Stats-ps12-solution

Kirthivasan Pandurangan Neelavathi

2023-12-09

1. Assume that the population of all sister-brother heights has a bivariate normal distribution and that the data in Table 14.4 were sampled from this population.

```
sister_height <- c(69, 64, 65, 63, 65, 62, 65, 64, 66, 59, 62)
brother_height <- c(71, 68, 66, 67, 70, 71, 70, 73, 72, 65, 66)

# (a) Sample coefficient of determination
correlation_coefficient <- cor(sister_height, brother_height)
cat("Correlation coeffient (r)", correlation_coefficient, "\n")
```

```
## Correlation coeffient (r) 0.5580547
```

```
r_squared <- correlation_coefficient^2
cat("R_squared (R^2) =",r_squared)
```

```
## R_squared (R^2) = 0.3114251
```

```
# (b)
```

Therefore we can say that, there is approximately 31.14% of variation in brother's and sister's height

1b:

Indicates Hypotheses test

alpha = 0.05

Null Hypothesis: Indicates that there is no relationship between sister's height (x) and brother's height (y)

Alternative Hypothesis: There is a significant relationship between the height of sisters and brothers, knowing a sister's height (x) allows for the prediction of her brother's height (y)

```
linear_model <- lm(brother_height ~ sister_height)
summary(linear_model)
```

```
##
## Call:
## lm(formula = brother_height ~ sister_height)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5909 -1.2273 -0.9545  1.1136  4.0000
```

```
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    31.1818    18.7584   1.662   0.1308
## sister_height   0.5909     0.2929   2.018   0.0744 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.379 on 9 degrees of freedom
## Multiple R-squared:  0.3114, Adjusted R-squared:  0.2349
## F-statistic:  4.07 on 1 and 9 DF,  p-value: 0.07442
```

From the data, p-val is < 0.05 so, we fail to reject null hypotheses which indicates that data doesn't provide convincing evidence that knowing a sister's height (x) helps one predict her brother's height (y).

(c)

```r
# Check the structure of the linear model summary
str(summary(linear_model)$coefficients)
```

```
##  num [1:2, 1:4] 31.182 0.591 18.758 0.293 1.662 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:2] "(Intercept)" "sister_height"
##   ..$ : chr [1:4] "Estimate" "Std. Error" "t value" "Pr(>|t|)"
```

```r
# Assuming "sister" is the correct column name in your data frame
# Replace it with the actual column name if it's different
se_slope <- summary(linear_model)$coefficients["sister_height", "Std. Error"]
df <- length(brother_height) - 2
critical_value <- qt(0.95, df)
margin_of_error <- critical_value * se_slope
confidence_interval <- coef(linear_model)["sister_height"] + c(-margin_of_error, margin_of_error)
cat("90% Confidence Interval for the Slope:", round(confidence_interval, 4), "\n")
```

```
## 90% Confidence Interval for the Slope: 0.054 1.1278
```

Question 2

```r
exam_anxiety_data <- data.frame(read.table("/Users/kirthipandu/Documents/Intro_to_stats/assignment/ps-1
exam_anxiety_data
```

```
##     Code Revise Exam Anxiety Gender
## 1      1      4   40  86.298   Male
## 2      2     11   65  88.716 Female
## 3      3     27   80  70.178   Male
## 4      4     53   80  61.312   Male
## 5      5      4   40  89.522   Male
## 6      6     22   70  60.506 Female
## 7      7     16   20  81.462 Female
```

```
## 8       8      21   55  75.820 Female
## 9       9      25   50  69.372 Female
## 10     10      18   40  82.268 Female
## 11     11      18   45  79.044   Male
## 12     12      16   85  80.656   Male
## 13     13      13   70  70.178   Male
## 14     14      18   50  75.014 Female
## 15     15      98   95  34.714   Male
## 16     16       1   70  95.164   Male
## 17     17      14   95  75.820   Male
## 18     18      29   95  79.044 Female
## 19     19       4   50  91.134 Female
## 20     20      23   60  64.536   Male
## 21     21      14   80  80.656   Male
## 22     22      12   75  77.432   Male
## 23     23      22   85  65.342 Female
## 24     24      84   90  56.116 Female
## 25     25      23   30  71.790 Female
## 26     26      26   60  81.462 Female
## 27     27      24   75  63.730   Male
## 28     28      72   75  27.460 Female
## 29     29      37   27  73.402 Female
## 30     30      10   20  89.522   Male
## 31     31       3   75  89.522 Female
## 32     32      36   90  75.014 Female
## 33     33      43   60  43.580   Male
## 34     34      19   30  82.268   Male
## 35     35      12   80  79.044   Male
## 36     36       9   10  79.044 Female
## 37     37      72   85  37.132   Male
## 38     38      10    7  81.462   Male
## 39     39      12    5  83.074 Female
## 40     40      30   85  50.834   Male
## 41     41      15   20  82.268   Male
## 42     42       8   45  78.238 Female
## 43     43      34   60  72.596   Male
## 44     44      22   70  74.208 Female
## 45     45      21   50  75.820 Female
## 46     46      27   25  70.984   Male
## 47     47       6   50  97.582   Male
## 48     48      18   40  67.760   Male
## 49     49       8   80  75.014   Male
## 50     50      19   50  73.402 Female
## 51     51       0   35  93.552 Female
## 52     52      52   80  58.894 Female
## 53     53      38   50  53.252 Female
## 54     54      19   49  84.686   Male
## 55     55      23   75  89.522 Female
## 56     56      11   25  71.790 Female
## 57     57      27   65  82.268   Male
## 58     58      17   80  69.372   Male
## 59     59      13   50  62.118   Male
## 60     60      42   70  68.566 Female
## 61     61       4   40  93.552   Male
```

```
## 62    62      8   80  84.686 Female
## 63    63      6   10  82.268   Male
## 64    64     11   20  81.462 Female
## 65    65      7   40  82.268   Male
## 66    66     15   40  91.134   Male
## 67    67      4   70  91.940 Female
## 68    68     28   52  86.298 Female
## 69    69     22   50  72.596   Male
## 70    70     29   60  63.730 Female
## 71    71      2   80  63.730   Male
## 72    72     16   60  71.790 Female
## 73    73     59   65  57.282   Male
## 74    74     10   15  84.686 Female
## 75    75     13   85  84.686   Male
## 76    76      8   20  77.432 Female
## 77    77      5   80  82.268 Female
## 78    78      2  100  10.000   Male
## 79    79     38  100  50.834 Female
## 80    80      4   80  87.910   Male
## 81    81     10   10  83.880   Male
## 82    82      6   70  84.686 Female
## 83    83     68  100  20.206 Female
## 84    84      8   70  87.104   Male
## 85    85      1   70  83.880 Female
## 86    86     14   65  67.760   Male
## 87    87     42   75  95.970 Female
## 88    88     13   85  62.118 Female
## 89    89      1   30  84.686   Male
## 90    90      3    5  92.746   Male
## 91    91      5   10  84.686 Female
## 92    92     12   90  83.074 Female
## 93    93     19   70  73.402   Male
## 94    94      2   20  87.910 Female
## 95    95     19   85  71.790   Male
## 96    96     11   35  86.298   Male
## 97    97     15   30  84.686 Female
## 98    98     23   70  75.820   Male
## 99    99     13   55  70.984 Female
## 100  100     14   75  78.238 Female
## 101  101      1    2  82.268   Male
## 102  102      9   40  79.044   Male
## 103  103     20   50  91.134 Female
```

