Zahra Mohamadkhani

Discrete data analysis of poverty in Brazil

Loading Data

In [2]:
```
library("readxl")
df <- read_excel("poverty_brazil.xlsx")
```

```
Warning message:
“Expecting numeric in F6988 / R6988C6: got 'NA'”
```

Contingency tables

In [3]:
```
independent_vars <- c("woman", "age", "education", "work", "metropolitan_area",
                      "non_white", "urban", "work_permit")
for (var in independent_vars) {
  cat("Contingency table for", var, "vs poverty:\n")
  print(table(df[[var]], df$poverty))
  cat("\n")
}
attach(df)
```

Contingency table for woman vs poverty:

	0	1
0	9433	3152
1	6651	1516

Contingency table for age vs poverty:

	0	1
14	3	2
15	16	5
16	27	9
17	46	23
18	129	34
19	164	55
20	197	62
21	234	77
22	265	83
23	307	78
24	305	91
25	301	86
26	302	86
27	326	100
28	337	89
29	397	105
30	408	120
31	367	122
32	379	106
33	438	139
34	413	153
35	423	140
36	423	162
37	419	162
38	484	146
39	464	128
40	473	152
41	467	151
42	402	136
43	398	143
44	398	142
45	405	107
46	388	103
47	421	102
48	352	81
49	396	103
50	428	108
51	374	94
52	352	83
53	336	97
54	352	66
55	330	72
56	286	74
57	259	75
58	235	44
59	219	59
60	204	49

```
 61 153  24
 62 159  30
 63 114  26
 64 111  21
 65  84  23
 66  65  13
 67  64  25
 68  40  11
 69  42  16
 70  31  14
 71  38   9
 72  26   9
 73  24   6
 74  12  10
 75  11   7
 76  14   7
 77  11   3
 78  11   1
 79   6   3
 80   4   2
 81   4   0
 82   3   1
 83   3   1
 84   2   0
 85   1   0
 86   0   1
 87   1   0
 89   1   0
 91   0   1
```

Contingency table for education vs poverty:

```
      0    1
 1  161  238
 2 2972 1941
 3 1129  454
 4  779  366
 5 5334 1310
 6 1041  140
 7 4668  219
```

Contingency table for work vs poverty:

```
       0    1
 1  1596 1291
 2  2003  462
 3   911  460
 4  2980  792
 5   711  181
 6   621  263
 7  1867  190
 8  1318  115
 9  2564  269
 10  773  232
 11  739  413
 12    1    0
```

```
Contingency table for metropolitan_area vs poverty:

      0    1
  0 9657 3406
  1 6427 1262

Contingency table for non_white vs poverty:

      0    1
  0 8113 1316
  1 7970 3352

Contingency table for urban vs poverty:

        0     1
  0  2553  1874
  1 13531  2794

Contingency table for work_permit vs poverty:

      0    1
  0 2304 1477
  1 6677 1019
  2 7103 2172
```

choose one of them

In [4]:
```
contingency_table <- table(df$poverty, df$non_white)
```

Print the contingency table

In [5]:
```
print(contingency_table)
```

```
      0    1
  0 8113 7970
  1 1316 3352
```

Add margins to the contingency table

In [6]:
```
ccontingency_table <- addmargins(contingency_table)
print(ccontingency_table)
```

```
         0     1   Sum
  0   8113  7970 16083
  1   1316  3352  4668
  Sum 9429 11322 20751
```

Perform Cross Table analysis

In [7]:
```
install.packages("gmodels")
library("gmodels")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘gtools’, ‘gdata’
```

