UBC Library Research Commons

Workshop: R for Statistical Analysis

Data Analysis Team:

- Matty Jullamon (GAA)
- Amir Michalovich (GAA)
- Jeremy Buhler (Data Librarian)
- Sarah Parker (Data Librarian)

Pre-workshop setup

Download and install R

For Windows:

1. Visit R Project (https://www.r-project.org/) to learn about R versions.
2. Download and install R from your preferred CRAN mirror here (https://cran.r-project.org/mirrors.html)
    A. Choose "0-Cloud" or a mirror site near you.

For Mac:

1. Check that your macOS system is up-to-date
2. Download and install R from The Comprehensive R Archive Network (https://cran.r-project.org/)

Download and install R studio

For Windows and Mac:

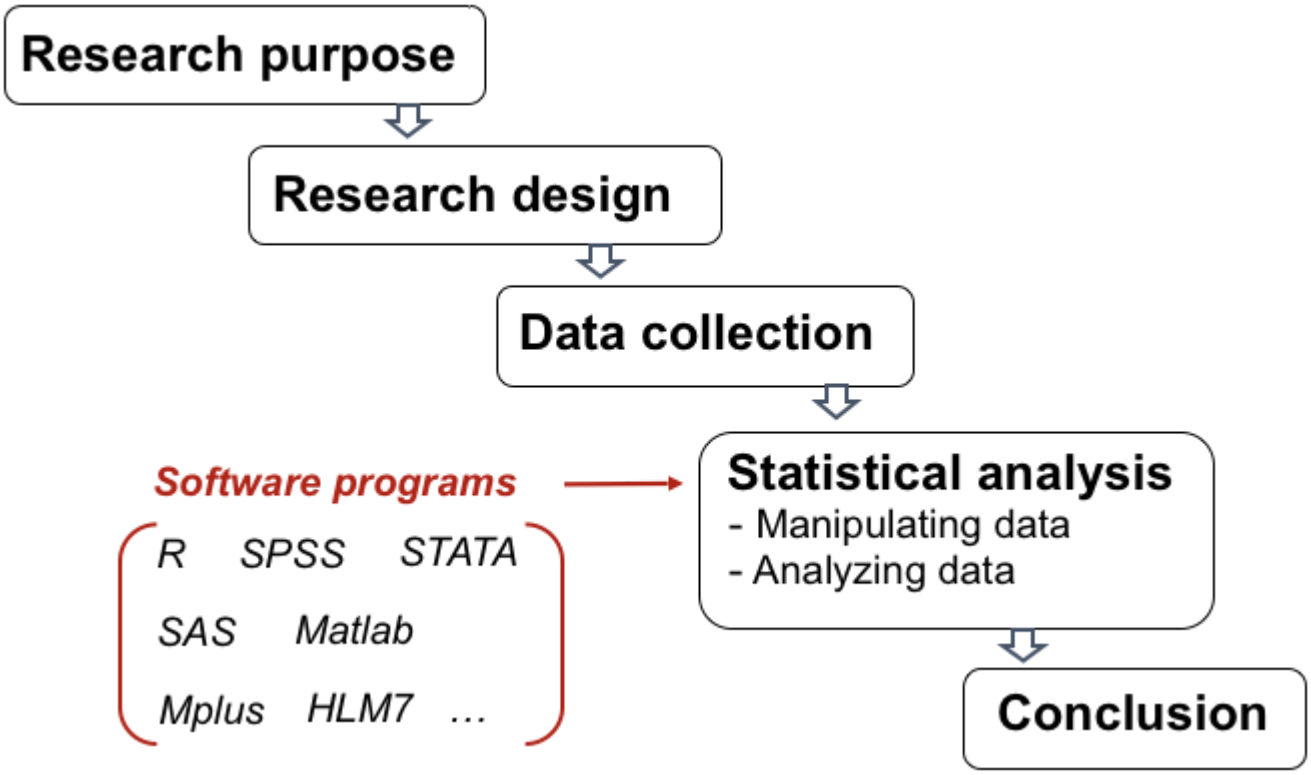
1. Download and install R Studio from here (https://rstudio.com/products/rstudio/download/#download)

Learning Objectives

- Learn how to identify the types of variables in R
- Learn the basic commands for descriptive statistics
- Learn the basic commands for inference statistics

Overview of Quantative Research Process

A systamatic research process that involves collecting objective, measureable data, using statistics to analyze the data, and generalizing the results to a larger population to explain a phenomena. Usually, software programs assist on data analysis.



