ISSN 2079-7141

UAEU- FBE-Working Paper Series

Title: Quasi-Likelihood Approach to Modeling Rate and Over-dispersion in the Analysis of Unemployment

Author(s): Ibrahim M. Abdalla Al-faki

Department: Statistics

No. 2012-04

Series Founding and Acting Editor: Prof. Dr. Abdulnasser Hatemi-J

Quasi-Likelihood Approach to Modeling Rate and Over-dispersion in the Analysis of Unemployment

Ibrahim M. Abdalla Al-faki
Associate Professor, Statistics
Faculty of Business and Economics
United Arab Emirates University
i.abdalla@uaeu.ac.ae

January 2011

Abstract

Setting a probability model that details data generating mechanism is an essential prerequisite for model estimation using the maximum likelihood methodology. Cases arise where there is insufficient information about the data to fully specify a probability model. Processes might generate continuous values between 0 and 1. In such circumstances neither a binary nor a binomial processes seem adequate in describing the data and providing appropriate variance functions. The former cannot take on values except for 0 and 1 and the latter requires a binomial denominator to form a proportion. In this study alternative variance functions and the quasi-likelihood approach (assuming data distribution unknown) are engaged to estimate unemployment counts and rates. The proposed models allow for over-dispersion in the data and accommodate effects of covariates. Model performance and fit are compared to alternatives using both real and simulated data.

Keywords: Quasi-likelihood; Over-dispersion; Poisson; Binomial; Unemployment.

1 Introduction

The focus of this research is on studying mechanisms that generate unemployment counts with emphasis on the UAE. Different processes such as the Poisson process or Bernoulli trials with binary outcomes will be discussed. Accounting for over-dispersion in the data as well as estimating effects of potential covariates will lead to an adjusted unemployment rates. Model development using quasi-likelihood methods allows modeling of the data without explicit specification of an underlying likelihood or log-likelihood function. Rather, using various diagnostics measures, one begins with mean and variance functions, which are not restricted to the collection of functions defined by single-parameter exponential family members, and abstract back to an implied log-likelihood function. The implied log-likelihood is then used to estimate parameters of the model. The real challenge in this scenario is abstraction of an appropriate new variance function that provides a better fit compared to incumbent functions.

Unemployment rates are usually estimated by dividing unemployment counts by the size of the economically active population or the work force without adjustment for covariates effects. Accurate estimates of unemployment counts and rate essentially reflect the real state of the economy. Under-or over-estimation of the rate give false impression of the economy's performance and might lead to economic strains and accumulation of social burdens and complications. High unemployment rates show a lack in the growth of the economy and vice versa. Also, high levels of unemployment result in a decrease in general consumption (people

have less money to spend as they are searching for jobs) and this will contribute to slow business growth.

The unemployment issue in the UAE and the other GCC countries is generally linked with fluctuation in oil prices. Recent drops in oil prices have lead to a decline in the share of non-UAE nationals in the labor force and an increase in the numbers of nationals looking for jobs, placing considerable burden on both the society and the economy (International Monetary Fund Report [1]; Albuainain [2]). According to the UAE Ministry of Planning mid-year estimates, the overall (unadjusted) crude unemployment rate in the country stands at 3.0 percent of the total estimated labor force by the end of 2004. The rate is higher for nationals (11.4 percent) compared to non-nationals (2.1 percent), and for females (19.7 percent) compared to males (8.2 percent). These sharp differences across nationality and gender boundaries are elucidated mainly by the increase in the numbers of national graduates particularly females, the quality of education which does not meet the labor market demand, local traditions and social norms which do not accept certain jobs, and the work environment at the private sector with respect to low pay, working hours and close performance measurements which do not attract the locals, (Albuainain [2]).

