24-1-2024

Διάγνωση Επιδραστικών Παρατηρήσεων (influential observations)

1) Κριτήριο leverage

h_i = leverage της παρατήρησης i (μόχλευση)

$$\frac{1}{n} \leq h_i \leq 1$$

Πίνακας κρίσιμων τιμών leverage

Stata: Μετά το regression

predict (name), leverage

---

2) Κριτήριο Cook's distance d_i

Κρίσιμες τιμές από πίνακες

Στον πίνακα A10

αν B είναι η κρίσιμη τιμή

τότε η παρατ - i είναι επιδραστική αν

$$d_i (n-k-1) > B \Rightarrow d_i > \frac{B}{n-k-1}$$

Stata : predict (name), cooksd

Αx αν $n=50$, $k=4$, $\alpha = 5\%$

από πίνακα : κρίσιμη τιμή $= 17{,}06$

$n-k-1 = 50-5 = 45$

Κριτήριο $45 d_i > 17.06 \Leftrightarrow d_i > \frac{17{,}06}{45} \simeq \underline{0{,}35}$

h_i, d_i : δεν είναι απαραίτητα συμβατά

Table A-9 Critical values for leverages, n = sample size, k = number of predictors

$\alpha = .10$

n \ k	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.626	0.759	0.847	0.911	0.956	0.984	0.997	1.000						
15	0.481	0.595	0.679	0.748	0.806	0.855	0.897	0.932	0.959	0.980				
20	0.394	0.491	0.565	0.627	0.682	0.731	0.775	0.815	0.851	0.883	0.988			
25	0.335	0.419	0.484	0.540	0.589	0.635	0.676	0.715	0.751	0.784	0.918	0.992		
30	0.293	0.366	0.424	0.474	0.519	0.560	0.599	0.635	0.669	0.701	0.837	0.937		
40	0.236	0.295	0.342	0.383	0.420	0.455	0.487	0.518	0.547	0.576	0.701	0.806		
60	0.172	0.214	0.248	0.279	0.306	0.332	0.356	0.380	0.402	0.424	0.524	0.612	0.888	
80	0.137	0.170	0.197	0.221	0.242	0.263	0.283	0.301	0.319	0.337	0.418	0.491	0.737	
100	0.114	0.141	0.164	0.183	0.201	0.219	0.235	0.250	0.266	0.280	0.348	0.410	0.625	0.941
200	0.064	0.079	0.091	0.102	0.111	0.121	0.130	0.138	0.146	0.155	0.192	0.227	0.353	0.568
400	0.036	0.043	0.050	0.055	0.060	0.065	0.070	0.075	0.079	0.083	0.104	0.122	0.190	0.311
800	0.020	0.024	0.027	0.030	0.032	0.035	0.037	0.040	0.042	0.044	0.055	0.065	0.100	0.164

$\alpha = .05$

αν $h_i > 0.268$ ⇒ i επιδραστική παρατήρηση

$k=3$

$n=60$

n \ k	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.683	0.802	0.879	0.933	0.969	0.990	0.999	1.000						
15	0.531	0.639	0.719	0.782	0.835	0.880	0.916	0.946	0.969	0.986				
20	0.436	0.531	0.602	0.662	0.714	0.761	0.802	0.839	0.872	0.901	0.991			
25	0.372	0.454	0.518	0.573	0.621	0.665	0.705	0.742	0.776	0.807	0.931	0.994		
30	0.325	0.398	0.455	0.505	0.549	0.589	0.627	0.662	0.695	0.726	0.855	0.947		
40	0.261	0.321	0.368	0.409	0.446	0.480	0.512	0.543	0.572	0.600	0.722	0.823		
60	0.190	0.233	0.268	0.298	0.326	0.352	0.376	0.400	0.422	0.444	0.543	0.630	0.898	
80	0.151	0.185	0.212	0.236	0.258	0.279	0.299	0.318	0.336	0.353	0.435	0.508	0.751	
100	0.126	0.154	0.176	0.196	0.215	0.232	0.248	0.264	0.279	0.294	0.363	0.425	0.638	0.946
200	0.070	0.085	0.098	0.108	0.119	0.128	0.137	0.146	0.154	0.162	0.201	0.236	0.362	0.570
400	0.039	0.047	0.053	0.059	0.064	0.069	0.074	0.079	0.083	0.088	0.108	0.127	0.196	0.317
800	0.021	0.025	0.029	0.032	0.034	0.037	0.039	0.042	0.044	0.046	0.057	0.067	0.103	0.168

