BIOS 522: Project 1

Falcons Group

2020-09-13

Contents

1 Standard Regression techniques questions

- Dr. Blum is interested in survival times of patients and would like to know the impact of treatment, age, and serum bilirubin as a categorical variable (<1.1, 1.1-3.3, and >3.3) on survival.
  - Use a linear model after excluding all censored observations
  - Use a linear model after treating censored times as death times
  - Use a logistic regression by defining a new outcomes as dead=1 and otherwise (survived or censored) as 0.
- For each of above models perform univariate and multivariate analyses(for the three covariates above).
- Interpret the estimates of coefficients of treatment, age, and serum bilirubin regardless of their significance.
- Now comment on the appropriateness of the data analyses Dr. Blum suggested. A critique.

2 Survival Regression techniques questions

- Now perform a parametric survival analysis (Weibull) and conduct the same univariate and multivariate analyses. Report same results for interpreting coefficients.
- Dr. Blum wants to know how to derive the i) estimates ii) standard errors of the coefficients that R outputs from the regression. What is the procedure? How can Dr. Blum recreate them herself? Attach this techinical section as an appendix to the report.

3 Analysis of the linear regression

3.1 Data input and cleaning

```
##
##   0   1
## 187 125
```

```
##
##   < 1.1 1.1-3.3    >3.3
##     116     113      83
```

3.2 Excluding censored observations

3.2.1 Regression on treatment (drug)

```
##
## Call:
## lm(formula = time ~ drug, data = data[data$dead == 1, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1477.7  -818.0  -327.7   737.3  2672.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1608.29     305.32   5.268 5.97e-07 ***
```

```
## drug          -89.63     195.46  -0.459    0.647
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1092 on 123 degrees of freedom
## Multiple R-squared:  0.001707,   Adjusted R-squared:  -0.00641
## F-statistic: 0.2103 on 1 and 123 DF,  p-value: 0.6474
```

3.2.2 Regression on age

```
##
## Call:
## lm(formula = time ~ ageinyear, data = data[data$dead == 1, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1568.2  -819.6  -286.4   729.0  2591.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2444.568    518.329   4.716  6.4e-06 ***
## ageinyear    -18.199      9.566  -1.902   0.0594 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1077 on 123 degrees of freedom
## Multiple R-squared:  0.02858,    Adjusted R-squared:  0.02069
## F-statistic: 3.619 on 1 and 123 DF,  p-value: 0.05945
```

3.2.3 Regression on serum bilirubin

```
##
## Call:
## lm(formula = time ~ bili_cat, data = data[data$dead == 1, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1739.10  -733.53   -67.53   474.41  2841.47
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2108.6      240.2   8.779 1.22e-14 ***
## bili_cat1.1-3.3   -259.5      279.5  -0.928    0.355
## bili_cat>3.3     -1111.1      272.1  -4.083 7.97e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 990.3 on 122 degrees of freedom
## Multiple R-squared:  0.1853, Adjusted R-squared:  0.172
## F-statistic: 13.88 on 2 and 122 DF,  p-value: 3.712e-06
```

3.2.4 Regression on all three covariates

```
##
## Call:
## lm(formula = time ~ drug + ageinyear + bili_cat, data = data[data$dead ==
##     1, ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2102.7  -624.4  -113.8   442.7  2865.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3675.04     627.60   5.856 4.24e-08 ***
## drug             -144.27     174.43  -0.827  0.40982
## ageinyear         -23.98       8.71  -2.753  0.00682 **
## bili_cat1.1-3.3  -336.07     274.49  -1.224  0.22321
## bili_cat>3.3    -1208.34     268.05  -4.508 1.53e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 967.4 on 120 degrees of freedom
## Multiple R-squared:  0.2354, Adjusted R-squared:  0.2099
## F-statistic: 9.237 on 4 and 120 DF,  p-value: 1.533e-06
```

3.3 Treating censored times as death times

3.3.1 Regression on treatment (drug)

```
##
## Call:
## lm(formula = time ~ drug, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1974.6  -824.6  -157.4   686.3  2540.4
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2034.38     200.66  10.139   <2e-16 ***
## drug          -18.76     127.40  -0.147    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1125 on 310 degrees of freedom
## Multiple R-squared:  6.992e-05,  Adjusted R-squared:  -0.003156
## F-statistic: 0.02168 on 1 and 310 DF,  p-value: 0.883
```

