TD 4 – Modèles de survie

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

Pierre Beaucoral

Contexte et données

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)

On construit les variables suivantes :

- une durée possiblement censurée :
  $time_i = \min(duration_i, 200)$
- une indicatrice d'événement (fin de grève observée) :

  $status_i = \mathbb{1}\ (duration_i \leq 200)$
- deux codages de la conjoncture :
  - `uoutput_sign` : cycle défavorable / favorable (signe de `uoutput`)
  - `uoutput_q` : choc faible / moyen / fort (terciles de `uoutput`)

Présentation des données et analyse descriptive

Exercice 1 – Correspondance entre codages du cycle

On construit :

- `uoutput_sign` : 2 modalités (cycle défavorable, favorable)
- `uoutput_q` : 3 modalités (choc faible, moyen, fort, par terciles)

1. Donner la **table de contingence** entre ces deux codages.
2. Calculer les **proportions par ligne**.
3. Commentez : les deux « échelles » du cycle sont-elles cohérentes ?

Indication R :

> **Indice :** commencer par construire un tableau de contingence avec `table(uoutput_sign, uoutput_q)`, puis utiliser `prop.table(..., margin = 1)` pour obtenir les proportions par ligne.

Exercice 2 – Corrélations entre variables explicatives

On considère comme variables explicatives potentielles :

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

1. Construire la **matrice de corrélation** entre ces deux variables.
2. Commentez : la version catégorielle apporte-t-elle une information réellement nouvelle ?

Indication R :

> utiliser la commande `corr()`

Exercice 3 – Courbes de survie brutes

On trace les **courbes de survie Kaplan–Meier** de la durée de grève selon `uoutput_q`.

1. Représenter graphiquement les courbes $\widehat{S}(t)$ pour chacun des trois niveaux de choc.
2. Commentez : que constate-t-on sur les différences de durée de grève selon le choc d'activité ?

Indication R :

Commencez par créer l'objet de survie avec `Surv(time, status)`, puis estimez un Kaplan–Meier par groupe avec `survfit(Surv_obj ~ uoutput_q, data = …)`. Utilisez ensuite `plot()` pour tracer les courbes et `legend()` pour afficher les trois niveaux de choc.

Exercice 4 – Test du log-rank

On souhaite tester l'effet global de `uoutput_q` sur la durée de grève à l'aide d'un **test du log-rank**.

1. Écrire précisément les hypothèses H_0 et H_1.
2. Calculer la statistique du test et la p-valeur.
3. Conclure au seuil de 5 % sur l'effet global du choc d'activité.

Indication R :

Utilisez la fonction `survdiff(Surv(time, status) ~ uoutput_q, data = ...)` pour réaliser le test du log-rank. La sortie vous donne la statistique de type chi^2 , les degrés de liberté et la p-valeur, à interpréter pour conclure au seuil de 5 %.

Exercice 5 – Vérification graphique de l'hypothèse de risques proportionnels

On trace les courbes de $\log(-\log(\hat{S}(t)))$ par niveau de `uoutput_q`.

1. Les courbes semblent-elles approximativement **parallèles** ?
2. Que suggère ce résultat sur la **plausibilité** du modèle de Cox avec cette variable ?

Indication R :

Utilisez `ggsurvplot()` du package **survminer** en lui passant un objet `survfit(Surv(...) ~ uoutput_q, ...)` et l'argument `fun = "cloglog"` pour obtenir les courbes de $\log(-\log(\hat{S}(t)))$ par groupe, puis regardez si elles sont à peu près parallèles.

Modèle de Cox

Exercice 6 – Modèle de Cox avec uoutput et uoutput_q_num

On ajuste un modèle de Cox ne contenant que les variables explicatives suivantes :

- `uoutput` (continu)
- `uoutput_q_num` (version numérique de `uoutput_q`)

1. Écrire l'équation du modèle sous la forme :

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

2. Estimer ce modèle dans R.
3. Interpréter le **hazard ratio** associé à `uoutput`.
4. Calculer le **critère AIC** du modèle et le noter pour comparaison.

Indication R :

> ajustez un modèle de Cox avec `coxph(Surv(time, status) ~ uoutput + uoutput_q_num, data = ...)`, puis utilisez `summary()` pour lire les coefficients et interpréter les hazard ratios, et `AIC()` pour comparer ce modèle aux autres spécifications.

