# Hw 3

*Ben Chu*

*April 1, 2018*

Loading some data,packages, and function.

```
load("C:/Users/Branly Mclanbry/Downloads/GSS1991HW.RData")
load("C:/Users/Branly Mclanbry/Downloads/TOP2003.RData")
hw1 <- GSS1991HW %>%
  janitor::clean_names()
options(contrasts = c('contr.helmert', 'contr.poly'))
pphehe <- function(x,var) {
 (qqnorm(x, main = var))
 (qqline(x))
}
denss <- function(x,var) {
  plot(density(x), main = var)
}

skurt.1 <-function(x,var) {
  print(var)
  print(round(DescTools::Kurt(x, method = 2, conf.level = .99, R = 2000),2))
  print(round(DescTools::Skew(x, method = 2, conf.level = .99, R = 2000),2))
}

transformer <- function(x,var){
  squareroot <- (x+1)^.5
  inverse <- 1/(x+1)
  log <- log10(x+1)
  print(var)
  print("squareroot")
  print(round(DescTools::Skew(squareroot,na.rm=TRUE, method=2,conf.level=.99, R=2000),2))
  print(round(DescTools::Kurt(squareroot,na.rm=TRUE, method=2,conf.level=.99, R=2000),2))
  print("log")
  print(round(DescTools::Skew(log,na.rm=TRUE, method=2,conf.level=.99, R = 2000),2))
  print(round(DescTools::Kurt(log,na.rm=TRUE, method=2,conf.level=.99, R = 2000),2))
  print("inverse")
  print(round(DescTools::Skew(inverse,na.rm=TRUE, method=2,conf.level=.99, R = 2000),2))
  print(round(DescTools::Kurt(inverse,na.rm=TRUE, method=2,conf.level=.99, R = 2000),2))
}

p_list <- list(hw1$educ,hw1$maeduc,hw1$prestg80)
p_names <- names(hw1[2:4])
```

# HW1

Sex and education both significantly predict prestige $R^2$ = .26, $F$(3,1158) = 134.32, $p$ < .001. However, not all variables contributed equally, education was a significant predictor ($b^*$ = .50, $p$ <.001) while sex was not ($b^*$ = -.11, $p$ = .73), and neither was the interaction ($b^*$ = .02, $p$ = .06)

Essentially, higher education predicts a higher occupational prestige, while sex and the interaction between sex and eduction do not.

```
gss.dat <- lm(prestg80~sex*educ, hw1)
summ(gss.dat,center = TRUE, digits = 5, confint = TRUE)
```

```
## MODEL INFO:
## Observations: 1162
## Dependent Variable: prestg80
##
## MODEL FIT:
## F(3,1158) = 134.3184, p = 0
## R-squared = 0.25815
## Adj. R-squared = 0.25622
##
## Standard errors: OLS
##                 Est.     2.5%    97.5%   t val.       p
## (Intercept) 26.80581 25.80841 27.80322 52.67521 0       ***
## sex         -0.23471 -1.55352  1.0841  -0.34882 0.72729
## educ         2.14216  1.8066   2.47771 12.51221 0       ***
## sex:educ     0.45576 -0.01272  0.92423  1.90676 0.0568    .
##
## All continuous predictors are mean-centered.
```

```
lm.beta(gss.dat)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling
var(x) on a factor x is deprecated and will become an error.
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.
```

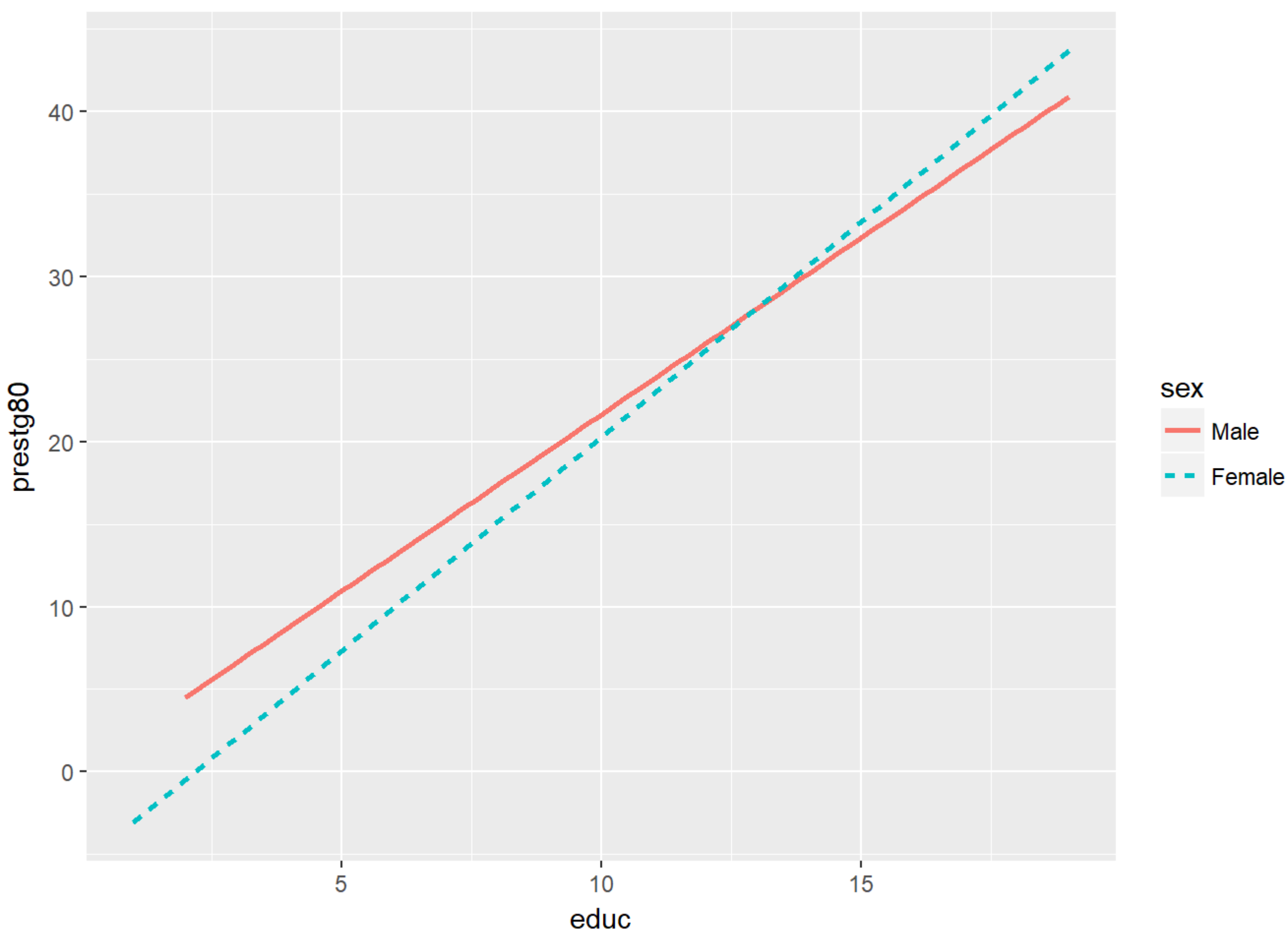
```
## Warning in b * sx: longer object length is not a multiple of shorter object
## length
```

```
##      sexFemale           educ sexFemale:educ
##    -0.22235158     0.45534725     0.01720257
```

```
modelEffectSizes(gss.dat)
```

```
## lm(formula = prestg80 ~ sex * educ, data = hw1)
##
## Coefficients
##                      SSR df pEta-sqr dR-sqr
## (Intercept)       1.2544  1   0.0000     NA
## sex             477.1217  1   0.0032 0.0024
## educ          20033.2177  1   0.1191 0.1003
## sex:educ        465.2390  1   0.0031 0.0023
##
## Sum of squared errors (SSE): 148180.5
## Sum of squared total  (SST): 199743.6
```

```
ggplot(gss.dat, aes(x = educ, prestg80, color = sex)) +
  geom_smooth(aes(linetype = sex), method = 'lm', se = F)
```

