Modèles de survie en R – Correction

Année universitaire 2025–2026 ▪ Parcours Économie de la santé & Développement durable

Pierre Beaucoral

Notions de base & censure

Données de durée de vie

On s'intéresse à une **durée aléatoire** (T) (positive) :

- durée d'une grève
- durée d'un épisode de chômage
- durée de séjour à l'hôpital
- durée jusqu'à un défaut de paiement, etc.

Pour chaque unité (grève, individu, contrat) $i = 1, \dots, n$:

- T_i : **durée vraie** (inobservable en général)
- on observe seulement :
  - t_i : **durée observée**
  - δ_i : **indicatrice d'événement**

avec :

- $\delta_i = 1$ si l'événement est observé (fin de grève pendant l'observation)
- $\delta_i = 0$ si la durée est **censurée** (on sait juste que la grève dure au moins jusqu'à t_i)

Encodage en R

En R, on encode le couple ((t_i, _i)) avec la fonction `Surv()` du package **survival**.

```
# Exemple jouet : 5 durées, 2 censures
t  <- c(3, 5, 8, 10, 4)          # durées observées
d  <- c(1, 0, 1, 0, 1)           # 1 = événement, 0 = censuré

S_ex <- Surv(time = t, event = d)
S_ex
```

```
[1]  3   5+  8  10+  4
```

- Les + indiquent les durées **censurées**.
- La classe `Surv` est le format standard utilisé par toutes les fonctions d'analyse de survie (`survfit`, `coxph`, `survreg`, etc.).

Fonction de survie, répartition et densité

Fonction de survie

La **fonction de survie** (S(t)) est définie par :

$S(t) = \Pr(T > t)$,

c'est-à-dire :

- la probabilité que la durée **dépasse** t,
- exemple : probabilité que la grève soit encore en cours au jour t,
- c'est une fonction **décroissante** de t,
- $S(0) = 1$, et $\lim_{t\to\infty} S(t) = 0$ (souvent).

Fonction de répartition

La **fonction de répartition** (CDF) est :

$F(t) = \Pr(T \leq t) = 1 - S(t)$.

- $F(t)$ augmente de 0 (au temps 0) vers 1 (quand $t \to \infty$).
- $F(t)$ = probabilité que l'événement soit arrivé **avant** ou **à** la date t.

Densité (cas continu)

Si (T) est une variable continue admettant une **densité** (f(t)) :

$f(t) = \frac{d}{dt}F(t)$.

Lien entre (f) et (S) :

$f(t) = -\frac{d}{dt}S(t) \quad \iff \quad S(t) = 1 - \int_0^t f(u), du$.

Intuition :

- (f(t)) décrit la répartition de la masse de probabilité dans le temps ;
- (S(t)) cumule **ce qui n'est pas encore arrivé**.

Fonction de risque (hazard)

Définition

La **fonction de risque** (hazard) (h(t)) est définie par :

$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$.

Interprétation :

- probabilité **instantanée** que l'événement se produise juste après (t),
- **conditionnellement** à ce que l'événement ne soit pas encore arrivé à (t).

Pour la durée d'une grève :

> $h(t)$ mesure, au temps (t), le risque que la grève se termine **immédiatement**, sachant qu'elle est toujours en cours à t.

Lien entre $h(t)$, $f(t)$ et $S(t)$

On peut montrer que :

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

Intuition :

- $f(t)$ = densité où l'événement arrive **à** (t),
- $S(t)$ = probabilité que l'événement ne soit pas encore arrivé **avant** t,
- donc $h(t)$ = probabilité de finir maintenant **sachant qu'on est encore en vie** à t.

On définit aussi la **cumulative hazard** :

$\Lambda(t) = \int_0^t h(u), du$.

On a alors le lien fondamental :

$S(t) = \exp\left(-\Lambda(t)\right) = \exp\left(-\int_0^t h(u), du\right)$.

Exemples de lois de durée

Exemple 1 – Risque constant : loi exponentielle

Supposons un risque constant $h(t) = \lambda > 0$.

Alors :

- $\Lambda(t) = \int_0^t \lambda, du = \lambda t$
- $S(t) = \exp(-\lambda t)$.

C'est la **loi exponentielle** :

- la probabilité de survie décroît de façon **exponentielle**.

Interprétation :

- Le « risque de fin de grève » est le même quel que soit t.
- C'est souvent trop simple mais utile comme cas de base.

Exemple 2 – Loi de Weibull

La loi de **Weibull** est très utilisée en analyse de survie.

