SOCI 30005_PS2_Hinojosa

Cintia Hinojosa

5/9/2019

CSIW Dataset

```
csiw <- read_sav("csiw_new.sav")
summary(csiw, plain.scii = FALSE)
```

```
##      school          teacher            id            group
##  Min.   : 1.000   Min.   : 1.00   Min.   : 1.00   Min.   :1.000
##  1st Qu.: 1.000   1st Qu.: 4.00   1st Qu.:20.00   1st Qu.:2.000
##  Median : 2.000   Median : 8.00   Median :35.00   Median :3.000
##  Mean   : 4.509   Mean   :10.91   Mean   :32.12   Mean   :2.659
##  3rd Qu.: 9.000   3rd Qu.:18.00   3rd Qu.:44.00   3rd Qu.:4.000
##  Max.   :13.000   Max.   :55.00   Max.   :61.00   Max.   :4.000
##                                                   NA's   :40
##     treatmt           cch1            cch2          pre_test
##  Min.   :1.000   Min.   :0.000   Min.   :0.00   Min.   : 0.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.: 4.000
##  Median :1.000   Median :1.000   Median :2.00   Median : 6.000
##  Mean   :1.302   Mean   :1.137   Mean   :1.66   Mean   : 6.015
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:2.00   3rd Qu.: 9.000
##  Max.   :2.000   Max.   :3.000   Max.   :3.00   Max.   :16.000
##                  NA's   :79      NA's   :81     NA's   :79
##    post_test         ccprod1          ccprod2           ccrdr1
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000
##  1st Qu.: 6.000   1st Qu.: 2.000   1st Qu.: 3.000   1st Qu.:0.0000
##  Median : 9.000   Median : 4.000   Median : 5.000   Median :0.0000
##  Mean   : 8.601   Mean   : 3.918   Mean   : 4.911   Mean   :0.8628
##  3rd Qu.:11.000   3rd Qu.: 6.000   3rd Qu.: 7.000   3rd Qu.:1.0000
##  Max.   :15.000   Max.   :17.000   Max.   :19.000   Max.   :8.0000
##  NA's   :81       NA's   :79       NA's   :81       NA's   :79
##      ccrdr2           grade            CSIW
##  Min.   : 0.000   Min.   :1.000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.:1.000   1st Qu.:0.0000
##  Median : 1.000   Median :1.000   Median :1.0000
##  Mean   : 2.477   Mean   :1.471   Mean   :0.6978
##  3rd Qu.: 3.000   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :12.000   Max.   :2.000   Max.   :1.0000
##  NA's   :82       NA's   :40
```

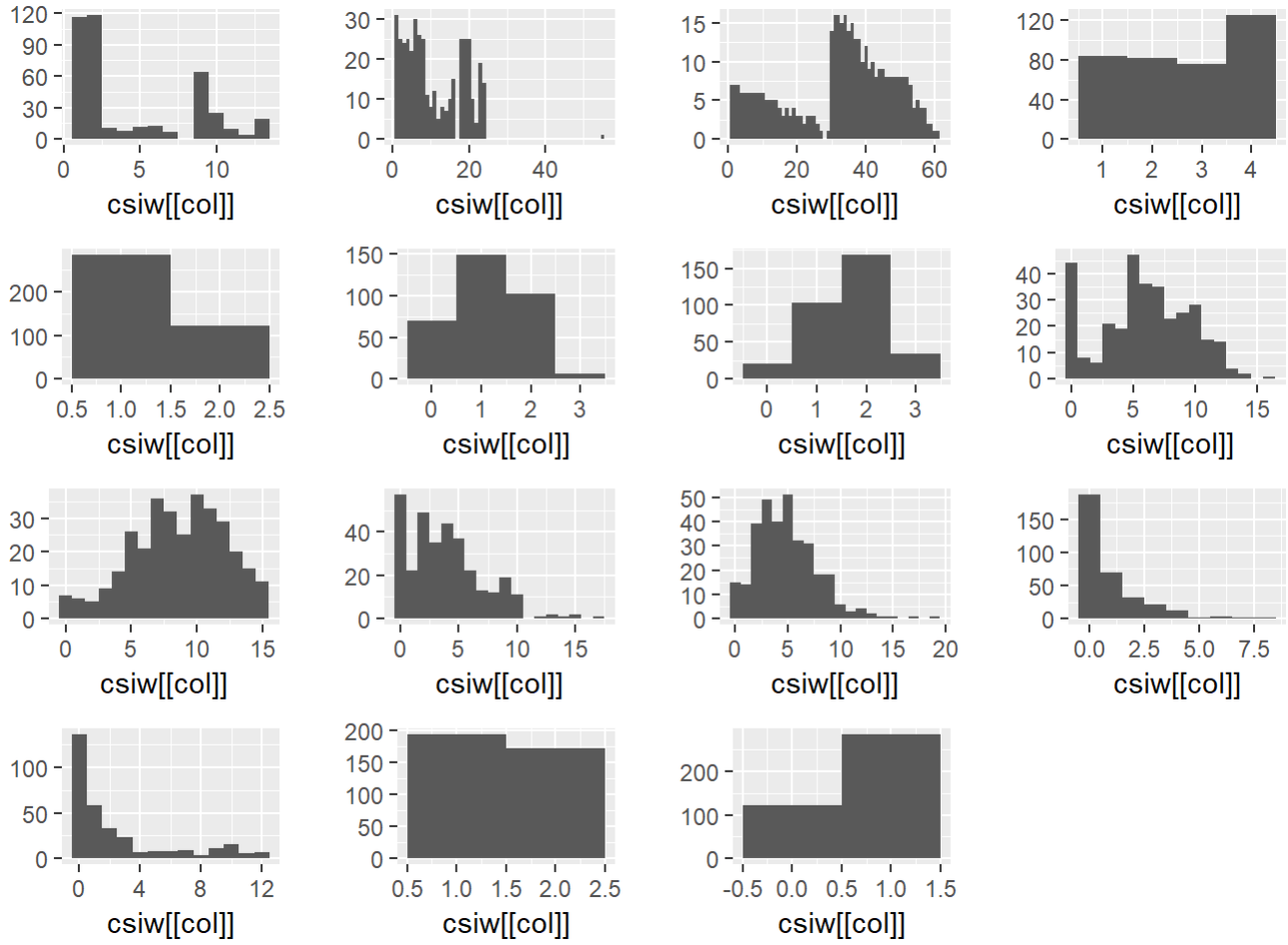
Key Variables

- CSIW (treatmt)
  - 1=CSIW
  - 0=control

- Achievement Level (group)
    - 1=High
    - 2=Average
    - 3=Low
    - 4=Learning Disability
- Holistic pretest (cch1)
    - pre-test on writing achievement
- Holistic posttest (cch2)
    - post-test on writing achievement
- Grade (grade)
    - 1=Grade 4
    - 2=Grade 5

```
# quick histograms
list <-lapply(1:ncol(csiw),
              function(col) ggplot2::qplot(csiw[[col]],
                                           geom = "histogram",
                                           binwidth = 1))

cowplot::plot_grid(plotlist = list)
```

```
## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to c
ontinuous.
## Don't know how to automatically pick scale for object of type haven_labelled. Defaulting to c
ontinuous.
```




```
## Class 'haven_labelled'  atomic [1:407] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "treatment condition"
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "labels")= Named num [1:2] 1 2
##   .. ..- attr(*, "names")= chr [1:2] "Experimental" "Control"
```

```
## Class 'haven_labelled'  atomic [1:407] 3 1 1 1 2 3 3 3 3 2 ...
##   ..- attr(*, "label")= chr "Achievement Level"
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "labels")= Named num [1:4] 1 2 3 4
##   .. ..- attr(*, "names")= chr [1:4] "High achieving" "Average achieving" "Low achieving" "Le
arning disabled"
```

```
## Class 'haven_labelled'  atomic [1:407] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "label")= chr "grade"
##   ..- attr(*, "format.spss")= chr "F2.0"
##   ..- attr(*, "labels")= Named num [1:2] 1 2
##   .. ..- attr(*, "names")= chr [1:2] "Fourth grade" "Fifth grade"
```

```
##  atomic [1:407] 2 NA 0 0 1 2 1 1 1 1 ...
##  - attr(*, "label")= chr "holistic, pretest"
##  - attr(*, "format.spss")= chr "F2.0"
```

```
##  atomic [1:407] NA NA 1 1 2 NA 1 1 2 2 ...
##  - attr(*, "label")= chr "holistic, posttest"
##  - attr(*, "format.spss")= chr "F2.0"
```

```
##
##   1   2
## 194 173
```

```
##
##   1   2
## 284 123
```

```
##
##   0   1
## 123 284
```

```
## ### Frequencies
## #### csiw$treat
## **Type:** Numeric
##
## |    &amp;nbsp; | Freq |      % | % Cum. |
## |----------:|-----:|-------:|-------:|
## |     **0** |  123 |  30.22 |  30.22 |
## |     **1** |  284 |  69.78 | 100.00 |
## | **Total** |  407 | 100.00 | 100.00 |
```

```
##
##     high  average      low learndis
##       84       82       76      125
```

```
##  [1] "treat"          "group"          "pretest"        "posttest"
##  [5] "grade"          "group_low"      "group_high"     "group_average"
##  [9] "group_NA"       "group_learndis"
```

```
##  [1] "treat"          "group"          "pretest"        "posttest"
##  [5] "grade"          "group_low"      "group_high"     "group_average"
##  [9] "group_NA"       "group_learndis" "grade_1"        "grade_2"
## [13] "grade_NA"
```

```
##
##   1   2
## 194 173
```

```
##
##   0   1
## 213 194
```

```
##
##   0   1
## 234 173
```

A. Naive Model

> Whenever you write down a model, make sure all terms are defined and all assumptions stated. If you want to discuss a table or plot, paste it directly into the text. Do not include any appendices. Do not include any tables or figures that you do not discuss in the text.

A1.Naive Population Model

$$ \begin{aligned} Y_i =*i* + {} +_i \ *i* ~ *N(0,*{y|x}) \end{aligned}

$$ In the naive model above, Y_i is the post-test score, our main outcome of interest for evaluating the effect of CSIW, β_1. The alpha parameter, α, or y-intercept of the regression line, will be the average predicted post-test score for an indivicual student in the population, Y_i, if the slope of the line, β, was 0. That is, if the beta coefficient representing participation in CSIW, β_i, were 0. The random error part of the model (2nd line) is the standard deviation parameter for the post-test scores for a particular subset of the population that has the same value for participation in CSIW and pre-test scores.

The model is naive because it assumes the post-test scores are a function of participation in CSIW, when the outcomes were likely to have been influenced by other related factors.

A naive estimate of the individual treatment effect of the CSIW program, τ_i, is the difference in achievement scores between a student in the CSIW program, $Y_i(1)$ and not in the CSIW program, $Y_i(0)$:

$$\tau_i = Y_i(1) - Y_i(0)$$

A2. Naive Model Estimate

```
# Remove rows with missing data
csiw <- csiw[which(complete.cases(csiw)), ]

# Range of post-test scores
summary(csiw$posttest) #0-3
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.737   2.000   3.000
```

```
# Run regression
naive.model <- lm(posttest~treat, data=csiw)
summary(naive.model)
```

```
##
## Call:
## lm(formula = posttest ~ treat, data = csiw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8325 -0.4706  0.1675  0.1675  1.5294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.47059    0.08477  17.348  < 2e-16 ***
## treat        0.36187    0.09871   3.666 0.000299 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.699 on 257 degrees of freedom
## Multiple R-squared:  0.04969,    Adjusted R-squared:  0.046
## F-statistic: 13.44 on 1 and 257 DF,  p-value: 0.0002994
```

```
1.47/3 # 49% control
```

```
## [1] 0.49
```

```
(1.47+.36)/3 # 61% csiw
```

```
## [1] 0.61
```

```
# save naive predicted values and residuals
csiw$naive.predicted <- predict(naive.model)
csiw$naive.residuals <- residuals(naive.model)
```

$$y_{\text{i\_posttest}} = a_i(1.47) + b_{\text{i\_CSIW}}(.36)$$

The naive model predicts that the average post-test score for an individual student from the population is significantly predicted by participating in the treatment condition. The model predicts that a CSIW student would score an average of .36 points higher than the average from a non-CSIW student, which is predicted to be 1.47.

