C. Silva Ribeiro *
J. M. Rolo **

Análise Social, vol. XIII (50), 1977-2.º, 461-475

A tecnologia estrangeira do sector farmacêutico português: um ensaio econométrico ***

Este artigo é produto de um estudo a que foram sujeitos 91 contratos de tecnologia estrangeira em vigor em algumas empresas do sector da indústria farmacêutica portuguesa.

Tais contratos foram inicialmente objecto de um primeiro tratamento, através de um inquérito, que deu lugar à obtenção de um volume apreciável de *informação empírica de base*, cujas possibilidades de tratamento mais elaborado estão a ser exploradas.

Boa parte dessa informação empírica foi já possível reduzi-la a *informação codificada*, susceptível de tratamentos mais sofisticados.

O presente trabalho é uma primeira tentativa neste sentido, que, sendo portadora de naturais limitações, nem por isso deixa de constituir, a exemplo doutros trabalhos do âmbito de um projecto de investigação mais vasto em curso no Gabinete de Investigações Sociais, um estudo pioneiro no domínio das transferências de tecnologia em Portugal, domínio que, por diversas razões, se desejaria objecto de maior atenção a vários níveis, sobretudo por parte das instituições do Estado mais directamente ligadas ao assunto [1].

1. DESCRIÇÃO DAS VARIÁVEIS

A informação de base, depois de devidamente codificada, em função não só da sua natureza, como também das exigências dos próprios modelos, forneceu resultados úteis para as seguintes variáveis:

Y_1 — Duração do contrato (anos).
Y_2 — *Royalties* (percentagem sobre as vendas).

* Assistente do Instituto Superior de Economia da Universidade Técnica de Lisboa; responsável pela cadeira de Programação Matemática do Curso de Pós-Graduação em Métodos Matemáticos do mesmo Instituto.

** Assistente do Instituto Superior de Economia da Universidade Técnica de Lisboa; investigador do Gabinete de Investigações Sociais.

*** O presente trabalho é mais um produto de um projecto subordinado ao tema «Investimento estrangeiro e tecnologia em Portugal», em curso no Gabinete de Investigações Sociais, que, desta vez, contou com a colaboração de C. Silva Ribeiro, a quem coube a tarefa do tratamento econométrico de boa parte da informação empírica de base já disponível.

Para melhor entendimento deste ensaio consulte-se J. M. Rolo, «Modalidades de tecnologia importada em Portugal», in *Análise Social*, n.º 47, 1976, pp. 541-561.

[1] Referimo-nos, entre outras, sobretudo ao Banco de Portugal.

X_1 — Antiguidade do contrato (anos).
X_2 — Participação no capital da empresa compradora de tecnologia (sim ou não).

Z — País da empresa vendedora de tecnologia (sim ou não):

Z_1 — Suíça.
Z_2 — Bélgica.
Z_3 — Itália.
Z_4 — Holanda.
Z_5 — Reino Unido.
Z_6 — França.
Z_7 — E. U. A.
Z_8 — R. F. A.
Z_9 — Luxemburgo.
Z_{10} — Outros.

W — Modalidades de tecnologia importada (sim ou não) [2]:

W_1 — Licenças de exploração de patentes (ETN).
W_2 — Conhecimentos técnicos (ETN).
W_3 — Marcas (ETN).
W_4 — I & D (ETS).
W_5 — Representação comercial (ETP).
W_6 — Estudos de viabilidade (ETP).
W_7 — Formação e troca de pessoal (ETS).
W_8 — Bens de capital (ETN).
W_9 — Manutenção de equipamentos (ETS).
W_{10} — Serviços de avaliação e controlo (ETS).
W_{11} — Aperfeiçoamento de processos produtivos (ETS).
W_{12} — Outras.

V — Cláusulas restritivas (sim ou não)[3]:

V_1 — Exportação só para colónias (CRE).
V_2 — Proibição total de exportar (CRE).
V_3 — Exportação com autorização prévia (CRE).
V_4 — Outras cláusulas restritivas das exportações (CRE).
V_5 — Imposição de regras severas ao exercício da propriedade industrial.
V_6 — Matérias-primas e bens intermediários.
V_7 — Segredo.
V_8 — Política de vendas.
V_9 — Controlo de qualidade.
V_{10} — Volumes de produção.
V_{11} — Processos de produção.
V_{12} — Política de preços.
V_{13} — Obrigações face a terceiros.
V_{14} — Outras.

