Exam 2: PSY V0500 Statistical Methods in Psychology

2024-04-19

Questions 1

Does age influence the likelihood of experiencing loneliness? To investigate this question, we analyze a dataset containing information on individuals' age grouped into intervals of 5 and their self-reported loneliness status. The data is categorized into different age brackets, and loneliness status is classified as either 'never' (coded as 0) or 'always' (coded as 1). The objective is to determine if there is a statistically significant association between age groups and loneliness status. Given that both variables are categorical in nature we will conduct a chi square test.

Pearson's Chi-squared test is a statistical method used to determine if there is a significant association between two categorical variables. In this analysis, the variables of interest are age groups and loneliness status. The test assesses whether the observed frequencies in each category deviate significantly from what would be expected if there were no association between the variables

Assumptions

1. *Independence of observations*: The observations in the contingency table must be independent of each other.
2. *Expected cell frequencies*: The expected frequency of observations in each cell of the contingency table should be sufficiently large

Null Hypothesis (H0): There is no association between age group and loneliness status. In other words, the proportion of individuals who are always lonely is the same across all age groups.

Alternative Hypothesis (H1): There is an association between age group and loneliness status. Specifically, the proportion of individuals who are always lonely varies across age groups.

```
lonely <- c(0, 0, 0, 0, 0, 0, 0, 0,1, 1, 1, 1, 1, 1, 1, 1)
age_group <- c("[15,25]", "(25,35]", "(35,45]", "(45,55]", "(55,65]", "(65,75]", "(75,85]", "(85,95]",
count <- c(1699, 7663, 10937, 9742, 11997, 11480, 3459, 329, 569, 1478, 1658, 1319, 1204, 758, 179, 28)

data <- data.frame(lonely, age_group, count)

Table = xtabs(data$count~data$age_group + data$lonely)

print(Table)
```

```
##                data$lonely
## data$age_group     0     1
##        (25,35]  7663  1478
##        (35,45] 10937  1658
##        (45,55]  9742  1319
##        (55,65] 11997  1204
```

```
##        (65,75] 11480   758
##        (75,85]  3459   179
##        (85,95]   329    28
##        [15,25]  1699   569
```

```
chisq.test(Table)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Table
## X-squared = 1240, df = 7, p-value < 2.2e-16
```

- Chi-squared statistic: 1240
- Degrees of freedom (df): 7
- p-value: < 0.000

Since the p-value is significantly less than the chosen significance level (typically 0.05), we reject the null hypothesis. This indicates strong evidence to suggest that there is an association between age groups and loneliness status. In other words, the proportion of individuals who are always lonely varies significantly different across all age groups.

Questions 2

Problem Statement Is there an association between marital status, education level, and loneliness among individuals?

Null Hypothesis (H0): There is no significant association between marital status, education level, and loneliness. In other words, the likelihood of experiencing loneliness is independent of both marital status and education level.

Alternative Hypothesis (H1): There is a significant association between marital status, education level, and loneliness. Specifically, married individuals and those with higher education levels are less likely to experience loneliness compared to unmarried individuals and those with lower education levels

```
educ_recode = c("not 4yr degree", "not 4yr degree", "not 4yr degree",
                "4yr or more", "4yr or more", "4yr or more")
marital_status = c("married", "divorced_or_separated", "never_married","married", "divorced_or_separate

lonely = c(15022, 7103, 6017, 20927, 4617, 6425)

data1 <- data.frame(educ_recode, marital_status, lonely)

Table = xtabs(data1$lonely~data1$educ_recode + data1$marital_status)

print(Table)
```

```
##                  data1$marital_status
## data1$educ_recode divorced_or_separated married never_married
##    4yr or more                     4617   20927          6425
##    not 4yr degree                  7103   15022          6017
```

```r
chisq.test(Table, le)
```

```
##
##  Pearson's Chi-squared test
##
## data:  Table
## X-squared = 1272.2, df = 2, p-value < 2.2e-16
```

- Chi-squared statistic: 1272.2
- Degrees of freedom (df): 2
- p-value: < 0.000

The p-value obtained is significantly less than the chosen significance level (typically 0.05), indicating strong evidence to reject the null hypothesis. Therefore, we conclude that there is a significant association between marital status, education level, and loneliness.

Questions 3

```r
library(tidyverse)
# load data
Household_Pulse_data = read.csv("hps_04_00_02_puf.csv")

#head(Household_Pulse_data)
```

```r
# For
Household_Pulse_data$lonely <- as.numeric(
(Household_Pulse_data$SOCIAL2 == 1) |
 (Household_Pulse_data$SOCIAL2 == 2) )
```

```r
# Select relevant variables
selected_data <- Household_Pulse_data %>%
  select(lonely, EGENID_BIRTH, RRACE, FRMLA_YN)


# OLS linear model
model <- lm(lonely ~., data = selected_data)

summary(model)
```

```
##
## Call:
## lm(formula = lonely ~ ., data = selected_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14058 -0.10309 -0.10309 -0.09031  0.91128
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   6.355e-02  5.383e-03  11.806  < 2e-16 ***
## EGENID_BIRTH  1.278e-02  2.278e-03   5.607 2.07e-08 ***
## RRACE         1.243e-02  1.407e-03   8.835  < 2e-16 ***
## FRMLA_YN     -1.769e-05  4.185e-05  -0.423    0.673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3019 on 71148 degrees of freedom
## Multiple R-squared:  0.001543,   Adjusted R-squared:  0.001501
## F-statistic: 36.66 on 3 and 71148 DF,  p-value: < 2.2e-16
```

a) Variables chosen as predictors:

