Title: Estimating the quality of life impact of diabetes related complications over time from the UKPDS

Authors: Maria Alva [1], Alastair Gray [1], Borislava Mihaylova [1], Philip Clarke[2]

Astract:

This paper explores the efficiency and consistency of different parametric and non-parametric estimators that deal with issues inherent to bounded outcome measures in a panel data context. Reliable estimates of the quality of life impact of diabetes-related complications are important for researchers conducting trial-based and model-based evaluations of the cost-effectiveness of interventions. We estimate the immediate and long term impact of ten diabetes-related complications on EQ-5D derived utilities, controlling for age, sex and diabetes duration. Previous estimates based on a single cross-sectional study of patients enrolled in the UK Prospective Diabetes Study (UKPDS) have been widely used. Here we report updated results drawing on the UKPDS Post Trial Monitoring Study, when the EQ-5D was administered a maximum of seven times over the 1996-2008 period. This allows us to greatly extend follow-up, examine a larger number and wider range of complications, and compare cross-sectional results with those from repeated measures of quality of life over time. The specific problems we consider are: heteroskedasticity; time invariant unobserved patient characteristics correlated with the presence of complications; and probability mass accumulating at the upper boundary of the tariff (27% of observations in this population). In the absence of a gold standard or true data generating process looking both at cross sections and panels, we consider the performance of the following approaches: OLS, GLS, Tobit, CLAD, Flogit and Two-Part models.

1) Introduction

The aim of this study is to report reliable estimates of the effect on utility values of major complications of type 2 diabetes, using data from a large trial of therapies for diabetes, the United Kingdom Prospective Diabetes Study (UKPDS). The study builds on previous work by Clarke, Gray et al. (2003) who used a single cross-sectional EQ-5D survey from the same study to estimate the short- and long-term effects of 6 pre-specified clinical events (myocardial infarction (MI), ischemic heart disease (IHD), stroke, heart failure, amputation, and blindness) on utility. We now benefit from 10 additional years of follow-up and survey data, permitting us to measure the effect of these complications on quality of life in both a cross-sectional and longitudinal setting. Further the additional years of data allow us to examine the effects of other complications previously excluded due to insufficient number of events (renal failure and vitreous haemorrhage) on utilities. We also look at the effect on utility of two treatments: retinal photocoagulation and cataract extraction.

We propose a set of empirical models to estimate these effects after controlling for demographic variables and other complications. In particular, we examine the merits of different statistical models in dealing with three problems specific to quality-of-life data: heteroskedasticity; time invariant unobserved patient characteristics correlated with the

[1] Health Economic Research Centre, Department of Public Health, University of Oxford, Oxford, UK.
[2] Sydney School of Public Heath, University of Sydney, Australia

presence of complications; and ceiling effects, especially probability mass accumulating at the upper boundary of the utility tariff when measured by the generic instrument the EQ-5D.

It is not straightforward to come up with an unambiguous criterion to identify the correct statistical model while uninformed about the data generating process. Two considerations should be borne in mind while choosing the appropriate regression model and functional form: producing unbiased and efficient estimates. As each estimator carries a set of assumptions, and so choosing an estimate with less restrictive assumptions represents an additional desirable characteristic.

Part 2 of the paper reports details of the data set. Part 3 sets out the candidate statistical methods. Part 4 reports results and Part 5 contains a brief discussion.

2) Data

The UK Prospective Diabetes Study (UKPDS) was a randomised trial of glycaemic therapies in 5,102 patients with newly diagnosed type 2 diabetes. The trial ran for twenty years (1977 to 1997) in 23 UK clinical sites and showed that the complications of type 2 diabetes could be reduced by improving blood glucose and/or blood pressure control. From September 1997 the intervention phase of the study closed and surviving patients were treated to best practice by their GPs. At that point patients entered a post-trial monitoring study, which continued to collect information on clinical events, resource use and quality of life. Clinical results from 10 years of post study monitoring were published in 2008 (Holman, Paul et al. 2008).

A maximum of 7 EQ-5D questionnaires were administered to each patient in the UKPDS: one in 1996/97 and 6 during the post-trial monitoring study between 2002/03 and 2007/08, when each patient was given the questionnaire annually. At the final visit everyone was administered a questionnaire (#7), irrespective of time since previous questionnaire.

A total of 4287 UKPDS patients were alive at the time of the first administered questionnaire. The response rate of fully completed questionnaires varied from 68% to 74%. We excluded from this analysis all questionnaires administered less than 4 months apart, corresponding to 233 observations. A total of 1425 patients died between 1997 and 2008; 1,046 of whom had answered at least one questionnaire.

Table 1 provides a cross-tabulation of the questionnaire number completed by each participant by calendar year. For example, most respondents filled in their first questionnaire between 1996 and 1997, but 165 eligible patients responded for the first time only in subsequent rounds.

Table 1: Questionnaire number per calendar date

	Questionnaire number							
Year	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
1996	1,919	0	0	0	0	0	0	1,919
1997	1,272	0	0	0	0	0	0	1,272
2002	1	49	0	0	0	0	0	50
2003	131	1,419	35	0	0	0	0	1,585
2004	24	330	1,231	50	0	0	0	1,635
2005	4	90	300	1,039	38	0	0	1,471
2006	4	35	111	283	910	36	0	1,379
2007	1	17	38	97	237	595	9	994
2008	0	15	31	58	139	392	526	1,161
Total	3,356	1,955	1,746	1,527	1,324	1,023	535	11,466

In total we have 11,466 observations from 3356 patients. The average number of questionnaires completed was 3.4, the median was 3 and the maximum was 7. 42% (1,401/3,356) of respondents only answered one questionnaire (Table 2).

Table 2: Panel structure

	Number of non-duplicate responses per patient						
number of questionnaires	1	2	3	4	5	6	7
number of patients	1401	209	219	203	301	488	535

Table 3 reports the frequency distribution of time between completing successive questionnaires for those patients completing more than one questionnaire in the 2002-2008 period. The overall time difference (including the first questionnaire) varies from a month to 12 years, with mean equal to approximately 1 year and 9 months and median equal to one year.

Table 3: Frequency distribution of number of days between consecutive questionnaires

Mean (SD)	826 (857)
Variance	734788
Skewness	1.457
Kurtosis	3.482
Min / max	30 / 4382
10% percentile	365
25% percentile	365
Median	365
75% percentile	731
90% percentile	2425
Number of Observations	8692

It is evident, therefore, that we have an unbalanced panel and that questionnaires are not equally spaced through time.

