UWES R Workshop: Econometrics in R

Andrew Girgis

University of Waterloo

March 12, 2025

Discord

Figure 1: Discord QR Code

Overview

By the end of today's Presentation you will(hopefully) have a better understanding of:

- The Data Journey
- Data Visualization
- Regression
- Economic Insights

Figure 2: The Data Journey[1]

---

[1](Government of Canada 2021)

Step 1: **Define** - Find - Gather

For today's presentation we will be studying what features can affect a students final grade. First we define our problem.

Objective: To derive what impact important features has on a students final grade.

Step 1: Define - **Find** - Gather

- Now that we have defined our problem the next step in the data journey is to **find and gather** our data. For this I will be using a free dataset from the UC Irvine machine learning repository[2].
- The dataset I will be using is the Student Performance dataset that can be found at archive.ics.uci.edu/dataset/320/student+performance.
- Other sources for free data:
    - Kaggle
    - Statistics Canada
    - US Federal Government Data
    - FRED Economic Data

---

[2]UC Irvine Machine Learning Repository: archive.ics.uci.edu/datasets

Import libraries

Libraries (also known as packages) in R are collections of functions, data sets, and other code that extend the functionality of the base R language. They can be downloaded and installed onto your machine using the install.packages() function, then loaded into your R session using the library() function.

```r
# install.packages('readr')
library(readr)
library(plyr)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(GGally)
library(stargazer)
```

readr: for reading files

Import libraries

Libraries (also known as packages) in R are collections of functions, data sets, and other code that extend the functionality of the base R language. They can be downloaded and installed onto your machine using the install.packages() function, then loaded into your R session using the library() function.

```
# install.packages('readr')
library(readr)
library(plyr)
library(dplyr)
library(ggplot2)
library(tidyverse)
library(GGally)
library(stargazer)
```

Import data

To import data we use the function read_csv() from the readr library.

```
#data_path = '[insert your path here]'
#student_df = read_csv(data_path)
```

Step 2: **Explore** - Clean - Describe

- We use the head() function to view the first 6 lines of our data.
- Ensure the data was imported properly.
- Ensure the variables and values make sense.

```
head(student_df[1:7])
```

```
## # A tibble: 6 x 7
##    school sex     age address famsize Pstatus  Medu
##    <chr>  <chr> <dbl> <chr>   <chr>   <chr>   <dbl>
## 1 GP     F        18 U       GT3     A           4
## 2 GP     F        17 U       GT3     T           1
## 3 GP     F        15 U       LE3     T           1
## 4 GP     F        15 U       GT3     T           4
## 5 GP     F        16 U       GT3     T           3
## 6 GP     M        16 U       LE3     T           4
```

Step 2: **Explore** - Clean - Describe

- After viewing the data we must understand the data.
- Some variables may be hard to interpret.
- See this link for full description of data.
- After viewing the descriptions of all the feature variables[3] I have chosen to focus on the following variables: absences, higher, activities, studytime, schoolsup, reason, address, sex, age, Pstatus
- For simplicity we will be using the target variable [4] G3 for this regression.

---

[3]Feature Variables: Variables that will be used as the independents for the regression (predictors for the target variable).

[4]Target Variable: Variable that will be used as our dependent in the regression.

Step 2: **Explore** - Clean - Describe

- Summary() is an incredibly useful function!
- Changes its output based on the input. If you input a list(column) the summary function will output basic summary stats. If you input an object(regression) the summary function will output important regression results/statistics.

```r
summary(student_df["age"])
```

```
##       age
##  Min.   :15.0
##  1st Qu.:16.0
##  Median :17.0
##  Mean   :16.7
##  3rd Qu.:18.0
##  Max.   :22.0
```

```
# Can use 'age' or column number
```

Step 2: **Explore** - Clean - Describe

- Table is a great way to get an overview of non-numeric variables.
- Provides a count of the unique values in the dataset.

```r
table(student_df[2])
```

```
## sex
##   F   M
## 208 187
```

Step 2: **Explore** - Clean - Describe

```
table(student_df["Mjob"])
```

```
## Mjob
##  at_home    health     other  services   teacher
##       59        34       141       103        58
```

Although we wont be using Mjob as a variable in our regression, I would like to show what the table function outputs with a non-binary categorical variable.