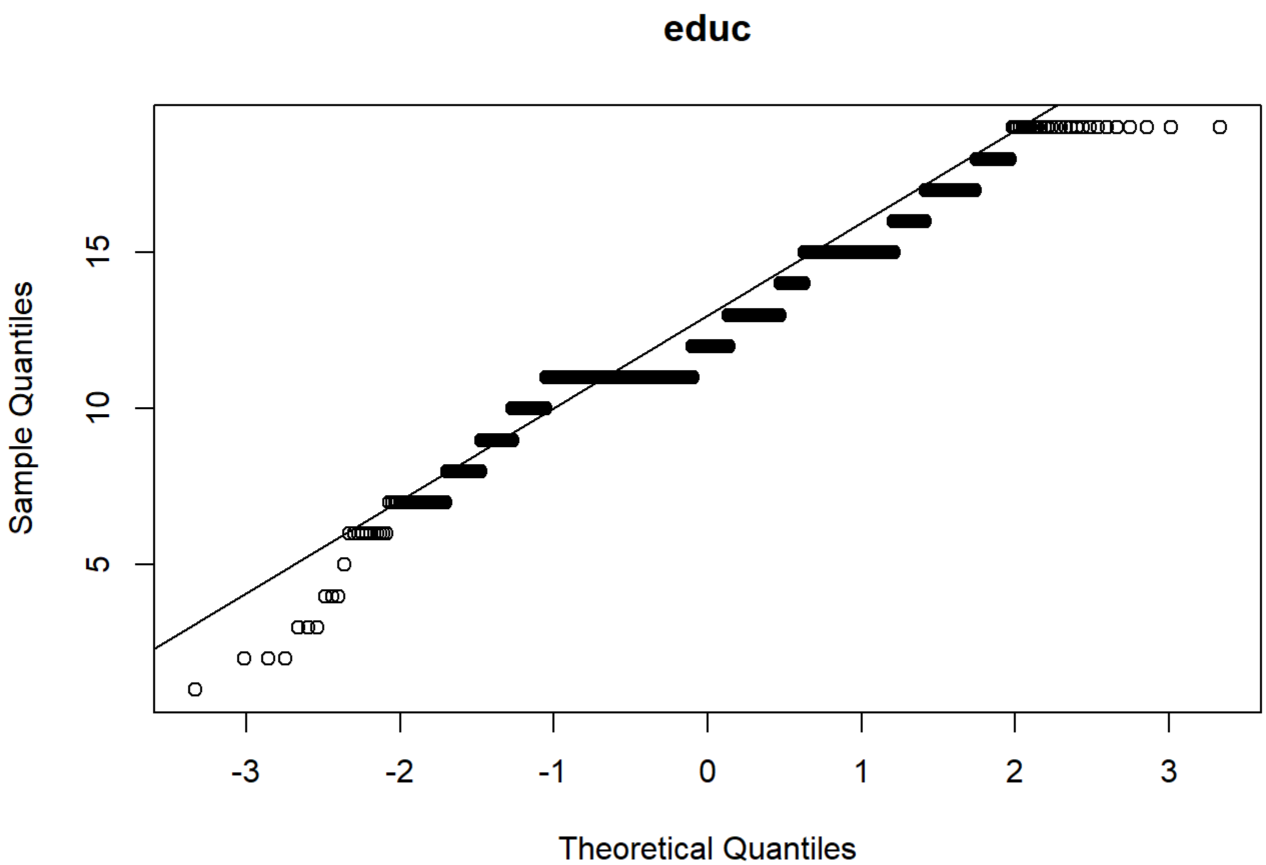



# HW2

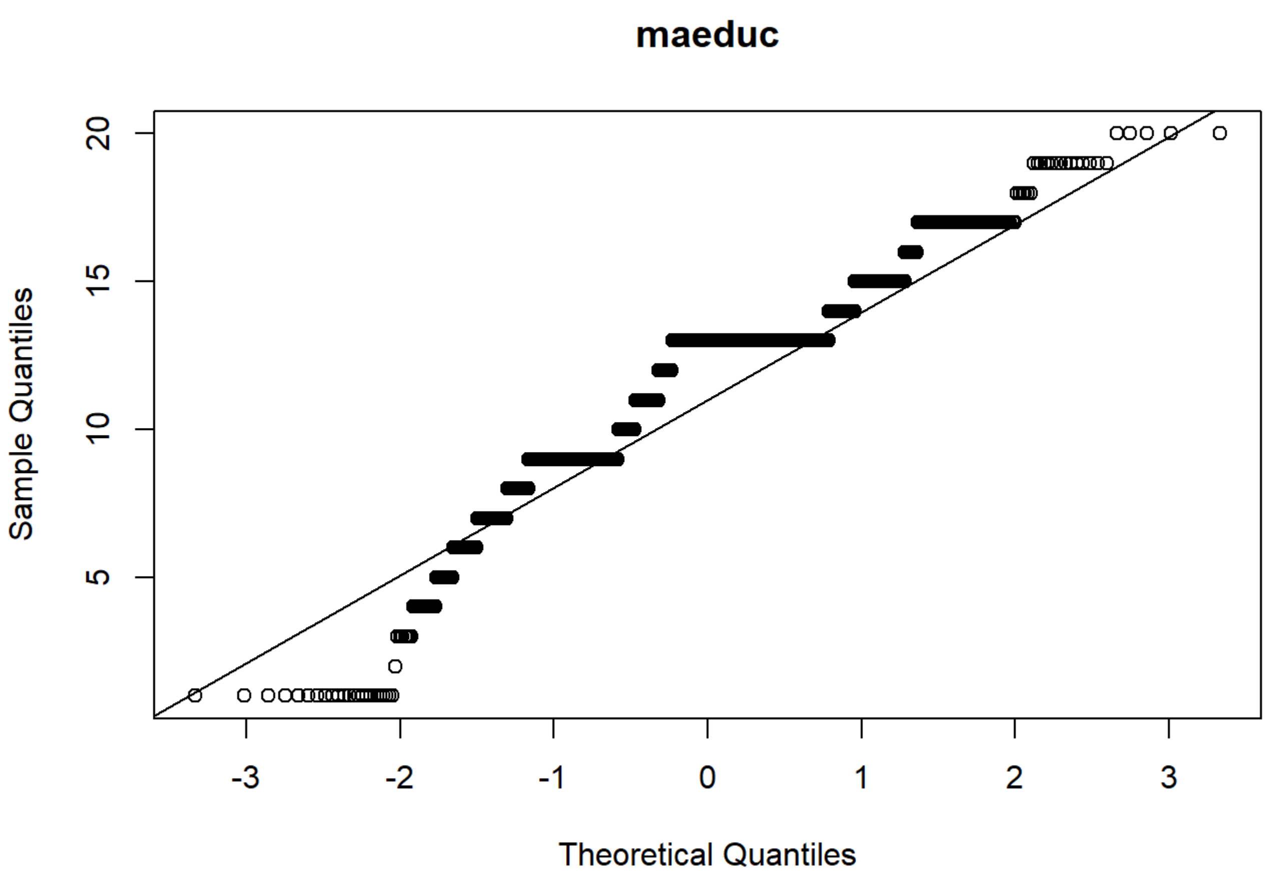
Mothers education and occupational prestige were transformed due to a violation in normality. Specifically a reflected square root for mother's education and a square root transformtion for prestige. Mothers education and education significantly predice prestige $R^2$ = .23, $F$(3,1158) = 115, $p$ < .001. Education level($b^*$ = .51, $p$ <.001) and mother's education ($b^*$ = .07, $p$ < . 01) significantly predicted occupational prestige. It is important to mention that mother's education was reflected so interpretation should be reversed.

The interaction was also significant ($b^*$ = .02, $p$ < .001) and the slope for mother's education one standard deviation above the mean (*b* = 0.22, *p* < .001), while at the mean (*b* = .25,*p* < .001) and one standard deviation below the mean (*b* = .27, p<.001). Specifically that lower education provides a lower amount of prestige compared to higher education, which provides a larger amount of prestive.

