Question 1

Part 1

To explore the effect of marijuana usage on wages with a log-level regression model, we need to consider how changes in the frequency of usage relate to wage changes. The equation for the model is as follows:

$$\log(\text{wage}) = \beta_0 + \beta_1 \times \text{marijuana\_usage} + \beta_2 \times \text{education} + \beta_3 \times \text{experience} + \beta_4 \times \text{gender} + u$$

Variable Definitions:

- log(wage): Natural logarithm of the wage, the dependent variable.
- marijuana_usage: Number of times marijuana was used in the last month.
- education and experience: Control variables representing the individual's educational attainment and work experience.
- gender: Dummy variable where 1 represents female and 0 male.
- u: Error term capturing unobserved factors affecting wages.

Coefficient Interpretation:

- β_0: The intercept term, representing the expected value of log(wage) when all independent variables are zero. This is the baseline wage for males (gender=0) with no marijuana usage, no education, and no experience.
- β_1: The estimated percentage change in wages associated with a one-unit increase in marijuana usage, holding all other variables constant. For instance, if $\beta_1 = 0.05$, then a one-time increase in marijuana use is associated with a 5% increase in wages.
- β_2: The estimated percentage change in wages associated with a one-unit increase in education, holding all other variables constant. This coefficient indicates the return on education in terms of wage increases.
- β_3: The estimated percentage change in wages associated with a one-unit increase in experience, holding all other variables constant. This coefficient indicates the return on work experience in terms of wage increases.
- β_4: The differential effect of being female on log(wage), holding all other variables constant. This coefficient indicates the gender wage gap, with positive values suggesting higher wages for females and negative values suggesting lower wages compared to males.

Example Calculation: If marijuana usage increases by five times, the increase in wages can be estimated by multiplying β_1 by five. For example, if $\beta_1 = 0.05$, then a five-time increase in marijuana usage is associated with a $5 \times 5\% = 25\%$ increase in wages. This calculation stems from the additive property of percentage changes in the logarithmic transformation, where an increase in usage by five times cumulatively leads to a wage change of $500 \times \beta_1\%$.

This analysis provides a quantitative framework for assessing the impact of varying levels of marijuana usage on wages, considering logarithmic changes directly correspond to percentage variations in the original wage scale.

Part 2

To assess whether the impact of marijuana usage on wages differs between men and women, we modify the previous log-level regression model to include interaction terms between marijuana usage and gender. This enhanced model setup allows for evaluating differential impacts across gender, with men as the reference group.

Model Specification:

$$\begin{aligned}\log(\text{wage}) = \beta_0 &+ \beta_1 \times \text{marijuana\_usage} + \beta_2 \times \text{gender} \\ &+ \beta_3 \times (\text{marijuana\_usage} \times \text{gender}) + \beta_4 \times \text{education} \\ &+ \beta_5 \times \text{experience} + u\end{aligned}$$

Variable Definitions:

- log(wage): Natural logarithm of the wage, serving as the dependent variable.
- marijuana_usage: Number of times marijuana was used in the last month.
- gender: Dummy variable coded as 1 for females and 0 for males, establishing males as the reference category.
- marijuana_usage × gender: Interaction term that assesses how the impact of marijuana usage differs between genders.
- education and experience: Control variables.
- u: Error term capturing unobserved factors affecting wages.

Interpretation of Coefficients:

- β_0: The intercept term, representing the expected value of log(wage) when all independent variables are zero. This is the baseline wage for males (gender=0) with no marijuana usage, no education, and no experience.
- β_1: The estimated change in log(wage) for males (gender=0) associated with a one-unit increase in marijuana usage, holding all other variables constant.
- β_2: The differential effect of being female on log(wage) when marijuana usage is zero, holding all other variables constant. This coefficient is key for this analysis as it indicates whether there is a gender wage gap independent of marijuana usage.
- β_3: The additional change in log(wage) for females relative to males due to marijuana usage, holding all other variables constant. This coefficient is crucial for determining if the effect of marijuana usage on wages is different for females compared to males.

- β_4: The estimated change in log(wage) associated with a one-unit increase in education, holding all other variables constant.
- β_5: The estimated change in log(wage) associated with a one-unit increase in experience, holding all other variables constant.

Statistical Testing for Gender Differences: We test both the direct effect of gender and the interaction effect to see if they are statistically significant.

Hypothesis Tests:

- Null Hypothesis (H_0): $\beta_2 = 0$ and $\beta_3 = 0$ (No differences in the effects of gender and its interaction with marijuana usage on wages).
- Alternative Hypothesis (H_1): Either $\beta_2 \neq 0$ or $\beta_3 \neq 0$ (Differences exist in the effects of gender or its interaction with marijuana usage on wages).

Restricted and Unrestricted Models:

- **Restricted Model:**

$$\begin{aligned}\log(\text{wage}) = \beta_0 &+ \beta_1 \times \text{marijuana\_usage} + \beta_4 \times \text{education} \\ &+ \beta_5 \times \text{experience} + u\end{aligned}$$

- **Unrestricted Model:**

$$\begin{aligned}\log(\text{wage}) = \beta_0 &+ \beta_1 \times \text{marijuana\_usage} + \beta_2 \times \text{gender} \\ &+ \beta_3 \times (\text{marijuana\_usage} \times \text{gender}) + \beta_4 \times \text{education} \\ &+ \beta_5 \times \text{experience} + u\end{aligned}$$

Calculation of the F-statistic:

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)}$$

where:

- RSS_R: Residual Sum of Squares for the Restricted Model
- RSS_{UR}: Residual Sum of Squares for the Unrestricted Model
- q: Number of restrictions (in this case, 2)
- n: Total sample size
- k: Number of parameters in the unrestricted model (excluding the intercept)

