Statistical Theory behind Models in "Power to Decarbonize" Report

Arun Ramamurthy | Environmental Progress

March 13, 2018

This document will review the assumptions and allowances of the models used in Environmental Progress' "Power to Decarbonize" report, which used historical data of 65 nations over 51 years to characterize the factors that affect `Carbon Intensity of Energy`, a metric for a country's progress towards deep decarbonization.

To provide context, the design matrix of the models in the report is displayed below. Each row represents a single country's electricity generation system and carbon intensity in a given year.

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    3563 obs. of  7 variables:
##  $ Year                                          : num  1965 1966 1967 1968 1969 ...
##  $ Country                                       : chr  "Algeria" "Algeria" "Algeria" "Algeria" ...
##  $ Carbon Intensity of Energy (g/kWh)            : num  229 235 231 229 231 ...
##  $ Solar Electricity Generation per Capita (MWh) : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Wind Electricity Generation per Capita (MWh)  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Nuclear Electricity Generation per Capita (MWh): num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Hydro Electricity Generation per Capita (MWh) : num  0.0311 0.0279 0.0304 0.0405 0.0256 ...
```

We have our response variable `Carbon Intensity of Energy (g/kWh)`, and several possible explanators. The predictors used in the report were `Electricity Generation per Capacity (MWh)` for each of `Solar`, `Wind`, `Nuclear`, and `Hydro`.

Of note, the models used in the report are simple non-parametric generalized additive models, which circumvent several of the stringent assumptions of linear regression. However, in light of interpretability, this follow-up report will provide four simple linear models for `Carbon Intensity of Energy`, and even show that under these more strict conditions, the assumptions still by-and-large hold. Note that all the models in the report are *simple* - that is, they each only ascertain the effect of one of the four electricity predictors on `Carbon Intensity of Energy`. Furthermore, given our evaluation of the variance-inflation factor for each of these predictors, there is a relatively low level of coliniearity between these four predictors - in real life too, deployment of each of these technologies are formed by orthogonal policies, although some energy policies may deploy sets of these technologies in conjunction.

First, we regress CIE on each of these electricity generation predictors, and display the results.

```
## $Solar
##
## Call:
## lm(formula = `Carbon Intensity of Energy (g/kWh)` ~ `Solar Electricity Generation per Capita (MWh)`,
##     data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -138.19  -23.55    5.51   21.27  104.68
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                     204.668      1.654
