Estudi de l'estat de Salut autopercebut: Modelització de l'índex d'utilitat EQ-5D mitjançant un model tobit

TREBALL DE RECERCA

Programa de doctorat en Estadística, Anàlisi de dades i bioestadística

Gemma Vilagut Saiz

Tutors: Dr. Josep Fortiana
Dr. Jordi Alonso

Data: 25 de Maig, 2008

Taula de continguts

Índex de taules i figures

Estudi de l'estat de Salut autopercebut: Modelització de l'índex d'utilitat EQ-5D mitjançant un model tobit

Resum

Objectiu: Les mesures sobre l'estat de salut tendeixen a ser asimètriques i tenir un elevat percentatge d'individus amb la puntuació més alta possible (*efecte sostre*). S'han proposat diferents mètodes de modelització d'aquest tipus de variables tenint en compte l'efecte sostre, com els models tobit, els models *Censored Least Absolute Deviations* (CLAD) o els models en dues parts. L'objectiu d'aquest treball es descriure el model tobit, i comparar-lo amb el model lineal amb mínims quadrats ordinaris, que ignora l'efecte sostre.

Mètodes: S'ha utilitzat dos conjunts de dades diferents per comparar els dos models: a) dades reals procedents de l'estudi poblacional sobre trastorns mentals ESEMeD, per modelitzar una de les mesures sobre l'estat de salut més utilitzades, l'índex de salut EQ5D; i; b) dades simulades. Per a la comparació de les prediccions dels dos models en l'anàlisi de les dades de l'estudi ESEMeD es va utilitzar una validació creuada i es van comparar diferents estimadors: el percentatge d'error absolut (R^1), el percentatge d'error al quadrat (R^2), l'error quadràtic mig (MSE), i l'error absolut de predicció mitjà (MAPE). Pel que fa a les dades simulades es van generar diversos conjunts de dades amb diferents proporcions d'individus amb la puntuació més alta possible i diferents valors de la variància dels errors. Es van comparar les estimacions dels coeficients, els percentatge de variabilitat explicada i els gràfics de residus respecte als valors previstos obtinguts en els models lineals i els models tobit estimats en les diferents situacions.

Resultats: Pel que fa a l'estudi ESEMeD, les prediccions de l'índex de salut EQ5D obtingudes amb el model lineal i les del model tobit van ser molt similars. Les estimacions dels coeficients de regressió en el model lineal van ser consistentment més petites que les del model tobit per a tots els predictors. En l'estudi de simulació s'observa que per una variància de l'error petita (σ=1), el model tobit va presentar estimacions no esbiaixades dels coeficients i prediccions acurades, especialment quan el percentatge d'individus amb la puntuació més alta possible era més petit. En canvi, per valors de la variància de l'error més grans (σ=10 o σ=20) el percentatge de variabilitat explicada pel model tobit i les prediccions obtingudes eren més semblants a les obtingudes amb el model de regressió lineal.

Conclusions: El percentatge de variabilitat estimada pels models, així com el percentatge d'individus amb la puntuació màxima sembla tenir un efecte important en el funcionament del model tobit en comparació amb el del model lineal.

Summary

Objective: Health status measures usually have an asymmetric distribution and present a high percentage of respondents with the best possible score (ceiling effect), specially when they are assessed in the overall population. Different methods to model this type of variables have been proposed that take into account the ceiling effect: the tobit models, the Censored Least Absolute Deviations (CLAD) models or the two-part models, among others. The objective of this work was to describe the tobit model, and compare it with the Ordinary Least Squares (OLS) model, that ignores the ceiling effect.

Methods: Two different data sets have been used in order to compare both models: a) real data comming from the European Study of Mental Disorders (ESEMeD), in order to model the EQ5D index, one of the measures of utilities most commonly used for the evaluation of health status; and b) data obtained from simulation. Cross-validation was used to compare the predicted values of the tobit model and the OLS models. The following estimators were compared: the percentage of absolute error (R^1), the percentage of squared error (R^2), the Mean Squared Error (MSE) and the Mean Absolute Prediction Error (MAPE). Different datasets were created for different values of the error variance and different percentages of individuals with ceiling effect. The estimations of the coefficients, the percentage of explained variance and the plots of residuals versus predicted values obtained under each model were compared.

Results: With regard to the results of the ESEMeD study, the predicted values obtained with the OLS model and those obtained with the tobit models were very similar. The regression coefficients of the linear model were consistently smaller than those from the tobit model. In the simulation study, we observed that when the error variance was small (σ=1), the tobit model presented unbiased estimations of the coefficients and accurate predicted values, specially when the percentage of individuals wiht the highest possible score was small. However, when the errror variance was greater (σ=10 or σ=20), the percentage of explained variance for the tobit model and the predicted values were more similar to those obtained with an OLS model.

Conclusions: The proportion of variability accounted for the models and the percentage of individuals with the highest possible score have an important effect in the performance of the tobit model in comparison with the linear model.

1. Antecedents i objectius

Les enquestes de salut poblacionals solen incloure mesures sobre l'estat de salut autopercebuda. Habitualment aquestes mesures s'obtenen a partir d'un qüestionari amb un nombre relativament petit d'ítems de tipus Likert (respostes categòriques ordinals) a partir dels quals s'obté una puntuació que resumeix l'estat de salut d'aquell individu.

Sigui quina sigui la metodologia utilitzada per obtenir la puntuació final sobre l'estat de salut autodeclarat, aquesta mesura es caracteritza per:

a) tenir una alta proporció d'individus amb la puntuació més alta possible (efecte sostre);

b) ser típicament asimètrica, amb una proporció relativament petita d'individus amb les puntuacions més baixes indicadores de mala salut, especialment quan s'avalua en població general;

c) tenir el seu rang restringit a un subconjunt de la recta real.

Una pràctica habitual en l'estudi de l'estat de salut autodeclarat és l'estimació de models de regressió amb l'objectiu de predir l'estat de salut o d'examinar la seva relació amb diferents determinants de la salut com l'edat, el sexe o el nivell socioeconòmic, per tal de quantificar com canvia el valor mig de l'estat de salut per a diferents valors en els determinants.

Els mètodes de regressió habitualment utilitzats per estudiar aquestes variables són mètodes de regressió lineal múltiple mitjançant mínims quadrats ordinaris. Però aquests mètodes no tenen en compte l'elevat efecte sostre ni l'asimetria que normalment presenta aquesta mesura i, com a conseqüència, poden proporcionar resultats esbiaixats. S'han proposat metodologies alternatives per analitzar aquest tipus de variables: com els models per a dades censurades Tobit (Tobin 1958;Greene 2003a), els models *Censored Least Absolute Deviations* (CLAD) (Powell 1984), o els models en dues parts (Mullahy 1998;Manning, Duan and Rogers 1987). En aquest treball es descriuen el primer d'aquests tres tipus de models (model Tobit) i es realitza un exercici pràctic de comparació del model tobit amb el model lineal en dos conjunts de dades diferents:

1.- Estimació dels dos models a partir d'unes dades reals, amb l'objectiu de comparar els resultats de la regressió lineal amb els del model de regressió tobit per quantificar l'associació dels trastorns mentals i malalties físiques amb una de les mesures d'utilitat més utilitzades per avaluar l'estat de salut autopercebut, l'índex EuroQol 5D (EQ-5D) (The EuroQol Group 1990).

2.- Utilització d'unes dades simulades per comparar el comportament del model tobit amb el del model lineal amb mínims quadrats ordinaris per diferents proporcions d'individus amb la puntuació més alta possible, i diferents valors de la variància dels errors (que fa modificar el percentatge de variabilitat explicada pels models). Aquest estudi de simulació va ser motivat pels resultats que obteníem en l'exercici amb dades reals que ens estranyaven ja que en la comparació entre els valors observats i els predits s'obtenien resultats molt similars entre els dos tipus de models, i fins i tot lleugerament millors pel que fa al model lineal.

2. Mètodes

2.1. Descripció de les dades reals utilitzades en l'exercici pràctic

Les dades utilitzades en l'exercici pràctic amb dades reals exposat en aquest treball provenen del *European Study of the Epidemiology of Mental Disorders (ESEMeD)*, una enquesta transversal que es va dur a terme en mostres representatives d'individus adults no institucionalitzats de 6 països europeus (Bèlgica, França, Alemanya, Itàlia, Holanda i Espanya) i que tenia com a principal objectiu determinar les prevalences dels trastorns mentals més comuns, és a dir, el percentatge d'individus d'aquests països que han patit els trastorns en algun moment de la seva vida (prevalença vida) i el percentatge d'individus que han patit algun d'aquests trastorns durant l'any previ a l'entrevista (prevalença 12 mesos). Els individus participants van ser entrevistats en persona a les seves pròpies llars per entrevistadors no especialitzats en psiquiatria i especialment entrenats per administrar el qüestionari.

Es va fer servir un mostreig probabilístic estratificat, multi-etàpic en tots els països. El mostreig estratificat implica la divisió de la població en grups relativament homogenis, anomenats estrats, i la selecció de mostres independents en cadascun d'aquests estrats. El mostreig multi-etàpic es refereix a seleccionar una mostra en dues o mes etapes successives, per exemple, en un mostreig en dues etapes, la primera etapa de mostreig podria consistir en seleccionar aleatòriament municipis dins de cadascun dels estrats i, en la segona etapa es seleccionarien de forma aleatòria individus dins dels municipis seleccionats en la primera etapa(Cochran 1977). En el cas de l'estudi ESEMeD els estrats van ser definits per les regions o les regions i la grandària del municipi, depenent del país. El nombre d'etapes de mostreig també va variar depenent del país, de 1 etapa (a Holanda) a 4 etapes de mostreig (Espanya). En total, es van entrevistar un total de 21.425 individus i la taxa de resposta global per als sis països (mitjana ponderada per la població global del país) va ser del 61,2%, variant de 78,6% a Espanya a 45,9% a França(Alonso, Angermeyer, *i cols* 2004).

L'instrument diagnòstic utilitzat per determinar la presència de trastorns mentals va ser una nova versió del Composite International Diagnostic Interview (Wittchen 1994), anomenat CIDI 3.0, que va ser desenvolupat i adaptat pel comitè coordinador del WHO World Mental Health Survey Initiative (Kessler and Ustun 2004).

Per tal d'optimitzar el procés i el cost de l'entrevista, es va dur a terme un procediment en dues fases. La primera fase es va administrar a tots els entrevistats i consistia en l'avaluació diagnòstica dels trastorns de l'estat d'ànim i d'ansietat més comuns, avaluació de la qualitat de vida relacionada amb la salut, informació sobre d'utilització de serveis sanitaris per als problemes emocionals i tractament, i característiques sociodemogràfiques més rellevants. Aquells individus que superaven un nombre determinat de símptomes d'algun dels trastorns d'ansietat o de l'estat d'ànim avaluats ("individus d'alt risc") i un 25% aleatori de la resta ("individus de baix risc") van continuar amb la segona fase de l'entrevista, que consistia amb una entrevista en profunditat per recollir informació sobre altres trastorns mentals addicionals, com trastorn per estrès posttraumàtic o trastorns relacionats amb el menjar, així com informació sobre malalties físiques autoreportades, sobre discapacitat i sobre factors de risc, entre d'altres. El nombre total d'individus dels 6 països als quals se'ls va administrar la segona part del qüestionari van ser n=8.796.

Variable dependent dels models - Mesura de l'estat de Salut:

L'estat de salut es va avaluar mitjançant el qüestionari EQ-5D (EuroQol 5D) (veure *apèndix 1*), que va ser administrat a tots els entrevistats al final de l'entrevista. El EQ-5D consisteix en 5 ítems diferents: mobilitat, cures personals, activitats quotidianes, dolor/malestar, i ansietat/depressió, que es coneixen també com a dimensions del EQ-5D. Cadascuna d'aquestes té 3 categories: (1) Cap problema, (2) Alguns problemes; (3) problemes extrems. Els entrevistats havien de seleccionar la categoria que millor descrivia el seu estat de salut actual per a cadascun dels ítems. A partir de les respostes d'un individu es construeix el seu corresponent estat de salut, seleccionant el nivell escollit de cadascuna de les dimensions. Així, per exemple, l'estat de salut 21111 defineix un individu amb alguns problemes de mobilitat, i cap tipus de problema en cap de les altres quatre dimensions de l'instrument. Els estats de salut del EQ-5D es poden convertir en un únic índex resum aplicant una fórmula que assigna un pes a cadascuna de les categories de cada dimensió. Aquesta quantificació va ser obtinguda en base a la valoració dels estats de salut del EQ-5D en mostres poblacionals. S'han obtingut quantificacions per a l'EQ-5D a partir de mostres poblacionals de diferents països (Bèlgica, Finlandia, Nova Zelanda, Eslovènia, Dinamarca, Alemanya, Japó, Espanya, Anglaterra, entre d'altres) i s'han aplicant diferents metodologies per obtenir les valoracions dels diferents estats de salut per part dels individus que conformaven les mostres. Froberg and Kane (Froberg and Kane 1989a;Froberg

and Kane 1989b) descriuen alguns dels mètodes més habituals per obtenir les valoracions dels estats de salut. En aquest estudi, per tal d'obtenir l'índex EQ-5D s'han aplicat les quantificacions d'Anglaterra, que van ser obtingudes a partir del mètode *Time trade-off (TTO)* (Dolan 1997). L'índex EQ-5D té un rang teòric de -0.59 a 1, on 0 és equivalent a la mort, 1 és equivalent a un estat de salut perfecte i els valors negatius representen estats de salut subjectivament apreciats com pitjors que la mort.

