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/riyaeliza123/554_lab4_prabhriya
- 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:

In a real-life context, let's consider the probability of a football team winning a match.

Odds: The odds of the team winning can be calculated as the ratio of the probability of winning to the probability of losing. For example, if the probability of winning is 0.6, then the odds of winning would be 0.6 / (1 - 0.6) = 0.6 / 0.4 = 1.5. This means that for every 1.5 times the team wins, it loses once.

Odds Ratio: The odds ratio compares the odds of an event happening in one group to the odds of the same event happening in another group. For example, if we want to compare the odds of winning for two teams, Team A and Team B, and the odds of winning for Team A are 1.5 and for Team B are 0.8, then the odds ratio would be 1.5 / 0.8 = 1.875. This means that Team A is 1.875 times more likely to win than Team B.

Log-Odds Ratio: The log-odds ratio is the natural logarithm of the odds ratio. Using the same example, if the odds ratio is 1.875, then the log-odds ratio would be ln(1.875) = 0.625. This means that the difference in the log-odds of winning between Team A and Team B is approximately 0.625.

Interpretation:

- Odds: For every 1.5 wins, the team loses once.

- Odds Ratio: Team A is 1.875 times more likely to win than Team B.

- Log-Odds Ratio: The difference in the log-odds of winning between Team A and Team B is approximately 0.625.

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: Does regular physical activity reduce the risk of developing type 2 diabetes in adults aged 40-60?

1. Cross-sectional (Contemporaneous): Conduct a survey of adults aged 40-60 to collect data on their physical activity levels and diabetes status at a single point in time. Analyze the data to determine if there is an association between physical activity and diabetes risk.

2. Case-control (Retrospective): Identify a group of adults aged 40-60 with type 2 diabetes (cases) and a control group without diabetes. Collect data on their past physical activity levels. Compare the historical physical activity levels between the two groups to assess the association between physical activity and diabetes risk.

3. Cohort (Prospective): Recruit a large cohort of adults aged 40-60 without diabetes and follow them over several years. Regularly assess their physical activity levels and track the development of new cases of type 2 diabetes. Analyze the data to determine if there is a relationship between baseline physical activity levels and the risk of developing diabetes over time.

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 the above research question, I would choose the cohort study design.

This is because cohort study allows for the assessment of physical activity levels at baseline and the subsequent development of diabetes over time, providing stronger evidence for causality compared to cross-sectional or case-control studies.

While cohort studies can be resource-intensive and require long-term follow-up, they are feasible for this research question given the relatively short time frame (40-60 years old) and the availability of tools for tracking physical activity and diabetes status over time.

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.

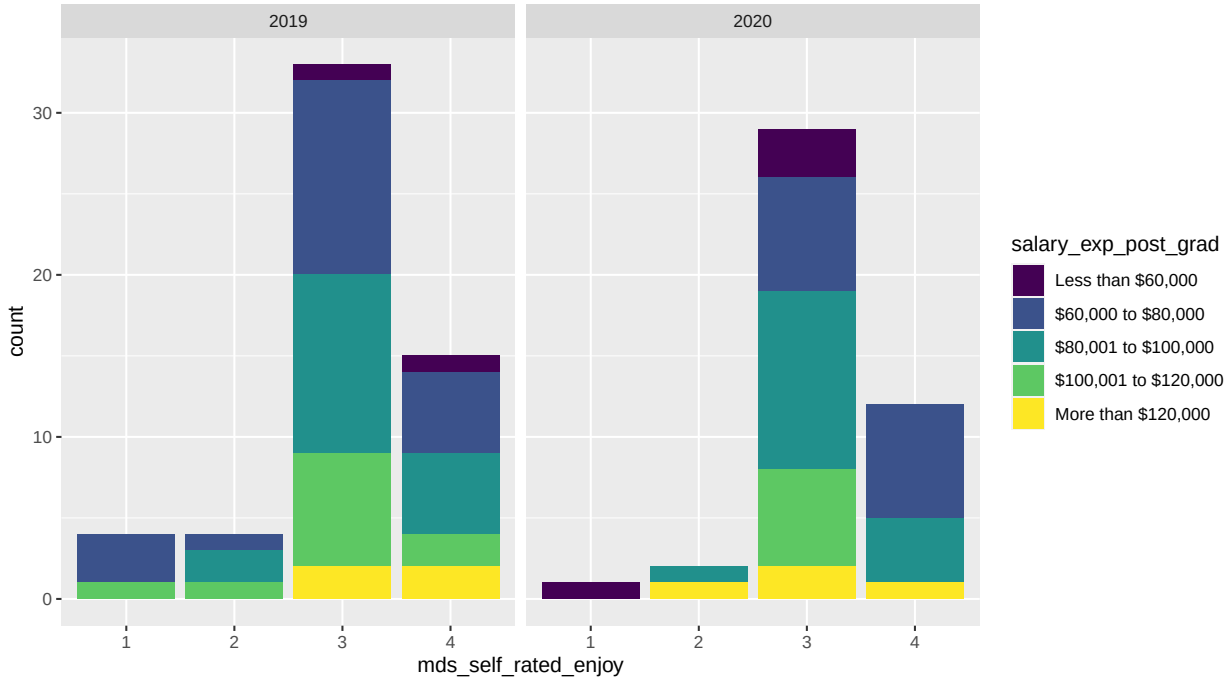
ANSWER:

