Survival Analysis and Cox Regression

(Jing Li, Miami University)

1. This handout introduces Cox Proportional Hazard Model (Cox Regression), which is commonly used to analyze time-to-event or duration data. In medical science, event is often death or recurrence of disease. In engineering, event can be breakdown of a bus. In economics, event can be finding a job after being unemployed.

2. The focus is T, a non-negative continuous random variable representing time spent in the initial state. T is survival time if the event is death and initial state is being alive. For a bus, the initial state is running without mechanic issue or accident, and $T > t$ means that a bus does not break down before time t. $t < T < t + \Delta t$ means that a bus breaks down during the interval of $(t, t + \Delta t)$.

3. The center of survival analysis is hazard function defined as

$$\begin{aligned} h(t) &= Hazard\ Function \equiv \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} && (1) \\ &= \frac{\lim_{\Delta t \to 0} \frac{P(t<T<t+\Delta t)}{\Delta t}}{P(T > t)} = \frac{\lim_{\Delta t \to 0} \frac{P(T<t+\Delta t)-P(T<t)}{\Delta t}}{P(T > t)} = \frac{\lim_{\Delta t \to 0} \frac{F(t+\Delta t)-F(t)}{\Delta t}}{P(T > t)} && (2) \\ &= \frac{f(t)}{s(t)} && (3) \end{aligned}$$

   where

$$F(t) \equiv P(T < t) \tag{4}$$

   is the cumulative distribution function (CDF);

$$f(t) \equiv \frac{dF(t)}{dt} = \lim_{\Delta t \to 0} \frac{P(T < t + \Delta t) - P(T < t)}{\Delta t} \tag{5}$$

   is the probability density function (PDF), and

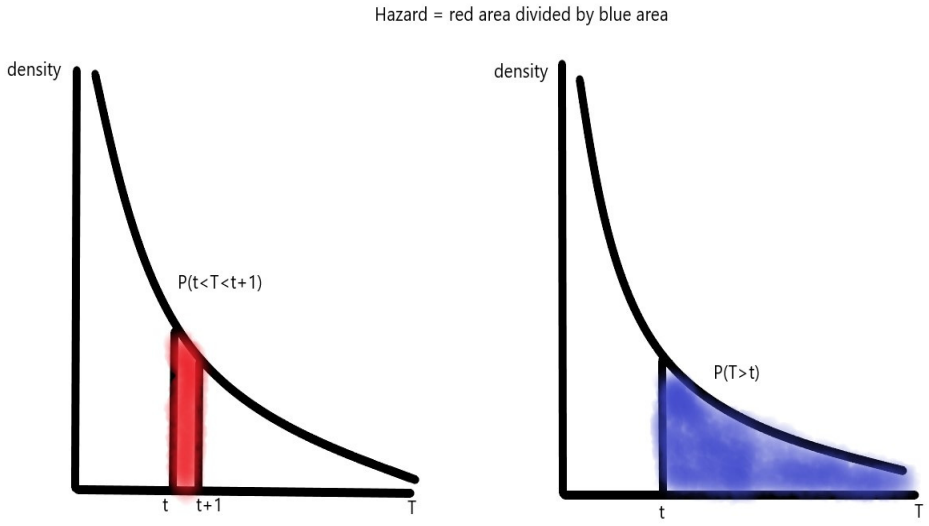
$$s(t) \equiv P(T > t) = 1 - P(T < t) = 1 - F(t) \tag{6}$$

   is survival function.

4. Hazard function (or just hazard) is the instantaneous rate of leaving the initial state (dying, breaking down, finding job...) conditional on that the event has not occurred

before time t. Or, according to (3), you can think of hazard as density normalized by survival function

5. The hazard can be roughly visualized as the red area divided by blue area in the graph (drawn by me, so kind of ugly) below



where we let $\Delta t = 1$, so the red area is $P(t < T < t + 1)$, and blue area is $P(T > t)$. Hazard is a ratio of two probabilities, or ratio of two shared areas. It is clear that hazard is NOT the probability that the event occurs in the next period—that unconditional probability is the red area. Remember, hazard is conditional probability. As $\Delta t \to 0$, hazard becomes conditional density, so could be great than one.

