Department
for Environment
Food & Rural Affairs

Data Science Accelerator

Ratio Imputation of the June Survey of Agriculture & Horticulture

Francesca Parrott, Defra

Background

About me:

- Statistician at Sky, ICNARC, UCL and Defra
- Have previously worked with web statistics, medical statistics using Stata
- Now agricultural statistics and learning R

Data Science Accelerator:

- Run by Government Digital Service, open to anyone across the public sector with a project that requires data science capabilities
- One day per week for 12 weeks at a local hub with a dedicated mentor
- New cohort every quarter

Background

June Survey of Agriculture & Horticulture - Farm Surveys Team - Defra

- Has run annually since 1866
- Legal obligation to complete it under Agricultural Statistics Act 1979
- Collects data on land, crops, livestock and labour ~100 variables




- Sample size of ~30,000 holdings (~15,000 responses)
- Population size of ~105,000 holdings

- 90,000 holdings with no response data
- Need data for all holdings in England ➔ imputation

Ratio Imputation (e.g. hectares of wheat)

Base data (2016 Final dataset)	Response data (2017 Survey dataset)	Imputed data (2017 Final dataset)
0	2	2
5	4	4
10	15	15
15	17	17
20	21	21
0		0
5		5.9
10		11.8
15		17.7
20		23.6

Ratio = 59 / 50 = 1.18

Multiply by Ratio = 1.18

- Holdings are split into six strata based on farm type/size
- Separate ratios are calculated for each strata, for each variable
- New holdings are dealt with separately as they don't have any base data

Current process

For each of the ~100 variables:

1. Run Genstat code to calculate the ratios
2. Copy & paste the results into Word docs called proving sheets
3. Manually inspect results for outliers or small strata
4. Edit Genstat code to remove outliers or combine strata

5. Re-do steps 1-4 until no more changes are needed
6. Manually log the removed outliers and the combined strata into an Excel spreadsheet

7. Run Genstat code to apply the finalised ratios and output the final imputed dataset

Oats A4 Original

```
Y-variate (response data):  a4
X-variate (base data):      xa4
Correlation:                0.750
Ratio method:               separate
Variance method:            Conventional (Taylor series)
Deff:                       0.4192 (wrt design based srs)
Deff ratio analysis:        (Not calculated due to missing X)
```

post_strat	Numbers of observations: total	imputed	Sampling: sample	excluded	fraction	Matched data: y>0	x>0
1	49871	46364	3507	0	0.070	65	59
2	15524	13871	1653	0	0.106	93	85
3	14999	12346	2653	0	0.177	253	246
4	7962	5916	2046	0	0.257	247	234
5	7759	5057	2702	0	0.348	305	282
6	7829	4705	3124	0	0.399	368	338
99	2989	2062	927	0	0.310	6	0
Total	106933	90321	16612	0	0.155	1337	1244

Estimated totals

post_strat	Matched sample: sum y	sum x	ratio	All data: sum x	Raising factor: ratio	expans'n	Estimated totals: imputed	all	s.e.	%r.s.e.
1	739	594	1.243	7878	13.259	14.220	9057	9795	2057	21.0
2	1424	1170	1.217	10321	8.824	9.391	11139	12563	1486	11.8
3	4860	4344	1.119	22012	5.067	5.654	19766	24626	1638	6.7
4	5302	4375	1.212	16458	3.761	3.891	14642	19944	1074	5.4
5	8103	6743	1.202	19558	2.901	2.872	15401	23504	1098	4.7
6	13550	11576	1.171	25597	2.211	2.506	16412	29962	1352	4.5
99	51	*	*	*	3.224	3.224	114	166	66	39.7
Total	34030	28802	1.181	101825	3.543	6.437	86532	120561	3649	3.0

95% confidence limits for total are 113409 to 127713

Estimates in strata with ratio=* are based on simple raising
The ratio shown in the total row is the combined ratio estimator

10 points with highest influence

Holding numbers with % influence

Percentage influence is calculated as the percentage change
in the grand total when each sampled observation is omitted.

