Small Area Estimation

Session 7

Model Dependent Methods

Cross-Sectional Models for Continuous Measurements II

Preface

In Session **6** we considered the **Area level model (Fay & Herriot)**, which assumes that the covariate information is only available at the **area level**.

Below we consider a model that uses **individual** covariate information.

Nested Error Unit Level Regression model

(**Battese, Harter and Fuller, 1988**).

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + u_i + \varepsilon_{ij} \; ; \; i = 1...M, \; j = 1...N_i$$ **(model assumed for Pop.)**

u_i and ε_{ij} are independent errors with variances σ_u^2 and σ_ε^2 respectively.

Here again, the u_i's represent the combined effect of **area characteristics**, not accounted for by the auxiliary variables.

- The area means, $\bar{\mathbf{X}}_i = \sum_{j=1}^{Ni} \mathbf{x}_{ij} / N_i$ are assumed to be **known**.

Unit level regression model (cont.)

Observations: $y_{ij} = x'_{ij}\beta + u_i + \varepsilon_{ij},\ \ i = 1...m,\ j = 1...n_i.$

Denoting by $\bar{Y}_i = \sum_{j=1}^{N_i} y_{ij} / N_i$ the population mean of the ***y***-values for area ***i*** under the model is $\bar{Y}_i = \bar{\mathbf{X}}_i\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i$, but for sufficiently large N_i, $\bar{\varepsilon}_i \approx 0$, [$E(\bar{\varepsilon}_i) = 0$ and $Var(\bar{\varepsilon}_i) = \sigma_\varepsilon^2 / N_i$], and hence the target parameter of interest is commonly defined as,

$$\theta_i = \bar{\mathbf{X}}'_i\boldsymbol{\beta} + u_i = E(\bar{Y}_i \mid \bar{\mathbf{X}}_i, u_i).$$

Under this model, ***Cov***$(y_{ij}, y_{kl}) = \begin{cases} \sigma_u^2 + \sigma_\varepsilon^2, & \text{if } (i,j) = (k,l) \\ \sigma_u^2, & \text{if } i = k,\ j \neq l \\ 0, & \text{if } i \neq k \end{cases}$

BLUP under the unit level regression model

For **known variances**, the **BLUP** of θ_i is,

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)'\hat{\boldsymbol{\beta}}_{GLS}] + (1-\gamma_i)\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS}$$

$$= \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS} + \gamma_i[\bar{y}_i - \bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{GLS}] = \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS} + \hat{u}_i$$

where,

$\bar{\mathbf{x}}_i$ and $\bar{y}_i$ are the sample means **;** $\gamma_i = \dfrac{\sigma_u^2}{\sigma_u^2 + (\sigma_\varepsilon^2 / n_i)}$ is the **shrinkage factor**.

[σ_{Di}^2 in the Fay-Herriot model is now replaced by $(\sigma_\varepsilon^2 / n_i) = Var(\bar{y}_i \mid u_i)$.**]**

BLUP for unit level regression model (cont.)

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = (\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}\sum_{i=1}^{m}\mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{y}_i$$

$$\mathbf{y}_i = (y_{i1},\ldots,y_{in_i})',\ \mathbf{y} = (\mathbf{y}_1',\ldots,\mathbf{y}_m')',\ \mathbf{X}_i = (\mathbf{x}_{i1},\ldots,\mathbf{x}_{in_i})',\ \mathbf{X} = (\mathbf{X}_1',\ldots,\mathbf{X}_m')'$$

$$\mathbf{V}_i = \sigma_\varepsilon^2\mathbf{I}_{n_i} + \sigma_u^2\mathbf{1}_{n_i}\mathbf{1}_{n_i}',\ \mathbf{V} = \mathbf{I}_m \otimes \mathbf{V}_i.$$