Part determinista dels models - Avaluació dels trastorns mentals, malalties físiques i informació sòcio-demogràfica:

Els trastorns mentals que s'han avaluat per aquest treball inclouen: trastorns de l'estat d'ànim (Episodi de Depressió Major, distímia), trastorns d'ansietat (fòbia específica, fòbia social, agorafòbia sense pànic, trastorn de pànic, trastorn d'ansietat generalitzada i trastorn per estrès posttraumàtic) i trastorns per consum d'alcohol (abús d'alcohol i dependència d'alcohol). Aquests trastorns es van definir a partir dels criteris del Manual Diagnòstic i Estadístic dels Trastorns Mentals, quarta Edició (DSM-IV). Aquests criteris són uns dels més utilitzats per a la definició de trastorns mentals, juntament amb els de la desena edició de la Classificació Internacional de Malalties (ICD-10).

Per tal d'obtenir informació sobre malalties físiques, es va administrar una llista de comprovació (*check-list*) de malalties físiques autodeclarades i malalties doloroses. Les malalties doloroses incloses a la llista eren: artrosi o reumatisme, problemes cervicals o d'esquena, mals de cap freqüents o molt forts, o altres tipus de dolor crònic. Pel que fa a la resta de malalties cròniques avaluades, eren les següents: al·lèrgies estacionals, embòlia cerebral, atac de cor, hipertensió,asma, tuberculosi,una altra malaltia pulmonar crònica, malaltia parasitària, diabetis, úlcera intestinal o estomacal, malaltia de la tiroides, problemes neurològics (esclerosi múltiple, pàrquinson, o atacs epilèptics), SIDA o infecció per VIH i càncer.

La presència dels trastorns mentals i de les malalties físiques es va avaluar pel que fa als 12 mesos previs a l'entrevista i en algun moment de la vida. En l'estudi pràctic s'estudiaran els trastorns mentals i malalties físiques cròniques en els 12 mesos previs.

La informació sòcio-demogràfica recollida i utilitzada inclou el sexe, l'edat, l'estat civil, l'educació, la grandària del municipi, la situació laboral, i el país.

2.2. Mètodes estadístics per a la modelització de dades amb elevat efecte sostre

En els estudis poblacionals habitualment hi ha un nombre important d'individus que obtenen la puntuació màxima possible pel que fa a la salut general autopercebuda. Aquests resultats es poden interpretar de dues maneres: o bé la mesura és acurada i realment una proporció elevada de la població té *salut perfecta*, o bé l'índex té un efecte sostre i no és capaç de discriminar entre nivells elevats d'estats de salut. En aquesta segona situació, la part superior de la distribució poblacional de l'estat de salut es col·lapsa a un valor sostre, que es correspon amb el valor màxim que pot prendre l'escala.

Quan hi ha efecte sostre, els models de regressió estàndard l'ignoren. Un procediment alternatiu és tractar l'índex com si estigués censurat, essent la puntuació de 1 la puntuació màxima observable. En aquest cas, per als individus amb una puntuació de 1 tot el que sabem és que l'estat de salut real d'aquests individus és com a mínim 1. El **model de regressió tobit** (Tobin 1958) i el model ***Censored Least Absolute Deviations* (CLAD)** (Powell 1984) tenen en compte de forma inherent la presència de censures.

Altres mètodes que s'han descrit adequats en aquests casos són els **models en dues parts** (Mullahy 1998; Manning et al. 1987), que permeten modelitzar separadament la probabilitat de tenir la puntuació màxima 1, i el valor esperat de la variable condicionat a que aquest valor és inferior a 1.

El model tobit

El model tobit (Tobin 1958;Greene 2003b;Bleda Hernández and Tobías Garcés 2002) és un model de regressió per a dades censurades.

La presentació usual dels models tobit és per a dades censurades inferiorment, però el punt de censura de la variable no té perquè estar necessàriament en el límit inferior. Podem plantejar també un model tobit amb d'altres esquemes de censura, com és el cas del problema pràctic que tractem en aquest treball on les dades tenen censura superior.

En aquest cas, suposem una variable observada y amb un únic **punt de censura superior** a. Postulem una variable latent subjacent y^* de forma que:

$y = y^*$, quan $y^* < a$,

$y = a_y$, quan $y^* \geq a$,

on a és el punt de censura, i a_y és el valor assignat a la variable y si y^* és més gran que el punt de censura. En general $a_y = a$, encara que no sempre ha de ser així necessàriament. D'aquí en endavant considerarem que $a_y = a$.

En el model de regressió Tobit s'assumeix que la variable subjacent segueix una distribució normal $y^* \sim N(\mu,\sigma^2)$ i per tant la probabilitat que una observació estigui censurada és:

$$Prob(y=a)=Prob(y^*\geq a)=Prob(N(\mu,\sigma^2)\geq a)=Prob\left(N(0,1)\geq\left(\frac{a-\mu}{\sigma}\right)\right)=1-\Phi\left(\frac{a-\mu}{\sigma}\right),$$

on $\boldsymbol{\Phi(\cdot)}$ representa la funció de distribució d'una N(0,1).

La probabilitat de no censura és:

$$Prob(y<a)=Prob(y^*\leq a)=Prob(N(\mu,\sigma^2)\leq a)=Prob\left(N(0,1)\leq\left(\frac{a-\mu}{\sigma}\right)\right)=\Phi\left(\frac{a-\mu}{\sigma}\right).$$

Així doncs, la funció de distribució de la variable observada y és:

$F(t)=\mathbb{1}_{\{t<a\}}\cdot F^*(t) + \mathbb{1}_{\{t\geq a\}}$, on F^* és la funció de distribució de la variable latent.

La distribució de la variable observada és una mixtura entre una part contínua i una part discreta; s'assigna al punt de censura a tota la probabilitat que correspon a l'àrea censurada.

La formulació general del model de regressió tobit per a una variable observada amb **censura superior** (i punt de censura a) és:

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i \ ,$$

$y = y^*$, quan $y^* < a$,

$y = a$, quan $y^* \geq a$.

És a dir, el valor mig de la variable latent és una funció lineal de les variables explicatives. L'esperança de la variable latent y* és:

$$E[y_i^* | \boldsymbol{x_i}] = \boldsymbol{x_i'\beta}$$

(Greene 2003b) demostra que el valor esperat de la variable observada és:

$$E[y_i | \boldsymbol{x_i}] = \left(1 - \Phi\left(\frac{a - \boldsymbol{x_i\beta}}{\sigma}\right)\right) \cdot a + \Phi\left(\frac{a - \boldsymbol{x_i\beta}}{\sigma}\right) \cdot \left(\boldsymbol{x_i\beta} + \sigma\lambda_i\right) \ ,$$

On

$$\lambda_i = -\frac{\phi\left(\frac{a - \boldsymbol{x_i\beta}}{\sigma}\right)}{\Phi\left(\frac{a - \boldsymbol{x_i\beta}}{\sigma}\right)} \ .$$

L'estimació de la contribució unitària de cada predictor *j* sobre el valor mig de la variable latent *y** ve donada per l'estimació del coeficient corresponent $\hat{\beta}_j$:

$$\frac{\partial E[y_i^* | \boldsymbol{x_i}]}{\partial x_{ij}} = \hat{\beta}_j \ .$$

En canvi, si el que volem és estimar la contribució unitària de cadascun dels predictors x_j sobre el valor mig de la variable observada *y*, conegut com *efecte marginal*, (Greene 2003a) mostra que, per obtenir-la cal ponderar el coeficient corresponent per la probabilitat de no censura:

$$\frac{\partial E[y_i | \boldsymbol{x_i}]}{\partial x_{ij}} = \hat{\beta}_j \cdot Prob[y_i^* < a] = \hat{\beta}_j \cdot \Phi\left(\frac{a - \boldsymbol{x_i'\beta}}{\sigma}\right) \ .$$

La probabilitat de no censura depèn dels valors que pren l'individu *i* en cadascuna de les variables *x*. El que se sol fer per trobar els efectes marginals és ponderar per la mitjana de la probabilitat de no censura dels individus de la mostra (Greene 1999).

Assumpcions del model tobit:

El model tobit assumeix que la distribució de la variable resposta, condicionada als valors de les variables explicatives és **Normal**, i la **variància és uniforme (homoscedasticitat)**.

(Austin, Escobar and Kopec 2000) van dur a terme un estudi de simulació de Montecarlo, a partir del qual va determinar que quan la distribució de la mesura de l'estat de salut condicionada als predictors té cues més llargues o és més asimètrica que la distribució normal, el model tobit encara proporciona resultats no esbiaixats, al contrari que el model lineal amb mínims quadrats ordinaris. Altres autors són més pessimistes respecte a l'efecte de la no normalitat en les estimacions del model tobit (Tobin 1958;Greene 2003a;Arabmazar and Schmidt 1982). Una alternativa és utilitzar un mètode d'estimació que sigui robust a canvis en la distribució, com el mètode *censored least absolute deviations* (CLAD) proposat per Powell (Powell 1984). El principal inconvenient de la implantació d'aquest mètode és la seva complexitat computacional.

Pel que fa a l'heteroscedasticitat, Austin (Austin et al. 2000) va mostrar que quan la variància de la distribució condicional no és uniforme, les estimacions obtingudes amb el model tobit poden arribar a ser fins i tot pitjors que les obtingudes amb el model lineal. En aquest mateix sentit, Maddala també va demostrar que en presència d'heteroscedasticitat les estimacions del model tobit són esbiaixades i va proposar una extensió del model tobit que tingués en compte l'heteroscedasticitat, sempre i quan fos possible determinar la forma funcional de la variància (per exemple, si aquesta augmenta linealment en funció d'una determinada covariable).

Estimació:

L'estimació es realitza mitjançant el mètode de màxima versemblança. La funció de versemblança que es maximitza conté informació de les observacions censurades i de les no censurades. La funció del logaritme de la versemblança per al model tobit és:

$$\ln L = -\frac{1}{2}\sum_{y_i<a}\left[\ln(2\pi)+\ln\sigma^2+\frac{(y_i-\boldsymbol{x}_i'\boldsymbol{\beta})^2}{\sigma^2}\right]+\sum_{y_i=a}\ln\left[1-\Phi\left(\frac{a-\boldsymbol{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right] .$$

La primera part del sumatori es correspon amb el model de regressió clàssica per a les observacions no censurades i la segona part mostra les probabilitats de les observacions censurades. Aquesta versemblança és una mixtura de distribucions continua i discreta.

(Olsen 1978) va proposar una reparametrització d'aquesta funció de versemblança defininint $\gamma = \beta/\sigma$ i $\theta = 1/\sigma$,

$$\ln L = -\frac{1}{2}\sum_{y_i<a}\left[\ln(2\pi) - \ln\theta^2 + (\theta y_i - \mathbf{x}_i'\boldsymbol{\gamma})^2\right] + \sum_{y_i=a}\ln\left[1-\Phi(\theta a - \mathbf{x}_i'\boldsymbol{\gamma})\right],$$

i va provar que la matriu Hessiana d'aquesta funció de versemblança és definida negativa quan la mostra és suficientment gran. Com a conseqüència, la funció de versemblança té un únic màxim global sempre i quant el procés iteratiu acabi obtenint una solució. Per aquesta raó, l'algoritme d'optimització utilitzat per la majoria de paquets estadístics, com el SAS versió 9.01 (Copyright © 2002-2003 SAS Institute Inc., Cary, NC, USA) o el programa Stata (StataCorp 2003) , per a l'estimació mitjançant màxima versemblança, és el de Newton-Raphson, que utilitza la inversa de la Hessiana. La matriu de covariàncies asimptòtica obtinguda per defecte es calcula a partir de la inversa de la matriu d'informació observada.

El model tobit està implementat en molts dels paquets estadístics més utilitzats. A l'apèndix 2 es dóna informació sobre les instruccions per a l'estimació en tres dels paquets estadístics més utilitzats: SAS, R(R Development Core Team 2006) i Stata.

2.3. Mètodes de l'exercici pràctic.

L'exercici amb les dades de l'estudi ESEMeD.

Un dels objectius secundaris de l'estudi ESEMeD és quantificar l'associació entre la presència de trastorns mentals i malalties físiques cròniques i l'estat de salut autopercebut mesurat amb l'índex EQ5D.

Com ja s'ha indicat prèviament, l'estimació d'aquestes associacions a partir d'un model de regressió lineal múltiple pot proporcionar estimacions esbiaixades dels coeficients degut a la distribució asimètrica i amb un elevat percentatge d'individus amb la puntuació més alta possible.

En aquest treball es van comparar les estimacions dels coeficients obtingudes amb el model tobit amb les del model lineal per predir l'estat de salut autopercebut avaluat amb l'índex EQ5D a

partir de la presència de diferents trastorns d'ansietat i malalties físiques cròniques i de determinades variables sociodemogràfiques dels individus.

Es va utilitzar una validació creuada per comparar les prediccions dels dos models dividint la mostra en dues parts de forma aleatòria: un 90% per a la mostra d'entrenament a partir de la qual s'obtenien els coeficients dels models estimats, i un 10% per a la mostra de validació. Es van calcular els coeficients R^1 i R^2, l'error quadràtic mitjà (*Mean Squared Error*, MSE) i l'error absolut de predicció mitjà (*Mean Absolute Prediction Error*, MAPE) per als dos models estimats:

$$R^1 = 1 - \frac{\sum |y_i - \hat{y}_i|}{\sum |y_i - \bar{y}|} \qquad MAPE = \frac{\sum |y_i - \hat{y}_i|}{n}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \qquad MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n}$$

Els valors predits del model tobit, en aquest cas són els que corresponen a la variable censurada $E(y_i \mid x_i)$, segons la fórmula que es presenta a la pàgina 9.

Aquest procés de selecció de les mostres de validació i d'entrenament i càlcul dels estimadors es va repetir 100 vegades (mètode *Random subsampling* de validació creuada). Es van calcular les mitjanes d'aquests estimadors en les 100 mostres de validació i les desviacions estàndard de les 100 estimacions.

Es van obtenir els gràfics dels residus respecte als valors previstos dels dos models i els gràfics de normalitat dels residus amb l'objectiu de testar les hipòtesis de normalitat dels residus i d'homoscedasticitat dels dos models.

