Factors influencing children's general knowledge assessment

Research Question

What are the key factors that is influencing/helpful to predict kindergarten children’s general knowledge achievement?

Preview

- Introduction
- Assumption checking
- Variable selection and model building
- Dignostics and evaluation
- Remidies
- Conclusions

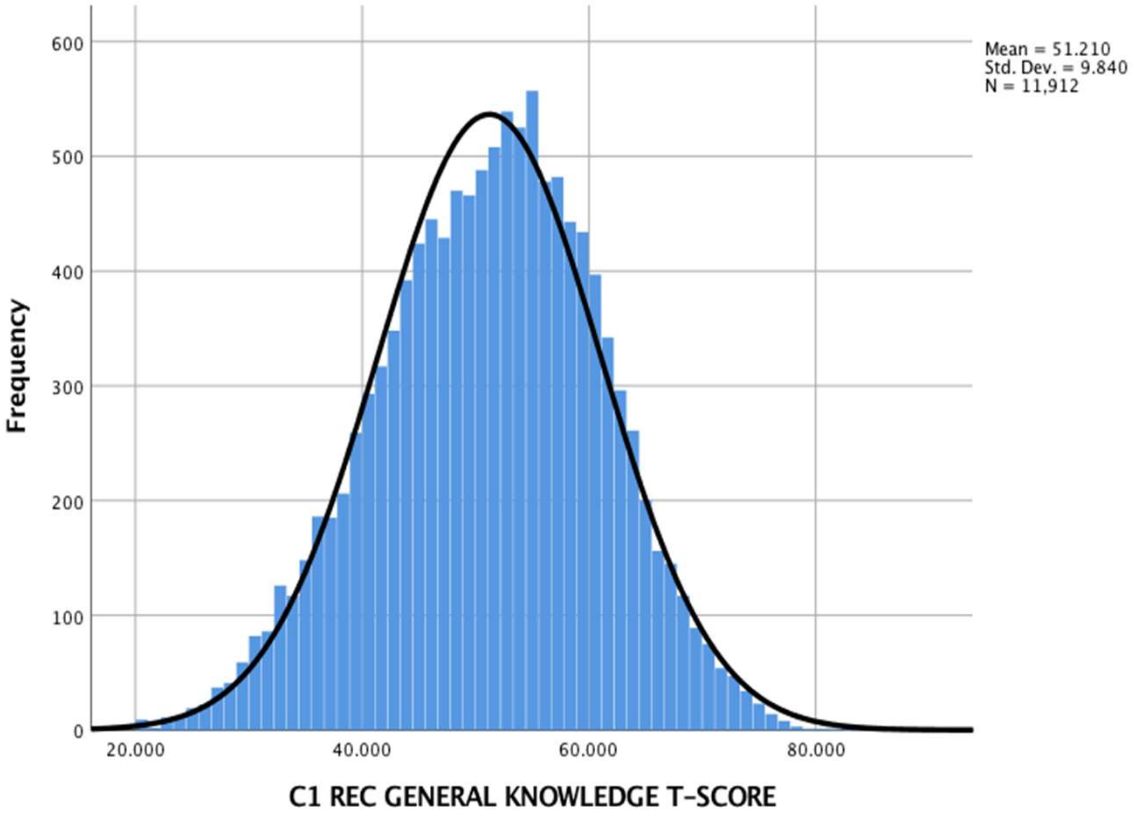
Sample & Data Source

- Data is from National Center for Education Statistics (NCES, publicly available)
- *n* = 11,912 kindergarten children from public and private schools across the United States of America with diverse background of gender, race, family social-economic and health status, etc,.

Candidate variables

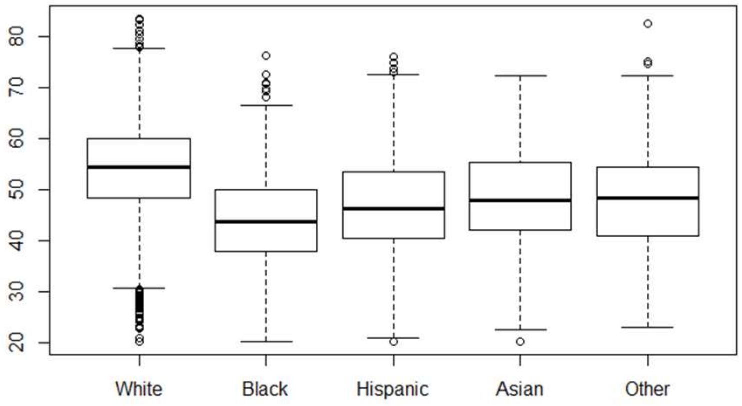
Variables	Label	Type
General Knowledge Score	T-score (Standardized, 0-100)	Continuous
Age	54 - 79 (months) -- 4.5 - 6.6-year old	
TV watched at weekdays	0 - 20 hours	
TV watched at weekends	0 - 44 hours	
Race	White, African American, Asian, Latino, Others	Nominial
Gender	Female, Male	
School	Public, Private	
Kindergarden Class	Morning, Afternoon, All-day	
Social Economic Status	Low to High (1-5)	Ordinal
Motor Skills	Low to High (1-17)	

