Applied Linear Modeling

November 12, 2019

To-do

- Warm-up/discussion
  - *Review a classmate's code*
- Data, slides, and workshop file in GitHub
- R packages used today:
  - *car*
  - *sjstats*
  - *tidyverse*

Do education level and sex predict technology use?

The two-way ANOVA process:

- Exploratory data analysis
- Assumption checking
- Model development
- Post-hoc tests
- Results reporting

We will go a bit out of order for demonstration purposes.

Data for learning two-way ANOVA

- The General Social Survey (GSS) includes many variables from many years of surveys of a random sample of people in the US
- I chose a few variables in the GSS Data Explorer (https://gssdataexplorer.norc.org/variables/vfilter) and downloaded the data set
  - *Requires making a free account*
- The data (not changed from download) are available in GitHub

Importing and exploring the data

- Data are saved in a .rda file, which can be imported directly:

```
# load GSS rda file
load(file = "gss2018.rda")
```

- The data are named whatever they were named when saved, in this case “GSS”
- To change the object name, use:

```
# assign the original objected called GSS to gss.2018
gss.2018 <- GSS

# remove the original object GSS
rm(GSS)
```

Data management

- Restrict data frame to the variables of interest and recode as needed:
  - *USETECH: During a typical week, about what percentage of your total time at work would you normally spend using different types of electronic technologies (such as computers, tablets, smart phones, cash registers, scanners, GPS devices, robotic devices, and so on)?*
  - *DEGREE: highest educational attainment (< high school, high school, junior college, college, grad school)*
  - *SEX: participant sex (male, female)*

Recode to clean the variables

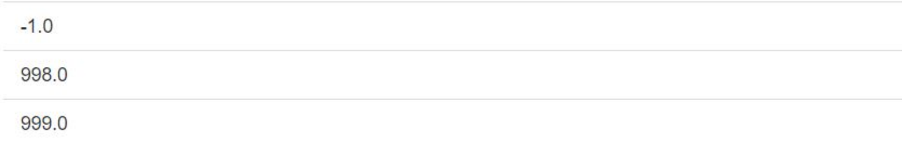
```
library(package = "tidyverse")

# select variables
gss.2018.cleaned <- gss.2018 %>%
  select(USETECH, DEGREE, SEX)

# examine the variables
summary(object = gss.2018.cleaned)
```

```
##     USETECH           DEGREE           SEX       
##  Min.   : -1.00   Min.   :0.000   Min.   :1.000  
##  1st Qu.: -1.00   1st Qu.:1.000   1st Qu.:1.000  
##  Median : 10.00   Median :1.000   Median :2.000  
##  Mean   : 48.09   Mean   :1.684   Mean   :1.552  
##  3rd Qu.: 80.00   3rd Qu.:3.000   3rd Qu.:2.000  
##  Max.   :999.00   Max.   :4.000   Max.   :2.000  
```

Recoding variables

Information on variable coding is found in the GSS Data Explorer: https://gssdataexplorer.norc.org/variables/vfilter

Missing values for USETECH:



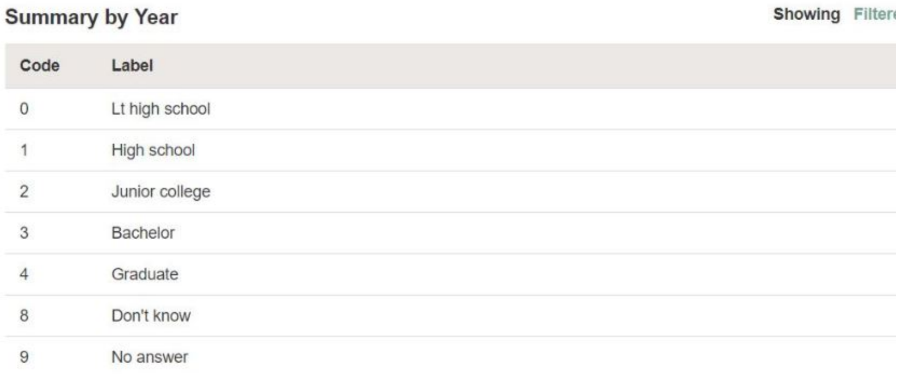
-1.0

998.0

999.0

Recoding variables

Categories for DEGREE:



Summary by Year

Code	Label
0	Lt high school
1	High school
2	Junior college
3	Bachelor
4	Graduate
8	Don't know
9	No answer

