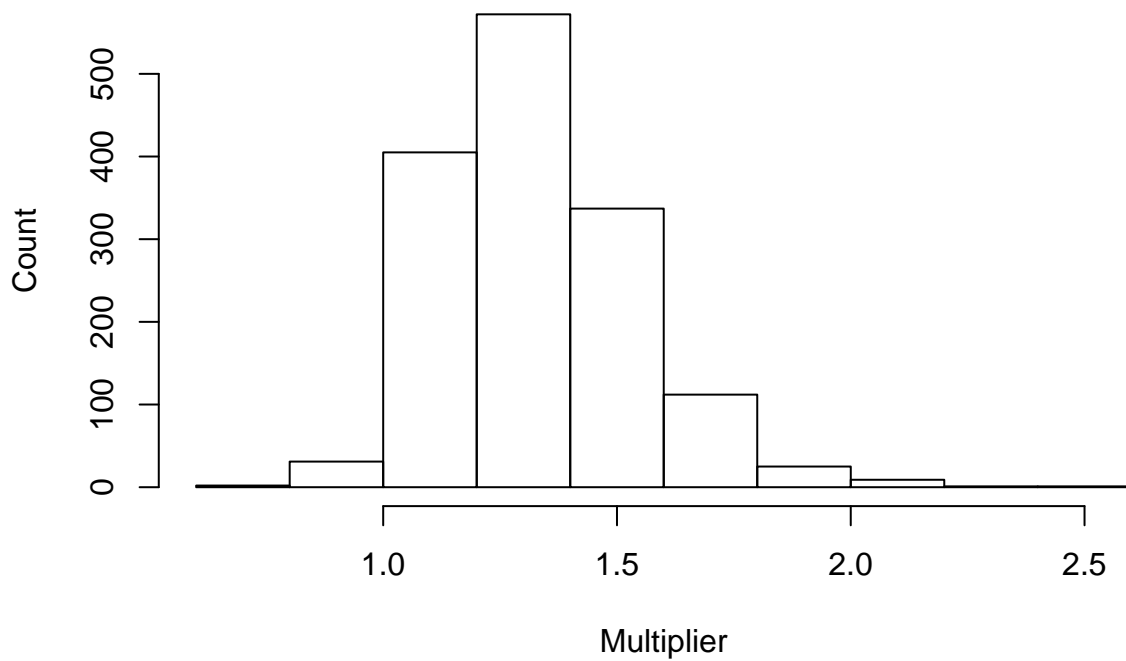
Relationship between lit and math pretests



