Investigation of key factors influencing survival in patients with advanced ovarian carcinoma

1 Introduction

In the study by Edmunson et al. (1979), published as "Different Chemotherapeutic Sensitivities and Host Factors Affecting Prognosis in Advanced Ovarian Carcinoma vs. Minimal Residual Disease" [1], the researchers examine how two treatments (single or combination therapy) and patient-specific factors influence the survival outcomes of individuals with advanced ovarian cancer compared to those with minimal residual disease. The ovarian cancer is a severe disease affecting the ovaries, a part of the female reproductive organ and is the fifth leading cause of cancer deaths among women [2]. To perform a survival analysis, the authors generated data from a clinical trial on 26 patients diagnosed with advanced ovarian cancer. This dataset includes the following variables:

- *futime*: The survival or censoring time in days. It represents the time until death, or the end of the study period for patients who are still alive. It is a continuous, numerical variable and the response variable in survival analysis.
- *fustat*: The censoring status, with the value 1 representing death and 0 representing censored data (patients still alive at the end of the study period). It is a categorical and nominal variable.
- *age*: The age of the patient in years. It is a continuous and numerical variable.
- *resid.ds*: The presence of residual disease, coded as 1 for no residual disease or 2 for residual disease. It is a categorical and nominal variable.
- *rx*: The treatment group to which the patient belongs. Either coded as 1 for single therapy or 2 for combination therapy. It is a categorical and nominal variable.
- *ecog.ps*: The Eastern Cooperative Oncology Group (ECOG) Performance Score, which assesses the patient's level of functioning. In our dataset, scores are either 1 (physically restricted) or 2 (ambulatory). This variable is categorical and nominal.

2 Explanatory data analysis

To begin with, the dataset is globally balanced since there are 26 observations for each variable, each one corresponding to a patient. Therefore, we can consider directly number of

counts rather than proportions. Four variables out of six are categorical, which implies that I will study them using frequency tables or bar charts. For the two other variables being continuous, I will analyze them using statistics table or graphical distributions. Secondly, I will investigate the pairwise interactions between variables. No higher order interactions will be considered since the data does not contain enough data-points.

2.1 Univariate analysis

In Table 1, we observe that survival time ranges from 59 to 1227 days, while patient ages range from 38.89 to 74.50 years. The close mean and median values for the *age* variable suggest a symmetric distribution while the *futime* variable shows a significant difference between the mean and median, indicating a right-skewed distribution. The interquartile range (IQR) highlights the dispersion, with survival time showing a wide range and age being less varied.

	Min	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max**	**IQR**
futime (in days)	59.0	368.0	476.0	599.5	794.8	1227.0	426.8
age (in years)	38.89	50.17	56.85	56.17	62.38	74.50	12.21

Table 1: Summary statistics for the *futime* and *age* variables.

As shown in Figure 1, the distribution of *futime* appears visually right-skewed, as indicated by a longer tail on the right side of the graph than on the left side. This highlights the presence of a subset of patients who survive longer than the typical survival period. The *age* variable shows an approximately normal distribution, with a peak around the mean value. This suggests a relatively symmetric spread of ages. However, there is a noticeable gap between 45 and 50 years, indicating a lack of uniformity in the dataset for this variable.

