

Proposal EDA

Thomas Mande

2022-12-02

```
#install.packages('haven')
library(haven)
sesame <- read_dta("sesame.dta")
library(tidyverse)

## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr 0.3.4
## v tibble 3.1.8       v dplyr 1.0.10
## v tidyr 1.2.1        v stringr 1.4.1
## v readr 2.1.2        v forcats 0.5.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
```

```
library(knitr)
library(broom)
```

```
head(sesame)
```

```
## # A tibble: 6 x 28
##   rowna-1 id site sex age viewcat setting viewenc prebody prelet preform
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     1     1     1     1   66     1     2     1     16     23     12
## 2     2     2     1     2   67     3     2     1     30     26     9
## 3     3     3     1     1   56     3     2     2     22     14     9
## 4     4     4     1     1   49     1     2     2     23     11    10
## 5     5     5     1     1   69     4     2     2     32     47    15
## 6     6     6     1     2   54     3     2     2     29     26    10
## # ... with 17 more variables: prenumb <dbl>, prerelat <dbl>, preclasf <dbl>,
## #   postbody <dbl>, postlet <dbl>, postform <dbl>, postnumb <dbl>,
## #   postrelat <dbl>, postclasf <dbl>, peabody <dbl>, agecat <dbl>,
## #   encour <dbl>, '_Isite_2' <dbl>, '_Isite_3' <dbl>, '_Isite_4' <dbl>,
## #   '_Isite_5' <dbl>, regular <dbl>, and abbreviated variable name 1: rownames
```

Data Cleaning + Super Basic Stats

```
#sesame 1 i'm using for gams
sesame <- sesame %>%
```

```

mutate(viewcat = as.factor(viewcat)) %>%
mutate(site = as.factor(site)) %>%
mutate(sex = as.factor(sex)) %>%
mutate(setting = as.factor(setting)) %>%
mutate(viewenc = as.factor(viewenc)) %>%
mutate(regular = as.factor(regular))
sesame1<- sesame
sesame1 <- sesame1 %>%
  mutate(difflet = postlet - prelet) %>%
  mutate(diffnumb = postnumb - prenumb)

```

```

levels(sesame$site) <- c("Disadv City", "Adv Sub", "Adv Rural", "Disadv Rural", "Disadv Spanish")

```

```

sesame <- sesame %>%
  mutate(diffbody = postbody - prebody) %>%
  mutate(difflet = postlet - prelet) %>%
  mutate(diffform = postform - preform) %>%
  mutate(diffnumb = postnumb - prenumb) %>%
  mutate(diffrelat = postrelat - prerelat) %>%
  mutate(diffclasf = postclasf - preclasf)

```

```

sesame %>%
  group_by(site) %>%
  count()

```

```

## # A tibble: 5 x 2
## # Groups:   site [5]
##   site          n
##   <fct>      <int>
## 1 Disadv City    60
## 2 Adv Sub       55
## 3 Adv Rural     64
## 4 Disadv Rural  43
## 5 Disadv Spanish 18

```

```

sesame %>%
  group_by(site) %>%
  count(encour)

```

```

## # A tibble: 10 x 3
## # Groups:   site [5]
##   site          encour    n
##   <fct>      <dbl> <int>
## 1 Disadv City         0   28
## 2 Disadv City         1   32
## 3 Adv Sub             0   19
## 4 Adv Sub             1   36
## 5 Adv Rural           0   14
## 6 Adv Rural           1   50
## 7 Disadv Rural        0   23
## 8 Disadv Rural        1   20
## 9 Disadv Spanish      0    4
## 10 Disadv Spanish     1   14

```

```
sesame %>%
  group_by(viewcat) %>%
  count(encour)
```

```
## # A tibble: 8 x 3
## # Groups:   viewcat [4]
##   viewcat encour     n
##   <fct>     <dbl> <int>
## 1 1         0     40
## 2 1         1     14
## 3 2         0     13
## 4 2         1     47
## 5 3         0     17
## 6 3         1     47
## 7 4         0     18
## 8 4         1     44
```

Question 1: Does watching sesame street impact learning?

#Created models to look for effects of being in different viewing categories on learning across categories

```
lm_body <- lm(diffbody ~ viewcat, data = sesame)
lm_let <- lm(difflet ~ viewcat, data = sesame)
lm_form <- lm(diffform ~ viewcat, data = sesame)
lm_numb <- lm(diffnumb ~ viewcat, data = sesame)
lm_relaf <- lm(diffrelaf ~ viewcat, data = sesame)
lm_clasf <- lm(diffclasf ~ viewcat, data = sesame)
summary(lm_body)
```

```
##
## Call:
## lm(formula = diffbody ~ viewcat, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.7833  -3.2460  -0.3253   3.2915  19.0938
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1667     0.6918   4.578 7.61e-06 ***
## viewcat2      0.6167     0.9535   0.647   0.518
## viewcat3      0.7396     0.9393   0.787   0.432
## viewcat4      1.3172     0.9462   1.392   0.165
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.083 on 236 degrees of freedom
## Multiple R-squared:  0.008225, Adjusted R-squared:  -0.004382
## F-statistic: 0.6524 on 3 and 236 DF, p-value: 0.5822
```

```
summary(lm_let)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -37.000  -6.383  -0.741   5.519  27.650
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.481      1.339   1.853  0.06507 .
## viewcat2       5.869      1.846   3.180  0.00167 **
## viewcat3      12.519      1.818   6.886 5.15e-11 ***
## viewcat4      13.615      1.831   7.435 1.93e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.839 on 236 degrees of freedom
## Multiple R-squared:  0.2337, Adjusted R-squared:  0.2239
## F-statistic: 23.99 on 3 and 236 DF, p-value: 1.386e-13
```

```
summary(lm_form)
```

```
##
## Call:
## lm(formula = diffform ~ viewcat, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.6333  -2.6333   0.1935   2.2222  13.1094
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.7778     0.5037   5.515 9.14e-08 ***
## viewcat2       0.8556     0.6943   1.232  0.21907
## viewcat3       1.1128     0.6839   1.627  0.10504
## viewcat4       2.0287     0.6890   2.945  0.00356 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.701 on 236 degrees of freedom
## Multiple R-squared:  0.03618, Adjusted R-squared:  0.02393
## F-statistic: 2.953 on 3 and 236 DF, p-value: 0.03331
```

```
summary(lm_numb)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat, data = sesame)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.906  -5.671   0.000   6.407  24.550
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.593      1.272   3.611 0.000372 ***
## viewcat2       3.857      1.753   2.200 0.028750 *
## viewcat3       6.314      1.727   3.656 0.000316 ***
## viewcat4       7.407      1.740   4.258 2.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.346 on 236 degrees of freedom
## Multiple R-squared:  0.08244,    Adjusted R-squared:  0.07077
## F-statistic: 7.068 on 3 and 236 DF,  p-value: 0.0001439
```

```
summary(lm_relata)
```

```
##
## Call:
## lm(formula = diffrelat ~ viewcat, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.167  -2.167   0.375   1.833  11.833
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.1667     0.4708   2.478  0.0139 *
## viewcat2       0.3667     0.6489   0.565  0.5726
## viewcat3       0.4583     0.6393   0.717  0.4741
## viewcat4       1.3978     0.6440   2.171  0.0310 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.46 on 236 degrees of freedom
## Multiple R-squared:  0.02192,    Adjusted R-squared:  0.009488
## F-statistic: 1.763 on 3 and 236 DF,  p-value: 0.1549
```

```
summary(lm_clasf)
```

```
##
## Call:
## lm(formula = diffclasf ~ viewcat, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.8594  -3.0926   0.1406   2.6833  10.1406
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    2.0926    0.6014    3.480 0.000598 ***
## viewcat2       1.2241    0.8289    1.477 0.141083
## viewcat3       1.7668    0.8166    2.164 0.031490 *
## viewcat4       2.4558    0.8226    2.986 0.003129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.419 on 236 degrees of freedom
## Multiple R-squared:  0.03867,    Adjusted R-squared:  0.02645
## F-statistic: 3.164 on 3 and 236 DF,  p-value: 0.02524
```

Mean in difference in each of these scores by view category. can skip for analysis

```
sesame %>%
  group_by(viewcat) %>%
  summarise(mean = mean(diffbody))
```

```
## # A tibble: 4 x 2
##   viewcat mean
##   <fct>   <dbl>
## 1 1      3.17
## 2 2      3.78
## 3 3      3.91
## 4 4      4.48
```

```
sesame %>%
  group_by(viewcat) %>%
  summarise(mean = mean(difflet))
```

```
## # A tibble: 4 x 2
##   viewcat mean
##   <fct>   <dbl>
## 1 1      2.48
## 2 2      8.35
## 3 3      15
## 4 4     16.1
```

```
sesame %>%
  group_by(viewcat) %>%
  summarise(mean = mean(diffform))
```

```
## # A tibble: 4 x 2
##   viewcat mean
##   <fct>   <dbl>
## 1 1      2.78
## 2 2      3.63
## 3 3      3.89
## 4 4      4.81
```

```
sesame %>%
  group_by(viewcat) %>%
  summarise(mean = mean(diffnumb))
```

```
## # A tibble: 4 x 2
##   viewcat mean
##   <fct>   <dbl>
## 1 1      4.59
## 2 2      8.45
## 3 3     10.9
## 4 4     12
```

```
sesame %>%
  group_by(viewcat) %>%
  summarise(mean = mean(diffrelat))
```

```
## # A tibble: 4 x 2
##   viewcat mean
##   <fct>   <dbl>
## 1 1      1.17
## 2 2      1.53
## 3 3      1.62
## 4 4      2.56
```

```
sesame %>%
  group_by(viewcat) %>%
  summarise(mean = mean(diffclasf))
```

```
## # A tibble: 4 x 2
##   viewcat mean
##   <fct>   <dbl>
## 1 1      2.09
## 2 2      3.32
## 3 3      3.86
## 4 4      4.55
```

#see what happens using regular instead of viewcat. still strongly significant for difflet and diffnumb

```
lm_let_reg <- lm(difflet ~ regular, data = sesame)
lm_numb_reg <- lm(diffnumb ~ regular, data = sesame)
lm_body_reg <- lm(diffbody ~ regular, data = sesame)
lm_form_reg <- lm(diffform ~ regular, data = sesame)
lm_relat_reg <- lm(diffrelat ~ regular, data = sesame)
lm_clasf_reg <- lm(diffclasf ~ regular, data = sesame)
summary(lm_body_reg)
```

```
##
## Call:
## lm(formula = diffbody ~ regular, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.0591  -3.0860  -0.0591   3.1640  18.9409
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.1667     0.6898   4.590 7.16e-06 ***
```

```
## regular1      0.8925      0.7836      1.139      0.256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.069 on 238 degrees of freedom
## Multiple R-squared:  0.005421, Adjusted R-squared:  0.001242
## F-statistic: 1.297 on 1 and 238 DF, p-value: 0.2559
```

```
summary(lm_let_reg)
```

```
##
## Call:
## lm(formula = difflet ~ regular, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -35.220  -7.220  -1.481   5.584  27.780
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.481      1.394   1.780  0.0764 .
## regular1      10.739      1.584   6.781 9.37e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.25 on 238 degrees of freedom
## Multiple R-squared:  0.1619, Adjusted R-squared:  0.1584
## F-statistic: 45.98 on 1 and 238 DF, p-value: 9.366e-11
```

```
summary(lm_form_reg)
```

```
##
## Call:
## lm(formula = diffform ~ regular, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1129  -2.1129  -0.1129   2.2222  12.8871
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.7778      0.5052   5.499 9.85e-08 ***
## regular1       1.3351      0.5738   2.327  0.0208 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.712 on 238 degrees of freedom
## Multiple R-squared:  0.02224, Adjusted R-squared:  0.01813
## F-statistic: 5.413 on 1 and 238 DF, p-value: 0.02083
```

```
summary(lm_numb_reg)
```



```
##
## Call:
## lm(formula = diffnumb ~ regular, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.478  -5.814   0.407   6.407  22.522
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.593      1.279   3.591 0.000399 ***
## regular1       5.886      1.453   4.052 6.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.397 on 238 degrees of freedom
## Multiple R-squared:  0.06454,    Adjusted R-squared:  0.06061
## F-statistic: 16.42 on 1 and 238 DF,  p-value: 6.876e-05
```

```
summary(lm_relav_reg)
```

```
##
## Call:
## lm(formula = diffrelat ~ regular, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1667  -1.9086   0.0914   2.0914  11.8333
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.1667      0.4721   2.471  0.0142 *
## regular1       0.7419      0.5363   1.383  0.1678
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.47 on 238 degrees of freedom
## Multiple R-squared:  0.007977,    Adjusted R-squared:  0.003809
## F-statistic: 1.914 on 1 and 238 DF,  p-value: 0.1678
```

```
summary(lm_clasf_reg)
```

```
##
## Call:
## lm(formula = diffclasf ~ regular, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.914  -2.914   0.086   3.086  10.086
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.0926      0.6018   3.477 0.000603 ***
```

```
## regular1      1.8214      0.6836      2.664 0.008244 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.423 on 238 degrees of freedom
## Multiple R-squared:  0.02896, Adjusted R-squared:  0.02488
## F-statistic: 7.098 on 1 and 238 DF, p-value: 0.008244
```

#Decided to focus in on variables with two strongest effects, and see whether those effects still held.

