STICHTING
MATHEMATISCH CENTRUM
2e BOERHAAVESTRAAT 49
AMSTERDAM

S 218 (M 78)

Het toetsen van lineaire tegen kwadratische regressie

MATHEMATISCH CENTRUM,
2e Boerhaavestraat 49,
A m s t e r d a m - O.

Statistische Afdeling
S 218 (M 78)

Het toetsen van lineaire tegen kwadratische regressie[1)]

1.

Het model, dat wij in dit geval beschouwen is:

$$\underline{y}_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \underline{u}_i \,. \tag{1.1}$$

De onafhankelijke variabele, x, neemt de bekende waarden x_i $(i = 1, \ldots, N)$ aan.
De afhankelijke variabele $\underline{y}$[2)] is stochastisch. De waarde van $\underline{y}$ die bij x_i behoort is $\underline{y}_i$ en deze heeft een variantie σ_i^2.
De stochastische termen $\underline{u}_i$ zijn onderling onafhankelijk normaal verdeeld met verwachting 0 en variantie σ_i^2.
De toetsing van lineaire tegen kwadratische regressie komt neer op het toetsen van

$$H_0 : \beta_3 = 0 \,. \tag{1.2}$$

Om de berekeningen te vereenvoudigen, wijzigen wij de notatie enigszins, waarbij wij gebruik maken van de volgende definities:

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i \qquad (i = 1, \ldots, N) \,, \tag{1.3}$$

$$v_i = x_i - \bar{x} \,, \tag{1.4}$$

$$\overline{v^2} = \frac{1}{N} \sum_{i=1}^{N} v_i^2 \,, \tag{1.5}$$

$$z_i = v_i^2 - \overline{v^2} \,. \tag{1.6}$$

---

1) Dit memorandum is slechts bedoeld ter oriëntatie en streeft niet naar volledigheid of volledige exactheid.
2) Stochastische grootheden worden door onderstreepte letters aangeduid.

Nu is $\sum_{i=1}^{N} v_i = 0$ en $\sum_{i=1}^{N} z_i = 0$ in verband met (1.3) en (1.5).
Het model (1.1) wordt nu in deze notatie geschreven als:

(1.1') $$\underline{y}_i = \alpha_1 + \alpha_2 v_i + \alpha_3 z_i + \underline{u}_i \ ,$$

waarbij de relaties tussen de coëfficiënten gegeven worden door:

(1.7) $$\beta_1 = \alpha_1 - \alpha_2 \bar{x} - \alpha_3 (\overline{x^2} - \bar{x}^2) \ ,$$

(1.8) $$\beta_2 = \alpha_2 - 2\alpha_3 \bar{x} \ ,$$

(1.9) $$\beta_3 = \alpha_3 \ .$$

Er kunnen zich nu 4 gevallen voordoen:

a. Bij iedere bekende waarde x_i hebben wij één waarneming van $\underline{y}_i$; de varianties σ_i^2 voor alle $\underline{y}_i$ zijn gelijk.

b. Bij iedere bekende waarde x_i hebben wij één waarneming van y_i ; de varianties σ_i^2 zijn ongelijk, maar hun verhoudingen zijn bekend.

c. Bij iedere bekende waarde x_i zijn n_i waarnemingen van $\underline{y}_i$ verricht. De varianties σ_i^2 zijn gelijk.

d. Bij iedere bekende waarde x_i zijn n_i waarnemingen van $\underline{y}_i$ verricht; de varianties σ_i^2 zijn ongelijk, maar hun verhoudingen zijn bekend.

Wij lichten de toets toe aan het eenvoudigste voorbeeld, geval a.

