*September 12, 2017*

From:Brian Fitzgerald, Ruizhi Qiao, Kiera Murphy

To: *Client*

Re: Final Report for *Sesame Street Data*

**PROJECT TITLE: Analyzing Factors and their Influence on Children’s Learning Scores**

**1.0 - PROJECT DESCRIPTION**

Penn State Statistics students were presented with data obtained by the Educational Testing Service regarding the popular television series Sesame Street and its influence on children aged 3 to 5 years old and their knowledge of several fundamental concepts. The study contained cognitive tests performed before and after the child viewed Sesame Street, as well as other descriptive variables regarding the child’s demographics.

The client provided the raw data from ETS and inquired about Sesame Street’s effectiveness in improving the students learning scores. Specifically, knowledge of letters, numbers and forms were of interest in the experiment. The client would also like to determine if the show was equally as effective in helping children across demographics, genders and other variates.

The objective here is determining which demographics are most likely to see improvement in their distinct cognitive skills from viewing Sesame Street. Within this, finding demographics that were less impacted by the show is of importance to determine how the Children’s Television Workshop, Sesame Street’s producer, can better cater to larger, more diverse audiences. The results of this study are intended to be generalized to children across America.

**1.1 - RESEARCH QUESTIONS**

Question 1: Was the Sesame Street television show effective in improving 3 to 5 year olds knowledge of letters, numbers and forms?

Question 2: If it was effective, was it equally effective in teaching across demographics, ages and genders?

Question 3: Were there any other interesting observations or correlations worth noting beyond the aforementioned?

**1.2 - STATISTICAL QUESTIONS**

Question 1: Were there significant improvements in the variables prelet, preform, or prenumb?

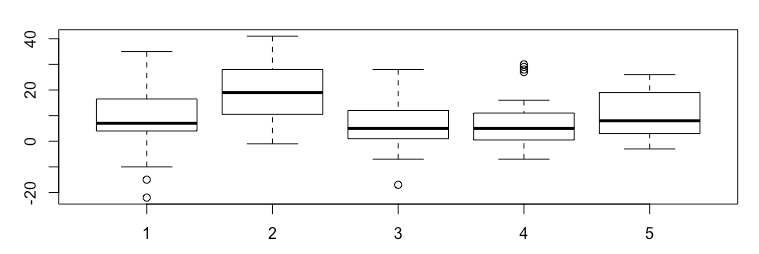
Question 2: Does demographic have a significant impact on the overall improvement?

**1.3 - VARIABLES**

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Description** | **Type** | **Levels** |
| **Prelet, Preform and Prenumb** | Score for child’s knowledge of letters numbers and forms before viewing Sesame Street | Ordinal | 58, 20 and 54 levels respectively |
| **Postlet, Postform and Postnumb** | Score for child’s knowledge of letters numbers and forms after viewing Sesame Street | Ordinal | 58, 20 and 54 levels respectively |
| **Improve\_let, Improveform and Improvenumb** | The difference between the students’ knowledge before and after viewing the program | Ordinal | Levels -22 to 41, -10 to 17 and -35 to 33, respectively |
| **Site** | Demographic the sampled child was from | Categorical | 1-5  1: Inner city disadvantaged children  2: Advantaged suburban children  3: Advantaged rural children  4: Disadvantaged rural children  5: Disadvantaged Spanish children |
| **Age** | Age in months of sampled child | Ordinal | Levels 34 – 69 months old |
| **Sex** | Gender of sampled child | Categorical | 2 levels,  1: Male  2: Female |
| **Peabody** | Mental age score for the children obtained prior to viewing Sesame Street | Ordinal | Levels 0 to 99 |
| **Viewcat** | How frequently the child viewed Sesame Street | Categorical | Levels 1 to 4 |

We are interested in the variables relating to the children's knowledge of letters, numbers and forms. These variables are prelet, postlet, prenumb, postnumb, preform and postform. To assess the improvement we created new variables called improvelet, improvenumb and improveform calculated by subtracting postx from prex. We are interested if the other variables, site, age and gender have an impact on these.

