ANOVA and Regression

Contents

```
library(mosaic)
library(ggplot2)
library(dplyr)
library(psych)
```

ANOVA

We are going to be using the Big Five Inventory dataset, `bfi`, to demonstrate ANOVA. This dataset it contained in the `psych` package.

```
data(bfi)
```

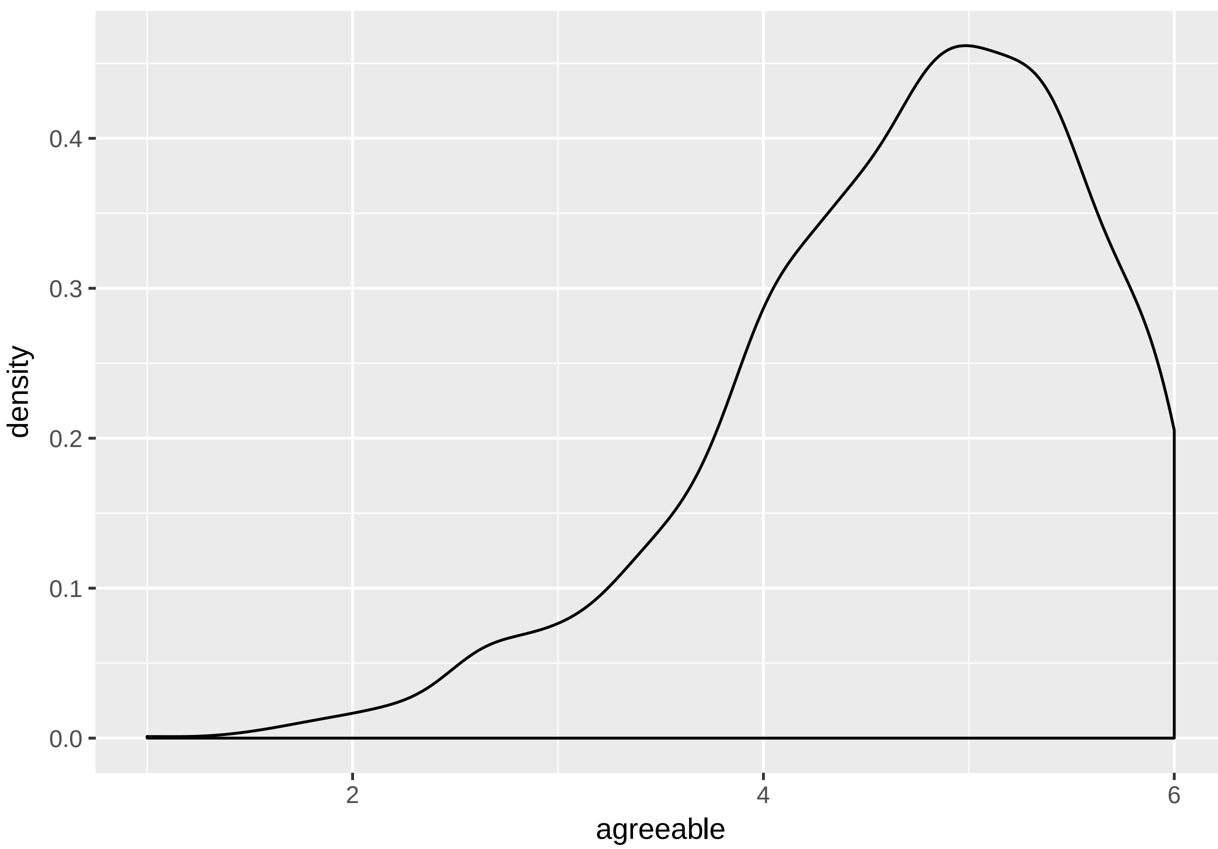
Before we get into the ANOVA, we should first create all of our scale scores.

```
corr.test(select(bfi, A1, A2, A3, A4, A5))
corr.test(select(bfi, C1, C2, C3, C4, C5))
corr.test(select(bfi, E1, E2, E3, E4, E5))
corr.test(select(bfi, N1, N2, N3, N4, N5))
corr.test(select(bfi, O1, O2, O3, O4, O5))

bfi <- bfi %>%
  mutate(A1.r = (min(A1, na.rm = TRUE) + max(A1, na.rm = TRUE)) - A1,
         C4.r = (min(C4, na.rm = TRUE) + max(C4, na.rm = TRUE)) - C4,
         C5.r = (min(C5, na.rm = TRUE) + max(C5, na.rm = TRUE)) - C5,
         E1.r = (min(E1, na.rm = TRUE) + max(E1, na.rm = TRUE)) - E1,
         E2.r = (min(E2, na.rm = TRUE) + max(E2, na.rm = TRUE)) - E2,
         O2.r = (min(O2, na.rm = TRUE) + max(O2, na.rm = TRUE)) - O2,
         O5.r = (min(O5, na.rm = TRUE) + max(O5, na.rm = TRUE)) - O5)
```

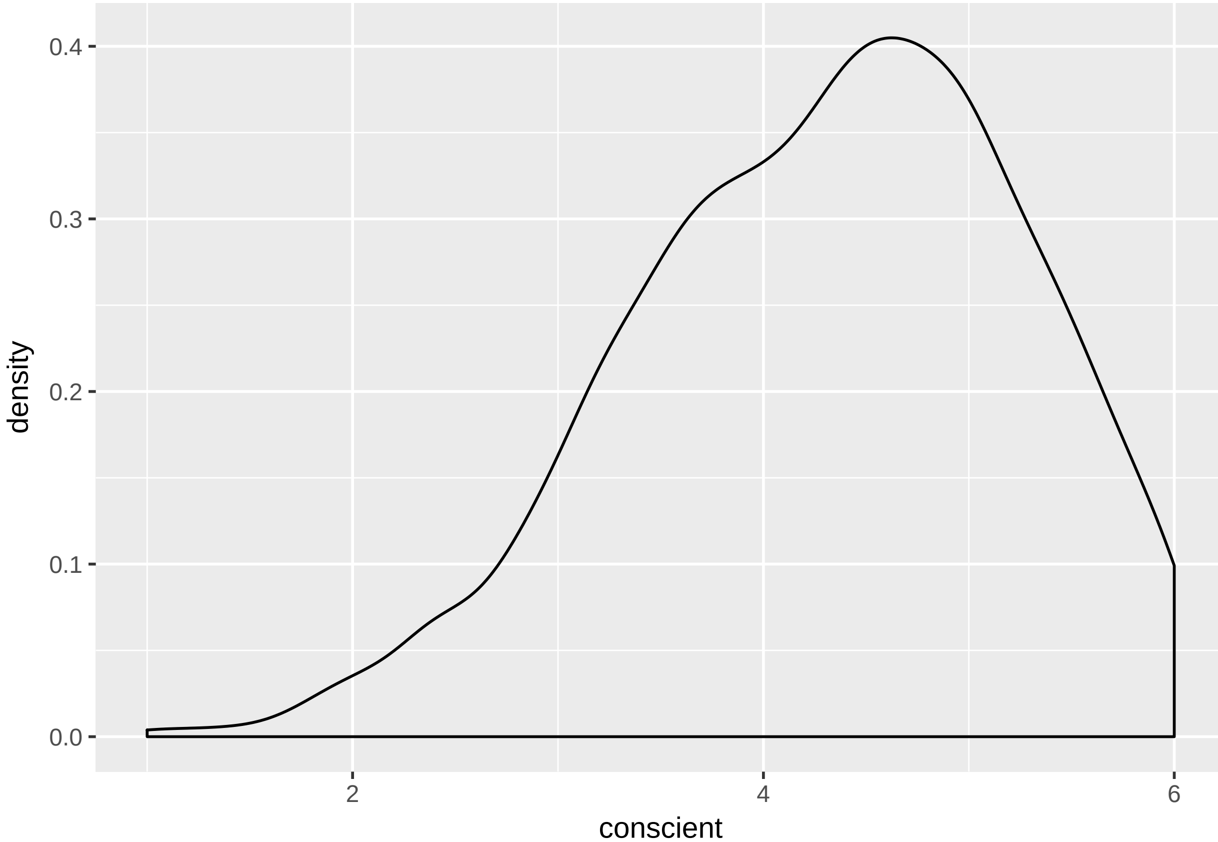
```r
alpha(select(bfi, A1.r, A2, A3, A4, A5))
alpha(select(bfi, C1, C2, C3, C4.r, C5.r))
alpha(select(bfi, E1.r, E2.r, E3, E4, E5))
alpha(select(bfi, N1, N2, N3, N4, N5))
alpha(select(bfi, O1, O2.r, O3, O4, O5.r))

bfi <- bfi %>%
  mutate(agreeable = (A1.r + A2 + A3 + A4 + A5)/5,
         conscient = (C1 + C2 + C3 + C4.r + C5.r)/5,
         extrov = (E1.r + E2.r + E3 + E4 + E5)/5,
         neurot = (N1 + N2 + N3 + N4 + N5)/5,
         openness = (O1 + O2.r + O3 + O4 + O5.r)/5) %>%
  filter(!is.na(education))

glimpse(bfi)
```

