Assessment Template: Statistical Inference

Masoud Rayat Zadeh

Apr 17th 2024

Instructions to students

Save this template as your studentID.Rmd; you will upload this file as part of your submission. Change the author information on line 3 of this file to your **student ID**. Do not change the authorship to your name.

You should knit this file to a document **Word** format. The Word document is what will be marked!

Any changes that you make to the data (e.g. variable name changes) should be made entirely within R.

The subsections labelled **Answer:** indicate where you should put in your written Answers. The template also provides blank code chunks for you to complete your Answers; you may choose to add additional chunks if required.

This is an individual assessment: do not work with any other person during this exam. Text-matching software will be used on all submissions.

Instructions for submission

You must submit your assignment before the stated deadline by electronic submission through Blackboard.

- It is a good idea to save your work early and frequently to ensure you have no issues with the submission portal. Multiple submissions can be made to the portal, but only the final one will be marked.
- It is your responsibility to submit the exam in a format stipulated above. Your marks may be affected if your tutor cannot open or properly view your submission.
- Do not leave submission to the very last minute. Always allow time in case of technical issues.
- The date and time of your submission is taken from the Blackboard server and is recorded when your submission is complete, not when you click Submit.

- It is essential that you check that you have submitted the correct file, and that each complete file was received. Submission receipts are accessed from the Coursework tab.

There is no late submission permitted on this timed assessment. Ensure that you submit in good time. Neither the module leader nor module team can accept late assessments, do not ask them to do so.

Background to the research

Wales is part of the United Kingdom. There has been a growing number of supporters in recent years for the idea that Wales leaves the United Kingdom and becomes an independent country.

The client is the Welsh Government who are interested in what impacts current residents' likelihood of supporting Welsh Independence.

Data has been obtained from a survey of residents, where respondents were asked a series of questions about themselves, and also to take an assessment of their Welsh language reading, writing, speaking and listening abilities.

Data instructions

Your individual data set is accessed via Blackboard >>> Assessments >>> Dewis Data For Exam.

You must only analyse the specified data. No other data is to be used for this assessment.

All data manipulation and analyses must be done within R.

Data structure

The variables collected for a sample of residents are:

resident - If respondent is 'Current' or 'Former' resident of Wales

area - If respondent is/was resident of 'North' Wales or 'South' Wales

support - Self-assessed level of support for Welsh independence, selected using a sliding scale (higher values mean stronger support for Welsh independence)

friends - Estimated number of friends or family members who support Welsh Independence

year - Year of birth

gender - Gender

read - Welsh language Reading ability score

write - Welsh language Writing ability score

speak - Welsh language Speaking ability score

listen - Welsh language Listening ability score

QUESTIONS START HERE

Question 1: Data Preparation

a) Ensure you have prepared your knitted Word document as per Instructions to Students

b) You should load the data in R, describe and perform any actions with respect to:

-any manipulation of the data structure

-ensuring data is valid and that only views of **current** residents are explored in this research.

(7 marks)

Answer:

I loaded the dataset from the specified file path into a data frame named wales_survey_data. To achieve this, I used the read.csv function in R, which reads a file in CSV format and creates a data frame. To get an initial understanding of the dataset, I displayed the first few rows. This aids in confirming the correct loading of the data and offers a preview of the dataset's structure and contents. For this, I used the head function. I successfully loaded the dataset into the R environment and displayed its first few rows by executing the above commands. This crucial initial exploration step ensures the correct import of the data and enables a preliminary examination of its structure and contents.