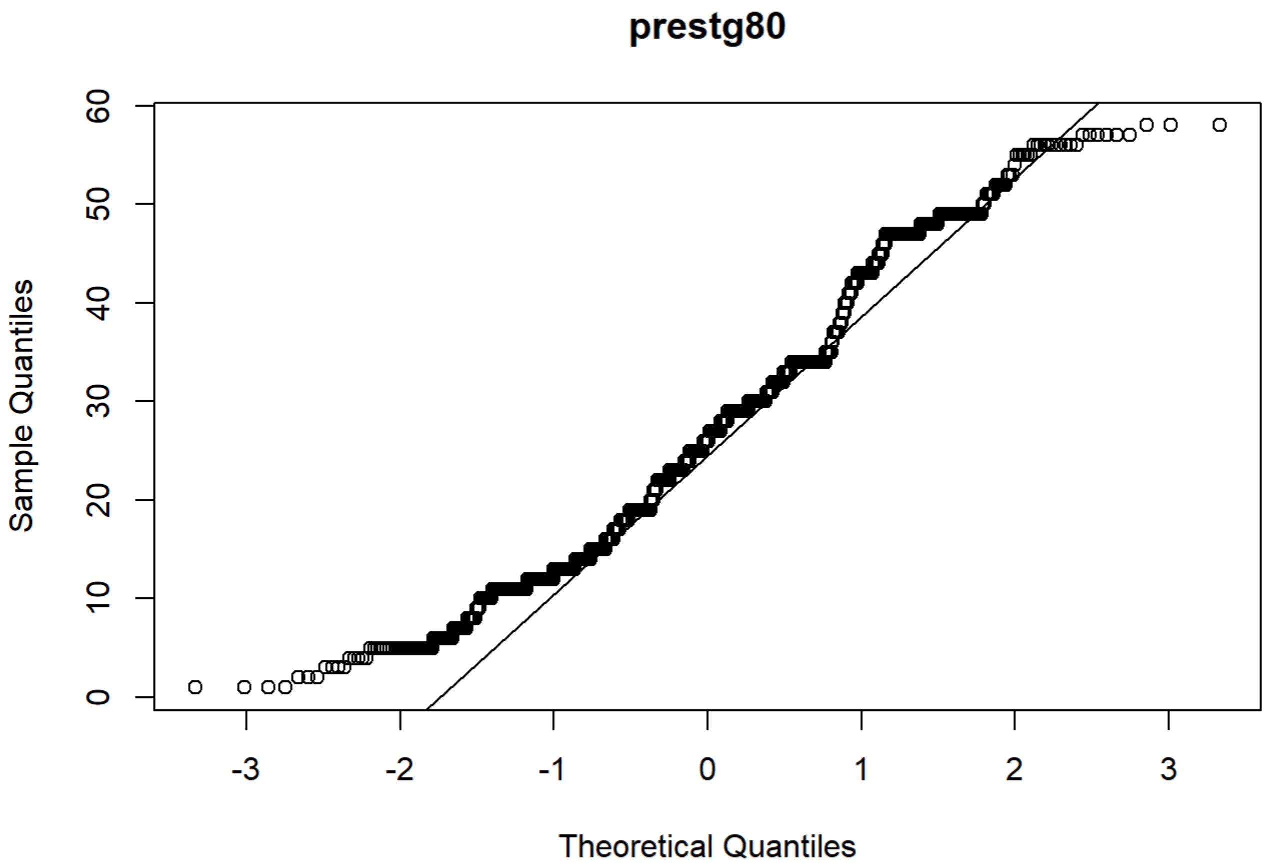
```
walk2(p_list,p_names,pphehe)
```



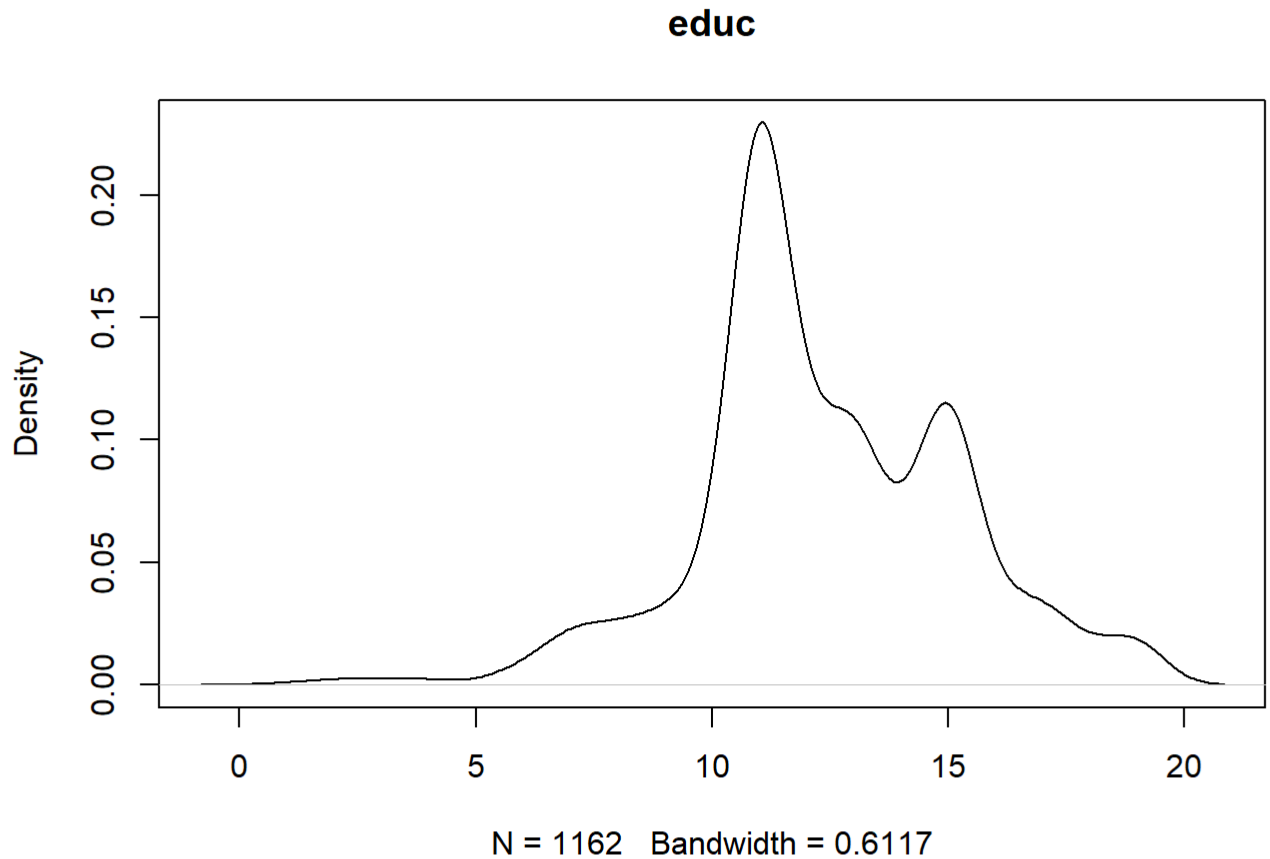
educ
Sample Quantiles
Theoretical Quantiles
5
10
15
-3
-2
-1
0
1
2
3