Data Analysis in Quantitative Research

Definitions

- Data refers to facts or pieces of information that can either be quantitative or qualitative.
- Variable refers to any property that can be observed or measured.

Types of Variables

It is important to understand the different types of variables because they will determine the statistical analysis method.

Type	Description	Example
Nominal	Labels or Descriptions that cannot be ordered	Gender
Ordinal	Labels or Descriptions that can be ordered	Education Level
Interval	Numeric values with equal magnitude, doesn't have absolute zero	SAT scores
Ratio	Numeric values with equal magnitude, does have absolute zero	Age

Categorize these variables in R

Nominal/Ordinal -> Character or Factor

Intervel/Ratio -> Numeric or Integer

Definitions

- Character: Text
- Factor: Integer associated with a specific category
- Numeric: Number with decimal point
- Integer: Number with no decimal point

Getting Started

Set working directory in R studio

You can set the working directory using **Session** > **Set Working Directory** > **Choose Directory**.

Loading a built-in R dataset

About the data

3 Measures Of Ability: SATV, SATQ, ACT: "Self reported scores on the SAT Verbal, SAT Quantitative and ACT were collected as part of the Synthetic Aperture Personality Assessment (SAPA) web based personality assessment project. Age, gender, and education are also reported. The data from 700 subjects are included here as a demonstration set for correlation and analysis" (Revelle et al., 2009). (https://www.rdocumentation.org/packages/psych/versions/1.9.12.31/topics/sat.act)

Format

A data frame with 700 observations on the following 6 variables.

`gender`

males = 1, females = 2

`education`

self reported education 1 = high school ... 5 = graduate work

`age`

age

`ACT`

ACT composite scores may range from 1 - 36. National norms have a mean of 20.

`SATV`

SAT Verbal scores may range from 200 - 800.

`SATQ`

SAT Quantitative scores may range from 200 - 800

Write and run the following commands to load the dataset

```
install.packages("psych") #install this if you haven't done so.

library(psych)

scores <- sat.act
```

In [ ]:
```
scores <-read.csv("sat.act.csv")
```

Identifying and Renaming Variables

str(df): To check the structure of your data

```
> str(scores)
'data.frame':   700 obs. of  6 variables:
 $ gender   : int  2 2 2 1 1 1 2 1 2 2 ...
 $ education: int  3 3 3 4 2 5 5 3 4 5 ...
 $ age      : int  19 23 20 27 33 26 30 19 23 40 ...
 $ ACT      : int  24 35 21 26 31 28 36 22 22 35 ...
 $ SATV     : int  500 600 480 550 600 640 610 520 400 730 ...
 $ SATQ     : int  500 500 470 520 550 640 500 560 600 800 ...
> 
```

Question: What do you notice?

as.factor(df$columnname): To change a variable to factor

In [64]:
```
scores$gender <- as.factor(scores$gender)
```

is.factor(df$columnname): To check if a variable is defined as factor

In [62]:
```
is.factor(scores$gender)
```

TRUE

Extra information

is.integer(df$columnname): To check if a variable is defined as integer

is.numeric(df$columnname): To check if a variable is defined as numeric

is.character(df$columnname): To check if a varible is defined as character

as.integer(df$columnname): To change a variable to integer

as.numeric(df$columnname): To change a variable to numeric

as.character($columnndfame): To change a variable to character

Exercise #1

- Using as.factor command, change 'education' to factor.
- Using is.factor command, check if 'education is defined as factor.

Answer to Exercise #1

In [75]:
```
scores$education <- as.factor(scores$education)
```

In [76]:
```
is.factor(scores$education)
```

TRUE

Descriptive Statistics

Descriptive statistics summarize the data in a meaningful way. The purpose of using descriptive statistics is to explore the observed data and not to draw inferences.