2. (a) Is there a significant difference between average anxiety for the population of male students and the population of female students? Perform an appropriate significance test, stating hypotheses, a P-value, and a substantive conclusion.

h0 <- Indicates no difference in the average anxiety between male and female students

h1 <- Significant difference in the average anxiety between male and female students

T-test is chosen because of two independent groups, continuous outcome, normality assumption, homogeneity of variances.

```r
male_anxiety <- exam_anxiety_data[exam_anxiety_data$Gender == "Male", "Anxiety"]
female_anxiety <- exam_anxiety_data[exam_anxiety_data$Gender == "Female", "Anxiety"]

t_test_result <- t.test(male_anxiety, female_anxiety)
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  male_anxiety and female_anxiety
## t = -0.32961, df = 100.41, p-value = 0.7424
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.147444  5.110827
## sample estimates:
## mean of x mean of y
##  74.38373  75.40204
```

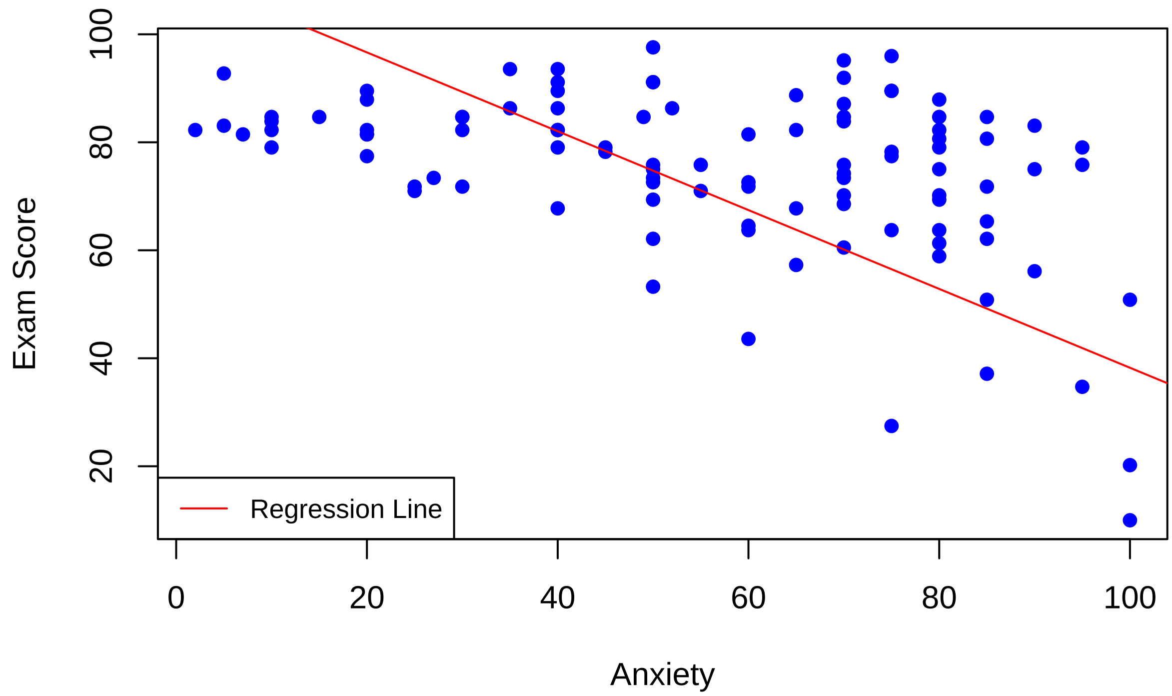
conclusion: The p-value of 0.7424 suggests to fail to reject the null hypothesis.

