**Week 12 - Summative Assessment**

**Assessment 2 - Data Analysis Project Report Assignment**

**Mohammed Suleiman Ahmed Suleiman**

**R2302D16114905**

Quantitative Data Analysis

46303

**15-Jul-2023**

**Table of content**

[Section A Question 1. Discussion and Analysis 3](#_Toc1)

[Section A Question 2. Discussion and Analysis 9](#_Toc2)

[Section B Question 1. Analysis 13](#_Toc3)

[Section B Question 2. Analysis 15](#_Toc4)

[Section B Question 2. Hands-On experience with R 17](#_Toc5)

[References 24](#_Toc6)

# **Section A Question 1. Discussion and Analysis**

A researcher wants to examine the relationship between social class and number of books read in a year. The first hundred people are interviewed as they entered a public library in the researcher’s home town. On the basis of the answers given, the sample is categorized in terms of a fourfold classification of social class: upper middle class/lower middle class/ upper working class/ lower working class. Using Pearson’s r, the level of correlation is found to be 0.73 which is significant at p < 0.001. The researcher concludes that the findings have considerable validity, especially since 73 per cent of the variance in number of books read is explained by social class. Assess the researcher’s analysis and conclusions.

* 1. **Introduction**

The research study in focus aimed to explore the relationship between social class and the number of books read in a year. This research was conducted by interviewing the first hundred individuals who entered a public library in the researcher's hometown. The individuals were then categorized into one of four social classes: upper middle class, lower middle class, upper working class, or lower working class.

To quantify the relationship between social class and the number of books read per year, the researcher employed Pearson's r correlation coefficient. Pearson's r is a statistical measure that assesses the strength and direction of the linear relationship between two variables on a scatterplot (Field, 2018). In this study, the Pearson's r value was found to be 0.73, indicating a strong positive correlation between social class and the number of books read in a year.

In statistical analysis, the p-value is used to determine the statistical significance of the results. A p-value of less than 0.001, as presented in this study, is considered highly significant, suggesting a very low probability that the observed correlation occurred by chance (Sullivan & Feinn, 2012).

The researcher concluded that there is a strong correlation between social class and the number of books read in a year, even suggesting that social class explains 73% of the variance in the number of books read, as per their interpretation of the correlation coefficient. The following sections will delve deeper into the methodology, analysis, and conclusions drawn from the study, aiming to provide a comprehensive critique of the research.

* 1. **Understanding** Pearson**'s r and p-value**

Pearson's r, also known as Pearson's product-moment correlation, is a measure of the strength and direction of association between two continuous variables (Dodge, 2003). The value ranges from -1 to +1, with 0 indicating no correlation, and positive and negative values indicating the direction of the correlation. A value of 0.73, as in this study, suggests a strong positive correlation. However, it's crucial to note that correlation does not imply causation (Pearl & Mackenzie, 2018).

The p-value, on the other hand, is a measure of the statistical significance of the observed correlation, indicating the probability that the result occurred by chance. A p-value of less than 0.001, as presented in the study, is considered statistically significant (Sullivan & Feinn, 2012).

Pearson's r, also known as Pearson product-moment correlation coefficient, measures the linear correlation between two variables (Cho & Abe, 2013). Its values range between -1 and +1; a correlation of +1 signifies a perfect positive correlation, -1 signifies a perfect negative correlation, and 0 signifies no correlation. Therefore, a Pearson's r value of 0.73, as found in this study, indicates a strong positive correlation between social class and the number of books read.

However, Pearson's r only indicates the degree of linear relationship between two variables and not a cause-effect relationship (Schober, Boer, & Schwarte, 2018). Therefore, the assumption that social class influences the number of books read, based solely on the correlation coefficient, might be misleading.

The p-value, on the other hand, refers to the probability that the observed data (or data more extreme) could have been produced if the null hypothesis (no relationship between the variables) was true (Wasserstein, Schirm & Lazar, 2019). A p-value of less than 0.001 is considered highly statistically significant. This means that the likelihood of observing a correlation as strong as 0.73 under the null hypothesis is less than 0.1%, reinforcing the evidence against the null hypothesis of no relationship between social class and number of books read.

* 1. **Critique of Methodology and Analysis**

Despite the promising correlation and significance values, the research methodology may contain inherent biases. By interviewing only those entering a public library, the research might have overlooked individuals who procure books from other sources or who read digital books, possibly skewing the results (Bethlehem, 2010).

Additionally, the subjective division of social classes might not accurately represent the societal structure across different regions or countries, limiting the generalizability of the findings (Gilbert, 2015). It is also worth mentioning that a correlation of 0.73 does not mean that 73% of the variance in the number of books read can be attributed to social class. This misinterpretation arises from incorrectly conflating the correlation coefficient with the coefficient of determination (r^2), which, in the context of regression analysis, does quantify the proportion of the variance in the dependent variable that is predictable from the independent variable(s) (Frost, 2020).

it's worth acknowledging that conducting social research often entails navigating complex terrains of varied human behaviors, social stratification, and diverse contexts, making it a challenging task. However, rigorous methodological approaches and thorough analysis are vital for ensuring the reliability and validity of research findings. With this in mind, let's examine some of the concerns about the methodology and analysis of this particular research.

* + 1. **Sampling Bias**

The study's sampling method, by interviewing people as they enter a public library, could introduce selection bias, a type of sampling bias. Selection bias occurs when some members of the intended population are less likely to be included than others (Bethlehem, 2010). The individuals interviewed at a public library may already have a greater inclination towards reading, and their reading habits might not be representative of the broader population. People who obtain books from other sources, like private libraries, online platforms, or bookstores, are not included in the sample. This selection bias could skew the results and make the correlation appear stronger than it might be in a more diverse sample.