B. ANCOVA Model

B1. ANCOVA Population Model

$$Y_i = \alpha + \beta_{\text{i\_CSIW}} * X_{\text{i\_pretest}} + \varepsilon_i$$
$$\varepsilon_i \ N(0, \sigma_{y|x})$$

In the above model, student pre-test scores are added as a covariate, $X_{\text{i\_pretest}}$. By adding this covariate, we are controling for the effect a student's pre-test score may have on their post-test score, $Y_{\text{i\_posttest}}$, so that we can get a better estimate of the CSIW treatment effect, $\beta_{\text{i\_CSIW}}$. The y-intercept, alpha α is the predicted average post-

test score for an individual who was not in the CSIW and scored a 0 on their pre-test (or at the average score if the pre-test variable is centered). The error term, ε_i, is still a parameter representing the distance between an individual's observed post-test score and the model's predicted average post-test score.

$$

$$Y_{\text{i\_posttest}} = \alpha_i + \beta_{\text{i\_CSIW}} + \gamma(X_{\text{i\_pretest}} - \bar{X}_{\text{i\_pretest}}) + \varepsilon_i$$
$$\beta = \mu_{y1} - \mu_{y0} - \gamma(\mu_{x1} - x_{x0})$$

$$

$$Y_i = \alpha + \tau_i + \beta x_i + \varepsilon_i$$

This ANCOVA model assumes a linear relationship between the expected writing post-test scores and our predictor variables, such that the slope of the CSIW beta coefficient, β, is equal across different values of the pre-test scores. The coefficient, β, is the difference between the naive estimated difference and the bias included in post-test score estimates due to influence from pre-test scores. The second section, $\gamma(\mu_{x1} - x_{x0})$, represents the bias from pre-test scores that is being subtracted from the naive estimate, $\mu_{y1} - \mu_{y0}$, of the CSIW program effect on post-test scores.

B2. ANCOVA Model Estimate

```
# Remove rows with missing data
csiw <- csiw[which(complete.cases(csiw)), ]

# ancova regression
ancova.model <- lm(posttest ~ treat + pretest, data=csiw)
summary(ancova.model)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest, data = csiw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42116 -0.42116  0.00896  0.26784  1.63840
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.99104    0.10158   9.757  < 2e-16 ***
## treat        0.43012    0.09051   4.752 3.36e-06 ***
## pretest      0.37056    0.05091   7.279 4.14e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6375 on 256 degrees of freedom
## Multiple R-squared:  0.2126, Adjusted R-squared:  0.2065
## F-statistic: 34.57 on 2 and 256 DF,  p-value: 5.138e-14
```

```
(.99 + 0 + 0)/3 # 33% control
```

```
## [1] 0.33
```

```
(.99 + .43 + 0)/3 # 47% + csiw
```

```
## [1] 0.4733333
```

```
(.99 + .43 + .37)/3 # 60% + pre-test
```

```
## [1] 0.5966667
```

```
# save ancova predicted values and residuals
csiw$ancova.predicted <- predict(ancova.model)
csiw$ancova.residuals <- residuals(ancova.model)
```

B3. ANCOVA Explanation

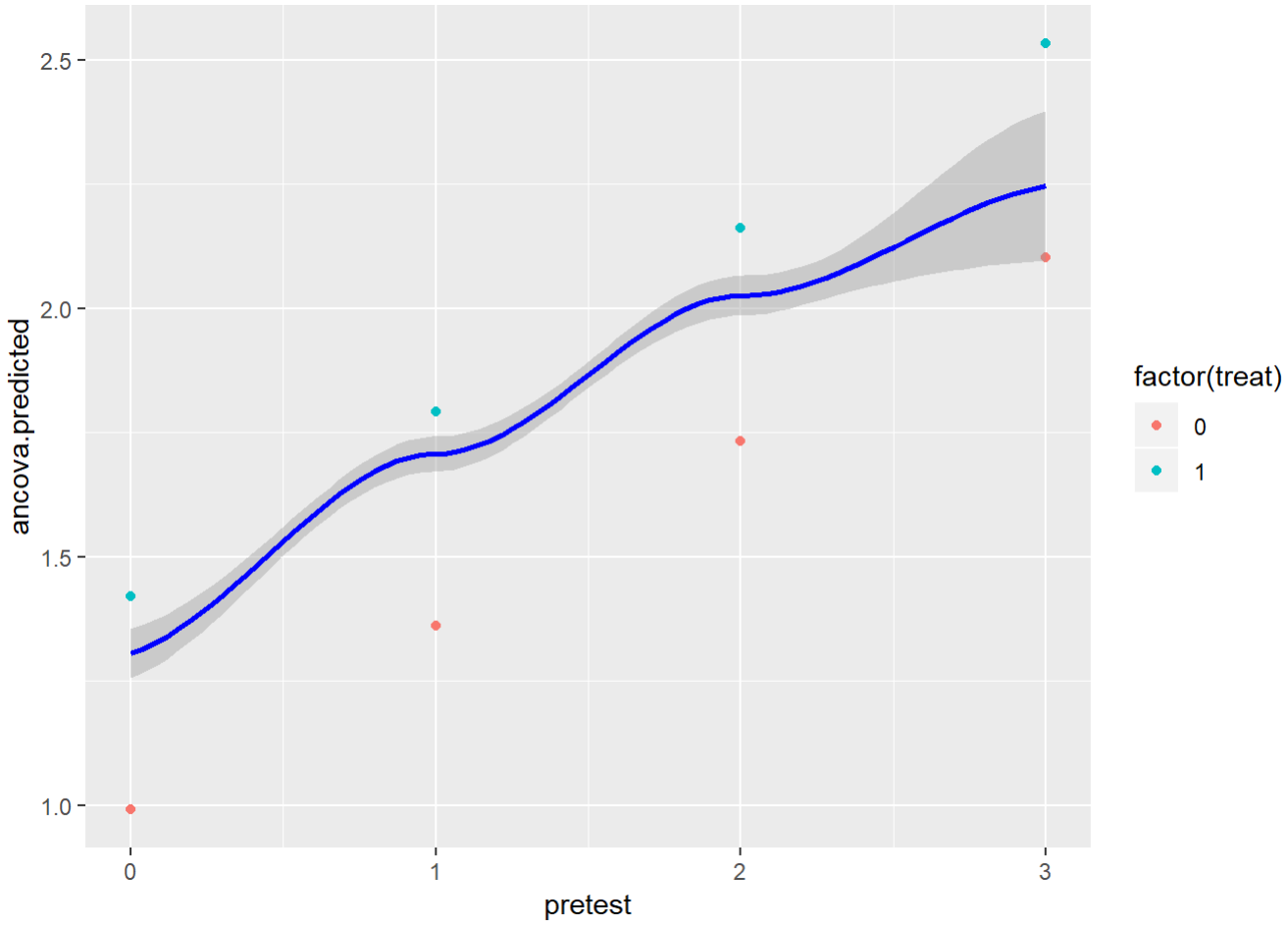
The ANCOVA model predicts that a CSIW student will score an estimated 2.95 points higher on average, compared to a non-CSIW students' average post-test score (t=6.35, p<0.0). The pre-test covariate is also a statistically significant predictor that estimates an additional 0.43 point increase on average in post-test score for every 1 point increase in the pre-test score.

The F-statistic is a measure of the error sum of squares over the total sum of squares. In the ANCOVA model (F=34.57, p<.001), the error sum of squares is smaller than that of the naive ANOVA model (F=13, p<.001), while the total sum of squares remains the same. Adding pre-test scores as a covariate reduced the error sum of squares by accounting for the variance due to pre-test scores.

B4. Plot ANCOVA Predicted Values

> Graph the predicted values as a function of pre-test and group membership

```
# Graph the predicted values as a function of pre-test and group membership
# y=predicted values; x=pre-test

ggplot(csiw, aes(x=pretest, y=ancova.predicted, color=factor(treat))) + geom_point()+geom_smooth
(method="loess", color="blue")
```

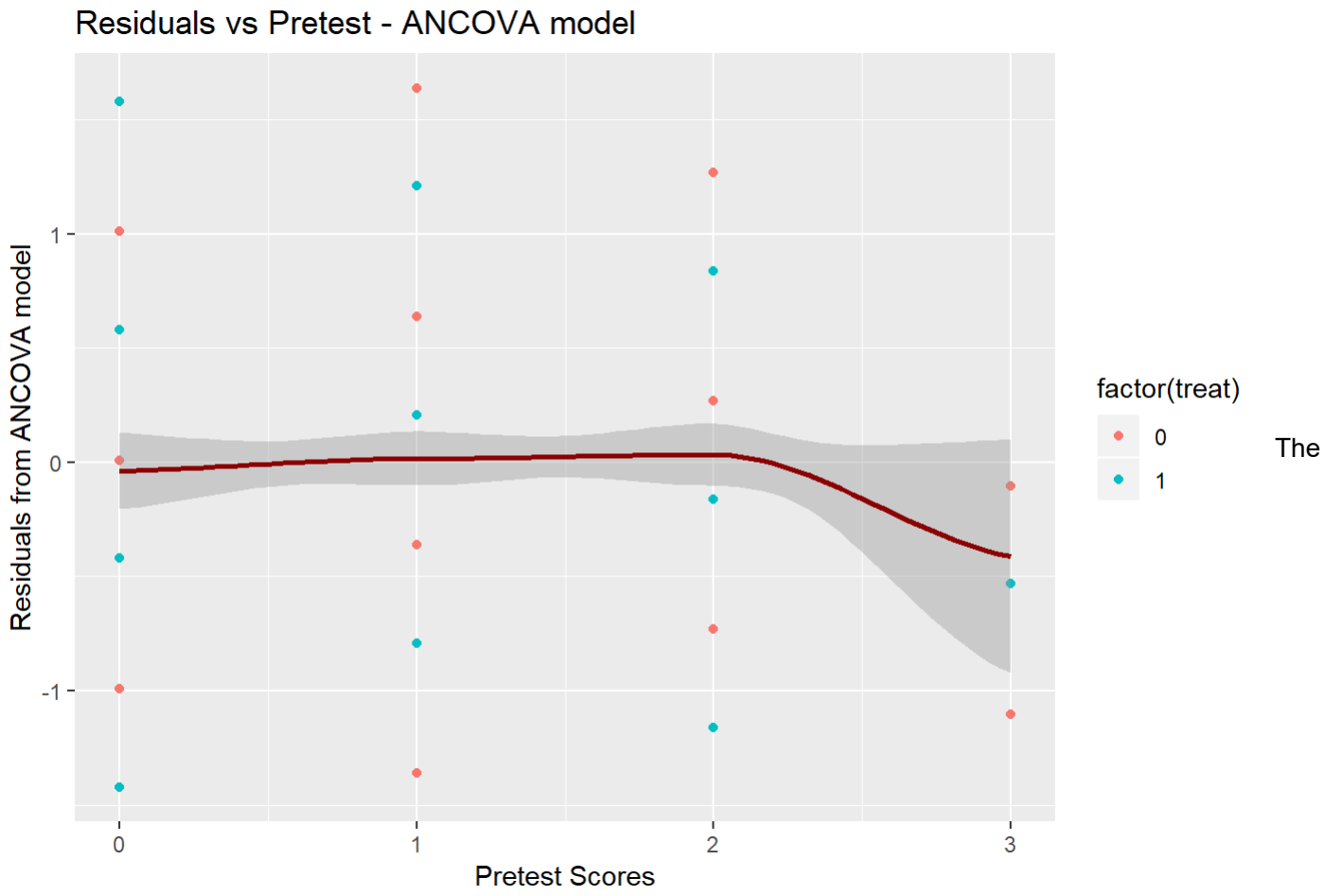


There is a mostly linear parallel pattern that shows a positive main effect of CSIW, where CSIW students are predicted to score higher than the non-CSIW students at about the same rate of change. The gap between the green CSIW data ponts and the red non-CSIW data points represents the estimated mean difference between groups, $\bar{Y}_1 - \bar{Y}_2$. There is a jump in the loess line that may suggest the data does not fit the linearity assumption.

C. ASSESSING LINEARITY

C1-2. Plot ANCOVA Residuals and Pre-test Covariate

```
# plot ancova residuals against pre-test covariate w/loess line
# y=residuals; x=pretest
ggplot(csiw, aes(x=pretest, y=ancova.residuals, color=factor(treat))) +geom_point()+geom_smooth
(method="loess", color="darkred")+
  xlab("Pretest Scores") + ylab("Residuals from ANCOVA model") + labs(title="Residuals vs Pretes
t - ANCOVA model")
```



The loess line is more straight than before but still has two slight curve point that seem to cancel each other out. The residuals are also in a trumpet shape with more variation at the low end of the pre-test scores.