Recoding

```
# recode variables of interest to valid ranges
gss.2018.cleaned <- gss.2018 %>%
  select(SEX, DEGREE, USETECH) %>%
  mutate(USETECH = na_if(x = USETECH, y = -1)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 999)) %>%
  mutate(USETECH = na_if(x = USETECH, y = 998)) %>%
  mutate(SEX = factor(x = SEX, labels = c("male","female"))) %>%
  mutate(DEGREE = factor(x = DEGREE, labels = c("< high school",
                                               "high school", "junior college",
                                               "college", "grad school")))
# check recoding
summary(object = gss.2018.cleaned)
```

```
##      SEX                    DEGREE         USETECH
##  male  :1051   < high school : 262   Min.   :  0.00
##  female:1294   high school   :1175   1st Qu.: 15.00
##                junior college: 196   Median : 60.00
##                college       : 465   Mean   : 55.15
##                grad school   : 247   3rd Qu.: 90.00
##                                      Max.   :100.00
##                                      NA's   :936
```

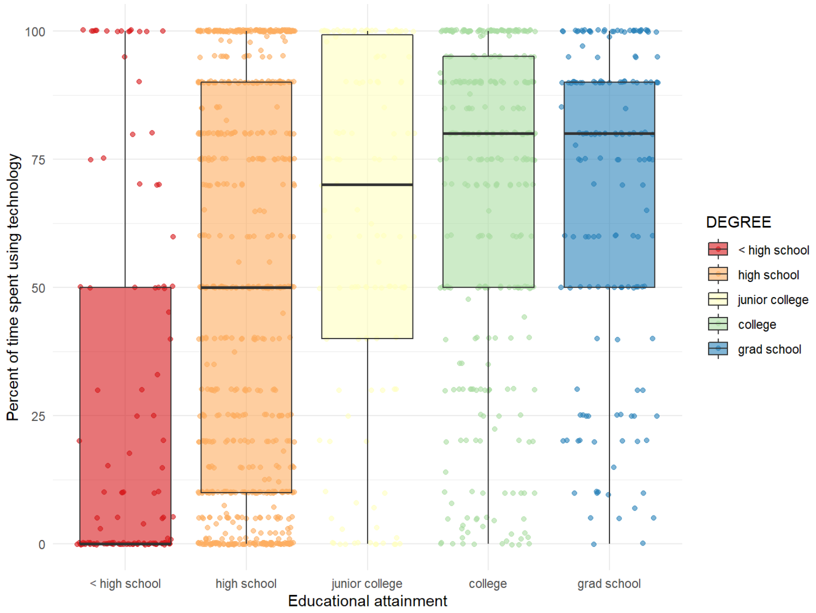
Understanding and conducting two-way ANOVA

- One-way ANOVA is useful for comparing the means of a continuous variable across 3 or more categories of a categorical variable
- What happens when there there are two categorical variables that may both be useful in explaining a continuous variable?

Examining technology use patterns by DEGREE

```
# graph usetech by degree
gss.2018.cleaned %>%
ggplot(aes(y = USETECH, x = DEGREE)) +
  geom_jitter(aes(color = DEGREE), alpha = .6) +
  geom_boxplot(aes(fill = DEGREE), alpha = .6) +
  scale_fill_brewer(palette = "Spectral") +
  scale_color_brewer(palette = "Spectral") +
  theme_minimal() +
  labs(x = "Educational attainment", y = "Percent of time spent using technology")
```

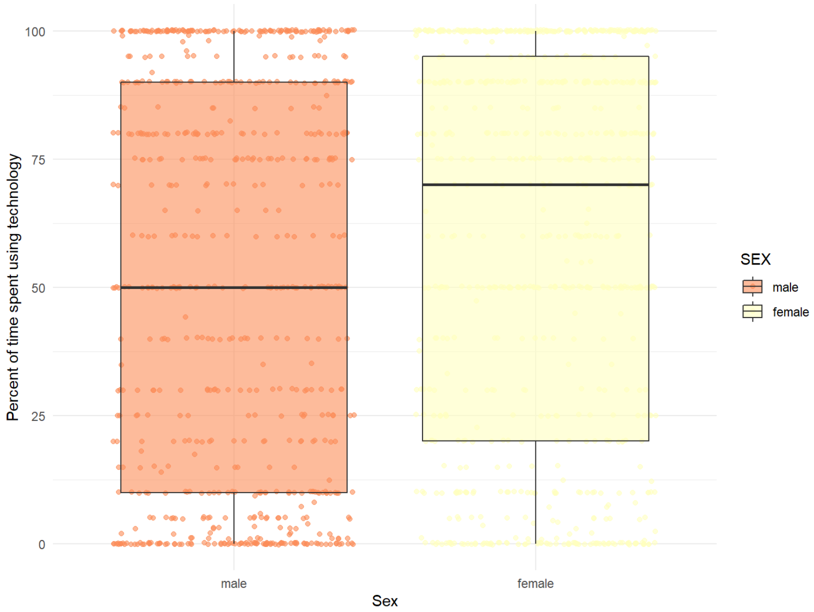
Examining technology use patterns by DEGREE