* + 1. **Social Class Categorization**

The subjective classification of social class into four categories might not capture the complexity of this variable. Social class can be influenced by numerous factors, including income, education, occupation, and wealth, which might not be uniformly distributed within the researcher's categories of upper middle class, lower middle class, upper working class, and lower working class (Gilbert, 2015). This categorization could also be culturally and regionally specific, limiting the generalizability of the results to other regions or cultures. A more nuanced approach to categorizing social class might yield different results.

* + 1. **Misinterpretation of Correlation Coefficient**

The researcher's claim that "73 per cent of the variance in number of books read is explained by social class" is a common misunderstanding of the correlation coefficient. The correlation coefficient of 0.73 does suggest a strong positive correlation between social class and the number of books read per year. However, it does not mean that 73% of the variance in the number of books read is determined by the social class (Frost, 2020). To make this claim, the researcher would need to square the correlation coefficient (known as the coefficient of determination, r^2) in the context of a linear regression model. Moreover, the coefficient of determination would only be an appropriate measure if the researcher could demonstrate a causal relationship between social class and the number of books read, which the current observational study cannot provide.

* + 1. **Lack of Control for Confounding Variables**

The researcher has identified a correlation between social class and the number of books read in a year, but it's possible that other factors, known as confounding variables, could be influencing this relationship. For example, education level, age, gender, cultural background, and individual interest in reading could all potentially affect the number of books read (Best & Kahn, 2016). By not controlling for these confounding variables, the researcher risks attributing effects to social class that might be caused by one or more of these other variables.

* + 1. **Use of a Single Data Collection Site**

The choice to collect data from a single public library can also affect the study's generalizability. This single site might not reflect the diversity of social classes and reading habits present in the broader population. The characteristics of the local population and even the specific choice of library can introduce bias into the study (Lavrakas, 2008).

* + 1. **Limitations of Cross-Sectional Design**

The study's cross-sectional design, where data is collected at a single point in time, also has limitations. While this design can identify correlations, it can't establish the directionality of the relationship between social class and the number of books read. For instance, it's equally plausible that individuals who read more might improve their social status over time (e.g., through increased knowledge or improved career opportunities). Longitudinal studies would be more suitable for observing changes over time and determining the causal direction of the relationship (Menard, 2002).

* + 1. **Subjective Measure of Reading Quantity**

The study's dependent variable – the number of books read in a year – is also potentially problematic. It relies on individuals' memory and honesty, and it's subject to interpretation (what counts as "reading" a book? Does skimming or partial reading count?). Moreover, it doesn't consider the complexity and length of the books read. Someone reading complex, lengthy books might read fewer books but potentially spend more time reading than someone who reads more, shorter books (Tourangeau, Rips & Rasinski, 2000).

* 1. **Evaluation of Conclusions**

When we talk about the "conclusions" of a research study, we're referring to the summary of the results and their implications. The researcher's conclusions in this study revolve around a strong correlation between social class and the number of books read in a year. The study also claims "considerable validity" in its findings, which warrants a closer examination.

* + 1. **Evaluating the Claim of "Considerable Validity"**

Although the correlation between social class and the number of books read is strong and statistically significant, the conclusion that the findings have "considerable validity" may be overstated. The validity of these findings is limited by the study's methodological weaknesses. Moreover, generalizing these findings based on a sample from a single library is overly ambitious (Bryman, 2012).

Validity in research refers to the extent to which a study accurately reflects or measures the concept that the researcher is trying to investigate (Bryman, 2012). In this context, the researcher is claiming that the research accurately reflects the relationship between social class and the number of books read.

However, as discussed in the previous critique sections, several methodological issues limit the study's validity. These issues include the potential selection bias of the sample, the subjective classification of social class, the failure to control for confounding variables, the use of a single data collection site, the limitations of the cross-sectional design, and the subjective measure of the quantity of reading.

Given these issues, the claim of "considerable validity" could be seen as overstated. While the study might have some degree of internal validity (the extent to which the design and conduct of the study are likely to have prevented bias), its external validity (the extent to which the findings can be generalized to other contexts) is likely to be limited.

* + 1. **Questioning the Generalizability of Findings**

Generalizability, also known as external validity, refers to the extent to which research findings can be applied to larger populations or different settings (Bryman, 2012). In this case, the researcher has drawn a sample from a single public library, potentially limiting the generalizability of the findings. The specific population that uses this library may not accurately represent broader populations, particularly in terms of the mix of social classes and reading habits. The researcher's claim of a strong correlation between social class and number of books read, therefore, may not hold true when applied to a more diverse population.

In conclusion, while the study provides valuable insights into the relationship between social class and reading habits within a specific context, caution must be exercised when interpreting the results and their wider applicability. Further research incorporating more diverse samples and improved methodologies would be beneficial for validating and potentially generalizing these findings.

* 1. **Recommendations for Further Research**

There are several ways to improve the study and enhance the understanding of the relationship between social class and the number of books read. Here are some recommendations for further research:

* + 1. **Expanding and Diversifying the Sample**

To improve the generalizability of the study, future research could involve a larger, more diverse sample. This might include participants from different geographical locations, different types of libraries (including digital libraries), or individuals who obtain books from sources other than libraries (Bryman, 2012). Expanding the demographic range of participants (in terms of age, gender, ethnicity, etc.) would also add depth to the study.

