UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO
Mestrado em Matemática Aplicada à Economia e à Gestão

Erros de Medida em Modelos Não Lineares

Montezuma Boaventura Guimarães Dumangane

Orientador: Professor Doutor João Manuel Caravana Santos Silva

Presidente do Júri: Professor Doutor João Manuel Andrade e Silva
Vogais: Professor Doutor José António Ferreira Machado
Professor Doutor João Manuel Caravana Santos Silva

Janeiro 1997

UNIVERSIDADE TÉCNICA DE LISBOA
INSTITUTO SUPERIOR DE ECONOMIA E GESTÃO
Mestrado em Matemática Aplicada à Economia e à Gestão

Erros de Medida em Modelos Não Lineares

Montezuma Boaventura Guimarães Dumangane

Orientador: Professor Doutor João Manuel Caravana Santos Silva

Presidente do Júri: Professor Doutor João Manuel Andrade e Silva
Vogais: Professor Doutor José António Ferreira Machado
Professor Doutor João Manuel Caravana Santos Silva

Janeiro 1997

Resumo

Em quase todos os trabalhos econométricos aplicados, os dados disponíveis para análise são apenas aproximações às grandezas que figuram na modelização económica que se pretende descrever na especificação estatística. Este problema, que se traduz na utilização de variáveis proxy, denomina-se erros de medida nas variáveis.

Este estudo pretende analisar as consequências da utilização destas variáveis contaminadas com erro de medida nos modelos de regressão não linear. Analisam-se os casos em que o erro de medida afecta as variáveis explicativas e quando a variável dependente. Nos modelos de escolha discreta, a presença de erro de medida na resposta conduz ao problema de má classificação na variável dependente.

Seja qual for a natureza do erro, a utilização de dados contaminados, quase sempre conduz a um enviesamento nos procedimentos estatísticos e à consequente inconsistência na estimação dos parâmetros das distribuições condicionais.

Dada a complexidade do problema, e a grande variedade de formulações paramétricas, pretende-se com esta dissertação analisar genericamente as distorções introduzidas nas densidades, fazendo-se recurso a aproximações às distribuições e formulando o problema da reespecificação das verosimilhanças, apresentando alguns resultados específicos para algumas especificações paramétricas. Utilizando a metodologia de aproximação às distribuições constroi-se um teste tipo score para detecção de erro de medida nas variáveis explicativas.

O problema da má classificação na variável dependente é estudado para os modelos de escolha binária e para um caso particular do modelo de regressão Poisson.

Palavras Chave: Erro de medida; Má classificação; Teste score; Modelos não lineares; Aproximação para variância do erro pequena; verosimilhança.

Abstract

In almost all applied econometric work the data available for analysis are just approximations to the constructs that figure in the economic model that underlies the statistical work. The problem of using this proxy variables is called error in variables.

This study analysis the consequences of using error contaminated variables in nonlinear regression models. It studies the cases when the measurement error affects both the covariates and the response variate. In discrete choice models, the presence of measurement error leads to the misclassification problem.

Whatever the nature of the measurement error, using this data almost always leads to biased statistical procedures and inconsistency in the estimation of parameters of conditional distributions.

Given the complexity of the problem and the enormous variety of parametric specifications, this dissertation pretends to analyze in a general way the distortions induced in the densities, using small variance approximation and using likelihood analysis, presenting results for some parametric specifications.

By using this methodology a score test to detect measurement error in covariates is developed.

Misclassification in response variate is studied in binary choice models and for a special case of the Poisson regression model.

Key words: Measurement error; Misclassification; Score test; Nonlinear models; Small variance approximation; likelihood.

Agradecimentos

Desejo deixar aqui expressa, a minha gratidão pelos contributos inestimáveis que recebi no decurso da realização do presente trabalho:

Ao meu orientador, Professor Doutor João Manuel Caravana Santos Silva, pelo acompanhamento, estímulo e rigor exigido.

À Universidade Católica Portuguesa, na pessoa do Professor Doutor Fernando Machado pelos meios colocados à minha disposição.

Ao Dr. Manuel Leite Monteiro, pelo auxílio prestado e disponibilidade demonstrada.

Aos meus pais e irmãos, pelo apoio e paciência revelados neste árduo processo de gestação.

Aos meus colegas de gabinete pelo apoio dado nos momentos difíceis.

Ao Programa PRAXIS XXI, pelo financiamento concedido para a realização deste curso.

Indice

Capítulo I- Introdução **1**

1.1 Origens do Erro Medida 3

1.1.1 Erros de Transcrição 4

1.1.2 Erros nas Respostas 4

1.1.3 Utilização de Proxies 5

1.2 Uma visão geral 5

Capítulo II- Tipologia dos Erros de Medida **8**

2.1 Modelos Funcionais e Estruturais 8

2.2 Modelos para o Erro de Medida 9

2.2.1 Modelos de Erro 9

2.2.2 Modelos de Regressão Calibrada 10

2.3 Transportabilidade 10

2.4 Origem dos dados 11

2.5 Diferenciabilidade do Erro de Medida 12

2.6 Primeira Abordagem: Efeito Atenuação 13

Capítulo III- Regressão Linear **15**

3.1 Introdução 15

3.2 Modelo de Regressão Linear 15

3.2.1 Análise do Modelo de Regressão Linear Simples . . . 15

3.2.2 Regressão Simples com estrutura de erro mais complexa 16

3.2.3 Regressão Múltipla com uma variável explicativa medida com erro 17

3.2.4 Múltiplas variáveis explicativas medidas com erro . . 18

3.3 Métodos Alternativos ao OLS 18

3.3.1 Método das Variáveis Instrumentais 19

3.3.2 Método dos Momentos 21

3.3.3 Regressão Ortogonal 21

Capítulo IV- Modelos não Lineares **23**

4.1 Introdução 23

4.2 Erro de Medida em Modelos Não Lineares: Uma abordagem geral 24

4.3 Aproximação para Variância do erro pequena 27

4.3.1 Introdução 27

4.3.2 Aproximação para erro nas variáveis explicativas . . . 29

4.4 Máxima Verosimilhança 35

4.4.1 Introdução 35

4.4.2 Verosimilhança quando não observamos X 37

4.4.3 Modelos de Erro 37

4.4.4 O Modelo de Berkson 39

4.4.5 Verosimilhança quando X é parcialmente observado . 40

4.5 Modelo Poisson 40

Capítulo V- Teste Score **45**

5.1 Introdução 45

5.2 A Estatística de Teste 45

5.3 Implementação 48

5.4 Simulações 52

5.4.1 Modelo Logit 53

5.4.2 Modelo Poisson 56

Capítulo VI- Erro na variável Dependente **59**

6.1 Introdução 59

6.2 Aproximação para Variância do Erro Pequena 61

6.3 Máxima Verosimilhança 66

6.3.1 Verosimilhança quando Y não é observado 66

6.3.2 Verosimilhança quando Y é parcialmente observado . 67

6.4 Má Classificação no Modelo de Regressão Poisson 68

6.4.1 Introdução 68

6.4.2 O modelo 69

6.5 Simulações 75

6.6 Má Classificação em Modelos de Escolha Binária. 77
6.6.1 Introdução 77
6.6.2 Modelo de Escolha Discreta Mal Classificado 78

Capítulo VII- Conclusões 84

Bibliografia 87

Capítulo I- Introdução

A presente dissertação tem por objectivo o estudo do problema dos erros de medida em modelos não lineares. Esta questão constitui um tópico de investigação da teoria Econométrica. Impõe-se por isso, antes de nos debruçarmos sobre esta temática, responder à consideração prévia de uma questão mais pertinente: O que é a Econometria?[1] A resposta a esta pergunta aparentemente simples pode tornar-se bastante complexa. A dificuldade decorre da relativa juventude deste ramo da ciência económica, a que se acresce o grande crescimento em direcções bastante diversificadas. Deste modo, ao invés de caminharmos para uma definição unânime, as definições sobre o que é a teoria econométrica divergiram cada vez mais. Daí que já tivesse havido quem como Goldeberg (1989), um proeminente econometrista, numa atitude conciliadora e assumindo uma postura não dogmática, acabou por definir a econometria, afirmando que "...hoje em dia, econometria é o que os econometristas fazem". Uma definição hoje em dia considerada standard, é que "a econometria é o estudo da aplicação de métodos de análise estatística a fenómenos económicos". A grande diversidade de respostas a esta questão, (onde se incluem visões muito cínicas da utilidade dos econometristas) tem origem no facto da classe dos econometristas ser bastante vasta e constituída por investigadores com diferentes motivações e formações. Primeiramente, pode-se considerar que eles são economistas, capazes de utilizar a teoria económica para melhorar as análises empíricas dirigidas a problemas concretos. Por vezes, são matemáticos que formulam teoria económica de forma a poder ser testada estatisticamente. Outras, são estatísticos aplicados, que através de métodos computacionais tentam estimar relações económicas, ou prever acontecimentos. Por vezes são estatísticos teóricos, que utilizam os seus conhecimentos no desenvolvimento de técnicas estatísticas apropriadas para a resolução de problemas empíricos que caracterizam a ciência económica. Correndo o risco de estarmos a elaborar uma análise demasiadamente simplista, é a esta última função, a que por conveniência, nos vamos referir quando pretendermos descrever

[1] O termo "econometria", assumiu forma aquando da formação nos anos 30, da "Econometric Society" e a fundação da revista *Econometrica*.

Etimologicamente, a econometria consistiria na medida, ou medição dos fenómenos económicos. Contudo a dimensão deste ramo da ciência económica não se compadece com uma visão tão puéril e redutora.

o papel do econometrista (Peter Kennedy 1992). Citando Malinvaud:

> "*The art of the econometrician consists in finding the set os assumptions wich are both sufficiently specific and sufficiently realistic to allow him to take the best possible advantage of the data available to him*" (Malinvaud, 1966).

Dada a natureza das relações económicas e a incapacidade de se controlar as experiências na ciência económica, as hipóteses estatísticas são frequentemente violadas. Por isso, não é de estranhar que a generalidade dos economistas sintam que o maior problema com que os econometristas se deparam é o facto de os dados com os quais têm de trabalhar serem muito pobres. Este facto decorre directamente do carácter não experimental da ciência económica. É dentro deste contexto que se insere o problema dos erros de medida. Antes de mais, ele surge precisamente do próprio carácter das relações que se pretendem estabelecer entre variáveis económicas[2] . Muito genericamente, o problema dos erros de medida preocupa-se com as implicações de se usar variáveis incorrectamente medidas.

Durante bastante tempo, estudos realizados sobre as mais variadas temáticas, detectaram a presença de erros na mensuração de certas variáveis, com particular incidência nos dados recolhidos por respostas e dados previamente tratados. Até à algum tempo, apenas uma fracção da literatura relativa a procedimentos estatísticos tinha em conta esta realidade e usava procedimentos concebidos para a utilização de variáveis explicativas medidas com erro nos modelos de regressão. A literatura neste campo tem-se desenvolvido significativamente. O objectivo é aumentar o número de técnicas estatísticas que explicitamente reconheçam a presença de erro de medida e que nessas condições possam estimar os parâmetros de interesse de forma consistente. Em resposta a Morgenstern (1963) sobre a precisão dos dados económicos Griliches (1985) notou, relativamente a este tipo de dados que, "*That is all there is-it is the only game in town and we have to make the best of it*". É dentro deste espírito que se vai desenvolver o estudo que se segue. Apesar das dificuldades e complexidades introduzida por este problema, há que rentabilizar os conhecimentos estatísticos de forma a se construirem técnicas de estimação que permitam extrair toda a informação relevante contida

[2] Embora o âmbito de aplicação deste problema se alargue a outras ciências que não a economia.

na amostra contaminada com erro de medida. Pretende-se diagnosticar os efeitos e em alguns apresentar soluções.

O estudo das consequências deste problema no modelo de regressão linear, absorve grande parte da literatura sobre erros de medida e está compilada no trabalho de Fuller (1987). Nos últimos anos tem aumentado o interesse pelo estudo das implicações nos Modelos de Regressão Não Linear, sendo a compilação efectudada por Carroll, Ruppert & Stefanski (1995) um excelente apanhado das técnicas relevantes mais avançadas. A presença deste tipo de problemas tem como área de aplicação de excelência as ciências naturais. Muitas vezes, os problemas dirigidos a estas ciências implicam a necessidade de se medir reacções de determinados organismos à absorção de compostos químicos, reacções biológicas despoletadas pela exposição a determinados factores de risco, etc. Quase sempre, nessas condições é difícil medir com rigor essas variáveis. Isto acontece quando analisamos o efeito de fertilizantes sobre colheitas, como ilustrado por Fuller (1987). Carroll, Ruppert & Stefanski (1995) fala em experiências biológicas, com herbicidas e utiliza de forma recorrente como exemplo ilustrativo o *estudo Framingham*, que analisa a relação entre a tensão arterial e doenças cardiovasculares, pondo ênfase no problema de mensuração da tensão arterial dos indivíduos. Quanto a questões económicas, Chesher (1990) refere a necessidade de aplicar estas técnicas quando se utilizam dados provenientes de inquéritos às famílias. Outro exemplo, é quando se pretende observar o consumo. Na maioria das vezes estamos interessados num conceito de taxa de consumo média de longo prazo (ao longo do tempo), mas o que a acabamos por utilizar é o consumo num determinado (geralmente curto) período de tempo. O mesmo problema acontece quando tentamos medir o rendimento e mais genericamente quando pretendemos medir qualquer tipo de fluxo económico. Outra situação a ter em conta é a análise de surveys. Dado que os indivíduos são observados por apenas um curto período de tempo, os surveys tendem a produzir dados muito variáveis no rendimento e despesas, acentuando desigualdades na distribuição do rendimento e na despesa.

1.1 Origens do Erro Medida

A questão econométrica da existência de erros de medida é cada vez mais relevante na análise dos procedimentos estatísticos. Os erros de medida são de particular

importância porque na maior parte dos casos podem enviesar profundamente a análise econométrica. Contudo, a expressão erros de medida pode esconder várias situações distintas, com consequências igualmente distintas nos métodos de análise utilizados. Por isso, interessa antes de mais considerar quais as diferentes fontes dos erros de medida.

1.1.1 Erros de Transcrição

Erros de medida nos dados utilizados pelo analista podem ocorrer muito naturalmente por falha do entrevistador no tratamento dos dados recolhidos. É de supor que este tipo de erros existam, mas que sejam também independentes dos verdadeiros valores das variáveis em causa. Os erros originados por este tipo de procedimento são geralmente bastante elevados e sem nenhum padrão de comportamento específico, daí que as técnicas de análise gráfica e a literatura relativa às observações influentes sejam as mais indicadas para a detecção deste fenómeno.

1.1.2 Erros nas Respostas

A presença de erros de medida pode ter origem nas próprias respostas dos entrevistados. Nestes casos, pode ser bastante complicado distinguir a relação entre o erro de medida e o valor correcto da variável em causa visto que perante algumas destas situação, o erro de medida está correlacionado com o valor da variável explicativa e a relação funcional que traduz essa correlação é quase sempre bastante difícil de identificar. Por vezes os entrevistados dão respostas que não são precisas por não terem vontade de fazer um esforço para se recordarem da resposta exacta. As motivações para esta situação podem ser as mais variadas: esquecimento, falta de paciência que se traduz em diversos casos em respostas arredondadas, etc.. Outras vezes os entrevistados, consoante o teor do inquérito, preferem esconder algumas das suas características, ou enviesar as respostas, sendo a magnitude do enviesamento função do verdadeiro atributo. Tais situações podem ocorrer quando se interroga sobre questões como rendimento, consumo de cigarros, registo de nascimentos em mães solteiras etc.

1.1.3 Utilização de Proxies

O tipo de erro de medida mais importante e mais comum acontece porque de alguma forma o analista utiliza a variável errada. Isto é, a variável observável não é a contrapartida empírica da variável que deveria estar a ser utilizada. Muitas vezes o analista não tem sequer noção de que está a cometer esse erro, ou seja, de que na verdade está a trabalhar com uma aproximação. A variável observável constitui neste caso uma proxy, ou uma aproximação à verdadeira variável, também denominada variável latente. Nesta situação não existe correlação entre o erro cometido e o verdadeiro valor da variável que se pretende observar, ou seja, a variável latente e o erro de medida são independentes. É sobre este tipo de situação que se vai centrar grande parte da análise subsequente.

Muitas vezes a questão que se coloca é saber se de facto existe um verdadeiro previsor. A modelização do comportamento de uma determinada variável dependente, é explicada teoricamente por um determinado conjunto de variáveis explicativas cuja relação é conhecida e se supõe verdadeira. Nestas situações muitas vezes acontece que as variáveis utilizadas apenas teoricamente podem ser definidas, (porque reflectem ou têm origem na teoria económica) não havendo uma variável que na prática seja a sua verdadeira realização. Nestes casos, que representam uma fracção importante quando falamos de variáveis económicas, o investigador é forçado a usar um definição operacional para o verdadeiro previsor que esteja o mais próximo possível dele, mas que simultaneamente seja mensurável. Perante este problema coloca-se uma dupla questão: primeiro, de saber o quão distante está esta definição operacional da verdadeira variável, segundo de saber com que precisão é possível medir a variável assim definida.

1.2 Uma visão geral

O principal objectivo deste estudo é a análise das consequências da utilização de variáveis medidas com erro nos modelos de regressão não linear. A motivação, decorre de muito frequentemente se considerar que, o efeito da existência de erro de medida na variável explicativa se esgota no efeito atenuação, segundo o qual, o parâmetro associado à variável contaminada é enviesado em direcção à origem e como tal os testes de significância individual dos parâmetros continuam a ser

válidos. Pretende-se com o presente trabalho considerar o efeito atenuação numa perspectiva mais abrangente.

O estudo divide-se em seis partes: considerações introdutórias, tipologia dos modelos para o erro de medida, análise da regressão linear, análise da classe dos modelos de regressão não lineares, construção de um teste para detecção de erro de medida nos regressores, análise do erro de medida na variável dependente e problema de má classificação em modelos de variável dependente discreta.

Definidas as hipóteses para o comportamento da relação entre o erro de medida, a variável contaminada e a verdadeira variável não observada no capítulo II, o capítulo III analisa os efeitos da presença de erro na variável explicativa no modelo de regressão linear (sob várias hipóteses para a sua estrutura), para concluir da necessidade de uma interpretação mais abrangente do efeito atenuação. Na secção 3.3, consideram-se métodos alternativos ao OLS, para estimação consistente dos parâmetros de interesse.

A análise dos modelos não lineares no capítulo IV, parte já do princípio de que a presença de erro na variável explicativa tem implicações bastante complexas que dependem da estrutura da especificação em causa e das hipóteses formuladas sobre o modelo para o erro de medida. Na secção 4.3, introduz-se uma aproximação para variância pequena às distribuições marginais, condicionais e respectivos momentos para a classe dos modelos não lineares. Esta metodologia, apresentada por Chesher (1991) permite, de uma forma genérica, i.e., sem especificar a forma funcional das distribuições em causa, estudar as consequências sobre as distribuições e momentos de interesse da presença de erro de medida. Mostra-se que o impacto do erro de medida depende da curvatura das densidades, representada pelas segundas derivadas, mas não depende da distribuição do erro de medida. Considera-se a título exemplificativo o caso do modelo de regressão linear normal e os modelo Logit e Probit. A secção 4.4, formula o problema num contexto de máxima verosimilhança, privilegiando uma análise estrutural onde é necessário o conhecimento das distribuições da verdadeira variável e do modelo para o erro de medida. Consideram-se diferentes estruturas do modelo para o erro de medida e o caso em que é possível observar a verdadeira variável para um subconjunto da amostra. Na secção 4.5 considera-se o caso do modelo de regressão Poisson, fazendo-se uma análise funcional, tirando partido das

propriedades relativas aos momentos que caracterizam esta distribuição.

O capítulo V, utiliza a técnica de aproximação das distribuições para construir um teste score para detecção de erro de medida clássico. A sua forma é um misto de teste de matriz de informação e teste de detecção de não linearidades nos regressores. Consideram-se os casos dos modelos Logit, onde o teste assume a forma de um teste de omissão de variáveis e do modelo de regressão Poisson, apresentando-se os resultados de simulações.

Finalmente, no capítulo VI considera-se o problema da presença de erro na variável dependente, repetindo-se a análise com recurso a uma aproximação para variância pequena, secção 6.2 e máxima verosimilhança, secção 6.3. Neste capítulo estuda-se ainda o caso particular de quando a variável dependente é discreta, i.e., má classificação na variável dependente. Nesta situação, a tipologia do erro de medida apresentada no capítulo II não se aplica, dada a estrutura do erro imposta pelo carácter discreto da variável mal medida, e a possibilidade de não verificação da hipótese de não diferenciabilidade do erro de medida. Apresenta-se o problema para os modelos de regressão binários, secção 6.5, e para um caso particular de má classificação no modelo Poisson, secção 6.4, onde apenas as observações com resposta zero podem estar mal classificadas nos uns. Nestes modelos, considera-se a possibilidade de reespecificação da logverosimilhança para a estimação dos parâmetros de interesses, apresentando-se os resultados de estudos de Monte Carlo, e no caso do Poisson mal classificado um estimador GMM.

Capítulo II- Tipologia dos Erros de Medida

A constatação da existência de erros de medida leva à necessidade de se definir uma estrutura para a relação entre W, a variável de facto observável, X a variável latente sobre a qual se deseja obter informação, u o erro de medida e Z as restantes variáveis explicativas medidas sem erro. O objectivo desta secção é fornecer o quadro das várias hipóteses que possibilitam a aplicação de uma análise estatística.

2.1 Modelos Funcionais e Estruturais

A modelização dos erros de medida assenta basicamente em duas características. A primeira diz respeito às hipóteses estocásticas que estabelecemos relativamente à variável não observável. De acordo com este critério, pode-se distinguir entre modelização estrutural, onde a distribuição da variável X é perfeitamente especificada, ou seja é modelizada com uma distribuição paramétrica conhecida, e modelização funcional onde X é considerado um vector de observações fixo ou aleatório, mas neste caso apenas se impõem hipóteses mínimas relativamente à sua distribuição (geralmente modeliza-se apenas a média e a variância).

Em rigor, a escolha do tipo de modelização a adoptar, deveria depender das características da amostra. Se ela contêm observações de todos os indivíduos relevantes, de modo a que amostra e universo coincidem, então deveríamos optar por uma modelização funcional. As observações da variável seriam sempre fixas, independentemente do processo de amostragem. Alternativamente, se a amostra é aleatoriamente retirada de um universo, o mais correcto será a utilização de uma modelização estrutural, onde se supõe conhecida a lei probabilística que gera os dados.

A opção pela modelização funcional pode facilmente justificar-se, pelo facto de ao encontrar-se um estimador ou método de estimação consistente, ele ser robusto relativamente às hipóteses que possamos levantar sobre a distribuição da verdadeira variável, ou seja, ele é robusto a erros de especificação na distribuição de X. Por outro lado, pode-se sempre argumentar que na maioria dos casos

é difícil conhecer a verdadeira distribuição de X, e nestes casos não é válida uma argumentação semelhante à que se utiliza para variáveis do tipo residual; em grandes amostras e em virtude do teorema do limite central degeneram na distribuição normal.

A modelização estrutural tem contudo como grande atractivo: permitir a obtenção de resultados mais precisos e eficientes, embora estejam sempre bastante dependentes das hipóteses que estabelecemos para a distribuição da variável latente.

2.2 Modelos para o Erro de Medida

A segunda característica a identificar, diz respeito à estrutura da relação entre a variável proxy a variável latente e o erro de medida, ou seja, o modelo para o erro de medida. Existem basicamente duas classes de modelos.

2.2.1 Modelos de Erro

Neste tipo de modelo a distribuição condicional de W dado X, Z e u é modelizada. Dentro desta classe de modelos, o caso mais simples e a que se dará maior ênfase durante o presente estudo, denomina-se Modelo de Erro Clássico, que é escrito como,

$$W = X + u \tag{2.1}$$

Este modelo é o mais apropriado quando se tenta medir directamente X e a variável observada W é uma medida não enviesada da verdadeira variável. Por isso, supõe-se adicionalmente que $E(u \mid X) = 0$. Caso não exista uma medida não enviesada para a variável latente, devido por exemplo a um enviesamento sistemático ou deliberado nas respostas dos entrevistados, então deve-se considerar a especificação,

$$W = \alpha_0 + \alpha_x X + \alpha_z Z + u \tag{2.2}$$

desde que se garanta que o enviesamento seja independente do erro cometido, o que equivale a dizer que $E(u \mid X, Z) = 0$. A ideia subjacente a esta modelização é a de que W é uma medida enviesada de X sendo necessário através da observação

de outras variáveis proceder à correcção do enviesamento. Desta forma uma medida não enviesada para a variável latente é dada por $(W - \alpha_0 - \alpha_z Z)/\alpha_x$.

2.2.2 Modelos de Regressão Calibrada

Neste tipo de modelos pretende-se estudar a distribuição condicional de X dado W, Z e u. O verdadeiro valor da variável é função do valor observado para a variável proxy. Sendo assim e contrariamente ao que sucedia na modelização anterior, esta é fixa e prédeterminada e determina o valor da variável latente.

$$X = \gamma_0 + \gamma_w W + \gamma_z Z + u, \; E(u \mid Z, W) = 0 \tag{2.3}$$

No caso em que X é não enviesado para W, então $\gamma_0 = 0$, $\gamma_w = 1$ e $\gamma_z = 0$ temos o denominado modelo de Berkson.

Note-se que em ambos os tipos de modelos a relação não tem necessariamente de ser do tipo aditiva, pode ser multiplicativa, ou ter outra qualquer forma eventualmente linearizável.

A escolha do tipo de modelo para o erro a considerar dependerá das características da observabilidade da amostra disponível. Um possível critério é considerar qual das duas variáveis é fixa determinando a aleatoriedade da outra[3] . De qualquer forma a modelização tipo regressão calibrada é sempre mais complexa. Muitas vezes a escolha entre as duas modelizações possíveis não é óbvia e nesses casos a escolha assenta basicamente em critérios de conveniência e considerações empíricas.