Step 2: Explore - **Clean** - Describe

- From the UC Irvine Machine Learning Repository we see that this dataset contains no missing values. However we can confirm that this is true.
- Depending where you gather your data you may not know whether there are missing values so it is considered good practice to always **inspect and clean** your data.

```r
# Check for NA values in the entire data
# frame
sum(is.na(student_df))
```

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

```r
# Check for NA values in specific columns
# (e.g., age column)
sum(is.na(student_df$age))
```

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

- As an example I will input a NA value in the data.

```
student_df[124, 23] <- NA
```

```
# Check for NA values in the entire data
# frame
any(is.na(student_df))
```

```
## [1] TRUE
```

```
# Find row and column indices with NA values
na_locations <- which(is.na(student_df), arr.ind = TRUE)
na_locations
```

```
##      row col
## [1,] 124  23
```
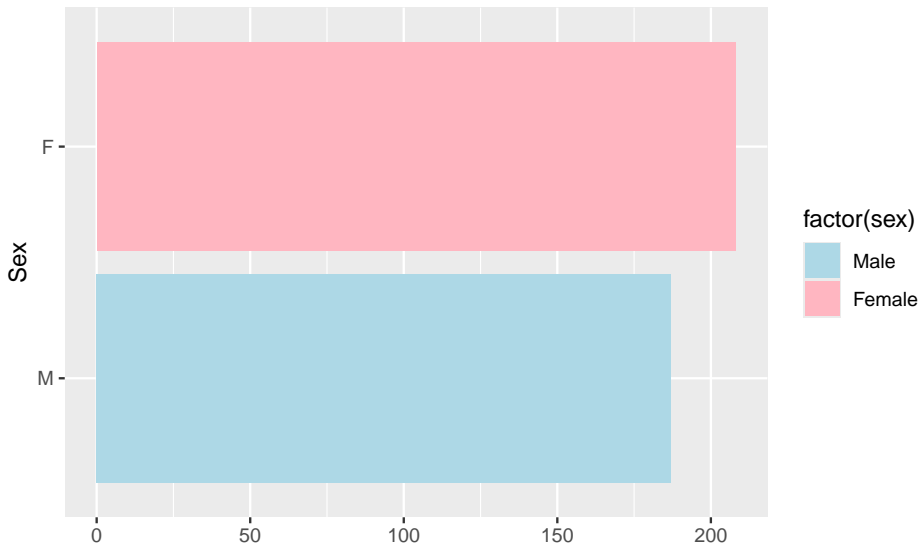
Step 2: Explore - **Clean** - Describe

