Final Data Analysis Project

Introduction and Data

Seroconversion is process by which HIV antibodies become detectable in the blood. Therefore, seroconversion can be a proxy for HIV infection and it is useful to know how time to seroconversion varies between different groups and risk factors. Mustanski et. al.[1] perform a survival analysis to determine serocvonersion risk factors for young men who have sex with men (MSM) and transgender women in the Chicago metropolitan area. ATN CARES is a Los Angeles and New Orleans based study with a longitudinal cohort of MSM and transgender women. The goal of this analysis is to replicate the results in Mustanski et. al. using bayesian methods in order to account for the rarity of seroconversion.

For the purpose of this analysis, time from birth to seroconversion (years), gender identity (trans/cisgender), self-reported history of syphilis (yes/no), and unwanted adolescent sexual experiences (yes/no) were collected on 888 participants. There were 28 seroconversions observed and the rest of the participants were censored at the time of data collection.

Summary statistics can be found in Table 1 and Kaplan-Meier estimates of survival curves can be found in Figure 1A-E. Note also that only 7 out 2^4 possible categories had uncensored events, 15 of which were in the cisgender, black, no syphilis, no unwanted adolescent experiences class.

Modeling

Time to seroconversion was assumed to follow a Weibull distribution with density function $f(t|r,\lambda) = r\lambda t^{r-1}\exp(-\lambda t^r) = h(t)\exp(-\lambda t^r)$ for $t > 0$,

$r > 0$, $\lambda > 0$. In order to add in covariates, the link function $\lambda = \exp(\eta)$ was used where η is the linear predictor.[2] The model can written as

$$T \sim h(t) \exp\left[-\exp(\beta_1 * \text{trans} + \beta_2 * \text{black} + \beta_3 * \text{syphilis} + \beta_4 * \text{unwanted}) t^r\right]$$

This is a parametric proportional hazards model. Note that the difference between this parametric model and Cox's proportional hazards model is that $h(t)$ is assumed to have a specific form, in this case the weibull hazard function. The weibull proportional hazards model was chosen based on the ability to mimic the more complicated model in Mustanski that allows the rate of seroconversions to change over time. Additionally, the model is easier to implement from a bayesian perspective compared to semi-parametric methods.

Priors were elicited from the results in Mustanski et. al. The model needed priors to be specified for r, $h(t)$, β_1, β_2, β_3, and β_4. First, note that $h(t)$ can thought of as an intercept β_0. To obtain the baseline β_0 prior and a prior for r, it was noted that in Mustanski et. al.'s model, the risk of seroconversion decreased for every year in age and ended with 44 seroconversions out of 809 subjects. Using an optimization procedure, a similar survival curve was estimated over 13 years to represent the time spread in the ATN CARES data. R code for this optimization algorithm is in the appendix. This gave $r = 0.63$, $\lambda = 1184$ for a weibull distribution that would produce a similar survival curve. r was then given the prior gamma$(0.63, 0.63)$. The prior for β_0 was set as $N(-4.4, 2^2)$ with the variance chosen to allow this prior to be fairly vague. The priors for the regression parameters were specified as normal with mean equal to the estimates in Mustanski's model and standard deviations

equal to the standard errors. Thus the priors are $\beta_1 \sim N(-0.83, 0.75^2)$, $\beta_2 \sim N(1.79, 0.41^2)$, $\beta_3 \sim N(0.48, (2 * 0.37)^2)$, $\beta_4 \sim N(0.65, 0.38^2)$. The prior for β_3 has an increased variance because the original model considered all STDs instead of just syphilis.

The model was fit using JAGS in R and JAGS code is provided in the appendix. I fit both a Weibull and a exponential model to see what would happen if r was fixed at 1. The same models were also fit with vague priors in order to see the effect of the informative priors. I also fit a Cox proportional hazards model and parametric accelerated failure time models using the coxph() and survreg() functions from the survival package in R to compare the frequentist approach to the bayesian methods.

