Solar Panels Analysis

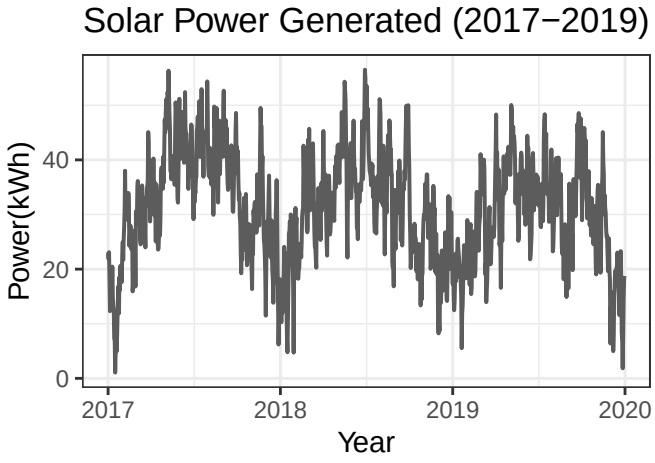
Carly Lundgreen

Introduction

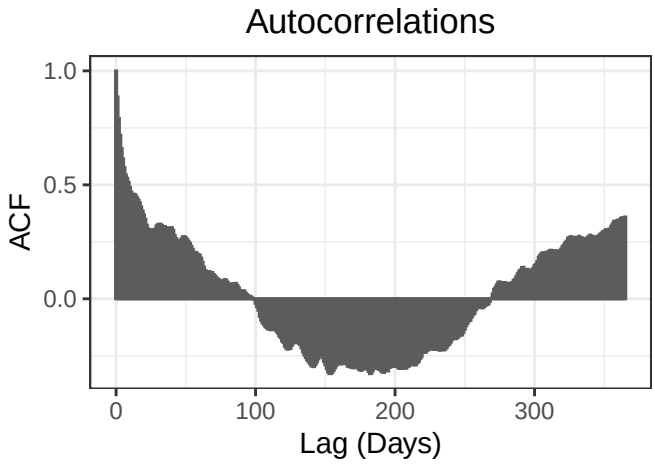
As the world moves further from fossil-fuel based sources of energy, research into the efficiency and viability of alternative sources of energy has become increasingly important. One particularly green source of energy is solar power, which is generated through solar panels. Solar panels are used everywhere from residential neighborhoods–solar panels can often be seen on rooftops in suburban areas–to massive solar farms that supply large quantities of energy to the power grid.

The power that is generated by solar panels, measured in kilowatt hours (kWh), can vary widely due to factors such as the time of year, angle of the panels, direction of the sun, cloud cover, age of the panels, etc. Solar panels degrade over time, which means that the power output decreases over time as well. We have a dataset of daily power generated from a single solar panel system on a single customer household. We will analyze this data to help the power company understand how these solar panels are degrading over time, discover approximately how many years it takes for the panels to lose 50% of their average power generating capability, and obtain projections of power for the following year.

This dataset has 1095 daily power measurements spanning three years: 2017, 2018, and 2019. We can see a clear seasonality to these data just from observing the time series plot below. Power generating capability appears to increase in the summer and decrease in the winter, on average. This is expected because there is much more sunlight in the summer than in the winter. Because solar power measurements are based on such a wide array of factors, a few of which were described above, it makes sense that there is a large variability in the individual power measurements. This is illustrated by the sharp increases and decreases in daily generated power. On average, however, we do see the clear seasonal trend of increasing average power in the warmer months and decreasing average power generated in the colder months.



A particularly obvious problem with this dataset is that daily power measurements are not independent of each other. This is something we will have to incorporate into a statistical model to be able to sufficiently understand how these solar panels are degrading over time and obtain accurate predictions for future years. The Auto-Correlation Function (ACF) plot below displays the extent to which present observations are correlated with past observations in these data. Along the x-axis is a time lag that allows us to see how correlated observations are when they are $t_1 - t_0$ days apart.