2.3 Transportabilidade

A consideração da especificação dos modelos para o erro de medida implica que seja necessário caracterizar os respectivos modelos mediante a estimação dos seus parâmetros. Dada a limitação dos dados e visto que não observamos a variável de interesse, o modelo para o erro de medida é muitas vezes construído com dados obtidos fora do estudo em causa. Tal só é possível, se as características dos

[3] Como exemplo considere-se o estudo sobre o efeito de herbicidas realizado por Rudemo, et al. (1989). Neste, uma quantidade de herbicida W era aplicada a uma planta. Contudo a verdadeira quantidade de herbicida absorvida, X difere de W, não só por causa do potencial erro na aplicação da dose, mas também devido ao próprio processo de absorção da planta. Neste caso W determina X e logo deveríamos considerar uma modelização tipo regressão calibrada.

fenómenos sob análise em ambos os estudos independentes forem iguais, o que se traduzirá na possibilidade de se poder transportar não só o modelo para o erro, mas também e principalmente os seus parâmetros sem que com isso haja enviesamento. A transportabilidade exige que não só a estrutura do modelo mas também os parâmetros relevantes sejam transportáveis. Se os universos em questão forem qualitativamente idênticos, então o conhecimento da distribuição de W dado (X, Z) ou de X dado (W, Z) na amostra relativa ao estudo independente pode ser transportado para o estudo em questão.

Contudo há que ter em atenção que a transportabilidade dos dados não pode ser aferida de ânimo leve. A utilização de estudos independentes para a construção de modelos erro de medida pode introduzir enviesamento no estudo do modelo inicial. É que, pelo teorema de Bayes, a distribuição de X dado W e Z depende não só da distribuição de W dado (X, Z) mas também da distribuição de X dado Z que muito seguramente diverge de população para população. Isto quer dizer que, mesmo que a estrutura do modelo para o erro de medida seja transportável a distribuição da verdadeira variável na maioria dos casos deve diferir.

2.4 Origem dos dados

Qualquer uma destas formas de modelização requer informação que nos possibilite estimar a distribuição de X dado (W, Z) e de W dado (X, Z). Esta informação é classificada de acordo com a sua origem e pode ser separada em duas grandes categorias:

• Subconjuntos internos de dados primários: Corresponde a observar a variável latente para alguns indivíduos do estudo em causa.

• Dados externos ou independentes: Dentro desta categoria existem três tipos de dados, que se supõe estarem disponíveis numa subamostra aleatória do estudo em causa:

• Dados de validação: Onde observamos a variável latente X num estudo externo.

• Dados repetidos: Onde dispomos de várias réplicas de W.

• Instrumentos: Onde além de W podemos observar uma variável instrumental T independente do erro da regressão e do erro de medida e correlacionada

com os regressores.

A construção de modelos para o erro de medida que caracterizem as distribuições condicionais de X dado (W, Z) e de W dado (X, Z) é por vezes efectuada recorrendo à utilização de réplicas. A vantagem na sua utilização pode ser justificada pelo facto da média das réplicas ser uma medida melhor do que apenas uma observação. Além disso se considerarmos o modelo clássico as réplicas podem ser utilizadas para estimar a variância do erro de medida. Contudo o conceito de réplicas implica que o entrevistado responda mais de uma vez à mesma pergunta. Tal facto origina que em algumas circunstâncias determinadas pelo teor da pergunta haja um enviesamento nas segundas respostas. Este fenómeno pode ser ultrapassado se juntarmos uma constante às segundas réplicas de modo a que a média das segundas réplicas iguale a das primeiras (Carroll, Ruppert & Stefanski, 1995).

2.5 Diferenciabilidade do Erro de Medida

Finalmente, na tipologia da modelização de erros de medida importa distinguir entre erro de medida diferencial e não diferencial. O erro de medida é denominado não diferencial quando a distribuição de Y dado (X, Z, W) depende apenas de X e Z. Dito de outra forma, W não pode conter nenhuma informação adicional para explicar Y que não esteja já contida em X e Z. Nestas condições diz-se que W é uma variável substituta. Caso tal condição não se verifique, o erro de medida é diferencial. Nestes casos as técnicas de análise são mais complexas, sendo por vezes necessário observar a verdadeira variável para pelo menos um subconjunto das observações. A hipótese de erro de medida não diferencial é absolutamente vital quando se utiliza o método da regressão calibrada (ver Carroll, Ruppert & Stefanski, 1995). Sob esta hipótese é possível simplificar a relação entre Y e W. No caso do modelo linear a hipótese da não diferenciabilidade permite-nos escrever a regressão dos dados observados como uma regressão linear de Y sobre $E(X \mid W)$. Partindo de

$$\begin{aligned} E(Y \mid W) &= E\{E(Y \mid X, W) \mid W\} = \\ &= E\{E(Y \mid X) \mid W\} = \\ &= E\{\beta_0 + \beta_x X \mid W\} = \\ &= \beta_0 + \beta_x E(X \mid W) \end{aligned} \tag{2.4}$$

A hipótese da não diferencialbilidade foi essencial para a passagem da primeira para a segunda expressão (2.4).

2.6 Primeira Abordagem : Efeito Atenuação

Até muito recentemente aceitava-se a ideia, de que o único efeito da presença do erro de medida era o de enviesar em direcção à origem o coeficiente da variável contaminada e como tal, poderíamos ignorar o problema para efeitos de teste de significância individual do parâmetro. Estas conclusões, na sua maioria são derivadas da análise da estimação pelo método dos mínimos quadrados do modelo de regressão linear simples. A conclusão que se retira neste modelo conduz à interpretação de que quando o regressor está contaminado com erro de medida (e considerando o modelo para erro de medida clássico) o estimador dos mínimos quadrados é enviesado em direcção ao zero, sendo o enviesamento tanto maior quanto maior a proporção da variação do regressor atribuível ao erro de medida, ou seja, quanto maior fôr a variância do erro relativamente à da variável latente. Da análise deste modelo conclui-se que perante a existência de erros de medida, o efeito das variáveis explicativas contaminadas sobre a variável dependente é atenuado.

Contudo nem sempre o efeito é assim tão simples. Quando este erro é considerado grave, i.e., de dimensão considerável, a primeira conclusão que se retira é que os habituais procedimentos estatísticos podem produzir estimadores inconsistentes e como tal os testes podem levar a resultados enganadores. Numa primeira abordagem, a contaminação origina que a variável explicativa pareça variar num intervalo de amplitude superior ou mais alargado do que na realidade é. Desta forma, altera a escala de variação dos regressores não tendo porém qualquer efeito sobre a escala de variação da variável explicada. Consequentemente, o efeito sobre a variável dependente de uma variação marginal duma variável medida com erro é inferior ao efeito da mesma variação da variável explicativa isenta de erro.

Apesar desta conclusão se traduzir para o modelo de regressão linear simples no efeito atenuação sobre o parâmetro associado à variável explicativa, ela não pode ser generalizada a outras formulações para a distribuição de Y dado X e Z. O efeito atenuação, tal como foi considerado neste modelo é um caso particular, devendo ser considerado num contexto mais alargado. Interessa pois saber qual a natureza do efeito atenuação em situações mais gerais.

O que se pode dizer é que o enviesamento introduzido pelo erro de medida pode ser bastante complexo. Nestas condições, quase não existem resultados para amostras finitas, mesmo sob a hipótese mais simples de erro de medida clássico. Quase todos os resultados conhecidos são assintóticos.

Capítulo III- Regressão Linear

3.1 Introdução

Uma grande parte da literatura sobre erros de medida tem sido conduzida em relação ao estudo das consequências da utilização do método dos mínimos quadrados no modelo de regressão linear[1] . O resultado desta análise levou a considerar-se que perante a existência de erro de medida clássico o estimador dos mínimos quadrados é enviesado em direcção à origem, originando o que se denominou de efeito "atenuação".

Esta conclusão, embora seja o resultado de uma análise em condições particulares, pode incorrectamente ser aceite como a consequência deste problema em todas as outras especificações. A veracidade desta afirmação depende contudo da especificação considerada para a relação de Y condicional em X e da relação entre X, W e U. Mesmo no modelo de regressão linear, os efeitos da existência de erro de medida dependem da estrutura ou dimensão da matriz das variáveis explicativas, e da distribuição conjunta do erro de medida e das restantes variáveis, nomeadamente da sua correlação com o erro da equação.

O objectivo desta secção é concluir que esta interpretação nem sempre é válida.

3.2 Modelo de Regressão Linear

3.2.1 Análise do Modelo de Regressão Linear Simples

Considere-se o modelo de regressão linear univariado, $Y_t = \beta_0 + \beta_x X_t + \varepsilon_t$, $t = 1, 2, ..., N$. Assuma-se ainda que estamos perante uma modelização funcional onde o único regressor tem média μ_x e variância σ_x^2 e as observações da variável latente $(X_1, X_2, ..., X_N)$ são consideradas fixas em amostras repetidas. Suponha-se ainda que ao invés de X_t apenas é possível observar uma sua medida W_t e que estas estão relacionadas através do modelo para o erro de medida clássico,

[1] Fuller (1987) faz uma análise exaustiva sobre as consequências no modelo de regressão linear.

$$W_t = X_t + u_t \tag{3.1}$$

onde u_i é uma variável aleatória com média nula e variância σ_u^2 independente de X_i e de ε_i. Sob estas condições é bem conhecido o resultado, de que a aplicação do método dos mínimos quadrados produz um estimador inconsistente para β_x, ou seja,

$$Plim\ \widehat{\beta}_{OLS} = \beta_x \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2} = \beta_x \frac{1}{1 + \frac{\sigma_u^2}{\sigma_x^2}} \tag{3.2}$$

Esta expressão permite determinar a origem e a dimensão do enviesamento provocado pela contaminação por erro de medida. O enviesamento será tanto maior, quanto maior a variância de u relativamente à de X_i, e quanto maior a volatilidade de u_i. Esta conclusão retirada da expressão (3.2) está de acordo com a célebre interpretação do efeito atenuação. Neste caso, além do enviesamento a presença de erro de medida tem como consequência, que a variância do erro da regressão naive seja dada por $\sigma_\varepsilon^2 + \frac{\beta_x \sigma_x^4}{\sigma_x^2 + \sigma_u^2}$, o que significa que não só o declive da recta de regressão é atenuado como os dados observados estão mais dispersos em torno da relação linear.

Outra consequência da aplicação do método dos mínimos quadrados nestas circunstâncias é que o enviesamento do $\widehat{\beta}_{OLS}$ deixa de poder ser escrito como uma combinação linear dos erros o que em termos da determinação da distribuição dos estimadores coloca algumas dificuldades.

3.2.2 Regressão Simples com estrutura de erro mais complexa

Contrariamente a algumas interpretações simplistas das consequências da existência de erro de medida, o efeito atenuação é uma característica da hipótese de erro de medida clássico e do tipo de regressores considerados. Quando levantada esta hipótese, permitindo que W_t seja uma medida enviesada de X_t e que exista correlação entre ε_t e u_t os efeitos da contaminação das variáveis explicativas são mais ambíguos. Se $(X_t, u_t, \varepsilon_t)$ tiverem distribuição conjunta normal, então e considerando ainda o modelo de regressão simples, mas onde o modelo erro de medida é dado por $W_t = \alpha_0 + \alpha_1 X_t + u_t$, mantendo-se por isso a estrutura aditiva, as estimativas para a intercepção e declive da regressão de Y_t sobre W_t vêm:

$$\beta_0^* = \beta_0 + \beta_x\mu_x - \beta_x^*(\alpha_0 + \alpha_1\mu_x)$$
e
$$\beta_x^* = \frac{\alpha_1\beta_x\sigma_x^2 + \rho_{\varepsilon u}\sqrt{\sigma_u^2\sigma_\varepsilon^2}}{\alpha_1^2\sigma_x^2 + \sigma_u^2} \tag{3.3}$$

Desta expressão é possível tirar duas ilacções bastante importantes. A primeira é que fica claro que o efeito atenuação é apenas um caso particular dum modelo para o erro de medida mais geral, nomeadamente $\alpha_0 = 0$, $\alpha_1 = 1$ e $\rho_{\varepsilon u} = 0$, ou seja ausência de correlação entre as variáveis residuais do modelo de regressão e do modelo erro de medida. A segunda conclusão que deve retirar-se, é que naquelas condições, não só o termo independente também vem enviesado, como é possível que $|\beta_x^*| > |\beta_x|$, o que significa que o efeito da contaminação seja exactamente o oposto ao efeito atenuação anteriormente considerado.

3.2.3 Regressão Múltipla com uma variável explicativa medida com erro

Neste tipo de formulação os efeitos da contaminação são ainda mais complexos. Considere-se o modelo de regressão linear onde X é um escalar e $\mathbf{Z}$ é uma matriz de dimensão superior a um. O modelo linear é agora dado por,

$$Y_t = \beta_0 + \beta_x X_t + \beta_z^t \mathbf{Z} + \varepsilon \tag{3.4}$$

Carroll, Ruppert & Stefanski (1995) demonstram que sob o modelo erro de medida clássico o estimador dos mínimos quadrados da regressão de Y sobre W e $\mathbf{Z}$ o coeficiente associado à variável medida com erro é consistente para,

$$Plim\ \widehat{\beta}_{wOLS} = \beta_x \frac{\sigma_{x|z}^2}{\sigma_{x|z}^2 + \sigma_u^2} \tag{3.5}$$

Este resultado análogo ao anteriormente obtido que caracterizava o efeito atenuação é igual ao anterior se e só se X e $\mathbf{Z}$ não estiverem correlacionados, visto que nesse caso $\sigma_{x|z}^2 = \sigma_x^2$, ou seja, a variância condicional é igual à variância incondiconal[4] . O efeito sobre a estimativa do parâmetro associado à variável medida com erro é também o de um enviesamento em direcção à origem. Além deste efeito,

[4] $\sigma_{x|z}^2$ é a variância dos resíduos da regressão de X sobre $\mathbf{Z}$.

o enviesamento provocado pelo erro de medida não se restringe ao coeficiente da variável contaminada. Demonstra-se que o estimador naive do método dos mínimos quadrados do coeficiente das variáveis medidas sem erro é consistente para $Plim\ \widehat{\beta_z^*} = \beta_z + \beta_x (1-\lambda)\Gamma_z$, onde Γ_z é o coeficiente da regressão de X sobre $\mathbf{Z}$, i.e., $E(X \mid \mathbf{Z}) = \Gamma_0 + \Gamma_z \mathbf{Z}$ e $\lambda = \sigma^2_{x|z} / \left(\sigma^2_{x|z} + \sigma^2_u\right)$ (ver Carroll, Ruppert & Stefanski, 1995). Logo o efeito sobre os coeficientes das variáveis observáveis sem erro será tanto maior quanto maior fôr o coeficiente daquela relação linear. O enviesamento em β_z depende da dependência linear entre X e $\mathbf{Z}$. As consequências da interpretação da estimativa naive é que o parâmetro pode parecer significativo quando de facto não o é, além de que o seu sinal e dimensão podem estar completamente alterados.

3.2.4 Múltiplas variáveis explicativas medidas com erro

Numa situação em que $Y = \beta_0 + \beta_x^t \mathbf{X} + \beta_z^t \mathbf{Z} + \varepsilon$ onde $\mathbf{X}$ é uma matriz, $\mathbf{W}$ é uma medida não enviesada para as variáveis latentes e $\sum_{ab}$ é a matriz de variâncias e covariâncias entre quaisquer duas variáveis aleatórias. O vector do estimador naive dos mínimos quadrados para os coeficientes dos regressores é,

$$\begin{bmatrix} \beta_x^* \\ \beta_z^* \end{bmatrix} = \begin{bmatrix} \sum_{xx} + \sum_{uu} & \sum_{xz} \\ \sum_{zx} & \sum_{zz} \end{bmatrix}^{-1} \left(\begin{bmatrix} \sum_{xx} & \sum_{xz} \\ \sum_{zx} & \sum_{zz} \end{bmatrix} \begin{bmatrix} \beta_x \\ \beta_z \end{bmatrix} + \begin{bmatrix} \sum_{u\varepsilon} \\ 0 \end{bmatrix} \right)$$

Logo o método do mínimos quadrados mais uma vez produz estimativas claramente enviesadas para o vector de parâmetros da regressão, a menos que $\sum_{uu}$ e $\sum_{u\varepsilon}$ sejam matrizes nulas.

3.3 Métodos Alternativos ao OLS

Como foi visto, o método dos mínimos quadrados é tipicamente enviesado quando estamos perante erro de medida nas variáveis explicativas, sendo a dimensão e direcção do seu enviesamento função do tipo de modelo de regressão considerado e da distribuição do erro de medida. Como tal, o efeito atenuação, na sua interpretação naive, é apenas um caso particular e uma das vertentes dos efeitos causados pela contaminação das variáveis explicativas. Interessa agora considerar outros métodos de estimação que tomem em consideração a existência de erro de medida.

3.3.1 Método das Variáveis Instrumentais

Genericamente a inconsistência do estimador dos mínimos quadrados para os coeficientes da regressão linear (seja qual for a sua estrutura) na presença de regressores contaminados com erro deriva do facto de na regressão das variáveis observadas, os regressores W estarem correlacionados com os erros, i.e, $Y = W\beta + v$ e $Plim\ W'v \neq 0$, onde $v = \varepsilon - u\beta$ e $W = X + u$.

Uma alternativa bastante popular é utilizar informação adicional, que não a contida nas variáveis Y e W e estimar os parâmetros recorrendo ao método das variáveis instrumentais.

Suponha-se que existe um conjunto de variáveis T, que estão correlacionadas com a variável latente X, não correlacionadas com o erro de medida u, e não correlacionadas com o erro da regressão ε. Sobre estas condições podemos escrever:

$$E(Y \mid T) = E(X \mid T)'\beta$$

e

$$E(W \mid T) = E(X \mid T) \tag{3.6}$$

A ideia básica desta técnica é que a regressão de W sobre T revela a relação entre T e o esperado de X dado T, porque a variável instrumental não está correlacionada com o erro de medida. Nestas condições podemos prever $E(X \mid T)$ usando a previsão de W da regressão linear de W sobre T, e depois, numa segunda fase executar a regressão dessa previsão sobre Y. Usando um modelo linear para a relação entre X (e logo W) e T (e.g. $W = T'\gamma + \eta$) e o correspondente estimador dos mínimos quadrados $\widehat{\gamma} = (T'T)^{-1}T'W$ temos,

$$\widehat{E}(X \mid T) = T(T'T)^{-1}T'W \tag{3.7}$$

Fazendo a regressão de Y sobre este valor esperado temos o seguinte estimador do parâmetro β,

$$\widehat{\beta}_{IV} = \left(W'T(T'T)^{-1}T'W\right)^{-1}W'T(T'T)^{-1}T'Y \tag{3.8}$$

Este é o estimador das variáveis instrumentais generalizado apresentado por Sargan (1958). Quando o número de instrumentos, T fôr igual ao número de

variáveis explicativas medidas com erro W do modelo, a expressão do estimador simplifica-se para,

$$\widehat{\beta}_{IV} = (T'W)^{-1} T'Y \tag{3.9}$$

que é o estimador apresentado por Reiersol (1941). Este estimador pode ser generalizado para casos em que a regressão de X sobre T é não linear. Uma possível escolha para instrumento de X é a observação desta variável num estudo independente, enquanto que as variáveis que não estejam contaminadas servirão de instrumentos para elas próprias.

Existem alguns problemas que tornam o método das variáveis instrumentais pouco atraente. O primeiro é que $\widehat{\beta}_{IV}$ pode ser bastante impreciso. O método baseia-se no facto de podermos conhecer a relação entre o instrumento e a variável latente, observando a relação entre T e W. Se a relação entre X e T fôr fraca, então a menos que a amostra seja bastante grande não será possível obter uma estimativa precisa daquela relação, o que por sua vez, pôem em causa a capacidade de estimação da relação entre Y e X. O segundo problema é que nem sempre é fácil encontrar instrumentos. Para que seja considerado um instrumento, T tem de verificar três condições:

$$Plim\frac{T'u}{N} = 0;\ Plim\frac{T'X}{N} \neq 0;\ Plim\frac{T'\varepsilon}{N} = 0 \tag{3.10}$$

A primeira condição diz-nos que a variável instrumental não deve emanar da mesma fonte que produziu a variável medida com erro. Quando não se verifica a segunda condição, ou seja, se o instrumento e a variável latente não estiverem correlacionadas a previsão de $E(X \mid T)$, não reflectirá a variação da variável X não observável, mas será apenas a consequência de uma variação da amostra absolutamente aleatória. Nestas condições o estimador das variáveis instrumentais será claramente inconsistente. A terceira condição diz-nos que, os instrumentos não podem ter um efeito sobre Y para além do efeito provocado pela variável latente, a menos que $E(\varepsilon \mid T)$ seja ortogonal a $E(X \mid T)$. Esta condição é necessária para garantir a consistência da estimação do parâmetro de interesse.

3.3.2 Método dos Momentos

No modelo de regressão linear simples a obtenção de um estimador corrigido para o coeficiente da regressão contaminada passa pelo conhecimento do "rácio de fiabilidade"-$\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ Fuller (1987).

Se fosse possível obter uma estimativa para a variância do erro de medida, $\widehat{\sigma}_u^2$ e se considerarmos $\widehat{\sigma}_w^2$ como sendo a variância da amostra da variável proxy, então uma estimativa consistente para o "rácio de fiabilidade" será $\widehat{\lambda} = (\widehat{\sigma}_w^2 - \widehat{\sigma}_u^2)/\widehat{\sigma}_w^2$. Este algoritmo corresponde a uma estimação pelo método dos momentos, porque tanto o método dos mínimos quadrados como o rácio de fiabilidade dependem apenas dos momentos dos dados observados. O estimador corrigido do declive é obtido dividindo o estimador naive por esta quantidade. As propriedades em amostras finitas deste estimador são bastante fracas. Em amostras pequenas a distribuição de $\widehat{\beta}_{x^*}/\widehat{\lambda}$ é bastante enviesada, o que motivou a consideração de uma versão modificada proposta por Fuller (87).

No caso multivariado e admitindo que as matrizes $\sum_{u\varepsilon}$ e $\sum_{uu}$ são conhecidas, ou podem ser estimadas através da matriz de correlações da amostra e com recurso a réplicas, o estimador do Método dos Momentos vem,

$$\begin{bmatrix} \mathbf{S}_{ww} - \sum_{uu} & \mathbf{S}_{wz} \\ \mathbf{S}_{zw} & \mathbf{S}_{zz} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_{wy} - \sum_{u\varepsilon} \\ \mathbf{S}_{zy} \end{bmatrix} \tag{3.11}$$

onde $\mathbf{S}_{ab}$ é a matriz de variâncias e covariâncias da amostra entre duas variáveis aleatórias. O estimador pode ser simplificado admitindo a hipótese mais usual que $\sum_{u\varepsilon} = 0$, ou seja, que o erro de medida e o erro da regressão não estão correlacionados.

3.3.3 Regressão Ortogonal

O método dos mínimos quadrados pode ser recuperado para produzir uma estimativa corrigida para erro de medida dos parâmetros do modelo. O problema clássico do método dos mínimos quadrados, consiste na estimação de β_0 e β_1, tendo como critério a minimização do quadrado da distância vertical dos pontos (Y_t, X_t) à linha $Y_t = \beta_0 + \beta_1 X_t$.

Na presença de erros de medida cada observação pode desviar-se da recta

de regressão não só de forma vertical, mas também na horizontal. Isto é, os pontos sobre a recta são dados por $(\beta_0 + \beta_1 X_t, X_t)$ e os pontos observados são $(\beta_0 + \beta_1 X_t + \varepsilon_t, W_t = X_t + u_t)$. Desta forma há que minimizar o quadrado da distância euclideana, devidamente standardizada,

$$\sum_{t=1}^{N} \left[(Y_t - \beta_0 - \beta_1 X_t)^2 + \eta \, (W_t - X_t)^2 \right] \tag{3.12}$$

ou o quadrado da distância estatística Fuller(87), onde $\eta = \frac{\sigma_\varepsilon^2}{\sigma_u^2}$ é o único parâmetro desconhecido.

A aplicabilidade deste método é limitada pela necessidade de se conhecer previamente o parâmetro η, que quase sempre é bastante difícil de especificar de forma a garantir a consistência dos procedimentos estatísticos subsequentes.

Capítulo IV- Modelos não Lineares

4.1 Introdução

A análise das consequências da contaminação por erro de medida é uma questão bastante complexa dada a multiplicidade de especificações que podemos considerar, não só para o modelo base mas também para o modelo para o erro de medida. Na prática todos os tipos de especificações são susceptíveis de serem estimadas com dados contaminados por erro de medida. O seu impacto dependerá do tipo de procedimento estatístico considerado, assim como da distribuição conjunta ou condicional das variáveis subjacentes ao estudo em causa.

No capítulo anterior, foi possível verificar que mesmo no caso do modelo de regressão linear, não havia uma característica única que descrevesse de forma inequívoca os efeitos da presença de erro de medida na estimação do parâmetro de interesse. O chamado efeito atenuação, na sua interpretação mais naive, muitas vezes considerado como a consequência genérica deste problema, é mesmo neste contexto apenas um caso particular e apenas um dos efeitos causados por este problema. Para esta classe de modelos o efeito atenuação, de enviesamento em direcção à origem apenas se verifica quando consideramos:

i) O modelo de regressão linear simples;

ii) O modelo para o erro de medida clássico;

iii) E ausência de correlação entre o erro de medida e o erro da regressão.

