Exercise 03

Exercise 3.1

Pakistan was chosen as the focal country, with a focus on **gender inequality**, representing a significant "boy preference" for the third child as the analysis outcome. The dataset comprises **13118** female observations. To handle missing values on crucial information (third child event and duration between the second birth and censoring), listwise deletion was applied, resulting in an analytical sample of 11,135 observations.

The sample statistics by urbanity are detailed in **Table 1**.

Table 1. The Sample Statistics by Urbanity

	Rural (N=5765)	Urban (N=5370)	Overall (N=11135)
Sex Composition			
Boys	1592 (27.6%)	1442 (26.9%)	3034 (27.2%)
Girls	1383 (24.0%)	1267 (23.6%)	2650 (23.8%)
Mix	2790 (48.4%)	2661 (49.6%)	5451 (49.0%)
Cohort			
1968-1983	2984 (51.8%)	2975 (55.4%)	5959 (53.5%)
1984-2001	2781 (48.2%)	2395 (44.6%)	5176 (46.5%)
Education			
No education	3884 (67.4%)	2139 (39.8%)	6023 (54.1%)
Some education	1881 (32.6%)	3231 (60.2%)	5112 (45.9%)
Age at First Birth			
<20	2533 (43.9%)	2096 (39.0%)	4629 (41.6%)
20-24	2325 (40.3%)	2264 (42.2%)	4589 (41.2%)
25-29	763 (13.2%)	853 (15.9%)	1616 (14.5%)
>=30	144 (2.5%)	157 (2.9%)	301 (2.7%)

Table 1 provides a detailed breakdown of sample statistics by urbanity. It categorizes data into urban and rural areas, including information on sex composition, cohort, education, and age at first birth.

Initially, the table explores the sex composition of two children, differentiating between two boys, two girls, and a mix. The proportions of these three categories are similar across both urban and rural settings. Further granularity is achieved by segmenting the data into age cohorts and education levels of the respondents, distinguishing between 1968–1983 and 1984–2001. This is also stratified by educational levels, including those with no formal education and those with some level of education. The cases of the two cohorts are roughly divided in half. Meanwhile, in rural areas, a substantial majority (67.4%) of mothers have no education. In contrast, urban areas exhibit a different pattern, with 39.8% of females having no education and 60.2% having some formal education. Additionally, the sample includes the age of the first birth, allowing for a nuanced examination of variations within different fertility patterns. This reveals a tendency for more respondents in rural areas to have their first birth at a younger age.

Exercise 3.2

A preference for sons and a sex selection against females are widespread in vast regions of the world, including a great number of Asian and East European countries. (Becquet et al., 2022) Pakistan, as a patriarchal society, places higher value on sons compared to daughters due to economic, social, and cultural factors. Sons are often expected to provide financial support, care for parents in old age, carry forward the family name, lineage, and fulfill religious duties. In fact, in any population with a sex preference, reproductive decision-making might be affected, influencing reproductive behaviors. (Becquet et al., 2022) Pakistani families may have a strong preference for having at least one son and may continue to have children until they achieve their desired sex composition. Consequently, my first hypothesis is:

H1: There is a significant "boy preference" for the third child in Pakistan, suggesting that mothers with two girls are more likely to have a third child compared to mothers with two boys or a mix.

Observing that rural areas generally have lower levels of education, income, and limited access to family planning services compared to urban areas, it is suggested that these factors may influence fertility preferences. Rural families, relying more on sons for agricultural labor and security, may face societal pressure and stigma for having only daughters. In contrast, urban families, with better opportunities and incentives to invest in their children's human capital, may challenge traditional gender roles and expectations. Therefore, my second hypothesis is:

H2: The "boy preference" for the third child is stronger in rural areas than in urban areas, indicating that the influence of having two girls on the hazard of having a third child is more pronounced for rural families than for urban families.

Considering the potential impact of education, cohort, and age of the first birth on gender preferences and reproductive behaviors in Pakistani society, these factors are assumed to play significant roles. This is especially negatively associated with women's education and urban residence (Chung & Gupta, 2007, 2011). Thus, my third hypothesis is:

H3: The strength of the "boy preference" for the third child in Pakistan is influenced by residency, education, cohort, and age of the first birth. Specifically, rural families with lower levels of education, belonging to older cohorts, and having younger mothers at the time of their first birth are more likely to exhibit a stronger preference for sons compared to urban families with higher education levels, younger cohorts, and older mothers at the time of their first birth.