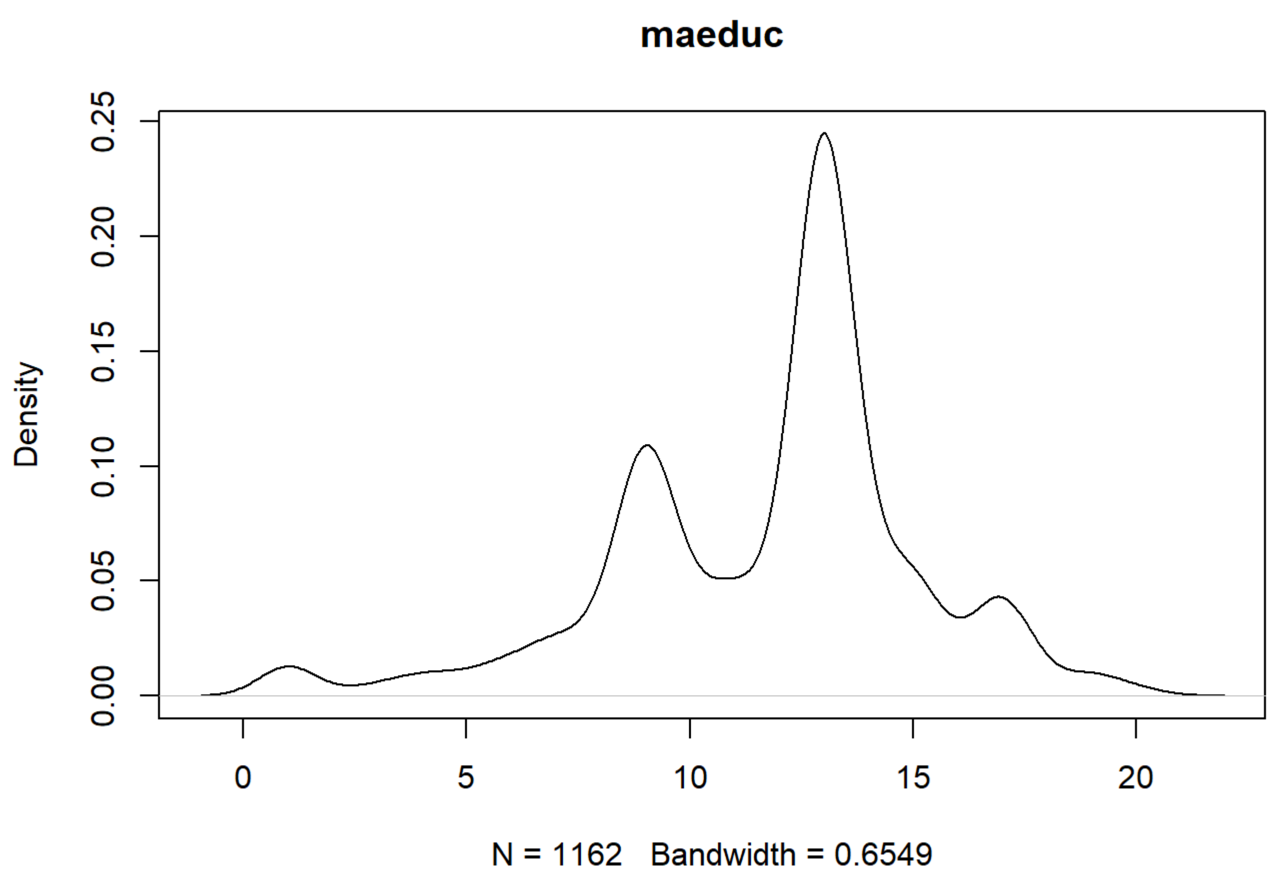



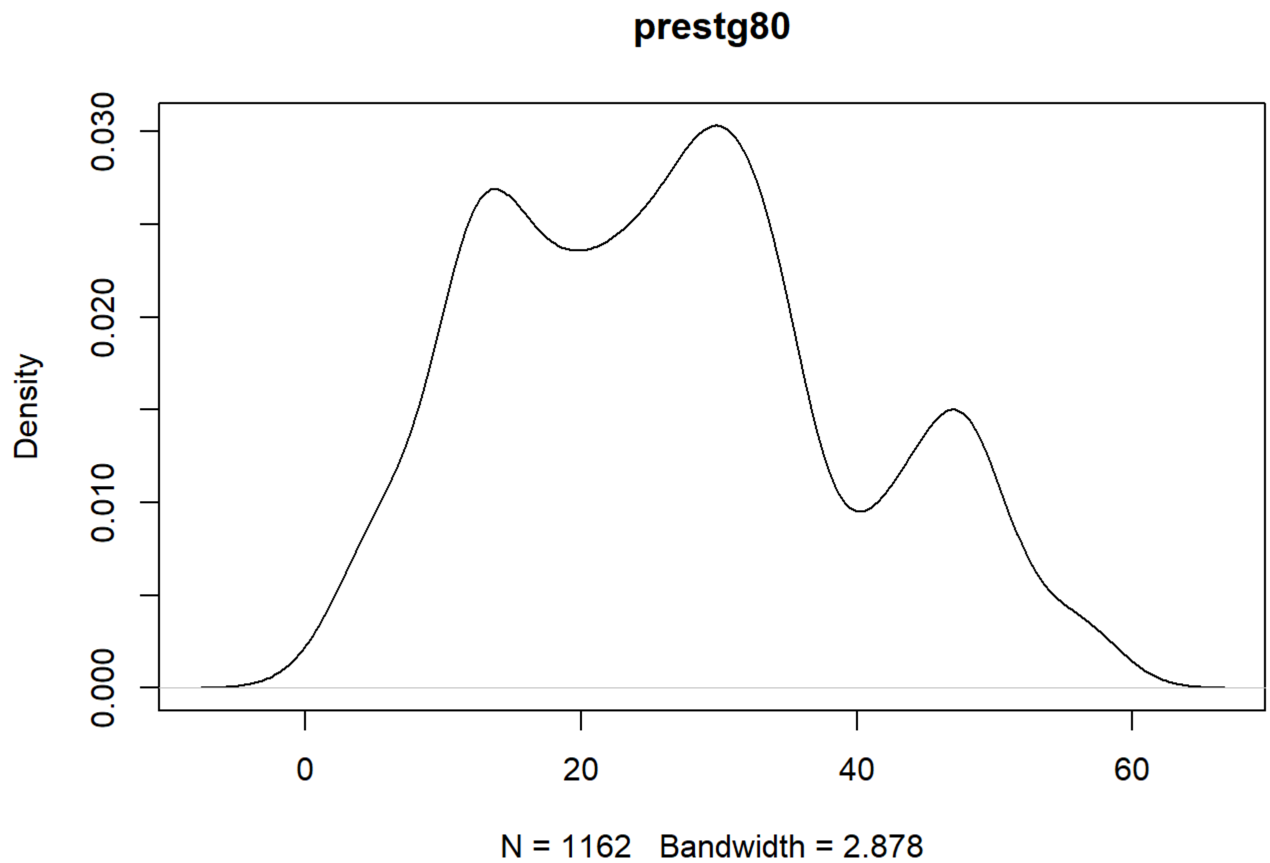
maeduc
Sample Quantiles
Theoretical Quantiles
5
10
15
20
-3
-2
-1
0
1
2
3




```
walk2(p_list,p_names,denss)
```


educ
Density
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
N = 1162   Bandwidth = 0.6117


maeduc
Density
0.00
0.05
0.10
0.15
0.20
0.25
0
5
10
15
20
N = 1162   Bandwidth = 0.6549




```
walk2(p_list,p_names,skurt.1)
```

```
## [1] "educ"
##   kurt lwr.ci upr.ci
##   0.77   0.23   1.54
##   skew lwr.ci upr.ci
##  -0.09  -0.37   0.15
## [1] "maeduc"
##   kurt lwr.ci upr.ci
##   1.05   0.65   1.59
##   skew lwr.ci upr.ci
##  -0.72  -0.92  -0.52
## [1] "prestg80"
##   kurt lwr.ci upr.ci
##  -0.66  -0.83  -0.49
##   skew lwr.ci upr.ci
##   0.33   0.23   0.43
```

```
walk2(p_list,p_names,transformer)
```

```
## [1] "educ"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##  -0.72  -1.26  -0.30
##   kurt lwr.ci upr.ci
##   2.68   1.23   4.77
## [1] "log"
##   skew lwr.ci upr.ci
##  -1.75  -2.70  -1.04
##   kurt lwr.ci upr.ci
##   8.84   4.06  15.18
## [1] "inverse"
##   skew lwr.ci upr.ci
##   6.81   3.82   9.24
##   kurt lwr.ci upr.ci
##  78.87  30.42 151.92
## [1] "maeduc"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##  -1.51  -1.73  -1.27
##   kurt lwr.ci upr.ci
##   3.61   2.79   4.64
## [1] "log"
##   skew lwr.ci upr.ci
##  -2.62  -2.88  -2.34
##   kurt lwr.ci upr.ci
##   9.33   7.61  11.45
## [1] "inverse"
##   skew lwr.ci upr.ci
##   5.13   4.27   5.98
##   kurt lwr.ci upr.ci
##  28.36  18.51  42.29
## [1] "prestg80"
## [1] "squareroot"
##   skew lwr.ci upr.ci
##  -0.20  -0.32  -0.06
##   kurt lwr.ci upr.ci
##  -0.54  -0.72  -0.30
## [1] "log"
##   skew lwr.ci upr.ci
##  -0.96  -1.24  -0.74
##   kurt lwr.ci upr.ci
##   1.20   0.38   2.52
## [1] "inverse"
##   skew lwr.ci upr.ci
##   4.92   3.12   6.03
##   kurt lwr.ci upr.ci
##  37.33  17.68  53.95
```

# Creating new variables

```
hw1 <- hw1 %>%
  mutate(prestg80_sqrt = sqrt(prestg80),
         maeduc_sqrt_ref = sqrt((max(maeduc) + 1) - maeduc))
```