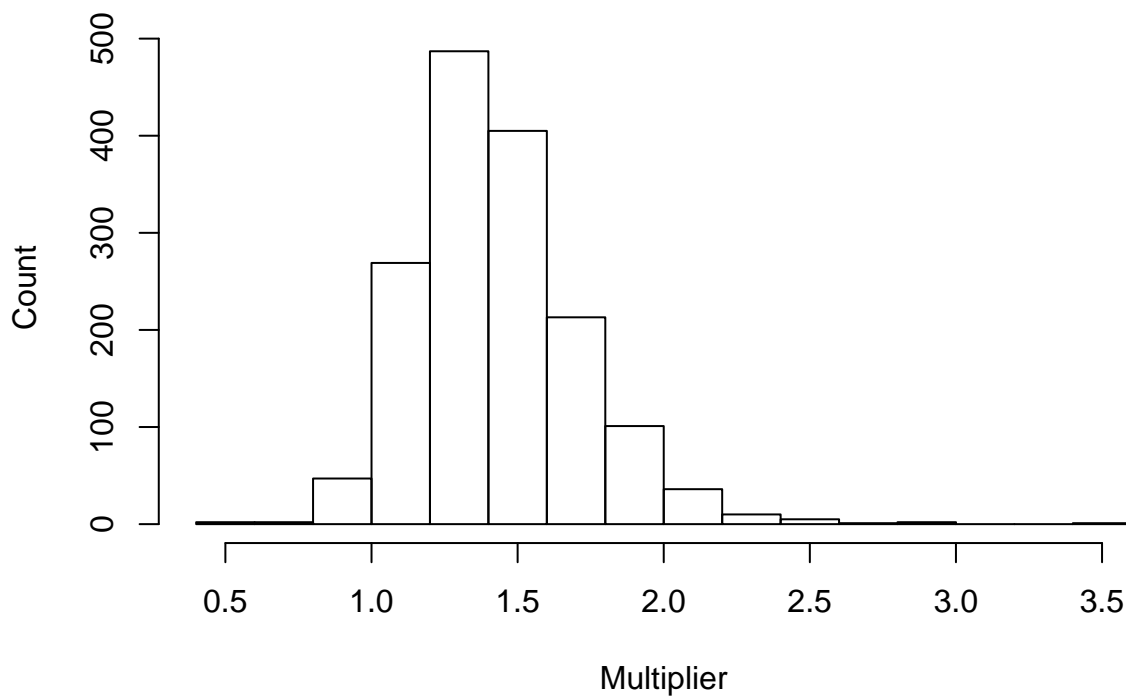
Growth

Histograms of growth

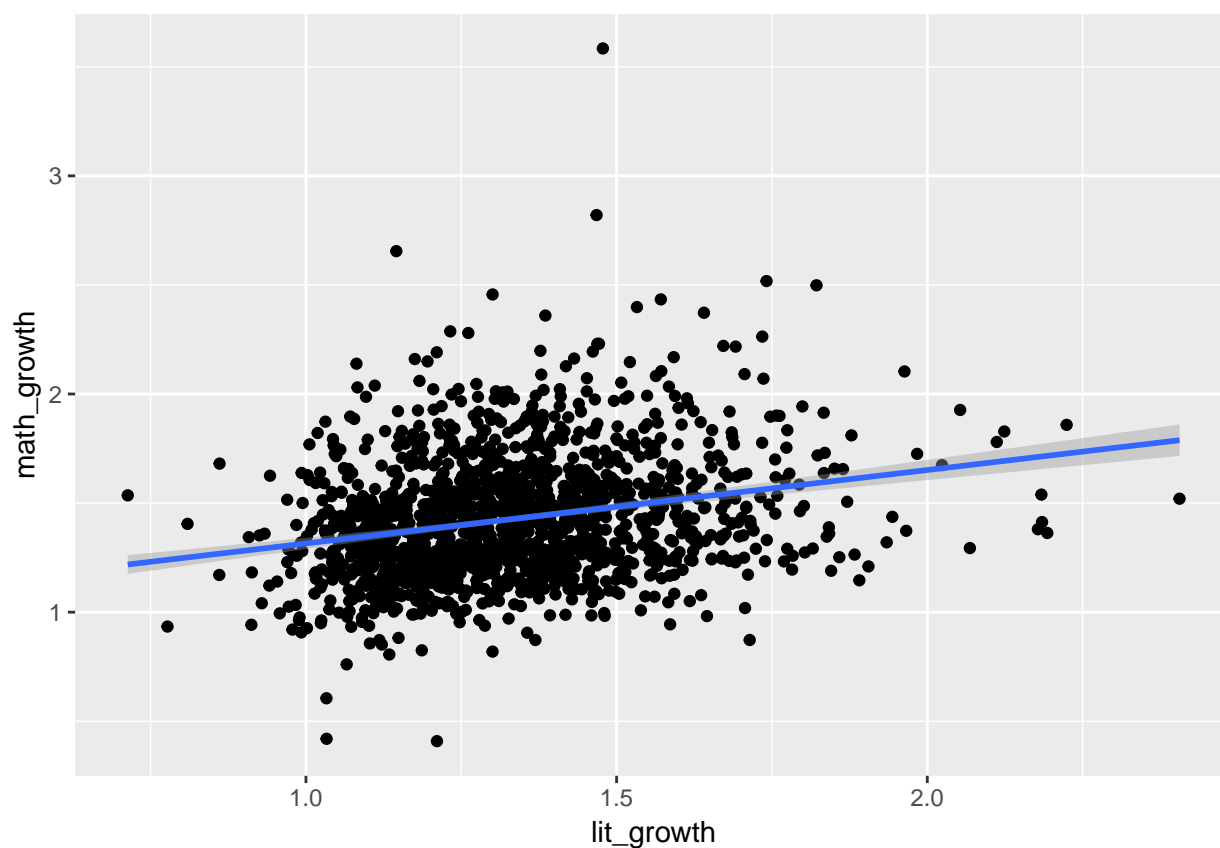
Lit Growth



Math Growth



Relationship between growth



Predicting math growth based on covariates

```
##
## Call:
## lm(formula = math_growth ~ ses + female + black + hisp + asian,
##     data = data_with_growth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03070 -0.19678 -0.03289  0.15373  2.18556
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.42888    0.01176 121.528  <2e-16 ***
## ses           -0.02348    0.01014  -2.315   0.0207 *
## female        -0.00540    0.01481  -0.365   0.7155
## black         -0.03991    0.02243  -1.780   0.0753 .
## hisp           0.04734    0.02054   2.305   0.0213 *
## asian          0.01108    0.03347   0.331   0.7407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.287 on 1512 degrees of freedom
```

```
## (164 observations deleted due to missingness)
## Multiple R-squared:  0.01254,    Adjusted R-squared:  0.00927
## F-statistic: 3.839 on 5 and 1512 DF,  p-value: 0.001847
```

For example, the model finds that the coefficient on SES is -0.02. This tells us that...

Ideas for things to add

- Model lit growth as well
- Only some covariates were included, can you include other covariates?
- Can you turn the insights from the model into crisp visualizations?