The motivation of this research is triggered by the need to investigate the mechanisms that generate unemployment counts/rates in the United Arab Emirates (UAE), taking into account the influence of different factors and covariates. The study seeks to identify models that allow for over-dispersion in the data and demonstrate better performance, with less restrictions on data and model assump-

tions. The rest of the paper is structured as follows. Section 2 outlines modeling strategy and the data used to estimate unemployment rate and to empirically demonstrate models' performance. Section 3 explains in details two possible alternatives to commonly used models, namely; the Quasi-likelihood and the extended Quasi-likelihood models. Empirical results are presented in Section 4. Section 5 displays some concluding remarks.

2 Methodology

A sensible assumption to start with is that unemployment counts are generated by a Poisson process. This is a commonly accepted assumption for count data, with a wide support and coverage in the literature. In many situations, such data are sparse in nature and inhibited by over or under-dispersion, with the consequences of underestimating the standard error and type I error. This might produce misleading inference about model parameters. One solution to the problem might be achieved by replacing the Poisson by the Negative binomial distribution. In this study, however, a further alternative scenario is presented; viz. the Quasi-likelihood approach (detailed in the following subsections), which allows for over-dispersion and modeling of the data without explicit specification of the distribution and the underlying likelihood or log-likelihood function. The data used in model fit and performance in this exercise is based on 52159 cases of unemployed individuals officially registered by the Ministry of Economy in the UAE in 2005. Gender, education and age are included as factor covariates and the size

of the economically active population is used as an offset. Bootstrap simulation samples are generated to further demonstrate models' fit and performance.

2.1 Quasi-Likelihood

Let $\mathbf{Y}$ be a vector of independent responses with mean vector μ and covariance matrix $\phi V(\mu)$. The mean μ is assumed a function of covariates, $\mathbf{x}$, and regression parameters, β. The covariates are expressed into the regression function by writing $\mu(\beta)$.

ϕ is typically unknown, and need to be estimated, and $V(\mu)$ is a matrix of known functions. The assumption of independence of $\mathbf{Y}$ implies that $V(\mu)$ must be diagonal.

$$V(\mu)=\text{diag}(V_1(\mu),\ldots,V_n(\mu))$$

It is further assumed that $V_i(\mu)$ only depends on the i^{th} component of μ. The $V_1(\mu),\ldots,V_n(\mu)$ may be taken to be identical, although they could have different arguments.

Given the above conditions considering a single component Y or y of $\mathbf{Y}$ the quasi-likelihood function is defined by Wedderburn [3] by

$$U = u(\mu|Y) = \frac{Y-\mu}{\phi V(\mu)} \tag{1}$$

There are several common properties for the function U in (1) with the log-likelihood derivative or the score function. In particular

$$\begin{aligned} E(U) &= 0 \\ Var(U) &= 1/(\phi V(\mu)) \\ -E(\tfrac{\partial U}{\partial \mu}) &= 1/(\phi V(\mu)) \end{aligned} \tag{2}$$

These properties contain most of the first order first-order asymptotic theory concerned with the likelihood. The function,

$$Q(\mu|y) = \int_y^\mu \mu(y|y)dt = \int_y^\mu \frac{y-t}{\phi V(t)}dt, \tag{3}$$

therefore behaves like a log-likelihood function. It is referred to as a *quasi-likelihood* or as a *log quasi-likelihood* for μ based on the data y. By independence of of the components of Y, the quasi-likelihood for the complete data is given by summing individual contributions

$$Q(\mu|\mathbf{y}) = \sum Q(\mu_i|y_i), \tag{4}$$

which depends multiplicatively on ϕ. Therefore, ϕ does not affect the *MLEs* of $\mu(\beta)$. The quasi-deviance function, which measures the discrepancy between the observation and its expected value, corresponding to a single observation is given as

$$D(y|\mu) = -2\phi Q(\mu|y) = 2\int_\mu^y \frac{y-t}{V(t)}dt. \tag{5}$$

The total deviance, D(**y**—μ), is the sum of the individual components, and only

depends on $\mathbf{y}$ and μ, but not ϕ.