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

# MAKE THIS A FUNCTION

plot_salary_vs_factor <- function(df, factor_var) {

  df[[factor_var]] <- factor(df[[factor_var]])

  plot <- ggplot(df, aes(x = .data[[factor_var]], fill = salary_exp_post_grad)) +
    geom_bar(position = "stack") +
    facet_wrap(~ year)

  return(plot)
}
```

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

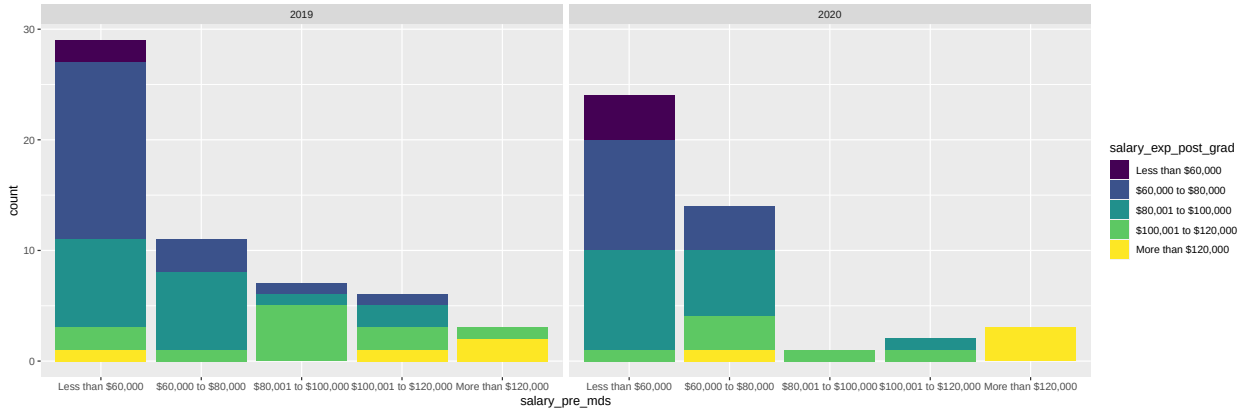
ANSWER (`mds_self_rated_enjoy` versus `salary_exp_post_grad`):

In 2019, people who enjoyed more had more salary expectation. In 2020, we see a similar trend. People who enjoyed less had lesser salary expectation in 2019 and 2020.



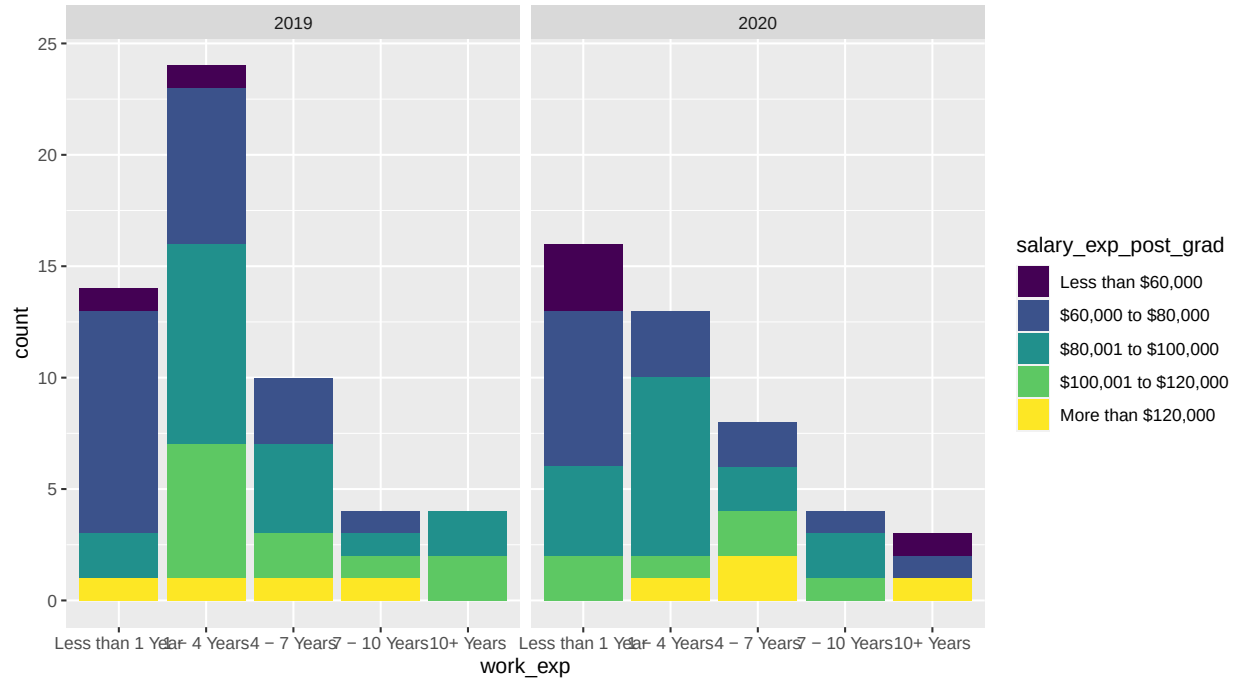
ANSWER (`salary_pre_mds` versus `salary_exp_post_grad`):

In both years, people who had a salary of 80K to 100K expect more after the program while all other categories expect the same salary or sometimes even lesser than their previous salary.



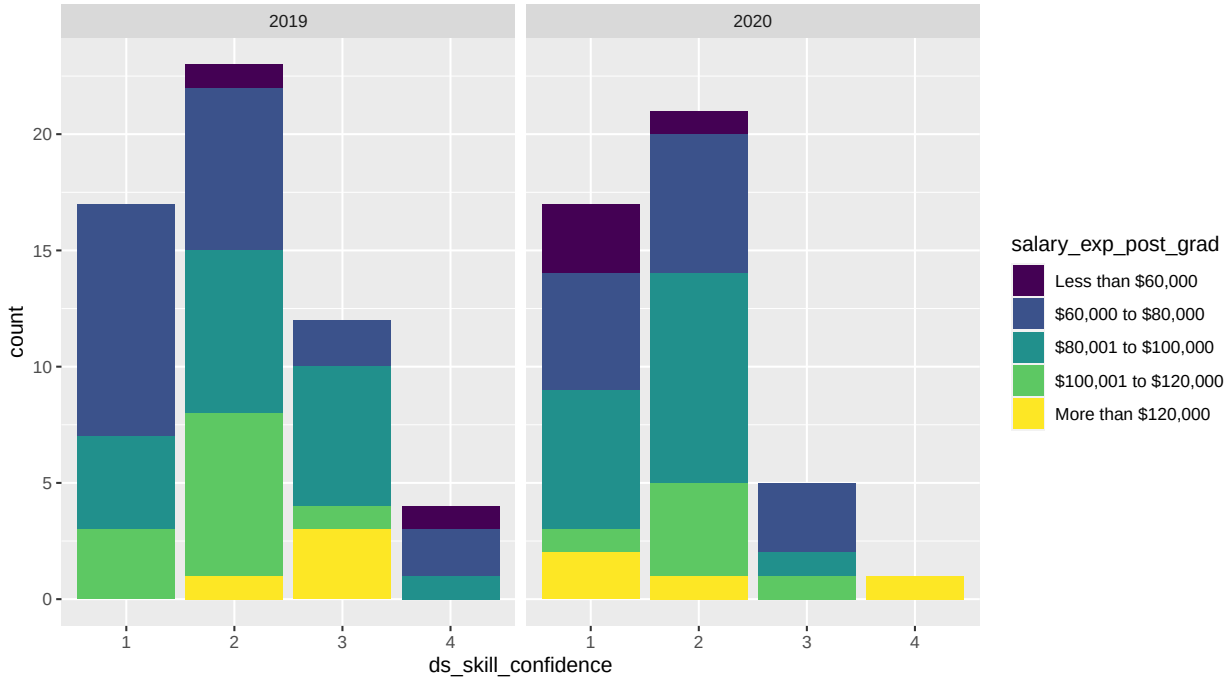