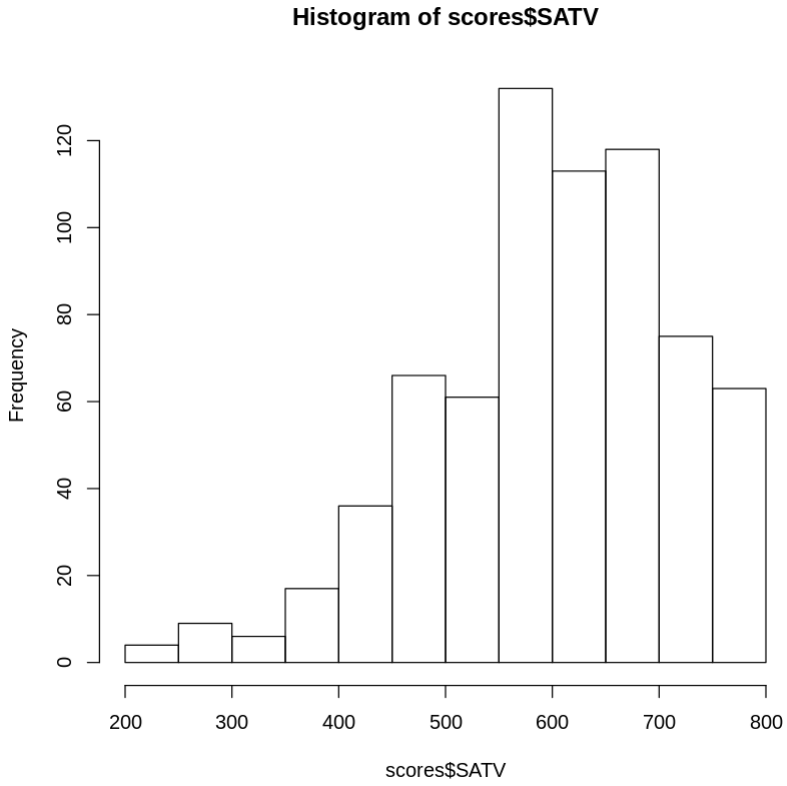
We will use the psych package functions to perform descriptive statistics.

describe(df): To obtain descriptive statistics for all variables

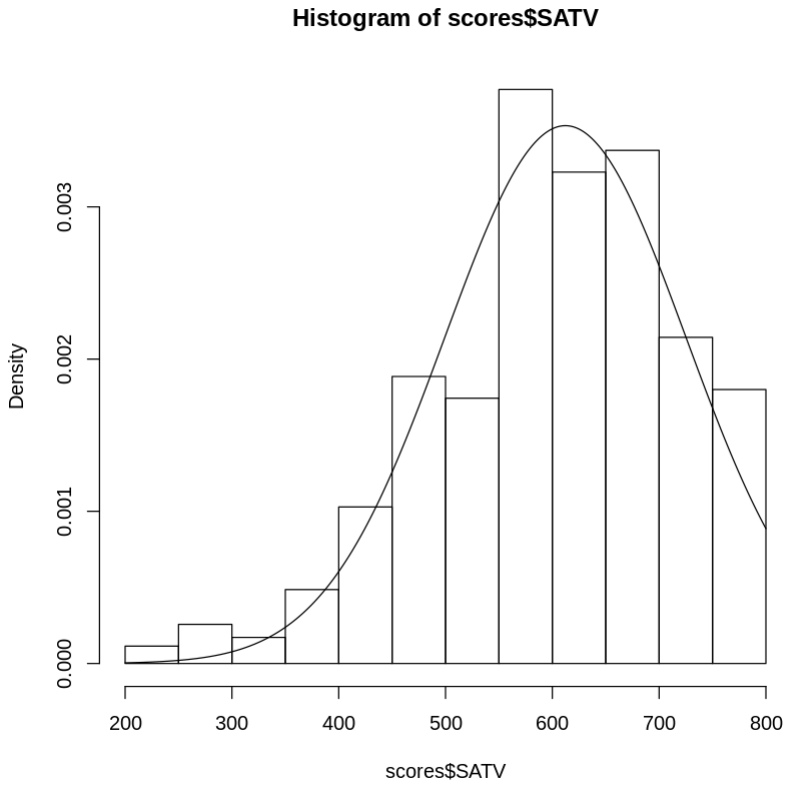
```
> describe(scores)
          vars   n   mean     sd median trimmed    mad min max range  skew kurtosis   se
gender       1 700   1.65   0.48      2    1.68   0.00   1   2     1 -0.61    -1.62 0.02
education    2 700   3.16   1.43      3    3.31   1.48   0   5     5 -0.68    -0.07 0.05
age          3 700  25.59   9.50     22   23.86   5.93  13  65    52  1.64     2.42 0.36
ACT          4 700  28.55   4.82     29   28.84   4.45   3  36    33 -0.66     0.53 0.18
SATV         5 700 612.23 112.90    620  619.45 118.61 200 800   600 -0.64     0.33 4.27
SATQ         6 687 610.22 115.64    620  617.25 118.61 200 800   600 -0.59    -0.02 4.41
> 
```

