Bivariate Poisson models with varying offsets: An application to the paired mitochondrial DNA dataset

Pei-Fang Su[a,*], Yu-Lin Mau[a], Yan Guo[b], Chung-I Li[a], Qi Liu[b], John D. Boice[c], Yu Shyr[b]
[a]Department of Statistics, National Cheng Kung University,Tainan 70101, Taiwan
[b]Center for Quantitative Sciences, Vanderbilt University, Nashville, TN 37232, USA
[c]National Council on Radiation Protection Measurements, Bethesda, MD 20814, USA
[*]pfsu@mail.ncku.edu.tw

Supplementary

1. Simulation setting of Bivariate Poisson distribution

In this section, we report the simulation results to assess the performance of the bivariate Poisson regression model when considering varying offset terms. We first discuss the estimation of the parameters of the bivariate Poisson distribution. Then, adding interested covariates, we discuss the estimation of the coefficient for the bivariate Poisson regression model.

Consider three independent random variables, X_0, X_1, and X_2 which follow an independent Poisson distributions with the parameters $X_{i0} \sim t_{i0}\theta_0$, $X_{i1} \sim t_{i1}\theta_1$, $X_{i2} \sim t_{i2}\theta_2 > 0$. We fixed $\theta_1 = 2$, $\theta_2 = 3$ as true values, and we varied $\theta_0 = 0$, 0.6, 2.5, 5.8, which correspond to the cases of independent, low correlation , medium correlation, and high correlation, respectively. Then, $Y_1 = X_1 + X_0$ and $Y_2 = X_2 + X_0$ is a bivariate Poisson distribution. To mimic real-world conditions, the offsets t_{ij} were generated from the $U(0,1)$ distribution.

The sample size, or the number of pairs, was set to be 30, 50, 100, 150, and 200. We showed the performance of the estimated coefficients using the bivariate Poisson distribution without considering the offset (denoted as BP model) and the bivariate Poisson regression with the offset (denoted as the BPO model). In Table S1, the estimation (denoted as Esti.), bias (denoted as Bias), and mean square error (MSE) of θ_0, θ_1 and θ_2 are shown based on 2000 simulation runs.

As seen in Table S1, the estimations θ_1 and θ_2 of the BPO models are close to the pre-specified parameters $\theta_1 = 2$, $\theta_2 = 3$, respectively. As expected, without considering the offset, the parameters of the BP model result in biased estimators. Overall, the MSE of the BP model is larger than the MSE of the BPO model. As

the correlation increases, the bias and MSE increase. Conversely, as the sample size increases, the bias and MSE show a decreasing pattern. In particular, the biases and MSE of the BP model become very large as the true correlation equals 0.7. The reason is that θ_0 mainly measures the dependence between the two count variables. When the offset variable is adjusted, the BPO provides an unbiased estimation.

2. Bivariate Poisson regression model

In the second simulation, we discuss the influence on the coefficients of the bivariate Poisson regression model with and without considering the offset term. Let Y_1 and Y_2 represent the number of mutation counts, and let t_{i1} and t_{i2} $(i = 1, 2, ..., n)$ be the number of accessible base pairs of mtDNA for the ith pair of the mother and offspring. Assume that Y_1 and Y_2 are drawing from a bivariate Poisson distribution, BPO$(\mu_{i0}, \mu_{i1}, \mu_{i2})$ and that the parameters $\mu_{ij} = t_{ij}\theta_j$ are functions of one explanatory variable Z_i:

$$\begin{aligned}
\log(\mu_{i1}) &= \log(t_{i1}) + \beta_{10} + \beta_{11} Z_i, \\
\log(\mu_{i2}) &= \log(t_{i2}) + \beta_{20} + \beta_{21} Z_i, \\
\log(\mu_{i0}) &= \log(t_{i0}) + \beta_{00}.
\end{aligned}$$

We set $\beta_{10} = \beta_{20} = 0$, and the effect $\beta_{11} = 1$, $\beta_{21} = 0.5$, and $\beta_{00} = 1$. The paired count data are influenced by Z_i through the effect of β_{11} and also through the effect of β_{21}. β_{00} represents the association among paired samples. The covariate Z_i was generated from a binary distribution (chemotherapy administered/not administered) with a probability of 0.5. The sample size was set to be 30, 50, 100, 150, and 200. In the 16569-bp-long human mitochondrial genome, the number of accessible base pairs was generated based on the following three scenarios. Scenario 1 is generated from the $U(9000/16569, 1)$ distribution; scenario 2 is generated from the $U(12000/16569, 1)$ distribution, and scenario 3 is generated from the $U(15000/16569, 1)$ distribution, indicating a low-quality read, medium-quality read, and high-quality read, respectively. In Table S2, we show the parameter estimation for β_{00}, β_{10}, β_{11}, β_{20}, and β_{21} under the BP and BPO regression model, with the bootstrap standard error in parentheses. All of the testing results are based on 2000 simulation runs. We also show the average estimated correlation (denoted as *Cor*), $E(Y_1)$ and $E(Y_2)$, under different scenarios.