The seasonality discussed previously is very obvious in this plot. This dataset begins on January 1st, so it appears that the correlation between observations changes after about 100 days. This is about 3 months, which means that observations in the spring and summer are still related to observations in January, only negatively so. The largest takeaway from this ACF plot, however, is that observations at successive time points are correlated, and this correlation takes a long time to "die out." We will attempt to capture this slowly-decaying correlation in the statistical model that we will outline in the following section. If the correlation is not accounted for, the standard errors associated with parameter estimates will be too small, which means that inference procedures such as confidence intervals on these parameters will be inaccurate. Predictions will also not be as accurate, because we are ignoring an important component the distribution of the response variable.

Model Specification

In order to understand how the solar panels are degrading over time and predict future daily solar power measurements, we will fit a multiple linear regression model with a lag-1 autogressive covariance structure. The selected model is outlined below. We will compare this model to a linear model that includes Time as the only covariate and does not account for any correlation between successive observations, i.e. $y_i \sim N(X\beta, \sigma^2 I)$. In following sections, we will outline where this linear model breaks down in the assumptions and how our model takes care of those issues. We will also compare evaluation metrics such as AIC, RMSE, and bias for the two models.

The following model was fit to the solar panel data:

$$y_i = \beta_0 + \beta_1 x_1 + \ldots + \beta_{12} x_{12} + \epsilon,\ \epsilon \sim N(0, \sigma^2 R)$$

Where y_i are daily solar power measurements taken from an individual home from the start of 2017 to the end of 2019. The variable x_1 is a numeric variable representing the date (year, month, and day) that the i^{th} observation occurred. The rest of the columns of the design matrix $\mathbf{X}$ $(x_2, \ldots, x_{12})$ are categorical variables indicating the month of the i^{th} observation such that x_2 corresponds to whether the observation was taken in February, x_3 corresponds to whether the observation was taken in March, and so on through December.

The design matrix is linear in the parameters $\beta_0, \ldots \beta_{12}$. Contextually, β_0 indicates the expected average daily power generated by these solar panels in January–which is the chosen "baseline" or comparison month for this particular model–and at time zero. Time equaling zero admittedly is impossible and uninterpretable, so we will refrain from further discussion of this parameter.

We interpret β_1 as the expected average daily increase in power generated by these panels, holding all else constant. Because $\beta_2, \ldots, \beta_{12}$ are categorical indicators for each month, with January the default comparison

month, they all have similar interpretations. For example, β_2 indicates how much more daily power in kWh is generated in February on average compared to January (holding all else constant) and β_3 indicates how much more daily power is generated in March on average compared to January, again holding all else constant. The remaining $\beta_i's$ are interpreted similarly.

Regarding the uncertainty of this model, ϵ is a 1xn vector of differences between the actual solar power measurements and the power measurements predicted by the model for observations at times $t = 1, \dots n$. We also denote σ^2 as the residual variance, or the variance between observations. Essentially, σ is the amount that we expect the model to differ from the true solar power measurements on average.

$\mathbf{R}$ is an nxn matrix where the ij^{th} elements are based on $\rho(t_i, t_j)$, or the correlations between solar power measurements at time points i, j. We define $\rho(\epsilon_{t_i}, \epsilon_{t_j}) = \phi^{|t_1 - t_2|}$ so that the covariance matrix $\mathbf{R}$ has the following structure:

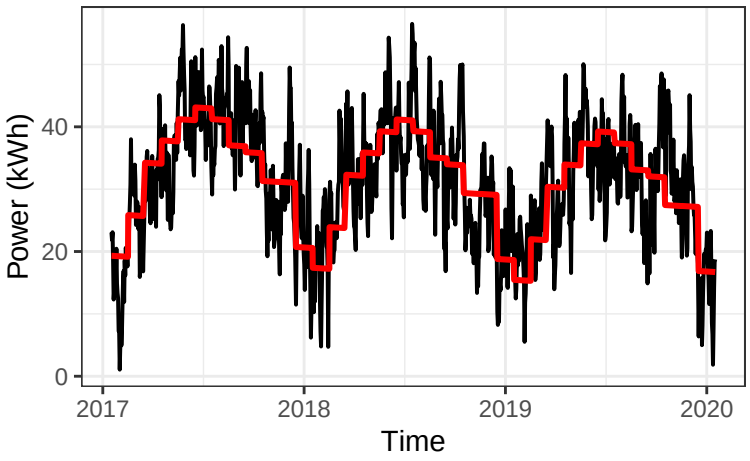
$$\begin{pmatrix} 1 & \phi & \phi^2 & \dots & \phi^{t-1} \\ \phi & 1 & \phi & \vdots & \phi^{t-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \phi^{t-1} & \dots & \dots & \ddots & 1 \end{pmatrix}$$

