Influential Observations in Regression Models

Ted Chang
University of Virginia

Phillip S. Kott
National Agricultural Statistical Service

NASS Census of Agriculture

Target population: all ‘farms’ defined to be entities with $1000 in annual agricultural sales or the potential for $1000 in sales.

The Census uses mailing lists maintained by the state level NASS offices.

To correct for under coverage in the mailing lists, NASS conducts an Area Frame Survey.

AREA FRAME SURVEY

Stratified sample of ‘segments’: usually 1 square mile each.

Strata constructed within each state based primarily upon the % of land devoted to agriculture.

- strata 11-19: >75% cultivated
- strata 21-29: 15-75% cultivated
- stratum 31: agri-urban
- stratum 32: dense urban
- strata 41-49: <15% cultivated

Using aerial photographs, segments are divided into ‘tracts’. All tracts in a sampled segment are enumerated.

No noncoverage/non response problems.

Project goal: develop a statistical model for the ‘probability’ that a farm is not on the mailing list.

Model variables: sales and stratum

(Original study also used variables related to crops produced, participation in USDA support programs, demographic characteristics, and horse ownership.)

Assumed model: $p_i=(1+\exp(-\beta^T X_i))^{-1}$

β: model coefficients

p_i: probability that ith farm is NOT on the mailing list (NML)

X_i: column vector of covariates (sales, stratum) for ith farm to be used to predict p_i

w_i: product of sampling weight and tract to farm acreage ratio ('fudge factor')

The most important variables are related to sales. Best model using the sales variables:

	int.	sales5K	sales50K	sales1M
coef. β_j	0.320	-1.465	-0.847	-1.449
st. error	0.170	0.218	0.257	0.708
s.e. total	0.171	0.219	0.258	0.713

sales5K = 1 if sales at least $5000, 0 otherwise

Here standard error is design based, denoted $\sqrt{\hat{V}_{db}}$, calculated using Binder (1983).

Consider the 'super population model':
$y_i \sim bin(1,p_i)$, $p_i=p_i(\beta)=(1+\exp(-\beta^T X_i))^{-1}$, $i \in U$
Model ignores cluster and stratum effects not explicitly incorporated into the X_i.

Finite population parameter B maximizes

$$\sum_{i\in U} y_i \log(p_i(B)) + (1-y_i)\log(1-p_i(B))$$

'total variance':

$$Var_{db,m}(\hat{\beta}) = E_m(Var_{db}(\hat{\beta})) + Var_m(E_{db}(\hat{\beta}))$$

$E_m(Var_{db}(\hat{\beta}))$ **is estimated by Binder's** $\hat{V}_{db}$**.**

$Var_m(E_{db}(\hat{\beta})) \approx Var_m(B) = O(N^{-1})$ **should be <<** $\hat{V}_{db}$**.**

Suppose we add indicator variables for the strata:

	sales5K	sales50K	sales1M	str11	str17
coef. βj	-1.302	-0.860	-1.612	0.082	-0.296
st. error	0.233	0.262	0.733	0.269	0.313
s.e. total	0.234	0.263	0.737	0.270	0.314

str19	str21	str27	str31	str32	str41	str45
0.689	0.677	0.011	2.285	17.395	0.584	1.566
1.935	0.295	0.311	0.824	1.037	0.254	0.388
1.940	0.296	0.312	0.827	2.018	0.255	0.394

Stratum 32 has 1 data point!

Suppose we recode with an intercept and remove str32:

	int	sales5K	sales50K	sales1M	str11
coef. βj	17.39	-1.302	-0.856	-1.612	-17.48
st. error	1.037	0.233	0.262	0.733	1.034
s.e. total	2.018	0.234	0.263	0.737	2.016

str17	str19	str21	str27	str31	str41	str45
-17.69	-16.71	-16.72	-17.38	-15.11	-16.81	-15.83
1.049	2.175	1.065	1.068	1.320	1.040	1.087
2.024	2.783	2.033	2.034	2.178	2.020	2.045

str32 = int – str11 - . . . . – str45 ≈ 0, so

$\sum_{i\in s} w_i p_i(\hat{\beta})(1-p_i(\hat{\beta}))X_i X_i^T$ **is close to singular,**

$\hat{V}_{db,m}(\hat{\beta}) - \hat{V}_{db} = \left[\sum_{i\in s} w_i p_i(\hat{\beta})(1-p_i(\hat{\beta}))X_i X_i^T\right]^{-1}$ **is large.**

In regression setting:

$$Var_{db,m}(\hat{\beta}) - E_m(Var_{db}(\hat{\beta})) = Var_m(E_{db}(\hat{\beta})) \approx Var_m(B) = \sum_{i \in U} X_i X_i^T$$

is estimated by $\sum_{i \in s} w_i X_i X_i^T$**.**

Recall, in weighted linear regression: $\sum_{i \in s} w_i X_i X_i^T$ **is used to detect**

- **multicolinearity and instability in** $\hat{\beta}$
- **high leverage**

A point is influential if it is high leverage and has a large residual.