Les anàlisis per aquest treball van ser realitzades amb el paquest estadístic SAS versió 9.01 del sistema SAS per Windows™ (Copyright © 2002-2003 SAS Institute Inc., Cary, NC, USA). La macro en SAS per a la validació creuada es pot consultar en l'*apèndix 3*.

Descripció de l'estudi de simulació

L'objectiu de l'estudi de simulació era comparar els resultats del model lineal amb mínims quadrats ordinaris amb els d'un model tobit per diferents proporcions de dades censurades i 3

valors diferents de la variància dels errors (σ^2). Es va generar dades en les que es fixava la part determinista dels models, establint una relació lineal entre la variable dependent i l'únic predictor inclòs. Així doncs, en les dades generades, es van fixar els dos coeficients dels models a dos valors arbitraris β_0=3, i β_1 =5,5. El predictor lineal x_i es va generar com una seqüència regular de 25 valors entre -4,0 i 4,0, en la que cadascun dels 25 valors es repetia 5 vegades (grandària de la mostra N=125). La variable latent no censurada (*y**) es va obtenir sumant un terme d'error aleatori $\varepsilon_i \sim N(0, \sigma^2)$, a la part determinista del model ($\beta_0 + \beta_1 x_i$).
Posteriorment, per a cada individu es va crear la variable censurada. En aquest cas, només es va considerar el cas de censura superior, en concordança amb les dades reals que hem fet servir, que únicament presenten censura superior. A la variable censurada se li va assignar el valor de la variable latent generada aleatòriament y_i* en cas que aquest fos inferior al valor del punt de censura *a*, i se li va assignar el valor de *a* en cas que y_i* fos superior o igual a *a*.

Es van generar 3 bases de dades diferents, per 3 valors de σ (σ=1, σ=10 i σ=20), per representar diferents nivells de variació al voltant de la recta de regressió. Per cadascun dels valors de σ, es van crear 2 variables censurades *y* diferents, modificant el punt de censura (*a* = -5 i *a* = 10). Escollint diferents punts de censura es modifica el percentatge d'individus amb la puntuació màxima possible (efecte sostre): amb *a*= -5 el percentatge de dades censurades era aproximadament un 65% (valor similar al percentatge de censura observat en la mostra d'ESEMeD), mentre que amb *a*=10 el percentatge d'individus amb dades censurades era aproximadament del 35%.

Per cadascuna de les bases de dades generades, es van estimar 5 models de regressió: a) el model de regressió lineal per la variable dependent *y** (variable latent), b) el model de regressió lineal per la variable censurada amb punt de censura *a=-5*, c) el model de regressió lineal per la variable censurada amb punt de censura *a=10*; d) el model de regressió tobit per la variable censurada amb punt de censura *a=-5*, i e) el model de regressió tobit per la variable censurada amb punt de censura *a=10*. Es van comparar les estimacions dels coeficients entre els diferents models, el percentatge de variabilitat explicada per cadascun dels models i els gràfics de residus respecte als valors previstos per cadascun dels models.

La simulació es va dur a terme amb el paquet estadístic R, versió 4.2.0 (R Development Core Team 2006). El programa en R corresponent a la simulació es pot consultar en l'*apèndix 4*.

3. Resultats

3.1. Resultats de l'exercici amb les dades de l'estudi ESEMeD

En primer lloc es presenta una descripció de les dades, tant pel que fa a les característiques sociodemogràfiques dels individus, com a la proporció d'individus amb els diferents tipus de trastorns mentals o malalties cròniques estudiades, i respecte a la mesura utilitzada per avaluar l'estat de salut autopercebut, l'índex de salut EQ5D.

Les característiques de la mostra es descriuen a la taula 1.

Taula 1. Característiques de la mostra de l'estudi ESEMeD (n=8.779)

		n*	% (EE)
Edat:	18-24	664	11,43 (0,60)
	25-34	1599	18,35 (0,64)
	35-49	2669	27,79 (0,73)
	50-64	2197	21,76 (0,67)
	>64	1667	20,67 (0,71)
Sexe:	Homes	3689	48,22 (0,83)
Anys d'educació: ≤12		5515	65,36 (0,78)
Estat civil:	Casat / convivint	5788	66,76 (0,81)
	Separat/divorciat/vidu	1327	11,14 (0,52)
	Mai casat	1681	22,09 (0,73)
Grandària de municipi: <10000 habitants		2525	33,17 (0,85)
	10000-100000 hab.	3840	38,72 (0,83)
	>100000 hab.	2431	28,11 (0,77)
Situació laboral:	Treball remunerat	4863	56,51 (0,83)
	No treball remunerat	3933	43,49 (0,83)
Nivell d'ingressos:	Baix	1590	18,97 (0,67)
	Mitjà-baix	2709	32,06 (0,80)
	Mitjà-alt	2976	33,23 (0,79)
	Alt	1521	15,74 (0,59)
País:	Bèlgica	1043	3,79 (0,24)
	França	1436	20,54 (0,54)
	Alemanya	1323	31,51 (0,60)
	Itàlia	1779	22,44 (0,70)
	Holanda	1094	6,11 (0,24)
	Espanya	2121	15,60 (0,44)
Depressió		905	4,09 (0,20)
Distímia		155	0,76 (0,09)
Qualsevol trastorn de l'estat d'ànim		991	4,52 (0,21)
Agorafòbia		72	0,42 (0,08)

Taula 1. Característiques de la mostra de l'estudi ESEMeD (n=8.779) *(continuació)*

	n*	% (EE)
Fòbia social	228	1,62 (0,17)
Trastorn d'ansietat generalitzada	118	0,60 (0,09)
Trastorn de Pànic	174	0,80 (0,09)
Trastorn per estrès posttraumàtic	214	1,13 (0,12)
Qualsevol trastorn d'ansietat	1197	8,70 (0,41)
Abús d'alcohol	55	0,50 (0,11)
Dependència d'alcohol	33	0,25 (0,10)
Abús/dependència d'alcohol	88	0,75 (0,15)
Qualsevol trastorn mental	1847	11,93 (0,46)
Qualsevol Malaltia dolorosa	3410	33,44 (0,77)
Qualsevol altra malaltia crònica	2888	30,83 (0,78)
Algun trastorn mental i físic	1232	7,56 (0,35)
No trastorn mental ni malaltia física	3358	45,47 (0,84)

*n ponderada, % no ponderat

La majoria dels individus (67%) estava casat o convivint amb algú en el moment de l'entrevista. Gairebé la meitat de la mostra tenia entre 35 i 64 anys. Un 57% tenia un treball remunerat. El 45% de la mostra no tenien cap trastorn mental ni malaltia física crònica. Un 4,5% van presentar algun trastorn de l'estat d'ànim (depressió o distímia) durant els 12 mesos previs a l'entrevista, i un 8,7% van patir algun dels trastorns d'ansietat estudiats en els 12 mesos previs.

La figura 1 mostra la distribució de l'índex EQ5D a la mostra de l'estudi ESEMeD. L'efecte sostre observat era molt elevat, ja que un 62% d'individus que va obtenir la puntuació més alta possible en aquest índex. Un 1% de la mostra tenia valors negatius.

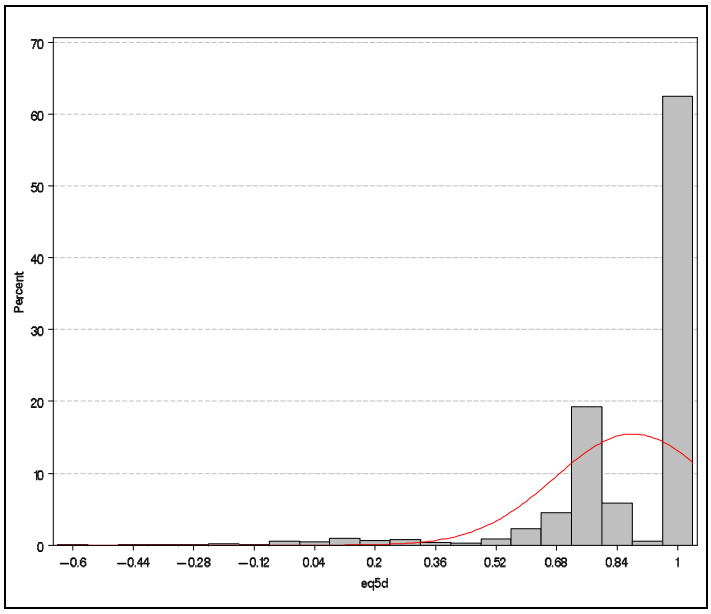

Figura 1. Distribució del índex EQ5D a la mostra de l'estudi ESEMeD

La taula 2 presenta un descriptiu de les puntuacions de l'índex EQ-5D, així com el percentatge d'individus amb algun problema o problemes extrems en cadascuna de les dimensions del EQ-5D, per a tota la mostra i segons determinades malalties físiques o trastorns mentals. La mitjana de l'índex EQ5D en tota la mostra va ser de 0.91, i més de la meitat de la mostra va obtenir la puntuació màxima en aquest índex. Per als individus amb algun tipus de trastorn de l'estat d'ànim, la mitjana de l'índex va ser de 0.77. Els individus que van presentar una puntuació mitjana més baixa van ser els que tenien alguna malaltia neurològica ($\bar{x}(IC95\%) = 0{,}60(\ 0{,}39 - 0{,}81)$). La dimensió de dolor o malestar és la que va presentar una proporció més gran d'individus amb alguns problemes o problemes extrems (25% del total de la mostra). En canvi, a la dimensió de cures personals només un 3.4% de la mostra va reportar algun tipus de problema. El percentatge d'individus amb algun tipus de problemes en alguna de les dimensions era substancialment més alt entre els que patien determinades malalties físiques, en especial embòlia cerebral, malaltia pulmonar o problema neurològic, o trastorns mentals (distímia, agorafòbia o trastorn d'ansietat generalitzada).

Taula 2. Descripció de l'índex EQ5D i % d'individus amb algun problema o problemes extrems en cadascuna de les dimensions del EQ-5D, estudi ESEMeD

	n	Índex EQ-5D		% individus amb problemes*					
		$\bar{X}$ (IC 95%)	Mediana(Q25,Q75)	Mo	CP	AQ	Do	DA	Cap[1]
Tota la Mostra	8779	0,91(0,91 - 0,92)	1,00(0,79 - 1,00)	12,6	3,4	9,7	25,0	7,6	68,7
Trastorns mentals:									
Depressió	903	0,77(0,74 - 0,79)	0,82(0,71 - 1,00)	21,0	9,7	23,1	42,1	41,9	36,8
Distímia	154	0,77(0,72 - 0,81)	0,79(0,69 - 0,87)	25,3	12,0	32,5	50,9	51,6	24,7
Qualsevol trastorn de l'estat d'ànim	*988*	*0,77(0,75 - 0,79)*	*0,82(0,71 - 1,00)*	*21,5*	*10,2*	*24,4*	*43,2*	*41,5*	*36,2*
Agorafòbia	72	0,80(0,75 - 0,84)	0,78(0,72 - 1,00)	20,6	9,3	22,8	63,6	36,1	25,5
Fòbia Específica	676	0,83(0,80 - 0,87)	1,00(0,75 - 1,00)	19,9	7,9	19,3	35,0	20,2	54,3
Fòbia social	227	0,81(0,76 - 0,86)	0,83(0,71 - 1,00)	17,7	7,1	21,6	40,6	32,4	44,1
Trastorn d'ansietat generalitzada	118	0,76(0,69 - 0,83)	0,78(0,69 - 1,00)	15,7	9,3	22,7	53,1	43,1	35,1
Trastorn de Pànic	173	0,74(0,68 - 0,81)	0,81(0,68 - 1,00)	18,2	6,4	28,0	40,1	40,3	38,2
Trastorn per estrès posttraumàtic	213	0,76(0,71 - 0,81)	0,78(0,69 - 1,00)	29,9	11,2	29,7	52,5	33,6	36,8
Qualsevol trastorn d'ansietat	*1192*	*0,83(0,80 - 0,85)*	*1,00(0,73 - 1,00)*	*18,9*	*7,1*	*19,0*	*37,6*	*23,4*	*51,5*
Abús d'alcohol	54	0,87(0,80 - 0,95)	1,00(0,75 - 1,00)	9,3	3,9	12,6	33,7	10,9	65,1
Dependència d'alcohol	33	0,89(0,79 - 1,00)	1,00(0,80 - 1,00)	7,9	7,1	12,3	18,0	17,7	70,1
Abús/dependència d'alcohol	87	0,88(0,82 - 0,94)	1,00(0,76 - 1,00)	8,8	5,0	12,5	28,5	13,2	66,8
Qualsevol trastorn mental	*1840*	*0,83(0,81 - 0,85)*	*1,00(0,73 - 1,00)*	*18,3*	*7,5*	*19,0*	*36,3*	*25,7*	*50,7*

