

Clustered Data Analysis Exercise (VER 20) Solution

1. Generate some descriptive statistics to get a feel for the data..

A: Left to you ... important things to note are:

- no missing values
- 6 farms - farm #6 had relatively few records
- 416 litters and 1114 observations, hence an average of 2.8 pigs per litter
- -dwg_wean-: outcome variable has an approximately bell-shaped distributed (not need to quantify further or test for normality because the normality assumption is for errors, not the outcome itself)
- -parity-: range 1 to 11, reasonable distribution from 1-6 but then sparse (but we are just going to assume linearity for this exercise so this wasn't investigated further)
- -ap2_t-: 118 pos, 996 neg
- -mp_t-: 305 pos, 809 neg
- -infl_t-: 200 pos, 914 neg
- -prrs_t-: 237 pos, 837 neg

2. Multiple linear regression model ignoring any clustering of pigs within a litter or farm.

```
. regress dwg_wean parity ap2_t mp_t infl_t prrs_t
```

Source	SS	df	MS			
Model	.200378376	5	.040075675	Number of obs =	1114	
Residual	5.01049207	1108	.004522105	F(5, 1108) =	8.86	
Total	5.21087044	1113	.004681824	Prob > F	= 0.0000	
				R-squared	= 0.0385	
				Adj R-squared	= 0.0341	
				Root MSE	= .06725	

dwg_wean	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parity	.0045047	.0009307	4.84	0.000	.0026786	.0063307
ap2_t	.0265452	.0065809	4.03	0.000	.0136328	.0394576
mp_t	-.0113186	.0045555	-2.48	0.013	-.020257	-.0023803
infl_t	.0025429	.0053185	0.48	0.633	-.0078926	.0129784
prrs_t	.001996	.00498	0.40	0.689	-.0077754	.0117673
_cons	.3774745	.0044933	84.01	0.000	.3686581	.3862909

A: Parity and diseases only explain about 4% of the variation in daily weight gain ($R^2=0.0385$). The predictors -parity-, -ap2_t- and -mp_t- are statistically significant ($P<0.05$).

3. Modeling with -farm_id- as a fixed effect.

(a) Effects on estimates and their significance.

```
. regress dwg_wean parity ap2_t mp_t infl_t prrs_t i.farm_id
```

Source	SS	df	MS			
Model	.557369208	10	.055736921	Number of obs =	1114	
Residual	4.65350124	1103	.004218949	F(10, 1103) =	13.21	
Total	5.21087044	1113	.004681824	Prob > F	= 0.0000	
				R-squared	= 0.1070	
				Adj R-squared	= 0.0989	
				Root MSE	= .06495	

dwg_wean	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
parity	.0045047	.0009307	4.84	0.000	.0026786	.0063307
ap2_t	.0265452	.0065809	4.03	0.000	.0136328	.0394576
mp_t	-.0113186	.0045555	-2.48	0.013	-.020257	-.0023803
infl_t	.0025429	.0053185	0.48	0.633	-.0078926	.0129784
prrs_t	.001996	.00498	0.40	0.689	-.0077754	.0117673
_cons	.3774745	.0044933	84.01	0.000	.3686581	.3862909

parity		.0043214	.0009098	4.75	0.000	.0025363	.0061065
ap2_t		.0327513	.0068534	4.78	0.000	.0193041	.0461985
mp_t		-.0049073	.0047918	-1.02	0.306	-.0143093	.0044948
infl_t		.0072444	.0051999	1.39	0.164	-.0029584	.0174471
prrs_t		.010505	.0052915	1.99	0.047	.0001225	.0208876
farm_id							
2		.0404278	.0075077	5.38	0.000	.0256968	.0551589
3		.0151046	.0064766	2.33	0.020	.0023968	.0278124
4		.0050318	.0072099	0.70	0.485	-.0091148	.0191784
5		-.014026	.0064168	-2.19	0.029	-.0266166	-.0014354
6		.0460065	.0131044	3.51	0.000	.0202941	.0717188
_cons		.3653756	.006605	55.32	0.000	.3524157	.3783355

A: The effects of -parity- and -ap2_t- are still significant and have not changed much in magnitude. The coefficient for _mp_t- is approximately halved and no longer significant. On the other hand, the coefficient for -prrs_t- has gone up 5-fold and is now weakly significant. We would expect these changes to be results of confounding effects of farms.