Calculation of the F-test in R:

```
# Assuming a linear model 'model' has been fitted with lm()
model_unrestricted <- lm(log(wage) ~ marijuana_usage + gender +
    marijuana_usage*gender + education + experience, data=data)
model_restricted <- lm(log(wage) ~ marijuana_usage + education +
    experience, data=data)

# Extract RSS
RSS_UR <- sum(resid(model_unrestricted)^2)
RSS_R <- sum(resid(model_restricted)^2)

# Calculate F-statistic manually
q <- 2
n <- nrow(data)
k <- length(coef(model_unrestricted)) - 1
F_statistic <- ((RSS_R - RSS_UR) / q) / (RSS_UR / (n - k - 1))

# Alternatively, using an inbuilt function
linear_hypothesis(model_unrestricted, c("gender = 0", "marijuana_usage*
    gender = 0"), test = "F")
```

Interpreting the F-statistic: The F-statistic tests whether the explained variance in a model with more parameters is significantly greater compared to a model with fewer parameters, under the null hypothesis that the additional parameters are zero. A significant F-statistic (where p-value is less than or equal to 0.05) indicates that the additional parameters significantly improve model fit.

Decision Rule:

- If the F-statistic's corresponding p-value is less than or equal to 0.05, reject the null hypothesis H_0.
- To further assess the strength of the test, compare the F-statistic to a critical value from the F-distribution table at a given significance level (e.g., 0.05) with appropriate degrees of freedom (df1 = number of parameters tested, df2 = total sample size minus number of parameters in the model - 1).
- A higher F-statistic relative to the critical value strongly supports the rejection of the null hypothesis, indicating that the effects of gender and its interaction with marijuana usage on wages are statistically significant.

Calculation of Critical Value in R:

```
# Assuming alpha = 0.05 and appropriate degrees of freedom
critical_value <- qf(0.95, df1 = q, df2 = n - k - 1)
```

Alternative Method: P-value Calculation:

```
# Calculate p-value for the F-statistic
p_value <- 1 - pf(F_statistic, df1 = q, df2 = n - k - 1)
```

If the p-value is less than or equal to 0.05, we reject the null hypothesis H_0, concluding that there are significant gender differences in the effect of marijuana usage on wages.

This detailed approach ensures a comprehensive understanding of whether policy interventions or program designs need to consider gender-specific impacts of marijuana usage on economic outcomes.

Part 3

To estimate the effects of marijuana usage on wages, categorized by usage levels, we employ a log-level regression model. In this model, marijuana usage is segmented into categorical levels defined as light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Nonusers serve as the reference category, which means the coefficients of the user categories measure the difference in log wages relative to nonusers.

Model Specification: The regression equation is structured to incorporate dummy variables for each marijuana usage category with nonusers as the baseline. This approach facilitates the estimation of differential wage impacts directly attributed to each usage level:

$$\begin{aligned}\log(\text{wage}) = \beta_0 & \\ & + \beta_1 \times \text{LightUser} \\ & + \beta_2 \times \text{ModerateUser} \\ & + \beta_3 \times \text{HeavyUser} \\ & + \beta_4 \times \text{education} \\ & + \beta_5 \times \text{experience} \\ & + \beta_6 \times \text{gender} \\ & + u\end{aligned}$$

Variable Definitions:

- log(wage): The natural logarithm of the wage, serving as the dependent variable in the model.
- LightUser, ModerateUser, HeavyUser: Dummy variables for each category of marijuana usage. Each variable takes a value of 1 if the individual falls into the respective category and 0 otherwise. The omitted category, nonusers, serves as the reference group against which the impacts of other categories are measured.
- education, experience, gender: These control variables are included to adjust for additional factors that may influence wage levels.
- u: The error term capturing all unobserved factors affecting wages.

Interpretation of Coefficients:

- β_0: The intercept term, representing the expected value of log(wage) when all independent variables are zero. This is the baseline wage for nonusers with no education, no experience, and male gender.
- β_1: The estimated percentage change in wages for light users compared to nonusers, holding education, experience, and gender constant. For instance, if $\beta_1 = 0.05$, this suggests that light users earn approximately $100 \times 0.05 = 5\%$ more than nonusers.
- β_2: The estimated percentage change in wages for moderate users compared to nonusers, holding education, experience, and gender constant. For instance, if $\beta_2 = -0.03$, this suggests that moderate users earn approximately $100 \times -0.03 = -3\%$ less than nonusers.

- β_3: The estimated percentage change in wages for heavy users compared to nonusers, holding education, experience, and gender constant. For instance, if $\beta_3 = 0.08$, this suggests that heavy users earn approximately $100 \times 0.08 = 8\%$ more than nonusers.

- β_4: The estimated percentage change in wages associated with a one-unit increase in education, holding marijuana usage, experience, and gender constant. This coefficient indicates the return on education in terms of wage increases.

- β_5: The estimated percentage change in wages associated with a one-unit increase in experience, holding marijuana usage, education, and gender constant. This coefficient indicates the return on work experience in terms of wage increases.

- β_6: The estimated percentage change in wages for females compared to males, holding marijuana usage, education, and experience constant. This coefficient indicates the gender wage gap.

Examples:

- β_1: If $\beta_1 = 0.05$, light users earn approximately 5% more than nonusers. For example, if a nonuser earns 1000, a light user would earn approximately 1050, assuming all other variables are held constant.

- β_2: If $\beta_2 = -0.03$, moderate users earn about 3% less than nonusers. For example, if a nonuser earns 1000, a moderate user would earn approximately 970, assuming all other variables are held constant.

- β_3: If $\beta_3 = 0.08$, heavy users earn about 8% more than nonusers. For example, if a nonuser earns 1000, a heavy user would earn approximately 1080, assuming all other variables are held constant.

Conclusion: By distinguishing between different levels of marijuana usage and using nonusers as the reference category, this analysis helps to elucidate the varied economic impacts of marijuana consumption on wages. Such insights are crucial for policymakers and industry stakeholders aiming to understand and address the broader economic implications of marijuana legalization and usage in the workforce.