Let's also take a look at the distributions of our new variables.

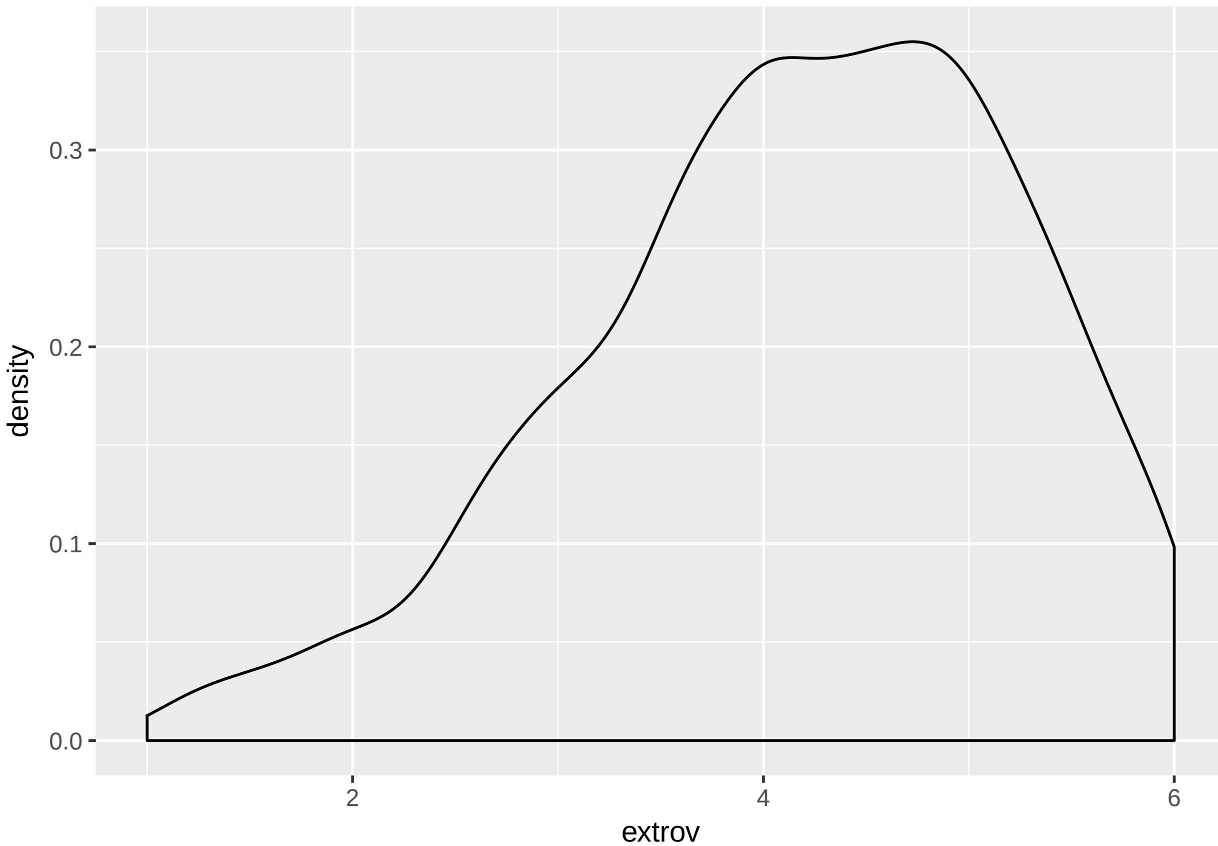
```r
ggplot(bfi, aes(x = agreeable)) + geom_density()
```



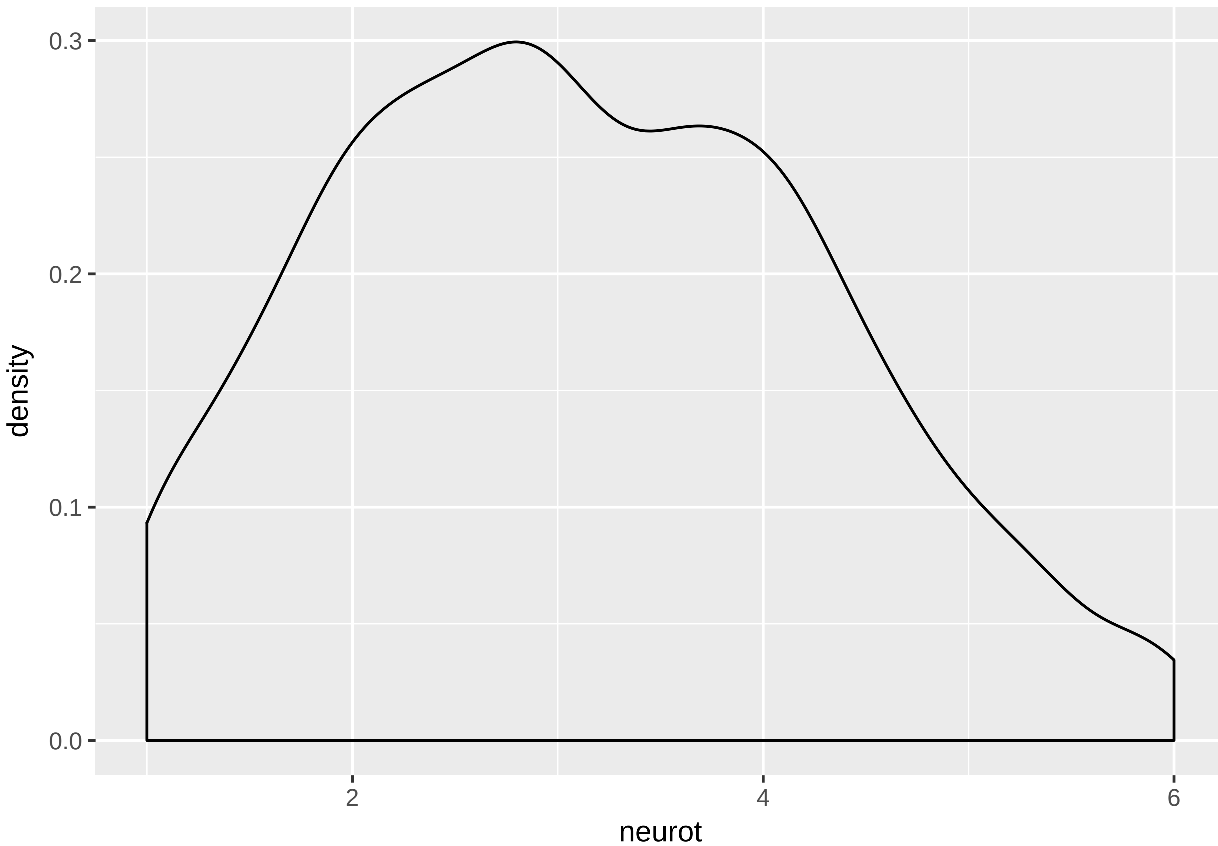
```r
ggplot(bfi, aes(x = conscient)) + geom_density()
```