Results

Summaries of the parameter posterior distributions are presented in Table 2 and densities are plotted in Figure 2. The weibull model was chosen as the final model because r had a posterior mean of 1.83 with the 95% posterior interval of (1.26, 2.44), signifying an increasing hazard. Coefficients estimates from the other bayesian and frequentist models are shown in Table 3 as part of a sensitivity analysis.

Coefficients for trans and black had very high posterior probability of not equaling 0. On the other hand β_3 and β_4 did not have posteriors that support non-equality with 0. The weibull model showed some improvement over the Cox PH in the length of the confidence intervals for the regression parameters. For example, the 95% posterior interval for the trans hazard ratio had length 0.95 compared to the Cox PH 95% confidence interval with

length 2.9.

Sensitivity Analysis

A sensitivity analysis was conducted by comparing different combinations of weibull and exponential ($r = 1$) models with informative and vague priors in order to judge the effect of prior choice on the inference. Also presented in Table 3 are the frequentist Cox proportional hazards, weibull, and exponential accelerated failure models. Note that Weibull accelerated failure model failed to converge and its inferences are faulty. Note also that the exponential accelerated failure model is specified differently and the coefficients have the opposite sign compared to the bayesian models. Overall, the inference does not change between different models and priors.

Since there were only 28 seroconversions, it will be useful to know the effect of removing these individual cases. Therefore, I performed a case deletion analysis using the weibull model with informative priors. Figure 3 shows the estimated 95% posterior intervals for each case deletion. The estimates for β_2 are stable but the variance of r fluctuates a little. Note that since there is only 1 trans seroconversion, deleting this case has a pronounced effect on the posterior interval.

Conclusions

My final conclusions are that the bayesian weibull model with informative priors is the best choice for this data since it can account for an increasing hazard as suggested in Mustanski and supported by a high posterior probability of being greater than 1. The model can also use the prior information

from Mustanski et. al. Additionally, the model produces smaller confidence intervals for the parameters compared to the frequentist Cox proportional hazards model.

The two most important predictors for seroconversion in this data were trans and black. Interpreting the model hazard ratio estimates, a trans-female is 0.36 as likely to seroconvert compared to a cisgender MSM and a black MSM is 3 to 9 times as likely to seroconvert compared to other races.

Appendix

References:

1. Mustanski B, Ryan DT, Newcomb ME, D'Aquila RT, Matson M. *Very High HIV Incidence and Associated Risk Factors in a Longitudinal Cohort Study of Diverse Adolescent and Young Adult Men Who Have Sex with Men and Transgender Women.* AIDS Behav. 2020 Jun;24(6):1966-1975. doi: 10.1007/s10461-019-02766-4. PMID: 31858300.

2. *Bayesian Survival Analysis*, Ibrahim, Joseph G., Chen, Ming-Hui, Sinha, Debajyoti.

3. *JAGS Version 4.3.0 User Manual*, Plummer, Martyn.

Variable	**Seroconverts**		**Gender**		**Race**		**Syphilis**		**Unwanted**	
Value	Yes	No	Trans	Cis	Black	Other	Yes	No	Yes	No
n	28	860	72	816	366	522	85	803	265	623
%	3.15	96.8	91.9	8.11	41.2	58.8	9.57	90.4	29.8	70.2

Table 1: Counts of categorical data.

	Mean	SD	2.5%	97.5%	HR	HR 2.5%	HR 97.5%	P>0	P<0
intercept	-8.24	0.71	-9.68	-6.89				0.00	1.00
trans	-1.02	0.58	-2.20	0.06	0.36	0.11	1.06	0.03	0.97
black	1.60	0.30	1.03	2.20	4.96	2.79	8.98	1.00	0.00
syphilis	0.15	0.46	-0.79	1.00	1.17	0.46	2.73	0.64	0.36
unwanted	0.17	0.28	-0.38	0.70	1.18	0.68	2.02	0.73	0.27
r	1.83	0.30	1.26	2.44				1.00	0.00

Table 2: Posterior summaries for Weibull model with informal priors. HR: hazard ratio.