ANSWER (`work_exp` versus `salary_exp_post_grad`):

People with 4-7 years of work experience have very high salary expectations while people with 10+ years of experience have lower expectations.



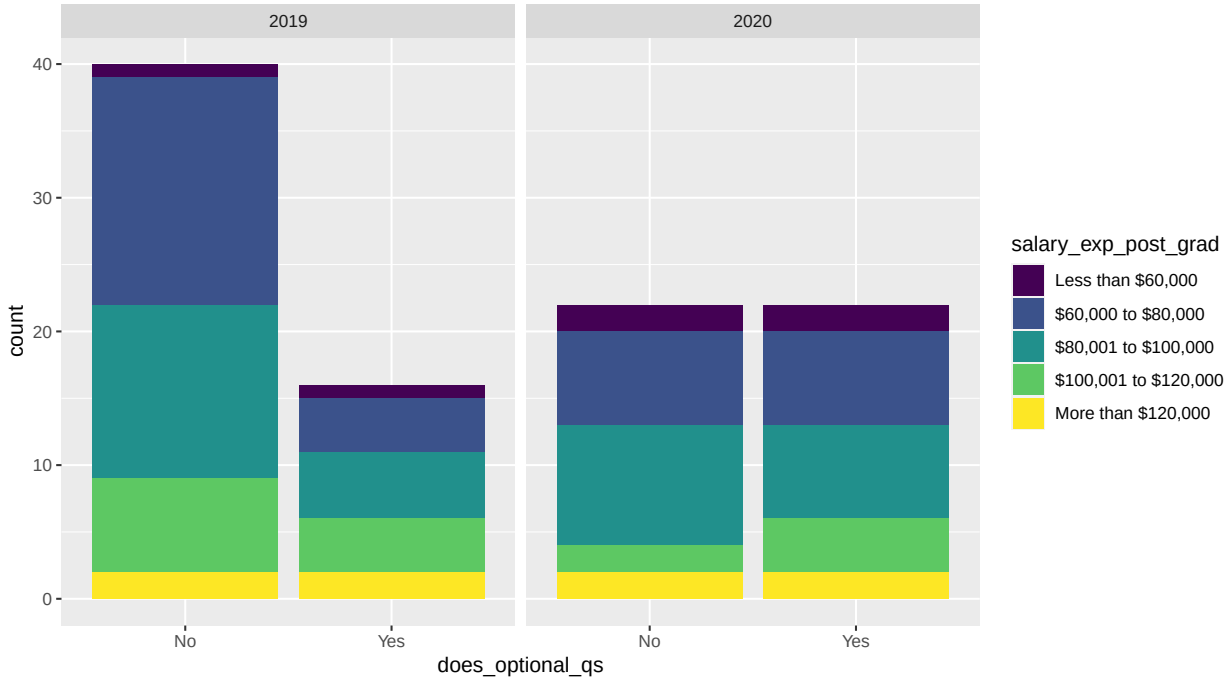
ANSWER (`ds_skill_confidence` versus `salary_exp_post_grad`):

People with more skill expect more salary. In 2020, we see some people with low skill confidence expecting 120K salary.



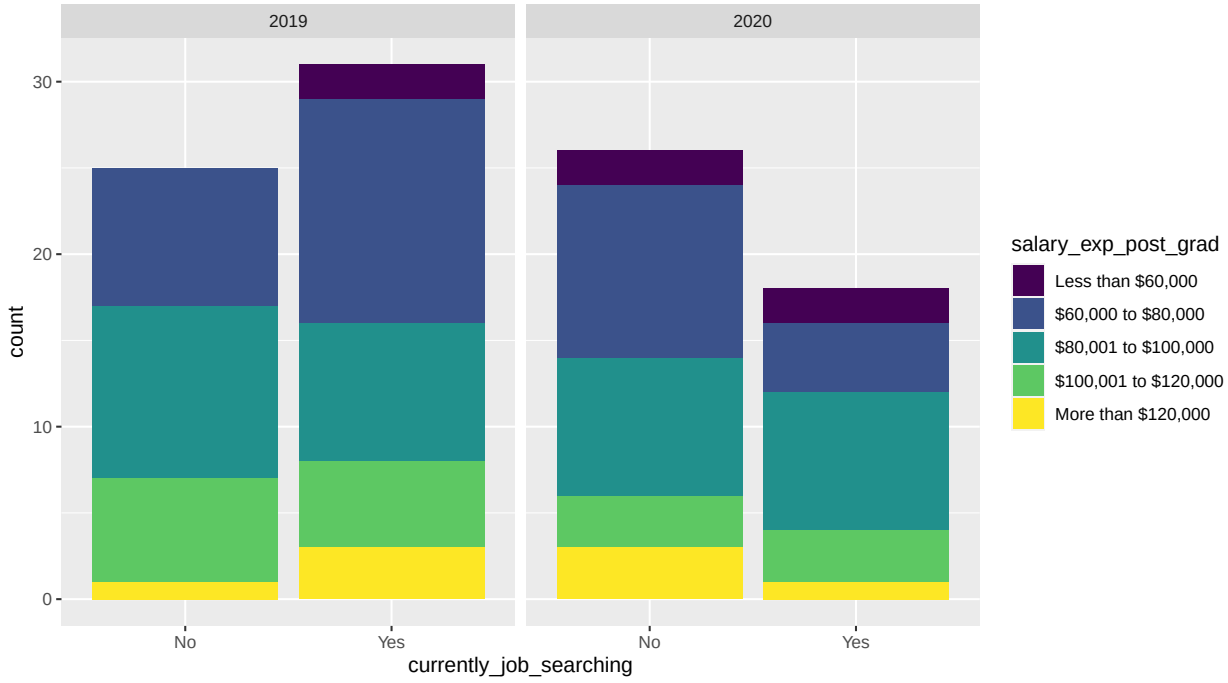
ANSWER (`does_optional_qs` versus `salary_exp_post_grad`):

We don't see much difference in proportion of people who do optional questions v/s their salary expectations. Regardless of whether or not they do the question, the expectations are mixed. No particular pattern.



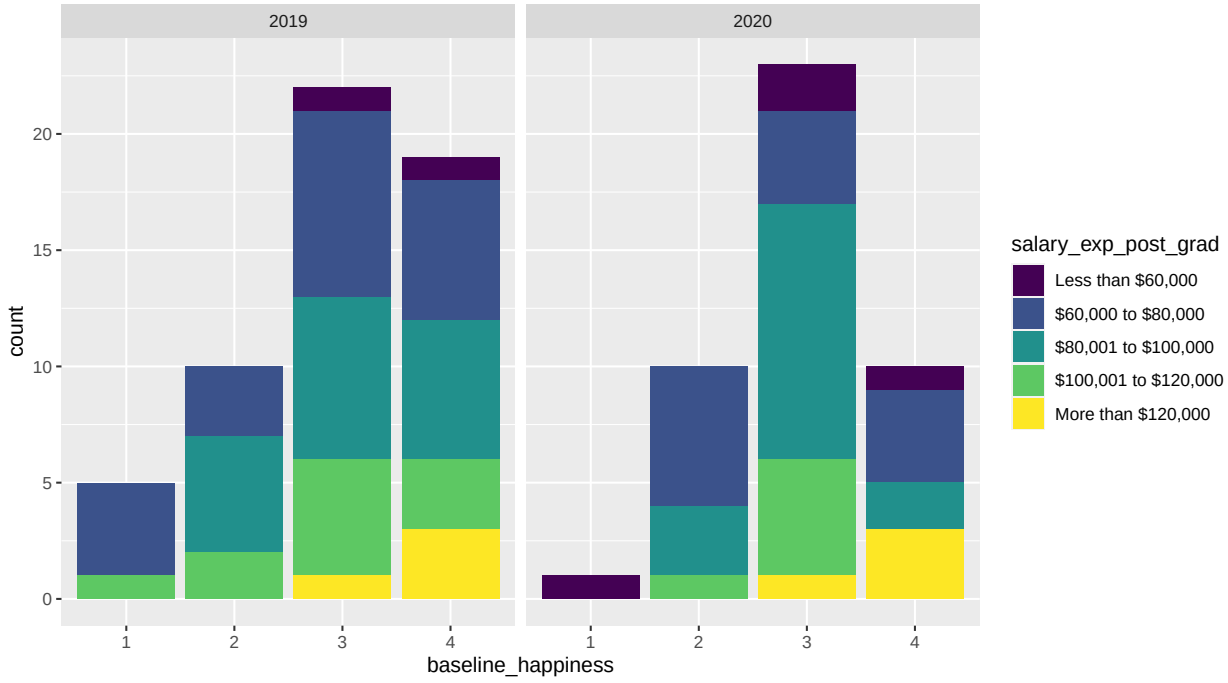
ANSWER (`currently_job_searching` versus `salary_exp_post_grad`):

People who aren't looking for a job have lower salary expectations as compared to people looking for a job. In 2019, about 50% of people looking for a job expected a 60K to 80K salary.