## `Solar Electricity Generation per Capita (MWh)`  -12.546     23.354
##                                                t value Pr(>|t|)
## (Intercept)                                    123.707   <2e-16 ***
```

```
## `Solar Electricity Generation per Capita (MWh)`  -0.537    0.591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 46.67 on 896 degrees of freedom
## Multiple R-squared:  0.000322,   Adjusted R-squared:  -0.0007937
## F-statistic: 0.2886 on 1 and 896 DF,  p-value: 0.5913
##
##
## $Wind
##
## Call:
## lm(formula = `Carbon Intensity of Energy (g/kWh)` ~ `Wind Electricity Generation per Capita (MWh)`,
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.459  -26.580    2.217   27.899  100.461
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                     208.889      1.590 131.357
## `Wind Electricity Generation per Capita (MWh)`  -19.068      5.328  -3.579
##                                                Pr(>|t|)
## (Intercept)                                     < 2e-16 ***
## `Wind Electricity Generation per Capita (MWh)`  0.00036 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.35 on 1133 degrees of freedom
## Multiple R-squared:  0.01118,    Adjusted R-squared:  0.0103
## F-statistic: 12.81 on 1 and 1133 DF,  p-value: 0.0003599
##
##
## $Hydro
##
## Call:
## lm(formula = `Carbon Intensity of Energy (g/kWh)` ~ `Hydro Electricity Generation per Capita (MWh)`,
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -113.891  -23.926   -5.359   24.818  104.340
##
## Coefficients:
##                                                Estimate Std. Error
## (Intercept)                                    231.6468     0.7606
## `Hydro Electricity Generation per Capita (MWh)`  -8.1449     0.1982
##                                                t value Pr(>|t|)
## (Intercept)                                     304.57   <2e-16 ***
## `Hydro Electricity Generation per Capita (MWh)`  -41.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 38.43 on 2919 degrees of freedom
## Multiple R-squared:  0.3664, Adjusted R-squared:  0.3662
## F-statistic:  1688 on 1 and 2919 DF,  p-value: < 2.2e-16
##
##
## $Nuclear
##
## Call:
## lm(formula = `Carbon Intensity of Energy (g/kWh)` ~ `Nuclear Electricity Generation per Capita (MWh)
##     data = .)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.447  -21.927    2.926   24.817   77.070
##
## Coefficients:
##                                                  Estimate Std. Error
## (Intercept)                                       237.211      1.323
## `Nuclear Electricity Generation per Capita (MWh)`  -17.257      0.562
##                                                  t value Pr(>|t|)
## (Intercept)                                       179.26   <2e-16 ***
## `Nuclear Electricity Generation per Capita (MWh)`  -30.71   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.33 on 1255 degrees of freedom
##   (23 observations deleted due to missingness)
## Multiple R-squared:  0.429,  Adjusted R-squared:  0.4286
## F-statistic:   943 on 1 and 1255 DF,  p-value: < 2.2e-16
```

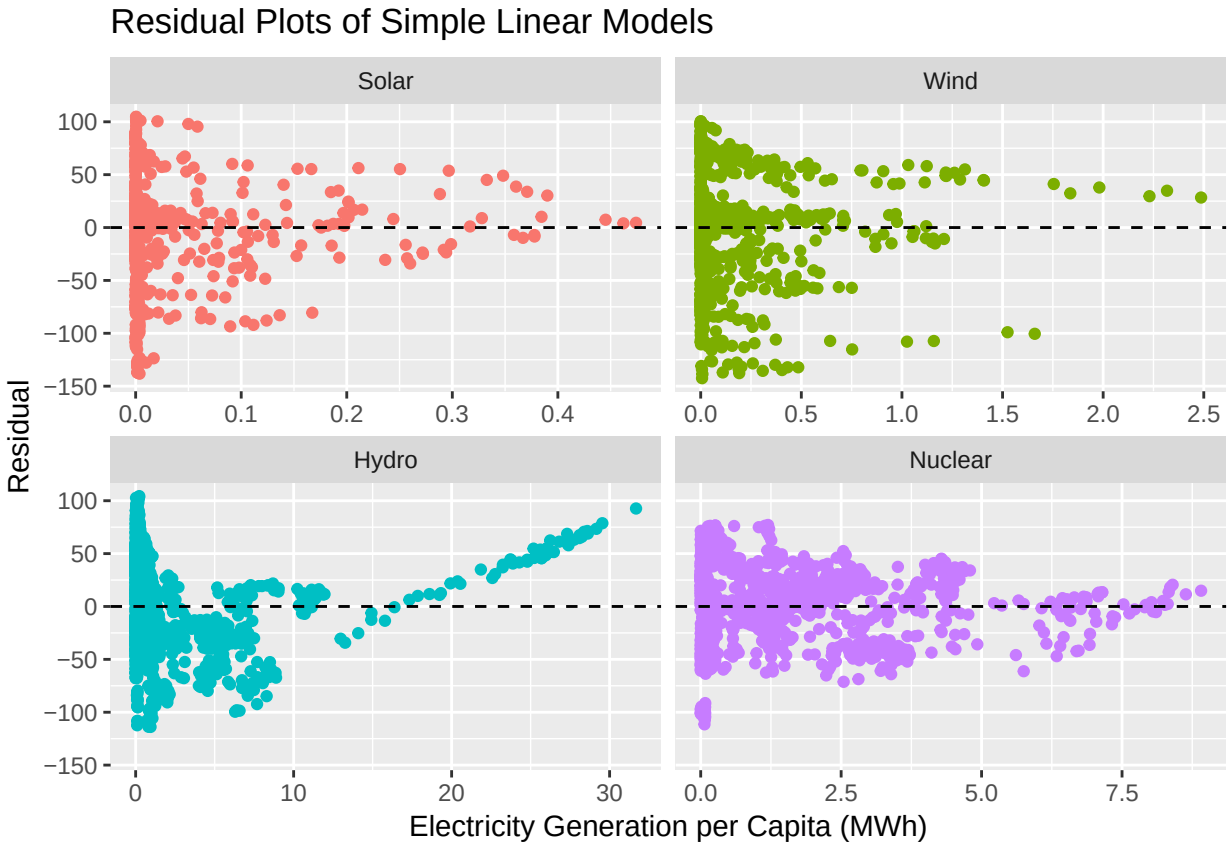
A summary of the slope estimates is also given below.