#Is there anything else we have to do to show that sesame street generally was associated with increase

```
lm_let <- lm(difflet ~ viewcat + sex + age + setting + + prelet + site, data = sesame)
lm_numb <- lm(diffnumb ~ viewcat + sex + age + setting + + prenumb + site, data = sesame)

summary(lm_let)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + sex + age + setting + +prelet +
##      site, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.2064  -5.6284  -0.2192   5.3259  21.4166
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -5.9873     4.9980  -1.198  0.232186
## viewcat2         5.8084     1.6945   3.428  0.000722 ***
## viewcat3        12.1372     1.7418   6.968  3.42e-11 ***
## viewcat4        12.5065     1.7748   7.047  2.15e-11 ***
## sex2             1.0614     1.1319   0.938  0.349386
## age              0.2717     0.1002   2.712  0.007206 **
## setting2         0.1724     1.2900   0.134  0.893826
## prelet          -0.3788     0.0732  -5.175  5.00e-07 ***
## siteAdv Sub       7.4776     1.6485   4.536  9.27e-06 ***
## siteAdv Rural    -5.2865     1.6468  -3.210  0.001518 **
## siteDisadv Rural -0.7470     1.8490  -0.404  0.686577
## siteDisadv Spanish 1.5424     2.4938   0.619  0.536852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.658 on 228 degrees of freedom
## Multiple R-squared:  0.4267, Adjusted R-squared:  0.3991
## F-statistic: 15.43 on 11 and 228 DF, p-value: < 2.2e-16
```

```
summary(lm_numb)
```

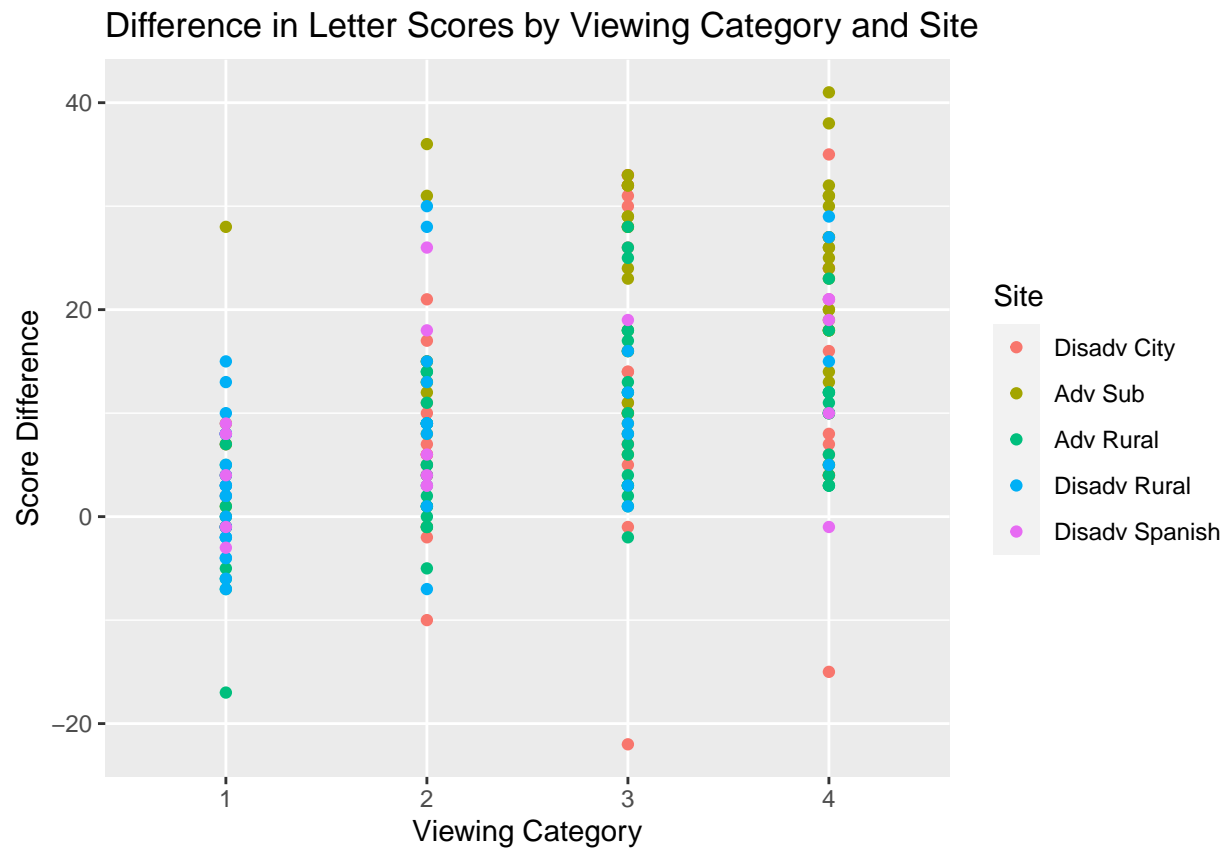
```
##
## Call:
## lm(formula = diffnumb ~ viewcat + sex + age + setting + +prenumb +
##      site, data = sesame)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -33.222  -5.345   0.101   5.788  21.250
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.64267     5.13107  -0.320  0.74915
## viewcat2         4.87825     1.69565   2.877  0.00440 **
## viewcat3         8.29293     1.75548   4.724 4.04e-06 ***
## viewcat4         9.22375     1.79041   5.152 5.58e-07 ***
## sex2             0.69555     1.13043   0.615  0.53897
## age              0.21322     0.10719   1.989  0.04788 *
## setting2         1.85877     1.28264   1.449  0.14866
## prenumb        -0.40162     0.06473  -6.205 2.55e-09 ***
## siteAdv Sub       5.00316     1.65403   3.025  0.00277 **
## siteAdv Rural    -0.84254     1.66932  -0.505  0.61424
## siteDisadv Rural  0.12581     1.84263   0.068  0.94563
## siteDisadv Spanish 4.19485     2.49180   1.683  0.09365 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.644 on 228 degrees of freedom
## Multiple R-squared:  0.2417, Adjusted R-squared:  0.2051
## F-statistic: 6.605 on 11 and 228 DF,  p-value: 1.47e-09
```

Question 2: How did the benefits of watching sesame street vary across demographic groups?

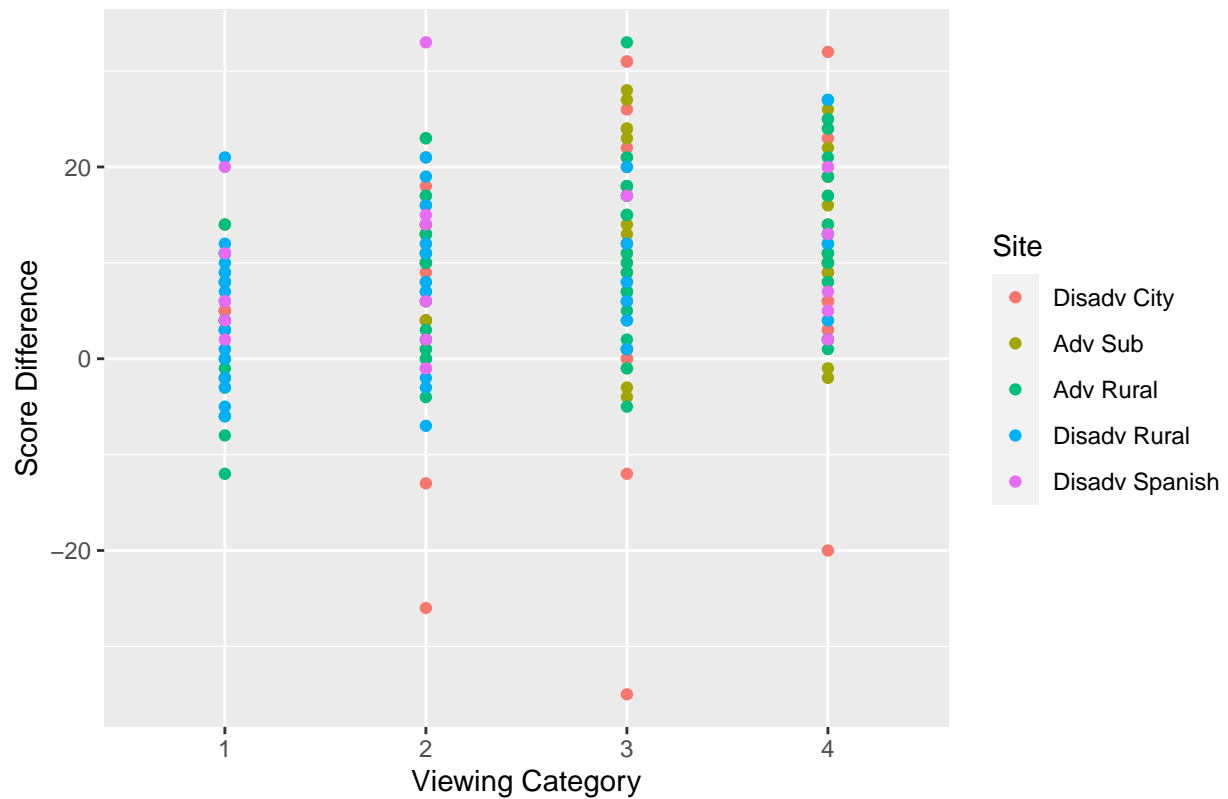
#Continuing focus on difflet and diffnumb, these graphs show how learning varied depending on how much

```
ggplot(data = sesame, mapping = aes(x = viewcat, y = difflet, color = site)) +
  geom_point() +
  labs(title = "Difference in Letter Scores by Viewing Category and Site",
       x = "Viewing Category", y = "Score Difference", color = "Site")
```



```
ggplot(data = sesame, mapping = aes(x = viewcat, y = diffnumb, color = site)) +
  geom_point() +
  labs(title = "Difference in Number Scores by Viewing Category and Site",
        x = "Viewing Category", y = "Score Difference", color = "Site")
```

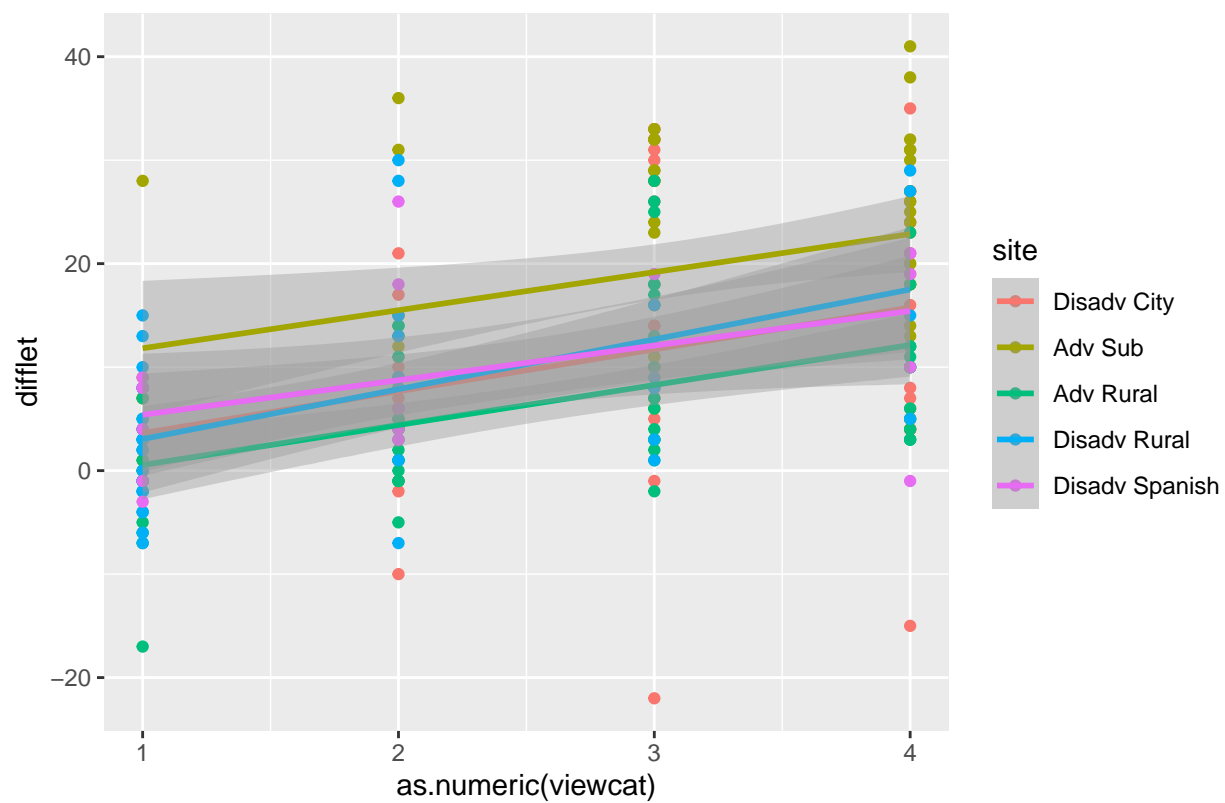
Difference in Number Scores by Viewing Category and Site



#Just made these graphs and I really like them. They confirm the general point that there isn't really

