

Un R^2 de 0,03, et le constat le plus utile du rapport

Douze caractéristiques de ménage expliquent trois pour cent de la variation du PB des enfants. Cibler le dépistage sur les 20 % les moins bien prédits trouve 31 % des enfants malnutris, contre 20 % au hasard. Le modèle a échoué, et son échec est la réponse opérationnelle.

Pierre Robentz Cassion

Leçon 8 sur 8

Régression appliquée aux données de programme — Le plan et l'ajustement

Ce que couvre cette leçon

- Un modèle bâti pour répondre à une question de ciblage
- Ne le supprimez pas
- Des diagnostics qui changent une décision
- Ce que le R^2 est et n'est pas
- Rapportez-le en entier
- La suite

Un modèle bâti pour répondre à une question de ciblage — En Python (suite)

```
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf

survey = pd.read_csv("household-survey-2025.v1.csv")
d = survey.dropna(subset=["child_muac_mm", "child_age_months", "child_sex",
                          "household_size", "displacement_status",
                          "main_livelihood", "food_insecure",
                          "improved_water_source"])

model = smf.ols(
    "child_muac_mm ~ child_age_months + child_sex + household_size"
    " + displacement_status + main_livelihood + food_insecure"
    " + improved_water_source", data=d).fit()

print(f"n = {int(model.nobs)}    R2 = {model.rsquared:.4f}")
```

Un modèle bâti pour répondre à une question de ciblage — En Python (suite)

```
f"    adjusted R2 = {model.rsquared_adj:.4f}"  
print(f"residual SD {np.sqrt(model.scale):.2f}    outcome SD {d['child_muac_mm'].std():.2f}")
```

Un modèle bâti pour répondre à une question de ciblage — En R

```
model <- lm(child_muac_mm ~ child_age_months + child_sex + household_size +  
            displacement_status + main_livelihood + food_insecure +  
            improved_water_source, data = d)  
summary(model)
```

Un modèle bâti pour répondre à une question de ciblage

- $n = 641$. $R^2 = 0,030$, R^2 ajusté = 0,011. Écart-type résiduel 15,95 mm contre un écart-type du résultat de 16,04 mm

Un modèle bâti pour répondre à une question de ciblage

Terme	Coefficient	IC à 95 %
Insécurité alimentaire	−4,70 mm	−7,45 à −1,95
Moyen d'existence salarié	−3,49 mm	−8,44 à +1,45
Moyen d'existence agricole	−2,76 mm	−6,16 à +0,64
Âge de l'enfant, par mois	−0,03 mm	−0,11 à +0,05
Source d'eau améliorée	−1,09 mm	−3,76 à +1,58

Ne le supprimez pas — En Python

```
d = d.assign(predicted=model.fittedvalues, mam=(d["child_muac_mm"] < 125))
for frac in (0.20, 0.30, 0.50):
    cut = d["predicted"].quantile(frac)
    caught = d.loc[d["predicted"] <= cut, "mam"].sum()
    print(f"screen worst {frac:.0%}: catches {caught}/{d['mam'].sum()}")
    f" = {caught / d['mam'].sum():.1%} of cases")
```

Ne le supprimez pas — En R

```
d |> mutate(predicted = fitted(model), mam = child_muac_mm < 125) |>  
  arrange(predicted) |>  
  summarise(caught = mean(head(mam, n() * 0.2)) * sum(mam))
```

Ne le supprimez pas

Règle de dépistage	Enfants dépistés	PB < 125 mm trouvés
20 % les moins bien prédits	129	28 sur 90 = 31,1 %
30 % les moins bien prédits	193	41 sur 90 = 45,6 %
50 % les moins bien prédits	321	54 sur 90 = 60,0 %
20 % au hasard	129	20 % attendus

Ne le supprimez pas

- Le ciblage par le modèle trouve 31 % des enfants malnutris en en dépistant 20 %. Dépister 20 % au hasard en trouve 20 %...
- C'est cela, le constat — Non pas « le modèle n'a pas marché » — *les caractéristiques des ménages n'identifient pas...

