**Assignment 4**

1. Comparison of Five Teaching Methods – (40 pts)

**A) Compute the pooled estimate of the standard deviation from these summary statistics.**

* Pooled Estimate of the Standard Deviation Calculation:

Given the formula for the pooled variance:

Taking the square root gives the pooled standard deviation:

The Pooled Estimate of the Standard Deviation is 4.483.

**B) Determine a set of coefficients that will contrast the methods using programmed text as part of the method with those that do not use programmed text.**

The aim is to compare methods including programmed text (Group 2: R, Group 3: R+L) against those without (Group 1: L+D, Group 4: C, Group 5: C+L).

We define the contrast *D* as:

With the following contrast coefficients:

*Group 1 = 1/3   
Group 2= (-1)/2  
Group 3 = (-1)/2  
Group 4 = 1/3  
Group 5 = 1/3*

​  
**C) Evaluate the comparison by calculating the g statistic using (b) and compute a 95% confidence interval.**

* Calculation of the g statistic:

Standard Error Calculation:

95% Confidence Interval Calculation:

With *t*40​(0.975) = 2.0221:

The 95% confidence interval for the contrast is (0.24, 5.76).

**2. Education and Future Income (60 pts)**

**A) Test for differences among the five population distributions**

|  |
| --- |
| > library(Sleuth3)  > data("ex0525")  > # ANOVA to test if mean incomes differ among education levels  > anova\_result <- aov(Income2005 ~ factor(Educ), data = ex0525)  > summary(anova\_result)  Df Sum Sq Mean Sq F value Pr(>F)  factor(Educ) 4 6.882e+11 1.721e+11 89.61 <2e-16 \*\*\*  Residuals 2579 4.952e+12 1.920e+09  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 |
|  |
| |  | | --- | | > | |

The ANOVA results show a very small p-value (< 2e-16), which provides strong evidence that at least one of the education categories has a different income distribution from the others.

**B) Multiple comparison procedure**

> # Multiple comparison using Tukey's HSD test

> TukeyHSD(anova\_result)

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = Income2005 ~ factor(Educ), data = ex0525)

$`factor(Educ)`

diff lwr upr p adj

13-15-12 8011.061 2002.372 14019.750 0.0025813

16-12 33132.077 26113.227 40150.927 0.0000000

<12-12 -8563.448 -19482.342 2355.447 0.2031235

>16-12 39990.566 32760.125 47221.008 0.0000000

16-13-15 25121.016 17550.263 32691.769 0.0000000

<12-13-15 -16574.508 -27856.100 -5292.917 0.0005961

>16-13-15 31979.506 24212.182 39746.830 0.0000000

<12-16 -41695.524 -53546.019 -29845.030 0.0000000

>16-16 6858.490 -1714.213 15431.193 0.1860668

>16-<12 48554.014 36576.984 60531.044 0.0000000

The Tukey HSD test identifies that every group significantly differs from at least one other group except the >16 years education group compared to the 16 years group (p = 0.1860668), indicating that income does not significantly differ between these two groups.

**C) Visual analysis and transformation selection**

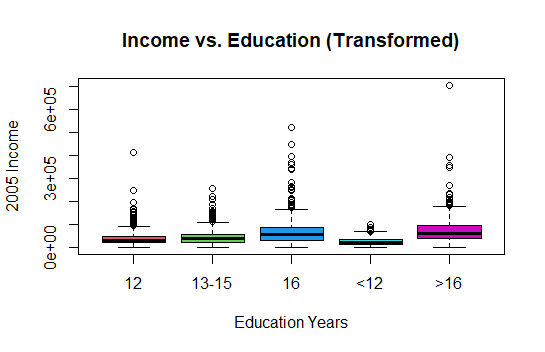
> # Visually analyze the data with a boxplot

> boxplot(Income2005 ~ factor(Educ), data = ex0525,

+ main = "Income vs. Education (Transformed)",

+ ylab = "2005 Income", xlab = "Education Years",

+ col = 2:8)



The spread of incomes seems wide, particularly for individuals with higher education. The distribution shows variability in income across different education levels with numerous outliers.

> #Based on skewness and spread, we perform a log transformation if appropriate

> ex0525$Income2005\_log <- log(ex0525$Income2005)

>

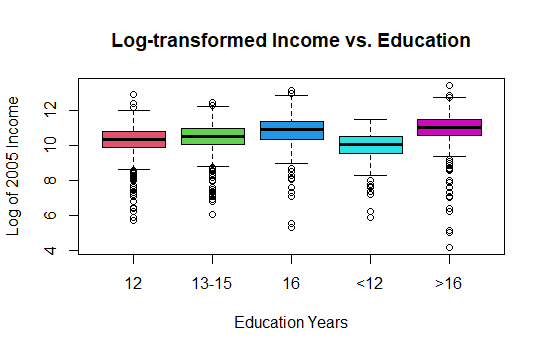
> # Boxplot to visualize the transformed data

> boxplot(Income2005\_log ~ factor(Educ), data = ex0525,

+ main = "Log-transformed Income vs. Education",

+ ylab = "Log of 2005 Income", xlab = "Education Years",

+ col = 2:8)



The log transformation has symmetrized the distribution, reducing the effect of outliers and making the spread within education groups more comparable.

**D) Calculate median income differences**

> # Calculate median incomes for each education level

> median\_incomes <- tapply(ex0525$Income2005, ex0525$Educ, median)

> median\_incomes

12 13-15 16 <12 >16

31000 38000 56500 23500 60500

> # Calculate differences between consecutive education levels

> median\_diffs <- diff(median\_incomes)

> median\_diffs

13-15 16 <12 >16

7000 18500 -33000 37000

> # If you need to calculate percentage differences

> median\_perc\_diffs <- median\_diffs / median\_incomes[1:length(median\_incomes)-1] \* 100

> median\_perc\_diffs

13-15 16 <12 >16

22.58065 48.68421 -58.40708 157.44681

* 13-15 years of education earns $7,000 more than 12 years (a 22.58% increase).
* 16 years of education earns $18,500 more than 13-15 years (a 48.68% increase).
* Individuals with <12 years of education earn $33,000 less than those with 16 years (a 58.41% decrease).
* Those with >16 years of education earn $37,000 more than those with <12 years (a 157.45% increase).

**E) Analyze for a linear trend**

> # Assign average years of education to each group

> ex0525$AvgYears <- factor(ex0525$Educ, labels = c(10, 12, 14, 16, 20))

> # Linear model to test for linear trend in income as a function of the number of years studied

> linear\_trend\_model <- lm(Income2005 ~ as.numeric(AvgYears), data = ex0525)

> summary(linear\_trend\_model)

Call:

lm(formula = Income2005 ~ as.numeric(AvgYears), data = ex0525)

Residuals:

Min 1Q Median 3Q Max

-73794 -24625 -8625 13259 629780

Coefficients:

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

(Intercept) 28566.6 1698.1 16.82 <2e-16 \*\*\*

as.numeric(AvgYears) 9058.1 629.7 14.39 <2e-16 \*\*\*

---

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

Residual standard error: 44970 on 2582 degrees of freedom

Multiple R-squared: 0.0742, Adjusted R-squared: 0.07384

F-statistic: 206.9 on 1 and 2582 DF, p-value: < 2.2e-16

The linear regression model shows a significant relationship between years of education and income (p-value < 2.2e-16), with an average increase of $9,058.1 in income for each additional year of education, as estimated by the model.