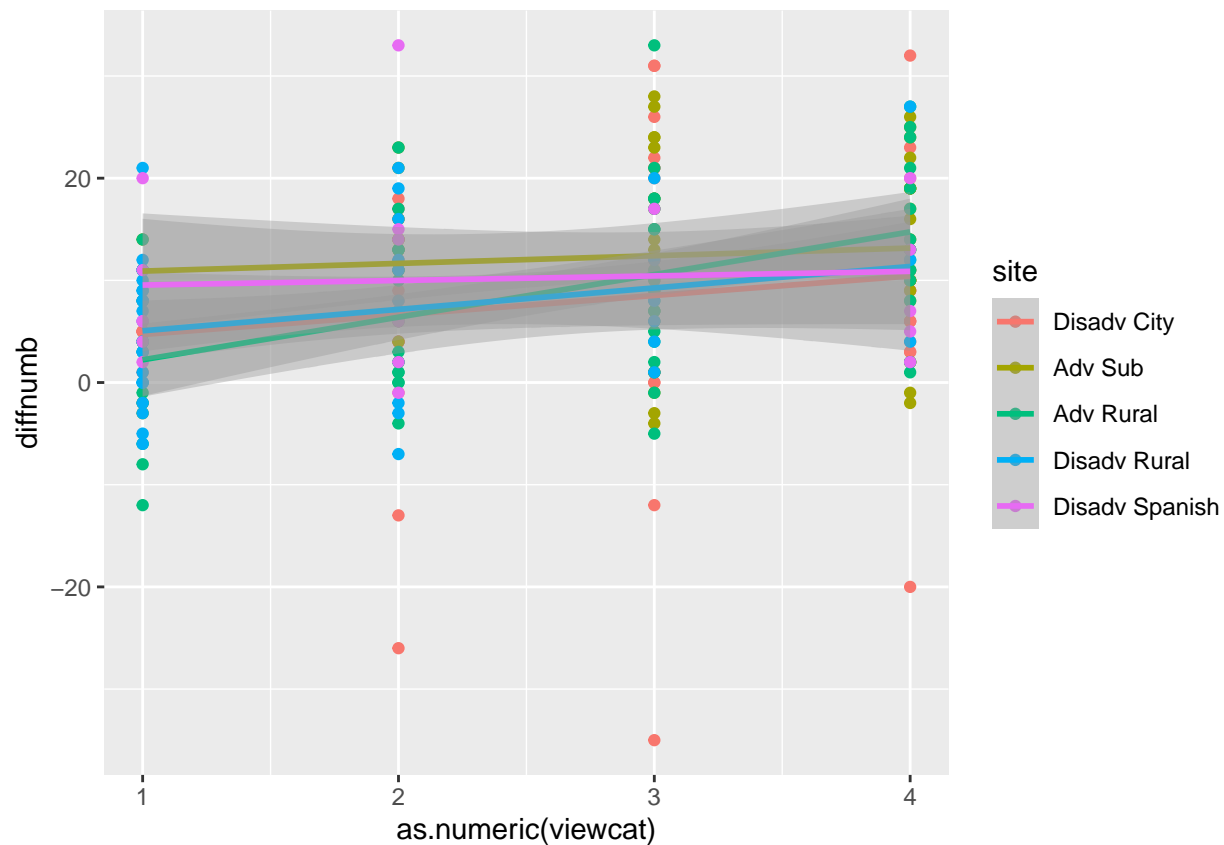
```
library(ggplot2)
qplot(x = as.numeric(viewcat), y = difflet, data = sesame, color = site) +
  geom_smooth(method = "lm") + labs(title = "")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



```
qplot(x = as.numeric(viewcat), y = diffnumb, data = sesame, color = site) +
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



#related summary stats for all categories

```
sesame %>%
  group_by(site) %>%
  summarise(mean = mean(diffbody))
```

```
## # A tibble: 5 x 2
##   site      mean
##   <fct>    <dbl>
## 1 Disadv City    2.92
## 2 Adv Sub        2.82
## 3 Adv Rural      4.56
## 4 Disadv Rural   4.72
## 5 Disadv Spanish 5.61
```

```
sesame %>%
  group_by(site) %>%
  summarise(mean = mean(difflet))
```

```
## # A tibble: 5 x 2
##   site      mean
##   <fct>    <dbl>
## 1 Disadv City   10.3
## 2 Adv Sub       19.6
## 3 Adv Rural      6.70
## 4 Disadv Rural   6.86
## 5 Disadv Spanish 9.67
```

```
sesame %>%
  group_by(site) %>%
  summarise(mean = mean(diffform))
```

```
## # A tibble: 5 x 2
##   site      mean
##   <fct>    <dbl>
## 1 Disadv City    3.27
## 2 Adv Sub       4.31
## 3 Adv Rural     4.16
## 4 Disadv Rural  2.98
## 5 Disadv Spanish 4.89
```

```
sesame %>%
  group_by(site) %>%
  summarise(mean = mean(diffnumb))
```

```
## # A tibble: 5 x 2
##   site      mean
##   <fct>    <dbl>
## 1 Disadv City    7.85
## 2 Adv Sub      12.5
## 3 Adv Rural     8.88
## 4 Disadv Rural  6.72
## 5 Disadv Spanish 10.1
```

```
sesame %>%
  group_by(site) %>%
  summarise(mean = mean(diffrelat))
```

```
## # A tibble: 5 x 2
##   site      mean
##   <fct>    <dbl>
## 1 Disadv City    1.15
## 2 Adv Sub       1.44
## 3 Adv Rural     2.44
## 4 Disadv Rural  2.42
## 5 Disadv Spanish 0.556
```

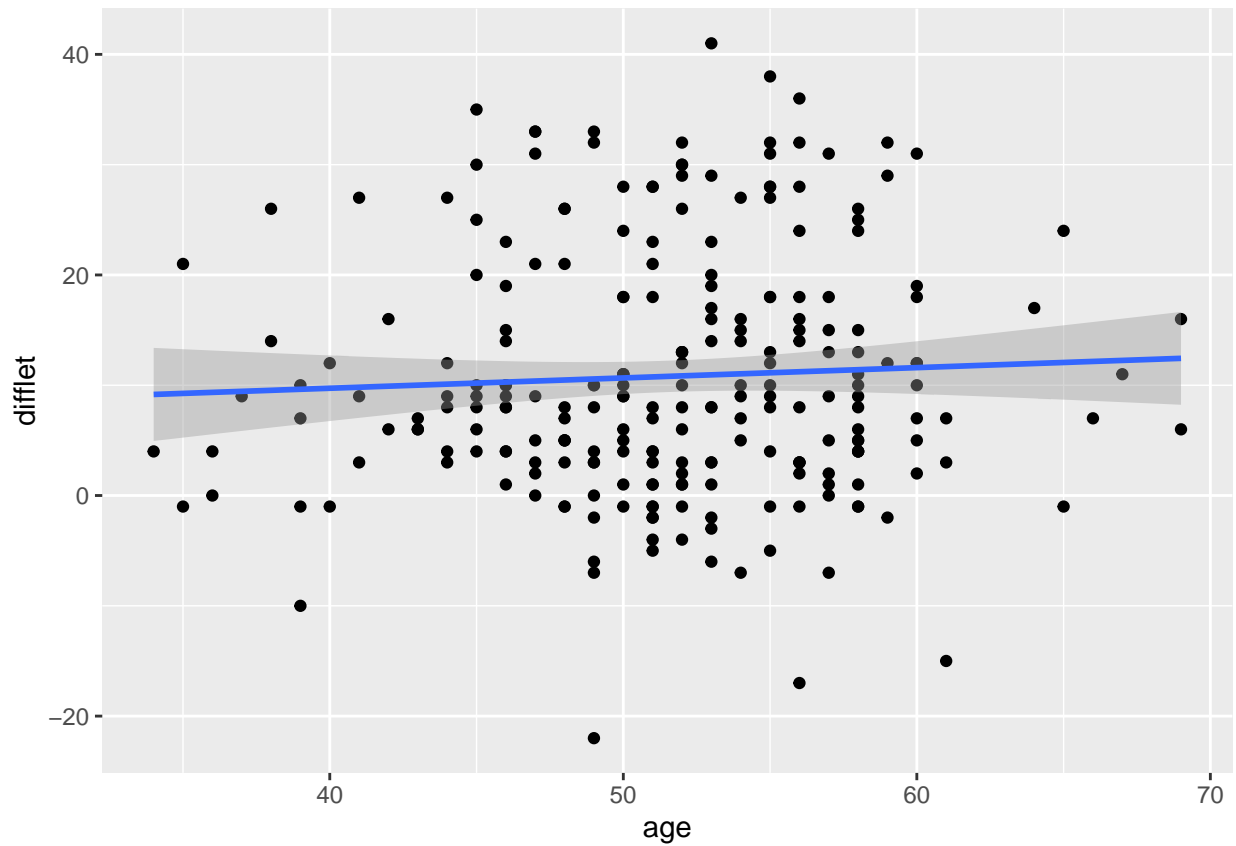
```
sesame %>%
  group_by(site) %>%
  summarise(mean = mean(diffclasf))
```

```
## # A tibble: 5 x 2
##   site      mean
##   <fct>    <dbl>
## 1 Disadv City    3.22
## 2 Adv Sub       4.53
## 3 Adv Rural     3.39
## 4 Disadv Rural  2.44
## 5 Disadv Spanish 4.28
```



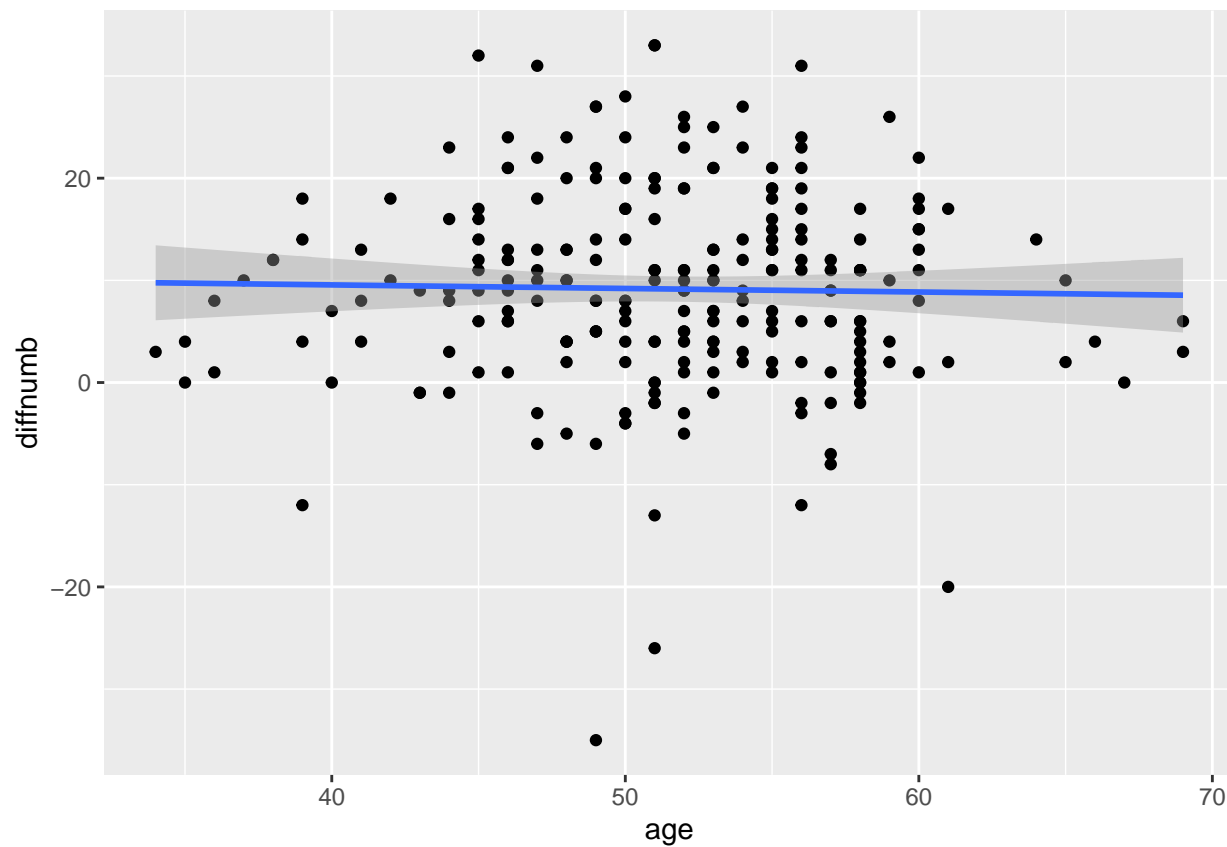
```
#age and difflet, diffnumb, since age was significant in earlier models. don't see much of a correlation
ggplot(data = sesame, mapping = aes(x = age, y = difflet)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



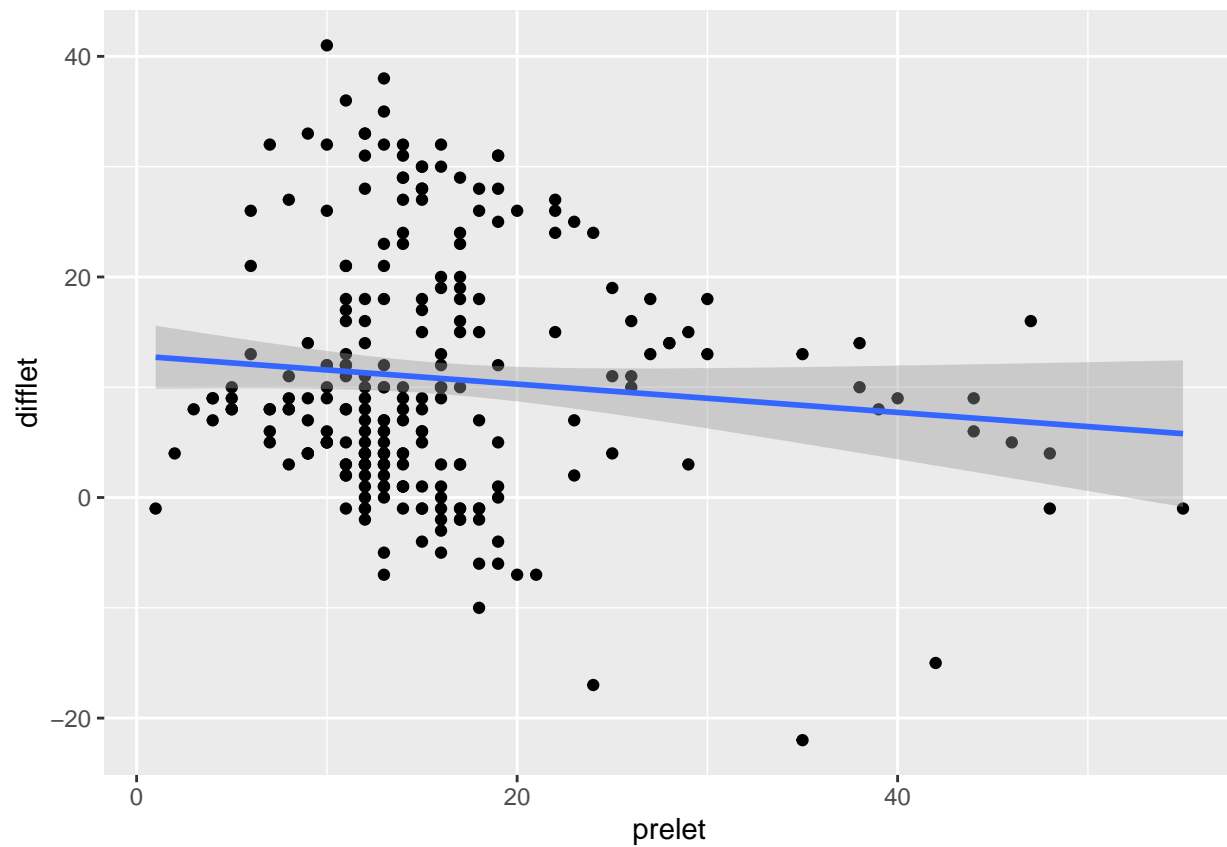
```
ggplot(data = sesame, mapping = aes(x = age, y = diffnumb)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