Risque :

$h(t) = \lambda p t^{p-1}$,

- si (p = 1) : on retrouve le cas exponentiel (risque constant),
- si (p > 1) : **risque croissant** (plus le temps passe, plus la fin est probable),
- si (p < 1) : **risque décroissant** (l'événement est surtout probable au début).

C'est un modèle paramétrique flexible pour décrire des durées où le risque évolue dans le temps.

Censure à droite

Pourquoi de la censure ?

En pratique, on ne connaît pas toujours la date exacte de l'événement :

- la période d'observation se termine alors que la grève continue,
- l'individu est **perdu de vue** (perte de suivi),
- les données sont tronquées (ex. base administrative clôturée à une date donnée).

On introduit une **durée de censure** (C_i) :

- temps jusqu'à la fin d'observation pour l'unité (i).

On observe alors :

$t_i = \min(T_i, C_i), \qquad \delta_i = \mathbb{1}(T_i \leq C_i).$

Hypothèse clé : censure non informative

On suppose généralement que la censure est **non informative** :

> Le mécanisme de censure ne dépend pas de la durée vraie, conditionnellement aux covariables.

Autrement dit :

- être censuré ou non ne doit pas apporter d'**information supplémentaire** sur la durée résiduelle, au-delà de ce qu'on connaît déjà (covariables).

Exemple :

- on suit des grèves jusqu'au 31 décembre,
- celles qui dépassent cette date sont censurées,
- mais la date du 31/12 est fixée **indépendamment** des caractéristiques des grèves.

Si la censure est informative (ex. on arrête d'observer quand l'entreprise fait faillite), les méthodes classiques (KM, Cox) peuvent être biaisées.

Exemple simple en R

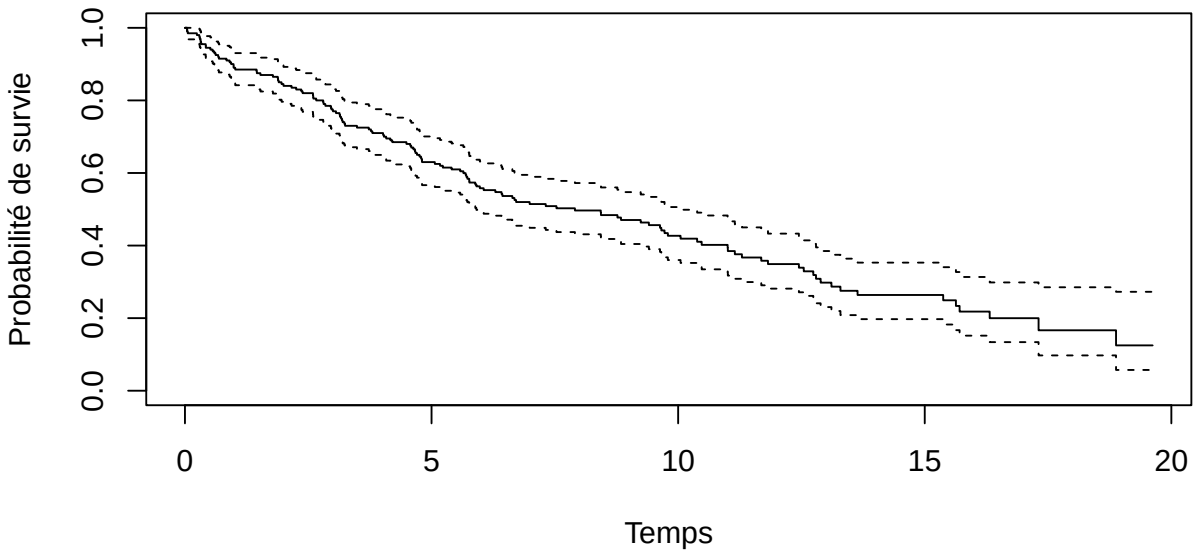
Simuler des durées exponentielles censurées :

```r
set.seed(123)

n      <- 200
T_true <- rexp(n, rate = 0.1)   # durées vraies (exponentielle)
C      <- runif(n, min = 5, max = 20)  # temps de censure

time   <- pmin(T_true, C)
status <- as.integer(T_true <= C)

S_sim  <- Surv(time = time, event = status)
head(S_sim)
```

```
[1]  8.4345726  5.7661027 13.2905487  0.3157736  0.5621098  3.1650122
```

Tracer un Kaplan–Meier :

```
km_sim <- survfit(S_sim ~ 1)
plot(km_sim, xlab = "Temps", ylab = "Probabilité de survie")
```



- la courbe tient compte à la fois des événements observés et des observations censurées.