```
##           Model   estimate  std.error   statistic p.value
## 1   CIE ~ Solar -12.545737 23.3536443  -0.5372068 0.59126
## 2    CIE ~ Wind -19.067755  5.3281573  -3.5786772 0.00036
## 3   CIE ~ Hydro  -8.144914  0.1982403 -41.0860617 0.00000
## 4 CIE ~ Nuclear -17.257098  0.5619600 -30.7087641 0.00000
```

Some notes about these models: although `Wind` and `Solar` have larger absolute slopes, they also have larger standard errors, and particularily for `Solar`, we do not have confidence in that these slope estimates are non-zero. Furthermore, the models for `Solar` and `Wind` have R^2 of 0.0003 and 0.01 respectively, much smaller than the R^2 for `Hydro` and `Nuclear` (0.37 and 0.43 respectively).

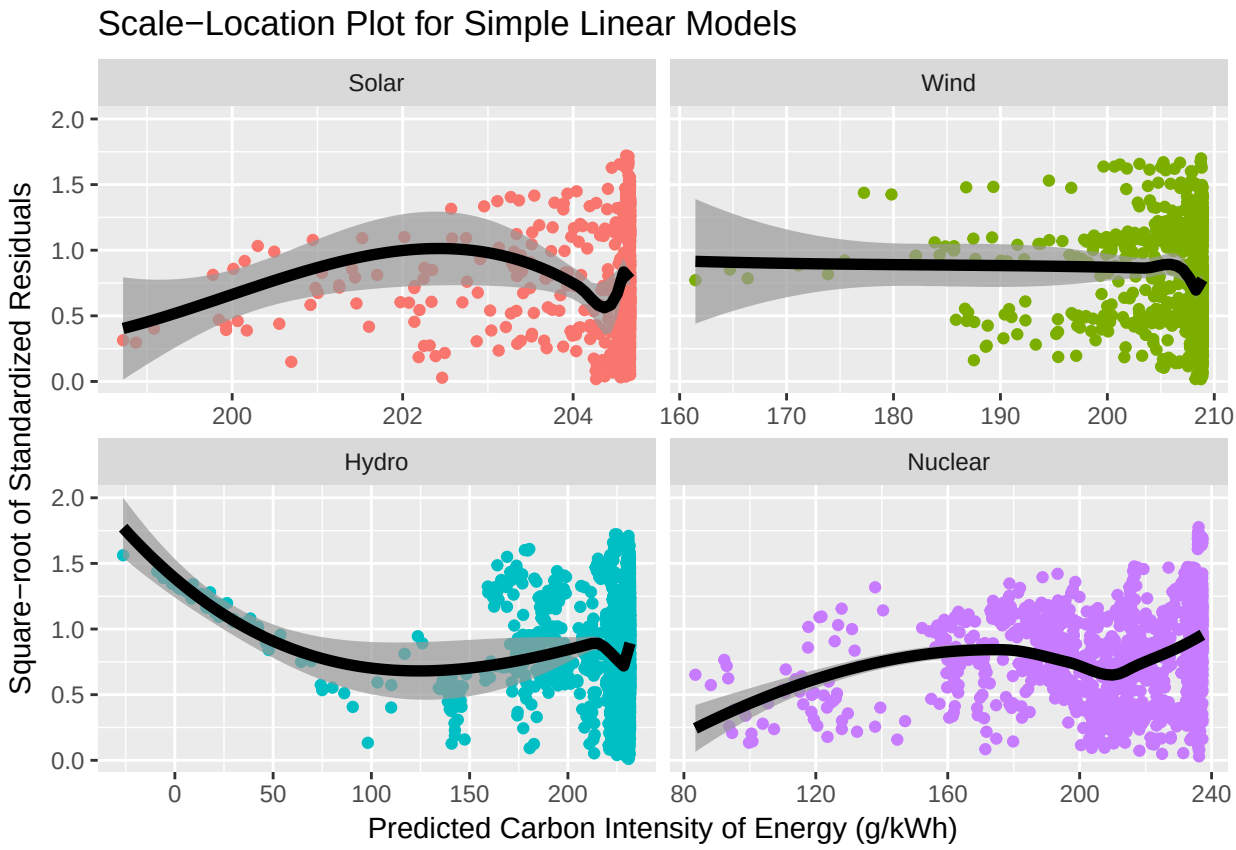
Given the strength of `Nuclear` as a predictor, as detailed in our report, we suspect that there may be fundamental properties about nuclear energy that cause nuclear energy deployment to be a more precise policy lever for carbon intensity than wind or solar energy. However, the models discussed in this document are purely descriptive, and do not extend into causality arguments nor statements about the nature of these technologies.