Na esmagadora maioria dos casos em que a distribuição de Y é apenas condicional numa variável explicativa, $f(Y \mid X)$ sob a hipótese de erro de medida não diferencial, a relação subjacente entre Y e X é preservada, no sentido em que, a correlação entre Y e W é positiva sempre que $E(Y \mid X)$ e $E(W \mid X)$ são funções crescentes de X (Carroll, Ruppert & Stefanski, 1995). Isto sucede porque com erro de medida não diferencial, Y e W dado X são não correlacionados e logo a covariância entre Y e W é apenas a covariância entre $E(Y \mid X)$ e $E(W \mid X)$ (Weinberg et. al., 1993).

O que este resultado nos indica é que no caso mais simples quando X representa apenas um regressor e não existem mais variáveis explicativas, o sentido

da relação entre os dados observados não é afectada sob a hipótese de erro de medida não diferencial. Contudo este resultado ilustra apenas uma parte dum problema mais complexo. Ele apenas descreve a correlação entre Y e W não dando qualquer indicação sobre a manutenção da estrutura da relação original.

Hwang & Stefanski (1994) demonstraram que à excepção do modelo de regressão linear com erro normal, se $E(Y \mid X)$ fôr crescente em X e caso se verifique $W = X + u$ com u independente de X e Y, então o $E(Y \mid W)$ não é necessariamente crescente em W. Isto constitui um problema sério visto a inferência sobre a relação de Y e X baseada na relação entre Y e W ser em geral enganadora.

Existem várias razões pelas quais não se deve ignorar que a presença de erro de medida pode acarretar graves problemas. Primeiro porque, embora por vezes consigamos estimar a direcção da relação entre Y e X o enviesamento em termos de magnitude dos parâmetros pode ser muito grande. Segundo, a discussão anterior restringe-se ao caso em que Y é condicional apenas numa variável explicativa. Com a introdução de mais variáveis explicativas medidas com ou sem erro, o "trend" entre as variáveis pode-se alterar e sendo assim a utilidade da estimação de $f(y \mid z, w)$ é nula. Finalmente, é possível quando estamos perante várias variáveis explicativas através de uma correcta modelização da estrutura do erro aumentar a potência da inferência.

4.2 Erro de Medida em Modelos Não Lineares: Uma abordagem geral

As metodologias estatísticas conhecidas do investigador, têm como objectivo estudar o comportamento de uma dada variável, tendo em conta a existência de um determinado conjunto de variáveis e parâmetros a elas associados, que de algum modo condicionam ou determinam o seu comportamento. O carácter da variável dependente, assim como a relação com o denominado conjunto de informação, determinam a forma concreta da especificação estatística que se pretende estimar. Contudo, seja qual fôr a especificação considerada, na sua generalidade, o problema da estimação de um modelo paramétrico pode muitas vezes colocar-se em termos da tentativa de identificação do $E(y_t \mid \Omega_t)$ onde Ω_t representa um dado conjunto de informação. O conjunto de informação assume, qualquer que seja a modelização considerada para o valor esperado condicional, a forma de

uma função de regressores e parâmetros que são geralmente representados respectivamente pela matriz X_t de dimensão $(n \times k)$ e o vector β de dimensão $(k \times 1)$. Sendo assim, uma especificação muito geral para o probema genérico da estimação do comportamento de uma dada variável pode ser escrito como:

$$E(y_t \mid \Omega_t) = E(y_t \mid X_t) = F(g(X_t, \beta)) \tag{4.1}$$

onde $F(\cdot)$ representa o modelo estatístico ou probabilístico em causa, que é (nos casos aqui considerados) uma transformação não linear de $g(\cdot)$, a função index, que determina como regressores e parâmetros de interesse se conjungam na especificação paramétrica. Dentro desta especificação genérica é comum considerar-se que a função index assume a forma de uma combinação linear de regressores e parâmetros, resultando numa expressão do tipo,

$$E(y_t \mid X_t) = F(X_t\beta) \tag{4.2}$$

Dentro desta formulação geral para modelos não lineares, interessa considerar quais os efeitos genéricos produzidos pela presença de erro de medida na especificação e estimação daquele esperado condicional. Mais particularmente, vai-se restringir a análise ao modelo para o erro de medida dado por,

$$W_t = X_t + u_t \tag{4.3}$$

que representa o modelo clássico. A variável verdadeiramente observada, denominada variável proxy, é igual à variável latente mais um erro, que aqui se supõe ter média nula e variância constante, e independente da verdadeira variável. A utilização naive da variável observável produz estimativas inconsistentes para os parâmetros de interesse. Esta afirmação deriva do facto de, dado que o modelo para o erro de medida é verdadeiro e a especificação correcta é dada pela expressão 4.2, a utilização da variável proxy implica que se devesse considerar a nova especificação,

$$E(y_t \mid X_t) = F(W_t\beta - v_t) \tag{4.4}$$

dado que β é um vector de constantes pode escrever-se, $v_t = u_t\beta$, tendo v_t as mesmas propriedades que u_t. Esta especificação comporta dois problemas sérios

que estão na origem da inconsistência provocada pela contaminação. De modo a manter-se a especificação, a utilização de W_t na regressão, só seria possível mediante o conhecimento do vector v_t. Dada a incapacidade de o observarmos, estamos perante um problema típico de variável omitida. Se o objectivo fosse estimar o parâmetro, apenas poderíamos ignorar o problema se W_t e v_t fossem ortogonais, contudo em virtude da especificação do modelo para o erro de medida tal não se verifica e logo não é possível estimar consistentemente o parâmetro β pela regressão naive.

O problema da dependência dos regressores desta expressão pode ser corrigido se alterarmos a função index, de modo a garantir ortogonalidade entre regressor e erro, tendo em conta a dependência entre W_t e u_t. Seja esta dependência dada por $E(u_t \mid W_t)$, que nestas condições se assume diferente de zero. Manipulando esta expressão é fácil mostrar-se que a expressão acima é equivalente a,

$$E(y_t \mid X_t) = F(Z_t\beta + \xi_t) \tag{4.5}$$

onde $Z_t = W_t - E(u_t \mid W_t)$ e $\xi_t = -(u_t - E(u_t \mid W_t))\beta$. Logo por construção é fácil concluir que $E(Z\xi \mid W) = 0$. Esta reinterpretação do problema, permite-nos observar outras consequências da presença de erro de medida. A primeira diz respeito à forma funcional da função index. A questão que se coloca é que embora na especificação inicial do modelo, o valor esperado era condicional numa combinação linear dos regressores com vector de parâmetros, a introdução do $E(u_t \mid W_t)$ faz com que na generalidade dos casos a linearidade da função index não se mantenha. Isto quer dizer que devemos escrever a expressão anterior como,

$$E(y_t \mid X_t) = F(h(W_t)\beta + \xi_t) \tag{4.6}$$

A dimensão deste problema assim como a forma da função $h(\cdot)$ dependerá da distribuição conjunta de u_t e W_t. Um caso particular, que tem interesse analisar dá-se quando condicional em W_t o erro de medida tem distribuição normal. Neste caso, é fácil mostrar que $h(W_t)$ ainda é uma função linear em W_t, embora os coeficientes que pré-multiplicam W_t não são idênticos aos da especificação naive. Além disto, a reespecificação do modelo considerada na expressão (4.6) encerra outro problema de maior gravidade. De acordo com aquela expressão,

a função index passa a ser a soma de uma função dos regressores observados e dos parâmetros e de uma variável aleatória ξ_t com uma dada função densidade probabilidade. Esta nova componente aleatória, e mais particularmente a sua distribuição põe em risco a manutenção da forma funcional do modelo base, ou seja, põe em causa a própria forma da expressão $F(\cdot)$. Isto porque se quisermos utilizar aquela reespecificação condicionada apenas em W_t, temos de marginalizar a função $F(h(W_t)\beta + \xi_t)$ em ordem a ξ_t. Supondo que conhecemos a sua distribuição, a resolução desse integral vai quase sempre alterar a especificação do modelo para a distribuição de Y_t condicional em X_t.

4.3 Aproximação para Variância do erro pequena

4.3.1 Introdução

A presença de erro de medida tem como principal efeito, fazer a distribuição que gera os dados diferir da distribuição da variável explicativa de interesse. Um argumento muito geral, leva-nos a esperar que o erro de medida provoque um atenuar do efeito das variáveis explicativas sobre a variável dependente. O erro de medida faz com que as variáveis explicativas pareçam ter uma amplitude de variação superior à que de facto têm. Desta forma altera a escala de variação dos regressores, não tendo contudo nenhum efeito sobre a variável explicada. Consequentemente, o efeito sobre a variável dependente de uma variação marginal da variável contaminada com erro, é inferior ao efeito da mesma variação da verdadeira variável explicativa. Este argumento simples, é verdadeiro para o modelo de regressão linear simples, contudo a questão que se levanta, dada a complexidade do problema é saber se este argumento também se aplica de forma tão simples, à classe dos modelos não lineares.

Quando observamos W ao invés de X, a análise estatística formal e informal dá-nos informação sobre a função densidade $f_W(W)$. Isto quer dizer que histogramas e outros instrumentos de análise descritiva, dão-nos informação sobre a forma de $f_W(W)$ e não de $f_X(X)$, a densidade de interesse e todos os procedimentos estatísticos dão-nos informação acerca da distribuição que gera os dados observados. Por isso, a análise estatística afectada pelo erro de medida abrange não apenas os procedimentos formais, como a análise de regressões, mas também

procedimentos informais, onde se inclui a análise gráfica.

Como existe um vasto conjunto de procedimentos estatísticos no universo de análise, importa partir de uma base comum e determinar primeiramente o impacto do erro de medida sobre as distribuições dos dados disponíveis ao analista. Desta forma o conhecimento das consequências da existência de erro de medida, necessita que se analise o efeito da contaminação nas funções densidade probabilidade de interesse. Para tal, a análise deve centrar-se não nas propriedades dos métodos de estimação e no impacto sobre uma classe específica de modelos paramétricos, mas sim na análise das distorções introduzidas nas funções distribuição que definem genericamente uma modelização paramétrica.

Esta abordagem sendo mais genérica é particularmente útil, visto a tipologia dos efeitos da contaminação, divergir consoante o modelo paramétrico que governa a relação entre Y e X, e dentro de um tipo de modelo, das características do vector das variáveis explicativas, como ficou demonstrado para o caso simples da regressão linear. Utilizando esta metodologia uma grande classe de modelos é susceptível de ser examinada. O impacto da existência de erro de medida será aferido através do recurso a uma aproximação para variância pequena às distribuições associadas às variáveis que tenham sido contaminadas pelo erro de medida.

Este tipo de metodologia é particularmente útil porque, as aproximações fornecem informação qualitativa bastante importante acerca dos efeitos do erro de medida sobre as densidades probabilidade. Elas são bastante precisas quando a variação devido ao erro de medida é pequena relativamente às outras fontes de variação do modelo e não dependem da forma precisa da distribuição do erro de medida, o que é de extrema utilidade visto que quase nunca conhecemos essa especificação. Contudo ela depende do conhecimento da distribuição da verdadeira variável explicativa, o que pode na maioria dos casos pôr entraves à operacionalização desta forma de abordar o problema. Esta metodologia permite ainda reconhecer as consequências que são comuns a muitos procedimentos e tirar conclusões mais genéricas acerca deste problema, e que de outra forma seriam negligenciados. Apesar dos méritos desta abordagem, deve-se sempre ter em conta que os resultados dela resultantes são apenas aproximações.

4.3.2 Aproximação para erro nas variáveis explicativas

Sejam Y e X um vector de variáveis aleatórias de componentes Y_t $(t = 1, ..., N)$ e X_t $(t = 1, ..., N)$ com distribuição conjunta dada por $f_{Y|X}(y \mid x) f_X(x)$ onde y e x são possíveis realizações das variáveis Y e X. Enquanto Y pode ter elementos discretos na sua distribuição, assume-se que X é uma variável contínua em todo o seu domínio. Assume-se também que o domíno de $f_{Y|X}(y \mid x)$ é independente de x.

A metodologia proposta e desenvolvida por Chesher (1991) pretende analisar uma situação em que se quer obter informação sobre $f_{Y|X}(y \mid x)$ e $f_X(x)$ quando não é possível observar as realizações da variável aleatória X. O que nos é dado a observar é uma variável proxy, W com componentes $W_{it} = X_{it} + \sigma_i u_{it}$ $(\sigma_i \geq 0,\ i = 1, ..., k)$. Assume-se que as variáveis aleatórias u_i, componentes do vector u são continuamente distribuídas, independentes de Y e X, com média nula, variância unitária e função densidade conjunta $f_u(u)$.

Às realizações de $\sigma_i u_{it}$ corresponde o erro de medida, às realizações da variável X_t corresponde a variável latente não observada, e às realizações de W_t corresponde a variável contaminada com erro de medida e que é efectivamente observada. Outra forma de caracterizar a relação entre estas variáveis é dizer que estamos perante o modelo para o erro de medida clássico.

Sob estas condições a aproximação de ordem $O(\sigma^2)$ à distribuição conjunta de Y e W, que constitui o ponto de partida desta análise, (ver Chesher 1990, para mais detalhes) é dada por,

$$f_{YW}(y, w) = f_{Y|X}(y \mid w) f_X(w) + \\ + \tfrac{1}{2}\sigma_{ij}\left[f_{Y|X}(y \mid w) f_X^{ij}(w) + 2 f_{Y|X}^{i}(y \mid w) f_X^{j}(w) + f_{Y|X}^{ij}(y \mid w) f_X(w)\right] + o(\sigma^2) \tag{4.7}$$

onde $f_X^j(w)$ e $f_{Y|X}^{ij}$ são respectivamente as primeira e segunda derivadas da densidade de X em ordem ao seu i-ésimo argumento avaliadas em $X = w$. De imediato se conclui que a aproximação para variância pequena a esta distribuição (e as aproximações a distribuições marginais e condicionais a ela associada) não dependem da forma funcional da distribuição do erro de medida. Contudo elas dependem intimamente da distribuição marginal da variável latente.

Neste contexto, interessa conhecer duas distribuiçoes associadas a $f_{YW}(y, w)$.

A primeira é a distribuição marginal da variável explicativa observada $f_W(w)$. O interesse no conhecimento desta distribuição reside no facto de muito frequentemente ser necessário realizar inferência sobre aspectos da distribuição marginal da variável cujas realizações são observadas com erro, além de que é da utilização destas, enquanto variáveis explicativas, que surgem os problemas de especificação das distribuições condicionais. Integrando $f_{YW}(y,w)$ em ordem a y e tendo em conta que $\int f_{Y|X}(y \mid w)\,dy = 1$ e logo $\int f^{i}_{Y|X}(y \mid w)\,dy = \int f^{ij}_{Y|X}(y \mid w)\,dy = 0$, chegamos à seguinte expressão:

$$f_W(w) = f_X(w) + \frac{1}{2}\sigma_{ij} f^{ij}_X(w) + o\left(\sigma^2\right) \tag{4.8}$$

Esta aproximação fornece informação qualitativa bastante importante para o estudo das consequências da presença de erro de medida. Contudo, não pode ser utilizada como um modelo que possibilite efectuar uma análise estatística formal dos dados contaminados dado que pode não representar uma função densidade. O problema é que quando os σ_{ij}'s são significativamente diferentes de zero e o termo $f^{ij}_X(w)$ é negativo a aproximação (4.8) pode assumir um valor negativo. Este problema pode ser ultrapassado mediante uma alteração da expressão (4.8) que a faz obedecer ao critérios que definem uma função densidade, (ver Chesher, 1990).

A aproximação assim definida, permite-nos explorar as consequências sobre a densidade dos dados observados contaminados com erro de medida. Seja $\sum = [\sigma_{ij}]$ a matriz definida positiva de variâncias e covariâncias do erro de medida, σu, e seja $f^{(2)}_X = \left[f^{ij}_X(w)\right]$ a Hessiana da densidade conjunta da variável latente. Sendo assim o termo $\frac{1}{2}\sigma_{ij} f^{ij}_X(w)$, pode ser reescrito como $\frac{1}{2} tr\left(\sum f^{(2)}_X\right)$. Em zonas onde f_X é côncava, esta expressão é negativa e em zonas em que f_X é convexa a ela é positiva. Consequentemente, o efeito de primeira ordem da presença de erro de medida, é o de elevar a densidade de X onde ela é convexa e de a baixar onde ela é côncava. Na maior parte dos casos de interesse onde as densidades são unimodais (exceptuando a distribuição Uniforme), elas são côncavas perto da moda e convexas nas abas, de modo que o efeito do erro de medida é o de "alisar" a curva da densidade da variável. Como consequência, as irregularidades da curva são diminuídas (o que equivale a dizer que se perde informação) e a sua dispersão aumenta visto introduzir-se exogenamente um elemento adicional de

variabilidade.

Um caso particular de interesse é quando f_X é a densidade multivariada normal com média μ e variância Ω. Nestas condições e utilizando uma aproximação equivalente a (4.8), i.e, da mesma ordem, é fácil mostrar que,

$$f_W(w) \propto \exp\left\{-\frac{1}{2}(w-\mu)'\Omega^{-1}(\Omega-\sum)\Omega^{-1}(w-\mu)\right\} + o\left(\sigma^2\right) \qquad (4.9)$$

Neste caso particular a aproximação à densidade de W é caracterizada pela manutenção da forma funcional da distribuição multivariada normal, reparametrizada para μ e $\Omega(\Omega-\sum)\Omega^{-1}$. Obviamente quando $\sum = 0$ as duas distribuições coincidem, caso contrário o efeito de primeira ordem da presença de erro de medida é o de aumentar a dispersão da distribuição normal em torno da mesma média.

Interessa ainda analisar o efeito da contaminação nos momentos de maior relevância. Suponha-se que se pretende calcular a média de uma função dos dados que estão contaminados com erro de medida. Esta média dá-nos informação enviesada sobre o primeiro momento da variável proxy, mas se se ignorar a presença do erro de medida, pode ser incorrectamente utilizada como uma estimativa do primeiro momento da variável latente. Utilizando a mesma metodologia, ou usando a aproximação definida em (4.8), é possível determinar a relação entre $E_W\{g(W)\}$ e $E_X\{g(X)\}$ para uma dada função $g(.)$ dos dados e considerando o modelo erro de medida clássico. Esta aproximação é dada por,

$$\begin{aligned} E_W\{g(W)\} &= E_X\{g(W)\} + \tfrac{1}{2}\sigma_{ij}E_X\{g^{ij}(W)\} + o(\sigma^2) \\ &= E_X\{g(W)\} + \tfrac{1}{2}tr\left[\sum E_X\left\{g^{(2)}(W)\right\}\right] + o(\sigma^2) \end{aligned} \qquad (4.10)$$

onde a matriz $g^{(2)}(W)$ é a Hessiana da função $g(W)$. Se esta função fôr linear, então, e como seria de esperar o efeito da presença do erro de medida é nulo. Quando $g(W)$ é uma função convexa, como o caso em que estimamos variâncias e outros momentos de ordem superior, a utilização dos dados contaminados conduz a um enviesamento positivo, quando $g(W)$ é côncava, como quando calculamos médias logarítmicas o enviesamento é negativo.

A outra distribuição cujo o efeito da contaminação interessa considerar, é a densidade de Y condicional em W, $f_{Y|W}(y \mid w)$ já que na maioria das situações a análise econométrica preocupa-se com a estimação do efeito de um conjunto de

variáveis sobre a distribuição de outro conjunto de variáveis. A aproximação é dada por:

$$f_{Y|W}(y \mid w) = f_{Y|X}(y \mid w) + \frac{1}{2}\sigma_{ij}\left[2f^{i}_{Y|X}(y \mid w) F^{j}_{X}(w) + f^{ij}_{Y|X}(y \mid w)\right] + o\left(\sigma^2\right) \tag{4.11}$$

ou

$$\begin{aligned} f_{Y|X}(y \mid w) \Big[1 + \tfrac{1}{2}\sigma_{ij}\Big[2F^{i}_{Y|X}(y \mid w) F^{j}_{X}(w) + F^{ij}_{Y|X}(y \mid w) + \\ + F^{i}_{Y|X}(y \mid w) F^{j}_{Y|X}(y \mid w)\Big]\Big] + o(\sigma^2) \end{aligned} \tag{4.12}$$

onde $F^{j}_{X}(w)$ é a primeira derivada da log densidade da variável latente relativamente ao seu j-ésimo argumento avaliada em $X = w$. À semelhança do que sucedia anteriormente esta expressão não representa uma função densidade adequada, visto poder para valores de σ_{ij} grandes, assumir valores negativos, (ver Chesher, 1990).

Mais uma vez é importante considerar o caso especial em que condicional no conjunto X a variável Y tem uma distribuição univariada normal com uma função de regressão linear, $E(Y \mid X = x) = x'\beta$ e função cedástica constante $Var(Y \mid X = x) = \omega^2$. Fazendo recurso à aproximação (4.12) e depois de algumas simplificações obtêm-se,

$$\begin{aligned} f_{Y|W}(y \mid w) \propto \exp\left[-\tfrac{1}{2\omega^2}\left(1 - \tfrac{\beta'\sum\beta}{\omega^2}\right)\left\{y - w'\beta - \tfrac{\omega^2}{\omega^2 - \beta'\sum\beta} F^{(1)}_{X}(w)'\sum\beta\right\}^2\right] \\ + o(\sigma^2) \end{aligned} \tag{4.13}$$

Da análise desta expressão conclui-se que condicional na variável contaminada W, Y tem até à ordem considerada uma distribuição normal com função cedástica constante, mas com uma função regressão que pode não ser linear. A eventual não linearidade surge devido à presença na aproximação do vector $F^{(1)}_{X}(w)$ na função regressão. No caso especial em que a variável latente também segue uma distribuição multivariada normal, aquela aproximação tem uma função regressão linear. Perante esta situação o único efeito da presença do erro de medida é a alteração dos coeficientes da regressão e da variância condicional. Contudo a forma funcional da distribuição condicional e a linearidade da regressão permanecem inalteradas.

Pode ainda aplicar-se esta aproximação a outras especificações, como os modelos de escolha discreta mais populares, Logit e Probit.

Considere-se o modelo Probit em que a distribuição de Y condicional em X segue a distribuição:

$$f_{Y|X}(y \mid x) = \Phi(x'\beta)^y (1 - \Phi(x'\beta))^{1-y} \tag{4.14}$$

onde $\Phi(\cdot)$ é a função distribuição normal. Utilizando a aproximação definida em (4.12) obtem-se a seguinte expressão:

$$f_{Y|W}(y \mid w) = \left[\Phi\left[\frac{\left(w' + F_X^{(1)}(w)'\sum\right)\beta}{\sqrt{1+\beta'\sum\beta}}\right]\right]^y \left[1 - \Phi\left[\frac{\left(w' + F_X^{(1)}(w)'\sum\right)\beta}{\sqrt{1+\beta'\sum\beta}}\right]\right]^{1-y} \tag{4.15}$$

Neste caso, a estrutura do modelo Probit mantem-se aplicável na presença de erro medida pequeno, mas o efeito sobre as variáveis na função "index" do Probit, novamente introduz não linearidade nos regressores. Mais uma vez no caso em que a distribuição da variável latente é normal a função index é linear embora o parâmetro estimado seja enviesado para β.

No modelo Logit em que a distribuição condicional de Y dado X é dada por

$$\begin{gathered} f_{Y|X}(y \mid x) = p(x'\beta)^y (1 - p(x'\beta))^{1-y} \\ \text{com } p(a) = \exp(a)/(1+\exp(a)) \end{gathered} \tag{4.16}$$

a aproximação conduz a

$$\begin{gathered} f_{Y|W}(y \mid w) = p(w'\beta)^y (1 - p(w'\beta))^{1-y} \times \\ \left[1 + (y - p(w'\beta)) F_X^{(1)}(w)'\sum\beta + \tfrac{1}{2}\beta'\sum\beta\left[(y - p(w'\beta))^2 - p(w'\beta)(1 - p(w'\beta))\right]\right] \\ +o(\sigma^2) \end{gathered} \tag{4.17}$$

Algebricamente não é possível manipular esta expressão de modo a obter-se uma estrutura Logit, mas a semelhança entre os dois modelos (Logit e Probit) permite-nos afirmar que as consequências serão bastante semelhantes.

Como anteriormente, interessa considerar as consequências do erro de medida nos momentos condicionais. Seja $g(Y)$ uma função de Y cujo valor esperado

condicional em X se pretende conhecer, $E_{Y|X}\{g(Y) \mid x\}$. Quando apenas temos disponíveis observações da variável contaminada com erro, a relação entre $g(Y)$ e W, é uma distorção da relação de interesse, a regressão entre $g(Y)$ e X. Os procedimentos aplicados às realizações de $g(Y)$ e W dão informação acerca de $E_{Y|W}\{g(Y) \mid w\}$, e dado que variável explicativa está medida com erro, interessa saber qual a relação daquele valor esperado com $E_{Y|X}\{g(Y) \mid x\}$.

Recorrendo novamente a uma aproximação a $E_{Y|W}\{g(Y) \mid W\}$ obtemos a expressão,

$$E_{Y|W}\{g(Y) \mid w\} = E_{Y|X}\{g(Y) \mid w\} + E^{(1)}_{Y|X}\{g(Y) \mid w\} \sum F^{(1)}_X(w) + \frac{1}{2} tr\left[\sum E^{(2)}_{Y|X}\{g(Y) \mid w\}\right] + o(\sigma^2) \tag{4.18}$$

Quando a regressão de $g(Y)$ sobre X é linear, o último termo desta aproximação desaparece e o enviesamento é apenas dado pelo segundo termo. Contudo este depende de $F^{(1)}_X(w)$. Se além da linearidade suposermos que X segue uma distribuição normal de modo que $F^{(1)}_X(w) = 0$ então $E_{Y|W}\{g(Y) \mid w\} = E_{Y|X}\{g(Y) \mid w\}$ até à ordem considerada. Isto acontece quando o valor de W corresponde à moda da distribuição de X. Para os outros valores de W, $E_{Y|W}\{g(Y) \mid w\}$ e $E_{Y|X}\{g(Y) \mid w\}$ podem diferir.

