DSCI 554 Lab 4

Practicum of Causality Through an Observational Study

Contents

Lab Overview

In this last lab, you will practice investigating causality through an observational study.

Lab Mechanics

rubric={mechanics:5}

- Paste the URL to your GitHub repo here: https://github.com/UBC-MDS/554-rory-kiersten
- Once you finish the assignment, you must **knit** this `R` markdown to create a `.pdf` file and push everything to your GitHub repo using `git push`. You are responsible for ensuring all the figures, texts, and equations in the `.pdf` file are appropriately rendered.
- **You must submit the rendered `.pdf` file to Gradescope.**

  Heads-up: You need to have a minimum of 3 commits.

Code Quality

rubric={quality:3}

The code that you write for this assignment will be given one overall grade for code quality. Check our **code quality rubric** as a guide to what we are looking for. Also, for this course (and other MDS courses that use `R`), we are trying to follow the `tidyverse` code style. There is a guide you can refer too: http://style.tidyverse.org/

Each code question will also be assessed for code accuracy (i.e., does it do what it is supposed to do?).

Writing

rubric={writing:3}

To get the marks for this writing component, you should:

- Use proper English, spelling, and grammar throughout your submission (the non-coding parts).
- Be succinct. **This means being specific about what you want to communicate, without being superfluous.**

Check our **writing rubric** as a guide to what we are looking for.

A Note on Challenging Questions

Each lab will have a few challenging questions. These are usually low-risk questions and will contribute to maximum 5% of the lab grade. The main purpose here is to challenge yourself or dig deeper in a particular area. When you start working on labs, attempt all other questions before moving to these questions. If you are running out of time, please skip these questions.

We will be more strict with the marking of these questions. If you want to get full points in these questions, your answers need to

- be thorough, thoughtful, and well-written,
- provide convincing justification and appropriate evidence for the claims you make, and
- impress the reader of your lab with your understanding of the material, your analytical and critical reasoning skills, and your ability to think on your own.

Setup

If you fail to load any packages, you can install them and try loading the library again.

```
library(tidyverse)
library(janitor)
library(tools)
library(scales)
library(broom)
library(MASS)
library(rmarkdown)
```

Exercise 1: Conceptual Part

Q1.1. Odds, Odds Ratio, and Log-odds Ratio

rubric={reasoning:6}

In your own words, explain what is meant by odds, odds ratio, and log-odds ratio. Use a numeric real-life framework in your explanation. This framework has to connect the three concepts. Based on your numeric example, provide interpretations on these metrics.

ANSWER:

Odds: The ratio of the probability of an event occurring to the probability of the event not occurring. Eg. If you have a biased coin where 25% of the coin flips land as heads, the odds of heads:tails would be 1:3.

Odds ratio: The ratio of the probability of an event occurring to the probability of the event not occurring, presented as a single number. Eg. If you have a biased coin where 25% of the coin flips land as heads, the odds ratio would be $1/3 = 0.3333$.

Log-odds ratio: The natural logarithm of the odds ratio. Eg. If you have a biased coin where 25% of the coin flips land as heads, the log-odds ratio would be $\ln(1/3)^* = -1.098612$.

*In R the log() function is the natural logarithm.

Q1.2. Creating a Research Question

rubric={reasoning:6}

Create a **single** research question you would be interested in answering via an observational study. Using this research question, describe how you would design three studies, each using the approaches listed below (**one or two sentences per approach**):

- Cross-sectional (Contemporaneous).
- Case-control (Retrospective).
- Cohort (Prospective).

ANSWER:

Research question: Do people who drink coffee have higher test scores than those who don't?

Cross-sectional (Contemporaneous): I take a random sample of 300 people from the population (Canadians above 18 years of age). From this population sample I would record whether they drank a coffee that day and their performance on a test.

Case-control (Retrospective): I would use a random sample of people from the population (Canadians above 18 years of age) who write a test. From this sample I would identify 150 who performed well and 150 subjects who performed poorly. Then I would ask them whether they drank coffee that day or not.

Cohort (Prospective): I would use a random sample of people from the population (Canadians above 18 years of age) who perform poorly on a test. I would then ask half of the subjects to drink coffee and half to refrain for a period of 1 week. I would then retest them and see how many from each group performed well on the test.

Q1.3. Design an Observational Study for Your Research Question

rubric={reasoning:6}

For your described research question above, which study design (out of the above three) would you choose (considering realistic study constraints for your example)? Justify your choice in three or four sentences.

ANSWER:

For this study I would use a cross-sectional design. It is the most realistic since there are many factors that can impact test scores. The use of this study design would most clearly identify whether there is an association between coffee consumption and test performance. The other two study designs are more convoluted and would not improve assessment of the question being asked.

Exercise 2: Analyze Causality in an Observational Study

Let us retake the question from `lab3-ex4`:

Does a person's self-rated enjoyment of the MDS program (X) had any causal influence on their expected salary upon graduation (Y)?

To answer this, a team ran a survey which included the following two questions to collect data to answer their question of interest (Y):

What is your salary expectation after graduation in CAD? (`salary_exp_post_grad`)

- `Less than $60,000`
- `$60,000 to $80,000`
- `$80,001 to $100,000`
- `$100,001 to $120,000`
- `More than $120,000`

The response above was subject to the following explanatory variable of interest (X):

What is your self rated enjoyment of MDS on a scale of `1 - 4`? With `4` being very happy with MDS and `1` being not happy at all with MDS. (`mds_self_rated_enjoy`)

- `1`
- `2`
- `3`
- `4`

Moreover, the team also collected the answers on the following questions:

What was your previous salary prior to MDS? (`salary_pre_mds`)

- `Less than $60,000`
- `$60,000 to $80,000`
- `$80,001 to $100,000`
- `$100,001 to $120,000`
- `More than $120,000`

How many years of professional work experience did you have prior to MDS? (`work_exp`)

- `0 - 1 Years`
- `1 - 4 Years`
- `4 - 7 Years`
- `7 - 10 Years`
- `10+ Years`

How confident in your data science skill set did you feel when first starting MDS on a scale of `1 - 4`? With `4` being very confident and `1` being not confident (`ds_skill_confidence`).

- `1`
- `2`
- `3`
- `4`

Do you typically do optional questions in labs? (`does_optional_qs`)

- `Yes`
- `No`

Are you currently applying for data science jobs? (`currently_job_searching`)

- `Yes`
- `No`

How would you rate your current happiness level on a scale of 1-4? With 4 being very happy and 1 being not happy (`baseline_happiness`).

- 1
- 2
- 3
- 4

How often do you attend MDS career events? For example, the panel on technical interview questions (`freq_attend_mds_career_events`).