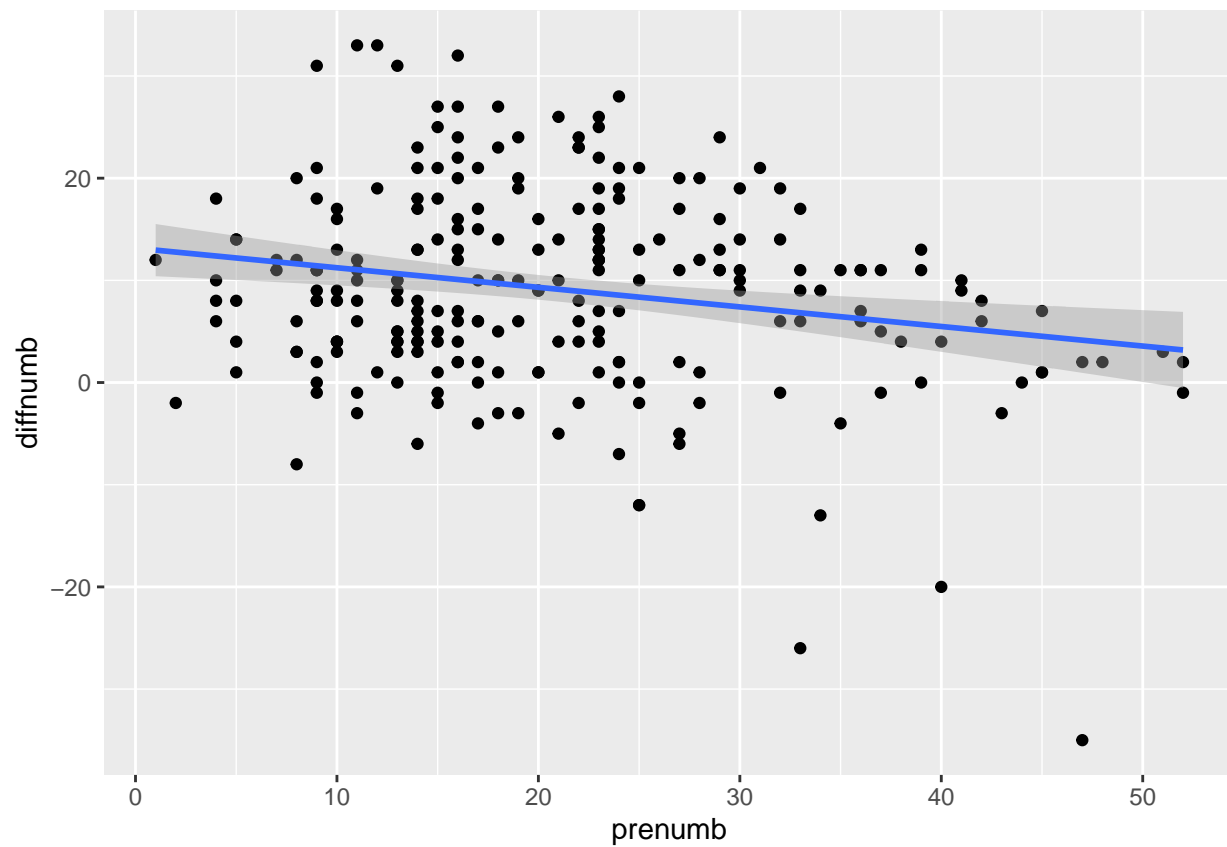
```
#prelet, prenumb and difflet, diffnumb, since age these were significant in earlier models. seems like
ggplot(data = sesame, mapping = aes(x = prelet, y = difflet)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## 'geom_smooth()' using formula 'y ~ x'
```



```
ggplot(data = sesame, mapping = aes(x = prenumb, y = diffnumb)) +  
  geom_point() +  
  geom_smooth(method = "lm")
```

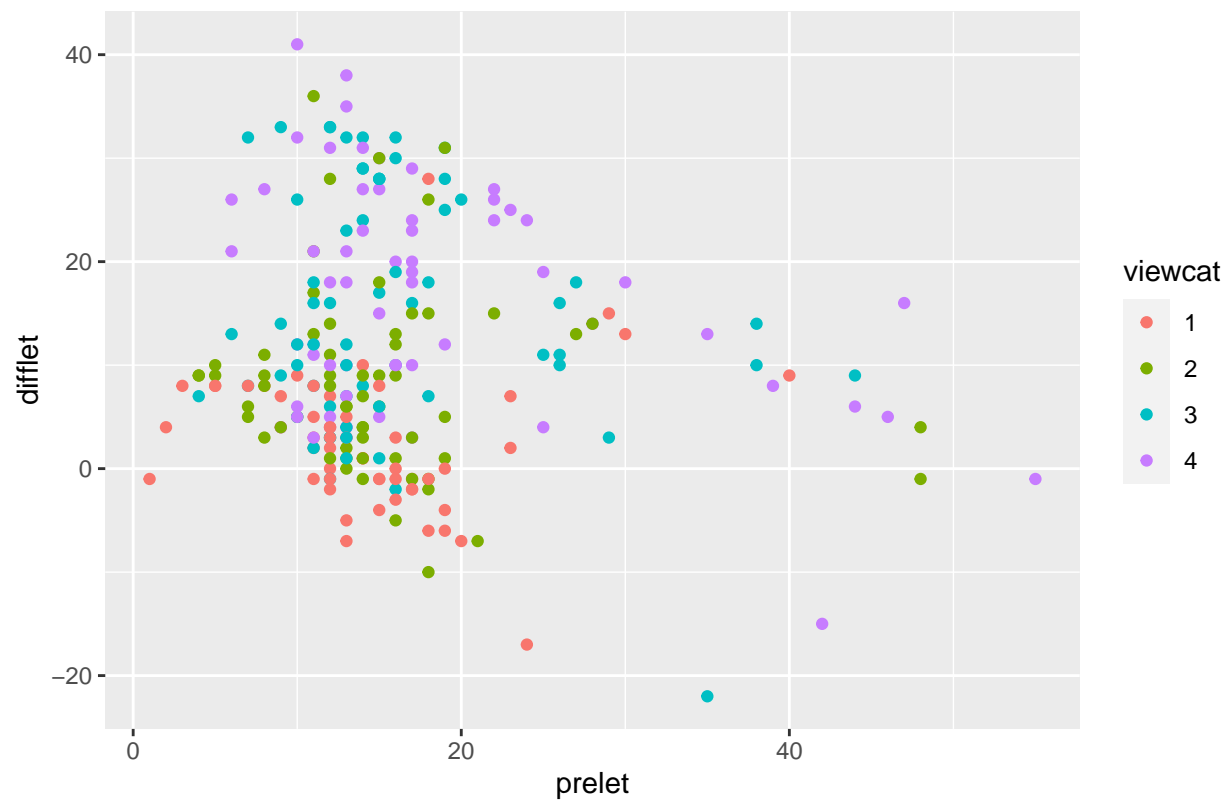
```
## 'geom_smooth()' using formula 'y ~ x'
```



#these give an idea of how groups compared on their intial test scores and how much they improved

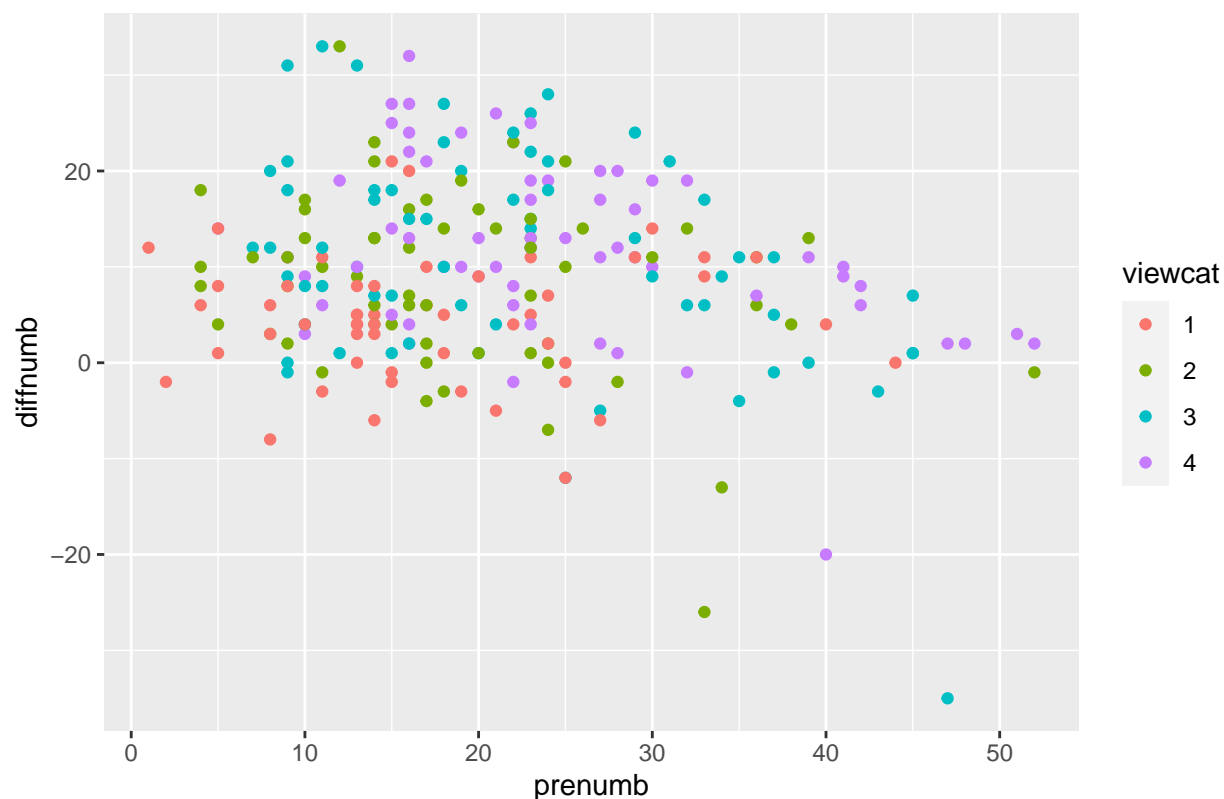
```
ggplot(data = sesame, mapping = aes(x = prelet, y = difflet, color = viewcat)) +  
  geom_point() + labs(title = "Distribution of Pretest vs. Improvement for Letters")
```

Distribution of Pretest vs. Improvement for Letters



```
ggplot(data = sesame, mapping = aes(x = prenumb, y = diffnumb, color = viewcat)) +  
  geom_point() + labs(title = "Distribution of Pretest vs. Improvement for Numbers")
```