Changes made to A

Holding Number

- Removed CPI om strata 6 as a higher influential figure and outstanding point on graph, this increased the total from 120,561 to 120,904, it also improved strata 6 %rse from 4.5 to 4.2

Oats A4 Final

```
Data summary
------------

Y-variate (response data):  a4
X-variate (base data):      xa4
Correlation:                0.765
Ratio method:               separate
Variance method:            Conventional (Taylor series)
Deff:                       0.4117 (wrt design based srs)
Deff ratio analysis:        (Not calculated due to missing X)
```

post_strat	Numbers of observations: total	imputed	Sampling: sample	excluded	fraction	Matched data: y>0	x>0
1	49871	46364	3507	0	0.070	65	59
2	15524	13871	1653	0	0.106	93	85
3	14999	12346	2653	0	0.177	253	246
4	7962	5916	2046	0	0.257	247	234
5	7759	5057	2702	0	0.348	305	282
6	7829	4704	3124	1	0.399	368	337
99	2989	2062	927	0	0.310	6	0
Total	106933	90320	16612	1	0.155	1337	1243

Estimated totals

post_strat	Matched sample: sum y	sum x	ratio	All data: sum x	Raising factor: ratio	expans'n	Estimated totals: imputed	all	s.e.	%r.s.e.
1	739	594	1.243	7878	13.259	14.220	9057	9795	2057	21.0
2	1424	1170	1.217	10321	8.824	9.391	11139	12563	1486	11.8
3	4860	4344	1.119	22012	5.067	5.654	19766	24626	1638	6.7
4	5302	4375	1.212	16458	3.761	3.891	14642	19944	1074	5.4
5	8103	6743	1.202	19558	2.901	2.872	15401	23504	1098	4.7
6	13550	11339	1.195	25360	2.237	2.506	16755	30305	1260	4.2
99	51	*	*	*	3.224	3.224	114	166	66	39.7
Total	34030	28565	1.188	101588	3.553	6.437	86875	120904	3615	3.0

95% confidence limits for total are 113817 to 127991

Estimates in strata with ratio=* are based on simple raising
The ratio shown in the total row is the combined ratio estimator
Totals and means include restricted (excluded) data

