How do literacy and age of marriage affect family size

ass1

Sakura Hu and Zhanyi Wang

Introduction

Method

Result

Portugal fertility survey 1979

- More information
- data source
- data dictionary

File `portugal.RData` on the course web site. Code in `Assignment1.Rmd`

```r
head(portugal)
```

```
##   age ageMarried monthsSinceM pregnancies children sons region literacy
## 1  43     22to25          242           3        3    2  lt10k      yes
## 2  32     22to25          124           1        1    0  lt10k      yes
## 3  22     15to18           59           1        1    1  lt10k      yes
## 4  28     22to25           63           1        1    0  lt10k      yes
## 5  30     15to18          169           2        2    2  lt10k      yes
## 6  37     18to20          226           2        2    1  lt10k      yes
```

```r
table(portugal$region)
```

```
## 
##  lt10k lisbon  porto   20k+ 10-20k 
##   3502    470    160    583    433
```

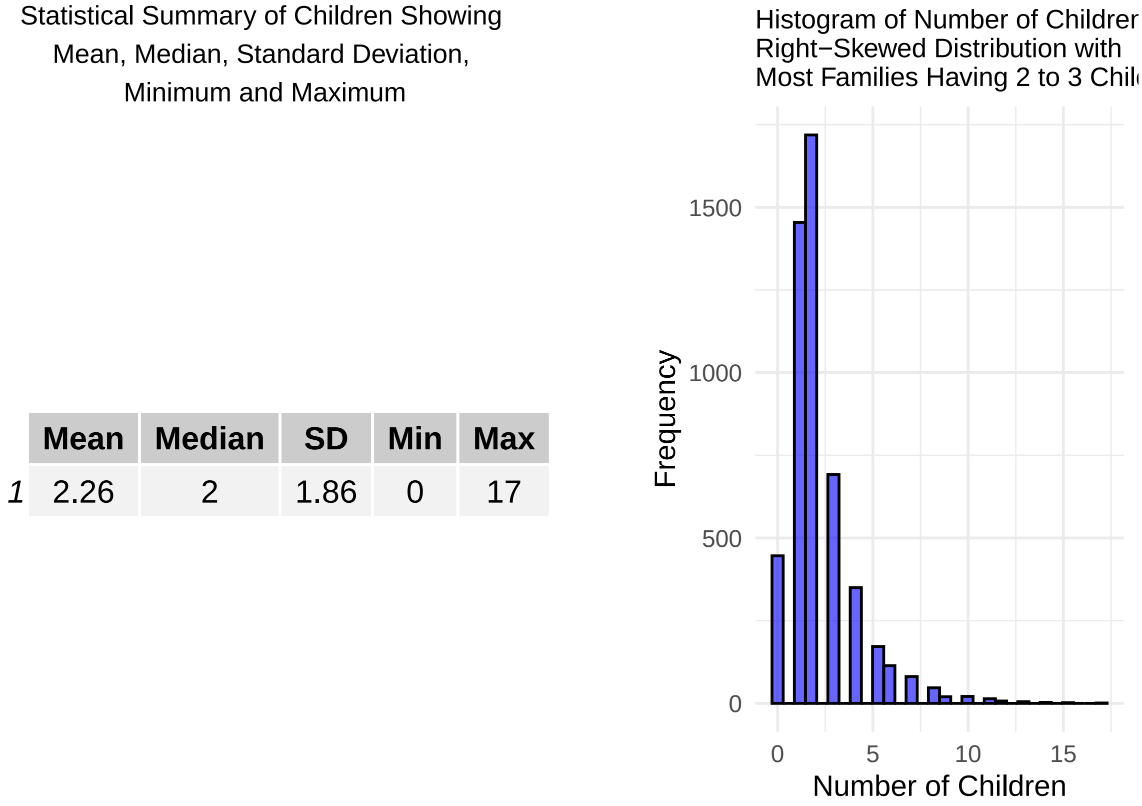
Figure 1 presents the statistical summary of the response variable, “Children.” The number of children per family ranges from 0 to 17, indicating that some families have no children, while others have as many as 17. The distribution is right-skewed, with most families having 2 to 3 children. The mean number of children is 2.26, the median is 2, and the standard deviation is 1.86, reflecting moderate variability. The histogram shows that the data is concentrated around 2 to 3 children, with fewer families having very high numbers of children.