(b) Let anxiety be your x-variable and exam score be your y-variable. Find the regression line to predict exam score from anxiety. Carefully explain (in words or using math) what your regression line means — do not just paste R output

```r
# Assuming "Exam" and "Anxiety" are column names in your data frame exam_anxiety_data
regression_model <- lm(Exam ~ Anxiety, data = exam_anxiety_data)

# Create a scatter plot with regression line
plot(Anxiety ~ Exam, data = exam_anxiety_data,
     xlab = "Anxiety", ylab = "Exam Score", pch = 16, col = "blue")
abline(regression_model, col = "red")

# Add legend
legend("bottomleft", legend = "Regression Line", col = "red", lty = 1, cex = 0.8)
```




```r
# Display regression summary
summary(regression_model)
```

```
##
## Call:
## lm(formula = Exam ~ Anxiety, data = exam_anxiety_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.185 -16.046   1.166  19.856  41.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 111.2444    11.3498   9.801 2.46e-16 ***
## Anxiety      -0.7300     0.1484  -4.920 3.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.41 on 101 degrees of freedom
## Multiple R-squared:  0.1933, Adjusted R-squared:  0.1853
## F-statistic:  24.2 on 1 and 101 DF,  p-value: 3.374e-06
```

```r
b0 <- 111.2444
b1 <- -0.7300
```

```r
anxiety_level <- 20
pred_exam_score <- b0 + b1 * anxiety_level
cat("anxiety_level: ", anxiety_level, "\t", "predicted_exam_score: ", pred_exam_score, "\n")
```

```
## anxiety_level:  20    predicted_exam_score:  96.6444
```

```r
anxiety_level <- 90
pred_exam_score <- b0 + b1 * anxiety_level
cat("anxiety_level: ", anxiety_level, "\t", "predicted_exam_score: ", pred_exam_score, "\n")
```

```
## anxiety_level:  90    predicted_exam_score:  45.5444
```

```r
cat("Conclusion: The higher the anxiety, the lower the exam score. \n \n")
```

```
## Conclusion: The higher the anxiety, the lower the exam score.
##
```

```r
anxiety.lm = lm(Exam ~ Anxiety, data = exam_anxiety_data)
summary(anxiety.lm)
```

```
##
## Call:
## lm(formula = Exam ~ Anxiety, data = exam_anxiety_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.185 -16.046   1.166  19.856  41.461
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 111.2444    11.3498   9.801 2.46e-16 ***
## Anxiety      -0.7300     0.1484  -4.920 3.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.41 on 101 degrees of freedom
## Multiple R-squared:  0.1933, Adjusted R-squared:  0.1853
## F-statistic:  24.2 on 1 and 101 DF,  p-value: 3.374e-06
```

The obtained p-value of 0.7424 we fail to reject the null hypothesis. Consequently,

This indicates that p-value aligns with the idea that there is no significant difference in the average anxiety levels between female and male students based on the available data.

However, it's essential to recognize the limitations of drawing broader conclusions from this analysis alone.

The current dataset may not be large enough or representative enough to confidently generalize the relationship between the average anxiety of the two groups. To arrive at a more robust conclusion, a substantially larger and more representative sample would be required.

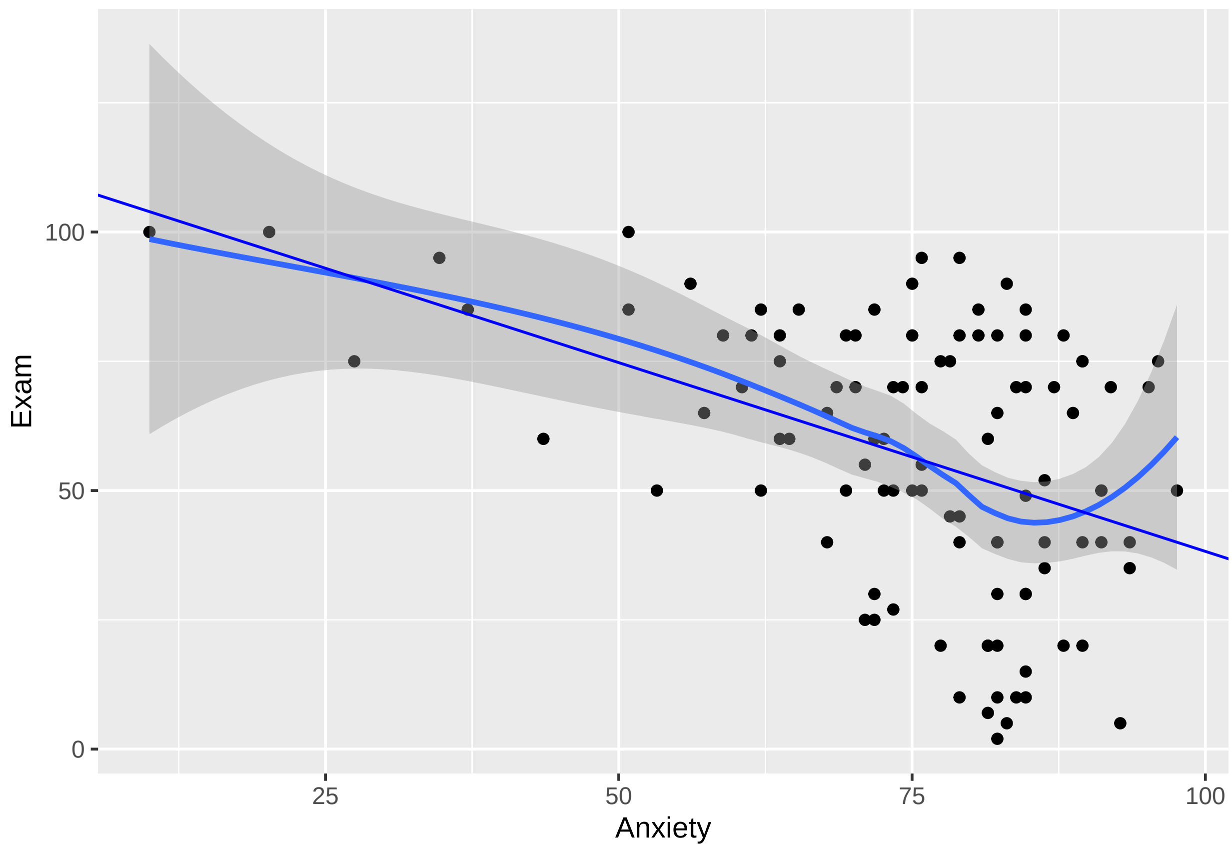
Moreover, greater uncertainty is introduced by the lack of information indicating whether the findings are the result of a controlled experiment. To reach a more solid and trustworthy conclusion regarding the correlation between the average anxiety levels of male and female students, a controlled trial would be better.

