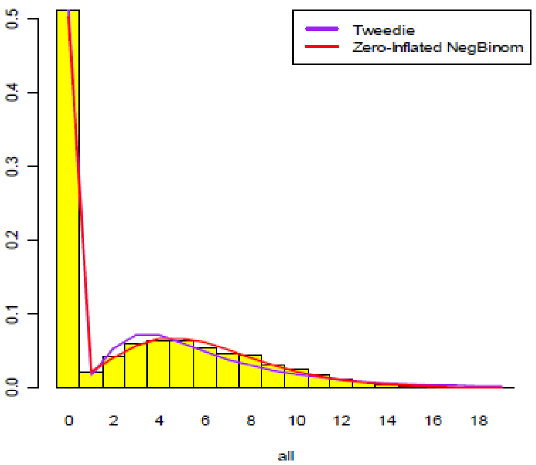
La Distribution Tweedie

En probabilité et en statistiques, les distributions Tweedie appartiennent à la classe des modèles de dispersion exponentielle, célèbres pour leur rôle dans les modèles linéaires généralisés. C'est une famille de distributions de probabilité qui comprend des distributions continues telles que la distribution Normale et Gamma, la distribution de Poisson exclusivement discrète, et la classe de distributions composées mixtes Poisson-Gamma qui ont une quantité importante de zéros.

On retrouve souvent les distributions Tweedie dans les études actuarielles, les analyses biométriques, de survie, en écologie, dans les analyses de consommation d'alcool chez les jeunes, applications médicales, en météorologie et climatologie voire dans les activités de la pêche.

STATISTICA intègre la distribution Tweedie parmi les nombreuses distributions proposées dans le module *Modèles Linéaires/Non-Linéaires Généralisés*.



D'une manière générale, lorsqu'on utilise une distribution gamma dans notre modèle final, la distribution tweedie apparaît comme le modèle théorique adéquat pour tenir compte du grand pourcentage de zéros dans la variable de perte. La famille de distributions Tweedie est une sous classe de la famille exponentielle avec une fonction de variance donnée par μ^p où μ représente la moyenne et p est supérieur ou égal à 0.

$$\mathrm{Var}[\underline{Y}] = \phi.V(\underline{\mu}) \,/\, \underline{\omega}$$

Normale: $\phi = \sigma^2$, $V(x) = 1 \Rightarrow \mathrm{Var}[\underline{Y}] = \sigma^2.\underline{1}$ p=0

Poisson: $\phi = 1$, $V(x) = x \Rightarrow \mathrm{Var}[\underline{Y}] = \underline{\mu}$ p = 1 avec $\varphi = 1$

Gamma: $\phi = k$, $V(x) = x^2 \Rightarrow \mathrm{Var}[\underline{Y}] = k\underline{\mu}^2$ p = 2

Tweedie: $\phi = k$, $V(x) = x^p \Rightarrow \mathrm{Var}[\underline{Y}] = k\underline{\mu}^p$ $1 < p < 2$

Les distributions Normale (p = 0), Poisson (p = 1 avec φ = 1), Gamma (p = 2) et Gaussienne inverse (p = 3) sont donc des cas spéciaux. Pour d'autres valeurs de puissance non

remarquables, les distributions sont toujours définies mais ne peuvent pas être écrites dans une forme finie, et par conséquent il est très difficile de les estimer. Quand 1 < p < 2, les distributions sont continues pour Y supérieur à zéro, avec une quantité positive pour Y = 0. Pour p > 2, les distributions sont continues pour Y supérieur à zéro. Le choix de p peut se faire simplement en analysant les résidus.
La variable dépendantes doit être numérique, avec des données supérieures ou égale à zéro.si une données est inférieure à zéro ou manquante, alors l'observation correspondante n'est pas utilisée dans l'analyse. La valeur fixe du paramètre de la distribution Tweedie peut être n'importe quelle valeur supérieure à 1 et inférieure à 2.