---

[2] ETN, elementos tecnológicos nucleares; ETS, elementos tecnológicos de suporte; ETP, elementos tecnológicos prospectivos (cf. artigo citado na nota *** supra).
[3] CRE, cláusulas restritivas das exportações.

Os valores assumidos por estas variáveis em cada um dos contratos são, no seu conjunto, uma matriz *(cross section)* de dados, que constitui o suporte sobre o qual operam ordenadamente os modelos de regressão linear adoptados.

2. DESCRIÇÃO DOS MODELOS UTILIZADOS

2.1 Com o objectivo de explicar algumas relações (globais) entre as variáveis descritas que nos parecia importante conhecer à partida, começámos por construir *tabelas de contingência,* a fim de verificar se existia «associação» entre os seguintes pares de atributos:

Países e modalidades de tecnologia importada;
Países e cláusulas restritivas das exportações;
Países e outras cláusulas restritivas;
Modalidades de tecnologia importada e cláusulas restritivas das exportações;
Modalidades de tecnologia importada e outras cláusulas restritivas.

Como se verá adiante, os respectivos ensaios do χ^2 (qui-quadrado) mostram que não existe associação (global) entre os contributos confrontados, o que, como veremos, não impede que existam outras relações (parcelares) cujo estudo reclama a utilização de modelos de regressão linear.

2.2 Numa segunda via de análise elaboraram-se numa primeira fase, vários *modelos de regressão linear múltipla,* considerando como *variáveis a explicar:*

A duração do contrato (Y_1);
As royalties (Y_2);
As modalidades de tecnologia importada (W);
As cláusulas restritivas (V).

As variáveis explicativas foram:

A antiguidade do contrato (X_1);
A participação no capital da empresa compradora de tecnologia (X_2);
Os países (Z).

Nesta sede surgiram-nos diversos problemas resultantes, nomeadamente, do facto de as variáveis W e V serem *variáveis artificiais (dummy variables).* Como se sabe, os valores calculados para estas variáveis podem ser interpretados como estimativas das probabilidades condicionadas pelos valores assumidos pelas variáveis independentes, podendo acontecer aqueles valores calculados não pertencerem ao intervalo [0,1]. Põe-se então o problema econométrico das *variáveis dependentes limitadas,* existindo diversos modelos para resolver esta questão [4]. Em relação aos casos em que se nos puseram problemas de limitação de variáveis, isso ficou assinalado no quadro das conclusões (cf. quadro n.º 6).

[4] Cf. «Apêndice».

Outra dificuldade importante reside no facto de a hipótese de homoescedasticidade dos resíduos não se manter. Daqui resulta que os estimadores dos mínimos quadrados dos coeficientes de regressão são enviesados [5].

Finalmente, o coeficiente de determinação perde muito da sua importância como medida do ajustamento proposto [6].

A leitura do quadro n.º 6 deverá ter em conta todas estas observações.

Nesta primeira fase, em relação aos regressandos (variáveis a explicar) não artificiais Y_1 (duração do contrato) e Y_2 *(royalties)*, obtiveram-se resultados relativamente modestos (cf. quadro n.º 6).

2.3 Com o intuito de por um lado, incrementar o grau de explicação destas duas variáveis e, por outro, verificar se as variáveis Y_1 e Y_2 são interdependentes, construíram-se, numa segunda fase, quatro modelos de regressão linear múltipla (cf. quadro n.º 7): nos dois primeiros, o regressando foi Y_1 e os regressores Y_2, as modalidades de tecnologia importada (W) ou as cláusulas restritivas (V); nos outros dois, as variáveis Y_1 e Y_2 trocaram os papéis, passando assim Y_2 a regressando e Y_1 a regressor.

De uma maneira geral, conclui-se que não há interdependência entre Y_1 e Y_2 (não há lugar a estimação simultânea), mas verifica-se que outros regressores além dos países contribuem, de algum modo, para a explicação do comportamento de Y_1 ou Y_2.