- When the **parameter** of interest is $\bar{Y}_i = \bar{\mathbf{X}}_i\boldsymbol{\beta} + u_i + \bar{\varepsilon}_i$,

$$\hat{\bar{Y}}_i = \frac{1}{N_i}[\sum_{j=1}^{n_i} y_{ij} + (N_i - n_i)\hat{u}_i + \sum_{j\notin s_i}\mathbf{x}_{ij}'\hat{\boldsymbol{\beta}}_{GLS}]$$

$$= f_i\bar{y}_i + (1-f_i)[\bar{\mathbf{X}}_i^{c\prime}\hat{\boldsymbol{\beta}}_{GLS} + \gamma_i(\bar{y}_i - \bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{GLS})]$$

$f_i = (n_i / N_i)$ → **sampling fraction**.

$\bar{\mathbf{X}}_i^c = \frac{\sum_{j\notin s_i}\mathbf{x}_{ij}}{(N_i - n_i)}$ → **mean of nonsampled units in area *i***.

BLUP for unit level regression model (cont.)

Consider the estimation of $\theta_i = \bar{\mathbf{X}}_i'\boldsymbol{\beta} + u_i = E(\bar{Y}_i \mid \bar{\mathbf{X}}_i, u_i)$,

$$\hat{\theta}_i = \gamma_i[\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)'\hat{\boldsymbol{\beta}}_{GLS}] + (1-\gamma_i)\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS}$$

$$= \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS} + \gamma_i[\bar{y}_i - \bar{\mathbf{x}}_i'\hat{\boldsymbol{\beta}}_{GLS}] = \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS} + \hat{u}_i$$

$\hat{\theta}_i$ is in this case a linear combination of the '**regression estimator**' $\bar{y}_i + (\bar{\mathbf{X}}_i - \bar{\mathbf{x}}_i)'\hat{\boldsymbol{\beta}}_{GLS}$, and the '**synthetic estimator**' $\bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS}$, or alternatively, the '**synthetic estimator**' + a '**correction factor**'.

- For $i \notin s, \hat{\theta}_i = \bar{\mathbf{X}}_i'\hat{\boldsymbol{\beta}}_{GLS}$ (**synthetic estimator** is used for **nonsampled areas**)
- For $\bar{\mathbf{x}}_i = \bar{\mathbf{X}}_i$ the estimator is the **same** as under the Fay and Herriot model.

Variance under unit level regression model

Variance of BLUP (**known variances**):

$$E(\hat{\theta}_i - \theta_i)^2 = (\sigma_\varepsilon^2 / n_i)\gamma_i + \mathbf{c}_i' \, Var(\hat{\boldsymbol{\beta}}_{GLS}) \, \mathbf{c}_i \; ; \quad \mathbf{c}_i' = (\bar{\mathbf{X}}_i' - \gamma_i \bar{\mathbf{x}}_i').$$

- For $\bar{\mathbf{x}}_i = \bar{\mathbf{X}}_i$ the variance is the **same** as under the Fay and Herriot model (with $\sigma_{Di}^2 = \sigma_\varepsilon^2 / n_i$).

Exercise: Compute the variance of the synthetic estimator $\bar{\mathbf{X}}_i' \hat{\boldsymbol{\beta}}_{GLS}$:

I- when predicting $\theta_i, i \notin s$; and

II- when predicting $\bar{Y}_i, i \notin s$.

The case of unknown variances

As with the Fay and Herriot model, **EBLUP** estimators are obtained by replacing the unknown variances by their sample estimators.

Battese *et al.* (1988) use the following sample estimators:

$$\hat{\sigma}_{\varepsilon}^{2} = \frac{1}{n-m-p+1}\sum_{i,j} e_{ij}^{2} \text{ ; } \quad n = \sum_{i=1}^{m} n_i \text{ (total sample size) , } \quad p = \dim(\mathbf{x}_{ij}).$$

The e_{ij}'s are the estimated residuals when regressing $\tilde{y}_{ij} = (y_{ij} - \bar{y}_i)$ on the transformed covariates $\tilde{\mathbf{x}}_{ij} = (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)$ in areas with n_i>1. $\hat{\sigma}_{\varepsilon}^{2}$ is also the ordinary variance estimator when regressing $y_{ij} = x_{ij}'\beta + u_i + \varepsilon_{ij}$ with **fixed effects** u_i's. (The u_i's are coefficients of dummies defining the areas).
$\bar{\mathbf{x}}_i, \bar{y}_i$ are the area sample means, $\boldsymbol{\beta}$ is estimated by **OLS**.