C3. ANCOVA Residual and Pre-Test Linearity Explaination

> A scattorplot of the ancova residuals against pre-test scores shows a curve in the loess line indicating the it may be a quadratic function rather than a linear function, so we should run further analyses to adjust.

D. QUADRATIC ANCOVA MODEL

D0. Center and square pre-test

```
# Create centered pretest variable and a centered squared
csiw$pretest.c <- csiw$pretest-mean(csiw$pretest)
csiw$pretest.csq <- (csiw$pretest.c)^2
```

D1. Quadratic ANCOVA Population Model

> Write down a model that uses CSIW, pretest_c, and pretest_csq" are predictors

$$Y_{\text{i\_posttest}} = \alpha_i + \beta Z_i + \varepsilon_i$$
$$\beta = Z_i + \gamma(Z_i) + \varepsilon_i$$

The naive estimate of the coefficient for a student's achievement as predicted by their participation in the CSIW program. The potential bias in the model is captured by $\gamma(Z_i) + \varepsilon_i$. The two conditions for bias to equal 0 is for $\gamma = 0$ or if there was no difference in the predicted writing post-test score across individual students with different pre-test scores.

D2. Quadratic ANCOVA Estimate and Explanation

> Tell us how to interpret all of the coefficients in the model (including the intercept)

```
# run quadratic ancova model
quadratic.ancova.model <- lm(posttest ~ treat+pretest.c+pretest.csq, data=csiw)
summary(quadratic.ancova.model)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest.c + pretest.csq, data = csiw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42394 -0.35734  0.05348  0.26214  1.64266
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.48452    0.09069  16.370  < 2e-16 ***
## treat        0.41081    0.09146   4.492 1.07e-05 ***
## pretest.c    0.36979    0.05083   7.275 4.26e-12 ***
## pretest.csq -0.08175    0.05993  -1.364    0.174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6365 on 255 degrees of freedom
## Multiple R-squared:  0.2183, Adjusted R-squared:  0.2091
## F-statistic: 23.74 on 3 and 255 DF,  p-value: 1.387e-13
```

```
# Save predicted values and residuals
csiw$quadratic.ancova.predicted <- predict(quadratic.ancova.model)
csiw$quadratic.ancova.residuals <- residuals(quadratic.ancova.model)
```

The y-intercept is the predicted average post-test score of a non-CSIW student, controlling for the pre-test performance (a=1.49). The effect of the CSIW program on post-test score is predicted to increase the average score by .41 points (t=4.49, p<0.0). The now centered pre-test covariate is also a statistically significant predictor of post-test scores in this model and it is predicted to add an average of .37 points (t=7.28,p<0.0).

E. SEARCHING FOR CONFOUNDERS

E1. Check for ommissions

> Check to see if you have omitted any confounders. Tell us what you found.

```
# Dummies from achievement level eg. achievement level = high
## corresponding dummy in the dataset = group_high

# LEARNING DISABILITY #
## Y(posttest) <-- Z(Learning disability)
## ANOVA: continuous vs. discrete
summary(aov(posttest~group_learndis, data=csiw))
```

```
##                 Df Sum Sq Mean Sq F value   Pr(>F)
## group_learndis   1   23.2  23.198   54.72 1.97e-12 ***
## Residuals      257  109.0   0.424
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## X(CSIW) <-- Z(Learning disability)
## CHI2: discrete vs. discrete
csiw$treat.factor <- factor(csiw$treat)
chisq.test(csiw$treat.factor,csiw$group_learndis)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  csiw$treat.factor and csiw$group_learndis
## X-squared = 7.2728, df = 1, p-value = 0.007001
```

```
# LOW ACHIEVEMENT #
## Y(posttest) <-- Z(low)
## ANOVA: continuous vs. discrete
summary(aov(posttest~group_low, data=csiw))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## group_low     1   1.45  1.4522   2.856 0.0923 .
## Residuals   257 130.69  0.5085
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## X(CSIW) <-- Z(low)
## CHI2: discrete vs. discrete
csiw$treat.factor <- factor(csiw$treat)
chisq.test(csiw$treat.factor,csiw$group_low)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  csiw$treat.factor and csiw$group_low
## X-squared = 0.27366, df = 1, p-value = 0.6009
```

```
# AVG ACHIEVEMENT #
## Y(posttest) <-- Z(avg)
## ANOVA: continuous vs. discrete
summary(aov(posttest~group_average, data=csiw))
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## group_average   1   2.47  2.4710   4.897 0.0278 *
## Residuals     257 129.68  0.5046
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
### X(CSIW) <-- Z(avg)
### CHI2: discrete vs. discrete
chisq.test(csiw$treat.factor,csiw$group_average)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  csiw$treat.factor and csiw$group_average
## X-squared = 1.3705e-29, df = 1, p-value = 1
```

```
# HIGH ACHIEVEMENT #
## Y(posttest) <-- Z(high)
## ANOVA: continuous vs. discrete
summary(aov(posttest~group_high, data=csiw))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## group_high    1  16.75  16.748    37.3 3.73e-09 ***
## Residuals   257 115.40   0.449
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## X(CSIW) <-- Z(high)
## CHI2: discrete vs. discrete
chisq.test(csiw$treat.factor,csiw$group_high)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  csiw$treat.factor and csiw$group_high
## X-squared = 9.4347, df = 1, p-value = 0.002129
```

A variable would a confound if it was a statistically significant predictor of both CSIW and post-test scores. According to the tests of association results, learning disability is a confound as it is a predictor of both post-test scores (F=54.72, p<.001) and participation in CSIW (x^2=7.27, p<.007). High achievement is the other confound that has a significant association with post-test scores (F=37.3, p<.001) and CSIW (x^2=9.44, p<.002) This imbalance across conditions may indicate that students with learning disabilities and high achieving students did not participate in CSIW by random and may also have an extra (dis)advantage in taking the writing post-test.

```
# Dummies from grade levels: grade_1, grade_2

# GRADE 1 #
## Y(posttest) <-- Z(4th grade)
## ANOVA: continuous vs. discrete
summary(aov(posttest~grade_1, data=csiw))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## grade_1       1   6.22   6.218   12.69 0.000438 ***
## Residuals   257 125.93   0.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## X(CSIW) <-- Z(4th grade)
## CHI2: discrete vs. discrete
chisq.test(csiw$treat.factor,csiw$grade_1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  csiw$treat.factor and csiw$grade_1
## X-squared = 0.9978, df = 1, p-value = 0.3178
```

```
# GRADE 2 #
## Y(posttest) <-- Z(5th grade)
## ANOVA: continuous vs. discrete
summary(aov(posttest~grade_2, data=csiw))
```

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## grade_2       1   6.22   6.218   12.69 0.000438 ***
## Residuals   257 125.93   0.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## X(CSIW) <-- Z(5th grade)
## CHI2: discrete vs. discrete
chisq.test(csiw$treat.factor,csiw$grade_2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  csiw$treat.factor and csiw$grade_2
## X-squared = 0.9978, df = 1, p-value = 0.3178
```

Regarding grade level, both 4th and 5th grade are significantly associated with post-test scores (F=12.69, p<0.001), but not with participation in CSIW so they are not confounds and should not be in our final model.

E2. Estimate with Confounders

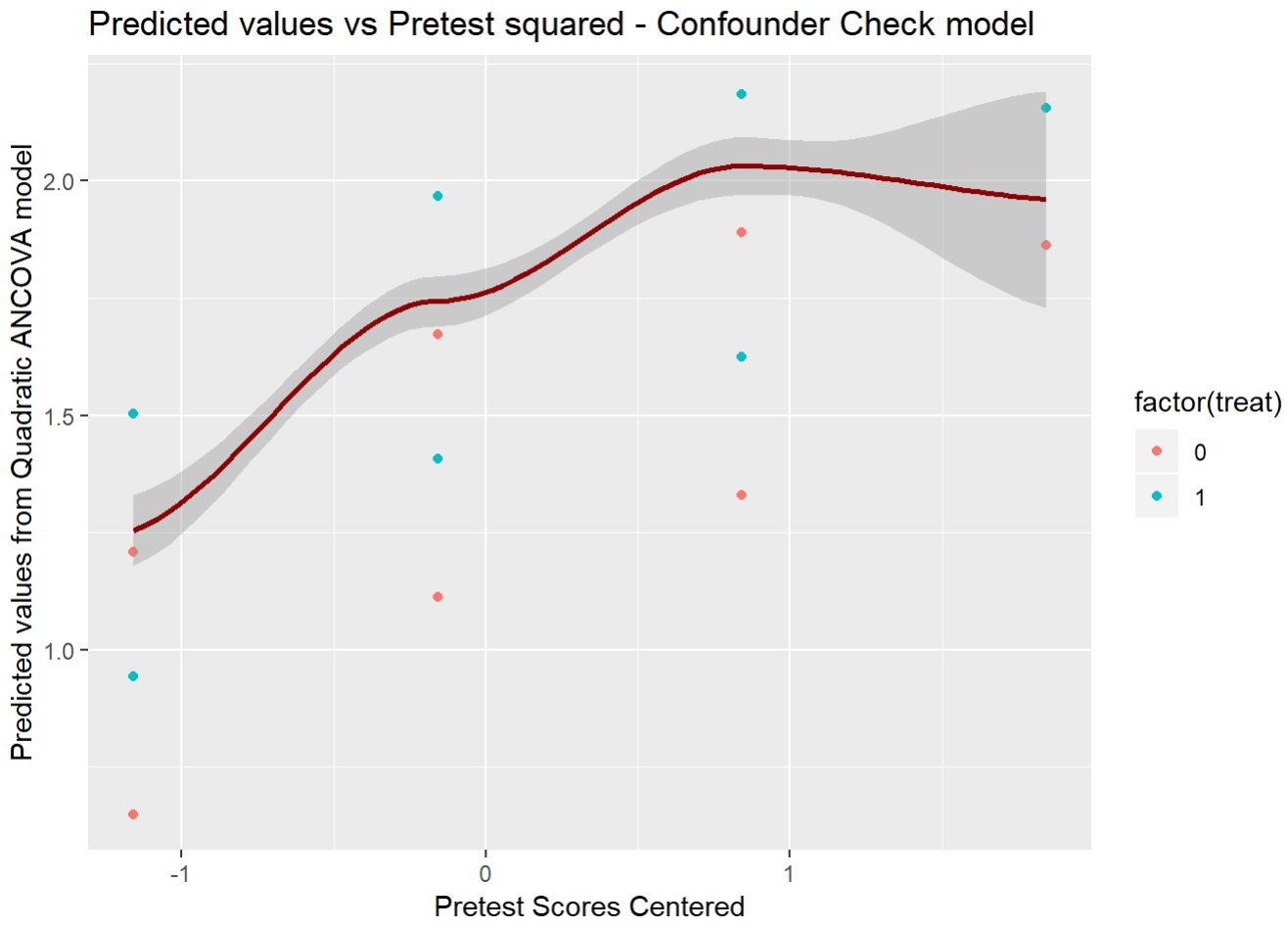
> Re-estimate the model now but add any confounders

Estiamte with learning disability confound

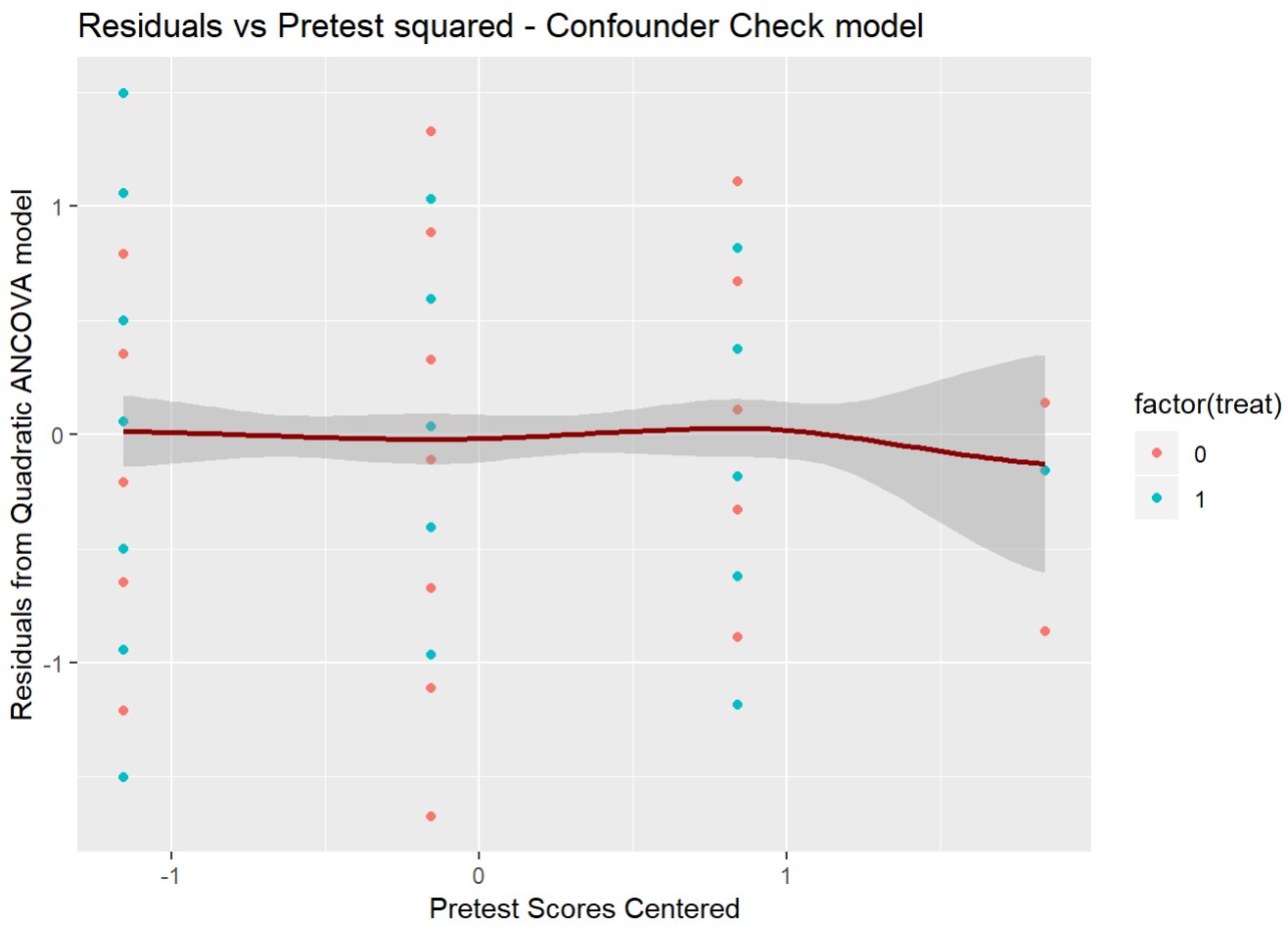
```
# Learning disability (LD) confound #
confounder.check.model.ld <- lm(posttest ~ treat+pretest.c+pretest.csq+group_learndis, data=csi
w)
summary(confounder.check.model.ld)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest.c + pretest.csq + group_learndis,
##     data = csiw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.67288 -0.33052  0.03353  0.36342  1.49709
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.72374    0.09378  18.380  < 2e-16 ***
## treat           0.29360    0.08789   3.341 0.000962 ***
## pretest.c       0.30185    0.04895   6.167 2.72e-09 ***
## pretest.csq    -0.12282    0.05657  -2.171 0.030849 *
## group_learndis -0.56028    0.09289  -6.032 5.69e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5964 on 254 degrees of freedom
## Multiple R-squared:  0.3163, Adjusted R-squared:  0.3055
## F-statistic: 29.37 on 4 and 254 DF,  p-value: < 2.2e-16
```