- Not often
- Sometimes
- Often

The data was collected **in 2019 and 2020** via a survey with 65 and 49 respondents each (**raw sample sizes before data wrangling**). The *raw datasets* `salary_df_2019` and `salary_df_2020` are the following:

```
# Run this code before proceeding.

salary_df_2019 <- read_csv("data/salary_survey_responses_2019.csv", skip = 2) %>%
  clean_names()

salary_df_2020 <- read_csv("data/salary_survey_responses_2020.csv", skip = 2) %>%
  clean_names()
```

Data Wrangling

The raw data needs some wrangling. Therefore, for both `salary_df_2019` and `salary_df_2020`, the below code does the following:

- Select only those columns **containing** `qid` in the header.
- Rename these columns as follows:
  - `import_id_qid172807697` as `salary_exp_post_grad`
  - `import_id_qid172807701_1` as `mds_self_rated_enjoy`
  - `import_id_qid96` as `salary_pre_mds`
  - `import_id_qid172807686` as `work_exp`
  - `import_id_qid98_1` as `ds_skill_confidence`
  - `import_id_qid172807685` as `does_optional_qs`
  - `import_id_qid92` as `currently_job_searching`
  - `import_id_qid99_1` as `baseline_happiness`
  - `import_id_qid93` as `freq_attend_mds_career_events`
- Create a column called `year` with the corresponding labels: `2019` or `2020`.
- Drop all observations with missing data.

Finally, we will merge `salary_df_2019` and `salary_df_2020` into a single data frame called `salary_df`.

Run the code before proceeding:

```
salary_df_2019 <- salary_df_2019 %>%
  dplyr::select(contains("qid")) %>%
  rename(
    salary_exp_post_grad = import_id_qid172807697,
    mds_self_rated_enjoy = import_id_qid172807701_1,
    salary_pre_mds = import_id_qid96,
    work_exp = import_id_qid172807686,
    ds_skill_confidence = import_id_qid98_1,
    does_optional_qs = import_id_qid172807685,
```

```
    currently_job_searching = import_id_qid92,
    baseline_happiness = import_id_qid99_1,
    freq_attend_mds_career_events = import_id_qid93
  ) %>%
  drop_na() %>%
  mutate(year = 2019)

salary_df_2020 <- salary_df_2020 %>%
  dplyr::select(contains("qid")) %>%
  drop_na() %>%
  mutate(year = 2020)
colnames(salary_df_2020) <- colnames(salary_df_2019)

salary_df <- bind_rows(salary_df_2019, salary_df_2020)
head(salary_df)
```

```
## # A tibble: 6 x 10
##   salary_exp_post_grad mds_self_rated_enjoy salary_pre_mds     work_exp
##   <chr>                               <dbl> <chr>              <chr>
## 1 $60,000 to $80,000                      3 $60,000 to $80,000 Less than 1 Year
## 2 $80,001 to $100,000                     3 $60,000 to $80,000 1 - 4 Years
## 3 $80,001 to $100,000                     3 $60,000 to $80,000 4 - 7 Years
## 4 Less than $60,000                       4 Less than $60,000  Less than 1 Year
## 5 $60,000 to $80,000                      3 $60,000 to $80,000 1 - 4 Years
## 6 $80,001 to $100,000                     3 Less than $60,000  1 - 4 Years
## # i 6 more variables: ds_skill_confidence <dbl>, does_optional_qs <chr>,
## #   currently_job_searching <chr>, baseline_happiness <dbl>,
## #   freq_attend_mds_career_events <chr>, year <dbl>
```

In `salary_df_test`, the below code does the following:

- Use `toTitleCase()` to change the level names of `freq_attend_mds_career_events` to *Title Style*.
- Change the factor level `0 - 1 Years` to `Less than 1 Year` in `work_exp`.
- Convert columns `does_optional_qs`, `currently_job_searching`, and `year` to **NOMINAL factor-type**.
- Convert the rest of the columns to **ORDERED factor-type**.
- Make sure that factors `salary_exp_post_grad`, `mds_self_rated_enjoy`, `salary_pre_mds`, `work_exp`, `ds_skill_confidence`, `baseline_happiness`, and `freq_attend_mds_career_events` have the correct level order from left to right via function `levels()`. If not, we will reorder these levels according to the order detailed at the beginning of this exercise.

Run the code before proceeding:

```
salary_df <- salary_df %>%
  mutate(
    freq_attend_mds_career_events =
      toTitleCase(freq_attend_mds_career_events)
  ) %>%
  mutate(work_exp = case_when(
    work_exp == "0 - 1 Years" ~ "Less than 1 Year",
    TRUE ~ work_exp
  )) %>%
  mutate(
    does_optional_qs = factor(does_optional_qs),
    currently_job_searching = factor(currently_job_searching),
    year = factor(year),
```

```
    salary_exp_post_grad = factor(salary_exp_post_grad, ordered = TRUE),
    mds_self_rated_enjoy = factor(mds_self_rated_enjoy, ordered = TRUE),
    salary_pre_mds = factor(salary_pre_mds, ordered = TRUE),
    work_exp = factor(work_exp, ordered = TRUE),
    ds_skill_confidence = factor(ds_skill_confidence, ordered = TRUE),
    baseline_happiness = factor(baseline_happiness, ordered = TRUE),
    freq_attend_mds_career_events = factor(freq_attend_mds_career_events,
      ordered = TRUE
    )
  ) %>%
  mutate(
    salary_exp_post_grad = fct_relevel(
      salary_exp_post_grad,
      "Less than $60,000",
      "$60,000 to $80,000",
      "$80,001 to $100,000",
      "$100,001 to $120,000",
      "More than $120,000"
    ),
    salary_pre_mds = fct_relevel(
      salary_pre_mds,
      "Less than $60,000",
      "$60,000 to $80,000",
      "$80,001 to $100,000",
      "$100,001 to $120,000",
      "More than $120,000"
    ),
    work_exp = fct_relevel(
      work_exp,
      "Less than 1 Year",
      "1 - 4 Years",
      "4 - 7 Years",
      "7 - 10 Years",
      "10+ Years"
    ),
    freq_attend_mds_career_events = fct_relevel(
      freq_attend_mds_career_events,
      "Not Often",
      "Sometimes",
      "Often"
    )
  )
```

Q2.1. Exploratory Data Analysis

rubric={accuracy:4,viz:9,reasoning:9}

Make eight suitable plots of `salary_exp_post_grad` versus each one of the rest of factor-type variables except `year`. Nonetheless, in these eight plots, include panels per `year`.

Note: If you are using the same class of plot eight times, build a function first.