In [8]:
```
TABLE <- CrossTable(contingency_table, expected=TRUE, chisq=TRUE,
                    prop.chisq=FALSE, prop.c=TRUE, prop.r=TRUE, fisher=TRUE)
print(TABLE)
```

```
   Cell Contents
|-------------------------|
|                       N |
|              Expected N |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|


Total Observations in Table:  20751


             |
             |         0 |         1 | Row Total |
-------------|-----------|-----------|-----------|
           0 |      8113 |      7970 |     16083 |
             |  7307.918 |  8775.082 |           |
             |     0.504 |     0.496 |     0.775 |
             |     0.860 |     0.704 |           |
             |     0.391 |     0.384 |           |
-------------|-----------|-----------|-----------|
           1 |      1316 |      3352 |      4668 |
             |  2121.082 |  2546.918 |           |
             |     0.282 |     0.718 |     0.225 |
             |     0.140 |     0.296 |           |
             |     0.063 |     0.162 |           |
-------------|-----------|-----------|-----------|
Column Total |      9429 |     11322 |     20751 |
             |     0.454 |     0.546 |           |
-------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  722.621     d.f. =  1     p =  3.603479e-159

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  721.7237     d.f. =  1     p =  5.647241e-159


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  2.592703

Alternative hypothesis: true odds ratio is not equal to 1
p =  2.174958e-164
95% confidence interval:  2.414365 2.785207

Alternative hypothesis: true odds ratio is less than 1
p =  1
```

```
95% confidence interval:  0 2.753579

Alternative hypothesis: true odds ratio is greater than 1
p =  1.54087e-164
95% confidence interval:  2.441542 Inf



$t

       0    1
  0 8113 7970
  1 1316 3352

$prop.row

            0         1
  0 0.5044457 0.4955543
  1 0.2819195 0.7180805

$prop.col

            0         1
  0 0.8604306 0.7039392
  1 0.1395694 0.2960608

$prop.tbl

             0          1
  0 0.39096911 0.38407788
  1 0.06341863 0.16153438

$chisq

        Pearson's Chi-squared test

data:  t
X-squared = 722.62, df = 1, p-value < 2.2e-16


$chisq.corr

        Pearson's Chi-squared test with Yates' continuity correction

data:  t
X-squared = 721.72, df = 1, p-value < 2.2e-16


$fisher.ts

        Fisher's Exact Test for Count Data

data:  t
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
```

```
 2.414365 2.785207
sample estimates:
odds ratio
  2.592703


$fisher.tl

	Fisher's Exact Test for Count Data

data:  t
p-value = 1
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 2.753579
sample estimates:
odds ratio
  2.592703


$fisher.gt

	Fisher's Exact Test for Count Data

data:  t
p-value < 2.2e-16
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 2.441542      Inf
sample estimates:
odds ratio
  2.592703
```

Summary of Findings:Significant Association: Both Pearson's chi-squared test and Fisher's exact test show a highly significant association between the two variables.

Odds Ratio: The odds ratio indicates a substantial difference in the likelihood of being in the response category between the two groups.

Proportions: The proportions in the contingency table help understand the distribution and independence of the variables.

Overall, the analysis provides strong evidence of a significant association between the two categorical variables, with a clear quantification of the strength of this association through the odds ratio.

Perform the Pearson Chi-Square test

In [9]:
```
chi_square_test <- chisq.test(contingency_table)
```

Print the result of the Chi-Square test

In [10]:
```
print(chi_square_test)
```

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 721.72, df = 1, p-value < 2.2e-16
```

very strong evidence against the null hypothesis.null hypothesis: indepedance

In [ ]:
```
lrt_result <- chisq.test(contingency_table, simulate.p.value = TRUE)
print(lrt_result)
```

If the p-value is less than the significance level (0.05), you reject the null hypothesis.

if the p-value is greater than the significance level, you fail to reject the null hypothesis.

In our case, the p-value is 0.0004998, which is much smaller than 0.05.

confidence interval

In [12]:
```
install.packages("epitools")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

In [13]:
```
library(epitools)
contingency_table <- table(df$poverty, df$non_white)
print(contingency_table)
pearson_test_result <- chisq.test(contingency_table)
print(pearson_test_result)
```

```
      0    1
  0 8113 7970
  1 1316 3352

	Pearson's Chi-squared test with Yates' continuity correction

data:  contingency_table
X-squared = 721.72, df = 1, p-value < 2.2e-16
```

Calculate the odds ratio and its confidence interval

In [14]:
```
odds_ratio_result <- oddsratio(contingency_table)
print(odds_ratio_result$measure)
```

```
 odds ratio with 95% C.I.
  estimate   lower   upper
0 1.000000      NA      NA
1 2.592536 2.41575 2.78362
```

Odds Ratio Estimate: 2.5925 Interpretation: Non-white individuals are approximately 2.59 times more likely to be in poverty compared to non-non-white (white) individuals.