Pour illustrer la distribution Tweedie, Ouvrons le module *Modèles Linéaires/Non-Linéaires Généralisés* et le fichier de données CarInsurance.sta représentant des indemnisations suite à des sinistres automobiles. Sélectionnons la variable *Montant du Sinistre* comme variable Dépendante et les variables *Age du Conducteur*, *Catégorie de Véhicule* et *Age du Véhicule* comme Facteurs catégoriels.

Pour cet exemple, spécifions 1,15 comme paramètre d'indice.

L'estimation des paramètres du modèle donne les coefficients significatifs suivants :

Données : Montant du Sinistre - Paramètres estimés (CarInsurance)*

Montant du Sinistre - Paramètres estimés (CarInsurance.sta)
Distribution : TWEEDIE(1,15)
Fonction de Liaison : LOG

Effet	Niveau Effet	Colonne	Estimat.	Erreur Type	Wald Stat.	LC Inf. 95, %	LC Sup. 95, %	p
Ord.Orig		1	5,32800	0,037434	20258,51	5,25464	5,40137	0,000000
Age du Conducteur	17-20	2	0,11812	0,090042	1,72	-0,05836	0,29460	0,189595
Age du Conducteur	21-24	3	0,05423	0,092138	0,35	-0,12636	0,23481	0,556171
Age du Conducteur	25-29	4	0,18857	0,087806	4,61	0,01647	0,36067	0,031746
Age du Conducteur	30-34	5	-0,04599	0,095560	0,23	-0,23328	0,14131	0,630362
Age du Conducteur	35-39	6	-0,19054	0,100797	3,57	-0,38810	0,00702	0,058714
Age du Conducteur	40-49	7	-0,04795	0,095629	0,25	-0,23538	0,13948	0,616067
Age du Conducteur	50-59	8	-0,02043	0,094672	0,05	-0,20598	0,16512	0,829135
Catégorie de Véhicule	A	9	-0,14052	0,064237	4,79	-0,26642	-0,01462	0,028704
Catégorie de Véhicule	B	10	-0,08117	0,063135	1,65	-0,20491	0,04258	0,198586
Catégorie de Véhicule	C	11	-0,02136	0,062062	0,12	-0,14300	0,10028	0,730758
Age du Véhicule	0-3	12	0,37258	0,057258	42,34	0,26035	0,48480	0,000000
Age du Véhicule	4-7	13	0,26514	0,058830	20,31	0,14984	0,38045	0,000007
Age du Véhicule	8-9	14	-0,11890	0,065328	3,31	-0,24694	0,00915	0,068764
Echelle			15,72107	1,866479		12,45729	19,83995	

Le test de Vraisemblance mesure la variation de la log-vraisemblance, imputable à l'effet respectif (actuel), tous les autres effets demeurant inchangés. Pour une valeur p < 0,01, les effets Âge du Véhicule et Catégorie de Véhicule sont significatifs.

Données : Montant du Sinistre - Test Vraisemblance Type 3 (CarIns...

Montant du Sinistre - Test Vraisemblance Type 3
Distribution : TWEEDIE(1,15)
Fonction de Liaison : LOG

Effet	Degré de Liberté	Log-Vraisbnc	Chi² Square	p
Age du Conducteur	7	-732,483	9,37624	0,226760
Catégorie de Véhicule	3	-736,439	17,28775	0,000617
Age du Véhicule	3	-761,774	67,95694	0,000000

Si vous comparez l'Écart Normé à son chi-deux asymptotique à 114 degrés de liberté, la valeur p est de 0,88. Ceci indique que l'ajustement du modèle est bon.

Données : Montant du Sinistre - Stats d...

Montant du Sinistre - Stats de
Distribution : TWEEDIE(1,15)
Fonction de Liaison : LOG

Stat.	dl	Stat.	Stat/dl
Ecart	114	2068,695	18,14645
Ecart Normé	114	131,587	1,15428
Chi² Pearson	114	1781,331	15,62571
P Chi² Normée	114	113,309	0,99393
AIC		1485,591	
BIC		1528,371	
LogVraisembl.		-727,795	

Le graphique confirme que le modèle ajuste relativement bien les données ; toutefois, il met également en évidence trois points atypiques avec des valeurs très importantes qui suggèrent de les analyser de plus près afin de valider leur présence dans le modèle.