Taula 2. Descripció de l'índex EQ5D i % d'individus amb algun problema o problemes extrems en cadascuna de les dimensions del EQ-5D, estudi ESEMeD

	n	Índex EQ-5D		% individus amb problemes*					
		$\overline{X}$ (IC 95%)	Mediana(Q25,Q75)	Mo	CP	AQ	Do	DA	Cap[1]
Malalties físiques cròniques:									
Artrosis/ reumatisme	1254	0,76(0,74 - 0,78)	0,79(0,69 - 1,00)	37,1	11,3	29,0	62,0	15,3	31,6
Problemes cervicals o d'esquena	2085	0,82(0,80 - 0,83)	0,79(0,73 - 1,00)	23,9	7,4	21,7	52,0	11,8	42,0
Mals de cap freqüents o molt forts	1169	0,85(0,83 - 0,87)	1,00(0,78 - 1,00)	16,4	5,7	15,8	42,0	13,8	53,2
Qualsevol altre dolor crònic	616	0,76(0,73 - 0,79)	0,77(0,69 - 1,00)	37,1	11,1	30,1	64,8	12,4	29,9
Al·lèrgies estacionals	857	0,92(0,90 - 0,93)	1,00(0,79 - 1,00)	9,2	2,2	8,8	26,4	7,0	68,6
Embòlia cerebral	66	0,69(0,55 - 0,82)	0,79(0,57 - 0,86)	58,0	15,9	48,8	50,5	8,1	24,6
Atac de cor	181	0,80(0,75 - 0,85)	0,81(0,73 - 1,00)	42,3	3,9	23,6	46,6	9,7	44,2
Malaltia cardíaca	398	0,76(0,71 - 0,82)	0,80(0,69 - 1,00)	44,4	13,0	37,6	48,0	13,0	38,6
Pressió alta	1147	0,83(0,81 - 0,86)	0,88(0,73 - 1,00)	27,0	6,9	20,8	42,2	10,8	49,8
Asma	330	0,86(0,83 - 0,89)	1,00(0,77 - 1,00)	25,7	4,4	16,1	33,7	9,1	52,6
Altres malalties pulmonars	169	0,75(0,69 - 0,81)	0,79(0,71 - 1,00)	36,4	11,5	27,2	51,5	21,9	29,5
Diabetis	367	0,82(0,79 - 0,85)	0,85(0,73 - 1,00)	32,7	6,1	27,1	43,5	8,2	44,6
Úlcera estomacal o intestinal	181	0,73(0,65 - 0,80)	0,79(0,69 - 1,00)	35,4	10,8	29,2	56,1	15,0	35,5
Malaltia de la tiroides	293	0,87(0,84 - 0,91)	1,00(0,77 - 1,00)	18,0	5,0	15,5	33,1	9,0	61,0
Problema Neurològic	72	0,60(0,39 - 0,81)	0,73(0,26 - 1,00)	61,0	31,9	50,8	56,6	13,3	28,6
Càncer	91	0,79(0,70 - 0,89)	0,79(0,69 - 1,00)	30,7	15,1	31,1	48,0	6,5	46,6
Qualsevol Malaltia dolorosa	*3403*	*0,83(0,82 - 0,84)*	*0,81(0,73 - 1,00)*	*23,0*	*6,8*	*19,6*	*48,6*	*11,4*	*45,5*
Qualsevol altra malaltia crònica	*2883*	*0,86(0,85 - 0,87)*	*1,00(0,78 - 1,00)*	*22,6*	*6,0*	*17,7*	*36,7*	*9,5*	*55,2*
Algun trastorn crònic	*4813*	*0,86(0,85 - 0,87)*	*1,00(0,78 - 1,00)*	*20,0*	*5,5*	*16,3*	*38,7*	*9,8*	*54,2*
Algun trastorn mental i físic	1227	0,78(0,75 - 0,81)	0,79(0,69 - 1,00)	24,8	10,5	25,4	49,1	29,2	40,2
No trastorn mental ni malaltia física	3353	0,96(0,96 - 0,97)	1,00(1,00 - 1,00)	5,0	1,2	2,7	11,0	4,0	84,6

*Dimensions del EQ5D: **Mo**: mobilitat, **CP**: cures personals, **AQ**: activitats quotidianes, **Do**: dolor/malestar, **AD**: ansietat/depressió

[1]**Cap=** % d'individus sense problemes en totes les dimensions del EQ5D

La taula 3 mostra els coeficients estimats del model de regressió tobit i del model de regressió lineal múltiple. Les variables sociodemogràfiques d'educació, grandària de municipi i nivell socioeconòmic no van ser incloses en els models perquè no eren estadísticament significatives i la seva presència tampoc no milloraven el model significativament en termes de la variància explicada (model lineal) o de disminució del *Akaike Information Criteria* (model tobit).

Taula 3. Resultats del model de regressió tobit i model de regressió lineal en la mostra de l'estudi ESEMeD

		Model tobit		Model lineal (OLS)	
		Coeficient (EE)	Khi-quadrat (p-valor)	Coeficient (EE)	t-test (p-valor)
Constant		1,358 (0,02)	6040,45 (<0,001)	0,980 (0,005)	185,09 (<0,001)
Sexe: Home		0,038 (0,01)	15,83 (<0,001)	0,013 (0,003)	4,00 (<0,001)
Edat (Ref. 35-49):	18-24	0,127 (0,02)	35,40 (<0,001)	0,025 (0,007)	3,88 (<0,001)
	25-34	0,085 (0,02)	30,42 (<0,001)	0,014 (0,005)	2,97 (0,003)
	50-64	-0,055 (0,01)	16,83 (<0,001)	-0,014 (0,005)	-2,87 (0,004)
	65 o més	-0,167 (0,02)	103,79 (<0,001)	-0,060 (0,006)	-10,21 (<0,001)
Estat civil (Ref. Casat): Prèviament casat		-0,047 (0,01)	11,63 (<0,001)	-0,026 (0,005)	-4,95 (<0,001)
	Mai casat	-0,016 (0,02)	1,20 (0,273)	-0,004 (0,005)	-0,77 (0,439)
Situació laboral: Treball no remunerat		-0,032 (0,01)	7,28 (0,007)	-0,010 (0,004)	-2,49 (0,013)
País (Ref. Espanya):	Bèlgica	-0,088 (0,03)	11,96 (<0,001)	-0,017 (0,009)	-1,86 (0,063)
	França	-0,124 (0,02)	62,92 (<0,001)	-0,023 (0,005)	-4,29 (<0,001)
	Alemanya	-0,058 (0,02)	15,07 (<0,001)	-0,012 (0,005)	-2,38 (0,018)
	Itàlia	-0,018 (0,02)	1,32 (0,251)	-0,001 (0,005)	-0,13 (0,897)
	Holanda	-0,082 (0,02)	14,20 (<0,001)	-0,016 (0,007)	-2,18 (0,029)
Depressió		-0,188 (0,02)	79,86 (<0,001)	-0,086 (0,008)	-10,18 (<0,001)
Distímia		-0,094 (0,04)	4,64 (0,031)	-0,032 (0,018)	-1,75 (0,081)
Fòbia específica		-0,066 (0,02)	12,82 (<0,001)	-0,036 (0,007)	-5,18 (<0,001)
Fòbia social		-0,147 (0,032)	20,93 (<0,001)	-0,058 (0,013)	-4,55 (<0,001)
Trastorn de pànic		-0,130 (0,04)	8,57 (0,003)	-0,074 (0,018)	-4,11 (<0,001)
Trastorn per estrès posttraumàtic		-0,086 (0,04)	5,34 (0,021)	-0,052 (0,015)	-3,46 (<0,001)
Agorafòbia sense pànic		-0,116 (0,06)	3,89 (0,049)	-0,029 (0,025)	-1,18 (0,237)
Trastorn d'ansietat generalitzada		-0,076 (0,05)	2,27 (0,132)	-0,045 (0,020)	-2,22 (0,027)
Abús d'alcohol		-0,137 (0,06)	4,95 (0,026)	-0,056 (0,022)	-2,51 (0,012)
Dependència d'alcohol		-0,028 (0,09)	0,09 (0,763)	-0,016 (0,031)	-0,50 (0,618)
Artrosis o reumatisme		-0,144 (0,01)	115,92 (<0,001)	-0,080 (0,005)	-14,75 (<0,001)
Problemes cervicals o d'esquena		-0,184 (0,01)	280,70 (<0,001)	-0,064 (0,004)	-15,33 (<0,001)
Mals de cap freqüents o molt forts		-0,093 (0,02)	40,82 (<0,001)	-0,036 (0,005)	-6,63 (<0,001)
Qualsevol altre dolor crònic		-0,193 (0,02)	130,62 (<0,001)	-0,091 (0,007)	-13,09 (<0,001)
Al·lèrgies estacionals		0,026 (0,016)	2,53 (0,111)	0,012 (0,006)	2,16 (0,031)
Embòlia cerebral		-0,138 (0,042)	10,54 (0,001)	-0,088 (0,018)	-4,86 (<0,001)
Atac de cor		-0,030 (0,033)	0,81 (0,367)	-0,015 (0,013)	-1,13 (0,259)
Malaltia cardíaca		-0,072 (0,022)	11,10 (<0,001)	-0,047 (0,009)	-5,30 (<0,001)
Pressió alta		-0,023 (0,013)	2,99 (0,084)	-0,012 (0,005)	-2,31 (0,021)
Asma		-0,042 (0,024)	3,15 (0,076)	-0,009 (0,009)	-0,96 (0,336)

Taula 3. Resultats del model de regressió tobit i model de regressió lineal en la mostra de l'estudi ESEMeD (*continuació)*

	Model tobit		**Model lineal (OLS)**	
	Coeficient (EE)	Khi-quadrat (p-valor)	Coeficient (EE)	t-test (p-valor)
Altres malalties pulmonars	-0,097 (0,032)	9,43 (0,002)	-0,061 (0,013)	-4,64 (<0,001)
Diabetis	-0,065 (0,021)	9,71 (0,002)	-0,023 (0,008)	-2,77 (0,006)
Úlcera estomacal o intestinal	-0,124 (0,031)	15,77 (<0,001)	-0,086 (0,013)	-6,77 (<0,001)
Malaltia de la tiroides	0,074 (0,024)	9,06 (0,003)	0,023 (0,009)	2,60 (0,009)
Problema Neurològic	-0,302 (0,044)	46,42 (<0,001)	-0,218 (0,019)	-11,60 (<0,001)
Càncer	-0,097 (0,039)	6,33 (0,012)	-0,064 (0,015)	-4,17 (<0,001)
R^2	*0.2691*		*0.2837*	

Es pot observar que els coeficients obtinguts en el model tobit eren més grans en magnitud que els del model lineal. Cal recalcar que els coeficients obtinguts en el model tobit representen la contribució de cadascuna de les variables independents sobre el valor mig de la variable latent *y**, que pot prendre valors més grans que 1 (valor observable màxim del EQ-5D). Per aquesta raó la constant del model és més gran que 1 en el cas del model tobit. Tant en el model tobit com en el lineal, la majoria de trastorns tenien un efecte estadísticament significatiu, excepte la dependència d'alcohol, les al·lèrgies estacionals, l'atac de cor, la pressió alta i l'asma en el model tobit, i la distímia, l'agorafòbia sense pànic, la dependència d'alcohol, l'atac de cor, i l'asma en el model lineal.

A la taula 4 es presenten els resultats de la validació creuada per als dos models de regressió que comparem. Els resultats van ser molt similars entre ambdós models, encara que el coeficient R^2, vas ser significativament més elevat (millor) per al model de regressió lineal. Pel que fa a la resta d'indicadors, aquests van ser lleugerament millors per al model de regressió lineal.

Taula 4. Resultats de la validació creuada

	R^1 (de)	**R^2 (de)**	**MSE (de)**	**MAPE (de)**
Model Tobit	0.2577 (0.0019)	0.2371 (0.0038)	0.0920 (0.0003)	0.0226 (0.0002)
Regressió lineal	0.2606 (0.0020)	0.2554 (0.0030)	0.0916 (0.0003)	0.0220 (0.0002)

La figura 2 mostra el l'histograma dels residus i el gràfic de residus respecte als valors previstos del model tobit. Els valors predits del model tobit corresponen a la variable censurada E(yi | xi),

segons la fórmula que es presenta a la pàgina 8. Els residus es van calcular també a partir dels valors predits de la variable censurada. Els dos gràfics van ser molt similars als corresponents obtinguts amb el model lineal, que es mostren a la figura 3. Tant els residus del model tobit com els del model lineal van presentar distribucions lleugerament aplanades a l'esquerra, el coeficient d'asimetria (c.a.) va ser negatiu en els dos casos (c.a=-2.19 per al model tobit i c.a=-2.42 per al model lineal). A més, ambdues distribucions tenien forma leptocúrtica, indicant una concentració important de les dades al voltant de la mitjana (coeficient de curtosis propers a 10 en els dos casos). Pel que fa als gràfics de residus respecte als valors predits, no es va observar que la variabilitat dels residus augmentés conforme augmentaven els valors predits, per tant no sembla observar-se heteroscedasticitat dels residus.