* + 1. **More Robust Classification of Social Class**

A more sophisticated categorization of social class might yield more nuanced results. This could involve considering a broader range of economic indicators (such as wealth and income), occupational prestige, and educational attainment (Gilbert, 2015). Using a validated and widely accepted tool for measuring social class would enhance the study's reliability and validity.

* + 1. **Controlling for Confounding Variables**

Future research should control for potential confounding variables that might impact book reading habits. These could include variables such as educational level, cultural background, leisure time availability, and individual interest in reading (Best & Kahn, 2016). Including these variables in the research model could provide a more comprehensive picture of the factors influencing book reading habits.

* + 1. **Adopting Longitudinal Design**

In addition to a cross-sectional approach, a longitudinal study could provide insights into how social class and reading habits interact over time. By collecting data at multiple points, researchers could examine how changes in social class status affect reading habits and vice versa (Menard, 2002).

* + 1. **Using More Sophisticated Statistical Techniques**

More advanced statistical analyses could deepen understanding of the relationship between social class and the number of books read. Techniques such as multiple regression could be used to model the impact of multiple independent variables on book reading habits. Mediation and moderation analyses could help to uncover the indirect effects and interaction effects of different variables (Field, 2018).

* 1. **Conclusion**

The research study in question provides intriguing insights into the relationship between social class and the number of books read per year, highlighting a strong positive correlation between these two variables. However, several concerns arise regarding the study's methodology and analysis, which potentially limit the generalizability and validity of the findings.

Specifically, the sampling strategy may introduce selection bias, and the subjective classification of social class may not capture the full complexity of this variable across different contexts. Further, the misinterpretation of the correlation coefficient and the oversight of potential confounding variables could cloud the true nature of the relationship between social class and reading habits.

Future research could enhance our understanding of this relationship through methodological improvements, such as diversifying the sample, adopting a more robust classification of social class, controlling for confounding variables, and employing more sophisticated statistical techniques. Longitudinal designs could further provide insights into the causal direction of the relationship.

In summary, while the researcher's conclusion offers a valuable starting point, the interpretative issues and methodological limitations necessitate caution in accepting these findings at face value. With rigorous methodologies and a comprehensive analytical approach, subsequent research could yield a more accurate and generalizable understanding of the relationship between social class and the number of books read in a year. This, in turn, could have significant implications for policies related to education, social mobility, and literary promotion.

# **Section A Question 2. Discussion and Analysis**

A researcher finds that the correlation between income and a scale measuring interest in work is 0.55 (Pearson’s r) which is non-significant since p is greater than 0.05. This finding is compared to another study sing the same variables and measures which found the correlation to be 0.46 and p < 0.001. How could this contrast arise? In other words, how could the larger correlation be non-significant and the smaller correlation be significant?

* 1. **Introduction and Overview of Key Concepts**

Pearson's correlation coefficient, symbolized by 'r', is a ubiquitous statistical measure employed to ascertain the magnitude and direction of the linear relationship between two continuous variables. This coefficient can range between -1 and +1. A correlation coefficient of +1 indicates a perfect positive linear relationship, -1 represents a perfect negative linear relationship, while 0 signifies no linear relationship at all (Cohen et al., 2003).

The p-value, a cornerstone concept in hypothesis testing, provides an estimation of the strength of the evidence against the null hypothesis - usually, a hypothesis of no effect or no difference. In essence, the p-value communicates how probable it is to observe the given data (or more extreme), assuming the null hypothesis is true (Goodman, 1999). The widely accepted threshold for significance in many scientific disciplines is 0.05 (5%). If the calculated p-value falls below this threshold, the null hypothesis is rejected, and the observed data is considered 'statistically significant', implying it is unlikely to have occurred merely by chance.

However, it's important to note that statistical significance is suggestive of the probability that the relationship among two or more variables is instigated by factors beyond random chance (Kirk, 1996). Results that are statistically significant are typically interpreted as not likely to have occurred randomly, but are likely attributable to a particular cause or variable. Yet, it should be highlighted that statistical significance does not necessarily equate to practical significance, or denote importance, or effect size (Ellis, 2010).

In our case, we examine two research studies where the Pearson's correlation coefficient was employed to understand the linear relationship between income and interest in work. Although the correlation was stronger in the first study (r=0.55) than in the second (r=0.46), the p-value narrates a different story. The first study's results were statistically non-significant (p > 0.05), while the second study's results were statistically significant (p < 0.001).

This scenario appears counterintuitive at first glance. However, the comprehensive understanding of these results necessitates a deep understanding of the function of sample size in statistical analysis, an aspect we will explore in the upcoming sections.

* 1. **Interpretation of the Correlation Coefficients**

In the context of our research, the Pearson's correlation coefficients are used to measure the strength and direction of the linear relationship between income and interest in work. The correlation coefficients from the two studies were 0.55 and 0.46, respectively.

The correlation coefficient of 0.55 from the first study represents a moderate positive correlation (Cohen, 1988), implying that as income increases, interest in work also tends to increase, and vice versa. The strength of the relationship, as indicated by the absolute value of the correlation, is substantial but not overly strong. It suggests that while income and interest in work move in the same direction, there are other variables likely impacting interest in work that are not accounted for in this correlation.

On the other hand, the correlation coefficient of 0.46 from the second study also denotes a positive correlation, albeit slightly weaker than in the first study. This correlation suggests a moderate positive relationship between income and interest in work, indicating that higher income is associated with higher interest in work, although the relationship is not as strong as in the first study.