Table 4 reports the proportion of patients at different levels of each EQ-5D question by each questionnaire number and overall. The table also shows the mean time between each questionnaire number. The proportion of patients reporting no problems fell markedly between questionnaire 1 and 2, which for most patients corresponded to 1996/97 and 2002/03 respectively. Changes across questionnaires 2 to 7 were less pronounced.

Table 4: Proportion of levels 1, 2 and 3 by dimension and questionnaire number

EQ-5D DIMENSION	level:	Questionnaires							
		1	**2**	**3**	**4**	**5**	**6**	**7**	**TOTAL**
MOBILITY	1	51.8%	47.3%	47.9%	48.5%	52.9%	53.0%	53.3%	50.3%
	2	47.6%	51.7%	51.5%	50.6%	46.9%	47.0%	46.7%	49.1%
	3	0.5%	1.0%	0.6%	0.9%	0.2%	0	0	0.5%
SELF-CARE	1	82.9%	79.3%	79.3%	81.1%	83.8%	84.6%	84.7%	81.9%
	2	15.7%	19.0%	18.7%	17.3%	15.0%	15.2%	15.1%	16.8%
	3	1.3%	1.7%	1.9%	1.6%	1.2%	0.3%	0.2%	1.4%
USUAL ACTIVITIES	1	59.7%	55.4%	54.9%	53.6%	58.2%	59.5%	57.9%	57.2%
	2	34.0%	37.5%	38.5%	40.7%	37.0%	35.9%	37.8%	36.9%
	3	6.3%	7.1%	6.6%	5.6%	4.8%	4.6%	4.3%	6.0%
PAIN / DISCOMFORT	1	41.2%	34.3%	34.7%	35.7%	37.2%	38.1%	38.9%	37.4%
	2	49.6%	54.7%	54.5%	53.7%	54.5%	52.8%	54.0%	52.8%
	3	9.2%	11.0%	10.9%	10.6%	8.3%	9.1%	7.1%	9.8%
ANXIETY/ DEPRESSION	1	64.7%	65.2%	63.3%	65.9%	67.0%	70.2%	67.7%	65.6%
	2	31.6%	31.8%	33.3%	31.2%	30.9%	27.8%	29.7%	31.3%
	3	3.7%	3.0%	3.4%	2.9%	2.1%	2.1%	2.6%	3.1%
Mean years between (SD):			6.34 (1.68)	1.15 (0.52)	1.12 (0.43)	1.12 (0.39)	1.19 (0.43)	1.01 (0.08)	

For utility valuations of EQ-5D health states we used the time trade off values derived from the most widely known UK population study (Dolan 1997). Table 5 shows the mean (SD) utility levels for each questionnaire number and overall, the corresponding Visual Analogue Scale results, and the proportion of patients at a tariff state of full health for each questionnaire number.

Table 5: Mean utility and Visual Analogue Scale results by questionnaire number

	Questionnaires							
	1	2	3	4	5	6	7	TOTAL
EQ-5D utility:								
Mean	0.759	0.666	0.664	0.660	0.663	0.666	0.686	0.693
SD	0.275	0.317	0.309	0.310	0.302	0.300	0.302	0.302
Percent at utility=1	35.6%	24.5%	22.6%	22.5%	21.7%	22.2%	25.6%	26.7%
VAS:								
Mean	0.732	0.700	0.692	0.695	0.694	0.691	0.708	0.707
SD	0.192	0.194	0.200	0.199	0.197	0.203	0.201	0.197
Percent with VAS of 100	6.2%	2.9%	2.5%	1.9%	1.8%	1.6%	2.5%	3.4%

A central objective of this study is to estimate the relationship between quality of life and a set of diabetes-related complications. Table 6 lists the 10 clinical events of interest, and their

rates in the study population. Note that an individual patient may experience more than one event in his or her clinical history.

Table 6: Rate of clinical events across groups for the 1996-2008 period

Events	Event rate for all patients (N=3356)	Event rate for patients reporting tariff=1during study (N=599)	Event rate for patients reporting tariff≠1 at least once during study (N=2757)
Myocardial Infarction	0.339	0.145	0.381
Ischemic heart disease	0.368	0.139	0.417
Stroke	0.136	0.033	0.158
Amputation	0.069	0	0.084
Blindness in 1 eye	0.180	0.060	0.206
Renal failure	0.018	0.002	0.022
Heart failure	0.101	0.043	0.113
Cataract extraction	0.803	0.275	0.918
Retinal photocoagulation	0.576	0.302	0.635
Vitreous haemorrhage	0.116	0.063	0.127

We investigated whether there were differences between respondents who answered one EQ-5D survey during the period of follow-up and those answering multiple questionnaires. Table 7 indicates that there are significant differences between these two groups: those answering multiple surveys tended to be older (68 vs. 63 years), had longer duration of diabetes, and had lower utility values. This is important, as any differences might provide a justification for including singletons in the regression analysis which would in turn restrict the use of fixed effects estimation methods even when these may be deemed appropriate for other reasons (e.g. to avoid omitted variables bias).

Table 7: Difference between singletons and multiple responders at first response (1996/7)

	Tariff	Current Age	Duration	Male (%)
Total (N=3191)				
Mean (SD)	0.701 (0.270)	66.068 (9.209)	14.378 (4.310)	0.590 (0.492)
Patients answering more than 1 questionnaire (N=1805)				
Mean (SD)	0.678 (0.252)	68.094 (8.410)	17.170 (2.921)	0.588 (0.492)
Patients answering only 1 questionnaire (N=1386)				
Mean (SD)	0.730 (0.288)	63.430 (9.533)	10.741 (2.876)	0.594 (0.491)
difference (SD)	0.052 (0.01)**	-4.66 (0.32)**	-6.43 (0.10)**	0.01 (0.02)

*p< 0.05 **p< 0.01

3) Statistical Issues in the Modelling of Utility Data over time

For simplicity, in our subsequent analysis we shall assume an additive specification for the independent variables in all models.

We want to estimate the following model:

$$tariff=f(xb)+u$$

in order to obtain the mean value of the tariff conditional on the covariates used, i.e. E(tariff|x)=f(xb). Where f(.) denotes a given functional form, xb is a matrix of coefficients and explanatory variables, and *u* denotes the error term vector. Without loss of generality, we avoid the use of individual and time subscripts in the above equation, to facilitate switching from pooled cross-section to panel analysis to test whether or not decomposing total variation within and between individuals is necessary. Clearly, the true functional form of E(tariff|x) is not known, but for the purposes of computing E(tariff|x), we focus here on linear index models only.