Considere-se o caso em que X é um escalar. Neste caso o impacto do erro de medida é pequeno quando os valores esperados são avaliados perto das modas da densidade de X. Se o valor esperado de $g(Y)$ dado que $X = x$ é crescente (decrescente) em x, então o efeito do erro de medida é aumentar o valor esperado de $g(Y)$ quando está abaixo (acima) da moda de X e de diminuir quando está acima (abaixo) dela. Para valores à esquerda de cada moda da distribuição de X, o escalar $F^{(1)}_X(w)$ é positivo e para valores à direita da sua moda, o escalar é negativo. Quando a distribuição é unimodal o declive da regressão de $g(Y)$ sobre W está mais perto de zero do que o da regressão de $g(Y)$ sobre X. Este é o típico efeito atenuação do erro de medida embora numa versão mais geral e completa.

No caso em que a distribuição de X é normal, o segundo termo da expressão (4.18) é linear o que implica que após contaminação a regressão linear permanece linear, seja qual fôr a distribuição do erro de medida. Quando a distribuição de

X é não linear o mesmo termo da aproximação é responsável pela introdução de não linearidades.

Em modelos de regressão não lineares, a presença de erro de medida tem um efeito atenuação introduzido pelo segundo termo da aproximação (4.18) e um efeito dependente da curvatura da regressão de $g(Y)$ sobre X e que é responsável por um efeito adicional não linear. Para valores de W em que a função $E_{Y|X}\{g(Y) \mid W\}$ é convexa este termo é positivo e o erro de medida tem como consequência um elevamento da regressão de $g(Y)$ sobre W relativamente à regressão de interesse. Para valores em que $E_{Y|X}\{g(Y) \mid W\}$ é côncava o efeito é oposto.

As densidades aproximadas (4.7), (4.8) e (4.12) fornecem a base para construção de funções verosimilhança aproximadas para uma grande variedade de especificações que permitam a existência de erro de medida, desde que se conheça a distribuição da variável latente. Esta aproximação é ainda de extrema utilidade para a construção de testes de hipótese $H_0 : \sigma_{ij} = 0$ para algun(s) i e j de interesse. Como estes testes de especificação apenas utilizam informação acerca da curvatura da função verosimilhança na vizinhança da hipótese nula a função verosimilhança aproximada fornece a mesma estatística de teste que seria obtida pela completa especificação do modelo para o erro de medida.

4.4 Máxima Verosimilhança

4.4.1 Introdução

Esta secção dedica-se à análise do método da máxima verosimilhança nos modelos não lineares. Uma das principais diferenças relativamente aos demais métodos de estimação reside na caracterização do modelo para o erro de medida. Os ditos métodos funcionais baseiam-se em modelos para o erro de medida aditivos (ou multiplicativos) em que a especificação da distribuição da variável latente não é necessária. A modelização funcional usa modelos paramétricos para a distribuição da variável dependente, mas não especifica qualquer hipótese para as variáveis explicativas não observadas. O método da verosimilhança pelo contrário exige uma total especificação paramétrica do modelo, permitindo um maior âmbito de aplicabilidade e podendo a análise estender-se a casos em que a variável explicativa

medida com erro é discreta - má classificação.

Desta forma, os métodos de verosimilhança requerem modelos estatísticos para a distribuição de X por vezes condicional na variável observada. Como estes modelos descrevem a estrutura de X, são denominados modelos estruturais. A análise destas especificações levanta sempre problemas acerca da robustez da estimação e inferência baseada em modelos estruturais para as variáveis não observadas. A questão é que os resultados de tal modelização dependem fortemente da hipótese relativa à distribuição de X. O método da máxima verosimilhança é bastante útil, mas a possível não robustez na inferência causada por má especificação é um problema grave que não deve ser ignorado.

A escolha entre o tipo de modelização a adoptar é essencialmente um acto de fé e coragem. Para uns, os ganhos em eficiência proporcionados por uma modelização estrutural são ofuscados pela necessidade de proceder a uma análise de especificação cuidadosa e muitas vezes bastante consumidora de tempo. Para outros, a análise estatística requer que se utilizem os instrumentos mais ricos ao dispôr do analista, justificando deste modo o recurso às técnicas da máxima verosimilhança.

Interessa considerar três situações de acordo com o tipo de dados que temos ao dispôr:

i) Não observamos X mas existem dados suficientes, internos ou externos, para caracterizar a distribuição de W dado X e Z.

ii) X não é observável mas acredita-se que se verifica o modelo de Berkson.

iii) É possível observar X para um subconjunto da amostra.

Em qualquer das hipóteses, a análise da máxima verosimilhança parte sempre da especificação da densidade de Y dados Z e X representada por $f_{Y|ZX}(y \mid z, x, \theta)$ onde o parâmetro de interesse é o vector θ. Por exemplo se Y tiver distribuição normal com média $\beta_0 + \beta_x x + \beta_z z$ e variância σ^2 então,

$$f_{Y|ZX}(y \mid z, x, \theta) = \sigma^{-1}\phi\left\{(y - \beta_0 - \beta_x x - \beta_z z)/\sigma\right\} \tag{4.19}$$

onde $\phi(u) = (2\pi)^{-1/2}\exp(-.5u^2)$ e o vector $\theta = (\beta_0, \beta_x, \beta_z, \sigma^2)$

4.4.2 Verosimilhança quando não observamos X

Considere-se o caso mais simples em que Y está apenas condicionado em X e a variável verdadeiramente observada é W. Considere-se ainda que não nos é possível observar a verdadeira variável explicativa. A análise do modelo de verosimilhança passa pela determinação da densidade conjunta de Y e W. Considere-se primeiro o caso mais simples em que Y W e X são variáveis discretas, então $\Pr(Y = y, W = w) = \sum_x \Pr(Y = y, W = w, X = x)$. Recorrendo ao Teorema de Bayes é fácil mostrar que,

$$\Pr(Y = y, W = w) = \sum_x \Pr(Y = y \mid W = w, X = x) . \Pr(W = w, X = x)$$

e tendo em conta a hipótese de erro de medida não diferencial prova-se que:

$$\Pr(Y = y, W = w) = \sum_x \Pr(Y = y \mid X = x, \theta) \Pr(W = w, X = x) \qquad (4.20)$$

Esta expressão constitui o ponto de partida para a análise da máxima verosimilhança quando não observamos a variável X mas apenas uma sua medida.

4.4.3 Modelos de Erro

Se a informação disponível nos permite caracterizar a distribuição de W dado X, então estamos perante modelos de erro. Desta forma sabendo que,

$$\Pr(W = w, X = x) = \Pr(W = w \mid X = x) \Pr(X = x)$$

a expressão (4.20) é dada por :

$$\sum_x \Pr(Y = y \mid X = x, \theta) \Pr(W = w \mid X = x) \Pr(X = x) \qquad (4.21)$$

Esta expressão é bastante útil, porque permite-nos decompor o modelo apropriado para uma análise da máxima verosimilhança quando a relação entre a variável latente e a variável proxy é dada pelo conhecimento da distribuição de W dado X. O modelo divide-se em três componentes:

i) O modelo paramétrico de interesse que se assume ser conhecido;

ii) O modelo para o erro de medida sobre o qual conjecturamos sobre a sua especificação;

iii) A distribuição da variável latente[5] .

No caso mais geral em que existe uma segunda medida da variável X, denominada T e variáveis explicativas medidas sem erro Z, a distribuição conjunta de Y W e T dado Z sofre as seguintes pequenas alterações,

$$f_{YWT|Z}(y,w,t \mid z,\theta,\tilde{\alpha}_1,\tilde{\alpha}_2) = \int f_{Y|ZX}(y \mid z,x,\theta)\, f_{WT|ZX}(w,t \mid z,x,\tilde{\alpha}_1)\, f_{X|Z}(x \mid z,\tilde{\alpha}_2)\, d\mu(x) \tag{4.22}$$

onde a notação $d\mu(x)$ indica que os integrais são somas se X é discreta e integrais se contínua. Nesta notação $f_{WT|ZX}(w,t \mid z,x,\tilde{\alpha}_1)$ corresponde ao modelo para o erro de medida estimado. Na maior parte das vezes ele não depende da variável Z. Por exemplo o modelo clássico em que o erro de medida segue uma distribuição normal com variância σ_ε (única componente do vector $\tilde{\alpha}_1$), $f_{WT|ZX}(w,t \mid z,x,\tilde{\alpha}_1)$ é dado por $\sigma_u^{-1}\phi\{(w-x)/\sigma_u\}$, onde $\phi(.)$ é a função densidade normal. A especificação do modelo para o erro de medida não se configura na maioria dos casos muito problemática. Apesar dos problemas de transportabilidade, a utilização de dados externos (se disponíveis) permite muitas vezes especificar com alguma precisão aquela componente.

Como já foi referido o grande problema reside na especificação e estimação de $f_{X|Z}(x \mid z,\tilde{\alpha}_2)$. As dificuldades surgem porque: X não é observável e na maior parte dos casos não é possível transpôr estas distribuições, visto que a estudos diferentes estão geralmente associadas distribuições diferentes para a variável não observável. Carroll, Ruppert & Stefanski (1995) e Davidian & Gallant, (1993) apresentam algumas sugestões para a estimação destas distribuições.

A verosimilhança desta especificação é o produto da expressão (4.22) para os valores da amostra.

Na maior parte dos casos interessa conhecer a distribuição de Y condicional em W T e Z, $f_{Y|ZWT}(y \mid z,w,t,\theta)$. Esta obtêm-se dividindo a expressão (4.22) pelo seu integral ou soma em relação aos valores de y. Esta densidade é de extrema importância porque permite-nos analisar estatisticamente a média e variância

[5] A obrigatoriedade de especificação desta expressão constitui o grande problema quando utilizamos os métodos da verosimilhança e é a principal causa da eventual fraca robustez da estimação e inferência sobre os parâmetros.

condicionais de Y dado (Z, W, T), ou seja, fazer inferência estatística nos moldes tradicionais.

4.4.4 O Modelo de Berkson

Se a informação disponível é de tal forma que permita construir um modelo para o erro de medida em que se modeliza a distribuição de X dado W através do modelo de Berkson, então teremos $X = W + u$. Geralmente assume-se que u, o erro de medida, segue uma distribuição normal independente de W, com média nula e variância σ_u^2. Esta especificação tem a particularidade de, no caso do modelo de regressão linear a análise naive, que ignora a presença de contaminação, permitir uma inferência correcta sobre os parâmetros da regressão, visto que $E(X \mid W) = W$ (Berkson, 1950). Sendo assim, o modelo de Berkson com erro homocedástico conduz a estimativas consistentes do termo independente em modelos loglineares (ver Carroll, Ruppert & Stefanski, 1995).

Na análise da verosimilhança perante este modelo para o erro de medida a expressão (4.20), da distribuição conjunta de Y e W, $\Pr(Y = y, W = w)$ passa a,

$$\sum_x \Pr(Y = y \mid X = x, \theta) \Pr(X = x \mid W = w) \Pr(W = w) \tag{4.23}$$

Admitindo sempre a hipótese de não diferenciabilidade, podemos dividir ambos os membros da igualdade por $\Pr(W = w)$ e assim, obter a distribuição de Y condicional em W. No caso mais geral, em que se admite também que as distribuições sejam contínuas temos,

$$f_{Y|ZW}(y \mid z, w, \theta, \tilde{\gamma}) = \int f_{Y|XZ}(y \mid x, z, \theta) f_{X|W}(x \mid w, \tilde{\gamma}) \, d\mu(x) \tag{4.24}$$

onde de acordo com a hipótese formulada sobre a distribuição do erro de medida $f_{X|W}(x \mid w, \tilde{\gamma})$ é dada por $\sigma_u^{-1} \phi \{(x - w)/\sigma_u\}$ e o vector γ corresponde ao parâmetro σ_u^2. Mais uma vez a verosimilhança desta especificação é o produto da expressão (4.24) para os valores da amostra. Em alguns casos, como o modelo de regressão linear não é possível identificar todos os parâmetros, enquanto que nos modelos não lineares a identificação é sempre possível através daquela densidade condicional (Carroll, Ruppert & Stefanski, 1995).

4.4.5 Verosimilhança quando X é parcialmente observado

A interpretação clássica do problema das variáveis omitidas diz-nos que, algumas variáveis não podem ser observadas para toda a amostra. O problema do erro de medida é caracterizado pela existência de um conjunto de variáveis explicativas, a que chamamos X, que nunca são observadas, i.e, estão sempre omitidas. Sendo assim, o problema do erro de medida é uma caso extremo de variáveis omitidas, em que nunca observamos uma variável mas temos informação suplementar, sobre a forma de uma variável proxy a que chamamos W e por vezes dum instrumento T.

Em casos em que a variável X é observada para um subconjunto da população é necessário proceder a uma correcção ao modelo de Berkson. Através dos dados observados, é possível modelizar a distribuição de X condicional em Z e W. Seja $f_{X|ZW}(x \mid z, w, \tilde{\gamma})$ essa distribuição estimada. Sendo assim a densidade de Y condicional em Z e W, para o conjunto da amostra onde é possível observar X é dada por,

$$f_{Y|ZW}(y \mid z, w, \theta, \tilde{\gamma}) = \int f_{Y|XZ}(y \mid x, z, \theta) f_{X|ZW}(x \mid z, w, \tilde{\gamma}) \, d\mu(x) \qquad (4.25)$$

Seja esse subconjunto indexado por $\Delta_t = 1$, sendo a restante amostra em que apenas observamos uma medida de X, indexada por $\Delta_t = 0$. Urge referir que para que se possa elaborar este raciocínio à que impôr certas hipóteses relativamente ao mecanismo de observabilidade de X: X tem de estar aleatoriamente omitido, ou seja, a sua observabilidade depende apenas dos valores de (Y, Z, W) e não dos próprios valores de X; e temos de assumir que a probabilidade de observarmos X é $\pi(Y, Z, W)$.

Sob esta hipótese podemos escrever a verosimilhança da amostra como,

$$\prod_{t=1}^{n} \left[\left\{ f_{Y|XZ}(y_t \mid x_t, z_t, \theta) f_{X|ZW}(x_t \mid z_t, w_t, \tilde{\gamma}) \right\}^{\Delta_t} \times f_{Y|ZW}^{1-\Delta_t}(y_t \mid z_t, w_t, \theta, \tilde{\gamma}) \right] \qquad (4.26)$$

4.5 Modelo Poisson

Considere-se o modelo de regressão Poisson.

$$\Pr(Y_t = j \mid X_t) = \exp[-\lambda(X_t\beta)]\frac{\lambda(X_t\beta)^j}{j!},\ t = 1, ..., N \qquad (4.27)$$

onde j é um inteiro não negativo, X_t é um vector de variáveis explicativas, Y_t é a variável que condicional em X_t segue uma distribuição Poisson, β é o vector de parâmetros e $\lambda(X_t\beta)$ é o parâmetro associado a esta distribuição. Devido à característica dos primeiro e segundo momentos específicos desta distribuição $E(Y_t \mid X_t) = Var(Y_t \mid X_t) = \lambda(X_t\beta)$. De modo a garantir que $\lambda(X_t\beta)$ seja uma função das variáveis explicativas monótona e positiva, assume-se que $\lambda(X_t\beta) = \exp(X_t\beta)$.

Considere-se ainda que só é possível observar uma medida da verdadeira variável explicativa contaminada por erro de medida e que este segue o modelo clássico.

$$W_t = X_t + u_t \text{ onde } E(u_t) = 0 \text{ e } E\left(uu^t\right) = \sum \qquad (4.28)$$

onde W_t é a matriz com as observações das variáveis explicativas contaminadas com erro de medida, X_t é a verdadeira matriz das variáveis explicativas e u_t é o erro de medida e $\sum (k \times k)$ é a matriz definida positiva de variâncias e covariâncias do erro de medida.

Nestas condições pretende-se analisar as consequências da utilização da variável proxy no modelo de regressão Poisson, ou seja, averiguar das implicações de condicionar a variável dependente apenas no conjunto de variáveis explicativas observadas.

Se o verdadeiro modelo fôr o que descreve a distribuição de Y_t condicional em X_t então a especificação correcta condicionalmente em W_t e u_t deveria ser,

$$\Pr(Y_t = j \mid W_t, u_t) = \exp[-\lambda(W_t\beta - u_t\beta)]\frac{\lambda(W_t\beta - u_t\beta)^j}{j!} \qquad (4.29)$$

Contudo, como não observamos o erro de medida, o analista apenas condiciona

em W_t. Sendo assim, o modelo relevante é definido por:

$$\Pr(Y_t = j \mid W_t) = E_{u|W}\left[\exp\left[-\lambda(W_t\beta - u_t\beta)\right]\frac{\lambda(W_t\beta - u_t\beta)^j}{j!}\right] = \tag{4.30}$$

$$= \int \exp\left[-\lambda(W_t\beta - u_t\beta)\right]\frac{\lambda(W_t\beta - u_t\beta)^j}{j!} f_{u|W}(u_t \mid W_t)\, du_t$$

Dado que u_t e W_t são não independentes[6] o conhecimento de $f_{u|W}(u_t \mid W_t)$ é fundamental para a construção do modelo operacional.

Uma forma de ultrapassar o problema da necessidade de especificar totalmente a distribuição do erro condicional nas variáveis explicativas passa pela reespecificação de $\lambda(X_t\beta)$ como $\lambda(Z_t\beta + \xi_t)$ onde $Z_t = W_t - E(u_t \mid W_t)$ e $\xi_t = -[u_t - E(u_t \mid W_t)]\beta$. Desta forma definidos os regressores deste modelo Poisson transformado são por construção ortogonais ao novo erro, ou seja, $E(Z\eta \mid W) = 0$ e o termo residual, ξ_t tem valor esperado nulo (ver secção 4.2). O problema da contaminação por erro de medida, reparametrizado desta forma, passa a ser intepretado como os efeitos individuais característicos da heterogeneidade negligenciada, o que equivale a ter um modelo de regressão Poisson composto. Nestas condições, a omissão de ξ_t do conjunto das variáveis condicionantes provoca sobredispersão, i.e, $Var(Y_t \mid Z_t) > E(Y_t \mid Z_t)$, (Santos Silva, J.M. & Andrade e Silva, 1994).

Apesar deste desenvolvimento a sua aplicabilidade está condicionada pelo conhecimento do $E(u_t \mid W_t)$[7] e de sabermos se esta expressão e por conseguinte a nova especificação de $\lambda(\cdot)$ para os dados observados continua a ser uma função linear da variável explicativa. À semelhança do proposto na secção 4.2, suponha-se o caso particular em que u_t e W_t têm distribuição conjunta normal, com matrizes de variâncias e covariâncias $\sum_{uu}$ e $\sum_{XX} + \sum_{uu}$ respectivamente. Neste caso teremos:

$$\begin{aligned} E(u_t \mid W_t) &= E(u_t) + \left(I + \sum_{XX}\sum_{uu}^{-1}\right)^{-1}(W_t - E(W_t)) \\ &= \left(I + \sum_{XX}\sum_{uu}^{-1}\right)^{-1}(W_t - E(W_t)). \end{aligned} \tag{4.31}$$

[6] Ao contrário do que acontece no caso da heterogeneidade negligenciada onde u_t representa o vector dos efeitos individuais.

[7] Que é uma hipótese menos restritiva do que supor o conhecimento de $f_{u|W}(u \mid W)$.

O argumento da função $\lambda(\cdot)$ sob esta hipótese vem igual a W_t mais uma função linear de W_t[8] e um termo independente com valor esperado nulo. Contudo, na generalidade dos casos não é possível garantir a linearidade do argumento da função $\lambda(\cdot)$, de maneira que persistir numa especificação linear para os dados contaminados por erro de medida resultará em má especificação do modelo.

Ainda no caso especial da normalidade na distribuição conjunta de u_t e W_t outro problema surge. Ainda que a especificação permaneça linear, os coeficientes que prémultiplicam W_t não são iguais aos coeficientes da especificação isenta de erro de medida, o que corresponderá a um previsível enviesamento dos parâmetros estimados.

Da observação do modelo condicional em u_t e em W_t patente na expressão (4.29) é possível identificar a alteração no modelo Poisson provocada pelo erro de medida clássico, em termos dos primeiro e segundo momentos e logo, a origem da inconsistência do estimador da máxima verosimilhança. Interessa averiguar se se continua a verificar a igualdade que caracteriza estes modelos, $E(Y_t \mid X_t) = Var(Y_t \mid X_t)$. Recorrendo à lei das expectativas iteradas e admitindo que $\lambda(X_t\beta) = \exp(X_t\beta)$ e $W_t = X_t + u_t$, é fácil mostrar que:

$$E(Y_t \mid W_t) = E_{u|W}\left[\exp(W_t\beta - u_t\beta) \mid W_t\right] = \exp(W_t\beta)\, E_{u|W}\left[\exp(-u_t\beta) \mid W_t\right] \tag{4.32}$$

Como W_t está medido com erro (e logo é não independente de u_t) temos de considerar que $E_{u|W}\left[\exp(-u_t\beta) \mid W_t\right] = m(W_t)$, uma função qualquer de W_t com $m(W_t) > 0$ e deste modo temos que:

$$E(Y_t \mid W_t) = \exp(W_t\beta)\, m(W_t) \tag{4.33}$$

Por outro lado

$$\begin{aligned} Var(Y_t \mid W_t) &= E_{u_t|W_t}\left[Var(Y_t \mid W_t, u_t)\right] + Var_{u|W}\left[E(Y_t \mid W_t, u_t)\right] = \\ &= \exp(W_t\beta)\, m(W_t) + \exp(2W_t\beta)\, Var_{u|W}\left[\exp(-u_t\beta) \mid W_t\right] \end{aligned} \tag{4.34}$$

[8] De facto é fácil mostrar que, $W - E(u \mid W) = W - (I + \Sigma_{XX}\Sigma_{uu})^{-1}(W - E(W))$
$= W\left(I - \left(I + \Sigma_{XX}\Sigma_{uu}^{-1}\right)^{-1}\right) - \left(I + \Sigma_{XX}\Sigma_{uu}^{-1}\right)^{-1} E(W)$.

utilizando o mesmo argumento, ou seja, considerando que $Var_{u|W}\left[\exp\left(-u_t\beta\right) \mid W_t\right] = h\left(W_t\right)$ onde $h\left(W_t\right) > 0$, vem,

$$Var\left(Y_t \mid W_t\right) = \exp\left(W_t\beta\right) m\left(W_t\right) + \exp\left(2W_t\beta\right) h\left(W_t\right) \tag{4.35}$$

Analisando as expressões (4.33) e (4.35) facilmente se conclui que a omissão de u_t do conjunto de variáveis a condicionar provoca sobredispersão, i.e., verifica-se a desigualdade $Var\left(Y_t \mid X_t\right) > E\left(Y_t \mid X_t\right)$, quaisquer sejam as funções $m\left(X_t\right)$ e $h\left(X_t\right)$ definidas anteriormente. Isto é o resultado típico do problema associado à heterogeneidade negligenciada. Contudo contrariamente ao que sucedia neste tipo de problema, a utilização do método da máxima verosimilhança, tem como consequência a obtenção de estimativas inconsistentes para todos os parâmetros do modelo, coeficientes associados às variáveis explicativas e termo independente. Isto porque da análise da expressão (4.33), podemos constatar que a forma como o valor esperado de Y_t depende de W_t é diferente da forma como valor o esperado de Y_t depende de X_t, o que vem corroborar a análise anteriormente efectuada. Este resultado é consequência do facto de quando utilizamos o conjunto de variáveis explicativas contaminadas com erro de medida, introduzimos na análise uma variável não observável que está correlacionada com os regressores.

Capítulo V- Teste Score

5.1 Introdução

Nos capítulos anteriores a análise do problema da existência de erro de medida centrou-se no estudo das consequências e na tentativa de obtenção de métodos alternativos de estimação consistentes sobre esta hipótese. Genericamente, a contaminação por erro de medida altera a forma como os regressores condicionam a variável dependente, podendo mesmo por em causa as formas funcionais das densidades condicionais. Quase sempre, o resultado é a obtenção de estimativas inconsistentes para os parâmetros das distribuições condicionais. Como a presença de erro de medida produz distorções nas distribuições das variáveis em estudo, a sua presença deverá ser detectada quando existe informação sobre as distribuições que moldariam a relação entre as variáveis medidas sem erro. Supondo que existe tal informação, é possível a construção de um teste de especificação que detecte, sob certas hipóteses, a presença de erro de medida em alguma(s) variáveis explicativas duma distribuição condicional.

Na verdade, tal informação nem sempre está disponível, nomeadamente a que diz respeito à especificação da distribuição das variáveis explicativas isentas de erro. Contudo a metodologia de aplicação do teste permite ultrapassar esse problema.

O teste introduzido nesta secção, é um teste clássico tipo score que considera a variância dos erros de medida. Como tal pretende ser um procedimento aplicável independentemente da distribuição do erro de medida. A metodologia adoptada, considera aproximações em série de Taylor para variância pequena a funções verosimilhança que envolvem variáveis potencialmente contaminadas por erro de medida (Chesher, 1990). Esta aproximação está na base da construção de um teste score cuja hipótese nula é a nulidade das variâncias dos erros de medida.

5.2 A Estatística de Teste

Suponha-se que a relação entre as realizações de uma variável Y condicionada nas realizações independentes de uma variável X é expressa pela densidade condicional de Y dado X, que pode ser escrita por $\Pr(Y \mid X) = f(Y, X\beta)$. A influência da variável explicativa dá-se através de uma combinação linear do vector

de parâmetros β. Considere-se ainda o problema clássico da presença de erro de medida nas variáveis explicativas, onde apenas observamos uma variável proxy cuja relação com a verdadeira variável é dada por $W_i = X_i + u_i$, com $E(u_i) = 0$, $E(u_j u_j) = \sigma_{ij}$ com $i = j = 1, \ldots, k$. Mais uma vez, W é a matriz com as observações das variáveis explicativas verdadeiramente observadas; X é a verdadeira matriz de observações das variáveis explicativas; e u é a matriz do erro de medida. As variáveis X e u têm dimensão respectivamente $(n \times (k+1))$ e $(n \times k)$.