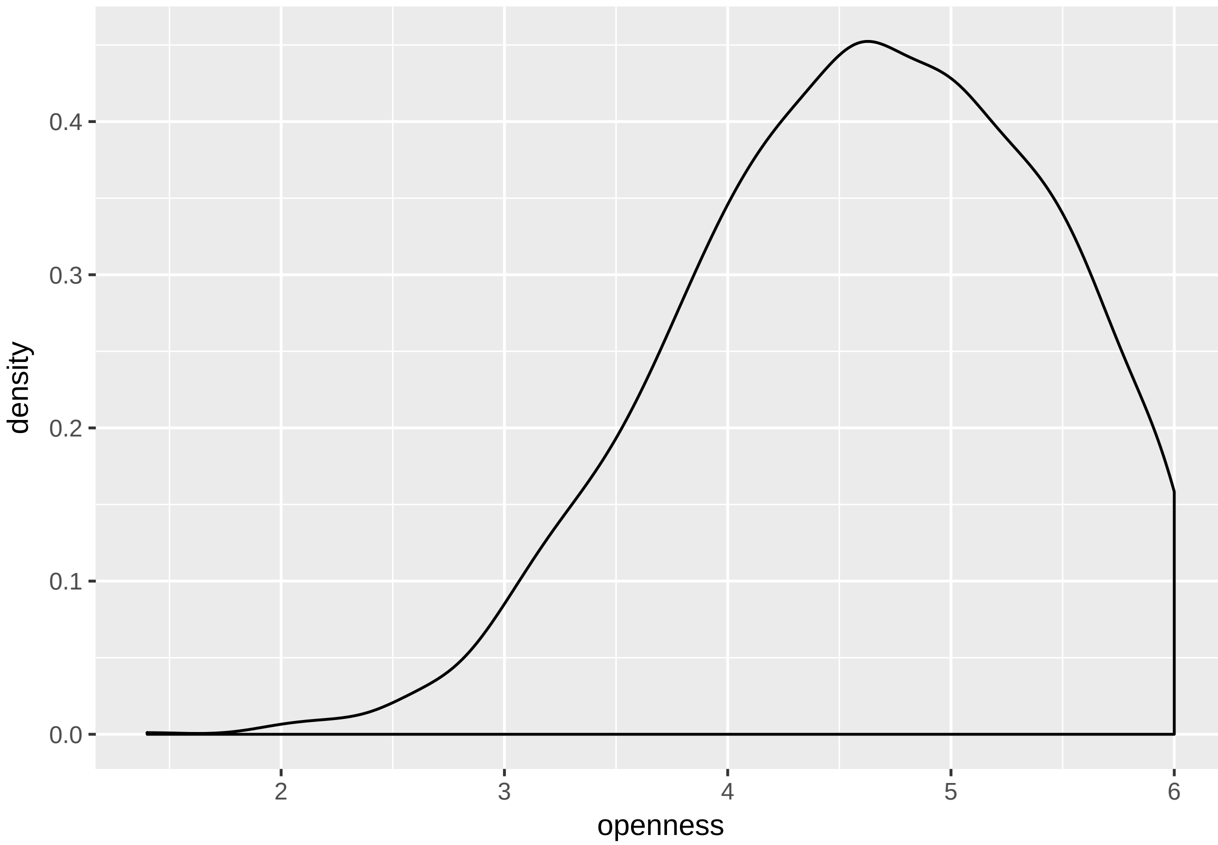

```r
ggplot(bfi, aes(x = extrov)) + geom_density()
```




```r
ggplot(bfi, aes(x = neurot)) + geom_density()
```




```r
ggplot(bfi, aes(x = openness)) + geom_density()
```



Yesterday we used `t.test()` to test for differences in conscientiousness between those who graduated college, and those who did not. Note that I did not create the variable again, I used a logical statement directly in the `t.test()` function.

```
t.test(conscient ~ (education > 3), data = bfi)
```

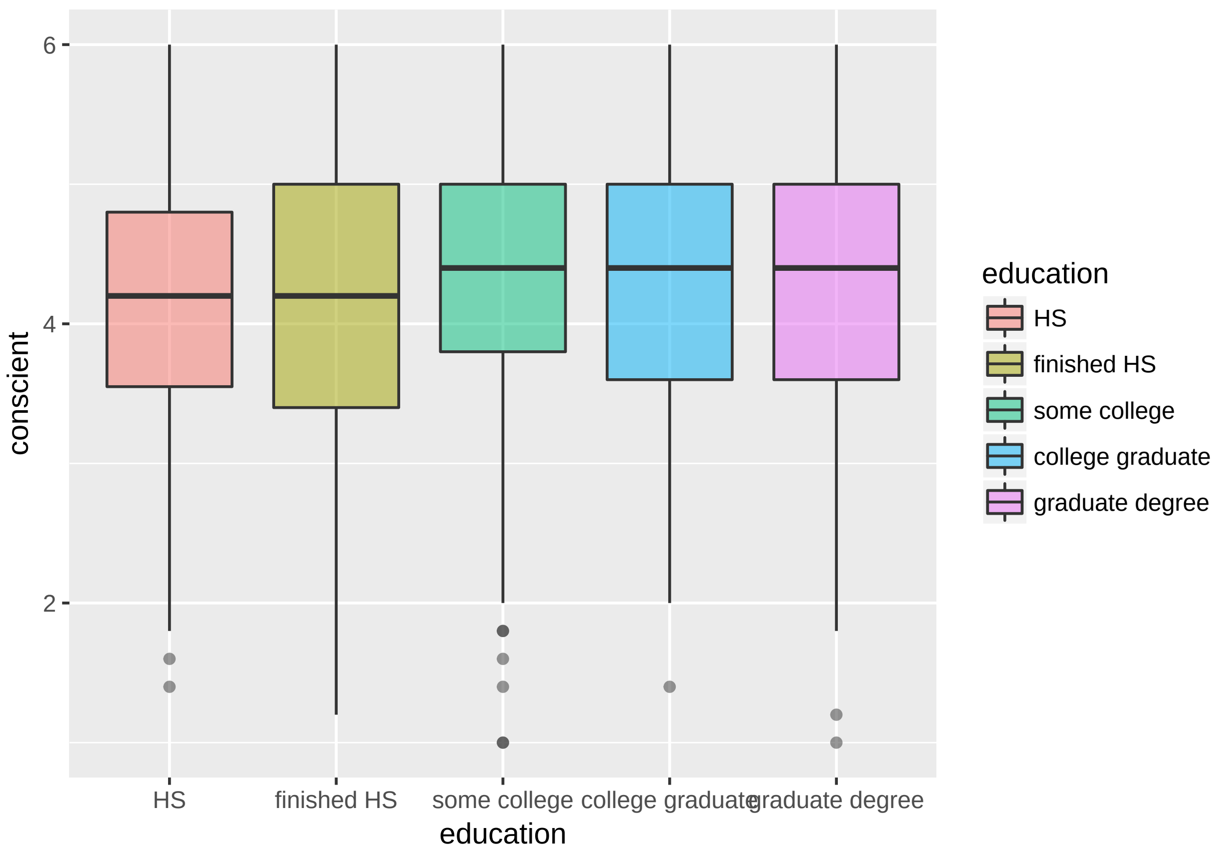
```
##
##  Welch Two Sample t-test
##
## data:  conscient by education > 3
## t = 1.77, df = 1496.6, p-value = 0.07694
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.007864586  0.153181568
## sample estimates:
## mean in group FALSE  mean in group TRUE
##            4.325250            4.252592
```

A better thing to do would be to check for differences with a one-way anova.