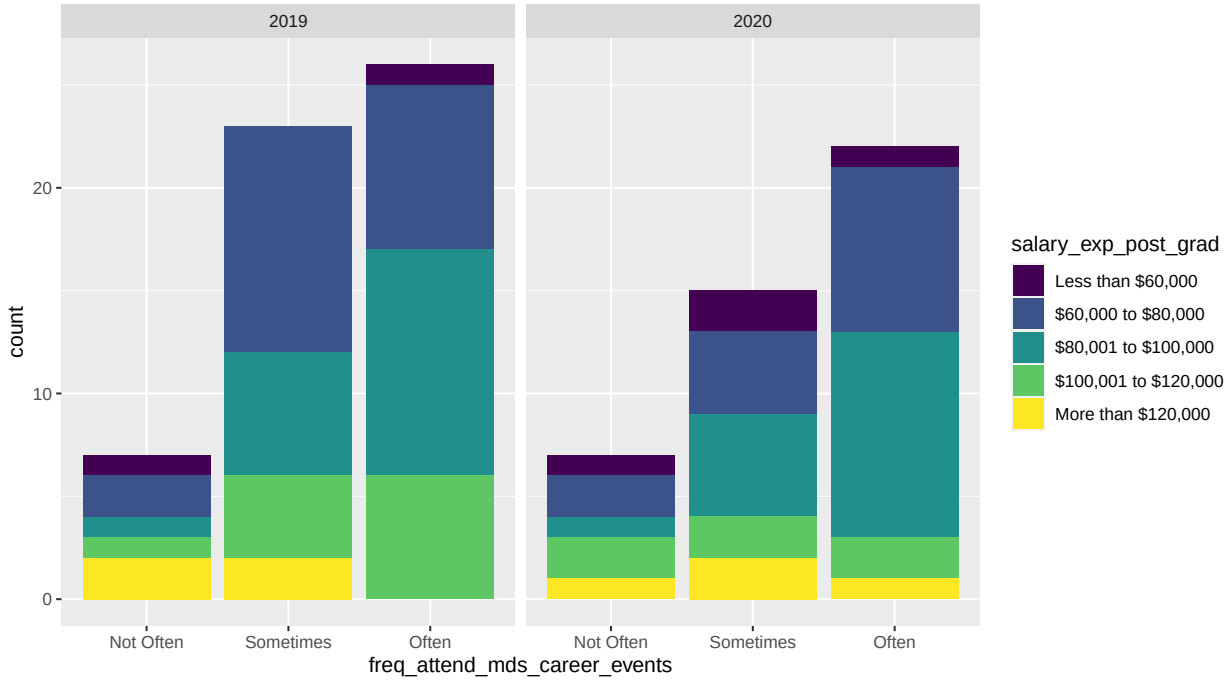
ANSWER (`baseline_happiness` versus `salary_exp_post_grad`):

Happier people have higher salary expectations. Moderately happy people have all ranges of expected salary.



ANSWER (`freq_attend_mds_career_events` versus `salary_exp_post_grad`):

People who attend career fairs often have a mid-range expectation of 80-100K. In 2019, the lesser you attender fairs, the higher the salary expectation is.



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 salary_exp_post_grad is an ordinal variable, we use ordinal regression. In R, we use the polr() function to conduct ordinal regression. This function comes from MASS package and we set argument Hess as TRUE.

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.

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

p_values <- pnorm(abs(tidy(initial_model)$statistic), lower.tail = FALSE) * 2

tidy_table <- cbind(tidy(initial_model), p_value = p_values) |>
  mutate_if(is.numeric, round, 2)

tidy_table$p.adjust <- p.adjust(tidy_table$p_value, method = "fdr")

tidy_table
```

```
##                                      term estimate std.error statistic
## 1                  mds_self_rated_enjoy2-1     2.30      1.16      1.98
## 2                  mds_self_rated_enjoy3-2    -0.68      0.75     -0.91
## 3                  mds_self_rated_enjoy4-3    -0.33      0.42     -0.79
## 4     Less than $60,000|$60,000 to $80,000    -2.66      0.47     -5.68
## 5   $60,000 to $80,000|$80,001 to $100,000    -0.17      0.32     -0.53
## 6 $80,001 to $100,000|$100,001 to $120,000     1.34      0.34      3.89
## 7  $100,001 to $120,000|More than $120,000     2.69      0.45      6.01
##     coef.type p_value  p.adjust
## 1 coefficient    0.05 0.0875000
## 2 coefficient    0.36 0.5016667
## 3 coefficient    0.43 0.5016667
## 4       scale    0.00 0.0000000
## 5       scale    0.60 0.6000000
## 6       scale    0.00 0.0000000
## 7       scale    0.00 0.0000000
```