Sendo assim, e dada a estrutura do modelo para o erro de medida, se o vector erro de medida fosse observável, o modelo dos dados observados seria escrito como $\Pr(Y \mid W, u) = f[Y, (W\beta - u\beta)]$. Expandindo esta expressão em torno de $u = 0$ até à segunda ordem e calculando o valor esperado em relação a u condicional em W obtêm-se;

$$\begin{aligned} f[Y,(W\beta - u\beta)] \doteq \Pr(Y \mid W, 0) - \sum_{i=1}^{k} E_{u|W}(u_i \mid W) \tfrac{\partial P(\cdot)}{\partial u_i} \mid_{u=0} + \\ + \tfrac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} E_{u|W}(u_i u_j \mid W) \tfrac{\partial^2 P(\cdot)}{\partial u_i \partial u_j} \mid_{u=0} + o(\sigma^2) \end{aligned} \tag{5.1}$$

onde k é o número de variáveis explicativas, u_i é a i-ésima coluna da matriz u e $E_{u|W}(\cdot)$ representa o valor esperado relativamente a u condicional em W.

A análise desta expressão permite-nos levantar duas questões preliminares. A primeira, é que o Score desta expressão para testar $H_0 : E_{u|W}(u_i u_j \mid W) = 0$, assemelha-se ao de um teste de heterogeneidade negligenciada. Contudo, naquela situação os momentos condicionais da variável que determina a aleatoriedade do parâmetro são independentes da variável explicativa. Tal não é verdade na presença de erro de medida, o próprio modelo erro de medida pressupõe correlação ou dependência entre a variável observada e o erro de medida. Para termos um problema semelhante ao da heterogeneidade negligenciada aqueles valores esperados teriam de ser incondicionais nas variáveis explicativas observadas, ou seja, teríamos de ter $E_u(u_i u_j)$. A segunda questão é que sob esta forma o segundo termo da expressão não tem qualquer interpretação.

De modo a obter-se um teste que respeite aquele princípio, há que escrever os momentos da variável erro de medida incondicionalmente na variável W. Desta forma, escrevendo aquele valor esperado condicional e sabendo que $f_{uW}(u, W) =$

$f_{W|u}(W \mid u) f_u(u)$ é fácil mostrar que,

$$E_{u|W}(u \mid W) = \int u \frac{f_{W|u}(W \mid u)}{f_W(W)} f_u(u)\, du = E_u\left(u \frac{f_{W|u}(W \mid u)}{f_W(W)}\right) \tag{5.2}$$

o que nos permite escrever o valor esperado de interesse incondicionalmente nos regressores W. Contudo embora este problema esteja resolvido, esta expressão depende do conhecimento de $f_{W|u}(W \mid u)$. Seguindo Santos Silva, J.M. (1993), podemos aproximar esta expressão em série de Taylor em torno de $u = 0$ apenas até à primeira ordem e tirar o valor esperado em ordem a u, obtendo a seguinte expressão,

$$E_u\left(u \frac{f_{W|u}(W \mid u)}{f_W(W)}\right) \doteq \frac{1}{f_W(W)}\left[E_u(u) f_{W|u}(W \mid 0) + \sum_{i=1}^{k} E_u(u_i u_j) f^{(1)}_{W|u}(W \mid 0)\right] \tag{5.3}$$

Sabendo que $E_u(u) = 0$ e que $E_u(u_i u_j) = \sigma_{ij}$ obtemos,

$$E_u\left(u \frac{f_{W|u}(W \mid u)}{f_W(W)}\right) \doteq \sum_{i=1}^{k} \sigma_{ij} \frac{f^{(1)}_{W|u}(W \mid 0)}{f_W(W)} \tag{5.4}$$

e dado que podemos também aproximar em série de Taylor em torno de $u = 0$ de grau zero $E_{u|W}(u_i u_j) \doteq \sigma_{ij} + o(\sigma)$. Voltando à expressão (5.1) e escrevendo $\frac{\partial P(\cdot)}{\partial u_i}|_{u=0} = P^{(1)}(Y \mid W, 0)$, $\frac{\partial^2 P(\cdot)}{\partial u_i \partial u_j}|_{u=0} = P^{(2)}(Y \mid W, 0)$, que representam respectivamente as primeira e segunda derivadas da função densidade condicional, obtemos a seguinte aproximação,

$$\begin{aligned} f[Y, (W\beta - u\beta)] &\doteq \Pr(Y \mid W, 0) + \\ + \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij} &\left[\tfrac{1}{2} P^{(2)}(Y \mid W, 0) - \tfrac{\partial \ln f_W(W)}{\partial W} P^{(1)}(Y \mid W, 0)\right] + o(\sigma^2) \end{aligned} \tag{5.5}$$

Esta expressão representa a *pseudo função verosimilhança condicional* através da qual podemos construir um teste tipo Score para testar $H_0 : \sigma_{ij} = 0$, $\forall_{i,j}$ $i = j = 1, 2, ..., k$. O teste é baseado numa logverosimilhança da forma,

$$\begin{aligned} \ln L_t(Y, W, u, \beta, \sigma_{ij}) &\doteq \ln\{\Pr(Y \mid W, 0) + \\ + \sum_{i=1}^{k}\sum_{j=1}^{k} \sigma_{ij} &\left[\tfrac{1}{2} P^{(2)}(Y \mid W, 0) - \frac{\partial \ln f_W(W)}{\partial W} P^{(1)}(Y \mid W, 0)\right]\Big\} + o(\sigma^2) \end{aligned} \tag{5.6}$$

E logo a contribuição para o Score da t-ésima observação vem:

$$S_{t\sigma_{ij}}(\beta)\mid_{\sigma_{ij}=0} \doteq \frac{1}{2}\left(\frac{\partial^2 \ln P(\cdot)}{\partial u_i \partial u_j} + \frac{\partial \ln P(\cdot)}{\partial u_i}\frac{\partial \ln P(\cdot)}{\partial u_j}\right) - \left(\frac{\partial \ln f_W(W)}{\partial W}\frac{\partial \ln P(\cdot)}{\partial u_i}\right) + o(\sigma^2) \tag{5.7}$$

Desta forma podemos claramente identificar neste teste duas componentes:

i) Uma primeira expressão que é claramente um teste de matriz de informação e que corresponde a testar a volatilidade da variável u assemelha-se ao teste de heterogeneidade negligenciada (com a diferença de que as derivadas são em ordem às variáveis ao invés de serem em ordem aos parâmetros).

ii) A segunda expressão é a de um teste tipo Score de omissão de variável, onde temos o vector Score a multiplicar por $\partial \ln f_W(W)/\partial W$. Neste caso pode interpretar-se como um teste de especificação da forma funcional dos regressores da função index causada pela contaminação da distribuição da variável explicativa[9] .

É fácil verificar que sob a hipótese nula o valor esperado deste score, avaliado em $\widehat{\beta}$, estimativa da máxima verosimilhança na ausência de erro de medida, é zero. O primeiro termo do score dá-nos a igualdade da matriz de informação, que sob a hipótese nula deve ser zero, o segundo termo sendo um teste tipo Score de especificação da forma funcional, neste caso, não linearidade dos regressores na função index, tem também sob a hipótese nula valor esperado nulo, daí que o score resultante seja a soma de duas expressões com esperança matemática nula.

5.3 Implementação

Como já foi referido, a expressão (5.6) representa a pseudo log-verosimilhança a partir da qual um teste tipo score pode ser construído para testar a hipótese nula $H_0 : \sigma_{ij} = 0, \forall_{ij}$ com $i = j = 1, ..., k$. A relevância da construção de um teste para detecção deste problema, justifica-se pelo facto de:

i) Os modelos não lineares, nomeadamente o Logit e o Poisson aqui considerados, são preferencialmente estimados pelo método da máxima verosimilhança. Para que da utilização deste método resultem estimativas consistentes para os parâmetros o modelo deve estar correcta e totalmente especificado, i.e, tem de se

[9] Contudo como não conhecemos a densidade de X, para podermos operacionalizar o teste Chesher (1990) sugere que se faça uma aproximação polinomial àquela expressão.

conhecer toda a distribuição condicional da variável dependente. A omissão do erro de medida resulta assim numa incorrecta especificação do modelo estimado.

ii) Quase nunca é possível especificar a distribuição do erro de medida e a forma funcional da distribuição conjunta do erro e da variável contaminada. Neste contexto a utilidade de um teste deste tipo é justificada pela vantagem de se trabalhar sob a hipótese nula e logo contra uma multiplicidade de alternativas na sua vizinhança.

O modelo representado por (5.6) está definido para um espaço paramétrico definido por $\theta = \{\beta, \sigma\}$ onde $dim(\beta) = k$ e σ representa um vector de dimensão $k \times (k+1)/2$. Sendo assim,

$$s\left(\widehat{\theta}\right) \widehat{\Im}^{-1} s\left(\widehat{\theta}\right)^{T} \text{ com } s\left(\widehat{\theta}\right) = \sum_{t=1}^{N} S_{\theta}\left(\widehat{\theta}\right) \tag{5.8}$$

é a estatística de teste adequada para testar a presença de erro de medida nas variáveis explicativas, onde $\widehat{\theta}$ representa a estimativa da máxima verosimilhança do modelo estimado sob a hipótese nula e $\Im\left(\widehat{\theta}\right)$ representa uma estimativa consistente da matriz de informação.

À semelhança do espaço dos parâmetros, o vector score de cada observação está dividido em $S_{\beta}\left(\widehat{\beta}\right)$ e $S_{\sigma}\left(\widehat{\beta}\right)$, ambos avaliados sobre a hipótese nula, i.e, com $\sigma = \overline{0}$. Sob a hipótese nula esta forma quadrática representa uma estatística de teste com distribuição assintótica $\chi^2_{(k\times(k+1)/2)}$[10].

A operacionalização do teste pressupõe a obtenção de uma estimativa para a matriz de informação. Considerem-se duas formas de estimar a matriz de informação:

i) Menos a Hessiana empírica[11], que se obtem derivando o score em ordem aos parâmetros avaliados sob a hipótese nula:

$$\Im^{1}\left(\widehat{\theta}\right) = -\sum \frac{\partial^2 L\left(\widehat{\theta}\right)}{\partial\widehat{\theta}_i \partial\widehat{\theta}_j} \tag{5.9}$$

ii) e ainda,

[10]Pode ser necessário corrigir os graus de liberdade da estatística de teste porque os scores relevantes para efectuar o teste são quase sempre funcionalmente dependentes.

[11]Este modo de estimar a matriz de informação vai ser apenas utilizado na construção do teste para o modelo de regressão Poisson. No caso do modelo Logit menos a Hessiana numérica é igual à matriz de informação estimada pelo que se usará como estimador, a matriz de informação avaliada no parâmetro estimado pela máxima verosimilhança sob a hipótese nula.

$$\Im^2\left(\widehat{\theta}\right) = s\left(\widehat{\theta}\right)^T s\left(\widehat{\theta}\right) \text{ com } \widehat{\theta} = \left\{\widehat{\beta}, 0\right\} \tag{5.10}$$

ou seja, o produto externo do vector gradiente, OPG proposto por Berndt, Hall, Hall & Hausman (1974).

Quando a matriz de informação é estimada pelas segundas derivadas analíticas, que corresponde ao método de Newton, a estatística de teste, NW, é a forma quadrática definida em (5.8). O seu cálculo exige a especificação da logverosimilhança e o cálculo das segundas derivadas avaliadas no vector $\widehat{\theta} = \left\{\widehat{\beta}, 0\right\}$[12] , onde o parâmetro $\widehat{\beta}$ foi previamente estimado pela máxima verosimilhança sob a hipótese nula.

As estatísticas de teste tipo score baseadas na estimativa da matriz de informação $\Im^2$ são bastante convenientes porque podem facilmente ser calculadas por recurso à regressão linear artificial denominada OPG. Conforme sugerido por Chesher (1983) e Lancaster (1984), os teste de matriz de informação podem ser executados por intermédio deste procedimento. Considere-se a seguinte regressão auxiliar:

$$\iota_t = S_{t\beta}\left(\widehat{\beta}\right) b + S_{t\sigma}\left(\widehat{\beta}\right) c + erro \tag{5.11}$$

onde ι_t representa um vector unitário $S_\beta\left(\widehat{\beta}\right)$ o vector score dos parâmetros associados aos regressores do modelo e $S_\sigma\left(\widehat{\beta}\right)$ o vector score associado aos parâmetros que determinam a existência de erro de medida avaliados sob a hipótese nula. Testar a hipótese nula $H_0 : \sigma_{ij} = 0, \forall_{ij}$ corresponde a testar $H_0 : c = \overline{0}$. Nestas condições é bem conhecido o resultado de que a estatística de teste proposta por Godfrey & Wickens (1981) pode ser calculada como,

$$LM_1 = N - SRQ \tag{5.12}$$

que tem, sobre a hipótese nula distribuição assintótica $\chi^2_{(k\times(k+1)/2)}$.

Para o modelo de regressão Poisson a matriz dos regressores associados ao parâmetro c é dada por ;

$$S_{\sigma_{ii}}\left(\widehat{\beta}\right) = \frac{1}{2}\widehat{\beta}_i^2\left[\left(Y-\widehat{\lambda}\right)^2 - \widehat{\lambda}\right] - \widehat{\beta}_i \frac{\partial \ln f_X(X)}{\partial X}\left(Y-\widehat{\lambda}\right) \tag{5.13}$$

[12]Esta forma de executar o teste é pouco prática.

e

$$S_{\sigma_{ij}}\left(\widehat{\beta}\right) = \frac{1}{2}\widehat{\beta}_i\widehat{\beta}_j\left[\left(Y-\widehat{\lambda}\right)^2 - \widehat{\lambda}\right] \tag{5.14}$$

onde $\widehat{\lambda} \equiv \exp\left(X\widehat{\beta}\right)$.

Para o modelo de regressão binária Logit os regressores auxiliares são dados por:

$$S_{\sigma_{ii}}\left(\widehat{\beta}\right) = \frac{1}{2}\widehat{\beta}_i^2\left[\left(Y-\widehat{F}\right)^2 - \widehat{F}\left(1-\widehat{F}\right)\right] - \widehat{\beta}_i\frac{\partial \ln f_X(X)}{\partial X}\left(Y-\widehat{F}\right) \tag{5.15}$$

e

$$S_{\sigma_{ij}}\left(\widehat{\beta}\right) = \frac{1}{2}\widehat{\beta}_i\widehat{\beta}_j\left[\left(Y-\widehat{F}\right)^2 - \widehat{F}\left(1-\widehat{F}\right)\right] \tag{5.16}$$

e $\widehat{F} \equiv \Pr\left(Y=1 \mid X\widehat{\beta}\right)$[13] .

No caso do modelo Logit é possível tirar-se partido do facto dos scores associados aos parâmetros que determinam a existência de erro de medida, i.e, os elementos típicos da matriz de informação poderem ser escritos como,

$$S_{\sigma_{ii}}\left(\widehat{\beta}\right) = \frac{1}{2}\widehat{\beta}_i^2\left(1-2\widehat{F} - \frac{\partial \ln f_X(X)}{\partial X}.\frac{1}{\widehat{\beta}_i}\right).\left(Y-\widehat{F}\right) \tag{5.17}$$

e

$$S_{\sigma_{ij}}\left(\widehat{\beta}\right) = \frac{1}{2}\widehat{\beta}_i\widehat{\beta}_j\left(1-2\widehat{F}\right).\left(Y-\widehat{F}\right) \tag{5.18}$$

Deste resultado é fácil verificar que as expressões (5.17) e (5.18) podem ser interpretadas como o score associado aos parâmetros de regressores definidos pela matriz **S** com colunas $\frac{1}{2}\widehat{\beta}_i^2\left(1-2\widehat{F} - \frac{\partial \ln f_X(X)}{\partial X}.\frac{1}{\widehat{\beta}_i}\right)$ e $\frac{1}{2}\widehat{\beta}_i\widehat{\beta}_j\left(1-2\widehat{F}\right)$. O problema de detecção de erro de medida no modelo Logit passa a ser interpretado como um teste de omissão do conjunto de variáveis representado por **S** da função index[14] . Sendo assim usando o resultado de Davidson & Mackinnon (1984b) o teste pode ser executado com recurso a uma regressão linear auxiliar que corresponde a uma versão modificada da regressão Gauss Newton -GNR- para modelos não lineares. O teste é executado fazendo a regressão,

$$\widehat{V}_t^{-1/2}\left(y_t - \widehat{F}_t\right) = \widehat{V}_t^{-1/2}\widehat{f}X_t b + \widehat{V}_t^{-1/2}\widehat{f}S_t c + erro \tag{5.19}$$

[13]Em qualquer um dos modelos de regressão não lineares aqui considerados é possível identificar nas expressões dos scores o elemento típico da matriz de informação.

[14]Para o efeito aqui considerado **S** é um regressor adicional que não depende de parâmetros desconhecidos, pelo que sob a hipótese alternativa o modelo continua a depender de X através de uma função index que é uma combinação linear de regressores originais e parâmetros mais o regressor adicional.

onde $\widehat{F} \equiv \Pr\left(Y = 1 \mid X\widehat{\beta}\right) \equiv F\left(X\widehat{\beta}\right)$, $\widehat{f} \equiv f\left(X\widehat{\beta}\right)$ e $\widehat{V} \equiv \widehat{F}\left(1 - \widehat{F}\right)$. Testar a hipótese nula $H_0 : \sigma_{ij} = 0, \forall_{ij}$ corresponde a testar a omissão do conjunto de variáveis **S** do conjunto de informação, ou seja $H_0 : c = 0$. Sob a hipótese nula, a estatística de teste, LM_2, obtêm-se calculando o SQE da regressão auxiliar que mais uma vez tem distribuição assintótica $\chi^2_{(k\times(k+1)/2)}$. Neste caso a estimativa da matriz de informação associada a LM_2 é a própria matriz de informação avaliada em $\widehat{\beta}$ da máxima verosimilhança.

À partida são conhecidos alguns resultados sobre o comportamento das estatísticas de teste aqui consideradas, nomeadamente o facto de em amostras finitas LM_1 sobrerejeitar a hipótese nula e LM_2 ter uma potência empírica inferior embora rejeite com menor frequência a hipótese nula quando ela é verdadeira (ver Davidson & Mackinnon, 1984b).

5.4 Simulações

De modo a avaliar a extensão do enviesamento nas estimativas naives e as propriedades para amostras finitas das diferentes estatísticas de teste aqui consideradas, realizou-se uma série de estudos de Monte Carlo. Dentro da classe dos modelos não lineares considerou-se o modelo de regressão binária Logit e o modelo de regressão Poisson para dados de contagem. Para qualquer uma das especificações, considerou-se o modelo para erro de medida clássico com distribuição normal,

$$W_t = X_t + u_t \text{ e } u_t \sim NID\,(0,1) \tag{5.20}$$

Como já foi mencionado, em rigor o teste exige o conhecimento da distribuição da variável medida sem erro para avaliar o termo $\partial \ln f_X\,(X)\,/\partial X$. Como na prática esta expressão quase nunca é conhecida, seguindo a sugestão de Chesher (1992) adoptou-se uma aproximação polinomial do tipo $\partial \ln f_X\,(X)\,/\partial X = X^2$. O objectivo é aferir das condições em que nestes modelos mais populares o erro de medida tem repercussões sérias na análise econométrica e da capacidade dos teste tipo score aqui construídos para detectar esse "vírus".

5.4.1 Modelo Logit

Para a análise das consequências do erro de medida nos modelos Logit e o comportamento das estatísticas de teste LM_1 e LM_2, construíram-se os seguintes modelos com base numa especificação para a $\Pr(Y = 1 \mid X, \beta) = F(X'\beta)$:

$$(I) : (\beta_0, \beta_1) = (1, -2), X_1 \sim \chi^2_{(1)}/\sqrt{2}$$
$$(II) : (\beta_0, \beta_1) = (1, -1), X_1 \sim \left(\chi^2_{(1)} - 1\right)/\sqrt{2}$$
$$(III) : (\beta_0, \beta_1) = (1, -2), X_1 \sim Normal(1, 1)$$
$$(IV) : (\beta_0, \beta_1, \beta_2) = (1, 2, -1), X_1 \sim \chi^2_{(1)}/\sqrt{2} \text{ e } X_2 \sim Poisson(2)$$

Os modelos foram construídos de forma a que a variável sujeita a contaminação com erro de medida tenha variância unitária e distribuição contínua. Todas as variáveis foram geradas independentemente do erro de medida. Além da distribuição das variáveis explicativas e dos parâmetros de cada um dos designs considerados, os modelos propostos são caracterizados pelo valor esperado de Y condicional nos regressores. Estas probabilidades são para cada um dos modelos respectivamente, 0,55 (I), 0,85 (II), 0,33 (III) e 0,51 (IV). Estas estatísticas dão-nos uma ideia de como as observações estão mais ou menos enviesadas para um dos extremos da distribuição condicional de Y.

O erro de medida u, foi gerado de uma distribuição $Normal(0, \sigma_u)$ para os valores de $\sigma_u = 0, 0,1, 0,3$, e $0,5$. Todas as experiências tiveram como base uma amostra de 500 observações e foram repetidas 1500 vezes.

Tabela 1: Estimativas dos Modelos Logit

	$\sigma_u = 0$	$\sigma_u = 0,1$	$\sigma_u = 0,3$		$\sigma_u = 0$	$\sigma_u = 0,1$	$\sigma_u = 0,3$
	Modelo I				**Modelo II**		
β_0	1,00545	0,97249	0,77765	β_0	0,99947	1,00021	1,00298
	(0,14374)	(0,14361)	(0,13645)		(0,10905)	(0,10904)	(0,10907)
β_1	-2,0261	-1,94426	-1,50867	β_1	-1,01293	-1,00045	-0,91233
	(0,23484)	(0,22590)	(0,17001)		(0,12766)	(0,12606)	(0,11624)
	Modelo III				**Modelo IV**		
β_0	1,00569	0,98274	0,81826	β_0	1,00792	1,02614	1,1383
	(0,17460)	(0,1731)	(0,16345)		(0,25926)	(0,25955)	(0,25778)
β_1	-2,01420	-1,98283	-1,76464	β_1	2,0298	1,95083	1,52110
	(0,18915)	(0,18479)	(0,16407)		(0,24647)	(0,23453)	(0,17628)
				β_2	-1,0113	-1,00277	-0,95614
					(0,125)	(0,12453)	(0,12097)

Desvios padrão estimados em parêntesis.

Os resultados da tabela 1 são bastante elucidativos quanto ao efeito da utilização de variáveis explicativas medidas com erro nas estimativas do modelo

Logit. Considerando apenas os modelos com variável explicativa $\chi^2_{(1)}$ constata-se que o enviesamento é tanto maior quanto maior a variância do erro de medida, mas que para um mesmo σ_u o efeito do erro de medida é mais severo no modelo I que nos restantes. Isto sugere que o enviesamento provocado pelo erro de medida sobre as estimativas dos parâmetros é máximo quando $E(Y \mid X) = 0,5$.

Dos resultados obtidos é possível constatar que o enviesamento é sempre em direcção à origem, pelo que se pode falar em efeito atenuação na sua interpretação naive. É de notar que em todos os modelos aqui considerados o termo independente também sofre um enviesamento já que a contaminação por erro de medida não altera a média da função index e logo o termo independente deve ter um movimento compensatório. A consideração do modelo IV, permite-nos verificar que a presença de erro de medida numa variável, afecta também a estimativa do parâmetro da variável medida sem erro. Neste caso o enviesamento é também em direcção à origem. Para além do efeito atenuação constata-se que todos os desvio padrão das estimativas são menores e tanto menores quanto maior a variância do erro de medida. Existe desta forma uma tradeoff entre enviesamento e variância do estimador naive[15] .

A tabela 2 mostra-nos o comportamento das estatísticas de teste LM_1 e LM_2 para estes designs para um nível de significância de 5% associado a uma distribuição $\chi^2_{(1)}$. Dos resultados obtidos para $\sigma_u = 0$ é evidente que a estatística de teste LM_2 é melhor comportada que LM_1, apresentando uma dimensão sempre inferior a 5% para este nível de significância. Pelo contrário, quando calculada por LM_1 esta estatística rejeita a hipótese nula verdadeira com maior frequência do que seria desejável, atingindo um valor de 7,667% no caso do modelo II. Este resultado não é de todo surpreendente dado que são bem conhecidas as propriedades da regressão OPG na construção de estatísticas de teste. Consequentemente, dadas as propriedades da regressão OPG atrás referidas, não é de estranhar que o teste LM_1, para um nível de significância de 5% pareça ter mais capacidade para rejeitar a hipótese nula falsa, dado ter para todos os modelos

[15]Esta questão pode ser importante já que a maioria dos estimadores corrigidos têm sempre um enviesamento inferior, mas uma variância superior ao estimador naive (ver Stefanski & Carroll, 1985 sobre estimadores corrigidos para o modelo Logit e suas propriedades). Isto sugere que devem-se utilizar tais técnicas alternativas de estimação à luz de critérios como o Erro Quadrático Médio.

e para todos os valores de σ_u uma frequência de rejeição superior à de LM_2. À semelhança do que sucedia com o enviesamento dos parâmetros, existe uma relação entre potência empírica e $E(Y \mid X)$. Novamente, quanto mais perto de $0,5$ maior a capacidade do teste detectar a presença de erro de medida.