2.2 Estimation

Differentiation of the function $Q(\mu|y)$ generates the quasi-likelihood estimating equations for the parameters β. Thus the quasi-likelihood score function is expressed as

$$U(\beta) = D^T V^{-1}(Y-\mu)/\phi = 0, \tag{6}$$

where D is a $n \times p$ matrix with elements $\partial\mu_i/\partial\beta_r$, the derivatives of $\mu(\beta)$ with respect to the parameters.

The covariance matrix of $U(\beta)$ is also the negative expected value of $\partial U(\beta)/\partial\beta$, and is

$$i_\beta = D^T V^{-1} D/\phi. \tag{7}$$

Under the usual limiting conditions on eigenvalues of i_β, the asymptotic variance-covariance matrix of $\hat{\beta}$ is

$$Var(\hat{\beta}) = i_\beta^{-1} = \phi(D^T V^{-1} D)^{-1}. \tag{8}$$

That is, i_β plays the same role as the Fisher's information for ordinary likelihood functions. The Fisher's score method can be employed to obtain the parameter estimate $\hat{\beta}$,

$$\hat{\beta}_{n+1} = \hat{\beta}_n + (\hat{D}_n^T \hat{V}_n^{-1} \hat{D}_n)^{-1} \hat{D}_n^T \hat{V}_n^{-1} (y - \hat{\mu}_n), n = 0, 1, \ldots, \tag{9}$$

where $\hat{D}_n = [D]_{\beta=\hat{\beta}_n}$, $\hat{V}_n = [V]_{\beta=\hat{\beta}_n}$, $\hat{\mu}_n = [\mu]_{\beta=\hat{\beta}_n}$

In the absence of an ML estimation for ϕ, the method of moments estimate can be utilized. Based on the residual vector $Y - \hat{\mu}$, the following statistic provide an estimate for ϕ,

$$\tilde{\phi} = \frac{1}{n-p} \sum_i \frac{(Y_i - \hat{\mu}_i)^2}{V_i(\hat{\mu}_i)} = \frac{\chi^2}{n-p}, \tag{10}$$

where χ^2 is the generalized Pearson statistic.

2.3 Extended Quasi-likelihood

Likelihood ratio and score tests are appropriate tools for testing hypotheses regarding nested subsets of covariates in the linear predictor and assessing hypothesized link functions in the generalized linear models with distributions in exponential family. The quasi-likelihood form described above allows similar application of these methods. However, it is not suitable for the comparison of different variance functions, this is because the properties of the quasi-likelihood that make it comparable to a likelihood refer only to derivatives of β and not ϕ. In order to assess and compare various variance functions together with the comparison of all other components of the generalized linear model, namely; the linear predictor and the link function as well as opening the possibility of modeling the dispersion parameter ϕ as a function of covariates, the definition of the quasi-likelihood given in

(3) need to be extended to terms for the variance.

Therefore, define a function $Q^+(\mu,\phi|Y)$ for a single component Y with mean μ and variance $\phi V(\mu)$ which is the same as $Q(\mu|y)$, but also has the properties of a log-likelihood with respect to derivatives of ϕ, (Neleder and Pregibon [4]).

$$Q^+(\mu,\phi|y) = -\frac{1}{2}log\{2\pi\phi V(y)\} - \frac{1}{2}D(y|\mu)/\phi, \tag{11}$$

