Effects of Top Coding and Aggregation

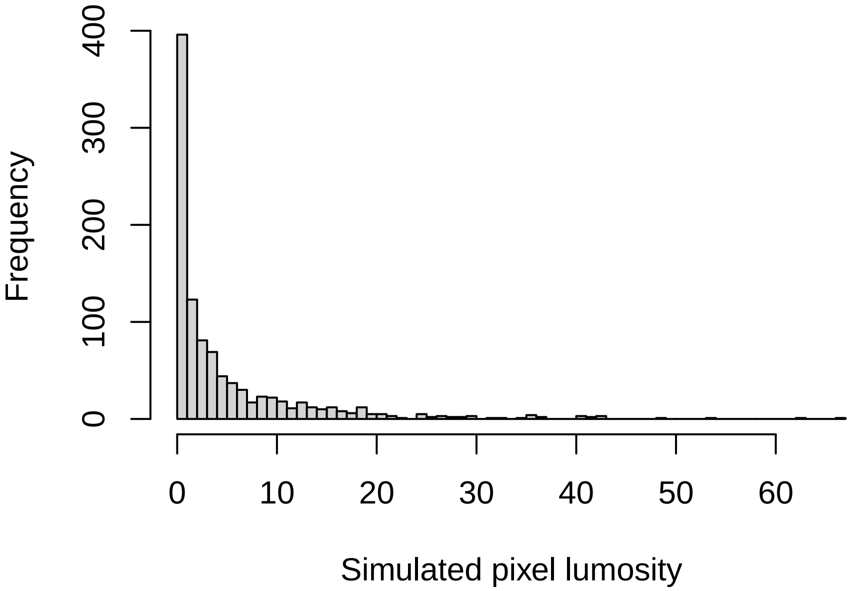
George Yang

2/2/2022

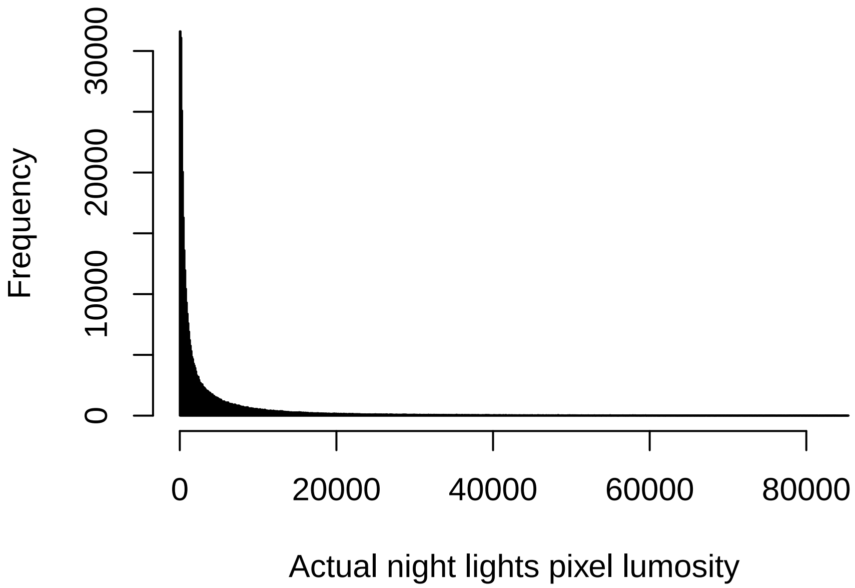
We know that one big difference between DMSP and VIIRS data is that DMSP pixels are top-coded, while VIIRS data is not.

Now, to explore how this might affect regressions, we can run a simulation with top coded data and without top coded data (with classical independent normally distributed measurement error).

So, first, we create our night lights variable. We get 1000 observations from random draws of a gamma distribution. The distribution is fit on the VIIRS night lights using maximum likelihood estimation and reflect the long-tail of the night lights distribution. So, it basically looks like the night lights data, but on a different scale.