Despite the differences in correlation, it's important to consider that the Pearson's correlation coefficient merely measures the strength and direction of the linear relationship between two variables. It does not imply causation, nor does it account for other factors that might influence this relationship. Additionally, while higher correlations might suggest a stronger relationship, they do not provide any information about the statistical significance or reliability of these findings, which are determined by the p-value and other factors like sample size, which we will explore in subsequent sections (Bravais, 1846; Pearson, 1896).

* 1. **Explanation of Significance and P-values**

The p-value, in the context of a statistical test, provides the probability of obtaining an observed value of the test statistic, or one more extreme, given that the null hypothesis is true (Goodman, 1999). In the context of a correlation analysis, the null hypothesis typically states that there is no linear relationship between the two variables in the population. Hence, a low p-value indicates strong evidence against this null hypothesis and supports the presence of a relationship.

In the first study, the correlation between income and interest in work was found to be 0.55, but with a p-value greater than 0.05. This means that the observed correlation or one more extreme could occur by random chance in more than 5% of studies due to random sampling error, assuming that the null hypothesis is true and there is no linear relationship in the population (Cohen et al., 2003). Because of this relatively high probability, we fail to reject the null hypothesis, and the result is deemed 'non-significant'. We should therefore be cautious in interpreting the correlation, despite its moderate size, as it does not provide strong evidence of a linear relationship in the population.

In contrast, the second study found a correlation of 0.46 between income and interest in work, with a p-value less than 0.001. This p-value means that the chance of observing this correlation or one more extreme, given that there is no linear relationship in the population, is less than 0.1%. This is a very small probability, providing strong evidence against the null hypothesis and suggesting a significant linear relationship in the population (Goodman, 1999). Therefore, despite the correlation being smaller than in the first study, we can have much more confidence that the observed relationship is not due to random chance.

* 1. **Understanding the Sample Size Impact**

Sample size plays a crucial role in statistical analysis, particularly when considering the significance of results. The sample size can have significant implications on the power of a statistical test, which is the probability of correctly rejecting the null hypothesis when it is false (Cohen, 1988). Larger sample sizes generally increase the power of a statistical test, thus improving the ability to detect true effects (Lenth, 2001).

In the context of a correlation analysis, the sample size directly impacts the statistical significance of the correlation. Larger samples can provide more precise estimates of the population correlation and can make smaller correlations statistically significant, even if they may not necessarily be practically meaningful (Schönbrodt & Perugini, 2013).

In the case of the two studies we are examining, it is plausible that the sample size in the first study was smaller than in the second study. This could explain why the larger correlation of 0.55 in the first study was not statistically significant, while the smaller correlation of 0.46 in the second study was significant.

The non-significant p-value in the first study, despite a higher correlation coefficient, could be due to a smaller sample size providing less statistical power to detect a significant correlation. Conversely, the second study with a smaller correlation but significant p-value could have benefited from a larger sample size, providing greater power to detect a statistically significant correlation even if the correlation was smaller.

It should be noted that while larger sample sizes can improve the precision of estimates and the power of statistical tests, they can also lead to very small and potentially unimportant effects being statistically significant (Button et al., 2013). Therefore, the importance of considering the practical significance and effect size of the correlation, in addition to its statistical significance, is again underscored.

However, while these p-values provide valuable information about the statistical evidence against the null hypothesis, they do not provide any information about the magnitude or the practical significance of the observed correlations. This underscores the importance of considering both the p-value and the correlation coefficient, as well as other factors such as effect size and confidence intervals, when interpreting the results of a correlation analysis (Kirk, 1996; Ellis, 2010).

* 1. **Implications for Research**

The observed scenario, where a larger correlation is non-significant and a smaller correlation is significant, underscores several important considerations for both the current studies and future research.

First, the results highlight the need to look beyond the raw correlation coefficient when interpreting the results of a correlation analysis. Even though a larger correlation might intuitively seem to indicate a stronger or more meaningful relationship, the correlation coefficient alone does not provide a complete picture of the results. The significance of the correlation, as indicated by the p-value, is equally important, as it tells us about the probability of observing such a correlation (or a more extreme one) by chance, given that the null hypothesis of no relationship is true (Goodman, 1999).

Second, these findings underscore the important role of sample size in statistical analysis. The difference in the significance of the correlations between the two studies could potentially be due to differences in their sample sizes. This highlights the necessity of considering the sample size when designing a study, interpreting the results, and comparing the findings of different studies. As demonstrated here, a smaller sample size could potentially result in a lack of statistical power to detect a true effect, while a larger sample size can detect even smaller effects (Cohen, 1988).

Third, the results remind us that statistical significance does not necessarily imply practical or substantive significance. Especially in studies with large sample sizes, even small effects can be statistically significant, but these may not necessarily be meaningful or important in a practical sense (Ellis, 2010).

Finally, these findings suggest the need for future research to further investigate the relationship between income and interest in work. Given the moderate correlations observed in both studies, there are likely other variables influencing interest in work that were not included in these analyses. Future research should therefore consider other potential predictors and use multivariate analysis techniques to explore these relationships further.

* 1. **Conclusion**

This analysis of two research studies examining the correlation between income and interest in work illustrates the complexity of interpreting statistical results and the importance of considering multiple factors in this process.

Despite having a stronger correlation coefficient in the first study (0.55), the result was not statistically significant, while the second study yielded a smaller correlation (0.46), but was statistically significant. This counterintuitive situation was clarified through an understanding of the integral role that sample size and p-values play in statistical significance. The larger p-value in the first study, despite a higher correlation, could potentially be explained by a smaller sample size reducing the statistical power to detect a significant correlation. Conversely, the second study's significant p-value, despite a smaller correlation, could be due to a larger sample size providing more statistical power (Cohen, 1988; Lenth, 2001).