We will now verify the assumptions of linear regression on each of these models. First, we evaluate the **linearity** assumption by plotting a residual plot for each model:



With the exception of Norway for hydroelectricity (seen as the sharp 45° line in the `Hydro` facet), the linearity assumption holds for all four of these models.

The second assumption of linear regression is **constant variance** of the errors. This assumption is a bit more stringent, and difficult to fulfill with this data due to the wide spread of CIE for countries with near-zero deployment of some energy source, which cause the residuals for near-zero `X` to be larger than the residuals for the rest of the feature space. However, disregarding near-zero values in our predictors, the residual bands seen above are roughly constant throughout. We can verify this with a scale-location plot for each model, as shown below.



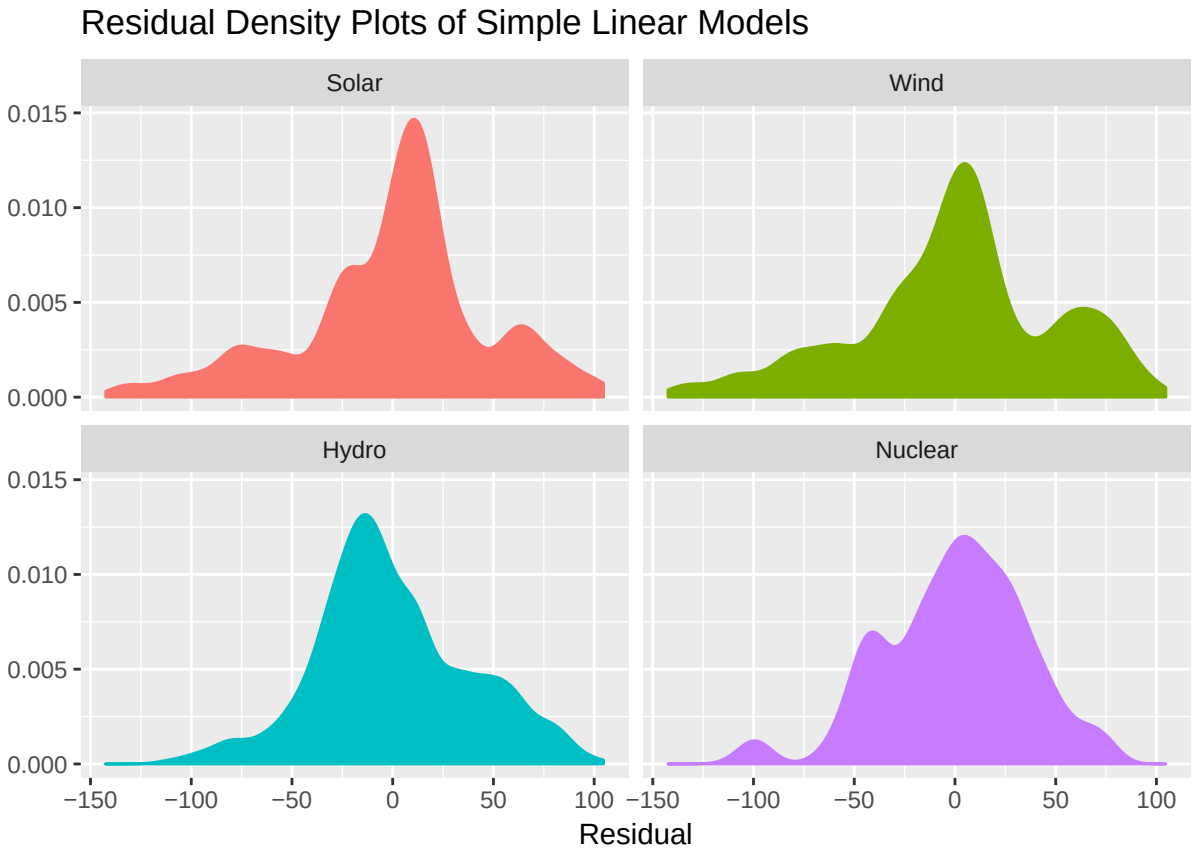
From the scale-locations plot, I suspect that outlier countries (like Norway, Sweden, Denmark, and Germany) do not fit into our simple linear models as well. Current energy analytics research at Environmental Progress is partially focused on the Scandanavian interconnected grids, as we wish to investigate the nature of these outlying cases. Additionally, we see that the uneven spread of our predictor variables, both within themselves and between energy sources, are causing some artifacts in our linear models. However, within the IQR of our predictors, our model seems to fit quite well, and there is evidence of homoskedasticity.