Examining technology use patterns by SEX

```
# graph usetech by sex
gss.2018.cleaned %>%
ggplot(aes(y = USETECH, x = SEX)) +
  geom_jitter(aes(color = SEX), alpha = .6) +
  geom_boxplot(aes(fill = SEX), alpha = .6) +
  scale_fill_brewer(palette = "Spectral") +
  scale_color_brewer(palette = "Spectral") +
  theme_minimal() +
  labs(x = "Sex", y = "Percent of time spent using technology")
```

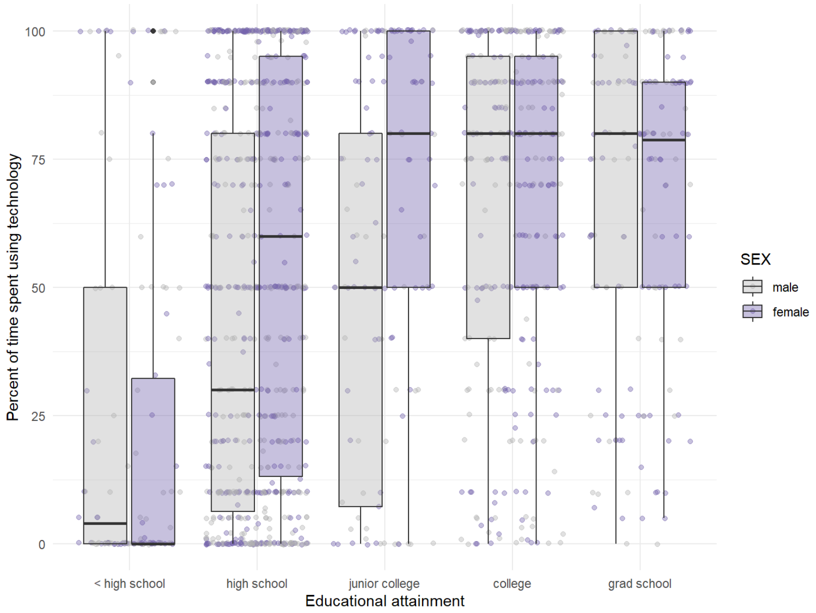
Examining technology use patterns by SEX



Exploratory data analysis for two-way ANOVA

```
# graph usetech by degree
gss.2018.cleaned %>%
ggplot(aes(y = USETECH, x = DEGREE)) +
  geom_jitter(aes(color = SEX), alpha = .4) +
  geom_boxplot(aes(fill = SEX), alpha = .4) +
  scale_fill_manual(values = c("gray70", "#7463AC")) +
  scale_color_manual(values = c("gray70", "#7463AC")) +
  theme_minimal() +
  labs(x = "Educational attainment", y = "Percent of time spent using technology")
```

Exploratory data analysis for two-way ANOVA



Interpreting the plot

- There is a different pattern of technology use for males and females
- Females with less than a high school degree were using technology a lower percentage of the time than males in this group
- Females use technology more of the time compared to the males in the high school group and for the junior college group
- Males and females seem to have relatively equal time spent with technology once a bachelor or graduate degree is earned

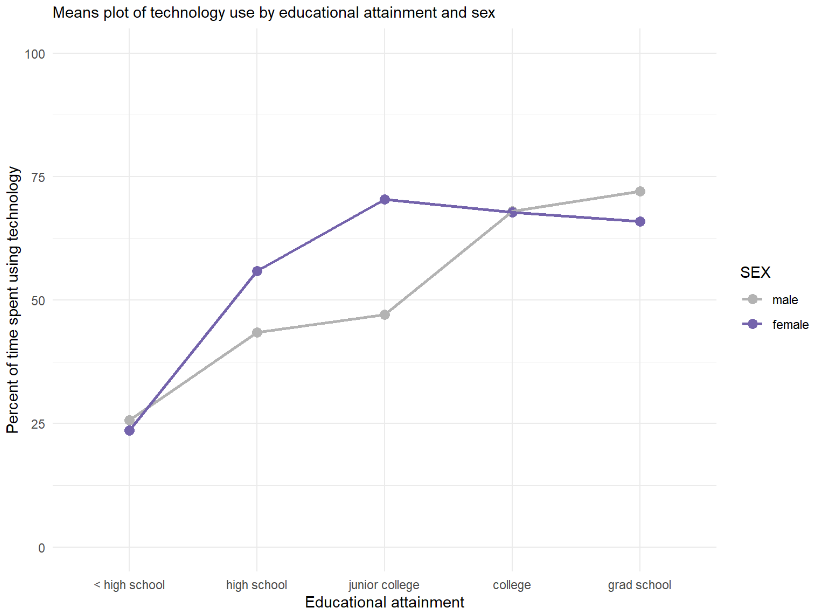
Checking for an interaction

- A pattern of differences where one variable behaves differently across the levels of another variable is consistent with an **interaction**