Where $D(y|\mu)$ is the deviance as given in (6) and ϕ is the dispersion parameter. Similar to Q, Q^+ does not assume a full distribution of the response but only needs specification of the first two moments. Being a linear function of Q, maximizing Q^+ results in the same estimates of β as those given by Q. The estimate of the dispersion parameter ϕ obtained from maximizing Q^+ is the mean deviance, $\hat{\phi} = D(y|\hat{\mu})/n$. When Q^+ coincides with the normal and inverse Guassian distributions, $\hat{\phi}$ is the maximum likelihood estimate of ϕ. For an exponential family distribution with a given variance function, the quasi-likelihood is equivalent to likelihood proper (Morris; 1982). This is true for the normal and inverse Gaussian distributions. For the gamma distribution Q^+ differs from the log-likelihood by a factor depending only on ϕ. For the Poisson, binomial and negative binomial distributions, Q^+ is obtainable from the respective log-likelihood function by replacing any factorial $k!$ by Stirling's approximation

$$k! = (2\pi k)^{\frac{1}{2}} k^k e^{-k}. \tag{12}$$

2.4 Modeling the variance as a function of the mean

The original quasi-likelihood model, Q, of Wedderburn requires the knowledge of the variance function, $V(\mu)$, up to a multiplicative constant ϕ. This requirement can be relaxed for Q^+ by embedding the variance function in a family indexed by an unknown parameter θ (Nelder and Pregibon, 1987), so that $var(y) = \phi V_\theta(\mu)$. A useful family is obtained by considering powers of μ

$$V_\theta(\mu) = \mu^\theta, \tag{13}$$

where most common values of θ are $0, 1, 2, 3$ which correspond to variance functions associated with normal, Poisson, gamma and inverse Gayssian distributions respectively.

$$V(\mu) = \mu^k (1-\mu)^l, \tag{14}$$

corresponds to the binomial distribution when $k = l = 1$ and to the famous Wedderburn square variance function when $k = l = 2$.

3 Empirical Results

Utilizing two R software functions, glm and glm.nb and the package EQL, the standard Poisson, negative binomial and the extended quasi-likelihood models are fitted, respectively, to the UAE 2005 unemployment data with gender (females as reference group), education level (Secondary or lower, Post-secondary,

College/University, and Postgraduate education level (reference)), and age (15 to 24, 25 to 34, 35 to 49, 50 to 59 and 60+ (reference)) as factor covariates. The EQL contains functions for the computation of the EQL for a given family of variance functions, in particular, extended binomial variance family, Power-logit, $V(\mu) = \mu^k(1-\mu)^l$, and the power variance family, Power-log, $V_\theta(\mu) = \mu^\theta$. The respective, estimated values for the parameters k and l are 2.6 and 2.35. The estimated value of the parameter θ for the power variance function is 1.88, see profile EQL plot in Figure 1.

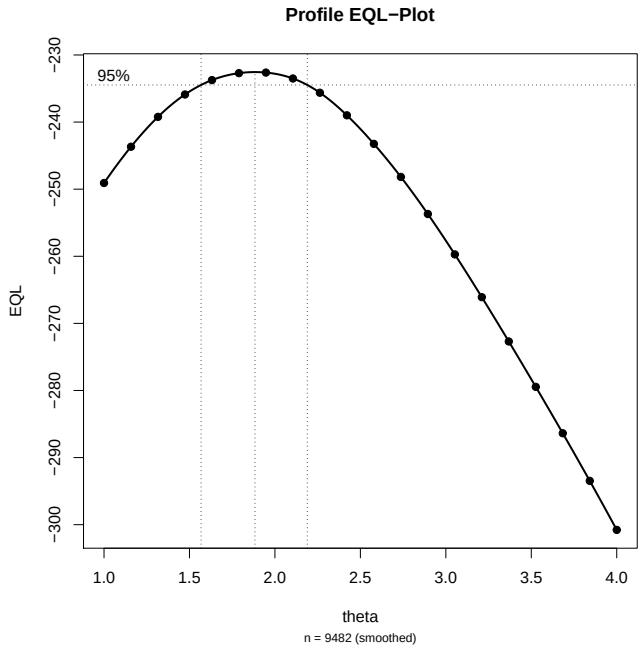

Figure 1: Estimation of the Structural Parameter, θ.