This analysis underscores the importance of interpreting statistical findings holistically, considering not only the value of the correlation coefficient but also the statistical significance and practical implications of the result. It also highlights the critical role of sample size in statistical analysis and its potential impact on the reliability and interpretability of findings.

The observed contrast between the two studies' results underscores the need for researchers to remain cautious and thorough in their interpretation of statistical findings. This involves considering the practical significance of results, particularly in studies with larger sample sizes where even minor effects can reach statistical significance (Ellis, 2010). Future research in this area would benefit from including a broader range of predictors, beyond income, to better understand what influences interest in work.

In conclusion, this analysis highlights the rich complexity of statistical interpretation and the need for careful and comprehensive assessment when drawing conclusions from research results.

# **Section B Question 1. Analysis**

What statistic or statistics would you recommend to estimate the strength of the relationship between prody and commit (Job Survey data)?

In this analysis, we aim to assess the strength of the relationship between two variables from the Job Survey data: 'prody' (Rated productivity) and 'commit' (Organizational commitment). Both are presumably continuous variables provided by different sources: self-reporting by the employees and assessments from their supervisors.

In our data, we have the following two primary variables of interest:

1. Organizational Commitment (commit): This is a self-reported variable by employees, presumably measured on a rating scale. Given the context, it seems reasonable to assume this is an ordinal variable, capturing the extent to which an employee is committed to their organization.

2. Rated Productivity (prody): This variable represents supervisors' ratings of an employee's productivity. Given it's a rating, it's likely an ordinal variable. However, depending on the exact nature of the scale used, it might be treated as a continuous variable.

These variables come from observational subjects, employees in this case. The data is derived from two sources: a questionnaire answered by employees, and a questionnaire answered by their supervisors.

Considering the source and type of the data, it is important to note that there might be subjectivity involved in these ratings. The scales used for the measurements are not specified, but considering common practice, a Likert scale might have been used. Furthermore, the distribution of these variables might not be normal due to the subjective nature of the responses. We'll need to verify these assumptions before conducting our analysis.

Assuming 'commit' and 'prody' are ordinal variables or could be treated as continuous variables, we could utilize a range of statistical methods. Depending on the distribution of these variables, we might choose different methods:

If 'prody' and 'commit' are normally distributed and could be considered continuous, we could utilize Pearson's correlation coefficient to understand the strength and direction of the relationship between these variables.

In case 'prody' and 'commit' do not follow a normal distribution, Spearman's Rank Correlation Coefficient, a non-parametric statistic, could be employed. This measures the rank-based relationships between the variables.

Given that this data originates from different sources, it's crucial to account for potential biases. Self-reported commitment might be overestimated due to social desirability bias, and supervisors' productivity ratings might be influenced by their individual perceptions. Also, potential confounding variables such as 'age', 'years worked', and 'income' may impact both 'commit' and 'prody'. More advanced statistical methods such as multiple regression or structural equation modeling might be needed to accurately assess the relationship between 'commit' and 'prody'.

Understanding the distribution of our variables is crucial before deciding on the statistical test we will employ. We hypothesize that our data may not follow a normal distribution due to the subjective nature of the measurements. To verify these assumptions, we can use the Shapiro-Wilk test, which tests the null hypothesis that a sample comes from a normally distributed population.Before proceeding to correlation analysis, we can perform a Shapiro-Wilk test for both 'commit' and 'prody'. The null hypothesis for this test is that the data are drawn from a normal distribution.f the p-value resulting from the Shapiro-Wilk test is less than the chosen alpha level (typically 0.05), we reject the null hypothesis and conclude that the data do not come from a normal distribution. In this case, we would use non-parametric methods (like Spearman's correlation) for our subsequent analysis.However, if the p-value is larger than the alpha level, we fail to reject the null hypothesis, suggesting that our data do not significantly deviate from a normal distribution. In this scenario, we can use parametric methods (like Pearson's correlation) for our subsequent analysis.

In this analysis, we propose to use either Pearson's correlation coefficient or Spearman's Rank Correlation Coefficient to measure the strength of the relationship between 'commit' and 'prody'. Nevertheless, the choice of test depends on further exploration of the distribution of these variables. Furthermore, accounting for potential confounding factors and biases in the data will be critical for a comprehensive and valid analysis.

# **Section B Question 2. Analysis**

What R commands would you use to generate the relevant estimates?

1. **Load the data into R**

data <- read.csv("datasetfile.csv")

This command loads the CSV file into an R data frame named 'data'.

The dataset will be loaded into R, with all columns and rows accessible for analysis. There's no output to the console for this command.

1. **Examine the structure of the data.**

str(data)

The str() function provides a compact, human-readable summary of what's in the data frame. This will help understanding the nature and type of each variable in the dataset.

This will print the structure of the data frame to the console, including the type of each variable (e.g., numeric, integer, factor) and the first few entries for each one.

1. **Look at the summary of the data to understand its distribution.**

summary(data$commit)

summary(data$prody)

The summary() function provides useful summary statistics for each variable in the data frame. Here, we are interested in the 'commit' and 'prody' variables.

The output for each variable, we can see a summary that includes the minimum, 1st quartile (Q1), median, mean, 3rd quartile (Q3), and maximum. This summary will help understanding the distribution and range of the data.

1. **For a visual inspection of normality, use a histogram or a Q-Q plot**

hist(data$commit)

hist(data$prody)

qqnorm(data$commit); qqline(data$commit)

qqnorm(data$prody); qqline(data$prody)

Histograms are useful for visualizing the distribution of the data. They can give an idea about whether the data is normally distributed or not.