# Running analysis

```
gss.dat.2 <- lm(prestg80_sqrt~educ*maeduc_sqrt_ref,hw1)
summ(gss.dat.2, center = TRUE, digits = 5, confint = TRUE)
```

```
## MODEL INFO:
## Observations: 1162
## Dependent Variable: prestg80_sqrt
##
## MODEL FIT:
## F(3,1158) = 115.0119, p = 0
## R-squared = 0.22956
## Adj. R-squared = 0.22756
##
## Standard errors: OLS
##                          Est.     2.5%    97.5%    t val.       p
## (Intercept)           4.95567  4.884    5.02733 135.53361 0       ***
## educ                  0.24652  0.21924  0.2738   17.71211 0       ***
## maeduc_sqrt_ref       0.17481  0.04341  0.30621   2.60749 0.00924  **
## educ:maeduc_sqrt_ref -0.03965 -0.07439 -0.00491  -2.23726 0.02546   *
##
## All continuous predictors are mean-centered.
```

```
gss.dat.2 %>% center_lm() %>% lm.beta()
```

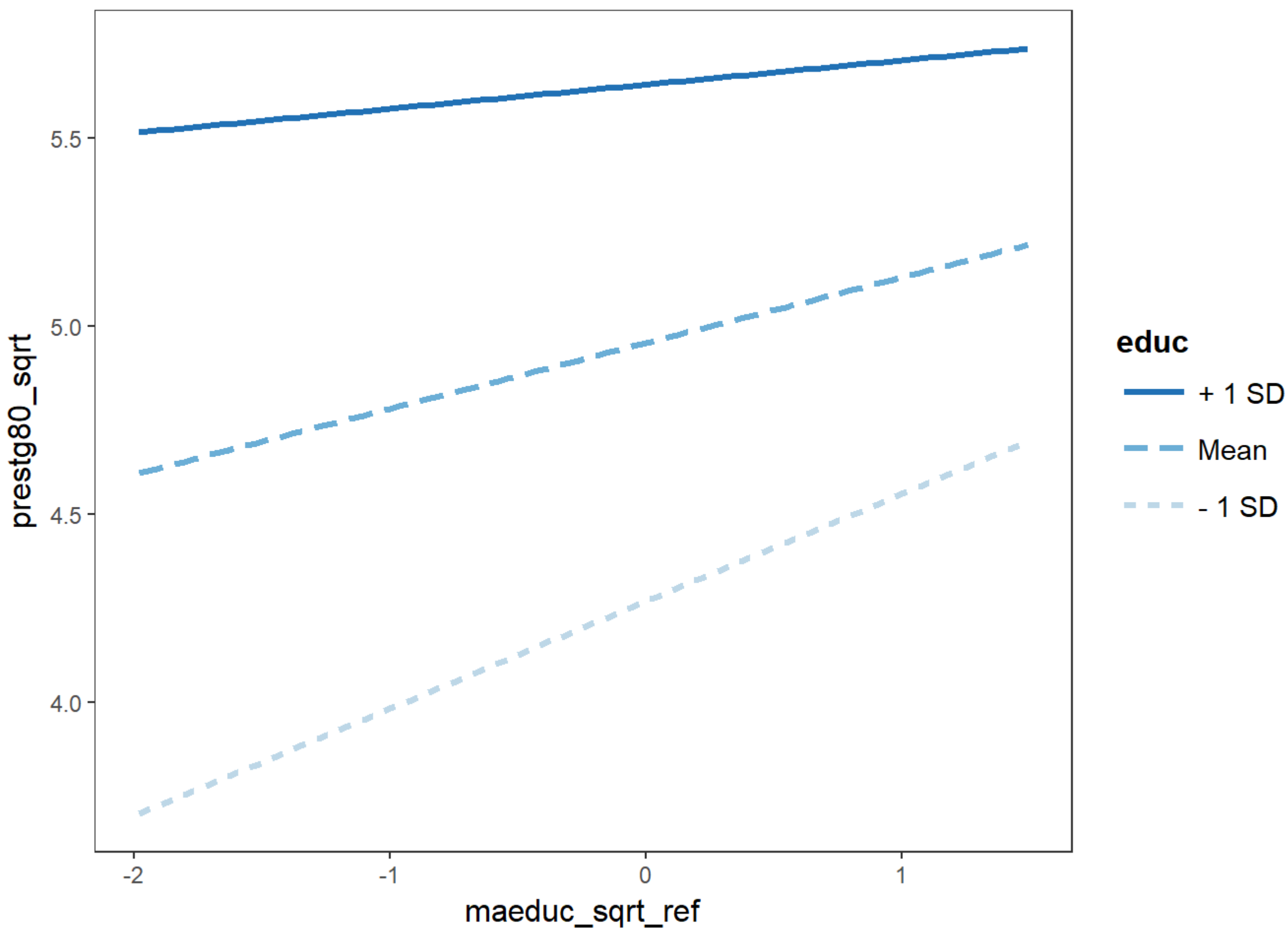
```
## Warning in b * sx: longer object length is not a multiple of shorter object
## length
```

```
##                 educ      maeduc_sqrt_ref educ:maeduc_sqrt_ref
##           0.51233115           0.07443852          -0.08240373
```

```
sim_slopes(gss.dat.2,maeduc_sqrt_ref,educ,johnson_neyman = FALSE, cont.int = TRUE, centered = c(
'educ','maeduc_sqrt_ref'), digits = 5)
```

```
## SIMPLE SLOPES ANALYSIS
##
## Slope of maeduc_sqrt_ref when educ = 2.78812 (+ 1 SD):
##    Est.    S.E.       p
## 0.06426 0.07953 0.41925
##
## Slope of maeduc_sqrt_ref when educ = 0 (Mean):
##    Est.    S.E.       p
## 0.17481 0.06704 0.00924
##
## Slope of maeduc_sqrt_ref when educ = -2.78812 (- 1 SD):
##    Est.    S.E.       p
## 0.28536 0.08688 0.00105
```

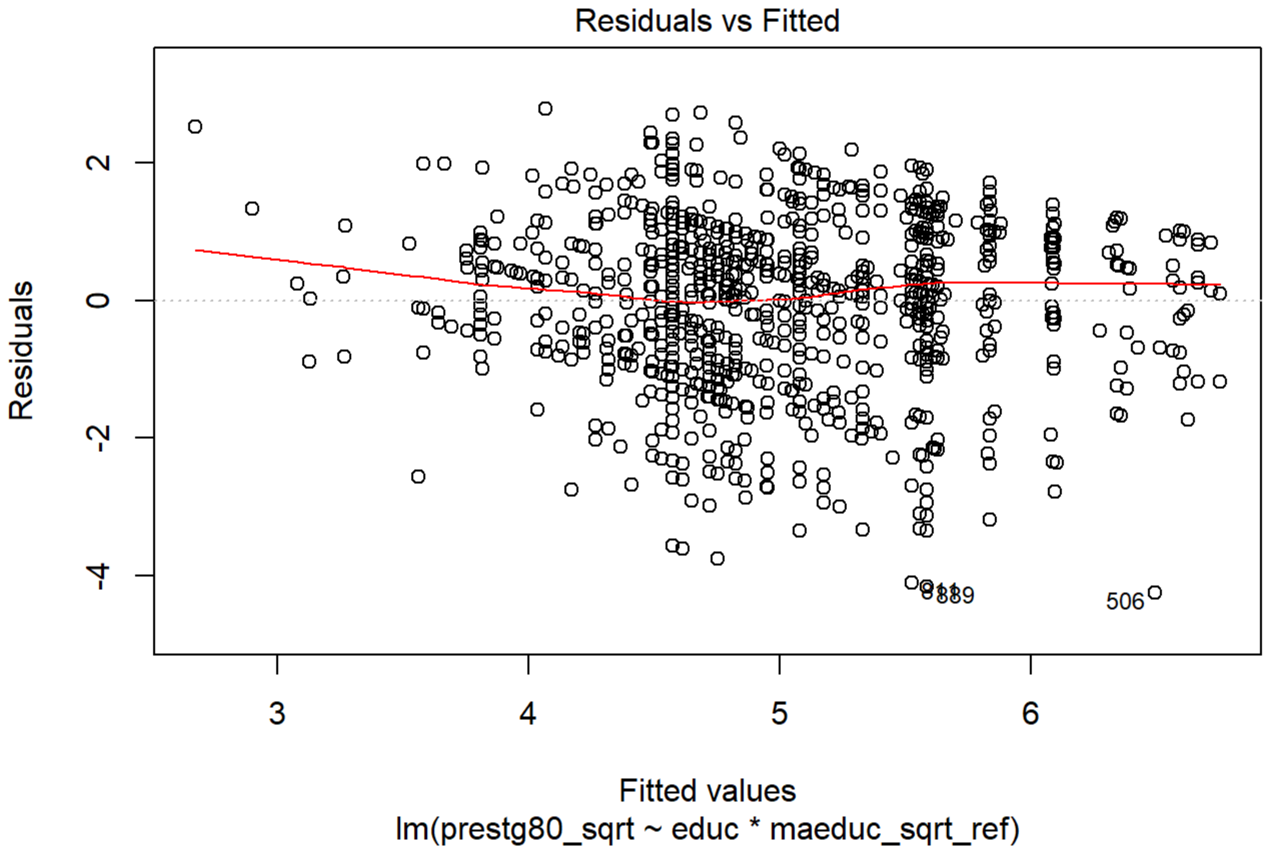
```
interact_plot(gss.dat.2,maeduc_sqrt_ref, educ,centered = c('educ','maeduc_sqrt_ref'))
```



