Analysis and Modeling of Student Grades

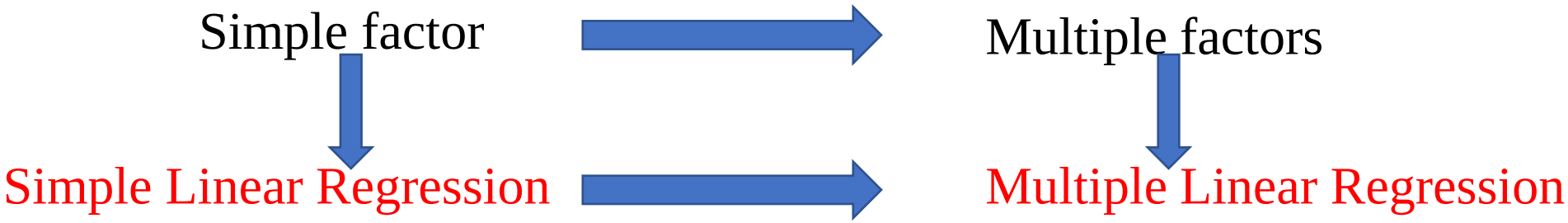
Backgrounds & Our data

- Nowadays, students often attach great importance to their grades. They spend a lot of time on study to obtain high scores. But are there any other indirect factors that make a big difference to the scores? We explore these factors in the project.

- 3 kinds of scores — Numerical Variables
  - math score
  - reading score
  - writing score
- 5 kinds of factors — Categorical Variables
  - gender (female, male)
  - race/ethnicity (group A B C D E)
  - parental level of education (associate's degree, bachelor's degree, master's degree, high school, some high school, some college)
  - lunch (standard, free/reduced)
  - test preparation course (completed, none)

Ideas & Research Questions

✓ Which factors are influential ? (focused on mean scores)



✓ Is our model good? How to assess and improve the model?

Do predictions

2/3 train data
1/3 test data

Full model : include all variables

```
## Coefficients:
##                                          Estimate Std. Error t value
## (Intercept)                               48.8167     1.9647  24.847
## genderfemale                               3.9043     0.9942   3.927
## race.ethnicitygroup B                      1.4708     1.9292   0.762
## race.ethnicitygroup C                      3.2672     1.7704   1.845
## race.ethnicitygroup D                      5.4096     1.8401   2.940
## race.ethnicitygroup E                      6.8479     2.0456   3.348
## parental.level.of.educationsome high school  1.7393     1.5741   1.105
## parental.level.of.educationsome college    5.6872     1.4865   3.826
## parental.level.of.educationassociate's degree  5.5783     1.5117   3.690
## parental.level.of.educationbachelor's degree  8.7805     1.8190   4.827
## parental.level.of.educationmaster's degree  9.1755     2.3781   3.858
## lunchstandard                              9.0802     1.0285   8.829
## test.preparation.coursecompleted           7.8593     1.0238   7.677
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## genderfemale                             9.51e-05 ***
## race.ethnicitygroup B                    0.446097
## race.ethnicitygroup C                    0.065424 .
## race.ethnicitygroup D                    0.003399 **
## race.ethnicitygroup E                    0.000862 ***
## parental.level.of.educationsome high school  0.269595
## parental.level.of.educationsome college  0.000143 ***
## parental.level.of.educationassociate's degree 0.000243 ***
## parental.level.of.educationbachelor's degree  1.73e-06 ***
## parental.level.of.educationmaster's degree  0.000125 ***
## lunchstandard                             < 2e-16 ***
## test.preparation.coursecompleted         5.98e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.69 on 653 degrees of freedom
## Multiple R-squared:  0.2438,  Adjusted R-squared:  0.2299
## F-statistic: 17.54 on 12 and 653 DF,  p-value: < 2.2e-16
```

Variable	Category	Pass significance level
gender	female	**0.001**
race/ethnicity	group B	**1**
	group C	**0.1**
	group D	**0.01**
	group E	**0.001**
parental level of education	some high school	**1**
	some colleg	**0.001**
	associate's degree	**0.001**
	bachelor's degree	**0.001**
	master's degree	**0.001**
lunch	standard	**0.001**
test preparation course	completed	**0.001**

Maybe removing variable **race/ethnicity** will optimize the model.

Reduced model : remove race/ethnicity

```
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   52.0991     1.3951  37.344
## genderfemale                                   3.8380     0.9986   3.843
## parental.level.of.educationsome high school    1.8237     1.5867   1.149
## parental.level.of.educationsome college        6.3380     1.4918   4.248
## parental.level.of.educationassociate's degree  6.0291     1.5225   3.960
## parental.level.of.educationbachelor's degree   9.1503     1.8333   4.991
## parental.level.of.educationmaster's degree     9.9637     2.3904   4.168
## lunchstandard                                  9.1695     1.0380   8.834
## test.preparation.coursecompleted               7.7895     1.0326   7.544
##                                              Pr(>|t|)
## (Intercept)                                   < 2e-16 ***
## genderfemale                                 0.000133 ***
## parental.level.of.educationsome high school  0.250807
## parental.level.of.educationsome college      2.46e-05 ***
## parental.level.of.educationassociate's degree 8.31e-05 ***
## parental.level.of.educationbachelor's degree  7.69e-07 ***
## parental.level.of.educationmaster's degree    3.48e-05 ***
## lunchstandard                                 < 2e-16 ***
## test.preparation.coursecompleted              1.52e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.82 on 657 degrees of freedom
## Multiple R-squared:  0.2228, Adjusted R-squared:  0.2134
## F-statistic: 23.55 on 8 and 657 DF,  p-value: < 2.2e-16
```

- A more credible model
  - Only one category failed to pass a t-test with significance level of 0.001
  - Most categories' p-values dropped. The smaller the Pr(> |t|), the more significant the variable.
  - Not perfect, but works better.

