This version: October, 2019

SUPPLEMENTARY MATERIALS FOR:
"Specification Curve: Descriptive and Inferential Statistics On All Reasonable Specifications"

Uri Simonsohn
Universitat Ramon Llull
ESADE Business School
urisohn@gmail.com

Joseph P. Simmons
University of Pennsylvania
The Wharton School
jpsimmo@wharton.upenn.edu

Leif D. Nelson
University of California, Berkeley
Haas School of Business
Leif_nelson@haas.berkeley.edu

OUTLINE.

Section	Pages
Supplement 1. Set of reasonable specifications for hurricanes study	2-10
Supplement 2. Set of reasonable specifications for racial discrimination study	11-13
Supplement 3. Descriptive specification curve for racial discrimination study	14
Supplement 4. The benefits of bootstrapping in specification-curve analysis (Poisson regression example).	15-18
References	19

Supplement 1. Set of reasonable specifications for hurricanes study.

Jung et al. (Jung, Shavitt, Viswanathan, & Hilbe, 2014) hypothesized that hurricanes with more feminine names are perceived as less threatening and hence lead to fewer precautionary measures by the general public. To convert this hypothesis into a testable prediction, Jung et al carried out various operationalizations. To construct a specification curve, we examine what we judged to be the five major operationalizations (most likely to be consequential), and consider sensible alternatives. In particular, we shall examine these operationalizations:

1. The set of storms to include in the analyses
2. How to measure the femininity of storms' names
3. What regression model to run (e.g., Negative-Binomial vs OLS)
4. What's the key prediction made by the authors' hypothesis
5. What to control for

Note that these five mirror the operationalizations in Figure 1 in the main paper.

1) The set of storms to include in the analyses

1.1) Universe of storms

Jung et al. included only Atlantic hurricanes included in a NOAA list (see page in their paper).[1] The universe of named storms that cause destruction is much larger than that. First, named tropical storms (of lesser intensity than hurricanes) also lead to deaths and have gendered names (these would more than double sample size). Second,

[1] The list has been saved on the WebArchive: http://web.archive.org/web/20140709120550/http://www.aoml.noaa.gov/hrd/hurdat/All_U.S._Hurricanes.html

hurricanes elsewhere in the world also receive names and also cause deaths.[2] Third, some Atlantic hurricanes are missing from the NOAA list used. For instance hurricane Diane from 1955, Isidore from 2002, and Ernesto from 2006 are missing from the list.

While it would be sensible and straightforward to assess the robustness of the results to different definitions of the universe of storms to study, this type of robustness is somewhat unusual (most papers cannot so easily expand their data-sources) and hence we have decided not to do so for our specification curve demonstration.

1.2) *Outliers*

The paper excludes hurricanes Katrina and Audrey from the analyses, considering them "outliers."[3] The exclusion decision was made after the authors run the regressions with them included, and is motivated in the paper as seeking to eliminate over-dispersion from the model (rather than seeking to eliminate observations that may be invalid). [4,5]

The two excluded observations have 1833 and 416 deaths, see solid circles in Figure S1. The same figure highlights other candidate outlier observations (dotted circles). The can also be though thought of as leverage points, observations with extreme predictor values.