For reference, this is actually what VIIRs night lights look like:

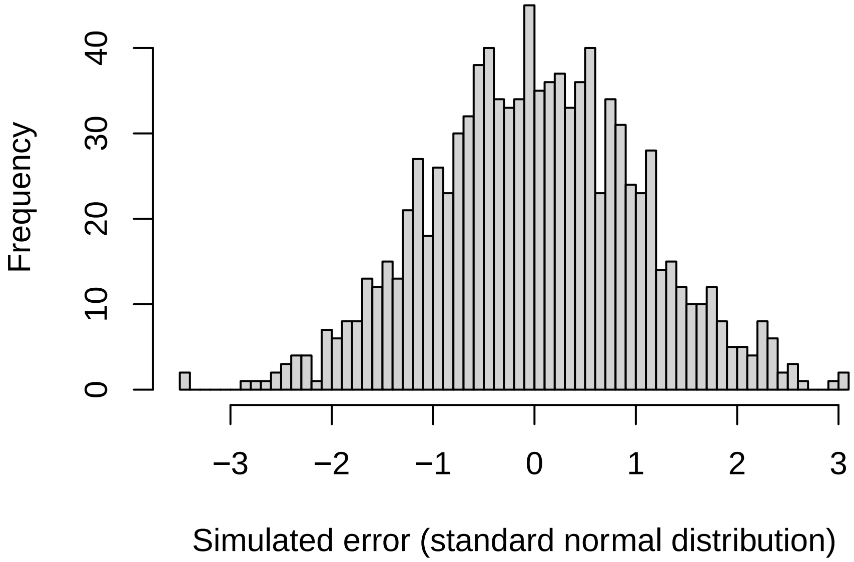


And here is the calculations that the computer did to get the shape and rate parameters.

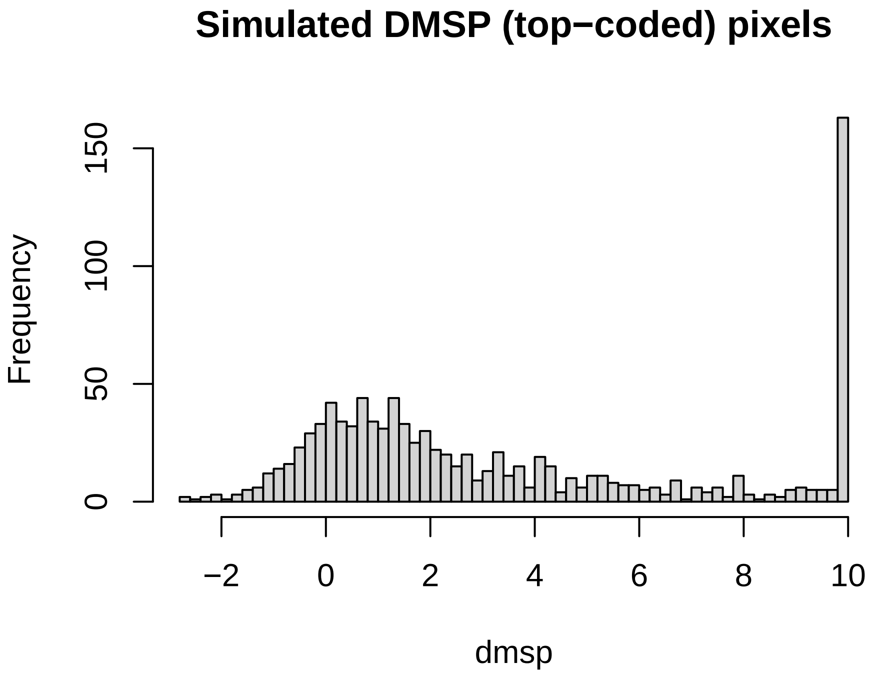
```
## Fitting of the distribution ' gamma ' by maximum likelihood
## Parameters :
##         estimate  Std. Error
## shape 0.46150185 0.005347860
## rate  0.08849264 0.001657559
## Loglikelihood:  -23714.81   AIC:  47433.61   BIC:  47448.03
## Correlation matrix:
##           shape      rate
## shape 1.0000000 0.6185178
## rate  0.6185178 1.0000000
```

```
## null device
##           1
```