As seen in Table S2, all of the estimations for the BPO model are close to the pre-specified parameters under three scenarios. As the number of pairs increases, the standard error decreases. The BP model does not account for the offset term. In fact, it does not influence the estimation of the slope coefficient β_{11} and β_{21}. However, it has a large impact on the estimations of the intercept terms β_{10} and

β_{20}. Thus, the biased estimations of the intercept β_{10} and β_{20} result in an underestimated correlation, and they overestimate $E(Y_1)$ and $E(Y_2)$. In particular, for scenario 1, (i.e., the accessible DNA fraction generated from $U(9000/16569, 1)$), the estimated correlation is 0.71 for the BPO model. However, the estimated correlation is only 0.66, which is lower than the true value. Note that the higher-quality sequence (i.e., the accessible DNA fraction generated from $U(15000/16569, 1)$) provides a better estimation of the paired correlation and a more correct expectation value of the paired counts. In summary, our method can adjust for unequal fractions of accessible mitochondrial DNA across samples and provide more precise estimations.

3. Vary the percentage of accessible base pairs

Next, we vary the percentage of accessible base pairs that is generated from uniform distributions to investigate the effect on the expectation and correlation estimation. We set the regression model

$$\begin{aligned}
\log(\mu_{i1}) &= \log(t_{i1}) + \beta_1 Z_i, \\
\log(\mu_{i2}) &= \log(t_{i2}) + 0.5 Z_i, \\
\log(\mu_{i0}) &= 1,
\end{aligned}$$

where Z_i is also generated from a binary outcome with probability of 0.5. We vary β_1 from 0.2 to 1.0, and we set the fixed sample size to be 200. In Figure S1, we showed the estimated $E(Y_1)$, $E(Y_2)$ and the correlation based on the different effects of β_1. In the upper panels of Figure S1, we vary the percentage of accessible base pairs that are generated from uniform distributions with $\mathrm{U}(0.3, 1)$, $\mathrm{U}(0.5, 1)$, $\mathrm{U}(0.7, 1)$, and $\mathrm{U}(0.9, 1)$ for the BPO model, which means that the sequence quality goes from bad to good. In the lower panels of Figure S1, we also vary the percentage of accessible base pairs that were generated from uniform distributions with $\mathrm{U}(0.3, 0.4)$, $\mathrm{U}(0.5, 0.6)$, $\mathrm{U}(0.7, 0.8)$, and $\mathrm{U}(0.9, 1.0)$ for the BPO model. This setting also means that the sequence quality goes from bad to good and has a small variance. The classical bivariate Poisson regression model (denoted as the BP model) is also shown in this figure. Overall, as β_1 increases, $E(Y_1)$ increases, as well. Because β_1 has no effect on Y_2, $E(Y_2)$ maintain a constant over the entire β_1. Conversely, as β_1 increases, the correlation decreases. However, when the quality of accessible base pairs is bad (i.e., $\mathrm{U}(0.3, 1)$), without considering offset, the expectation of $E(Y_1)$ and $E(Y_2)$ is overestimated. Conversely, the estimated paired correlation is underestimated. In summary, a higher-quality sequence provides a better estimation of the paired correlation and a more correct expectation value of the paired counts. Our proposed method can adjust for the unequal lengths

of the accessible mitochondrial DNA across samples and provides a more precise estimation, especially for the situation in which the sequencing depth is not adequate or the sequencing quality is poor, with a limited fraction of mtDNA available for variant calling.

Table S1: Parameter estimations for the bivariate Poisson distributions with and without considering the offset terms. $\theta_1 = 2$, $\theta_2 = 3$, and vary $\theta_0 = 0$, 0.6, 2.5, 5.8, which correspond to $\rho = 0, 0.2, 0.5$, and 0.7.