Part 4

We will use the F-test to statistically evaluate the null hypothesis that the coefficients of the categorical marijuana usage variables (Light User, Moderate User, and Heavy User) are all zero. This test will tell us if these variables, when added to the model, provide any significant explanatory power in predicting wages.

Models for Comparison:

- **Unrestricted Model:** Includes all predictors, including the marijuana usage categories.

$$
\begin{aligned}
\log(\text{wage}) = \beta_0 & \\
&+ \beta_1 \times \text{LightUser} \\
&+ \beta_2 \times \text{ModerateUser} \\
&+ \beta_3 \times \text{HeavyUser} \\
&+ \beta_4 \times \text{education} \\
&+ \beta_5 \times \text{experience} \\
&+ \beta_6 \times \text{gender} \\
&+ u
\end{aligned}
$$

- **Restricted Model:** Excludes the marijuana usage categories, serving as the baseline model.

$$\log(\text{wage}) = \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{experience} + \beta_3 \times \text{gender} + u$$

Hypothesis Testing:

- Null Hypothesis (H_0): $\beta_1 = \beta_2 = \beta_3 = 0$ (No effect of marijuana usage categories on wages).
- Alternative Hypothesis (H_1): At least one $\beta_i \neq 0$ (At least one category of marijuana usage affects wages).

Calculation of the F-statistic: The F-statistic is calculated using the formula:

$$F = \frac{(\text{SSR}_r - \text{SSR}_{ur})/q}{\text{SSR}_{ur}/(n-k-1)}$$

where SSR_r is the sum of squared residuals from the restricted model, SSR_{ur} is from the unrestricted model, n is the number of observations, k is the number of predictors in the unrestricted model, and q is the number of restrictions (3 in this case).

Performing the F-test in R: We use R to fit both models and compute the F-statistic:

```r
# Fit the unrestricted model
model_ur <- lm(log(wage) ~ LightUser + ModerateUser + HeavyUser +
    education + experience + gender, data = dataset)

# Fit the restricted model
model_r <- lm(log(wage) ~ education + experience + gender, data =
    dataset)

# Extract the sum of squared residuals
ssr_ur <- sum(resid(model_ur)^2)
ssr_r <- sum(resid(model_r)^2)

# Number of parameters in unrestricted model and number of restrictions
k <- length(coef(model_ur)) - 1  # minus one for intercept
q <- 3  # number of usage categories

# Calculate F-statistic
F_value <- ((ssr_r - ssr_ur) / q) / (ssr_ur / (nrow(dataset) - k - 1))
p_value <- pf(F_value, q, nrow(dataset) - k - 1, lower.tail = FALSE)
```

Decision Rule:

- If the computed F-statistic exceeds the critical value from the F-distribution (or if the p-value is less than the chosen significance level, typically 0.05), reject the null hypothesis, indicating that the marijuana usage categories do have a significant effect on wages.
- If not, we do not reject the null hypothesis, suggesting that these variables may not be necessary for the wage prediction model.

This structured approach using an F-test provides a rigorous statistical method to assess the impact of marijuana usage categories on wage prediction, ensuring model adequacy and the relevance of predictors included in the regression analysis.

Part 5

Drawing causal inference from survey data often involves significant challenges, especially when the data is observational rather than experimental. Here, we discuss some potential problems with establishing causality from the survey data collected on marijuana usage and wage.

1. Confounding Variables: One of the primary issues in drawing causal inferences is the presence of confounding variables that might affect both the independent variable (marijuana usage) and the dependent variable (wage). Factors such as mental health, personal motivation, and access to education might influence both an individual's propensity to use marijuana and their earning potential. If these confounders are not adequately controlled for, the estimated effect of marijuana usage on wages may be biased. Suppose an individual with higher anxiety and stress levels (a confounding variable) are more likely to use marijuana to cope and also tend to have lower wages due to the impact of stress on job performance. Hence, if stress levels are not controlled for, the analysis might falsely attribute the lower wages to marijuana usage rather than the underlying stress.

Model Specification:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + u$$

Where:

- Y: Wages
- X: Marijuana usage
- Z: Confounding variables (e.g., mental health, education)
- u: Error term

2. Reverse Causality: Another issue is the possibility of reverse causality, where higher wages could potentially lead to increased disposable income, thus increasing marijuana usage. This reverse causality can lead to erroneous conclusions about the direction of the effect between the variables. Notably, individuals with higher wages might have more disposable income to spend on recreational activities and socials that may include the usage of marijuana. This then makes it challenging to determine whether its usage affects wages or higher wages lead to more usages.

3. Measurement Errors: Survey data on sensitive topics such as drug use are prone to measurement errors, as respondents might underreport or overreport their usage due to stigma or social desirability bias. Such inaccuracies in the independent variable can lead to attenuation bias, where the estimated effect of marijuana on wages is systematically biased towards zero. Indeed, if respondents underreport their marijuana usage due to fear or judgment or legal repercussions, the analysis might show a weaker relationship between marijuana usage and wages than there truly is.

4. Omitted Variable Bias: Omitted variable bias occurs when a model leaves out one or more relevant variables that influence both the independent and dependent variables. For example, personality traits such as risk-taking or innovation might lead to both higher marijuana usage and higher wages but are often difficult to measure and hence omitted. An individual's risk-taking behavior could lead them to both engage in marijuana due to curiosity and excitement and at the same time pursue high-risk and high-reward job opportunities. That said, if this type of behavior and personality trait is not accounted for, the model might then incorrectly attribute higher wages solely to marijuana usage.

5. Selection Bias: Selection bias might occur if the sample is not representative of the general population. For instance, if the survey disproportionately samples from particular demographics or economic sectors more likely to use marijuana, the findings may not generalize to the broader population. For example, if the survey is mainly conducted among younger individuals from urban areas where the use of marijuana is more socially accepted and common, the result may not be applicable to older adults or people in rural areas.