```
# LD Confound model predicted values
csiw$confounder.check.model.ld <- predict(confounder.check.model.ld)

# LD Confound model residuals
csiw$confounder.check.residuals.ld <- residuals(confounder.check.model.ld)

# LD Confound model predicted values vs pretest confounds
par(mfrow=c(1,2))

ggplot(csiw, aes(x=pretest.c, y = confounder.check.model.ld, color=factor(treat))) + geom_point
()+
  geom_smooth(method="loess", color="darkred")+  xlab("Pretest Scores Centered") + ylab("Predict
ed values from Quadratic ANCOVA model") + labs(title="Predicted values vs Pretest squared - Conf
ounder Check model")
```




```
# Confound model residuals vs pretest confounds
ggplot(csiw, aes(x=pretest.c, y = confounder.check.residuals.ld, color=factor(treat))) + geom_po
int()+
  geom_smooth(method="loess", color="darkred")+  xlab("Pretest Scores Centered") + ylab("Residua
ls from Quadratic ANCOVA model") + labs(title="Residuals vs Pretest squared - Confounder Check m
odel")
```



I added learning disability and high achievement student status as confounding variables.

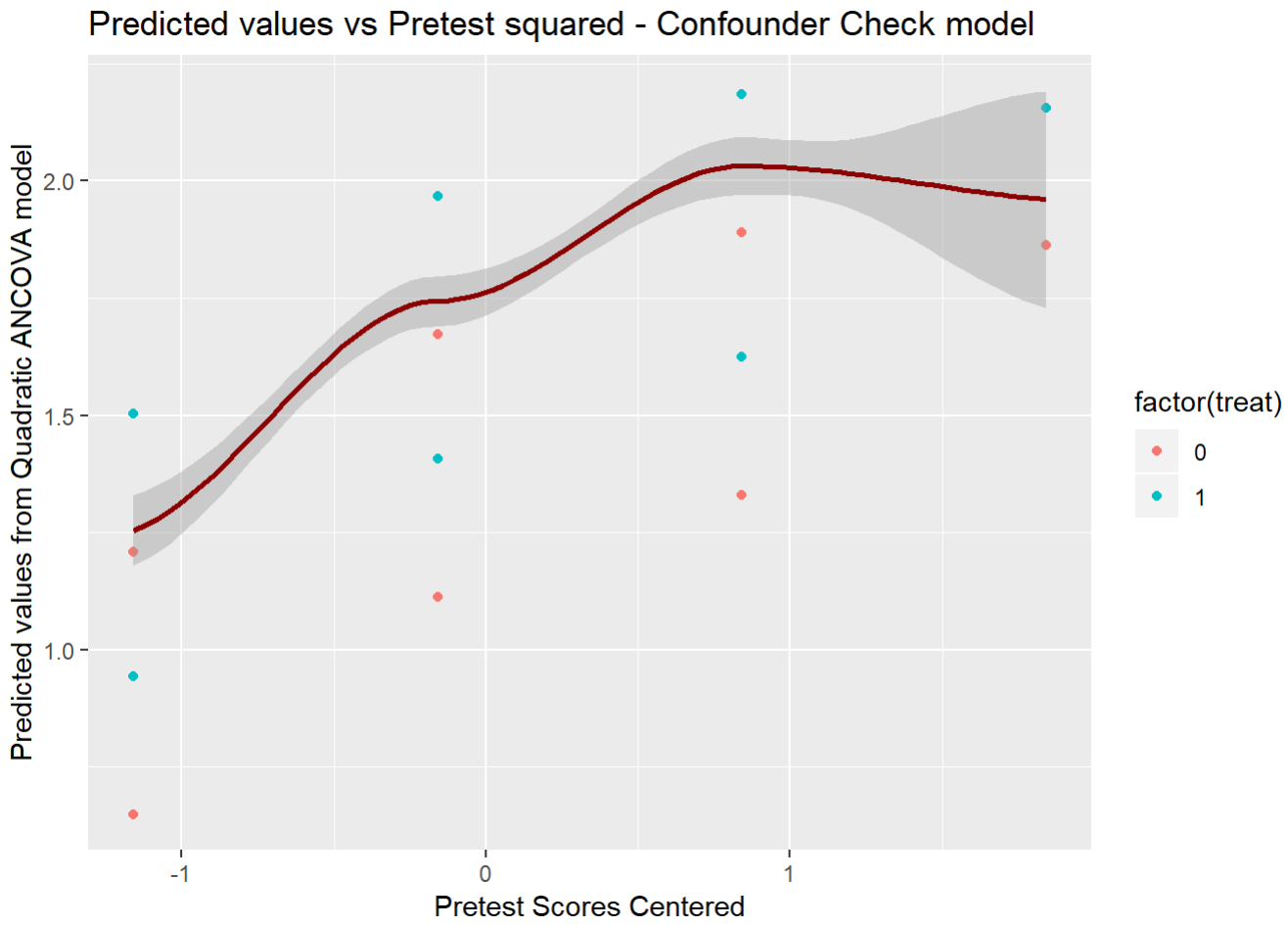
The model adjusted for learning disability status shows that CSIW participation is still a significant predictor of writing post-test scores (t=3.341, p<.001) that estimates an average increase of .29 points. Pre-test scores are also statistically significant and estimate an additional average .30 points for each 1 unit increase in pre-test score (t=6.167, p<.0001). Lastly, having a learning disability is also statistically significant (-6.03, p<.001) and which tacks on an average deduction of 0.56 points from the post-test score.

Estimate with high achieving confound

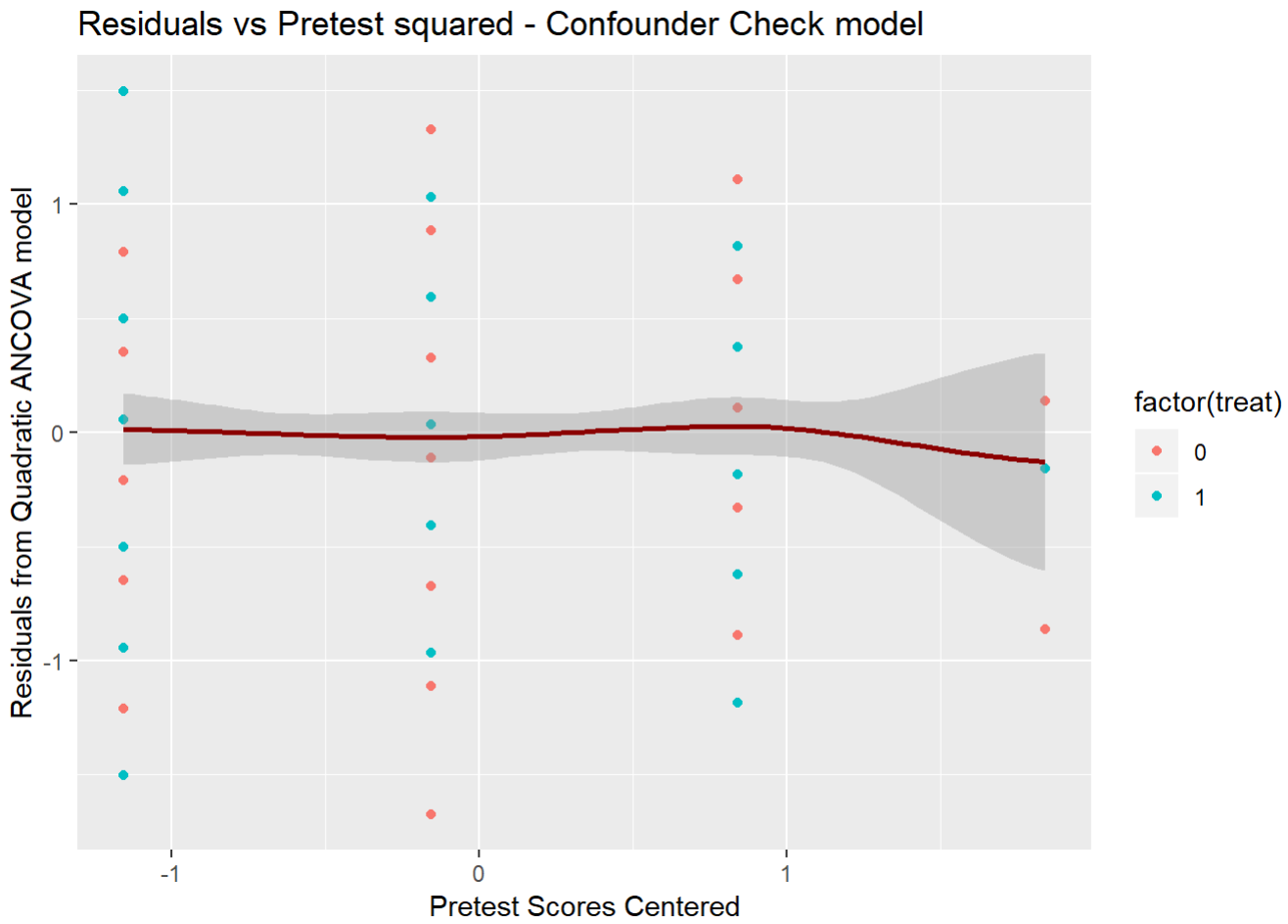
```
# High achieveing (HA) confound #
confounder.check.model.ha <- lm(posttest ~ treat+pretest.c+pretest.csq+group_high, data=csiw)
summary(confounder.check.model.ha)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest.c + pretest.csq + group_high,
##     data = csiw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.40749 -0.39813  0.01481  0.34724  1.59251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.45779    0.08911  16.359  < 2e-16 ***
## treat        0.33063    0.09246   3.576 0.000418 ***
## pretest.c    0.30416    0.05322   5.715 3.06e-08 ***
## pretest.csq -0.08616    0.05868  -1.468 0.143272
## group_high   0.33243    0.09563   3.476 0.000598 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6231 on 254 degrees of freedom
## Multiple R-squared:  0.2538, Adjusted R-squared:  0.2421
## F-statistic:  21.6 on 4 and 254 DF,  p-value: 2.352e-15
```