It turns out that a slightly different comparison of variances is more sensitive.

Let $MSE_0 = Var_m(\hat{\beta})$**. For linear regression**

$$\hat{\beta} = \left[\sum_{i\in s} w_i X_i X_i^T\right]^{-1} \sum_{i\in s} w_i X_i y_i$$

$$MSE_0 = \left[\sum_{i\in s} w_i X_i X_i^T\right]^{-1} \left[\sum_{i\in s} w_i^2 X_i X_i^T\right] \left[\sum_{i\in s} w_i X_i X_i^T\right]^{-1}$$

Now $E_m(\hat{\beta}) = \beta$**, so** $Var_{db,m}(\hat{\beta}) = E_{db}(MSE_0)$ **and hence** MSE_0 **estimates total variance.**

Let $MSE_L = E_m(\hat{V}_{db})$

(complicated design dependent formula).

Compare MSE_0 **to** MSE_L**.**

	sales5K	sales50K	sales1M	str10s	str20s	str30s	str40s
$\hat{\beta}$	-1.358	-0.765	-1.528	-0.158	0.338	2.918	0.704
$\hat{V}_{db}^{1/2}$	0.231	0.267	0.702	0.230	0.252	0.955	0.238
$\hat{V}_{db,m}^{1/2}$	0.232	0.268	0.707	0.230	0.253	0.958	0.239
$MSE_0^{1/2}$	0.233	0.271	0.620	0.190	0.205	1.097	0.231
$MSE_L^{1/2}$	0.230	0.266	0.606	0.187	0.198	0.936	0.227

Notice $\hat{V}_{db}^{1/2}$ and $\hat{V}_{db,m}^{1/2}$ are fairly close, but MSE_0 is about 37% bigger than MSE_L in `str30s`**.**

This is because strata 31-39 have 11 data points out of 1468 ($\hat{N}$ 1803.6 out of 66731.5).

Ex: Suppose n draws with replacement, weights d_i

$$U = U_1 \cup U_2$$

Let $X_i = 1 \quad i \in U_1; \; X_i = 0 \quad i \in U_2$

$$\hat{N}_1 = \sum_{s_1} d_i$$

$$\hat{\beta} = \hat{N}_1^{-1} \sum_{s_1} d_i y_i$$

$$\hat{V}_{db} = \frac{n}{n-1} \hat{N}_1^{-2} \sum_{s_1} d_i^2 (y_i - \hat{\beta})^2$$

Model: $E(y_i) = \beta \quad i \in U_1; \quad E(y_i) = 0 \quad i \in U_2$

$$Var(y_i) = \sigma^2$$

$$MSE_0 = Var_m(\hat{\beta}) = \hat{N}_1^{-2} \sum_{s_1} d_i^2 \sigma^2$$

$$MSE_L = E_m(\hat{V}_{db}) = \frac{n}{n-1} \hat{N}_1^{-2} \sum_{s_1} d_i^2 \left(\sigma^2 + 2Cov_m(y_i, \hat{\beta}) + V_m(\hat{\beta}) \right)$$

$$= \frac{n}{n-1} \left[MSE_0 - \frac{2\sigma^2}{\hat{N}_1^3} \sum_{s_1} d_i^3 + \frac{MSE_0}{\hat{N}_1^2} \sum_{s_1} d_i^2 \right]$$

Suppose $d_i = O(n^{-1}N)$ **so** $n_1 N_1^{-1} \approx nN^{-1}$. **Then**

$$MSE_0 = O(\frac{n_1 N^2}{N_1^2 n^2}) = O(\frac{N}{nN_1})$$

$$MSE_L = \frac{n}{n-1} \left[MSE_0 + O(\frac{N^2}{n^2 N_1^2}) \right]$$

so that if $N_1 N^{-1} \to 0$ **as** $n \to \infty$**, second term of** MSE_L **is not small relative to** MSE_0

Recall: Given two symmetric matrices A ($=MSE_L$) and B ($=MSE_0$), with A positive definite, there is are matrices P and L, L diagonal, such that

$$A = PP^T$$

$$B = PLP^T$$

P, L are the eigenvectors and eigenvalues of B in a coordinate system which orthogonalizes A.

Ex: model sales5K, sales50K, sales1000K, str10s, str20s, str30s, str40s

L = diag(1.41, 1.06, 1.05, 1.04, 1.03, 1.02, 1.02)

1st col of P:

sales5K	sales50K	sales1M	str10s	str20s	str30s	str40s
-1.300	-1.288	-0.003	-0.074	-0.032	-1.078	-0.031

	str11	str17	str19	str21	str27	str31	str32	str41	str45
<5K	76	21	1	43	47	6	0	46	9
5K-50K	91	48	1	31	43	1	0	45	2
50K-1M	292	88	1	60	37	2	1	63	2
>1M	288	27	14	40	14	1	0	25	2

$\hat{N}$	str11	str17	str19	str21	str27	str31	str32	str41	str45
<5K	4270.	2649.	37.39	5119.	5058.	983.3	0.000	3710.	923.1
5K-50K	4203.	5195.	86.00	2037.	3881.	220.1	0.000	2912.	199.0
50K-1000K	6203.	5112.	86.00	2645.	2156.	262.6	334.1	2497.	93.05
>1000K	2327.	441.5	339.8	1092.	515.5	3.418	0.000	1027.	113.3

The farm in str 31 with sales >1000K has low weight.