Figure 2 provides the statistical summary of the independent variables, “Age Married” and “Literacy.” Among the 5,148 samples, the majority of individuals marry between ages 20 to 25. Specifically, 1,126 individuals married between 20 to 22, and 1,468 individuals married between 22 to 25. In contrast, marriage before age 15 and after 30 is uncommon, with only 52 individuals marrying between 0 to 15 and 217 marrying after 30. The distribution of “Age Married” follows an approximately bell-shaped pattern, slightly left-skewed, with the peak occurring in the 22 to 25 category.

For “Literacy,” the majority of individuals are literate. Out of the 5,148 samples, 4,567 individuals reported being literate, while only 581 reported otherwise. The distribution is highly skewed, with literacy being the dominant category.

These statistical summaries provide key insights into the dataset, highlighting the central tendency, spread, and shape of each variable while contextualizing the findings within the sample population.

Figure 1: Statistical Summary and Histogram of response variable 'Children'

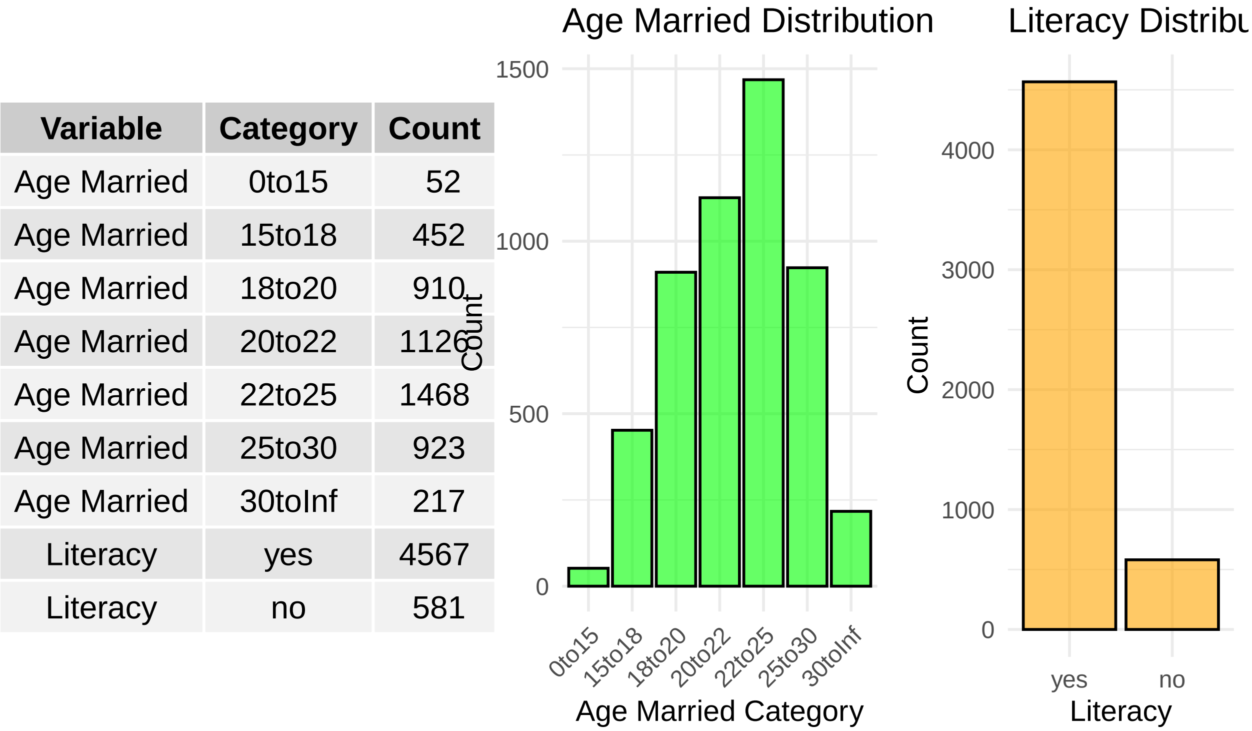


Statistical Summary of Children Showing Mean, Median, Standard Deviation, Minimum and Maximum