- A traditional **means plot** is often used to visualize the idea of an interaction

```
# means plots graph
gss.2018.cleaned %>%
  ggplot(aes(y = USETECH, x = DEGREE, color = SEX)) +
  stat_summary(fun.y = mean, geom="point", size = 3) +
  stat_summary(fun.y = mean, geom="line", aes(group = SEX), size = 1) +
  scale_color_manual(values = c("gray70", "#7463AC")) +
  theme_minimal() +
  labs(x = "Educational attainment",
       y = "Percent of time spent using technology",
       subtitle = "Means plot of technology use by educational attainment and sex") +
  ylim(0, 100)
```

Checking for an interaction



Interpreting a means plot

- When the lines in means plots like this one are parallel, it indicates that the mean of the continuous variable is consistently higher or lower for certain groups compared to others
- When the means plot shows lines that cross or diverge, this indicates that there is an interaction between the categorical variables
  - *The mean of the continuous variable is different at different levels of one categorical variable depending on the value of the other categorical variable*
- For example, mean technology use is lower for females compared to males for the lowest and highest educational attainment categories, but female technology use is higher than male technology use for the three other categories of educational attainment
- The two variables are working *together* to influence the value of technology use

Checking the group means

```
# means by degree and sex
use.stats.2 <- gss.2018.cleaned %>%
  group_by(DEGREE, SEX) %>%
  drop_na(USETECH) %>%
  summarize(m.techuse = mean(USETECH),
            sd.techuse = sd(USETECH))
use.stats.2
```

```
## # A tibble: 10 x 4
## # Groups:   DEGREE [5]
##    DEGREE         SEX    m.techuse sd.techuse
##    <fct>          <fct>      <dbl>      <dbl>
##  1 < high school  male        25.7       35.4
##  2 < high school  female      23.7       37.5
##  3 high school    male        43.5       37.8
##  4 high school    female      55.9       38.6
##  5 junior college male        47.0       36.8
##  6 junior college female      70.4       31.7
##  7 college        male        68.0       33.1
##  8 college        female      67.7       31.2
##  9 grad school    male        72.1       29.2
## 10 grad school    female      65.9       30.9
```

Conduct the ANOVA

- Since this is not a one-way ANOVA, a different function is used, `aov()`

```
# two-way ANOVA technology use by degree and sex
techuse.by.deg.sex <- aov(formula = USETECH ~ DEGREE + SEX + DEGREE * SEX,
                          data = gss.2018.cleaned)
summary(techuse.by.deg.sex)
```

```
##                 Df  Sum Sq Mean Sq F value   Pr(>F)    
## DEGREE           4  221301   55325  44.209  < 2e-16 ***
## SEX              1   16473   16473  13.163 0.000296 ***
## DEGREE:SEX       4   26510    6627   5.296 0.000311 ***
## Residuals     1399 1750775    1251                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 936 observations deleted due to missingness
```

Examining the output

- There are three F-statistics for this ANOVA
  - *One for each of the two individual variables, also called the main effects*
  - *One for the interaction term*

Interpreting the main effects

- A main effect is the relationship between only one of the independent variables and the dependent variable, ignoring the impact of any additional independent variables or interaction effects
- When the interaction term is significant, it is important to examine the nature of the interaction before interpreting main effects to see if the main effects can or should be interpreted clearly

