Biostatistics M215: Final Project

Joanna Boland, Chad Pickering, and Matt Ponzini

1. INTRODUCTION

Besides skin cancer, breast cancer (BC) is the most commonly diagnosed cancer among American women; in 2017, it is estimated that about 30 percent of newly diagnosed cancers in women will be breast cancers. BC incidence rates have been decreasing steadily in the U.S. since 2000, but an estimated 266,120 new cases of invasive BC are expected to be diagnosed in women in 2018, and about 1 in 8 U.S. women will develop invasive BC over the course of their lifetime.

It is of interest to understand the survival behavior of women with recent BC diagnoses stratified by demographic characteristics. We will primarily explore models that fit data that fail to meet proportional hazards criteria. For this analysis, the SEER BC registry has been reduced to women who were diagnosed with BC for the first time between 2004 and 2015 and live in the greater Los Angeles region, totaling 51,422 women.

Covariates considered for analysis include marital status, age, and year at diagnosis, tumor size (mm), distant metastasis status, stage summary, surgical procedure, behavior, histology type, race, Hispanic ethnicity, total number of malignant tumors, and total number of benign tumors.

2. METHODS

For preliminary survival analysis results with and without race stratification, we fit non-parametric Kaplan-Meier (Product-Limit) survival and Nelson-Aalen cumulative hazard functions (Figures 1-3). A Cox proportional hazards model was fit to the complete data and is found to be in strong violation of the proportional hazards assumption globally and for many individual predictors (Figures 4-6). For variable selection on the Cox model, we use the Minimum-Maximum Concave Penalty (MCP) procedure, which only removes 2 predictors. An accelerated failure time (AFT) model was fit as well, and the Buckley-James procedure yielded the same large set of predictors that MCP found to be significant. We fit several parametric models to check for agreement using the same subset of covariates determined by the MCP and Buckley-James procedures (Figure 7). We endorse the AFT model over the proportional hazards model because BC itself has a well-studied sequence of intermediary stages, and AFT models are known to do well with these underpinnings. Also, unlike proportional hazards models, the regression parameter estimates from AFT models are robust to omitted predictors. It should be noted that competing risks (causes of death of individual unrelated to BC) are not considered at any point in the analysis.

3. RESULTS

Without race stratification, the Kaplan-Meier (Product-Limit) curve shows that for women in the subgroup defined above, survival probability remains relatively high (>85%) that an individual will remain alive approximately 12 years (144 months) after diagnosis (Figure 1). The restricted mean survival for this group of women is 10.9 years (130 months). At approximately 5 years (60 months) after diagnosis, at which point a woman in this subgroup has a 90% chance of surviving until, the rate of decrease of the survival probability per month begins to slow somewhat, which suggests that death by BC occurs more often within the first few years after diagnosis. This slowing rate can also be seen in the Nelson-Aalen cumulative hazard function (Figure 2). When stratifying by race, the estimated survival functions show that Black women have a much lower probability of survival than White women through any number of months (approximately 79% at 12 years versus 86%; Figure 3). Asian/other women have the highest probability of survival after diagnosis at any chosen time (approximately 88% at 12 years).

When a Cox proportional hazards model was fit to all predictors and all subjects in the data, we found a strong pattern in the deviance residuals (Figure 4) as well as a non-linear violation of the Cox-Snell residuals (Figure 5). The hazard ratios and their 95% confidence intervals are therefore not reliable, but nevertheless displayed for completeness in Figure 6. The Minimum-Maximum Concave Penalty (MCP) procedure for variable

selection was performed on the Cox model, which removed only two predictors, surgery on lymph nodes and number of malignant tumors.

In an attempt to generate a model with non-proportional hazards, accelerated failure time (AFT) models were fit to all predictors. The Buckley-James procedure for model selection yielded the same large set of predictors that MCP found to be significant (the same two were removed), which was a notable consistency, and quite convenient. Four AFT models, exponential, Weibull, log-normal, and log-logistic, were fit, and their AICs and BICs found. The Weibull model had the lowest AIC and BIC values (37004 and 37243, respectively) and fits the closest to the non-parametric estimate (Table 1; Figure 7). The Weibull model is also preferred over the others as it is the most flexible in terms of proportional hazards violations.

Using this final Weibull approximation of the AFT model, several coefficient estimates can be interpreted to provide a more thorough exploration of the effect of various demographic, tumor, and treatment-related characteristics on BC survival, marital status, and histology.

The relative risk of death by BC for a Black woman as compared to a White woman is 1.595 (1.465, 1.736), and 0.867 (0.786, 0.957) for Asian/other women as compared to White women adjusting for all other covariates, which agrees with our cursory analysis of the race-stratified Kaplan-Meier curve (Table 2; both $p<0.01$). Women of Hispanic ethnicity also face a greater relative risk of death by BC, 1.154 (1.071, 1.243) than women of non-Hispanic ethnicity ($p<0.001$). This is likely correlated with lack of sufficient health care services.