Rather than assume a once and for all time effect, we investigate the EQ-5D tariffs within 12 months of each type of clinical event and more than 12 months following each type of event. The first set of responses should capture any acute consequences of the complication, and the second should capture any long-term effects (Clarke, Gray et al. 2002).

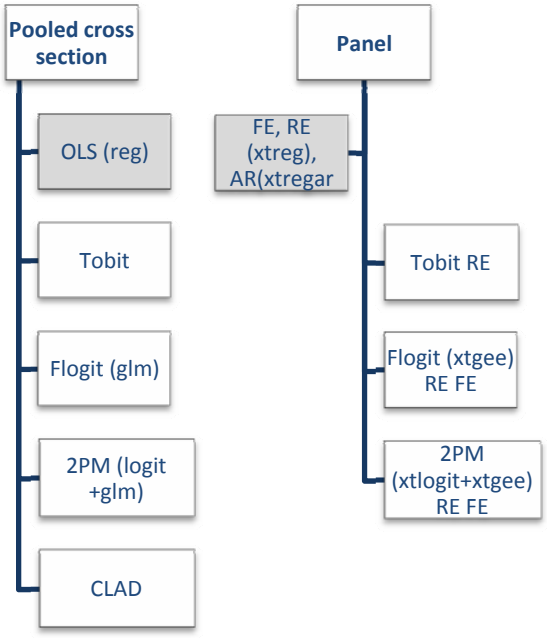
We shall test a variety of methods (Figure 1) while being explicit about their underlying assumptions. We seek to model the impact of diabetes complication on the actual tariff. Our ideal model should accommodate the following three data characteristics:

1. Probability mass accumulating at the upper limit
2. Heteroskedasticity

And allow us to examine:

3. The presence of other and possible unobserved time invariant characteristic(s), that cause(s) both complications and utility loss, formally, the situation in which $E(u_i\ x) \neq 0$

Figure 1: Heuristic conceptualization of the modelling strategy adopted



The non trivial probability mass accumulating at the upper limit of the outcome measure renders OLS inappropriate or of limited value because, firstly, it implies constant partial effects so that the probability of a tariff value exactly equal to one is very small relative to the observed distribution, and secondly because it does not guaranteed the predicted value will fall within the observed boundaries: i.e. the predicted tariff values may be greater than 1. This is also a problem with feasible generalized least squares (GLS) procedures like the RE or within regression estimators, like FE, albeit not an issue in our case as our estimations feature predominantly dichotomous explanatory variables. Models that account for ceiling effects, such as Tobit, may fail to address the problem of heteroskedasticity, i.e. variance of the error is not constant but depends on one or more of the independent variables. All models have been tested for $E(u_i\ x) \neq 0$ by exploiting the panel data structure available.

We estimate and test models with both pooled cross sectional and panel specification. The models in the shaded boxes do not account for the probability mass at 1, but all other models attempt to address this issue. Finally we include CLAD, which is known to be robust to heteroskedasticity and non-normality, but unlike all other models reported in this study, reports effects on the median of the distribution rather than the mean (Powell 1984).

Pooled OLS, Panel and GLS

Using pooled cross sections might disregard much useful information contained across questionnaires. If individual specific unobservables are correlated with the explanatory variables (i.e. omitted variables bias), the beta coefficients will be inconsistent. Moreover, if we define the error term as $=u_i+v_{it,}$ because are the composite error terms in each time

period, they will be serially correlated across time. In fact under the random effect specification $Corr(\varepsilon_{it}\varepsilon_{is}) = \frac{\sigma_u^2}{\sigma_v^2+\sigma_u^2}, t \neq s$ .Where $\sigma_v^2 = Var(v_{it})$ and $\sigma_u^2 = Var(u_i)$. This positive serial correlation in the error term can be substantial: because the usual pooled OLS standard errors ignore this correlation, they will be incorrect, as will be the usual test statistics. To eliminate serial correlation in the errors the following transformation is required (Wooldridge 2002):

$$y_{it} - \lambda\bar{y}_i = \beta_0(1-\lambda) + \beta_1(x_{it} - \lambda\bar{x}_i) + (v_{it} - \lambda\bar{v}_i)$$

where $\lambda = 1 - \left[\frac{\sigma_v^2}{\sigma_v^2+T\sigma_u^2}\right]^{1/2}$ is between zero and one. The over bar denotes the time averages. The parameter λ is never known in practice but can be estimated through feasible GLS. Pooled OLS is obtained when $\lambda = 0$, and fixed effects is obtained when $\lambda = 1$. The random effects specification subtracts a fraction of the time average, where the fraction depends on σ_u^2 and σ_v^2 and the number of time periods. When λ is close to zero, the RE estimates will be close to the OLS estimates. This would be the case when the unobserved effect u_i is relatively unimportant. If, however, σ_u^2 is larger relative to σ_v^2 , $\hat{\lambda}$ will be closer to unity. While correlation between u_i and one or more explanatory variables causes inconsistency in the RE estimation, the correlation is attenuated by the factor 1- λ.

When using fixed effects it is not possible to estimate coefficients for variables that do not vary within individuals (gender). One can perform different analyses for males and females, which would be equivalent to running a model with interaction effects. More importantly however, while one may be addressing bias in the estimates in a fixed effects model, the respondents with only one observation do not contribute to the calculation of the coefficients. This is a severe trade-off, as respondents who answered just one questionnaire represent more than 40% of our sample.

The prebuilt Stata program does not allow the user to compare FE and RE models that have been adjusted for intergroup correlation through clustering, i.e. models with robust variance estimates (the Hausman test assumes that the error term is homoskedastic which is not the case with a limited depended variable). To overcome this difficulty we use Wooldridge's (2005) test, an alternative to the Hausman test that is robust to heteroskedasticity and at the same time provides consistent coefficients if the strict exogeneity assumption does not hold (i.e $u_i \perp x_{it}$ is rejected). This is carried out through a regression on de-meaned time-varying regressors: $tariff_{it} = \beta' x_{it} + \delta\,\bar{x}_i + \varepsilon_{it}$

Tobit

$$tariff_{it}^{*} = \beta' x_{it} + \varepsilon_{it}$$

note that the dependent variable is now $tariff_i^*$, not $tariff_i$.

$$tariff_{it} = \begin{cases} a \; if \; tariff_{it}^{*} < a \\ tariff_i^{*} \; if \; a \leq tariff_{it}^{*} \leq 1 \\ 1 \; if \; tariff_{it}^{*} > 1 \end{cases}$$

where $tariff_{it}^{*}$ is assumed to be a latent measure of quality of life, ε_i is normally distributed with constant variance and t_{it} is the utility as measured by the tariff score; a is equal to the lowest value (-.594) but in our case it is not binding. Estimates of Tobit models are notoriously sensitive (non- robust) to departures from baseline assumptions (normality and homoskedasticity) (Wooldridge (2002)).