6. (Optional) Equations (5) and (6) imply that

$$f(t) = -\frac{ds(t)}{dt} \tag{7}$$

Then equation (3) implies that

$$h(t) = \frac{f(t)}{s(t)} = \frac{-\frac{ds(t)}{dt}}{s(t)} = -\frac{dln(s(t))}{dt} \tag{8}$$

Given that, we can define the cumulative hazard function as

$$H(t) \equiv \int h(t) = -\ln(s(t)) \tag{9}$$

The math above indicates that a distribution can be characterized by any one of PDF, CDF, hazard function, and survival function. For instance, we can obtain survival function from hazard as follows

$$s(t) = e^{-\int h(t)} \tag{10}$$

and after that, obtain density using (7). The density and survival functions are useful for constructing the likelihood function.

7. Consider a special case of hazard function. If T follows exponential distribution with parameter λ, it follows that

$$\begin{aligned} f(t) &= \lambda e^{-\lambda t} && (11)\\ F(t) &= \int f(t) = 1 - e^{-\lambda t} && (12)\\ s(t) &= 1 - F(t) = e^{-\lambda t} && (13)\\ h(t) &= \frac{f(t)}{s(t)} = \lambda && (14) \end{aligned}$$

   So hazard function is constant λ for exponential distribution.

8. The unknown λ can be estimated by maximum likelihood method. Before specifying the likelihood, we need to distinguish uncensored (actual) survival time and censored survival time. Survival time is censored if information is incomplete—we do not know (i) when the initial state starts (left-censoring) or (ii) when the initial state ends (right-censoring). For a bus, left-censoring means we do not know when is the previous breakdown; right-censoring meaning we do not know when is the next breakdown. Both left censoring and right censoring imply that actual survival time is greater than the reported survival time.

9. Suppose there are j uncensored survival times $t_1, t_2, \ldots, t_j$; and k censored survival times $c_1, c_2, \ldots, c_k$. The log-likelihood for t_i based on (11) is

$$\ln(\lambda e^{-\lambda t_i}) = \ln \lambda - \lambda t_i$$

   The log-likelihood for c_i based on (13) is

$$\ln(e^{-\lambda c_i}) = -\lambda c_i$$

The log likelihood for the whole sample is

$$\sum_{i=1}^{j}(\ln\lambda - \lambda t_i) + \sum_{i=1}^{k}(-\lambda c_i) = j\ln\lambda - \lambda\left(\sum_{i=1}^{j} t_i + \sum_{i=1}^{k} c_i\right)$$

Taking derivative with respect to λ and setting the derivative to zero lead to the maximum likelihood estimate

$$\hat{\lambda} = \frac{j}{\sum_{i=1}^{j} t_i + \sum_{i=1}^{k} c_i} \tag{15}$$

(a) Formula (15) implies that for an exponential distribution, hazard λ can be interpreted as the number of event per period. For instance, suppose total survival time is 100 days (denominator) and only one bus breaks down (numerator), then $\lambda = \frac{1}{100} = 0.01$, or there is 0.01 breakdown per day.

(b) Formula (15) also implies that the estimate would be biased if censored survival time had not been accounted for

$$\frac{j}{\sum_{i=1}^{j} t_i} > \frac{j}{\sum_{i=1}^{j} t_i + \sum_{i=1}^{k} c_i}$$

10. To show how to use (15), consider the sorted bus breakdown data. The first six observations are

	wodate	bus	mileage	driver	route	duration	censor
1	3/24/2023	1101	273136.7	5590	28	82	0
2	12/13/2023	1101	280973.8	5879	19	264	1
4	2/9/2023	1102	314733.6	NULL	NULL	39	0
3	3/1/2023	1102	316468.5	5989	19	20	1
5	9/15/2023	1102	318076.7	6263	41	198	1
6	12/4/2023	1102	322715.2	6289	33	80	1

(a) This is panel data—we observe the same bus several times in year 2023 whenever a working order is placed. The focus is on duration (dependent variable)

(b) For bus 1101, the first duration 82 is difference between March 24 2023 and January 1 2023. It is left-censored since we do not know when is the last breakdown in year 2022. For this censored duration, the binary indicator censor equals 0

(c) For bus 1101, the second duration 264 is difference between December 13 and March 24. This value is not censored, and censor equals 1

(d) Mileage is numeric, while bus, driver, and route are categorical. Those variables can be covariates (predictors)

(e) It is clear that a bus does NOT have fixed route or driver.