Distribution of Pretest vs. Improvement for Numbers



Here I was trying to see whether or not any interaction terms are significant. First I tried to create

```
lm_let_interact <- lm(difflet ~ viewcat + site + viewcat*site, data = sesame)
lm_numb_interact <- lm(diffnumb ~ viewcat + site + viewcat*site, data = sesame)
lm_let_interact_full <- lm(difflet ~ viewcat + site + viewcat*site + age + prelet + viewcat*age + viewcat*prelet, data = sesame)
lm_numb_interact_full <- lm(diffnumb ~ viewcat + site + viewcat*site + age + prenumb + viewcat*age + viewcat*prenumb, data = sesame)

summary(lm_let_interact)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + site + viewcat * site, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -38.200  -5.235  -0.183   5.609  22.857
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.33333     3.05517   0.764  0.44584
## viewcat2         3.72549     3.77831   0.986  0.32521
## viewcat3        13.86667     3.67891   3.769  0.00021 ***
## viewcat4         9.80952     3.91593   2.505  0.01297 *
## siteAdv Sub       8.91667     5.50778   1.619  0.10690
## siteAdv Rural    -3.33333     4.04161  -0.825  0.41040
## siteDisadv Rural  0.05797     3.60368   0.016  0.98718
```

```
## siteDisadv Spanish      1.83333    4.83065    0.380    0.70467
## viewcat2:siteAdv Sub      0.12451    6.60893    0.019    0.98499
## viewcat3:siteAdv Sub     -5.05784    6.28312   -0.805    0.42169
## viewcat4:siteAdv Sub      1.44048    6.31161    0.228    0.81968
## viewcat2:siteAdv Rural    1.50980    5.12032    0.295    0.76838
## viewcat3:siteAdv Rural   -1.51667    4.97345   -0.305    0.76069
## viewcat4:siteAdv Rural     0.65714    5.28540    0.124    0.90117
## viewcat2:siteDisadv Rural  5.38321    5.13115    1.049    0.29527
## viewcat3:siteDisadv Rural -8.09130    5.58462   -1.449    0.14880
## viewcat4:siteDisadv Rural  6.79917    6.32365    1.075    0.28346
## viewcat2:siteDisadv Spanish 2.10784    6.50214    0.324    0.74611
## viewcat3:siteDisadv Spanish 0.96667   10.56134    0.092    0.92716
## viewcat4:siteDisadv Spanish 0.02381    6.79242    0.004    0.99721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.166 on 220 degrees of freedom
## Multiple R-squared:  0.38, Adjusted R-squared:  0.3265
## F-statistic: 7.097 on 19 and 220 DF, p-value: 1.194e-14
```

```
summary(lm_numb_interact)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + site + viewcat * site, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.600  -4.780   0.000   5.965  23.286
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.0000     3.1335   1.277   0.2031
## viewcat2          1.9412     3.8752   0.501   0.6169
## viewcat3          6.6000     3.7732   1.749   0.0817
## viewcat4          4.7143     4.0163   1.174   0.2418
## siteAdv Sub       2.5000     5.6490   0.443   0.6585
## siteAdv Rural    -1.8333     4.1452  -0.442   0.6587
## siteDisadv Rural  0.8261     3.6961   0.224   0.8234
## siteDisadv Spanish 4.1667     4.9545   0.841   0.4013
## viewcat2:siteAdv Sub  5.5588     6.7784   0.820   0.4131
## viewcat3:siteAdv Sub -0.3353     6.4442  -0.052   0.9586
## viewcat4:siteAdv Sub  1.4524     6.4734   0.224   0.8227
## viewcat2:siteAdv Rural 2.6569     5.2516   0.506   0.6134
## viewcat3:siteAdv Rural 1.2833     5.1010   0.252   0.8016
## viewcat4:siteAdv Rural 8.1857     5.4209   1.510   0.1325
## viewcat2:siteDisadv Rural 1.4327     5.2627   0.272   0.7857
## viewcat3:siteDisadv Rural -2.9261     5.7278  -0.511   0.6100
## viewcat4:siteDisadv Rural 1.7096     6.4858   0.264   0.7923
## viewcat2:siteDisadv Spanish 1.3922     6.6688   0.209   0.8348
## viewcat3:siteDisadv Spanish 2.2333    10.8321   0.206   0.8368
## viewcat4:siteDisadv Spanish -3.4810     6.9666  -0.500   0.6178
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 9.401 on 220 degrees of freedom
## Multiple R-squared:  0.1346, Adjusted R-squared:  0.05987
## F-statistic: 1.801 on 19 and 220 DF,  p-value: 0.02397

# Check for significant interaction with age
lm_let_age_interact <- lm(difflet ~ viewcat + age + viewcat*age, data = sesame)
lm_num_age_interact <- lm(diffnumb ~ viewcat + age + viewcat*age, data = sesame)
summary(lm_let_age_interact)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + age + viewcat * age, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -36.807  -6.789  -0.780   5.572  25.446
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.64425    11.48801   0.143  0.8863
## viewcat2      -19.44626    15.91007  -1.222  0.2228
## viewcat3       9.47865    15.07198   0.629  0.5300
## viewcat4      29.05524    15.17602   1.915  0.0568 .
## age            0.01625     0.22149   0.073  0.9416
## viewcat2:age    0.49855     0.30886   1.614  0.1078
## viewcat3:age    0.05894     0.29035   0.203  0.8393
## viewcat4:age   -0.29603     0.29085  -1.018  0.3098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.755 on 232 degrees of freedom
## Multiple R-squared:  0.2595, Adjusted R-squared:  0.2371
## F-statistic: 11.61 on 7 and 232 DF,  p-value: 1.214e-12
```

```
summary(lm_num_age_interact)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + age + viewcat * age, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.581  -5.653   0.122   6.047  24.533
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    11.3288    11.0211   1.028  0.305
## viewcat2       -7.1187    15.2634  -0.466  0.641
## viewcat3       -6.9615    14.4593  -0.481  0.631
## viewcat4      15.1344    14.5592   1.040  0.300
## age           -0.1308     0.2125  -0.615  0.539
## viewcat2:age    0.2142     0.2963   0.723  0.470
```



```
## viewcat3:age 0.2576 0.2785 0.925 0.356
## viewcat4:age -0.1464 0.2790 -0.525 0.600
##
## Residual standard error: 9.358 on 232 degrees of freedom
## Multiple R-squared: 0.09563, Adjusted R-squared: 0.06835
## F-statistic: 3.505 on 7 and 232 DF, p-value: 0.001356
```

```
lm_let_age_interact <- lm(difflet ~ viewcat + age + viewcat*age, data = sesame)
lm_num_age_interact <- lm(diffnumb ~ viewcat + age + viewcat*age, data = sesame)
summary(lm_let_age_interact)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + age + viewcat * age, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -36.807  -6.789  -0.780   5.572  25.446
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.64425    11.48801   0.143  0.8863
## viewcat2     -19.44626    15.91007  -1.222  0.2228
## viewcat3       9.47865    15.07198   0.629  0.5300
## viewcat4     29.05524    15.17602   1.915  0.0568 .
## age           0.01625     0.22149   0.073  0.9416
## viewcat2:age   0.49855     0.30886   1.614  0.1078
## viewcat3:age   0.05894     0.29035   0.203  0.8393
## viewcat4:age  -0.29603     0.29085  -1.018  0.3098
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.755 on 232 degrees of freedom
## Multiple R-squared: 0.2595, Adjusted R-squared: 0.2371
## F-statistic: 11.61 on 7 and 232 DF, p-value: 1.214e-12
```

```
summary(lm_num_age_interact)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + age + viewcat * age, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -45.581  -5.653   0.122   6.047  24.533
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.3288    11.0211   1.028  0.305
## viewcat2     -7.1187    15.2634  -0.466  0.641
## viewcat3     -6.9615    14.4593  -0.481  0.631
## viewcat4     15.1344    14.5592   1.040  0.300
## age          -0.1308     0.2125  -0.615  0.539
```

```
## viewcat2:age 0.2142 0.2963 0.723 0.470
## viewcat3:age 0.2576 0.2785 0.925 0.356
## viewcat4:age -0.1464 0.2790 -0.525 0.600
##
## Residual standard error: 9.358 on 232 degrees of freedom
## Multiple R-squared: 0.09563, Adjusted R-squared: 0.06835
## F-statistic: 3.505 on 7 and 232 DF, p-value: 0.001356

lm_let_prelet_interact <- lm(difflet ~ viewcat + prelet + viewcat*prelet, data = sesame)
lm_num_prenumb_interact <- lm(diffnumb ~ viewcat + age + viewcat*prenumb, data = sesame)
summary(lm_let_prelet_interact)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + prelet + viewcat * prelet, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -32.142  -6.656  -0.535   5.788  27.102
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.74762    3.11898   0.881 0.379264
## viewcat2       7.69597    4.12142   1.867 0.063119 .
## viewcat3      16.29944    4.16306   3.915 0.000119 ***
## viewcat4      19.39873    3.97221   4.884 1.94e-06 ***
## prelet        -0.01845    0.19595  -0.094 0.925072
## viewcat2:prelet -0.12206    0.25308  -0.482 0.630041
## viewcat3:prelet -0.23598    0.25031  -0.943 0.346779
## viewcat4:prelet -0.31230    0.22795  -1.370 0.171991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.686 on 232 degrees of freedom
## Multiple R-squared: 0.2699, Adjusted R-squared: 0.2479
## F-statistic: 12.25 on 7 and 232 DF, p-value: 2.583e-13
```

```
summary(lm_num_prenumb_interact)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + age + viewcat * prenumb, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -34.694  -5.863  -0.034   6.104  22.421
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.8828    5.1930  -0.363  0.7173
## viewcat2       8.3950    3.5450   2.368  0.0187 *
## viewcat3      14.6948    3.5223   4.172 4.28e-05 ***
## viewcat4      15.0735    3.7488   4.021 7.85e-05 ***
```

```
## age            0.1498      0.1020    1.468    0.1433
## prenumb        -0.0729      0.1345   -0.542    0.5883
## viewcat2:prenumb -0.2249      0.1747   -1.287    0.1993
## viewcat3:prenumb -0.3624      0.1639   -2.211    0.0280 *
## viewcat4:prenumb -0.2934      0.1657   -1.771    0.0779 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.829 on 231 degrees of freedom
## Multiple R-squared:  0.1985, Adjusted R-squared:  0.1707
## F-statistic: 7.151 on 8 and 231 DF,  p-value: 1.871e-08
```