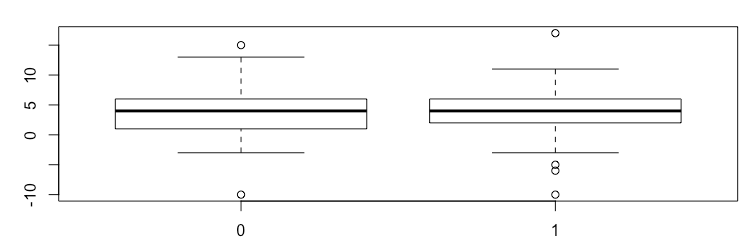
**2.0 - EXPLORATORY DATA ANALYSIS (EDA)**

***Figure 1:* Boxplot of improvement of letter score by site**

* Site 1: Disadvantaged inner city children
* Site 2: Advantaged suburban children
* Site 3: Advantaged rural children
* Site 4: Disadvantaged rural children
* Site 5: Disadvantaged Spanish speaking children

From this we can see that advantaged suburban children improved more than any other demographic. Also viewing Sesame Street was of least value to both advantaged and disadvantaged rural children.

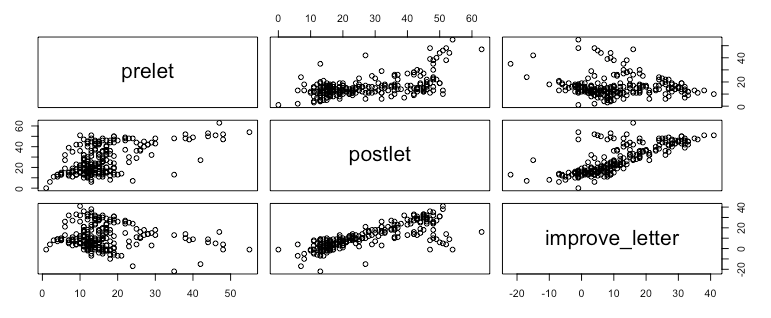
**Figure 2: Boxplot of improve form by agecat**



* Interesting improved by -10, interesting outliers that have negative improvements
* Agecat 0: 0-51 months, agecat 1: 52 months and above
* Similar average improvement

From this we can see that Sesame Street was roughly equally as effective in helping younger and older children in improving their knowledge of letters. The plots for number and form were very similar.

**Figure 3: Correlation matrix for pre, post and improvement in letter scores**



From this we can see the negative correlation between the child’s letter score before viewing Sesame Street and their improvement. This can be interpreted as Sesame Street is more effective at teaching children with a lower level of understanding.

**3.0 –STATISTICAL ANALYSIS**

To answer the client’s first question, which is if the Sesame Street television show significant to improve 3 to 5 year-old children with their knowledge on letters, numbers and forms, we use multiple linear regression model to fit each response variables. This model can figure out if other variables have significant linear relationship with each response variables.

Improveclasf, improvebody, improverelat, age, sex, viewcat(4 levels as 3 new variables), sites(5 levels as 4 new variables), setting, viewenc, peabody, agecat, encour and regular are set as explanatory variables. And improvenumb, improvelet, and improveform are set independently as 3 response variables for 3 independent multiple linear regression models.

After statistical processes( fit model, model selection), we get three model for three different response variables. From the models, we can see that all of the response variables(improvenumb, improvelet, and improveform) are significant. Thus, it is significant that Sesame Street is helpful to improve children’s ability on numbers, letters and forms.

In the first model, we can see that improvenumb is related to age, setting, improveclasf, improvebody, improverelat, regular, parts of sites(site1, site3 and site4), part of viewcat(vc2), and prenumb.

Model: improvenumb=0.101+0.221age-2.180setting+0.696improveclasf+0.153improvebody+0.527improverelat+5.825regular-3.820s1-4.980s3-4.631s4-2.441vc2-0.276renumb

This means for each one unit increase in age, improveclasf, improvebody, improverelat, and prenumb( all ordinary variables), there will be 0.221, 0.696, 0.153, 0.527, -0.276 increasing in improvenumb( the response variable). Also, for categorical variables, if the children are watching Sesame Street at home, the improvenumb will increase 2.180; if the children are watching regularly, their improve score will increase 5.825; if the children is in site1, site3 or site4, their score will decrease 3.820, 4.980 or 4.631; and if the children are vc2, their score will drop 2.441.