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:

The adjusted p-values of the 3 comparisons of levels of enjoyments (mds_self_rated_enjoy) are all greater than 0.05, which suggests that the differences in self-rated enjoyment of MDS are not statistically significant in predicting salary expectation after 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:

1. Including other survey questions as confounders helps control for potential confounding variables that could influence the relationship between the predictor and outcome variables. This improves the accuracy of the estimated effect and enhances the study's ability to identify causal relationships.

2. By accounting for additional variables, the study can reduce bias that might arise from unmeasured or omitted variables.

3. Considering a broader range of factors in the analysis can lead to more generalizable results.

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.

```
full_model <- polr(salary_exp_post_grad ~ mds_self_rated_enjoy +
                     salary_pre_mds + work_exp + ds_skill_confidence +
                     does_optional_qs + currently_job_searching +
                     baseline_happiness + freq_attend_mds_career_events + year,
                   data = salary_df,
                   Hess = TRUE)

full_model_summarized <- cbind(tidy(full_model),
                               p.value = pnorm(abs(tidy(full_model)$statistic),
                                               lower.tail = FALSE) * 2 ) |>
  mutate_if(is.numeric, round, 2) |>
  mutate(p.adjust = round(p.adjust(p.value, method="fdr"), 3))

full_model_summarized
```

```
##                                                    term estimate std.error
## 1                              mds_self_rated_enjoy2-1     3.36      1.59
## 2                              mds_self_rated_enjoy3-2    -2.94      1.01
## 3                              mds_self_rated_enjoy4-3    -0.53      0.62
## 4      salary_pre_mds$60,000 to $80,000-Less than $60,000     1.16      0.56
## 5    salary_pre_mds$80,001 to $100,000-$60,000 to $80,000     2.66      0.92
## 6  salary_pre_mds$100,001 to $120,000-$80,001 to $100,000    -0.22      1.04
## 7   salary_pre_mdsMore than $120,000-$100,001 to $120,000     4.45      1.60
## 8                   work_exp1 - 4 Years-Less than 1 Year     1.58      0.58
```

```
## 9                          work_exp4 - 7 Years-1 - 4 Years    -0.34      0.62
## 10                        work_exp7 - 10 Years-4 - 7 Years    -0.61      0.90
## 11                          work_exp10+ Years-7 - 10 Years    -1.19      1.15
## 12                                  ds_skill_confidence2-1     0.37      0.50
## 13                                  ds_skill_confidence3-2     0.86      0.63
## 14                                  ds_skill_confidence4-3    -1.19      1.24
## 15                                     does_optional_qsYes     0.43      0.49
## 16                               currently_job_searchingYes     0.17      0.47
## 17                                   baseline_happiness2-1     0.26      1.42
## 18                                   baseline_happiness3-2     0.92      0.57
## 19                                   baseline_happiness4-3    -0.30      0.60
## 20       freq_attend_mds_career_eventsSometimes-Not Often     0.60      0.75
## 21           freq_attend_mds_career_eventsOften-Sometimes    -0.20      0.55
## 22                                                year2020     0.45      0.48
## 23                  Less than $60,000|$60,000 to $80,000    -5.64      0.80
## 24                $60,000 to $80,000|$80,001 to $100,000    -2.32      0.67
## 25               $80,001 to $100,000|$100,001 to $120,000     0.23      0.62
## 26                $100,001 to $120,000|More than $120,000     2.86      0.71
##    statistic   coef.type p.value p.adjust
## 1       2.11 coefficient    0.04    0.116
## 2      -2.90 coefficient    0.00    0.000
## 3      -0.86 coefficient    0.39    0.634
## 4       2.08 coefficient    0.04    0.116
## 5       2.89 coefficient    0.00    0.000
## 6      -0.21 coefficient    0.83    0.850
## 7       2.79 coefficient    0.01    0.037
## 8       2.70 coefficient    0.01    0.037
## 9      -0.55 coefficient    0.58    0.754
## 10     -0.68 coefficient    0.50    0.684
## 11     -1.03 coefficient    0.30    0.631
## 12      0.75 coefficient    0.45    0.650
## 13      1.37 coefficient    0.17    0.402
## 14     -0.96 coefficient    0.34    0.631
## 15      0.87 coefficient    0.38    0.634
## 16      0.36 coefficient    0.72    0.780
## 17      0.19 coefficient    0.85    0.850
## 18      1.62 coefficient    0.11    0.286
## 19     -0.49 coefficient    0.62    0.768
## 20      0.80 coefficient    0.42    0.642
## 21     -0.36 coefficient    0.72    0.780
## 22      0.95 coefficient    0.34    0.631
## 23     -7.08       scale    0.00    0.000
## 24     -3.46       scale    0.00    0.000
## 25      0.37       scale    0.71    0.780
## 26      4.04       scale    0.00    0.000
```

```
significant_results <- full_model_summarized |>
  filter(p.adjust < 0.05)
significant_results
```

```
##                                                term estimate std.error
## 1                                mds_self_rated_enjoy3-2    -2.94      1.01
## 2  salary_pre_mds$80,001 to $100,000-$60,000 to $80,000     2.66      0.92
## 3 salary_pre_mdsMore than $120,000-$100,001 to $120,000     4.45      1.60
## 4                   work_exp1 - 4 Years-Less than 1 Year     1.58      0.58
```

```
## 5                  Less than $60,000|$60,000 to $80,000    -5.64      0.80
## 6                $60,000 to $80,000|$80,001 to $100,000    -2.32      0.67
## 7               $100,001 to $120,000|More than $120,000     2.86      0.71
##   statistic   coef.type p.value p.adjust
## 1     -2.90 coefficient    0.00    0.000
## 2      2.89 coefficient    0.00    0.000
## 3      2.79 coefficient    0.01    0.037
## 4      2.70 coefficient    0.01    0.037
## 5     -7.08       scale    0.00    0.000
## 6     -3.46       scale    0.00    0.000
## 7      4.04       scale    0.00    0.000
```