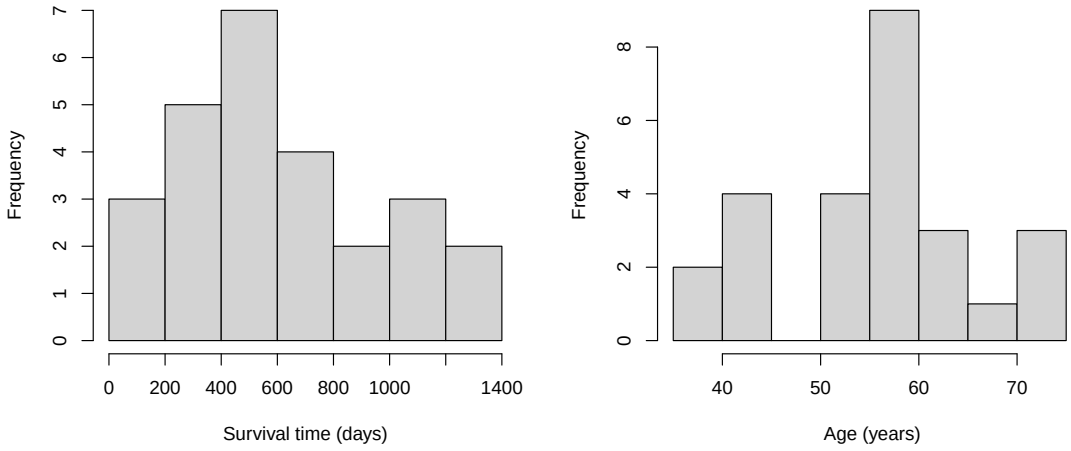

Figure 1: Histograms of the distributions of the *futime* and *age* variables.

In Table 2, we can analyze the occurences' frequencies for each categorical variable in the dataset, while Figure 2 provides a clear visualization. We observe that slightly more patients died during the follow-up period than were censored. Additionally, there are more patients that have residual disease present after an initial treatment compared to those who have no residual disease in the dataset. The treatment groups are balanced, with an equal number of patients receiving single therapy and combination therapy. Finally, there is an even split between patients when considering the ECOG performance status. The balanced distribution of treatments and ECOG performance status ensures comparability between groups.

	0	**1**
fustat (censoring status)	14	12

	1	**2**
resid.ds (residual disease)	11	15
rx (treatment group)	13	13
ecog.ps (ECOG perf. status)	14	12

Table 2: Frequency tables for all four categorical variables of the dataset.

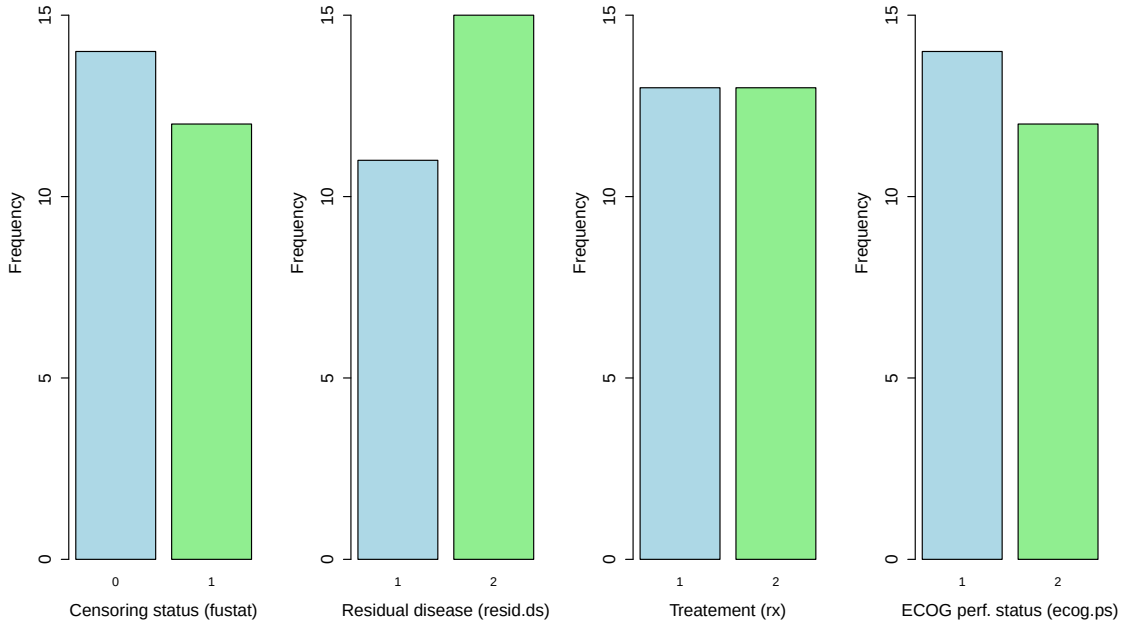

Figure 2: Bar charts of all four categorical variables of the dataset.