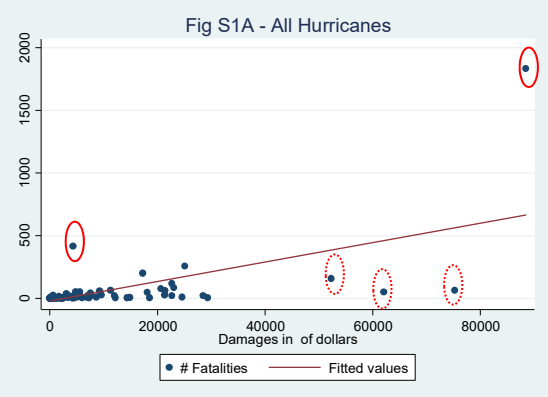
---

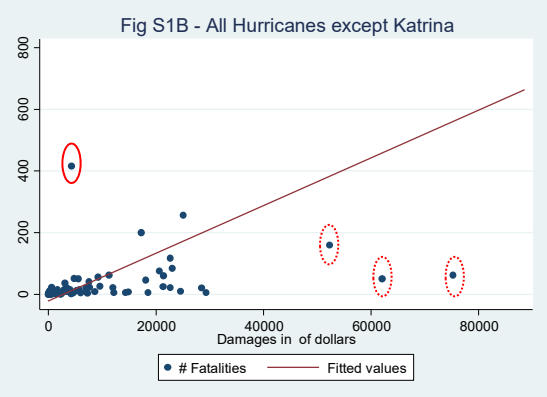
[2] See e.g. this list saved on the WebArchive:
http://web.archive.org/web/20150216102357/http://www.nhc.noaa.gov/aboutnames.shtml

[3] Jung et al. reads *"We removed two hurricanes, Katrina in 2005 (1833 deaths) and Audrey in 1957 (416 deaths), leaving 92 hurricanes for the final data set. Retaining the outliers leads to a poor model fit due to overdispersion*." (p.4)

[4] There are many other ways to address over-dispersion. For example, a textbook on binomial regressions, written by one of the authors of the PNAS paper, suggests 35 different methods to deal with overdispersion; adjusting for outliers is just one of them, and there are in turn many ways to deal with outliers (Hilbe (2011) "*Negative Binomial Regressions*," Second Edition, p. 158)

[5] Because those observations are known to be legitimate, the overdispersion these "outliers" create probably is better addressed by modifying the model rather than ignoring those valid datapoints.





Notes: The solid circles are the two "outliers" drop from the analyses by Jung et al., the dashed circles identify additional potential outliers (or leverage points, extreme values of X variable).

In light of the above, we examine the following alternatives to deal with outliers (new operationalizations in light blue, original in black):

Alternative operationalizations for dealing with "outliers:"

1) Exclude 0 observations.
2) Exclude 1 most extreme on deaths (drop Katrina with 1833, keep Audrey)
3) Exclude 2 most extreme on deaths (drop Katrina & Audrey)
4) Exclude 2 most extreme on deaths and remaining 1 most extreme on damages
5) Exclude 2 most extreme on deaths and remaining 2 most extreme on damages
6) Exclude 2 most extreme on deaths and remaining 3 most extreme on damages

2) How to measure the femininity of storms' names

The authors used 9 raters to judge in a 1-11 scale the femininity of the storm names, and also a binary (1=female, 0=male) gender indicator. [6,7]

Operationalizations for quantifying femininity: [8]

[6] Throughout we use the term "linearity" abstracting from the fact that estimated regression is a negative binomial and hence a linear term is not really assuming a linear effect. The key point is that a linear term imposes a strong functional form assumption, rather than that assumption is of a linear effect per-se.

[7] An additional concern worth mentioning is that femininity of a name may be correlated with other attributes of the name, such as how strong, evil, or harmful names are perceived. E.g., male name Adolf vs female name Angel. This would require controlling for other name attributes, something that would be a distraction for our purposes but necessary to properly interpret the original results.

[8] Because the authors do not report femininity for Katrina and Audrey, we conducted an MTUrk survey with 32 participants asking them to rate all 94 storms using the same scale from Jung et al. The ratings were correlated $r = .98$ with those used by Jung et al. However, within gender the ratings are much lower: $r = .67$

1) Femininity rating
2) Binary gender indicator

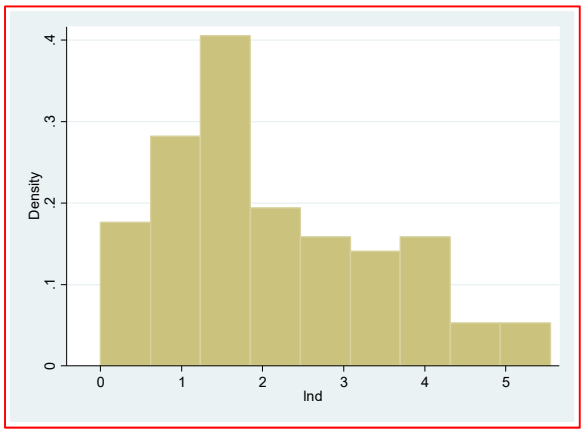
3) What regression model to run

Jung et al. estimated a negative binomial regression which is often used for count data with over-dispersion (that is, where the Poisson assumption of $\mu=\sigma$ does not apply). Some papers examining deaths from natural disasters employ a zero-inflated negative binomial (Czajkowski, Simmons, & Sutter, 2011), where a higher proportion of observation with 0 values (higher than that implied by a regular negative binomial) is observed. To perform a zero inflated one needs to identify variables that predict whether there are any deaths but not how many there are. This complication, paired with just 10% of observations having 0 deaths, leads us to not include a zero-inflated negative binominal in the specification curve.