Calculate the risk ratio

In [15]:
```
risk_ratio_result <- riskratio(contingency_table)
print(risk_ratio_result$measure)
```

```
 risk ratio with 95% C.I.
  estimate    lower    upper
0 1.000000       NA       NA
1 1.449045 1.414971 1.483939
```

Risk Ratio Estimate: 1.4490 Interpretation: The risk of poverty is approximately 1.45 times higher for non-white individuals compared to non-non-white (white) individuals.

difference of proportions

In [16]:
```
contingency_table <- table(df$poverty, df$non_white)
print(contingency_table)
```

```
     0    1
0 8113 7970
1 1316 3352
```

Extract counts for the two proportions

In [17]:
```
count_poverty_non_white <- contingency_table[2, 2]  # Non-White and in Poverty
count_non_poverty_non_white <- contingency_table[1, 2]  # Non-White not in Poverty
count_poverty <- sum(contingency_table[2, ])  # Total in Poverty
count_non_poverty <- sum(contingency_table[1, ])  # Total not in Poverty
```

Create vectors of successes and trials

In [18]:
```
successes <- c(count_poverty_non_white, count_non_poverty_non_white)
trials <- c(count_poverty, count_non_poverty)
```

Successes: The number of occurrences of the event of interest in each group.

Trials: The total number of observations or trials in each group.

Perform the two-proportion test

In [19]:
```
prop_test_result <- prop.test(successes, trials)
print(prop_test_result)
```

```
	2-sample test for equality of proportions with continuity correction

data:  successes out of trials
X-squared = 721.72, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.2073446 0.2377078
sample estimates:
   prop 1    prop 2 
0.7180805 0.4955543 
```

X-squared = 721.72: This is the chi-squared test statistic. It measures the difference between the observed and expected proportions under the null hypothesis.

3. Degrees of Freedom df = 1: The degrees of freedom for this test. For a two-sample proportion test, the degrees of freedom are typically 1.

4. P-value: p-value < 2.2e-16: The p-value is extremely small, indicating that the observed difference in proportions is highly unlikely to have occurred by chance. This strong evidence suggests rejecting the null hypothesis that the proportions are equal.

5. Alternative Hypothesis: alternative hypothesis: two.sided: The test is two-sided, meaning it checks for differences in both directions (whether one proportion is greater or less than the other).

6. Confidence Interval: 95 percent confidence interval: 0.2073446 to 0.2377078: This interval estimates the range within which the true difference in proportions lies with 95% confidence. Since the interval does not include 0 and is entirely positive, it indicates a significant difference between the two proportions, with the first group having a higher proportion than the second group.

7. Sample Estimates:

prop 1 = 0.7180805: The proportion of successes in the first group.

prop 2 = 0.4955543: The proportion of successes in the second group.

Measures of Association (e.g., Phi Coefficient)

In [20]:
```
install.packages("vcd")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘zoo’, ‘lmtest’
```

In [21]:
```r
library(vcd)
contingency_table <- table(df$poverty, df$non_white)
print(contingency_table)
```

```
Loading required package: grid


Attaching package: ‘vcd’


The following object is masked from ‘package:epitools’:

    oddsratio

```

```
      0    1
  0 8113 7970
  1 1316 3352
```

Compute association statistics

In [22]:
```r
association_stats <- assocstats(contingency_table)
print(association_stats)
```

```
                    X^2 df P(> X^2)
Likelihood Ratio 746.91  1        0
Pearson          722.62  1        0

Phi-Coefficient   : 0.187
Contingency Coeff.: 0.183
Cramer's V        : 0.187
```

Phi coefficient for 2x2 tables

In [23]:
```r
phi_coefficient <- association_stats$phi
print(paste("Phi Coefficient:", phi_coefficient))
```

```
[1] "Phi Coefficient: 0.18661037442791"
```

Perform Fisher's Exact Test

In [24]:
```r
contingency_table <- table(df$poverty, df$non_white)
print(contingency_table)
fisher_test <- fisher.test(contingency_table)
print(fisher_test)
```

```
     0    1
  0 8113 7970
  1 1316 3352

	Fisher's Exact Test for Count Data

data:  contingency_table
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.414365 2.785207
sample estimates:
odds ratio 
  2.592703 
```

