1 Recovering the parameter of a link function

A simulation experiment was carried out to assess how reliably the parameter of a parametric link can be recovered with the help of `glmx`. A related question was if the use of a parametric link introduces bias in the regression weights.

1.1 Outline of the simulation experiment

Data from a binomial GLM with one continuous predictor X was generated. All models were of the form

$$t(\mathsf{E}(Y_i|X_i)) = \beta_0 + \beta_1 X_i, \quad i = 1, ..., n.$$

Here t is a link function, β_1 is the true coefficient (fixed at 1) and the intercept was fixed at 0. The link functions studied were the Gosset link based on the t-distribution, the t_α family and the second family of Aranda-Ordaz (1981) restricted to positive parameters, see Table 1. These choices of link functions were motivated by the fact that all three have unbounded range; this was necessary in the sense that `glm`'s fitting procedure relies on the ability to invert arbitrary values of the linear predictor.

Table 1: Parametric links used in simulations

Name	$\mu =$	parameter	symmetric	unbounded	sigmoid
Gosset[a]	$\psi_\nu^{-1}(\eta)$	$0 < \nu < \infty$	+	+	+
Aranda-Ordaz II	$\log([(1-\eta)^{-\phi}]/\phi)$	$-\infty < \phi < \infty$	-[b]	if $\phi \geq 0$	if $\phi > -1$
t_α[c]	$\alpha \log(\eta) - (2-\alpha)\log(1-\eta)$	$0 \leq \alpha \leq 2$	-[c]	+[c]	+[c]

[a] ψ_ν is the cumulative density function of the t-distribution with ν degrees of freedom.

[b] For $\phi = 0$ this is the complementary log-log and for $\phi = 1$ it is the logit. If $\phi < 0$ the transformation is bounded at 1.

[c] t_α = logit for $\alpha = 1$, otherwise t_α is asymmetric, unbounded and sigmoid, with the exception of the degenerate $\alpha = 0$, $\alpha = 2$.

Besides the link the following three variables were varied: the true parameter of the link (three reasonable values for each link), the sample size ($n = 500, 1000, 5000$) and the way the covariates X were generated (a medium range and a wide range). We had three hypotheses: (a) extreme values for the parameters are more difficult to recover, (b) as sample size increases recovery improves and (c) when the range of the covariates is wide the recovery of the parameters benefits, because more information on the tails of the distribution of $\mathsf{E}(Y_i|X_i)$ is available.

1.2 Results

Tables 2, 3 and 4 report the results of the simulations. Note that for convenience and numerical stability the numerical optimization of the parameters of the link functions was carried out for transformed parameters. In the case of the Gosset and Aranda-Ordaz links, the log-transformed parameters were optimized and in the case of the t_α link the logit of $\alpha/2$ was used. Both these transformation are supported by `glmx` via the `xlink` argument.

For the Gosset link very low values of ν are hard to recover. Initial simulations with values of ν below 0.5 produced even larger bias for the regression coefficients. The value of $\nu = 1$ amounts to the Cauchit link, i.e. the inverse of the cdf of the heavy-tailed Cauchy distribution. In general the cumulative density function of the t-distribution is numerically challenging for $\nu < 0.5$, leading to unstable coefficients. Notice that contrary to our hypothesis on the influence of sample size, a sample size of 5000 was no guarantee to recover ν correctly when the spread of X was low: a true ν of 7.39 is already close to a probit link, so the bias for $\log(\nu)$ of 20.70 for $X \sim \mathrm{U}(-2, 2)$ does not represent a large difference in links.

In the case of the Aranda-Ordaz II link, the value of 1 leads to the logit transformation and values beyond 1 lead to a family of asymmetric transformations similar in shape to $x \mapsto -\log(-\log(x))$. Acceptable recovery is here already possible for $n = 1000$ and a large spread of X.

The t_α family is reasonably well-behaved, though admittedly the true values for α shy away from the extreme values. Similarly to the Aranda-Ordaz II link the RMSE of $\mathsf{logit}(\alpha/2)$ decreases as sample size or the spread of X increases.

In all three cases the recovery of the regression coefficients was more reliable than the recovery of the link. In sum the sample size should be fairly large if a binomial GLM with a parametric link is used. If the sample is too small, the a parametric link might just overfit the data.

Table 2: Parameter Recovery for the Gosset link

Parameters of Simulation				Bias			RMSE		
ν	$\log(\nu)$	n	X	β_0	β_1	$\log(\nu)$	β_0	β_1	$\log(\nu)$
1.00	0.00	500	U(-2,2)	7.21	373.89	0.83	246.96	10920.87	2.49
2.72	1.00	500	U(-2,2)	-1.33	9.80	1.32	41.75	304.70	3.08
7.39	2.00	500	U(-2,2)	0.00	0.17	1.55	0.12	0.92	3.40
1.00	0.00	1000	U(-2,2)	0.00	0.12	0.41	0.09	0.49	1.58
2.72	1.00	1000	U(-2,2)	-0.00	0.05	1.30	0.07	0.27	2.96
7.39	2.00	1000	U(-2,2)	-0.00	0.05	2.90	0.06	0.19	4.88
1.00	0.00	5000	U(-2,2)	-0.00	0.00	0.24	0.03	0.16	1.31
2.72	1.00	5000	U(-2,2)	0.00	-0.12	15.22	0.02	0.19	18.08
7.39	2.00	5000	U(-2,2)	0.00	-0.03	20.70	0.02	0.10	27.92
1.00	0.00	500	U(-4,4)	0.04	0.56	0.17	1.05	11.28	1.05
2.72	1.00	500	U(-4,4)	0.01	0.08	0.43	0.14	0.34	1.60
7.39	2.00	500	U(-4,4)	-0.00	0.05	1.49	0.11	0.22	3.05
1.00	0.00	1000	U(-4,4)	-0.01	0.08	0.05	0.12	0.37	0.48
2.72	1.00	1000	U(-4,4)	-0.00	0.02	0.14	0.08	0.20	0.66
7.39	2.00	1000	U(-4,4)	-0.00	0.02	0.95	0.07	0.14	2.34
1.00	0.00	5000	U(-4,4)	0.00	0.02	0.00	0.05	0.13	0.15
2.72	1.00	5000	U(-4,4)	0.00	0.01	0.01	0.04	0.08	0.19
7.39	2.00	5000	U(-4,4)	-0.00	-0.00	0.14	0.03	0.06	0.55