$\alpha = .01$

n \ k	1	2	3	4	5	6	7	8	9	10	15	20	40	80
10	0.785	0.875	0.930	0.965	0.986	0.997	1.000	1.000						
15	0.629	0.724	0.792	0.844	0.887	0.921	0.948	0.969	0.984	0.994				
20	0.524	0.612	0.677	0.731	0.777	0.817	0.852	0.883	0.910	0.933	0.996			
25	0.450	0.529	0.589	0.640	0.685	0.724	0.761	0.794	0.824	0.851	0.953	0.997		
30	0.394	0.466	0.521	0.568	0.610	0.648	0.683	0.716	0.746	0.774	0.889	0.964		
40	0.318	0.377	0.424	0.464	0.501	0.534	0.565	0.595	0.622	0.649	0.763	0.855		
60	0.231	0.275	0.310	0.341	0.369	0.395	0.420	0.443	0.465	0.487	0.584	0.668	0.917	
80	0.183	0.218	0.246	0.271	0.293	0.314	0.334	0.353	0.372	0.389	0.471	0.543	0.778	
100	0.152	0.181	0.205	0.225	0.244	0.262	0.279	0.295	0.310	0.325	0.394	0.456	0.666	0.956
200	0.085	0.100	0.113	0.124	0.135	0.145	0.154	0.163	0.172	0.180	0.219	0.255	0.383	0.598
400	0.046	0.054	0.061	0.067	0.073	0.078	0.083	0.088	0.092	0.097	0.118	0.138	0.208	0.330
800	0.025	0.029	0.033	0.036	0.039	0.041	0.044	0.046	0.049	0.051	0.062	0.073	0.110	0.175

Table A-10 Critical values for the maximum of N values of Cook's $\boxed{d(i) \times (n-k-1)}$ (Bonferroni correction used) n observations and k predictors

!!

$\alpha = 0.1$

k	n=5	10	15	20	25	50	100	200	400	800
1	14.96	11.13	11.84	12.68	13.46	16.39	19.97	23.94	28.70	33.80
2	40.53	12.21	12.09	12.63	13.22	15.65	18.64	22.09	25.96	30.12
3		13.30	12.09	12.35	12.79	14.84	17.48	20.52	23.86	27.50
4		15.21	12.18	12.14	12.45	14.23	16.62	19.36	22.30	25.97
5		19.33	12.44	12.03	12.21	13.76	15.95	18.49	21.39	24.51
6		31.06	12.94	12.01	12.04	13.39	15.43	17.81	20.36	23.51
7		96.01	13.79	12.08	11.94	13.10	15.02	17.27	19.75	22.42
8			15.26	12.26	11.90	12.85	14.70	16.83	19.20	21.73
9			18.00	12.55	11.91	12.66	14.40	16.52	18.62	21.45
10			23.93	13.02	11.97	12.50	14.16	16.16	18.43	20.55
15				27.66	13.60	12.01	13.39	15.16	17.00	19.34
20					30.94	11.83	12.92	14.53	16.31	18.35
40						15.95	12.26	13.56	15.10	16.83
80							13.49	13.05	14.39	15.85