- Finally we will filter our data to hone in on the variables we want to focus on.

```
filtered_df <- student_df %>%
    select(absences, higher, activities, studytime,
        schoolsup, sex, age, G3)
```
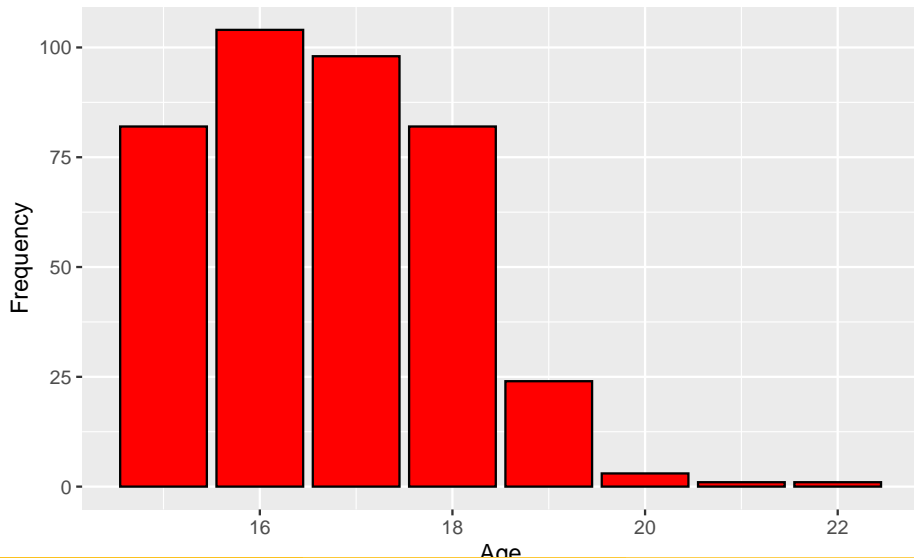
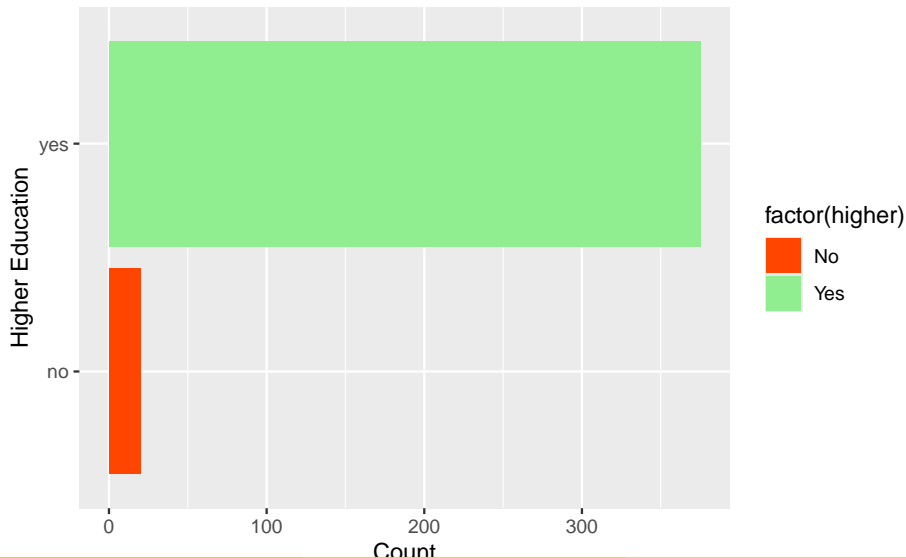# Step 3: **Analyze** - Model

Horizontal Bar Plot of Sex
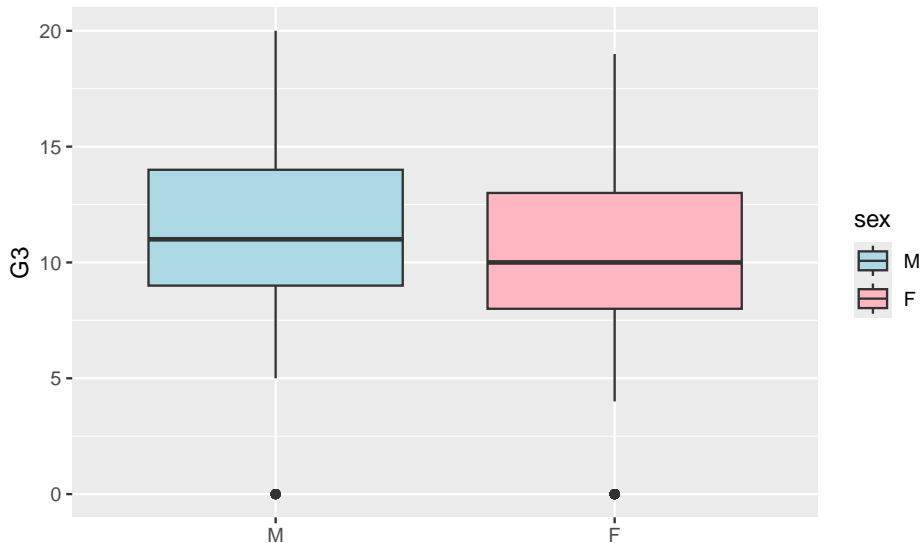
Step 3: **Analyze** - Model