2.4 A terceira fase consistiu na elaboração de dois modelos de regressão linear múltipla, um para Y_2, outro para Y_1, como variáveis a explicar. As variáveis explicativas, contudo, foram seleccionadas entre aquelas que apresentaram melhor comportamento na 1.ª e 2.ª fases (cf. quadro n.º 8). Este processo de «decantação» contribuiu para melhorar consideravelmente a explicação de Y_1 e Y_2.

3. RESULTADOS

3.1 ANÁLISE GLOBAL: TABELAS DE CONTINGÊNCIA E χ^2 (QUI-QUADRADO) [7]

Na construção das tabelas de contingência agruparam-se, por vezes, certas modalidades de um atributo de forma que as frequências absolutas observadas em cada «célula» sejam, pelo menos, 3 ou 4, para que os ensaios do χ^2 tenham significado.

---

[5] Id.

[6] Id.

[7] Representando por F^o_{ij} ($i=1, 2, \ldots, m$; $j=1, 2, \ldots, n$) as frequências observadas e por F^c_{ij} as frequências calculadas, seja

$$F_{i\cdot} = \sum_{j=1}^{n} F_{ij} \quad \text{e} \quad F_{\cdot j} = \sum_{i=1}^{m} F_{ij}$$

respectivamente a soma das linhas e das colunas da tabela de contigência. Tem-se

$$F^c_{ij} = \frac{F_{i\cdot} \times F_{\cdot j}}{N}$$

Nos quadros apresentados, as frequências observadas aparecem no canto superior esquerdo e as frequências calculadas no canto inferior direito.

Associação entre países e modalidades de tecnologia importada

[QUADRO N.º 1]

W	Z_1	Z_2	Z_3	Z_6	Z_7	Z_8	Outros	Total
W_1	4 6,16	13 10,84	7 6,17	15 14,78	5 7,88	9 10,35	15 12,81	68
W_2	9 7,79	17 13,71	7 6,54	19 18,7	7 9,97	11 13,09	16 16,20	86
W_3	9 5,98	7 10,52	4 5,02	17 14,35	6 7,65	11 10,04	12 12,43	66
Outras ...	3 5,07	7 8,93	3 4,26	9 12,17	14 6,49	11 8,52	9 10,55	56
Total ...	25	44	21	60	32	42	52	276

Número de graus de liberdade: 18.

$\chi^2(18)$: 10,29
$\chi^2_{0,05}(18)$: 28,87
$\chi^2_{0,01}(18)$: 34,81

Conclusão: O ensaio aceita a hipótese de independência entre os dois atributos a um nível de significação quer de 1 % quer de 5 %.

Associação entre países e cláusulas restritivas à exportação

[QUADRO N.º 2]

V	Z_1	Z_2	Z_6	Z_8	Outros	Total
V_1	6 6,81	16 11,58	12 12,95	7 8,18	21 22,48	62
Outras	4 3,19	1 5,42	7 6,05	5 3,82	12 10,52	29
Total	10	17	19	12	33	91

onde N é o número total de observações. Prova-se que

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(F^o_{ij} - F^c_{ij})^2}{F^c_{ij}} \cap \chi^2\,[(m\text{-}1).(n\text{-}1)]$$

Com um nível de significação de α %, se $\chi^2 < \chi^2_\alpha$ (valor dado pela tabela), a hipótese de independência é aceite; se $\chi^2 > \chi^2_\alpha$, a hipótese é rejeitada, existe associação. Neste caso, o grau de associação é medido pelo coeficiente de contingência $\sqrt{\frac{\chi^2}{\chi^2+N}}$

Número de graus de liberdade: 4.

$\chi^2(4)$: 5,27
$\chi^2_{0,05}(4)$: 9,49
$\chi^2_{0,01}(4)$: 13,28

Conclusão: O ensaio aceita a hipótese de independência entre os dois atributos a um nível de significação quer de 1 % quer de 5 %.