6. Non-random Assignment: Unlike experimental designs, observational studies like surveys do not involve random assignment to treatment and control groups. This non-random assignment means that any observed differences in wages could be due to pre-existing differences among respondents rather than marijuana usage itself. Individuals who choose to use marijuana might differ systematically from those who do not in ways that also affect their wages, such as their overall health, work ethic, or social environment. Without random assignment, these pre-existing differences can confound the relationship between marijuana usage and wages.

Question 2

Part 1

To examine the impact of law school rank on the median starting salary of new law school graduates, we fit a log-linear regression model with the following specification:

$$\log(\text{salary}) = \beta_0 + \beta_1\text{LSAT} + \beta_2\text{GPA} + \beta_3\log(\text{libvol}) + \beta_4\log(\text{cost}) + \beta_5\text{rank} + u$$

Variable Definitions:

- LSAT: Median LSAT score of the graduating class.
- GPA: Median college GPA of the graduating class.
- log(libvol): Logarithm of the total volumes in the law school library, indicating library size.

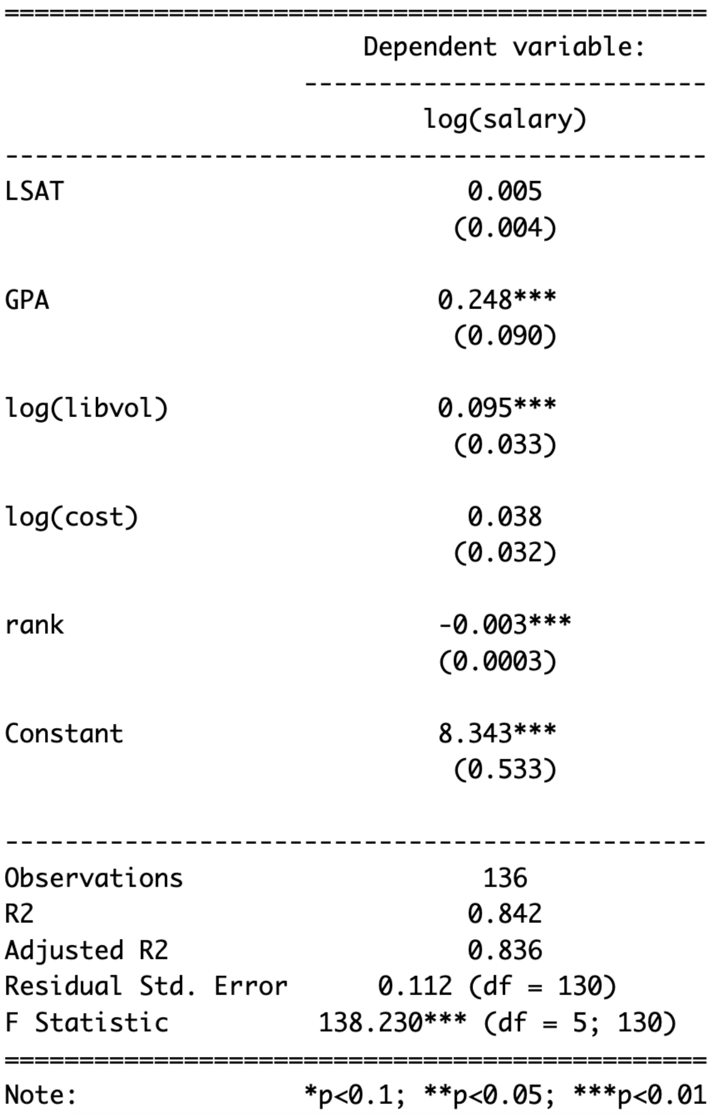
- log(cost): Logarithm of the annual cost of attending law school, reflecting the financial investment.

- rank: Ranking of the law school, with a lower number indicating a higher prestige and presumed quality of education.

Step 1: Fit the Model To understand how the variables such as LSAT scores, GPA, library volumes, cost, and rank influence the log-transformed starting salaries of new law school graduates, we first prepare and analyze the data using R. The steps involve subsetting the necessary columns, handling missing data, fitting the regression model, and finally, presenting the model summary in a clean format using the stargazer package.

R Code:

```r
# Load necessary data columns
data <- data[c("salary", "LSAT", "GPA", "libvol", "cost", "rank")]

# Remove any observations with missing values to avoid errors in the analysis
data.new <- na.omit(data)

# Fit the regression model
model <- lm(log(salary) ~ LSAT + GPA + log(libvol) + log(cost) + rank, data = data.new)

# Install and load stargazer for beautiful table outputs
if (!require(stargazer)) {
    install.packages("stargazer", dependencies = TRUE)
    library(stargazer)
}

# Output the regression table in text format for use in documents
stargazer(model, type = "text")

# Alternatively, view a detailed summary of the model
summary(model)
```



```
===============================================
                        Dependent variable:
                    ---------------------------
                            log(salary)
-----------------------------------------------
LSAT                           0.005
                              (0.004)

GPA                          0.248***
                              (0.090)

log(libvol)                  0.095***
                              (0.033)

log(cost)                      0.038
                              (0.032)

rank                         -0.003***
                             (0.0003)

Constant                     8.343***
                              (0.533)

-----------------------------------------------
Observations                    136
R2                             0.842
Adjusted R2                    0.836
Residual Std. Error      0.112 (df = 130)
F Statistic          138.230*** (df = 5; 130)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Figure 1: Regression Output

Step 2: Hypothesis Testing The objective of the hypothesis test is to determine whether the rank of law schools, when considered among other factors like LSAT scores, GPA, library volume, and cost of education, has a significant causal impact on the median starting salary of law school graduates. This is formally tested by examining the statistical significance of the coefficient associated with the rank variable in the regression model.

Formulating the Hypotheses:

- **Null Hypothesis** (H_0): The null hypothesis posits that the coefficient for the rank variable (β_5) is zero. Mathematically, this is represented as:

$$H_0 : \beta_5 = 0$$

The null hypothesis implies that the rank of a law school does not have a causal effect on the median starting salary of its graduates, after controlling for other factors such as LSAT scores, GPA, library volumes, and education costs. In other words, under the null hypothesis, the prestige or ranking of the school is assumed to have no influence on the earnings potential of the graduates. For example[2], the starting salaries of graduate from Stanford Law School, ranked 1st, and those from

Florida International University, ranked 68th, should be similar when controlling for the other factors mentioned above.

- **Alternative Hypothesis** (H_a): The alternative hypothesis contends that the rank coefficient is not zero:

  $$H_a : \beta_5 \neq 0$$

  Acceptance of the alternative hypothesis would suggest that the law school's rank does significantly affect the median starting salary, either positively or negatively, indicating that the rank is a predictor of salary outcome. This can be interpreted as the higher-ranked (i.e., numerically lower) law schools potentially offering better career prospects, or vice versa, depending on the sign of β_5.

Testing the Hypothesis: To test these hypotheses, we calculate the t-statistic for β_5 using the estimated coefficient and its standard error from the regression output. The t-statistic will help us determine if the observed coefficient for the rank is statistically significantly different from zero.

R Code to Compute the Test Statistic:

```r
# Extracting the coefficient and standard error for 'rank'
rank_coef <- coef(summary(model))["rank", "Estimate"]
rank_se <- coef(summary(model))["rank", "Std. Error"]

# Calculating the t-statistic
t_statistic <- rank_coef / rank_se

# Display t-statistic
cat("The t-statistic for the rank coefficient is:", t_statistic, "\n")
```

Given the computed t-statistic of -9.540787, we can infer a strong influence of rank on starting salary, opposite to what might be assumed with lower ranks typically indicating better outcomes. In other words, the effect of law school rank on starting salary is statistically significant. For exanple, a graduate form a lower ranked school might earn significantly less than a graduate from a higher-ranked university.

Critical Value: The critical value for a 5% significance level from a two-tailed t-distribution is computed to assess the statistical significance of the t-statistic.

```r
# Critical value for a two-tailed test at alpha = 0.05
critical_value <- qt(0.975, df = nrow(lawsch_data) - length(coef(model)))

# Display critical value and p-value
cat("The critical value at 5% significance level is:", critical_value, "\n")
```

Decision Rule: With a t-statistic of -9.540787, which is greater in absolute value than the critical value of 1.975905, we reject the null hypothesis. By rejecting the null hypothesis, we have established that the rank of a law school has a statistically significant impact on the median starting salary of its graduates, after controlling for other relevant factors such as LSAT scores, GPA, library volumes, and education costs.

This rigorous application of statistical hypothesis testing validates the importance of the law school's ranking in influencing graduate outcomes in terms of salary. Indeed, according to U.S. News[2], The starting salaries of 2020 law school graduates showed a strong correlation with the rank of the law school they attended. For each of the top 15 schools in the 2023 Best Law Schools rankings, the median starting salary in the private sector was \$190,000. Although seven lower-ranked schools also reported a median private sector salary of \$190,000, no school reported a figure higher than \$190,000. Conversely, all of the 10 schools with the lowest private sector starting salaries were ranked in the bottom quarter of the law school rankings, listed within a ranking range. This data indicates that, overall, the lower the rank of a law school, the lower the median private sector salary of its graduates.

This difference in starting salaries based on rank will likely have significant long-term financial impacts for graduate as well such as lifetime earnings, savings and debt. Notably, graduates from higher-ranked schools will often have access to more prestigious law firms opportunities, high-profile cases and networks. This then leads to faster career advancement as well, making the tuition fees a worthwhile investment. That said, our finding can in a way help prospective law school applicants to make informed decisions, balancing potential financial outcomes with their career goals.

Part 2

In this section, we analyze the impact of individual and combined predictors, specifically LSAT scores and GPAs, on the log-transformed median starting salary of law school graduates. We evaluate the significance of these predictors both individually and jointly through hypothesis testing.

Individual Significance Testing

For each predictor, we perform t-tests to evaluate whether the coefficients of LSAT and GPA are statistically different from zero, which would suggest an effect on the log-transformed salary.

Hypotheses for LSAT:

- **Null Hypothesis** (H_0): $\beta_1 = 0$. This hypothesis states that LSAT scores do not influence the starting salary.
- **Alternative Hypothesis** (H_a): $\beta_1 \neq 0$. This suggests that LSAT scores have a significant impact on starting salary.

In other words, consider a law school graduate with a high LSAT score of 180. If the LSAT score significantly impacts starting salary, the latter with the high score should consistently earn more than those with lower scores. Hence, if our analysis supports the alternative hypothesis, it would mean that the effort putting to achieve a high LSAT score can translate into financial benefits after graduation.

Hypotheses for GPA:

- **Null Hypothesis** (H_0): $\beta_2 = 0$. This implies that GPA does not affect the starting salary.

- **Alternative Hypothesis** (H_a): $\beta_2 \neq 0$. This indicates that GPA significantly affects the starting salary.

Similarly, if our analysis supports the alternative hypothesis, a student with a higher GPA should earn more than a student with a lower GPA, assuming all other factors are equal. This would indicate that academic performance throughout the law school plays a crucial role in determining starting salary.

R Code to Compute t-statistics and Compare with Critical Values:

```r
# Extract coefficients and standard errors for LSAT and GPA
coef_summary <- summary(model)$coefficients
LSAT_coef <- coef_summary["LSAT", "Estimate"]
LSAT_se <- coef_summary["LSAT", "Std. Error"]
GPA_coef <- coef_summary["GPA", "Estimate"]
GPA_se <- coef_summary["GPA", "Std. Error"]