			θ_0			θ_1			θ_2		
ρ		n	Esti.	Bias	MSE	Esti.	Bias	MSE	Esti.	Bias	MSE
0	BP	30	0.073	0.073	0.018	0.923	-1.077	1.214	1.426	-1.574	2.564
		50	0.056	0.056	0.010	0.941	-1.059	1.153	1.442	-1.558	2.478
		100	0.040	0.040	0.005	0.961	-1.039	1.095	1.461	-1.539	2.394
		150	0.032	0.032	0.003	0.968	-1.032	1.077	1.467	-1.533	2.367
		200	0.028	0.028	0.002	0.973	-1.027	1.604	1.473	-1.527	2.346
	BPO	30	0.064	0.064	0.018	1.930	-0.070	0.152	2.931	-0.069	0.217
		50	0.046	0.046	0.009	1.950	-0.050	0.089	2.953	-0.047	0.128
		100	0.003	0.003	0.004	1.973	-0.027	0.043	2.970	-0.030	0.064
		150	0.023	0.023	0.002	1.997	-0.023	0.029	2.976	-0.024	0.041
		200	0.019	0.019	0.002	1.981	-0.019	0.021	2.981	-0.019	0.032
0.2	BP	30	0.233	-0.367	0.177	1.065	-0.935	0.958	1.570	-1.430	2.161
		50	0.220	-0.380	0.173	1.082	-0.918	0.896	1.581	-1.419	2.087
		100	0.209	-0.391	0.168	1.092	-0.908	0.853	1.590	-1.410	2.025
		150	0.206	-0.394	0.166	1.095	-0.905	0.838	1.594	-1.406	2.002
		200	0.205	-0.395	0.164	1.092	-0.908	0.838	1.595	-1.405	1.993
	BPO	30	0.598	-0.002	0.105	1.998	-0.002	0.202	3.006	0.006	0.268
		50	0.605	0.005	0.061	2.000	0.000	0.116	2.993	-0.007	0.162
		100	0.603	0.003	0.020	1.998	-0.002	0.059	2.994	-0.006	0.079
		150	0.602	0.002	0.002	2.001	0.001	0.040	2.998	-0.002	0.051
		200	0.601	0.001	0.015	1.994	-0.006	0.029	2.998	-0.002	0.040
0.5	BP	30	1.070	-1.430	2.179	1.185	-0.815	0.799	1.684	-1.316	1.899
		50	1.067	-1.433	2.134	1.183	-0.817	0.748	1.681	-1.319	1.841
		100	1.057	-1.443	2.122	1.193	-0.807	0.690	1.693	-1.307	1.759
		150	1.057	-1.443	2.187	1.193	-0.807	0.677	1.692	1.308	1.743
		200	1.056	-1.444	2.106	1.191	-0.809	0.675	1.690	-1.310	1.741
	BPO	30	2.518	0.018	0.274	1.992	-0.008	0.240	0.994	-0.006	0.305
		50	2.514	0.014	0.156	1.989	-0.011	0.137	2.982	-0.018	0.177
		100	2.501	0.001	0.081	1.999	-0.001	0.068	2.998	-0.002	0.086
		150	2.503	0.003	0.052	2.000	0.000	0.045	2.998	-0.002	0.060
		200	2.497	-0.003	0.039	1.997	-0.003	0.034	2.996	-0.004	0.043
0.7	BP	30	2.699	-3.101	9.885	1.202	-0.798	0.783	1.703	-1.297	1.865
		50	2.689	-3.111	9.834	1.207	-0.793	0.716	1.708	-1.295	1.790
		100	2.684	-3.116	9.789	1.213	-0.787	0.662	1.716	-1.284	1.704
		150	2.685	-3.115	9.756	1.215	-0.785	0.645	1.714	-1.286	1.691
		200	2.684	-3.118	9.760	1.216	-0.784	0.636	1.713	-1.287	1.682
	BPO	30	5.817	0.017	0.503	1.986	-0.014	0.256	2.988	-0.012	0.328
		50	5.812	0.012	0.308	1.988	-0.012	0.151	2.987	-0.013	0.193
		100	5.799	-0.001	0.154	1.991	-0.009	0.074	2.995	-0.005	0.095
		150	5.805	0.005	0.101	1.998	-0.002	0.050	2.996	-0.004	0.063
		200	5.800	0.000	0.075	1.997	-0.003	0.037	2.994	-0.006	0.047

Table S2: Parameter estimations (with the bootstrap standard error in parentheses) for the BP and BPO models under different numbers of accessible base pairs. $\beta_{10} = \beta_{20} = 0$, the effect $\beta_{11} = 1$, $\beta_{21} = 0.5$, and $\beta_{00} = 1$.

