STAT 437 - Assignment 3
Due: Friday, March 18 on Crowdmark
Resubmit: Friday, April 1 on Crowdmark

This assignment covers the materials contained in **Lecture 029** through to **Lecture 045**. Reminder that you are permitted to discuss to these problems with classmates, but every student must submit their own solutions which are their own work (including any code, figures, etc.). **Please indicate any students that you discussed solutions with on your submission**. Please ensure that your submissions on Crowdmark are legible, and separated based on the problems included at the submission link. Submissions can be handwritten or typeset.

Part 1: True or False (20 Marks)

For each of the following problems indicate whether the statement is true or false, and give a short justification for your answer. Correct answers without justification will receive only partial credit.

PROBLEM 1. Suppose you are presented with the following life table results from a discrete time study.

Time	Number at Risk	Number of Events
1	250	22
2	228	13
3	215	5
4	210	26
5	184	11
6	A	B

True or False: From the given information we can conclude that $A = 173$.

Solution 1: [1] False. [1] While there is no censoring observed before time 6, censoring is a key feature in time-to-event data, and without the explicit acknowledgment that there is no censoring, we are not able to determine what A is.

PROBLEM 2. A study was conducted which measured disease-free survival for patients following treatment, measured in years. The study commenced with 337 individuals, and information regarding the number of observed events and individuals being censored during the first four years are included in the following table.

Time	Number Censored	Number of Events
1	7	18
2	15	41
3	48	46
4	24	36

True or False: We can conclude that $\widehat{h}(3) - \widehat{h}(2) = \frac{241}{4992}$.

Solution 2: [1] True. [1] With 337 individuals at risk at time 1, that means there are 312 at risk at time 2 and 256 at risk at at time 3. Then $\widehat{h}(2) = \frac{41}{312}$ and $\widehat{h}(3) = \frac{46}{256}$, which differencing gives the result.

Problem 3. Suppose that T is a continuous, time-to-event, random variable. Moreover, suppose you are given the hazard function of T as $h(t)$. **True or False: This would be sufficient information to compute $E[T]$.**

Solution 3: [1] True. [1] From the hazard function, we can compute the density of the random variable as

$$f(t) = h(t)\exp\left(-\int_0^t h(s)ds\right).$$

As a result, we can write that

$$E[T] = \int_0^\infty th(t)\exp\left(-\int_0^t h(s)ds\right)dt,$$

which is (in principle) able to be solved.

Problem 4. A proportional odds model is fit to some data, which results in the following estimating results:

$$\text{logit}\left\{\widehat{h}(j)\right\} = 0.51D_1 + 0.99D_2 + 0.48D_3 - 0.13D_4.$$

Here, $D_j = I(t = j)$. **True or False: Based on the estimated model, we can find that $\widehat{S}(4) = 0.02068788$.**

Solution 4: [1] True. [1] Recall that $\widehat{S}(4) = \prod_{j=1}^4(1 - \widehat{h}(j))$. We find that

$$\begin{aligned}
\widehat{h}(1) &= \text{expit}(\alpha_1) = \text{expit}(0.51) = 0.6248065 \\
\widehat{h}(2) &= \text{expit}(\alpha_2) = \text{expit}(0.99) = 0.7290879 \\
\widehat{h}(3) &= \text{expit}(\alpha_3) = \text{expit}(0.48) = 0.6177479 \\
\widehat{h}(4) &= \text{expit}(\alpha_4) = \text{expit}(-0.13) = 0.4675457.
\end{aligned}$$

From here, plugging these estimates into the results gives us a cumulative product of 0.02068788.

Problem 5. Suppose you are conducting a study to determine the age at which individuals begin to smoke, and so you begin to survey individuals aged 18+, and then continue following them for the duration of your study. You note that the data you observe may be subject to both left-censoring and right-censoring. Upon completion of the study period, you compute that the sample mean age for beginning to smoke is 23.2. **True or False: This is an underestimate of the true value, since censored data push the mean estimate to be lower than it otherwise would be.**

Solution 5: [1] False. [1] While it is true that right censored data tend to bias the mean downwards, left censored data would have the opposite effect. Without knowing the data in the study it is impossible to say whether the mean is likely to be lower, higher, or near the estimated value.