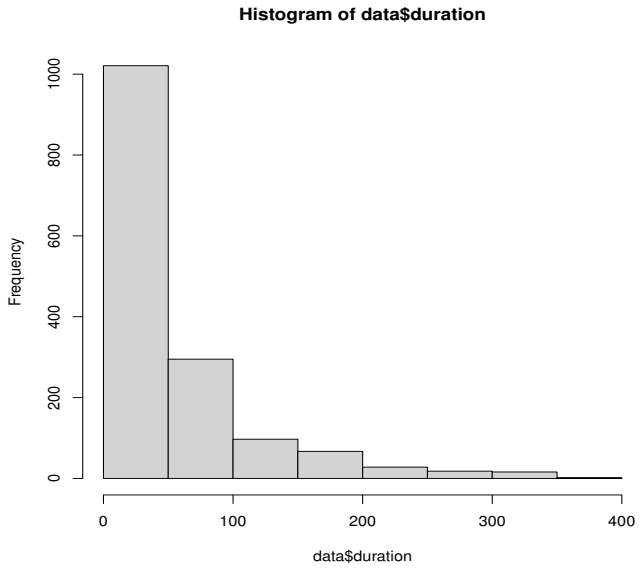
11. The descriptive statistics for duration are

```
> summary(data$duration)
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00    9.00   28.00   51.14   66.00  362.00
```

The minimum of 0 raises a red flag! Average duration is 51, different from the median 28. Next we can use this function

```
hist(data$duration)
```

to draw the histogram of duration. The distribution is NOT bell-shaped (so is not a normal distribution). Instead, it is skewed to right, like an exponential distribution.



12. Assuming duration follows an exponential distribution, then we can estimate λ with formula (15). But first, let's examine data carefully

(a) The data is "dirty" — for instance, there are two working orders placed for bus 1213 on Feb 22. As a result, duration is zero.

```
> data[data$duration==0,][1,]
      wonum    wodate  bus  mileage driver route wodatelag duration censor
144 1875930 2/22/2023 1213 385204.1   5289    78 2/22/2023        0      1
> data[data$bus==1213,][2:3,]
      wonum    wodate  bus  mileage driver route wodatelag duration censor
140 1875931 2/22/2023 1213 385204.1   5289    78  2/4/2023       18      1
144 1875930 2/22/2023 1213 385204.1   5289    78 2/22/2023        0      1
```

(b) Cleaning data amounts to removing observations with 0 duration using **subset** function. After cleaning, we are ready to apply formula (15)

```
> sdata=subset(data, duration>0)
> sum(sdata$censor==1)
[1] 1234
> sum(sdata$duration)
[1] 78960
> lambdahat = sum(sdata$censor==1)/sum(sdata$duration)
> lambdahat
[1] 0.01562817
```

The estimated λ is 0.01562817, implying that there is 0.01562817 breakdown per day (across all buses in the sample).

(c) The same estimate can be obtained with **survreg** in **survival** package.

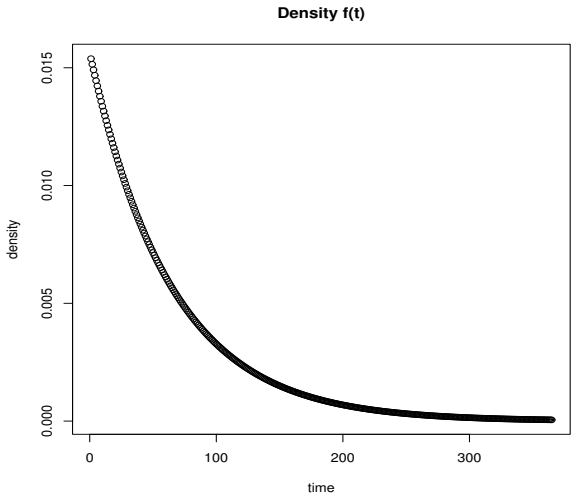
```
> # install.packages("survival")
> library(survival)
> m = survreg(Surv(duration, censor) ~ 1,sdata,dist="exponential")
> lambdahat = exp(-m$coef)
> lambdahat
(Intercept)
 0.01562817
```

(d) Note that $\hat{\lambda} = 0.01562817$ overestimates the true hazard because (i) there may be bus that never broke down in 2023, and their censored survival time 365 is absent in the denominator in (15); (ii) for buses that broke down at least once in

2023, their right-censored survival times are absent in the denominator as well. For instance, for bus 1101, that missing right-censored survival time is difference between December 31, 2023 and December 13, 2023 (date for its last breakdown).