Another alternative to the negative binomial is to run OLS with log(count+1) as the dependent variable. Logging count data has been discouraged (O'Hara & Kotze, 2010). The discouragement is based on simulations that show that if the true data are binomial or Poisson, then log(count) performs worse than a Poisson or negative binomial model, but the whole point is that one may doubt the underlying data are adequately captured by a negative binomial or Poisson model, and wishes to run the log model for robustness. As the authors of the paper discouraging log(counts) write "…*our result may not generalize to real data, which rarely has (sic) as balanced a design as our simulations"*

for male names, $r = .83$ for female names. This suggests the measure of femininity beyond the binary gender variable adds considerable noise to the model. We use the MTurk ratings in our analyses.

Simple linear models are known to be robust to a broad range of violations of assumptions, this is not the case for non-linear models like the negative binomial. In addition, log(deaths) has a rather reasonable distribution, that a linear model should obtain valid estimates form.



Moreover, a well cited paper of predicted hurricane deaths uses a OLS with log(deaths+1) as the dependent variable (Toya & Skidmore, 2007). [9] We therefore include OLS regressions with log(count+1) as the dependent variable in our specification curve.

Regression models
1) Negative binomial
2) OLS regression with log(deaths+1) as the dependent variable

4) What's the key prediction made by the authors' hypothesis

The key hypothesis in the paper is that "*a hurricane with a feminine vs. masculine name will lead to less protective action and more fatalities*." (p.1) This prediction

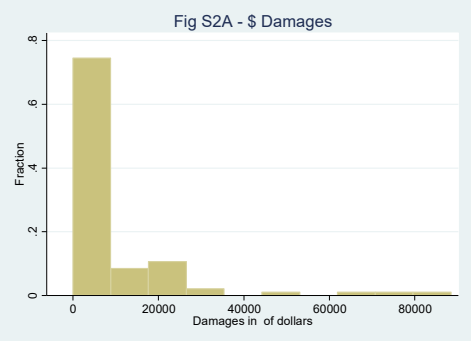
[9] 279 Google cites as of January 2015.

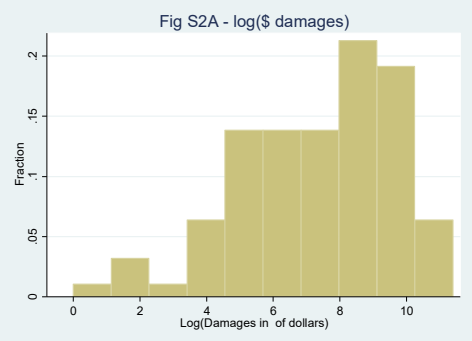
suggests a **main effect** of gender of hurricane name on fatalities but the paper tests instead an interaction of femininity with dollar damages (such that the effect is stronger for hurricanes high in dollar damages).

One can justify this specification by considering that fatalities is a proxy for insufficient protective measures, and the proxy is only sensible when people who fail to take protective measures have a higher probability of death. Failing to protect against harmless storms should not lead to higher probability of death. Nevertheless, there are many alterantive ways to operationalize this potential dependency of the effect of gender on dollar damages (including ignoring it). We discuss these alternatives below.

4.1 Damages vs log(damages)

The operationalization in the paper has gender interacted with damages measured in dollars "linearly" [10] mapping on to deaths. As the histograms in Figure S2 show, damages measured in dollars is an extremely skewed distribution, while log(damages) has a reasonable degree of skew. We hence shall include specifications with log(damages) also.