Problem 6. Statistics Canada reports several key statistics regarding post-secondary students' educational attainment, including average time to graduation[1]. The above page mentions that

> The average time to graduation indicator represents the average number of elapsed academic years new students took to complete the credential in a given educational qualification. [...] students who did not complete the educational qualification within the period of observation were excluded from the average time to graduation calculation.

In other words, students who have not graduated at the end of the observation period are excluded from the calculation of average time to graduation. The observation window is defined to be 1.5 times the typical duration of a program. Assume that no students are lost to follow-up in this study (that is, every student is observed until they either graduate or drop out of post-secondary education). **True or False: It is accurate to describe this as the average time to graduation for post-secondary students in Canada.**

Solution 6: [1] False. [1] This is an example of right-truncated data. In order to be included in the calculation of the average, you must have completed your schooling within the observation window. As a result, in order to accurately report on these results, this **must** be reported as a conditional mean. It would be more accurate to describe this as "the average time to graduation for post-secondary students in Canada who complete their schooling within 1.5 times the typical program duration."

Problem 7. Suppose that you use a proportional odds model where the variate of interest is treatment group. You find that, for the individuals receiving the experimental treatment, $\widehat{h}_E(1) = 0.958$, $\widehat{h}_E(2) = 0.685$, and $\widehat{h}_E(3) = 0.237$. For those who received the placebo treatment, you compute $\widehat{h}_P(2) = 0.312$. **True or False: From this information you know that the proportional odds model would estimate that $\widehat{h}_P(1) = 0.4363445$ and $\widehat{h}_P(3) = 0.1079474$?**

Solution 7: [1] False. [1] In a proportional odds model, the **odds** are proportional to each other, not the hazards themselves. That is,

$$\frac{h_E(j)/(1-h_E(j))}{h_P(j)/(1-h_P(j))} = \beta,$$

for some constant β. These values are computed as though

$$\frac{h_E(j)}{h_P(j)} = \beta,$$

[1] https://www150.statcan.gc.ca/n1/pub/37-20-0001/372000012021004-eng.htm

which is not the case. You could use this relationship to determine that

$$\begin{aligned}
\widehat{\beta} &= 0.2085384 \\
\frac{\widehat{h}_P(1)}{1-\widehat{h}_P(1)} &= 0.2085384 \times \frac{0.958}{1-0.958} = 4.756663 \implies \widehat{h}_P(1) \approx 0.826 \\
\frac{\widehat{h}_P(3)}{1-\widehat{h}_P(3)} &= 0.2085384 \times \frac{0.237}{1-0.237} = 0.06477538 \implies \widehat{h}_P(3) \approx 0.061.
\end{aligned}$$

Problem 8. Suppose that you fit a proportional hazards model and find that the coefficient on a treatment indicator is equal to -2.94. **True or False: This means that the hazard ratio for an individual who received the treatment, compared to one who did not receive the treatment, is equal to -2.94.**

Solution 8: [1] False. [1] The hazard ratio is given by $\exp(-2.94) = 0.05286573$.

Problem 9. Suppose that you get the following output from a model fit using servreg in R.

```
## Call:
## survreg(formula = Surv(time, status) ~ age + factor(ses),
##         data = birth, dist = "lognormal")
##                          Value       Std. Error
## (Intercept)              0.2305      0.07599
## age                      0.0187      0.00276
## factor(ses)lower         0.0512      0.03001
## factor(ses)unknown       0.0824      0.03712
## factor(ses)upper        -0.0188      0.09099
## Log(scale)              -0.5962      0.01756
```

True or False: The implied error distribution from this model is normally distributed with $\sigma^2 = 0.35545444$.

Solution 9: [1] False. [1] The reported value for the log(Scale) parameter is -0.5962, which implies that the scale parameter is $\sigma = \exp(-0.5962) = 0.5509$. As a result, the error distribution will be normal (since it is a lognormal model), however, the variance is going to be given by $0.5509^2 = 0.3035$.

Problem 10. Suppose that you fit both a proportional hazards model and an AFT model to the same dataset. Some of the resulting coefficients are summarized in the following table.