Histogram of Age

Horizontal Bar Plot of students who want to pursue higher education

Boxplot of FInal Average by Sex
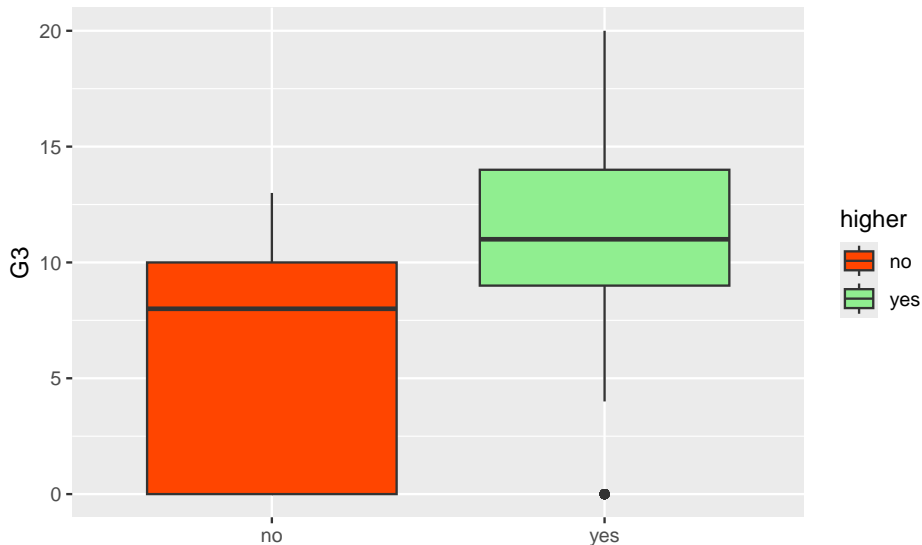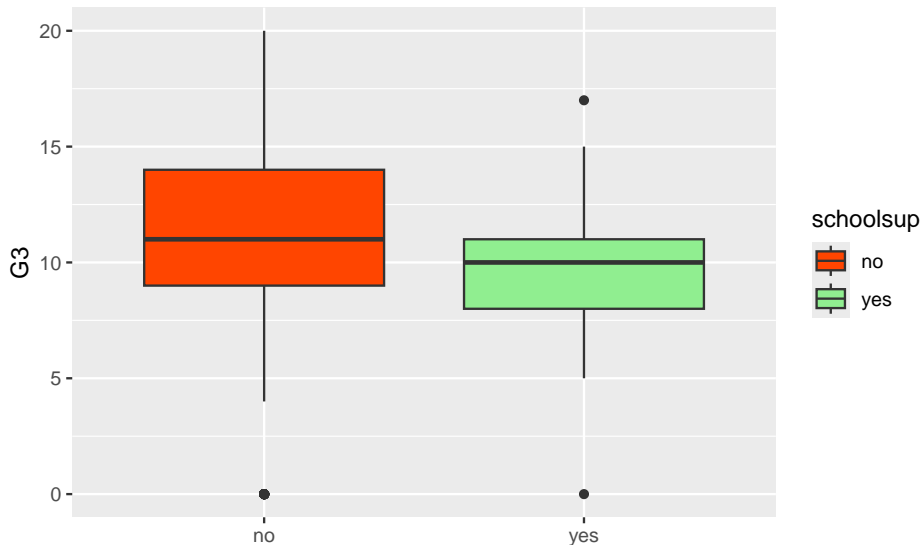
Step 3: **Analyze** - Model

