Résumé

Ce guide[1] vous aidera à déterminer quand il est judicieux d'essayer de "contrôler d'autres choses" lors de l'estimation de l'effet du traitement à l'aide de données expérimentales. Nous nous concentrons sur les grandes idées et fournissons des exemples en R.

1 Qu'est-ce que l'ajustement de covariable ?

Les "covariables" sont les caractéristiques de base de vos sujets expérimentaux. Lorsque vous exécutez une expérience, vous êtes principalement intéressé par la collecte de données des variables de résultat que votre intervention peut affecter, i.e. les décisions de dépenses, les attitudes envers la démocratie ou les contributions à un bien public dans une expérience de laboratoire. Mais c'est aussi une bonne idée de collecter des données sur les caractéristiques de base des sujets avant l'assignation du traitement, i.e. le sexe, le niveau d'éducation ou le groupe ethnique. Si vous faites cela, vous pouvez explorer comment l'effet du traitement varie en fonction de ces caractéristiques (voir 10 choses à savoir sur les effets de traitement hétérogènes (https://egap.org/resource/10-things-to-know-about-heterogeneous-treatment-effects)). Mais cela vous permet également d'effectuer un ajustement de covariable.

L'ajustement de covariable est un autre nom pour contrôler les variables de base lors de l'estimation de l'effet du traitement. Souvent, cela est fait pour améliorer la précision. Les résultats des sujets sont susceptibles d'avoir une certaine corrélation avec des variables qui peuvent être mesurées avant l'assignation aléatoire. La prise en compte de variables telles que le sexe vous permettra de mettre de côté la variation des résultats qui est prédite par ces variables de base, afin que vous puissiez isoler l'effet du traitement sur les résultats avec plus de précision et de puissance.

L'ajustement des covariables peut être un moyen moins coûteux d'améliorer la précision plutôt que d'augmenter le nombre de sujets dans l'expérience. C'est en partie pour cette raison que les chercheurs recueillent souvent de nombreuses données sur les covariables avant l'assignation aléatoire. Les pré-tests (mesures analogues à la variable de résultat mais limitées aux périodes précédant l'assignation aléatoire) peuvent être particulièrement utiles pour prédire les résultats, et les enquêtes initiales peuvent interroger les sujets sur d'autres caractéristiques de base.

2 Contrôle des covariables au stade du design (découpage par bloc)

La meilleure façon de contrôler les covariables est d'utiliser la randomisation par bloc pour le faire au stade du design avant même de commencer votre expérience. La randomisation par bloc vous permet de créer des groupes de traitement et de contrôle équilibrés pour certaines covariables. Par exemple, vous pourriez vous attendre à ce que le sexe et le revenu aident à prédire la variable de résultat. La randomisation par bloc peut

garantir que les groupes de traitement et de contrôle ont des proportions égales de femmes/à revenu élevé, de femmes/à faible revenu, d'homme/à revenu élevé et d'homme/à faible revenu. Lorsque les variables utilisées pour les blocs aident à prédire les résultats, le découpage par bloc améliore la précision en empêchant les corrélations aléatoires entre l'assignation du traitement et les covariables de base.

Pour plus d'informations sur le découpage par bloc et comment l'implémenter dans R, voir 10 choses à savoir sur la randomisation (https://egap.org/resource/10-things-to-know-about-randomization). Les gains de précision du découpage par bloc (par rapport à l'ajustement de covariable sans bloc) ont tendance à être plus importants lorsque la taille des échantillons est petite.[2]

Lorsque le découpage par bloc est effectué pour améliorer la précision, l'erreur type estimée doit tenir compte du découpage. (Sinon, l'erreur type aura tendance à être conservatrice car elle ne vous donnera pas le crédit de l'amélioration de la précision obtenue par le découpage.) Une méthode simple et couramment utilisée consiste à régresser le résultat sur la variable muette d'assignation du traitement ainsi que sur les variables muettes de bloc. Lorsque la probabilité d'assignation au traitement est constante d'un bloc à l'autre, l'inclusion des variables muettes de bloc dans la régression ne modifie pas l'effet du traitement estimé, mais tend à donner une estimation plus précise de l'erreur type.[3]

Si la probabilité d'assignation au traitement varie d'un bloc à l'autre, vous devez alors contrôler ces probabilités inégales afin d'obtenir des estimations non biaisées de l'effet moyen du traitement. 10 choses à savoir sur la randomisation (https://egap.org/resource/10-things-to-know-about-randomization) l'aborde de manière pratique.

3 Comment le faire dans une régression ?

Parfois, vous n'avez pas la possibilité de mettre en œuvre un design expérimental par bloc (i.e., si vous rejoignez un projet après l'assignation aléatoire) ou vous préférez simplifier votre schéma de randomisation pour réduire les risques d'erreur administrative. Vous pouvez toujours ajuster les covariables de base en utilisant la régression multiple. N'oubliez pas que dans une régression bivariée — lorsque vous régressez votre résultat uniquement sur votre indicateur de traitement — le coefficient de traitement n'est qu'une différence des moyennes. Cette méthode simple donne une estimation non biaisée de l'effet moyen du traitement (ATE). Lorsque nous ajoutons au modèle des covariables de base corrélées aux résultats, le coefficient de traitement est une estimation approximativement non biaisée de l'ATE qui a tendance à être plus précise que la régression bivariée.

