Parametric_Regression

yc4384_Yangyang_Chen

2024-12-03

Failed of Log-Rank Tests

```
## Call:
## survdiff(formula = surv_object ~ trt, data = pbcseq_cleaned,
##     rho = 0)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=0 140       63     66.6     0.194     0.401
## trt=1 136       66     62.4     0.207     0.401
##
##  Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

Why we should consider some alternative approaches based on parametric models:

The assumption of proportional hazards might not be appropriate (based on major departures).

I. Exponential Regression

Assume T_i follows an exponential distribution with a parameter λ that depends on $\mathbf{Z}_i$, say $\lambda_i = \Psi(\mathbf{Z}_i)$. Then we can write:

$$T_i \sim \text{exponential}(\Psi(\mathbf{Z}_i))$$

(1) Fit Exponential Regression model by Stepwise Selection

a. Multivariate Analysis

- Fit a full model with all candidate variables, then use bidirectional stepwise selection to identify the optimal subset of predictors based on AIC.

```
##
## Call:
## survreg(formula = Surv(time, status) ~ bili + albumin + copper +
##     protime + stage + sex, data = pbcseq_cleaned, dist = "exponential")
##                 Value Std. Error     z       p
## (Intercept) 10.949086   1.830518  5.98 2.2e-09
## bili        -0.076750   0.016120 -4.76 1.9e-06
## albumin      0.681309   0.237017  2.87  0.0040
## copper      -0.002735   0.000933 -2.93  0.0034
## protime     -0.263338   0.094047 -2.80  0.0051
## stage2      -1.668938   1.050203 -1.59  0.1120
## stage3      -1.952091   1.027593 -1.90  0.0575
## stage4      -2.279886   1.025429 -2.22  0.0262
## sexf         0.447154   0.250923  1.78  0.0747
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -1144.2   Loglik(intercept only)= -1206.3
##  Chisq= 124.19 on 8 degrees of freedom, p= 4.5e-23
## Number of Newton-Raphson Iterations: 6
## n= 276
```

b. Fit Exponential model

```
##
## Call:
## survreg(formula = Surv(time, status) ~ bili + albumin + copper +
##     protime + stage + sex, data = pbcseq_cleaned, dist = "exponential")
##                 Value Std. Error     z       p
## (Intercept) 10.949086   1.830518  5.98 2.2e-09
## bili        -0.076750   0.016120 -4.76 1.9e-06
## albumin      0.681309   0.237017  2.87  0.0040
## copper      -0.002735   0.000933 -2.93  0.0034
## protime     -0.263338   0.094047 -2.80  0.0051
## stage2      -1.668938   1.050203 -1.59  0.1120
## stage3      -1.952091   1.027593 -1.90  0.0575
## stage4      -2.279886   1.025429 -2.22  0.0262
## sexf         0.447154   0.250923  1.78  0.0747
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -1144.2   Loglik(intercept only)= -1206.3
##  Chisq= 124.19 on 8 degrees of freedom, p= 4.5e-23
## Number of Newton-Raphson Iterations: 6
## n= 276
```

(2) Perform Likelihood Ratio Test