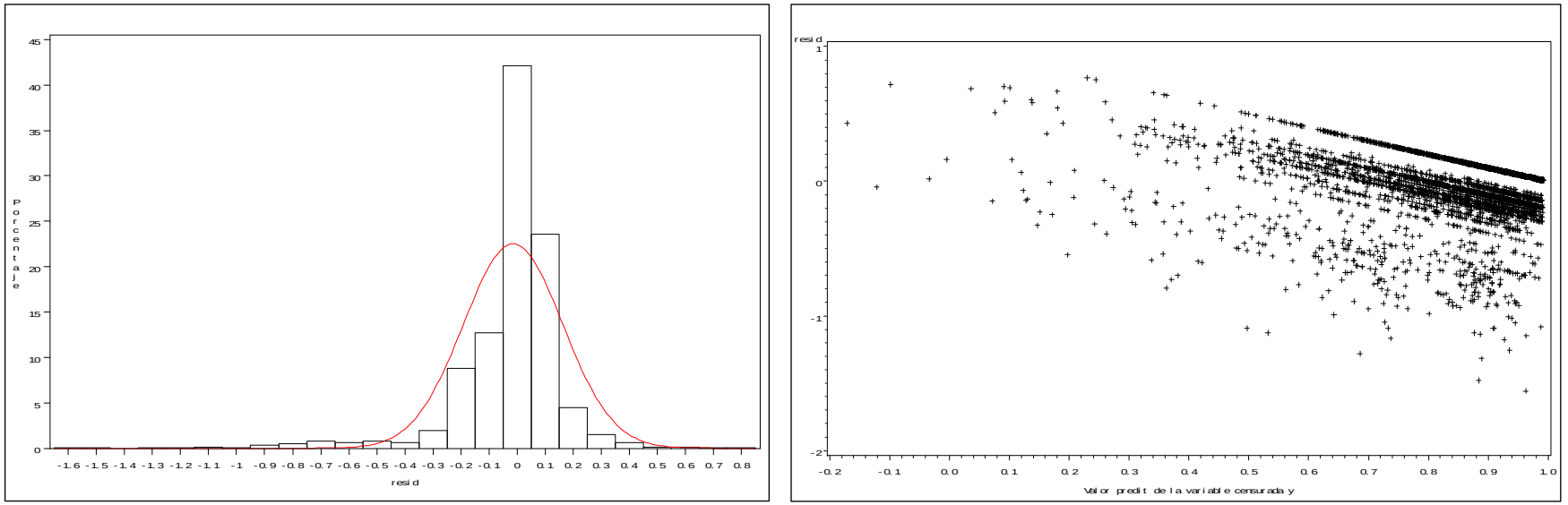

Figura 2. Histograma dels residus del model tobit i gràfic de residus respecte els valors predits del model tobit, estudi ESEMeD

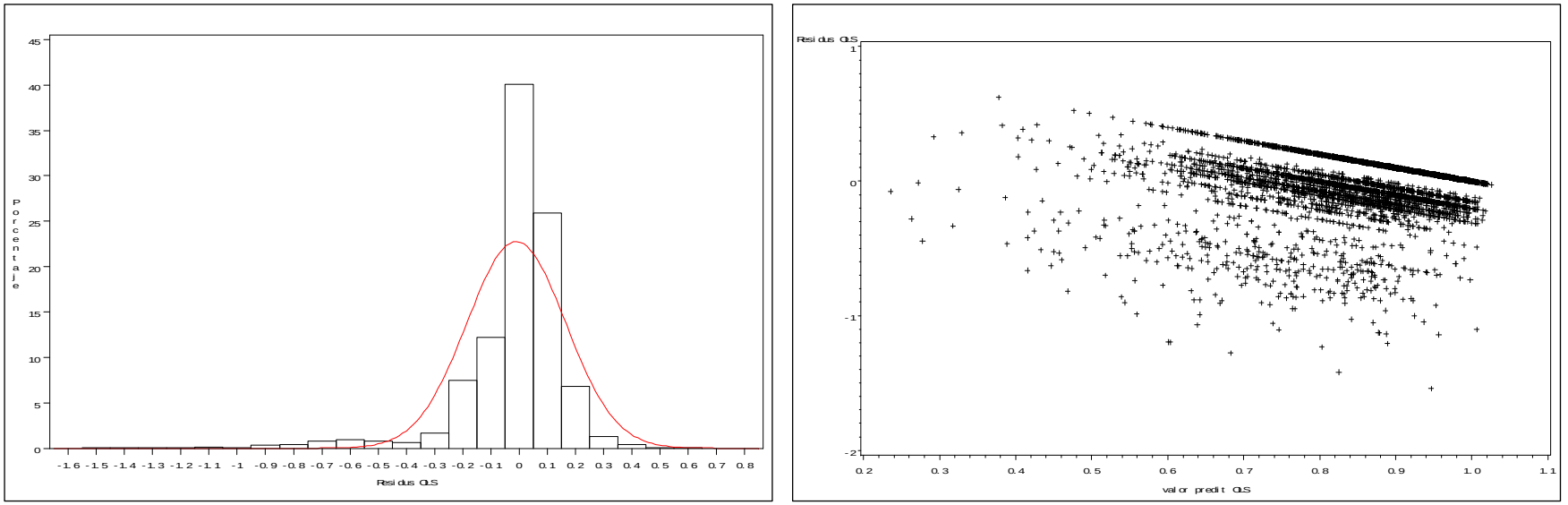

Figura 3. Histograma dels residus del model de regressió lineal i gràfic de residus respecte els valors predits del model lineal, estudi ESEMeD

3.2. Resultats de la simulació

Els resultats de la simulació es resumeixen a la taula 5.

Taula 5. Resultats dels models amb dades simulades. Valors fixats: β_0=3, β_1=5.5

Valor de σ	Model	$\hat{\beta}_0(EE)$	$\hat{\beta}_1(EE)$	Biaix Relatiu β_1	Residus EE	R^2	Censures n (%)
σ = 1	m. lineal de y*	2.9113 (0.0967)*	5.4898 (0.0402)*	0.19%	1.081	0.993	
	m. lineal de y *(a=-5)*	-7.4756 (0.2457)*	1.3800 (0.1022)*	74.91%	2.747	0.597	84 (67%)
	m. tobit *(a=-5)*	2.489 (0.603)*	5.346 (0.213)*	2.80%	1.20	0.974	
	m. lineal de y *(a=10)*	0.1779 (0.2653)	3.9754 (0.1104)*	27.72%	2.966	0.913	43 (34%)
	m. tobit *(a=10)*	2.908 (0.1567)*	5.484 (0.0741)*	0.29%	1.11	0.992	
σ = 10	m. lineal de y*	2.7106 (0.896)*	6.1155 (0.373)*	-11.19%	10.02	0.686	
	m. lineal de y *(a=-5)*	-9.0048 (0.5490)*	1.9740 (0.228)*	64.11%	6.138	0.378	80 (64%)
	m. tobit *(a=-5)*	2.67 (2.032)*	6.18 (0.801)*	-12.36%	9.98	0.526	
	m. lineal de y *(a=10)*	-1.3086 (0.7202)	4.216 (0.300)*	23.35%	8.052	0.617	46 (37%)
	m. tobit *(a=10)*	2.99 (1.1136)*	6.14 (0.0815)*	-11.64%	10.2	0.669	
σ = 20	m. lineal de y*	4.2261 (1.618)*	5.3102 (0.673)*	3.45%	18.09	0.336	
	m. lineal de y *(a=-5)*	-10.099 (0.728)*	1.4113 (0.303)*	74.34%	8.145	0.150	79 (63%)
	m. tobit *(a=-5)*	1.88 (2.524)*	4.54 (0.918)*	17.45%	16.6	0.162	
	m. lineal de y *(a=10)*	-2.134 (1.108)	3.046 (0.461)*	44.62%	12.39	0.262	49 (39%)
	m. tobit *(a=10)*	3.83 (1.842)*	5.08 (0.760)*	7.64%	17.6	0.270	

EE: Error estàndard; * p-valor<0.05

Quan la variància de l'error era petita (σ=1), el model tobit va explicar més del 97% de la variabilitat, mentre que en el cas del model lineal aquest percentatge va ser força inferior, arribant a una R^2=60% en el cas de *a=-5* (67% de dades censurades). Les estimacions dels coeficients del model tobit van ser molt semblants a les teòriques, amb un biaix relatiu de la pendent del model inferior al 3%. Tot i així, el biaix en les estimacions dels coeficients va ser més gran conforme augmentava el percentatge d'individus censurats, passant d'un biaix del 0,3% amb una censura del 34%, a un biaix del 2,8% amb una censura del 67%. En el cas del model lineal, es va obtenir un biaix del 75% quan el percentatge de censura era del 67%. A més a més, les estimacions dels coeficient de la pendent del model lineal van ser consistentment més baixes que el valor teòric

fixat i que les obtingudes amb el model tobit.

Conforme augmentava el valor de σ, el percentatge de variabilitat explicada pels models disminuïa considerablement, i les R^2 dels models tobit s'allunyaven de les obtingudes amb la regressió lineal amb la variable latent *y** com a dependent i s'aproximaven a les dels models lineals per predir la variable censurada *y*, especialment quan el percentatge de censura era elevat. Així, per σ=10, el percentatge de variabilitat del model lineal per estimar la variable latent va ser del 68.6%, mentre que el del model tobit amb *a = 10* (37% de censura) va ser lleugerament inferior (R^2 = 67%) i per *a = -5* va ser del 53%. Les R del model lineal van ser força inferiors, arribant al 38% quan el percentatge de censura era elevat. Per σ=20, la R^2 del model tobit va ser molt semblant a la del model de regressió lineal (aproximadament 16% quan el percentatge de censura és del 63% i 26% quan el percentatge de censura és del 39%), i força inferiors a la del model lineal de la variable latent y* (R^2 = 33,6%). Per $\sigma \geq 10$ el biaix de les estimacions de β_1 del model tobit va augmentar força, però en cap cas va superar el 20%. A més, quan la proporció de dades censurades era al voltant del 40%, el biaix de les estimacions de la pendent del model tobit es va aproximar al del model lineal per a la variable latent. En canvi, el biaix de les estimacions del model lineal per la variable censurada va ser molt més elevat, especialment quan la proporció de dades censurades era alta. Aquest percentatge va disminuir força quan la proporció d'individus amb censura era del 40%.

En els gràfics de residus respecte als valors previstos dels diferents models, que es poden consultar en l'apèndix 5, s'observa que per σ=1, els residus van mostrar un comportament totalment aleatori tant en el model lineal per a la variable latent *y** com en els dos models tobit estimats, però no en el cas dels 2 models lineals per a les variables censurades. En canvi, per valors de σ més grans, els gràfics de residus respecte a valors previstos dels models tobits es van assemblar força als dels models lineals per a la variable censurada corresponents.

4. Conclusions

En aquest treball hem descrit el model tobit, que en estudis previs s'ha presentat com una alternativa adequada als models lineals en l'estimació de dades amb un elevat percentatge d'individus amb la puntuació més alta (o més baixa) possible. Un exemple d'aquest tipus de dades són les mesures de l'estat de salut, que tendeixen a ser asimètriques i tenir un elevat efecte sostre.

Cal destacar que, a part del model tobit, diversos estudis han proposat altres mètodes alternatius per a l'anàlisi d'aquest tipus de variables, entre els quals destaca el model Censored Least Absolute Deviations (CLAD), que no ha estat avaluat en aquest treball. El model CLAD, en certes circumstàncies han demostrat ser millors que els models tobit, ja que es tracta d'una estimació no paramètrica que, per tant, és robusta a la violació de l'assumpció de la distribució normal dels residus. A més, també s'ha demostrat que és robusta a l'incompliment de l'assumpció sobre l'homoscedasticitat dels residus (Powell 1984).

En les dades simulades presentades, s'observa que quan la variància de l'error era petita (σ=1), el model tobit presentava estimacions no esbiaixades dels coeficients i prediccions acurades, especialment quan el percentatge d'individus amb la puntuació més alta possible era més petit (aproximadament del 35%), per bé que les prediccions eren lleugerament pitjors a les del models lineal en què la variable dependent era la variable latent *y**. En canvi, per valors de la variància de l'error més grans (s=10 o s=20) el percentatge de variabilitat explicada i les prediccions obtingudes amb el model tobit empitjoraven substancialment, arribant a ser similars a les obtingudes amb el model de regressió lineal per predir la variable censurada *y*. A més, els resultats eren pitjors conforme augmentava el percentatge de dades censurades. Aquest resultat concorda amb l'obtingut amb les dades d'ESEMeD, en què els valors predits del model tobit i el percentatge de variabilitat explicada eren molt similars als del model lineal, possiblement degut al fet que en les dades d'ESEMeD el percentatge d'individus amb efecte sostre era molt elevat (62%) i a més, els models estimats tenien una R^2 inferior al 30% tant en el cas del model tobit com en el cas del model lineal.

Així doncs, sembla que tant el percentatge de variabilitat estimat pels models, com el percentatge d'individus amb la puntuació màxima tenen un efecte important en el funcionament del model tobit en comparació amb el del model lineal.

Referències bibliogràfiques

Alonso, J., Angermeyer, M. C., *i cols.* (2004), "Sampling and methods of the European Study of the Epidemiology of Mental Disorders (ESEMeD) project," Acta Psychiatrica Scandinavica (Suppl), 8-20 [http://www.epremed.org (darrer accès: 23/05/2008].

Arabmazar, A. and Schmidt, P. (1982), "An Investigation of the Robustness of the Tobit Estimator to Non-Normality" Econometrica, 50, 1055-1064.

Austin, P. C., Escobar, M., and Kopec, J. A. (2000), "The use of the Tobit model for analyzing measures of health status," Quality of Life Research, 9, 901-910.

Bleda Hernández, M. J. and Tobías Garcés, A. (2002), "Application of tobit regression models in modelling censored epidemiological variables," Gaceta Sanitaria, 16, 188-194.

Cochran, W. G. (1977), "Sampling Techniques, 3rd Edition". Wiley: New York.

Dolan, P. (1997), "Modeling Valuations for EuroQol Health States" Medical Care, 35, 1095-1108.

Froberg, D. G. and Kane, R. L. (1989a), "Methodology for measuring health-state preferences -II: Scaling methods," Journal of Clinical Epidemiology, 42, 459-471.

-----(1989b), "Methodology for measuring health-state preferences -III: Population and context effects," Journal of Clinical Epidemiology, 42, 585-592.

Greene, W. (1999), "Marginal effects in the censored regression model," Economics Letters, 64, 43-49.

Greene, W. H. (2003a), Econometric Analysis (5th), New Jersey: Prentice Hall.

-----(2003b), "Limited Dependent Variable and Duration Models" in Econometric Analysis (5th ed.), New Jersey: Prentice Hall.

Imai, K., King, G., and Lau, O. (2005), "Zelig: Everyone's Statistical Software," R package version, 2-4.

Kessler, R. C. and Ustun, T. B. (2004), "The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite International Diagnostic Interview

(CIDI)," International Journal of Methods in Psychiatric Research, 13, 93-121.

Manning, W. G., Duan, N., and Rogers, W. H. (1987), "Monte Carlo evidence on the choice between sample selection and two-part models," Journal of Econometrics, 35, 59-82.

Mullahy, J. (1998), "Much ado about two: reconsidering retransformation and the two-part model in health econometrics," Journal of Health Economics, 17, 247-281.

Olsen, R. (1978), "Note on the Uniqueness of the Maximum Likelihood Estimator for the Tobit Model," Econometrica, 46, 1211-1215.

Powell, J. L. (1984), "Least absolute deviations estimation for the censored regression model," Journal of Econometrics, 25, 303-325.

