EC475 Problem set 4

Numerical optimization

Antoine Goujard

26/11/09

We are interested in a counting model where for all t, $Y_t \in \mathbb{N} \bigcup \{0\}$.
We are assuming that the data are generated by:
$Prob(Y_t = y|Z_t = z) = exp(-\lambda_t)\frac{\lambda_t^y}{y!} = p_\theta(y|z)$
With, $\lambda_t = exp(\alpha + \beta.z_{1t} + \gamma.z_{2t}) = exp(z_t'\theta)$
Then, $E(Y_t|Z_t) = Var(Y_t|Z_t) = \lambda_t$.

We want to use MLE method.
$\hat{\theta} = argmax_\theta L_T = argmax_\theta \prod_{t=1}^T p_\theta(y_t|z_t)$
Or, $\hat{\theta} = argmax_\theta LL_T$
With $LL_T = \sum_{t=1}^T ln(p_\theta(y_t|z_t)) = \sum_{t=1}^T \ell_t$
$\ell_t = -\lambda_t + y_t.ln(\lambda_t) - ln(y_t!)$

To compute the FOCs, we need: $\frac{\partial \ell_t}{\partial \theta'}(\theta) = (y_t - \lambda_t(\theta)).z_t'$
So, $\frac{\partial LL_T}{\partial \theta'}(\hat{\theta}) = \sum_{t=1}^T (y_t - \lambda_t(\hat{\theta})).z_t' = 0$.

Contrary to OLS we find no-closed form solution for $\hat{\theta}$.

The score vector is: $LL_{T,\theta} = (\frac{\partial LL_T}{\partial \theta})_{3\times 1} = \sum_{t=1}^{T}(y_t - \lambda_t(\hat{\theta})).z_t$

Note that the Hessian is: $\mathcal{H}(\theta) = \frac{\partial^2 LL_T}{\partial\theta\partial\theta'}(\theta) = -\sum_{t=1}^{T} exp(z_t'\theta)z_t.z_t'$ n.d.
The log-likelihood is globally $(\forall\theta)$ and strictly (no flat area) concave.

By definition of our problem, $LL_{T,\theta}(\hat{\theta}) = 0$. This set of FOCs are called the maximum likelihood equations.

We apply the Newton-Raphson algorithm to find the solution of our non-linear optimization problem. This proceeds by iterations from a starting value $\theta^{(0)}$. After m steps, we get $\theta^{(m)}$ and we define $\theta^{(m+1)}$ by:

$$\theta^{(m+1)} = \theta^{(m)} - \underbrace{[S^{(m)}]^{-1}}_{\text{step}} \cdot \underbrace{\frac{\partial LL}{\partial\theta}(\theta^{(m)})}_{\text{slope}}$$

Where $S^{(m)} = \mathcal{H}(\theta^{(m)}) = \sum_{t=1}^{T} \frac{\partial^2 \ell_t}{\partial\theta\partial\theta'}(\theta^{(m)})$ (Newton Raphson or NR).
Or, $S^{(m)} = -\sum_{t=1}^{T} \frac{\partial \ell_t}{\partial\theta}(\theta^{(m)})\frac{\partial \ell_t}{\partial\theta'}(\theta^{(m)})$ (Berndt-Hall-Hall-Hausman or BHHH).

In the GAUSS iteration loop of the file `optpoiss.gcf`, 4 stopping criterions are defined:

1. $c1 = maxc(abs(b - b0))$; /* $max|\theta^{(m)} - \theta^{(m-1)}|$ */
2. $c2 = maxc(abs(g))$; /* Slope/score (FOCs) $g = LL_{T,\theta}(\theta^{(m-1)})$ */
3. $c3 = abs(l - l0)$; /* $|LL_T(\theta^{(m)}) - LL_T(\theta^{(m-1)})|$ */
4. $c4 = abs(gradstep)$; /* $gradstep = g'.[S^{(m-1)}]^{-1}.g$ */