Model Working - Sites:

```
site1 <- sesame %>%
  filter(site == "Disadv City")

lm_let_site1 <- lm(difflet ~ viewcat + age + prelet, data = site1)
lm_numb_site1 <- lm(diffnumb ~ viewcat + age + prenumb, data = site1)
summary(lm_let_site1)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + age + prelet, data = site1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.7869  -5.9767  -0.3158   7.5079  20.2906
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -9.8730     8.6555  -1.141  0.259046
## viewcat2       6.2168     4.2814   1.452  0.152275
## viewcat3      16.5824     4.1826   3.965  0.000218 ***
## viewcat4      14.8584     4.6101   3.223  0.002153 **
## age           0.3618     0.1712   2.113  0.039262 *
## prelet       -0.5043     0.1408  -3.581  0.000734 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.25 on 54 degrees of freedom
## Multiple R-squared:  0.3444, Adjusted R-squared:  0.2836
## F-statistic: 5.672 on 5 and 54 DF,  p-value: 0.0002831
```

```
summary(lm_numb_site1)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + age + prenumb, data = site1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -29.0001 -5.7258 0.3417 5.9551 19.7258
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12.7543     8.9238  -1.429  0.15869
## viewcat2      5.1136     4.2733   1.197  0.23667
## viewcat3     11.6874     4.2348   2.760  0.00788 **
## viewcat4     11.2735     4.5647   2.470  0.01671 *
## age           0.5450     0.1839   2.964  0.00451 **
## prenump      -0.6732     0.1306  -5.155 3.69e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.25 on 54 degrees of freedom
## Multiple R-squared:  0.3571, Adjusted R-squared:  0.2976
## F-statistic:      6 on 5 and 54 DF,  p-value: 0.0001748

site2 <- sesame %>%
  filter(site == "Adv Sub")

lm_let_site2 <- lm(difflet ~ viewcat + age + prelet, data = site2)
lm_numb_site2 <- lm(diffnumb ~ viewcat + age + prenump, data = site2)
summary(lm_let_site2)

##
## Call:
## lm(formula = difflet ~ viewcat + age + prelet, data = site2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.2258  -4.4533  -0.9108   5.8521  17.4240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.3742    16.8365  -0.319  0.75093
## viewcat2       5.5870     5.4024   1.034  0.30613
## viewcat3      12.4952     5.1527   2.425  0.01904 *
## viewcat4      14.2799     4.9777   2.869  0.00606 **
## age           0.4548     0.3095   1.469  0.14814
## prelet       -0.5288     0.1554  -3.403  0.00134 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.085 on 49 degrees of freedom
## Multiple R-squared:  0.3062, Adjusted R-squared:  0.2354
## F-statistic: 4.326 on 5 and 49 DF,  p-value: 0.002442

summary(lm_numb_site2)

##
## Call:
## lm(formula = diffnumb ~ viewcat + age + prenump, data = site2)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0032  -3.6010  -0.1088   6.0638  13.7954
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.4829     15.1373   0.098  0.92236
## viewcat2       9.6995     4.8527   1.999  0.05119 .
## viewcat3       9.9029     4.7108   2.102  0.04070 *
## viewcat4       8.7640     4.4861   1.954  0.05647 .
## age            0.2269     0.2902   0.782  0.43803
## prenumb       -0.3553     0.1258  -2.824  0.00684 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.096 on 49 degrees of freedom
## Multiple R-squared:  0.1773, Adjusted R-squared:  0.0933
## F-statistic: 2.111 on 5 and 49 DF,  p-value: 0.07975
```

```
site3 <- sesame %>%
  filter(site == "Adv Rural")

lm_let_site3 <- lm(difflet ~ viewcat + age + prelet, data = site3)
lm_numb_site3 <- lm(diffnumb ~ viewcat + age + prenumb, data = site3)
summary(lm_let_site3)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + age + prelet, data = site3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6379  -4.7124  -0.7541   4.1823  16.1924
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.76612     8.23972   0.821  0.414918
## viewcat2       5.35658     2.66531   2.010  0.049118 *
## viewcat3      13.14796     2.72910   4.818 1.08e-05 ***
## viewcat4      11.21109     2.86240   3.917 0.000239 ***
## age          -0.15009     0.16822  -0.892  0.375971
## prelet       -0.03014     0.21002  -0.144  0.886390
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.061 on 58 degrees of freedom
## Multiple R-squared:  0.3277, Adjusted R-squared:  0.2697
## F-statistic: 5.654 on 5 and 58 DF,  p-value: 0.0002595
```

```
summary(lm_numb_site3)
```

```
##
## Call:
```

```
## lm(formula = diffnumb ~ viewcat + age + prenumb, data = site3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.3511  -4.8791   0.3004   5.1427  20.3917
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.0735     8.7263   1.613 0.112223
## viewcat2       5.8707     2.8321   2.073 0.042635 *
## viewcat3       9.9257     2.8629   3.467 0.000998 ***
## viewcat4      15.9004     3.0545   5.206 2.66e-06 ***
## age          -0.1444     0.1821  -0.793 0.430866
## prenumb       -0.3659     0.1449  -2.526 0.014307 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.423 on 58 degrees of freedom
## Multiple R-squared:  0.3534, Adjusted R-squared:  0.2977
## F-statistic: 6.341 on 5 and 58 DF,  p-value: 9.267e-05
```

```
site4 <- sesame %>%
  filter(site == "Disadv Rural")

lm_let_site4 <- lm(difflet ~ viewcat + age + prelet, data = site4)
lm_numb_site4 <- lm(diffnumb ~ viewcat + age + prenumb, data = site4)
summary(lm_let_site4)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + age + prelet, data = site4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5327  -4.5441  -0.6393   3.3524  18.8723
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -33.27329   15.55204  -2.139 0.039060 *
## viewcat2      9.00398    2.90671   3.098 0.003712 **
## viewcat3      7.93997    3.61512   2.196 0.034413 *
## viewcat4     15.60407    4.11761   3.790 0.000539 ***
## age          0.70394    0.31092   2.264 0.029522 *
## prelet       -0.08054    0.22437  -0.359 0.721665
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.539 on 37 degrees of freedom
## Multiple R-squared:  0.4285, Adjusted R-squared:  0.3513
## F-statistic: 5.548 on 5 and 37 DF,  p-value: 0.000655
```

```
summary(lm_numb_site4)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + age + prenumb, data = site4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.7355  -4.7945  -0.3455   4.6429  15.2194
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.25730    16.94198   0.605   0.549
## viewcat2      2.93896     2.96535   0.991   0.328
## viewcat3      3.26818     3.72163   0.878   0.386
## viewcat4      6.97850     4.23123   1.649   0.108
## age          -0.04229     0.34611  -0.122   0.903
## prenumb      -0.16877     0.15874  -1.063   0.295
##
## Residual standard error: 7.763 on 37 degrees of freedom
## Multiple R-squared:  0.121, Adjusted R-squared:  0.002179
## F-statistic: 1.018 on 5 and 37 DF,  p-value: 0.4209
```

```
site5 <- sesame %>%
  filter(site == "Disadv Spanish")

lm_let_site5 <- lm(difflet ~ viewcat + age + prelet, data = site5)
lm_numb_site5 <- lm(diffnumb ~ viewcat + age + prenumb, data = site5)
summary(lm_let_site5)
```

```
##
## Call:
## lm(formula = difflet ~ viewcat + age + prelet, data = site5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.8034  -6.9687  -0.2015   4.7214  15.6128
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.3599    23.6778  -0.100   0.9223
## viewcat2       5.6551     5.0828   1.113   0.2877
## viewcat3      13.3929     9.7227   1.377   0.1935
## viewcat4      11.9715     5.6032   2.137   0.0539 .
## age           0.1969     0.4688   0.420   0.6819
## prelet        -0.2405     0.2156  -1.116   0.2864
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.524 on 12 degrees of freedom
## Multiple R-squared:  0.3434, Adjusted R-squared:  0.06986
## F-statistic: 1.255 on 5 and 12 DF,  p-value: 0.3439
```

```
summary(lm_numb_site5)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + age + prenumb, data = site5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.682  -5.022  -2.042   3.466  21.053
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.50383    30.99077   0.016   0.987
## viewcat2      3.81313     6.00059   0.635   0.537
## viewcat3      8.97897    11.31622   0.793   0.443
## viewcat4      2.55655     7.87541   0.325   0.751
## age           0.16778     0.64050   0.262   0.798
## prenumb      -0.07726     0.32429  -0.238   0.816
##
## Residual standard error: 9.867 on 12 degrees of freedom
## Multiple R-squared:  0.07267,    Adjusted R-squared:  -0.3137
## F-statistic: 0.1881 on 5 and 12 DF,  p-value: 0.9615
```

```
sitemodel1 <- lm(difflet ~ site, data = sesame)
sitemodel2 <- lm(diffnumb ~ site, data = sesame)
summary(sitemodel1)
```

```
##
## Call:
## lm(formula = difflet ~ site, data = sesame)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -32.300  -6.370  -1.685   6.327  24.700
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.3000     1.2989   7.930 8.86e-14 ***
## siteAdv Sub      9.2818     1.8782   4.942 1.47e-06 ***
## siteAdv Rural   -3.5969     1.8080  -1.989  0.0478 *
## siteDisadv Rural -3.4395     2.0103  -1.711  0.0884 .
## siteDisadv Spanish -0.6333     2.7039  -0.234  0.8150
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 235 degrees of freedom
## Multiple R-squared:  0.202,    Adjusted R-squared:  0.1884
## F-statistic: 14.87 on 4 and 235 DF,  p-value: 7.573e-11
```

```
summary(sitemodel2)
```

```
##
## Call:
## lm(formula = diffnumb ~ site, data = sesame)
##
```



```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -42.850  -5.548   0.150   5.279  24.150
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.850      1.234   6.361 1.04e-09 ***
## siteAdv Sub       4.641      1.785   2.601  0.0099 **
## siteAdv Rural     1.025      1.718   0.597  0.5513
## siteDisadv Rural  -1.129      1.910  -0.591  0.5550
## siteDisadv Spanish  2.261      2.569   0.880  0.3797
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.559 on 235 degrees of freedom
## Multiple R-squared:  0.04409,    Adjusted R-squared:  0.02782
## F-statistic:  2.71 on 4 and 235 DF,  p-value: 0.03091
```

Question 3 Work: Can we accurately predict how students' test scores might change based on their demographic characteristics and how much they watch sesame street?

My first attempt is through using regression trees with the target of predicting both difflet and diffnum based off of the demographic characteristics and how much they actually watch the program.

```
library(tree)
library(MASS)
```

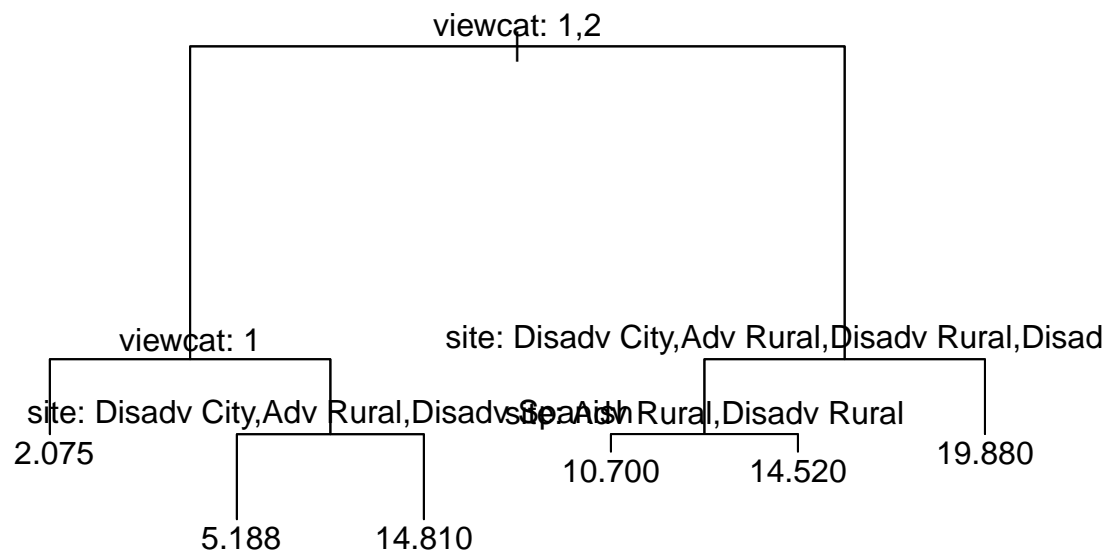
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##      select
```

```
set.seed(4)
train <- sample(1:nrow(sesame), nrow(sesame)*.7)
tree.letters <- tree(difflet ~ site + viewcat, sesame, subset = train)
summary(tree.letters)
```

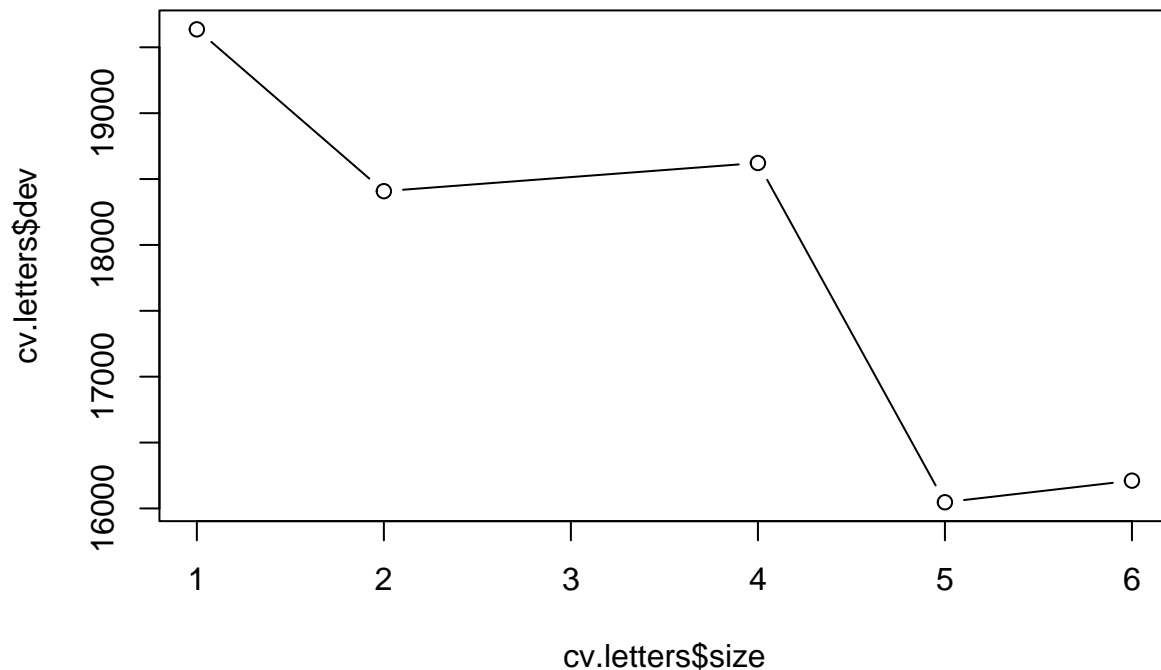
```
##
## Regression tree:
## tree(formula = difflet ~ site + viewcat, data = sesame, subset = train)
## Number of terminal nodes:  6
## Residual mean deviance:  78.97 = 12790 / 162
## Distribution of residuals:
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -36.5200 -5.7310 -0.8894  0.0000  5.8120  21.1900
```

```
plot(tree.letters)
text(tree.letters, pretty = 0)
```



Attempt at pruning...

```
cv.letters <- cv.tree(tree.letters)
plot(cv.letters$size, cv.letters$dev, type = "b")
```



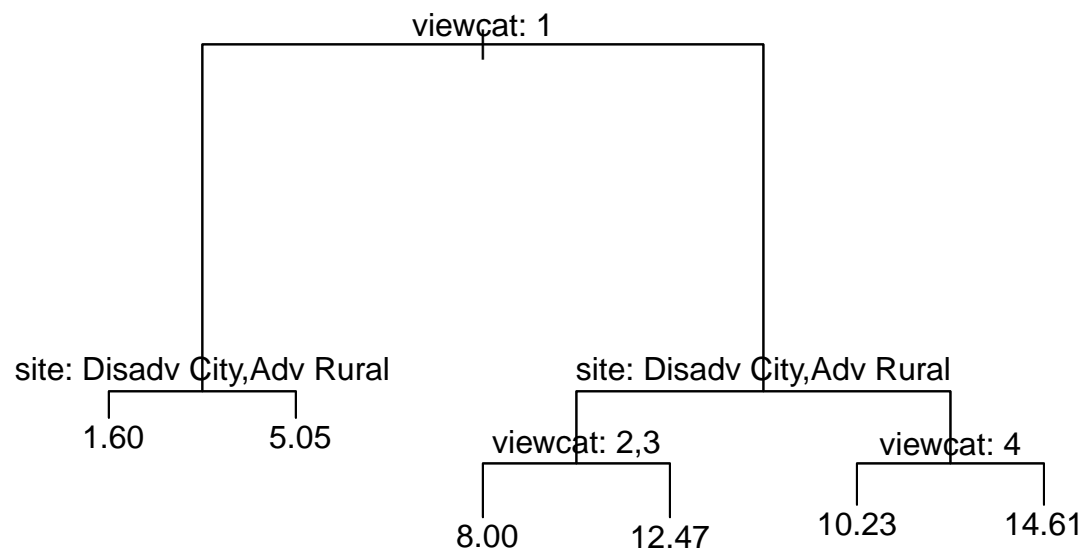
Based off of this, the CV shows that the best tree is one with 4 nodes. I made that tree below...