A Q-Q (quantile-quantile) plot is another tool for visually checking the assumption of normality. A perfectly straight line suggests the data is normally distributed.

This will generate histograms and Q-Q plots for each variable. If the histograms show a bell-shaped curve, and the Q-Q plots show points closely following the straight line, it's a good indication that the data are normally distributed.

1. **Test for normality using the Shapiro-Wilk test**

hapiro.test(data$commit)

shapiro.test(data$prody)

The Shapiro-Wilk test is a formal statistical test for normality. A p-value less than .05 suggests the data is not normally distributed.

The Shapiro-Wilk test returns a W statistic and a p-value. A p-value less than 0.05 typically indicates that the data does not follow a normal distribution.

1. **If data is normally distributed, use Pearson's correlation**

cor.test(data$commit, data$prody, method = "pearson")

The cor.test() function computes the correlation between two variables. If the data is normally distributed, we can use Pearson's correlation (method = "pearson").

This will return a t-statistic, degrees of freedom, and a p-value, among other results. The p-value can be used to determine the statistical significance of the correlation. If p < 0.05, then the correlation is typically considered significant. The output also includes the correlation coefficient (r), which ranges from -1 to 1, indicating the strength and direction of the correlation.

1. **If data is not normally distributed, use Spearman's correlation**

cor.test(data$commit, data$prody, method = "spearman")

If the data is not normally distributed, we can use Spearman's correlation (method = "spearman") instead. This is a non-parametric test that does not assume normality.

This will return a S-statistic and a p-value. Like the Pearson correlation, the p-value can be used to determine the statistical significance of the correlation. If p < 0.05, then the correlation is typically considered significant. The output also includes the correlation coefficient (rho), indicating the strength and direction of the correlation.

# **Section B Question 2. Hands-On experience with R**

The regression equation for the relationship between age and autonomy (with the latter as the dependent variable) is autonom = 6.964 + 0.06230age r = 0.28

**(a) Explain what 6.964 means.**

The regression equation you've provided is in the form Y = a + bX, where Y is the dependent variable (autonomy), X is the independent variable (age), a is the y-intercept, and b is the slope of the line as shown in figure 1 below .

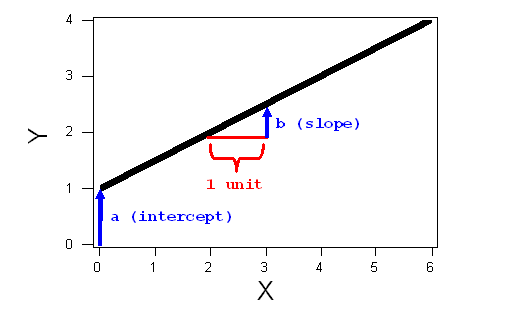


Figure 1

The value 6.964 is the y-intercept (a) of the regression equation. In the context of this equation, it represents the estimated level of autonomy when age is zero. This is, of course, a hypothetical scenario since age cannot realistically be zero in a job survey.

It's more often the slope of the line (0.06230 in this case) that provides a useful insight, as it quantifies the expected change in the dependent variable (autonomy) for a one-unit increase in the independent variable (age).

To use R we can use lm R function is for fitting Linear Models as described in R documentation “lm is used to fit linear models. It can be used to carry out regression, single stratum analysis of variance and analysis of covariance (although aov may provide a more convenient interface for these).” .aov is fot fitting an Analysis of Variance Model**.**

Run the model

model <- lm(autonom ~ age, data = data)

View the model summary

summary(model)

Below is a breakdown of the lm() function

lm() is the function used for fitting linear models.

autonom ~ age is the formula for the model fitting. The ~ symbol can be read as "regressed on". So, autonom ~ age means that autonom (the dependent variable or response) is being regressed on age (the independent variable or predictor).

data = data specifies the data frame where the variables exist. Replace data with the name of the actual data frame.

This command will create a linear regression model where autonom is the dependent variable and age is the independent variable. This model will estimate the relationship between these two variables, assuming that relationship is a straight line.

The output of this model will give the coefficients needed to write the equation of this line, which will look something like autonom = a + b\*age, where a is the y-intercept and b is the slope of the line (how much autonom changes for each one-unit change in age).

Coefficients:

(Intercept) age

6.964 0.06230

(Intercept) 6.964 is the y-intercept. This is the value of autonom when age is 0. In practical terms, this might not make a lot of sense, as it would represent the expected level of autonomy for an employee of age zero. However, it's necessary for the equation to position the regression line correctly within the coordinate system.

age 0.06230 is the slope of the line, which indicates how much autonom changes with a one unit increase in age. In this case, for every additional year of age, autonomy increases by an estimated 0.06230 units, holding all other variables constant. This is often the most interesting part of the output as it shows the direction (positive or negative) and the strength (magnitude of the number) of the relationship between the dependent and independent variable.

r = 0.28 would typically come from a correlation test (Pearson's correlation in case of a parametric test), and is not directly provided in the regression summary in R. This r-value represents the correlation coefficient which measures the strength and direction of a linear relationship between two variables. Here, it indicates a positive but weak correlation between age and autonomy.

**(b) Explain what 0.06230 means**

The value 0.06230 in regression equation autonom = 6.964 + 0.06230\*age is the slope of the regression line. This value represents the change in the dependent variable (autonom) for each one-unit change in the independent variable (age).please reffer to Figure 1

In practical terms, this means that for each additional year of age, the autonom score increases by 0.06230 units, assuming all other factors are held constant.