This covariance structure is referred to as a Lag-1 Autoregressive (AR1) process where measurements are correlated in time acccording to a parameter ϕ, but the "further out" in time the measurements go, the correlation between observations decreases according to $\phi^{|t_1 - t_2|}$, for $\phi \in \{-1, 1\}$.

In this scenario, we assume that the diagonal elements of $\mathbf{R}$, $\phi^{|t_i - t_i|}$, are equal to 1. We can make this assumption because the observations in this dataset were taken at uniformly varying time points (i.e. one measure of solar power taken per day). If we had a dataset with non-uniform time measurements, it may be possible to have more than one measurement of the response variable taken at the same time. In this situation, we would want to account for potential sampling variability between these "frozen time" measurements where the correlation is not necessarily equal to 1. This would introduce an alternative covariance structure that is different from the AR1 structure that we have chosen. It is important to recognize that our chosen structure works because we have uniform time measurements.

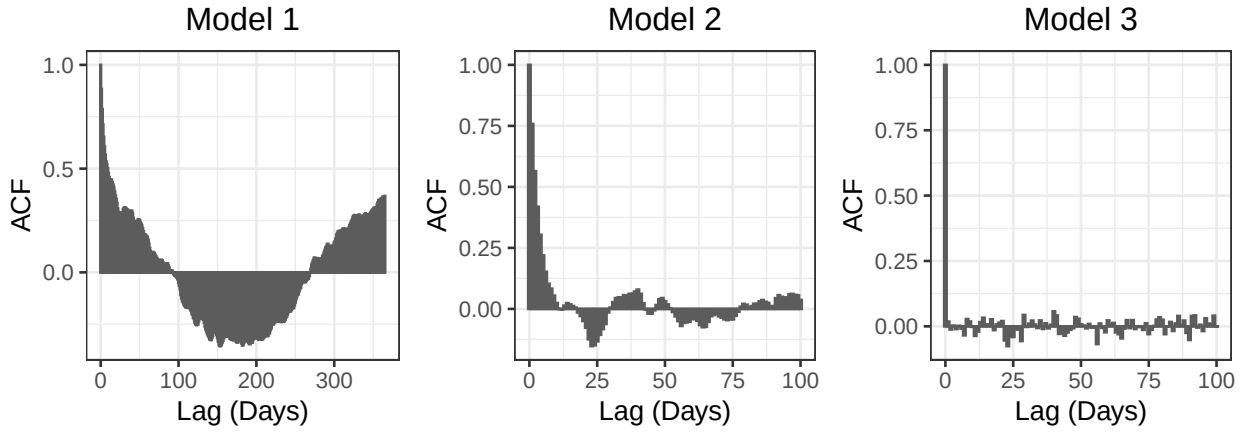
Model Justification

This model assumes that the residuals follow a normal distribution and that residuals vary equally about zero according to the constant variance σ^2. We are building the assumption of dependence into the model, so we depart from the typical linear regression assumption of independence between observations. First, however, we will plot the fitted values from our model aover the top of the observed solar power measurements. This plot is below.