Volgens de methode der kleinste kwadraten zijn door minimaliseren van de vorm:

(1.10) $$Q = \sum_{i=1}^{N} (y_i - \alpha_1 - \alpha_2 v_i - \alpha_3 z_i)^2$$

de schattingen a_1 , a_2 en a_3 van α_1 , α_2 en α_3 te verkrijgen en wel de volgende:

(1.11) $$a_1 = \bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i \ ,$$

(1.12) $$a_2 = \frac{S_{12} S_{33} - S_{13} S_{23}}{S_{22} S_{33} - S_{23}^2} \ ,$$

(1.13) $$a_3 = \frac{S_{13}S_{22} - S_{12}S_{23}}{S_{22}S_{33} - S_{23}^2} .$$

Hierin is:

(1.14)
$$\begin{cases} S_{11} = \sum_{i=1}^{N} (y_i - \bar{y})^2 , \\ S_{12} = \sum_{i=1}^{N} y_i v_i , \\ S_{13} = \sum_{i=1}^{N} y_i z_i , \\ S_{22} = \sum_{i=1}^{N} v_i^2 , \\ S_{23} = \sum_{i=1}^{N} v_i z_i , \\ S_{33} = \sum_{i=1}^{N} z_i^2 . \end{cases}$$

Wij toetsen nu de hypothese $H_0 : \alpha_3 = 0$. De toetsingsgrootheid wordt berekend volgens onderstaande formule:

(1.15) $$F = \frac{(N-3)\, a_3^2 \,(S_{22}S_{33} - S_{23})^2}{(S_{11} - S_{12}a_2 - S_{13}a_3)S_{22}} .$$

Deze bezit, onder H_0, een F-verdeling met 1 en $N-3$ vrijheidsgraden.
Wij kunnen ook $t, = \sqrt{F}$ berekenen, deze bezit een Student-verdeling met $N-3$ vrijheidsgraden. De hypothese H_0 wordt verworpen als de eenzijdige (resp. tweezijdige) overschrijdingskans van de gevonden waarde van F (resp. $t,$) kleiner is dan de van tevoren gekozen onbetrouwbaarheidsdrempel α.
Is deze overschrijdingskans $> \alpha$, dan wordt de hypothese: $H_0 : \alpha_3 = 0$, niet verworpen.

Vervolgens beschouwen wij het algemene geval d. Bij eenzelfde waarde x_i zijn n_i waarnemingen van $\underline{y}_i$ verricht en voor de varianties binnen de groepen geldt:

(1.16) $$\sigma_i^2 = \sigma^2\{\underline{y}_i\} = d_i \sigma^2$$

met onbekende σ^2 , maar bekende d_i ($i = 1, \ldots, k$; $j = 1, \ldots, n_i$; $\sum_{i=1}^{k} n_i = N$).

Het model (1.1) kan nu geschreven worden als:

(1.1") $$\underline{y}_{ij} = \alpha_1 + \alpha_2 v_i + \alpha_3 z_i + \underline{u}_{ij} .$$

Wij voeren nu de volgende gewichten in:

(1.17) $$g_i = \frac{n_i}{d_i}$$

en definiëren

(1.18) $$g = \sum_{i=1}^{k} g_i ,$$

(1.19) $$\bar{x} = g^{-1} \sum_{i=1}^{k} g_i x_i .$$

Vervolgens berekenen wij

(1.20) $$v_i = x_i - \bar{x} ,$$

zodat

(1.21) $$\bar{v} = g^{-1} \sum_{i=1}^{k} g_i v_i = 0$$

en

(1.22) $$z_i = v_i^2 - \overline{v^2} = v_i^2 - g^{-1} \sum_{i=1}^{k} g_i v_i^2 ,$$

zodat ook

(1.23) $$\bar{z} = g^{-1} \sum_{i=1}^{k} g_i z_i = 0 .$$

Verder berekenen wij ook:

(1.24) $$y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

en

(1.25) $$\bar{y} = g^{-1} \sum_{i=1}^{k} g_i y_{i.}$$

De vergelijkingen (1.14) gaan nu over in

(1.14') $$\begin{cases} S_{11} = \sum_{i,j} \frac{1}{d_i} (y_{ij} - \bar{y})^2 = \sum_{i,j} \frac{1}{d_i} (y_{ij} - y_{i.})^2 + \sum_{i=1}^{k} g_i (y_{i.} - \bar{y})^2, \\ S_{12} = \sum_{i=1}^{k} g_i y_{i.} v_i , \\ S_{13} = \sum_{i=1}^{k} g_i y_{i.} z_i , \\ S_{22} = \sum_{i=1}^{k} g_i v_i^2 , \\ S_{23} = \sum_{i=1}^{k} g_i v_i z_i , \\ S_{33} = \sum_{i=1}^{k} g_i z_i^2 . \end{cases}$$