	WBI 1	WBV	EBI	EBV	Cox	WAF*	EAF*
intercept	**-8.24**	**-8.66**	**-6.53**	**-6.31**		**102.55**	6.22
trans	-1.02	-1.56	-0.97	-1.60	-0.93	**21.34**	0.91
black	**1.60**	**1.54**	**1.63**	**1.51**	**1.46**	**-10.61**	**-1.47**
syphilis	0.15	-0.04	0.18	-0.01	0.11	**-25.53**	-0.15
unwanted	0.17	-0.44	0.19	-0.40	-0.37	**4.15**	0.35
r	**1.83**	**2.10**				**0.00**	

Table 3: Coefficient estimates for each parameter by each model. W: Weibull. B: Bayesian, I: informative priors. V: vague priors. AF: frequentist accelerated failure. *: broken model and/or different parameterization. **Bold**: 95% posterior or confidence interval does not include 0 (or 1 for r).

Figure 1A: Kaplan-Meier Estimate for Full Data

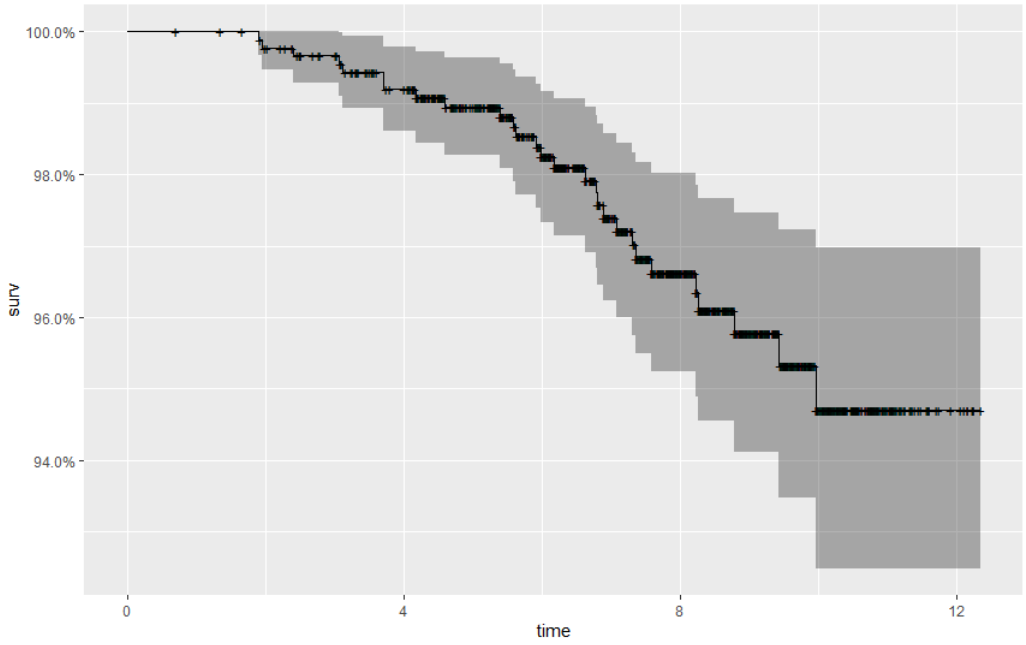


Figure 1B: Kaplan-Meier Estimate for Trans vs. Cisgender