As depicted in Table 1, it is clear that the Poisson distribution assumption for the data is not a suitable choice. The Poisson (Ps) fit is highly over-dispersed, esti-

mated dispersion parameter is 82.57, see Table 1. Implying that the significance of some of the parameters, particularly the education parameters, are over-estimated when adjusting for unemployment rate in the UAE. The apparent high Poisson over-dispersion is clearly reversed when the data is fitted using an extended Quasi-likelihood model with a power variance function (Power-log). The model is highly under-dispersed, implying that the significance of some of the factors used to adjust for unemployment rate is highly under-estimated, Table 1. Both the standard binomial model and the extended Quasi-likelihood model with the power logit variance function (Power-logit) seem to compete on the same ground. Both are producing an acceptable, close to 1, dispersion parameter estimates. The residual plots presented in Figure 2 demonstrate the performance of the four models. The Power-logit model seems to produce the best fit compared to the other three models. Results and parameter estimates, generally, indicate that unemployment rate is significantly higher for females compared to males. The rate decreases with the increase in the level of education and that they are higher among younger populations compared to the older populations.

	Ps	Std	NB	Std	Power-log	Std	Power-logit	Std
(Intercept)	-3.021^*	0.008	-2.721^*	0.150	-2.833^*	0.168	-2.431^*	0.236
Gender2	0.982^*	0.009	1.366^*	0.104	1.356^*	0.124	1.510^*	0.120
Educate2	0.189^*	0.020	-0.281	0.148	-0.161	0.172	-0.331^*	0.156
Educate3	0.618^*	0.011	0.008	0.141	0.111	0.165	-0.193	0.174
Educate4	0.423^*	0.034	-0.098	0.148	-0.035	0.177	-0.286	0.157
Age2	-1.429^*	0.010	-1.446^*	0.156	-1.430^*	0.181	-1.614^*	0.249
Age3	-2.180^*	0.014	-2.269^*	0.156	-2.233^*	0.186	-2.487^*	0.234
Age4	-2.101^*	0.028	-2.285^*	0.162	-2.228^*	0.197	-2.553^*	0.233
Age5	-1.484^*	0.054	-1.718^*	0.180	-1.630^*	0.207	-2.025^*	0.241
Dispersion	82.57		1.33		0.30		1.34	

Table 1: Parameter estimation: Poisson(Ps), Negative binomial (NB), Power-log($\theta = 1.88$), and Power-logit ($k = 2.6, l = 2.35$)