Associação entre países e outras cláusulas restritivas

[QUADRO N.º 3]

V	Z_1	Z_2	Z_6	Z_7	Z_8	Outros	Total
V_5	8 6,97	16 16,16	14 15,82	7 7,99	9 11,40	20 15,65	74
V_6	7 5,75	12 13,32	13 13,04	6 6,59	10 9,39	13 12,9	61
V_7	4 5,47	16 12,67	11 12,4	6 6,27	11 8,93	10 12,27	58
V_8	6 5,56	17 12,88	12 12,61	4 6,37	9 9,09	11 12,48	59
V_9	4 4,43	9 10,26	10 10,05	6 5,08	8 7,24	10 9,94	47
V_{10}	4 3,86	6 8,95	11 8,76	4 4,43	7 6,31	9 8,67	41
V_{11}	4 2,83	5 6,55	6 6,41	5 3,24	4 4,62	6 6,34	30
Outras ...	4 6,13	14 14,19	16 13,9	9 7,02	9 10,01	13 13,75	65
Total	41	95	93	47	67	92	435

Número de graus de liberdade: 35.

$\chi^2(35)$: 9,54
$\chi^2_{0,05}(35)$: 49,80
$\chi^2_{0,01}(35)$: 57,34

Conclusão: O ensaio aceita a hipótese de independência entre os dois atributos a um nível de significação quer de 1 % quer de 5 %.

Associação entre modalidades de tecnologia importada e cláusulas restritivas das exportações

[QUADRO N.º 4]

V	W_1	W_2	W_3	W_6	W_7	Outras	Total
V_1	54 46,85	64 42,23	48 46,19	11 13,86	4 7,26	13 20,46	194
V_2	11 14	15 17,76	12 13,81	5 4,14	4 2,17	11 6,12	58
Outras ...	6 10,14	11 12,86	10 10	5 3	3 1,57	7 4,43	42
Total... ...	71	90	70	21	11	31	294

Número de graus de liberdade: 10.

$\chi^2(10)$: 6,79
$\chi^2_{0,05}(10)$: 18,31
$\chi^2_{0,01}(10)$: 23,21

Conclusão: O ensaio aceita a hipótese de independência entre os dois atributos a um nível de significação quer de 1 % quer de 5 %.

Associação entre modalidades de tecnologia importada e outras cláusulas restritivas

[QUADRO N.º 5]

V	W_1	W_2	W_3	W_6	W_7	Outras	Total
V_5	57 54,91	69 68,26	55 51,57	13 15,36	9 11,52	31 32,38	234
V_6	44 43,88	58 54,55	45 41,21	10 12,27	8 9,2	22 25,86	187
V_7	46 40,6	53 50,47	22 38,13	14 11,35	10 8,51	28 23,94	173
V_8	46 42,24	55 52,51	40 39,67	12 11,81	6 8,86	21 24,91	180
V_9	33 36,61	44 45,51	42 34,38	10 10,24	7 7,68	20 21,59	156
V_{10}	29 30,51	38 37,92	32 28,65	9 8,53	7 6,4	15 17,99	130
V_{11}	22 23,7	28 29,46	24 22,26	7 6,63	6 4,97	14 13,98	101
V_{12}	17 20,18	25 25,09	20 18,95	6 5,64	3 4,23	15 11,9	86
V_{13}	13 15,72	16 19,55	15 14,77	5 4,4	5 3,3	13 9,27	67
V_{14}	22 20,65	23 25,67	14 19,39	6 5,77	8 4,33	15 12,18	88
Total... ...	329	409	309	92	69	194	1 402

Número de graus de liberdade: 45.

$\chi^2(45)$: 14,6
$\chi^2_{0,05}(45)$: 61,66
$\chi^2_{0,01}(45)$: 69,96

Conclusão: O ensaio aceita a hipótese de independência entre os dois atributos a um nível de significação quer de 1 % quer de 5 %.

Conclusão geral

As hipóteses de independência dos pares de atributos referidos são aceites, isto é, existe independência entre os atributos referidos. Assim, por

exemplo, as classificações dos contratos por países e por modalidades de tecnologia importada são independentes: globalmente, isto significa que qualquer modalidade de tecnologia pode ser importada de qualquer país e qualquer país pode exportar para Portugal qualquer modalidade de tecnologia. O mesmo se passa com os outros quatro pares de atributos.