	Mean	Median	SD	Min	Max
1	2.26	2	1.86	0	17

Histogram of Number of Childrer
Right−Skewed Distribution with
Most Families Having 2 to 3 Chil

Figure 2: Statistical Summary and Histogram of independnet variables 'Age Married' and 'Literacy'.
Out of 5148 samples, most samples married between 20 to 25,
and there are 4567 samples who can read literature



Variable	Category	Count
Age Married	0to15	52
Age Married	15to18	452
Age Married	18to20	910
Age Married	20to22	1126
Age Married	22to25	1468
Age Married	25to30	923
Age Married	30toInf	217
Literacy	yes	4567
Literacy	no	581

To answer the research question, the data was first modeled using a Poisson regression with an offset for the log-transformed years married. The offset is included to account for the duration of marriage, which may influence the number of children over time. The model is specified as follows:

$$Y_i \sim \text{Poisson}\left(\exp\left(\text{offset}\left(\log\left(\max(1, \text{monthsSinceM}_i)/12\right)\right) + \beta_1 \cdot \text{literacy}_i + \beta_2 \cdot \text{ageMarried}_i\right)\right)$$

Where:

- children is the count of children,
- $\text{offset}(\log(\max(1, \text{monthsSinceM})/12))$ adjusts for the number of years married,
- literacy is a binary variable indicating whether the person is literate or not,
- ageMarried is the categorical variable representing the age at which the person married, with "25to30" as the reference category.

Figure 3 shows the summary of this model. According to the results, the variable `literacyno` (indicating illiteracy) has a significant effect on the number of children, with an estimate of 0.159 and a standard error of 0.024. The corresponding z-value is 6.770, and the p-value is less than 0.005, indicating strong evidence against the null hypothesis and suggesting that being illiterate (compared to being literate) is associated with a higher log count of children. Specifically, being illiterate increases the expected log count of children by 0.159, all else being equal. In contrast, the coefficients for the `ageMarried` categories (relative to the reference group of 25-30 years) do not show significant effects. For example, the coefficient for `ageMarried15to18` is 0.062 with a standard error of 0.037, yielding a z-value of 1.702 and a p-value of 0.089, which is above the standard significance level of 0.05, indicating that this result is not statistically significant. Similarly, other age categories, such as `ageMarried20to22` and `ageMarried30toInf`, show p-values of 0.598 and 0.891, respectively, suggesting no significant relationship between age at marriage and the number of children.

Overall, while literacy is a statistically significant predictor, age at marriage does not appear to have a meaningful effect on the number of children in this model.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.789	0.023	-77.858	0.000
literacyno	0.159	0.024	6.770	0.000
ageMarried0to15	0.036	0.081	0.448	0.654
ageMarried15to18	0.062	0.037	1.702	0.089
ageMarried18to20	0.048	0.031	1.557	0.120
ageMarried20to22	0.016	0.030	0.528	0.598
ageMarried22to25	-0.013	0.029	-0.468	0.640
ageMarried30toInf	0.008	0.060	0.136	0.891

Figure 4 presents the means and variances of literacy and age at marriage. In several cases, the variance is at least twice as large as the mean, indicating potential overdispersion. Since the Poisson model assumes equal mean and variance, this violation suggests the need for a more flexible model. To address this, a Negative Binomial model is applied, which allows the variance to exceed the mean by introducing an additional dispersion parameter. The model is specified as follows:

$$Y_i \sim \text{NegBin}\left(\exp\left(\text{offset}\left(\log\left(\max(1, \text{monthsSinceM}_i)/12\right)\right) + \beta_1 \cdot \text{literacy}_i + \beta_2 \cdot \text{ageMarried}_i\right), \theta\right)$$

Where:

- children represents the number of children,
- $\text{offset}\left(\log\left(\max(1, \text{monthsSinceM})/12\right)\right)$ adjusts for the number of years married,
- literacy is a binary variable indicating whether the person is literate or not,
- ageMarried is the categorical variable representing the age at which the person married.