Contingency 3

In [25]:
```
install.packages("questionr")
install.packages("DescTools")
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘R.methodsS3’, ‘R.oo’, ‘R.utils’, ‘proxy’, ‘R.cach
e’, ‘e1071’, ‘styler’, ‘classInt’, ‘labelled’


Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘rootSolve’, ‘lmom’, ‘mvtnorm’, ‘expm’, ‘Exact’, ‘g
ld’
```

In [26]:
```
library(dplyr)
library(questionr)
library(DescTools)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union


Registered S3 method overwritten by 'DescTools':
  method         from
  reorder.factor gdata
```

Create the contingency table with poverty as the dependent variable and woman and non_white as independent variables

In [27]:

```
TabPoverty = table(df$woman, df$non_white, df$poverty)
dimnames(TabPoverty) = list(
  Woman = c("No", "Yes"),
  NonWhite = c("No", "Yes"),
  Poverty = c("No", "Yes")
)

print(TabPoverty)
```

```
, , Poverty = No

     NonWhite
Woman    No  Yes
  No   4649 4783
  Yes  3464 3187

, , Poverty = Yes

     NonWhite
Woman    No  Yes
  No    886 2266
  Yes   430 1086
```

Two-way tables for different combinations

In [28]:

```
TabWomanNo = TabPoverty["No", , ]
TabWomanYes = TabPoverty["Yes", , ]
TabNonWhiteNo = TabPoverty[, "No", ]
TabNonWhiteYes = TabPoverty[, "Yes", ]
```

Combined table

In [29]:
```
TabCombined = TabWomanNo + TabWomanYes
```

Odds ratios

In [31]:
```
oddsratio_combined = (TabCombined[1,1] * TabCombined[2, 2]) / (TabCombined[1,2]
                      * TabCombined[2, 1])
print(oddsratio_combined)
```

```
[1] 2.592813
```

In [32]:
```
odds.ratio(TabCombined)
```

A odds.ratio: 1 × 4

	OR	2.5 %	97.5 %	p
	<dbl>	<dbl>	<dbl>	<dbl>
Fisher's test	2.592703	2.414365	2.785207	2.174958e-164

In [44]:
```
oddsratio_woman_no = (TabWomanNo[1, 1] * TabWomanNo[2, 2]) /
                      (TabWomanNo[1, 2] * TabWomanNo[2, 1])
print(oddsratio_woman_no)
odds.ratio(TabWomanNo)
```

```
[1] 2.48591
```

A odds.ratio: 1 × 4

	OR	2.5 %	97.5 %	p
	<dbl>	<dbl>	<dbl>	<dbl>
Fisher's test	2.485741	2.275791	2.716327	1.79216e-98

In [45]:
```
oddsratio_woman_yes = (TabWomanYes[1, 1] * TabWomanYes[2, 2]) /
                       (TabWomanYes[1, 2] * TabWomanYes[2, 1])
print(oddsratio_woman_yes)
odds.ratio(TabWomanYes)
```

```
[1] 2.745094
```

A odds.ratio: 1 × 4

	OR	2.5 %	97.5 %	p
	<dbl>	<dbl>	<dbl>	<dbl>
Fisher's test	2.744768	2.427277	3.107434	2.499882e-64

In [46]:
```
oddsratio_nonwhite_no = (TabNonWhiteNo[1, 1] * TabNonWhiteNo[2, 2]) /
                         (TabNonWhiteNo[1, 2] * TabNonWhiteNo[2, 1])
print(oddsratio_nonwhite_no)
odds.ratio(TabNonWhiteNo)
```

```
[1] 0.651353
```

A odds.ratio: 1 × 4

	OR	2.5 %	97.5 %	p
	<dbl>	<dbl>	<dbl>	<dbl>
Fisher's test	0.651384	0.5743909	0.7379622	5.334033e-12

In [47]:
```
oddsratio_nonwhite_yes = (TabNonWhiteYes[1, 1] * TabNonWhiteYes[2, 2]) /
                          (TabNonWhiteYes[1, 2] * TabNonWhiteYes[2, 1])
print(oddsratio_nonwhite_yes)
odds.ratio(TabNonWhiteYes)
```

```
[1] 0.7192639
```