3.3.2 Regression on age

```
##
## Call:
```

```
## lm(formula = time ~ ageinyear, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1989.2  -855.8  -144.4   694.6  2699.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2716.200    305.468   8.892   <2e-16 ***
## ageinyear    -14.191      5.975  -2.375   0.0182 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1115 on 310 degrees of freedom
## Multiple R-squared:  0.01787,    Adjusted R-squared:  0.0147
## F-statistic: 5.641 on 1 and 310 DF,  p-value: 0.01816
```

3.3.3 Regression on serum bilirubin

```
##
## Call:
## lm(formula = time ~ bili_cat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1996.5  -721.5  -132.4   645.0  2664.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2504.20      92.25  27.145  < 2e-16 ***
## bili_cat1.1-3.3  -397.73     131.33  -3.028  0.00267 **
## bili_cat>3.3    -1329.90     142.85  -9.310  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 993.6 on 309 degrees of freedom
## Multiple R-squared:  0.2226, Adjusted R-squared:  0.2176
## F-statistic: 44.24 on 2 and 309 DF,  p-value: < 2.2e-16
```

3.3.4 Regression on all three covariates

```
##
## Call:
## lm(formula = time ~ drug + ageinyear + bili_cat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2095.3  -736.2  -107.8   638.0  2605.0
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3233.048    348.597   9.274  < 2e-16 ***
## drug              -40.120    113.243  -0.354  0.72337
```

```
## ageinyear         -13.394      5.339  -2.509  0.01262 *
## bili_cat1.1-3.3  -400.348    130.840  -3.060  0.00241 **
## bili_cat>3.3    -1322.395    141.897  -9.319  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 986.8 on 307 degrees of freedom
## Multiple R-squared:  0.2382, Adjusted R-squared:  0.2283
## F-statistic:    24 on 4 and 307 DF,  p-value: < 2.2e-16
```

4 Analysis of the logistic regression

4.0.1 Regression on treatment (drug)

```
##
## Call:
## glm(formula = dead ~ drug, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0296  -1.0296  -0.9936   1.3328   1.3730
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.26747    0.36312  -0.737    0.461
## drug        -0.09074    0.23118  -0.393    0.695
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 420.12  on 311  degrees of freedom
## Residual deviance: 419.97  on 310  degrees of freedom
## AIC: 423.97
##
## Number of Fisher Scoring iterations: 4
```

4.0.2 Regression on age

```
##
## Call:
## glm(formula = dead ~ ageinyear, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6231  -0.9929  -0.7585   1.2390   1.7978
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.95147    0.61139  -4.828 1.38e-06 ***
## ageinyear    0.05045    0.01177   4.287 1.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 420.12  on 311  degrees of freedom
## Residual deviance: 400.30  on 310  degrees of freedom
## AIC: 404.3
##
## Number of Fisher Scoring iterations: 4
```

4.0.3 Regression on serum bilirubin

```
##
## Call:
## glm(formula = dead ~ bili_cat, family = "binomial", data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6021  -1.0517  -0.5630   0.8056   1.9598
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -1.7619     0.2625  -6.711 1.93e-11 ***
## bili_cat1.1-3.3   1.4587     0.3243   4.499 6.84e-06 ***
## bili_cat>3.3      2.7208     0.3593   7.573 3.64e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 420.12  on 311  degrees of freedom
## Residual deviance: 348.73  on 309  degrees of freedom
## AIC: 354.73
##
## Number of Fisher Scoring iterations: 4
```

4.0.4 Regression on all three covariates

```
##
## Call:
## glm(formula = dead ~ drug + ageinyear + bili_cat, family = "binomial",
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3879  -0.7720  -0.4455   0.8424   2.2815
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -5.15026    0.94393  -5.456 4.86e-08 ***
## drug             0.07332    0.27610   0.266    0.791
## ageinyear        0.06316    0.01389   4.547 5.44e-06 ***
## bili_cat1.1-3.3  1.56164    0.33894   4.607 4.08e-06 ***
## bili_cat>3.3     2.91444    0.38083   7.653 1.97e-14 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 420.12  on 311  degrees of freedom
## Residual deviance: 325.33  on 307  degrees of freedom
## AIC: 335.33
##
## Number of Fisher Scoring iterations: 4
```