The relative risk of death increases by 0.6% for each additional year of age at BC diagnosis ($p<0.001$). Interestingly (but not statistically significant at the 0.05 level), for every additional year that passes, the relative risk of death by BC decreases by 0.6%, which aligns with the decreasing BC incidence rates after the turn of the century. The relative risk of death by BC for married or partnered women as compared to single women is 0.79 (0.742, 0.841), giving evidence to the prevalent claim that women who have social support after the diagnosis of BC have better quality of life and health outcomes ($p<0.001$).

Compared to the reference ductal and lobular neoplasm histology group, women whose cancer histology is a cystic, mucinous, or serous neoplasm have a much lower relative risk of death, at 0.473 (0.332, 0.674) ($p<0.001$). All summary stages had substantially higher relative risks of death by BC as compared to the reference in situ/localized only group (all $p<0.001$). Observed metastasis at the time of diagnosis increases relative risk of death by about 64.8% compared to when metastasis is not found ($p<0.05$). Lastly, malignancy (behavior) of the cancer increases the relative risk of death by a factor of around 16.054 (9.922, 25.975) compared to a non-malignancy group that includes benign, in situ, and borderline malignant cancers ($p<0.001$).

4. LIMITATIONS

The scope of this analysis with respect to inference is limited to those women in the narrow subset defined in the introduction: those who were diagnosed with BC for the first time between 2004 and 2015 and live in the Los Angeles metro area. Those who were diagnosed outside this time interval and in other regions of the U.S. and world will have different health outcomes and survival behavior than is presented here. Additionally, due to our relative inexperience in the field, the predictors we considered at the beginning of this analysis are likely a less than optimal set; other available data in the registry may be more predictive of survival behavior of the selected population. Further, the massive sample size of the registry subset could have the adverse effect of artificially shrinking p-values such that conclusions we deem statistically significant here are truly mild at best.

5. CONCLUSION

Women in this particular subset of the population have a relatively high overall chance of survival from BC even several years after diagnosis, regardless of the variable demographic characteristics under study, although some have a higher relative risk of death than others. Our parametric Weibull AFT model yields significant risk disparities in demographic variables such as race, Hispanic ethnicity, age at diagnosis, and marital/partner status, as well as cancer-specific variables such as metastasis status, summary stage, malignancy/behavior, and histological group. This analysis of the SEER BC registry can be broadened to target increasingly generic subgroups, allowing for more inclusive inference and intricate predictive modeling.