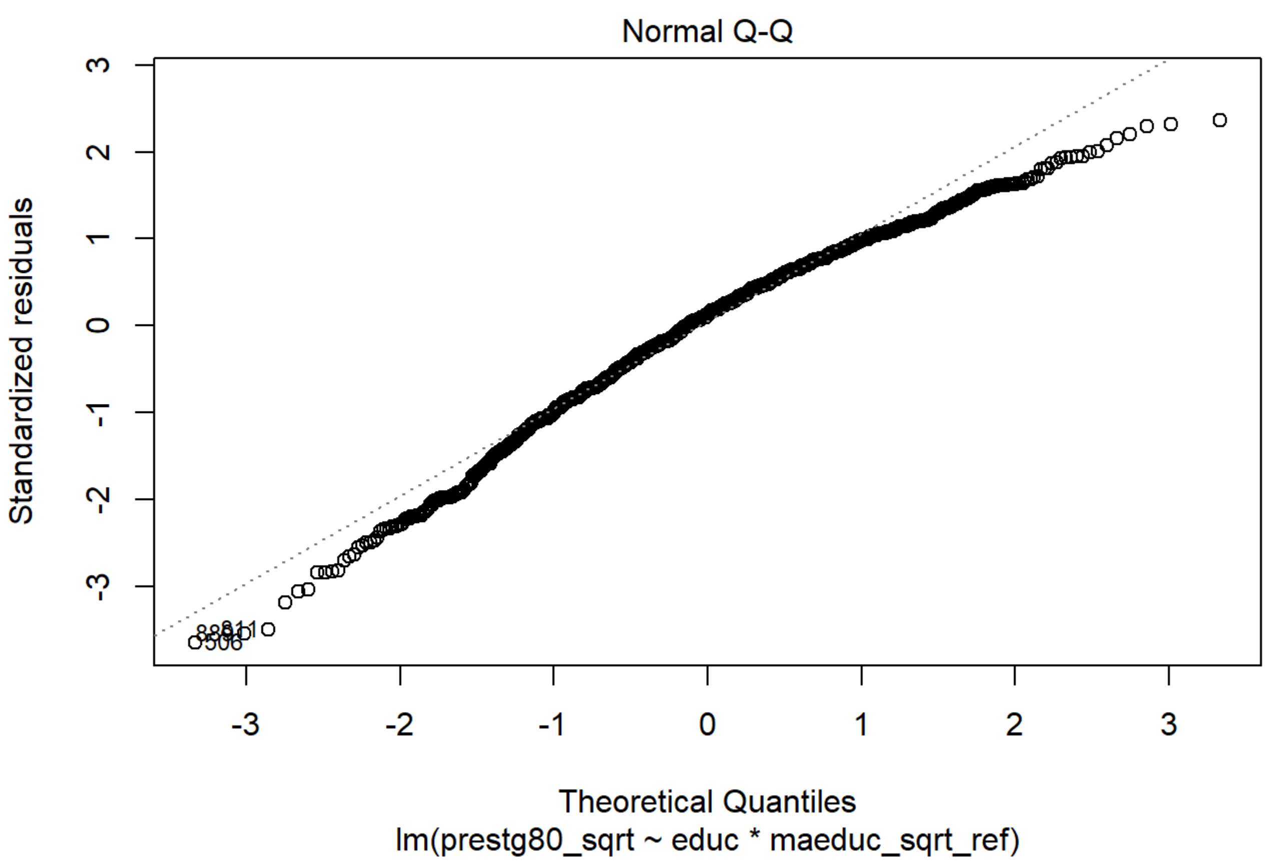

# HW3

Normality of residuals suggests that multiple points stray from the line, which might suggest a problem, further analysis are necessary. Linearity of the residuals from the scale-location, and residuals vs fitted graph are somewhat straight but require further analysis. Homoscedasticity is examined through the Breusch-pagan test which is significant, which suggests that homoscedasticity is met. Multicollinearity is examined through inflation and tolerance factors which are within margins.