5 Critique of Dr. Blum's analyses

6 Parametric survival analysis

6.1 Parametric survival analysis on drug

```
## Call:
## flexsurvreg(formula = Surv(time, dead) ~ drug, data = data, dist = "Weibull")
##
## Estimates:
##        data mean  est        L95%       U95%       se         exp(est)
## shape         NA     1.1269     0.9659     1.3148     0.0887         NA
## scale         NA  4334.5558  2642.1801  7110.9361  1094.7492         NA
## drug      1.4936     0.0418    -0.2696     0.3531     0.1589     1.0426
##        L95%       U95%
## shape         NA         NA
## scale         NA         NA
## drug      0.7637     1.4235
##
## N = 312,  Events: 125,  Censored: 187
## Total time at risk: 625985
## Log-likelihood = -1188.718, df = 3
## AIC = 2383.436
```

6.2 Parametric survival analysis on age

```
## Call:
## flexsurvreg(formula = Surv(time, dead) ~ ageinyear, data = data,
##     dist = "Weibull")
##
## Estimates:
##            data mean  est        L95%       U95%       se         exp(est)
## shape             NA   1.14e+00   9.81e-01   1.33e+00   8.95e-02         NA
## scale             NA   2.65e+04   1.11e+04   6.36e+04   1.18e+04         NA
## ageinyear   5.00e+01  -3.44e-02  -5.00e-02  -1.88e-02   7.96e-03   9.66e-01
##            L95%       U95%
## shape             NA         NA
## scale             NA         NA
## ageinyear   9.51e-01   9.81e-01
##
## N = 312,  Events: 125,  Censored: 187
```

```
## Total time at risk: 625985
## Log-likelihood = -1178.733, df = 3
## AIC = 2363.467
```