6. TABLES FIGURES

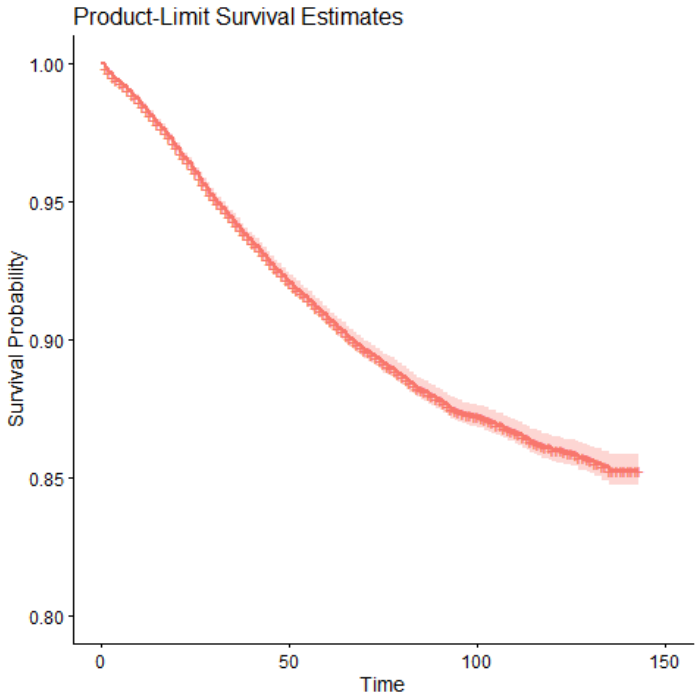

Figure 1. Product-Limit Survival Estimates

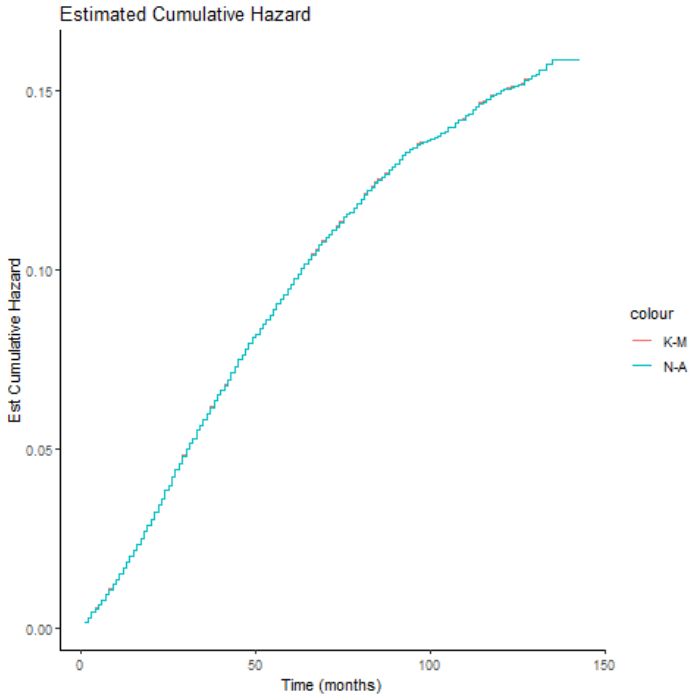

Figure 2. Estimated Cumulative Hazard

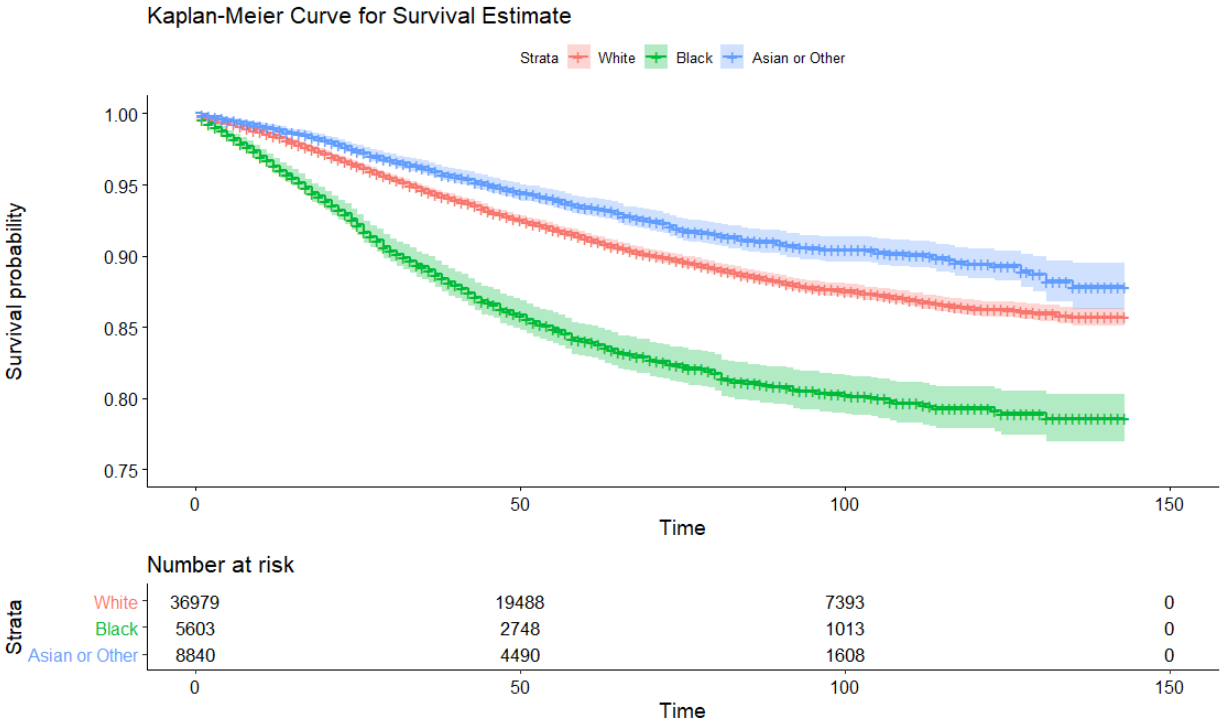

Figure 3. Kaplan-Meier Curve for Survival Estimate by Race