# Calculate t-statistics
LSAT_t <- LSAT_coef / LSAT_se
GPA_t <- GPA_coef / GPA_se

# Output t-statistics
cat("The t-statistic for LSAT is:", LSAT_t, "\n")
cat("The t-statistic for GPA is:", GPA_t, "\n")
```

Given the standard critical value for a two-tailed test at a 5% significance level is approximately ±1.975905, we find:

- The t-statistic for LSAT: 1.171045 does not exceed the critical value, indicating insufficient evidence to reject the null hypothesis, suggesting that LSAT scores may not independently influence the starting salary significantly.
- The t-statistic for GPA: 2.749133 exceeds the critical value, providing strong evidence against the null hypothesis and indicating that GPA has a significant positive effect on starting salary.

Joint Significance Testing (F-test)

To rigorously evaluate the combined impact of LSAT and GPA on starting salaries, we conduct an F-test. This statistical test compares two models: a restricted model that excludes LSAT and GPA, hypothesizing that these variables do not affect the outcome, and an unrestricted model that includes all predictors, hypothesizing their potential influence.

Model Specifications:

- **Unrestricted Model**: Includes all variables (LSAT, GPA, library volume, cost, and rank). This model assumes that each factor may play a role in predicting starting salaries.
- **Restricted Model**: Excludes LSAT and GPA, focusing only on library volume, cost, and rank. This model tests the hypothesis that excluding LSAT and GPA does not significantly reduce the explanatory power of the model.

Hypotheses for the F-test:

- **Null Hypothesis** (H_0): $\beta_1 = \beta_2 = 0$. This hypothesis claims that neither LSAT scores nor GPA significantly affects the salary when considered together with other variables.
- **Alternative Hypothesis** (H_a): At least one of β_1 or β_2 is not zero. This suggests that these academic metrics do have a significant effect on the salary.

R Code to Perform the F-test and Output Results:

```r
# Fit the unrestricted model (including all variables)
unrestricted_model <- lm(log(salary) ~ LSAT + GPA + log(libvol) + log(cost) + rank, data = data.new)

# Fit the restricted model (excluding LSAT and GPA)
restricted_model <- lm(log(salary) ~ log(libvol) + log(cost) + rank, data = data.new)

# Calculate SSR for both models
SSR.ur <- sum(unrestricted_model$residuals^2)
SSR.r <- sum(restricted_model$residuals^2)

# Number of observations and predictors
n <- nrow(data.new)
k <- length(coef(unrestricted_model))  # Number of predictors in the unrestricted model
p <- 2  # Number of parameters (LSAT and GPA) being tested

# Compute the F-statistic
F.statistic <- ((SSR.r - SSR.ur) / p) / (SSR.ur / (n - k))

# Determine the critical value and p-value
critical_value <- qf(0.95, df1 = p, df2 = n-k)
p_value <- pf(F.statistic, df1 = p, df2 = n-k, lower.tail = FALSE)

# Output the results
cat("F-statistic:", F.statistic, "\n")
cat("Critical value at 95% confidence:", critical_value, "\n")
cat("P-value:", p_value, "\n")
```

F-test Results and Conclusion: The calculated F-statistic is 9.95174, which greatly exceeds the critical value of 3.065839 at the 95% confidence level. Furthermore, the extremely low p-value of 0.0000518231 strongly supports rejecting the null hypothesis. This result confirms the joint significance of LSAT and GPA in influencing starting salaries, indicating that both metrics are crucial determinants of salary outcomes.

Hence, a law school student with both a high LSAT score and GPA is likely to receive a higher starting salary than one with only one of these metrics being strong. This could be due to the fact that while a high LSAT score indicates strong analytical and logical reasoning skills, a high GPA demonstrates consistent academic performance and a strong work ethic. Employers might view candidates with both high LSAT scores and GPA as more well-rounded and capable, thus offering them a higher starting salary.

Conclusion: The results from the individual t-tests combined with the F-test provide a comprehensive view of how LSAT scores and GPA jointly influence the starting salaries of law school graduates. The analysis shows that GPA has a significant individual effect, and when combined with LSAT scores, the two create a robust predictor of salary outcomes. This underlines the importance of academic performance in legal education as it relates to career success. Prospective law school students can use this finding to guide their preparation strategies to maximize their future earning potential. Ultimately, a high LSAT score can facilitate admission to a prestigious law school, which, as we have established prior, positively impacts starting salary. Additionally, maintaining a high GPA can attract competitive job offers upon graduation, further enhancing earning potential.

Part 3

In this section, we assess the necessity of incorporating two additional predictors, the size of the class (`clsize`) and the size of the faculty (`faculty`), into the regression model used to predict the log-transformed starting salary of law school graduates.

Model Specifications:

- **Unrestricted Model**: Includes all previously considered variables (LSAT, GPA, logarithm of library volumes, logarithm of cost, rank) as well as the size of the class and the size of the faculty. This model hypothesizes that these additional variables might have a significant effect on the starting salary.
- **Restricted Model**: Excludes the `clsize` and `faculty` variables, maintaining only the original set of predictors. This model tests the hypothesis that the excluded variables do not contribute significantly to the predictive power regarding starting salaries.

Hypotheses for Joint Significance Test:

- **Null Hypothesis** (H_0): The coefficients for both `clsize` and `faculty` are zero ($\beta_{\text{clsize}} = \beta_{\text{faculty}} = 0$), implying that these variables do not significantly impact the salary.
- **Alternative Hypothesis** (H_a): At least one of the coefficients (β_{clsize} or β_{faculty}) is non-zero, suggesting that these variables do contribute to explaining variations in the salary.

Statistical Analysis with R: To test these hypotheses, we conduct an F-test to compare the fit of the unrestricted and restricted models. The following R code outlines the steps taken to perform this analysis:

```
# Load necessary packages and data
load("/Users/grovercastroduenas/Desktop/Stats & Econometrics/PS3/lawsch85.RData")
data <- data[c("salary", "LSAT", "GPA", "libvol", "cost", "rank", "clsize", "faculty")]
data.new <- na.omit(data)  # Remove missing data

# Fit the unrestricted model including clsize and faculty
unrestricted_model <- lm(log(salary) ~ LSAT + GPA + log(libvol) + log(cost) + rank + clsize + faculty, data = data.new)

# Fit the restricted model excluding clsize and faculty
```

```r
restricted_model <- lm(log(salary) ~ LSAT + GPA + log(libvol) + log(
    cost) + rank, data = data.new)

# Calculate the sum of squared residuals for both models
SSR.ur <- sum(unrestricted_model$residuals^2)
SSR.r <- sum(restricted_model$residuals^2)

# Calculate the F-statistic
n <- nrow(data.new)  # Number of observations
p <- 2  # Number of additional parameters tested
k.ur <- length(coef(unrestricted_model))  # Number of predictors in the
     unrestricted model
F.statistic <- ((SSR.r - SSR.ur) / p) / (SSR.ur / (n - k.ur))

# Calculate critical value and p-value for the F-test
critical_value <- qf(0.95, df1 = p, df2 = n - k.ur)
p_value <- pf(F.statistic, df1 = p, df2 = n - k.ur, lower.tail = FALSE)
```

Results and Conclusion: The F-statistic obtained from the test is 0.9483693, with a critical value of 3.069894 and a p-value of 0.3901861. Since the F-statistic is lower than the critical value and the p-value exceeds the conventional threshold of 0.05, we fail to reject the null hypothesis. This indicates that the addition of `clsize` and `faculty` to the model does not significantly improve the model's ability to predict log-transformed starting salaries.

For example, in a scenario where a law school has a large faculty and a small class size. Our analysis suggests that these factors do not significantly influence starting salary. Even if a law school has a favorable student-to-faculty ratio, this does not necessarily translate into better career outcomes if the school's overall prestige, quality of education, network opportunities, and student performance are not taken into account. Therefore, the size of the class and faculty might not have a direct impact on starting salaries.

Conclusion: The analysis suggests that, within the context of this data and model specification, neither the size of the class nor the size of the faculty significantly affects the starting salary of law school graduates. This result helps to streamline the model by indicating that these factors can be excluded, thus simplifying the predictive analysis without losing significant explanatory power.

Question 3

We know that two variables can be individually insignificant while also being jointly significant, however, is it possible that a variable is individually significant but when tested with other variables, they are jointly insignificant? We find that it is indeed possible to arrive at both these outcomes, because we use different testing methods to test the significance of a variable (t-test) or set of variables (f-test). In the case where a variable is individually significant, but jointly insignificant with other variables, we can conclude that in the f-test the significance of the single variable's coefficient is being hidden by the coefficient of the other variables as a set, which is why it's important for peer-reviewed studies to take this into account when designing regressions.

Key Concepts of Variable Significance

Individual Significance: This term refers to the statistical significance of each predictor's coefficient, determined via a t-test, which checks if a predictor exerts a statistically significant influence on the dependent variable when controlling for other factors. The t-test is best suited for testing a single hypothesis because it considers one-sided alternatives on the distribution curve, whereas a f-test needs two-sided alternatives to reach the same result.

T-statistic Formula:

$$t = \frac{(\hat{\beta 1} - \hat{\beta 2})}{se(\hat{\beta 1} - \hat{\beta 2})}$$

In this equation, we are testing whether the estimator difference divided by the sampling error is significantly different than zero to reject the null hypothesis.

Joint Significance: This concept involves an F-test to assess if a set of variables, when considered together, significantly enhances the model's explanatory power compared to a model without these predictors. The null hypothesis in this case posits that all such coefficients collectively offer no additional predictive value.

F-statistic Formula:

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)}$$

We will explore this equation in detail in the sections below.

Let's start by looking at an example where one variable is significant and the other is not:

Let us imagine that we have the following example, where we have 2 variables ‘’ Hours Slept” and ‘’Hours Worked” and that we have collected data from 8 students:

Student	**Hours Slept**	**Hours Worked**	**Test Score**
1	7	2	84
2	6	3	76
3	8	1	89
4	5	4	70
5	6	2	80
6	8	1	95
7	9	1	98
8	6	3	82

Table 1: Predicting Test Score using Hours Slept and Hours Worked

We will perform multiple linear regression and examine the results.

The following python code has been used:

```
#Input data
data = { 'Hours Slept': [7, 6, 8, 5, 6, 8, 9, 6],
     'Hours Worked': [2, 3, 1, 4, 2, 1, 1, 3],
     'Test Score': [84, 76, 89, 70, 80, 95, 98, 82]
}

df = pd.DataFrame(data)