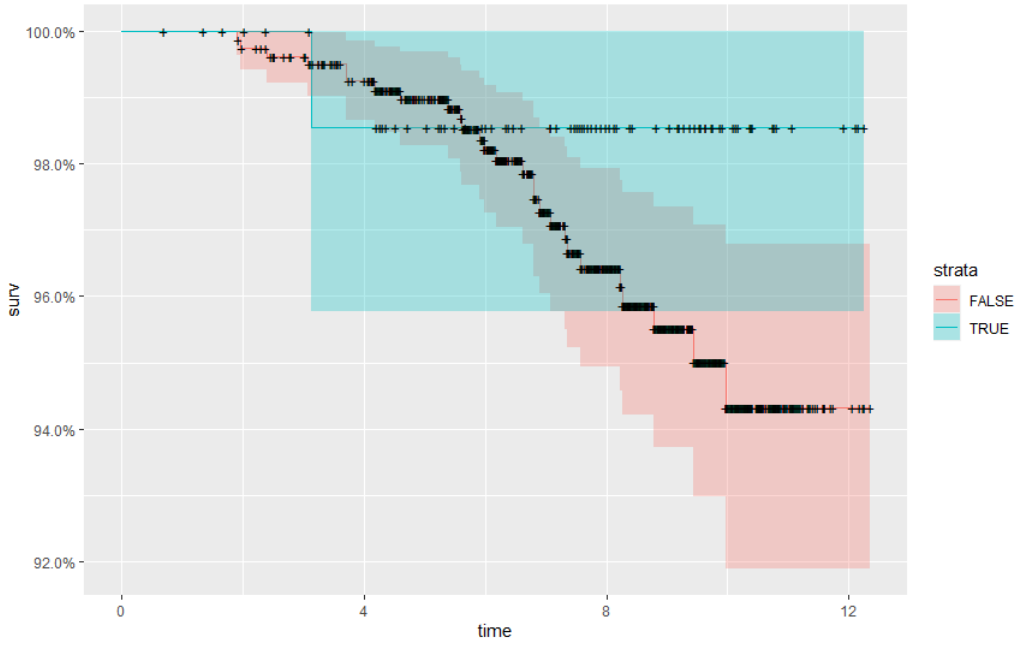


Figure 1C: Kaplan-Meier Estimate for Black vs. Other

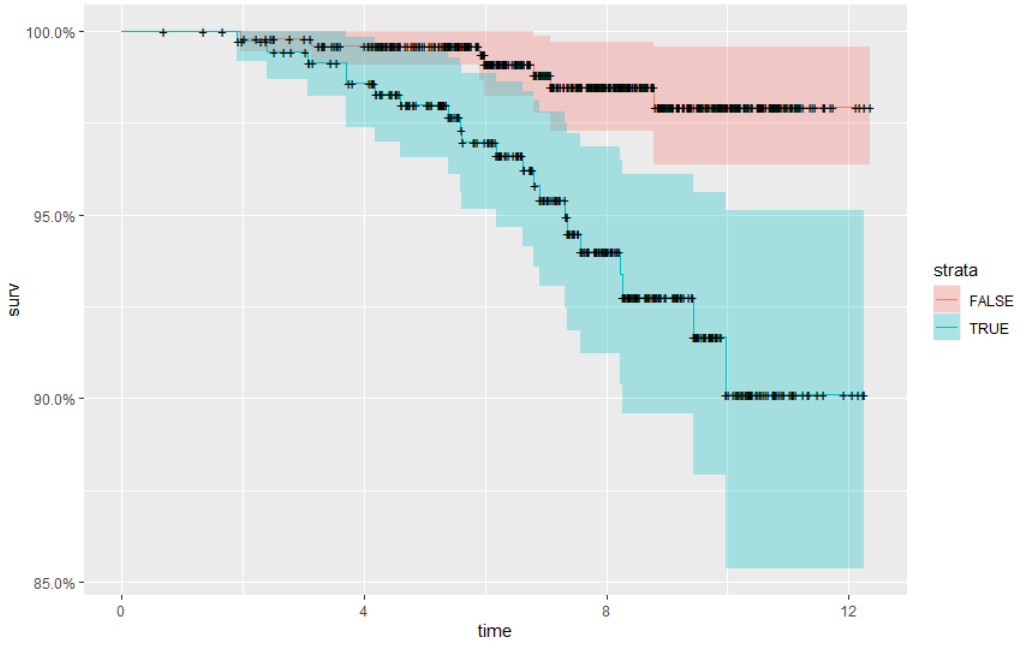


Figure 1D: Kaplan-Meier Estimate for Syphilis

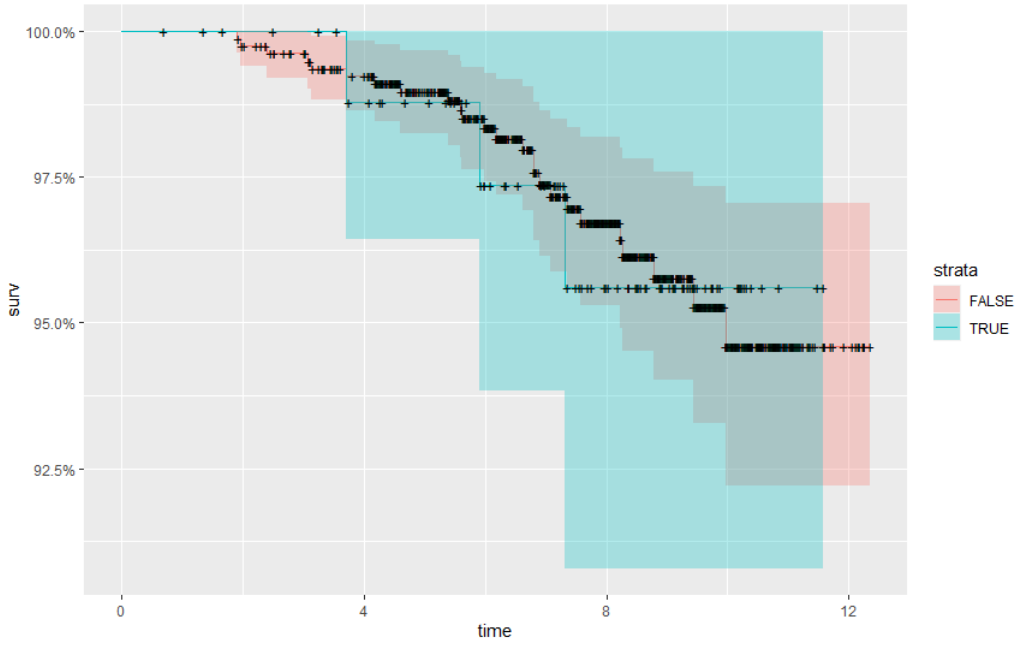


Figure 1E: Kaplan-Meier Estimate for Unwanted

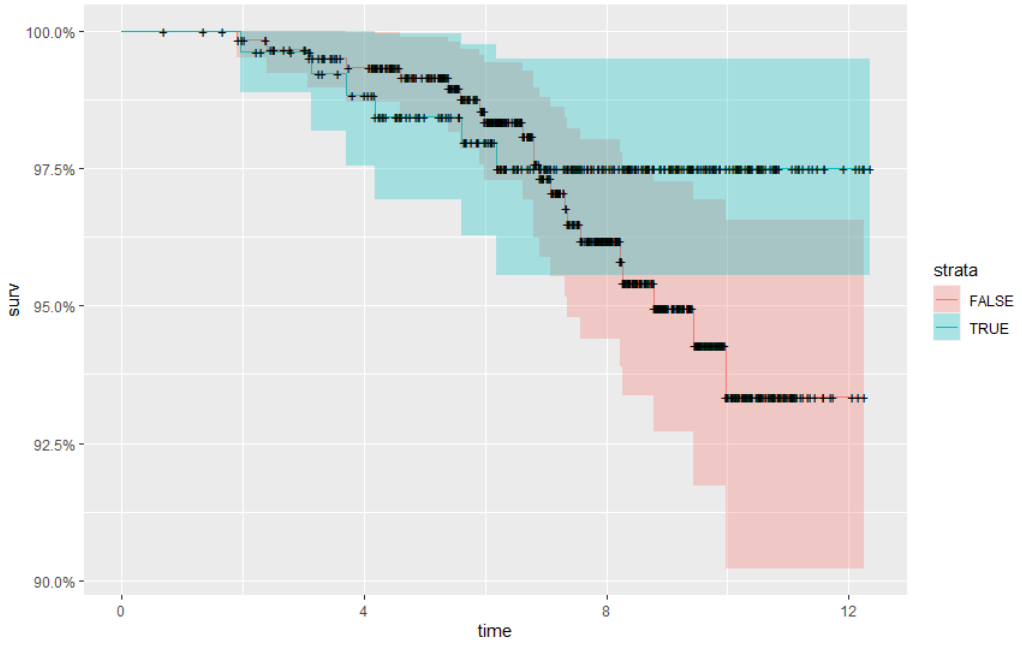


Figure 2: Priors(grey) vs. Posteriors(red)