Strictly speaking, however, the following test shows that we are reasonably certain of heteroskedasticity for `Nuclear` and `Hydro`, if we take into account these outlier countries. Further research is required to determine *how* exactly variance in our model errors is related to large-scale deployments of `Nuclear` and `Hydro`.

```
## $Solar
##          BP
## 0.003555062
##
## $Wind
##       BP
## 0.598099
##
## $Hydro
##           BP
## 7.358042e-05
##
## $Nuclear
##           BP
## 3.015168e-15
```

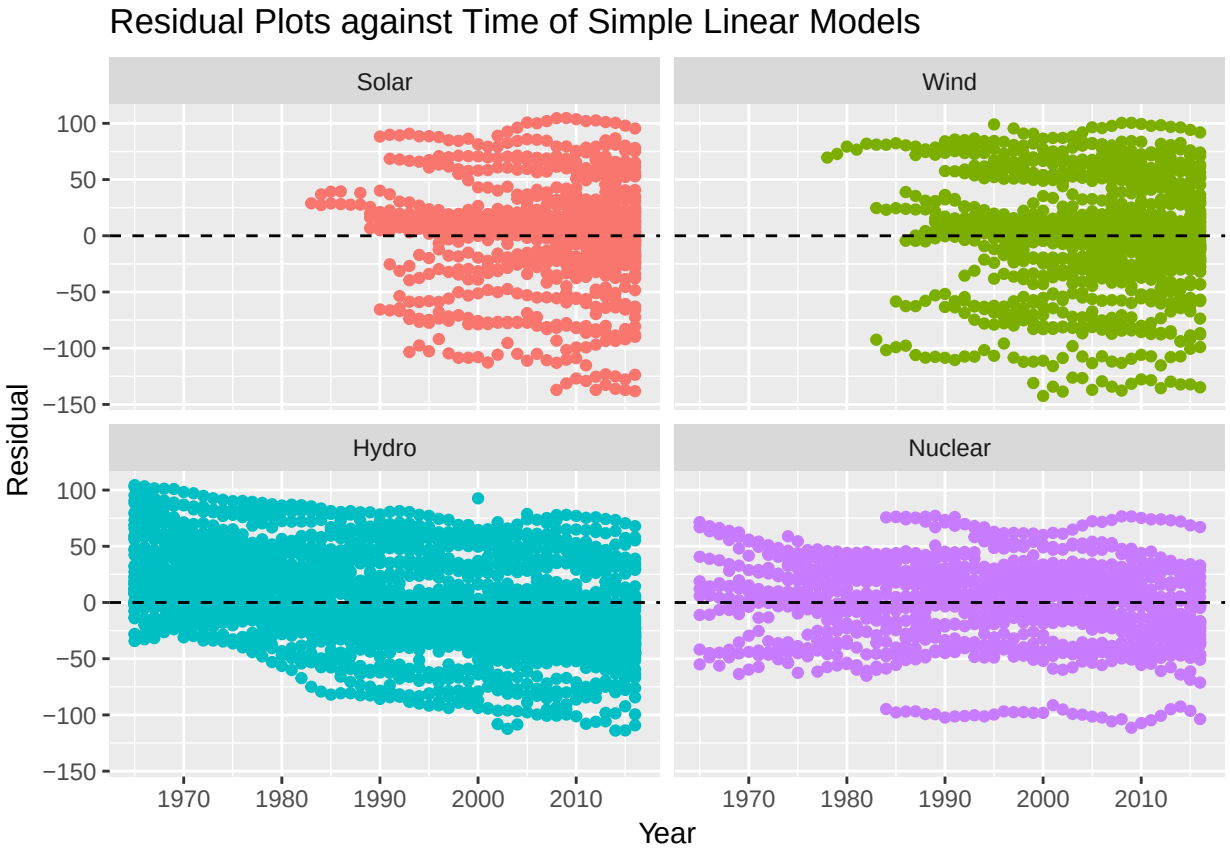
The final assumption of linear regression is that the error terms are **identically & independently normal**.

We can verify the normality assumption first by graphing a density plot of the residuals of each simple model.