literacy	agemarried	Mean	Variance
yes	25to30	1.97	1.98
no	25to30	3.27	5.49
yes	0to15	2.79	2.64
no	0to15	4.46	2.27
yes	15to18	2.40	4.10
no	15to18	4.31	6.22
yes	18to20	2.15	2.98
no	18to20	4.87	12.98
yes	20to22	2.12	2.68
no	20to22	3.98	7.91
yes	22to25	1.97	1.98
no	22to25	3.92	7.05
yes	30toInf	1.42	1.79
no	30toInf	1.68	3.41

This model shows that the there are The summary of this negative binomial is shown in Figure 5. According to the model results, literacy has a statistically significant effect on the number of children. Specifically, the coefficient for 'literacyno' (indicating illiteracy) is 0.148 with a standard error of 0.027, yielding a z-value of 5.570. The associated p-value is less than 0.05, providing strong evidence against the null hypothesis and suggesting that being illiterate is associated with a higher expected number of children. This suggests that, holding all else constant, being illiterate increases the expected log count of children by 0.148.

In contrast, the effects of age at marriage do not appear to be statistically significant in this model. The coefficients for different ageMarried categories are estimated relative to the reference group 25–30 years old. with all levels in ageMarried has p value greater than 0.05, which suggest that in this population, the timing of marriage alone may not be a key determinant of fertility outcomes once other factors, such as education, are taken into account.

The estimated dispersion measure is 0.2645, indicating that the variance exceeds the mean, which violates the Poisson model's assumption of equidispersion. This justifies the use of a Negative Binomial model, which introduces an additional dispersion parameter to account for the excess variability.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.772	0.025	-71.292	0.000
literacyno	0.148	0.027	5.570	0.000
ageMarried0to15	0.057	0.090	0.631	0.528
ageMarried15to18	0.073	0.040	1.810	0.070
ageMarried18to20	0.059	0.034	1.730	0.084
ageMarried20to22	0.025	0.033	0.754	0.451
ageMarried22to25	-0.011	0.031	-0.339	0.735
ageMarried30toInf	0.011	0.064	0.165	0.869

```r
# Load the necessary library

# Get summary of the model (coefficients and their standard errors)
coef_summary <- summary(portugalNB2)$coefficients

# Get the confidence intervals for the model parameters
conf_intervals <- confint(portugalNB2)
```

```
## Waiting for profiling to be done...
```

```r
# Display results using knitr::kable
knitr::kable(
  rbind(
    coef_summary[, c(1, 2)],  # Estimates and Standard Errors
    CI_lower = conf_intervals[, 1],  # Lower bound of CI
    CI_upper = conf_intervals[, 2]   # Upper bound of CI
  ),
  caption = "Model Coefficients, Standard Errors, and Confidence Intervals"
)
```

```
## Warning in rbind(coef_summary[, c(1, 2)], CI_lower = conf_intervals[, 1], :
## number of columns of result is not a multiple of vector length (arg 2)
```

Table 4: Model Coefficients, Standard Errors, and Confidence Intervals

	Estimate	Std. Error
(Intercept)	-1.7724866	0.0248623
literacyno	0.1476753	0.0265124
ageMarried0to15	0.0569934	0.0902874
ageMarried15to18	0.0730457	0.0403580
ageMarried18to20	0.0586101	0.0338884
ageMarried20to22	0.0245097	0.0325176
ageMarried22to25	-0.0105426	0.0311231