Objet de survie en R et interprétation

Surv() en pratique

Pour des données de grève, on construirait :

```
# Exemple conceptuel (ici, on simule juste pour illustrer)
time   <- c(7, 9, 13, 14, 26, 29)
status <- c(1, 1, 1, 1, 1, 0)

S_strikes_ex <- Surv(time = time, event = status)
S_strikes_ex
```

```
[1]  7   9  13  14  26  29+
```

- Chaque ligne est soit `durée` (événement) soit `durée+` (censure).
- Cet objet est ensuite utilisé par toutes les fonctions de survie.

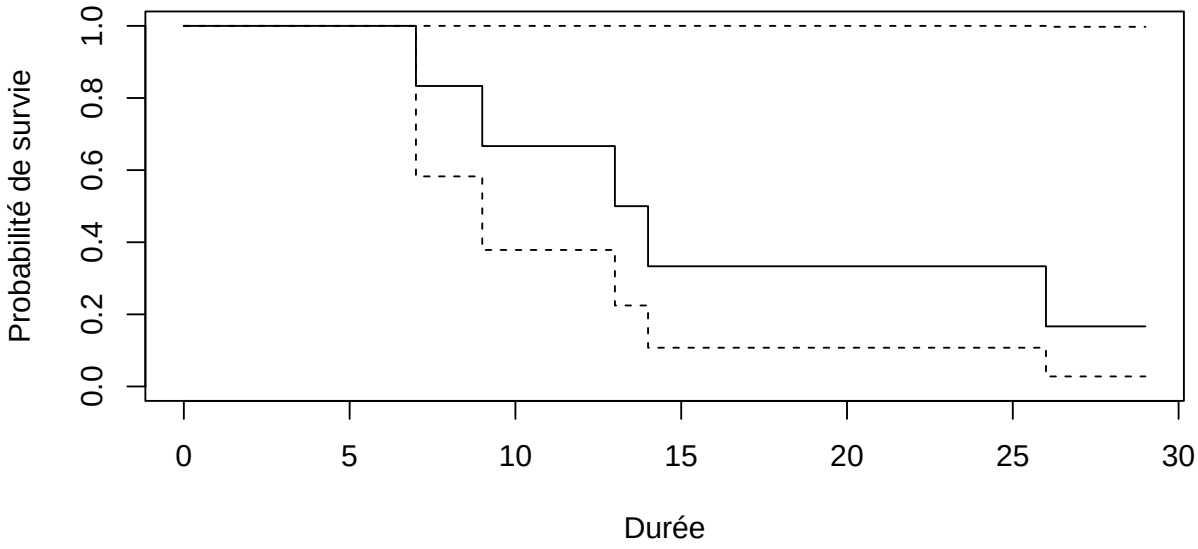
Kaplan–Meier dans R

Pour estimer la fonction de survie :

```
km_fit <- survfit(S_strikes_ex ~ 1)
km_fit
```

```
Call: survfit(formula = S_strikes_ex ~ 1)

     n events median 0.95LCL 0.95UCL
[1,] 6      5   13.5       9      NA
```

```
plot(km_fit, xlab = "Durée", ylab = "Probabilité de survie")
```



- Les « marches » de la courbe correspondent aux événements observés.
- Les observations censurées contribuent au **dénominateur** (nombre à risque), mais pas au numérateur (nombre d'événements).

Résumé des notions clés

À retenir

1. **Durée** (T) et censure :
   - on observe (t_i, δ_i), pas toujours la durée vraie.
2. **Fonction de survie** $S(t) = \Pr(T > t)$:
   - probabilité d'être encore « en vie » au temps (t).
3. **Risque (hazard)** $h(t)$) :
   - probabilité instantanée de l'événement, sachant qu'il n'est pas encore arrivé.
4. Liens fondamentaux :
   - $F(t) = 1 - S(t)$,
   - $h(t) = f(t)/S(t)$,
   - $S(t) = \exp\left(-\int_0^t h(u), du\right)$.
5. **Censure à droite** :
   - on observe $\min(T_i, C_i))et(\mathbb{1}(T_i \leq C_i)$,
   - hypothèse clé : censure **non informative**.

Pour la suite du cours

- Estimation de $S(t)$ par l'**estimateur de Kaplan–Meier**.
- Comparaison de courbes de survie entre groupes (test du **log-rank**).
- Modèles à covariables :
  - modèle de Cox (risques proportionnels),
  - modèles paramétriques (exponentiel, Weibull, …).