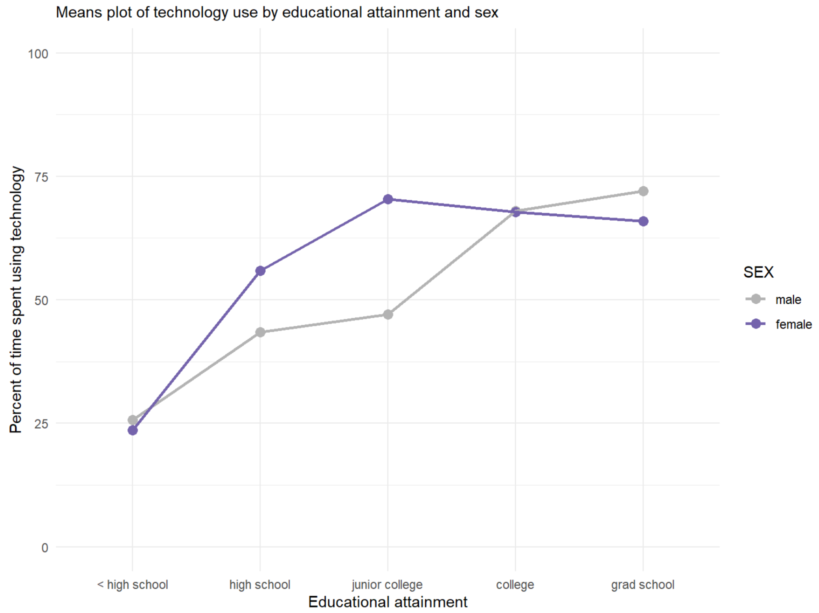
Interpreting the interaction

- The interaction is the relationship between the two variables taken *together*
- Are the group means different from one another when each group is a unique combination of DEGREE and SEX (e.g., high-school males, < high school females)

There was a statistically significant interaction between degree and sex on technology use [F(4,1399) = 5.3; p < .001].

Can we interpret the significant main effects?

- Are there clear patterns of difference in tech use by education **alone** or by sex **alone**



ANOVA effect size (model fit)

- Omega-squared (ω^2) is the proportion of variability in data that is explained by the group means (like the R^2)

- Effect sizes are interpreted:
  - *ω^2 = .01 to < .06 is a small effect*
  - *ω^2 = .06 to < .14 is a medium effect*
  - *$\omega^2 \geq$ .14 is a large effect*

```
# get effect size
library(package = "sjstats")
omega_sq(model = techuse.by.deg.sex)
```

```
##         term omegasq
## 1     DEGREE   0.107
## 2        SEX   0.008
## 3 DEGREE:SEX   0.011
```

Write a conclusion

There was a statistically significant interaction between degree and sex on mean percent of time using technology at work [F(4,1399) = 5.3; p < .001]. The effect size of the interaction was small (ω^2 = .01). The highest mean was 72.06% of time used for technology for males with graduate degrees. The lowest mean was 23.67% of the time for females with less than a high school diploma. The interaction between degree and sex shows that time spent on technology use increases more quickly for females with both males and females eventually having high tech use in the top two educational attainment groups.

Post-hoc test for two-way ANOVA

- The Bonferroni post-hoc test is not available in R for two-way ANOVA and contrasts are very complex and rarely used

- Instead use Tukey's HSD ("Honestly Significant Difference") post-hoc test

  - *Tukey's HSD compares the means of each pair of two groups using the same formula as a t-test, but then uses a different distribution (the **q** distribution) to find the p-value*

  - *The **q** distribution requires a larger value of the test-statistic in order to reach statistical significance, so means have to be further apart to be significant with Tukey's HSD than with a t-test (so the Type I error is not inflated)*

```
# Tukey's HSD post-hoc test
TukeyHSD(x = techuse.by.deg.sex)
```

The output

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = USETECH ~ DEGREE + SEX + DEGREE * SEX, data = gss.2018.cleaned)
##
## $DEGREE
##                                   diff       lwr       upr     p adj
## high school-< high school    24.8247754 15.244768 34.404783 0.0000000
## junior college-< high school 37.6070312 25.329478 49.884584 0.0000000
## college-< high school        43.0859568 32.760484 53.411429 0.0000000
## grad school-< high school    43.9107249 32.376284 55.445165 0.0000000
## junior college-high school   12.7822558  3.459487 22.105024 0.0017563
## college-high school          18.2611813 11.719691 24.802671 0.0000000
## grad school-high school      19.0859494 10.766152 27.405746 0.0000000
## college-junior college        5.4789255 -4.608337 15.566188 0.5733923
## grad school-junior college    6.3036936 -5.018002 17.625389 0.5490670
## grad school-college           0.8247681 -8.343540  9.993076 0.9991960
##
## $SEX
##                diff      lwr      upr     p adj
## female-male 6.80899 3.108699 10.50928 0.0003174
##
## $`DEGREE:SEX`
##                                                 diff          lwr
## high school:male-< high school:male         17.8132060   2.7275183
## junior college:male-< high school:male      21.3181818  -0.4992077
## college:male-< high school:male             42.3151914  25.7902764
## grad school:male-< high school:male         46.3538961  27.5496712
## < high school:female-< high school:male     -2.0378788 -22.6075109
## high school:female-< high school:male       30.1500000  15.0344692
## junior college:female-< high school:male    44.7418831  26.3028236
## college:female-< high school:male           42.0396406  25.8082011
## grad school:female-< high school:male       40.1813241  22.0984520
## junior college:male-high school:male         3.5049758 -14.4610385
## college:male-high school:male               24.5019854  13.5542915
## grad school:male-high school:male           28.5406901  14.3851943
## < high school:female-high school:male      -19.8510848 -36.2793820
## high school:female-high school:male         12.3367940   3.6616307
## junior college:female-high school:male      26.9286771  13.2619985
## college:female-high school:male             24.2264346  13.7269673
## grad school:female-high school:male         22.3681181   9.1859540
## college:male-junior college:male            20.9970096   1.8065820
## grad school:male-junior college:male        25.0357143   3.8508477
## < high school:female-junior college:male   -23.3560606 -46.1224714
## high school:female-junior college:male       8.8318182  -9.1592621
## junior college:female-junior college:male   23.4237013   2.5622868
## college:female-junior college:male          20.7214588   1.7831557
## grad school:female-junior college:male      18.8631423  -1.6841193
## grad school:male-college:male                4.0387047 -11.6416301
## < high school:female-college:male          -44.3530702 -62.1121183
## high school:female-college:male            -12.1651914 -23.1539720
```