One-Way ANOVA

Perform a one-way ANOVA for education level on conscientiousness. Let's first look at the distributions of conscientiousness by education level. Note that first we're changing the education variable's type from integer to factor with the `as.factor()` function and giving nice labels to the factor levels.

```
bfi <- bfi %>%
  mutate(education = as.factor(education),
         education = factor(education, labels=c('HS',
                                                  'finished HS',
                                                  'some college',
                                                  'college graduate',
                                                  'graduate degree')))

ggplot(bfi, aes(x = education, y = conscient, fill = education)) +
  geom_boxplot(alpha = .5)
```



Also, let's get descriptives by education level.

```
#descriptives
```

We can perform a Levene's test to test the homogeneity of variance assumption with the `leveneTest()` function that's in the `car` package.

```r
#install.packages("car")
library(car)

leveneTest(bfi$conscient, bfi$education)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    4  1.6167 0.1673
##       2485
```

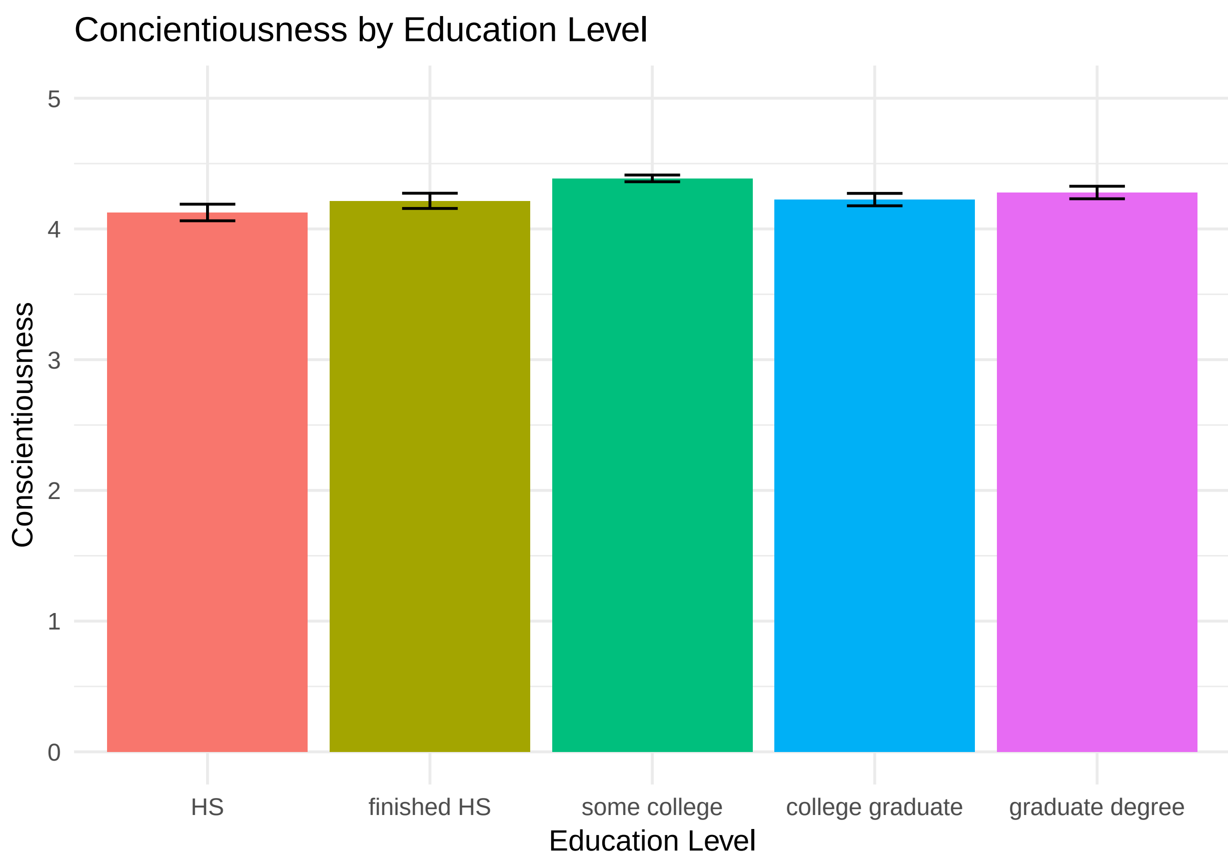
Now that we've checked out assumptions, finally, we can run the one-way ANOVA.

```r
mod1 <- aov(conscient ~ education, data = bfi)

summary(mod1)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## education      4     20   4.998   5.665 0.000155 ***
## Residuals   2485   2192   0.882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 87 observations deleted due to missingness
```

There are statistically significant differences between people with different education levels in conscientiousness, $F(4, 2485) = 5.67$, $p < .001$. We might want a bar graph for publication!

```r
#A small companion dataset for making error bars
plotdata <- bfi %>%
  group_by(education) %>%
  summarise(mean = mean(conscient, na.rm = TRUE),
            stdv = sd(conscient, na.rm = TRUE),
            n = n()) %>%
  mutate(se = stdv/sqrt(n))

#Making the Bar Graph
ggplot(plotdata, aes(x = education,
                     y = mean,
                     fill = education)) +
  geom_bar(stat = "identity") +
  geom_errorbar(aes(ymax = mean + se, ymin = mean - se),
                position = position_dodge(0.9), width = 0.25) +
  labs(x = "Education Level", y = "Conscientiousness") +
  ylim(0, 5) +  #the scale is really from 1 to 6
  ggtitle("Concientiousness by Education Level") +
  scale_fill_discrete(guide = FALSE) +
  theme_minimal()
```



Two-Way ANOVA

First let's do some data stuff we will need. For the Levene's Test we will create the gender X education levels with the `unite()` function.

```
#install.packages("tidyr")
library(tidyr)

bfi <- bfi %>%
  unite(gen_edu, gender, education, remove = FALSE) %>%
  mutate(gender = as.factor(factor(gender, labels=c('Men', 'Women'))),
         gen_edu = as.factor(gen_edu))
```

Two-way ANOVA for gender by education level on conscientiousness. We can get `favstats()` split by another categorical variable with the | symbol. It's above your `return` key.

```
favstats(conscient ~ education|gender, data = bfi)
```

```
##                   gender min   Q1 median  Q3 max     mean        sd   n
## 1                 HS.Men 2.4 3.60    4.2 5.0   6 4.238636 0.8597556  88
## 2        finished HS.Men 1.2 3.40    4.0 4.8   6 4.001980 1.0274220 101
## 3       some college.Men 1.0 3.60    4.4 5.0   6 4.259824 0.9434286 341
```

```
## 4    college graduate.Men 2.0 3.40    4.0 4.8    6 4.127692 0.9567399  130
## 5     graduate degree.Men 1.2 3.60    4.2 4.8    6 4.128000 1.0052960  150
## 6                HS.Women 1.4 3.40    4.2 4.8    6 4.048438 1.0007072  128
## 7       finished HS.Women 1.6 3.75    4.4 5.0    6 4.334444 0.9578445  180
## 8      some college.Women 1.0 3.80    4.6 5.2    6 4.437166 0.8884339  861
## 9  college graduate.Women 1.4 3.60    4.4 5.0    6 4.274603 0.9356112  252
## 10  graduate degree.Women 1.0 3.80    4.4 5.0    6 4.366023 0.9587237  259
## 11                    Men 1.0 3.40    4.2 5.0    6 4.179753 0.9612954  810
## 12                  Women 1.0 3.80    4.4 5.0    6 4.361190 0.9282123 1680
##    missing
## 1        5
## 2        2
## 3       15
## 4        4
## 5        2
## 6        3
## 7        9
## 8       32
## 9        8
## 10       7
## 11      28
## 12      59
```