R Development Core Team . R: A Language and Environment for Statistical Computing. 2006. Vienna, Austria, R Foundation for Statistical Computing. Computer program

Started, G. "The LIFEREG Procedure".

StataCorp . Stata Statistical Software. [Release 8]. 2003. StataCorp LP, College Station, TX. 2003. Computer program

The EuroQol Group (1990), "EuroQoL -A new facility for the measurement of health-related quality of life," Health Policy., 16, 199-208. [http://www.euroqol.org (darrer accès: 23/05/2008)]

Tobin, J. (1958), "Estimation of relationships for limited dependent variables," Econometrica, 26, 24-36.

Wittchen, H. U. (1994), "Reliability and validity studies of the WHO--Composite International Diagnostic Interview (CIDI): a critical review," Journal of Psychiatric Research, 28, 57-84.

Apèndix 1. l'Índex EQ-5D en anglès

HS2. I'm going to read several sets of statements. I want you to tell me which statement in each set best describes your health today. Here's the first set:

"I have no problems in walking around."
"I have some problems in walking around."
"I am confined to a bed today."

Which of these statements best describes you today?

NO PROBLEMS..... 1
SOME PROBLEMS 2
CONFINED TO BED 3

HS3. Here's the next set:

"I have no problems with self care today."
"I have some problems washing or dressing myself today."
"I am unable to wash or dress myself today."

(Which of these statements best describes you today?)

NO PROBLEMS..... 1
SOME PROBLEMS 2
UNABLE TO WASH OR DRESS 3

HS4. The next set deals with usual activities like work, study, homework, and leisure activities.

"I have no problems with performing my usual activities today."
"I have some problems with performing my usual activities today."
"I am unable to perform my usual activities today."

(Which of these statements best describes you today?)

NO PROBLEMS.....1
SOME PROBLEMS 2
UNABLE TO PERFORM ACTIVITIES.....3

HS5. The next set is about pain and discomfort.

"I have no pain or discomfort today."
"I have moderate pain or discomfort today."
"I have extreme pain or discomfort today."

(Which of these statements best describes you today?)

NO PAIN 1
MODERATE PAIN..... 2
EXTREME PAIN..... 3

*HS6. And here is the last set.

"I am not anxious or depressed today."
"I am moderately anxious or depressed today."
"I am extremely anxious or depressed today."

(Which of these statements best describes you today?)

NO ANXIOUS/DEPRESSED.....1
MODERATE ANXIOUS/DEPRESSED 2
EXTREME ANXIOUS/DEPRESSED 3

Apèndix 2. El model tobit en diferents paquets estadístics

El model tobit està implementat en molts dels paquets estadístics comunment utilitzats. Presentem la formulació del model tobit en els paquets estadístics SAS, R i STATA.

SAS v9.1:

El procediment en SAS per estimar el model tobit és el *PROC LIFEREG* (Started), que ajusta models paramètrics per a dades censurades, que poden ser tant censurades a la dreta, a l'esquerra o en un interval. La distribució de l'error s'assumeix normal en el model tobit, per bé que el procediment també permet escollir entre altres classes de distribucions, com l'Exponencial, la Weibull o la Lognormal, entre d'altres. El procediment estima els paràmetres amb el mètode de màxima versemblança mitjançant l'algoritme d'optimització de *Newton-Raphson*.
Sintaxi:

```
data <BD>;
set <BD>;
upper=<VARDEPENDENT>; /* creo la variable indicadora de la censura superior upper que pren valor igual a la
variable dependent si aquesta no està censurada i missing si està censurada*/
if <VARDEPENDENT>=<punt de censura> then upper=.;
run;

PROC lifereg DATA=<BD>;
    CLASS <llistat de variables categoriques>;
    WEIGHT <pes>;
    MODEL (<VARDEPENDENT>,<upper>)= <llistat de variables predictores>/ d=normal;
    OUTPUT out=out xbeta=<predlin>; /* guardem el predictor lineal*/
RUN;
```

R:

Existeixen tres procediments en R per estimar el model tobit:

Package survival (http://cran.r-project.org/web/packages/survival/survival.pdf):

- Instrucció **survreg**:

  La instrucció survreg del paquet survival permet estimar un model tobit. Fa servir un mètode d'estimació per *màxima versemblança* mitjançant l'algorisme d'optimització de *Newton-Raphson*.

  Exemple:

  ```
  tfit <- survreg(Surv(durable, durable>0, type='left') ~age + quant,
  data=tobin, dist='gaussian')
  > predict(tfit,type="response")
  ```

Package VGAM (cran.r-project.org/web/packages/VGAM/VGAM.pdf) :

- Instrucció **vglm**:

  La instrucció **vglm** inclou la funció *tobit*, que permet estimar el model tobit amb dades censurades, tant a la dreta com a l'esquerra. El mètode d'estimació utilitzat és el *Iteratively weighted least squares (IRLS).*

  Sintaxi:

```
tobit(Lower = 0, Upper = Inf, lmu="identity", lsd="loge",emu=list(),
esd=list(), imethod=1, zero=2)
```

  Exemple:

```
fit <- vglm(y ~ x, tobit(Lower=Lower, Upper=Upper), trace=TRUE)
```

Package Zelig (Imai, King and Lau 2005) :

Dins d'aquest paquet de R, hi ha dos tipus de models tobit que es poden estimar:

- **model =”tobit”**:

  Permet estimar models tobit per a dades censurades a l'esquerra.

  Sintaxi:

```
> z.out <- zelig(Y ~ X1 + X2, below = 0, above = Inf,
model = "tobit", data = mydata)
> x.out <- setx(z.out)
> s.out <- sim(z.out, x = x.out)
```

- **model =”tobit.bayes”:**

  Estima models de regressió lineal Bayesiana per a dades censurades, tant per l'esquerra com per la dreta, utilitzant la mostreig de Gibbs

  Sintaxi:

```
> z.out <- zelig(Y ~ X1 + X2, below = 0, above = Inf,
model = "tobit.bayes", data = mydata)
> x.out <- setx(z.out)> s.out <- sim(z.out, x = x.out)
```

STATA:

La instrucció ***tobit*** del paquet estadístic STATA permet estimar un model tobit (http://www.stata.com/help.cgi?tobit). Utilitza el mètode d'estimació de màxima versemblança i l'algoritme d'optimització de Newton-Raphson.

Sintaxi:

```
tobit depvar [indepvars] [if] [in] [weight] , ll[(#)] ul[(#)]  [vce(vcetype)]
[options]
```

on:

```
  * ll[(#)]             valor de la censura per l'esquerra
  * ul[(#)]             valor de la censura per la dreta
  * vce(vcetype) especifica el tipus d'error estàndard reportat, que inclou
errors estàndard assimptòtics (que són robustos a algun tipus de mala
especificació).
```