hist(df$columnname): To graphically descibe the distribution using histogram

In [78]:
```
hist(scores$SATV)
```



In [95]:
```
# To add the distribution curve

hist(scores$SATV, freq = F)
x <- 200:800
y <- dnorm(x = x, mean = 612.23, sd = 112.90)
lines(x = x, y = y)
```



Inferential Statistics

Unlike descriptive statistics, inferential statistics use the observed data to make inferences about the population.

In this workshop, we will cover four parametric tests: Independent t-test, One-way ANOVA, Pearson's correlation, & Simple linear regression. These tests are called parametric because they meet the assumptions of probability distribution.

Model assumptions

Common model assumptions found in parametric tests:

1. Independence
2. Normality
3. Equal variance

Both Pearson's correlation and Simple linear regression have some additional assumptions. For more information, click on the following links:

Pearson's correlation assumptions (https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/)

Simple linear regression assumptions (https://www.statisticssolutions.com/assumptions-of-linear-regression/)

Independent T-test

It is used to see whether there are group difference in numeric data between two groups.

For example, do males and females have different average SAT verbal scores?

H0: Mean SATV for males = Mean SATV for females

H1: Mean SATV for males != Mean SATV for females

In [ ]:
```
#Write and run this command:

t.test(scores$SATV ~ scores$gender, data = scores, var.eq = TRUE)
```

```
        Two Sample t-test

data:  scores$SATV by scores$gender
t = 0.49792, df = 698, p-value = 0.6187
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -13.09357  21.99137
sample estimates:
mean in group 1 mean in group 2
       615.1134        610.6645
```

Interpreting the results

- t: computing statistics
- df: degrees of freedom
- p-value: Statistical significance. 0.6187 is bigger than α = 0.05 so that means we must retain the null hypothesis

Conclusion

There was no statistically significant difference in SAT verbal scores between males and females, t(698) = 0.50, p = 0.62.

One-way ANOVA

It is used to determine whether there are group differences in numeric data between more than two groups

For example, do SAT verbal scores significantly differ by educational levels (1= HS, 2= some college degree, 3 = 2-year college degree, 4= 4-year college degree, 5= graduate work)?

H0: Mean SATV of students who have HS degree = Mean SATV of students who have some college degree = ...

H1: Mean SATV of students who have HS degree != Mean SATV of students who have some college degree != ...

In [ ]:
```
#Write and run this command:

m1 <- aov(scores$SATV ~ scores$education, data = scores)
summary(m1)
```

```
> m1 <- aov(scores$SATV ~ scores$education, data = scores)
> summary(m1)
                   Df  Sum Sq Mean Sq F value Pr(>F)
scores$education    5   80754   16151   1.269  0.275
Residuals         694 8829392   12722
```