Alternatively we can use `dplyr`

```
bfi %>%
  group_by(gender, education) %>% #could also use gen_edu here
  summarise(M = mean(conscient, na.rm = TRUE),
            Md = median(conscient, na.rm = TRUE),
            SD = sd(conscient, na.rm = TRUE))
```

```
## Source: local data frame [10 x 5]
## Groups: gender [?]
##
## # A tibble: 10 x 5
##    gender        education        M    Md        SD
##    <fctr>           <fctr>    <dbl> <dbl>     <dbl>
##  1    Men               HS 4.238636   4.2 0.8597556
##  2    Men      finished HS 4.001980   4.0 1.0274220
##  3    Men     some college 4.259824   4.4 0.9434286
##  4    Men college graduate 4.127692   4.0 0.9567399
##  5    Men  graduate degree 4.128000   4.2 1.0052960
##  6  Women               HS 4.048438   4.2 1.0007072
##  7  Women      finished HS 4.334444   4.4 0.9578445
##  8  Women     some college 4.437166   4.6 0.8884339
##  9  Women college graduate 4.274603   4.4 0.9356112
```

```
## 10  Women  graduate degree 4.366023   4.4 0.9587237
```

Let's check our assumption of homogeneity of variance.

```r
leveneTest(bfi$conscient, bfi$gen_edu)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    9  0.9593  0.472
##       2480
```

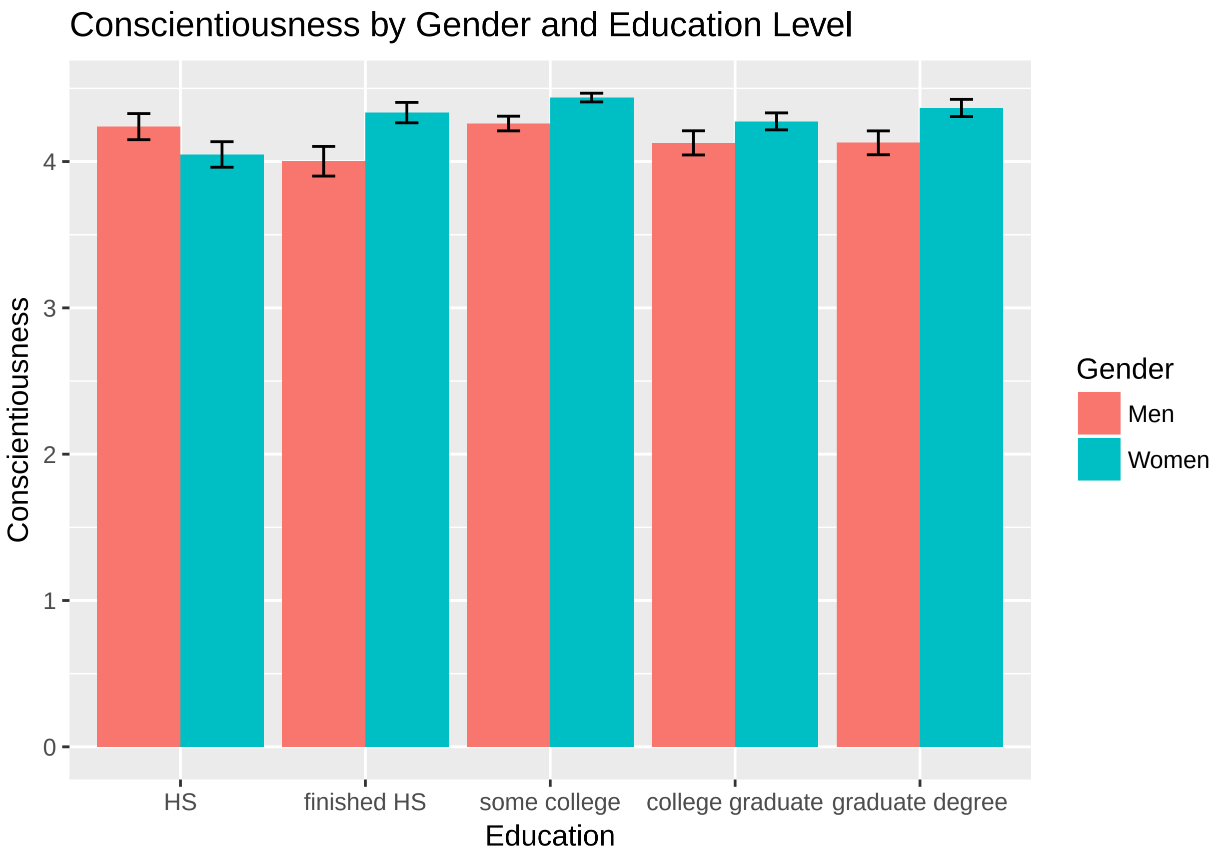
Because Levene's Test is non-significant we have evidence of homogeneity of variance (really, no evidence of heterogeneity of variance). Now we perform the two-way ANOVA.

```r
mod2 <- aov(conscient ~ education*gender, data = bfi)

summary(mod2)
```

```
##                    Df Sum Sq Mean Sq F value   Pr(>F)    
## education           4     20   4.998   5.716 0.000141 ***
## gender              1     15  14.988  17.143 3.58e-05 ***
## education:gender    4      9   2.241   2.563 0.036625 *  
## Residuals        2480   2168   0.874                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 87 observations deleted due to missingness
```

There is a statistically significant main effect of gender, $F(1, 2480) = 17.14$, $p < .001$, such that, women ($M = 4.36$, $SD = 0.93$) are more conscientious than men ($M = 4.18$, $SD = 0.96$). These main effects are qualified by a statistically significant two-way interaction of education and gender, $F(4, 2480) = 2.56$, $p = .037$. Let's make a bar graph!

```r
plotdata <- bfi %>%
  group_by(gender, education) %>%
  summarise(mean = mean(conscient, na.rm = TRUE),
            stdv = sd(conscient, na.rm = TRUE),
            n = n()) %>%
  mutate(se = stdv/sqrt(n))

ggplot(plotdata, aes(x = education, y = mean, fill = gender)) +
  geom_bar(stat = "identity", position = position_dodge(0.9)) +
  geom_errorbar(aes(ymax = mean + se, ymin = mean - se), position = position_dodge(0.9)
  labs(x = "Education", y = "Conscientiousness") +
  ggtitle("Conscientiousness by Gender and Education Level") +
  scale_fill_discrete(name = "Gender")
```



As we see in the graph (and you saw yesterday from the t-tests), women are higher than men in conscientiousness for every level of education expect for those participants that did not finish high school.

Try fiddling with the graph. Change the theme, change the labels. See if you can find on the internet how to change the colors.

```
#Copy and paste the bar graph code above to start fiddling with it.
```

We will see another example of a two-way ANOVA in the reproducible APA style document.