```
#prune.letters <- prune.tree(tree.letters, best = 5)
#summary(prune.letters)
#plot(prune.letters)
#text(prune.letters, pretty = 0)
#yhat.tree.let <- predict(prune.letters, newdata = sesame.test)
#mean((yhat.tree.let - sesame.test$diffnumb)^2)
```

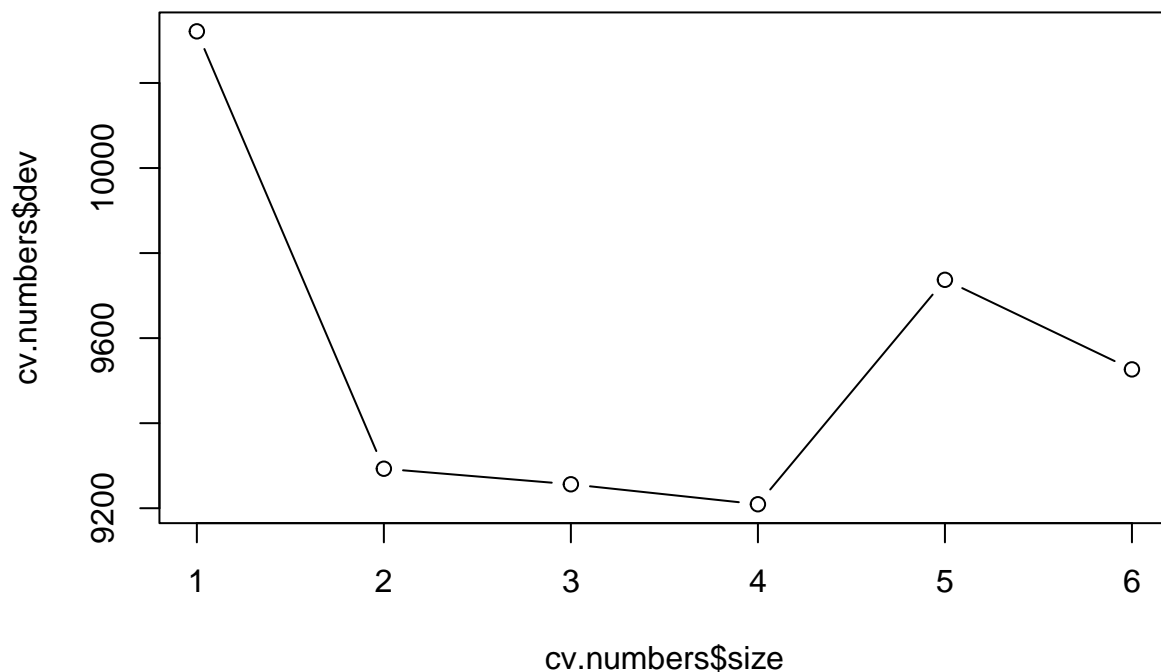
```
set.seed(4)
train <- sample(1:nrow(sesame), nrow(sesame)/2)
tree.num <- tree(diffnumb ~ site + viewcat, sesame, subset = train)
summary(tree.num)
```

```
##
## Regression tree:
## tree(formula = diffnumb ~ site + viewcat, data = sesame, subset = train)
## Number of terminal nodes: 6
## Residual mean deviance: 68.81 = 7844 / 114
## Distribution of residuals:
##      Min.   1st Qu.   Median     Mean  3rd Qu.    Max.
## -43.0000  -5.0120   -0.4154    0.0000   5.0980   23.0000
```

```
plot(tree.num)
text(tree.num, pretty = 0)
```



```
cv.numbers <- cv.tree(tree.num)
plot(cv.numbers$size, cv.numbers$dev, type = "b")
```

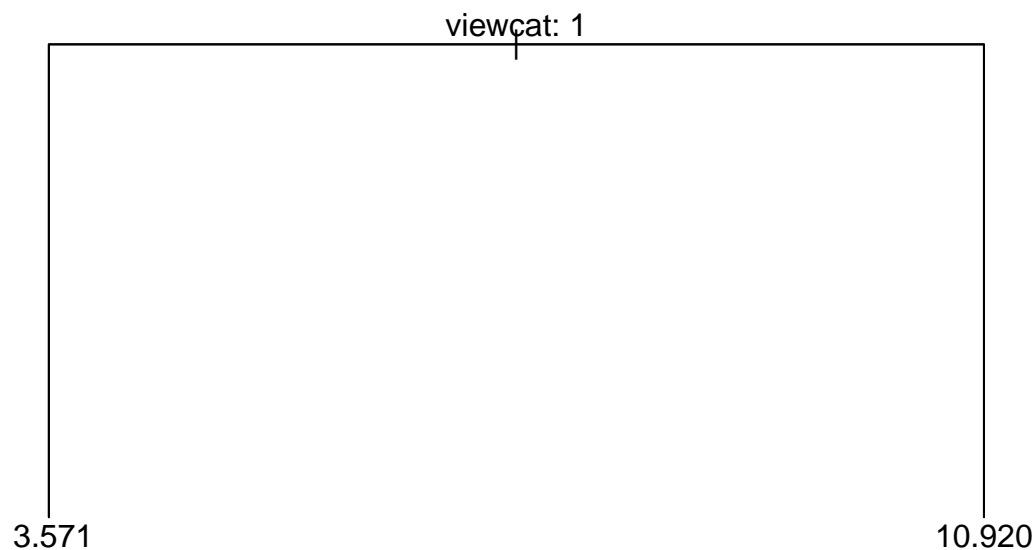


Here, the optimal nodes determined by cross validation is 2.

```
prune.numbers <- prune.tree(tree.num, best = 2)
summary(prune.numbers)
```

```
##
## Regression tree:
## snip.tree(tree = tree.num, nodes = 2:3)
## Variables actually used in tree construction:
## [1] "viewcat"
## Number of terminal nodes: 2
## Residual mean deviance: 72.81 = 8591 / 118
## Distribution of residuals:
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -45.9200 -4.9180   0.2555   0.0000  5.5920  20.0800
```

```
plot(prune.numbers)
text(prune.numbers, pretty = 0)
```



Ok so I fit the two trees above with site (which is the level of how economically disadvantaged the children are) and viewcat (which is how frequently they watch Sesame Street).

The other way that we had proposed answering this question was through GAMs. So this is what I worked with on those...

Random Forest and Boosting

```
set.seed(2)
train <- sample(1:nrow(sesame), nrow(sesame)*.7)
sesame.test <- sesame[-train,]
```

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
##
```

```
## Attaching package: 'randomForest'
```

```
## The following object is masked from 'package:dplyr':
```

```
##
```

```
##      combine
```

```
## The following object is masked from 'package:ggplot2':
```

```
##
```

```
##     margin
```

```
set.seed(1)
```

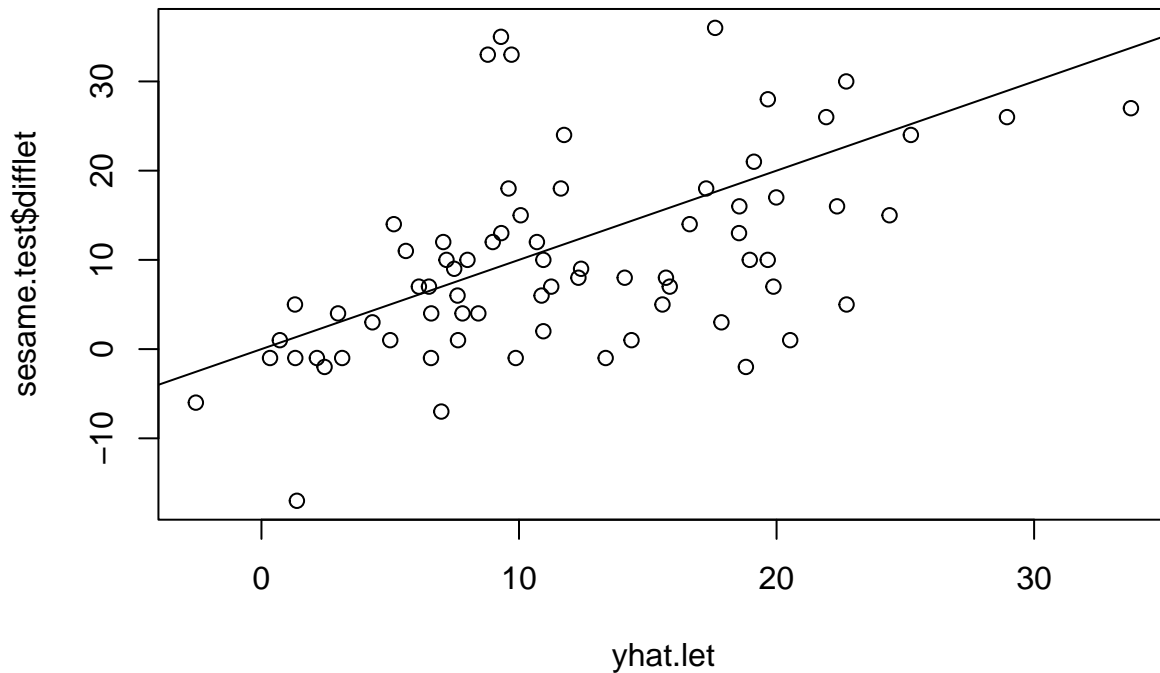
```
bag.sesame.let <- randomForest(difflet ~ site + sex + age + viewcat + setting + viewenc + prelet,  
                               data = sesame, subset = train, mtry = 7, importance = TRUE)
```

```
bag.sesame.numb <- randomForest(diffnumb ~ site + sex + age + viewcat + setting + viewenc +  
                                prenumb, data = sesame, subset = train, mtry = 7, importance = TRUE)
```

```
yhat.let <- predict(bag.sesame.let, newdata = sesame.test)
```

```
plot(yhat.let, sesame.test$difflet)
```

```
abline(0,1)
```



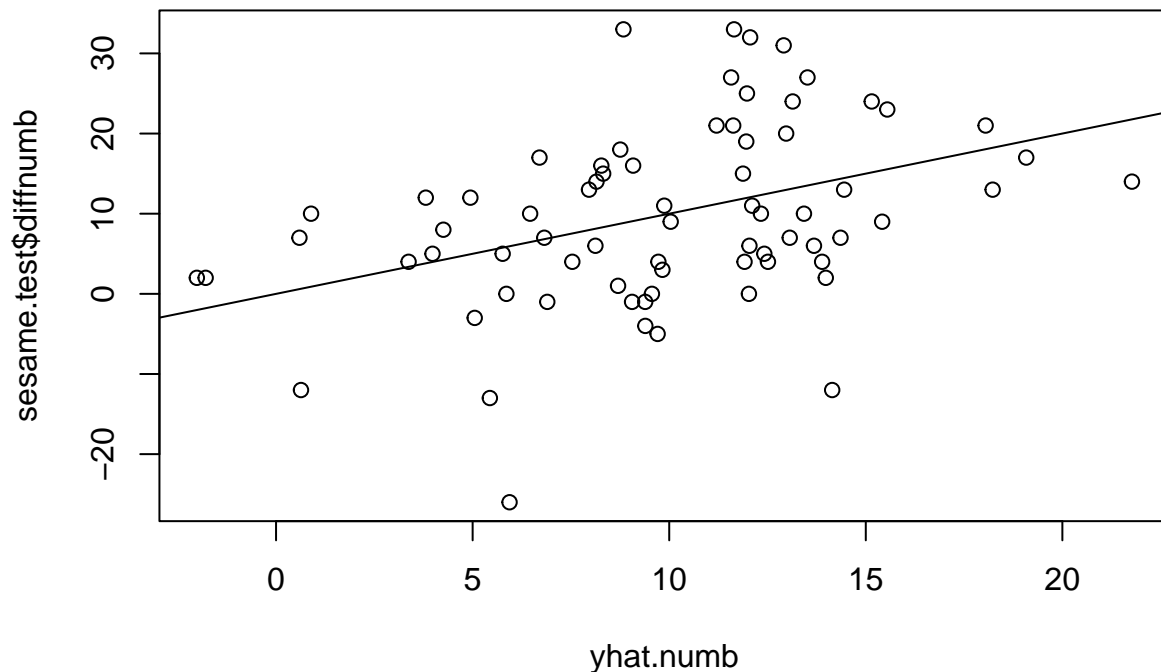
```
mean((yhat.let - sesame.test$difflet)^2)
```

```
## [1] 88.51915
```

```
yhat.numb <- predict(bag.sesame.numb, newdata = sesame.test)
```

```
plot(yhat.numb, sesame.test$diffnumb)
```

```
abline(0,1)
```



```
mean((yhat.numb - sesame.test$diffnumb)^2)
```

```
## [1] 107.5757
```

```
# Check basic linear models for prediction accuracy
```

```
q3_let <- lm(difflet ~ viewcat + site + sex + age + setting + viewenc + prelet, data = sesame, subset = train)
q3_numb <- lm(diffnumb ~ viewcat + site + sex + age + setting + viewenc + prenumb, data = sesame, subset = train)
summary(q3_let)
```

```
##
```

```
## Call:
```

```
## lm(formula = difflet ~ viewcat + site + sex + age + setting +  
##      viewenc + prelet, data = sesame, subset = train)
```

```
##
```

```
## Residuals:
```

```
##      Min      1Q  Median      3Q      Max  
## -29.138 -5.008 -0.670   5.043  18.026
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept)   -9.48541    6.14156  -1.544  0.124515  
## viewcat2       6.66729    2.20688   3.021  0.002947 **  
## viewcat3      13.33289    2.22926   5.981  1.48e-08 ***  
## viewcat4      12.39338    2.20363   5.624  8.48e-08 ***  
## siteAdv Sub    7.89499    1.95596   4.036  8.51e-05 ***
```



```
## siteAdv Rural      -5.70694    2.08981   -2.731 0.007050 **
## siteDisadv Rural   0.84153    2.23509    0.377 0.707053
## siteDisadv Spanish 1.78796    3.09170    0.578 0.563894
## sex2               1.74369    1.37627    1.267 0.207065
## age                0.32872    0.11742    2.800 0.005769 **
## setting2           -1.28778    1.62365   -0.793 0.428910
## viewenc2           -1.41415    1.64975   -0.857 0.392664
## prelet             -0.32581    0.08938   -3.645 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.707 on 155 degrees of freedom
## Multiple R-squared:  0.4568, Adjusted R-squared:  0.4148
## F-statistic: 10.86 on 12 and 155 DF,  p-value: 1.736e-15
```