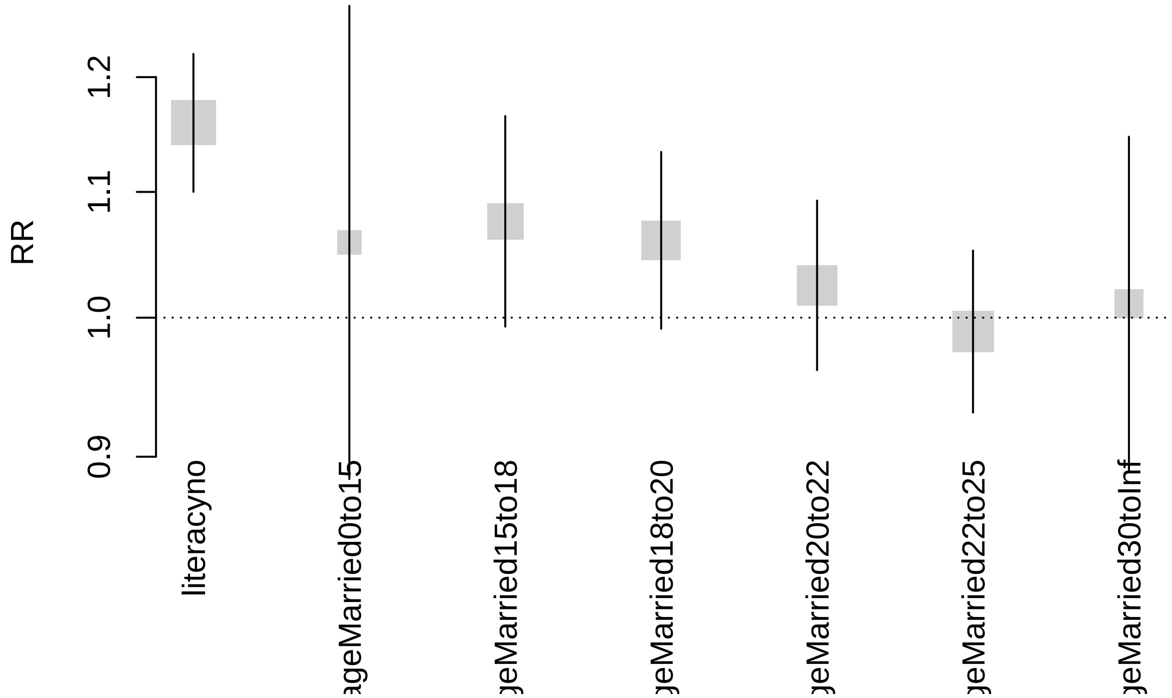
	Estimate	Std. Error
ageMarried30toInf	0.0105638	0.0641347
CI_lower	-1.8218114	0.0952789
CI_upper	-1.7236914	0.1997605

	Estimate	2.5 %	97.5 %
(Intercept)	-1.7724898	-1.8216692	-1.7233104
literacyno	0.1476897	0.0953899	0.1999895
ageMarried0to15	0.0570068	-0.1223318	0.2363453
ageMarried15to18	0.0730426	-0.0067813	0.1528666
ageMarried18to20	0.0586100	-0.0084053	0.1256252
ageMarried20to22	0.0245153	-0.0397481	0.0887786
ageMarried22to25	-0.0105444	-0.0719655	0.0508767
sd	0.2650000	0.2650000	0.2650000

```
##                 2.5 %    97.5 %  Estimate
## (Intercept) 0.1617555 0.1784744 0.1699094
```

```
##           2.5 %    97.5 %  Estimate
## sigma 0.2977129 0.2350557 0.2645356
```

```
##                       2.5 %   97.5 % Estimate             level variable x
## literacyno        1.1000877 1.221390 1.159153        literacyno          1
## ageMarried0to15   0.8848547 1.266612 1.058663   ageMarried0to15          2
## ageMarried15to18  0.9932416 1.165170 1.075776  ageMarried15to18          3
## ageMarried18to20  0.9916299 1.133857 1.060362  ageMarried18to20          4
## ageMarried20to22  0.9610315 1.092839 1.024818  ageMarried20to22          5
## ageMarried22to25  0.9305630 1.052193 0.989511  ageMarried22to25          6
## ageMarried30toInf 0.8904295 1.147020 1.010614 ageMarried30toInf          7
##                        cex
## literacyno        3.091968
## ageMarried0to15   1.669737
## ageMarried15to18  2.502755
## ageMarried18to20  2.731481
## ageMarried20to22  2.789352
## ageMarried22to25  2.853160
## ageMarried30toInf 1.987250
```



```
## [1] 0.2645351
```

Conclusion