2. (c)

i. Linearity Assessment:

```
library(vctrs)
library(broom)
library(ggplot2)
ggplot(exam_anxiety_data, aes(x = Anxiety, y = Exam)) + geom_point() +
geom_smooth() +
geom_abline(intercept = 111.24, slope = -0.73, color = "blue")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



To test the linearity assumption, a scatter plot is made with the residuals on the y-axis and anxiety on the x-axis.

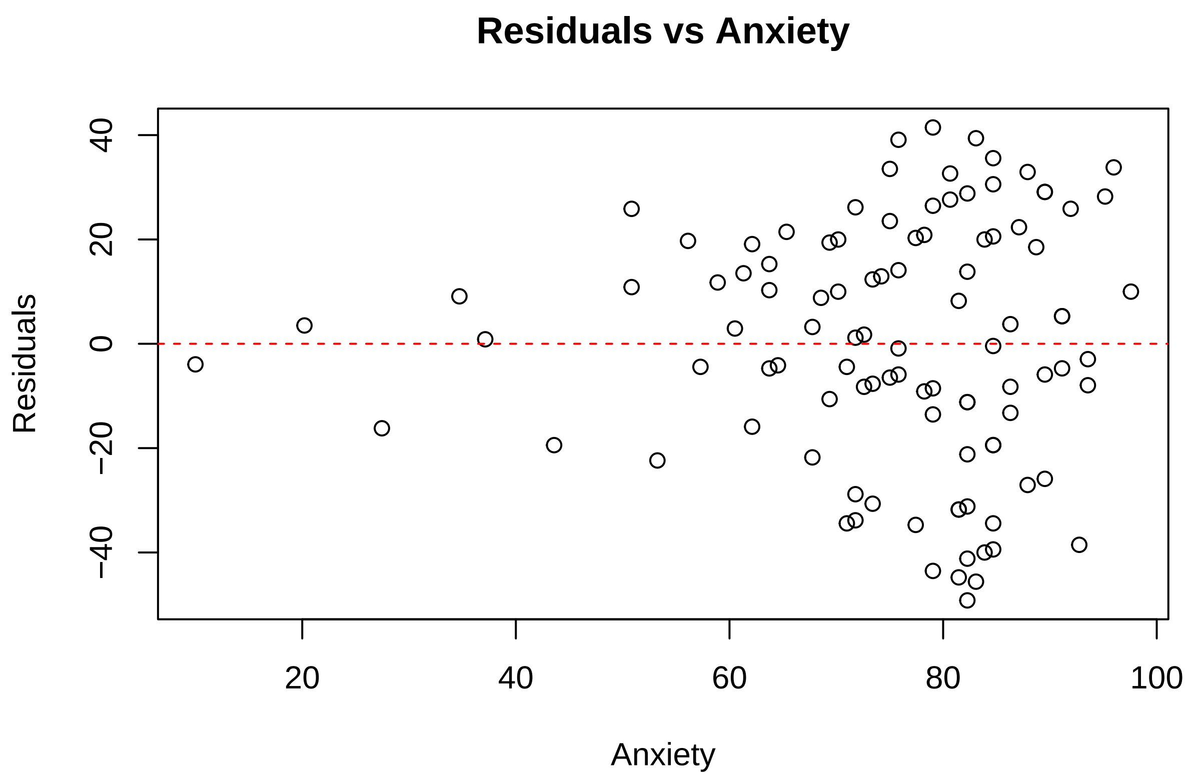
Relatives are a statistical measure of the differences between the observed exam outcomes and the predicted exam results derived from the regression line.

With this graphical depiction, we can visually assess whether the link between Anxiety and the residuals appears random and doesn't indicate any clear pattern.

A lack of pattern in the scatter plot would support the linearity assumption and show that the linear regression model is suitable for displaying the association between exam scores and anxiety.

