Education 240 Final Project

Professor Kisha Daniels

Matthew Cui

3/31/2020

Contents

1. Introduction

In this project, I will be using statistical testing method, data visualization techniques, and content covered in lectures and readings to carry out an analysis on this dataset collected and made available by the University of California, Irvine Department of Machine Learning.

First, let's read and load the data. In this chunk, I will also load the libraries that I will need to perform the necessary analysis used in the rest of this report.

```
schools <- read.csv("student-por.csv")

library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(infer)
library(broom)
```

This dataset has 649 rows and 33 columns, meaning that the researchers took a sample of 649 students and collected information about them on 33 different variables. The original data was titled `Student Alcohol Consumption` as it categorizes students into different consumption levels from 1-5 both on workdays and weekends, as shown by the variables `dalc` and `walc` respectively.

However, this dataset also contains valuable insights into the students' academic performance as it contains their grades from year 1 to 3, shown in the variables `g1`, `g2`, and `g3` respectively. I will be creating a series of visualizations using this information and attempting to explain the trends I see using information learned in class and in my readings.

The first step of data wrangling I need to perform is converting the categorical variables into factors. Factors are essentially different levels of a categorical variable. I also changed the other numerical categorical variables into factors as they are ordinal, meaning the ranking matters, but it doesn't contain any explicit information numerically. Essentially, the numbers are only meaningful in relative to one another.

1.1 Manipulating variables

```
schools <- schools %>%
  mutate_if(is.character, as.factor)
  # mutate_at(vars(Medu,
  #                Fedu,
  #                traveltime,
  #                studytime,
  #                famrel,
  #                freetime,
  #                goout,
  #                Dalc,
  #                Walc,
  #                health), factor)
```

2 The role of play

The first concept I want to investigate is Piaget's developmental theory on play. To do this, I will create a linear regression model of how the explanatory variables `goout`, `freetime` (categorical variables with levels of frequency from 1-5), and `activities` (binary variable of yes/no to extracirricculars) affect the response variable `g3` (final school grade).

2.1 Data

```
lm_g3_play <- lm(G3 ~ goout + freetime + activities, data = schools)

tidy(lm_g3_play) %>%
  select(term, estimate)
```

```
## # A tibble: 4 x 2
##   term          estimate
##   <chr>            <dbl>
## 1 (Intercept)     13.3
## 2 goout           -0.150
## 3 freetime        -0.357
## 4 activitiesyes    0.530
```

2.2 Interpretation

From this, we get a linear model in the form of $Y = m_1X_1 + m_2X_2 + m_3X_3 + c$. The linear model from this regression is `G3 = 13.3 - 0.150 * goout - 0.357 * freetime + 0.530 * activities`.

By interpreting the signs of the coefficients, we see that both going out and having free time negatively correlates with the final course grade. Essentially, if an individual has a free time score of 5, his final course grade would be calculated by multiplying 5 with -0.357. The intercept means that, given all other variable values to be 0, would be 13.3 out of 20. This means that if a student had no activities nor free time and doesn't go out, the average score in that group would be 13.3.

Extracurricular activities, on the other hand, has the only positive coefficient in the model, suggesting a positive correlation between `G3` and `activities`. This means that when a student has extracirricular activities, their predicted final course grades are to increase by 0.530 points on average.

2.3 Discussion

After understanding how different attributes affect this sample of students, we can now discuss how this supports or challenges Piaget's theories.

Scales, et al. (1991) defined play as "that absorbing activity in which healthy young children participate with enthusiasm and abandon," whereas Csikszentmihalyi (1981) described play as "a subset of life. . . , an arrangement in which one can practice behavior without dreading its consequence." In both cases, these researchers argue that play enables the learner to take on and understand simulated roles that they normally cannot be to expand their knowledge in different circumstances.

One reason, therefore, that could explain the effects of the variables in the model is the nature of the students' play. When they are engaged in extracurricular activities, it's often endorsed by the school, suggesting that there is a series of guidance provided by more knowledgeable others (MKO), whether that be a teacher, or older students. Their guidance is crucial in these students' learning of new knowledge, especially those that they have not mastered on their own. On the other hand, when students have `freetime`, or `goout`, this time might not be used productively and constructively to positively contribute towards their final grade.

3. Does quality of family relationships affect academic performance?

We have previously discussed how every student faces a different challenge. Whether it's suffering from physical or mental disabilities or financial struggles, these factors could all affect a student's academic performance in school. One other factor that I will be investigating is whether the quality of relationships a student has with their family affect their academic performance. I will do this through simlulation-based hypothesis testing.