I made, cleaned, and filtered the dataset to prepare it for further analysis. Using the str function, I analyzed the dataset's structure to provide a detailed summary of the data frame. Missing and unknown values in the 'area' column were counted using the is.na function and "Unknown" values. The dataset was cleaned by removing rows with missing values, "Unknown", or "Missing" to ensure only valid entries were retained. I then filtered the cleaned dataset to include only current residents, ensuring the analysis focused on the relevant subset of the population. I inspected the resulting dataset to verify changes and understand its new composition. Finally, I used the summary function to generate a summary of the cleaned and filtered dataset, which provided a comprehensive overview of the dataset's statistics, including measures like mean, median, and range for numeric variables and counts for categorical variables.

```r
# Load the dataset
wales_survey_data <- read.csv("D:/welsh-independence-support-
analysis/data/contents.csv")

# Display the first few rows of the dataset
head(wales_survey_data)
```

```
  resident  area year gender support friends read write speak listen
1   Former South 1992   Male    2.00       3   77    56    27     78
2   Former South 1989 Female    2.10       4   75    55    25     75
3  Current South 1995 Female    2.19       8   75    54    30     77
4  Current South 1988   Male    2.24       6   73    53    30     75
5  Current South 1990 Female    2.32       8   72    53    39     74
6  Current South 1987   Male    2.41       1   73    54    21     74
```

```r
# Inspect the structure of the dataset
str(wales_survey_data)
```

```
'data.frame':   471 obs. of  10 variables:
 $ resident: chr  "Former" "Former" "Current" "Current" ...
 $ area    : chr  "South" "South" "South" "South" ...
 $ year    : int  1992 1989 1995 1988 1990 1987 1991 1989 1987 1991 ...
 $ gender  : chr  "Male" "Female" "Female" "Male" ...
 $ support : num  2 2.1 2.19 2.24 2.32 2.41 2.42 2.62 2.65 2.73 ...
 $ friends : int  3 4 8 6 8 1 7 7 0 4 ...
 $ read    : int  77 75 75 73 72 73 74 77 76 77 ...
 $ write   : int  56 55 54 53 53 54 55 56 54 55 ...
 $ speak   : int  27 25 30 30 39 21 29 32 34 23 ...
 $ listen  : int  78 75 77 75 74 74 78 80 76 78 ...
```

```r
# Count missing values
missing_count <- sum(is.na(wales_survey_data$area))

# Count unknown values
unknown_count <- sum(wales_survey_data$area == "Unknown", na.rm = TRUE)

# Remove rows with missing, unknown, or Missing values in the 'area' column
cleaned_data <- wales_survey_data %>%
  filter(!is.na(area) & area != "Unknown" & area != "Missing")

# Filter to include only current residents
current_residents <- cleaned_data %>% filter(resident == "Current")

# Display the structure of the filtered dataset
str(current_residents)
```

```
'data.frame':   406 obs. of  10 variables:
 $ resident: chr  "Current" "Current" "Current" "Current" ...
 $ area    : chr  "South" "South" "South" "South" ...
 $ year    : int  1995 1988 1990 1987 1991 1987 1991 1989 1994 1995 ...
```

```
 $ gender  : chr  "Female" "Male" "Female" "Male" ...
 $ support : num  2.19 2.24 2.32 2.41 2.42 2.65 2.73 2.76 2.77 2.83 ...
 $ friends : int  8 6 8 1 7 0 4 7 8 12 ...
 $ read    : int  75 73 72 73 74 76 77 73 77 75 ...
 $ write   : int  54 53 53 54 55 54 55 54 57 56 ...
 $ speak   : int  30 30 39 21 29 34 23 36 37 29 ...
 $ listen  : int  77 75 74 74 78 76 78 78 79 76 ...
```

```
# Summary of the cleaned dataset
summary(current_residents)
```

```
  resident             area                year         gender
 Length:406         Length:406         Min.   :1986   Length:406
 Class :character   Class :character   1st Qu.:1990   Class :character
 Mode  :character   Mode  :character   Median :1993   Mode  :character
                                       Mean   :1993
                                       3rd Qu.:1995
                                       Max.   :2000
    support         friends            read           write
 Min.   :2.190   Min.   : 0.000   Min.   :69.00   Min.   :51.00
 1st Qu.:4.072   1st Qu.: 3.000   1st Qu.:75.00   1st Qu.:55.00
 Median :4.865   Median : 6.000   Median :76.00   Median :57.00
 Mean   :4.874   Mean   : 5.783   Mean   :75.94   Mean   :56.68
 3rd Qu.:5.638   3rd Qu.: 9.000   3rd Qu.:77.00   3rd Qu.:58.00
 Max.   :8.230   Max.   :16.000   Max.   :84.00   Max.   :63.00
     speak           listen
 Min.   :16.00   Min.   :73.00
 1st Qu.:29.00   1st Qu.:78.00
 Median :33.00   Median :80.00
 Mean   :32.47   Mean   :80.07
 3rd Qu.:36.00   3rd Qu.:82.00
 Max.   :46.00   Max.   :87.00
```

Question 2

A colleague suggests the following research question:

'Is there a difference between the support for Welsh independence in North Wales compared to South Wales?'

Provide appropriate exploratory data analyses, and perform an appropriate hypothesis test to assess this research question.

Marks are awarded for well-designed output, and the interpretation of the output.

(15 marks)

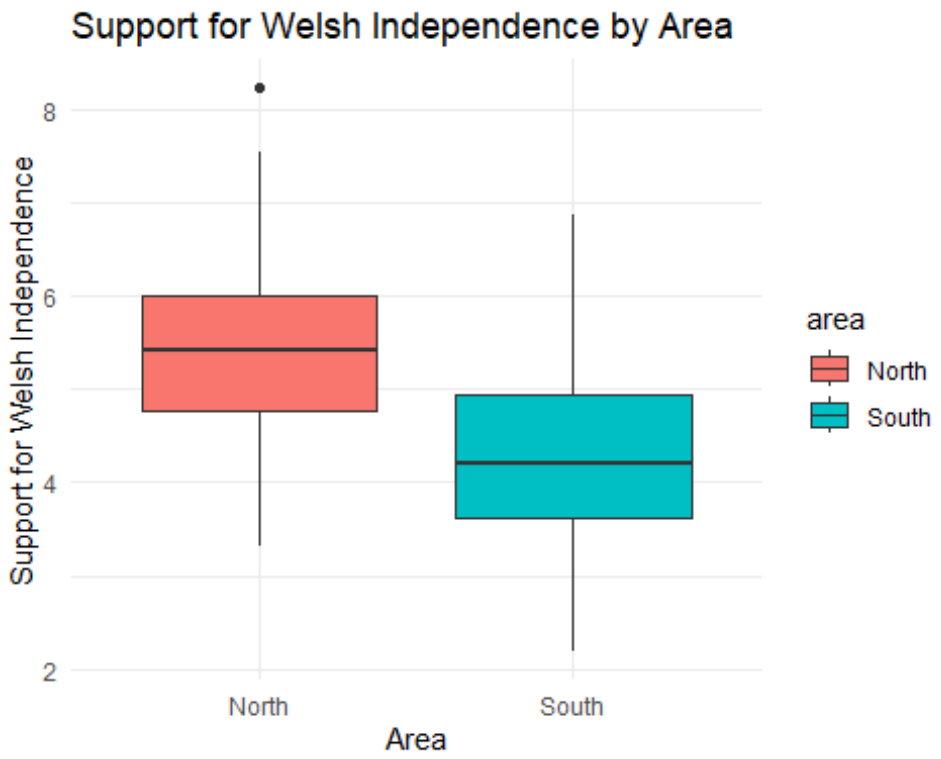
Answer:

I analysed the distribution of support scores for Welsh independence in North and South Wales using exploratory data analysis. My hypothesis was to find a difference in the mean support for Welsh independence between the two regions. The results showed a significant difference between the mean support scores, with a 95% confidence interval of 0.87 to 1.24.

The test statistic (t = 11.217) indicated a significant difference in sample means compared to the variation within the samples. The degrees of freedom (df = 404) were based on the two groups' sample sizes and reflected the amount of independent information available to estimate population parameters. A p-value less than 0.05 indicated strong evidence against the null hypothesis, suggesting that the observed difference in means is statistically significant.