```
##                                            Terms Resid. Df    -2*LL Test Df
## 1                                              1       275 2412.563      NA
## 2 bili + albumin + copper + protime + stage + sex       267 2288.371       8
##   Deviance     Pr(>Chi)
## 1       NA           NA
## 2  124.192 4.510631e-23
```

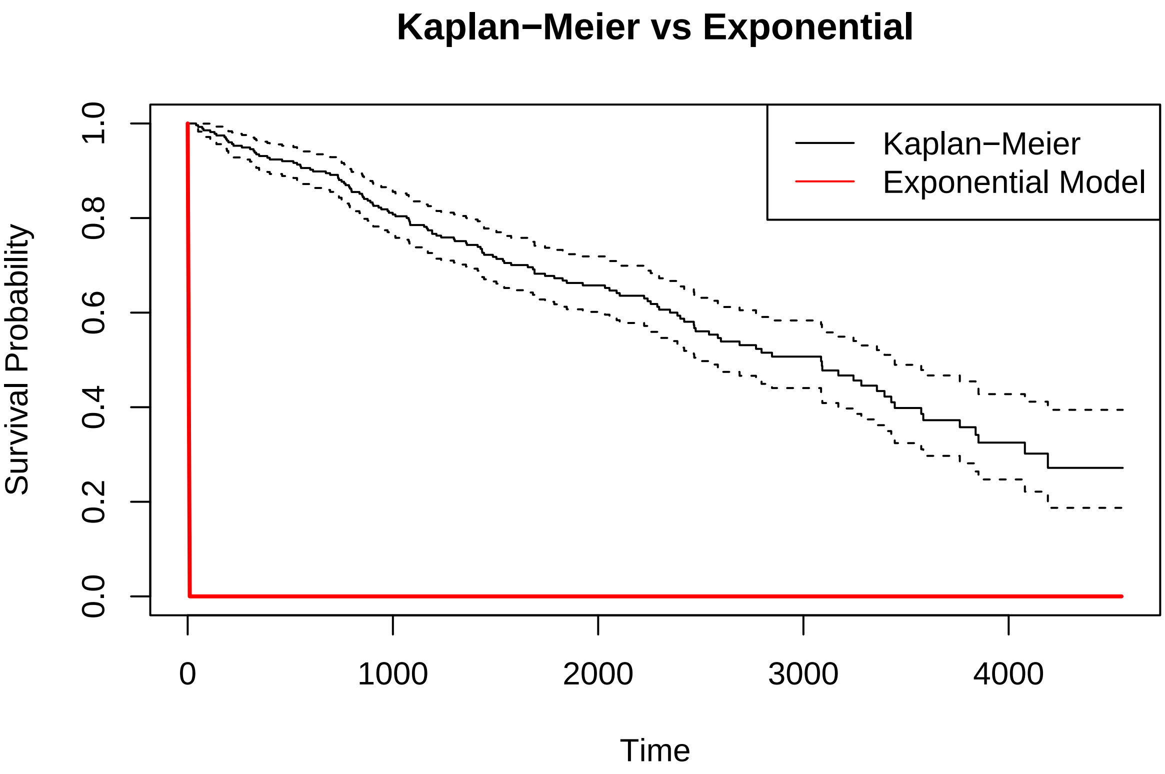
Interpretation of Results:

```
-   Model Significance: The large deviance difference $(\Delta D = 124.192)$ and small p-value $( p < 0
```

- Predictors Significance: The predictors in the exponential model contribute significantly to explaining the survival times, as evidenced by the highly significant p-value.
  - Model Fit: The exponential model improves the fit over the null model by reducing deviance, which measures unexplained variation in the data.

(3) Model Fit Comparison using AIC

```
##            df      AIC
## null_model  1 2414.563
## exp_model   9 2306.371
```



Interpretation of Results:

- The issue in the Kaplan-Meier vs Exponential plot seems to be that the Exponential Model line (red) is a constant hazard (flat line) and does not fit the Kaplan-Meier curve (black) well. This happens because:
  1. Exponential Model Assumption: The exponential distribution assumes a constant hazard rate over time, which may not match the actual survival pattern in your dataset.
  2. Poor Fit: The Kaplan-Meier curve indicates non-constant hazard rates (e.g., survival probabilities decrease differently over time), which suggests that the exponential model may not be appropriate for this data.
- Consider Alternative Parametric Models:
  - The Weibull model is a more flexible parametric survival model that allows for non-constant hazard rates. If the hazard rate varies over time, the Weibull model may provide a better fit.

II. Welbull Regression

Weibull Survival Function:

- The Weibull survival function is calculated as:

$$S(t) = \exp\left(-\left(\frac{t}{\lambda}\right)^k\right)$$

(1) Fit Weibull model by Stepwise Selection

a. Multivariate Analysis:

- Fit a full model with all candidate variables, then use bidirectional stepwise selection to identify the optimal subset of predictors based on AIC.