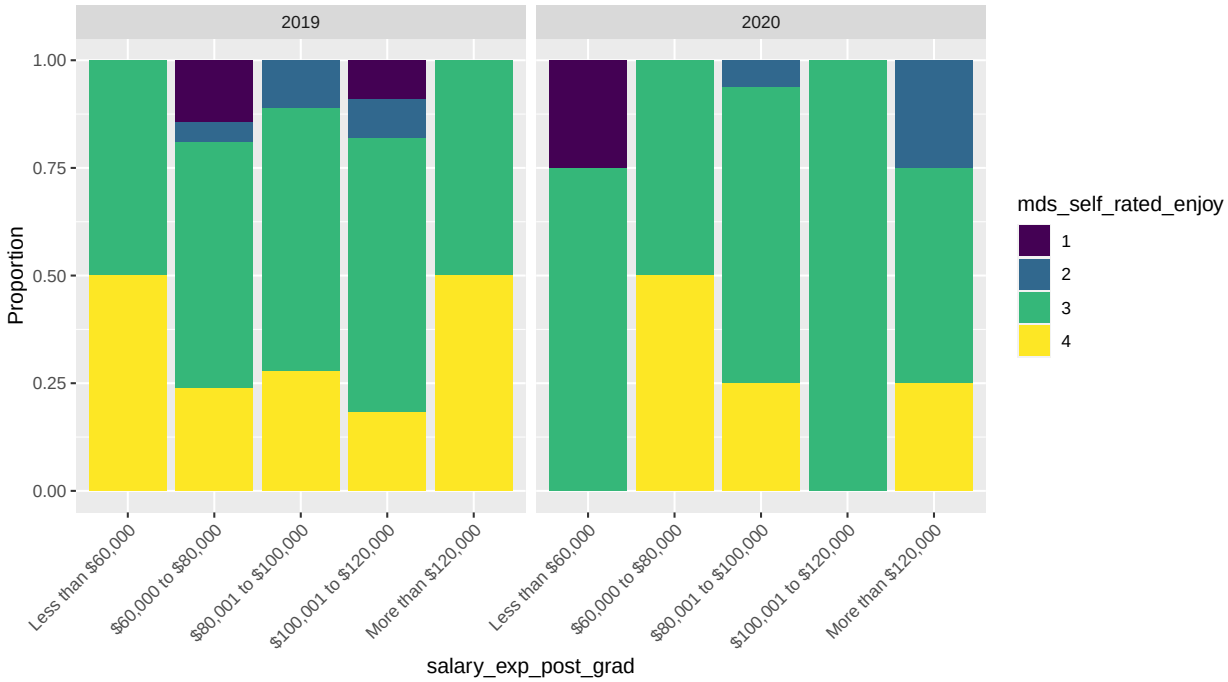
ANSWER:

```r
# Your plotting function(s).

# YOUR CODE HERE

createPlot <- function(col) {
  plot <- ggplot(salary_df, aes(x = salary_exp_post_grad, fill = .data[[col]])) +
    geom_bar(stat = "count", position = "fill") +
    facet_wrap(~year) +
    labs(y="Proportion") +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
  print(plot)
}
```

In one or two sentences BY PLOT, comment on your graphical findings by plot:

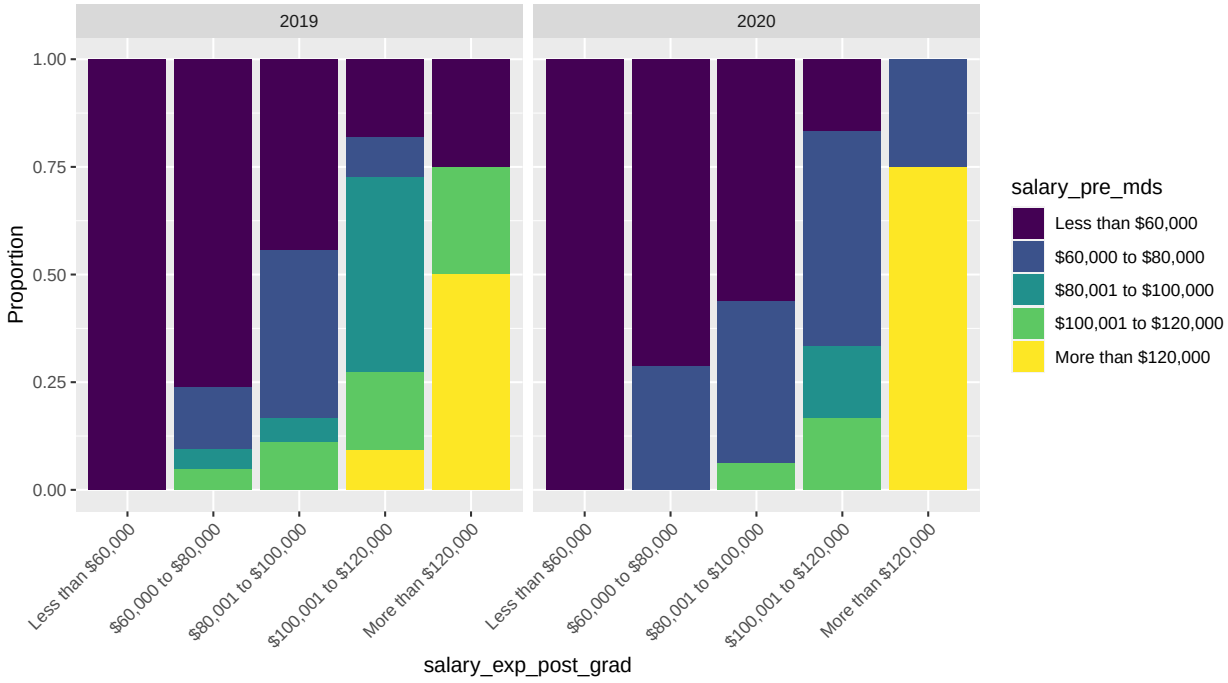
ANSWER (`mds_self_rated_enjoy` versus `salary_exp_post_grad`):

Most respondents self rated their enjoyment of MDS as either 3 or 4 and at a graphical level this does not appear to be associated with salary expectations post income or year.



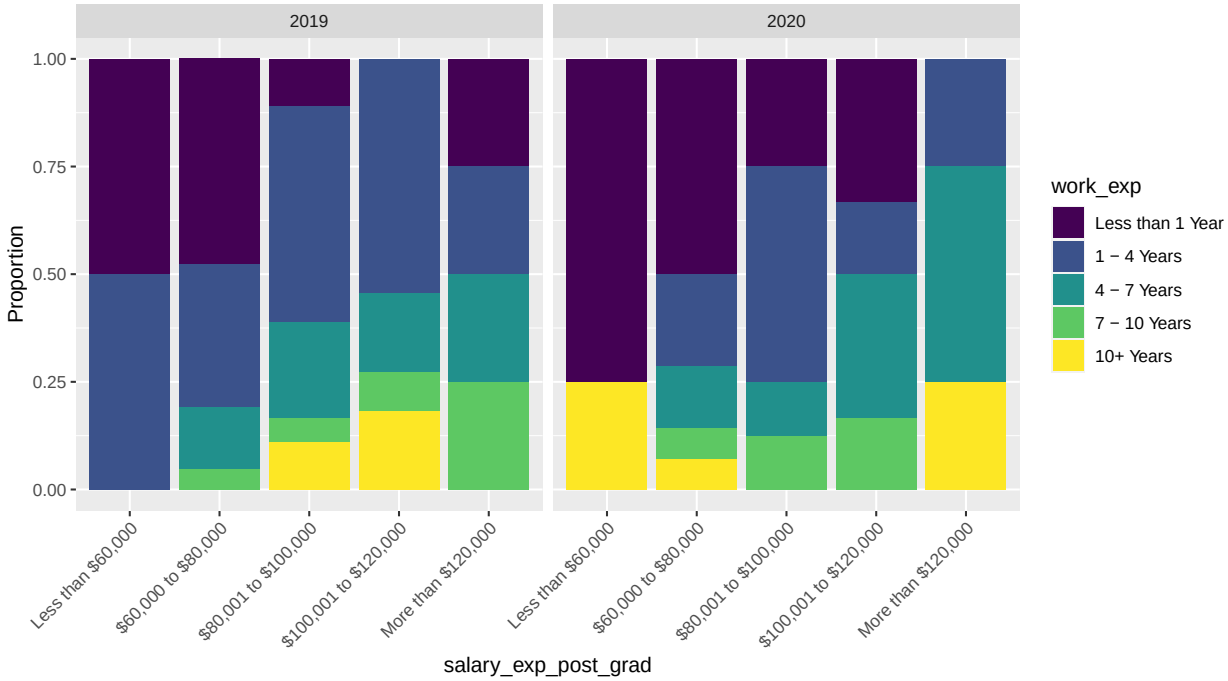
ANSWER (`salary_pre_mds` versus `salary_exp_post_grad`):

There appears to be a relationship between salary expectations post grad and pre mds, with lower salaries in before MDS associated with lower salaries expectations after graduation, and higher salaries before MDS associated with higher salaries after graduation. This is consistent across years.