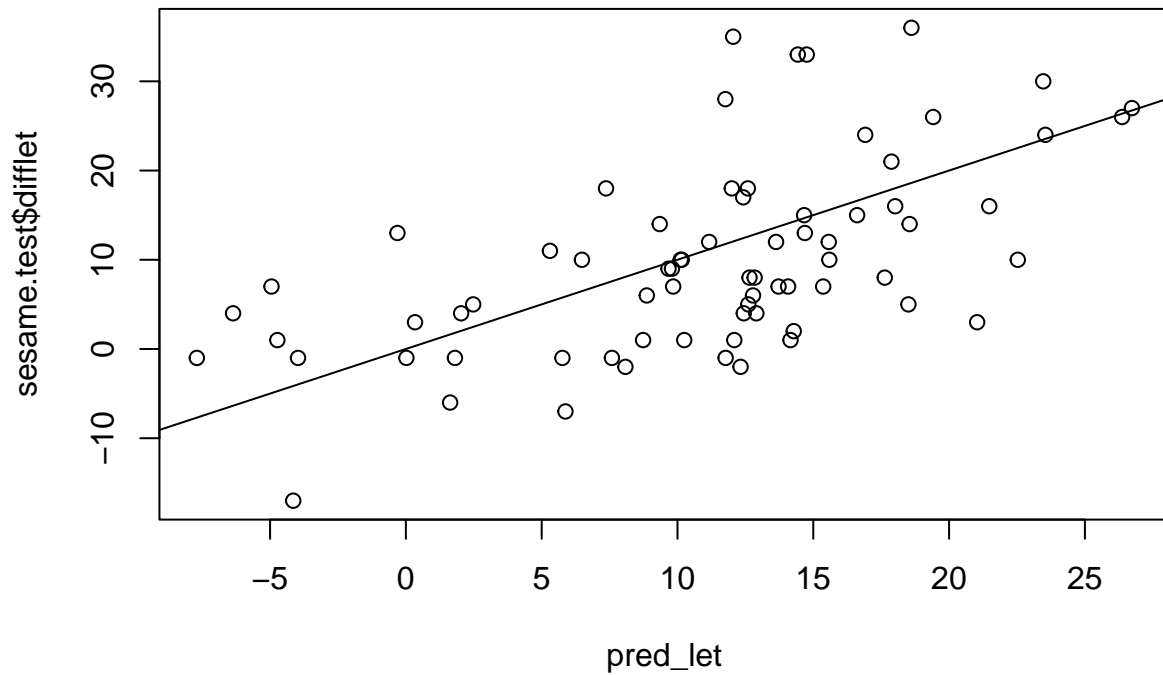
```
summary(q3_numb)
```

```
##
## Call:
## lm(formula = diffnumb ~ viewcat + site + sex + age + setting +
##     viewenc + prenumb, data = sesame, subset = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -35.699  -4.687   0.212   4.921  17.856
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -0.1301     5.9161  -0.022 0.982485
## viewcat2         4.9816     2.0637   2.414 0.016946 *
## viewcat3         7.4752     2.1163   3.532 0.000543 ***
## viewcat4         6.6205     2.0965   3.158 0.001911 **
## siteAdv Sub      5.3467     1.8316   2.919 0.004032 **
## siteAdv Rural   -1.8434     1.9683  -0.937 0.350435
## siteDisadv Rural -1.2737     2.0695  -0.615 0.539138
## siteDisadv Spanish 3.2050     2.8932   1.108 0.269688
## sex2             0.5003     1.2836   0.390 0.697237
## age             0.2003     0.1193   1.679 0.095169 .
## setting2        0.6793     1.5064   0.451 0.652652
## viewenc2       -0.2734     1.5399  -0.178 0.859336
## prenumb        -0.3444     0.0728  -4.731 4.99e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.139 on 155 degrees of freedom
## Multiple R-squared:  0.2374, Adjusted R-squared:  0.1783
## F-statistic:  4.02 on 12 and 155 DF,  p-value: 1.998e-05
```

```
# Look at MSPE for linear models
pred_let <- predict(q3_let, newdata = sesame.test)
mean((pred_let - sesame.test$difflet)^2)
```

```
## [1] 78.3376
```

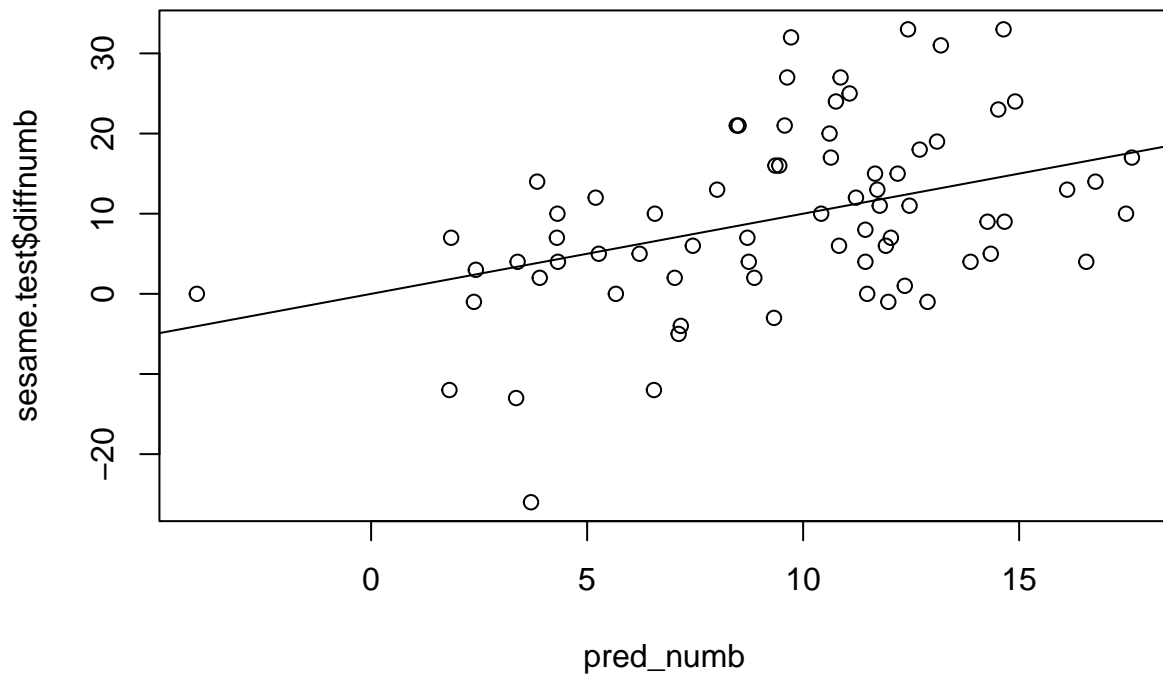
```
plot(pred_let, sesame.test$difflet)
abline(0,1)
```



```
pred_numb <- predict(q3_numb, newdata = sesame.test)
mean((pred_numb - sesame.test$diffnumb)^2)
```

```
## [1] 99.73218
```

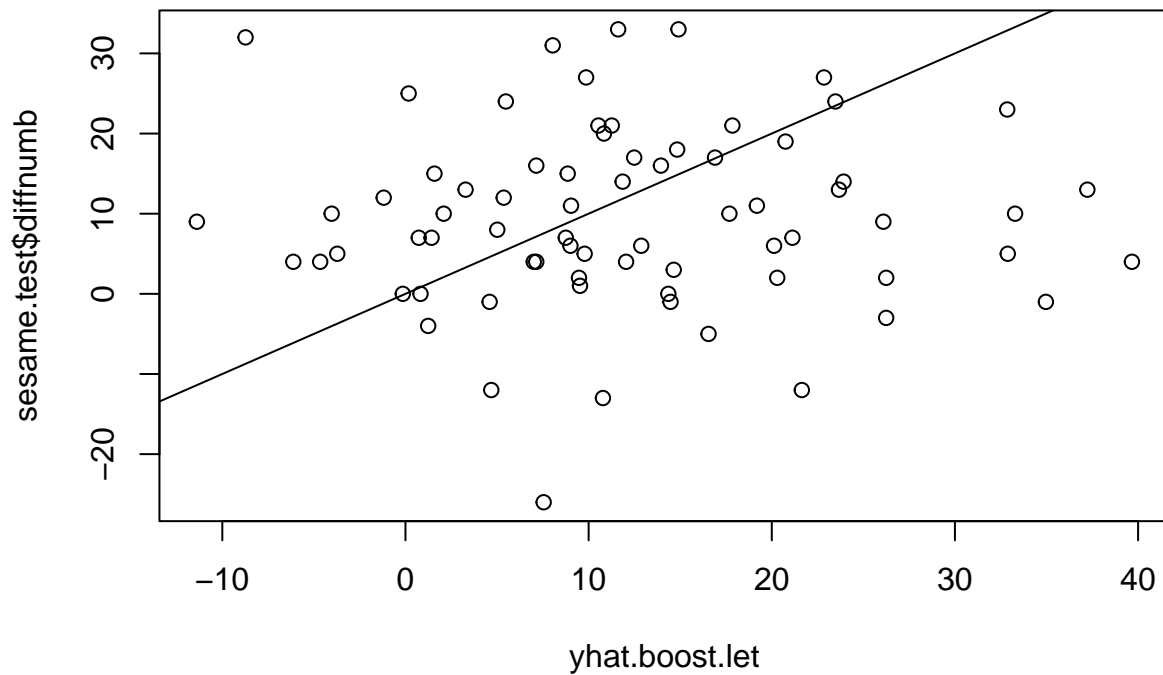
```
plot(pred_numb, sesame.test$diffnumb)
abline(0,1)
```



```
library(gbm)
```

```
## Loaded gbm 2.1.8.1
```

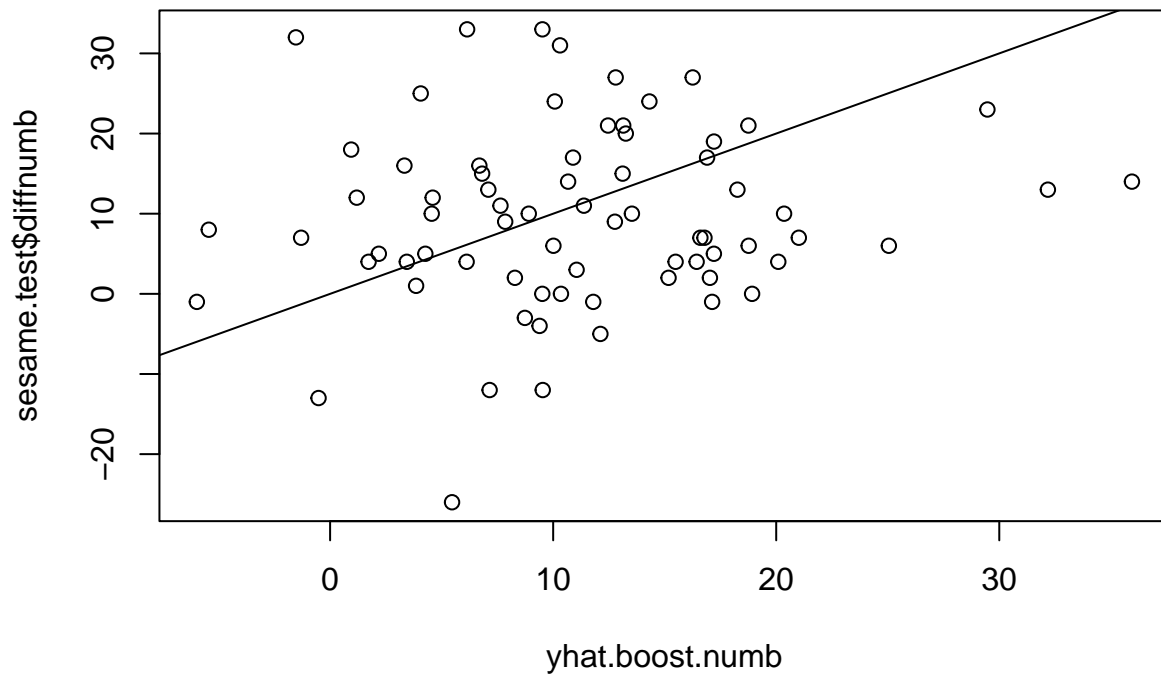
```
set.seed(2)
boost.sesame.let <- gbm(difflet ~ site + sex + age + viewcat + setting + viewenc + prelet, data =
  sesame[train,], distribution = "gaussian", n.trees = 5000,
  interaction.depth = 4)
yhat.boost.let <- predict(boost.sesame.let, newdata = sesame.test, n.trees = 5000)
plot(yhat.boost.let, sesame.test$diffnumb)
abline(0,1)
```



```
mean((yhat.boost.let - sesame.test$difflet)^2)
```

```
## [1] 157.3473
```

```
boost.sesame.numb <- gbm(diffnumb ~ site + sex + age + viewcat + setting + viewenc + prenumb,  
  data= sesame[train,], distribution = "gaussian", n.trees = 5000,  
  interaction.depth = 4)  
yhat.boost.numb <- predict(boost.sesame.numb, newdata = sesame.test, n.trees = 5000)  
plot(yhat.boost.numb, sesame.test$diffnumb)  
abline(0,1)
```



```
mean((yhat.boost.numb - sesame.test$diffnumb)^2)
```

```
## [1] 167.7286
```

```
#In the lab I was looking at it just did not explain why the degrees of freedom were chosen, but I did  
#There are 5 sites and 4 viewcats
```

```
library(gam)
```

```
## Loading required package: splines
```

```
## Loading required package: foreach
```

```
##
```

```
## Attaching package: 'foreach'
```

```
## The following objects are masked from 'package:purrr':
```

```
##
```

```
## accumulate, when
```

```
## Loaded gam 1.20.2
```

```
#gam.lets <- gam(difflet ~ ns(site, 6) + ns(viewcat, 5), data = sesame1)
#gam.nums <- gam(diffnumb ~ ns(site, 6) + ns(viewcat, 5), data = sesame1)
#summary(gam.lets)
#summary(gam.nums)

#I keep getting this error
#Error in (1 - h) * qs[i] : non-numeric argument to binary operator
```