Aryan Sharma

Final Exam

Statistical Programming

Part A: Short Answer Questions (each question is 10 points) 50 points.

1. Suppose you wish to measure the impact of smoking on the weight of newborns. You are planning to use the following model,

$$\log(bw_i) = \beta_0 + \beta_1 male_i + \beta_2 order_i + \beta_3 y_i + \beta_4 cig_i + \epsilon_i$$

where bw is the birth weight, male is a dummy variable assuming the value 1 if the baby is a boy or 0 otherwise, order is the birth order of the child, y is the log income of the family, cig is the amount of cigarettes per day smoked during pregnancy, i indexes the observation and the β's are the unknown parameters.

(a) What could be the problem in using OLS to estimate the above model?

The problem with using OLS is endogeneity—smoking may be correlated with unobserved factors like health or stress that also affect birth weight. This violates OLS assumptions and can bias the estimates.

(b) Suppose you have data on the average price of cigarettes in the state of residence. Would this information help to identify the true parameters of the model? How and Why?

Yes, the average price of cigarettes could serve as an instrumental variable. It affects cigarette smoking behavior but likely doesn't directly affect birth weight, which helps identify the causal effect of smoking through instrumental variable estimation.

2(a) What is the key difference between OLS regression and Quantile regression?

OLS estimates the mean effect, while quantile regression estimates the effect at different points in the distribution like median, 25th percentile, etc.

(b) What are advantages of using quantile regression model? Can one use quantile regression for making causal inference in the presence of endogeneity? Explain.

Quantile regression is helpful when the effect of predictors varies across the outcome distribution. It is robust to outliers and non-normality. However, to make causal inference with quantile regression in the presence of endogeneity, you'd still need valid instruments.

Aryan Sharma

3 (a) Write the logistic regression model and explain what an **odds ratio** means in logistic regression.

Logistic regression models the **log odds** of a binary outcome. The **odds ratio** tells us how the odds of the outcome change with a one-unit increase in a predictor.

(c) Explain what the coefficients in a logistic regression tell us (i) for a continuous predictor variable and (ii) for an indicator variable.
(i) For a continuous variable, the coefficient shows the change in log odds for a one-unit increase.
(ii) For a binary variable, the coefficient shows how being in one group changes the odds compared to the reference group.

4 (a) When do you use Principal Component Analysis? How is this method different from regression models?

PCA is used when we want to reduce dimensionality while retaining as much variance as possible. It's different from regression as PCA doesn't predict an outcome—it transforms the features into uncorrelated components.

(b) What do you understand by Linear Independence? Describe the core principle of Principal Component Analysis (PCA) in your own words. Provide one real-world example where one can apply PCA.

Linear independence means no variable can be written as a combination of others. PCA finds new axes (principal components) that are orthogonal and explain the most variance. Example: PCA can be used to compress image data while keeping key patterns like in facial recognition.

5 (a) What do you understand by model inference? Explain through an example.

Model inference is about drawing conclusions from the estimated model, like checking if a predictor is significant.
Example: In a regression of salary on education, if the coefficient of education is significant, we infer education affects income.

(b) What do you understand by Bayesian methods and how are they different from traditional or frequentist methods?

Bayesian methods update prior beliefs using data to get a posterior distribution. Unlike frequentist methods that rely only on the sample data, Bayesian approaches combine prior info and data for inference, and results are probabilistic for example., "there's a 95% chance β is in this range".

Aryan Sharma

PART – B (60 Points)

Objective:

The exam's objective is to evaluate your approach of estimating the causal effect of Minimum Legal Drinking age on mortality rates among young adults using the regression discontinuity design (RDD) approach.

The assignment should be done using R or Python or SAS or Stata or SPSS programming languages/software packages.

Include in your Word document the question, its result, and a clear, precise interpretation. Marks will be deducted for inadequate responses. Ensure that screenshots are clear, and interpretations are well-articulated.