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 to two paragraphs.

ANSWER:

For the full model, we have **4** statistically significant associations here where the p-values are less than 0.05 (alpha):

Main Association: When we jump from self rated enjoyment of 2 to 3 then the coefficient for that is significant. This indicates that is **some relationship** between explanatory variable self-rated enjoyment of MDS and the salary expectation after MDS program once we include the confounding variables.

Other Significant Associations for Salary expectation after MDS:

a) For salary before MDS we found comparison of $80,001-$100,000 to $60,000-$80,000 to be significant.

b) Similarly for salary before MDS we also found comparison of more than $120,000 to $100,001-$120,000 to be significant.

c) On comparing the work experience of less than 1 year to 1-4 years we found the association to be significant.

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

ANSWER:

There is no evidence of statistical difference between the years 2019 and 2020 as the adjusted p-value for that coefficient is greater than 0.05.

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`.

In our full model the respondents are stratified or grouped using our various confounders like work experience, salary before MDS etc. Following are the assumptions:

- All our confounding factors influence both our input self-rated enjoyment and the dependent variable salary expectations after graduation.
- No interaction between the strata which indicates a simple and smooth structure across strata for the variation in salary expectations after MDS.

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:

Null hypothesis: `initial_model` fits the data better as compared to `full_model`

Alternate hypothesis: `full_model` fits the data better as compared to `initial_model`

We see below that the p-value from Anova is less than 0.05 and hence we can reject the null hypothesis in favor of alternate hypothesis. Hence, we can say our `full_model` which includes confounding variables explains variance of our response significantly better as compared to the `initial_model`.

```
anova(full_model, initial_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 `reduced_model` has significant association between our response and all the included explanatory variables for different contrasts or ordinal value jumps.
- In comparison to our `full_model` , we have more no of significant coefficients in our `reduced_model` especially more levels of self-rated enjoyment and pre-MDS salary being significant in this new model.

```
reduced_model <- polr(salary_exp_post_grad ~ mds_self_rated_enjoy +
                        salary_pre_mds + work_exp,
                      data = salary_df,
                      Hess = TRUE)

reduced_model_summarized <- cbind(tidy(reduced_model),
                                  p.value = pnorm(abs(tidy(reduced_model)$statistic),
                                                  lower.tail = FALSE) * 2) |>
  mutate_if(is.numeric, round, 2) |>
```

```
  mutate(p.adjust = round(p.adjust(p.value, method="fdr"), 3))