Mixed Effects ANOVA (Split-Plot)

To demonstrate the mixed effects ANOVA we'll use the `sat.act` dataset. Recall that the `sat.act` dataset has information for 700 people on their SAT verbal, SAT quantitative, and ACT scores.

```
#?sat.act
data(sat.act)
```

Recall from yesterday that boys were higher on the SAT quantitative than the SAT verbal (on average) and the girls had the opposite test patterns.

```r
sat.act %>%
  filter(gender == 1) %>%
  t.test(SATV, SATQ, data = ., paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  SATV and SATQ
## t = -3.4934, df = 244, p-value = 0.0005661
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -32.08114  -8.94743
## sample estimates:
## mean of the differences
##               -20.51429
```

```r
sat.act %>%
  filter(gender == 2) %>%
  t.test(SATV, SATQ, data = ., paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  SATV and SATQ
## t = 3.1813, df = 441, p-value = 0.00157
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   5.604249 23.721543
## sample estimates:
## mean of the differences
##                 14.6629
```

We can test if these two patterns are statistically different from each other with a gender (**between-subjects**) by topic (**within-subjects**) two-way mixed effects ANOVA. Where topic has two levels: 1) Verbal and 2) quantitative. We have to restructure our data first.

```r
n <- nrow(sat.act)

sat.act <- sat.act %>%
  mutate(id = seq.int(n))

sat.act_long <- sat.act %>%
  gather(topic, score, SATV, SATQ) %>%
  mutate(gender = as.factor(factor(gender, labels=c('Boys', 'Girls'))),
         topic = as.factor(topic)) %>%
  arrange(id)
```

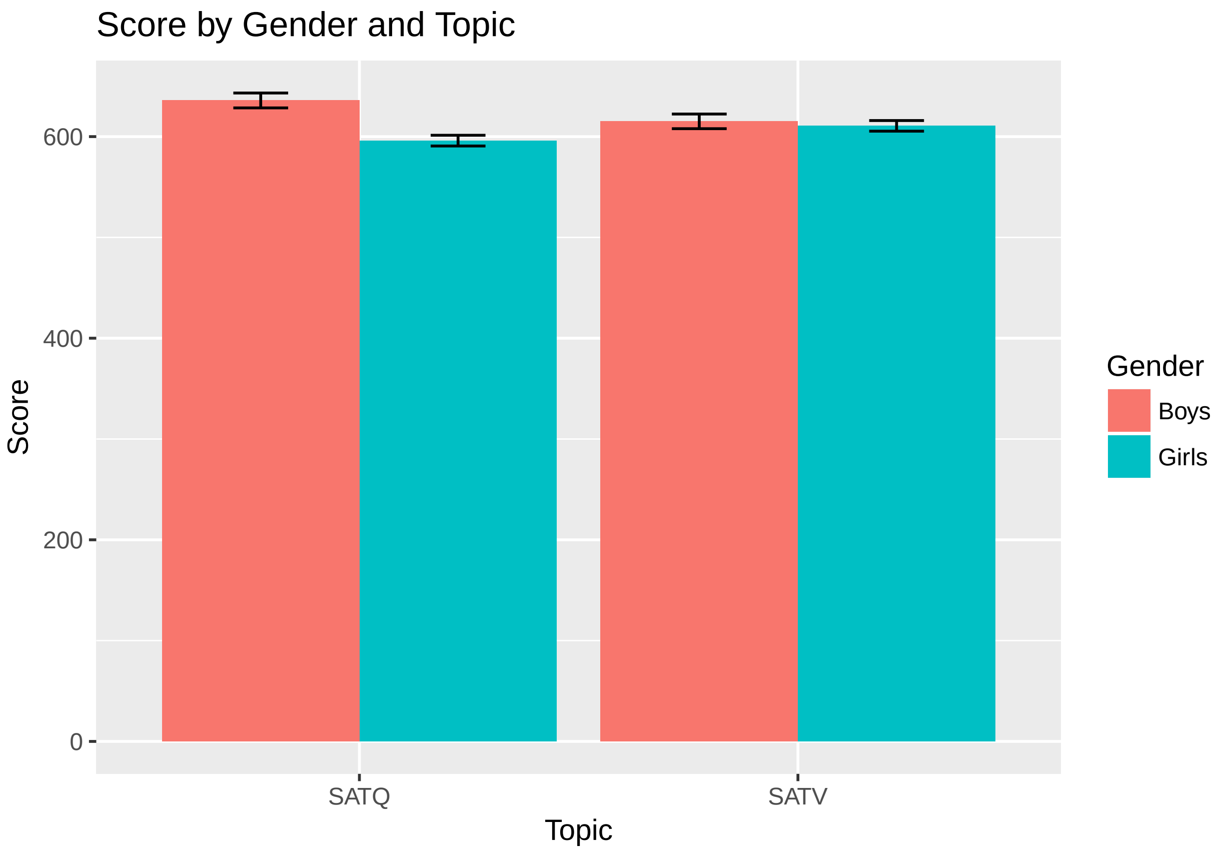
After you've taken a good look at the data, it's time for the ANOVA.

```
mod3 <- aov(score ~ gender + Error(id/gender) + topic*gender,
            data = sat.act_long)

summary(mod3)
```

```
##
## Error: id
##        Df Sum Sq Mean Sq
## gender  1  51294   51294
##
## Error: id:gender
##        Df Sum Sq Mean Sq
## gender  1  71083   71083
##
## Error: Within
##                Df   Sum Sq Mean Sq F value  Pr(>F)
## gender          1   107331  107331   8.348 0.00392 **
## topic           1     1440    1440   0.112 0.73791
## gender:topic    1    99154   99154   7.712 0.00556 **
## Residuals    1381 17754753   12856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As we may have guessed from our preliminary analyses, there is a main effect of gender, $F(1, 1381) = 8.35$, $p = .004$, no main effect of topic, $F(1, 1381) = 0.11$, $p = .738$, but a significant interaction of gender and topic, $F(1, 1381) = 7.71$, $p = .006$.

```
plotdata <- sat.act_long %>%
  group_by(gender, topic) %>%
  summarise(mean = mean(score, na.rm = TRUE),
            stdv = sd(score, na.rm = TRUE),
            n = n()) %>%
  mutate(se = stdv/sqrt(n))

ggplot(plotdata, aes(x = topic, y = mean, fill = gender)) +
  geom_bar(stat = "identity", position = position_dodge(0.9)) +
  geom_errorbar(aes(ymax = mean + se, ymin = mean - se), position = position_dodge(0.9)
  labs(x = "Topic", y = "Score") +
  ggtitle("Score by Gender and Topic") +
  scale_fill_discrete(name = "Gender")
```



It looks as though there are only gender differences on the SATQ, but not the SATV. Follow this up with two t-test of gender for each topic. Hint: use the `sat.act` data and the `filter()` function.

```
#two independent samples t-tests here.
```

Regression

Back to the `bfi` dataset. What if we wanted to treat `education` like an interval measured variable instead of ordinal? We could then regress conscientiousness on education in a simple linear regression model. First, let's clear our environment and re-load the data.

```
data(bfi)

bfi <- bfi %>%
  mutate(C4.r = (min(C4, na.rm = TRUE) + max(C4, na.rm = TRUE)) - C4,
         C5.r = (min(C5, na.rm = TRUE) + max(C5, na.rm = TRUE)) - C5,
         conscient = (C1 + C2 + C3 + C4.r + C5.r)/5)
```

It's good practice to make a scatter plot before running a regression. Do this using `ggplot2`. You might want to try using `geom_jitter()`. Also, add a linear regression line using `geom_smooth(method = "lm")`.