Dataset Description:

The dataset used in this exam contains information on the mortality rates and causes of death for young adults aged 19 to 22. You can find it in canvas (Final Exam_dataset_MLDA.csv).

agecell: Age in years (with a decimal point, as ages are binned).

all: Mortality rates per 100,000 individuals for each age group.

internal: Mortality rates from internal causes, such as diseases or medical conditions.

external: Mortality rates from external causes, such as accidents, homicides, or suicides.

alcohol: Mortality rates directly linked to alcohol-related causes.

homicide: Mortality rates from homicides.

suicide: Mortality rates from suicides.

mva: Mortality rates from motor vehicle accidents (MVAs).

drugs: Mortality rates due to drug-related causes.

External other: Mortality rates from other external causes, like falls, burns, or drownings.

All the fitted values for all mortality rates are predicted by the regression model. These values help to smooth the data and show the overall trend.

The cutoff point for the study is age 21, which distinguishes between individuals with and without legal access to alcohol.

Aryan Sharma

Loading the data

```
# Load data
file_path <- ("C:/Users/91884/Desktop/BAIS/Advance data science/Final exam/Final Exam Dataset_MLDA.xlsx")
data <- read_excel(file_path, sheet = "Data")
summary(data)
str(data)
```

A. Data Cleaning and Manipulation: (5 points)

Columns like all, internal, external, alcohol, etc., have 2 missing values each (out of 50). Since the missingness is small (only 4% of the data) and spread across multiple columns, the simplest and safest approach is to drop those rows.

```
# Create treatment variable: 1 if age >= 21, else 0 and Data Cleaning
df_clean <- data %>%
  filter(!is.na(all) & !is.na(internal) & !is.na(external) &
           !is.na(alcohol) & !is.na(homicide) & !is.na(suicide) &
           !is.na(mva) & !is.na(drugs) & !is.na(externalother))

# Create treatment variable: 1 if age ≥ 21, else 0
df_clean <- df_clean %>%
  mutate(treatment = ifelse(agecell >= 21, 1, 0))
```

```
> str(df_clean)
tibble [50 × 20] (S3: tbl_df/tbl/data.frame)
 $ agecell             : num [1:50] 19.1 19.2 19.2 19.3 1
 $ all                 : chr [1:50] "92.825401310000004"
 $ allfitted           : num [1:50] 91.7 91.9 92 92.2 92.
 $ internal            : chr [1:50] "16.617589949999999"
 $ internalfitted      : num [1:50] 16.7 16.9 17.1 17.3 1
 $ external            : chr [1:50] "76.207817079999998"
 $ externalfitted      : num [1:50] 75 75 75 74.9 74.9 ..
 $ alcohol             : chr [1:50] "0.63913804299999999'
 $ alcoholfitted       : num [1:50] 0.794 0.838 0.878 0.
 $ homicide            : chr [1:50] "16.31681824" "16.85
 $ homicidefitted      : num [1:50] 16.3 16.3 16.3 16.3 1
 $ suicide             : chr [1:50] "11.20371437" "12.19
 $ suicidefitted       : num [1:50] 11.6 11.6 11.6 11.6 1
 $ mva                 : chr [1:50] "35.829326629999997"
 $ mvafitted           : num [1:50] 34.8 34.6 34.4 34.3 1
 $ drugs               : chr [1:50] "3.8724246029999998"
 $ drugsfitted         : num [1:50] 3.45 3.47 3.49 3.51 1
 $ externalother       : chr [1:50] "8.5343732830000008"
 $ externalotherfitted: num [1:50] 8.39 8.53 8.66 8.79 8
 $ treatment           : num [1:50] 0 0 0 0 0 0 0 0 0 0 .
```

B. Exploratory Data Analysis (EDA): (5 Points)

```
#3. Summary Statistics
str(df_clean)
summary(select(df_clean, all, internal, external, alcohol,
               homicide, suicide, mva, drugs, externalother))
```

```
> summary(select(df_clean, all, internal, external, alcohol,
+                homicide, suicide, mva, drugs, externalother))
     all              internal           external           alcohol
 Length:50          Length:50          Length:50          Length:50
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
   homicide           suicide              mva               drugs
 Length:50          Length:50          Length:50          Length:50
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character
 externalother
 Length:50
 Class :character
 Mode  :character
```

