

La covariable qui améliore chaque statistique et détruit la réponse

Ajoutez le fait qu'une orientation ait été émise et l'AIC chute de 595, le pseudo- R^2 quadruple, et 42,5 % de l'écart selon le handicap disparaît. Tout critère de sélection de modèle dit de la garder. Toute considération causale dit qu'elle doit sortir, et aucune statistique d'ajustement ne vous dira laquelle a raison.

Pierre Robentz Cassion

Leçon 5 sur 8

Régression appliquée aux données de programme — Quelles covariables ont leur place

Ce que couvre cette leçon

- Une covariable qui paraît manifestement juste
- Pourquoi elle retire l'effet
- Quatre sortes de covariables, et ce que chacune fait
- Le tracer avant d'ajuster
- Le collisionneur, qui est pire
- Rapportez les deux, et dites lequel est lequel
- La suite

Une covariable qui paraît manifestement juste — En Python

```
import statsmodels.formula.api as smf

BASE = ("completed ~ disability + case_category + age_band + sex"
        " + service_requested + admin1")

adjusted = smf.logit(BASE, data=d).fit(dis=0)
plus_gate = smf.logit(BASE + " + referral_made", data=d).fit(dis=0)

for name, m in [("adjusted", adjusted), (" + referral_made", plus_gate)]:
    print(f"{name:16} AIC {m.aic:7.1f}    pseudo-R2 {m.prsquared:.3f}")
```

Une covariable qui paraît manifestement juste — En R

```
adjusted <- glm(completed ~ disability + case_category + age_band + sex +  
                service_requested + admin1, data = d, family = binomial())  
plus_gate <- update(adjusted, . ~ . + referral_made)  
AIC(adjusted, plus_gate)
```

Une covariable qui paraît manifestement juste

Modèle	Rapport de cotes	Effet marginal moyen	AIC	Pseudo-R ²
Brut	0,430	−19,05 pts	2138,5	0,012
Ajusté	0,388	− 19,41 pts	1999,6	0,089
+ referral_made	0,494	− 11,15 pts	1404,2	0,365

Une covariable qui paraît manifestement juste

- L'AIC s'améliore de 595 points. Le pseudo- R^2 quadruple. Et 42,5 % de l'effet s'évanouit

Pourquoi elle retire l'effet — En Python

```
print(d.groupby("disability")["referral_made"].agg(["mean", "size"]).round(4))
```

Pourquoi elle retire l'effet — En R

```
d |> summarise(made = mean(referral_made), n = n(), .by = disability)
```

Pourquoi elle retire l'effet

- Une orientation est émise pour 71,5 % des cas sans handicap signalé et 53,8 % des cas avec
- Conditionner sur `referral_made` demande : parmi les cas où une orientation a été émise, reste-t-il un écart ? — La...
- Une covariable située sur le chemin causal entre exposition et résultat est un médiateur, et ajuster dessus retire exactement la part de l'effet qui passe par elle

Quatre sortes de covariables, et ce que chacune fait

Sorte	Se situe	Ajuster ?	Si vous vous trompez
Facteur de confusion	Cause l'exposition et le résultat	Oui	Estimation biaisée, dans les deux sens
Médiateur	Entre l'exposition et le résultat	Non (pour un effet total)	Effet sous-estimé
Collisionneur	Causé par l'exposition et le résultat	Jamais	Biais créé là où il n'y en avait pas
Cause concurrente	Ne cause que le résultat	Facultatif	Précision seulement

Quatre sortes de covariables, et ce que chacune fait

- Seule la première ligne est une raison d'ajouter une variable — La quatrième est une raison pour laquelle vous pouvez. . .
- La sorte d'une variable n'est pas visible dans les données — Facteur de confusion, médiateur et collisionneur. . .

Le tracer avant d'ajuster — Exemple

```
disability -----> completion
|                               ^
+-----> referral made -----+
```

department --> disability department --> completion

- Le département est un facteur de confusion — le lieu où naît un cas influe à la fois sur qui est enregistré avec un. . .

Le tracer avant d'ajuster — En Python

```
TOTAL_EFFECT = ["disability", "case_category", "age_band", "sex",  
                "service_requested", "admin1"]          # confounders only  
MECHANISM = TOTAL_EFFECT + ["referral_made"]           # a different question
```

Le tracer avant d'ajuster — En R

Two named model specifications, two questions, both legitimate.

Le tracer avant d'ajuster

- Nommez les modèles d'après la question plutôt que d'après les variables — Un modèle appelé MECHANISM ne sera pas. . .

Le collisionneur, qui est pire — En Python

```
closed = d[d["case_status"].isin(["closed-resolved", "closed-lost-contact"])]  
print(f"closed cases: {len(closed)}")  
print(closed.groupby("disability")["completed"].agg(["mean", "size"]).round(4))
```

Le collisionneur, qui est pire — En R

```
d |> filter(case_status %in% c("closed-resolved", "closed-lost-contact")) |>  
  summarise(rate = mean(completed), n = n(), .by = disability)
```

Le collisionneur, qui est pire

Échantillon	n	Écart (effet marginal moyen)
Tous les cas consentis	1 581	–19,4 points
Cas clos seulement	860	–21,1 points

Le collisionneur, qui est pire

- Se restreindre aux cas clos porte l'écart à 21,1 points — et le mouvement est produit par la restriction plutôt que par...
- Sélectionner des lignes, c'est ajuster sur une variable — Écarter les cas ouverts, ne garder que les formulaires...

Rapportez les deux, et dites lequel est lequel — Exemple

Referral completion by disability status, decomposition

Total gap	-19.4 points
of which runs through referral-making	-8.3 points (42.5%)
remaining, among referrals made	-11.2 points

Referrals are made for 53.8% of cases reporting a disability against 71.5% of others. Both gates contribute; neither explains the other away.

The -11.2 figure is conditional on a referral having been made and must not be reported as the effect of disability on reaching a service.

Rapportez les deux, et dites lequel est lequel

- La dernière ligne est celle qui garde la décomposition honnête — Les deux nombres sont réels, ils répondent à des...

- Chaque modèle de ce cours a jusqu'ici supposé que chaque ligne était une observation indépendante.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/
reg-05-the-covariate-that-hides-the-answer](https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/reg-05-the-covariate-that-hides-the-answer)