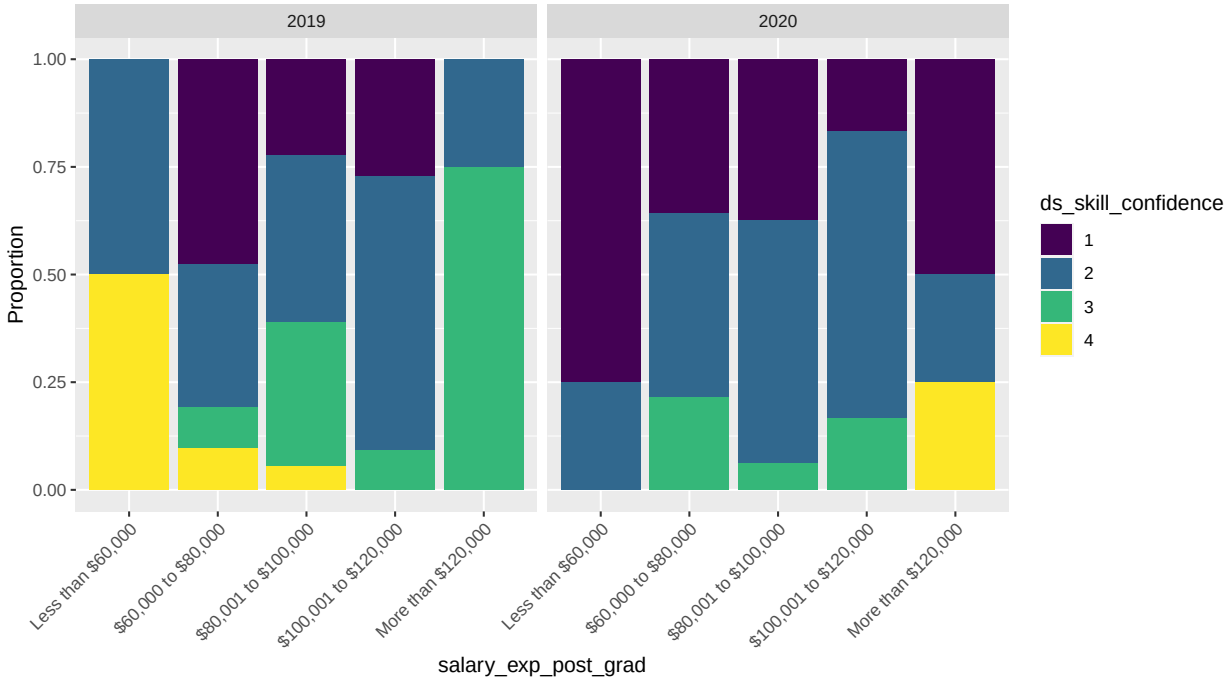
ANSWER (`work_exp` versus `salary_exp_post_grad`):

There does not appear to be a clear graphical association between the number of years of work experience and salary expectations post-graduation.



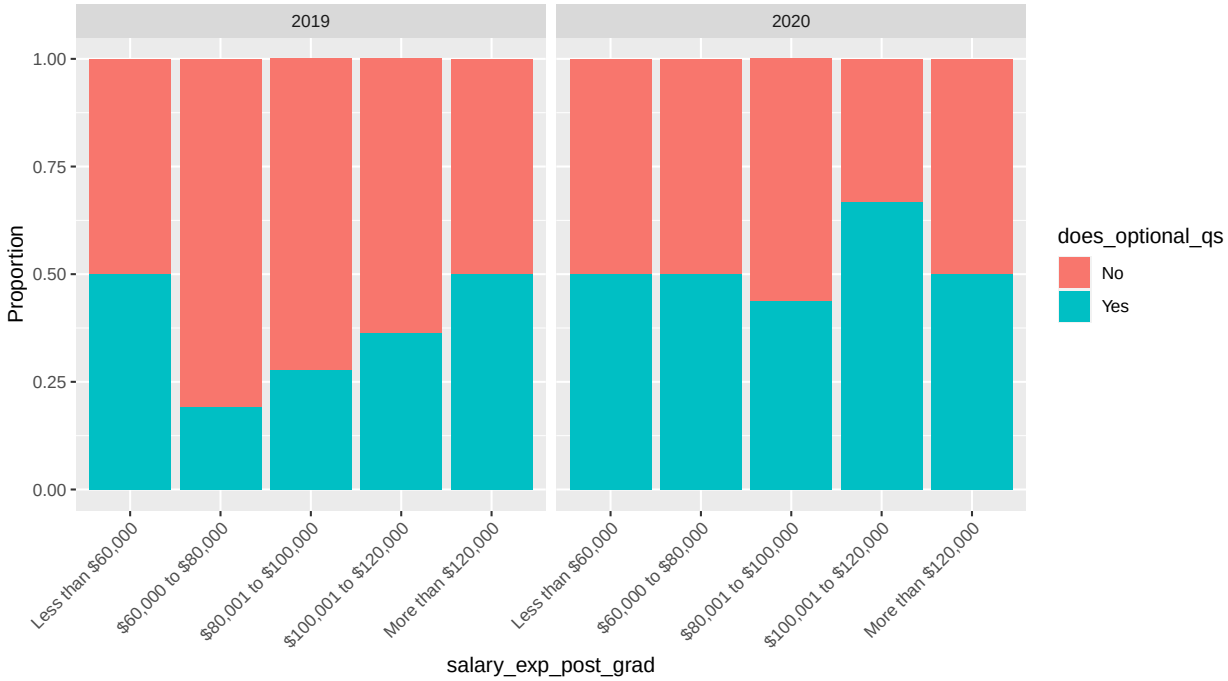
ANSWER (`ds_skill_confidence` versus `salary_exp_post_grad`):

There does not appear to be a clear graphical association between skill confidence and salary expectations post-graduation.



