Stats 101a
Final Project

By Amanda Li, Carolyn Wang, and Devin Yao

01

Introduction

Purpose and Guidelines

Introduction:

- For our report, we will be analyzing and summarizing the results of the math scores of 1,000 high school students in the United States
- We will be studying the effects of predictor variables such as writing scores, reading scores, lunch, parental education level, and gender on a student's math score
- Our data's relationship will be captured through the Multiple Linear Regression Model

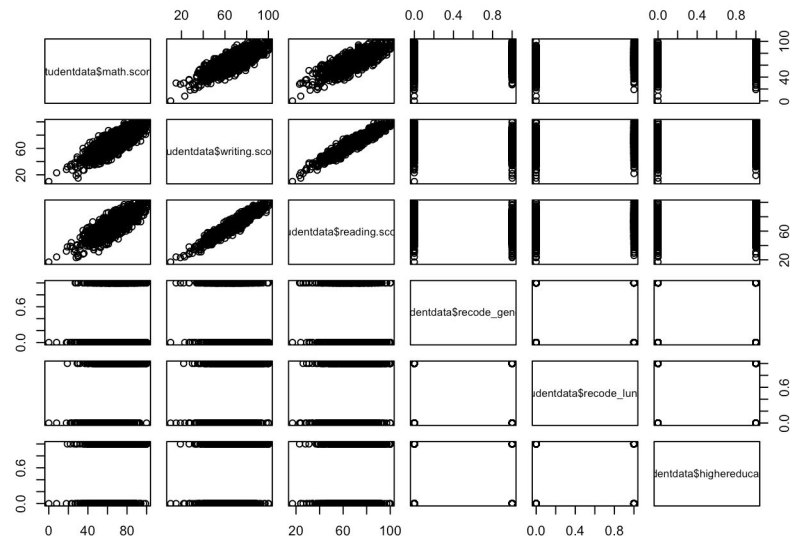
General Overview:

- 1) Discuss summary statistics and correlation of variable relationships
- 2) Check if all predictor variables are statistically significant
- 3) Verify all model assumptions are met and transform data if needed
- 4) Perform Partial F-Test and check Adjusted R-Squared, AIC, AICc, BIC to find the best subset model and come to our final model!

02

Data Description

Statistical Summary and Correlation



Summary Statistics:

	Mean	Standard Deviation
Math	66.09	15.1631
Writing	68.05	15.1957
Reading	69.17	14.6002

	0	1
Higher Education	375	625
Reduced Lunch	355	645
Male	518	482

- Each predictor variable besides Lunch has a relatively strong, positive, linear correlation with math score.
- The lunch predictor variable, instead, has a relatively strong, negative, linear correlation with math score.
- There is also correlation between the predictor variables writing score and reading score

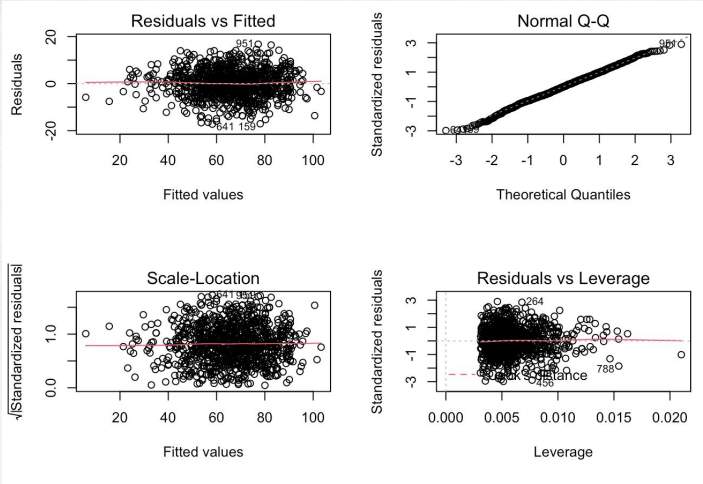


03

Results/Interpretation

Original Model

```
## Call:
## lm(formula = studentdata$math.score ~ studentdata$reading.score +
##     studentdata$writing.score + studentdata$gender + studentdata$reducedlunch +
##     studentdata$highereducation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.2305  -3.7311   0.0352   3.8788  16.7617
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2.36311    1.05197  -2.246   0.0249 *
## studentdata$reading.score    0.39430    0.04306   9.158   <2e-16 ***
## studentdata$writing.score    0.53638    0.04281  12.528   <2e-16 ***
## studentdata$gendermale      12.73499    0.39210  32.479   <2e-16 ***
## studentdata$reducedlunch    -3.88305    0.39862  -9.741   <2e-16 ***
## studentdata$highereducation -0.13336    0.39333  -0.339   0.7346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.795 on 994 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8539
## F-statistic:  1169 on 5 and 994 DF,  p-value: < 2.2e-16
```



math.score = −2.363 + 0.536writing.score + 0.394reading.score + 12.735gendermale−3.883reducedlunch−0.133highereducation