In the second model, we get that improvelet is related to age, improveclasf, improvebody, peabody, regular, parts of sites(site2 and site3), part of viewcat(vc2) and prelet.

Model: improvelet=-9.757 +0.260age+0.543improveclasf+0.224improvebody+0.064peabody+10.632regular+6.384s2-4.878s3-5.842vc2-0.345prelet

This means for each one unit increase in age, improveclasf, improvebody, peabody, and prelet( all ordinary variables), there will be 0.260, 0.543, 0.224, 0.064, -0.345 increasing in improvelet( the response variable). Also, for categorical variables, if the children are watching Sesame Street regularly, their improve score will increase 10.632; if the children is in site2 or site3, their score will increase 6.384 or decrease 4.878; and if the children are vc2, their score will drop 5.842.

In the third model, we can see that improveform is related to age, improveclasf, improvebody, regular, parts of sites(site2 and site3), parts of viewcat(vc2 and vc3) and preform.

Model:1.695+0.092age+0.319improveclasf+0.076improvebody+2.937regular+0.906s2-0.911s3-1.609vc2-1.065vc3-0.564preform

This means for each one unit increase in age, improveclasf, improvebody, and preform( all ordinary variables), there will be 0.092, 0.319, 0.076, -0.564 increasing in improveform( the response variable). Also, for categorical variables, if the children are watching Sesame Street regularly, their improve score will increase 2.937; if the children is in site2 or site3, their score will increase 0.906 or decrease 0.911; and if the children are vc2 or vc3, their score will drop 1.609 or 1.065.

**4.0 - RECOMMENDATIONS**

Question1: Was the Sesame Street television show effective in improving 3 to 5 year olds knowledge of letters, numbers and forms?

From boxplots and scatterplots of improvenumb, improvelet and improveform vs other predictors, we can see that although there are some potential outliers, there are significant increases on the improvement of each scores. The average improvement for knowledge of letters, numbers and forms are 9.2, 10.8 and 3.8, respectively. Thus, it is considered effective in improving 3 to 5 year old children’s knowledge on numbers, letters and forms.

Question 2: If Sesame Street was effective, was it equally effective in teaching across demographics, ages and genders?

From the optimal models of each three response variables, we can find some of the explanatory variables are all significant for all of three response variables, such as age, improveclasf, improvebody, regular, and prescores for each response variable. Some of the explanatory variables, like sites and viewcat, are all significant for every response variable; however, only parts of the sites and parts of the viewcat are significant to influence the result of improvement of the response variables’ scores. In addition, site3 and viewcat2 are important levels from sites and viewcat because they are all significant in all three models. There also exist some explanatory variables that only affect the result of one or two response variables. For example, watching in the school or at home only affect the improvement of ability on numbers. It should be considered when we talk about the response variables individually. Also, some of the variables are not significant for all of the response variables( sex). As a result, sex is not an important characteristic of measuring if Sesame Street is beneficial.

**5.0 - RESOURCES**

R markdown was used to analyze the data.

**6.0 - CONSIDERATIONS**

Despite our findings in this analysis, it is important to note that we do not know how these children were selected. For example, for demographic, we are told that ‘children representative of these populations were sampled from 5 different sites in the United States’. We do not know how the children were selected. They may have been randomly selected, or may have been biased selections.

We are not provided with a time frame when this study was collected, which is holding us back from drawing stronger conclusions. There is the potential for other outside, lurking variables that may be taking place during this time which may have affected our data. Given the time frame, we would be able to draw stronger conclusions from our data.

We observed several outliers in the improvement categories. Despite noticing these, we left them in the model because those with negative improvements were typically children with high initial scores and not a lot of room for improvement. There also was a positive correlation between letters, numbers and forms improvement. This shows that Sesame Street was equally as effective in teaching all three topics.