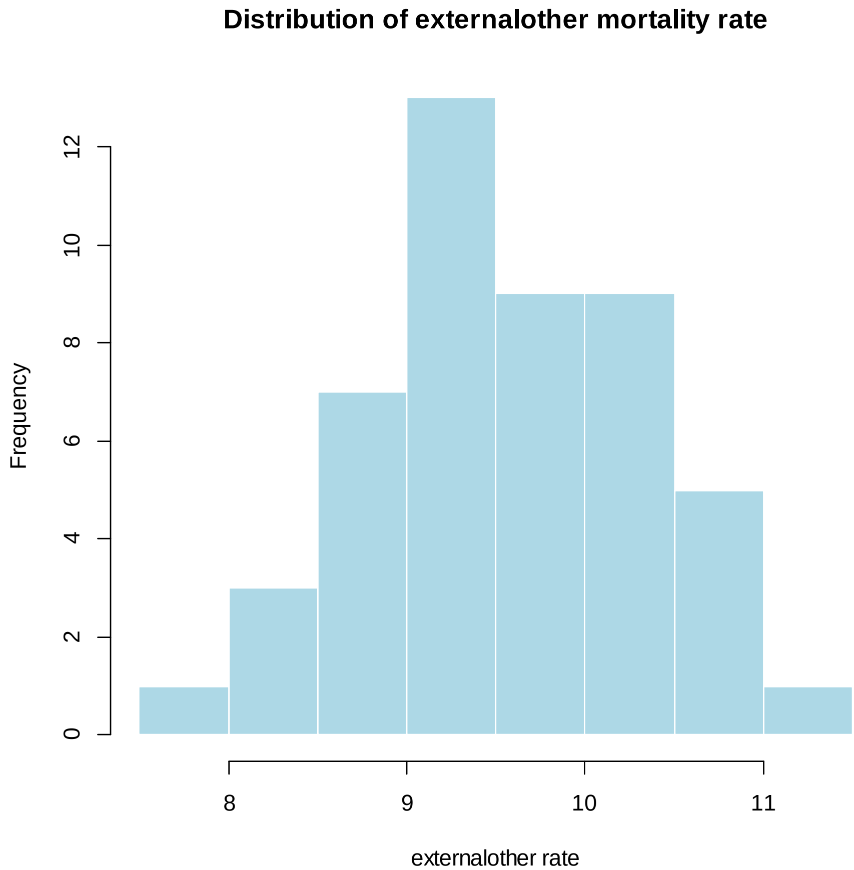
The columns are read as character variables instead of numeric variables hence we would have to convert them into numeric variables for our analysis.

```
df_clean$all <- as.numeric(df_clean$all)
df_clean$internal <- as.numeric(df_clean$internal)
df_clean$external <- as.numeric(df_clean$external)
df_clean$alcohol <- as.numeric(df_clean$alcohol)
df_clean$homicide <- as.numeric(df_clean$homicide)
df_clean$suicide <- as.numeric(df_clean$suicide)
df_clean$mva <- as.numeric(df_clean$mva)
df_clean$drugs <- as.numeric(df_clean$drugs)
df_clean$externalother <- as.numeric(df_clean$externalother)
str(df_clean)
```

- Explore the distribution of each category of mortality rates.

```
# Histogram for each mortality category
for (var in mortality_vars) {
  hist(df_clean[[var]], main = paste("Distribution of", var, "mortality rate"),
       xlab = paste(var, "rate"), col = "lightblue", border = "white")
}
```

Aryan Sharma



Highest mean mortality is from external causes mean ≈ 75, with substantial variability ,range from ~71 to 83. These include alcohol, homicide, suicide, and MVA.
Internal causes have the lowest variance and relatively stable distribution.
Alcohol-related deaths have a small mean =1.26 but increase sharply around the legal drinking age.
MVAs show a clear mid-to-high variance, suggesting potential age-related shifts.
Suicide and drugs show moderate means with consistent patterns, although not as clearly tied to age as alcohol or MVA.