Apèndix 3. Macro en SAS per a la validació creuada

```
/*****************************************************************************/
/* CROSSVALIDACIÓ                                                            */
/*****************************************************************************/
options nofmterr nomprint nomlogic nosymbolgen nonotes nosource nostimer noprintmsglist;

data all4g;
set all4;
upper=eq5d; /* creo la variable indicadora de la censura superior upper */
if eq5d=1 then upper=.;
run;

%macro crossvalid;
  %do i=1 %to 100;
   %put "SPLIT= " &i;
   %*Seleccio de la mostra de training ;
     proc surveyselect data= all4g out=training&i method=srs samprate=75;
     run;
   %*seleccio de la mostra de validació;
     proc sql;
      create table validar&i as select * from all4g where caseid not in (select caseid from training&i);
     quit;
   %*Estimacio del model tobit en la mostra de training;
     ods output ModelInfo=model&i(keep=Label1 nValue1 rename=(nValue1=LogL)) ParameterEstimates=Pars&i;
      PROC lifereg DATA=training&i outest=outtest&i(keep=_scale_);
         WEIGHT wt5part2;
         MODEL (eq5d,upper)= sex_d age1_d age2_d age4_d age5_d mstat2_d mstat3_d empw_d country1_d country2_d
                             country3_d country4_d country5_d d_mde12_d d_dysh12_d d_sp12_d d_so12_d d_pds12_d
                             d_pts12_d d_agp12_d d_gadh12_d d_alah12_d d_ald12_d cc4a_d cc4b_d cc4c_d cc4d_d
                             cc4f_d cc4h_d cc4kl_d cc4n_d cc4o_d cc4p_d cc4q_d cc4s_d cc4i_d cc4e_d cc4j_d
                             cc4g_d/d=normal;
         output out=out&i xbeta=xbeta&i;
      RUN;
     ods output close;
   %*Guardem els parametres (excepte el descala) i els numerem;
     data redoparms ;
     set pars&i;
       retain time ;
       time + 1 ;
       if substr(Parameter,1,5) ne 'Scale';
     run ;
   %*Obtinc una fila amb els valors dels parametres de p1 a p40 (el nombre total de parametres);
     data redoparams ;
       array v [*]   p1-p40 ;
       retain p1-p40 ;
       format   p1-p40 10.6;
       set redoparms end=last ;
         v[time]=estimate ;
       if last then output ;
       keep p1-p40;
     run;
   %*Afegeixo a la mostra de validacio 40 columnes noves p1-p40 que contenen els parametres per calcular xbeta
    (a la mostra de validacio) ;
     data validar&i.tob;
       if _n_=1 then set redoparams ;
       set validar&i;
         xbetaval&i =  p1+sex_d * p2+ age1_d * p3+ age2_d * p4+ age4_d * p5 +age5_d* p6+ mstat2_d* p7+
                       mstat3_d* p8+ empw_d* p9+ country1_d* p10+ country2_d* p11+ country3_d* p12+
                       country4_d* p13+ country5_d* p14+ d_mde12_d* p15+ d_dysh12_d* p16+ d_sp12_d* p17+
                       d_so12_d* p18+ d_pds12_d* p19+ d_pts12_d* p20+ d_agp12_d* p21+ d_gadh12_d* p22+
                       d_alah12_d* p23+ d_ald12_d* p24+ cc4a_d* p25+cc4b_d* p26+ cc4c_d* p27+ cc4d_d* p28+
                       cc4f_d* p29+ cc4h_d* p30+ cc4kl_d* p31+ cc4n_d* p32+ cc4o_d* p33+ cc4p_d* p34+ cc4q_d*
                       p35+ cc4s_d* p36+ cc4i_d* p37+ cc4e_d* p38+ cc4j_d* p39+ cc4g_d* p40;
     run;
   %*Obtencio dels valors predits, residus, residus al quadrat, valor absolut dels resid, per a la variable
     censurada en la mostra de validacio;
     data predval&i;
       if _n_=1 then set outtest&i;
       set validar&i.tob;
         lambda&i = -pdf('NORMAL' ,(1-Xbetaval&i)/_scale_) / cdf('NORMAL' ,(1-Xbetaval&i)/_scale_);
         Predval&i = (1-cdf('NORMAL' , (1-Xbetaval&i)/_scale_))+ (cdf('NORMAL' , (1-Xbetaval&i)/_scale_)
                    *(Xbetaval&i + _scale_*lambda&i));
         Probnocensval&i= cdf('NORMAL' ,(1-Xbetaval&i)/_scale_);
         residval&i=eq5d-predval&i;
         absresidv&i=abs(residval&i);
         residvalsq&i=residval&i**2;
     run;
     proc univariate data=predval&i;
         var eq5d;
         weight wt5part2;
         output out=meaneq5dval&i mean=meanval&i;
     run;
```

```
  data predval&i.2;
    if _n_=1 then set meaneq5dval&i;
    set predval&i;
      abserrval&i=abs(eq5d-meanval&i);
      errsqval&i=(eq5d-meanval&i)**2;
      keep caseid eq5d wt5part2 lambda&i xbetaval&i predval&i Probnocensval&i residval&i absresidv&i
             residvalsq&i abserrval&i errsqval&i _scale_;
  run;

%*Obtencio dels valors predits, residus, residus al quadrat, valor absolut dels resid, per a la variable
  censurada en la mostra de training;
  data predtrain&i;
    if _n_=1 then set outtest&i;
    set out&i;
      lambda&i = -pdf('NORMAL', (1-Xbeta&i)/_scale_) / cdf('NORMAL',(1-Xbeta&i)/_scale_);
      Predtrain&i = (1-cdf('NORMAL', (1-Xbeta&i)/_scale_))+ (cdf('NORMAL', (1-Xbeta&i)/_scale_)* (Xbeta&i +
                    _scale_*lambda&i));
      Probnocenstrain&i= cdf('NORMAL', (1-Xbeta&i)/_scale_);
      residtrain&i=eq5d-predtrain&i;
      absresidtr&i=abs(residtrain&i);
      residtrsq&i=residtrain&i**2;
  run;
  proc univariate data=predtrain&i;
      var eq5d;
      weight wt5part2;
      output out=meaneq5dtrain&i mean=meantrain&i;
  run;
  data predtrain&i.2;
      if _n_=1 then set meaneq5dtrain&i;
      set predtrain&i;
      abserr&i=abs(eq5d-meantrain&i);
      errsq&i=(eq5d-meantrain&i)**2;
      keep caseid eq5d wt5part2 lambda&i xbeta&i predtrain&i Probnocenstrain&i residtrain&i absresidtr&i
           residtrsq&i abserr&i errsq&i _scale_;
  run;

%*ESTIMACIo DEL MODEL OLS AMB la mostra de training;
  ods output parameterestimates=par_ols&i  FitStatistics=fits_ols&i  ANOVA=anova_ols&i;
  PROC reg DATA=training&i;
     WEIGHT wt5part2;
     MODEL eq5d= sex_d age1_d age2_d age4_d age5_d mstat2_d mstat3_d empw_d country1_d country2_d
                 country3_d country4_d country5_d d_mde12_d d_dysh12_d d_sp12_d d_so12_d d_pds12_d
                 d_pts12_d d_agp12_d d_gadh12_d d_alah12_d d_ald12_d cc4a_d cc4b_d cc4c_d cc4d_d cc4f_d
                 cc4h_d cc4kl_d cc4n_d cc4o_d cc4p_d cc4q_d cc4s_d cc4i_d cc4e_d cc4j_d cc4g_d;
     output out=outols&i predicted=pred_ols&i residual=res_ols&i cookd=cook_ols&i h=leverage_ols&i;
  RUN;
  ods output close;
%*Guardem els parametres i els numerem;
  data redoparms ;
  set par_ols&i;
    retain time ;
    time + 1 ;
  run ;
%*Obtinc una fila amb els valors dels paràmetres de p1 a p40 (el nombre total de parametres);
  data redoparams ;
     array v [*]   p1-p40 ;
     retain p1-p40 ;
     format   p1-p40 10.6;
     set redoparms end=last ;
         v[time]=estimate ;
         if last then output ;
         keep p1-p40;
  run;
%*Afegeixo a la mostra de validacio 40 columnes noves p1-p40 que contenen els parametres per calcular xbeta
  (a la mostra de validacio) ;
  data predval_ols&i;
    if _n_=1 then set redoparams ;
    set validar&i;
      xbetaols&i =  p1+sex_d * p2+ age1_d * p3+ age2_d * p4+ age4_d * p5 +age5_d* p6+ mstat2_d* p7+
                    mstat3_d* p8+ empw_d* p9+ country1_d* p10+ country2_d* p11+ country3_d* p12+
                    country4_d* p13+ country5_d* p14+d_mde12_d* p15+ d_dysh12_d* p16+ d_sp12_d* p17+
                    d_so12_d* p18+ d_pds12_d* p19+ d_pts12_d* p20+ d_agp12_d* p21+ d_gadh12_d* p22+
                    d_alah12_d* p23+ d_ald12_d* p24+ cc4a_d* p25+cc4b_d* p26+ cc4c_d* p27+ cc4d_d* p28+
                    cc4f_d* p29+ cc4h_d* p30+ cc4kl_d* p31+ cc4n_d* p32+ cc4o_d* p33+ cc4p_d* p34+
                    cc4q_d* p35+ cc4s_d* p36+ cc4i_d* p37+ cc4e_d* p38+ cc4j_d* p39+ cc4g_d* p40;
      resid&i=eq5d-xbetaols&i;
      absresid&i=abs(resid&i);
      residsq&i=resid&i**2;
  run;
  data predval_ols&i.2;
    if _n_=1 then set meaneq5dval&i;
    set predval_ols&i;
      abserr&i=abs(eq5d-meanval&i);
      errsq&i=(eq5d-meanval&i)**2;
      keep caseid eq5d wt5part2 xbetaols&i  resid&i absresid&i residsq&i abserr&i errsq&i;
  run;
```

```
%*Calculo a la mostra de training els resid, absresid, residsq;
  data predtrain_ols&i.2;
    if _n_=1 then set meaneq5dtrain&i;
    set outols&i;
      absresidtr&i=abs(res_ols&i);
      residsqtr&i=res_ols&i**2;
      abserrtr&i=abs(eq5d-meantrain&i);
      errsqtr&i=(eq5d-meantrain&i)**2;
      keep caseid eq5d wt5part2 pred_ols&i res_ols&i cook_ols&i leverage_ols&i absresidtr&i residsqtr&i
           abserrtr&i errsqtr&i;
  run;
%*Faig la suma dels residus al quadrat, dels valors absoluts dels residus, de la dif entre valor observat i
  la mitjana al quadrat (errsq) i en valor absolut (abserr);
  proc univariate data=predval&i.2;
      var absresidv&i residvalsq&i abserrval&i errsqval&i ;
      weight wt5part2;
      output out=valtobitsum&i sum= absrestv_sum ressqtv_sum abserrtv_sum errsqtv_sum;
  run;
  proc univariate data=predval&i.2;
      var absresidv&i residvalsq&i ;
      weight wt5part2;
      output out=valtobitmean&i mean= absrestv_mean ressqtv_mean ;
  run;
  data valtobitmean&i;
    set valtobitmean&i;
      split=&i;
  run;
  data valtobit&i;
    merge valtobitsum&i valtobitmean&i;
  run;
  proc univariate data=predtrain&i.2;
      var absresidtr&i residtrsq&i abserr&i errsq&i ;
      weight wt5part2;
      output out=traintobitsum&i sum= absrestt_sum ressqtt_sum abserrtt_sum errsqtt_sum;
  run;
  proc univariate data=predtrain&i.2;
      var absresidtr&i residtrsq&i ;
      weight wt5part2;
      output out=traintobitmean&i mean= absrestt_mean ressqtt_mean ;
  run;
  data traintobitmean&i;
    set traintobitmean&i;
      split=&i;
  run;
  data traintobit&i;
      merge traintobitsum&i traintobitmean&i;
  run;
  proc univariate data=predval_ols&i.2;
      var absresid&i residsq&i abserr&i errsq&i ;
      weight wt5part2;
      output out=valolssum&i sum= absresov_sum ressqov_sum abserrov_sum errsqov_sum;
  run;
  proc univariate data=predval_ols&i.2;
      var absresid&i residsq&i ;
      weight wt5part2;
      output out=valolsmean&i mean= absresov_mean ressqov_mean;
  run;
  data valolsmean&i;
    set valolsmean&i;
      split=&i;
  run;
  data valols&i;
      merge valolssum&i valolsmean&i;
  run;
  proc univariate data=predtrain_ols&i.2;
      var absresidtr&i residsqtr&i abserrtr&i errsqtr&i ;
      weight wt5part2;
      output out=trainolssum&i sum= absresot_sum ressqot_sum abserrot_sum errsqot_sum;
  run;
  proc univariate data=predtrain_ols&i.2;
      var absresidtr&i residsqtr&i ;
      weight wt5part2;
      output out=trainolsmean&i mean= absresot_mean ressqot_mean;
  run;
  data trainolsmean&i;
    set trainolsmean&i;
      split=&i;
  run;
  data trainols&i;
      merge trainolssum&i trainolsmean&i;
  run;
  data split&i;
      merge valtobit&i traintobit&i valols&i trainols&i; by split;
  run;
```

```
   %*Acumulem splits;
         %if &i=1 %then %do;
              data allsplits;
                    set split&i;
              run;
         %end;
         %else %do;
              data allsplits;
                    set allsplits split&i;
              run;
         %end;
   %*Esborro els arxius q no em fan falta;
     proc datasets nolist;
         delete valtobitsum&i valtobitmean&i valtobit&i traintobitsum&i traintobitmean&i traintobit&i
                valolssum&i valolsmean&i valols&i trainolssum&i trainolsmean&i trainols&i ;
     run;quit;
     dm "out;clear;";
  %end;
%mend;
%crossvalid;

libname outl 'H:\5_ESTADISTICA\doctoratUB\treballs de recerca\tobit\SAS\crossvalidacio';
proc datasets library=work;
 copy out=outl;
run;
/* Càlcul dels coeficients R1 i R2 per al model tobit i per al model OLS */
data allsplits;
set outl.allsplits;
   R1_tv=1-(absrestv_sum/abserrtv_sum);
   R2_tv=1-(ressqtv_sum/errsqtv_sum);
   R1_ov=1-(absresov_sum/abserrov_sum);
   R2_ov=1-(ressqov_sum/errsqov_sum);
run;
/* Càlcul de la mitjana i la desviació estàndard dels coeficients en les 100 rèpliques*/
proc univariate data=allsplits;
  var R1_tv R2_tv R1_ov R2_ov absrestv_mean ressqtv_mean absresov_mean ressqov_mean;
  *weight wt5part2;
  output out=R1R2 mean= meanR1_tv meanR2_tv meanR1_ov mean_R2ov meanMSE_tv meanMAPE_tv meanMSE_ov meanMAPE_ov
                  stdmean= seR1_tv seR2_tv seR1_ov se_R2ov seMSE_tv seMAPE_tv seMSE_ov seMAPE_ov;
run;
proc print;
run;
```

Apèndix 4. Programa en R utilitzat en l'estudi de simulació

```
n<-25
b0<-3 #coeficient intercept
b1<-5.5  #coeficient de x
k<-5
N<-n*k
#Genero una sequencia de n=25 ind repetits 5 vegades cadascun, de -4 a 4
x<-rep(seq(-4,4,length=n),each=k)
#Genero part determinista del model:
y.pred<-b0+b1*x
#genero una base de dades amb els valors de la y.latent,y.predit,y.resid
d<-data.frame(y.pred=y.pred, x=x)
#1a iteració: Fixo la SIGMA (desviació estàndard) a 1:
sigma<-1
#genero els errors, amb una distribució normal amb N ind amb mitjana zero
#i amb DE=SIGMA
d$e.sd1<-rnorm(N,mean=0,sd=sigma)
d$ystar.sd1<-d$y.pred+d$e.sd1
#regressió lineal per a la variable latent:
ols.ystar<-lm(d$ystar.sd1~d$x)
summary(ols.ystar)
#gràfic de residus respecte als valors predits:
plot(ols.ystar$fitted.values,ols.ystar$residuals)
#creo la y* (que és la que observem) ystar:
b<-(-5)
d$y.sd1.b5n<-pmin(d$ystar.sd1,b)
#Regressió lineal per a la variable censurada:
ols.y<-lm(d$y.sd1.b5n~d$x)
summary(ols.y)
#gràfic de residus respecte als valors predits:
plot(ols.y$fitted.values,ols.y$residuals)
#Model tobit en R
tobit<-survreg(formula = Surv(d$y.sd1.b5n, d$y.sd1.b5n < (-5), type = "right")
~ d$x, dist = "gaussian")
stobit<-summary(tobit)
#Calculo el valor predit de la variable censurada (right)
FI<-pnorm((b-tobit$linear.predictors)/stobit$scale)
fi<-dnorm((b-tobit$linear.predictors)/stobit$scale)
lambda<- (-fi/FI)
d$pred.y.sd1.b5n<- (1-FI)*b+FI*(tobit$linear.predictors+stobit$scale*lambda))
#residus de la variable censurada
d$res.y.sd1.b5n<-d$y.sd1.b5n-d$pred.y.sd1.b5n
#plot dels residus respecte als valors predits
plot(d$pred.y.sd1.b5n,d$res.y.sd1.b5n)
R.squared<-1-(sum(d$res.y.sd1.b5n^2)/sum((d$y.sd1.b5n-mean(d$y.sd1.b5n))^2))
R.squared
#Canvio el punt de censura i repeteixo
b<-10
d$y.sd1.b10<-pmin(d$ystar.sd1,b)
#Regressió lineal per a la variable censurada:
ols.y<-lm(d$y.sd1.b10~d$x)
summary(ols.y)
#gràfic de residus respecte als valors predits:
plot(ols.y$fitted.values,ols.y$residuals)
#Model tobit en R
tobit<-survreg(formula = Surv(d$y.sd1.b10, d$y.sd1.b10 < 10, type = "right") ~
              d$x, dist = "gaussian")
stobit<-summary(tobit)
```

```
#Calculo el valor predit de la variable censurada (right)
FI<-pnorm((b-tobit$linear.predictors)/stobit$scale)
fi<-dnorm((b-tobit$linear.predictors)/stobit$scale)
lambda<- (-fi/FI)
d$pred.y.sd1.b10<- (1-FI)*b+FI*(tobit$linear.predictors+(stobit$scale*lambda))
#residus de la variable censurada
d$res.y.sd1.b10<-d$y.sd1.b10-d$pred.y.sd1.b10
#plot dels residus respecte als valors predits
plot(d$pred.y.sd1.b10,d$res.y.sd1.b10)
R.squared<-1-(sum(d$res.y.sd1.b10^2)/sum((d$y.sd1.b10-mean(d$y.sd1.b10))^2))
R.squared
```

Apèndix 5. Gràfics de residus respecte valors predits dels models estimats amb les dades simulades

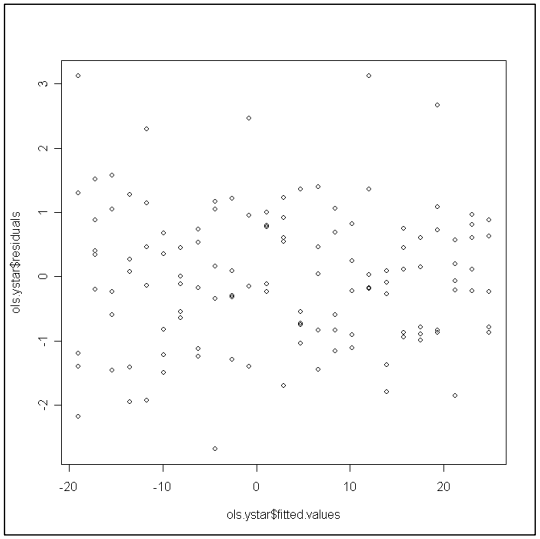

Figura 4. Residus versus predits model lineal de *y (σ=1)**

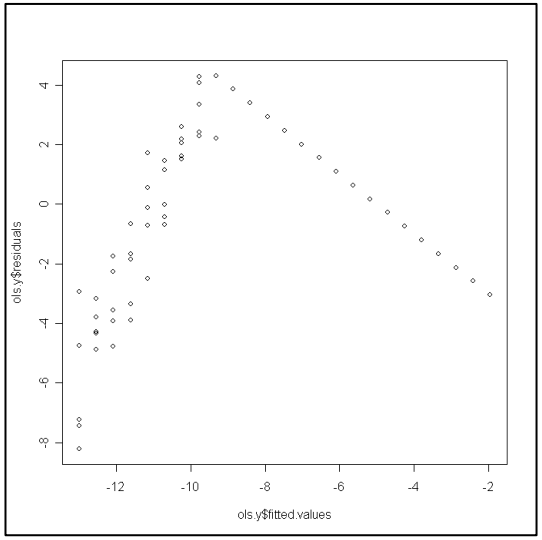

Figura 5. Residus versus predits model lineal de *y* (σ=1, *a*= -5)