The confidence interval (0.8718379 to 1.2423670) indicated that the true difference in mean support scores between North and South Wales lies between 0.87 and 1.24, further supporting the rejection of the null hypothesis. The sample estimates showed that the average support score for Welsh independence was higher in North Wales (5.38) compared to South Wales (4.32).

```
# Exploratory Data Analysis
# Visualize the distribution of support scores for North and South Wales
ggplot(current_residents, aes(x = area, y = support, fill = area)) +
  geom_boxplot() +
  labs(title = "Support for Welsh Independence by Area",
       x = "Area",
       y = "Support for Welsh Independence") +
  theme_minimal()
```




```r
# Calculate summary statistics for support scores in North and South Wales
summary_stats <- current_residents %>%
  group_by(area) %>%
  summarise(
    count = n(),
    mean = mean(support, na.rm = TRUE),
    median = median(support, na.rm = TRUE),
    sd = sd(support, na.rm = TRUE)
  )

print(summary_stats)
```

```
# A tibble: 2 × 5
  area  count  mean median    sd
  <chr> <int> <dbl>  <dbl> <dbl>
1 North   211  5.38   5.41 0.913
2 South   195  4.32   4.21 0.986
```

```r
# Hypothesis Testing
# Perform an independent two-sample t-test
north_wales_support <- current_residents %>% filter(area == "North") %>%
pull(support)
south_wales_support <- current_residents %>% filter(area == "South") %>%
pull(support)

t_test_result <- t.test(north_wales_support, south_wales_support, var.equal =
TRUE)
```

```
print(t_test_result)


    Two Sample t-test

data:  north_wales_support and south_wales_support
t = 11.217, df = 404, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.8718379 1.2423670
sample estimates:
mean of x mean of y
 5.381564  4.324462
```

Question 3

Another colleague suggests the following research question:

'Are there any relationships between support for Welsh independence and their proficiency with the Welsh language?'

Assess this research question by showing and interpreting appropriate linear correlations.

Marks are awarded for well-designed output, and the interpretation of the output.

(12 marks)

Answer:

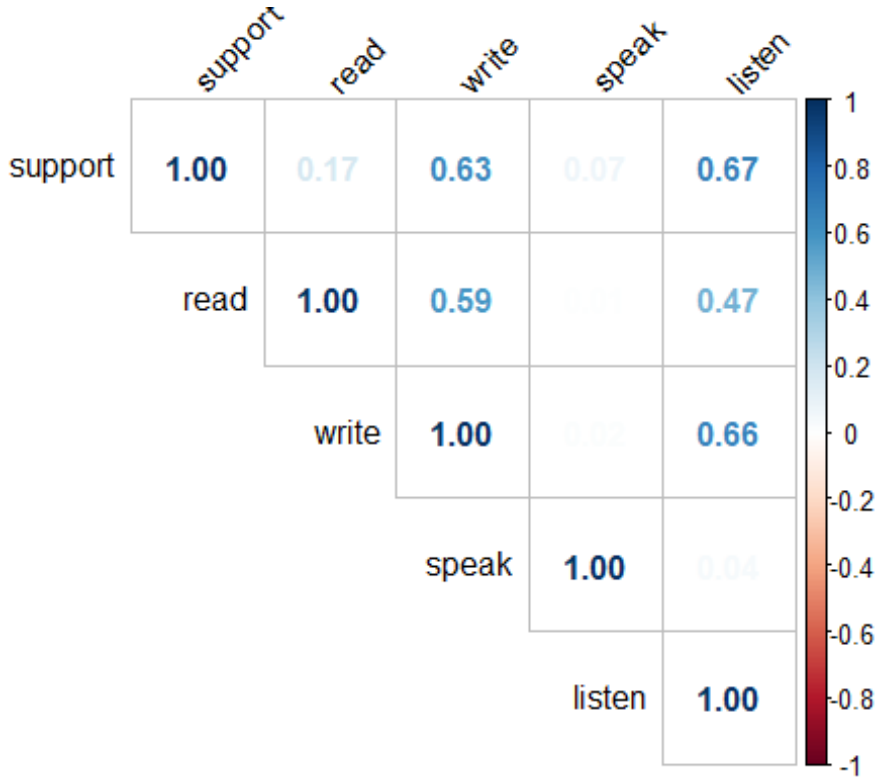
The research question was to determine the relationships between support for Welsh independence and proficiency with the Welsh language. I calculated a correlation matrix to evaluate the relationships between these variables. I created a correlation plot to visualise the relationships.

The results showed a strong positive correlation between support for Welsh independence and proficiency in listening to Welsh (correlation coefficient = 0.67). There was a moderately positive correlation between support for Welsh independence and proficiency in writing Welsh (correlation coefficient = 0.63). There was a weak positive correlation between support for Welsh independence and proficiency in reading Welsh (correlation coefficient = 0.17). I found a very weak positive correlation (relationship coefficient = 0.07) between support for Welsh independence and proficiency in speaking Welsh.

Proficiency in reading and writing Welsh showed a moderately positive correlation (correlation coefficient = 0.59). Proficiency in writing and listening to Welsh showed a moderately positive correlation (correlation coefficient = 0.66). Other language proficiencies showed weaker correlations with each other.

