I

Understanding ENSO Risk

1 Estimating Disaster Damages

The first step in scoping teleconnections markets involves estimating the economic damages associated with target indexes. That estimation requires a large database of disaster damage statistics. Unfortunately, many of the leading disaster databases are missing substantial data. So, before turning to ENSO in chapter 2, this chapter presents:

- a statistical process using country-level statistics from the World Bank to supplement the EM-DAT disaster statistics database; and
- analysis of the quality of that statistical process through cross-validation.

EM-DAT database

The three databases most popular for academic disaster research are, the Emergency Events Database (EM-DAT) maintained by Center for Research on the Epidemiology of Disasters in Brussels, NatCat maintained by the reinsurance company MunichRe, and Sigma by the reinsurer SwissRe. EM-DAT, the database I use here, offers fewer recent records than NatCat, the largest of the three at roughly 15,000 entries. But it has more complete historical records and is easily accessible to researchers[1]. Established as resource for epidemiological studies, and supported by the United Nation's World Health Organization (WHO) and the Belgian Government, EM-DAT is the database that academics have used most often to estimate ENSO related damages[2,3].

EM-DAT contains data on roughly 12,000 natural disasters from 1900 to present. Each disaster in the database contains some combination of the following information:

- Country: Country(ies) in which the disaster has occurred.
- Disaster type: Description of the disaster according to a pre-defined classification.

[1] Debarati Guha-Sapir and Regina Below. The quality and accuracy of disaster data: A comparative analyses of three global data sets. Working paper, 2002

[2] MJ Bouma, RS Kovats, SA Goubet, J.S.H. Cox, and A. Haines. Global assessment of El Niño's disaster burden. *The Lancet*, 350(9089): 1435–1438, 1997

[3] L. Goddard and M. Dilley. El Niño: Catastrophe or opportunity. *Journal of Climate*, 18(5):651–665, 2005

- Date: The date when the disaster occurred.
- Killed: Persons confirmed as dead, missing, and presumed dead.
- Total affected: A sum including:

  Injured: People suffering from physical injuries, trauma or an illness requiring medical treatment as a direct result of a disaster.

  Homeless: People needing immediate assistance for shelter.

  Affected: People requiring immediate assistance during a period of emergency; it can also include displaced or evacuated people.
- Estimated Damage: Several institutions have developed methodologies to quantify these losses. However, there is no standard procedure for determining losses. Estimated damages are given in million USD current to the reporting of the disaster.

Those data points are compiled from various sources, including UN agencies, non-governmental organizations, insurance companies, research institutes and press agencies. For a disaster to be entered into the database it must fulfill at least one of the following conditions:

- Ten or more people reported killed.
- Hundred or more people reported affected.
- Declaration of a state of emergency.
- Call for international assistance.

EM-DAT is large enough to provide a solid foundation for investigating the consequences of regional climate anomalies. But the database suffers from some glaring omissions. For example, the northern Peruvian state of Piura was the epicenter of flooding during the 1997/1998 El Niño. As noted earlier, an estimated 200,000 people were affect by that flooding. However, the EM-DAT database registers the greatest El Niño in the modern era as only an outbreak of disease in Piura and flooding in nearby Ecuador. For that reason, I supplement the EM-DAT statistics with regressions to fill in missing data in this chapter. I also use informative Bayesian priors on impact parameters in chapters 2 and 10.

Supplemental data sources and initial exploratory analysis

EM-DAT includes roughly 11,000 disasters dated between 1960 and 2010. Of those 11,000 entries, most included estimates of the people killed and affected[4], but only 32 percent included estimates of economic damages. I randomly selected 75 percent of those roughly 3,600

[4] I supplemented the entries missing estimates of the numbers affected or killed with the median value for that disaster type.

disasters with damage estimates as the training set for a Bayesian regression that would extend expected damage estimates to all disasters in the sample, a process called bootstrapping in statistics.

EM-DAT gives disaster estimates in terms of USD current to the disaster's reporting (i.e. generally in the year that the disaster occurred.) Using a base month of June 2010, I adjusted those economic damage estimates in the database values for inflation. Inflation estimates came from the U.S. Department of Labor's Bureau of Labor Statistics via the Federal Reserve Bank of St. Louis' FRED database[5].