Exercise: Show**:** $(y_{ij} - \bar{y}_i) = (\mathbf{x}_{ij} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{ij} - \bar{\varepsilon}_i)$ and compute the variance and covariances of the residuals $\varepsilon_{ij}^{*} = (\varepsilon_{ij} - \bar{\varepsilon}_i)$ (The u_i's cancel out.)

- $E(\hat{\sigma}_{\varepsilon}^{2}) = \sigma_{\varepsilon}^{2}$ by classical regression theory.

The case of unknown variances (cont.)

$$\tilde{\sigma}_u^2 = \frac{1}{n^*}[\sum_{i,j} \hat{v}_{ij}^2 - (n-p)\hat{\sigma}_\varepsilon^2];$$

$$n^* = n - trace[(\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^{m} n_i^2 \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i'] \text{ , } \mathrm{X}'\mathrm{X} = \sum_{\mathrm{i}=1}^{\mathrm{m}} \sum_{\mathrm{j}=1}^{\mathrm{n_i}} \mathbf{x}_{ij}\mathbf{x}_{ij}'.$$

The $\hat{v}_{ij}$'s are the estimated residuals from the ordinary least squares regression of y_{ij} on $\mathbf{x}_{ij}$. $\hat{\sigma}_u^2 = \mathbf{max}(0, \tilde{\sigma}_u^2)$.

Other estimators like **MLE** or **REML** can be considered instead. See the book by Rao (2003) and Rao and Molina (2015).

- MSE estimation of **EBLUP** is considered in Session 8.

Application (Battese, Harter and Fuller, 1988)

Battese *et al*. used the model to estimate the number of **hectares** of **corn** and **soybeans** for **12 counties** in the State of Iowa, U.S.A.

(**1 Hectare = 10,000m^2 ≅ 2.47 acre**).

Each **county** was divided into **area segments** with the number of **segments** in a county ranging from **394** to **687,** (see **Table 3;** a **segment** consists of about **250 hectares**).

In this application $\mathbf{x}'_{ij} = (1, x_{1,ij}, x_{2,ij})$ where $x_{1,ij}$ ($x_{2,ij}$) defines the number of **pixels** of **corn** (**soybeans**) in segment ***j*** of county ***i*** , as obtained from **Satellite readings** (**1 pixel = 0.45 hectare**).

These values are known for every segment in every county (and hence also the true county means, $\bar{\mathbf{X}}_i$).

Application (cont.)

For a **sample** of segments, the areas under **corn** and **soybeans** have been evaluated by farm operators (the **dependent variable** values, y_{ij}).

The number of sample **segments**, n_i in a **county** varies from **1** to **5**, and $n = \sum_{i=1}^{m} n_i = 37$.

The model:

$$y_{ij} = \beta_0 + x_{1,ij}\beta_1 + x_{2,ij}\beta_2 + u_i + \varepsilon_{ij}, \quad i = 1...12, \quad j = 1...n_i$$

Some of the results of this study are summarized in **Tables 3-5** and **Figures 1-2**.