Exercices – Correction (survie de grève)

Présentation des données et analyse descriptive

On s'intéresse à la **durée de grèves** dans l'industrie manufacturière américaine.

Les données proviennent de :

Kennan, J. (1985), *The Duration of Contract Strikes in U.S. Manufacturing*, Journal of Econometrics.

Nous utiliserons la base `StrikeDuration` du package **AER** dans R, qui contient :

- `duration` : durée de la grève (en jours)
- `uoutput` : choc d'activité non anticipé (mesure de conjoncture macroéconomique)

Q1 – Correspondance entre codages du cycle

On considère les 62 grèves de la base `StrikeDuration`.

On a construit :

- `uoutput_sign` : 3 modalités (cycle défavorable, neutre, favorable)
- `uoutput_q` : 3 modalités (choc faible, moyen, fort, par terciles)

```
with(strikes, table(uoutput_sign, uoutput_q))
```

```
                  uoutput_q
uoutput_sign       Choc faible Choc moyen Choc fort
  Cycle défavorable          25          0         0
  Cycle favorable             0         17        20
```

```
prop.table(with(strikes, table(uoutput_sign, uoutput_q)), 1)
```

```
                  uoutput_q
uoutput_sign       Choc faible Choc moyen Choc fort
  Cycle défavorable   1.0000000  0.0000000 0.0000000
  Cycle favorable     0.0000000  0.4594595 0.5405405
```

Commentaires :

- Tous les épisodes en **cycle défavorable** sont classés en `Choc faible`.
- Les épisodes en **cycle favorable** se répartissent entre `Choc moyen` et `Choc fort`.
- La matrice est quasi diagonale : les deux codages racontent la **même histoire** (choc défavorable vs favorable), l'un en version "signe", l'autre en version "force du choc".
- On peut donc considérer que les deux échelles sont **cohérentes**.

Q2 – Corrélations entre variables explicatives

Variables explicatives potentielles :

- `uoutput` (continu)
- `uoutput_q_num` : version numérique de `uoutput_q` (1 = faible, 2 = moyen, 3 = fort)

```
strikes <- strikes |>
  mutate(uoutput_q_num = as.numeric(uoutput_q))

cor(dplyr::select(strikes, uoutput, uoutput_q_num), use = "complete.obs")
```

```
               uoutput uoutput_q_num
uoutput       1.0000000     0.9245881
uoutput_q_num 0.9245881     1.0000000
```

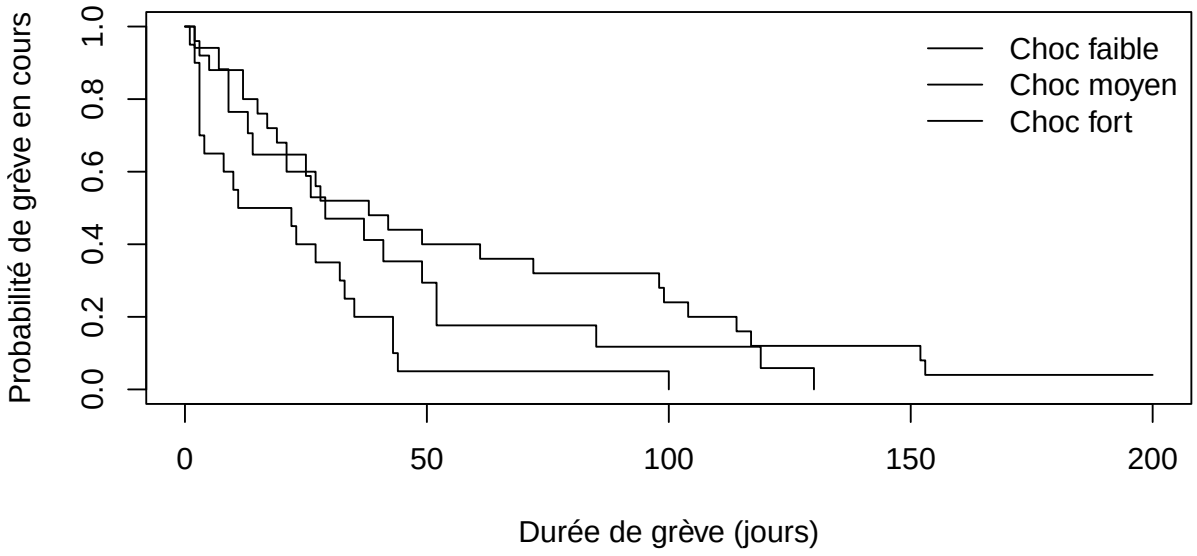
Commentaires :

- On obtient une **corrélation positive forte** entre `uoutput` et `uoutput_q_num` (la variable en terciles est une discrétisation monotone de `uoutput`).