- Race: Different racial or ethnic backgrounds may have varying social support systems and experiences of loneliness.
- Gender: Gender might influence loneliness due to societal expectations and differences in social interactions.
- Infants_under_18_months: The presence of infants in the household could impact loneliness due to increased caregiving responsibilities and changes in social dynamics.

b) The estimates from the linear model suggest that variables such as age and race, areassociated with reported loneliness. Specifically, older age and certain racial backgrounds appear to be positively associated with loneliness, while the presence of infants in the household does not seem to have a significant effect on loneliness scores.

c)

```
anova(model)
```

```
## Analysis of Variance Table
##
## Response: lonely
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## EGENID_BIRTH     1    2.9  2.9082 31.9126 1.619e-08 ***
## RRACE            1    7.1  7.0971 77.8792 < 2.2e-16 ***
## FRMLA_YN         1    0.0  0.0163  0.1786    0.6726
## Residuals    71148 6483.7  0.0911
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Questions 4

```
# logistic regression model
logit_model <- glm(lonely ~ ., data = selected_data, family = binomial)

summary(logit_model)
```

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

```
## glm(formula = lonely ~ ., family = binomial, data = selected_data)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5890882  0.0595567 -43.473  < 2e-16 ***
## EGENID_BIRTH  0.1414175  0.0252051   5.611 2.02e-08 ***
## RRACE         0.1245529  0.0141459   8.805  < 2e-16 ***
## FRMLA_YN     -0.0001927  0.0004596  -0.419    0.675
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 46754  on 71151  degrees of freedom
## Residual deviance: 46649  on 71148  degrees of freedom
## AIC: 46657
##
## Number of Fisher Scoring iterations: 5
```

In the logistic regression model, the coefficients for "EGENID_BIRTH" and "RRACE" remain positive and statistically significant, indicating that older age and certain racial backgrounds are associated with higher odds of loneliness.

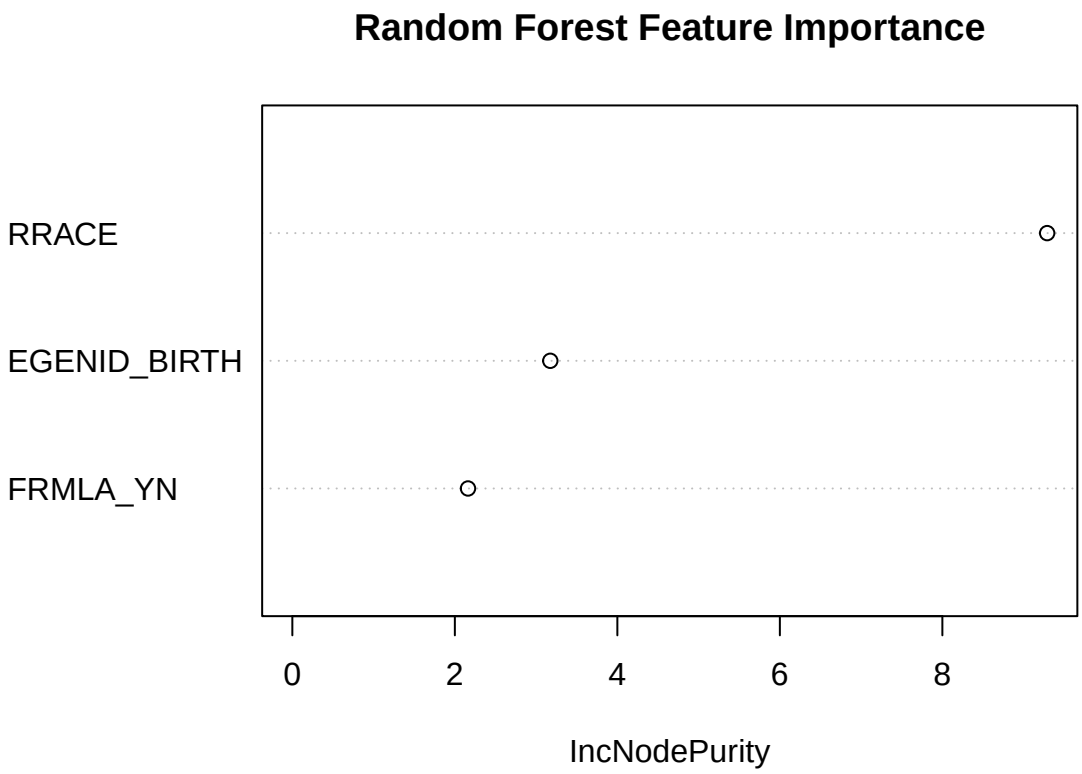
Questions 5

Tree models such as Random forest are usually best for feature extraction give that we can leverage the model's feature importance scores.These scores quantify the contribution of each predictor variable to the model's predictive performance.

```
# Random forest
library(randomForest)
```

```
## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
# Random Forest model
rf_model <- randomForest(lonely ~ ., data = selected_data)

# Plot feature importance
varImpPlot(rf_model, main = "Random Forest Feature Importance")
```



When comparing the three select ed features Race emerges as the most influencial variable to lonliness.