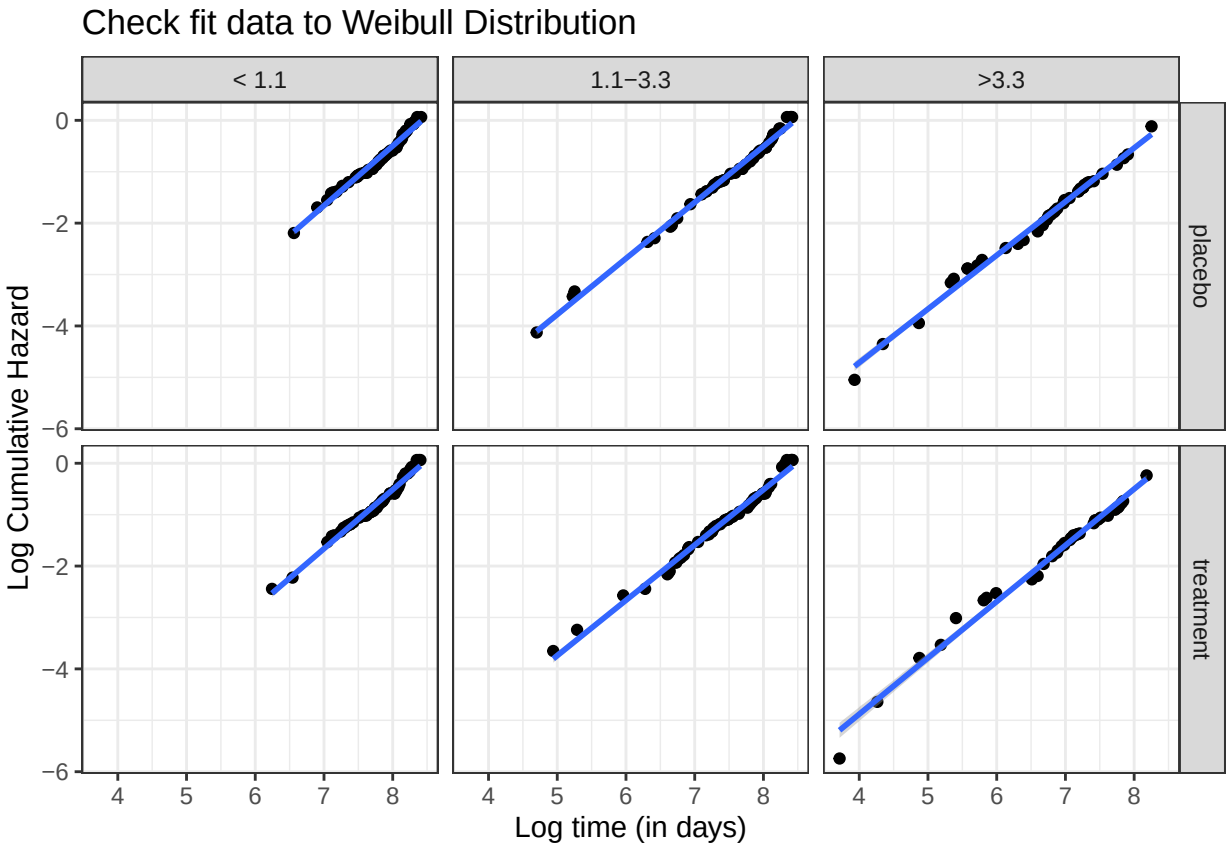
6.3 Parametric survival analysis on serum bilirubin

```
## Call:
## flexsurvreg(formula = Surv(time, dead) ~ bili_cat, data = data,
##     dist = "Weibull")
##
## Estimates:
##                   data mean  est        L95%       U95%       se
## shape                    NA   1.34e+00   1.15e+00   1.55e+00   1.01e-01
## scale                    NA   1.08e+04   7.18e+03   1.62e+04   2.24e+03
## bili_cat1.1-3.3    3.62e-01  -9.43e-01  -1.38e+00  -5.09e-01   2.22e-01
## bili_cat>3.3       2.66e-01  -1.90e+00  -2.35e+00  -1.45e+00   2.30e-01
##                   exp(est)   L95%       U95%
## shape                    NA         NA         NA
## scale                    NA         NA         NA
## bili_cat1.1-3.3    3.90e-01   2.52e-01   6.01e-01
## bili_cat>3.3       1.49e-01   9.51e-02   2.34e-01
##
## N = 312,  Events: 125,  Censored: 187
## Total time at risk: 625985
## Log-likelihood = -1136.177, df = 4
## AIC = 2280.354
```

6.4 Parametric survival analysis on all three covariates

```
## Call:
## flexsurvreg(formula = Surv(time, dead) ~ drug + ageinyear + bili_cat,
##     data = data, dist = "Weibull")
##
## Estimates:
##                   data mean  est        L95%       U95%       se
## shape                    NA   1.37e+00   1.18e+00   1.58e+00   1.02e-01
## scale                    NA   6.10e+04   2.31e+04   1.61e+05   3.02e+04
## drug               1.49e+00  -1.67e-01  -4.31e-01   9.75e-02   1.35e-01
## ageinyear          5.00e+01  -2.92e-02  -4.14e-02  -1.71e-02   6.21e-03
## bili_cat1.1-3.3    3.62e-01  -9.60e-01  -1.39e+00  -5.34e-01   2.17e-01
## bili_cat>3.3       2.66e-01  -1.90e+00  -2.34e+00  -1.46e+00   2.25e-01
##                   exp(est)   L95%       U95%
## shape                    NA         NA         NA
## scale                    NA         NA         NA
## drug               8.46e-01   6.50e-01   1.10e+00
## ageinyear          9.71e-01   9.59e-01   9.83e-01
## bili_cat1.1-3.3    3.83e-01   2.50e-01   5.86e-01
## bili_cat>3.3       1.50e-01   9.62e-02   2.33e-01
##
## N = 312,  Events: 125,  Censored: 187
## Total time at risk: 625985
## Log-likelihood = -1124.358, df = 6
## AIC = 2260.716
```

6.5 Checking Weibull distribution is a good fit with log-log plot



The plot of log time against the log cumulative hazard is mostly linear in each of the treatment / serum bilirubin groups and deviates mostly at the earlier time points. Due to the linearity within each of the groups, we conclude the Weibull distribution is a good fit for this dataset. It is harder to visualize along with age (our third covariate), but based on this diagnostic it seems Weibull is a good fit regardless.