Our interest is in reporting the marginal effect on the actual tariff rather than on the latent variable. These can be obtained by multiplying β by the probability of a non-limit observation (Greene 1999).

Similarly to the pooled OLS model, unobserved characteristics that are related to the tariffs as well as the observed covariates might lead to biased estimates. A panel data approach allows these unobservables to be decomposed into a time invariant and time varying component $\varepsilon_{it} = u_i + v_{it}$. A likelihood-ratio test comparing the panel tobit against pooled tobit model is performed showing support for $\sigma_u \neq 0$ (Prob>chibar^2=0.000; chibar^2=3240.82).

Unfortunately there is no command for a conditional fixed-effects model, as there does not exist a sufficient statistic allowing the fixed effects to be conditioned out of the likelihood (Stata Manual). **xttobit** fits only random-effects tobit models.

FLogit

The Fractional Logit model (Flogit) has been proposed by Papke and Wooldridge (1996) as an alternative to OLS and Tobit.

Flogit combines features of logit transformation and binomial regression approaches, i.e. it considers the tariff a count rather than a continuous dependent variable. Levy et al. (2008) use Flogit to measure the impact on utility for different frequencies of hypoglycaemia episodes, as their elicitation mechanism never leads to negative utilites.

Because we have tariff<0, in order to apply Flogit, we work with the transformed variables: $q = \frac{tariff - a}{1 - a}$ so that $q \epsilon [0,1]$. Predicted values for q can then be retransformed to produce predicted values for the underlying untransformed variable t as tariff=q*(1-a) +a.

For a panel specification of Flogit we follow (Papke and Wooldridge 2008) in the use of generalized estimating equation (GEE) approach. The Flogit model ensures consistency in prediction (i.e. all fractions sum to one). Also, the logit transformation allows a dimishing marginal effect of any explanatory variable on the conditional expectation of q_{it} due to the fact that the first derivative of $E(q_{it} \mid x_{it})$ with respect to x_{it} is close to zero as x_{it} approaches infinity.

Two Part Model:

In the two-part models the first part addresses the entire sample and the second a subset, so that the distribution of utilities in the total population is a mixture distribution (Dow and Norton 2003).

$$\Pr[tariff = 1|x] = \Phi(\beta_i x_{it}, u_{it})$$

$$E[d|d > 0, x] = \beta_i x_{it} + v_{it}$$

$d = 1 - tariff$. The reason for translating the tariff into a strictly positive domain is in order to be able to model the second equation using glm (generalized linear model) with a gamma family, which means the transformed tariff is bounded from below, but not above.

When there is a nontrivial and possibly analytically interesting (conditional) probability mass at a given extreme point, an attractive method is the two-part model (Manning et al. 1987 & Mullahy 1998), which is commonly used in the analysis of health care expenditure with data that has a nontrivial proportion of zero outcomes and a positively skewed distribution of the positive outcomes. A two-part modelling process would have in the first part a binary logit or probit model to describe the probability that an individual achieves a certain utility value. We use a logit model for the first part as it requires less stringent conditions on the error term than the probit model.

We apply the following decomposition to E(d|x):

$$\mathrm{E}(d|x) = \Pr(t = 1) \times 0 + (1 - \Pr(t = 1)) \times E[d|d > 0, x]$$

This carries a stronger assumption than Heckman's however, i.e. independence of the error terms. If any factors that affect both the selection and substantive equations are omitted from the model, these factors will enter both error terms and induce correlation among them. If omitted variables are likely to be a problem then the estimates will be biased, with the size of the bias depending on the correlation between the errors in the two models. Unfortunately, the independence of errors cannot be tested directly. We can however test the performance of the cross-sectional 2PM against a panel specification.

CLAD

The censored least absolute deviations estimator (CLAD) is a generalization of the least absolute deviations (LAD) estimator which uses Powell's (1984) estimator.

$$\sum_{i=1}^{n} |y_{it} - \min(x_{it}\beta)|$$

The CLAD estimator has been shown to perform well when the distributional assumptions of the tobit model are violated (Deaton 1998). Although CLAD is less efficient than the Tobit model, it is robust to changes in the distribution of the error term.

Statistical considerations alone however should not dictate the choice of the appropriate model. Theoretically we want to be able to aggregate the overall effect associated with a particular event across the whole population. So we care about the mean regardless of the degree of skewness. On these grounds, the CLAD regressor may not be appropriate as it gains the desirable property of consistency through the use of median regression that does not account for the strength of all individuals(Clarke, Gray et al. 2002). Errors are bootstrapped in a two-stage simple random draw where the patient is the primary sampling unit.

Unfortunately in our sample the CLAD estimates are unstable and convergence is obtained sporadically. The estimation technique in Stata proposed by Dean Jolliffe[3] and colleagues for the CLAD estimator is (Buchinsky 1994) iterative linear programming algorithm (ILPA). With ILPA, comvergence occurs when there are no predicted values exceeding the boundaries of the upper and lowering censoring value in two consecutive iterations.

4) Results

Effect of Clinical Events on Tariffs

Table 8 provides the parameter estimates for the cross-sectional models and Table 9 those for the panel models. Coefficients in the tobit models are marginal effects on the actual tariff. The Flogit coefficients denote the rate of change in the log odds for changes in the covariates. Herein, we present the marginal effect evaluated at the predicted mean $\frac{\partial E(y|x)}{\partial x}|_{x=\bar{x}}$. In the Flogit and 2PM the dependent variable is not the same as in the other 3 models reported in Table 8. in the former case, tariffs have been shrunk so as to fit the [0,1] domain compatible with a fractional model estimator, while in the case of the 2PM, the dependent variable was transformed to 1 minus the tariff so as to allow for positive values only, i.e. from 0 to 1.594.

[3] http://users.starpower.net/djolliffe/stb/clad.pdf

Throughout all estimators, consistently, the occurrence of events has a negative impact on utility, with the exception of renal failure and vitreous haemorrhage, neither of which are statistically significant in models explaining the mean (i.e. all except for CLAD). Moreover all effects are statistically significant, with the exception of retinal photocoagulation and cataract extraction which are procedures rather than events.