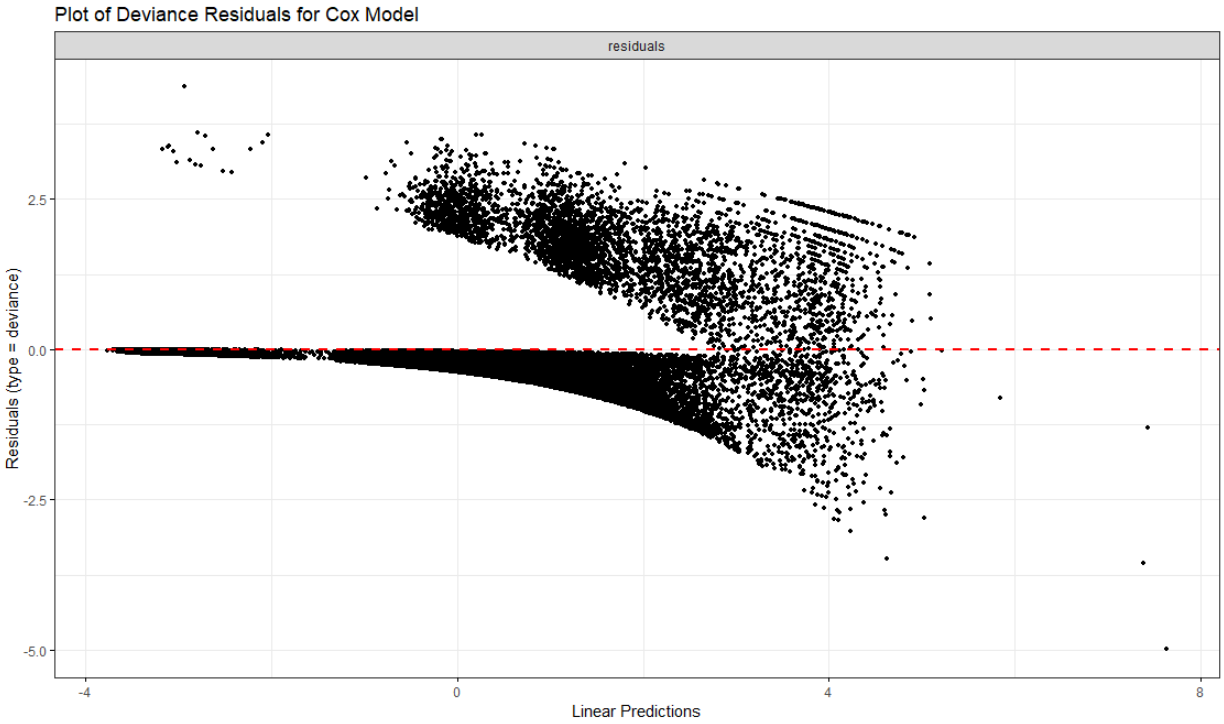

Figure 4. Deviance Residuals for Cox Model

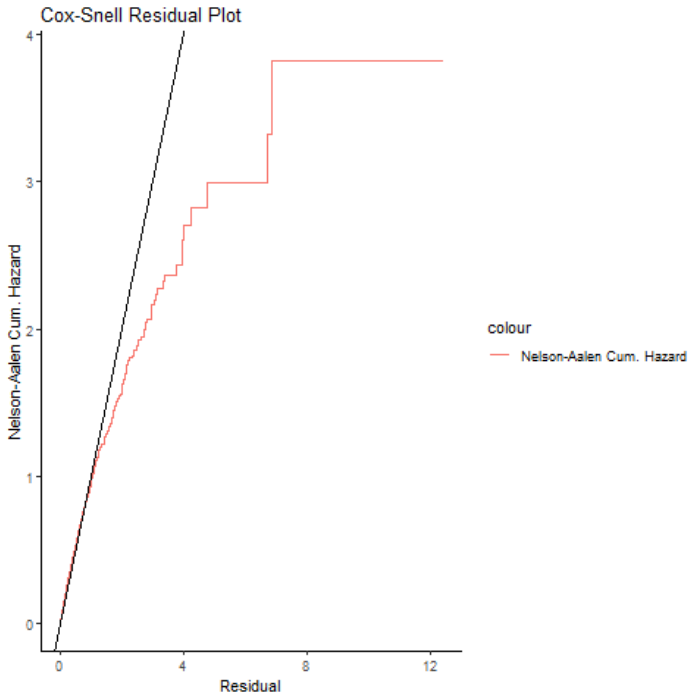

Figure 5. Cox-Snell Residual Plot for Cox Model

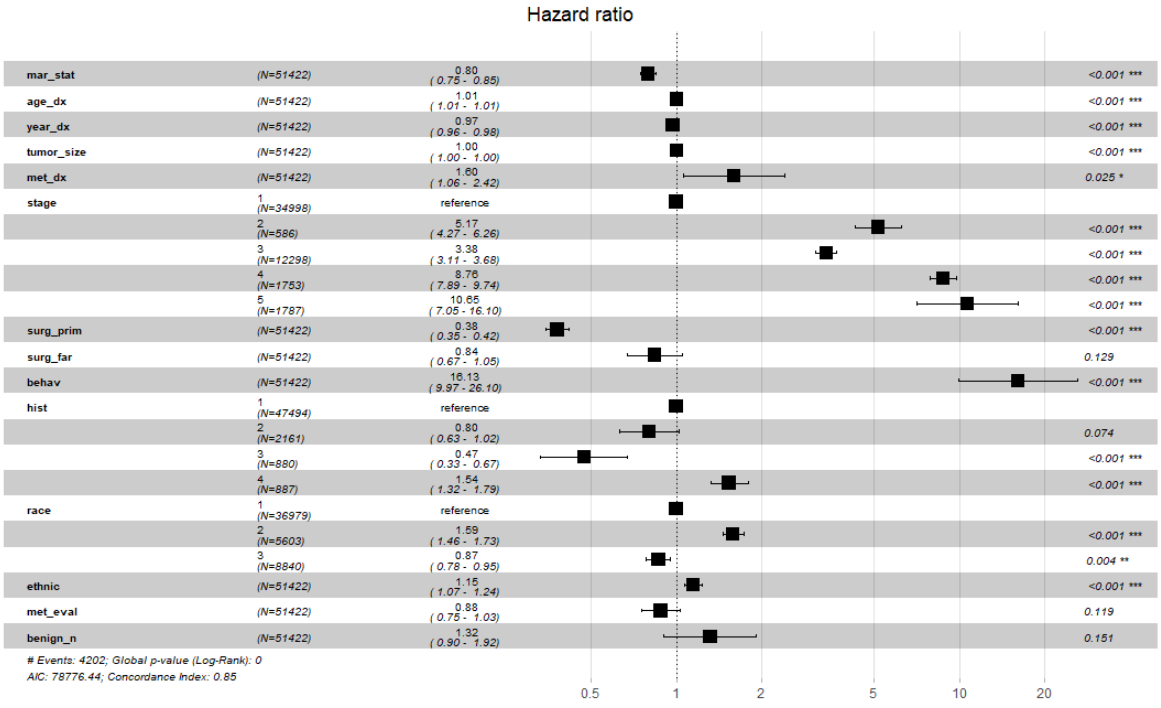

Figure 6. Hazard Ratios and CIs for Cox Model