- La version catégorielle n'apporte donc pas d'information entièrement nouvelle, mais elle permet :
  - de capturer plus facilement des **effets non linéaires** du choc,
  - de produire des comparaisons simples (`Choc faible` vs `Choc fort`) en termes de **hazard ratios**.

Q3 – Courbes de survie brutes (Kaplan–Meier)

```
km_cycle <- survfit(S_strikes ~ uoutput_q, data = strikes)
plot(km_cycle, xlab = "Durée de grève (jours)", ylab = "Probabilité de grève en cours")
legend("topright", legend = levels(strikes$uoutput_q), lty = 1, bty = "n")
```



Commentaires :

- Pour les **chocs forts** (conjoncture très favorable), la courbe de survie décroît plus vite : les grèves ont tendance à être **plus courtes**.
- Pour les **chocs faibles** (conjoncture défavorable), la courbe décroît plus lentement : les grèves restent **plus longtemps en cours**.
- Les courbes sont bien séparées, ce qui suggère un **effet du contexte macroéconomique** sur la durée des grèves.

Q4 – Test du log-rank

```
survdiff(S_strikes ~ uoutput_q, data = strikes)
```

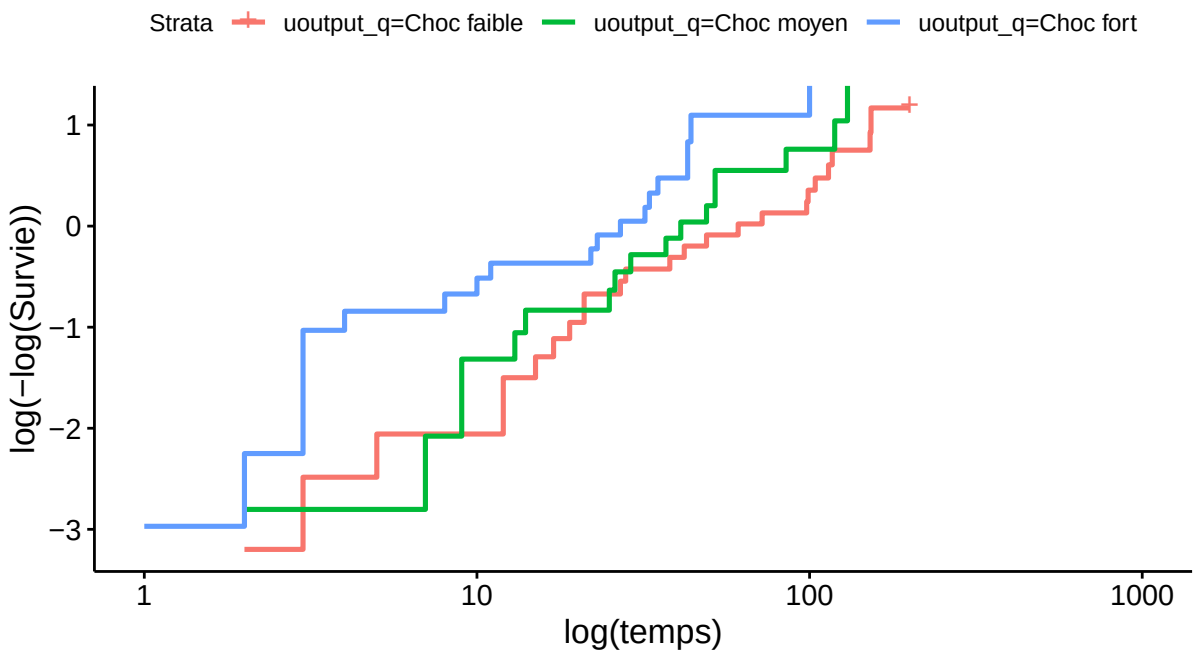
```
Call:
survdiff(formula = S_strikes ~ uoutput_q, data = strikes)

                       N Observed Expected (O-E)^2/E (O-E)^2/V
uoutput_q=Choc faible 25       24     32.4  2.198928  5.143950
uoutput_q=Choc moyen  17       17     16.9  0.000153  0.000219
uoutput_q=Choc fort   20       20     11.6  6.074740  8.183882

 Chisq= 9.2  on 2 degrees of freedom, p= 0.01
```