A likelihood ratio test was used to examine whether there was a significant difference between the acute (events within a year prior to the survey) and the long term coefficients. The test statistics are presented in Table 10. With the exception of renal failure and amputations, events occurring within the 12 months preceding a questionnaire and events occurring more than one year previously did not have a significantly different impact on the tariff values (at the 95% level).

Table 8: Cross sectional Regression Analysis –Dependent variable: EQ-5D tariff values

Cross Sections	Pooled OLS		Tobit (Marginal Effects)		Flogit (Marginal Effects) dependent variable=q		2PM (part1 Pr tariff=1)		2PM part2 dependent variable=1-tariff if tariff<1 (Marginal Effects)		CLAD	
Variables	β_{OLS}	Robust SE	β_{Tobit}	Robust SE	β_{Flogit}	Robust SE	β_{1PM}	Robust SE	β_{2PM}	Robust SE	β_{clad}	Robust SE
Constant	0.839**	(0.036)					0.388*	-0.161			0.949**	(0.012)
Age (current)	-0.001	(0.001)	-0.001**	(0.000)	-0.001**	(0.000)	-0.014**	-0.003	-0.001	(0.000)	-0.002**	(0.000)
Duration of Diabetes	-0.006**	(0.001)	-0.005**	(0.000)	-0.004**	(0.000)	-0.036**	-0.005	0.005**	(0.001)	-0.003**	(0.000)
Male	0.080**	(0.010)	0.060**	(0.004)	0.051**	(0.004)	0.558**	-0.046	-0.051**	(0.006)	0.060**	(0.003)
Previous events												
Myocardial Infarction												
previous year	-0.088*	(0.035)	-0.062*	(0.026)	-0.054**	(0.020)	-0.438	-0.295	0.068*	(0.033)	-0.024	(0.021)
prior to previous year	-0.027	(0.018)	-0.023**	(0.007)	-0.017**	(0.006)	-0.332**	-0.087	0.007	(0.011)	-0.032**	(0.006)
Ischemic Heart Disease												
previous year	-0.101**	(0.032)	-0.081**	(0.025)	-0.067**	(0.021)	-0.845**	-0.296	0.056	(0.033)	-0.041**	(0.016)
prior to previous year	-0.071**	(0.017)	-0.054**	(0.007)	-0.044**	(0.006)	-0.574**	-0.086	0.038**	(0.010)	-0.066**	(0.005)
Stroke												
previous year	-0.225**	(0.054)	-0.162**	(0.041)	-0.135**	(0.033)	-1.099*	-0.533	0.174**	(0.052)	-0.245**	(0.027)
prior to previous year	-0.180**	(0.031)	-0.128**	(0.014)	-0.110**	(0.011)	-0.857**	-0.165	0.147**	(0.018)	-0.109**	(0.009)
Heart Failure												
previous year	-0.085	(0.050)	-0.066	(0.036)	-0.055	(0.032)	-0.59	-0.422	0.070	(0.050)	-0.112**	(0.028)
prior to previous year	-0.173**	(0.034)	-0.122**	(0.015)	-0.102**	(0.012)	-0.885**	-0.201	0.138**	(0.018)	-0.106**	(0.010)
Renal Failure												
previous year	0.495**	(0.050)	0.351**	(0.002)	0.188**	(0.002)	omitted		omitted		omitted	
prior to previous year	-0.158**	(0.052)	-0.036**	(0.010)	-0.094**	(0.021)			0.061	(0.034)	-0.094**	(0.019)
Amputation												
previous year	-0.320**	(0.063)	-0.235**	(0.049)	-0.189**	(0.040)	-2.085*	-1.021	0.256**	(0.061)	-0.553**	(0.027)
prior to previous year	-0.153**	(0.042)	-0.119**	(0.023)	-0.093**	(0.017)	-1.831**	-0.422	0.077**	(0.026)	-0.089**	(0.015)
Blindness in 1 eye												
previous year	-0.050	(0.045)	-0.035	(0.034)	-0.030	(0.026)	-0.325	-0.427	0.035	(0.044)	-0.052*	(0.023)
prior to previous year	-0.049*	(0.024)	-0.124**	(0.033)	-0.030**	(0.008)	-0.365**	-0.13	0.034*	(0.014)	-0.045**	(0.008)
Cataract Extraction												
previous year	-0.030	(0.021)	-0.018	(0.013)	-0.018	(0.011)	-0.047	-0.15	0.033	(0.019)	-0.039**	(0.011)
prior to previous year	-0.015	(0.016)	-0.014*	(0.006)	-0.009	(0.005)	-0.300**	-0.076	-0.004	(0.008)	-0.007	(0.005)
Retinal Photocoagulation												
previous year	-0.047	(0.033)	-0.030	(0.018)	-0.031	(0.018)	-0.04	-0.196	0.071*	(0.032)	-0.015	(0.015)
prior to previous year	-0.008	(0.015)	-0.008	(0.006)	-0.005	(0.005)	-0.178*	-0.07	-0.007	(0.008)	-0.031**	(0.005)
Vitreous Haemorrhage												
previous year	0.013	(0.043)	0.010	(0.028)	0.008	(0.023)	0.117	-0.36	-0.008	(0.040)	-0.051*	(0.025)
prior to previous year	0.057	(0.031)	0.037**	(0.012)	0.031**	(0.010)	0.379*	-0.168	-0.038*	(0.018)	0.025*	(0.011)