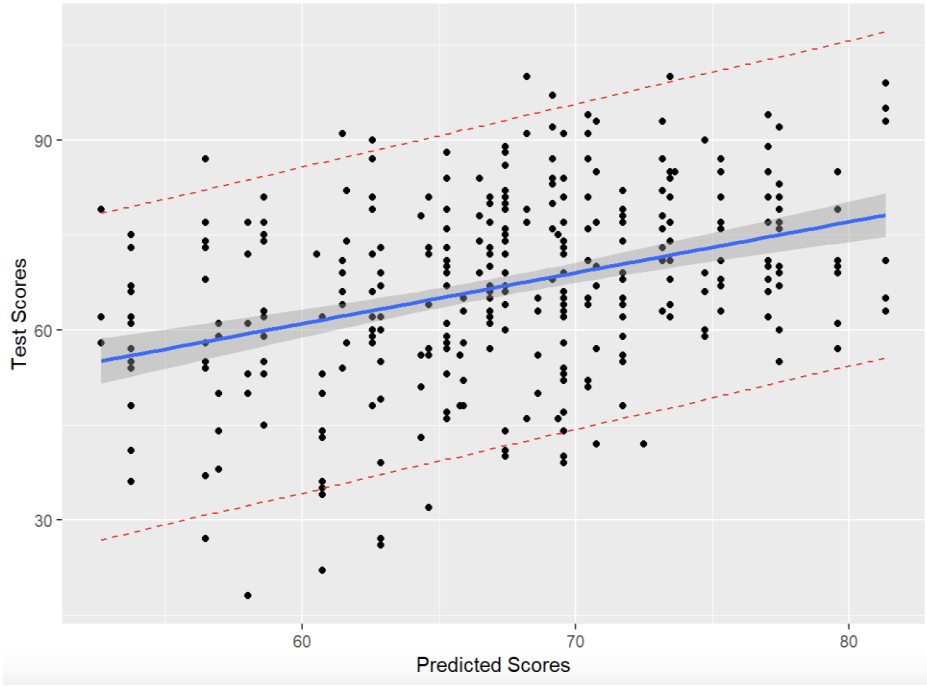
Using this model, the estimate formula for mean score is :

$$\hat{Y} = 52.0991 + 3.8380 \cdot G + 1.8237 \cdot P_1 + 6.3380 \cdot P_2 + 6.0291 \cdot P_3 + 9.1503 \cdot P_4 + 9.9637 \cdot P_5 + 9.1695L + 7.7895 \cdot T$$

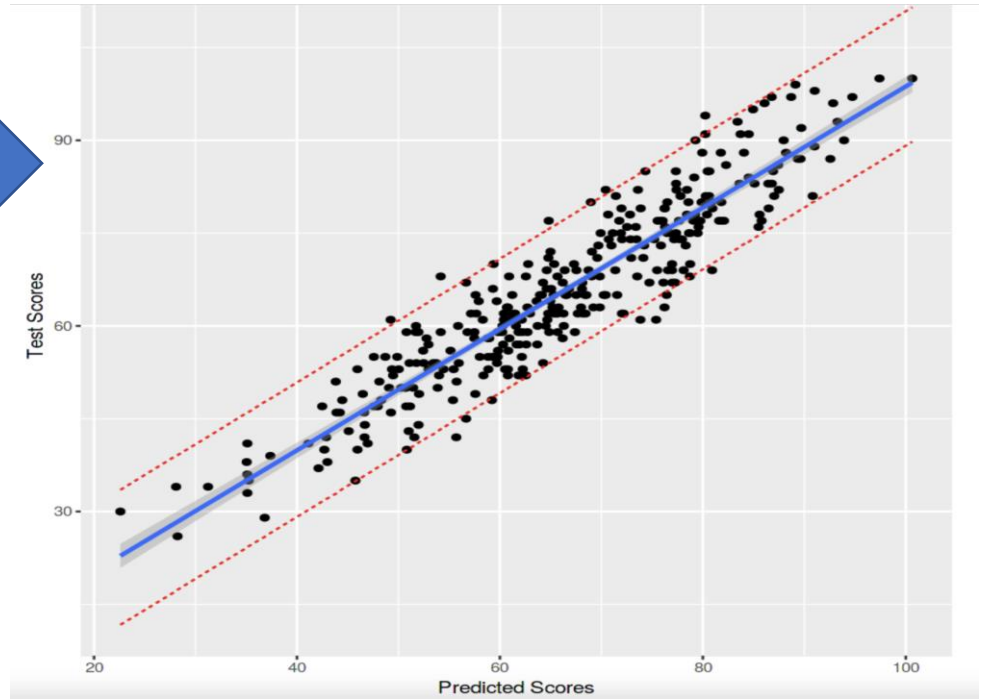
Estimate for mean score of reference group:

Gender : male;
Parental level of education : high school;
Lunch : free/reduced;
Test preparation course : none.

Prediction



- Analysis
  - The prediction is not strong enough. The 95% CI interval is quite broad.
  - We only use categorical variables.
  - The data set miss some strongly relevant variables. eg. Studying time.



- The prediction is stronger when we add reading and writing scores as variables to predict math score.