3.1 Formalizing research hypothesis and operationalizing variables

3.2 Simulating the sampling process

```
famsuccess_schools <- schools %>%
  mutate(categorized_famrel = case_when(
    famrel >= 4 ~ "high-quality",
    famrel < 4 ~ "low-quality"
  )) %>%
```

```
  mutate(categorized_g3 = case_when(
    G3 >= 12 ~ "above average",
    G3 < 12 ~ "below average"
  )) %>%
  mutate(categorized_famrel = as.factor(categorized_famrel)) %>%
  mutate(G3 = as.factor(G3))
```

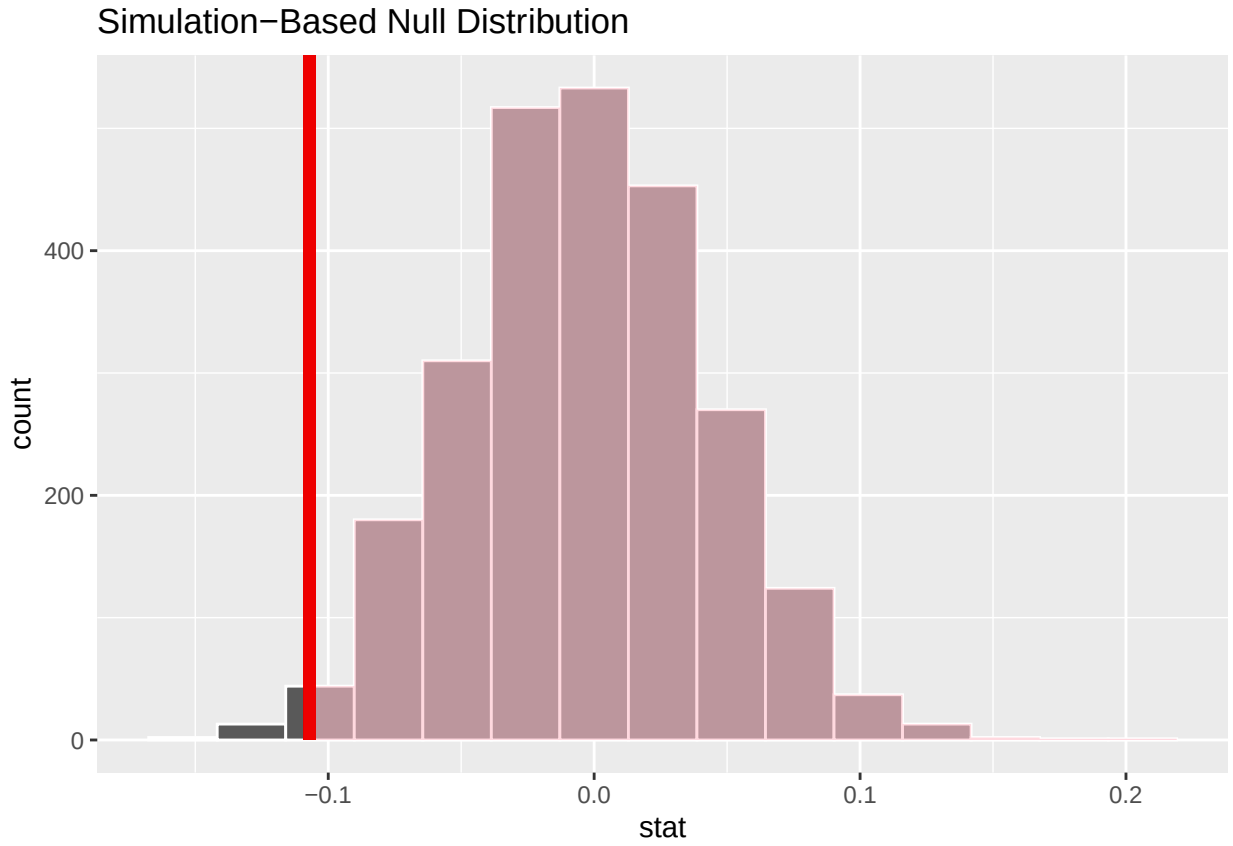
```
p_hat_diff <- famsuccess_schools %>%
  count(categorized_famrel, categorized_g3) %>%
  group_by(categorized_famrel) %>%
  mutate(p_hat = n / sum(n)) %>%
  filter(categorized_g3 == "above average") %>%
  pull(p_hat) %>%
  diff()

p_hat_diff
```

```
## [1] -0.1074208
```

```
famgrade_dist <- famsuccess_schools %>%
  specify(response = categorized_g3, explanatory = categorized_famrel,
          success = "above average") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 2500, type = "permute") %>%
  calculate(stat = "diff in props",
            order = c("low-quality", "high-quality"))
```

```
visualize(famgrade_dist) +
  shade_p_value(obs_stat = -0.107, direction = "greater")
```




```
famgrade_dist %>%
  filter(stat >= -0.107) %>%
  summarise(p_value = n() / nrow(famgrade_dist))
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.990
```