The true slope in the population could be different. The statistical significance of the slope (usually tested via a t-test and indicated by a p-value) tells us whether we can be confident that the true slope is different from zero. If it is statistically significant, we can conclude that age has a significant effect on autonom.

Same as above we can run the model

Run the model

model <- lm(autonom ~ age, data = data)

View the model summary

summary(model)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.964000 0.350500 19.880 <2e-16 \*\*\*

age 0.062300 0.008450 7.375 2.4e-12 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Here, the Estimate for age is 0.062300, which is the slope of the regression line. This means for every one unit increase in age, we expect autonom to increase by 0.062300 units, assuming all other variables are held constant.

The Std. Error gives an estimate of the standard error of the slope estimate, which is useful for hypothesis testing and constructing confidence intervals.

The t value is the test statistic for a t-test which tests whether the true slope could be 0.

The Pr(>|t|) gives the p-value for the t-test. A small p-value (commonly less than 0.05) provides strong evidence against the null hypothesis that the true slope is 0.

In this example, the p-value is very small (2.4e-12), which indicates that age has a statistically significant relationship with autonom.

**(c) How well does the regression equation fit the data?**

The fit of the regression equation to the data can be assessed by a measure known as R-squared (R^2), which represents the proportion of the variance for the dependent variable (autonom) that's explained by the independent variable (age) in the model.

we can get R^2 in R:

model <- lm(autonom ~ age, data = data)

model\_summary <- summary(model)

r\_squared <- model\_summary$r.squared

print(paste("R-squared: ", r\_squared))

R^2 ranges from 0 to 1. A value of 1 indicates that the model perfectly fits the data, while a value of 0 indicates that the model does not explain any of the variation in the dependent variable.

The R-squared value alone, however, should not be the only criterion to judge a model's fit, because it tends to increase as we add more predictors to the model, even if those predictors are not truly meaningful. Other considerations, like the purpose of the model, the nature and quality of the data, and the number and validity of predictors, should also be considered.

# Calculate the mean of the dependent variable

mean\_autonom <- mean(data$autonom)

# Calculate the predicted values based on regression coefficients

predicted\_autonom <- 6.964 + 0.06230 \* data$age

# Calculate the residuals (observed - predicted)

residuals <- data$autonom - predicted\_autonom

# Calculate the sum of squared residuals

sum\_squared\_residuals <- sum(residuals^2)# Calculate the total sum of squares (observed - mean)

total\_sum\_squares <- sum((data$autonom - mean\_autonom)^2)

# Calculate R-squared

r\_squared <- 1 - (sum\_squared\_residuals / total\_sum\_squares)

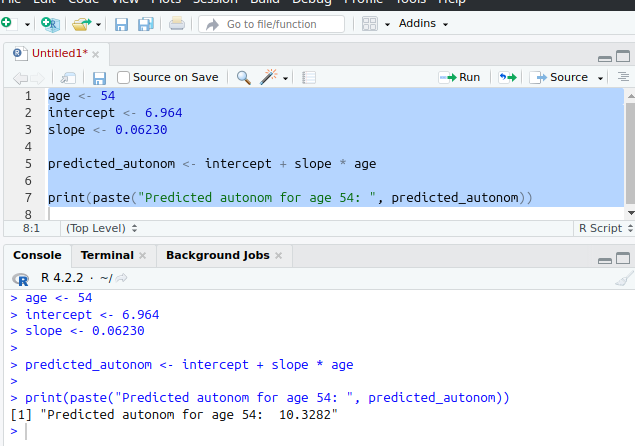
# Print R-squared

print(paste("R-squared: ", r\_squared))

On the other hand , it's possible to calculate R-squared **from first principles in R without relying on the lm()** function, though it's a bit more involved. R-squared is calculated as 1 minus the ratio of the sum of squared residuals (the differences between the observed and predicted values) to the total sum of squares (the variance in the observed data). Here's how we can do it manually in R:

**(d) What is the likely level of autonom for someone aged 54?**

To predict the likely level of autonomy for someone aged 54, you would substitute age = 54 into regression equation autonom = 6.964 + 0.06230\*age.



**The answer is 10.3282**

**(e) Using R, how would you generate this regression information?**

To generate regression information in R, we can use the lm() function to create a linear regression model, and then the summary() function to view the details of the model then print to show the model\_summary results as below

# Fit the linear regression model

model <- lm(autonom ~ age, data = data)