	n	β_{00}	β_{10}	β_{11}	β_{20}	β_{21}	Cor	$E(Y_{i1})$	$E(Y_{i2})$
Scenario 1. $U(9000/16569, 1)$									
BP	30	0.994 (0.143)	-0.428 (1.132)	1.132 (1.127)	-0.398 (0.663)	0.524 (0.499)	0.661	4.586	3.846
	50	0.992 (0.107)	-0.338 (0.350)	1.065 (0.358)	-0.321 (0.326)	0.500 (0.228)	0.661	4.575	3.853
	100	0.992 (0.077)	-0.274 (0.240)	1.015 (0.242)	-0.274 (0.218)	0.492 (0.151)	0.661	4.579	3.852
	150	0.993 (0.065)	-0.263 (0.206)	1.008 (0.205)	-0.258 (0.178)	0.492 (0.126)	0.661	4.578	3.852
	200	0.994 (0.054)	-0.259 (0.171)	1.008 (0.174)	-0.248 (0.150)	0.489 (0.103)	0.661	4.577	3.851
BPO	30	1.001 (0.139)	-0.185 (1.126)	1.143 (1.123)	-0.162 (0.622)	0.531 (0.472)	0.716	4.162	3.588
	50	1.000 (0.103)	-0.099 (0.345)	1.075 (0.355)	-0.087 (0.319)	0.508 (0.226)	0.715	4.150	3.595
	100	1.000 (0.075)	-0.036 (0.234)	1.026 (0.238)	-0.043 (0.216)	0.501 (0.150)	0.716	4.154	3.594
	150	1.001 (0.063)	-0.023 (0.201)	1.018 (0.205)	-0.023 (0.178)	0.499 (0.126)	0.716	4.153	3.593
	200	1.001 (0.053)	-0.019 (0.168)	1.017 (0.171)	-0.014 (0.148)	0.496 (0.102)	0.716	4.151	3.592
Scenario 2. $U(12000/16569, 1)$									
BP	30	1.002 (0.162)	-0.349 (0.850)	1.152 (0.859)	-0.318 (0.485)	0.554 (0.345)	0.661	4.579	3.852
	50	1.000 (0.118)	-0.223 (0.352)	1.046 (0.342)	-0.227 (0.316)	0.505 (0.225)	0.661	4.573	3.847
	100	0.999 (0.081)	-0.184 (0.228)	1.024 (0.233)	-0.184 (0.213)	0.507 (0.144)	0.661	4.576	3.852
	150	1.002 (0.066)	-0.166 (0.196)	1.007 (0.192)	-0.172 (0.174)	0.506 (0.114)	0.662	4.572	3.853
	200	0.997 (0.057)	-0.164 (0.163)	1.011 (0.163)	-0.163 (0.148)	0.502 (0.101)	0.661	4.578	3.852
BPO	30	1.005 (0.156)	-0.208 (0.850)	1.155 (0.859)	-0.179 (0.480)	0.557 (0.343)	0.693	4.322	3.696
	50	0.996 (0.117)	-0.067 (0.351)	1.053 (0.341)	-0.073 (0.313)	0.510 (0.226)	0.693	4.317	3.691
	100	1.002 (0.080)	-0.044 (0.227)	1.028 (0.232)	-0.045 (0.211)	0.509 (0.143)	0.692	4.320	3.696
	150	1.005 (0.065)	-0.025 (0.194)	1.011 (0.191)	-0.032 (0.172)	0.509 (0.114)	0.693	4.316	3.696
	200	1.000 (0.057)	-0.023 (0.161)	1.014 (0.162)	-0.024 (0.147)	0.505 (0.100)	0.693	4.322	3.696
Scenario 3. $U(15000/16569, 1)$									
BP	30	1.005 (0.155)	-0.198 (1.119)	1.102 (1.108)	-0.357 (2.034)	0.636 (1.472)	0.661	4.581	3.853
	50	1.002 (0.120)	-0.125 (0.564)	1.053 (0.564)	-0.134 (0.318)	0.519 (0.222)	0.662	4.573	3.851
	100	1.004 (0.088)	-0.107 (0.228)	1.046 (0.224)	-0.104 (0.208)	0.515 (0.104)	0.662	4.582	3.851
	150	1.000 (0.066)	-0.070 (0.190)	1.014 (0.183)	-0.073 (0.166)	0.506 (0.111)	0.662	4.576	3.850
	200	1.002 (0.060)	-0.064 (0.158)	1.011 (0.156)	-0.065 (0.145)	0.504 (0.096)	0.661	4.577	3.854
BPO	30	1.005 (0.155)	-0.151 (1.123)	1.103 (1.112)	-0.309 (1.845)	0.636 (1.336)	0.672	4.493	3.799
	50	1.002 (0.120)	-0.078 (0.560)	1.053 (0.560)	-0.087 (0.318)	0.519 (0.222)	0.672	4.485	3.797
	100	1.004 (0.088)	-0.059 (0.227)	1.046 (0.224)	-0.057 (0.208)	0.515 (0.140)	0.672	4.494	3.797
	150	1.000 (0.066)	-0.002 (0.190)	1.014 (0.183)	-0.025 (0.165)	0.507 (0.111)	0.672	4.488	3.796
	200	1.003 (0.060)	-0.016 (0.158)	1.012 (0.156)	-0.018 (0.145)	0.504 (0.095)	0.672	4.488	3.800