```r
# ii. Independence


plot(exam_anxiety_data$Anxiety, residuals(regression_model),
     main = "Residuals vs Anxiety", xlab = "Anxiety", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
```

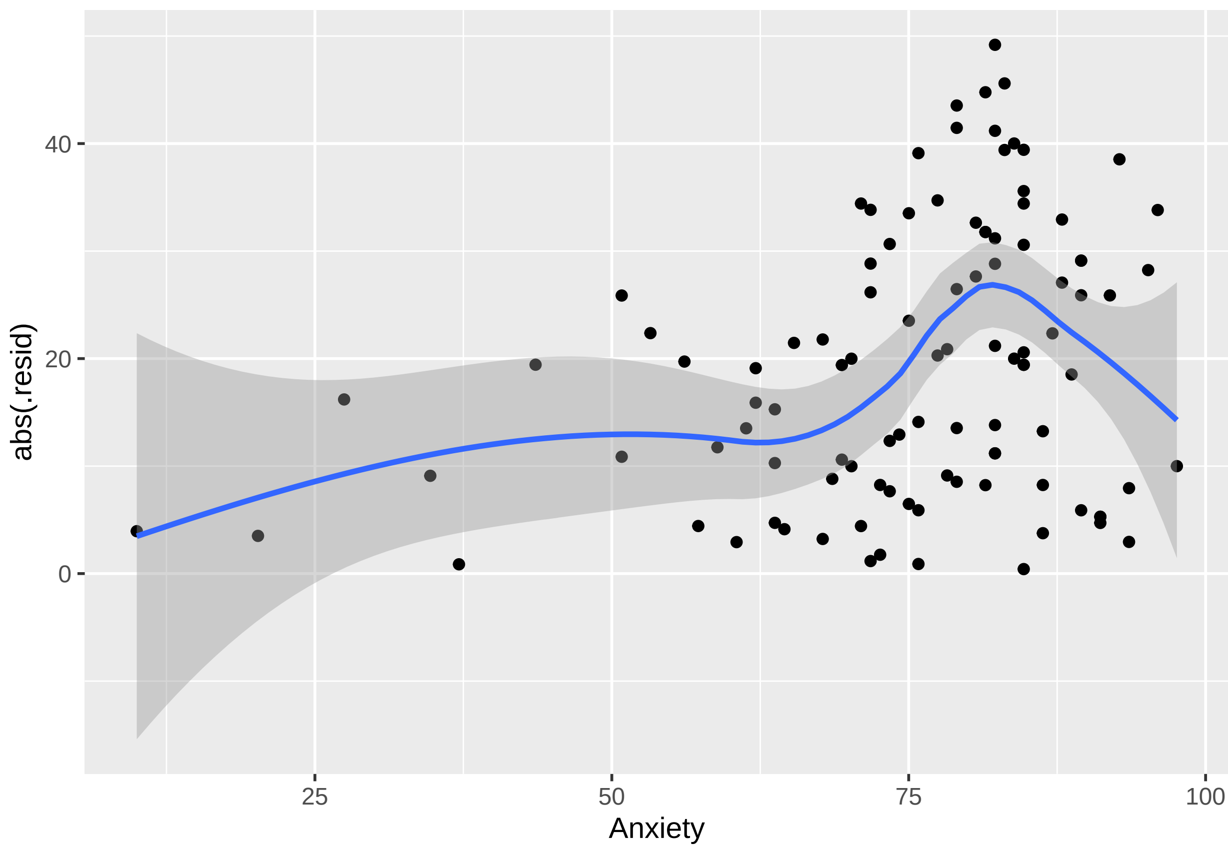



```r
# Since the red line is roughly horizontal and the points are randomly scattered, it suggests independe
```

The procedure used to collect the data affects how independent errors are. The data was taken at random from a larger population as stated, confirming that the assumption of independence is satisfied.

```r
library(broom)
anxiety.lm.df <- augment(anxiety.lm)
ggplot(anxiety.lm.df, aes(x = Anxiety, y = abs(.resid) )) + geom_point() +
geom_smooth()
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



A scatter plot is created with the absolute values of the residuals on the y-axis and anxiety on the x-axis in order to assess homoskedasticity. It is easier to determine whether the residual spread stays mostly constant across various anxiety levels with the use of this graphical representation. Homoskedasticity, or the assumption that the variance of errors stays constant across the range of anxiety values, is supported by a consistent spread.

iv. Normality of errors

A quantile-quantile (qq) plot is created in order to examine the assumption of normality in errors. A distinct curve that deviates from the straight line that would suggest a normal distribution of errors is visible in the qq plot. Consequently, we deduce that the normal distribution of errors assumption is not satisfied. This suggests that when drawing conclusions from the data, other approaches besides those predicated on the assumption of normalcy should be taken into account.

Question 3a:

The analysis of variance (ANOVA) F-test several assumptions based on the follows:

Normality: The sample sizes are relatively large (35 rats per group), so the normality assumption level of significance is less. This assumption can be taken under consideration and but not for large samples due to central limit theorem.

Homogeneity of Variances: Homogeneity of variances involves examining the spread of residuals for each group. Other better ways of illustrating is by utilizing residual plots or statistical tests.

Independence and Random Sampling: These assumptions are based on If the rats were randomly assigned to the different feed groups and each rat was kept in a separate cage, these assumptions are likely met.

Question 3b:

```r
h0 <- "The means of all groups are equal."
cat("Null Hypothesis (H0):", h0)
```

```
## Null Hypothesis (H0): The means of all groups are equal.
```

```r
h1 <- "At least one of the rat feeds has a different effect on weight gain compared to the others."
cat("Alternative Hypothesis (Ha):", h1)
```

```
## Alternative Hypothesis (Ha): At least one of the rat feeds has a different effect on weight gain comp
```

```r
# Sample means and standard deviations
means <- c(83.5, 92.3, 88.6, 99.4)
stds <- c(16.9, 14.6, 14.2, 14.1)

# Number of rats in each group
n_per_group <- 35

# Calculate total number of rats
n_total <- n_per_group * length(means)

# Calculate grand mean
grand_mean <- mean(means)

# Calculate between-group sum of squares
ss_between <- n_per_group * sum((means - grand_mean)^2)

# Calculate within-group sum of squares
ss_within <- sum((n_per_group - 1) * stds^2)

# Calculate total sum of squares
ss_total <- sum((means - grand_mean)^2) * n_per_group

# Calculate degrees of freedom
df_between <- length(means) - 1
df_within <- n_total - length(means)
df_total <- n_total - 1

# Calculate mean squares
ms_between <- ss_between / df_between
ms_within <- ss_within / df_within

# Calculate F-statistic
f_statistic <- ms_between / ms_within

# Calculate p-value
p_value <- pf(f_statistic, df_between, df_within, lower.tail = FALSE)

# Create ANOVA table
anova_table <- data.frame(
  Variation = c("Between", "Within", "Total"),
  `Sum of squares` = c(ss_between, ss_within, ss_total),
  DF = c(df_between, df_within, df_total),
```

```r
  `Mean square` = c(ms_between, ms_within, NA),
  F = c(f_statistic, NA, NA),
  `P-value` = c(p_value, NA, NA)
)

# Print ANOVA table
print(anova_table)
```

```
##   Variation Sum.of.squares  DF Mean.square        F      P.value
## 1   Between        4698.75   3    1566.250 6.967149 0.0002140835
## 2    Within       30573.48 136     224.805       NA           NA
## 3     Total        4698.75 139          NA       NA           NA
```

```r
cat("The p-value is less than your chosen significance level (0.05) and hence rejecting the null hypothe
```