Exercice 7 – Version purement catégorielle du cycle

On ajuste un modèle de Cox où la conjoncture est mesurée seulement par la variable qualitative `uoutput_q` :

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

1. Écrire l'équation du modèle :

   $h_i(t) = h_0(t), \exp\left(\beta 2, \mathbb{1}uoutputq, i = \text{Choc moyen} + \beta 3, \mathbb{1}uoutputq, i = \text{Choc fort}\right).$

2. À partir de `summary(cox_cat)`, dire s'il existe une **différence significative** de durée de grève entre au moins deux niveaux de choc.
3. Comparer la **déviance** $(-2 \log L)$ de ce modèle à celle du modèle précédent.

Exercice 8 – Nombre de paramètres et emboîtement des modèles

On considère trois modèles :

- M1 : `S ~ uoutput`
- M2 : `S ~ uoutput + uoutput_q_num`
- M3 : `S ~ uoutput_q` (facteur 3 modalités)

1. Donner le **nombre de paramètres de pente** (hors baseline) dans chacun des trois modèles.
2. Quelles relations d'**emboîtement** existe-t-il entre ces modèles (qui est inclus dans qui) ?

Exercice 9 – Choix de la meilleure mesure du cycle

On veut déterminer quelle représentation de la conjoncture (continu vs factorisé) prédit le mieux la durée des grèves.

1. Utiliser des **tests de rapport de vraisemblance** (LR) pour comparer :
   - M1 vs M2
   - M1 vs M3
2. Combiner l'information des tests LR et des **AIC** pour choisir la mesure du choc d'activité que vous retiendriez.
3. Justifier votre choix d'un point de vue **économique** et **économétrique**.

Indication R :

> estimez séparément les trois modèles de Cox correspondant à M1, M2 et M3, puis utilisez des tests de rapport de vraisemblance via `anova(mod_simple, mod_complex, test = "LRT")` pour comparer les spécifications. Interprétez la p-valeur et l'évolution de l'AIC pour décider si le modèle plus riche améliore significativement l'ajustement par rapport au modèle plus simple.

Modèles paramétriques

Exercice 10 – Modèle de Weibull

On ajuste un modèle de Weibull avec la (ou les) variable(s) explicative(s) retenue(s) à la question 9.

1. Estimer un modèle de Weibull à l'aide de `survreg()` :
2. Préciser le **nombre de paramètres** estimés (y compris les paramètres de forme / échelle).
3. Comparer son **AIC** à celui du modèle de Cox retenu précédemment.

Indication R (exemple avec uoutput seul) :

> utilisez `survreg(Surv(time, status) ~ uoutput, data = ..., dist = "weibull")` pour estimer un modèle de Weibull, puis regardez dans `summary()` quels paramètres sont estimés (intercept, coefficient de `uoutput`, paramètre de scale) et utilisez `AIC()` pour comparer ce modèle au modèle de Cox retenu.

Exercice 11 – Courbes de survie et risque instantané

À partir du modèle de durée retenu (Cox ou Weibull) :

1. Tracer les **courbes de survie** pour différents scénarios de conjoncture (par ex. valeurs faibles, médianes et fortes de `uoutput`).
2. Tracer (ou commenter) les **taux de risque instantané** correspondants.
3. Commenter la forme du risque dans le temps (croissant ? décroissant ?).

Indication possible avec le modèle de Cox simple :

> Commencez par définir un petit `data.frame newdata` avec plusieurs valeurs représentatives de `uoutput` (par exemple des quantiles de la distribution), puis utilisez `survfit(mod_cox, newdata = newdata)` pour obtenir les courbes de survie prédites. Tracez-les avec `plot()` et ajoutez une `legend()` pour comparer visuellement la probabilité de fin de grève selon le niveau de choc.

Exercice 12 – Interprétation économique

Sur la base des résultats obtenus (Kaplan–Meier, modèles de Cox, Weibull) :

1. Dans quel contexte conjoncturel les grèves ont-elles le plus de chances d'être **courtes** ?
2. Dans quel contexte ont-elles le plus de chances d'être **longues** ?
3. Discuter ces résultats en termes :
   - de **hazard ratios** (probabilité instantanée de fin de grève)
   - et de **fonctions de survie** (probabilité de grève encore en cours).

Relier cette interprétation aux mécanismes économiques de **négociation salariale** et de **coûts de la grève** pour les syndicats et les employeurs.