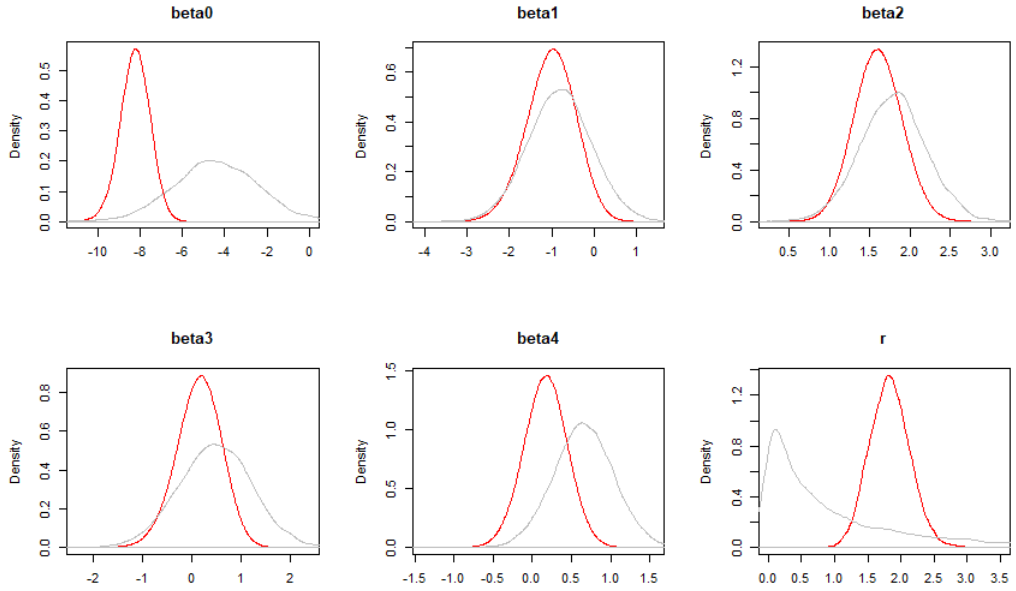


Figure 3a: Case Deletion for β_1

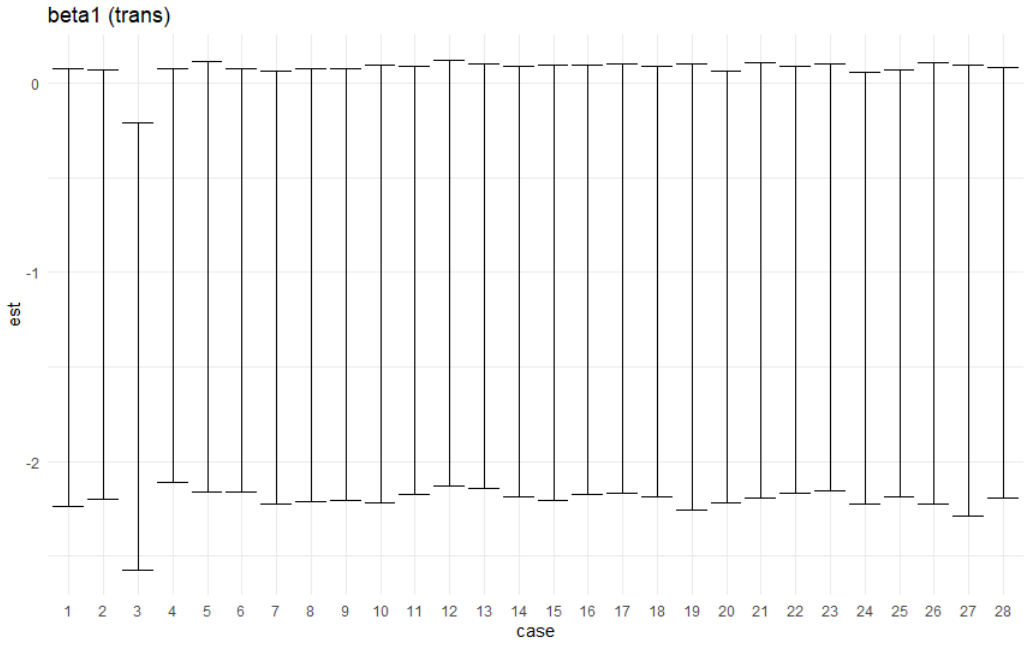