[10] See footnote 4

Functional form for $ damages:

1) $ Damages
2) Log($ Damages)

4.2 Interactions

Beyond functional form, note that damages is being used as a proxy for potential fatalities. There are various other ways to proxy how deadly a storm would be expected to be. Rather than use only dollar damages, for instance, one may include hurricane characteristics as well, such as its category, it maximum wind speed, its year, etc. Indeed Jung et al used an interaction with minimum pressure as a covariate. We consider operationalizations that add or replace wind, and hurricane category as proxies for expected fatalities. [11]

4.3 Main effect

Finally, while the justification for the interaction prediction is reasonable, it seems ex-ante also reasonable to examine the *main effect* of gender, on various grounds. First, as shown above, the stated hypothesis in the paper stipulated a main effect of femeninity. Second, 90% of hurricanes lead to at least one fatality, suggesting lack of protective measures could be fatal for most observations. Third, when the authors control for third variables (e.g., hurricane year), for which the same logic to require an interaction applies, they do not include the interaction, suggesting further than main effect estimates are sensible. Fourth, prior papers running models predicting fatalities from natural disasters like hurricanes and earthquakes test main effects (of variables like per-capita GDP) rather

[11] Jung et al indicate that wind data are not available prior to 1979, but we were able to locate such data and use it in our analyses.

than interactions with disaster intensity (Czajkowski et al., 2011; Kahn, 2005; Toya & Skidmore, 2007). It hence seems ex-ante sensible to estimate a gender main effect to test the hypothesis of interest.

Operationalizations for key prediction.

Codebook

Zmin:[12] -1*(standardize minimum pressure)
Fem: femininity of name of hurricane
Dam: damages of hurricane (log or $)
zCat: standardize category of hurricane (1-5)
zWin: standardized maximum wind of hurricane
z3: average(zMin, zCat, ZWin)

Specifications to test the impact of femininity.
Via interactions:

1. Fem*Dam
2. Fem*Dam Fem*zMin
3. Fem*Dam Fem*zWind
4. Fem*Dam Fem*zCat
5. Fem*Dam Fem*z3

Via main effect:

6. Fem Dam z3

5) What to control for

Hurricane names are randomly assigned, hence covariates might be expected to not play an important role in the regression (omitted variables shouldn't correlate with randomly assigned names). The key predictor in most specifications, however, involves an interaction, and the interaction term, Damages, is not randomly assigned and hence could have confounds. Moreover, as shown above it is highly skewed, introducing possible specification error into some models, additional controls interacted with damages may alleviate or at least facilitate identifying these problems.

[12] Minimum pressure is multiplied by -1 so that a higher number is associated with higher intensity and is hence easy to combine with the other indicators.

The dataset only contains year as a plausible covariate (in addition to the hurricane intensity variables from (3)). Considering that time effects can often be non-linear, and that year has an important discontinuity in 1979, prior to 1979 all hurricanes had male names we consider the following set of covariate:

Specifications for covariates

1) No covariate[13]
2) Year * Damages
3) Dummy for year after 1979 * Damages

[13] Jung et al. (2014) did also run models controlling for year and write that it "was dropped for the main analysis as its effect was nonsignificant in all models." (p.4)

Supplement 2. Set of reasonable specifications for racial discrimination study.

Bertrand and Mullainathan (Bertrand & Mullainathan, 2004) manipulated both the name of the fictitious candidates whose names they used and the quality of the resumes they sent. To manipulate perceived race, they used 18 distinctively African American names, and 18 non-distinctively African-American names. Half the names were male. The way in which resumes' quality was manipulated varied from ad to ad, creating some ambiguity in terms of how to quantify the quality of any given resume. This ambiguity is the primary source of the alternative specifications we consider.

We generate the set or reasonable specifications by considering alternative operationalizations involving:

1) How to deal with potential heterogeneity of the main effect across genders
2) How to measure quality of resume
3) What regression model to employ (OLS vs Probit)

1) How to deal with potential heterogeneity of the main effect across genders

Considering Bertrand and Mullainathan report some results broken down by gender, that it would be reasonable to observe and report discrimination only in one of the genders, or of different magnitude across gender, we report results for the entire sample, only for males, and only for females.