** $p<0.01$, * $p<0.05$

Table 9: Panel Regression Analysis –Dependent variable: EQ-5D tariff values

	Panel FE		Panel RE		Tobit (Marginal Effects) RE		Flogit (Marginal Effects) dependent variable=q		2PM (RE) (part1 Pr tariff=1)		2PM (RE) part2 dependent variable=1-tariff if tariff<1 (Marginal Effects)	
Variables	β_{FE}	R.SE	β_{RE}	R.SE	$\beta_{TOBIT(RE)}$	R.SE	$\beta_{Flogit(RE)}$	R.SE	$\beta_{1PM(RE)}$	R.SE	$B_{2PM(RE)}$	R.SE
Constant	0.964**	(0.011)	0.919**	(0.030)					1.488**	(0.444)		
Age (current)	omitted because of collinearity		-0.001*	(0.001)	-0.001**	(0.000)	-0.001**	(0.000)	-0.032**	(0.008)	0.000	(0.001)
Duration of Diabetes	-0.015**	-0.015**	-0.011**	(0.001)	-0.008**	(0.001)	-0.007**	(0.001)	-0.122**	(0.011)	0.010**	(0.001)
Male			0.089**	(0.009)	0.068**	(0.007)	0.057**	(0.006)	1.240**	(0.141)	-0.060**	(0.010)
Previous events												
Myocardial Infarction												
previous year	-0.060*	(0.031)	-0.078**	(0.026)	-0.048**	(0.019)	-0.048**	(0.016)	-0.456	(0.492)	0.072*	(0.028)
prior to previous year	0.012	(0.024)	-0.019	(0.015)	-0.014	(0.010)	-0.010	(0.009)	-0.621**	(0.231)	0.006	(0.015)
Ischemic Heart Disease												
previous year	-0.051	(0.032)	-0.071*	(0.029)	-0.057**	(0.017)	-0.046*	(0.019)	-1.500**	(0.476)	0.030	(0.033)
prior to previous year	-0.025	(0.022)	-0.055**	(0.015)	-0.040**	(0.009)	-0.033**	(0.009)	-1.093**	(0.218)	0.028	(0.015)
Stroke												
previous year	-0.157**	(0.039)	-0.172**	(0.036)	-0.127**	(0.028)	-0.099**	(0.022)	-2.470**	(0.885)	0.109**	(0.036)
prior to previous year	-0.162**	(0.034)	-0.177**	(0.026)	-0.126**	(0.016)	-0.101**	(0.016)	-1.905**	(0.387)	0.136**	(0.029)
Heart Failure												
previous year	-0.107*	(0.053)	-0.096*	(0.044)	-0.066*	(0.027)	-0.060*	(0.027)	-0.578	(0.649)	0.090*	(0.045)
prior to previous year	-0.097**	(0.033)	-0.120**	(0.025)	-0.069**	(0.016)	-0.064**	(0.015)	-1.075*	(0.435)	0.104**	(0.026)
Renal Failure												
previous year	omitted		0.487**	(0.042)	0.350	(1.515)	0.199**	(0.003)	28.730	(17,587.136)	omitted	
prior to previous year	-0.029	(0.138)	-0.113	(0.064)	-0.074	(0.039)	-0.058	(0.040)	-20.411	(1,649.699)	0.024	(0.047)
Amputation												
previous year	-0.198**	(0.072)	-0.242**	(0.065)	-0.160**	(0.031)	-0.139**	(0.040)	-3.019*	(1.396)	0.187**	(0.062)
prior to previous year	-0.134**	(0.043)	-0.166**	(0.035)	-0.117**	(0.024)	-0.099**	(0.020)	-2.727**	(0.794)	0.124**	(0.036)
Blindness in 1 eye												
previous year	0.009	(0.054)	-0.027	(0.044)	-0.017	(0.024)	-0.016	(0.025)	-0.165	(0.682)	0.020	(0.043)
prior to previous year	0.037	(0.027)	-0.021	(0.019)	-0.013	(0.013)	-0.010	(0.011)	-0.484	(0.325)	0.016	(0.019)
Cataract Extraction												
previous year	0.009	(0.014)	0.003	(0.013)	0.005	(0.009)	0.002	(0.008)	0.376	(0.238)	0.010	(0.015)
prior to previous year	0.005	(0.013)	-0.000	(0.011)	0.002	(0.007)	0.004	(0.006)	-0.299	(0.176)	-0.007	(0.011)
Retinal Photocoagulation												
previous year	-0.064**	(0.023)	-0.058**	(0.022)	-0.038**	(0.014)	-0.041**	(0.014)	0.118	(0.323)	0.084**	(0.027)
prior to previous year	-0.038*	(0.016)	-0.029*	(0.012)	-0.020**	(0.007)	-0.019**	(0.007)	-0.346*	(0.172)	0.021	(0.012)
Vitreous Haemorrhage												
previous year	-0.032	(0.030)	-0.026	(0.031)	-0.019	(0.022)	-0.016	(0.017)	-0.332	(0.586)	0.035	(0.032)
prior to previous year	0.029	(0.029)	0.034	(0.025)	0.020	(0.014)	0.017	(0.013)	0.122	(0.393)	-0.031	(0.022)

** $p<0.01$, * $p<0.05$

Table 10: Stability of Coefficients over time (post OLS reg)

H0: Acute=Long term	Test F(1, 11442)	Prob > F
Myocardial Infarction	3.24	0.0718
Ischemic Heart Disease	0.87	0.3516
Stroke	0.89	0.3442
Heart Failure	3.38	3.38
Renal Failure	4.94	0.0262
Amputation	8.42	0.0037
Blindness in 1 eye	0	0.9817
Cataract Extraction	0.52	0.4707
Retinal Photocoagulation	2.38	0.1226
Vitreous Haemorrhage	0.81	0.3671

Model Selection Criteria and Diagnostics

There is no single empirical test that can prove that one approach is better than another. Table 1 provides means, minimum and maximum of actual and estimated tariffs conditional on the covariates, mean squared error (MSE) and mean absolute error (MAE). All predicted values have been transformed into the original scale of interest for purposes of comparison.

Table 11: Estimated predicted values compared to actual tariffs

Model	Obs	Mean	Min	Max	MSE	MAE
tariff	11466	0.693	-0.594	1		
Predicted values Cross Section Models						
OLS	11466	0.693	0.064	1	0.0839	0.2133
Tobit	11466	0.756	0.002	1	0.0887	0.2102
Flogit	11466	0.693	-0.055	1	0.0839	0.2132
2PM	11466	0.694	-0.100	1	0.0840	0.2134
Predicted values Panel Models						
FE	11466	0.693	0.234	0.915	0.0881	0.2200
RE	11466	0.675	0.091	1	0.0851	0.2198
Tobit (RE)	11466	0.734	0.044	1	0.0889	0.2147
Flogit	11466	0.673	-0.024	1	0.0855	0.2206
2PM	11466	-0.635	-0.020	1	0.0890	0.2353

OLS by construction has the lowest MSE, this feature however no longer holds when we look at rare and more extreme values, as illustrated in Table 12.