- What do the summary statistics reveal about the distribution of different categories of mortality rates?
  The summary statistics reveal that mortality rates from external causes are more common and variable, suggesting strong behavioral influence. In contrast, internal causes are stable and largely unaffected by policy changes. The skew and spikes in alcohol-related mortality near age 21 justify the use of regression discontinuity design (RDD) to estimate the causal impact of the legal drinking age on youth mortality.

Aryan Sharma

C. Visualizations: (10 Points)

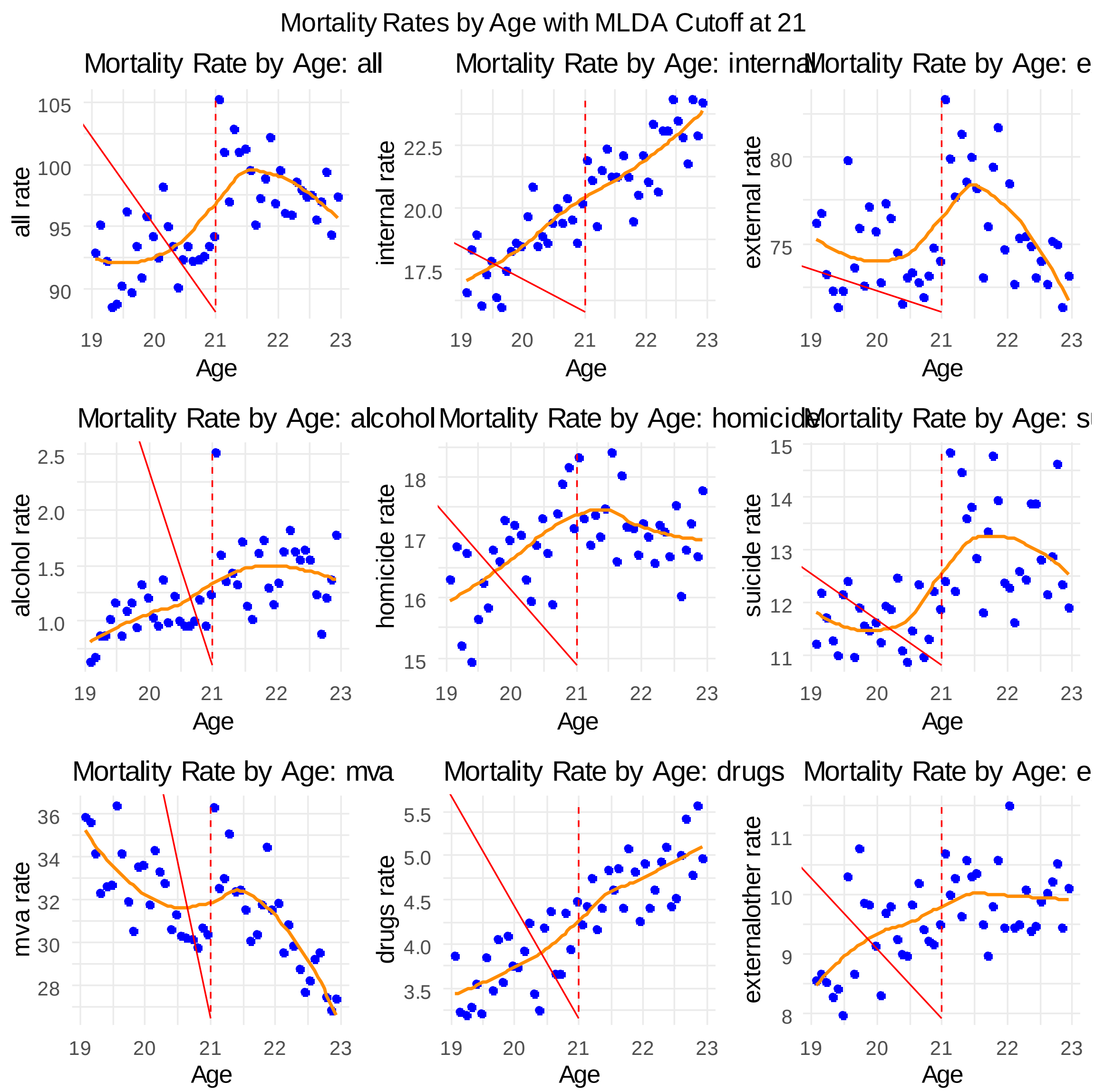
- Create scatter plots or line graphs to visualize the age profile for different types of mortality rates.

```r
#C. visualisation

plots <- list()
for (var in mortality_vars) {
  p <- ggplot(df_clean, aes(x = agecell, y = .data[[var]])) +
    geom_point(color = "blue") +
    geom_smooth(method = "loess", se = FALSE, color = "darkorange") +
    geom_vline(xintercept = 21, linetype = "dashed", color = "red") +
    labs(title = paste("Mortality Rate by Age:", var),
         x = "Age", y = paste(var, "rate")) +
    theme_minimal()

  plots[[var]] <- p
}

#grid layout 3x3
grid.arrange(grobs = plots, ncol = 3, top = "Mortality Rates by Age with MLDA Cutoff at 21")
```