2.2 Bivariate analysis

Since survival analysis evaluates which factor influences significantly survival time, which is contained in the variable *futime*, we will now investigate the pairwise interactions between this variable and all the others present in the dataset.

Figure 3 allows us to analyze the interaction between age and survival time, as both variables are numerical and continuous. The smoothed red line shows the trend implied by the points

in the graph. We observe that as age increases, survival time decreases, suggesting that older patients have shorter survival times. This agrees with the biological intuition that older individuals are more vulnerable to diseases. A correlation coefficient of -0.65 describes the relationship between the two variables, indicating a significant negative correlation.

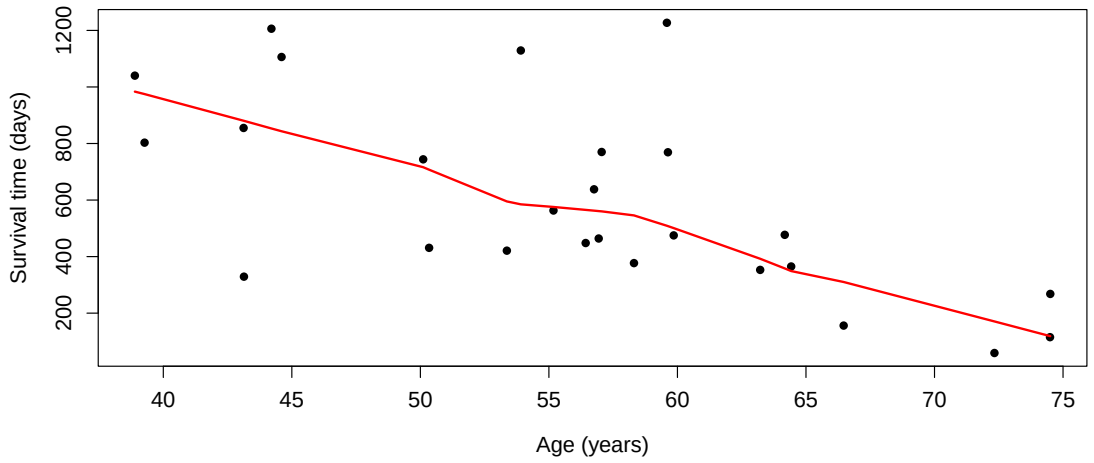

Figure 3: Scatter plot of the follow-up time (*futime*) vs. patients' age.

Figure 4 illustrates the relationships between survival time (*futime*) and four categorical variables: censoring status (*fustat*), presence of residual disease (*resid.ds*), treatment group (*rx*), and ECOG performance status (*ecog.ps*). Since all these variables are categorical and nominal, comparing boxplots allows us to detect their influence on survival time. We observe a significant decrease in the mean survival time for censored patients, suggesting that patients who were censored tended to have shorter survival times. We also see a decrease in the mean survival time for patients with residual disease, suggesting that the presence of residual disease is associated with shorter survival times. On the other hand, there is an increase in the mean survival time for patients receiving combination therapy compared to single therapy, indicating that combination therapy may be more effective. Lastly, mean survival times are similar between the two levels of ECOG performance status, indicating that ECOG performance status may not significantly impact survival time.
To further assess the influence of these factors on survival time, I aimed to perform a t-test to determine whether the differences in means between groups were statistically significant. However, this was not feasible due to insufficient data points. Therefore, the characterization of interactions between survival times and the other categorical factors relies on graphical observations.