```
# HA Confound model predicted values
csiw$confounder.check.model.ha <- predict(confounder.check.model.ha)

# HA Confound model residuals
csiw$confounder.check.residuals.ha <- residuals(confounder.check.model.ha)

# HA Confound model predicted values vs pretest confounds
par(mfrow=c(1,2))

ggplot(csiw, aes(x=pretest.c, y = confounder.check.model.ld, color=factor(treat))) + geom_point
()+
  geom_smooth(method="loess", color="darkred")+  xlab("Pretest Scores Centered") + ylab("Predict
ed values from Quadratic ANCOVA model") + labs(title="Predicted values vs Pretest squared - Conf
ounder Check model")
```




```
# Confound model residuals vs pretest confounds
ggplot(csiw, aes(x=pretest.c, y = confounder.check.residuals.ld, color=factor(treat))) + geom_po
int()+
  geom_smooth(method="loess", color="darkred")+  xlab("Pretest Scores Centered") + ylab("Residua
ls from Quadratic ANCOVA model") + labs(title="Residuals vs Pretest squared - Confounder Check m
odel")
```



The model adjusted to include high achieving status as a confound shows that CSIW (t=16.36, p<.001), pre-test (t=5.72, p<.001), and high achieving status (t=3.48, p<.001) are significant predictors of average writing post-test scores. A non-CSIW student with an average pre-test score who is not considered a high achieving student is estiamted to have an average post-test score of 1.46 (out of max 3). Participating in CSIW is predicted to add an average of .33 points, each unit increase in the pre-test score is predicted to add an average of .30 points, and being a high achiever subpopulation is also predicted to add .33 points on average.

F. HETEROGENEITY

F1. By Grade Level

Does the treatment effect depend on the grade level of the child? confounder check for grade dummy called grade_2, grade_2 = 1, if Grade = 5 and grade_2 = 0 if Grade = 4.

```
# Subset by grade

csiw4 <-  filter(csiw, csiw$grade_1 == 1)
csiw5 <- filter(csiw, csiw$grade_2 == 1)

model.grade4 <- lm(posttest ~ treat+pretest.c+pretest.csq, data=csiw4)
summary(model.grade4)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest.c + pretest.csq, data = csiw4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30976 -0.62414  0.09538  0.37586  1.69024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.34208    0.12671  10.592  < 2e-16 ***
## treat        0.42457    0.12828   3.310   0.0012 **
## pretest.c    0.36008    0.07346   4.902 2.72e-06 ***
## pretest.csq -0.02967    0.08495  -0.349   0.7275
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6557 on 133 degrees of freedom
## Multiple R-squared:  0.1958, Adjusted R-squared:  0.1777
## F-statistic:  10.8 on 3 and 133 DF,  p-value: 2.137e-06
```

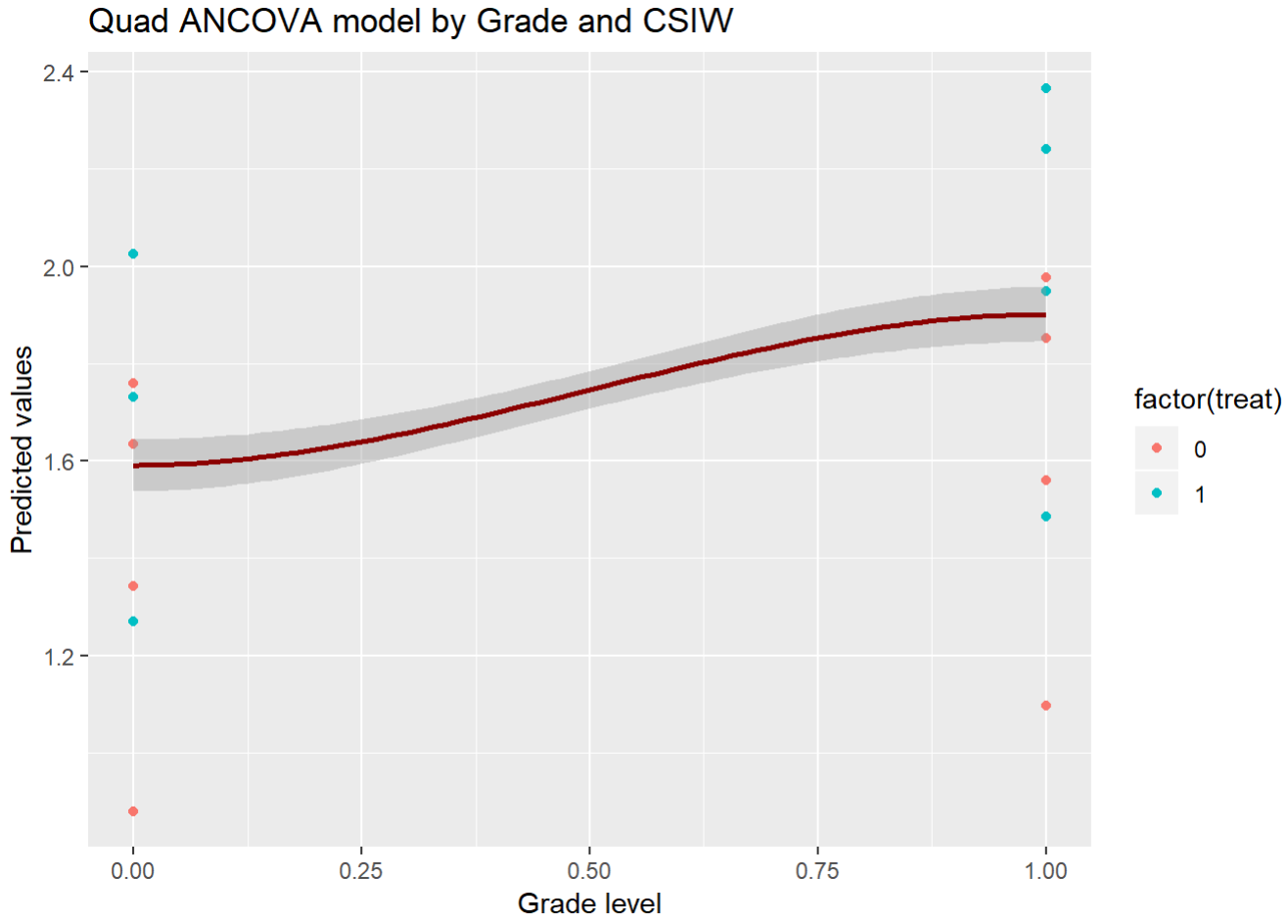
```
model.grade5 <- lm(posttest ~ treat+pretest.c+pretest.csq, data=csiw5)
summary(model.grade5)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest.c + pretest.csq, data = csiw5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43631 -0.23062  0.02192  0.38536  1.38536
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.67389    0.12814  13.063  < 2e-16 ***
## treat        0.36344    0.13015   2.793   0.0061 **
## pretest.c    0.35137    0.07054   4.981 2.18e-06 ***
## pretest.csq -0.14462    0.08376  -1.727   0.0869 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6019 on 118 degrees of freedom
## Multiple R-squared:  0.2202, Adjusted R-squared:  0.2003
## F-statistic:  11.1 on 3 and 118 DF,  p-value: 1.788e-06
```

```
model.grade <- lm(posttest ~ treat+pretest.c+pretest.csq+grade_2, data=csiw)
summary(model.grade)
```

```
##
## Call:
## lm(formula = posttest ~ treat + pretest.c + pretest.csq + grade_2,
##     data = csiw)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48640 -0.34186  0.05145  0.36461  1.73099
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.39956    0.09474  14.773  < 2e-16 ***
## treat        0.38931    0.09065   4.295 2.49e-05 ***
## pretest.c    0.35115    0.05065   6.933 3.38e-11 ***
## pretest.csq -0.08431    0.05918  -1.425  0.15550
## grade_2      0.21739    0.07915   2.747  0.00645 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6284 on 254 degrees of freedom
## Multiple R-squared:  0.2409, Adjusted R-squared:  0.2289
## F-statistic: 20.15 on 4 and 254 DF,  p-value: 1.991e-14
```

```
csiw$ancova.grade.predicted <- predict(model.grade)
csiw$ancova.grade.residuals <- residuals(model.grade)

ggplot(csiw, aes(x=grade_2, y=ancova.grade.predicted, color=factor(treat))) +geom_point()+geom_s
mooth(method="loess", color="darkred")+
  xlab("Grade level") + ylab("Predicted values") + labs(title="Quad ANCOVA model by Grade and CS
IW")
```



According to the above regressions tables that show the treatment coeffecient within grade 4 students (b=.42, p<.001) and within grade 5 students (b=.39, p<.001), participating in CSIW has a significant effect on average post-test scores in both grades. A regression model that includes grade as a covariate, it is a significant predictor of post-test scores (b=.22, p<0.0001), where 5th grade students have an average estimated increase of .22 on post-test scores compared to 4th grade students. The graph of the predicted values by CSIW participation shows a small slope increase from grade 4 to grade 5 where the gap in post-test performance by CSIW is larger for 5th graders.

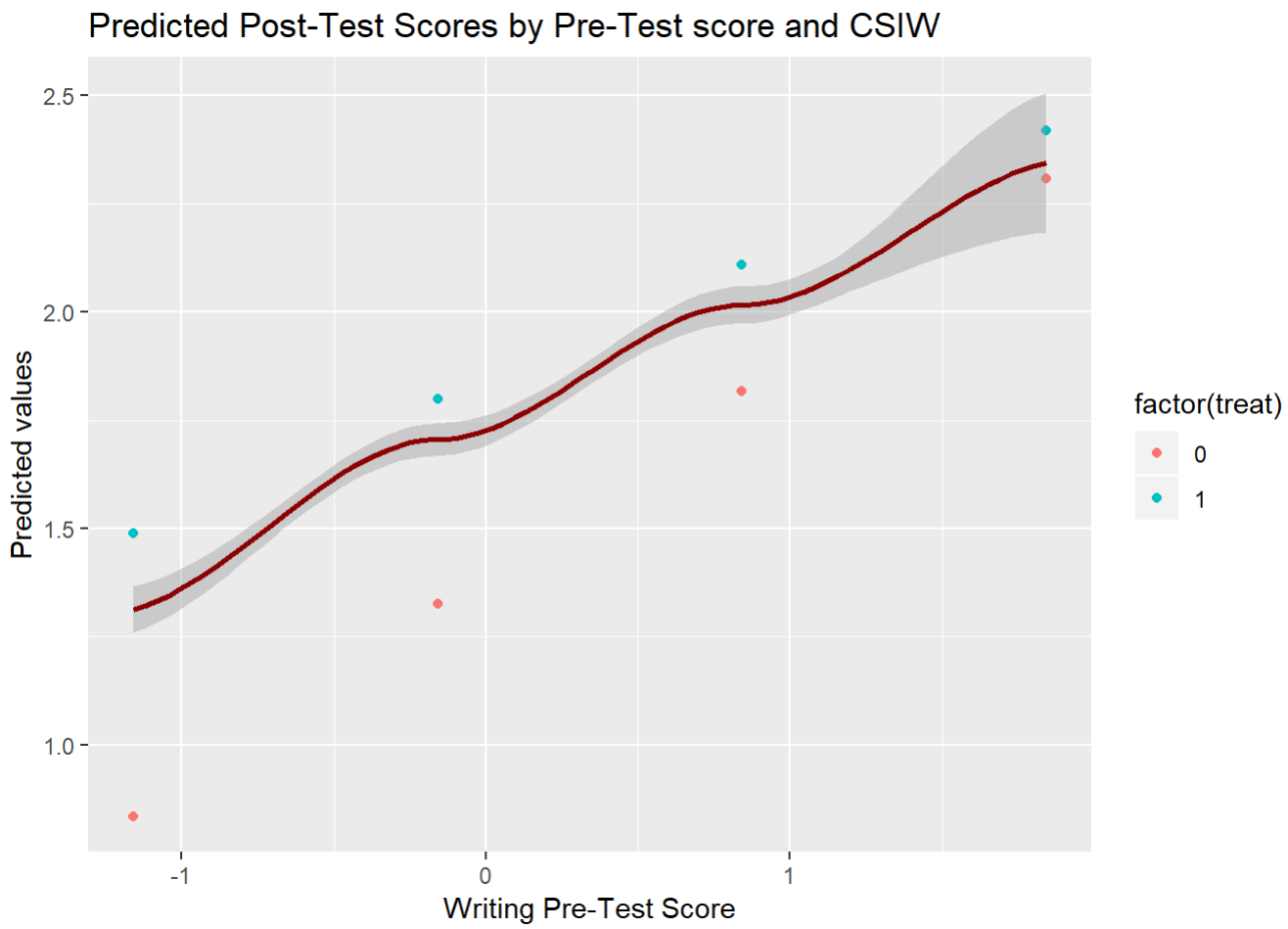
F2. By Pre-test

> Does the treatment effect depend on the child's prior achievement level? Provide statistical evidence.