2) How to measure quality of resume

To most help-wanted ads, Bertrand and Mullainathan (Mullainathan, 2002) sent four resumes. Two high and two low quality ones (orthogonally varying race and gender also). Whether to be of high and low quality is randomly assigned, but how to implement

how vs low quality is decided on a case-by-case basis in light of the nature of the ad. Resumes were made of higher/lower quality by varying holes in employment history, having a certification degree, possessing foreign language skills, etc.

Bertrand and Mullainathan operationalize quality in their regression results in two main ways: using a 1/0 predictor for having been randomly assigned to high vs low quality, using a continuous predictor of quality (see e.g. Panels A and B in their Table 4). The continuous predictor was created by estimating the regressions in two stages. In the first stage, using 1/3 of the sample, they predict call-back rates using all measures of quality they manipulated. They then use the fitted values for call-back probabilities as the quality index for the remaining 2/3 of the sample, using a median split for high vs low predicted call-back rates as the alternative measure of quality.

We expand these two to fifteen alternative operationalizations of quality. We modify their two-stage estimation in a way that increases power. In particular, rather than use 1/3 of the sample to obtain fitted values, we use 1/2. In addition, we do not drop observations, our second stage includes all observations, one half of fitted values are obtained from the other half. One could increase power further with more refined techniques (e.g, jackknife) but it is not necessary for the purposes of our demonstration.

As operationalizations of quality we added the simple sum of 0/1 quality indicators (that is, the number of quality variables that were changed to create a higher quality resume). A second alternative was the median split of this variable.

The remaining alternatives are rely on the two-stage estimation approach alluded to above. We varied whether this first stage included covariates or not (the specification in the paper includes as covariates gender, city, occupation code for the job, and dummy

variables for required skills), and whether it was estimated on all names, or only distinctively Black or White names. The logic for this later variation in operationalizations is that Blacks and non-Blacks may benefit differently from different quality measures (e.g., some quality measures may alleviate negative stereotypes for Black names, but have no effect on White names).

Each of these three 2-stage approaches was implemented with and without covariates in the 1st stage , and entered as a continuous or median split predictor in the 2nd stage, resulting in 3x2x2=12 operationalizations. Combined with the previously mentioned 3 we arrive at the 15 alternative specifications of quality.

3) Regression model

The paper (Bertrand & Mullainathan, 2004) reported probit regression for the 1st stage, classified observations into a high and low quality bin, and conducted simple $\chi^2(1)$ difference of proportion tests on the resulting cells (see their Table 4). Because we consider continuous predictors of quality we rely on regression models throughout, reporting results for both probit and OLS regressions. We should note that they key prediction is one of an interaction: is the benefit of a quality resume higher for nonblack names? (Gelman & Stern, 2006) But the authors only report the two simple effects (significant effect for nonBlacks, but not for Blacks). The non-reported interaction is not significant in either of the specifications included in the paper.

Supplement 3. Descriptive Specification Curves for Discrimination Study

Figure S3. Descriptive Specification Curve – Main effect of distinctively black name on call back rate.