$\alpha = 0.05$

k	n=5	10	15	20	25	50	100	200	400	800
1	24.97	15.24	15.55	16.37	17.18	20.41	24.31	28.83	33.88	40.15
2	82.06	16.56	15.63	16.01	16.56	19.08	22.33	26.05	30.20	33.96
3		18.16	15.50	15.49	15.85	17.93	20.72	24.14	27.57	32.06
4		21.28	15.59	15.14	15.33	17.06	19.63	22.49	25.83	29.31
5		28.40	15.94	14.95	14.96	16.41	18.70	21.39	24.42	28.24
6		50.22	16.70	14.91	14.70	15.91	17.97	20.54	23.48	26.68
7		192.90	17.99	15.00	14.55	15.50	17.49	20.00	22.35	25.67
8			20.32	15.25	14.48	15.19	17.05	19.31	22.06	24.44
9			24.78	15.69	14.49	14.92	16.69	18.85	21.34	24.29
10			34.72	16.38	14.58	14.70	16.38	18.42	20.49	23.33
15				39.98	16.94	14.03	15.36	17.16	19.39	21.75
20					44.63	13.79	14.81	16.52	18.46	20.32
40						19.50	13.92	15.22	16.83	18.76
80							15.55	14.58	15.99	17.52

$\alpha = 0.01$

p	n=5	10	15	20	25	50	100	200	400	800
1	77.29	28.72	26.88	27.24	27.92	31.46	36.10	41.22	49.42	68.39
2	415.27	30.97	26.13	25.65	25.81	28.12	32.61	37.34	44.99	57.70
3		35.12	25.66	24.22	24.33	26.17	29.15	34.23	37.55	52.58
4		44.09	25.82	23.58	23.20	24.56	27.31	31.26	35.28	40.60
5		66.83	26.66	23.20	22.49	23.39	25.84	29.44	34.14	36.91
6		150.47	28.48	23.12	22.00	22.55	24.35	28.42	31.04	36.91
7		964.09	31.80	23.34	21.71	21.79	24.19	26.87	31.04	33.55
8			37.84	23.93	21.59	21.26	23.28	25.83	29.31	33.55
9			50.10	24.93	21.64	20.76	22.23	25.62	28.21	30.50
10			80.67	26.54	21.83	20.37	22.11	24.53	28.21	30.50
15				92.09	27.02	19.16	20.22	22.40	25.64	27.73
20					102.32	18.82	19.18	21.32	23.31	25.21
40						29.95	18.04	19.32	21.17	22.91
80							20.67	18.57	20.12	22.90

ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ

Βήματα επιλογής

1) Μέγιστο μοντέλο
2) Κριτήρια Σύγκρισης
3) Στρατηγική Επιλογής Μεταβλητών
4) Ανάλυση τελικού μοντέλου
5) Ερμηνείες - Προβλέψεις

① Μέγιστο Μοντέλο (Full Model)

Περιέχει όλες τις υποψήφιες μεταβλητές κ' όλους τους επιπλέον όρους (π.χ. μεγαλύτερης τάξης, αλληλεπιδράσεις κτλ).

$$Y = b_0 + b_1 X_1 + \dots + b_k X_k + \varepsilon$$

df_{er} "μεγάλο"

$df_{er} \geq 30$

πιο αυστηρός κανόνας $n \geq 5k$ ή $n \geq 10k$.

Συνήθως τα δημογραφικά στοιχεία περιλαμβάνονται στο μοντέλο (π.χ. ηλικία, ύψος, βάρος κτλ)

Αν περιλάβουμε interaction terms τότε οπωσδήποτε κ' τα main effects.

$$Y = b_0 + b_1 X_1 X_2 \qquad \underline{\text{όχι}}$$

$$Y_0 = b_0 + \underbrace{b_1 X_1 + b_2 X_2}_{\text{main effects}} + b_3 X_1 X_2$$