reduced_model_summarized
```

```
##                                                  term estimate std.error
## 1                                mds_self_rated_enjoy2-1     3.79      1.22
## 2                                mds_self_rated_enjoy3-2    -2.32      0.87
## 3                                mds_self_rated_enjoy4-3    -0.39      0.47
## 4     salary_pre_mds$60,000 to $80,000-Less than $60,000     1.35      0.53
## 5    salary_pre_mds$80,001 to $100,000-$60,000 to $80,000     2.07      0.83
## 6  salary_pre_mds$100,001 to $120,000-$80,001 to $100,000    -0.13      0.97
## 7   salary_pre_mdsMore than $120,000-$100,001 to $120,000     4.33      1.46
## 8                    work_exp1 - 4 Years-Less than 1 Year     1.32      0.53
## 9                         work_exp4 - 7 Years-1 - 4 Years    -0.46      0.59
## 10                       work_exp7 - 10 Years-4 - 7 Years    -0.61      0.84
## 11                         work_exp10+ Years-7 - 10 Years    -0.59      1.07
## 12                   Less than $60,000|$60,000 to $80,000    -5.63      0.70
## 13                 $60,000 to $80,000|$80,001 to $100,000    -2.53      0.53
## 14               $80,001 to $100,000|$100,001 to $120,000    -0.17      0.47
## 15                $100,001 to $120,000|More than $120,000     2.40      0.60
##    statistic   coef.type p.value p.adjust
## 1       3.10 coefficient    0.00    0.000
## 2      -2.69 coefficient    0.01    0.017
## 3      -0.83 coefficient    0.41    0.575
## 4       2.53 coefficient    0.01    0.017
## 5       2.51 coefficient    0.01    0.017
## 6      -0.13 coefficient    0.89    0.890
## 7       2.97 coefficient    0.00    0.000
## 8       2.48 coefficient    0.01    0.017
## 9      -0.78 coefficient    0.44    0.575
## 10     -0.73 coefficient    0.46    0.575
## 11     -0.56 coefficient    0.58    0.669
## 12     -8.06       scale    0.00    0.000
## 13     -4.72       scale    0.00    0.000
## 14     -0.37       scale    0.71    0.761
## 15      3.98       scale    0.00    0.000
```

```
significant_results <- reduced_model_summarized |>
  filter(p.adjust < 0.05) |>
  filter(coef.type == 'coefficient')
significant_results
```

```
##                                                  term estimate std.error
## 1                                mds_self_rated_enjoy2-1     3.79      1.22
## 2                                mds_self_rated_enjoy3-2    -2.32      0.87
## 3     salary_pre_mds$60,000 to $80,000-Less than $60,000     1.35      0.53
## 4   salary_pre_mds$80,001 to $100,000-$60,000 to $80,000     2.07      0.83
## 5  salary_pre_mdsMore than $120,000-$100,001 to $120,000     4.33      1.46
## 6                    work_exp1 - 4 Years-Less than 1 Year     1.32      0.53
##   statistic   coef.type p.value p.adjust
## 1      3.10 coefficient    0.00    0.000
## 2     -2.69 coefficient    0.01    0.017
## 3      2.53 coefficient    0.01    0.017
## 4      2.51 coefficient    0.01    0.017
```

```
## 5      2.97 coefficient    0.00    0.000
## 6      2.48 coefficient    0.01    0.017
```

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:

- The adjusted p-values below indicate that the `full_model` and `reduced_model` are significantly better than `initial_model` as the p-values are less than 0.05 (alpha).
- On comparing `reduced_model` (simpler model) with `full_model` the p-value was found to be greater than 0.05 (alpha), hence we would choose `reduced_model` as we cannot reject the null hypothesis here.

```
anova_initial_full <- anova(initial_model, full_model)
anova_initial_reduced <- anova(initial_model, reduced_model)
anova_full_reduced <- anova(full_model, reduced_model)

p_vals <- c(anova_initial_full$"Pr(Chi)"[2], anova_initial_reduced$"Pr(Chi)"[2], anova_full_reduced$"Pr

adjusted_p_vals <- p.adjust(p_vals, method = "BH")
adjusted_p_vals
```

```
## [1] 7.017901e-09 4.418499e-11 5.245731e-01
```

```
names <- c("initial versus full", "initial versus reduced", "full versus reduced")
data.frame(comparison = names, adjusted_p_value = adjusted_p_vals) |>
  mutate_if(is.numeric, round, 9)
```

```
##               comparison adjusted_p_value
## 1    initial versus full      0.000000007
## 2 initial versus reduced      0.000000000
## 3    full versus reduced      0.524573057
```

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:

We have opted for the reduced model in our analysis. This model demonstrated a statistically significant correlation between the primary outcome, MDS enjoyment, and the expected salary after MDS completion. Significant associations were also observed between the expected post-MDS salary and two confounders: pre-MDS salary and work experience. This correlation is logical, as prior salary and work experience likely influence anticipated earnings.

We assume that our model includes all relevant confounders for our key relationship, allowing us to discuss potential causality between post-graduate salary expectations and self-rated enjoyment of the MDS program. In the full model, we accounted for all potential confounders and identified those that are significant. These significant confounders were then incorporated into our reduced model. We discovered that the reduced model outperformed the full model, leading us to conclude that omitted confounders do not obscure the relationship of interest. Further, our reduced model identified a significant association between pre-MDS salary and post-MDS salary expectations, particularly between the salary ranges of less than $60K and $60K-80K, as well as between $60K-80K and $80K-100K. Additionally, the influence of another confounder was notable: comparing less than 1 year of work experience to 1-4 years of work experience was significantly associated. We cannot assert causality for these confounders because our study wasn't designed to probe these specific links and additional confounders might exist between these variables. However, the significant association warrants further investigation.

Now, we'll address the question, "Does a person's self-rated enjoyment of the MDS program have any causal impact on their expected salary upon graduation?" Our analysis indicates that an increase in self-rated enjoyment of the MDS from 1 to 2 leads to a rise in expected post-graduation salary. However, an increase from 2 to 3 appears to result in a decrease in expected salary.

(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:

Type your answer here, replacing this text.

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.