```
##
## Call:
## survreg(formula = Surv(time, status) ~ edema + bili + albumin +
##     copper + protime + stage + sex, data = pbcseq_cleaned, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept)  9.080176   1.183689  7.67 1.7e-14
## edema0.5    -0.103982   0.175908 -0.59  0.5544
## edema1      -0.590555   0.200062 -2.95  0.0032
## bili        -0.060572   0.010043 -6.03 1.6e-09
## albumin      0.473766   0.156000  3.04  0.0024
## copper      -0.001935   0.000589 -3.29  0.0010
## protime     -0.116627   0.066168 -1.76  0.0780
## stage2      -0.991583   0.641922 -1.54  0.1224
## stage3      -1.148605   0.630116 -1.82  0.0683
## stage4      -1.436687   0.630014 -2.28  0.0226
## sexf         0.311077   0.161156  1.93  0.0536
## Log(scale)  -0.494453   0.071260 -6.94 4.0e-12
##
## Scale= 0.61
##
## Weibull distribution
## Loglik(model)= -1122.9   Loglik(intercept only)= -1203.8
##  Chisq= 161.8 on 10 degrees of freedom, p= 1.4e-29
## Number of Newton-Raphson Iterations: 7
## n= 276
```

b. Fit Weibull Regression Model

```
##
## Call:
## survreg(formula = Surv(time, status) ~ edema + albumin + protime +
##     stage + sex, data = pbcseq_cleaned, dist = "weibull")
##               Value Std. Error     z       p
## (Intercept)  8.4985     1.3834  6.14 8.1e-10
## edema0.5    -0.2588     0.1903 -1.36 0.17385
## edema1      -1.1007     0.2119 -5.19 2.1e-07
## albumin      0.7644     0.1654  4.62 3.8e-06
## protime     -0.1733     0.0746 -2.32 0.02015
## stage2      -1.2469     0.7412 -1.68 0.09250
## stage3      -1.5517     0.7257 -2.14 0.03249
## stage4      -1.8103     0.7245 -2.50 0.01246
## sexf         0.5776     0.1639  3.52 0.00043
## Log(scale)  -0.3639     0.0705 -5.16 2.4e-07
##
## Scale= 0.695
##
## Weibull distribution
## Loglik(model)= -1145.6   Loglik(intercept only)= -1203.8
##  Chisq= 116.38 on 8 degrees of freedom, p= 1.8e-21
## Number of Newton-Raphson Iterations: 6
## n= 276
```

(2) Perform Likelihood Ratio Test

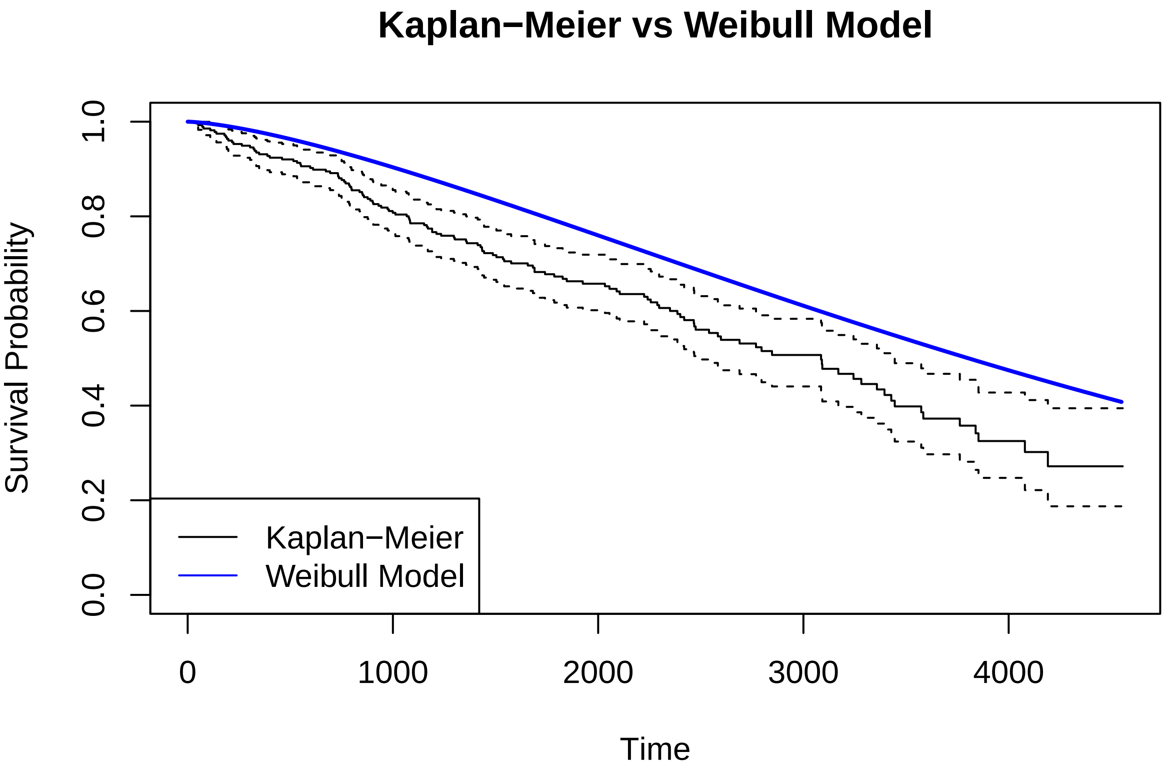
```
##                                              Terms Resid. Df    -2*LL Test Df
## 1                                                1       274 2407.517      NA
## 2 bili + albumin + copper + protime + stage + sex       267 2288.371       7
##   Deviance     Pr(>Chi)
## 1       NA           NA
## 2 119.1457 1.153937e-22
```

Interpretation of Results:

```
-   Model Significance: The large deviance difference $(\Delta D = 119.1457)$ and small p-value $( p < 
```

- Predictors Significance: The predictors in the weibull model contribute significantly to explaining the survival times, as evidenced by the highly significant p-value.
  - Model Fit: The exponential model improves the fit over the null model by reducing deviance, which measures unexplained variation in the data.

(3) Diagnostic Plot



Interpretation of the Plot

```
1.  Model Fit: The Weibull model generally provides a good fit to the data. Its alignment with the Kapl

2.  Weibull Assumptions: The plot supports the assumption of a Weibull distribution for most of the dat
```

III. Parametric Models Comparison

(1) AIC Comparisons

- To assess whether the Weibull model provides a significantly better fit than the exponential model, compare their AIC values:

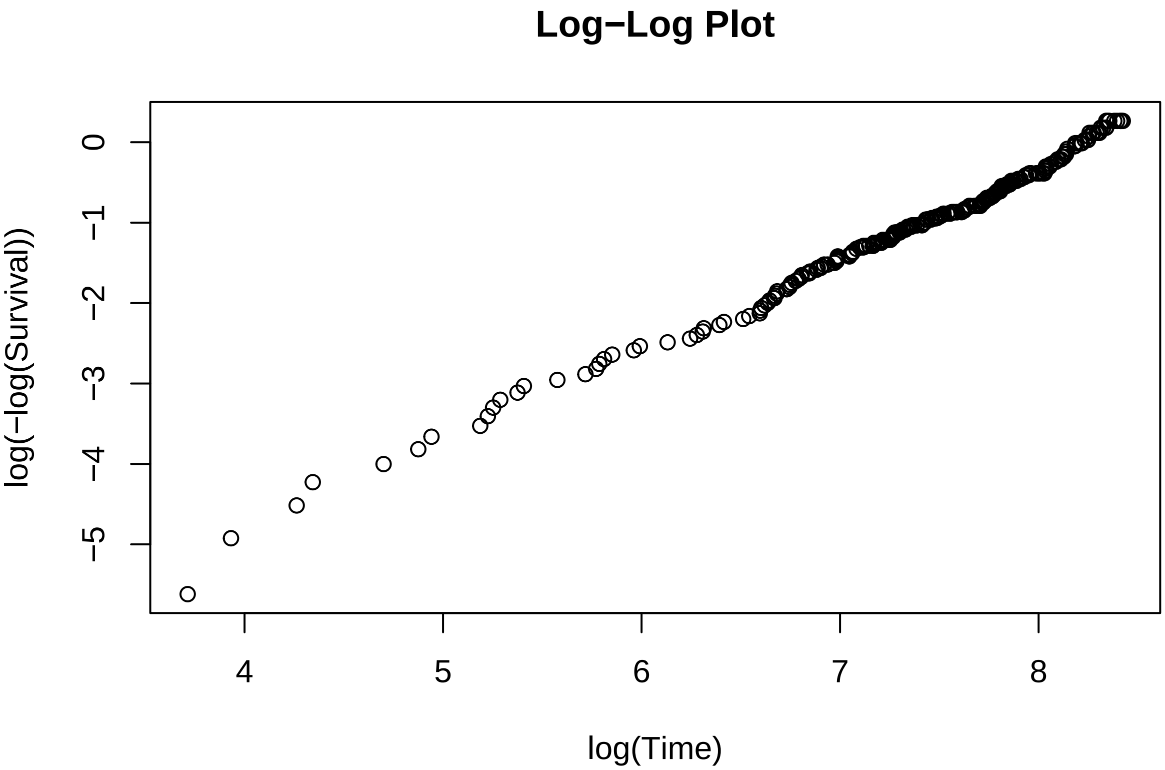
```
##               df      AIC
## exp_model      9 2306.371
## weibull_model 10 2311.133
```

Interpretation:

While the exponential model achieves a slightly lower AIC, the Weibull model is slightly more flexible and is therefore the preferred choice based on these results.

(2) Graphical Diagnostics