Application, summary

m=12 counties in north-central Iowa (U.S.A).

i=county, ***j***=area segment; n_i = **1,…,5**.

y_{ij} = number of hectares of corn (or soybeans) in segment ***j*** of county ***i*** as evaluated in the farm interview survey.

x_{1ij} = number of pixels (picture elements of 0.45 hectares) classified as **corn** in segment (***i,j***) (satellite data).

x_{2ij} = number of pixels classified as **soybeans** in segment (***i,j***) (satellite data).

$$\mathbf{x}'_{ij}\boldsymbol{\beta} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} \quad ; \quad \theta_i = \beta_0 + \beta_1 \bar{X}_{1i} + \beta_2 \bar{X}_{2i} + u_i .$$

Application, summary (cont.)

Estimates (standard errors) for **corn:**

$\hat{\beta}_0 = 51(25),\ \hat{\beta}_1 = 0.329(0.05)\ \beta_2 = -0.134(0.056)$

$\hat{\sigma}_\varepsilon^2 = 150(45),\ \hat{\sigma}_u^2 = 140(89)$

Estimates (standard errors) for **soybeans:**

$\hat{\beta}_0 = -16(29),\ \hat{\beta}_1 = 0.028(0.058)\ \beta_2 = 0.494(0.065)$

$\hat{\sigma}_\varepsilon^2 = 195(59),\ \hat{\sigma}_u^2 = 272(49)$

Plot of reported hectares of corn versus satellite pixels of corn by county

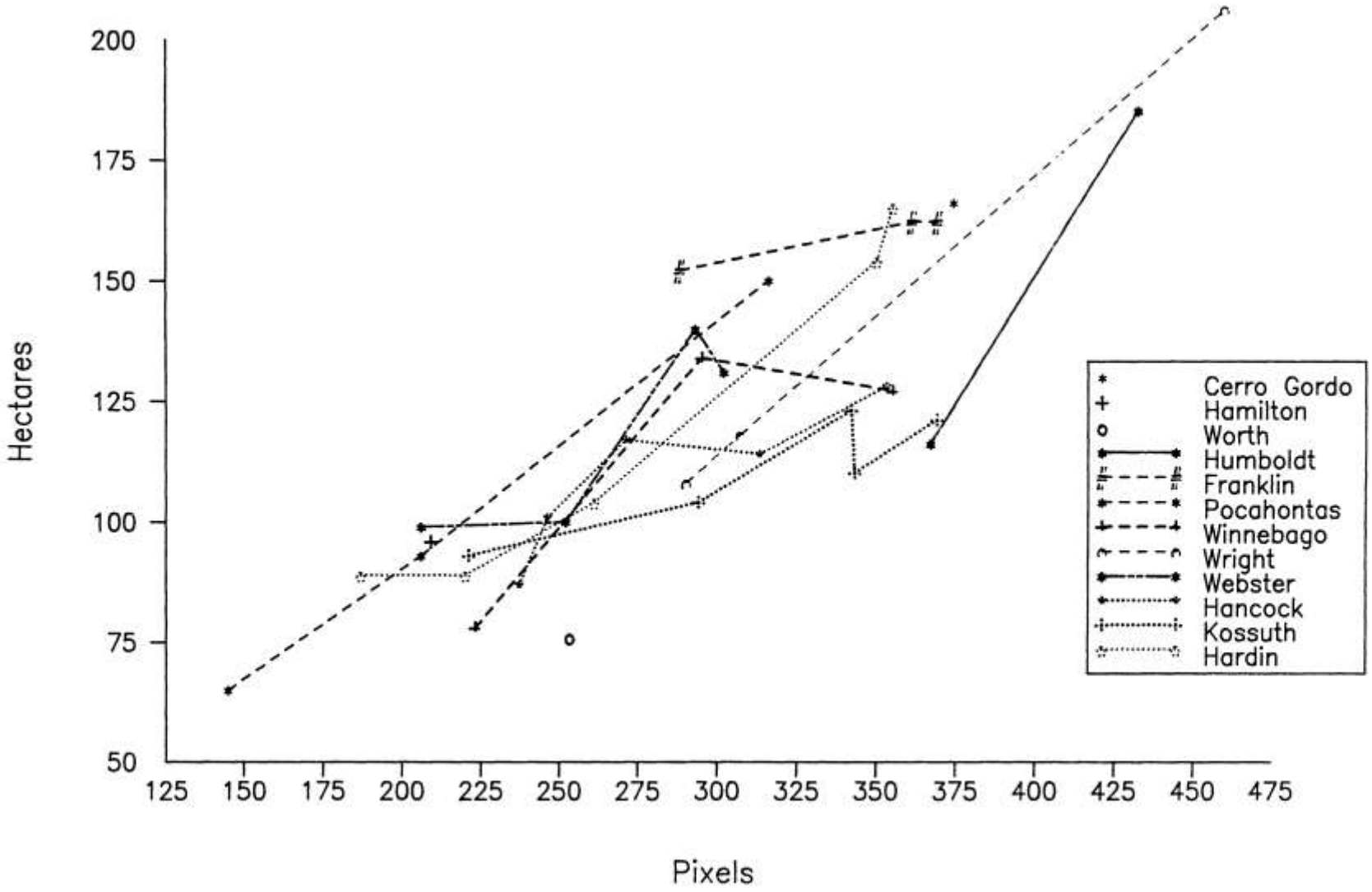

Figure 1. Plot of Corn Hectares Versus Corn Pixels by County.

Plot of reported hectares of soybeans versus satellite pixels of soybeans by county