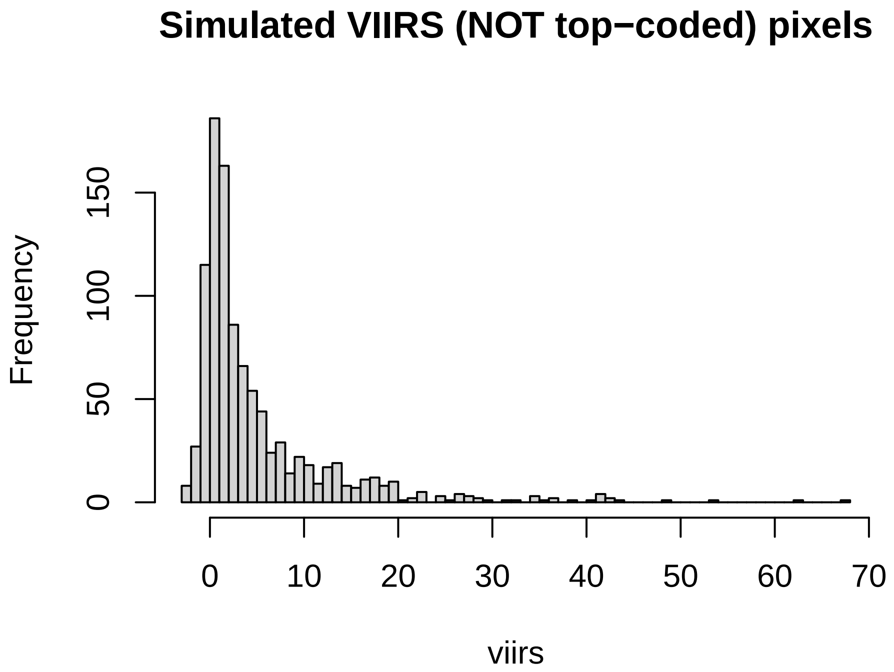
We introduce classical measurement error by taking 1000 draws of a standard normal distribution. The errors look like so.



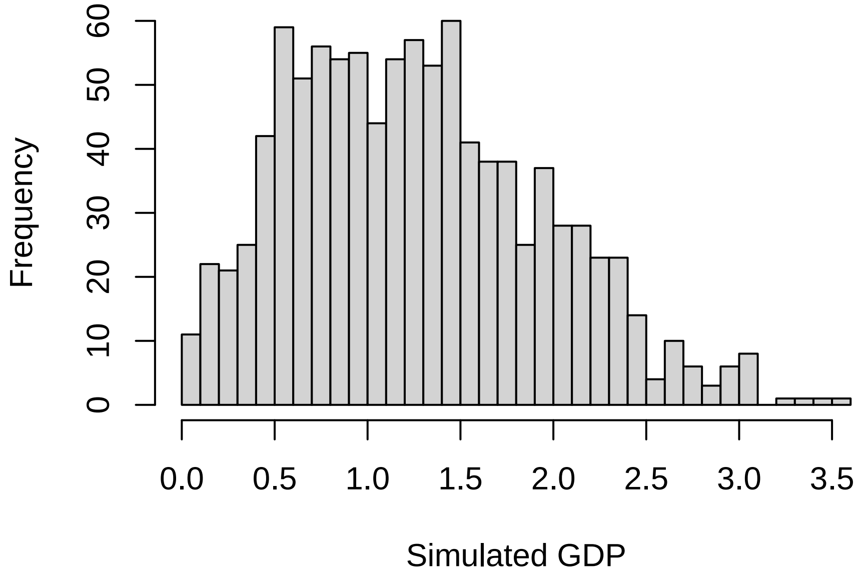
Adding the "true" night lights to the error term give us what's observed. For DMSP, we top-code values at 10.



For VIIRS, we do not.



And we code the true relationship between GDP and lights as just the night lights variable raised to 0.3 (i.e. Log GDP = 0.3*Log Lights).



	Model 1	Model 2
(Intercept)	1.979	2.052
	(0.112)	(0.125)
log(lights_area_viirs)	0.328	
	(0.071)	
log(lights_area_dmsp)		0.346
		(0.100)
Num.Obs.	100	100
R2	0.179	0.109
R2 Adj.	0.171	0.100
AIC	104.8	113.1
BIC	112.7	120.9
Log.Lik.	−49.424	−53.526
F	21.353	11.952

We can then use another random number generator to get us arbitrary groupings of these pixels (which represent countries). The number generator spits out 1000 numbers from 1 to 100 (with each number from 1 to 100 representing a separate country). The top of our dataset now looks like this.

