R^2 of 0.03, and the most useful result in the report

Twelve household characteristics explain three per cent of the variation in child MUAC. Targeting screening on the model's worst-predicted 20% finds 31% of the malnourished children, against 20% at random. The model failed, and its failure is the operational answer.

Pierre Robentz Cassion

Lesson 8 of 8

Regression for Programme Data — The design and the fit

What this lesson covers

- A model built to answer a targeting question
- Do not delete it
- Diagnostics that change a decision
- What R^2 is and is not
- Report it whole
- What comes next

A model built to answer a targeting question — In Python (cont.)

```
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}"
```

A model built to answer a targeting question — In Python (cont.)

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

A model built to answer a targeting question — In R

```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)
```

A model built to answer a targeting question

- $n = 641$. $R^2 = 0.030$, adjusted $R^2 = 0.011$. Residual SD 15.95 mm against an outcome SD of 16.04 mm

A model built to answer a targeting question

Term	Coefficient	95% CI
Food insecure	**−4.70 mm**	−7.45 to −1.95
Salaried livelihood	−3.49 mm	−8.44 to +1.45
Farming livelihood	−2.76 mm	−6.16 to +0.64
Child age, per month	−0.03 mm	−0.11 to +0.05
Improved water source	−1.09 mm	−3.76 to +1.58

Do not delete it — In Python

```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")
```

Do not delete it — In R

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

Do not delete it

Screening rule	Children screened	MUAC < 125 mm found
Worst-predicted 20%	129	28 of 90 = **31.1%**
Worst-predicted 30%	193	41 of 90 = 45.6%
Worst-predicted 50%	321	54 of 90 = 60.0%
Random 20%	129	**20%** expected

Do not delete it

- Targeting on the model finds 31% of the malnourished children while screening 20% of them. Screening 20% at random...
- That is the finding — Not "the model did not work" — *household characteristics do not identify malnourished children...

Diagnostics that change a decision

- Fitted values outside the possible range — For a binary outcome fitted with linear regression, this is immediate

Diagnostics that change a decision — In 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()}")
```

Diagnostics that change a decision — In R

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

Diagnostics that change a decision

- Eight of 1,581 fitted probabilities fall below zero, the lowest at -0.060 — The logistic model on the same data stays...
- Residuals against fitted values — Look for a fan shape, which means the spread depends on the level

Diagnostics that change a decision — In Python

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

Diagnostics that change a decision — In R

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

Diagnostics that change a decision

- The decision: heteroscedasticity does not bias the coefficient, it biases the standard error — so the fix is a robust...
- Influence — A handful of rows can carry a coefficient

Diagnostics that change a decision — In Python

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

Diagnostics that change a decision — In R

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

Diagnostics that change a decision

- The decision: look at the flagged rows before doing anything — The WASH course's eleven households with a unit error. . .
- Linearity of a continuous predictor — Fit it in bands and see whether the coefficients step evenly

Diagnostics that change a decision — In 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))
```

Diagnostics that change a decision — In R

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

Diagnostics that change a decision

- The decision: if the bands do not step evenly, the linear term is answering the wrong question — and banding is usually. . .

What R^2 is and is not

- R^2 is the share of variance the model accounts for — In programme data it is routinely small and that is a fact about...
- It is not a measure of whether a coefficient is right — A randomised trial with a decisive treatment effect can have an...
- It is a measure of whether predictions are worth making — That is the use above, and it is the honest one: the...
- Adjusted R^2 penalises added terms, and it went from 0.030 to 0.011 here — the gap between the two is a measure of how...

Report it whole — Example (cont.)

```
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
```

Report it whole — Example (cont.)

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.

Report it whole

- The last paragraph is why the section exists — A model that does not fit is evidence, and the file drawer it usually. . .

What comes next

- That is the course.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/regression-for-programmes/

reg-08-the-model-that-does-not-fit