Two coefficients cross the line and one crosses back

The rural-remote stratum holds 11% of households and a third of the interviews. Fit the model unweighted and livestock livelihoods look like a significant risk factor; weight it and they stop being one, while household size and returnee status start.

Pierre Robentz Cassion

Lesson 7 of 8

What this lesson covers

- The design the sample was drawn under
- What the weights do to an estimate
- The same model, weighted and not
- Weights, clusters and strata are three separate things
- When to leave the weights out
- Report it whole
- What comes next

The design the sample was drawn under — In Python

```
import pandas as pd

survey = pd.read_csv("household-survey-2025.v1.csv")
frame = pd.read_csv("household-survey-frame-2025.v1.csv")

print(frame.groupby("stratum")["households"].sum())
print(survey.groupby("stratum").size())
```

The design the sample was drawn under — In R

```
library(dplyr)
frame |> summarise(households = sum(households), .by = stratum)
survey |> count(stratum)
```

The design the sample was drawn under

Stratum	Households in frame	Share	Interviews	Share
Rural accessible	28,958	51.3%	337	33.8%
Urban	21,270	37.7%	316	31.7%
Rural remote	**6,200**	**11.0%**	**343**	**34.4%**

The design the sample was drawn under

- Rural remote is 11.0% of the population and 34.4% of the sample — That is a deliberate design choice — the survey...

What the weights do to an estimate — In Python

```
stratum_households = frame.groupby("stratum")["households"].sum()
base_weight = stratum_households / (25 * 14)
interviewed = frame[frame["selected"]].set_index("ea_id")["households_interviewed"]

survey["weight"] = (survey["stratum"].map(base_weight)
                    * 14 / survey["ea_id"].map(interviewed))
assert abs(survey["weight"].sum() - frame["households"].sum()) < 1

unweighted = survey["food_insecure"].mean()
weighted = np.average(survey["food_insecure"], weights=survey["weight"])
print(f"unweighted {unweighted:.1%}   weighted {weighted:.1%}")
```

What the weights do to an estimate — In R

```
survey <- survey |>
  mutate(weight = base_weight[stratum] * 14 / households_interviewed)

c(unweighted = mean(survey$food_insecure),
  weighted   = weighted.mean(survey$food_insecure, survey$weight))
```

What the weights do to an estimate

- 32.8% unweighted, 29.1% weighted — Rural remote has the highest food insecurity (46.4%), and over-representing it by...
- A regression fitted on this file without weights inherits exactly that problem — and it is not fixed by adding...

The same model, weighted and not — In Python

```
import statsmodels.api as sm
import statsmodels.formula.api as smf

FORMULA = ("food_insecure ~ improved_water_source + displacement_status"
           " + main_livelihood + household_size")

plain = smf.glm(FORMULA, data=d, family=sm.families.Binomial()).fit(
    cov_type="cluster", cov_kwds={"groups": d["ea_id"]})

weighted = smf.glm(FORMULA, data=d, family=sm.families.Binomial(),
                   freq_weights=d["weight"]).fit(
    cov_type="cluster", cov_kwds={"groups": d["ea_id"]})
```

The same model, weighted and not — In R

```
library(survey)

design <- svydesign(ids = ~ea_id, strata = ~stratum, weights = ~weight,
                    data = survey, nest = TRUE)
svyglm(food_insecure ~ improved_water_source + displacement_status +
         main_livelihood + household_size, design = design,
       family = quasibinomial())
```

The same model, weighted and not

Term	Unweighted OR	Weighted OR	Moves?
Improved water source	0.75 (0.55–1.03)	0.74 (0.51–1.08)	—
Displacement: returnee	0.56 (0.25–1.26)	**0.39 (0.15–0.97)**	**becomes significant**
Livelihood: farming	1.55 (1.03–2.34)	1.69 (1.04–2.75)	—
Livelihood: livestock	**1.51 (1.04–2.18)**	1.26 (0.73–2.18)	**stops being significant**
Livelihood: salaried	0.41 (0.22–0.77)	0.45 (0.24–0.87)	—
Household size	1.04 (0.98–1.11)	**1.08 (1.01–1.16)**	**becomes significant**

The same model, weighted and not

- Three coefficients change which side of the line they sit on, in both directions — Weighting is not a correction that...
- Livestock livelihoods are concentrated in the over-sampled remote stratum — Unweighted, they carry a third of the...

Weights, clusters and strata are three separate things

Design feature	Fixes	If you omit it
Weights	The estimate	Estimates describe the sample, not the population
Clusters	The standard error	Standard errors too small, as in the last lesson
Strata	The standard error	Standard errors slightly too large — the conservative error

Weights, clusters and strata are three separate things

- Weights change the coefficient; clusters and strata change its interval — That is the same distinction as...

Weights, clusters and strata are three separate things — In Python

```
print(f"clusters: {d['ea_id'].nunique()} enumeration areas")
print(f"strata:   {d['stratum'].nunique()}")
print(f"weights:  {d['weight'].min():.1f} to {d['weight'].max():.1f}")
```

Weights, clusters and strata are three separate things — In R

```
# survey::svydesign wants all three; a glm with weights= handles only one.
```

Weights, clusters and strata are three separate things

- `glm(..., weights=)` is not survey-weighted regression — It applies the weights and computes standard errors as though...

When to leave the weights out

- A within-stratum question — "Among rural remote households, does an improved water source go with lower food...
- A model whose covariates include everything the weights are built from — If stratum and selection probability are fully...

When to leave the weights out — In Python

```
print(f"unweighted {plain.params['main_livelihood[T.livestock]']:.3f}")
print(f"weighted   {weighted.params['main_livelihood[T.livestock]']:.3f}")
```

When to leave the weights out — In R

```
# A large gap between the two is evidence the design is not in the model.
```

When to leave the weights out

- A large difference between the weighted and unweighted fit is itself diagnostic — and it is the check to run before...

Report it whole — Example (cont.)

```
Food insecurity, household survey 2025

  976 households with complete covariates (of 996 interviewed),
  75 enumeration areas, 3 strata.
  Survey-weighted logistic regression; weights sum to 56,428 households,
  which is the frame total. Standard errors account for clustering by
  enumeration area and for stratification.

  Weighted prevalence      29.1%   (unweighted 32.8%)

  Adjusted odds ratios (weighted):
    Improved water source          0.74   95% CI 0.51 to 1.08
    Returnee household             0.39   95% CI 0.15 to 0.97
    Farming livelihood             1.69   95% CI 1.04 to 2.75
    Livestock livelihood           1.26   95% CI 0.73 to 2.18
    Salaried livelihood            0.45   95% CI 0.24 to 0.87
```

Report it whole — Example (cont.)

```
  Household size, per person      1.08   95% CI 1.01 to 1.16

The unweighted model makes livestock livelihoods significant (OR 1.51,
95% CI 1.04 to 2.18) and household size not. The rural remote stratum is
11% of households and 34% of interviews, and livestock is concentrated
there; the unweighted result is an artefact of the sampling design.

Odds ratios are reported because food insecurity is 29% and the ratios are
moderate; the weighted risk difference for salaried livelihoods is -14.2
points against the observed mix of livelihoods, which is the number in
the summary.
```

Report it whole

- The paragraph naming the artefact is what a reviewer will check first — and writing it yourself is better than having. . .

What comes next

- Every model so far has been fitted and read.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/regression-for-programmes/reg-07-weighting-the-model