Pour ajuster les covariables par régression multiple, utilisez le modèle :

$$Y_i = \alpha + \beta Z_i + \gamma X_i + \epsilon_i$$

où Y_i est la variable de résultat, Z_i est l'indicateur de traitement et X_i est un vecteur d'une ou plusieurs covariables. Le reste ϵ_i est votre terme de perturbation — le bruit inexpliqué restant.

Lorsque les groupes de traitement et de contrôle sont de taille inégale, les gains de précision de l'ajustement des covariables peuvent être plus importants si vous incluez les interactions entre le traitement et les covariables (voir ce blog (https://web.archive.org/web/20151024055802%20/http://blogs.worldbank.org/impactevaluations/node/847) pour plus d'information). Pour faciliter l'interprétation, recentrez les covariables pour avoir une moyenne nulle :

$$Y_i = \alpha + \beta Z_i + \gamma W_i + \delta Z_i * W_i + \epsilon_i$$

où $W_i = X_i - \overline{X}$ et $\overline{X}$ est la valeur moyenne de X_i pour l'ensemble de l'échantillon.

Si les sujets reçoivent différentes probabilités d'assignation au traitement en fonction de leurs covariables, alors notre méthode d'estimation doit en tenir compte (encore une fois, voir 10 choses à savoir sur la randomisation (https://egap.org/resource/10-things-to-know-about-randomization) pour plus de détails).

4 Pourquoi le faire ?

Il n'est pas absolument nécessaire de contrôler les covariables lors de l'estimation de l'effet moyen du traitement dans un ECR qui attribue à chaque sujet la même probabilité de recevoir le traitement. La différence des moyennes traitement-contrôle non ajustée pour les résultats est un estimateur sans biais de l'ATE. Cependant, l'ajustement des covariables a tendance à améliorer la précision si les covariables sont de bons prédicteurs du résultat.[4]

Dans les grands échantillons, l'assignation aléatoire a tendance à produire des groupes de traitement et de contrôle avec des caractéristiques de base similaires. Pourtant, par "chance du tirage au sort", un groupe peut être légèrement plus éduqué, ou un groupe peut avoir des taux de vote légèrement plus élevés lors des élections précédentes, ou un groupe peut être légèrement plus âgé en moyenne. Pour cette raison, l'ATE estimé est soumis à une "variabilité d'échantillonnage", ce qui signifie que vous obtiendrez des estimations de l'ATE qui ont été produites par une méthode non biaisée, mais qui ont manqué la cible.[5] Une variabilité d'échantillonnage élevée contribue au bruit (imprécision), pas au biais.

Le contrôle des covariables a tendance à améliorer la précision si les covariables sont prédictives des résultats potentiels. Jetez un œil à l'exemple suivant, qui est vaguement basé sur l'expérience de Giné et Mansuri sur le comportement électoral des femmes au Pakistan.[6] Dans cette expérience, les auteurs ont randomisé une campagne d'information auprès de femmes au Pakistan pour étudier ses effets sur leur participation, l'indépendance de leur choix de candidat et leurs connaissances politiques. Ils ont réalisé une enquête de base qui leur a fourni plusieurs covariables.

Le code suivant imite cette expérience en créant de fausses données pour quatre des covariables qu'ils collectent : la possession d'une carte d'identité, la scolarité, l'âge et l'accès à la télévision. Cela crée également deux résultats potentiels (https://egap.org/resource/10-things-to-know-about-causal-inference) (les résultats qui se produiraient si elle était assignée au traitement ou non) pour mesurer à quel point la femme a voté indépendamment de l'opinion des hommes de sa famille. Les résultats potentiels sont corrélés avec les quatre covariables, et l'effet de traitement "réel" sur la mesure d'indépendance est ici de 1. Pour déterminer si notre estimateur est biaisé ou non, nous simulons 10 000 répétitions de notre expérience. À chaque répétition, nous assignons le traitement de manière aléatoire, puis régressons le résultat observé Y sur l'indicateur de traitement Z, avec et sans contrôle des covariables. Ainsi, nous simulons deux méthodes (non corrigée et corrigée par les covariables) pour estimer l'ATE. Pour estimer le biais de chaque méthode, nous prenons la différence des moyennes pour 10 000 estimations simulées et le "vrai" effet de traitement.

```
rm(list=ls())

set.seed(20140714)
N = 2000
N.treated = 1000
Replications = 10000

true.treatment.effect = 1

# Créer des covariables de pré-traitement
owns.id.card = rbinom(n = N, size = 1, prob = .18)
has.formal.schooling = rbinom(n = N, size = 1, prob = .6)
age = round(rnorm(n = N, mean = 37, sd = 16))
age[age<18] = 18
age[age>65] = 65
TV.access = rbinom(n = N, size = 1, prob = .7)
epsilon = rnorm(n = N, mean = 0, sd = 2)

# Créer des résultats potentiels corrélés aux covariables pré-traitement
Y0 = round(owns.id.card + 2*has.formal.schooling + 3*TV.access + log(age) + epsilon)
Y1 = Y0 + true.treatment.effect

# Répéter l'assignation du traitement
Z.mat = replicate(Replications, ifelse(1:N %in% sample(1:N, N.treated), 1, 0))

# Générer des résultats observés
Y.mat = Y1 * Z.mat + Y0 * (1 - Z.mat)

diff.in.means = function(Y, Z) {
  coef(lm(Y ~ Z))[2]
}

ols.adjust = function(Y, Z) {
  coef(lm(Y ~ Z + owns.id.card + has.formal.schooling + age + TV.access))[2]
}

unadjusted.estimates = rep(NA, Replications)
adjusted.estimates   = rep(NA, Replications)

for (i in 1:Replications) {
  unadjusted.estimates[i]  =  diff.in.means(Y.mat[,i], Z.mat[,i])
  adjusted.estimates[i]    =  ols.adjust(Y.mat[,i], Z.mat[,i])
}

# Variabilité estimée (écart type) de chaque estimateur
sd.of.unadj = sd(unadjusted.estimates)
sd.of.unadj
sd.of.adj   = sd(adjusted.estimates)
sd.of.adj

# Biais estimé de chaque estimateur
mean(unadjusted.estimates) - true.treatment.effect
mean(adjusted.estimates) - true.treatment.effect

# Marge d'erreur (avec un niveau de confiance de 95 %) pour chaque biais estimé
```

```
1.96 * sd.of.unadj / sqrt(Replications)
1.96 * sd.of.adj   / sqrt(Replications)
```