De formule voor de toetsingsgrootheid (1.15) en het toetsen van de hypothese $H_0 : \alpha_j = 0$ blijven onveranderd.
Zijn de varianties binnen de groepen constant, dan worden de gewichten in het bovenstaande geval vereenvoudigd tot (geval c):

(1.17') $$g_i = n_i$$

Verder is:

(1.24) $$y_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$$ (1.25') $$\bar{y} = \frac{1}{N} \sum_{i,j} y_{ij} = \frac{1}{N} \sum_{i=1}^{k} n_i y_{i.}$$

en

(1.14'') $$\begin{cases} S_{11} = \sum_{i,j} (y_{ij} - \bar{y})^2 = \sum_{i,j} y_{ij}^2 - \frac{(\sum y_{ij})^2}{N}, \\ S_{12} = \sum_{i=1}^{k} n_i y_{i.} v_i , \\ S_{13} = \sum_{i=1}^{k} n_i y_{i.} z_i , \\ S_{22} = \sum_{i=1}^{k} n_i v_i^2 , \\ S_{23} = \sum_{i=1}^{k} n_i v_i z_i , \\ S_{33} = \sum_{i=1}^{k} n_i z_i^2 . \end{cases}$$

Is nu van iedere x_i één waarneming y_i bekend en geldt voor de varianties (1.16) (geval b), dan krijgen wij de volgende formules:

(1.17") $$g_i = \frac{1}{d_i},$$

(1.25") $$\bar{y} = g^{-1} \sum_{i=1}^{N} g_i y_i,$$

(1.14''') $$\begin{cases} S_{11} = \sum_{i=1}^{N} \frac{1}{d_i} (y_i - \bar{y})^2, \\ S_{12} = \sum_{i=1}^{N} g_i y_i v_i, \\ S_{13} = \sum_{i=1}^{N} g_i y_i z_i, \\ S_{22} = \sum_{i=1}^{N} g_i v_i^2, \\ S_{23} = \sum_{i=1}^{N} g_i v_i z_i, \\ S_{33} = \sum_{i=1}^{N} g_i z_i^2. \end{cases}$$

De formule voor de toetsingsgrootheid (1.15) blijft onveranderd.

2.

De regressielijnen kunnen getekend worden door horizontaal v uit te zetten en bij iedere v_i verticaal:

1. De gevonden waarden y_i. ,
2. $Y_i = a_1 + a_2 v_i + a_3 z_i$,
3. $Y_i^{(1)} = a_1 + a_2^{(1)} v_i$,

} de "regressiewaarden" voor het kwadratische en het lineaire model.

waarin de schatting $a_2^{(1)}$ van α_2 in het lineaire model gegeven wordt door:

(2.1) $$a_2^{(1)} = \frac{S_{12}}{S_{22}} = a_2 + \frac{S_{23}}{S_{22}} a_3 \ .$$ (laatste lid ter controle).

De toetsingsgrootheid F wordt voor gelijke varianties binnen de groepen (σ_i^2) eveneens gegeven door de onderstaande formule:

(2.2) $$F = \frac{(N-3)\sum_{i=1}^{k} n_i (Y_i^{(1)} - Y_i)^2}{\sum_{i=1}^{k} n_i (y_{i.} - Y_i)^2 + \sum_{ij} (y_{ij} - y_{i.})^2}$$

en indien voor de varianties σ_i^2 geldt:

(1.16) $$\sigma_i^2 = d_i \sigma^2 \quad \text{en (1.17)} \quad g_i = \frac{n_i}{d_i}$$

door:

(2.3) $$F = \frac{(N-3)\sum_{i=1}^{k} g_i (Y_i^{(1)} - Y_i)^2}{\sum_{i=1}^{k} g_i (y_{i.} - Y_i)^2 + \sum_{ij} \frac{1}{d_i} (y_{ij} - y_{i.})^2} \ .$$

Voor kleine k is deze wijze van berekenen sneller dan volgens formule (1.14') of (1.14") en (1.15).

De uitkomsten volgens beide methoden moeten gelijk zijn. Wij kunnen dus eventueel een van de beide berekeningen als controle-berekening gebruiken.