**7.0 - ACKNOWLEDGEMENT OF WORK**

*We would like to thank the Educational Testing Service for the data. Additionally we would like to thank the client!*

**APPENDIX A - Recoded Variates**

SS=read.xlsx('sesame.xlsx', sheetIndex = 1)

SS$improvenumb=SS$postnumb-SS$prenumb

SS$improvelet=SS$postlet-SS$prelet

SS$improveform=SS$postform-SS$preform

SS$improvebody=SS$postbody-SS$prebody

SS$improverelat=SS$postrelat-SS$prerelat

SS$improveclasf=SS$postclasf-SS$preclasf

site1=SS[SS$site==1,]

site2=SS[SS$site==2,]

site3=SS[SS$site==3,]

site4=SS[SS$site==4,]

site5=SS[SS$site==5,]

sesame=SS

sesame$s1[sesame$site==1]=1

sesame$s1[sesame$site>1]=0

sesame$s2[sesame$site==2]=1

sesame$s2[sesame$site>2]=0

sesame$s2[sesame$site<2]=0

sesame$s3[sesame$site==3]=1

sesame$s3[sesame$site>3]=0

sesame$s3[sesame$site<3]=0

sesame$s4[sesame$site==4]=1

sesame$s4[sesame$site>4]=0

sesame$s4[sesame$site<4]=0

sesame$sex[sesame$sex==2]=0

sesame$setting[sesame$setting==2]=0

sesame$viewenc[sesame$viewenc==2]=0

sesame$vc1[sesame$viewcat==1]=1

sesame$vc1[sesame$viewcat>1]=0

sesame$vc2[sesame$viewcat==2]=1

sesame$vc2[sesame$viewcat>2]=0

sesame$vc2[sesame$viewcat<2]=0

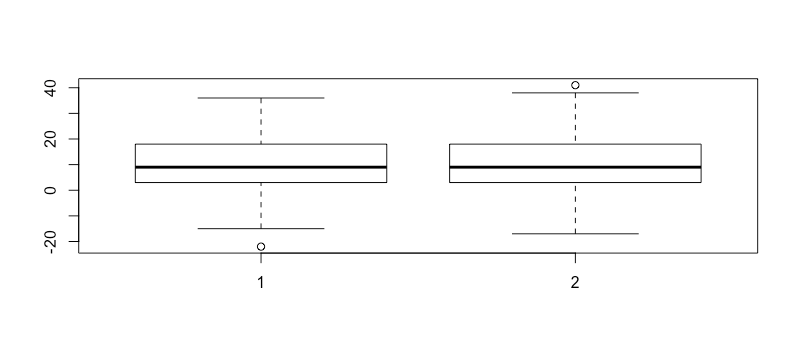
sesame$vc3[sesame$viewcat==3]=1

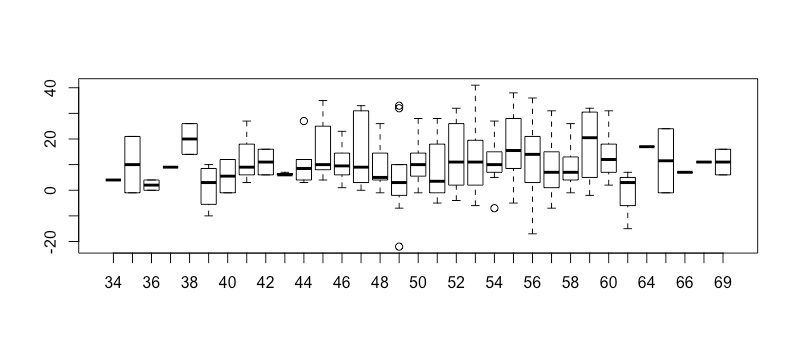
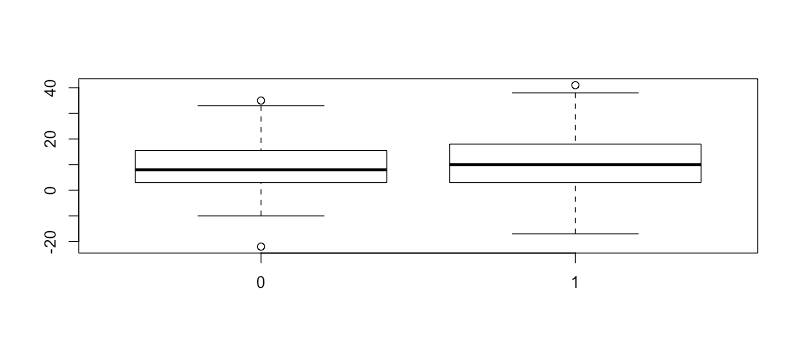
sesame$vc3[sesame$viewcat>3]=0

