HW 5

March 3, 2025

1 Exercise 1

Use the dataset `hodg` from the package `KMsurv` for this exercise. You can load the data via the following sequence of commands: library(KMsurv); data("hodg")

1. Fit a Weibull regression using the `survreg` function in the `survival` package to the data with main effects for `gtype`, `dtype` and an interaction between them. You'll need to change `gtype`, `dtype` to factors before you run your analysis because they are coded as integers in the dataset.

2. Run a composite score test (see section 5.1.4 in the notes) to determine if the shape parameter (note that this corresponds to 1/`scale` in `survreg`'s parameterization) is different than 1. Hint: This may involve fitting `survreg` with a different distribution than Weibull.

3. Test the proportional hazards assumption with a likelihood ratio test (see class notes, section 5.3.2). Be explicit about what null hypothesis you'll test, and the asymptotic distribution of the test statistic under the null hypothesis of proportional hazards. Hint: You can use the `loglik` element returned from `survreg` as part of this test.

4. Examine the fit of the model using the Cox-Snell residual model checking algorithm outlined in class notes section 5.4.2. You will have to define a cumulative hazard function manually; use the `pweibull(q, shape, scale, lower.tail=FALSE)` function for a shortcut.

5. Fit a log-logistic AFT model using the `survreg` function with the `dist = "loglogistic"` option specified with main effects for `gtype`, `dtype` and an interaction between them.

6. Examine the Cox-Snell residuals for the log-logistic model. R implements the survival function of a logistic distribution as `plogis(q, location, scale, lower.tail=FALSE)` and the AFT model parameters returned in the `survreg` object correspond directly to the location and scale parameters in the `plogis` R function.

2 Exercise 2

Use the dataset `kidtran` from the package `KMsurv` for this exercise. You can load the data via the following sequence of commands: library(KMsurv); data("kidtran")

1. Create a new categorical predictor variable `age_gp` which is a transformation of the `age` variable into age categories: $(0, 18], (18, 40], (40, 55], (55, \infty]$. Use the base `R` function `cut` for this task.

2. Fit a Weibull regression using the `survreg` function in the `survival` package to the data with a main effect for `age_gp`

3. Use the `residuals` function with the `type` argument set to `"dfbetas"`. This function returns a matrix where each row corresponds to equation 5.47 in the class notes for the datapoint with index equal to the row index. Each column is then scaled by the estimated asymptotic standard error of the parameter, which is the diagonal of the inverse of the observed Fisher information. Make a plot of the scaled influence statistic for the coefficient corresponding to the $(18, 40]$ age group. Do you see any outliers? If so, what is anomalous about this (these) datapoint(s)?

4. Refit the model without the outlier datapoint(s). How do the estimated regression coefficients change? If they change, why do the regressionc coefficients change? See if you can glean any insight from looking at the number of failures within each age group before and after you remove the outliers, and thinking about how you would solve the score equations for the Weibull rate parameter (inverse of the base R's scale parameter) in both scenarios.

5. How might you modify your analysis to avoid this issue?