Once the sample's damage estimates were in comparable units, I added in the World Bank and IMF estimates in 2010 US dollars of the GDP per capita in the country-year of each disaster[6]. Figure 1.1 shows a scatter plot of disasters in EM-DAT with official damage estimates plotted against the annual GDP per capita.

[5] Federal Reserve Bank of St. Louis. Consumer price index for all urban consumers: All items (cpiaucsl). http://research.stlouisfed.org/fred2/series/CPIAUCSL, 2012

[6] The World Bank Group. World Bank Data. http://data.worldbank.org/, 2012

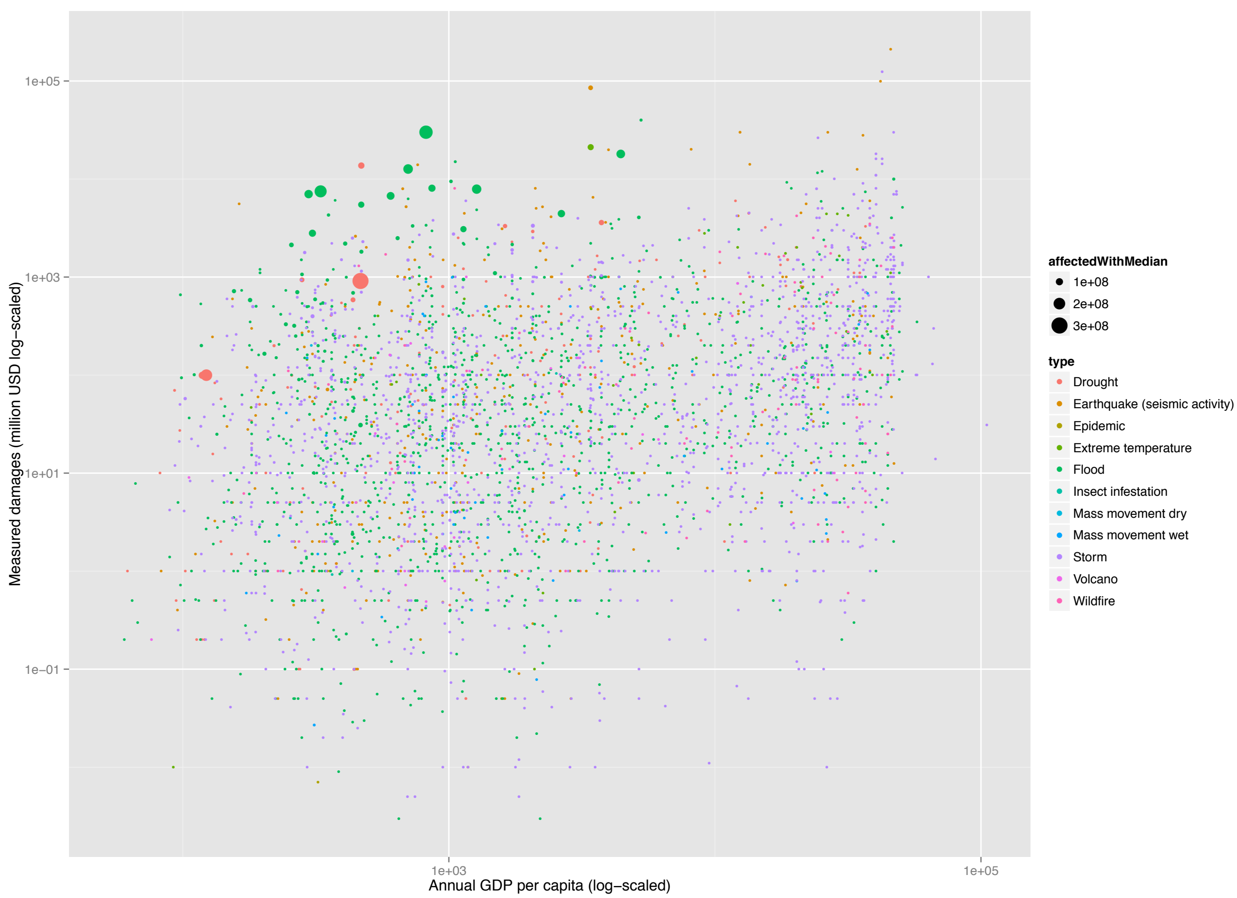

Figure 1.1: Scatter plot of available disaster damage estimates from EM-DAT database. The disaster type is indicated by the color of the point. The number of individuals affected is indicated by the diameter of the point. In this figure, the median number affected for that disaster type filled in for missing estimates of the number affected.