Figure 3b: Case Deletion for β_2

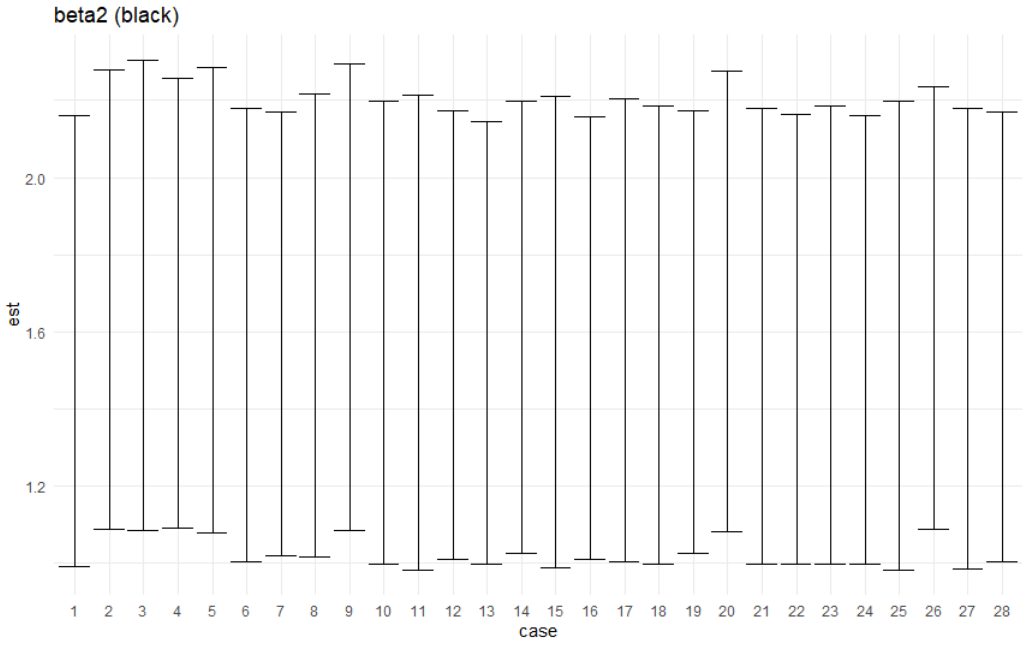
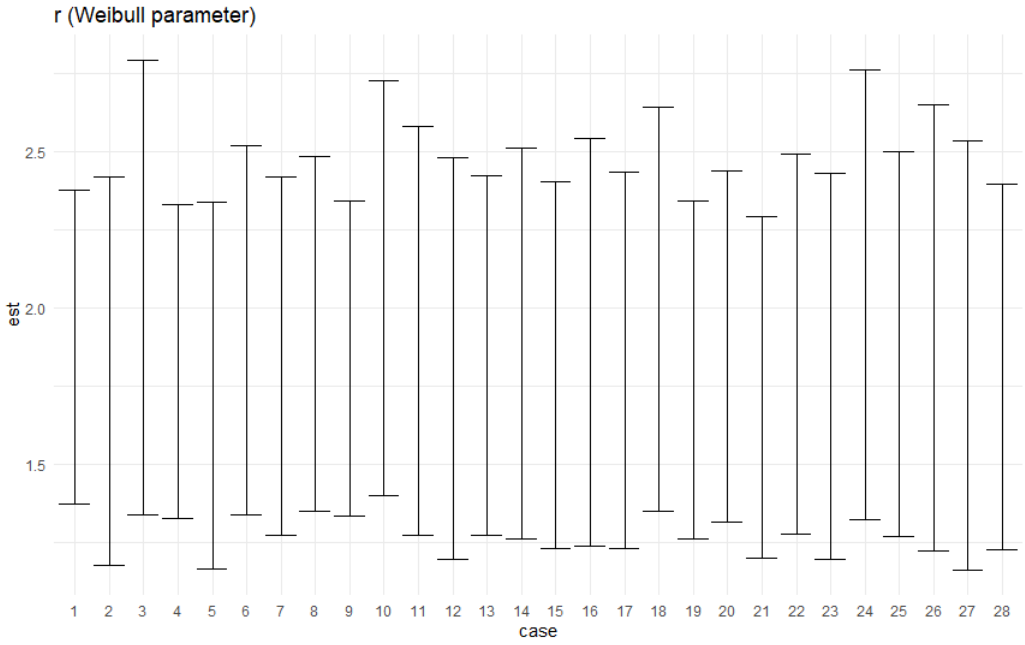