For each of them the level of tolerance is set to `1.e-6`.
The algorithm stops at step `m` (ie, $\hat{\theta} = \theta^{(m)}$) if one of these conditions is satisfied or if the number of iterations is greater than `100`.

```
/* step 1:  initial values + parameters */
maxiter= 50 ; /* our maximum number of steps */
tol= 10^(-6) ; /* when we will decide that the change does
not matter */
Let theta= 0 0 0 ; /* initial values for theta 3*1 */
l= -100000000 ; /* initial value for log-likelihood */
crit= 1000 ; /* initial value of criterion that we
compare to tol to decide to stop */
ITER= 0 ; /* we initialize the ITER */

/* step 2:  loop with specific stopping rules */
do until crit<tol OR ITER==maxiter ;
theta0=theta ; l0=l ; /* previous values of theta and
log-lik */
l= f1(theta) ; /* log-lik (rk.  here theta=theta0) */
g= f2(theta) ; /* analytical score */
h= f3(theta) ; /* analytical hessian */
```

```
/*step 3:  iteration of the algorithm */
theta=theta0-inv(h)*g ; /* Newton Raphson iteration */
crit=abs(l-l0) ; /* criterion to stop, if we use c3 */
ITER=ITER+1 ; /* count the iterations */
endo ; /* go back to the start of the loop "do until" */

/*step 4:  printing the final results */
print "number of iterations:  " ;; ITER ;
print "log-likelihood:  " ;; l ;
print "theta:" ;; theta' ;
print "Score :  " ;; g' ;
print "Var-covar(theta):  " ;; -1*inv(h) ;
```

Rk. The problem is now to write f_1, f_2 and f_3 in Gauss matrix language.

Suppose we store the explanatory variable into $x_{n\times k}$ and the dependent variable into $y_{n\times 1}$.

$f_1(\theta) = LL_T(\theta) = \sum_{t=1}^{T} \ell_t(\theta)$

$\ell_t(\theta) = \underbrace{-\lambda_t}_{a_t(\theta)} + \underbrace{y_t.ln(\lambda_t)}_{b_t(\theta)} \underbrace{-ln(y_t!)}_{c_t}$

(c_t) is fixed wrt θ, `in Gauss c=-ln(y!) ;`

$(a_t)(\theta)$, `in Gauss a=-exp(x*theta) ;`

$(b_t)(\theta)$, `in Gauss b=(x*theta).*y ;`

So the log-likelihood at θ is `in Gauss l=sumc(a+b+c) ;`

Similarly $f_2(\theta) = g$ is `in Gauss g=x'*(y-exp(x*theta)) ;`

And $f_3(\theta) = h$ is `in Gauss h=-x'*(exp(x*theta).*x) ;`

We want to test: $H_0 : \alpha = \gamma = 0$ vs $H_1 : \alpha \neq 0$ or $\gamma \neq 0$.

We can rewrite this test as $H_0 : R.\theta = 0$ vs $H_1 : R.\theta \neq 0$.

With $R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

We know that the unconstrained estimator $\hat{\theta}$ is such that:
$\sqrt{T}.(\hat{\theta}_T - \theta) \Rightarrow \mathcal{N}(0, V)$

The variance V is given by: $-[E(\frac{\partial^2 \ell_t}{\partial\theta\partial\theta'}(\theta))]^{-1} = [E(\frac{\partial \ell_t}{\partial\theta}(\theta)\frac{\partial \ell_t}{\partial\theta'}(\theta))]^{-1}$

rk. This equality only holds at the true value of θ if the model is well specified (see last slide). Failures of this $=$ are the basis of White Information Matrix tests.

This can be estimated by $\hat{V}_1 = -T.\mathcal{H}(\hat{\theta})^{-1} = -(\frac{1}{T}\sum_{t=1}^{T} \frac{\partial^2 \ell_t}{\partial\theta\partial\theta'}(\hat{\theta}))^{-1}$.

Or by $\hat{V}_2 = (\frac{1}{T}\sum_{t=1}^{T} \frac{\partial \ell_t}{\partial\theta}(\hat{\theta})\frac{\partial \ell_t}{\partial\theta'}(\hat{\theta}))^{-1}$

Thus the asymptotic variance of the MLE estimator is estimated by:
$\widehat{Vas}(\hat{\theta}) = \frac{1}{T}\hat{V}_1 = -\mathcal{H}(\hat{\theta})^{-1}$ or $(\sum_{t=1}^{T} \frac{\partial \ell_t}{\partial\theta}(\hat{\theta})\frac{\partial \ell_t}{\partial\theta'}(\hat{\theta}))^{-1}$.