# Define the predictors and the response variable
X = df[['Hours Slept', 'Hours Worked']]
Y = df['Test Score']

# Add a constant term for the intercept
X = sm.add_constant(X)

# Fit the regression model
model = sm.OLS(Y, X).fit()

# Print the summary of the model
print(model.summary())
```

We will proceed by analyzing the results we have derived from the above.

Individual Significance (t-tests)

Hours Slept: The t-statistic for Hours Slept is 4.700 with a p-value of 0.005. Since the p-value is less than 0.05, Hours Slept is individually significant.

Hours Worked: The t-statistic for Hours Worked is -1.364 with a p-value of 0.233. Since the p-value is greater than 0.05, Hours Worked is not individually significant.

Joint Significance (F-test)

The F-statistic is 21.98 with a p-value of 0.0008. Since the p-value is much less than 0.05, the combination of Hours Slept and Hours Worked is jointly significant, meaning that at least one of these variables significantly contributes to predicting Test Score when considered together.

Summary:

Hours Slept is individually significant (p is less than 0.05). The variable Hours Slept alone significantly explains variation in Test Scores. When we put it together with the variable Hours Worked (joint significance) the model is significantly better explained.

Hours Worked is not individually significant (p is greater than 0.05). However, when we put it together with the variable Hours Slept, we find that the model significantly better explains the variation Test Scores.

From the analysis, we can see that it is possible for a variable to be individually significant while another is not. Individual and joint significance may differ due to the interplay of predictors in explaining the dependent variable.

Expanded Scenarios Demonstrating Variable Significance Discrepancies

1. **High Correlation Among Predictors (Multicollinearity):** *Explanation:* When variables that are highly correlated with each other are included in a model, it may be difficult to distinguish their individual effects because they contribute shared information to the model's explanatory power.

   Example: In health research, variables like cholesterol levels and dietary fat intake may both individually correlate with heart disease risk. However, their joint significance may not be as pronounced since both are related to dietary habits and collectively do not add unique information.

2. **Redundancy and Overlapping Information:** *Explanation:* This occurs when multiple variables convey similar information, causing the regression model to fail in demonstrating their combined utility.

   Example: Economic models might find that consumer confidence and disposable income both forecast consumer spending effectively alone. However, their overlap in explaining economic well-being could make their combined effect non-significant.

3. **Inclusion of Non-Informative Variables:** *Explanation:* Adding variables that do not directly impact the dependent variable can clutter the model, reducing the overall significance of the set.

   Example: In predicting employee productivity, including the variable 'day of the week' alongside hours worked and training sessions might obscure the significance of more relevant predictors in joint tests.

4. **Change in Model Context or Dependent Variable:** *Explanation:* Variables significant under one model specification may lose significance under another if the dependent variable or the context shifts, altering the dynamics within the model.

   Example: Variables like advertising spend and seasonality might predict short-term sales effectively but fail to show joint significance in a long-term sales model influenced more by market trends and macroeconomic factors.

5. **Control Variables:** *Explanation:* The introduction of control variables can absorb some of the explanatory power of primary variables if they share a common underlying relationship with the dependent variable.

   Example: The significance of location in real estate pricing models may decrease when control variables such as zoning laws or economic conditions are included, as they might capture aspects of location benefits more effectively.

Effects on F-statistic When Removing/Adding Variables

The F-statistic is used to test the joint significance of multiple coefficients. When we remove or add variables to the model, the F-statistic will be affected based on how these changes alter the residual sum of squares (RSS) in the restricted and unrestricted models.

F-statistic Formula:

$$F = \frac{(RSS_R - RSS_{UR})/q}{RSS_{UR}/(n-k-1)}$$

where:

- RSS_R: Residual Sum of Squares for the restricted model (fewer variables).
- RSS_{UR}: Residual Sum of Squares for the unrestricted model (more variables).
- q: Number of restrictions (number of variables removed or added).
- n: Total sample size.
- k: Number of parameters in the unrestricted model (excluding the intercept).

Removing Variables: When removing a variable from the unrestricted model, the RSS_{UR} typically increases because the model loses explanatory power. The RSS for the restricted model (RSS_R) remains unchanged. If the removed variable(s) were contributing significantly to the model, the increase in RSS_{UR} will be substantial, leading to a lower F-statistic. A lower F-statistic might indicate that the remaining variables are jointly insignificant.

- *Example:* If removing 'education' from the model significantly increases RSS_{UR}, it suggests that 'education' was an important predictor. Consequently, the F-statistic will decrease, indicating a loss in joint significance of the remaining variables.

Adding Variables: When adding a variable to the restricted model, the RSS_R typically decreases because the model gains explanatory power. The RSS for the unrestricted model (RSS_{UR}) remains unchanged. If the added variable(s) contribute significantly, the decrease in RSS_R will be substantial, leading to a higher F-statistic. A higher F-statistic enhances the joint significance of the included variables.

- *Example:* If adding 'education' to the model significantly decreases RSS_R, it suggests that 'education' is an important predictor. Consequently, the F-statistic will increase, indicating an enhancement in joint significance of the variables.

Advanced Considerations in Model Building

Upon further research, we identified several advanced techniques that can effectively address issues highlighted in the scenarios above:

- **Strategic Variable Selection:** It is crucial to critically evaluate the necessity and contribution of each variable. Employing a theoretically informed selection process helps in avoiding redundancy and enhances model interpretability.

- **Diagnostic Tools for Multicollinearity:** Techniques like the Variance Inflation Factor (VIF) provide quantifiable measures to assess the degree of multicollinearity among predictors, guiding decisions on which variables to retain or exclude.
- **Sophisticated Modeling Techniques:** Methods such as ridge regression and elastic net incorporate penalties that reduce the impact of multicollinearity and help in selecting the most relevant predictors by shrinking coefficients of less important variables, thus stabilizing the model's predictions.

Conclusion

This refined understanding of individual versus joint significance of variables is critical for constructing robust regression models. A detailed analysis and the appropriate application of advanced statistical techniques ensure that the resulting models are not only statistically valid but also practically sound, providing meaningful insights into the relationships they are intended to capture. In conclusion, only the t-test can find the significance of a single coefficient being more than zero, whereas the f-test is only suited to determine the significance of a whole set of variables's coefficient being more than zero. Therefore, it is entirely possible that we can have an outcome where we find a variable to be individually significant but jointly insignificant. This outcome would be cause for concern for a regression, because it would mean that a variable we likely want to focus on is getting drowned out by the coefficient of the set of other variables in the f-test, and we can have misleading conclusions.

References

[1] U.S. News & World Report, *Law School Cost and Starting Salary*, Retrieved from `https://www.usnews.com/education/best-graduate-schools/top-law-schools/articles/law-school-cost-starting-salary`, 2024.

[2] Wooldridge, J. M. (2018). *Introductory Econometrics: A Modern Approach.* 7th edition. Boston: Cengage. ISBN: 9781337558860. pp- 136-145.