Reading: GLM.Basics
Model: 2013.Q2
Problem Type: Define the design matrix and vector of responses

Given An actuary is building a log-link generalized linear model to create a Homeowners Hurricane Severity model using the data below.

Dollars of loss

Hurricane Shutters	Distance to Coast	
	≤ 25 miles	> 25 miles
Yes	$5,968,000	$5,092,000
No	$5,609,000	$1,133,000

Number of Claims

Hurricane Shutters	Distance to Coast	
	≤ 25 miles	> 25 miles
Yes	14	35
No	23	18

Average Coverage A Amount

Hurricane Shutters	Distance to Coast	
	≤ 25 miles	> 25 miles
Yes	$446,000	$350,000
No	$251,000	$269,000

The model will include four parameters: β_0, β_1, β_2, and β_3, where β_0 is the intercept, β_1 is the average severity for homes with hurricane shutters, β_2 is the average severity for homes greater than 25 miles from the coast, and β_3 is the average severity for the natural log of the average Coverage A Amount (continuous variable).

Find

a. Define the design matrix [X].

b. Define the vector of responses [Y].

Solution

Average Severity = Dollars of loss / Number of Claims

Average Severity

Hurricane Shutters	Distance to Coast	
	≤ 25 miles	> 25 miles
Yes	$426,286	$145,486
No	$243,870	$62,944

We have four distinct data points in the GLM, one for each combination of Hurricane Shutters and Distance to Coast. As such, the design matrix will have four rows.

Alice: "Although there are only four records in the data set, remember these likely came from many observations that were aggregated to this level."

The design matrix consists of a column for each parameter. In the case of a categorical variable the value is either 1 or 0 depending on whether or not the record has that rating characteristic. For a continuous variable, the actual value is used after any transformation needed.

	β0	β1	β2	β3
	1	1	0	LN(446,000)
X =	1	1	1	LN(350,000)
	1	0	0	LN(251,000)
	1	0	1	LN(269,000)

Alice: "Notice there's a row for every observation and all rows have a 1 in the intercept column (β_0)."

Alice: "It's important you follow the given description of the parameters because this tells you the base levels. Here it's implicit the base levels are:
1. Homes with no hurricane shutters
2. Homes less than or equal to 25 miles from the coast.
Remember the base level is usually the one with the most exposures. This makes sense here as people tend to live close to the coast and not always have hurricane shutters."

The vector of responses is a column vector of the average severities. It's important you write them down in the same order you processed the records when building the design matrix.

	$426,286
Y =	$145,486
	$243,870
	$62,944