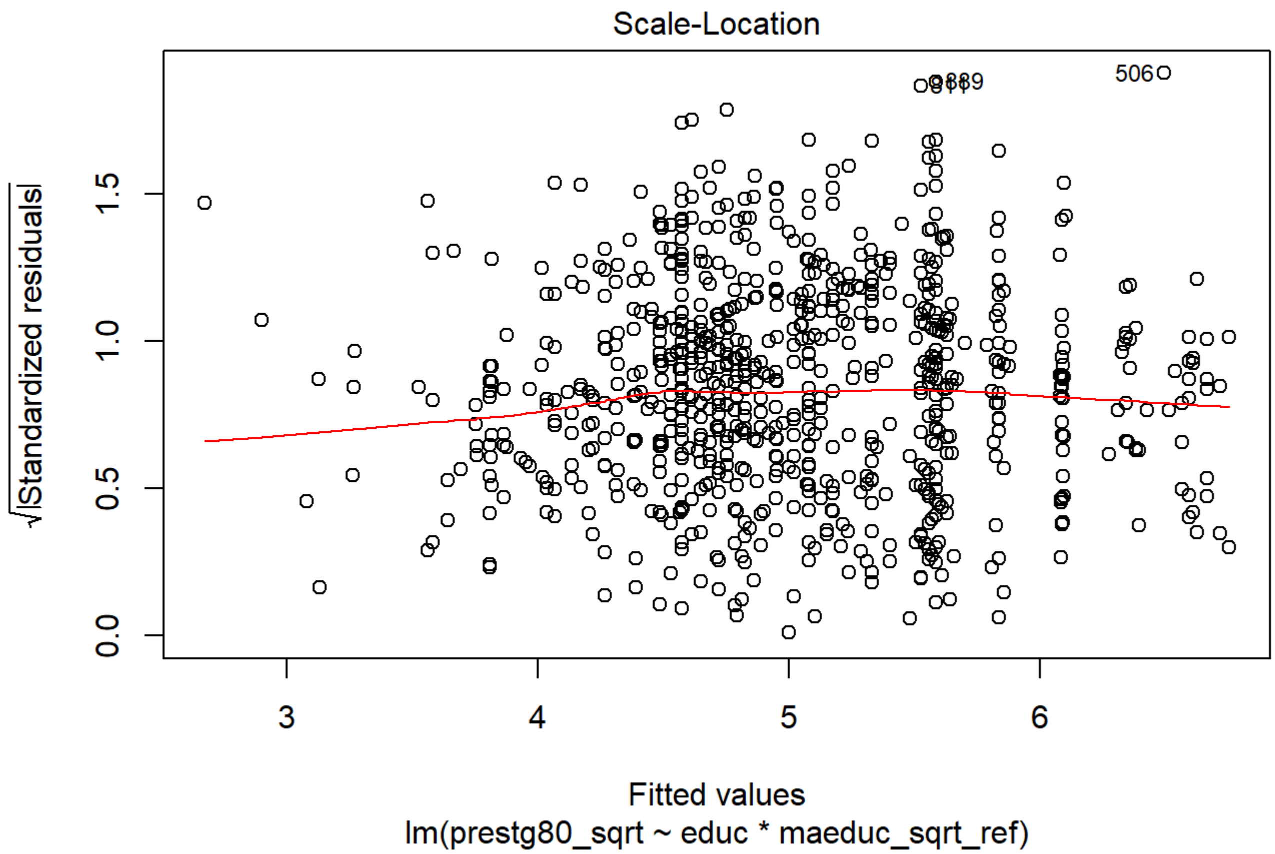
```
plot(gss.dat.2)
```

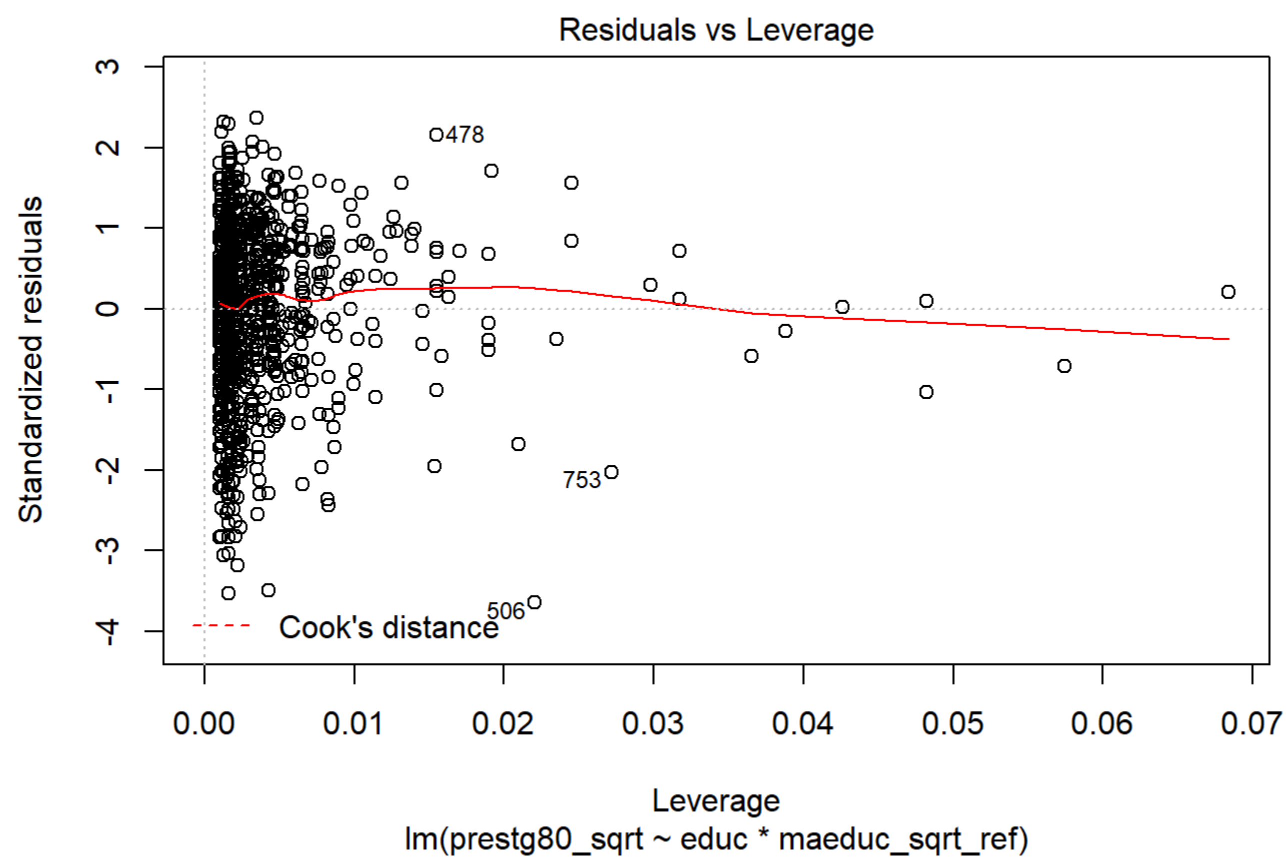


Residuals vs Fitted
Residuals
Fitted values
lm(prestg80_sqrt ~ educ * maeduc_sqrt_ref)
506
889

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(prestg80_sqrt ~ educ * maeduc_sqrt_ref)
506


Scale-Location
√|Standardized residuals|
Fitted values
lm(prestg80_sqrt ~ educ * maeduc_sqrt_ref)
506
0.0
0.5
1.0
1.5
3
4
5
6


Residuals vs Leverage
Standardized residuals
Leverage
lm(prestg80_sqrt ~ educ * maeduc_sqrt_ref)
478
753
506
Cook's distance
3
2
1
0
-1
-2
-3
-4
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07

```
n.2 <- 1162
hat <- hatvalues(gss.dat.2)
mahun <- ((n.2-1)*(hat))-1
tail(sort(mahun),10)
```

```
##        2       32      275      747      183      647        4     1013
## 35.87404 35.87404 35.87404 41.43713 44.02739 48.49484 54.99446 54.99446
##      634      641
## 65.67489 78.39294
```

```
1-pchisq(201.45,df = 3)
```

```
## [1] 0
```

```
vif(gss.dat.2)
```

```
##                 educ      maeduc_sqrt_ref educ:maeduc_sqrt_ref
##             21.14208             15.15147             20.85662
```

```
1/vif(gss.dat.2)
```

```
##                 educ      maeduc_sqrt_ref educ:maeduc_sqrt_ref
##           0.04729904           0.06600021           0.04794641
```

```
lmtest::bptest(gss.dat.2,varformula = ~fitted.values(gss.dat.2),FALSE,hw1)
```

```
## 
##  Breusch-Pagan test
## 
## data:  gss.dat.2
## BP = 1.5656, df = 1, p-value = 0.2109
```

# HW4

# 4A

Results from the ANCOVA suggests that employees make a larger begining salary based off of education $F$ = 302.80(1,471), $p$ <.001. It is important to look at the effect sizes, specfically that education (\(\eta^2\) = .38) accounts for a larger amount of explanation compared to minority status (\(\eta^2\) = .01)

Loading in data

```
load("C:/Users/Branly Mclanbry/Downloads/employee (7).RData")
hw2 <- employee %>%
  mutate(educ.num = as.numeric(educ))
```

Ancova model

```
salary.dat2 <- aov(salbegin~educ+minority,hw2)
Anova(salary.dat2, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: salbegin
##                 Sum Sq  Df  F value    Pr(>F)
## (Intercept) 6.1545e+08   1  16.6670 5.234e-05 ***
## educ        1.1181e+10   1 302.7943 < 2.2e-16 ***
## minority    1.6093e+08   1   4.3582   0.03737 *
## Residuals   1.7392e+10 471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
etaSquared(salary.dat2)
```

```
##               eta.sq eta.sq.part
## educ     0.381592056 0.391311067
## minority 0.005492375 0.009168272
```

# 4B

Significant interaction suggests that ANCOVA assumptions are violated.

```
salary.dat2 <- aov(salbegin~educ*minority,employee)
Anova(salary.dat2, type = "III")
```

```
## Anova Table (Type III tests)
##
## Response: salbegin
##                   Sum Sq  Df F value    Pr(>F)
## (Intercept)   1.1412e+09   1  32.221 2.412e-08 ***
## educ          1.1554e+10   1 326.211 < 2.2e-16 ***
## minority      5.8602e+08   1  16.545 5.571e-05 ***
## educ:minority 7.4509e+08   1  21.036 5.787e-06 ***
## Residuals     1.6647e+10 470
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# 4C

Still, running the model as a linear model provides slightly different results $R^2$ = .43, $F$(3.470) = 119.08, $p$ < .001. However, this suggests that education ($b^*$ = .69, $p$ <.001), minority ($b^*$ = .71, $p$ < .001) and the interaction ($b^*$ = -.42, $p$<.001) predicts beginning salary.

Essentially, higher levels of education lead to a higher beginning salary especially if you are a non-minority. Those in the minority status with equivalent education start at a lower salary.