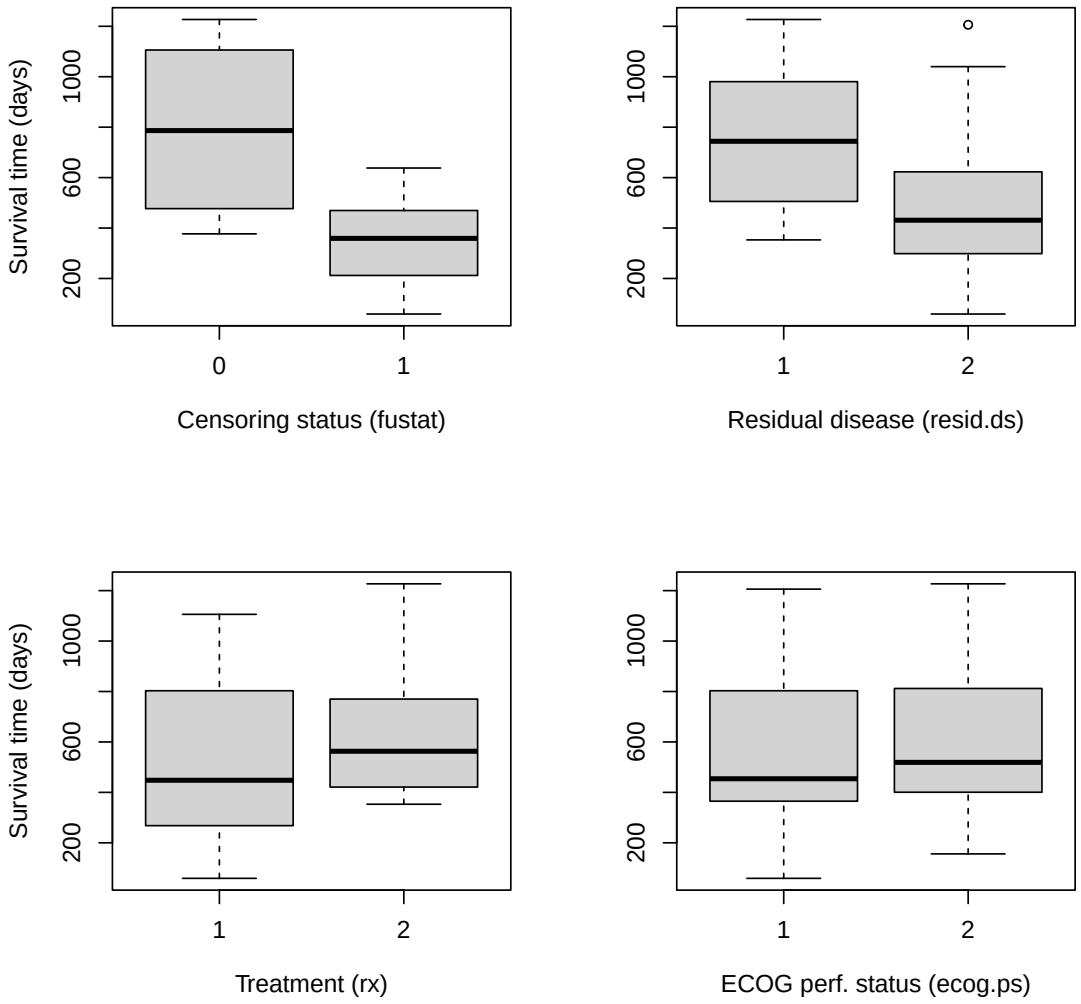

Figure 4: Boxplots of the follow-up time (*futime*) vs. the four categorical variables.