Les deux méthodes — avec et sans covariables — donnent le véritable effet de traitement de 1 en moyenne. Lorsque nous avons exécuté la régression sans covariables, notre ATE estimé était en moyenne de 1,0008 sur les 10 000 répétitions, et avec les covariables, il était en moyenne de 1,0003. Notez que l'estimation ajustée par régression est essentiellement sans biais même si notre modèle de régression est mal spécifié — nous contrôlons l'âge de manière linéaire lorsque le véritable processus de génération de données implique le logarithme de l'âge.[7]

Les vrais gains viennent de la précision de nos estimations. L'erreur type (l'écart type de la distribution d'échantillonnage) de notre ATE estimé lorsque nous ignorons les covariables est de 0,121. Lorsque nous incluons des covariables dans le modèle, notre estimation devient un peu plus stricte : l'erreur type est de 0,093. Parce que nos covariables étaient prédictives de nos résultats, les inclure dans la régression a expliqué une partie du bruit présent dans nos données afin que nous puissions resserrer notre estimation de l'ATE.

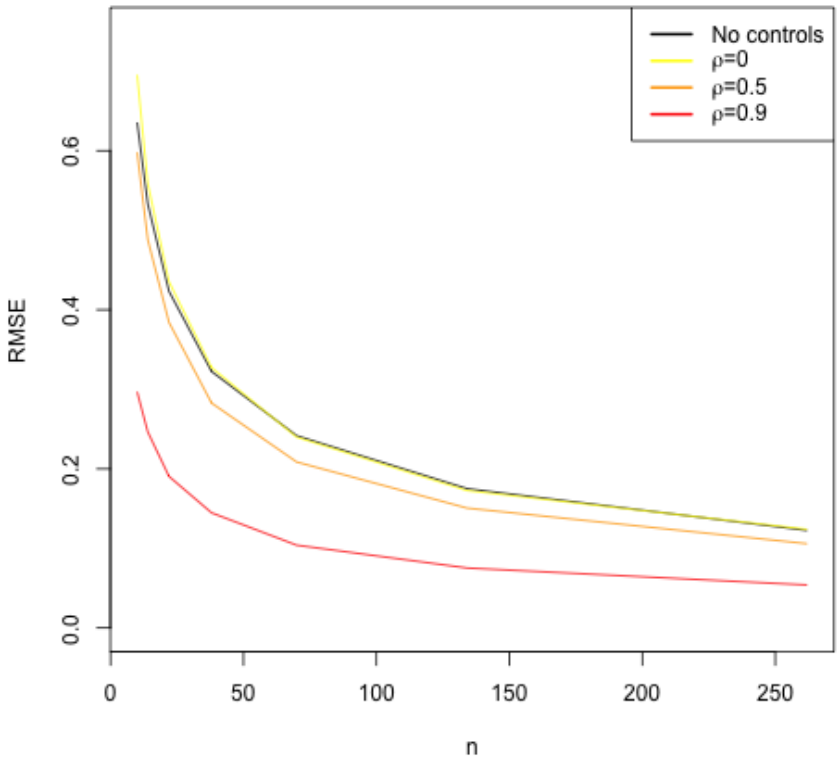
5 Quand cela sera-t-il utile ?

Quand l'ajustement pour les covariables est-il le plus susceptible d'améliorer la précision ?

L'ajustement des covariables sera plus utile lorsque vos covariables sont fortement prédictives (ou "pronostiques") de vos résultats. L'ajustement des covariables vous permet essentiellement d'utiliser des informations sur les relations entre les caractéristiques de base et votre résultat afin que vous puissiez mieux identifier la relation entre le traitement et le résultat. Mais si les caractéristiques de base ne sont que faiblement corrélées avec le résultat, l'ajustement des covariables ne vous sera d'aucune utilité. Les covariables que vous voudrez ajuster sont celles qui sont fortement corrélées aux résultats.