```
## junior college:female-college:male          2.4266917 -12.8138117
## college:female-college:male                -0.2755508 -12.7548798
## grad school:female-college:male            -2.1338673 -16.9414427
## < high school:female-grad school:male     -48.3917749 -68.2892584
## high school:female-grad school:male       -16.2038961 -30.3911918
## junior college:female-grad school:male     -1.6120130 -19.2981376
## college:female-grad school:male            -4.3142555 -19.6849976
## grad school:female-grad school:male        -6.1725720 -23.4870269
## high school:female-< high school:female    32.1878788  15.7321731
## junior college:female-< high school:female 46.7797619  27.2270154
## college:female-< high school:female        44.0775194  26.5912218
## grad school:female-< high school:female    42.2192029  23.0019908
## junior college:female-high school:female   14.5918831   0.8922699
## college:female-high school:female          11.8896406   1.3473395
## grad school:female-high school:female      10.0313241  -3.1849820
## college:female-junior college:female       -2.7022425 -17.6240305
## grad school:female-junior college:female   -4.5605590 -21.4777217
## grad school:female-college:female          -1.8583165 -16.3376501
##                                                   upr     p adj
## high school:male-< high school:male        32.8988937 0.0072699
## junior college:male-< high school:male     43.1355713 0.0619111
## college:male-< high school:male            58.8401064 0.0000000
## grad school:male-< high school:male        65.1581210 0.0000000
## < high school:female-< high school:male    18.5317533 0.9999995
## high school:female-< high school:male      45.2655308 0.0000000
## junior college:female-< high school:male   63.1809427 0.0000000
## college:female-< high school:male          58.2710800 0.0000000
## grad school:female-< high school:male      58.2641962 0.0000000
## junior college:male-high school:male       21.4709901 0.9998264
## college:male-high school:male              35.4496792 0.0000000
## grad school:male-high school:male          42.6961858 0.0000000
## < high school:female-high school:male      -3.4227876 0.0052315
## high school:female-high school:male        21.0119573 0.0003049
## junior college:female-high school:male     40.5953557 0.0000000
## college:female-high school:male            34.7259018 0.0000000
## grad school:female-high school:male        35.5502821 0.0000039
## college:male-junior college:male           40.1874372 0.0192892
## grad school:male-junior college:male       46.2205808 0.0071871
## < high school:female-junior college:male   -0.5896498 0.0389231
## high school:female-junior college:male     26.8228985 0.8690307
## junior college:female-junior college:male  44.2851158 0.0141081
## college:female-junior college:male         39.6597618 0.0192858
## grad school:female-junior college:male     39.4104039 0.1039186
## grad school:male-college:male              19.7190396 0.9983501
## < high school:female-college:male         -26.5940220 0.0000000
## high school:female-college:male            -1.1764108 0.0167764
## junior college:female-college:male        17.6671952 0.9999688
## college:female-college:male                12.2037783 1.0000000
## grad school:female-college:male            12.6737082 0.9999867
## < high school:female-grad school:male     -28.4942914 0.0000000
## high school:female-grad school:male        -2.0166004 0.0113631
## junior college:female-grad school:male     16.0741116 0.9999998
## college:female-grad school:male            11.0564866 0.9967894
## grad school:female-grad school:male        11.1418829 0.9816675
## high school:female-< high school:female    48.6435845 0.0000000
## junior college:female-< high school:female 