```
model.pretest <- lm(posttest ~ treat*pretest.c, data=csiw)
summary(model.pretest)
```

```
##
## Call:
## lm(formula = posttest ~ treat * pretest.c, data = csiw)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4877 -0.4877  0.1648  0.2017  1.6738
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.40391    0.07796  18.008  < 2e-16 ***
## treat            0.44357    0.09055   4.898 1.72e-06 ***
## pretest.c        0.49097    0.08800   5.579 6.16e-08 ***
## treat:pretest.c -0.18036    0.10770  -1.675   0.0952 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6353 on 255 degrees of freedom
## Multiple R-squared:  0.2212, Adjusted R-squared:  0.212
## F-statistic: 24.14 on 3 and 255 DF,  p-value: 8.744e-14
```

```
csiw$ancova.pretest.predicted <- predict(model.pretest)

ggplot(csiw, aes(x=pretest.c, y=ancova.pretest.predicted, color=factor(treat))) +geom_point()+ge
om_smooth(method="loess", color="darkred")+
  xlab("Writing Pre-Test Score") + ylab("Predicted values") + labs(title="Predicted Post-Test Sc
ores by Pre-Test score and CSIW")
```

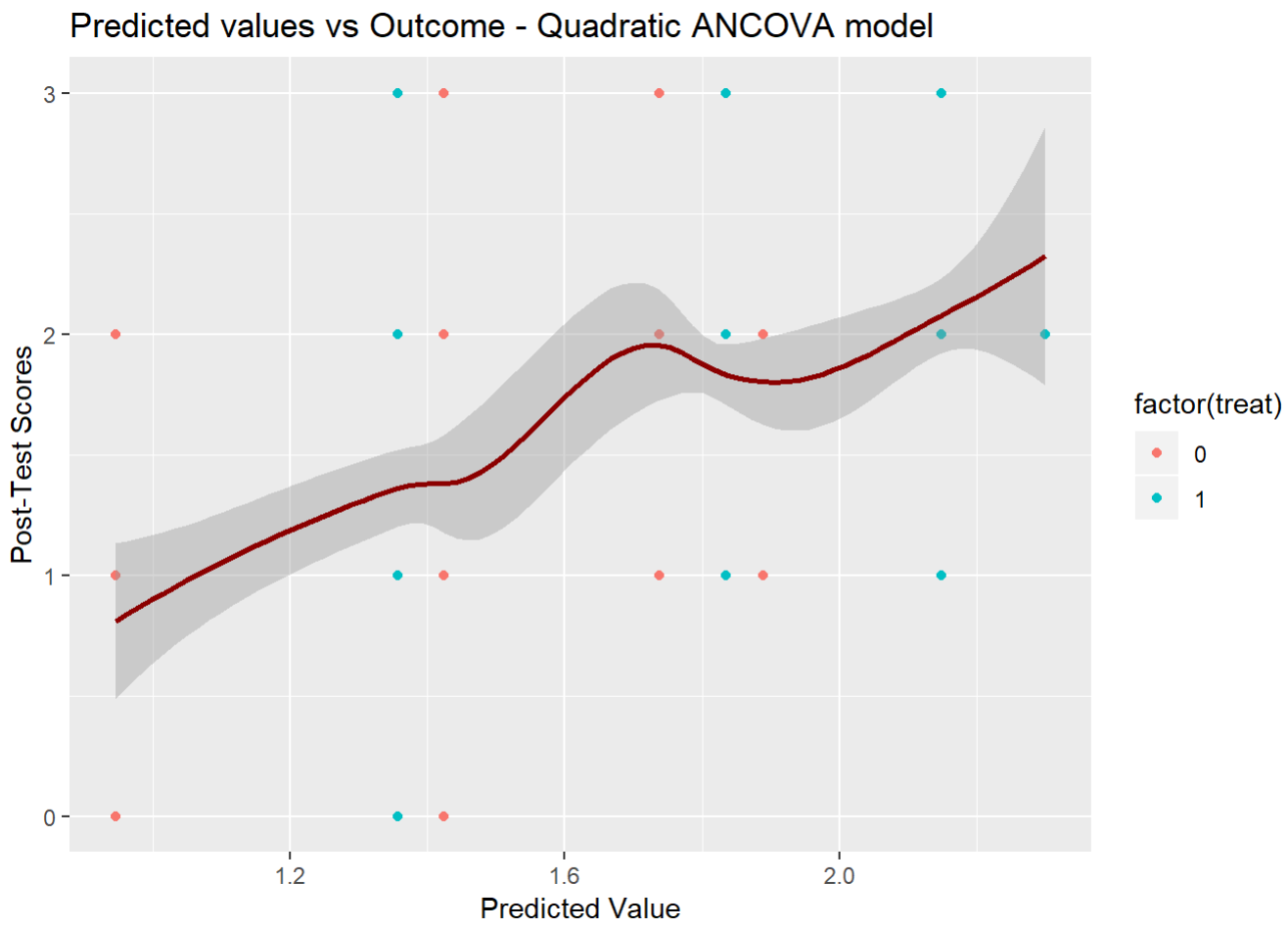


In the above regression model, both CSIW particiation (b=0.44, p<.001) and pretest scores (b=.49, p<.001) are significant predictors of post-test scores.The plot of predicted post-test scores by pre-test scores and CSIW participation show that there is a larger gap in post-test scores between CSIW and non-CSIW students among those who performed worse on the writing pre-test, compared to those who scored higher.

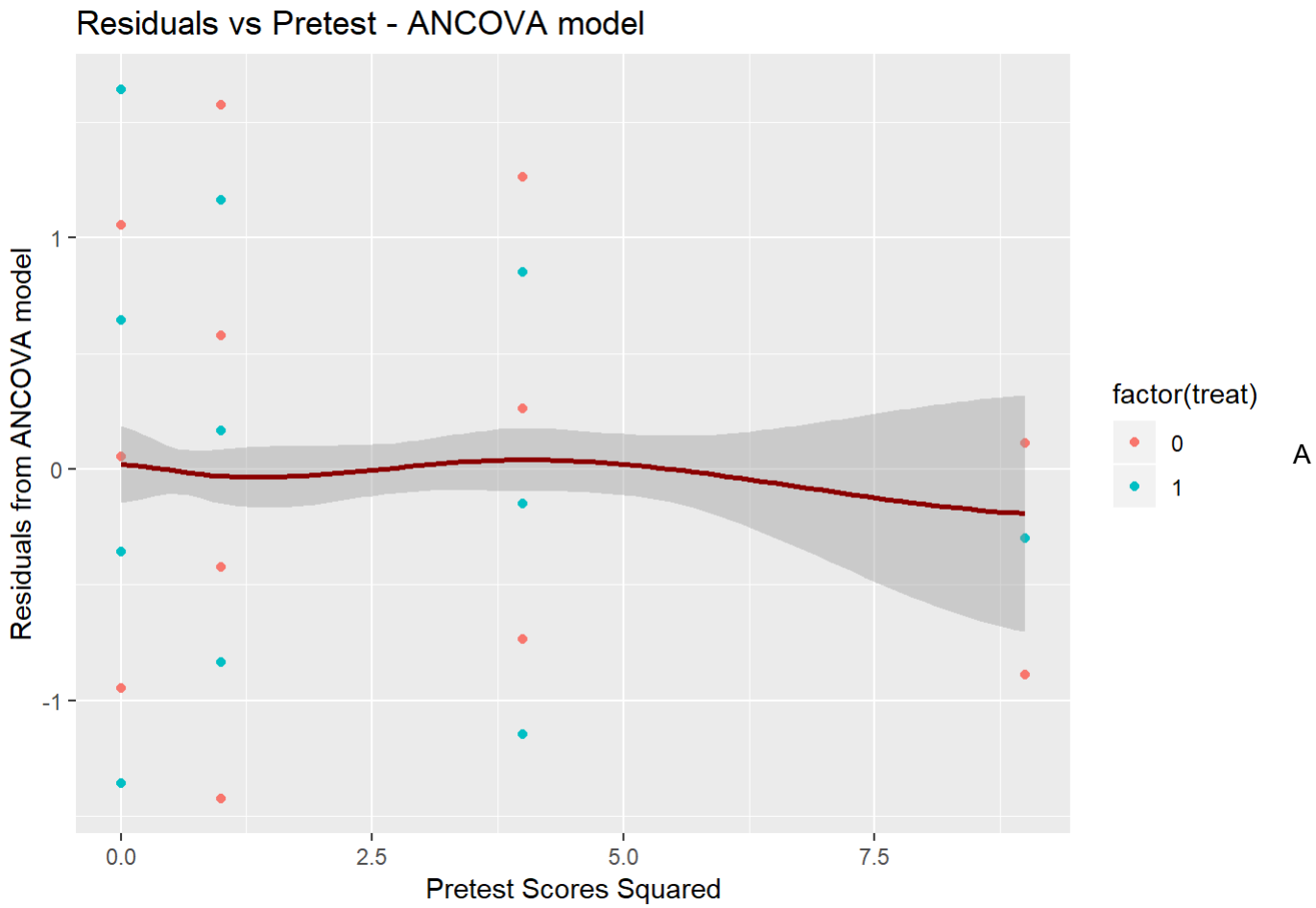
G. CHECKING ASSUMPTIONS ON RANDOM ERROR

Use a Loess Line and a Quadratic fit to check linearity of quadratic ancova model.

```
csiw$pretest.sq <- csiw$pretest^2

csiw$quadratic.ancova.predicted <- predict(quadratic.ancova.model)
csiw$quadratic.ancova.residuals <- residuals(quadratic.ancova.model)

# Linearity
ggplot(csiw, aes(x=csiw$quadratic.ancova.predicted, y = csiw$posttest, color=factor(treat))) + g
eom_point()+
  geom_smooth(method="loess", color="darkred")+  xlab("Predicted Value") + ylab("Post-Test Score
s") + labs(title="Predicted values vs Outcome - Quadratic ANCOVA model")
```




```
## Scatter plot
# Make a scatter plot in which the vertical axis has the residuals from the ANCOVA model and the
horizontal axis is the covariate.
# Use a Loess Line and a Quadratic fit to check linearity.