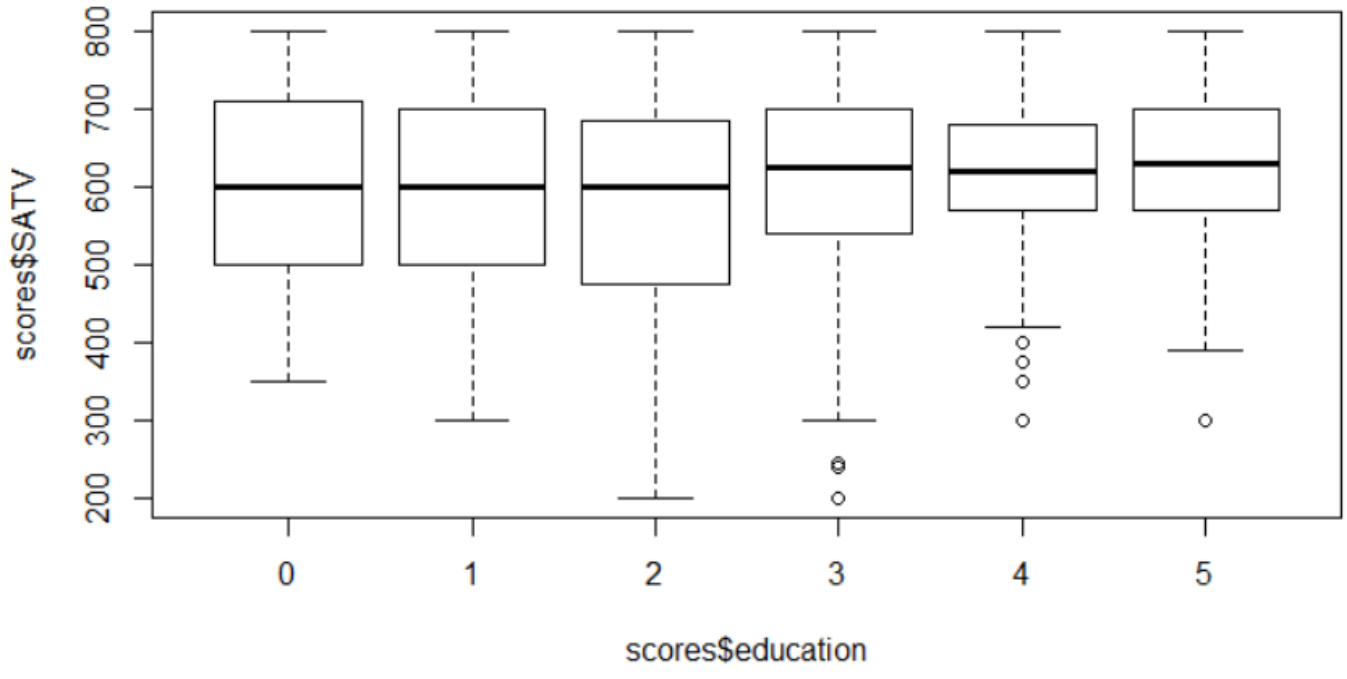
Interpreting the results

- df: degree of freedom
- sum sq: sum of squares
- mean sq: mean squares
- F value: computing statistics
- Pr(>F): statistical significance. 0.275 is bigger than α = 0.05 so that means we must retain the null hypothesis. We do not have to run the post hoc tests because the group differences are not significant.

Does this make sense?

Check using boxplot.

In [ ]:
```
# Write and run this command:

boxplot(scores$SATV ~ scores$education)
```



Conclusion

There were no significant group differences in SAT verbal scores according to students' educational levels, F(5, 694) = 1.269, p = 0.275.

Extra information

There are different types of post hoc tests (https://www.rdocumentation.org/packages/DescTools/versions/0.99.36/topics/PostHocTest), but the Tukey's HSD is the most popular post hoc test for comparing multiple pairings.

In [2]:
```
# R command for Tukey's HSD:

# TukeyHSD(aov(scores$SATV ~ scores$education, data = scores), conf.level=.95)
```

Pearson's Correlation

It is used to examine relationships between variables (represented by numeric data)

For example, Is there a relationship between SATV and SATQ?