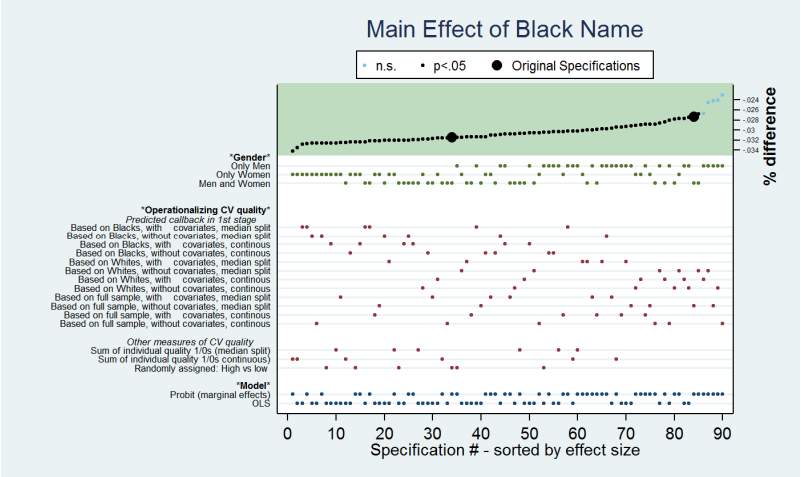
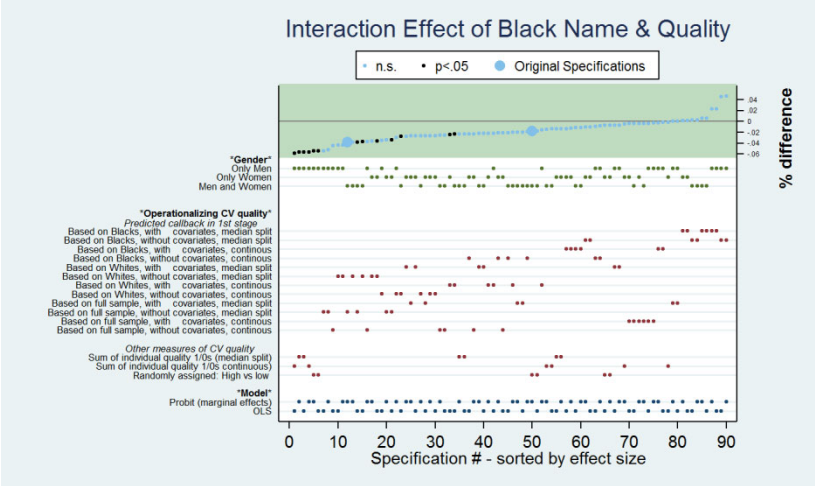


Figure S4. Descriptive Specification Curve – Interaction of distinctively black name and resume quality, on call back rate.



Supplement 4. The benefits of bootstrapping in specification-curve analysis.

To illustrate the robustness to misspecification in specification-curve analysis, we conducted Montecarlo simulations estimating Poisson regressions, because they are known to have inflated false-positive rates when assumptions are not met.

Specifically, we generated data where there is a true linear relationship between *y* and x, but the analyst does not observe *x*, and observes instead five alternative proxies: x_1, x_2, x_3, x_4, and x_5. Each is correlated r=.85 with the latent variable *x*. The dependent variable consists of count data (see Figure S5), but which does not follow a Poisson distribution.

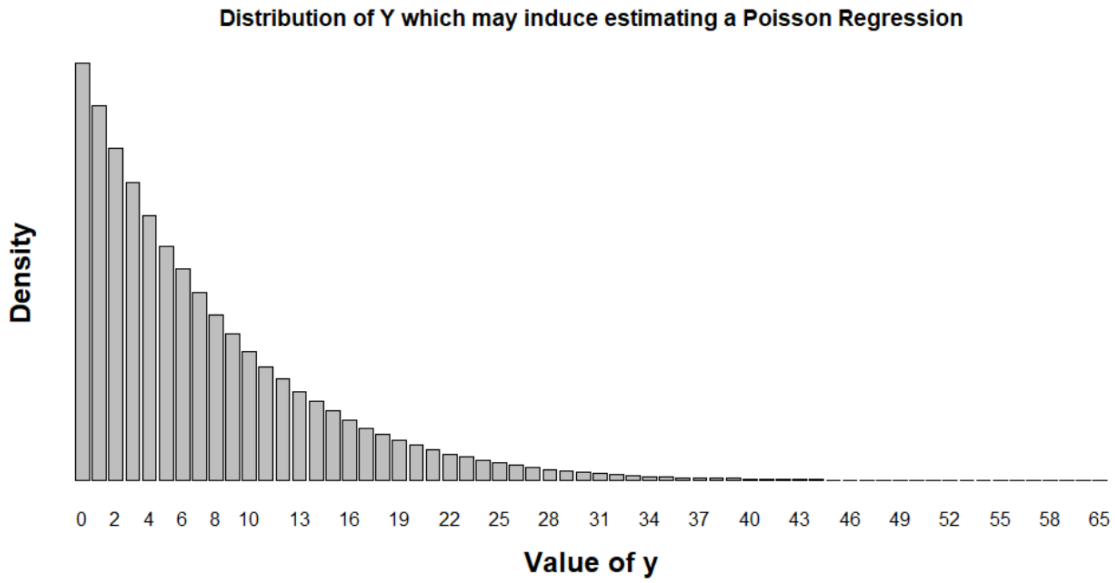


Figure S5. Distribution of y-variable in Montecarlo simulations later analyzed via Poisson Regression.