Commentaires :

- Hypothèses du test du log-rank :
  - H_0 : les fonctions de survie sont **identiques** dans les 3 groupes de choc (`Choc faible`, `Choc moyen`, `Choc fort`).
  - H_1 : au moins une fonction de survie **diffère**.
- On obtient une statistique χ^2 d'environ 9,2 avec 2 ddl et une p-valeur 0,01.
- Au seuil de 5 %, on **rejette** H_0 : la durée des grèves dépend significativement du niveau de choc d'activité.
- En regardant les contributions :
  - `Choc faible` : moins de fins de grève observées que prévu → grèves **plus longues**.
  - `Choc fort` : plus de fins de grève observées que prévu → grèves **plus courtes**.

Q5 – Vérification graphique des risques proportionnels

```
if (requireNamespace("survminer", quietly = TRUE)) {
  ggsurvplot(
    survfit(S_strikes ~ uoutput_q, data = strikes),
    fun = "cloglog",
    xlab = "log(temps)",
    ylab = "log(-log(Survie))"
  )
}
```



Commentaires :

- Les courbes `log(-log(S(t)))` par niveau de `uoutput_q` sont grossièrement **quasi parallèles** (sans croisements extrêmes).
- L'hypothèse de **risques proportionnels** pour `uoutput_q` est donc **raisonnablement plausible** dans ce jeu de données.
- On peut donc utiliser un **modèle de Cox** avec cette variable, tout en gardant en tête que l'échantillon est petit.

Modèle de Cox – Correction

Q6 – Modèle de Cox continu

Modèle ajusté :

```
cox_full <- coxph(S_strikes ~ uoutput + uoutput_q_num, data = strikes)
summary(cox_full)
```

```
Call:
coxph(formula = S_strikes ~ uoutput + uoutput_q_num, data = strikes)

  n= 62, number of events= 61

                 coef exp(coef) se(coef)     z Pr(>|z|)
uoutput         5.5249  250.8571  10.3206 0.535    0.592
uoutput_q_num   0.2007    1.2223   0.5385 0.373    0.709

               exp(coef) exp(-coef) lower .95 upper .95
uoutput          250.857   0.003986 4.117e-07 1.529e+11
uoutput_q_num      1.222   0.818155 4.254e-01 3.512e+00

Concordance= 0.608  (se = 0.041 )
Likelihood ratio test= 8.57  on 2 df,   p=0.01
Wald test            = 8.5  on 2 df,   p=0.01
Score (logrank) test = 8.83  on 2 df,   p=0.01
```

```
AIC(cox_full)
```

```
[1] 389.1666
```

Équation du modèle :

$h_i(t) = h_0(t), \exp\left(\beta_1 uoutput_i + \beta_2 uoutput_{q,i}\right)$

Interprétation du coefficient de `uoutput` :

- Le terme $\exp(\eta_1)$ est le **hazard ratio** associé à une augmentation d'une unité de `uoutput`.
- Comme `uoutput` est un choc macro **petit en amplitude**, on interprète plutôt de petites variations (par ex. aller d'un quantile faible à un quantile élevé).
- Si $\exp(\eta_1) > 1$, un choc plus favorable est associé à une **fin plus rapide** de la grève (durées plus courtes).

AIC :

- On retient la valeur d'AIC affichée et on la comparera aux autres spécifications (M1, M3).
- Un AIC **plus faible** signale un meilleur compromis ajustement / complexité.

Q7 – Version purement catégorielle du cycle

```
cox_cat <- coxph(S_strikes ~ uoutput_q, data = strikes)
summary(cox_cat)
```

```
Call:
coxph(formula = S_strikes ~ uoutput_q, data = strikes)

  n= 62, number of events= 61

                      coef exp(coef) se(coef)     z Pr(>|z|)
uoutput_qChoc moyen 0.3556    1.4271   0.3253 1.093  0.27435
uoutput_qChoc fort  0.9643    2.6229   0.3265 2.954  0.00314 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                    exp(coef) exp(-coef) lower .95 upper .95
uoutput_qChoc moyen     1.427     0.7007    0.7543     2.700
uoutput_qChoc fort      2.623     0.3813    1.3832     4.974

Concordance= 0.604  (se = 0.039 )
Likelihood ratio test= 8.47  on 2 df,   p=0.01
Wald test            = 8.81  on 2 df,   p=0.01
Score (logrank) test = 9.3  on 2 df,   p=0.01
```

Équation du modèle :

$h_i(t) = h_0(t)\, \exp\left(\eta_2\, \mathbb{1}\{uoutput_q = \text{"Choc moyen"}\} + \eta_3\, \mathbb{1}\{uoutput_q = \text{"Choc fort"}\}\right).$

avec `Choc faible` comme catégorie de référence.