(b) Significance of the farm predictor.

```
. testparm i.farm_id

( 1) 2.farm_id = 0
( 2) 3.farm_id = 0
( 3) 4.farm_id = 0
( 4) 5.farm_id = 0
( 5) 6.farm_id = 0

F( 5, 1103) = 16.92
Prob > F = 0.0000
```

A: The farm effects are strongly significant.

(c) Explained variance by farms.

A: Adding farm raised the R² from 0.0385 to 0.107, an increase of 0.068. Relatively speaking this is a quite large increase, but we're still talking about models with very low predictive ability.

(d) Variation of weight gains across farms.

A: The estimated farm effects vary from a low of -0.014 (relative to the baseline) for farm=5 to +0.046 for farm=6, so the range of -dwg_wean- across farms is 0.06. This is about the same as the residual standard deviation (root MSE), so compared to the natural variability in the weight gains there is not a really large amount of variability between farms (but an average effect of 1 standard deviation can still be very significant).

4. Modeling with random litter effects.

(a) Effects on estimates and their significance.

```
. mixed dwg_wean parity ap2_t mp_t infl_t prrs_t i.farm_id || litt_id:, reml

Performing EM optimization:

Performing gradient-based optimization:

Iteration 0: log restricted-likelihood = 1511.5121
Iteration 1: log restricted-likelihood = 1511.5121

Computing standard errors:

Mixed-effects REML regression          Number of obs    =    1114
Group variable: litt_id                Number of groups  =     416
```

```

Obs per group: min =      1
                avg =      2.7
                max =      3

Log restricted-likelihood = 1511.5121      Wald chi2(10) = 75.87
                                          Prob > chi2 = 0.0000
-----+-----
   dwg_wean |      Coef.   Std. Err.   z   P>|z|   [95% Conf. Interval]
-----+-----
   parity   |   .0042685   .0012138   3.52  0.000   .0018895   .0066476
   ap2_t    |   .0264896   .0078248   3.39  0.001   .0111534   .0418259
   mp_t     |  -.0048876   .0047293  -1.03  0.301   -.0141569   .0043816
   infl_t   |   .0086204   .0050384   1.71  0.087   -.0012548   .0184955
   prrs_t   |   .0079813   .0057995   1.38  0.169   -.0033856   .0193481
   farm_id  |
   2        |   .0404373   .0098319   4.11  0.000   .0211672   .0597074
   3        |   .014821    .0084985   1.74  0.081   -.0018357   .0314777
   4        |   .0024839   .0094104   0.26  0.792   -.0159601   .0209278
   5        |  -.0141871   .0084382  -1.68  0.093   -.0307258   .0023515
   6        |   .0453742   .0169869   2.67  0.008   .0120805   .078668
   _cons    |   .3674336   .0085091  43.18  0.000   .3507561   .3841111
-----+-----
Random-effects Parameters |      Estimate   Std. Err.   [95% Conf. Interval]
-----+-----
litt_id: Identity
   var(_cons) |   .0018959   .0002028   .0015374   .0023381
-----+-----
   var(Residual) |   .0023565   .0001262   .0021216   .0026174
-----+-----
LR test vs. linear regression: chibar2(01) = 180.35 Prob >= chibar2 = 0.0000

```

A: The effect and statistical significance of -infl_t- get stronger, and they get weaker for -prrs_t- compared to the model without random effect of litter. These two variables probably show some clustering at the litter level. Other parameters do not change much.

(b) What proportion of variance resides at the litter level?