The Wald test statistic is:
$W = \hat{\theta}'.R'.(R.\frac{\hat{V}(\hat{\theta})}{T}.R')^{-1}.R.\hat{\theta}$

The Likelihood ratio test statistic is:
$LR = 2.(LL_T(\hat{\theta}) - LL_T(\hat{\theta}_r))$

The score or Lagrange multiplier test statistic is given by:
$LM = LL'_{T,\theta}(\hat{\theta}_r).[\frac{\hat{V}(\hat{\theta}_r)}{T}]^{-1}.LL_{T,\theta}(\hat{\theta}_r)$
ie this is one of the stopping criterion of our algorithm evaluated at the constrained estimator ($LL_{T,\theta}(\hat{\theta}_r)$ is the score vector at $\hat{\theta}_r$).

All these statistics are $\Rightarrow \chi^2_2$ as T increases.

The point estimates are:

Estimates	NR	BHHH
$\hat{\alpha}$	0.487	0.487
$\hat{\beta}$	0.897	0.897
$\hat{\gamma}$	-0.286	-0.286
LL	-420.685	-420.685
# iterations	13	39

We also obtain:

1. True Hessian (N.R.): $\hat{W} = 272.92$ $\widehat{LR} = 259.43$ $\widehat{LM} = 261.94$
2. BHHH: $\hat{W} = 186.71$ $\widehat{LR} = 259.43$ $\widehat{LM} = 59.97$

- The points estimates are nearly the same for NR and BHHH.
- The log-likelihood at $\hat{\theta}$ is also very similar.
- The BHHH approximation is quite good.
- The NR algorithm converges faster than BHHH, because it has better theoretical properties. In particular, in well-behaved cases, it has a **quadratic rate of convergence** in a nbhd of θ: $||\theta^{(m)} - \theta|| \sim_{m \to +\infty} \mu.||\theta^{(m-1)} - \theta||^2$ (for some $\infty > \mu > 0$).
- The fact that the $\widehat{LM}$ statistics are different for NR and BHHH is not surprising. We can see (next technical slide) that the approximation of the Hessian by the outer product of the score is justified at the true value of θ which we estimate by $\hat{\theta}$. The equality does not hold for other values of θ. In particular if the restriction is **false**, $\hat{\theta}_r$ will be $\neq \theta$.

Justification for the Newton Raphson method. Under regularity conditions, if x is in a neighborhood of $\theta^{(m)}$:

$LL_T(x) \simeq$

$LL_T(\theta^{(m)}) + LL_{T,\theta}(\theta^{(m)})(x-\theta^{(m)}) + \frac{1}{2}(x-\theta^{(m)})'\mathcal{H}(\theta^{(m)})(x-\theta^{(m)}) := G(x)$

Maximizing $G(x)$ gives us: $x^* = \theta^{(m)} - \mathcal{H}(\theta^{(m)})^{-1} LL_{T,\theta}(\theta^{(m)})$.

Fisher information equality. The BHHH idea may be seen as an approximation of the Hessian matrix. Under regularity conditions, at the **true parameter value** θ_0, we have:

$$I = -E(\frac{\partial^2 ln(p_\theta(y|z))}{\partial\theta\partial\theta'}(\theta_0)|Z) = E(\frac{\partial ln(p_\theta(y|z))}{\partial\theta}(\theta_0).\frac{\partial ln(p_\theta(y|z))}{\partial\theta'}(\theta_0)|Z).$$

The proof is based on the fact that we can rewrite (in the discrete case):

$$I = -\sum_y \frac{\partial^2 p_\theta(y|z)}{\partial\theta\partial\theta'}(\theta_0) + \sum_y [\frac{\partial ln(p_\theta(y|z))}{\partial\theta}(\theta_0).\frac{\partial ln(p_\theta(y|z))}{\partial\theta'}(\theta_0) p_{\theta_0}(y|z)]$$

$$I = -[\frac{\partial^2}{\partial\theta\partial\theta'}\sum_y p_\theta(y|z)](\theta_0) + E(\frac{\partial ln(p_\theta(y|z))}{\partial\theta}(\theta_0).\frac{\partial ln(p_\theta(y|z))}{\partial\theta'}(\theta_0)|Z)$$

Where the first term is 0 because, for all θ, $\sum_y p_\theta(y|z) = 1$.