Interprétation :

- $\exp(\eta_2)$:hazard ratio "Choc moyen" vs "Choc faible".
- $\exp(\eta_3)$: hazard ratio "Choc fort" vs "Choc faible".
- Si au moins un coefficient est significatif, on conclut à une **différence de durée** entre au moins deux niveaux de choc.
- En pratique, avec ce petit échantillon, les intervalles de confiance sont larges → significativité parfois fragile, mais la direction va vers des **grèves plus rapides** en cas de choc plus fort.

Comparaison des déviances :

- On lit les log-vraisemblances de `cox_full` et `cox_cat` dans les sorties et l'on compare :
  - une **déviance plus faible** ($-2\log L$) indique un meilleur ajustement (à nombre de paramètres donné).

Q8 – Nombre de paramètres

Modèles :

- M1 : `S ~ uoutput`
  - 1 paramètre de pente (η_1).
- M2 : `S ~ uoutput + uoutput_q_num`
  - 2 paramètres de pente ((η_1, η_2).
- M3 : `S ~ uoutput_q` (facteur 3 modalités)
  - 2 paramètres de pente (deux dummies vs référence : `Choc moyen`, `Choc fort`).

Emboîtement :

- M1 est **inclus dans** M2 si l'on impose $eta_2 = 0$.
- M3 n'est pas emboîté dans M1/M2 (spécification complètement catégorielle vs continue).

Q9 – Choix de la meilleure mesure du cycle

```
cox_M1 <- coxph(S_strikes ~ uoutput, data = strikes)
cox_M2 <- coxph(S_strikes ~ uoutput + uoutput_q_num, data = strikes)
cox_M3 <- coxph(S_strikes ~ uoutput_q, data = strikes)

anova(cox_M1, cox_M2, test = "LRT")
```

```
Analysis of Deviance Table
 Cox model: response is  S_strikes
 Model 1: ~ uoutput
 Model 2: ~ uoutput + uoutput_q_num
   loglik  Chisq Df Pr(>|Chi|)
1 -192.65
2 -192.58 0.1389  1     0.7094
```

```r
anova(cox_M1, cox_M3, test = "LRT")
```

```
Analysis of Deviance Table
 Cox model: response is  S_strikes
 Model 1: ~ uoutput
 Model 2: ~ uoutput_q
   loglik  Chisq Df Pr(>|Chi|)
1 -192.65
2 -192.63 0.0382  1      0.845
```

Commentaires :

- LR(M1 vs M2) : si la p-valeur est élevée → l'ajout de `uoutput_q_num` ne permet pas une amélioration significative → on peut préférer le modèle plus simple M1.
- LR(M1 vs M3) : si la p-valeur est faible → la version **catégorielle** du choc (M3) capture un effet non linéaire utile.
- On combine ces tests avec les AIC pour choisir :
  - soit une spécification **continue simple** (M1),
  - soit une spécification **catégorielle** (M3) si les effets par classe sont plus lisibles et mieux ajustés.

---

3. Modèles paramétriques – Correction

Q10 – Modèle de Weibull

```r
weib_best <- survreg(S_strikes ~ uoutput, data = strikes, dist = "weibull")
summary(weib_best)
```

```
Call:
survreg(formula = S_strikes ~ uoutput, data = strikes, dist = "weibull")
               Value Std. Error     z      p
(Intercept)  3.79012    0.13944 27.18 <2e-16
uoutput     -9.67700    3.00825 -3.22 0.0013
Log(scale)   0.00631    0.10180  0.06 0.9506

Scale= 1.01

Weibull distribution
Loglik(model)= -285.4   Loglik(intercept only)= -290.2
    Chisq= 9.6 on 1 degrees of freedom, p= 0.002
Number of Newton-Raphson Iterations: 6
n= 62
```

Nombre de paramètres estimés :

- Intercept (α)
- Coefficient de `uoutput` (η_1)
- Paramètre de **scale** (lié au paramètre de forme du Weibull)

→ Au total, **3 paramètres** dans ce modèle.