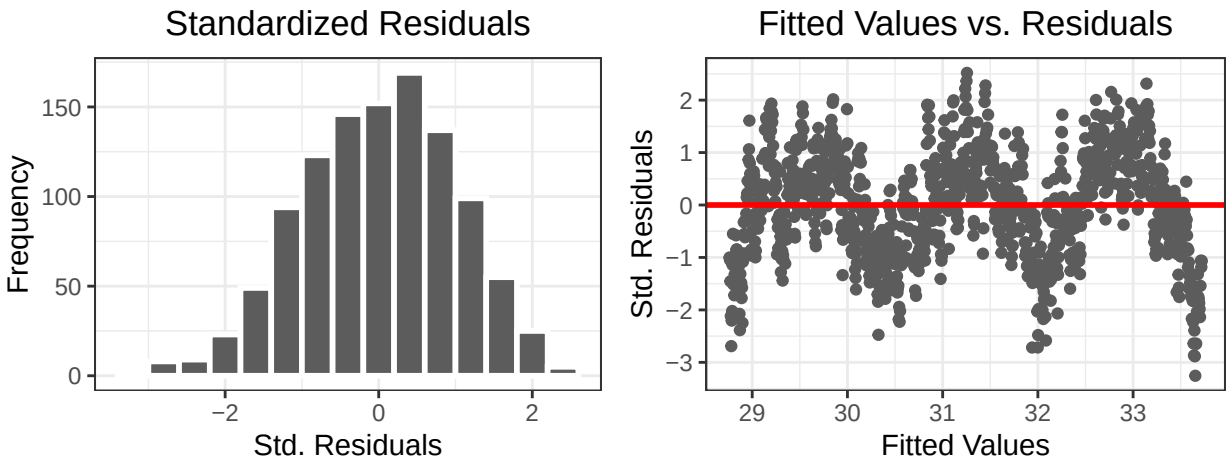


We can see that our model, which is a step function that changes month-to-month, follows a similar undulating pattern that is seen in the solar power measurements. The model is predicting higher daily generated power in the spring and summer, which is just what we observe in the data.

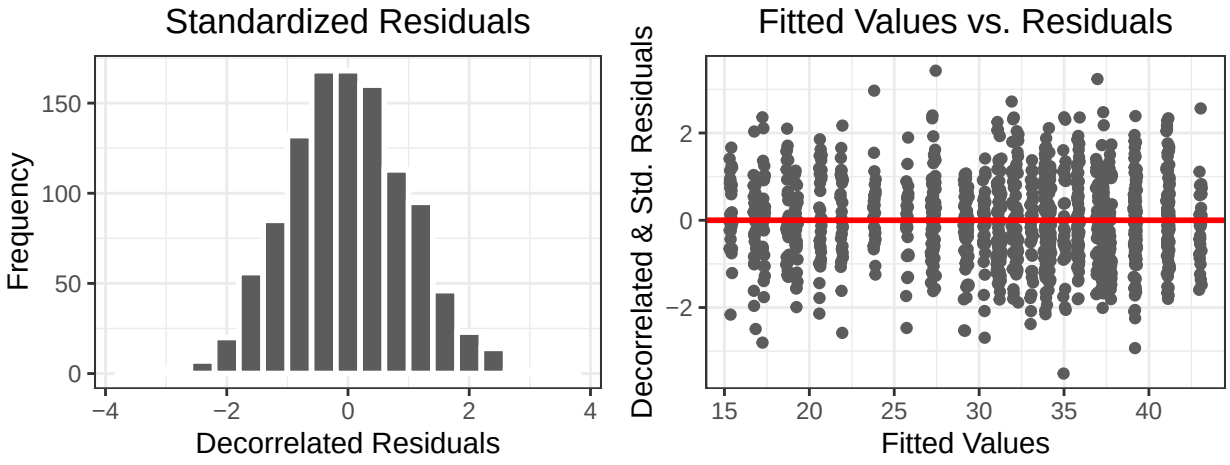
In order to understand how our model captures the correlation in these data, we have displayed ACF plots for three models below. Model 1 is the ACF plot for the linear model with Time as the only covariate that does not account for any correlation. It is clear that successive days and months are correlated with previous days and months when using this model, which is expected since no correlation structure was incorporated. Model 2 was fit with the same design matrix as our selected model, but no correlation structure was included. It is clear that simply including an effect for month accounted for much of the correlation between observations, but there is still some lingering correlation effect that needed to be accounted for. Model 3 is our chosen model, outlined in the previous section. The ACF plot from our model shows that present observations are now only correlated with themselves and any lingering correlation over time has been explained by our AR1 structure.