```r
#make a scatter plot here.
```

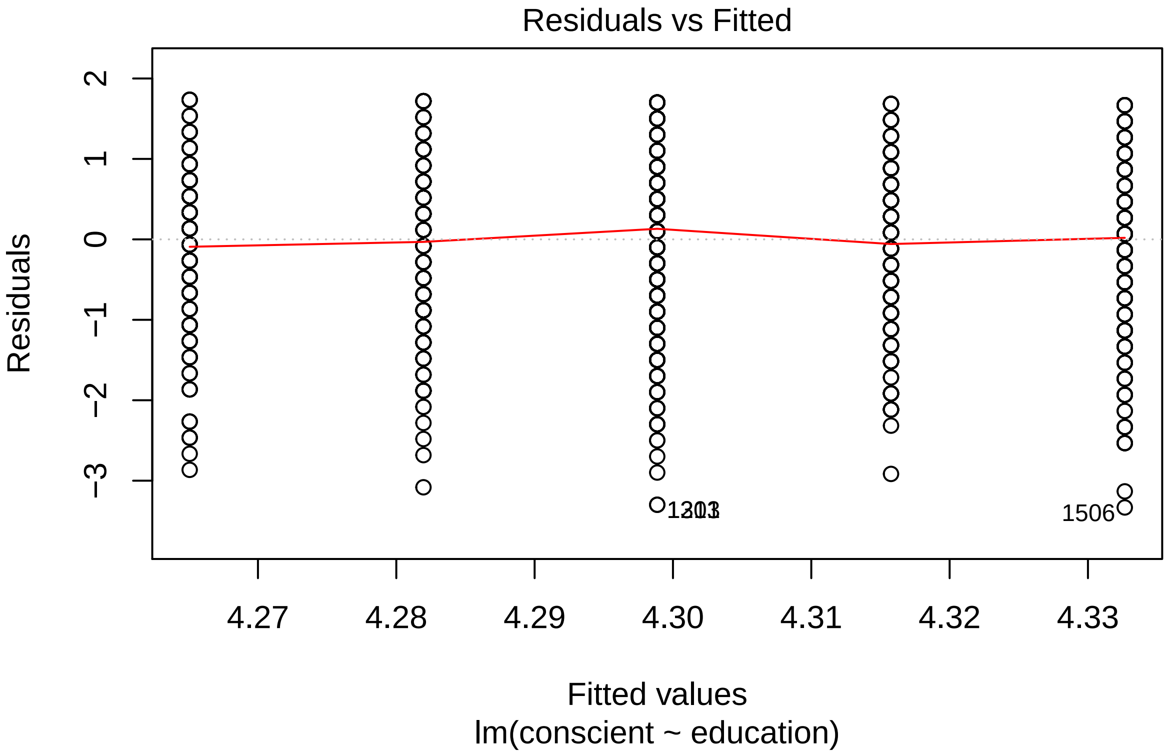
Now let's run our model using the `lm()` function.

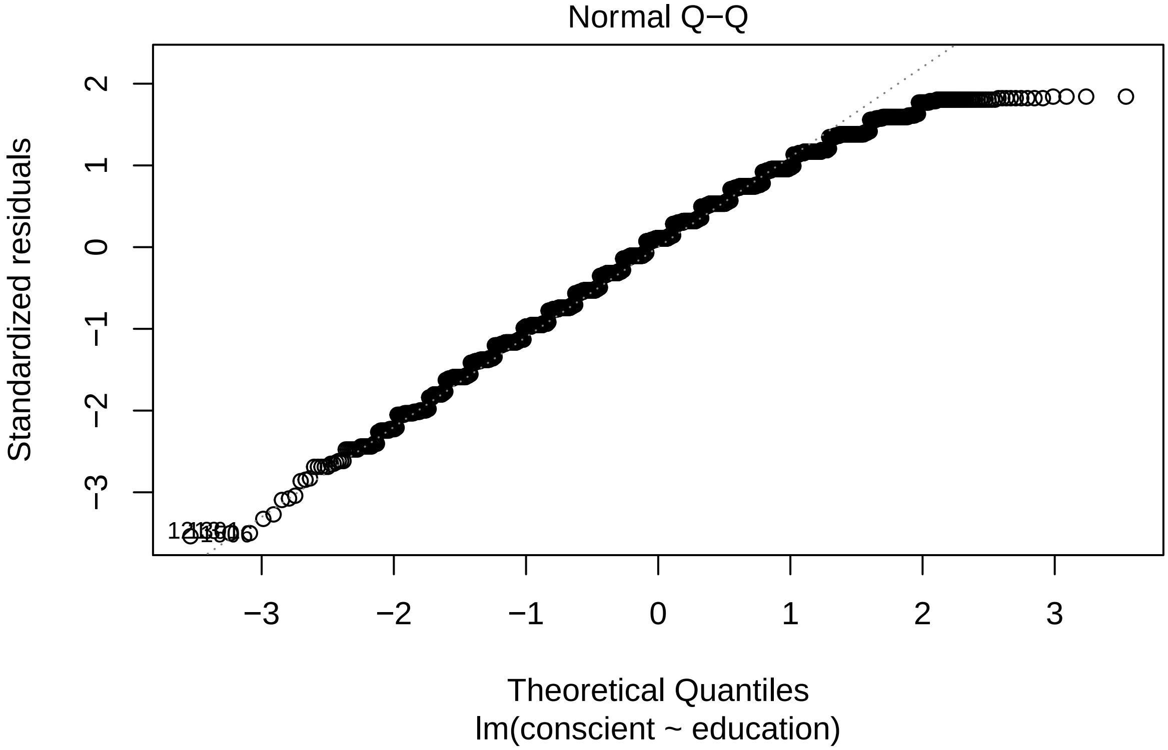
```r
mod4 <- lm(conscient ~ education, data = bfi)