Descriptive statistics (gcdpf.dta):

Var	Mean	Sd	Min	Max
y	102492.898	66874.492	20968.5	376429
yprob	0.6	0.491	0	1
ytob	111012.836	60336.461	73000	376429
k	5.810	3.389	2	10.433
l	3.908	1.619	1.998	6.032
e	5.221	3.752	0.527	17.039
ly$= ln(Y)$	11.364	0.581	9.951	12.838
lytob	11.517	0.413	11.198	12.838
lk	1.554	0.677	0.693	2.345
ll	1.267	0.452	0.692	1.797
le	1.399	0.741	-0.64	2.835
# Obs.	160	160	160	160

a) Plain OLS, dependent variable, $ln(Y)$:

Var	$\hat{\beta}^1_{ols}$	Se
lk	0.253	(0.058)
ll	0.491	(0.087)
le	0.121	(0.053)
Intercept	10.178	(0.156)
R^2	0.289	
# Obs.	160	

b) Probit and Logit model ($1_{Y>73000}$): $\hat{\beta}_l \simeq 1.6\hat{\beta}_p$ Amemiya, 1981.

Var	Probit	Se	Logit	Se	$1.6\hat{\beta}_p$
lk	0.703	(0.162)	1.177	(0.278)	1.125
ll	0.908	(0.242)	1.519	(0.415)	1.453
le	0.149	(0.147)	0.262	(0.245)	0.239
Intercept	-2.149	(0.456)	-3.618	(0.809)	-3.438
LL	-89.181		-89.153		
# Obs.	160		160		

c) Censored with respect to OLS, full sample:

Dependent variable, $ln(Y_{tob}) = max(ln(Y), ln(73000))$

Var	$\hat{\beta}_{ols}$	Se	$\hat{\beta}_{tobit}$	Se
lk	0.147	(0.043)	0.290	(0.069)
ll	0.331	(0.064)	0.533	(0.103)
le	0.078	(0.039)	0.109	(0.060)
Intercept	10.759	(0.115)	10.058	(0.200)
σ			0.515	(0.040)
R^2	0.364			
LL			-114.579	
# Obs.	160		160	
# $Y \leq 73000$	64		64	

$ln(Y_i) = y_i^* = x_i'\beta + \varepsilon_i$
But we observe: $y_i = y_{tob,i} = max(ln(73000), y_i^*)$ and x_i.
We can go back to the lecture note case (p13), using,
$y_i - ln(73000) = max(0, y_i^* - ln(73000))$

d) Truncated with respect to OLS, sample $Y > 73000$:

Var	$\hat{\beta}_{ols}$	Se	$\hat{\beta}_{trunc}$	Se
lk	0.065	(0.065)	0.152	(0.158)
ll	0.331	(0.096)	0.914	(0.328)
le	0.076	(0.052)	0.190	(0.131)
Intercept	11.047	(0.188)	9.483	(0.783)
σ			0.578	(0.098)
R^2	0.387			
LL			-22.782	
# Obs.	96		96	

$ln(Y_i) = y_i^* = x_i'\beta + \varepsilon_i$
But we observe: (y_i^*, x_i) if and only if $y_i^* > ln(73000)$.
To use the lecture note results (p13), we can focus on: $y_i^* - ln(73000)$.
$\hat{\beta}_{ols}$ is biased toward 0 with respect to the regression of y on x in the full sample, $\hat{\beta}^1_{ols}$. $\hat{\beta}_{trunc}$ seems to over correct this, the magnitudes of the point estimates are larger than those in $\hat{\beta}^1_{ols}$ (except for the intercept).

a) 2-step estimation of the censored model:

Var	$\hat{\beta}_{hl}$	Se
Intercept	-2.624	(1.993)
lk	0.534	(0.384)
ll	0.939	(0.486)
le	0.184	(0.092)
γ	1.239	(1.090)
R^2	0.543	
# Obs.	160	

$$q_i = y_i - \overbrace{ln(73000)}^{c} = x_i'\beta\Phi(\widehat{\frac{x_i'\beta - c}{\sigma}}) + \gamma\phi(\widehat{\frac{x_i'\beta - c}{\sigma}}) + \zeta_i$$

Rk: OLS standard errors' estimates are incorrect. The 2^{nd} stage OLS are not corrected for the 1^{st} estimation that adds a complicated estimation error: $-x_i'\beta[\Phi(\widehat{\frac{x_i'\beta - c}{\sigma}}) - \Phi(\frac{x_i'\beta - c}{\sigma})] - \gamma[\phi(\widehat{\frac{x_i'\beta - c}{\sigma}}) - \phi(\frac{x_i'\beta - c}{\sigma})]$, to the 2^{nd} stage disturbance term.