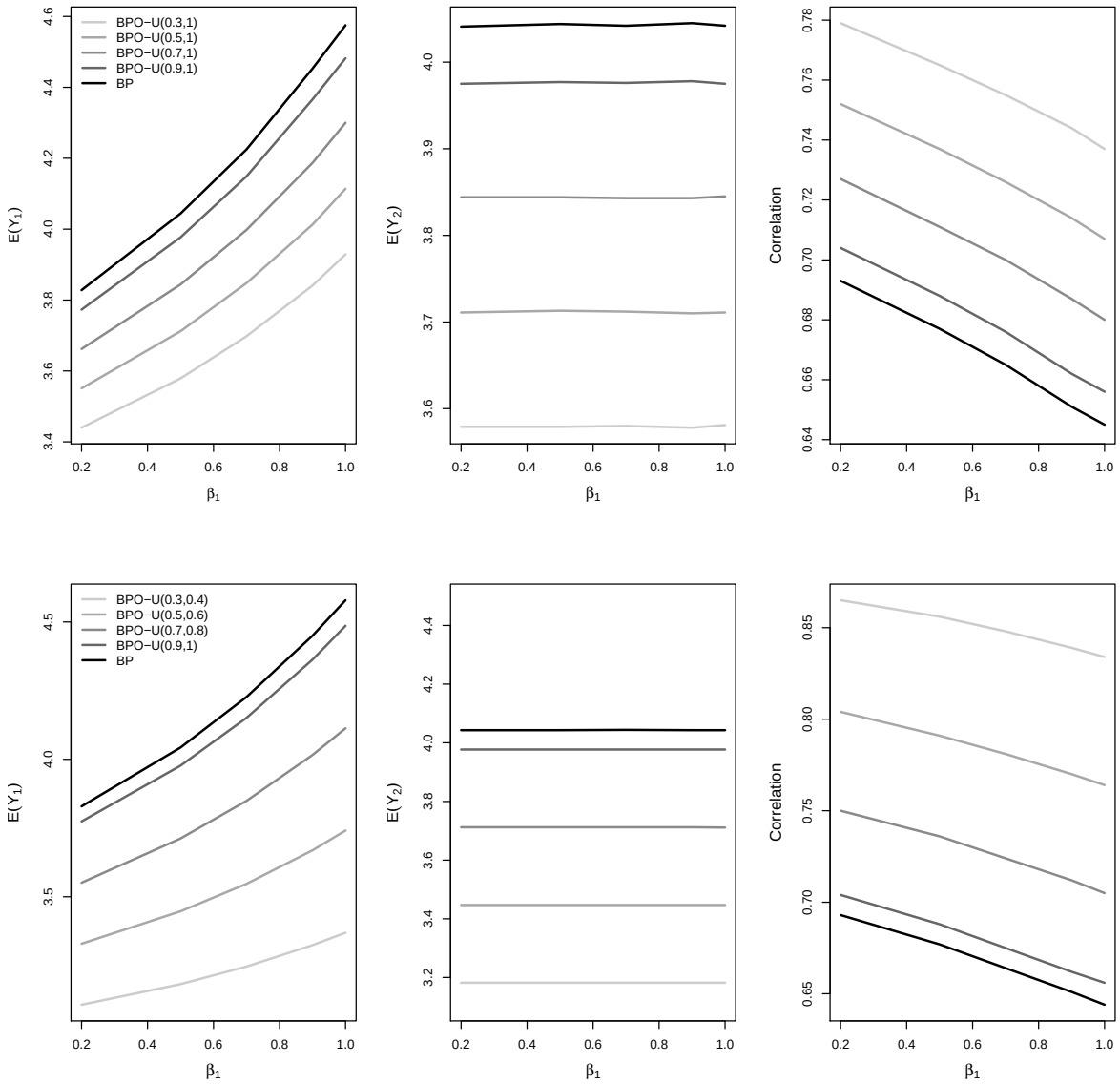

Figure S1: The estimated $E(Y_1)$, $E(Y_2)$ and the correlation based on different effects of β_1.