Ex (artificial data): Data generated according to the model

	int	sales1K	sales5K	sales50K	sales1M	age	hisp	str10s
	3.286	-1.348	-0.613	-0.772	-1.722	-0.041	1.059	-0.895

Mean of results from 1000 runs fitting correct model:

Bhat	3.340	-1.389	-0.604	-0.789	-1.927	-0.042	1.061	-0.915
mse0	0.337	0.115	0.095	0.081	0.595	0.000090	0.101	0.048
MSEL	0.321	0.108	0.090	0.078	0.566	0.000086	0.094	0.046
Binder	0.326	0.110	0.092	0.078	0.353	0.000087	0.097	0.046

Notice the difference between V_{db} (Binder) and MSEL in sales1000K

Mean of 1000 runs, unweighted (MLE fit):

Bhat	3.317	-1.383	-0.598	-0.778	-1.778	-0.041	1.052	-0.902
t	2.097	-3.746	1.769	-0.857	-4.738	-1.093	-0.919	-1.227

Conclusion: Sample size is insufficient even for MLE asymptotics!
Why should it be sufficient for any other asymptotic calculation?

Would we see a problem with one run? Data from first run:

Bhat	3.623	-1.874	-0.310	-1.095	-2.447	-0.038	1.104	-1.255
mse0	0.367	0.126	0.093	0.085	0.898	0.000094	0.108	0.053
MSEL	0.349	0.119	0.089	0.082	0.851	0.000090	0.101	0.050
Binder	0.327	0.170	0.108	0.098	0.402	0.000103	0.119	0.048

Ex (artificial data): Data generated according to the model

int	sales1K	sales5K	sales50K	sales1M	age	hisp	str10s
3.286	-1.348	-0.613	-0.772	-1.000	-0.041	1.059	-0.895

Mean of 1000 runs fitting correct model:

	int	sales1K	sales5K	sales50K	sales1M	age	hisp	str10s
Bhat	3.326	-1.373	-0.614	-0.785	-1.073	-0.041	1.062	-0.897
mse0	0.330	0.114	0.094	0.081	0.265	0.000088	0.098	0.046
MSEL	0.314	0.108	0.090	0.077	0.253	0.000084	0.092	0.044
Binder	0.316	0.109	0.092	0.078	0.221	0.000085	0.094	0.045

Mean of 1000 runs fitting incorrect model:

	sales5K	sales50K	sales1M	str10s	str20s	str30s	str40s
Bhat	-1.361	-0.688	-0.975	-0.172	0.529	1.325	0.505
mse0	0.055	0.073	0.255	0.036	0.042	0.889	0.052
MSEL	0.053	0.070	0.244	0.035	0.040	0.686	0.050
Binder	0.055	0.074	0.210	0.036	0.066	0.648	0.051

The first run:

	sales5K	sales50K	sales1M	str10s	str20s	str30s	str40s
Bhat	-1.413	-0.327	-1.090	-0.236	0.541	-0.048	0.479
mse0	0.056	0.066	0.202	0.035	0.041	0.619	0.051
MSEL	0.054	0.064	0.194	0.034	0.039	0.483	0.049
Binder	0.041	0.055	0.197	0.035	0.037	0.746	0.029

Question: Why is the difference between MSE_L and $\hat{V}_{db}$ in the variable `str20s`?

Hypothesis: Hispanics tend to cluster in strata 21 and 27 and not in the others.

Fisher exact test: 2 x n_h table of farms

- **rows = Hispanic status**
- **columns = PSU's (segments)**
- **test is conditional on row and column totals**
- **H_0: row and column classification are independent**

stratum		p-value
11	>75% cultivated	0.107[1]
17	>75% cultivated: fruit & nut	0.653
19	>75% cultivated: vegetable	1.000
21	15-75% cultivated	0.00011
27	15-75% cultivated: fruit & nut	0.042
31	agri-urban: > 100 homes/sqmi	1.000
32	dense urban: > 100 homes/sqmi	no test[2]
41	<15% cultivated	0.078
45	<15% cultivated: public no-ag, desert	1.000

[1]SAS monte carlo estimate of Fisher exact p-value
[2]only 1 sampled PSU has farms

EXECUTIVE SUMMARY

- **Discrepancy between MSE_0 and MSE_L indicates small cells (more general, multicolinearity).**
- **Discrepancy between $\hat{V}_{db}$ and MSE_L indicates model failure.**
- **Useful in model fitting in which many candidate models are considered and looking at individual data and cell statistics not practical. Especially important to avoid excess interaction terms which create instability.**

  Example: National AFS: 45991 farms, final model had 39 main effects and 3 two-way interactions.