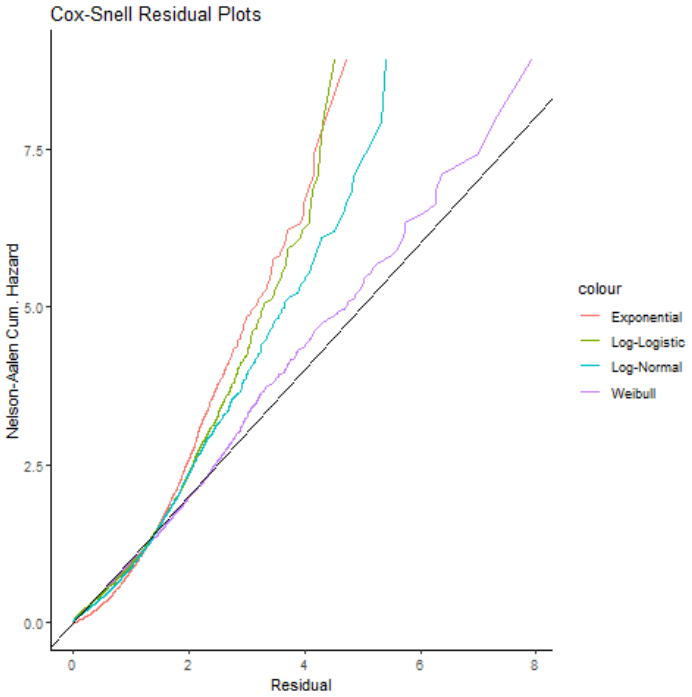

Figure 7. Hazard Ratio for Cox Model

Table 1. AIC/BIC Estimates for AFT Models

	Exponential	Weibull	Log-Logistic	Log-Normal
AIC	38098.08	37004.27	37645.15	37826.19
BIC	38328.12	37243.16	37884.04	38065.08