sesame$vc3[sesame$viewcat<3]=0

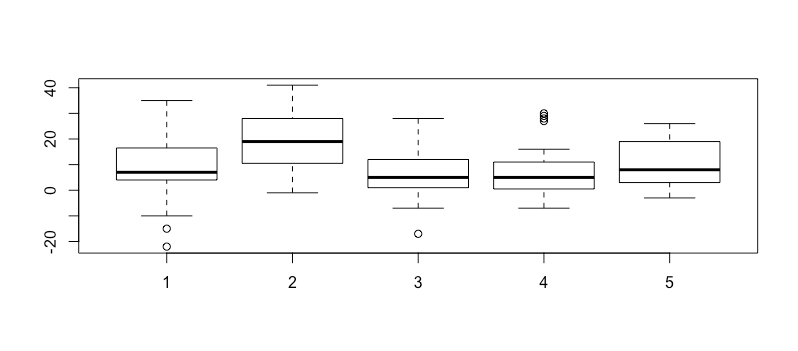
**APPENDIX B - Boxplots & Correlations**

pairs(~age+improveclasf+improvebody+improverelat+peabody+prenumb+prelet+preform+improvenumb+improvelet+improveform)

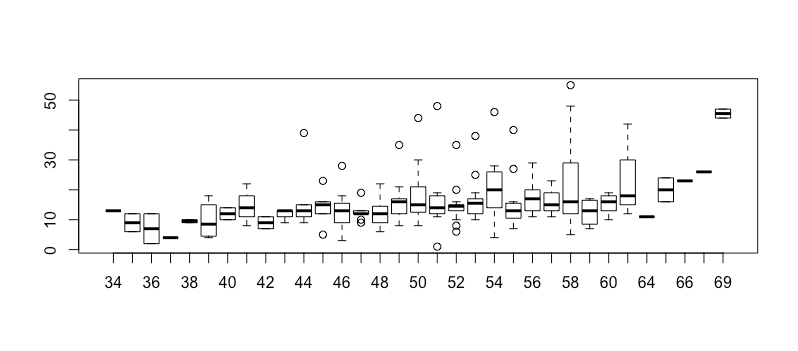
boxplot(improve\_letter ~ sex

boxplot(improve\_letter ~ age)boxplot(improve\_letter ~ agecat)

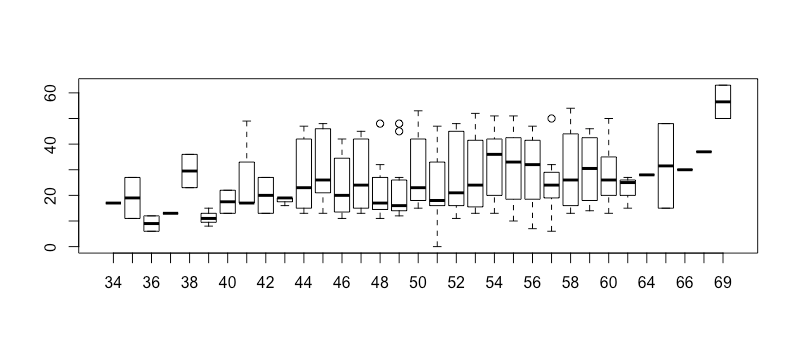
boxplot(improve\_letter ~ site)