Exercise 3.3

Table 2 presents the hazard ratio of Sex Composition in Pakistani females, revealing that families with two girls face a significantly higher risk of having a third child compared to families with two boys. The hazard ratio (HR) is 1.20, indicating a 20% higher hazard for females with two girls. The 95% confidence interval suggests that the true HR lies between 1.14 and 1.28, with a p-value less than 0.001, signifying a statistically significant effect. This

aligns with hypothesis **H1**. Conversely, mothers with a mix of boys and girls exhibit a similar risk of having a third child as mothers with two boys, with an HR of 0.98 and a non-significant p-value of 0.5.

Table 2. Hazard Ratio of Sex Composition (Pakistan)

	Model01		
Characteristic	**HR**[1]	**95% CI**[1]	**p-value**
Sex Composition			
Boys	—	—	
Girls	1.20	1.14, 1.28	<0.001
Mix	0.98	0.93, 1.03	0.5

[1] HR = Hazard Ratio, CI = Confidence Interval

Table 3. Hazard Ratio of Sex Composition, Cohort, Education, Age at First Birth, and Residency (Pakistan)

	Model02		
Characteristic	**HR**[1]	**95% CI**[1]	**p-value**
Sex Composition			
Boys	—	—	
Girls	1.23	1.16, 1.31	<0.001
Mix	1.00	0.96, 1.06	0.8
Cohort			
1968-1983	—	—	
1984-2001	0.78	0.74, 0.81	<0.001
Education			
No education	—	—	
Some education	0.78	0.74, 0.81	<0.001
Age at First Birth			
<20	—	—	
20-24	0.86	0.82, 0.90	<0.001
25-29	0.70	0.65, 0.75	<0.001
>=30	0.49	0.42, 0.57	<0.001
URBAN			
Rural	—	—	
Urban	0.89	0.85, 0.93	<0.001

[1] HR = Hazard Ratio, CI = Confidence Interval

In **Table 3**, the Cox model incorporates covariates, such as sex composition, cohort, education, and age at first birth. Families with two boys or two girls demonstrate a significantly higher risk of having a third child than families with a mix of boys and girls, with an HR of 1.23, a p-value less than 0.001, supporting the hypothesis H1. Additionally, females from the cohort of 1984 to 2001 have a significantly lower risk than those from 1968 to 1983 (HR = 0.78, $p < 0.001$). Respondents with some education exhibit a significantly lower risk (HR = 0.78, $p < 0.001$), and older mothers at the time of their first birth have decreased risks with increasing age ($p < 0.001$). Urban families have a significantly lower risk than rural families (HR = 0.89, $p < 0.001$).

These results support hypotheses **H2** and **H3**, suggesting that the "boy preference" strength in Pakistan is influenced by residency, education, cohort, and age of the first birth. Specifically, rural families with lower education levels, older cohorts, and younger mothers at first birth are more likely to exhibit a stronger preference for sons compared to urban families with higher education levels, younger cohorts, and older mothers at first birth.

Exercise 3.4

The results of separate models for urban and rural settings are displayed in **Table 4**, showcasing hazard ratios for sex composition, cohort, education, and age at first birth.

Table 4. Hazard Ratio of Sex Composition, Cohort, Education, and Age at First Birth (Urban vs Rural)

	Model03_URBAN			Model03_RURAL		
Characteristic	**HR**[1]	**95% CI**[1]	**p-value**	**HR**[1]	**95% CI**[1]	**p-value**
Sex Composition						
Boys	—	—		—	—	
Girls	1.29	1.18, 1.40	<0.001	1.19	1.10, 1.29	<0.001
Mix	0.97	0.90, 1.04	0.4	1.03	0.97, 1.11	0.3
Cohort						
1968-1983	—	—		—	—	
1984-2001	0.77	0.72, 0.82	<0.001	0.79	0.74, 0.84	<0.001
Education						
No education	—	—		—	—	
Some education	0.71	0.67, 0.76	<0.001	0.84	0.79, 0.90	<0.001
AGEKID1						
<20	—	—		—	—	
20-24	0.86	0.81, 0.92	<0.001	0.87	0.82, 0.93	<0.001
25-29	0.68	0.62, 0.76	<0.001	0.72	0.65, 0.79	<0.001
>=30	0.58	0.47, 0.72	<0.001	0.41	0.32, 0.52	<0.001

[1] HR = Hazard Ratio, CI = Confidence Interval

In both urban and rural settings, having two girls as prior children increases the hazard of having a third child compared to having two boys or a mix. However, the effect is more pronounced in urban settings (HR = 1.29) than in rural settings (HR = 1.19). The later cohort (1984-2001) exhibits a lower hazard of having a third child than the earlier cohort (1966-1983), with a slightly stronger effect in urban settings (HR = 0.77) compared to rural settings (HR = 0.79). Similarly, having some education increases the hazard of having a third child compared to having no education. The effect is more robust in urban settings (HR = 0.71) than in rural settings (HR = 0.84). In both settings, older age at first birth decreases the hazard of having a third child compared to younger age, with a stronger overall effect in urban settings. Notably, only the group with an age larger than 30 has a higher HR in urban settings, likely due to the small sample size.