Normality of Dependent Variable

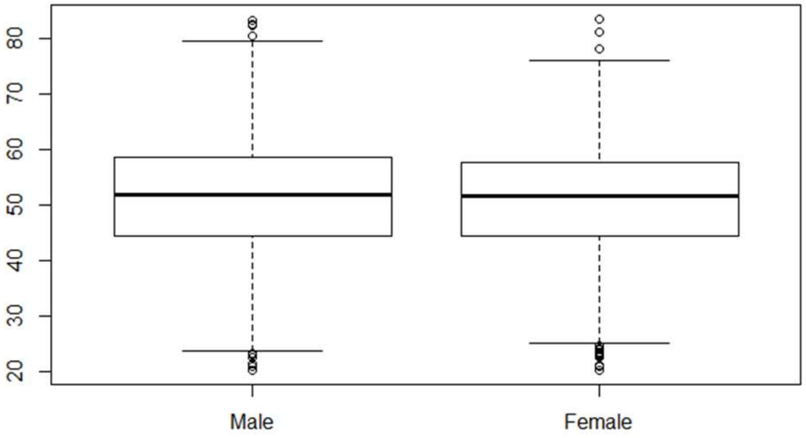


General analysis

- Boxplot of race against Y



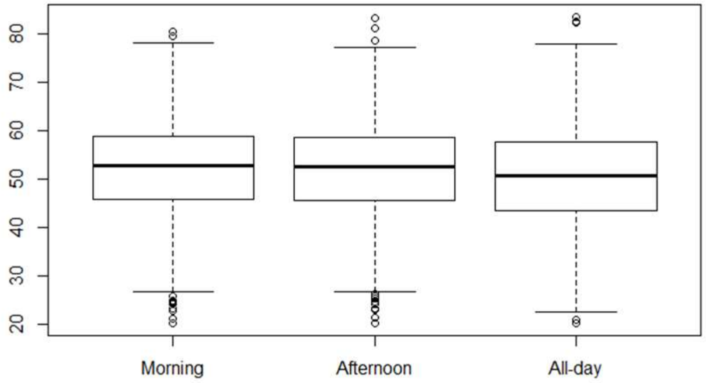
- Boxplot of gender against Y



General analysis

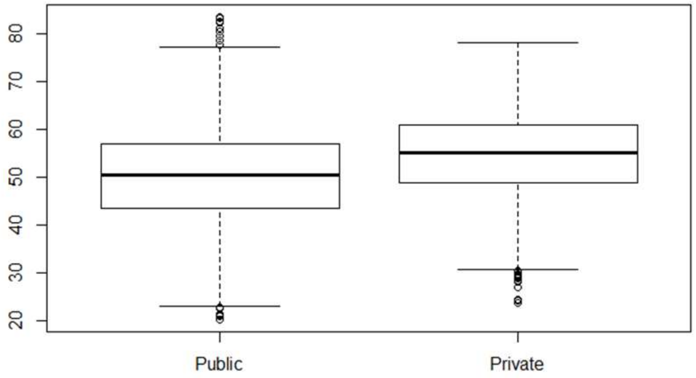
- Boxplot of kindergarten against Y

(in terms of morning, afternoon, or all day)



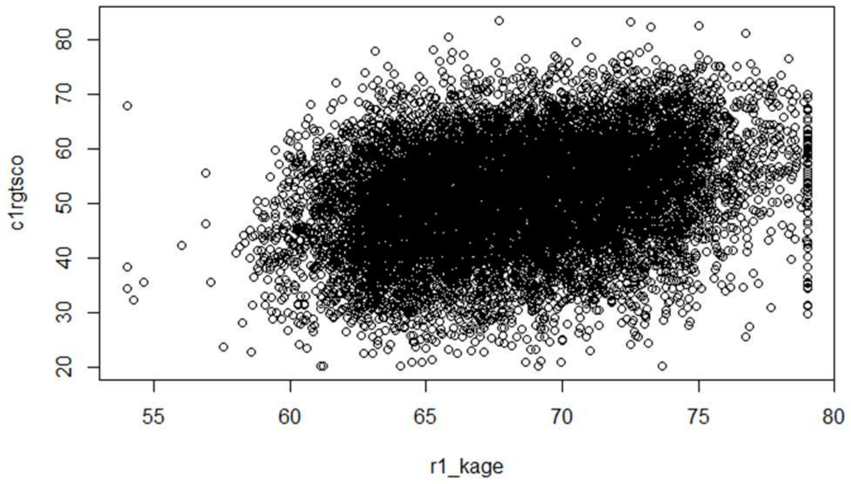
- Boxplot of kindergartden against Y

(in terms of private or public)