Model	Parameter	Value
Proportional Hazards	Coefficient for Age	0.41
Proportional Hazards	Log(Scale)	0.246
Accelerated Failure Time	Coefficient for Age	-0.89

True or False: It is possible that you used a Weibull distribution for both models.

Solution 10: [1] False. [1] When using a Weibull distribution for AFT and PH models there is a direct correspondence between the estimated parameters. Namely, we find that $\beta^* = -\beta/\kappa$, which in this case would suggest that we should have seen an estimated coefficient for age as -1.6666, in the accelerated failure time model, had a Weibull distribution been used for both models.

Part 2: Conceptual Question (10 Marks)

For the following questions, provide your answers with justification and clear communication. The answers do not need to be long, but correct responses without complete justification will receive only partial credit.

Working as a data scientist, you are tasked to review and help to implement an analysis of your company's sales data. In particular, you have records of the different sales interactions that prospective customers had over a period of time, as well as relevant underlying demographic information. Each customer is recorded as having either made a purchase, having been lost to follow-up, or still active in the sales funnel at the time you are running your study. Sales in your company normally proceed over the course of several months. Your company is interested in using the data collected so far to get insight into the sales funnel, and try to establish best practices going forward. As a result, they want you to make any inferences and suggestions based on the data that you have available today.

PROBLEM 11. (4 Marks) Indicate how the analysis of this data can be formulated in terms of survival analysis. Pay particular attention to specifying the necessary components for survival analysis, and ensuring that you indicate how each possible status of an individual corresponds to those which we can analyze in this framework.

Solution 11: Recall that in survival analysis, the three main components are a baseline, a carefully defined event of interest, and a timescale to work on. In this situation, [0.5] the most natural choice for a baseline is the moment that a prospective customer enters the funnel, [0.5] and months seems to be a reasonable timescale to work on. [1] The event of interest is that a customer makes a purchase of our company's products.
[1] Customers who made purchases are the customers for whom we will have observed an event. [1] Customers who were lost to follow-up, or those who had not yet made a purchase when the data were collected, are considered to be left-censored.

PROBLEM 12. (2 Marks) Your boss thinks that the sample mean time is a good way to capture customers' purchase behaviour. Is this the case? Explain using course concepts.

Solution 12: [1] No, generally we should not rely on means whenever we have data which are subject to censoring. [1] We are going to be underestimating, on average, the amount of time that a customer spends in the funnel since for censored individuals all we know is that the purchase time would have been beyond the censoring time. A better measure of central tendency would be the median, or else a parametric estimate for the mean.

PROBLEM 13. (2 Marks) Part of the sales process involves some customers receiving an in-person sales pitch. This is an incredibly costly process, which management is convinced helps the sales funnel, owing to a few high-profile customers who have made purchases shortly after receiving the pitch. Suppose that you fit an accelerated failure time model, which appears to be a good fit. The estimated coefficient for the in-person sales pitch is 0.70. Does this provide evidence that it is an effective strategy at increasing the speed of the sale?

Solution 13: [1] No. [1] In an accelerated failure time model we can view the survival time as being scaled approximately by e^{η}. If we compare two prospective customers, one who receives the sales pitch and one who does not, the customer that receives the sales pitch will proceed through the funnel approximately half as fast as the customer who did not.
Formally, if T_1 is the time for the individual receiving the sales pitch, and T_0 is the time for the customer not receiving the sales pitch, then $P(T_1 \leq t) = P(T_0 \exp(0.7) \leq t) = P(T_0 \leq 0.5t)$. That means that for any time t, the probability that T_0 has purchased by t will be greater than the probability that T_1 has purchased by t.

Problem 14. (2 Marks) A coworker of yours, who has learned a little bit of survival analysis, suggests that you drop any individuals who are still in the sales funnel at the time of your analysis from consideration. How would this suggestion change your ability to draw conclusions?

Solution 14: [1] This is an example of truncation, and as a result we would need to draw conclusions which are conditional on the truncating criteria. [1] In this case, that would be only people who were observed to have made a purchase **or** lost to follow-up before today's data.
Note: this differs from our usual conception of truncation. Here, individuals are being truncated on a different scale (e.g., related to calendar date) than our problem is being analyzed on. The impact of this may serve to be less severe (if, for instance, the times that customers enter our sales funnel are relatively homogeneous) or it could be more severe (if, for instance, those who were most likely to make purchases were early adopters).