Boxplot of G3 by Higher Education Aspiration

Boxplot of G3 vs wheter a student is getting extra help

Step 3: **Analyze** - Model

- Lets look into this

```
table(filtered_df$schoolsup)
```

```
##
##  no yes
## 344  51
```

- The lower average can be explained by multiple factors including the smaller sample, the quality in the extra support or the students natural ability.
- An interesting analysis to look into (since we have the data) is to see the affect the extra help had on the difference in average from first year in hs to last.
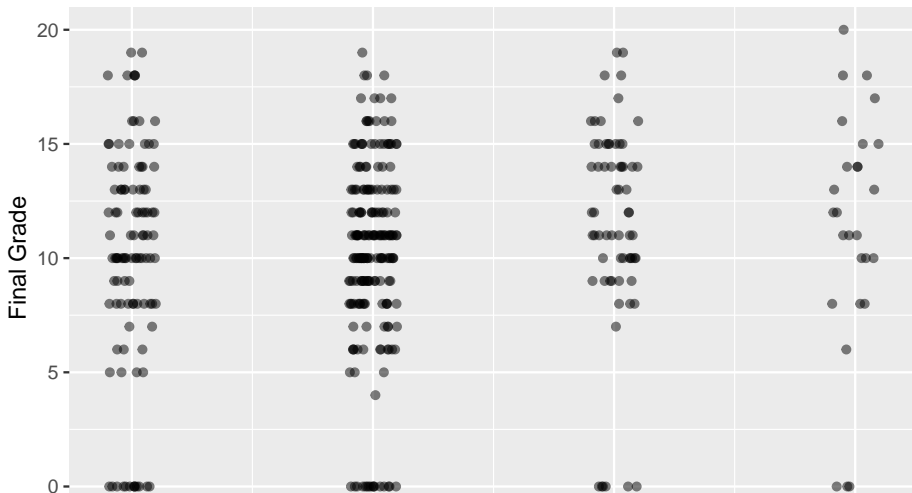
Step 3: **Analyze** - Model

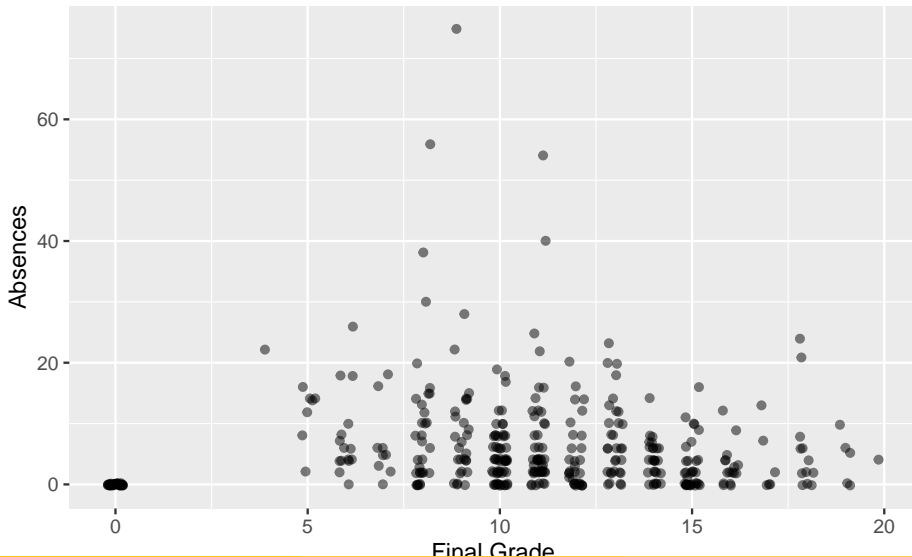Scatter Plot of Study Time vs. Final Grade
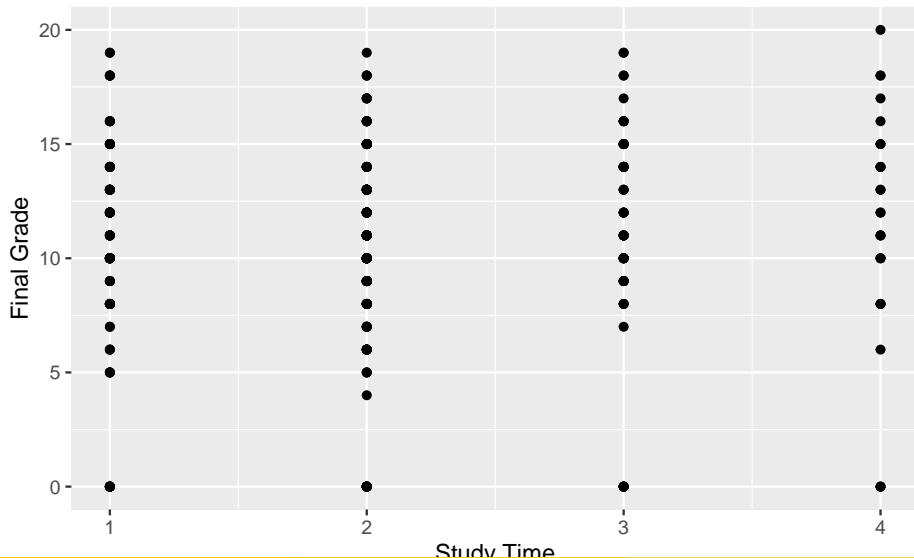
Step 3: **Analyze** - Model