3.3 Alternative method of investigating causes for academic performance

This method invovles the use of linear regression as discussed in Section 2. Instead of only picking two variables, however, we create a linear model that has one response variable, `G3`, and a lot of explanatory variables. Within these explanatory variables, we aim to find what combination of them has the highest adjusted R^2 value (similar to R^2 but takes into account of the number of explanatory variables).

```
lm_full <- lm(G3 ~ school + G1 + studytime + higher + internet + famrel +
              health + absences, data = schools) # creating model
```

Once the model has been created, the function `step()` automates a backward-selection process and ends the output with the combination of variables with the highest adjusted R^2 value. It essentially brute-forces the process by starting with the full model, records the $adj.R^2$, removes one variable, determines whether the $adjR^2$ is higher without that variable. If it is higher, the current model becomes the 'best model', until removing another variable yields a higher value.

```
best_model <- step(lm_full) # using the step function
```

```
## Start:  AIC=772.2
```

```
## G3 ~ school + G1 + studytime + higher + internet + famrel + health +
##     absences
##
##             Df Sum of Sq    RSS     AIC
## - internet   1       2.7 2077.4  771.06
## - absences   1       4.7 2079.3  771.67
## - studytime  1       5.3 2079.9  771.85
## - famrel     1       5.5 2080.2  771.93
## <none>                   2074.6  772.20
## - school     1       7.3 2081.9  772.48
## - higher     1      13.7 2088.3  774.47
## - health     1      24.7 2099.3  777.89
## - G1         1    3196.9 5271.5 1375.42
##
## Step:  AIC=771.06
## G3 ~ school + G1 + studytime + higher + famrel + health + absences
##
##             Df Sum of Sq    RSS     AIC
## - absences   1       5.1 2082.4  770.65
## - studytime  1       5.2 2082.5  770.68
## - famrel     1       6.2 2083.6  771.00
## <none>                   2077.4  771.06
## - school     1       9.5 2086.9  772.03
## - higher     1      14.0 2091.3  773.41
## - health     1      25.4 2102.8  776.96
## - G1         1    3225.3 5302.6 1377.24
##
## Step:  AIC=770.65
## G3 ~ school + G1 + studytime + higher + famrel + health
##
##             Df Sum of Sq    RSS     AIC
## - studytime  1       4.3 2086.7  769.98
## - famrel     1       5.4 2087.8  770.31
## <none>                   2082.4  770.65
## - higher     1      12.7 2095.1  772.59
## - school     1      13.8 2096.3  772.93
## - health     1      26.6 2109.1  776.89
## - G1         1    3259.6 5342.0 1380.04
##
## Step:  AIC=769.98
## G3 ~ school + G1 + higher + famrel + health
##
##          Df Sum of Sq    RSS     AIC
## - famrel  1       5.2 2091.9  769.59
## <none>                2086.7  769.98
## - higher  1      14.5 2101.2  772.47
## - school  1      14.9 2101.6  772.60
## - health  1      27.8 2114.5  776.58
## - G1      1    3421.4 5508.2 1397.92
##
## Step:  AIC=769.59
## G3 ~ school + G1 + higher + health
##
##          Df Sum of Sq    RSS     AIC
```

```
## <none>                2091.9  769.59
## - higher  1      15.0 2107.0  772.24
## - school  1      15.1 2107.0  772.26
## - health  1      25.5 2117.5  775.47
## - G1      1    3436.2 5528.2 1398.27
```

```
tidy(best_model) # displaying in a tidy manner
```

```
## # A tibble: 5 x 5
##   term        estimate std.error statistic   p.value
##   <chr>          <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)    1.43     0.405       3.52 4.55e-  4
## 2 schoolMS      -0.336    0.156      -2.15 3.16e-  2
## 3 G1             0.931    0.0286     32.5  5.07e-138
## 4 higheryes      0.528    0.245       2.15 3.18e-  2
## 5 health        -0.138    0.0492     -2.80 5.19e-  3
```

From the output of `best_model`, we get a linear model of

`final period grade = 1.43 - 0.336 * MS + 0.931 * first period grade + 0.528 * higher - 0.138 * health`

Comparing this to the full model we started with, the variables `studytime`, `internet`, `famrel`, and `absences` were all discarded from the best model. From the code below, we see that approximately `68.9%` of the variability in the response variable `G3` can be explained by the response variables.

```
glance(best_model) %>%
  select(adj.r.squared)
```

```
## # A tibble: 1 x 1
##   adj.r.squared
##           <dbl>
## 1         0.689
```

It is indeed a bit awkward that the variable that I hypothesized in the beginning of this section will directly cause variances in the final course grade did not even make this final correlational model. This goes to say that, at least in this sample, the quality of relationships with family does not play a relatively big role in affecting students' final academic performances. This also explains why the p-value we obtained in the previous statistical test is so ridiculously high.

3.4 Discussion