10 points with highest influence

Holding numbers with % influence

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	question	desc	Post-ImpJur	Post-ImpNH	published	targetimp	se	basename	section	sumvar	no01	heading	pub01	special	nwith07	combimp	comb01	stratimp	stratfac	Outback[1]	Outback[2]	Outback[3]	Outback[4]	Outback[5]
2	a1	wheat		4,491.0	1,652,143.0	1,647,652	1.0	xa1	a	0	0	S04a1		0	30798.5	0	0	post_strat	post_strat					
3	a2	winter barley		1,037.0	360,873.0	359,836	1.0	xa2	a	0	0	S04a2		0	14902.4	0	0	post_strat	post_strat					
4	a3	spring barley		1,996.0	481,587.0	479591	1.0	xa3	a	0	0	S04a3		0	13498.3	0	0	post_strat	post_strat					
5	a4	oats		166.0	120,904.0	120,738	1.0	xa4	a	0	0	S04a4		0	6275	0	0	post_strat	post_strat					
6	r5	mixed corn		3.7	4,895.0	4,891	1.0	xr5	r	0	0	S04r5		0	240.84	0	0	post_strat1234	post_strat					
7	a6	rye		0.0	29,887.0	29,887	1.0	xa6	a	0	0	S04a6		0	261.7	0	0	post_strat123	post_strat					
8	a7	triticale		18.0	9,518.0	9,500	1.0	xa7	a	0	0	S04a7		0	1239.01	0	0	post_strat123	post_strat					
9	r10	early potatoes		0.0	10,053.0	10,053	1.0	xr10	r	0	0	S04r10		0	1961.24	0	0	post_strat123	post_strat					
10	a11	maincrop potatoes		64.0	98,151.0	98,087	1.0	xa11	a	0	0	S04a11		0	5384.84	0	0	post_strat123	post_strat					
11	a12	sugar beet		79.0	111,269.0	111,190	1.0	xa12	a	0	0	S04a12		0	5177.6	0	0	post_strat123	post_strat					
12	a14	leguminous forage crops		21.6	16,088.0	16,066	1.0	xa14	a	0	0	S04a14		0	10	0	0	post_strat123	post_strat					
13	a18	other crops for stockfeeding		26.0	6,851.0	6,825	1.0	xa18	a	0	0	S04a18		0	2263.86	0	0	post_strat1234	post_strat					
14	a19	root crops, brassicas and fodder beet		72.0	23,533.0	23,461	1.0	xa19	a	0	0	S04a19		0	3184.67	0	0	post_strat	post_strat					
15	a20	Borage		0.0	837.6	838	2.0	xa20	a	0	0	S04a20		0	10	0	0	post_strat0	post_strat					
16	a21	field beans		382.0	188,718.0	188,336	1.0	xa21	a	0	0	S04a21		0	5627.12	0	0	post_strat	post_strat					
17	a22	peas for harvesting dry		0.0	39,150.0	39,150	1.0	xa22	a	0	0	S04a22		0	1962.7	0	0	post_strat123	post_strat					
18	r231	maize - grain		0.0	7,759.0	7,759	2.0	xr231	r	0	0	S04r231		0	7459.84	0	0	post_strat	post_strat					
19	a232	maize - fodder		320.0	118,141.0	117,821	1.0	xa232	a	0	0	S04a232		0	7459.84	0	0	post_strat	post_strat					
20	r233	maize - AD		148.0	57,382.0	57,234	2.0	xr233	r	0	0	S04r233		0	7459.84	0	0	post_strat	post_strat					
21	a24	winter osr		2,135.0	514,794.0	512,659	1.0	xa24	a	0	0	S04a24		0	14115	0	0	post_strat	post_strat					
22	a25	spring osr		0.0	8,330.0	8,330	1.0	xa25	a	0	0	S04a25		0	690.38	0	0	post_strat123	post_strat					
23	a27	linseed		195.0	26,374.0	26,179	1.0	xa27	a	0	0	S04a27		0	592.78	0	0	post_strat12	post_strat					
24	a31	all other crops		17.0	14,272.0	14,255	1.0	xa31	a	0	0	S04a31		0	2989.8	0	0	post_strat12	post_strat					
25	a32	bare fallow/ land withdrawn from agric. production		475.0	200,049.0	199,574	1.0	xa32	a	0	0	S04a32		0	13687.9	0	0	post_strat	post_strat					
26	a33	short rotation coppice		0.0	2,966.0	2,966	1.0	xa33	a	0	0	S04a33		0	10	0	0	post_strat0	post_strat					
27	r34	miscanthus		0.0	7,366.0	7,366	1.0	xr34	r	0	0	S04r34		0	10	0	0	post_strat0	post_strat					
28	a35	crops for aromatic or medicinal use		0.0	2,178.6	2,179	1.0	xa35	a	0	0	S04a35		0	10	0	0	post_strat1234	post_strat					
29	b5	vining peas for processing		23.0	26,769.0	26,746	1.0	xb5	b	0	0	S04b5		0	905.46	0	0	post_strat123	post_strat					
30	b14	other peas and beans		0.2	2,430.4	2,430	1.0	xb14	b	0	0	S04b14		0	659.4	1	0	post_strat	post_strat					
31	b15	Culinary plants		1.7	2,807.0	2,805	1.0	xb15	b	0	0	S04b15		0	10	0	0	post_strat0	post_strat					
32	b21all	all other veg (inc b6 and b7)		72.0	63,797.0	63,725	1.0	xb21all	b	0	0	S04b21all		0	4380.56	0	0	post_strat1234	post_strat					
33	t3	orchards (combined in 2011)		9.5	22,237.0	22,228	1.0	xt3	t	0	0	S04t3		0	10	0	0	post_strat	post_strat					
34	c5	strawberries		0.2	3,061.0	3,061	1.0	xc5	c	0	0	S04c5		0	820	0	0	post_strat123	post_strat					
35	t6	raspberries		2.2	1,450.6	1,448	1.0	xt6	t	0	0	S04t6		0	836.43	0	0	post_strat123	post_strat					
36	t7	blackcurrants		1.4	2,072.0	2,071	1.0	xt7	t	0	0	S04t7		0	530.6	0	0	post_strat0	post_strat					
37	t10	wine grapes		3.2	1,993.8	1,991	1.0	xt10	t	0	0	S04t10		0	243.8	0	0	post_strat123	post_strat					
38	t11	other small fruit (inc. gooseberries & blackberries)		0.1	1,246.5	1,246	1.0	xt11	t	0	0	S04t11		0	613.73	0	0	post_strat1234	post_strat					
39	u6	christmas trees		1.4	1,979.7	1,978	1.0	xu6	u	0	0	S04u6		0	1105.79	0	0	post_strat12	post_strat					
40	u8	perennial herbaceous plants		0.0	327.0	327	1.0	xu8	u	0	0	S04u8		0	632.62	0	0	post_strat12	post_strat					
41	d10	other hardy nursery stock		0.0	2,319.8	2,320	1.0	xd10	d	0	0	S04d10		0	1037.35	0	0	post_strat123	post_strat					
42	d13	bulbs and flowers grown in the open		21.0	6,355.0	6,334	1.0	xd13	d	0	0	S04d13		0	760.24	0	0	post_strat1234	post_strat					
43	f1	glasshouse area used for veg & fruit		26.0	6,972,375.0	6,972,349	1.0	xf1	f	0	0	S04f1		0	1415.25	0	0	post_strat123	post_strat					
44	f2	glasshouse area used for flowers & foliage		9,389.0	4,782,012.0	4,772,623	1.0	xf2	f	0	0	S04f2		0	2075.18	0	0	post_strat12	post_strat					
45	f7	glasshouse not in use on census day		19.0	1,246,921.0	1,246,902	1.0	xf7	f	0	0	S04f7		0	1493.23	0	0	post_strat0	post_strat					
46	f11	mushroom sheds		0.0	36,467.0	36,467	1.0	xf11	f	0	0	S04f11		0	10	0	0	post_strat0	post_strat					
47	g1	grass under 5 yrs old		4,915.0	640,099.0	635,184	1.0	xg1	g	0	0	S04g1		0	36276.2	0	0	post_strat	post_strat					
48	g2	grass over 5 yrs old		44,017.0	3,278,641.0	3,234,624	1.0	xg2	g	0	0	S04g2		0	136498	0	0	post_strat	post_strat					
49	s5	rough grazing		6,613.0	478,773.0	472,160	1.0	xs5	s	0	0	S04s5		0	13080.6	0	0	post_strat	post_strat					