# plot quadratic ancova residuals against pre-test covariate w/loess line
ggplot(csiw, aes(x=pretest.sq, y=quadratic.ancova.residuals, color=factor(treat))) +
  geom_point()+geom_smooth(method="loess", color="darkred") +
  xlab("Pretest Scores Squared") + ylab("Residuals from ANCOVA model") +
  labs(title="Residuals vs Pretest - ANCOVA model")
```



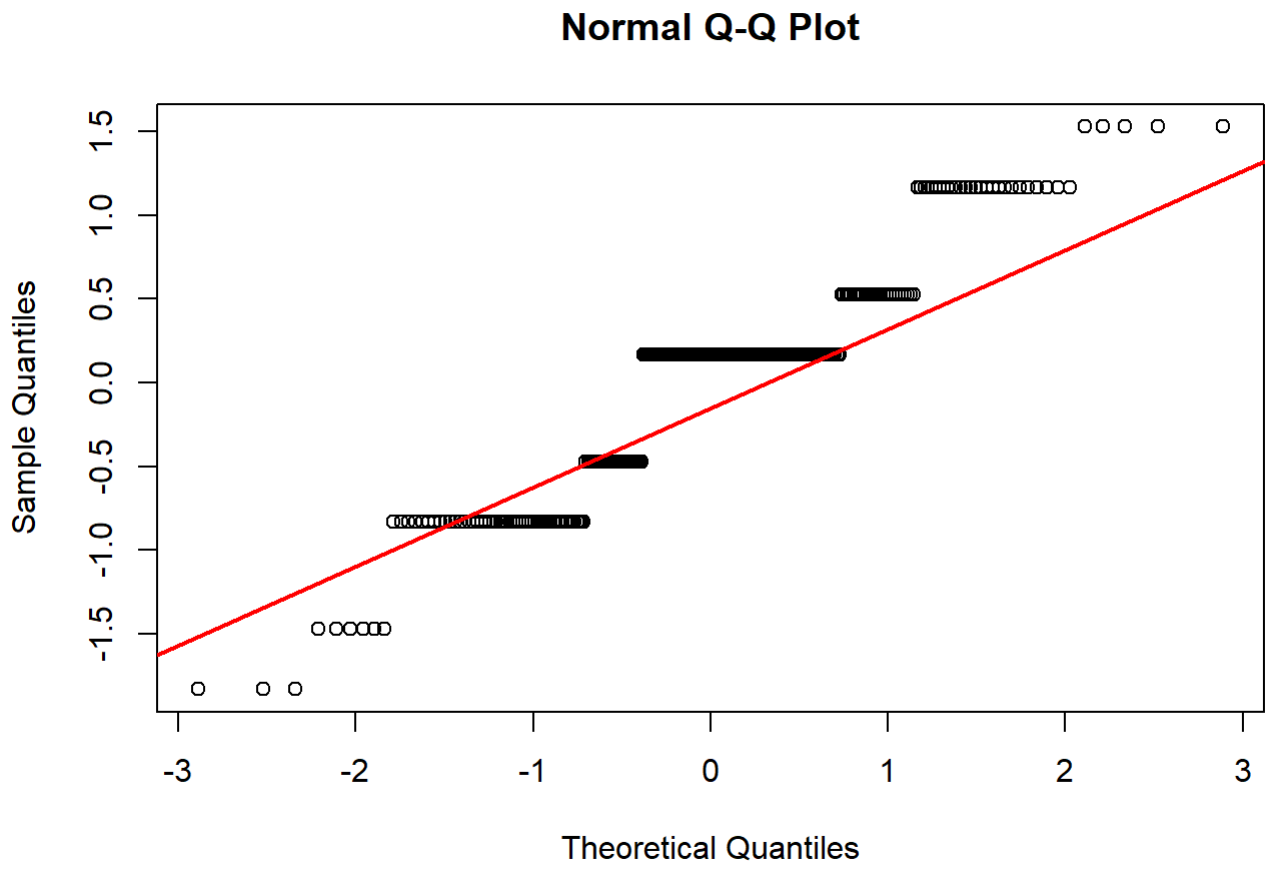
plot of the model's predicted value agains the pre-test scores shows that the model violates the linearity assumption because it has about 3 kinks across the predicted values. The data fit the line much better in this model, which is seen in the narrow standard errors between the predicted values and pretest scores but it is not straight.

The plot of the residuals against the pre-test covariate have a more even distribution of residual errors on the top and bottom sides of the loess line, but it is still slightly curved.

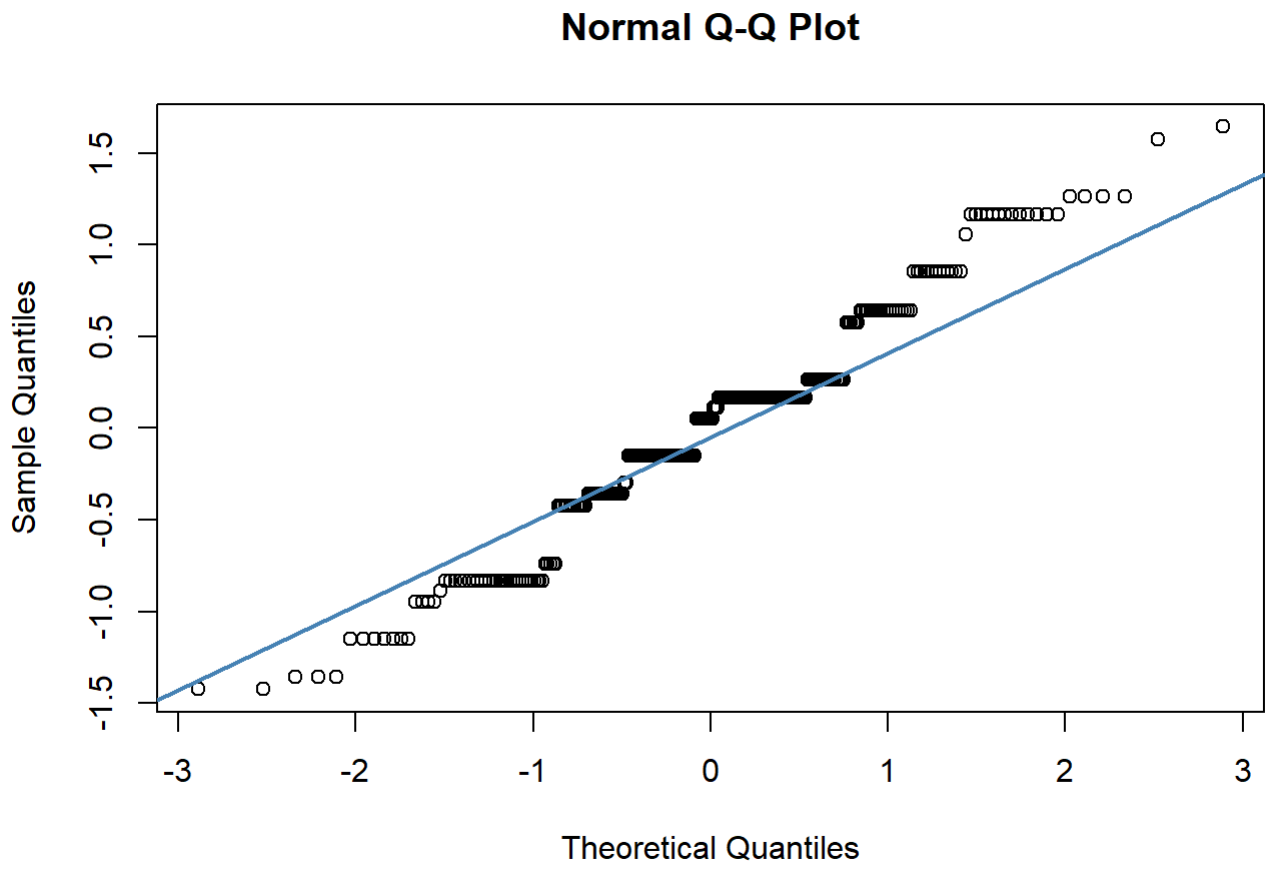
G1. Normality of Residuals

> Use a plot to check the normality of the residuals and explain

```
# Naive
qqnorm(csiw$naive.residuals, pch = 1)
qqline(csiw$naive.residuals, col = "red", lwd = 2)
```




```
# Quadratic
qqnorm(csiw$quadratic.ancova.residuals, pch = 1)
qqline(csiw$quadratic.ancova.residuals, col = "steelblue", lwd = 2)
```

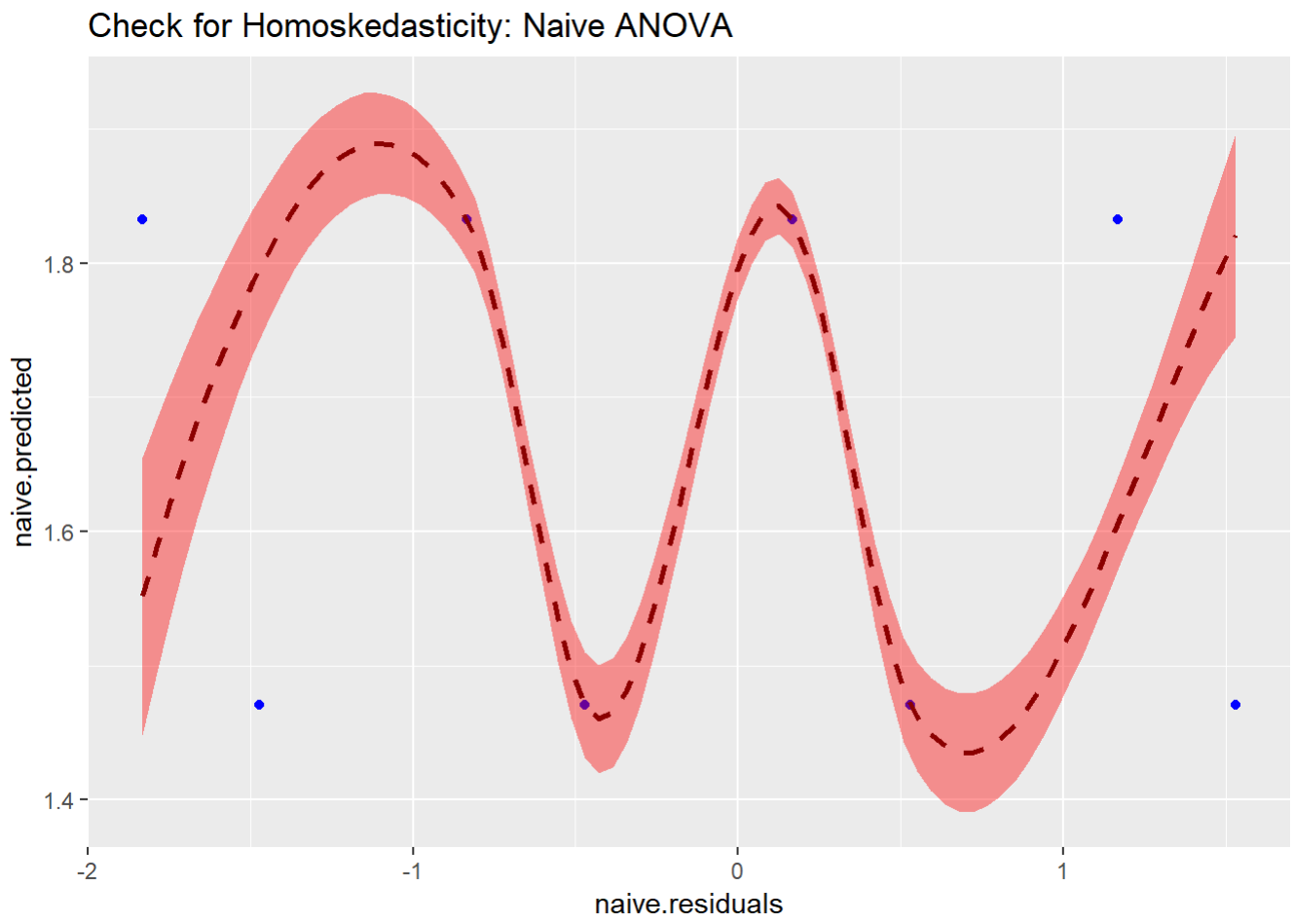


A Q-Q plot of the model residuals shows that both the naive and quadratic models don't converge around a straight line which is cause for invalidating the normality assumption, although the quadratic model residual seem to fit tighter.

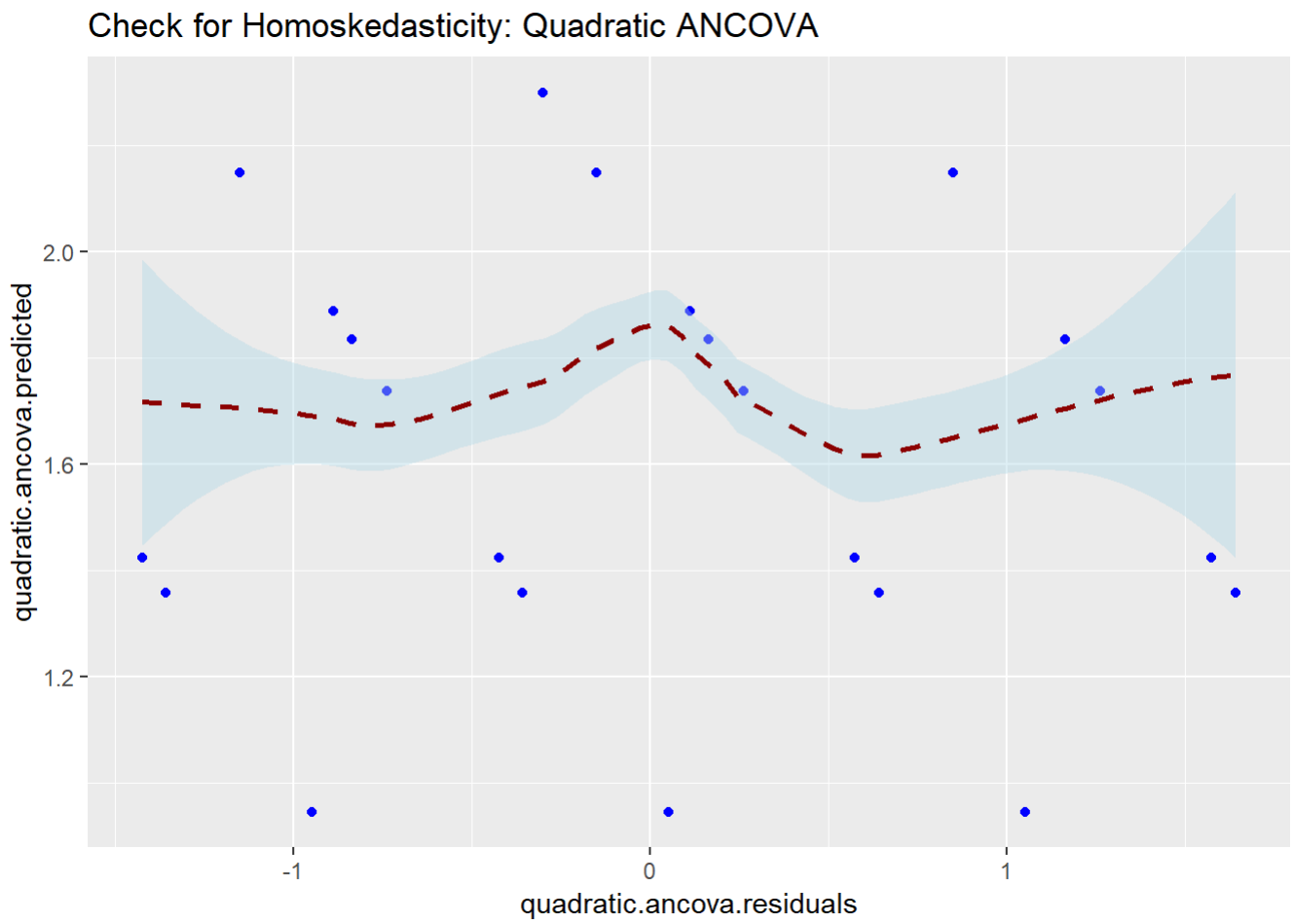
G2. Homoskedasticity

> Use a plot to check homoscedasticity and explain