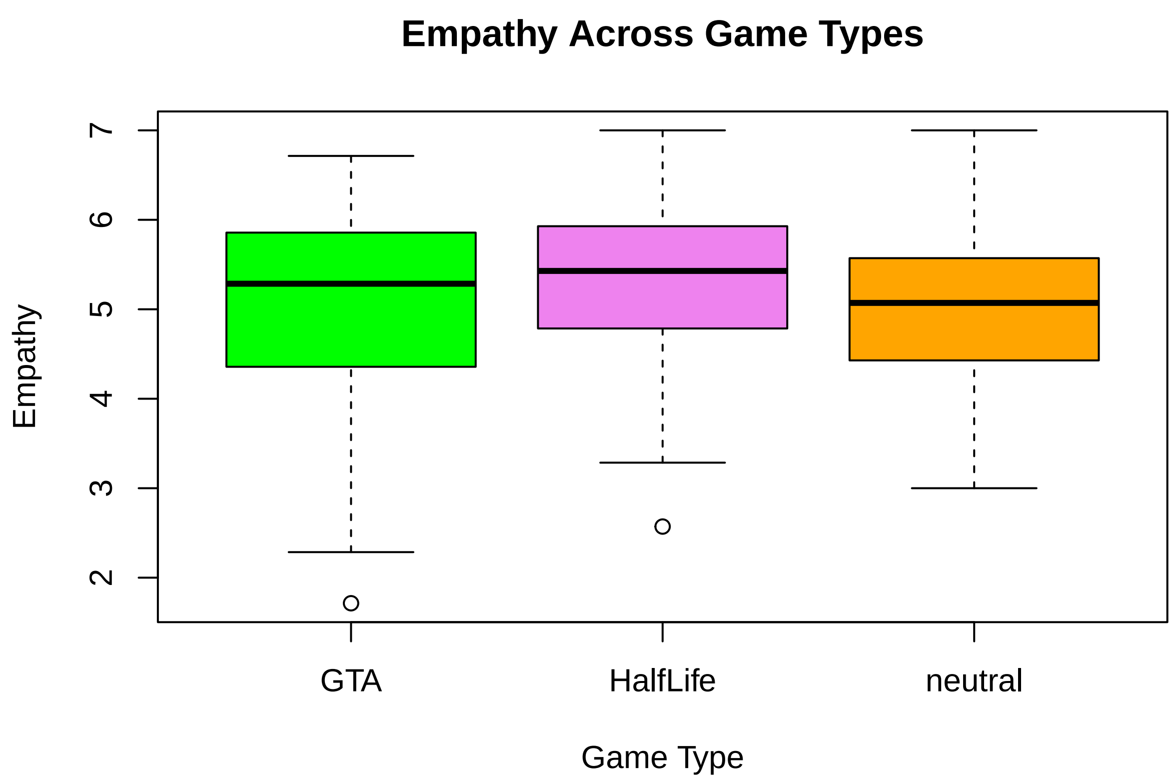
```
## The p-value is less than your chosen significance level (0.05) and hence rejecting the null hypothesi
```

```r
cat("Therfore,", h1)
```

```
## Therfore, At least one of the rat feeds has a different effect on weight gain compared to the others
```

Question 4

```r
game_data <- read.table("/Users/kirthipandu/Documents/Intro_to_stats/assignment/ps-12/gameEmpathy.txt",s
boxplot(empathy ~ game.type, data = game_data, col = c("green", "violet", "orange"), main = "Empathy Acr
```