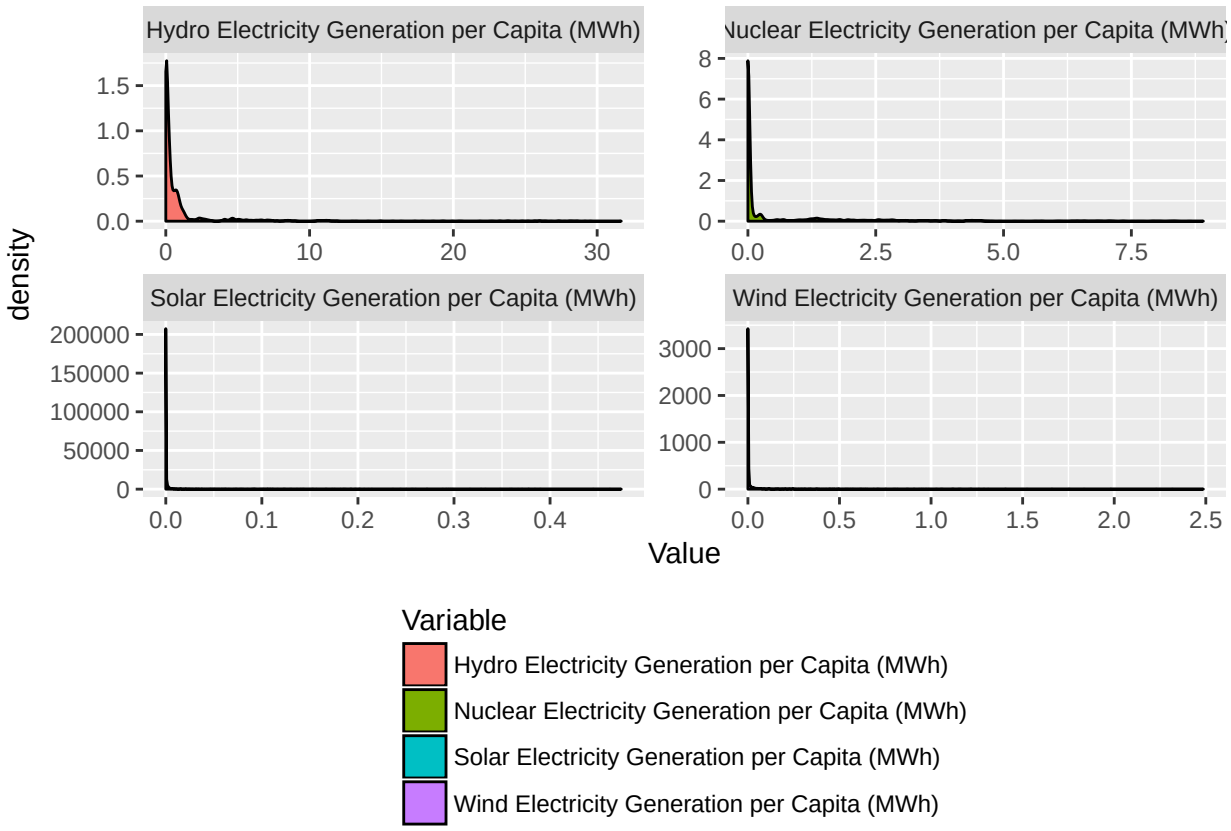
The error terms are in fact close to normally-distributed for each model, and the plot shows nearly consistent variance between them. In our residual plots above, we also see that the errors are independent of X for the most part, with the previously noted exception of some outlier countries.

Finally, we wish to test if the errors are independent of each other. Due to the time-series nature of our data, this assumption will not strictly hold. However, graphs of residuals over time show that the variance does not shift much over time, and we further note that we can fix this particular deviation from the assumptions of linear regression if we included `Year` as a predictor in each of the simple models.



Finally, a couple of *desirable* properties of our data is a nice spread to our explanatory variables, and lack of colinearity in our predictors.

Unfortunately, as the density plots below indicate, our data is heavily right-skewed - however, we do have a large range in explanatory outcomes, so they characterize the effects of a sizeable portion of the feature space. Having tried various Box-Cox transformations, I wasn't able to find a single interpretable transformation for all four predictors. Secondly, as stated above, investigation into the variance-inflation factors of each predictor show only slight colinearity, in the case of `Solar` and `Wind`, with the other predictors, and lack of colinearity for `Hydro` and `Nuclear`.



This concludes our analysis of the validity of simple linear models as applied to the central problem detailed in the "Power to Decarbonize" report. Please send any questions to this author's inbox or via Twitter.