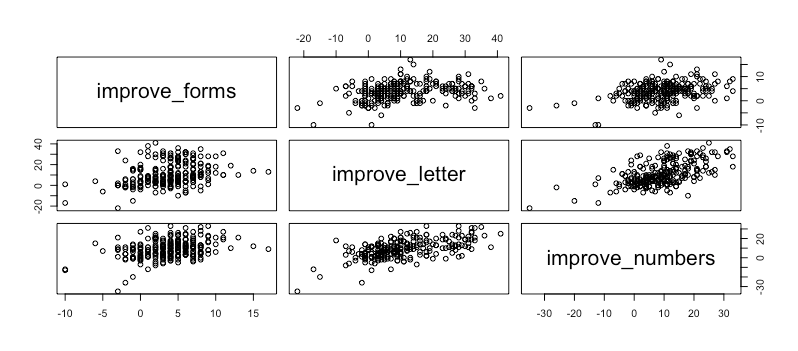
boxplot(prelet ~ age)



boxplot(postlet ~ age)



pairs(~improve\_forms + improve\_letter + improve\_numbers)



**APPENDIX C - Model #1 (Number score improvement)**

fit1=lm(improvenumb~sex+age++setting+viewenc+improveclasf+improvebody+improverelat+peabody+agecat+encour+regular+s1+s2+s3+s4+vc1+vc2+vc3+prenumb)

summary(fit1)

MSE1=(summary(fit1)$sigma)^2

step(fit1, scale=MSE1, direction="backward")

none=lm(improvenumb~1)

step(none, scope=list(upper=fit1), scale=MSE1)

## Call:

## lm(formula = improvenumb ~ improveclasf + improverelat + regular +

## prenumb + s2 + age + vc2 + improvebody + setting)

##

## Coefficients:

## (Intercept) improveclasf improverelat regular prenumb

## -4.2240 0.7350 0.4572 5.4536 -0.2574

## s2 age vc2 improvebody setting

## 4.2616 0.2145 -2.1799 0.1763 -1.7083

fitnumb=lm(improvenumb ~ age + setting + improveclasf + improvebody +

improverelat + regular + s1 + s3 + s4 + vc2 + prenumb)

summary(fitnumb)

##

## Call:

## lm(formula = improvenumb ~ age + setting + improveclasf + improvebody +

## improverelat + regular + s1 + s3 + s4 + vc2 + prenumb)

##

## Residuals:

## Min 1Q Median 3Q Max

## -29.6134 -5.2796 0.4747 5.1154 19.6492

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 0.10084 4.74760 0.021 0.983073

## age 0.22069 0.09215 2.395 0.017434 \*

## setting -2.18031 1.07794 -2.023 0.044275 \*

## improveclasf 0.69644 0.12418 5.608 5.89e-08 \*\*\*

## improvebody 0.15334 0.10674 1.437 0.152215

## improverelat 0.52718 0.16396 3.215 0.001492 \*\*

## regular 5.82453 1.40346 4.150 4.69e-05 \*\*\*

## s1 -3.81957 1.34987 -2.830 0.005077 \*\*

## s3 -4.98011 1.40685 -3.540 0.000485 \*\*\*

## s4 -4.63062 1.62785 -2.845 0.004851 \*\*

## vc2 -2.44077 1.21910 -2.002 0.046457 \*

## prenumb -0.27622 0.05911 -4.673 5.07e-06 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 7.518 on 228 degrees of freedom

## Multiple R-squared: 0.4263, Adjusted R-squared: 0.3986

## F-statistic: 15.4 on 11 and 228 DF, p-value: < 2.2e-16

**APPENDIX D - Model #2 (Letter improvement Score)**

fit2=lm(improvelet~sex+age++setting+viewenc+improveclasf+improvebody+improverelat+peabody+agecat+encour+regular+s1+s2+s3+s4+vc1+vc2+vc3+prelet)

summary(fit2)

MSE2=(summary(fit2)$sigma)^2

step(fit2, scale=MSE2, direction="backward")

none=lm(improvelet~1)

step(none, scope=list(upper=fit2), scale=MSE2)