Literatuur over een speciaal geval van deze toets kan men vinden in G. TINTNER.

Literatuur

1 FISHER, R.A. en F. YATES, Statistical Tables for Biological, Agricultural and Medical Research, 3rd ed., Oliver and Boyd, London 1948.

2 HALD, A., Statistical Tables and Formulas, J. Wiley and Sons, New York, 1952.

3 KENDALL, M.G., The advanced theory of statistics, Vol. II, 2nd ed., Griffin, London 1948, p. 145 e.v.

4 TINTNER, G., Econometrics, J. Wiley and Sons, New York, 1952, p. 83-92.

Tabel 4.

Tweezijdige overschrijdingskansen van de toetsen van Student.

ν	k = .9	.8	.7	.6	.5	.4	.3	.2	.1	.05	.02	.01
1	.158	.325	.510	.727	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	.142	.289	.445	.617	.816	1.061	1,386	1.886	2.920	4.303	6.965	9.925
3	.137	.277	424	584	.765	.978	1.250	1.638	2.353	3.182	4.541	5.841
4	.134	.271	.414	.569	.741	.941	1.190	1.533	2.132	2.776	3.747	4.604
5	.132	.267	.408	.559	.727	.920	1.156	1.476	2.015	2.571	3.365	4.032
6	.131	.265	.404	.553	.718	.906	1.134	1.440	1.943	2.447	3.143	3.707
7	.130	.263	.402	.549	.711	.896	1.119	1.415	1.895	2.365	2.998	3.499
8	.130	.262	.399	.546	.706	.889	1.108	1.397	1.860	2.306	2.896	3.355
9	.129	.261	.398	.543	.703	.883	1.100	1.383	1.833	2.262	2.821	3.250
10	.129	.260	.397	.542	.700	.879	1.093	1.372	1.812	2.228	2.764	3.169
11	.129	.260	.396	.540	.697	.876	1.088	1.363	1.796	2.201	2.718	3.106
12	.128	.259	.395	.539	.695	.873	1.083	1.356	1.782	2.179	2.681	3.055
13	.128	.259	.394	.538	.694	.870	1.079	1.350	1.771	2.160	2.650	3.012
14	.128	.258	.393	.537	.692	.868	1.076	1.345	1.761	2.145	2.624	2.977
15	.128	.258	.393	.536	.691	.866	1.074	1.341	1.753	2.131	2.602	2.947
16	.128	.258	.392	.535	.690	.865	1.071	1.337	1.746	2.120	2.583	2.921
17	.128	.257	.392	.534	.689	.863	1.069	1.333	1.740	2.110	2.567	2.898
18	.127	.257	.392	.534	.688	.862	1.067	1.330	1.734	2.101	2.552	2.878
19	.127	.257	.391	.533	.688	.861	1.066	1.328	1.729	2.093	2.539	2.861
20	.127	.257	.391	.533	.687	.860	1.064	1.325	1.725	2.086	2.528	2.845
21	.127	.257	.391	.532	.686	.859	1.063	1.323	1.721	2.080	2.518	2.831
22	.127	.256	.390	.532	.686	.858	1.061	1.321	1.717	2.074	2.508	2.819
23	.127	.256	.390	.532	.685	.858	1.060	1.319	1.714	2.069	2.500	2.807
24	.127	.256	.390	.531	.685	.857	1.059	1.318	1.711	2.064	2.492	2.797
25	.127	.256	.390	.531	.684	.856	1.058	1.316	1.708	2.060	2.485	2.787
26	.127	.256	.390	.531	.684	.856	1.058	1.315	1.706	2.056	2.479	2.779
27	.127	.256	.389	.531	.684	.855	1.057	1.314	1.703	2.052	2.473	2.771
28	.127	.256	.389	.530	.683	.855	1.056	1.313	1.701	2.048	2.467	2.763
29	.127	.256	.389	.530	.683	.854	1.055	1.311	1.699	2.045	2.462	2.756
30	.127	.256	.389	.530	.683	.854	1.055	1.310	1.697	2.042	2.457	2.750
∞	.12566	.25335	.38532	.52440	.67449	.84162	1.03643	1.28155	1.64485	1.95996	2.32634	2.57582