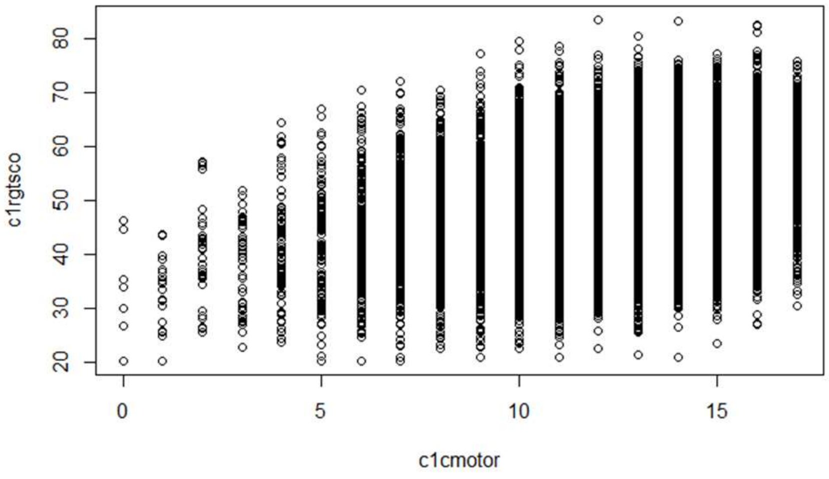


General analysis

- Plot of child's age against Y

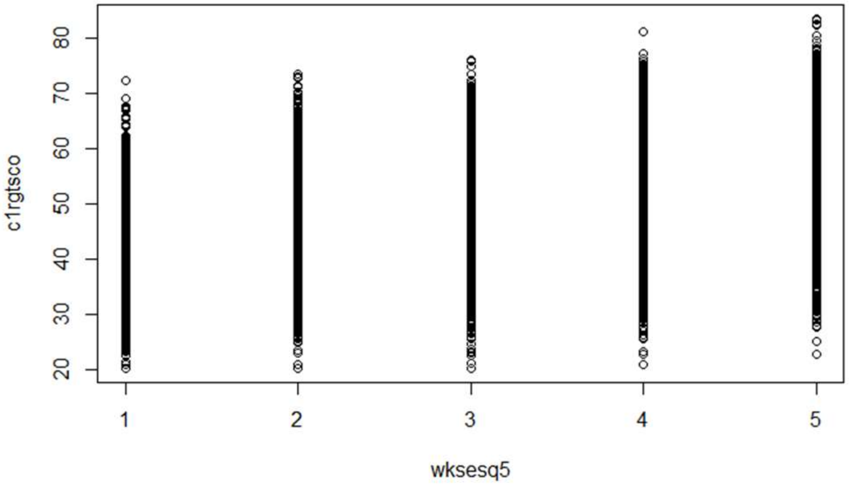


- Plot of child's assessment on composite motor skills against Y

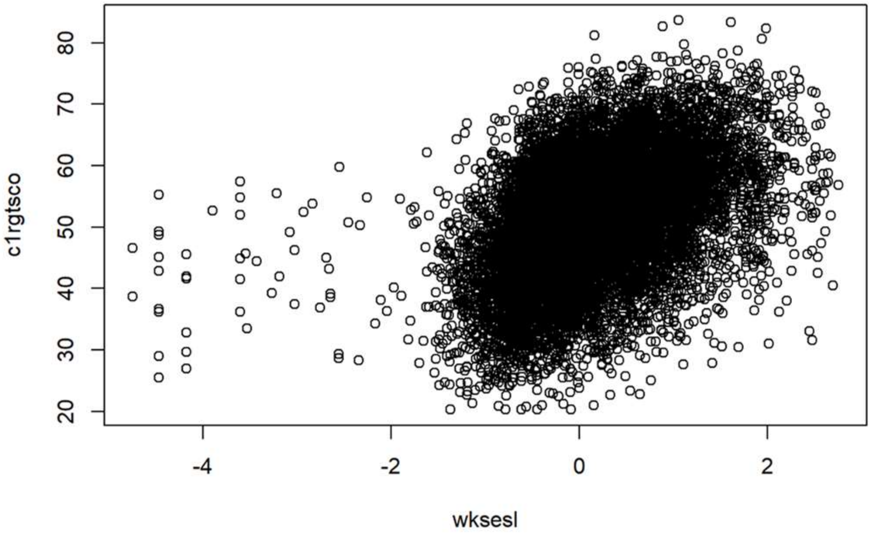


General analysis

- Plot of SES scale of the family against Y (categorical)



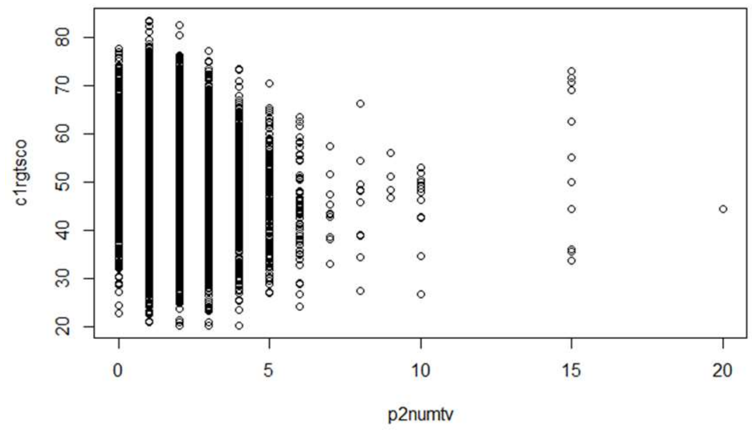
- Plot of SES scale of the family against Y (continuous)