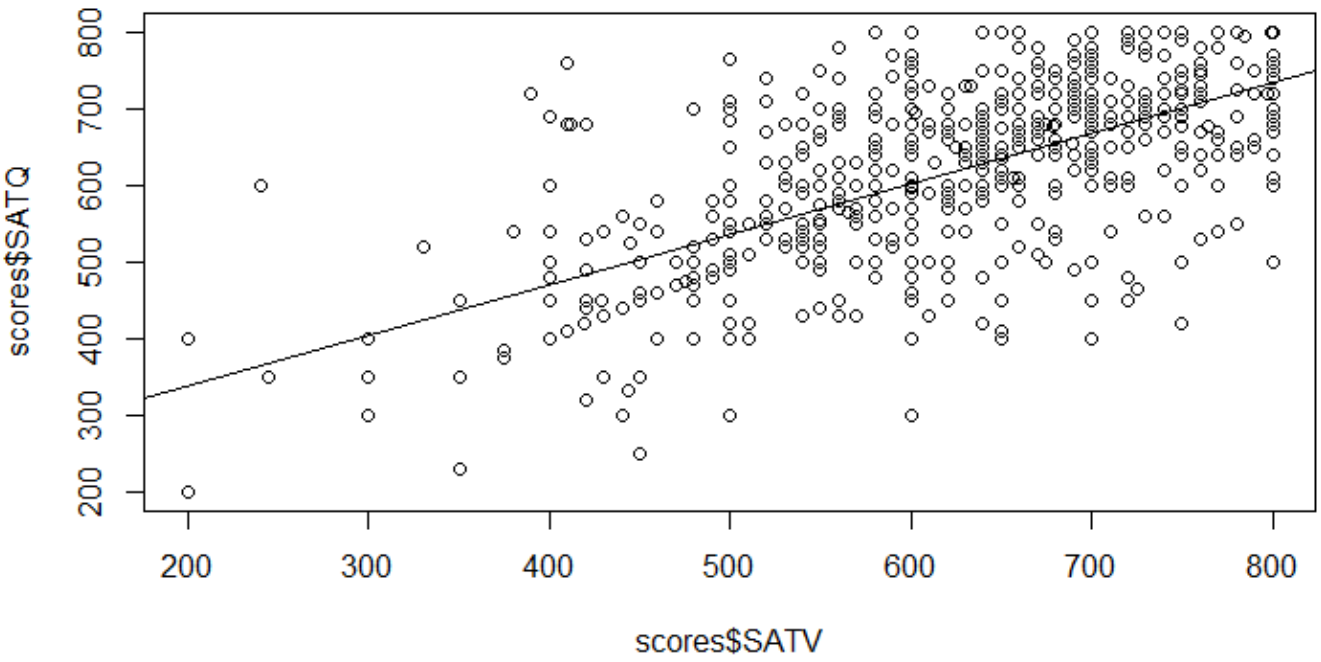
H0: There is no relationship between SATV and SATQ.

H1: There is a relationship between SATV and SATQ.

In [ ]:
```
# Write and run this command:

cor.test(scores$SATV,scores$SATQ)
```

```
        Pearson's product-moment correlation

data:  scores$SATV and scores$SATQ
t = 22.05, df = 685, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5983352 0.6860379
sample estimates:
      cor
0.6442999
```

In [ ]:
```
# Write and run this command to see scatterplot:

plot(scores$SATV,scores$SATQ)
abline(lm(scores$SATQ ~ scores$SATV)) #to add regression line
```



Conclusion

There was statistically significant positive correlation between SAT verbal scores and SAT Quantitative scores (r = 0.644, p < 0.001).

Simple Linear Regression (for your reference)

It is used to explain/predict the phenomenon of interest based one independent variable.

For example, do ACT scores predict SAT verbal scores?

In [ ]:
```
# Write and run this command:

m2 <- lm(scores$SATV ~ scores$ACT)
summary(m2)
```

```
Call:
lm(formula = scores$SATV ~ scores$ACT)

Residuals:
    Min      1Q  Median      3Q     Max
-496.98  -47.73    8.38   62.42  247.48

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 237.3432    21.2319   11.18   <2e-16 ***
scores$ACT   13.1324     0.7334   17.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 93.53 on 698 degrees of freedom
Multiple R-squared:  0.3148,    Adjusted R-squared:  0.3138
F-statistic: 320.7 on 1 and 698 DF,  p-value: < 2.2e-16
```