A odds.ratio: 1 × 4

	OR	2.5 %	97.5 %	p
	<dbl>	<dbl>	<dbl>	<dbl>
Fisher's test	0.7192879	0.6599156	0.7837614	2.348492e-14

Chi-squared tests

In [48]:
```
chisq.test(TabCombined, correct = TRUE)
chisq.test(TabWomanNo, correct = TRUE)
chisq.test(TabWomanYes, correct = TRUE)
chisq.test(TabNonWhiteNo, correct = TRUE)
chisq.test(TabNonWhiteYes, correct = TRUE)
```

```
	Pearson's Chi-squared test with Yates' continuity correction

data:  TabCombined
X-squared = 721.72, df = 1, p-value < 2.2e-16
	Pearson's Chi-squared test with Yates' continuity correction

data:  TabWomanNo
X-squared = 429.31, df = 1, p-value < 2.2e-16
	Pearson's Chi-squared test with Yates' continuity correction

data:  TabWomanYes
X-squared = 277.46, df = 1, p-value < 2.2e-16
	Pearson's Chi-squared test with Yates' continuity correction

data:  TabNonWhiteNo
X-squared = 46.502, df = 1, p-value = 9.151e-12
	Pearson's Chi-squared test with Yates' continuity correction

data:  TabNonWhiteYes
X-squared = 57.511, df = 1, p-value = 3.361e-14
```

GLM

In [50]:
```
library(readxl)
attach(df)
```

```
The following objects are masked from df (pos = 3):

    age, education, metropolitan_area, non_white, poverty, urban,
    woman, work, work_permit


The following objects are masked from df (pos = 5):

    age, education, metropolitan_area, non_white, poverty, urban,
    woman, work, work_permit


The following objects are masked from df (pos = 13):

    age, education, metropolitan_area, non_white, poverty, urban,
    woman, work, work_permit
```

In [51]:
```
install.packages("fastDummies")
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

In [52]:
```
library(fastDummies)
df1 <- dummy_cols(df,select_columns = c("education","work","work_permit"),
                  remove_selected_columns = TRUE,
                  remove_most_frequent_dummy = TRUE)
```

In [53]:
```
full_model <- glm(poverty ~ .,data = df1[1:24],
              family = binomial(link = "logit" ))
reduced_model <- glm(poverty ~ .,data = df1[2:24],
                family = binomial(link = "logit" ))
```

In [54]:
```
install.packages("jtools")
```

```
Installing package into '/usr/local/lib/R/site-library'
(as 'lib' is unspecified)
```

In [55]:
```
library(jtools)
summ(full_model , exp = TRUE)
```

```
Attaching package: 'jtools'


The following object is masked from 'package:DescTools':

    %nin%
```

```
MODEL INFO:
Observations: 20751 (1 missing obs. deleted)
Dependent Variable: poverty
Type: Generalized linear model
  Family: binomial
  Link function: logit

MODEL FIT:
χ²(23) = 3970.13, p = 0.00
Pseudo-R² (Cragg-Uhler) = 0.27
Pseudo-R² (McFadden) = 0.18
AIC = 18202.83, BIC = 18393.39

Standard errors:MLE
------------------------------------------
                          exp(Est.)   2.5%
----------------------- ----------- ------
(Intercept)                    0.41   0.34
woman                          1.15   1.05
age                            0.99   0.98
metropolitan_area              0.99   0.91
non_white                      2.12   1.96
urban                          0.51   0.46
education_1                    4.39   3.49
education_2                    2.11   1.91
education_3                    1.43   1.25
education_4                    1.42   1.22
education_6                    0.59   0.48
education_7                    0.29   0.25
work_1                         1.32   1.15
work_2                         0.85   0.74
work_3                         1.34   1.15
work_5                         0.94   0.78
work_6                         1.36   1.14
work_7                         0.60   0.50
work_8                         0.41   0.33
work_9                         0.67   0.56
work_10                        1.19   0.99
work_11                        0.79   0.67
work_12                        0.00   0.00
work_permit_0                  2.31   2.11
------------------------------------------

------------------------------------------------------------------------------------
--------------------------------------------

97.5%
----------------------- ------------------------------------------------------------
--------------------------------------------
(Intercept)
0.49
woman
1.26
age
0.99
metropolitan_area
```