Table 2. Parametric Weibull AFT Model Estimates

	Coefficient Estimates (SE)	Relative Risk (95% CI)
(Intercept)	-1.33	
	(10.85)	
Marital Status	0.21***	0.79
	(0.03)	(0.742, 0.841)
Age at Diagnosis	-0.01***	1.006
	(0.00)	(1.006, 1.007)
Year of Diagnosis	0.01	0.994
	(0.01)	(0.983, 1.006)
Tumor Size	-0.00***	1.004
	(0.00)	(1.004, 1.005)
Metastasis at Diagnosis	-0.45*	1.648
	(0.19)	(1.092, 2.487)
Stage 2 (Regional - direct extension only)	-1.51***	5.313
	(0.09)	(4.387, 6.434)
Stage 3 (Regional - lymph nodes only)	-1.11***	3.413
	(0.04)	(3.142, 3.707)
Stage 4 (Regional - direct extension and lymph)	-1.98***	8.993
	(0.05)	(8.092, 9.995)
Stage 5 (NOS/distant sites)	-2.15***	10.881
	(0.19)	(7.203, 16.438)
Primary Location Surgery	0.88***	0.379
	(0.04)	(0.344, 0.417)
Non-Local Surgery	0.15	0.846
	(0.10)	(0.676, 1.059)
Behavior of Tumor	-2.50***	16.054
	(0.22)	(9.922, 25.975)
Histology 2 (Adenomas and adenocarcinomas)	0.21	0.796
	(0.11)	(0.626, 1.012)
Histology 3 (Cystic, mucinous and serous neoplasms)	0.67***	0.473
	(0.16)	(0.332, 0.674)
Histology 4 (Other neoplasms)	-0.40***	1.554
	(0.07)	(1.335, 1.809)
Race 2 (Black)	-0.42***	1.595
	(0.04)	(1.465, 1.736)
Race 3 (Asian/Other)	0.13**	0.867
	(0.05)	(0.786, 0.957)
Ethnicity	-0.13***	1.154
	(0.03)	(1.071, 1.243)
Metastasis at Evaluation	0.13	0.863
	(0.07)	(0.736, 1.011)
Number of Benign Tumors	-0.26	1.33
	(0.17)	(0.913, 1.937)
Log(scale)	-0.10***	
	(0.01)	
Log Likelihood	-27658.21	
Num. obs.	51422	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