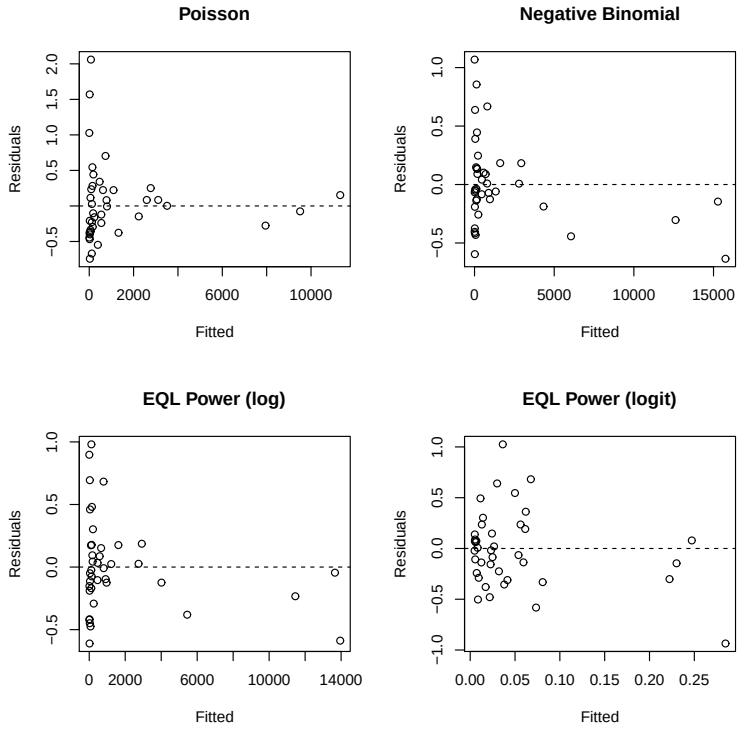

Figure 2: Residual Plots

Bootstrap simulations with 1000 replications are generated to compare the performance of the proposed models. The standard Poisson and the negative binomial models are fitted using the R functions glm and glm.nb. The extended quasi-likelihood models are also estimated using the R software, employing the EQL function. All replications are conducted treating model factors (predictors) as fixed and model response as random. Table 2 displays the Bootstrap estimates, bias and related standard errors (Std). Apparently, the power-logit variance family class has outperformed other variance families, generally producing the lowest

bias, implying that treating the data as binomial like distribution is the appropriate choice.

	Ps	NB	Power(log)	Power(logit)
Intercept	-3.0210	-2.7208	-2.8323	-2.4308
Bias	-0.0336	-0.0277	0.3635	0.0061
Std	0.2564	0.2438	0.6186	0.3157
Gender	0.9820	1.3657	1.3564	1.5097
Bias	0.1076	-0.0176	-0.0069	-0.0010
Std	0.2328	0.1195	0.2062	0.1615
Educate2	0.1894	-0.2810	-0.1614	-0.3311
Bias	-0.0573	0.0314	-0.1975	0.0200
Std	0.2164	0.2109	0.4921	0.2596
Educate3	0.6183	0.0079	0.1104	-0.1927
Bias	-0.0646	0.0370	-0.2545	0.0363
Std	0.2273	0.2289	0.5539	0.2957
Educate4	0.4233	-0.0976	-0.0353	-0.2864
Bias	-0.0676	0.0257	-0.2783	0.0048
Std	0.2081	0.2227	0.5595	0.3013
Age2	-1.4286	-1.4458	-1.4305	-1.6137
Bias	0.0337	0.0152	-0.0314	-0.0044
Std	0.2192	0.1854	0.2386	0.2396
Age3	-2.1801	-2.2694	-2.2335	-2.4867
Bias	0.0245	0.0147	-0.0727	-0.0094
Std	0.2522	0.1737	0.2609	0.2225
Age4	-2.1006	-2.2845	-2.2286	-2.5529
Bias	-0.0020	-0.0018	-0.1768	-0.0268
Std	-0.3535	0.2314	0.3601	0.2404
Age5	-1.4842	-1.7178	-1.6303	-2.0251
Bias	-0.0567	-0.0159	-0.2637	-0.0372
Std	0.4584	0.3036	0.4932	0.3326

Table 2: Bootstrap Simulation Results, $R = 1000$

Based on the Power-logit model fits, see Table 3, it seems that the highest unemployment rate, 28.5%, is reported among young (15-24 years) females with secondary or lower education level. Young females college or university graduates have also scored high unemployment rate, 24.7%, followed by young female postgraduates, 23.0% and then young females with post secondary education 22.2%. Male unemployment rate as reported in the table are much lower, ranging from a minimum of 1.0% to a maximum of 8.0% across all age and education levels, see Table 3.