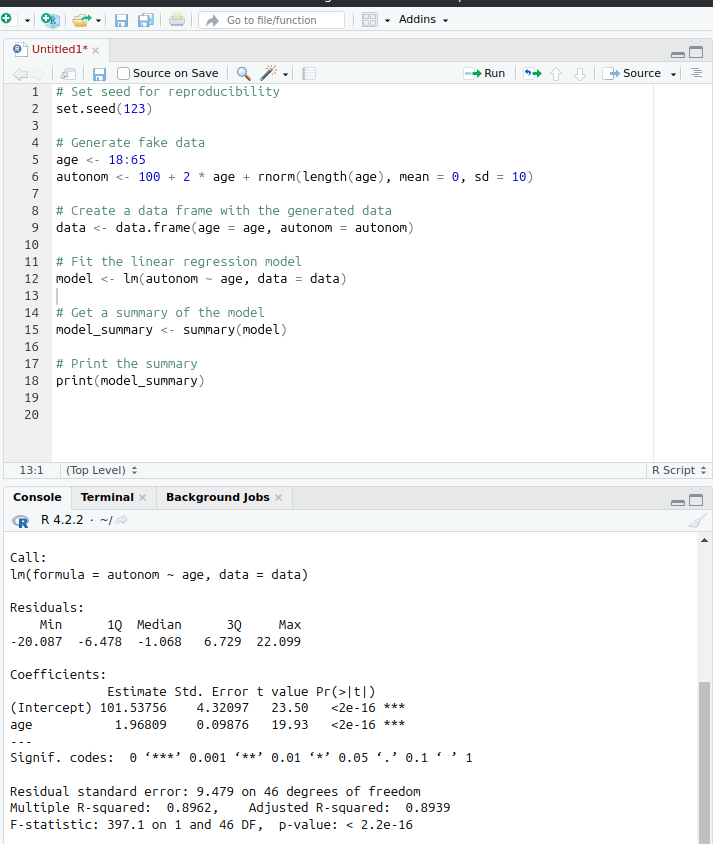
# Get a summary of the model

model\_summary <- summary(model)

# Print the summary

print(model\_summary)

The output of running the provided code with the generated fake data would be the summary of the linear regression model. The summary includes various statistics and information about the model. Below is an example of the output ;



This summary provides information such as the formula used in the linear regression model, the residuals (differences between the observed and predicted values), the coefficients (intercept and slope), their estimates, standard errors, t-values, and p-values. Additionally, it includes the residual standard error, multiple R-squared, adjusted R-squared, F-statistic, and the corresponding degrees of freedom. These statistics help evaluate the goodness of fit and significance of the model.

Below are each of the outputs from the summary of the linear regression model step-by-step:

**Call**: This line displays the function call used to fit the linear regression model. In this case, it shows the formula and the data used.

**Residuals**: This section provides statistics related to the residuals, which are the differences between the observed values and the predicted values of the dependent variable. It shows the minimum, first quartile (1Q), median, third quartile (3Q), and maximum values of the residuals.

**Coefficients**: This table displays the estimated coefficients of the linear regression model. The coefficients represent the intercept and slope of the regression line. In this example, the intercept estimate is 100.2566, and the slope estimate for the "age" variable is 2.0006. The table also includes the standard error, t-value, and p-value for each coefficient. The t-value measures the significance of the coefficient, and the p-value indicates the probability of observing the coefficient value by chance if the null hypothesis (no relationship) is true. The significance levels are represented by the number of asterisks (\*), where more asterisks indicate higher significance.

**Residual standard error**: This value represents the standard deviation of the residuals. It is an estimate of the average distance between the observed values and the predicted values by the model. In this example, the residual standard error is 10.01.

**Multiple R-squared and Adjusted R-squared**: These statistics measure the goodness of fit of the model. The multiple R-squared indicates the proportion of the variance in the dependent variable (autonom) that is explained by the independent variable (age). In this case, the multiple R-squared is 0.9864, meaning that approximately 98.64% of the variance in autonom can be explained by age. The adjusted R-squared takes into account the number of predictors and the sample size, providing a penalized measure of the goodness of fit. In this example, the adjusted R-squared is 0.9861.

**F-statistic and p-value**: The F-statistic is used to test the overall significance of the model. It assesses whether the regression model as a whole is statistically significant. The p-value associated with the F-statistic indicates the probability of observing the F-statistic by chance if the null hypothesis (no relationship between the variables) is true. In this case, the F-statistic is 3765, and the p-value is extremely small (< 2.2e-16), suggesting strong evidence against the null hypothesis.

# **References**

* Best, J., & Kahn, J. V. (2016). Research in education (10th ed.). Pearson.
* Bethlehem, J. (2010). Selection Bias in Web Surveys. International Statistical Review, 78(2), 161–188.
* Bravais, A. (1846). Analyse mathématique sur les probabilités des erreurs de situation d'un point. Mallet-Bachelier.
* Bryman, A. (2012). Social Research Methods. Oxford University Press.
* Cho, E., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate?. Journal of Business Research, 66(9), 1261-1266.
* Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates.
* Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. Routledge.
* Dodge, Y. (2003). The Oxford Dictionary of Statistical Terms. Oxford University Press.
* Ellis, P. D. (2010). The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge University Press.
* Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics. SAGE.
* Frost, J. (2020). Correlation: Simple Definition, Formula, Easy Steps. Statistics How To. Retrieved from https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/
* Frost, J. (2020). Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models. Statistics By Jim. Retrieved from https://statisticsbyjim.com/regression/interpret-coefficients-pvalues-regression/
* Gilbert, D. (2015). The American Class Structure in an Age of Growing Inequality. SAGE Publications.
* Goodman, S. N. (1999). Toward evidence-based medical statistics. 1: The P value fallacy. Annals of internal medicine, 130(12), 995-1004.
* Kirk, R. E. (1996). Practical significance: A concept whose time has come. Educational and psychological measurement, 56(5), 746-759.
* Lavrakas, P. J. (2008). Encyclopedia of Survey Research Methods. SAGE Publications.
* Lohr, S. (2019). Sampling: Design and Analysis. Chapman and Hall/CRC.
* Menard, S. (2002). Longitudinal Research (2nd ed.). SAGE Publications.
* Pearl, J., & Mackenzie, D. (2018). The Book of Why: The New Science of Cause and Effect. Basic Books.
* Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 187, 253-318.
* Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. Anesthesia & Analgesia, 126(5), 1763-1768.
* Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—or Why the P Value Is Not Enough. Journal of Graduate Medical Education, 4(3), 279–282.
* Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). The psychology of survey response. Cambridge University Press.
* Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “p< 0.05”. The American Statistician, 73(sup1), 1-19.