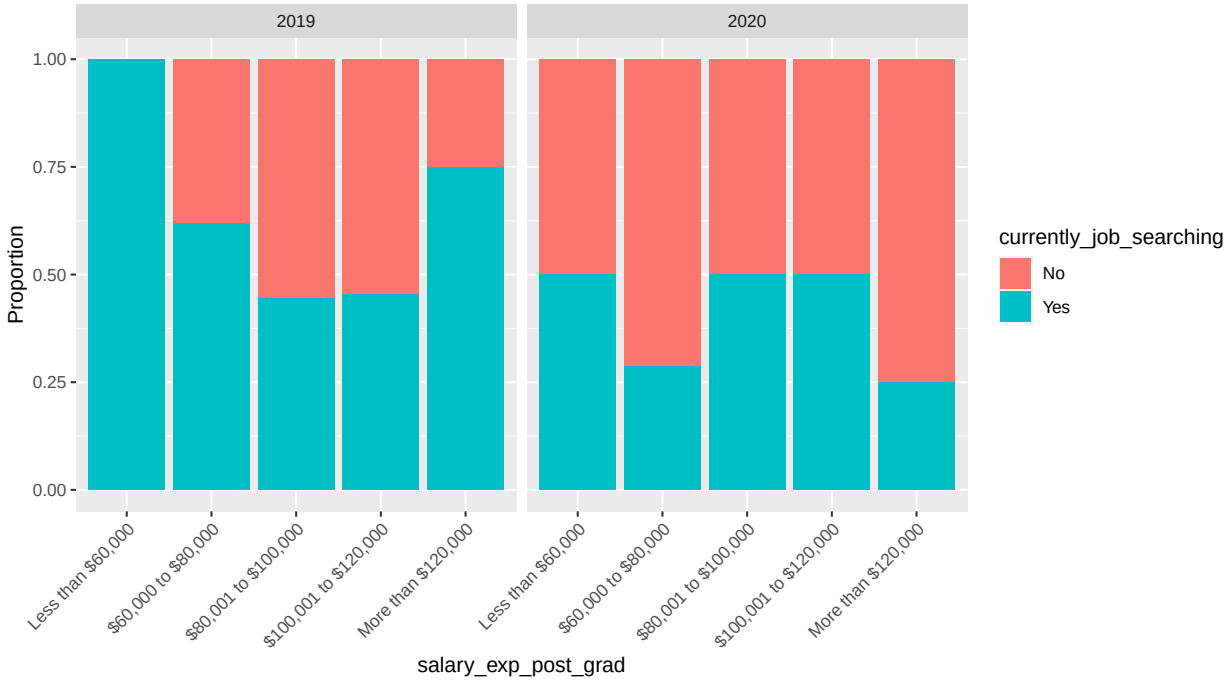
ANSWER (`does_optional_qs` versus `salary_exp_post_grad`):

There does not appear to be a clear graphical association between doing the optional questions and salary expectations post-graduation.



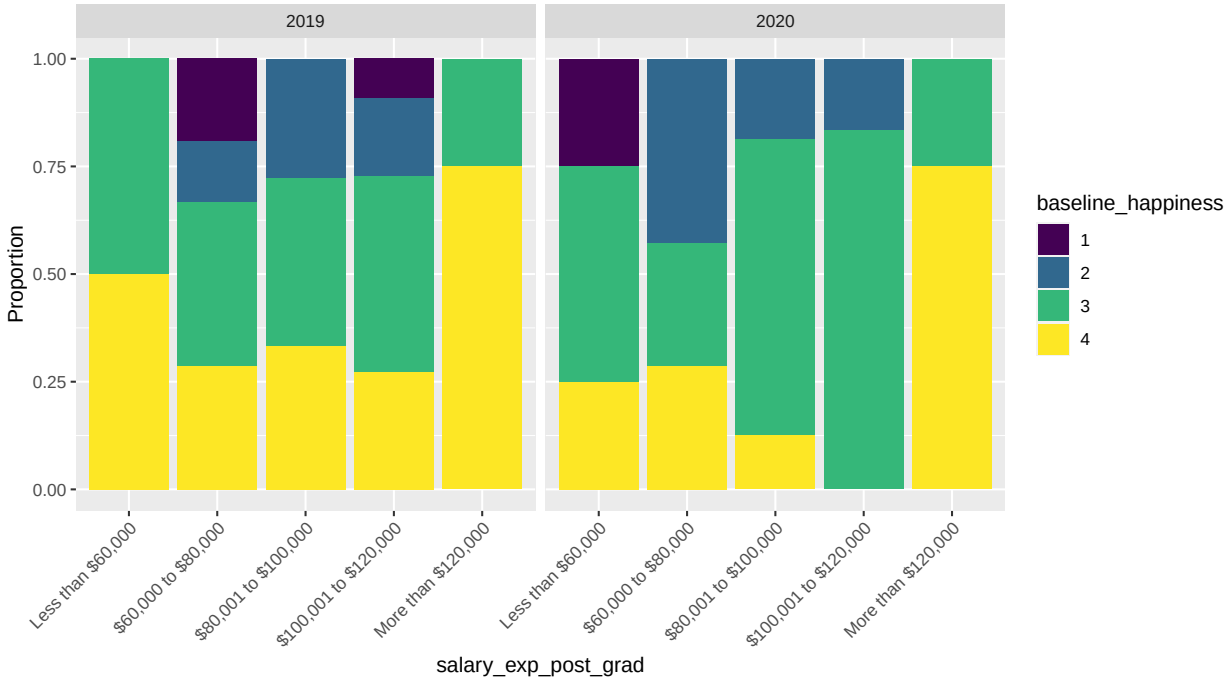
ANSWER (`currently_job_searching` versus `salary_exp_post_grad`):

There does not appear to be a clear graphical association between currently searching for a job and salary expectations post-graduation. However, in 2019, all of the respondent with salary expectations less than 60k were currently searching for work.



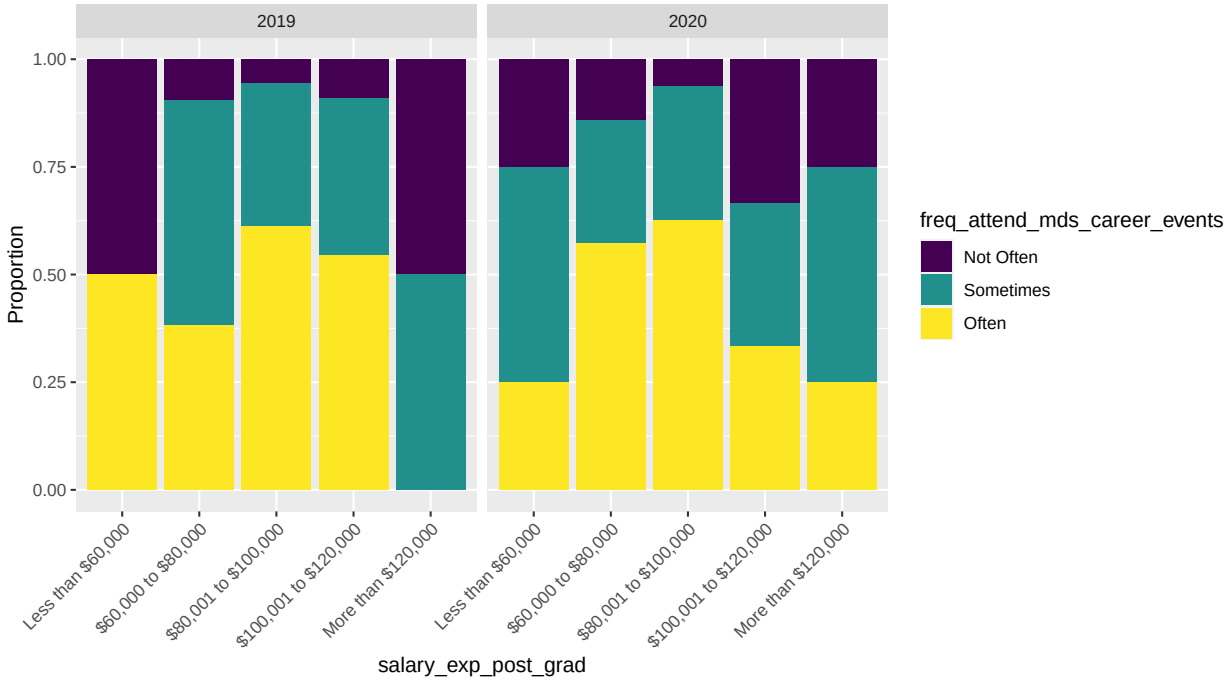
ANSWER (`baseline_happiness` versus `salary_exp_post_grad`):

There does not appear to be a clear graphical association between currently searching for a job and salary expectations post-graduation. However, graduates with the highest salary expectations tend to have a higher proportion of students with a baseline happiness of 4.