② Κριτήρια Σύγκρισης

Ⓐ Nested Models

$b_{p+1}X_{p+1} + \dots + b_k X_k$

(full) $Y = b_0 + b_1 X_1 + \dots + b_p X_p + \dots + b_k X_k$

(partial) $Y = b_0 + \underbrace{b_1 X_1 + \dots + b_p X_p}_{\text{υποσύνολο των αρχικών}}$ $\qquad (p<k)$

nested στο full model

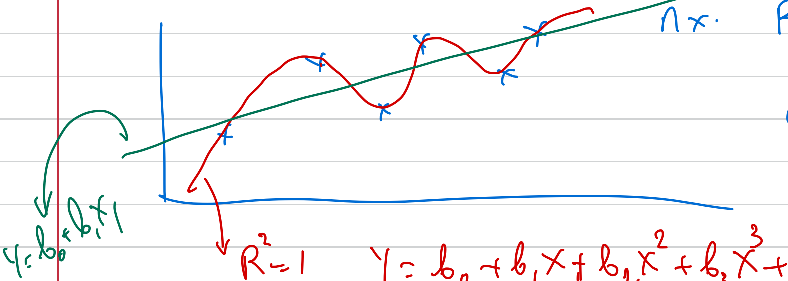
① F-test για part of model

$H_0: b_{p+1} = \dots = b_k = 0 \qquad H_1:$ τουλάχ. ένα $\neq 0$.

② Adjusted R^2

$$R^2 = \frac{SSR}{SST} = \%$$ μεταβ. της Y που εξηγείται από το μοντέλο.

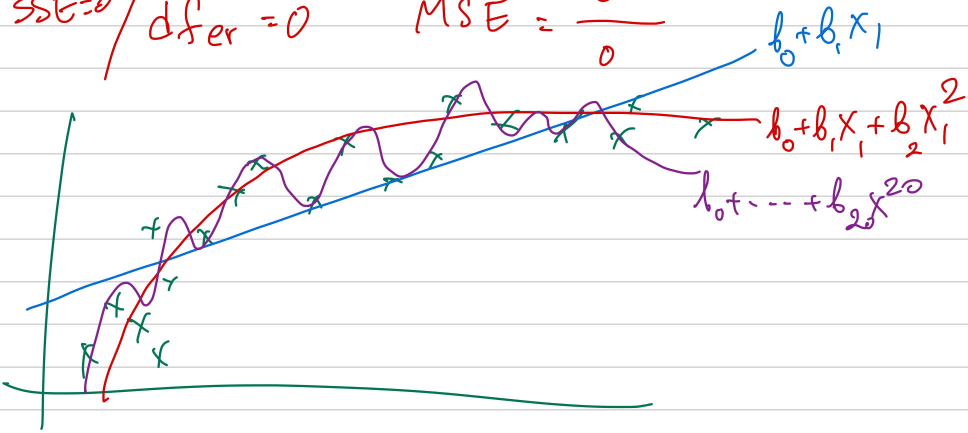
Σε nested models $R^2_{full} \ge R^2_{partial}$



π.χ. $R^2 \approx 1$ όσο [illegible] αυξάνοντας την τάξη

$R^2 = 1 \quad Y = b_0 + b_1 X + b_2 X^2 + b_3 X^3 + b_4 X^4 + b_5 X^5$

SSE=0 / $df_{er} = 0 \qquad MSE = \frac{0}{0}$



Adjusted R^2

$$adj\text{-}R^2 = 1 - (1 - R^2)\overbrace{\underbrace{\frac{n-1}{n-p-1}}_{df_{er}}}^{>1}$$

p = αρ. μεταβλ.

$p+1$ = αρ. παραμ.

$\le R^2$

Επίσης σαν $p \ll n$ δηλ. $df_{er} \approx n-1$

$adj\text{-}R^2 \approx R^2$

Αν adj-R^2 << R^2 $\Rightarrow$ ένδειξη overfitting

3) Στατιστικό C_p-Mallows

Full : (k)

Partial: (p) $\quad p<k$ (nested)

$$C_p = \frac{SSE(p)}{MSE(k)} - [n - 2(p+1)]$$

Αν Full model ~ Partial Model

$$MSE(p) \approx MSE(k)$$

Επίσης $MSE(p) = \dfrac{SSE(p)}{df_{er}(p)} = \dfrac{SSE(p)}{n-(p+1)} \approx MSE(k)$