These results indicate that factors influencing the hazard of having a third child differ between urban and rural settings. Moreover, the "boy preference" appears to be more prominent in urban settings than in rural settings.

Exercise 3.5

Omitted variable bias could indeed be a concern in the analysis of Sex Composition and “boy preference”, particularly if the economic status is excluded from the model. Pakistan, characterized by historical instability, political unrest, and instances of conflict, has experienced fluctuations in economic conditions that might significantly influence reproductive behaviors.

The omission of economic status as a covariate may lead to biased estimates. Economic stability can impact families' decisions regarding family size, fertility preferences, and the perceived value of having sons or daughters. During periods of economic uncertainty, families might prioritize economic considerations over gender preferences, potentially altering the observed "boy preference" for the third child.

Unstable governments and conflicts, prevalent in Pakistan's history, can affect economic conditions, exacerbating the importance of economic status in shaping reproductive behaviors. Families facing economic hardships might prioritize factors like financial security and stability over gender preferences.

Therefore, excluding economic status from the model may result in an incomplete understanding of the factors influencing the hazard of having a third child. To mitigate omitted variable bias, it is crucial to consider and include economic status in the analysis, ensuring a more accurate assessment of the relationship between Sex Composition and reproductive behaviors in the context of Pakistan's unique socio-economic challenges.

References:

Becquet, V., Sacco, N., & Pardo, I. (2022). Disparities in Gender Preference and Fertility: Southeast Asia and Latin America in a Comparative Perspective. *Population Research and Policy Review*, *41*(3), 1295–1323. https://doi.org/10.1007/s11113-021-09692-1

Chung, W., & Gupta, M. D. (2007). The Decline of Son Preference in South Korea: The Roles of Development and Public Policy. *Population and Development Review*, *33*(4), 757–783. https://doi.org/10.1111/j.1728-4457.2007.00196.x

Chung, W., & Gupta, M. D. (2011). Factors influencing 'missing girls' in South Korea. *Applied Economics*, *43*(24), 3365–3378. https://doi.org/10.1080/00036841003636284

National Institute of Population Studies (NIPS) [Pakistan] and ICF. 2019. Pakistan Demographic and Health Survey 2017-18. Islamabad, Pakistan, and Rockville, Maryland, USA: NIPS and ICF.

Yoo, S. H., Hayford, S. R., & Agadjanian, V. (2017). Old Habits Die Hard? Lingering Son Preference in an Era of Normalizing Sex Ratios at Birth in South Korea. *Population Research and Policy Review*, *36*(1), 25–54. https://doi.org/10.1007/s11113-016-9405-1

R Packages:

Enzmann JAIRscadwbD, Schwartz M, Jain N, Kraft S (2023). _descr: Descriptive Statistics_. R package version 1.1.8, <https://CRAN.R-project.org/package=descr>.

Göran Broström (2023). eha: Event History Analysis. R package version 2.10.3. URL https://cran.r-project.org/package=eha

Göran Broström (2022). Event History Analysis with R, Second Edition. Chapman and Hall/CRC, Boca Raton.

Rich B (2023). _table1: Tables of Descriptive Statistics in HTML_. R package version 1.4.3, <https://CRAN.R-project.org/package=table1>.

Sjoberg DD, Whiting K, Curry M, Lavery JA, Larmarange J. Reproducible summary tables with the gtsummary package. The R Journal 2021;13:570–80. https://doi.org/10.32614/RJ-2021-053.

Therneau T (2023). _A Package for Survival Analysis in R_. *R package version 3.5-7,* <https://CRAN.R-project.org/package=survival>.