Figure 3c: Case Deletion for r



Optimization algorithm for setting prior $S(t)$

```
# solve recurrence relation numerically to get
# total proportion left at each time point.
optim(1,
      function(x) {
        n = 1000
        ns = numeric(14)
        ns[1] = n
        for (i in 2:14) {
          ns[i] = ns[i-1]-x*.87^i
        }
        return(abs(946-ns[14]))
      },
      method = "Brent", lower = 0, upper = 44) # 11


# objective function
# abs val error loss
objective <- function(param) {

  k = param[1]
  lam = param[2]
  # target values
  year_targets = c( # values from recurrence solved earlier
    0.9916070,
    0.9843051,
    0.9779525,
    0.9724256,
    0.9676173,
    0.9634341,
    0.9597947,
    0.9566284,
    0.9538737,
    0.9514771,
    0.9493921,
    0.9475781,
    0.9460000
  )

  t = 1:13 # 1-13 years based on ATN time spread
```

```
  probs = exp(-(t/lam)^k) # S(t)

  return(sum(abs(year_targets-probs)))
}

# find weibull params that minimize loss function
optim(c(1,10), objective)
```

JAGS code for the weibull model with informative priors

```
model
{
for (i in 1:N) {
  isCensored[i] ~ dinterval(t[i], censorLimitVec[i]) # apply censoring
  log(mu[i]) <- beta0 + beta1*trans[i] + beta2*black[i] +
  beta3*syphilis[i] + beta4*unwanted[i] # link
  t[i] ~ dweibull(r, mu[i])
  #t[i] ~ dweibull(1, mu[i]) # exp likelihood
}
# priors
beta0 ~ dnorm(-4.4, 1/4)
beta1 ~ dnorm(-0.83, 1/(0.75*0.75))
beta2 ~ dnorm(1.79, 1/(.41*.41))
beta3 ~ dnorm(0.48, 1/(4*0.37*0.37))
beta4 ~ dnorm(0.65, 1/(.38*.38))
r ~ dgamma(0.63, 0.63)
}
```