Contudo, pode muito bem acontecer que uma dada modalidade de tecnologia importada, quando tomada isoladamente, possa depender significativamente de um ou vários países. Neste caso, estamos em presença do ensaio da seguinte hipótese: a variável binária que representa a modalidade de tecnologia importada em causa tem como regressores certos países, também variáveis binárias. É esta análise parcelar que se passa a fazer.

3.2 ANÁLISE PARCELAR: OS MODELOS DE REGRESSÃO LINEAR

3.2.1 No quadro n.º 6 apresenta-se para cada modelo (a cada linha corresponde um modelo) a respectiva variável dependente, as estimativas dos coeficientes de regressão, bem como os respectivos desvios-padrão (entre-parêntesis), o R^2, o valor da estatística F para todas as variáveis e o valor da estatística F para as variáveis Z (países). Na coluna das «Observações» refere-se a dimensão da amostra que esteve na base da respectiva regressão e os casos em que pode surgir o problema da limitação de variáveis. Quanto aos ensaios de hipóteses, estabeleceram-se as seguintes convenções:

Hipótese de *um* coeficiente ser nulo:

Ensaio do t — Student a 5 %: as variáveis aceites têm o respectivo coeficiente sublinhado a traço contínuo.
Ensaio do t — Student a 10 %: as variáveis aceites têm o respectivo coeficiente sublinhado a tracejado.

Hipótese de *todos* os coeficientes ou de os coeficientes relativos aos *países* serem nulos:

Ensaio do F a 10 %: no caso de a hipótese ser rejeitada, o valor do F é sublinhado a traço contínuo.
Ensaio do F a 5 %: no caso de a hipótese ser rejeitada, o valor do F é sublinhado a tracejado.

A variável Z_5 foi incluída na variável Z_{10}. A influência da variável Z_0 (França) está incluída no termo independente; a França é o país com mais contratos. O contributo dos outros países deve ser interpretado em termos diferenciais [8].

---

[8] Representando por β_0 o termo constante dos modelos e por δ_i o coeficiente de Z_i (países), evidentemente que o desvio-padrão do contributo total do país Z_i para a explicação da variável dependente é dado por

$$\overset{\sigma}{\beta_0+\delta_i} = \sqrt{V(\beta_0+\delta_i)} = \sqrt{V(\beta_0)+V(\delta_i)+2\ \mathrm{cov}(\beta_0, \delta_i)}$$

Algumas conclusões

a) EXPLICAÇÃO DA DURAÇÃO DO CONTRATO (Y_1)

São os países que explicam, com algum significado, o comportamento de Y_1, com destaque para a *Bélgica*.

b) EXPLICAÇÃO DAS «ROYALTIES» (Y_2)

Os países não dão grande contributo para explicar as *royalties*. Destacam-se, contudo, os E. U. A. e a R. F. A.

c) EXPLICAÇÃO DAS MODALIDADES DE TECNOLOGIA IMPORTADA (W)

Elementos tecnológicos nucleares (ETN)

Licenças de exploração de patentes (W_1)

Tem algum significado a participação no capital da empresa compradora de tecnologia. Apesar de os países no seu conjunto não terem importância, destaca-se a Suíça.

Conhecimentos técnicos (W_2)

Apesar de os países no seu conjunto não terem grande significado, destacam-se a Itália e o Luxemburgo.

Marcas (W_3)

Todas as variáveis relativas aos países têm um certo impacte na explicação desta variável, com destaque para a Bélgica, a Itália e a Holanda.

Elementos tecnológicos prospectivos (ETP)

Estudos de viabilidade (W_6)

Tem um impacte importante a antiguidade do contrato. Quanto aos países, é de referir a Bélgica, apesar de no seu conjunto não terem grande significado.

Os outros elementos tecnológicos prospectivos são quase inexistentes.

Elementos tecnológicos de suporte (ETS)

Formação e troca de pessoal (W_7)

As variáveis todas, em conjunto, têm significado para explicar W_7. São de destacar X_2 e os E. U. A.

Os outros elementos tecnológicos de suporte são quase inexistentes.

d) EXPLICAÇÃO DAS CLÁUSULAS RESTRITIVAS DAS EXPORTAÇÕES (V)

Exportação só para as colónias (V_1)

É relativamente importante o contributo dos países, com destaque para a Bélgica, Holanda e outros.