Le graphique suivant montre la relation entre le pronostic de votre covariable et les gains que vous obtenez en l'ajustant. En abscisse se trouve la taille de l'échantillon, et en ordonnée se trouve la racine de l'erreur quadratique moyenne (https://fr.wikipedia.org/wiki/Racine_de_l%27erreur_quadratique_moyenne) (root-mean-square deviation, RMSE), entre l'estimateur et le vrai ATE. Nous voulons que notre RMSE soit faible, et l'ajustement des covariables devrait nous aider à le réduire.

```
rm(list=ls())
library(MASS)  # pour mvrnorm()
set.seed(1234567)
num.reps = 10000

# Le véritable effet du traitement est de 0 pour chaque unité

adj.est = function(n, cov.matrix, treated) {
    Y.and.X  =  mvrnorm(n, mu = c(0, 0), Sigma = cov.matrix)
    Y   =  Y.and.X[, 1]
    X   =  Y.and.X[, 2]
    coef(lm(Y ~ treated + X))[2]
}

unadj.est = function(n, treated) {
    Y = rnorm(n)
    coef(lm(Y ~ treated))[2]
}

rmse = function(half.n, rho = 0, control = TRUE) {
    treated  =  rep(c(0, 1), half.n)
    n = 2 * half.n

    if (control) {
        cov.matrix  =  matrix(c(1, rho, rho, 1), nrow = 2, ncol = 2)
        return( sqrt(mean(replicate(num.reps, adj.est(n, cov.matrix, treated)) ^ 2)) )
    }
    else {
        return( sqrt(mean(replicate(num.reps, unadj.est(n, treated)) ^ 2)) )
    }
}

half.n = c(5, 7, 11, 19, 35, 67, 131)
n = 2 * half.n
E  = sapply(half.n, rmse, control = FALSE)
E0 = sapply(half.n, rmse, rho = 0)
E1 = sapply(half.n, rmse, rho = 0.5)
E2 = sapply(half.n, rmse, rho = 0.9)

plot(n, E, type = "l", ylab = "RMSE", xlim = c(min(n),max(n)), ylim = c(0,.75))
lines(n, E0, col = "yellow")
lines(n, E1, col = "orange")
lines(n, E2, col = "red")
legend(x = 'topright',
       c("Pas de contrôle",
         expression(paste(rho, "=0")), expression(paste(rho, "=0.5")),
         expression(paste(rho, "=0.9"))),
         col=c("black", "yellow","orange", "red"), lty = 1, lwd=2)
```



La ligne noire montre le RMSE lorsque nous n'ajustons pas pour une covariable. La ligne rouge montre le RMSE lorsque nous ajustons pour une covariable fortement prédictive (la corrélation entre la covariable et le résultat est de 0,9). Vous pouvez voir que la ligne rouge est toujours en dessous de la ligne noire, c'est-à-dire que le RMSE est plus faible lorsque vous ajustez pour une covariable prédictive. La ligne orange représente le RMSE lorsque nous ajustons pour une covariable de pronostic modéré (la corrélation entre la covariable et le résultat est de 0,5). Nous obtenons toujours des gains de précision par rapport à la ligne noire, mais pas autant qu'avec la ligne rouge. Enfin, la ligne jaune montre ce qui se passe si vous contrôlez une covariable qui n'est pas du tout prédictive du résultat. La ligne jaune est presque identique à la ligne noire. Vous n'avez obtenu aucune amélioration de la précision en contrôlant une covariable non prédictive ; en fait, vous avez payé une légère pénalité parce que vous avez perdu un degré de liberté, ce qui est particulièrement coûteux lorsque la taille de l'échantillon est petite. Cet exercice démontre que vous obtiendrez le plus de gains de précision en contrôlant les covariables qui prédisent fortement les résultats.

Comment savoir quelles covariables sont susceptibles d'être prédictives avant de lancer votre expérience ? Des expériences antérieures ou même des études d'observation peuvent offrir des indications sur les caractéristiques de base qui prédisent le mieux les résultats.

6 Contrôle des covariables prédictives, qu'elles présentent ou non des déséquilibres

Les covariables doivent généralement être choisies en fonction de leur capacité attendue à aider à prédire les résultats, qu'elles présentent ou non des "déséquilibres" (i.e., qu'il existe ou non des différences notables entre les groupes de traitement et de contrôle dans les valeurs moyennes ou d'autres aspects des distributions de covariables). Il y a deux raisons à cette recommandation :

1. L'inférence fréquentiste (erreur type, intervalles de confiance, p-valeurs, etc.) suppose que l'analyse suit une stratégie prédéfinie. Choisir des covariables sur la base des déséquilibres observés rend plus difficile l'obtention d'inférences qui reflètent votre stratégie réelle. Par exemple, supposons que vous choisissiez de ne pas contrôler le genre parce que les groupes de traitement et de contrôle ont une composition similaire

pour le genre, mais vous *auriez* contrôlé le genre s'il y avait eu un déséquilibre notable. Les méthodes typiques d'estimation de l'erreur type supposeront à tort que vous ne contrôlerez jamais le sexe, quel que soit le déséquilibre que vous constatez.

2. L'ajustement pour une covariable fortement prédictive a tendance à améliorer la précision, comme nous l'avons expliqué ci-dessus. Pour recevoir le crédit dû pour cette amélioration de la précision, vous devez ajuster la covariable même s'il n'y a pas de déséquilibre. Par exemple, supposons que le sexe soit fortement corrélé avec votre résultat, mais il arrive que les groupes de traitement et de contrôle aient exactement la même composition de genre. Dans ce cas, l'estimation non ajustée de l'ATE sera exactement la même que l'estimation ajustée à partir d'une régression du résultat sur le traitement et le sexe, mais leurs erreurs types seront différentes. L'erreur type de l'estimation non ajustée a tendance à être plus grande car elle suppose que même si les groupes de traitement et de contrôle avaient des compositions de genre très différentes, vous utiliseriez toujours la différence des moyennes traitement-contrôle non ajustée (qui serait probablement loin du vrai ATE dans ce cas). Si vous spécifiez à l'avance que vous ajusterez pour le sexe, quel que soit l'ampleur du déséquilibre que vous constatez, vous aurez tendance à obtenir une erreur type plus petite, un intervalle de confiance plus serré et un test de signification plus puissant.