Now that we know that the residuals have been sufficiently "decorrelated" (i.e. the correlation sufficiently captured by our model), we will assess the assumptions of normality and equal variance among these decorrelated residuals. First, however, displayed below is a histogram of the standardized residuals from the simple linear model along with a plot of the fitted values vs. standardized residuals for this simple model. The histogram has a slight left skew and it is difficult to tell whether the assumption of normality is broken by this model, but there is a clear and obvious pattern in the fitted values vs. residuals plot. This is the same undulating pattern seen in the data, and is clear evidence that the linear model does not sufficiently meet the assumption of equal variance among the residuals.



Contrasting the above plots with the two plots displayed below, we can see a clear improvement in the fitted values vs. residuals plot. The standardized (and decorrelated) residuals appear to vary constantly around zero, and the histogram of standardized residuals appears sufficiently normal to conclude that our model has met these assumptions.



Model Performance Evaluation

In order to assess model fit, we calculated a "Pseudo R^2" using the following formula: $R^2_{pseudo} = Cor(y, fitted(mod))^2$. This value turned out to be 0.526. Ideally, this would be closer to 1, if we interpret this as a typical R^2 value, which is the percent of variation in the response variable explained by the model. Most likely this metric for model fit is catching on to the step-function aspect of our model. The variation in the solar power measurements may be better explained by a more smooth function. However, splines are not ideal due to the predictive goals of this analysis because splines do not behave well beyond the range of the observed data. We will assess predictive accuracy to get a better idea of how our model answers the prediction questions posed at the beginning of this analysis.

We calculated AIC to begin assessing the predictive accuracy of our model, and the AIC was found to be 6388.822. This was much lower than the AIC for a simple linear model, which was found to be 8160.493. AIC is an estimator of prediction error (lower AIC quantities are better) and is mostly used as a means of model selection. If it wasn't already clear before, our model is much better at predicting than an SLR with a single Time covariate.

The table below compares predictive accuracy metrics calculated from 30 cross-validated studies using a training set of 80% of the data. We are reporting the cross-validated root mean square error (RMSE), bias, average prediction interval coverage (i.e. what percent of prediction intervals actually contain the true value), and average prediction interval width for our model and for a simple linear regression (SLR) model that includes Time as the only covariate. We denote our model as "GLS" in the below table because our inclusion of a correlation structure required us to work in the generalized least squares framework rather than ordinary least squares.

	RMSE	**Bias**	**Coverage**	**Width**
GLS	3.714	0.016	0.971	16.170
SLR	10.029	-0.066	0.957	39.416

The RMSE of our model is clearly lower than that of the SLR model. Our RMSE of 3.7, indicating we are off in predictions by 3.7 kwH on average, is much smaller than the interquartile range of the solar power measurements, which is 14.5. Thus the predictive accuracy of the model is excellent. The bias for the two models is comparable in magnitude, and the bias of 0.016 from our model means we overpredict, on average, by this amount. The linear model, on average, underpredicts by a slightly greater magnitude. This bias is very small in comparison to the spread of the solar power measurements, indicating excellent predictive accuracy. The average prediction interval width of our model is slightly larger than the IQR, which is relatively large in comparison to the spread of the response variable. However, there is a large spread in the observed solar power values from 2017-2019, so the standard errors are likely not very far off what we would expect to see in comparison to what we have already observed. The coverage of 0.97 means that 97% of all the 95% prediction interval widths calculated actually contained the true observed value. This is close to 0.95, so it appears that our model is working as expected and predicting quite accurately.

Results

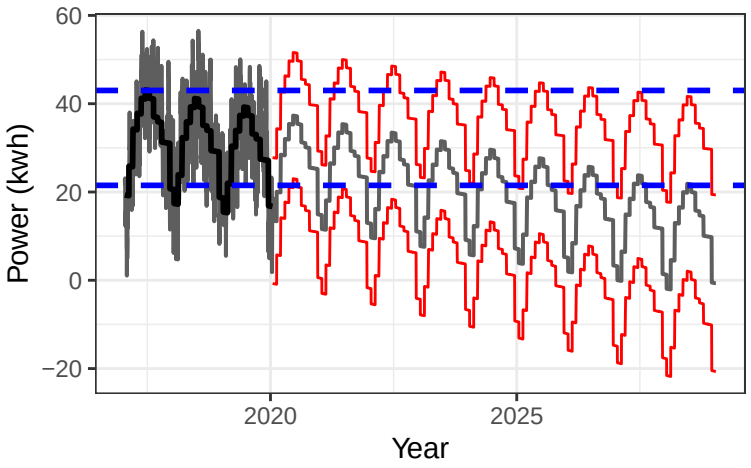
The table below displays estimates of each of the β_i values discussed in the Model Selection section, along with 95% confidence interval bounds.