```
salary.lm<- lm(salbegin ~ educ * minority,employee)
summ(salary.lm, center = TRUE, digits = 5, confint = TRUE)
```

```
## MODEL INFO:
## Observations: 474
## Dependent Variable: salbegin
##
## MODEL FIT:
## F(3,470) = 119.086, p = 0
## R-squared = 0.43186
## Adj. R-squared = 0.42823
##
## Standard errors: OLS
##                      Est.        2.5%       97.5%   t val.          p
## (Intercept)   17286.88133 16679.02414 17894.73851 55.73952     0      ***
## educ           1901.81596  1695.43607  2108.19585 18.06131     0      ***
## minority      -2070.7473  -3406.1674   -735.3272  -3.03919     0.0025  **
## educ:minority -1158.12344 -1653.02404  -663.22284 -4.58654 1e-05      ***
##
## All continuous predictors are mean-centered.
```

```
lm.beta(salary.lm)
```

```
## Warning in var(if (is.vector(x) || is.factor(x)) x else as.double(x), na.rm = na.rm): Calling
var(x) on a factor x is deprecated and will become an error.
##   Use something like 'all(duplicated(x)[-1L])' to test for a constant vector.
```

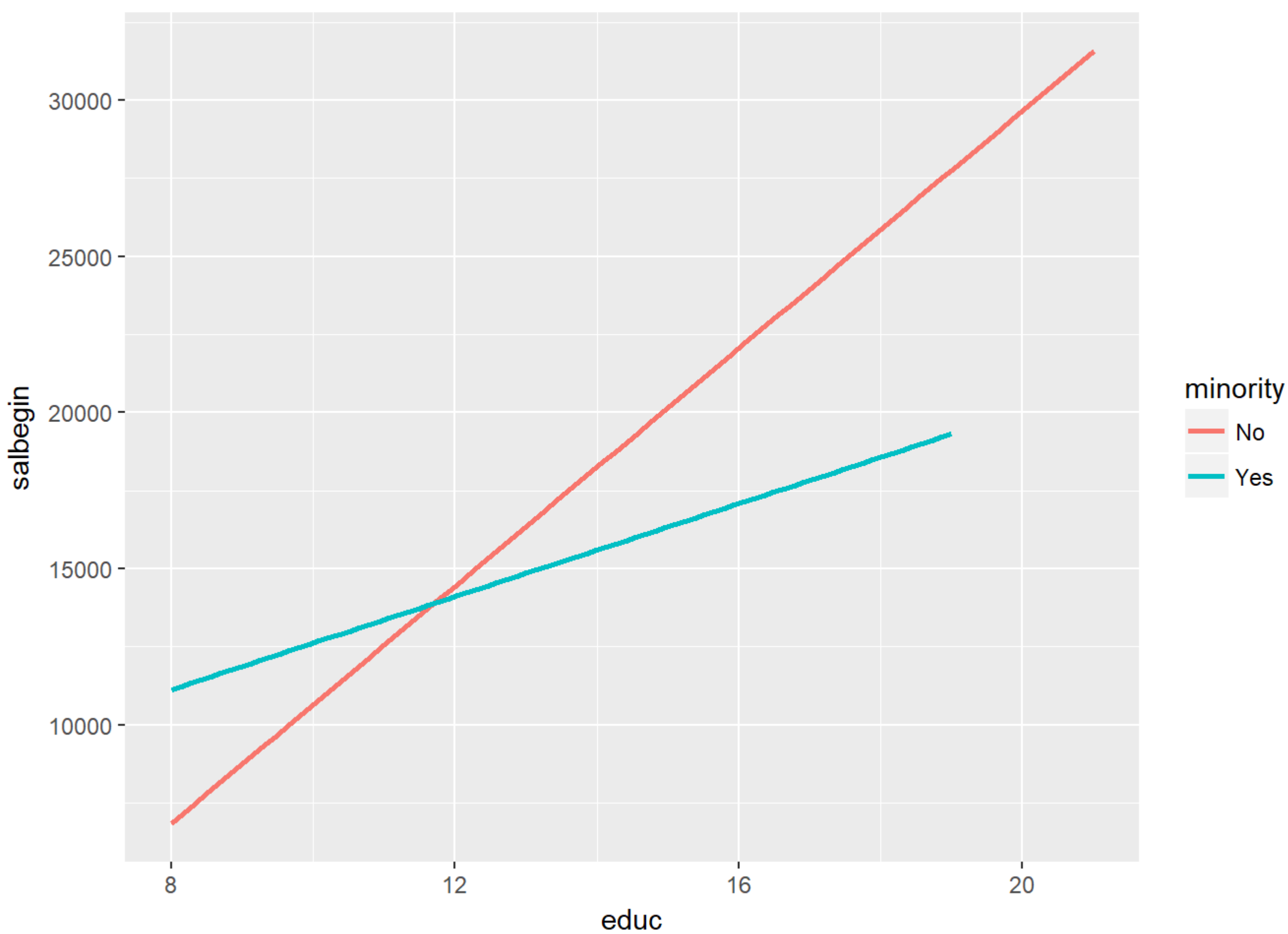
```
## Warning in b * sx: longer object length is not a multiple of shorter object
## length
```

```
##            educ      minorityYes educ:minorityYes
##       0.6970778        0.7134442       -0.4244901
```

```
sim_slopes(salary.lm,minority,educ,johnson_neyman = FALSE, cont.int = TRUE, centered = c('educ',
'minority'), digits = 5)
```

```
## SIMPLE SLOPES ANALYSIS
##
## Slope of minority when educ = 2.88485 (+ 1 SD):
##       Est.      S.E.         p
## -5411.756  1096.077     0.000
##
## Slope of minority when educ = 0 (Mean):
##        Est.       S.E.          p
## -2070.7473   681.3493     0.0025
##
## Slope of minority when educ = -2.88485 (- 1 SD):
##        Est.       S.E.          p
## 1270.26095  887.88164    0.15319
```

```
ggplot(employee, aes(educ, salbegin)) +
  geom_smooth(aes(color = minority), method = "lm", se = F)
```




# HW5

Authors tested ANCOVA by utilizing regression to test interactions between grade and condition on performance. This was met because the interaction was not significant $p = .40$ . ANCOVA suggests fairly similar results $F(1,22) = 3.92$, $p = .06$

```
h5 <- TOP2003 %>% janitor::clean_names()
h5.2 <- aov(quiz2 ~ current + condit, data = h5)
Anova(h5.2)
```

```
## Anova Table (Type II tests)
##
## Response: quiz2
##           Sum Sq Df F value  Pr(>F)
## current    9.323  1  3.3971 0.07883 .
## condit    10.745  1  3.9151 0.06051 .
## Residuals 60.377 22
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
h5.3 <- aov(quiz2 ~ current * condit, data = h5)
summary(h5.3)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## current         1   5.84   5.838   2.102 0.1619
## condit          1  10.74  10.745   3.868 0.0626 .
## current:condit  1   2.05   2.048   0.737 0.4002
## Residuals      21  58.33   2.778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1 observation deleted due to missingness
```

# HW6

Power analysis suggests sample size around 140 for significant beta for quality.

Analysis

```
cor.dat <- cor(grants)
round(cor.dat,2)
```

```
##         SUBMIT QUALITY UNIVERS MONEY
## SUBMIT    1.00   -0.80   -0.60 -0.24
## QUALITY  -0.80    1.00    0.72  0.45
## UNIVERS  -0.60    0.72    1.00  0.56
## MONEY    -0.24    0.45    0.56  1.00
```

```
pwr.MRC_all(-.24,.45,.56,-.80,-.60,.72,140)
```

```
## [1] "Sample size is  140"
## [1] "Power R2 =  1"
## [1] "Power b1 =  0.885"
## [1] "Power b2 =  0.8089"
## [1] "Power b3 =  0.9988"
## [1] "Proportion Rejecting None =  0"
## [1] "Proportion Rejecting One =  0.0705"
## [1] "Proportion Rejecting Two =  0.1663"
## [1] "Power ALL (Proportion Rejecting All) =  0.7632"
```

```
pwr.MRC_all(-.80,-.60,-.24,.72,.45,.56,150)
```

```
## [1] "Sample size is  150"
## [1] "Power R2 =  1"
## [1] "Power b1 =  1"
## [1] "Power b2 =  0.4875"
## [1] "Power b3 =  0.9094"
## [1] "Proportion Rejecting None =  0"
## [1] "Proportion Rejecting One =  0.0696"
## [1] "Proportion Rejecting Two =  0.4639"
## [1] "Power ALL (Proportion Rejecting All) =  0.4665"
```