Proibição total de exportação (V_2)

É relativamente importante o contributo dos países, com destaque para a Holanda e E. U. A.

As outras cláusulas restritivas das exportações não chegam a representar 10 % dos casos.

e) EXPLICAÇÃO DAS OUTRAS CLÁUSULAS RESTRITIVAS *(V)*

Imposição de regras severas ao exercício da propriedade industrial (V_5)

Importância dos países, com destaque para a Bélgica, Holanda e outros.

Matérias-primas e bens intermediários (V_6)

Sem significado.

Segredo (V_7)

Grande importância das variáveis em globo. Sobressaem X_2, a Bélgica, a Itália, a R. F. A. e outros países.

Política de vendas (V_8)

Grande importância dos países, com destaque para a Bélgica e a Itália.

Controlo de qualidade (V_9)

Quase sem significado.

Volumes de produção (V_{10})

Sem significado.

Processos de produção (V_{11})

Sem significado.

Política de preços (V_{12})

Quase sem significado.

As restantes cláusulas restritivas são quase inexistentes.

3.2.2 As convenções do quadro n.º 7 são as mesmas que do quadro n.º 6, apenas com uma alteração: a coluna da estatística *F*, para subconjuntos de variáveis, não se refere a países, mas, conforme os casos, a:

ETN ou ETS+ETP;
Cláusulas restritivas das exportações ou outras cláusulas restritivas.

Dentre as variáveis *W* e *V* escolheram-se as mais significativas para incluir nos modelos.

Algumas conclusões

a) EXPLICAÇÃO DA DURAÇÃO DO CONTRATO (Y_1)

Não depende das *royalties*. Os elementos tecnológicos nucleares são os principais responsáveis pela explicação da duração do contrato, dentre as modalidades de tecnologia importada. Destacam-se daquele conjunto as marcas. Também é importante a formação e a troca de pessoal.

Quanto às cláusulas restritivas, só têm significado as outras cláusulas restritivas, com evidência para «Matérias-primas e bens intermediários», «Política de vendas» e «Volumes de produção».

b) EXPLICAÇÃO DAS «ROYALTIES» (Y_2)

Não depende da duração dos contratos. Quanto às modalidades de tecnologia importada, destacam-se os ETN, com relevo para as licenças de exploração de patentes.

Quanto às cláusulas restritivas, só têm importância as outras cláusulas restritivas, com destaque para os volumes de produção e a política de vendas.

3.2.3 Agrupando as variáveis com mais impacte explicativo sobre Y_1 e Y_2 nas fases 1 e 2, construíram-se mais dois modelos, que, depois de estimados, forneceram os resultados constantes do quadro n.º 8.

Algumas conclusões

a) EXPLICAÇÃO DAS ROYALTIES (Y_2)

As variáveis mais importantes são:

País	E. U. A.
ETN...	Licenças de exploração de patentes Conhecimentos técnicos
Outras cláusulas restritivas	Imposição de regras severas ao exercício da propriedade industrial Volumes de produção

b) EXPLICAÇÃO DA DURAÇÃO DOS CONTRATOS (Y_1)

As variáveis mais importantes são:

País	Bélgica
ETN...	Marcas
ETS	Formação e troca de pessoal

4. CONSIDERAÇÕES FINAIS

Em boa verdade, não se pode dizer que os resultados obtidos neste estudo sejam muito ricos de indicações úteis com vista à explicação integrada do fenómeno das transferências de tecnologia em Portugal. Também não se

encontram nesses resultados subsídios excepcionalmente importantes para a fundamentação de uma política coerente na matéria.

Importa, porém, não perder de vista que estamos perante um primeiro ensaio de exploração, quer da informação de base disponível que pode ser melhorada, quer das potencialidades dos modelos de regressão linear múltipla aplicados a fenómenos que envolvem informações qualitativas que levantam dificuldades, algumas das quais já referenciadas, cuja superação não é fácil, exigindo bastante prática e sensibilidade. A este propósito chama-se a atenção para o percurso seguido neste trabalho na utilização dos modelos de regressão linear múltipla, sobretudo, para a solução da terceira fase, que, de algum modo, é indicador de uma relativa margem de manobra imaginativa que pode dar bons resultados.