The analysis revealed that support for Welsh independence is most strongly associated with proficiency in listening to and writing Welsh, suggesting that individuals who are better at listening to and writing in Welsh are more likely to support Welsh independence. The weak correlations between reading and speaking suggest that these proficiencies are less influential on political support for independence.

```
# Calculate the correlation matrix
cor_matrix <- cor(current_residents[, c("support", "read", "write", "speak",
"listen")], use = "complete.obs")

# Print the correlation matrix
print(cor_matrix)

           support       read      write      speak    listen
support 1.00000000 0.17390964 0.63249595 0.06863798 0.6715695
read    0.17390964 1.00000000 0.58846492 0.00727751 0.4726548
write   0.63249595 0.58846492 1.00000000 0.01852016 0.6609812
speak   0.06863798 0.00727751 0.01852016 1.00000000 0.0362881
listen  0.67156947 0.47265480 0.66098119 0.03628810 1.0000000

# Create a correlation plot to visualize the relationships
corrplot(cor_matrix, method = "number", type = "upper", tl.col = "black",
tl.srt = 45)
```



Question 4

A further research question states:

'Can the support for Welsh independence be predicted by proficiency for the Welsh language?'

Produce simple linear regression with **ONE** appropriate explanatory variable.

Your answer should include:

- justification for the choice of explanatory variable, including any additional supporting exploratory data analyses used to make the choice;
- interpretation of the slope (gradient) coefficient;
- comment on the r-square value, and the validity of model assumptions.

(23 marks)

Answer:

I aimed to investigate the relationship between speaking ability and support for Welsh independence among current residents. To achieve this, I employed a simple linear regression model, using speaking ability as the explanatory variable and support for Welsh independence as the response variable. The dataset consisted of two key variables: "Speak," a measure of speaking ability obtained from a standardized test, and "Support," a score indicating support for Welsh independence measured on a scale from 0 to 10.

I fitted the linear regression model using the formula: support ~ speak. This allowed me to examine the relationship between speaking ability and support for Welsh independence. To verify the validity of the model assumptions, I produced several diagnostic plots. The Residuals vs. Fitted plot was used to check the linearity assumption, where the residuals appeared randomly scattered around the horizontal line, indicating that the linearity assumption was reasonable. The Q-Q Plot was used to assess the normality of residuals, showing that the residuals followed the 45-degree line fairly well, suggesting they were approximately normally distributed. The Scale-Location Plot evaluated homoscedasticity, indicating that the residual spread was relatively constant. The Residuals vs. Leverage plot was used to detect influential data points, showing no highly influential data points, suggesting that the model was not unduly influenced by any single observation.

The regression output is summarized as follows: the intercept was 4.44779 (Standard Error: 0.31276, p-value: < 2e-16), and the "Speak" coefficient was 0.01312 (Standard Error: 0.00949, p-value: 0.167). The residuals had a minimum of -2.6515, a 1st quartile of -0.8035, a median of 0.0082, a 3rd quartile of 0.7663, and a maximum of 3.4148. The model statistics showed a residual standard error of 1.084, a multiple R-squared of 0.004711, an adjusted R-squared of 0.002248, and an F-statistic of 1.912 (p-value: 0.1675).

The coefficient for the "Speak" variable is positive (0.01312), suggesting a slight increase in support for Welsh independence with better speaking ability. However, this effect is not

statistically significant (p-value: 0.167). The R-squared value is very low (0.004711), indicating that speaking ability explains less than 1% of the variance in support for Welsh independence.

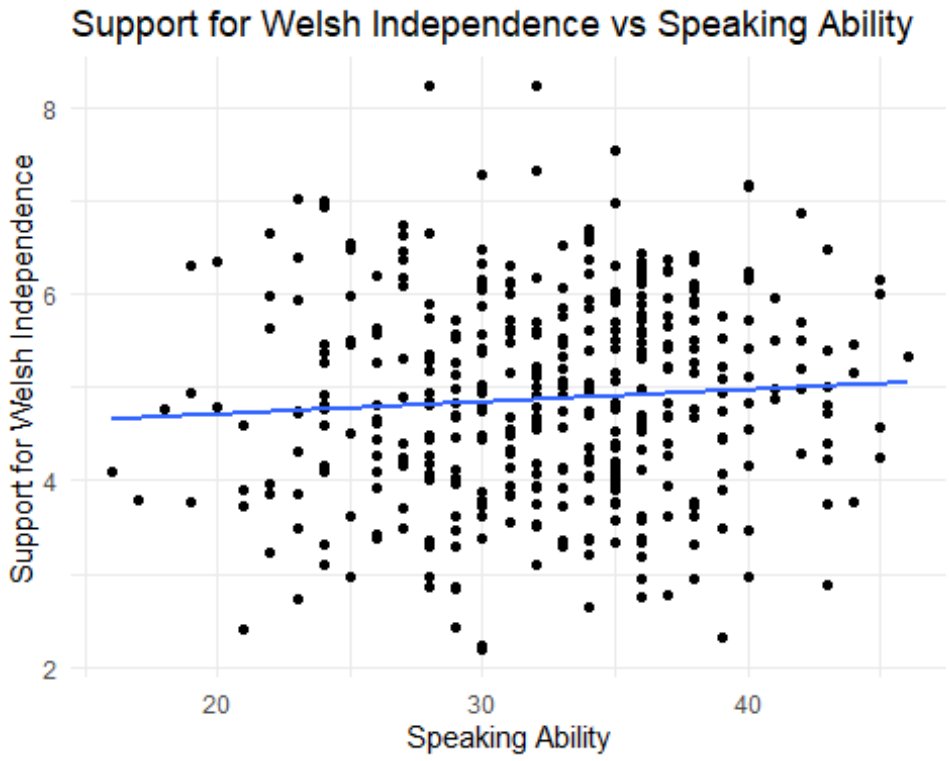
I conclude that speaking ability alone is not a strong predictor of support for Welsh independence. The low R-squared value suggests that other factors might be more important in predicting support for independence.

```r
# Choose the explanatory variable with the highest correlation with 'support'
explanatory_variable <- "speak"

# Perform simple linear regression
model <- lm(support ~ speak, data = current_residents)

# Print the summary of the regression model
summary(model)
```