```
1.08
non_white
2.29
urban
0.56
education_1
5.54
education_2
2.33
education_3
1.63
education_4
1.64
education_6
0.71
education_7
0.34
work_1
1.51
work_2
0.98
work_3
1.55
work_5
1.14
work_6
1.62
work_7
0.72
work_8
0.51
work_9
0.79
work_10
1.42
work_11
0.94
work_12                    315359464247047997159099788793749956339671893036809837798514069572351098203145559531994359575609344.00
work_permit_0
2.53
-------------------------------------------------------------------------------------------------------------------------------

---------------------------------------
                          z val.      p
----------------------- -------- ------
(Intercept)                -9.76   0.00
woman                       3.06   0.00
age                        -9.08   0.00
metropolitan_area          -0.14   0.89
non_white                  18.95   0.00
urban                     -13.45   0.00
education_1                12.54   0.00
education_2                14.60   0.00
education_3                 5.24   0.00
```

```
education_4                  4.66   0.00
education_6                 -5.36   0.00
education_7                -15.36   0.00
work_1                       3.93   0.00
work_2                      -2.32   0.02
work_3                       3.75   0.00
work_5                      -0.60   0.55
work_6                       3.39   0.00
work_7                      -5.55   0.00
work_8                      -7.95   0.00
work_9                      -4.73   0.00
work_10                      1.88   0.06
work_11                     -2.71   0.01
work_12                     -0.06   0.95
work_permit_0               17.88   0.00
---------------------------------------
```

Odds Ratio: 0.31 Interpretation: The baseline odds of poverty (for reference categories of all variables) is significantly less than 1 ($p < 0.001$), indicating a lower likelihood of poverty for the reference group.

woman1:Odds Ratio: 1.13Interpretation: Women are 13% more likely to be in poverty compared to men, significant at the 0.01 level ($p < 0.01$) others the same.

The logistic regression model indicates that several factors significantly influence poverty status. Non-white individuals have more than twice the odds of being in poverty compared to white individuals (OR = 2.17, $p < 0.001$). Women are also slightly more likely to be in poverty (OR = 1.13, $p < 0.01$). Higher education levels reduce the odds of poverty, with the highest education level (education_7) having the most substantial effect (OR = 0.28, $p < 0.001$). Urban living significantly decreases the odds of poverty (OR = 0.52, $p < 0.001$). Other factors such as age, metropolitan area, and work status show varying levels of influence on poverty.

In [56]:
```
library(lmtest)
lrtest(full_model, reduced_model)
```

```
Loading required package: zoo


Attaching package: ‘zoo’


The following objects are masked from ‘package:base’:

    as.Date, as.Date.numeric
```

A anova: 2 × 5

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	24	-9077.413	NA	NA	NA
2	23	-9082.093	-1	9.361035	0.002216469

In [57]:
```
anova(full_model, reduced_model, test = "LRT")
```

A anova: 2 × 5

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	20727	18154.83	NA	NA	NA
2	20728	18164.19	-1	-9.361035	0.002216469

The log-likelihood ratio test and the analysis of deviance table both show that including the `woman` variable significantly improves the model's fit for predicting poverty, with p-values of 0.0067 and 0.0006, respectively. This indicates that the `woman` variable has a statistically significant effect on poverty status. Excluding this variable leads to a poorer model fit, underscoring its importance in the analysis.

All Contingency tables

In [58]:
```
independent_vars <- c("woman", "age", "education", "work", "metropolitan_area",
                      "non_white", "urban", "work_permit")
for (var in independent_vars) {
  cat("Contingency table for", var, "vs poverty:\n")
  print(table(df[[var]], df$poverty))
  cat("\n")
}
```

Contingency table for woman vs poverty:

	0	1
0	9433	3152
1	6651	1516

Contingency table for age vs poverty:

	0	1
14	3	2
15	16	5
16	27	9
17	46	23
18	129	34
19	164	55
20	197	62
21	234	77
22	265	83
23	307	78
24	305	91
25	301	86
26	302	86
27	326	100
28	337	89
29	397	105
30	408	120
31	367	122
32	379	106
33	438	139
34	413	153
35	423	140
36	423	162
37	419	162
38	484	146
39	464	128
40	473	152
41	467	151
42	402	136
43	398	143
44	398	142
45	405	107
46	388	103
47	421	102
48	352	81
49	396	103
50	428	108
51	374	94
52	352	83
53	336	97
54	352	66
55	330	72
56	286	74
57	259	75
58	235	44
59	219	59
60	204	49

61	153	24
62	159	30
63	114	26
64	111	21
65	84	23
66	65	13
67	64	25
68	40	11
69	42	16
70	31	14
71	38	9
72	26	9
73	24	6
74	12	10
75	11	7
76	14	7
77	11	3
78	11	1
79	6	3
80	4	2
81	4	0
82	3	1
83	3	1
84	2	0
85	1	0
86	0	1
87	1	0
89	1	0
91	0	1

Contingency table for education vs poverty:

	0	1
1	161	238
2	2972	1941
3	1129	454
4	779	366
5	5334	1310
6	1041	140
7	4668	219

Contingency table for work vs poverty:

	0	1
1	1596	1291
2	2003	462
3	911	460
4	2980	792
5	711	181
6	621	263
7	1867	190
8	1318	115
9	2564	269
10	773	232
11	739	413
12	1	0