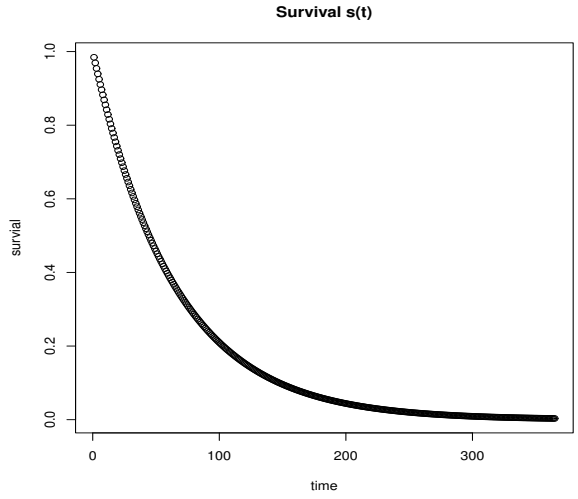
(e) We can use these functions to plot the density function (again, assuming distribution is exponential)

```
time = 1:365
density = lambdahat*exp(-lambdahat*time)
plot(time,density, main="Density f(t)")
```



The vertical axis of density function is hard to interpret. By contrast, interpretation is easier for survival function, which can be computed based on (13)

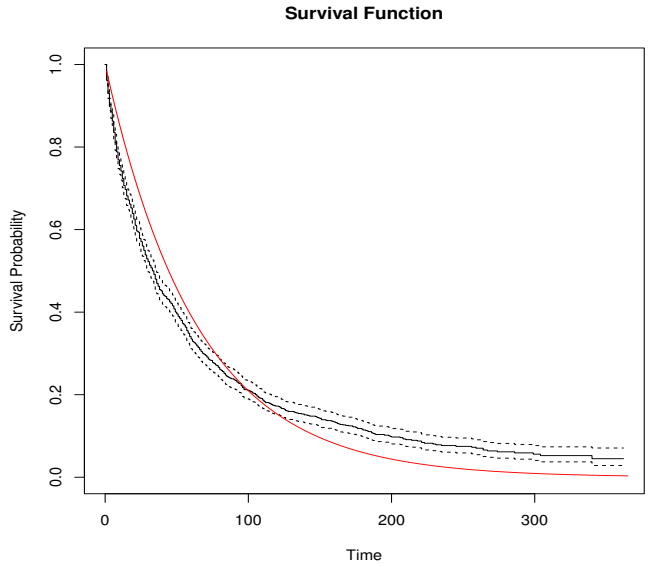
```
time = 1:365
survial = exp(-lambdahat*time)
plot(time,survial, main="Survival s(t)")

> exp(-lambdahat*200)
(Intercept)
0.04390912
> sum(sdata$duration>200)/length(sdata$duration)
[1] 0.04204993
```



The vertical axis tells us the probability of not having breakdown (survival) by that time. For instance, at duration=200, the predicted survival probability is 4.39 percent. The actual survival proportion is 4.20 percent.

(f) Even better, we can put together the actual survival function (with its confidence intervals, black) and predicted survival function (red) assuming exponential distribution

```
fit = survfit(Surv(duration, censor) ~ 1, sdata)
plot(fit, xlab = "Time", ylab = "Survival Probability", main = "Survival Functi
lines(exp(-lambdahat*time),col = "red")
```



13. So far the analysis is univariate—we only look at duration. By contrast, a multivariate analysis is more interesting, and aims to predict duration using covariates x

14. We can start with a simple regression that relates duration to mileage. The pro of linear regression is easy interpretation—the slope coefficient measures the change in duration when mileage changes by one unit. The con is that the regression fails to account for two important facts: (i) the dependent variable duration follows a skewed distribution (could be exponential, Weibull, gamma...in short, non-normality), and (ii) duration can be censored (i.e., actual duration may exceed reported duration).