summary(mod4)
```

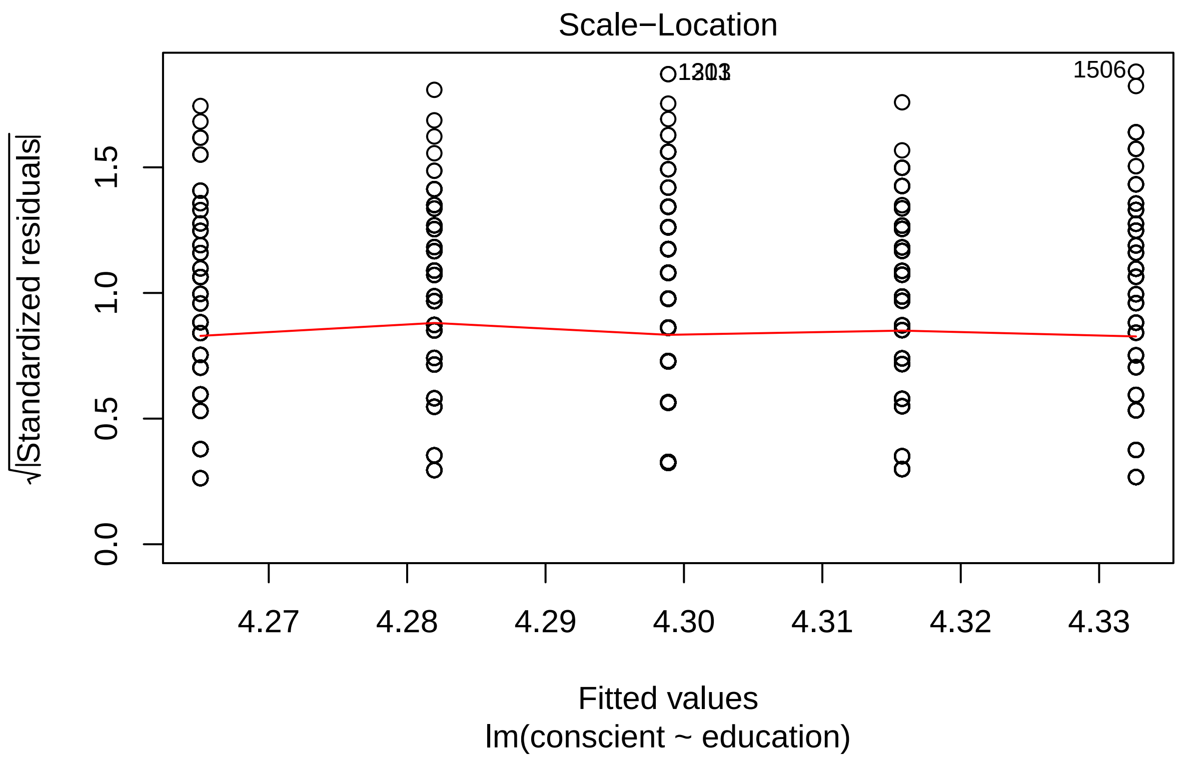
```
## 
## Call:
## lm(formula = conscient ~ education, data = bfi)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.3327 -0.6989  0.1011  0.7011  1.7349 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  4.24816    0.05758  73.778   <2e-16 ***
## education    0.01690    0.01702   0.993    0.321    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.9428 on 2488 degrees of freedom
##   (310 observations deleted due to missingness)
## Multiple R-squared:  0.0003961,  Adjusted R-squared:  -5.714e-06 
## F-statistic: 0.9858 on 1 and 2488 DF,  p-value: 0.3209
```

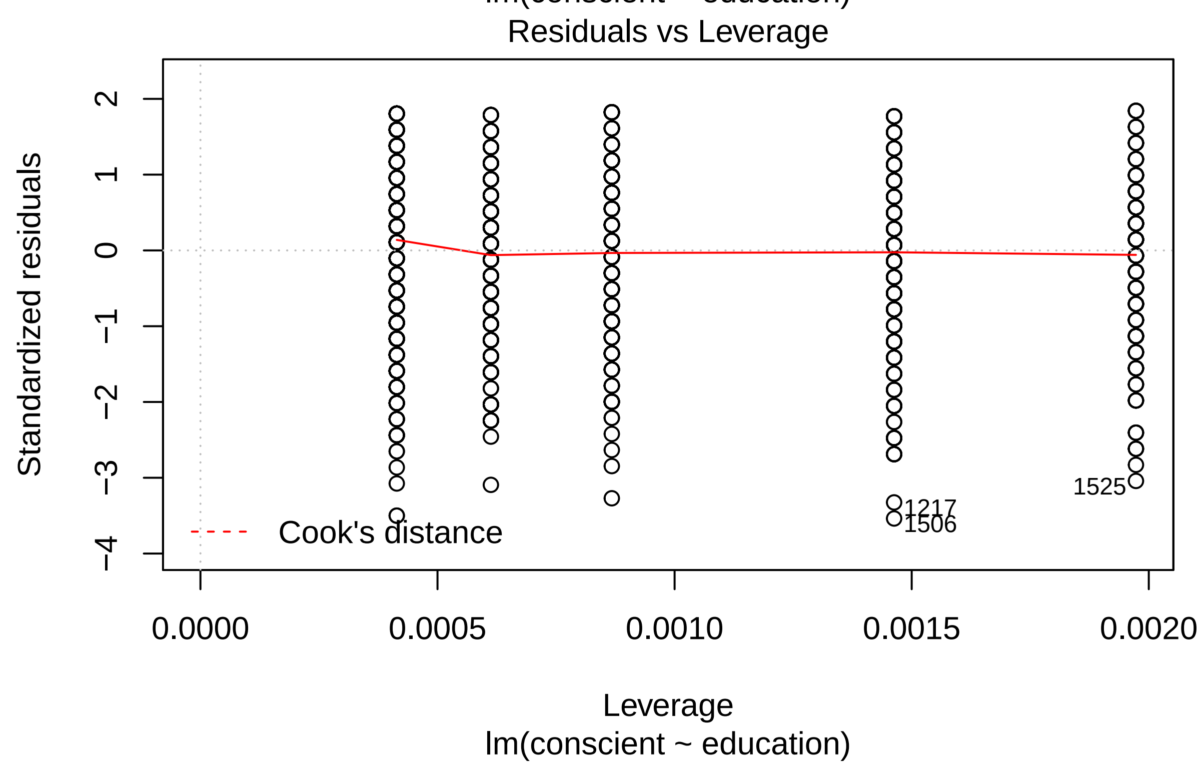
There is no statistically significant effect of education on conscientiousness using a linear regression model. Let's check our regression diagnostics. This is one of the benefits of R, we can use the `plot()` function.

```r
plot(mod4)
```


Residuals vs Fitted
Residuals
2
1
0
−1
−2
−3
1303
1506
4.27
4.28
4.29
4.30
4.31
4.32
4.33
Fitted values
lm(conscient ~ education)


Normal Q−Q
Standardized residuals
2
1
0
−1
−2
−3
1506
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(conscient ~ education)





Let's add gender as a factor to make it a multiple regression model. Note that we can also get the confidence intervals.

```
mod5 <- lm(conscient ~ education*as.factor(gender), data = bfi)
```

```r
summary(mod5)
```

```
##
## Call:
## lm(formula = conscient ~ education * as.factor(gender), data = bfi)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4229 -0.6229  0.0456  0.6456  1.8433
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.22038    0.09411  44.845   <2e-16 ***
## education                    -0.01274    0.02764  -0.461    0.645
## as.factor(gender)2            0.03122    0.11873   0.263    0.793
## education:as.factor(gender)2  0.04700    0.03499   1.343    0.179
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.939 on 2486 degrees of freedom
##   (310 observations deleted due to missingness)
## Multiple R-squared:  0.009233,   Adjusted R-squared:  0.008037
## F-statistic: 7.722 on 3 and 2486 DF,  p-value: 3.917e-05
```

```r
confint(mod5)
```

```
##                                    2.5 %     97.5 %
## (Intercept)                   4.03583456 4.40491891
## education                    -0.06693699 0.04145875
## as.factor(gender)2           -0.20159288 0.26402948
## education:as.factor(gender)2 -0.02162153 0.11562241
```

Check the residuals for `mod5`

```r
#use the plot() function here
```

Try making a figure that splits the slope of education on conscientiousness by gender. Copy and past from yesterday's code if you'd like.

```r
#regression lines split by gender
```

Logistic Regression

Because personality is relatively stable, we may instead ask are people who are more conscientious more likely to graduate from college? This would be a logistic regression model. We can use the code we wrote yesterday for creating the dichotomous variable 1 = yes college, 0 = no college.

```
bfi <- bfi %>%
  mutate(coll_grad = (education > 3))
```

Then we can run our model by using the `glm()` function and adding `family = binomial`.

```
mod6 <- glm(coll_grad ~ conscient, data = bfi, family = binomial)

summary(mod6)
```

```
##
## Call:
## glm(formula = coll_grad ~ conscient, family = binomial, data = bfi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9750  -0.8829  -0.8535   1.4818   1.5773
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.41527    0.19925  -2.084   0.0371 *
## conscient   -0.08142    0.04549  -1.790   0.0735 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3113.0  on 2489  degrees of freedom
## Residual deviance: 3109.8  on 2488  degrees of freedom
##   (310 observations deleted due to missingness)
## AIC: 3113.8
##
## Number of Fisher Scoring iterations: 4
```

```
confint(mod6)
```

```
##                  2.5 %       97.5 %
## (Intercept) -0.8074043 -0.026019218
## conscient   -0.1705590  0.007841277
```

It's handy to look at the exponentiated estimates using the `exp()` function.

```
exp(coef(mod6))
```

```
## (Intercept)   conscient
##   0.6601588   0.9218058
```

You are 0.92 times as likely to go graduate from college for every 1 unit increase in conscientiousness, but this is only marginally significant, $p = .073$.