ANSWER (`freq_attend_mds_career_events` versus `salary_exp_post_grad`):

There does not appear to be a clear graphical association between frequently attending career events for a job and salary expectations post-graduation.



Q2.2. Choosing a Regression Model

rubric={reasoning:3}

Given the form in the response `salary_exp_post_grad`, what is the most suitable regression model for this survey data? What function on `R` would you use here? **Answer in two o three sentences.**

ANSWER:

Since we have an ordinal response with increasing salary expectations, we could consider using an ordinal logistic regression model. We could use the R function polr().

Q2.3. Naive Data Modelling

rubric={accuracy:6,reasoning:6}

Let us begin by just using the X (`mds_self_rated_enjoy`) and Y (`salary_exp_post_grad`) of interest in this observational study. Given your response in **Q2.2**, estimate a regression model with these two variables only and call it `initial_model`.

However, before starting with the model fitting, it is important to highlight something regarding `mds_self_rated_enjoy` and the rest of the **non-binary** survey questions . These variables are ordinal. Thus, when fitting a regression with them as explanatory variables (regardless of whether the regression is ordinary least-squares, Binary Logistic, count-type, etc.), we are likely interested in assessing the statistical significance of the difference between their ordered levels along with the corresponding interpretation. This will take us to the concept of contrasts.

By default, `R` reports the coefficients for the contrasts based on orthogonal polynomials **in ordered-type factors**. Roughly speaking, using polynomial contrasts in an ordered factor of k levels, we would fit $k-1$ polynomials. Then, we would statistically assess whether any of the polynomial fits of that factor vary with the response Y. For instance, if we have four levels in an ordered factor, we would fit linear (`L`), quadratic (`Q`), and cubic (`C`) polynomials. Then, for each one of these polynomials, we would ask how much the relationship with the response looks like a line (`L`), a parabola (`Q`), and a cubic (`C`). Note that this class of modelling assumes **equally-spaced levels** in the ordered factor.

However, to make interpretation more straightforward, there is an alternative contrast modelling. This modelling is called the **successive difference** contrasts. If we want to answer whether there are differences between the ordered levels, we will check these successive difference contrasts. The model estimates in these contrasts are the differences between the means of the second and first levels, the third and second levels, etc. We have to set up the `R` contrasts setting as follows:

```
# Run this code before proceeding.
options(contrasts = c("contr.treatment", "contr.sdif"))
```

Show the model's summary via the function `tidy()`. Do not forget to calculate the p-values for the regression coefficients along with the adjusted p-values using a 5% false discovery rate adjustment via the Benjamini-Hochberg procedure (**check the below note**).

> **Note:** In terms of regression analysis, it is usual to work with the nominal p-values (i.e., the raw p-values obtained from estimating the regression model) when testing multiple regression coefficients. Nonetheless, there has been work in the literature on adjusting for multiple testing in these frameworks to prevent false positives. **For the sake of this case study**, let us assume we want to control for this. You can find more information on this matter in **Mundfrom et al. (2006)**.

Depending on the regression model you indicated in **Q2.2**, **DSCI 562 lecture notes** will be helpful with the `R` syntax.

```
# YOUR CODE HERE
library(MASS)

initial_model <- MASS::polr(salary_exp_post_grad ~ mds_self_rated_enjoy, data=salary_df, Hess = TRUE)

tidy_initial_model <- tidy(initial_model)

tidy_initial_model <- tidy_initial_model |>
  mutate(p_raw = 2*pnorm(abs(statistic), lower.tail = F)) |>
  mutate(p_adj = p.adjust((p_raw), method="fdr"))

tidy_initial_model
```

```
## # A tibble: 7 x 7
##   term                    estimate std.error statistic coef.type   p_raw   p_adj
##   <chr>                      <dbl>     <dbl>     <dbl> <chr>       <dbl>   <dbl>
## 1 mds_self_rated_enjoy2-1    2.30      1.16      1.98  coeffici~ 4.72e-2 8.26e-2
## 2 mds_self_rated_enjoy3-2   -0.681     0.748    -0.910 coeffici~ 3.63e-1 4.99e-1
## 3 mds_self_rated_enjoy4-3   -0.335     0.422    -0.793 coeffici~ 4.28e-1 4.99e-1
## 4 Less than $60,000|$60,~   -2.66      0.468    -5.68  scale     1.36e-8 4.76e-8
## 5 $60,000 to $80,000|$80~   -0.170     0.322    -0.528 scale     5.97e-1 5.97e-1
## 6 $80,001 to $100,000|$1~    1.34      0.344     3.89  scale     1.02e-4 2.37e-4
## 7 $100,001 to $120,000|M~    2.69      0.448     6.01  scale     1.87e-9 1.31e-8
```

By looking at the model's summary in `initial_model` on the adjusted p-values, what can you conclude on the statistical relationship between self-rated enjoyment of MDS with the salary expectation after MDS graduation? **Answer in two o three sentences.**

ANSWER:

None of the levels of self-rated enjoyment of MDS were statistically significant under this model. This suggests that there was no relationship between self-rated enjoyment of MDS and salary expectation after MDS graduation.

Given your results in the `initial_model` summary, **comment in two o three sentences** on how adding the rest of the survey questions (as stratified confounders per se) will benefit your observational study in

identifying causality.

ANSWER:

The addition of stratified confounders to the model will help separate MDS enjoyment from other factors known to influence salary expectations. This will help us separate participants into similar groups based on the confounders, which is similar to blocking and helps us identify the effects of the variable of interest.

Q2.4. Full Data Modelling

rubric={accuracy:4,reasoning:5}

Now, estimate regression model called `full_model` of `salary_exp_post_grad` versus `mds_self_rated_enjoy` **along with the rest of the survey questions and year as STANDALONE stratified confounders per se (i.e., no need to do any other strata manipulation)**. Show the model's summary via the function `tidy()`.