## Call:

## lm(formula = improvelet ~ age + improveclasf + improvebody +

## peabody + regular + s2 + s3 + vc2 + prelet)

##

## Coefficients:

## (Intercept) age improveclasf improvebody peabody

## -9.75733 0.25973 0.54348 0.22414 0.06387

## regular s2 s3 vc2 prelet

## 10.63175 6.38403 -4.87814 -5.84246 -0.34504

fitlet=lm(improvelet ~ age + improveclasf + improvebody +

peabody + regular + s2 + s3 + vc2 + prelet)

summary(fitlet)

##

## Call:

## lm(formula = improvelet ~ age + improveclasf + improvebody +

## peabody + regular + s2 + s3 + vc2 + prelet)

##

## Residuals:

## Min 1Q Median 3Q Max

## -26.1732 -4.9991 -0.4583 4.6262 20.8643

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) -9.75733 4.64297 -2.102 0.036683 \*

## age 0.25973 0.09226 2.815 0.005298 \*\*

## improveclasf 0.54348 0.12420 4.376 1.83e-05 \*\*\*

## improvebody 0.22414 0.11202 2.001 0.046567 \*

## peabody 0.06387 0.04340 1.472 0.142471

## regular 10.63175 1.42342 7.469 1.67e-12 \*\*\*

## s2 6.38403 1.55046 4.118 5.34e-05 \*\*\*

## s3 -4.87814 1.31538 -3.709 0.000261 \*\*\*

## vc2 -5.84246 1.30867 -4.464 1.26e-05 \*\*\*

## prelet -0.34504 0.07049 -4.895 1.85e-06 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 8.133 on 230 degrees of freedom

## Multiple R-squared: 0.4896, Adjusted R-squared: 0.4696

## F-statistic: 24.52 on 9 and 230 DF, p-value: < 2.2e-16

**APPENDIX E - Model #3 (Form score improvement)**

fit3=lm(improveform~sex+age++setting+viewenc+improveclasf+improvebody+improverelat+peabody+agecat+encour+regular+s1+s2+s3+s4+vc1+vc2+vc3+preform)

summary(fit3)

MSE3=(summary(fit3)$sigma)^2

step(fit3, scale=MSE3, direction="backward")

none=lm(improveform~1)

step(none, scope=list(upper=fit3), scale=MSE3)

## Call:

## lm(formula = improveform ~ improveclasf + preform + regular +

## vc2 + vc3 + age + s3 + improvebody + s2)

##

## Coefficients:

## (Intercept) improveclasf preform regular vc2

## 1.69450 0.31890 -0.56394 2.93696 -1.60872

## vc3 age s3 improvebody s2

## -1.06510 0.09217 -0.91110 0.07620 0.90603

fitform=lm(improveform ~ age + improveclasf + improvebody +

regular + s2 + s3 + vc2 + vc3 + preform)

summary(fitform)

##

## Call:

## lm(formula = improveform ~ age + improveclasf + improvebody +

## regular + s2 + s3 + vc2 + vc3 + preform)

##

## Residuals:

## Min 1Q Median 3Q Max

## -7.4959 -1.4593 0.4011 1.7237 7.3122

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 1.69450 1.48989 1.137 0.25658

## age 0.09217 0.02891 3.188 0.00163 \*\*

## improveclasf 0.31890 0.04074 7.828 1.81e-13 \*\*\*

## improvebody 0.07620 0.03637 2.095 0.03726 \*

## regular 2.93696 0.56839 5.167 5.15e-07 \*\*\*

## s2 0.90603 0.45738 1.981 0.04879 \*

## s3 -0.91110 0.43640 -2.088 0.03792 \*

## vc2 -1.60872 0.50323 -3.197 0.00158 \*\*

## vc3 -1.06510 0.47599 -2.238 0.02620 \*

## preform -0.56394 0.05909 -9.544 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 2.627 on 230 degrees of freedom

## Multiple R-squared: 0.5267, Adjusted R-squared: 0.5082

## F-statistic: 28.44 on 9 and 230 DF, p-value: < 2.2e-16