Tabela 2: Resultado das Simulações Logit[a]
Frequências de rejeição ao nível de 5%

	$\sigma_u = 0$		$\sigma_u = 0,1$		$\sigma_u = 0,3$		$\sigma_u = 0,5$	
	LM_1	LM_2	LM_1	LM_2	LM_1	LM_2	LM_1	LM_2
Modelo I								
F.R.	0,06733	0,03933	0,14067	0,0533	0,66867	0,49	0,886	0,772
Média	1,21397	0,91481	1,80507	1,145986	7,67797	3,00109	13,24778	6,57046
	(1,80057)	(1,52298)	(2,59752)	(1,45986)	(6,55372)	(3,00109)	(8,41476)	(3,45980)
Modelo II								
F.R.	0,07667	0,046	0,092667	0,04	0,21533	0,104	0,44533	0,30333
Média	1,27750	0,94296	1,37050	0,95547	2,43958	1,51090	4,58427	2,98653
	(1,95058)	(1,39555)	(2,13094)	(1,35004)	(3,33247)	(1,74516)	(4,62954)	(2,70335)
Modelo III								
F.R.	0,070667	0,038	0,077333	0,040667	0,06733	0,042	0,07333	0,042
Média	1,28739	0,90101	1,26414	0,89009	1,15270	0,89725	1,19069	0,96895
	(2,08694)	(1,28002)	(1,98126)	(1,25606)	(1,60934)	(1,26111)	(1,61730)	(1,35743)
Modelo IV								
F.R.	0,10667	0,04733	0,12400	0,05000	0,26533	0,13133	0,48867	0,314
Média	2,69441	1,97537	2,84535	1,98043	4,69062	3,01356	7,24091	4,79152
	(3,29631)	(1,95425)	(3,53974)	(1,90730)	(4,95025)	(2,53544)	(5,73748)	(3,27408)
Modelo I: Teste MI								
F.R.	0,13267	0,03133	0,166	0,026	0,40067	0,03533	0,61467	0,12467
Média	1,96126	0,92694	2,31069	0,87736	5,32689	1,18282	8,93337	1,93180
	(3,76549)	(1,36307)	(4,41703)	(1,19456)	(7,35191)	(1,165)	(9,18506)	(1,50832)

[a]Desvios padrão assintóticos de LM_1 e LM_2 em parêntesis.
F.R.: Frequência de Rejeição.

É importante analisar o caso em que a variável explicativa contaminada com erro tem distribuição normal- modelo IV. Embora o efeito atenuação sobre os parâmetros persista, ambas as estatísticas de teste denotam um clara incapacidade para a um nível de significância de 5% rejeitar a hipótese nula quando ela é falsa. Este resultado está de acordo com o apresentado na secção (4.3.2) relativamente ao efeito no modelo de regressão Probit (e por semelhança no Logit), quando a variável contaminada com erro tem distribuição normal, segundo o qual a função index contínua a ser linear nos regressores. Este resultado suscita outra questão igualmente pertinente. Dado que o teste é um misto de teste de matriz de informação e de detecção de não linearidades na função index, resta averiguar qual a influência das linearidades introduzidas neste tipo de especificação e qual o peso desta vertente do teste em ambas as estatísticas[16] . Os resultados são

[16]Note-se que estes resultados estão condicionados pelo tipo de aproximação polinomial a $\partial \ln f_X(X)/\partial X$ aqui considerada. A utilização de outras expressões para aquele termo da estatística de teste e a comparação das frequências de rejeição podem neste contexto fornecer uma pista para averiguação do tipo de não linearidades introduzidas nesta especificação.

bastante reveladores. Se considerarmos apenas o score do teste de matriz de informação, em ambas as estatísticas de teste a frequência de rejeição baixa consideravelmente, com maior incidência na estatística de teste LM_2. Isto leva-nos a concluir que não só o efeito das não linearidades nos regressores é importante nesta classe de modelos como, a estatística de teste LM_2 é particularmente sensível à detecção deste problema, contrariamente LM_1 parece ser mais sensível á parte do teste referente á variabilidade do erro de medida.

Finalmente interessa considerar o cálculo da estatística de teste quando o modelo tem duas variáveis explicativas- modelo V. Esta especificação é importante porque alerta-nos para a necessidade de proceder à correcção dos graus de liberdade da estatística de teste, devido à redundância de alguns regressores[17] . Comparando os resultados como o modelo I que tem aproximadamente o mesmo $E(Y \mid X)$ o teste apresenta uma frequência de rejeição menor para o mesmo nível de significância. Da análise dos resultados não resulta uma clara supremacia de uma estatística de teste sobre a outra. LM_1 comporta-se melhor sob a hipótese alternativa enquanto LM_2 tem uma melhor performance sob a hipótese nula em termos de frequência de rejeição e de desvio padrão da média da estatística de teste nas réplicas consideradas.

5.4.2 Modelo Poisson

Repetiu-se o mesmo procedimeto para o modelo de regressão Poisson, considerando-se as estatísticas de teste LM_1 e NW nos seguintes designs:

$$(I): \ (\beta_0, \beta_1) = (2, -1), \ X_1 \sim \left(\chi^2_{(1)} - 1\right)/\sqrt{2}$$
$$(II): \ (\beta_0, \beta_1) = (-1, -1, 5), \ X_1 \sim \left(\chi^2_{(1)} - 1\right)/\sqrt{2}$$

O modelo I tem uma média incondicional da variável dependente de $9,22$ e o modelo II de $0,56$. Mais uma vez as variáveis explicativas têm distribuição contínua e variância unitária. O modelo para o erro de medida é o mesmo considerado na secção anterior com $\sigma_u = 0,\ 0,2$, e $0,3$. Dados os maiores custos computacionais associados a este tipo de especificação consideraram-se apenas 1000 réplicas para uma amostra de 500 observações.

[17] O score associado ao parâmetro σ_{12} é sempre colinear ao score de σ_{11}, daí que tenha sido necessário considerar como nível crítico a 5% uma variável $\chi^2_{(2)}$.

Tabela 3: Resultado das Simulações Poisson

	$\sigma_u = 0$	$\sigma_u = 0,2$	$\sigma_u = 0,3$
	Modelo I		
β_0	1,99999	2,0444	2,07642
	(0,019567)	(0,019185)	(0,019386)
β_1	-1,00031	-0,84068	-0,71770
	(0,033349)	(0,027809)	(0,024974)
	Modelo II		
β_0	-1,01234	-0,88648	-0,81516
	(0,099843)	(0,07804)	(0,071423)
β_1	-1,5133	-1,16363	-0,94810
	(0,17619)	(0,11592)	(0,092067)

Desvios padrão estimados em parêntesis.

Da análise da tabela 3 é novamente possível observar o efeito atenuação sobre as estimativas dos parâmetros afectos à variável explicativa medida com erro, assim como a redução dos respectivos desvios padrão estimados. Em ambos os modelos considerados o enviesamento provocado pela presença de erro de medida pode ser bastante severo e tanto maior quanto maior a variância do erro de medida. Os resultados sugerem também que a dimensão do enviesamento é em termos absolutos crescente com a dimensão do parâmetro e relativamente independente da média de Y e que mais uma vez os termos independentes ajustam-se de modo a manter a média da função index inalterada.

Tabela 4: Resultado das Simulações Poisson[a]
Frequências de rejeição ao nível de 5%

	$\sigma_u = 0$		$\sigma_u = 0,2$		$\sigma_u = 0,3$	
	LM_1	NW	LM_1	NW	LM_1	NW
			Modelo I			
F.R.	0,068	0,071	0,799	0,798	0,975	0,975
Média	1,18105	1,20677	7,75085	8,13744	13,95583	15,33509
	(1,75685)	(1,85278)	(4,45786)	(4,91839)	(5,76200)	(7,09242)
			Modelo II			
F.R.	0,075	0,102	0,124	0,163	0,367	0,338
Média	1,17738	4,76281	1,66148	9,16762	4,00216	22,31631
	(1,77443)	(63,54257)	(2,58821)	(86,59483)	(4,92682)	(293,69479))
			Modelo I: Teste MI			
F.R.	0,068	0,070	0,799	0,977	0,975	1
Média	1,18105	1,18935	7,75085	14,32925	13,95583	35,46388
	(1,75685)	(1,81239)	(4,45786)	(6,43402)	(5,76200)	(10,95791)

[a] Desvios padrão estimados de LM_1 e NW em parêntesis.
F.R.: Frequência de Rejeição

O comportamento das estatísticas de teste LM_1 e NW está patente na tabela 4. Para um nível de significância de 5% ambas as estatísticas de teste tendem a rejeitar a hipótese nula verdadeira mais do que o desejável. Neste capítulo NW tende a ser particularmente mal comportada. Relativamente à potência empírica, quando se considera o teste na sua dupla vertente as duas estatísticas de teste apresentam resultados bastante semelhantes, sendo as frequências de

rejeição no modelo I bastante elevadas. Como seria de esperar, a frequência de rejeição aumenta com σ_u e á luz destes resultados são função da média do modelo Poisson (97,5% para $\sigma_u = 0,3$ no modelo I)[18] .

À semelhança do efectuado para o modelo Logit interessa analisar o comportamento das estatísticas de teste relativamente à importância e detecção das não linearidades nos regressores introduzidas neste tipo de especificação pelo erro de medida. Considerando apenas o modelo I, os resultados da estatística LM_1 quando se considera apenas a parte relativa ao teste de matriz de informação são exactamente iguais ao do teste completo que considera a aproximação polinomial de segundo grau. Dadas as elevadas frequências de rejeição, tal facto sugere que o efeito de introdução de não linearidades nos regressores provocado pelo erro de medida não é (neste design) muito severo e que, quando calculado pela regressão OPG o teste não é muito sensível a este problema de especificação. Quando calculado por NW a dimensão do teste quase não se altera mas a capacidade do teste para rejeitar uma hipótese nula falsa aumenta, detectando sempre a presença de erro quando a hipótese alternativa é $H_1 : \sigma_u = 0,3$.

Os resultados destas experiências parecem sugerir que a estatística de teste LM_1 é preferível a NW (embora seja ela própria bastante mal comportada) já que se comporta melhor sob ambas as hipóteses, excepto quando consideramos apenas o teste MI onde esta apresenta uma maior frequência de rejeição.

[18]Note-se que o enviesamento do parâmetro estimado era mais acentuado no modelo II com média menor, o que reforça a ideia que o enviesamento é função da dimensão do verdadeiro parâmetro e a capacidade de detecção é função da média do Poisson. Contudo respostas mais concretas a estas questões só podem ser alcançadas no âmbito de um estudo de simulação mais alargado que não cabe a este trabalho realizar.

Capítulo VI- Erro na variável Dependente

6.1 Introdução

A análise efectuada nos capítulos anteriores considerava apenas as consequências da existência de erro de medida na variável explicativa sobre a distribuição condicional de Y dado X. Esta preocupação quase exclusiva prestada á análise da presença de erro nos regressores é explicada pelo facto de na maioria das vezes a contaminação na variável dependente poder ser ignorável, i.e., os métodos estatísticos anteriormente utilizados continuam a ser válidos. Ignorar-se a presença do erro de medida pressupõe que o modelo válido para a verdadeira "resposta" continua a ser válido para a variável proxy, não se alterando os parâmetros afectos às variáveis explicativas, introduzindo-se apenas um elemento adicional de variabilidade que se junta ao erro da regressão.

A maior parte da análise aplicável às consequências ou distorções introduzidas nas distribuições marginais, transita para o estudo do erro na variável dependente de um modelo de regressão condicional num conjunto de variáveis explicativas. Ao invés de estarmos interessados nos efeitos sobre a distribuição marginal de Y, agora consideramos que é a distribuição de Y condicional num conjunto de variáveis explicativas, que aparecem na equação da regressão, que está contaminada com erro.

Quando Y é uma variável contínua e considerando um modelo genérico dado por:

$$Y_t = g(X_t; \beta) + \varepsilon_t \tag{6.1}$$

e

$$S_t = Y_t + u_t \tag{6.2}$$

a utilização da variável proxy na regressão de interesse, tem como efeito um aumento da variância da variável dependente, mas na sua forma benigna, o "vírus" não provoca nenhum enviesamento das médias dos coeficientes da regressão.

Considere-se o caso mais simples dos modelos de regressão linear sob as hipóteses clássicas, com erro de medida aditivo dado por (6.2). Neste caso, é

fácil verificar que o erro de medida pura e simplesmente se confunde com o erro da equação. O efeito da contaminação é apenas, o de aumentar a volatilidade na estimativa dos parâmetros. Deste modo a utilização do método dos mínimos quadrados contínua a garantir a consistência dos estimadores, perdendo-se contudo alguma eficiência.

Estes resultados animadores dependem contudo do pressuposto de que o erro de medida é independente dos regressores da equação. Suponha-se o seguinte modelo nas condições do modelo de regressão linear clássico,

$$Y_t = X_t\beta + \varepsilon_t \tag{6.3}$$

e o modelo para o erro de medida (6.2). Sendo assim a regressão dos dados observados corresponde a,

$$\begin{aligned} E(S_t \mid X_t) &= X_t\beta + E(\varepsilon_t \mid X_t) + E(u_t \mid X_t) \\ &= X_t\beta + E(u_t \mid X_t) \end{aligned} \tag{6.4}$$

a menos que $E(u_t \mid X_t)$ seja igual a zero a regressão linear sobre os dados observados não produzirá estimativas consistentes para os parâmetros de interesse. Há que garantir que o efeito atribuível às variáveis explicativas, e que se estima com os dados observados, é o efeito sobre a variável latente subjacente no modelo base e não um efeito do erro de medida que opera através da relação existente entre a variável latente e a variável proxy. Embora na maior parte dos casos se imponha a hipótese da independência, nestas situações a presença de erro de medida não pode ser ignorada visto o erro de medida conter informação que é explicada pelos os regressores.

Em modelos de regressão mais complicados, como é o caso de modelos de regressão com variável dependente qualitativa a presença de erro de medida na variável dependente não pode ser ignorada. Situações em que variáveis discretas são medidas com erro denominam-se de má classificação. A abordagem introdutória à tipologia dos modelos para o erro de medida não se aplica a estes casos. Os modelos para o erro de medida anteriormente considerados, não são aqui aplicáveis já que pela característica da variável dependente o erro de medida só pode assumir valores discretos (ver Klepper, S., 1988 sobre variáveis discretas explicativas medidas com erro). Neste tipo de situação, a probabilidade de existir

erro de medida é em geral função do valor assumido pela variável dependente. Nestas condições estamos perante erro de medida diferencial. A análise nestas condições torna-se mais complexa visto que a variável S deixa de ser uma substituta, passando a ser uma variável que contem outro tipo de informação que não existe em Y. Assim, na presença de observações mal classificadas, erro de medida diferencial impõem dificuldades adicionais, tanto em detalhes técnicos como na formulação do problema. Toda a abordagem é diferente à excepção da análise da máxima verosimilhança, onde o conhecimento das várias distribuições relevantes continua a ser imprescindível. Em modelos de regressão não linear como o Logit e o Poisson a presença de má classificação produz sempre estimativas inconsistentes para os parâmetros de interesse.

6.2 Aproximação para Variância do Erro Pequena

A análise dos efeitos da presença de erro de medida na variável dependente, quase sempre conduz à conclusão de que a sua presença não põe em causa a estrutura do verdadeiro modelo. Os procedimentos e métodos estatísticos utilizados, continuam a ser válidos, sendo apenas introduzido um elemento adicional de variabilidade nas estimativas. O custo da utilização deste tipo de dados mede-se em termos de perca de eficiência.

Esta secção tem como objectivo, partindo de um modelo paramétrico para as variáveis não observadas construir uma aproximação para variância pequena às distribuições e momentos contaminados por erro de medida. Pretende-se desta forma saber como é que algumas distribuições estatisticamente relevantes e alguns momentos se alteram quando usamos a variável dependente contaminada com erro. Mais uma vez a metodologia adoptada é semelhante à de Chesher (1990) apresentada na secção 4.3.

Sejam Y e X vectores de variáveis aleatórias com componentes Y_t e X_t $(t = 1, ..., N)$ com distribuição conjunta dada por $f_{Y|X}(y \mid x) f_X(x)$ onde y e x são possíveis realizações das variáveis Y e X. Esta metodologia aplica-se a casos em que Y é continuamente distribuída[19] .

Esta secção preocupa-se com situações em que queremos obter informação

[19]O que de certa forma limita esta análise porque não engloba importantes estruturas como os modelos de variável dependente discreta.

sobre $f_{Y|X}(y \mid x)$ mas não observamos as realizações da variável Y. Ao invés, observamos uma variável proxy $S_t = Y_t + \sigma u_t$ com componentes S_t e com $\sigma \geq 0$.

A variável u tem distribuição contínua $f_u(u)$ independente de Y e de X com $E(u) = 0$ e $Var(u) = 1$, enquanto σ é um escalar positivo. Às realizações de σu, Y e S correspondem respectivamente o erro de medida, a variável dependente livre de erro e a variável dependente contaminada com erro.

Nestas condições a distribuição conjunta de Y, X, e u é dada por,

$$f_{YXu}(y, x, u) = f_{Y|X}(y \mid x) f_X(x) f_u(u) \tag{6.5}$$

O estudo das implicações da utilização das variáveis observadas leva-nos a considerar as seguintes transformações: $Y = S - \sigma u$; $X = X$ e $u = u$ que nos conduz ao seguinte Jacobiano da transformação:

$$J = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\sigma & 1 \end{vmatrix} = 1 \tag{6.6}$$

Desta forma a distribuição conjunta de S, X e u é

$$f_{SXu}(s, x, u) = f_{Y|X}(s - \sigma u \mid x) f_X(x) f_u(u) \tag{6.7}$$

onde σ é um escalar e s e w são respectivamente vectores coluna de realizações da variável dependente observada e do erro de medida.

Usando novamente a metodologia proposta em Chesher (1990), esta densidade pode ser aproximada através de uma expansão em série de Taylor em torno de $\sigma = 0$, reunindo-se os termos até à segunda ordem,

$$f_{SXu}(s, x, u) = f_X(x) f_u(u) \left[f_{Y|X}(s \mid x) - u\sigma f^{i}_{Y|X}(s \mid x) + \frac{1}{2} u^2 \sigma^2 f^{ij}_{Y|X}(s \mid x) \right] \tag{6.8}$$

onde os superescritos representam respectivamente a primeira e segunda derivadas da densidade em relação ao vector y. Dado que se assume que X e u são variáveis aleatórias independentes, a aproximação à densidade condicional da variável proxy S dadas as variáveis X e u é dada pela expressão dentro dos parêntesis.

Sendo o objectivo central da econometria o estudo dos efeitos de um conjunto de variáveis sobre uma dada variável será de máximo interesse achar a aproximação à distribuição da variável observada condicional apenas no conjunto das variáveis explicativas

$$\begin{aligned} f_{S|X}(s \mid x) &= \int f_{S|Xu}(s \mid x,u)\, f_{u|X}(w \mid x)\, du = \\ &= f_{Y|X}(s \mid x) + \tfrac{1}{2}\sigma^2 f^{ij}_{Y|X}(s \mid x) + o(\sigma^2) \end{aligned} \tag{6.9}$$

Esta expressão dá-nos a aproximação à distribuição condicional de S quando consideramos que o erro de medida é pequeno, i.e., $\sigma \to 0$.

Apesar deste ter sido o resultado encontrado para a aproximação à densidade sobre análise é possível encontrar outras aproximações para $f_{S|X}(s \mid x)$ que até à ordem considerada lhe são equivalentes,

$$f_{Y|X}(s \mid x) \exp\left[\frac{\sigma^2}{2}\frac{f^{ij}_{Y|X}(s|x)}{f_{Y|X}(s|x)}\right] + o(\sigma^2)$$

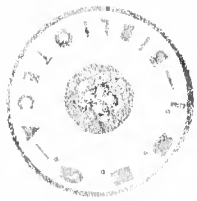
$$f_{Y|X}(s \mid x)\left[1 + \frac{\sigma^2}{2}\frac{f^{ij}_{Y|X}(s|x)}{f_{Y|X}(s|x)}\right] + o(\sigma^2)$$

$$f_{Y|X}(s \mid x)\left[1 + \tfrac{\sigma^2}{2}\left(F^{ij}_{Y|X}(s \mid x) + F^{i}_{Y|X}(s \mid x)\, F^{j}_{Y|X}(s \mid x)\right)\right] + o(\sigma^2)$$

onde $F_{Y|X}(s \mid x)$ representa o logaritmo da densidade de Y condicional em X avaliada na variável medida com erro.

Estas expressões vão permitir estudar de uma forma genérica as consequências sobre a forma funcional da relação entre as variáveis, ou seja, os desvios que ocorrem na densidade dos dados observados relativamente à que efectivamente gera os dados. Contudo quando se equaciona a possibilidade de utilização desta expressão para a realização de procedimentos estatísticos formais, onde se utiliza o facto de estarmos perante uma função densidade, mais uma vez aquela aproximação não constitui um modelo válido. O problema consiste no facto daquela expressão não integrar um. Não nos é possível garantir que o integral $\int f^{ij}_{Y|X}(s \mid x)\, ds$ integre zero. Tal, deve-se ao facto do domínio de integração coincidir com o domínio de diferenciação e deste modo não ser possível permutar as duas operações.

Além deste problema, quando σ fôr significativamente diferente de zero, nada nos garante que aquela expressão não possa assumir valores negativos. Contudo esta questão pode ser contornada se considerarmos outras aproximações da mesma ordem que garantam a positividade daquela expressão.



Como já foi referido, a aproximação mostra-nos como é que, quando estamos perante erro de medida nas condições definidas, a densidade condicional da amostra observável se relaciona com a densidade do verdadeiro modelo.

Da análise da expressão é imediatamente perceptível que mais uma vez a aproximação não depende do conhecimento da forma funcional da distribuição do erro de medida. O que separa esta aproximação da verdadeira densidade condicional é o segundo termo da expressão. Este representa a concavidade da distribuição condicional da verdadeira densidade avaliada na variável proxy.

Se a variável estiver correctamente medida, situação em que o erro de medida seria nulo, então as duas densidades condicionais coincidiríam e todos os procedimentos estatísticos formais seriam válidos possibilitando a estimação consistente dos parâmetros. Caso contrário, quando $\sigma \neq 0$, há que ter em conta a interpretação do segundo termo da expressão e as suas consequências.

Nas zonas onde a densidade $f_{Y|X}(y \mid x)$ é convexa o termo de correcção é positivo e nas zonas onde a densidade $f_{Y|X}(s \mid x)$ é côncava este termo é negativo. Consequentemente o efeito de primeira ordem da presença de erro de medida traduz-se (graficamente) na elevação da densidade observável onde ela é convexa e na sua depressão onde é côncava. Como na maior parte dos casos de interesse estamos perante distribuições unimodais (côncavas perto da moda e convexas nas abas), este efeito traduz-se na elevação das abas e na diminuição do pico da densidade, onde o efeito atinge o sua amplitude máxima. Desta forma, o efeito de primeira ordem, reflecte-se num suavizamento da curva que representa a distribuição condicional da variável não observada, retirando-lhe "informação" ao aumentar a dispersão. Quanto à manutenção ou não da forma funcional do modelo original esta questão depende do caso particular que se está a estudar.

Um caso importante e bastante especial é quando Y condicional em X segue uma distribuição univariada normal com função de regressão linear $E_{Y|X}(y \mid x) = x\beta$ e $Var_{Y|X}(y \mid x) = \omega^2$. A função densidade é dada por,

$$f_{Y|X}(y \mid x, \beta, \omega) = \frac{1}{\omega\sqrt{2\pi}} \exp\left[-\frac{1}{2\omega^2}(y - x\beta)'(y - x\beta)\right] \qquad (6.10)$$

utilizando uma das aproximações possíveis, após algumas simplificações obtemos a seguinte aproximação,

$$f_{S|X}(s \mid x,\beta,\omega) = \frac{1}{\omega\sqrt{2\pi}} \exp\left\{-\frac{1}{2\omega^2}\left[(s-x\beta)'(s-x\beta)\left(1-\frac{\sigma^2}{\omega^2}\right)+\sigma^2\right]\right\} \tag{6.11}$$

É de notar que quando $\sigma^2 = 0$, situação em que não existe erro de medida a densidade condicional em X da variável dependente observada coincide com a verdadeira densidade.

Neste exemplo está bem patente o efeito de suavizamento. Quando estamos na moda da distribuição, onde $s - x\beta = 0$ o valor da densidade aproximada é inferior ao da verdadeira densidade, o que se traduz na depressão do pico da distribuição. Logo, e de acordo com o resultado anteriormente estabelecido a distribuição condicional de S em X está abaixo no pico e a cima nas abas, da distribuição de Y condicional em X.

O caso da distribuição normal é muito particular dado que a forma funcional da distribuição não se altera. Até à ordem aqui considerada podemos escrever,

$$f_{S|X}(s \mid x,\beta,\omega) = \frac{1}{\omega\sqrt{2\pi}} \exp\left(-\frac{\sigma^2}{2\omega^2}\right) \exp\left[-\frac{1}{2\omega^2}(s-x\beta)'(s-x\beta)\left(1-\frac{\sigma^2}{\omega^2}\right)\right] \tag{6.12}$$

que equivale a,

$$f_{S|X}(s \mid x,\beta,\omega) \propto \exp\left[-\frac{1}{2}\frac{\omega^2-\sigma^2}{\omega^4}(s-x\beta)'(s-x\beta)\right] \tag{6.13}$$

Desta forma mostra-se que, não só a forma funcional da densidade permanece inalterada, mas também a regressão linear que caracteriza o esperado condicional $E_{S|X}(s \mid x) = x\beta$. Contudo a presença de erro de medida tem como consequência o aumento da variância condicional - $Var_{S|X}(s \mid x) = \frac{\omega^2-\sigma^2}{\omega^4}$ que apesar de tudo continua a ser uma função cedástica e independente dos regressores.

Seria também interessante analisar o efeito sobre o valor esperado condicional provocado pela contaminação por erro de medida.