- **Question**: What trends do you observe in visualizations? How do these trends support or contradict the expected impact of the legal drinking age?

Category	Trend Observed	Interpretation
All Mortality	Small jump around age 21	Slight increase in overall risk, possibly due to alcohol-sensitive causes
Internal Causes	Smooth upward trend, no jump	No impact from MLDA; confirms stability of non-behavioral causes
External Causes	Noticeable upward spike at 21	Strong evidence of MLDA impact—external causes reflect risky behaviors
Alcohol	Clear increase post-21	Direct evidence of alcohol access increasing mortality
Homicide	Mild increase after 21	Possible secondary effect of alcohol on violence
Suicide	Some rise post-21, but noisy	May imply alcohol's mental health effects, though variable
MVA (Motor Vehicle Accidents)	Significant drop before 21, then rise	Alcohol-related driving risks surge after legal access
Drugs	Gradual rise with age	Likely unrelated to MLDA; more age-driven
External Other	Stable to mild increase	No strong linkage to alcohol access

D. Linear Regression Discontinuity (RD) Model: (40 Points)

Implement a linear RD model to analyze the causal effect of legal access to alcohol on death rates. Include the following steps:

- Create a binary variable treatment that indicates whether an individual is 21 years or older.

```
# Create treatment variable: 1 if age ≥ 21, else 0
df_clean <- df_clean %>%
  mutate(treatment = ifelse(agecell >= 21, 1, 0))
```

project.R Final_exam.R* df_clean d_lag d

Filter

ohol	alcoholfitted	homicide	homicidefitted	suicide	suicidefitted	mva	mvafitted	drugs	drugsfitted	externalother	externalotherfitted	treatment
6391380	0.7943445	16.31682	16.28457	11.20371	11.59210	35.82933	34.81778	3.872425	3.448835	8.534373	8.388236	0
6774093	0.8375749	16.85996	16.27070	12.19337	11.59361	35.63926	34.63389	3.236511	3.470022	8.655786	8.530174	0
8664426	0.8778347	15.21925	16.26288	11.71581	11.59513	34.20565	34.44674	3.202071	3.492069	8.513741	8.662681	0
8673084	0.9151149	16.74282	16.26115	11.27501	11.59665	32.27896	34.25630	3.280689	3.514980	8.258285	8.785728	0
0191631	0.9494066	14.94773	16.26551	10.98431	11.59819	32.65097	34.06259	3.548198	3.538755	8.417533	8.899288	0
1713219	0.9807007	15.64281	16.27599	12.16663	11.59973	32.72144	33.86558	3.211689	3.563399	7.972546	9.003332	0
8699163	1.0089884	16.26365	16.29260	12.40576	11.60128	36.38520	33.66527	3.857890	3.588913	10.287705	9.097831	0
0979514	1.0342605	15.82565	16.31537	10.97951	11.60284	34.18793	33.46165	3.483156	3.615300	8.670031	9.182756	0
1748509	1.0565081	16.78900	16.34431	11.90010	11.60441	31.91047	33.25470	4.055130	3.642563	10.763150	9.258080	0
9484129	1.0757217	16.61619	16.37944	11.57064	11.60598	30.57683	33.04441	3.566033	3.670704	9.863494	9.323772	0
3291142	1.0918926	17.27848	16.42077	11.46836	11.60756	33.53165	32.83079	4.101267	3.699727	9.835445	9.379805	0
2164142	1.1050112	16.95377	16.46834	11.63196	11.60915	33.60344	32.61381	3.763282	3.729633	9.123107	9.426147	0