We consider an analyst who performs Poisson regression on each of the proxies separately, 5 specifications, and then combines them into a specification-curve analysis. Beginning with a scenario where the true effect of the latent variable on y is zero, thus it

is also zero for each of the five proxies, should result in a 5% false-positive rate, and this should also occur with specification-curve analysis overall, a 5% false-positive rate.

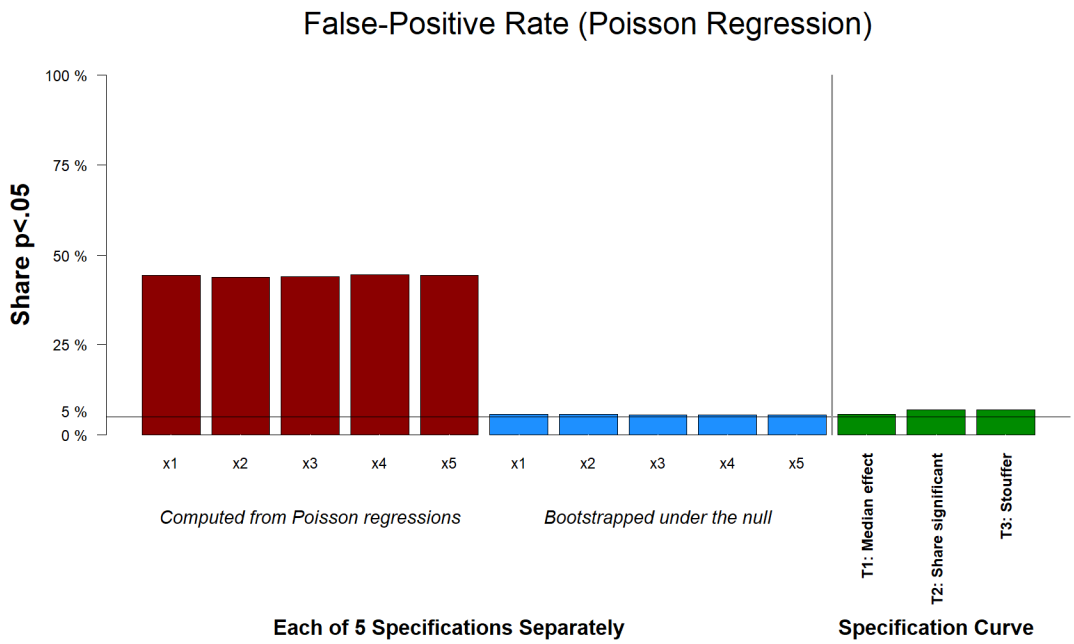


Fig S6. False-positive rate with Poisson regression, bootstrapped Poisson regression, and specification-curve analysis.

The first five bars in Figure S6 show that Poisson regressions have an elevated false-positive rate for these data, because they are not actually distributed Poisson, rejecting the null over 40% of the time for each proxy, due to the violation of assumptions. The next five bars shows that if *p*-values are computed via bootstrapping, bootstrapping those same Poisson regressions, the nominal 5% false-positive rate is attained. The next three bars show that combining *the red bars*, via each of the three test-statistics proposed for specification-curve analysis, the False-Positive Rate remains at or just slightly above 5%. This demonstrates that submitting a set of specifications to specification-curve analysis, some of which may have inflated false-positive rates due to un-met assumption, does not lead the overall test, which relies on bootstrapping, to be inflated.

We next consider power, returning to the same scenario as before, but now the latent variable has an effect on the dependent variable, and thus each of the 5 proxies is truly correlated with it as well and thus we expect to reject the null more often than 5% of the time.