Introducing jitter!

Scatter Plot of Study Time vs. Final Grade

Scatter Plot of Final Grade vs. Absences

Step 3: **Analyze** - Model

```
## Warning: No shared levels found between `names(values)` of
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
## No shared levels found between `names(values)` of the manua
## data's colour values.
```

Step 3: Analyze - **Model**

- Regression analysis is a way of mathematically identifying which independent variables has an impact on our dependant variable.(Gallo 2022)
- It helps us answer the questions:
    - Which factors(independent variables) matter most?
    - Which can we ignore?
    - How do those factors interact with one another?
    - How certain are we about all these factors?

Step 3: Analyze - **Model**

Simple Linear Regression

Our True Model:

$$y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$

β_0: True Intercept

β_1: True beta coefficient that quantifies the exact strength and direction of the relationship between each independent variable and the dependent variable.

ϵ: Error term

Our Estimated Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i}$$

$\hat{\beta}_0$: Estimated Intercept

$\hat{\beta}_1$: Estimated coefficient that quantifies the strength and direction of the

Table 1: Simple Regression Results - Part 1

	Dependent variable:
	G3
studytime	0.534*
	(0.274)
Constant	9.328***
	(0.603)

Table 2: Simple Regression Results - Part 2

R^2	0.010
Adjusted R^2	0.007
Residual Std. Error	4.565 (df = 393)
F Statistic	3.797* (df = 1; 393)