15. The results of a simple OLS regression are

```
> cor(sdata$mileage,sdata$duration)
[1] -0.3267051
> summary(lm(duration~mileage,sdata))
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.063e+02  4.316e+00   24.62   <2e-16 ***
mileage     -1.768e-04  1.312e-05  -13.48   <2e-16 ***
Multiple R-squared:  0.1067,Adjusted R-squared:  0.1061
F-statistic: 181.6 on 1 and 1520 DF,  p-value: < 2.2e-16
```

    (a) The negative correlation -0.3267051 between duration and mileage agrees with common sense—a bus with high mileage is more likely to break down (having shorter duration) than a bus with low mileage

    (b) The estimated slope coefficient -1.768e-04 is easy to interpret—it implies that mileage rising by 10000 miles is associated with duration falling by 1.768 days

    (c) We can compare a bus with average mileage to a new bus with 0 mileage

```
> summary(sdata$mileage)
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   903.1 211939.1 352272.7 307594.4 397456.2 519648.0
```

    For a new bus, the duration between breakdowns is $307594.4(1.768e-04) = 54.38269$ days longer than the bus with average mileage.

    (d) The intercept term 1.063e+02 implies that the average duration for a bus with 0 mileage is 106.3 days

(e) R-squared is only 0.1067. So mileage does not have much predicative power.

(f) It is easy to add categorical covariate such as route. Again, we need to clean data before running the regression

```
> # add route (remove obs with route=NULL or 20-)
> sort(unique(sdata$route))
 [1] "1"    "11"   "12"   "14"   "15"   "16"   "17"   "19"   "2"    "20"   "20-
[15] "24"   "25"   "27"   "28"   "29"   "3"    "30"   "31"   "32"   "33"   "34"
[29] "40"   "41"   "42"   "43"   "436"  "46"   "49"   "5"    "50"   "51"   "52"
[43] "65"   "67"   "71"   "72"   "74"   "75"   "77"   "78"   "81"   "82"   "85"
> s1data = subset(sdata, route!="NULL" & route!="20-")
> summary(lm(duration~mileage+factor(route),s1data))

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        9.110e+01  2.497e+01   3.648 0.000274 ***
mileage           -1.720e-04  1.394e-05 -12.340  < 2e-16 ***
factor(route)11    2.089e+01  2.570e+01   0.813 0.416525
factor(route)12   -3.699e+01  6.503e+01  -0.569 0.569592
...
factor(route)38    9.392e+01  4.259e+01   2.205 0.027588 *
...
factor(route)999 -4.996e+01  6.504e+01  -0.768 0.442503
---
Multiple R-squared:  0.1341,Adjusted R-squared:  0.1008
F-statistic: 4.034 on 54 and 1407 DF,  p-value: < 2.2e-16
```

i. The results are too long, so only some results are shown here.

ii. Pay attention that **factor** function is used for route in **lm** function because route is categorical (the number has no numerical meaning)

iii. Route 1 is reference group, and its dummy variable is excluded (to avoid dummy variable trap)

iv. The coefficient of other route dummies are difference between that route and route 1. Only the dummy for route 38 is significant at 5% level. Holding mileage constant, route 38 is associated with increasing duration by 93.92

days relative to route 1.

v. The R-squared 0.1341 is still small. Thus, route does not add much predicative power. This finding is not unexpected—a bus does not run a fixed route. In the data, we only know the route when a bus breaks down. We do not know the route when a bus does not breaks down, but those non-breaking-down-routes may contribute to duration, the dependent variable.

(g) It is instructive to compare the R-squared when driver or bus is added as covariates

```
> s2data = subset(sdata, driver!="NULL")
> summary(lm(duration~mileage,sdata))$r.squared
[1] 0.1067362
> summary(lm(duration~mileage+factor(route),s1data))$r.squared
[1] 0.1340521
> summary(lm(duration~mileage+factor(driver),s2data))$r.squared
[1] 0.4266715
> summary(lm(duration~mileage+factor(bus),sdata))$r.squared
[1] 0.4495417
```

According to R-squared, it seems that duration can be explained by driver and bus more than mileage and route.

(h) You should know how to predict duration based on the regression results

16. We are done with the easy but flawed linear regression. Next, we try the nonlinear semi-parametric Cox Proportional Hazard Model (Cox Regression), which is theoretically better than linear regression by accounting for non-normality and censoring of duration. The downside is that interpretation of slope coefficient becomes tricky, and the model is heavy in math (hard to teach/learn).

17. The Cox regression considers a special multiplicative form of hazard function—It does not treat hazard as a constant, instead, it lets the hazard vary when time and covariate change:

$$h(t, x) = \lambda_0(t)e^{\beta x} \tag{16}$$

where $\lambda_0(t)$ is baseline hazard; and $e^{\beta x}$ captures the effect of x on hazard. Note that we can obtain $\lambda_0(t)$ by letting $x = 0$. Thus baseline hazard is the hazard when all covariates are zero.