CODE APPENDIX

```
library(survival)
library(tidyverse)
library(ncvreg)
library(AdapEnetClass)
library(survminer)
library(flexsurv)
library(SurvRegCensCov)
library(texreg)

dat0$hist <- factor(dat0$hist)
## 1 - ductal and lobular neoplasms
## 2 - adenomas and adenocarcinomas
## 3 - cystic, mucinous and serous neoplasms
## 4 - Other neoplasms

dat0 <- dat0[dat0$race != "9",]
dat0 <- dat0 %>%
  mutate(race = if_else(race == "3" | race == "4", "3", race))
dat0$race <- factor(dat0$race)
# 1 - White
# 2 - Black
# 3 - Asian or Other

# ethnic
# 1 - hispanic
# 0 - not hispanic


dat0 <- dat0[dat0$cod != "41000" & dat0$cod != "50000" & dat0$cod != "50030"
             & dat0$cod != "50040" & dat0$cod != "50050" & dat0$cod != "50051"
             & dat0$cod != "50120" & dat0$cod != "50140" & dat0$cod != "50150"
             & dat0$cod != "50160" & dat0$cod != "50170" & dat0$cod != "50180"
             & dat0$cod != "50200" & dat0$cod != "50210" & dat0$cod != "50220"
             & dat0$cod != "50230" & dat0$cod != "50300",]
dat0 <- dat0 %>%
  mutate(cod = if_else(cod == "00000", 0,
                       if_else(cod == "26000", 1,
                               if_else(cod == "50060" | cod == "50070"
                                       | cod == "50090" | cod == "50100"
                                       | cod == "50110" | cod == "50130", 2,
                                       3))))
# 0 - Alive
# 1 - Breast Cancer
# 2 - Heart Disease
# 3 - Other Cancer

dat0 <- dat0 %>%
  mutate(delta = if_else(delta == "4", 1, 0))

dat0 <- dat0[dat0$met_eval != "9",]
dat0 <- dat0 %>%
  mutate(met_eval = if_else(dat0$met_eval == "0" | dat0$met_eval == "1", 0, 1))
# 0 - no metastasis
# 1 - metastasis

dat0$time <- as.numeric(dat0$time)
dat0 <- dat0[dat0$time != 9999,]
```

```
  mutate(stage2 = if_else(stage == 2, 1, 0),
         stage3 = if_else(stage == 3, 1, 0),
         hist2 = if_else(hist == 2, 1, 0),
         hist3 = if_else(hist == 3, 1, 0),
         hist4 = if_else(hist == 4, 1, 0),
         race2 = if_else(race == 2, 1, 0),
         race3 = if_else(race == 3, 1, 0),
         )

  breast <- readRDS("~/Fall 2018/BIOSTAT M215/Final Project/breast.Rds")

breast <- breast[breast$time != 0,]
breast <- breast[breast$cod != 2 & breast$cod != 3,]
breast$ethnic <- as.numeric(breast$ethnic)

### KME and NAE
#MKE
kme.fit <- survfit(Surv(time, delta) ~ 1, data = breast)

png("KM.png")
ggsurvplot(kme.fit, ylim = c(0.8, 1), title = "Product-Limit Survival Estimates",
     ylab = "Survival Probability", xlab = "Time", legend = "none")
dev.off()

#KME CumHaz
H.km <- -log(kme.fit$surv)

#Nelson-Aalen Estimator
h <- kme.fit$n.event / kme.fit$n.risk
H.na <- cumsum(h)
s <- kme.fit$n.event / kme.fit$n.risk ^ 2
V.na <- cumsum(s)
NAest <- cbind(H.na, sqrt(V.na))
colnames(NAest) <- c("NA-Est", "Std Err")

#compare KME and NAE for cum. hazard
cumhaz.plot <- ggplot() +
  geom_step(aes(x = kme.fit$time, y = H.km, colour = "K-M")) +
  geom_step(aes(x = kme.fit$time , y = H.na, colour = "N-A")) +
  labs(x = "Time (months)", y = "Est Cumulative Hazard", title = "Estimated Cumulative Hazard") +
  theme_classic()

png("CumHaz.png")
cumhaz.plot
dev.off()

#### Best subset selection using Cox Model

survfit <- survfit(Surv(time, delta) ~ race, data = breast)
ggsurvplot(survfit, data = breast, conf.int = TRUE,
           title = "Kaplan-Meier Curve for Survival Estimate",
           risk.table = TRUE, ylim = c(.75, 1),
           legend.labs = c("White", "Black", "Asian or Other"))
```

```r
X <- breast %>%
  select(-time, -delta, -cod, -delta_cr) %>%
  as.data.frame(.)

mcp.fit <- ncvsurv(X, Surv(breast$time, breast$delta), penalty = "MCP",
                   nlambda = 25)
mcp.bic <- AIC(mcp.fit, k = log(nrow(breast)))
mcp.est <- mcp.fit$beta[, which.min(mcp.bic)]

fit.cox <- coxph(Surv(time, delta) ~ mar_stat + age_dx + year_dx + tumor_size +
               met_dx + stage + surg_prim + surg_far + behav + hist + race +
                 ethnic + met_eval + benign_n, data = breast, ties = "breslow")
cox.zph(fit.cox)
ggcoxdiagnostics(fit.cox, type = "deviance",
                 title = "Plot of Deviance Residuals for Cox Model")
ggforest(fit.cox, data = breast)

cox.snell.cox <- ggplot() +
  geom_step(aes(x = fit.cs$time, y = H.cs, colour = "Nelson-Aalen Cum. Hazard")) +
  geom_abline() +
  labs(x = "Residual", y = "Nelson-Aalen Cum. Hazard", title = "Cox-Snell Residual Plot") +
  theme_classic()

png("cs.cox.png")
cox.snell.cox
dev.off()

X <- X %>%
  mutate(stage2 = if_else(stage == 2, 1, 0),
         stage3 = if_else(stage == 3, 1, 0),
         stage4 = if_else(stage == 4, 1, 0),
         stage5 = if_else(stage == 5, 1, 0),
         hist2 = if_else(hist == 2, 1, 0),
         hist3 = if_else(hist == 3, 1, 0),