66.3325084 0.0000000
## college:female-< high school:female        61.5638170 0.0000000
## grad school:female-< high school:female     61.4364150 0.0000000
## junior college:female-high school:female    28.2914963 0.0261888
## college:female-high school:female           22.4319416 0.0133486
## grad school:female-high school:female       23.2476303 0.3233313
## college:female-junior college:female        12.2195454 0.9999069
## grad school:female-junior college:female    12.3566037 0.9976459
## grad school:female-college:female           12.6210171 0.9999951
```

Interpreting the output

- The first section under `$DEGREE` was comparing groups of `DEGREE` to each other
- The second section under `$SEX` was comparing males and females to each other
- The third section was the interaction, comparing groups of `DEGREE·SEX` with each other
- For example, the first row in this last section is *high school:male-< high school:male*, which compares *high school male* to *< high school male*
  - *The numbers that follow are the difference between the means (diff = 17.81), the confidence interval around the difference (95% CI: 2.73 to 32.90), and the p-value for the difference between the means (p = 0.007)*

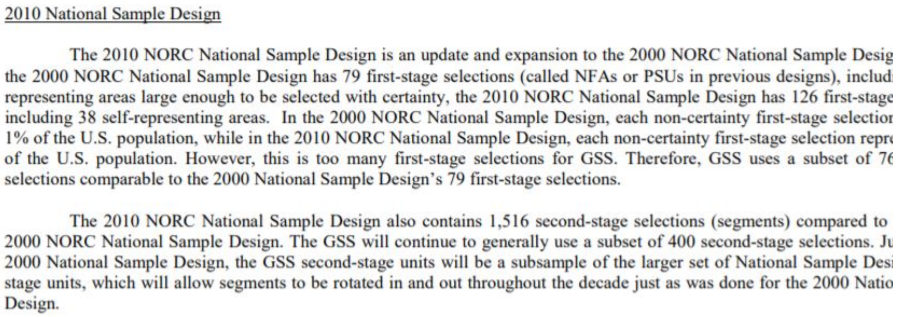
There is a statistically significant (p < .05) difference of 17.81 between the mean percentage time of technology use for males with less than high school (25.70%) compared to males with high school (43.52%) in the sample. The confidence interval shows that difference between the means of these two groups is somewhere between 2.73 and 32.90 in the population this sample came from.

Two-way ANOVA assumptions

- Same assumptions
  - *Continuous outcome and independent groups*
  - *Independent observations*
  - *Normality within groups*
  - *Equal variances within groups*

Data types and indepence assumptions

- Percent time tech use is continuous
- Degree groups are independent because nobody is in more than one of these groups; they do not overlap or conflict
- Independent observations can be checked via the GSS website:



2010 National Sample Design

The 2010 NORC National Sample Design is an update and expansion to the 2000 NORC National Sample Desig
the 2000 NORC National Sample Design has 79 first-stage selections (called NFAs or PSUs in previous designs), includ
representing areas large enough to be selected with certainty, the 2010 NORC National Sample Design has 126 first-stage
including 38 self-representing areas. In the 2000 NORC National Sample Design, each non-certainty first-stage selectior
1% of the U.S. population, while in the 2010 NORC National Sample Design, each non-certainty first-stage selection repr
of the U.S. population. However, this is too many first-stage selections for GSS. Therefore, GSS uses a subset of 7
selections comparable to the 2000 National Sample Design's 79 first-stage selections.

The 2010 NORC National Sample Design also contains 1,516 second-stage selections (segments) compared to
2000 NORC National Sample Design. The GSS will continue to generally use a subset of 400 second-stage selections. Ju
2000 National Sample Design, the GSS second-stage units will be a subsample of the larger set of National Sample Des
stage units, which will allow segments to be rotated in and out throughout the decade just as was done for the 2000 Natio
Design.

Testing the normality assumption

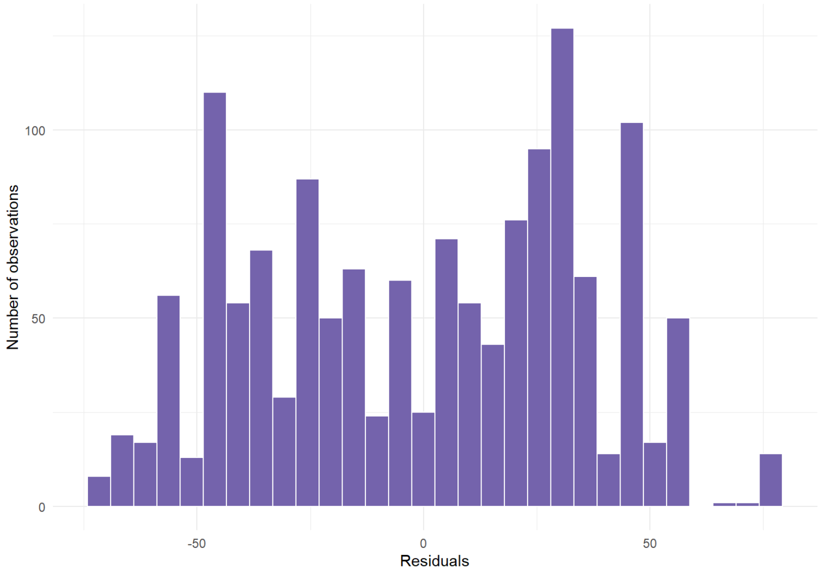
- With the enormous number of distinct groups, the normality assumption can be checked by examining residuals
  - *Normally distributed residuals indicate the differences between the observed values and the predicted/group means have a normal distribution*
  - *Normally distributed residuals would meet the assumption*

