

Course website	http://www.stat.wisc.edu/courses/st572-larget/
TA	Xu He(Mario)
Email	hexu@stat.wisc.edu
Website of TA	www.stat.wisc.edu/~hexu
Office	1275F, MSC
Office hour	Th 11:00-12:00, 4:00-6:00
Phone	(608)334-9792

1 A Generalized linear regression problem

For 30 Galapagos islands, we have a count of the number of species of tortoise found on each island. We also have five geographic variables for each island. The original dataset contained several missing values which have been filled for convenience.

“Species” the number of species of tortoise found on the island

“Area” the area of the island (km²)

“Elevation” the highest elevation of the island (m)

“Nearest” the distance from the nearest island (km)

“Scruz” the distance from Santa Cruz island (km)

“Adjacent” the area of the adjacent island (square km)

- Make plots of each potential input variable versus the outcome of Species. Which variables appear to be most predictive? Are there some variables where a transformation of the predictor variable might be warranted to lessen the influence of extreme points?
- Fit a Poisson regression model to predict Species versus the outcome of Species based on the Area, Elevation, Nearest, Scruz and Adjacent. Report the coefficients of this model. Do the signs of the coefficients make sense?
- Find the mean and standard deviation for each input variable in the previous model. What change in the predicted Species is associated with a one standard deviation increase in the variable while holding other variables constant?
- Fit a Poisson regression model with the same variables used in part(b), but without the Elevation variable. Report the coefficients of this model. Which model is better?
- Predict Species with all variables be the mean but Elevation be 1000, using model in part(b) and part(d). Does adding the deleting the predictor Elevation have much influence on these predicted values? What if Elevation be 30?
- Refit the model from part(d) accounting for possible overdispersion. Report the overdispersion coefficient and explain what it means.
- An alternative model would include Area as an offset or exposure variable. Explain why or why not treating Area in this fashion would be justified.

	Species	Area	Elevation	Nearest	Scruz	Adjacent
Baltra	58	25.09	346	0.6	0.6	1.84
Bartolome	31	1.24	109	0.6	26.3	572.33
Caldwell	3	0.21	114	2.8	58.7	0.78
Champion	25	0.10	46	1.9	47.4	0.18
Coamano	2	0.05	77	1.9	1.9	903.82
Daphne.Major	18	0.34	119	8.0	8.0	1.84
Daphne.Minor	24	0.08	93	6.0	12.0	0.34
Darwin	10	2.33	168	34.1	290.2	2.85
Eden	8	0.03	71	0.4	0.4	17.95
Enderby	2	0.18	112	2.6	50.2	0.10
Espanola	97	58.27	198	1.1	88.3	0.57
Fernandina	93	634.49	1494	4.3	95.3	4669.32
Gardner1	58	0.57	49	1.1	93.1	58.27
Gardner2	5	0.78	227	4.6	62.2	0.21
Genovesa	40	17.35	76	47.4	92.2	129.49
Isabela	347	4669.32	1707	0.7	28.1	634.49
Marchena	51	129.49	343	29.1	85.9	59.56
Onslow	2	0.01	25	3.3	45.9	0.10
Pinta	104	59.56	777	29.1	119.6	129.49
Pinzon	108	17.95	458	10.7	10.7	0.03
Las.Plazas	12	0.23	94	0.5	0.6	25.09
Rabida	70	4.89	367	4.4	24.4	572.33
SanCristobal	280	551.62	716	45.2	66.6	0.57
SanSalvador	237	572.33	906	0.2	19.8	4.89
SantaCruz	444	903.82	864	0.6	0.0	0.52
SantaFe	62	24.08	259	16.5	16.5	0.52
SantaMaria	285	170.92	640	2.6	49.2	0.10
Seymour	44	1.84	147	0.6	9.6	25.09
Tortuga	16	1.24	186	6.8	50.9	17.95
Wolf	21	2.85	253	34.1	254.7	2.33