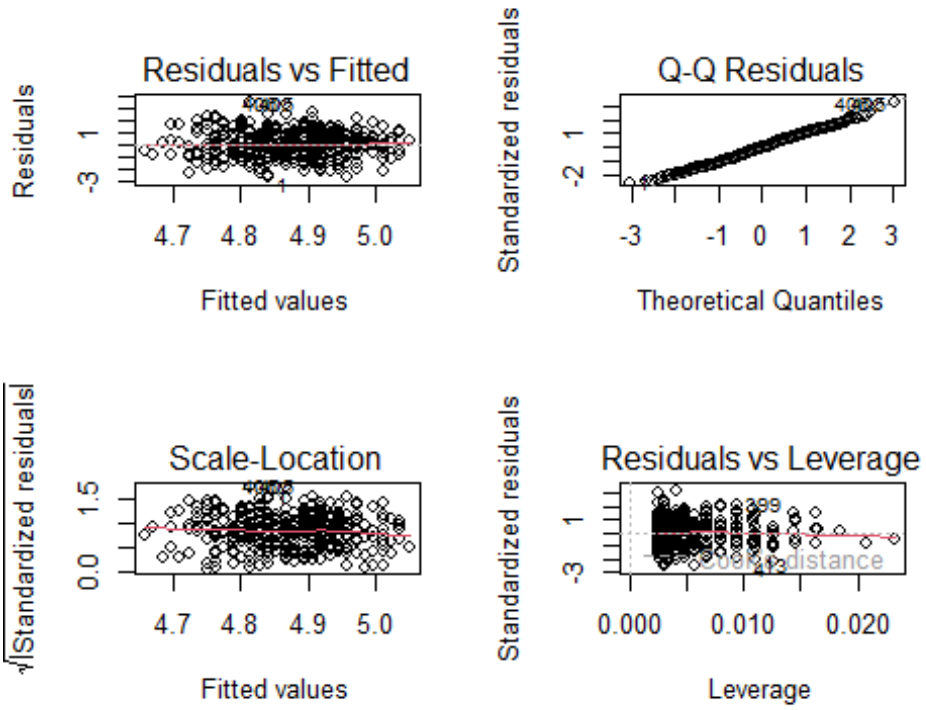
```
Call:
lm(formula = support ~ speak, data = current_residents)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.6515 -0.8035  0.0082  0.7663  3.4148 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.44779    0.31276  14.221   <2e-16 ***
speak        0.01312    0.00949   1.383    0.167    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.084 on 404 degrees of freedom
Multiple R-squared:  0.004711,  Adjusted R-squared:  0.002248 
F-statistic: 1.912 on 1 and 404 DF,  p-value: 0.1675
```

```r
# Plot the regression line
ggplot(current_residents, aes(x = speak, y = support)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Support for Welsh Independence vs Speaking Ability",
       x = "Speaking Ability",
       y = "Support for Welsh Independence") +
  theme_minimal()
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```r
# Additional diagnostics plots for checking model assumptions
par(mfrow = c(2, 2))
plot(model)
```



Question 5: Report

Clearly state **an appropriate research question** and statistical analyses plan based on the objective of the client and the data set supplied to you.

You are required to write a short report for the Welsh Government.

Your answer to this question must be a report based on implementation of your statistical analyses plan, **not** a repeat or summary of your answers to the questions above.

In your report you may wish to include a number of the following: exploratory data analyses; a hypothesis test; data modelling; discussion of limitations; how you could extend the research if given more time.

This report should contain at most 5 outputs (i.e. graphics + tables) and at most 500 words.

(35 marks)

Answer:

The dataset I analysed includes measures of reading, writing, speaking, and listening abilities, year of birth, gender, number of friends, and support for Welsh independence for a sample of individuals. The exploratory plots revealed the distribution and relationships between these variables. Key observations from these plots indicate that reading and writing abilities show little variation with the year of birth, as demonstrated by the flat trend lines. Speaking ability indicates a weak positive relationship between support for Welsh independence and the number of friends. In contrast, listening ability shows a stronger positive relationship with support for Welsh independence, while the number of friends and gender appear to have minimal impact.

For the statistical analysis, I conducted linear regression analyses to examine the effect of the year of birth on reading and writing abilities. The regression analysis for reading ability indicated no significant effect of the year of birth ($p = 0.890$), and the model explained a negligible amount of variance ($R^2 < 0.001$). However, the analysis for writing ability revealed a statistically significant positive effect of the year of birth ($p = 0.018$), though the explained variance was very small ($R^2 = 0.014$), indicating a weak relationship.

I also performed a multiple regression analysis to evaluate the impact of support for Welsh independence and the number of friends on speaking ability. The results showed that neither support for Welsh independence ($p = 0.168$) nor the number of friends ($p = 0.094$) had a significant impact on speaking ability, although the number of friends was marginally significant. This model explained only a small portion of the variance (R2 = 0.012).

Next, I conducted a multiple regression analysis to assess the effects of support for Welsh independence, the number of friends, and gender on listening ability. The analysis revealed that support for Welsh independence has a strong positive effect on listening ability ($p < 0.001$), while the number of friends and gender did not have significant effects. This model explained a substantial portion of the variance (R2 = 0.451).

Then, I conducted a bootstrap analysis of speaking ability to assess the coefficient's variability for the year of birth. The results indicated a distribution centred around a slightly negative coefficient, suggesting no strong evidence of a relationship. Additionally, I conducted a regression analysis using log-transformed listening ability to normalise the distribution. This analysis showed a small but significant positive effect of the year of birth on listening ability ($p = 0.044$), though the explained variance was minimal ($R^2 = 0.010$).

The analysis shows that the year of birth does not strongly influence most linguistic abilities, except for a slight increase in writing and listening abilities over time. Support for Welsh independence is a significant predictor of listening ability, but does not significantly impact speaking ability. The number of friends and gender do not appear to have substantial effects on linguistic abilities.