	2.5%	**Estimate**	**97.5%**
Intercept	1089.040	3914.880	6740.721
Date	-3.331	-1.931	-0.531
February	2.924	6.682	10.440
March	11.055	15.247	19.438
April	14.652	18.984	23.316
May	18.195	22.547	26.899
June	20.211	24.597	28.983
July	18.498	22.879	27.261
August	14.449	18.844	23.239
September	13.470	17.894	22.319
October	9.038	13.443	17.848
November	9.106	13.462	17.818
December	-0.867	3.181	7.228

Based on these estimates, we conclude the following:

We can expect between a 0.531 and 3.331 kWh average daily decrease in generated power over time, assuming other factors are held constant. It is clear that the solar panels are degrading over time because the average daily power generated decreases over time.

Below is a plot displaying the solar power data along with our model's predicted average daily power measurements in kWh for the years 2020-2028. The red lines indicate the standard errors of our predicted power measurements, and we can see that the measurements become more uncertain as time goes on. It seems that the average maximum power generated each year (which appears to be in the middle of the year for all the predicted years) decreases over time. The top horizontal blue line indicates the maximum average daily power generated for the year 2017, based on the data and our model. This maximum appears to be about 40 kWh per day. The bottom blue line indicates the maximum average daily power generated in 2028, and this maximum appears to be about 20 kWh per day. Based on this, it appears that it takes approximately 11 years for the panels to lose 50% of their average power generating capability.



Though in the plot above we can see the predicted average daily power generated by these solar panels for the year 2020, the table below displays the average predicted daily generated power for each month of 2020 (the year just following our dataset). Because these are averages of predictions, they are in a sense "averages of averages", but they provide a snapshot of how much power these solar panels will generate for the next year. Average 95% prediction interval bounds are also included to get a sense of the uncertainty around each estimate.

	2.5%	**Estimate**	**97.5%**
January	-10.835	5.722	22.280
February	-4.386	12.251	28.888
March	4.008	20.647	37.286
April	7.578	24.226	40.875
May	10.981	27.626	44.271
June	12.867	29.517	46.168
July	10.991	27.636	44.282
August	6.793	23.440	40.086
September	5.679	22.332	38.985
October	1.064	17.717	34.370
November	0.899	17.578	34.256
December	-9.624	7.133	23.890

Conclusion

By modeling power by time along with a categorical month effect, as well as including a lag-1 autoregressive covariance structure, we were able to capture the variation in solar power measurements along with the

correlation between successive observations. The estimates of β, the vector of effects of each of the covariates in the model, provide an understanding of how the solar panels are degrading over time. We projected that it will take around 11 years for the solar panels to lose 50% of their power generating capability. The relevance of this projection depends on factors such as how much the panels cost to install/replace, how much money is actually saved by having the solar panels installed, and how well these panels perform relative to panels from other companies.

Though the model is able to predict quite accurately, it does have a shortcoming in the quality of the model fit. Using a "Pseudo R^2", we found that our covariates (linear time variable and categorical month variable) explained about 53% of the variation in generated power under the traditional R^2 interpretation. This could potentially be improved by including other basis function expansions in $\mathbf{X}$ and running a Lasso regression as a means of variable selection. Another option is to use only Time as the covariate and apply a more complicated correlation structure that accounts for the obvious summer-winter seasonality that shows up in this data. Also, these data were gathered from a single customer household, which probably does not give the power company the ability to generalize these results to their entire population of customers. Data from more households would provide more insight into the function of the actual product because variation attributed to individual homes would be ironed out as the sample size (of homes) became sufficiently large.

A final point of further analysis may be to gather outside covariates that might help explain sunlight (and thus the amount of power generated by the solar panels) to see how variables other than time, and basis expansions of time, affect the generating power of these solar panels. This would allow the customer to get further insight into the financial benefit that the solar panels provide and give the power company insight into how well the solar panels actually function.