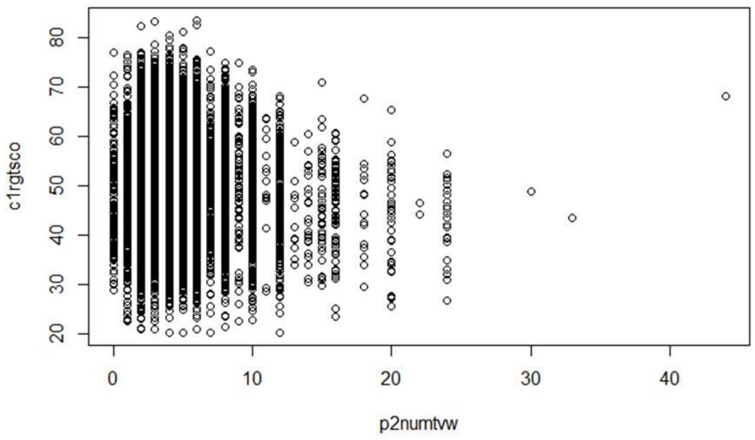
(SES) : Social Economic Status

General analysis

- Plot of number of hours TV watched on the weekdays against Y



- Plot of number of hours TV watched on the weekends against Y



Model building and variable selection

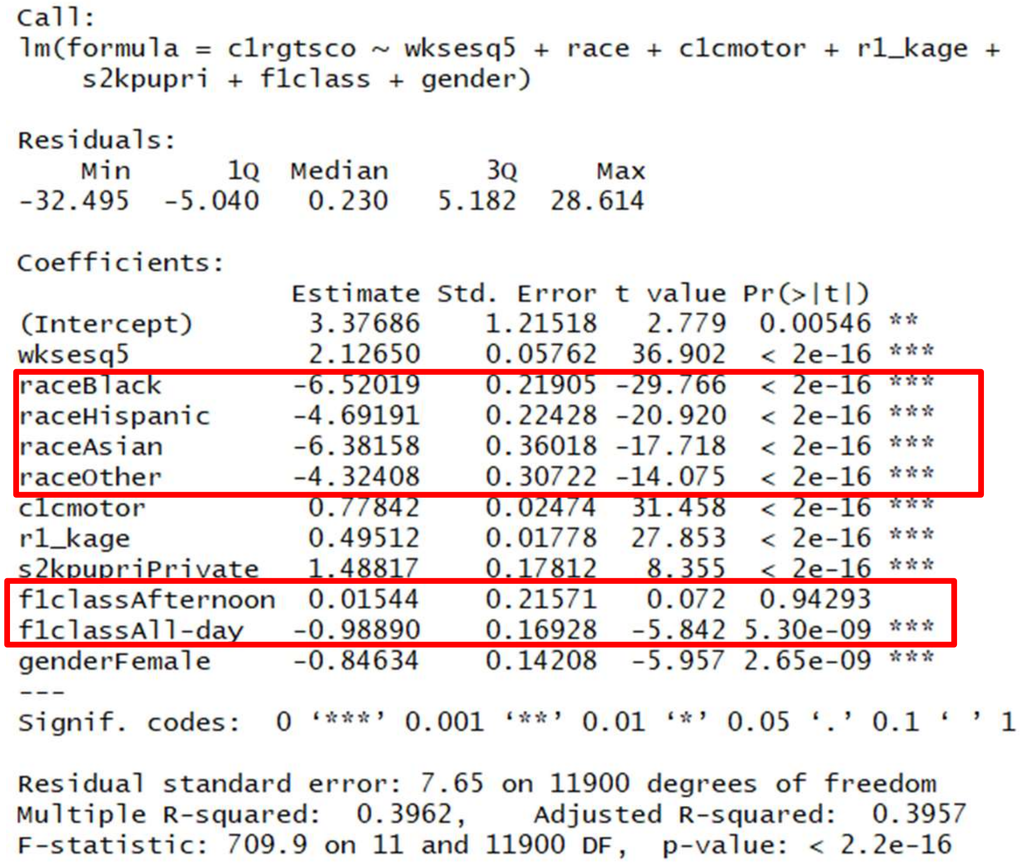
- AIC, BIC Forward/Backward
- Interprete the Model
- Adjust Significance Level

Model building and variable selection

Our Model: (Same Result for both AIC and BIC; and for all Both Direction)

General Knowledge Score ~ Social Economic Status + Race + Motor Skill + Age + Class Type + School Type + Gender

Model building and variable selection

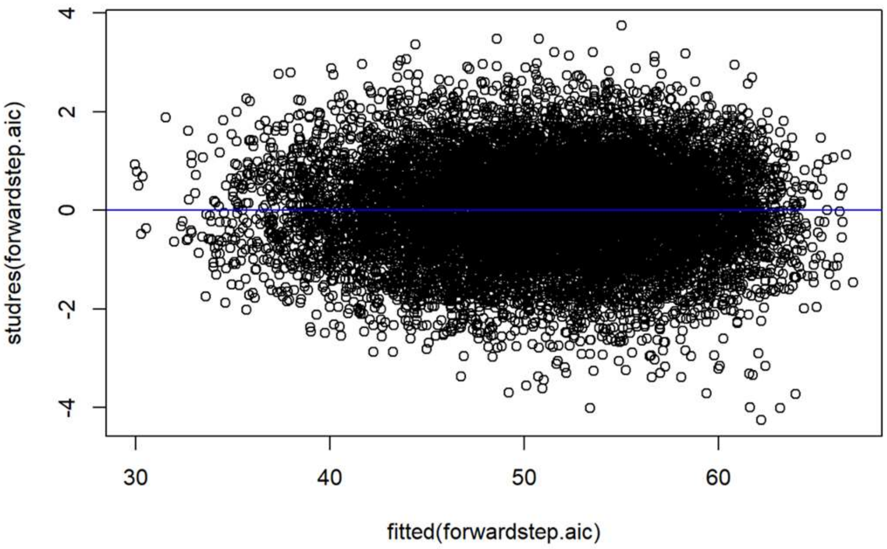


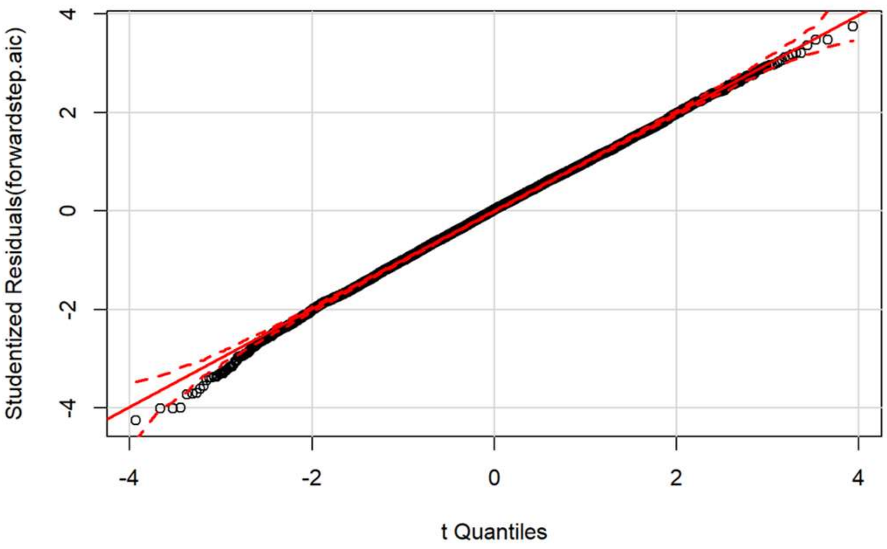
```
Call:
lm(formula = c1rgtsco ~ wksesq5 + race + c1cmotor + r1_kage +
    s2kpupri + f1class + gender)

Residuals:
    Min      1Q  Median      3Q     Max
-32.495  -5.040   0.230   5.182  28.614

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.37686    1.21518   2.779  0.00546 **
wksesq5           2.12650    0.05762  36.902  < 2e-16 ***
raceBlack        -6.52019    0.21905 -29.766  < 2e-16 ***
raceHispanic     -4.69191    0.22428 -20.920  < 2e-16 ***
raceAsian        -6.38158    0.36018 -17.718  < 2e-16 ***
raceOther        -4.32408    0.30722 -14.075  < 2e-16 ***
c1cmotor          0.77842    0.02474  31.458  < 2e-16 ***
r1_kage           0.49512    0.01778  27.853  < 2e-16 ***
s2kpupriPrivate   1.48817    0.17812   8.355  < 2e-16 ***
f1classAfternoon  0.01544    0.21571   0.072  0.94293
f1classAll-day   -0.98890    0.16928  -5.842 5.30e-09 ***
genderFemale     -0.84634    0.14208  -5.957 2.65e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.65 on 11900 degrees of freedom
Multiple R-squared:  0.3962,	Adjusted R-squared:  0.3957
F-statistic: 709.9 on 11 and 11900 DF,  p-value: < 2.2e-16
```