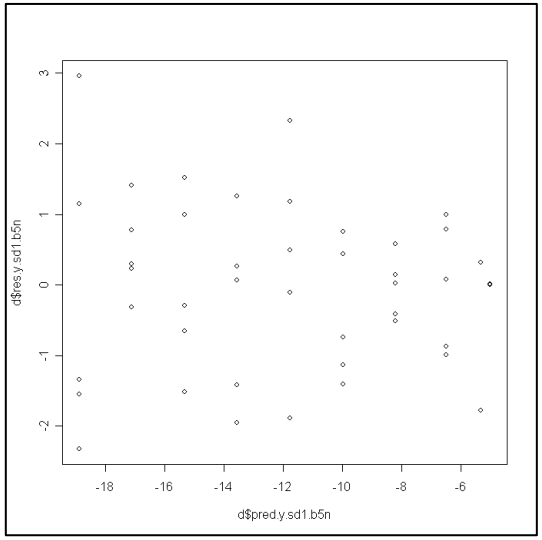

Figura 6. Residus versus predits model tobit (σ=1, *a*= -5)

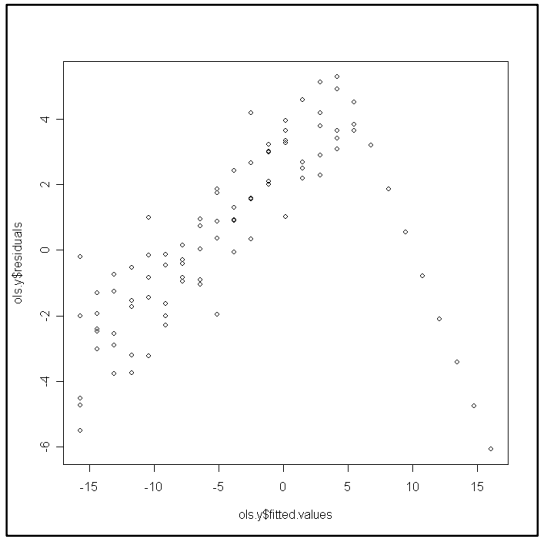

Figura 7. Residus versus predits model lineal de y (σ=1, *a*=10)

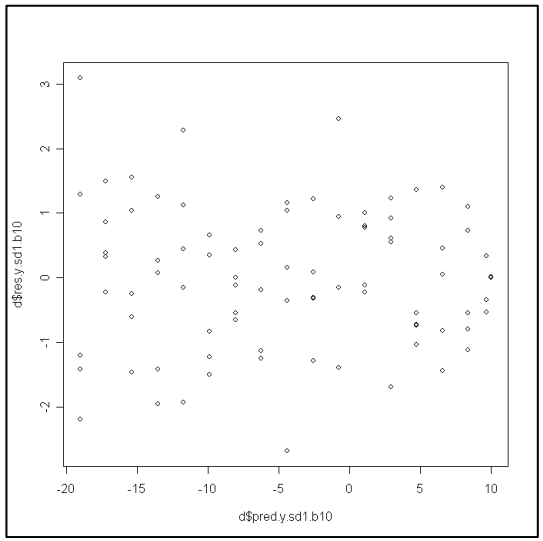

Figura 8. Residus versus predits model tobit (σ=1, *a*=10)

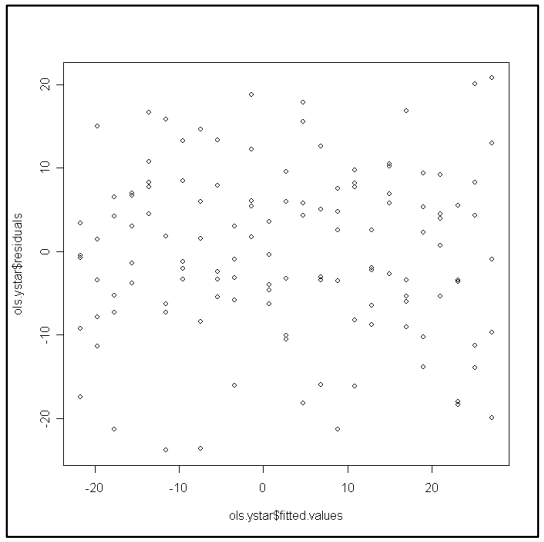

Figura 9. Residus versus predits model lineal de *y** **(σ=10)**

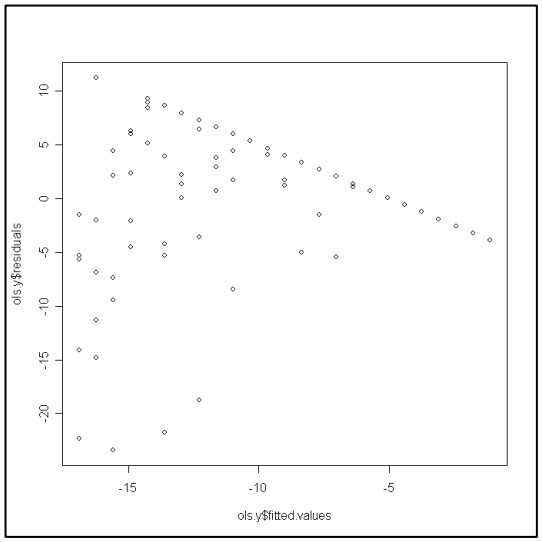

Figura 10. Residus versus predits model lineal de *y* **(σ=10, *a*= -5)**

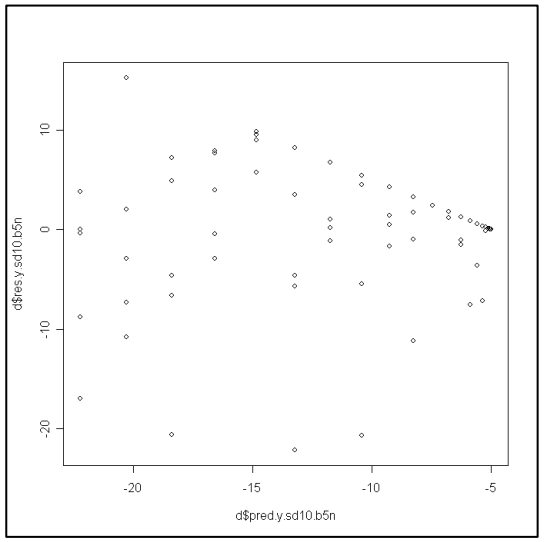

Figura 11. Residus versus predits model tobit (σ=10, *a*= -5)

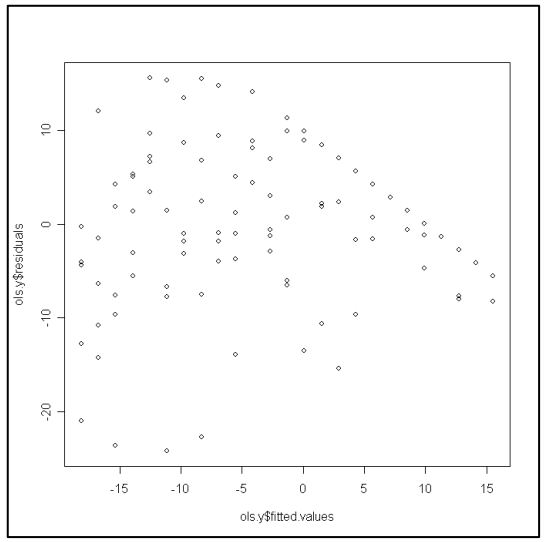

Figura 12. Residus versus predits model lineal de *y* **(σ=10, *a*= 10)**

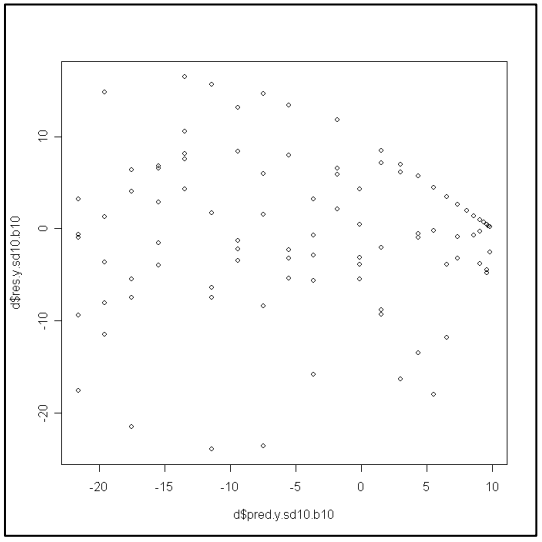

Figura 13. Residus versus predits model tobit (σ=10, *a*= 10)

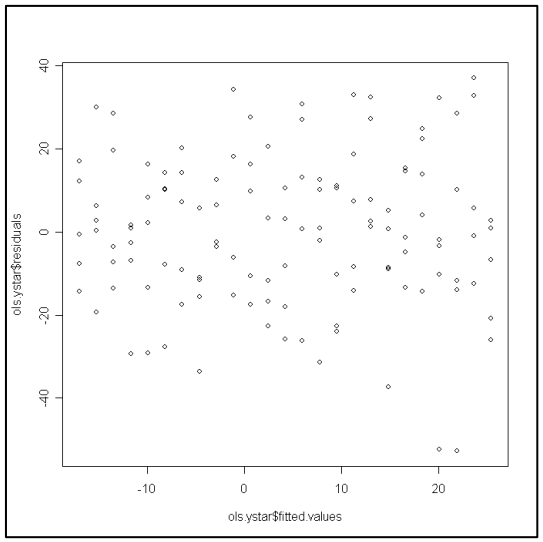

Figura 14. Residus versus predits model lineal de *y (σ=20)**

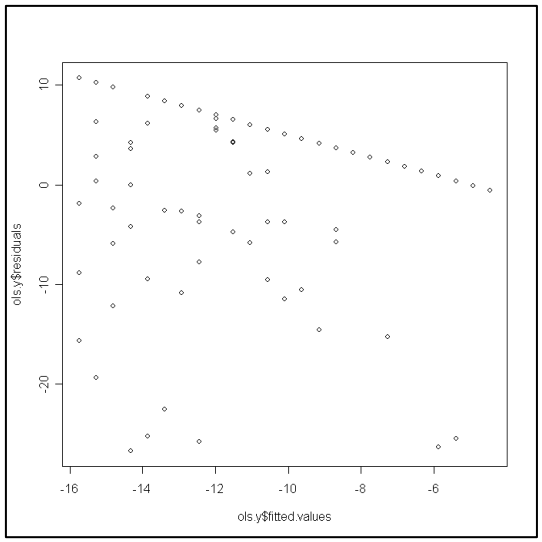

Figura 15. Residus versus predits model lineal de *y* (σ=20, *a*= -5)

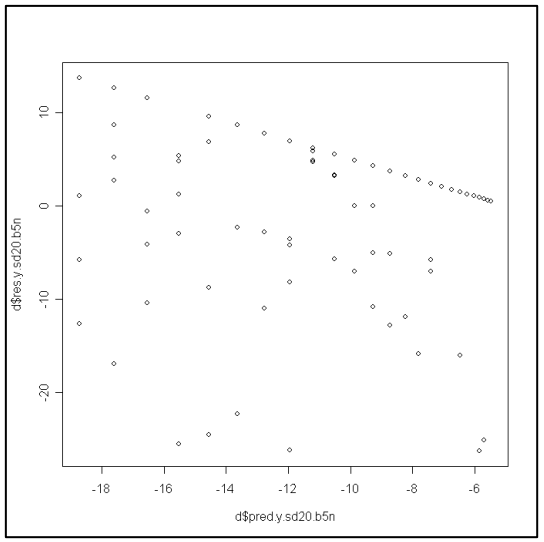

Figura 16. Residus versus predits model tobit (σ=20, *a*= -5)

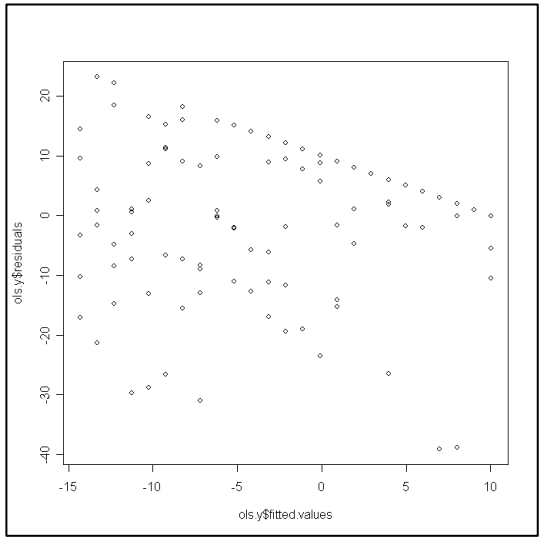

Figura 17. Residus versus predits model lineal de *y* (σ=20, *a*= 10)

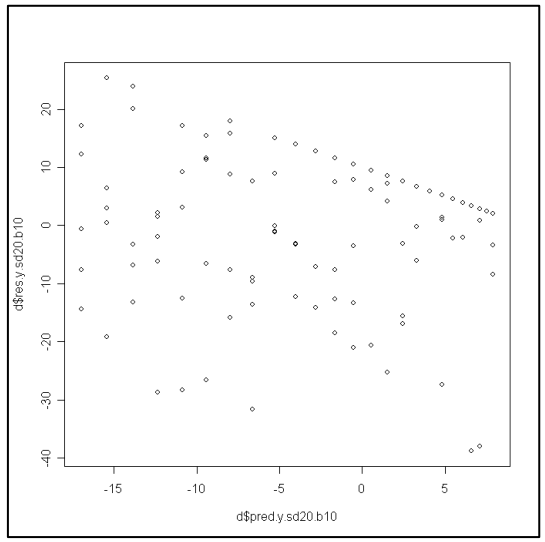

Figura 18. Residus versus predits model tobit (σ=20, *a*= 10)