Des diagnostics qui changent une décision

- Des valeurs ajustées hors de la plage possible — Pour un résultat binaire ajusté par régression linéaire, c'est immédiat

Des diagnostics qui changent une décision — En Python

```
lpm = smf.ols("completed ~ disability + case_category + age_band + sex"
              " + service_requested + admin1", data=referrals).fit()
print(f"fitted values from {lpm.fittedvalues.min():.3f}"
      f" to {lpm.fittedvalues.max():.3f}")
print(f"outside [0, 1]: {((lpm.fittedvalues < 0) | (lpm.fittedvalues > 1)).sum()}")
```

Des diagnostics qui changent une décision — En R

```
range(fitted(lm(completed ~ ., data = referrals)))
```

Des diagnostics qui changent une décision

- Huit des 1 581 probabilités ajustées tombent sous zéro, la plus basse à $-0,060$ — Le modèle logistique sur les mêmes. . .
- Les résidus contre les valeurs ajustées — Cherchez une forme en éventail, qui signifie que la dispersion dépend du. . .

Des diagnostics qui changent une décision — En Python

```
import matplotlib.pyplot as plt  
plt.scatter(model.fittedvalues, model.resid, s=6)
```

Des diagnostics qui changent une décision — En R

```
plot(model, which = 1)
```

Des diagnostics qui changent une décision

- La décision : l'hétéroscédasticité ne biaise pas le coefficient, elle biaise l'erreur-type — le remède est donc une...
- L'influence — Une poignée de lignes peut porter un coefficient

Des diagnostics qui changent une décision — En Python

```
influence = model.get_influence().cooks_distance[0]
print(f"rows with Cook's D > 4/n: {(influence > 4 / len(d)).sum()}")
```

Des diagnostics qui changent une décision — En R

```
sum(cooks.distance(model) > 4 / nobs(model))
```

Des diagnostics qui changent une décision

- La décision : regardez les lignes signalées avant de faire quoi que ce soit — Les onze ménages du cours EAH avec une. . .
- La linéarité d'un prédicteur continu — Ajustez-le par classes et voyez si les coefficients progressent régulièrement

Des diagnostics qui changent une décision — En Python

```
d["age_band"] = pd.cut(d["child_age_months"], [0, 12, 24, 36, 48, 60])  
print(smf.ols("child_muac_mm ~ C(age_band)", data=d).fit().params.round(2))
```

Des diagnostics qui changent une décision — En R

```
lm(child_muac_mm ~ cut(child_age_months, c(0, 12, 24, 36, 48, 60)), data = d)
```

Des diagnostics qui changent une décision

- La décision : si les classes ne progressent pas régulièrement, le terme linéaire répond à la mauvaise question — et le...

Ce que le R^2 est et n'est pas

- Le R^2 est la part de variance dont le modèle rend compte — Dans des données de programme il est couramment petit, et...
- Ce n'est pas une mesure de la justesse d'un coefficient — Un essai randomisé avec un effet de traitement tranché peut...
- C'est une mesure de la valeur des prédictions — C'est l'usage ci-dessus, et c'est l'usage honnête : la question du...
- Le R^2 ajusté pénalise les termes ajoutés, et il est passé de 0,030 à 0,011 ici — l'écart entre les deux mesure combien...

Rapportez-le en entier — Exemple (suite)

Predicting child MUAC from household characteristics

Linear model, 641 children with complete data.

R-squared 0.030, adjusted 0.011. Residual SD 15.95 mm against an outcome SD of 16.04 mm.

Only food insecurity is distinguishable from zero: -4.70 mm (95% CI -7.45 to -1.95). Household size, displacement status, water source, child sex and child age are not.

Used as a targeting rule, screening the 20% of children the model ranks worst would find 31% of the children with MUAC below 125 mm, against 20% expected from screening 20% at random. Screening half the children would find 60%.

Recommendation: do not target screening on household characteristics. The

Rapportez-le en entier — Exemple (suite)

survey's 14.0% prevalence of MUAC below 125 mm is not concentrated in households that a registration form can identify, and blanket screening remains the only rule that reaches the caseload.

This model is reported because it does not fit. A negative result about targeting is a budget decision, and dropping the model would have left the proposal to assume targeting works.

- Le dernier paragraphe est la raison d'être de cette section — Un modèle qui s'ajuste mal est une preuve, et le tiroir. . .

- C'est le cours.

Lire la leçon complète, avec le code exécutable

[https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/
reg-08-the-model-that-does-not-fit](https://data-analysis.cassion.dev/fr/cours/regression-for-programmes/reg-08-the-model-that-does-not-fit)