À semelhança do anteriormente definido, seja Y a variável dependente não observada e seja $g(Y)$ uma função escalar dessa variável cujo valor esperado condicional em $X = x$ queremos conhecer, $E_{Y|X}[g(Y) \mid x]$. A presença de erro de medida origina que os procedimentos estatísticos aplicados a $g(S)$ e x nos dêem informação sobre $E_{S|X}[g(S) \mid x]$. Posto isto é importante saber qual a relação

entre o valor esperado condicional desejado e o valor esperado condicional sobre o qual temos informação. Sabendo que,

$$E_{S|X}\left[g\left(S\right) \mid x\right] = \int g\left(S\right) f_{S|X}\left(s \mid x\right) ds \tag{6.14}$$

e usando a aproximação à densidade condicional (6.9) obtem-se

$$E_{S|X}\left[g\left(S\right) \mid x\right] = E_{Y|X}\left[g\left(S\right) \mid x\right] + \frac{1}{2}\sigma^2 \int g\left(S\right) f^{ij}_{Y|X}\left(s \mid x\right) ds + o\left(\sigma^2\right) \tag{6.15}$$

onde $f^{ij}_{Y|X}\left(s \mid x\right)$ representa a derivada de segunda ordem da densidade em ordem ao vector y avaliada em s. Integrando o segundo termo por partes duas vezes e assumindo que nos extremos da distribuição de Y dado X a densidade e a primeira derivada são nulas, obtemos o seguinte resultado:

$$E_{S|X}\left[g\left(S\right) \mid x\right] = E_{Y|X}\left[g\left(S\right) \mid x\right] + \frac{1}{2}\sigma^2 E_{Y|X}\left[g^{ij}\left(S\right) \mid x\right] + o\left(\sigma^2\right) \tag{6.16}$$

Através desta expressão é possível concluir que quando a função $g\left(S\right)$ é linear, a presença de erro de medida não altera o valor esperado. Quando $g\left(S\right)$ é uma função convexa (como quando calculamos variâncias e outros momentos de ordem superior) o uso de dados contaminados para calcular esses momentos provoca um enviesamento positivo. O contrário acontece quando a função é côncava (médias logarítmicas), onde o enviesamento vem negativo.

6.3 Máxima Verosimilhança

6.3.1 Verosimilhança quando Y não é observado

Suponha-se que determinada variável dependente tem distribuição condicional num dado conjunto de variáveis explicativas dada por $f_{Y|ZX}\left(y \mid z, x, \beta\right)$. Suponha-se ainda que a variável dependente Y não é observável mas apenas uma sua proxy S, cuja distribuição condicional na verdadeira variável dependente e nas restantes variáveis explicativas é representada por $f_{S|YZX}\left(s \mid y, z, x, \gamma\right)$ sendo o parâmetro γ o vector que caracteriza essa distribuição. À semelhança do que sucedia com o erro de medida nos regressores, S é uma variável substituta se a sua distribuição depender apenas da verdadeira resposta, ou seja, $f_{S|YZX}\left(s \mid y, z, x, \gamma\right) = f_{S|Y}\left(s \mid y, \gamma\right)$. Esta definição implica que toda a informação patente na relação entre S e os regressores é explicada ou está patente na variável dependente latente.

Partindo da distribuição conjunta das duas variáveis dependentes condicionais nos regressores,

$$f_{SY|ZX}(s, y \mid z, x, \beta, \gamma) = f_{S|YZX}(s \mid y, z, x, \gamma) f_{Y|ZX}(y \mid z, x, \beta) \qquad (6.17)$$

Considerando o caso mais geral, em que o modelo erro de medida depende também das variáveis explicativas, a distribuição condicional dos dados observados é dada por,

$$f_{S|ZX}(s \mid z, x, \beta, \widehat{\gamma}) = \int f_{Y|ZX}(y \mid z, x, \beta) f_{S|YZX}(s \mid y, z, x, \widehat{\gamma}) \, d\mu(y) \qquad (6.18)$$

No caso mais comum de S ser uma variável substituta, ou, sob a hipótese de erro de medida não diferencial, $f_{S|YZX}(s \mid y, z, x, \widehat{\gamma})$ é substituída por $f_{S|Y}(s \mid y, \widehat{\gamma})$. Neste caso particular, mas mais usual, se não houver uma relação estatística entre a verdadeira variável dependente e os regressores, ($\beta = 0$) também não haverá entre a variável proxy e as variáveis explicativas. Isto acontece porque nenhum dos termos no integral (somatório) depende dos regressores. O primeiro porque $\beta = 0$, e o segundo porque S é uma variável substituta. Este resultado é bastante interessante porque, se o objectivo fôr determinar se os regressores de alguma forma explicam a variável dependente, a validade dos testes de hipótese baseados na regressão naive não é posta em causa pela presença de erro de medida. Novamente, a utilidade desta análise é limitada pela necessidade de conhecimento ou construção de um modelo para o erro de medida. Quando conhecidas todas as componentes a verosimilhança da amostra é o produto da expressão (6.18) para cada observação.

6.3.2 Verosimilhança quando Y é parcialmente observado

Suponha-se agora que para um subconjunto da amostra é possível observar a verdadeira resposta. Estas observações estão indexadas por $\Delta_t = 1$. Para este subconjunto é possível modelizar a distribuição de S condicional em Y, Z e X, ou seja, $f_{S|YZX}(s \mid y, z, x, \widehat{\gamma})$. A verosimilhança da amostra passa então a ser e dada por,

$$\prod_{t=1}^{n}\left[\{f\left(S_t \mid Y_t, Z_t, X_t, \widehat{\gamma}\right) f\left(Y_t \mid Z_t, X_t, \beta\right)\}^{\Delta_t} \times \{f\left(S_t \mid Z_t, X_t, \widehat{\gamma}, \beta\right)\}^{1-\Delta_t}\right] \tag{6.19}$$

onde o segundo termo é dado pela expressão (6.18). Desta forma é possível construir o modelo erro de medida, que possibilita a utilização de toda a informação disponível para caracterizar a distribuição de S dado Z e X. O modelo para a distribuição de S dado (Y, Z, X) assume uma importância fundamental na função verosimilhança. No caso particular em que S é uma variável discreta, uma das soluções para a sua modelização consiste na utilização de um Logit Multinomial onde a variável S pode assumir os níveis $(1, 2, ..., S)$, Carroll, Ruppert & Stefanski (1995).

A utilização dos métodos associados à máxima verosimilhança são em princípio de fácil aplicação e interpretação. Contudo existem duas dificuldades associadas a esta abordagem. A primeira é que é necessário construir-se um modelo para a distribuição de S dado (Y, Z, X). Esta relação tem de ser estimada levantando-se todo o tipo de questões ligadas à robustez. O segundo é que a resolução da expressão (6.18) requer uma integração ou somatório numérico, o que pode tornar estas expressões extremamente complexas necessitando do recurso a pesados métodos computacionais.

6.4 Má classificação no modelo de Regressão Poisson

6.4.1 Introdução

Considere-se o caso em que existe uma variável aleatória discreta que assume valores $0, 1, 2, ...,$. Os valores assumidos pela variável são dados de contagem, ou seja, o número de vezes que determinado fenómeno ocorre num determinado período de tempo. Sejam as realizações dessa variável aleatória representadas por, $Y_1, Y_2, ..., Y_N$ e admita-se que sejam processos Poisson mutuamente independentes com média $\lambda_1, \lambda_2, ..., \lambda_N$ respectivamente, onde N corresponde ao número de observações na amostra. Neste caso a função densidade probabilidade da t-ésima observação é dada por,

$$\Pr(Y_t = j) = \exp(-\lambda_t)\frac{(\lambda_t)^j}{j!} \quad j = 1, 2, ... \tag{6.20}$$

Interessa considerar situações em que o parâmetro de interesse da distribuição Poisson é função de um conjunto de variáveis explicativas e de um vector de parâmetros. Sendo assim, Y_t é uma variável que condicional em X_t, segue uma distribuição Poisson, com valor esperado condicional $\lambda(X_t, \beta)$[20] . Dada a existência de uma amostra constituída por um vector de dados de contagens e observações do conjunto das variáveis explicativas, a utilização do método da máxima verosimilhança permite a estimação consistente do vector de parâmetros β.

Suponha-se que existe um problema de má classificação na variável dependente que faz com que uma proporção dos zeros seja registada como assumindo o valor um. A utilização desta amostra mal classificada, tem como consequência, inconsistência nas estimativas dos parâmetros da regressão Poisson. Este tipo de problema com a amostra pode acontecer, se alguns indivíduos por algum motivo não queiram responder o valor zero quando questionados e por isso reportem o valor um. Este tipo de comportamento pode muito facilmente ocorrer se a não realização de um determinado acontecimento acarretar alguma forma de estigma social. Considerem-se a título de exemplo as seguintes questões: Número de livros que leu no último mês, número de jornais que leu no último fim-de-semana, número de idas ao estrangeiro nos últimos anos etc. Em qualquer um destes casos, a resposta "zero" representa sempre uma situação social inferior que os entrevistados podem não estar dispostos a revelar. Neste contexto, a resposta "um" representa um mínimo socialmente aceitável e é de supor que alguns deles estejam de facto mal classificados.

6.4.2 O modelo

Seja S_t com $t = 1, 2, ..., N$ a amostra mal classificada e θ a probabilidade da observação com valor zero estar correctamente classificada. Sendo assim é necessário

[20]Geralmente assume-se que $\lambda(X_t, \beta) = \exp(X_t\beta)$ de modo a garantir a positividade e monoticidade daquele parâmetro condicional. Por conveniência na notação assuma-se que $\lambda(X_t, \beta) \equiv \lambda_t$.

reespecificar o modelo tendo em conta que a probabilidade da t-ésima observação ser zero é dada por:

$$\begin{aligned} \Pr(S_t = 0) &= \theta . \Pr(Y_t = 0) \\ &= \theta . \exp(-\lambda_t) \end{aligned} \tag{6.21}$$

e que, paralelalemente os uns observados derivam de:

$$\begin{aligned} \Pr(S_t = 1) &= (1-\theta) . \Pr(Y_t = 0) + \Pr(Y_t = 1) \\ &= \exp(-\lambda_t) . [\lambda_t + (1-\theta)] \end{aligned} \tag{6.22}$$

Definido como parâmetro que determina a presença de má classificação na amostra, θ está obviamente sujeito à restrição de estar entre zero e um[21] . A correcção das probabilidades afectadas pelo erro mostra-nos que a probabilidade de encontrarmos um zero na amostra é de facto inferior ao que na realidade acontece, traduzindo o facto de os zeros observados na amostra contaminada estarem subrepresentados, dado que a verdadeira variável seguir um processo Poisson. Paralelamente, estando os zeros mal classificados nos uns, na ausência de correcção a probabilidade de ocorrer o valor um na amostra está sobreavaliada.

É de supor que as distorções introduzidas pela presença de má classificação tenham consequências sob os momentos condicionais. Tendo Y_t, a verdadeira variável dependente, uma distribuição Poisson, λ_t o parâmetro associado a esta distribuição, representa simultaneamente o valor esperado e variância da distribuição condicional nos valores de X_t. A utilização de dados contaminados no cálculo do valor esperado condicional, conduz porém às expressões para os primeiro e segundo momentos,

$$E(S_t) = (1-\theta) . \exp(-\lambda_t) + \lambda_t \tag{6.23}$$

e

$$Var(S_t) = (1-\theta) . \exp(-\lambda_t) . [1 - (1-\theta) . \exp(-\lambda_t) - 2\lambda_t] + \lambda_t \tag{6.24}$$

A análise destas expressões permite-nos explicar a inconsistência da utilização do modelo Poisson nestas condições. A estimação pela máxima verosimilhança dos

[21]No limite, quando $\theta = 0$ esta especificação corresponde a um tipo de modelo censurado nos zeros muito especial porque, a massa de probabilidade correspondente a $Y_t = 0$ está toda concentrada em $Y_t = 1$.

parâmetros de interesse usando a especificação que não considera a presença de dados mal classificados não está correcta. Não só o valor esperado condicional difere de λ_t e simultaneamente a forma como o valor esperado depende de X, como é fácil provar-se que $E(S_t) > Var(S_t)$. Estamos perante um problema que provoca subdispersão[22] .

O problema da inconsistência das estimativas da máxima verosimilhança produzidas pela utilização da amostra mal classificada pode ser também estudado por recurso a resultados básicos da teoria da máxima verosimilhança. De acordo com esta, o estimador da máxima verosimilhança é a solução de uma equação não linear que iguala o vector do valor esperado do score a zero. Daí resulta que o vector score da t-ésima observação respeita a condição $E\{S_t(\beta)\} = 0$. Na ausência de erro o score de uma regressão Poisson é dado por:

$$S_t(\beta) = (y_t - \lambda_t)\frac{\partial \ln \lambda_t}{\partial \beta_s} \tag{6.25}$$

sendo fácil mostrar que aquela condição verifica-se para todas as observações. Dado que as probabilidades alteraram-se e consequentemente o valor esperado condicional da variável dependente, na presença de uma amostra com este tipo de erro de medida o valor esperado do Score do modelo Poisson vem,

$$E\{S_t(\beta)\} = ((1-\theta).\exp(-\lambda_t))\frac{\partial \ln \lambda_t}{\partial \beta_s} \tag{6.26}$$

Quando $\theta = 1$, aquele valor esperado anula-se. Com $\theta \neq 1$ para todas as observações o valor esperado do score muito dificilmente será igual a zero, traduzindo-se a utilização da especificação Poisson em estimativas inconsistentes para os parâmetros.

Sendo assim a constatação da existência deste problema exige que o modelo válido para a estimação dos parâmetros de interesse seja reespecificado. Considerando as probabilidades anteriormente definidas, a logverosimilhança da t-ésima observação do modelo Poisson com este tipo de erro na variável dependente é dada por:

[22] É fácil mostrar que enquanto o valor esperado condicional da variável observada aumenta sempre, a variância aumenta sempre menos e dependendo do valor de λ_t pode até diminuir, e que desde que λ_t seja superior a 0,5 as observações são artificialmente (via erro de medida) aproximadas à média.

$$\ln L_t(\theta, \lambda_t, \beta) = (1 - D_{t0}) \cdot (\ln \theta - \lambda_t) + (1 - D_{t1}) \cdot [\ln(\lambda_t + (1 - \theta)) - \lambda_t] + \\ + (D_{t0} + D_{t1} - 1) \cdot (j \cdot \ln \lambda_t - \lambda_t - \ln j!) \tag{6.27}$$

onde D_{t0} e D_{t1} representam duas variáveis binárias tais que:

$$D_{t0} = \begin{cases} 0 \text{ se } j = 0 \\ 1 \text{ se } j \neq 0 \end{cases} \quad \text{e} \quad D_{t1} = \begin{cases} 0 \text{ se } j = 1 \\ 1 \text{ se } j \neq 1 \end{cases} \tag{6.28}$$

A logverosimilhança da amostra é a soma para todas as observações daquela expressão. Note-se que a última fracção da logverosimilhança, aplicável para valores de $j \neq 0, 1$ é a expressão usual para o problema do modelo de regressão Poisson, dado que assumiu-se que o problema de má classificação não atinge as observações que não assumam aquele valor.

A maximização desta verosimilhança em ordem aos parâmetros de interesse, onde se inclui o parâmetro da má classificação é fácil de se encontrar. Diferenciando a expressão (6.27) em ordem a β_s e a θ e igualando a zero, obtêm-se o sistema de equações não lineares que representam as condições de primeira ordem para este problema:

$$\frac{\partial \ln L_t(\cdot)}{\partial \beta_s} = (1 - D_{t0}) \cdot (-\lambda_t) \cdot \frac{\partial \ln \lambda_t}{\partial \beta_s} + (1 - D_{t1}) \cdot \left(\frac{\lambda_t}{\lambda_t + (1 - \theta)} - \lambda_t \right) \cdot \frac{\partial \ln \lambda_t}{\partial \beta_s} + \\ + (D_{t0} + D_{t1} - 1) \cdot (j - \lambda_t) \cdot \frac{\partial \ln \lambda_t}{\partial \beta_s} \tag{6.29}$$

e,

$$\frac{\partial \ln L_t(\cdot)}{\partial \theta} = (1 - D_{t0}) \cdot \frac{1}{\theta} + (1 - D_{t1}) \cdot \left(\frac{-1}{\lambda_t + (1 - \theta)} \right) \tag{6.30}$$

À excepção dos termos indexados por $D_{t1} = 0$, as contribuições para o score das demais observações são idênticos aos do modelo Poisson para a verdadeira variável dependente. A presença deste tipo de má classificação, apenas induz alterações nas condições de primeira ordem dos parâmetros β, para as observações em que $j = 1$. Esta constatação permite-nos simplificar aquela expressão e simultaneamente dar uma nova interpretação ao score,

$$\frac{\partial \ln L_t(\cdot)}{\partial \beta_s} = (j - \lambda_t) \cdot \frac{\partial \ln \lambda_t}{\partial \beta_s} + (1 - D_{t1}) \cdot \left(\frac{\theta - 1}{\lambda_t + (1 - \theta)} \right) \cdot \frac{\partial \ln \lambda_t}{\partial \beta_s} \qquad (6.31)$$

Desta forma não só a expressão é simplificada, sendo apenas necessário recorrer a um indicador, como o score de β passa a ser apenas a soma do score habitual de um modelo Poisson, com a diferença entre o mesmo score para as observações em que $j = 1$ e o score corrigido para o mesmo valor de j. É fácil demonstrar que dado θ este score é uma função côncava em todo o domínio o que garante a identificação global do parâmetro β_s (ver Davidson & Mackinnon,1993 sobre identificação).

A expressão do score de θ permite-nos observar que este parâmetro é estimado usando apenas as observações para as quais $j = 0, 1$. Tal já seria de esperar dadas as características deste tipo de erro de medida. Contudo este facto pode constituir uma séria limitação em amostras finitas, visto que se houver poucas observações cuja variável proxy assuma este valor a estimação deste parâmetro com o mínimo de precisão pode ser bastante difícil.

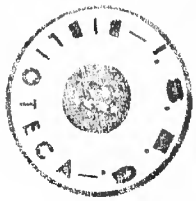
Dado que ambos os scores dependem dos parâmetros β e θ, a sua estimação terá de ser feita em simultâneo, através da resolução do sistema de equações não linear constituído pelas condições de primeira ordem. Outra implicação deste problema é que o valor esperado da matriz de informação de Fisher não é diagonal por blocos. É fácil mostrar que,

$$E\left[\frac{\partial^2 \ln L_t(\cdot)}{\partial \beta_s \partial \theta} \right] = (1 - D_{t1}) \cdot \left(\frac{\lambda_t}{[\lambda_t + (1 - \theta)]^2} \right) \cdot \frac{\partial \ln \lambda_t}{\partial \beta_s} \qquad (6.32)$$

que nunca iguala a zero.

Alternativamente, é possível estimar consistentemente os parâmetros do modelo utilizando o Método dos Momentos Generalizado. De acordo com esta técnica de estimação, se um modelo estiver correctamente especificado, haverão momentos condicionais, que dependem dos parâmetros de interesse, que serão zero. Tendo-se previamente definido os primeiro e segundo momentos condicionais desta especificação, considerem-se as seguintes condições de momentos[23] :

[23]Sob algumas condições de regularidade que garantem a existência de um único vector $\widehat{\beta}$ e $\widehat{\theta}$ que anula aquelas condições.



$$E\left(S_t - \mu_t\left(X_t, \beta, \theta\right)\right) = 0$$

e

$$E\left(S_t^2 - \vartheta_t\left(X_t, \beta, \theta\right)\right) = 0 \tag{6.33}$$

onde $\mu_t\left(X_t, \beta, \theta\right)$ é a média da variável mal classificada definida na expressão (6.23) e $\vartheta_t\left(X_t, \beta, \theta\right) = E(S_t^2 \mid X)$. Garantindo-se a identificação dos parâmetros do modelo, as suas estimativas são a solução das contrapartidas empíricas das condições de momentos, que neste caso são dadas por:

$$\frac{1}{n}\sum_{t=1}^{n}\left[S_t - \left(\left(1-\widehat{\theta}\right).\exp\left(-\widehat{\lambda}_t\right) + \widehat{\lambda}_t\right)\right].X_{ts} = 0 \text{ com } s = 1, ..., k$$

e

$$\frac{1}{n}\sum_{t=1}^{n}\left[S_t^2 - \left(\left(1-\widehat{\theta}\right).\exp\left(-\widehat{\lambda}_t\right) + \widehat{\lambda}_t + \widehat{\lambda}_t\right)\right].X_{ts} = 0 \text{ com } s = 1, ..., k \tag{6.34}$$

Estas equações representam condições de ortogonalidade, segundo as quais, $S_t^m - E\left(S_t^m\right)$ e $m = 1, 2$ devem ser ortogonais aos regressores[24] . Considerando estas duas equações estamos perante o caso em que o número de condições de momentos empíricas é superior ao número de parâmetros. Sendo assim, se as condições de momentos não forem funcionalmente dependentes, aquele sistema de equações é sobreidentificado. De modo a utilizar-se toda a informação da amostra e sobre a estrutura da especificação (nomeadamente admitindo a existência de subdispersão), é necessário reconciliar as estimativas dos parâmetros produzidas pelo sistema sobreidentificado, mediante a aplicação do Método dos Momentos Generalizados. Recorrendo ao resultado de Hansen (1982), este conduz à minimização de uma função objectivo quadrática da forma,

$$Argmin_{\beta,\theta}\ q = \overline{m}' V^{-1} \overline{m} \tag{6.35}$$

onde $\overline{m}$ representa o vector de condições de ortogonalidade, constituído por $\overline{m}_j$ $j = 1, ..., 2k$ tal que

$$\overline{m}_j = \frac{1}{n}\sum_{t=1}^{n} m_j\left(s_t, x_t, \beta, \theta\right) \tag{6.36}$$

[24]Pode-se considerar adicionalmente os scores definidos pelas condições de primeira ordem do problema da máxima verosimilhança, já que o seu valor esperado é nulo e logo constituem condições de momentos que definem um estimador GMM.

e $V = E\left[\frac{1}{n}\sum_{t=1}^{n}\overline{m.m'}\right]$ é a matriz definida positiva de variâncias e covariâncias das condições de ortogonalidade escolhida de forma a minimizar a matriz de variâncias e covariâncias assintótica do estimador GMM. São conhecidas as propriedades assintóticas deste estimador donde se destaca a normalidade das estimativas dos parâmetros, sendo ainda possível testar a validade das condições de ortogonalidade utilizando o teste **D** de Newey & West (1987b).

É ainda possível considerar uma especificação mais genérica, segundo a qual a probabilidade de má classificação é função de um conjunto de variáveis explicativas, que pode ou não ser o mesmo que condiciona a variável dependente. Seguindo a sugestão de Lambert (1992) para o modelo ZIP (Zero Inflated Poisson), suponha-se que essa dependência é especificada por um modelo Logit tal que:

$$logit\,(\theta) = \ln\left(\theta/\left(1-\theta\right)\right) = Z_t\delta \tag{6.37}$$

As variáveis que condicionam o modelo Poisson podem ou não ser as mesmas que condicionam o Logit para o parâmetro θ[25] .

A especificação do problema decorrente da introdução desta hipótese permanece idêntica, bastando substituir θ por $\exp\left(Z_t\delta\right)/\left\{1+\exp\left(Z_t\delta\right)\right\}$ e ter em conta que os vectores score de interesse são agora $\partial \ln L_t\left(\cdot\right)/\partial\delta_s = \partial \ln L_t\left(\cdot\right)/\partial\theta \cdot \partial\theta/\partial\delta_s$ e $\partial \ln L_t\left(\cdot\right)/\partial\beta_s$ que igualados a zero conduzem ás condições de primeira ordem de interesse.

6.5 Simulações

De modo a avaliar as consequências deste tipo de má classificação da variável dependente, sobre as estimativas dos parâmetros e a capacidade da logverosimilhança corrigida para estimar os verdadeiros parâmetros, com particular relevo para θ, procedeu-se a um pequeno estudo de Monte Carlo. Gerou-se um mod-

[25]Esta especificação permite, a consideração de várias hipóteses quanto à relação entre θ e λ:(i) as variáveis que afectam o modelo Poisson podem ser as mesmas que condicionam o parâmetro de má classificação, mas λ e θ não estão relacionados, e sendo assim temos $2k$ parâmetros a estimar (se $dim(X_t) = dim(Z_t)$);(ii) quando Z_t é uma coluna de uns, o que equivale a ter a especificação anteriormente considerada;(iii) e $Z_t = X_t$ e λ e θ estão relacionados sendo possível reduzir o número de parâmetros a estimar a metade. Esta hipótese é discutida em Lambert (1992) assumindo a forma de uma reparametrização do modelo.

elo Poisson condicional numa variável explicativa e seguidamente criou-se uma amostra para a variável dependente onde uma percentagem dos zeros estão classificados como um. Consideraram-se dois modelos com apenas uma variável explicativa e onde $\lambda(X\beta) = \exp(\beta_0 + \beta_1 X_1)$,

$$(I): \ (\beta_0, \beta_1) = (1, 5, -2) \text{ e } X_1 \sim U(0,1)$$
$$(II): \ (\beta_0, \beta_1) = (1, -1) \text{ e } X_1 \sim Normal(0,1)$$

O modelo I tem média incondicional 1,94 e 22,5% de zeros, enquanto o modelo II tem média 4,64 e apenas 15,14% de zeros. Consideraram-se ainda duas probabilidades de má classificação $1 - \theta = 0,3$ e $1 - \theta = 0,1$. As experiências foram realizadas para amostras de dimensão $N = 100, 300$ e 500 e repetidas 1500 vezes. Os resultados estão patentes na tabela 5.

Da análise dos resultados, é fácil constatar que a existência de observações da variável dependente mal classificadas, provoca um enviesamento nas estimativas de todos os parâmetros, com particular incidência sobre o parâmetro afecto à variável explicativa. Nos designs aqui considerados parece repetir-se o efeito atenuação, no sentido em que o enviesamento do parâmetro β_1 é em direcção à origem. Embora seja lícito considerar que o enviesamento é função da distribuição da variável explicativa, no caso aqui considerado mesmo uma pequena proporção de zeros mal classificados, $1 - \theta = 0,1$ é susceptível de provocar um enviesamento significativo nos parâmetros. A dimensão deste é relativamente constante nas amostras consideradas e tanto maior quanto maior $1 - \theta$.