```
# make a data frame
tech.deg.sex <- data.frame(techuse.by.deg.sex$residuals)

# plot the residuals
tech.deg.sex %>%
ggplot(aes(x = techuse.by.deg.sex.residuals)) +
  geom_histogram(fill = "#7463AC", col = "white") +
  theme_minimal() +
  labs(x = "Residuals",
       y = "Number of observations",
       subtitle = "Distribution of residuals from ANOVA explaining tech use\nbased on educational
         attainment and sex (GSS, 2018)")
```

Checking residuals

- Well, rats

Distribution of residuals from ANOVA explaining tech use
based on educational attainment and sex (GSS, 2018)



Testing the homogeneity of variances assumption

- Use the `leveneTest()` function to test the null hypothesis that the variances are equal
- Rats, again :-(

```
# Levene test for ANOVA
library(package = "car")
leveneTest(y = USETECH ~ DEGREE*SEX,
                data = gss.2018.cleaned)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)    
## group    9  7.1573 3.324e-10 ***
##       1399                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Alternatives when ANOVA assumptions fail

- Report descriptive and visual statistics only
- Report ANOVA results but do not generalize outside the sample
- Recode or transform the outcome variable and try again or try a different type of model
- Use an alternate test:
  - *One-way ANOVA assumptions fail*
    - Use Welch's or Brown-Forsythe when normality is met but equal variances fails
    - Use Kruskal-Wallis to examine ranks across groups when normality fails (regardless of equal variances)
  - *Two-way ANOVA assumptions fail*
    - Transform the outcome into ranks (like Kruskal-Wallis does) and conduct the ANOVA on the ranks

Example of conclusion without generalizing

There was a statistically significant interaction between degree and sex on mean percent of time using technology at work [F(4,1399) = 5.3; p < .001]. The effect size of the interaction was small (ω^2 = .01). The highest mean was 72.06% of time used for technology for males with graduate degrees. The lowest mean was 23.67% of the time for females with less than a high school diploma. The interaction between degree and sex shows that time spent on technology use increases more quickly for females with both males and females eventually having high tech use in the top two educational attainment groups. The ANOVA failed the assumptions of normally distributed residuals and equal variances in groups, so the results should not be generalized outside the sample.

Exercise 9

Use the GSS data and the appropriate tests to examine technology use by happiness and sex.

- Create your own R Markdown file
- Conduct this analysis:
  - *Open the data*
  - *Clean the happiness, sex, and tech use variables so they have clear variable names, category labels, and missing value coding (See GSS Data Explorer for category labels if needed)*
  - *Select the clean happiness, sex, and tech use variables into a smaller data frame*
  - *Use graphics and descriptive statistics to examine tech use by sex and happiness*
  - *Check assumptions for a two-way ANOVA of tech use by happiness and sex*
  - *Even if you do not meet assumptions:*
    - Conduct a two-way ANOVA with technology use by sex and happiness
    - Compute the effect size and conduct post-hoc analysis
- Format your results into a short report that reports the results, showing **only**:
  - *A figure comparing group means*
  - *A table of group means and standard deviations (Use tidyverse or tableone, your choice)*

  - *One or two paragraphs explaining your descriptive and ANOVA results (including any assumption checking and post-hoc testing)*
  - *Hide all R chunks using “echo = FALSE” in the code chunk curly braces {r}*
  - *Hide any results from R that are not needed for the final report by using “results = FALSE” in the code chunk {r}*
  - *Supress warnings and messages from code chunks when needed using “warning = FALSE, message = FALSE” in the code chunk {r}*

- Upload to Canvas anytime before class on **December 3**