En supposant que l'assignation aléatoire a été mise en œuvre correctement, l'examen des déséquilibres devrait-il jouer un rôle *quelconque* dans le choix des covariables à corriger ? Voici un échantillon de points de vue :

- Mutz, Pemantle et Pham (2016) soutiennent que, à moins qu'il n'y ait une attrition différentielle, la pratique consistant à sélectionner des covariables sur la base des déséquilibres observés est "non seulement inutile" mais "même pas utile ... et peut en fait être préjudiciable", car cela invalide les intervalles de confiance, détériore la précision (par rapport à l'ajustement prédéfini pour les covariables prédictives) et ouvre la porte à tout.[8]

- Permutt (1990), en utilisant la théorie et des simulations pour étudier des scénarios spécifiques, constate que lorsqu'un test d'équilibre est utilisé pour décider s'il faut ajuster pour une covariable, le test de signification pour l'effet du traitement est conservatif (c'est-à-dire qu'il a une probabilité réelle d'erreur de type I inférieure à son niveau nominal). Il écrit : "une plus grande puissance statistique peut être obtenue en ajustant toujours pour une covariable qui est fortement corrélée avec la réponse quelle que soit sa distribution entre les groupes". Cependant, il n'exclut pas complètement de considérer les déséquilibres observés : "choisir des covariables sur la base de la différence entre les moyennes des groupes de traitement et de contrôle n'est pas irrationnel. Après tout, certaines erreurs de type I peuvent être plus graves que d'autres. Signaler une différence significative dans les résultats qui peut être expliquée comme l'effet d'une covariable peut être une erreur plus embarrassante que de signaler une erreur qui disparaît lors de la répétition mais sans explication simple. Des considérations similaires peuvent s'appliquer aux erreurs de type II. Un résultat positif qui dépend de l'ajustement de covariable peut de toute façon être considéré comme moins convaincant qu'un test positif pour deux échantillons, de sorte que l'erreur de ne pas conclure aussi assurément peut être moins grave. Ces justifications, cependant, sont extérieures à la théorie formelle des tests d'hypothèses."[9]

- Altman (2005) écrit : "il semble de loin préférable de choisir les variables à ajuster sans tenir compte de l'ensemble de données réelles à disposition". Il recommande de contrôler les covariables fortement prédictives, ainsi que celles qui ont été utilisées dans le découpage par bloc. Cependant, il aborde également un dilemme : "dans la pratique, un déséquilibre peut survenir lorsque le besoin éventuel d'ajustement n'a pas été anticipé. Que doivent faire les chercheurs ? Ils pourraient choisir d'ignorer le déséquilibre ; comme indiqué, ce serait tout à fait approprié. La difficulté est alors une question de crédibilité. Les lecteurs de leur article (y compris les examinateurs et les éditeurs) peuvent se demander si le résultat observé a été influencé par la distribution inégale d'une ou plusieurs covariables de base. Il est toujours possible, et sans doute conseillé, d'effectuer une analyse ajustée, mais maintenant avec la reconnaissance explicite qu'il s'agit d'une analyse exploratoire plutôt que définitive, et que l'analyse non ajustée doit être considérée comme la principale. Évidemment, si les analyses simples et ajustées donnent sensiblement le même résultat, alors il n'y a aucune difficulté d'interprétation. Ce sera généralement le cas. Cependant, si les résultats des deux analyses diffèrent, alors il y a un vrai problème. L'existence d'un

tel écart doit jeter un doute sur la véracité du résultat global (non ajusté). La situation est similaire aux difficultés d'interprétation qui surviennent avec des comparaisons de sous-groupes non planifiées. Une suggestion dans de telles circonstances est d'essayer d'imiter ce qui aurait été fait si le problème avait été anticipé, à savoir d'ajuster non pas les variables qui sont observées comme étant déséquilibrées, mais toutes les variables qui auraient été identifiées à l'avance comme prédictives. Une source indépendante pourrait être utilisée pour identifier ces variables. Alternativement, les données de l'essai pourraient être utilisées pour déterminer quelles variables sont prédictives. Cette stratégie pourrait également être pré-spécifiée dans le protocole de l'étude. Étant donné que cette analyse serait effectuée sous réserve d'un déséquilibre observé, elle n'élimine pas le biais et ne peut donc pas être considérée comme pleinement satisfaisante."[10]

- Tukey (1991) note que les déséquilibres observés peuvent justifier un ajustement en tant que contrôle de robustesse : bien que "la plupart des statisticiens" accepteraient une analyse d'un essai clinique randomisé qui ne tient pas compte des covariables, "certains cliniciens, et certains statisticiens, semble-t-il, aimeraient être plus sceptiques (peut-être à titre d'analyse supplémentaire) en demandant une analyse qui tienne compte des déséquilibres observés pour ces covariables. Gagner plus de confiance dans les résultats d'une telle analyse est en effet approprié, car le degré de protection contre les conséquences d'une randomisation inadéquate ou la survenue (aléatoire) d'une randomisation inhabituelle est considérablement augmenté par l'ajustement. *Une plus grande sécurité, plutôt qu'une précision accrue ... sera souvent la raison principale de l'ajustement de la covariance dans un essai randomisé. ...* L'objectif principal de permettre [l'ajustement] pour les covariables dans un essai *randomisé* est défensif : pour indiquer clairement que l'analyse a rempli ses obligations scientifiques."[11]