Table 12: MSE and MAE by tariff range

	MSE					
N (%)	4010 (35%)	4800 (41.9%)	961 (8.4%)	258 (2.3%)	844 (7.4%)	593 (5.2%)
ranges	**0.8<= tariff<=1**	**0.6<= tariff<0.8**	**0.4<= tariff<6**	**0.2<= tariff<0.4**	**0<= tariff<0.2**	**tariff<0**
Predicted values Cross Section Models						
OLS	0.0677	0.0086	0.0268	0.1458	0.2965	0.5659
Tobit	0.0415	0.0135	0.0488	0.1944	0.3590	0.6511
Flogit	0.0676	0.0086	0.0274	0.1459	0.2969	0.5658
2PM	0.0677	0.0085	0.0274	0.1481	0.2970	0.5655
Predicted values Panel Models						
FE	0.0726	0.0105	0.0255	0.1569	0.3105	0.5751
RE	0.0770	0.0119	0.0225	0.1376	0.2771	0.5381
Tobit (RE)	0.0521	0.0157	0.0412	0.1857	0.3341	0.6165
Flogit	0.0779	0.0125	0.0232	0.1367	0.2755	0.5367
2PM	0.1000	0.0187	0.0169	0.1142	0.2364	0.4806
	MAE					
Predicted values Cross Section Models						
OLS	0.2434	0.0733	0.1432	0.3599	0.5328	0.7386
Tobit	0.1772	0.0931	0.1945	0.4109	0.5829	0.7897
Flogit	0.2438	0.0723	0.1457	0.3594	0.5322	0.7379
2PM	0.2434	0.0728	0.1449	0.3648	0.5333	0.7381
Predicted values Panel Models						
FE	0.2479	0.0825	0.1360	0.3761	0.5465	0.7475
RE	0.2571	0.0875	0.1252	0.3457	0.5137	0.7197
Tobit (RE)	0.1958	0.1004	0.1706	0.3963	0.5603	0.7677
Flogit	0.2590	0.0881	0.1275	0.3425	0.5106	0.7181
2PM	0.2947	0.1125	0.1005	0.3096	0.4715	0.6781

While model fit is a matter of interest, the central issue for us is to be able to estimate the causal effects appropriately and to minimize bias. We start with investigating bias on the coefficient estimates, defined as the absolute value of the difference between the mean value of the coefficients of interest across 500 bootstraps (equally sized re-samples with replacement of 3356 individuals) and the value of the coefficient obtained from the original data presented in tables 8 and 9 divided by the observed coefficient.

We present the results of this exercise in Table 13 and highlight all those instances where we have achieved the minimum amount of bias for all the coefficients of interest (i.e. complications).

Table 14 provides an intuition for model selection on the different dimensions of interest: heteroskedasticity; time invariant unobserved patient's characteristics correlated with complications and utility; probability mass accumulating at the upper boundary of the tariff; and any additional restrictions assumed by the method

Table 13: Percentage Bias across models

Variables	Pooled OLS % Bias	Tobit % Bias	Flogit % Bias	2PM (part1 Pr tariff=1) % Bias	2PM part2 if tariff<1 % Bias	Panel FE % Bias	Panel RE % Bias	Tobit RE % Bias	Panel Flogit % Bias	2PM (RE) (part1 Pr tariff=1) % Bias	2PM (RE) part2 if tariff<1 % Bias
Myocardial Infarction											
previous year	2.484%	3.125%	3.388%	11.508%	0.123%	2.498%	0.708%	2.768%	0.677%	14.709%	0.277%
prior to previous year	0.768%	3.562%	4.576%	8.606%	8.580%	0.226%	2.278%	7.230%	5.752%	2.921%	8.561%
Ischemic Heart Disease											
previous year	0.787%	0.592%	0.678%	8.204%	2.119%	2.081%	2.180%	0.585%	2.991%	25.655%	7.067%
prior to previous year	0.572%	0.515%	0.504%	2.955%	0.669%	0.207%	0.113%	2.189%	0.579%	16.165%	1.797%
Stroke											
previous year	0.256%	1.300%	1.060%	9.702%	1.003%	0.939%	1.904%	1.493%	2.246%	59.883%	0.545%
prior to previous year	0.958%	1.335%	1.050%	6.067%	1.104%	0.043%	0.485%	2.581%	1.103%	27.363%	0.714%
Heart Failure											
previous year	1.872%	1.611%	3.089%	15.107%	4.029%	0.058%	0.736%	2.948%	2.609%	13.351%	2.827%
prior to previous year	0.292%	0.159%	0.062%	9.601%	0.479%	0.109%	0.059%	3.683%	0.142%	1.448%	0.361%
Renal Failure											
previous year	0.380%	0.101%	3.816%				0.503%	2.114%	0.107%	20.704%	
prior to previous year	1.020%	1.629%	2.808%		1.126%	11.192%	3.528%	2.670%	4.929%	18.407%	16.668%
Amputation											
previous year	1.547%	1.295%	1.233%	18.827%	0.668%	0.467%	0.624%	7.451%	0.585%	248.703%	0.895%
prior to previous year	2.166%	0.933%	0.766%	8.875%	0.856%	0.582%	1.032%	1.135%	0.872%	43.054%	0.369%
Blindness in 1 eye											
previous year	2.889%	2.901%	0.131%	18.566%	7.276%	11.934%	1.621%	19.845%	5.324%	5.900%	4.381%
prior to previous year	2.219%	2.093%	1.511%	5.359%	4.730%	4.225%	4.710%	0.134%	8.104%	18.055%	8.332%
Cataract Extraction											
previous year	0.203%	2.334%	0.968%	3.653%	1.740%	2.659%	14.847%	61.366%	2.564%	27.872%	6.014%
prior to previous year	2.841%	2.335%	2.415%	0.712%	9.781%	7.272%	4188.948%	34.986%	3.273%	0.582%	4.622%
Retinal Photocoagulation											
previous year	0.718%	1.511%	0.743%	4.672%	0.723%	1.803%	0.418%	0.402%	0.096%	33.919%	0.620%
prior to previous year	7.575%	10.186%	16.769%	1.337%	11.042%	0.242%	1.093%	4.353%	0.792%	35.758%	0.401%
Vitreous Haemorraghe											
previous year	18.733%	20.666%	35.692%	36.171%	19.191%	3.305%	8.392%	11.353%	11.538%	47.524%	3.849%
prior to previous year	1.273%	0.043%	1.007%	6.979%	0.586%	6.484%	4.894%	8.192%	10.013%	93.986%	0.077%