18. The Cox model is "proportional" since the hazard ratio given by

$$hazard\ ratio \equiv \frac{h(t,x_2)}{h(t,x_1)} = \frac{\lambda_0(t)e^{\beta x_2}}{\lambda_0(t)e^{\beta x_1}} = \frac{e^{\beta x_2}}{e^{\beta x_1}} = e^{\beta(x_2-x_1)} \tag{17}$$

is independent of t. We can call this assumption of time-invariant hazard ratio. It is the key assumption for Cox regression. There are methods that can be used to check this assumption (google or chatgpt)

19. One interpretation of β is based on taking log and derivative of (16)

$$\beta = \frac{d \ln h(t,x)}{dx} \tag{18}$$

So beta measures the effect on log hazard when x changes by one unit. Notice that the numerator in (18) is NOT duration, implying that the magnitude of beta in Cox regression is NOT comparable to linear regression.

20. Log hazard is kind of abstract. An alternative interpretation of β is based on (17): by letting $x_2 - x_1 = 1$, we have

$$e^{\beta} = \frac{h(t,x_2)}{h(t,x_1)} = hazard\ ratio \tag{19}$$

So e^{β} measures the hazard ratio when x changes by one unit. Keep in mind the following

$$\beta = 0 \Rightarrow e^{\beta} = 1 \Rightarrow hazard\ ratio = 1 \tag{20}$$

$$\beta > 0 \Rightarrow e^{\beta} > 1 \Rightarrow hazard\ ratio > 1 \tag{21}$$

$$\beta < 0 \Rightarrow e^{\beta} < 1 \Rightarrow hazard\ ratio < 1 \tag{22}$$

In words, a positive beta implies that hazard increases after x rises by one unit; a negative beta implies that hazard decreases after x rises by one unit. Note that (19) implies

$$h(t,x_2) = e^{\beta} h(t,x_1) \tag{23}$$

Thus new hazard is proportional to old hazard. In practice, many people interpret e^{β} based on (23). For instance, suppose $\beta = 0.2$, then changing x by one unit effectively multiplies hazard by $e^{0.2} = 1.22$

21. The Cox regression is fitted with semi-parametric method since it does not impose particular parametric form for the baseline hazard $\lambda_0(t)$. The R function **coxph** in **survival** package can estimate Cox regression

    (a) First, we use mileage as the sole predictor

```
> library(survival)
> m1 = coxph(Surv(duration, censor) ~ mileage, sdata)
> summary(m1)
n= 1522, number of events= 1234

             coef exp(coef)  se(coef)     z Pr(>|z|)
mileage 3.195e-06 1.000e+00 2.524e-07 12.66   <2e-16 ***

        exp(coef) exp(-coef) lower .95 upper .95
mileage         1          1         1         1

Concordance= 0.593  (se = 0.009 )
Likelihood ratio test= 173.4  on 1 df,   p=<2e-16
Wald test            = 160.2  on 1 df,   p=<2e-16
Score (logrank) test = 164.6  on 1 df,   p=<2e-16

> sum(sdata$censor==1)
[1] 1234
> length(sdata$duration)
[1] 1522
> 12.66^2
[1] 160.2756
```

    R reports that the sample size is 1522, and there are 1234 actual (uncensored) duration. The estimated beta is 3.195e-06. Its sign is the same as linear regression—the positive beta implies that rising mileage leads to greater hazard of breaking-down. The magnitude is not comparable to linear regression. Actually, because 3.195e-06 is close to zero, the hazard ratio reported as exp(coef) is close to unity. The Wald test is just the squared z value (t test). Both the Wald and T tests imply that mileage is statistically significant.

(b) Ask Chatgpt "what are Concordance and Score test reported by coxph in R"

(c) **basehaz** function can show baseline hazard $\lambda_0(t)$ for a bus with zero mileage

```
> head(basehaz(m1))
      hazard time
1 0.03649503    1
2 0.06793957    2
3 0.09589317    3
4 0.11482501    4
5 0.13942846    5
6 0.17019696    6
```

It is clear that the baseline hazard rises as duration rises—indicating that it is more likely for this new bus to break down as duration rises.