- Certains statisticiens soutiennent que nos inférences devraient être conditionnelles à une mesure du déséquilibre des covariables — en d'autres termes, lors de l'évaluation du biais, de la variance et de l'erreur quadratique moyenne d'une estimation ponctuelle ou de la probabilité de couverture d'un intervalle de confiance, au lieu de considérer toutes les randomisations possibles, il peut être plus pertinent de ne considérer que les randomisations qui produiraient un déséquilibre des covariables similaire à celui que nous observons. De ce point de vue, les déséquilibres observés peuvent être pertinents pour le choix de l'estimateur.[12]

- Lin, Green et Coppock (2016) écrivent : "Les covariables doivent généralement être choisies sur la base de leur capacité attendue à aider à prédire les résultats, qu'elles semblent bien équilibrées ou déséquilibrées entre les bras de traitement. Mais il peut y avoir des occasions où la liste de covariables spécifiée dans le PAP [plan de pré-analyse] a omis une covariable potentiellement importante (en raison d'un oubli ou de la nécessité de garder la liste courte lorsque N est petit) avec un déséquilibre non négligeable. La protection contre les biais ex post (conditionnés au déséquilibre observé) est alors une préoccupation légitime. Cependant, ils recommandent :"si les déséquilibres observés sont autorisés à influencer le choix des covariables, les vérifications de l'équilibre et les décisions d'ajustement doivent être finalisées à l'aveugle, avant de voir les données de résultat. La *direction* du déséquilibre observé (par exemple, si le groupe de traitement ou le groupe de contrôle semble plus avantagé au départ) ne devrait pas être autorisé à influencer les décisions concernant l'ajustement". Enfin l'estimateur pré-spécifié devrait"toujours être déclaré et étiqueté comme tel, même si des estimations alternatives sont également déclarées". ."[13]

7 Quand ne pas le faire ?

C'est une mauvaise idée d'ajuster les covariables lorsque vous pensez que ces covariables ont pu être influencées par votre traitement. C'est l'une des raisons pour lesquelles de nombreuses covariables sont collectées à partir des enquêtes de base ; parfois, les covariables recueillies à partir d'enquêtes après l'intervention pourraient refléter l'effet du traitement plutôt que les caractéristiques sous-jacentes du sujet. L'ajustement pour les covariables qui sont affectées par le traitement — les covariables "post-traitement" — peut entraîner un biais.

Supposons, par exemple, que Giné et Mansuri aient collecté des données sur le nombre de rassemblements politiques auxquels une femme a assisté après avoir reçu le traitement. En estimant l'effet du traitement sur l'indépendance du choix politique, vous pourriez être tenté d'inclure cette variable comme covariable dans votre régression. Mais inclure cette variable, même si elle prédit fortement le résultat, peut fausser l'effet estimé du traitement.

Créons cette fausse variable, qui est corrélée (comme la mesure des résultats) avec les covariables de base et également avec le traitement. Ici, par construction, l'effet du traitement sur le nombre de rassemblements politiques suivis est de 2. Lorsque nous avons inclus la variable des rallyes comme covariable, l'effet moyen du traitement estimé sur l'indépendance du choix du candidat était en moyenne de 0,54 sur les 10 000 répétitions. Rappelons que le véritable effet du traitement sur ce résultat est 1. Il s'agit d'un biais important, tout cela parce que nous avons contrôlé avec une covariable post-traitement ![14] Ce biais résulte du fait que la covariable est corrélée au traitement.

```
# Créer une covariable post-traitement corrélée avec des covariables pré-traitement
rallies0 = round(.5*owns.id.card + has.formal.schooling + 1.5*TV.access + log(age))
rallies1 = rallies0 + 2
rallies.mat = rallies1 * Z.mat + rallies0 * (1-Z.mat)

# Estimer l'ATE avec un nouveau modèle qui inclut la covariable post-traitement

adjust.for.post = function(Y, Z, X) {
  coef(lm(Y ~ Z + X + owns.id.card + has.formal.schooling + age + TV.access))[2]
}

post.adjusted.estimates = rep(NA, Replications)

for (i in 1:Replications) {
  post.adjusted.estimates[i]  =  adjust.for.post(Y.mat[,i], Z.mat[,i], rallies.mat[,i])
}

# Biais estimé du nouvel estimateur
mean(post.adjusted.estimates) - true.treatment.effect

# Marge d'erreur (au niveau de confiance de 95 %) pour le biais estimé
1.96 * sd(post.adjusted.estimates) / sqrt(Replications)
```

Ce n'est pas parce que vous ne devez pas ajuster les covariables post-traitement que vous ne pouvez pas collecter de données sur les covariables post-traitement, mais vous devez faire preuve de prudence. Certaines mesures pourraient être recueillies après le traitement, mais il est peu probable qu'elles soient affectées par le traitement (par exemple, l'âge et le sexe). Soyez prudent avec les mesures qui peuvent être soumises à des effets d'évaluation : par exemple, les femmes traitées peuvent être plus conscientes des attentes de participation politique et peuvent déclarer rétrospectivement qu'elles étaient plus actives politiquement qu'elles ne l'étaient réellement plusieurs années auparavant.