- All variables are significant at 0.001 level (after adjusting significance level for multiple testing)
- Race
- SES
- Motor skill score and age are positively related to RGT-score
- On average, children in private schools have higher RGT-score than those in public schools
- Time of Class
- Gender
- R^2 is less than 0.4

Diagnostics

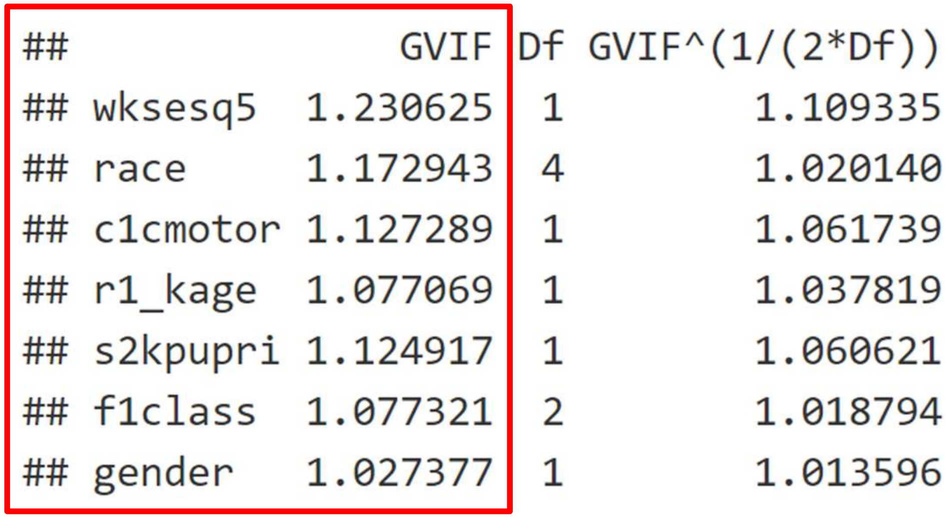
- Constant Variance
- Normality
- Independence (Assumed)





Diagnostics

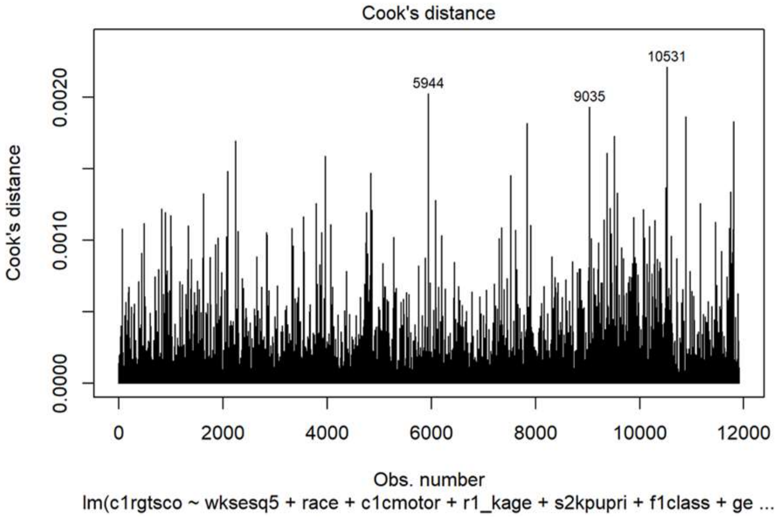
- Multicollinearity Problem
- VIF

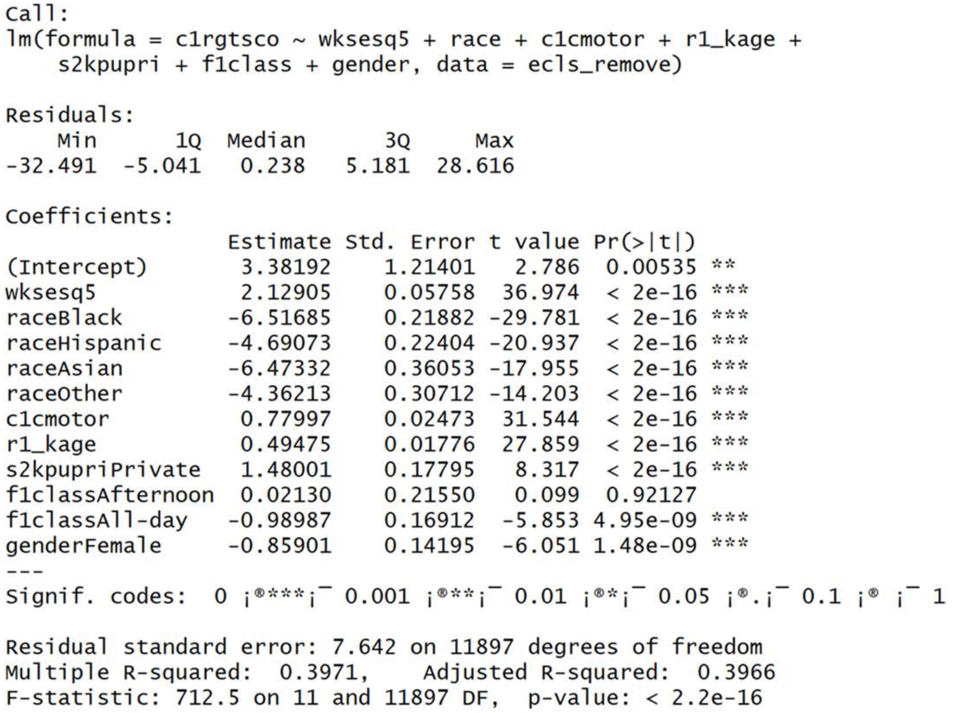


```
##              GVIF Df GVIF^(1/(2*Df))
## wksesq5  1.230625  1        1.109335
## race     1.172943  4        1.020140
## c1cmotor 1.127289  1        1.061739
## r1_kage  1.077069  1        1.037819
## s2kpupri 1.124917  1        1.060621
## f1class  1.077321  2        1.018794
## gender   1.027377  1        1.013596
```