Part 3: Theoretical or Applied Question (20 Marks)

Please **pick one** of the following two problems and solve it. **If you solve both** only the first problem will be looked at. The first problem corresponds to an **application** problem, while the second is a **theoretical question**. For the application question, please provide the code and relevant output (consider using a software like RMarkdown, or being highly selective with what output you copy to ensure your solution is legible). For the theoretical question, please include enough of your work to justify the steps you have taken.

Problem 15. (Application) On the course website you will find a dataset called `stroke.csv`. These data consist of 518 acute stroke patients from a Copenhagen Stroke Study (Jørgensen et al., Stroke, 1996, 27(10): 1765–1769). The survival times are measured in years from the time of hospital admission. The data are recorded in a summary format, summarizing the number of individuals who had a stroke in each year (`d` in the data) out of all individuals at risk (`n` in the data), based on two separate grouping factors. First, individuals are indicated as either having had a previous stroke at the time of admission (`prevstroke` = 1 if yes, otherwise 0) and whether or not the individuals have a history of ischemic heart disease (`ihd` = 1 if yes, otherwise 0). Each row then consists of the number of at risk individuals and the number of events, for each year, within each group.

1. Estimate the survival probabilities for those who experienced a previous stroke, and those who did not, regardless of their IHD status. What is the estimated median survival time for both sets of patients?

2. Consider fitting a proportional odds discrete time model for these data which includes history for IHD, history of stroke, and the interaction. Does the proportional odds assumption seem valid? Discuss using a formal hypothesis test.

   Hint: Note that for grouped data, the GLM call in R is given by

   ```
   glm(cbind(d, n-d) ~ ..., ...)
   ```

   where d is the number of events and n is the total number at risk.

3. Is the model from (2) an appropriate model? Can it be simplified? Considering selection of the most appropriate model starting from the model you fit for (2).

4. Based on the fitted model in (3), discuss the impacts of disease and stroke history on the risk of death at a given year. Provide the corresponding estimates, 95% confidence intervals for these estimates, and a brief interpretation in words.

5. What is the median survival time for patients who have both a history of IHD and a history of previous strokes?

Solution 15:

1. [5 Marks] 2 Marks for the correct values for each grouping, 1 Mark for the correct medians. First, we want to get the summation for each year, regardless of IHD status.

```r
stroke_grouped <- reshape(stroke,
                          v.names = c("n", "d"),
                          idvar = c("prevstroke", "year"),
                          direction = "wide",
                          timevar = "ihd")

# Hazards can be estimated based on total events
# divided by total at risk
hazards <- (stroke_grouped$d.0 + stroke_grouped$d.1)/
            (stroke_grouped$n.0 + stroke_grouped$n.1)

# We include a '1' at the start of the list as S(0) = 1
survival_nostroke <- cumprod(c(1, 1 - hazards[1:10]))
survival_stroke <- cumprod(c(1, 1 - hazards[11:20]))
```

Outputting this in a table for easy reading we see that the corresponding survival probabilities are:

Time	No Previous Stroke	Previous Stroke
0	1.0000	1.0000
1	0.8251	0.7684
2	0.7400	0.6421
3	0.6738	0.5263
4	0.5957	0.4105
5	0.5248	0.3368
6	0.4303	0.2632
7	0.3593	0.2000
8	0.3239	0.1789
9	0.2742	0.1474
10	0.2388	0.1368

In order to get the median, we can see that that $m = 5$ for those without a previous stroke, and $m = 3$ for those with a previous stroke. Applying the corresponding corrections looks like:

```r
m_nostroke <- 5 + (survival_nostroke[6]-0.5)/(survival_nostroke[6]-survival_nostroke[5])
m_stroke <- 3 + (survival_stroke[4] - 0.5)/(survival_stroke[4] - survival_stroke[3])
```

This gives us an estimated median survival time of 5.2625 for those without a history of strokes, and an estimated median of 3.2273 for those with a history of strokes.

2. [5 Marks] 2 Marks for the correct model fit. 2 Marks for the correct hypothesis test and p-value. 1 Mark for the correct conclusion related back to the proportional odds assumption.