Figure 1.1 shows a small cluster of major disasters impacting populous and relatively poor counties. These are represented by the large dots in upper left quadrant of the figure. These are mostly major floods in India, China, and Bangladesh. While they caused some of the largest losses in the database, they affected orders of magnitude more people than other expensive disasters. In other words, the ratio between the human and measured economic tolls of these disasters was distinct from the others in the sample. The regression is most revealing when there is a stable relationship between economic damage and the combination of people affected and country-year GDP per capita. To achieve this, I separated this cluster of events from the rest of the sample and estimated the same regression on just that outlier sub-sample.

To identify this cluster of observations, I ranked all disasters by their ratio of affected individuals (including median estimates where no others were available) to the log of their GDP per capita in the country-year where they occurred. I separated out observations with high ratios of the people affected to log GDP per capita, those with ratios at or above the sample's 98th percentile.

Apart from segregating that cluster of events into its own regression, my other data cleaning procedures were of minor importance. I discarded the two disasters with irregular entries, one in the early 1960s in Uganda and another in Luxembourg. I also replaced zeros for disaster impact estimates with a nominally low value (0.1) so that all variables could be analyzed in log form. I also replaced zero estimates of GDP per capita (due mostly to lapses in record-keeping, as in Afghanistan under the Taliban) with the lowest recorded value for GDP per capita in the sample.

The final step before running the bootstrap was to divide the sample with damage estimates into a training set and a set for cross-validation. To do so, I randomly selected 25 percent of the disasters with damage estimates and set them aside to cross-validate my fitted regression. This left me with a total of six data sets. The number of disaster events in each set is available in table 1.1.

Table 1.1: Number of disasters in each data subset

	Training set	Cross-validation set	Prediction set	Total
Basic set	2646	862	7680	11188
High affected:GDP/capita	110	44	73	227
Total	2756	906	7753	11415

Bootstrapping additional estimates of economic damages

I bootstrapped economic damage estimates using the Bayesian statistical program JAGS and a varying-intercept model (equation 2.1) adapted from Gelman and Hill [2007]. This model, where each disaster is subscripted i, estimates separate damage equations for each disaster type based on diffuse priors. The model is not hierarchical, insofar as there is no linkage of data across disaster types. I decided against modeling with informative priors or a hierarchical model, two potential advantages of using Bayesian techniques. Nevertheless, I preferred Bayesian methods because they do not impose an assumption of stationarity in the underlying model parameters and they facilitate simple bootstrapping.

$$\begin{array}{lcl} \log \text{damage}_i & \sim & \mathcal{N}(\hat{y}_i, \sigma_y^2) \\ \hat{y}_i & = & a_{\text{disaster type},i} \\ & & +b_1 * \log \text{GDP per capita}_i \\ & & +b_2 * \log \text{affected}_i \\ & & +b_3 * \log \text{killed}_i \\ a_{\text{disaster type}} & \sim & \mathcal{N}(\mu, \sigma_a^2) \\ \mu, b & \sim & \mathcal{N}(0, 1000) \\ \sigma^2 & \sim & \mathcal{U}(0, 100) \end{array} \tag{1.1}$$

Table 1.2 presents parameter estimates for equation 2.1 fit to the main training set (i.e. excluding the outlier training set.) Trace and density plots of the regression parameters after $50,000$ iterations, with no thinning on the basic data set are shown in Miscellaneous Appendix's figures 21 through 25. They indicate good mixing of the simulation chains with the $\hat{R}$ parameter in 1.2 at 1.001 or below for all parameters.

The regression indicates that the most important factor associated with a disaster's damages is log GDP per capita in the country-year where it occurred. The mean estimate from the model indicates that a 1 percent increase in GDP per capita is associated with a 0.53 percent increase in economic damages. The second strongest factor in the model is log of people killed in the disaster, with a 1 percent increase in the number of people killed in a disaster associated with 0.42 percent more economic damage. Finally, a 1 percent increase in the number of people affected by the disaster was associated with an increase in economic damages of 0.2 percent. All three parameters have 95 percent probability intervals well above zero and all three parameters appear largely distinct from one another, with only a slight overlap of the 95 percent probability intervals for the log GDP per capita and log of people killed in the disaster. This means that they are above zero and their order of relative importance is stable, with high probability.