Diagnostic

- Outliers and Influential Points
- Find 3 Influential Points
- Model After Remove Influential Points is The Same





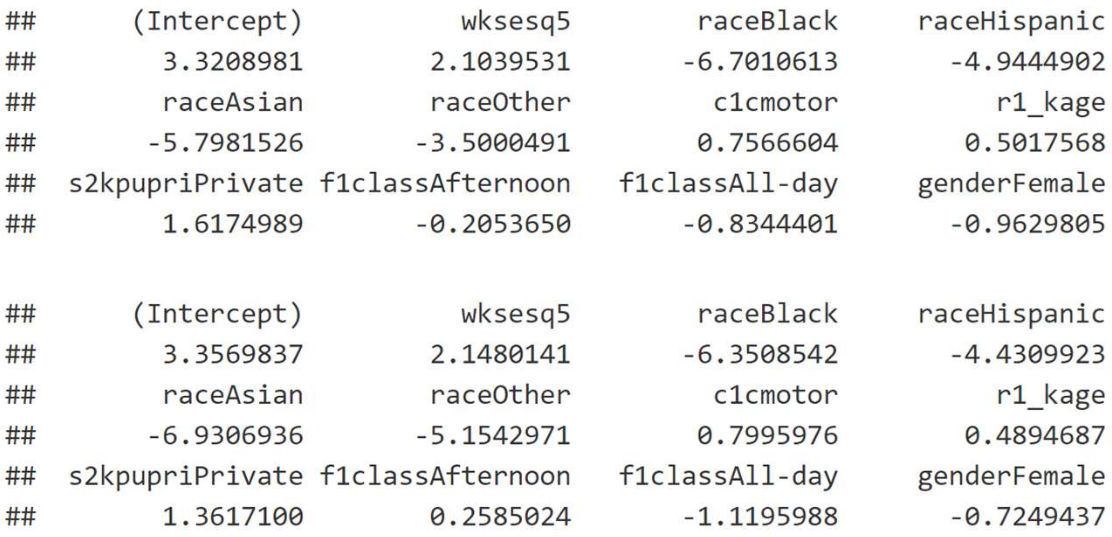
```
Call:
lm(formula = c1rgtsco ~ wksesq5 + race + c1cmotor + r1_kage +
    s2kpupri + f1class + gender, data = ecls_remove)

Residuals:
    Min      1Q  Median      3Q     Max 
-32.491  -5.041   0.238   5.181  28.616 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)       3.38192    1.21401   2.786  0.00535 ** 
wksesq5           2.12905    0.05758  36.974  < 2e-16 ***
raceBlack        -6.51685    0.21882 -29.781  < 2e-16 ***
raceHispanic     -4.69073    0.22404 -20.937  < 2e-16 ***
raceAsian        -6.47332    0.36053 -17.955  < 2e-16 ***
raceOther        -4.36213    0.30712 -14.203  < 2e-16 ***
c1cmotor          0.77997    0.02473  31.544  < 2e-16 ***
r1_kage           0.49475    0.01776  27.859  < 2e-16 ***
s2kpupriPrivate   1.48001    0.17795   8.317  < 2e-16 ***
f1classAfternoon  0.02130    0.21550   0.099  0.92127    
f1classAll-day   -0.98987    0.16912  -5.853 4.95e-09 ***
genderFemale     -0.85901    0.14195  -6.051 1.48e-09 ***
---
Signif. codes:  0 ¡®***¡¯ 0.001 ¡®**¡¯ 0.01 ¡®*¡¯ 0.05 ¡®.¡¯ 0.1 ¡® ¡¯ 1

Residual standard error: 7.642 on 11897 degrees of freedom
Multiple R-squared:  0.3971,	Adjusted R-squared:  0.3966 
F-statistic: 712.5 on 11 and 11897 DF,  p-value: < 2.2e-16
```

Cross Validation

- Double CV
- PRESS
- K-fold

Cross Validation - Double Cross Validation

- MSPR value for two sample are very close (58.01853 and59.49036)
- Coefficient Comparison: (Also Fairly Close)



```
##      (Intercept)         wksesq5       raceBlack    raceHispanic
##        3.3208981       2.1039531      -6.7010613      -4.9444902
##        raceAsian       raceOther        c1cmotor         r1_kage
##       -5.7981526      -3.5000491       0.7566604       0.5017568
##   s2kpupriPrivate f1classAfternoon  f1classAll-day    genderFemale
##        1.6174989      -0.2053650      -0.8344401      -0.9629805

##      (Intercept)         wksesq5       raceBlack    raceHispanic
##        3.3569837       2.1480141      -6.3508542      -4.4309923
##        raceAsian       raceOther        c1cmotor         r1_kage
##       -6.9306936      -5.1542971       0.7995976       0.4894687
##   s2kpupriPrivate f1classAfternoon  f1classAll-day    genderFemale
##        1.3617100       0.2585024      -1.1195988      -0.7249437
```