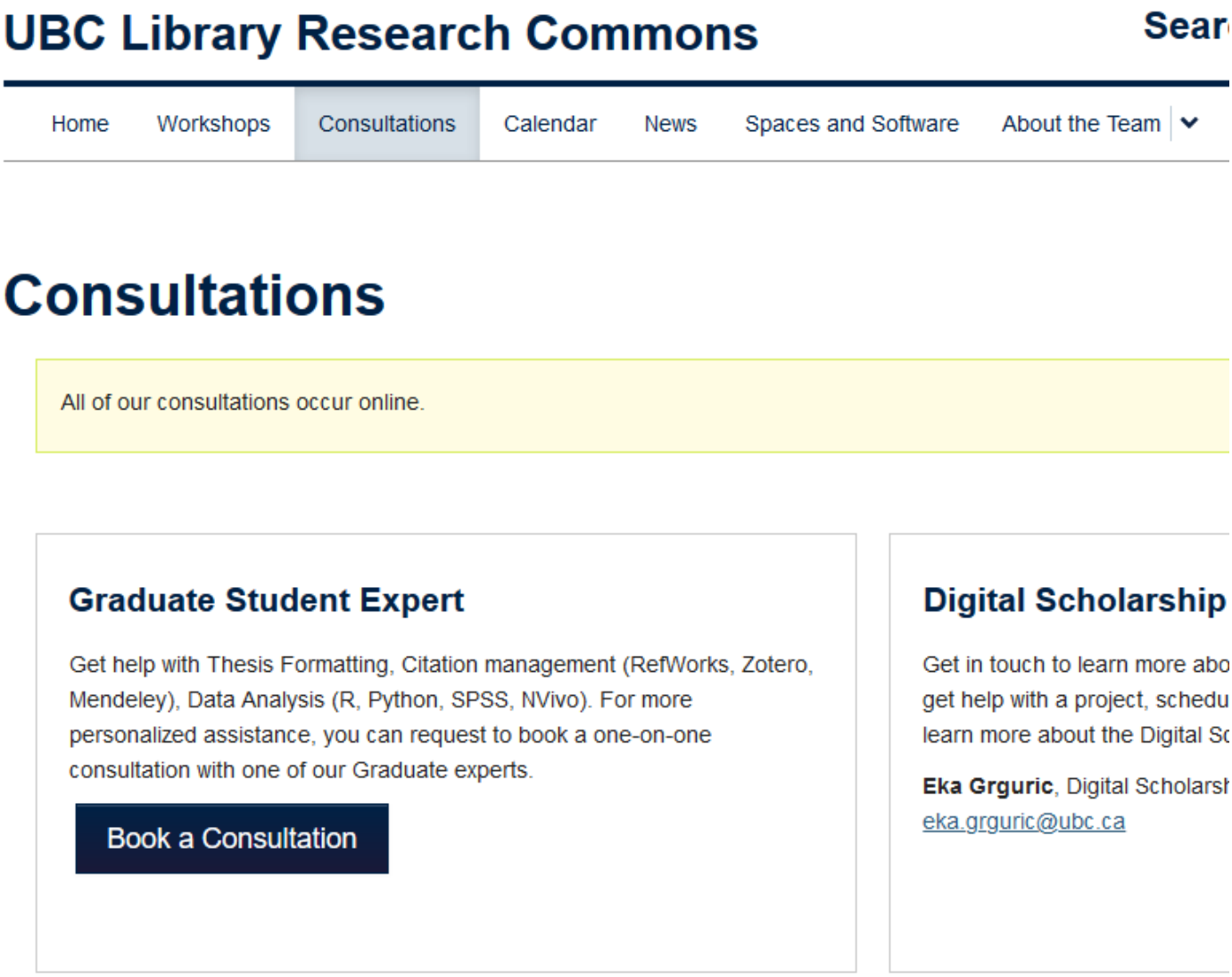
Interpreting the results

- The estimated regression line equation: SATV = 237.34 + 13.13(ACT scores)
- 31% of the variability in the SAT verbal scores was explained by the variables in the regression model.
- The overall regression model significantly explained the SAT verbal scores.

Conclusion

ACT scores significantly predicted SAT verbal scores. We would expect 13.13 points increase in SAT verbal scores for every one point increase in ACT score, assuming all the other variables are held constant.

Questions?


UBC Library Research Commons
Sear
Home
Workshops
Consultations
Calendar
News
Spaces and Software
About the Team
Consultations
All of our consultations occur online.
Graduate Student Expert
Get help with Thesis Formatting, Citation management (RefWorks, Zotero, Mendeley), Data Analysis (R, Python, SPSS, NVivo). For more personalized assistance, you can request to book a one-on-one consultation with one of our Graduate experts.
Book a Consultation
Digital Scholarship
Get in touch to learn more abo
get help with a project, schedu
learn more about the Digital Sc
Eka Grguric, Digital Scholarsh
eka.grguric@ubc.ca

Reference(s);

Revelle, William, Wilt, Joshua, and Rosenthal, Allen. (2009). Personality and Cognition: The Personality-Cognition Link. In Gruszka, Alexandra and Matthews, Gerald and Szymura, Blazej (Eds.) Handbook of Individual Differences in Cognition: Attention, Memory and Executive Control, Springer.