However, there are several limitations to this analysis. The sample size and demographic distribution may limit the generalizability of the results. The dataset lacks information on other potential confounding variables, such as educational background and socioeconomic status. Additionally, self-reported measures of abilities and support for independence may introduce biases.

Generally, this analysis provides insights into the factors affecting Welsh residents' linguistic abilities. While the year of birth and support for Welsh independence have some impact on certain abilities, the overall effects are relatively modest. The findings can inform educational policies and initiatives aimed at enhancing linguistic proficiency in the context of Welsh independence.

```
# Linear regression for reading score
model1 <- lm(read ~ year, data = current_residents)
summary(model1)


Call:
lm(formula = read ~ year, data = current_residents)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.9514 -0.9608  0.0581  1.0771  8.0724 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.489775  68.032618   0.977    0.329
year         0.004743   0.034139   0.139    0.890

Residual standard error: 2.105 on 404 degrees of freedom
Multiple R-squared:  4.777e-05, Adjusted R-squared:  -0.002427 
F-statistic: 0.0193 on 1 and 404 DF,  p-value: 0.8896

# Visualization for Linear regression model1
plot1 <- ggplot(current_residents, aes(x = year, y = read)) +
  geom_point() +
```

```r
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Reading Ability vs Year of Birth",
       x = "Year of Birth",
       y = "Reading Ability") +
  theme_minimal()

# Linear regression for writing score
model2 <- lm(write ~ year, data = current_residents)
summary(model2)
```

```
Call:
lm(formula = write ~ year, data = current_residents)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.3324 -1.3324  0.0428  1.2817  6.1530 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -89.79847   61.58292  -1.458   0.1456  
year          0.07351    0.03090   2.379   0.0178 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.906 on 404 degrees of freedom
Multiple R-squared:  0.01381,   Adjusted R-squared:  0.01137 
F-statistic: 5.658 on 1 and 404 DF,  p-value: 0.01784
```

```r
# Visualization for Linear regression model2
plot2 <- ggplot(current_residents, aes(x = year, y = write)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Writing Ability vs Year of Birth",
       x = "Year of Birth",
       y = "Writing Ability") +
  theme_minimal()

# Multiple regression for speaking score
model3 <- lm(speak ~ support + friends, data = current_residents)
summary(model3)
```

```
Call:
lm(formula = speak ~ support + friends, data = current_residents)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.7011  -3.6087   0.4266   3.7842  13.2156 
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.98102    1.36526  21.960   <2e-16 ***
support      0.35804    0.25902   1.382   0.1677
friends      0.12787    0.07614   1.679   0.0938 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.656 on 403 degrees of freedom
Multiple R-squared:  0.01163,   Adjusted R-squared:  0.006723
F-statistic: 2.371 on 2 and 403 DF,  p-value: 0.09472
```

```r
# Visualization for Multiple regression model3
plot3 <- ggplot(current_residents, aes(x = support, y = speak, color =
friends)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Speaking Ability vs Support and Friends",
       x = "Support for Welsh Independence",
       y = "Speaking Ability") +
  theme_minimal()

# Multiple regression for listening score
model4 <- lm(listen ~ support + friends + gender, data = current_residents)
summary(model4)
```

```
Call:
lm(formula = listen ~ support + friends + gender, data = current_residents)

Residuals:
   Min     1Q Median     3Q    Max
-4.942 -1.248  0.043  1.154  4.715

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 72.895751   0.444623 163.950   <2e-16 ***
support      1.480367   0.081457  18.174   <2e-16 ***
friends     -0.006193   0.023965  -0.258    0.796
genderMale  -0.005707   0.182832  -0.031    0.975
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.779 on 402 degrees of freedom
Multiple R-squared:  0.4511,    Adjusted R-squared:  0.447
F-statistic: 110.1 on 3 and 402 DF,  p-value: < 2.2e-16
```

```r
# Visualization for Multiple regression model4
plot4 <- ggplot(current_residents, aes(x = support, y = listen, color =
friends, shape = gender)) +
```