```
ggplot(csiw, aes(x=naive.residuals, y=naive.predicted)) + geom_point(color="blue") +
  labs(title="Check for Homoskedasticity: Naive ANOVA") +
  geom_smooth(method="loess", linetype="dashed",color="darkred", fill="red")
```




```
ggplot(csiw, aes(x=quadratic.ancova.residuals, y=quadratic.ancova.predicted)) + geom_point(color
="blue") +
  labs(title="Check for Homoskedasticity: Quadratic ANCOVA") +
  geom_smooth(method="loess", linetype="dashed",color="darkred", fill="lightblue")
```



The homoskedasticity assumption doesn't hold for the quadratic ANCOVA model as the residuals spread more at the tail ends of the x-axis, similarly to the shape of the naive model's residuals against predicted values but with less dramatic curves.

H. CONCLUSION

H1. Final Estimate

> What is your best estimate of the impact of CSIW on writing achievement (provide a confidence interval).

```
# Summary
export_summs(naive.model, ancova.model, quadratic.ancova.model, confounder.check.model.ld, confo
under.check.model.ha, confint = TRUE, model.names = c("Naive", "ANCOVA", "Quad ANCOVA", "Q-ANCOV
A-LD", "Q-ANCOVA-HA"))
```

	Naive	ANCOVA	Quad ANCOVA	Q-ANCOVA-LD	Q-ANCOVA-HA
(Intercept)	1.47 ***	0.99 ***	1.48 ***	1.72 ***	1.46 ***
	(0.08)	(0.10)	(0.09)	(0.09)	(0.09)

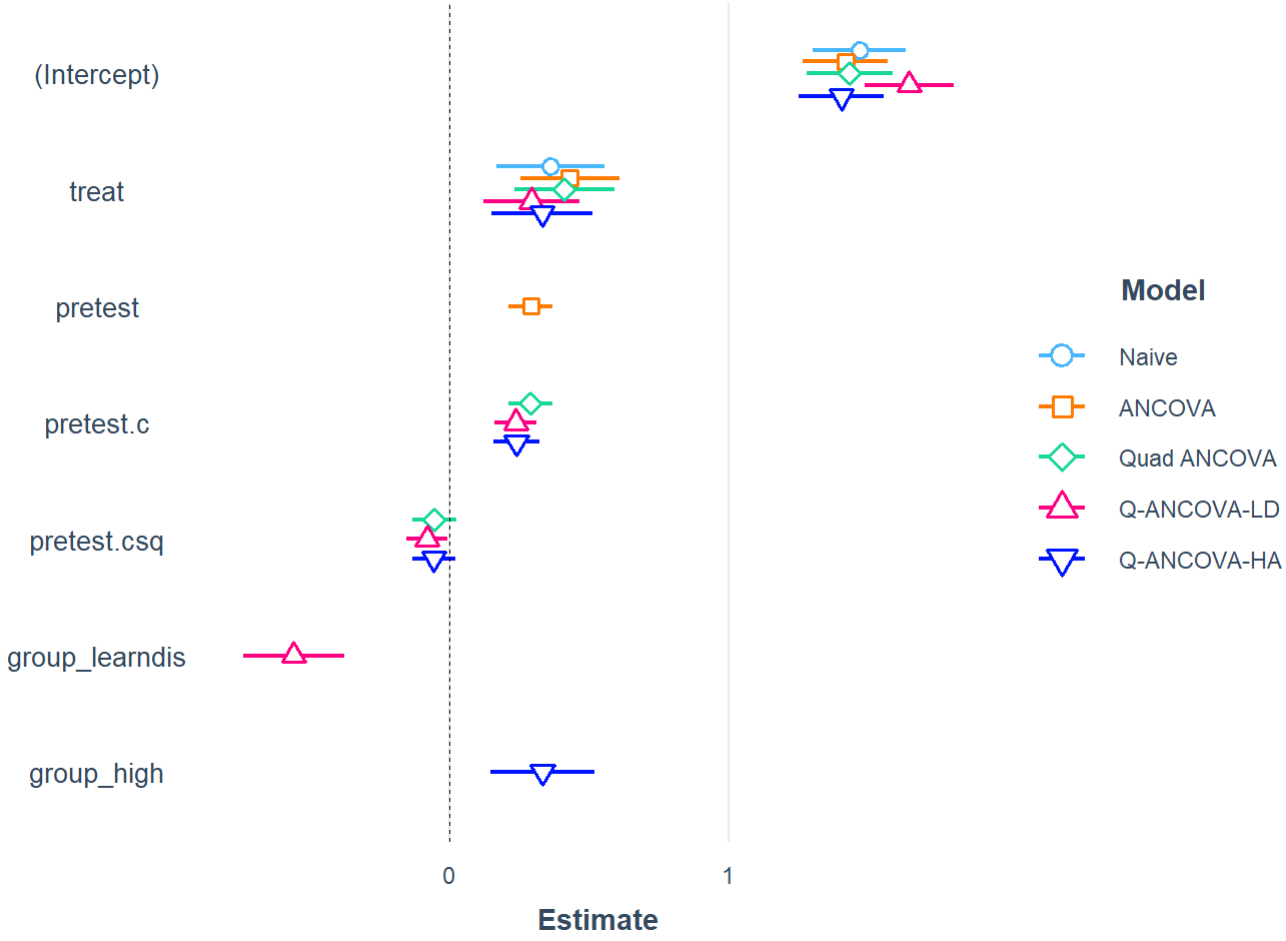
treat	0.36 ***	0.43 ***	0.41 ***	0.29 ***	0.33 ***
	(0.10)	(0.09)	(0.09)	(0.09)	(0.09)
pretest		0.37 ***			
		(0.05)			
pretest.c			0.37 ***	0.30 ***	0.30 ***
			(0.05)	(0.05)	(0.05)
pretest.csq			-0.08	-0.12 *	-0.09
			(0.06)	(0.06)	(0.06)
group_learndis				-0.56 ***	
				(0.09)	
group_high					0.33 ***
					(0.10)
N	259	259	259	259	259
R2	0.05	0.21	0.22	0.32	0.25

*** p < 0.001; ** p < 0.01; * p < 0.05.

```
# Confidence intervanls
confint(confounder.check.model.ld, level = 0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept)     1.5390491  1.90842572
## treat           0.1205175  0.46667451
## pretest.c       0.2054619  0.39824138
## pretest.csq    -0.2342309 -0.01141429
## group_learndis -0.7431988 -0.37735251
```

```
plot_summs(naive.model, ancova.model, quadratic.ancova.model, confounder.check.model.ld, confoun
der.check.model.ha, scale = TRUE, model.names = c("Naive", "ANCOVA", "Quad ANCOVA", "Q-ANCOVA-L
D", "Q-ANCOVA-HA"), omit.coefs= FALSE)
```



The first table above compares the estimates and model fit for the naive ANOVA model, ANCOVA, Quadratic ANCOVA, and Quadratic ANCOVA controlling for learning disability and high achievement status. The R^2 estimate shows how much of the variation the model is able to explain, or how well the model fits the sample data. The highest R^2 is in the quadratic ANCOVA model that includes learning disability status as a counfound (R^2=.32) and the naive model explains the least variance in the post-test scores (R^2=.05).

Under the quadratic ANCOVA model that controls for learning disability status, the beta coefficient for CSIW participation is b=.29 and there is 95% confidence that a student randomly selected from the population will get an estimated CSIW coefficient between 0.12 and 0.46.

The next figure is a plot of the coefficient estimates with 95% confidence intervals across each models so we can visually compare each variable. We can see that particiating in CSIW and having a learning disability are the larger absolute coefficients that predict a bigger increase or decrease on average relative to not participating or not having a learning disability.

H2. Assumptions

> Under what assumptions is this a valid estimate of the causal effect?

The predicted estimates from the repored models depends on passing assumptions about the models' residual error. A randomized controlled trial in which students were randomly assigned to participate in the CSIW program would help make a valid causal inference from these estimates because it removes potential selection bias from

unobserved variables. We tested achievement level and grade level as potential confounds but there may be other unobserved variables that are predictive of students' writing post-test scores and participation in CSIW.

1) Structural: Linearity

To satisfy linearity, the relationship between our outcome of interest, writing post-test scores, and the predictor variables, CSIW, pre-test scores, and achievement level must each have a linear relationship, respectively.

2) Random: Normal residuals

The residual errors from the model must be normally distributed.

3) Random: Homogeneous residual variance

The variance in the model's residual errors must be constant across all levels of the predictor variables.

4) Random: Independent residuals

The model's residual error for a student must be must be independent from the residual errors of any other student. If the residual errors are shown to be correlated with each other, then there is a related unobserved factor in the error term that could bias the estimate.