É, porém, o campo da estimação simultânea que está no horizonte próximo das preocupações dos autores, adivinhando estes, contudo, as dificuldades que os esperam.

Lisboa, 9 de Março de 1977.

Apêndice

VARIÁVEIS DEPENDENTES QUALITATIVAS [9]

Considere-se o modelo de regressão linear múltipla, onde o regressando é uma variável binária. Seja

$$Y = X\beta + U$$

onde Y representa a coluna das observações da variável dependente y:

$$(i=1, 2, \ldots, n) \quad y_i = \begin{cases} 1, \text{ se o acontecimento se verifica} \\ 0, \text{ se o acontecimento se não verifica} \end{cases}$$

X é a matriz $n \times k$ das observações das variáveis independentes x_j $(j=1,\ldots,k)$: n observações de cada variável k variáveis. O vector dos resíduos é representado por $U=[u_1,\ldots, u_n]'$. Supõe-se que $EU=O$ (valores esperados nulos).

Vamos verificar que a hipótese de homoescedasticidade dos resíduos não se mantém. Do modelo (1) tira-se que

$$u_i = y_i - X_i\beta \quad (i=1,2,\ldots, n)$$

onde X_i é a linha i da matriz X (observação de ordem i de todas as variáveis).

---

[9] Cf. Goldberger, *Econometric Theory*, Wiley, 1964, pp. 248-255; M. A. Kooyman, *Dummy variables in Econometrics*, Tilburg University Press, 1976, pp. 61-72; Thomas Johnson, «Qualitative and limited dependent variables in Economic Relationships», *Econometrica*, vol. 40, n.º 3, Maio de 1972; John G. Cragg, «Some statistical models for limited dependent variables with application to the demand for durable goods», in *Econometrica*, vol. 39, n.º 5, Setembro de 1971.

Conforme a observação de ordem i de y é 1 ou O, assim se tem

$$\begin{cases} u_i = 1 - X_i \beta \\ u_i = X_i \beta \end{cases}$$

Como $E(u_i) = 0$, a distribuição de cada u_i é dada por

u_i	$f(u_i)$
$-X_i \beta$	$1 - X_i \beta$
$1 - X_i \beta$	$X_i \beta$

e, portanto,

$$E(u_i) = (-X_i \beta)(1 - X_i \beta) + (1 - X_i \beta)(X_i \beta) = 0$$

A variância é igual a

$$E(u_i^2) = (-X_i \beta)^2 (1 - X_i \beta + (1 - X_i \beta)^2 (X_i \beta) = \\ = (X_i \beta)(1 - X_i \beta) = E(y_i)[1 - E(y_i)]$$

Vê-se então que as variâncias dos u_i dependem das observações das variáveis independentes. Então, se estimarmos β com o método dos mínimos quadrados, obtêm-se estimativas enviesadas. Teria então de se utilizar outro método para estimar β. À primeira vista, seria de aplicar o método dos mínimos quadrados generalizado, que não necessita da hipótese da homoescedasticidade. Contudo, este método não pode ser aplicado directamente, uma vez que a matriz das covariâncias dos resíduos envolve o vector desconhecido β. Um método iterativo de mínimos quadrados generalizados permite, teoricamente, justificar a estimativa de β [10].

Vejamos, no entanto, mais de perto o significado do valor esperado de y_i condicionado por X_i: $E[y_i \mid X_i]$. Para certo X_i, $f(u_i) = 1 - X_i \beta$ é a probabilidade de u_i quando $y_i = 0 : u_i = -X_i \beta$; para certo X_i, $f(u_i) = X_i \beta$ dá a probabilidade de u_i quando $y_i = 1 : u_i = 1 - X_i \beta$. Como as probabilidades de u_i aparecem relacionadas com os valores assumidos por y_i, então $1 - X_i \beta$ é igual à probabilidade de $y_i = 0$; de modo semelhante, $X_i \beta$ é a probabilidade de $y_i = 1$. Portanto, $E[y_i \mid X_i] = X_i \beta$ mede a probabilidade de $y_i = 1$. Então, os valores estimados y_i são, geralmente, interpretados como estimativas das probabilidades de $y_i = 1$, dado X_i.