Τότε $C_p \approx \dfrac{[n-(p+1)] \cdot \cancel{MSE(k)}}{\cancel{MSE(k)}} - [n - 2(p+1)] =$

$$= \cancel{n} - (p+1) - \cancel{n} + 2(p+1) = p+1$$

Όταν partial ≈ full $\Rightarrow$ $C_p \approx p+1$

Όταν partial υποδεέστερο του full

$MSE(p) > MSE(k) \Rightarrow C_p > p+1$

B) Γενικά (όχι nested)

π.χ. 1) $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_1^2$ $k=4$

2) $Y = b_0 + b_1 X_2 + b_2 X_3$ $k=3$

Κριτήρια Πληροφορίας

AIC = Akaike Information Criterion

BIC = Bayesian " "

$$AIC = 2k - 2\log L$$

L = πιθανοφάνεια δείγματος κάτω από τις LSE b.

$$\varepsilon \sim N(0, \sigma^2)$$

$$Y \sim N(b_0 + b_1 x, \sigma^2)$$

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\left(\boxed{y - (b_0 + b_1 x)}\right)^2}{2\sigma^2}}$$

r

$$L = \prod_{i=1}^{n} f(y_i) = (\quad) \cdot e^{-\frac{SSE}{2\sigma^2}}$$

$\max L \iff \min SSE$

$MLE \iff LSE$

"Καλό" μοντέλο $\iff$ AIC μικρό.

---

③ Μέθοδοι Επιλογής Μεταβλητών

Full model: $Y = b_0 + b_1 X_1 + \cdots + b_k X_k + \varepsilon$

Πόσα μοντέλα δημιουργούνται από το full υπάρχουν?

$\{X_1, X_2, \ldots, X_k\}$

Μοντέλο: υποσύνολο του

Πόσα υποσύνολα υπάρχουν?

$$\left.\begin{array}{lll} X_1 & : & \text{ναι / όχι} \\ X_2 & : & \text{ναι / όχι} \\ \vdots & & \\ X_k & : & \text{ναι / όχι} \end{array}\right\} 2^k$$

$Y = b_0 \quad \cdots\cdots \quad Y = \text{full}$

Αν $k = 10 \Rightarrow 2^k = 1024$

Μέθοδοι Stepwise για επιλογή μεταβλητών.

① Forward method ($X_1, \ldots, X_k$ υποψήφιες μεταβλητές)

ⓐ Όλα τα μονοπαραγοντικά μοντέλα

	stat	R
$Y = b_0 + b_1 X_1$	0.01 p-value	AIC
$Y = b_0 + b_1 X_2$	0.02 p	$\vdots$
$\vdots$		
$Y = b_0 + b_1 X_k$	p	AIC

$\downarrow$ t-test(b_1)

Επιλέγω το μικρότερο p-value (έστω π.χ. X_1)

Έχω επιλέξει μια παράμετρο P_{enter}

Αν το μικρότερο p-value $< P_{enter}$

η μεταβλητή X_1 μπαίνει στο μοντέλο

$Y = b_0 + b_1 X_1$

υπόλοιπες (εκτός)

$X_2 \leftarrow 0.7$

X_3

$\vdots$

X_k

$$Y = b_0 + b_1 X_1 + \boxed{b_2 X_2}$$

pvalue X_2 (t-test)

$$Y = b_0 + b_1X_1 + b_3X_3$$

p-v. X_3 ← min

$$\vdots$$

$$\underbrace{Y = b_0 + b_1X_1 + b_kX_k}_{k-1 \text{ μοντέλα}}$$

p-va. X_k

το μικρότερο (π.χ. X_3)

Αν το min p-value < Penter ⇒ X_3 μπαίνει

$$Y = b_0 + b_1X_1 + b_2X_3 \begin{cases} + b_3X_2 \\ + b_3X_4 \\ \vdots \\ + b_3X_k \end{cases}$$