3 Model fitting

3.1 Kaplan-Meier estimator

To characterize the survival time, the survival function S(t) has been defined. It allows to compute the probability of survival past time t. Since we are in the presence of censoring, we typically estimate the survival function using the Kaplan-Meier estimator formulated as follows:

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{r_i}\right)$$

In order to make confidence intervals for the survival function, the variance also needs to be estimated. The variance associated with the Kaplan-Meier estimator can be computed using Greenwood's formula:

$$\mathrm{Var}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - r_i)}$$

To assess whether there is a difference in survival time depending on the treatment taken by patients, I will carry out a log-rank test that allows to compare the observed number of deaths in each group to the expected number using a χ^2 statistical test. I will consider the following hypotheses:

- **Null Hypothesis** (H_0): There is no difference in survival between treatment groups.

- **Alternative Hypothesis** (A) : There is a difference in survival between treatment groups.

As we can see in Table 3, the number of observed death considering patients following a single therapy *(rx = 1)* is greater than the expected number. For patients following a combination therapy *(rx = 2)*, there are less observed death than what is expected. The χ^2 value obtained is equal to 1.1 on 1 degree of freedom. The p-value is equal to 0.3. Thus, the p-value remain bigger than the significance level α of 5%. Thus, we do not reject the null hypothesis H_0, which implies that according to the Kaplan-Meier estimator, there are no difference in survival between the two treatment groups (single or combination therapy).

	N	**Observed**	**Expected**	$(O-E)^2/E$	$(O-E)^2/V$
rx = 1	13	7	5.23	0.596	1.06
rx = 2	13	5	6.77	0.461	1.06

Table 3: Results of the log-rank test.

3.2 Cox regression

The survival function can also be estimated by parametric models, such as the Cox regression or Cox proportional hazards model. Variables considered in this model are *rx* (treatment group), *age* and *resid.ds* (presence of residual disease) since those factors have proven to have an influence on survival time in the exploratory data analysis.

Table 4 reveals that combination therapy is associated with a lower hazard of death (negative coefficient). But because of the large p-value (0.20), the effect of this treatment compared to single therapy is not statistically significant at a 5% significance level. Age is associated with a greater hazard of death (positive coefficient), which means that the older the patients, the more risk of death there is. The p-value associated to this coefficient (0.0014) being smaller than a significance level of 5%, the effect of age on survival time is statistically

significant. Similarly, the presence of residual disease increases greatly the chances of death (positive coefficient greater than 0.5), however this effect is not statistically significant since the p-value (0.36) is greater than a significance level of 5%.

	coef	**exp(coef)**	**se(coef)**	**z**	$Pr(>\|z\|)$
rx	-0.80	0.45	0.63	-1.27	0.20
age	0.15	1.16	0.046	3.19	0.0014
resid.ds	0.70	2.00	0.76	0.92	0.36

Table 4: Results of the Cox regression fitting.

To assess the overall significance of the model, we can look at the results of several tests such as the likelihood ratio test , the Wald test or the log-rank test. All these tests have very low p-values, respectively 0.0008, 0.002 and 0.0001, all smaller than a significance level of 5%. This shows that the Cox regression model is statistically significant. Additionally, parametric models such as the Cox regression allow to estimate more precisely the survival function S(t). Therefore, I will proceed with this model rather than the Kaplan-Meier estimator, which shows a large p-value in the log-rank test.

4 Model assessment

4.1 Cox Proportionnal Hazards Assumptions

For the Cox regression model to be valid, we have to verify the Cox proportional hazards assumptions (Cox PH assumptions):

- **Proportional Hazards Assumption:** the regression effect β is constant over time.
- **Linearity in Log-Hazards:** the log-hazard function is a linear combination of the covariates (including possibly higher order terms, interactions).
- **Exponential Link Function**: the link function is exponential.

To know whether those assumptions hold, we can examine different kinds of residuals such as Schoenfeld, Martingale or Deviance residuals.

4.2 Schoenfeld residuals

The Schoenfeld residuals allow to verify the first Cox PH assumption by representing the difference between the observed and expected covariates. As we can observe in Figure 5, all three plots for the *rx*, and *age* variables show no clear trend around zero and no random patterns. On the other hand, the *resid.ds* variable show some extreme residuals values but

overall do not reveal any systemic trend. Thus, as all Schoenfeld residuals are randomly scattered around zero, the proportional hazards assumption holds.