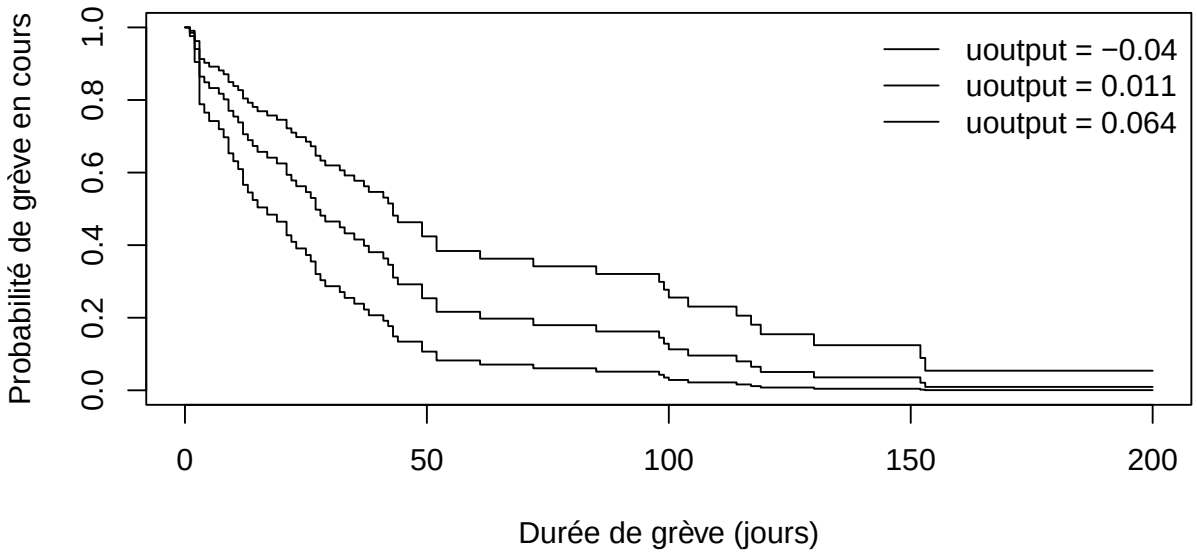
Comparaison AIC :

- On calcule l'AIC du Weibull et on le compare à l'AIC du Cox retenu :
  - AIC(Weibull) plus faible → modèle paramétrique préféré.
  - AIC(Cox) similaire ou plus faible → on peut rester en Cox, plus souple (baseline non paramétrique).

Q11 – Courbes de survie pour différents scénarios

On utilise ici le modèle de Cox simple `cox_uoutput` pour illustrer les différences de survie selon le choc :

```r
newdata <- data.frame(
  uoutput = quantile(strikes$uoutput, probs = c(.2, .5, .8))
)

surv_cox <- survfit(cox_uoutput, newdata = newdata)

plot(
  surv_cox,
  xlab = "Durée de grève (jours)",
  ylab = "Probabilité de grève en cours"
)

legend(
  "topright",
  legend = paste0("uoutput = ", round(newdata$uoutput, 3)),
  lty = 1,
  bty = "n"
)
```



Commentaires :

- Pour les valeurs de `uoutput` **élevées** (choc favorable), la courbe de survie est en dessous des autres : les grèves ont une probabilité plus faible de "survivre" longtemps → elles se terminent plus vite.

- Pour les valeurs **faibles / négatives** (choc défavorable), la courbe de survie est au-dessus : les grèves ont une probabilité plus forte de durer.
- La forme du risque implicite (hazard) augmente avec le temps au début, puis peut se stabiliser, ce qui est cohérent avec un modèle de type Weibull.

Q12 – Interprétation économique

En combinant :

- les **hazard ratios** des modèles de Cox,
- les **courbes de survie** (Kaplan–Meier et prédictions de modèles),

on peut conclure :

- En **période de conjoncture favorable** (choc d'activité positif / fort), les grèves ont plus de chances de se terminer rapidement → **hazard plus élevé**, survie plus faible.
- En **période de conjoncture défavorable**, les grèves ont tendance à durer plus longtemps → **hazard plus faible**, survie plus élevée.

En termes de négociation :

- Quand l'activité est bonne, les employeurs ont davantage de marges (les pertes liées à la grève sont plus coûteuses à court terme), ce qui peut favoriser des **accords plus rapides**.
- Quand la conjoncture est mauvaise, les coûts d'opportunité sont plus faibles et les rapports de force différents, ce qui peut conduire à des grèves **plus longues**.