Note:	*p<0.1; **p<0.05; ***p<0.01

Step 3: Analyze - **Model**

Visually

Scatterplot of Final Grade vs. Study Time
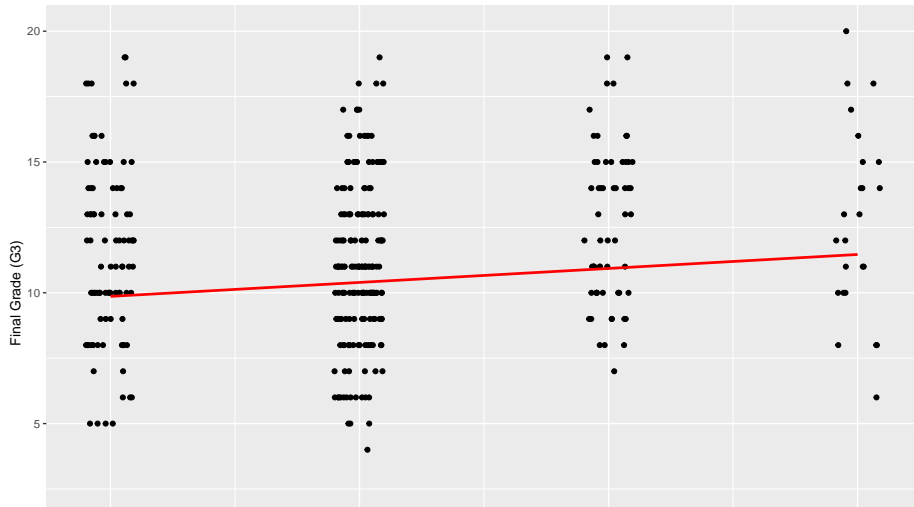
Step 3: Analyze - **Model**

Multiple Linear Regression

Our True Model:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \beta_6 X_{6i} + \beta_7 X_{7i} + \epsilon_i$$

β_0: True Intercept

β_{1-7}: True beta coefficient that quantifies the exact strength and direction of the relationship between each independent variable and the dependent variable.

ϵ: Error term

Our Estimated Model:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \hat{\beta}_4 X_{4i} + \hat{\beta}_5 X_{5i} + \hat{\beta}_{6i} X_{6i} + \hat{\beta}_7 X_{7i}$$

$\hat{\beta}_0$: Estimated Intercept

Step 3: Analyze - **Model**

Interpretation:

For every one unit increase in X_{ki} our y increases (or decreases, depending on sign) by $\hat{\beta}_k$, on average, while holding other variables constant. where k is the independent variable we are interpreting.

Step 3: Analyze - **Model**

```
filtered_df$sex <- factor(filtered_df$sex, levels = c("M",
    "F"))
filtered_df$sex <- as.numeric(filtered_df$sex) -
    1
filtered_df$sex
# M:0 F:1

# Encode 'higher' variable
filtered_df$higher <- factor(filtered_df$higher,
    levels = c("no", "yes"))
filtered_df$higher <- as.numeric(filtered_df$higher) -
    1
filtered_df$higher
# 'no': 0, 'yes': 1
```

```r
head(filtered_df[1:6])
```

```
## # A tibble: 6 x 6
##   absences higher activities studytime schoolsup   sex
##      <dbl>  <dbl>     <dbl>     <dbl>     <dbl> <dbl>
## 1        6      1         0         2         1     1
## 2        4      1         0         2         0     1
## 3       10      1         0         2         1     1
## 4        2      1         1         3         0     1
## 5        4      1         0         2         0     1
## 6       10      1         1         2         0     0
```

Table 3: Regression Results - Part 1

	Dependent variable:
	G3
absences	0.054*
	(0.028)
higher	3.393***
	(1.057)
activities	−0.348
	(0.452)
studytime	0.712**
	(0.282)

Table 4: Regression Results - Part 2

	Dependent variable:
	G3
schoolsup	-1.622^{**}
	(0.690)
sex	-1.436^{***}
	(0.479)
age	-0.621^{***}
	(0.187)
Constant	16.947^{***}
	(3.471)

Table 5: Regression Results - Part 3

	Dependent variable:
	G3
Observations	395
R^2	0.099
Adjusted R^2	0.083
Residual Std. Error	4.388 (df = 387)
F Statistic	6.078*** (df = 7; 387)
Note:	*p<0.1; **p<0.05; ***p<0.01

Figure 3: Behold the Dragon Scroll

References I

Gallo, Amy. 2022. "A Refresher on Regression Analysis." *Harvard Business Review*. https://hbr.org/2015/11/a-refresher-on-regression-analysis.

Government of Canada, Statistics Canada. 2021. "Data Journey." *Government of Canada, Statistics Canada*. https://www.statcan.gc.ca/en/wtc/data-literacy/journey.