Transformation

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## Y1    0.9611            1       0.8495       1.0728
##
## Likelihood ratio test that transformation parameter is equal to 0
##  (log transformation)
##                              LRT df       pval
## LR test, lambda = (0) 393.3511  1 < 2.22e-16
##
## Likelihood ratio test that no transformation is needed
##                               LRT df   pval
## LR test, lambda = (1) 0.4589779  1 0.4981
```

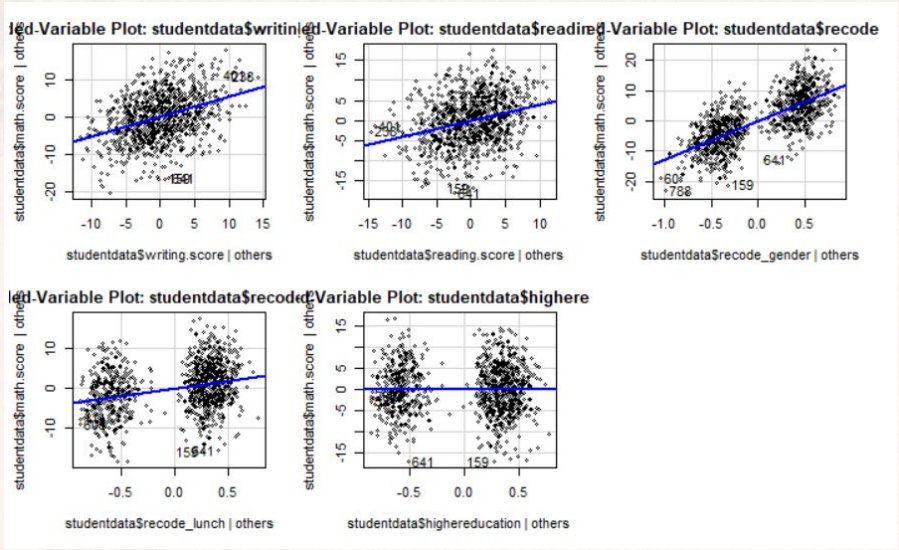
- Because our original model met the model assumptions, it seems like transformation is not needed. Also, to prove our sense of the data, we used box cox to see if there was a better candidate.
- The results of the box cox transformation confirmed that no transformation is needed so we proceeded with our analysis using the original model.

Variable Selection

```
studentdata$writing.score  studentdata$reading.score
                12.591824                  11.756230
      studentdata$gender                studentdata$lunch
                 1.143093                   1.083484
studentdata$highereducation
                 1.079763
```

- Variable “higher education ” was not significant.
- VIFs suggests great collinearity between variable reading score and writing score.
- Used the added variable plot to explore the pure relationship between each variable. The results are consistent with the R output.

```
## Call:
## lm(formula = studentdata$math.score ~ studentdata$reading.score +
##     studentdata$writing.score + studentdata$gender + studentdata$reducedlunch +
##     studentdata$highereducation)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.2305  -3.7311   0.0352   3.8788  16.7617
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -2.36311    1.05197  -2.246   0.0249 *
## studentdata$reading.score       0.39430    0.04306   9.158   <2e-16 ***
## studentdata$writing.score       0.53638    0.04281  12.528   <2e-16 ***
## studentdata$gendermale         12.73499    0.39210  32.479   <2e-16 ***
## studentdata$reducedlunch       -3.88305    0.39862  -9.741   <2e-16 ***
## studentdata$highereducation -0.13336    0.39333  -0.339   0.7346
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.795 on 994 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8539
## F-statistic:  1169 on 5 and 994 DF,  p-value: < 2.2e-16
```



All possible subset

```
Subset selection object
5 Variables  (and intercept)
  Forced in Forced out
a     FALSE      FALSE
b     FALSE      FALSE
c     FALSE      FALSE
d     FALSE      FALSE
e     FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         a   b   c   d   e
1  ( 1 ) " " "*" " " " " " "
2  ( 1 ) "*" " " "*" " " " "
3  ( 1 ) "*" " " "*" "*" " "
4  ( 1 ) "*" "*" "*" "*" " "
5  ( 1 ) "*" "*" "*" "*" "*"
```