Interestingly, the regression shows significant overlap between all the 95 percent probability intervals of the regression intercepts, indicating that none of the disaster types distinguished themselves as being particularly devastating, independent of the country or people they impacted. While the various disaster types are not distinct from one another with high probability, the gap between the mean estimate of the least (Mass movement wet) and most (Drought) impactful disaster type is large enough to be of fundamental economic importance. Independent of specific impacts (people killed etc.), a drought is associated with 94.8 percent more economic damage than a mudslide (Mass movement wet) according to the mean parameter estimates from the model.

	mean	sd	2.50%	25.00%	50.00%	75.00%	97.50%	$\hat{R}$	n.eff
a[1] Flood	-3.147	0.353	-3.837	-3.383	-3.150	-2.909	-2.444	1.0010	24000
a[2] Storm	-2.856	0.359	-3.559	-3.097	-2.858	-2.615	-2.147	1.0010	23000
a[3] Earthquake	-2.593	0.378	-3.332	-2.847	-2.594	-2.343	-1.844	1.0010	24000
a[4] Drought	-2.416	0.416	-3.220	-2.699	-2.415	-2.138	-1.599	1.0010	18000
a[5] Extreme temperature	-2.906	0.461	-3.809	-3.217	-2.909	-2.598	-1.999	1.0010	24000
a[6] Mass movement wet	-3.364	0.464	-4.298	-3.675	-3.356	-3.048	-2.465	1.0012	7600
a[7] Wildfire	-2.852	0.425	-3.684	-3.140	-2.854	-2.569	-2.017	1.0010	24000
a[8] Volcano	-3.287	0.512	-4.323	-3.623	-3.271	-2.938	-2.324	1.0010	24000
a[9] Epidemic	-3.177	0.693	-4.715	-3.560	-3.129	-2.728	-1.965	1.0010	24000
a[10] Insect infestation	-3.025	0.616	-4.302	-3.399	-3.007	-2.629	-1.846	1.0010	20000
a[11] Mass movement dry	-3.072	0.661	-4.488	-3.458	-3.037	-2.651	-1.851	1.0010	24000
b1 Log GDP per capita	0.533	0.030	0.473	0.512	0.533	0.553	0.592	1.0009	24000
b2 Log affected	0.200	0.019	0.162	0.187	0.200	0.213	0.238	1.0010	18000
b3 Log killed	0.422	0.028	0.368	0.404	0.422	0.441	0.478	1.0009	24000
mu.a	-2.972	0.409	-3.787	-3.243	-2.970	-2.698	-2.181	1.0010	24000
sigma.a	0.481	0.233	0.175	0.323	0.435	0.584	1.066	1.0009	24000
sigma.y	2.391	0.033	2.327	2.368	2.391	2.413	2.457	1.0011	10000

Table 1.2: Diagnostics for Bayesian regression of economic damages given population affected, population killed, and GDP per capita the disaster's country-year

Figure 1.2 shows the in-sample prediction of the model based on equation 2.1. It includes sub-sample of high disaster with high affected to GDP/capita ratios that I estimated using its own regression. The straight black line in the figure is a benchmark for a one to one correspondence between the model's predicted damages for a disaster and the actual observed values in the EM-DAT database. The black line runs directly through the cluster, indicating that the model provides a reasonable in-sample fit. It reliably infers the order of magnitude of a disaster's damages using only the disaster type, GDP per capita, and the numbers of people affected and killed.

When I used the model to model damages in the cross-validation set, it produced predictions that were similarly in line with observed damages. Figure 1.3 presents observed and predicted damages for the cross-validation set. As with the in-sample prediction in figure 1.2, the straight black line indicates perfect correspondence between modeled and observed damages. It runs directly through the main cluster of observations. That suggests that the fitted models offer reasonable inference on out-of-sample disasters.