8 Préoccupations concernant le biais lié aux petits échantillons

Dans les petits échantillons, l'ajustement par régression peut produire une estimation biaisée de l'effet moyen du traitement.[15] Certaines simulations ont suggéré que ce biais a tendance à être négligeable lorsque le nombre d'unités assignées de manière aléatoire est supérieur à vingt.[16] Si vous travaillez avec un petit échantillon, vous pouvez utiliser une méthode d'ajustement de covariable sans biais telle que la post-stratification (diviser l'échantillon en sous-groupes sur la base des valeurs d'une ou plusieurs covariables de base, calculer la

différence des moyennes traitement-contrôle pour les résultats de chaque sous-groupe, et prendre une moyenne pondérée de ces estimations d'effet du traitement spécifiques au sous-groupe, avec des poids proportionnels à la taille de l'échantillon).[17]

9 Comment rendre vos décisions d'ajustement de covariable transparentes ?

Dans un souci de transparence, si vous ajustez les covariables, pré-spécifiez vos modèles et déclarez à la fois les estimations non corrigées et corrigées des covariables.

Les simulations ci-dessus ont montré que les résultats peuvent changer légèrement ou pas si légèrement selon les covariables que vous choisissez d'inclure dans votre modèle. Nous avons mis en évidence quelques règles empiriques ici : n'incluez que les covariables de pré-traitement qui sont prédictives des résultats. Décider quelles covariables inclure, cependant, est souvent une entreprise subjective plutôt qu'objective, donc une autre règle de base est d'être totalement transparent sur vos décisions de covariables. Incluez toujours le modèle le plus simple — la régression simple des résultats du traitement sans contrôle des covariables — dans votre article ou annexe pour compléter les résultats de votre modèle, y compris les covariables.

Pour minimiser la crainte de vos lecteurs qui pensent que vous avez recherché la combinaison particulière de covariables qui produit des résultats favorables à votre hypothèse, pré-spécifiez vos modèles dans un plan de pré-analyse.[18] Cela vous donne l'opportunité d'expliquer, avant de voir les résultats, quelles covariables de pré-traitement seront prédictives du résultat. Vous pouvez même écrire ces régressions dans R en utilisant de fausses données, comme ici, de sorte que lorsque vos résultats du terrain arrivent, tout ce que vous avez à faire est d'exécuter votre code sur les données réelles. Ces efforts sont un moyen utile de vous lier les mains en tant que chercheur et d'améliorer votre crédibilité.

10 Les covariables peuvent vous aider à enquêter sur l'intégrité de l'assignation aléatoire

Parfois, il n'est pas clair si l'assignation aléatoire a réellement eu lieu (ou si elle s'est produite en utilisant la procédure envisagée par le chercheur). Par exemple, lorsque les chercheurs analysent des assignations aléatoires naturelles (par exemple, celles effectuées par un organisme gouvernemental), il est utile d'évaluer statistiquement si le degré de déséquilibre entre les groupes de traitement et de contrôle se situe dans la marge d'erreur attendue. Un test statistique consiste à régresser l'assignation du traitement sur toutes les covariables et à calculer la statistique F. La degré de signification de cette statistique peut être évalué en simulant un grand nombre d'assignations aléatoires et en calculant pour chacune la statistique F ; la distribution résultante peut être utilisée pour calculer la p-valeur de la statistique F observée. Par exemple, si 10 000 simulations sont effectuées et que 30 simulations seulement génèrent une statistique F supérieure à celle réellement obtenue à partir des données, la valeur p est de 0,003, ce qui suggère que le niveau de déséquilibre observé est très inhabituel. Dans de tels cas, on peut souhaiter étudier de plus près la procédure de randomisation.

Pour approfondir

Athey, Susan et Guido W. Imbens (2017). "The Econometrics of Randomized Experiments." Dans le *Handbook of Economic Field Experiments*, vol. 1 (E. Duflo and A. Banerjee, eds.). arXiv (http://arxiv.org/abs/1607.00698) DOI (http://dx.doi.org/10.1016/bs.hefe.2016.10.003)

Gerber, Alan S. et Donald P. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*, chapitre 4.

Hennessy, Jonathan, Tirthankar Dasgupta, Luke Miratrix, Cassandra Pattanayak et Pradipta Sarkar (2016). "A Conditional Randomization Test to Account for Covariate Imbalance in Randomized Experiments." *Journal of Causal Inference* 4: 61–80.

Judkins, David R. et Kristin E. Porter (2016). “Robustness of Ordinary Least Squares in Randomized Clinical Trials.” *Statistics in Medicine* 35: 1763–1773.

Lin, Winston (2012). “Regression Adjustment in Randomized Experiments: Is the Cure Really Worse than the Disease?” Blog sur l’impact du développement, partie I (https://web.archive.org/web/20151024055802/http://blogs.worldbank.org/impactevaluations/node/847) et partie II (https://web.archive.org/web/20151024022122/http://blogs.worldbank.org/impactevaluations/node/849).

Raudenbush, Stephen W. (1997). “Statistical Analysis and Optimal Design for Cluster Randomized Trials.” *Psychological Methods* 2: 173–185.

Wager, Stefan, Wenfei Du, Jonathan Taylor et Robert Tibshirani (2016). “High-Dimensional Regression Adjustments in Randomized Experiments.” *Proceedings of the National Academy of Sciences* 113: 12673–12678. arXiv (https://arxiv.org/abs/1607.06801) DOI (http://doi.org/10.1073/pnas.1614732113)