To get the model fit, we can use

```
stroke$fyear <- factor(stroke$year)
model.PO <- glm(cbind(d, n-d) ~ -1 + fyear + prevstroke*ihd,
                data = stroke,
                family = binomial)
```

In order to test the PO assumption, we would want to test this model against a saturated model.

```
model.saturated <- glm(cbind(d, n-d) ~ -1 + fyear*prevstroke*ihd,
                       data = stroke,
                       family = binomial)
```

Mathematically, the hypothesis test is equivalent to $H_0 : \beta_4 = \beta_5 = \cdots = \beta_{30} = 0$, where we have ordered the β's based on those that interact with time (there are 27 of them). We can run this in R as

```
anova(model.saturated, model.PO, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(d, n - d) ~ -1 + fyear * prevstroke * ihd
## Model 2: cbind(d, n - d) ~ -1 + fyear + prevstroke * ihd
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1         0      0.000
## 2        27     13.902 -27  -13.902   0.9822
```

The p-value is 0.9822, which means that we do not reject H_0, and as such can conclude that we *can* zero out those coefficients. Note that the saturated model need not be actually fit, as the deviance in a saturated model will always be 0. We conclude that the proportional odds assumption is valid, and can proceed with this model.

3. [3 Marks] 2 marks for correctly testing significance – either Wald or deviance tests are acceptable. 1 mark for correctly selecting the main effects model.

We test whether the proportional odds model can be simplified. Looking at the coefficient summary, we see that the interaction `prevstroke:ihd` does not appear to be significant. We can also test this through a deviance test.

```
model.ME <- update(model.PO, formula = cbind(d, n - d) ~ -1 + fyear + prevstroke + ihd)
anova(model.ME, model.PO, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(d, n - d) ~ fyear + prevstroke + ihd - 1
## Model 2: cbind(d, n - d) ~ -1 + fyear + prevstroke * ihd
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        28     14.449
## 2        27     13.902  1  0.54695   0.4596
```

Considering the results of this model, we see that there are no coefficients that appear as though they can be dropped according to Wald tests. Confirming this via deviance tests we get

```
model.ME.IHD <- update(model.PO, formula = cbind(d, n - d) ~ -1 + fyear + ihd)
model.ME.stroke <- update(model.PO, formula = cbind(d, n - d) ~ -1 + fyear + prevstroke)
anova(model.ME, model.ME.IHD, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(d, n - d) ~ fyear + prevstroke + ihd - 1
## Model 2: cbind(d, n - d) ~ fyear + ihd - 1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        28     14.449
## 2        29     23.741 -1  -9.2915 0.002302 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model.ME, model.ME.stroke, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: cbind(d, n - d) ~ fyear + prevstroke + ihd - 1
## Model 2: cbind(d, n - d) ~ fyear + prevstroke - 1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        28     14.449
## 2        29     18.630 -1  -4.1806  0.04089 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As a result, we select the main effects model as the most appropriate model.

4. [5 Marks] 1 mark for using odds ratios as interpretation. 3 marks for correctly getting odds ratios and confidence intervals. 1 mark for a relevant concluding statement.

Working from the main effects model from (3) we want to consider the effects of the covariates on `prevstroke` and `ihd`. Recall that in a proportional odds model, exponentiating the coefficients gives an odds ratio, and so we consider the odds ratio estimates and corresponding confidence intervals in the following table.

Factor	Odds Ratio Estimate	Lower 95% CI	Upper 95% CI
Previous Stroke	1.5427	1.1763	2.0232
IHD	1.3215	1.0164	1.7181

We would conclude from these results that the odds ratio of having a stroke for individuals with previous stroke history, compared to those without previous stroke history, is 1.5427 with a 95% CI of $(1.1763, 2.0232)$. This suggests a sizable increase in the odds of mortality based on previous stroke history.

We would conclude from these results that the odds ratio of having a stroke for individuals with IHD, compared to those without IHD, is 1.3215 with a 95% CI of $(1.0164, 1.7181)$. This suggests an increase in the odds of mortality based on IHD status.

5. [2 Marks] 1 mark for using the correct process. 1 mark for the correct answer.

We can compute this all directly, based on the estimated survival probabilities. Recall that to get the hazards we take the `expit()` of the estimated quantity.