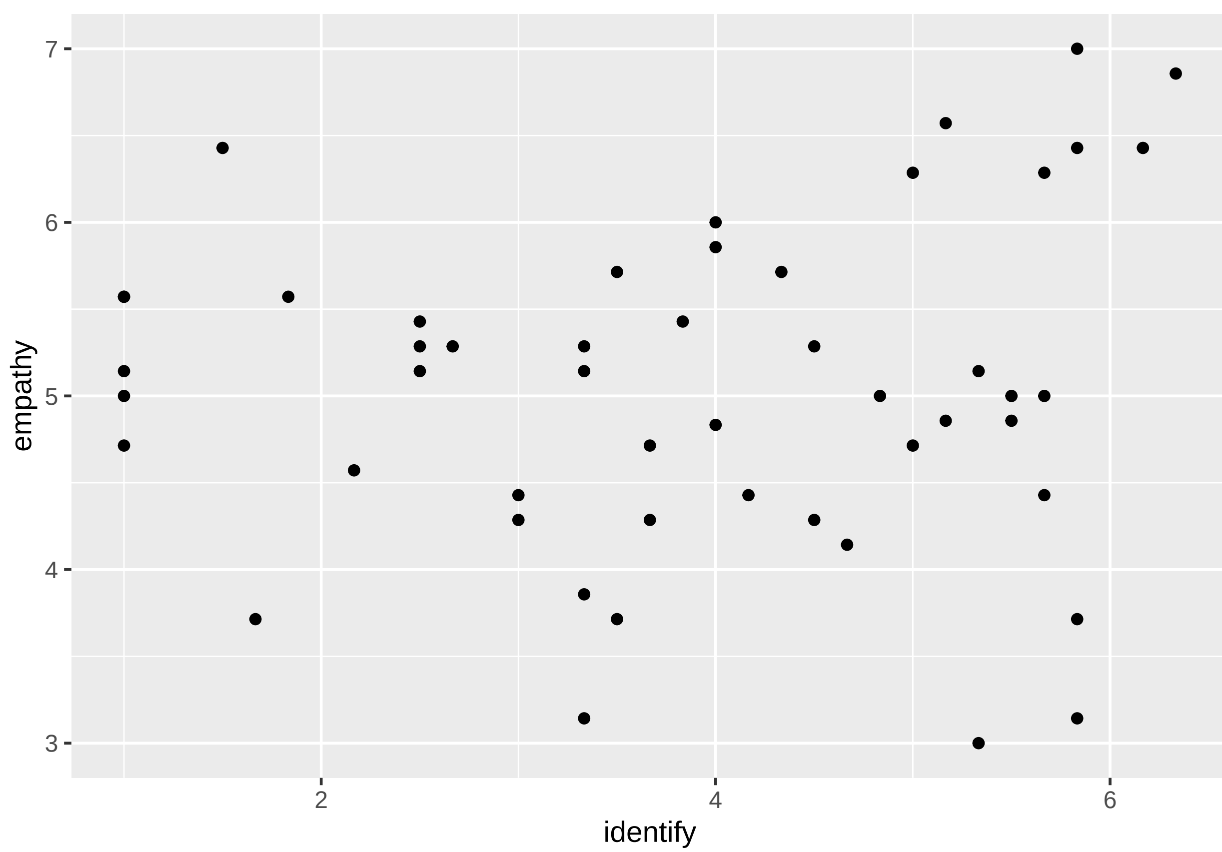
h0 <- "There is no significant difference in average empathy scores among the game types. h1 <-"There are significant differences in average empathy scores among the game types.

```
anova_result <- aov(empathy ~ game.type, data = game_data)
summary(anova_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## game.type     2   2.25   1.125   1.092  0.338
## Residuals   150 154.47   1.030
```

Q4 b)

```
if (!requireNamespace("ggplot2", quietly = TRUE)) {
  install.packages("ggplot2")
}
library(ggplot2)
neutral.players <- subset(game_data, game.type == "neutral")
ggplot(neutral.players, aes(x = identify, y = empathy)) + geom_point()
```



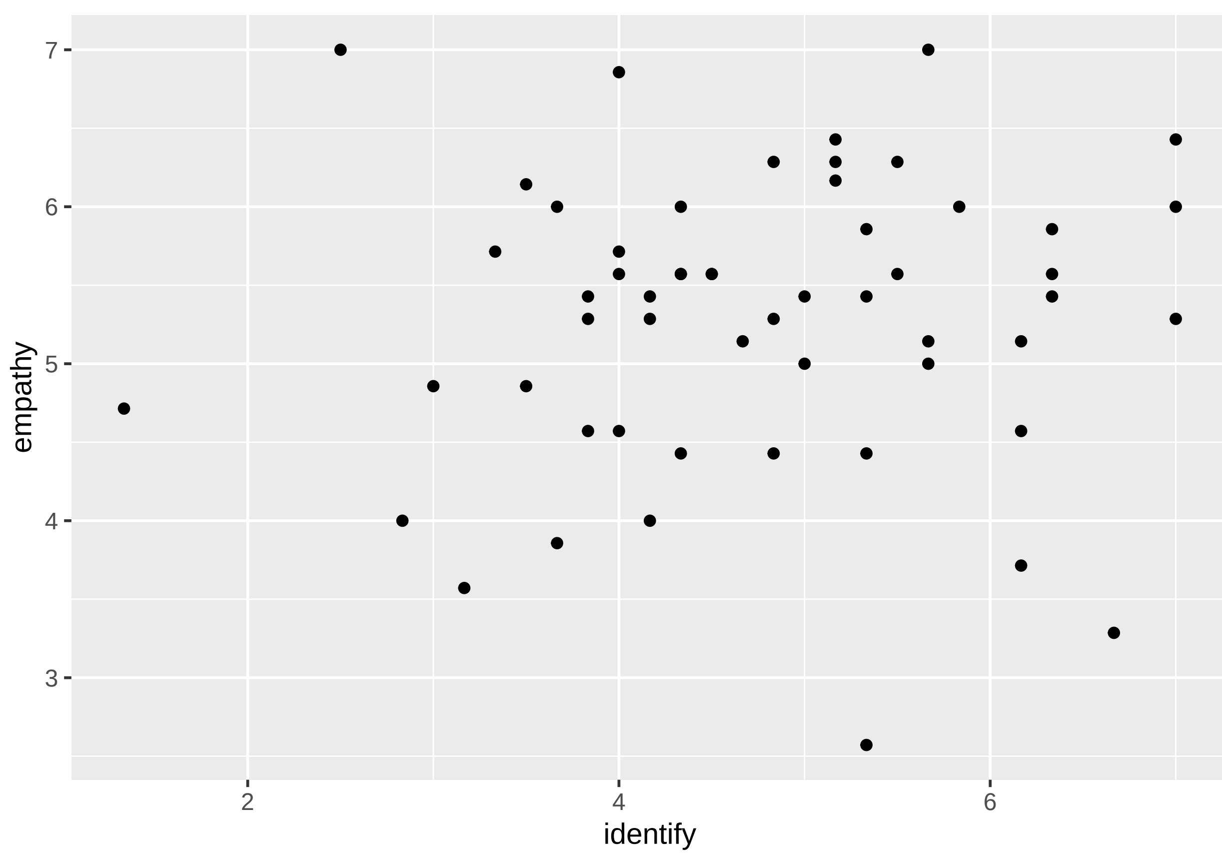

```
neutral.lm <- lm(empathy ~ identify, data = neutral.players)
summary(neutral.lm)
```

```
##
## Call:
```

```
## lm(formula = empathy ~ identify, data = neutral.players)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13273 -0.62533  0.05582  0.66360  1.84025
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.84445    0.36120  13.412   <2e-16 ***
## identify     0.05405    0.08641   0.626    0.535
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9602 on 48 degrees of freedom
## Multiple R-squared:  0.008085,   Adjusted R-squared:  -0.01258
## F-statistic: 0.3913 on 1 and 48 DF,  p-value: 0.5346
```