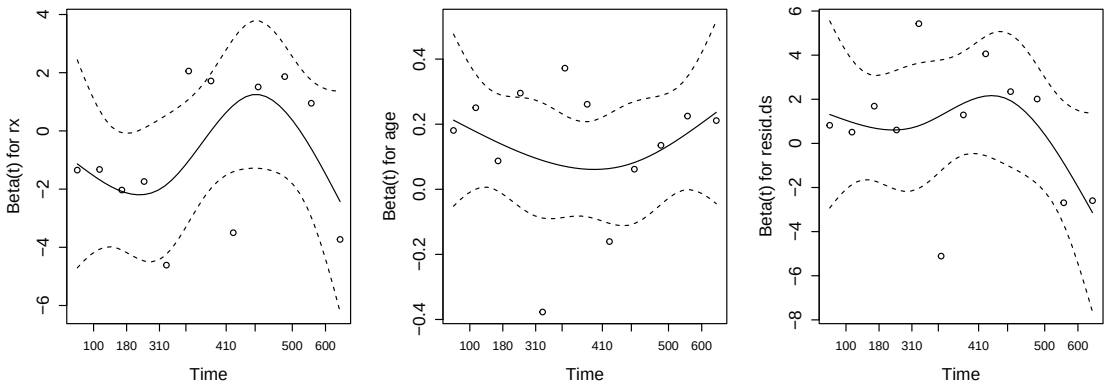

Figure 5: Schoenfeld residuals' plots.

4.3 Martingale residuals

To further test the Cox PH assumptions, we can analyze Martingale residuals. Those help to determine if a variable's relationship with the hazard is linear, as stated in the second Cox PH assumption. For the *rx* variable describing the treatment group of patients, we observe that for each level, the median is near zero and that the box plot is not widespread. For the *age* variable, points do not reveal any particular trend or pattern as we can see in Figure 6. For the *resid.ds*, we can make the same observations as for the rx variable: for each level, the median is near zero and the box plot is not widespread. Therefore, the relationships between all three variables and the hazard appear to be linear. However, we cannot affirm that the second Cox PH assumption holds since Martingale residuals can be difficult to interpret due to their asymmetry arising from their bounds. To further verify the second Cox PH assumption, we will plot other residuals, which are more easily interpretable.

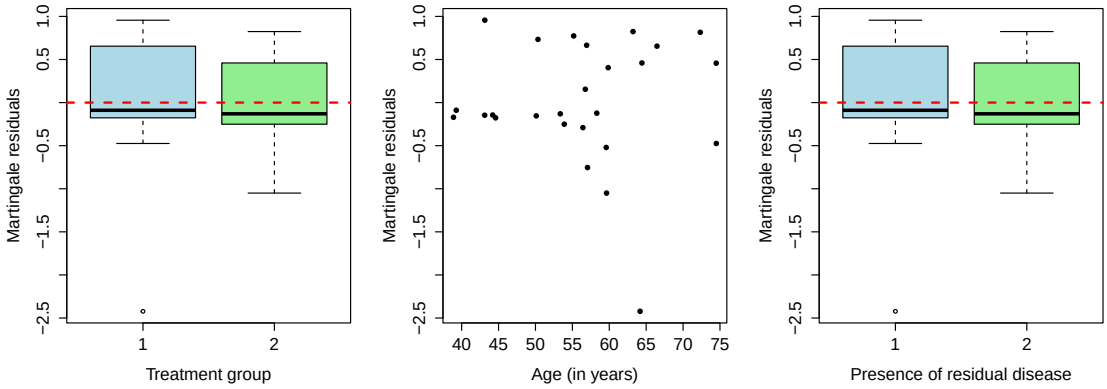

Figure 6: Martingale residuals' plots.

