Final Report

Research Question

Does laptop operating system influence a person's self-rated level of stress in the MDS program?

We think this is an important question to research as this question might help future students be best set up for success in the MDS program. It might also help program instructors in designing better learning activities. If instructors know that a student's choice of OS influences one's stress levels, instructors may be able to keep OS-specific issues in mind while designing learning activities in order to mitigate sources of student stress.

Methods

Survey Study Design

Our survey asks respondents what operating system they use on their laptop (Linux, Mac, or Windows) and what their level of stress is on a scale from 0 to 10, where 0 represents "little or no stress" and 10 represents "a great deal of stress." The remaining questions targetted what we thought could be potential confounders: age, work experience, gender, self-reported coding proficiency, and academic background.

For work experience and age we asked users to enter a number, rather than having pre-determined bins. We did this so that we could have more detailed data for our analysis. For academic background, we chose to provide some pre-specified categories to simplify data wrangling later. Coding proficiency was also rated on a scale. In order to preserve survey respondents' anonymity, we gave them a "prefer not to say" option for questions about age, work experience, and gender.

Data Collection

We collected the survey data using the UBC-hosted version of Qualtrics, as recommended by the UBC Office of Research Ethics. We sent the survey out to MDS students during lab time and received 52 responses.

Analysis Methods

After cleaning the data, we produced some exploratory plots to see if there was a relationship between operating system and stress and to see if the potential confounders had any association with operating system or stress. Our exploratory data analysis can be found here.

To determine if operating system affects stress level, we fit linear regression and proportional odds models and tested the significance of `os` on `stress`.

Analysis, Results, and Discussion

Our first approach was to try a simple linear regression relating the response variable - stress levels - with the explanatory variable of interest, the operating system.

```
# linear regression with stress and os only
tidy(anova(lm(stress ~ os, data = data)))
```

```
## # A tibble: 2 x 6
##   term        df  sumsq meansq statistic p.value
##   <chr>    <int>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 os           2   4.63   2.32     0.479   0.623
```

```
## 2 Residuals    49 237.    4.84    NA      NA
```

A p-value greater than 0.6 for **os** suggests that there is no significant relationship between operating system and stress in the data. We then proceeded to refine our model by conditioning the model to an ordinal output. This represents an improvement given that respondents provided their self-assessed stress levels from a discrete scale of 1 to 10. Here's what we obtained:

```r
# proportional odds model, stress and os only
results <- tidy(polr(as.factor(stress) ~ os, data = data))
results <- results %>%
  mutate(p_value = pnorm(statistic, lower.tail = FALSE))

results
```

```
## # A tibble: 11 x 6
##    term       estimate std.error statistic coefficient_type  p_value
##    <chr>         <dbl>     <dbl>     <dbl> <chr>               <dbl>
##  1 osMac         0.603     0.891     0.677 coefficient         0.249
##  2 osWindows     1.14      0.922     1.24  coefficient         0.108
##  3 1|2          -3.23      1.26     -2.56  zeta                0.995
##  4 2|3          -1.53      0.905    -1.69  zeta                0.955
##  5 3|4          -1.15      0.880    -1.31  zeta                0.905
##  6 4|5          -0.727     0.859    -0.846 zeta                0.801
##  7 5|6          -0.0911    0.834    -0.109 zeta                0.543
##  8 6|7           0.264     0.827     0.319 zeta                0.375
##  9 7|8           1.42      0.851     1.67  zeta                0.0477
## 10 8|9           2.85      0.927     3.07  zeta                0.00107
## 11 9|10          4.75      1.30      3.64  zeta                0.000134
```

We *approximated* the p-values of each parameter by calculating the probability of observing a more extreme statistic in a standard normal distribution. Again, although Windows users seem to have higher stress levels, the results are far from being statistically significant.

Finally, we tried a model that included som confounders, in the hope of finding some surprising insight on what may be associated with higher stress levels. Our hypothesis is that perhaps people with better coding skills (whether it be due to their background, previous experience or just more confidence in their abilities), will show less signs of stress.

```r
# linear regression with coding and background
tidy(anova(lm(stress ~ os + coding + background, data = data)))
```

```
## # A tibble: 4 x 6
##    term          df  sumsq meansq statistic p.value
##    <chr>      <int>  <dbl>  <dbl>     <dbl>   <dbl>
## 1 os             2   4.63   2.32     0.460   0.635
## 2 coding         4   7.29   1.82     0.362   0.834
## 3 background     5  28.4    5.69     1.13    0.360
## 4 Residuals     40 201.     5.04    NA      NA
```

By fitting the linear model and analyzing the results in an ANOVA table, we see that no covariate seems to have a statistically significant relationship with our response variable, stress level. The p-values are all clearly above any reasonable significance level, which indicates that no variable, when added to the model, provides any extra insight on predicting the stress level of students. The graph below shows this conclusion visually: there is no clear pattern or relationship between stress level and our hypothesized explanatory variable, operating system.