(d) It is easy to compute the hazard for nonzero value of mileage based on (16)—we just multiply baseline hazard by $e^{\beta x}$. For instance, for a bus with average mileage, the hazards for the first six duration are

```
> head(basehaz(m1))[,1]*exp((3.195e-06)*mean(sdata$mileage))
[1] 0.09750856 0.18152308 0.25621039 0.30679306 0.37252935 0.45473760

> 0.09750856/0.03649503
[1] 2.671831
```

Focus on duration = 1. Increasing mileage from 0 to the average level more than doubles the hazard

(e) There are two ways to include categorical covariate. First, we can put it in the exponential

$$h(t, x, c) = \lambda_0(t)e^{\beta x+\gamma c} \tag{24}$$

where c denotes a set of dummy variables for the categorical variable and γ is coefficient. For instance, consider adding route as covariate

```
> m2a = coxph(Surv(duration, censor) ~ mileage + factor(route), s1data)
> summary(m2a)

n= 1462, number of events= 1187

                       coef  exp(coef)   se(coef)      z Pr(>|z|)
mileage           3.161e-06  1.000e+00  2.682e-07 11.784  < 2e-16 ***
factor(route)11  -6.632e-01  5.152e-01  4.338e-01 -1.529  0.12632
factor(route)12   8.359e-01  2.307e+00  1.081e+00  0.773  0.43944
...
factor(route)23  -2.260e+00  1.043e-01  1.081e+00 -2.092  0.03647 *
...
Concordance= 0.607  (se = 0.009 )
Likelihood ratio test= 219.6  on 54 df,   p=<2e-16
Wald test            = 66.26  on 54 df,   p=0.1
Score (logrank) test = 265.5  on 54 df,   p=<2e-16

> head(basehaz(m2a))
      hazard time
1 0.05591634    1
2 0.10693185    2
3 0.15137606    3
4 0.18110977    4
5 0.22010828    5
6 0.26681637    6
```

We see that route 23 is significantly different from route 1 (the reference group). Now Concordance rises to 0.607, a good news. The baseline hazard (when mileage = 0, and all route dummies equal zero, or we are talking about route 1) changes accordingly.

(f) Exercise: how to compute the hazard for a bus with average mileage that runs route 23 when duration is 100? Hint: see formula (24)

(g) Alternatively, we can put categorical covariate in the baseline hazard

$$h(t, x, c) = \lambda_0(t, c)e^{\beta x} \tag{25}$$

To do so, we need to apply **strata** function other than **factor** function to route

```
> m2b = coxph(Surv(duration, censor) ~ mileage + strata(route), s1data)
> summary(m2b)

n= 1462, number of events= 1187

             coef exp(coef)  se(coef)     z Pr(>|z|)
mileage 3.027e-06 1.000e+00 2.736e-07 11.06   <2e-16 ***

Concordance= 0.6  (se = 0.011 )
Likelihood ratio test= 132.4  on 1 df,   p=<2e-16
Wald test            = 122.4  on 1 df,   p=<2e-16
Score (logrank) test = 125.5  on 1 df,   p=<2e-16

> head(basehaz(m2b))
      hazard time strata
1 0.34385555    3      1
2 0.58684208    7      1
3 0.91089100   50      1
4 1.59716483   54      1
5 2.98994350  102      1
6 0.02819465    1     11
```

Now the **basehaz** function shows how baseline hazard changes for each given route (strata). Also note that no coefficient is reported for route dummies.

(h) For forecasting purpose, specification (24) allows easier coding than (25)

(i) We can use **cox.zph** to test the assumption of time-invariant hazard ratio (or assumption of proportional hazard)

```
> m1 = coxph(Surv(duration, censor) ~ mileage, sdata)
> cox.zph(m1)
chisq df    p
mileage  2.41  1 0.12
GLOBAL   2.41  1 0.12
```

In this case, the big p-value implies no rejection of that assumption.

22. To learn more about Cox regression, please check out

    https://stats.oarc.ucla.edu/r/seminars/introduction-to-survival-analysis-in-r/

    https://stats.oarc.ucla.edu/wp-content/uploads/2022/05/intro_survival_r_code.r

    In particular, see how to do prediction with Cox regression in the second link.

23. Finally, there are several parametric alternative models to Cox model. For instance, Weibull Model assumes a particular parametric form for the baseline hazard

$$h(t, x) = \lambda_0(t)e^{\beta x}, \quad \lambda_0(t) = \theta\alpha t^{\alpha-1} \tag{26}$$

    The **survreg** function in survival package can be used to estimate Weibull model. I prefer the Cox model over Weibull model because the former is less restrictive.