Tabela 5: Estimativas dos Modelos Poisson e Verosimilhana corrigida

	β_0	β_1	θ	$\beta_0(\theta)$	$\beta_1(\theta)$
Modelo I: Prob(má classificação)=0,1					
N=100	1,48736	-1,94368	0,89620	1,49848	-2,01672
	(0,11907)	(0,26163)	(0,15600)	(0,12136)	(0,29183)
N=300	1,48823	-1,94123	0,89833	1,49844	-2,00652
	(0,069252)	(0,15532)	(0,091819)	(0,070573)	(0,17242)
N=500	1,48794	-1,93769	0,90195	1,49741	-1,99852
	(0,051529)	(0,12058)	(0,074181)	(0,052531)	(0,13309)
Modelo I: Prob(má classificação)=0,3					
N=100	1,46969	-1,82688	0,70018	1,49857	-2,01698
	(0,11749)	(0,24378)	(0,15094)	(0,12193)	(0,29274)
N=300	1,47007	-1,82480	0,69822	1,49844	-2,00653
	(0,068308)	(0,14539)	(0,087305)	(0,070856)	(0,17313)
N=500	1,46982	-1,82189	0,70135	1,49725	-1,99810
	(0,050743)	(0,11204)	(0,070847)	(0,052636)	(0,13328)
Modelo II: Prob(má classificação)=0,3					
N=300	1,02697	-0,98310	0,69810	0,99797	-1,00067
	(0,035511)	(0,022867)	(0,099483)	(0,038522)	(0,024040)

Desvios padrão estimados em parêntesis.

Relativamente á verosimilhança corrigida, a sua capacidade para estimar os parâmetros de interesse está bem patente nos resultados. Independentemente da amostra tanto o vector $\beta(\theta)$ como o parâmetro θ são estimados perto do verdadeiro valor, aumentando obviamente a precisão das estimativas com a dimensão da amostra. Note-se que para qualquer valor de $1-\theta$ aqui considerado os desvios padrão dos parâmetros de interesse são praticamente idênticos. Os resultados obtidos para o modelo II, cuja média da variável dependente é maior e consequentemente a percentagem de zeros na amostra menor, parece sugerir que o enviesamento provocado por uma mesma quantidade de má classificação é menor quanto maior fôr a média da distribuição de Y[26] . Neste modelo as estimativas dos parâmetros $\beta(\theta)$ continuam centradas e estimadas com maior precisão, á excepção do parâmetro θ, cujo desvio padrão é agora superior dado que foi estimado com menor número de observações que no modelo I.

Em qualquer dos casos a logverosimilhança corrigida estima com rigor os parâmetros de interesse.

6.6 Má Classificação em Modelos de Escolha Binária.

6.6.1 Introdução

Num modelo de escolha discreta, os valores da variável dependente Y_i apenas podem assumir dois valores, 1 e 0. À variável dependente está normalmente associada a ocorrência ou não de um determinado acontecimento. Neste contexto existem vários modelos de escolha discreta, donde se destacam pela sua fácil aplicabilidade e popularidade os modelos Probit e Logit. Genericamente, estes modelos são caracterizados por uma função $F(x)$ com as propriedades,

$$F(-\infty)=0,\ F(\infty)=1$$

e

$$f(x)\equiv\frac{\partial F(x)}{\partial x}>0 \tag{6.38}$$

Sendo assim $F(x)$ é uma transformação monótona crescente que projecta números reais num espaço limitado por zero e um. Os modelos de escolha discreta aqui

[26]Embora este resultado seja bastante intuitivo deve se ter em conta que nos modelos I e II a variável X_1 foi gerada de universos estatísticos diferentes o que limita a comparabilidade dos dados.

considerados consistem numa função transformação aplicada a uma função index que depende de um conjunto de regressores e dum vector de parâmetros, $F(X_t\beta)$.

Nesta secção pretende-se analisar as consequências sobre estes modelos, da presença de erro de medida na variável dependente. Sendo a variável dependente binária, a presença deste tipo de problema traduz-se no facto de em alguns casos a resposta ser registada na categoria errada; por exemplo a variável é registada em um quando o seu verdadeiro valor é zero. Este tipo de erro de medida pode facilmente ocorrer devido a erro de compreensão por parte do entrevistado, traduzindo-se numa resposta errada, ou pura e simplesmente devido a um erro por parte de quem trata os dados no seu registo.

Quando a variável dependente em causa está condicionada num conjunto de variáveis explicativas, a presença de dados mal classificados provoca inconsistência na estimativa dos parâmetros da distribuição condicional. A análise das consequências e soluções para este problema serão feitas num contexto de estimação pela Máxima Verosimilhança.

6.6.2 Modelo de Escolha Discreta Mal Classificado

Considere-se o modelo de regressão binário derivado do célebre problema da existência de uma variável latente não observável. Seja y_t^* a variável latente:

$$y_t^* = X_t\beta + \varepsilon_t \tag{6.39}$$

Ao invés de y_t^* apenas nos é dado a observar o seu sinal, que determina o valor da variável observada binária y_t de acordo com a seguinte relação:

$$\begin{aligned} y_t &= 1 \text{ se } X_t\beta + \varepsilon_t \geq 0 \\ y_t &= 0 \text{ se } X_t\beta + \varepsilon_t \leq 0 \end{aligned} \tag{6.40}$$

Defina-se π com sendo a probabilidade da variável dependente estar correctamente classificada. Assuma-se ainda que π é independente dos regressores e constante na amostra considerada em ambos os tipos de resposta. Considere-se ainda que o modelo de escolha discreta $F(X_t\beta)$ é a probabilidade de $y_t = 1$. A consideração de que existem observações da variável dependente mal classificadas sugere as seguintes alterações nas probabilidades associadas ao acontecimento:

$$\begin{aligned} \Pr(y_t = 1) &= \pi \cdot \Pr(y_t > 0) + (1 - \pi) \cdot \Pr(y_t \leq 0) \\ &= \pi \cdot F(X_t\beta) + (1 - \pi) \cdot (1 - F(X_t\beta)) \end{aligned} \tag{6.41}$$

Calculando o valor esperado de y_t obtemos:

$$E(y_t \mid X_t) = 1 - \pi + (2\pi - 1) \cdot F(X_t\beta) \tag{6.42}$$

Esta equação pode ser consistentemente estimada através dos Mínimos Quadrados Não Lineares sob a forma,

$$y_t = \alpha + (1 - 2\alpha) \cdot F(X_t\beta) + \eta_t \text{ com } \alpha = 1 - \pi \tag{6.43}$$

O mesmo problema pode ainda ser abordado considerando o outro tipo de respostas:

$$\begin{aligned} \Pr(y_t = 0) &= \pi \cdot \Pr(y_t \leq 0) + (1 - \pi) \cdot \Pr(y_t > 0) \\ &= \pi - (2\pi - 1) \cdot F(X_t\beta) \end{aligned} \tag{6.44}$$

ou, escrito em termos de α, que representa o parâmetro de má classificação,

$$\Pr(y_t = 0) = (1 - \alpha) - (1 - 2\alpha) \cdot F(X_t\beta) \tag{6.45}$$

Note-se que mais uma vez tem-se que:

$$E(y_t \mid X) = 1 - \pi + (2\pi - 1) \cdot F(X_t\beta) \tag{6.46}$$

A estimação da especificação usual do modelo de escolha discreta não considerando a presença de erro de medida quando de facto a variável dependente está mal classificada, conduz a estimativas dos parâmetros enviesadas e inconsistentes. Este resultado contrasta com a presença de erro de medida clássico, na variável dependente no modelo de regressão linear, que se traduz em estimativas dos parâmetros consistentes mas não eficientes.

Sendo assim a logverosimilhança da amostra do modelo probabilístico com variável dependente mal classificada é escrita como:

$$\begin{aligned} L_t = \sum_{t=1}^{n} \{ & y_t \cdot log[\alpha + (1 - 2\alpha) \cdot F(X_t\beta)] + \\ & (1 - y_t) \cdot log[(1 - \alpha) + (2\alpha - 1) \cdot F(X_t\beta)] \} \end{aligned} \tag{6.47}$$

Note-se que quando $\alpha = 0$, ou seja não existe erro de classificação, a logverosimilhança da amostra assume a sua expressão usual para um modelo binário[27] . Ao contrário dos modelos de regressão habituais, Logit e Probit, onde a logverosimilhança é côncava em todo o seu domínio visto ser a soma de duas funções côncavas, esta logverosimilhança corrigida tem de verificar algumas condições para que seja côncava[28] . Além disso o valor esperado da matriz de informação de Fisher para (α, β) não é diagonal por blocos nos parâmetros (ver Hausman, J.A. & Morton, F.M., 1994).

Uma questão interessante do problema da má classificação neste tipo de modelos, é que, basta uma observação mal classificada para provocar um enviesamento considerável na estimação dos parâmetros. Considerem-se as condições de primeira ordem, onde o primeiro termo é o somatório das observações indexadas por $y_t = 1$, e o segundo para as quais $y_t = 0$,

$$y \cdot \sum_t \frac{\hat{f}_t X_t}{\hat{F}_t} - (1-y) \cdot \sum_j \frac{\hat{f}_j X_j}{1-\hat{F}_j} = 0 \tag{6.48}$$

onde $\hat{F}_t \equiv F\left(X_t\hat{\beta}\right)$ e $\hat{f}_t \equiv f\left(X_t\hat{\beta}\right)$ e $\hat{\beta}$ é o vector de estimativas da máxima verosimilhança. A origem da inconsistência pode ser analisada através desta expressão. Na presença de observações incorrectamente classificadas, estas, são adicionadas à função verosimilhança e ao score através do termo incorrecto. A inconsistência nos parâmetros estimados pelo modelo surge porque as observações mal classificadas vão prever o resultado oposto àquele efectivamente verificado. Por exemplo, uma observação com um index elevado prevê um, com a probabilidade $F(X_t\beta)$ perto de um. Contudo, se essa observação estiver mal classificada, o seu valor observado será zero. A observação será incluída no segundo termo da expressão da condição de primeira ordem, onde um valor para $F(X_t\beta)$ perto de um aproxima a zero o denominador da expressão, e logo todo o termo tenderá para a infinito. Desta forma, a soma das condições de primeira ordem na presença deste erro de medida pode ser bastante elevada (quando deveria ser perto

[27]Para que a equação possa ser estimada pela máxima verosimilhança o parâmetro α tem de ser inferior a 0.5, caso contrário os dados estão de tal modo mal classificados que é impossivel identificar os parâmetros do modelo (Hausman & Scott, 1994).

[28]Hausman & Scott, (1994) definiram as condições necessárias de concavidade da logverosimilhança para os modelo Logit e Probit.

de zero) e consequentemente as estimativas dos parâmetros de interesse podem ser extremamente enviesadas.

Outra forma de se aferir da inconsistência provocada pela presença de observações mal classificadas é usar o facto de a máxima verosimilhança igualar o valor esperado do score a zero na ausência de erro na variável dependente. Reescrevendo o score da t-ésima observação como:

$$S_t(\beta) = \frac{[y_t - F(\cdot)]}{F(\cdot)[1 - F(\cdot)]} f(\cdot) X_t \tag{6.49}$$

é fácil notar que na ausência de má classificação $E[S_t(\beta)]$ é igual a zero. Contudo na presença de erro de medida desta natureza, as probabilidades alteram-se produzindo valores esperados diferentes dos habituais para uma variável binária. Usando a expressão para o valor esperado de y_t, é fácil mostrar que o valor esperado do score do modelo binário é diferente de zero. De facto,

$$E[S_t(\beta)] = \frac{f_t X_t}{F_t}[(1-\alpha)F_t + \alpha(1-F_t)] - \frac{f_t X_t}{1-F_t}[(1-\alpha)(1-F_t) + \alpha F_t] \tag{6.50}$$

Na ausência de erro, $\alpha = 0$ o valor esperado da equação acima anula-se como seria de esperar neste tipo de modelos. Se $\alpha \neq 0$, então o valor esperado daquele score, é diferente de zero, donde que estimativas para os parâmetros nestas condições utlilizando o score tradicional de um modelo binário são inconsistentes.

É interessante analisar a inconsistência nos coeficientes do Probit e do Logit produzidas por poucas observações mal classificadas, visto ser este o caso mais usual. Usando o Score corrigido para avaliar a variação nos coeficientes estimados produzida pela má classificação avaliada em $\alpha = 0$, Hausman & Scott, (1994) chegam à expressão para o caso da especificação Probit,

$$\frac{\partial \beta_k}{\partial \alpha} |_{\alpha=0} = \frac{1 - 2F(X_t\beta)}{f(X_t\beta) X_t} \tag{6.51}$$

Note-se que no caso em que a observação tem um valor para $F(X_t\beta)$ perto de zero ou de um, então $f(X_t\beta)$ tende também a estar perto de zero e logo aquela derivada pode assumir valores bastante significativos. Sendo assim, uma observação mal classificada pode, teoricamente, ser suficiente para introduzir in-

consistência nas estimativas dos parâmetros[29] .

Numa situação em que se suspeite de erro de medida desta natureza, o investigador pode suspeitar que as probabilidades de má classificação de uma categoria para outra possam não ser simétricas, como estava implícito até ao momento. Neste caso devem introduzir-se as seguintes alterações. Sejam,

$$\begin{aligned} \pi_0 &= \text{probabilidade de correcta classificação dos } 0's \\ \pi_1 &= \text{probabilidade de correcta classificação dos } 1's \end{aligned} \tag{6.52}$$

Sendo assim, sob a hipótese de probabilidades de má classificação assimétricas, a probabilidade de se observar uma dada resposta vem,

$$\begin{aligned} \Pr(y_t = 1 \mid X) &= \pi_1 \cdot \Pr(y_t > 0) + (1 - \pi_0) \cdot \Pr(y_t \leq 0) \\ &= (1 - \pi_0) + (\pi_0 + \pi_1 - 1) \cdot F(X_t\beta) \end{aligned} \tag{6.53}$$

Da mesma forma,

$$\begin{aligned} \Pr(y_t = 0 \mid X) &= \pi_0 \cdot \Pr(y_t \leq 0) + (1 - \pi_1) \cdot \Pr(y_t > 0) \\ &= \pi_0 - (\pi_0 + \pi_1 - 1) \cdot F(X_t\beta) \end{aligned} \tag{6.54}$$

A logverosimilhança da amostra sob a hipótese de diferentes probabilidades de má classificação é dada por,

$$\begin{aligned} logL_t = &\sum_{t=1}^{n} y_t \ln\left[(1 - \pi_0) + (\pi_0 + \pi_1 - 1) \cdot F(X_t\beta)\right] + \\ &+ \sum_{t=1}^{n} (1 - y_t) \ln\left[\pi_0 - (\pi_0 + \pi_1 - 1) \cdot F(X_t\beta)\right] \end{aligned} \tag{6.55}$$

Mais uma vez, para que seja possível estimar esta equação, tem de se impôr a restrição de que $\pi_0 + \pi_1$ deve ser superior a um. Caso não se verifique esta condição, a amostra estará tão contaminada por erro, que é impossível estabelecer qualquer relação entre regressores e variável dependente, mesmo recorrendo a este tipo de especificação.

Esta análise pode ser ainda mais refinada se de alguma forma se suspeitar que o parâmetro de má classificação esteja correlacionado com os y_t de uma forma linear. Desta forma as probabilidades de interesse escrevem-se como:

[29]Geralmente, a dimensão da inconsistência nas estimativas devido à má classificação depende também da distribuição dos $X's$.

$$\begin{aligned}\Pr(y_t = 1) &= \pi(y_t) \cdot \Pr(y_t > 0) + (1 - \pi(y_t)) \cdot \Pr(y_t \leq 0) \\ &= \pi(X_t\delta) \cdot F(X_t\beta) + (1 - \pi(X_t\delta)) \cdot (1 - F(X_t\beta))\end{aligned} \tag{6.56}$$

e,

$$E(y_t \mid X) = 1 - \pi(X_t\delta) + (2\pi(X_t\delta) - 1) \cdot F(X_t\beta) \tag{6.57}$$

Supondo que a probabilidade $\pi(X_t\delta)$ possa ser descrita por um modelo Logit, com função probabilidade dada por $\Lambda(X_t\delta)$. Deste modo o valor esperado de y_t dado X pode ser escrito como:

$$E(y_t \mid X) = 1 - \Lambda(X_t\delta) + (2\Lambda(X_t\delta) - 1) \cdot F(X_t\beta) \tag{6.58}$$

Se condicionarmos em X, todos os resultados anteriores continuam a ser válidos, embora a estimação dos parâmetros nesta especificação seja mais complexa.

Capítulo VII- Conclusões

O problema da presença de erros de medida nos dados que o analista utiliza pode ser considerado sob várias vertentes. O estudo que aqui se conclui, tinha como principal objectivo (embora não exclusivo) elaborar uma análise das consequências deste problema nos modelos não lineares, não abordando o problema das técnicas de estimação alternativas. Nos modelos de regressão, a utilização de dados contaminados conduz quase sempre a estimativas inconsistentes para os parâmetros do modelo. Esta inconsistência, traduzia-se no caso do modelo de regressão linear com erro de medida clássico, no efeito atenuação, onde as estimativas dos parâmetros contaminados são enviesadas em direcção à origem. Um dos objectivos deste tudo era dar um enquadramento mais genérico ao que se denominou por efeito atenuação. A interpretação naive deste conceito, originária do que sucede naquela classe de modelos, não encerra todos os efeitos da presença do erro de medida. Mesmo considerando o modelo para o erro de medida clássico existem outras consequências que interessa considerar.

Em termos genéricos, foi possível verificar que a contaminação induz dois efeitos em qualquer modelo de regressão: Na função index, pondo em causa a linearidade dos regressores e na própria densidade da distribuição condicional em causa. Os resultados alcançados pelo recurso à técnica de aproximação às distribuições quando a variância do erro se supõe pequena, permitiu dar uma nova dimensão à interpretação do efeito atenuação. Conclui-se que a gravidade da distorção introduzida pela contaminação dependia da curvatura das densidades e da magnitude do erro de medida. Apesar do carácter estrutural desta abordagem, ao contrário da formalização pela máxima verosimilhança o conhecimento da distribuição do erro de medida não é necessário. Ao se considerar que a variância do erro é pequena, a curvatura das funções densidade implícitas nas aproximações será semelhante à das verdadeiras densidades desconhecidas.

No decorrer deste estudo, privilegiou-se a análise estrutural. A passagem da análise das consequências para propostas de especificação alternativas susceptíveis de produzir estimativas para os parâmetros foi impossibilitada pelo desconhecimento da distribuição marginal da verdadeira variável[30] (ver secções 4.3,

[30]No caso da análise da máxima verosimilhança é também necessário especificar a distribuição

4.4, 6.2 e 6.3). O conhecimento desta distribuição assume um papel primordial no estudo deste problema. É por esta ser contaminada por erro de medida que todos os outros problemas surgem. O desafio que se coloca é de como usar a informação contida em W para descrever $f_X(X)$. A analogia com o problema das variáveis omitidas está aqui bem patente, X nunca é observado mas existe informação sobre a forma de uma variável proxy, que pode ser utilizada para descrever o seu comportamento. A solução deste problema (que não foi alvo de análise neste estudo) pode passar pela utilização de métodos semiparamétricos (ver Carroll, Ruppert & Stefanski, 1995 e Sepanski, J.H & Carroll, R.J, 1993) ou, recorrendo a uma estimação não paramétrica daquela densidade. Chesher (1990) propõe a utilização da aproximação à densidade marginal (4.8), para partindo de uma amostra contaminada W, estimar-se não parametricamente $\widehat{f}_W(W)$ e resolvendo uma equação diferencial estimar a densidade de interesse $\widehat{f}_X(X)$.

Quanto ao problema da contaminação por erro de medida na variável dependente, a conclusão que se retira é que, sob algumas condições se pode ignorar. Na presença de erro de medida não diferencial e quando a variável dependente é discreta, a contaminação por este "vírus" é benigna no sentido em que não põe em causa a consistência dos estimadores das médias. A análise da máxima verosimilhança não se altera, sendo sempre necessário o conhecimento das distribuições que compõem a modelização paramétrica. Pelo recurso à metodologia de aproximação às distribuições, conclui-se que os resultados obtidos para as distribuições marginais são em quase tudo idênticos. Contudo, à semelhança do que sucedia para a variável explicativa, é necessário ter em conta o enviesamento introduzido em alguns momentos condicionais estatísticamente relevantes. Mais uma vez a curvatura das densidades determina a amplitude dos efeitos.

O problema da má classificação em modelos de variável dependente discreta, embora esteja inserido na problemática dos erros de medida carece de uma abordagem distinta. O carácter discreto da variável dependente e a eventual não verificação da não diferenciabilidade do erro de medida, transformam quase sempre este problema numa questão de deficiente parametrização do modelo base. As abordagens efectuadas para os modelos de regressão binária e para o modelo Poisson, traduzem a necessidade de introdução de parâmetros adicionais, que po-

do modelo para o erro de medida para além do modelo base.

dem ser especificados dependendo ou não de regressores. Desta forma é possível reespecificar as densidades e verosimilhanças dos modelos base tendo em conta o comportamento da amostra e estimar consistentemente todos os parâmetros. Nos modelos de regressão binária este problema pode ser grave mesmo para um erro de medida reduzido. Teoricamente basta uma observação para enviesar gravemente os parâmetros. No modelo Poisson, o enviesamento potencial dos parâmetros depende, no caso considerado, do peso das observações com valor zero e um na amostra e da percentagem de dados mal classificados. Os resultados alcançados são específicos para o tipo de má classificação aqui apresentado. Contudo, a metodologia adoptada fornece um quadro metodológico geral para a abordagem de problemas semelhantes com outras realizações de processos Poisson.

A grande dificuldade na análise deste problema é que os efeitos da presença de erro de medida nos modelos de regressão não lineares, são específicos para cada especificação paramétrica. O esforço desenvolvido nesta dissertação foi o de encontrar algumas características comuns a este problema reunindo os contributos feitos neste vasto campo de investigação.

Bibliografia

Berkson, J. (1950). Are there two Regressions? *Journal of the American Statistical Association,* 45, pp. 164-180.

Berndt, E. R., Hall, B. H., Hall, R. E. & Hausman, J. A. (1974). Estimation and Inference in Nonlinear Structural Models. *Annals of Economic and Social Measurement,* 3, pp. 653-65.

Carroll, R. J., Ruppert, D. & Stefanski, L. A. (1995). *Measurement Error in Nonlinear Models.* London: Chapman & Hall.

Chesher, A. (1983). The Information Matrix Test: Simplified Calculation via a Score Test Interpretation. *Economic Letters,* 13, pp. 45-48.

Chesher, A. (1990). The Effects of Measurement Error and Measurement Error Sensitive Specification Test. University of Bristol, Discussion Paper No. 90/274.

Chesher, A. (1991). The Effects of Measurement Error. *Biometrika,* 78, p.p 451-462.

Davidian, M. & Gallant, A. R. (1993). The Nonlinear Mixed Effects Model with a Smooth Random Effects Density. *Biometrika,* 80, p.p 475-488.

Davidson, R. & Mackinnon, J. G. (1984b). Convenient Specification Testes for Logit and Probit Models. *Journal of Econometrics,* 25, pp. 241-262.

Davidson, R. & Mackinnon, J. G. (1993). *Estimation and Inference in Econometrics.* Oxford: Oxford University Press.

Fuller, W. A. (1987). *Measurement Error Models.* New York: John Wiley & Sons.

Godfrey, L. G. & Wickens, M. R. (1981). Testing Linear and Log-linear Regressions for Functional Form. *Review of Economic Studies,* 48, pp.487-96.

Goldeberger, A. S. (1989). The ET Interview. *Econometric Theory,* 5, pp. 133-160.

Griliches, Z. (1985). Data and Econometricians- the Uneasy Alliance. *American Economic Review,* 74, pp. 196-200.

Hansen, L. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica,* 50, pp. 1029-1054.

Hausman, J. A. & Morton, F. M., (1994). Misclassification of a Dependent Variable in a Discrete Response Setting. Massachusetts Institute of Technology, Working Paper.

Hwang, J. T. & Stefanski, L. A. (1994). Monotonicity of Regression Functions in Structural Measurement Error Models. *Statistics & Probability Letters,* 20, pp. 113-116.

Kennedy, P. (1993). A Guide to Econometrics. 3rd ed. Oxford: Blackwell Publishers.

Klepper, S. (1988). Bounding the Effects of Measurement Error in Regressions Involving Dichotomuos Variables. *Journal of Econometrics*, 37, pp. 343-359.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics,* 34, pp. 1-14.

Lancaster, T. (1984). The Covariance Matrix of the Information Matrix Test. *Econometrica,* 52, pp. 1051-1053.

Malinvaud, E. (1966). *Statistical Methods of Econometrics.* Amsterdam: North Holland.

Morgenstern, O. (1963). *On the Accuracy of Economic Observations.* Princeton, NJ: Princeton University Press.

Newey, W. & West, K. (1987b). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review,* 28, pp.777-787.

Reiersol, O. (1941). Confluence Analysis by Means of Lag Moments and other Methods of Confluence Analysis. *Econometrica,* 9, pp. 1-23.

Santos Silva, J. M. (1993). A Note on the Score Test for Neglected Heterogeneity in the Truncated Normal Regression Model. *Economic Letters*, 43, 11-14.

Santos Silva, J. M. & Andrade e Silva, J. M. (1994). Misspecification in Models for Positive Count Data, ISEG/UTL, Comunicação apresentada no ESEM'94.

Sargan, J. D. (1958). The Estimation of Economic Relationships using Instrumental Variables. *Econometrica,* 26, pp. 393-415.

Sepanski, J. H. & Carroll, R. J. (1993). Semiparametric Quasilikelihood and Variance Function Estimation in Measurement Error Models. *Journal of Econometrics*, 58, pp. 226-253.

Weinberg, C. R., Umbach, D. M. & Greenland, S. (1993). When will Non-differential Misclassification Preserve the Direction of a Trend? Preprint.