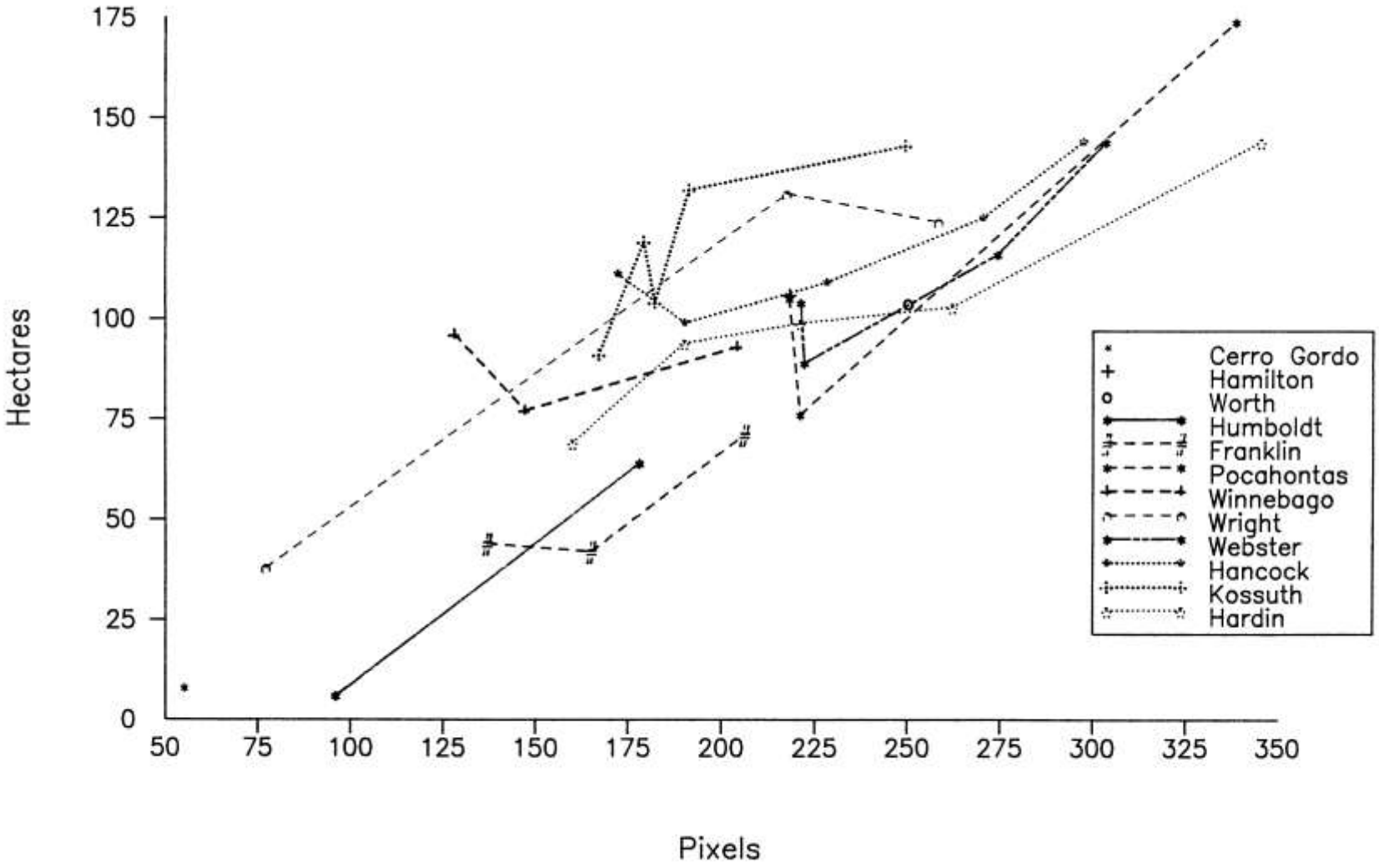

Figure 2. Plot of Soybean Hectares Versus Soybean Pixels by County.

Survey and satellite data

Table 1. Survey and Satellite Data for Corn and Soybeans in 12 Iowa Counties

County	No. of segments		Reported hectares		No. of pixels in sample segments		Mean number of pixels per segment*	
	Sample	County	Corn	Soybeans	Corn	Soybeans	Corn	Soybeans
Cerro Gordo	1	545	165.76	8.09	374	55	295.29	189.70
Hamilton	1	566	96.32	106.03	209	218	300.40	196.65
Worth	1	394	76.08	103.60	253	250	289.60	205.28
Humboldt	2	424	185.35	6.47	432	96	290.74	220.22
			116.43	63.82	367	178		
Franklin	3	564	162.08	43.50	361	137	318.21	188.06
			152.04	71.43	288	206		
			161.75	42.49	369	165		
Pocahontas	3	570	92.88	105.26	206	218	257.17	247.13
			149.94	76.49	316	221		
			64.75	174.34	145	338		
Winnebago	3	402	127.07	95.67	355	128	291.77	185.37
			133.55	76.57	295	147		
			77.70	93.48	223	204		
Wright	3	567	206.39	37.84	459	77	301.26	221.36
			108.33	131.12	290	217		
			118.17	124.44	307	258		
Webster	4	687	99.96	144.15	252	303	262.17	247.09
			140.43	103.60	293	221		
			98.95	88.59	206	222		
			131.04	115.58	302	274		
Hancock	5	569	114.12	99.15	313	190	314.28	198.66
			100.60	124.56	246	270		
			127.88	110.88	353	172		
			116.90	109.14	271	228		
			87.41	143.66	237	297		
Kossuth	5	965	93.48	91.05	221	167	298.65	204.61
			121.00	132.33	369	191		
			109.91	143.14	343	249		
			122.66	104.13	342	182		
			104.21	118.57	294	179		
Hardin	6	556	88.59	102.59	220	262	325.99	177.05
			88.59	29.46	340	87		
			165.35	69.28	355	160		
			104.00	99.15	261	221		
			88.63	143.66	187	345		
			153.70	94.49	350	190		

* The mean number of pixels of a given crop per segment in a county is the total number of pixels classified as that crop, divided by the number of segments in that county.