---

1. Auteur d’origine : Lindsay Dolan. Révisions : Don Green et Winston Lin, 1er novembre 2016. Le guide est un document vivant et susceptible d’être mis à jour par les membres de EGAP à tout moment ; les contributeurs répertoriés ne sont pas responsables des modifications ultérieures. Merci à Macartan Humphreys et Diana Mutz pour les discussions utiles.↩
2. Miratrix, Luke W., Jasjeet S. Sekhon et Bin Yu (2013). “Adjusting Treatment Effect Estimates by Post-Stratification in Randomized Experiments.” *Journal of the Royal Statistical Society, Series B* 75: 369–396.↩
3. Voir, e.g., pages 217–219 de Miriam Bruhn et David McKenzie (2009), “In Pursuit of Balance: Randomization in Practice in Development Field Experiments,” *American Economic Journal: Applied Economics* 1 (4): 200–232.↩
4. Un bref examen du biais et de la précision : imaginez répéter l’expérience plusieurs fois (sans changer l’échantillon expérimental et les conditions, mais en refaisant l’assignation aléatoire du traitement chaque fois). Un estimateur sans biais peut surestimer ou sous-estimer l’ATE sur une répétition donnée, mais son espérance (la moyenne sur toutes les répétitions possibles) sera égale à l’ATE réel. Nous préférons généralement des estimateurs sans biais ou approximativement sans biais, mais nous valorisons également la précision (qui est formellement définie comme l’inverse de la variance). Imaginez que vous lancez une fléchette sur un jeu de fléchettes. Si vous frappez le centre de la cible en moyenne mais que vos tirs sont souvent loin du but, vous disposez d’un estimateur non biaisé mais imprécis. Si vous frappez près du centre à chaque fois, votre estimateur est plus précis. Un chercheur peut choisir d’accepter un petit biais en échange d’une grande amélioration de la précision. Un critère possible pour évaluer les estimateurs est l’erreur quadratique moyenne (https://fr.wikipedia.org/wiki/Erreur_quadratique_moyenne), qui est égale à la variance plus le carré du biais. Voir, par exemple, Sharon Lohr (2010), *Sampling: Design and Analysis*, 2e éd., pp. 31–32.↩
5. la “variabilité d’échantillonnage” fait référence à la dispersion des estimations qui sera produite simplement en raison des différentes assignations aléatoires qui auraient pu être tirées. Lorsque le tirage au sort pour l’assignation aléatoire produit un groupe de traitement avec plus d’As et un groupe de contrôle avec plus de Bs, il est plus difficile de séparer les caractéristiques de base (A et B) de l’assignation de traitement en tant que prédicteur des résultats observés.↩
6. Giné, Xavier et Ghazala Mansuri (2012). “Together We Will: Experimental Evidence on Female Voting Behavior in Pakistan.” (http://siteresources.worldbank.org/DEC/Resources/gine_mansuri_voting_ReStat.pdf)↩
7. Le biais estimé est de 0,0003 avec une marge d’erreur (au niveau de confiance de 95 %) de 0,0018.↩

8. Diana C. Mutz, Robin Pemantle et Philip Pham (2016), “Model Choice in Experimental Design: Messy Analyses of Clean Data.” (https://www.math.upenn.edu/~pemantle/papers/Preprints/balance.pdf)↩
9. Thomas Permutt (1990), “Testing for Imbalance of Covariates in Controlled Experiments,” *Statistics in Medicine* 9: 1455–1462.↩
10. Douglas G. Altman (2005), “Covariate Imbalance, Adjustment for,” (http://doi.org/10.1002/0470011815.b2a01015) dans *Encyclopedia of Biostatistics*.↩
11. John W. Tukey (1991), “Use of Many Covariates in Clinical Trials,” *International Statistical Review* 59: 123–137. Italique dans l’original.↩
12. Voir, par exemple : D. R. Cox et N. Reid (2000), *The Theory of the Design of Experiments*, pp. 29–32 ; D. Holt et T. M. F. Smith (1979), “Post Stratification”, *Journal of the Royal Statistical Society, Series A (General)* 142: 33–46; Richard M. Royall (1976), “Current Advances in Sampling Theory: Implications for Human Observational Studies,” *American Journal of Epidemiology* 104: 463–474. Pour une introduction aux désaccords philosophiques sur l’inférence statistique, voir Bradley Efron (1978), “Controversies in the Foundations of Statistics,” (http://www.maa.org/programs/maa-awards/writing-awards/controversies-in-the-foundations-of-statistics) *American Mathematical Monthly* 85: 231–246.↩
13. Winston Lin, Donald P. Green et Alexander Coppock (2016), “Standard Operating Procedures for Don Green’s Lab at Columbia,” (https%20:%20//github.com/acoppock/Green-Lab-SOP) version 1.05, 7 juin. Italique dans l’original.↩
14. Le biais estimé est de − 0,459 avec une marge d’erreur (au niveau de confiance de 95 %) de 0,002.↩
15. David A. Freedman (2008), “On Regression Adjustments in Experiments with Several Treatments,” *Annals of Applied Statistics* 2: 176–196. Voir aussi les articles de blog de Winston Lin (partie I (https://web.archive.org/web/20151024055802/http://blogs.worldbank.org/impactevaluations/node/847) et partie II (https://web.archive.org/web/20151024022122/http://blogs.worldbank.org/impactevaluations/node/849)) à propos de sa réponse à Freedman.↩
16. Green, Donald P. et Aronow, Peter M., Analyzing Experimental Data Using Regression: When Is Bias a Practical Concern? (7 Mars 2011). Article en cours : http://ssrn.com/abstract=1466886 (http://ssrn.com/abstract=1466886)↩
17. Miratrix, Sekhon et Yu (2013), cités ci-dessus.↩
18. Pour plus de détails sur les plans de pré-analyse, voir, par exemple, Benjamin A. Olken (2015), “Promises and Perils of Pre-Analysis Plans,” (http://doi.org/10.1257/jep.29.3.61) *Journal of Economic Perspectives* 29 (3) : 61–80.↩