Do not forget to calculate the p-values for the regression coefficients along with the adjusted p-values using a 5% false discovery rate adjustment via the Benjamini-Hochberg procedure.

```
# YOUR CODE HERE

full_model <- MASS::polr(salary_exp_post_grad ~ mds_self_rated_enjoy + salary_pre_mds + work_exp + ds_sl

tidy_full_model <- tidy(full_model)

tidy_full_model <- tidy_full_model |>
  mutate(p_raw = 2*pnorm(abs(statistic), lower.tail = F)) |>
  mutate(p_adj = p.adjust((p_raw), method="fdr"))

tidy_full_model
```

```
## # A tibble: 26 x 7
##    term                    estimate std.error statistic coef.type   p_raw  p_adj
##    <chr>                      <dbl>     <dbl>     <dbl> <chr>       <dbl>  <dbl>
##  1 mds_self_rated_enjoy2-1    3.36      1.59      2.11  coeffici~ 0.0350  0.108
##  2 mds_self_rated_enjoy3-2   -2.94      1.01     -2.90  coeffici~ 0.00371 0.0199
##  3 mds_self_rated_enjoy4-3   -0.530     0.616    -0.861 coeffici~ 0.389   0.633
##  4 salary_pre_mds$60,000 ~    1.16      0.558     2.08  coeffici~ 0.0373  0.108
##  5 salary_pre_mds$80,001 ~    2.66      0.920     2.89  coeffici~ 0.00382 0.0199
##  6 salary_pre_mds$100,001~   -0.221     1.04     -0.213 coeffici~ 0.831   0.852
##  7 salary_pre_mdsMore tha~    4.45      1.60      2.79  coeffici~ 0.00529 0.0229
##  8 work_exp1 - 4 Years-Le~    1.58      0.584     2.70  coeffici~ 0.00687 0.0255
##  9 work_exp4 - 7 Years-1 ~   -0.344     0.620    -0.555 coeffici~ 0.579   0.753
## 10 work_exp7 - 10 Years-4~   -0.612     0.900    -0.680 coeffici~ 0.496   0.679
## # i 16 more rows
```

```
tidy_full_model |>
  filter(p_adj < 0.05)
```

```
## # A tibble: 7 x 7
##   term                  estimate std.error statistic coef.type    p_raw    p_adj
##   <chr>                    <dbl>     <dbl>     <dbl> <chr>        <dbl>    <dbl>
## 1 mds_self_rated_enjoy~    -2.94     1.01      -2.90 coeffici~ 3.71e- 3 1.99e- 2
## 2 salary_pre_mds$80,00~     2.66     0.920      2.89 coeffici~ 3.82e- 3 1.99e- 2
## 3 salary_pre_mdsMore t~     4.45     1.60       2.79 coeffici~ 5.29e- 3 2.29e- 2
## 4 work_exp1 - 4 Years-~     1.58     0.584      2.70 coeffici~ 6.87e- 3 2.55e- 2
## 5 Less than $60,000|$6~    -5.64     0.796     -7.08 scale     1.42e-12 3.69e-11
```

```
## 6 $60,000 to $80,000|$~    -2.32     0.669     -3.46 scale     5.39e- 4 4.67e- 3
## 7 $100,001 to $120,000~     2.86     0.709      4.04 scale     5.40e- 5 7.02e- 4
```

By looking at the model's summary in `full_model` on the adjusted p-values, what can you conclude on the statistical relationship between self-rated enjoyment of MDS with the salary expectation after MDS graduation? Also, comment on the relationship between the salary expectation after MDS graduation and the confounders.

Comment in one or two paragraphs.

ANSWER:

The transition from level 2 to 3 in self-rated enjoyment shows a significant negative relationship with salary expectation (p-adj = 0.0199), indicating that increasing enjoyment from level 2 to level 3 decreases the likelihood of expecting a higher salary bracket. For the other transitions in self-rated enjoyment of the MDS program, such as moving from level 1 to 2 or from level 3 to 4, the relationship with salary expectations after graduation was not statistically significant.

Is there a statistical difference between the data of both years? **Answer in one or two sentences.**

ANSWER:

No, there is no statistical difference between the years as the p-value was higher than the significance level.

Q2.5. Confounding Stratification

rubric={reasoning:4}

In three or four sentences, explain the respondents' stratification in the `full_model` made by the confounders (with `year` included). State any assumptions made on the `full_model`.

The respondents were stratified into categories based on their responses to the confounding variables: salary_pre_mds, work_exp, ds_skill_confidence, does_optional_qs, currently_job_searching, baseline_happiness, freq_attend_mds_career_events, and year. Each confounder was split into 2-4 groups depending on the type of data. The assumption is that the groups are balanced, with equal variance between groups.

Q2.6. Primary Model Selection

rubric={accuracy:1,reasoning:3}

It is necessary to statistically check whether `full_model` provides a better data fit than the `initial_model`. Using $\alpha = 0.05$, conduct the corresponding hypothesis testing (**provide the necessary code**). Do not forget to specify the corresponding hypotheses and conclusion **in your written answer**.

ANSWER:

The goodness of fit test gave a p-value < 0.05, meaning we can reject the null hypothesis in favour of the alternative. Therefore, we can conclude that the full model fits the data better than the initial model.

```
# YOUR CODE HERE

anova(initial_model, full_model)
```

```
## Likelihood ratio tests of ordinal regression models
##
## Response: salary_exp_post_grad
##
## 1
## 2 mds_self_rated_enjoy + salary_pre_mds + work_exp + ds_skill_confidence + does_optional_qs + curren
##   Resid. df Resid. Dev   Test    Df LR stat.    Pr(Chi)
```

```
## 1        93   276.6276
## 2        74   198.9610 1 vs 2    19 77.66663 4.6786e-09
```

Q2.7. Reduced Data Modelling

rubric={accuracy:4,reasoning:3}

Based on your results in the `full_model` obtained in **Q2.4**, estimate a third regression model called `reduced_model`. This model should only contain those **STANDALONE** explanatory variables that turned out to have at least one significant regression coefficient.

Show the model's summary via the function `tidy()`. Do not forget to calculate the p-values for the regression coefficients along with the adjusted p-values using a 5% false discovery rate adjustment via the Benjamini-Hochberg procedure.

Comment on your statistically significant results **in two or three sentences**.

ANSWER:

The transition from satisfaction level 1 to 2 now leads to a statistically significant increase in salary expectation, which was not detected in the full model. Removing the stratified confounders that were present in the full model resulted in more statistically significant findings. The other comparison that become statistically significant was the difference between earning 60-80k and less than 60k.

```
# YOUR CODE HERE

reduced_model <- MASS::polr(salary_exp_post_grad ~ mds_self_rated_enjoy + salary_pre_mds + work_exp, da

tidy_reduced_model <- tidy(reduced_model)

tidy_reduced_model <- tidy_reduced_model |>
  mutate(p_raw = 2*pnorm(abs(statistic), lower.tail = F)) |>
  mutate(p_adj = p.adjust((p_raw), method="fdr"))

tidy_reduced_model
```

```
## # A tibble: 15 x 7
##    term                 estimate std.error statistic coef.type    p_raw    p_adj
##    <chr>                   <dbl>     <dbl>     <dbl> <chr>        <dbl>    <dbl>
##  1 mds_self_rated_enjo~    3.79      1.22      3.10  coeffici~ 1.92e- 3 7.21e- 3
##  2 mds_self_rated_enjo~   -2.32      0.865    -2.69  coeffici~ 7.23e- 3 1.81e- 2
##  3 mds_self_rated_enjo~   -0.390     0.472    -0.827 coeffici~ 4.08e- 1 5.79e- 1
##  4 salary_pre_mds$60,0~    1.35      0.533     2.53  coeffici~ 1.15e- 2 2.18e- 2
##  5 salary_pre_mds$80,0~    2.07      0.827     2.51  coeffici~ 1.21e- 2 2.18e- 2
##  6 salary_pre_mds$100,~   -0.128     0.968    -0.132 coeffici~ 8.95e- 1 8.95e- 1
##  7 salary_pre_mdsMore ~    4.33      1.46      2.97  coeffici~ 2.94e- 3 8.81e- 3
##  8 work_exp1 - 4 Years~    1.32      0.533     2.48  coeffici~ 1.31e- 2 2.18e- 2
##  9 work_exp4 - 7 Years~   -0.456     0.588    -0.775 coeffici~ 4.38e- 1 5.79e- 1
## 10 work_exp7 - 10 Year~   -0.614     0.837    -0.734 coeffici~ 4.63e- 1 5.79e- 1
## 11 work_exp10+ Years-7~   -0.593     1.07     -0.556 coeffici~ 5.79e- 1 6.68e- 1
## 12 Less than $60,000|$~   -5.63      0.699    -8.06  scale     7.45e-16 1.12e-14
## 13 $60,000 to $80,000|~   -2.53      0.535    -4.72  scale     2.32e- 6 1.74e- 5
## 14 $80,001 to $100,000~   -0.172     0.468    -0.366 scale     7.14e- 1 7.65e- 1
## 15 $100,001 to $120,00~    2.40      0.602     3.98  scale     6.75e- 5 3.38e- 4
```

```
tidy_reduced_model |>
  filter(p_adj < 0.05)
```

```
## # A tibble: 9 x 7
##   term                  estimate std.error statistic coef.type    p_raw    p_adj
##   <chr>                    <dbl>     <dbl>     <dbl> <chr>        <dbl>    <dbl>
## 1 mds_self_rated_enjoy~     3.79     1.22       3.10 coeffici~ 1.92e- 3 7.21e- 3
## 2 mds_self_rated_enjoy~    -2.32     0.865     -2.69 coeffici~ 7.23e- 3 1.81e- 2
## 3 salary_pre_mds$60,00~     1.35     0.533      2.53 coeffici~ 1.15e- 2 2.18e- 2
## 4 salary_pre_mds$80,00~     2.07     0.827      2.51 coeffici~ 1.21e- 2 2.18e- 2
## 5 salary_pre_mdsMore t~     4.33     1.46       2.97 coeffici~ 2.94e- 3 8.81e- 3
## 6 work_exp1 - 4 Years-~     1.32     0.533      2.48 coeffici~ 1.31e- 2 2.18e- 2
## 7 Less than $60,000|$6~    -5.63     0.699     -8.06 scale     7.45e-16 1.12e-14
## 8 $60,000 to $80,000|$~    -2.53     0.535     -4.72 scale     2.32e- 6 1.74e- 5
## 9 $100,001 to $120,000~     2.40     0.602      3.98 scale     6.75e- 5 3.38e- 4
```

Q2.8. Secondary Model Selection

rubric={accuracy:4,reasoning:3}

As in **Q2.6**, make pairwise comparisons between `initial_model`, `full_model`, and `reduced_model`. Do not forget to correct for multiple testing (using a 5% false discovery rate adjustment via the Benjamini-Hochberg procedure).

Based on these comparisons, which model would you finally choose? **Answer in one or two sentences.**

ANSWER:

After the three goodness of fit tests, it was found that at a significance level of 0.05, the reduced and full models fit the data better than the initial model. However, there the full model does not provide a statistically significant improvement on the reduced model. Therefore, I would opt for the reduced model as it is simpler.

```
first <- as.data.frame(anova(initial_model, full_model))
first$Comparison <- "Full vs Initial"

second <- as.data.frame(anova(initial_model, reduced_model))
second$Comparison <- "Reduced vs Initial"

third <- as.data.frame(anova(reduced_model, full_model))
third$Comparison <- "Full vs Reduced"

df <- rbind(first, second, third)

df <- df |>
  rename(p_val = `Pr(Chi)`) |>
  mutate(
    p_adj = p.adjust(p_val, method="fdr"),
    method="fdr"
  )

df
```

```
##
## 1
## 2 mds_self_rated_enjoy + salary_pre_mds + work_exp + ds_skill_confidence + does_optional_qs + curren
## 3
## 4
## 5
## 6 mds_self_rated_enjoy + salary_pre_mds + work_exp + ds_skill_confidence + does_optional_qs + curren
##   Resid. df Resid. Dev   Test    Df LR stat.        p_val         Comparison
```

```
## 1        93   276.6276           NA       NA           NA    Full vs Initial
## 2        74   198.9610 1 vs 2    19 77.66663 4.678600e-09    Full vs Initial
## 3        93   276.6276           NA       NA           NA Reduced vs Initial
## 4        85   209.0257 1 vs 2     8 67.60189 1.472833e-11 Reduced vs Initial
## 5        85   209.0257           NA       NA           NA    Full vs Reduced
## 6        74   198.9610 1 vs 2    11 10.06474 5.245731e-01    Full vs Reduced
##          p_adj method
## 1           NA    fdr
## 2 7.017901e-09    fdr
## 3           NA    fdr
## 4 4.418499e-11    fdr
## 5           NA    fdr
## 6 5.245731e-01    fdr
```

Q2.9. Inferential Conclusions

rubric={reasoning:8}

Suppose you **attempt** to use the chosen model in **Q2.8 to explain causality** between `salary_exp_post_grad` and `mds_self_rated_enjoy`, along with the corresponding confounders. Interpret and communicate this chosen model in 300-500 words.

> **Heads-Up:** Interpreting the regression coefficients is optional in this part. You can do it if you consider it necessary for your general conclusion.

ANSWER:

In this analysis, the chosen reduced model was utilized to explore associations between the self-rated enjoyment of the MDS program and expected salary upon graduation, while accounting for key confounders like previous salary and work experience. This model was selected for its balance of simplicity, allowing us to focus on the most impactful predictors.

Moving from level 1 to 2 satisfaction is associated statistically significant increase in salary expectations. This suggests that initial improvements in program satisfaction can positively impact students' salary expectations, potentially indicating that initial positive experiences in the program might boost confidence or perceived value from the education, influencing their career outlook positively.

Interestingly, increasing satisfaction from level 2 to 3 correlates with a decrease in salary expectations. One possible explanation for this is that less skilled students enjoy the program more as it challenges them, but they have lower salary expectations because they have less confidence in their skills. However, this is merely a hypothesis for future investigation and would require more potentially cofounding variables such as confidence in abilities and skill level to be included.

While this model provides insights into the relationships between the studied variables, the observational nature of the study introduces limitations to the causal interpretations we can confidently make. Despite controlling for previous salary and work experience, which influence salary expectations, it is likely that other unmeasured confounders exist.

(Challenging) Q2.10. Study Critique

rubric={reasoning:4}

From the MDS student study above, write **one or two paragraphs** criticizing the study design and analysis. If you were to run this same study again, how would you improve it?

ANSWER:

We could rely less on self-reported data but instead try to find more objective variables, such as exam performance.

Objective measures like job placement rates after graduation or actual starting salaries

Control for country, or average salary in country of employment

Submission

CONGRATULATIONS!!!! You are done with the last lab of the statistical stream in MDS!

- Knit the assignment to generate the `.pdf` file and push everything to your Github repo.
- Double check all the figures, texts, equations are rendered properly in the `.pdf` file
- **Submit the `.pdf` file to Gradescope.**

Attribution

The question, data, and analysis that makes up the questions in Exercise 2 were derived from a survey and analysis performed by the following past MDS students:

- Carrie Cheung.
- Alex Pak.
- Talha Siddiqui.
- Evan Yathon.