Y = 4.84+0.0541*X, where 4.84 represents the intercept, and 0.0541 is the slope

```r
halflife.players <- subset(game_data, game.type == "HalfLife")
ggplot(halflife.players, aes(x = identify, y = empathy)) + geom_point()
```



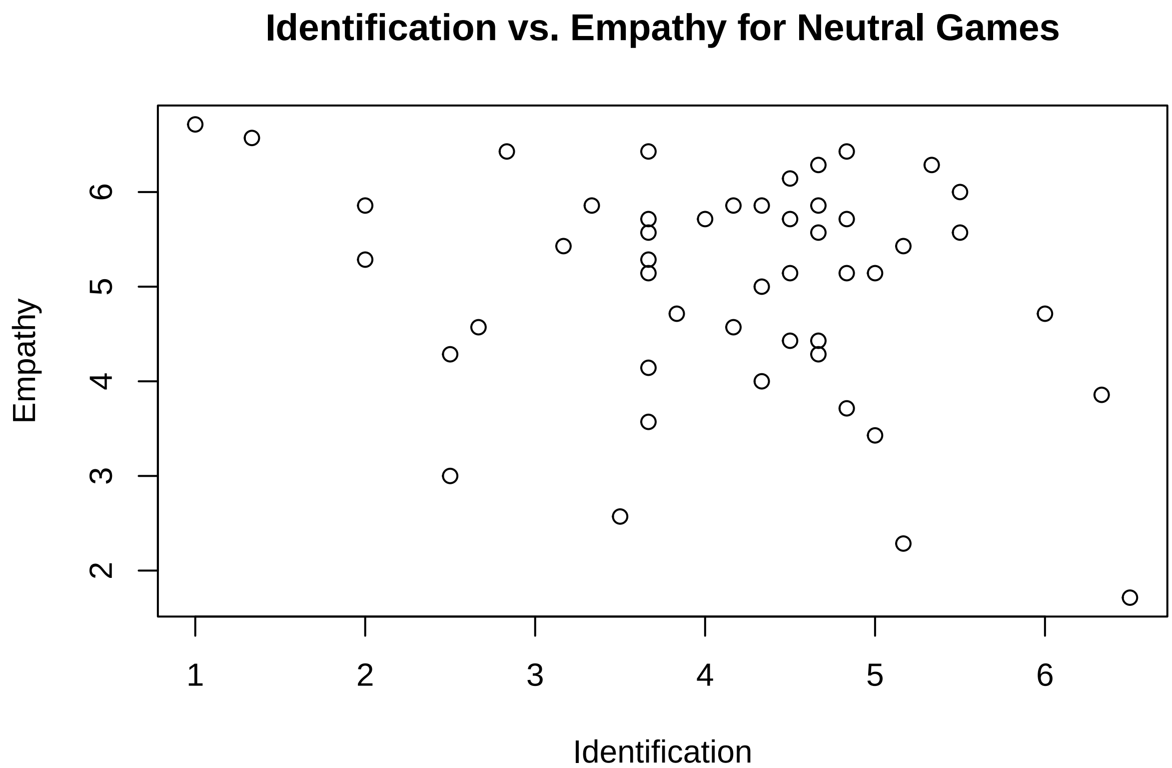

```r
halflife.lm <- lm(empathy ~ identify, data = halflife.players)
summary(halflife.lm)
```

```
##
## Call:
## lm(formula = empathy ~ identify, data = halflife.players)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7538 -0.3804  0.1216  0.5579  1.8294
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.03427    0.51213   9.830 1.53e-13 ***
## identify     0.05455    0.10431   0.523    0.603
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.929 on 53 degrees of freedom
## Multiple R-squared:  0.005133,   Adjusted R-squared:  -0.01364
## F-statistic: 0.2734 on 1 and 53 DF,  p-value: 0.6032
```

Y = 5.03 + 0.0546*X, where 5.03 is the intercept and 0.0546 is the slope.

```r
gta_players <- subset(game_data, game.type == "GTA")

plot(gta_players$identify,
     gta_players$empathy,
     main = "Identification vs. Empathy for Neutral Games",
     xlab = "Identification", ylab = "Empathy")
```




```r
cor(gta_players$identify, gta_players$empathy)
```

```
## [1] -0.2722745
```

Y = 6.13 - 0.267*X, where 6.13 is the intercept and -0.267 is the slope.