Fig.1: Operating System vs. Stress Level
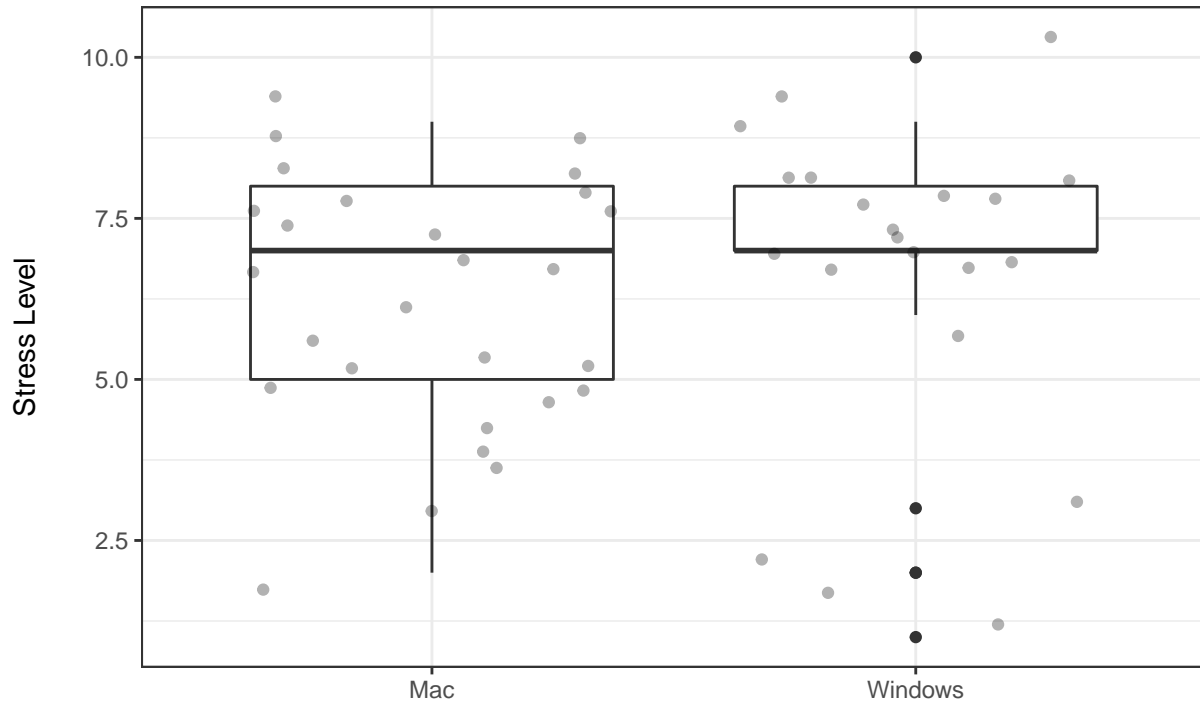
If anything, we could argue that most Windows users are concentrated around higher stress levels, whereas a significant part of Mac users present mid-to-low levels of stress (< 7 on a 0-10 scale). However, our linear regression suggests that this could well have happened just due to random chance, especially given the small amount of data that we have.

Survey & Study Design and Discussion

In the proposal phase, we tried to be exhaustive in identifying potential confounders for our response variables and hypothesized cause. We brainstormed based on intuition and experience. We came up with a clear, plausible and interesting hypothesis, with the aim of exploring causality.

Nevertheless we identified a few bumps during our study that affected our result:

1. We value our classmates' privacy and gave this a lot of thought in the design process. We received valuable feedback from some of our classmates who noted that they were concerned about identifiability while responding to our survey, and some mentioned not completing the survey due to this concern. A few respondents chose "Prefer not to say" for both "gender" and "age" and this reinforced for us that the combination of these two data points could be used to uniquely identify some respondents. As a result, we note that our survey questions may have affected our response rate and the quality of the data we collected.

2. Following the point above, we hit a technical bump during the survey. We used regexes to avoid invalid inputs for age and work experience. Being unfamiliar with the tool and having not had as much time as we would have wanted for thorough survey testing, our survey returned errors for some of our respondents. For example, if a respondent selected "Prefer not to say", the platform still required them to input a number for "age". This was pointed out by a respondent and may have resulted in a loss of responses or lower quality responses.

3. Based on instructor feedback, we realized that the question about work experience was too general. We should have rephrased the question to link working experience with operating system to make it more relevant to our question.

If we got the chance to do this survey again, here are some improvements we would implement:

1. We might consider asking respondents for either "age" or "gender," instead of both, to address the privacy concerns of some respondents. And for the survey question about "age", we would bin the "age" variable to make it more unidentifiable. Hopefully this could improve our response rate and perhaps the quality of our data.

2. Since the result shows that there were no significant relationship between OS and stress level, there is still chances that we missed some confounders. We would add an open-ended question like "What causes the stress in your opinion?" or "Why do you use your OS of choice?" to see if there was anything we missed.