Wickham H, Miller E, Smith D (2023). _haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files_. *R package version 2.5.4,* <https://CRAN.R-project.org/package=haven>.

```r
library(haven)
library(descr)
library(questionr)
library(survival)
library(table1)
library(eha)
library(gtsummary)


#### Step 1: Load data ####
rm(list=ls())
DATA01 <- read_dta("assets/PKIR71FL_SMALL.DTA") # 13118 obs.

DATA02 <- subset(DATA01, !(is.na(DATA01$EVENT)))
DATA02 <- subset(DATA02, !(is.na(DATA02$TIME)))
# 11135 obs. left

# Update TIME=0 to TIME=0.1 because TIME=0 is not allowed in the Weibull
model
# TIME=0 probably means that the event occurred at the same year of the
second birth
DATA02$TIME <- ifelse(DATA02$TIME==0, 0.1, DATA02$TIME)

# Survival plot
SURVIVAL01 <- survfit(Surv(DATA02$TIME,DATA02$EVENT) ~ 1)
plot(SURVIVAL01, xlab = "duration between 2nd and 3rd birth")


#### Step 2: Variable construction ####

#New Variable: GENKIDS01
DATA02$GENKIDS01 <- "NA"
DATA02$GENKIDS01[DATA02$GENKIDS==1] <- "Boys"
DATA02$GENKIDS01[DATA02$GENKIDS==2] <- "Girls"
DATA02$GENKIDS01[DATA02$GENKIDS==3] <- "Mix"
DATA02$GENKIDS01 <- as.factor(DATA02$GENKIDS01)

#New Variable: EDUCATION
DATA02$EDU <- "NA"
DATA02$EDU[DATA02$v106==0] <- "No education"
DATA02$EDU[DATA02$v106==1] <- "Some education"
DATA02$EDU[DATA02$v106==2] <- "Some education"
DATA02$EDU[DATA02$v106==3] <- "Some education"
DATA02$EDU <- as.factor(DATA02$EDU)

#New Variable: URBAN
DATA02$URBAN <- "NA"
```

```r
DATA02$URBAN[DATA02$v025==1] <- "Urban"
DATA02$URBAN[DATA02$v025==2] <- "Rural"
DATA02$URBAN <- as.factor(DATA02$URBAN)

#New Variable: COHORT
DATA02$COHORT <- "NA"
DATA02$COHORT[DATA02$v010 >= 1968 & DATA02$v010 <= 1983] <- "1968-1983"
DATA02$COHORT[DATA02$v010 >= 1984 & DATA02$v010 <= 2001] <- "1984-2001"
DATA02$COHORT <- as.factor(DATA02$COHORT)

#New Variable: AGEKID1
DATA02$AGEKID1 <- "NA"
DATA02$AGEKID1[DATA02$v212 < 20] <- "<20"
DATA02$AGEKID1[DATA02$v212 >= 20 & DATA02$v212 < 25] <- "20-24"
DATA02$AGEKID1[DATA02$v212 >= 25 & DATA02$v212 < 30] <- "25-29"
DATA02$AGEKID1[DATA02$v212 >= 30] <- ">=30"
DATA02$AGEKID1 <- as.factor(DATA02$AGEKID1)
DATA02$AGEKID1 <- factor(DATA02$AGEKID1, levels = c("<20", "20-24", "25-
29", ">=30"))

# Urban and Rural separated datasets
DATA_URBAN <- subset(DATA02,subset = DATA02$v025==1)
DATA_RURAL <- subset(DATA02,subset = DATA02$v025==2)

# Sample Statistics
table1::label(DATA02$GENKIDS01) <- "Sex Composition"
table1::label(DATA02$EDU) <- "Education"
table1::label(DATA02$COHORT) <- "Cohort"
table1::label(DATA02$AGEKID1) <- "Age at First Birth"
table1(~ GENKIDS01 + COHORT + EDU + AGEKID1 | URBAN, data = DATA02)



#### Step 3: Analysis ####

# Model 1: GENKIDS01

# fit Cox model
coxph(Surv(TIME,EVENT) ~ GENKIDS01, data=DATA02)
MODEL01 <- coxph(Surv(TIME,EVENT) ~ GENKIDS01, data=DATA02)

# display regression model results
OUTPUT01 <- tbl_regression(MODEL01, exponentiate = TRUE)
tbl_merge(
  tbls = list(OUTPUT01),
  tab_spanner = c("Model01"))

# Model 2: GENKIDS01 + COHORT + EDU + AGEKID1 + URBAN
MODEL02 <- coxph(Surv(TIME,EVENT) ~ GENKIDS01 + COHORT + EDU + AGEKID1 +
URBAN, data=DATA02)
```

```r
OUTPUT02 <- tbl_regression(MODEL02, exponentiate = TRUE)

tbl_merge(
  tbls = list(OUTPUT02),
  tab_spanner = c("Model02"))


#### Step 4: Sub-datasets Analysis ####

# Model 3: GENKIDS01 + COHORT + EDU + AGEKID1
MODEL03_URBAN <- coxph(Surv(TIME,EVENT) ~ GENKIDS01 + COHORT + EDU +
AGEKID1, data=DATA_URBAN)
OUTPUT03_URBAN <- tbl_regression(MODEL03_URBAN, exponentiate = TRUE)

MODEL03_RURAL <- coxph(Surv(TIME,EVENT) ~ GENKIDS01 + COHORT + EDU +
AGEKID1, data=DATA_RURAL)
OUTPUT03_RURAL <- tbl_regression(MODEL03_RURAL, exponentiate = TRUE)

tbl_merge(
  tbls = list(OUTPUT03_URBAN, OUTPUT03_RURAL),
  tab_spanner = c("Model03_URBAN", "Model03_RURAL"))
```