```
expit <- function(w) { 1/(1 + exp(-w)) }
pred_df <- data.frame(fyear = factor(1:10),
                      ihd = 1,
                      prevstroke = 1)
hazards <- c(0, expit(predict(model.ME, newdata = pred_df)))
survival_probs <- cumprod(1 - hazards)

survival_probs
```

```
##                     1          2          3          4          5          6
## 1.00000000 0.71426645 0.58238108 0.48251021 0.37955112 0.30026756 0.21239707
##          7          8          9         10
## 0.15342861 0.12757937 0.09507222 0.07543634
```

Judging from the results of the survival probabilities, we take $m = 2$ and do the linear correction as

```
m <- 2
med <- m + (survival_probs[3] - 0.5)/(survival_probs[3] - survival_probs[4])
```

As a result the estimated median survival time for an individual with both previous stroke history and a history of IHD is given by 2.825 years, after admission.

Problem 16. (Theoretical) Let T be a continuous random variable with $T \in [0, \infty)$. Take $f(t)$, $F(t)$, $S(t)$, $h(t)$, and $H(t)$ to be the density, cumulative density function, survival function, hazard function, and cumulative hazard function, respectively.

1. Demonstrate that $h(t)F(t) = h(t) - f(t)$.
2. Let $X = \min(T, C)$ where $C > 0$ is a pre-defined constant, and let $Y = H(X)$. Show that $E[Y] = F(C)$.

   Hint: the result from part (1) may be helpful.
3. Show the the cumulative distribution function of $Y = S(T)$ is uniformly distributed on $[0, 1]$.

   Hint: consider what we know about the monotonicity of $F(t)$.
4. Suppose that V has a uniform distribution on $[0, 1]$. Show that $X = -\log V$ has an exponential distribution with hazard rate equal to 1.

Solution 16: 1. [2 Marks] To get this note that

$$h(t)F(t) = \frac{f(t)}{S(t)}(1 - S(t)) = h(t) - f(t).$$

2. [8 Marks] Note that $E[Y] = E[H(X)] = E[H(\min(T, C))] = H(C)P(T > C) + E[H(C)|T < C]P(T < C)$. Also, recall that $H'(t) = h(t)$, by definition. From this we get

$$\begin{aligned}
&E[Y] \\
&= P(T \geq C)H(C) + P(T < C)E[H(T)|T < C] \\
&= S(C)H(C) + F(C)\int_0^C H(t)\frac{f(t)}{F(C)}dt \\
&= [1 - F(C)]\, H(C) + \int_0^C H(t)f(t)dt \\
&= H(C) - F(C)H(C) + H(t)F(t)\Big|_{t=0}^{C} - \int_0^C H'(t)F(t)dt \\
&= H(C) - F(C)H(C) + H(C)F(C) - H(0)F(0) - \left[\int_0^C h(t)dt - \int_0^C f(t)dt\right] \\
&= H(C) - H(C) + F(C) \\
&= F(C).
\end{aligned}$$

3. [5 Marks] We can get this via direct calculation. First note that since T is continuous, we can invert $F(T)$. $F(t)$ is strictly increasing, so $F^{-1}(t)$ is also strictly increasing. Moreover, this implies that $S(T)$ has an inverse, which will be strictly decreasing. Then note that

$$\begin{aligned}
P(Y \leq y) &= P(S(T) \leq y) \\
&= P(T > S^{-1}(y)) && \text{Sign flips since } S^{-1}(t) \text{ is decreasing.} \\
&= S(S^{-1}(y)) \\
&= y.
\end{aligned}$$

This is exactly the uniform CDF on $[0, 1]$.
4. [5 Marks] Consider $X = -\log V$. Then we get that

$$\begin{aligned} P(X \leq x) &= P(-\log V \leq x) \\ &= P(\log V > -x) \\ &= P(V > \exp(-x)) \\ &= 1 - P(V \leq \exp(-x)) \\ &= 1 - \exp(-x). \end{aligned}$$

This is exactly the CDF of an Exp(1) random variable, and as such is exponential with hazard rate 1.