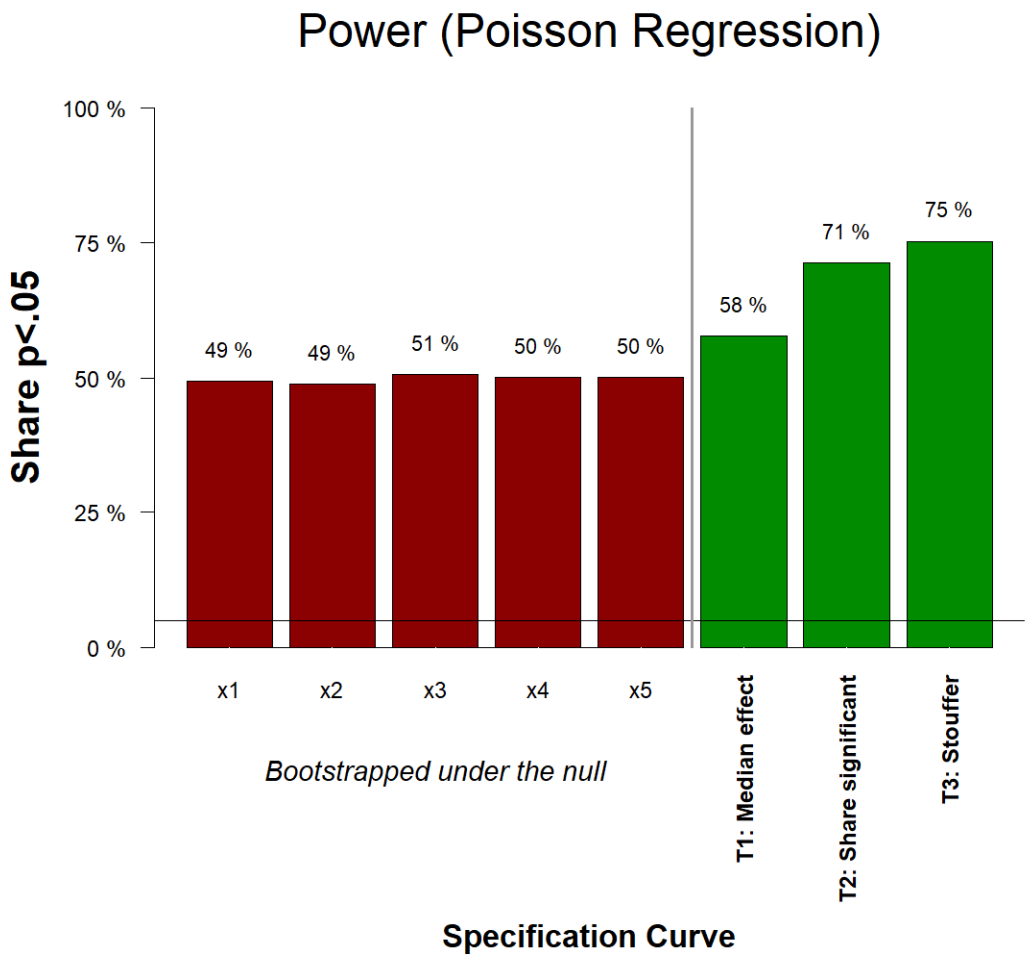


Fig. S7. *Power for individual specifications and the joint specification-curve analysis*

Figure S7 shows that test statistics in specification-curve analysis have more power than the underlying individual specifications, especially so the Stouffer test which combines all *p*-values using the Stouffer method, an then assesses significance by comparing the observed Stouffer value to the bootstrapped ones.

In short, this supplement demonstrates the value of Bootstrapping in specification-curve analysis, serving two purposes. First, it corrects for many consequences of specification errors that may inflate the false-positive rate. Second, it allows conducting joint inference across specifications, enhancing their power.

References

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991-1013.

Czajkowski, J., Simmons, K., & Sutter, D. (2011). An analysis of coastal and inland fatalities in landfalling US hurricanes. *Natural hazards, 59*(3), 1513-1531.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician, 60*(4), 328-331.

Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. *Proceedings of the National Academy of Sciences*, 201402786.

Kahn, M. E. (2005). The death toll from natural disasters: the role of income, geography, and institutions. *Review of Economics and Statistics, 87*(2), 271-284.

Mullainathan, S. (2002). A memory-based model of bounded rationality. *Quarterly Journal of Economics, 117*(3), 735-774.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution, 1*(2), 118-122.

Toya, H., & Skidmore, M. (2007). Economic development and the impacts of natural disasters. *Economics Letters, 94*(1), 20-25.