Holding numbers

Task

For each of the ~100 variables:

R Markdown

1. Run Genstat code to calculate the ratios
2. Copy & paste the results into Word docs called proving sheets
3. Manually inspect results for outliers or small strata
4. Edit Genstat code to remove outliers or combine strata

5. Re-do steps 1-4 until no more changes are needed
6. Manually log the removed outliers and the combined strata into an Excel spreadsheet

7. Run Genstat code to apply the finalised ratios and output the final imputed dataset

Code

1_SetUp.R

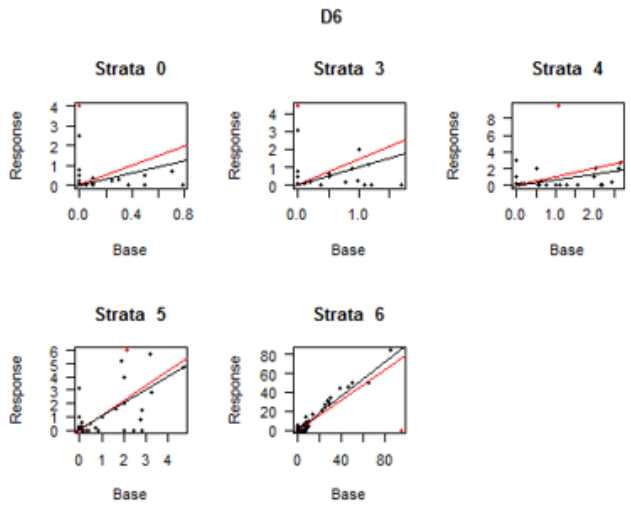
- Loads all the datasets
- Runs basic checks on the data
- Re-organises the datasets ready for analysis