4.4 Deviance residuals

Plotting the Deviance residuals for each covariate allows to assess whether the relationship between the variables and the log-hazard is linear. However, unlike Martingale residuals, Deviance residuals are symmetric and normalized, which makes them easier to analyze. In Figure 7, box plots of each level for treatment groups are around the zero line and have a similar spread of residuals. The Deviance residuals in function of patients' age do not reveal any trend or pattern and are randomly scattered around zero. The Deviance residuals of the *resid.ds* variable are around zero. However, there is a difference in the spread of residuals between the two levels of this covariate and the box plot associated with the level 1 reveals some outliers. However, we can consider these differences non-significant given the small magnitude of the difference in spread and the few number of outliers. Therefore, as all residuals show no significant deviation from zero or systemic patterns, the second Cox PH assumption holds.

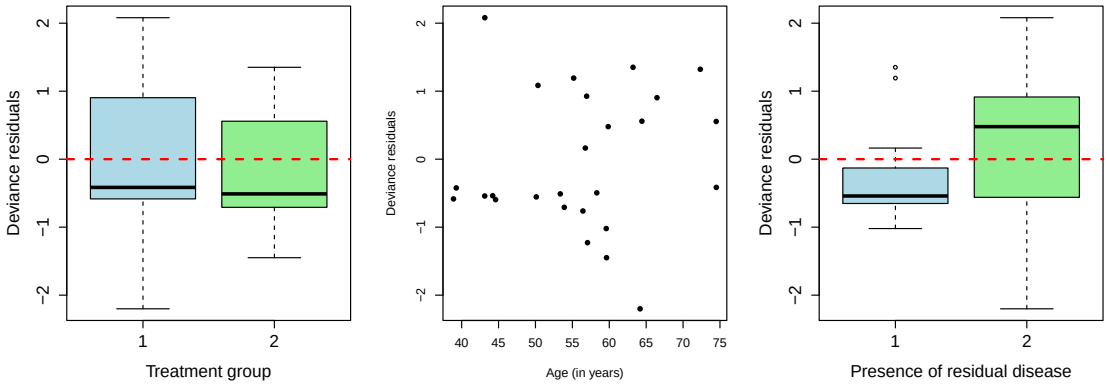

Figure 7: Deviance residuals' plots.

5 Final estimated model

As the main Cox proportional hazards assumptions hold, we can formulate a final Cox proportional hazards model estimating survival time in patients with advanced ovarian cancer compared to those with minimal residual disease. The Cox regression general formula is as follows:

$$\boxed{\widehat{h(t)} = \quad \mathrm{h}_0(t) \cdot \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}$$

The final mathematical model is obtained by substituting the betas and variables with their respective values:

$$\boxed{\widehat{h(t)} = h_0(t) \cdot \exp(-0.80 rx + 0.15 age + 0.70 resid.ds)}$$

6 Conclusion

In the study performed by Edmunson et al., the authors concluded that survival times were comparable between the two treatments for patients with advanced ovarian carcinoma. Based on this survival analysis, we obtain a similar result since neither the type of treatment nor the presence of residual disease significantly influences survival. The only host factor showing a statistically significant effect on survival times is patients' age, reflecting that older individuals tend to be more vulnerable to diseases. However, I have decided to retain all variables selected for the model fitting in the final mathematical model, as this model exhibits statistical significance. Despite the treatment group and presence of residual disease not being statistically significant, the coefficients associated with these factors indicate their substantial contribution to survival time estimation. Finally, we could discuss the conclusions drawn from this analysis since we considered clinical data from only 26 patients, which could bias our final results.

References

[1] J. Edmunson, T. Fleming, D. Decker, G. Malkasian, J. Jefferies, M. Webb, and L. Kvols. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma vs. minimal residual disease. *Cancer Treatment Reports*, 63:241–247, 1979.

[2] R. L. Siegel, K. D. Miller, and A. Jemal. Cancer statistics, 2020. *CA: a cancer journal for clinicians*, 70(1):7–30, 2020.