- Plot the log cumulative hazard ($log(-log(S(t)))$) against the log of time. If the points form a straight line, it confirms that the Weibull model is appropriate:



Interpretation:

In the log-log plot, the points approximately form a straight line, it suggests that the Weibull distribution is a good fit for the data. Minor deviations in the tails may warrant further exploration, but the Weibull model appears to capture the main survival patterns effectively.

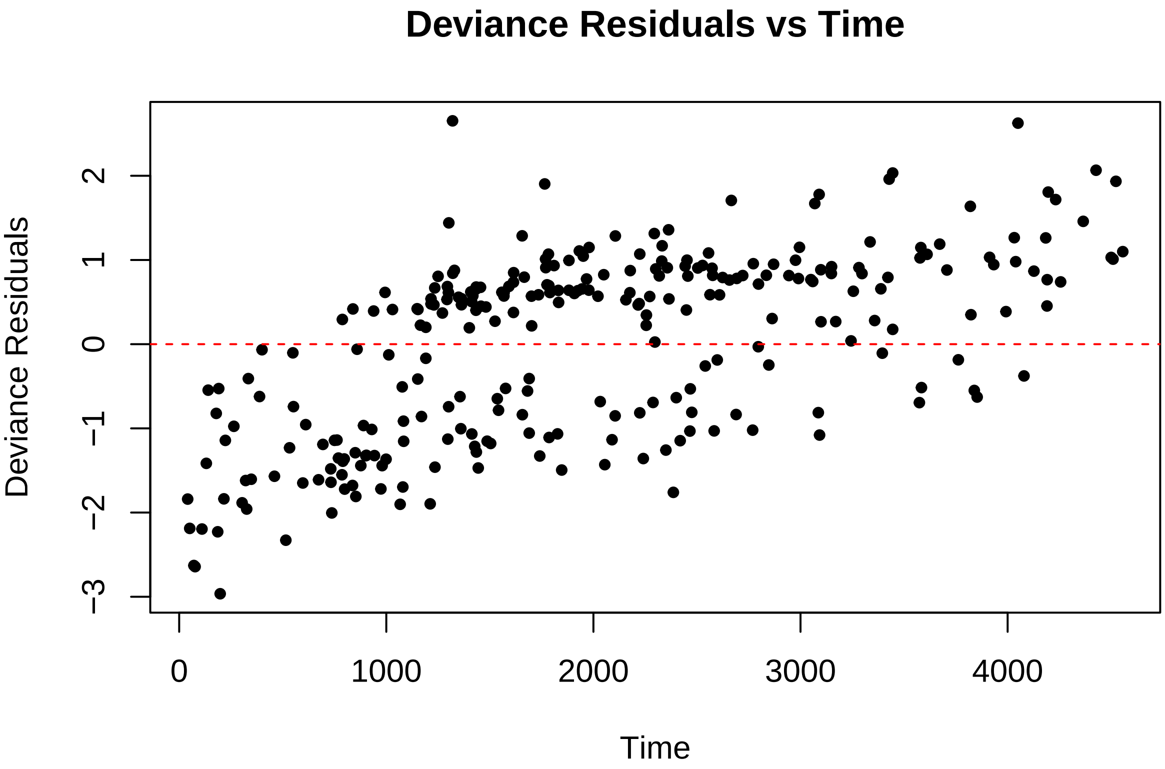
IV. Model Diagnostics

(1) Residual diagnostics

- Residual diagnostics help assess whether the Weibull model fits the survival data well. For survival models, we commonly use Deviance residuals and Cox-Snell residuals.

a. Deviance Residuals

- Deviance residuals can be computed for parametric models fitted with `survreg()`:

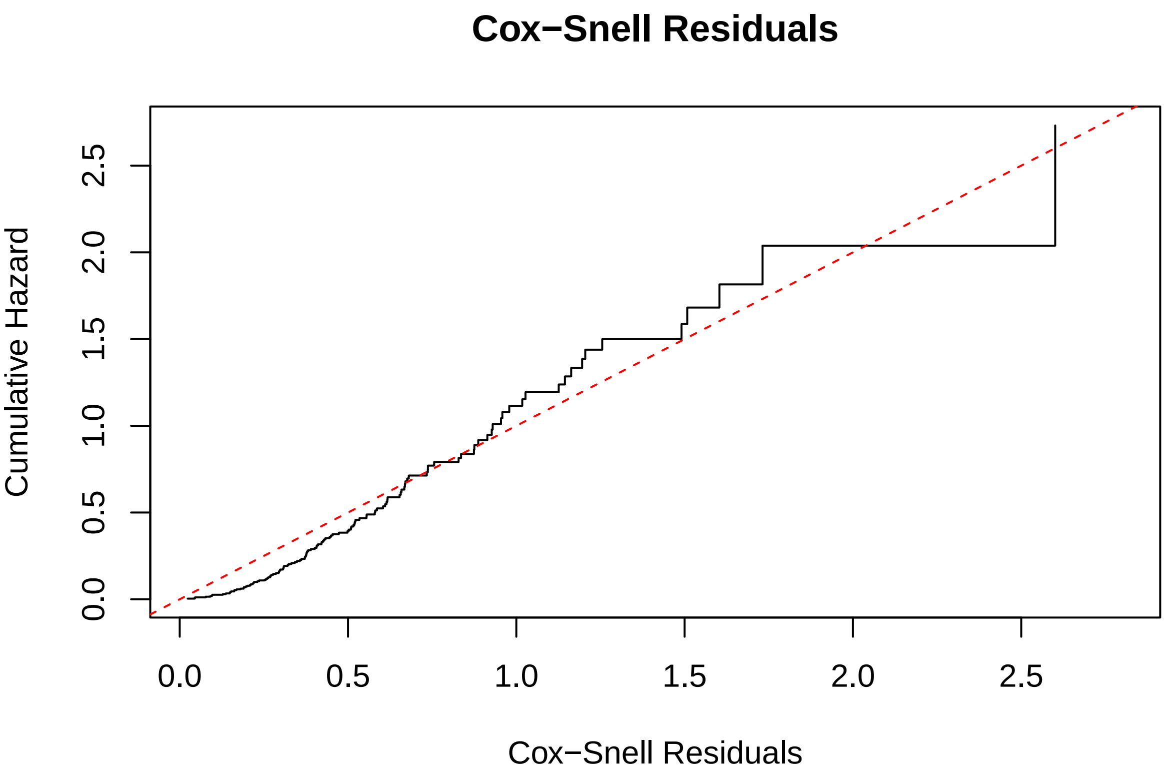


Interpretation:

- Deviance residuals scatter randomly around 0, indicating good model fit.

b. Cox-Snell Residuals

- Cox-Snell residuals are used to assess overall goodness-of-fit for parametric models. They should follow an exponential distribution with a mean of 1 if the model fits well.



Interpretation:

- The cumulative hazard of the Cox-Snell residuals lie close to the 45-degree line (red line), indicating the Weibull model fits well.

(2) Goodness-of-fit tests

a. Likelihood Ratio Test

- Use the anova() function to compare the Weibull model to a simpler model – the exponential model:

```
##                                              Terms Resid. Df    -2*LL    Test Df
## 1 bili + albumin + copper + protime + stage + sex       267 2288.371         NA
## 2         edema + albumin + protime + stage + sex       266 2291.133 1 vs. 2  1
##    Deviance Pr(>Chi)
## 1        NA       NA
## 2 -2.761586       NA
```

Interpretation:

- While the exponential model has a slightly lower deviance (2288.371) compared to the Weibull model (2291.133), the Weibull model is more flexible and is therefore the preferred choice based on these results.

b. Compare AIC values

- To assess whether the Weibull model provides a significantly better fit than the exponential model, compare their AIC values:

```
##               df      AIC
## exp_model      9 2306.371
## weibull_model 10 2311.133
```

Interpretation:

- While the exponential model achieves a slightly lower AIC, the Weibull model is more flexible and is therefore the preferred choice based on these results.