επαναλαμβάνουμε μέχρι όλες οι εκτός μεταβλητές p_i > Penter

---

② Backward Method

Full model

$$Y = b_0 + b_1X_1 + \dots + b_kX_k$$

p_1 ... p_k : p-values t-test

Έστω p_1 = max των p-values

Παράμετρος P_{remove}

Αν max-pvalue $= p_1 > P_{remove}$ ⇒ X_1 αφαιρείται

Τώρα $Y = b_0 + b_1 X_2 + \dots + b_{k-1} X_k$

$b_1 X_2 \downarrow p$ … $b_{k-1} X_k \downarrow p \rightarrow \max.$

Επανάληψη μέχρι $\max p < P_{remove}$.

---

③ Stepwise

Σε κάθε βήμα

$Y = b_0 + b_1 X_1 + b_2 X_2$

forward step

① Εξετάζουμε τις μεταβλητές εκτός κι αποφασίζουμε αν θα μπει κάποια.

backward step

② Αν μπει μια νέα, εξετάζουμε αν πρέπει να αφαιρεθεί κάποια από τις άλλες που υπάρχουν ήδη μέσα

Προσοχή Αν θέσω $P_{enter} = 0.07$

$P_{rem} = 0.05$

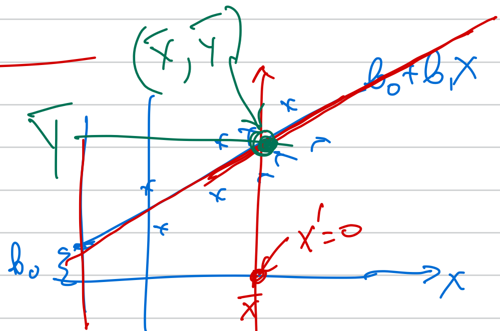
Μπορεί σε ένα βήμα X να έχει $p = 0.06$

$\Rightarrow$ μπαίνει $< P_{enter}$ $\Rightarrow$ αμέσως βγαίνει $> P_{remove}$

$P_{enter} < P_{remove}$

Προσοχή Μετά την stepwise χρειάζεται να ελέγξουμε το μοντέλο $\left\{ \begin{array}{l} \text{διαγνωστικά} \\ \text{ως προς τη φύση των μεταβλητών (εφαρμογή)} \end{array} \right.$

Κεντρικοποίηση



$$Y = b_0 + b_1 X.$$

$b_0 = E(Y \mid X=0)$ έχει νόημα όταν στο δείγμα $\exists$ παρατ. με $X=0$.

Έστω $\bar{X}$: μέση τιμή του X.

$$X' = X - \bar{X}$$

$Y = \gamma_0 + \gamma_1 X'$ ίδιο R^2, ίδιο p-value
όλα τα στατιστικά

$$Y = \gamma_0 + \gamma_1 (X - \bar{X}) = \underline{\gamma_0 - \gamma_1 \bar{X}} + \gamma_1 X$$

αρχικό $Y = \underline{b_0} + b_1 X$

$\gamma_1 = b_1$

$\hat{\gamma}_1 = \hat{b}_1$

$\hat{b}_0 = \hat{\gamma}_0 - \hat{\gamma}_1 \bar{X} \Rightarrow \hat{\gamma}_0 = \hat{b}_0 + \hat{\gamma}_1 \bar{X} = \hat{b}_0 + \hat{b}_1 \bar{X} = \bar{Y}$

$\hat{b}_0 = \bar{Y} - \hat{b}_1 \cdot \bar{X}$ (δηλ. $\bar{Y} = \hat{b}_0 + \hat{b}_1 \bar{X}$)

$(\bar{X}, \bar{Y})$ (κέντρο ο βάρους)

$Y = \bar{Y} + b_1 (X - \bar{X})$

$Y - \bar{Y} = b_1 (X - \bar{X})$ κεντρικοποιημένο