2_Run.R

- Calls the following two pieces of code
  - 2a_Functions.R
  - 2b_ProvingSheet.Rmd
- Uses these to run the imputation one variable at a time in a loop and creates a proving sheet for each variable, a summary excel spreadsheet and the final imputed dataset

D_hns

D6



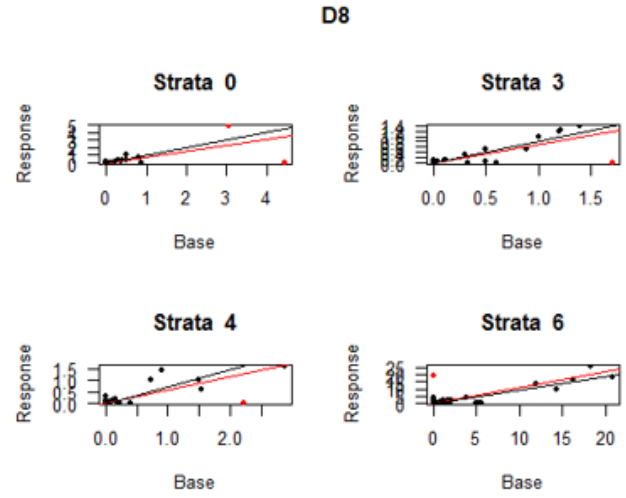
Before outlier removal

strata	ratio_n	base	resp	ratio	sd	est_base	est
0	20	4.55	11.11	2.44	1.4	15.71	38.33
3	18	10.5	14.79	1.41	1.66	59.01	83.2
4	25	25.45	25.85	1.02	2.17	152.89	155.95
5	29	37.97	41.78	1.1	1.67	184.85	203.34
6	81	773.86	621.49	0.8	9.48	1720.7	1376.56
99	945	NA	1.35	3.16	0	1.35	4.27

After outlier removal

strata	ratio_n	base	resp	ratio	sd	est_base	est
0	19	4.55	7.11	1.56	1.07	15.71	24.51
3	17	10.5	10.38	0.99	1.28	59.01	58.42
4	24	24.39	16.32	0.67	1.29	152.89	102.44
5	28	35.84	35.78	1	1.55	184.85	184.85
6	80	679.5	621.49	0.91	4.17	1720.7	1565.84
99	945	NA	1.35	3.16	0	1.35	4.27

D8



Before outlier removal

strata	ratio_n	base	resp	ratio	sd	est_base	est
0	25	12.7	10.06	0.79	0.92	33.36	26.35
3	18	11.03	7.63	0.69	0.39	43.95	30.33
4	16	11	6.43	0.58	0.47	55.49	32.18
6	55	126.69	137.19	1.08	3.36	220.9	238.57
99	945	NA	1.11	3.16	0	1.11	3.51
Total	-	-	-	-	-	354.81	330.94

After outlier removal

strata	ratio_n	base	resp	ratio	sd	est_base	est
0	23	5.21	5.19	1	0.23	33.36	33.36
3	17	9.32	7.63	0.82	0.26	43.95	36.04
4	15	8.77	6.43	0.73	0.33	55.49	40.51
6	54	126.69	117.69	0.93	2.06	220.9	205.44
99	945	NA	1.11	3.16	0	1.11	3.51
Total	-	-	-	-	-	354.81	318.86

Challenges

1. Re-creating the statistical methods that were used by GenStat.
   - Standard error and % influence

2. De-bugging over all ~100 variables
   - E.g. code runs fine for variables 1-50 but throws an error on variable 51….

3. Changing my plan as I went on

Still to do

- Test different methods of outlier removal
- Imputation model for new holdings
- Use this new method for the June Survey 2018!

Data Science Accelerator

+ Dedicated time away from my day-to-day job
+ Fresh pair of eyes and an objective opinion
+ Coding help if I needed it

– Lack of specialist statistical knowledge
– Only 3 accelerants in the Sheffield hub

Department
for Environment
Food & Rural Affairs

Data Science Accelerator

Ratio Imputation of the June Survey of Agriculture & Horticulture

Francesca Parrott, Defra

Any questions?