Contudo, esta interpretação dá lugar a uma objecção importante:

Como uma função linear não é limitada, haverá sempre um certo conjunto de valores de X_i, que fazem que $E[y_i \mid X_i]$ caia fora do intervalo [0, 1]. Na amostra poderá haver X_i para os quais $\hat{y}_i$ não pertença ao intervalo [0, 1]. Então, quando isto acontece, há incompatibilidade com aquela interpretação probabilística.

O modelo clássico para resolver esta questão é o modelo Probit [11].

---

[10] Ver Goldberger, op. cit., pp. 245 e 250.

[11] Cf. D. J. Finney, *Probit Analysis*, Cambridge University Press, 1952. Vamos seguir de perto Goldberger, op. cit., pp. 250-251.

Seja I um índice que é função linear dos regressores: $I_i = X_i \beta$. Seja I^* uma variável normal estandardizada: N (0,1). Considere-se

$$y_i = \begin{cases} 1 & \text{se } I_i \geqslant I_i^* \\ 0 & \text{se } I_i < I_i^* \end{cases}$$

Cada y_i é, portanto, função de X_i (via I_i) e de I_i^*. Os I_i^*, que desempenham o papel de variâncias residuais, podem ser interpretados como valores críticos do índice (ex.: $y_i = 1$ significa, por exemplo, que o contrato de tecnologia importada incide sobre conhecimentos técnicos; $y_i = 0$ significa o caso contrário. O contrato incidirá sobre conhecimentos técnicos se os valores das variáveis independentes forem tais que $I_i \geqslant I_i^*$).

Tem-se

$$\begin{aligned} \text{Prob } (y = 1 \mid I) &= \text{Prob } (I^* \leqslant I \mid I) = F(I) \\ \text{Prob } (y = 0 \mid I) &= \text{Prob } (I^* > I \mid I) = 1 - F(I) \end{aligned}$$

Pelo facto de os I_i, e portanto as probabilidades anteriores, serem função de β, é sugerida a estimativa de máxima verosimilhança de β. Sem perda de generalidade, suponhamos que nas primeiras p observações se tem $y = 1$ e nas restantes $n - p$ observações se tem $y = 0$. Então, a função de verosimilhança da amostra é

$$L = F(I_1) \ldots F(I_p)\,[1 - F(I_{p+1})] \ldots [1 - F(I_n)]$$

e a verosimilhança logarítmica

$$L^* = \sum_{i=1}^{p} \log F(I_i) \sum_{i=p+1}^{n} \log [1 - F(I_i)]$$

com

$$F(I_i) = \frac{1}{2\pi} \int_{-\infty}^{X_i \beta} e^{-\frac{u^2}{e^2}}\, du$$

Calculando as derivadas parciais de L^* em ordem aos β e igualando a zero, obtêm-se as equações normais que determinam os estimadores de máxima verosimilhança dos β, isto é, $\hat{\beta}$. As equações normais, evidentemente, não são lineares.

No modelo Probit, o valor esperado de y_i condicionado por I_i é dado por

$$E(y_i \mid I_i) = \text{Prob } (y_i = 1 \mid I_i).\ 1 + \text{Prob } (y_i = 0 \mid I_i).\ 0 = F(I_i)$$

que é a ordenada da distribuição cumulativa normal estandardizada: $0 \leqslant F(I_i) \leqslant 1$. O valor esperado estimado $\hat{y}_i = F(I_i) = F(X_i \hat{\beta})$ tem as mesmas propriedades. Teoricamente fica resolvido o problema da limitação de y_i ao intervalo [0,1].

A dificuldade do método está na resolução das equações normais.

Finalmente, uma observação sobre o coeficiente de determinação. Com o objectivo de simplificar, consideremos o seguinte modelo de regressão linear simples:

$$y_i = \alpha + \beta x_i + u_i$$

onde y_i pode tomar os valores 1 ou 0 (observações de uma variável binária) e x_i pode tomar qualquer valor (observações de uma variável real). Como, neste caso, o diagrama de dispersão é constituído por pontos sobre as duas rectas $y = 0$ e $y = 1$, não tem grande significado o ajustamento de uma recta, sendo reduzida a importância do coeficiente de determinação.