- Conduct separate RD models for each mortality rate category.

```
#D. RDD Model
# Store RD model results
rd_results <- data.frame(
  Category = character(),
  Intercept = numeric(),
  Treatment = numeric(),
  Age = numeric(),
  Interaction = numeric(),
  P_Treatment = numeric(),
  stringsAsFactors = FALSE
)

# Run RD model for each mortality category
for (var in mortality_vars) {
  formula_str <- as.formula(paste(var, "~ treatment + agecell + treatment:agecell"))
  model <- lm(formula_str, data = df_clean)
  summary_model <- summary(model)

  rd_results <- rbind(rd_results, data.frame(
    Category = var,
    Intercept = round(summary_model$coefficients[1, 1], 3),
    Treatment = round(summary_model$coefficients[2, 1], 3),
    Age = round(summary_model$coefficients[3, 1], 3),
    Interaction = round(summary_model$coefficients[4, 1], 3),
    P_Treatment = round(summary_model$coefficients[2, 4], 3)
  ))
}
print(rd_results)
```

- For each model, describe the coefficients, their statistical significance, and the effect of the legal drinking age.

```
> print(rd_results)
       Category Intercept Treatment    Age Interaction P_Treatment
1           all    76.251    83.333  0.827      -3.603       0.001
2      internal   -13.876     1.155  1.618      -0.036       0.918
3      external    90.127    82.179 -0.791      -3.567       0.001
4       alcohol    -1.811     6.237  0.142      -0.276       0.031
5      homicide     0.737    24.160  0.795      -1.145       0.000
6       suicide    11.059    10.619  0.029      -0.420       0.212
7           mva    83.849    28.945 -2.568      -1.162       0.043
8         drugs    -4.079     0.801  0.392      -0.028       0.814
9 externalother    -0.547    14.029  0.488      -0.647       0.041
```

- **Question**: What conclusions can be drawn about the causal impact of legal age on different types of mortality rates?

  The RD model results show that the legal drinking age of 21 has a significant impact on certain mortality rates. Specifically, alcohol-related deaths, external causes, motor vehicle accidents, and homicides all show statistically significant increases right after age 21. This suggests that gaining legal access to alcohol leads to higher risks in these areas. On the other hand, internal causes and drug-related deaths do not show significant changes, which supports the idea that the increase in risky behaviors is linked to alcohol access and not just age. Overall, the results support the effectiveness of the MLDA policy in delaying alcohol-related harm. For example, alcohol-related mortality increases by **6.24 deaths per 100,000** at age 21, which is a **significant jump of over 500%** compared to the baseline intercept –1.81. Similarly, motor vehicle accident deaths rise by **28.95**, which is a **34.5%**

increase over the baseline of 83.85. These large discontinuities at the cutoff highlight the sharp behavioral shift due to legal access. Looking at the age coefficients, a one-year increase in age is associated with a **positive slope** in alcohol deaths **+0.14 per year** and external causes **–0.79**, suggesting a natural decline without the MLDA jump. These trends show that while age alone has a gradual effect, the policy-driven shift at 21 causes an abrupt and notable increase in mortality.