Cross Validation - PRESS

- PRESS = 697814.1
- SSE = 696355
- PRESS is Only Slightly Larger Than SSE

Cross Validation - 10-fold

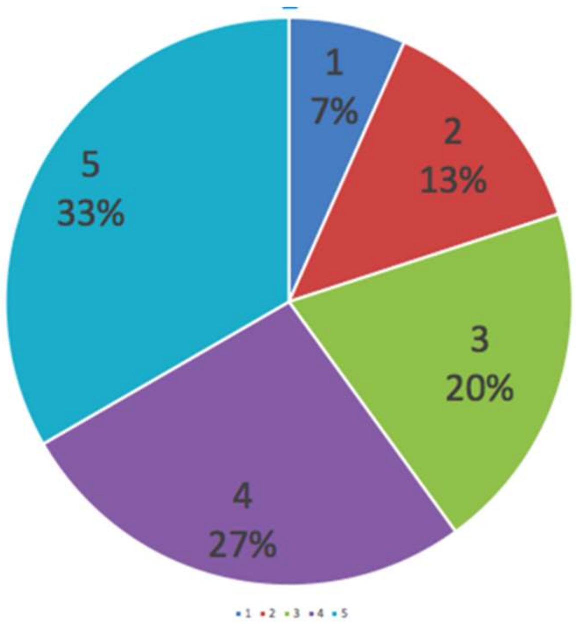
- MSE = 58.51721
- Average MSPR = 58.59942
- They are Fairly Close

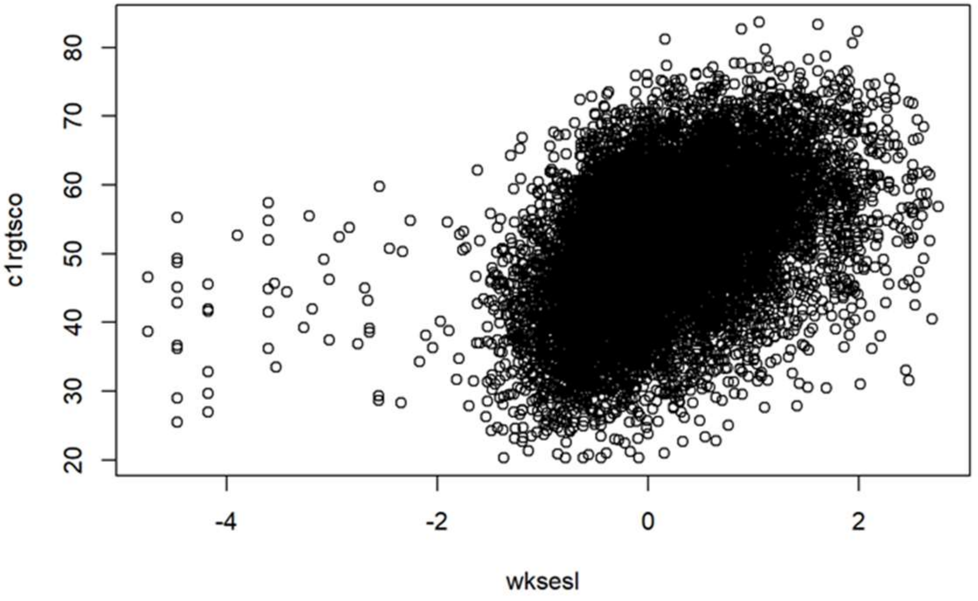
Conclusion From CV:

- Our Model Has Good Predictability
- Above Statements are Subjective Comment

Remedial Actions

- Try Other Variables: Continuous SES

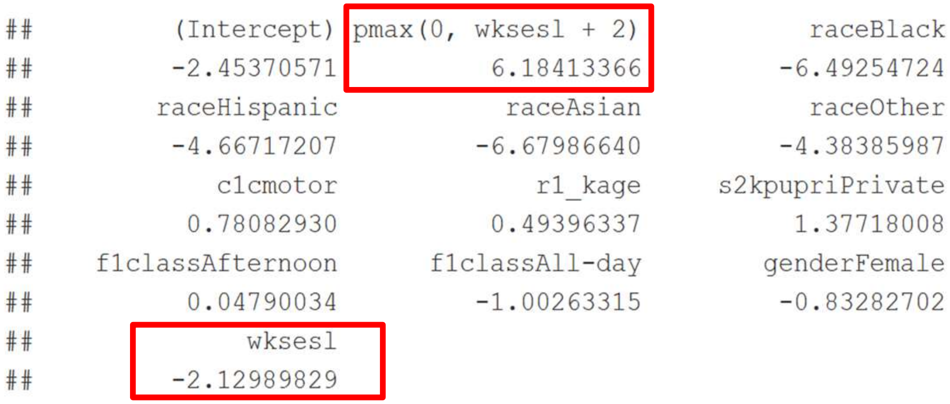




Continuous SES

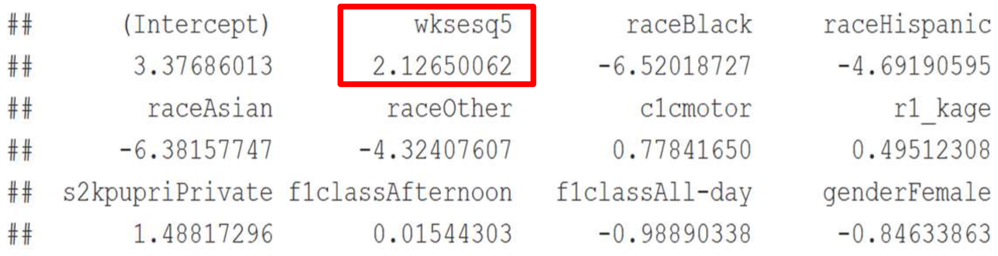
- Compare Coefficients
- Except Terms About SES, Others are Very Similar

Continuous SES:



```
##         (Intercept) pmax(0, wksesl + 2)           raceBlack
##         -2.45370571          6.18413366         -6.49254724
##        raceHispanic           raceAsian           raceOther
##         -4.66717207         -6.67986640         -4.38385987
##           c1cmotor              r1_kage    s2kpupriPrivate
##          0.78082930          0.49396337          1.37718008
##   f1classAfternoon        f1classAll-day        genderFemale
##          0.04790034         -1.00263315         -0.83282702
##              wksesl
##         -2.12989829
```