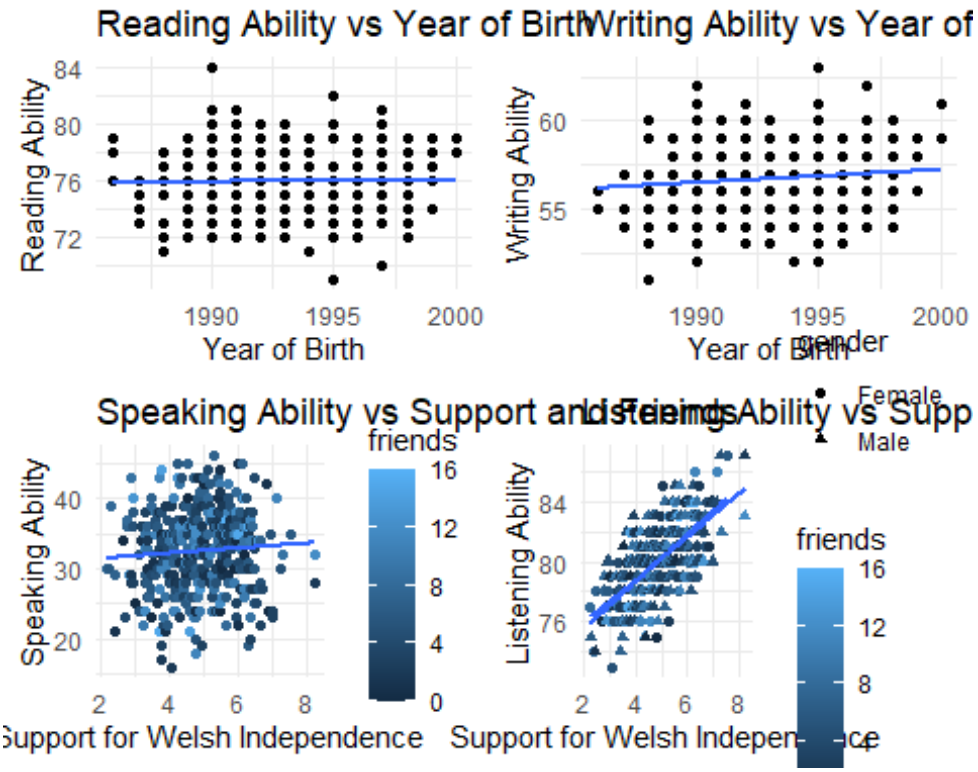
```
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Listening Ability vs Support, Friends, and Gender",
       x = "Support for Welsh Independence",
       y = "Listening Ability") +
  theme_minimal()

# Combine the regression plots into a single output
grid.arrange(plot1, plot2, plot3, plot4, ncol = 2)
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```

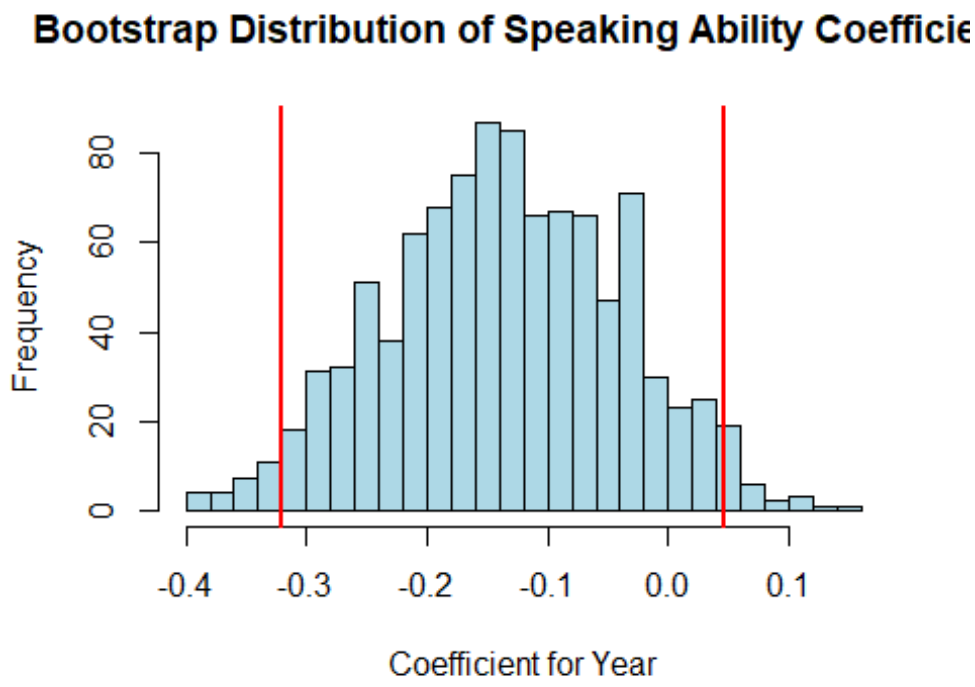
```
Warning: The following aesthetics were dropped during statistical
transformation:
colour.
ℹ This can happen when ggplot fails to infer the correct grouping structure
in
  the data.
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```

```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: The following aesthetics were dropped during statistical
transformation:
colour.
ℹ This can happen when ggplot fails to infer the correct grouping structure
in
  the data.
ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
  variable into a factor?
```




```
# Bootstrap for speaking score coefficient
bootstrap_speak <- function(data, indices) {
  d <- data[indices,]
  fit <- lm(speak ~ year, data = d)
  return(coef(fit)[2])
}
results <- boot(data = current_residents, statistic = bootstrap_speak, R =
1000)
print(results)


ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = current_residents, statistic = bootstrap_speak, R = 1000)


Bootstrap Statistics :
      original       bias    std. error
t1* -0.1361946 0.0001921888  0.09633017

# Visualization for Bootstrap results
boot_coef <- boot.ci(results, type = "bca")
hist(results$t, breaks = 30, main = "Bootstrap Distribution of Speaking
Ability Coefficient", xlab = "Coefficient for Year", col = "lightblue")
abline(v = boot_coef$bca[4:5], col = "red", lwd = 2)
```