```
The table of 5 models values is showed below:
     [,1] [,2]                [,3]                [,4]                [,5]
[1,] "m1" "0.66810427850066"  "4336.79061033197" "4338.97242851379" "4337.92050904689"
[2,] "m2" "0.828547754772404" "3677.27205061972" "3681.27205061972" "3678.9668986921"
[3,] "m3" "0.841578700766174" "3599.22169909601" "3605.88836576267" "3601.48149652585"
[4,] "m4" "0.854074134697404" "3518.05796431775" "3528.55796431775" "3520.88271110506"
[5,] "m5" "0.85394422077126"  "3519.94231122516" "3535.94231122516" "3523.33200736993"
```

The All possible subset suggest that the “best” model is
math.score ~
writing.score +reading.score + gendermale +reducedlunch

Forward/backward AIC ,BIC, Partial F test

The Forward, Backward AIC BIC suggest that our “best” model is:

math.score ~
writing.score +reading.score + gendermale +reducedlunch

```
Start:  AIC=3518.06
Y ~ x1 + x2 + x3 + x4

       Df Sum of Sq   RSS    AIC
<none>              33383 3518.1
- x2    1      2895 36278 3599.2
- x4    1      3223 36606 3608.2
- x1    1      5431 38814 3666.8
- x3    1     35508 68891 4240.5

The best model recomendation from AIC is:Y ~ x1 + x2 + x3 + x4
```

```
Start:  AIC=3542.6
Y ~ x1 + x2 + x3 + x4

       Df Sum of Sq   RSS    AIC
<none>              33383 3542.6
- x2    1      2895 36278 3618.9
- x4    1      3223 36606 3627.9
- x1    1      5431 38814 3686.4
- x3    1     35508 68891 4260.2

The best model recomendation from AIC is:Y ~ x1 + x2 + x3 + x4
```

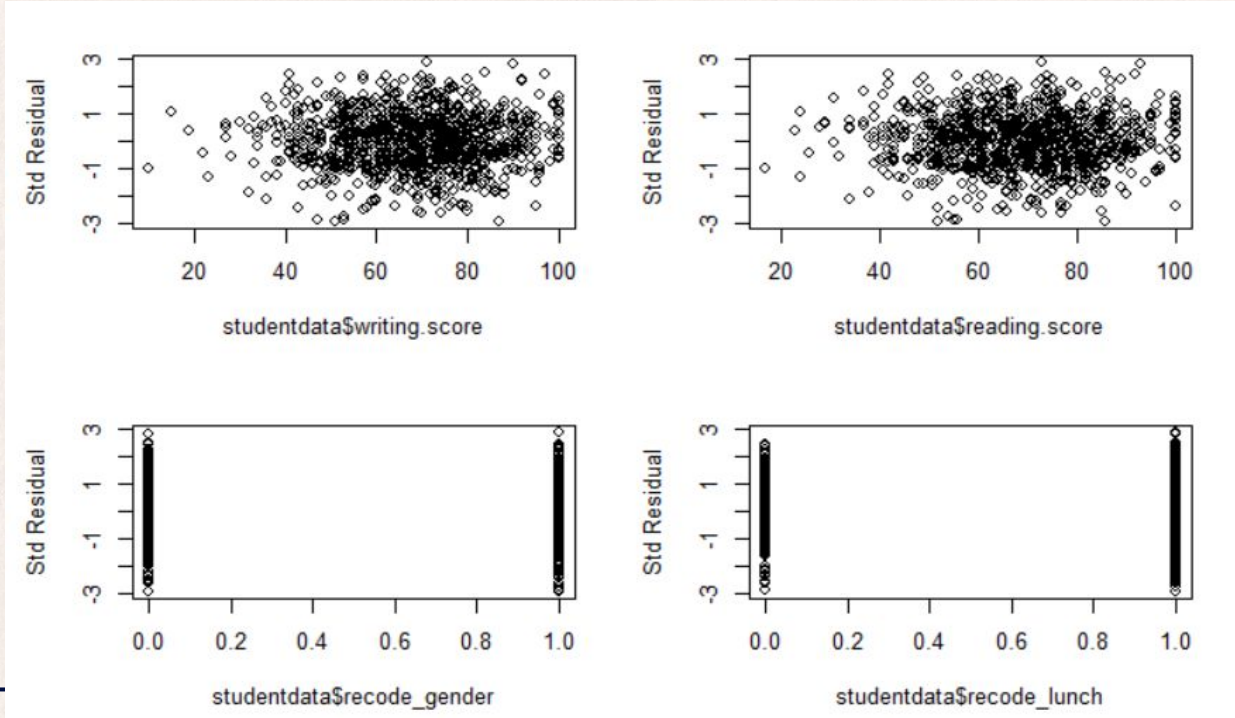
The partial F test proves that the selected model is better