         hist4 = if_else(hist == 4, 1, 0),
         race2 = if_else(race == 2, 1, 0),
         race3 = if_else(race == 3, 1, 0)) %>%
  select(-stage, -hist, - race)
X$ethnic <- as.numeric(X$ethnic)

weight <- mrbj(cbind(breast$time, breast$delta) ~ X$mar_stat +
                 X$age_dx + X$year_dx + X$tumor_size + X$met_dx + X$stage2 +
                 X$stage3 + X$stage4 + X$stage5 + X$surg_prim + X$sur_lymph +
                 X$surg_far + X$behav + X$hist2 + X$hist3 + X$hist4 + X$race2 +
                 X$race3 + X$ethnic + X$met_eval + X$malig_n + X$benign_n,
               mcsize = 100, trace = FALSE, gehanonly = FALSE)
wt <- round(weight$enet, 5)

X <- X %>%
  select(-sur_lymph, -malig_n) %>%
  as.data.frame(.)
```

```r
breast <- breast %>%
  mutate(stage2 = if_else(stage == 2, 1, 0),
         stage3 = if_else(stage == 3, 1, 0),
         stage4 = if_else(stage == 4, 1, 0),
         stage5 = if_else(stage == 5, 1, 0),
         hist2 = if_else(hist == 2, 1, 0),
         hist3 = if_else(hist == 3, 1, 0),
         hist4 = if_else(hist == 4, 1, 0),
         race2 = if_else(race == 2, 1, 0),
         race3 = if_else(race == 3, 1, 0))

### Fitting Full AFT Models
breast3 <- breast %>%
  select(-delta_cr)
##### Parametric exponential model:
fit.exp <- survreg(Surv(time, delta) ~ ., data = breast3, dist="exponential")
summary(fit.exp)

##### Parametric Weibull model:
fit.weibull <- survreg(Surv(time, delta) ~ ., data = breast3, dist="weibull")
summary(fit.weibull)

##### Parametric Log-normal model:
fit.lognormal <- survreg(Surv(time, delta) ~ ., data = breast3, dist="lognormal")
summary(fit.lognormal)

##### Parametric Log-logistic model:
fit.loglogistic <- survreg(Surv(time, delta) ~ ., data = breast3, dist="loglogistic")
summary(fit.loglogistic)


### Cox-Snell Residuals
# Exp distribtion
sigma.exp <- fit.exp$scale
eta.exp <- -fit.exp$linear.predictors / sigma.exp

r.exp <- breast3$time * exp(eta.exp)
fit.exp.cs <- survfit(Surv(r.exp, breast3$delta) ~ 1)
H.exp <- cumsum(fit.exp.cs$n.event / fit.exp.cs$n.risk)

# Weibill distribution
sigma.wb <- fit.weibull$scale
alpha.wb <- 1 / sigma.wb
eta.wb <- -fit.weibull$linear.predictors / sigma.wb
r.wb <- breast3$time ^ alpha.wb * exp(eta.wb)

fit.wb.cs <- survfit(Surv(r.wb, breast3$delta) ~ 1)
H.wb <- cumsum(fit.wb.cs$n.event / fit.wb.cs$n.risk)

# Log-Logistics
sigma.ll <- fit.loglogistic$scale
alpha.ll <- 1 / sigma.ll
eta.ll <- -fit.loglogistic$linear.predictors / sigma.ll
```

```
r.ll <- -log(1 / (1 + breast3$time ^ alpha.ll * exp(eta.ll)))

fit.ll.cs <- survfit(Surv(r.ll, breast3$delta) ~ 1)
H.ll <- cumsum(fit.ll.cs$n.event / fit.ll.cs$n.risk)

# Log-Normal
eta.ln <- -fit.lognormal$linear.predictors / fit.lognormal$scale
r.ln <- -log(1 - pnorm((log(breast3$time)
- fit.lognormal$linear.predictors) / fit.lognormal$scale))

fit.ln.cs <- survfit(Surv(r.ln, breast3$delta) ~ 1)
H.ln <- cumsum(fit.ln.cs$n.event / fit.ln.cs$n.risk)

# plot of residuals vs cumhaz
cs.reg.plot <- ggplot() +
  geom_line(aes(x = fit.wb.cs$time, y = H.wb, colour = "Weibull")) +
  geom_line(aes(x = fit.exp.cs$time, y = H.exp, colour = "Exponential")) +
  geom_line(aes(x = fit.ll.cs$time, y = H.ll, colour = "Log-Logistic")) +
  geom_line(aes(x = fit.ln.cs$time, y = H.ln, colour = "Log-Normal")) +
  geom_abline() +
  labs(x = "Residual",
  y = "Nelson-Aalen Cum. Hazard",
  title = "Cox-Snell Residual Plots") +
  theme_classic()

png("cs.reg.png")
cs.reg.plot
dev.off()

AIC.exp <- AIC(fit.exp); AIC.exp
AIC.wb <- AIC(fit.weibull); AIC.wb
AIC.ll <- AIC(fit.loglogistic); AIC.ll
AIC.ln <- AIC(fit.lognormal); AIC.ln

BIC.exp <- AIC(fit.exp, k = log(nrow(breast3))); BIC.exp
BIC.wb <- AIC(fit.weibull, k = log(nrow(breast3))); BIC.wb
BIC.ll <- AIC(fit.loglogistic, k = log(nrow(breast3))); BIC.ll
BIC.ln <- AIC(fit.lognormal, k = log(nrow(breast3))); BIC.ln

### Fitting Weibull AFT Model with selected vars
fit.weibull <- survreg(Surv(breast$time, breast$delta) ~ X$mar_stat + X$age_dx
                   + X$year_dx + X$tumor_size + X$met_dx + X$stage2 + X$stage3
                   + X$stage4 + X$stage5 + X$surg_prim + X$surg_far + X$behav
                   + X$hist2 + X$hist3 + X$hist4 + X$race2 + X$race3 + X$ethnic
                   + X$met_eval + X$benign_n, dist="weibull")
summary(fit.weibull)
RR <- ConvertWeibull(fit.weibull, conf.level = 0.95)$HR
RR <- RR %>%
  as.data.frame(.) %>%
  round(., 3) %>%
  mutate("Relative Risk (95% CI)" = paste(round(RR$HR, 3), " (",
                                          round(RR$LB, 3), "," ,
                                          round(RR$UB, 3), ")", sep = "")) %>%
```

```
  select(-HR, -LB, -UB)

texreg(fit.weibull, dcolumn = TRUE, booktabs = TRUE, float.pos = "hb!",
       caption = "Parametric Weibull AFT Model Estimates")
```