Based on this cross-validation, I determined that the model was robust and used it to infer damages for the subset of the EM-DAT database without observed damages. After excluding disasters that I did not believe could be linked to regional climate anomalies (985

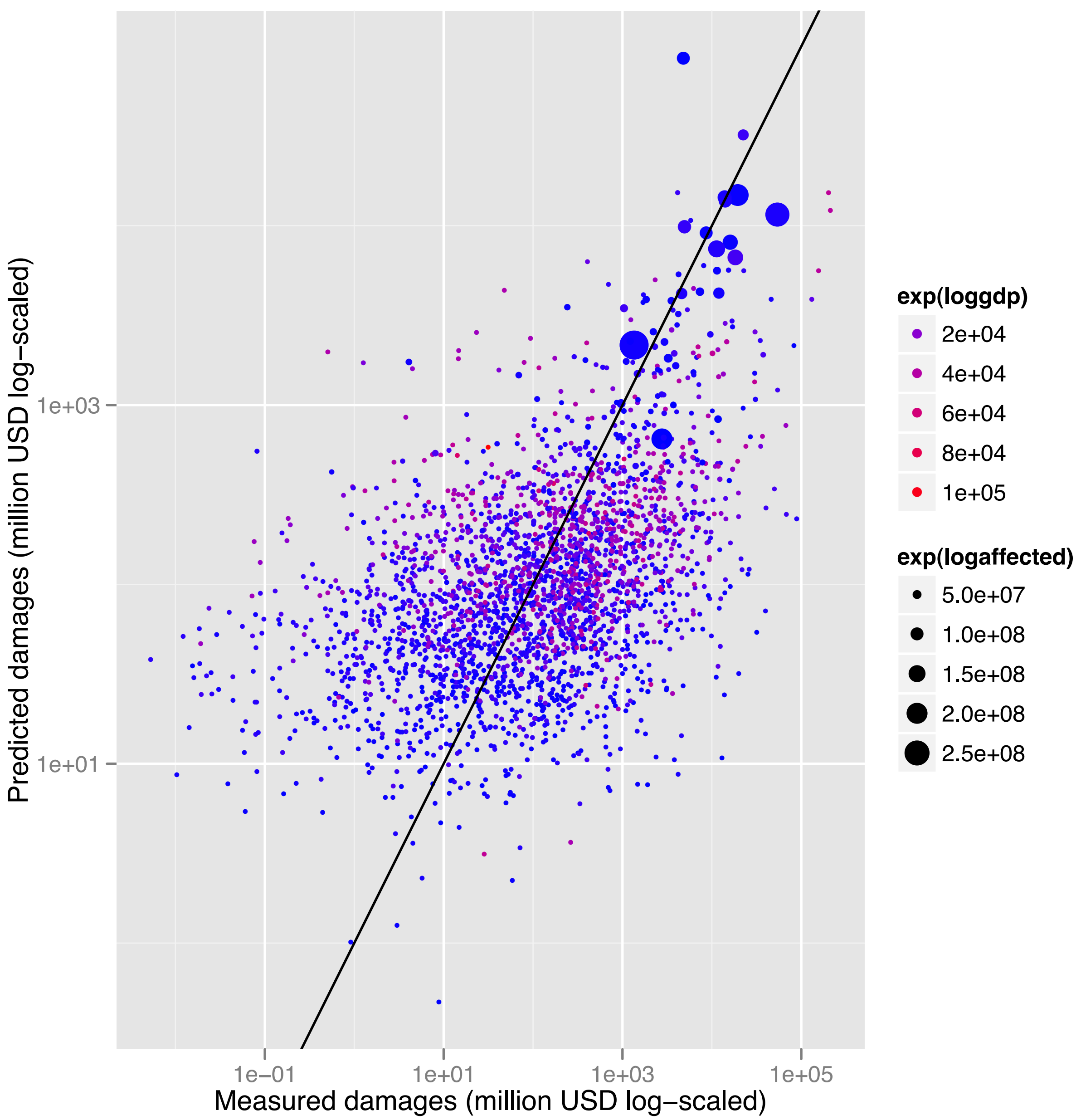

Figure 1.2: In-sample fit of estimated economic damages vs. actual economic damages from EM-DAT