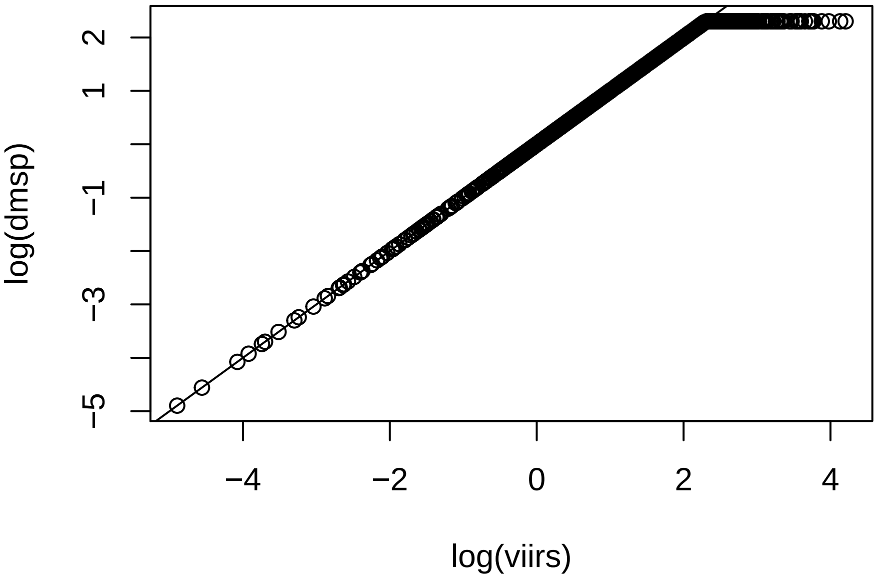
```
##     country       gdp      error       dmsp      viirs
##       <int>     <num>      <num>      <num>      <num>
##  1:      41 1.5643179  0.9167050  5.3604595  5.3604595
##  2:      54 2.1801524 -0.3646684 10.0000000 13.0719348
##  3:      97 0.6225621  1.0660761  1.2721117  1.2721117
##  4:      43 0.4193575  0.4452967  0.5004978  0.5004978
##  5:      75 1.3200111 -1.2553990  1.2676478  1.2676478
##  6:      10 2.0903236  1.5024891 10.0000000 13.1807557
##  7:      96 1.5872730  1.6464546  6.3113163  6.3113163
##  8:      19 1.1323667 -1.1755210  0.3378903  0.3378903
##  9:      86 1.7536728  1.4350762  7.9387970  7.9387970
## 10:      21 1.1724958  0.1025502  1.8022452  1.8022452
## 11:      73 0.5398514 -2.5513814 -2.4232720 -2.4232720
## 12:      69 1.4117273 -0.4756698  2.6805657  2.6805657
## 13:      79 2.0263576 -0.8386644  9.6903315  9.6903315
## 14:      42 1.1502362  0.6858586  2.2803548  2.2803548
## 15:      79 0.1088872 -0.1952031 -0.1945866 -0.1945866
```

We can collapse values by country. And finally, we run a regression of log GDP on the left hand side and log lights per area on the right hand side, with the top coded variable and the non-top-coded lights variable.

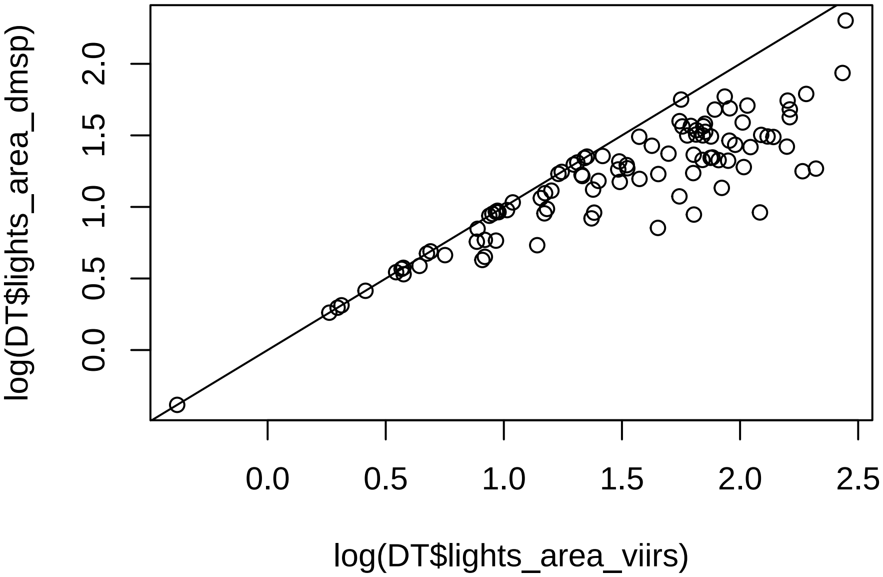
And oddly enough, we are able to replicate the result from the paper—the coefficient on the top-coded simulated night lights (DMSP) is greater than that of VIIRS. This is exactly what we find in the pre-2013 period.

Moreover, if we plot the non-top coded lights variable and the top coded lights variable, we see a broad corresspondence after we've aggregated things above the pixel level. This is exactly what we find in the country comparisons between DMSP and VIIRS.

This is what it looks like if we plot log-log of *pixels*.



This is what it looks like if we plot log *aggregate* pixels on log *aggregate* pixels.


log(DT$lights_area_dmsp)
log(DT$lights_area_viirs)