Gender	Education Level	Age	Crude rate	Ps	NB	Power-log	Power-logit
Male	Secondary or lower	15 - 24	0.056	0.049	0.066	0.059	0.081
		25 - 34	0.011	0.012	0.016	0.014	0.017
		35 - 49	0.006	0.006	0.007	0.006	0.007
		50 - 59	0.007	0.006	0.007	0.006	0.007
		60+	0.017	0.011	0.012	0.012	0.011
	Some post secondary	15 - 24	0.052	0.059	0.050	0.050	0.059
		25 - 34	0.011	0.014	0.012	0.012	0.012
		35 - 49	0.006	0.007	0.005	0.005	0.005
		50 - 59	0.005	0.007	0.005	0.005	0.005
		60+	0.008	0.013	0.009	0.010	0.008
	College/University	15 - 24	0.111	0.090	0.066	0.066	0.068
		25 - 34	0.018	0.022	0.016	0.016	0.014
		35 - 49	0.006	0.010	0.007	0.007	0.006
		50 - 59	0.005	0.011	0.007	0.007	0.006
		60+	0.007	0.021	0.012	0.013	0.009
	Postgraduate	15 - 24	0.083	0.074	0.060	0.057	0.062
		25 - 34	0.016	0.018	0.014	0.014	0.013
		35 - 49	0.006	0.008	0.006	0.006	0.005
		50 - 59	0.006	0.009	0.006	0.006	0.005
		60+	0.004	0.017	0.011	0.011	0.009
Female	Secondary or lower	15 - 24	0.094	0.130	0.258	0.228	0.285
		25 - 34	0.034	0.031	0.061	0.055	0.073
		35 - 49	0.025	0.015	0.027	0.024	0.032
		50 - 59	0.049	0.016	0.026	0.025	0.030
		60+	0.076	0.030	0.046	0.045	0.050
	Some post secondary	15 - 24	0.170	0.157	0.195	0.194	0.222
		25 - 34	0.051	0.038	0.046	0.047	0.054
		35 - 49	0.023	0.018	0.020	0.021	0.023
		50 - 59	0.012	0.019	0.020	0.021	0.022
		60+	0.072	0.036	0.035	0.038	0.036
	College/University	15 - 24	0.262	0.242	0.260	0.255	0.247
		25 - 34	0.072	0.058	0.061	0.061	0.061
		35 - 49	0.027	0.027	0.027	0.027	0.027
		50 - 59	0.023	0.030	0.026	0.027	0.025
		60+	0.029	0.055	0.047	0.050	0.042
	Postgraduate	15 - 24	0.204	0.199	0.234	0.221	0.230
		25 - 34	0.069	0.048	0.055	0.053	0.056
		35 - 49	0.028	0.022	0.024	0.024	0.024
		50 - 59	0.019	0.024	0.024	0.024	0.023
		60+	0.025	0.045	0.042	0.043	0.038

Table 3: UAE Crude and Adjusted Unemployment Rates Based on Different Models Fit

4 Conclusion

The study is motivated by the need to estimate unemployment rates while adjusting for the effect of important factors. The assumption that a Poisson process generates unemployment counts is faced up with over-dispersion in the data, resulting in over-estimates of the significance of model parameters and in questionable estimates of unemployment rates. Assuming no distribution for the data, just binomial like, seems to provide the right solution. An Extended Likelihood (EQL) model based on logit link and a power variance function produced a better fit and generally lower bias.

Acknowledgements

I thank the Faculty of Busness and Ecomomics at the United Arab Emirates University for supporting this research through the college *summer research grant*, 2010.

References

[1] International Monetary Fund Report (IMF). (2005). *United Arab Emirates: Selected Issues and Statistical Appendix*, IMF Country Report No. 05/268, Washington, D.C. 20431.

[2] Albuainain, R. M. (2005). *Unemployment rate in the United Arab Emirates: The case of Abu-Dhabi*, The General Woment Union, Abu-Dhabi, UAE.

[3] R. W. M. Wedderburn, R. W. M. (1974). *Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method*, Biometrika, Vol. 61, No. 3 (Dec., 1974), pp. 439-447.

[4] Nelder, J. A. and Pregibon, D. (1987). *An extended quasi-likelihood function*, Biometrika, Vol. 74, pp. 221-232.

[5] Thaler, T. (2009). *Package EQL: Extended-Quasi-Likelihood-Function (EQL)*, R software packages, Repository CRAN.