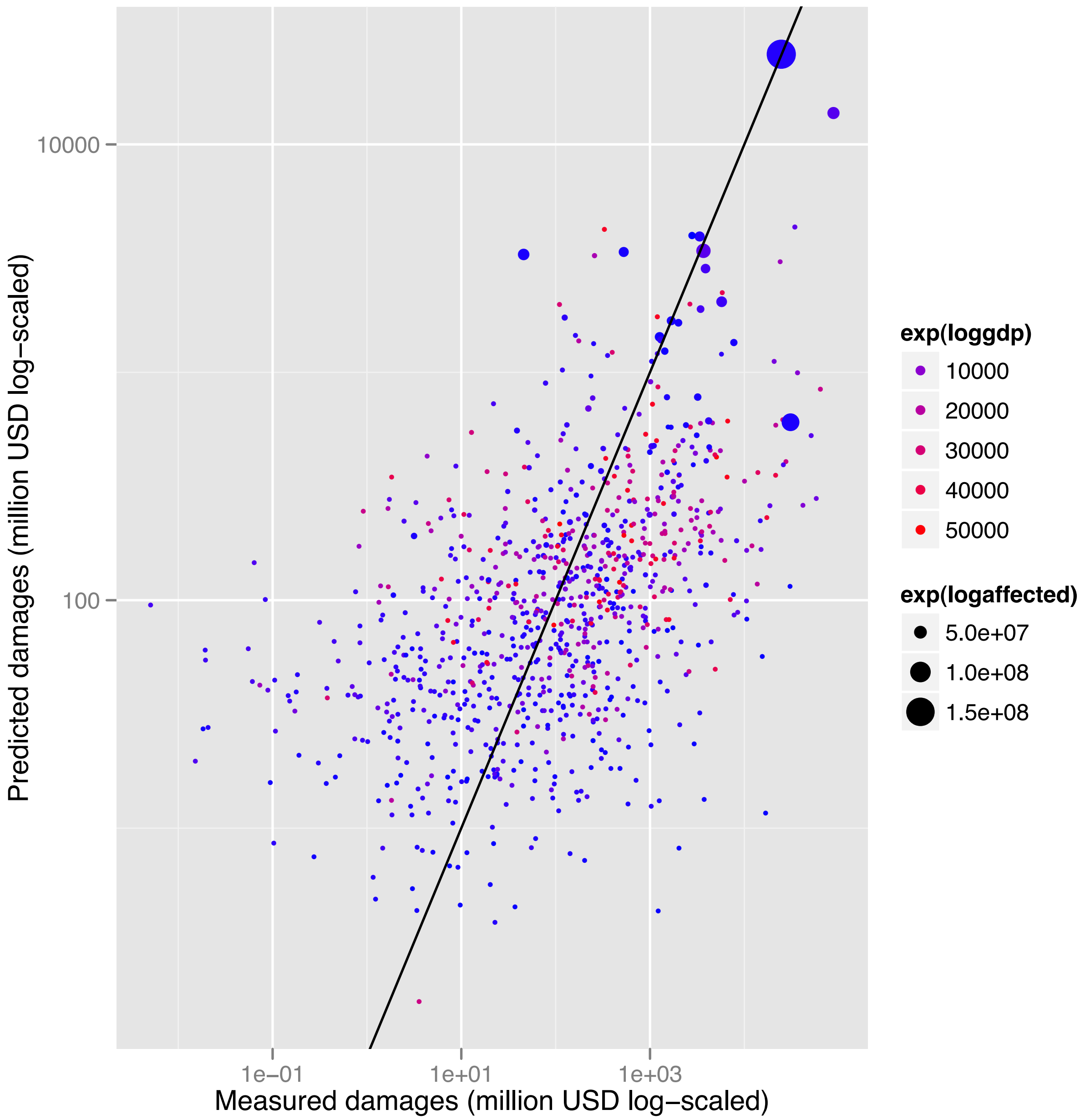

Figure 1.3: Cross-validation of estimated economic damages vs. actual economic damages from EM-DAT

earthquakes and 191 volcanic eruptions) and discarding disasters without a reported start month, I was left with a sample of 9979 events between 1960 and 2010, disasters with economic damage estimates that I considered relevant to teleconnection indexes. Roughly two-thirds of those damage estimates came from the bootstrapping outlined in this chapter.