A: We calculate: $(.00190) / (.00190 + 0.00236) = 0.446 = 45\%$. This also means the correlation between weight gains of two pigs from the same litter is 0.45. That is, -dwg_wean- is appreciably clustered at the litter level.

(c) Variation of weight gains across litters.

A: We can compute a 95% range for litter effects as $\pm 1.96 * \sqrt{0.0019} = \pm 0.085$. That is, 95% of the litter effects will be within $2 * 0.085 = 0.17$. We can also look at the estimated litter random effects (see do-file), and their range is from -0.108 to 0.090. Considering that this range corresponds to 416 litters, it is more narrow than the 95% predicted range computed above. Generally speaking, the estimated random effects are closer to the overall mean than the simple means (for litters) and the predictions; this phenomenon is called shrinkage towards of the mean of the estimates. The predicted range is therefore the most adequate range to compute.

5. Modeling with both random litter and farm effects.

(a) Arguments in favour of, and against, treating -farm_id- as a random effect?

A: If the farms were specifically selected for this study and are not representative of farms in the general population, then treating them as fixed effects may be more appropriate. On the other hand, the linear mixed model gives us an estimate of the amount of variation at the farm level.

(b) Proportions of variance at each of the three levels.

```

. mixed dwg_wean parity ap2_t mp_t infl_t prrs_t || farm_id: || litt_id:, reml

Performing EM optimization:
Performing gradient-based optimization:

```

```
Iteration 0: log restricted-likelihood = 1522.3627
Iteration 1: log restricted-likelihood = 1522.3627
```

Computing standard errors:

```
Mixed-effects REML regression          Number of obs   =   1114
```

Group Variable	No. of Groups	Observations per Group		
		Minimum	Average	Maximum
farm_id	6	30	185.7	293
litt_id	416	1	2.7	3

```
Log restricted-likelihood = 1522.3627          Wald chi2(5)      =   27.44
                                                Prob > chi2      =   0.0000
```

dwg_wean	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
parity	.0042416	.0012123	3.50	0.000	.0018654	.0066177
ap2_t	.0262083	.0077942	3.36	0.001	.010932	.0414846
mp_t	-.005472	.0047072	-1.16	0.245	-.0146979	.0037539
infl_t	.0082594	.0050338	1.64	0.101	-.0016067	.0181254
prrs_t	.0071656	.0057612	1.24	0.214	-.0041262	.0184574
_cons	.3809482	.0103724	36.73	0.000	.3606187	.4012776

Random-effects Parameters	Estimate	Std. Err.	[95% Conf. Interval]	
farm_id: Identity				
var(_cons)	.0004433	.000331	.0001026	.0019158
litt_id: Identity				
var(_cons)	.0018977	.000203	.0015388	.0023404
var(Residual)	.0023563	.0001262	.0021214	.0026172

```
LR test vs. linear regression:          chi2(2) = 243.19   Prob > chi2 = 0.0000
```

A: We compute the total (unexplained) variance as $0.00044+0.00190+0.00236 = .00470$. Then the proportion of variance at each of the levels is obtained by dividing with 0.0047. This yields

```
. di "propn. of var. at farm level  " .0004433 / .0047
propn. of var. at farm level  .09431915
. di "propn. of var. at litter level  " .0018977 / .0047
propn. of var. at litter level  .40376596
. di "propn. of var. at pig level  " .0023563 / .0047
propn. of var. at pig level  .50134043
```

There is only little variance at the farm level (compared to the litter and pig levels).

(c) Intraclass correlations for (i) pigs (from different litters) within a herd and (ii) two pigs within the same litter.

A: We can use the built-in Stata command to compute those values:

```
. estat icc
```

```
Residual intraclass correlation
```

Level	ICC	Std. Err.	[95% Conf. Interval]	
farm_id	.094373	.0640448	.0234289	.3115963
litt_id farm_id	.4983722	.0448083	.411491	.5853519

The correlation between weight gains of two pigs in the same litter is 0.50 (so pretty high), and the correlation between two pigs from the same herd but different litters is 0.09 (quite low).