7 Technical Appendix: How to derive coefficient estimates and standard errors from parametric survival analyses

8 Code

```
knitr::opts_chunk$set(
  echo = FALSE,           # don't show code
  warning = FALSE,        # don't show warnings
  message = FALSE,        # don't show messages (less serious warnings)
  cache = FALSE,          # set to TRUE to save results from last compilation
  fig.align = "center"    # center figures
)
library(flexsurv)
library(ggplot2)
library(mice)
set.seed(1)             # make random results reproducible
# Reading in data
data <- read.table("data.csv",
                   quote="\"", comment.char="")

# Renaming variables
colnames(data) <- c("caseid", "time", "status", "drug", "age", "sex", "ascites",
                    "hepatomegaly", "spiders", "edema", "bilirubin", "cholesterol",
                    "albumin", "urine_copper", "alk_phosphatase", "sgot",
                    "triglicerides", "platelets", "prothrombin", "hist_stage")

# Combining censored cases
data$dead <- ifelse(data$status == 0 | data$status == 1, 0, 1)
table(data$dead)

# Creating catergorical factor for bilirubin
data$bili_cat <- ifelse(data$bilirubin < 1.1, "< 1.1",
                        ifelse(data$bilirubin >= 1.1 & data$bilirubin <= 3.3, "1.1-3.3", ">3.3"))
data$bili_cat <- factor(data$bili_cat, levels = c("< 1.1", "1.1-3.3", ">3.3"))
table( data$bili_cat)

# Creating age in years
data$ageinyear <- data$age/365.25
# univariate analysis for the treatment (drug)
summary(lm(time ~ drug, data = data[data$dead==1,]))
# univariate analysis for age
summary(lm(time ~ ageinyear, data = data[data$dead==1,]))
# univariate analysis for serum bilirubin
summary(lm(time ~ bili_cat, data = data[data$dead==1,]))
# multivariate analysis for all three covariates
summary(lm(time ~ drug + ageinyear + bili_cat, data = data[data$dead==1,]))
# univariate analysis for the treatment (drug)
summary(lm(time ~ drug, data = data))
# univariate analysis for age
summary(lm(time ~ ageinyear, data = data))
# univariate analysis for serum bilirubin
summary(lm(time ~ bili_cat, data = data))
# multivariate analysis for all three covariates
summary(lm(time ~ drug + ageinyear + bili_cat, data = data))
# univariate analysis for the treatment (drug)
summary(glm(dead ~ drug, family = "binomial", data = data))
```

```r
# univariate analysis for age
summary(glm(dead ~ ageinyear, family = "binomial", data = data))
# univariate analysis for serum bilirubin
summary(glm(dead ~ bili_cat, family = "binomial", data = data))
# multivariate analysis for all three covariates
summary(glm(dead ~ drug + ageinyear + bili_cat, family = "binomial", data = data))
# here we used library "flexsurv" for the analysis
# univariate analysis for the treatment (drug)
flexsurvreg(Surv(time, dead) ~ drug, data = data, dist = "Weibull")
# univariate analysis for age
flexsurvreg(Surv(time, dead) ~ ageinyear, data = data, dist = "Weibull")
# univariate analysis for serum bilirubin
flexsurvreg(Surv(time, dead) ~ bili_cat, data = data, dist = "Weibull")
# multivariate analysis for all three covariates
flexsurvreg(Surv(time, dead) ~ drug + ageinyear + bili_cat, data = data, dist = "Weibull")
# Create cumulative hazard
data$ch <- nelsonaalen(data, time, dead)
data$drug <- ifelse(data$drug == 1, "treatment", "placebo")
ggplot(data, aes(log(time), log(ch))) +
  geom_point() +
  geom_smooth(method="lm") +
  facet_grid(drug ~ bili_cat) +
  theme_bw() +
  ylab("Log Cumulative Hazard") +
  xlab("Log time (in days)") +
  ggtitle("Check fit data to Weibull Distribution")
# this R markdown chunk generates a code appendix
```