```
Contingency table for metropolitan_area vs poverty:

       0    1
  0 9657 3406
  1 6427 1262

Contingency table for non_white vs poverty:

       0    1
  0 8113 1316
  1 7970 3352

Contingency table for urban vs poverty:

        0     1
  0  2553  1874
  1 13531  2794

Contingency table for work_permit vs poverty:

       0    1
  0 2304 1477
  1 6677 1019
  2 7103 2172
```

Odds Ratio

In [59]:

```
coef_full_model <- coef(full_model)
odds_ratios <- exp(coef_full_model)
print(odds_ratios)
se <- sqrt(diag(vcov(full_model)))
ci_lower <- exp(coef_full_model - qnorm(0.975) * se)
ci_upper <- exp(coef_full_model + qnorm(0.975) * se)
or_ci <- data.frame(
  Odds_Ratio = odds_ratios,
  CI_Lower = ci_lower,
  CI_Upper = ci_upper
)
print(or_ci)
```

```
      (Intercept)             woman               age metropolitan_area
     0.4087823204      1.1522428462      0.9852994599      0.9939481241
        non_white             urban       education_1       education_2
     2.1195817384      0.5079970362      4.3928268979      2.1066156679
      education_3       education_4       education_6       education_7
     1.4283515652      1.4183202231      0.5884957715      0.2878207704
           work_1            work_2            work_3            work_5
     1.3178767291      0.8509202369      1.3353505274      0.9425452000
           work_6            work_7            work_8            work_9
     1.3557118379      0.6017205252      0.4088619493      0.6661569647
          work_10           work_11           work_12     work_permit_0
     1.1855283513      0.7896671139      0.0006417893      2.3082679877
                   Odds_Ratio       CI_Lower      CI_Upper
(Intercept)        0.4087823204  3.415635e-01 4.892296e-01
woman              1.1522428462  1.052338e+00 1.261632e+00
age                0.9852994599  9.821559e-01 9.884531e-01
metropolitan_area  0.9939481241  9.121938e-01 1.083030e+00
non_white          2.1195817384  1.961148e+00 2.290815e+00
urban              0.5079970362  4.602655e-01 5.606785e-01
education_1        4.3928268979  3.485461e+00 5.536407e+00
education_2        2.1066156679  1.906100e+00 2.328224e+00
education_3        1.4283515652  1.249887e+00 1.632298e+00
education_4        1.4183202231  1.224321e+00 1.643060e+00
education_6        0.5884957715  4.847792e-01 7.144020e-01
education_7        0.2878207704  2.455348e-01 3.373893e-01
work_1             1.3178767291  1.148260e+00 1.512549e+00
work_2             0.8509202369  7.423226e-01 9.754051e-01
work_3             1.3353505274  1.148188e+00 1.553021e+00
work_5             0.9425452000  7.761927e-01 1.144550e+00
work_6             1.3557118379  1.137035e+00 1.616445e+00
work_7             0.6017205252  5.029206e-01 7.199300e-01
work_8             0.4088619493  3.279191e-01 5.097846e-01
work_9             0.6661569647  5.630281e-01 7.881758e-01
work_10            1.1855283513  9.925781e-01 1.415987e+00
work_11            0.7896671139  6.654939e-01 9.370095e-01
work_12            0.0006417893 1.306108e-105 3.153595e+98
work_permit_0      2.3082679877  2.105981e+00 2.529985e+00
```

In [60]:

```
model <- glm(poverty ~ .,data = df1[1:24], family = binomial(link = "logit" ))
interaction_model <- glm(poverty ~ .+(df$woman*df$non_white),data = df1[1:24],
                      family = binomial(link = "logit" ))
anova(interaction_model, model, test = "LRT")
```

A anova: 2 × 5

	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	20726	18153.09	NA	NA	NA
2	20727	18154.83	-1	-1.733961	0.1879059

based on this analysis, Model 2 (without the interaction term) does not significantly worsen the fit compared to Model 1, suggesting that the interaction term (df$woman * df$

non_white) may not be necessary in explaining the variation in the response variable (poverty).

In [61]:
```
install.packages("ordinal")
library(ordinal)
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)

also installing the dependencies ‘ucminf’, ‘numDeriv’



Attaching package: ‘ordinal’


The following object is masked from ‘package:dplyr’:

    slice
```

In [62]:
```
df$poverty <- as.factor(df$poverty)
df$woman <- as.factor(df$woman)
df$age <- as.numeric(df$age)
df$education <- as.factor(df$education)
df$work <- as.factor(df$work)
df$metropolitan_area <- as.factor(df$metropolitan_area)
df$non_white <- as.factor(df$non_white)
df$urban <- as.factor(df$urban)
df$work_permit <- as.factor(df$work_permit)
model_cloglog <- glm(poverty ~ woman + age + education + work +
                 metropolitan_area + non_white + urban + work_permit,
                     family = binomial(link = cloglog), data = df)
summary(model_cloglog)
```

```
Call:
glm(formula = poverty ~ woman + age + education + work + metropolitan_area +
    non_white + urban + work_permit, family = binomial(link = cloglog),
    data = df)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.935684   0.102939   9.090  < 2e-16 ***
woman1              0.098265   0.038891   2.527  0.01152 *
age                -0.016635   0.001347 -12.349  < 2e-16 ***
education2         -0.493887   0.074571  -6.623 3.52e-11 ***
education3         -0.819951   0.087367  -9.385  < 2e-16 ***
education4         -0.823583   0.092642  -8.890  < 2e-16 ***
education5         -1.097914   0.080794 -13.589  < 2e-16 ***
education6         -1.574094   0.116652 -13.494  < 2e-16 ***
education7         -2.337803   0.106348 -21.983  < 2e-16 ***
work2              -0.193134   0.061706  -3.130  0.00175 **
work3              -0.029214   0.061011  -0.479  0.63205
work4              -0.153550   0.055362  -2.774  0.00554 **
work5              -0.249172   0.085302  -2.921  0.00349 **
work6               0.031828   0.076460   0.416  0.67722
work7              -0.472983   0.087064  -5.433 5.55e-08 ***
work8              -1.062738   0.103159 -10.302  < 2e-16 ***
work9              -0.511679   0.080541  -6.353 2.11e-10 ***
work10             -0.166799   0.081103  -2.057  0.03972 *
work11             -0.210487   0.073150  -2.877  0.00401 **
work12             -7.725666 114.217763  -0.068  0.94607
metropolitan_area1 -0.014745   0.036857  -0.400  0.68912
non_white1          0.626344   0.033438  18.732  < 2e-16 ***
urban1             -0.489622   0.039576 -12.372  < 2e-16 ***
work_permit1       -0.998529   0.044109 -22.638  < 2e-16 ***
work_permit2       -0.343475   0.039339  -8.731  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22125  on 20750  degrees of freedom
Residual deviance: 17903  on 20726  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 17953

Number of Fisher Scoring iterations: 9
```

Significance of Predictors: Many predictors are highly significant, suggesting they have a strong association with poverty status.

The coefficients show the direction and magnitude of the effect each predictor has on the log odds of being in poverty. Positive coefficients indicate increased odds, while negative coefficients indicate decreased odds.

The residual deviance is lower than the null deviance, indicating that the model with predictors fits the data better than the model with only the intercept.