```
tudentdata$math.score ~ studentdata$writing.score + studentdata$reading.score +
tdata$gender + studentdata$lunch
tudentdata$math.score ~ studentdata$writing.score + studentdata$reading.score +
tdata$writing.score + studentdata$gender + studentdata$reducedlunch +
tdata$highereducation
 RSS Df Sum of Sq      F Pr(>F)
3383
3380  1    3.8607 0.115 0.7346
```

```
## Start:  AIC=5438.73
## Y ~ 1
##
##         Df Sum of Sq    RSS    AIC
## + x2     1    153533  76157 4336.8
## + x1     1    147974  81716 4407.2
## + x4     1     28278 201411 5309.3
## + x5     1      6543 223146 5411.8
## + x3     1      6481 223208 5412.1
## <none>               229689 5438.7
##
## Step:  AIC=4336.79
## Y ~ x2
##
##         Df Sum of Sq   RSS    AIC
## + x3     1     33031 43126 3770.1
## + x4     1      6457 69699 4250.2
## + x1     1      1274 74883 4321.9
## <none>              76157 4336.8
## + x5     1       108 76049 4337.4
##
## Step:  AIC=3770.12
## Y ~ x2 + x3
##
##         Df Sum of Sq   RSS    AIC
## + x1     1    6519.2 36606 3608.2
## + x4     1    4311.7 38814 3666.8
## <none>              43126 3770.1
## + x5     1      85.9 43040 3770.1
##
## Step:  AIC=3608.22
## Y ~ x2 + x3 + x1
##
##         Df Sum of Sq   RSS    AIC
## + x4     1    3223.1 33383 3518.1
## <none>              36606 3608.2
## + x5     1      40.4 36566 3609.1
##
## Step:  AIC=3518.06
## Y ~ x2 + x3 + x1 + x4
##
##         Df Sum of Sq   RSS    AIC
## <none>              33383 3518.1
## + x5     1    3.8607 33380 3519.9
```

Final Model Diagnosis

Standardized residual Plot



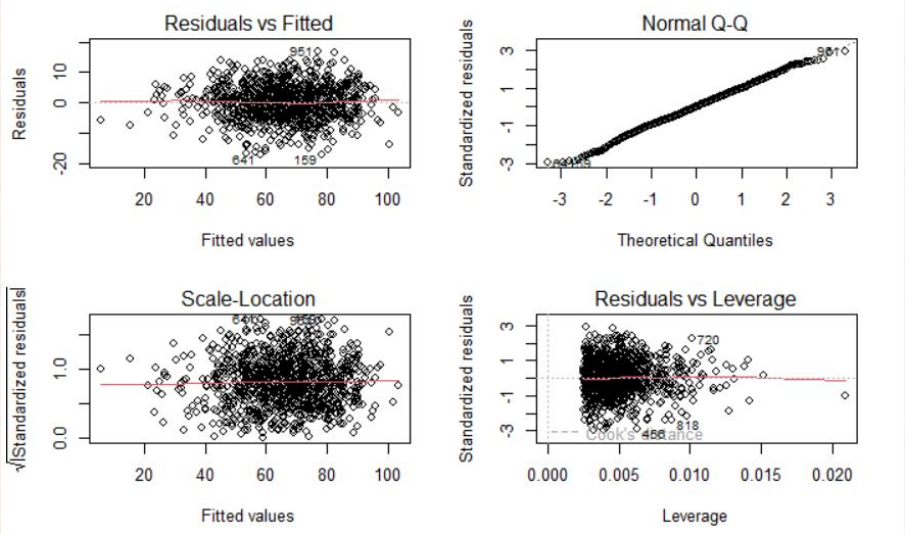
Final Model Diagnosis

VIF