Table 14: Qualitative Outline of Model Selection Issues

Method/ Problem	Heteroskedasticity $Var(e_{it})=E(e_{it}^2)\neq\sigma^2$	$E(u_i\ x)\neq 0$	Probability mass accumulating in the upper limit	Other issues brought by the method - additional restrictions
Pooled OLS with clustered errors	☑ standard errors can be adjusted to be asymptotically robust to both heteroskedasticity and serial correlation. Alternatively feasible generalized least squares (FGLS) can be used	✘ does not account for time invariant unobservable typical to the panel setting	Not accounting for the upper and lower bounds of outcome variable could create bias	
Panel FE, RE	☑ FGLS offers more elegant ways to account not only for heteroskedasticity but also serial correlation.	☑ RE accounts for this partially while FE deals with the porblem fully. However in the case of singletons, with FE one de facto discards all those observations.	✘ Not well suited to impose a bounded effect of the covariates on tariffs	
Tobit	✘ Tobit requires normality of the error term and constant variance.	✘ It assumes strict exogeneity and if this is not satisfied it becomes an inconsistent estimator like pooled OLS.	☑ Tobit assumes that the observed utilities are censored at 1 and hence that the true utility can be greater than 1. One can correct for this easily by calculating marginal effects within the observed boundaries.	
Panel Tobit RE	☑ Clustering on the panel variable produces a consistent VCE estimator when the disturbances are not identically distributed over the panels or there is serial correlation	✘ same as above	☑ same as above	
CLAD	☑ this approach is a solution to the (cross-section) Tobit's sensitivity to heteroskedasticity.	✘	☑	The estimated coefficients correspond to median values which are of limited use in economic evaluations where the aim is to maximize the mean health effects.
Fractional Regression – Flogit (Papke & Wooldridge 2006)(Papke and Wooldridge 1996)	☑ ✘ the errors have bounded heteroskedastic structures	✘	☑ While holding the theoretical advantage property of bounding utility at 1, the coefficients are less straightforward to interpret . It requires a transformation of the dependent variable in the 0-1 interval.	A binomial stochastic structure results in too much weight being given to the tails relative to the truth. So in some cases, standard errors may be too wide.
Panel Flogit (Papke and Wooldridge 2008)	☑ We also account for an autoregressive process in order to improve the efficiency of the estimator.	☑	☑ same as above	Same as above
Two part model (logit + glm) cross section	✘	✘	☑	Requires that the error terms in the selection and outcome equations are uncorrelated.
Two part model (xtlogit + gee) panel	☑	☑ ✘	☑	Same as above

Discussion

This paper shows the importance of exploring the panel dimension of patient level datasets. We formally test pooled versus panel estimators using a likelihood ratio test and reject the hypothesis that there is no within patient correlation across questionnaires (Prob>chi^2=0.000 for all pairs of pooled and panel models). Both bias tests and the qualitative analysis of models suggest that patient specific variance is important, suggesting that the panel structure should not be ignored. For reasons of prediction, we are particularly concerned with problems of omitted variable bias and ceiling effects. For this we favour two type of models: FE and Panel Flogit. The former is linear in the coefficients, which helps considerably with the interpretation of results. The marginal effects of Flogit change with the vector of explanatory variables and at the point in which we held all other variables constant. Moreover, Flogit requires a retransformation of the dependent variable. The problem, however, with performing FE is that we discard approximately 40% of our data, creating a different sort of bias, because people who report only 1 questionnaire are different from patients answering more than one questionnaire (Table 7). In order to discriminate between RE and FE, we estimated a Hausman test. The problem with this test is that it assumes that the RE estimator is fully efficient, which is not the case under the conditions of heteroskedasticity and clustering effects. We therefore used the Wooldridge (2002) robust version of Hausman test, which allows us to estimate our models with clustering, in two different settings: with all the data available and in comparable samples (i.e. only among patients answering more than 1 questionnaire). The Wooldridge approach tests that the coefficients on the meaned data are jointly equal to zero. In both situations, we cannot reject the null hypothesis that the combined coefficients are equal to zero, as such we cannot rule out that the $corr(u_i, \bar{x}_i) \neq 0$, suggesting that a within estimator would be appropriate. The shortcoming is that in order to perform such models we have to forgo some of our data. The 2PM panel appears to be a promising specification. Further investigation is needed to establish the best possible model for the second part, whether a log link gamma family, where we use the information from the overall sample, or a Flogit or FE is most suitable. An extension of this paper would explore both dynamic panels and the issue of reverse causality.

This paper focuses on an exploration of different statistical methods for estimating the utility impact of diabetes-related complications, using data from a large trial of therapies for diabetes. The wide choice of candidate methods reflects the advantages of having repeated measures, although the panel is not balanced. Following further exploration and discussion, we hope to be able to produce a set of recommended values for researchers who require reliable estimates of the utility impact of these complications.

References

Buchinsky, M. (1994). "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression." Econometrica **62**(2): 405-458.

Clarke, P., A. Gray, et al. (2002). "Estimating Utility Values for Health States of Type 2 Diabetic Patients Using the EQ-5D (UKPDS 62)." Med Decis Making **22**(4): 340-349.

Clarke, P., A. Gray, et al. (2003). "The impact of diabetes-related complications on healthcare costs: results from the United Kingdom Prospective Diabetes Study (UKPDS Study No. 65)." Diabetic Medicine **20**(6): 442-450.

Deaton, A. (1998). The analysis of household surveys: a microeconometric approach to development policy, The Johns Hopkins University Press.

Dolan, P. (1997). "Modeling Valuations for EuroQol Health States." Medical Care **35**(11): 1095-1108.

Dow, W. and E. Norton (2003). "Choosing Between and Interpreting the Heckit and Two-Part Models for Corner Solutions." Health Services and Outcomes Research Methodology **4**(1): 5-18.

Greene, W. (1999). "Marginal effects in the censored regression model." Economics Letters **64**(1): 43-49.

Holman, R. R., S. K. Paul, et al. (2008). "Long-Term Follow-up after Tight Control of Blood Pressure in Type 2 Diabetes." N Engl J Med **359**(15): 1565-1576.

Papke, L. E. and J. M. Wooldridge (1996). "Econometric Methods for Fractional Response Variables With an Application to 401 (K) Plan Participation Rates." Journal of Applied Econometrics **11**(6): 619-632.

Papke, L. E. and J. M. Wooldridge (2008). "Panel data methods for fractional response variables with an application to test pass rates." Journal of Econometrics **145**(1-2): 121-133.

Powell, J. L. (1984). "Least absolute deviations estimation for the censored regression model." Journal of Econometrics **25**(3): 303 - 325.

Wooldridge, J. M. (2005). "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity." Journal of Applied Econometrics **20**(1): 39-54.