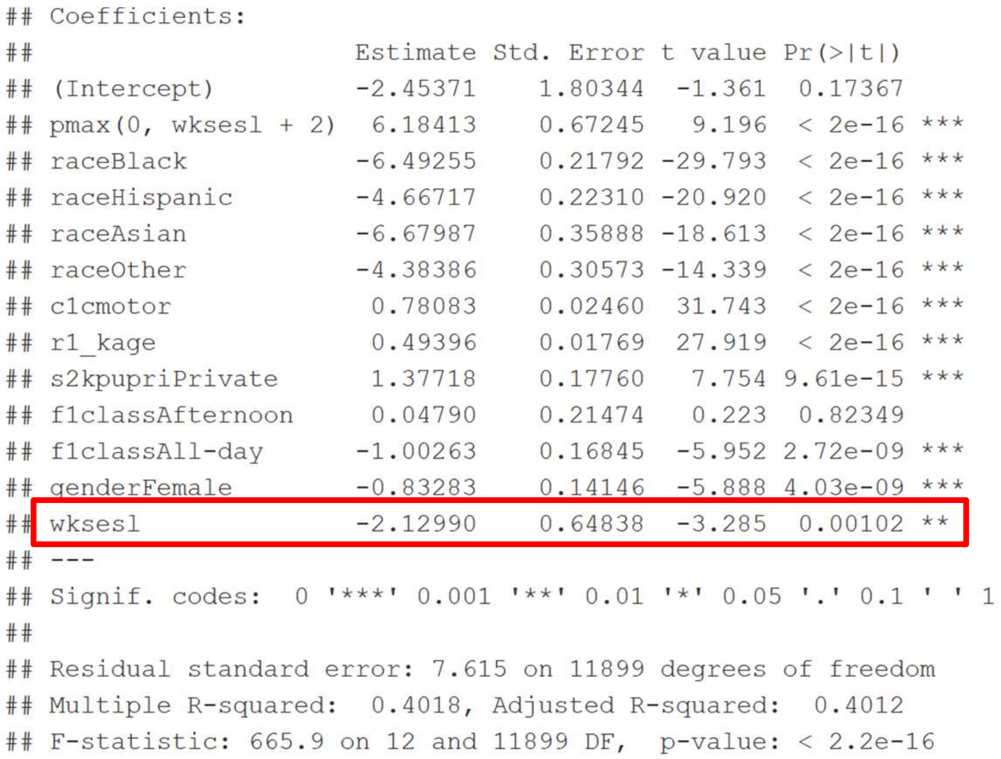
Categorical SES:



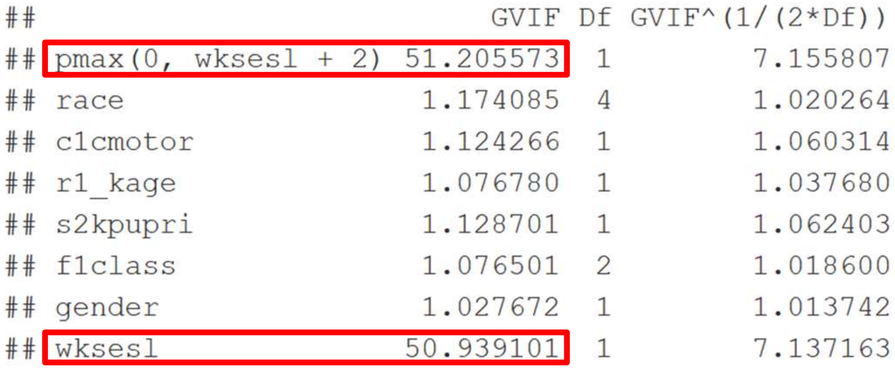
```
##       (Intercept)          wksesq5          raceBlack     raceHispanic
##        3.37686013       2.12650062        -6.52018727      -4.69190595
##         raceAsian        raceOther           c1cmotor          r1_kage
##       -6.38157747      -4.32407607         0.77841650       0.49512308
##   s2kpupriPrivate f1classAfternoon     f1classAll-day     genderFemale
##        1.48817296       0.01544303        -0.98890338      -0.84633863
```

Continuous SES

- VIF and Summary
- Should We Make SES Piecewise?



```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.45371    1.80344  -1.361  0.17367
## pmax(0, wksesl + 2)   6.18413    0.67245   9.196  < 2e-16 ***
## raceBlack            -6.49255    0.21792 -29.793  < 2e-16 ***
## raceHispanic         -4.66717    0.22310 -20.920  < 2e-16 ***
## raceAsian            -6.67987    0.35888 -18.613  < 2e-16 ***
## raceOther            -4.38386    0.30573 -14.339  < 2e-16 ***
## c1cmotor              0.78083    0.02460  31.743  < 2e-16 ***
## r1_kage               0.49396    0.01769  27.919  < 2e-16 ***
## s2kpupriPrivate       1.37718    0.17760   7.754 9.61e-15 ***
## f1classAfternoon      0.04790    0.21474   0.223  0.82349
## f1classAll-day       -1.00263    0.16845  -5.952 2.72e-09 ***
## genderFemale         -0.83283    0.14146  -5.888 4.03e-09 ***
## wksesl               -2.12990    0.64838  -3.285  0.00102 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.615 on 11899 degrees of freedom
## Multiple R-squared:  0.4018, Adjusted R-squared:  0.4012
## F-statistic: 665.9 on 12 and 11899 DF,  p-value: < 2.2e-16
```



```
##                          GVIF Df GVIF^(1/(2*Df))
## pmax(0, wksesl + 2) 51.205573  1        7.155807
## race                 1.174085  4        1.020264
## c1cmotor             1.124266  1        1.060314
## r1_kage              1.076780  1        1.037680
## s2kpupri             1.128701  1        1.062403
## f1class              1.076501  2        1.018600
## gender               1.027672  1        1.013742
## wksesl              50.939101  1        7.137163
```

Conclusion

- Important Findings

➢ Race, Gender, SES, Time of Class, Age, School, Motor Skill are Important Factors Influencing Test Score

➢ The Data Set May Should Include More Factor/Factors

➢ SES May Deserve Some Further Study

- Future Research

➢ More Factor/Factors' Data Should be Collected

➢ The Unevenly Distributed Data Points of Variables Could Cause Problems

➢ Interaction Term Investigation

➢ More Different and Advanced Analysis