b) 2-step estimation of the truncated model:

var	$\hat{\beta}_{hl}$	Se
lk	0.608	(0.281)
ll	1.026	(0.364)
le	0.176	(0.072)
γ	1.454	(0.734)
Intercept	-2.975	(1.437)
R^2	0.186	
# Obs.	96	

$$q_i = y_i - c = x_i'\beta + \gamma\phi\widehat{\left(\frac{x_i'\beta - c}{\sigma}\right)}.\Phi\widehat{\left(\frac{x_i'\beta - c}{\sigma}\right)}^{-1} + \zeta_i$$

Rk: OLS standard errors' estimates are incorrect. The 2^{nd} stage OLS are not corrected for the 1^{st} estimation that adds an estimation error to the 2^{nd} stage disturbance term.

c) 2-step estimation and MLE:

- For both cases, $\hat{\beta}_{hl}$ is larger than $\hat{\beta}_{mle}$ of the corresponding model.
- $\hat{\beta}_{mle}$ are closer to our benchmark $\hat{\beta}^1_{ols}$ on the full sample using the true "y^*".
- 2-steps estimators are supposed to be less efficient than MLE. Here their standard errors are in general larger than the corresponding ones for $\hat{\beta}_{mle}$. However, to compare them directly we should correct for the 1^{st} stage estimation errors.
- 2-step estimators have the advantage of being more easy to compute as they only require to find the probit point estimates and to apply OLS. They may be useful as a starting point if the corresponding MLE model fails to converge.

a) NLLS, probit case

	NLS		Probit	
Var	$\hat{\beta}_{nls}$	Se	$\hat{\beta}_{p}$	Se
Intercept	-2.241	(0.525)	-2.149	(0.456)
ll	0.921	(0.256)	0.908	(0.242)
lk	0.730	(0.172)	0.703	(0.162)
le	0.167	(0.144)	0.149	(0.147)
R^2	0.687			
LL			-89.181	
# Obs.	160		160	

- The point estimates are close (using `nl in STATA`).
- The ses of the NLLS point estimates are slightly larger.
- The MLE is the optimally weighted NLLS estimator in this case, so we know that the standard NLLS estimator will be inefficient.

b) NLLS, censored case

	NLS		TOBIT (MLE)	
Var	$\hat{\beta}_{nls}$	Se	$\hat{\beta}_{mle}$	Se
Intercept	-6.845	(44.323)	-1.140	(0.200)
ll	1.889	(8.794)	0.533	(0.103)
lk	0.833	(3.831)	0.290	(0.069)
le	0.453	(2.080)	0.109	(0.060)
σ	2.706	(16.515)	0.515	(0.040)
R^2	0.543			
LL			-114.579	
# Obs.	160		160	

- $q_i = y_i - c = x_i'\beta\Phi(\frac{x_i'\beta}{\sigma}) + \sigma\phi(\frac{x_i'\beta}{\sigma}) + \zeta_i$
- The point estimates are not close (using `nl in STATA` and $\sigma^{(0)} = 0.5$). The NLLS point estimates are very different from the benchmark $\hat{\beta}^1_{ols}$.
- The ses of the NLLS estimates are huge.

c) NLLS, truncated case

	NLS		TRUNCATED(MLE)	
Var	$\hat{\beta}_{nls}$	Se	$\hat{\beta}_{mle}$	Se
Intercept	-197.157	(8648.687)	-1.715	(0.098)
ll	52.018	(2256.676)	0.914	(0.328)
lk	14.563	(629.776)	0.152	(0.158)
le	13.548	(585.779)	0.190	(0.131)
σ	6.222	(137.275)	0.578	(0.098)
R^2	0.690			
LL			-22.782	
# Obs.	96		96	

- $q_i = y_i - c = x_i'\beta + \sigma\phi(\frac{x_i'\beta}{\sigma}).\Phi(\frac{x_i'\beta}{\sigma})^{-1} + \zeta_i$
- The point estimates are not close (using `nl in STATA` and $\sigma^{(0)} = 0.5$). The NLLS point estimates are very different from the benchmark $\hat{\beta}^1_{ols}$.
- The ses of the NLLS estimates are huge. Convergence is slow.