Estimates for Corn

Table 2. Predicted Hectares of Corn With Standard Errors of Alternative Predictors

County	Sample segments	Predicted hectares	Standard errors		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	122.2	9.6	13.7	30.5
Hamilton	1	126.3	9.5	12.9	30.5
Worth	1	106.2	9.3	12.4	30.5
Humboldt	2	108.0	8.1	9.7	21.5
Franklin	3	145.0	6.5	7.1	17.6
Pocahontas	3	112.6	6.6	7.2	17.6
Winnebago	3	112.4	6.6	7.2	17.6
Wright	3	122.1	6.7	7.3	17.6
Webster	4	115.8	5.8	6.1	15.2
Hancock	5	124.3	5.3	5.7	13.6
Kossuth	5	106.3	5.2	5.5	13.6
Hardin	5	143.6	5.7	6.1	13.6

Estimates for Soybeans

Table 3. Predicted Hectares of Soybeans With Standard Errors of Alternative Predictors

County	Sample segments	Predicted hectares	Standard errors		
			Best predictor	Survey regression predictor	Sample mean
Cerro Gordo	1	77.8	12.0	15.6	29.1
Hamilton	1	94.8	11.8	14.8	29.1
Worth	1	86.9	11.5	14.2	29.1
Humboldt	2	79.7	9.7	11.1	20.6
Franklin	3	65.2	7.6	8.1	16.8
Pocahontas	3	113.8	7.7	8.2	16.8
Winnebago	3	98.5	7.7	8.3	16.8
Wright	3	112.8	7.8	8.4	16.8
Webster	4	109.6	6.7	7.0	14.6
Hancock	5	101.0	6.2	6.5	13.0
Kossuth	5	119.9	6.1	6.3	13.0
Hardin	5	74.9	6.6	6.9	13.0

Model testing in application

Battese *et al.* performed several goodness of fit tests:

1. They added the variables $x_{1,ij}^2$ and $x_{2,ij}^2$ as **additional explanatory variables** and tested their significance using **conventional t-tests**. **Neither** variable was found significant.

2. To test that the errors ε_{ij} and u_i have **normal** distributions (assumed for estimation of **MSE**), the authors **computed adjusted residuals**,

$$y_{ij}^* = (y_{ij} - \alpha_i \bar{y}_i) - (\mathbf{x}_{ij} - \alpha_i \bar{\mathbf{x}}_i)'\boldsymbol{\beta} \; ; \; \alpha_i = 1 - \left(\frac{\sigma_\varepsilon^2 / n_i}{\sigma_\varepsilon^2 / n_i + \sigma_u^2}\right)^{1/2}, \quad \mathbf{x}_{ij} = (1, x_{1,ij}, x_{2,ij})'.$$

The adjusted residuals (with **known parameters**) are **uncorrelated** with **mean zero** and **variance** σ_ε^2.

Exercise: Show that $Var(y_{ij}^*) = \sigma_\varepsilon^2$. **Hint**: Write y_{ij}^* as function of ε_{ij} and u_i only. **Note**: $Cov(\varepsilon_{ij}, \varepsilon_{il}) = 0$ for $j \neq l$ under the model.

Model testing, continued

Estimated adjusted residuals are obtained by replacing the unknown variances by the sample estimates, and $\boldsymbol{\beta}$ by $\hat{\beta}_{G\hat{L}S}$ ($\hat{\beta}_{G\hat{L}S}$ is the same as $\hat{\beta}_{GLS}$ but with **estimated variances**).

The normality of the error terms was tested by applying the Shapiro-Wilks test to the estimated adjusted residuals, yielding large **p-values** for both the **corn** and the **soybeans** data. (The normality assumption **is not rejected**.)

- One can study also the behaviour of the estimated standardized residuals, $(y_{ij} - \mathbf{x}'_{ij}\hat{\boldsymbol{\beta}}_{G\hat{L}S} - \tilde{u}_i)/\hat{\sigma}_\varepsilon$; $\tilde{u}_i = \hat{\gamma}_i[\bar{y}_i - \bar{\mathbf{x}}'_i\hat{\boldsymbol{\beta}}_{G\hat{L}S}]$, to detect **outlying** observations.

- Many other tests for SAE models are proposed in the literature.