```
> vif(m1_reduced)
studentdata$writing.score studentdata$reading.score        studentdata$gender 
                 12.088952                 11.548979                  1.138949 
       studentdata$lunch 
                  1.076974 
```

Model diagnosis plot

From the assumption of the standardized residual plot and the diagnosis plot. We conclude that the reduced model is valid



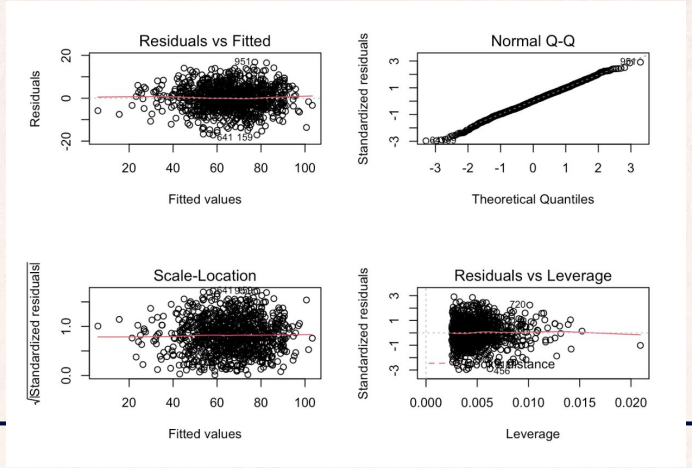
Final Model

```
## Call:
## lm(formula = studentdata$math.score ~ studentdata$reading.score +
##     studentdata$writing.score + studentdata$gender + studentdata$reducedlunch)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1629  -3.7738   0.0457   3.8532  16.8466
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2.37553    1.05086  -2.261    0.024 *
## studentdata$reading.score   0.39624    0.04266   9.289   <2e-16 ***
## studentdata$writing.score   0.53347    0.04193  12.722   <2e-16 ***
## studentdata$gendermale     12.72698    0.39122  32.532   <2e-16 ***
## studentdata$reducedlunch   -3.89352    0.39725  -9.801   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.792 on 995 degrees of freedom
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8541
## F-statistic:  1463 on 4 and 995 DF,  p-value: < 2.2e-16
```

- This model is stronger than the original model since all variables coefficients are significant.
- Meets model assumptions, and has undergone variable selection, removing redundancy in the model.

- We removed the variable on parental higher education after applying the 4 methods of variable selection, resulting in this final model:

math.score = -2.376 + 0.533writing.score + 0.396reading.score + 12.727gendermale - 3.894reducedlunch



04

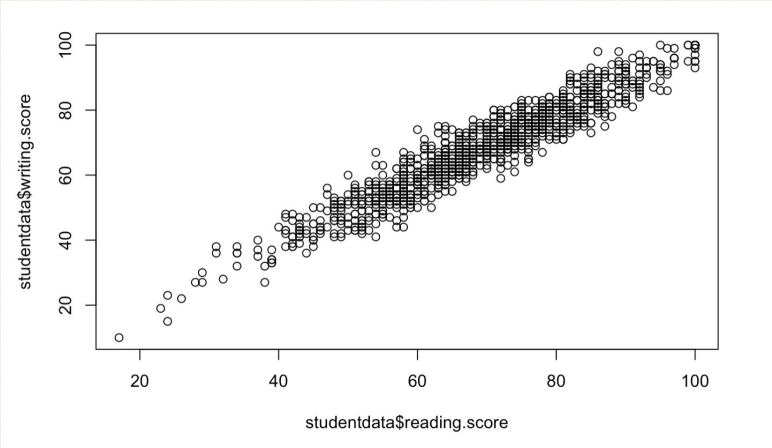
Discussion

Discussion

- From our final model, we found a student's math score:
  - Increases by about 5.33 when writing score increases by 10 points, and about 3.96 when reading score increases by 10 points
  - Is about 12.727 points higher when the student is male
  - Is about 3.894 points lower when the student is on free reduced lunch
- Our findings align with research in the field of education
  - Strong reading comprehension and vocabulary skills are necessary in math assessments
  - Students on free/reduced lunch come from lower socioeconomic backgrounds, and have less parental involvement in their educations, putting them at a disadvantage
  - Gender stereotypes surrounding math have been found to shape women's perceptions of math and negatively impact their performance

Limitations

The main limitation of our model is multicollinearity, or strong correlations among predictor variables. Reading and writing scores are highly correlated, yet after variable selection techniques we were unable to remove either of them. If we were to continue our project, we would use more advanced statistical methods to deal with the multicollinearity in our model.



Variable Inflation Factors (VIFs)

studentdata$writing.score	studentdata$reading.score
12.088952	11.548979

Thank You!