References

Aranda-Ordaz, F. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, *68*, 357–363.

Table 3: Parameter Recovery for the Aranda-Ordaz II link

Parameters of Simulation				Bias			RMSE		
ϕ	log(ϕ)	n	X	β_0	β_1	log(ϕ)	β_0	β_1	log(ϕ)
0.20	-1.61	500	U(-2,2)	0.07	0.06	-1.55	0.30	0.24	3.23
1.00	0.00	500	U(-2,2)	0.80	0.45	-0.75	8.63	4.45	2.36
4.00	1.39	500	U(-2,2)	10.83	4.92	-0.32	24.30	11.13	2.92
0.20	-1.61	1000	U(-2,2)	0.03	0.03	-0.85	0.18	0.15	2.19
1.00	0.00	1000	U(-2,2)	0.12	0.08	-0.31	0.53	0.35	1.43
4.00	1.39	1000	U(-2,2)	8.17	3.70	-0.04	20.86	9.44	2.29
0.20	-1.61	5000	U(-2,2)	0.00	0.00	-0.24	0.08	0.07	0.94
1.00	0.00	5000	U(-2,2)	0.02	0.01	-0.02	0.17	0.12	0.36
4.00	1.39	5000	U(-2,2)	0.51	0.24	-0.00	4.36	1.96	0.61
0.20	-1.61	500	U(-4,4)	0.02	0.02	-1.73	0.23	0.15	3.28
1.00	0.00	500	U(-4,4)	0.21	0.10	-0.13	4.72	1.84	0.83
4.00	1.39	500	U(-4,4)	2.74	1.01	0.07	12.60	4.56	1.06
0.20	-1.61	1000	U(-4,4)	0.00	0.01	-1.25	0.16	0.11	2.77
1.00	0.00	1000	U(-4,4)	0.02	0.02	-0.05	0.26	0.15	0.41
4.00	1.39	1000	U(-4,4)	0.54	0.20	0.02	4.98	1.70	0.57
0.20	-1.61	5000	U(-4,4)	-0.00	-0.00	-0.26	0.08	0.05	0.97
1.00	0.00	5000	U(-4,4)	0.01	0.00	-0.00	0.10	0.06	0.16
4.00	1.39	5000	U(-4,4)	0.04	0.02	0.01	0.27	0.12	0.19

Table 4: Parameter Recovery for the t_α link

Parameters of Simulation				Bias			RMSE		
α	$\log(\alpha/2)$	n	X	β_0	β_1	$\log(\alpha/2)$	β_0	β_1	$\log(\alpha/2)$
1.00	-0.69	500	U(-1,1)	-0.10	-0.01	0.35	1.13	0.16	4.17
1.12	-0.58	500	U(-1,1)	0.07	0.01	0.12	1.12	0.17	4.11
1.24	-0.47	500	U(-1,1)	0.20	0.02	0.15	1.14	0.22	4.10
1.00	-0.69	1000	U(-1,1)	0.02	-0.01	-0.10	0.99	0.12	3.29
1.12	-0.58	1000	U(-1,1)	0.06	-0.00	0.09	1.00	0.14	3.27
1.24	-0.47	1000	U(-1,1)	0.16	0.01	0.02	1.01	0.16	3.09
1.00	-0.69	5000	U(-1,1)	-0.01	-0.01	0.03	0.60	0.05	1.08
1.12	-0.58	5000	U(-1,1)	0.03	-0.01	0.02	0.60	0.07	1.09
1.24	-0.47	5000	U(-1,1)	0.06	0.00	0.00	0.56	0.09	1.04
1.00	-0.69	500	U(-2,2)	0.02	-0.01	-0.03	0.64	0.10	1.23
1.12	-0.58	500	U(-2,2)	0.02	-0.00	0.00	0.66	0.11	1.20
1.24	-0.47	500	U(-2,2)	0.05	-0.00	0.00	0.64	0.12	1.00
1.00	-0.69	1000	U(-2,2)	0.01	0.00	-0.01	0.49	0.07	0.69
1.12	-0.58	1000	U(-2,2)	0.01	0.00	0.01	0.49	0.08	0.72
1.24	-0.47	1000	U(-2,2)	-0.00	-0.00	0.06	0.46	0.09	0.67
1.00	-0.69	5000	U(-2,2)	-0.00	-0.00	0.01	0.22	0.03	0.27
1.12	-0.58	5000	U(-2,2)	-0.00	-0.00	0.01	0.21	0.03	0.27
1.24	-0.47	5000	U(-2,2)	-0.00	-0.00	0.01	0.21	0.04	0.26