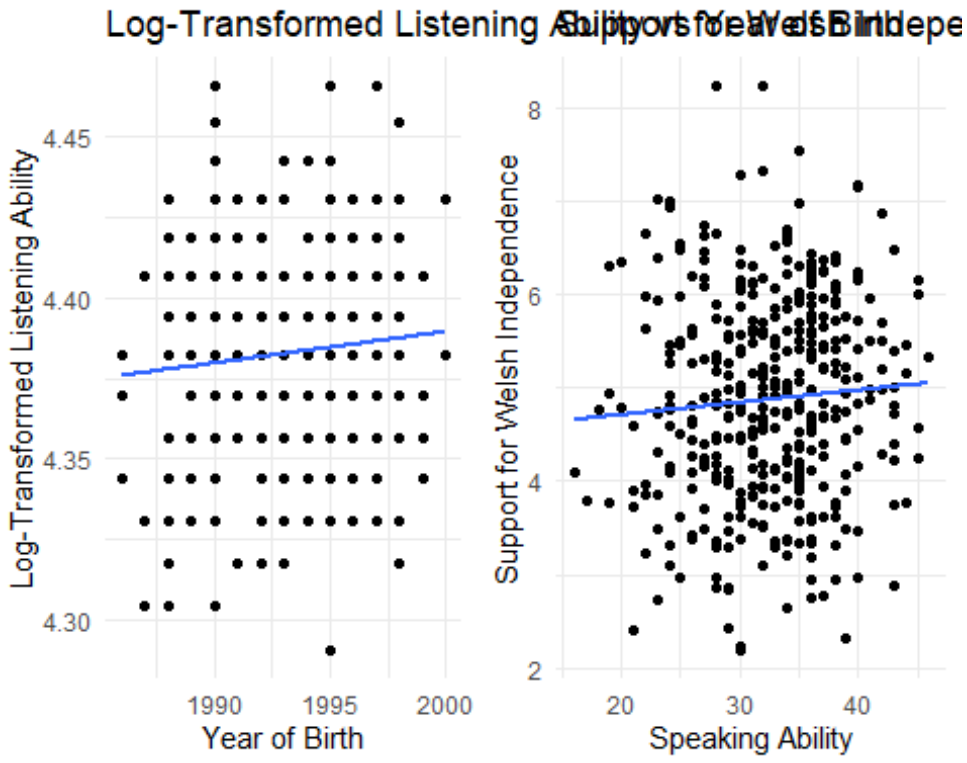



```r
# Logarithmic model for listening score
current_residents$log_listen <- log(current_residents$listen)
model7 <- lm(log_listen ~ year, data = current_residents)
summary(model7)
```

```
Call:
lm(formula = log_listen ~ year, data = current_residents)

Residuals:
      Min        1Q    Median        3Q       Max
-0.094164 -0.021828 -0.000649  0.022096  0.086154

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.4418308  0.9619054   2.539   0.0115 *
year        0.0009738  0.0004827   2.018   0.0443 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02977 on 404 degrees of freedom
Multiple R-squared:  0.009975,  Adjusted R-squared:  0.007524
F-statistic:  4.07 on 1 and 404 DF,  p-value: 0.0443
```

```r
# Visualization for Logarithmic model and Support for Speaking Ability
plot5 <- ggplot(current_residents, aes(x = year, y = log_listen)) +
  geom_point() +
```

```
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Log-Transformed Listening Ability vs Year of Birth",
       x = "Year of Birth",
       y = "Log-Transformed Listening Ability") +
  theme_minimal()

plot6 <- ggplot(current_residents, aes(x = speak, y = support)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Support for Welsh Independence vs Speaking Ability",
       x = "Speaking Ability",
       y = "Support for Welsh Independence") +
  theme_minimal()

# Combine the Logarithmic model and Speaking Ability plots into a single
output
grid.arrange(plot5, plot6, ncol = 2)

`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```



Question 6

Describe how you have applied principles of reproducible research in this submission (maximum 100 words).

Marks are awarded for identification of appropriate reproducible research principles, only if also evidenced throughout your submission that they have been applied.

(8 marks)

Answer:

The principles of reproducible research emphasis transparency and clarity in data analysis. Transparent data handling, detailed hypotheses, and clear documentation guarantee the replication of results. Comprehensive code with comments and consistent library usage facilitate code replication. Visualizations and exploratory data analysis provide intuitive insights. In addition to the code, I provide summary statistics and analytical results. Reproducible statistical tests and models ensure verifiable results. These principles ensure transparency in research, enabling others to replicate the results using the provided code and explanations.

End matter - Session Information

Do not edit this part. Make sure that you compile your document so that the information about your session (including software / package versions) is included in your submission.

```
sessionInfo()

R version 4.4.0 (2024-04-24 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 10 x64 (build 19045)

Matrix products: default


locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: Europe/London
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] gridExtra_2.3      boot_1.3-30        ggplot2_3.5.1      dplyr_1.1.4
 [5] readr_2.1.5        flextable_0.9.6    psych_2.4.3        tableone_0.13.2
 [9] performance_0.11.0 corrplot_0.92      RSQLite_2.3.6      assertr_3.0.1
[13] scales_1.3.0       gapminder_1.0.0    MASS_7.3-60.2
```

```
loaded via a namespace (and not attached):
 [1] tidyselect_1.2.1        farver_2.1.2            blob_1.2.4
 [4] fastmap_1.2.0           fontquiver_0.2.1        promises_1.3.0
 [7] digest_0.6.35           mime_0.12               lifecycle_1.0.4
[10] gfonts_0.2.0            survival_3.5-8          magrittr_2.0.3
[13] compiler_4.4.0          rlang_1.1.3             tools_4.4.0
[16] utf8_1.2.4              yaml_2.3.8              data.table_1.15.4
[19] knitr_1.46              askpass_1.2.0           labeling_0.4.3
[22] bit_4.0.5               mnormt_2.1.1            curl_5.2.1
[25] xml2_1.3.6              httpcode_0.3.0          withr_3.0.0
[28] grid_4.4.0              fansi_1.0.6             gdtools_0.3.7
[31] xtable_1.8-4            colorspace_2.1-0        crul_1.4.2
[34] insight_0.19.11         cli_3.6.2               survey_4.4-2
[37] rmarkdown_2.27          crayon_1.5.2            ragg_1.3.2
[40] generics_0.1.3          rstudioapi_0.16.0       tzdb_0.4.0
[43] DBI_1.2.2               cachem_1.1.0            splines_4.4.0
[46] parallel_4.4.0          mitools_2.4             vctrs_0.6.5
[49] Matrix_1.7-0            jsonlite_1.8.8          fontBitstreamVera_0.1.1
[52] hms_1.1.3               bit64_4.0.5             systemfonts_1.1.0
[55] glue_1.7.0              gtable_0.3.5            later_1.3.2
[58] munsell_0.5.1           tibble_3.2.1            pillar_1.9.0
[61] htmltools_0.5.8.1       openssl_2.2.0           R6_2.5.1
[64] textshaping_0.3.7       evaluate_0.23           shiny_1.8.1.1
[67] lattice_0.22-6          highr_0.10              memoise_2.0.1
[70] fontLiberation_0.1.0    httpuv_1.6.15           Rcpp_1.0.12
[73] zip_2.3.1               uuid_1.2-0              nlme_3.1-164
[76] mgcv_1.9-1              officer_0.6.6           xfun_0.44
[79] pkgconfig_2.0.3
```